Master's Thesis of Engineering

Stock Short Sales Fee Prediction

주식 공매도 비용 예측

August 2022

Graduate School of Engineering
Seoul National University
Industrial Engineering

Sung-Yeon Lim

Stock Short Sales Fee Prediction

Professor Sungzoon Cho

Submitting a master's thesis of
Industrial Engineering

August 2022

Graduate School of Engineering
Seoul National University
Industrial Engineering

Sung-Yeon Lim

Confirming the master's thesis written by
Sung-Yeon Lim
August 2022

Chair	Jaewook Lee	(Seal)
Vice Chair	Sungzoon Cho	(Seal)
Examiner	Woojin Chang	(Seal)

Abstract

Stock short positions in financial investment can be achieved by borrowing and selling stocks. Such activities involve fees including commissions and stock loan fees. Prediction of such fees is valuable in two ways; historical data enables rigorous back-testing of investment strategies, and predicting the future fees contributes to risk management and execution planning. The fees are highly positively skewed, so that the fees are formed around 0 under normal regime. Such stocks are referred to as 'general collateral'. On the other hand, those with abnormally high loan fees are said to be 'special'. As a contribution to the stock short sales fee prediction, the thesis focuses on predicting such specialness via data mining and machine learning techniques. As a result, the models are proposed to predict the specialness, and performance baselines are produced by comparing well-established machine learning techniques.

Keyword: Stock Short Sales, Stock Loan Fee, Machine Learning, Data Mining
Student Number: 2019-24079

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1. Motivation

A short sale in financial investment is one of the widely used methods to construct a short position, from which an investor profits when the value of the financial instrument falls. Increasing short interest indicates execution of bigger short investments, and the need for the market analysis is growing, especially data-driven approaches. Figure 1 pictures the daily average short interest values in dollars, computed from a proprietary database of short interest.

Short selling activity involves borrowing stocks to sell, and the accompanying fees include not only commission fees but also stock borrowing fees. The stock short sales fee could affect the investment returns [4]. Identifying stock short sales fees is useful in several ways.

First, historical short fees are crucial for accurate back-testing. The most important goal of back-testing is to reproduce a hypothetical investment environment as realistically as possible to evaluate investment strategies. Realistic back-testing leads to better investment decisions. Any kind of short selling strategy must account for short fees within back-testing, and sometimes the accuracy of

fees used can lead to such undesired ramifications [4, 18].

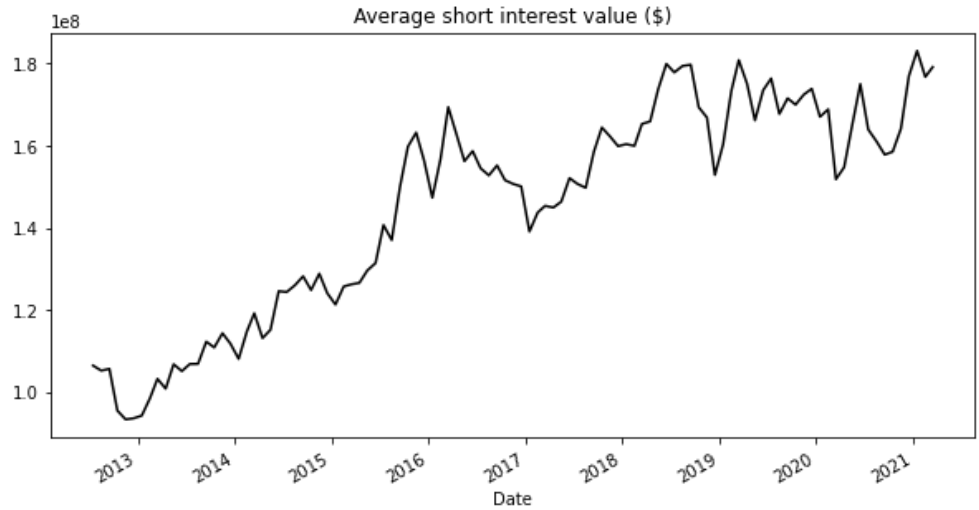


Figure 1: Average Short Interest Value

Second, future short fees are crucial for constructing execution strategies. Fees affect future returns, and a good execution strategy is expected to take the fees under account during optimization process. If the investors would have known the short fees in advance, the quality of trade execution boosts.

Either way, the prediction of the stock short fees is highly in need. In the case of historical short fees, the data is not widely available. Even if one could get access to it, the data does not cover a complete set of assets, and the time horizon is also relatively short, compared with the period conventionally used in back-testing. A model could provide fees for assets under uncovered periods. In the second case, future short fees should be predicted for risk management in executing short trades. This is when the prediction model for the short fee should be put into production.

However, not many data-driven modeling approaches have been taken towards predicting stock short sales fees. In the next section,

we look at the prediction problem and define the scope of the thesis.

1.2. Problem Formulation – Stock Short Fee Regimes

When a stock loan occurs, the borrower posts collateral for the stock to be borrowed to the lender. Normally the value of the collateral is greater than the market value of the borrowed share. In the case of cash collateral, the lender pays the borrower the interest for the collateral. It is called the 'rebate'. A stock short sales fee is determined implicitly by the rebate rate. Normally the stock loan fee is said to be the shortcoming of the rebate from the base rate, such as the Federal Funds rate [10].

In other cases where the collaterals are other assets (e.g. Treasury Bills), the stock loan fee is paid directly from the short seller to the lender. Figure 2 depicts the transactions associated with stock short sales. In the thesis, we refer to the stock loan fee as "stock short sales fee", "short fee", or just "fee".

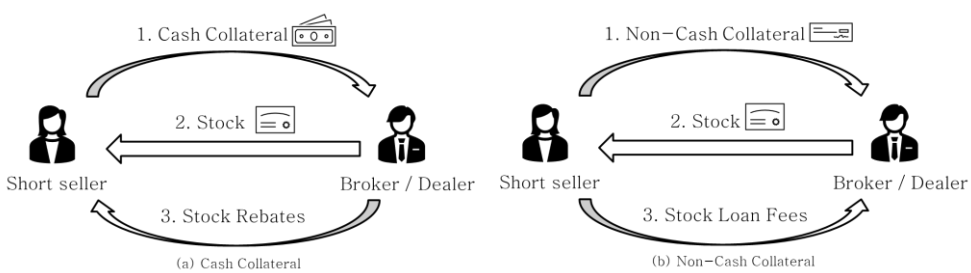


Figure 2: Transactions of Stock Short Sales

The short fees are highly positively skewed. Almost all the time the fees remain at relatively normal, low levels. This is because, in theory, there is an infinite number of stocks to be lent. Borrowed stocks can be listed for loan an infinite number of times [34].

However, on some special occasions, the fees are abnormally high. In works of literature, the normal times are referred to as the "general collateral" (GC) regime, and the abnormal times are referred to as the "special" regime [10].

The behavior of the fees is different depending on which regime a stock is at. Hence, separate models would be employed to predict the stock loan fees under different circumstances. Before utilizing multiple models, a data point should be classified into either of the two regimes. Only then the corresponding model can be used for prediction.

The thesis focuses on this preliminary problem – predicting which regime a data point falls under. Of course, such model serves as preliminary research, located at the earlier part of a bigger prediction pipeline, but it also has a value of its own.

Abnormally high short fees imply a practically low possibility of borrowing stocks. Predicting if a stock will be "special" in the future brings high values to investors so that the investors can plan the execution. Furthermore, most practitioners tend not to even consider shorting assets whose short fees are not in their general collateral regime, unless they are professional short investors. This leads to a discussion of contributions.

1.3. Contributions

The thesis takes data-driven modeling approaches to predict stock short sales fees, implementing machine learning techniques. The main contributions follow two-fold:

1. We provide machine learning modeling approaches to predict the special regime of publicly traded firms. Models built can be used to classify if a company is under such regime.

2. We provide baselines for stock short sales fees and a data-driven understanding of the stock short sales market. Feature importance is analyzed, and the model can be used for future research on predicting short fees with greater accuracy.

The thesis consists of six chapters. In Chapter 2, we review the literature related to the stock short sales market and modeling the stock markets. In Chapter 3 we introduce the framework proposed, including both the predictors and dependent variables. In Chapter 4 we list the models used to construct baseline results. Chapter 5 describes the experimental settings and reports the results with discussions. We conclude the thesis with further discussions in Chapter 6.

Chapter 2

Literature Review

2.1. Stock Short Sales Market

The stock short sales market has been investigated in a wide range of literature [4, 10, 11, 12, 13, 18, 26, 32, 36, 37]. D'Avolio [10] focuses on the shorting activity and fees associated with it, whereas previous studies focused on the short interest (quantity) only. 1% is used as the threshold to tell if a stock is special – stocks whose short loan fees are over the 1% threshold are thought to be special.

The author investigated how the fees are formed in equilibrium and provided the framework to assist that the fees move away from the general collateral regime when the investor opinions diverge. The risk of the borrower not being able to extend the stock loan contract under the same terms, called the recall risk, also builds up in the variance of the spread between optimistic non-lenders and participants of stock loan trades.

Also, it is pointed out that firm size and institutional ownership effects are negatively correlated to the fees, but positively correlated with the proxies for differences of opinion. This includes high share turnover, high dispersion of analyst forecasts, high price multiples,

and low cash flows. The model successfully analyzed the factors correlating with the short fee and we develop the idea to focus more on the actual prediction.

Building upon this, Beneish et. al. [4] tried to model if the stock is special or not, as a part of a two-stage model to identify determinants of lendable stock inventory. The lendable inventory gets affected by the borrowing costs and company characteristics, and the borrowing cost also gets affected by the company characteristics. Moreover, it is shown that short side returns of nine well-known market anomalies can be attributed to the special stocks, and when the short fees are taken into account while back-testing, several of these anomalous returns disappear.

Duffie et. al. [12] suggested mathematical models for the asset value regarding the stock lending market, using market microstructures. It is concluded that if the stock is hard to 'locate' (succeed to agree on the terms of lending), the price of an asset rise initially, even higher than the prospect of all investors, and declines over time. This was explained by the fact that the short fee acts as the source of income for lenders, and the optimists reflect such expectations to the price. Duffie et. al. [12] backed D'Avolio [10] that the differences in investor consensus on stock valuation elevate the short fee, by validating such in the model proposed.

The thesis has a similar interest to the mentioned literature, but the primary focus lies more on the prediction of the specialness itself, while possibly investigating the applicable use of machine learning techniques. We try to produce results that are closer to the practitioners instead of finding economic explanations of the phenomena.

Other literature viewed short interest as one of the variables to

explain stock returns. Often regression analysis and investment simulations are accompanied [11, 13, 36, 37]. Especially, Kot [26] asserted that the shorting activity is positively related to arbitrage opportunities and negatively related to stock returns.

Muravyev et. al. [32] viewed the uncertainty of short fees as the source of risk and calculated the risk premium by using option-implied short fees. The implications from Engelberg et. al. [13] were that the short investors would pay risk premia to avoid the variance in short fees and recall risk. The term stock loan fee was estimated with the option-implied borrowing fee which reflects the risk premium.

The option-implied fee is a reasonable estimator for the indicative short fee, but the purpose is not to predict but to derive a risk premium. In addition, it is also likely that option data would not be widely available for the stocks whose short fees are missing.

This would be the case as they are likely to be the market underdogs, meaning either the market capitalizations are among the bottom percentiles, or/hence the trading volumes are dry. Hence the explicit mining model gets its significance.

The abundance of studies viewed the stock loans from an economic perspective, assisting economic explanation of investor behavior or stock returns. The thesis asserts the value in explicitly predicting stock short sales fees. Accordingly, we contribute by taking a novel approach to machine learning techniques and providing experimental results.

2.2. Modelling the Stock Market

There have been modeling approaches to predict the stock market, namely stock return and volatility. However, there is little literature that tried to predict stock short sales fees, especially using machine learning techniques.

Traditionally financial literature has a huge interest in explaining stock returns. Fama & French [14] has proposed that there are characteristics of firms, called 'factors', which were said to explain the returns of the stock of the firms. Since then, a series of studies tried to expand their work to explain the variance of the residuals [15].

The volatility of stock returns is also highlighted among finance researchers. Volatility is especially important in risk management and portfolio construction. Stocks exhibit volatility clustering, which is the empirical characteristic that volatilities of stock returns are autocorrelated with themselves [30]. Such observations led to the use of statistical time series models such as GARCH (Generalized Autoregressive Conditional Heteroskedasticity) [5]. Hwang et. al. [25] have implemented GARCH-X which incorporates external factors while predicting the volatility.

Such topics are also active in the machine learning community. Various models have been used to predict the return or volatility of stock returns. Company characteristics or technical indicators are normally used as predictors.

Gu et. al. [20] extensively compared several machine learning methods to measure asset risk premia. 94 company characteristics are used, together with industry indicators. Gu et. al. [21] proposed an autoencoder-based model which can learn deep latent factors.

Non-linear conditional latent factor exposures are estimated, and economic discussions were made.

Abe & Nakayama [1] used company characteristics as inputs and deep learning models to predict cross-sectional stock returns one month ahead. They used five points of time for 25 factors which adds up to 125 input features. Alonso-Monsalve et. al. [2] used 18 market technical indicators and deep learning models to predict the direction of crypto assets. Mohan et. al. [31] and Vargas et. al. [40] made use of alternative unstructured data is used as well as other commonly used company characteristics for stock market prediction.

Christensen el. al. [8], Feng et. al. [15], and Yang et. al. [42] used company characteristics, macro indicators, and a range of unstructured data with machine learning models to predict stock returns volatility. The thesis makes use of the techniques established by the literature and expands them to the stock loan fee prediction.

2.3. Other Literature

Other literature includes Nashikkar & Pedersen [33] which claims that the specialness of corporate bonds and equity are correlated, so that given that a firm's stocks are special, its bonds are likely to be special and vice versa. As bad credit indices support bond specialness, we tried to include credit-related features to predict the specialness.

Psillaki et. al. [35] produced an early-warning model for evaluating credit default risk. While proposing a novel efficiency measuring technique, hints are found on what credit-related features could be. They are grouped as firm size, profitability, liquidity, leverage, turnover, or tangible collateral.

Especially, Moody's KMV model is widely used to assess credit risk by estimating the default probability based on market data [39]. The thesis incorporates the features that appeared in the above literature to construct a concise but information-wise extensive set of features.

Chapter 3

Proposed Framework

The thesis presents machine learning models as baselines for predicting the special regime. Since machine learning models are built upon data, we describe the predictor and target variables used to build models. Stock short sales fee data is used as the dependent variable, whereas company characteristics and asset characteristics comprise the input data. A total of 1.3M data points are used for the analysis. The overall framework is depicted in Figure 3.

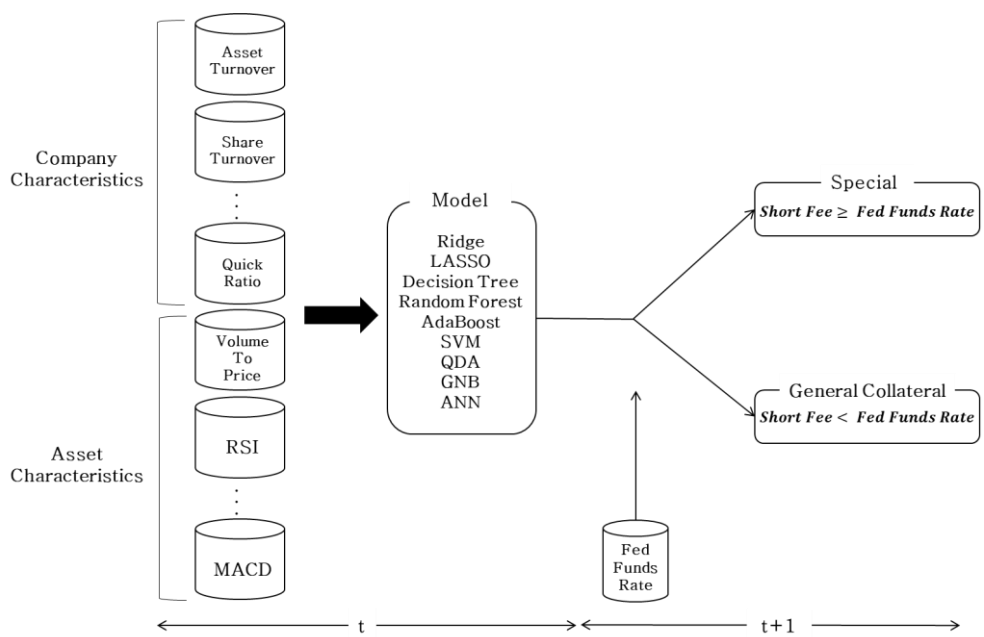


Figure 3: Proposed Framework

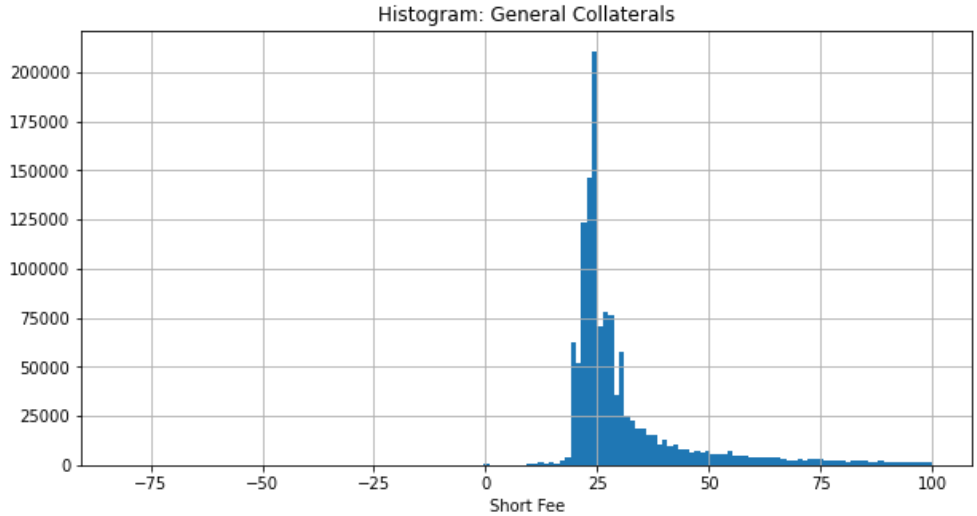
3.1. Stock Short Fee as Target Variable

The dependent variable is a stock short fee, which is the fee the broker charges in return for serving stock short sales. The data is collected from the brokers through surveys and is averaged on a daily frequency. Among the universe, 70% of the stocks are covered on average.

Figure 4 is a histogram of stock short fees. The skewness can be observed from the data. The plots have been separated at 100bps following D'Avolio [10], and the bottom plot contains the larger fees.

The labels are distributed from the short fees. There is no agreed hard boundary for a stock being special. Pedersen [34] has empirically mentioned those with the top 10% of short fees are hard to borrow, Beneish et. al. [4] used scores obtained from a professional data vendor. Geczy et. al. [18] estimated the 'GC rate' (rebate rate of a GC stock) depending on the size of the loan and calculated the specialness of the rebate. 25bps of a buffer from the GC rate was given to tell if a stock is special.

Following D'Avolio [10] and Muravyev et. al. [32] viewed 1% as the hard boundary of special stocks. As D'Avolio [10] pointed out, any stock with a rebate less than the GC rate can be considered a special stock, and the threshold is chosen to stratify the economically significant sense of "costliness". Geczy et. al. [18] used around 33bps to 40bps on average as the threshold, depending on the size of the loans. Moreover, the average short fee of the stocks falling into the boundary score used in Beneish et. al. [4] was over 270bps. This implies that the thresholds can be set in a malleable manner, as far as the rebates of the specials stay below the GC rate.



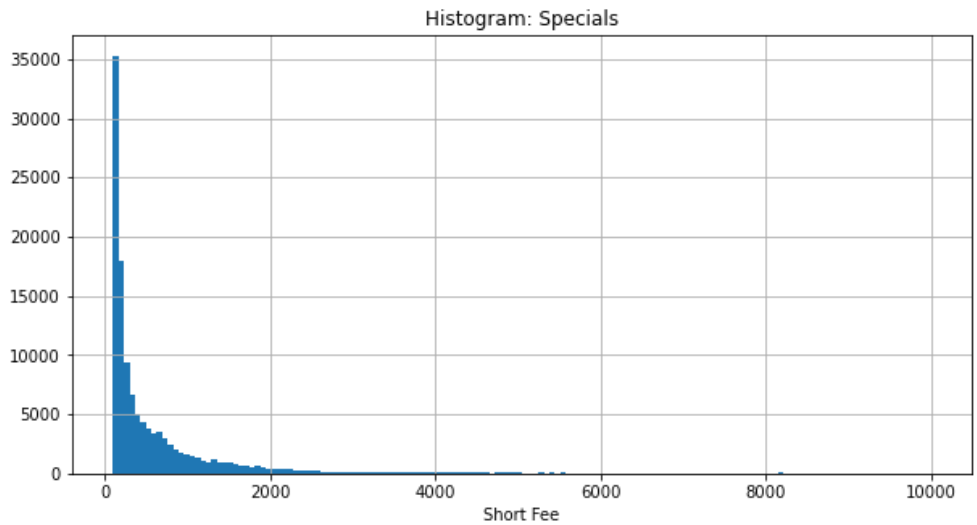


Figure 4: Stock Short Sales Fee Distribution

In the thesis, the base rate is used as the threshold for the ‘specialness’, which indicates that the special stocks would be regarded to incur negative rebates. The stocks with negative rebates have also been of a focus. Similar to the stocks with loan fees exceeding 100bps, those with negative rebates are hard to borrow, often with a greater magnitude especially in rising interest rate regimes, and hence referred to as “very special” [10]. Such rare incidents have encouraged previous studies to focus on the 1%−rule

instead. However, the same argument also suggests taking the 1% convention with caution during the low-interest rate regime. Under such circumstances, special stocks defined conventionally exhibit higher specialness than those with negative rebates. In other words, if the base rate stays lower than 1%, the portion of specials gets smaller if we were to follow the 1%-rule.

Other than the empirical reasons, thresholds involving the base rates can be considered from economic perspectives. The base rates tend to be positively correlated to the default risks of firms, and it would lift the loan fees of the special stocks. It can also be seen in the perspective of cost of capital, as the base rate provides the baseline for the risk-free nature of any investments or financing and the stock lending market should also be impacted by the money market. Hence the base rate can be seen as the threshold for the significance in the economic sense [10], point in time.

This leads to the point from the practitioner's view. Investments that give up more than the risk-free rate require extra care since it is guaranteed that the returns start below the risk-free level. This motivates additional research and risk assessment of such investments, and such prediction brings value before the execution.

The Federal Funds rates are used as the proxy for the base rate. It is to be noted that during the period of the sample the rates do not diverge too far from the 100bps line, hence the label distributions under both circumstances remain close to each other.

Other than the exceptional circumstances which require a more delicate approach, such as periods of zero or negative interest rates, we believe the thresholds set by the base rate can reflect economic conditions to a certain extent. Other techniques to set thresholds that provide better perspectives of the market and adjust for the time-

variability of the label distribution at the same time, are to be investigated in future research.

As a result, short fees which exceed the Federal Funds rates are labeled as special, and others are labeled as GC. The labels are then shifted a day so that the model gets to predict the abnormality of the fees the day after. The datasets are defined in a time-serial manner and the distribution of the label changes in the time dimension. We observed 7.4% of special instances in total.

3.2. Company Characteristics as Predictor Variables

Company Characteristics, sometimes called company fundamentals, are chosen according to two criteria. First, regarding the companies susceptible to credit risk are likely to go under the special regime [33], credit-related features are selected. As discussed in Psillaki et. al. [35] firm size measured in sales, profitability, liquidity, leverage, turnover, and tangible collateral features are calculated. Second, other company characteristics which are widely used in literature to represent a company are used. [1, 20] Table 1 includes the complete list of company characteristics used as input features. All features are calculated so that they are point-in-time.

Table 1: Company Characteristics

Feature	Description
Asset Turnover	Net sales over total assets
Book to Price	Book value over market capitalization

Table 1: Company Characteristics – Continued

Feature	Description
Cash Flow Ratio	Operating cash flow over current liabilities
Cash Ratio	Cash and equivalents over current liabilities
Current Liabilities to Price	Current liabilities over market capitalization
Current Ratio	Current assets over current liabilities
Debt to Assets	Total debt over total assets
Debt to Price	Total debt over market capitalization
Earnings to Price	EBITDA over market capitalization
Equity Value to Liabilities	Total market value of equity to total liabilities
Interest Coverage Ratio	EBIT over interest expenses
Long-Term Debt to Assets	Long-term debt over total assets
Market Capitalization	Market Capitalization
Net Cash Flow to Price	Net cash flow over Market capitalization
Net Current Assets to Price	(Current asset - current liabilities) over market capitalization

Table 1: Company Characteristics – Continued

Feature	Description
Net Profit Margin	Net profit over revenue
Quick Ratio	Liquid assets (cash & equivalents, accounts receivables) over current liabilities
Receivables Turnover Ratio	Net credit sales over average accounts receivable
Return on Assets	Net income over total assets
Return on Equity	Net income over shareholder’s equity
Return on Invested Capital	Net operating tax over invested capital
Sales to Enterprise Value	Sales over enterprise value
Sales to Price	Sales over market capitalization
Share Turnover	Average trade volume for the trailing month over the number of shares outstanding
Total Liabilities to Assets	Total liabilities to total assets
Working Capital Accruals	(Non−cash assets – current liabilities) over total assets
Working Capital to Assets	(Current assets – current liabilities) over total assets
Working Capital Turnover Ratio	Net sales over working capital

Table 1: Company Characteristics – Continued

Feature	Description
YoY Total Debt	Year-on-year growth in total debt
YoY Asset Growth	Year-on-year growth in total assets
YoY Debt to Assets	Year-on-year growth in debt to assets

3.3. Asset Characteristics as Predictor Variables

Asset Characteristics are mainly calculated from the information collected from the market. Basic market features as well as technical indicators, commonly used to predict short-term stock returns, are used as inputs. Volume and volatility-related features are expected to represent the interest of the market in the equity. Other technical indicators are expected to represent if general technical traders see if the equity is under/overvalued at the time. The features are described in Table 2.

Furthermore, as Moody's KMV credit model implies, the actual credit risk of a company, namely the distance to default, can be estimated using market features [41]. We expect the model can retrieve some hints of the market sentiment as well as fundamental implications from the market data.

All features have been passed through z-score transformation before being fed into the model for a stable optimization process. A total of 43 features are used as inputs.

Table 2: Asset Characteristics

Feature	Description
1 Month Price Momentum	Price change in a month
1 Year Return Volatility	Volatility of stock returns for the trailing year
3 Months Return Volatility	Volatility of stock returns for the trailing 3 months
5-Day Money Flow/Volume	Return-weighted dollar volume over dollar volume for the trailing 5 days
6 Months Return Volatility	Volatility of stock returns for the trailing 6 months
9MA MACD	9 days moving average of MACD
Volume to Price	Daily dollar volume over market capitalization
MACD	Moving average convergence/divergence [3]
MACD cross-section	Cross-sectional z score for MACD given the universe
Relative Volume Momentum	Average volume for the trailing 10 days over average volume for the trailing 50 days
RSI	Relative strength index [41]
Volume	Daily trade volume

3.4. Universe

We replicate Russell 3000 Index to build a universe comprising stocks of the 3000 largest companies being traded in the US markets at a given point in time. We then filter out those with no short fees available, which is around 30% of the data.

Then some more are lost due to periodic calculations such as YoY's, as some stocks just do not have enough history being included in the universe. For example, YoY features have values only after such stock has been in the index for a year. Others have been taken out if we simply do not have access to data for that stock at that date. This leaves an average of 1800 stocks at a given date.

The fact that the short fees of only 70% of the stocks are available supports the need for the model. The strategies involving short positions over such equities cannot be rigorously evaluated. As the literature suggests, some well-known 'market anomalies' lose their abnormal returns when the short fees are fully taken into account [4].

Moreover, as mentioned, the loss of data during preprocessing would have little if not a negative impact on the modeling, as if there were more data available, the imbalance of the specials would be mitigated, helping the model learn the distributions.

Chapter 4

Classification Models

Here we introduce techniques used for prediction.

4.1. Penalized Logistic Regression

Penalized logistic regression adds a regularization term to the objective function of logistic regression fitting to reduce the variance of the model. Logistic Regression is a linear model which computes the log odds and uses a logistic function to translate the log odds into probability. Let's say we have the input data $x_i \in \mathbb{R}^{n+1}$ where n is the number of features, containing 1 as its first element. Logistic Regression models class probability of a data point x, p(x) such that

$$\frac{p(x)}{1-p(x)} = e^{\theta^T x}$$

where θ is the parameter vector. The parameters are learned by solving the following optimization problem over the parameters.

$$\min \sum_i \left(y_i - \frac{1}{1+e^{-\theta^T x}} \right) + \lambda R(\theta)$$

where R is the regularization function. In the case of $R = \|\cdot\|_1$, we call the model LASSO [38], and in the case of $R = \|\cdot\|_2$, we call the model Ridge logistic regression [23]. λ is the hyperparameter which determines the strength of regularization. In the thesis, a validation dataset is used to determine the hyperparameter. L-BFGS solver

[43] is used to solve the optimization problem.

4.2. Decision Tree

Decision tree [6] finds decision rules depending on the information gain in a greedy way. Typically, a decision rule consists of a feature and decision boundary, and the process of finding the decision rule occurs recursively. The branches grow as the decision rule adds on. There are many ways of measuring information gain, and in this thesis, we used the Gini index, defined as follows:

$$G = \sum_{i} p(i)(1 - p(i))$$

where $p(i)$ is the empirical frequency of class i. The boundaries are chosen in a way that the information gain before and after the potential split is the greatest. The depth of the tree is a hyperparameter to decide. The thesis uses 5 as the maximum depth of a tree, and the ratio of classes in the leaf nodes is used as a prediction probability for plotting the ROC curve.

4.3. Random Forest

Random forest [7] is an ensemble method that uses multiple decision tree models to arrive at the final decision. During training, randomly sampled subsets of data are used to construct multiple weak trees which vote for the result. This is the basic idea of bootstrap aggregating (bagging), which results in lower model variance. Following Breiman [7], and Zhu et. al. [44], the importance

of each input feature can be calculated to enable some interpretation of the model and help understand the behavior of the dataset. Several sub-trees to make, or the maximum depth of each tree are hyperparameters. We use 1000 trees with a maximum depth of 10.

4.4. Adaptive Boosting

Adaptive boosting [17] is also an ensemble method that uses subsequent weak learners. The weights of previously incorrectly predicted data points are adjusted, so the subsequent model gets rewarded for predicting such data points. It can be used with any kind of base estimator. The family of model series then is combined to return the final output of the boosted model. The thesis used a decision tree with a maximum depth of 5 as the base estimator, with 1000 sequential learners.

4.5. Support Vector Machine

Support vector machine [9] for classification finds the hyperplane that separates data points into classes. Especially, the hyperplane between two classes is found by maximizing the average distance from the hyperplane to the support vectors of each class. Sometimes kernel functions can be used to transform the input space to a higher dimension to enable non-linear separations. The thesis used a linear SVM classifier.

4.6. Quadratic Discriminant Analysis

Quadratic discriminant analysis [22] uses a quadratic decision surface to separate data into classes. Assuming the data points are normally distributed given class, the decision boundary of QDA can be shown to be quadratic. The likelihood ratio is used to tell if the data falls into a category, and some threshold is used to make a classification decision.

4.7. Gaussian Naïve Bayes

Gaussian naïve bayes [22] also assumes that the data is normally distributed given the class. On top of that, the independence assumption is added. That is, given class information the features are independent, which leaves the same technique as QDA but with a diagonal covariance matrix. The class probability is then calculated using the Bayes Rule.

4.8. Artificial Neural Network

Artificial neural network [27] consists of a module called the perceptron, which is a linear operation followed by a non-linear activation function, which normally has significant values over some threshold. Such perceptrons build a layer of the neural network, and multiple layers can be used to make compositions. Normally artificial neural networks have more than 3 layers, namely the input layer, hidden layers, and the output layer.

The output of the output layer then passes through the softmax function so that the output gets translated into the probabilities of the class represented by each node of the output layer. The softmax values are compared with the label to produce the objective function to optimize. The backpropagation algorithm is widely used so that the parameters of the perceptrons are fitted.

The thesis used a 43-dimensional input layer, one 43-dimensional hidden layer, and a 2-dimensional output layer. AdamW [29] optimizing algorithm is used to find the parameters. The learning rate is determined to be 1e-4. Early stopping is applied so that if the cost does not enhance for 5 epochs the training stops, and the model with the highest validation score up to the point of stop is selected.

4.9. Isolation Forest

Isolation forest [28] detects anomalies by isolating anomalies from the normal points. By randomly selecting a feature and randomly determining a value to be used for splits, partitions are made within the data space. The split value is bounded by the minimum and maximum values of the selected feature. Such split takes place recursively until a partition contains only one value or data points inside a partition have the same values. Data points which require less partitioning are more likely to be anomalies.

As the number of splits conducted to isolate a data point is equivalent to the length of the path from the root to the leaf node of a tree representing a recursive partitioning, the average length over random trees is used to compute the anomaly score.

Chapter 5

Experimental Results

5.1. Experimental Settings

Experiments are conducted in two settings. Setup 1 simply divides train and test datasets from the entire dataset. The train set consists of data from 2017-01-01 to 2019-01-01. The test set follows from 2019-01-01 to 2019-12-31.

Setup 2 follows rolling-window settings, where multiple train/test set pairs are constructed in a rolling window sense. The experiment is designed to construct a train set of 6 months and a test set of the following 3 months of data. The size of the window is fixed to 1 month. As a result, 30 pairs of datasets are used to build and evaluate models.

In Figure 5, both experiment settings are shown. The green batch represents the train dataset, and the blue batch represents the test datasets. The hyperparameters are found using the validation dataset built from the train dataset.

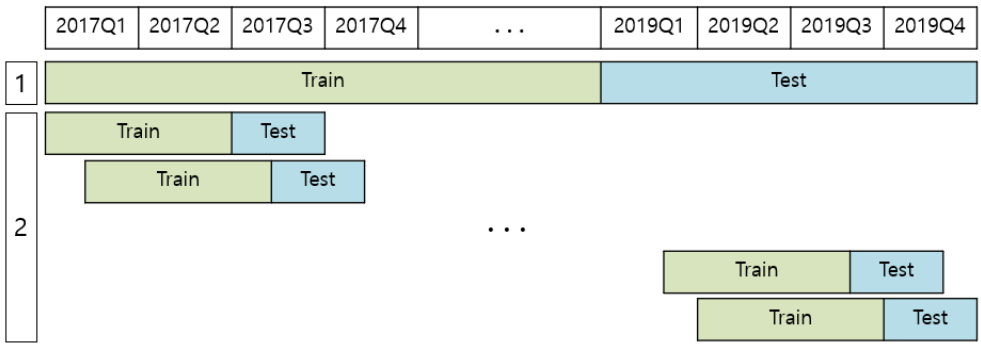


Figure 5: Dataset Preparation

As we are dealing with an imbalanced problem, we use another evaluation metric than accuracy. Normally recall or precision are used, but one class is not particularly more important than the other class, we consider an average classification performance, the AUROC. One other reason why this is considered the right metric is that there are a series of models built. We then conclude that the model with the highest average AUROC exhibits the best fit with the problem.

We formulate the modeling problems as follows. Let's say $D_t^{raw} = \{(X_t^{train}, s_t^{train}), (X_t^{test}, s_t^{test})\}$ is the train−test set pair of input data and adjusted short rates, time t denoting the index of the test dataset. The label is distributed such that

$$y_t^{train} = \begin{cases} 1, & s_t^{train} \geq \alpha \\ 0, & s_t^{train} < \alpha \end{cases} \text{ and } y_t^{test} = \begin{cases} 1, & s_t^{test} \geq \alpha \\ 0, & s_t^{test} < \alpha \end{cases},$$

where $\alpha = FedFundsRate_t$.

The goal is to learn a function $f: \mathbb{R}^{43} \to \{0,1\}$, with series of datasets: $\{\{(X_1^{train}, s_1^{train}), (X_1^{test}, s_1^{test})\}, \dots,$
$\{(X_t^{train}, s_t^{train}), (X_t^{test}, s_t^{test})\} \dots, \{(X_t^{train}, s_t^{train}), (X_t^{test}, s_t^{test})\}\}$.

5.2. Results

We report test evaluation results of the aforementioned models under both settings. Classification accuracies are calculated with a decision threshold of 0.5.

Table 3: Experiment Results (Setup 1)
Evaluation Metrics of Short Fees Prediction under Setup 1

Models	Test AUROC	Test Acc.
AdaBoost	0.928	**0.958**
ANN	0.920	0.939
Decision Tree	0.889	0.832
Logistic Regression (LASSO)	0.909	0.796
Logistic Regression (Ridge)	0.909	0.796
Naïve Bayes	0.854	0.889
QDA	0.789	0.838
Random Forest	**0.939**	0.917
SVM	0.906	0.810

Random forest showed the best test AUROC among the models tested. Although the AdaBoost model exhibited higher test accuracy, we take into account that average classification performance (ARUOC) is our prime performance indicator. It is inspiring that majority of baseline models achieved AUROC greater than 0.9, serving as strong baselines.

Table 4: Experiment Results (Setup 2)
Evaluation Metrics of Short Fees Prediction under Setup 2

Models	Test AUROC	Test Acc.
AdaBoost	0.911	**0.945**
ANN	0.909	0.932
Decision Tree	0.869	0.826
Logistic Regression (Ridge)	0.886	0.828
Logistic Regression (LASSO)	0.887	0.827
Naïve Bayes	0.841	0.898
QDA	0.830	0.900
Random Forest	**0.943**	0.935
SVM	0.883	0.838

Table 2 shows the same indicators as Table 1 but under setup 2. Unlike setup 1, multiple models are built during the training process, and the figures reported are the average metrics over the time horizon.

Random Forest again gained the highest AUROC of all, and the highest accuracy again goes to the AdaBoost method. The average AUROC of all baseline models sustained over 0.8, which again forms solid baselines.

For both setups, Naïve Bayes and QDA are the worst models to use, and it is thought to be due to the assumptions made by the models, that the data is normally distributed given the label. As we did not perform any kind of transformation to fit normality, the shape of the data might act as a hindrance to effective training. Of course,

the discrepancy exists, nevertheless.

We observed better models under setup 1 compared to setup 2, in general. It is suspected that the covariate shift in the data is not strong enough, as if there were an extreme distribution shift over the time horizon, models are hard to be built. This is due to the fact that the consistency in data distribution between the train and test sets is widely assumed. However, such results can justify that the effect of covariate shift is weaker than the advantage models gain from outnumbered training data.

5.3. Discussions

The results give few implications to discuss. First of all, we plot the evaluation metrics over time under setup 2. Figure 6 shows AUROC over time, and Figure 7 plots the test accuracies over time. The x-axis is the time index of the start of each test dataset.

Decision Tree and QDA model seem to show high variances of AUROC over time, but those with higher performance metrics, such as Random Forest or ANN models show smaller variances.

Also, in general, models tend to perform better for more recent data. This could be considered in conjunction with the distribution of the labels. From the assumptions, the thresholds have increased during the period of the sample (as rates rise) and have induced severer imbalance. The result is counterintuitive – it is known that models are hard to build under a greater degree of imbalance for general problems.

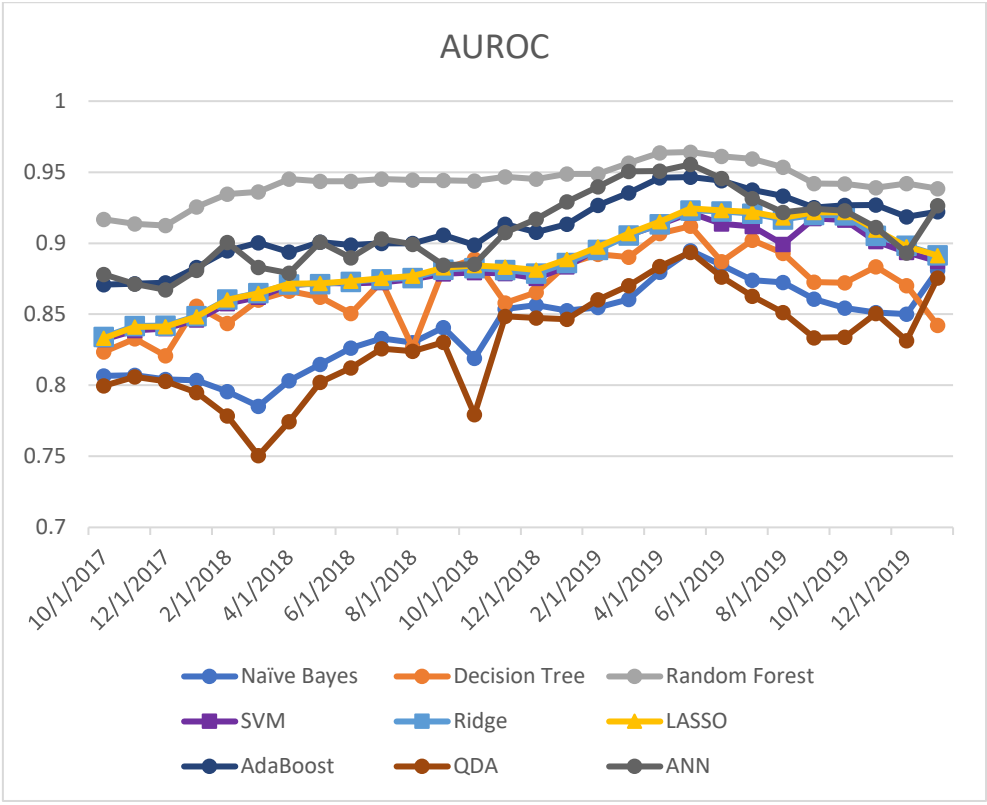


Figure 6: Test AUROC of Baseline Models

This leads to the discussion of the anomalous behavior of the special stocks, and there could be a possibility that the earlier datasets might have included general collateral stocks in the special class. The need for further studies on finding better measurement of specialness rises, although the thesis already took a step further from the previous studies. Figure 8 shows the ROC curves of all baseline models, for reference.

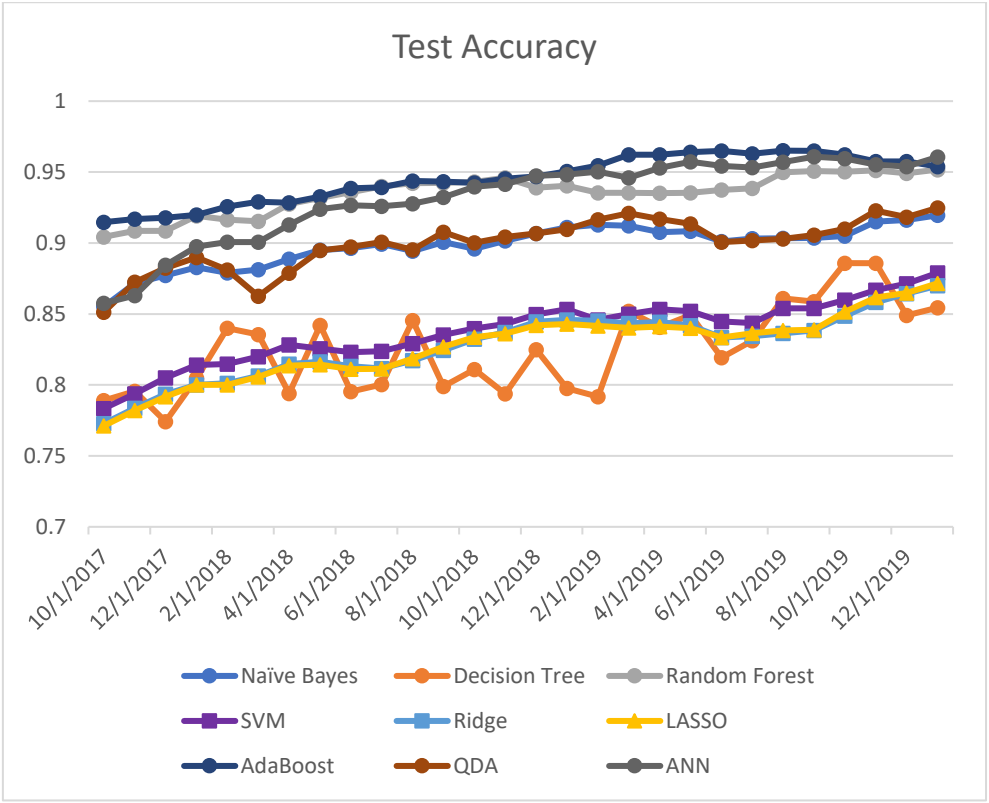


Figure 7: Test Accuracy of Baseline Models

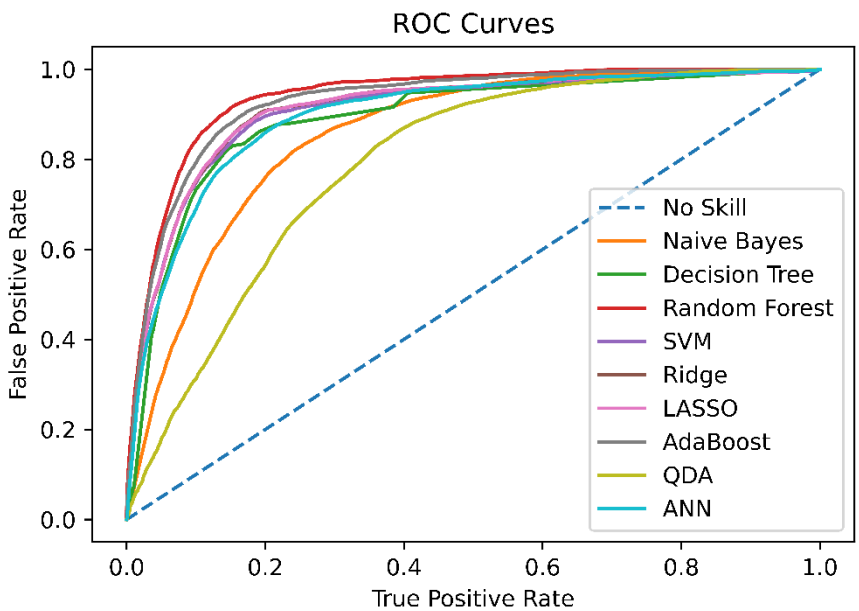


Figure 8: ROC Curves of Baseline Models

Next, we look at the feature importance of AdaBoost and Random Forest models built under setup 2. The feature importances are averaged over time. In Figure 7, the AdaBoost model saw share turnover and market capitalization as the most important features, which agrees with both the short fee studies [10] and credit studies [35]. Also, high-ranking features involve asset turnover, receivables turnover, working capital to total assets, working capital accruals, and cash ratio, which are used to evaluate how efficiently a company uses its assets. These have been pointed out in Psillaki et. al. [35] as factors for credit risks.

In Figure 8, feature importances from Random Forest are shown. Random Forest used market volatility and market capitalization factors to make decisions. The market capitalization factor agrees with the literature, and the volatility is not a surprise, as volatility captures much information about an asset, and related features are directly used in the KMV model to estimate the default probability. This also applies to AdaBoost results, that 1 Year Return Volatility can be found as one of the most important features.

For both models, we can again observe that the technical indicators (eg. RSI, MACD, and their variants) are far less important than other features and are not considered as useful when making decisions. It is interesting that volatility features and other technical indicators contain different information, and different contributions are made to the prediction.

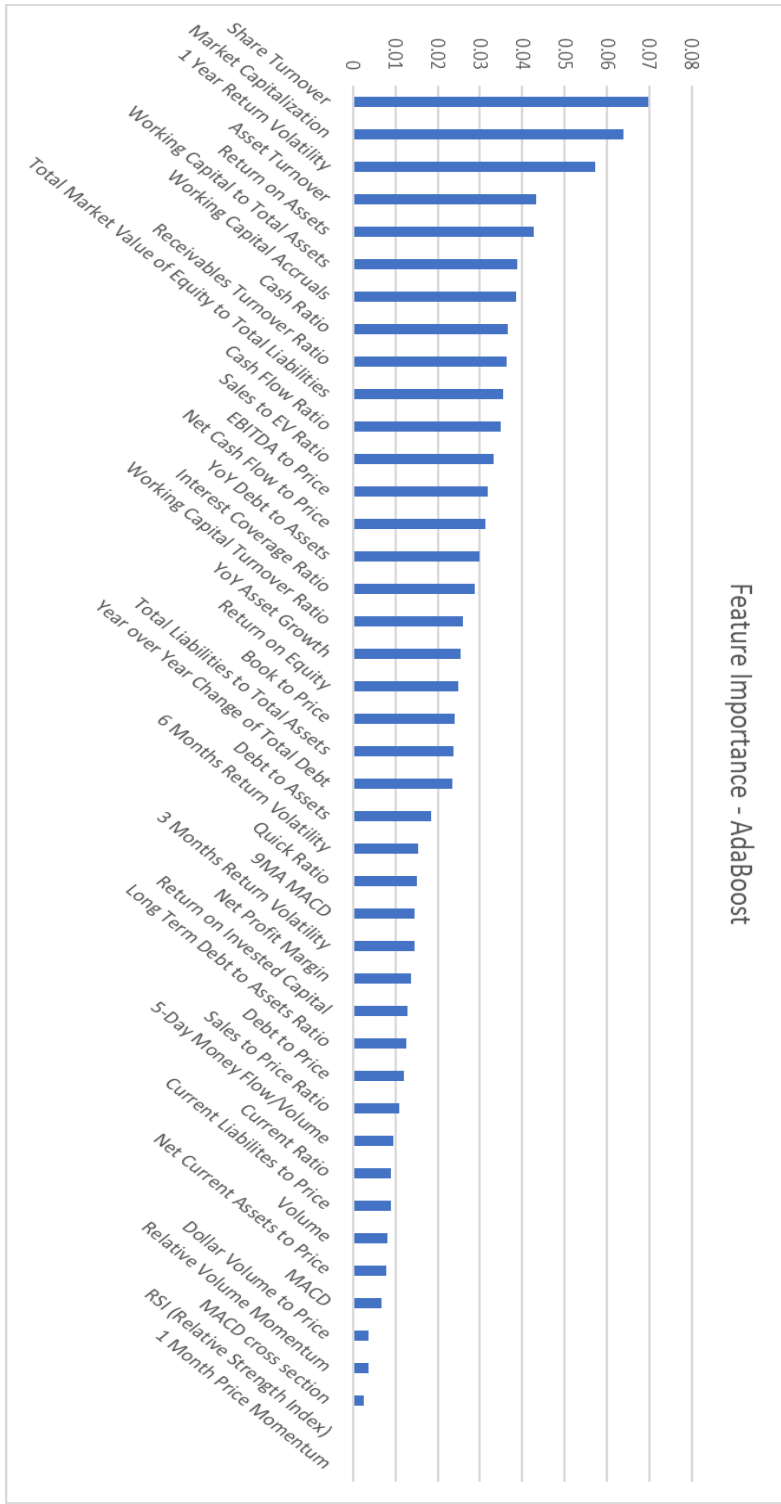


Figure 9: Feature Importance of AdaBoost Model

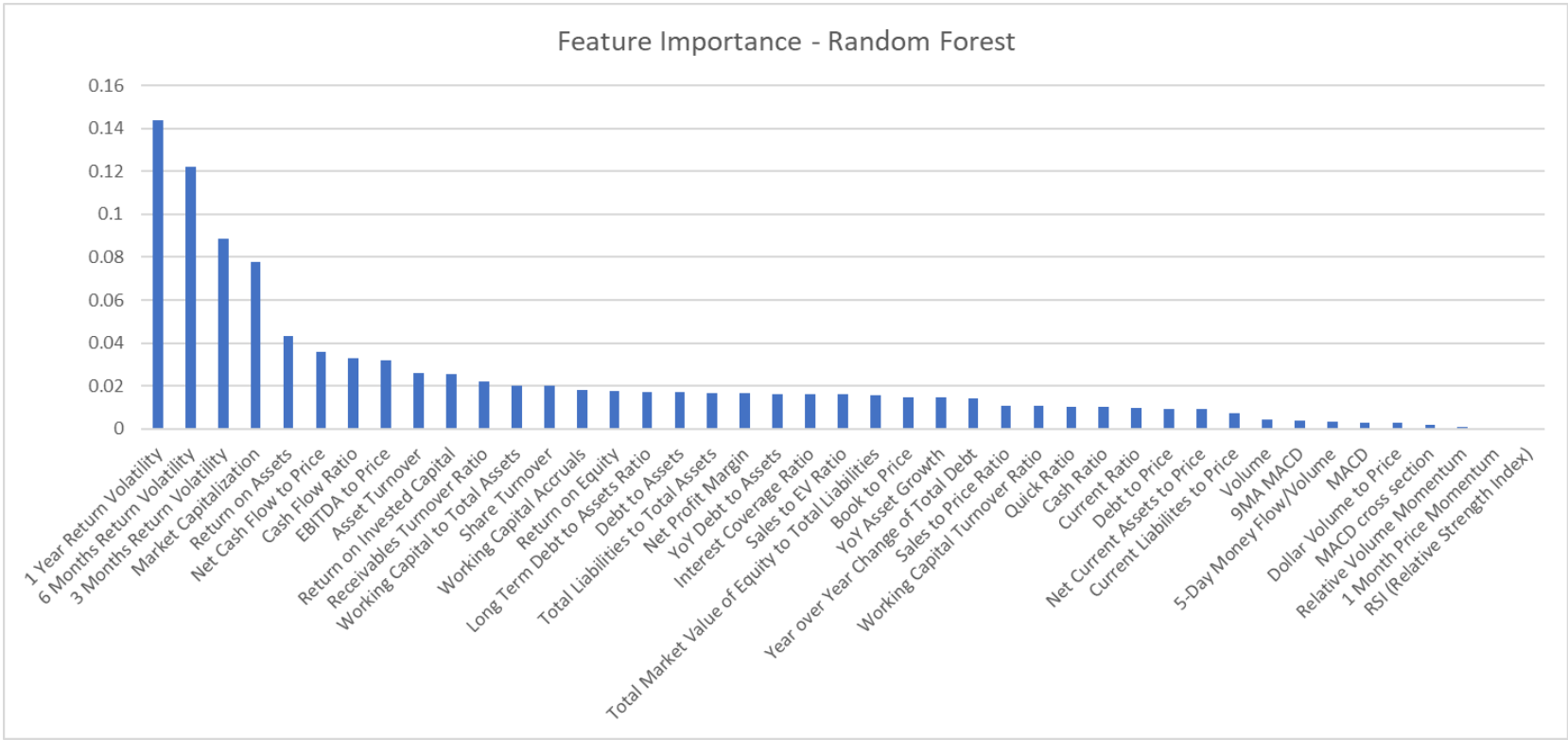


Figure 10: Feature Importance of Random Forest

5.4. Anomaly Detection

The last thing to note is an ad-hoc experiment conducted. We have used Isolation Forest [28] technique, which is widely used for anomaly detection. The experiment settings are quite similar to experiments discussed above, but there is a crucial difference. As Isolation Forest is trained in an unsupervised fashion, each train dataset is used without labels.

Under the anomaly detection settings, we assume that the special stocks are 'anomalies'. Isolation Forest can be trained with or without the anomalies, but there is no agreed caveat. In the thesis, all data points in a train dataset including the anomalies are used to train the anomaly detection model.

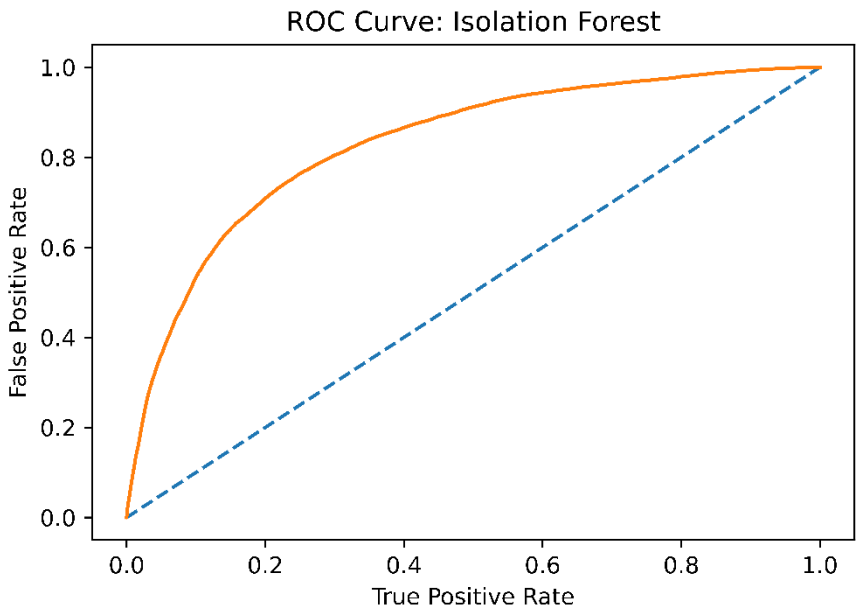


Figure 11: ROC Curve of Isolation Forest Model

The test data points are then fed into the model so that the model predicts if the data point is an anomaly or not. The output lies between −1 and 1, the anomaly being closer to −1. The output space has been linearly transformed between 0 and 1. We use this as the predicted probability of being special, to compute the AUROC.

The resulting metric shows an AUROC of 0.829 under setup 1 and 0.817 under setup 2. Although the AUROC is lower than any of the supervised models, we could see some predictive power of abnormal short fees. Interestingly, the anomalies determined only by looking at the input variables tend to fall under the special regime. Figure 9 shows the ROC (Receiver Operating Characteristics) curve for Isolation Forest under setup 1, and Figure 10 is the test AUROC obtained under setup 2.

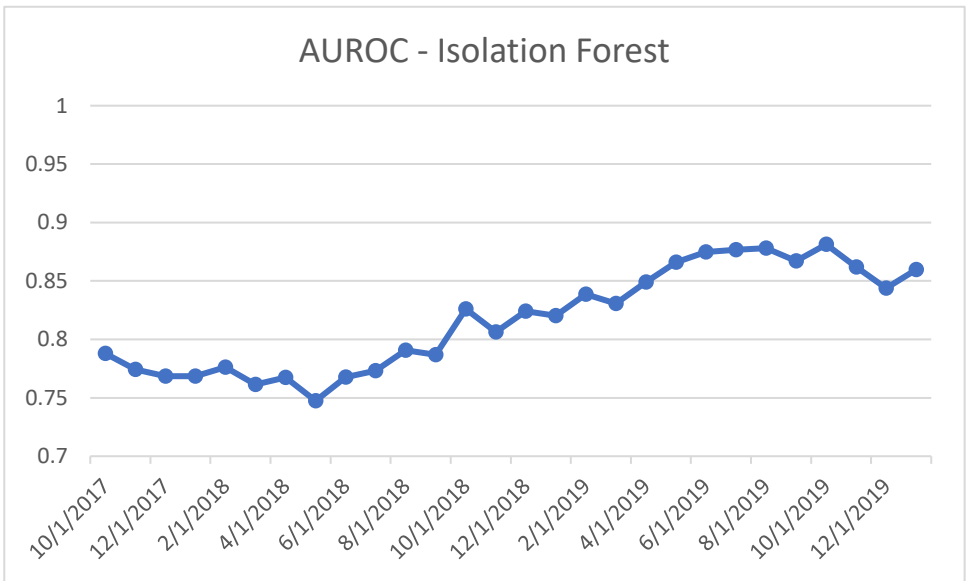


Figure 12: Test AUROC of Isolation Forest Model

Chapter 6

Conclusion

Understanding the stock short sales fee has big implications for both practitioners and academia. Under the general collateral regime, short sales fees show fairly consistent behavior, remaining on low levels near 0. However, the special regime occurs, where the short fee skyrockets, potentially threatening investment returns.

For practitioners, it is valuable to predict how the short fee is going to behave shortly and how it behaved in the past. They are used for trade execution planning and strategy evaluations through back-testing. Especially, we formulated the problem so that the prediction has bigger implications for the practitioners compared with the previous conventions in academia.

The thesis provides preliminary research for data-driven stock short sales fee prediction. This is because we assumed that different modeling approaches are to be taken for different regimes. Furthermore, viewing prediction of the specialness as a sole problem, we provided baselines for such prediction problem. Predicting the special regime itself has values, as it gives the practitioners a guide to screening a universe for building short investments.

Also, limited literature is available that investigated the behaviors of the short fees using data mining/machine learning techniques. The thesis provides initiatives for further studies on data-driven short fee analysis.

There is room for improvement as far as the threshold of the specialness threshold is concerned. Although we have not used a

fixed boundary for it, more research is to be done on how to infer the proper measure of specialness from either the stock market conditions or stock loan market microstructure, more realistic problems can be defined.

For one example, if the broker's Hard-To-Borrow (HTB) catalogue is available, the problem can also be formulated in a way to predict the hard-to-borrow stocks on the list, which are not available for immediate loans.

There can be two streams of future research stepping on this thesis. First, as the model serves as a preliminary stage of stock short fee prediction, the model suggested can be used as the indicator for which subsequent model to use. Different features and models are expected to be used to solve regression problems to predict the short fees, under each regime.

Furthermore, to solve the problem defined in the thesis better, state-of-the-art machine learning techniques could be tried, although studies are saying that deep learning and tree-based ensemble methods do not show significant performance differences when dealing with tabular data [19, 24]. One could formulate the input data in another format, for example, a sequence, to apply various modeling techniques from different fields. In addition, the market-related features could be built more delicately from Moody's KMV model, namely the distance to default.

Bibliography

[1] Abe, Masaya, and Hideki Nakayama. "Deep learning for forecasting stock returns in the cross-section." In Pacific-Asia conference on knowledge discovery and data mining, pp. 273-284. Springer, Cham, 2018.

[2] Alonso-Monsalve, S., Suárez-Cetrulo, A.L., Cervantes, A. and Quintana, D., 2020. Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators. Expert Systems with Applications, 149, p.113250.

[3] Appel, Gerald. Technical analysis: power tools for active investors. FT Press, 2005.

[4] Beneish, Messod Daniel, Charles MC Lee, and D. Craig Nichols. "In short supply: Short-sellers and stock returns." Journal of accounting and economics 60, no. 2-3 (2015): 33-57.

[5] Bollerslev, Tim. "Generalized autoregressive conditional heteroskedasticity." Journal of econometrics 31, no. 3 (1986): 307-327.

[6] Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Classification and regression trees. Routledge, 2017.

[7] Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.

[8] Christensen, K., Siggaard, M. and Veliyev, B., 2021. A machine learning approach to volatility forecasting. Available at SSRN.

[9] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20, no. 3 (1995): 273-297.

[10] D'avolio, Gene. "The market for borrowing stock." Journal of

financial economics 66, no. 2-3 (2002): 271-306.

[11] Diether, Karl B., Kuan-Hui Lee, and Ingrid M. Werner. "Can short-sellers predict returns? Daily evidence." Daily Evidence (July 14, 2005) (2005).

[12] Duffie, Darrell, Nicolae Garleanu, and Lasse Heje Pedersen. "Securities lending, shorting, and pricing." Journal of Financial Economics 66, no. 2-3 (2002): 307-339.

[13] Engelberg, Joseph E., Adam V. Reed, and Matthew C. Ringgenberg. "Short-selling risk." The Journal of Finance 73, no. 2 (2018): 755-786.

[14] Fama, Eugene F., and Kenneth R. French. "The cross-section of expected stock returns." the Journal of Finance 47, no. 2 (1992): 427-465.

[15] Feng, Guanhao, Stefano Giglio, and Dacheng Xiu. "Taming the factor zoo: A test of new factors." The Journal of Finance 75, no. 3 (2020): 1327-1370.

[16] Filipović, Damir, and Amir Khalilzadeh. "Machine Learning for Predicting Stock Return Volatility." Swiss Finance Institute Research Paper 21-95 (2021).

[17] Freund, Yoav, Robert Schapire, and Naoki Abe. "A short introduction to boosting." Journal-Japanese Society For Artificial Intelligence 14, no. 771-780 (1999): 1612.

[18] Geczy, Christopher C., David K. Musto, and Adam V. Reed. "Stocks are special too: An analysis of the equity lending market." Journal of Financial Economics 66, no. 2-3 (2002): 241-269.

[19] Gorishniy, Yury, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. "Revisiting deep learning models for tabular data." Advances in Neural Information Processing Systems 34 (2021).

[20] Gu, S., Kelly, B. and Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), pp.2223-2273.

[21] Gu, S., Kelly, B. and Xiu, D., 2021. Autoencoder asset pricing models. Journal of Econometrics, 222(1), pp.429-450.

[22] Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.

[23] Hoerl, Arthur E., and Robert W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems." Technometrics 12, no. 1 (1970): 55-67.

[24] Huang, Xin, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. "Tabtransformer: Tabular data modeling using contextual embeddings." arXiv preprint arXiv:2012.06678 (2020).

[25] Hwang*, Soosung, and Steve E. Satchell. "GARCH model with cross-sectional volatility: GARCHX models." Applied Financial Economics 15, no. 3 (2005): 203-216.

[26] Kot, Hung Wan. "What determines the level of short-selling activity?." Financial Management (2007): 123-141.

[27] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015): 436-444.

[28] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." In 2008 eighth ieee international conference on data mining, pp. 413-422. IEEE, 2008.

[29] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).

[30] Mandelbrot, Benoit. "The variation of some other speculative prices." The Journal of Business 40, no. 4 (1967): 393-413.

[31] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. and Anastasiu, D.C., 2019, April. Stock price prediction using news sentiment analysis. In 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 205-208). IEEE.

[32] Muravyev, Dmitriy, Neil D. Pearson, and Joshua M. Pollet. "Is there a risk premium in the stock lending market? evidence from equity options." The Journal of Finance (2016).

[33] Nashikkar, Amrut J., and Lasse Heje Pedersen. "Corporate bond specialness." In EFA 2007 Ljubljana Meetings Paper. 2007.

[34] Pedersen, Lasse Heje. "Efficiently inefficient." In Efficiently Inefficient. Princeton University Press, 2015.

[35] Psillaki, Maria, Ioannis E. Tsolas, and Dimitris Margaritis. "Evaluation of credit risk based on firm performance." *European journal of operational research* 201, no. 3 (2010): 873-881.

[36] Rapach, David E., Matthew C. Ringgenberg, and Guofu Zhou. "Short interest and aggregate stock returns." Journal of Financial Economics 121, no. 1 (2016): 46-65.

[37] Reed, Adam V. "Connecting supply, short-sellers and stock returns: Research challenges." Journal of Accounting and Economics 60, no. 2-3 (2015): 97-103.

[38] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58, no. 1 (1996): 267-288.

[39] Valášková, K., Gavláková, P. and Dengov, V., 2014. Assessing credit risk by Moody's KMV model. In 2nd International Conference on Economics and Social Science (ICESS 2014), Information Engineering Research Institute, Advances in Education Research (Vol. 61, pp. 40-44).

[40] Vargas, M.R., Dos Anjos, C.E., Bichara, G.L. and Evsukoff, A.G., 2018, July. Deep learning for stock market prediction using technical indicators and financial news articles. In 2018 international joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.

[41] Wilder, J. Welles. New concepts in technical trading systems. Trend Research, 1978.

[42] Yang, L., Ng, T.L.J., Smyth, B. and Dong, R., 2020, April. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In Proceedings of The Web Conference 2020 (pp. 441-451).

[43] Zhu, Ciyou, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization." ACM Transactions on mathematical software (TOMS) 23, no. 4 (1997): 550-560.

[44] Zhu, Ruoqing, Donglin Zeng, and Michael R. Kosorok. "Reinforcement learning trees." Journal of the American Statistical Association 110, no. 512 (2015): 1770-1784.

Abstract

주식 공매도를 위한 주식 대여는 비용이 발생한다. 해당 비용을 예측하는 것은 두 가지 측면에서 투자자에게 유리하다. 먼저, 과거 공매도 비용 데이터가 정확하다면 투자 전략 백테스팅의 정확도 향상을 기대할 수 있다. 또한, 미래의 공매도 데이터를 예측한다면 투자 위험 관리와 전략 실행 최적화의 재료가 된다. 주식 대여 비용의 분포는 아주 치우쳐져 있다. (양의 왜도) 일반적으로 0에 가까운 값을 가지는데 이를 문헌에선 일반 담보 (General Collateral)의 상태에 있다고 한다. 하지만 공매도 수요가 높은 상황에서는 공매도 비용이 크게 증가하는 것을 관찰할 수 있는데 이를 특이한 (Special) 상태에 있다고 한다. 본 연구는 주식 공매도 비용 예측에 공헌하고자 특이 주식과 일반 담보 주식을 분류하는 모델을 개발한다. 특히, 머신러닝 및 데이터 마이닝 방법을 사용한다. 분류 모델을 제안하는 것과 더불어 다양한 기법을 적용함으로써 해당 문제의 베이스라인을 제공한다.