Enhancing Option Pricing and Stock Return Predictions: Integrating Machine Learning with Firm Characteristics and Option Greeks

LI, NAN

Enhancing Option Pricing and Stock Return Predictions: Integrating Machine Learning with Firm Characteristics and Option Greeks

Author: Nan Li

Thesis submitted for the degree of
Doctor of Philosophy in Finance

Supervisors:
Dr. Chulwoo Han, Prof. Panayiotis C. Andreou, Prof. Damian Damianov

2024

Durham, United Kingdom

Abstract

This thesis explores the use of machine learning in financial derivatives, particularly stock options, to improve understanding and prediction of option pricing. It includes three empirical chapters. Chapter 1 evaluates machine learning models that integrate various firm characteristics to predict stock option prices. It introduces two semi-parametric models: a variant of Andreou, Charalambous, and Martzoukos (2010) generalized parametric function model (GPF) and Lajbcygier and Connor (1997)'s hybrid model (HBD), applied to U.S. stock options from 1996 to 2021. The GPF model consistently outperforms the HBD model, with specific firm characteristics emerging as key predictors of option prices. Chapter 2 explores the predictive capacity of option market characteristics, especially implied volatility and Greeks, in forecasting extreme stock returns of the underlying assets. The study employs the LightGBM algorithm, which significantly outperforms traditional logistic regression in predicting stock market trends, emphasizing the value of a comprehensive approach to option dynamics. Chapter 3 builds upon the insights from Chapter 1, focusing on an in-depth analysis of option Greeks and specific firm characteristics within three semi-parametric frameworks: GPF, HBD, and AFFT (Almeida, Fan, Freire, and Tang, 2023). This chapter also explores how these three frameworks maintain consistency with various input features throughout the Pandemic period. Notably, the GPF framework shows exceptional resilience and adaptability when integrated with option Greeks and firm characteristics during the Pandemic. Overall, this thesis underscores the efficacy of incorporating firm characteristics and option Greeks in option pricing and stock return prediction, highlighting the superiority and adaptability of machine learning models in volatile market scenarios.

Keywords: Machine learning; Option pricing; Stock extreme return; Firm characteristics; Option Greeks; Pandemic analysis; Portfolio management.

Declaration

I, Nan Li, declare that this thesis titled, "Enhancing Option Pricing and Stock Return Predictions: Integrating Machine Learning with Firm Characteristics and Option Greeks", and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree at Durham University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. Except where referenced otherwise, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Acknowledgements

Lifelong learning defines my doctoral journey, culminating in a richly rewarded period.

I am deeply grateful to my supervisors for their invaluable advice, support and wisdom.

Thanks to Chulwoo Han, Panayiotis Andreou, and Damian Damianov, I appreciate it.

Acknowledging my wife, Hongyan Tao, for her enduring love and unwavering support.

Owing deep gratitude to my parents, Fujiang Li and Weiyun Zhang, for their support.

Many thanks to friends Zhaodong He and Xinyi Zhang, for their cherished friendship.

Appreciating everyone who is involved in this journey, I extend my deepest gratitude.

Overall, I am sincerely grateful to all, and looking forward to the future with optimism.

Contents

List of Figures

List of Tables

Chapter 0

Introduction

0.1 Introduction

With the advent of machine learning techniques, the field of finance, particularly option pricing, is evolving rapidly. Traditional models like the Black-Scholes (BS) model, though foundational, have been challenged by the complex realities of financial markets, including the non-normality of stock returns and time-varying volatility. The incorporation of machine learning in this context offers a novel perspective, addressing the limitations of traditional parametric models by capturing nonlinear and intricate relationships between option prices and their underlying factors.

This thesis aims to explore the integration of machine learning methods in the realm of financial derivatives, focusing on stock options. The objective is threefold: to assess the effectiveness of machine learning models in pricing options, to explore the predictive power of option characteristics (such as implied volatility and Greeks) for stock returns, and to evaluate the performance of machine learning structures in option pricing during periods of heightened market volatility, like the Pandemic.

0.1.1 Introduction of 'Stock options pricing via machine learning combined with firm characteristics'

Option valuation is a pivotal area in finance, attracting both academic and practitioner interest. The Black-Scholes model, despite its limitations like assuming constant volatility and normal distribution of stock returns, remains a cornerstone in this field. However, models accommodating the non-normality of returns and time-varying volatility, such as those by Heston (1993) and Corrado and Su (1996), provide more realistic assumptions.

Traditional option pricing has evolved from models like Black-Scholes (Black and Scholes, 1973) and binomial models (Cox, Ross, and Rubinstein, 1979) to more complex ones like the LSM algorithm (Samimi, Mardani, Sharafpour, and Mehrdoust, 2017) and Markov chain techniques (Duan and Simonato, 2001). However, these often struggle with real-world market behaviours due to their reliance on rigid assumptions that may not align with the dynamic and unpredictable nature of financial markets. Machine learning offers an alternative, adaptable approach to option pricing by providing data-driven, non-parametric models (as introduced by Hutchinson, Lo, and Poggio (1994)) and semi-parametric models that integrate machine learning with traditional parametric models, thereby offering a dual advantage. Semi-parametric models like the GPF (Andreou et al., 2010) and HBD (Lajbcygier and Connor, 1997) structures use machine learning to enhance traditional models' inputs or correct pricing errors. Recent machine learning advancements have shown promise in derivative pricing, with models like generative Bayesian learning (Jang and Lee, 2018) and support vector regression (Madhu, Rahman, Mukherjee, Islam, Roy, and Ali, 2021) improving accuracy and efficiency.

Moreover, a growing body of research underscores the impact of firm characteristics on option pricing. Research by Rubinstein (1983) and Figlewski (1989) note how individual asset risks, including firm debt and dividend policy, and market imperfections like uncertain volatility and transaction costs, affect option pricing. Subramanian (2004) extend this to options in mergers and acquisitions, suggesting firm-specific events significantly influence

pricing. Recently, Trigeorgis and Lambertides (2014); Andreou (2015); Vasquez and Xiao (2023); Chen, Guo, and Zhou (2023) demonstrate the influence of firm characteristics on asset dynamics and option values. Factors such as firm-specific business volatility, managerial flexibility, market default risk, firm leverage, asset volatility, and firm fundamentals significantly shape option valuation and implied volatility curves. Additionally, Zhan, Han, Cao, and Tong (2022) highlights the correlation between firm characteristics and delta-hedged equity option returns, emphasizing the relevance of these factors in predicting market behaviour and informing investment strategies. These findings inspire our exploration of an option pricing model that extends beyond standard variables to include a broad set of firm characteristics, potentially indicative of a firm's financial health and future performance.

In this chapter, we incorporate a wide range of firm characteristics into option pricing models, using advanced machine learning algorithms, with the objective of utilizing historical data, such as realized volatility and other relevant factors, to accurately determine current option prices. Specifically, we compare two semi-parametric methods - a generalized parametric function (GPF) model Andreou et al. (2010) and a hybrid (HBD) model Lajbcygier and Connor (1997) - against a parametric benchmark model. Our data comprises 15.2 million U.S. stock options from 1996 to 2021, with 111 firm characteristics (Jensen, Kelly, and Pedersen, 2021) considered.

We find that semi-parametric models outperform the benchmark even before incorporating firm characteristics. Moreover, adding firm characteristics further enhances their performance. Notably, conditional skewness, Dimson Beta, dividend yield, one-year-lagged annual stock return, and downside beta are among the most important firm characteristics in option pricing.

Our research makes significant contributions by integrating a wider range of firm characteristics into option pricing models than has been previously done, and by evaluating the performance of machine learning-based models on individual stock options. By addressing challenges in parametric models and leveraging machine learning, we provide a comprehen-

sive and nuanced framework for option pricing in modern financial markets. To make the model more practical, our study demonstrates that a model incorporating the five most crucial firm characteristics within the GPF framework significantly improves root mean square error by 15%, approaching the 19% improvement seen with models that use a comprehensive set of 111 characteristics. Although models with fewer characteristics are not as accurate, they are simpler to implement and reduce uncertainties associated with data collection. Furthermore, we observe that pricing performance remains stable even when the training/validation dataset is limited by option type, maturity, and moneyness. Smaller datasets not only expedite the training time of machine learning algorithms but also speed up the entire pricing process.

0.1.2 Introduction of 'Can option characteristics provide leading information about stocks' extreme returns?'

This study investigates the potential of option-implied volatility and Greeks (Delta, Gamma, Theta, and Vega) in predicting stock returns using machine learning. It builds on extensive research suggesting that option characteristics, particularly implied volatility, contain relevant information for forecasting stock returns. Studies like Bates (1991) and Ofek, Richardson, and Whitelaw (2004) indicate that higher implied volatilities in options, especially out-of-the-money puts, often precede significant market movements. Moreover, the Greeks serve as vital tools in risk management and portfolio optimization, with research by Cao and Han (2013) and Zhan et al. (2022) indicating their influence on stock returns.

Our research utilizes U.S. stock and option market data from January 1996 to December 2022, focusing on the predictive power of these option characteristics for stock returns. We categorize stocks as experiencing jumps, crashes, or normal conditions and create portfolios based on these predictions. Machine learning, particularly the gradient-boosting LightGBM model, is employed to analyze the data, exploring the intricate relationships between option characteristics and stock market returns. This approach is compared against traditional

logistic regression and control variables used in prior studies like Chen, Hong, and Stein (2001) and Jang and Kang (2019).

The results indicate that implied volatility and Greeks are significantly useful features for predicting extreme returns. Statistically, the LightGBM model, which integrates control variables, implied volatility, and Greeks (LGBM-$SDGTV$), outperforms the control variable-only benchmarks (LGBM-B) in forecasting extreme stock returns, demonstrating higher accuracy with impressive AUCs of 0.766, and 0.755 for jumps, and crashes, respectively, compared to the benchmark's AUCs of 0.687, and 0.672. Financially, portfolios based on the LightGBM model's predictions significantly outperform traditional market portfolio benchmarks in Sharpe ratio and average annual return, with the LGBM-$SDGTV$ model achieving Sharpe ratios of 1.42 and 1.58, and average annual returns of 33% and 31% for value-weighted and equal-weighted portfolios, respectively, compared to the benchmark portfolio's more modest Sharpe ratios of 0.75 and 1.22 and returns of 20% and 23%.

In summary, this chapter highlights the critical role of option characteristics in forecasting extreme stock returns. It shows that machine learning, especially models like LightGBM, can effectively harness these indicators, offering significant advancements in stock market prediction. Our findings not only confirm the predictive power of option market indicators but also open new possibilities for integrating machine learning in financial analysis and portfolio management. In terms of practical application, Our method for standardizing option characteristics examines all market options, effectively avoiding selection bias (Byun and Kim, 2016; Zhan et al., 2022; Bali, Beckmeyer, Moerke, and Weigert, 2023). However, given our portfolio's emphasis on extreme returns, higher standard deviations and maximum drawdowns are inevitable.

0.1.3 Introduction of 'Comparative analysis of stock option pricing: Machine learning and Pandemic impact'

In this chapter, we extend the discussion of 'Stock options pricing via machine learning combined with firm characteristics, focusing on a comparative analysis of three distinct machine learning architectures: GPF (Andreou et al., 2010), HBD (Lajbcygier and Connor, 1997), and AFFT (Almeida, Fan, Freire, and Tang, 2022). We examine each architecture through four submodels that incrementally integrate various inputs: basic factors, option sensitivities (Greeks), firm characteristics, and a combination of these elements. A key part of our analysis is the assessment of the resilience of these pricing models during the Pandemic, characterized by extreme market volatility, specifically evaluating their effectiveness and accuracy in adapting to rapid market changes and providing reliable predictions during this tumultuous period.

Our exploration revolves around the incorporation of option Greeks and firm characteristics into semi-parametric option pricing models. We test the three aforementioned semi-parametric models against a parametric benchmark model (PARA) using a dataset of U.S. stock options belonging to the stocks listed in the SP500. The primary focus is on evaluating how the inclusion of option Greeks (Delta, Gamma, Theta, Vega) and firm characteristics (based on the criteria set by Jensen et al. (2021)) can enhance the predictive accuracy of these models.

We find that all semi-parametric models outperform the benchmark model, with improvements in their root mean squared errors (RMSEs) when option Greeks and firm characteristics are incorporated. The GPF model, in particular, shows the best performance, especially during the Pandemic. This could be due to the narrower range of implied volatility compared to the option price residual, leading to more stable predictions. The models' performance is consistent across various subsets of options and stocks with different return patterns, although stocks with extremely high or low returns show a higher mean absolute percentage error (MAPE).

In practical terms, beyond the elements discussed in Chapter 1, integrating option Greeks into the model enhances its applicability for short-term option pricing. The daily updates of option Greeks improve access and the model's ability to reflect current market conditions. Using only option Greeks as features in the machine learning algorithm can greatly decrease training time, though their time-sensitive nature may sometimes limit their effectiveness.

This research enhances semi-parametric option pricing models by incorporating option Greeks and firm characteristics, and evaluating their effects on pricing predictions over different timeframes. It compares the GPF and AFFT models, emphasizing the superior efficiency of the GPF. The models utilize daily updates of option Greeks for real-time market dynamics and firm characteristics for in-depth asset analysis, aiming to boost accuracy under diverse market conditions. The study aims to provide a comprehensive tool for option pricing that merges the immediate relevance of option Greeks with insights from firm characteristics, improving model selection and applicability across various market situations.

0.1.4 Introduction to the machine learning methods used in the thesis

In this thesis, we utilize LightGBM (Ke, Meng, Finley, Wang, Chen, Ma, Ye, and Liu, 2017), a gradient-boosting framework that uses decision trees for both classification and regression tasks. Its advantages, especially in finance, include speed and low memory requirements (Chang, Chang, and Wu, 2018; Basak, Kar, Saha, Khaidem, and Dey, 2019; Sun, Liu, and Sima, 2020; Ivașcu, 2021). LightGBM's unique leaf-wise tree growth leads to faster convergence and less resource use, making it ideal for managing large and complex datasets common in options research. Additionally, its default hyperparameters simplify replication and generalization, while advanced tuning can further adapt the model to specific needs.

Previous studies have primarily employed deep neural networks (DNNs) (Hornik, Stinchcombe, and White, 1989), support vector machines (SVMs) (Hearst, Dumais, Osuna, Platt, and Scholkopf, 1998), and random forests (RFs) (Breiman, 2001), demonstrating their effectiveness in financial analysis (Tay and Cao, 2001; Feng, He, Polson, and Xu, 2018; Gu,

Kelly, and Xiu, 2020; Han, 2021; Bali et al., 2023; Chen, Pelger, and Zhu, 2024). However, each method has limitations for this thesis's specific context. DNNs, while adept at detecting complex data patterns, require significant computational power and large datasets, and are prone to overfitting with extended training durations. SVMs, although effective in classification and regression, struggle with large-scale financial data, increasing computational demands. RFs are reliable for similar tasks but are comparatively slower and more memory-intensive than alternatives like LightGBM, with their level-wise tree growth leading to longer training periods.

Selecting LightGBM for this thesis considers factors like model complexity, computational speed, and handling large datasets typical in option markets. LightGBM excels in quickly processing extensive financial data and managing complex, non-linear relationships. It also effectively addresses multicollinearity in datasets by choosing the most relevant features during tree splitting, minimizing redundancy and overfitting. This ability improves LightGBM's accuracy and reliability with highly correlated data, such as firm characteristics, providing a balanced approach to sophisticated modelling within computational and data complexity constraints.

Chapter 1

Stock options pricing via machine learning combined with firm characteristics

Abstract

This paper proposes machine learning-based option pricing models that incorporate firm characteristics. We employ two semi-parametric models, one that uses machine learning to predict the implied volatility and a hybrid that corrects the pricing error of the binomial model, and use 111 firm characteristics as well as option-related variables as the input features. Our empirical analysis is conducted using a sample including 15,247,956 stock option observations in the period January 1996 to December 2021. We find that both semi-parametric models outperform the parametric one even without firm characteristics, whilst firm characteristics significantly enhance the performance of these models. Conditional skewness, Dimson Beta, dividend yield, one-year-lagged annual stock return, and downside beta are found to be the most important features.

Keywords: Option pricing; Semi-parametric; Firm characteristics; Machine learning.

1.1 Introduction

Option valuation has garnered significant attention in the field of finance, both from theoretical and practical perspectives. Academics are intrigued by the potential for a robust option pricing model to shed light on financial market operations, while market makers aim to use an efficient pricing model to determine prices in the derivatives market. The Black-Scholes (BS) model (Black and Scholes, 1973) is the earliest and most well-known model for pricing the European options. Despite its simplicity, it provides a good estimate of option prices and remains as one of the most important option pricing models. The BS model relies on the assumption that stock returns are normally distributed and have constant volatility over time. However, there is abundant evidence that stock returns have a fat-tailed distribution and exhibit time-varying volatility. To account for the non-normality of stock returns, Corrado and Su (1996) augment the BS model with skewness and kurtosis. To account for the time-varying volatility, Heston (1993) and Hagan, Kumar, Lesniewski, and Woodward (2002) propose stochastic volatility models, which assume the volatility to be a random variable. These models allow a more realistic representation of the underlying asset's volatility and provide more accurate pricing of options. For American options, which face premature exercise and dividend payment issues, tree-based (Cox et al., 1979; Boyle, 1986) or Monte Carlo simulation-based (Broadie and Glasserman, 1997; Longstaff and Schwartz, 2001; Andersen and Broadie, 2004) methods have been proposed. A closed-form solution can also be obtained for an American option in case of known absolute dividends (Roll, 1977; Geske, 1979; Whaley, 1981), or for options with short or long maturity (Barone-Adesi and Whaley, 1987), or for options with proportional dividends (Villiger, 2006).

The majority of option pricing research including those mentioned above adopts a parametric method, *i.e.*, they assume a certain distribution for the underlying asset return and derive the fair price of the option under the no-arbitrage condition. While parametric models are preferable as they are established on a solid economic foundation and often allow an analytic option pricing formula, the reality could deviate from their underlying assumptions, and parametric models often fail to accurately fit the actual option prices observed in the market.

More recently, machine learning-based models have been proposed as an alternative approach.

These models have the ability to capture the nonlinear and complex relations between option prices and their underlying factors, making them more flexible than traditional parametric models. Machine learning-based models can be categorized into two types: non-parametric and semi-parametric models. Ruf and Wang (2019) provide a comprehensive review of the application of neural networks on option pricing.

Non-parametric models are agnostic about the economic theory behind option pricing and predict the option price by learning purely from the data the relations between the input variables and the option price. One of the earliest studies in this category is the work of Hutchinson et al. (1994), which treats option pricing as a regression problem. They demonstrate that the learning network is superior to traditional parametric methods in option valuation. With advances in machine learning algorithms, researchers have proposed ways to improve the generalization of non-parametric methods, such as Bayesian adjustment, early stopping, and bagging, which allow more robust pricing of options (Gençay and Qi, 2001). For example, Gradojevic, Gençay, and Kukolj (2009) use modular neural networks to improve prediction performance, while Liang, Zhang, Xiao, and Chen (2009) use a combination of neural networks and support vector regression to reduce pricing errors in traditional option pricing methods such as Monte Carlo simulation, binomial trees, and finite difference methods. Park, Kim, and Lee (2014) demonstrate that the Gaussian process model significantly outperforms parametric methods in both in-sample and out-of-sample pricing of KOSPI 200 Index options. Liu, Oosterlee, and Bohte (2019) introduce an Artificial Neural Network (ANN)-based, data-driven method to accelerate financial option valuation and implied volatility calculations, demonstrating through testing on various solvers that the trained ANN significantly reduces computation time. Buehler, Gonon, Teichmann, and Wood (2019) apply reinforcement learning to price and hedge derivative contracts, including options. Ruf and Wang (2022) propose a neural network designed to generate a hedging strategy for options, focusing on minimizing hedging errors rather than pricing errors.

Whilst non-parametric models can fit the data flexibly without making any assumptions about the function underlying option price, the very flexibility can be toxic and expose the models to the risk of overfitting. Semi-parametric models address this risk by combining a parametric model with machine learning. Guided by economic theory, a semi-parametric model can fit the data with a

more parsimonious structure. One approach is to employ machine learning to predict unobservable variables such as implied volatility, which are then used as input to a parametric model. Andreou, Charalambous, and Martzoukos (2006) and Andreou et al. (2010) predict volatility for the BS model and volatility, skewness, and kurtosis for Corrado and Su (1996)'s model via a neural network and use them as input to the corresponding parametric option pricing model. Wang (2009) apply a novel hybrid approach, merging the grey forecasting model with GARCH, to estimate volatility in a neural network option-pricing model, demonstrating that this hybrid method can effectively forecast derivative securities prices and outperforms other approaches in the neural network option-pricing model. Audrino and Colangelo (2010) introduce a semi-parametric technique for predicting implied volatility surfaces, utilizing a regression tree as the foundational model and progressively reducing the difference between model predictions and actual values by incorporating additional trees. Zheng, Yang, and Chen (2019) introduce a gated deep neural network model for predicting implied volatility surfaces, incorporating conventional financial conditions and empirical volatility evidence, demonstrating superior performance over the commonly used surface stochastic volatility model. Lajbcygier and Connor (1997) employ neural networks to correct the pricing error of the BS model.

An advantage of machine learning is that it can accommodate any features that potentially carry information about the option price. The conventional theory suggests that option prices are primarily determined by standard variables; however, emerging research indicates that additional factors, such as investor sentiment and firm characteristics, also play a critical role.

The impact of firm characteristics on option pricing within an arbitrage-free framework is an intricate area of financial study. Several research papers provide insights into this topic. Rubinstein (1983) develop an option pricing formula considering the risk of individual assets of the firm, encompassing differential riskiness and effects of firm debt and dividend policy. Figlewski (1989) highlight that market imperfections, such as uncertain volatility and transaction costs, limit the practical application of arbitrage in option pricing. This underscores the potential impact of firm-specific factors, like financial stability and market behaviour, on option prices in reality, beyond idealized arbitrage-based models. Subramanian (2004) develop an arbitrage-free framework for pricing options on stocks of firms in mergers and acquisitions, accounting for discontinuous impacts

on stock prices. This model outperforms the Black-Scholes model in explaining observed option prices, indicating that firm-specific events like mergers can significantly influence option pricing in an arbitrage-free context. Recently, Trigeorgis and Lambertides (2014); Andreou (2015); Vasquez and Xiao (2023); Chen et al. (2023) demonstrate the significant influence of firm characteristics on underlying asset dynamics and option values. Trigeorgis and Lambertides (2014) emphasize the importance of firm-specific business volatility and managerial flexibility in growth option assessment. Market default risk, as shown by Andreou (2015), and firm leverage and asset volatility, as highlighted by Vasquez and Xiao (2023), are directly linked to option pricing. Chen et al. (2023) document how firm fundamentals shape the implied volatility curve, crucial in option valuation. Additionally, Zhan et al. (2022) illustrate the significant correlation between firm characteristics and delta-hedged equity option returns, further underscoring the relevance of these factors not only in option pricing but also in predicting market behaviour and investment strategies. These findings inspire our exploration of an option pricing model that extends beyond standard variables to include a broad set of firm characteristics, potentially indicative of a firm's financial health and future performance.

Our research pioneers the integration of an extensive array of firm characteristics with advanced machine learning algorithms to ascertain their utility in option pricing. We are among the first to systematically analyze and test a wide spectrum of firm characteristics to determine which are most predictive of option values by using semi-parametric methods. This innovative approach not only underscores the significance of firm characteristics in option valuation but also sets a precedent for future explorations into more sophisticated and nuanced financial models.

We employ two semi-parametric methods to employ firm characteristics. The first model is based on Andreou et al. (2010)'s generalized parametric function (GPF) model, which employs machine learning for the prediction of implied volatility. The second model is based on Lajbcygier and Connor (1997)'s hybrid (HBD) model, which employs machine learning for pricing error correction. These models are compared against a parametric benchmark model. We select the binomial model with lag-implied volatility (BI) as our benchmark, as it utilizes the same input variables as the semi-parametric methods.

We evaluate the models using 15,247,956 U.S. stock options from January 1996 to December

2021. As to the firm characteristics, we choose 111 firm characteristics from Jensen et al. (2021) after removing firm characteristics with many missing values. Specifically, we exclude firm characteristics with missing values greater than 20% over the sample period. We first find that the semi-parametric models, GPF and HBD, outperform the benchmark, BI, even before incorporating firm characteristics. The root mean squared errors (RMSEs) of GPF and HBD are respectively 1.701 and 1.710, whereas that of BI is 2.023. Second, firm characteristics can further improve the performance of GPF and HBD: the RMSEs of the two models are reduced to 1.376 and 1.459, respectively. GPF consistently renders a smaller error than HBD. We conjecture that this is because the range of implied volatility is narrower than the range of option price residual, making a more stable prediction possible. We assess the models with various subsets of options defined by the option type, time-to-maturity, and moneyness, and find that the models perform consistently across these options groups. The usual option features such as time-to-maturity, strike price, lagged implied volatility and underlying price are the most important features for pricing options, but firm characteristics are also deemed to play a non-trivial role. Firm characteristics collectively have a feature importance score of 49.19 out of 100 in GPF. In particular, conditional skewness, Dimson Beta, dividend yield, year-1-lagged annual return, and downside beta turn out to be the most important firm characteristics in option pricing.

The study on using firm characteristics in machine learning models for option pricing, particularly with the GPF framework, reveals that including key firm features—conditional skewness, Dimson Beta, dividend yield, one-year-lagged annual return, and downside beta—enhances model accuracy by 15% in root mean square error. This is close to the 19% improvement seen with models that incorporate a wider set of 111 firm characteristics. These results suggest that models can be simplified effectively without major loss in predictive accuracy, balancing complexity and operational efficiency. However, limiting firm characteristics can neglect important details that might impact option pricing in certain markets or options, potentially overlooking critical market dynamics.

The model's flexibility with smaller, specific datasets—sorted by option type, maturity, and moneyness—highlights its effectiveness and efficiency. Training the algorithms on focused datasets ensures accurate pricing predictions and speeds up computation. However, this method may com-

promise dataset representativeness, as using limited data might not fully reflect market dynamics, challenging the accuracy of option price predictions in volatile markets. Operational issues arise due to reliance on accurate and complete data collection about firm characteristics. Faulty or partial data can distort model results, highlighting the need for strict data management. Moreover, although using smaller datasets can improve efficiency, it may restrict the model's learning ability and its capacity to perform well across various market conditions. Future efforts should aim to refine these models, broaden the scope of firm characteristics examined, and improve data handling techniques. Tackling these challenges is crucial for enhancing model robustness and adaptability in dynamic market environments.

Above all, the first contribution of our paper is the integration of an extensive and diverse set of firm characteristics into option pricing models, using 111 firm characteristics compared to previous studies like Zhan et al. (2022) which employ 10 firm characteristics and Chen et al. (2023) which employ 94 firm characteristics. This extensive range of characteristics offers a more nuanced understanding of the factors influencing option pricing, extending beyond traditional models. Our second contribution lies in evaluating the performance of machine learning-based models for individual stock options. We underscore the challenges in pricing stock options with parametric models, notably due to their inability to fully capture the nuances of investor perspectives on firm growth and the impact of firm-specific factors. Our methodology is aligned with the work of Chen et al. (2023), who employed machine learning tools to document the influence of firm fundamentals on the shape of the option implied volatility curve, both economically and statistically. They utilized a LASSO approach to select relevant firm fundamentals from a large set, thus avoiding overfitting issues common in such analyses. Inspired by their methodology, we utilize LightGBM, a highly efficient algorithm in managing multicollinearity, as it employs a histogram-based approach during training, enabling independent consideration of features and reduced sensitivity to multicollinearity. This choice enhances our ability to dissect the complex relationships in financial data, potentially offering a more nuanced understanding of how firm fundamentals impact the implied volatility curve. Our research advances the understanding of option pricing by incorporating a broader range of firm characteristics than has been previously considered. By leveraging machine learning techniques and addressing the complexities introduced by default risk and firm-specific

factors, we provide a more nuanced and comprehensive framework for option pricing that aligns with current financial market complexities.

This paper is structured as follows. Section 1.2 describes the option pricing models used in this study, Section 1.4 details the data and methodology for the empirical analysis, Section 1.5 presents the empirical results, Section 1.6 presents robustness tests, and Section 1.7 concludes.

1.2 Models

1.2.1 Parametric model

The binomial option pricing model (Cox et al., 1979) is a widely used method for calculating the fair value of an option (Rubinstein, 1994; Broadie and Detemple, 1996; Jiang and Dai, 2004). It is based on the assumption that the underlying asset price can only move up or down by a fixed percentage at each time step. The model is widely used for pricing European-style options (exercisable only at expiration), it can also be adapted for American-style options, which can be exercised at any point before expiration.

In the case of American-style options, the binomial option pricing model requires a modification to account for the possibility of early exercise. At each time step, the option holder has the option to exercise the option and receive the intrinsic value of the option, which is the difference between the current underlying asset price and the strike price. The option holder may choose to exercise the option if the intrinsic value is greater than the current value of the option as determined by the binomial tree model.

To calculate the fair value of an American option using the binomial model, we follow the following steps provided by Cox et al. (1979):

1. Calculate the size of each time step (dt) using the following formula:

$$dt = \frac{T}{M}, \tag{1.1}$$

where T is the maturity, M is the total number of time steps[1].

[1] For the purpose of saving computation time, we set the value of M to 100.

2. Calculate the up factor (u) and the down factor (d) using the following formulas:

$$u = \exp(\sigma \times \sqrt{dt}) \tag{1.2}$$

$$d = \frac{1}{u}, \tag{1.3}$$

where σ is the volatility of the underlying asset.

3. Calculate the probability of an up move (p) using the following formula:

$$p = \frac{r \times dt - d}{u - d}, \tag{1.4}$$

where r is the risk-free interest rate, q is the dividend yield.

4. Calculate the value of the option at expiration for each possible price of the underlying asset. For an underlying asset price S with dividend rate q [2] and strike price X at a given point in time, the price of a call option is equal to $max(S \cdot (1 - q) - X, 0)$, and the price of a put option is equal to $max(X - S \cdot (1 - q), 0)$.

5. Working backwards from expiration, calculate the value of the option at each time step for each possible price of the underlying asset using the following formulas:

$$Price_u = \frac{p \times Price_{uu} + (1 - p) \times Price_{ud}}{1 + rf}, \tag{1.5}$$

$$Price_d = \frac{p \times Price_{du} + (1 - p) \times Price_{dd}}{1 + rf}, \tag{1.6}$$

where $Price_{uu}$, $Price_{ud}$, $Price_{du}$, and $Price_{dd}$ are the option values if the underlying asset price goes up twice, up then down, down then up, and down twice, respectively.

6. At each time step, the option's value is determined by choosing the higher of the two: the immediate payoff from exercising the option (intrinsic value[3]) or the expected value of holding

[2] We adopt the Cox et al. (1979) framework for extending the binomial tree model to dividend-paying stocks, assuming that the stock exhibits a constant yield, denoted as q, on each ex-dividend date.

[3] The intrinsic value of a call option is equal to $max(S - X, 0)$, and the intrinsic value of a put option is equal to $max(X - S, 0)$

the option (theoretical value [4]). This choice accounts for the flexibility American options offer in terms of early exercise. This evaluation is key due to the option's early exercise feature. We assess at each step whether exercising the option for its intrinsic value is preferable over retaining it for potential future value, considering the risk-free rate and the likelihood of price movements.

To find the market implied volatility for option i on t, $\sigma_{i,t}^{mrk}$, we solve an optimization problem that has the following form:

$$\sigma_{i,t}^{mrk} = \arg\min_{\sigma_{i,t}^{mrk}} [P_{i,t}^{mrk} - BI(\sigma_{i,t}^{mrk}, S_{i,t}, X_{i,t}, T_{i,t}, rf_{i,t})]^2, \tag{1.7}$$

where $P_{i,t}^{mrk}$ is the market price for option i on t, $BI(\sigma_{i,t}^{mrk}, S_{i,t}, X_{i,t}, T_{i,t}, rf_{i,t})$ is the binomial model that employs the parameters $S_{i,t}$, $X_{i,t}$, $T_{i,t}$, and $rf_{i,t}$, which are the stock price, the strike price, the maturity, and the risk-free rate for option i on t.

To estimate the price of option i on day t, we use the binomial model that employs σ_t^{avg} as our benchmark model (BI), where σ_t^{avg} is the average of the implied volatilities of all the options with the same underlying asset in the period from $t-10$ to $t-1$ (i.e., over the past 10 trading days) as Equation (1.8).

$$\sigma_{i,t}^{avg} = \frac{1}{10} \sum_{j=1}^{j=10} \sigma_{i,t-j}^{imp}, \tag{1.8}$$

where the $\sigma_{i,t-j}^{avg}$ is the implied volatility of option i on day $t-j$ where $j \in [1,2,3,4,5,6,7,8,9,10]$.

1.2.2 GPF model

Andreou et al. (2010) propose a semi-parametric option pricing model (GPF model) and show that it exhibits an enhanced pricing performance when applied to index options. The GPF model predicts unobservable option variables (volatility for the BS model and volatility, skewness, and kurtosis for the Corrado and Su (1996) model) via a neural network and uses them as input to the corresponding parametric option pricing model. An advantage of the GPF model is that it combines the benefits of both parametric and non-parametric methods. The parametric model provides a

4 $\frac{p \cdot Price_u + (1-p) \cdot Price_d}{1+rf}$

theoretical framework for option pricing, while the machine learning algorithm generates a more accurate prediction of the input variables. By combining these two approaches, the GPF model can effectively address the limitations of traditional parametric and non-parametric methods in option pricing. The GPF model is flexible and can be adapted to meet the needs of different option pricing scenarios, making it a versatile tool for option pricing research.

For our study, we leverage the inherent flexibility of the GPF by employing firm characteristics into the prediction process as Equation (1.9).

$$P_{GPF}^{pre} = BI(\sigma^{pre}, S, X, T, q, rf), \tag{1.9}$$

where P_{GPF}^{pre} is the predicted price through the GPF structure, S is the close price of the underlying asset, X is the strike price, T is the maturity, q is the dividend rate, rf is the risk free rate, and σ^{pre} is the predicted implied volatility obtained through Equation (1.11). The function $BI(\cdot)$ represents the process of the binomial tree model.

The original GPF model employs a one-step method, training the machine learning algorithm to minimize option pricing error through a loss function defined by the mean squared error of the option prices. In our approach, tailored to accommodate the extensive dataset of stock options, we optimize the loss function based on implied volatility. This refined focus allows for a faster and more relevant minimization of pricing errors, particularly suited to the scale and nature of the data we handle. The loss function we optimize in the GPF is shown in Equation (1.10).

$$Loss(GPF)_t = \min \sum_{i=1}^{N} (\sigma_{i,t}^{mrk} - \sigma_{i,t}^{pre})^2, \tag{1.10}$$

where t is the date of the training, i is the i-th observation, N represents total observations on t. $\sigma_{i,t}^{mrk}$ represents the market implied volatility of option observation on t, and the $\sigma_{i,t}^{pre}$ represents the predicted implied volatility of option observation i on t.

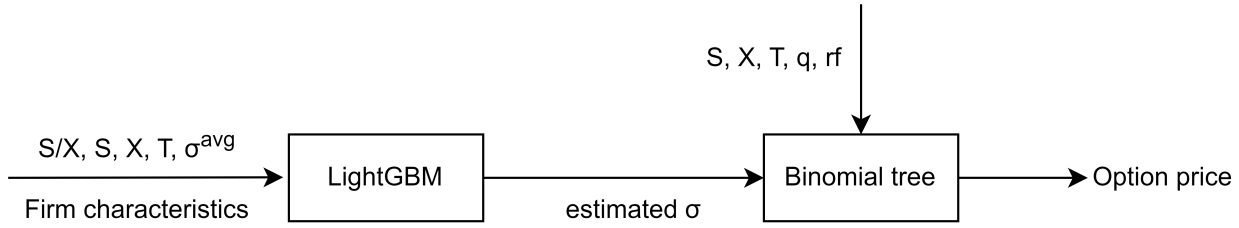
The estimation of the σ^{pre} can be summarized as follow:

$$\sigma_{i,t}^{pre} = \text{LightGBM}(F_{i,t}; \theta), \tag{1.11}$$

where, $\sigma_{i,t}^{pre}$ is the predicted implied volatility for the i-th observation on t, as determined by the LightGBM model. $F_{i,t}$ represents the vector of input features for the i-th observation, such as the moneyness (S/X), close price (S), strike price (X), maturity (T), σ^{avg}, and firm characteristics. θ denotes the set of parameters of the LightGBM model.

Our study enhances the GPF by incorporating firm characteristics, leveraging its ability to integrate relevant features and enriching its practicality. We replace the neural network with LightGBM, a gradient-boosting method known for its superior performance in diverse machine-learning applications, thus evolving the GPF into a more comprehensive and precise predictive tool. The schematic structure of our GPF model is depicted in Figure 1.1.

Figure 1.1. GPF model structure.



1.2.3 Hybrid model

The hybrid model (HBD model) proposed by Lajbcygier and Connor (1997) also combines the advantages of a parametric option pricing model and machine learning to improve option pricing accuracy. However, it is different from GPF in that it predicts the pricing error of a parametric model via machine learning. HBD first calculates the option price using a parametric model such as the binomial model (We use the lagged implied volatility as an input to the binomial model as a parametric model for the HBD.). This model price is then contrasted with the market price to obtain the pricing error (residual). A machine learning algorithm is then employed to predict the residual. The final option price is the sum of the model price and the predicted residual. We summarize the HBD as follows.

$$P_{HBD}^{pre} = Res^{pre} + BI(\sigma^{avg}, S, X, T, q, rf), \tag{1.12}$$

where P^{pre}_{HBD} is the predicted price through the HBD structure, S is the close price of the underlying asset, X is the strike price, T is the maturity, q is the dividend rate, and rf is the risk free rate. The function $BI(\cdot)$ represents the process of the binomial tree model, σ^{avg} is the average implied volatility according to Equation (1.8).

The loss function we optimize in the HBD is shown in Equation 1.13).

$$Loss(HBD)_t = \min \sum_{i=1}^{N} (Res^{mrk}_{i,t} - Res^{pre}_{i,t})^2, \tag{1.13}$$

where i is the i-th observation, N represents total observations on t. Res^{mrk} represents the residual between the market option price and the binomial tree-based option pricing[5], Res^{pre} is residual that the HBD model seeks to predict.

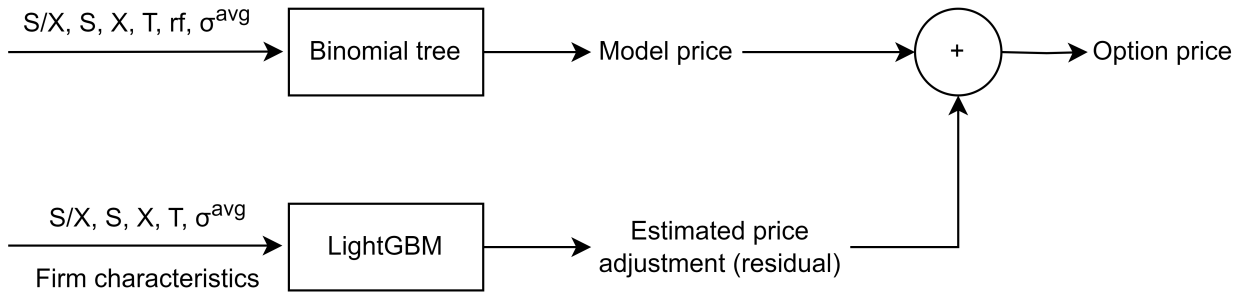
The estimation of the Res^{pre} is summarized as follow:

$$Res^{pre}_{i,t} = \text{LightGBM}(F_{i,t}; \theta), \tag{1.14}$$

where, $Res^{pre}_{i,t}$ is the predicted residuals for the i-th observation on t, as determined by the LightGBM model. $F_{i,t}$ represents the vector of input features for the i-th observation, such as the moneyness (S/X), close price (S), strike price (X), maturity (T), σ^{avg}, and firm characteristics. θ denotes the set of parameters of the LightGBM model.

We extend the HBD model by incorporating firm characteristics and employing LightGBM instead of a neural network. Figure 1.2 describes the schematic structure of the HBD model.

Figure 1.2. HBD model structure.



[5] $P^{mrk} - BI(\sigma^{avg}_{i,t}, S_{i,t}, X_{i,t}, T_{i,t}, rf_{i,t})$

1.3 LightGBM

The machine learning algorithm we employ is LightGBM proposed by Ke et al. (2017). LightGBM is a gradient-boosting framework that employs decision tree-based learning algorithms for both classification and regression tasks. It is designed for efficiency and scalability when dealing with large datasets. Here is an outline of the algorithm and the key formulas for using LightGBM for regression:

1. Initialize the model: Start with an initial model, which is usually the average of the target variable values for regression tasks. Mathematically, the initial model can be represented as:

$$F_0(x) = arg\min_c \sum L(y_i, c) \tag{1.15}$$

where $L(y_i, c)$ is the loss function for the target variable y_i corresponding to observation x_i and the constant value c.

2. Gradient boosting: Iteratively construct weak learners (decision trees) and combine them to create a strong learner. For each iteration $k = 1, 2, ..., K$:

   (a) Calculate the gradients g_i and Hessians h_i for each observation x_i in the dataset:

$$g_i = \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \tag{1.16}$$

$$h_i = \frac{\partial^2 L(y_i, F(x_i))}{\partial F(x_i)^2} \tag{1.17}$$

   where $F(x_i)$ is the current model's prediction for observation x_i.

   (b) Build a new decision tree to fit the negative gradients: Fit a new decision tree $f_k(x)$ using the dataset (x_i, g_i, h_i). In LightGBM, the decision tree is built using the 'leaf-wise' strategy with depth limitation, which splits the tree node with the highest loss reduction first.

(c) Determine the optimal learning rate (shrinkage factor) η_k:

$$\eta_k = arg\min_{\eta} \sum [L(y_i, F(x_i) + \eta \times f_k(x_i)] \tag{1.18}$$

Utilize line search or another optimization method to find the learning rate that minimizes the loss function.

(d) Update the model:

$$F_k(x) = F_{t-1}(x) + \eta_k \times f_k(x) \tag{1.19}$$

3. Final model: After K iterations, the final model can be represented as:

$$F(x) = F_0(x) + \sum_{k=1}^{K} \eta_k \times f_k(x) \tag{1.20}$$

1.4 Data and Methodology

1.4.1 Option data

Options data are obtained from OptionMetrics' IvyDB US and cover the period in the U.S. from January 1996 to December 2021. We use the first 365 calendar days of the sample to train the machine learning-based models and set the out-of-sample period to be from January 1997 to December 2021. Following Dumas, Fleming, and Whaley (1998), we define the option price as the midpoint of the bid and ask prices to reduce the estimation noise of implicit parameters. The underlying stocks' prices ($S_{i,t}$) are collected from the Securities table in OptionMetrics' IvyDB US, and the 3-month Treasury bill rate, which is used as a proxy for the risk-free rate ($rf_{i,t}$), is obtained from the St. Louis Federal Reserve Economic Data. Following Bakshi, Cao, and Chen (1997) and Andreou et al. (2010), we filter the option data using the following criteria.

1. Options with a trading volume of less than 100 contracts are eliminated as they are deemed to be illiquid and their prices may not represent the actual market price.

2. The time to maturity should be at least six days and no longer than 365 calendar days as options near expiration may induce liquidity-related biases.

3. Options with price quotes less than 0.1 U.S. dollars are eliminated.

4. The moneyness (S/X) of an option should be between 0.8 and 1.2.

5. Observations that violate the usual no-arbitrage condition are dropped.

After filtering, the final data set contains 15,247,956 observations. Sample characteristics of the dataset are reported in Table 1.1 and 1.2. Table 1.1 shows that there are more observations of out-of-the-money options than in-the-money options and more observations of call options than put options. The volatility smile is observed in all maturity groups and it is more pronounced in the near-term options. Table 1.2 reports the number of options per underlying stock year by year. The number of options per stock has increased over time. On average, the average number of options increases from 2 in 1996 to 11 in 2021. However, the median number of options remains small at 2 even in 2021, while the maximum number of options increases to 298, implying that there are only a handful of stocks with many options and the rest have only a few options associated with them.

1.4.2 Firm characteristics

Integrating firm characteristics into option pricing is supported by substantial evidence showing their significant impact on the dynamics of underlying assets and option values. For instance, Andreou (2015) demonstrate that factors like market default risk, derived from firm-level financial health, significantly influence the risk-neutral moments of major index options such as volatility and skewness. Vasquez and Xiao (2023) confirm the effect of default risk on the returns of delta-hedged equity options, directly linking firm leverage and asset volatility to option pricing.

Chen et al. (2023) document how firm fundamentals, such as profitability and market power, shape implied volatility curves and influence option prices. Their findings suggest that an array of firm-specific attributes, such as profitability and market power, substantially influence the cross-sectional variation in option prices. Additionally, Trigeorgis and Lambertides (2014) emphasize the value of firm-specific business volatility and managerial flexibility in assessing growth options, further advocating for the incorporation of such characteristics in option pricing models. Zhan et al. (2022) also show that delta hedged option portfolio is inversely related to factors such as stock price profit margin and firm profitability, while being positively correlated with cash holdings, cash flow

Table 1.1. Characteristics of option data. This table describes the characteristics of the option data. The data is obtained from OptionMetrics' IvyDB US and covers the stock options in the US market from January 1996 to December 2021. 'Price', 'Implied volatility', and 'Observations' respectively refer to the average price, average implied volatility, and the number of observations in each subset.

Panel A. Call options							
Moneyness	DOTM	OTM	JOTM	ATM	JITM	ITM	DITM
S/X	0.80-0.90	0.90-0.95	0.95-0.99	0.99-1.01	1.01-1.05	1.05-1.10	1.10-1.20
Near-term (6-60 days)							
Price	1.522	1.783	2.574	4.327	5.581	7.383	10.196
σ^{mrk}	0.533	0.446	0.365	0.345	0.393	0.466	0.529
Observations	584,412	1,214,045	1,737,246	928,081	916,704	457,736	261,016
Mid-term (60-180 days)							
Price	2.393	3.110	4.541	6.753	7.364	8.740	11.085
σ^{mrk}	0.427	0.343	0.323	0.331	0.352	0.391	0.449
Observations	638,112	577,593	518,574	227,643	277,091	178,228	153,012
Long-term (181-365 days)							
Price	3.964	5.007	6.840	9.277	9.772	11.028	13.488
σ^{mrk}	0.359	0.316	0.311	0.322	0.338	0.360	0.399
Observations	242,411	168,113	134,848	58,920	74,444	54,149	55,433
Panel B. Put options							
Moneyness	DOTM	OTM	JOTM	ATM	JITM	ITM	DITM
S/X	1.10-1.20	1.05-1.10	1.01-1.05	0.99-1.01	0.95-0.99	0.90-0.95	0.80-0.90
Near-term (6-60 days)							
Price	1.636	1.951	2.720	4.215	5.303	7.658	17.488
σ^{mrk}	0.509	0.442	0.374	0.352	0.400	0.440	0.521
Observations	548,159	884,809	1,185,264	662,459	579,464	93,865	22,592
Mid-term (60-180 days)							
Price	2.692	3.479	4.566	6.044	6.449	7.916	11.028
σ^{mrk}	0.417	0.369	0.347	0.347	0.369	0.427	0.469
Observations	367,382	292,084	284,545	144,687	181,842	75,981	19,482
Long-term (181-365 days)							
Price	4.759	5.745	7.065	8.889	8.805	10.113	14.641
σ^{mrk}	0.374	0.351	0.342	0.344	0.358	0.394	0.456
Observations	135,843	88,731	79,142	40,563	55,304	32,431	15,516

Table 1.2. Number of options per stock. This table describes the average number of firms per year and the descriptive statistics (Mean, Std, Min, 25%, 50%, 75% and Max) of the number of options per stock in each year in the sample period.

Year	No. firms	Mean	Std	Min	25%	50%	75%	Max
1996	147	2	3	1	1	1	3	26
1997	170	3	4	1	1	1	3	34
1998	180	3	5	1	1	1	4	32
1999	188	4	5	1	1	1	4	36
2000	196	4	5	1	1	2	4	35
2001	206	4	4	1	1	2	5	28
2002	207	4	5	1	1	2	5	29
2003	230	4	5	1	1	2	6	28
2004	262	5	5	1	1	2	6	28
2005	286	5	5	1	1	3	7	34
2006	317	5	6	1	1	3	8	43
2007	352	6	7	1	1	3	8	55
2008	342	6	7	1	1	3	7	58
2009	340	6	8	1	1	3	8	62
2010	334	7	9	1	1	3	9	72
2011	339	8	13	1	1	3	10	146
2012	313	8	16	1	1	3	10	236
2013	346	8	15	1	1	3	9	210
2014	363	8	17	1	1	2	8	251
2015	356	7	17	1	1	2	6	235
2016	358	7	15	1	1	2	6	172
2017	402	7	16	1	1	2	6	170
2018	446	8	19	1	1	2	6	195
2019	450	8	19	1	1	2	6	190
2020	492	10	24	1	1	2	7	278
2021	611	11	27	1	1	2	7	298

variance, new shares issuance, total external financing, distress risk, and the dispersion of analysts' forecasts.

In our research, we use 111 firm characteristics[6] that include accounting ratios, momentum features, stock return volatility, and other characteristics related to the firm's operation, growth, risk, and performance. These characteristics are selected from the comprehensive list of firm characteristics in Jensen et al. (2021) after eliminating firm characteristics with more than 20% missing values. We fill the remaining missing values with the cross-sectional median value following Gu et al. (2020). The list of the firm characteristics with their description can be found in the appendix. The exact definitions of the firm characteristics can be found in Jensen et al. (2021) or the references therein. To avoid any forward-looking bias, we align the firm characteristics from six months prior with the option data from the current month.

[6] These firm characteristics are generated using PyAnomaly, a powerful Python library for firm characteristics generation and asset pricing study. https://pyanomaly.readthedocs.io/en/latest/index.html The package provides easy access to various financial data sources and ensures data accuracy and consistency.

1.4.3 Model specifications and evaluation metrics

We evaluate the out-of-sample option pricing performance of five models: the binomial model (BI), the GPF model with standard features (GPF), the GPF model enhanced with firm characteristics (GPF_F), the HBD model with standard features (HBD), and the HBD model incorporating firm characteristics (HBD_F).

The models are evaluated using the metrics root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) as follows:

$$RMSE = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(z_l - \hat{z}_l)^2}, \tag{1.21}$$

$$MAE = \frac{1}{L}\sum_{l=1}^{L}|z_l - \hat{z}_l|, \tag{1.22}$$

$$MAPE = \frac{1}{L}\sum_{l=1}^{L}\left|\frac{z_l - \hat{z}_l}{z_l}\right|, \tag{1.23}$$

where L represents the number of observations in the dataset, z_l is the market price of the l-th option, and $\hat{z}_l$ is the predicted price for the l-th option, as given by the model.

We use the Model Confidence Set (MCS) method, a statistical technique designed for model comparison and selection in time series analysis, to evaluate a group of predictive models. Introduced by Hansen, Lunde, and Nason (2011), the MCS helps identify models with statistically comparable performance. Our research implements MCS through a series of defined steps.

1. **Selection of Prediction Models**: Initiate the MCS procedure by selecting models based on their mean square error (MSE) for comparison.

2. **Null Hypothesis Formulation**: Assume no significant MSE differences among models and test this assumption.

3. **Loss Differential Analysis**: Create a matrix of MSE differences between models for statistical analysis.

4. **Statistical Testing**: Apply statistical tests to assess if MSE differences are significant.

5. **Interpreting p-values**: Analyze p-values using a 1% significance threshold to determine model inclusion in the MCS.

1.4.4 Implementation Details

In our study using the LightGBM gradient boosting method, we apply it to regression tasks with default hyperparameters such as number of leaves, tree depth, and learning rate, choosing a conservative setup without extensive tuning. The loss function selected is MSE.

The GPF model operates daily by first calculating the average implied volatility (σ^{avg}), using it along with standard option features and firm characteristics to feed into LightGBM. The model trains on the past year's data, with the most recent 30 days as the validation set. This trained model predicts the next day's implied volatility, which is then input into a binomial tree model to forecast the option price.

The HBD approach calculates the option price via a binomial tree, then determines the price residuals—differences between the market and model prices. These residuals serve as targets for LightGBM training. The output is used to adjust the next day's binomial model-based option price.

For validation, models are assessed using out-of-sample metrics to identify the one with the lowest MSE through the MCS method, indicating the most accurate option pricing performance.

1.5 Empirical Analysis

We evaluate the out-of-sample performance of the models using the evaluation metrics defined in the previous section. We first compare the pricing errors of all options and then examine the pricing errors of different subsets defined by option type, moneyness, and maturity to validate the consistency of the pricing performance. We also identify informative firm characteristics by analyzing feature importance.

1.5.1 Pricing performance

1.5.1.1 All options

Table 1.3 reports the pricing performance of the models evaluated using all options. First, comparing the three models that do not incorporate firm characteristics, BI, GPF, and HBD, we find that GPF performs best, followed by HBD. The RMSEs of GPF, HBD, and BI are 1.701, 1.710, and 2.023, respectively. The superiority of GPF is also reflected in the other evaluation metrics. It is notable that the machine learning-based models can reduce the pricing error of BI by 15%. This result suggests that employing machine learning can effectively reduce pricing error, and the GPF architecture is more effective than that of HBD.

When firm characteristics are incorporated, both GPF and HBD yield smaller pricing errors: the RMSEs of GPF_F and HBD_F are respectively 1.376 and 1.459, which are considerably smaller than those of GPF and HBD. Firm characteristics improve the pricing performance of both GPF and HBD by about 19.11% and 14.68%, respectively, making GPF_F the best performer among all five models, followed by HBD_F. The MCS results indicate that GPF_F is the only model that is included in the MCS.

The reason for the superior performance of GPF over HBD appears to be related to the way it integrates machine learning. GPF predicts implied volatility via machine learning while HBD predicts residual. The same error in implied volatility can lead to an error of different magnitudes in option price depending on the moneyness and the maturity of the option. This added complexity makes the prediction of pricing residual more challenging.

Table 1.3. Pricing performance for all options. This table presents the out-of-sample pricing errors of each model for all the options in the sample. The out-of-sample period is from January 1997 to December 2021. RMSE, MAE, and MAPE refer to the root mean squared error, the mean absolute error, and the mean absolute percentage error, respectively. MCS-p refers to the p-value of the model confidence set (MCS) test, where a high p-value indicates a model is more likely to be part of the MCS. The details of each model can be found in Section 1.4.3.

	BI	GPF	HBD	GPF_F	HBD_F
RMSE	2.023	1.701	1.710	1.376	1.459
MAE	0.566	0.470	0.473	0.381	0.385
MAPE	0.208	0.178	0.186	0.144	0.156
MCS-p	0.000	0.000	0.000	1.000	0.015

1.5.1.2 Call vs put options

Table 1.4 reports the pricing performance of the models for call options (Panel A) and put options (Panel B), separately. It shows that the differences in pricing errors between call options and put options are almost negligible, and GPF_F and HBD_F remain as the best performers for both types of options. For instance, the RMSEs of GPF_F, HBD_F, GPF, and HBD for call options are 1.365, 1.453, 1.700, and 1.702, respectively, while the RMSEs for put options are 1.394, 1.468, 1.717, and 1.722. These models significantly outperform BI for both option types, whose RMSE is 2.065 for call options and 1.954 for put options. This result suggests that machine learning can enhance the pricing performance for both option types and the performance can be further improved by incorporating firm characteristics. The smaller MAPE of put options can be attributed to the fact that the prices of out-of-the-money put options are higher than the prices of out-of-the-money call options in our sample, as shown in Table 1.1. Out-of-the-money options are cheaper and bear larger percentage errors. The higher prices of out-of-the-money put options result in smaller overall percentage errors.

Table 1.4. Pricing performance for call and put options. This table presents the out-of-sample pricing errors of each model for call and put options. The out-of-sample period is from January 1997 to December 2021. RMSE, MAE, and MAPE refer to the root mean squared error, the mean absolute error, and the mean absolute percentage error, respectively. MCS-p refers to the p-value of the model confidence set (MCS) test, where a high p-value indicates a model is more likely to be part of the MCS. The details of each model can be found in Section 1.4.3.

	BI	GPF	HBD	GPF_F	HBD_F
Panel A. Call options (9,379,625 observations)					
RMSE	2.065	1.700	1.702	1.365	1.453
MAE	0.561	0.465	0.467	0.372	0.379
MAPE	0.212	0.180	0.189	0.143	0.157
MCS-p	0.000	0.000	0.000	1.000	0.015
Panel B. Put options (5,775,718 observations)					
RMSE	1.954	1.717	1.722	1.394	1.468
MAE	0.574	0.480	0.483	0.394	0.396
MAPE	0.201	0.175	0.182	0.145	0.154
MCS-p	0.000	0.000	0.000	1.000	0.009

1.5.1.3 Different moneyness options

It is well known that implied volatility differs across moneyness, the phenomenon known as volatility smile, and moneyness can have a significant impact on option pricing performance. For instance, in-the-money options may require larger price adjustments compared to out-of-the-money options as the former is more likely to be exercised prior to maturity, whereas the latter may be held to maturity. To examine the impact of moneyness on option pricing, we analyze the models' performances for different moneyness options.

Table 1.5 reports the pricing errors for each moneyness group. GPF_F is the best performer for all moneyness groups[7] with the lowest RMSEs: 1.148 for DOTM, 1.253 for OTM, 1.468 for JOTM, 1.735 for ATM, 1.404 for JITM, 1.117 for ITM, 1.046 for DITM. It is notable that the benchmark BI model has larger errors for all moneyness groups, whereas the proposed models perform consistently across all moneyness groups. Moreover, the performance improvement by firm characteristics is particularly evident when pricing DOTM options. This result suggests that firm characteristics can explain volatility smile to some extent. Given the larger trading volume of out-of-the-money options and the fact that investors commonly use these options as a form of protection due to their lower cost, accurate pricing of these options is of greater significance than that of in-the-money options. Out-of-the-money options are also more difficult to price due to volatility smile. Both GPF_F and HBD_F exhibit greater performance improvement when pricing these options.

1.5.1.4 Different maturity options

Table 1.6 reports the pricing performance of the models for different maturity groups. Again, GPF_F performs best in all maturity groups in terms of RMSE, followed by HBD_F, GPF, and HBD. The errors tend to increase with time to maturity for BI, where the RMSE of BI increases dramatically from 1.816 (near-term) to 3.050 (long-term). In contrast, the proposed models perform consistently throughout the maturity groups. For instance, the RMSE of GPF_F for near-term, mid-term and long-term are 1.470, 1.1077, and 1.352, respectively. Moreover, firm characteristics appear to have a more significant impact on options with longer maturities. For near-term options,

[7] DOTM stands for deep out-the-money; OTM stands for out-the-money; JOTM stands for just out-the-money; ATM stands for at-the-money; JITM stands for just in-the-money; ITM stands for in-the-money; DITM stands for deep in-the-money. The specific definition of the moneyenss can be found in Table 1.1.

Table 1.5. Pricing performance for different moneyness options. This table presents the out-of-sample pricing errors of each model for different moneyness options.

	BI	GPF	HBD	GPF_F	HBD_F
Panel A. Deep out-the-money (2,501,427 observations)					
RMSE	2.115	1.449	1.456	1.148	1.242
MAE	0.603	0.426	0.427	0.316	0.330
MAPE	0.329	0.245	0.266	0.185	0.211
MCS-p	0.000	0.000	0.000	1.000	0.003
Panel B. Out-the-money (3,203,805 observations)					
RMSE	1.904	1.573	1.576	1.253	1.347
MAE	0.545	0.440	0.449	0.350	0.359
MAPE	0.283	0.245	0.260	0.196	0.215
MCS-p	0.000	0.000	0.000	1.000	0.010
Panel C. Just Out-the-money (3,918,073 observations)					
RMSE	1.996	1.827	1.830	1.468	1.552
MAE	0.571	0.494	0.494	0.403	0.404
MAPE	0.223	0.199	0.203	0.165	0.173
MCS-p	0.000	0.000	0.000	1.000	0.016
Panel D. At-the-money (2,052,942 observations)					
RMSE	2.339	2.102	2.154	1.735	1.816
MAE	0.657	0.556	0.566	0.474	0.470
MAPE	0.145	0.134	0.135	0.114	0.117
MCS-p	0.000	0.000	0.000	1.000	0.008
Panel E. Just in-the-money (2,071,177 observations)					
RMSE	1.970	1.640	1.698	1.404	1.472
MAE	0.533	0.475	0.484	0.404	0.405
MAPE	0.085	0.086	0.087	0.072	0.076
MCS-p	0.000	0.000	0.000	1.000	0.008
Panel F. In-the-money (882,661 observations)					
RMSE	1.765	1.338	1.347	1.117	1.182
MAE	0.456	0.404	0.418	0.345	0.351
MAPE	0.054	0.057	0.059	0.047	0.052
MCS-p	0.000	0.000	0.000	1.000	0.014
Panel G. Deep in-the-money (517,296 observations)					
RMSE	1.760	1.223	1.237	1.046	1.101
MAE	0.430	0.375	0.377	0.306	0.317
MAPE	0.039	0.040	0.043	0.033	0.039
MCS-p	0.000	0.000	0.000	1.000	0.009

the RMSE of GPF is reduced by 18.92% when firm characteristics are incorporated, whereas it is reduced by 20.30% for mid-term options and by 21.94% for long-term options. Similarly, the RMSE of HBD is reduced by 14.74% for near-term options, by 15.42% for mid-term options, and by 17.85% for long-term options.

The pricing performance of long-term options is likely to be improved more when firm characteristics are taken into account. Toft and Prucyk (1997) argue that long-term options are more closely related to a firm's financial traits, like its capital structure and leverage. This connection suggests that considering these aspects can lead to more accurate option pricing. Similarly, Kahle and Shastri (2005) find a significant link between long-term debt ratios, a key firm characteristic, and the value of long-term options, highlighting the importance of these factors in option pricing. Bakshi, Cao, and Chen (2000) observe that long-term options contain vital information about the firm, which can be used to improve the accuracy of pricing models for these options. In summary, considering firm characteristics in pricing long-term options is essential for better performance, as these factors significantly impact long-term returns, risk/reward profiles, and strategic decisions.

Table 1.6. Pricing performance for different maturity options. This table presents the out-of-sample pricing errors of each model for different maturity options.

	BI	GPF	HBD	GPF_F	HBD_F
Panel A. Near-term (10,004,961 observations)					
RMSE	1.816	1.813	1.825	1.470	1.556
MAE	0.522	0.480	0.481	0.395	0.398
MAPE	0.220	0.202	0.212	0.167	0.181
MCS-p	0.000	0.000	0.000	1.000	0.006
Panel B. Mid-term (3,905,431 observations)					
RMSE	2.118	1.389	1.401	1.107	1.185
MAE	0.582	0.426	0.431	0.330	0.336
MAPE	0.184	0.136	0.140	0.101	0.110
MCS-p	0.000	0.000	0.000	1.000	0.008
Panel C. Long-term (1,228,960 observations)					
RMSE	3.050	1.732	1.737	1.352	1.427
MAE	0.866	0.537	0.543	0.423	0.435
MAPE	0.180	0.119	0.122	0.090	0.097
MCS-p	0.000	0.000	0.000	1.000	0.009

1.5.1.5 Year-by-year performance

Table 1.7 reports the MAPEs of each model year by year over the out-of-sample period. We report MAPE instead of RMSE or MAE because both option prices and (absolute) errors increase with time and therefore it is difficult to compare the performance across years with absolute evaluation metrics. As before, GPF_F and HBD_F outperform GPF and HBD, respectively, and BI performs the worst. The proposed models perform particularly well in comparison to BI in recent years when there are significantly more options in the market. For instance, the MAPEs of GPF_F and BI in 1997 are 0.131 and 0.153, respectively, whereas those in 2021 are 0.146 and 0.190. The improved performance in recent years can be attributed to the increased volume of the training set. Machine learning algorithms learn patterns from historical data and having a large amount of data for training is critical to avoid overfitting and make more accurate predictions.

The increasing trend of pricing errors corresponds to the growth of trading volume. As more participants engage in the options market and trading activity intensifies, prices can swing more due to more frequent changes of demand and supply dynamics. Increased liquidity also allows for quicker transactions and increased price sensitivity to new information. These factors can contribute to the increasing trend of pricing errors.

1.5.2 Importance of firm characteristics

The LGBM evaluates the importance of input features through two methods: Frequency (Number of Splits) and Gain (Split Importance). The Frequency counts the total number of times a feature is used to split the data across all trees in the model. Features used more often in the tree construction process are considered more important. The Gain measures the total improvement in the loss function that results from each split based on a particular feature. A higher gain indicates that the feature is more important for making accurate predictions. We assess feature importance using the Gain method as it quantifies the actual contribution of each feature. To determine the overall importance of a feature, we compute its importance in each month and use the time-series average.

Figure 1.3 and 1.4 present the 20 most important features in GPF_F and HBD_F, respectively.

Table 1.7. Pricing performance in each year. This table presents the MAPE of each model year by year. The out-of-sample period is from January 1997 to December 2021. The details of each model can be found in Section 1.4.3.

Year	BI	GPF	HBD	GPF_F	HBD_F	Obs.
1997	0.153	0.153	0.158	0.131	0.134	153,229
1998	0.172	0.157	0.163	0.133	0.136	181,157
1999	0.133	0.124	0.128	0.105	0.107	199,367
2000	0.134	0.128	0.132	0.101	0.103	220,678
2001	0.161	0.132	0.135	0.098	0.100	213,662
2002	0.191	0.137	0.139	0.099	0.102	232,965
2003	0.168	0.122	0.123	0.099	0.100	280,989
2004	0.151	0.121	0.120	0.101	0.102	334,722
2005	0.168	0.140	0.140	0.117	0.118	388,334
2006	0.157	0.135	0.135	0.114	0.117	474,555
2007	0.171	0.146	0.149	0.118	0.123	561,669
2008	0.184	0.162	0.167	0.122	0.128	558,326
2009	0.163	0.122	0.125	0.103	0.106	598,001
2010	0.178	0.137	0.141	0.109	0.113	632,137
2011	0.204	0.152	0.158	0.127	0.135	733,218
2012	0.214	0.154	0.161	0.130	0.141	676,854
2013	0.209	0.171	0.180	0.139	0.159	743,229
2014	0.218	0.201	0.213	0.162	0.177	765,182
2015	0.244	0.212	0.227	0.172	0.183	717,363
2016	0.265	0.215	0.226	0.175	0.193	685,442
2017	0.250	0.213	0.226	0.178	0.191	775,183
2018	0.244	0.221	0.233	0.175	0.195	954,575
2019	0.273	0.231	0.248	0.189	0.208	1,000,970
2020	0.232	0.225	0.237	0.169	0.188	1,323,542
2021	0.190	0.177	0.185	0.146	0.162	1,749,994

The scores in the figures are the average scores of all monthly training results. Although not reported here, we find that the importance scores are stable over the sample period, which illustrates that the most important features consistently aid in predicting option prices.

Figure 1.3. Feature importance score of the 20 most important features in GPF$_F$. The scores in the figure are the average scores of all monthly training results extracted from the LightGBM algorithm. A higher score means the feature is used more often in the tree construction process and the feature is considered more important. T, S/X, σ^{avg}, S, and X, are the most important features and take about 50.81% of total importance. Conditional skewness (coskew_21d), Dimson Beta (beta_dimson_21d), dividend yield (div12m_me), One-year-lagged annual stock return (seas_1_1an), and downside beta (betadown_252d) are the most important firm characteristics.

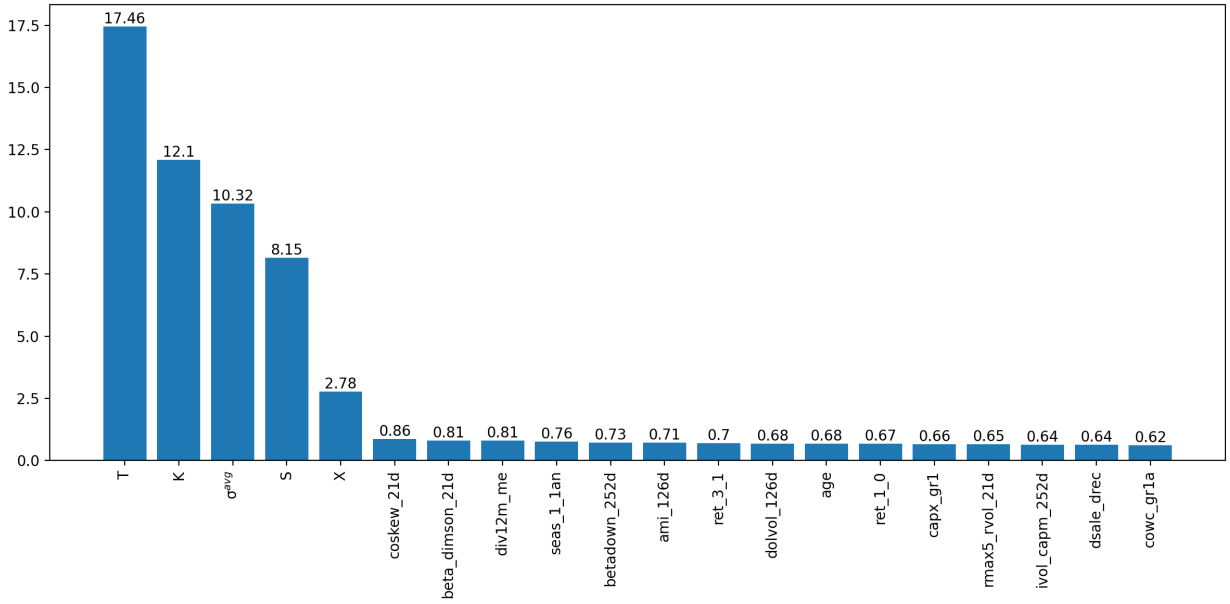


Figure 1.3 demonstrates that option-related variables, including T, S/X, σ^{avg}, S, and X, are the most critical, accounting for 50.81% of the total feature importance. This aligns with expectations since the target of the algorithm in GPF$_F$ is implied volatility, and these parameters are crucial for its calculation in the binomial model. S/X helps the model assess option moneyness, potentially enhancing volatility smile detection.

The remaining 49.19% of feature importance is attributed to firm characteristics such as conditional skewness (coskew_21d), Dimson Beta (beta_dimson_21d), dividend yield (div12m_me), one-year-lagged annual stock return (seas_1_1an), and downside beta (betadown_252d). These features' predictive capabilities are further discussed in Section 1.6.

Conditional skewness (coskew_21d) significantly impacts option pricing by affecting the return of the underlying asset, with a notable influence score of 0.86%. Our model, based on a binomial distribution, considers both increases and decreases in the underlying asset's price at each time

step. Incorporating the distribution details of the asset into our machine learning algorithm has proven beneficial for addressing volatility smiles more effectively. Other option pricing researches also show that the skewness of stock could help the prediction of option price. For instance, Corrado and Su (1996) extend the Black-Scholes model to include the skewness of the underlying asset, highlighting its value in option pricing. Studies also indicate that option implied volatility skewness (Bates, 1991; Doran, Peterson, and Tarrant, 2007; Van Buskirk, 2009; Xing, Zhang, and Zhao, 2010) the stock skewness (Chen et al., 2001; Byun and Kim, 2016; Jang and Kang, 2019) can predict the underlying stock return, making it reasonable to assume that the skewness also helpful for the option price prediction since the value of a call (put) option is a positive (negative) function of the stock price (Byun and Kim, 2016).

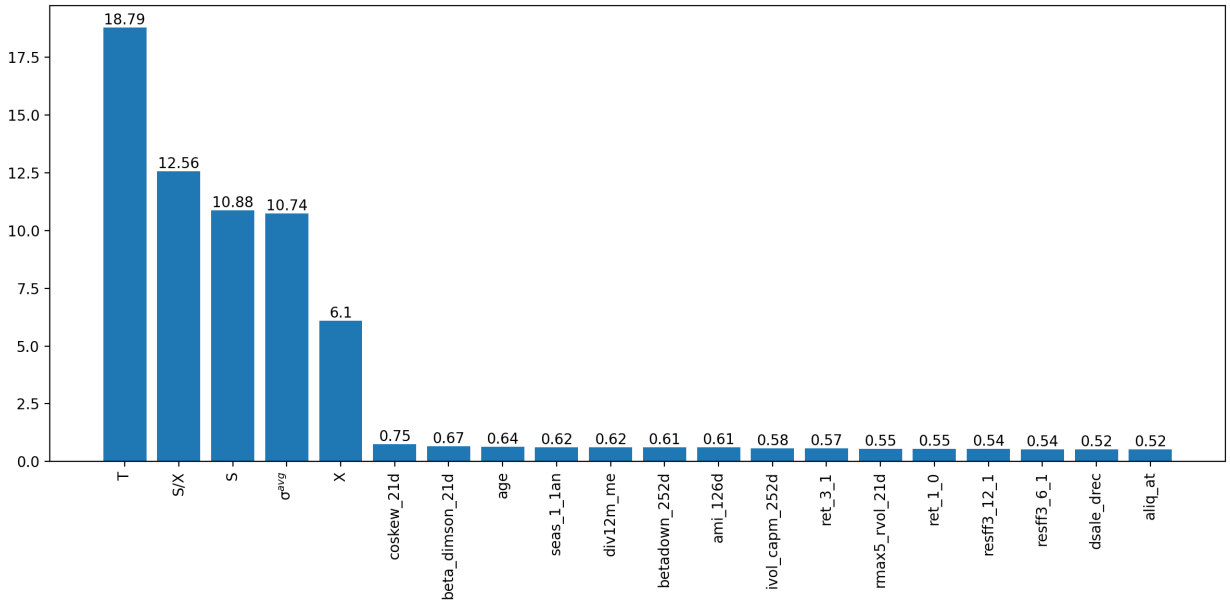
As for the Dimson beta (beta_dimson_21d) proposed by Dimson (1979) has an important value of 0.81%. Dimson (1979) argues that the traditional beta estimate (the sensitivity of stock returns to market returns) may be biased in thinly traded markets because of infrequent trading. To adjust for this, Dimson (1979) suggests using a model that incorporates leading and lagging factors in the calculation of beta. The adjustment of Dimson Beta can indirectly affect the binomial option pricing model through the estimation of volatility. If we assume that the beta adjustment more accurately reflects the risk of a given stock, it may lead to a more accurate estimate of the stock's future price volatility, which is an important input in the binomial option pricing model. Since we use machine learning to predict implied volatility, Dimson beta can be valuable in the prediction process.

Figure 1.4 reveals that the option-related variables are also the most important features in HBD_F. Their importance scores are higher and firm characteristics play a less important role, with an aggregate score of 40.93% in HBD_F. This is because the machine learning algorithm in HBD_F predicts the residuals from the BI model that may have more noise in the price dimension, and the option-related variables are directly related to the option price, which is at the same dimension of the residual.

Conditional skewness (coskew_21d), Dimson Beta (beta_dimson_21d), firm age (age), One-year-lagged annual stock return (seas_1_1an) and dividend yield (div12m_me) are the most important features among firm characteristics. The fact that these firm characteristics are identified as impor-

tant features in both models implies that they are not picked by chance but do contain information about the option price.

Figure 1.4. Feature importance score of the 20 most important features in HBD$_F$. The scores in the figure are the average scores of all monthly training results extracted from the LightGBM algorithm. A higher score means the feature is used more often in the tree construction process and the feature is considered more important. T, S/X, σ^{avg}, S, and X, are the most important features and take about 59.07% of total importance. Conditional skewness (coskew_21d), Dimson Beta (beta_dimson_21d), firm age (age), One-year-lagged annual stock return (seas_1_1an) and dividend yield (div12m_me)



From an economic standpoint, firm characteristics that affect stock performance, such as profitability, leverage, market value, and volatility, are inherently linked to the company's financial health and growth prospects. These characteristics influence investors' perceptions and expectations, which are generally reflected in the stock price movements and, by extension, in the pricing of options on those stocks. For example, the most important skewness condition in firm characteristics measures the skewness of stock returns. Several studies (Chen et al., 2001; Dennis and Mayhew, 2002; Conrad, Dittmar, and Ghysels, 2013; Boyer and Vorkink, 2014; Byun and Kim, 2016; Del Viva, Kasanen, and Trigeorgis, 2017; Schneider, Wagner, and Zechner, 2020) show that the skewness could relate to the stock returns. The call (put) option price is a positive (negative) function of the stock price, the change in the stock returns could affect the call (put) option price. Thus, the factors that are related to stock return performance could also related to the option

price prediction. This is consistent with the finding in Byun and Kim (2016) that high-skewness stocks usually have higher stock returns and overpriced call options. On the other hand, several measures of beta also have high characteristic importance. Beta is a measure of the riskiness of a stock's performance relative to certain benchmark returns. It also affects the expectation of future stock returns (Latane and Rendleman, 1976; Chang, Christoffersen, Jacobs, and Vainberg, 2012; Boloorforoosh, Christoffersen, Fournier, and Gouriéroux, 2020) and thus the expectation of option prices. As for dividend yield and annual yield with a one-year lag, they are both measures of option profitability, where dividend yield correlates with option value because option holders are not eligible for dividends, and annual yield gives some momentum insight into the stock, which may make investors prefer options on high-momentum stocks and thus affect the option price (Byun and Kim, 2016).

The price of call (put) options is positively (negatively) influenced by stock prices, though these relationships are typically nonlinear. Parametric models often restrict the source of nonlinearity to standard variables like volatility, maturity, strike price, closing price, and risk-free rate. Corrado and Su (1996) expands these variables to include skewness and kurtosis. Machine learning models, particularly semi-parametric ones discussed in the thesis, effectively incorporate both standard and additional variables, blending the strengths of parametric and machine learning approaches. These semi-parametric models retain the structure of parametric models while leveraging machine learning to address nonlinearities and firm characteristics. They also manage multicollinearity, allowing for a wider exploration of firm characteristics, including operational aspects, as noted by Gu et al. (2020); Han (2021); Bali et al. (2023). This capability enables models like LightGBM to select the most relevant features for modelling, even among correlated characteristics.

1.6 Robustness tests

1.6.1 Training models by the most important firm characteristics

The results presented in Table 1.8 detail the iteration variations of the GPF, each progressively integrating additional firm characteristics. The initial model, GPF_f^1, incorporates conditional

skewness (coskew_21d). Successive versions build on this by adding Dimson Beta (beta_dimson_21d) in GPF_f^2, then dividend yield (div12m_me) in GPF_f^3, followed by one-year-lagged annual stock return (seas_1_1an) in GPF_f^4, and culminating with downside beta (betadown_252d) in GPF_f^5. Each additional firm characteristic aims to test the balance of complexity and predictive accuracy.

The greatest RMSE reduction is seen with the inclusion of the coskew_21d, dropping from 1.701 to 1.532. The next significant improvement comes from adding beta_dimson_21d, further reducing RMSE to 1.487. The smallest reduction occurs with the fifth characteristic, bringing the RMSE down from 1.453 to 1.446. These findings indicate that the first two characteristics significantly improve prediction accuracy, with subsequent characteristics offering progressively lesser benefits.

Table 1.8. Pricing performance for all options. This table presents the out-of-sample pricing errors of each model for all the options in the sample. The out-of-sample period is from January 1997 to December 2021. RMSE, MAE, and MAPE refer to the root mean squared error, the mean absolute error, and the mean absolute percentage error, respectively. MCS-p refers to the p-value of the model confidence set (MCS) test, where a high p-value indicates a model is more likely to be part of the MCS. The details of each model can be found in Section 1.4.3.

	GPF_F	GPF_f^5	GPF_f^4	GPF_f^3	GPF_f^2	GPF_f^1	GPF
RMSE	1.376	1.446	1.453	1.463	1.487	1.532	1.701
MAE	0.381	0.401	0.404	0.409	0.417	0.430	0.470
MAPE	0.144	0.151	0.152	0.153	0.157	0.161	0.178
MCS-p	1.000	0.000	0.000	0.000	0.000	0.000	0.000

1.6.2 Training models for each option group

As shown in Table 1.1, the distribution of the sample is highly imbalanced across different option groups, *e.g.*, there are significantly more near-term options than long-term options. The imbalance in the training set can lead to a biased result as the algorithm may prioritize minimizing the pricing error of near-term options over long-term options. If the relations between the input and the output are different between these two option groups, training one model for all options will result in a relatively poor performance for long-term options. On the other hand, training a machine algorithm individually for each option group can suffer from a small data problem, which can cause overfitting. The small data problem can be particularly severe in the early years of the sample, where available options are significantly fewer. To test the trade-off between the data

imbalance problem and the small data problem, we train the models individually for each option group and compare the results with those from the previous one-model-fits-all case. The results are presented below. Overall, it appears that the relationship between input characteristics and option prices for different types of options is not so obvious, so instead of training separate models for each group of options, one model for all types of options is sufficient.

1.6.2.1 Call vs put options

Table 1.9 reports the pricing performance of the models that are separately trained on call and put options. The proposed models yield slightly smaller errors when trained individually. For call options, the RMSEs of GPF_F and HBD_F are reduced from 1.365 to 1.335 and from 1.453 to 1.433, respectively, and for put options, the RMSEs are reduced from 1.394 to 1.365 and from 1.468 to 1.449. Although the improvements are minor, the result suggests that it is worth training the models separately for each option type.

Table 1.9. Pricing performance of individual models for call and put options. This table presents the out-of-sample pricing errors of the models that are trained on call and put options individually. The out-of-sample period is from January 1997 to December 2021. RMSE, MAE, and MAPE refer to the root mean squared error, the mean absolute error, and the mean absolute percentage error, respectively. MCS-p refers to the p-value of the model confidence set (MCS) test, where a high p-value indicates a model is more likely to be part of the MCS. The details of each model can be found in Section 1.4.3.

	BI	GPF	HBD	GPF_F	HBD_F
Panel A. Call options (9,379,625 observations)					
RMSE	2.065	1.685	1.696	1.335	1.433
MAE	0.561	0.448	0.450	0.346	0.356
MAPE	0.212	0.169	0.184	0.132	0.147
MCS-p	0.000	0.000	0.000	1.000	0.000
Panel B. Put options (5,775,718 observations)					
RMSE	1.954	1.687	1.689	1.365	1.449
MAE	0.574	0.453	0.451	0.367	0.371
MAPE	0.201	0.167	0.177	0.135	0.146
MCS-p	0.000	0.000	0.000	1.000	0.209

1.6.2.2 Different moneyness options

Table 1.10 reports the pricing performance of the models that are trained on different moneyness groups individually. The results of training the model alone were better in the DOTM group, but

in the other groups, the results were mixed, and even worse in the DITM group. It appears that dividing the sample into many groups results in not enough sample size for each group and the benefit of specifying a model for each group cannot dominate the small data problem.

1.6.2.3 Different maturity options

Table 1.11 reports the pricing performance of the models that are trained on different maturity groups individually. The models perform slightly better for near-term and mid-term options, *e.g.*, the RMSEs of GPF_F and HBD_F are respectively 1.448 and 1.532 for near-term options when trained individually, whereas they are 1.470 and 1.556 when trained using the entire sample. However, the models perform worse for long-term options when trained individually. The poor performance can be attributed to the small sample size of long-term options.

1.6.3 Black-Scholes model-based machine learning models

In this section, we replace the binomial model in the GPF and HBD with the BS model to test whether machine learning in the BS model setting helps improve pricing accuracy and whether firm characteristics help predict option prices. In addition, we seek to test whether machine learning can address the early exercise of American options that cannot be explained by the BS model. The implied volatilities calculations, the inputs to the GPF and HBD, and the hyperparameters of the machine learning model based on the BS model are the same as those of the machine learning model based on the binomial model. The benchmark model is the deterministic volatility function (DVF) proposed by Dumas et al. (1998) and extensively tested by Andreou, Charalambous, and Martzoukos (2014), where the DVF is an ad-hoc BS whose volatility is estimated using the function equation:

$$\sigma_{DVF} = \max(0.01, a_0 + a_1 K + a_2 K^2 + a_3 T + a_4 T^2 + a_5 KT, \tag{1.24}$$

where $K = S/X$, S is the close price, and T is the maturity.

Specifically, we tested five models: the DVF model (DVF), using lagged implied volatility as input to the BS model; the GPF, GPF_F, HBD, and HBD_F, replacing the binomial model with the BS model in the GPF and HBD structures (described in 1.4.3). Table 1.12 reports the pricing

Table 1.10. Pricing performance of individual models for different moneyness options. This table presents the out-of-sample pricing errors of the models that are trained on different maturity options individually.

	BI	GPF	HBD	GPF_F	HBD_F
Panel A. Deep out-the-money (2,501,427 observations)					
RMSE	2.115	1.324	1.416	1.055	1.256
MAE	0.603	0.394	0.403	0.301	0.328
MAPE	0.329	0.235	0.256	0.182	0.210
MCS-p	0.000	0.000	0.000	1.000	0.000
Panel B. Out-the-money (3,203,805 observations)					
RMSE	1.904	1.426	1.520	1.256	1.368
MAE	0.545	0.425	0.428	0.346	0.360
MAPE	0.283	0.236	0.253	0.195	0.216
MCS-p	0.000	0.000	0.000	1.000	0.012
Panel C. Just Out-the-money (3,918,073 observations)					
RMSE	1.996	1.636	1.753	1.461	1.581
MAE	0.571	0.475	0.478	0.398	0.408
MAPE	0.223	0.192	0.201	0.163	0.175
MCS-p	0.000	0.000	0.000	1.000	0.002
Panel D. At-the-money (2,052,942 observations)					
RMSE	2.339	2.002	2.126	1.721	1.802
MAE	0.657	0.563	0.566	0.498	0.508
MAPE	0.145	0.133	0.136	0.120	0.130
MCS-p	0.000	0.000	0.000	1.000	0.035
Panel E. Just in-the-money (2,071,177 observations)					
RMSE	1.970	1.637	1.702	1.488	1.589
MAE	0.533	0.466	0.467	0.408	0.417
MAPE	0.085	0.081	0.084	0.072	0.078
MCS-p	0.000	0.000	0.000	1.000	0.002
Panel F. In-the-money (882,661 observations)					
RMSE	1.765	1.461	1.553	1.278	1.470
MAE	0.456	0.403	0.410	0.359	0.377
MAPE	0.054	0.053	0.057	0.048	0.056
MCS-p	0.000	0.000	0.000	1.000	0.000
Panel G. Deep in-the-money (517,296 observations)					
RMSE	1.760	1.465	1.499	1.306	1.478
MAE	0.430	0.374	0.375	0.338	0.351
MAPE	0.039	0.038	0.041	0.035	0.042
MCS-p	0.000	0.000	0.000	1.000	0.000

Table 1.11. Pricing performance of individual models for different maturity options. This table presents the out-of-sample pricing errors of the models that are trained on different maturity options individually. The out-of-sample period is from January 1997 to December 2021. RMSE, MAE, and MAPE refer to the root mean squared error, the mean absolute error, and the mean absolute percentage error, respectively. MCS-p refers to the p-value of the model confidence set (MCS) test, where a high p-value indicates a model is more likely to be part of the MCS. The details of each model can be found in Section 1.4.3.

	BI	GPF	HBD	GPF_F	HBD_F
Panel A. Near-term (10,004,961 observations)					
RMSE	1.816	1.611	1.675	1.448	1.532
MAE	0.522	0.452	0.454	0.384	0.395
MAPE	0.220	0.195	0.210	0.166	0.184
MCS-p	0.000	0.000	0.000	1.000	0.011
Panel B. Mid-term (3,905,431 observations)					
RMSE	2.118	1.396	1.419	1.135	1.191
MAE	0.582	0.427	0.437	0.340	0.348
MAPE	0.184	0.139	0.141	0.105	0.113
MCS-p	0.000	0.000	0.000	1.000	0.002
Panel C. Long-term (1,228,960 observations)					
RMSE	3.050	1.801	1.805	1.355	1.544
MAE	0.866	0.545	0.544	0.436	0.463
MAPE	0.180	0.121	0.128	0.099	0.111
MCS-p	0.000	0.000	0.000	1.000	0.019

performance of the models evaluated using all options.

Firstly, when comparing three models that do not incorporate firm characteristics (DVF, GPF, and HBD), GPF demonstrates the best performance, followed by HBD. The RMSEs for GPF, HBD, and DVF are 3.396, 3.536, and 3.743, respectively. The superiority of GPF is evident across other evaluation metrics as well.

Incorporating firm characteristics, both GPF and HBD exhibit smaller pricing errors compared to their counterparts. The RMSEs for GPF_F and $\text{HBD}_F$are 1.449 and 1.755, respectively, which are lower than the RMSEs of GPF and HBD at 3.396 and 3.536. The inclusion of firm characteristics enhances the pricing performance of both GPF and HBD, making GPF_F the top-performing model among all five. The MCS p-value confirms that GPF_F is the only model included in the MCS.

Significantly, the GPF models outperform the HBD models and BS in the BS-based models. This implies that GPF exhibits greater stability than HBD when utilizing different parametric models. However, it should be noted that the accuracy of the pricing error prediction method is heavily reliant on the accuracy of the parametric model, such as HBD.

Table 1.12. Pricing performance for all options (BS models). This table presents the out-of-sample pricing errors of each model for all the options in the sample. The out-of-sample period is from January 1997 to December 2021. RMSE, MAE, and MAPE refer to the root mean squared error, the mean absolute error, and the mean absolute percentage error, respectively. MCS-p refers to the p-value of the model confidence set (MCS) test, where a high p-value indicates a model is more likely to be part of the MCS. The details of each model can be found in Section 1.4.3.

	BI	GPF	HBD	GPF_F	HBD_F
RMSE	3.743	3.396	3.536	1.449	1.755
MAE	1.125	0.995	1.041	0.412	0.500
MAPE	0.620	0.567	0.578	0.174	0.201
MCS-p	0.000	0.000	0.000	1.000	0.019

1.7 Conclusion

This paper proposes machine learning-based option pricing models that incorporate firm characteristics. Individual stock option prices can be affected by the prospects of the underlying firm and our aim is to assess whether firm characteristics can enhance the pricing of stock options. We employ two semi-parametric models, a variant of Andreou et al. (2010)'s generalized parametric function model (GPF) and a variant of Lajbcygier and Connor (1997)'s hybrid model (HBD), and evaluate them using individual stock options in the US market in the period from 1996 to 2021. The results suggest that both GPF and HBD are effective in pricing American options and firm characteristics can significantly improve the performance of these models. Between the two models, GPF consistently performs better. Among the firm characteristics, conditional skewness, Dimson Beta, dividend yield, one-year-lagged annual return, and downside beta are found to be the most important features in predicting option prices. We contribute to the option pricing literature by making the first attempt to incorporate firm characteristics into option pricing and demonstrating its effectiveness.

Appendices

.1 Firm characteristics

Abbreviation	Description	Author	Year	Journal
age	Firm age	Jiang, Lee, and Zhang	2005	RAS
aliq_at	Asset liquidity to book assets	Ortiz-Molina and Phillips	2014	JFQA
ami_126d	Illiquidity	Amihud	2002	JFM
at_be	Book leverage	Fama and French	1992	JF
at_gr1	Asset growth	Cooper, Gulen, and Schill	2008	JF
at_me	Assets-to-market	Fama and French	1992	JF
at_turnover	Capital turnover	Haugen and Baker	1996	JFE
be_gr1a	Change in common equity	Richardson et al.	2005	JAE
be_me	Book-to-market	Rosenberg, Reid, and Lanstein	1985	JF
beta_dimson_21d	Dimson Beta	Dimson	1979	JFE
betadown_252d	Downside beta	Ang, Chen, and Xing	2006	RFS
bev_mev	Book-to-market enterprise value	Penman, Richardson, and Tuna	2007	JAR
bidaskhl_21d	High-low bid-ask spread	Corwin and Schultz	2012	JF
capx_gr1	CAPEX growth	Xie	2001	AR
cash_at	Cash-to-assets	Palazzo	2012	JFE
chcsho_12m	Net stock issues	Pontiff and Woodgate	2008	JF
coa_gr1a	Change in current operating assets	Richardson et al.	2005	JAE
col_gr1a	Change in current Operating liabilities	Richardson et al.	2005	JAE
cop_at	Cash-based operating profitability	Ball et al.	2016	JFE
cop_atl1	Cash-based operating profits to lagged assets	Ball et al.	2016	JFE
coskew_21d	Coskewness	Harvey and Siddique	2000	JF
cowc_gr1a	Change in net non-cash working capital	Richardson et al.	2005	JAE
dbnetis_at	Net debt finance	Bradshaw, Richardson, and Sloan	2006	JAE
debt_me	Debt to market	Bhandari	1988	JFE
dgp_dsale	Gross margin growth to sales growth	Abarbanell and Bushee	1998	AR
div12m_me	Dividend yield	Litzenberger and Ramaswamy	1979	JF
dolvol_126d	Dollar trading volume	Brennan, Chordia, and Subrahmanyam	1998	JFE
dolvol_var_126d	Volatility of dollar trading volume	Chordia, Subrahmanyam, and Anshuman	2001	JFE

Continued on the next page

dsale_drec	Sales growth to receivable growth	Abarbanell and Bushee	1998	AR
ebit_bev	Return on net operating assets	Soliman	2008	AR
ebit_sale	Profit margin	Soliman	2008	AR
ebitda_mev	Enterprise multiple	Loughran and Wellman	2011	JFQA
emp_gr1	Employment growth	Belo, Lin, and Bazdresch	2014	JPE
eqnetis_at	Net equity finance	Bradshaw, Richardson, and Sloan	2006	JAE
eqnpo_12m	Composite equity issuance	Daniel and Titman	2006	JF
eqnpo_me	Net payout yield	Boudoukh et al.	2007	JF
eqpo_me	Payout yield	Boudoukh et al.	2007	JF
f_score	Piotroski F-score	Piotroski	2000	AR
fcf_me	Cash flow-to-price	Lakonishok, Shleifer, and Vishny	1994	JF
fnl_gr1a	Change in financial liabilities	Richardson et al.	2005	JAE
gp_at	Gross profits-to-assets	Novy-Marx	2013	JFE
gp_atl1	Gross profits-to-lagged assets	Novy-Marx	2013	JFE
inv_gr1a	Inventory change	Thomas and Zhang	2002	RAS
iskew_capm_21d	Idiosyncratic skewness (CAPM)	Bali, Engle, and Murray	2016	BOOK
iskew_ff3_21d	Idiosyncratic skewness (FF3)	Bali, Engle, and Murray	2016	BOOK
iskew_hxz4_21d	Idiosyncratic skewness (q-factor)	Bali, Engle, and Murray	2016	BOOK
ivol_capm_21d	Idiosyncratic volatility (CAPM)	Ang et al.	2006	JF
ivol_capm_252d	Idiosyncratic volatility	Ali, Hwang, and Trombley	2003	JFE
ivol_ff3_21d	Idiosyncratic volatility (FF3)	Ang et al.	2006	JF
ivol_hxz4_21d	Idiosyncratic volatility (q-factor)	Ang et al.	2006	JF
lnoa_gr1a	Change in long-term net operating assets	Fairfield, Whisenant, and Yohn	2003	AR
lti_gr1a	Chagne in long-term investments	Richardson et al.	2005	JAE
market_equity	Market equity	Banz	1981	JFE
mispricing_mgmt	Mispricing factor: Management	Stambaugh and Yuan	2016	RFS
mispricing_perf	Mispricing factor: Performance	Stambaugh and Yuan	2016	RFS
ncoa_gr1a	Change in non-current operating assets	Richardson et al.	2005	JAE
ncol_gr1a	Change in non-current operating liabilities	Richardson et al.	2005	JAE
netdebt_me	Net debt-to-price	Penman, Richardson, and Tuna	2007	JAR
netis_at	Net external finance	Bradshaw, Richardson, and Sloan	2006	JAE
nfna_gr1a	Change in net financial assets	Richardson et al.	2005	JAE
ni_be	Return on equity	Haugen and Baker	1996	JFE
ni_inc8q	Number of consecutive quarters with earnings in...	Barth, Elliott, and Finn	1999	JAR

Continued on the next page

ni_me	Earnings to price	Basu	1983	JFE
niq_at	Quarterly return on assets	Balakrishnan, Bartov, and Faurel	2010	JAE
niq_at_chg1	Change in quarterly return on assets	Balakrishnan, Bartov, and Faurel	2010	JAE
niq_be	Return on equity (quarterly)	Hou, Xue, and Zhang	2015	RFS
niq_be_chg1	Change in quarterly return on equity	Balakrishnan, Bartov, and Faurel	2010	JAE
niq_su	Earnings surprise	Foster, Olsen, and Shevlin	1984	AR
nncoa_gr1a	Change in net non-current operating assets	Richardson et al.	2005	JAE
noa_at	Net operating assets	Hirshleifer et al.	2004	JAE
noa_gr1a	Change in net operating assets	Hirshleifer et al.	2004	JAE
oaccruals_at	Operating accruals	Sloan	1996	AR
oaccruals_ni	Percent operating accruals	Hafzalla, Lundholm, and Van Winkle	2011	AR
ocf_at	Operating cash flow to assets	Bouchard et al.	2019	JF
ocf_at_chg1	Change in operating cash flow to assets	Bouchard et al.	2019	JF
ocf_me	Operating Cash flows to price	Desai, Rajgopal, and Venkatachalam	2004	AR
op_at	Operating profits-to-assets	Ball et al.	2016	JFE
op_atl1	Operating profits-to-lagged assets	Ball et al.	2016	JFE
opex_at	Operating leverage	Novy-Marx	2011	JFE
prc_highprc_252d	52-week high	George and Hwang	2004	JF
qmj_prof	Quality minus Junk: Profitability	Asness, Frazzini, and Pedersen	2018	RAS
qmj_safety	Quality minus Junk: Safety	Asness, Frazzini, and Pedersen	2018	RAS
resff3_12_1	12 month residual momentum	Blitz, Huij, and Martens	2011	JEF
resff3_6_1	6 month residual momentum	Blitz, Huij, and Martens	2011	JEF
ret_1_0	Short-term reversal	Jegadeesh	1990	JF
ret_12_1	Momentum (12 month)	Jegadeesh and Titman	1993	JF
ret_12_7	Intermediate momentum (7-12)	Novy-Marx	2012	ROF
ret_3_1	Momentum (3 month)	Jegadeesh and Titman	1993	JF
ret_6_1	Momentum (6 month)	Jegadeesh and Titman	1993	JF
ret_9_1	Momentum (9 month)	Jegadeesh and Titman	1993	JF
rmax1_21d	Maximum daily return	Bali, Cakici, and Whitelaw	2011	JFE
rmax5_21d	Highest 5 days of return	Bali, Brown, and Tang	2017	JFE
rmax5_rvol_21d	Highest 5 days of return to volatility	Assness et al.	2020	JFE
rskew_21d	Return skewness	Bali, Engle, and Murray	2016	BOOK
rvol_21d	Return volatility	Ang et al.	2006	JF

Continued on the next page

sale_bev	Asset turnover	Soliman	2008	AR
sale_gr1	Annual sales growth	Lakonishok, Shleifer, and Vishny	1994	JF
sale_me	Sales to price	Barbee, Mukherji, and Raines	1996	FAJ
saleq_su	Revenue surprise	Jegadeesh and Livnat	2006	JFE
seas_1_1an	Year 1-lagged return, annual	Heston and Sadka	2008	JFE
seas_1_1na	Year 1-lagged return, nonannual	Heston and Sadka	2008	JFE
sti_gr1a	Change in short-term investments	Richardson et al.	2005	JAE
taccruals_at	Total accruals	Richardson et al.	2005	JAE
taccruals_ni	Percent total accruals	Hafzalla, Lundholm, and Van Winkle	2011	AR
tangibility	Tangibility	Hahn and Lee	2009	JF
tax_gr1a	Tax expense surprise	Thomas and Zhang	2011	JAR
turnover_126d	Share turnover	Datar, Naik, and Radcliffe	1998	JFM
turnover_var_126d	Volatility of share turnover	Chordia, Subrahmanyam, and Anshuman	2001	JFE
zero_trades_126d	Zero-trading days (6 months)	Liu	2006	JFE
zero_trades_21d	Zero-trading days (1 month)	Liu	2006	JFE
zero_trades_252d	Zero-trading days (12 months)	Liu	2006	JFE

Chapter 2

Can option characteristics provide leading information about stocks' extreme returns?

Abstract

This study examines the predictive ability of option-implied volatility and Greeks via machine learning model (LightGBM) to forecast stock extreme returns. We analyze U.S. stock market data from January 1996 to December 2022. Our findings reveal that option characteristics significantly enhance prediction accuracy. The LightGBM model surpasses logistic regression in forecasting stock extreme returns. We also construct investment portfolios that notably outperform in Sharpe ratios and average returns, demonstrating the value of option characteristics in predicting extreme stock returns and advancing market prediction strategies. This research enriches financial market analysis and suggests areas for further enhancement.

Keywords: Option Greeks; Multi-class; Stock Extreme Returns; Machine Learning.

2.1 Introduction

Studies have demonstrated that valuable insights about future stock returns can be obtained from the options market (Chiras and Manaster, 1978; Bates, 1991; Claessen and Mittnik, 2002; Ofek et al., 2004; Xing et al., 2010; Cremers and Weinbaum, 2010). Investors seek to maximize profits by trading options when they anticipate significant price changes in the underlying asset. Such investment activities can lead options to contain information about their underlying stock's extreme return. We examine whether option characteristics (implied volatility and Greeks) can predict jumps and crashes of the underlying stocks.

Studies have shown that the implied volatilities of options contain leading information about stock prices. Bates (1991) finds that out-of-the-money put options became abnormally expensive prior to the 1987 crash. Doran et al. (2007) demonstrate that the option-implied volatility skew can predict short-term stock performance, with the put volatility skew being particularly useful in forecasting short-term market declines. Ni, Pan, and Poteshman (2008) find evidence to support the hypothesis that option traders trade on volatility information, which has implications for the stock market. Specifically, they find that the demand for volatility in the options market predicts the future realized volatility of the underlying stock. They also propose that investors trade on volatility information in the options market, which is subsequently reflected in the underlying stock. Van Buskirk (2009) investigate the relationship between the firm-level option-implied volatility skew and the probability of extreme negative events and finds that negative jumps in earnings announcements during short-window periods can be predicted when the option-implied volatility skew is high. Xing et al. (2010) also find that the shape of the volatility smirk is a strong predictor of future stock returns and that stocks with steep option volatility smirks perform worse than those with flatter option volatility smirks. Ofek et al. (2004) conduct a study to examine the put-call parity under no-arbitrage conditions and with restrictions on short sales. They find that the extent of the violation of put-call parity and the cost of short selling provide valuable predictive power for future stock returns. Similarly, Cremers and Weinbaum (2010) propose that the deviation of put-call parity can be used to forecast stock returns. Bali and Hovakimian (2009) investigate the relationship between implied volatility and expected stock returns. They first demonstrate a

negative and significant relationship between expected returns and the realized-implied volatility spread, which can be viewed as a proxy variable for volatility risk. They also show evidence of a significant positive link between expected returns and the call-put option implied volatility spread, which can be viewed as a proxy for jump risk. These studies suggest that option-implied volatility can be a valuable predictor of future stock returns.

Besides implied volatility, option Greeks may contain additional information about future stock returns. As illustrated in Section 2.2, when stocks experience jumps or crashes in the subsequent month, their options exhibit notably different implied volatility and Greeks compared to those associated with stocks performing normally. The implied volatilities of both calls and puts are significantly higher when the underlying stock jumps or crashes in the next month. Specifically, the average implied volatilities of the call and put options of jump stocks are respectively 0.70 and 0.72, whereas those of crash stocks are 0.73 and 0.77. In contrast, the average implied volatilities of the call and put options of normal stocks are respectively 0.44 and 0.45 and are statistically significantly lower than those of jump or crash stocks. These observations are consistent with Bates (1991), who finds that out-of-the-money put options were overly expensive before the 1987 crash. Bakshi and Kapadia (2003) also note that investors tend to pay higher prices for options under volatile market conditions. The deltas of call (put) options are significantly higher (lower) when their underlying stocks jump or crash. The delta of a call (put) option increases (decreases) with the volatility and this result can be partially attributed to the higher implied volatility of jump and crash stocks. The gammas of both call and put options associated with jump or crash stocks are significantly higher than the gammas of the options associated with normal stocks. Theta's lower value in options for stocks that experience jumps or crashes in the next month, relative to normal stocks, indicates they suffer less time decay before such events. Increased market interest and trading in volatile stocks may lead to more options trading, potentially mitigating the impact of time decay. Stocks experiencing jumps or crashes in the coming month typically have lower Vega values, suggesting their option prices are less affected by underlying stock volatility changes. This may be because their higher implied volatilities already account for significant inherent volatility due to natural price instability. Therefore, the impact of additional volatility shifts on these stocks' option prices is relatively muted.

Given the notable difference in implied volatility and option Greeks across stocks experiencing jump, crash, and normal, our objective is to predict the jump and crash of a stock, exploiting the information drawn from its options' characteristics. To address the nonlinearities between option characteristics and the underlying stock's price, We employ a nonlinear multi-class classification algorithm, LightGBM (LGBM), which is a gradient-boosting algorithm developed by Microsoft. We also employ a logit model as a benchmark, following Jang and Kang (2019). The details of these algorithms are described in Section 2.2.

Our results demonstrate that the gradient-boosting machine learning model (LightGBM in our research) outperforms traditional logistic regression. The outperformance can be seen from both the model-based benchmark models (LGBM-B and Logit-B) which only contain control variables as well as the combined models (LGBM-$SDGTV$ and Logit-$SDGTV$) which contain control variables and option characteristics. Based on the raw probabilities of jump and crash, the benchmark LGBM model (LGBM-B) predicts jumps and crashes with AUCs of 0.687 and 0.672, respectively, while the benchmark Logit model (Logit-B) predicts the corresponding with AUCs of 0.667 and 0.650, respectively. In addition, the AUCs of jumps and crashes under the LGBM-$SDGTV$ model are 0.766 and 0.755, while the AUCs of jumping and crashing under the Logit-$SDGTV$ model are 0.693 and 0.703, respectively. After reclassification using the probability difference between the probability of jump and crash, the benchmark LGBM model (LGBM-B) yields AUCs of 0.523 and 0.536 respectively for jump and crash, whereas the benchmark Logit model (Logit-B) yields AUCs of 0.498 and 0.522 for the corresponding predictions. The AUCs for jump and crash are respectively 0.523 and 0.568 under LGBM-$SDGTV$, whereas they are 0.478 and 0.549 under Logit-$SDGTV$.

In addition, the financial performance of LGBM-$SDGTV$ is outstanding: the Sharpe ratio of the value-weighted long/short portfolio of LGBM-$SDGTV$ is 1.42, with an annualized return of 33% (t-statistic = 3.7), and that of the equal-weighted long/short portfolio is 1.58, with an annualized return of 31% (t-statistic = 3.34). These figures greatly exceed the non-model based benchmarks[1], which have Sharpe ratios of 0.62 (value-weighted) and 0.52 (equal-weighted), with corresponding average annual returns of 10% and 11%.

The methods outlined in this chapter effectively use option trading features, such as implied

[1] Take long positions in all available stocks.

volatility and Greeks, to predict significant stock returns. Utilizing all available market options avoids selection bias, offering a comprehensive view and capturing diverse market dynamics. The application of machine learning, notably the LightGBM model, to this extensive data set underscores the methods' practicality. These models, with their superior predictive capabilities and financial performance over logit models, provide a strong framework for forecasting stock market trends, benefiting investors and analysts (Byun and Kim, 2016; Zhan et al., 2022; Bali et al., 2023).

However, the study also acknowledges certain limitations. A primary concern is the inherent risk associated with portfolios focused on extreme return investments, as indicated by higher standard deviations and maximum drawdowns, emphasizing the importance of cautious investment and robust risk management. Additionally, using extensive options data reduces selection bias but raises issues with data handling and model complexity. Accurate and timely data is essential for model effectiveness. Furthermore, as Conrad, Kapadia, and Xing (2014); Jang and Kang (2019); Andreou, Andreou, and Lambertides (2021) demonstrate, varying definitions of market jumps and crashes can affect model performance based on the thresholds set.

Our study makes two contributions to stock market forecasting. First, we demonstrate the role of option Greeks in predicting extreme stock returns, providing a new perspective on the options market. Second, we illustrate the effectiveness of machine learning in predicting extreme stock returns. By combining option characteristics with the firm characteristics used in Jang and Kang (2019), the long-short portfolio of our approach produces a Sharpe ratio of 1.42, which exceeds Jang and Kang (2019)'s 0.80. In summary, we have made a first attempt to demonstrate the predictive potential of option Greeks, and we expect future advances in feature engineering to produce even more impressive results.

Given the potential financial implications of extreme stock market movements, it is crucial to accurately predict stock jumps and crashes. By doing so, investors can strategically avoid stocks predicted to crash, thereby mitigating potential losses. Conversely, identifying stocks likely to experience a jump can present lucrative investment opportunities. Therefore, the ability to accurately forecast these market movements is not just beneficial but essential for informed investment decision-making. This predictive capability can be enhanced through the use of sophisticated analytical tools and algorithms, which can analyse market trends and other relevant data to generate

more reliable predictions. By integrating these tools into their investment strategies, investors can navigate the market more effectively, maximising gains and minimising losses.

The structure of the study is as follows: Section 1 provides an overview of the research. In Section 2, the methodology is explained. In Section 3, the model construction process is detailed. Section 4 illustrates the main findings of the analysis on the predictive power of the option characteristics and robustness checks. Finally, Section 5 concludes the study.

2.2 Data and methodology

Our objective is to predict the jump and crash of a stock using the information drawn from the stock's options as well as other predictors in the literature. We employ two multi-class classification algorithms, Logit and LightGBM (LGBM). We choose LGBM because the option's price and other characteristics are nonlinear functions of the underlying stock's price. There can also be nonlinear interactions between the input features. The logit model has been widely used in the literature and is chosen as a benchmark. Based on the predicted probabilities of jump and crash, we construct jump, crash, and jump-minus-crash portfolios and evaluate their performance.

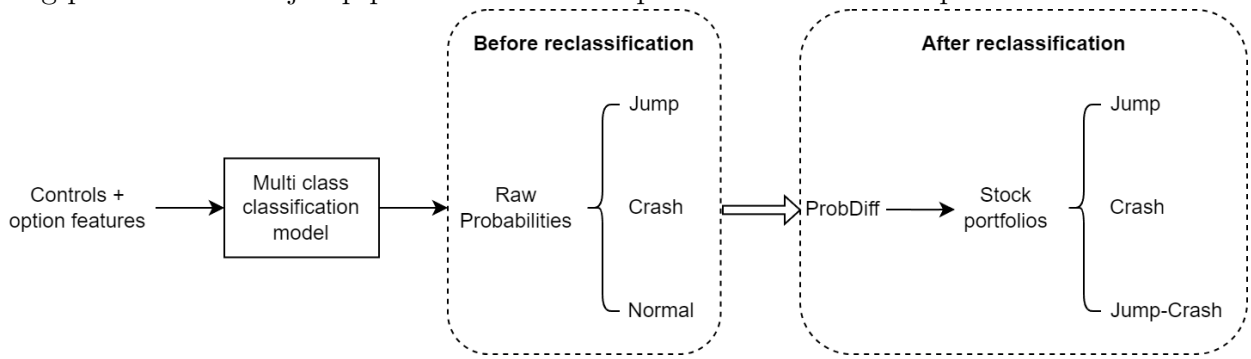
The operational workflow of our model, encompassing the entire process from data input to portfolio construction and performance evaluation, is illustrated in the flowchart presented in Figure 2.1.

2.2.1 Data

We obtain option data from OptionMetrics and it covers the period from January 1996 to December 2022. The data includes information on call and put options of exchange-listed US stocks, such as the end-of-day bid and ask prices, volume, implied volatility, Delta, Gamma, Theta, and Vega. OptionMetrics employs a binomial tree model to determine the implied volatility and Greeks of the options.

We gather stock price and accounting data from the Center for Research in Security Prices (CRSP) and Compustat. Our sample comprises common stocks that are listed on NYSE, Amex, or Nasdaq and have traded options. We exclude stocks with a closing price below one dollar at the

Figure 2.1. The flowchart illustrates the prediction procedure. It involves operating a multi class classification model using either multinomial logit regression or LightGBM. This model produces three raw probabilities corresponding to jump, crash, and normal scenarios. The probability difference (ProbDiff) is used for reclassification and is equal to the Jump probability minus the Crash probability. The ProbDiff sorting results are used to generate stock portfolios, where a higher ProbDiff means a higher probability of a jump and a lower ProbDiff means a higher probability of a crash. In each month, the jump portfolio contains long positions in stocks with higher ProbDiff, and the crash portfolio contains long positions in stocks with lower ProbDiff. The jump-crash portfolio is a long-short portfolio that contains long positions in the jump portfolio and short positions in the crash portfolio.



end of the previous month to avoid extremely illiquid stocks. The three-month Treasury bill rate from the St. Louis Federal Reserve Economic Data is used as a proxy for the risk-free rate.

Following Zhan et al. (2022), we focus exclusively on options that do not expire till the end of the following month and have a positive trading volume in order to ensure that the options are active and hold potential significance in the market during the period of interest. In addition, we apply the following filters to the option data. First, to avoid biases related to market microstructure, we only retain options, of which the trading volume and bid quote are positive, the bid price is strictly smaller than the ask price, and the mid-point of the bid and ask quote is at least 0.1 US dollar. Second, we exclude all the options that violate obvious no-arbitrage conditions.[2] Third, We only select options with maturity between 31 and 60 days at the end of the month t to ensure that the options do not expire during the prediction period. Fourth, we only retain stocks that have at least one call and one put option after filtering. Finally, We only select the last observation for each optionid for each month that meets the above requirements.

Our final sample comprises a total of 397,048 stock-month observations, with an average of

[2] For instance, we exclude anomalies such that an out-of-the-money option is more expensive than the underlying stock or an in-the-money option.

1,229 stocks per month. Table 2.1 reports the number of options per stock year by year. The number of options per stock has increased over time. On average, the number of call options has increased from 3 in 1996 to 11 in 2022, while the number of put options has increased from 2 to 11. Nevertheless, the minimum number of options remains at 1 throughout the sample period and the median number has increased only moderately (from 3 to 6 in case of call and 2 to 5 in case of put). In contrast, the maximum number has increased from 16 to 363 (call) and from 14 to 376 (put). This result implies that except for a handful of stocks with many associated options, the number of options of a stock has not increased significantly and is usually under 10.

Table 2.2 reports the implied volatility and Greeks of the options. We first group stocks into jump, crash, and normal, and calculate the implied volatilities and Greeks of the option as described in. The implied volatilities of both calls and puts are significantly higher when the underlying stock jumps or crashes in the next month. Specifically, the average implied volatilities of the call and put options of jump stocks are respectively 0.70 and 0.72, whereas those of crash stocks are 0.73 and 0.77. In contrast, the average implied volatilities of the call and put options of normal stocks are respectively 0.44 and 0.45 and are statistically significantly lower than those of jump or crash stocks. These observations are consistent with Bates (1991), who finds that out-of-the-money put options were overly expensive before the 1987 crash. Bakshi and Kapadia (2003) also note that investors tend to pay higher prices for options under volatile market conditions. The deltas of call (put) options are significantly higher (lower) when their underlying stocks jump or crash. The delta of a call (put) option increases (decreases) with the volatility and this result can be partially attributed to the higher implied volatility of jump and crash stocks. The gammas of both call and put options associated with jump or crash stocks are significantly higher than the gammas of the options associated with normal stocks. Theta's lower value in options for stocks that experience jumps or crashes in the next month, relative to normal stocks, indicates they suffer less time decay before such events. Increased market interest and trading in volatile stocks may lead to more options trading, potentially mitigating the impact of time decay. Stocks experiencing jumps or crashes in the coming month typically have lower Vega values, suggesting their option prices are less affected by underlying stock volatility changes. This may be because their higher implied volatilities already account for significant inherent volatility due to natural price instability. Therefore, the impact of

additional volatility shifts on these stocks' option prices is relatively muted.

Table 2.1. Number of options per stock. This table displays the average annual stock count (in the 'Obs.' column) and descriptive statistics—Mean, Standard Deviation, Minimum (Min), 25th (25%), 50th (50%), and 75th (75%) percentiles, and Maximum(Max)—for the number of options per stock over the sample period. The last row, 'Avg.', reports the average value for each column.

	Call							Put							
YEAR	MEAN	STD	MIN	25%	50%	75%	MAX	MEAN	STD	MIN	25%	50%	75%	MAX	OBS
1996	3	2	1	2	3	4	16	2	2	1	1	2	3	14	362
1997	3	2	1	2	3	4	17	2	2	1	1	2	3	14	485
1998	3	2	1	2	3	4	21	3	2	1	1	2	3	18	585
1999	4	4	1	2	3	4	38	3	3	1	1	2	4	33	617
2000	5	5	1	2	4	6	44	4	4	1	1	2	5	40	628
2001	4	3	1	2	3	5	23	3	3	1	1	2	4	24	638
2002	3	2	1	2	3	4	15	3	2	1	1	2	4	16	669
2003	3	2	1	2	3	4	15	3	2	1	1	2	4	14	690
2004	3	2	1	2	3	4	24	3	2	1	1	2	4	22	800
2005	3	3	1	2	3	4	37	3	2	1	1	2	4	26	874
2006	4	3	1	2	3	5	45	3	3	1	1	2	4	38	1,008
2007	4	3	1	2	3	5	40	3	3	1	2	3	4	35	1,187
2008	5	4	1	2	3	6	54	5	5	1	2	4	6	56	1,247
2009	4	4	1	2	3	5	40	5	5	1	2	3	5	43	1,243
2010	5	4	1	2	3	6	50	4	4	1	2	3	6	46	1,260
2011	6	5	1	2	4	7	64	5	5	1	2	4	7	60	1,232
2012	5	6	1	2	4	7	107	5	6	1	2	4	7	97	1,192
2013	6	7	1	2	4	7	114	5	7	1	2	3	7	108	1,351
2014	7	11	1	2	4	8	148	7	10	1	2	3	7	136	1,439
2015	8	12	1	2	4	8	162	8	12	1	2	4	8	145	1,501
2016	7	11	1	2	4	8	156	7	11	1	2	3	8	155	1,558
2017	8	11	1	2	4	8	169	7	11	1	2	3	7	157	1,741
2018	9	14	1	2	4	9	309	8	14	1	2	4	8	282	1,917
2019	9	15	1	2	4	10	280	8	14	1	2	4	9	288	1,929
2020	12	21	1	3	6	12	363	11	20	1	2	4	11	376	2,024
2021	12	20	1	3	6	12	340	10	18	1	2	4	10	325	2,622
2022	11	18	1	2	5	11	299	11	19	1	2	5	11	312	2,287
AVG.	6	7	1	2	4	7	111	5	7	1	2	3	6	107	1,225

2.2.2 Input variables

2.2.2.1 Control variables

To account for relevant factors affecting stock returns, we include control variables following Jang and Kang (2019). These variables are firm age (AGE), detrended turnover ($DTURN$), past individual return in excess of the market return ($EXRET$), tangible asset ($TANG$), total skewness ($TSKEW$), total volatility ($TVOL$), past market return ($RM12$), sales growth (SG), and firm size ($SIZE$). They are also motivated by the research of Chen et al. (2001), Hutton, Marcus, and Tehranian (2009), Boyer, Mitton, and Vorkink (2010) and Conrad et al. (2014). Variables that are

Table 2.2. Cross-sectional descriptive statistics of the implied volatility and Greeks. The stocks are categorized by monthly log-returns: above 20% as jumps, below -20% as crashes. There are 18,908 jumps, 25,064 crashes, and 352,665 normal observations in the total sample period, averaging 59 jumps, 78 crashes, and 1,092 normal monthly. The columns belonging to Call, Put, and Spread are the descriptive status of calls, puts, and the difference between calls and puts, respectively. The row t1 in each Panel reports the t-statistic of jump versus normal in the Jump column and of crash versus normal in the Crash column. The row t2 in each Panel reports the t-statistic of Jump versus Crash.

		Call			Put		
		Jump	Crash	Normal	Jump	Crash	Normal
Sigma	MEAN	0.70	0.73	0.44	0.72	0.77	0.45
	t1	40.24	38.19		38.53	38.63	
	t2	-3.13			-4.82		
Delta	MEAN	0.56	0.56	0.54	-0.44	-0.44	-0.46
	t1	31.56	31.66		34.31	35.55	
	t2	-2.91			-4.40		
Gamma	MEAN	0.14	0.14	0.11	0.14	0.13	0.11
	t1	10.60	12.13		10.46	12.07	
	t2	0.97			1.99		
Theta	MEAN	-8.53	-8.63	-9.73	-8.12	-8.38	-9.23
	t1	6.46	5.66		5.89	4.32	
	t2	0.38			1.04		
Vega	MEAN	3.96	3.94	7.14	3.94	3.92	7.12
	t1	-25.48	-26.36		-25.63	-26.46	
	t2	0.17			0.19		

updated on a monthly basis; AGE, $DTURN$, $EXRET$, $TSKEW$, $TVOL$, $RM12$, and $SIZE$, are matched with the monthly stock returns with a one-month lag, and those that are updated quarterly or annually; $TANG$ and SG, are matched with the returns with a six-month lag. The definitions of these variables are provided in the Appendix. We fill the missing values with the cross-sectional median value following Gu et al. (2020).

2.2.2.2 Option related variables

Previous studies find that implied volatility embeds valuable information for predicting stock returns (Ofek et al., 2004; Cremers and Weinbaum, 2010). We explore the predictive power of not only implied volatility but also option Greeks. Stocks have different numbers of options with varying degrees of moneyness. This poses a problem when we are to use option characteristics as input variables for the classification algorithms as they require the same number of input variables and every observation of an input variable should have the same meaning. For instance, if a stock has only an at-the-money option and another stock has only an in-the-money option, their Deltas should not be used as the same input variable as they have different implications. We need to define common input variables for all stocks through some sort of standardization. We address this challenge by assuming hypothetical at-the-money options. Specifically, for each option, we fix the implied volatility and recalculate the Greeks assuming the moneyness is 1.[3] We then calculate the average implied volatility and Greeks of all call options available in a month. We repeat the same procedure for put options. Consequently, there are ten option-related variables: the average implied volatility and Greeks (Delta, Gamma, Theta, and Vega) of call options and those of put options. The input variables are defined as the five call option characteristics and the spreads of the five characteristics between the call and put options. We use the spreads instead of the put option characteristics to avoid the multicollinearity problem while retaining the information from both options.

[3] The details can be found in the Appendix.

2.2.3 Definitions of jump and crash

We classify stock returns into three groups: jump, normal, and crash. Inspired by the method proposed in Jang and Kang (2019), We define stocks with monthly log returns above 20% as jumps and stocks below −20% as crashes. In-between returns are classified as normal. This classification method assigns approximately 5% of all stock returns to jump and crash groups, respectively, which are similar to the distribution pattern observed in Jang and Kang (2019). Further supporting our threshold selection, the study by Andreou et al. (2021) indicates that the average crash return was −17.56% (equivalent to a log return of −19.31%) for the period from 1990 to 2018. This empirical evidence further satisfies the appropriateness of our chosen thresholds of 20% and −20% for defining jump and crash in our analysis. In addition, previous studies such as Conrad et al. (2014) and Jang and Kang (2019) have shown that the results remain robust even when different threshold values are used. Therefore, slight variations in the threshold levels are not expected to impact the findings significantly (We test this in the robust tests.).

As a robustness check, we also explore dynamic thresholds for jump and crash, where a return is classified as jump (crash) if it is above (below) the 95th (5th) percentile of all stock returns over the past 24 months. Using these dynamic thresholds, we aim to capture extreme stock returns based on their relative performance within the market.

Table 2.3 reports the descriptive statistics of the returns of jump and crash stocks during the sample period. When the market is bullish, *e.g.*, the dotcom bubble in 2000 and the stock market rally during the pandemic in 2020, the proportion of jump stocks exceeds 10%, whereas when the market is bearish, *e.g.*, the burst of the dotcom bubble in 2001 and the global financial crisis in 2008, the proportion of crash stocks exceeds 10%. When the proportion of jump stocks is high, the proportion of crash stocks tends to be also high, which implies that the market becomes more volatile and stocks jump and crash more frequently in both bullish and bearish periods. By definition, the maximum returns of jump stocks (107% on average) are greater in magnitude than the minimum returns of crash stocks (-55% on average). Nevertheless, the proportion of crash stocks in a bearish market (maximum 15.50% in 2008) usually exceeds the proportion of jump stocks in a bullish market (maximum 10.87% in 2020), which implies that stocks tend to move more synchronously in a bearish market.

Table 2.3. Descriptive statistics of the jump and crash from 1996 to 2022. The table provides an overview of descriptive statistics for jump and crash between 1996 and 2022, categorized using a threshold where log-returns exceeding 20% over the next one month are considered jumps, and those below −20% over the next one month are classified as crashes. '%' denotes the percentage of the number of jumps or crashes to the total observations. 'Mean' and 'Median' refer to the average and median monthly returns, respectively. 'Std' indicates the standard deviation of the monthly returns, while 'Min' and 'Max' represent the minimum and maximum monthly returns. The last row, 'Avg.', reports the average value for each column. All these values are calculated from monthly observations for the corresponding years.

	Jump						Crash						
YEAR	%	MEAN	50%	STD	MIN	MAX	%	MEAN	50%	STD	MIN	MAX	OBS
1996	4.81	0.34	0.30	0.11	0.24	0.62	5.27	-0.26	-0.23	0.07	-0.43	-0.19	362
1997	5.83	0.32	0.29	0.09	0.23	0.63	5.48	-0.25	-0.23	0.07	-0.48	-0.19	485
1998	8.38	0.34	0.30	0.13	0.23	0.88	10.37	-0.26	-0.24	0.07	-0.48	-0.18	585
1999	9.85	0.37	0.31	0.20	0.22	1.50	6.66	-0.27	-0.24	0.08	-0.55	-0.18	617
2000	10.59	0.38	0.32	0.18	0.22	1.13	15.45	-0.29	-0.27	0.10	-0.61	-0.18	628
2001	8.56	0.34	0.30	0.14	0.23	0.93	12.17	-0.29	-0.26	0.09	-0.60	-0.19	638
2002	4.32	0.32	0.28	0.10	0.23	0.63	10.78	-0.29	-0.25	0.11	-0.68	-0.18	669
2003	4.80	0.33	0.29	0.12	0.23	0.73	2.29	-0.26	-0.23	0.08	-0.46	-0.19	690
2004	2.79	0.30	0.28	0.07	0.23	0.55	3.24	-0.25	-0.23	0.06	-0.40	-0.18	800
2005	2.16	0.30	0.28	0.07	0.23	0.46	2.63	-0.25	-0.24	0.07	-0.44	-0.19	874
2006	2.49	0.31	0.27	0.08	0.23	0.54	2.26	-0.26	-0.22	0.08	-0.48	-0.19	1,008
2007	2.17	0.32	0.28	0.12	0.22	0.74	3.43	-0.26	-0.23	0.09	-0.55	-0.18	1,187
2008	3.96	0.34	0.30	0.12	0.22	0.83	15.50	-0.27	-0.24	0.09	-0.64	-0.18	1,247
2009	9.01	0.36	0.30	0.18	0.23	1.29	6.18	-0.27	-0.25	0.08	-0.56	-0.19	1,243
2010	3.78	0.30	0.28	0.09	0.23	0.71	3.08	-0.27	-0.26	0.08	-0.48	-0.20	1,260
2011	3.68	0.32	0.27	0.12	0.23	0.73	5.77	-0.25	-0.23	0.07	-0.49	-0.18	1,232
2012	2.69	0.34	0.28	0.15	0.23	0.87	3.21	-0.26	-0.23	0.09	-0.55	-0.18	1,192
2013	2.81	0.34	0.29	0.15	0.22	0.98	2.04	-0.27	-0.24	0.10	-0.59	-0.18	1,351
2014	2.28	0.36	0.27	0.19	0.23	0.92	3.89	-0.27	-0.23	0.10	-0.61	-0.18	1,439
2015	2.90	0.34	0.30	0.13	0.22	0.86	4.73	-0.27	-0.24	0.10	-0.67	-0.18	1,501
2016	4.03	0.33	0.29	0.14	0.22	0.92	4.58	-0.28	-0.24	0.11	-0.67	-0.18	1,558
2017	2.68	0.35	0.29	0.18	0.22	1.18	2.56	-0.26	-0.24	0.08	-0.57	-0.18	1,741
2018	2.97	0.34	0.29	0.17	0.22	1.08	6.13	-0.26	-0.23	0.09	-0.62	-0.18	1,917
2019	4.57	0.35	0.30	0.19	0.22	1.47	5.09	-0.28	-0.24	0.11	-0.69	-0.18	1,929
2020	10.87	0.42	0.32	0.35	0.22	3.42	9.09	-0.28	-0.25	0.11	-0.71	-0.18	2,024
2021	5.55	0.38	0.31	0.26	0.22	2.70	6.98	-0.26	-0.23	0.09	-0.73	-0.18	2,622
2022	5.43	0.39	0.31	0.22	0.22	1.52	13.07	-0.28	-0.25	0.10	-0.78	-0.18	2,287
AVG.	4.96	0.34	0.29	0.15	0.23	1.07	6.37	-0.27	-0.24	0.09	-0.57	-0.18	1,225

2.2.4 Classification algorithms

2.2.4.1 Logit

We predict a total of three classes: normal (class 0), jump (class 1) and crash (class 2). The reference class is assumed to be the normal class (class 0). The multinomial logistic regression is summarized as follows:

$$\ln \frac{P_{1,t}}{P_{0,t}} = a_1 + \sum_{j=1}^{J} b_{1,j} X_{j,t-1}, \tag{2.1}$$

$$\ln \frac{P_{2,t}}{P_{0,t}} = a_2 + \sum_{j=1}^{J} b_{2,j} X_{j,t-1}, \tag{2.2}$$

$$P_{0,t} + P_{1,t} + P_{2,t} = 1, \tag{2.3}$$

where a_1 and a_2 refer to the intercepts for jump and crash prediction, respectively. $X_{j,t-1}$ denotes the feature j known at month $t-1$, $b_{1,j}$ is the coefficient of the feature j for jump, $b_{2,j}$ is the coefficient of the feature j for crash. $P_{0,t}$, $P_{1,t}$, and $P_{2,t}$ represent the probabilities for normal, jump and crash at month t, respectively. J represents the total number of the features.

2.2.4.2 LGBM

The machine learning algorithm we employ is LGBM proposed by Ke et al. (2017). LGBM is a gradient-boosting framework that employs decision tree-based learning algorithms for both classification and regression tasks. It is designed for efficiency and scalability when dealing with large datasets. An outline of the algorithm for multi-class classification follows:

1. Initialize the model: The initial model can be represented as:

$$F_0(x) = arg\min_c \sum L(y_i, c) \tag{2.4}$$

   where $L(y_i, c)$ is initial loss function for the target label y_i corresponding to observation i and the constant value c.

2. Gradient boosting: Iteratively construct weak learners (decision trees) and combine them to

create a strong learner. For each iteration $t = 1, 2, ..., T$:

(a) Compute the gradients g_i for each observation i and each class k:

$$g_i = -\frac{\partial L(y_i, F^k(x_i))}{\partial F^k(x_i)} \tag{2.5}$$

where $L(y_i, F_k(x_i))$ is the loss function, y_i is the true label of observation i, and $F^k(x_i)$ is the current prediction for observation i and class k.

(b) A decision tree $f_t(x)$ is fit using (x_i, g_i, h_i), where x_i is the feature vector, g_i the gradient, and h_i the Hessian for observation i. The fitting process is as follows:

i. Split Selection: Splits are chosen to maximize the gain in loss reduction, calculated as:

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma$$

where I_L and I_R are the sets of indices of observations in the left and right splits, and λ, γ are regularization parameters [4].

ii. Leaf-wise Growth: The tree grows leaf-wise by repeating the split process, aiming to maximize Gain.

(c) Update the model:

$$F_t(x) = F_{t-1}(x) + \eta_t \times f_t(x) \tag{2.6}$$

3. Final model: After T iterations, the final model can be represented as:

$$F(x) = F_0(x) + \sum_{t=1}^{T} \eta_t \times f_t(x) \tag{2.7}$$

In our research, the multi-log loss is chosen as the loss function. The LGBM includes many hyperparameters that can be tuned to optimize the performance of the model, such as the number of leaves, the depth of the tree, and the learning rate. In our study, to address the issue of unbalanced

[4] In LightGBM for multi-class prediction, λ helps smooth the model by penalizing large leaf weights, while γ controls tree complexity by imposing a cost on adding more leaves, both aiming to reduce overfitting and improve generalization. We keep the regularization parameters as default to keep conservative.

distribution, we adjusted the parameter for class sampling to ensure a balanced representation of different classes. We retained the default values for all other hyperparameters in LGBM to assess the model from a conservative standpoint.

2.2.5 Prediction evaluation

2.2.5.1 Statistical measures

We use the area under curve (AUC) as our statistical measure to evaluate the performance of the predicted models. Since we are doing a multi-class classification task, the AUC is computed using the One-vs-Rest (also known as One-vs-All) strategy. The One-vs-Rest strategy and the calculation of AUC for multi-class classification using this strategy allow for the assessment of model performance in scenarios where there are more than two classes.

The AUC is the area under the receiver operating characteristic (ROC) curve and represents the average predictive power of the model (Hanley and McNeil, 1982). The ROC curve is a graphical interpretation of the classifier's true positive rate (TPR) and the false positive rate (FPR). The TPR is the number of true positive samples divided by the sum of true positive and false negative samples, also known as sensitivity. The FPR is the number of false positive samples divided by the sum of false positive and true negative samples. The ROC curve represents a compromise between the TPR and FPR for different classifier performances at different decision thresholds. Each point on the curve represents the corresponding threshold in the classifier's decision function. The AUC represents the mean value, which can objectively reflect the ability to predict positive and negative samples in aggregate and consider the effect of removing sample skew, especially for unbalanced data. In addition, AUC can better describe the learning effect of the model when the sample data requires attention to cost sensitivity. A value of 1.0 for AUC means that the model is a perfect classifier. In contrast, 0.5 implies a random classifier.

To calculate the AUC for multi-class classification using the One-vs-Rest strategy, the following steps are typically followed:

1. A separate binary classifier is trained for each class, where the objective is to predict whether an instance belongs to that particular class or not. The training data for each classifier

comprises instances from the positive class (the specific class of interest) and instances from the negative class (the remaining classes).

2. Predicted probabilities or scores are generated after training the binary classifiers. These probabilities or scores reflect the model's confidence in assigning an instance to the positive class (the specific class) or the negative class (the rest of the classes).

3. The AUC is computed for each class by comparing the predicted probabilities or scores of that class with the scores of the negative class. A commonly used method for this calculation is the binary classification AUC computation, such as the ROC curve.

4. To obtain an overall measure of the multi-class AUC, the AUC scores for each class are averaged. This average provides an indication of the model's discriminative ability across all classes, summarising its performance in distinguishing between different classes.

2.2.5.2 Financial measures

We test the financial performance of the portfolios we proposed in Section 2.2.6. Five metrics are employed to assess the financial performance of the portfolio: average return (MEAN), standard deviation (Std), Sharpe ratio (SR), cumulative return (Cum), and maximum drawdown (MDD). The Sharpe ratio and maximum drawdown are calculated based on the following equations.

$$SR = \frac{excess\ return}{standard\ deviation}, \tag{2.8}$$

$$MDD = the\ maximum\ observed\ loss\ from\ a\ peak\ to\ a\ trough. \tag{2.9}$$

Our study primarily emphasizes the performance of value-weighted investments, with equal-weighted investments being reported in the robustness check for comparison. Given our strategy involves monthly re-balancing, transaction costs are an inevitable part of the process. These costs are taken into account and tested in our robustness check to ensure the validity of our findings.

2.2.6 Implementation details

We train the models every month starting from January 1997 using all available data on the training date and use them to predict the returns in the subsequent month: To predict the returns in January 1997, the models are trained using the data from January 1996 to December 1996; and for the returns in February 1997, they are trained with the data from January 1996 to January 1997, and so on. Logit models use all the samples in the training set to estimate the parameters, whereas LGBM models use the first 80% of the set for training and the rest for validation.

Table 2.4 describes the input variables and the models. The input variables are cross-sectionally normalized by subtracting the cross-sectional mean and dividing by the cross-sectional standard deviation. The models that start with 'Logit' are logit-based models and those that start with 'LGBM' are LGBM-based models. For each classification algorithm, we consider seven variants. The 'B' denotes a benchmark model that employs only the control variables; 'SDGTV' denotes a model that employs the control variables and all the option-related variables; and 'S' ('D', 'G', 'T', 'V') denotes a model that employs the control variables and the implied volatility (Delta, Gamma, Theta, Vega).

LGBM has many hyperparameters that can be tuned to optimize the model's performance, such as the number of leaves, the depth of the tree, and the learning rate. To assess the models from a conservative standpoint, we opt to retain the default values for all the hyperparameters except *class_weight*, which we set to 'balanced' to address the imbalanced data issue: the number of jump and crash stocks being significantly smaller than that of normal stocks.

The models predict the probabilities for jump, crash, and normal. Then a stock is classified as jump (crash) when its jump (crash) probability is the highest. Instead of following this conventional method, we use the probability difference between jump and crash (ProbDiff), as suggested by Han (2021). Han (2021) points out that a stock with a high jump probability often has a high crash probability. If a stock's jump, normal, and crash probabilities are respectively 40%, 21%, and 39%, the conventional method will classify the stock into jump, although it is equally likely to crash. Using ProbDiff avoids such misclassification.[5]

[5] Jang and Kang (2019) rely solely on the crash probability to categorize stocks. Such a method does not fully exploit all the predictive information.

Table 2.4. Model specifications. This table presents the specifications of the test models. Panel (a) lists the input features of the models, and panel (b) lists the models tested in the empirical study.

(a) Input features

Call features	
σ_c	The average implied volatility of call options.
Δ_c	The average Delta of call options.
Γ_c	The average Gamma of call options.
Θ_c	The average Theta of call options.
V_c	The average Vega of call options.
Spread features	
σ_s	$\sigma_c - \sigma_p$
Δ_s	$\Delta_c - \Delta_p$
Γ_s	$\Gamma_c - \Gamma_p$
Θ_s	$\Theta_c - \Theta_p$
V_s	$V_c - V_p$
Control Variables	
	financial leverage (LEV), return of asset (ROA), market-to-book ratio (MB), firm size ($SIZE$), past market return ($RM12\_VW$), firm age (AGE), past individual return in excess of the market return ($EXRET12$), total volatility ($TVOL$), total skewness ($TSKEW$), tangible asset ($TANG$), sales growth (SG) and detrended turnover ($DTURN$)

(b) Test models

Model	Input features	Algorithm
Logit-B	Controls	Logit
Logit-$SDGTV$	Controls + All option features	Logit
Logit-S	Controls + $\sigma_c + \sigma_s$	Logit
Logit-D	Controls + $\Delta_c + \Delta_s$	Logit
Logit-G	Controls + $\Gamma_c + \Gamma_s$	Logit
Logit-T	Controls + $\Theta_c + \Theta_s$	Logit
Logit-V	Controls + $V_c + V_s$	Logit
LGBM-B	*Controls*	LGBM
LGBM-$SDGTV$	Controls + All option features	LGBM
LGBM-S	Controls + $\sigma_c + \sigma_s$	LGBM
LGBM-D	Controls + $\Delta_c + \Delta_s$	LGBM
LGBM-G	Controls + $\Gamma_c + \Gamma_s$	LGBM
LGBM-T	Controls + $\Theta_c + \Theta_s$	LGBM
LGBM-V	Controls + $V_c + V_s$	LGBM

We create three types of portfolios based on the ProbDiff measure. A jump portfolio consists of the top 5% stocks in terms of ProbDiff, a crash portfolio consists of the bottom 5% stocks, and a jump-crash portfolio is a long-short portfolio that buys a jump portfolio and shorts a crash portfolio. Each portfolio is rebalanced every month.

2.3 Empirical results

In this section, we evaluate the predictive power of the proposed models using various measures and focusing on the informativeness of the option-related variables. All the results presented in this section are out-of-sample results.

2.3.1 Statistical Performance

Table 2.5 reports the statistical performance measures of the models. Both LGBM and Logit models demonstrate predictive power for stocks' jump and crash, as evidenced by the AUCs significantly greater than 0. Between LGBM and Logit, LGBM outperforms Logit across all the sub-models. It also shows that adding option-related variables increase the AUC under both models.

Before reclassification, the benchmark LGBM model (LGBM-B) predicts jumps and crashes with AUCs of 0.687 and 0.672, respectively, while the benchmark Logit model (Logit-B) predicts the corresponding with AUCs of 0.667 and 0.650, respectively. In addition, the AUCs of jumps and crashes under the LGBM-$SDGTV$ model are 0.766 and 0.755, while the AUCs of jumping and crashing under the Logit-$SDGTV$ model are 0.693 and 0.703, respectively.

After reclassification, the benchmark LGBM model (LGBM-B) yields AUCs of 0.523 and 0.536 respectively for jump and crash, whereas the benchmark Logit model (Logit-B) yields AUCs of 0.498 and 0.522 for the corresponding predictions. The AUCs for jump and crash are respectively 0.523 and 0.568 under LGBM-$SDGTV$, whereas they are 0.478 and 0.549 under Logit-$SDGTV$.

While reclassification reduces accuracy, it provides more valuable insights for forecasting. Reclassification is based on the sorting of ProbDif, which internally assigns stocks with a high probability of a jump or crash to the normal group, whereas the raw probabilities assign such stocks to

both the jump and crash groups. It is true that the reclassification method will move some stocks that actually jump or crash to the normal group, but this method is meaningless if we categorize a stock into both the jump and crash groups, as it does not provide valuable insights for portfolio construction.

Among the sub-models, the models containing all option-related variables (LGBM-$SDGTV$ and Logit-$SDGTV$) perform best before and after reclassification. Among the sub-models using a single option feature, all option features provided useful information for predicting jumps and crashes before reclassification and useful information for predicting crashes after reclassification.

In conclusion, the above data suggests that Option-related variables can improve jump and crash performance, and the improvement is more significant under LGBM.

2.3.2 Financial performance

2.3.2.1 Overall financial performance

Table 2.6 reports the financial performance of the value-weighted portfolios formed by the LGBM and Logit models over the out-of-sample period from January 1997 to December 2022. In the table, VW refers to a value-weighted portfolio that invests in all the stocks in the sample. As we are mainly interested in the informativeness of the option characteristics, we calculate the t-statistics of the mean return against the benchmark models, LGBM-B and Logit-B.

The LGBM-based models consistently outperform the Logit-based models across all sets of input variables in terms of the Sharpe ratio, and the difference becomes more evident when option-related variables are included. For instance, the jump-crash portfolios of LGBM-B and LGBM-$SDGTV$ respectively yield annualized Sharpe ratios of 0.75 and 1.42, whereas those of Logit-B and Logit-$SDGTV$ yield annualized Sharpe ratios of 0.21 and 0.49. The Sharpe ratio of the Logit-B and -$SDGTV$ are even lower than that of the market portfolio VW, which is 0.62.

The results from the LGBM models reveal that all the models incorporating option characteristics outperform the benchmark LGBM-B. The most prominent one is LGBM-$SDGTV$, which achieves a Sharpe ratio of 1.42 and a mean return of 33% (t-statistic = 3.7). These values are higher than the Sharpe ratio and mean return of LGBM-B, which are 0.75 and 20%, respectively.

Table 2.5. Accuracy of jump and crash predictions. The table presents the statistical performance of various models over the test period of January 1997 to December 2022. The results are presented regarding the AUCs of jump and crash stocks for LGBM (Panel A) and Logit (Panel B), respectively. The dataset contains 18,908 jumps, 25,064 crashes, and 352,665 normal observations within the out-of-sample period. PPV: Positive predictive value (precision) FDR: False discovery rate

Panel A. Before reclassification

		B	*SDGTV*	*S*	*D*	*G*	*T*	*V*
LGBM								
Jump	AUC	0.687	0.766	0.754	0.749	0.703	0.704	0.716
	PPV	0.129	0.156	0.148	0.152	0.133	0.138	0.133
	FDR	0.106	0.154	0.147	0.152	0.120	0.120	0.128
Crash	AUC	0.672	0.755	0.741	0.742	0.676	0.681	0.698
	PPV	0.143	0.210	0.200	0.204	0.153	0.152	0.162
	FDR	0.095	0.121	0.119	0.123	0.098	0.095	0.098
Logit								
Jump	AUC	0.667	0.693	0.699	0.690	0.667	0.666	0.668
	PPV	0.200	0.178	0.192	0.197	0.200	0.197	0.200
	FDR	0.220	0.244	0.246	0.237	0.216	0.220	0.222
Crash	AUC	0.650	0.703	0.708	0.704	0.651	0.646	0.655
	PPV	0.189	0.248	0.258	0.257	0.193	0.186	0.189
	FDR	0.267	0.260	0.264	0.255	0.263	0.261	0.264

Panel B. After reclassification

		B	*SDGTV*	*S*	*D*	*G*	*T*	*V*
LGBM								
Jump	AUC	0.523	0.523	0.523	0.518	0.531	0.529	0.532
	PPV	0.123	0.151	0.141	0.144	0.133	0.134	0.129
	FDR	0.107	0.140	0.135	0.140	0.114	0.116	0.120
Crash	AUC	0.536	0.568	0.559	0.568	0.540	0.548	0.549
	PPV	0.149	0.201	0.194	0.197	0.151	0.151	0.160
	FDR	0.086	0.113	0.112	0.113	0.091	0.088	0.091
Logit								
Jump	AUC	0.498	0.478	0.483	0.488	0.498	0.497	0.492
	PPV	0.149	0.141	0.152	0.156	0.152	0.150	0.144
	FDR	0.143	0.166	0.149	0.142	0.142	0.152	0.149
Crash	AUC	0.522	0.549	0.555	0.559	0.526	0.520	0.529
	PPV	0.155	0.210	0.227	0.227	0.159	0.154	0.163
	FDR	0.220	0.241	0.248	0.242	0.222	0.214	0.221

Table 2.6. Performance of Value-Weighted Portfolios. The table presents the annually measured financial performance of the models from January 1997 to December 2022. The performance metrics include average return (MEAN), standard deviation (Std), Sharpe ratio (SR), and maximum drawdown (MDD). The Newey-West tests (column $t1$ and $t2$) serve as significant test for return, with B as the benchmark ($t1$) and VW as the benchmark ($t2$). The dataset contains 18,908 jumps, 25,064 crashes, and 352,665 normal observations within the test period.

Panel A. LGBM models

	Jump						Crash						Jump-Crash					
	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD
VW	0.10			0.16	0.62	-0.51	0.10			0.16	0.62	-0.51	0.10			0.16	0.62	-0.51
B	0.20		(2.29)	0.30	0.65	-0.57	-0.00		(-2.51)	0.33	-0.01	-0.84	0.20		(1.49)	0.26	0.75	-0.47
$SDGTV$	0.22	(0.54)	(2.44)	0.34	0.64	-0.72	-0.12	(-3.25)	(-4.6)	0.34	-0.34	-0.99	0.33	(2.67)	(3.7)	0.23	1.42	-0.47
S	0.18	(-0.28)	(1.99)	0.32	0.57	-0.70	-0.08	(-1.88)	(-3.84)	0.37	-0.23	-0.99	0.27	(1.2)	(2.92)	0.26	1.04	-0.47
D	0.19	(-0.04)	(2.1)	0.32	0.60	-0.73	-0.03	(-0.74)	(-2.76)	0.37	-0.08	-0.95	0.22	(0.57)	(1.99)	0.26	0.88	-0.47
G	0.21	(0.5)	(2.74)	0.32	0.67	-0.65	0.03	(1.2)	(-2.03)	0.29	0.10	-0.76	0.18	(-0.33)	(1.64)	0.22	0.84	-0.51
T	0.20	(0.2)	(2.23)	0.31	0.66	-0.63	-0.01	(-0.23)	(-2.88)	0.31	-0.04	-0.90	0.21	(0.31)	(1.9)	0.23	0.92	-0.36
V	0.17	(-0.71)	(1.6)	0.32	0.52	-0.70	-0.02	(-0.45)	(-3.01)	0.32	-0.06	-0.90	0.19	(-0.25)	(1.56)	0.22	0.87	-0.33

Panel B. Logit models

	Jump						Crash						Jump-Crash					
	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD
VW	0.10			0.16	0.62	-0.51	0.10			0.16	0.62	-0.51	0.10			0.16	0.62	-0.51
B	0.16		(1.24)	0.32	0.50	-0.75	0.09		(-0.23)	0.28	0.34	-0.82	0.07		(-0.47)	0.32	0.21	-0.90
$SDGTV$	0.17	(0.14)	(1.54)	0.30	0.56	-0.60	0.02	(-2.75)	(-1.84)	0.31	0.07	-0.94	0.14	(1.56)	(0.55)	0.30	0.49	-0.62
S	0.17	(0.48)	(1.38)	0.33	0.52	-0.75	-0.01	(-2.83)	(-2.53)	0.33	-0.03	-0.92	0.18	(2.5)	(1.02)	0.33	0.54	-0.63
D	0.14	(-0.65)	(0.83)	0.34	0.42	-0.78	0.02	(-1.98)	(-2.01)	0.34	0.05	-0.90	0.13	(1.37)	(0.36)	0.32	0.39	-0.76
G	0.17	(0.51)	(1.43)	0.32	0.52	-0.73	0.09	(-0.42)	(-0.39)	0.29	0.30	-0.80	0.08	(0.64)	(-0.3)	0.32	0.25	-0.78
T	0.18	(1.01)	(1.93)	0.30	0.60	-0.55	0.11	(0.76)	(0.3)	0.25	0.44	-0.78	0.07	(0.18)	(-0.43)	0.30	0.25	-0.80
V	0.17	(0.35)	(1.49)	0.32	0.53	-0.71	0.06	(-1.42)	(-1.34)	0.27	0.22	-0.84	0.11	(1.11)	(0.08)	0.31	0.35	-0.73

A separate examination of the performance of the jump and crash portfolios reveals that the superior performance of LGBM-$SDGTV$ can be attributed to both jump and crash components while mostly from the crash side.

All the models that incorporate a single option characteristic increase the Sharpe ratio compared to the benchmark model, which suggests that each option characteristic provides valuable information about the underlying stock's future return. However, they do not quite measure up to LGBM-$SDGTV$ that combines all the option characteristics. This result suggests that each option characteristic carries different information from each other. Among these models, LGBM-S emerges as the top performer with a Sharpe ratio of 1.04, followed by LGBM-T (Sharpe ratio = 0.92), LGBM-D (Sharpe ratio = 0.88), LGBM-V (Sharpe ratio = 0.87), and LGBM-G (Sharpe ratio = 0.84).

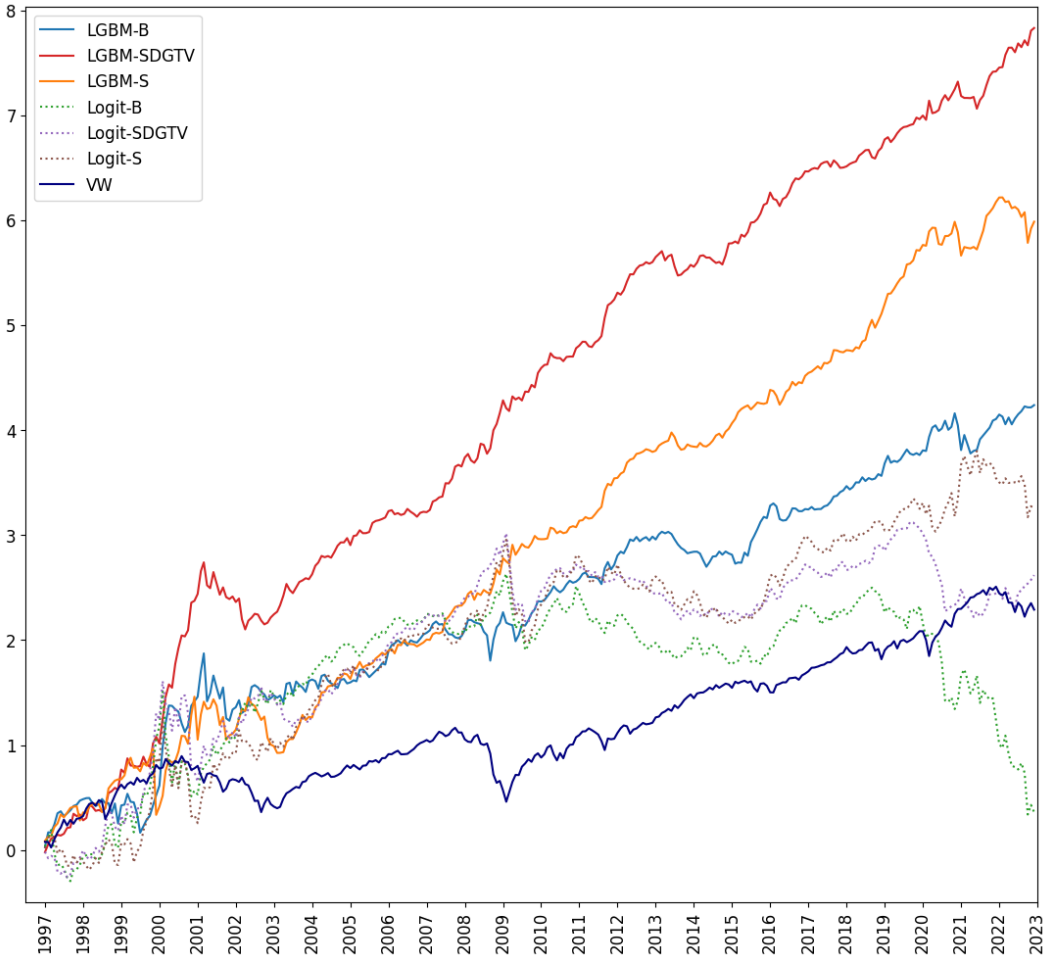
Unlike the LGBM-based models, the gain from the option characteristics is marginal in the Logit-based models. The Logit-S yields the highest Sharpe ratio of 0.54 and mean return of 0.18, while the Logit-$SDGTV$ yields the second highest Sharpe ratio of 0.49 and mean return of 0.14. Only Logit-S has significant improvement over Logit-B, whose Sharpe ratio and mean return are 0.21 and 0.07, but has insignificant improvement over VW, whose Sharpe ratio and mean return are 0.62 and 0.10. This underwhelming result has been anticipated because option characteristics are nonlinear functions of the underlying stock's price. It appears that a linear model cannot fully exploit the information contained in the option characteristics.

The long-short portfolios based on LGBM are formed using only a small portion of stocks that are predicted to yield extreme returns. As a result, the portfolios have the potential for substantial gains, but they are inherently less diversified and more volatile than the market portfolio. Nevertheless, their maximum drawdowns (MDD) are smaller than or comparable to that of the market portfolio, indicating that these portfolios do not suffer from extreme losses.

Figure 2.2 displays the log-scale cumulative returns of selected long-short portfolios. It shows that LGBM-$SDGTV$ long-short portfolio grows steadily throughout the sample period without any significant and prolonged losses. Especially in the later stages, as the amount of data increases, which is beneficial for the machine learning algorithm, the performance gradually stabilizes. While the performance of many machine learning-based prediction models declines in recent years, it is

impressive that the log-scale cumulative return of LGBM-*SDGTV* increases almost linearly until the last day of the sample period. It is also notable that LGBM-*SDGTV* is not affected by the global financial crisis in 2007. In contrast, LGBM-*B* and Logit-based models suffer significant losses in this period.

Figure 2.2. Cumulative returns (logarithmic scale) of the value-weighted long-short portfolios. This figure compares the cumulative returns (logarithmic scale) of the value-weighted long-short portfolios obtained from the LGBM- and Logit-based models. VW represents a benchmark portfolio invested in all available stocks each month.



2.3.2.2 Financial performance by year

Table 2.7 reports the Sharpe ratios of the LGBM models in each year over the out-of-sample period. LGBM-*SDGTV* performs consistently throughout the sample period yielding positive Sharpe ratios in 24 out of 26 years. The other LGBM-based models also perform well yielding

positive Sharpe ratios in most years.

For example, in 2022, the implied volatility-based LGBM model (LGBM-S) had a negative Sharpe ratio (-0.38), while the Vega-based LGBM model (LGBM-V) had a positive Sharpe ratio (1.78). Taking advantage of the combination of option features, the Sharpe ratio of LGBM-$SDGTV$ reaches 1.98. In comparison, in 2021, the Sharpe ratios of LGBM-S and LGBM-V are 1.10 and -0.29, respectively, and the Sharpe ratio of LGBM-$SDGTV$ is 0.51. This suggests that each option feature provides different information for the prediction according to the market, and the combination of option features captures more comprehensive information in the option features.

Table 2.7. Performance of value-weighted portfolios for LGBM-based models by year. The table displays annual Sharpe ratios for LGBM models applied to jump-crash portfolios between 1997 and 2022 in Panel A, and monthly return skewness and kurtosis in Panel B. "Skew", and "Kurt" labels in the rows refer to the skewness and kurtosis of each model's monthly return distribution, respectively.

Year	VW	B	$SDGTV$	S	D	G	T	V
Panel A. Sharpe ratios by year								
1997	1.78	2.74	2.23	1.45	0.76	2.06	2.50	1.43
1998	1.33	-0.78	1.17	1.83	0.46	1.74	0.26	-0.61
1999	1.59	0.94	1.55	-0.27	1.83	0.37	1.08	0.64
2000	-0.12	1.89	3.33	3.08	2.18	2.06	2.19	2.02
2001	-0.41	0.21	0.29	-0.34	-0.30	0.37	-0.03	-0.44
2002	-1.08	0.54	-0.66	-0.16	-0.39	-0.20	1.25	1.05
2003	2.49	0.68	2.23	1.53	0.74	0.10	0.64	0.57
2004	1.37	-0.03	3.64	2.85	1.27	0.96	0.71	1.00
2005	0.83	1.35	1.37	1.39	2.45	-0.16	0.43	0.36
2006	2.40	1.88	0.54	0.99	1.15	1.63	1.80	0.23
2007	0.94	-0.09	2.64	2.97	0.27	-0.07	1.82	1.21
2008	-1.90	0.51	1.98	1.45	2.07	-0.20	1.16	1.33
2009	1.25	0.91	1.56	1.30	1.53	0.97	2.43	1.73
2010	0.86	1.83	1.62	0.91	2.46	0.20	0.55	0.98
2011	-0.00	0.85	2.31	2.85	2.54	1.08	1.86	1.38
2012	1.38	1.92	3.09	2.63	1.99	3.08	3.63	3.09
2013	3.19	-1.37	-0.04	0.44	-1.21	-0.25	0.21	0.80
2014	1.11	-0.01	1.26	1.62	0.53	0.87	0.74	1.57
2015	-0.03	1.69	2.88	2.57	2.34	1.50	1.52	0.56
2016	1.06	0.52	1.78	1.33	1.40	2.51	0.49	0.88
2017	5.44	2.56	0.37	1.84	0.85	2.77	1.65	0.83
2018	-0.32	1.34	1.62	2.14	2.08	2.59	2.17	0.95
2019	2.01	1.20	2.37	4.19	2.13	2.32	1.25	1.17
2020	0.88	1.16	1.45	0.75	2.19	1.08	0.80	1.35
2021	2.14	0.34	0.51	1.10	-0.01	0.82	0.37	-0.29
2022	-0.82	0.93	1.98	-0.38	-0.09	0.52	-1.39	1.78
Panel B. Distributions of the monthly returns								
Skew	-0.53	0.28	1.74	1.08	0.31	0.83	0.75	0.72
Kurt	0.93	5.74	9.40	8.94	8.51	6.24	4.27	2.58
Support		jump 25,498		crash 29,464		normal 433,629		

2.3.2.3 Factor regression

To investigate if systematic risk factors account for returns of long-short portfolios, we use three-factor and five-factor models on value-weighted portfolios from LGBM-based models, focusing on the top-performing LGBM-SDGTV. Table 2.8 lists the models used: Fama and French's three-factor model (FF3) Fama and French (1996), FF3 with momentum (FF3m) (Carhart, 1997), and their five-factor models (FF5 and FF5m) Fama and French (2015) [6]. These factors are obtained from the Fama-French Portfolios & Factors dataset via Wharton Research Data Services.

The FF5m model, exhibiting the highest R^2 values, indicates significant alphas for jump-crash, jump, and crash portfolios. The jump-crash portfolio shows a monthly alpha of 2.7% (t-statistic of 5.880), the jump portfolio 1.2% (t-statistic of 3.296), and the crash portfolio -1.7% (t-statistic of -5.372). These findings imply that differences in returns among these portfolios cannot be explained solely by traditional risk factors, highlighting the potential for abnormal returns through our investment strategy. The effectiveness of different factor regression models further supports these conclusions.

2.3.3 Features sensitivities

2.3.3.1 Feature importance of LGBM

The LGBM evaluates the importance of input features through two methods: Frequency (Number of Splits) and Gain (Split Importance). The Frequency counts the total number of times a feature is used to split the data across all trees in the model. Features used more often in the tree construction process are considered more important. The Gain measures the total improvement in the loss function that results from each split based on a particular feature. A higher gain indicates that the feature is more important for making accurate predictions.

We assess feature importance using the Gain method as it quantifies the actual contribution of each feature. To determine the overall importance of a feature, we compute its importance in each month and use the time-series average.

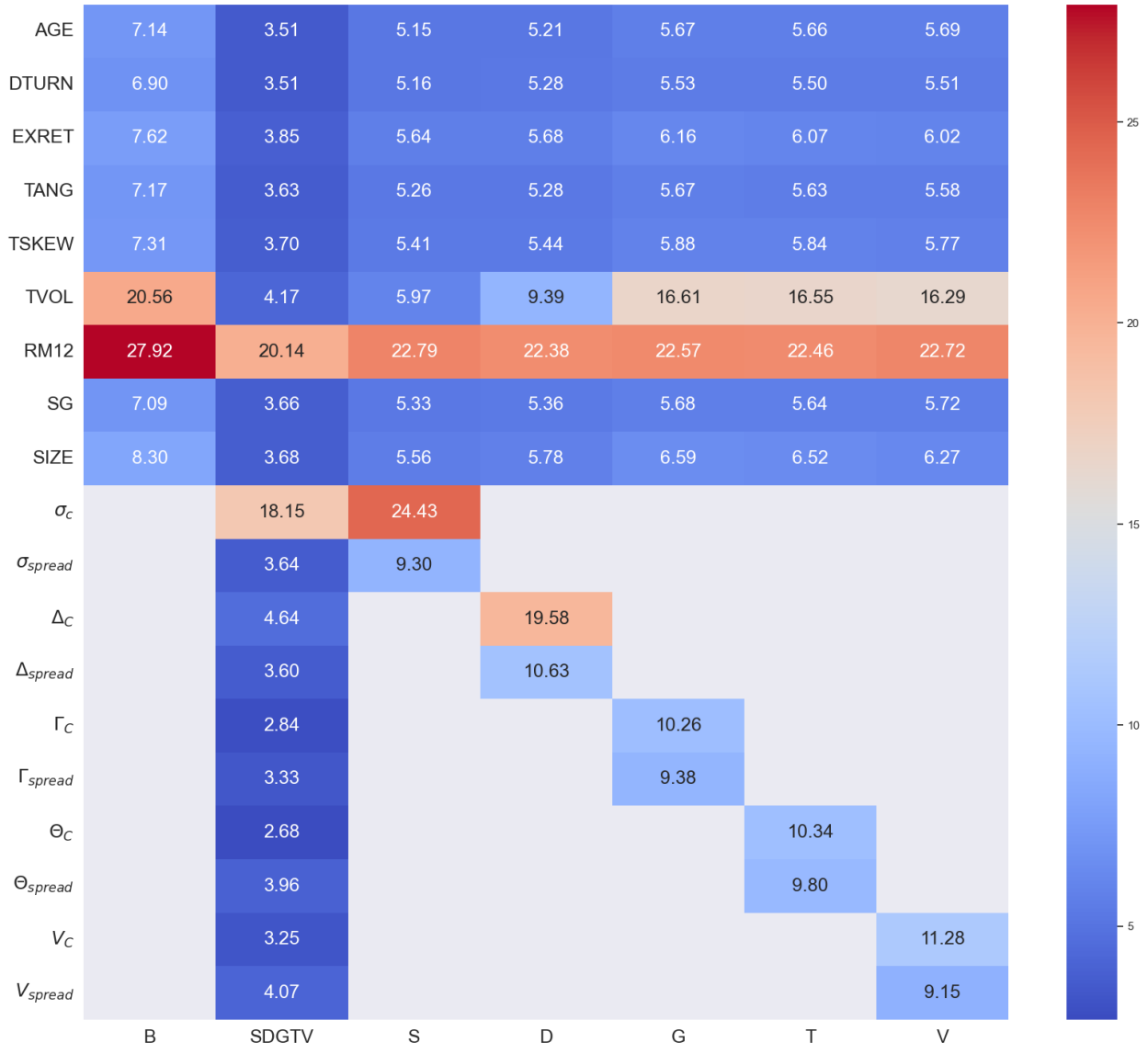
[6] Mktrf is the market factor, SMB is the small minus big, HML is the high minus low, RMW is the robust minus weak, and CMA is the conservative minus aggressive.

Table 2.8. Factor regressions. This table reports the factor regression results of the LGBM-based value-weighted jump-crash portfolio. FF3, FF3m, FF5, and FF5m respectively denote the Fama-French three factors, FF3 plus momentum, Fama-French five factors, and FF5 plus momentum. The sample period is from January 1997 to December 2021.

	Intercept	Mktrf	SMB	HML	RMW	CMA	Momentum	R^2
Panel A. LGBM-*SDGTV* (jump-crash)								
FF3	0.027	-0.175	0.173	0.111				0.017
	(6.211)	(-1.745)	(1.278)	(0.785)				
FF3m	0.027	-0.134	0.158	0.150			0.101	0.023
	(6.002)	(-1.252)	(1.161)	(1.030)			(1.138)	
FF5	0.028	-0.194	0.168	0.120	-0.039	-0.077		0.018
	(6.016)	(-1.619)	(1.058)	(0.608)	(-0.186)	(-0.271)		
FF5m	0.027	-0.161	0.144	0.185	-0.067	-0.102	0.106	0.023
	(5.880)	(-1.305)	(0.898)	(0.899)	(-0.318)	(-0.361)	(1.182)	
Panel B. LGBM-*SDGTV* (jump)								
FF3	0.007	1.463	0.853	-0.013				0.650
	(2.046)	(17.497)	(7.574)	(-0.106)				
FF3m	0.009	1.393	0.879	-0.079			-0.172	0.658
	(2.365)	(15.793)	(7.835)	(-0.652)			(-2.345)	
FF5	0.011	1.264	0.622	0.163	-0.729	-0.251		0.675
	(3.113)	(13.103)	(4.869)	(1.025)	(-4.303)	(-1.103)		
FF5m	0.012	1.221	0.653	0.081	-0.693	-0.218	-0.136	0.680
	(3.296)	(12.384)	(5.097)	(0.496)	(-4.086)	(-0.961)	(-1.888)	
Panel C. LGBM-*SDGTV* (crash)								
FF3	-0.021	1.640	0.682	-0.128				0.738
	(-6.776)	(22.607)	(6.981)	(-1.247)				
FF3m	-0.020	1.527	0.724	-0.234			-0.277	0.758
	(-6.356)	(20.521)	(7.646)	(-2.303)			(-4.476)	
FF5	-0.018	1.460	0.456	0.039	-0.687	-0.177		0.759
	(-5.613)	(17.557)	(4.146)	(0.283)	(-4.705)	(-0.901)		
FF5m	-0.017	1.382	0.513	-0.110	-0.622	-0.118	-0.246	0.774
	(-5.372)	(16.683)	(4.763)	(-0.793)	(-4.363)	(-0.617)	(-4.071)	

Figure 2.3 depicts the feature importance. In each model, RM12 (past market return) consistently emerges as the top feature with larger than 20% importance except model S where σ_c is the top importance feature with importance 24.43% while RM12 ranks second with importance 20.14%. The second important feature varies across models: TVOL (total volatility) leads in models B, G, T, and V, while σ_c takes precedence in model SDGTV, Δ_c in model D, and RM12 in model S.

Figure 2.3. The feature importance of LGBM-based models.



Apparently, the inclusion of σ and Δ significantly reduces the importance of TVOL, which decreases from 20.56% to 5.97% in Model S when σ_c and σ_{spread} are considered, and from 20.56%

to 9.39% in Model D when Δ_c and Δ_{spread} are considered. In Models S and D, the spreads of σ and Δ rank third after the values of σ and Δ for call options, implying that there is useful information in both call and put options. The above indicates that σ and Δ reveal more information of the underlying stock than TVOL. To illustrate this, Table 2.9 presents the relationships between σ_c, σ_{spread}, Δ_c, Δ_{spread} and TVOL.

Panel A of Table 2.9 reports the results of sorting based on TVOL, where σ_c and Δ_c are positively correlated with TVOL, while σ_{spread} and Δ_{spread} are negatively correlated. Meanwhile, sorted by TVOL, next-month returns (re) show a U-shaped pattern, with stocks located at both the low and high ends of the distribution receiving relatively higher returns. This suggests that TVOL does not adequately address bimodality.

Focusing on the sorting results based on the σ_c in Panel B of Table 2.9 reveals a consistent positive relationship between the σ_c and TVOL, with the monthly return relationship appearing more linear. Specifically, stocks with higher σ_c typically show higher monthly returns, while those with lower σ_c show lower returns. This suggests that the σ_c can not only provide a risk measurement that is both similar to and potentially more informative than TVOL but also partially solves the bimodality issue. The reason is that higher implied volatility in options indicates market expectations of larger stock price movements, leading to higher returns as compensation for increased risk. In contrast, stocks associated with options of lower implied volatility are expected to see smaller price movements, and thus lower returns, reflecting the lower risk. Unlike TVOL, which relies on historical volatility, the implied volatility of options directly reflects investor expectations for the stock, capturing speculative and insurance activities. This makes implied volatility a more precise measure of risk, as it encompasses expectations not readily apparent from historical volatility data.

Panel C of Table 2.9 reports the sorting results based on the σ_{spread} [7]. σ_{spread} reflects different risk perceptions among investors. Call and put options facilitate both speculations on stock price movements and protection against unfavourable price changes. When investors anticipate notable price fluctuations, they may utilize calls or puts to hedge or capitalize on these expectations, influencing the spread in implied volatilities. Furthermore, a greater absolute difference in implied volatility between call and put options, associated with higher TVOL and returns, suggests that

[7] call's σ - put's σ

Table 2.9. Feature relations by decile. The table shows relationships based on decile sorting of TVOL, σ_c, σ_{spread}, Δ_c, Δ_{spread}, and re. TVOL represents the stock total volatility. σ_c and σ_{spread} denote the implied volatility of call options and the implied volatility spread between call and put options, respectively. Δ_c, Δ_{spread} denote the Delta of call options and the Delta spread between call and put options, respectively. re refers to the monthly stock return. Groups 1 to 10 are organized by ascending values of the variable specified at the top of each panel.

	1	2	3	4	5	6	7	8	9	10
Panel A. TVOL										
TVOL	**0.011**	**0.015**	**0.018**	**0.020**	**0.022**	**0.025**	**0.028**	**0.032**	**0.040**	**0.061**
σ_c	0.229	0.292	0.353	0.371	0.403	0.455	0.506	0.574	0.665	0.898
σ_{spread}	0.006	0.000	-0.004	-0.005	-0.005	-0.006	-0.005	-0.003	-0.008	-0.031
Δ_c	0.525	0.529	0.532	0.534	0.537	0.540	0.543	0.548	0.554	0.570
Δ_{spread}	1.005	1.003	1.002	1.002	1.002	1.001	1.001	1.001	1.000	0.998
re	0.008	0.005	0.007	0.005	0.005	0.000	0.008	0.008	0.010	0.022
Panel B. σ_c										
TVOL	0.014	0.017	0.019	0.021	0.023	0.026	0.029	0.033	0.039	0.050
σ_c	**0.191**	**0.249**	**0.292**	**0.333**	**0.377**	**0.428**	**0.491**	**0.578**	**0.714**	**1.094**
σ_{spread}	-0.015	-0.011	-0.009	-0.009	-0.007	-0.007	-0.007	-0.006	-0.005	0.016
Δ_c	0.523	0.526	0.529	0.531	0.534	0.538	0.542	0.548	0.557	0.584
Δ_{spread}	1.004	1.003	1.002	1.002	1.001	1.001	1.001	1.000	1.000	1.001
re	0.007	0.007	0.007	0.006	0.008	0.008	0.008	0.009	0.008	0.010
Panel C. σ_{spread}										
TVOL	0.037	0.029	0.026	0.024	0.024	0.023	0.024	0.025	0.027	0.032
σ_c	0.657	0.469	0.416	0.390	0.376	0.377	0.394	0.433	0.501	0.734
σ_{spread}	**-0.190**	**-0.056**	**-0.034**	**-0.022**	**-0.012**	**-0.003**	**0.007**	**0.021**	**0.049**	**0.181**
Δ_c	0.554	0.540	0.537	0.535	0.535	0.535	0.536	0.538	0.543	0.559
Δ_{spread}	0.987	0.998	1.000	1.001	1.002	1.002	1.003	1.004	1.005	1.014
re	-0.001	0.009	0.010	0.009	0.008	0.007	0.008	0.008	0.011	0.011
Panel D. Δ_c										
TVOL	0.016	0.018	0.020	0.022	0.025	0.027	0.028	0.029	0.037	0.050
σ_c	0.228	0.282	0.321	0.351	0.403	0.456	0.465	0.497	0.663	1.080
σ_{spread}	-0.014	-0.010	-0.009	-0.008	-0.007	-0.008	-0.008	-0.006	-0.004	0.016
Δ_c	**0.512**	**0.523**	**0.528**	**0.532**	**0.536**	**0.541**	**0.546**	**0.551**	**0.559**	**0.585**
Δ_{spread}	1.000	1.000	1.001	1.001	1.001	1.001	1.003	1.004	1.002	1.002
re	0.009	0.010	0.009	0.007	0.005	0.008	0.007	0.008	0.007	0.009
Panel E. Δ_{spread}										
TVOL	0.037	0.030	0.027	0.025	0.025	0.025	0.025	0.025	0.025	0.027
σ_c	0.660	0.490	0.431	0.412	0.415	0.424	0.429	0.441	0.454	0.589
σ_{spread}	-0.173	-0.049	-0.028	-0.015	-0.006	0.001	0.010	0.022	0.040	0.140
Δ_c	0.549	0.539	0.535	0.534	0.535	0.537	0.539	0.543	0.547	0.555
Δ_{spread}	**0.984**	**0.996**	**0.998**	**0.999**	**1.000**	**1.002**	**1.003**	**1.005**	**1.007**	**1.022**
re	0.002	0.010	0.009	0.009	0.008	0.007	0.007	0.007	0.010	0.010

increased options trading activity occurs in anticipation of significant stock movements, leading to greater implied volatility discrepancies. Byun and Kim (2016) observe that options associated with lottery-like stocks exhibit deviations from put-call parity, indicating unique trading behaviours and risk assessments by traders of these options. Bali and Hovakimian (2009) also point out that the implied volatility spread is highly repeated to the stock returns.

Sorting results by the Δ_c in Panel D of Table 2.9 reveals a positive relationship between TVOL and the Δ_c, yet only options with the highest Δ_c correspond to higher monthly returns. This suggests Δ_c may reflect TVOL trends to some extent and with additional information. A higher Δ_c indicates greater sensitivity of the option price to changes in the underlying stock, a condition we attribute to assumed moneyness equal to 1. This sensitivity likely results from increased trading activity, signalling investor expectations.

For the sorting results based on Δ_{spread} [8] in Panel E of Table 2.9, values less than 1 indicate that the Δ of call options is lower than the absolute value of the delta of put options. Conversely, a delta spread greater than 1 suggests the call option's delta exceeds the absolute value of the put option's delta. These patterns may arise from trading preferences among investors for either call or put options, influencing the respective option's price sensitivity to changes in the stock price. Analysis of sorting results by Δ_{spread} reveals that higher Δ_{spread} are associated with potentially greater stock returns.

Overall, σ and Δ offer insights beyond TVOL, with the σ of call options mirroring TVOL trends but demonstrating clearer relationships with stock returns thus reducing the bimodality phenomenon. The other three features each contribute unique insights into TVOL's representation, offering different perspectives. These unique insights of option-related features could be explained by their nature to reflect more current market sentiments than TVOL, which captures a longer time range past stock volatility. These option-related features are grounded in recent one-month option trading activities and are influenced by the options' maturity, requiring investors to make more considered decisions. Consequently, option features present more immediate information compared to TVOL. Similarly, Bali et al. (2023) also show that the σ and Δ are two of the most important option features combined with the variance risk premium when predicting the return of delta-

[8] call's Δ - put's Δ

neutral option portfolios. Our results provide more details regarding why σ and Δ could be more helpful when predicting stock-related returns.

Regarding the Γ, Θ, and V of call options, they rank third in each of the sub-models (G, T, and V), followed by the spreads of these three features. However, when used in conjunction with σ and Δ, these three option features are less important but still provide some unique importance. For example, Θ measures the rate of time decay in option value, highlighting the impact of option maturity. This factor reminds investors to account not just for potential changes in option value or profits from excising options but also for the time-sensitive nature of options, where value decreases over time. However, in scenarios where investors anticipate significant stock movements in the future, the concern over time decay might be secondary to the potential profits. Bali et al. (2023) also identify Θ as an important feature for predictions of the returns of delta-neutral option portfolios.

2.3.3.2 Logit regression results

Table 2.10 shows the in-sample estimation results of the Logit models. To account for the correlation between observations across both firms and time, we cluster standard errors by firm and month following Petersen (2008) and Thompson (2011) and set the normal predictions as the reference class.

In B containing only control variables, all control variables are highly significant at the 5% level except for the SG (sales growth) for the jump prediction. The estimated coefficients on the control variables indicate that stocks that are younger, have lower de-trended turnover, have lower past excess returns, are more tangible, have lower skewness, have increased volatility, have reduced past market returns, have increased sales growth, and are smaller are more prone to extreme future returns in both directions. However, for both jump and crash predictions, all coefficients have the same sign, implying that jump and crash stocks generally have the same firm characteristics, which may explain the presence of bimodal phenomena in the classification task mentioned in Han (2021). Our results are different compared to Jang and Kang (2019). There are three reasons for this. First, the time horizon of our dataset is from 1996 to 2022, while that of Jang and Kang (2019) is from 1952 to 2014. Second, we keep only option stocks that meet the filtering rules. This may reduce

Table 2.10. Parameter estimates of the Logit-based models of extreme stock returns. The table presents the estimated parameters of the Logit-based models for predicting extreme stock returns. A jump is defined as a log return greater than 20% over the next month and a crash is defined as a log return lower than -20% over the next month. The z-stats are calculated based on the standard errors clustered by firm and month. The sample includes all available firm-month observations during the period from January 1996 to December 2022.

	jump							crash						
	B	*SDGTV*	*S*	*D*	*G*	*T*	*V*	*B*	*SDGTV*	*S*	*D*	*G*	*T*	*V*
Intercept	-2.79	-2.94	-2.89	-2.89	-2.8	-2.8	-2.83	-3.09	-3.24	-3.17	-3.16	-3.09	-3.09	-3.14
z-stat	(-372.67)	(-351.66)	(-365.24)	(-366.51)	(-372.57)	(-372.46)	(-361.88)	(-358.32)	(-331.39)	(-352.16)	(-353.24)	(-358.21)	(-358.27)	(-344.66)
AGE	-0.23	-0.13	-0.17	-0.18	-0.23	-0.23	-0.23	-0.16	-0.06	-0.1	-0.12	-0.17	-0.16	-0.16
z-stat	(-26.24)	(-14.59)	(-18.71)	(-20.34)	(-27.03)	(-26.1)	(-26.87)	(-16.76)	(-6.22)	(-10.24)	(-11.75)	(-17.52)	(-16.75)	(-17.09)
DTURN	-0.04	-0.02	-0.02	-0.02	-0.04	-0.04	-0.03	-0.03	-0.01	-0.01	-0.01	-0.02	-0.03	-0.02
z-stat	(-8.65)	(-4.91)	(-4.15)	(-4.68)	(-7.56)	(-8.62)	(-6.82)	(-5.74)	(-2.55)	(-2.03)	(-2.55)	(-4.76)	(-5.68)	(-3.96)
EXRET	-0.02	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	-0.04	-0.03	-0.03	-0.02	-0.03	-0.04	-0.03
z-stat	(-4.6)	(-2.04)	(-1.78)	(-1.37)	(-4.04)	(-4.41)	(-3.38)	(-6.47)	(-4.93)	(-4.53)	(-4.01)	(-5.78)	(-6.47)	(-5.47)
TANG	0.06	0.01	0.02	0.03	0.06	0.06	0.06	0.08	0.03	0.04	0.05	0.08	0.08	0.08
z-stat	(9.73)	(1.34)	(2.5)	(4.24)	(9.6)	(9.85)	(9.38)	(12.42)	(4.88)	(6.14)	(7.62)	(12.18)	(12.41)	(11.93)
TSKEW	0.03	-0.01	-0.01	-0.0	0.03	0.03	0.03	0.03	0.0	0.0	0.01	0.03	0.03	0.04
z-stat	(5.08)	(-0.86)	(-1.03)	(-0.61)	(5.08)	(5.19)	(6.2)	(5.46)	(0.6)	(0.47)	(0.81)	(5.54)	(5.46)	(6.49)
TVOL	0.42	0.06	0.07	0.12	0.41	0.42	0.39	0.41	0.08	0.1	0.14	0.41	0.41	0.39
z-stat	(63.23)	(7.58)	(9.72)	(16.65)	(62.52)	(63.05)	(59.51)	(57.08)	(9.72)	(11.97)	(18.13)	(56.52)	(56.9)	(53.29)
RM12	-6.35	-6.62	-6.61	-6.54	-6.36	-6.36	-6.37	-4.29	-4.52	-4.53	-4.49	-4.3	-4.29	-4.3
z-stat	(-48.18)	(-49.03)	(-49.09)	(-48.65)	(-48.2)	(-48.24)	(-48.24)	(-27.58)	(-28.63)	(-28.81)	(-28.57)	(-27.61)	(-27.54)	(-27.61)
SG	0.01	-0.01	-0.0	-0.0	0.01	0.01	0.01	0.01	0.0	0.01	0.01	0.02	0.01	0.02
z-stat	(1.54)	(-1.03)	(-0.51)	(-0.4)	(1.85)	(1.55)	(1.99)	(3.02)	(0.59)	(1.04)	(1.16)	(3.3)	(3.04)	(3.38)
SIZE	-0.17	-0.06	-0.05	-0.04	-0.14	-0.17	-0.07	-0.25	-0.15	-0.14	-0.13	-0.22	-0.25	-0.14
z-stat	(-20.59)	(-6.38)	(-5.59)	(-4.26)	(-16.42)	(-18.91)	(-8.05)	(-25.92)	(-13.73)	(-13.75)	(-12.87)	(-22.07)	(-24.46)	(-13.98)
σ_c		0.49	0.68						0.43	0.63				
z-stat		(22.3)	(99.73)						(16.74)	(83.99)				
σ_{spread}		-0.11	-0.18						-0.1	-0.12				
z-stat		(-9.18)	(-37.1)						(-7.43)	(-22.27)				
Δ_c		0.09		0.65					0.08		0.61			
z-stat		(4.06)		(91.81)					(3.08)		(76.95)			
Δ_{spread}		-0.1		-0.27					-0.04		-0.19			
z-stat		(-7.22)		(-39.91)					(-2.58)		(-24.37)			
Γ_c		-0.01			0.08				-0.0			0.08		
z-stat		(-2.0)			(14.31)				(-0.2)			(12.57)		
Γ_{spread}		-0.01			0.01				-0.02			-0.03		
z-stat		(-1.69)			(2.69)				(-1.82)			(-4.27)		
Θ_c		-0.5				0.01			-0.63				0.0	
z-stat		(-27.24)				(1.09)			(-30.67)				(0.52)	
Θ_{spread}		0.07				0.05			0.01				-0.02	
z-stat		(7.87)				(7.85)			(1.21)				(-2.12)	
V_c		-0.67					-0.38		-0.87					-0.42
z-stat		(-27.19)					(-29.56)		(-30.23)					(-27.36)
V_{spread}		0.02					0.02		0.0					-0.0
z-stat		(2.09)					(2.61)		(0.11)					(-0.01)
R^2	0.068	0.112	0.107	0.103	0.069	0.068	0.074							

the number of observations by more than half compared to Jang and Kang (2019). Third, our predictions are based on one month into the future, while Jang and Kang (2019) predicts returns for the next 12 months. We use the same definition as Jang and Kang (2019), which could give the same results.

Regarding the coefficients of option features, σ_c and Δ_c are positively significant for predicting both jumps and crashes, whether applied individually or combined. Conversely, σ_{spread} and Δ_{spread} are negatively significant for predicting both jumps and crashes, whether applied individually or combined. Γ_c has positively significant coefficients when used alone for jump and crash prediction, but switches to negatively significant for jump predictions and becomes insignificant for crashes when combined. Γ_{spread} coefficients are insignificant for both jumps and crashes. Θ_c transitions from insignificance when solo to negative significance in combination for both types of predictions. Θ_{spread} maintains positive significance for jump predictions alone or combined, and only negatively significant for crash predictions when used solo. $Vega_c$ is negatively significant for predicting both jumps and crashes, whether applied individually or combined. $Vega_{spread}$ only positively significant for jump predictions when used solo or combined.

In the comprehensive model $SDGTV$, which includes both control variables and option characteristics, most control variables continue to exhibit significant and directional influence. However, the significance of TSKEW changes, with its coefficients being positive and significant for jump and crash predictions in model B but becoming insignificant in $SDGTV$, S, and D. Similarly, the coefficients of TVOL for predicting jumps and crashes decrease substantially in $SDGTV$, S, and D, while they remain stable in G, T, and V. These changes underscore that option features, particularly σ and Δ, provide valuable information about the skewness and volatility of the underlying stock, aligning with findings from feature importance analyses in LGBM.

Furthermore, the value of R^2 improves from 0.068 in the control-only B model to 0.112 in the $SDGTV$, and in S, the addition of σ alone improves R^2 to 0.107, while in D, the addition of Δ improves R^2 to 0.103.

Overall, the most stable relationships among the option features are seen with σ and Δ, both showing consistent and significant effects whether applied individually or in combination. Other option features display less stable relationships, likely due to their different interactions with stock

returns, where simple linear relationships may not adequately capture their dynamics.

2.3.4 Robustness check

To ensure the robustness of our findings, we conduct a host of robustness checks. These include evaluating the performance of equally weighted portfolios, analyzing the impacts of transaction costs, examining dynamic thresholds for jumps and crashes, testing outcomes across portfolios with varying numbers of stocks, exploring the efficacy of different classification methods, and using alternative data filtering rules. Given the superior performance of the LGBM-$SDGTV$ model, we conduct the robustness checks using only LGBM-$SDGTV$.

2.3.4.1 Equal-weight portfolios

Table 2.11 reports the financial performance of equally weighted portfolios and Figure 2.4 displays their cumulative returns. Prior studies, *e.g.*, Gu et al. (2020); Han (2021); Han, He, and Toh (2023), equally weighted portfolios outperform value-weighted portfolios in terms of the Sharpe ratio and mean return. Our results are partly consistent with them. For instance, the Sharpe ratio of LGBM-$SDGTV$ jump-crash portfolio improves from 1.42 to 1.58, while the mean return decreases from 0.33 to 0.31. The difference in our results compared to prior studies could be attributed to the stock selection criteria. Previous research typically utilizes all available stocks, while our analysis focuses exclusively on stocks with available options, which are generally larger firms. These larger firms tend to exhibit more synchronous performance patterns. Although the value-weighted portfolio does make some adjustments to the final performance, these adjustments are relatively minor.

2.3.4.2 Transaction costs

Panel A of Table 2.12 reports the financial performance of the LGBM-$SDGTV$ portfolios under the assumption of transaction costs of 30 basis points. The annualized return of the jump-crash portfolio is reduced from 33% to 26% but remains statistically significant. The Sharpe ratio is also reduced but is still significant at 1.11.

Table 2.11. Performance of Equal-Weighted Portfolios. The table presents the annually measured financial performance of the models from January 1997 to December 2022. The performance metrics include average return (MEAN), standard deviation (Std), Sharpe ratio (SR), and maximum drawdown (MDD). The Newey-West tests (column $t1$ and $t2$) serve as significant tests for return, with B as the benchmark ($t1$) and EW as the benchmark ($t2$). The dataset contains 18,908 jumps, 25,064 crashes, and 352,665 normal observations within the out-of-sample period.

Panel A. LGBM models

	Jump						Crash						Jump-Crash					
	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD
EW	0.11			0.21	0.52	-0.55	0.11			0.21	0.52	-0.55	0.11			0.21	0.52	-0.55
B	0.26		(4.35)	0.31	0.82	-0.51	0.03		(-2.67)	0.34	0.08	-0.86	0.23		(1.94)	0.19	1.22	-0.25
SDGTV	0.24	(-0.54)	(3.55)	0.34	0.71	-0.56	-0.07	(-3.8)	(-5.0)	0.35	-0.20	-0.98	0.31	(2.9)	(3.34)	0.20	1.58	-0.26
S	0.24	(-0.99)	(4.06)	0.32	0.73	-0.54	-0.04	(-3.85)	(-4.72)	0.36	-0.12	-0.95	0.28	(2.0)	(2.99)	0.18	1.53	-0.23
D	0.23	(-0.99)	(4.1)	0.32	0.73	-0.54	-0.03	(-2.6)	(-3.91)	0.37	-0.09	-0.94	0.26	(1.3)	(2.57)	0.20	1.31	-0.40
G	0.23	(-1.08)	(3.98)	0.32	0.71	-0.55	0.02	(-0.19)	(-3.09)	0.31	0.08	-0.80	0.21	(-0.78)	(1.75)	0.18	1.15	-0.32
T	0.25	(-0.48)	(3.94)	0.32	0.77	-0.52	-0.00	(-1.63)	(-4.55)	0.31	-0.00	-0.83	0.25	(0.78)	(2.46)	0.18	1.38	-0.20
V	0.22	(-1.36)	(4.03)	0.32	0.70	-0.55	-0.00	(-1.48)	(-3.85)	0.34	-0.00	-0.86	0.22	(-0.18)	(2.08)	0.17	1.31	-0.24

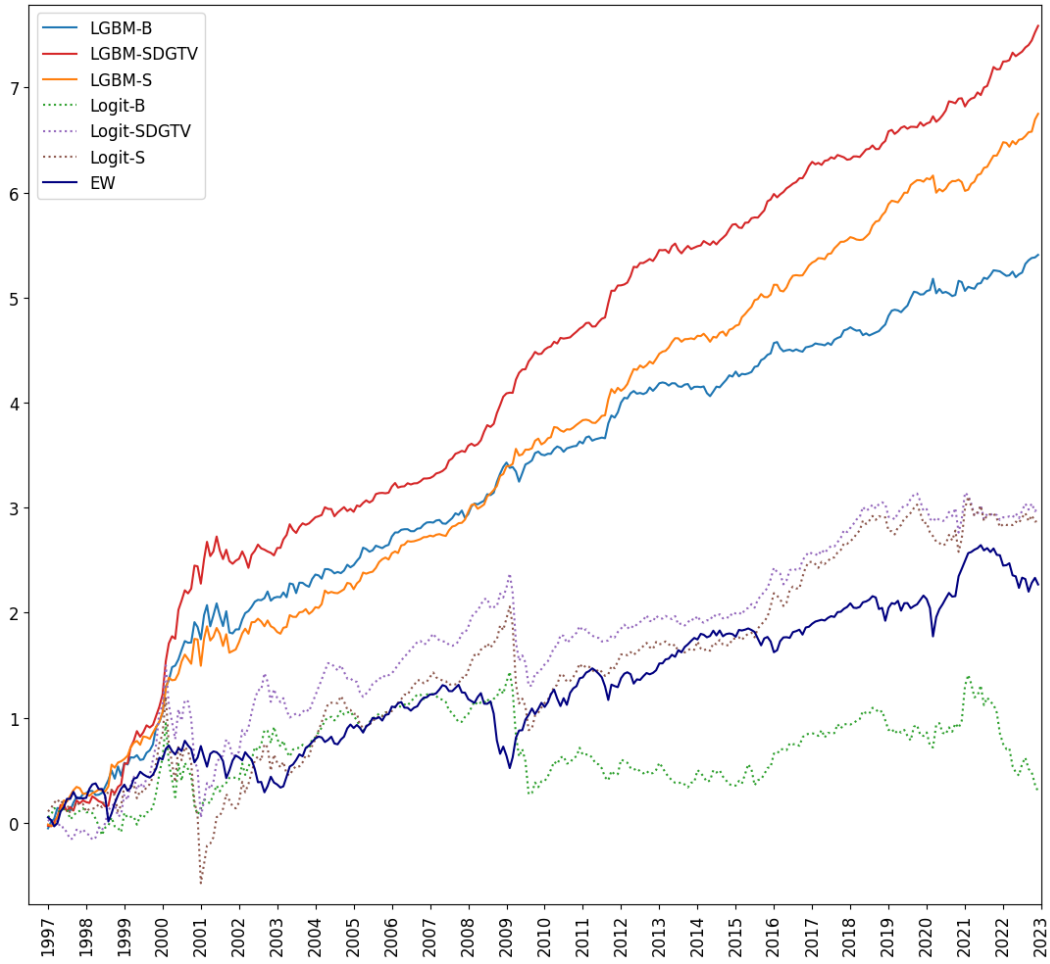
Panel B. Logit models

	Jump						Crash						Jump-Crash					
	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD
EW	0.11			0.21	0.52	-0.55	0.11			0.21	0.52	-0.55	0.11			0.21	0.52	-0.55
B	0.18		(1.73)	0.34	0.52	-0.70	0.13		(0.48)	0.30	0.42	-0.70	0.05		(-0.85)	0.29	0.18	-0.69
SDGTV	0.16	(-0.82)	(1.34)	0.32	0.50	-0.65	-0.00	(-4.46)	(-2.6)	0.36	-0.00	-0.94	0.16	(2.71)	(0.65)	0.30	0.54	-0.76
S	0.17	(-0.52)	(1.53)	0.33	0.51	-0.68	0.01	(-4.14)	(-2.31)	0.38	0.02	-0.94	0.16	(2.9)	(0.65)	0.31	0.52	-0.83
D	0.16	(-0.81)	(1.4)	0.33	0.50	-0.67	0.02	(-3.43)	(-1.96)	0.38	0.06	-0.92	0.14	(2.37)	(0.42)	0.30	0.48	-0.78
G	0.19	(1.63)	(2.05)	0.34	0.56	-0.67	0.09	(-3.66)	(-0.63)	0.31	0.29	-0.72	0.10	(3.72)	(-0.11)	0.28	0.36	-0.70
T	0.21	(1.92)	(2.24)	0.35	0.59	-0.70	0.12	(-0.43)	(0.35)	0.29	0.42	-0.72	0.08	(1.97)	(-0.36)	0.30	0.29	-0.67
V	0.18	(0.38)	(1.84)	0.33	0.56	-0.70	0.11	(-1.59)	(-0.02)	0.31	0.35	-0.77	0.07	(1.14)	(-0.51)	0.30	0.25	-0.70

Table 2.12. Performance of Value-Weighted Portfolios. The table presents the annually measured financial performance for LGBM-based models from January 1997 to December 2022. The performance metrics include average return (MEAN), standard deviation (Std), Sharpe ratio (SR), and maximum drawdown (MDD). The Newey-West tests (column $t1$ and $t2$) serve as significant tests for return, with B as the benchmark ($t1$) and VW as the benchmark ($t2$).

	Jump						Crash						Jump-Crash					
	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD
Panel A. Transaction cost																		
VW	0.07			0.16	0.40	-0.53	0.07			0.16	0.40	-0.53	0.07			0.16	0.40	-0.53
B	0.16		(2.29)	0.30	0.53	-0.58	-0.04		(-2.51)	0.33	-0.12	-0.92	0.13		(0.94)	0.26	0.48	-0.50
$SDGTV$	0.18	(0.54)	(2.44)	0.34	0.53	-0.73	-0.15	(-3.25)	(-4.6)	0.34	-0.44	-1.00	0.26	(2.67)	(3.12)	0.23	1.11	-0.51
Panel B. Dynamic threshold																		
B	0.17		(1.71)	0.32	0.55	-0.73	0.05		(-1.21)	0.31	0.16	-0.88	0.12		(0.34)	0.22	0.56	-0.51
$SDGTV$	0.21	(1.07)	(2.02)	0.36	0.59	-0.71	-0.08	(-3.86)	(-4.33)	0.32	-0.25	-0.98	0.29	(3.3)	(3.39)	0.24	1.23	-0.26
Panel C. Invest in the top and bottom 50 stocks																		
B	0.19		(2.11)	0.31	0.61	-0.62	0.02		(-1.96)	0.34	0.05	-0.81	0.17		(1.18)	0.27	0.65	-0.48
$SDGTV$	0.19	(0.01)	(1.96)	0.33	0.57	-0.61	-0.12	(-4.02)	(-4.57)	0.35	-0.33	-0.99	0.30	(2.85)	(3.21)	0.25	1.24	-0.34
Panel D. Invest in the top and bottom 2.5% stocks																		
B	0.19		(1.73)	0.34	0.55	-0.72	-0.01		(-2.19)	0.38	-0.03	-0.92	0.20		(1.4)	0.32	0.62	-0.71
$SDGTV$	0.21	(0.4)	(1.9)	0.36	0.57	-0.73	-0.17	(-3.47)	(-4.61)	0.40	-0.42	-1.00	0.37	(2.51)	(3.71)	0.29	1.28	-0.32
Panel E. Invest in the top and bottom 10% stocks																		
B	0.17		(1.98)	0.27	0.63	-0.62	0.04		(-2.06)	0.29	0.13	-0.75	0.13		(0.61)	0.22	0.62	-0.49
$SDGTV$	0.20	(1.23)	(2.45)	0.31	0.64	-0.60	-0.05	(-3.51)	(-3.82)	0.31	-0.16	-0.94	0.25	(3.24)	(2.47)	0.23	1.11	-0.32
Panel F. Invest based on the probability of jump only																		
B	0.15		(1.1)	0.33	0.46	-0.76	0.10		(0.17)	0.13	0.79	-0.40	0.05		(-1.2)	0.28	0.17	-0.69
$SDGTV$	0.11	(-1.26)	(0.08)	0.37	0.29	-0.79	0.10	(-0.47)	(-0.09)	0.13	0.75	-0.35	0.01	(-1.1)	(-1.89)	0.33	0.02	-0.85
Panel G. Invest based on the probability of crash only																		
B	0.11		(0.24)	0.13	0.82	-0.33	0.07		(-0.68)	0.35	0.21	-0.80	0.04		(-0.86)	0.30	0.11	-0.86
$SDGTV$	0.11	(0.47)	(0.52)	0.13	0.84	-0.37	-0.11	(-3.93)	(-4.06)	0.37	-0.30	-0.99	0.22	(3.83)	(1.35)	0.31	0.71	-0.62
Panel H. Invest based on the probability of jump and crash, respectively																		
B	0.15		(1.1)	0.33	0.46	-0.76	0.07		(-0.68)	0.35	0.21	-0.80	0.08		(-0.44)	0.22	0.37	-0.55
$SDGTV$	0.11	(-1.26)	(0.08)	0.37	0.29	-0.79	-0.11	(-3.93)	(-4.06)	0.37	-0.30	-0.99	0.22	(2.41)	(2.0)	0.20	1.05	-0.30

Figure 2.4. Cumulative returns (logarithmic scale) of the equal-weighted long-short portfolios. This figure compares the cumulative returns (logarithmic scale) of the equal-weighted long-short portfolios obtained from the LGBM- and Logit-based models. EW represents a benchmark portfolio invested in all available stocks each month.



2.3.4.3 Dynamic thresholds for jump and crash

Previously, we define jump as a log return over 20% and crash as a log return below -20%. Here we examine alternative definitions: jump is defined as a return in the top 5% and crash as a return in the bottom 5% of all returns over the past 24 months. These dynamic thresholds aim to capture extreme returns based on their relative performance within the market. Panel B of Table 2.12 reports the financial performance of the portfolios constructed using the dynamic thresholds. LGBM-$SDGTV$ continues to outperform LGBM-B: the annualized return and the Sharpe ratio of the jump-crash portfolio are respectively 29% and 1.23.

2.3.4.4 Sensitivity to the investment percentile

In our primary analysis, we concentrate on the highest and lowest 5% of stocks according to their ProbDif, aligning with the observation that jumps and crashes make up about 5% of all cases. We assess portfolio outcomes across different percentiles to verify if returns are consistently advantageous. The results, as shown in Table 2.12 Panels C to E, are based on portfolios created using LGBM models with investment thresholds of the top and bottom 50 stocks, 2.5%, and 10% based on ProbDif ranking. Specifically, the LGBM-$SDGTV$ model demonstrates Sharpe ratios of 1.24, 1.28, and 1.11 for Value-weighted portfolios at the 50 stocks, 2.5%, and 10% thresholds, respectively. Across all investment thresholds, the LGBM-$SDGTV$ model outperforms the LGBM-B model and VW consistently.

2.3.4.5 Different classification criteria

In the primary analysis, we sort stocks on ProbDiff (probability of jump - probability of crash) to address bimodality, following Han (2021). ProbDiff has the advantage of utilizing information from both the jump and crash probabilities. In this part, we test three alternative sorting criteria: jump probability (Panel F in Table 2.12), crash probability (Panel G in Table 2.12), and jump probability for jump portfolio and crash probability for crash portfolio (Panel H in Table 2.12). All three criteria underperform ProbDiff: When only the jump probability is used, the Sharpe ratio is mere 0.02, when only the crash probability is used, it is 0.71, and when both probabilities are used, it is 1.05. In comparison, the Sharpe ratio from ProbDiff is 1.42. The superior performance when using both probabilities can be attributed to the fact that some stocks that are included in the jump portfolio have high crash probabilities and are also included in the crash portfolio, effectively removing the stocks that have both high jump and crash probabilities from the long-short portfolio.

2.3.4.6 Using only ATM options

In this section, we extract option characteristics from only one ATM call option and one ATM put option at the end of the month, following Zhan et al. (2022). As this approach requires at least one ATM call and one ATM put for each stock, the sample size is reduced to 282,371. Table 2.13 and 2.14 report the performance of the portfolios under this approach. The performance is slightly

worse than using all available options in the month. The inferior performance can be attributed to i) loss of information contained in ITM and OTM options and ii) smaller sample size.

Table 2.13. Accuracy of jump and crash predictions. The table presents the statistical performance of various models over the test period of January 1997 to December 2022. The results are presented regarding the AUCs of jump and crash stocks for LGBM (Panel A) and Logit (Panel B), respectively. The dataset contains 14,459 jumps, 10,669 crashes, and 257,243 normal observations within the out-of-sample period. PPV: Positive predictive value (precision) FDR: False discovery rate

Panel A. Before reclassification

		B	*SDGTV*	*S*	*D*	*G*	*T*	*V*
LGBM								
Jump	AUC	0.683	0.773	0.767	0.718	0.706	0.708	0.715
	PPV	0.092	0.136	0.131	0.111	0.111	0.107	0.112
	FDR	0.094	0.149	0.146	0.113	0.108	0.113	0.115
Crash	AUC	0.667	0.749	0.734	0.689	0.668	0.676	0.692
	PPV	0.106	0.168	0.158	0.126	0.117	0.120	0.126
	FDR	0.087	0.115	0.118	0.093	0.083	0.084	0.087
Logit								
Jump	AUC	0.669	0.706	0.710	0.671	0.667	0.668	0.668
	PPV	0.203	0.195	0.201	0.199	0.201	0.194	0.200
	FDR	0.226	0.257	0.263	0.230	0.223	0.227	0.226
Crash	AUC	0.645	0.700	0.705	0.647	0.643	0.642	0.647
	PPV	0.182	0.247	0.252	0.181	0.184	0.179	0.182
	FDR	0.264	0.268	0.275	0.263	0.263	0.261	0.262

Panel B. After reclassification

		B	*SDGTV*	*S*	*D*	*G*	*T*	*V*
LGBM								
Jump	AUC	0.484	0.500	0.494	0.507	0.522	0.512	0.515
	PPV	0.089	0.125	0.122	0.108	0.110	0.107	0.108
	FDR	0.094	0.133	0.134	0.107	0.100	0.107	0.110
Crash	AUC	0.520	0.555	0.544	0.535	0.526	0.531	0.533
	PPV	0.116	0.159	0.148	0.120	0.110	0.116	0.122
	FDR	0.078	0.106	0.109	0.085	0.078	0.080	0.082
Logit								
Jump	AUC	0.440	0.448	0.441	0.450	0.456	0.464	0.455
	PPV	0.122	0.141	0.138	0.127	0.135	0.129	0.124
	FDR	0.149	0.169	0.157	0.151	0.153	0.159	0.148
Crash	AUC	0.502	0.537	0.525	0.522	0.528	0.527	0.518
	PPV	0.158	0.207	0.222	0.147	0.162	0.150	0.159
	FDR	0.206	0.220	0.235	0.196	0.192	0.202	0.202

2.4 Conclusion

Our study conclusively demonstrates that employing advanced machine learning techniques, specifically the LightGBM model, significantly enhances the predictive accuracy of stock market movements by leveraging option characteristics like implied volatility and Greeks. This approach

Table 2.14. Performance of Value-Weighted Portfolios. The table presents the annually measured financial performance of the models from January 1997 to December 2022. The performance metrics include average return (MEAN), standard deviation (Std), Sharpe ratio (SR), and maximum drawdown (MDD). The Newey-West tests (column $t1$ and $t2$) serve as significant tests for return, with B as the benchmark ($t1$) and VW as the benchmark ($t2$). The dataset contains 14,459 jumps, 10,669 crashes, and 257,243 normal observations within the out-of-sample period.

Panel A. LGBM models

	jump						crash						jump-crash					
	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD
VW	0.10			0.16	0.63	-0.50	0.10			0.16	0.63	-0.50	0.10			0.16	0.63	-0.50
B	0.19		(2.02)	0.30	0.63	-0.66	-0.01		(-2.3)	0.32	-0.04	-0.94	0.20		(1.56)	0.24	0.85	-0.40
$SDGTV$	0.20	(0.25)	(2.35)	0.31	0.64	-0.61	-0.07	(-1.46)	(-3.82)	0.32	-0.22	-0.97	0.27	(1.2)	(2.8)	0.24	1.13	-0.24
S	0.16	(-0.94)	(1.37)	0.33	0.48	-0.79	-0.05	(-0.95)	(-3.18)	0.32	-0.15	-0.95	0.21	(0.19)	(1.91)	0.24	0.88	-0.59
D	0.22	(0.99)	(2.68)	0.31	0.71	-0.58	-0.02	(-0.14)	(-2.91)	0.30	-0.06	-0.93	0.23	(0.79)	(1.98)	0.24	0.97	-0.33
G	0.21	(0.8)	(2.83)	0.30	0.71	-0.62	0.04	(1.71)	(-1.73)	0.29	0.15	-0.69	0.17	(-0.76)	(1.39)	0.22	0.78	-0.34
T	0.25	(1.96)	(3.33)	0.29	0.85	-0.58	0.02	(0.6)	(-2.15)	0.30	0.05	-0.80	0.24	(0.67)	(2.41)	0.22	1.06	-0.38
V	0.15	(-1.26)	(1.29)	0.30	0.50	-0.68	-0.06	(-1.34)	(-3.92)	0.31	-0.20	-0.96	0.21	(0.21)	(2.19)	0.24	0.90	-0.44

Panel B. Logit models

	jump						crash						jump-crash					
	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD	Mean	t1	t2	Std	SR	MDD
VW	0.10			0.16	0.63	-0.50	0.10			0.16	0.63	-0.50	0.10			0.16	0.63	-0.50
B	0.16		(1.27)	0.32	0.51	-0.79	0.15		(1.26)	0.30	0.49	-0.72	0.01		(-1.26)	0.34	0.04	-0.92
$SDGTV$	0.15	(-0.38)	(1.09)	0.30	0.49	-0.66	0.02	(-3.41)	(-1.76)	0.31	0.08	-0.88	0.13	(1.99)	(0.33)	0.32	0.39	-0.64
S	0.13	(-1.35)	(0.47)	0.33	0.38	-0.80	0.06	(-2.91)	(-0.89)	0.34	0.17	-0.86	0.07	(1.37)	(-0.46)	0.35	0.19	-0.87
D	0.14	(-1.12)	(0.85)	0.32	0.44	-0.79	0.07	(-2.96)	(-0.95)	0.28	0.24	-0.79	0.08	(1.96)	(-0.37)	0.32	0.24	-0.81
G	0.14	(-1.15)	(0.65)	0.33	0.42	-0.76	0.12	(-1.15)	(0.44)	0.30	0.40	-0.71	0.02	(0.07)	(-1.18)	0.34	0.05	-0.91
T	0.19	(0.93)	(1.88)	0.31	0.62	-0.58	0.14	(-0.43)	(0.95)	0.28	0.50	-0.66	0.05	(0.98)	(-0.75)	0.32	0.16	-0.77
V	0.18	(0.38)	(1.48)	0.32	0.54	-0.72	0.10	(-1.36)	(0.03)	0.28	0.37	-0.74	0.07	(1.16)	(-0.43)	0.35	0.20	-0.82

substantially outperforms traditional logistic regression models, achieving strikingly higher Sharpe ratios and average annual returns in both value-weighted and equal-weighted portfolio constructions. The integration of multiple option characteristics into the LightGBM framework proved particularly effective, suggesting that a comprehensive view of option dynamics is crucial for accurate market predictions.

These findings highlight the critical role of daily option characteristics fluctuations in understanding and anticipating market trends. They also underscore the superiority of machine learning methods over traditional models in capturing complex patterns inherent in option data, thus offering more effective tools for forecasting stock returns. This research opens avenues for further refinement in stock market prediction strategies, especially through advanced feature engineering of option characteristics, cementing the value of machine learning in financial market analysis.

Appendices

.1 Control variables

- AGE: Firm age is the number of months since the firm's first appearance on the CRSP monthly stock file.
- $DTURN$: Detrended turnover is the six-month average of monthly share turnover subtracted by the prior 18-month average.
- $EXRET$: Past individual return in excess of the market return, measured for each stock at the end of month t as the log return on the stock from month t - 11 to month t, in excess of the same return on the CRSP value-weighted index.
- $TANG$: Tangible assets at the end of June in year t to May in year t + 1 are property, plant, and equipment (Compustat annual item PPEGT) divided by total assets (item AT) for the end of year t - 1.
- $TSKEW$: Total skewness at the end of month t is calculated using daily log returns from month t - 5 to month t. We exclude stocks with fewer than 50 daily log returns during the six-month period.
- $TVOL$: Total volatility at the end of month t is the standard deviation of daily log returns from month t - 5 to month t. We exclude stocks with fewer than 50 daily log returns during the six-month period.
- $RM12$: Past market return. Measured at the end of month t as the log return of the CRSP value-weighted index from month t - 11 to month t.
- SG: Sales growth at the end of June in year t to May in year t + 1 is the log difference between sales (Compustat annual item SALE) at the end of year t - 1 and sales at the end of year t - 2.
- $SIZE$: The natural logarithm of the firm's market capitalization.

.2 *Binomial tree option pricing model and hypothetical greeks*

The Binomial option pricing model is a widely used method for calculating the fair value of an option. It is based on the assumption that the underlying asset price can only move up or down by a fixed percentage at each time step. The model is widely used for pricing European-style options, which can only be exercised at expiration. However, it can also be adapted to price American-style options, which can be exercised at any time before expiration.

In the case of American-style options, the Binomial option pricing model requires a modification to account for the possibility of early exercise. At each time step, the option holder has the option to exercise the option and receive the intrinsic value of the option, which is the difference between the current underlying asset price and the strike price. The option holder may choose to exercise the option if the intrinsic value is greater than the current value of the option.

To calculate the fair value of an American option using the Binomial model, we follow the following steps:

1. Calculate the size of each time step (Δt) using the following formula:

$$\Delta t = \frac{T}{N}, \tag{10}$$

   where T is the maturity, N is the total number of time steps[9].

2. Calculate the up factor (u) and the down factor (d) using the following formulas:

$$u = \exp(\sigma \times \sqrt{\Delta t}) \tag{11}$$

$$d = \frac{1}{u}, \tag{12}$$

   where σ is the volatility of the underlying asset.

3. Calculate the probability of an up move (p) using the following formula:

$$p = \frac{\exp(r - q) \times \Delta t - d}{u - d}, \tag{13}$$

[9] For the purpose of saving computation time, we set the value of N to 100.

where r is the risk-free interest rate, q is the dividend yield.

4. Calculate the value of the option at expiration for each possible price of the underlying asset. For an underlying asset price S and strike price X at a given point in time, the price of a call option is equal to $max(S - X, 0)$, and the price of a put option is equal to $max(X - S, 0)$.

5. Working backwards from expiration, calculate the value of the option at each time step for each possible price of the underlying asset using the following formulas:

$$Price_u = \frac{p \times Price_{uu} + (1-p) \times Price_{ud}}{1 + rf}, \tag{14}$$

$$Price_d = \frac{p \times Price_{du} + (1-p) \times Price_{dd}}{1 + rf}, \tag{15}$$

where $Price_{uu}$, $Price_{ud}$, $Price_{du}$, and $Price_{dd}$ are the option values if the underlying asset price goes up twice, up then down, down then up, and down twice, respectively.

6. The fair value of the American option is the maximum of the intrinsic value and the calculated value at each time step.

With respect to the Greeks, the binomial tree model provides discrete approximations.

Delta: Delta measures the change in the option price for a unit change in the underlying asset price, which can be calculated according to:

$$Delta = \frac{Price_u - Price_d}{S \times (u - d)} \tag{16}$$

Gamma: Gamma measures the change in Delta for a unit change in the underlying asset price, which can be calculated according to:

$$Gamma = \frac{\Delta_u - \Delta_d}{S \times (u - d)} \tag{17}$$

Theta: Theta measures the change in the option price for a unit change in time, which can be calculated according to:

$$Theta = \frac{Price_{T+\Delta t} - Price_T}{\Delta t} \tag{18}$$

Vega: Vega measures the sensitivity of the option price to changes in volatility. To approximate Vega in a binomial model, we perturb the volatility (increase it slightly, say by 0.01) and then compute the difference in the option price as follows:

$$Vega = \frac{Price(\sigma + 0.01) - Price(\sigma)}{0.01} \tag{19}$$

Chapter 3

Comparative analysis of stock option pricing: Machine learning and Pandemic impact

Abstract

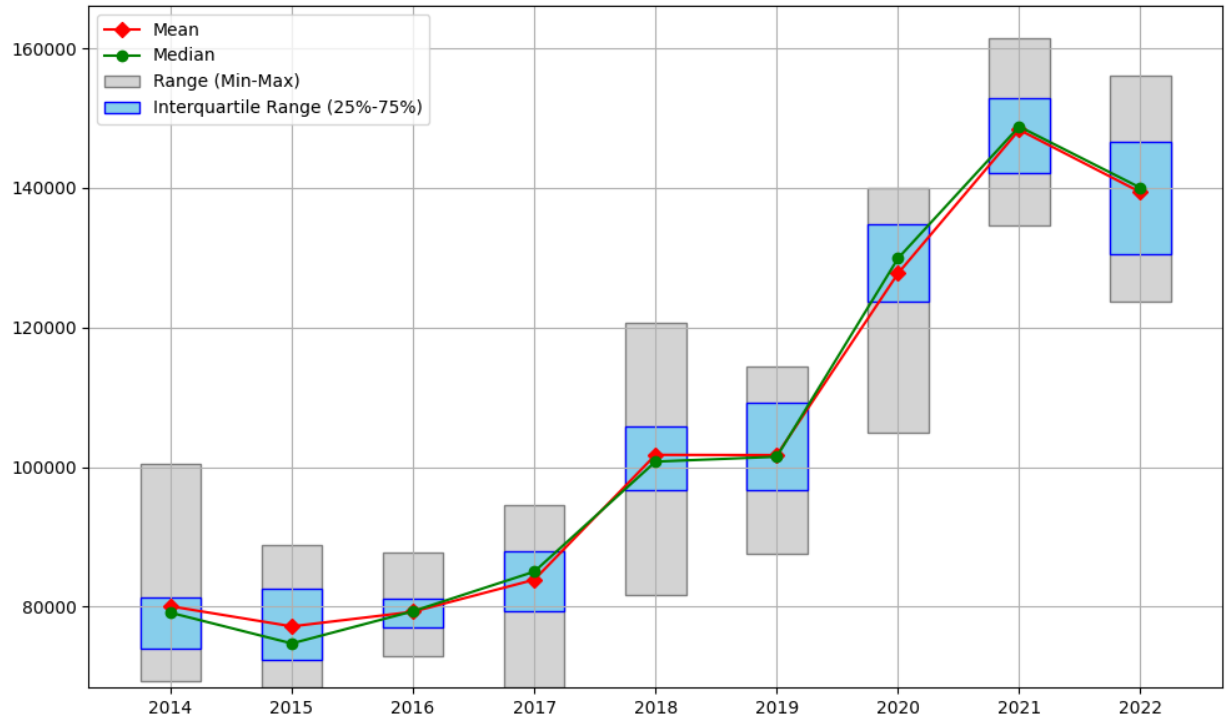
This study evaluates machine learning models' effectiveness in stock option pricing during the Pandemic's market challenges. It explores three architectures, integrating basic factors, historical Greeks, and firm characteristics to improve accuracy. The research contrasts traditional parametric methods with adaptable machine learning approaches. Analyzing data from over 10 million U.S. stock options for SP500 stocks from 2014 to 2022, it demonstrates these models' superior performance in volatile markets. Contributions include incorporating Greeks and firm characteristics into pricing models, identifying models that maintain robustness in turbulence, and revealing how stock return patterns impact accuracy. The findings highlight machine learning's crucial role in advancing option pricing under challenging conditions.

Keywords: Option pricing; Option Greeks; Firm characteristics; Machine learning.

3.1 Introduction

In the fluctuating financial market, robust option pricing models are essential, particularly in variable conditions such as those presented by the Pandemic, which has significantly increased options trading activity as shown in 3.1. This study evaluates the efficacy of machine learning methods in pricing stock options during the Pandemic, exploring three different architectures and their four respective submodels. The first submodel incorporates basic elements like option type and asset metrics. The second adds historical sensitivities of options (Greeks) for improved precision. The third integrates firm characteristics to broaden the market analysis, and the fourth combines all elements, providing a thorough approach to option pricing in a market impacted by the Pandemic.

Figure 3.1. Yearly statistics. Candlestick charts give yearly observational statistics for options with a daily trading volume of more than 100 contracts. Statistics are calculated based on monthly observations for each year. The mean and median monthly trading volumes have increased substantially since the beginning of the pandemic (2020) and begin to decline in 2022, but are still higher than in previous years.



3.1.1 Traditional parametric approaches

The Black-Scholes (BS) model, introduced by Black and Scholes (1973), is a foundational approach for pricing European options. Despite its simplicity and reliance on assumptions like normally distributed stock returns with constant volatility, it remains a critical model in option pricing. However, given that stock returns often exhibit fat tails and variable volatility, enhancements such as the incorporation of skewness and kurtosis by Corrado and Su (1996) and stochastic volatility models by Heston (1993) and Hagan et al. (2002) have been developed to more accurately reflect underlying asset volatility. For American options, which require considerations for early exercise and dividends, methods like tree-based algorithms (Cox et al., 1979; Boyle, 1986), finite difference method (Crank and Nicolson, 1947), and Monte Carlo simulations (Broadie and Glasserman, 1997; Longstaff and Schwartz, 2001; Andersen and Broadie, 2004) have been proposed, alongside closed-form solutions like the Barone-Adesi and Whaley model (BAW) (Barone-Adesi and Whaley, 1987) for handling the early exercise feature.

Recent advancements continue to enhance precision and speed, with methods like the accelerated binomial model (Breen, 1991) and Markov chain techniques within the GARCH framework (Duan and Simonato, 2001). The regime-switching models proposed by Duan, Popova, and Ritchken (2002) further refine the understanding of volatility's asymmetric response to market changes. These address the limitations of traditional models, offering robust, adaptable solutions for modern financial markets. Innovations such as the de-Americanization method (Burkovska, Gaß, Glau, Mahlstedt, Schoutens, and Wohlmuth, 2018) reflect the ongoing effort to reconcile theoretical models with the volatile nature of actual markets, broadening the analytical tools available for option pricing.

3.1.2 Machine learning in option pricing

Parametric methods use predetermined asset return distributions to determine an option's fair value under the no-arbitrage principle. While economically robust and sometimes offering direct pricing formulas, these models often falter when confronted with real-world market data deviations. In contrast, machine learning-based models provide a more flexible approach by excelling in cap-

turing complex relationships between option prices and influencing factors. These models diverge into non-parametric and semi-parametric types. Non-parametric models, highlighted in seminal works like those by Hutchinson et al. (1994), are purely data-driven, freeing them from conventional economic constraints but also making them prone to overfitting. Advances such as neural network-based pricing by Gençay and Qi (2001) and combined neural networks and support vector regression techniques by Gradojevic et al. (2009) and Liang et al. (2009) showcase the potential of modular neural networks.

Semi-parametric models merge the economic rationale of parametric models with the flexibility of machine learning, offering a balanced approach exemplified by the Generalized Pricing Framework (GPF) (Andreou et al., 2010) and the Hybrid (HBD) (Lajbcygier and Connor, 1997) models as well as the model in Almeida et al. (2022) (AFFT). These models integrate machine learning predictions with parametric frameworks, effectively addressing the drawbacks of each method in isolation.

Recent advances in machine learning have also propelled developments in option pricing techniques. Works like De Spiegeleer, Madan, Reyners, and Schoutens (2018) on Gaussian process regression, Gan, Wang, and Yang (2020) on a model-free approach for pricing mean options, and Jang and Lee (2018) on generative Bayesian models for American options, demonstrate the field's ongoing evolution, highlighting machine learning's capacity to improve prediction accuracy and computational efficiency in derivative pricing.

3.1.3 Motivations and contributions

The Greeks provide investors with references to different dimensions of risk management for option pricing in addition to implied volatility. Papahristodoulou (2004) propose a Linear Programming approach that utilizes key Greek variables from the Black-Scholes equation, such as delta, gamma, theta, rho, and kappa, to ascertain the most effective hedging strategy. Gao (2009) suggest a broad linear programming framework that adjusts risk limits for all Greeks, demonstrating that widening these limits enhances potential returns, thus facilitating a customizable risk-return equilibrium. Chen, Lee, and Shih (2010) utilizes the Greek letters for call and put options on dividend-paying stocks and non-dividend-paying stocks, suggesting that the Greek letters reflect various types of risk and are valuable for risk management. Given the usefulness of the Greeks for

risk management, combining the Greeks with machine learning-based option pricing models that extract potential information from the Greeks may generate more accurate option pricing models.

As for firm characteristics, the impact of firm characteristics on option valuation in the absence of arbitrage is a complex research topic in finance. Many academic papers have provided valuable insights into this area. Rubinstein (1983) construct an option pricing formula that takes into account the risk associated with an individual firm's assets, incorporates varying degrees of risk and considers the effects of the firm's debt and dividend policies. Figlewski (1989) emphasize that market imperfections, such as unpredictable volatility and transaction costs, limit the practical application of arbitrage in option valuation. This emphasizes the potential impact of firm-specific factors (e.g., financial stability and market behaviour) on real-world option prices beyond idealized models based on arbitrage. Subramanian (2004) establish an arbitrage-free framework for pricing stock options for firms in mergers and acquisitions that takes into account discontinuous effects on stock prices. The model outperforms the Black-Scholes model in explaining observed option prices, suggesting that specific events such as corporate mergers and acquisitions can have a significant impact on option pricing in the absence of arbitrage. Recently, Trigeorgis and Lambertides (2014); Andreou (2015); Vasquez and Xiao (2023); Chen et al. (2023) demonstrate the profound impact of firm characteristics on the dynamics of the underlying asset and option values. Trigeorgis and Lambertides (2014) emphasize the importance of firm-specific business volatility and managerial flexibility in growth option valuation. As shown by Andreou (2015), market default risk as well as firm leverage and asset volatility as emphasized by Vasquez and Xiao (2023), are directly related to option pricing. Chen et al. (2023) document how firm fundamentals shape the implied volatility curve, which is crucial in option valuation. In addition, Zhan et al. (2022) illustrates the significant correlation between firm characteristics and delta-hedged stock option returns, further emphasizing the relevance of these factors not only to option pricing but also to predicting market behaviour and investment strategies. Andreou, Han, and Li (2023) also demonstrates that some firm characteristics can improve the accuracy of stock option pricing predictions when utilizing machine learning models.

Motivated by the interconnection between option Greeks and firm characteristics with option pricing, along with the strengths of machine learning algorithms and the adaptability of semi-

parametric option pricing models, this study broadens the scope of semi-parametric option pricing. We extend Andreou et al. (2023) to include Greek as novel input features and test the stability of different semi-parametric models under volatile market conditions.

Specifically, We begin with fundamental input variables closely linked to the parametric model, including factors like the moneyness (S/X), asset price (S), strike price (X), maturity (T), option type (c/p), and past implied volatility (σ^{avg}). We then expand our focus to include option Greeks and firm characteristics, examining their effectiveness in predicting option prices. We employ three semi-parametric methods to employ the additional input features. The first model is based on Andreou et al. (2010)'s GPF model, which employs machine learning for the prediction of implied volatility. The second model is based on Lajbcygier and Connor (1997)'s HBD model, which employs machine learning for pricing error correction. The third model is based on Almeida et al. (2022)'s AFFT model, which employs machine learning for implied volatility error correction. These models are compared against a parametric benchmark model. We select the binomial model with past implied volatility (σ^{avg}) as our benchmark (PARA).

We have also focused on the pricing performance of the above models during the Pandemic period. The impact of a global pandemic outbreak brings unprecedented market volatility and challenges the established norms of financial markets. The economic uncertainty during this period provided a rare opportunity to critically evaluate the performance and resilience of our machine-learning models. The pandemic causes a significant increase in financial market risk, and the stock market's response varies according to the intensity of the outbreak, thus increasing volatility and unpredictability (Baek, Mohanty, and Glambosky, 2020; Zhang, Hu, and Ji, 2020). By examining the pricing performance of the models we test during pandemics, we aim to gain insights into the predictive accuracy, adaptability, and overall robustness of these models in the face of unexpected market disturbances. In particular, we seek to understand how these models withstand the stress of such black swan events under different input configurations (Zhan, Fang, and Xu, 2017).

We evaluate the models using 10,110,377 U.S. stock options from January 2014 to December 2022. These options are associated with companies listed in the SP 500 index. Regarding the option Greeks, our focus is on Delta, Gamma, Theta, and Vega. Concerning the firm characteristics, we select 111 attributes as delineated in Jensen et al. (2021), excluding any characteristic that exhibits

a substantial amount of missing data. Specifically, we exclude firm characteristics with missing values greater than 20% over the sample period.

We first find that all semiparametric models, *i.e.*, GPF, HBD, and AFFT, outperform the benchmark model, PARA, even before incorporating additional features. The root mean square errors (RMSEs) for GPF, HBD, and AFFT are 2.100, 2.017, and 2.139, respectively, compared to 2.567 for PARA. Second, option Greeks improve the performance of GPF, HBD, and AFFT by reducing their RMSEs to 1.673, 1.723, and 1.733, respectively. Third, firm characteristics further improve the performance of GPF, HBD, and AFFT by reducing their RMSEs to 1.624, 1.688, and 1.678, respectively. Lastly, the combination of option Greeks and firm characteristics improves the performance of GPF, HBD, and AFFT the most: the RMSEs of the three models are reduced to 1.535, 1.622 and 1.580, respectively. The GPF model consistently outshines the others, particularly during the Pandemic. Our study also indicates that the precision of option valuation varies among stocks with different return distributions. Specifically, stocks with extremely high or low returns exhibited a higher mean absolute percentage error (MAPE) than those with more moderate returns.

Using option Greeks and firm characteristics in machine learning models for stock option pricing improves their accuracy and usability. Option Greeks enable real-time applications, especially in short-term pricing, by reflecting daily market fluctuations. This makes the models more responsive to current market conditions. Moreover, prioritizing option Greeks as key features can significantly reduce the training time for these algorithms by simplifying the data processed. Incorporating both option Greeks and firm characteristics into the model enhances its predicting accuracy but lengthens the training time. Managing a wide range of inputs requires substantial computational resources and time, particularly with the need for regular updates to reflect current market conditions. Consequently, this frequent retraining may strain computational capabilities, hindering real-time or near-real-time predictions for certain users or applications. To balance these detailed insights with the practicality of model deployment, developers face the challenge of managing model complexity without compromising computational efficiency. Mitigating strategies may involve selecting only the most influential option Greeks and firm characteristics, employing more streamlined machine learning algorithms, or adopting incremental learning methods to minimize the need for comprehensive retraining. In summary, including option Greeks and firm characteristics in stock

option pricing models enhances precision. In the GPF framework, incorporating option Greeks reduces the RMSE from 2.100 to 1.673, and adding firm characteristics further reduces it to 1.624. Combining both factors decreases the RMSE to 1.535. However, adding more variables increases computational demands. Careful selection of variables is essential for accuracy and practicality.

Our research makes contributions to the field of option valuation in four ways. First, it breaks new ground by combining option Greeks with machine learning methods, demonstrating the unexplored potential of option Greeks in stock option valuation. Second, it combines firm characteristics with option Greeks to construct a more efficient and resilient option pricing framework. Third, by examining three distinct machine-learning-based semi-parametric option valuation models during the Pandemic, this study identifies the model that is most adept at volatile markets. Finally, it enriches the existing knowledge in this area by emphasizing the critical impact of stock return distributions on option pricing effectiveness.

This paper is structured as follows. Section 3.2 describes the option pricing models used in this study, Section 3.3 details the data and methodology for the empirical analysis, Section 3.4 presents the empirical results, and Section 3.5 concludes.

3.1.4 Upgrades in this chapter compared to Chapter 1

This study builds on Chapter 1 by exploring the integration of option Greeks and firm characteristics into machine learning-enhanced semi-parametric option pricing models. Option Greeks, updated daily, are pivotal for short-term option price predictions, reflecting current market sentiments and risk assessments. In contrast, firm characteristics, which change less frequently, are crucial for long-term predictions, offering insights into the underlying asset's health and prospects.

We also investigate the distinct impacts of option Greeks and firm characteristics on pricing performance and assess the consistency of three semi-parametric models' performance before and during the pandemic to identify the most stable model for practical selection. Specifically, we examine the GPF and AFFT models, both predicting implied volatility but differing in complexity. The GPF model, predating the AFFT, employs a simpler mechanism that we hypothesize to be more efficient than the AFFT's addition of implied volatility residuals, which increases model complexity.

By combining option Greeks and firm characteristics, the models enhance prediction accuracy across short-term market fluctuations and long-term fundamental changes. This synergy between daily updated Greeks and the more stable firm characteristics creates a comprehensive tool for option pricing, beneficial for rapid decision-making by traders and analysts.

Another extension in Chapter 3, as compared to Chapter 1, is the examination of two additional parametric models for pricing equity options. These models are tested to see how they compare with the binomial tree approach, providing a broader perspective on the methodologies available for option pricing. This exploration allows for a more comprehensive comparison and understanding of how different models handle the integration of option Greeks and firm characteristics, potentially leading to improved prediction accuracy and practical application in varying market environments.

Overall, this study aims to demonstrate that a balanced approach, leveraging the immediate relevance of option Greeks and the depth of firm characteristics, improves the applicability and accuracy of machine learning models in option pricing, offering valuable insights for model development and selection in varying market conditions.

3.2 The option pricing models

3.2.1 Binomial tree method

The binomial option pricing model, as described by Cox et al. (1979), is a common approach for determining the fair market value of an option, further supported by studies like Rubinstein (1994); Broadie and Detemple (1996). This model assumes that the price of the underlying asset changes in a fixed proportion at each interval. Primarily applied to European-style options, which are exercisable only at their expiry, it is also adaptable for American-style options that can be exercised anytime before their expiration.

For American options, the binomial model incorporates adjustments to consider early exercise. At every interval, the investor has the choice to exercise the option, receiving its intrinsic value - the difference between the current price of the underlying asset and the strike price. Exercising is favourable if the intrinsic value surpasses the option's present value.

To estimate the fair value of an American option using the binomial approach, we employ the methodology outlined by Cox et al. (1979).

1. Calculate the size of each time step (dt) using the following formula:

$$dt = \frac{T}{M}, \tag{3.1}$$

where T is the maturity, M is the total number of time steps[1].

2. Calculate the up factor (u) and the down factor (d) using the following formulas:

$$u = \exp(\sigma \times \sqrt{dt}) \tag{3.2}$$

$$d = \frac{1}{u}, \tag{3.3}$$

where σ is the volatility of the underlying asset.

3. Calculate the probability of an up move (p) using the following formula:

$$p = \frac{r \times dt - d}{u - d}, \tag{3.4}$$

where r is the risk-free interest rate, q is the dividend yield.

4. Determine the option's value at its expiry for each potential price of the underlying asset. Considering an asset with price S and a constant dividend rate q, as expanded in Cox et al. (1979) for dividend-paying stocks, and a strike price X at a specific time, the value of a call option is the $max(S \cdot (1 - q) - X, 0)$. Similarly, for a put option, the value is the greater of $max(X - S \cdot (1 - q), 0)$.

5. Working backwards from expiration, calculate the value of the option at each time step for each possible price of the underlying asset using the following formulas:

$$Price_u = \frac{p \times Price_{uu} + (1 - p) \times Price_{ud}}{1 + rf}, \tag{3.5}$$

[1] For the purpose of saving computation time, we set the value of M to 100.

$$Price_d = \frac{p \times Price_{du} + (1-p) \times Price_{dd}}{1+rf}, \tag{3.6}$$

where $Price_{uu}$, $Price_{ud}$, $Price_{du}$, and $Price_{dd}$ are the option values if the underlying asset price goes up twice, up then down, down then up, and down twice, respectively.

6. At each time step, the option's value is determined by choosing the higher of the two: the immediate payoff from exercising the option (intrinsic value[2]) or the expected value of holding the option(theoretical value[3]). This decision considers the advantage of early exercise available in American options. Such evaluation is crucial because of the possibility of exercising the option early. At every stage, we determine whether it's more beneficial to exercise the option for its current intrinsic value rather than holding onto it for potential future gains, taking into account the risk-free interest rate and the probability of price changes.

To find the market implied volatility for option i on t, $\sigma_{i,t}^{mrk}$, we solve an optimization problem that has the following form:

$$\sigma_{i,t}^{mrk} = \arg\min_{\sigma_{i,t}^{mrk}} [P_{i,t}^{mrk} - PARA(\sigma_{i,t}^{mrk}, S_{i,t}, X_{i,t}, T_{i,t}, rf_{i,t})]^2, \tag{3.7}$$

where $P_{i,t}^{mrk}$ is the market price for option i on t, $PARA(\sigma_{i,t}^{mrk}, S_{i,t}, X_{i,t}, T_{i,t}, rf_{i,t})$ is the binomial model that employs the parameters $S_{i,t}$, $X_{i,t}$, $T_{i,t}$, and $rf_{i,t}$, which are the stock price, the strike price, the maturity, and the risk-free rate for option i on t.

For the Greeks of option i on t, we follow the following equations under the binomial tree model:

Delta (Δ): The delta of an option measures the sensitivity of the option's price to changes in the underlying asset's price. It can be calculated as follows:

$$\Delta = \frac{Price_u - Price_d}{S \cdot (u-d)}. \tag{3.8}$$

Gamma (Γ): The gamma of an option measures the rate of change of delta concerning changes

[2] The intrinsic value of a call option is equal to $max(S \cdot (1-q) - X, 0)$, and the intrinsic value of a put option is equal to $max(X - S \cdot (1-q), 0)$

[3] $\frac{p \cdot Price_u + (1-p) \cdot Price_d}{1+rf}$

in the underlying asset's price. It can be calculated as follows:

$$\Gamma = \frac{\Delta_u - \Delta_d}{S \times (u - d)}. \tag{3.9}$$

Theta (Θ): The theta of an option measures the rate of change of the option's price concerning time decay. It can be calculated as follows:

$$\Theta = \frac{Price_{t+dt} - Price_t}{dt}. \tag{3.10}$$

Vega (V): The vega of an option measures the sensitivity of the option's price to changes in implied volatility. It can be calculated as follows:

$$V = \frac{Price(\sigma^{mrk} + 0.01) - Price(\sigma^{mrk})}{0.01}. \tag{3.11}$$

To estimate the price of option i on day t, we use the binomial model that employs σ_t^{avg} as our benchmark model (PARA), where σ_t^{avg} is the average of the implied volatilities of all the options with the same underlying asset in the period from $t-10$ to $t-1$ (i.e., over the past 10 trading days) as Equation (3.12).

$$\sigma_{i,t}^{avg} = \frac{1}{10} \sum_{j=10}^{j=1} \sigma_{i,t-j}^{mrk}, \tag{3.12}$$

where the $\sigma_{i,t-j}^{mrk}$ is the market implied volatility of option i on $t-j$.

We apply the same process to calculate the average Greeks:

$$G_{i,t}^{avg} = \frac{1}{10} \sum_{j=10}^{j=1} G_{i,t-j}^{mrk}, \tag{3.13}$$

where the $G_{i,t-j}^{mrk}$ is the implied Greeks of option i on $t-j$ and $G \in [\Delta, \Gamma, \Theta, V]$.

3.2.2 Machine learning structures

The binomial option pricing model is extended by three machine learning structures: GPF, HBD and AFFT.

3.2.2.1 GPF structure

Andreou et al. (2010) introduce a semi-parametric option pricing model known as the GPF model, demonstrating its superior ability to price index options. This model utilizes a neural network to estimate unseen variables critical for option pricing (such as volatility in the Black-Scholes model and volatility, skewness, and kurtosis in the Corrado and Su (1996) model) and then integrates these predictions into established parametric option pricing models. The strength of the GPF model lies in its synthesis of parametric and non-parametric techniques: the parametric component provides a foundational theoretical structure for pricing options, while the machine learning aspect enhances the accuracy of predicting input variables. This approach allows the GPF model to effectively overcome the shortcomings inherent in solely parametric or non-parametric methods. Additionally, its adaptability makes the GPF model a valuable and flexible asset in various option pricing contexts.

In our research, we utilize the adaptable nature of the GPF by employing firm characteristics in the forecasting model, as shown in Equation (3.14).

$$P_{GPF}^{pre} = PARA(\sigma^{pre}, S, X, T, q, rf), \tag{3.14}$$

where P_{GPF}^{pre} is the predicted price through the GPF structure, S is the close price of the underlying asset, X is the strike price, T is the maturity, q is the dividend rate, and rf is the risk free rate. The function $PARA(\cdot)$ represents the process of the binomial tree model, and σ^{pre} is obtained through Equation (3.16).

The initial GPF model uses a single-step process, wherein it trains a machine learning algorithm to reduce the error in option pricing. This reduction is measured using a loss function based on the mean squared error of the option prices. In our modified approach, designed specifically for the large-scale stock options dataset, we recalibrate the loss function to focus on implied volatility. This adjustment enables a quicker and more relevant reduction in pricing errors, particularly effective for the size and characteristics of our dataset. Our GPF model's loss function is presented in Equation (3.15):

$$Loss(GPF)_t = \min \sum_{i=1}^{N} (\sigma_{i,t}^{mrk} - \sigma_{i,t}^{pre})^2, \tag{3.15}$$

where t is the date of the training, i is the i-th observation, N represents total observations on t. $\sigma_{i,t}^{mrk}$ represents the market implied volatility of option observation on t, and the $\sigma_{i,t}^{pre}$ represents the predicted implied volatility of option observation i on t.

The estimation of the σ^{pre} can be summarized as follow:

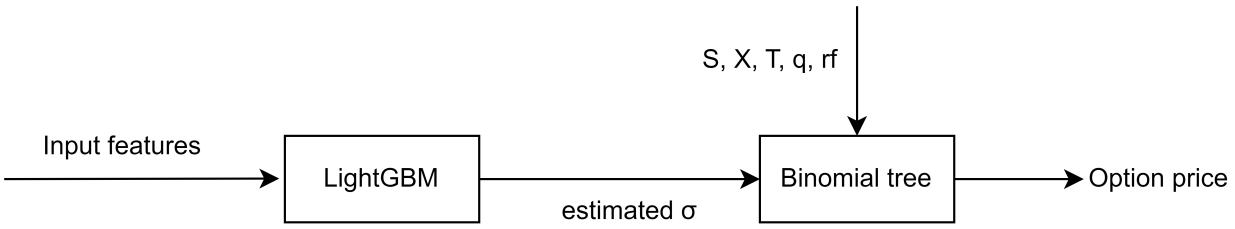
$$\sigma_{i,t}^{pre} = \text{LightGBM}(F_{i,t}; \phi), \tag{3.16}$$

where, $\sigma_{i,t}^{pre}$ is the predicted implied volatility for the i-th observation on t, as determined by the LightGBM model. $F_{i,t}$ represents the vector of input features for the i-th observation, such as the moneyness (S/X), close price (S), strike price (X), maturity (T), σ^{avg}, Greeks, and firm characteristics. θ denotes the set of parameters of the LightGBM model.

This equation represents how LightGBM utilizes input variables (F) to estimate implied volatility, which is a crucial component in the GPF model. The variables can encompass several pertinent elements such as moneyness (S/X), prices of the underlying asset (S), strike prices (X), maturity (T), metrics from option Greeks, and additional corporate factors. Throughout its training phase, the parameters of the LightGBM model (ϕ) are refined to reduce the loss function previously detailed as Equation (3.15).

Our research extends the adaptability of the GPF by incorporating firm characteristics into its predictive mechanism, as shown in Equation (3.14). This advancement utilizes the GPF's proficiency in integrating extra, pertinent variables smoothly. The inclusion of firm characteristics is intended to augment the GPF model, allowing it to encompass a wider array of data that may influence option pricing. This expansion of the GPF model retains its inherent accuracy in option valuation while broadening its capability to incorporate a variety of data points, enhancing the model's predictive accuracy. Instead of a neural network, our modified version employs LightGBM, a gradient-boosting framework noted for its exceptional efficacy in diverse machine-learning scenarios. The conceptual diagram of our revised GPF model is illustrated in Figure 3.2.

Figure 3.2. Schematic description of the GPF structure. In the GPF-based models, the option value is provided by a parametric model. Input features denote the set of all input parameters for the LightGBM model. S, X, T, q, and rf are directly passed to the binomial tree model.



3.2.2.2 HBD structure

The hybrid (HBD) model, introduced by Lajbcygier and Connor (1997), also merges the strengths of a parametric approach and machine learning to enhance the precision of option pricing. This model differs from the GPF by focusing on predicting the price residual of a parametric model using machine learning techniques. Initially, HBD utilizes a parametric model like the binomial tree model, with past implied volatility as an input, to calculate the option's price. This calculated price is then compared with the market price to determine the pricing residual. Subsequently, a machine learning algorithm predicts this residual. The final value of the option is determined by adding the predicted residual to the model's price. The process of the HBD model can be outlined as follows.

$$P_{HBD}^{pre} = Res^{pre} + PARA(\sigma^{avg}, S, X, T, q, rf), \tag{3.17}$$

where P_{HBD}^{pre} is the predicted price through the HBD structure, S is the close price of the underlying asset, X is the strike price, T is the maturity, q is the dividend rate, and rf is the risk free rate. The function $PARA(\cdot)$ represents the process of the binomial tree model, σ^{avg} is the average implied volatility according to Equation (3.12).

The loss function we optimize in the HBD is shown in Equation (3.18) and (3.19).

$$Loss(HBD)_t = \min \sum_{i=1}^{N} (Res_{i,t}^{P,mrk} - Res_{i,t}^{P,pre})^2, \tag{3.18}$$

$$Res_{i,t}^{P,mrk} = P_{i,t}^{mrk} - PARA(\sigma_{i,t}^{avg}, S_{i,t}, X_{i,t}, T_{i,t}, rf_{i,t}), \tag{3.19}$$

where t is the date of the training, i is the i-th observation, N represents total observations on t. Res^{mrk} represents the residual between the market option price (P^{mrk}) and the binomial tree-based option pricing, Res^{pre} represents the residual that the HBD model seeks to predict.

The estimation of the Res^{pre} can be summarized as follow:

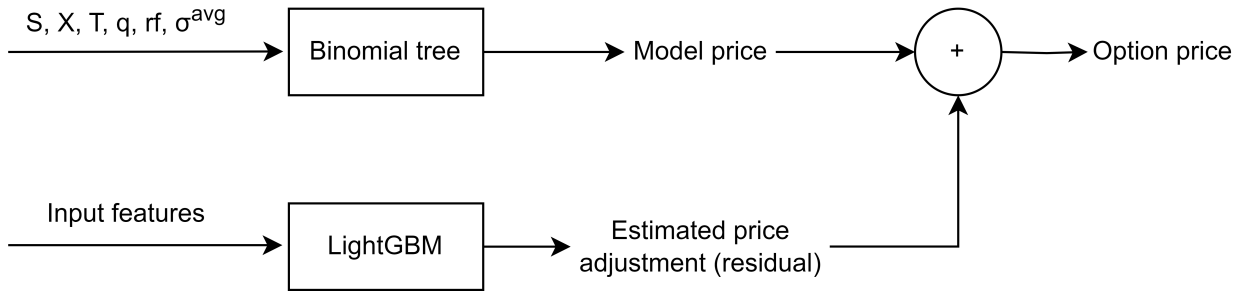
$$Res_{i,t}^{P,pre} = \text{LightGBM}(F_{i,t}; \phi), \tag{3.20}$$

where, $Res_{i,t}^{P,pre}$ is the predicted price residuals for the i-th observation on t, as determined by the LightGBM model. $F_{i,t}$ represents the vector of input features for the i-th observation, such as the moneyness (S/X), close price (S), strike price (X), maturity (T), σ^{avg}, Greeks, and firm characteristics. θ denotes the set of parameters of the LightGBM model.

This equation demonstrates how the LightGBM technique uses a set of features, denoted as F, to predict the price deviation, termed as (Res^{pre}), a crucial element in the HBD model. The feature set, F, may include various relevant data points like the moneyness (S/X), the prices of the underlying assets (S), strike prices (X), maturity (T), various option Greeks, and additional company-specific attributes. During the training process, LightGBM's parameters, represented by ϕ, are adjusted to reduce the loss outlined in Equation (3.18).

We enhance the HBD model by adding firm characteristics and using LightGBM instead of a neural network. Figure 3.3 presents a diagrammatic representation of the HBD model's structure.

Figure 3.3. Schematic description of the HBD structure. In the HBD-based models, the option value is provided by the combination of binomial tree model-based price and estimated price adjustment. Input features denote the set of all input parameters for the LightGBM model. S, X, T, q, rf, and σ_{avg} are directly passed to the binomial tree model.



3.2.2.3 AFFT structure

The AFFT Structure, as referenced in Almeida et al. (2022), serves as an alternative to the HBD structure. Unlike the HBD, which employs a machine learning model to forecast the residual of option price in a parametric model, the AFFT focuses on predicting the residual of implied volatility. It begins by determining the implied volatility residual through a comparison between the past implied volatility used in the parametric model (σ^{avg}) and the market's implied volatility (σ^{mrk}). This residual is then used as the target for the machine-learning model. Once we have a trained model, it predicts the implied volatility residual. This predicted residual, combined with past implied volatility data, aids in calculating the predicted implied volatility. This predicted value is then fed into the parametric model to generate the option price. The AFFT can be outlined in the following manner.

$$P_{AFFT}^{pre} = PARA(\sigma^{avg} + Res^{\sigma,pre}, S, X, T, q, rf), \tag{3.21}$$

where P_{AFFT}^{pre} is the predicted price through the AFFT structure, σ^{avg} is the average implied volatility provided by Equation (3.12), $Res^{\sigma,pre}$ is the predicted σ residual, S is the price of the underlying asset, X is the strike price, T is the maturity, q is the dividend rate, and rf is the risk free rate. The function $PARA(\cdot)$ represents the process of the binomial tree model, σ^{avg} is the average implied volatility according to Equation (3.12).

The loss function we optimize in the AFFT is shown in Equation (3.22) and (3.23).

$$Loss(AFFT)_t = \min \sum_{i=1}^{N} (Res_{i,t}^{\sigma,mrk} - Res_{i,t}^{\sigma,pre})^2, \tag{3.22}$$

$$Res_{i,t}^{\sigma,mrk} = \sigma_{i,t}^{mrk} - \sigma_{i,t}^{avg}, \tag{3.23}$$

where t is the date of the training, i is the i-th observation, N represents total observations on t. $Res^{\sigma,mrk}$ represents the residual between the market implied volatility provided by the Equation (3.7) (σ^{mrk}) and the average implied volatility given by the Equation (3.12), σ^{avg} represents the residual that the AFFT model seeks to predict.

The estimation of the Res^{σ} can be summarized as follow:

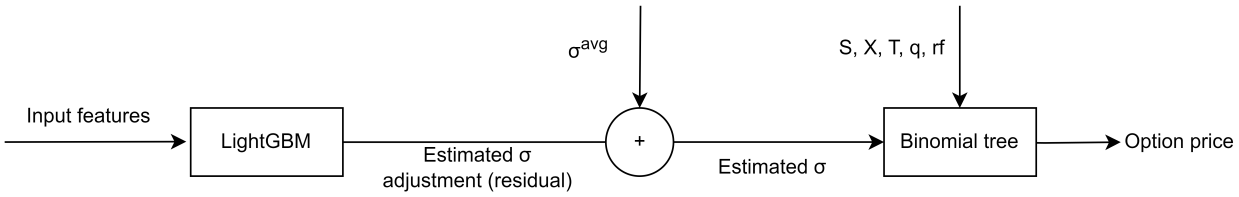
$$Res_{i,t}^{\sigma,pre} = \text{LightGBM}(F_{i,t}; \phi), \tag{3.24}$$

where, $Res_{i,t}^{\sigma,pre}$ is the predicted implied volatility residuals for the i-th observation on t, as determined by the LightGBM model. $F_{i,t}$ represents the vector of input features for the i-th observation, such as the moneyness (S/X), close price (S), strike price (X), maturity (T), σ^{avg}, Greeks, and firm characteristics. θ denotes the set of parameters of the LightGBM model.

This equation demonstrates how LightGBM uses the feature set F to estimate the implied volatility residual, $(Res^{\sigma,pre})$, a key element in the AFFT model. The feature set F may include a range of relevant data points, such as moneyness (S/X), prices of the underlying asset (S), strike prices (X), maturity (T), option Greeks, and other firm characteristics. During its training process, LightGBM's parameters, ϕ, are optimized to reduce the specified loss function, as detailed in Equation (3.22).

The AFFT framework is enhanced by adding option Greeks and firm characteristics into its model and applying the LightGBM technique. The architectural layout of the AFFT is depicted in Figure 3.4.

Figure 3.4. Schematic description of the AFFT structure. In the AFFT-based models, the option value is provided by the binomial tree model. Input features denote the set of all input parameters for the LightGBM model. S, X, T, q, and rf, along with the combination of σ_{avg} and the estimated σ adjustment, are directly passed to the binomial tree model.



3.2.3 LightGBM

In our study, we utilize the LightGBM machine learning algorithm, introduced by Ke et al. (2017). LightGBM. LightGBM stands for Light Gradient Boosting Machine and is a framework that uses decision tree algorithms for both classification and regression problems. It is specifically

crafted to be effective and scalable for large datasets. Below, we provide a summary of the algorithm and the essential formulas for implementing LightGBM in regression tasks.

1. Initialize the model: Start with an initial model, which is usually the average of the target variable values for regression tasks. Mathematically, the initial model can be represented as:

$$F_0(x) = arg\min_c \sum L(y_i, c) \tag{3.25}$$

where $L(y_i, c)$ is the loss function for the target variable y_i corresponding to observation x_i and the constant value c.

2. Gradient boosting: Iteratively construct weak learners (decision trees) and combine them to create a strong learner. For each iteration $k = 1, 2, ..., K$:

   (a) Calculate the gradients g_i and Hessians h_i for each observation x_i in the dataset:

$$g_i = \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \tag{3.26}$$

$$h_i = \frac{\partial^2 L(y_i, F(x_i))}{\partial F(x_i)^2} \tag{3.27}$$

   where $F(x_i)$ is the current model's prediction for observation x_i.

   (b) Build a new decision tree to fit the negative gradients: Fit a new decision tree $f_k(x)$ using the dataset (x_i, g_i, h_i). In LightGBM, the decision tree is built using the 'leaf-wise' strategy with depth limitation, which splits the tree node with the highest loss reduction first.

   (c) Determine the optimal learning rate (shrinkage factor) η_k:

$$\eta_k = arg\min_\eta \sum [L(y_i, F(x_i) + \eta \times f_k(x_i)] \tag{3.28}$$

   Utilize line search or another optimization method to find the learning rate that minimizes the loss function.

(d) Update the model:

$$F_k(x) = F_{k-1}(x) + \eta_k \times f_k(x) \tag{3.29}$$

3. Final model: After T iterations, the final model can be represented as:

$$F(x) = F_0(x) + \sum_{k=1}^{K} \eta_k \times f_k(x) \tag{3.30}$$

3.3 Data and methodology

3.3.1 Option data

Options data are obtained from OptionMetrics' IvyDB US and cover the options in the US market from January 2014 to December 2022. We use the first 365 calendar days of the sample to train the machine learning-based models and set the out-of-sample period to be from January 2015 to December 2022. Following Dumas et al. (1998), we define the option price as the midpoint of the bid and ask prices to reduce the estimation noise of implicit parameters. The underlying stocks' prices are collected from the Securities table in OptionMetrics' IvyDB US, and the 3-month Treasury bill rate, which is used as a proxy for the risk-free rate, is obtained from the St. Louis Federal Reserve Economic Data. Following Bakshi et al. (1997) and Andreou et al. (2010), we filter the option data using the following criteria.

1. Options with a trading volume of less than 100 contracts are eliminated as they are deemed to be illiquid and their prices may not represent the actual market price.

2. The time to maturity should be at least six days and no longer than 365 calendar days as options near expiration may induce liquidity-related biases.

3. Options with price quotes less than 0.1 U.S. dollars are eliminated.

4. The moneyness (S/X) of an option should be between 0.8 and 1.2.

5. Observations that violate the usual no-arbitrage condition are dropped.

After filtering, the final data set contains 10,110,377 observations. The characteristics of this dataset are detailed in Tables 3.1 and 3.2. Table 3.1 indicates a higher count of out-of-the-money option observations compared to in-the-money option observations and a greater number of call options relative to put options. Across all maturity groups, a volatility smile pattern is evident, with a more distinct presence in options with shorter maturities. Table 3.2 presents the annual distribution of options per underlying stock. Stocks in the SP500 with a daily options trading volume exceeding 100 contracts have risen from 265 in 2014 to 303 in 2022. Concurrently, the average number of options associated with each stock has grown over the years, from an average of 12 in 2014 to 20 in 2022. Despite the median number of actively traded options remaining relatively low at 5 in 2022, there's a slight increase from 4 in 2014. Moreover, the peak number of options has escalated from 254 in 2014 to 351 in 2022.

Between January 2020 and December 2022, coinciding with the Pandemic period, there was a notable increase in options trading, demonstrated by the escalated average daily trading volume of options as well as the peak numbers reached on certain days. This heightened trading activity likely indicates the market's response to the global uncertainties and evolving investment prospects brought about by the Pandemic.

Furthermore, the fact that the peak trading volumes of options far exceed the average suggests that a small selection of stocks have a disproportionately large number of options, while the majority are associated with fewer options.

3.3.2 Firm characteristics

The integration of firm characteristics into option pricing is supported by extensive research demonstrating their significant impact on the behaviour of underlying assets and, consequently, on the value of options. The relationship between characteristics unique to a firm and overall market trends is vital. This is shown by Andreou (2015), who discovered that the market's perception of a firm's default risk profoundly affects the risk-neutral characteristics, such as volatility and skewness, of SP 500 Index options. Vasquez and Xiao (2023) also point out the influence of default risk on the expected returns of delta-hedged equity options, thus establishing a direct connection between a firm's debt levels, asset volatility, and option pricing. Chen et al. (2023) further this

Table 3.1. Characteristics of option data. This table describes the characteristics of the option data. The data is obtained from OptionMetrics' IvyDB US and covers the stock options in the US market from January 2014 to December 2022, with out-of-sample data covering the period January 2015 to December 2022. "Price", "Implied volatility", "Obs (Total)" and "Obs (Out)" respectively refer to the average price, average implied volatility, the number of total observations, and the number of out-of-sample observations in each subset.

Moneyness	**DOTM**	**OTM**	**JOTM**	**ATM**	**JITM**	**ITM**	**DITM**
Panel A. Call options							
S/X	0.80-0.90	0.90-0.95	0.95-0.99	0.99-1.01	1.01-1.05	1.05-1.10	1.10-1.20
Near-term (6-60 days)							
Price	1.977	2.320	3.007	4.875	7.063	11.097	17.457
Implied Volatility	0.389	0.352	0.321	0.309	0.318	0.333	0.341
Obs (Total)	330,559	671,159	1,304,056	800,659	592,793	196,766	98,343
Obs (Out)	321,891	639,602	1,211,439	733,610	538,122	180,286	91,229
Mid-term (60-180 days)							
Price	2.176	2.548	3.214	4.813	7.051	12.246	25.063
Implied Volatility	0.386	0.358	0.337	0.329	0.345	0.372	0.400
Obs (Total)	352,923	542,265	908,281	571,124	389,568	113,081	48,747
Obs (Out)	340,772	511,163	836,955	521,703	354,768	105,769	46,696
Long-term (181-365 days)							
Price	3.283	3.911	5.636	8.774	9.534	11.872	15.874
Implied Volatility	0.351	0.317	0.307	0.306	0.308	0.310	0.316
Obs (Total)	284,259	321,380	319,067	139,373	150,835	79,712	57,664
Obs (Out)	269,715	291,437	281,316	122,348	131,915	70,462	51,435
Panel B. Put options							
S/X	1.10-1.20	1.05-1.10	1.01-1.05	0.99-1.01	0.95-0.99	0.90-0.95	0.80-0.90
Near-term (6-60 days)							
Price	3.644	4.484	5.899	7.965	9.307	11.905	18.327
Implied Volatility	0.348	0.333	0.331	0.334	0.344	0.355	0.378
Obs (Total)	190,281	169,882	175,963	87,902	92,948	41,780	25,502
Obs (Out)	174,963	151,357	155,687	77,721	82,720	37,719	23,720
Mid-term (60-180 days)							
Price	7.405	8.256	10.693	14.039	14.521	16.114	18.984
Implied Volatility	0.335	0.316	0.314	0.312	0.315	0.314	0.318
Obs (Total)	193,858	134,855	115,069	52,050	64,241	45,476	48,158
Obs (Out)	178,932	120,538	101,927	45,722	56,317	39,740	42,256
Long-term (181-365 days)							
Price	7.677	9.323	11.800	14.772	16.354	18.829	28.609
Implied Volatility	0.338	0.336	0.339	0.343	0.355	0.366	0.381
Obs (Total)	127,894	79,447	69,379	33,855	39,544	25,476	24,203
Obs (Out)	115,566	71,556	62,399	30,310	35,517	23,081	22,299

Table 3.2. Number of options per stock. This table describes the descriptive statistics of the number of options per stock per day in the sample period.

Year	No. firms	No. options						
		Mean	Std	Min	25%	50%	75%	Max
2014	265	12	21	1	2	5	13	254
2015	262	12	21	1	1	4	13	243
2016	281	11	18	1	1	4	13	179
2017	290	12	20	1	1	4	13	172
2018	312	14	25	1	2	4	14	213
2019	308	14	24	1	1	4	15	202
2020	313	17	33	1	1	4	18	316
2021	314	20	40	1	2	5	20	349
2022	303	20	41	1	2	5	19	351

idea by showing how a firm's basic financial health shapes the implied volatility curve, a crucial factor in determining option prices. Their research indicates that various firm-specific aspects like profitability and market dominance significantly affect the differences in option prices across different firms. In a similar vein, Trigeorgis and Lambertides (2014) highlight the importance of a firm's specific business volatility and the adaptability of management in evaluating growth options, reinforcing the need to consider these factors in option pricing models. Firm characteristics are also key in studies on predicting the returns of options. Research has found links between these traits and the returns of delta-hedged equity options. Zhan et al. (2022) show that the expected returns from selling delta-hedged call options are inversely related to elements like stock price profit margin and firm profitability, but positively associated with factors such as cash reserves, variability in cash flow, issuing new shares, total external financing, risk of financial distress, and the range of analysts' forecasts. This underlines the significant role that specific firm attributes play in option returns, emphasizing their relevance not only in pricing options but also in forecasting market behaviour and shaping investment strategies.

Our study utilizes 111 firm characteristics that include accounting ratios, momentum features, stock return volatility, and other characteristics related to the firm's operation, growth, risk, and performance by following Jensen et al. (2021). To be more practical and consistent, these firm characteristics are collected using the PyAnomaly Python library, which is designed for creating firm characteristics and aiding in asset pricing research[4]. We exclude any firm characteristic with over 20% missing observations. We address the remaining gaps in data by replacing missing values

[4] This tool is accessible at [https://pyanomaly.readthedocs.io/en/latest/index.html]. PyAnomaly offers straightforward access to a range of financial datasets, ensuring both accuracy and consistency in data.

with the cross-sectional median value following Gu et al. (2020). A detailed list and description of these firm characteristics are available in the appendix. Jensen et al. (2021) and other cited references offer complete definitions of these characteristics. To prevent any forward-looking bias, we align the firm characteristics from six months prior with the option data from the current month.

3.3.3 Model specifications and evaluation metrics

We have organized our inputs into three main categories. The first, the basic group, encompasses variables such as $Moneyness$ (S/X), S (close price), X (strike price), T (maturity), c/p (a binomial indicator with 0 for call option and 1 for put option), and σ^{avg}. The second category, the Greeks group, includes Δ^{avg}, Γ^{avg}, Θ^{avg}, and V^{avg}. The third, the firm characteristics group, encapsulates 111 firm characteristics.

Our research assesses the accuracy of different models in predicting out-of-sample option prices. We begin with the binomial model, labeled PARA. We then examine various configurations of the GPF structure. These variations are: GPF, using only the basic group; GPF_G, which combines the basic group with the Greeks group; GPF_F, merging the basic group with firm characteristics; and GPF_{GF}, integrating all three groups. Similarly, we evaluate the HBD structure: HBD, focused on the basic group; HBD_G, including both basic and Greek groups; HBD_F, blending the basic group with firm characteristics; and HBD_{GF}, employing all three groups. Lastly, we consider the AFFT structure: AFFT uses only the basic group; AFFT_G includes the Greeks along with the basic inputs; AFFT_F combines the basic group with firm characteristics; and AFFT_{GF} integrates all three groups. Each model's pricing performance is compared to determine which input combination most accurately reflects market prices.

Table 3.3 summarizes the combinations of input features and the tested models in the empirical analysis.

The evaluation metrics of the models include root mean squared error (RMSE), mean absolute error (MAE), root mean squared percentage error (RMSPE), and mean absolute percentage error

Table 3.3. Model specifications. This table presents the specifications of the test models. Panel A lists the input features of the models, and panel B lists the models tested in the empirical study.

Panel A. Input features		
Basic group	Moneyness (S/X), S (close price), X (strike price), T (maturity), c/p (a binomial indicator with 0 for call option and 1 for put option), and σ^{avg} (past implied volatility)	
Greeks group	Δ^{avg}, Γ^{avg}, Θ^{avg}, V^{avg}	
Firm characteristics	111 firm characteristics six months prior.	
Panel B. Test models		
Model	Input features	Structure
GPF	Basic group	GPF
GPF_G	Basic group + Greeks Group	GPF
GPF_F	Basic group + Firm characteristics	GPF
GPF_{GF}	Basic group + Greeks Group + Firm characteristics	GPF
HBD	Basic group	HBD
HBD_G	Basic group + Greeks Group	HBD
HBD_F	Basic group + Firm characteristics	HBD
HBD_{GF}	Basic group + Greeks Group + Firm characteristics	HBD
AFFT	Basic group	AFFT
AFFT_G	Basic group + Greeks Group	AFFT
AFFT_F	Basic group + Firm characteristics	AFFT
AFFT_{GF}	Basic group + Greeks Group + Firm characteristics	AFFT

(MAPE) as detailed in Equations (3.31), (3.32), (3.33), and (3.34).

$$RMSE = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(z_l - \hat{z}_l)^2}, \tag{3.31}$$

$$MAE = \frac{1}{L}\sum_{l=1}^{L}|z_l - \hat{z}_l|, \tag{3.32}$$

$$RMSPE = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\frac{z_l - \hat{z}_l}{z_l}\right)^2}, \tag{3.33}$$

$$MAPE = \frac{1}{L}\sum_{l=1}^{L}\left|\frac{z_l - \hat{z}_l}{z_l}\right|, \tag{3.34}$$

where L represents the number of observations in the dataset, z_l is the actual value of the j-th observation, and $\hat{z}_l$ is the predicted value for the l-th observation, as given by the model.

We use the Model Confidence Set (MCS) method, a statistical technique designed for model comparison and selection in time series analysis, to evaluate a group of predictive models. In-

troduced by Hansen et al. (2011), the MCS helps identify models with statistically comparable performance. Our research implements MCS through a series of defined steps.

1. **Selection of Prediction Models**: Initiate the MCS procedure by selecting models based on their mean square error (MSE) for comparison.

2. **Null Hypothesis Formulation**: Assume no significant MSE differences among models and test this assumption.

3. **Loss Differential Analysis**: Create a matrix of MSE differences between models for statistical analysis.

4. **Statistical Testing**: Apply statistical tests to assess if MSE differences are significant.

5. **Interpreting p-values**: Analyze p-values using a 1% significance threshold to determine model inclusion in the MCS.

3.3.4 Implementation Details

In our research, we employ the LightGBM gradient boosting method for regression tasks, using default hyperparameters such as the number of leaves, tree depth, and learning rate to maintain simplicity. The mean squared error is chosen as the loss function.

For the GPF model, we daily compute the average implied volatility, σ^{avg}, and integrate it with basic group features to form the input for LightGBM. Additional features like Greeks and firm characteristics are included in variations of the GPF model. LightGBM is trained on data from the 365 days, using the last 30 days for validation and the earlier period for training, and then used to estimate the next day's implied volatility, which feeds into a binomial tree model to determine option prices.

The HBD approach calculates option prices using a binomial tree, then derives the price residuals—differences between market and parametric model-derived prices—which serve as training targets for LightGBM. The output is an estimated price residual used to adjust the parametric model-based option price.

Similarly, the AFFT approach calculates the implied volatility residual, which is then used as a target for training LightGBM. The resulting estimated implied volatility residual is added to σ^{avg} to compute the predicted implied volatility for option pricing.

All models are evaluated using out-of-sample metrics, with the MCS approach employed to compare their performance and identify the model with the lowest MSE and highest p-value, indicating significant superior accuracy.

3.4 Empirical analysis

3.4.1 Pricing performance for all options

3.4.1.1 Overall performance

Table 3.4 shows the performance of four option pricing models: PARA, GPF, HBD, and AFFT, tested across all options.

Table 3.4. Pricing performance for all options. This table shows the out-of-sample pricing errors for each model for the period January 2015 to December 2022 for all the options, containing a total of 9,302,680 observations. RMSE, MAE, RMSPE, and MAPE respectively refer to the root mean squared error, the mean absolute error, the root mean squared percentage error, and the mean absolute percentage error. MCS-p refers to the p-value of the model confidence set test. Columns labelled "%" indicate the proportion of months during the entire forecast period in which each model ranked first according to the error metric specified in the previous column.

	RMSE	%	MAE	%	RMSPE	%	MAPE	%	MCS-p
GPF	2.100	0.000	0.648	0.000	0.367	0.000	0.186	0.000	0.000
GPF_G	1.673	2.083	0.528	0.000	0.297	12.500	0.152	0.000	0.000
GPF_F	1.624	16.667	0.483	10.417	0.303	5.208	0.141	1.042	0.001
GPF_{GF}	1.535	34.375	0.458	63.542	0.281	62.500	0.132	83.333	1.000
HBD	2.017	0.000	0.635	0.000	0.414	0.000	0.199	0.000	0.000
HBD_G	1.723	2.083	0.532	0.000	0.361	0.000	0.164	0.000	0.000
HBD_F	1.688	7.292	0.500	3.125	0.372	0.000	0.157	0.000	0.001
HBD_{GF}	1.622	14.583	0.477	16.667	0.341	0.000	0.146	0.000	0.004
AFFT	2.139	0.000	0.654	0.000	0.370	0.000	0.187	0.000	0.000
AFFT_G	1.733	0.000	0.536	0.000	0.303	7.292	0.152	0.000	0.000
AFFT_F	1.678	3.125	0.493	2.083	0.307	3.125	0.142	0.000	0.000
AFFT_{GF}	1.580	20.833	0.469	9.375	0.287	18.750	0.134	30.208	0.002
PARA	2.567	0.000	0.858	0.000	0.418	0.000	0.250	0.000	0.000

Focusing solely on RMSE and excluding Greeks and firm characteristics, HBD is the best with a 2.017 RMSE, followed by GPF (2.100), AFFT (2.139), and PARA (2.567). This order is the same

for MAE, confirming HBD's top accuracy. For MAPE, a relative error metric, GPF leads with 0.186, closely followed by AFFT (0.187), HBD (0.199), and PARA (0.250). The variation across these metrics underscores the importance of a comprehensive evaluation approach to understand each model's strengths and limitations. All machine learning models improve RMSE and MAPE by over 20% compared to PARA, indicating their effectiveness in reducing pricing errors.

Incorporating option Greeks significantly enhances all models: GPF_G, HBD_G, and AFFT_G, with GPF_Gshowing superior performance across all error metrics. Their RMSE values (1.673, 1.723, 1.733) and MAPE values (0.152, 0.164, 0.152) are considerably lower than those of GPF, HBD, and AFFT. The improvement is more notable in GPF_G, and AFFT_G, which focus on implied volatility, whereas HBD_G, which targets pricing residuals, shows lesser but still notable benefits.

Adding firm characteristics further reduces errors in GPF_F, HBD_F, and AFFT_F, with GPF_Fleading in all metrics. Their RMSEs (1.624, 1.688, 1.678) and MAPEs (0.141, 0.157, 0.142) also surpass their Greek-inclusive counterparts, except in RMSPE, indicating higher relative errors in lower-priced options.

Combining both firm characteristics and option Greeks further boosts model performance in GPF_F, HBD_F, and AFFT_F, with $\text{GPF}_F$with RMSEs 1.535, 1.622, 1.580, respectively. The overall best performance model is GPF_{GF}, which shows the smallest RMSE (1.535) and also stands out in MCS results.

The GPF structure's superior performance across various feature combinations stems from its structural differences with HBD and AFFT. Unlike the HBD structure's focus on predicting pricing residuals, and the AFFT structure's focus on predicting implied volatility residuals, GPF targets directly implied volatility prediction. The differences significantly influence the complexity and effectiveness of each model.

Predicting implied volatility, as done by GPF, is inherently less complex than predicting its residual, as in AFFT. Implied volatility is an extracted metric, that captures essential market expectations about option price. By directly predicting implied volatility, the GPF model utilizes a simpler, more direct metric that has filtered out many of the confounding factors in option prices. This approach balances the complexity of the model with the ability to capture the underlying market dynamics. In contrast, the AFFT model adds a layer of complexity by focusing on the

residuals of implied volatility. The residuals represent deviations from the expected values, which requires the model to first understand and predict the difference from the baseline. This additional complexity can create challenges for accurate modelling, especially when integrating additional features such as option Greeks and firm characteristics.

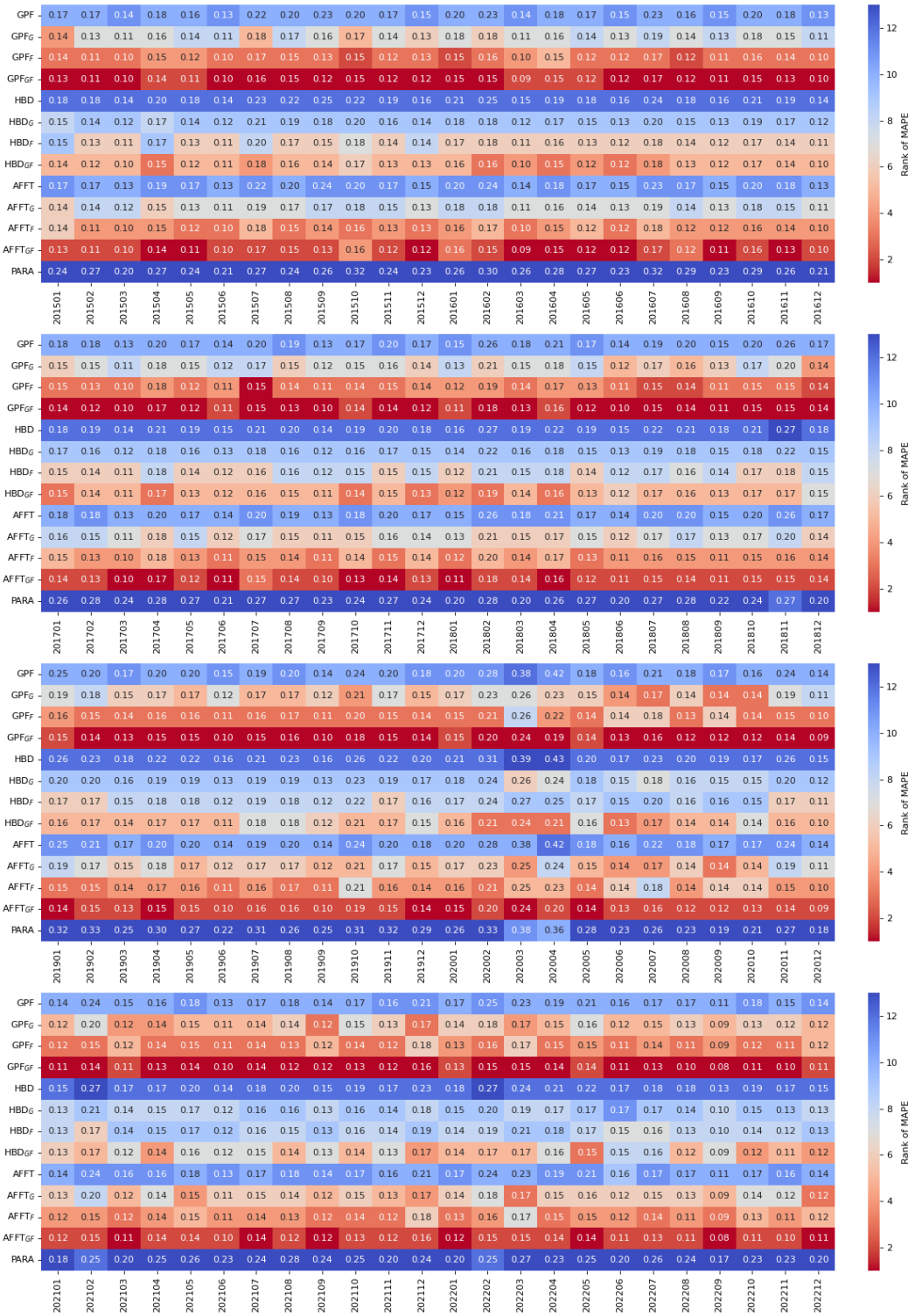
On the other hand, HBD models directly predict price residuals, which is intuitive but introduces more noise. Implied volatility prediction models in option pricing, such as GPF and AFFT, effectively reduce noise through the use of implied volatility, which is a metric extracted from parametric models that inherently takes into account different option characteristics such as moneyness, maturity, and option type (call/put). In contrast, a pricing residual model like HBD, which directly predicts price residuals, may introduce more noise due to the poor filtering nature of its methodology, which does not effectively separate these specific option attributes.

Overall, GPF's method of predicting implied volatility strikes a balance, being complex enough to incorporate additional valuable information, such as option Greeks and firm characteristics, without being overly complicated. Table 3.4 also illustrates the consistent superiority of the GPF_{GF} model over others, with its performance leading most of the time. Specifically, GPF_{GF} ranks first in approximately 34.38% of the months for RMSE, 63.54% for MAE, 62.5% for RMSPE, and an impressive 83.33% for MAPE. This dominance is observed across all error measures, indicating the overwhelming advantage of GPF_{GF} over competing models.

Figure 3.5 further shows the month-by-month MAPE performance for all models over the entire out-of-sample period. We choose to present MAPE rather than RMSE because option prices and their absolute errors tend to grow over time, making it challenging to fairly compare year-over-year performance using absolute error metrics. For other error measures, please check the Appendix. Over time, GPF_{GF} has been the dominant model, with the lowest MAPE in 80 out of 96 months, and maintaining the second lowest MAPE in the remaining 16 months. This was followed by AFFT_{GF}, which has recorded the best MAPE in 29 out of 96 months[5] and the second-best in 63 months. When considering other metrics such as RMSPE, RMSE and MAE, GPF_{GF} also continues to impress. This highlights GPF_{GF}'s ability to consistently provide stable and accurate pricing predictions.

[5] In some cases, GPF_{GF} and AFFT_{GF} share the same MAPE.

Figure 3.5. Monthly MAPEs and rankings are presented for each model. The legend indicates the corresponding colour for each model's rank in every month. Within each rectangle, the displayed number represents the MAPE for that specific month and model. The GPF_{GF} model consistently proved to be the most accurate, achieving the lowest MAPE in 80 out of 96 months and securing the second-lowest MAPE for the remaining 16 months.



3.4.1.2 The pricing performance before and during the Pandemic

In terms of overall performance, GPF_{GF} performs best over the entire forecasting period. In this section, we further check whether the GPF_{GF} model can continue to perform best over the Pandemic period when the market is facing great turbulence.

Table 3.5 reports the pricing performance before and during the Pandemic period. It's noteworthy that the GPF_{GF} consistently outperforms PARA and other models, especially during the Pandemic period from 2020 to 2022. Before and during the Pandemic period, the RMSEs for GPF_{GF} are 1.068 and 3.691, HBD_{GF} are 1.183 and 4.131, AFFT_{GF} are 1.136 and 3.907, PARA are 2.845 and 10.466, respectively. Regarding MAPE before and during the pandemic: GPF_{GF} is 0.133 and 0.131, HBD_{GF} is 0.146 and 0.147, AFFT_{GF} is 0.135 and 0.132, while PARA is 0.260 and 0.240.

Table 3.5. The pricing performance before and during the Pandemic. The table presents a comparison of various option pricing models' performance, specifically before and during the Pandemic, covering January 2015 to 2019.12 (Before) and January 2020 to December 2022 (During), respectively. Key performance metrics such as RMSE, MAE, RMSPE, and MAPE. The total number of observations in the study amounts to 9,302,680, with 4,731,816 of these recorded before the Pandemic and 4,570,864 during the Pandemic period.

	RMSE		MAE		RMSPE		MAPE	
	Before	During	Before	During	Before	During	Before	During
GPF	1.702	7.213	0.454	0.849	0.129	0.141	0.184	0.188
GPF_G	1.299	4.349	0.385	0.675	0.094	0.083	0.154	0.149
GPF_F	1.143	4.184	0.341	0.631	0.098	0.084	0.140	0.142
GPF_{GF}	1.068	3.691	0.328	0.591	0.087	0.070	0.133	0.131
HBD	1.624	6.601	0.451	0.825	0.164	0.179	0.196	0.202
HBD_G	1.316	4.677	0.389	0.680	0.130	0.131	0.166	0.163
HBD_F	1.230	4.528	0.355	0.651	0.129	0.149	0.154	0.160
HBD_{GF}	1.183	4.131	0.344	0.614	0.120	0.113	0.146	0.147
AFFT	1.837	7.408	0.461	0.854	0.132	0.142	0.185	0.189
AFFT_G	1.436	4.625	0.393	0.683	0.098	0.086	0.155	0.150
AFFT_F	1.233	4.456	0.349	0.643	0.102	0.087	0.142	0.143
AFFT_{GF}	1.136	3.907	0.338	0.604	0.091	0.074	0.135	0.132
PARA	2.845	10.466	0.634	1.091	0.182	0.167	0.260	0.240

The heightened RMSE and MAE observed in option pricing during the Pandemic are largely attributable to the increased volatility and unpredictability that characterized financial markets during this period. The Pandemic brought in a level of uncertainty not seen in previous times, leading to rapid and often drastic fluctuations in stock prices. These circumstances present a

significant challenge to predictive models used in option pricing. Typically, these models rely on historical data and established market trends to predict future prices. However, the Pandemic disrupted these trends, making historical data less reliable for predictions. As a result, the absolute pricing errors, as measured by RMSE and MAE, which quantify the magnitude of prediction errors without considering the direction, were notably higher. This reflects the difficulty models face in accurately predicting option prices in the face of sudden market changes and increased volatility.

Conversely, the relative error metrics like MAPE and RMSPE did not show a proportional increase during the Pandemic. This discrepancy arises from the intrinsic nature of these metrics as relative measures of error, which consider the size of the actual values when assessing errors. During periods of high volatility, like the Pandemic, the actual values of options can undergo significant changes. In such scenarios, even if the absolute errors (as captured by RMSE and MAE) are large, the relative errors might not increase to the same extent. This is because these relative errors are normalized by the actual values, which were also experiencing substantial fluctuations.

The enhanced pricing performance of machine learning-based models over the PARA model can largely be attributed to the size of their training datasets (The findings in robust test 3.4.4.2 also support this.). Machine learning models are particularly effective when they have access to extensive historical data, as this allows them to identify and learn from underlying patterns more efficiently. A larger dataset not only reduces the risk of overfitting - a common issue where a model learns the details and noise in the training data to the extent that it negatively impacts the performance of the model on new data - but also significantly improves the model's ability to make accurate predictions. This abundance of data is a key factor in the superior performance of machine learning models in option pricing compared to the PARA model, which may not utilize as extensive a dataset for its predictions.

3.4.2 Pricing performance for different groups

3.4.2.1 Call vs put options

Table 3.6 presents the models' pricing performance separately for call options (as shown in Panel A) and put options (as shown in Panel B). The results show that the variation in pricing errors

between these option types is insubstantial. The GPF_{GF} consistently emerges as the top performer, followed by AFFT_{GF} and then HBD_{GF}. For call options, the RMSEs of GPF_{GF}, HBD_{GF}, and AFFT_{GF} are 1.541, 1.642, and 1.584, respectively. For put options, the RMSEs of GPF_{GF}, HBD_{GF}, and AFFT_{GF} are 1.526, 1.592, and 1.574, respectively. These RMSEs are well below the RMSEs of PARA, which are 2.587 for calls and 2.538 for puts.

The superiority of GPF_{GF}, HBD_{GF}, and AFFT_{GF} is also evident across other measures, including MAE, RMSE, and MAPE. Notably, GPF_{GF} dominates in all error metrics. These findings emphasize the potential of machine learning to improve the accuracy of call and put option pricing. This accuracy can be further improved if firm characteristics and option Greeks are integrated into the model.

3.4.2.2 Different moneyness options

The well-established observation that implied volatility varies with moneyness, a phenomenon referred to as the volatility smile, is pivotal in recognizing that moneyness significantly influences option pricing performance. Specifically, ITM options may necessitate more significant price alterations than OTM options. This is because ITM options are more prone to be exercised before their expiration, while OTM options are often retained until maturity. To provide insight into how moneyness affects option pricing, we evaluate the pricing performance of various models across different moneyness groups.

As presented in Table 3.7, the results detail the pricing errors associated with each moneyness group[6]. Among all groups, GPF_{GF} consistently emerges as the superior performer, recording the lowest values in RMSE, MAE, RMSPE, and MAPE. Specifically, for GPF_{GF}, the RMSE is 1.312 for DOTM options, 1.413 for OTM options, 1.590 for JOTM options, 1.791 for ATM options, 1.607 for JITM options, 1.330 for ITM options, and 1.240 for DITM options. RMSE of 1.240 for DITM options. What is particularly striking is that the benchmark PARA model has a large margin of error across all moneyness groups. In contrast, the proposed model maintains consistent performance regardless of moneyness. An important observation is that Greek letters and firm

[6] DOTM stands for deep out-the-money; OTM stands for out-the-money; JOTM stands for just out-the-money; ATM stands for at-the-money; JITM stands for just in-the-money; ITM stands for in-the-money; DITM stands for deep in-the-money. The specific definition of the moneyenss can be found in Table 3.1.

Table 3.6. Pricing performance for call and put options. This table shows the out-of-sample pricing errors for each model for the period January 2015 to December 2022 for call and put options. RMSE, MAE, RMSPE, and MAPE respectively refer to the root mean squared error, the mean absolute error, the root mean squared percentage error, and the mean absolute percentage error. MCS-p refers to the p-value of the model confidence set test. Columns labelled "%" indicate the proportion of months during the entire forecast period in which each model ranked first according to the error metric specified in the previous column.

	RMSE	%	MAE	%	RMSPE	%	MAPE	%	MCS-p
Panel A. Call options (5,520,239 observations)									
GPF	2.112	0.000	0.644	0.000	0.379	0.000	0.189	0.000	0.000
GPF_G	1.685	1.042	0.525	1.042	0.305	9.375	0.154	0.000	0.000
GPF_F	1.625	15.625	0.475	7.292	0.306	3.125	0.141	0.000	0.019
GPF_{GF}	1.541	37.500	0.450	68.750	0.285	60.417	0.132	83.333	1.000
HBD	2.023	0.000	0.632	0.000	0.432	0.000	0.204	0.000	0.000
HBD_G	1.747	0.000	0.532	0.000	0.373	0.000	0.169	0.000	0.000
HBD_F	1.698	5.208	0.496	2.083	0.383	0.000	0.159	0.000	0.019
HBD_{GF}	1.642	13.542	0.473	9.375	0.354	0.000	0.149	0.000	0.019
AFFT	2.156	0.000	0.651	0.000	0.382	0.000	0.189	0.000	0.000
AFFT_G	1.753	0.000	0.533	0.000	0.312	8.333	0.155	0.000	0.000
AFFT_F	1.681	4.167	0.485	1.042	0.312	4.167	0.142	0.000	0.003
AFFT_{GF}	1.584	23.958	0.461	17.708	0.293	20.833	0.134	31.250	0.019
PARA	2.587	0.000	0.885	0.000	0.471	0.000	0.276	0.000	0.000
Panel B. Put options (3,782,441 observations)									
GPF	2.082	0.000	0.654	0.000	0.349	0.000	0.182	0.000	0.000
GPF_G	1.654	3.125	0.531	0.000	0.286	13.542	0.149	0.000	0.001
GPF_F	1.622	14.583	0.495	8.333	0.298	4.167	0.141	1.042	0.002
GPF_{GF}	1.526	35.417	0.468	54.167	0.275	47.917	0.132	78.125	1.000
HBD	2.009	0.000	0.640	0.000	0.386	0.000	0.193	0.000	0.000
HBD_G	1.686	4.167	0.531	0.000	0.343	1.042	0.158	0.000	0.000
HBD_F	1.675	5.208	0.507	2.083	0.356	0.000	0.154	0.000	0.004
HBD_{GF}	1.592	16.667	0.482	26.042	0.322	0.000	0.143	3.125	0.057
AFFT	2.113	0.000	0.660	0.000	0.352	0.000	0.183	0.000	0.000
AFFT_G	1.702	0.000	0.538	0.000	0.290	9.375	0.149	0.000	0.000
AFFT_F	1.674	4.167	0.505	1.042	0.301	0.000	0.142	0.000	0.000
AFFT_{GF}	1.574	18.750	0.479	13.542	0.279	20.833	0.134	34.375	0.045
PARA	2.538	0.000	0.820	0.000	0.325	12.500	0.213	0.000	0.000

characteristics significantly improve performance when pricing options with deeper moneyness. Such outcomes imply that option Greeks and firm characteristics can provide insights into the volatility smile to a certain degree.

3.4.2.3 Different maturity options

Table 3.8 summarizes the performance of various pricing models in different maturity groups. Among them, GPF_{GF} performs the best in all maturity groups, taking into account the metrics RMSE, MAE, RMSPE and MAPE. It is closely followed by AFFT_{GF} and HBD_{GF}.

The RMSEs of the PARA model increase with time to maturity, where the RMSE jumps from 2.204 for the near-term options to 2.504 for the mid-term options, and then rises sharply to 4.359 for the long-term options. This is in contrast to the performance of the proposed model, which is stable across maturity groups. For example, the RMSEs for GPF_{GF} are 1.604 (near-term), 1.242 (mid-term), and 1.598 (long-term), respectively. On examining the MAPE, a relative error measure, a decline is observed with increasing time to maturity for PARA, showcasing values of 0.263 (near-term), 0.222 (mid-term), and 0.217 (long-term). This decreasing trend is even more pronounced for the proposed models. Specifically, the MAPEs for GPF_{GF} are 0.161 (near-term), 0.076 (mid-term), and 0.052 (long-term), respectively. One potential explanation for this is that with time to maturity increase, the time value of the option increases causing the absolute value of the option to increase. This could cause the absolute measure, like the RMSE, to increase with time, and cause the relative measure, like MAPE, to decrease with time.

The results also suggest that firm characteristics are more indicative of long-term options, while option Greek letters are more helpful for short-term options. This can be seen in the difference in improvement between GPF and GPF_G and between GPF and GPF_F in terms of short-term performance and long-term performance. For short-term options, the RMSEs for GPF, GPF_G, and GPF_F are 2.133, 1.696, and 1.707, respectively; for long-term options, the RMSEs for GPF, GPF_G, and GPF_F are 2.400, 1.916, and 1.645, respectively. Regarding the MAPE, for short-term options, the MAPEs for GPF, GPF_G, and GPF_F are 0.218, 0.178, and 0.172, respectively; for long-term options, the MAPEs for GPF, GPF_G, and GPF_F are 0.093, 0.074, and 0.053, respectively.

Table 3.7. Pricing performance for different moneyness options. This table shows the out-of-sample pricing errors for each model for the period January 2015 to December 2022 for different moneyness options. RMSE, MAE, RMSPE, and MAPE respectively refer to the root mean squared error, the mean absolute error, the root mean squared percentage error, and the mean absolute percentage error. MCS-p refers to the p-value of the model confidence set test. Columns labelled "%" indicate the proportion of months during the entire forecast period in which each model ranked first according to the error metric specified in the previous column.

	RMSE	%	MAE	%	RMSPE	%	MAPE	%	MCS-p	RMSE	%	MAE	%	RMSPE	%	MAPE	%	MCS-p	RMSE	%	MAE	%	RMSPE	%	MAPE	%	MCS-p
	Panel A. Deep out-the-money (1,401,839 observations)									**Panel B. Out-the-money (1,785,653 observations)**									**Panel C. Just out-the-money (2,649,723 observations)**								
GPF	1.873	0.000	0.635	0.000	0.522	0.000	0.267	0.000	0.000	1.966	0.000	0.633	0.000	0.484	0.000	0.255	0.000	0.000	2.157	0.000	0.649	0.000	0.370	1.042	0.210	0.000	0.000
GPF_G	1.481	1.042	0.507	1.042	0.400	5.208	0.212	0.000	0.000	1.545	2.083	0.511	0.000	0.394	13.542	0.207	0.000	0.000	1.724	2.083	0.531	0.000	0.312	15.625	0.174	1.042	0.001
GPF_F	1.382	20.833	0.445	17.708	0.398	8.333	0.191	7.292	0.010	1.487	17.708	0.462	11.458	0.405	7.292	0.191	2.083	0.013	1.689	16.667	0.493	10.417	0.321	3.125	0.164	1.042	0.001
GPF_{GF}	1.312	43.750	0.423	62.500	0.368	55.208	0.180	70.833	1.000	1.413	37.500	0.437	66.667	0.374	50.000	0.180	82.292	1.000	1.590	36.458	0.465	53.125	0.299	53.125	0.154	79.167	1.000
HBD	1.814	0.000	0.649	0.000	0.657	0.000	0.317	0.000	0.000	1.897	0.000	0.628	0.000	0.534	0.000	0.276	0.000	0.000	2.089	0.000	0.631	0.000	0.382	0.000	0.216	0.000	0.000
HBD_G	1.601	0.000	0.539	0.000	0.585	0.000	0.257	0.000	0.000	1.618	0.000	0.521	0.000	0.461	0.000	0.227	0.000	0.000	1.767	2.083	0.529	0.000	0.329	1.042	0.179	0.000	0.000
HBD_F	1.528	0.000	0.495	0.000	0.576	0.000	0.240	0.000	0.000	1.578	4.167	0.487	2.083	0.493	0.000	0.215	0.000	0.002	1.736	5.208	0.501	4.167	0.345	1.042	0.173	0.000	0.002
HBD_{GF}	1.491	7.292	0.475	1.042	0.541	0.000	0.226	0.000	0.001	1.508	10.417	0.462	6.250	0.439	1.042	0.201	1.042	0.013	1.665	20.833	0.476	28.125	0.318	5.208	0.161	5.208	0.003
AFFT	1.880	0.000	0.640	0.000	0.528	0.000	0.268	0.000	0.000	2.009	0.000	0.640	0.000	0.489	1.042	0.256	0.000	0.000	2.203	0.000	0.656	0.000	0.371	0.000	0.211	0.000	0.000
AFFT_G	1.518	0.000	0.513	0.000	0.417	1.042	0.214	0.000	0.000	1.610	0.000	0.519	0.000	0.401	6.250	0.208	0.000	0.000	1.789	0.000	0.539	0.000	0.313	10.417	0.174	0.000	0.000
AFFT_F	1.411	8.333	0.454	2.083	0.409	4.167	0.194	2.083	0.010	1.551	2.083	0.473	2.083	0.411	2.083	0.194	1.042	0.002	1.746	2.083	0.503	1.042	0.324	1.042	0.165	0.000	0.001
AFFT_{GF}	1.336	20.833	0.434	18.750	0.384	29.167	0.183	29.167	0.252	1.459	26.042	0.448	19.792	0.382	22.917	0.182	30.208	0.009	1.642	14.583	0.476	12.500	0.303	19.792	0.155	31.250	0.002
PARA	2.790	0.000	0.995	0.000	0.605	1.042	0.403	0.000	0.000	2.383	0.000	0.840	0.000	0.523	1.042	0.338	0.000	0.000	2.464	0.000	0.835	0.000	0.435	0.000	0.277	0.000	0.000
	Panel D. At-the-money (1,531,414 observations)									**Panel E. Just in-the-money (1,199,359 observations)**									textbfPanel F. In-the-money (457,057 observations)								
GPF	2.442	0.000	0.710	0.000	0.213	0.000	0.139	0.000	0.000	2.128	0.000	0.649	0.000	0.121	0.000	0.082	0.000	0.000	1.797	0.000	0.582	0.000	0.072	0.000	0.048	0.000	0.000
GPF_G	1.943	2.083	0.584	1.042	0.180	7.292	0.115	1.042	0.000	1.716	2.083	0.533	0.000	0.100	2.083	0.068	0.000	0.002	1.454	1.042	0.476	0.000	0.057	0.000	0.039	0.000	0.002
GPF_F	1.909	14.583	0.547	7.292	0.183	0.000	0.109	0.000	0.000	1.692	17.708	0.495	6.250	0.099	2.083	0.063	0.000	0.002	1.367	15.625	0.428	13.542	0.054	16.667	0.035	18.750	0.375
GPF_{GF}	1.791	27.083	0.517	36.458	0.171	63.542	0.103	82.292	1.000	1.607	28.125	0.469	57.292	0.093	84.375	0.060	90.625	1.000	1.330	30.208	0.410	54.167	0.051	82.292	0.033	89.583	1.000
HBD	2.346	0.000	0.680	0.000	0.214	0.000	0.139	0.000	0.000	2.021	0.000	0.625	0.000	0.123	0.000	0.084	0.000	0.000	1.646	0.000	0.564	0.000	0.077	0.000	0.050	0.000	0.000
HBD_G	1.978	7.292	0.576	0.000	0.183	2.083	0.117	0.000	0.000	1.722	4.167	0.526	0.000	0.105	0.000	0.070	0.000	0.000	1.446	2.083	0.474	0.000	0.065	0.000	0.041	0.000	0.000
HBD_F	1.952	8.333	0.549	4.167	0.193	0.000	0.114	0.000	0.002	1.717	8.333	0.502	5.208	0.111	0.000	0.067	0.000	0.005	1.424	8.333	0.446	3.125	0.067	0.000	0.039	0.000	0.156
HBD_{GF}	1.870	23.958	0.523	45.833	0.177	15.625	0.106	14.583	0.019	1.641	20.833	0.476	32.292	0.100	6.250	0.063	6.250	0.280	1.373	12.500	0.423	13.542	0.060	2.083	0.036	3.125	0.375
AFFT	2.493	0.000	0.717	0.000	0.214	0.000	0.140	0.000	0.000	2.183	0.000	0.656	0.000	0.122	0.000	0.083	0.000	0.000	1.805	1.042	0.586	0.000	0.072	0.000	0.048	0.000	0.000
AFFT_G	2.003	1.042	0.593	0.000	0.180	4.167	0.116	0.000	0.000	1.788	0.000	0.541	0.000	0.101	0.000	0.068	0.000	0.000	1.530	1.042	0.484	0.000	0.057	0.000	0.039	0.000	0.000
AFFT_F	1.974	6.250	0.559	1.042	0.184	0.000	0.110	0.000	0.000	1.749	7.292	0.506	1.042	0.100	0.000	0.064	0.000	0.000	1.413	7.292	0.437	3.125	0.054	5.208	0.035	7.292	0.046
AFFT_{GF}	1.841	11.458	0.529	6.250	0.173	22.917	0.104	35.417	0.001	1.660	13.542	0.482	6.250	0.095	30.208	0.060	42.708	0.005	1.361	21.875	0.420	17.708	0.052	45.833	0.033	57.292	0.375
PARA	2.771	0.000	0.914	0.000	0.252	0.000	0.178	0.000	0.000	2.509	0.000	0.781	0.000	0.132	0.000	0.096	0.000	0.000	2.330	0.000	0.709	0.000	0.084	0.000	0.056	0.000	0.000
	Panel G. Deep in-the-money (277,635 observations)																										
GPF	1.733	0.000	0.562	0.000	0.051	0.000	0.032	0.000	0.000																		
GPF_G	1.385	2.083	0.456	0.000	0.040	0.000	0.026	0.000	0.001																		
GPF_F	1.310	19.792	0.402	18.750	0.036	35.417	0.022	40.625	0.289																		
GPF_{GF}	1.240	26.042	0.387	48.958	0.034	69.792	0.021	86.458	1.000																		
HBD	1.571	0.000	0.553	0.000	0.058	0.000	0.034	0.000	0.000																		
HBD_G	1.407	0.000	0.463	0.000	0.048	0.000	0.028	0.000	0.001																		
HBD_F	1.370	6.250	0.427	3.125	0.048	0.000	0.026	0.000	0.107																		
HBD_{GF}	1.344	11.458	0.412	4.167	0.045	2.083	0.025	0.000	0.289																		
AFFT	1.714	0.000	0.562	0.000	0.051	0.000	0.032	0.000	0.000																		
AFFT_G	1.398	1.042	0.458	0.000	0.040	0.000	0.026	0.000	0.000																		
AFFT_F	1.322	15.625	0.405	12.500	0.036	20.833	0.022	34.375	0.289																		
AFFT_{GF}	1.262	18.750	0.391	28.125	0.035	46.875	0.022	73.958	0.498																		
PARA	2.924	0.000	0.789	0.000	0.069	0.000	0.042	0.000	0.000																		

Table 3.8. Pricing performance for different maturity options. This table shows the out-of-sample pricing errors for each model for the period January 2015 to December 2022 for different maturity options. RMSE, MAE, RMSPE, and MAPE respectively refer to the root mean squared error, the mean absolute error, the root mean squared percentage error, and the mean absolute percentage error. MCS-p refers to the p-value of the model confidence set test. Columns labelled "%" indicate the proportion of months during the entire forecast period in which each model ranked first according to the error metric specified in the previous column.

	RMSE	%	MAE	%	RMSPE	%	MAPE	%	MCS-p
Panel A. Near-term (6,434,005 observations)									
GPF	2.133	1.042	0.647	0.000	0.414	0.000	0.218	0.000	0.000
GPF_G	1.696	3.125	0.530	1.042	0.336	19.792	0.178	1.042	0.005
GPF_F	1.707	13.542	0.513	8.333	0.352	4.167	0.172	0.000	0.001
GPF_{GF}	1.604	31.250	0.482	51.042	0.326	51.042	0.161	76.042	1.000
HBD	2.108	0.000	0.646	0.000	0.471	0.000	0.235	0.000	0.000
HBD_G	1.766	4.167	0.541	0.000	0.413	0.000	0.194	0.000	0.000
HBD_F	1.759	6.250	0.527	5.208	0.431	0.000	0.191	0.000	0.003
HBD_{GF}	1.672	21.875	0.498	21.875	0.395	0.000	0.177	1.042	0.019
AFFT	2.192	0.000	0.654	0.000	0.418	0.000	0.219	0.000	0.000
AFFT_G	1.760	2.083	0.539	0.000	0.343	9.375	0.179	0.000	0.000
AFFT_F	1.766	2.083	0.522	2.083	0.357	2.083	0.174	0.000	0.000
AFFT_{GF}	1.648	16.667	0.491	17.708	0.333	13.542	0.162	34.375	0.005
PARA	2.204	0.000	0.764	0.000	0.424	4.167	0.263	0.000	0.000
Panel B. Mid-term (1,922,515 observations)									
GPF	1.811	0.000	0.580	0.000	0.222	0.000	0.126	0.000	0.000
GPF_G	1.449	1.042	0.466	0.000	0.177	0.000	0.101	0.000	0.000
GPF_F	1.296	16.667	0.383	16.667	0.141	15.625	0.079	21.875	0.018
GPF_{GF}	1.242	39.583	0.368	56.250	0.136	69.792	0.076	71.875	1.000
HBD	1.663	0.000	0.556	0.000	0.242	0.000	0.133	0.000	0.000
HBD_G	1.466	0.000	0.460	0.000	0.199	0.000	0.108	0.000	0.000
HBD_F	1.371	7.292	0.404	2.083	0.186	0.000	0.092	0.000	0.017
HBD_{GF}	1.347	19.792	0.389	15.625	0.170	0.000	0.087	0.000	0.018
AFFT	1.843	0.000	0.586	0.000	0.222	0.000	0.126	0.000	0.000
AFFT_G	1.517	0.000	0.473	0.000	0.178	0.000	0.101	0.000	0.000
AFFT_F	1.337	6.250	0.394	3.125	0.145	3.125	0.081	5.208	0.003
AFFT_{GF}	1.288	10.417	0.381	9.375	0.140	31.250	0.078	31.250	0.018
PARA	2.504	0.000	0.846	0.000	0.382	0.000	0.222	0.000	0.000
Panel C. Long-term (946,160 observations)									
GPF	2.400	0.000	0.795	0.000	0.243	0.000	0.093	0.000	0.000
GPF_G	1.916	1.042	0.634	0.000	0.192	1.042	0.074	0.000	0.000
GPF_F	1.645	25.000	0.483	23.958	0.139	30.208	0.053	42.708	0.292
GPF_{GF}	1.598	31.250	0.471	54.167	0.135	37.500	0.052	76.042	1.000
HBD	2.044	0.000	0.724	0.000	0.242	0.000	0.094	0.000	0.000
HBD_G	1.899	1.042	0.616	0.000	0.206	0.000	0.077	0.000	0.000
HBD_F	1.783	2.083	0.517	2.083	0.160	1.042	0.061	0.000	0.014
HBD_{GF}	1.784	9.375	0.509	9.375	0.159	2.083	0.059	0.000	0.031
AFFT	2.326	0.000	0.794	0.000	0.243	0.000	0.093	0.000	0.000
AFFT_G	1.946	0.000	0.638	0.000	0.196	0.000	0.074	0.000	0.000
AFFT_F	1.688	11.458	0.499	2.083	0.137	15.625	0.054	12.500	0.091
AFFT_{GF}	1.651	18.750	0.490	11.458	0.133	35.417	0.053	29.167	0.292
PARA	4.359	0.000	1.528	0.000	0.447	0.000	0.217	0.000	0.000

3.4.2.4 Pricing performance for the stocks with different returns

In this section, we will further analyze how different return profiles of the underlying stock affect the pricing of the underlying option. To achieve this, we categorize stocks into 10 groups based on their cross-sectional daily return distributions. Each group represents 10 percentile stock returns, ranging from the lowest returns (0-10th percentile) to the highest returns (90-100th percentile). This categorization allows us to analyze and compare the pricing performance of options linked to stocks with different return percentiles.

We report the MAPEs of the models for stocks with different return profiles in Table 3.9, 3.10, 3.11, and 3.12. Table 3.9 provides a report for all options; Table 3.10 categorizes the data by call and put options; Table 3.11 groups the performance based on moneyness, and Table 3.12 presents the results organized by maturity.

Table 3.9. Pricing performance for all options with different stock returns (MAPE). This table presents the out-of-sample pricing errors of each model for all options. The out-of-sample period is from January 2015 to December 2022. The columns represent groups categorized according to percentile returns, ranging from the lowest 10 percentile (10) to the highest 10 percentile (100). The observations (Obs) are given in thousands.

	10	20	30	40	50	60	70	80	90	100
GPF	0.212	0.187	0.179	0.181	0.177	0.177	0.177	0.181	0.182	0.210
GPF_G	0.180	0.151	0.146	0.146	0.142	0.142	0.143	0.145	0.147	0.177
GPF_F	0.175	0.140	0.133	0.134	0.129	0.129	0.129	0.131	0.135	0.174
GPF_{GF}	0.166	0.131	0.125	0.124	0.121	0.121	0.120	0.123	0.127	0.166
HBD	0.227	0.200	0.194	0.194	0.189	0.189	0.190	0.193	0.196	0.223
HBD_G	0.197	0.165	0.158	0.158	0.155	0.153	0.153	0.157	0.162	0.189
HBD_F	0.194	0.157	0.149	0.150	0.146	0.144	0.145	0.146	0.152	0.190
HBD_{GF}	0.185	0.145	0.138	0.139	0.135	0.134	0.133	0.137	0.142	0.178
AFFT	0.213	0.188	0.180	0.181	0.177	0.177	0.178	0.182	0.183	0.211
AFFT_G	0.182	0.152	0.146	0.147	0.143	0.142	0.143	0.145	0.148	0.179
AFFT_F	0.178	0.141	0.135	0.135	0.130	0.130	0.130	0.133	0.136	0.177
AFFT_{GF}	0.169	0.132	0.126	0.127	0.122	0.122	0.122	0.124	0.128	0.169
PARA	0.243	0.244	0.252	0.261	0.259	0.258	0.256	0.252	0.240	0.238
Obs	935,370	936,367	936,142	936,386	936,545	935,984	936,179	936,349	936,160	937,199

In Table 3.9, it's evident that the GPF_{GF} model demonstrates the best overall performance across all stock return profiles, followed by AFFT_{GF} and HBD_{GF}. When comparing models that incorporate either firm characteristics, option Greeks, or neither, we find that models based on firm characteristics generally outperform those based on option Greeks. They all outperform models that include neither firm characteristics nor option Greek letters. In addition, all of these models

Table 3.10. Pricing performance for call and put options with different stock returns (MAPE). This table presents the out-of-sample pricing errors of each model for the call and put options. The out-of-sample period is from January 2015 to December 2022. The columns represent groups categorized according to percentile returns, ranging from the lowest 10 percentile (10) to the highest 10 percentile (100). The observations (Obs) are given in thousands.

	10	20	30	40	50	60	70	80	90	100
Panel A. Call options										
GPF	0.228	0.196	0.188	0.187	0.182	0.179	0.178	0.179	0.178	0.197
GPF_G	0.194	0.158	0.152	0.151	0.146	0.144	0.144	0.143	0.143	0.166
GPF_F	0.186	0.144	0.137	0.136	0.131	0.129	0.129	0.129	0.131	0.159
GPF_{GF}	0.176	0.136	0.128	0.127	0.122	0.121	0.120	0.121	0.122	0.152
HBD	0.250	0.213	0.205	0.203	0.196	0.193	0.193	0.192	0.192	0.209
HBD_G	0.216	0.175	0.167	0.166	0.161	0.158	0.156	0.157	0.158	0.178
HBD_F	0.211	0.165	0.156	0.156	0.150	0.146	0.146	0.145	0.147	0.175
HBD_{GF}	0.202	0.153	0.145	0.145	0.140	0.136	0.134	0.135	0.138	0.165
AFFT	0.229	0.197	0.188	0.188	0.182	0.180	0.179	0.180	0.179	0.198
$AFFT_G$	0.195	0.159	0.153	0.152	0.147	0.144	0.144	0.144	0.145	0.167
$AFFT_F$	0.189	0.146	0.139	0.138	0.132	0.130	0.130	0.130	0.132	0.162
$AFFT_{GF}$	0.180	0.137	0.130	0.129	0.124	0.122	0.122	0.122	0.124	0.155
PARA	0.277	0.276	0.285	0.295	0.291	0.286	0.281	0.273	0.254	0.242
Obs	516,912	530,848	538,130	543,127	552,178	556,869	563,053	572,151	583,644	599,040
Panel B. Put options										
GPF	0.191	0.174	0.168	0.173	0.170	0.173	0.175	0.184	0.189	0.233
GPF_G	0.164	0.142	0.137	0.139	0.137	0.139	0.140	0.147	0.152	0.196
GPF_F	0.162	0.133	0.128	0.130	0.126	0.129	0.130	0.136	0.143	0.200
GPF_{GF}	0.153	0.125	0.120	0.121	0.118	0.121	0.121	0.128	0.134	0.190
HBD	0.200	0.183	0.178	0.182	0.179	0.183	0.185	0.195	0.202	0.248
HBD_G	0.173	0.151	0.145	0.147	0.145	0.147	0.147	0.156	0.168	0.208
HBD_F	0.173	0.146	0.139	0.142	0.140	0.141	0.142	0.149	0.159	0.216
HBD_{GF}	0.165	0.135	0.129	0.130	0.129	0.130	0.131	0.139	0.149	0.200
AFFT	0.192	0.175	0.168	0.173	0.170	0.174	0.175	0.185	0.190	0.234
$AFFT_G$	0.165	0.143	0.138	0.140	0.137	0.139	0.140	0.146	0.153	0.199
$AFFT_F$	0.164	0.135	0.129	0.131	0.127	0.130	0.131	0.137	0.144	0.204
$AFFT_{GF}$	0.156	0.126	0.121	0.123	0.119	0.121	0.122	0.128	0.135	0.193
PARA	0.200	0.201	0.207	0.214	0.215	0.217	0.217	0.220	0.217	0.229
Obs	418,458	405,519	398,012	393,259	384,367	379,115	373,126	364,198	352,516	338,159

Table 3.11. Pricing performance for different moneyness options with different stock returns (MAPE). This table presents the out-of-sample pricing errors of each model for different moneyness options. The out-of-sample period is from January 2015 to December 2022. The columns represent groups categorized according to percentile returns, ranging from the lowest 10 percentile (10) to the highest 10 percentile (100). The observations (Obs) are given in thousands.

	10	20	30	40	50	60	70	80	90	100	10	20	30	40	50	60	70	80	90	100	10	20	30	40	50	60	70	80	90	100
	Panel A. Deep out-the-money										**Panel B. Out-the-money**										**Panel C. Just out-the-money**									
GPF	0.302	0.260	0.249	0.249	0.248	0.250	0.252	0.256	0.263	0.322	0.289	0.254	0.242	0.246	0.240	0.241	0.241	0.250	0.255	0.297	0.231	0.209	0.205	0.208	0.202	0.202	0.203	0.208	0.208	0.236
GPF_G	0.249	0.204	0.197	0.197	0.194	0.195	0.198	0.201	0.208	0.260	0.247	0.206	0.196	0.196	0.192	0.193	0.193	0.199	0.204	0.249	0.203	0.172	0.168	0.169	0.165	0.164	0.166	0.167	0.169	0.205
GPF_F	0.230	0.182	0.175	0.174	0.170	0.174	0.176	0.177	0.186	0.247	0.240	0.188	0.178	0.179	0.172	0.175	0.174	0.179	0.186	0.245	0.203	0.162	0.156	0.157	0.151	0.151	0.152	0.155	0.159	0.208
GPF_{GF}	0.218	0.172	0.165	0.163	0.161	0.164	0.164	0.168	0.174	0.235	0.226	0.177	0.166	0.166	0.161	0.164	0.162	0.168	0.175	0.233	0.192	0.152	0.146	0.146	0.141	0.141	0.141	0.144	0.149	0.199
HBD	0.357	0.308	0.303	0.300	0.293	0.296	0.300	0.304	0.316	0.372	0.309	0.275	0.264	0.267	0.261	0.262	0.262	0.270	0.276	0.316	0.236	0.215	0.210	0.213	0.208	0.208	0.209	0.212	0.213	0.239
HBD_G	0.301	0.248	0.239	0.240	0.236	0.237	0.237	0.244	0.262	0.303	0.268	0.225	0.216	0.216	0.211	0.211	0.210	0.219	0.225	0.266	0.208	0.179	0.173	0.175	0.171	0.170	0.170	0.173	0.176	0.209
HBD_F	0.285	0.231	0.221	0.223	0.220	0.218	0.224	0.222	0.241	0.297	0.266	0.212	0.201	0.204	0.198	0.196	0.199	0.202	0.210	0.267	0.212	0.172	0.165	0.168	0.163	0.160	0.161	0.164	0.168	0.216
HBD_{GF}	0.275	0.215	0.207	0.208	0.208	0.205	0.204	0.209	0.229	0.275	0.252	0.197	0.188	0.187	0.182	0.183	0.181	0.190	0.197	0.250	0.201	0.160	0.153	0.154	0.150	0.148	0.149	0.152	0.156	0.203
AFFT	0.306	0.261	0.250	0.249	0.247	0.250	0.254	0.257	0.264	0.324	0.291	0.255	0.243	0.246	0.240	0.242	0.242	0.251	0.256	0.299	0.232	0.210	0.205	0.209	0.203	0.203	0.205	0.209	0.210	0.237
$AFFT_G$	0.251	0.205	0.199	0.199	0.196	0.195	0.197	0.202	0.211	0.267	0.249	0.206	0.197	0.197	0.193	0.193	0.194	0.200	0.206	0.252	0.204	0.173	0.169	0.169	0.166	0.164	0.166	0.167	0.170	0.206
$AFFT_F$	0.235	0.185	0.178	0.176	0.173	0.174	0.177	0.179	0.188	0.255	0.245	0.190	0.181	0.180	0.173	0.175	0.175	0.180	0.187	0.250	0.205	0.164	0.157	0.159	0.152	0.152	0.153	0.156	0.160	0.212
$AFFT_{GF}$	0.224	0.174	0.167	0.166	0.163	0.164	0.167	0.170	0.178	0.241	0.231	0.178	0.168	0.168	0.163	0.164	0.164	0.170	0.177	0.238	0.195	0.152	0.147	0.149	0.143	0.142	0.143	0.145	0.150	0.201
PARA	0.388	0.386	0.400	0.415	0.415	0.416	0.415	0.411	0.394	0.397	0.325	0.326	0.335	0.350	0.347	0.349	0.344	0.344	0.330	0.328	0.254	0.265	0.280	0.291	0.291	0.288	0.287	0.281	0.267	0.262
Obs	169	149	139	135	132	129	131	132	139	155	188	185	183	180	179	177	177	177	178	175	244	266	275	278	279	281	276	273	261	237
	Panel D. At-the-money										**Panel E. Just in-the-money**										**Panel F. In-the-money**									
GPF	0.155	0.139	0.134	0.137	0.134	0.133	0.134	0.138	0.138	0.157	0.098	0.083	0.079	0.078	0.076	0.076	0.077	0.079	0.081	0.096	0.060	0.048	0.043	0.042	0.041	0.042	0.042	0.044	0.046	0.058
GPF_G	0.135	0.115	0.111	0.112	0.111	0.109	0.110	0.112	0.112	0.136	0.084	0.068	0.064	0.064	0.062	0.062	0.063	0.063	0.066	0.082	0.049	0.038	0.035	0.035	0.034	0.034	0.034	0.035	0.037	0.047
GPF_F	0.135	0.110	0.104	0.106	0.102	0.102	0.101	0.104	0.106	0.136	0.082	0.063	0.059	0.058	0.056	0.056	0.057	0.058	0.061	0.080	0.046	0.034	0.031	0.031	0.030	0.030	0.030	0.031	0.033	0.044
GPF_{GF}	0.128	0.102	0.097	0.098	0.096	0.095	0.095	0.097	0.099	0.130	0.078	0.060	0.056	0.055	0.053	0.053	0.053	0.054	0.057	0.076	0.044	0.032	0.030	0.029	0.028	0.028	0.028	0.029	0.032	0.042
HBD	0.156	0.139	0.134	0.137	0.134	0.134	0.134	0.138	0.137	0.157	0.099	0.084	0.080	0.079	0.077	0.078	0.078	0.080	0.082	0.097	0.062	0.050	0.046	0.045	0.043	0.044	0.044	0.046	0.048	0.060
HBD_G	0.137	0.117	0.112	0.114	0.112	0.111	0.111	0.113	0.114	0.137	0.086	0.070	0.066	0.065	0.064	0.063	0.063	0.065	0.068	0.084	0.052	0.041	0.037	0.037	0.036	0.035	0.035	0.037	0.040	0.050
HBD_F	0.140	0.114	0.107	0.110	0.107	0.106	0.105	0.108	0.110	0.140	0.087	0.068	0.063	0.063	0.061	0.060	0.061	0.062	0.065	0.084	0.051	0.038	0.035	0.035	0.034	0.033	0.034	0.034	0.037	0.049
HBD_{GF}	0.132	0.106	0.100	0.101	0.099	0.098	0.098	0.100	0.102	0.132	0.082	0.063	0.059	0.058	0.056	0.055	0.056	0.057	0.060	0.079	0.047	0.036	0.033	0.032	0.031	0.031	0.031	0.032	0.035	0.046
AFFT	0.156	0.140	0.134	0.138	0.134	0.134	0.134	0.138	0.138	0.158	0.098	0.084	0.079	0.078	0.076	0.076	0.077	0.079	0.082	0.097	0.060	0.048	0.043	0.043	0.041	0.042	0.042	0.044	0.047	0.058
$AFFT_G$	0.136	0.116	0.111	0.113	0.111	0.109	0.110	0.112	0.113	0.137	0.084	0.068	0.064	0.064	0.063	0.062	0.063	0.064	0.066	0.082	0.049	0.038	0.036	0.035	0.034	0.034	0.034	0.035	0.037	0.047
$AFFT_F$	0.137	0.111	0.105	0.107	0.103	0.102	0.102	0.105	0.107	0.139	0.083	0.064	0.060	0.059	0.057	0.057	0.057	0.059	0.061	0.081	0.046	0.034	0.032	0.031	0.030	0.030	0.030	0.031	0.033	0.045
$AFFT_{GF}$	0.130	0.103	0.097	0.100	0.097	0.096	0.096	0.098	0.100	0.132	0.079	0.060	0.056	0.056	0.054	0.053	0.054	0.055	0.058	0.078	0.044	0.033	0.030	0.030	0.029	0.029	0.028	0.030	0.032	0.043
PARA	0.166	0.171	0.178	0.185	0.186	0.186	0.183	0.180	0.171	0.172	0.101	0.094	0.094	0.096	0.095	0.094	0.095	0.095	0.093	0.101	0.065	0.055	0.054	0.053	0.053	0.052	0.052	0.053	0.053	0.061
Obs	130	148	158	164	166	168	164	161	152	130	119	117	115	116	117	118	121	124	127	134	53	44	42	40	40	40	42	44	49	64
	Panel G. Deep in-the-money																													
GPF	0.041	0.032	0.028	0.027	0.027	0.027	0.027	0.029	0.030	0.038																				
GPF_G	0.032	0.025	0.024	0.023	0.023	0.022	0.023	0.023	0.024	0.031																				
GPF_F	0.029	0.022	0.020	0.019	0.019	0.019	0.019	0.020	0.021	0.027																				
GPF_{GF}	0.028	0.021	0.020	0.019	0.018	0.018	0.019	0.019	0.020	0.026																				
HBD	0.044	0.035	0.031	0.030	0.030	0.029	0.030	0.031	0.033	0.042																				
HBD_G	0.035	0.028	0.026	0.025	0.025	0.024	0.024	0.025	0.027	0.034																				
HBD_F	0.034	0.026	0.024	0.023	0.023	0.022	0.023	0.023	0.025	0.033																				
HBD_{GF}	0.031	0.024	0.022	0.022	0.021	0.021	0.021	0.022	0.023	0.030																				
AFFT	0.041	0.032	0.029	0.027	0.028	0.027	0.027	0.029	0.030	0.038																				
$AFFT_G$	0.032	0.025	0.024	0.023	0.023	0.022	0.022	0.023	0.024	0.031																				
$AFFT_F$	0.029	0.022	0.020	0.019	0.019	0.019	0.019	0.020	0.021	0.028																				
$AFFT_{GF}$	0.028	0.021	0.020	0.019	0.019	0.018	0.019	0.019	0.020	0.027																				
PARA	0.049	0.042	0.040	0.039	0.040	0.039	0.039	0.040	0.039	0.045																				
Obs	34	27	24	24	24	24	25	26	30	43																				

Table 3.12. Pricing performance for different maturity options with different stock returns (MAPE). This table shows the out-of-sample pricing errors for each model for the period January 2015 to December 2022 for different maturity options. The columns represent groups categorized according to percentile returns, ranging from the lowest 10 percentile (10) to the highest 10 percentile (100). The observations (Obs) are given in thousands.

	10	20	30	40	50	60	70	80	90	100
Panel A. Near-term										
GPF	0.247	0.219	0.211	0.214	0.208	0.207	0.207	0.211	0.212	0.241
GPF_G	0.214	0.178	0.171	0.172	0.167	0.167	0.167	0.169	0.171	0.206
GPF_F	0.216	0.171	0.163	0.163	0.157	0.158	0.158	0.160	0.164	0.213
GPF_{GF}	0.204	0.159	0.151	0.151	0.146	0.147	0.146	0.149	0.153	0.202
HBD	0.267	0.237	0.229	0.231	0.223	0.223	0.224	0.227	0.230	0.258
HBD_G	0.235	0.195	0.186	0.187	0.182	0.181	0.180	0.184	0.190	0.220
HBD_F	0.239	0.191	0.180	0.183	0.176	0.174	0.175	0.176	0.183	0.230
HBD_{GF}	0.227	0.176	0.167	0.167	0.163	0.161	0.160	0.164	0.171	0.214
AFFT	0.248	0.221	0.211	0.214	0.208	0.208	0.208	0.212	0.213	0.242
$AFFT_G$	0.216	0.179	0.172	0.173	0.168	0.167	0.167	0.170	0.173	0.209
$AFFT_F$	0.220	0.172	0.164	0.165	0.158	0.158	0.158	0.161	0.165	0.217
$AFFT_{GF}$	0.208	0.160	0.152	0.153	0.147	0.147	0.147	0.150	0.154	0.205
PARA	0.257	0.258	0.267	0.277	0.274	0.273	0.270	0.264	0.252	0.246
Obs	653,681	647,822	645,338	641,194	640,560	641,871	643,049	647,998	655,742	663,038
Panel B. Mid-term										
GPF	0.141	0.124	0.121	0.123	0.120	0.121	0.122	0.123	0.123	0.141
GPF_G	0.113	0.099	0.099	0.099	0.098	0.097	0.098	0.098	0.097	0.112
GPF_F	0.090	0.079	0.077	0.078	0.076	0.076	0.076	0.076	0.076	0.088
GPF_{GF}	0.087	0.076	0.075	0.075	0.074	0.073	0.073	0.073	0.074	0.086
HBD	0.148	0.131	0.129	0.130	0.129	0.128	0.128	0.130	0.128	0.147
HBD_G	0.119	0.107	0.106	0.106	0.105	0.104	0.104	0.104	0.104	0.120
HBD_F	0.102	0.091	0.090	0.091	0.090	0.088	0.088	0.088	0.088	0.102
HBD_{GF}	0.097	0.087	0.085	0.086	0.085	0.084	0.083	0.084	0.084	0.098
AFFT	0.142	0.124	0.121	0.122	0.121	0.121	0.122	0.123	0.123	0.142
$AFFT_G$	0.113	0.100	0.100	0.099	0.098	0.097	0.098	0.098	0.098	0.113
$AFFT_F$	0.092	0.081	0.078	0.080	0.078	0.077	0.077	0.077	0.077	0.091
$AFFT_{GF}$	0.089	0.077	0.076	0.077	0.076	0.075	0.075	0.075	0.075	0.089
PARA	0.212	0.213	0.222	0.230	0.231	0.228	0.227	0.226	0.213	0.214
Obs	189,143	192,664	193,505	196,238	197,048	196,507	195,653	194,356	190,266	188,021
Panel C. Long-term										
GPF	0.108	0.091	0.086	0.087	0.086	0.086	0.088	0.092	0.093	0.115
GPF_G	0.084	0.073	0.071	0.071	0.071	0.069	0.070	0.073	0.074	0.089
GPF_F	0.060	0.052	0.051	0.051	0.051	0.050	0.050	0.052	0.052	0.062
GPF_{GF}	0.058	0.052	0.050	0.050	0.050	0.049	0.049	0.051	0.051	0.061
HBD	0.108	0.092	0.089	0.089	0.087	0.087	0.088	0.092	0.093	0.116
HBD_G	0.086	0.076	0.073	0.073	0.074	0.072	0.072	0.076	0.076	0.095
HBD_F	0.069	0.060	0.059	0.059	0.059	0.058	0.057	0.060	0.059	0.074
HBD_{GF}	0.067	0.058	0.057	0.057	0.057	0.056	0.055	0.058	0.058	0.073
AFFT	0.108	0.092	0.086	0.087	0.086	0.086	0.088	0.091	0.093	0.116
$AFFT_G$	0.083	0.073	0.071	0.071	0.071	0.069	0.070	0.072	0.073	0.090
$AFFT_F$	0.061	0.054	0.052	0.053	0.052	0.051	0.052	0.053	0.053	0.064
$AFFT_{GF}$	0.059	0.053	0.051	0.052	0.051	0.050	0.051	0.052	0.052	0.063
PARA	0.206	0.211	0.215	0.221	0.221	0.218	0.219	0.225	0.212	0.223
Obs	92,546	95,881	97,299	98,954	98,937	97,606	97,477	93,995	90,152	86,140

outperform the PARA model. This pattern also can be seen for different types of options (call and put), different moneyness and different maturities from Table 3.10, 3.11 and 3.12, which implies that the good performance of GPF_{GF} is consistent for different groups of options.

Another key observation is that while PARA shows relatively uniform MAPEs across stocks with varying returns, the machine learning-based models exhibit a smile pattern in their distribution. This pattern indicates higher MAPEs for stocks at the extreme ends of returns - both lowest and highest. For example, for the GPF_{GF} model, the MAPE is 0.166 for the 0-10th percentile model, 0.166 for the 90-100th percentile model, and 0.121 for the 40-50th percentile and 50-60th percentile models. This pattern exists for other machine learning-based models. The smile pattern in pricing errors, more pronounced for stocks with very low or high returns, arises mainly due to their higher volatility and the market's reaction to them. Stocks at these extremes are more unpredictable due to factors like speculative trading, market news, or significant events, leading to abrupt price changes. High-return stocks might be overvalued due to market optimism, while low-return stocks could be undervalued or face negative sentiment, both leading to discrepancies between model predictions and actual prices. Moreover, machine learning models, reliant on historical data, may not accurately predict these outliers, as extreme returns are less represented in typical market conditions. While machine learning models generally enhance pricing accuracy, their effectiveness varies with the stock's return profile, being more accurate for stocks with moderate, predictable returns and less so for those with extreme returns. This highlights the need for continuous refinement of these models to improve forecasting accuracy for different return profiles.

In Table 3.10, the smile pattern is evident in both call and put options, with a more pronounced effect in call options, particularly for stocks with the highest and lowest returns. For example, For the GPF_{GF} model, the MAPEs for call options are 0.176 (0-10th percentile), 0.152 (90-100th percentile), 0.122 (40-50th percentile), and 0.121 (50-60th percentile), while for put options, they are 0.153 (0-10th percentile), 0.190 (90-100th percentile), 0.118 (40-50th percentile), and 0.121 (50-60th percentile). This pattern exists for other machine learning-based models. The higher MAPE for call options in these extreme return categories can be attributed to several factors. For stocks with high returns, call options may carry inflated expectations of continued upward movement, leading to overvaluation and larger pricing errors. Conversely, for stocks with low returns, the pessimism

surrounding these stocks might result in the underpricing of call options, again causing larger errors. Additionally, call options inherently carry more risk and uncertainty in extreme market conditions compared to put options, as they represent a right to buy in a potentially volatile market. This increased uncertainty in predicting future prices of stocks with extreme returns leads to higher MAPEs for call options compared to put options.

In Table 3.11, the smile pattern observed across all moneyness categories of options can be largely attributed to the characteristics of machine learning algorithms and the nature of option moneyness. Machine learning models excel at capturing patterns in data, but their performance can vary based on the complexity and volatility of the underlying asset. For each moneyness group, stocks with either lower or higher returns exhibit higher MAPE compared to those with mid-range returns. This trend is more pronounced in options further out-of-the-money (DOTM and OTM, both with MAPE above 0.20), and gradually lessens as we move towards options that are DITM (with MAPE around 0.028). This variation in MAPE can be attributed to the increasing predictability and stability of options as they move from being more speculative (DOTM) to more intrinsic-value-based (DITM). Machine learning algorithms are more effective in situations where the pricing dynamics are stable and predictable, as seen in the lower MAPEs for ITM and DITM options. In contrast, the higher volatility and less predictable behaviour of DOTM and OTM options make it more challenging for these algorithms to accurately predict pricing, resulting in a higher MAPE. This overall smile pattern across different moneyness categories underscores the sensitivity of machine learning-based option pricing models to the inherent risks and volatility associated with the underlying assets' returns.

In Table 3.12, the smile pattern is also observed across different maturities of options, being more pronounced in near-term options and less evident in long-term options. One reason for this is the larger sample size of near-term options compared to mid-term and long-term options, which impacts the performance of machine learning algorithms. Machine learning models are data-driven; they excel when they have more data to learn from. A larger dataset for near-term options provides these models with more information, enabling them to capture and learn from the nuances and complexities of short-term market movements more effectively. This leads to a more pronounced smile pattern as the models can detect and reflect the volatility and pricing errors more accurately

in these options. On the other hand, long-term options tend to lean more towards their intrinsic value over time. This intrinsic value is often less volatile and more predictable, which may explain why the smile pattern is less evident in these options. Moreover, long-term options' pricing is less influenced by short-term market fluctuations and more by the underlying asset's fundamental value, which machine learning models might capture with less variability.

3.4.3 Feature importance

LightGBM provides two methods to calculate the importance of each input feature: Frequency (Number of Splits) and Gain (Split Importance). The Frequency counts the total number of times a feature is used to split the data across all trees in the model. Features used more often in the tree construction process are considered more important. The Gain (Split Importance) measures the total improvement in the loss function that results from each split based on a particular feature. A higher gain indicates that the feature is more important for making accurate predictions. LightGBM's default method for feature importance is Frequency, we only report the importance of features given by Frequency. For comparison purposes, we calculate the importance of each feature as a percentage of the importance of all features per day and finally derive the average of all samples for each feature.

Figure 3.6 presents the feature importance for each model. For the models using firm characteristics, we first report the total feature importance of the firm characteristics during the prediction process, and then give the detailed feature importance of each firm characteristic in Figure 3.7.

Figure 3.6. Feature importance for each model. The scores in the figure are the average scores of all monthly training results extracted from the LightGBM algorithm. A higher score means the feature is used more often in the tree construction process and the feature is considered more important. The specifications of each model can be found in Table 3.3.

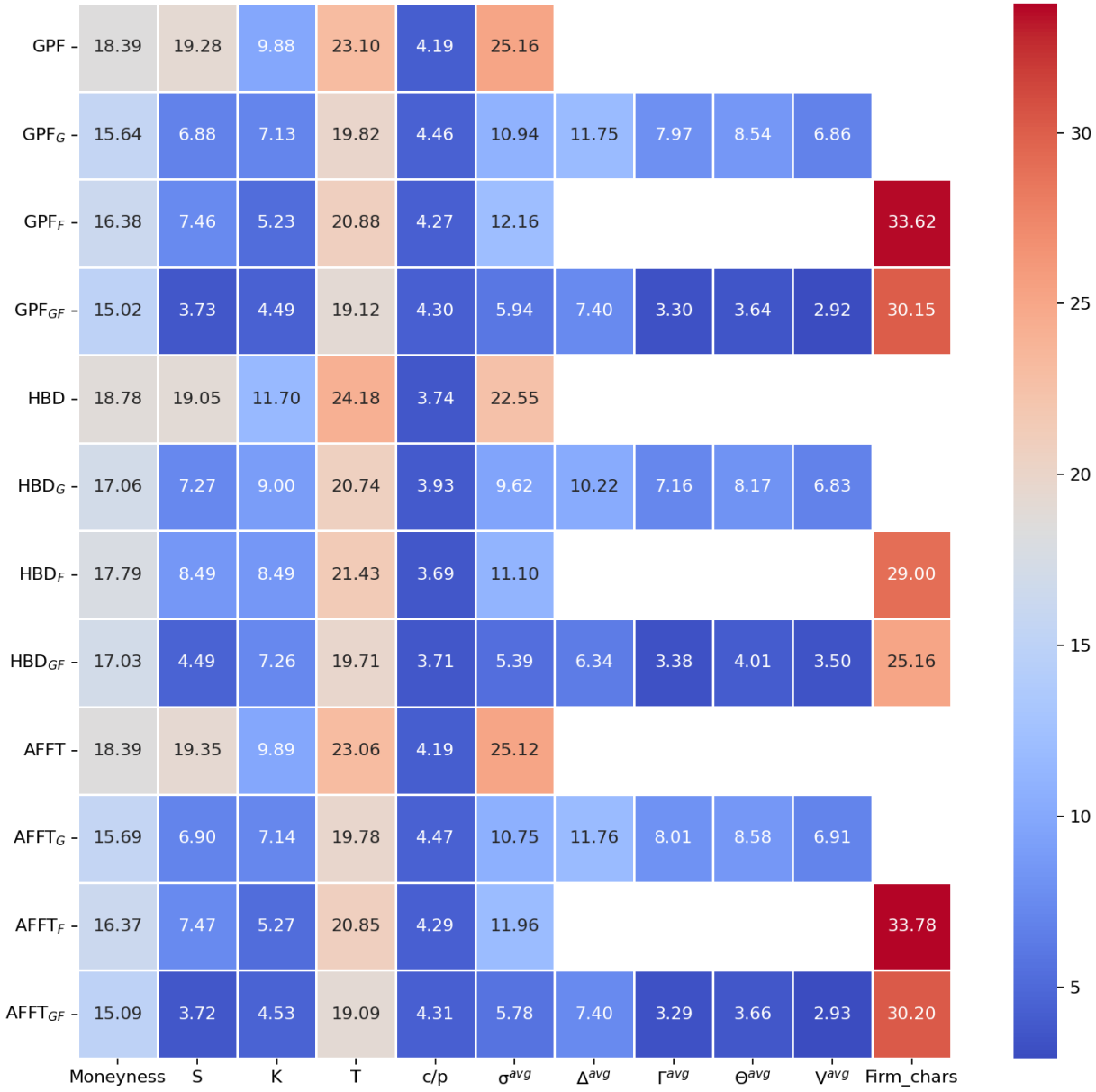
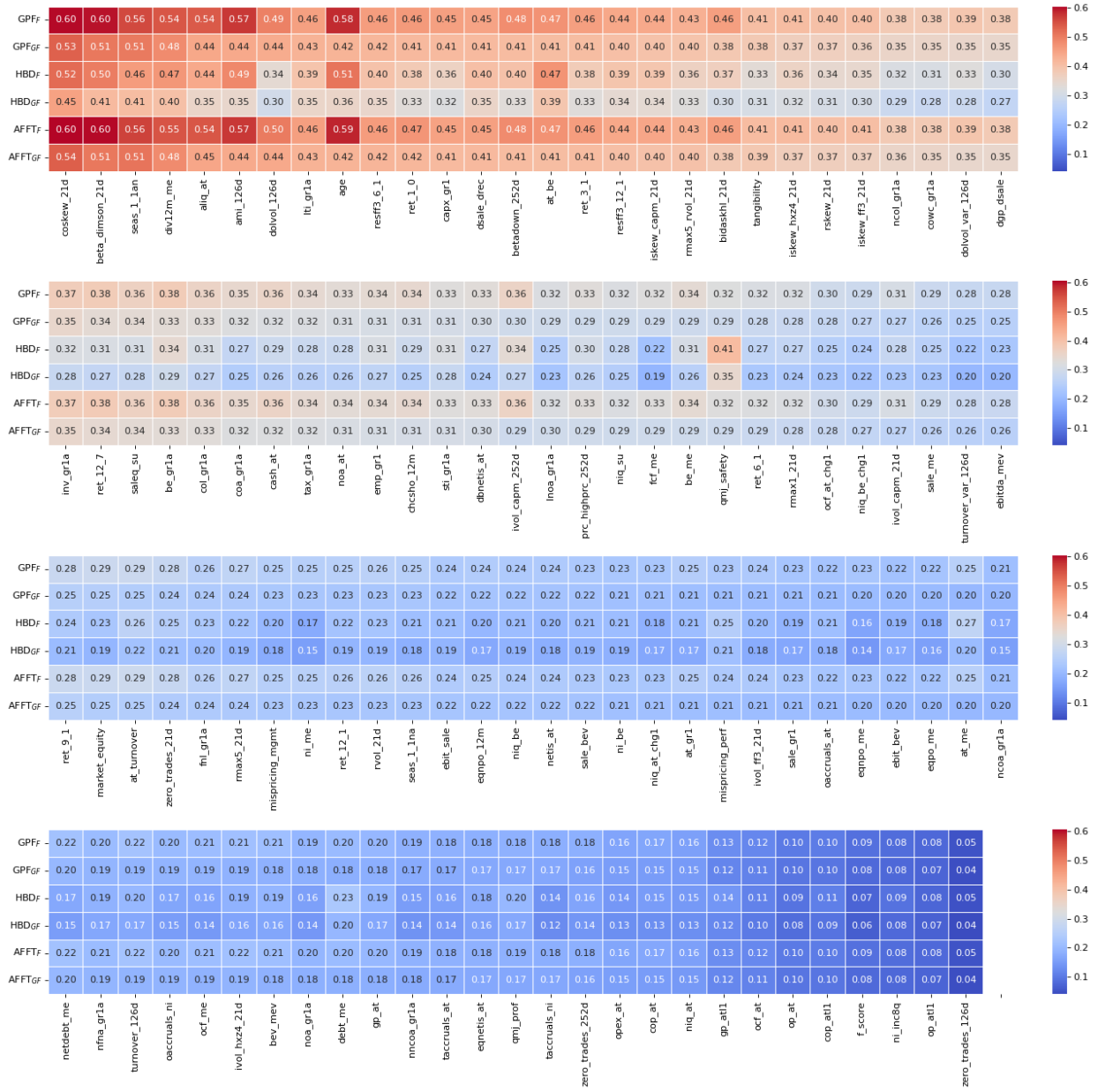


Figure 3.7. Feature importance of the firm characteristics. The definitions of the firm characteristics are outlined in the Appendix. The ranking of these characteristics is determined based on their feature importance within the GPF_{GF} model, identified as the top-performing model overall.



Firm characteristics play an important role in understanding the performance of firms and their prospective growth. As shown in our previous research Andreou et al. (2023), the firm characteristics can help to increase the prediction accuracy of the option price for all stocks. The firm characteristics still show their help in the period between January 2015 to December 2022 for the SP500 stocks, and the prediction power exists in all pricing structures (GPF, HBD, and AFFT).

The option Greeks also show their help in the pricing process. Furthermore, the combination of the option Greeks and firm characteristics further increases the pricing accuracy.

In Figure 3.6, the GPF_{GF} demonstrates that the basic variables, namely $Moneyness$, S, X, T and c/p, have importance values of 15.02, 3.73, 4.49, 19.12, and 4.30, respectively. The σ^{avg} holds an importance of 5.94. Meanwhile, the Greek groups: Δ^{avg}, Γ^{avg}, Θ^{avg}, and V^{avg}, possess respective importance scores of 7.40, 3.30, 3.64, and 2.92; the firm characteristics group accounts for 30.15 importance in total.

When we compare the feature importance of GPF and GPF_G models, we observe that σ^{avg} has a heightened importance of 25.16 in the GPF. However, in the GPF_G, the importance of the σ^{avg}, Δ^{avg}, Γ^{avg}, Θ^{avg}, and V^{avg}, are 10.94, 11.75, 7.97, 8.54, and 6.86, respectively. The superior performance of the GPF_G over GPF, suggests that the option Greeks contain additional critical information not solely captured by the σ^{avg}. It has been shown that this additional information helps to predict option prices. The efficacy of the Greeks in option pricing isn't restricted to the GPF structure alone; it's also evident within the HBD and AFFT structures. However, the GPF structure stands out, indicating its proficiency in harnessing the nuanced information embedded within the option Greeks. Furthermore, when analyzing the importance of the variable S, we find an interesting phenomenon. In GPF, the importance of S is 19.28. However, in GPF_G, this importance decreases to 6.88. The decrease in importance implies that option Greeks may contain some predictive power originally associated with S, revealing interdependencies in the input features.

The importance of firm features is prominent in all models, not only evident in the GPF structure but also in the HBD and AFFT structures. It is noteworthy that when we examine the GPF_{GF} in depth, the combined importance of the firm features reaches 30.15, which exceeds the impact of any individual feature. However, a closer look at Figure 3.7 reveals an interesting phenomenon: when evaluated individually, the importance of each firm characteristic is relatively low (below 0.5). In the field of machine learning, particularly with tree-based models like LightGBM, feature importance reveals the role of each input in the model's decisions. LightGBM, a gradient boosting framework, excels in handling datasets with both categorical and continuous features and is adept at navigating multicollinearity—where features are interrelated, a challenge for linear

models. LightGBM in GPF_{GF}efficiently captures the synergy between firm characteristics, illustrating how combined features can significantly impact the pricing accuracy, even if they seem minor individually.

3.4.4 Robustness tests

3.4.4.1 Pricing performance for different parametric models

We examine the robustness of our findings by comparing the pricing performance of the GPF, HBD and AFFT structures under other parametric models. These parametric models include the Crank-Nicolson model and the Barone-Adesi and Whaley model (BAW). Details of each model are provided in the Appendix.

Table 3.13 provides a comparison of the performance of the Crank-Nicholson and BAW models with the binomial model. For the Crank Nicolson-based models, the RMSEs for GPF_{GF}, HBD_{GF}, and AFFT_{GF} are 1.537, 1.624, and 1.564, respectively. The BAW models have RMSEs for GPF_{GF}, HBD_{GF}, and AFFT_{GF} are 1.542, 1.627, and 1.568, respectively. These values are similar to those of the binomial-based models, where the RMSEs for GPF_{GF}, HBD_{GF}, and AFFT_{GF} are 1.535, 1.622, and 1.580, respectively. The RMSEs of the PARA models for each parametric approach are also similar: 2.562 for the Crank-Nicholson model, 2.566 for the BAW model, and 2.567 for the binomial model.

This highlights the robustness of our findings. This suggests that the choice of parametric model - be it Crank Nicolson, BAW or binomial - does not have a significant impact on the overall error profile. This applies not only to the benchmark PARA model, but also to the proposed structures (GPF, HBD, and AFFT), suggesting that their effectiveness in option pricing is not affected by changes in the underlying parameter modelling approach

.

3.4.4.2 Training models for each return percentile

Table 3.14, 3.15, 3.16, and 3.17 illustrate a smile pattern in option pricing performance across stocks from different return profiles. The smile pattern is consistent across different types of options,

Table 3.13. Pricing performance for all options. This table presents the out-of-sample pricing errors of each model for all the options. The out-of-sample period is from January 2015 to December 2022 and encompasses 9,302,680 data points. RMSE, MAE, RMSPE, and MAPE respectively refer to the root mean squared error, the mean absolute error, the root mean squared percentage error, and the mean absolute percentage error. MCS-p refers to the p-value of the model confidence set test. Columns labelled "%" indicate the proportion of months during the entire forecast period in which each model ranked first according to the error metric specified in the previous column.

	RMSE	%	MAE	%	RMSPE	%	MAPE	%	MCS-p
Panel A. Crank-Nicolson model									
GPF	2.108	0.000	0.649	0.000	0.366	0.000	0.186	0.000	0.000
GPF_G	1.674	3.125	0.527	0.000	0.297	15.625	0.151	0.000	0.000
GPF_F	1.630	15.625	0.484	11.458	0.303	3.125	0.141	0.000	0.001
GPF_{GF}	1.537	38.542	0.457	61.458	0.280	55.208	0.132	78.125	1.000
HBD	2.014	0.000	0.635	0.000	0.414	0.000	0.199	0.000	0.000
HBD_G	1.728	3.125	0.532	0.000	0.366	0.000	0.164	0.000	0.000
HBD_F	1.691	4.167	0.501	1.042	0.372	1.042	0.157	0.000	0.001
HBD_{GF}	1.624	16.667	0.476	16.667	0.341	0.000	0.146	0.000	0.011
AFFT	2.135	0.000	0.653	0.000	0.370	0.000	0.187	0.000	0.000
AFFT_G	1.735	0.000	0.536	0.000	0.303	9.375	0.152	0.000	0.000
AFFT_F	1.667	4.167	0.493	1.042	0.309	2.083	0.142	0.000	0.000
AFFT_{GF}	1.564	17.708	0.467	12.500	0.286	31.250	0.134	39.583	0.011
PARA	2.562	0.000	0.857	0.000	0.417	0.000	0.250	0.000	0.000
Panel B. Barone-Adesi and Whaley model									
GPF	2.093	0.000	0.647	0.000	0.365	0.000	0.186	0.000	0.000
GPF_G	1.684	3.125	0.527	0.000	0.297	12.500	0.151	0.000	0.000
GPF_F	1.638	19.792	0.484	12.500	0.304	1.042	0.141	0.000	0.000
GPF_{GF}	1.542	35.417	0.458	58.333	0.282	54.167	0.132	79.167	1.000
HBD	2.014	0.000	0.635	0.000	0.414	0.000	0.199	0.000	0.000
HBD_G	1.732	1.042	0.533	0.000	0.367	0.000	0.165	0.000	0.000
HBD_F	1.691	6.250	0.501	1.042	0.372	1.042	0.157	0.000	0.000
HBD_{GF}	1.627	16.667	0.477	14.583	0.343	0.000	0.146	0.000	0.004
AFFT	2.143	0.000	0.654	0.000	0.370	0.000	0.187	0.000	0.000
AFFT_G	1.718	3.125	0.534	0.000	0.303	6.250	0.152	0.000	0.000
AFFT_F	1.667	2.083	0.492	1.042	0.309	1.042	0.142	0.000	0.000
AFFT_{GF}	1.568	15.625	0.468	16.667	0.288	31.250	0.134	37.500	0.052
PARA	2.566	0.000	0.858	0.000	0.418	0.000	0.250	0.000	0.000

moneyness, and maturities. However, due to the nature of machine learning, the unbalance of the training datasets can lead to biased outcomes. Typically, these algorithms may focus more on minimizing pricing errors for stocks in middle return percentiles (10 to 90) compared to those at the extremities (0-10 and 90-100). This bias arises if the relationship between input factors and output pricing differs significantly among various stock return percentiles. Training a single model for all options could result in suboptimal performance for stocks at the extremes. On the flip side, creating individual machine learning models for each option group poses its own challenges, particularly the risk of overfitting due to limited data in each specific category. To evaluate this balance between the issues of data unbalance and limited data availability, we trained models for each stock return percentile separately. We then compared these results to those obtained from a more generalist, one-model-fits-all approach.

Table 3.14 presents the option pricing performance for each stock return percentile based on models trained individually on these percentiles. The results show that the proposed model exhibits a higher MAPE compared to the results presented in Table 3.9. All machine learning-based models show an increase in MAPE except for the PARA model, which is unaffected due to its insensitivity to sample size limitations. It is worth noting that the best performing GPF_{GF}, identified in both the previous analysis and Table 3.14, has a MAPE of approximately 0.190 when trained on the segmented stock return group; however, this MAPE drops to between 0.120 and 0.166 when the model is trained to include all available options. This significant difference could be due to the limitations imposed by the smaller sample size, which limits the ability of the machine learning model to capture the full range of information compared to the more inclusive primary analysis.

Comparing the results from Tables 3.9 and 3.14, it becomes evident that training machine learning algorithms on non-extreme stocks enhance the pricing performance for extreme return stocks. This improvement suggests a stable internal relationship between option pricing and input variables, effectively captured by these algorithms when provided with sufficient data. This finding indicates that the variations in the relationship between input features and option pricing across different option types are not substantial enough to necessitate separate models for each group. Consequently, a unified model approach for all types of options seems to be sufficient and effective.

Despite inferior pricing performance in Table 3.14 compared to Table 3.9, machine learning

Table 3.14. Pricing performance for all options with different stock returns (MAPE) - Robust test. This table presents the out-of-sample pricing errors of each model for all options. The out-of-sample period is from January 2015 to December 2022. The columns represent groups categorized according to percentile returns, ranging from the lowest 10 percentile (10) to the highest 10 percentile (100). The observations (Obs) are given in thousands.

	10	20	30	40	50	60	70	80	90	100
GPF	0.227	0.214	0.210	0.211	0.210	0.206	0.204	0.207	0.207	0.211
GPF_G	0.211	0.202	0.200	0.200	0.199	0.197	0.195	0.197	0.197	0.200
GPF_F	0.202	0.194	0.192	0.191	0.190	0.188	0.186	0.189	0.188	0.192
GPF_{GF}	0.199	0.192	0.190	0.189	0.189	0.187	0.184	0.187	0.187	0.190
HBD	0.252	0.239	0.233	0.233	0.233	0.228	0.227	0.228	0.228	0.238
HBD_G	0.233	0.223	0.222	0.222	0.222	0.217	0.216	0.216	0.217	0.224
HBD_F	0.251	0.234	0.241	0.235	0.233	0.229	0.227	0.226	0.227	0.234
HBD_{GF}	0.238	0.223	0.229	0.225	0.225	0.218	0.218	0.217	0.220	0.225
AFFT	0.230	0.213	0.211	0.209	0.209	0.205	0.204	0.206	0.205	0.214
$AFFT_G$	0.212	0.200	0.199	0.198	0.198	0.195	0.193	0.195	0.194	0.202
$AFFT_F$	0.211	0.195	0.193	0.191	0.191	0.189	0.186	0.189	0.190	0.199
$AFFT_{GF}$	0.206	0.193	0.191	0.189	0.189	0.187	0.185	0.186	0.188	0.197
PARA	0.243	0.244	0.252	0.261	0.259	0.258	0.256	0.252	0.240	0.238
Obs	935,370	936,367	936,142	936,386	936,545	935,984	936,179	936,349	936,160	937,199

models consistently outperform the PARA across all return profiles, demonstrating their effectiveness even with limited sample sizes. However, this effectiveness is constrained. Furthermore, the presence of the smile pattern in these models, where stocks with extreme returns exhibit higher MAPEs than those with non-extreme returns, underscores the influence of stock returns on pricing performance.

These observations are further supported by Tables 3.15, 3.16, and 3.17, which show performance metrics across different option types, moneyness, and maturities. These results collectively highlight the impact of limited sample sizes and emphasize the importance of using sufficient data for training machine learning algorithms.

3.5 Conclusion

This paper proposes machine learning-based option pricing models that incorporate firm characteristics and option Greeks. We aim to assess whether firm characteristics and option Greeks can enhance the pricing of stock options. We employ three semi-parametric models, a variant of Andreou et al. (2010)'s generalized parametric function model (GPF), a variant of Lajbcygier and

Table 3.15. Pricing performance for call and put options with different stock returns (MAPE) - Robust test. This table presents the out-of-sample pricing errors of each model for the call and put options. The out-of-sample period is from January 2015 to December 2022. The columns represent groups categorized according to percentile returns, ranging from the lowest 10 percentile (10) to the highest 10 percentile (100). The observations (Obs) are given in thousands.

	10	20	30	40	50	60	70	80	90	100
Panel A. Call options										
GPF	0.247	0.227	0.220	0.219	0.216	0.209	0.205	0.205	0.203	0.200
GPF_G	0.228	0.214	0.209	0.208	0.205	0.201	0.196	0.195	0.193	0.189
GPF_F	0.218	0.204	0.200	0.197	0.194	0.190	0.187	0.186	0.183	0.179
GPF_{GF}	0.214	0.203	0.197	0.195	0.193	0.189	0.185	0.184	0.182	0.177
HBD	0.279	0.257	0.249	0.247	0.245	0.235	0.232	0.228	0.225	0.226
HBD_G	0.257	0.241	0.238	0.235	0.233	0.224	0.221	0.216	0.213	0.212
HBD_F	0.275	0.251	0.255	0.247	0.244	0.236	0.231	0.226	0.224	0.220
HBD_{GF}	0.260	0.239	0.243	0.237	0.235	0.225	0.222	0.217	0.215	0.212
AFFT	0.250	0.226	0.221	0.218	0.215	0.208	0.205	0.205	0.201	0.202
$AFFT_G$	0.229	0.213	0.209	0.206	0.203	0.198	0.194	0.193	0.190	0.190
$AFFT_F$	0.227	0.206	0.201	0.198	0.195	0.191	0.187	0.186	0.184	0.185
$AFFT_{GF}$	0.222	0.203	0.199	0.196	0.193	0.189	0.185	0.183	0.182	0.183
PARA	0.277	0.276	0.285	0.295	0.291	0.286	0.281	0.273	0.254	0.242
Obs	516,912	530,848	538,130	543,127	552,178	556,869	563,053	572,151	583,644	599,040
Panel B. Put options										
GPF	0.204	0.198	0.197	0.199	0.201	0.201	0.202	0.210	0.214	0.231
GPF_G	0.189	0.186	0.187	0.190	0.191	0.193	0.193	0.199	0.205	0.220
GPF_F	0.183	0.181	0.182	0.182	0.184	0.186	0.186	0.194	0.198	0.216
GPF_{GF}	0.180	0.179	0.179	0.181	0.183	0.185	0.184	0.192	0.196	0.213
HBD	0.220	0.215	0.211	0.215	0.217	0.217	0.219	0.227	0.232	0.258
HBD_G	0.203	0.200	0.201	0.204	0.206	0.206	0.208	0.215	0.224	0.244
HBD_F	0.222	0.212	0.222	0.218	0.218	0.219	0.220	0.226	0.233	0.258
HBD_{GF}	0.212	0.202	0.212	0.209	0.210	0.209	0.212	0.218	0.227	0.250
AFFT	0.206	0.197	0.197	0.197	0.200	0.200	0.202	0.209	0.213	0.235
$AFFT_G$	0.191	0.185	0.186	0.187	0.189	0.190	0.191	0.198	0.202	0.222
$AFFT_F$	0.191	0.181	0.183	0.183	0.185	0.186	0.185	0.193	0.199	0.223
$AFFT_{GF}$	0.187	0.179	0.181	0.181	0.183	0.184	0.184	0.191	0.197	0.221
PARA	0.200	0.201	0.207	0.214	0.215	0.217	0.217	0.220	0.217	0.229
Obs	418,458	405,519	398,012	393,259	384,367	379,115	373,126	364,198	352,516	338,159

Table 3.16. Pricing performance for different moneyness options with different stock returns (MAPE) - Robust test. This table presents the out-of-sample pricing errors of each model for different moneyness options. The out-of-sample period is from January 2015 to December 2022. The columns represent groups categorized according to percentile returns, ranging from the lowest 10 percentile (10) to the highest 10 percentile (100). The observations (Obs) are given in thousands.

	10	20	30	40	50	60	70	80	90	100	10	20	30	40	50	60	70	80	90	100	10	20	30	40	50	60	70	80	90	100
	Panel A. Deep out-the-money										**Panel B. Out-the-money**										**Panel C. Just out-the-money**									
GPF	0.333	0.308	0.298	0.299	0.300	0.290	0.293	0.299	0.306	0.324	0.309	0.290	0.285	0.284	0.285	0.280	0.277	0.285	0.288	0.298	0.245	0.237	0.236	0.239	0.238	0.236	0.233	0.236	0.234	0.238
GPF_G	0.304	0.287	0.281	0.284	0.282	0.277	0.279	0.285	0.292	0.308	0.287	0.273	0.271	0.270	0.271	0.268	0.265	0.271	0.274	0.282	0.230	0.224	0.225	0.228	0.226	0.227	0.223	0.224	0.223	0.226
GPF_F	0.288	0.272	0.268	0.268	0.267	0.263	0.266	0.270	0.273	0.291	0.276	0.261	0.259	0.255	0.256	0.255	0.252	0.258	0.260	0.272	0.222	0.218	0.218	0.218	0.218	0.217	0.214	0.218	0.216	0.221
GPF_{GF}	0.283	0.269	0.265	0.266	0.265	0.262	0.263	0.268	0.272	0.287	0.272	0.258	0.256	0.253	0.255	0.254	0.249	0.255	0.258	0.268	0.219	0.216	0.215	0.216	0.216	0.215	0.212	0.215	0.215	0.218
HBD	0.405	0.384	0.375	0.379	0.380	0.370	0.369	0.373	0.378	0.407	0.342	0.326	0.321	0.321	0.322	0.314	0.316	0.321	0.320	0.335	0.258	0.250	0.247	0.250	0.251	0.247	0.246	0.246	0.245	0.252
HBD_G	0.372	0.359	0.360	0.362	0.363	0.351	0.355	0.355	0.362	0.384	0.317	0.305	0.305	0.305	0.307	0.299	0.300	0.303	0.306	0.315	0.238	0.234	0.236	0.239	0.238	0.236	0.234	0.233	0.233	0.238
HBD_F	0.407	0.376	0.401	0.394	0.386	0.382	0.377	0.377	0.378	0.393	0.340	0.317	0.329	0.320	0.320	0.316	0.313	0.314	0.318	0.329	0.255	0.246	0.252	0.248	0.249	0.245	0.244	0.242	0.245	0.252
HBD_{GF}	0.385	0.357	0.374	0.373	0.374	0.358	0.360	0.359	0.367	0.381	0.324	0.302	0.314	0.307	0.309	0.301	0.301	0.303	0.307	0.317	0.243	0.235	0.243	0.240	0.240	0.236	0.235	0.235	0.236	0.243
AFFT	0.340	0.303	0.298	0.296	0.297	0.287	0.292	0.297	0.302	0.331	0.313	0.289	0.286	0.282	0.282	0.278	0.278	0.285	0.284	0.301	0.247	0.237	0.236	0.238	0.238	0.236	0.234	0.235	0.233	0.240
$AFFT_G$	0.307	0.281	0.280	0.279	0.278	0.271	0.274	0.281	0.284	0.310	0.289	0.272	0.269	0.267	0.268	0.264	0.264	0.269	0.270	0.285	0.231	0.225	0.225	0.227	0.226	0.225	0.222	0.223	0.222	0.228
$AFFT_F$	0.306	0.271	0.268	0.269	0.266	0.264	0.263	0.268	0.274	0.304	0.288	0.262	0.261	0.256	0.255	0.257	0.252	0.257	0.263	0.281	0.230	0.220	0.219	0.219	0.219	0.217	0.215	0.218	0.218	0.226
$AFFT_{GF}$	0.297	0.267	0.264	0.267	0.263	0.260	0.261	0.266	0.270	0.301	0.281	0.258	0.259	0.252	0.253	0.254	0.250	0.254	0.260	0.279	0.225	0.218	0.218	0.217	0.217	0.216	0.213	0.215	0.216	0.224
PARA	0.388	0.386	0.400	0.415	0.415	0.416	0.415	0.411	0.394	0.397	0.325	0.326	0.335	0.350	0.347	0.349	0.344	0.344	0.330	0.328	0.254	0.265	0.280	0.291	0.291	0.288	0.287	0.281	0.267	0.262
Obs	169	149	139	135	132	129	131	132	139	155	188	185	183	180	179	177	177	177	178	175	244	266	275	278	279	281	276	273	261	237
	Panel D. At-the-money										**Panel E. Just in-the-money**										**Panel F. In-the-money**									
GPF	0.166	0.158	0.156	0.157	0.157	0.156	0.154	0.156	0.154	0.159	0.105	0.094	0.091	0.090	0.090	0.089	0.088	0.090	0.090	0.097	0.064	0.055	0.052	0.051	0.051	0.050	0.050	0.051	0.052	0.058
GPF_G	0.155	0.149	0.148	0.150	0.149	0.149	0.148	0.148	0.146	0.151	0.096	0.088	0.086	0.086	0.085	0.085	0.084	0.085	0.086	0.091	0.058	0.051	0.049	0.049	0.048	0.048	0.047	0.049	0.049	0.054
GPF_F	0.150	0.146	0.144	0.145	0.144	0.144	0.143	0.145	0.142	0.146	0.092	0.085	0.083	0.082	0.082	0.081	0.081	0.081	0.082	0.087	0.054	0.048	0.046	0.045	0.045	0.045	0.044	0.045	0.046	0.050
GPF_{GF}	0.149	0.145	0.143	0.143	0.143	0.142	0.142	0.143	0.141	0.145	0.091	0.084	0.082	0.081	0.081	0.080	0.080	0.081	0.081	0.086	0.053	0.048	0.045	0.045	0.045	0.044	0.044	0.045	0.046	0.049
HBD	0.171	0.162	0.158	0.159	0.159	0.158	0.157	0.157	0.156	0.164	0.110	0.098	0.094	0.094	0.093	0.092	0.092	0.092	0.094	0.102	0.070	0.061	0.056	0.057	0.056	0.055	0.055	0.056	0.057	0.064
HBD_G	0.158	0.151	0.150	0.152	0.151	0.151	0.150	0.149	0.148	0.155	0.100	0.090	0.089	0.088	0.088	0.087	0.086	0.087	0.088	0.095	0.061	0.056	0.052	0.052	0.052	0.051	0.051	0.052	0.053	0.059
HBD_F	0.169	0.160	0.160	0.158	0.158	0.157	0.156	0.156	0.156	0.163	0.108	0.097	0.097	0.096	0.094	0.091	0.092	0.092	0.093	0.100	0.067	0.061	0.059	0.061	0.057	0.054	0.056	0.057	0.057	0.061
HBD_{GF}	0.160	0.153	0.154	0.153	0.153	0.151	0.152	0.151	0.151	0.158	0.102	0.092	0.092	0.090	0.090	0.087	0.088	0.088	0.089	0.096	0.062	0.056	0.053	0.054	0.054	0.051	0.053	0.053	0.053	0.059
AFFT	0.167	0.158	0.157	0.157	0.157	0.156	0.155	0.156	0.153	0.160	0.106	0.094	0.091	0.090	0.090	0.088	0.088	0.089	0.090	0.098	0.065	0.054	0.051	0.050	0.050	0.049	0.049	0.051	0.052	0.058
$AFFT_G$	0.156	0.149	0.149	0.149	0.149	0.149	0.147	0.148	0.145	0.151	0.097	0.087	0.086	0.085	0.085	0.083	0.083	0.084	0.085	0.092	0.058	0.050	0.048	0.048	0.047	0.046	0.046	0.048	0.048	0.054
$AFFT_F$	0.155	0.147	0.146	0.146	0.145	0.144	0.144	0.145	0.144	0.149	0.096	0.086	0.084	0.082	0.082	0.081	0.080	0.081	0.083	0.090	0.056	0.048	0.046	0.045	0.045	0.044	0.044	0.045	0.045	0.052
$AFFT_{GF}$	0.152	0.146	0.144	0.144	0.144	0.143	0.143	0.143	0.142	0.148	0.094	0.084	0.083	0.081	0.081	0.080	0.079	0.080	0.082	0.089	0.055	0.047	0.045	0.045	0.044	0.044	0.043	0.044	0.045	0.051
PARA	0.166	0.171	0.178	0.185	0.186	0.186	0.183	0.180	0.171	0.172	0.101	0.094	0.094	0.096	0.095	0.094	0.095	0.095	0.093	0.101	0.065	0.055	0.054	0.053	0.053	0.052	0.052	0.053	0.053	0.061
Obs	130	148	158	164	166	168	164	161	152	130	119	117	115	116	117	118	121	124	127	134	53	44	42	40	40	40	42	44	49	64
	Panel G. Deep in-the-money																													
GPF	0.044	0.038	0.036	0.034	0.035	0.034	0.034	0.036	0.036	0.039																				
GPF_G	0.039	0.035	0.034	0.033	0.034	0.033	0.032	0.034	0.034	0.036																				
GPF_F	0.035	0.033	0.031	0.030	0.031	0.030	0.030	0.031	0.031	0.033																				
GPF_{GF}	0.035	0.033	0.031	0.030	0.030	0.030	0.029	0.031	0.031	0.032																				
HBD	0.052	0.044	0.042	0.042	0.041	0.041	0.040	0.041	0.042	0.046																				
HBD_G	0.044	0.040	0.039	0.039	0.038	0.038	0.037	0.038	0.039	0.042																				
HBD_F	0.048	0.044	0.045	0.045	0.042	0.041	0.041	0.041	0.042	0.044																				
HBD_{GF}	0.044	0.041	0.041	0.041	0.040	0.038	0.039	0.039	0.040	0.042																				
AFFT	0.044	0.037	0.035	0.034	0.035	0.034	0.033	0.035	0.035	0.039																				
$AFFT_G$	0.039	0.034	0.033	0.032	0.033	0.032	0.031	0.033	0.032	0.036																				
$AFFT_F$	0.037	0.032	0.031	0.030	0.030	0.030	0.029	0.030	0.030	0.033																				
$AFFT_{GF}$	0.036	0.032	0.031	0.030	0.030	0.030	0.029	0.030	0.030	0.033																				
PARA	0.049	0.042	0.040	0.039	0.040	0.039	0.039	0.040	0.039	0.045																				
Obs	34	27	24	24	24	24	25	26	30	43																				

Table 3.17. Pricing performance for different maturity options with different stock returns (MAPE) - Robust test. This table shows the out-of-sample pricing errors for each model for the period January 2015 to December 2022 for different maturity options. The columns represent groups categorized according to percentile returns, ranging from the lowest 10 percentile (10) to the highest 10 percentile (100). The observations (Obs) are given in thousands.

	10	20	30	40	50	60	70	80	90	100
Panel A. Near-term										
GPF	0.260	0.246	0.243	0.244	0.242	0.238	0.235	0.238	0.237	0.238
GPF_G	0.242	0.231	0.230	0.232	0.229	0.228	0.224	0.226	0.226	0.226
GPF_F	0.238	0.227	0.226	0.225	0.223	0.222	0.219	0.222	0.220	0.224
GPF_{GF}	0.234	0.224	0.223	0.223	0.221	0.220	0.217	0.219	0.219	0.220
HBD	0.292	0.277	0.271	0.272	0.271	0.265	0.263	0.264	0.263	0.271
HBD_G	0.271	0.260	0.259	0.260	0.259	0.254	0.252	0.251	0.252	0.257
HBD_F	0.293	0.273	0.282	0.275	0.273	0.269	0.265	0.263	0.263	0.269
HBD_{GF}	0.279	0.260	0.270	0.265	0.265	0.257	0.256	0.254	0.256	0.261
AFFT	0.264	0.246	0.244	0.243	0.241	0.238	0.235	0.237	0.235	0.242
$AFFT_G$	0.244	0.231	0.231	0.230	0.229	0.227	0.223	0.225	0.224	0.229
$AFFT_F$	0.248	0.228	0.227	0.225	0.224	0.222	0.219	0.221	0.222	0.230
$AFFT_{GF}$	0.241	0.225	0.224	0.223	0.221	0.220	0.216	0.217	0.219	0.228
PARA	0.257	0.258	0.267	0.277	0.274	0.273	0.270	0.264	0.252	0.246
Obs	653,681	647,822	645,338	641,194	640,560	641,871	643,049	647,998	655,742	663,038
Panel B. Mid-term										
GPF	0.162	0.151	0.149	0.148	0.149	0.145	0.145	0.146	0.144	0.152
GPF_G	0.148	0.141	0.142	0.141	0.141	0.137	0.138	0.139	0.136	0.143
GPF_F	0.129	0.127	0.127	0.126	0.127	0.124	0.124	0.124	0.122	0.125
GPF_{GF}	0.129	0.127	0.126	0.125	0.126	0.124	0.123	0.124	0.122	0.124
HBD	0.174	0.162	0.161	0.161	0.160	0.156	0.156	0.156	0.154	0.165
HBD_G	0.156	0.148	0.149	0.150	0.149	0.146	0.146	0.145	0.143	0.151
HBD_F	0.166	0.157	0.161	0.161	0.156	0.154	0.154	0.152	0.152	0.155
HBD_{GF}	0.154	0.148	0.151	0.149	0.148	0.144	0.145	0.144	0.143	0.145
AFFT	0.163	0.148	0.147	0.146	0.148	0.143	0.144	0.145	0.142	0.153
$AFFT_G$	0.147	0.137	0.138	0.137	0.138	0.135	0.135	0.135	0.133	0.142
$AFFT_F$	0.137	0.129	0.128	0.127	0.127	0.125	0.124	0.125	0.124	0.131
$AFFT_{GF}$	0.135	0.127	0.128	0.126	0.126	0.124	0.124	0.125	0.123	0.130
PARA	0.212	0.213	0.222	0.230	0.231	0.228	0.227	0.226	0.213	0.214
Obs	189,143	192,664	193,505	196,238	197,048	196,507	195,653	194,356	190,266	188,021
Panel C. Long-term										
GPF	0.132	0.126	0.122	0.120	0.125	0.120	0.120	0.125	0.121	0.133
GPF_G	0.120	0.120	0.116	0.115	0.118	0.114	0.115	0.119	0.115	0.125
GPF_F	0.100	0.101	0.097	0.098	0.101	0.098	0.097	0.100	0.097	0.103
GPF_{GF}	0.099	0.103	0.097	0.097	0.102	0.099	0.097	0.101	0.098	0.103
HBD	0.136	0.132	0.129	0.127	0.131	0.127	0.128	0.131	0.129	0.140
HBD_G	0.123	0.124	0.121	0.120	0.123	0.119	0.120	0.124	0.121	0.131
HBD_F	0.129	0.126	0.126	0.125	0.125	0.120	0.122	0.126	0.126	0.133
HBD_{GF}	0.121	0.120	0.118	0.118	0.121	0.115	0.115	0.120	0.118	0.127
AFFT	0.132	0.123	0.120	0.117	0.122	0.117	0.118	0.122	0.119	0.132
$AFFT_G$	0.119	0.116	0.113	0.111	0.115	0.110	0.111	0.116	0.111	0.123
$AFFT_F$	0.106	0.102	0.100	0.099	0.102	0.099	0.098	0.101	0.099	0.107
$AFFT_{GF}$	0.105	0.102	0.100	0.098	0.102	0.099	0.098	0.101	0.099	0.107
PARA	0.206	0.211	0.215	0.221	0.221	0.218	0.219	0.225	0.212	0.223
Obs	92,546	95,881	97,299	98,954	98,937	97,606	97,477	93,995	90,152	86,140

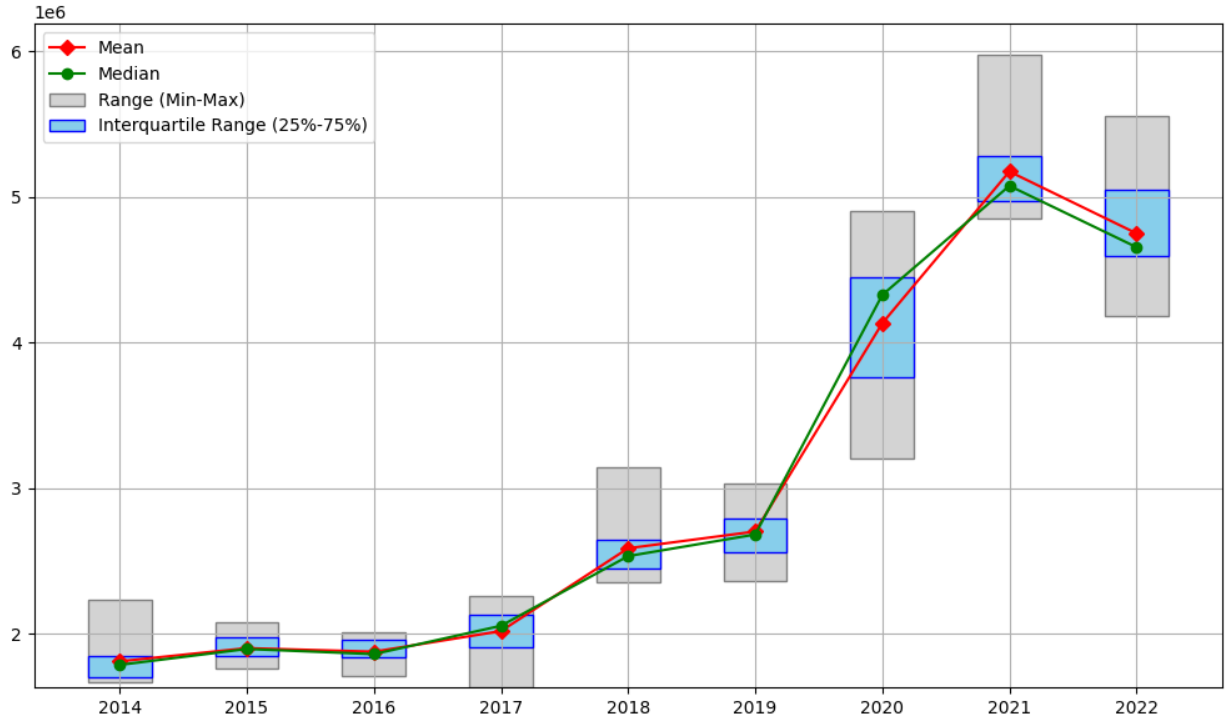
Connor (1997)'s hybrid model (HBD), and a variant of Almeida et al. (2022)'s machine learning based model (AFFT). We evaluate them using individual stock options belonging to the SP500 index in the period from 2014 to 2023. The results suggest that all GPF, HBD, and AFFT are effective in pricing American options, and the combination of firm characteristics and option Greeks can significantly improve the performance of these models. Among the three models, the GPF structure consistently performs better, especially during the Pandemic period. Our analysis also reveals that option pricing performance varies among stocks with different return profiles. Specifically, stocks at the extreme percentiles of returns exhibit higher MAPE compared to those within the middle return percentiles.

This study makes four contributions to the literature on option pricing. Firstly, we pioneer the combination of option Greeks with machine learning techniques to demonstrate the potential of option Greek in stock option pricing. Secondly, we combine option firm characteristics and option Greeks to develop a more efficient and robust option pricing model. Thirdly, through a comparative analysis of three distinct machine learning-based option pricing structures during the Pandemic period, our research identifies the model best suited for coping with such turbulent market conditions. Lastly, the study emphasizes the pivotal role of stock return profiles in determining the effectiveness of option pricing, adding a new dimension to the current understanding of this domain.

Appendices

.1 *Yearly statistic for positive trading options*

Figure 8. Yearly statistic for positive trading options. Candlestick charts give yearly observational statistics for options with a positive daily trading volume. Statistics are calculated based on monthly observations for each year. The mean and median monthly trading volumes have increased substantially since the beginning of the pandemic (2020) and begin to decline in 2022, but are still higher than in previous years.



.2 *Crank-Nicolson model*

The Crank-Nicolson method is a finite difference method used to solve partial differential equations (PDEs). It's an implicit method and it's known for its high accuracy and stability as it's unconditionally stable. Applying the Crank-Nicolson method to the discretized Black-Scholes PDE involves a combination of explicit and implicit methods [7]. It takes an average between the ex-

[7] The explicit Euler method, also known as the Forward Euler method, is known as a "first-order" method because it provides a numerical approximation to the exact solution of the ODE that is accurate to terms of the first order in the size of the time step. This method is also explicit, meaning that the value of the function at a given time step is expressed explicitly in terms of the value at previous time steps. The Implicit Euler method, also known as the Backward Euler method, is a first-order numerical procedure for solving

plicit and implicit approximations at each time step, offering a good balance between accuracy and stability.

To use the Crank-Nicolson model to price American options, first, we need to consider the Black-Scholes differential equation:

$$\frac{\partial V}{\partial t} + 0.5\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0 \tag{35}$$

where V is the option price, S is the underlying asset price, t is the time to expiration, σ is the volatility of the underlying asset, r is the risk-free interest rate.

Then, we denote V_j^n as the price of the option at time step n and spot price j. Then we discretize the asset price and time, and express the equation in finite difference form using the Crank-Nicolson method as follows:

$$\begin{aligned} V_j^n =& V_j^{n+1} \\ & - 0.5dt\left(\sigma^2 j^2 (V_{j+1}^n - 2V_j^n + V_{j-1}^n) + rj(V_{j+1}^n - V_j^n) - rV_j^n\right) \\ & - 0.5dt\left(\sigma^2 j^2 (V_{j+1}^{n+1} - 2V_j^{n+1} + V_{j-1}^{n+1}) + rj(V_{j+1}^{n+1} - V_j^{n+1}) - rV_j^{n+1}\right) \end{aligned} \tag{36}$$

for $n = 0, 1, 2, ..., N-1$ and $j = 0, 1, ..., J$, where: dt is the time step size, σ is the volatility of the underlying asset, r is the risk-free interest rate, V_{j+1}^n and V_{j-1}^n are the option prices at the next time step and neighbouring spot prices.

This is a system of linear equations. At each time step, we must solve this system to find V_j^{n+1}. The Crank-Nicolson method is unconditionally stable and second-order accurate in time, making it a good choice for this type of problem.

For American options, we also need to consider early exercise. Specifically, we use the *Quantlib.PlainVanillaPay* handles the payoff structure of the American option and the *Quantlib.AmericanExercise* to handle the early exercise of the American option. In addition, we use the *Quantlib.BlackScholesMertonProcess* to construct the BS process and use the *Quantlib.FdBlackScholesVanillaEngine* as the pricing en-

ordinary differential equations (ODEs) with a given initial value. It is more stable than the explicit Euler method, as it is unconditionally stable.

gine and specify the method as the Crank Nicolson[8] to price the American option.

.3 Barone-Adesi and Whaley model

The Barone-Adesi and Whaley (BAW) model provides an approximation to the pricing of American options. It uses an analytical formula, which improves computational efficiency compared to methods that require a binomial or trinomial tree.

The BAW method assumes that the early exercise feature is valuable, and that this value can be calculated separately. It combines the Black-Scholes model with an adjustment for the early exercise feature, which is assumed to behave as a binary option with a strike at a critical asset price.

The BAW model computes the price of an American option as a sum of the European option price and an early exercise premium. The formulas for the option prices are given as:

American call option price:

$$C = C_{BS} - B_c + A_c(1 - N(d_1)) \left(\frac{S}{q_c}\right)^{\frac{2r}{\sigma^2}}, \tag{37}$$

American put option price:

$$P = P_{BS} - B_p + A_p(1 - N(d_1)) \left(\frac{S}{q_p}\right)^{\frac{2r}{\sigma^2}}, \tag{38}$$

where C_A and P_A are the American call and put option prices, C_{BS} and P_{BS} are the Black-Scholes prices for European call and put options, B_c and B_p are the present value of the strike price X for call and put options, A_c and A_p are the adjustment coefficients for the call and put options, $N()$ is the standard normal cumulative distribution function, and $d_1 = \frac{\ln(\frac{S}{X}) + (r - q + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}}$.

The parameters q_c, q_p, A_c, and A_p are found by solving a system of non-linear equations.

The parameters q_c and q_p are the roots of the following equations respectively:

For q_c (calls):

$$f_1(x) = 1 + \frac{2r}{\sigma^2}\left(\frac{x}{1-x} - \log(1-x)\right) - \frac{X}{S}, \tag{39}$$

[8] Quantlib.FdmSchemeDesc.CrankNicolson()

For q_p (puts):

$$f_1(x) = 1 + \frac{2r}{\sigma^2}\left(\frac{x}{1-x} - \log(1-x)\right) - \frac{S}{X}, \tag{40}$$

In these equations, r is the risk-free interest rate, σ is the volatility of the underlying asset, S is the spot price of the underlying asset, and X is the strike price.

For A_c (calls):

$$A_c = -\frac{S}{q_{1_c}} N(d_1)(1 - N(d_2)), \tag{41}$$

For q_p (puts):

$$A_p = -\frac{S}{q_{1_p}} N(-d_1)(1 - N(-d_2)), \tag{42}$$

where $N()$ denotes the cumulative distribution function of standard normal distribution, $d_1 = \frac{\ln\left(\frac{S}{X}\right) + \left(r - q + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}$, and $d_2 = d_1 - \sigma\sqrt{T}$.

This is a rough outline of the calculations. The full calculation is a bit more complex and involves some iterative processes. In our research, we use the Quantlit to employ the BAW model. Specifically, we use the *Quantlib.PlainVanillaPayoff* handles the payoff structure of the American option and the *Quantlib.AmericanExercise* to handle the early exercise of the American option. In addition, we use the *Quantlib.BlackScholesMertonProcess* to construct the BS process and use the *Quantlib.BaroneAdesiWhaleyApproximationEngine* as the pricing engine to price the American option.

.4 *Month by month performances*

Figure 9, 10, and 11 show the month-by-month performance of all the models along with the rankings for each month.

Figure 9. Monthly RMSEs and rankings are presented for each model. The legend indicates the corresponding colour for each model's rank in every month. Within each rectangle, the displayed number represents the RMSE for that specific month and model. The GPF_{GF} model consistently proved to be the most accurate, achieving the lowest RMSE in 33 out of 96 months and the second-lowest RMSE in 26 months.

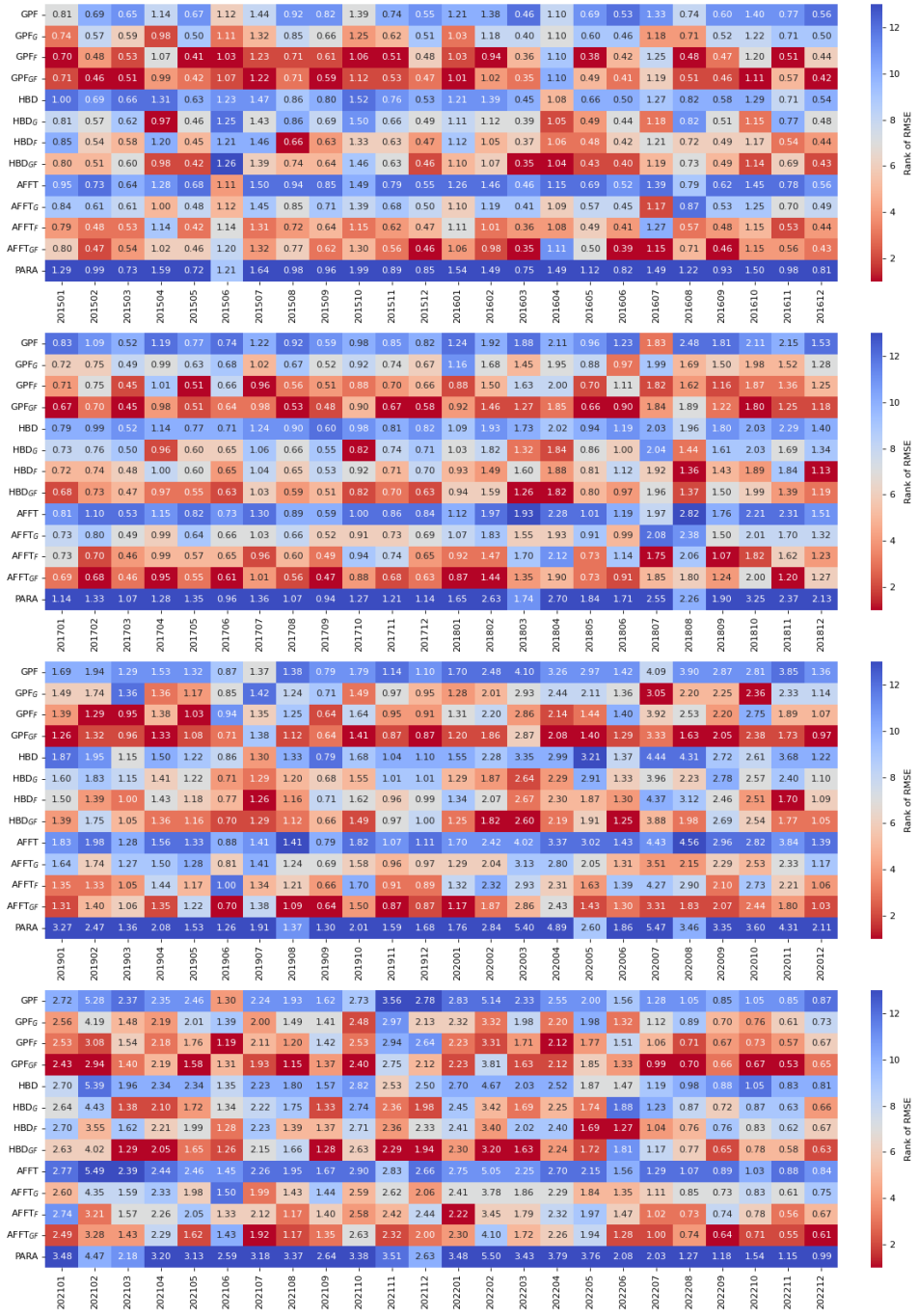


Figure 10. Monthly MAEs and rankings are presented for each model. The legend indicates the corresponding colour for each model's rank every month. Within each rectangle, the displayed number represents the MAE for that specific month and model. The GPF_{GF} model consistently proved to be the most accurate, achieving the lowest MAE in 61 out of 96 months and the second-lowest MAE in 17 months.

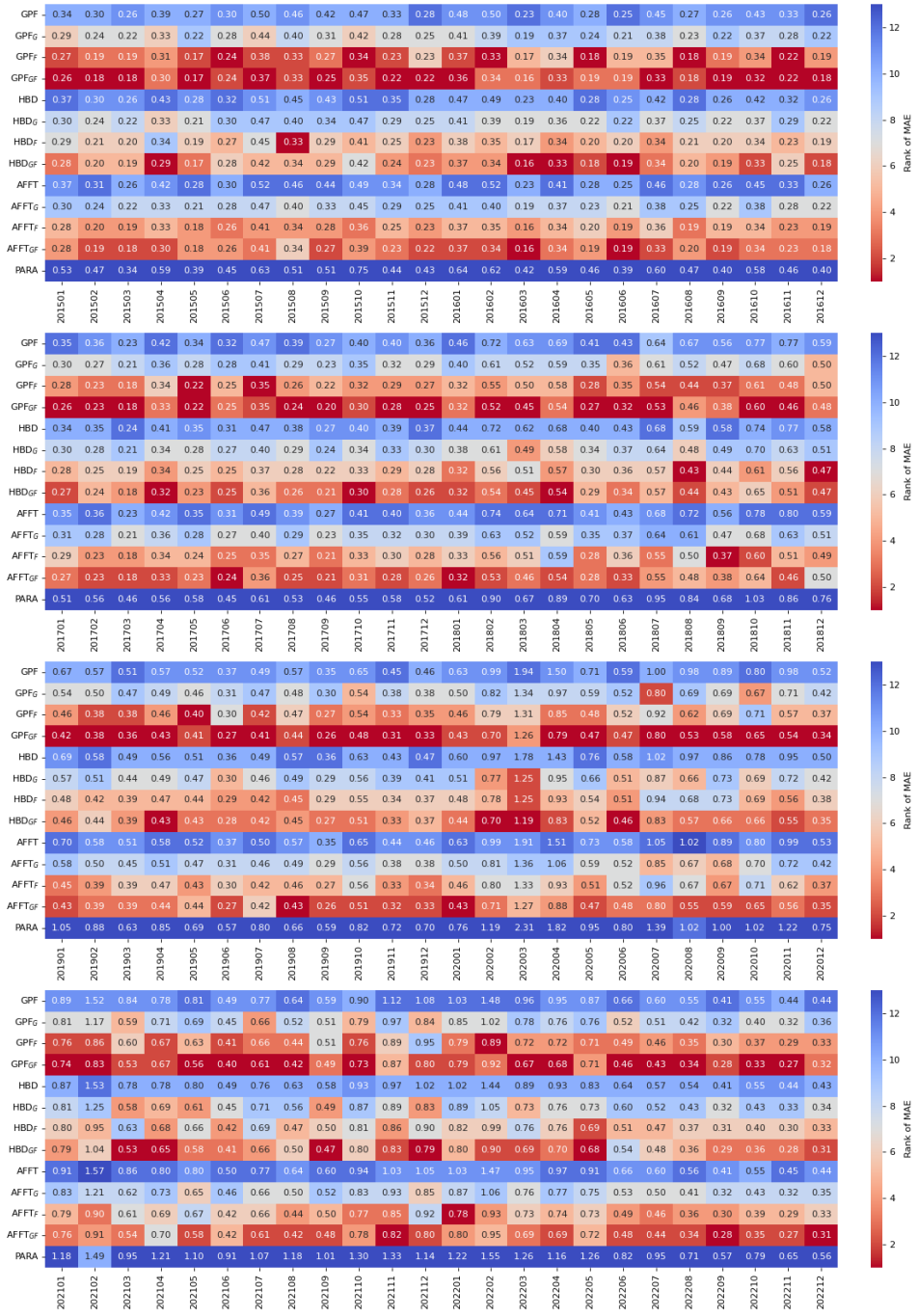
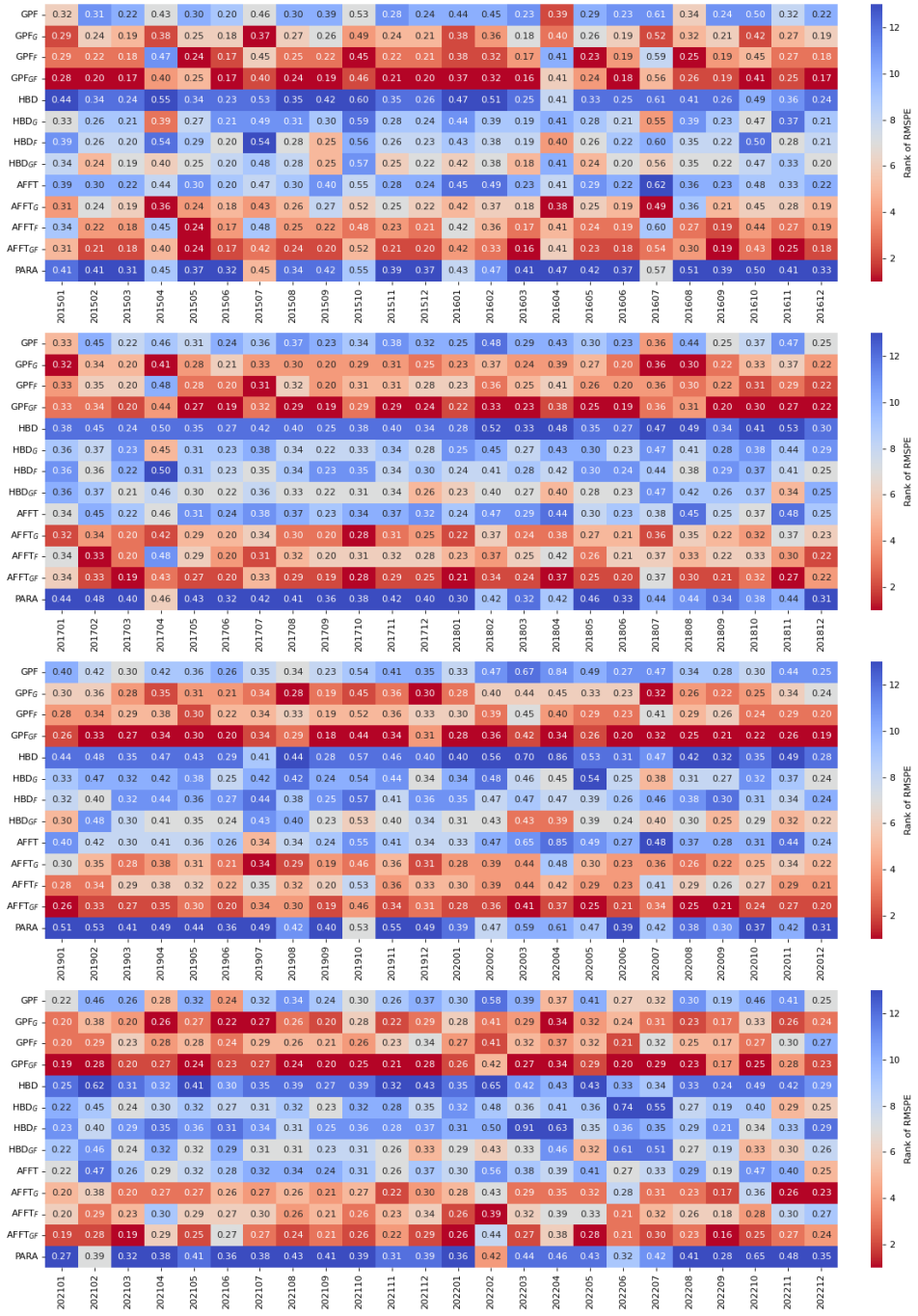


Figure 11. Monthly RMSPEs and rankings are presented for each model. The legend indicates the corresponding colour for each model's rank in every month. Within each rectangle, the displayed number represents the RMSPE for that specific month and model. The GPF_{GF} model consistently proved to be the most accurate, achieving the lowest RMSPE in 60 out of 96 months and the second-lowest RMSPE in 19 months.



Chapter 4

Conclusion and future research directions

The key finding of this thesis includes three aspects. First, the integration of machine learning with traditional option pricing models has shown notable improvements in accuracy. Firm characteristics and option Greeks, when combined with machine learning techniques, provide a more nuanced view of option valuation, reflecting the complex interplay of market factors. Second, option characteristics, particularly implied volatility and Greeks, possess significant predictive power for extreme stock returns. The use of machine learning models, such as LightGBM, has demonstrated superior performance in utilizing these characteristics to predict extreme stock returns. Third, the Pandemic period provided a unique testing ground for machine learning models in option pricing. The study reveals the resilience and adaptability of these models under extreme market conditions, with certain models like the GPF structure showing consistent good performance.

While the thesis successfully integrates machine learning with option pricing and demonstrates its effectiveness, questions remain about the models' performance under varying market conditions other than the Pandemic. In addition, the precise impact of individual firm characteristics and their interactions in option pricing is not fully explored due to the limi-

tations of machine learning algorithms, leaving room for deeper analysis.

This thesis also lays the groundwork for various future research directions. One potential area is testing the scalability and generalizability of the proposed models in different markets, including currency and commodities. Another avenue involves integrating diverse data types, like high-frequency trading data or social media sentiment, to augment the models' predictive capabilities. Additionally, exploring advanced machine learning architectures like reinforcement learning and generative adversarial networks could lead to more sophisticated approaches in option valuation and stock return prediction. Last but not least, merging behavioural finance with machine learning presents a promising approach to asset pricing and stock return prediction, combining psychological insights on investor behaviour with the analytical power of machine learning for more accurate and comprehensive financial models.

Bibliography

Almeida, C., Fan, J., Freire, G., Tang, F., 2022. Can a Machine Correct Option Pricing Models? *Journal of Business & Economic Statistics* pp. 1–14.

Almeida, C., Fan, J., Freire, G., Tang, F., 2023. Can a machine correct option pricing models? *Journal of Business & Economic Statistics* 41, 995–1009.

Andersen, L., Broadie, M., 2004. Primal-dual simulation algorithm for pricing multidimensional American options. *Management Science* 50, 1222–1234.

Andreou, C. K., Andreou, P. C., Lambertides, N., 2021. Financial distress risk and stock price crashes. *Journal of Corporate Finance* 67, 101870.

Andreou, P. C., 2015. Effects of market default risk on index option risk-neutral moments. *Quantitative Finance* 15, 2021–2040.

Andreou, P. C., Charalambous, C., Martzoukos, S. H., 2006. Robust artificial neural networks for pricing of European options. *Computational Economics* 27, 329–351.

Andreou, P. C., Charalambous, C., Martzoukos, S. H., 2010. Generalized parameter functions for option pricing. *Journal of Banking & Finance* 34, 633–646.

Andreou, P. C., Charalambous, C., Martzoukos, S. H., 2014. Assessing the performance of symmetric and asymmetric implied volatility functions. *Review of Quantitative Finance and Accounting* 42, 373–397.

Andreou, P. C., Han, C., Li, N., 2023. Stock options pricing via machine learning methods combined with firm characteristics. In *Stock options pricing via machine learning methods combined with firm characteristics: Andreou, Panayiotis C.— uHan, Chulwoo— uLi, Nan*, [Sl]: SSRN.

Audrino, F., Colangelo, D., 2010. Semi-parametric forecasts of the implied volatility surface using regression trees. *Statistics and Computing* 20, 421–434.

Baek, S., Mohanty, S. K., Glambosky, M., 2020. COVID-19 and stock market volatility: An industry level analysis. *Finance Research Letters* 37, 101748.

Bakshi, G., Cao, C., Chen, Z., 1997. Empirical performance of alternative option pricing models. *Journal of Finance* 52, 2003–2049.

Bakshi, G., Cao, C., Chen, Z., 2000. Pricing and hedging long-term options. *Journal of Econometrics* 94, 277–318.

Bakshi, G., Kapadia, N., 2003. Delta-hedged gains and the negative market volatility risk premium. *The Review of Financial Studies* 16, 527–566.

Bali, T. G., Beckmeyer, H., Moerke, M., Weigert, F., 2023. Option return predictability with machine learning and big data. *The Review of Financial Studies* 36, 3548–3602.

Bali, T. G., Hovakimian, A., 2009. Volatility spreads and expected stock returns. *Management Science* 55, 1797–1812.

Barone-Adesi, G., Whaley, R. E., 1987. Efficient analytic approximation of American option values. *Journal of Finance* 42, 301–320.

Basak, S., Kar, S., Saha, S., Khaidem, L., Dey, S. R., 2019. Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance* 47, 552–567.

Bates, D. S., 1991. The crash of 87: was it expected? The evidence from options markets. *Journal of Finance* 46, 1009–1044.

Black, F., Scholes, M., 1973. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81, 637–654.

Boloorforoosh, A., Christoffersen, P., Fournier, M., Gouriéroux, C., 2020. Beta risk in the cross-section of equities. *The Review of Financial Studies* 33, 4318–4366.

Boyer, B., Mitton, T., Vorkink, K., 2010. Expected idiosyncratic skewness. *The Review of Financial Studies* 23, 169–202.

Boyer, B. H., Vorkink, K., 2014. Stock options as lotteries. *Journal of Finance* 69, 1485–1527.

Boyle, P. P., 1986. Option valuation using a tree-jump process. *International Options Journal* 3, 7–12.

Breen, R., 1991. The accelerated binomial option pricing model. *Journal of Financial and Quantitative Analysis* 26, 153–164.

Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32.

Broadie, M., Detemple, J., 1996. American option valuation: new bounds, approximations, and a comparison of existing methods. *The Review of Financial Studies* 9, 1211–1250.

Broadie, M., Glasserman, P., 1997. Pricing American-style securities using simulation. *Journal of Economic Dynamics and Control* 21, 1323–1352.

Buehler, H., Gonon, L., Teichmann, J., Wood, B., 2019. Deep hedging. *Quantitative Finance* 19, 1271–1291.

Burkovska, O., Gaß, M., Glau, K., Mahlstedt, M., Schoutens, W., Wohlmuth, B., 2018. Calibration to American options: numerical investigation of the de-Americanization method. *Quantitative Finance* 18, 1091–1113.

Byun, S.-J., Kim, D.-H., 2016. Gambling preference and individual equity option returns. *Journal of Financial Economics* 122, 155–174.

Cao, J., Han, B., 2013. Cross section of option returns and idiosyncratic stock volatility. *Journal of Financial Economics* 108, 231–249.

Carhart, M. M., 1997. On persistence in mutual fund performance. *Journal of Finance* 52, 57–82.

Chang, B.-Y., Christoffersen, P., Jacobs, K., Vainberg, G., 2012. Option-implied measures of equity risk. *Review of Finance* 16, 385–428.

Chang, Y.-C., Chang, K.-H., Wu, G.-J., 2018. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing* 73, 914–920.

Chen, D., Guo, B., Zhou, G., 2023. Firm fundamentals and the cross-section of implied volatility shapes. *Journal of Financial Markets* 63, 100771.

Chen, H.-Y., Lee, C.-F., Shih, W., 2010. Derivations and applications of Greek letters: Review and integration. *Handbook of Quantitative Finance and Risk Management* pp. 491–503.

Chen, J., Hong, H., Stein, J. C., 2001. Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of Financial Economics* 61, 345–381.

Chen, L., Pelger, M., Zhu, J., 2024. Deep learning in asset pricing. *Management Science* 70, 714–750.

Chiras, D. P., Manaster, S., 1978. The information content of option prices and a test of market efficiency. *Journal of Financial Economics* 6, 213–234.

Claessen, H., Mittnik, S., 2002. Forecasting stock market volatility and the informational efficiency of the DAX-index options market. *The European Journal of Finance* 8, 302–321.

Conrad, J., Dittmar, R. F., Ghysels, E., 2013. Ex ante skewness and expected stock returns. *Journal of Finance* 68, 85–124.

Conrad, J., Kapadia, N., Xing, Y., 2014. Death and jackpot: Why do individual investors hold overpriced stocks? *Journal of Financial Economics* 113, 455–475.

Corrado, C. J., Su, T., 1996. Skewness and kurtosis in S&P 500 index returns implied by option prices. *Journal of Financial Research* 19, 175–192.

Cox, J. C., Ross, S. A., Rubinstein, M., 1979. Option pricing: A simplified approach. *Journal of Financial Economics* 7, 229–263.

Crank, J., Nicolson, P., 1947. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, vol. 43, pp. 50–67.

Cremers, M., Weinbaum, D., 2010. Deviations from put-call parity and stock return predictability. *Journal of Financial and Quantitative Analysis* pp. 335–367.

De Spiegeleer, J., Madan, D. B., Reyners, S., Schoutens, W., 2018. Machine learning for quantitative finance: fast derivative pricing, hedging and fitting. *Quantitative Finance* 18, 1635–1643.

Del Viva, L., Kasanen, E., Trigeorgis, L., 2017. Real options, idiosyncratic skewness, and diversification. *Journal of Financial and Quantitative Analysis* 52, 215–241.

Dennis, P., Mayhew, S., 2002. Risk-neutral skewness: Evidence from stock options. *Journal of Financial and Quantitative Analysis* 37, 471–493.

Dimson, E., 1979. Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics* 7, 197–226.

Doran, J. S., Peterson, D. R., Tarrant, B. C., 2007. Is there information in the volatility skew? *Journal of Futures Markets: Futures, Options, and Other Derivative Products* 27, 921–959.

Duan, J.-C., Popova, I., Ritchken, P., 2002. Option pricing under regime switching. *Quantitative Finance* 2, 116.

Duan, J.-C., Simonato, J.-G., 2001. American option pricing under GARCH by a Markov chain approximation. *Journal of Economic Dynamics and Control* 25, 1689–1718.

Dumas, B., Fleming, J., Whaley, R. E., 1998. Implied volatility functions: Empirical tests. *Journal of Finance* 53, 2059–2106.

Fama, E. F., French, K. R., 1996. Multifactor explanations of asset pricing anomalies. *Journal of Finance* 51, 55–84.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116, 1–22.

Feng, G., He, J., Polson, N. G., Xu, J., 2018. Deep learning in characteristics-sorted factor models. *Journal of Financial and Quantitative Analysis* pp. 1–36.

Figlewski, S., 1989. Options arbitrage in imperfect markets. *Journal of Finance* 44, 1289–1311.

Gan, L., Wang, H., Yang, Z., 2020. Machine learning solutions to challenges in finance: An application to the pricing of financial products. *Technological Forecasting and Social Change* 153, 119928.

Gao, P.-w., 2009. Options strategies with the risk adjustment. *European Journal of Operational Research* 192, 975–980.

Gençay, R., Qi, M., 2001. Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping, and bagging. *IEEE Transactions on Neural Networks* 12, 726–734.

Geske, R., 1979. The valuation of compound options. *Journal of Financial Economics* 7, 63–81.

Gradojevic, N., Gençay, R., Kukolj, D., 2009. Option pricing with modular neural networks. *IEEE Transactions on Neural Networks* 20, 626–637.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33, 2223–2273.

Hagan, P. S., Kumar, D., Lesniewski, A. S., Woodward, D. E., 2002. Managing smile risk. *The Best of Wilmott* 1, 249–296.

Han, C., 2021. Bimodal characteristic returns and predictability enhancement via machine learning. *Management Science* .

Han, C., He, Z., Toh, A. J. W., 2023. Pairs trading via unsupervised learning. *European Journal of Operational Research* 307, 929–947.

Hanley, J. A., McNeil, B. J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36.

Hansen, P. R., Lunde, A., Nason, J. M., 2011. The model confidence set. *Econometrica* 79, 453–497.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. *IEEE Intelligent Systems and Their Applications* 13, 18–28.

Heston, S. L., 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies* 6, 327–343.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* 2, 359–366.

Hutchinson, J. M., Lo, A. W., Poggio, T., 1994. A nonparametric approach to pricing and hedging derivative securities via learning networks. *Journal of Finance* 49, 851–889.

Hutton, A. P., Marcus, A. J., Tehranian, H., 2009. Opaque financial reports, R2, and crash risk. *Journal of Financial Economics* 94, 67–86.

Ivașcu, C.-F., 2021. Option pricing using machine learning. *Expert Systems with Applications* 163, 113799.

Jang, H., Lee, J., 2018. Generative Bayesian neural network model for risk-neutral pricing of American index options. *Quantitative Finance* 19, 587 – 603.

Jang, J., Kang, J., 2019. Probability of price crashes, rational speculative bubbles, and the cross-section of stock returns. *Journal of Financial Economics* 132, 222–247.

Jensen, T. I., Kelly, B. T., Pedersen, L. H., 2021. Is there a replication crisis in finance? Tech. rep., National Bureau of Economic Research.

Jiang, L., Dai, M., 2004. Convergence of binomial tree methods for European/American path-dependent options. *SIAM Journal on Numerical Analysis* 42, 1094–1109.

Kahle, K. M., Shastri, K., 2005. Firm performance, capital structure, and the tax benefits of employee stock options. *Journal of Financial and Quantitative Analysis* 40, 135–160.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pp. 3146–3154.

Lajbcygier, P. R., Connor, J. T., 1997. Improved option pricing using artificial neural networks and bootstrap methods. *International Journal of Neural Systems* 8, 457–471.

Latane, H. A., Rendleman, R. J., 1976. Standard deviations of stock price ratios implied in option prices. *Journal of Finance* 31, 369–381.

Liang, X., Zhang, H., Xiao, J., Chen, Y., 2009. Improving option price forecasts with neural networks and support vector regressions. *Neurocomputing* 72, 3055–3065.

Liu, S., Oosterlee, C. W., Bohte, S. M., 2019. Pricing options and computing implied volatilities using neural networks. *Risks* 7, 16.

Longstaff, F. A., Schwartz, E. S., 2001. Valuing American options by simulation: a simple least-squares approach. *The Review of Financial Studies* 14, 113–147.

Madhu, B., Rahman, M. A., Mukherjee, A., Islam, M., Roy, R., Ali, L. E., 2021. A Comparative Study of Support Vector Machine and Artificial Neural Network for Option Price Prediction. *Journal of Computational Chemistry* 09, 78–91.

Ni, S. X., Pan, J., Poteshman, A. M., 2008. Volatility information trading in the option market. *Journal of Finance* 63, 1059–1091.

Ofek, E., Richardson, M., Whitelaw, R. F., 2004. Limited arbitrage and short sales restrictions: Evidence from the options markets. *Journal of Financial Economics* 74, 305–342.

Papahristodoulou, C., 2004. Option strategies with linear programming. *European Journal of Operational Research* 157, 246–256.

Park, H., Kim, N., Lee, J., 2014. Parametric models and non-parametric machine learning models for predicting option prices: Empirical comparison study over KOSPI 200 Index options. *Expert Systems with Applications* 41, 5227–5237.

Petersen, M. A., 2008. Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies* 22, 435–480.

Roll, R., 1977. An analytical valuation formula for unprotected american call options. *Journal of Financial Economics* 5, 251–258.

Rubinstein, M., 1983. Displaced diffusion option pricing. *Journal of Finance* 38, 213–217.

Rubinstein, M., 1994. Implied binomial trees. *Journal of Finance* 49, 771–818.

Ruf, J., Wang, W., 2019. Neural networks for option pricing and hedging: a literature review. *arXiv preprint arXiv:1911.05620* .

Ruf, J., Wang, W., 2022. Hedging with linear regressions and neural networks. *Journal of Business & Economic Statistics* 40, 1442–1454.

Samimi, O., Mardani, Z., Sharafpour, S., Mehrdoust, F., 2017. LSM algorithm for pricing American option under Heston–Hull–White's stochastic volatility model. *Computational Economics* 50, 173–187.

Schneider, P., Wagner, C., Zechner, J., 2020. Low-risk anomalies? *Journal of Finance* 75, 2673–2718.

Subramanian, A., 2004. Option pricing on stocks in mergers and acquisitions. *Journal of Finance* 59, 795–829.

Sun, X., Liu, M., Sima, Z., 2020. A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters* 32, 101084.

Tay, F. E., Cao, L., 2001. Application of support vector machines in financial time series forecasting. *Omega* 29, 309–317.

Thompson, S. B., 2011. Simple formulas for standard errors that cluster by both firm and time. *Journal of Financial Economics* 99, 1–10.

Toft, K. B., Prucyk, B., 1997. Options on leveraged equity: Theory and empirical tests. *Journal of Finance* 52, 1151–1180.

Trigeorgis, L., Lambertides, N., 2014. The role of growth options in explaining stock returns. *Journal of Financial and Quantitative Analysis* 49, 749–771.

Van Buskirk, A., 2009. Implied volatility skew and firm-level tail risk. *Unpublished paper, University of Chicago* .

Vasquez, A., Xiao, X., 2023. Default risk and option returns. *Management Science* .

Villiger, R., 2006. Valuation of American call options. *Wilmott Magazine* 3, 64–67.

Wang, Y.-H., 2009. Using neural network to forecast stock index option price: a new hybrid GARCH approach. *Quality & Quantity* 43, 833–843.

Whaley, R. E., 1981. On the valuation of American call options on stocks with known dividends. *Journal of Financial Economics* 9, 207–211.

Xing, Y., Zhang, X., Zhao, R., 2010. What does the individual option volatility smirk tell us about future equity returns? *Journal of Financial and Quantitative Analysis* pp. 641–662.

Zhan, T., Fang, L., Xu, Y., 2017. Prediction of thermal boundary resistance by the machine learning method. *Scientific Reports* 7, 7109.

Zhan, X., Han, B., Cao, J., Tong, Q., 2022. Option return predictability. *The Review of Financial Studies* 35, 1394–1442.

Zhang, D., Hu, M., Ji, Q., 2020. Financial markets under the global pandemic of COVID-19. *Finance Research Letters* 36, 101528.

Zheng, Y., Yang, Y., Chen, B., 2019. Gated deep neural networks for implied volatility surfaces. *arXiv preprint arXiv:1904.12834* 7.