CATÓLICA
LISBON
BUSINESS & ECONOMICS

Earnings Prediction using Machine Learning Methods and Analyst Comparison

Alexandre Inês Martins

Dissertation written under the supervision of Dan Tran

Dissertation submitted in partial fulfilment of requirements for the MSc in Finance, at the Universidade Católica Portuguesa, January 2022.

Earnings Prediction using Machine Learning Methods and Analyst Comparison

Alexandre Inês Martins

Abstract

In the course of this dissertation we propose an experimental study on how technical, macroeconomic, and financial variables, alongside analysts' forecasts, can be used to optimize the prediction for the subsequent quarter's earnings results using machine learning, comparing the performance of the models to analysts' forecasts. The dissertation includes three steps. In step one, an event study is conducted to test abnormal returns in firms' stock prices in the day following earnings announcement, grouped by earnings per share (EPS) growth in classes of size 3, 6 and 9, computed for each quarter. In step two, several machine learning models are built to maximize the accuracy of EPS predictions. In the last step, investment strategies are constructed to take advantage of investors' expectations, which are closely correlated with analysts' predictions. In the backdrop of an exhaustive analysis on quarterly earnings predictions using machine learning methods, conclusions are drawn related to the superiority of the CatBoost classifier. All machine learning models tested underperform analyst predictions, which could be explained by the time and privileged information at analysts' disposal, as well as their selection of firms to cover. Regardless, machine learning models can be used as a confirmation for analyst predictions, and statistically significant investment strategies are pursued with those fundamentals. Importantly, high confidence predictions by machine learning models are significantly more accurate than the average accuracy of forecasts.

Keywords: Earnings Announcements; Analyst errors; Event Study; machine learning; Technical Analysis;

Previsão the anuncios de resultados financeiros utilizando modelos de Machine Learning e subsequentemente comparando com previsões de analistas.

Alexandre Inês Martins

Resumo

No decorrer desta dissertação, realiza-se um estudo experimental sobre a forma como análises técnicas, macroeconómicas, fundamentais e as previsões dos analistas podem ser utilizadas em conjunto para otimizar a previsão dos resultados de lucros do próximo trimestre de empresas A dissertação inclui três etapas. Na primeira etapa, é efetuado um estudo de evento para testar os retornos anormais nas ações no dia seguinte aos anúncios de lucros, sendo estes agrupados pelo crescimento do lucro por ação nas classes de 3, 6 e 9, calculado para cada trimestre. Na etapa dois, vários modelos de machine learning (ML) são concebidos para maximizar a precisão das previsões de crescimento de lucros de empresas. Na última etapa, estratégias de investimento são construídas para tirar proveito das expectativas do investidor, que estão relacionadas com as previsões dos analistas. Uma vez que um dos projetos de pesquisa mais exaustivos sobre previsões de lucros para o próximo trimestre, conclusões podem ser retiradas relacionadas com a superioridade do modelo CatBoost nas previsões de lucros. Todos os modelos de testados apresentam desempenho inferior às previsões dos analistas, o que pode ser explicado pelo tempo e pelas informações privilegiadas a que os analistas têm acesso, bem como pela escolha da empresa sob a qual as suas previsões incidem. Os modelos de podem ser utilizados como uma confirmação para as previsões dos analistas criando estratégias de investimento estatisticamente significativas. Além disso, as previsões com alta confiança por modelos de são mais precisas do que a precisão média das previsões dos analistas.

Palavras-chave: Anúncio de Resultados; Erros dos analistas; Estudos de Eventos; Machine Learning; Análise Tecnica;

Acknowledgements

“No man ever steps in the same river twice, for it's not the same river and he's not the same man.” said Heraclitusin in the old days. It is not less true than it used to be, it is not truer, just intemporal as it is this small piece of research, forever carved in the river of knowledge.

First of all, I want to thank my parents and grandparents for their true support in all moments, especially in the difficult ones when time was scarce.

Also, to my girlfriend, Ana Sofia, for the unconditional support on all occasions.

Second, I would like to thank my great friends (Diogo Santos and Bernardo Garcês) for all the fun times and more stressed ones that we went true together during this beautiful journey and for the great conversations, ones more serious than others.

Third, I would like to thank to all my professors for all the knowledge that I could learn with then. I would also like to thank Professor José Faias for the support during all the masters and for helping me to develop better Python skills. A special thanks to my advisor Dan Tran for all the help and for introducing me to the world of ML.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation and Contextual Analysis

The predictability of events is one of the most popular topics of research in the field of finance, and the predictability of earnings is no exception. The company returns in the stock market reflect, at least in the long-term, the firm performance and for that reason, investment banks spend millions of dollars in compensations for analysts to predict earnings every quarter for specific firms that drive the upcoming path of the stock. During the year, there are few days that can really shake someone's portfolio, and earnings announcements are known for that exactly, as one of the events that are responsible for the biggest moves, i.e., for the moments with the highest volatility and where fortunes are made and lost. If prediction models are robust in earnings predictions, then investors can use those models to invest or divest in a specific company before earnings announcements. This can both protect them from extreme events and enable to invest in companies before earnings. This paper focuses on a quarterly timeframe to permit investment strategies, but the models could also be extended to other timeframes to predict earnings of companies trading in the stock market about to be acquired, either by the acquiring firm or for a long-term investment.

According to past research (Easton & Harris, 1991) there are many important earnings accounting figures such as EBIT, earnings per share (EPS), or earnings yield highly correlated to cumulative abnormal returns. In this dissertation, earnings growth is the target variable, due to its higher amount of information compared to the earnings per share metric. EPS depend on the number of shares, on buyback amounts, varying with the number of shares. The same cannot be said about earnings growth, getting from an absolute to relative variable.

In the prediction of earnings literature, analysts' predictions are intensively scrutinized, focusing on topics from the timing of the predictions to analyst error in predictions and its consequences to stock returns. Since the release of the study by Ball and Brown (1968), many explored the relations between earnings accounting figures and stock returns, where analyst estimates of earnings have been a key proxy for EPS. According to Brown et al. (1987) analyst predictions are superior to time-series forecasts due to both time and information advantages. Ever since, the research uses analysts' mean predictions as an estimate for EPS and considers it to be in accordance with investors' expectations. The focus around analysts' predictions shifted to using realized earnings to explain analysts' errors in predictions, to state the consequences of this errors in post-earning returns and whether analysts have a bias towards optimistic or negative predictions. Although it is relevant to understand when and why analysts fail or get right their predictions, all this research is undertaken as a post-event analysis, meaning it will not help investors to predict what will happen in the next earnings to a specific company. For a lack of better alternatives, analysts are still widely used as an estimate of earnings, some may trust a specific analyst more than the others, but most people use the mean analyst prediction. Initially, models like multiple linear regressions and logistic regressions were used to predict earnings, without success compared to analysts. With the increase of the computing power and data in the society, also came machine learning (ML) models, much more complex and diverse models. ML relighted all the debate around a way to outperform analyst predictions with the need to create a real-time unbiased prediction of earnings. Some models predict the sign of earnings, but investors need models that account not only for the sign but also for the magnitude of each earnings report. That's not all, because investors also need to know the probability or confidence of each prediction.

This research focuses on the use of ML models, from the simple Logistic regression to the more complex Neural Networks, in order to create a real-time prediction of earnings. The research does not focus so much on why analysts' predictions are right or wrong or what's the determinants for those distinct moments, but the confirmation that analyst errors create significant abnormal returns. Therefore, this research is conducted with earnings divided in equal size bins and an attempt to catch those errors in the models by creating better forecasts than analysts.

In order to account for the variability of financial information around earnings, this problem is going to be treated as a classification problem with the earnings growth in each period to be divided in 3, 6 and 9 equal sized bins. The decision was taken in order to have more robust ML models, although the change from a regression to classification problem may give more relevance to extreme events (responsible for a clear change of class) and may be difficult to account for small changes that can make the values be just over the beginning of the next bin, without a significant change.

Further detailed hypotheses of this research are explained in the next section, "1.2. Hypothesis and Structure".

1.2. Hypothesis and Structure

1.2.1 Investors expectations and the cross-section of stock returns.

The first main goal of the paper is to prove that investors' expectations are closely related to analyst expectations, which are biased and off-target in many situations. It should be observed that the earnings predictions are not as much correlated with post-earning stock returns. On the other hand, differences between mean analyst predictions and actual earnings results should yield abnormal post-earning returns, which increase with the difference between the two and are consistent for different timeframes.

Hypothesis 1: Differences between analyst predictions and actual earnings yield abnormal returns which increase with the difference between the two.

1.2.2 ML models vs analyst predictions.

If the ML forecast for earnings represents the unbiased earnings prediction, it should present a higher accuracy than mean analyst expectations. Models should not use biased analyst predictions as inputs in order to beat analysts. If analyst present more accurate predictions, then they could contain information that is not present in the markets and their predictions need to be incorporated into ML models.

Hypothesis 2: ML models can be used as a real-time unbiased earnings predictor more accurate than analysts.

1.2.3 ML models & analyst prediction confirmations.

Whether or not more accurate than analysts, this research presents a totally different approach to earnings predictions based on technical variables, macroeconomic variables, and financial ratios variables. Combining this perspective with the analysts' mean perspective should act as a potential confirmation for the analyst predictions.

Hypothesis 3: ML models can be used as a confirmation for analyst predictions, with higher earnings accuracy for when analyst and this research agree on the prediction.

1.2.4 Structure of the research

Regarding the structure of this paper, the next chapter will focus on the literature review, providing an extensive overview of the different methods explored for the topic at hand. Furthermore, in Chapter 3, the dataset, data treatment and data exploration will be reviewed. In Chapter 4, the methodology for both the benchmark models and ML models will be explained. In Chapter 5, the model evaluation metrics will be presented alongside the variable importance metrics. The results of the analysis will be displayed and discussed in Chapter 6. Finally, in Chapter 7 the research will be concluded.

2. Literature Review

2.1. Early Empirical Approaches

Earnings announcements carry important information about the firms. In the 1950s Modigliani-Miller theorem stated that the market value of a company is correctly calculated as the present value of its future earnings and its underlying assets, being independent of its capital structure. As one of the key elements to the constitution of the "fair" price of company stocks it is natural that earnings announcements may change the entire picture around a company future or simply confirm previous predictions. In some cases, the initial reaction to earnings announcements may last several weeks or even months. This anomaly is called Post-Earnings-Announcement drift (PEAD). It describes the drift of a firm's stock price in the direction of the firm's earnings surprise for an extended period of time. Ball and Brown (1968) and Beaver (1968) reported the first well-documented relationship between earnings and stock market reactions, and since then many tried to explain this phenomenon.

The first discussion came in the form of the analysis of analyst predictions. Brown and Rozeff (1978) compared analysts forecasts to univariate time series models for a period from 1 to 4 quarters ahead, observing the overall superiority of analyst predictions. It's also important to state the use of only 50 firms from 1972 through 1975 in the paper. Even the most extensive sample used in research at the time (Fried & Givoly, 1982) accounted for only 424 firms from 1969 through 1979, which is fairly low for today's standards. These choices can be justified by the data requirements of ARIMA models or filters to use only stocks traded on the NYSE, but independently of the reasons it does not change the fact that it's difficult to take broad conclusions from a limited sample of data. Brown et al. (1987) also concluded that analysts' forecasts are superior to time-series forecasts for quarterly estimate windows. Bhushan (1989) identified factors responsible for analysts following from firm size to institutional ownership, all significant, proving again that should not be generalized at this point that analyst have better forecasts than all other models. As an alternative method, Ou and Penman (1989) attempt to predict the sign of earnings changes. Their general research question is whether and to what extent standard financial ratios can be useful for financial statement analysis.

Abarbanell and Bushee (1997) explained the importance of fundamental signals to predict future earnings changes. They also asserted the heterogeneity in the importance of different fundamental signals and the importance of distinguishing relevant fundamental information from non-relevant information. Myers et al. (2007) explore the momentum properties of company earnings, by computing the probability of random binomial variables through time-series models to predict split-adjusted EPS. Fundamental data in the form of financial ratios and momentum variables are going to be applied to the research.

Hess and Kreutzmann (2010) provide evidence that unexpected macroeconomic news are captured when individual analysts revise their earnings forecasts, which implies that analysts use macroeconomic information when forecasting. Shu et al. (2013), using data related to U.S. firms over the period from 1962 to 2009, indicates that when predicting the future earnings of firms, the predictive accuracy of model-based approaches is improved by incorporating macroeconomic information. Although it's stated that the effect is much stronger for longer horizons and this research only uses a short-term horizon, macroeconomic variables are going to be used, nonetheless.

Although the focus of this research are quarterly predictions, analysts' forecasts have been used as a proxy for earnings even for long-term timeframes, as long as 2 to 5 years ahead, without proper prior research. The same applies to small firms that were almost not present in the early literature and to which analysts' predictions are used as well. All these papers misuse of analyst predictions gives even more credit to them without proof of the cause. In research from Bradshaw et al. (2012) it was reexamined whether time-series forecasts could outperform analysts. Surprisingly, it was found that time series forecasts provide the most accurate estimate of long-term (2- and 3-year-ahead) earnings predictions. One can ask if more advanced models can also outperform analysts in the short-term?

The basic tenor of results from many prior studies is that analyst forecasts are predictably biased and forecast bias appears consistent with several stock price anomalies.

2.2. ML Models

As computing power and ML techniques have advanced drastically, allowing researchers to examine whether additional independent variables and more computer intensive methodologies might be useful to predict future earnings is of crucial importance. ML models give a flexibility in the structure of the models completely opposite to the linear regression models used in previous research. The higher the complexity of the model the harder it is to interpret it beyond accuracy and other basic measures. Earnings prediction is a problem of supervised leaning, in which the data is trained in the model with each observation in order to get the best predictive model. Several ML techniques were tried in the past to predict earnings.

Research from Zhang et al. (2004) showed the that the application of the neural network approach incorporating fundamental accounting variables results in forecasts that are more accurate than linear forecasting models for 1-quarter ahead forecasts. To estimate the neural network weights of their neural network models, Zhang et al. (2004) used backward propagation (BP). Cao and Parry (2009) found that the genetic algorithm produces models that are significantly more accurate than the models examined by Zhang et al. (2004) using backward propagation. According to more recent research (Etemadi et al., 2014), rule extraction from neural network by genetic algorithm technique is significantly more accurate than multi-layer perceptron (MLP). There is no comparation between these models and analysts or other ML models as it was not the focus of the papers, giving less relevance to the research as it would be expected that complex neural networks would outperform linear forecasting in the first place.

According to new research (Fischer et al., 2020), support vector machine performs better than BR ARIMA developed in Brown and Rozeff (1979) as the premier univariate statistical model for the prediction of quarterly earnings. Again, there is lack of comparison between this model and other ML models.

In recent research, van Binsbergen et al. (2020) introduces a real-time measure of conditional biases in firms' earnings forecasts for different horizons. The measure is defined as the difference between analysts' expectations and a statistically optimal unbiased machine-learning benchmark. Van Binsbergen et al. (2020) uses Random Forest to predict earnings and analyst forecasts as inputs of the model, following the approach of creating a conditional prediction of earnings starting with analyst predictions until getting to an unbiased prediction. They observed the superiority of their models for different timeframes (from 1 quarter ahead to 2 years).

Xinyue et al. (2020) decided to investigate the use of LightGBM (a Gradient Boost Decision Tree model introduced by Microsoft in 2017) in earnings prediction. The paper compares earnings predictions using LightGBM (without analyst prediction data as inputs) to analyst predictions for 1-quarter and 1-year ahead forecasts. Contrary to previous papers, they chose to transform the problem into a qualitative one, transforming analyst predictions and actual earnings into bins of different size (3, 6 and 9) and comparing accuracies. The paper concluded the superiority of analyst predictions, but also observed that when LightGBM and analysts have the same prediction for earnings, accuracy is much higher. LightGBM can be then used as a confirmation for earnings predictions and for the prediction of the earnings sign, also computed during the research for a limited sample. Xinyue et al. (2020) also has its downfalls, with only one ML model used and without a deeper analysis to each prediction class to check the model performance side by side to analysts.

Hunt et al. (2019) uses a non-parametric ML technique, random forest, to predict the sign of earnings changes in annual returns. They find that Random Forest method significantly improves out-of-sample forecast accuracy and that these forecasts are useful to generate abnormal returns.

While the application of ML techniques is becoming increasingly popular in finance and in earnings predictions, only one article created a real time variable to predict forecast errors and outperforming analysts at the same time. Other papers use several methods that outperform linear models in earnings forecasts, but with no mention of analysts. Most papers do not use more than one ML model, justifying the choice with broad comparisons between models' performance while it should be the opposite, problem specific. Alternative approaches can be found, with papers predicting the sign of the change in earnings. This is important, but is only one part of earnings prediction, with the other being the magnitude of the change. Accurate earnings predictions are of the upmost importance.

3. Data

3.1 Data Retrieval

For the purpose of this paper, the US stock market is examined over a 20-year time span, from the end of the third quarter of 2000 to the end of 2020, and the sample only includes firms traded on NASDAQ, NYSE and AMEX, excluding others. Financial input variables are retrieved with quarterly frequency from Financial Ratios Firm Level by WRDS, namely Price/Earnings, Long-term Debt/Invested Capital, Cash Ratio. Analyst EPS estimates are retrieved from IBES database, as well as actual EPS values. Macroeconomic variables are taken from FRED (Economic Research Federal Reserve Bank of St. Louis). CRSP is used to retrieve daily stock returns, closing price, daily volume, number of shares outstanding and sic codes. Moreover, Kenneth French Data Library is used to retrieve risk-free rates and Fama-French Factors (Market Risk Premium, Size, Value, Operating Profitability and Investment), to run regressions presented in later sections. Beta Suite by WRDS is used to retrieve the betas of the Fama French 3 Factors, as well as the idiosyncratic volatilities (ivol) and alphas associated with the models for the US market. Those WRDS computations assumed an estimation window of 252 days and a minimum window of 126 days.

3.2 Variable Construction

In order to make an extensive analysis into earnings predictions past financial variables, macroeconomic variables and market movement variables are used.

3.2.1 Financial Variables

For each quarter earnings of each firm, 31 financial ratios are retrieved corresponding to the ratios of the firm in the corresponding quarter of the previous year, lagging 4 quarters to the present. To add to that, an additional 6 financial ratios for the previous quarter are also retrieved. These financial ratios were chosen to include a complete diversity of the firm fundamentals, such as cash flows, dividends, margins, return on assets, inventory, debt, interest and current ratio. In order to choose the variables to lag one quarter, the correlation between all financial variables from t-4 and EPS growth is computed, choosing the 6 most correlated variables to construct a similar measure, but only lagged one quarter. A complete set of variable definitions can be found in appendix 1.

3.2.2 Macroeconomic Variables

The economy can have an overwhelming effect on company earnings, in some industries more than another's and in specific periods. Economic data is most of the times only known in the quarter after the occurrence, same as EPS. For that reason, all macroeconomic data used in the research is lagged one period, for the information available at the moment of the prediction to be the same as the one at the public hands of investors and analysts. These variables are:

Variable	Computation
GDP	US Personal Consumption Expenditures: Durable Goods (PCDG) data from t-1
PC	US GDP data from t-1

***Table 1 | Macroeconomic variables used and their computation**: In this table the construction of the variable is described alongside their names.*

3.2.3 Market Movement Variables

The market may have more information than past information about the firm financials or the economy. Then, it is imperative to have information related to the stock market movements before company earnings. This research uses the ones detailed in table 2:

Variable	Computation
volume	5 trading days cumulative volume (from day 6 to day 1)/ number of shares outstanding
pre_earnigs_return	5 trading days cumulative return (from day 6 to day 1)

***Table 2 \ Market Movement variables used and their computation**: In this table the construction of the variable is described alongside their names.*

The day of the report is not considered, having too much volatility, making it difficult to drive conclusions from the data.

3.2.4 Dependent Variable and Analyst's variables

The models are defined to use a modificated version of EPS growth as the dependent variable, defined as EPS in quarter t minus EPS in quarter t-1, all divided by the absolute value of EPS in quarter t-1. EPS growth is defined that way to capture both the effect of the magnitude of the earnings announcement and the direction of the change. For instance, a firm with a negative EPS in quarter t-1 of -10 followed by a change to a positive EPS value of 5 in quarter t would give a negative EPS growth of -300% when in fact the result would be extremely positive for the firm in that scenario. The modified growth formula would give a contrasting result of 300% growth. On all occasions referring to EPS growth in this research, in fact, the modified EPS growth is always the one being referred to.

After the computation of EPS growth, this variable is then divided in 3, 6 and 9 equal size bins in each quarter, based on the values of EPS growth of each firm in each quarter. All firms with EPS information and analysts' predictions are considered at this stage. The final sample has different bin sizes, because even though some EPS do not have other crucial information and were deleted after, bins were still computed on this stage not to create a bias towards firms with more information. This way the class of earnings used are even closer to the true classes for all observations. The different bin size EPS variables are used in the respective prediction for different class sizes. The same is done with analyst predictions. First, analyst predictions done until the end of each quarter are taken into account and used to compute the mean EPS analyst forecast. From there, in each quarter, it's subtracted the EPS in the quarter before, dividing all by the absolute value of EPS in quarter t-1. Then the values are divided in similar class sizes to the actual EPS growth, meaning in a prediction of 6 classes all variables divided in bins have 6 classes. Also related to EPS, the values for EPS growth from quarter t-1 and quarter t-4 are also computed (using the absolute value of EPS growth in t-2 and t-5 respectively in the division part of the growth formula). The difference between the actual earnings bin and analyst prediction bin in the previous quarter is also computed.

3.3 Data Treatment

Variable standardisation is a common practice in ML, it is normally done by subtracting the mean and dividing by the unit variance, though this is prone to some outlier influence. Thus, all variables suffer a process named Standardisation using the Scikit-Learn Scale method. The analysis for all models is conducted on a four-year rolling window basis. With this, 68 overlapping out-off-sample windows are analysed, and the metrics' results are averaged for each model. In order not to contaminate the train set with test set information, the variable standardization is done on the train set first and then the parameters of the procedure are used on the test set.

Variable deletion was used on variables with over 75% missing rate (dpr_t-4, Intcov_ratio_t-4, Fcf_ocf_t-4, Int_debt_t-4, efftax_t-4). Although some high missing variables can be significant to earnings prediction, it would be a waste of resources to investigate repetitive missing values due to multiple formats. To be able to make the predictions and for the data to be comparable, only firms with Q1 to Q4 ending in month 3,6, 9 and 12 of each year are considered. All earnings reports before market open are considered as if the report was announced in the previous trading day after market close. The price of the firm stock needs to be higher than 5 in the end of the quarter that is going to be reported in order to be considered in the analysis. Dividend variable was turned into a dummy equal to 1 if the firm pays dividends, 0 otherwise. All observations with missing information remaining are not considered in the analysis. Then a correlation matrix is computed. Variables with high correlation with each other and low correlation to the dependent variable are deleted. In total 8 variables were deleted (ps_t-4, Debt_at_t-4, Debt_invcap_t-4, GProf_t-4, Invt_act_t-4, Cash_debt_t-4, Totdebt_invcap_t-4, Debt_capital_t-4). The remaining variables correlation matrix can be found in appendix 2. In the end there were 59715 observations, with an average of 711 earnings studied per quarter. Those earnings were associated with 2663 firms that remained in the analysis, trading on NASDAQ (50.29%), NYSE (48.85%) and AMEX (0,86%).

3.4 Data Exploration

For further understanding of the data at hand, data exploration techniques had to be applied and their results analysed before transforming the EPS into bins. The first method is to conduct a summary statistics analysis which is displayed in Figures 1 and 2.

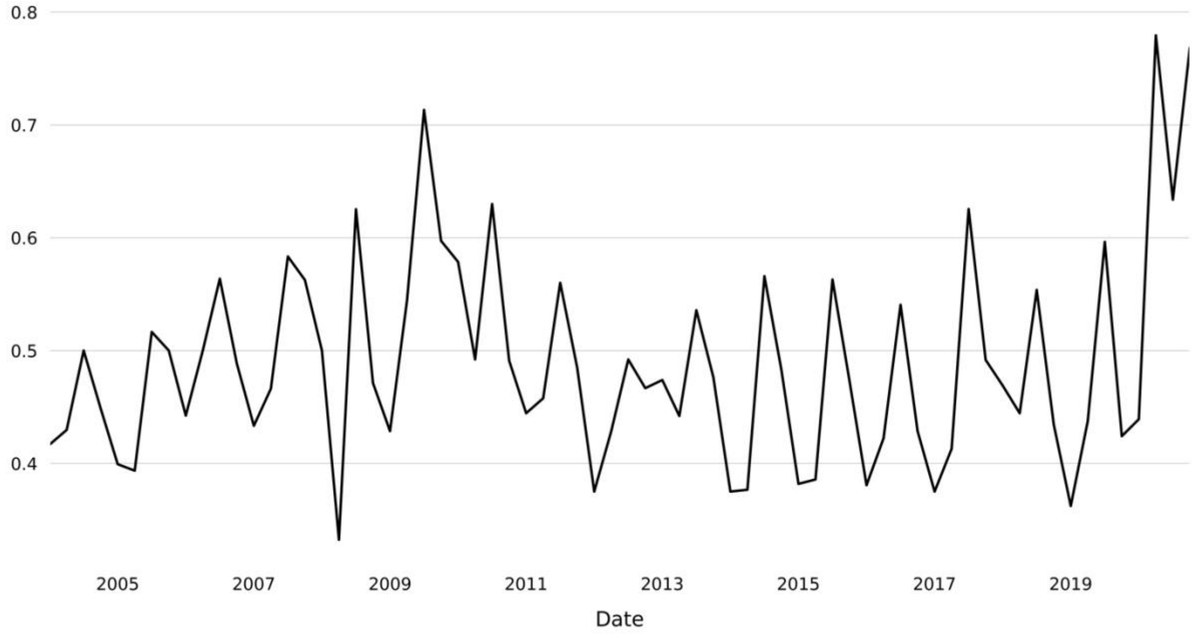


Figure 1 | Analyst Error in EPS Forecast per quarter: *This graph represents the magnitude of analyst errors per quarter in the dataset throughout the out-of-sample time period studied.*

The magnitude of analyst errors consistently along the entire sample for the one quarter ahead shows huge error, with expectations far away from reality, on average. There was a peak in this discrepancy during covid in 2020, in which the errors were at the highest ever. In 2009 during the financial crisis there was another peak at 70% error, but still lower than covid. Although these values seem to be higher during crisis, the lowest analyst error periods show a median miss of about 40%, with an overall mean of 48.98%. This indicates the need for a more accurate earnings prediction than analysts.

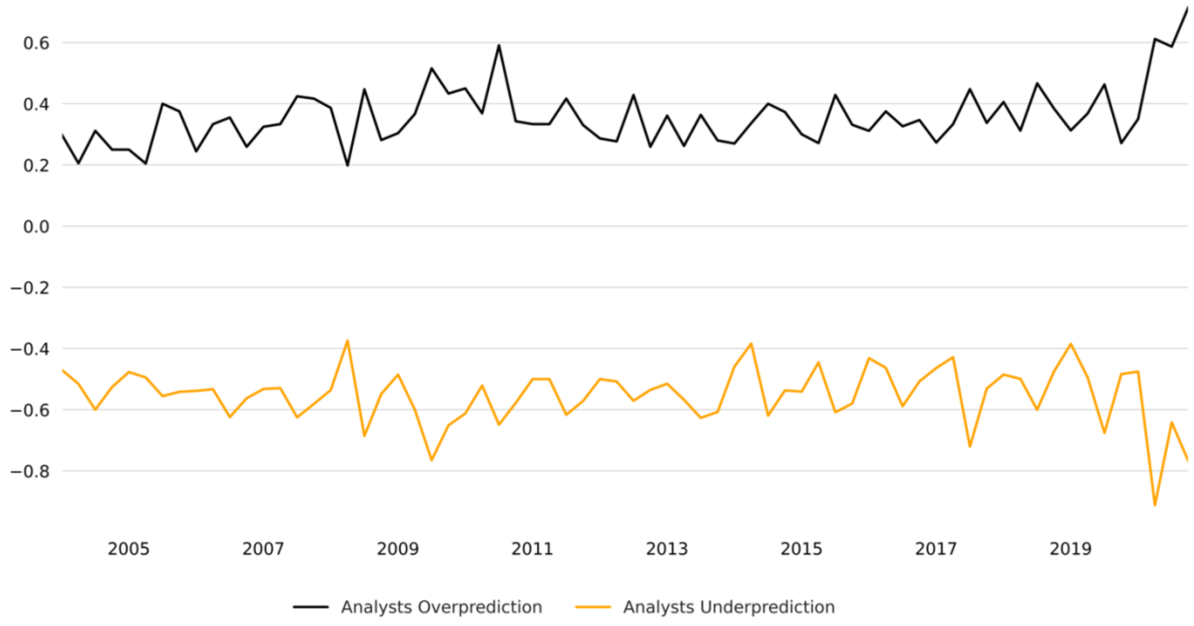


Figure 2 | Analyst Error in EPS Forecast per quarter by over and under estimation: *This graph represents the magnitude of analyst errors per quarter in the dataset throughout the out-of-sample time period studied, when analysts under and over estimate EPS, Under(over)prediction is defined as EPS Analyst predictions below(above) the true values.*

About 2/3 of analyst misses correspond to overestimations (38935 observations), with only 1/3 of observations (20629 observations) presenting as underestimations. For most of the sample the median magnitude of the misses was much higher for underestimations of earnings than overestimation (55.20% and 35.76% respectively), suggesting earnings surprises are much bigger when analysts are proven wrong in the favor of the firm, than when the firms cannot keep up with analysts' expectations.

To further develop the analysis, all EPS variables are analyzed in classes, which is what is used in the analysis. The reference class for the analysis is the class of six bins, although the results for the three and nine classes can also be found in the appendix. The questions remain, are investors expecting analyst errors, since they occur on a frequent basis, or is there a shock between expectations and reality? Are analyst errors more important than EPS growth in the post-earnings return? Is the evolution of EPS growth relevant? To help to answer those questions the correlation between analyst class error/EPS growth/EPS evolution and returns in the trading day after earnings is computed. The linear correlation is high for analyst errors and lower for the class representing earnings growth, even lower for EPS class evolution from the previous quarter (0.1569, 0.1290 and 0.0892 respectively). Then the Spearman non-linear correlation was computed, and it was substantially higher for analyst errors than earnings classes, even lower for EPS class evolution from the previous quarter (0.1530, 0. 1222 and 0.0638), all significant at the 1% significance level.

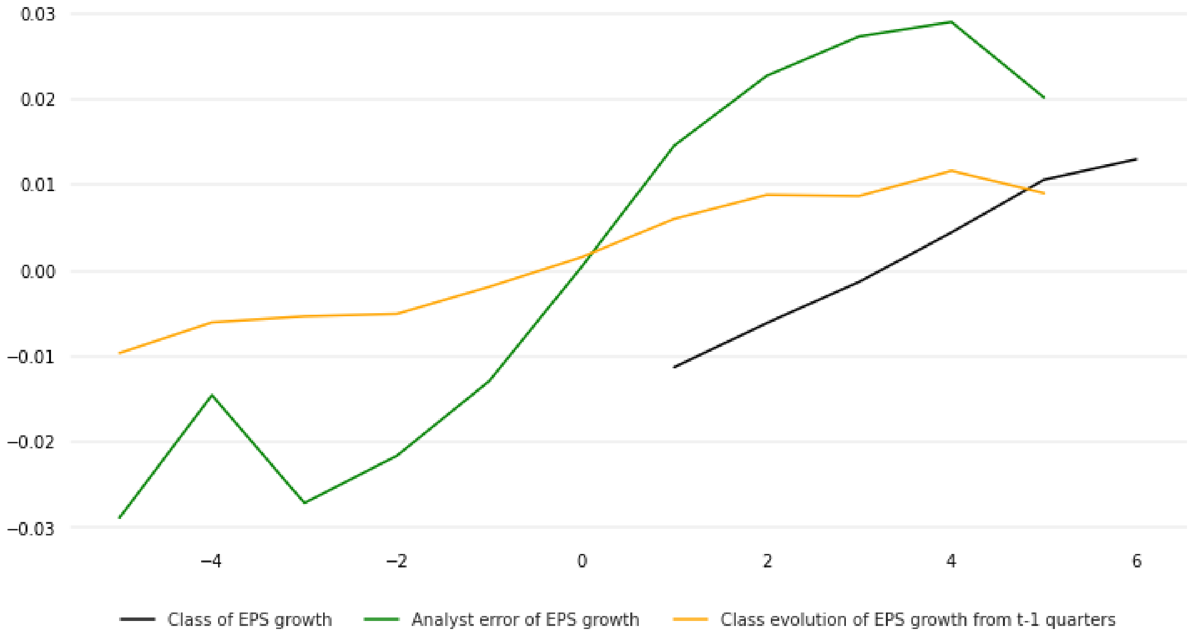


Figure 3 | Return of classes of EPS variables: *This graph represents the average return of each true class of EPS variables after the earnings announcements for all the sample.*

In Figure 3 it is possible to observe the positive effect that higher EPS growth, bigger differences between EPS growth compared to the past and better results than analyst predictions have on the 1-day return in the day after earnings, on average across the entire sample. EPS growth class appears to have the most straightforward relation between the true class and stock returns, although having a low difference of returns from the lowest to highest class. Next the evolution of the EPS class growth from the previous quarter appears to also have a very strong relation with higher returns associated with a positive class progression. The same can be observed for analyst class errors, except with a much higher magnitude of returns showing a clear difference between the lowest and highest classes. The increasing relation from class to class is somewhat broken for when analysts predict 1 and the EPS growth class is 5 and when the difference between analysts' predictions and the EPS growth class is -4. The returns are still in line with class expectations in terms of sign but break the trajectory of the trendline. This could be explained by other factors affecting a huge miss by analysts (e.g., very high growth but close to 0 from a company that started to increase EPS from a very small number could have other factors interfere with returns such as long-term guidance.).

4. Methodology

4.1 Dependent variable type decision & Train/Test procedure

EPS growth classes are the chosen target variable of the research. Focusing on a categorical variable instead of numerical continuous variable may come with advantages and some disadvantages. Financial markets are very volatile, especially around earnings announcements. This instability makes it very challenging to compute a robust earnings forecast, especially without recurring to analyst's predictions as a baseline. Moreover, percentage error of analysts' predictions or EPS growth measures need to take into consideration the overall performance of the other firms. For instance, earnings growth of 2% when the average quarterly firm growth is 2% is nothing extraordinary. If the average was 0.5%, 2% growth would be appreciated. Following the method of dividing earnings into classes introduced by Xinyue et al. (2020) on one hand more focus is given to EPS growth out of the ordinary and on the other hand analyst percentage errors are minimized by the class division. Class of 3 is divided into 1 (underperform), 2 (Neutral) and 3 (overperform). Class of size 6 and 9 do not have names but follow the same logic. In order to be able to train and test time-series data out-of-sample a 4-year rolling window was adopted, illustrated by figure 4:

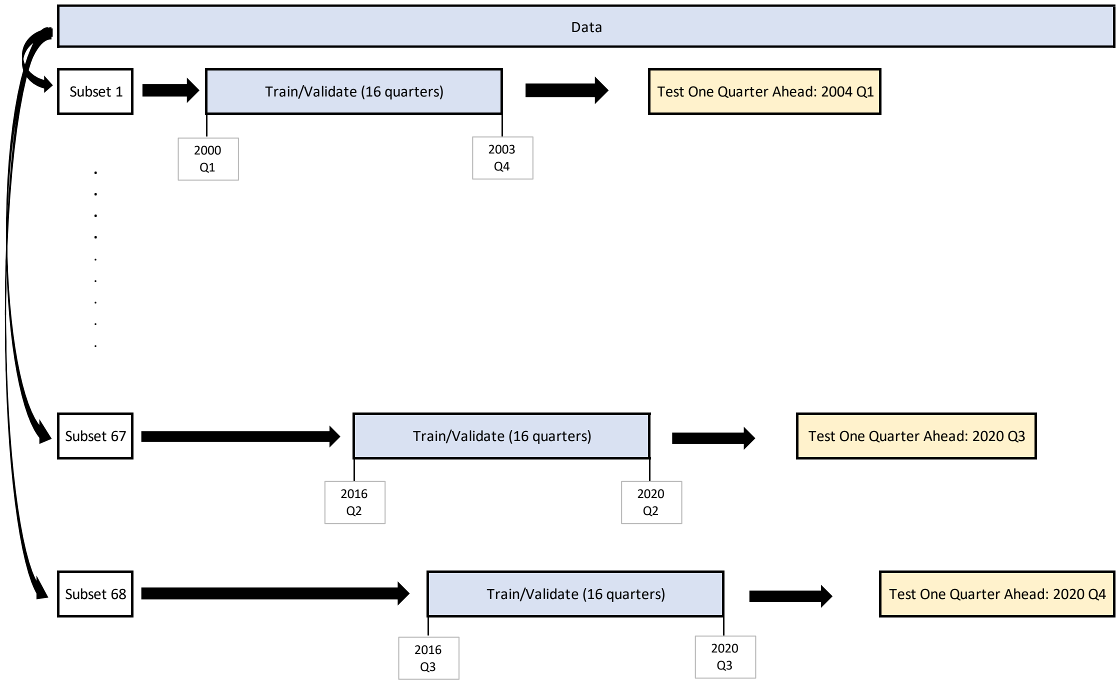

Data
Subset 1
Train/Validate (16 quarters)
Test One Quarter Ahead: 2004 Q1
2000 Q1
2003 Q4
Subset 67
Train/Validate (16 quarters)
Test One Quarter Ahead: 2020 Q3
2016 Q2
2020 Q2
Subset 68
Train/Validate (16 quarters)
Test One Quarter Ahead: 2020 Q4
2016 Q3
2020 Q3

Figure 4 | Rolling-Window illustration: *This figure represents the 68 overlapping Rolling-Windows from the beginning of 2000 to the end of 2020.*

The first rolling window had quarterly data from Q1 2000 to Q4 2003. This data is used to train the ML models to compute the forecast for Q1 2004. Then the window moves one period and the process starts again. The procedure is done until reaching the forecast for 2020 Q4.

4.2 Event Study

In research from Binder (1998) event study methodology and advances of the past 20 years were summarized. In the end, this methodology compares the normal return that a firm would expect in a certain event to the actual return perceived during that period. This research studies Earnings announcements as an event study to understand the significance of the abnormal returns of EPS growth classes compared to analyst error classes to EPS class progression, but more importantly if the aggregation of EPS growth in classes would generate abnormal returns and how significant. Then it is tested how well this variable could explain the returns on the day after earnings announcements. Many papers focus on different time windows since even before the event until some time after. This research only focused on 1 day, the day after the event. There are many models to estimate the abnormal return, but the one chosen is illustrated in the equation bellow:

$$Rai,t = Ri,t - \alpha it - \beta 1 \, (RMt{-}Rft) - \beta 2 \, SMBt - \beta 3 \, HML \tag{4.1}$$

Where Rai,t is the abnormal return, since it is the difference between the actual excess return (Ri,t) and the Fama and French Three Factor Model.

4.3. Introduction to ML Applied in the Research

ML, an application of artificial intelligence (AI), is a category of algorithms that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. The process of learning begins with observations, such as financial ratios observed after earnings, lagged one period and four periods, in this paper. Then the models look for patterns in the data in order to make predictions for the future, such as earnings predictions for the next quarter in the case of this paper. The primary aim is to allow the computer to learn automatically without human intervention or assistance and adjust actions accordingly.

Supervised ML algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly. In contrast, unsupervised ML algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. This thesis focuses, however, on supervised learning as the best method to predict earnings from carefully chosen variables.

All the following models are conducted on a 4-year rolling window analysis with a robust standardisation of the variables.

4.3.1 Shared ML Techniques and Concepts

4.3.1.1 Validation

Validation is a procedure used to evaluate ML models on unseen data. It is essential for model validation and hyperparameter tuning. Its goal is to test the model's ability to predict new data that is not used in estimating it, in order to give insight on how the model will generalize to an independent dataset and find problems such as overfitting. For each different rolling window there is a distinct train, validation, and test sets. The dataset is a time-series model. We decided to use the Hold-Out Validation Mechanism, in which the last quarter of the training set is used exclusively for validation (15 quarters of training, one of validation and one of testing).

4.3.1.2 Hyperparameter Tuning

ML models have parameters, which are the internal coefficients set by training or optimizing the model on a training dataset, but they also have hyperparameters. Hyperparameters are points of configuration that allow a ML model to be customized to guide the learning process for a specific dataset. Further, many ML models have a range of hyperparameters and they may interact in nonlinear ways. An optimization procedure involves defining a search space. This can be thought of geometrically as an n-dimensional volume, where each hyperparameter represents a different dimension and the scale of the dimension are the values that the hyperparameter may take on. The goal of the optimization procedure is to find a vector that results in the best performance of the model after learning, such as maximum accuracy or minimum error. A range of different optimization algorithms may be used, although two of the simplest and most common methods are:

- Random Search. Define a search space as a bounded domain of hyperparameter values and randomly sample points in that domain.
- Grid Search. Define a search space as a grid of hyperparameter values and evaluate every position in the grid.

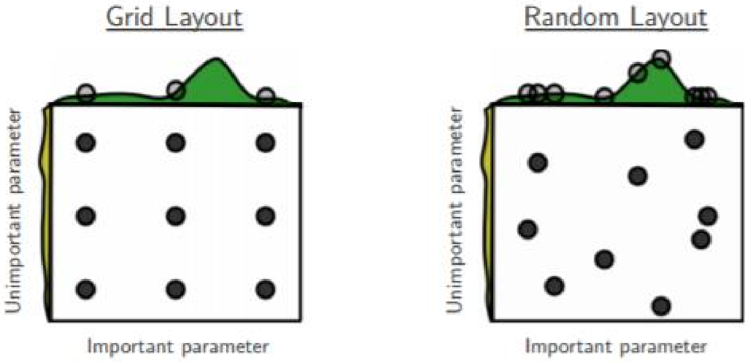


Figure 5 | Grid Search vs Random Search: This figure is a visual representation of the Grid Search and Random Search procedures. From analyticsindiamag.com

The RandomizedSearchCV tool from the Scikit-Learn library is applied for this tuning process. This tool conducts a random search of pre-specified hyperparameter values for a certain model to find the one that results in the best score for a certain pre-determined metric. The metric used for every following hyperparameter tuning is the score. RandomizedSearchCV also incorporates cross-validation into the procedure.

4.3.2 Benchmark

To provide a benchmark for the performance of the models, the Logistic Regression classification model is selected. The selection of this method as a benchmark arises from the fact that it is relatively simple and easy to implement, does not need many hyperparameters tunning and is known to achieve good results in classification tasks.

4.3.2.1 Logistic Classification- Benchmark

Logistic classification is a classification algorithm, used when the value of the target variable is categorical in nature. Logistic classification is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1. Logistic classification, by default, is limited to two-class classification problems. Some extensions like one-vs-rest can allow logistic classification to be used for multi-class classification problems, although they require that the classification problem first be transformed into multiple binary classification problems. Expectations are low and we expect it to be the worst performing model because it was not naturally built for multi class classification. For multinomial classification the loss minimised is the multinomial loss fit across the entire probability distribution. LogisticRegression from library sklearn was used with max_iter equal to 100 (Maximum number of iterations of the optimization algorithm). Using solver equal to 'saga' allows for the benefits of performing better in bigger libraries and still handle multinomial loss. Hyperparameter tuning is performed to penalty and C values. Penalty represents the regularization of the model variables to ensure overfitting does not happen, penalizing the use of unimportant variables to the prediction in the loss function. The main difference between regularization methods is the type of norm used, per table 3:

Regularization Name	**Formula**
L1	$\sum_{i=1}^{N} \lvert wi \rvert$
L2	$\sum_{i=1}^{N} wi^2$
Elasticnet	Combines L1 & L2 methods

***Table 3 \ Regularization**: This table represents the different methods for regularization in the Logistic model.*

4.3.3 KNN

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised ML algorithm that can be used to solve classification problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

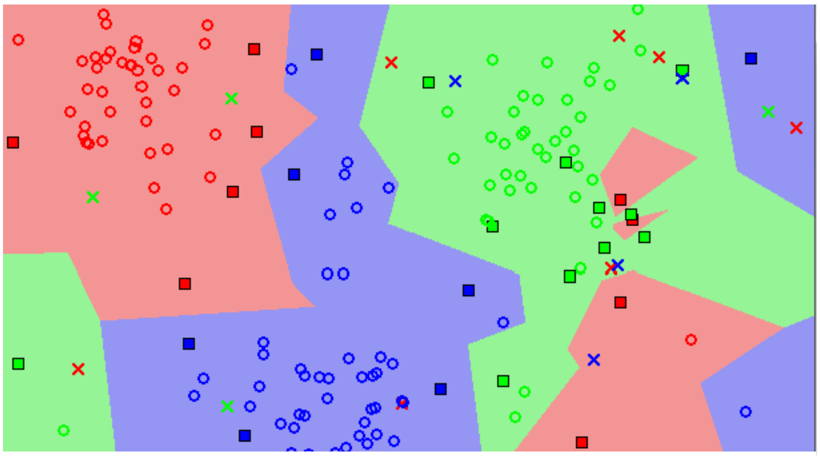


Figure 6 | KNN classification in three different classes: *This figure represents the visual representation of KNN classification, dividing the data into three different classes based on their relative position. From wikipedia*

It can be noticed in the image above that most of the time, similar data points are close to each other. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity, calculating the distance between points on a graph. First, let's understand the working of the KNN classification algorithm. In the classification problem, the K-nearest neighbor algorithm essentially said that for a given value of K algorithm will find the K nearest neighbor of unseen data point and then it will assign the class to unseen data point by having the class which has the highest number of data points out of all classes of K neighbors. KNeighborsClassifier from library sklearn.neighbors is applied to the research. We decided to consider different hyperparameters for n_neighbors, weights, and the metric. Hyperparameter n_neighbors represent the number of neighbors to use by default for queries. A range of values between 1 and 21 is chosen to be tested, with too little not being able to make an accurate distinction either for the three, six or nine classes and too many divisions resulting in overfitting. The weights can be uniform or distance. In the case of uniform weights all points in each neighborhood are weighted equally. Weight points by the inverse of their distance. in this case, closer neighbors of a query point will have a greater influence than neighbors which are further away. For distance metrics, the research will test Euclidean, Manhattan and Minkowski Distance with formulas:

Metric Name	Formula
Euclidean Distance	$\sqrt[2]{\sum_{i=1}^{N}(x-y)^2}$
Manhattan Distance	$\sum_{i=1}^{N}\lvert x-y\rvert$
Minkowski Distance	$(\sum_{i=1}^{N}(\lvert x-y\rvert)^p)^{\frac{1}{p}}$

Table 4 \ KNN classification metrics: *This table represents the different metrics for classification in the KNN model.*

4.3.4 CatBoost Classifier

Decision Trees (DTs) are a non-parametric supervised learning method used for classification A decision tree is a decision support technique that forms a tree-like structure. A decision tree consists of three components: decision nodes, leaf nodes, and a root node. A decision tree algorithm divides a training dataset into branches, which further segregate into other branches. This sequence continues until a leaf node is attained. The leaf node cannot be segregated further. The nodes in the decision tree represent attributes that are used for predicting the outcome. Decision nodes provide a link to the leaves. The following diagram shows the three types of nodes in a decision tree:

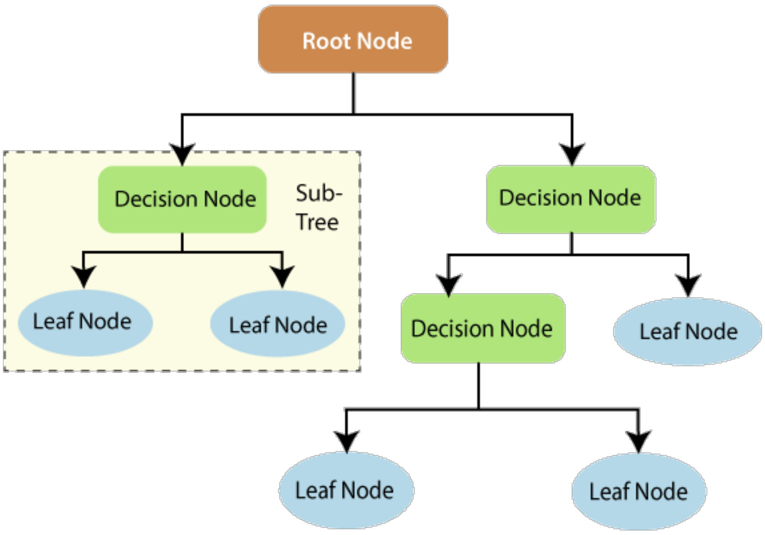


Figure 7 | Decision Tree structure: *This figure represents the three different elements of decision trees (Nodes, Leaf Nodes and the Root Node) and their place in the tree. From datacamp.com*

Random forests is a supervised learning algorithm used for classification. A forest is comprised of trees. The more trees it has, the more robust a forest is. A Random Forest system relies on various decision trees. Every decision tree consists of decision nodes, leaf nodes, and a root node. The leaf node of each tree is the final output produced by that specific decision tree. The selection of the final output follows the majority-voting system. In this case, the output chosen by the majority of the decision trees becomes the final output of the rain forest system. In general, the more trees used the better get the results. However, the improvement decreases as the number of trees increases, i.e. at a certain point the benefit in prediction performance from learning more trees will be lower than the cost in computation time for learning these additional trees.

In the case of random forests, the collection is made up of many decision trees. Random forests are considered “random” because each tree is trained using a random subset of the training data (referred to as bagging in more general ensemble models), and random subsets of the input features (coined feature bagging in ensemble model speak), to obtain diverse trees. Bagging decreases the high variance and tendency of a weak learner model to overfit a dataset. For random forests, both types of bagging are necessary. Without both types of bagging, many of the trees could create similar “if” conditions and essentially highly correlated trees. Instead of bagging and creating many weak learner models to prevent overfitting, often, an ensemble model may use a so-called boosting technique to train a strong learner using a sequence of weaker learners. In the case of decision trees, the weaker learners are underfit trees that are strengthened by increasing the number of “if” conditions in each subsequent model.

XGBoost, CatBoost, and LightGBM have emerged as the most optimized boosting techniques for gradient-boosted tree algorithms, all based on the Random Forest. The algorithms differ from one another in the implementation of the boosted trees algorithm and their technical compatibilities and limitations. XGBoost was the first to try improving GBM’s training time, followed by LightGBM and CatBoost, each with their own techniques, mostly related to the splitting mechanism. Some important aspects of the algorithm:

- Splits: Catboost offers a new technique called Minimal Variance Sampling (MVS), which is a weighted sampling version of Stochastic Gradient Boosting. In this technique, the weighted sampling happens in the tree-level and not in the split-level. The observations for each boosting tree are sampled in a way that maximizes the accuracy of split scoring.

- Leaf growth: Catboost grows a balanced tree. In each level of such a tree, the feature-split pair that brings to the lowest loss (according to a penalty function) is selected and is used for all the level's nodes.

CatBoost distinguishes itself from LightGBM and XGBoost by focusing on optimizing decision trees for categorical variables, or variables whose different values may have no relation with each other (eg. cars and airplanes). To compare cars and airplanes in XGBoost, it would be needed to split them into two one-hot encoded variables representing "is car" and "is airplane," but CatBoost determines different categories automatically with no need for preprocessing (LightGBM does support categories, but has more limitations than CatBoost). Catboostclassifier is used as part of the library catboost. The loss fuction (loss_function in the model) defines the metric to use in training. The specified value also determines the ML problem to solve. It was used the MultiClass loss_function. Number of interactions are set to 500. Hyperparameter tuning is performed on the depth (Depth in the model) and learning rate (learning_rate in the model). Depth represents the depth of the tree and values between 1 and 10 are applied. The learning rate is used for reducing the gradient step, defining the trade-off between the rate of convergence and overshooting. Values between 0.01 and 2 were tested.

4.3.5 SVM

Support Vector Machine is a supervised ML algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of all classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3. A kernel transforms an input data space into the required form. SVM uses a technique called the kernel trick. Here, the kernel takes a low-dimensional input space and transforms it into a higher dimensional space.

Poly kernel is chosen in this research. In order to improve the performance of the model hypermeter tuning is performed to C and to the degree of the model. C is the regularization parameter and degree is the degree of the polynomial kernel function.

4.3.6 Neural Networks

Neural networks, also known as artificial neural networks (ANNs) are a subset of ML and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another. Artificial neural networks (ANNs) are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

A multilayer perceptron (MLP) is a class of feedforward artificial neural network. In a feedforward network, information always moves one direction, it never goes backwards. A MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable. In the research MLP is used with hyperparameter optimization. The number of neurons in the hidden layer is an arbitrary design decision tested empirically to find the optimal design. The design chosen is comprised of two hidden layers with 30 and 15 neurons, respectively. The maximum number of iterations is set to 200 not to overfit. Hyperparameter tuning is performed regarding values of alpha representing the magnitude of the L2 penalty, solver for the type of weight optimization, activation function type of the hidden layers and the learning rate for the schedule of the weights updates.

4.3.7 Alternative Methods and Other Improvements

Time Series data, such as earnings results by quarter imply there could be temporal relations between quarterly observations and more specifically between the different periods in the 4-year rolling window. On the other hand, 5-Fold cross-validation allows for better validation and possibly generalization of the model. In cross-validation the train set is divided into five equal size bins of observations, with five train and validation methods performed for each rolling-window. All ML models are also computed with cross-validation and results presented in the appendix.

Although a possible method to predict earnings would be to start from technical, fundamental, and macro variables and past earnings information, it's a difficult task to predict the earnings class only from past information. An alternative method could be to start from a biased analyst earnings prediction and use all the previous variables discussed to detect and discard the analyst bias towards a more unbiased estimate.

4.3.8 Evaluation

The main measure to classify the performance of each classification model is the accuracy of the prediction, but the accuracy alone is not even half of the picture. A multi-class confusion matrix is computed to represent the performance of each class prediction in each model. A confusion matrix shows the combination of the actual and predicted classes. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class. It is a good measure of whether models can account for the overlap in class properties and understand which classes are most easily confused. Additional measures are computed based on the confusion matrix values adapted to multiple class predictions:

Metric	Formula	Evaluation focus
Average Accuracy	$\frac{\sum_{i=1}^{k} \frac{tp_i+tn_i}{tp_i+tn_i+fp_i+tn_i}}{k}$	The average per-class effectiveness of the classifier
Precision$_M$	$\frac{\sum_{i=1}^{k} \frac{tp_i}{tp_i+fp_i}}{k}$	Average per-class agreement of the true class labels with those of the classifier's
Recall$_M$	$\frac{\sum_{i=1}^{k} \frac{tp_i}{tp_i+fn_i}}{k}$	Average per-class effectiveness of a classifier to identify class labels

Metric	Formula	Evaluation focus
F1-score$_M$	$\frac{2 * Precision_M * Recall_M}{Precision_M + Recall_M}$	The harmonic mean of the macro-average precision and recall

***Table 5 | Result Performance metrics in classification:** This table presents the different metrics in classification to assess the performance of the models. Note: In the formulas below, k = total number of classes; µ and M indices represent micro- and macro-averaging, respectively*

4.3.9 Variable Predictive Power

For further understanding of the predictive power of each variable, different methods are applied to assess such information depending on the model at hand. For each of the five types of model (Logistic Classification, KNN and CatBoostClassifier, SVM and Neural Networks) only one evaluation was conducted, as each model variations are rather similar. This is done on the best predicting model.

5. Results and Discussions

This section is divided in the results for the proof that the difference between earnings prediction by analysts and the actual values present abnormal returns, the results for multi-class earnings prediction, results for the models used as a confirmation for analyst earnings prediction and investment strategies.

5.1 Earnings Event Study

From the literature is clear that earnings announcements generate abnormal returns and that there is a clear relation between the information the firm presents, and the return generated after earnings. In this research, earnings are divided into bins from the EPS growth to analyst predictions. The first necessary step in this research is to realize whether the created bins and variables create clusters of returns significant. In order to achieve that goal a 1-day event study is computed for the day of earnings, only with data that is going to be out-of-sample in the earnings ML forecasts, from 2004-03 until the end of 2020. To compute the abnormal returns the 3-Fama-French model is computed for an estimated window of 252 days with a minimum of 126 days of data for every firm in the day after they are going to present earnings. To the return in the day after earnings is subtracted the "normal" return given by the 3-Fama-French model. First, it's important to take a look at the earnings growth class represented by table 6:

						Percentiles					
Sample	Mean	Std	Skew	Kurt	T test	1%	25%	Median	75%	99%	#obs
EC= 1	-1,30%	7,03%	-0,76	4,64	(-14,95)***	-24,34%	-3,96%	-0,66%	1,81%	16,85%	6503
EC= 2	-0,80%	6,60%	-0,66	5,91	(-11,46)***	-22,02%	-3,26%	-0,42%	2,04%	16,21%	8861
EC= 3	-0,17%	6,17%	-0,31	6,40	(-2,51)**	-18,81%	-2,60%	-0,11%	2,44%	17,24%	8422
EC = 4	0,45%	6,14%	-0,14	5,05	(6,84)***	-17,25%	-2,26%	0,19%	3,07%	18,24%	8674
EC = 5	1,05%	6,51%	0,15	5,26	(14,39)***	-16,40%	-1,98%	0,34%	3,73%	20,78%	8021
EC = 6	1,36%	7,30%	0,76	6,23	(15,62)***	-17,21%	-2,06%	0,55%	3,98%	23,65%	7034

Table 6 \ EPS Growth Class: *This table presents a one-day event study in the day after earnings announcement for different classes of EPS growth (EC). Kurt refers to kurtosis.* The t-statistics are in parenthesis. *, **, *** indicates significance at 10%, 5%, 1%, respectively. *The t-statistics are in parenthesis. *, **, *** indicates significance at 10%, 5%, 1%, respectively.*

Looking at the table 6, the mean return in each class increases with every class increase, from almost negative 1.30% in class 1 to more than 1.30% in class 6. The standard deviation is very high across all classes, which would be expected around earnings. The bigger and lower classes are more significant than the classes in the middle, with lower returns by firms that were classified to a more neutral position. Although the extreme positive observations increase from classes 1 to 6 and the opposite occurs to extreme negative observations, it's still possible to observe extreme returns in all classes in the 1% and 99% percentiles. This shows that earnings growth is far from being the only factor influencing firm returns after earnings. The next table looks at the class evolution of EPS from the last quarter until the present.

						Percentiles					
Sample	Mean	Std	Skew	Kurt	T test	1%	25%	Median	75%	99%	#obs
EQ = -5	-1,14%	7,01%	-0,67	4,99	(-7,65)***	-23,37%	-3,80%	-0,57%	1,76%	17,32%	2216
EQ = -4	-0,78%	7,04%	-0,53	6,05	(-5,51)***	-24,18%	-3,44%	-0,45%	2,04%	18,54%	2472
EQ = -3	-0,61%	6,77%	-0,81	6,54	(-5,25)***	-22,68%	-3,05%	-0,28%	2,24%	16,61%	3391
EQ = -2	-0,64%	6,78%	-0,52	6,14	(-6,27)***	-22,55%	-3,18%	-0,31%	2,18%	17,75%	4460
EQ = -1	-0,21%	6,45%	-0,48	5,77	(-2,71)***	-19,55%	-2,85%	-0,11%	2,60%	17,92%	7192
EQ = 0	0,11%	6,63%	0,11	7,74	(-1,58)	-20,23%	-2,66%	-0,02%	2,82%	19,49%	9608
EQ = 1	0,54%	6,51%	0,09	4,98	(6,44)***	-18,16%	-2,27%	0,17%	3,22%	19,97%	6051
EQ = 2	0,91%	6,14%	0,00	5,09	(9,08)***	-15,58%	-1,85%	0,39%	3,42%	19,03%	3792
EQ = 3	0,86%	6,18%	0,53	3,55	(7,65)***	-15,39%	-2,00%	0,19%	3,45%	20,86%	2997
EQ = 4	1,09%	6,99%	0,52	3,61	(7,40)***	-16,82%	-2,15%	0,33%	3,81%	22,03%	2263
EQ = 5	0,95%	7,02%	0,50	4,00	(7,51)***	-17,72%	-2,31%	0,32%	3,38%	22,69%	3073

Table 7 \ EPS Growth Class evolution from t-1 quarters: *This table presents a one-day event study in the day after earnings announcement for all EPS class progressions from the previous quarter (EQ). Kurt refers to kurtosis.* The t-statistics are in parenthesis. *, **, *** indicates significance at 10%, 5%, 1%, respectively.

Looking at the table 7, the mean return in each class evolution increases with every class increase, from around negative 1% in class evolution -5 to more than 1% in class evolution 4, with the exception of class evolution 5. The standard deviation is very high across all classes, which would be expected around earnings. Class evolution 0 is not significant at the 10% level. The number of observations is much higher in the middle showing that an extreme evolution in both directions is not the most likely scenario. The next table looks at the Analyst EPS class error.

						Percentiles					
Sample	Mean	Std	Skew	Kurt	T test	1%	25%	Median	75%	99%	#obs
EA = -5	-3,15%	9,16%	-0,79	3,06	(-5,98)***	-33,71%	-6,56%	-2,14%	1,36%	19,82%	307
EA = -4	-1,77%	7,61%	-0,82	2,99	(-4,23)***	-28,55%	-4,77%	-0,86%	2,01%	16,16%	334
EA = -3	-3,15%	8,55%	-1,22	3,71	(-9,00)***	-31,55%	-5,94%	-1,31%	1,29%	16,57%	601
EA = -2	-2,48%	7,31%	-0,86	6,40	(-11,77)***	-24,86%	-5,19%	-1,39%	1,08%	15,69%	1210
EA = -1	-1,40%	6,34%	-0,76	6,17	(-16,80)***	-22,14%	-3,77%	-0,74%	1,57%	15,34%	5787
EA = 0	0,00%	6,26%	-0,17	7,16	(-0,06)	-18,94%	-2,48%	-0,05%	2,51%	18,21%	27650
EA = 1	1,44%	6,77%	0,38	3,99	(19,86)***	-15,86%	-1,93%	0,67%	4,37%	21,52%	8773
EA = 2	2,20%	7,62%	0,28	2,08	(12,22)***	-17,50%	-1,92%	1,31%	5,94%	22,91%	1801
EA = 3	2,69%	7,85%	0,36	2,56	(8,48)***	-14,32%	-1,60%	1,12%	5,92%	24,98%	615
EA = 4	3,37%	8,26%	0,44	1,91	(6,83)***	-16,48%	-1,69%	2,08%	7,95%	26,83%	285
EA = 5	1,95%	8,66%	0,65	3,13	(2,74)***	-17,55%	-2,33%	0,98%	5,98%	29,88%	152

Table 8 \ ***EPS Growth Analyst error:*** *This table presents a one-day event study in the day after earnings announcement for all differences between analyst class predictions and the true class of earnings (EA). Kurt refers to kurtosis. The t-statistics are in parenthesis. *, **, *** indicates significance at 10%, 5%, 1%, respectively.*

Looking at the table 8, the mean return shows a tendency to increase from the bottom to the top classes, from more than negative 3% in analyst error class -5 to more than 3% in class evolution 4, with the exception of analyst error class 5, with a return around 2%. The table shows that positive surprises from analyst expectations are perceived as positive by the markets and negative news compared to analyst predictions are accompanied by negative returns. The standard deviation is very high across all classes, which would be expected around earnings. Class evolution 0 is not significant at the 10% level and all other classes are significant at the 1% level, although having a very high number of observations. The number of observations is much higher in the middle showing that an extreme surprise relatively to analysts in both directions is not the most likely scenario. The very low number of observations may explain the drop from the trend of returns in class error 5 and -4.

What the previous tables also suggest is that the initial and ending class of earnings may be very important to mitigate to fewer extreme returns in certain clusters of returns (e.g. it's very different to increase from class 1 to 2 or 3 to 5). Now let's look at the explanatory power of this variables in the day after earnings return:

EPS Growth class	EPS Growth analyst error prediction	EPS Growth evolution from t-1 quarters	IVOL	TVOL	R²
0,0056 (30,20)***					1,90%
	0,0098 (36,06)***				2,70%
		2,40E-03 (20,56)***			0,90%
			-0,0166 (-0,47)		0,00%
				-0,0092 (-0,31)	0,00%
0,004 (11,99)***	7,90E-03 (23,24)***	-1,00E-04 (-0,55)	0,0231 (0,20)	-0,0399 (-0,43)	2,40%

Table 9 | ***EPS Growth Analyst error:*** *This table presents a regression of the after earnings return to the subsets of predictors chosen. In each row, the table reports the averages of the regression slope coefficients and their associated Newey-West (1987) adjusted t-statistics.* The t-statistics are in parenthesis. *, **, *** indicates significance at 10%, 5%, 1%, respectively. R-squared is in percentage.

All variables related to EPS were significant when tested individually with the return after earnings as the dependent variable. Analyst errors is the most significant variable explaining earnings, followed by EPS class and EPS evolution from the previous quarter class, respectively, looking at the single regressions. The multiple regression using all the variables shows the negligent effect of the evolution of EPS growth from the previous quarter class of EPS growth. A possible explanation is that EPS growth already represents the EPS growth from the previous quarter. In the EPS growth evolution in the end EPS of the current quarter are being compared to the two previous quarters. The multiple regression only explains 2.4% of the variation in returns in the day after earnings. This low number could be an effect of the choice of using the class variables as quantitative variables when they have very low variability compared to returns. Idiosyncratic Volatility (IVOL) is not significant in explaining the return after earnings, showing that these returns have no business with the return in the firm stocks that usually cannot be explained by models like the 3 Fama-French model.

5.2 Multi-class Prediction Results

A summary of the results is presented in the table and graph bellow where the models are compared to analyst prediction accuracy for each class of prediction (3, 6 and 9).

Number of Classes	Model					
	Logistic Classification	KNN	CatBoost	SVM	NN	Analysts Consensus (Mean)
3	58,00%	57,06%	60,95%	56,70%	59,02%	75,28%
6	35,70%	36,96%	42,62%	33,92%	40,13%	59,50%
9	24,69%	27,08%	31,96%	24,56%	30,01%	48,75%

Table 10 | Classification accuracy using ML models and analysts' predictions, by class.

As the number of classes increases the accuracy of all the models decrease. The volatility of the models increases as well. Higher analyst accuracy suggest that they have information not present in the markets. All the analysis will be done on the groups of six.

Other metrics were computed regarding the performance of the ML models, besides overall accuracy. Other metrics were computed not focusing on the size of each class, but on the average class performance, which can be seen in table 11:

	Logistic Classification	KNN	CatBoost	SVM	NN	Analysts Consensus
Accuracy (%)	31,39%	38,33%	41,67%	35,00%	41,67%	60,00%
Precision (%)	34,37%	41,78%	42,73%	36,35%	39,89%	64,77%
Recall (%)	33,70%	38,13%	42,59%	34,09%	40,93%	65,51%
F1-Score (%)	34,04%	39,87%	42,66%	35,18%	40,40%	65,14%

Table 11 \ Metrics of ML prediction classification when stratified by 6 classes.

5.2.1 Logistic Classification

The Logistic Classification with hyperparameter tunning showed a low overall accuracy of 35.70%. The multiclass average accuracy was lower at 31.39% and the Precision (true positives among all positives predicted) and Recall (true positives predicted among all positives) were higher, but still lower than the overall accuracy. Although having low accuracy the average class error is 1.25, implying that in most of the cases even when the prediction is wrong the prediction is not too far off from the true class. This makes the Logistic Classification a good benchmark for the prediction. The model performs the best when classifying extreme classes (1 and 6). The same can be said about analysts, being able to better identify extremes and still be better than the Logistic Classification model. Classes from 2 to 5 present a very high number of misclassifications, with class 3 showing predictions almost equally spread along all class of predictors.

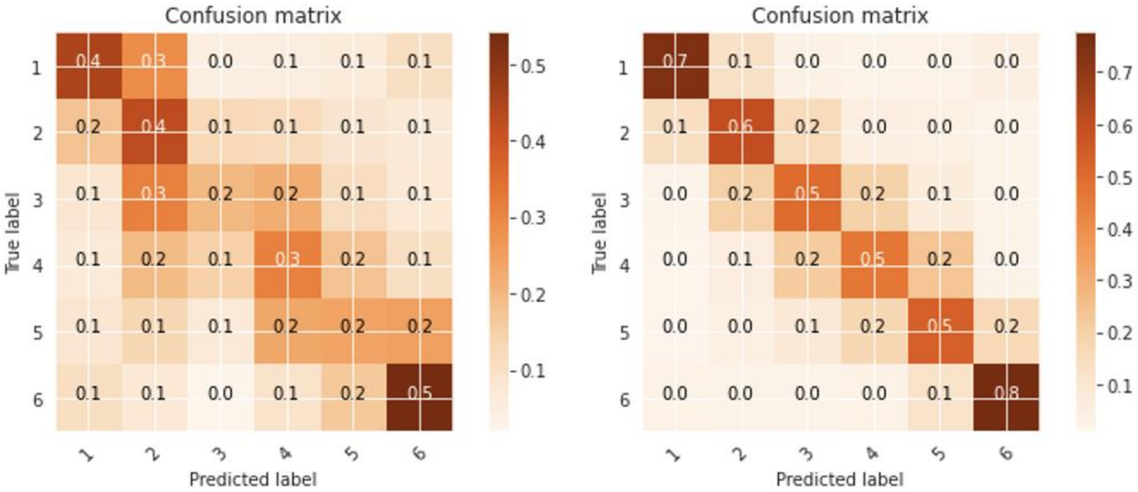


Figure 8 | Confusion matrix for the Logistic Classification model (left) vs. Analyst Forecasts confusion matrix (right).

5.2.2 KNN

KNN has a very similar performance compared to Logistic Classification with an overall accuracy of 36.96% and it's still worse than analysts' performance. These measures improve in the metrics per class with the surprising result of the second highest precision (41.78%) among all the ML models used. KNN improves the classification of the middle classes from 3 to 4 and performs worse classifying extreme events that are represented by the class 1 and 6.

This reality is exemplified by the average of the error across the out-of-sample data with an average of 1.21, lower than Logistic Classification.

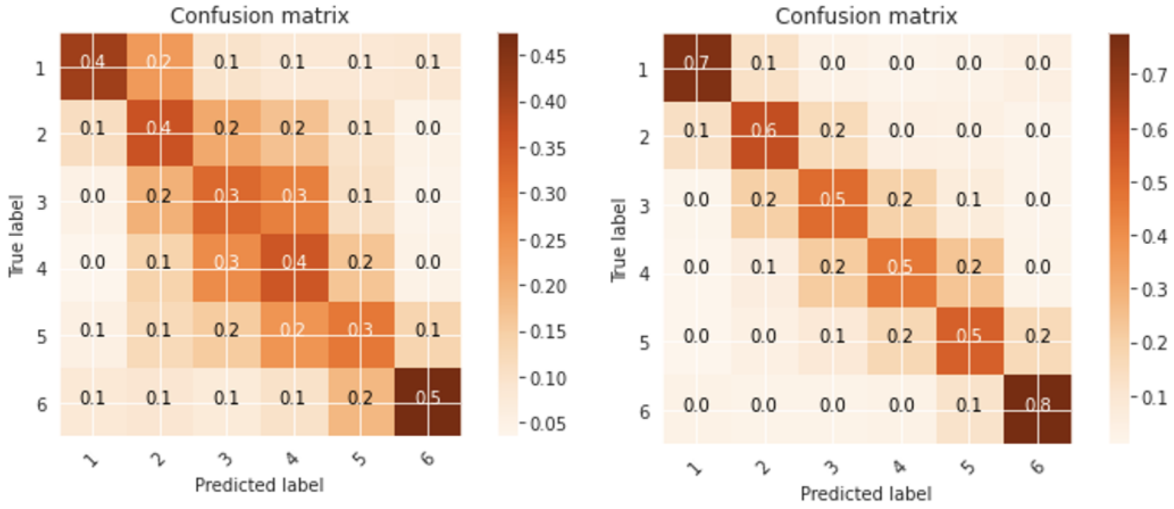


Figure 9 | Confusion matrix for the KNN Classification model (left) vs. Analyst Forecasts confusion matrix (right).

5.2.3 CatBoost

The CatBoost Classification with hyperparameter tunning, which is ultimately based on Random Forest, showed the highest overall accuracy of 42.62% from all ML models, still far away from analyst mean of 59.50%. Among the multiclass measures it's also the best model for all of them. The lowest value comes in the multiclass accuracy at 41.67%, although this measure consistently sees lower values. The average class error decreases to 1.15, implying that in most of the cases even when the prediction is wrong the prediction is not too far off from the true class, with most predictions within one class of the true class value. The model performs the best when classifying extreme classes (1 and 6), with an accuracy of 50% and 70% considering a one class error as still accurate. The same can be said about analysts, being able to better identify extremes and still be better than the CatBoost Classification model. The model performs as good as Logistic Classification for extreme EPS growth classification and as good as KNN for the intermediate observations. Since CatBoost is the best model, it was chosen to be used with analyst prediction and volatility of those predictions as an input to the model forecast.

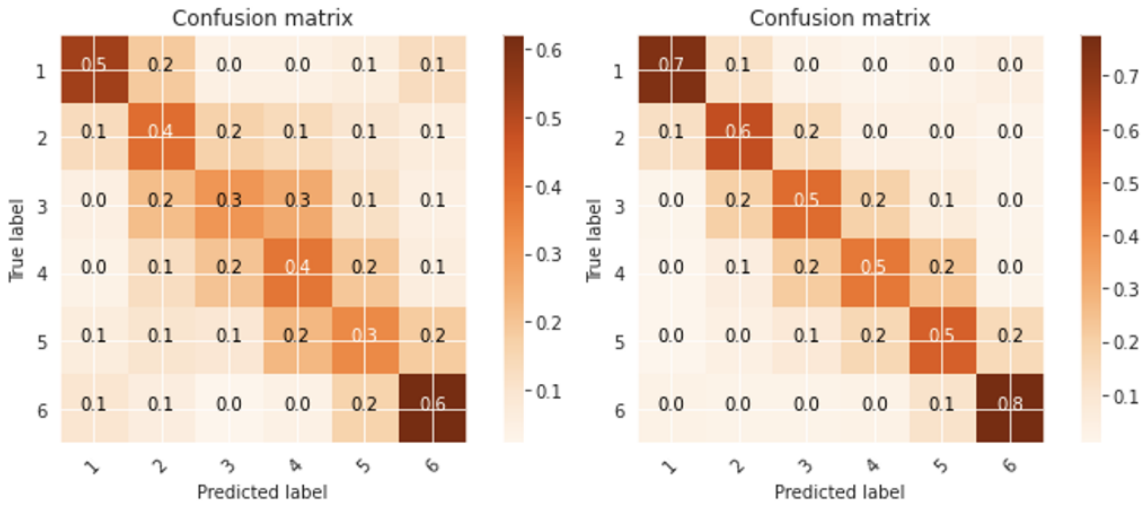


Figure 10 | Confusion matrix for the CatBoost Classification model (left) vs. Analyst Forecasts confusion matrix (right).

5.2.4 SVM

SVM is a model capable of capturing non-linear relationships between the independent variables and the dependent variable. SVM low average accuracy of 33.92% is surprising having in consideration the complexity of the model and the amount of computational power used in multiple hyperparameter tunings. A possible explanation could be the kernel used (), being a Polynomial Kernel to distinguish non-linear relations. The model performs better than Logistic Classification when looked at the multiclass measures, but it’s still far behind all the other models.

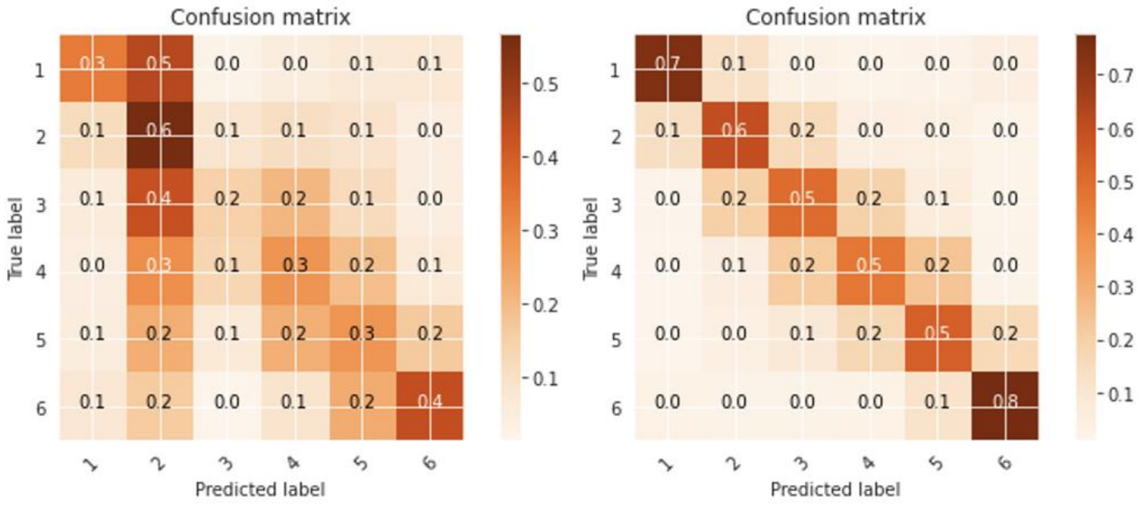


Figure 11 | Confusion matrix for the SVM Classification model (left) vs. Analyst Forecasts confusion matrix (right).

5.2.5 Neural Networks

The second-best performing model is a MLP with an overall accuracy of 40.13%, still behind analysts. The average error 1.16 is slightly higher than Catboost Classification, making the model perform slightly worse in the situations in which the prediction is wrongly classified. The model has a very similar performance to Catboost in terms of multiclass accuracy. Precision and recall are lower, meaning lower true positives among all positive predicted values for each class and lower true positives predicted among all positive values for each class of values.

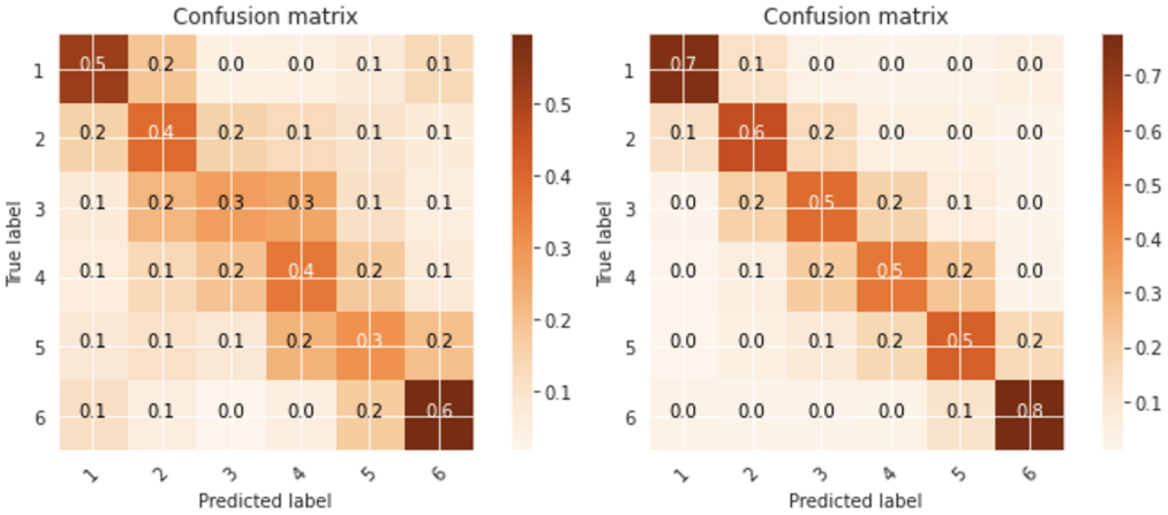


Figure 12 | Confusion matrix for the NN Classification (left) model vs. Analyst Forecasts confusion matrix (right).

5.2.6 Analyst forecasts incorporation into model

Putting together analyst expectations into the model the performance of the model underperforms analysts (57.46% versus 59.50% from analysts), but not by much, suggesting that analysts could already be taking into account the information used as input by the models (macro, financial and technical variables).

5.3 Can ML models confirm analysts' expectations?

In addition, further research was conducted on how the results of the models can be utilized to advance the consensus predictions, where we calculated the conditional accuracy rate of prediction under the circumstances that the result of each model converges (i.e., having the same forecast) or diverge with the consensus. Results are presented in the table below:

Number of Class	Analysts & Model agree					
	Logistic Classification	KNN	CatBoost	SVM	NN	Analysts Consensus (Mean)
3	82,25%	81,87%	82,72%	81,89%	82,13%	75,28%
6	68,63%	68,95%	70,72%	67,10%	69,56%	59,50%
9	58,46%	59,81%	61,62%	56,79%	60,40%	48,75%
Number of Class	Analysts & Model disagree					
	Logistic Classification	KNN	CatBoost	SVM	NN	Analysts Consensus (Mean)
3	64,71%	65,81%	62,59%	65,73%	64,45%	75,28%
6	52,19%	51,80%	48,63%	53,39%	50,05%	59,50%
9	44,01%	43,02%	40,75%	44,63%	41,84%	48,75%

Table 12 \ Analyst Prediction accuracy when ML models and analysts agree on class classification.

It is easily recognizable that when the consensus prediction has converged with our model results, the conditional accuracy rate is relatively higher than the normal average. Disagreements in predictions show the oppositive effect with lower analyst's accuracy than their average accuracy.

5.4 Investment Strategy

In order to take advantage of the previous results two investment strategies were computed. The first strategy is based on the situations in which the Catboost forecast, and Analysts agree on the prediction, which increases the accuracy of the prediction. In the beginning of each day equal amounts of capital are going to be employed in the number of earnings in each class. Daily returns are computed and based on the chronologic unfold of events monthly returns are computed. Other investment strategy consists in investing only in earnings events with Catboost predicting with more than 70% of probability a class of earnings from the 6. In that situation mean analyst prediction is going to be used, because analysts were always superior to the ML models. The assumption is that when the ML model is certain about an earnings event, then analysts are also more certain in their predictions. In a way this investment strategy tries to predict analyst errors, in order to minimize the earnings class forecast error.

Panel A: 3FF Portfolios	C1	C2	C3	C4	C5	C6	C6-C1
3FF alpha	-0,02	-0,01	0,01	0,01	0,02	0,06	0,08
	(-2,28)**	(-1,36)	(1,60)	(0,75)	(1,86)*	(4,05)***	(4,74)***
Mkt	0,76	0,50	0,40	0,03	0,18	0,48	-0,23
	(2,93)***	(2,11)**	(1,76)*	(0,11)	(0,66)	(1,32)	(-0,55)
SMB	0,02	0,71	-0,25	0,99	0,47	0,28	0,27
	-0,05	(1,61)	(-0,60)	(2,22)**	(0,92)	(0,42)	(0,34)
HML	-0,41	0,31	0,93	-0,35	-0,16	0,96	1,28
	(-1,05)	-0,87	(2,74)***	(-0,95)	(-0,39)	(1,76)*	(2,03)**
R-squared	5,00%	7,00%	7,60%	3,20%	1,10%	4,00%	2,20%
Panel B: Performances	C1	C2	C3	C4	C5	C6	C6-C1
Return (%)	-18,69%	-12,42%	17,54%	10,87%	26,34%	72,13%	89,91%
Standard Deviation (%)	48,81%	46,27%	42,59%	46,05%	50,86%	70,00%	80,85%
Sharpe Ratio	-0,38	-0,27	0,41	0,24	0,52	1,03	1,11
Skewness	0,14	0,30	0,89	0,64	1,05	2,20	0,94
Kurtosis	1,05	1,56	4,57	1,88	5,39	11,00	2,67
Panel C: Characteristics	C1	C2	C3	C4	C5	C6	C6-C1
Average number of firms	55,81	57,04	37,90	46,38	38,07	66,16	121,97
Average Book to Market of firms	0,63	0,46	0,38	0,35	0,43	0,59	0,61

Table 13 \ Investment strategy based on analysts and Catboost models agreeing*: In Panel A, 3FF alphas (in percentage) and betas are calculated by running time-series regressions of the EPS growth class prediction on the excess returns of the market, size and value factors. In Panel B, Excess Returns, Standard Deviation, Sharpe Ratio, Skewness and Kurtosis are computed for each quintile. First three measures are annualized. Panel C firm characteristics are computed. The t-statistics are in parenthesis. *, **, *** indicates significance at 10%, 5%, 1%, respectively. R-squared is in percentage.*

In Table 13, regarding the lowest EPS growth stocks (C1), the research finds an annualized excess return of -18.69% with a volatility of 48.81%. The annualized SR is found to be at -0.38, meaning that these stocks underperform, making them potential candidates for entering a short position. In direct comparison with the 3FF, we can observe that C1 yields negative abnormal returns of -0.01, respectively, being statistically significant (5%) with a t-stat of -2.28. Moving to long portfolio, including the highest forecasted EPS growth stocks (C5), there is a positive excess return of 72.13% with an annualized volatility of 70.00%. The SR yields 1.03 and is therefore extremely better compared to C1. In the regression with the 3FF, C6 gives statistically significant results with t-stat of 4,05, being significant at the 1% level. Moving to long-short portfolio (C6-C1) there is a positive excess return of 89.91% with an annualized volatility of 80.85%. The SR yields 1.11 and is therefore marginally better compared to C6.

The skewness of 0.94 and kurtosis of 2.67 are a good indication. In the regression with the 3FF, C6-C1 gives statistically significant results with t-stat of 4,74, significant at the 1% level.

Panel A: 3FF Portfolios	C1	C2	C3	C4	C5	C6	C6-C2
3FF alpha	-0,01	-0,03	0,00	0,02	0,03	0,05	0,08
	(-1,11)	(-2,39)**	(0,25)	(1,63)	(2,52)**	(3,47)***	(4,05)***
Mkt	0,49	0,40	0,29	0,30	0,13	0,32	-0,08
	(1,72)*	(1,48)	(1,07)	(1,00)	(0,43)	(0,83)	(-0,17)
SMB	0,54	0,64	0,51	1,19	0,56	0,77	0,09
	(1,05)	(1,29)	(1,01)	(2,19)**	(1,04)	(1,07)	(0,11)
HML	-0,42	-0,30	0,58	0,01	0,41	0,85	1,16
	(-0,98)	(-0,75)	(1,44)	(0,03)	(0,92)	(1,45)	(1,58)
R-squared	3,20%	3,10%	3,70%	4,60%	1,80%	3,30%	1,30%
Panel B: Performances	C1	C2	C3	C4	C5	C6	C6-C2
Return (%)	-8,19%	-26,31%	3,99%	27,32%	35,89%	65,53%	91,12%
Standard Deviation (%)	53,63%	51,56%	52,30%	57,28%	56,20%	75,00%	93,77%
Sharpe Ratio	-0,15	-0,51	0,08	0,48	0,64	0,87	0,97
Skewness	0,31	0,45	0,51	0,66	0,45	1,38	1,02
Kurtosis	1,39	0,87	1,85	0,95	1,17	5,00	3,71
Panel C: Characteristics	C1	C2	C3	C4	C5	C6	C6-C2
Average number of firms	13,26	3,50	2,08	2,31	3,94	21,19	34,46
Average Book to Market of firms	0,69	0,66	0,62	0,65	0,64	0,68	0,68

Table 14 \ Investment strategy based on analysts predictions for high confidence Catboost predictions: *In Panel A, 3FF alphas (in percentage) and betas are calculated by running time-series regressions of the EPS growth class prediction on the excess returns of the market, size and value factors. In Panel B, Excess Returns, Standard Deviation, Sharpe Ratio, Skewness and Kurtosis are computed for each quintile. First three measures are annualized. Panel C firm characteristics are computed. The t-statistics are in parenthesis. *, **, *** indicates significance at 10%, 5%, 1%, respectively. R-squared is in percentage.*

In the second strategy, regarding the predicted second lowest EPS growth stocks (C2), the research finds an annualized excess return of -26.31% with a volatility of 51.56%. The annualized SR is found to be at -0.51, meaning that these stocks underperform, making them potential candidates for entering a short position. In direct comparison with the 3FF, we can observe that C2 yields negative abnormal returns of -0.03, respectively, being statistically significant (5%) with a t-stat of -2.39. Moving to long portfolio, including the highest forecasted EPS growth stocks (C6), there is a positive excess return of 65.53% with an annualized volatility of 75.00%. The SR yields 0.87 and is therefore extremely better compared to C2. In the regression with the 3FF, C6 gives statistically significant results with t-stat of 3,47, being significant at the 1% level. Moving to long-short portfolio (C6-C2) there is a positive excess return of 91.12% with an annualized volatility of 93.77%. The SR yields 0.97

and is therefore marginally better compared to C6. Looking at the cumulative performance of the two strategies, the first is characterized by moments of extreme volatility making it not perform consistently. During the same period the S&P 500 increased by more than 6-fold. This strategy would bring investors funds almost to zero twice and finished 2020 with shy of a 10% return.

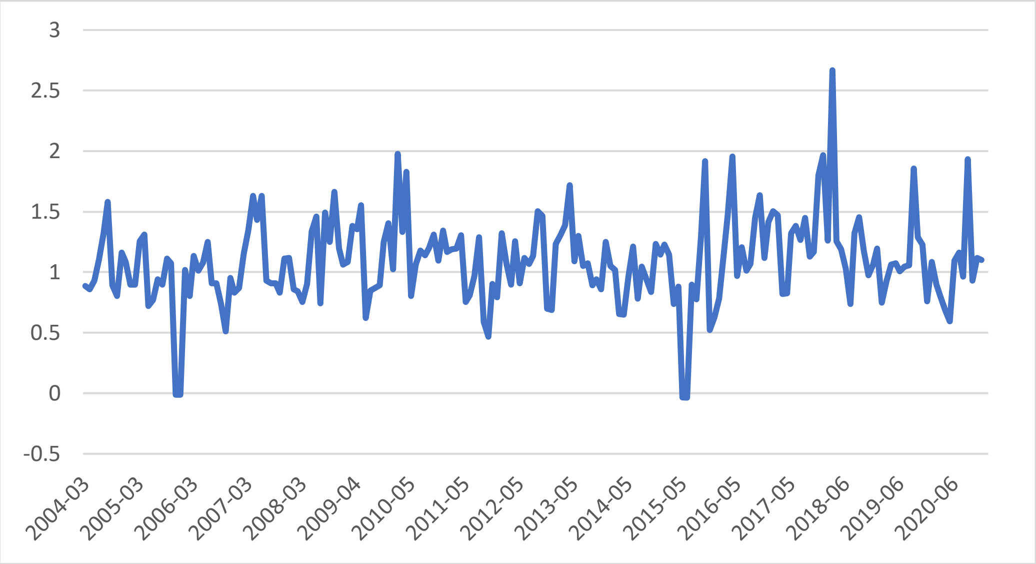


Figure 13 | Cumulative return of the Investment Strategy 1

The second strategy is characterized by an initial period with a bad performance until gaining consistency and reaching an incredible performance overall, but still very volatile.

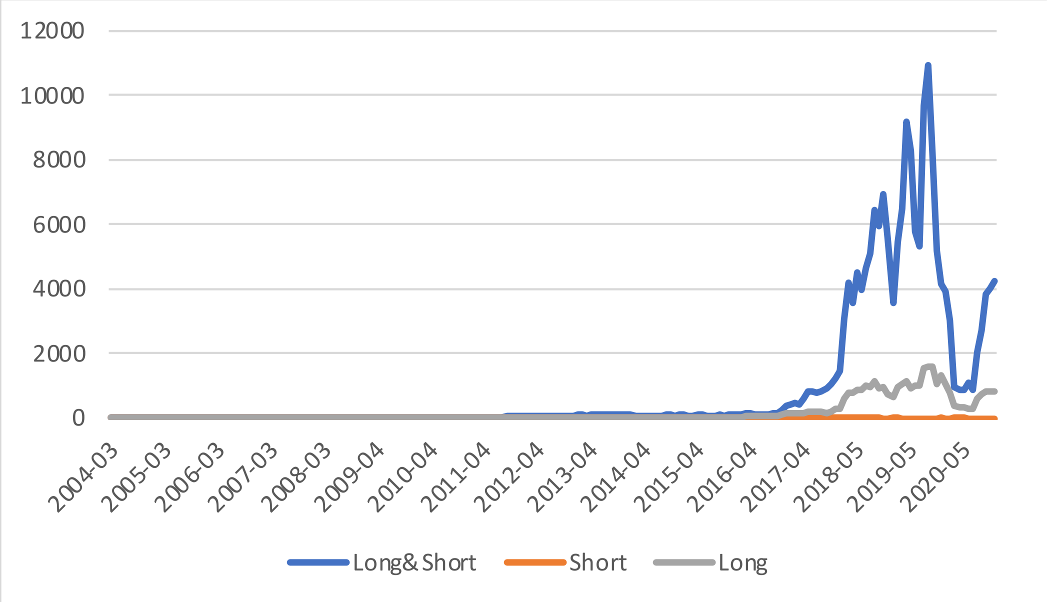


Figure 14 / Cumulative return of the Investment Strategy 2, namely the Long-short strategy and its two legs (long and short)

6. Conclusion

With the goal of creating a methodically and exhaustive research around earnings, collating pieces from different models discussed in different previous research, this research had the main goal to predict earnings and compare this prediction to analyst errors of prediction.

More specifically, it aimed at answering the following three questions:

- Can returns following earnings reports be explained by the growth class of earnings, the evolution of the class of earnings, or analysts' mistakes?
- Can ML models outperform analyst accuracy in earnings prediction?
- Can ML models confirm analysts' predictions?

Looking at the first step, it is clear that EPS growth is fundamental in explaining the markets' reaction to earnings reports. This can be explained by two effects – the magnitude of growth and the sign of earnings. We conclude that when analysts are right and EPS growth is small, no abnormal returns are captured in the day following earnings announcement. The same applies when EPS growth has been stable for at least one year. For EPS growth class, analyst class errors, and EPS class progression from the previous quarter, we capture abnormal returns when firms are stratified into six EPS growth classes. Of note, EPS class progression loses significance when considered in aggregate with the remaining two variables. The heightened volatility after earnings reports suggests that the market's reaction to earnings is a much more complex phenomenon than a linear relation between high EPS and high returns.

The answer pertaining to the second question is consistent across all ML models. Indeed, analysts' predictions are more accurate than ML models. One possible explanation could be the extensive research analysts perform in each prediction combined with the use of non-public information. In addition, the results suggest that perhaps analysts incorporate ML inputs in their computations, once we find low gains in accuracy when joining analysts' expectations with ML models. Either way, among all ML models used, the Catboost classification method is consistently the best performing model for all metrics. Actually, for predictions with high conviction, accuracy is much higher, and that fact is used in an investment strategy.

The results related to the last question indicate ML models' predictions as a good confirmation for analysts' predictions, consistent with existent literature. This study complements Xinyue et al. (2020) findings by adding investment strategies with satisfactory results. The ML models can identify situations in which analyst's accuracy is higher and lower, which indirectly can be used to identify situations on analysts' errors.

Finally, the study is not without limitations and improvements can be made upon this research in some respects. First, the research is limited by firms with analysts' predictions. In fact, there is an issue with sample selection bias, since it is only reasonable to predict earnings of firms targeted by analysts. Therefore, for further research in this topic to correctly predict earnings of firms, it should be included the ones without analysts' coverage. Second, the research could use more technical analysis variables and be tested for longer timeframes, rather than quarterly predictions. Third, the research could be replicated for a continuous EPS growth prediction, instead of dividing EPS in classes. Lastly, other ML models could be applied such as more complex neural network models and more extensive hyperparameter tuning could also be performed.

7. Appendix

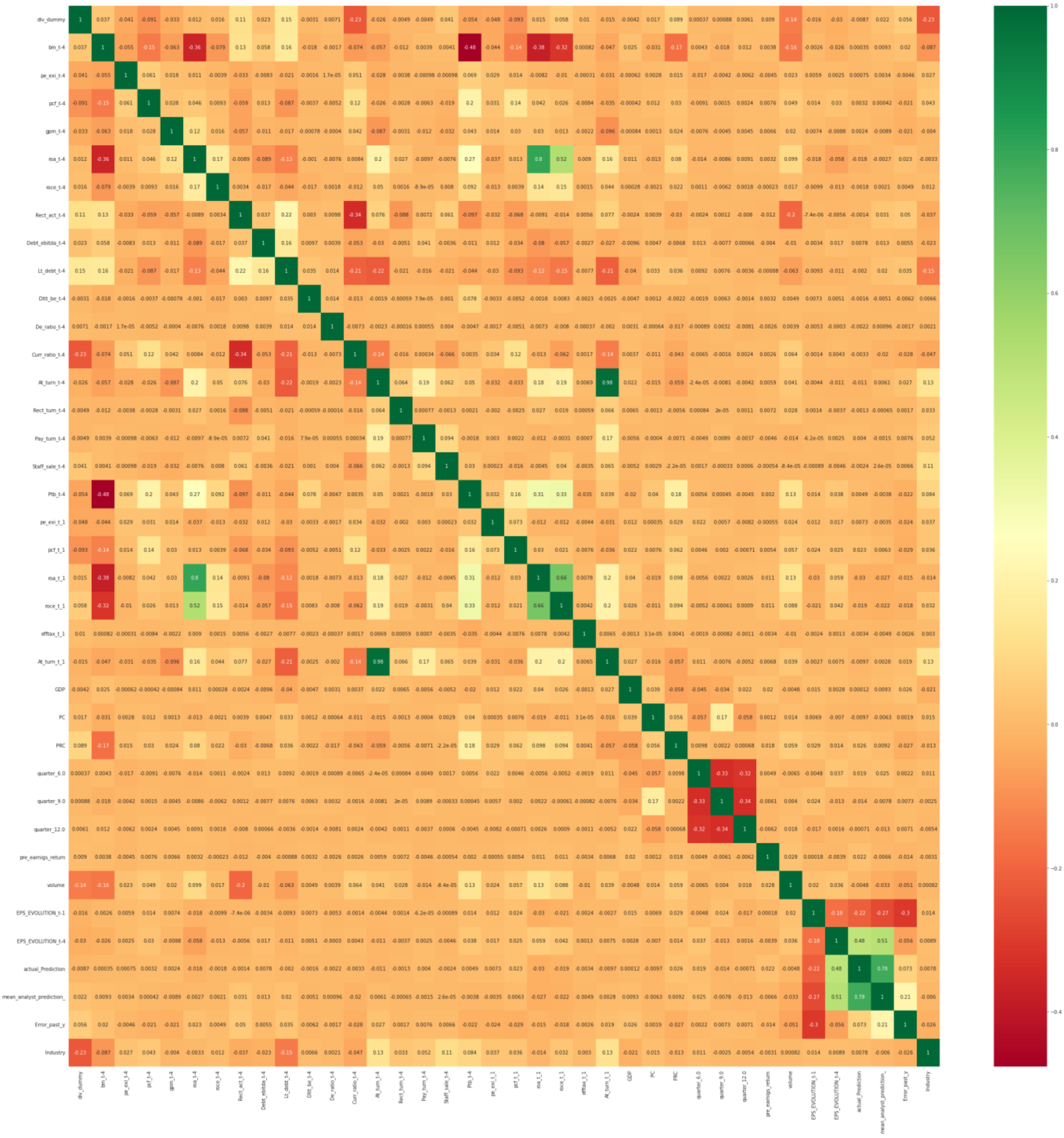
Appendix 1- variable names and description

Variable	Formula	Lagged 1 year	Lagged 1 quarter	Lagged 5 days
Dividend_ dummy	1 if paid dividend, 0 otherwise	-	✓	-
Current Quarter	1, if 1st quarter, 2 if 2nd quarter…	-	-	-
roa	Return on Assets	✓	✓	-
roce	Return on Capital Employed	✓	✓	-
pe_exi	P/E (Diluted, Excl. EI)	✓	✓	-
bm	Book/Market	✓	-	-
pcf	Price/Cash flow	✓	✓	-
gpm	Gross Profit Margin	✓	-	-
Rect_act	Receivables/Current Assets	✓	-	-
Debt_ebitda	Total Debt/EBITDA	✓	-	-
Lt_debt	Long-term Debt/Total Liabilities	✓	-	-
Dltt_be	Long-term Debt/Book Equity	✓	-	-
De_ratio	Total Debt/Equity	✓	-	-
Curr_ratio	Current Ratio	✓	-	-
At_turn	Asset Turnover	✓	✓	-
Rect_turn	Receivables Turnover	✓	-	-
Pay_turn	Payables Turnover	✓	-	-
Staff_sale	Labor Expenses/Sales	✓	-	-
Ptb	Price/Book	✓	-	-
efftax	Effective Tax Rate	✓	✓	-

dpr	Dividend Payout Ratio	✓	-	-
Intcov_ratio	Interest Coverage Ratio	✓	-	-
Fcf_ocf	Free Cash Flow/Operating Cash Flow	✓	-	-
Int_debt	Interest/Average Long-term Debt	✓	-	-
ps	Price/Sales	✓	-	-
Debt_at	Total Debt/Total Assets	✓	-	-
Debt_invcap	Long-term Debt/Invested Capital	✓	-	-
GProf	Gross Profit/Total Assets	✓	-	-
Invt_act	Inventory/Current Assets	✓	-	-
Cash_debt	Cash Flow/Total Debt	✓	-	-
Totdebt_invcap	Total Debt/Invested Capital	✓	-	-
Debt_capital	Total Debt/Capital	✓	-	-
GDP	US gross domestic product (GDP) data	-	✓	-
PC	US Personal Consumption Expenditures: Durable Goods (PCDG) data	-	✓	-
volume	5 trading days cumulative volume (from day 6 to day 1)/ number of shares outstanding	-	-	✓
pre_earnigs_return	5 trading days cumulative return (from day 6 to day 1)	-	-	✓
EPS_EVOLUTION	EPS growth class (t) - EPS growth class (t-1)	✓	✓	-
Error Past	EPS growth class (t-1) - EPS Analyst growth class Forecast for (t-1)	-	✓	-

Industry	Industries information based on the first 2 digits of the NAICS code	-	-	-
Actual value	EPS growth class	-	-	-

Appendix 2 – Correlation matrix of the independent variable



Appendix 3 – Event Study for number of classes equal to 3

Class 3 - EPS Growth class											
						Percentiles					
Sample	Mean	Std	Skewness	Kurtosis	T test	1%	25%	Median	75%	99%	#obs
Earnings class = 1											
Out-of-sample data	-1,01%	6,79%	-0,71	5,33	(-18,53)***	-23,34%	-3,57%	-0,51%	1,94%	16,54%	15364
Earnings class = 2											
Out-of-sample data	0,15%	6,17%	-0,23	5,71	(3.09)***	-18,13%	-2,43%	0,04%	2,76%	17,77%	17096
Earnings class = 3											
Out-of-sample data	1,19%	6,89%	0,50	5,96	(21,23)***	-16,80%	-2,01%	0,44%	3,84%	22,52%	15055

Class 3 - EPS Growth analyst error prediction											
						Percentiles					
Sample	Mean	Std	Skewness	Kurtosis	T test	1%	25%	Median	75%	99%	#obs
Earnings class-analyst class forecast = -2											
Out-of-sample data	-2,57%	8,16%	-0,88	3,25	(-9,80)***	-30,37%	-5,49%	-1,26%	1,52%	18,22%	969
Earnings class-analyst class forecast = -1											
Out-of-sample data	-1,86%	6,79%	-0,98	5,88	(-17,52)***	-24,26%	-4,26%	-0,95%	1,46%	14,70%	4083
Earnings class-analyst class forecast = 0											
Out-of-sample data	0,08%	6,40%	-0,10	6,66	(2.27)**	-19,02%	-2,51%	-0,03%	2,65%	18,77%	35679
Earnings class-analyst class forecast = 1											
Out-of-sample data	1,67%	7,10%	0,24	3,16	(18,12)***	-17,02%	-1,92%	0,84%	4,90%	21,83%	5972
Earnings class-analyst class forecast = 2											
Out-of-sample data	2,78%	7,94%	0,60	2,23	(9,95)***	-15,07%	-1,70%	1,38%	6,32%	27,55%	812

Class 3 - EPS Growth evolution from t-1 quarters											
						Percentiles					
Sample	Mean	Std	Skewness	Kurtosis	T test	1%	25%	Median	75%	99%	#obs
Earnings class - Previous Quarter Earnings class = -2											
Out-of-sample data	-0,88%	6,90%	-0,73	6,08	(-10,54)***	-23,85%	-3,40%	-0,43%	2,00%	17,06%	6873
Earnings class - Previous Quarter Earnings class = -1											
Out-of-sample data	-0,46%	6,73%	-0,48	5,53	(-6,36)***	-21,70%	-3,10%	-0,27%	2,38%	17,69%	8638
Earnings class - Previous Quarter Earnings class = 0											
Out-of-sample data	0,12%	6,53%	-0,03	6,88	(2,33)**	-19,57%	-2,59%	0,01%	2,82%	19,30%	17267
Earnings class - Previous Quarter Earnings class = 1											
Out-of-sample data	0,76%	6,43%	0,03	5,14	(10,04)***	-16,87%	-2,10%	0,29%	3,39%	19,68%	7213
Earnings class - Previous Quarter Earnings class = 2											
Out-of-sample data	0,98%	6,71%	0,56	3,92	(12,65)***	-16,79%	-2,14%	0,28%	3,46%	21,89%	7524

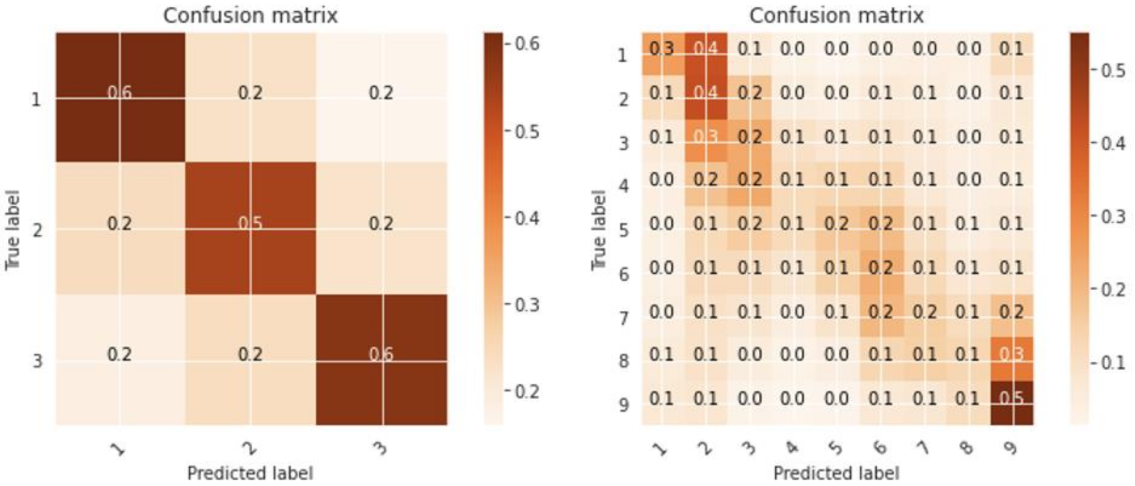
Appendix 4 – Average accuracy using Cross-Validation

Number of Class	Model				
	Logistic Classification	KNN	CatBoost	SVM	NN
3	57,89%	56,83%	59,64%	48,22%	56,83%
6	35,74%	36,91%	42,80%	27,59%	40,73%
9	24,59%	31,02%	31,02%	19,33%	29,70%

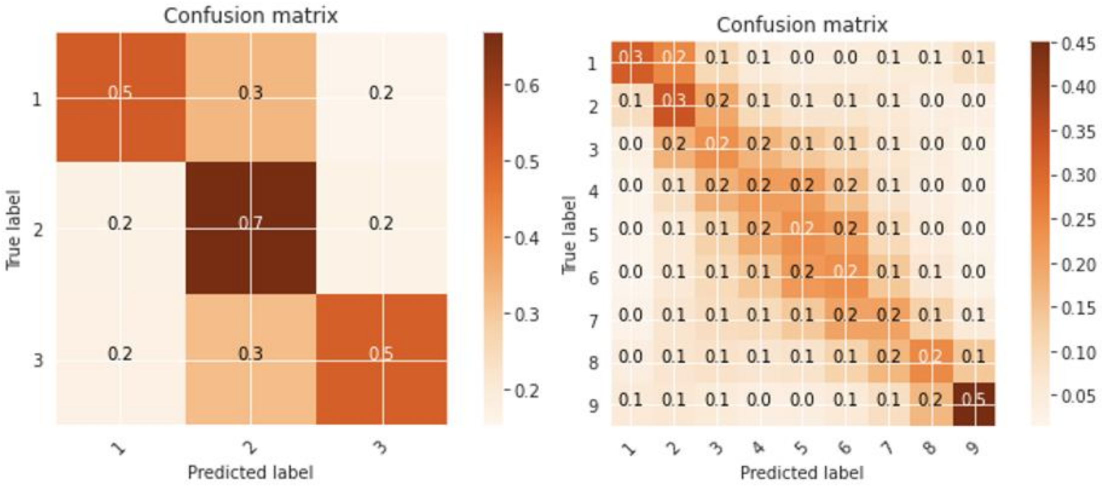
Appendix 5 – Variable importance

Variable	Importance
EPS_EVOLUTION_t-4	1
EPS_EVOLUTION_t-1	2
roce_t_1	3
roa_t_1	4
pcf_t_1	5

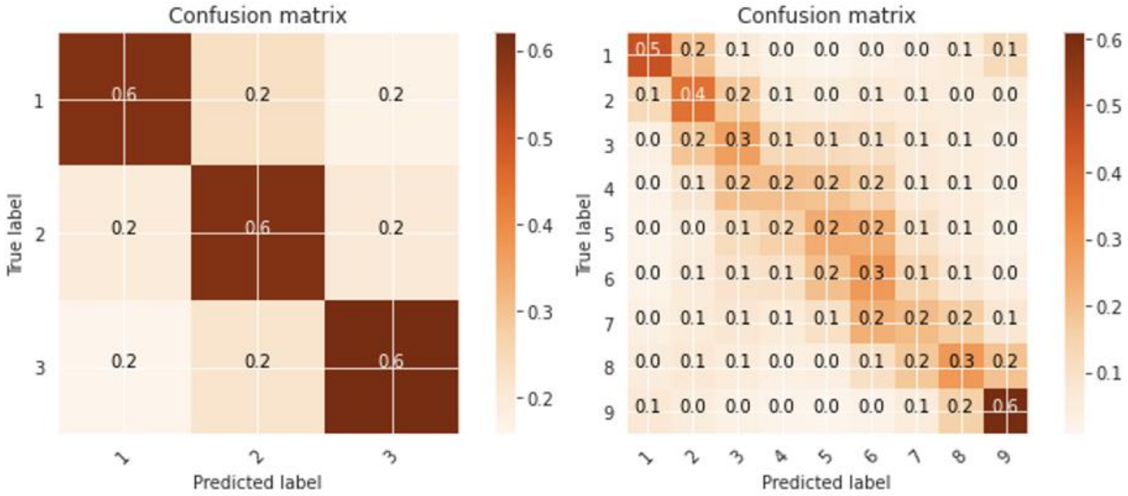
Appendix 6- Confusion matrices for the Logistic Classification for 3 classes of prediction on the left and 9 classes on the right



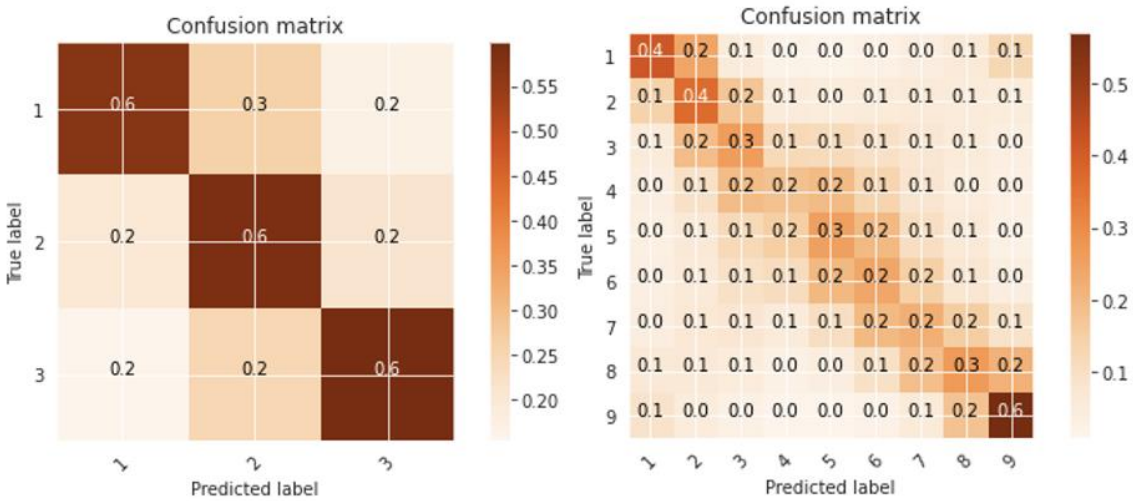
Appendix 7- Confusion matrices for the KNN Classification for 3 classes of prediction on the left and 9 classes on the right



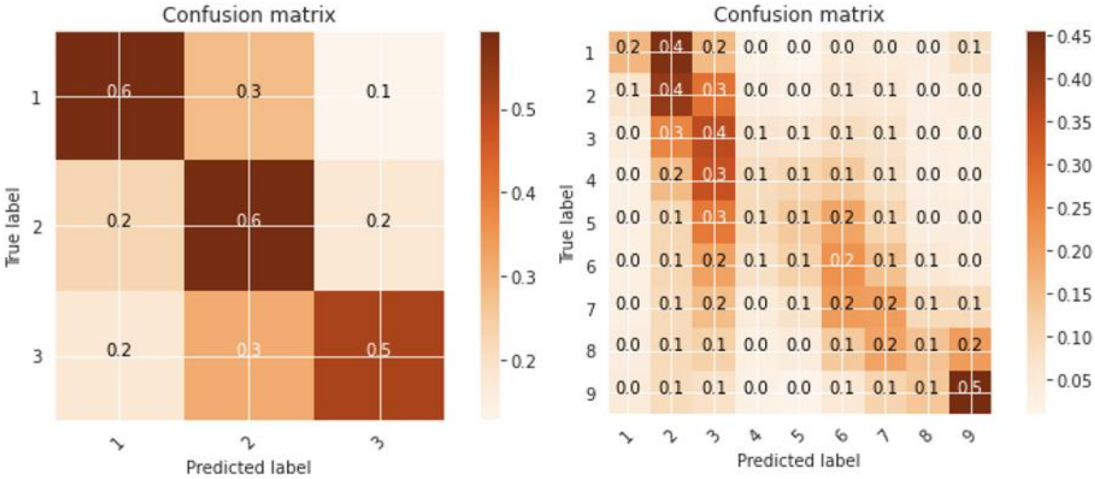
Appendix 8- Confusion matrices for the Catboost Classification for 3 classes of prediction on the left and 9 classes on the right



Appendix 9- Confusion matrices for the Neural Networks Classification for 3 classes of prediction on the left and 9 classes on the right



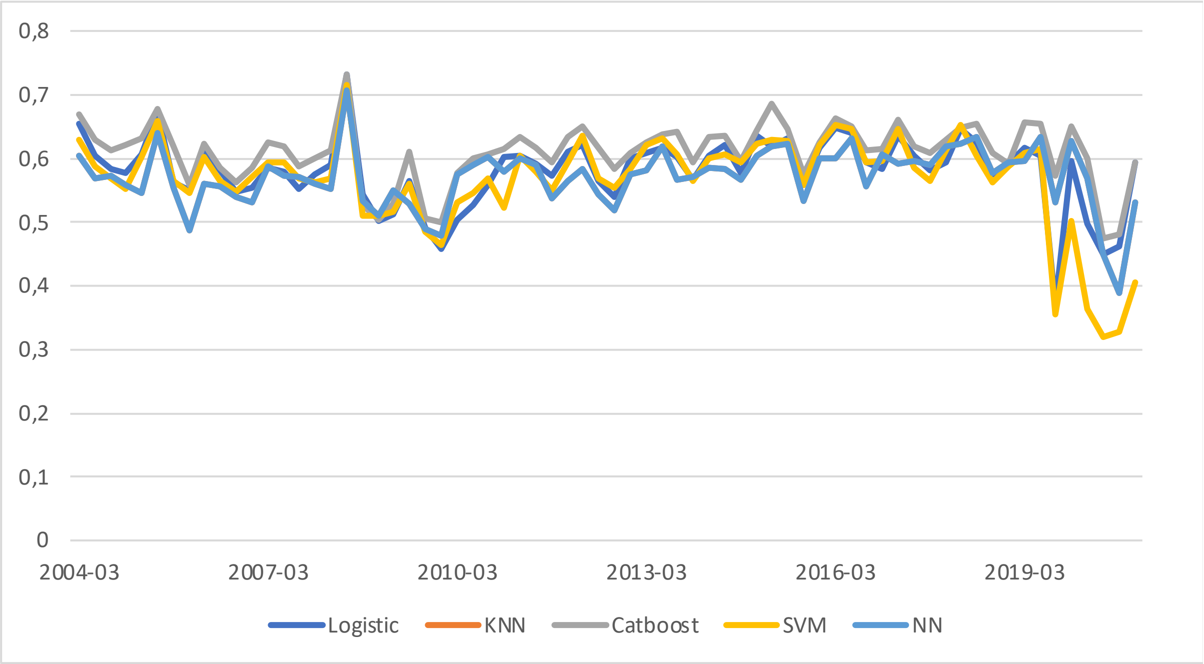
Appendix 10- Confusion matrices for the SVM Classification for 3 classes of prediction on the left and 9 classes on the right



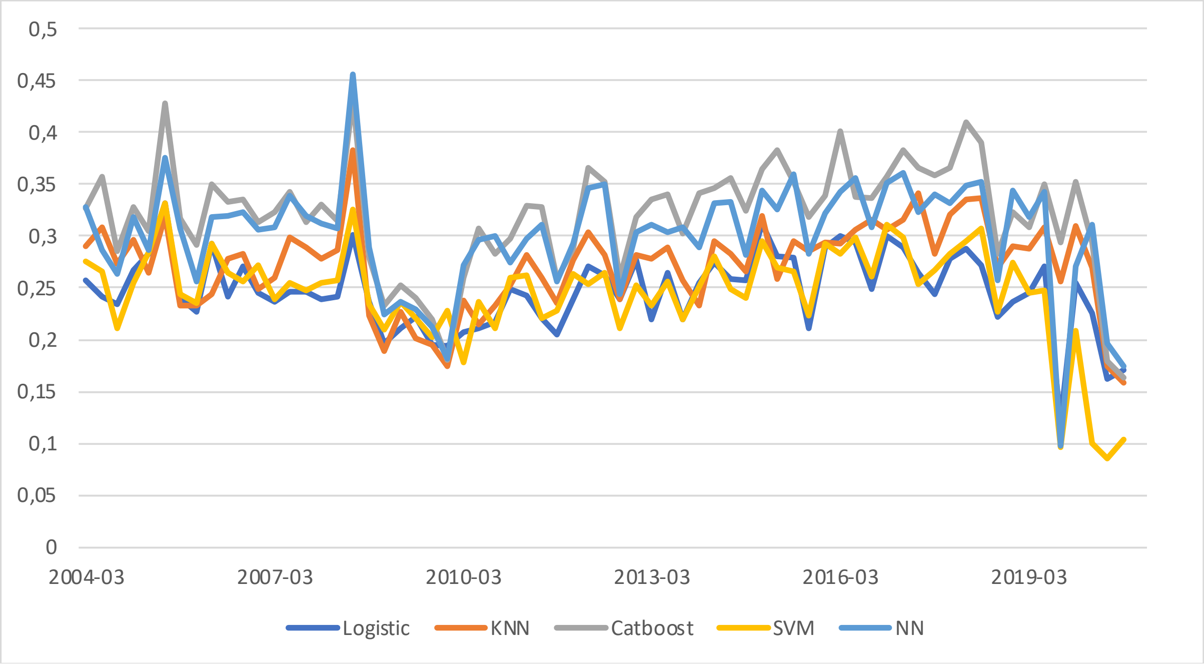
	1	2	3
1	0.6	0.3	0.1
2	0.2	0.6	0.2
3	0.2	0.3	0.5

	1	2	3	4	5	6	7	8	9
1	0.2	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.1
2	0.1	0.4	0.3	0.0	0.0	0.1	0.1	0.0	0.0
3	0.0	0.3	0.4	0.1	0.1	0.1	0.1	0.0	0.0
4	0.0	0.2	0.3	0.1	0.1	0.1	0.1	0.0	0.0
5	0.0	0.1	0.3	0.1	0.1	0.2	0.1	0.0	0.0
6	0.0	0.1	0.2	0.1	0.1	0.2	0.1	0.1	0.0
7	0.0	0.1	0.2	0.0	0.1	0.2	0.2	0.1	0.1
8	0.0	0.1	0.1	0.0	0.0	0.1	0.2	0.1	0.2
9	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.5

Appendix 11- Temporal accuracy of the Logistic, KNN Catboost, SVM and NN models for number of classes equal to 3



Appendix 12- Temporal accuracy of the Logistic, KNN Catboost, SVM and NN models for number of classes equal to 9



8. References

Abarbanell, J. S., & Bushee, B. J. (1997). Fundamental Analysis, Future Earnings, and Stock Prices. *Journal of Accounting Research*, *35*(1), 1. https://doi.org/10.2307/2491464

AFFLECK-GRAVES, J., DAVIS, L. R., & MENDENHALL, R. R. (1990). Forecasts of earnings per share: Possible sources of analyst superiority and bias. *Contemporary Accounting Research*, *6*(2), 501–517. https://doi.org/10.1111/j.1911-3846.1990.tb00771.x

Alwathnani, A. M., Dubofsky, D. A., & Al-Zoubi, H. A. (2017). Under-or-overreaction: Market responses to announcements of earnings surprises. *International Review of Financial Analysis*, *52*, 160–171. https://doi.org/10.1016/j.irfa.2017.07.006

Ball, R., & Brown, P. (1968). An Empirical Evaluation of Accounting Income Numbers. *Journal of Accounting Research*, *6*(2), 159. https://doi.org/10.2307/2490232

Beaver, W. H. (1968). The Information Content of Annual Earnings Announcements. *Journal of Accounting Research*, *6*, 67. https://doi.org/10.2307/2490070

Bhushan, R. (1989). Firm characteristics and analyst following. *Journal of Accounting and Economics*, *11*(2–3), 255–274. https://doi.org/10.1016/0165-4101(89)90008-6

Binder, J. (1998). The Event Study Methodology Since 1969. *Review of Quantitative Finance and Accounting*, *11*(2), 111–137. https://doi.org/10.1023/a:1008295500105

Bradshaw, M. T., Drake, M. S., Myers, J. N., & Myers, L. A. (2012). A re-examination of analysts' superiority over time-series forecasts of annual earnings. *Review of Accounting Studies*, *17*(4), 944–968. https://doi.org/10.1007/s11142-012-9185-8

Brown, L. D., Richardson, G. D., & Schwager, S. J. (1987). An Information Interpretation of Financial Analyst Superiority in Forecasting Earnings. *Journal of Accounting Research*, *25*(1), 49. https://doi.org/10.2307/2491258

Brown, L. D., & Rozeff, M. S. (1978). THE SUPERIORITY OF ANALYST FORECASTS AS MEASURES OF EXPECTATIONS: EVIDENCE FROM EARNINGS. *The Journal of Finance*, *33*(1), 1–16. https://doi.org/10.1111/j.1540-6261.1978.tb03385.x

Cao, Q., & Parry, M. E. (2009). Neural network earnings per share forecasting models: A comparison of backward propagation and the genetic algorithm. *Decision Support Systems*, *47*(1), 32–41. https://doi.org/10.1016/j.dss.2008.12.011

CHIANG, C., DAI, W., FAN, J., HONG, H., & TU, J. (2019). Robust Measures of Earnings Surprises. *The Journal of Finance*, *74*(2), 943–983. https://doi.org/10.1111/jofi.12746

Chopra, V. K. (1998). Why So Much Error in Analysts' Earnings Forecasts? *Financial Analysts Journal*, *54*(6), 35–42. https://doi.org/10.2469/faj.v54.n6.2223

Dbouk, B. (2017). Financial Statements Earnings Manipulation Detection Using a Layer of Machine Learning. *International Journal of Innovation, Management and Technology*, 172–179. https://doi.org/10.18178/ijimt.2017.8.3.723

Easton, P. D., & Harris, T. S. (1991). Earnings Asan Explanatory Variable for Returns. *Journal of Accounting Research*, *29*(1), 19. https://doi.org/10.2307/2491026

Etemadi, H., Ahmadpour, A., & Moshashaei, S. M. (2014). Earnings Per Share Forecast Using Extracted Rules from Trained Neural Network by Genetic Algorithm. *Computational Economics*, *46*(1), 55–63. https://doi.org/10.1007/s10614-014-9455-6

Falas, T., Charitou, A., & Charalambous, C. (1994). The application of artificial neural networks in the prediction of earnings. *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*. https://doi.org/10.1109/icnn.1994.374920

Fischer, J. A., Pohl, P., & Ratz, D. (2020). A machine learning approach to univariate time series forecasting of quarterly earnings. *Review of Quantitative Finance and Accounting*, *55*(4), 1163–1179. https://doi.org/10.1007/s11156-020-00871-3

Fried, D., & Givoly, D. (1982). Financial analysts' forecasts of earnings. *Journal of Accounting and Economics*, *4*(2), 85–107. https://doi.org/10.1016/0165-4101(82)90015-5

Givoly, D., & Lakonishok, J. (1984). The Quality of Analysts' Forecasts of Earnings. *Financial Analysts Journal*, *40*(5), 40–47. https://doi.org/10.2469/faj.v40.n5.40

Gu, Z., & Wu, J. S. (2003). Earnings skewness and analyst forecast bias. *Journal of Accounting and Economics*, *35*(1), 5–29. https://doi.org/10.1016/s0165-4101(02)00095-2

Kim, J. H. (2018). Market earnings expectation, measurement error in analysts' consensus forecasts and prediction of stock returns. *Accounting Research Journal*, *31*(2), 249–266. https://doi.org/10.1108/arj-03-2016-0031

Myers, J. N., Myers, L. A., & Skinner, D. J. (2007). Earnings Momentum and Earnings Management. *Journal of Accounting, Auditing & Finance*, *22*(2), 249–284. https://doi.org/10.1177/0148558x0702200211

O'brien, P. C. (1988). Analysts' forecasts as earnings expectations. *Journal of Accounting and Economics*, *10*(1), 53–83. https://doi.org/10.1016/0165-4101(88)90023-7

Rees, L. (2005). Abnormal Returns from Predicting Earnings Thresholds. *Review of Accounting Studies*, *10*(4), 465–496. https://doi.org/10.1007/s11142-005-4210-9

Shen, K. Y. (2012). The Modeling of Earnings Prediction by Time-Delay Neural Network. *Advanced Materials Research*, *433–440*, 907–911. https://doi.org/10.4028/www.scientific.net/amr.433-440.907

Shu, Y., Broadstock, D. C., & Xu, B. (2013). The heterogeneous impact of macroeconomic information on firms' earnings forecasts. *The British Accounting Review*, *45*(4), 311–325. https://doi.org/10.1016/j.bar.2013.06.011

van Binsbergen, J., Han, X., & Lopez-Lira, A. (2020). Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases. *NBERG*. https://doi.org/10.3386/w27843

Xinyue, C., Zhaoyu, X., & Yue, Z. (2020). Using Machine Learning to Forecast Future Earnings. *Atlantic Economic Journal*, *48*(4), 543–545. https://doi.org/10.1007/s11293-020-09691-1

Ye, Z. J., & Schuller, B. W. (2021). Capturing dynamics of post-earnings-announcement drift using a genetic algorithm-optimized XGBoost. *Expert Systems with Applications*, *177*, 114892. https://doi.org/10.1016/j.eswa.2021.114892

Zhang, W., Cao, Q., & Schniederjans, M. J. (2004). Neural Network Earnings per Share Forecasting Models: A Comparative Analysis of Alternative Methods. *Decision Sciences*, *35*(2), 205–237. https://doi.org/10.1111/j.00117315.2004.02674.x