FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ECONOMIA DE SÃO PAULO

ANDRÉ BINA POSSATTO

PAINTING THE BLACK BOX WHITE: A FUNDAMENTALIST-BASED TRADING STRATEGY USING INTERPRETABLE TREES

SÃO PAULO - SP
2020

ANDRÉ BINA POSSATTO

PAINTING THE BLACK BOX WHITE: A FUNDAMENTALIST-BASED TRADING STRATEGY USING INTERPRETABLE TREES

Dissertação apresentada à Escola de Economia de São Paulo da Fundação Getulio Vargas, como requisito para obtenção do Título de Mestre em Economia.

Área de concentração: Investimentos

Orientador: Prof. Marcelo Fernandes

SÃO PAULO - SP

2020

Possatto, André Bina.

Painting the black box white : a fundamentalist-based trading strategy using interpretable classification trees / André Bina Possatto. - 2020.

96 f.

Orientador: Marcelo Fernandes.

Dissertação (mestrado profissional MPFE) – Fundação Getulio Vargas, Escola de Economia de São Paulo.

1. Aprendizado do computador. 2. Ações (Finanças) - Preços - Previsão. 3. Investimentos - Análise. 4. Mercado financeiro - Brasil. I. Fernandes, Marcelo. II. Dissertação (mestrado profissional MPFE) – Escola de Economia de São Paulo. III. Fundação Getulio Vargas. IV. Título.

CDU 336.76::007.52(81)

Ficha Catalográfica elaborada por: Isabele Oliveira dos Santos Garcia CRB SP-010191/O
Biblioteca Karl A. Boedecker da Fundação Getulio Vargas - SP

ANDRÉ BINA POSSATTO

PAINTING THE BLACK BOX WHITE: A FUNDAMENTALIST-BASED TRADING STRATEGY USING INTERPRETABLE TREES

Dissertação apresentada à Escola de Economia de São Paulo da Fundação Getulio Vargas, como requisito para obtenção do Título de Mestre em Economia.

Área de Concentração: Investimentos

Data de Aprovação: 16/06/2020

Banca Examinadora:

_____

Prof. Marcelo Fernandes (Orientador)
FGV-EESP

_____

Prof. Ricardo Pereira Masini
FGV-EESP

_____

Prof. Alan de Genaro Dario
FGV-EAESP

SÃO PAULO - SP

2020

This work is dedicated to all the people
who found out that waiting for the future
is not enough.

ACKNOWLEDGEMENTS

First I would like to thank my dissertation supervisor Professor Marcelo Fernandes. Your help, guidance and coaching were paramount to this work's conclusion, helping shape raw ideas into a full dissertation. I extend my acknowledgement to all FGV-EESP professors with whom I had contact while pursuing my master's degree. Thank you for the generosity in sharing your knowledge.

I would also like thank all my colleagues in the different institutions in which I had the pleasure to study throughout my life. Although of different nations, backgrounds, beliefs and opinions I owe you for much of the development of my reasoning, especially when you proved me wrong. Likewise, I thank all my colleagues in the great companies I have worked and particularly at Cielo, where I have been working throughout my master's degree. The understanding and support of my team was essential to get this work done.

I am also grateful to all my friends for being there when I needed and for understanding when I could not be present due to the efforts necessary to obtain my degree. Without you, none of this would have any meaning.

I continue by thanking my family, whose encouragement was of utmost importance to start and finish this course. May this work be a testament of all your effort in raising me and teaching the ways of life.

Finally, I finish by thanking my girlfriend Heloisa, who was able to endure the countless times that I couldn't be present and has nonetheless helped me in these years of early mornings, late nights and weekends spent working and studying without break.

The past is the past
and is here to stay
(Nick Cave, We Real Cool)

ABSTRACT

Difficulty understanding how a black box model makes predictions has undermined machine learning's success in financial markets, according to a recent article from Bloomberg (2019b). Our work shows how model-agnostic methods to interpret machine learning predictions turn these models more transparent to a human investor. We benchmark three tree-based algorithms between themselves, creating long-short investment strategies with independent models for each leg and using only fundamentalist analysis. We then apply the models to the Brazilian stock market (Bovespa) and achieve an out-of-sample expected annual return of 26.4% with a Sharpe ratio of 0.50. Ensembles between the long and short legs improve this result for a Sharpe ratio of up to 1.26, comparable to other works in the literature reported by Avramov *et al.* (2019) when considering real-world constraints. Our strategy has low asset turnover and transaction costs do not explain the results. All models achieve positive risk premiums and two are statistically significant. Interpretation shows differences on the key predictors for over- and underperformance, with the first focusing on price-to-value and the second on size and liquidity. Local interpretation is discussed in the case of Magazine Luiza, showing how model explanation helps an investor to understand and decide which stocks to buy or sell based on the models' output. We argue that different performance and interpretation between long and short models and the possibility of ensembling are key advantages of modeling these positions separately.

Keywords: Machine Learning. Stock returns prediction. Model interpretation. Long-short trading strategy

RESUMO

A dificuldade em entender como modelos de aprendizado de máquina "caixa preta"fazem previsões prejudica sua adoção no mercado financeiro, de acodo com um artigo recente da Bloomberg (2019b). Nosso trabalho discute como métodos de interpretação agnósticos são capazes de tornar os modelos de aprendizado de máquina mais transparentes para investidores humanos. Nós comparamos três algoritmos baseados em árvores na geração de estratégias de investimento *long-short*, criando modelos independentes para as pernas de compra ("*long*") e venda ("*short*") e usando apenas variáveis fundamentalistas. Nós aplicamos esses modelos para o mercado de ações brasileiro (Bovespa) e atingimos um resultado anualizado esperado de 26.4% fora da amostra, com um índice de Sharpe de 0.50. Combinações[1] entre as pernas de compra e venda melhoram o resultado, atingindo um índice de Sharpe de até 1.26, comparável com os resultados de Avramov *et al.* (2019) quando restrições econômicas são levadas em conta. Nossa estratégia tem baixo nível de substituição de ativos ao longo do tempo e custos transacionais não explicam o resultado. Todos os modelos conseguem prêmios de risco positivos e dois são estatisticamente significativos. A interpretação dos modelos mostra diferenças nos preditores entre performances acima e abaixo do esperado, onde no primeiro caso prevalecem razões (*"ratios"*) entre preço e valor, e no segundo liquidez. A interpretação local dos modelos é discutida através do exemplo da Magazine Luiza e mostra como as explicações obtidas ajudam um investidor a entender e decidir pela compra ou venda do ativo de acordo com o resultado do modelo. Nós argumentamos que diferenças tanto na performance quanto na interpretação dos preditores e a possibilidade de combinações são vantagens de se modelar separadamente compra e venda.

Palavras-chave: Aprendizado de máquina. Previsão de retorno de ações. Interpretação de modelos. Estratégias long-short.

[1] Ensembles na literatura em inglês.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

APR	Annual Percentage Rate
AUC	Area Under Curve
CAPM	Capital Asset Pricing Model
CB	CatBoost
D&A	Depreciation and Amortization
DL	Deep Learning
DT	Decision Tree
EAR	Effective Annual Rate
EBITDA	Earnings Before Interest, Taxes, Depreciation and Amortization
EBT	Earnings Before Taxes
EMT	Efficient Market Theory
FA	Fundamentalist Analysis
IQR	Interquartile Range
KS	Kolmogorov-Smirnov
LASSO	Least Absolute Shrinkage Selection Operator
LG	LightGBM
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
NN	Neural Network
OLS	Ordinary Least Squares
OM	Order of Magnitude
PCA	Principal Component Analysis

P/E	Price/Earnings
P/EBITDA	Price/EBITDA
ROC	Receiver Operation Characteristic
ROE	Return On Equity
SR	Sharpe Ratio
SDF	Stochastic Discount Factor
SVM	Support Vector Machine
TA	Technical Analysis

CONTENTS

1 INTRODUCTION

Machine learning (ML) is one of the undisputed trends of recent years, with web searches for this term having more than tripled in the last five years, according to Google Trends (2020). Machine learning, a major subset of the field of artificial intelligence (AI), can be broadly described as the ability of a computer to learn from data, predict behavior, and ultimately act on it. If data is "the new oil"[1] machine learning could be called the "new combustion engine," with great potential to accelerate and change many industries.

Finance is a natural field in which to apply machine learning. It has large amounts of data (think of stock-market quotes changing in real time in a typical Wall-Street movie) and excellent incentives to pursue any technique that may help achieve higher financial returns.

Despite ML's promises of higher gains and a high adoption among hedge funds,[2] a recent article by Bloomberg (2019b) suggests that the actual results of using ML are falling short of expectations. Bloomberg's warning is not isolated, as a similar article in the Wall Street Journal (2020) proves. And while both articles suggest many reasons why these expectations are not being met,[3] one common issue is the black-box nature of machine learning algorithms. This issue generates a lack of trust among investors, thus impeding adoption of machine-learning algorithms. As Bloomberg (2019b) puts it: "If firms fail to fully understand what effect their algorithm is capturing, they may not know when to switch it off."

While the aforementioned news pieces from Bloomberg and the Wall Street Journal give a sense of the relevance of the trust problem, it also resonates in the ML literature. Gilpin *et al.* (2018) conclude that while there has been a surge of research on explainable AI, existing models are still inadequate. Biran *et al.* (2017) show the potential of this approach in finance. They demonstrate how providing human-centered explanations improves the performance of an investor picking stocks in an experimental setting.

To predict stock performance using fundamentalist analysis, we propose the using model-agnostic interpretation methods to tackle the "trust problem." These methods help an investor understand and compare how each model predicts stock results, both at a global (model as a whole) and local level (individual predictions). We make interpretations over a long-short trading strategy constructed by independently modeling the long and short

[1] This expression is now widespread in the field and is usually attributed to Clive Humby.
[2] 58% according to Bloomberg (2019b).
[3] Most of them are well discussed in the work of Avramov *et al.* (2019)

legs.

We compare three tree-based algorithms. The first is the simplest Decision Tree, an algorithm built by successively splitting the dataset in a hierarchical structure. The other two algorithms are boosted trees, meaning that they ensemble multiple trees to improve accuracy, focusing on the errors of the prior interaction. The comparison shows the potential of the two boosting tree algorithms (LightGBM and CatBoost) over the Decision Tree, which is inherently interpretable but has poor performance.

We separate long and short models instead of creating a single model and selecting the higher scores for the long model and the lower for the short model.[4] We choose this path both as a strategy to require a minimum excess return but also from a suspicion that the forces driving overperformance are different from the ones driving underperformance. This combination between three algorithms with two legs each adds to a total of six models.

We use only fundamentalist data present in the quarterly results releases of the traded companies. The six models are then trained and tested in out-of-sample data separated by a time-split, avoiding overfitting and reproducing a real-setting investment. Results are reported for the base models, the long-short strategies derived from them, and for combinations between models (known as ensembles). We test two different ensembling strategies: between different algorithms and combining the long and short models of the same algorithm.

The global interpretation of the three best models is discussed using Shapley values (the other three are also briefly discussed in the appendixes). This approach allows for the comparison both between long and short models and two different algorithms for the same objective (overperformance), showing the model-agnostic properties of this technique.

The local interpretation is then applied to the case of Magazine Luiza's performance based on its 2017Q3 results. This particular case is chosen because at the time the stock of this retailer had attained an impressive surge and there was excitement in the Brazilian stock market as to how high it could still go, or if it faced risk of a sharp fall. This situation provides fruitful ground to understand how the models weight these factors and generate a recommendation to buy, sell or do nothing with this stock.

Testing the models in out-of-sample data, we achieve an expected annualized return (EAR) of 26.4% in the long-short strategy based on the LightGBM algorithm (LG). This

[4] One common strategy would be to take the top decile for the long model and the bottom decile for the short model. The work of Gu *et al.* (2018) is an example of this approach.

algorithm also shows the best short performance overall with 10.5% and a strong long performance with 14.6%, though the CatBoost algorithm (CB) beat it with 19.0%. On the other hand, CB failed in the short position with only 1.0% while the Decision Tree algorithm (DT) presented a poor performance in the long position and a negative performance in the short (10.0% and -4.5% respectively).

These returns obtained by LG and CB's long position are still economically relevant when trading costs are accounted for, with LG's long-short portfolio having 22.1% EAR. Taking risk into account, Sharpe ratios of up to 0.50 are obtained. Two of the models also achieve statistically significant intercepts in a five-factor model, while the market risk coefficients show that the algorithms differ in risk appetite, with DT having the lowest βs and CB the highest.

We implement combinations between algorithms and between long and short positions for the same algorithm. While none of the ensembles among algorithms is successful, combining the long and short CatBoost-based models achieves better results for both legs, with EARs of 31.7% and 17.8% for the over- and underperformance models, respectively (vs. 19.0% and 1.0% without the ensemble). A long-short portfolio based on CB ensembles achieves a Sharpe ratio of 1.26.

Global model interpretation using Shapley values shows that valuation metrics dominate as predictors for the long position. Price/earnings (P/E), price/EBITDA (P/EBITDA) and book-to-market value all appear at the top of the list with the expected direction.[5] An increase in valuation also contributes positively to overperforming.

The conflict between an already high valuation coupled with a positive trend in the valuation itself is further explored using Magazine Luiza's example. Applying local interpretation, we show that the model correctly predicts Magazine Luiza's overperformance due to its rise in valuation having been accompanied by a growth in both EBITDA and return on equity (ROE).

Global interpretation of the short leg shows strong contributions of size and current liquidity. Stocks of larger companies have a higher chance of underperforming, as do the stocks of companies with low or decreasing current liquidity. We use this striking difference between the key features of the long and short models as an argument in favor of modeling the two legs separately. The inferior performance of the short model (both in accuracy and financial return) reinforces this conclusion. Together they provide evidence of the need to treat the long and short models separately, and not simply as two sides of the same coin.

[5]The lower the better for the first two, and the higher the better for the last one.

We compare our results with other machine-learning-based works in light of the critique made by Avramov *et al.* (2019). These authors argue that works such as Gu *et al.* (2018) often exploit economically unfeasible positions and achieve far lower returns when faced with real-world restrictions. With these restrictions in mind, we consider only the stocks that comprise Brazil's main stock index, Ibovespa, thus limiting the financial return benefits of choosing stocks of distressed companies and eliminating microcaps altogether. We also make predictions based on the results sheet from each quarter to the next 2 months. This is a low-frequency strategy and therefore has lower transaction costs. Finally, the trust problem of these methods is tackled with model interpretation.

Gu *et al.* (2018) report achieving a monthly return of up to 2.26% with a long-short strategy, an EAR of almost 31%. After reproducing this result, Avramov *et al.* (2019) then exclude microcaps, stocks without credit rating coverage and distressed firms. They find that the signals reduce by 64%, 52% and 77%, respectively. In this scenario the 1.26 Sharpe ratio of our long-short portfolio based on CB ensembles is superior to all Gu, Kelly and Xu's scenarios with economic restrictions, while a long-short strategy with the best original models has a comparable Sharpe ratio.

The rest of this dissertation is divided as follows. Chapter 2 reviews the literature on predicting stocks using machine learning and model interpretation. Chapter 3 presents methodology and discusses the data, with a descriptive analysis of both features and target. Chapter 4 contains the financial returns brought by the trading strategy and discusses how each feature contributes to this result. Chapter 5 concludes.

2 LITERATURE REVIEW

This chapter is divided in four sections. The first gives a broad perspective on predicting stock returns. The second focuses on how machine learning accomplishes this task. The third section discusses the need for interpretability. It presents the model interpretation methods we apply and how they are used to predict returns. The final section summarizes the main contributions of this dissertation.

2.1 Predicting Stock Returns

While the motivation to predict stock returns is clear, there is ample evidence of the difficulty of accomplishing this task (see for example Tay (2001)). In fact, according to Malkiel and Fama (1970) it could be argued that under the Efficient Market Theory (EMT) in perfect conditions it should not even be possible to consistently predict returns. The reason is that the price system would self-correct and leave no room for arbitrage. Fama (1991) later showed how the addition of risk factors explains abnormal returns, but the EMT does make predicting future returns quite a challenge.

The problem of predicting stock returns is closely related to predicting asset value. The expectation that rational agents will bring prices close to a "fair" value drives prices and consequently returns. The contributions of Sharpe (1964), Lintner (1965) and Black (1972) are cornerstones of the return prediction literature. Fama and French (1992) show that the single-factor model can be enhanced when size and book-to-market equity ratio (BE/ME) are accounted for. The same authors expand their research in subsequent articles, which propose a five-factor model (FAMA; FRENCH, 1993) and discuss the relation between earnings and BE/ME (FAMA; FRENCH, 1995), respectively. Regression of stock returns over BE/ME has also been tested in Brazil with similar results, as shown by Araújo and Machado (2018).

Two classes of prediction techniques can be used alone or in combination to predict asset prices and stock returns: technical analysis (TA) and and fundamentalist analysis (FA). TA is exemplified by the work of Lahmiri (2014). It focuses on constructing indicators based on past price information, which may or may not have an economical rationale. FA, on the other hand, is backed by economic theory. It usually constructs indicators based on company data provided on balance sheets, combined with other indicators (CAVALCANTE et al., 2016). A good review of the pros and cons of each approach is presented by Atsalakis and Valavanis (2009).

Since the methods used for these predictions are diverse, it is not our goal to present an exhaustive list. But before diving into machine learning, it is worth noting at least two standard approaches: structural regressions and time series.

Models such as the Capital Asset Pricing Model (CAPM) (BLACK, 1972) or Fama's multi-factor model (FAMA; FRENCH, 1993) use standard econometric techniques. They often resort to different variations of regressions to test hypotheses regarding market structure and price determinants. These models usually have a greater focus on testing the hypothesis than the predictive power itself, which is also why they are limited to naturally-interpretable methods such as logistic regressions.

Time series are arguably the natural path for predicting financial data. A well-established literature exists on the use of time series in stock prediction (see Ariyo *et al.* (2014) and Mondal *et al.* (2014), for example). There are also great challenges when using these methods (TSAY, 2014), since financial data does not always behave well, usually showing heteroskedasticity, for example. While they are not the focus of this dissertation, it is worth mentioning the growing literature on the use of neural networks coupled with time series. For a detailed discussion of these models, refer to Zhang (2003), Azoff (1994) and Kaastra and Boyd (1996).

2.2 Applications of Machine Learning in Stock Return Prediction

The idea of using machine learning to predict stock returns has wide support in the literature with a range of different methods being used (HENRIQUE; SOBREIRO; KIMURA, 2019). Taking in account this diversity, this section will highlight key works in the field and the distinctive contributions of this work. At the end of the section a broader comparison with the rest of the literature is made using a quantitative approach.

Based on the work of Fama and French (1993) many authors have searched for different asset-pricing factors, leading to a factor "zoo" (FENG; GIGLIO; XIU, 2017). Feng, Polson and Xu (2018) show the potential of deep learning to tame this zoo. Deep learning (DL) is a term generally applied to neural networks with many hidden layers. Feng *et al.* (2018) advocate that the great advantage of these algorithms is their ability to tame large sets of data and combine different features in a way that a human would not think to specify in a standard ordinary least squares (OLS) regression. They apply DL in the construction of risk-factors in a way that considers potential non-linear interactions between features.

With a similar goal of dealing with large and sparse datasets with multiple factors,

Chinco *et al.* (2019) show the benefit of using the Least Absolute Shrinkage Selection Operator (LASSO) to reduce the feature space. LASSO allows the authors to use fewer data points to predict returns for the next minute based on unexpected and short-lived factors. Kozak *et al.* (2020), on the other hand, argue that the Stochastic Discount Factor (SDF) outperforms LASSO when trained with economically-motivated priors. Regardless of the technique, both works endorse the conclusion that machine-learning-based approaches have good performance due to their capacity of identifying non-linear interactions between variables.

Moving away from factor-based asset pricing, Sirignano (2019) builds a deep learning architecture to tackle prediction price in a novel away. He applies it not only to transaction price data but also to analyzing all the orders in the book for each stock ("limit order book"). With this approach he succesfully models the joint distribution between best bid and best ask prices. In a related work, Sirignano *et al.* (2019) also use deep learning to find evidence that a stable price formation mechanism exists within order books.

The use of machine learning as a empirically successful way to price assets and predict stock returns, which is part of the goal of this dissertation, is demonstrated by Gu, Kelly and Xiu (2018). These authors compare different algorithms and report an out-of-sample Sharpe ratio of 0.77 for a machine-learning-based strategy, compared to 0.51 for a traditional buy-and-hold investment. They also show that this ratio can be improved to 1.35 if a decile approach is used for a long-short strategy, compared to only 0.61 of OLS. Comparing different models and features, Gu *et al.* (2018) find that the best predictors of stock returns are price trends, liquidity and volatility. They also show that the superior results of machine learning probably come from finding non-linear interactions between variables that would be hard to specify in a parametric model. The algorithms that perform best are neural networks and, to a lesser extent, regression trees. One of their important findings is that "shallow" methods outperform "deep" methods, which they attribute to a high noise-to-signal ratio (GU; KELLY; XIU, 2018).

This dissertation builds on the work of Gu *et al.* (2018) by using a long-short strategy and trees-based models. Nonetheless, we make some key distinctions. We focus only on fundamentalist indicators, obtained from financial demonstrations.[1] We also use classification algorithms instead of regression, modeling the long and short legs separately. Finally, we model the Brazilian market instead of that of the United States (which poses

[1]It could be argued that the year-over-year (YoY) and quarter-over-quarter (QoQ) evolution of asset prices, included in this dissertation, could be considered technical indicators. Even though they do represent a price trend, they are also implied in many common fundamentalist indicators such as Price/Earnings and are easily available in financial demonstrations (the closing price of the Quarter), justifying their use in this dissertation.

additional challenges since far less historical data is available).

Chen, Pelger and Zhu (2019) develop different deep learning models based on the principles of Gu *et al.* (2018) and Kozak *et al.* (2020), but add macroeconomic variables and a no-arbitrage condition to the estimation of the SDF. They report performance superior to other deep-learning-based methods with a Sharpe ratio of up to 2.6 and attribute it to their innovations. They also stress their method's ability to deal both with high dimensionality and non-linear interactions as the main advantage of machine learning over traditional methods. Analyzing the weights on the SDF, the typical technical analysis variables short-term reversal, momentum and volume are the elements that best explain their predictions (CHEN; PELGER; ZHU, 2019). This dissertation differs from their work by its focus only on fundamentalist analysis and tree-based methods.

The use of classification methods is supported by Leung, Daounk and Chen (2000). These authors compare multiple regression and classification approaches and show that classification scores with threshold trading rules outperform level prediction. Though more modern approaches have been developed since then, their conclusions are still valuable as a validation of the potential of classification scores in producing profitability. Two important distinctions between this dissertation and Leung *et al.* (2000) are the prediction of individual stocks rather than market index and the use of tree-based methods of classification.

More recently, Ballings *et al.* (2015) also approach the problem of predicting stock returns with classification methods. They perform a benchmark between single-classifier and ensemble methods for the European market. They test three ensembles (Ada-Boost, Kernel Factory and Random Forests) and conclude that ensemble methods, though uncommon in the literature, perform better than single-classifier methods such as Support Vector Machine (SVM) and neural networks. Random Forest has the best result overall in terms of Area Under (the Receiver Operating Characteristic) Curve (AUC or ROC-AUC), further encouraging the use of tree-based methods. Even though this dissertation shares some characteristics with Ballings *et al.* (2015), such as the use of FA and classification methods, it differentiates itself not only by the market addressed, but also by the use of gradient-boosted trees; validation with a time split of the data; and the use of a long-short trading strategy based on classification percentiles.

Moritz and Zimmermann (2016) use tree-based algorithms to create portfolios and develop a long-short strategy based on top and bottom percentiles. While the trading strategy is similar, this dissertation uses only fundamentalist indicators while Moritz and Zimmermann use different accounts of past returns.

Even though all machine-learning-based methods reviewed report results better than traditional approaches, a shadow of doubt is cast by Avramov, Cheng and Metzker (2019). These authors argue that despite superior results found by machine-learning strategies, many times they are economically unfeasible and tend to focus on illiquid stocks, small-caps and distressed companies, while also using strategies with high turnover (and consequently trading costs). After reproducing the strategies of other authors (including those of Gu *et al.* (2018) and Chen *et al.* (2019)) they show that the superior results vanish in the face of economic restrictions, with short legs having the worst deterioration. The authors conclude that machine learning is still profitable for the algorithms tested in long positions and state that "even though machine learning offers unprecedented opportunities to improve investment outcomes, it is important to consider the common economic restrictions in assessing the success of a newly developed machine learning method" (AVRAMOV; CHENG; METZKER, 2019).

Based on the conclusions of Avramov *et al.* (2019), another key distinction of this dissertation compared to other works is the limitation only to equities that compose the main stock index (in this case, the Ibovespa index). These assets are considered "blue chips," representing large companies with liquid stocks. Thus we avoid the pitfalls reported by these authors.

Another branch of research using machine learning in stock prediction is related to natural language processing (NLP) applied to relevant textual information. Si *et al.* (2014), for example, use semantic stock networks based on published news articles to predict changes in stock prices. Al Nasseri, Tucker and Cesare (2015), on the other hand, use decision trees and sentiment analysis on financial forums postings. They obtain evidence that these posts present predictive power over the actual behavior of prices. Oliveira, Cortez and Areal (2017) reach similar conclusions with Twitter data and a regression framework. Although this dissertation does not use textual analysis, it remains a possible path for future research. One could, for example, analyze administration communications from both financial statements and relevant facts as predictors of future stock behavior.

Applying machine-learning methods to the Brazilian stock market also has support from the literature. Raposo and Cruz (2002), for example, use fuzzy neural networks to tackle stock buy-hold-sell recommendations. They show that these models work better with distributed classes and propose a validated set of fundamentalist features, two contributions that will be exploited in this article. Another relevant contribution comes from Laboissiere, Fernandes and Lage (2015) who predict daily maximum and minimum prices, developing a strategy for intra-day trading.

Finally, Henrique, Sobreiro and Kimura (2019) perform an extensive literature review

and give a quantitative view of the main publications on this topic and their respective classification on key aspects of the modeling approach. The preferred choices for each aspect are summarized in Table 1.

Table 1 also compares the most common choices in the literature with those of this

Table 1 – Key characteristics of machine learning models of stock performance

Characteristic	**Preferred approach**	**Choice for this dissertation**
Market (Country)	USA (47%) and Taiwan (17%)	Brazil
Asset	Index (60%)	Individual Stocks
Predictors	Technical Analysis (37%)	Fundamentalist Analysis
Methods	Neural Networks (70%)	Classification Trees
Prediction	Direction (42%)	Direction
Metrics	Depends on Prediction	Accuracy

Characterization of past research on machine learning for stock return prediction and comparison with present work.
Source: Elaborated by the author based on Henrique *et al.* (2019)

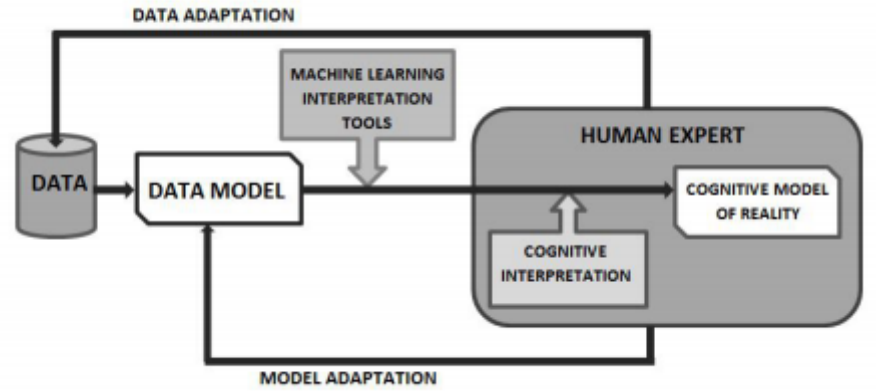
dissertation, yielding interesting insights. First, while 47% of the articles focus on the US stock market, the Brazilian market is studied in less than 4% of the articles. Also differing from the norm, we predict individual stocks rather than indexes, which are studied in 60% of the works. Regarding the predictors (features) used, TA is the preferred source of 37% of articles, while FA is second with 26%; the rest are divided between past prices and returns, which could also be considered a form of TA. The most common method used is neural networks, present in 70% of the works (usually more than one method is used). Our approach instead focuses on tree-based methods. Finally, both this dissertation and most publications in the literature favor the prediction of direction of change and not level, treating the prediction as a classification rather than a regression problem.

2.3 Model Interpretability

Despite the potential benefits of using machine learning models in finance, one common criticism is that their "black box" approach stands as a barrier to their adoption. Leading management consulting firm McKinsey & Company, for example, cites the lack of model "explainability" as a practical challenge to the adoption of AI (MANYIKA; BUGHIN, 2018).

The challenges of interpreting machine-learning models have also been discussed in the literature. After a broad overview of the subject of explanatory artificial intelligence, Gilpin *et al.* (2018) conclude that desite a surge of research on the topic, current methods are inadequate.

Figure 1 – Human-Centered Model Adaptation Process



Source: Vellido *et al.* (2012)

Despite its importance, interpretability does not have a universally-shared definition. Lipton (2016) discusses this issue thoroughly and presents many aspects of interpretability, while distinguishing between two classes of characteristics to help understand the concept. The first class is desired outcome, which can be be broken into categories of trust, causality, transferability, informativeness, and fair and ethical decision making. The second broad class is transparency, divided into simulatability, decomposability, algorithmic transparency, post hoc interpretability, text explanations, visualization, local explanations and explanation by example.

While the exact notion of interpretability may be elusive and context-dependent, one clear distinction between an explainable model and a pure "black box" model is that the former is inherently preoccupied with a decision-making actor to whom it is accountable and therefore "deserves explanation" while the second approach could be understood as focused on "results only." In that sense, Vellido, Martín-Guerrero and Lisboa (2012) stress the cognitive factor in their discussion of interpretability. This becomes evident in their representation of the model adjustment process. After emphasizing the human component, these authors go so far as to say that these models are "powerless" if unable to be understood.

Figure 1, reproduced from Vellido *et al.* (2012), shows the role of humans in cognitively interpretating the model and adjusting the model and the data to better suit their needs. Their representation gives model interpretation the role of an interface that allows the black box model to be comprehended by the "human expert."

The relevance of explainability in financial models is discussed by Biran and McKeown (2017). These authors propose the creation of human-centered narratives based on the predictions of the model. They test if these narratives help investors achieve better

results when the investors are asked if they would invest in different stocks of the S&P 500. They conclude that when given textual or graphical information based on the results of a machine learning model, the human decision-makers perform better in terms of returns and accuracy. This dissertation draws upon the work of Biran and McKeown (2017) as a key motivator. It builds on their work by focusing on fundamentalist analysis (they include TA features), using more complex models (they use logistic regression, which is already inherently interpretable) and formulating a trading strategy that comprises both long and short positions.

Given the need to understand the "reasoning" of the models, some of the works reviewed in the previous section also evaluate the importance of features in some way, though always in the scope of the whole model and not to explain an individual prediction. Gu *et al.* (2018), for example, measure both the impact on R^2 of setting the variable to zero and the squared partial derivatives. Moritz and Zimmerman (2016) follow a similar approach and measure the increase in mean squared error (MSE) when permutations are performed between feature values. Chen *et al.* (2019) use partial derivatives on weights of the stochastic discount function to create a sensitivity variable from which they interpret feature importance.

While these works include general feature importance evaluation, this dissertation focuses on two model-agnostic methods of interpreting machine learning models: LIME and Shapley Values. These methods are chosen because together they can explain a model on both a general and a local level, while also being readily applicable to any classifier. Both methods can also be applied for both long and short positions, allowing for a comparison that shows whether the justifications to buy a stock are the exact opposites of those used to sell, or if different factors apply to each decision.

LIME was proposed by Ribeiro, Singh and Guestrin (2016). It has the benefit of explaining "the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model." In short, this means that it takes samples around the point being explained and fits a linear model that represents an interpretable approximation of the original model in that region. While fitting the linear model, it weights the samples according to their distance to the observation being interpreted, allowing for an assessment of the feature's importance that is locally faithful.

Shapley Values, on the other hand, actually come from coalitional game theory presented by Shapley (1953). They have been adopted by Štrumbelj and Kononenko (2010) in the machine-learning context as a feature interpretation method. The idea is to give fair payoffs (importance) to the players (the features explaining the model), considering the collaborative manner in which the features interact to explain results.

The Shapley Values represent the difference between what would be predicted without a subset of features and what is actually predicted by then. All the different combinations must be assessed to account for the interactions, making the algorithm's complexity exponential. Štrumbelj and Kononenko (2010) solve this by proposing a sampling strategy that consistently estimates the Shapley Values without having to go through every combination. The model is interpreted by both the Shapley Value's average and it's distribution, which may show a non-linear effect.

2.4 Main Contributions

This dissertation differentiates itself from the existing international literature on predicting stock returns by offering a fundamentalist-based long-short trading strategy supported by top and bottom percentiles of separated classification tree scores. Another key distinction is the limitation to only those equities that compose the main stock index (in this case, the Ibovespa index), avoiding the pitfalls of other works reported by Avramov *et al.* (2019).

It is also novel in applying this general approach to Brazil, a developing country with a young stock market, compared to the United States. Brazil provides therefore far less data to train and validate such models, and is subject to stronger internal and international fluctuations. Applying machine learning in Brazil, Russia, India, China and South Africa (BRICS) stock markets has been cited by Henrique, Sobreiro and Kimura (2019) and Ballings *et al.* (2015) as a topic of interest for future research.

Finally, this dissertation's main contribution is the incorporation of general and local interpretation methods in a trading strategy in a manner that allows for the investor to decide to trust (or not) the model and the individual predictions and to make an informed decision. While some forms of interpretation and feature importance are discussed in other works, this is the first, to the best of our knowledge, to integrate a model-agnostic general interpretation method with local explanations of each buy and sell position.

3 METHODS AND ANALYSIS

This chapter is divided in five sections. The first discuss the modeling goal and the predicted data. The second deals with the features and describes the data collection and feature engineering processes. It also gives a basic descriptive analysis of the data. The third section explains the preprocessing steps used with the data. The fourth section describes the implementation of the models themselves. The final section discusses measurement of the models and the use of model interpretation techniques.

3.1 Modeling Goal and Predicted Data

3.1.1 Objective Definition

The goal of our model is two-fold. First we want to implement a successful trading strategy based on machine learning that focuses on fundamentalist indicators. Second, we want to understand the relationship between these indicators and the financial returns. To accomplish theses goals a long-short trading framework using the Ibovespa index as a benchmark is implemented, predicting which stocks will perform significantly better than the benchmark (and should be bought by an investor) and which will perform significantly worse (and should be sold).

As discussed in the literature review, the problem of predicting stock returns using machine learning can be approached either as a regression or a classification problem. A regression may seem at first glance the natural option, since it will try to predict the exact return. However, the classification definition of the problem, where the goal is to try to predict which stocks will perform better (or worse) than a certain threshold, is also possible. Leung, Daouk and Chen (2000), for example, show that classification scores coupled with threshold rules outperform level prediction when applied to an index-based trading strategy. Based on their conclusions and seeking to take advantage of the different algorithms available for classification, this dissertation defines the problem as one of classification.

The choice of a classification algorithm does not limit the strategy to using only the binary output of the model. In fact, all the classification algorithms used are capable of giving a probability score that a certain data point belongs to the target class, allowing the trading strategy to be based on the probabilities rather than on the predictions themselves.

Instead of just separating positive from negative returns, we set our target with a minimum

return that works as a safety margin. This safety margin can be set as a δ in percentage points versus the benchmark. Since we are using this margin, the over- and underperforming results are not complementary (which would be the case if δ were 0) and therefore two separate models are needed, one for each direction.

The positive and negative binary predicted values are defined as follows.

$$BPos_i = \begin{cases} 0, \text{if stock of company } i \text{ has a return below or equal to Ibovespa plus } \delta \\ 1, \text{if stock of company } i \text{ has a return greater than Ibovespa plus } \delta. \end{cases}$$

And for the negative.

$$BNeg_i = \begin{cases} 0, \text{if stock of company } i \text{ has a return greater than or equal to Ibovespa minus } \delta \\ 1, \text{if stock of company } i \text{ has a return below Ibovespa minus } \delta. \end{cases}$$

3.1.2 Data Collection

All raw data for the predicted values of the model was obtained from Bloomberg Terminal (2019a) in the form of daily adjusted closing prices. The chosen scope are stocks that compose the Ibovespa index at the time of retrieval (2019). We thus avoid the problem of large variations in returns seen in stocks with low liquidity. This choice is also very practical for modeling, since it naturally produces a balanced class around the index itself. The stocks considered are listed in Appendix A.

Since we take the stocks that comprise the index today, it could be argued that this would add selection bias to the data. Selection bias could then reduce the model's ability to generalize, leading to overfitting (when a model works well in trained data, but performs poorly in unseen data). To avoid this generalization problem, we make sure to validate the model with out-of-sample data from a time-split, mimicking a real-world setting.

The period of study for the Brazilian stock market is from 1995 to 2018. The decision to start in 1995 is due to the currency stabilization that occurred in Brazil in the prior years. Machine learning works when the conditions of the past are similar to those of the future. This choice of period allows for the model to be trained in somewhat stable economic conditions, avoiding the overfitting problem.

3.1.3 Data Processing and Treatment

The fundamentalist indicators used in our model are calculated from companies' financial statements. These statements are released each quarter, for the Brazilian stock market. So although data was collected daily, each data point must correspond to the interval between results releases, which is quarterly. If predictions are to be made based

on the balance sheet results, one must account for the publication gap between the end of the quarter and the release of the results, while avoiding any spillovers between data from one quarter to the next. To avoid this problem, only a two-month window is considered for results prediction, allowing for a safety margin between publication for the considered quarter, the results accrual, and the publication of the results of the next quarter.[1]

With this approach it should be clear that the intention is not to capture the effect of the release of results itself (since the day of release is not contained in the two-month window), but rather the predictive power of the indicators in the evolution of the stock price between one released result and the next.

To obtain the desired returns, daily log-returns are calculated with the adjusted closing prices. These returns are summarized for the two-month prediction window after each statement is published. The number of data points is determined by the number of financial statement releases for each company, starting in the first quarter of 1995 and ending in the last quarter of 2018. Considering these limitations, the maximum number of data points for each company is 96.[2]

Furthermore, when the price information was unavailable for at least 10 days in each period, the information was discarded.

3.1.4 Descriptive Analysis

Figure 2 contains three different histograms of quarterly returns from the period studied. First, the histogram of stock returns is presented in Figure 2a. The distribution of returns has a bell-shaped and symmetric appearance, with a positive median.

Since there are factors that affect the stock market as a whole, the histogram of index returns (Ibovespa) is presented in Figure 2b. The distribution of returns also has a bell-shaped and symmetric appearance with a positive median, although the graph for this distribution has a less-defined appearance. This is expected since Figure 2a includes data for more than 60 assets while Figure 2b has only one asset.

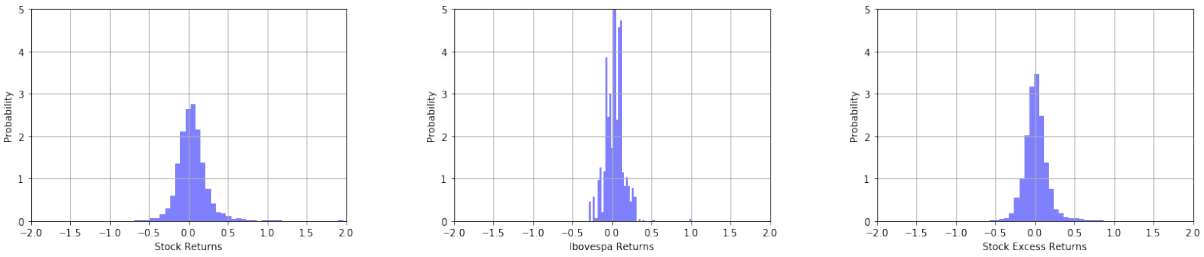
Finally, the histogram of excess return of each stock against Ibovespa is presented in Figure 2c. This distribution also has a bell-shaped and symmetric appearance with a median much closer to zero, as expected. This last set of returns is the one used to compute the predicted variable.

[1] The period considered is from 60 to 100 days after the end of the quarter. This implies that the results are measured after the following quarter's end, but before the release of the results related to it.

[2] Companies that went public after 1995Q1, closed capital or ceased to exist have fewer data points.

Figure 2 – Histograms of Stocks, Ibovespa and Excess Returns

(a) Stock Returns (b) Ibovespa Returns (c) Stock Excess Returns



Histograms for (a) the log-returns of each stock for each quarter; (b) the log-returns of Ibovespa for each quarter; and (c) log-returns of each stock against Ibovespa for each quarter
Source: Author

Table 2 – Descriptive metrics of the returns

Metric	Equity Return	Index Return	Excess Return
mean	4.8%	3.4%	1.4%
std	19.1%	11.7%	15.7%
min	-68.3%	-28.9%	-63.8%
D1	-14.5%	-10.3%	-15.2%
D2	-8.5%	-5.9%	-9.1%
D3	-3.9%	-1.9%	-5.4%
D4	0.1%	0.9%	-2.2%
D5 (median)	3.6%	3.7%	0.6%
D6	7.1%	6.9%	3.4%
D7	11.1%	9.6%	6.7%
D8	16.3%	11.2%	10.7%
D9	24.9%	17.5%	17.3%
max	254.5%	101.1%	250.6%

Summary metrics for the stock returns, index (Ibov) return and excess stock returns. Mean refers to the mean value for each variable, std refers to the standard deviation, min and max refer to the minimum and maximum values respectively, and D1 to D9 represent the deciles.
Source: Author

Table 2 compares the mean, standard deviation, minimum, maximum and the deciles (including the 5th decile, which is the median) of each metric presented in Figure 2. The excess returns series has a median close to zero, as expected, and a positive mean[3] of 1.4% equal to the equity returns minus the index. The standard deviation is also considerably smaller in the excess return than in the returns themselves, which supports the use of excess returns to mitigate the impact of market risk.

For the binary classification goal, a threshold δ must be chosen to separate over- and underperforming stocks from the rest. In this study a threshold of one quartile was chosen to maintain a safety margin while avoiding the problems associated with unbalanced

[3]The index is not equally weighted. Although the stocks presented here are those that constitute the Ibovespa index, it is not expected that both returns must match.

datasets. This choice gives a positive δ_P of 8.4% and a negative δ_N of 7.1%.

Summary statistics for the overperforming, underperforming and other stocks are shown in Table 3. Positive and negative δ correspond to the minimum of the overperforming and maximum of underperforming sets as expected. While the median is similar (13.1% in absolute values) between the two sets of interest, another key characteristic is the asymmetry of both mean and standard deviation between over- and underperforming stocks, with the first having a higher mean and deviation. The maximum of 250.6% for the overperforming stocks is also significant, since there is no limit for stock growth while any fall is limited to -100%, which could also explain the small variance in the latter set. Furthermore, this division serves as a reference to performance evaluation. A perfect model, which could single out all stocks that would over- or underperform, would have an expected excess return between 15 to 20 p.p. over the Ibovespa. A real model will most likely perform significantly worse than that, but the reference is valid as the best performance achievable.

Table 3 – Descriptive metrics of the excess returns of the performance sets

Metric	Overperforming Stocks	Underperforming Stocks	Other Stocks
mean	19.8%	-15.4%	0.6%
std	16.4%	7.9%	4.3%
min	8.4%	-63.8%	-7.1%
Q1	11.3%	-18.9%	-3.0%
Q2 (median)	13.1%	-13.1%	0.6%
Q3	21.5%	-9.6%	4.1%
max	250.6%	-7.1%	8.4%

Summary metrics for the overperforming, underperforming and other stocks' returns. Mean refers to the mean value for each variable, std refers to the standard deviation, min and max refer to the minimum and maximum values respectively and Q1 to Q3 represent the quartiles.
Source: Author

3.2 Features' Data and Engineering

3.2.1 The Features

The key predictors used are based on the work of Raposo and Cruz (2002) and complemented with factors identified by Fama and French (1992). Except for the size, all other metrics are ratios (which includes margins and multiples) making them comparable between different companies. By this construction, the size feature carries all the information on different scales, though it is still possible to analyze the interaction between this feature and others.

Dynamic components are also added in the form of the QoQ and YoY[4] growth of the ratios mentioned above. Since growth rates are inherently comparable, QoQ and YoY evolution of revenue, net income, EBITDA, total equity, book value and adjusted closing price are also added. All this data can be obtained from a single release of results.

Categorical variables representing industry and sector are also included. For illustrative purpose, Azul S.A. (AZUL4) is in the airlines industry and the industrial sector while Companhia Siderúrgica Nacional (CSNA4) is in the materials sector and steel industry. A full list of companies and respective industries and sectors is presented in Appendix B.

Since the goal is to identify fundamentalist factors and not company- or time-specific trends, the underlying company of each stock is left out of the model. For the same reason, any time-related information (except for the QoQ and YoY growth) is also left out, which also helps to avoid overfitting the model with past trends that may not repeat in the future.

The final list of features is shown in Table 4 and divided in six categories for didactic purposes: Size, Firm Value, Liquidity, Performance, Dynamic (evolution between quarters) and Characteristic, which contains the categorical variables.

3.2.2 Data Collection

All balance sheet and results information was obtained from Capital IQ's Database (2019) for each quarter and company. The categorization in sectors and industries was also obtained from Capital IQ.

The adjusted closing price data, used in part of the ratios and as input for the QoQ and YoY growths, was obtained from Bloomberg to be consistent with the information on returns discussed in Section 3.1. The prices used are always of the period preceding returns measurement.

3.2.3 Data Processing and Treatment

After obtaining the raw data from Capital IQ, some treatments were made to better suit the purpose of the model.

For the four companies marked as Banks (Santander, Bradesco, Itaú and Banco do Brasil) the Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) was substituted for the Earnings Before Taxes (EBT). This change is necessary because the EBITDA is meant to represent the cash flow generated from the operation of the company.

[4]Quarter-over-quarter and year-over-year respectively.

Table 4 – List of features used in the model

Type	Feature
Size	Size
Firm Value	Book-to-Market
Firm Value	Price/Earnings
Firm Value	Price/EBITDA
Liquidity	Current Liquidity Ratio
Performance	Asset Turnover
Performance	EBITDA Margin
Performance	Net Margin
Performance	Return on Assets
Performance	Return on Equity
Dynamic	YoY and QoQ Growth of Above Metrics
Dynamic	YoY and QoQ Growth of Adjusted Closing Price
Dynamic	YoY and QoQ Growth of Book Value
Dynamic	YoY and QoQ Growth of EBITDA
Dynamic	YoY and QoQ Growth of Net Income
Dynamic	YoY and QoQ Growth of Revenue
Dynamic	YoY and QoQ Growth of Total Equity
Characteristic	Primary Industry
Characteristic	Primary Sector

List of features divided in six categories: Size, Firm Value, Liquidity, Performance, Dynamic (evolution between quarters) and Characteristic (categorical variables). YoY and QoQ refer to Year-over-Year and Quarter-over-Quarter growth, respectively.
Source: Author

For other industries, financing activities are secondary and should not be compared in cash flow terms, but this is not true for banks, making this adjustment necessary.

The size feature was created taking the base 10 *log* from revenue.[5] Applying the logarithm can be understood as taking the order of magnitude of the revenue. This also makes variations in size more comparable between different companies since a variation in *log* is an approximation of the percentage variation.

Finally, the YoY and QoQ growth were calculated with variation between times t and the reference period (t-4 and t-1, respectively). In both cases, the variation was divided by the absolute value of the metric in the reference period.

After processing, when data was unavailable for any feature, the entire row was discarded.

[5]The *log* was taken from the revenue in thousands. This does not make any substantial difference since it only changes the position on the scale. Furthermore, since re-scaling is applied in the pre-processing stage, any difference would also be eliminated.

3.2.4 Descriptive Analysis

Table 5 presents the mean, standard deviation and quartile information (including the 50% quartile, which is the median) for all numeric features of the model that are not YoY and QoQ growth. This subsection comments on key aspects of this data. The full table with all features is presented in Appendix C.

Table 5 – Descriptive Statistics of Model Features

Metric	mean	std	min	Q1	Q2	Q3	max	IQR	range
Asset Turnover	0.78	0.64	0.01	0.39	0.60	0.97	4.62	0.58	4.61
Book-to-Market	1.09	1.47	-0.08	0.37	0.70	1.26	30.13	0.89	30.21
Current Liquidity Ratio	1.81	1.83	0.34	1.15	1.56	2.07	36.20	0.92	35.86
EBITDA Margin	27.75	23.57	-169.10	11.30	22.00	39.40	324.10	28.10	493.20
Net Margin	13.68	31.61	-249.70	2.28	9.14	18.40	741.30	16.12	991.00
Price/Earnings	695.02	31,473.37	-36,779.59	16.06	39.83	77.37	1,576,513.35	61.31	1,613,292.94
Price/EBITDA	24.77	233.82	-10,467.70	9.75	19.98	36.29	3,284.50	26.54	13,752.20
Return on Assets	6.76	5.74	-35.90	3.36	5.69	9.18	62.70	5.82	98.60
Return on Equity	10.23	282.92	-14083.90	5.65	13.70	24.80	200.90	19.15	14,284.80
Size	3.32	0.59	1.56	2.92	3.31	3.69	4.95	0.77	3.39

Summary metrics for key numeric features. Mean refers to the mean value for each variable. Std refers to the standard deviation. Min and max refer to the minimum and maximum values, respectively. Q1 to Q3 represent the quartiles. IQR is the inter-quartile range. Range is the range itself, *i.e.* the difference between the maximum and minimum values.

Source: Author

From the descriptive statistics, some key characteristics of the data are readily observable and must be accounted for in the pre-processing stage.

First, except for size, most of the metrics have large dispersions, which shows not only in relatively large standard deviations compared to the mean (return on equity, for example, has a standard deviation of 282.92, approximately 28 times larger than the mean, which is 10.23), but also in the large ranges (difference between maximum and minimum values). This characteristic suggests the presence of outliers in the data, a potential problem for good model performance. Considering this, the interquartile range (IQR, or the difference between Q3 and Q1) is also shown as a dispersion metric, since it is robust to extreme values.

A possible explanation of these extreme values consists in the use of ratios. A small denominator will make the ratio become very large. A company which in a particular quarter has earnings close to zero will have a high price/earnings ratio, for example.

Another striking characteristic of the data is that except for size, asset turnover and current liquidity ratio, the minimum value of the metrics is negative. Even though this situation is not desirable for the company, both earnings and EBITDA can be negative, which would lead to the associated metrics also being negative.

Finally, it is interesting to compare price/earnings and price/EBITDA, since they are not only similar metrics but also traditionally associated with company valuation. The price/EBITDA ratio is smaller than its earnings counterpart, as expected, since earnings are usually smaller than EBITDA after interest, taxes and D&A are deducted. The variation on price/EBITDA is also smaller than that of price/earnings. In fact, not only is its standard deviation 9.4 times larger than its mean (vs 45.3 for earnings), if the more robust comparison between IQR and median is made, price/EBITDA's IQR is only 33% greater than the median vs 54% for earnings.

3.3 Dataset Pre-Processing

This section deals with the pre-processing operations done to the dataset prior to training them. These steps may be optional, used only to try to improve the performance of the model (e.g. scaling), or mandatory (such as encoding categorical variables for most models).[6] It is also the case that more than one alternative is available for the same operation (e.g. standard scale or robust scale) or that the same operation may be applied with different parameters.

[6] One notable exception is CatBoost, where it is not only possible but advisable to use the categorical features without encoding.

In view of the multiple possibilities presented, and to avoid repeating a description in each case of the full treatment used, a default configuration is defined throughout this text. Only when a different configuration is applied will it be highlighted. Henceforth, any model run mentioned, unless otherwise noted, considers that data points with missing values were excluded, the categorical variables were encoded,[7] winsorization with a 2.5% threshold was applied, a robust scale was fitted, and no feature reduction technique was used.

3.3.1 Missing Values

As mentioned in Sections 3.1 and 3.2, the data collected for both the predictors and predicted variables contains missing values. These absences happen mostly due to the nonexistence of the data point (for example, when the company had not gone public at that time). Potential problems with the data source are not discarded, especially for older financial statements. In both cases, the consequence is that most missing values are concentrated in the early years of collected data.

Even though alternatives to filling in missing values exist (for example, substituting for the mean or median of the dataset), they should not be applied to the predicted variable. Also, a trade-off exists between inaccurate data versus insufficient data. Therefore, taking in account that most missing data is in older data points and the goal is to predict future ones, no missing value replacement will be used. The rows will be discarded altogether.

A direct consequence of this choice is a natural overweight of the data closer to the present, which is not necessarily undesirable.[8]

3.3.2 Categorical Variables

Most classification models deal only with numerical data. Therefore a method to transform categorical variables into numbers must be applied. Since both categorical variables (industry and sector) are not ordinal[9] we use one-hot encoding to transform the data. One-hot encoding means that each value of the feature is converted into a new boolean feature which indicates if that data point belongs to that category or not. To avoid the multicollinearity trap, one of the newly-created variables is discarded. The omitted value corresponds to the case where all other variables of the same original feature are

[7] We apply one-hot encoding, where each categorical variable is substituted by one binary variable for each possible value except the first, which is dropped and corresponds to all other being zero

[8] Actually, it is sometimes even desirable to over- or underweight some characteristics of the dataset whilst training the model, which is usually performed through over- or undersampling the data.

[9] An ordinal variable has a set of values that have a natural order, such as "Very Bad," "Bad," "Neutral," "Good," and "Very Good" which could be converted, for example, into a 1 to 5 scale.

zero. Henceforth it is considered that the categorical variables have been one-hot encoded. The noteworthy exception is the Catboost algorithm, which does not require this step.

3.3.3 Extreme Values

As was stressed in Section 3.2, many of the features present large ranges with extreme values that affect both the mean and standard deviation of each feature's data. In order to avoid "modeling the exception" (which would lead to overfitting), these data points can be eliminated or transformed. One way to transform the data is winsorization, where the extreme values, defined by a percentage of the data on each side, are substituted by the most extreme value in the rest of the data, thus creating a cap for the maximum and minimum values. All features in this dissertation are winsorized at a 2.5% percentile.

3.3.4 Scaling

While some machine-learning algorithms work independently of the scale of the data, others may overweight the importance of a feature with a larger OM. To counter these problems, both linear and non-linear transformations can be used to center and scale the data (non-linear scales also may have other effects, depending on the algorithm).

While the standard scaling method consists of removing the mean and dividing by the standard deviation, in consideration of the previously-mentioned large variability and presence of extreme values in the data, the default method considered is robust scaling, where each data point is decreased by its median and divided by the IQR. Robust scaling is applied to all features in this dissertation.

3.4 Model Implementation

This section describes how each of the classification models was implemented. All the models were implemented using appropriate libraries in the programming language Python. Since all the models allow for some form of hyper-parameter tuning, some of them are discussed in the following subsections. If a parameter is not cited, the library's default value was used. The values set in this section are the ones used when a model is cited throughout the text, unless otherwise explicitly noted.

3.4.1 Decision Tree

The Decision Tree Classifier was implemented using scikit-learn library and its DecisionTreeClassifier object.

The following hyper-parameters were set:

- **criterion:** The split criterion was set to "entropy" representing the information gain.
- **max_depth:** The maximum depth of the tree was set to 10.

3.4.2 LightGBM

The gradient-boosting algorithm LightGBM was implemented using "lightgbm" library and its LGBMClassifier object.

The following hyper-parameters were set:

- **learning_rate:** The learning rate was set to 0.001.
- **boosting_type:** The boosting type was set to "gbdt," representing the "Gradient Boosting Decision Tree."
- **objective:** The objective was set to "binary," since the problem was defined as a classification.
- **metric:** The metric to be evaluted on was set to "binary_logloss."
- **sub_feature:** The fraction of the features used in each interaction was set to 0.5, meaning that 50% of the features are left out.
- **num_leaves:** The number of leaves was set to 100.
- **min_data:** The minimum number of data points in each leaf was set to 20.
- **max_depth:** The maximum depth of the tree was set to 20.
- **n_estimators:** The number of estimators was set to 20000.

3.4.3 CatBoost

The CatBoost Classifier was implemented using "catboost" library and its CatBoostClassifier object.

One key difference between CatBoost and other models is that it does not require one-hot

encoding of the categorical features. Therefore it was not applied whenever CatBoost was used.

The following hyper-parameters were set:

- **loss_function:** The loss function was set to "Logloss."
- **eval_metric:** The evaluation metric was set to "AUC."
- **cat_features:** The categorical feature were set to represent the "Primary Industry" and "Primary Sector" features.
- **iterations:** The number of iterations was set to 1000.

3.5 Results Measurement and Model Interpretation

This section describes how the performance of the models was assessed. It is divided into three subsections: Validation and Test, Performance Evaluation, and Model Interpretation.

3.5.1 Validation and Test

Validation and Test are required in a machine-learning model to avoid overfitting of the data and improve model generalization. To perform them, one must separate training data from validation and testing data which has not been seen by the model and is used for evaluation. The difference between validation and testing is that the former is used to tune the model's parameters and the second to assess the tuned model.

Though models specific to time-series data are not used, the data used is inherently temporal and thus subject to variations conditional to the time. In that context, a simple random separation of train and test sets may allow the model to learn and take advantage of a structural change that it would not have seen in a real-world context. Therefore, in order to correctly assess the performance of the models, the dataset is divided in two parts: an in-sample dataset which contains data from the beginning of the series until the 2$^{\text{nd}}$ quarter of 2017 and an out-of-sample dataset from the 3$^{\text{rd}}$ quarter of 2017 onwards.

3.5.2 Performance Evaluation

The models can be evaluated both in a strict machine-learning sense using standard metrics and with domain-specific metrics which represent the business goal that drives the

modeling.

Since the ultimate goal of this dissertation is to aid stock selection in a long-short strategy framework, more emphasis will be put on the business metrics. Standard metrics are also presented though, since they allow for easier comparison with other works in the literature and similar models in different domains.

This subsection is therefore divided in these two complementary approaches.

3.5.2.1 Standard Metrics

As mentioned in the literature review, the most common assessment metric when dealing with classification problems is accuracy. It is intuitively defined as the ratio between correctly classified data points and the totality of observations. It is therefore very simple to calculate.

3.5.2.2 Business Metrics

Looking from the perspective of the "client" of the model, which in this dissertation is an investor, he is arguably not interested in the "technical" metrics but rather in the end result of following this strategy to improve financial return. In that sense, one argument in favor of business metrics is that similar to the way that standard modeling metrics are comparable between models, any investing strategy is comparable in terms of returns, regardless of the mechanism that drives the actions of the investor.

As mentioned in Section 3.1, classification methods calculate the probability that each observation belongs to the modeled class. The probability can then be transformed into a binary label using a cut-off threshold. Since the output is a probability, it is in fact continuous and allows for different strategies, such as taking any percentile of best- (or worst-) classified stocks, regardless of the label. One key advantage of this approach is that the same amount of stocks is always chosen, a characteristic that does not hold true if only the labeled stocks are chosen.

We therefore take the average return of the top quartile of each model as the standard metric for financial return. The quartile is taken for each quarter and not across quarters, an approach consistent with what would happen if an investor used the model in a real-world setting.

3.5.3 Model Interpretation

Model interpretation is performed with two complementary approaches: Shapley Values and LIME.

3.5.3.1 Shapley Values

Shapley Values give an overview of the importance of each feature to predict the outcome of the model in the given dataset. They can consequently be used by the modeler or the investor to decide if he "trusts" the model, i.e., if the most important features make sense and are behaving in the expected fashion. If a model states, for example, that a low P/E ratio is a sign of an underperforming stock, the investor may be doubtful of its performance in a real setting.

The Shapley Values method is implemented with the library called "shap." We show for each model a summary plot, where each feature is presented with its associated shap values and a distribution of the sampled data points with corresponding feature values in color code. For more details refer to Section 2.3 of the literature review or Štrumbelj and Kononenko (2010).

3.5.3.2 LIME

The Local Interpretable Model-agnostic Explanation (LIME) method, as the name itself states, is applied locally. It can be used to get a sense of how the entire model works by application of the method in multiple data-points but is more readily applicable to explain why an individual prediction was made.

This characteristic makes it useful to explain to the investor why the model is predicting that a particular asset will overperform (or underperform) in that particular quarter. The investor is consequently able to decide, based on his own domain knowledge of the market, the underlying company or other data that is not available to the model, if he trusts that prediction or not. For more details refer to Section 2.3 of the literature review or Ribeiro, Singh and Guestrin (2016).

The method is implemented with the library "lime" and the tabular explainer is used.

4 RESULTS

The first section discusses the results of the models tested and combinations between them (ensembles). The second discusses how model interpretation can be incorporated in the decision to follow (or not) the models' trading recommendations.

4.1 Model Performance and Financial Return

First, we compare the models tested among themselves, discussing both the accuracy and, more importantly, the financial return in the out-of-sample dataset. Next we review the models' performance in the face of real-world constraints and considerations. We conclude by testing different ensemble-learning trading strategies based on the original models and compare their performance with the original models themselves.

4.1.1 Model Performance

This subsection presents the results of the models, starting with accuracy, but focusing mostly on financial results. Since the approach chosen is to model the long and short positions separately, the results of the pure strategies are shown first. The subsection concludes with the results of the long-short strategy. All results shown are calculated based on the out-of-sample time-split test set, thus avoiding overfitting.

Table 6 presents the accuracy for the three models tested in both the long and short objectives. CatBoost performs better overall, achieving 78.5% for the long position and 73.4% for the short. It is closely followed by LightGBM with 77.5% and 73.0%, a difference of less than or equal to 1 p.p.. The standard Decision Tree is the worst-performing model overall, especially in the short position where it achieves only 68.9%, more than 4 p.p. below its competitors.

Prediction performance for the underperforming stocks is also worse for all three models, with the differences ranging from 4.5 to 6.8 p.p.. This hints at a possible difference between the factors that predict which stocks will perform particularly well and those that predict which will perform particularly badly. This topic will be further explored in this dissertation. It also implies that predicting which stocks will not perform well (which could be achieved simply using the complementary probability scored in the long model) is different than predicting which ones will perform badly, which can only be achieved by building separate models.

Since the classification algorithms used give a classification probability score, one way to

Table 6 – Accuracy of the Compared Models

Objective	Decision Tree	LightGBM	CatBoost
Top Performing Stocks (Long)	74.7%	77.5%	78.5%
Bottom Performing Stocks (Short)	68.9%	73.0%	73.4%

Accuracy for the classification of top and bottom performing stocks used for the long (buy) and short (sell) strategies, respectively.
Source: Author

represent modeling performance is to measure the results of an equally-weighted portfolio of stocks ranked by this score. This cumulative result represents the return of a portfolio composed of the n-best scoring stocks according to that model, where n is the ranking. All models will therefore converge to the dataset average return, but while a "bad model" should be close to a horizontal line, a "good model" should have better results for the lowest rankings. These results are presented in Figure 3.

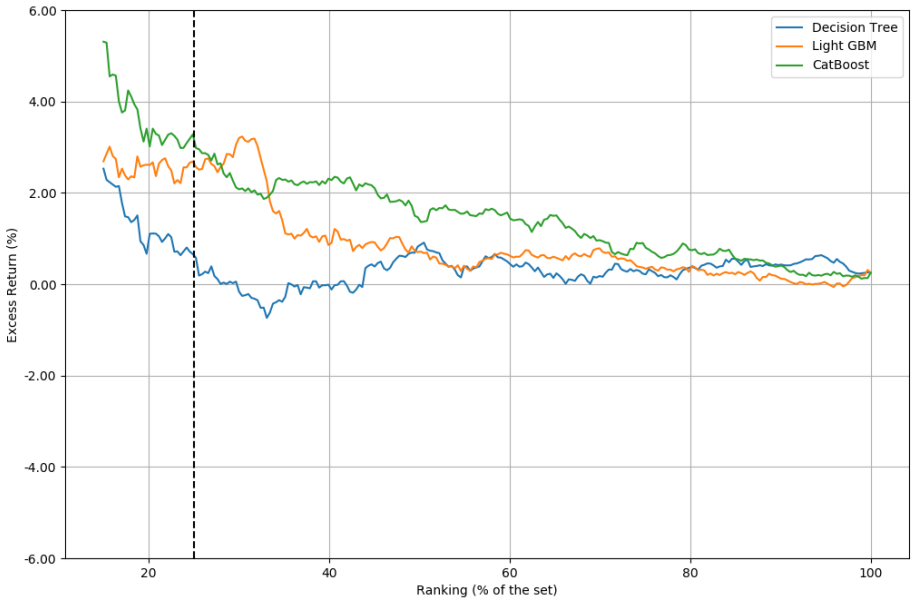
Performance of each model for the long position is shown in Figure 3a. As expected, all models show a peak in the best-ranked stocks and a downward trend that converges to the average return, which is close to zero since the excess returns against the Ibovespa are shown. Among the models, the Decision Tree has the poorest performance, trending quickly to the average and oscillating around it. CatBoost performs better in general, with its line standing above the others except in the 20-35% zone, where LightGBM has a peak. Both boosting methods show average excess returns of approximately 2.5% around the first quartile (25% on the horizontal axis, dashed line), which would lead to an excess APR (Annual Percentage Rate) of 15%, since each observation corresponds to a 2-month period.

Figure 3b shows the cumulative average performance for the short position. It is not as clear as in the long case and stays closer to zero. The boosting methods show a peak (now negative, since underperforming stocks are desired) in the best-ranked stocks and an upward trend that converges to the average. However this time the graph oscillates more, with overall result becoming close to zero or even positive. The Decision Tree has an even worse performance, being positive for most of the graph. The maximum performance is also less impressive, with only LightGBM going beyond 2.0% while in the long position (Figure 3a), CatBoost achieves more than 5.0%. At the first quartile reference (dashed line), LightGBM is the winner, but CatBoost performance is more consistent from the 35% mark onward.

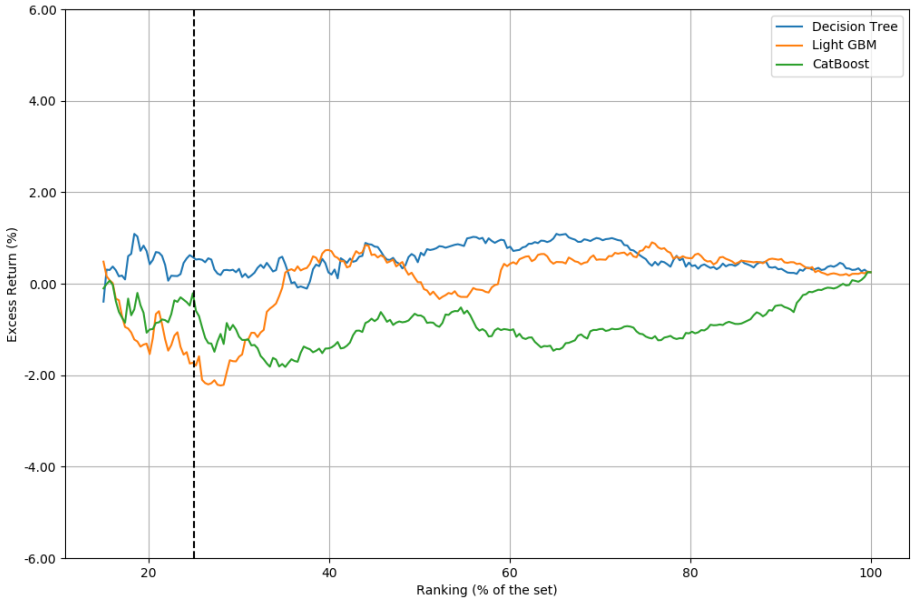
Taken together, Figures 3a and 3b present noteworthy common features. Both figures show that the cumulative average is non-monotonic for all models, showing that the ranking is not perfect. They also show the presence of "dents" and local "valleys" and "peaks," hinting

Figure 3 – Cumulative Return over Each Model's Score

(a) Long Models



(b) Short Models



Cumulative equally weighted average excess return (against the Ibovespa) for the portfolio according to the ranking[1] of stocks for each model in the (a) long objective; and (b) short objetive. The dashed line represent the first quartile in both cases.
Source: Author

at the high variance of the underlying excess returns of the stocks. Finally, they converge quickly to the mean. This may be a direct consequence of the chosen target variable, which focuses only on the extreme returns (first quartile of positive and negative returns), not the entire cross-section.

Even though the evolution of the cumulative average returns helps to show how the ranking can be used to build a strategy, it also makes it more difficult for a straightforward comparison. In order to use a single metric for comparison, financial results from now on are always reported based on the first quartile of each model, as described in Sub-section 3.5.2. This approach is not only consistent with the literature (GU; KELLY; XIU, 2018; MORITZ; ZIMMERMANN, 2016), but also has the advantage of choosing the same number of stocks regardless of the model or the objective. This is a desirable characteristic since it implies the same transaction costs, thus making the models directly comparable.

As the model's output is a probability score, a strategy could also be built to put more weight on stocks with a higher probability of overperforming. We choose not to do this because the number of companies in the first quartile each quarter is very limited and over-weighting could hurt diversification. Also, the model does not distinguish different returns as long as they are above the required threshold, and therefore a higher score does not mean a higher return. An equally-weighted portfolio of the stocks is therefore used. We leave the construction of classification-based models optimized for weighted portfolios for future research.

Following the equally-weighted portfolio strategy, the annualized[2] excess returns of the portfolios consisting of the stocks in the first quartile of each model are presented in Table 7. The portfolios built with the predicted underperforming stocks already consider a short strategy, in the sense that a positive portfolio return comes from a negative stock return. All results should be read then as "the higher the better."

Table 7 shows the Expected Annual Return (EAR) for the first quartile of each model.[3] As demonstrated in Figure 3, CatBoost prevails in the long position and LightGBM in the short position, with the Decision Tree lagging behind in both cases. As for the long-short strategy, LightGBM obtains the highest excess return, 26.4% against 20.0% of CatBoost. This can be explained by its superior result in the short position (10.5% vs 0.9%) out-weighting the advantage obtained by CatBoost in the long position (14.6% vs 19.0%).

[2] As mentioned in Section 3 each position is held for only two months to avoid potential data leakages. The result is therefore annualized based on this time interval.

[3] While for the long position the return calculation is straightforward, the short and long-short strategies calculate the return considering that the capital received by selling the asset is totally kept as collateral without compensation until the position is cleared. Therefore the short position return can be calculated taking simply the negative return of the asset.

Table 7 – Annual Excess Returns of the Compared Models

Strategy	Decision Tree	LightGBM	CatBoost
Pure Long	10.0%	14.6%	**19.0%**
Pure Short	-4.5%	**10.5%**	0.9%
Long-Short	5.1%	**26.4%**	20.0%

Expected Annualized Excess Return for the first quartile of the long (buy), short (sell) and long-short (buy the long stocks and sell the short) strategies respectively, with the results for the short strategy considered positive when the stock goes down. For the long-short strategy, the strategy is first applied and then compounded, and therefore it is different than simply summing the long and short strategies directly. The best result is marked in bold.
Source: Author

For all the models, the result of the long position is superior to the short (especially for Decision Tree and CatBoost), further showing that these two problems are not symmetrical.

Since the test set spans 6 quarters, from 2017Q3 to 2018Q4, the evolution of the compounded results for each of the pure strategies and the long-short strategy are shown in Figure 4. The results shown do consider that no investment was made in the one-month intervals between quarters that were left out of the results accrual. Each quarter represents, in fact, a two-month period.

The compounded return of the pure long strategy [4] is shown in Figure 4a. As expected from the average returns, CatBoost has the highest result and clearly dominates the Decision Tree. LightGBM has good performance in the first quarters and wins up until the fifth quarter, when CatBoost takes the lead, leaving LightGBM in second place. Despite its better performance, CatBoost has significant losses in the second and fourth quarters, while the other models experience only minor losses in the quarter-over-quarter evolution. All models have positive returns, with LightGBM staying positive in all quarters.

The pure short strategy,[5] on the other hand, presents far worse results, as shown in Figure 4b. Decision Tree and CatBoost have negative performances. LightGBM manages to achieve the best result overall. As in the long strategy, it stays in the positive zone at all times.

Finally, Figure 4c presents the compounded return of the long-short strategy.[6] While all models again achieve positive returns, Decision Tree has a poor performance of only 5.0%

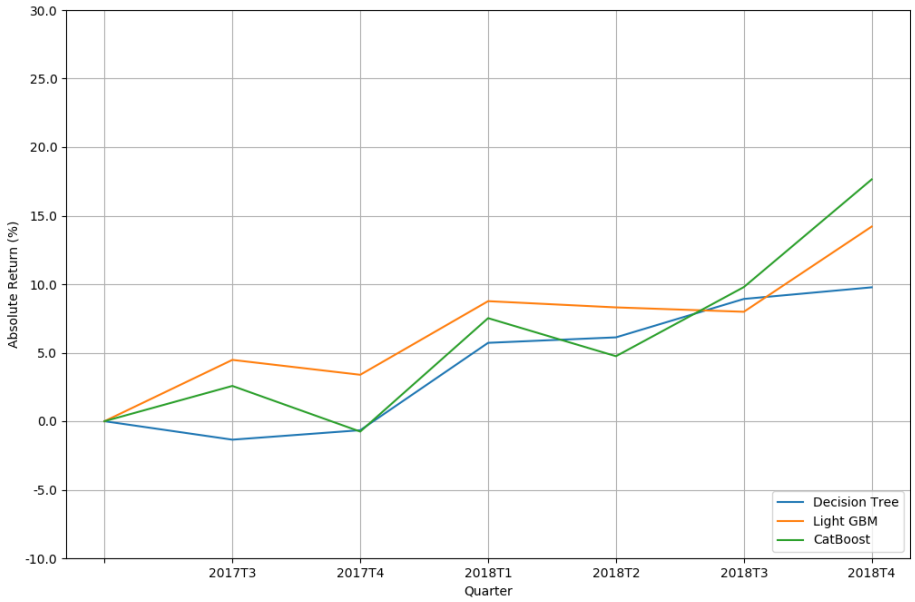
[4] Where the investor buys the recommended assets and "sells" the Ibovespa index.
[5] Where the investor sells the recommended assets and "buys" the Ibovespa index.
[6] Where the investor buys the recommended assets from the long model and sells the recommended assets from the short model.

while CatBoost and LightGBM exceed 15.0% and 25.0%, respectively. LightGBM is the clear winner, dominating the other models at all times.

While Figure 4 shows the evolution of the excess returns, an investor could decide

Figure 4 – Compounded Quarter-over-Quarter Excess Results Evolution

(a) Pure Long Strategy



(b) Pure Short Strategy

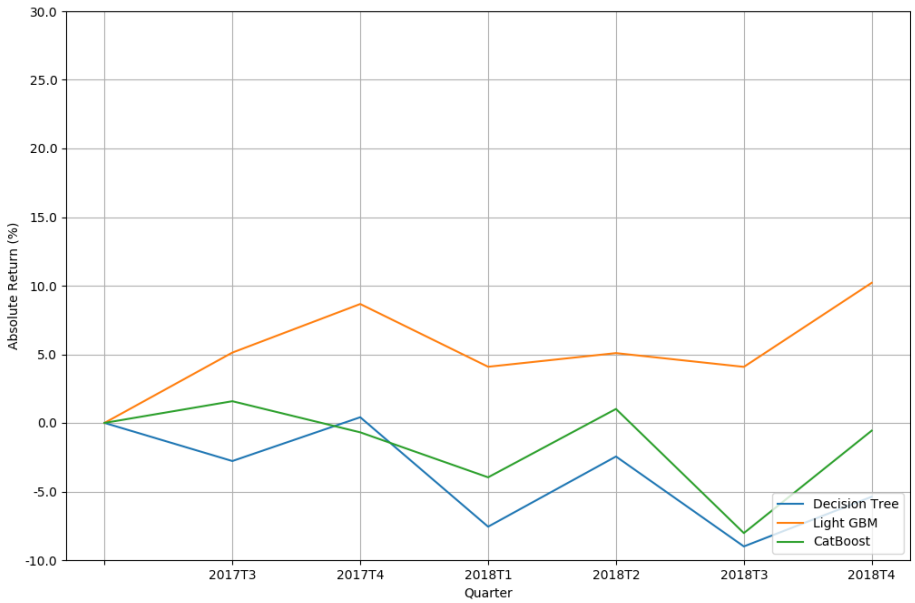
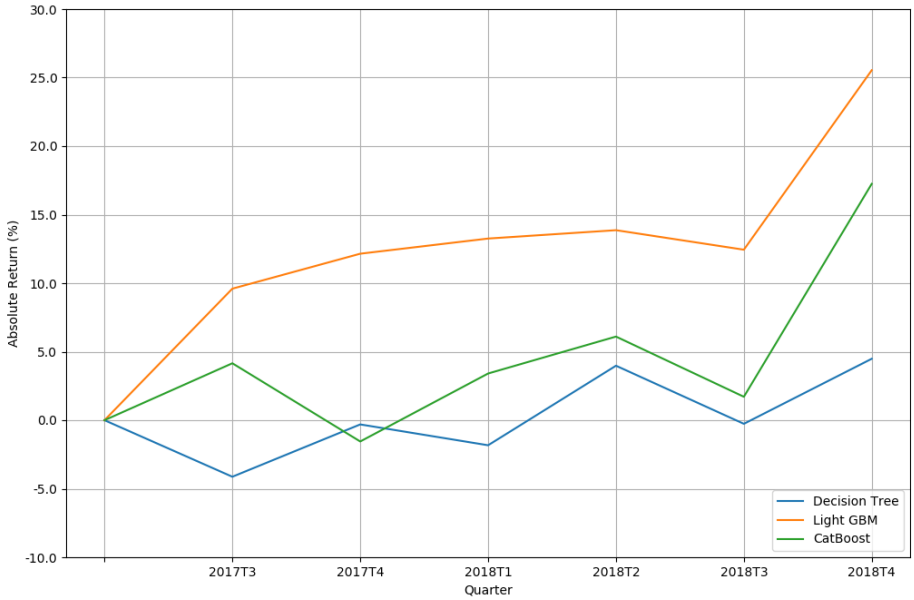


Figure 4 – Compounded Quarter-over-Quarter Excess Results Evolution (cont.)

(c) Long-Short Strategy



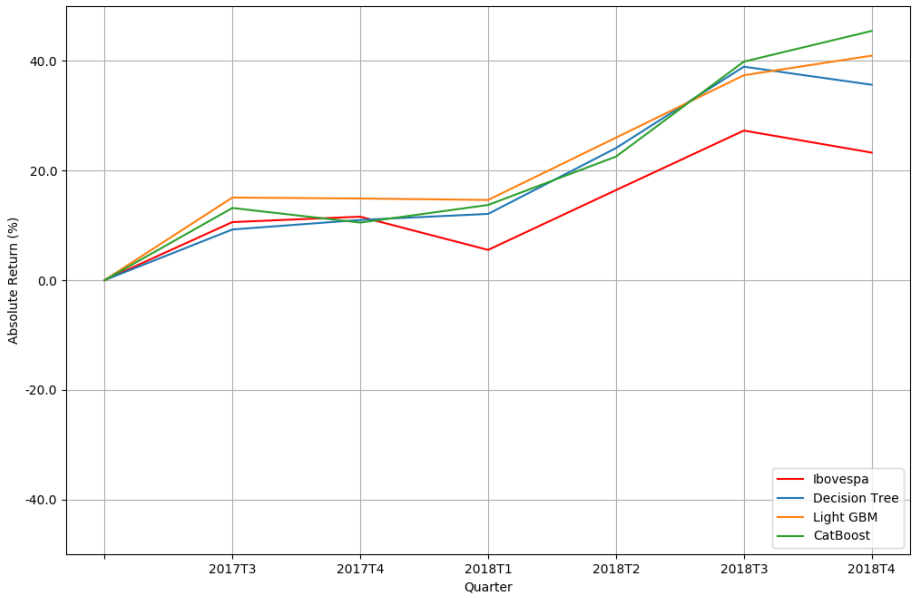
Compounded evolution of the excess returns of the three portfolios based on the models developed for (a) the long strategy; (b) the short strategy; and (c) the long-short strategy.
Source: Author

to also take the market risk while also trying to "beat" the index. Figure 5 compares the absolute return of the long and short strategies compared to a modified Ibovespa index that considers only the periods when the model is active (i.e. only the two-month "safe period" between results releases). The long position is compared with a long position in the Ibovespa and similarly, the short position is compared with a short position in the Ibovespa. Since in the test period the index has an overall positive return, the result of a short position in the Ibovespa is negative. The long-short strategy is not shown because it does not depend on the index and therefore yields the same results with absolute or excess returns.

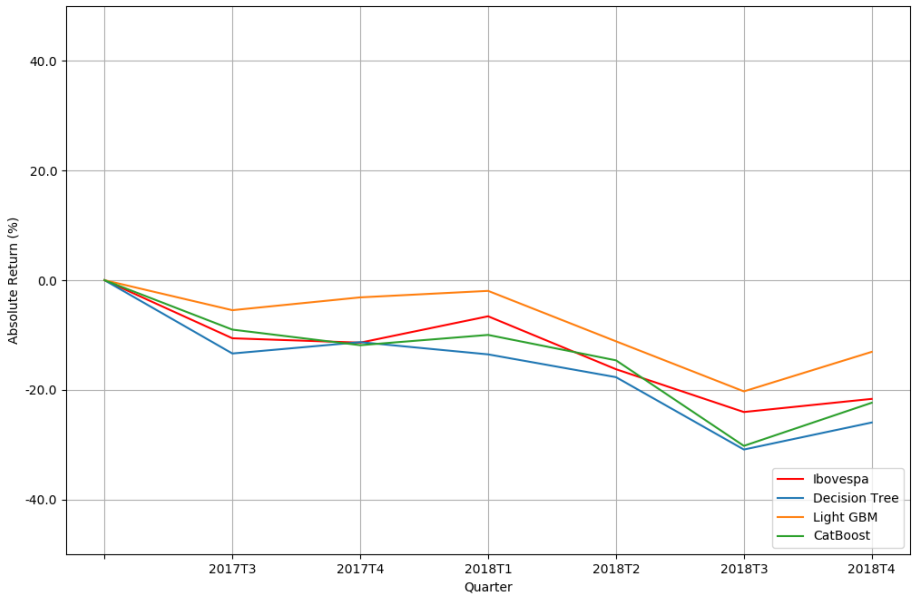
As expected from the previous results, all models outperform the index in the long strategy (Figure 5a). The boosting methods yield an approximate return of 40.0% while Ibovespa yields close to 25.0%. Figure 5b shows, nevertheless, that LightGBM consistently outperforms a short position on the Ibovespa, though it does not achieve positive results at any time, stressing the importance of the market risk in the overall result if it is not hedged.

Figure 5 – Compounded Quarter-over-Quarter Absolute Results Evolution

(a) Long Strategy



(b) Short Strategy



Compounded evolution of the absolute returns of the three portfolios based on the models developed for (a) the long position and the equivalent return of Ibovespa in the period; and (b) the short position and the equivalent return of Ibovespa in the period (also with a short position).
Source: Author

4.1.2 Transaction Costs and Risk

Even though these initial results are promising, economic constraints and costs must be accounted for. This subsection presents their impact. Although these considerations are common in the finance literature dealing with trading strategies, they are often ignored by the machine-learning literature that seeks to predict stock returns. This subsections helps bridge these two streams.

The first consideration is the trading costs associated with the transactions. Since these costs to buy and sell stocks are associated with each transaction, they must be accounted for, especially with higher-frequency strategies that induce a higher portfolio turnover. Frazzini, Israel and Moskowitz (2018) estimate the realized trading costs in international markets for recent years at 15.97 bps. Since their work does not consider the case of Brazil specifically, a conservative approach is taken and 20 bps are used for the long position and 40 bps for the short. This difference takes into consideration the more restricted short market in this country specifically.

Comparison of the two cost scenarios with the original results, shown in Table 8, does not

Table 8 – Annual Excess Returns with Transaction Costs

Cost Scenario	Strategy	Decision Tree	LightGBM	CatBoost
Without Costs	Pure Long	10.0%	14.6%	**19.0%**
	Pure Short	-4.5%	**10.5%**	0.9%
	Long-Short	5.1%	**26.4%**	20.0%
Frazzini *et al.* (2018)	Pure Long	8.9%	13.6%	**17.9%**
	Pure Short	-5.4%	**9.5%**	0.0%
	Long-Short	3.1%	**24.1%**	17.8%
Conservative	Pure Long	8.7%	13.3%	**17.6%**
	Pure Short	-6.8%	**8.0%**	-1.5%
	Long-Short	1.4%	**22.1%**	15.9%

Expected Annualized Excess Return in three transaction costs scenarios for the first quartile of the long (buy), short (sell) and long-short (buy the long stocks and sell the short) strategies respectively, with the results for the short strategy considered positive when the stock goes down. The best result is marked in bold. Both literature and conservative scenarios are reported.
Source: Author

alter the winning algorithm for each trading strategy. This is expected since the algorithm does not change the turnover of equities. The impact is greater on the long-short strategy because it doubles the number of operations. In the conservative scenario, Decision Tree's return in the long-short strategy decreases almost to zero, but it is still profitable in the long position. Catboost faces an interesting inversion, with its short position becoming

negative and its long-short strategy inferior to its pure long strategy when transaction costs are considered. Despite the costs, LightGBM is still the overall winner, with all its positions profitable.

The second concern not commonly addressed in the machine-learning literature is risk-adjusting the returns, since a higher return may come at the expense of higher volatility of the portfolio. A common benchmark metric of volatility is the Sharpe ratio, which takes the return subtracted from the risk-free rate[7] and divides by the standard deviation of this same metric.

Table 9 presents the Sharpe ratio for each of the pure strategies and long-short strategies.

Table 9 – Sharpe Ratios of the Compared Models

Strategy	Decision Tree	LightGBM	CatBoost
Pure Long	0.12	0.24	**0.34**
Pure Short	-0.38	**0.13**	-0.17
Long-Short	-0.29	**0.39**	0.17

Sharpe ratios of the long (buy), short (sell) and long-short (buy the long stocks and sell the short) strategies respectively. The Sharpe ratio is calculated in the annualized results. The best result is marked in bold.
Source: Author

Decision Tree has the worst risk-adjusted performance, as expected. It has negative ratios in the short and long-short strategies, and achieves only 0.12 in the long position. Repeating the previous results, LightGBM has the best performance in the short and long-short strategies, achieving Sharpe ratios of 0.13 and 0.39 respectively. LightGBM's long position is surpassed by CatBoost (0.34 vs 0.24). The latter model has a negative performance in the short position (-0.17) which also brings down its long-short performance (0.17).

While one of the key advantages of the Sharpe ratio is how easily it can be compared with other works, the evaluation must be contextualized. Gu *et al.* (2018) study 30,000 stocks over 60 years and report an additional Sharpe ratio in the long position versus the S&P 500 index of 0.26,[8] which is slightly better than LG's performance and inferior to CB's. On the other hand, the long-short strategy of Gu *et al.* achieves impressive 2.45 with an equally-weighted portfolio and 1.35 with the value-weighted one (GU; KELLY; XIU, 2018).

As discussed in Chapter 2, Avramov *et al.* (2019) show that the results obtained by

[7] The risk-free rate considered is usually the return on sovereign bonds, which in the Brazilian market poses an extra challenge, since its rate is comparatively high (ARIDA; BACHA; LARA-RESENDE, 2005).

[8] Using neural networks.

Gu *et al.* (2018) rely heavily on anomalies that are difficult to exploit in the face of real-world restrictions. Avramov *et al.* reproduce the results with the restrictions and produce a Sharpe ratio of 0.94, which is further reduced to 0.64 without micro-caps or with credit rating companies and 0.45 without credit downgrades. While the comparison is made without costs, a high portfolio turnover is also reported (AVRAMOV; CHENG; METZKER, 2019).

Since this dissertation considers only roughly 60 companies that compose the Ibovespa Index and not the whole cross-section, the stocks used are large-cap and liquid. The construction of a quarterly portfolio also generates a low turnover. With that perspective, the 0.38 achieved is close to the more conservative scenario reported by Avramov *et al.* using Gu, Kelly and Xu's (2018) algorithm. Combining LghtGBM's long model and CatBoost's short yields a Sharpe ratio of 0.50, surpassing the conservative scenario.

Another necessary consideration of the finance literature that is often neglected by the machine-learning research on predicting stock returns is risk factors. The decomposition of the cross-section established by Fama and French (1993) has assumed many configurations with authors arguing for more or less factors (the so-called "factor zoo") while investors try to find stocks with risk premiums not explained by the factors ("postive α").

To assess if the proposed models bring results not accounted for by the risk factors, a Fama-French five factor model (FF5) is estimated. The risk factors are taken from the finance research group Nefin, which has constructed a database with daily values for the five-factors model in the Brazilian equity market (NEFIN, 2020).

The FF5 results of each model are shown in Table 10. All the models have an intercept with the expected sign (positive for the long model, negative for the short). While the absolute values may seen small, an α of 0.07% on a daily basis amounts to approximately a 1.5% monthly return or 19.3% annual return. Nevertheless, only two models achieve a statistically significant intercept (at a 5% level). The first is CatBoost's long model, showing that it managed to be the best long model by finding stocks whose return is not explained by the risk factors. The second, surprisingly, is Decision's Tree short model, with its previous poor performance being associated with a lower market risk (its coefficient is 0.94 while all others are above 1.00) and not an inability to find negative α.

The market risk is significant in all models. It shows a progression among algorithms in both long and short legs. DT has lower βs, followed by LG and CB. WML is also significant in all long positions and in DT's short, but with the sign opposite of that expected. The other factors are not significant, except for SMB and HML, exclusively in the DT long

Table 10 – Portfolio Regression on Risk Factors

Strategy	**Long**			**Short**		
Algorithm	**DT**	**LG**	**CB**	**DT**	**LG**	**CB**
Intercept	0.07	0.07	0.09	-0.09	-0.03	-0.06
	(1.77)	(1.69)	(2.09)	(-2.26)	(-0.62)	(-1.26)
MF	1.06	1.09	1.18	-0.94	-1.03	-1.16
	(28.17)	(27.56)	(27.14)	(-23.3)	(-22.32)	(-23.98)
SMB	0.28	0.19	0.24	-0.03	-0.27	-0.10
	(2.12)	(1.35)	(1.57)	(-0.22)	(-1.68)	(-0.58)
IML	-0.17	-0.14	-0.12	0.05	0.27	0.03
	(-1.34)	(-1.08)	(-0.85)	(0.38)	(1.8)	(0.18)
HML	-0.13	0.00	-0.14	0.06	-0.02	0.09
	(-2.03)	(0.01)	(-1.83)	(0.88)	(-0.22)	(1.01)
WML	-0.12	-0.15	-0.37	0.13	0.13	0.13
	(-1.99)	(-2.38)	(-5.52)	(2.07)	(1.77)	(1.72)

Coefficients for the regression on risk factors in Fama-French five factor model fashion. Daily Factors are obtained from Nefin and estimated in the out-of-sample period. The factors are MF (Market Factor, or the excess return of the market portfolio), SMB (Small Minus Big, representing Size), HML (High Minus Low, which refers to the book-to-market value), WML (Winners Minus Losers, which accounts for past returns) and IML (Illiquid Minus Liquid, associated with stock liquidity in the market). The intercept represents the premium not explained by these factors. The t-statistic is presented in parentheses.
Source: Author

position.

While other researchers report FF5 adjusted returns with consistent significant intercepts, the decision to only use large-caps coupled with the limitations of the Brazilian equity market severely limited the sample size. Giovanetti *et al.* (2016), for example, compare the US market with Brazil's and show the difficulty of robustly estimating risk premiums with the lower number of stocks and historic data available in Brazil.

4.1.3 Ensembling

This subsection proposes ensembles between the models already discussed in this section.

4.1.3.1 Ensembling Different Algorithms

While LightGBM achieves the best performance in the long-short strategy, as demonstrated in Subsection 4.1.1, the other models also show positive results. CatBoost, for example, outperforms LightGBM in the long strategy. Taking this into account, we explore potential ensembles between algorithms and compare them with the original results. The rationale is that different algorithms may have comparative advantages in spotting different kinds of opportunities. An ensemble could potentially be the best of two worlds.

Each ensemble is constructed by taking the average ranking across the models and taking the first quartile of the new ranking for each quarter. This approach is applied to all possible combinations between models, which amounts to three pairs of algorithms and a final combination that considers the three of them. The average return for these combinations is shown in Table 11, along with the original models that were already discussed in Table 7.

Analysis of the four new combinations presented in Table 11 shows that only LG + CB in the long position achieves (slightly) better results than the best original model for each strategy (LG for long and CB for short). For the long position, ensembles with DT have performance between those of DT and the other model that compose them. The best return comes from the combination between the two best source models (LG + CB) as already mentioned, but it fails to improve much over the original CB performance. One possible explanation is that CB effectively dominates the other models, and therefore adding them does not help performance.

Table 11 – Return of the Ensembles across Algorithms

Algorithm	**Long**	**Short**	**Long-short**
Decision Tree	10.0%	-4.5%	5.1%
LightGBM	14.6%	**10.5%**	**26.4%**
CatBoost	19.0%	0.9%	20.0%
Decision Tree + LightGBM (DT+LG)	10.4%	1.8%	12.4%
Decision Tree + CatBoost (DT+CB)	12.7%	-1.5%	11.1%
LightGBM + CatBoost (LG+CB)	**19.3%**	2.3%	22.0%
All algorithms (DT+LG+CB)	14.1%	5.3%	19.9%

Expected Annualized Excess Return for the first quartile of the long (buy), short (sell), and long-short (buy the long stocks and sell the short) strategies for each algorithm and ensembles among them, with the results for the short strategy considered positive when the stock goes down. The best result is marked in bold.
Source: Author

As for the short strategy, one ensemble has negative results (DT + CB). Again the LG + CB is the best of the two-way combinations. The ensemble DT + LG + CB has the best

performance overall, but is still far inferior to the original LG. Since the ensembles of both legs are inferior to the originals, the long-short strategy based on them is also disappointing.

Sharpe ratios were also calculated and are shown in Table 12. They show results analogous

Table 12 – Sharpe Ratios of the Ensembles Between Algorithms

Algorithm	**Long**	**Short**	**Long-short**
Decision Tree	0.12	-0.38	-0.29
LightGBM	0.24	**0.13**	**0.39**
CatBoost	0.34	-0.17	0.17
Decision Tree + LightGBM (DT+LG)	0.12	-0.16	-0.04
Decision Tree + CatBoost (DT+CB)	0.19	-0.29	-0.09
LightGBM + CatBoost (LG+CB)	**0.38**	-0.13	0.24
All algorithms (DT+LG+CB)	0.23	-0.05	0.21

Sharpe ratios of the long (buy), short (sell) and long-short (buy the long stocks and sell the short) strategies respectively. The Sharpe ratio is calculated in the annualized results. The best result is marked in bold.
Source: Author

to Table 11 in terms of winning algorithms. Therefore, this ensemble strategy does not have a significant impact on risk reduction either. From the results presented above we conclude that this strategy has generally poor performance.

4.1.3.2 Ensembling Long and Short Models

Sub-section 4.1.1 showed that the long and short models perform very differently. Since different models are developed for each strategy, this sub-subsection proposes ensembles between the long and short versions of each model, again comparing them with the originals. The potential advantage of such an ensemble is that it selects stocks with a high probability of overperforming and a low probability of underperforming. While these would be complementary events if a single model was developed, since two models are constructed, one particular stock could have a high probability in both models. One example would be a cheap stock of a company in financial difficulties, which could have a good potential to have a price increase, but it could also collapse. In a case like this, the ensembling strategy would give a medium score to the stock and remove it from the portfolio. One argument against this strategy is that the models were constructed focusing on the first quartile and not necessarily on having a good distinction between performance in the other quartiles.

Each ensemble is constructed taking the average between the ranking of the original strategy and the inverse of the ranking of the symmetric strategy. We then take the first quartile of the new ranking for each quarter. This approach is applied to the three algorithms and two strategies. The average return for these combinations is shown in Table 13, along with the original models that were shown in Table 7.

A first glance at Table 13 indicates that the effect of ensembling the long and short

Table 13 – Return of the Ensembles across Strategies

Algorithm	**Long**	**Short**	**Long-short**
Decision Tree	10.0%	-4.5%	5.1%
LightGBM	14.6%	10.5%	26.4%
CatBoost	19.0%	0.9%	20.0%
Decision Tree Long-short ensemble	6.2%	0.4%	6.6%
LightGBM Long-short ensemble	15.0%	-4.2%	10.3%
CatBoost Long-short ensemble	**31.7%**	**17.8%**	**54.0%**

Expected Annualized Excess Return for the first quartile of the long (buy) and short (sell) strategies for each algorithm and ensembles between the long and short strategies, with the results for the short strategy considered positive when the stock goes down. The best result is marked in bold.
Source: Author

positions is different across the algorithms and positions. The Decision Tree, for example, has worse performance using the ensemble in the long position than it originally had, while its short position becomes slightly positive. LightGBM has a mixed result. Ensembling improves its long-strategy performance (15.0% versus 14.6%) but completely reverses its results on the short strategy, moving from positive 10.5% to negative 4.3%. CatBoost, on the other hand, boosts its result by 12.7 and 16.9 p.p. for the long and short strategies, respectively. Figure 3 presents a potential explanation for this behavior, with CB maintaining a consistent performance across the rankings (not just the first quartile) while LG converges to zero around the 40% percentile. The long-short results are a consequence of the individual legs, with CatBoost again achieving the best financial return overall.

Sharpe ratios were also calculated and are shown in Table 14. As expected from previous

Table 14 – Sharpe Ratios of the Long-Short Ensemble Models

Algorithm	**Long**	**Short**	**Long-short**
Decision Tree	0.12	-0.38	-0.29
LightGBM	0.24	0.13	0.39
CatBoost	0.34	-0.17	0.17
Decision Tree Long-short ensemble	0.00	-0.21	-0.21
LightGBM Long-short ensemble	0.25	-0.34	-0.10
CatBoost Long-short ensemble	**0.77**	**0.37**	**1.26**

Sharpe ratios of the long (buy), short (sell) and long-short (buy the long stocks and sell the short) strategies respectively. The Sharpe ratio is calculated in the annualized results. The best result is marked in bold.
Source: Author

results, the CatBoost long-short ensemble achieves the best performance overall. Its long-short strategy reaches a Sharpe ratio of 1.26, superior to Gu, Kelly and Xu's (2018) result

for the full sample with economic restrictions (AVRAMOV; CHENG; METZKER, 2019). We conclude that this ensembling strategy, unlike the one between algorithms, produces better results than the original models.

4.2 Model Interpretation

This section first compares the tested models in terms of global interpretation and discusses the economic explanations of the results. Then it briefly proposes local interpretation and how it would be used by an investor to make an informed decision.

4.2.1 Global Interpretation

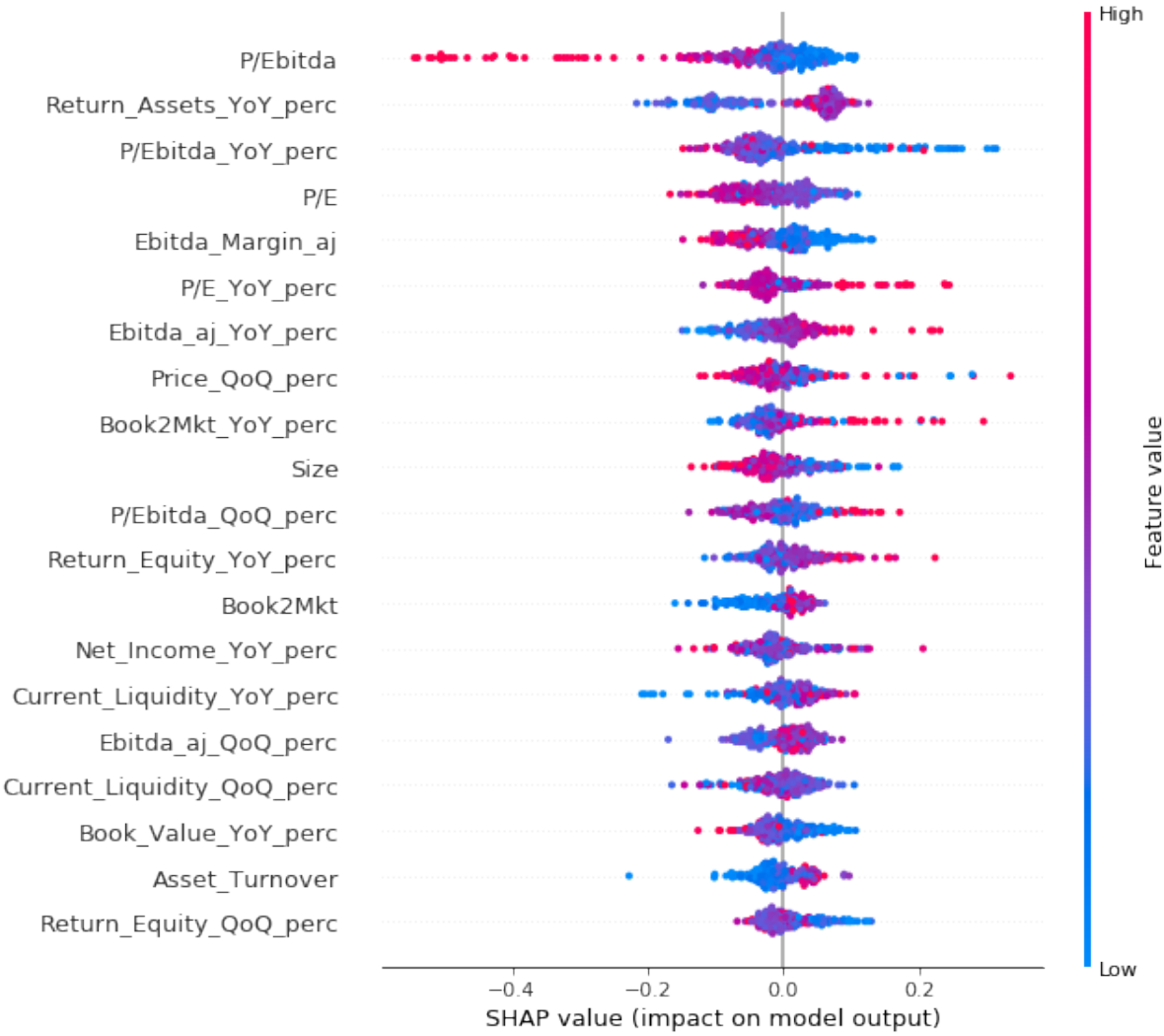
An investor is more confident in using a machine-learning model and can even improve the model's result if it is interpretable, as was thoroughly discussed in Section 2.3. To tackle this issue, this dissertation takes the approach of Vellido *et al.* (2012) shown in Figure 1 and interprets how each successful model predicts overperforming (or underperforming) stocks on a global level and discusses the economic interpretation of each feature.

The global interpretation approach we use, Shapley Values, could be applied to all algorithms since it is based on agnostic interpreters. On the other hand, Subsection 4.1.1 showed that CatBoost has the best performance for the long strategy and LightGBM performs better in the short strategy and also has the best long-short performance overall. Interpretation is therefore focused first on comparing CatBoost with LightGBM for the long strategy and then comparing the LightGBM's long and short models. The combinations whose figures are not presented in the main text can be found in Appendix D.

The Shapley Values are presented in the form of summary plots. Each plot shows the top twenty features in terms of average absolute impact to the predictions of the model being analyzed. It also shows the distribution of the impact in the form of a scatter plot. The color scale stands for the feature value, with a high value being associated with red and a low value with blue. A positive Shapley value increases the chance of a stock being selected to the portfolio and a negative one decreases it. This means that for a overperformance model a positive value correlates to a higher return, while for a underperformance model it correlates with a lower return. This means that if a particular feature shows a concentration of red dots on the negative side of the Shapley Value's axis of a overperformance model, a higher feature value is associated with lower chances of a stock overperforming.

The most important feature of CatBoost's long strategy is P/EBITDA. A higher ratio is associated with a lower chance of overperforming, as shown in Figure 6. This feature, along

Figure 6 – Shapley Values for CatBoost - Long Strategy



Shapley values plot relating features' values shown in a color scale from low values (blue) to high values (red) to their respective impact on the predicted variable. Only the top 20 variables are shown. In this model a higher shap value indicates a higher chance of a stock overperforming.
Source: Author

with P/E which is fourth on the list, has the expected direction of impact, based on the literature. A relatively low price, compared to the company's ability to generate cash flows and ultimately produce a profit for shareholders, is a standard valuation metric. Still, this relationship is not linear (especially for P/E), an effect that is captured by the model and can be witnessed by the red dots on the right-hand side of the shap plot. Even though it is less important than the other two, the model also associates higher book-to-market ratios with better performance, which is also expected from a fundamentalist view based on value.

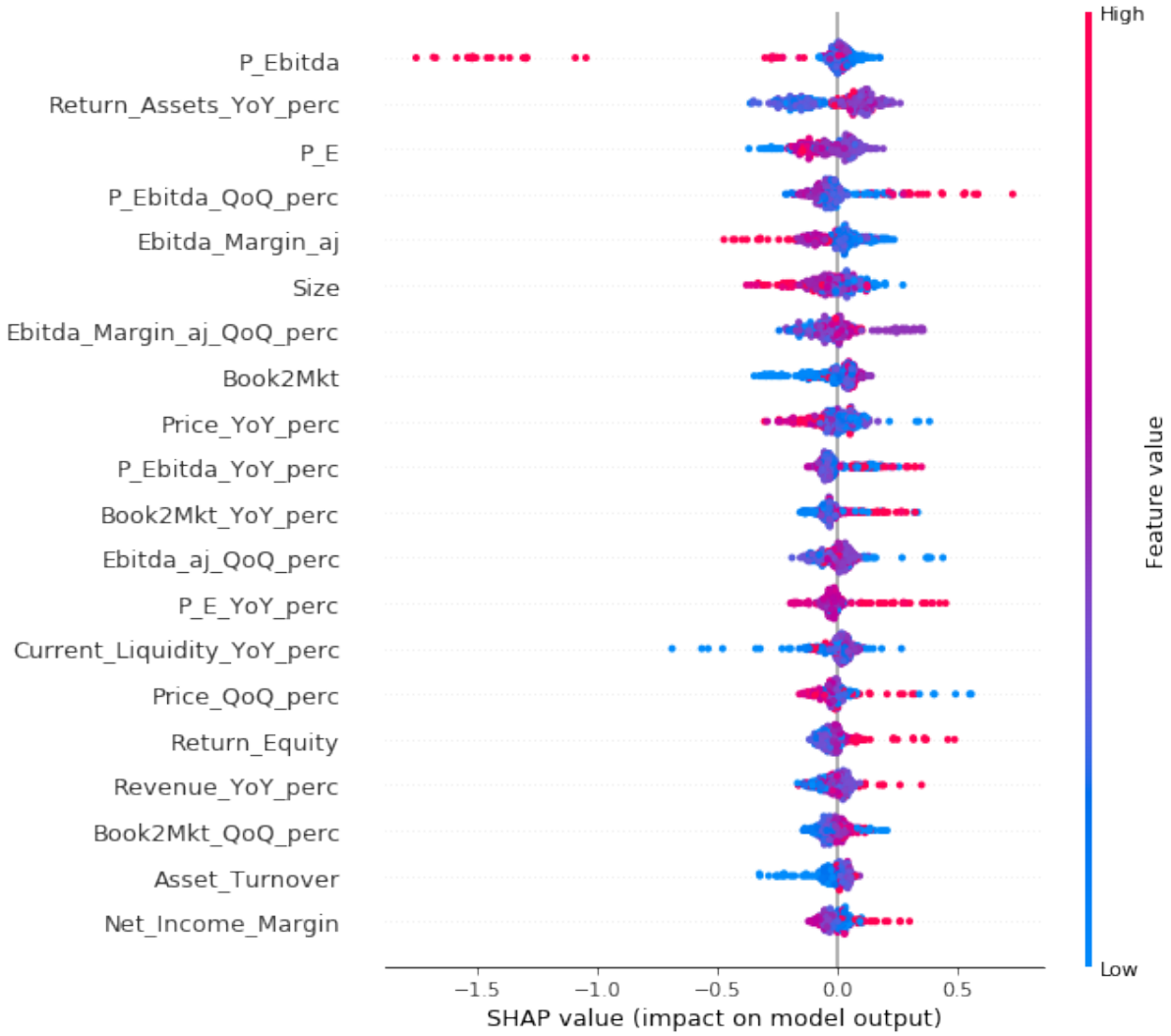
Although a low P/E or P/EBITDA is desired, and correlated with a higher chance of outperforming, the model also shows a positive impact of higher growth of these ratios in a QoQ or YoY view. This somewhat strange pattern can be interpreted as signaling a valuation momentum, when the company still has a low price-to-value but is moving towards a higher ratio. This dynamic is further explored in Sub-subsection 4.2.2.

As expected, the evolution of the company's return also improves its chance of overperforming. Growths on ROA, EBITDA and ROE are also top of the list of features (ROA YoY growth is actually second). Size, on the other hand, is associated with a lower chance of overperforming, an impact that is consonant with previous research. The fifth most important metric is EBITDA Margin, which is associated with a lower chance of overperforming. Even though this may seen counter-intuitive, a higher EBITDA Margin associated with the same P/EBITDA and P/E may mean a riskier business or a difficulty to transform cash-flows into profits, two aspects that could be seen as negative.

While Price is indirectly present in other metrics (especially the price to value ratios), it is also directly present in the form of QoQ growth. A rise in price could mean both momentum (with a positive impact) and a lower potential to further growth (with negative impact). In this model, the second effect outweighs the first, and this feature hurts a stock's chance of overperforming.

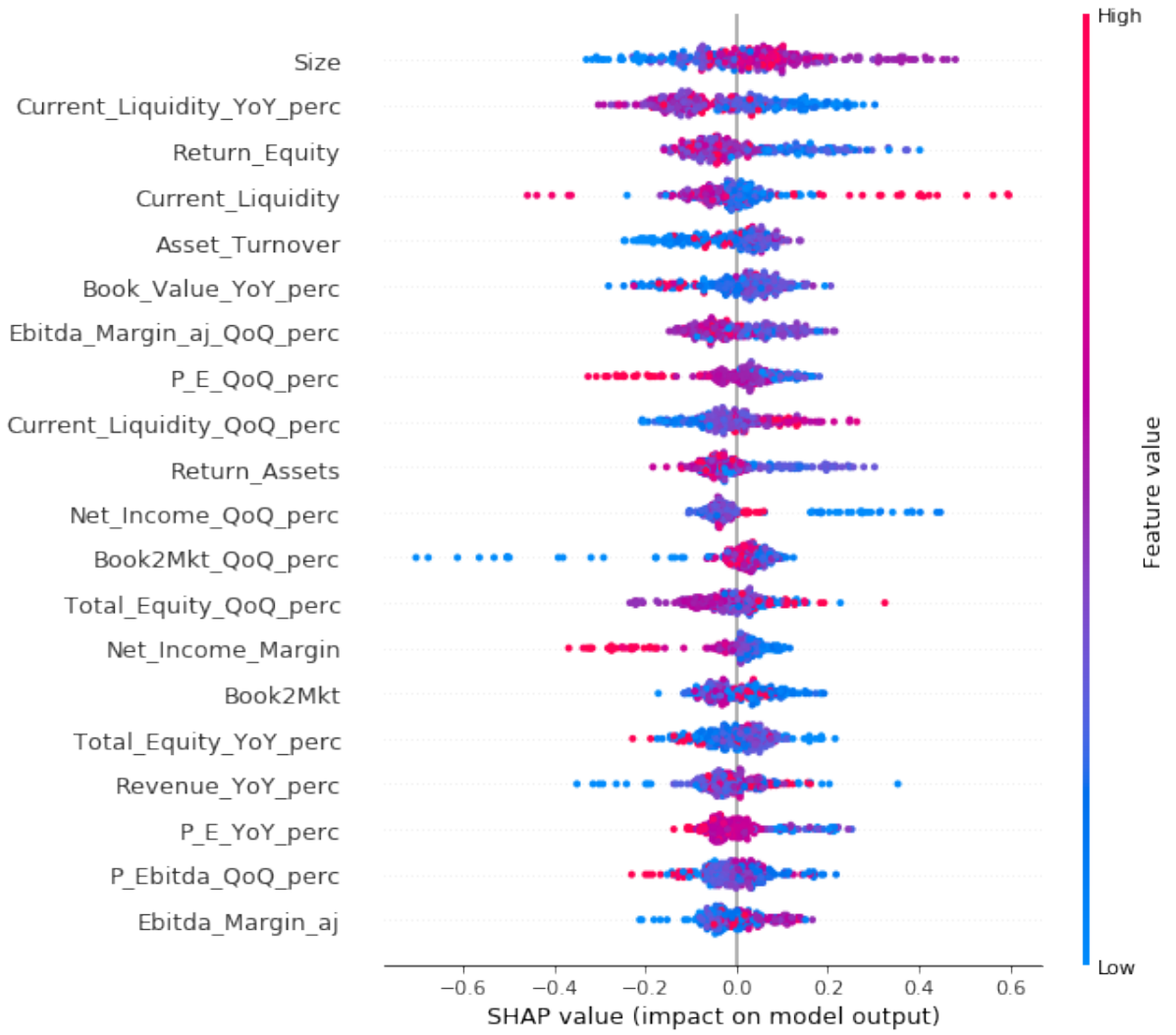
Comparing the Shapley plot of LightGBM's long strategy (Figure 7) with CatBoost's, the top five features are very similar, with only an inversion of P/E and P/EBITDA growth and the change from YoY to QoQ for the latter. The impacts are also similar, which shows consistency between the models. The models also had similar performance in terms of portfolio return. Size is more prominent in LG, taking the sixth position instead of tenth. Both QoQ and YoY growth of price are present, again with a negative impact. Both models also show a positive impact of a growth in EBITDA or its margin (QoQ Margin in LG and YoY of the absolute in CB), even though the impact of a high EBITDA Margin is negative, which further supports the explanation given above. In short, the models are similar in terms of how they build on the features for the prediction, even though the

Figure 7 – Shapley Values for LightGBM - Long Strategy



Shapley values plot relating features' values shown in a color scale from low values (blue) to high values (red) to their respective impact on the predicted variable. Only the top 20 variables are shown. In this model a higher shap value suggests a higher chance of a stock overperforming.
Source: Author

Figure 8 – Shapley Values for LightGBM - Short Strategy



Shapley values plot relating features' values shown in a color scale from low values (blue) to high values (red) to their respective impact in the predicted variable. Only the top 20 variables are shown. In this model a higher shap value suggests a higher chance of a stock underperforming.
Source: Author

underlying black-box algorithms are distinct.

This is not the case, however, for the long and short models. While one might expect that the short model could be built taking the inverse of the long, the Shapley plot in Figure 8 tells another story. First and foremost, P/E and P/EBITDA ratios are not among the top-20 most important features, even though they were key in the long models. Some of these ratios' dynamic components are present, with P/E QoQ growth in seventh place, but with an increase in P/E associated with a lower chance of underperforming. The price-to-value ratios therefore do not work in the same way in the short model.

Even though Size was already relevant in the long positions, it becomes the single most important feature, with bigger companies having higher chances of underperforming. This is consistent both with literature and the conclusions of the long models. ROA and ROE also have the expected negative impact. In the long model ROA growth was more relevant, but ROE takes its place in the short model and ROA appears in its absolute value.

Finally, Current Liquidity, both YoY and absolute, and Asset Turnover gained prominence reaching second, fourth and fifth place, respectively. Both features have the expected impact of lower values associated with a greater chance of underperforming. This difference shows that a difficulty in generating cash flows to maintain a healthy liquidity is more important to predict that a stock will plummet than the valuation metrics that predict it may overperform.

4.2.2 Local Interpretation

While an investor could follow the whole strategy constructed by the model, it is also possible to take each individual recommendation as an input for a stock-picking strategy. In that scenario, the decision to trust the model is therefore made hierarchically. First the investor decides if he globally trusts the model and then if he will follow each of the recommendations based on the output of the model, its interpretation, and other information he may have. The power of a local approach has been shown by Biran and McKeown (2017) and in this dissertation the use of LIME for this end is demonstrated.

For each feature, LIME gives the coefficients of the linear model, which is fitted around the desired data point. They can then be interpreted as contributions to the classification probability score. Since we work with both long and short models, the interpretation of the sign of the coefficients changes accordingly. In the long model a positive score means a higher chance of overperforming and in the short a higher chance of underperforming.

In a real-world application, the investor would normally want to understand all the individual recommendations. This dissertation shows the use of LIME through an example, but the exact same procedure can be used in any other case.

The example presented is Magazine Luiza's (MGLU3) performance in the third quarter of 2017, when it outperformed the Ibovespa. This particular example was chosen because it shows how the valuation ratios' absolute values interact with their YoY and QoQ growth, a dynamic discussed in the last subsection. It is also an interesting example because the company by then had shown the highest price rise of 2016. This fact was widely publicized (one example is this news piece by UOL (2017), a popular internet site in Brazil). There were doubts how long it could still rise, making it an insightful example.

Table 15 shows the ten best ranked features in terms of absolute contribution to the

Table 15 – LIME coefficients for MGLU3's 2017Q3 performance

Feature	Impact
High P/E YoY Growth	5.56%
High P/EBITDA QoQ Growth	2.98%
High Price YoY Growth	-2.97%
High P/EBITDA YoY Growth	2.77%
High P/E	-2.49%
High P/E YoY Growth	2.41%
Low EBITDA Margin	2.14%
High ROE YoY Growth	2.02%
High EBITDA YoY Growth	1.87%
High Total Equity YoY Growth	1.79%

LIME coefficients for LightGBM's prediction of Magazine Luiza's (MGLU3) performance in the third quarter of 2017. Only the ten higher coefficients for numeric features are shown. A positive coefficient indicates that that feature improved the outerforming chances of that stock in that quarter while a negative one lowered its chances.
Source: Author

model's score,[9] which predicted correctly (though by a close margin, the probability was 56%) that MGLU3 would overperform in that quarter. Among the first six features, the conflict between price-to-value and momentum is clear. Growth of P/E and P/EBITDA indicates a chance of overperforming, while an already-high P/E and a Price that had strongly risen recently point the other way.

The two penultimate features, on the other hand, contribute to the overperforming prediction of the model. They show that the price rise came with a better financial performance of the company, with a rise of both ROE and EBITDA.

Based on this information, the investor could then choose to buy the shares, believing that the recent price rise was justified by the company's better performance. Also important, it probably would not create a short position on this scenario, since the model showed a strong probability of large losses in case the stock did rise. In either case, the interpretation helps the investor by not only showing the reasons to invest in that stock but also by reassuring that potential negative factors were also considered. This position is far more comfortable than a list of buy and sell stocks being delivered by a black-box model. Finally, even if the investor blindly believed the model, it would also help him to

[9] The interpretation was based on LightGBM's Long Model.

justify his decision to others. This is of particular importance if he is investing the money of a client or has to report to a boss in a financial firm.

5 CONCLUSION

The main goals of this dissertation were twofold. First we wanted to propose a machine-learning-based trading strategy capable of achieving relevant financial returns using only fundamentalist indicators. Second, we wished to use interpretation methods to open the "black box" which machine learning is associated with and to explain how these results were achieved.

The first goal was achieved through the comparison between three tree-based classification algorithms trained to predict which stocks will have abnormal returns. The use of classification is not new in the literature of stock price prediction[1] even though level prediction is arguably the "natural" choice since price is a continuous variable. Nevertheless the decision to separately model the stocks that will over- and underperform (stocks in the first and last quartiles of returns) is different from other works, that usually define only one target, whether a stock did better or worse than the index. This dissertation was also based only on fundamentalist analysis to predict individual stocks in the under-developed Brazilian stock market, while the literature is stronger in technical analysis (37%) applied to the stock index (60%) for developed markets, with USA and Taiwan representing 47% and 17% each (HENRIQUE; SOBREIRO; KIMURA, 2019).

We compare the accuracy of the three models: Decision Tree (DT), LightGBM (LG) and CatBoost (CB). CatBoost performs better for both long and short models, followed closely by LightGBM. Decision Tree had a significantly worse result. The accuracy was also better for the long than for the short model in all cases.

More importantly, CatBoost achieved an Expected Annualized Excess Return of 19.0% in the long strategy (observed in out-of-sample time-split data), beating LightGBM and the Decision Tree, with 14.6% and 10.0%, respectively. In the short strategy, however, both DT and CB had poor performance (DT was actually negative) while LG achieved an EAR of 10.5%. The superior performance of LightGBM in the short model paved its way to a better performance overall in the long-short strategy, with 26.4% versus 20.0% for CB and 5.1% for DT. Adding transaction costs does not eliminate the gain (a key advantage of a low-turnover strategy), with LG's long-short strategy still achieving an annual return of 22.1%.

The relationship between return and risk was translated into Sharpe ratios. A com-

[1] See Ballings *et al.* (2015) for example.

bination between CB's long model and LG's short achieves 0.50. Machine-learning models often exploit economically unfeasible positions (AVRAMOV; CHENG; METZKER, 2019), a characteristic that this dissertation does not share. In this setting, the results are comparable with the model of Gu *et al.* (2018) with restrictions, which achieves Sharpe ratios between 0.45 and 0.64.

The portfolios were also assessed using a Fama-French five-factor model. All models have risk premiums of the desired sign, though only two were statistically significant (CB long and DT short). While the main text dwells on the difficulties of robust estimation in Brazil, one additional finding of the regression is that the models show a progression in terms of market risk. CB shows higher βs and DT has lower ones.

Two ensemble strategies were also tested: among different algorithms and using the long and short legs of the same algorithm. Ensembling algorithms yields no benefit, since the results of the ensembles fall between those of the original models. The same cannot be said of the second strategy. Combining the overperforming and underperforming models based on CatBoost achieved an EAR of 31.7% and 17.8% for the long and short legs, respectively. Although LG also experienced a gain in the long model, its short model and both DT's models had worse performance. This is a consequence of CB having more consistent performance in different percentiles (Figure 3), while LG is particularly strong in the first quartile, which was defined as the target. The Sharpe ratios for the long-short ensembles were also calculated, with CB achieving 1.26 in its long-short portfolio based on the long-short ensembles.

The second goal, opening the black box, was achieved through the use of global and local interpretation methods. Shapley values were used to first understand what the key features of each model were and how they contributed to the predictions. In the long position the price-to-value metrics prevail, with low P/E and P/EBITDA and high book-to-market values signaling a chance of overperforming. Interestingly, a QoQ or YoY rise in P/E or P/EBITDA also improved the chances of a superior result, hinting at a momentum factor. A YoY growth in ROA and a small size were also among the most important features.

The best short model, LightGBM, was also interpreted using Shapley Values but with different results. Size becomes the most relevant feature, with larger companies presenting higher chances of underperforming. Current Liquidity gains prominence, appearing in both in the second and fourth places with its YoY growth and absolute value, respectively. The absence of valuation metrics such as P/E from the first positions is also striking, especially considering how relevant they are in the long model.

Finally, an illustrative example was chosen for the use of local interpretation through LIME. The case is of Magazine Luiza's stock. In the prior years, it had developed a track record of impressive price rise. Some questioned if it had reached its peak and would fall (since it was already "expensive" in the eyes of some analysts) or if there was still room for further growth. The correct prediction that MGLU3 would overperform based on its 2017Q3 result's sheet indicators is analyzed. The LIME-based interpretation shows the dynamic between the absolute P/E and P/EBITDA ratios and their respective growth rates. In the case of Magazine Luiza, its stock performance was accompanied by a rise in both EBITDA and ROE wich influenced the prediction of overperformance even though its price had already risen significantly and its P/E ratio was high.

Both the financial returns and model interpretation show key differences between the long and short legs. This difference means that the problem of predicting which stocks will underperform is not directly complementary to predicting which will overperform. Predicting underperformance has different relevant features and yields a worse accuracy and financial return. These conclusions reinforce the choice to model long and short positions separately. This is one one of the key findings of this dissertation.

Further research on this topic could analyze other stocks that do not belong to the Ibovespa Index and see if the difference between long and short models persists. Other ensembles could also be tested, since those seen here do not have good results in general. The construction of weighted portfolios using the classification probability score is another option, to pursue even higher returns.

Since one of the restrictions of this dissertation was using only the published results sheet indicators, an interesting evolution would be to apply Natural Language Processing to the actual text of the release. An investor's report usually has explanations and messages from the company. While an analyst would normally consider those, the model developed here is restricted to the "hard numbers." It would therefore be interesting to see if a NLP model could bring other features to the mix. The use of speech-to-text tools together with NLP could also bring information from the results call that companies have after release, with a similar goal.

BIBLIOGRAPHY

ALPHABET INC. *Google Trends search for "Machine Learning", period set to last five years*. 2020. Available at: <https://trends.google.com/>. Accessed on: 06 May 2020. 23

ARAÚJO, R. C. C.; MACHADO, M. A. V. Book-to-market ratio, return on equity and Brazilian stock returns. *RAUSP Management Journal*, Emerald Publishing Limited, v. 53, n. 3, p. 324–344, 2018. 27

ARIDA, P.; BACHA, E.; LARA-RESENDE, A. Credit, interest, and jurisdictional uncertainty: Conjectures on the case of Brazil. *Inflation targeting, debt, and the Brazilian experience, 1999 to 2003*, The MIT Press Cambridge, MA, p. 265–293, 2005. 62

ARIYO, A. A.; ADEWUMI, A. O.; AYO, C. K. Stock price prediction using the ARIMA model. In: IEEE. *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*. [S.l.], 2014. p. 106–112. 28

ATSALAKIS, G. S.; VALAVANIS, K. P. Surveying stock market forecasting techniques–part II: Soft computing methods. *Expert Systems with Applications*, Elsevier, v. 36, n. 3, p. 5932–5941, 2009. 27

AVRAMOV, D.; CHENG, S.; METZKER, L. Machine learning versus economic restrictions: Evidence from stock return predictability. *Available at SSRN 3450322*, 2019. 11, 13, 23, 26, 31, 35, 62, 63, 68, 78

AZOFF, E. M. *Neural network time series forecasting of financial markets*. [S.l.]: John Wiley & Sons, Inc., 1994. 28

BALLINGS, M. et al. Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, Elsevier, v. 42, n. 20, p. 7046–7056, 2015. 30, 35, 77

BIRAN, O.; MCKEOWN, K. R. Human-centric justification of machine learning predictions. In: *IJCAI*. [S.l.: s.n.], 2017. p. 1461–1467. 23, 33, 34, 73

BLACK, F. Capital market equilibrium with restricted borrowing. *The Journal of Business*, JSTOR, v. 45, n. 3, p. 444–455, 1972. 27, 28

BLOOMBERG. *Bloomberg Professional*. 2019. Available at: Subscription Service. Accessed on: 12 Nov. 2019. 38

BLOOMBERG. *Why Machine Learning Hasn't Made Investors Smarter*. 2019. Available at: <https://www.bloomberg.com/news/articles/2019-07-11/why-machine-learning-hasn-t-made-investors-smarter-quicktake>. Accessed on: 11 Apr. 2020. 11, 13, 23

CAVALCANTE, R. C. et al. Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, Elsevier, v. 55, p. 194–211, 2016. 27

CHEN, L.; PELGER, M.; ZHU, J. Deep learning in asset pricing. *Available at SSRN 3350138*, 2019. 30, 31, 34

CHINCO, A.; CLARK-JOSEPH, A. D.; YE, M. Sparse signals in the cross-section of returns. *The Journal of Finance*, Wiley Online Library, v. 74, n. 1, p. 449–492, 2019. 29

FAMA, E. F. Efficient capital markets: II. *The Journal of Finance*, Wiley Online Library, v. 46, n. 5, p. 1575–1617, 1991. 27

FAMA, E. F.; FRENCH, K. R. The cross-section of expected stock returns. *The Journal of Finance*, Wiley Online Library, v. 47, n. 2, p. 427–465, 1992. 27, 41

FAMA, E. F.; FRENCH, K. R. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, Elsevier, v. 33, n. 1, p. 3–56, 1993. 27, 28, 63

FAMA, E. F.; FRENCH, K. R. Size and book-to-market factors in earnings and returns. *The Journal of Finance*, Wiley Online Library, v. 50, n. 1, p. 131–155, 1995. 27

FENG, G.; GIGLIO, S.; XIU, D. Taming the factor zoo. *Chicago Booth research paper*, n. 17-04, 2017. 28

FENG, G.; POLSON, N. G.; XU, J. Deep factor alpha. *arXiv preprint arXiv:1805.01104*, v. 2, 2018. 28

FRAZZINI, A.; ISRAEL, R.; MOSKOWITZ, T. J. Trading costs. *Available at SSRN 3229719*, 2018. 61

GILPIN, L. H. et al. Explaining explanations: An overview of interpretability of machine learning. In: IEEE. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. [S.l.], 2018. p. 80–89. 23, 32

GIOVANNETTI, B. C. et al. *Risk premia estimation in Brazil: Wait until 2041.* [S.l.], 2016. 64

GU, S.; KELLY, B.; XIU, D. *Empirical Asset Pricing via Machine Learning.* [S.l.], 2018. (Working Paper Series, 25398). 24, 26, 29, 30, 31, 34, 56, 62, 63, 67, 78

HENRIQUE, B. M.; SOBREIRO, V. A.; KIMURA, H. Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, Elsevier, 2019. 28, 31, 32, 35, 77

KAASTRA, I.; BOYD, M. Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, Elsevier, v. 10, n. 3, p. 215–236, 1996. 28

KOZAK, S.; NAGEL, S.; SANTOSH, S. Shrinking the cross-section. *Journal of Financial Economics*, Elsevier, v. 135, n. 2, p. 271–292, 2020. 29, 30

LABOISSIERE, L. A.; FERNANDES, R. A.; LAGE, G. G. Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks. *Applied Soft Computing*, Elsevier, v. 35, p. 66–74, 2015. 31

LAHMIRI, S. Entropy-based technical analysis indicators selection for international stock markets fluctuations prediction using support vector machines. *Fluctuation and Noise Letters*, World Scientific, v. 13, n. 02, p. 1450013, 2014. 27

LEUNG, M. T.; DAOUK, H.; CHEN, A.-S. Forecasting stock indices: A comparison of classification and level estimation models. *International Journal of Forecasting*, Elsevier, v. 16, n. 2, p. 173–190, 2000. 30, 37

LINTNER, J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, v. 47, n. 1, p. 13–37, 1965. 27

LIPTON, Z. C. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016. 33

MALKIEL, B. G.; FAMA, E. F. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, Wiley Online Library, v. 25, n. 2, p. 383–417, 1970. 27

MANYIKA, J.; BUGHIN, J. The promise and challenge of the age of artificial intelligence. *McKinsey Global Institute Executive Briefing*, 2018. 32

MONDAL, P.; SHIT, L.; GOSWAMI, S. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, Academy & Industry Research Collaboration Center (AIRCC), v. 4, n. 2, p. 13, 2014. 28

MORITZ, B.; ZIMMERMANN, T. Tree-based conditional portfolio sorts: The relation between past and future stock returns. *Available at SSRN 2740751*, 2016. 30, 34, 56

NASSERI, A. A.; TUCKER, A.; CESARE, S. de. Quantifying stocktwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Systems with Applications*, Elsevier, v. 42, n. 23, p. 9192–9210, 2015. 31

NEFIN. 2020. Available at: <http://www.nefin.com.br/risk_factors.html>. Accessed on: 24 Mar. 2020. 63

OLIVEIRA, N.; CORTEZ, P.; AREAL, N. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, Elsevier, v. 73, p. 125–144, 2017. 31

RAPOSO, R.; CRUZ, A. D. O. Stock market prediction based on fundamentalist analysis with fuzzy-neural networks. In: *Proceedings of 3rd WSES International Conference on Fuzzy Sets*. [S.l.: s.n.], 2002. 31, 41

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Why should I trust you?: Explaining the predictions of any classifier. In: ACM. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.], 2016. p. 1135–1144. 34, 52

SHAPLEY, L. S. A value for n-person games. *Contributions to the Theory of Games*, v. 2, n. 28, p. 307–317, 1953. 34

SHARPE, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, Wiley Online Library, v. 19, n. 3, p. 425–442, 1964. 27

SI, J. et al. Exploiting social relations and sentiment for stock prediction. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1139–1145. 31

SIRIGNANO, J.; CONT, R. Universal features of price formation in financial markets: Perspectives from deep learning. *Quantitative Finance*, Taylor & Francis, v. 19, n. 9, p. 1449–1459, 2019. 29

SIRIGNANO, J. A. Deep learning for limit order books. *Quantitative Finance*, Taylor & Francis, v. 19, n. 4, p. 549–570, 2019. 29

S&P CAPITAL IQ. *Capital IQ Database*. 2019. Available at: Subscription Service. Accessed on: 16 Nov. 2019. 42, 87

STRUMBELJ, E.; KONONENKO, I. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, v. 11, n. Jan, p. 1–18, 2010. 34, 35, 52

TAY, F. E.; CAO, L. Application of support vector machines in financial time series forecasting. *Omega*, Elsevier, v. 29, n. 4, p. 309–317, 2001. 27

TSAY, R. S. Financial time series. *Wiley StatsRef: Statistics Reference Online*, Wiley Online Library, p. 1–23, 2014. 28

UOL. *Magazine Luiza ganha 502% na Bolsa em 2016*. 2017. Available at: <https://economia.uol.com.br/cotacoes/noticias/redacao/2017/01/02/magazine-luiza-ganha-502-na-bolsa-em-2016-veja-as-acoes-que-mais-subiram.htm>. Accessed on: 05 Apr. 2020. 73

VELLIDO, A.; MARTÍN-GUERRERO, J. D.; LISBOA, P. J. Making machine learning models interpretable. In: CITESEER. *ESANN*. [S.l.], 2012. v. 12, p. 163–172. 33, 68

WALL STREET JOURNAL. *Use AI for Picking Stocks? Not So Fast*. 2020. Available at: <https://www.wsj.com/amp/articles/use-ai-for-picking-stocks-not-so-fast-11578279960>. Accessed on: 11 Apr. 2020. 23

ZHANG, G. P. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, Elsevier, v. 50, p. 159–175, 2003. 28

APPENDIX A – LIST OF COMPANIES

Table 16 presents the companies and respective exchange tickers used for trading on the Brazilian stock market.

Table 16 – Companies and Exchange Tickers

Company Name	**Ticker**
Ambev S.A.	ABEV3
Azul S.A.	AZUL4
B2W - Companhia Digital	BTOW3
B3 S.A. - Brasil, Bolsa, Balcão	B3SA3
Banco Bradesco S.A.	BBDC4
Banco do Brasil S.A.	BBAS3
Banco Santander (Brasil) S.A.	SANB4
BB Seguridade Participaçõe S.A.	BBSE3
BR Malls Participações S.A.	BRML3
Bradespar S.A.	BRAP3
Braskem S.A.	BRKM5
BRF S.A.	BRFS3
CCR S.A.	CCRO3
Centrais Elétricas Brasileiras S.A. - Eletrobrás	ELET6
Cielo S.A.	CIEL3
Companhia Brasileira de Distribuição	PCAR4
Companhia de Saneamento Básico do Estado de São Paulo - SABESP	SBSP3
Companhia Energética de Minas Gerais	CMIG4
Companhia Siderúrgica Nacional	CSNA3
Cosan S.A.	CSAN3
CVC Brasil Operadora e Agência de Viagens S.A.	CVCB3
Cyrela Brazil Realty S.A. Empreendimentos e Participações	CYRE3
EcoRodovias Infraestrutura e Logística S.A.	ECOR3
EDP - Energias do Brasil S.A.	ENBR3
Embraer S.A.	EMBR3
Engie Brasil Energia S.A.	EGIE3
Equatorial Energia S.A.	EQTL3
Estácio Participações S.A.	YDUQ3
Fleury S.A.	FLRY3
Gerdau S.A.	GGBR4
Gol Linhas Aéreas Inteligentes S.A.	GOLL4
Hypera S.A.	HYPE3

Continued on next page

Table 16 – *Continued from previous page*

Company Name	**Ticker**
Iguatemi Empresa de Shopping Centers S.A.	IGTA3
IRB-Brasil Resseguros S.A.	IRBR3
Itaú Unibanco Holding S.A.	ITUB4
Itaúsa - Investimentos Itaú S.A.	ITSA4
JBS S.A.	JBSS3
Klabin S.A.	KLBN4
Kroton Educacional S.A.	KROT3
Localiza Rent a Car S.A.	RENT3
Lojas Americanas S.A.	LAME4
Lojas Renner S.A.	LREN3
Magazine Luiza S.A.	MGLU3
Marfrig Global Foods S.A.	MRFG3
Metalurgica Gerdau S.A.	GOAU4
MRV Engenharia e Participações S.A.	MRVE3
Multiplan Empreendimentos Imobiliários S.A.	MULT3
Natura Cosméticos S.A.	NATU3
Petrobras Distribuidora S.A.	BRDT3
Petróleo Brasileiro S.A. - Petrobras	PETR4
Qualicorp Consultoria e Corretora de Seguros S.A.	QUAL3
Raia Drogasil S.A.	RADL3
Rumo S.A.	RAIL3
Smiles Fidelidade S.A.	SMLS3
Suzano S.A.	SUZB3
Telefônica Brasil S.A.	VIVT4
TIM Participações S.A.	TIMP3
Transmissora Aliança de Energia Elétrica S.A.	TAEE11
Ultrapar Participações S.A.	UGPA3
Usinas Siderúrgicas de Minas Gerais S.A.	USIM5
Vale S.A.	VALE3
Via Varejo S.A.	VVAR3
WEG S.A.	WEGE3

Source: Author

APPENDIX B – LIST OF SECTORS AND INDUSTRIES

Table 17 presents the companies' tickers and respective sectors and industries considered in this work. The list is based on S&P Capital IQ's (2019) Database classification.

Table 17 – Companies, Sectors and Industries

Ticker	Primary Sector	Primary Industry
ABEV3	Consumer Staples	Brewers
AZUL4	Industrials	Airlines
BTOW3	Consumer Discretionary	Internet and Direct Marketing Retail
B3SA3	Financials	Financial Exchanges and Data
BBDC4	Financials	Diversified Banks
BBAS3	Financials	Diversified Banks
SANB4	Financials	Diversified Banks
BBSE3	Financials	Multi-line Insurance
BRML3	Real Estate	Real Estate Operating Companies
BRAP3	Materials	Steel
BRKM5	Materials	Commodity Chemicals
BRFS3	Consumer Staples	Packaged Foods and Meats
CCRO3	Industrials	Highways and Railtracks
ELET6	Utilities	Electric Utilities
CIEL3	Information Technology	Data Processing and Outsourced Services
PCAR4	Consumer Staples	Hypermarkets and Super Centers
SBSP3	Utilities	Water Utilities
CMIG4	Utilities	Electric Utilities
CSNA3	Materials	Steel
CSAN3	Energy	Oil and Gas Refining and Marketing
CVCB3	Consumer Discretionary	Hotels, Resorts and Cruise Lines
CYRE3	Consumer Discretionary	Homebuilding
ECOR3	Industrials	Highways and Railtracks
ENBR3	Utilities	Electric Utilities
EMBR3	Industrials	Aerospace and Defense
EGIE3	Utilities	Renewable Electricity
EQTL3	Utilities	Electric Utilities
YDUQ3	Consumer Discretionary	Education Services
FLRY3	Health Care	Health Care Services

Continued on next page

Table 17 – *Continued from previous page*

Ticker	Primary Sector	Primary Industry
GGBR4	Materials	Steel
GOLL4	Industrials	Airlines
HYPE3	Health Care	Pharmaceuticals
IGTA3	Real Estate	Real Estate Operating Companies
IRBR3	Financials	Reinsurance
ITUB4	Financials	Diversified Banks
ITSA4	Financials	Diversified Banks
JBSS3	Consumer Staples	Packaged Foods and Meats
KLBN4	Materials	Paper Packaging
KROT3	Consumer Discretionary	Education Services
RENT3	Industrials	Trucking
LAME4	Consumer Discretionary	General Merchandise Stores
LREN3	Consumer Discretionary	Department Stores
MGLU3	Consumer Discretionary	General Merchandise Stores
MRFG3	Consumer Staples	Packaged Foods and Meats
GOAU4	Materials	Steel
MRVE3	Consumer Discretionary	Homebuilding
MULT3	Real Estate	Real Estate Operating Companies
NATU3	Consumer Staples	Personal Products
BRDT3	Consumer Discretionary	Automotive Retail
PETR4	Energy	Integrated Oil and Gas
QUAL3	Health Care	Managed Health Care
RADL3	Consumer Staples	Drug Retail
RAIL3	Industrials	Railroads
SMLS3	Communication Services	Advertising
SUZB3	Materials	Paper Products
VIVT4	Communication Services	Integrated Telecommunication Services
TIMP3	Communication Services	Wireless Telecommunication Services
TAEE11	Utilities	Electric Utilities
UGPA3	Energy	Oil and Gas Storage and Transportation
USIM5	Materials	Steel
VALE3	Materials	Steel
VVAR3	Consumer Discretionary	Home Improvement Retail
WEGE3	Industrials	Electrical Components and Equipment

Source: Author

APPENDIX C – KEY STATISTICS OF NUMERIC FEATURES

Table 18 presents key statistics of all the features used in the model. This is a complete version of Table 5 shown in Section 3.2.

Table 18 – Descriptive Statistics for all Model Features

Metric	**mean**	**std**	**min**	**Q1**	**Q2**	**Q3**	**max**	**iqr**	**range**
Asset Turnover	0.78	0.64	0.01	0.39	0.60	0.97	4.62	0.58	4.61
Book-to-Market	1.09	1.47	-0.08	0.37	0.70	1.26	30.13	0.89	30.21
Current Liquidity Ratio	1.81	1.83	0.34	1.15	1.56	2.07	36.20	0.92	35.86
EBITDA Margin	27.75	23.57	-169.10	11.30	22.00	39.40	324.10	28.10	493.20
Net Margin	13.68	31.61	-249.70	2.28	9.14	18.40	741.30	16.12	991.00
Price/Earnings	695.02	31,473.37	-36,779.59	16.06	39.83	77.37	1,576,513.35	61.31	1,613,292.94
Price/EBITDA	24.77	233.82	-10,467.70	9.75	19.98	36.29	3,284.50	26.54	13,752.20
Return on Assets	6.76	5.74	-35.90	3.36	5.69	9.18	62.70	5.82	98.60
Return on Equity	10.23	282.92	-14,083.90	5.65	13.70	24.80	200.90	19.15	14,284.80
Size	3.32	0.59	1.56	2.92	3.31	3.69	4.95	0.77	3.39
Asset Turnover QoQ growth	0.06	1.61	-0.97	-0.07	0.01	0.08	78.50	0.15	79.47
Asset Turnover YoY growth	0.04	0.70	-0.98	-0.10	0.00	0.10	26.60	0.20	27.58
Book-to-Market QoQ growth	0.03	0.43	-1.70	-0.13	-0.01	0.12	16.49	0.25	18.19
Book-to-Market YoY growth	0.06	0.63	-1.19	-0.24	-0.05	0.22	11.28	0.46	12.47
Book Value QoQ growth	0.04	0.48	-1.90	-0.01	0.02	0.05	21.77	0.06	23.67
Book Value YoY growth	0.14	0.60	-1.27	0.00	0.08	0.18	24.14	0.18	25.41
Current Liquidity Ratio QoQ growth	0.06	0.88	-0.97	-0.08	0.00	0.09	29.28	0.17	30.25
Current Liquidity Ratio YoY growth	0.08	0.53	-0.95	-0.14	0.00	0.19	9.98	0.33	10.93
Earnings QoQ growth	6.39	312.10	-806.63	-0.48	-0.03	0.36	15,596.50	0.84	16,403.13
Earnings YoY growth	8.45	303.90	-165.45	-0.41	0.05	0.47	14,155.00	0.88	14,320.45

Continued on next page

Table 18 – *Continued from previous page*

Metric	mean	std	min	Q1	Q2	Q3	max	iqr	range
EBITDA QoQ growth	0.19	1.85	-15.55	-0.15	0.03	0.22	44.24	0.37	59.79
EBITDA YoY growth	0.44	5.84	-21.33	-0.07	0.13	0.37	271.24	0.44	292.57
EBITDA Margin QoQ growth	0.12	1.83	-14.09	-0.14	-0.01	0.13	39.10	0.27	53.19
EBITDA Margin YoY growth	0.13	2.23	-23.05	-0.14	-0.01	0.14	70.51	0.28	93.56
Net Margin QoQ growth	4.31	208.30	-521.00	-0.48	-0.06	0.29	10,399.00	0.77	10,920.00
Net Margin YoY growth	7.16	270.69	-206.56	-0.48	-0.07	0.26	12,799.00	0.74	13,005.56
Price QoQ growth	0.06	0.23	-0.60	-0.07	0.03	0.16	2.88	0.23	3.48
Price YoY growth	0.29	1.11	-0.85	-0.09	0.15	0.45	30.12	0.54	30.97
Price/Earnings QoQ growth	5.10	232.67	-719.15	-0.52	-0.08	0.37	11,566.01	0.89	12,285.16
Price/Earnings YoY growth	7.39	338.18	-1,066.97	-0.49	-0.03	0.44	16,834.41	0.93	17,901.38
Price/EBITDA QoQ growth	0.16	2.89	-60.35	-0.22	0.00	0.27	71.98	0.49	132.33
Price/EBITDA YoY growth	0.33	3.06	-21.75	-0.24	0.04	0.39	106.76	0.63	128.51
Return on Assets QoQ growth	0.41	5.31	-48.58	-0.25	-0.01	0.25	165.82	0.50	214.40
Return on Assets YoY growth	1.39	58.36	-189.33	-0.27	-0.01	0.27	2,906.69	0.54	3,096.02
Return on Equity QoQ growth	0.78	39.91	-149.10	-0.50	-0.07	0.34	1,969.00	0.84	2,118.10
Return on Equity YoY growth	-0.87	47.33	-2,157.80	-0.51	-0.09	0.30	741.86	0.81	2,899.66
Revenue QoQ growth	0.10	1.68	-0.97	-0.04	0.03	0.11	81.02	0.15	81.99
Revenue YoY growth	0.20	0.75	-0.98	0.03	0.13	0.26	30.85	0.23	31.83
Size QoQ growth	0.01	0.04	-0.49	-0.01	0.00	0.01	1.15	0.02	1.64
Size YoY growth	0.02	0.04	-0.49	0.00	0.02	0.03	0.90	0.03	1.39

Continued on next page

Table 18 – *Continued from previous page*

Metric	mean	std	min	Q1	Q2	Q3	max	iqr	range
Total Equity QoQ growth	0.05	0.24	-1.90	-0.01	0.02	0.05	4.36	0.06	6.26
Total Equity YoY growth	0.19	0.51	-1.27	0.01	0.08	0.20	12.51	0.19	13.78

Summary metrics for the key numeric features. Mean refers to the mean value for each variable, std refers to the standard deviation, min and max refer to the minimum and maximum values respectively, Q1 to Q3 represent the quartiles, iqr is the inter-quartile range and "range" is the range itself, *i.e.* the difference between the maximum and minimum values.

Source: Author

APPENDIX D – SHAPLEY VALUES TO MODELS NOT DISCUSSED IN THE MAIN TEXT

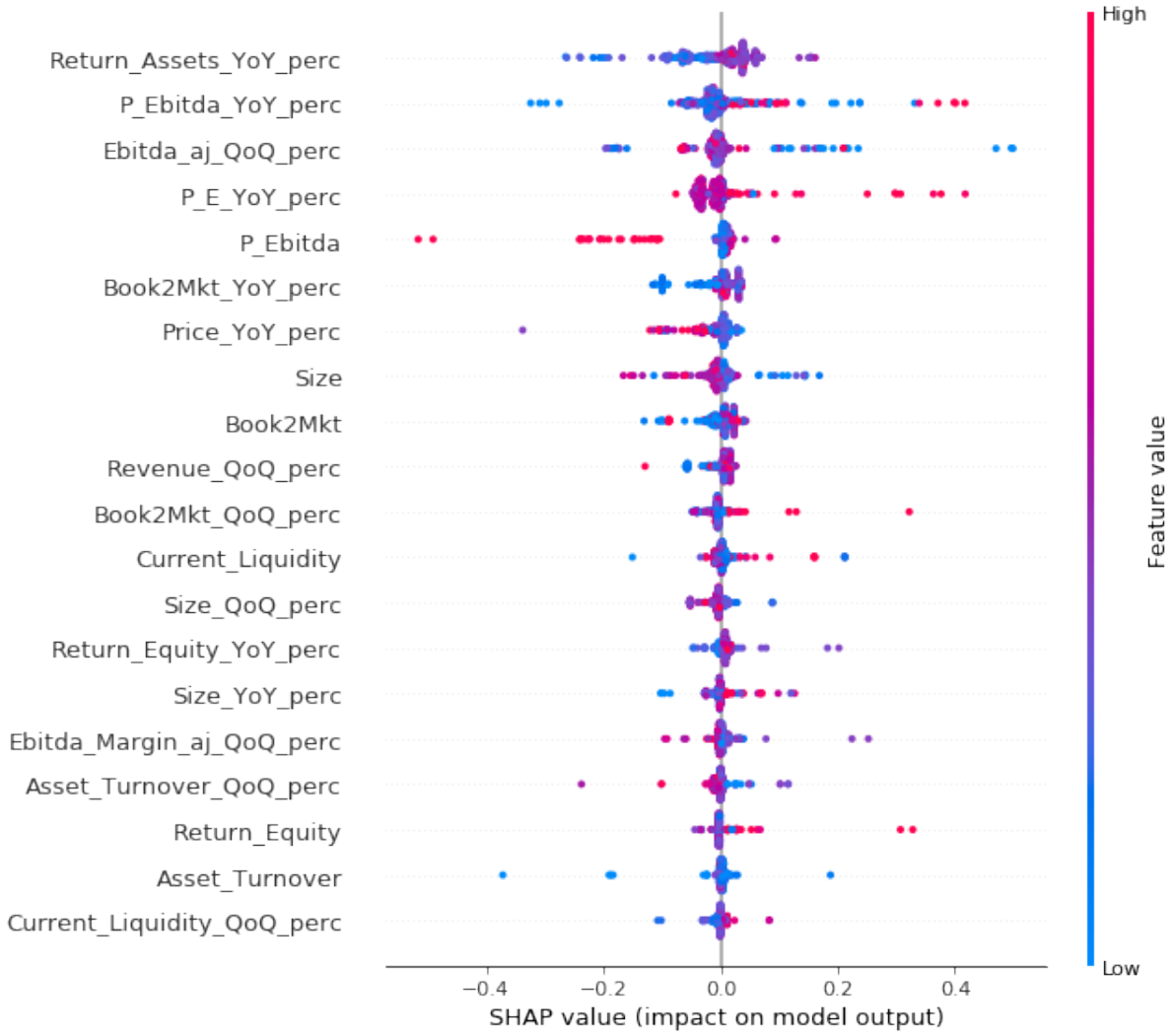
Section 4.2 focused only in 3 of the 6 models developed. This appendix presents the shap plots for the remaining models, allowing for a deeper understand of their performance.

Decision Tree's long model is shown in Figure 9. As discussed in the results section, the models' results are closer in the long strategy and the same happens with their features' interpretation. Many of the key feature are the same between DT and the other models, such as company size and YoY growth of ROA, book-to-market ratio and price. But the most remarkable difference of DT is that the P/E ratio is not directly included in the top list, even though it is a traditional fundamentalist ratio. This could explain the poorer performance of this model. Since similar indicator (such as P/EBITDA) are present, one possible explanation is that Decision Tree was not as able as the other models to deal with similar features and multiple correlations among them.

The most important features of Decision Tree's short model are not the ones of the long model. Size is the most important feature for the short strategy, with a larger size associated with a higher chance of under-performing as shown in Figure 10. Net income is not in the top 20 of the long strategy in any of its forms, but its QoQ growth is second in the short model. Also, none of the features associated with a good price-to-value ratio (P/E, P/EBITDA and their growth ratios) appear in the top 10, even though Book-to-Market does figure in fifth.

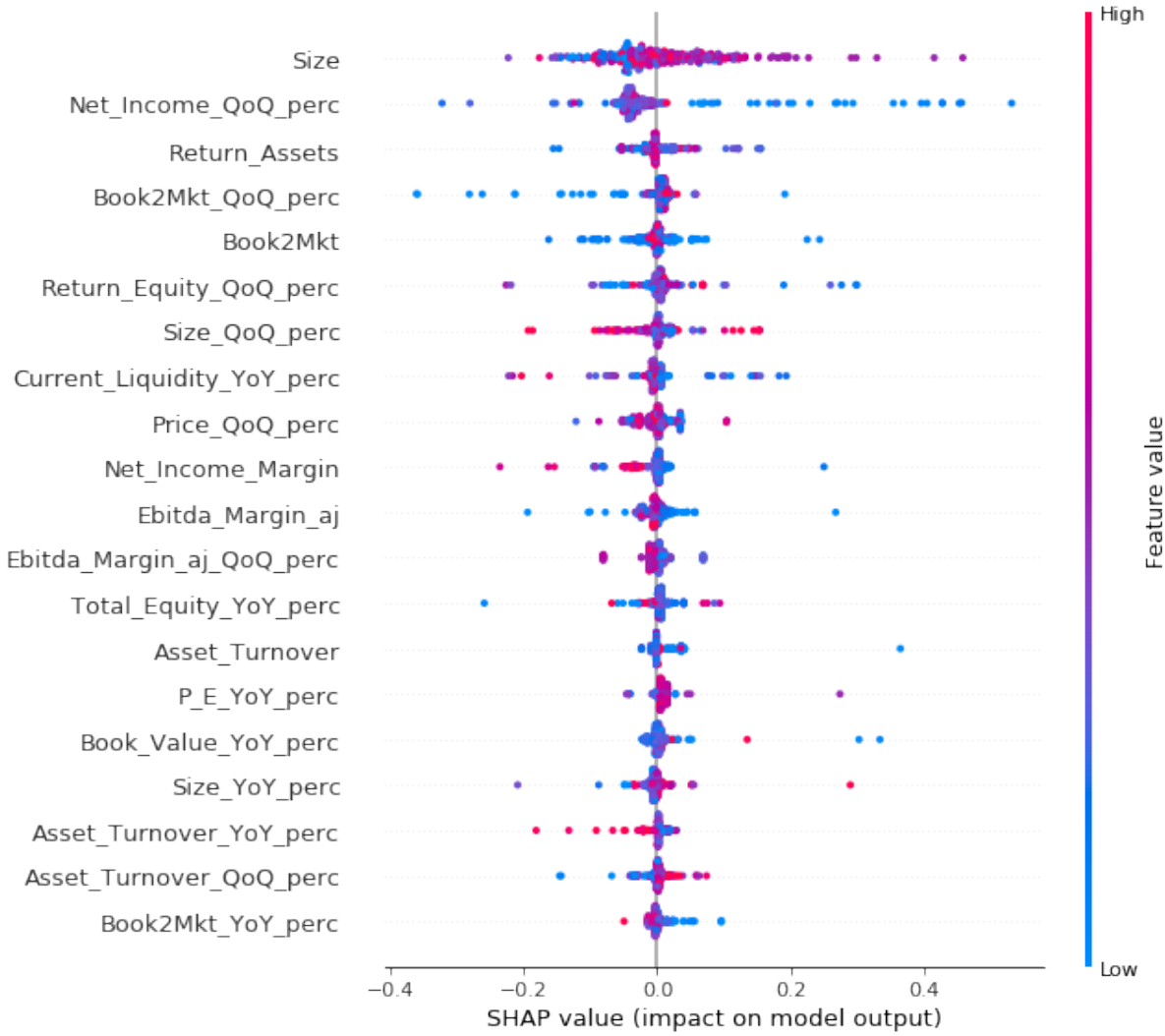
Finally, the shap plot for CatBoost's short model is shown in Figure 11. It heavily depends on the net income and its margin, with four of the top five features based on them. They associate higher margins ow margin growths with lower chances of under-performing. It is also remarkable that both size and current liquidity, two prominent features in LightGBM, do not appear in the top-20 list except for their growths. Finally, P/E has the opposite sign than expected, with higher values associated with smaller chances of underperforming. As a general conclusion, this model does not present the expected features, which might explain its poor performance.

Figure 9 – Shapley Values for the Decision Tree - Long Strategy



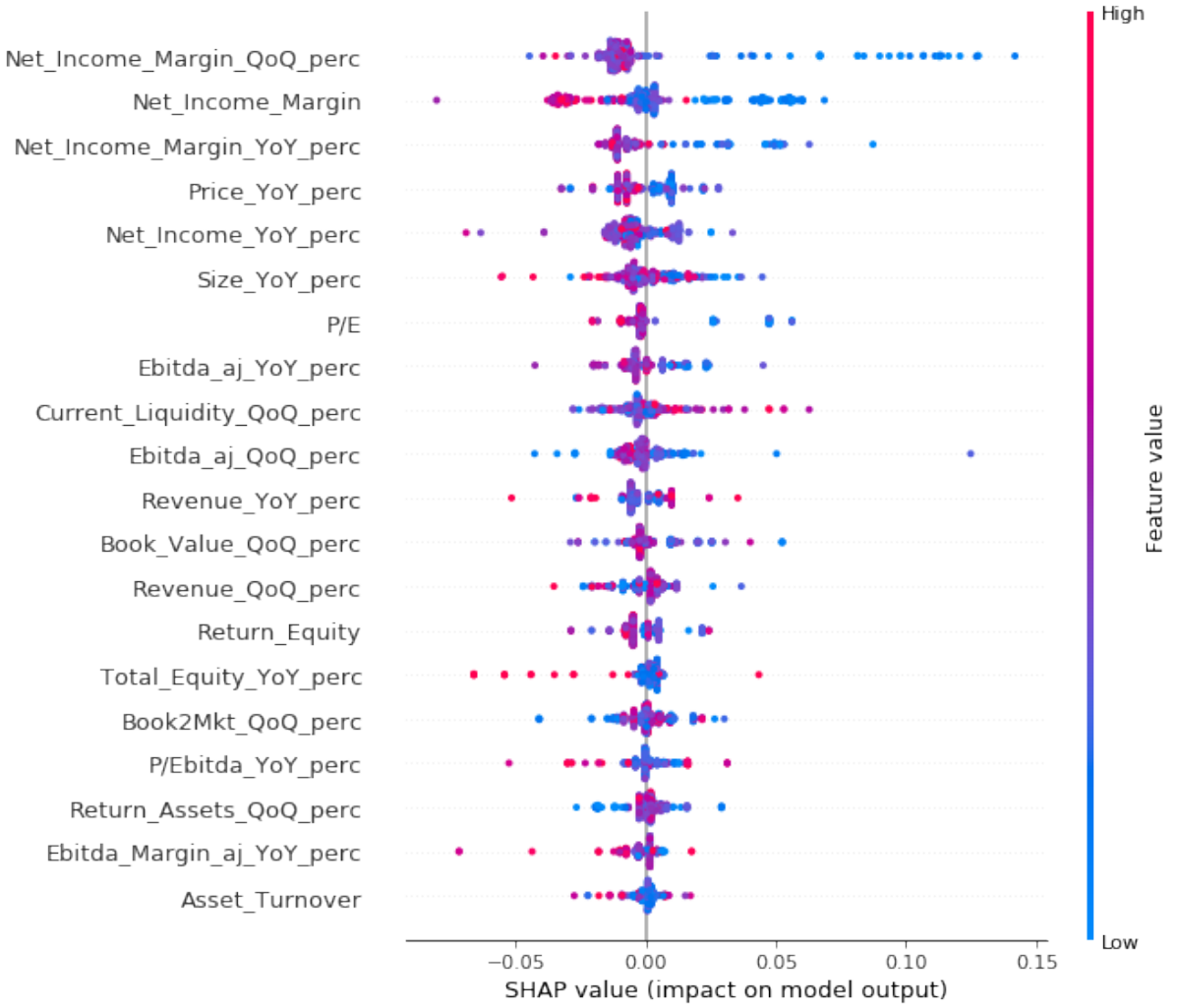
Shapley values plot relating features' values shown in a color scale from low values (blue) to high values (red) to their respective impact in the predicted variable. Only the top 20 variables are shown. In this model a higher shap value implies in a higher chance of a stock over-performing.
Source: Author

Figure 10 – Shapley Values for the Decision Tree - Short Strategy



Shapley values plot relating features' values shown in a color scale from low values (blue) to high values (red) to their respective impact in the predicted variable. Only the top 20 variables are shown. In this model a higher shap value implies in a higher chance of a stock under-performing.
Source: Author

Figure 11 – Shapley Values for CatBoost - Short Strategy



Shapley values plot relating features' values shown in a color scale from low values (blue) to high values (red) to their respective impact in the predicted variable. Only the top 20 variables are shown. In this model a higher shap value implies in a higher chance of a stock under-performing.
Source: Author