Escola de Camins
Escola Tècnica Superior d'Enginyeria de Camins, Canals i Ports
UPC BARCELONATECH

Developing a Machine Learning based Systematic Investment Strategy: A Case Study of the Construction Industry

Treball realitzat per:
Eric Brea Garcia

Dirigit per:
Maribel Ortego Martinez
Alvar Garola Crespo

Màster en:
Enginyeria de Camins, Canals y Ports

Barcelona, 28 de Gener de 2020

Departament d'Enginyeria Civil i Ambiental

TREBALL FINAL DE MÀSTER

Abstract

In this research work, an end-to-end systematic investment strategy based on machine learning models and leveraging the construction industry operational and management practices knowledge, is implemented. First, a literature research in the field of behavioral finance is done, presenting the current state of the knowledge and trends in the industry. A suitable investment opportunity exploiting prevailing market inefficiencies around earnings announcements is identified. Second, an extensive literature research is performed identifying the most relevant characteristics of construction companies' operations and major risk factors they are exposed to. These insights are used to engineer a set of relevant variables. Third, advanced statistical techniques are used to select the most relevant subset of features, which includes market and analysts' expectation data, macroeconomic indicators, the delay in reporting earnings, and the most important financial dimensions for construction firms. Fourth, the earnings' surprise classification problem is characterized by a class imbalance and asymmetric misclassification costs. These issues are a consequence of the desired business application, and are addressed by selecting an appropriate evaluation metric. Additionally, considerations on the temporal dimension and generative process of the data are made to select an appropriate validation scheme. Five different state-of-the-art machine learning algorithms are considered: a multinomial logistic regression, a bagging classifier, a random forest, an XGBoost and a linear Support Vector Machine. The multinomial logistic regression is found to be the most suitable model, exhibiting a bias towards predicting positive earnings' surprises over the rest of classes. The firm size, and the profitability and valuation measures, portrayed by the Return on Assets and Enterprise Value multiples, are found to be the most important variables when predicting earnings surprises. To conclude, the systematic investment strategy based on the investment signals produced by the selected machine learning model is back-tested, being the performance of the long-short portfolio driven by the positive surprise one as a consequence of the selected model bias.

Keywords: Quantitative Investing, Machine Learning, Behavioral Finance

Resumen

En este trabajo de fin de máster se diseña una estrategia de inversión sistemática para las empresas de la construcción basada en técnicas de aprendizaje automático, sacando ventaja del conocimiento específico de como operan las mismas. Primero, se realiza una investigación bibliográfica en el campo de las finanzas del comportamiento, presentándose el estado actual del conocimiento y tendencias en la industria. Esta lleva a identificar una oportunidad de inversión viable alrededor de las sorpresas en la publicación de resultados trimestrales, aprovechando ineficiencias existentes en el mercado de valores. Segundo, se realiza un estudio en profundidad de las principales características de las empresas constructoras y los factores de riesgo a los que están expuestas, transformándose la información obtenida en diferentes conjuntos de variables representativos de los mismos. Tercero, se emplean técnicas estadísticas avanzadas para seleccionar el subconjunto de variables más relevante. Estas variables representan datos del mercado y expectativas de los analistas, indicadores macroeconómicos, el retraso de las empresas en reportar resultados y las dimensiones financieras más importantes para evaluar a las constructoras. Cuarto, el problema de clasificación de sorpresas resultante está caracterizado por un desequilibrio en las clases de la variable dependiente y una asimetría en los costes de clasificación errónea. Estas cuestiones son consecuencia del caso de uso escogido y se abordan con el diseño de una medida de evaluación adecuada. Además, se realizan consideraciones con respecto a la dimensión temporal y al proceso generativo de los datos, escogiendo un esquema de evaluación de modelos acorde. Se consideran cinco modelos diferentes: una regresión logística multinomial, un classificador *bagging*, un *random forest*, un *XGBoost* y una *Support Vector Machine* lineal. Finalmente, la regresión logística multinomial es seleccionada como el modelo más adecuado, teniendo este modelo un sesgo hacia la predicción de sorpresas positivas. Las variables identificadas como más importantes a la hora de identificar sorpresas son la capitalización bursátil, y los múltiplos financieros *Return on Assets* y *Enterprise Value* que representan las dimensiones de rentabilidad y *valuation*. Una vez evaluada la estrategia de inversión sistemática propuesta, se observa que el retorno de la cartera *long-short* está dominado por las sorpresas positivas como consecuencia del sesgo en el modelo.

Palabras clave: Inversión Cuantitativa, Aprendizaje Automático, *Behavioral Finance*

Acknowledgements

The following research paper, while an individual work, benefited from the insights and direction of UPC professor Mrs. Maria Isabel Ortego and Mr. Alvar Garola. The author feels an enormous debt of respectful gratitude to them, for their valuable guidance, keen interest and patience throughout the course of this work.

The author deeply appreciates the affection and support he enjoyed from his family and friends during this research study.

Index

1. INTRODUCTION 11

2. LITERATURE REVIEW 14

2.1. EFFICIENT MARKETS HYPOTHESIS AND BEHAVIORAL FINANCE 14
2.2. INVESTOR SENTIMENT AND INATTENTION 17
2.3. EARNINGS ANNOUNCEMENTS AND SURPRISES 18
2.4. PREVIOUS RESEARCH 22
2.5. PERFORMANCE OF THE PREVIOUS MODEL FOR CONSTRUCTION STOCKS 27

3. UNDERSTANDING CONSTRUCTION COMPANIES' OPERATIONS 29

3.1. INDUSTRY OVERVIEW 30
3.2. CONSTRUCTION STOCKS, ONE OF A KIND 31
3.3. RISK IN THE CONSTRUCTION INDUSTRY 34
3.3.1. Intrinsic Risk Factors 34
3.3.2. Global Risk Factors 35
3.4. FINANCIAL DISTRESS IN THE CONSTRUCTION INDUSTRY 37
3.5. POLITICAL RISK IN THE CONSTRUCTION INDUSTRY 43

4. RESEARCH METHODOLOGY 46

4.1. DATA SOURCES 46
4.1.1. CRSP Database 48
4.1.2. I/B/E/S Database 48
4.1.3. WRDS Financial Ratios Suite 49
4.1.4. Federal Reserve Economic data 49
4.1.5. World Bank Data 49
4.2. THE INVESTMENT FRAMEWORK 50
4.3. VARIABLE DEFINITION 52
4.3.1. Target variables 52
4.3.2. Features 53
4.4. EXPLORATORY ANALYSIS AND VARIABLE SELECTION 57
4.4.1. Univariate Analysis 58
4.4.2. Recursive Feature Elimination 61
4.5. MODEL SELECTION AND ASSESSMENT 61

4.5.1. Imbalanced data and asymmetric misclassification costs 64

4.5.2. Performance Metrics 65

4.5.3. Machine Learning Algorithms 67

5. RESULTS 70

5.1. DATA SAMPLE CHARACTERISTICS 70

5.2. FEATURE SELECTION 71

5.2.1. Market data and Expectations 71

5.2.2. Global Risk Factors 72

5.2.3. Intrinsic Variables 75

5.3. MODEL SELECTION AND ASSESSMENT 79

5.3.1. Model Selection 79

5.3.2. Final Model Assessment 81

5.4. INVESTMENT STRATEGY VIABILITY 84

6. CONCLUSIONS 86

6.1. SUMMARY OF THE RESULTS 86

6.2. CONTRIBUTIONS OF THIS WORK 89

6.3. ASSESSMENT OF THE PROJECT GOALS 90

6.4. FURTHER RESEARCH 90

REFERENCES 92

Figure Index

Figure 1. Different price reactions to an information shock: efficient market (left), under-reaction (center) and over-reaction (right). (Source: Landier [3])................................. 18

Figure 2. Cumulative Abnormal Returns for SUE based decile portfolios between 1974 and 1986. (Source: Thaler [7]) 20

Figure 3. Cumulative abnormal returns for long-short SUE decile portfolios formed after the earnings day and kept for the next 4 quarters. (Source: Thaler [7]) 21

Figure 4. Normalized confusion matrix for the final model (left), and derived feature importance(right) (Source: Brea Garcia [10]). 25

Figure 5. Cumulative abnormal returns of the positive surprise (car_port_3), negative surprise (car_port_1) and long-short portfolio (car_long_short) (Source: Brea Garcia [10]) 26

Figure 6. Cumulative abnormal returns of the positive surprise, negative surprise and long-short portfolio based on the previous model investment signals for construction firms.. 28

Figure 7. World Governance Indicators as reported by the World Bank for the United States from 1996 to 2018 (Source: World Bank [34]) 44

Figure 8. Timeline of the investment strategy... 50

Figure 9. Diagram of a 5-fold Cross Validation model selection strategy (Source: Pedragosa et al. [54]) 63

Figure 10. Example of a confusion matrix (Source: Chawla et al. [57])........................ 65

Figure 11. Distributions of the numerical market and expectations variables under the different categories of the target variable... 71

Figure 12. Distribution of the Global Risk Factor variables under the categories of the target variable. Note that the GNI (gni) and Political Instability variables had annual frequency and the annual value was propagated across all months, as opposed to the rest of variables which had monthly frequency... 74

Figure 13. Distribution of the delay in reporting earnings under the categories of the target variable. 75

Figure 14. Distributions of the considered financial ratios under the categories of the target variable with outliers. 77

Figure 15. Confusion matrix (left), and its normalized version (right). 82

Figure 16. Relative feature importance for the multinomial logistic regression model. 83

Figure 17. Cumulative abnormal returns of the positive surprise, negative surprise and long-short portfolio 85

Table Index

Table 1. Summary statistics of the numerical features used to feed the previous machine learning based model (Source: Brea Garcia [10]). 23

Table 2. Model Summary for the XGBoost with oversampling (Source: Brea Garcia [10]). 24

Table 3. SIC classification of the firms in the data sample. 27

Table 4. Previous model performance when it comes to predicting earnings' surprises in construction stocks. 27

Table 5. Financial ratios considered for evaluating the capital structure of a company. 39

Table 6. Financial ratios considered for evaluating the efficiency of a company. 39

Table 7. Financial ratios considered for evaluating the financial soundness of a company. 40

Table 8. Financial ratios considered for evaluating the liquidity of a company. 41

Table 9. Financial ratios considered for evaluating the profitability of a company. 41

Table 10. Financial ratios for evaluating the profitability of a company. 42

Table 11. Standard Industrial Classification Division and Major Groups for construction companies. 47

Table 12. SIC classification of the firms in the data sample. 70

Table 13. Results for the univariate statistical tests on the market and expectations variables. 72

Table 14. Results for the univariate statistical tests and Recursive Feature Elimination procedure on the Global Risk factors. 73

Table 15. Results for the univariate statistical tests on the different financial ratios. The selected financial ratios have been identified with an asterisk. 78

Table 16. Summary statistics of the numerical features used to feed the machine learning models. 79

Table 17. F-score with beta 0.5 micro- and macro-averaged across classes. The performance in the training and test set are shown. 80

Table 18. Precision and Recall macro-averaged for the relevant classes, positive and negative earnings' surprises. The performance in the training and test set are shown. . 80

Table 19. Model Summary for the Logistic Regression Model. 81

1. Introduction

Systematic investment strategies based in quantitative models have grown in popularity since the 2007 financial crisis, reshaping the asset management industry. This investment style relies on quantitative models or rule-based algorithms to make investment decisions and capital allocations. With traditional active management having underperformed the market in recent years, investors have begun to shift towards quantitative strategies in the search of more consistently superior returns. Moreover, the increase in computational power, the proliferation of alternative data and recent developments in artificial intelligence are profoundly changing the way we do business in what is already known as the fourth industrial revolution, being the investment industry no different. Is in that context that this research work is originated, aiming to provide *alpha* through a quantitative approach that takes advantage of novel machine learning techniques.

To further understand the context in which this research is born, a brief review of the author's background is needed. After graduating from the BSc in Civil Engineering at UPC, the author passion for numerical methods and statistics brought him to choose the Computational Engineering specialization during the MSc in Civil Engineering which this work concludes. It was during the latter, that the author got introduced into the machine learning field during the course *Models per a la Presa de Decisions y Optimització en Enginyeria;* which led him to enroll in the *MSc Data Science for Business* double-degree program between École Polytechnique and HEC Paris. During the last two years in Paris, the author not only got the opportunity to expand his knowledge in statistics and learn state-of-the art machine learning techniques, but also developed a passion for financial markets and the investment world. This led the author to write his thesis in HEC Paris around them, which he successfully defended in June 2019.

When looking for a topic for the current research work, the author aimed to bring together his civil engineering background with its masters' domain of specialization and interest for the financial world. As a result, the current topic was seen ss the ideal intersection between the three worlds and found relevant given the recent interest in cross-industry applications of the so-called data revolution.

Therefore, the goals set for this research project are the following:

- Get an in-depth understanding of the structure and operations of civil engineering companies, and the risk factors surrounding them.
- Gain more knowledge in the domain of behavioral finance.
- Develop an end-to-end data science project experiencing all the different steps required to bring a model into production, from data sourcing to variable and model selection.
- Get to implement the state-of-the-art machine learning techniques learned during the masters.

The goals set for this research project are ambitious, and the author aims to fulfill them within the limits and scope that characterize a master thesis. In practice, developing a systematic investment strategy is a complex task which involves several highly specialized teams. Sales teams will identify the client needs and potential products with market demand, equity research teams and strategists will provide industry specific information and a cross-asset strategic view, while structuring and quantitative research teams will be in charge of developing and implementing the strategy and traders will assess the its implementation viability and costs. As a result, developing a systematic investment strategy is a multi-disciplinary task seldomly done by a single individual that requires different capabilities. Given the complexity of the task in hand, the author aims to get a thorough understanding of the process as a whole and of all the different parts and stakeholders involved; at the expense of not being able to go in depth into each one of the different steps of the process.

The rest of this research work is structured as follows. In section 2, a brief introduction to the field of behavioral finance is done, and the investment opportunity identified. Also, the previous work performed by the author is presented, and particularized for the construction industry, laying the perfect starting point for this research. In section 3, an extensive literature research on the civil engineering industry and the operations of construction enterprises is performed. This includes an overview of the risks surrounding the construction business, and a deep dive in the industry's exposure to financial distress and political risk. In section 4, the research methodology is exposed presenting the investment framework, variable definitions, methodologies for variable and model selection and assessing the performance of the final model. In section 5, the results for the different phases described in the research methodology are presented and analyzed.

To finish, in the conclusions the research results are summarized, the main contributions of this work presented, the fulfillment of this project goals assessed, and further lines of research outlined.

2. Literature review

In the following section an introduction to the theoretical and empirical grounds of the problem of predicting earnings' surprises is given. This includes a brief introduction to the efficient markets' hypothesis and the behavioral finance literature, focusing in the earnings' surprises events and the market inefficiencies that surround them. Also, the state of the art when it comes to predicting earnings' surprises and construction companies' financial distress is presented.

Next, the previous research carried out by the author is presented. First, the more general model developed, and approach taken, are briefly introduced; and then the results derived particularized to construction stocks, the ones relevant for this work. This section, far from summarizing the previous research, aims to give the reader the context and background in which this work lays and set the perfect ground for the research developed in the following sections.

2.1. Efficient Markets Hypothesis and Behavioral Finance

The main purpose of capital markets is channeling the wealth of savers, retail or institutional investors, into organizations that can invest it to put it into productive use, such as governments, companies and individuals. In the optimal case, the companies in the market make their production-investment decisions and their market price will reflect all available information regarding them, so investors can make their capital allocation decisions in an 'efficient' way [1].

The *Efficient Markets Hypothesis* (EMH) has been one of the most studied and tested theories in financial economics. It was proposed by Eugene Fama in 1970, where he defined an efficient financial market as 'one in which security prices always fully reflect available information' [1]. In a more practical sense, this definition implies that it's impossible to earn superior risk-adjusted returns when trading on available information, since it‘s already incorporated in security prices.

In his original formulation of the EMH, Fama distinguished between three different types of information sets, yielding to different forms of market efficiency [1]. First, the *weak-form* of market efficiency when only market data is considered (i.e. past prices, trading volumes and returns). Under this hypothesis future returns cannot be predicted based on past returns. Second, the *semi-strong from* of market efficiency when considering all publicly available information (i.e. earnings and dividend announcements, stock splits, executive compensation, etc.). This form of market efficiency implies that when information is released to the public it's immediately incorporated into market prices, thus making it impossible to earn abnormal returns trading on it [2]. Third, the *strong form* of market efficiency when all kinds of relevant information for price formation, including inside information, are considered. This assumption implies that it is not possible to make a profit on inside information, since it is also quickly leaked and incorporated into prices. It is important to note that these information sets are subsets of each other, being market data a subset of all public information, which is in turn a subset of all available information. As a result, the strong form of market efficiency is also semi-strong efficient and obviously weak efficient.

The formulation of the EMH spurred in the 1970s an effort from the academic community to develop a theoretical framework for it and test its predictions, leading to findings that supported the weak and semi-strong form of market efficiency. As Michael Jensen put it in 1978, 'there is no other proposition in economics which has a more solid empirical evidence supporting it than the Efficient Markets Hypothesis' [2].

The theoretical foundations of the EMH lay on three basic principles. First, investor rationality which makes markets efficient by definition. If investors are rational, they value securities by their fundamental value[1] and by quickly reacting to new information by trading on it, they incorporate it to prices leading to an immediate update of its fundamental value. Second, when investors do not behave in a rational manner their trades are uncorrelated, thus they cancel out without affecting security prices. This argument is heavily dependent on the non-correlation between irrational investors trading strategies, which is unlikely to hold. Third, in the case of irrational investors trading in similar ways, they are met by arbitrageurs bringing prices back to their fundamental values. In this case, and when perfect substitute securities are available, the competition between arbitrageurs prevents the price to deviate substantially from its fundamental value, thus in turn limiting their capacity to earn substantial abnormal returns. Moreover, irrational investors activity

[1] The fundamental value of a security can be defined as 'the net present value of its future cash flows, discounted using their risk characteristics' [2].

results in lower returns than arbitrageurs and passive investors leading to a reduction in their wealth and ultimately their extinction, thus disappearing with them the mispricing [2].

In a more practical way, the EMH implied that news affecting a security are incorporated in its price fast and accurately without leading to under or overreaction (see section 2.2); and that security prices should not react to non-information. In other words, there shouldn't be price trends nor price reversals after the news release, and security prices shouldn't be affected by variations in supply or demand if there are no news that alter their fundamental value [2].

The main implication of market efficiency is that *stale* information, understood as one of the information sets mentioned above, can't be used to *make money*. As Shleifer [2] points out, in a financial context making money is defined as earning 'a superior return after an adjustment for risk', since earning a profit as a result of trading on a set of information might just be a fair market compensation for the risk incurred. This leads us to the complicated and controversial task of measuring risk, another well studied problem in the financial literature, where several models have been proposed. The most used risk model is the *Capital Asset Pricing Model* proposed by Sharpe in 1964 [2]. As a result, most tests of market efficiency are dependent on the model of risk used.

In the 1980s, the EMH theoretical base was challenged and new empirical evidence against market efficiency appeared. First, psychological findings proved that investors do not behave in a rational manner, and that deviate from it in several fundamental ways. The main deviations from the standard decision-making model are: investors attitude towards risk displaying loss-aversion, a non-Bayesian expectation formation and investors experiencing a framing bias in their decision-making process [2]. Second, the psychological evidence also revealed that investors' deviations from rationality are far from being random and are highly correlated. Furthermore, these deviations are not only affecting retail investors but also professional money managers, who are affected by the same biases. Third, risk-less arbitrage does not exist in practice limiting arbitrageurs' capacity to bring back prices to fundamental values. Arbitrageurs' activity is heavily reliant on the existence of securities' perfect substitutes to use as a risk-less hedge. In the real world these close substitutes seldomly exist making arbitrage risky, thus diminishing the interest in such trades and limiting the ability of arbitrageurs to bring prices back to fundamental values [2]. Also, arbitrageur's wealth is finite and their capacity to bear losses limited, thus their ability to maintain a position until the mispricing disappears. This makes, in turn, arbitrage risky even in the case of the existence of perfect substitutes.

To the theoretical challenges exposed above, there were also empirical findings challenging the EMH. The main evidence against it were: findings of stock prices overreaction, success in predicting future returns based on past returns, evidence of size and market to book ratios as predictors of returns, and security price reaction to non-news, among others [2].

The appearance of the theoretical challenges and empirical evidence contradicting the *Efficient Markets Hypothesis*, has given rise to a new area of research: behavioral finance. Behavioral finance is a field of behavioral economics that studies the 'human fallibility in competitive markets, […] and examines what happens to prices and other dimensions of market performance when different types of investors trade with each other' [2]. The behavioral finance theory sits along two axes: limited arbitrage and a theory for investor sentiment to make predictions about security prices and returns. The former focused on studying the limits of arbitrage activities, and the latter on analyzing how market participants form their beliefs and valuations and thus structure their demand.

2.2. Investor Sentiment and Inattention

Traditionally, we understand the finance industry as an information processor in a Bayesian setting, where market participants have a *prior* belief which update with new information to generate a *posterior* belief [3]. The behavioral finance literature proposes a model for investor sentiment in line with the psychological evidence and deviating from the traditional model. In this model, when in presence of new information investors fail to sufficiently incorporate the new information and 'stick' to their prior beliefs, leading to price under-reaction. At the same time, when receiving similar news they tend to give more importance to the new information than their prior beliefs leading to over-reaction [4] [2].

Inattention to publicly available information (i.e. news) has been well documented by the finance literature [2], leading to two major behaviors when investors process an information shock: under-reaction and over-reaction. On the one hand, under-reaction consists in prices trending up/down after a positive/negative information shock as a result of slow processing of information by market participants. This implies that future returns can be predicted from current news. On the other hand, over-reaction consists in an over/under pricing of securities as a result of a sequence of positive/negative news over time which eventually experiences a reversion to its fundamental value. Inattention, and its derived phenomena of under and over reaction to news violate the *weak* and *semi-strong* form of market efficiency, leading to inefficiencies that can be exploited by traders.

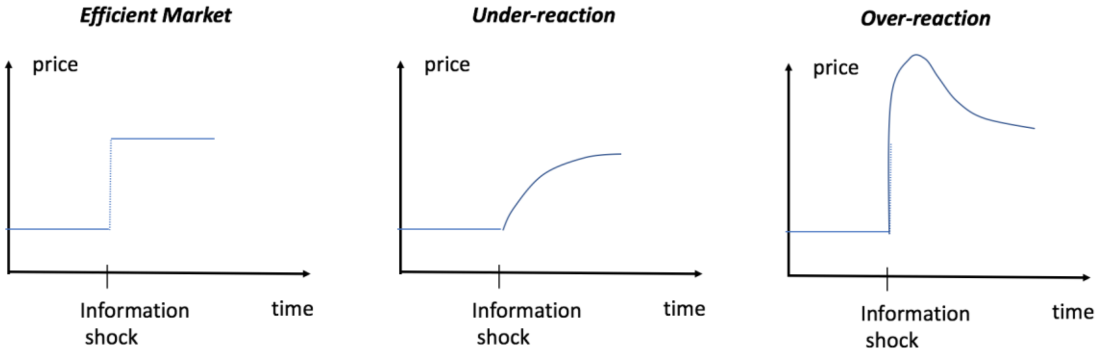


Figure 1. Different price reactions to an information shock: efficient market (left), under-reaction (center) and over-reaction (right). (Source: Landier [3]*)*

In figure 1, the different price reactions to an information shock are shown, to illustrate the exposed phenomena. The fast and accurate price adjustment predicted by the EMH (on the left), contrasts with the price drifts experimented as a result of inattention to news by market participants.

2.3. Earnings Announcements and Surprises

Public companies have periodic information disclosure obligations with the *Securities and Exchange Commission* (SEC) aiming to keep shareholders informed of the company's operations and financial health in a regular basis. These reporting obligations include the filing of Annual and Quarterly reports among others, where accounting information and the management's strategic view is released [5]. Each company can choose their reporting time after the end of the fiscal period within the SEC restrictions, being companies with a market capitalization over 75 million required to report within 40 days [6].

When it comes to publicly traded companies one of the most relevant and periodic news releases is the *Earnings Announcement Day* (EAD), where Quarterly reports are published. The EAD date is communicated in advance, a few weeks after the end of the fiscal period, and forecasts about the different metrics released are posted by equity analysts before the announcement. An earnings surprise takes place when the reported metrics differ significantly from investors' expectations, known as a positive or negative surprise depending on its direction.

When there is an earnings surprise, stock prices experiment price drifts around the EAD depending on the nature of the surprise:

(a) *Pre-announcement drift:* On average before a positive/negative surprise, firms experience an upward/downward price drift up to sixty days before the EAD, depending the abnormal returns delivered on the magnitude of the surprise.

(b) *Announcement day:* when a positive/negative surprise is posted, stock prices tend to experience a steep jump in the same direction of the surprise as the news are incorporated in the price.

(c) *Post-announcement drift:* On average, after the earnings day, stocks that posted a positive/negative surprise keep on drifting in the same direction for up to sixty days.

The above-mentioned phenomenon can be seen in figure 2, where decile portfolios are formed based on the *Standardized Unexpected Earnings*[2] (SUE) metric for earnings' surprises sixty days prior to the EAD and kept for sixty days after it. On average firms with a positive surprise experience a 4% upwards drift before the EAD and a 2% after it. For those affected by negative surprises the pre- and post-earnings downward drift are of -6% and -2% on average.

The price drift that precedes the earnings day, and the jump on the same day, can be explained as market participants incorporating new information regarding the earnings release to stock prices, as its common practice by managers to give hints on the results to boil down its impact. What it's striking, and violates market efficiency, it's the post-announcement drift, which can be explained in terms of under-reaction from market participants to earnings news and them being slowly incorporated in market prices [7].

[2] Here the SUE are calculated by producing a statistical forecast and computing the normalized forecast error. An alternative way of computing it is using analysts' consensus and I/B/E/S actuals. [65]

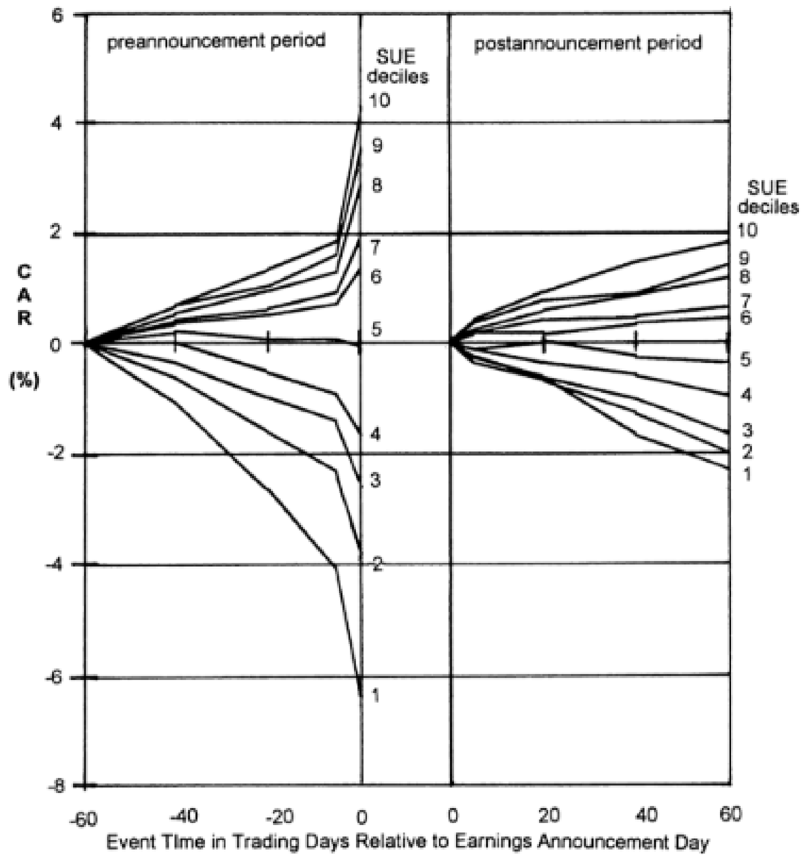


Figure 2. Cumulative Abnormal Returns for SUE based decile portfolios between 1974 and 1986. (Source: Thaler [7]*)*

There is empirical evidence that the earnings announcement drift is also influenced by the size of the company, being it in average larger for smaller companies [7]. This can be explained in terms of the degree of inattention given that analyst coverage it's inversely related to company size.

As commented above, the post-announcement drift can be explained as market participants failing to fully incorporate in prices the impact of earnings in future earnings announcements. This was corroborated by Bernard & Thomas [8] who found that earnings surprises are positively autocorrelated the next three quarters with decreasing magnitude and negatively autocorrelated with the fourth. This, as well as the size effect, can be seen in figure 3.

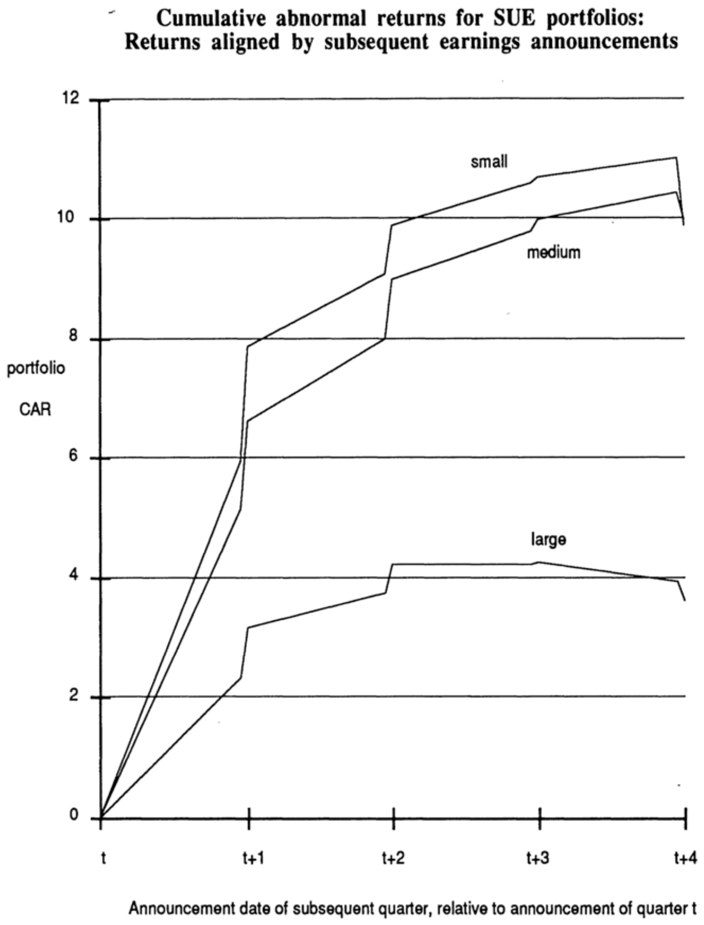


Figure 3. Cumulative abnormal returns for long-short SUE decile portfolios formed after the earnings day and kept for the next 4 quarters. (Source: Thaler [7])

As the reader might infer, the correlation between earnings surprises and returns sets up an opportunity to earn abnormal returns if one can forecast earnings surprises. The problem of predicting earnings surprises is complex and not new, and it has been widely studied by the finance academic community.

This problem has not only been studied by traditional linear models, but also by more novel non-linear methodologies. Dhar and Chou [9], compare the performance of four machine learning models: neural networks, induction algorithms, naïve Bayesian learning and genetic algorithms. The use of these techniques allows them to 'discover' the two non-linear relationships above mentioned: between the earnings surprise and the company size and the autocorrelation of surprises across adjacent quarters [9]. These relationships do not represent a novel finding, since they are well documented in the literature but allows them to showcase the power of machine learning techniques versus linear methods when it comes to unveiling interesting non-linear relationships in the data.

2.4. Previous research

As previously mentioned in the introduction, this research project sits on top of previous work from the author. Despite this being an independent work by itself, the author feels that it’s of interest for the reader to get an overview of the previous research, since one model is built on top of the results and conclusions extracted from the previous one. Moreover, in this work the previous model results are particularized to the civil engineering companies in the previous data sample and used as a benchmark to assess the performance of the new model.

The previous work aimed to develop a profitable investment strategy that took advantage of the pre- and post-earnings announcement price drifts by predicting earnings surprises at the end of each fiscal period and building long-short portfolios based on the earnings surprise direction. With that goal in mind, two different data sources were used: the IBES and CRSP databases. Both datasets are widely used in the financial industry and academic community, and provide different market information about security prices, trading volumes, company information and descriptive statistics of analyst expectations and revisions. These data sources are also used in this research and will be presented in more detail (see sections 4.1.1 and 4.1.2).

Two different approaches where developed to achieve that goal, each one of them using different levels of granularity of the data. On the one hand, the first approach aimed to leverage the historical performance of analysts to create a better estimate than the consensus by removing bad analysts and keeping only timely estimates. The improved estimate was used to forecast earnings’ surprises by comparing it to the consensus and assuming the divergence to be a good proxy of earnings surprises. This approach proved to have no predicting power and the investment strategy based on it failed to achieve abnormal returns. On the other hand, the second approach aimed to directly predict earnings’ surprises from market and analysts’ expectations data using state-of-the-art machine learning techniques. Is this approach that is relevant to this research work and will be now presented in more detail.

For prediction purposes, surprises were measured as a standardized version of the analyst’s consensus forecast errors, and categorized as: positive surprise, no surprise and negative surprise. Two different sets of features were developed to feed the model, which can be classified as market data and analysts' expectation based. The first group included the firm's size, company and industry returns trends approaching the decision day and past surprises information. The second set of features included a proxy for the quality of the analysts' consensus relative to their peers, analyst's revisions trends and consensus characteristics (i.e. standard deviation, number of estimates, etc).

The sample used to feed the model consisted of the 3,000 largest U.S. listed firms of each year for a 20-year interval ranging from January 1999 to December 2018. It included 7,574 stocks covered by 14,148 analysts and 776 brokers. In the following table the different variables and its descriptive statistics are shown:

Table 1. Summary statistics of the numerical features used to feed the previous machine learning based model (Source: Brea Garcia [10]).

feature	***count***	***mean***	***stdev***	***min***	***p25***	***p50***	***p75***	***max***
StarRatio	*133,561*	*0.359*	*0.116*	*0.08*	*0.286*	*0.333*	*0.400*	*1.000*
AFD1	*219,743*	*-0306*	*2.342*	*-43*	*0*	*0*	*0*	*37*
AFD2	*216,908*	*-0.907*	*5.020*	*-46*	*-2*	*0*	*1*	*41*
AFD3	*213,374*	*-0.303*	*2.711*	*-42*	*-1*	*0*	*0*	*41*
STDEV1	*193,292*	*0.044*	*0.413*	*0*	*0.010*	*0.020*	*0.040*	*151*
STDEV2	*189,935*	*0.045*	*0.302*	*0*	*0.010*	*0.020*	*0.050*	*84.85*
STDEV3	*185,311*	*0.051*	*0.397*	*0*	*0.010*	*0.030*	*0.050*	*138.56*
NUMEST1	*219,743*	*7.571*	*6.509*	*1*	*3*	*6*	*10*	*50*
NUMEST2	*216,908*	*7.411*	*6.442*	*1*	*3*	*5*	*10*	*50*
NUMEST3	*213,374*	*6.975*	*6.135*	*1*	*2*	*5*	*9*	*49*
size	*218,062*	*5,139,894*	*2.11e+7*	*1530*	*243,147*	*741,924*	*2,6e+6*	*8.96e+8*
car1	*218,062*	*0.002*	*0.138*	*-2.694*	*-0.059*	*-0.003*	*0.053*	*5.175*
car2	*217,431*	*0.001*	*0.196*	*-1.623*	*-0.092*	*-0.009*	*0.075*	*9.332*
car3	*216,757*	*-0,004*	*0.252*	*-5.843*	*-0.126*	*-0.020*	*0.088*	*10.924*
indtr1	*173,928*	*0.019*	*0.097*	*-0.588*	*-0.036*	*0.010*	*0.059*	*2.964*
indtr2	*173,909*	*0,040*	*0.153*	*-0.689*	*-0.045*	*0.020*	*0.099*	*4.1020*
indtr3	*173,864*	*0,089*	*0.252*	*-0.849*	*-0.054*	*0.041*	*0.174*	*4.912*

This yield to a multi-class classification problem with an imbalance dataset and eight different regressors introduced for different time periods before the observation time stamp aiming to capture trends in the data. This is a challenging setting for machine learning algorithms, since our categories of interest are the minority classes, and they will tend to overfit the majority class. An important assumption made, was that by introducing the regressors for different time periods all the important information at a given point in time was already incorporating without need to consider the temporal dimension. This allowed considering some state-of-the-art resampling techniques that aimed to rebalance the dataset, which were introduced as a preprocessing step for the machine learning algorithms. The selected learning technique was an XGBoost model on a previously rebalanced dataset by synthetically oversampling the minority classes using SMOTE methodology. The results of the model are presented in the following table:

Table 2. Model Summary for the XGBoost with oversampling (Source: Brea Garcia [10]).

Target Variable	***Precision***	***Recall***	***F1 score***	***Support***
Negative Surprise	0.40	0.32	0.36	2,226
No Surprise	0.84	0.94	0.89	18,654
Positive Surprise	0.49	0.24	0.32	3,074
Micro avg.	0.79	0.79	0.79	23,954
Macro avg.	0.58	0.50	0.52	23,954
Weighted avg.	0.76	0.79	0.77	23,954

The resulting model was able to classify surprises, either positive or negative, slightly less than one third of the cases, and of those returned only classified correctly around 45% of the cases. Regarding the majority class, no surprise, the model was able to return almost all the cases (94% recall) and to correctly classify 84% of them. The macro-averaged F_1score was of 0.52, and its divergence with the micro-averaged one portrayed the different classification performance across classes. The macro-average being significantly lower than the micro-average, indicates that the minority classes are being poorly classified, while achieving good classification results in the majority class.

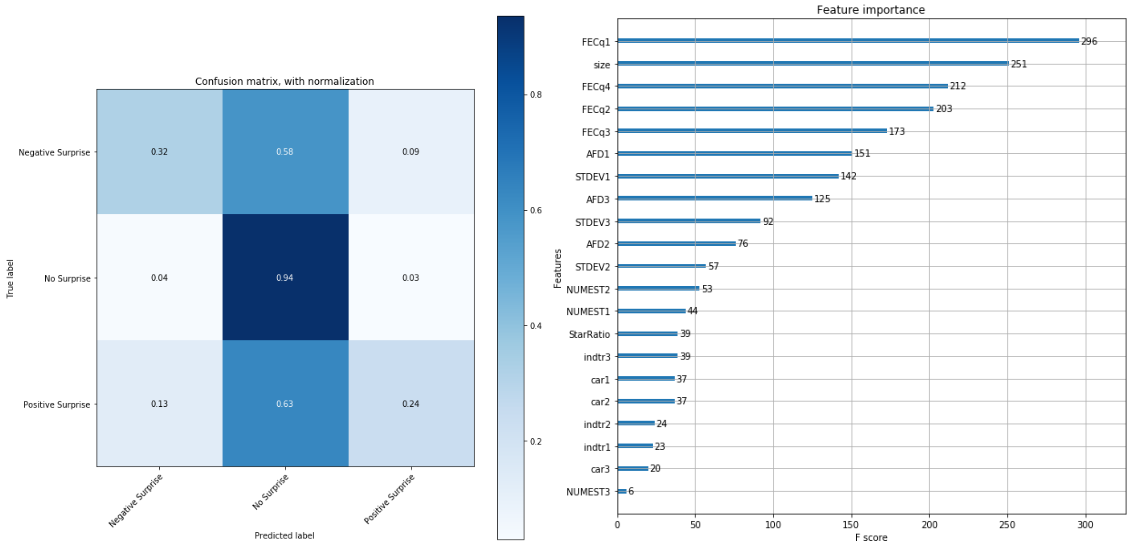


Figure 4. Normalized confusion matrix for the final model (left), and derived feature importance(right) (Source: Brea Garcia [10]).

It’s worth noting, that in general the model did not misclassify surprises’ in the opposite direction and that it had a strong bias towards predicting the dominant class, as illustrated in the confusion matrix in figure 4. Also, the model had the same performance for predicting positive and negative surprises, without exhibiting a better performance in one group.

An important result of the above presented model was the variable importance, being it able to identify some relationships well documented in the literature (figure 4, right). The most important variables when trying to predict earnings’ surprises were the firm size and past surprises for the previous four quarters. This is consistent with the behavioral finance literature, where the autocorrelation of earnings surprises and firm size effect is well-documented. Apart from those, only two other features prove to somehow important: the standard deviation and analysts’ forecast differences of the previous consensus. A possible explanation for the standard deviation relevance is that a higher dispersion in the consensus conveys more uncertainty regarding the earnings’ release, leaving more room for a surprise. Regarding the difference in analysts’ revisions, high values in any direction may convey the occurrence of an information shock that if not accounted for by the majority of analysts, may lead to a surprise.

The investment strategy based on the signals coming from that approach yielded poor results (figure 6). The cumulative abnormal returns of the positive surprise portfolio are slightly better than the ones of the negative surprise portfolio for most part of the considered time span, reversing for the last year considered. The resulting long-short portfolio achieved slightly positive abnormal returns for the first to years of the sample, dropping into negative territory for the rest of the back test. Overall, the designed investment strategy fails to consistently earn abnormal returns for the time period considered.

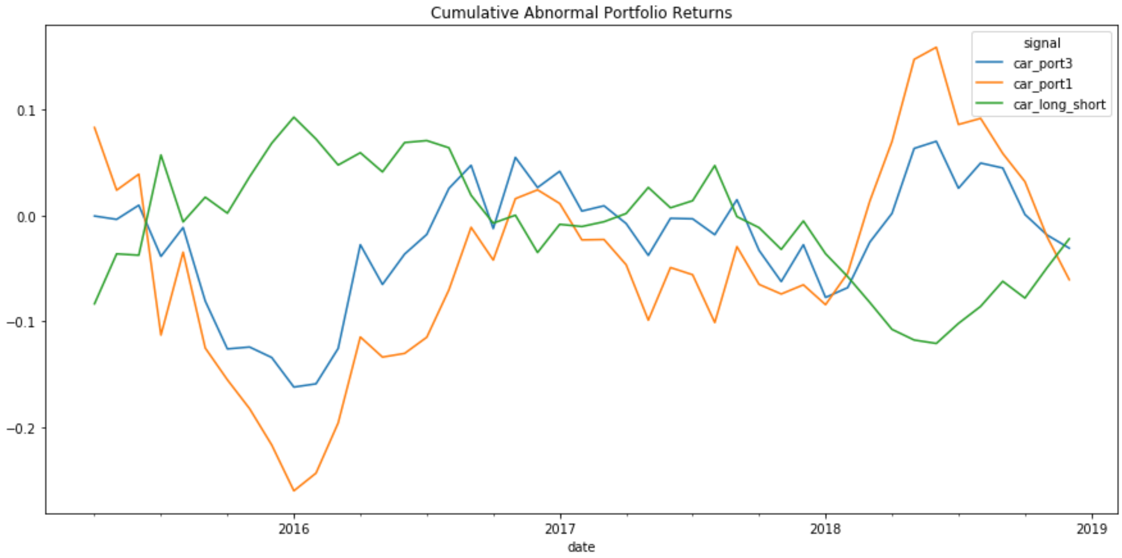


Figure 5. Cumulative abnormal returns of the positive surprise (car_port_3), negative surprise (car_port_1) and long-short portfolio (car_long_short) (Source: Brea Garcia [10]*)*

2.5. Performance of the previous model for construction stocks

In the previous study the data sources used can be deemed as generic and of common use among market participants, thus being its predictive ability for earnings' surprises limited. The question that the previous research raises is if an improved model developed for a particular industry and leveraging industry specific knowledge will fare better. In this research, the main goal is developing an investment strategy that profits from earnings surprises in the construction industry, by leveraging the civil engineering industry specific knowledge acquired during the degree.

As a starting point, the first thing to do is to see how the previous model performs when it comes to forecasting construction firms' earnings surprises, so it can be used as a benchmark for the model developed in this research. The original data sample is restricted to only construction stocks according to the SIC classification[3], and the resulting test set consists of 32 construction firms with the following classification:

Table 3. SIC classification of the firms in the data sample.

Division	*Major Group*	*Number of firms*
C Construction	*15 Building Construction, General Contractors and Operative Builders*	18
	16 Heavy Construction other than Building construction contractors	7
	17 Construction special trade contractors	7

The model performance achieved when restricted to the construction firms is presented in the following table:

Table 4. Previous model performance when it comes to predicting earnings' surprises in construction stocks.

Target Variable	*Precision*	*Recall*	*F1 score*	*Support*
Negative Surprise	0.29	0.26	0.27	47
No Surprise	0.82	0.91	0.86	286
Positive Surprise	0.20	0.07	0.11	42
Micro avg.	0.73	0.73	0.73	375
Macro avg.	0.44	0.41	0.41	375
Weighted avg.	0.68	0.73	0.70	375

[3] Companies classified as main groups 15, 16 and 17 according to SIC classification (see section 4.1).

We can see that the model performs worst when only considering the construction stocks results. The macro-averaged F_1score has dropped to 0.41, being able to classify better negative rather than positive surprises. The model is able to return 26% of the negative surprises and only 7% of the positive ones, classifying correctly 29% and 20% of the cases respectively. Again, the difference between the macro- and micro-averaged scores, indicates a good performance for the majority class, while performing poorly in our classes of interest. The macro-averaged precision and recall for the surprise categories are also computed, being the model able to identify 16% of the surprises and correctly classify 25% of them.

Regarding variable importance, since the model has still been trained with the full sample it remains the same. In ensemble methods variable importance is retrieved as a measure of feature relevance in the construction of the sequence of trees. Hence, the most important features are the firm size, past earnings surprises, the standard deviation of the consensus and the analysts' forecast differences. These variables are going to be used in the model developed in this research. And it will be expanded with industry specific variables as developed in the following sections. The resulting investment strategy, when only construction stocks are considered, is presented in figure 6. The long-short portfolio does not appear to achieve cumulative abnormal returns during the back test.

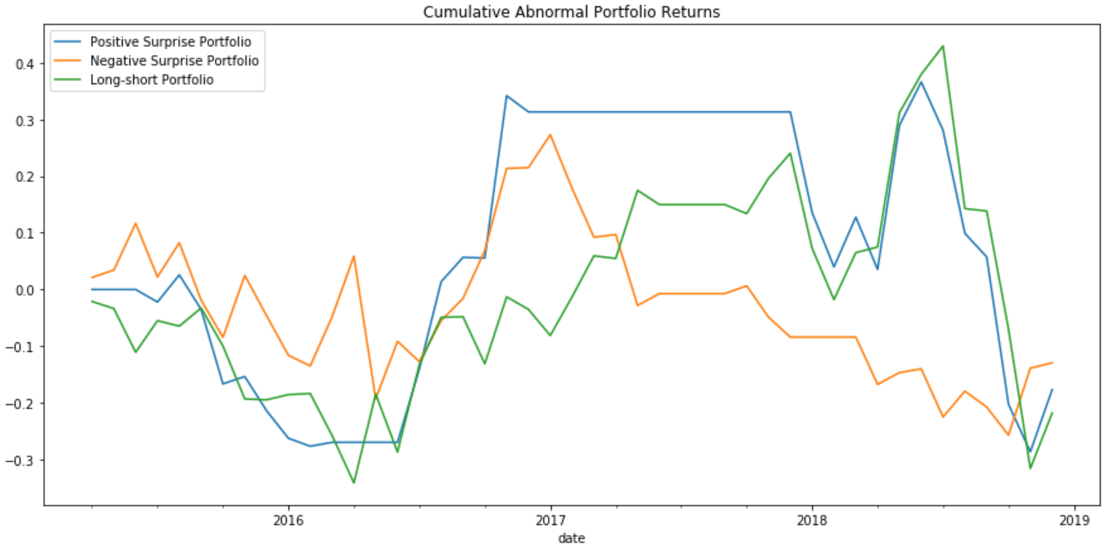


Figure 6. Cumulative abnormal returns of the positive surprise, negative surprise and long-short portfolio based on the previous model investment signals for construction firms.

3. Understanding Construction companies' operations

In order to properly develop relevant variables to assess construction companies' performance and support the concerning investment decision, an in-depth knowledge of the market structure, operational characteristics and sources of risk of the construction industry is needed. The following section analyzes the above-mentioned factors for all construction companies across all geographic regions, despite in the upcoming model we are going to restrict ourselves to publicly listed companies in the United States. The reason behind this limitation, is not only our desired investment strategy which relies in an inefficiency of public markets, but also data availability.

Extrapolating the identified risk factors and operational characteristics of construction firms across developed and emerging markets, company sizes and trades is not done without previous considerations. Listed construction stocks will in average tend to be bigger in size, have different managerial approaches and financial structures, different competitive positioning and internalization degree than private ones. Nonetheless, they share the same exposure to the economic cycle, political instability, financial fragility and project derived risks than their private peers. As a result, the performance of listed construction firms can be assumed to be representative of the overall industry, and the different risks presented to be shared among all players in the industry.

First, and overview of the civil engineering industry is made and its relevance for the world economy stated. Second, the special features that distinguish construction stocks from other industries are exposed in detail. Third, the main sources of uncertainty in construction projects are described. Finally, a deep dive is taken into the most relevant sources of risk for construction enterprises: economic and political risks.

Based on the above analysis several features are developed that aim to reflect the financial performance of construction enterprises and its exposure to the environment it operates in.

3.1. Industry overview

The construction and civil engineering industry is one of the largest and most important sectors of the world economy. Annually $ 10 trillion are spent worldwide in construction projects, which amounts to 13% of the world GDP, and it’s expected to grow up to $ 14 trillion by 2025. Also, construction endeavors employ 7% of the working population globally [11].

In the United States, the overall growth of the industry is expected to accelerate in 2019 over 5%, with a lot of mergers and acquisition activity in the sector driven by mega projects and the impact of new technologies. Despite the opportunities that arise in the industry from the impact of digital transformation and artificial intelligence, it faces considerable challenges like their inability to attract talent, low productivity, managing raw materials price volatility as a result of recent geopolitical events (i.e. trade war between the US and China) and its ability to keep pace with the technological developments and the new infrastructure needs it generates [12].

The civil engineering and construction sector are the backbone of the economy as they provide and update the infrastructure needed to support all business and technological advances, playing a key role in every national economy. For example, the construction sector is usually targeted by government during economic downturns to stimulate the economy and reinforce the country’s financial health [13]. Its importance in both, developed and emerging markets has increased in recent years, overweighting the role it plays in the economy [14]. In recent times, the industry has undergone several major structural changes as a result of the effects of globalization, the technological disruption in traditional business models and increased regulation.

Developments in transportation and communications, which enabled the creation of a global market, have presented construction companies in developed markets with new business opportunities and access to foreign markets. The main drivers behind these internationalization endeavors being the existence of growth opportunities unavailable in their domestic market, and a way of capitalizing their knowledge and expertise in a specific practice or technology [15]. Firms in developed countries have adopted internalization strategies in order to profit from the globalization [14]. In emerging markets, globalization has provided firms with access to technology, management expertise and financial knowledge helping them narrow the gap with their developed countries peers. As a result, nowadays construction companies face a revamped competitive environment.

The construction sector can be characterized as traditional, and often reluctant to embrace new technological developments. The industry tradition, and its inter-disciplinary and complex nature make difficult and slow the implementation of technological developments [14]. This poses an interesting opportunity for those market players willing to embrace cutting edge technology (i.e. BIM, machine learning, risk management tools) for disrupting the industry and gaining market share to the traditional players.

On the regulatory front, changes in national regulations and the appearance of supranational directives, such as the *World Trade Organization* and the *European Union Public Procurement Directives*, has led to selection improvements and increased transparency in the bidding process [14]. Also, nowadays increased attention is being paid to the *Environmental, Social and Governance* dimensions of companies' operations worldwide, leading to an increased scrutiny from the regulatory front. Construction companies are particularly exposed to environmental issues, as one of the main sources of environmental pollution, waste generation and contribution to climate change and depletion of natural capital [16]. As a result, the civil engineering industry is thought to have more social and environmental responsibility than other industries [17].

3.2. Construction stocks, one of a kind

The construction sector differs from other industries in several fundamental ways, as a result of the nature of the business. The industry is very fragmented, with large dependency on public-sector demand and very exposed to the economic cycle [11]. Companies usually operate with a fragile financial status, and its activity is inherently risky, being characterized by complex large-scale ventures that span over long periods of time. Furthermore, the risk management practice remains undeveloped relying in managers experience rather than sophisticated tools.

As mentioned before, the construction industry plays a key role in every country's economy being highly dependent on it, thus being exposed to macroeconomic fluctuations. These fluctuations influence both the supply and demand side. On the one hand, in a recessionary environment supply is reduced by the weakening of the purchasing power of consumers and tightening public budgets. On the other hand, in an economic downturn the increase in lending rates and liquidity squeeze can jeopardize companies' future [18]. As a result, the performance and sustainability of construction enterprises is heavily affected by the government financial policies and planning [13]. It's worth mentioning that the degree of internationalization of each company will influence its performance dependency to the global economy and exposure to geopolitical events.

The construction sector is a very fragmented and polarized market. On one side, we have few big national champions with international coverage developing a broad range of activities, from design and planning consulting to execution services [15] [13]. On the other side, a huge proportion of small companies that operate in small geographic areas with a high degree of specialization. The existence of small players in the industry can be, in part, explained by the extended practice of subcontracting which favors the existence of small specialized companies [14]. This disparity in company sizes, range of operations and geographical focus, makes up a large variety of company organizational and financial structures.

Construction firm's financial status, capital structure and revenue characteristics are relatively different to other industries, operating usually with high leverage and weak financial positions [19]. When it comes to listed stocks, construction companies' prices have been found to be more volatile than the general market index and their risk-adjusted performance to be in line with the broader market. New studies disagree, with Australian evidence showing the construction sector has outperformed the market in recent years [19]. While the increased volatility can be a deterrent for investors, it is partly compensated by the lower correlation to the market and its diversification possibilities, making it overall an attractive choice for portfolio managers.

Probably, the most singular characteristic of construction firms is the dependency of their financial performance on a handful of big projects. These firms usually engage in projects larger in size than their corporate assets, thus being their overall corporate performance mainly driven by the success of those projects [20]. As a matter of fact, the failure of a single project can bring the company to bankruptcy without apparent of signs financial distress at the corporate level.

Consequently, an in-depth understanding of the nature of construction projects and its operations is needed if one wishes to understand the performance drivers of construction stocks. The main characteristics of construction projects are its singularity, duration and multi-agency nature. Construction endeavors are usually unique, even in the case of design similarities, since boundary conditions are always different and hardly repeatable increasing the uncertainty surrounding them. This uncertainty is magnified by the unusual duration of construction projects, which generally span more than a year, being comparatively long when compared to other industries time horizons. The average project duration is of 1.4 years, with big projects being divided in phases with durations up to 5 or more years [21]. Moreover, there is a positive correlation between the project duration and its monetary value, thus a bigger revenue dependency comes in hand with longer durations increasing uncertainty.

Another important feature of a construction project is the involvement of many agents, such as the project owner, engineering companies, contractors and subcontractors, material suppliers and financing entities [14]. This multi-relational facet adds complexity and uncertainty to the project, since the bad management of contractual relationships and risk allocation among agents might lead to quality problems, and costs and duration overrun. Moreover, its common practice in big projects to divide it in smaller parts to diversify the counterparty risk over several contractors, which in turn increases the multi-agency derived risks.

Additionally, the construction industry is a labor-intensive sector with a predominant low qualified work force. The industry has an endemic problem to source qualified workers, and fails in the training of current employees due to the cyclical nature of the industry and an elevate churn [14]. Usually employees are hired by a temporary consortium formed by the contractors, being their employment contract linked to the project and terminated upon completion. Furthermore, the lack of management skills and capital of the smaller firms hinders their ability to source qualified human capital.

3.3. Risk in the construction industry

The nature of construction activities and processes makes it an inherently high-risk business. It is widely known that construction projects are constantly dealing with uncertainty in several fronts, making risk[4] another variable of a construction project which can substantially adversely affect the final costs, quality and duration of the project. Time and cost overruns have been found to be an endemic problem of the industry [22].

There are a wide range of risk factors which can affect the costs, duration and quality of a project, which can be classified in two broad categories: intrinsic and global risk factors [22]. Intrinsic risk factors are those inherent to the construction processes and operations, being the company responsible for its management. In contrast, global risk factors are those derived from the economic, geopolitical, sociologic and environmental events surrounding the companies' operations. The latter set of risks are out of control of the contractor and should be allocated among the different parties involved, being those handled by the party best suited to deal with them. However, in developed countries those risks are usually allocated to the contractor despite their inability to deal with them [22].

As a result, risk management plays an important role in construction operations to maximize the projects' profitability. Although, the high risk nature of the business being known by practitioners, the use of a sophisticated approach and tools to risk management (i.e. statistical decision theory models) remains scarce, relying mostly on intuition, judgement and experience [23]. As Baloi et al. [22] demand, there is a need for structuring and accumulating the knowledge and experience of individual project managers and combine it with normative models to advance the risk management practice in the industry.

3.3.1. Intrinsic Risk Factors

Intrinsic risks are those specific to the organization management and resources [22], and the most relevant ones can be grouped as follows:

- Technical (i.e. design, project complexity)

[4] Risk is an abstract concept difficult to quantify, which may have different meanings depending on the context. In construction projects risk is usually defined as 'the likelihood of a detrimental event occurring to the project' [22].

- Construction (i.e. productivity, material quality and availability, geological conditions, site conditions, site safety)
- Logistics (i.e. lack of equipment, accessibility)
- Legal and Contractual (i.e. responsibilities and risk allocation, local regulations)

The lack of managerial skills and contractual risks are identified as the ones having the biggest impact on the successful completion of the construction endeavor. On the one hand, management-related problems affect the project profitability by jeopardizing the project circumstances and client behavior, and having contractors to deal with partial information [22]. On the other hand, legal and contractual risks are those associated with the building contract which establishes the relationship between the parts and the allocation of risks. The main issues derived from it are contractual flaws or inappropriate contractual relationships and allocation of risks, which can lead to disputes, delays and costs overruns [23].

Despite these risk factors can jeopardize the company performance at a corporate level they are intrinsic to the project. In the event of the project being privately owned the details are confidential and are not disclosed to public, and even in the case of publicly owned projects were some information might be disclosed during the tender process, it’s not readily available and extracting investment signals from it wouldn’t be cost efficient. For this reason, no variables representing the project specific risks that the company faces have been included in the model.

3.3.2. Global Risk Factors

Construction enterprises operate in a dynamic environment, hence being its activity affected by its interactions with it. For this reason, their efficiency and profitability depend in great measure on their ability to understand the boundary conditions and adapt accordingly. Global risk factors are identified as those related to the environment the company operates in. The contractor has little control on them and are usually not accounted for in the cost estimates but have a huge impact in them. Global risk factors are more challenging to deal with than the others in developed countries, as a result of globalization [22].

Below, the major global risk factors affecting the construction industry are described in more detail:

- Environmental (i.e. weather)
- Social (i.e. ESG scandal exposure, availability of skilled labor)
- Economic (i.e. interest rates, price fluctuations, foreign exchange rates, inflation)
- Financial (i.e. leverage, liquidity, solvency, counterparty risk)
- Political (i.e. corruption, project desirability, nationalization, strikes, influence of power groups, labor restrictions, changes in labor costs, taxation, government relations)

Financial and Economic risks are identified as the ones having the biggest impact on the successful completion of construction projects [23]. They include the default risk of all the parties involved in the project, foreign exchange exposure and interest rates exposure, which have a huge impact in the contractors' cash flows. The exposure to these risks is a consequence of the industry's dependency to the global economy, which are usually disregarded by managers.

Another set of important external risks are those derived from the political conjuncture, which are less common but usually have a bigger impact in the company financial performance. This type of risk is more common of emerging markets than developed countries, being big international groups particularly exposed in their abroad They are by nature difficult to anticipate and manage, having been object of continuous research with several methodologies developed.

There is a subset of the risk factors, which can be categorized as *black swans*. The *black swan* concept was introduced first by Nassim N. Taleb [24], to describe extreme outlier events, or as he describes them: 'extreme rare event[s] with a huge impact and retrospective (though not prospective) predictability'. Natural disasters (i.e. floods, earthquakes, etc), economic downturns (i.e. financial crisis) and some political events (i.e. project cancelation, coup d'etat, nationalizations) fall in this category. This set of risks are usually object of high scrutiny in the contract draft and are deemed as excusable despite not being compensable [22].

3.4. Financial distress in the construction industry

As exposed above, construction projects entail, by nature, a high risk due to the high degree of uncertainty surrounding them and exposure to global risk factors, which can convert their high leverage into unbearable losses [15]. When we combine this high-risk profile with the unusual long duration of projects and their singularity, the industry sensitivity to economic cycles and a generalized lack of financial management knowledge by project managers, it leads to a higher vulnerability to financial failure than in other industries [21] [25]. As a matter of fact, the construction industry has had historically one of the highest bankruptcy rates across all industries [19].

Several studies have been performed around financial crises for construction companies, analyzing their causes and trying to predict them, since those events are critical for the project stakeholders. As Choi et al. [21] point out, construction companies usually experience financial distress around one to three years in advance of the financial crisis in terms of legal events (i.e. bankruptcy, default and delisting). A characteristic of financial distress is the absence of information before the legal events, and as a matter of fact it has been found in previous studies that the reporting of financial information is usually delayed for troubled firms [21]. This evidence will be used in the variable engineering process, and the delay in reporting earnings will considered as an explanatory variable (see section 4.3.2).

A traditional way of evaluating the performance of a company is financial ratio analysis. This methodology consists in assessing the strengths and weaknesses of a company by comparing its financial ratios[5] with the average industry value [25]. The interest for forecasting financial distress from financial ratios dates back to the 1960s, and since then different sets of financial measures have been found to be relevant depending on the type of business, definition of financial distress and research methodology employed [20].

When it comes to the civil engineering firms, the literature in financial analysis and prediction is more recent dating the first studies from the 1990s, and it still remains undeveloped [19]. The pioneer were *Kangari et al.* who proposed in 1992 the first quantitative model using financial ratios to evaluate the financial performance and

[5] Financial ratios are 'relative magnitude[s] of two selected numerical values taken from an enterprise's financial statements' and used for comparative purposes [13].

possibility of business failure for construction firms. Since then, several studies have proved the feasibility of predicting construction firms' financial distress from financial measures, despite its limitations for assessing the company's overall performance as a result of the increasing complexity and multi-disciplinary nature of their operations [14].

In his SFNN model to forecast financial distress, Chen [20] used as input variables 25 financial ratios, deemed as of common usage among bankers, when conducting financial analyses of construction companies. These ratios cover the profitability, solvency and liquidity dimensions of construction enterprises. Apostola et al. [13] developed a model to assess the financial performance of British construction companies using composite factors from nine financial ratios out of 24 considered. They identified liquidity, activity and profit margin and development as to have higher sensitivity to fluctuations of the United Kingdom economy. Horta & Camanho [14] developed a framework to assess construction companies' competitive positioning and assess their performance, using four financial ratios as KPIs to evaluate their financial health along the following dimensions: profitability productivity, financial autonomy and liquidity. Choi & Kim [21], in their model to predict financial distress, used 21 different financial ratios to evaluate the financial performance of construction contractors in Korea. The financial ratios used aimed to characterize the company across four dimensions: activity, leverage, liquidity and profitability. As a result, when it comes to construction firms the most important dimensions to assess its financial performance are: profitability, capitalization and liquidity.

There are a broad range of financial metrics that can be used to asses an enterprise financial performance. In this study, and to keep the problem tractable, only a small subset that have been proved in the literature to be relevant for construction firms are considered. This leaves us with a subset of 21 financial metrics, which can be classified according the WRDS categorization [26] into: capitalization, efficiency, financial soundness, liquidity, profitability and valuation ratios.

Capitalization, or leverage ratios, assess the debt component of a firm's capital structure; and are of upmost importance when assessing a company's financial status. Enterprises have two different ways of raising capital to finance their operations; through equity or debt. Issuing debt has some advantages over equity, since its usually cheaper, easier to access, non-dilutive and its interest payments are tax deductible [27]. In practice companies rise money using both options, and it's important to know the proportion of debt employed, also called leverage, since it increases the probability of financial failure. A company might struggle to pay its debt liabilities and its profitability could be hurt by interest rate payments [27]. The most common metrics for measuring the capitalization of a company are presented in the following table:

Table 5. Financial ratios considered for evaluating the capital structure of a company.

Financial Ratio	***Definition***	***Industry Median***
Debt-to-Equity	Total Liabilities / Shareholders Equity	1.28
Debt-to-Assets	Total Debt / Total Assets	0.56
Capitalization Ratio	Total L-T Debt / (Total L-T Debt + Common Equity + Preferred Stock)	0.38
Interest Coverage	EBIT / Interest and related expenses	3.22

The debt-to-equity ratio measures the proportion of a company's balance sheet that is financed by creditors versus the one financed by shareholders, the debt-to-assets ratio is an indicator of the degree of leverage in the balance sheet, the capitalization ratio assesses the leverage of the overall company and the interest coverage ratio evaluates the ability of the company to meet interest payments. Capitalization ratios levels vary across industries and tend to be higher in capital intensive sectors like construction. The capital requirements of civil engineering operations combined with the vulnerability of their cash flows makes them a high-risk business.

Efficiency, or activity ratios, are used to analyze the effectiveness of the company when employing its assets and liabilities. They usually quantify the operations of the company and are a good assessment of the firm's performance in the short-term [28]. Operational efficiency is closely related with profitability, thus an increase in efficiency usually translates in an increase in profitability. In the following table the efficiency ratios considered in this work are defined:

Table 6. Financial ratios considered for evaluating the efficiency of a company.

Financial Ratio	***Definition***[6]	***Industry Median***
Asset Turnover	Sales/Average Total Assets	1.21
Inventory Turnover	COGS/Average Inventories	5.41
Payables Turnover	(COGS + Δ Inv)/Average Accounts Payable	12.04
Receivables Turnover	Sales/Average Accounts Receivable	6.32

[6] These ratios are based on the average of the two most recent periods.

The asset turnover ratio measures how effective the company is in generating revenue from its assets, and the inventory turnover gives an idea of the company's ability to sell their products. Regarding the payables and receivable accounts turnover ratios, the former measures short-term liquidity in terms of the company paying its suppliers, and the latter its ability to manage collect its short-term debt.

Financial Soundness ratios compare the firm's profitability to its liabilities, measuring its ability to meet long-term financial obligations. These ratios are a relevant indicator of a company's financial health and its sustainability in the long term [29]. In the upcoming table the financial soundness metrics considered are listed:

Table 7. Financial ratios considered for evaluating the financial soundness of a company.

Financial Ratio	***Definition***	***Industry Median***
Cash Flow-to-Debt	Operating CF / Total Debt	0.05
Debt-to-EBITDA	Gross Debt/EBITDA	2.24
Interest-to-Debt	Interest / Average Total Debt	0.07
Long-term Debt-to-Total Liabilities	Long Term Debt / Total Liabilities	0.48

The cash flow-to-debt ratio assesses the ability of a firm's cash flows to cover its short and long-term obligations, and it has been found to be the best predictor for bankruptcy [20]. The debt-to-EBITDA measures the proportion of the debt to EBITDA, being the latter the best proxy of cash; and the long-term debt-to-Total Liabilities ratio the proportion of long-term debt. The interest-to-debt ratio is used as a comparative measure to evaluate a company's cost to finance through debt.

Liquidity ratios measure the company's ability to meet short-term financial obligations without resorting to external capital. They evaluate the ability of the company to convert its assets into cash in a cost-effective manner to cover short-term obligations [30]. They differ from the financial soundness ratios exposed above in the sense that they focus in the ability of the company to stay afloat by meeting its immediate liabilities, rather than in the company's overall ability to pay all its debt. The most used financial metrics when it comes to evaluating liquidity are defining in the table below:

Table 8. Financial ratios considered for evaluating the liquidity of a company.

Financial Ratio	***Definition***	***Industry Median***
Current Ratio	Current Assets / Current Liabilities	1.65
Quick Ratio	(Cash & Cash Equivalents + Marketable Securities + Accounts Receivable) / Current liabilities	1.44

The main difference between the current and quick ratio it's the exclusion of inventories in the latter's numerator, as it focuses only in the most liquid assets of a company. The quick ratio is a more extreme measure in the sense that only considers the assets that are most easily converted in cash in the case of a severe economic crunch.

Profitability ratios aim to evaluate the ability of the company to generate profit [26]. This dimension is indicated in the literature as the most important criteria when it comes to construction stocks, since it is closely related with earnings quality and positively correlated with dividend distributions [19]. Furthermore, profit growth is well regarded by investors as it reflects the company's ability to increase revenues while controlling costs. Next, the metrics selected to measure the profitability dimension are defined:

Table 9. Financial ratios considered for evaluating the profitability of a company.

Financial Ratio	***Definition***	***Industry Median***
Net Profit Margin	Net Income / Sales	0.03
Return on Assets	Operating Income before Depreciation/Total Assets	0.09
Return on Capital Employed	EBIT / (L-T and Current Liabilities + Common Equity)	0.09
Return on Equity	Net Income/Book Equity[7]	0.09

The net profit margin reflects the final profit value, the return on assets measures how efficient is a firm in generating profit from their assets, the return on capital employed measures the ability of a company to make profit from the capital employed and the return on equity the true return to investors in the company's equity. On the one hand, Balatbat et. al [19] found evidence that construction firms in the Australian market performance had been stagnated, with a net profit margin around staying at 3% versus the strong growth experienced by the main Australian groups. On the other hand, well established

[7] Book Equity is defined as the sum of Total Parent Stockholder's Equity, Deferred Taxes and Investment Tax Credit [26].

firms had strong and consistent profitable results, with ROE in line with the main groups; and as effective in generating profits with ROAs of 6-8%.

Valuation Ratios are used to get an understanding in the company's share price and determine its investment potential. These ratios are fundamental for investors as they give them insights in whether a company is under or over-valued with respect of its fundamentals. In the following table the valuation ratio considered are presented:

Table 10. Financial ratios for evaluating the profitability of a company.

Financial Ratio	***Definition***	***Industry Median***
Enterprise Value	EV/EBITDA	7.93
Price-to-Earnings	Share Price/EPS (excluding extraordinary items)	8.57
Price-to-Book	Share Price/Book Value of Equity[8]	1.43

The enterprise value multiple is one of the financial metrics most used by bankers when valuing a company. It's a convenient way of comparing firms since it ignores taxes, interest and non-cash flow items [19]. The price-to-earnings ratio (P/E) is also a widely used indicator in the financial community to evaluate if a company is under or over-valued relative to its earnings as a result of market dynamics. In the literature, construction firms have been found to not be over-valued with respect to the market [19]. Here the P/E ratio has been computed without considering extraordinary items[9]. Last, the price-to-book ratio is another valuation metric with respect to the net asset value of the firm, and gives an estimate of the its value in the case of liquidation [31].

[8] The Book Value of Equity is defined as the 'net asset value of a company' [67].

[9] Extraordinary items are defined as 'gains or losses from unusual events that are separately classified, presented and disclosed in the financial reports' [69].

3.5. Political Risk in the construction industry

One of the major sources of uncertainty in construction projects is the political environment in which the company operates, being international groups more sensitive to this risk category. For example, we can have extreme events like the expropriation by the government of Argentina of YPF, Repsol's subsidiary in the country, in 2012 [32], or less extreme ones like the cancelation of Ferrovial's flagship project in the Denver International Airport by the local government in 2019 [33]. Both events were driven by national and local changes in the political environment, with a huge impact in the companies' financial performance.

Given the abstractness and subjectivity of political risks, there are several definitions available in the academic literature. A widely accepted one defines them as 'the risk or probability of occurrence of some political event(s) that will change the prospects for the profitability of a given investment' [15]. Political risks can be classified in two broad categories: macro-risks when the political events in question affect the company's operations in a general way, and micro-risks when the event in question impacts a particular firm [15]. As an illustration, a change in labor costs will belong to the former, and a nationalization to the latter.

Multinational construction firms are particularly sensitive to micro-risks that can hinder the project's expected return, like currency exchange restrictions or policies requiring local joint ventures. Early project termination as a result of these risks is unlikely since it might be troublesome for the project owner, being the cash flows the most vulnerable. The project cash flows are extremely exposed to the political risks, and even small changes can convert the typical leverage incurred by the company in loses or even financial failure.

It's commonly assumed the existence of a positive correlation between the probability of facing political risks and the political instability in the country of operation. The stability of a country's political system depends on several social (i.e. religious and racial conflicts), economic (i.e. local businesses interest), and political (i.e. forthcoming elections) factors, which can usually be quantified and wrapped into a political stability index [15]. A drawback of these kind of indexes is their generality, since political instability does not always mean added risk for the project, and construction ventures are not equally sensitive to all political events.

An example of these indices are the ones elaborated by the World Bank, known as the *World Governance Indicators*. These indices are composite indicators based on over 30 data sources[10], that aim to monitor and evaluate the government of a country and its policies across six different dimensions: Voice and Accountability, Political Stability and Absence of Violence/Terrorism, Government Effectiveness, Regulatory quality, Rule of Law and Control of Corruption [39].

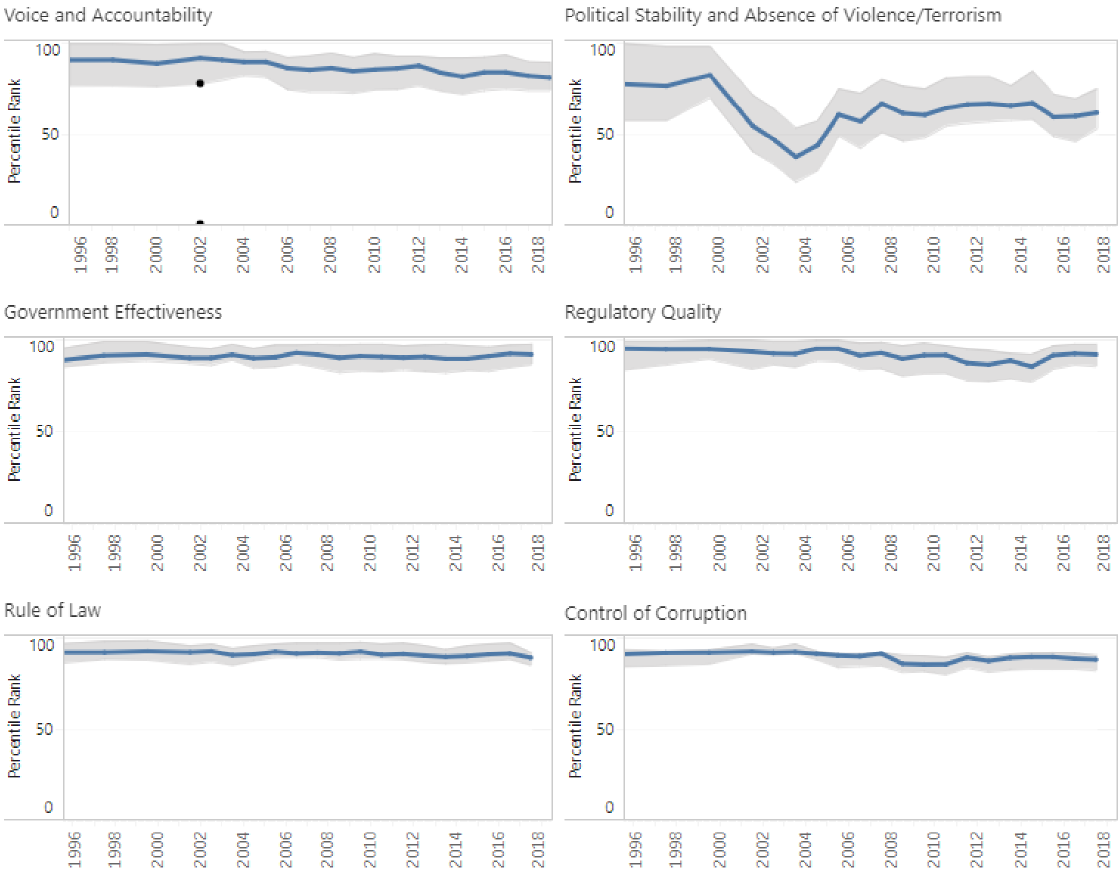


Figure 7. World Governance Indicators as reported by the World Bank for the United States from 1996 to 2018 (Source: World Bank [34]*)*

In figure 7, the values of the indices can be observed for the United States. Since the U.S. constitutes one of the most advanced countries in the world in terms of governance and political stability, the indices score really high with seldom variability. The exception to

[10] The data sources include surveys, think tanks, non-governmental organizations, international institutions and private firms [34].

the norm, is the *Political stability and Absence of Violence/Terrorism* which can be deemed as more volatile, with the percentile rank ranging from the 80 to the 40 levels in the sample. This is the result of the 9/11 terrorist attacks to the World Trade Center and the Pentagon on the 11th of September of 2001, leading to a decade troubled by the terrorist threat and the Iraq War between 2003 and 2011.

4. Research Methodology

In this section the research methodology followed is presented in detail. First, the data sources used are described, including the different hurdles encountered during the preprocessing. Second, a systematic investment strategy is developed from the investment opportunity presented by the market inefficiencies described in section 2. A detailed description of its implementation is given. Third, the model setting is presented including descriptions of the target variable and the different features extracted from the main performance drivers identified in the previous section. Fourth, the variable selection procedure and the methodologies used are introduced. To conclude, model selection approach and the learning techniques considered are outlined.

The investment signals from the model are used to build the systematic investment strategy, building long-short portfolios based on the direction of the surprise.

4.1. Data sources

Here, the different data sources used in this work are presented. These databases are widely used by the financial industry and academic community and provide information regarding market data, investors' expectations, macroeconomic and geopolitical measures of the factors affecting civil engineering firms' operations, and financial performance indicators for construction firms.

In this study, and as a result of the investment framework selected, we are going to focus in the *Earnings Per Share* (EPS) metric for quarterly earnings and U.S. listed construction firms. The geographical restriction to U.S. listed companies is due to data availability reasons. Detailed market and investors' expectations data is more abundant and readily available in the U.S., but not that accessible for other geographic areas. On top of that, complexity arises when combining datasets from different regions for a long time-window given the different data collection procedures and regulations. Even with the selected subset the author encountered difficulties combining datasets from different providers.

This research is focalized in a very special and characteristic type of company inside the civil engineering industry: construction enterprises. The task of classifying companies according to their trade is not novel, with several classifications available (i.e. SIC, NAICS, Fama-French). Indeed, industry classification are not static, and are constantly evolving or being substituted by new ones to adapt to an ever-changing world.

The *Standard Industrial Classification* (SIC) is an industry classification system used by governmental agencies established in the U.S. in 1937; and adopted also in other countries. This classification system was substituted for the *North American Industry Classification System* (NAICS) by the U.S. government departments in 1997 but was still used by the SEC until at least 2014 [35]. Here, the SIC classification system has been used, instead of the newer NAICS, given that it was the one valid during the most part of the data sample. It is organized in a nested structure in Divisions, Major Groups and Industry Groups. As an illustration the levels for the construction industry are shown in figure 11.

Table 11. Standard Industrial Classification Division and Major Groups for construction companies.

Division	*Major Group*	*Industry Group*
C Construction	15 Building Construction, General Contractors and Operative Builders	General Building Contractors-nonresidential
		General Building Contractors-residential
		Operative Builders
	16 Heavy Construction other than Building construction contractors	Carpentry and Floor Work
		Electrical Work
		Masonry, Stonework, Tile Setting, And Plastering
		Miscellaneous Special Trade Contractors
		Plumbing, Heating and Air-conditioning
	17 Construction special trade contractors	Heavy Construction, Except Highway and Street
		Highway and Street Construction

From the CRSP and IBES datasets universe of companies, the data sample used has been built as the construction firms in the intersection of both datasets for the last 20 years, covering a time interval ranging from 01/01/1999 to 31/12/2018. Once created different features are added from other data sources, expanding the individual companies' available information, as well as macroeconomic indicators surrounding them.

4.1.1. CRSP Database

The *Center for Research in Security Prices* (CRSP) stock database provides market data for individual US securities traded in the New York Stock Exchange (NYSE), the NYSE American (NYSE MKT) and the NASDAQ. In this research the historical monthly stock data has been used including stock information, price, shares outstanding, and returns for the stock and the S&P 500. Only common class shares[11] have been considered, and the S&P 500 returns used as a proxy of the market return. In order to combine the CRSP database with the IBES dataset a linking table to match both database unique identifiers has been created as instructed by WRDS [36].

4.1.2. I/B/E/S Database

The *Institutional Broker's Estimate System*, commonly known in the financial arena as I/B/E/S dataset, is a historical earnings estimates database containing analysts' forecasts for 23 financial metrics covering 60,000 companies across over a 100 markets for several time horizons. It includes estimates from 3,000 of the largest global and regional brokers, comprising over 30,000 individual analysts [37].

The database is composed by several files corresponding to different levels of data aggregation. For our purposes, the highest level of aggregation is used: the IBES Summary Statistics file. This file contains data at a company level, with descriptive statistics of the analysts' consensus[12]. The dataset has monthly frequency, being reported every Thursday before the third Friday of each month, following the Thompson Reuters production cycle.

As Bouchard et al. [4] point out, analysts included in the IBES database are professional forecasters and their estimates can be assumed to be representative of investor's earnings expectations. It's worth mentioning, that inclusion in the database is not mandatory, but voluntary, thus it does not include all the estimates available at a current moment in time[13]. Also, for a company to be included in the database there must have been at least

[11] Identified by share codes 10 and 11.

[12] The author has access to the academic version of IBES offered by Thompson Reuters. Due to restrictions of certain brokers, their estimates are not available in the academic version (only in the Institutional) but included in the consensus of the summary statistics file [37].

[13] The Summary statistics dataset includes an approximation of the number of analysts covering a certain stock, excluding estimates marked as outliers and stopped estimates due to inactivity [66].

one analyst covering it. For companies that no longer exist or analysts which stop providing estimates the forecasts are kept in the database to avoid survivorship bias [37].

4.1.3. WRDS Financial Ratios Suite

The *Wharton Research Data Service* (WRDS) research platform has a devoted suite that gives access to the most widely used financial ratios in academia. It is based in the CRSP common stock universe and includes different options for industry aggregation. Since the SIC classification is not available, the *French-Fama 49 Industries* classification has been used without it being a major drawback. The data for computing the ratios is sourced from the following databases: market data from CRSP, accounting data from Compustat[14] and earnings data from IBES [26]. Also, all the data has already been lagged to ensure that all data was publicly available for each time stamp.

The WRDS Financial Ratio suit contains 74 different financial ratios which are classified according to their financial meaning in the following groups: capitalization, efficiency, financial soundness, solvency, liquidity, profitability, valuation and others [26].

4.1.4. Federal Reserve Economic data

The *Federal Reserve Economic Data* (FRED) research data service, is one of the many research platforms and tools developed by the Federal Reserve Bank of St. Louis in the U.S. It offers access to more than 500,000 data series from over 87 public and private data providers, including governmental agencies like the *U.S Census* or *Bureau of Labor Statistics* [38] [39]. The measures included in the database encompass a broad range of geopolitical, economic and financial indicators; including among them producer price indexes, employment levels, demographic data, interest rates and monetary data.

4.1.5. World Bank Data

The *World Bank* (WB) is a global financial institution with the participation of 189 countries around the world, based in Washington D.C. (United States). It was funded in 1944, and is now comprised by five different institutions, with the goal of reducing extreme poverty around the world. The WB offering to developing countries is not limited

[14] Compustat is a database of financial and market data, with global coverage, provided by S&P Global Market Intelligence division. It covers more than 99,000 firms since 1964, and its use is widespread in the Finance industry [68].

to financing, but does also include policy advisory to governments and technical assistance in the implementation of projects [40].

The WB *Development Data Group* maintains and develops several financial, macroeconomic and sectorial databases in an open-data initiative. The data is sourced from the member countries national statistical institutes, depending the quality on the sophistication of each country data collection procedures [41].

Among the different indicators and statistical measures developed by the WB to promote effective policies and assess the impact of their projects are the *World Governance Indicators*. These indices are composite indicators based on over 30 data sources[15], that aim to monitor and evaluate the government of a country and its policies across six different dimensions: Voice and Accountability, Political Stability and Absence of Violence/Terrorism, Government Effectiveness, Regulatory quality, Rule of Law and Control of Corruption [34].

4.2. The investment framework

As outlined before, in this research application we are trying to build a systematic investment strategy that will profit of earnings surprises by forecasting them in advance of the earnings announcement and building portfolios that benefit from the pre- and post-announcement drift. The strategy will consist in making a forecast of the possible earnings surprise at the end of each fiscal period, building long-short portfolios based on the direction of the surprise, and holding them for two months after the earnings announcement. The portfolio will be long on firms with an expected positive surprise and short on those with an expected negative one. The following timeline shows all the relevant events regarding the strategy:

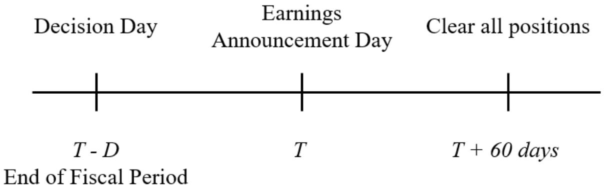


Figure 8. Timeline of the investment strategy.

[15] The data sources include surveys, think tanks, non-governmental organizations, international institutions and private firms [34].

As Dhar & Chou [9] point out, the choice of how much time in advance portfolios are formed with respect of the announcement day is a critical decision for the investment manager, since it represents a tradeoff between forecasting accuracy and profits. One the one hand, when forming the portfolio too early we make sure to capture the whole pre-announcement drift, but at the expense of a higher forecast error since there is more room for either analysts revisions that eliminate the surprise or and information shock that makes the forecast obsolete. On the other hand, the closer the decision day is to the earnings release the higher accuracy we can expect but we leave some money on the table. In this research, and to simplify the investment strategy, the decision day is fixed at the end of the fiscal period and portfolios are held for three months[16]. It's important to note that there is no overlap between portfolios given that its three-month frequency matches the quarterly earnings one[17], by the time new positions are formed previous ones have already been cleared.

Once the portfolios are formed their abnormal returns are computed as defined by Dhar & Chou [9]. The abnormal returns (AR_{it}) for stock *i* at time *t* are defined as the 'difference between the estimated normal returns and the actual returns' [9], and expressed[18] as follows:

$$AR_{it} = R_{it} - (\alpha_i + \beta_i \cdot R_{mt}) \tag{1}$$

And the cumulative abnormal returns (CAR_{it}) from month T_a to T_b as:

$$CAR_{it}(T_a, T_b) = \left(1 + AR_{i,T_a}\right) \cdot \left(1 + AR_{i,T_{a+1}}\right) \cdot \ldots \cdot \left(1 + AR_{i,T_b}\right) - 1 \tag{2}$$

[16] Firms in the sample, in average post earnings 34 days after the end of the fiscal period, with a standard deviation of 13 days.

[17] Only firms with fiscal periods ending in March, June, September and December are considered for the investment strategy. This restriction is not applied for the training and test of models.

[18] With the market model for return of stock *i* in month *t* being $R_{it} = \alpha_i + \beta_i \cdot R_{mt} + \epsilon_{it}$

4.3. Variable Definition

Now that we have defined the systematic investment strategy designed to exploit the previously identified market inefficiency, the goal is to find a model that allows us to predict the surprises in a reliable manner. To do so we are going to define the target variable as proxy of an earnings' surprise and leverage the knowledge on the operations of construction companies to engineer a set of explanatory variables with predictive power.

4.3.1. Target variables

The goal is to be able to predict surprises, and as exposed before those are defined as a substantial deviation of the reported earnings from market expectations. A way to measure it is by the normalized forecast error of analysts' consensus. There are various definitions of the forecast error in the literature, here we defined it as proposed by Dhar & Chou [9]. The Forecast Error for company *i* in quarter *q* is expressed as follows:

$$FE_{iq} = \frac{e_{iq} - E(e_{iq})}{|e_{iq}| + |E(e_{iq})|} \qquad (3)$$

where e_{iq} are the actual quarterly earnings for firm *i* at quarter q, and $E(e_{iq})$ the expected earnings defined as the last analysts' consensus before the EAD. The consensus is defined by Thompson Reuters as 'the average of all (subjects to I/B/E/S exclusion rules) estimates, from all analysts, for a given issue and time period' [42]. This 'normalized' definition of the forecast error ranges from [-1,1] and emphasizes the sign of the surprise over its magnitude. As noted by Dhar & Chou [9] it has the disadvantage of taking extreme values when the actual and expected earnings have different sign without accounting for the degree.

The Forecast Error (FE_{iq}) has been categorized depending on its Z-score, being labeled as a negative surprise (label 1) if it was lower than -0,5, no surprise (label 2) if it belonged to [-0.5,0.5], and positive surprise (label 3) if it was greater than 0.5. Therefore, we are in a multi-class classification problem with three levels: negative surprise, no surprise and positive surprise. It is important to point out that our goal is to accurately predict earnings surprises, positive or negative, since we are going to build our portfolios based on the surprise category. This means that we are particularly sensitive to misclassifying a positive surprise as negative and vice versa, while being a lesser problem misclassifying a surprise as no surprise.

4.3.2. Features

The features included in the model can be classified in three broad groups depending on their origin: market and expectations data, global risk factors and performance related variables. The first group of variables are derived from market data and analysts' expectations, the second represents all the risk factors related to the context in which a company operates and that can seriously affect its performance, and the last one all those variables relative to the company financial performance. Features are introduced as the year-on-year change before the end of the fiscal period (or decision day), except for those were indicated differently. This is done to capture trends in the data, taking special care in not incorporating future information. Different imputation approaches have been applied depending on the type of missing data and the variable nature.

The first group of variables are selected based on the results from a previous work from the author [10], as exposed in the introduction. In fact, these features and its relationship with the earnings' surprises phenomena is well documented in the literature [9]. The standard deviation of the analysts' consensus on the months prior to the earnings' release, being relatively important in the precedent model, hasn't been included in this model due to the amount of missing data for the current sample. Next, the variables and their definition are exposed in detail:

(a) *Firm size (size_pct)*: The firm size is computed monthly from the price on the last trading date of the month and the number of shares outstanding coming from the CRSP database, and it's expressed as the percentile rank. This variable has been found to be negatively correlated with the magnitude of earnings surprises, which can be explained in terms of inattention. Small firms usually have less analyst coverage than their bigger peers, leading to larger divergences with the consensus. Also, the firm size has been found to be correlated with a firms' financial performance [9]. A widely accepted explanation for this phenomenon is that the bigger liquidity of large firms has value for investment managers, since it reduces their costs. Moreover, given the disparity of sizes of construction companies it's also a discriminator for different competitive positioning across companies [14]. Indeed, it's standard practice for governments to categorize the construction companies by their size and restrict the projects they can undertake based on it.

(b) *Previous quarter's earnings surprise (FEqX):* This variable is simply the earnings surprise from the previous four quarters. As presented in the introduction, underreaction to past earnings is a well-documented market anomaly and there exists an autocorrelation of earnings surprises across adjacent quarters. To avoid missing data, as a result of the variable definition, the first four quarters for each stock are discarded[19].

(c) Analysts forecast differences (AFDX): This variable is defined as the difference between the number of analysts that reviewed their estimate upwards and those who did it downwards with respect to the previous consensus. Values for the consensus one, two and three months before the end of the fiscal period are included in the data. It aims to capture possible information shocks that made some analysts adjust their estimates and a possible lag of the consensus, as most analysts may fail to adjust in time, leading to a surprise.

While assessing the impact of external factors in construction companies' performance (section 3.3.2) the importance of its ability to dynamically adapt to their environment was stated. Among the most important global risk factors are those of economic nature and those associated with political risk. The following variables have been engineered to reflect those risks, thus as predictors of the company's financial performance.

(d) Gross National Income (GNI): One of the most relevant characteristics of the construction industry is its close relation with the national economy, and as a result of globalization tis exposure to the world economy. The *Gross National Income*[20] has been introduced to account for the economic fluctuations in the U.S. economy. As discussed by Kim et al. [18], the GNI has been used instead of the *Gross Domestic Product* to consider the impact that the global economic context has in the national economy.

(e) *Interest Rates (tbill1Y):* In this work we are only considering U.S. based stocks, thus the 1-year Treasury Bill has been used as an interest rate indicator. Interest

[19] Not all firms are always present through the 20 years considered.

[20] The Gross National Income (GNI) is the 'total domestic and foreign output claimed by residents of a country' [71], and can be computed from the GDP by adding the actual trading loss [18].

rates play a key role in the construction business as they impact the financing of projects and the purchasing power of project owners [18].

(f) *Construction Material and Machinery Producer Price Indices (ppi_mat/ppi_mach):* The *Price Producer Indices* for Construction Materials and Construction Machinery and Equipment are introduced as economic indicators of the changes in the costs in the industry. The PPIs measure the average change in sale prices over time for the domestic market [43].

(g) *Public and Private Construction Spending (cons_spe_pub/cons_spe_pri):* The spending in the construction activities, being public or private, has a big impact in the industry's performance, being also closely related with economic fluctuations. These indicators are constructed from the *Value of Construction Put in Place Survey* (VIP) carried out by the U.S. Census Bureau and estimate the total value[21] of the construction work performed on existing and new structures in the U.S. [44]. The data is divided into public and private sector, with monthly frequency, and it has been adjusted for seasonality.

(h) *Construction Unemployment Rate (unemp_excess_ma):* As a labor-intensive industry, the financial performance of the company will be exposed to fluctuations in the labor market. The construction unemployment as excess of the national unemployment rate is included as an indicator of the impact in the industry of labor fluctuations. The time series is built as the difference between the *U.S. Unemployment Rate* and the *Unemployed Rate: Construction Industry, Private Wage and Salary Workers* indices from the U.S. Bureau of Statistics. The time series are not adjusted by seasonality, and by computing the difference we take out common trends and seasonality between them. Also, the construction unemployment rate data is not available for the whole sample timespan, and the missing values are imputed using the mean of the unemployment excess variable.

Trying to find an indicator of political risk to monitor governance changes over short periods of time or relating to a single country is a complex task. Here, the *World Governance Indicators* as defined by the World Bank are considered, aiming to capture geopolitical events that may have implications for the construction activity as outlined in section 3.5. According to the World Bank's *Development Data Group* [34] these indices

[21] The value estimates include labor, materials, design and engineering costs; as well as overheads, profit margins, taxes and interest expenses [44].

are valuable to assess cross-country comparisons and monitor trends over long-periods of time (ten years), but meaningless for shorter time horizons and country-specific analysis. Moreover, they advise to be careful in identifying small changes over short time periods as statistically significant since the margin of error is relatively big. Also, they point out that variability in the data can be explained by changes in the source data, the addition of new data sources or changes in the aggregation weights [34].

From the six indicators that constitute the WGI, only the *Political stability and Absence of Violence/Terrorism* exhibits significant variability during the time period we are considering. These indicators are shown for the U.S. and relevant time sample in figure 7. Hence, it will be the only one included in the model aiming to capture any changes in the political landscape:

(i) *Political stability and Absence of Violence/Terrorism Index (pr_pv):* This index measures the 'perceived likelihood of political instability and/or political motivated violence, including terrorism' [45]. It's introduced aiming to control for the effects of political instability in the assignation of public construction works. Also, to reflect events such as the 9/11 which translated to increased government spending.

As mentioned before, during the assessment of the intrinsic risks faced by construction companies (section 3.3.1), most of these set of risks are project specific and the data is not published or not accessible in a cost-efficient way. Conversely, public companies have periodic information disclosure obligations with the SEC which is available to investors and shareholders. The periodicity, reliability and availability of these information makes it suitable to serve as inputs for a systematic investment strategy.

In section 3.4, the main drivers of financial distress for construction companies are presented, and 21 financial ratios that proved to be relevant to assess construction companies' financial performance were selected from the literature. These metrics allow us to evaluate the following dimensions: capitalization, efficiency, financial soundness, liquidity, profitability and valuation.

(j) Financial Ratios: The financial ratios defined in tables 4-9 are introduced as performance indicators. As pointed out by *Kangari et al.* [25], financial ratios are

meaningless when presented alone, and need to be benchmarked by industry levels. In this research work the median is used instead of the mean as industry-level as noted in the literature to be the best practice to avoid aggregation issues with negative denominator ratios [26]. The different financial ratios are smoothened by taking the 3-month moving average, and standardized by subtracting the industry median and dividing by the interquartile range. Missing data is handled by imputing the industry median.

(k) *Delay in reporting earnings (ndelay):* An interesting phenomenon identified in the literature is the delay in reporting earnings by failing companies [21]. The standardized delay in reporting for each company is introduced as a proxy of financial distress. The information regarding the day when the company will be reporting its quarterly earnings its usually not known by the last day of each fiscal period and might be updated by the company in the following weeks. Since the decision day in our strategy is fixed at the end of each fiscal period, only the information for the previous quarters will be available.

During the preprocessing of the different data sources and variable engineering process several measures have been taken to prevent missing data and lose the least amount of information possible. For the variables where moving averages or percentual changes where used, longer time series than the sample duration where employed. Mismatches between different data sources where considered case by case to avoid losing firms in the sample. The vast majority of missing values where in the financial ratios data, and where imputed by the industry median to avoid losing observations. For the remaining missing values, being its quantity not significant, mean imputation was used.

4.4. Exploratory Analysis and Variable Selection

In this research a classical approach to variable selection has been taken, based in empirical evidence from the literature and statistical considerations. First, an extensive literature research in earnings' surprises prediction, construction firms' financial performance assessment and construction business related risks has been done, yielding to an initial set of variables. Second, some statistical procedures are used to evaluate the relationship between each feature and the target variable. The first part was described, and its results presented, in the first two sections of this work. The second part of the process is described in this section, and the results included in the next one.

Balcaen & Ooghe [46], prevent us of the limitations of this methodology, and the implications in its implementation. They point out that popular features in the literature with demonstrated predictive ability in previous research might be unreliable as a result of *window dressing*[22] by managers. Also, they warn that this approach for variable selection, and the model derived, might be sample specific; leading to a poor generalization ability of the results. Moreover, they outline that simpler models, with a reduced amount of features, tend to perform better than more complex approaches when it comes to classification accuracy; since most of the predictive power of additional features is already accounted for by existing ones through their correlations.

In this section, different techniques are used for feature selection: filter methods and a wrapper method. On the one hand, filter methods are used to select variables based on their correlation or dependency with the target variable, without depending on a predictive model assumptions. Hence, they are more robust to overfitting, but have an inclination to select redundant features since they do not consider the relationship among them. On the other hand, wrapper methods take into account the relationship between features by considering different feature subsets and optimizing them for a given learning technique, which makes the resulting selection model-specific and prone to overfitting [47].

First, a univariate analysis is done to assess each feature correlation and dependency with the dependent variable. Three different tests are run: One-Way ANOVA Kruskal-Wallis test for numerical variables and a Pearson Chi-squared test for the categorical ones, and the Mutual Information between variables as a measure to assess dependency. Second, a Recursive Feature Elimination (RFE) algorithm is implemented with a linear and a *tree ensemble* model. The results the above-mentioned techniques are analyzed and a final set of features is selected.

4.4.1. Univariate Analysis

In this research we are dealing with a categorical target variable with three levels: negative surprise, no surprise and positive surprise. A common way of testing for variable significance will be to run a One-Way ANOVA F-test to compare the distributions of each feature under the different target variable groups. In short, the ANOVA procedure

[22] The practice by some managers of making decisions that affect commonly monitored KPIs to improve the appearance of their performance.

relies on a statistical test to compare two or more population means to analyze group differences in a sample [48]. In the One-Way ANOVA a commonly used statistical test is the F-test, which relies in four basic assumptions: the dependent variable to be measured in an equal interval scale, the independence of samples, the residuals to be normally distributed and homoscedasticity (homogeneity of variances among groups) [49].

As explained by Lowry [49], the F-test is robust with respect of the last two assumptions listed above, normality and homoscedasticity, as long as the sample groups are of the same size. In our case, as a result of the nature of the target variable we are dealing with an imbalanced dataset, since earnings' surprises are the exception rather than the norm, being no surprise the majority class. This makes the one-way analysis of variance with the F-test unsuitable for our purpose.

As a result we are going to use as an alternative the *Kruskal-Wallis H-test*, a non-parametric procedure for testing if sample groups come from the same distribution [50]. This test is an extension of the *Mann-Whitney-Wilcoxon* test for more than two groups and does not rely in the normality assumption for the residuals. The null hypothesis is that all groups are originated from the same distribution. The data is ranked ignoring the groupings, and the sample statistic *H* is defined as [50]:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i \left(\sum_{j=1}^{n_i} r_{ij} - \frac{1}{2}(N+1)\right)^2}{\sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(r_{ij} - \frac{1}{2}(N+1)\right)^2} \tag{4}$$

where n_i is the number of observations in group *i*, r_{ij} the rank of observation *j* from group *i*, and *N* de total number of observations. It is important to note that this test, when significant, it just signals that the data does not come from the same distribution; but it does not indicate which group does not belong [50]. Also, the *H* statistic is assumed to be chi-square distributed with g-1 degrees of freedom for groups with more than 5 observations for computational ease, being the group sizes an important consideration when selecting this test.

The majority of variables extracted from the literature research are numerical, except one: the past earnings' surprises for the last four quarters. In order to assess the relationship between the past earnings' surprises and the target variable, a *Pearson Chi-squared test* for categorical data is used. The test is used to assess if there is a statistically significant difference between the expected and observed frequencies of two categorical variables in a contingency table[23] [51]. In our case the null hypothesis will be that the current earnings' surprise is independent of the previous surprise, and the test statistic defined as follows [52]:

$$\chi^2 = \sum_{i \, \epsilon \, \{1,2,3\}} \sum_{j \, \epsilon \, \{1,2,3\}} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \tag{5}$$

where *i* and *j* are the categories of the current and past surprises, *O* and *E* are the observed and expected frequencies.

A non-parametric test for assessing the dependence of two random variables is to estimate the *Mutual Information* (MI) between them. Mutual information is a more general concept than correlation, not being limited to linear dependence. It is a dimensionless quantity measured in bits, which is defined for two discrete variables *X* and *Y* as [53]:

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)} \tag{6}$$

where $P_{XY}(x,y)$ it's their joint probability distribution and $P_X(x)$ and $P_Y(y)$ their marginal distributions. Mutual Information assesses the 'reduction of uncertainty regarding variable *X* after observing *Y'* [53], and can be interpreted as the amount of information a variable adds to making the correct classification decision [52]. The MI estimate is a non-negative value, being 0 when the variables are independent and higher values implying a higher dependency. For more information on the subject the reader is encouraged to check Latham & Roudi [53].

[23] Matrix exhibiting the frequency distribution of variables.

4.4.2. Recursive Feature Elimination

The *Recursive Feature Elimination* (RFE) methodology is also a popular procedure for variable selection, with the advantage over the univariate tests of accounting for the interaction between variables and helping identify the redundant ones. This procedure, for a given estimator, starts with all the features and recursively eliminates the least important ones leading to a smaller set of features in each iteration, until reaching the number of features stipulated [54]. The variable importance is determined through the coefficients in a linear model and the feature importance attribute for tree-based models. This methodology is combined with 5-fold cross validation, with temporal and grouping considerations (see section 4.5), to determine the optimal set of features.

One of the most important parameters for the RFE methodology is the selection of the evaluation metric, as it will heavily influence the selected subset. Here, the F-score measure for a 0.5 beta and macro averaged across classes is selected, given the imbalance in the data and the asymmetric misclassification costs (see section 4.5.2 a more detailed discussion on the selected performance metrics). I am going to consider two different models: a logistic regression and a random forest. In both cases the features are scaled, and the class weights are modified to account for class imbalance (see section 4.5.1).

4.5. Model Selection and Assessment

Once we have selected the subset of features that we are going to use to predict the forecast error, the next stage is selecting the appropriate model and assessing its performance. This procedure involves two different steps: the model selection and the final model performance evaluation. First, different models are proposed and the one with the best estimated performance is chosen; then the final model's performance is evaluated by estimating its prediction error. Selecting the correct metrics for evaluating the *generalization*[24] performance of an algorithm is of upmost importance as it influences both steps, the model selection and final model assessment [55].

The *test error*, or *generalization error*, is the prediction error in an independent test sample for a specific training set, and the *expected test error* is its expected value, which is not conditional on the training set. As pointed out by Hastie et al. [55], our goal when

[24] The *generalization* performance of a model is related to its predictive ability in unseen data [55].

evaluating a model is estimating the former, but most methods estimate the latter since it does not seem possible to estimate the conditional error with only the information on the training set. The *training error* is defined as the average loss over the training sample and it's not a good estimate for the test error, since it's decreasing with model complexity [55].

Using the same data for evaluating different models and assessing the final model generalization capabilities will be a mistake, since the model will overfit the data and our estimate will be underestimating the real test error. As a result, the best approach for performing the above-mentioned tasks is to divide the data into training, validation and test sets. Nevertheless, this approach requires abundance of data and the selected model might be dependent on the choice of train and validation sets [54].

In our case, and as a consequence of having restricted the sample to listed construction firms in the United States, we are dealing with a limited amount of data. For that reason, we are going to split the dataset into a train and tests sets with an 80/20 split and use *Cross Validation* for feature and model selection purposes, considering the temporal dimension of the data. As Hastie et al. [55] point out, there is no general rule for the train-test split, depending it on the signal-to-noise ratio of the data and the size of the training set.

Cross validation (CV) is a widely used methodology for replacing the validation step and directly estimating the expected test error. In a K-fold CV the train dataset is divided into *K* parts, with *K-1* parts being used to train the model and the *k*-th part as test data to compute the prediction error. This is done recursively over all folds and the prediction error is calculated as the average prediction error across them. This procedure is illustrated in figure 9.

It's important to note that, the split into train and test sets was done after the data preprocessing, and the test sample has been left out since then. Moreover, the cross-validation procedure has been applied to the whole pipeline, including the feature selection process. It's important to extend the cross-validation to the feature selection procedure, since when selecting features with the whole training data the latter model selection does not correctly simulate its application to an independent set [55]. This has been done for the RFE procedure but not for the univariate tests.

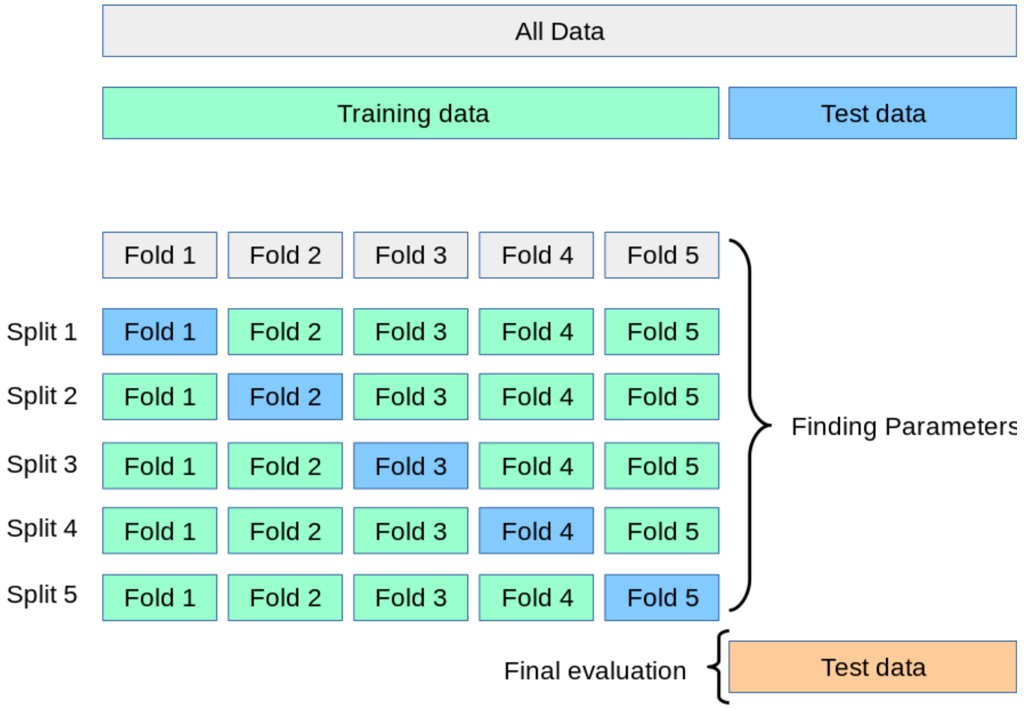


Figure 9. Diagram of a 5-fold Cross Validation model selection strategy (Source: Pedragosa et al. [54]*)*

The selection of the number of folds (K) in cross validation involves a Variance-Bias tradeoff. On the one hand, we can select as many folds as observations available, a procedure also known as *Leave-One-Out cross validation*, having the resulting estimator a low bias and large variance. On the other hand, using 5 or 10 folds will lead to an estimator with lower variance that can be overestimating the true prediction error depending on the relationship of the model learning curve and sample size [55]. Here we are going to use 5-fold cross validation.

Moreover, most K-fold cross validation procedures assume the data to be independent and identically distributed (i.i.d.). In our case, we can't make that assumption since the samples have been generated from a time-dependent process. A variation of K-fold CV which considers the temporal aspect when creating the folds is used. It deals with the time dimension by creating successive training sets as supersets of the preceding ones, to avoid the *look ahead* bias. Furthermore, the generative process has a grouped structure, since we have observations coming from different firms. In this situation, we are interested in avoiding overfitting to the subset of firms available in the training sample and ensure that the model will generalize well for all construction stocks. As a result, a variation of K-Fold cross-validation considering the time-series and grouping dimensions has been implemented.

4.5.1. Imbalanced data and asymmetric misclassification costs

Earnings' surprises are an exception rather than the norm, being the most common outcome earnings' reports in-line with market expectations. This inherent characteristic of the target variable leads us to a particular modelling setting: an imbalanced dataset[25]. This has a major impact in the prediction power of algorithms since they will have a bias towards selecting the most frequent class. In our case the majority class is *no surprise* with a proportion of 1:4:1.

Dealing with imbalanced datasets is not a novel problem for the predictive analytics community, and several techniques have been developed to improve algorithms' performance. These methodologies can be broadly classified according to their different approaches: under- and over-sampling address the problem by generating a balanced training set through re-sampling, while an alternative approach consists in giving more importance to the misclassification of minority classes by assigning different costs.

While resampling techniques are the simplest and most efficient method for dealing with class imbalance, the new observations created through the resampling procedure in current techniques fail to account for the time dependency among the observed values. Moniz et al. [56] claim, this temporal dimension should be taken into account when changing the distribution of the data, and propose extensions to current resampling methodologies, by introducing a temporal and relevance selection bias, in an attempt to deal with this problem. As a result, in this work the latter approach has been chosen, adjusting class weights to place more weight on the minority classes.

Moreover, we are dealing with asymmetric misclassification costs as a result of our desired investment strategy. The output of the classifier is going to be used as a signal to build long-short portfolios aiming to profit from earnings' surprises. In simple terms this means that, when detecting a positive surprise we will be buying the stock, and when identifying a negative one we will be short selling it. Therefore, when predicting a surprise the investment strategy will be putting money down the line, as opposed to not predicting one and staying passive. It can be seen that we are particularly sensitive to misclassifying a surprise in the opposed direction, being a less harmful error classifying

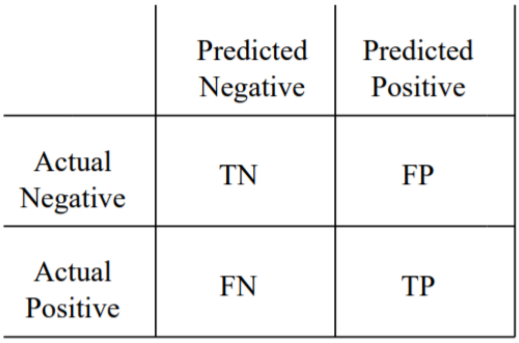
[25] A dataset is said to be imbalanced when one or more of the target classes is under-represented with respect of the others [57].

it as a no surprise. The least damaging misclassification will be predicting surprises as no surprise, since we will be just leaving money on the table.

In the next section, a performance metric that accounts for the imbalance in the data and the above-mentioned misclassification cost asymmetry is designed, aiming to be able to identify the most relevant model for our desired business application.

4.5.2. Performance Metrics

The most universal way of evaluating classification algorithms is by using the confusion matrix. This matrix, as illustrated in figure 10, has the actual classes as rows and the predicted classes as columns, and classifies observations as: *True Negatives*, *False Positives*, *False Negatives* and *True Positives* [57].



	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Figure 10. *Example of a confusion matrix (Source: Chawla et al.* [57]*)*

The typical metric for evaluating classification performance is predictive accuracy, nonetheless it's not appropriate for imbalanced data or asymmetric misclassification importance [57]. In an imbalanced setting, like the one we face in this study, a pertinent way of measuring a classifier performance is by using the Precision, Recall and F_β score metrics. Precision measures how relevant the results are and Recall how many truly relevant results are returned. The F_β score is defined as the harmonic mean between precision and recall parametrized by β. The parameter β is defined as the importance factor between Recall and Precision, being Recall β times as important as precision [58]. The mathematical definition of the above metrics are the following:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F_{\beta} = (1+\beta)^2 \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{9}$$

These metrics are defined for binary classification but are extended to a multi-class setting by applying them to each class independently and computing the average across classes. There are several ways of performing this average: *micro*-, *macro*- and *weighted average*. The *macro-average* gives equal weight to each class and is preferred for imbalanced classes if we want to bias the metric towards the minority classes; while the *micro-average* considers the contribution of each class consequently biasing it to the majority classes. By comparing the *micro*- and *macro-averages* the classification of the majority and minority classes can be assessed. A micro average significantly above the macro average will indicate poor classification ability across the minority classes as opposed to probably correctly classified majority classes; conversely a macro average significantly larger than the micro average will signal the opposite.

As presented in the previous section, we are dealing with an imbalanced dataset with asymmetric misclassification importance. On the one hand, earnings' surprises, either positive or negative, are rare occurrences, and at the same time the ones we are interested in being able to predict in a reliable manner. Being able to predict well *no surprises* is of no value to our investment purpose, thus we have a bias towards the minority classes. On the other hand, there is a different cost associated with misclassifying each of the categories. We are really sensitive to misclassifying surprises in the opposite direction and less sensitive to *no surprise* misclassification or not identifying them.

In short, we are interested in the minority classes and we are more sensitive to a certain kind of False Positives than False Negatives, being more important the precision than the recall for the minority classes. For that reason, I am going to choose as evaluation metric $\beta = \frac{1}{2}$ giving Precision twice the importance as Recall and use the *macro-average* to emphasize the model performance for earnings' surprises. Both averages are reported for misclassification comparison purposes between the majority and minority classes.

4.5.3. Machine Learning Algorithms

In this research work, five different machine learning algorithms are considered: a multinomial logistic regression, a bagging classifier, a random forest, a boosting algorithm and a Support Vector Machine. The first one is a linear method, the next three are *decision tree*-based ensemble methods, and the last one a discriminant classifier. In this section, each algorithm is briefly presented aiming to give the reader the intuition behind each one of them. The interested reader can find a more detailed explanation in *The Elements of Statistical Learning* by Hastie et al. [55].

Algorithms are benchmarked with a baseline model which consists in a zero-rule classifier which always predicts the majority class: *no surprise*. For settings with imbalanced data this baseline model definition is more appropriate than random guessing, since it achieves better results. This is equivalent to assuming that the analysts' consensus it's always accurate enough so there shouldn't be any earnings surprise.

Multinomial Logistic Regression

A *Logistic Regression* is a statistical model where the log-odds of a binary dependent variable are modelled as a linear combination of independent variables, being the mapping between the log-odds and probability known as the logistic function. It's worth mentioning that the logistic regression does not constitute a classifier by itself, being its output probabilities from which a classifier can be built. The *Multinomial Logistic Regression* is an extension of the former model to the multi-class setting, being used to predict the probabilities for each level from a set of independent variables [59].

Bagging

Bagging, with Random Forests and Boosting, belong to a family of learning techniques known as *Ensemble Learning*. Ensemble methods rely in a simple yet powerful idea, 'combining the output of many *weak* classifiers[26] to produce a powerful *committee*' [55]. They work by developing a population of simple classifiers from the training set and then combining them to produce a predictor.

[26] A *weak* classifier is defined as 'one whose error rate is only slightly better than random guessing' [55].

A bagging classifier fits a series of base learners to bootstrapped[27] samples of the training set and gives a prediction by either averaging individual predictions or taking the majority vote class. The use of bootstrapped samples helps reducing the variance of the classifier, and despite the bootstrapped trees may include different subsets of features they are not completely independent [55]. In this research application, a bagging classifier is combined with a previous under sampling rebalancing step to deal with the class imbalance and boost its performance, leading to a balanced bagging classifier. This is done as an experiment despite the temporal considerations exposed in section 4.5, and has noted that under-sampling is not believed to be as problematic as over-sampling.

Random Forest

The *Random Forest* technique constitutes a modification of the bagging procedure aiming to improve the variance reduction by reducing the correlation among the bootstrapped trees. The de-correlation among decision trees is achieved by randomly selecting the features considered at each split from a subset of the input variables during the tree growing process. This reduces the pairwise correlation among the trees, therefore reducing the variance on average [55].

Extreme gradient boosting (XGBoost)

Despite boosting is based in the same simple idea as other ensemble methodologies, it diverges from them in two crucial ways: the population of base classifiers used evolves over time and their vote is a weighted average [55]. It works by sequentially applying those simple classifiers to modified versions of the data and combining their predictions through a weighted majority vote. The data is sequentially modified to introduce a penalty for misclassified observations, and the weighted prediction gives more importance to the classifiers with better accuracy [55]. Boosting classification algorithms have proven to be very successful for a diverse range of problems, with one algorithm standing out for its effectiveness, low computational cost and scalability : XGBoost [60].

[27] Bootstrapping consists in creating datasets of the same size as the training set by randomly drawing with replacement [55].

Support Vector Machine

The *Support Vector Machine* methodology consists in building a set of hyperplanes in a high-dimensional space which can be used to classify new observations. The original procedure was designed to find the optimal separating hyperplane for linearly separable data, but it can be extended to non-linear classification using data transformations. These data transformations, known as *kernel* functions, map the data into a high-dimensional feature space to make it linearly separable [61]. The problem of finding the optimal hyperplane is a QP constrained optimization problem.

5. Results

In this section, the results of this research work are presented. First, an overview of the resulting data sample and its characteristics is done. Second, the feature selection procedure results are exposed following the methodology presented in the previous section. The results are grouped according the variable origin and analyzed independently from the other sets of variables. Third, the model selection results are analyzed, and the final model selected. Then, the chosen model performance is assessed, and the identified variable importance analyzed. Finally, the proposed investment strategy is implemented and back tested, studying its performance.

This research project has been implemented using Python, and the different methodologies used in this section with the *SciPy* and *Scikit-learn* packages[28]. For the sake of reproducibility, all the code necessary to replicate the results of this research work are available in the following website: https://tinyurl.com/yx3ykom7 .

5.1. Data sample characteristics

The complete dataset consisted of 2,068 observations and 33 features covering the last 20 years, which were divided into a train and test set with an 80/20 split. The train set consisted of 1609 records and was used for feature and model selection. The test set consists of 459 observations and is used to test the predictive performance of the machine learning models and back-test the investment strategy based on the model signals. The dataset is imbalanced with a 1:4:1 ratio, being no surprise the dominant class.

Table 12. SIC classification of the firms in the data sample.

Division	***Major Group***	***Number of firms***	
		Train	*Test*
C Construction	*15 Building Construction, General Contractors and Operative Builders*	*30*	*19*
	16 Heavy Construction other than Building construction contractors	*9*	*7*
	17 Construction special trade contractors	*11*	*7*

[28] More information on the above mentioned packages can be found in: https://docs.scipy.org/ and https://scikit-learn.org/.

As can be seen in table 11, out of the 57 construction firms in the data sample, there are 47 in the train set and 33 in the test set. The proportion of major groups is the same for the train and test sets with 60% of the firms belonging to Residential Building contractors, 20% to heavy construction contractors and 20% to construction related trades. The data sample is dominated by the residential building contractors category.

5.2. Feature Selection

In this section the results for the feature selection methodology introduced in section 4.4 are presented. The independent variables that have been developed are different in nature and come from different data sources; thus they will be analyzed separately.

5.2.1. Market data and Expectations

As can be seen in figure 11, the size variable distribution under the different earnings' announcement scenarios we are considering is in line with the evidence found in the academic literature: smaller firms tend to be more prone to earnings' surprises as a consequence of inattention by market participants. The analyst forecast differences variable does not appear to have different distributions, with outliers in the one and three months look back periods.

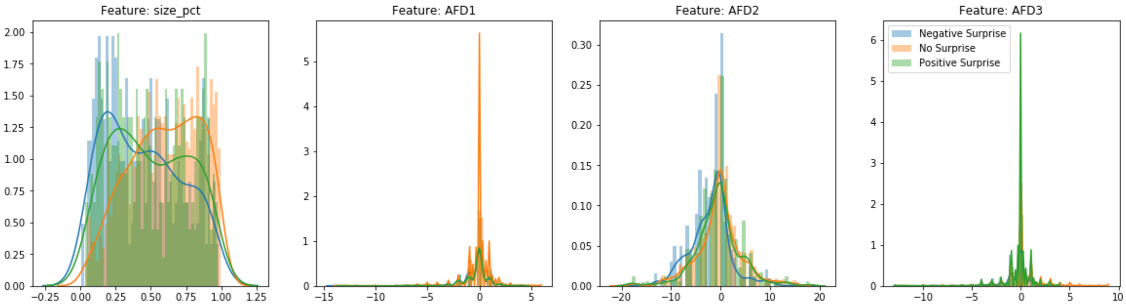


Figure 11. Distributions of the numerical market and expectations variables under the different categories of the target variable.

In the following table, the results for the Kruskal-Wallis test on the numerical variables, and the chi-squared test on the categorical ones, as well as the mutual information, are summarized:

Table 13. Results for the univariate statistical tests on the market and expectations variables.

Feature	***Kruskal-Wallis test***		***Chi-Squared Test***		***Mutual Information***
	H statistic	*p-value*	*statistic*	*p-value*	
Size	88.295	0.000	-	-	0.042
AFD1	13.450	0.001	-	-	0.018
AFD2	51.155	0.000	-	-	0.018
AFD3	30.810	0.000	-	-	0.011
FECq1	-	-	18.826	0.000	0.126
FECq2	-	-	5.983	0.050	0.055
FECq3	-	-	3.924	0.14	0.082
FECq4	-	-	5.389	0.07	0.065

The size and analyst forecast differences are found to be statistically significant; while for the lagged surprises we fail to reject the null hypothesis for the past surprises in the last two, three and four quarters. In the literature surprises are found to be autocorrelated among four adjacent quarters (figure 3), being the results of the test not in line with it. This might be interpreted as the previous past surprises already incorporating all the relevant information of the previous three for predicting the next one. Regarding the mutual information all variables add little information when classifying earnings' surprises, but none is identified to be independent. As a result, only the past surprise from the previous quarter is going to be considered, with the size and all analyst forecast differences variables.

5.2.2. Global Risk Factors

In figure 12, the distributions of the different global risk factors under the different surprise categories are shown. Both construction spending variables appear to have different distributions under the surprise categories. Public spending appears to be distributed differently for positive surprises, while private spending looks like is under the negative surprise category that is distributed in a different way. As a result, both variable might have discriminative power on the sign of the surprises. Also, for the Machinery PPI, positive surprises appear to have a slightly different distribution than the rest of classes. The difference among classes for the unemployment indicator appears to be in the distribution kurtosis. There can be seen much difference for the rest of global risk factors. In the following table the results for the univariate tests and the RFE procedure on the global risk factors are presented:

Table 14. Results for the univariate statistical tests and Recursive Feature Elimination procedure on the Global Risk factors.

Feature	***Kruskal-Wallis test***		***Mutual Information***	***Recursive Feature Selection***	
	H statistic	*p-value*		*Logistic Regression*	*Random Forest*
GNI	50.710	0.000	0.073	✓	
Interest Rates	47.567	0.000	0.053	✓	
PPI for Construction Materials	3.137	0.208	0.084	✓	
PPI for Construction Machinery	2.750	0.253	0.080	✓	✓
Public Construction Spending	37.892	0.000	0.097	✓	
Private Construction Spending	85.274	0.000	0.074	✓	✓
Unemployment Excess	23.165	0.000	0.089	✓	
Political Instability	0.538	0.764	0.079	✓	

The results for the Kruskal-Wallis test show that the considered PPIs and the Political Instability index are not statistically significant. The mutual information measure shows the same weak degree of dependency for all variables, with no variable deemed as independent from the earnings' surprises. The Recursive Feature Elimination procedure with the logistic regression identifies all variables as relevant while the random forest model only keeps the machinery PPI and private spending. Private construction spending relevance over the public one is expected, when one thinks of the behavior of both play during recessionary environments in the United States. Public spending is usually seen as a tool for economic stimulation during downturns, being less sensitive to the macroeconomic fluctuations than the private one.

As a result, only the Interest Rates, Machinery PPIs, public and private construction spending and unemployment excess are kept. The GNI and Political Instability index are discarded given their different temporal scale. This pair of macroeconomic and geopolitical indicators are better suited for cross-country comparisons, and over periods of time of the order of decades; being them unsuitable for our purpose. The machinery PPI is kept since its not discarded by both wrapper methodologies, despite not being found significant by the univariate tests.

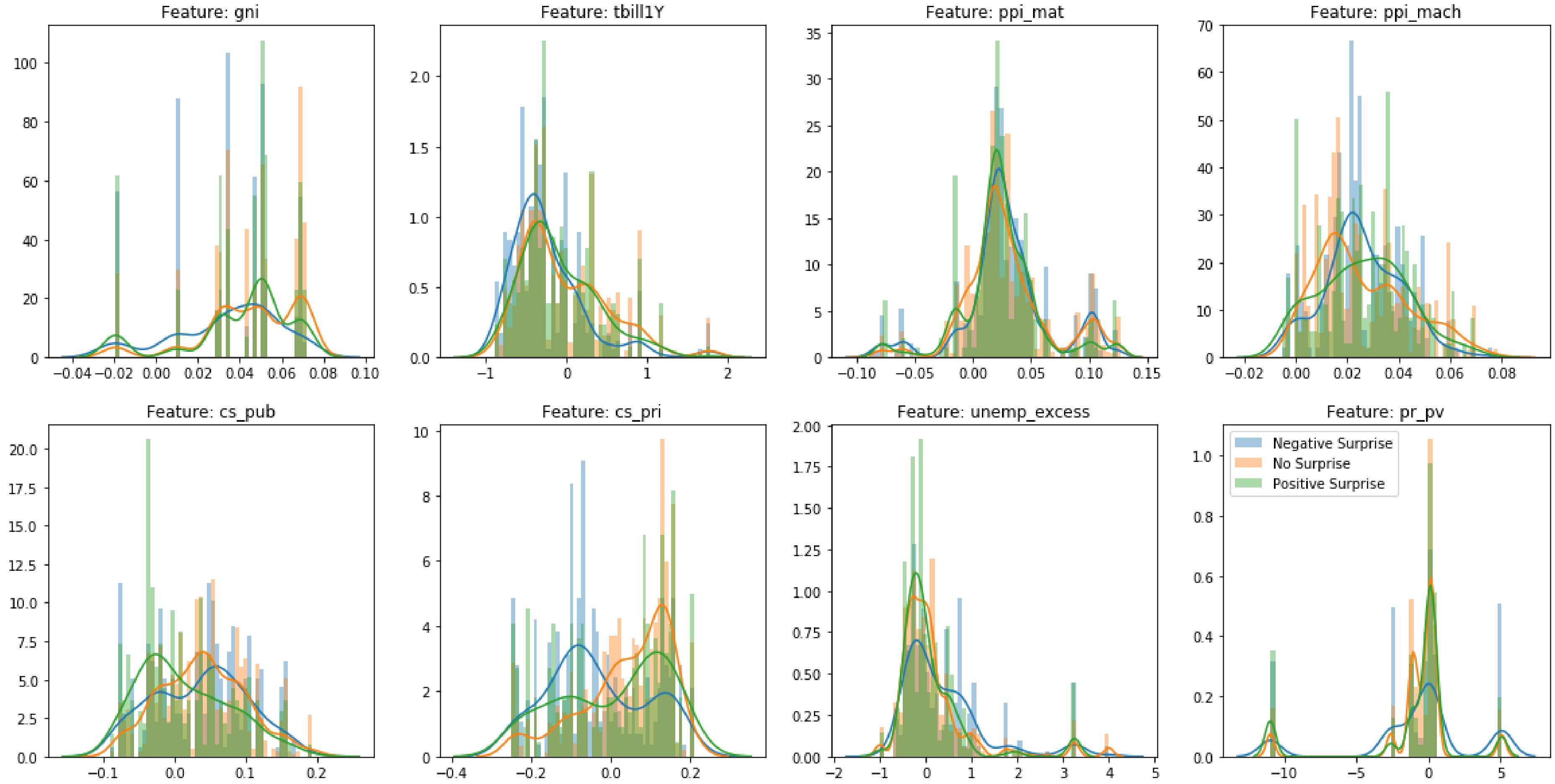


Figure 12. Distribution of the Global Risk Factor variables under the categories of the target variable. Note that the GNI (gni) and Political Instability variables had annual frequency and the annual value was propagated across all months, as opposed to the rest of variables which had monthly frequency.

5.2.3. Intrinsic Variables

First, we are going to analyze the normalized delay variable. If we take a close look at the figure 13, there appears to be seldom difference between surprises and no surprise, being the former more prone to experience a delay. This gives us a slight hint towards supporting the view that non-expected results reporting tends to be delayed with respect of the normal timing for a given firm. When it comes to the statistical tests, the normalized delay is found to be significant and mutual information non-zero.

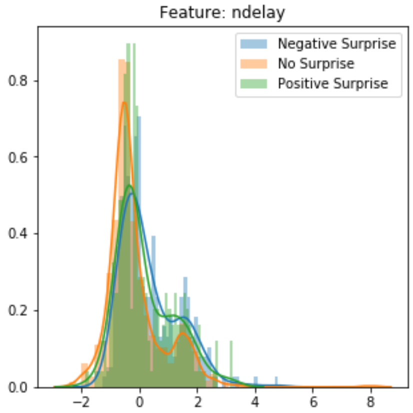


Figure 13. Distribution of the delay in reporting earnings under the categories of the target variable.

Regarding the financial ratios, their distributions under the different target variable classes are shown in figure 14 and the results of the hypothesis tests are presented in table 15. When looking at the financial dimensions these ratios portray, it seems that the most relevant ones are profitability and valuation.

When looking at figure 14, apparently the asset turnover, net profit margin, return on assets, return on capital employed and return on equity are the only variables that showcase a different distribution under the surprises' categories. For the Asset turnover negative surprises appear to be differently distributed versus the other groups. The valuation ratios seem to be slightly different distributed under both surprise categories, with no discriminative power between positive and negative surprises.

The proposed statistical test does not find statistically significant the debt-to-equity ratio, payables and receivables turnover, cash flow-to-debt ratio, interest-to-debt ratio, and current and quick ratios. It's worth mentioning that the last three ratios are sparsely populated and were imputed by the industry median. The proposed test is sensitive to this fact, being it a probable cause for failing to reject the null hypothesis. The accounts payable turnover and current ratio were found to be independent of the target variable according the mutual information metric.

When it comes to the Recursive Feature Elimination methodology the proposed linear model happens to keep only one variable per financial dimension, with two exceptions: it selected two valuation metrics and dropped both liquidity ratios. As mentioned above, liquidity ratios are sparsely populated and imputed by the mean; which may be the cause behind its rejection from the model despite the relevance found in the literature. Conversely, the *tree ensemble* model keeps 15 out of the 21 financial ratios considered, discarding the debt-to-equity, receivables turnover, cash flow-to-debt, interest-to-debt, and current and quick ratios. It's worth mentioning that the model keeps all the profitability and valuation metrics, reinforcing their relevance, while discarding six across the rest of financial dimensions. Furthermore, the selection of inventories over assets turnover is consistent with the construction firms' operational risks. Unfinished construction work is considered as an inventory, thus being this ratio a crucial indicator of a construction firm's activity. Again, both liquidity ratios have been discarded by the model.

As a result, and aiming to keep the interpretability of the model, only one financial ratio per dimension will be considered. The financial ratios kept for the final model, and identified with an asterisk in figure 15, are the intersection of the subsets selected by both RFE models with two exceptions. On the one hand, the quick ratio is included as a measure of liquidity. Liquidity it's identified as a critical dimension for construction firms in the literature, and it not being statistically significant might just be a product of data sparsity and imputation. On the other hand, only one valuation metric is selected: the enterprise value. This multiple is a common metric employed by investment bankers, while the P/E and price-to-book ratios might be a more used ratio among investment managers. The main difference among them is that the former takes into account the firm's capital structure; thus incorporating information about its leverage. Since the capitalization dimension is of utmost importance for construction firms, selecting the enterprise value over the price-to-book ratio appears to be more appropriate.

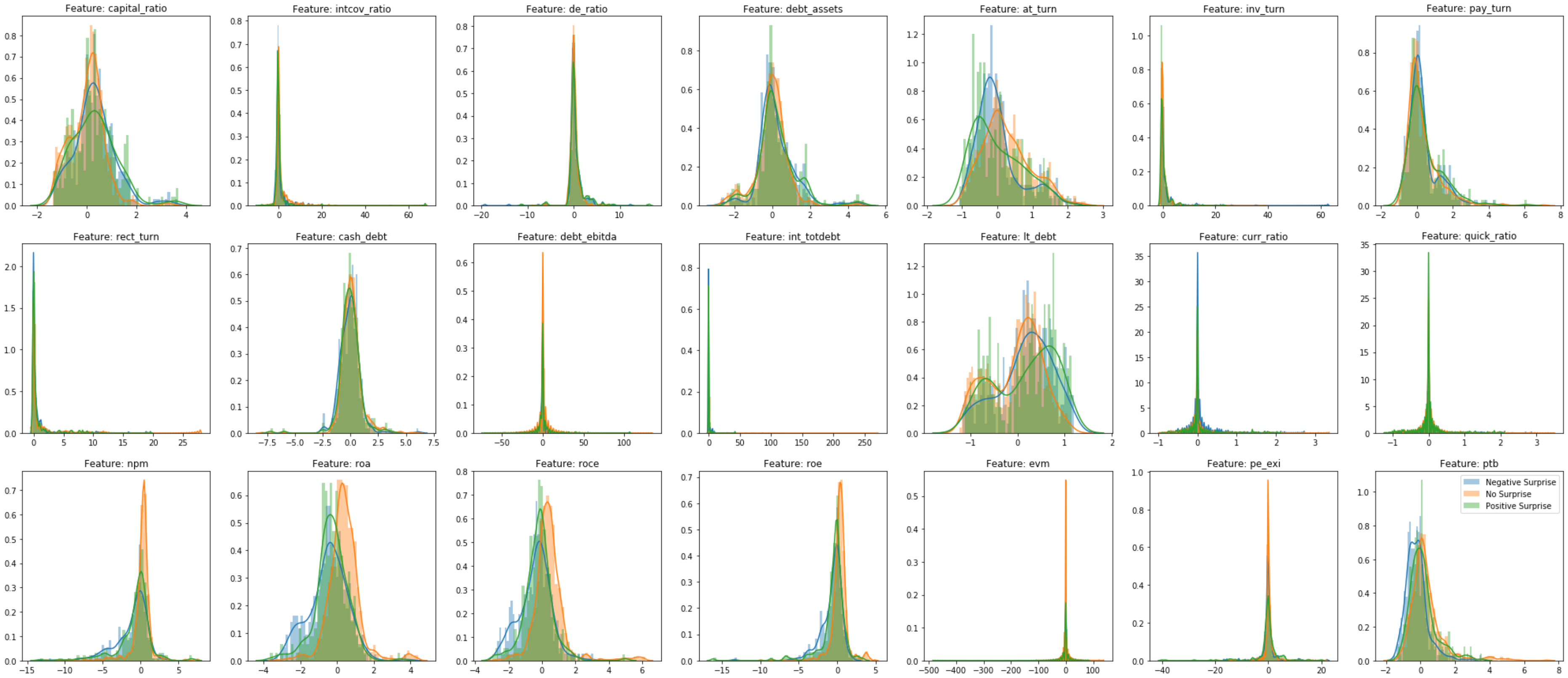


Figure 14. Distributions of the considered financial ratios under the categories of the target variable with outliers.

Table 15. Results for the univariate statistical tests on the different financial ratios. The selected financial ratios have been identified with an asterisk.

Financial Dimension	***Feature***	***Feature Name***	***Kruskal-Wallis test***		***Mutual Information***	***Recursive Feature Selection***	
			H statistic	*p-value*		*Logistic Regression*	*Random Forest*
Capitalization	Capitalization Ratio*	capital_ratio	49.227	0.000	0.028	✓	✓
	Interest Coverage	intcov_ratio	190.833	0.000	0.089		✓
	Debt-to-Equity	de_ratio	7.394	0.025	0.043		
	Debt-to-Assets	debt_assets	12.673	0.002	0.039		✓
Efficiency	Asset Turnover	at_turn	64.402	0.000	0.035		✓
	Inventory Turnover*	inv_turn	28.121	0.000	0.053	✓	✓
	Payables Turnover	pay_trun	7.015	0.030	0.000		✓
	Receivables Turnover	rect_trun	1.884	0.390	0.039		
Financial Soundness	Cash Flow-to-Debt	cash_debt	8.695	0.013	0.027		
	Debt-to-EBITDA*	debt_ebitda	17.328	0.000	0.096	✓	✓
	Interest-to-Debt	int_totdebt	1.536	0.464	0.010	✓	
	Long-term Debt-to-Total Liabilities	lt_debt	52.021	0.000	0.007		✓
Liquidity	Current Ratio	curr_ratio	0.370	0.831	0.000		
	Quick Ratio*	quick_ratio	1.495	0.474	0.013		
Profitability	Net Profit Margin	npm	230.241	0.000	0.078		✓
	Return on Assets*	roa	282.164	0.000	0.090	✓	✓
	Return on Capital Employed	roce	301.434	0.000	0.116		✓
	Return on Equity	roe	275.798	0.000	0.094		✓
Valuation	Enterprise Value*	evm	23.641	0.000	0.114	✓	✓
	Price-to-Earnings	pe_exi	48.225	0.000	0.062		✓
	Price-to-Book	ptb	134.052	0.000	0.054	✓	✓

5.3. Model Selection and Assessment

As a result of the feature selection procedure implemented in the previous section from the original 28 features considered, at the end only 17 are used to feed the machine learning model. This includes five market and expectation variables, five global risk factors, the delay in reporting earnings' and six financial ratios representing each one of the financial dimensions considered. The summary statistics of the final feature subset for the training set are presented in table 16.

Table 16. Summary statistics of the numerical features used to feed the machine learning models.

feature	*count*	*mean*	*stdev*	*min*	*p25*	*p50*	*p75*	*max*
Size	1609	0,552	0,263	0,005	0,335	0,557	0,783	0,994
AFD1	1609	-0,346	1,574	-14,000	-1,000	0,000	0,000	6,000
AFD2	1609	-0,669	4,699	-20,000	-3,000	0,000	1,000	18,000
AFD3	1609	-0,207	1,804	-13,000	-1,000	0,000	0,000	9,000
Interest Rates	1609	-0,071	0,525	-0,889	-0,429	-0,263	0,270	1,792
PPI Machinery	1609	0,026	0,017	-0,005	0,013	0,023	0,037	0,078
Public CS	1609	0,035	0,062	-0,091	-0,015	0,036	0,084	0,193
Private CS	1609	0,021	0,122	-0,251	-0,073	0,043	0,127	0,207
Unemployment Excess	1609	0,180	0,839	-1,000	-0,299	-0,040	0,381	4,000
Delay	1609	0,005	1,022	-2,427	-0,636	-0,314	0,472	8,204
Capitalization Ratio	1609	0,117	0,765	-1,337	-0,367	0,143	0,503	3,710
Inventories Turnover	1609	0,753	3,478	-0,693	-0,328	-0,078	0,704	70,201
Debt-to-EBITDA	1609	0,178	7,671	-67,450	-0,410	0,003	0,399	121,925
Quick Ratio	1609	0,097	0,569	-1,217	0,000	0,000	0,000	3,493
Return on Equity	1609	0,098	0,998	-4,090	-0,408	0,156	0,651	4,393
Enterprise Value	1609	-0,341	14,476	-436,908	-0,353	-0,017	0,420	117,967

In the following sections, the results from the model selection strategy are presented and the final model selected. Then, its performance in the training set and the variable importance analyzed.

5.3.1. Model Selection

In table 17, the cross-validation results for the five different learning techniques considered are presented:

Table 17. F-score with beta 0.5 micro- and macro-averaged across classes. The performance in the training and test set are shown.

Model	$F_{.5}$ *Score (Micro)*		$F_{.5}$ *Score (Macro)*		Balanced Accuracy
	Train Score	*Test Score*	*Train Score*	*Test Score*	
Baseline Model	0.77	0.62	0.27	0.22	0.33
Logistic Regression	0.68	0.49	0.54	0.40	0.43
Balanced Bagging Classifier	0.83	0.52	0.72	0.40	0.44
Random Forest	1.00	0.65	1.00	0.37	0.39
XGBoost	0.93	0.62	0.91	0.39	0.40
Linear SVM	0.80	0.64	0.48	0.39	0.41

As it can be seen, the best performing models according to the $F_{.5}$ macro score are the logistic regression and balanced bagging classifier, closely followed by the XGBoost, the linear SVM and Random Forest. All proposed models are found the have better predictive capabilities than the baseline classifier when it comes to predicting earnings surprises. It's worth noting than *tree ensemble* techniques achieve high scores in the training set that do not translate to the test folds, they appear to be overfitting the data. Moreover, the Random Forest, XGBoost and SVM have significantly better micro- than macro-averaged scores indicating that they are most likely predicting poorly the minority classes and correctly the majority class. That behavior is to be expected as a result of the class imbalance commented before, which is introducing a bias in the classifier and hindering its predictive power. Since we are in the business of predicting earnings surprises, let's take a closer look at each algorithm performance when it comes to the minority classes before making a final decision.

Table 18. Precision and Recall macro-averaged for the relevant classes, positive and negative earnings' surprises. The performance in the training and test set are shown.

Model	Precision		Recall	
	Train Score	*Test Score*	*Train Score*	*Test Score*
Baseline Model	0.00	0.00	0.00	0.00
Logistic Regression	0.34	0.29	0.64	0.38
Balanced Bagging Classifier	0.56	0.26	0.96	0.38
Random Forest	1.00	0.29	1.00	0.11
XGBoost	0.94	0.30	0.77	0.16
Linear SVM	0.41	0.38	0.21	0.18

In table 18, the estimated precision and recall when predicting earnings surprises is presented. On the one hand, we can see that the recall when predicting surprises is quite low for the random forest, XGBoost and linear SVM techniques, returning less than 20% of the relevant cases. The logistic regression and bagging classifier fare a bit better, identifying around 40% of the relevant cases. On the other hand, the precision of all algorithms is around 30% for the minority classes, being the linear SVM the best one with 38%.

Overall, the best performing models are the logistic regression and the linear SVM. The logistic regression returns twice the amount of relevant cases when compared to the SVM, despite the latter achieves slightly higher precision. The final decision comes out to be a tradeoff between precision and recall. As mentioned before, we have asymmetric misclassification costs that lead us to prefer precision over recall, however the misclassification error we are more sensible to is classifying surprises in the opposite direction being a less critical mistake identifying as a no surprise an actual surprise. For this reason, the logistic regression model has been chosen over the linear SVM, since it is able to return more relevant cases without a dramatic decrease in precision hoping the misclassification mistake won't be critical.

5.3.2. Final Model Assessment

In the following table the results for the multinomial logistic regression in the test set are presented:

Table 19. Model Summary for the Logistic Regression Model.

Target Variable	***Precision***	***Recall***	***$F_{.5}$ score***	***Support***
Negative Surprise	0.06	0.13	0.07	53
No Surprise	0.85	0.43	0.71	344
Positive Surprise	0.22	0.61	0.25	62
Micro avg.	0.42	0.42	0.42	459
Macro avg.	0.38	0.39	0.35	459
Balanced Accuracy	-	-	0.39	459

The overall model performance in the test set, as measured by the $F_{.5}$ macro score is in line with the one estimated through cross validation, being 0.35 versus the 0.40 obtained before. This proves that the model selection procedure has been properly done and it didn't overestimate the test score. The resulting model exhibits a biased behavior towards the positive surprise category, being able to return around 60% of the cases with low precision (22%). Conversely, it fails to identify almost any negative surprise returning only 13% of those with 6% precision. Regarding the majority class the model achieves high precision (85%), but only identifies less than half of the cases. This hints us that the algorithm is unable to distinguish between positive and negative surprises and is returning as positive surprises a high percentage of no surprises. Let's analyze further the misclassification behavior by looking at the confusion matrix in figure 15.

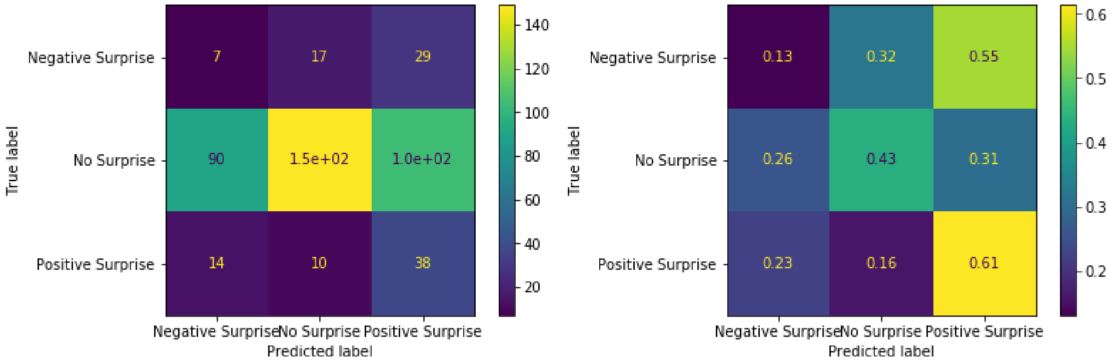


Figure 15. Confusion matrix (left), and its normalized version (right).

In the confusion matrix, the hinted behavior is clearly observable. From the 62 true positive surprises in the test data, 38 have been correctly classified, but a 105 of the predicted positive surprises are no surprise and 29 of them a true negative surprise. In short, only 22% of the predicted positive surprises are correctly classified, being 61% of the true positive surprises. For the negative surprises, only 7 are correctly classified, being 29 of them classified in the opposite direction and 17 identified as no surprise. Ergo, only 6% of the predicted negative surprises are correctly classified, being only correctly classified a 13% of the negative surprises in the test data. Regarding the majority class, it correctly predicts 43% of the existing ones, misclassifying 31% as positive surprises and 26% as negative surprises. Overall, the model is biased towards returning positive surprises and misclassifying the negative and no surprise classes as such. Regarding the sensitive cases for our business purposes, around 13% of the predicted negative surprises are actual positive ones; while 17% of the predicted positive surprises are negative ones.

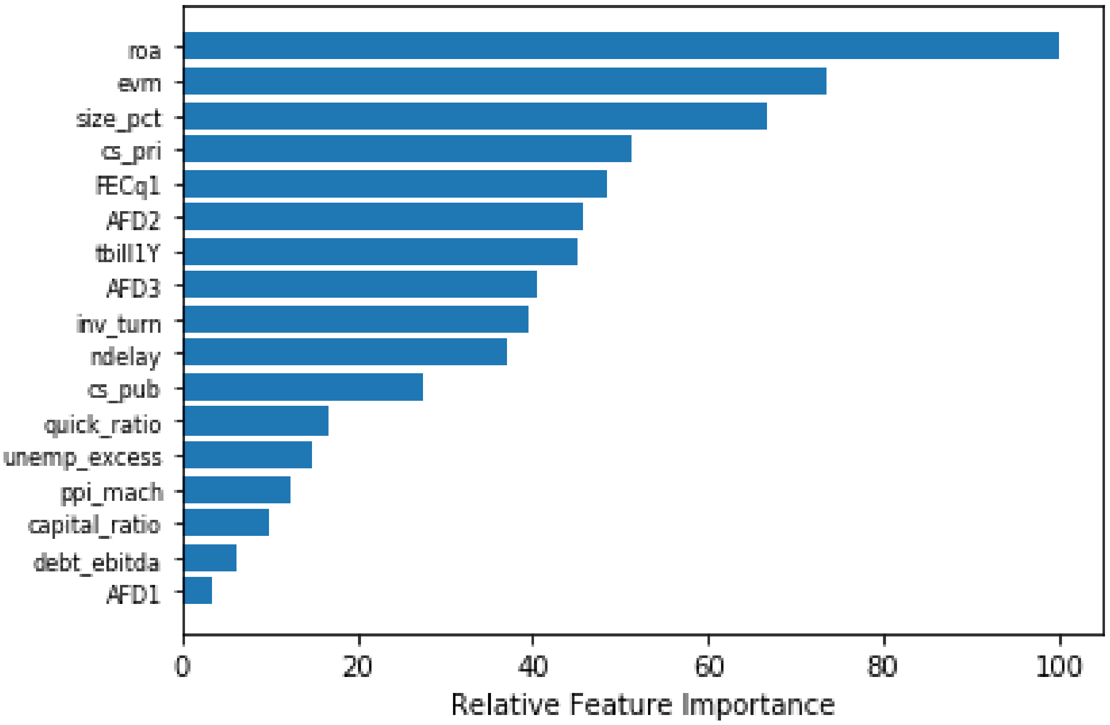


Figure 16. Relative feature importance for the multinomial logistic regression model.

In figure 16, the relative variable importance of the different features is presented. It can be observed, the most important variables when it comes to predicting earnings surprises are the profitability and valuation metrics, as represented by the return on assets and enterprise value multiples, and the company size. The profitability and valuation dimensions relevance are consistent with previous studies on construction firm's financial performance. The former, is identified as the most important metric when analyzing construction stocks, since it's closely related with earnings quality and has a positive correlation with dividend distributions [19]. The latter dimension importance its derived from the metric nature, overvalued firms will tend to disappoint when reporting results while undervalued ones have more potential for positive surprises and upward price drifts. The relationship between earnings surprises and the size was expected and is consistent with the literature (see section 2.3).

The next group of variables in relevance include two global risk factors, the inventory turnover, the delay in reporting earnings' and the rest of market and expectation features, but AFD1. Regarding the global risk factors, the private construction spending is found to be the 4th ranked variable in importance with the interest rates indicator (7th) which is consistent with the stated close relationship with the national economies, as both are relevant macroeconomic indicators. The delay in reporting earnings is found reasonably relevant, somewhat supporting the view that managers delay earnings' reporting when results are not in line with market expectations. The inventories turnover is a measure of the company's efficiency and particularly important for construction firms as stated before. The market and expectation data are also found to be reasonably relevant consistent with previous the previous research findings.

The rest of variables are not found the be particularly relevant in identifying earnings' surprises, which include: the public construction spending, quick ratio, unemployment excess, the machinery PPIs, the capitalization, the debt-to-ebitda ratios and the AFD1. The reason for the AFD not being relevant can be a result of data flaws or the lack of analyst revisions when approaching the end of the fiscal period.

Overall, the proposed model does not achieve better classification results than the particularization of the previous one. The proposed model is able to achieve a higher recall for the minority classes, while substantially reducing it for the majority class. When it comes to precision, its slightly improved for the no surprise category while reduced in our categories of interest.

5.4. Investment Strategy Viability

Once, the final model is selected and its performance has been evaluated, now it's time to analyze if the investment strategy resulting from it is able to achieve abnormal returns. The investment strategy is back tested in a for a four-year period, starting in January 2015 until December 2018, and it includes 28 construction companies. In figure 17, the cumulative abnormal returns for the different portfolios are presented.

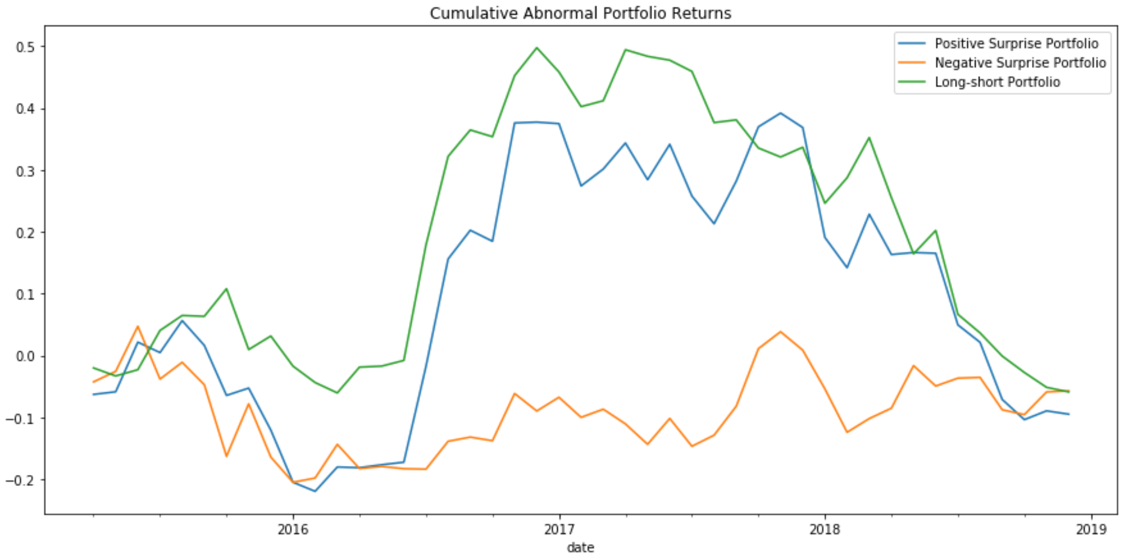


Figure 17. Cumulative abnormal returns of the positive surprise, negative surprise and long-short portfolio

As it can be seen, the portfolios showcase the same behavior as the model output, and the long-short portfolio performance is mainly driven by the one of the positive surprise portfolio. During the first year the strategy cumulative abnormal returns stay around zero, as result of the poor performance of the long portfolio and a reasonably good performance of the negative one. For the next two years, the positive surprise portfolio is able to harvest cumulative abnormal returns while the negative one fails to. Towards the end of the back-test there is a decrease in the performance of the long portfolio, thus of the overall strategy. This can be due to the nature of the time series being the relationships found not extensible for long timespans, requiring a recursive training of the model. In addition, the negative surprise portfolio returns has a consistent behavior over the whole back test, oscillating the returns around zero, and failing to achieve cumulative abnormal returns. This behavior is consistent with the inability of the model for predicting negative surprises. Overall, the performance of the long and short portfolios is not severely hindered by the misclassification of observations in the opposite direction.

6. Conclusions

The conclusions are structured as follows: first the research results are summarized, then the main contributions of this research work are exposed, which are followed by an assessment of the fulfillment of the goals established for this project, and this section concluded by outlining further research lines.

6.1. Summary of the results

In this research work a systematic investment strategy, which uses machine learning techniques and leverages an in-depth understanding of the construction industry operational and management practices, is successfully carried out. First, a literature research in the field of behavioral finance yielded several potential investment opportunities, being the anomalies around earnings' announcements the one found to be the most suitable for our purposes. Second, the extensive research work on the operational characteristics of construction companies and its risk management practices, was used to successfully identify two broad groups of risk factors affecting their performance: global risk factors and company specific ones. Third, univariate tests and a recursive feature selection procedure were implemented to select the most relevant subset of features, composed by five market and analysts' expectation variables, five macroeconomic indicators and six financial ratios representing the most important financial dimensions of construction firms. Fourth, the data temporal dimension and generative process, imbalance across classes and asymmetric classification costs that characterize the modelling setting are addressed by using the macro-averaged F-score with β=0.5 and an appropriate cross-validation scheme. Five state-of-the-art machine learning algorithms are tested, being found the most suitable one the multinomial logistic regression. The resulting model exhibits a bias towards predicting positive earnings' surprises over negative ones and identifies as the most relevant variables the profitability and valuation measures, portrayed by the Return on Assets and Enterprise Value multiples, and the firm size. To conclude, the systematic investment strategy based on the investment signals produced by the machine learning model is back tested in a four-year period failing to consistently achieve abnormal returns, and sharing with the model the same bias for positive surprises.

Market participants deviating from the customary rational decision-making model with common behavioral biases, and the limits of arbitrage, make financial markets inherently inefficient. Quarterly earnings' announcements are identified as one of the most important recurrent releases of information by public companies, spotting a market anomaly around them: the pre- and post-earnings announcement drifts. When reported returns diverge significantly from market expectations, an earnings' surprise takes place. As a consequence, the stock price experiences a drift in de direction of the surprise from 60 days before to 60 days after the information shock, being the *post* drift explained in terms of inattention by market participants. The selected systematic investment strategy aims to capture this market anomaly by predicting earnings' surprises at the end of each fiscal period and building long-short portfolios on the expected positive and negative surprise firms over a holding period equal to the quarter length.

The construction industry is found to be an inherently high-risk business being closely dependent to a country's national economy and operating in a fragile financial situation. It's considered to be the backbone of a country's economy, thus being it exposed to macroeconomic fluctuations and geopolitical events. Indeed, economic and political risks are found to be the most relevant sources of risk for the construction business. Additionally, construction firms usually operate with high leverage and carry out relatively long projects larger in value than their corporate assets; which makes their cash flows and financial health vulnerable. Therefore, risk management is identified as a critical dimension of the construction business and enhancing a firm's financial performance. In general, the risk management practice in the industry is found to be underdeveloped and reliant mostly on project manager's experience and intuition. This lack of proper risk management, combined with a vulnerable financial position and its exposure to global risk factors can convert the usual high leverage in unbearable losses, yielding the unusually high bankruptcy rate that plagues the industry.

Out of the initial 33 potential variables that resulted from the extensive literature research only 17 are finally introduced in the model. From the market and expectation variables the surprises for two, three and four past quarters were not found to be statistically significant and were removed from the model. Regarding the macroeconomic and geopolitical indicators considered, only the interest rates, the PPI for Construction Machinery, the public and private construction spending and the unemployment excess are found relevant and introduced in the model. The delay in reporting earning is found to be statistically significant and introduced in the model. When it comes to evaluating the financial performance of construction firms, out of the 21 different ratios proposed only six were selected as significant portraying each one of them one of the six financial dimensions considered. It's worth mentioning that the most important ratios seemed to be the ones related to profitability and valuation dimensions, and that the recursive feature elimination procedure based on a multinomial logistic regression only kept one ratio per dimension.

Five different state-of-the-art machine learning techniques were considered, including a linear model, 3 *tree ensemble* methods and a discriminative classifier. The time dimension of the data and generative process behind the observations are given due consideration, and a 5-fold cross-validation procedure with temporal splits and group aware is implemented to respect the structure in the data. Considerations regarding the data imbalance as a result of the nature of earnings' surprises, and the asymmetric classification costs derived from our desired investment strategy are made to appropriately select a performance metric that suits the modelling setting. The *macro-averaged* F-score with beta=0.5 is selected, to give more weight to the minority classes, the earnings' surprises, and prioritize precision over recall to account for the investment decisions triggered by each predicted category. The multinomial logistic regression is found to be a good compromise between precision and recall, with a reasonable performance across the cross-validation folds. When tested in the test set, the multinomial logistic model exhibits a bias towards returning positive surprises, failing to identify negative ones.

The most important variables are found to be the profitability and valuation metrics, with the company size. The relationship between earnings' surprises and the company size is well known in the literature and by the financial community, being often explained in terms of inattention, or understood as a liquidity premium [9]. The relevance of profitability ratios when it comes to assessing construction companies' financial performance had also been previously identified in the literature. In fact, is said to be the most important criteria when it comes to the construction industry, since its closely related to earnings' quality and positively correlated with dividend distributions [19], which is consistent with this research findings. Earnings' surprises have been defined in terms of divergences between the reported and expected earnings per share (EPS) measure, and the model identified as the most relevant feature the profitability dimension which is a proxy of earnings' quality. Valuation measures are also closely related with a firms' market performance, giving investors an idea of a firms' current valuation with respect of its peers; hence its relevance for predicting earnings' surprises is to be expected.

To conclude, the proposed investment strategy is back tested over a 4-year time span being its performance mainly driven by the positive surprise portfolio. The negative surprise portfolio fails to take advantage of the negative surprises, as a result of the model poor classification capabilities in this category. Nonetheless, the long-short portfolio doesn't seem to be harmed by the misclassification of surprises in opposite directions, achieving cumulative abnormal returns for the first three years of the back-test sample. This decrease in performance over time can probably be explained by the change in the relationship among variables over time and alleviated by recursive training of the model.

6.2. Contributions of this work

To the best of the author's knowledge, there is no previous academic work or live systematic investment strategy designed specifically for the construction industry stocks. Most of the currently available systematic investment strategies are based on theoretical investment factors, like momentum, or in alternative investment strategies, such as merger arbitrage; but are not industry specific. In that sense leveraging industry specific knowledge, and a deep understanding of the companies' operations, to enhance the predictive capabilities of a model-based investment strategy is novel.

Moreover, in this work the model evaluation metric has been customized to serve the specific business purpose we were considering: correctly classifying earnings surprises. The modelling setting in which the multi-classification problem lays is particularly challenging as a result of class imbalance and asymmetric misclassification costs. Precision, recall and F-score metrics of the information retrieval field are selected over more common metrics in the machine learning arena, such as the AUC or accuracy. In order to design an evaluation metric, more importance was given to precision over recall to fit the investment decision cost asymmetries, by selecting the F-score beta parameter as 0.5; which translates in giving precision twice the importance of recall. Additionally, to deal with the scarcity of relevant past observations, the macro-average across classes was selected biasing the measure towards the minority classes.

6.3. Assessment of the project goals

The goals set for this research project, which serves as the culmination of the MSc in Civil Engineering, having been successfully achieved. The extensive literature research endured to identify the main risk sources for construction stocks, and understand civil engineering operations, have provided the author with a deep understanding of the construction business. Moreover, the process of finding a suitable investment opportunity for a systematic investment strategy has provided the perfect setting for gaining more knowledge about the investment world and getting introduced in the behavioral finance literature. Finally, the completion of this work is by itself prove of having successfully dealt with all the hurdles that involve carrying out an end-to-end data science project; which included implementing five different cutting-edge machine learning techniques.

6.4. Further research

During the development of this research work the author encountered several research questions that despite being considered of great interest where found to be out of the scope of this work. Nonetheless, they are listed below and might be of interest for future research endeavors:

- There are many different definitions of earnings surprises available in the literature, being the one used in this research work from a mathematical perspective a normalized version of the forecast error for the EPS metric, and in a more conceptual way a proxy for a significant divergence between reported earnings' and market expectations. The question that arises is if a different definition of the target variable will yield better results, and how the feature importance across variables will vary under different definitions. Given that the end goal is predicting abnormal returns, a natural candidate to start with will be the CAR in a small window around the earnings' announcement day, aiming to directly predict the price jump.
- From the identified sources of risk for construction companies there is a subset of information that has the potential for having great predictive power: project specific information. The main hurdle regarding this information is its

accessibility and processing. An interesting way of continuing this work will be to systematically harvest this information by web scrapping companies' websites and trying to generate investment signals from them. Another possible source of information could be governmental information regarding public project characteristics disclosed in the tender process and the final assignations.

- The current research was restricted to a single geographic area given data availability restrictions. An interesting extension of this work will consist in testing the developed model across different regions and identifying variable importance variations across geographical areas. Is in the cross-country context that political risk indices as the ones developed by the World Bank will gain relevance and may be able to control for political instability affecting construction projects.

References

[1] E. Fama, "Efficient Capital Markets: A Review of Theory and Empirical work," *The Journal of Finance,* vol. 25, no. 2, pp. 383-417, 1970.

[2] A. Shleifer, Inefficient markets: An introduction to behavioral finance, 2000.

[3] A. Landier, *Behavioral Finance (course slides),* HEC Paris, 2018.

[4] J.-P. Bouchard, P. Krüger, A. Landier and D. Thesmar, "Sticky Expectations and the Profitability Anomaly," 2016.

[5] U.S. Securities and Exchange Commission, "Public Companies," [Online]. Available: https://www.investor.gov/introduction-investing/basics/how-market-works/public-companies.

[6] U.S. Securities and Exchange Commission, "Form 10-Q," 2 September 2011. [Online]. Available: https://www.sec.gov/fast-answers/answersform10qhtm.html.

[7] R. H. Thaler, Advances in behavioral finance, 1993.

[8] V. L. Bernand and J. K. Thomas, "Evidence that stock prices do not fully reflect the implications of current earnings for future earnings," *Journal of Accounting and Economics,* vol. 13, pp. 305-340, 1990.

[9] V. Dhar and D. Chou, "A comparision of Nonlinear Methods for Predicting Earnings Surprises and Returns," *IEEE Transactions on Neural Networks,* vol. 12, no. 4, 2001.

[10] E. Brea Garcia, "Using Machine Learning to predict Earnings surprises from Analysts' expectations," HEC Paris, 2019.

[11] McKinsey Global Institute, "Reinventing construction: a route to higher productivity," 2017.

[12] Deloitte, "2019 Engineering and Construction industry outlook," 2019.

[13] T. Apostola, G. Aretoulis, P. Papaioannou and G. Kalfakakou, "Performance Analysis of Construction Enterprises using Financial Ratios' groupings: An application in the British Construction Industry," *Seventh International Conference on Construction in the 21st Century (CITC-VII),* 2013.

[14] I. Horta and A. Camanho, "Competitive positioning and performance assessment in the construction industry," *Expert Systems with Applications,* vol. 41, no. 4, pp. 974-983, 2014.

[15] D. Ashley and J. Bonner, "Political Risks in International Construction," *Journal of Construction Engineering and Management,* vol. 113, no. 3, pp. 447-467, 1987.

[16] M. Kucukvar and O. Tatari, "Towards a triple bottom-line sustainability assessment of the U.S. construction industry," *Int J Life Cycle Assess,* vol. 18, p. 958–972, 2013.

[17] S. Kassim and N. H. Noordin, "Corporate Governance and Financial Performance: Empirical Evidence from Public Listed Construction Companies in Malaysia," 2015.

[18] S. Kim, S. Lee and J. Kim, "Relationship between the financial crisis of Korean construction firms and macroeconomic fluctuations," *Engineering, Construction and Architectural Management,* vol. 18, no. 4, pp. 407-422, 2011.

[19] M. Balatbat, C.-y. Lin and D. Carmichael, "Comparative performance of publicly listed construction companies: Australian evidence," *Construction Management and Economics,* vol. 28, no. 9, pp. 919-932, 2010.

[20] J.-H. Chen, "Developing SFNN models to predict financial distress of construction companies," *Expert Systems with Applications,* vol. 39, no. 1, pp. 823-827, 2012.

[21] H. Choi, H. Son and C. Kim, "Predicting financial distress of contractors in the construction industry using ensemble learning," *Expert Systems With Applications,* vol. 110, pp. 1-10, 2018.

[22] D. Baloi and A. Price, "Modelling global risk factors affecting construction cost performance," *International Journal of Project Management*, vol. 21, pp. 261-269, 2003.

[23] A. Akintoye and M. MacLeod, "Risk analysis and management in construction," *International Journal of Project Management*, vol. 15, no. 1, pp. 31-38, 1997.

[24] N. N. Taleb, The Black Swan, New York: Random House, 2007.

[25] R. Kangari, F. Farid and H. Elgharib, "Financial performance analysis for Construction Industry," *Journal of Construction Engineering and Mana,* vol. 118, no. 2, pp. 349-361, 1992.

[26] WRDS Research Team, "WRDS Industry Financial Ratio Manual," August 2016. [Online]. Available: https://wrds-www.wharton.upenn.edu/.

[27] Investopedia, "Capitalization Ratios," 10 October 2019. [Online]. Available: https://www.investopedia.com/terms/c/capitalization-ratios.asp.

[28] Investopedia, "Efficiency Ratio Definition," 21 May 2019. [Online]. Available: https://www.investopedia.com/terms/e/efficiencyratio.asp.

[29] Investopedia, "Analyzing Investments With Solvency Ratios," 25 June 2019. [Online]. Available: https://www.investopedia.com/articles/investing/101613/analyzing-investments-solvency-ratios.asp.

[30] Investopedia, "Liquidity Ratio Definition," 13 May 2019. [Online]. Available: https://www.investopedia.com/terms/l/liquidityratios.asp.

[31] DST Systems Inc, "Company valuation ratios," 2013. [Online]. Available: https://www.fidelity.com/learning-center/trading-investing/fundamental-analysis/company-valuation-ratios.

[32] El Pais, "Argentina expropia a Repsol su filial YPF," 17 April 2012. [Online]. Available: https://elpais.com/economia/2012/04/16/actualidad/1334590509_507539.html.

[33] El Pais, "El aeropuerto de Denver rescinde el contrato a Ferrovial," 14 August 2019. [Online]. Available: https://elpais.com/economia/2019/08/14/actualidad/1565785650_043440.html.

[34] The World Bank - Development Data Group, "World Governance Indicators," [Online]. Available: https://info.worldbank.org/governance/wgi/Home/Documents#doc-intro.

[35] Wikipedia, "Standard Industrial Classification," [Online]. Available: https://en.wikipedia.org/wiki/Standard_Industrial_Classification.

[36] WRDS, "Linking IBES and CRSP data (python)," [Online]. Available: https://wrds-www.wharton.upenn.edu/pages/support/applications/linking-databases/linking-ibes-and-crsp-data-python/.

[37] WRDS, "WRDS Overview of IBES," [Online]. Available: https://wrds-www.wharton.upenn.edu/pages/support/manuals-and-overviews/i-b-e-s/ibes-estimates/general/wrds-overview-ibes/.

[38] Federal Reserve Bank of St. Louis - Economic Research Department, "About Economic Research at the St. Louis Fed," [Online]. Available: https://research.stlouisfed.org/about.html.

[39] Wikipedia, "Federal Reserve Economic Data," [Online]. Available: https://en.wikipedia.org/wiki/Federal_Reserve_Economic_Data#cite_note-1.

[40] The World Bank, "Who we are," [Online]. Available: https://www.worldbank.org/en/who-we-are.

[41] The World Bank - Development Data Group, "About us," [Online]. Available: https://data.worldbank.org/about.

[42] WRDS, "Overview of I/B/E/S Historical Earnings Estimate Data," [Online]. Available: https://wrds-www.wharton.upenn.edu/pages/support/data-overview/wrds-overview-ibes-historical-earnings-estimate-database/.

[43] U.S. Bureau of Labor Statistics, "Producer Price Indexes," [Online]. Available: https://www.bls.gov/ppi/.

[44] U.S. Census Bureau, "Construction Spending," [Online]. Available: https://www.census.gov/construction/c30/c30index.html.

[45] The World Bank - Development Data Group, "Political Stability and Absence of Violence/Terrorism," [Online]. Available: https://info.worldbank.org/governance/wgi/Home/downLoadFile?fileName=pv.pdf.

[46] S. Balcaen and H. Ooghe, "35 years of studies on business failure: an overview," *The British Accounting Review of the classic statistical methodologies and their related problems,* vol. 38, pp. 63-93, 2006.

[47] Wikipedia, "Feature Selection," [Online]. Available: https://en.wikipedia.org/wiki/Feature_selection.

[48] Wikipedia, "Analysis of variance," [Online]. Available: https://en.wikipedia.org/wiki/Analysis_of_variance#Textbook_analysis_using_a_normal_distribution.

[49] R. Lowry, "Concepts & Applications of Inferential Statistics," 1999. [Online]. Available: http://vassarstats.net/textbook/index.html.

[50] Wikipedia, "Kruskal–Wallis one-way analysis of variance," [Online]. Available: https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance.

[51] Wikipedia, "Pearson's chi-squared test," [Online]. Available: https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test.

[52] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[53] P. E. Latham and Y. Roudi, "Mutual Information," *Scholarpedia,* vol. 4, no. 1, p. 1658, 2009.

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[55] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning, Springer, 2009.

[56] N. Moniz, P. Branco and L. Torgo, "Resampling strategies for imbalanced time series forecasting," *International Journal of Data Science and Analytics,* vol. 3, pp. 161-181, 2017.

[57] N. Chawla, W. K. Bowyer, O. L. Hall and w. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[58] Scikit-learn developers, "Model evaluation: quantifying the quality of predictions," [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html.

[59] Wikipedia, "Multinomail Logistic Regression," [Online]. Available: https://en.wikipedia.org/wiki/Multinomial_logistic_regression.

[60] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," 2016.

[61] Wikipedia, "Support-vector machine," [Online]. Available: https://en.wikipedia.org/wiki/Support-vector_machine.

[62] M. López de Prado, Advances in Financial Machine Learning, Wiley, 2018.

[63] J. Livnat and R. R. Mendenhall, "Comparing the Post-Earnings Announcement Drift for Surprises Calculated from Analyst and Time Series Forecasts," *Journal of Accounting Research,* vol. 44, no. 1, pp. 177-205, 2006.

[64] Thompson Reuters, *I/B/E/S Detail History User Guide,* 2016.

[65] D. Glushkov, "Post-Earnings Announcement Drift," February 2008. [Online]. Available: https://wrds-www.wharton.upenn.edu/pages/support/applications/portfolio-construction-and-market-anomalies/post-earnings-announcement-drift/.

[66] WRDS, "Calculating the Number of Analysts Making Earnings Forecasts for Each Firm," [Online]. Available: https://wrds-support.wharton.upenn.edu/hc/en-us/articles/115003391112-Calculating-the-Number-of-Analysts-Making-Earnings-Forecasts-for-Each-Firm.

[67] Investopedia, "Book Value," 1 Oct 2019. [Online]. Available: https://www.investopedia.com/terms/b/bookvalue.asp.

[68] Wikipedia, "Compustat," [Online]. Available: https://en.wikipedia.org/wiki/Compustat.

[69] Investopedia, "Extraordinary Items," 22 May 2019. [Online]. Available: https://www.investopedia.com/terms/e/extraordinaryitem.asp.

[70] Wikipedia, "F-test," [Online]. Available: https://en.wikipedia.org/wiki/F-test.

[71] Wikipedia, "Gross National Income," [Online]. Available: https://en.wikipedia.org/wiki/Gross_national_income#cite_note-1.

[72] Wikipedia, "Mutual information," [Online]. Available: https://en.wikipedia.org/wiki/Mutual_information#Definition.

[73] H. Zhang, F. Yang, Y. Li and H. Li, "Predicting profitability of listed construction companies based on principal component analysis and support vector machine—Evidence from China," *Automation in Construction,* vol. 53, pp. 22-28, 2015.