

MEF UNIVERSITY

**PREDICTION OF UP AND DOWN SIGNALS IN
SELECTED BLUE CHIP STOCKS**

Capstone Project

Mustafa Yıldız

İSTANBUL, 2019

MEF UNIVERSITY

**PREDICTION OF UP AND DOWN SIGNALS IN
SELECTED BLUE CHIP STOCKS**

Capstone Project

Mustafa Yıldız

Advisor: Prof. Dr. Utku Koç

İSTANBUL, 2019

MEF UNIVERSITY

Name of the project: Prediction of Up and Down Signals in Selected Blue Chip Stocks

Name/Last Name of the Student: Mustafa Yıldız

Date of Thesis Defense: 09/09/2019

I hereby state that the graduation project prepared by Mustafa Yıldız has been completed under my supervision. I accept this work as a “Graduation Project”.

09/09/2019

Asst. Prof. Dr. Utku Koç

I hereby state that I have examined this graduation project by Mustafa Yıldız which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

09/09/2019

Prof. Dr. Özgür Özlük

Director
of

Big Data Analytics Program

We hereby state that we have held the graduation examination of Mustafa Yıldız and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member

Signature

1. Asst. Prof. Dr. Utku Koç

.....

2. Prof. Dr. Özgür Özlük

.....

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Mustafa Yıldız

09/09/2019

Signature

TABLE OF CONTENTS

Academic Honesty Pledge	v
TABLE OF CONTENTS.....	vi
LIST OF FIGURES	viii
LIST OF TABLES.....	ix
EXECUTIVE SUMMARY	x
ÖZET	xi
1. INTRODUCTION	1
2. ANALYZE OF SELECTED BLUE CHIP STOCKS' PRICE MOVEMENTS.....	3
2.1 Data and Sources	3
2.2 Equity Market Data Analytics.....	3
2.3 Features	3
2.4 Labels	4
2.5 Stock Selection.....	4
2.6 Data Preparation.....	5
2.7 Data Transformation	5
2.8 Feature Scale Reducing.....	5
2.9 Algorithms.....	5
2.10 Analysis.....	6
2.11 Methodology & Cloud Computing	7
3. RESULT, DISCUSSION AND CONCLUSION.....	10
3.1 Result.....	10
3.2 Discussion	11
3.3 Conclusion.....	13
APPENDIX-A: LIST OF THE EQUITY MARKET DATA ANALYTICS	15
APPENDIX-B: DESCRIPTIVE STATISTICS OF DATA SET	18
Yapi Kredi Order Arrival Analytics	18
Order Cancellation Analytics.....	18
Order Flow Analytics.....	19
Volume Weighted Average Price Analytics	19
Buyer Seller Analytics	20

Garanti Bankasi Order Arrival Analytics	20
Order Cancellation Analytics.....	21
Order Flow Analytics.....	21
Volume Weighted Average Price Analytics	22
Buyer Seller Analytics	22
THY A.O Order Arrival Analytics	23
Order Cancellation Analytics.....	23
Order Flow Analytics.....	24
Volume Weighted Average Price Analytics	24
Buyer Seller Analytics	25
Aselsan Order Arrival Analytics.....	25
Order Cancellation Analytics.....	26
Order Flow Analytics.....	26
Volume Weighted Average Price Analytics	27
Buyer Seller Analytics	27
APPENDIX-C: ALGORITHM RESULTS	28
Multi Class Supervised Classification Algorithm Results for THY	28
Multi Class Supervised Classification Algorithm Results for Yapi Kredi	29
Multi Class Supervised Classification Algorithm Results for Garanti	30
Multi Class Supervised Classification Algorithm Results for Aselsan	31
REFERENCES	32

LIST OF FIGURES

Figure 1. Comparing Algorithms on MS Azure Cloud Computing	8
Figure 2. PCA and Cross Validation Structure on MS Azure Cloud Computing	9
Figure 3. PCA and Tune Model Hyper Parameters Structure on MS Azure Cloud Computing	9
Figure 4. Labels of Each Data Set	12
Figure 5. p,d,q values of ARIMA models of Each Stock	13

LIST OF TABLES

Table 1. Average Volume of BIST Stocks in 2018	4
Table 2. Confusion Matrix Draft	7
Table 3. Selected Multi Class Supervised Ensemble Algorithm Results over Standardized and Scale Reduced Data	11

EXECUTIVE SUMMARY

PREDICTION OF UP AND DOWN SIGNALS IN SELECTED BLUE CHIP STOCKS

Mustafa Yıldız

Advisor: Asst. Prof. Dr. Utku Koç

SEPTEMBER , 2019, 33 pages

Efforts have been made to predict the direction in which equity stocks will move in the capital markets. In most of these studies, Technical Analysis and Fundamental Analysis based models have been used. For daily price estimations, macroeconomic variables or financial ratios of financial instruments are used. On the other hand trade book data are taken into consideration in intraday price estimates.

In this study, equity market data analytics, which are created by Borsa İstanbul as a benchmark for intraday price signals, are used. These analytics are derived from trade and order book data. For 5 minute periods, intraday price and equity market data analytics data sets are created, and different algorithms are tried over these data sets. The study is carried out using one-week data of 4 selected blue chip stocks. The signals for increase is 1, for decreases is -1 and 0 for non-change signals. As a result of the study, the decision jungle algorithm is the most successful algorithm. In addition this, the lack of volatility and liquidity in the market have caused overfitting problems in ensemble algorithms.

According to the multiclass decision jungle confusion matrix, the positive true results for 1 (or increase of the price) are promising. If an investors can just use the algorithm for the price increase, it will be meaningful. The true positive ratio of 1, 54.5%, is too high when it is compared with its false trues value for decrease (or -1), which is just 13.6%. The difference between true positive and false negative (54.5% - 13.6%) will be the earning ratio for the investor, if he/she decides to invest the price increase of Yapi Kredi stock with the decision jungle algorithm.

Although it is stated that big data algorithms (machine learning techniques) can give the best results for the data, domain knowledge related to the data is still very important. As it is seen in the study, in order to overcome the problems of overfitting or bias that occur in other studies, it is necessary to obtain sufficient domain knowledge in consultation with the experts and practitioners of the subject. In addition, the increase in the studies on intraday trading, which is a shallow area in the literature, will provide better results in the studies conducted on price forecasts in the future. In the results of this study, parallel with the literature, it is revealed that there is difficulty in estimating the stock price movements.

Key Words: Multiclass supervised algorithms, Machine Learning, BIST stocks.

ÖZET

BIST 30 ENDEKSİNDE YER ALAN BAZI ŞİRKETLERE AİT FİYATLARDAKİ YUKARI VE AŞAĞI YÖNLÜ HAREKETLERİN TAHMİNİ

Mustafa Yıldız

Tez Danışmanı: Dr. Öğretim Üyesi Utku Koç

EYLÜL, 2019, 33 sayfa

Sermaye piyasalarında pay senetlerinin hangi yönde hareket edeceği sürekli olarak araştırmalara konu olmuştur. Bu çalışmaların çoğunda, Teknik Analiz ve Temel Analiz bazlı modeller kullanılmıştır. Günlük fiyat tahminleri için makroekonomik değişkenler veya finansal oranlar kullanılmıştır. Diğer yandan, gün içi fiyat tahminlerine yönelik çalışmalarda ise işlem defteri verileri dikkate alınmıştır.

Bu çalışmada, Borsa İstanbul tarafından gün içi fiyatlara yönelik sinyaller olarak oluşturulan pay senedi piyasası veri analitikleri kullanılmıştır. Bu analitikler, işlem ve emir defteri verilerinden elde edilmiştir. Söz konusu analitikler kullanılarak 5 dakikalık dönemler için, gün içi fiyat ve hisse senedi piyasası veri analizi veri setleri oluşturulmuş ve bu veri setleri üzerinden farklı algoritmalar denenmiştir. Çalışma, BIST 30 endeksinde yer alan şirketler arasından seçilen 4 pay senedinin bir haftalık verileri kullanılarak gerçekleştirilmiştir. Fiyat artışları 1, azalışları -1, fiyat değişikliği olmayan dönemler ise 0 olarak alınmıştır. Çalışma sonucunda fiyat hareketlerini tahmininde en başarılı algoritma olarak karar ağacı/ormanı algoritması çıkmıştır. Buna ek olarak, piyasadaki oynaklık ve likidite eksikliği, eş zamanlı toplu olarak çalışan makine öğrenimi algoritmalarında aşırı uyumlamaya neden olmuştur.

Çok sınıflı karar ağacı/ormanı algoritmasından elde edilen hata matrisine göre, 1 için olumlu gerçek sonuçlar (veya fiyat artışı) umut vericidir. Dolayısıyla bir yatırımcı söz konusu algoritmayı sadece fiyat artışı için kullanması halinde, anlamlı sonuçlar alabilecektir. % 54.5 olarak elde edilen gerçek pozitif değeri (fiyat artışı veya 1), sadece % 13.6 olarak elde edilen (fiyat düşüşü veya -1) yanlış negatif değerleriyle karşılaştırıldığında çok yüksektir. Gerçek pozitif değeri ile yanlış negatif değeri arasındaki oran, , Yapı Kredi pay senedinin fiyat artışına karar ormanı algoritması ile yatırım yapmaya karar vermesi halinde yatırımcı için kazanç oranı olacaktır.

Makine öğrenimi tekniklerinin veri seti analizleri için en iyi sonuçları verebileceği belirtilmesine rağmen, verilerle ilgili uzman bilgisi hala çok önemlidir. Çalışmada görüldüğü gibi, diğer çalışmalarda ortaya çıkan aşırı uyuma/uyum veya taraflılık sorunlarının üstesinden gelmek için, konunun uzmanları ve uygulayıcıları ile istişare içinde yeterli uzman bilgisinin elde edilmesi gerekmektedir. Ayrıca literatürde henüz sığ bir alan olan gün içi alım satım işlemlerine yönelik yapılan çalışmaların artması ilerleyen dönemlerde fiyat tahminlerine yönelik yapılan çalışmalarda daha iyi sonuçlar alınabilmesini sağlayacaktır. Çalışma sonuçları literatür ile paralel pay senedi fiyat hareketlerini tahmin etmede güçlüğü ortaya koymaktadır.

Anahtar Kelimeler: Çok sınıflı etiketli algoritmalar, Makine öğrenmesi, BIST pay senetleri.

1. INTRODUCTION

Future price forecasts of financial products traded in the capital markets have always been a subject of interest and tried to be modeled. In this study, it has been aimed to predict the future prices of some selected blue chip stocks, listed in the blue chip index of Borsa İstanbul, BIST 30 Index.

In general, two main methods are used to estimate the price of stocks. These are Technical Analysis and Fundamental Analysis. Fundamental Analysis makes a price prediction for stocks by considering macroeconomic variables or financial ratios such as interest, income, growth, exchange rate, EBITDA, P/E, liquidity ratio. In addition, this analysis method works based on the intrinsic value of the stock. Technical Analysis, on the other hand, generally produces price prediction models taking into account intraday prices and stock volumes or analytics derived from these data (Investopedia, 2019). In addition to these analyzes, time series depended methods can be used to predict/forecast the closing prices of the financial products such as stocks, derivative contracts and debt based instruments (Shang, 2017).

Various algorithms are formed by using a wide range of technical analysis methods for stock price estimations. Algo-trading is performed by using these algorithms. Algo-trading transactions are automatically generated orders or buy-sell transactions by means of algorithms prepared by considering market data. Many different analytical groups such as price trends, chart patterns, volume and momentum indicators, oscillators, moving averages, support and resistance levels are included in such algorithms. Furthermore, market rules such as order cancellation, short selling, commission rates and different order types are taken into consideration while developing these algorithms (Yingsaeree, 2012).

High Frequency Trading (HFT), which is a subset of algo-trading transactions, is automated methods that can perform high-speed orders and transactions based on an algorithm. In recent years, 60-70% of the total transaction volume in large stock exchanges such as NASDAQ, NYSE, LSE and Deutsche Borse occurs in HFT transactions (Investopedia, 2019). Also in Borsa İstanbul approximately %30 of the total trade has been executed by HFTs.

In this study, a model that can predict intraday stock prices is formed similar to the algorithms used in HFT and algo-trading transactions. In the literature (Gunduz, Yaslan, &

Cataltepe, 2017), intraday prices in stock markets are generally forecasted by using trade book¹ data. There are two important reasons for this. Firstly, the fact that the algorithms in the literature were based on trade book data shifted the subsequent studies to this direction (Barclay, 1993). The second is that the order book² data is larger and more complex than the trade book data. In this study, it is tried to make price predictions by using data analytics which are derived from both trade book and order.

¹ Trade Book: A trade book is an electronic list of realized orders which are saved with some extra information about transactions such as price, quantity, volume, sides, date, time etc.

² Order Book: An order book is an electronic list of buy and sell orders for a specific security or financial instrument organized by price level.

2. ANALYZE OF SELECTED BLUE CHIP STOCKS' PRICE MOVEMENTS

2.1 Data and Sources

In the study, the equity market data analytics data set derived from order book data is used. This data set is obtained from Borsa İstanbul database. The closed price data of stocks at 5-minute intervals are drawn over the Bloomberg terminal and added to this data set. The data set consists of files prepared for each working day of the first week of 2018. The data set obtained by merging csv files each of which has over 250 MB volume, consists of more than 3 millions rows and 43 columns.

2.2 Equity Market Data Analytics

Equity Market Data Analytics are real time data analytics derived from order book and trade data of Borsa İstanbul Equity Market. Market participants can get in-depth information about BIST order book by using these analytics. These analytics present extra information about patterns, market conditions and trends.

These analytics have 5 different categories as follows (Borsa İstanbul, 2019);

- Order arrival analytics,
- Order cancellation analytics,
- Order flow analytics,
- Volume weighted average price analytics and
- Buyer-seller analytics.

Equity Market Data Analytics are calculated and distributed for the BIST 100 Stocks, and they are calculated at 1-second periods.

2.3 Features

39 data analytics calculated by Borsa İstanbul for the stock market were used as features in the study. Analytical prepared in 5 categories in daily csv files were consolidated. At consolidation, 39 data analytics are aligned in a row with separate Stock ID by time (Appendix). (Borsa İstanbul, 2019) These analytics are related to market depth such as buying, selling, trading, cancellation and volume. Similar analytics for intraday trading patterns are considered in the literature (Hautsch, 2001).

2.4 Labels

Labels are five minute closing prices. Data were obtained from Bloomberg terminal (Bloomberg, 2019). The data set is finalized by adding labels to the features data set. In the final case of the data set, there are 39 features, Label, Stock ID, Day and Time columns.

2.5 Stock Selection

Labels of the data analytics, which have acquired from the order and trade book data sets, have downloaded from the Bloomberg terminal. It have been decided to use a small subset from the whole BIST Stocks, because of two following reasons.

1. Liquidity is an important factor for intraday trading. It is too difficult to get healthy results by using an illiquid stock intraday price. It can result with too much noise and the explaining power of the algorithm would be too low.
2. The second reason is related with the terminal's presentation of the stock prices. Bloomberg terminal generally presents some of the intraday prices of the stock with 2 digits after point. In this condition, prices of the consecutive periods may not change, but in reality these prices changed but can not be monitored, because terminal shows just 2 or 1 digits after point of these prices.

Table 1. Average Volume of BIST Stocks in 2018

Ticker	Name	Weight	Average Volume
KRDMD TI Equity	Kardemir Karabuk Demir Celik Sanayi	0.643615	102,546,719
GARAN TI Equity	Turkiye Garanti Bankasi AS	8.461386	100,210,590
YKBNK TI Equity	Yapi ve Kredi Bankasi AS	1.768231	74,801,374
PETKM TI Equity	Petkim Petrokimya Holding AS	1.54793	67,213,170
THYAO TI Equity	Turk Hava Yollari AO	3.713529	60,799,978
EKGYO TI Equity	Emlak Konut Gayrimenkul Yatirim Ortaklig	1.025224	47,241,675
HALKB TI Equity	Turkiye Halk Bankasi AS	1.534406	44,578,494
KARSN TI Equity	Karsan Otomotiv Sanayii Ve Ticaret AS	0.091497	43,845,592
AKBNK TI Equity	Akbank T.A.S.	8.594504	40,236,126
DOHOL TI Equity	Dogan Sirketler Grubu Holding AS	0.524581	40,115,175
ISCTR TI Equity	Turkiye Is Bankasi AS	3.677633	35,878,877
IHLGM TI Equity	Ihlas Gayrimenkul Proje Gelistirme	0.116873	31,165,830
ZOREN TI Equity	Zorlu Enerji Elektrik Uretim AS	0.160015	30,725,488
VAKBN TI Equity	Turkiye Vakiflar Bankasi TAO	1.287006	27,891,080
TSKB TI Equity	Turkiye Sinai Kalkinma Bankasi AS	0.39626	26,958,000
TTKOM TI Equity	Turk Telekomunikasyon AS	1.036403	26,370,636
SASA TI Equity	Sasa Polyester Sanayi AS	0.400397	25,192,323
SKBNK TI Equity	Sekerbank Turk AS	0.173653	23,093,355
KOZAA TI Equity	Koza Anadolu Metal Madencilik Isletmeler	0.530044	22,125,388
IHLAS TI Equity	Ihlas Holding AS	0.135461	20,404,086
ASELS TI Equity	Aselsan Elektronik Sanayi Ve Ticaret AS	2.45047	18,731,678

Source: Bloomberg, bold ones are selected for this study.

Four stocks are selected according to the liquidity, weight of the each stock in the BIST 30 Index, and the stock balance in the selected group between finance and real sectors (Table 1).

2.6 Data Preparation

The analytics included in the features dataset contain data for the last 5 minutes. If there is no change in the pre-determined time interval that will affect the value of the relevant analytics, that analytics is not created. Therefore, if the analytic is not created, the analytic value is taken as zero and the missing values at the data set is completed.

2.7 Data Transformation

In some of the supervised algorithms, the data may need to be standardized or normalized. As some of these algorithms (such as Neural Networks and PCA-Principal Component Analyze) are used in the study, features have been standardized (Ibrahim, 2014). At the result of standardization, whole columns of the data set turn into values that have a mean of zero and a standard deviation of one.

2.8 Feature Scale Reducing

Principal Component Analysis (PCA) was used in order to reduce the scale of the features. As data analytics comprise of five groups, the number of components in PCA was selected as 5.

PCA: PCA is a dimesion reduction technique. With this technique, a data set with a high size or a lots of variables can be reduced to a smaller size or to a smaller number of variables. This method removes the necessary parts with high variation in the data, while removing unnecessary parts with low variation.

2.9 Algorithms

Due to the characteristics of data set four ensemble supervised classification algorithms were used. These are Decision Forest, Decision Jungle, Neural Network, Logistic Regression. By comparing the results of these algorithms, the highest accuracy/recall/precision ratios are revealed. The brief description of these algorithms as follows.

Decision Forest: The decision forest, a supervised learning algorithm, can be used for both classification and regression. This is one of the easiest and most flexible algorithms. This algorithm can increase the number of decision trees to provide more robust results. It is possible to obtain better results than the single decision tree by averaging the decision trees that are formed with the randomly selected pieces from the master data with the voting technique (for regression arithmetic average, for classification mod) (Navlani, 2018).

Decision Jungle: The working principle and structure of this algorithm is very similar to the decision forest algorithm. Decision Jungle reduces memory consumption tremendously, it also led to consistently better generalization performance than decision forest. The logic behind this comes from the structure of decision jungle. It uses directed acyclic graphs (DAGs) instead of nodes in the decision forests. Unlike conventional decision trees that only allow one path to every node, a DAG in a decision jungle allows multiple paths from the root to each leaf (Shotton, Sharp, Nowozin, Winn, & Criminisi, 2013).

Neural Network: A Neural Network (or Artificial Neural Network) , which is a learning technique through examples, is adapted from the structure of biological neuron system. This algorithm follows a non-linear path using nodes. It is a complex adaptive system. Since it can change its internal structure by changing the input weights, it is considered as an adaptive system (Navlani, Neural Network Models in R, 2019).

Logistic Regression: Labels of logistic regression, which is a statistical method, can be at ordinal or nominal scale. This algorithm, which is generally used to solve binary classification problems, can also solve multinomial classification problems. (Navlani, 2018).

2.10 Analysis

5 minutes price of analytics are used as features, and also signs of price changes are used as labels. Signals used for labeling transform into 1, -1 and 0 as up, down and non-change respectively³. In parallel with the literature, transaction costs are considered for non-change or 0 values (Vella V. &, 2014). As for features, values of data analytics are used. Consequently, analysis was conducted as follows.

³ The price changes of stocks which are more than %0.2 have been taken as up (if the change is increase) and down (if the change is decrease) movements, on the other hand, the price change less than % 0.2 assume as non-change. %0.2 ratio have been determined as the commission used in Turkish capital markets.

$$\text{sign}(\Delta y_t) = \beta_0 + \left(\sum_{i=1}^{39} \beta_i X_i \right)_t$$

On the other hand, 10 fold cross validation is performed and the model is created according to the optimum test train set according to the accuracy ratios. Cross validation is useful for reducing bias in a model that can be caused by using a single training set.

2.11 Methodology & Cloud Computing

In the study, firstly, all columns with zero values were omitted. Afterwards, Principal Component Analyze (PCA) was applied in order to separate features which are comprising of analytics into five components. Since the data set consists of five sub-groups as domain knowledge, five is chosen as the number of components in PCA.

Whole dataset loaded into Microsoft Azure Cloud and processed in the Azure Studio by the help of selected machine learning algorithms and tools⁴.

The results of the selected algorithms are determined according to the following metrics:

- Overall/Average accuracy: It's the ratio of the correctly labeled subjects to the whole pool of subjects.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})^5$$

- Micro-averaged / Macro-averaged precision: Precision is the ratio of the correctly positive labeled by the model/selected algorithm to all positive labels.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

- Micro-averaged / Macro-averaged recall: Recall is the ratio of the correctly positive labeled by the model/selected algorithm to all true labels.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

In addition to these metrics, results were examined according to the confusion matrix.

Table 2. Confusion Matrix Draft

	Predicted Class		
	1	0	-1

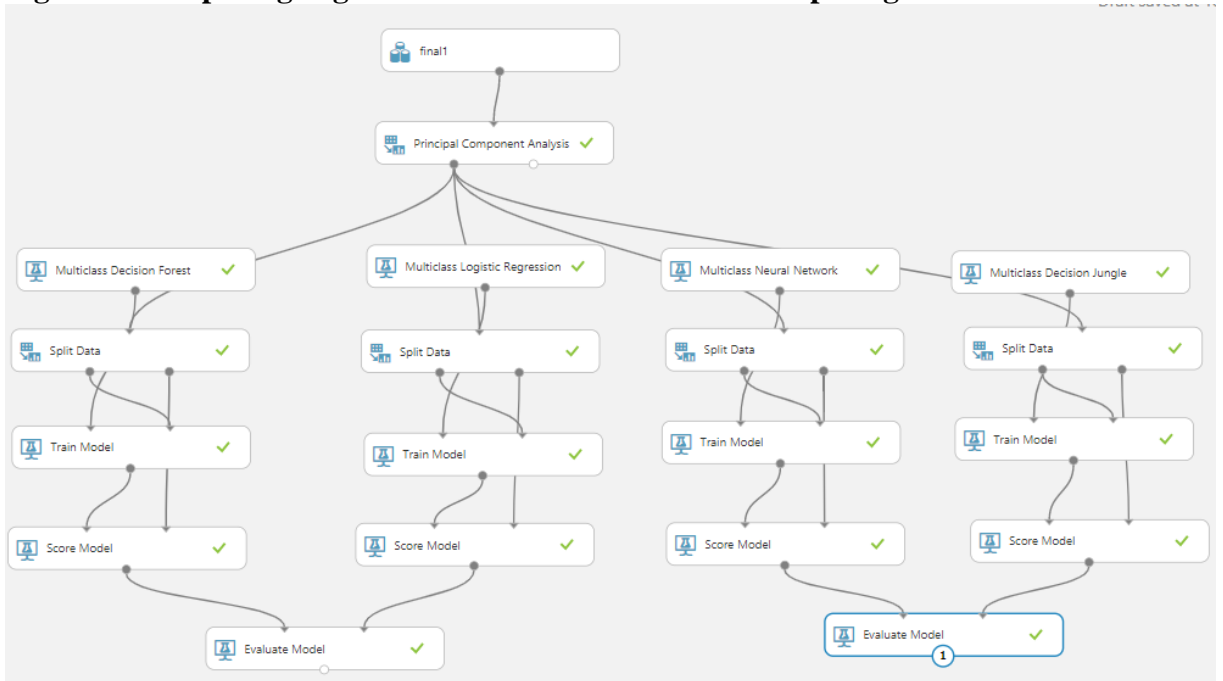
⁴ These tools are created by the MS Azure Studio, they can be used by drag&drop. Azure Studio converts ML algorithms, ML data analyzing methods, ML graphs, ML data selecting techniques etc. into tools or buttons in order to ease the use of these methods and techniques by users instead of typing heavy scripts in programming languages.

⁵ TP: True Positive / TN: True Negative / FP: False Positive/ FN: False Negative

Actual Class	1			
	0			
	-1			

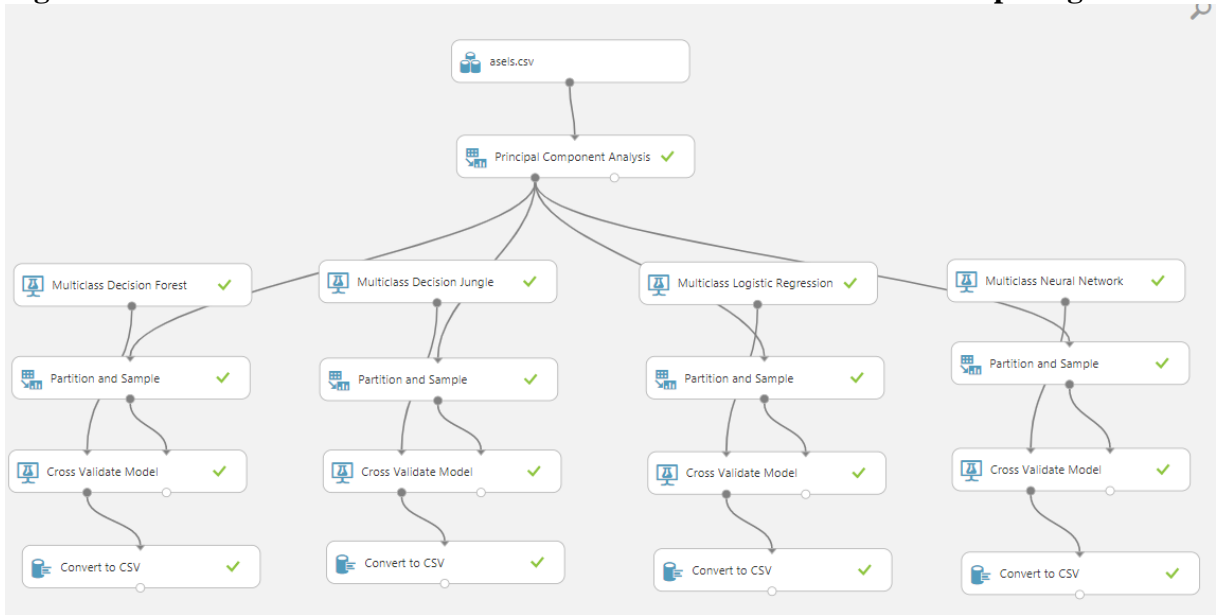
The methods used in the cloud as follows.

Figure 1. Comparing Algorithms on MS Azure Cloud Computing



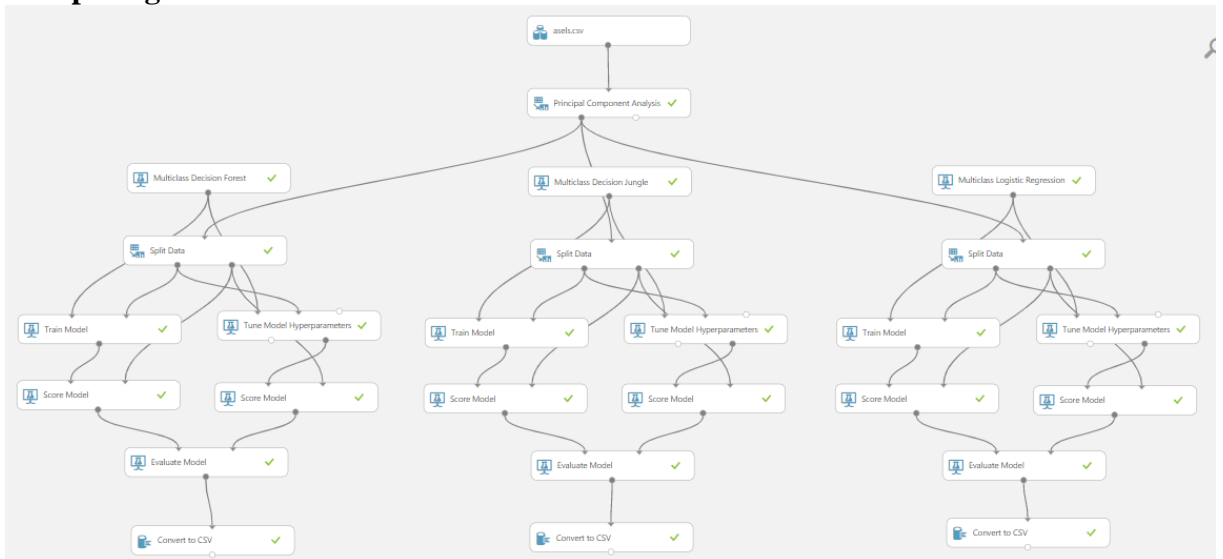
Source: Microsoft Azure Machine Learning Studio

Figure 2. PCA and Cross Validation Structure on MS Azure Cloud Computing



Source: Microsoft Azure Machine Learning Studio

Figure 3. PCA and Tune Model Hyper Parameters Structure on MS Azure Cloud Computing



Source: Microsoft Azure Machine Learning Studio

3. RESULT, DISCUSSION AND CONCLUSION

3.1 Result

10 fold cross validation is performed to increase the accuracy ratio in the test data set and to obtain a more effective training model. Since the data set is data for the first week of 2018, ie 4 days, the data set is divided into 3 days train and 1 day test. Since the Labels of the data set consists of 3 classes, multi-class supervised classification algorithms were preferred.

The results of 4 algorithms with 4 selected blue chip stocks are given in Appendix-C. Overfitting problems are noteworthy. Aselsan and Garanti shares showed overfitting with both standard parameters and hyper parameter tuning. Although the hyper parameter tuning approaches of Yapı Kredi and THY shares were also caused overfitting, the analysis with standard parameters yielded more meaningful and interpretable results in the confusion matrix. The reason for this is that the label of these two shares contains less 0 than the others.

Observing the confusion matrixes, both Yapı Kredi and THY obtained the most significant results according to Decision Jungle Algorithm. According to both results, it is seen that the loss risks are very low if the estimations are made in the direction of increase in Yapı Kredi and THY, and the probability of increasing estimation is 50% in Yapı Kredi and 20% in THY. On the other hand, for both shares, the remaining possibilities will at least not be harmed because it creates situations where the share price will not change.

PCA has contributed to the increase of accuracy ratio. In addition, better accuracy ratio have been acquired by using feature transformation or standardization of features. The results of Yapı Kredi Stocks with the standardized features are as follows.

Table 3. Selected Multi Class Supervised Ensemble Algorithm Results over Standardized and Scale Reduced Data

Decision Forest	Logistic Regression	Neural Network	Decision Jungle																																																																																																																																
<p>Predicted Class</p> <table border="1"> <tr> <td></td> <td>-1</td> <td>0</td> <td>1</td> </tr> <tr> <td>Actual Class</td> <td></td> <td></td> <td></td> </tr> <tr> <td>-1</td> <td>26.1%</td> <td>47.8%</td> <td>26.1%</td> </tr> <tr> <td>0</td> <td>27.5%</td> <td>60.0%</td> <td>12.5%</td> </tr> <tr> <td>1</td> <td>13.6%</td> <td>40.9%</td> <td>45.5%</td> </tr> </table> <p>Metrics</p> <table border="1"> <tr><td>Overall accuracy</td><td>0.470588</td></tr> <tr><td>Average accuracy</td><td>0.647059</td></tr> <tr><td>Micro-averaged precision</td><td>0.470588</td></tr> <tr><td>Macro-averaged precision</td><td>0.440548</td></tr> <tr><td>Micro-averaged recall</td><td>0.470588</td></tr> <tr><td>Macro-averaged recall</td><td>0.438472</td></tr> </table>		-1	0	1	Actual Class				-1	26.1%	47.8%	26.1%	0	27.5%	60.0%	12.5%	1	13.6%	40.9%	45.5%	Overall accuracy	0.470588	Average accuracy	0.647059	Micro-averaged precision	0.470588	Macro-averaged precision	0.440548	Micro-averaged recall	0.470588	Macro-averaged recall	0.438472	<p>Predicted Class</p> <table border="1"> <tr> <td></td> <td>-1</td> <td>0</td> <td>1</td> </tr> <tr> <td>Actual Class</td> <td></td> <td></td> <td></td> </tr> <tr> <td>-1</td> <td>4.3%</td> <td>91.3%</td> <td>4.3%</td> </tr> <tr> <td>0</td> <td>2.5%</td> <td>90.0%</td> <td>7.5%</td> </tr> <tr> <td>1</td> <td>4.5%</td> <td>72.7%</td> <td>22.7%</td> </tr> </table> <p>Metrics</p> <table border="1"> <tr><td>Overall accuracy</td><td>0.494118</td></tr> <tr><td>Average accuracy</td><td>0.662745</td></tr> <tr><td>Micro-averaged precision</td><td>0.494118</td></tr> <tr><td>Macro-averaged precision</td><td>0.46068</td></tr> <tr><td>Micro-averaged recall</td><td>0.494118</td></tr> <tr><td>Macro-averaged recall</td><td>0.39025</td></tr> </table>		-1	0	1	Actual Class				-1	4.3%	91.3%	4.3%	0	2.5%	90.0%	7.5%	1	4.5%	72.7%	22.7%	Overall accuracy	0.494118	Average accuracy	0.662745	Micro-averaged precision	0.494118	Macro-averaged precision	0.46068	Micro-averaged recall	0.494118	Macro-averaged recall	0.39025	<p>Predicted Class</p> <table border="1"> <tr> <td></td> <td>-1</td> <td>0</td> <td>1</td> </tr> <tr> <td>Actual Class</td> <td></td> <td></td> <td></td> </tr> <tr> <td>-1</td> <td></td> <td></td> <td>100.0%</td> </tr> <tr> <td>0</td> <td></td> <td></td> <td>100.0%</td> </tr> <tr> <td>1</td> <td></td> <td></td> <td>100.0%</td> </tr> </table> <p>Metrics</p> <table border="1"> <tr><td>Overall accuracy</td><td>0.470588</td></tr> <tr><td>Average accuracy</td><td>0.647059</td></tr> <tr><td>Micro-averaged precision</td><td>0.470588</td></tr> <tr><td>Macro-averaged precision</td><td>NaN</td></tr> <tr><td>Micro-averaged recall</td><td>0.470588</td></tr> <tr><td>Macro-averaged recall</td><td>0.333333</td></tr> </table>		-1	0	1	Actual Class				-1			100.0%	0			100.0%	1			100.0%	Overall accuracy	0.470588	Average accuracy	0.647059	Micro-averaged precision	0.470588	Macro-averaged precision	NaN	Micro-averaged recall	0.470588	Macro-averaged recall	0.333333	<p>Predicted Class</p> <table border="1"> <tr> <td></td> <td>-1</td> <td>0</td> <td>1</td> </tr> <tr> <td>Actual Class</td> <td></td> <td></td> <td></td> </tr> <tr> <td>-1</td> <td>17.4%</td> <td>56.5%</td> <td>26.1%</td> </tr> <tr> <td>0</td> <td>17.5%</td> <td>67.5%</td> <td>15.0%</td> </tr> <tr> <td>1</td> <td>13.6%</td> <td>31.8%</td> <td>54.5%</td> </tr> </table> <p>Metrics</p> <table border="1"> <tr><td>Overall accuracy</td><td>0.505882</td></tr> <tr><td>Average accuracy</td><td>0.670588</td></tr> <tr><td>Micro-averaged precision</td><td>0.505882</td></tr> <tr><td>Macro-averaged precision</td><td>0.453394</td></tr> <tr><td>Micro-averaged recall</td><td>0.505882</td></tr> <tr><td>Macro-averaged recall</td><td>0.464789</td></tr> </table>		-1	0	1	Actual Class				-1	17.4%	56.5%	26.1%	0	17.5%	67.5%	15.0%	1	13.6%	31.8%	54.5%	Overall accuracy	0.505882	Average accuracy	0.670588	Micro-averaged precision	0.505882	Macro-averaged precision	0.453394	Micro-averaged recall	0.505882	Macro-averaged recall	0.464789
	-1	0	1																																																																																																																																
Actual Class																																																																																																																																			
-1	26.1%	47.8%	26.1%																																																																																																																																
0	27.5%	60.0%	12.5%																																																																																																																																
1	13.6%	40.9%	45.5%																																																																																																																																
Overall accuracy	0.470588																																																																																																																																		
Average accuracy	0.647059																																																																																																																																		
Micro-averaged precision	0.470588																																																																																																																																		
Macro-averaged precision	0.440548																																																																																																																																		
Micro-averaged recall	0.470588																																																																																																																																		
Macro-averaged recall	0.438472																																																																																																																																		
	-1	0	1																																																																																																																																
Actual Class																																																																																																																																			
-1	4.3%	91.3%	4.3%																																																																																																																																
0	2.5%	90.0%	7.5%																																																																																																																																
1	4.5%	72.7%	22.7%																																																																																																																																
Overall accuracy	0.494118																																																																																																																																		
Average accuracy	0.662745																																																																																																																																		
Micro-averaged precision	0.494118																																																																																																																																		
Macro-averaged precision	0.46068																																																																																																																																		
Micro-averaged recall	0.494118																																																																																																																																		
Macro-averaged recall	0.39025																																																																																																																																		
	-1	0	1																																																																																																																																
Actual Class																																																																																																																																			
-1			100.0%																																																																																																																																
0			100.0%																																																																																																																																
1			100.0%																																																																																																																																
Overall accuracy	0.470588																																																																																																																																		
Average accuracy	0.647059																																																																																																																																		
Micro-averaged precision	0.470588																																																																																																																																		
Macro-averaged precision	NaN																																																																																																																																		
Micro-averaged recall	0.470588																																																																																																																																		
Macro-averaged recall	0.333333																																																																																																																																		
	-1	0	1																																																																																																																																
Actual Class																																																																																																																																			
-1	17.4%	56.5%	26.1%																																																																																																																																
0	17.5%	67.5%	15.0%																																																																																																																																
1	13.6%	31.8%	54.5%																																																																																																																																
Overall accuracy	0.505882																																																																																																																																		
Average accuracy	0.670588																																																																																																																																		
Micro-averaged precision	0.505882																																																																																																																																		
Macro-averaged precision	0.453394																																																																																																																																		
Micro-averaged recall	0.505882																																																																																																																																		
Macro-averaged recall	0.464789																																																																																																																																		

According to the multiclass decision jungle confusion matrix, the positive true results for 1 (or increase of the price) are promising (Table 3). If an investors can just use the algorithm for the price increase, it will be meaningful. The true positive ratio of 1, % 54.5, is too high when it is compared with its false trues value for decrease (or -1), which is just %13.6. On the other hand, at the case of false trues of 1 for non-change, investor will not lost any money, so it is not needed to be taken into consideration. Consequently, if the difference between true positive and false negative (just for -1), %54.5-%13.6=%40.9 can be acquired. This ratio will be the earning ratio for the investor, if he/she decides to invest the price increase of Yapi Kredi stock with the decision jungle algorithm.

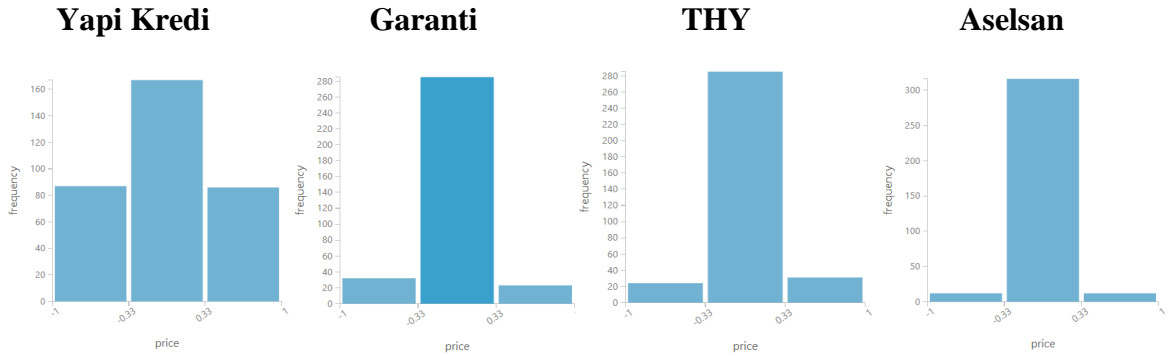
3.2 Discussion

Macroeconomic variables such as inflation, interest, exchange rate, or financial ratios and capital market processes such as EBITDA, earning per share, dividend payment, dilution of stocks, capital volatility, liquidity, depth, order types are not used in predictions of intraday trading prices. Instead of these features market microstructure data such as volatility, liquidity, depth, order types are taken as features for predicting the intraday financial instrument prices (Admati, 1988). However, in some papers, intraday trading price forecasts are based on instant news rather than market microstructure data (Mittermayer, 2004).

According to the results, decision jungle algorithm, which is the best performing model, has come forward. However, the consistency of the results is controversial due to the limited period of the data set and limited stock selection. In addition, decision tree

approaches tend to have a high tendency to overfitting especially in the framework of parameter tuning.

Figure 4. Labels of Each Data Set



Foster and Viswanathan showed that volatility and trading volume were higher during certain hours of the day about an intraday study of the New York Stock Exchange stock market (Foster, 1993). It is thought that there will be no problem of overfitting especially in the studies carried out for hours with high volatility. This type of hours can be determined and the study can be repeated by using the data of these hours.

Vella, in his neural network-based AI study, considered risk and volatility when making intraday price estimation (Vella V. , 2014). For this study, it can be repeated especially considering the volatility measures/features of stocks.

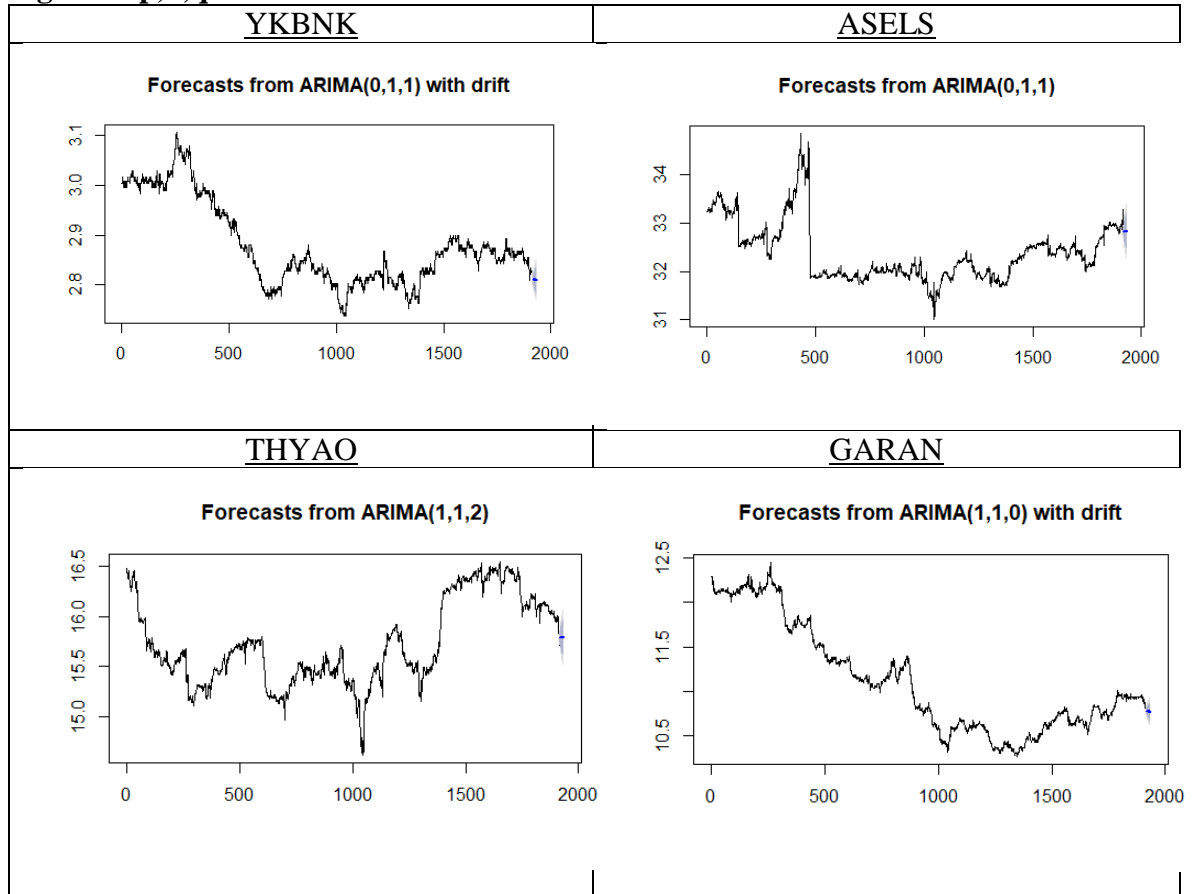
On the other hand, if higher volatility is obtained, repeating the same study with data analytics and share prices for 1 minute periods will result in higher accuracy ratios.

In addition to these approaches, autoregressive models in which the lags of labels are taken into consideration can be applied to the labels and lags taken from these models can be added as features to the datasets. For Yapi Kredi and Aselsan AR values is zero, so it is not needed to add any lags of labels to the data sets of these stocks. ARIMA analysis should be preprocessed in order to learn the number of lags and check the stationary. The ARIMA model can be designed as follows

$$y_t / \Delta y_t / \text{sign}(\Delta y_t) = \beta_0 + \left(\sum_{i=1}^{39} \beta_i X_i \right)_t + \left(\sum_{i=1}^{39} \beta_i X_i \right)_{t-1} + \left(\sum_{i=1}^{39} \beta_i X_i \right)_{t-2}$$

The ARIMA results of each stock are as follows.

Figure 5. p,d,q values of ARIMA models of Each Stock



As a summary the causes of high overfittng in the results may be as follws.

- The lack of volalility for selected period
- The lack of liquidity for selected period
- Omitting of risk and volatility related features
- The instrictive values of labels contains some of the values of features
- Omitting of lags of labels as features
- The lack of data set long for enough machine learning
- The lack of sensivitiy in stock prices

By considering these factors, this study can be rerun, and better metrics can be obtained.

3.3 Conclusion

Although it is stated that big data algorithms (machine learning techniques) can give the best results for the data, domain knowledge related to the data is still very important. As

it is seen in the study, in order to overcome the problems of overfitting or bias that occur in other studies, it is necessary to obtain sufficient domain knowledge in consultation with the experts and practitioners of the subject. In addition, the increase in the studies on intraday trading, which is a shallow area in the literature, will provide better results in the studies conducted on price forecasts in the future. In the results of this study, parallel with the literature, it is revealed that there is difficulty in estimating the stock price movements.

APPENDIX-A: LIST OF THE EQUITY MARKET DATA ANALYTICS

LIST OF THE EQUITY MARKET DATA ANALYTICS						
No.	Name	Explanation	Category	Period	Update frequency	Order-related / Trade-related
1	Number of arrived orders	Number of orders arriving per 60 seconds	ORDER ARRIVAL	60 Sec	Every second	Order-related
2	Cumulative number of arrived orders	Total number of arrived orders up to that time	ORDER ARRIVAL	Day (up to the calculation time)	Every second	Order-related
3	Quantity of arrived orders	Quantity of arrived orders per 60 seconds	ORDER ARRIVAL	60 Sec	Every second	Order-related
4	Cumulative quantity of arrived orders	Total quantity of arrived orders up to that time	ORDER ARRIVAL	Day (up to the calculation time)	Every second	Order-related
5	Number of arrived buy orders	Number of buy orders arriving per 60 seconds	ORDER ARRIVAL	60 Sec	Every second	Order-related
6	Number of arrived sell orders	Number of sell orders arriving per 60 seconds	ORDER ARRIVAL	60 Sec	Every second	Order-related
7	Quantity of arrived buy orders	Quantity of buy orders arriving per 60 seconds	ORDER ARRIVAL	60 Sec	Every second	Order-related
8	Quantity of arrived sell orders	Quantity of sell orders arriving per 60 seconds	ORDER ARRIVAL	60 Sec	Every second	Order-related
9	Number of arrived fill and kill orders	Number of fill and kill orders arriving per 60 seconds	ORDER ARRIVAL	60 Sec	Every second	Order-related
10	Number of cancelled orders	Number of cancelled orders per 60 seconds	ORDER CANCELLATION	60 Sec	Every second	Order-related
11	Quantity of cancelled orders	Quantity of cancelled orders per 60 seconds	ORDER CANCELLATION	60 Sec	Every second	Order-related
12	Number of cancelled buy orders	Number of cancelled buy orders per 60 seconds	ORDER CANCELLATION	60 Sec	Every second	Order-related
13	Number of cancelled sell orders	Number of cancelled sell orders per 60 seconds	ORDER CANCELLATION	60 Sec	Every second	Order-related

14	Quantity of cancelled buy orders	Quantity of cancelled buy orders per 60 seconds	ORDER CANCELLATION	60 Sec	Every second	Order-related
15	Quantity of cancelled sell orders	Quantity of cancelled sell orders per 60 seconds	ORDER CANCELLATION	60 Sec	Every second	Order-related
16	Cumulative number of cancelled orders	Total number of cancelled orders up to that time	ORDER CANCELLATION	Day (up to the calculation time)	Every second	Order-related
17	VWAP of cancelled orders	Volume weighted average price of cancelled orders up to that time	ORDER CANCELLATION	Day (up to the calculation time)	Every second	Order-related
18	VWAP of cancelled buy orders	Volume weighted average price of cancelled buy orders up to that time	ORDER CANCELLATION	Day (up to the calculation time)	Every second	Order-related
19	VWAP of cancelled sell orders	Volume weighted average price of cancelled sell orders up to that time	ORDER CANCELLATION	Day (up to the calculation time)	Every second	Order-related
20	Cancel / order ratio 1	The ratio of the number of cancelled orders to the number of arrived orders per 60 seconds	ORDER CANCELLATION	60 Sec	Every second	Order-related
21	Cancel / order ratio 2	The ratio of the quantity of cancelled orders to the quantity of arrived orders per 60 seconds	ORDER CANCELLATION	60 Sec	Every second	Order-related
22	Cumulative cancel / order ratio 1	The ratio of the total number of cancelled orders to the total number of arrived orders up to that time	ORDER CANCELLATION	Day (up to the calculation time)	Every second	Order-related
23	Cumulative cancel / order ratio 2	The ratio of the total quantity of cancelled orders to the total quantity of arrived orders up to that time	ORDER CANCELLATION	Day (up to the calculation time)	Every second	Order-related
24	Average quantity of arrived buy orders	Average quantity of buy orders per 5 minutes	ORDER FLOW	5 min	Every second	Order-related
25	Average quantity of arrived sell orders	Average quantity of sell orders per 5 minutes	ORDER FLOW	5 min	Every second	Order-related
26	Volatility of arrived buy order quantities	Volatility of buy order quantity per 5 minutes	ORDER FLOW	5 min	Every second	Order-related
27	Volatility of arrived sell order quantities	Volatility of sell order quantity per 5 minutes	ORDER FLOW	5 min	Every second	Order-related
28	VWAP of trades	Volume weighted average price (VWAP) of trades per 5 minutes	VWAP	5 min	Every second	Trade-related
29	VWAP of all trades	Volume weighted average price (VWAP) of trades up to that time	VWAP	Day (up to the calculation time)	Every second	Trade-related

30	VWAP of buyer-initiated trades	Volume weighted average price (VWAP) of trades where a pending sell order(s) is matched with a new buy order per 5 minutes	VWAP	5 min	Every second	Trade-related
31	VWAP of seller-initiated trades	Volume weighted average price (VWAP) of trades where a pending buy order(s) is matched with a new sell order per 5 minutes	VWAP	5 min	Every second	Trade-related
32	Number of buyer-initiated trades	Number of trades where a pending sell order(s) is matched with a new buy order per 60 seconds	BUYER SELLER	60 Sec	Every second	Trade-related
33	Number of seller-initiated trades	Number of trades where a pending buy order(s) is matched with a new sell order per 60 seconds	BUYER SELLER	60 Sec	Every second	Trade-related
34	Quantity of buyer-initiated trades	Quantity of trades where a pending sell order(s) is matched with a new buy order per 60 seconds	BUYER SELLER	60 Sec	Every second	Trade-related
35	Quantity of seller-initiated trades	Quantity of trades where a pending buy order(s) is matched with a new sell order per 60 seconds	BUYER SELLER	60 Sec	Every second	Trade-related
36	Buyer / seller ratio 1	The ratio of the number of buyer-initiated trades to the number of seller-initiated trades per 60 seconds	BUYER SELLER	60 Sec	Every second	Trade-related
37	Buyer / seller ratio 2	The ratio of the quantity of buyer-initiated trades to the quantity of seller-initiated trades per 60 seconds	BUYER SELLER	60 Sec	Every second	Trade-related
38	Cumulative buyer / seller ratio 1	The ratio of the total number of buyer-initiated trades to the total number of seller-initiated trades up to that time	BUYER SELLER	Day (up to the calculation time)	Every second	Trade-related
39	Cumulative buyer / seller ratio 2	The ratio of the total quantity of buyer-initiated trades to the total quantity of seller-initiated trades up to that time	BUYER SELLER	Day (up to the calculation time)	Every second	Trade-related

Source: BIST Database

APPENDIX-B: DESCRIPTIVE STATISTICS OF DATA SET

Yapi Kredi Order Arrival Analytics

Statistics	Number of arrived orders	Cumulative number of arrived orders	Quantity of arrived orders	Cumulative quantity of arrived orders	Number of arrived buy orders	Number of arrived sell orders	Quantity of arrived buy orders	Quantity of arrived sell orders	Number of arrived fill and kill orders
Mean	14.1971	2265.9676	340501.9176	47903219.03	4.8265	3.8471	91409.5618	123551.1441	0
Median	10	0	67458	0	0	0	0	0	0
Min	0	0	0	0	0	0	0	0	0
Max	175	10733	6311345	227577797	101	55	3822707	3801362	0
Standard Deviation	20.0846	3052.0293	851968.7912	63172478.28	9.8491	7.9941	316081.7726	482230.6861	0
Unique Values	52	148	247	148	32	29	150	130	1
Coefficient of Variance	1.414697368	1.346898914	2.502096896	1.318752258	2.040629856	2.077954823	3.457863339	3.903085557	0

Order Cancellation Analytics

Statistics	Number of cancelled orders	Quantity of cancelled orders	Number of cancelled buy orders	Number of cancelled sell orders	Quantity of cancelled buy orders	Quantity of cancelled sell orders	Cumulative number of cancelled orders	VWAP of cancelled orders	VWAP of cancelled buy orders	VWAP of cancelled sell orders	Cancel / order ratio 1	Cancel / order ratio 2	Cumulative cancel / order ratio 1	Cumulative cancel / order ratio 2
Mean	3.3029	64694.7882	0.6529	1.4529	9081.9206	30007.5618	259.5735	0.2939	0.1139	0.2214	0.2665	0.3831	0.1173	0.1014
Median	0	0	0	0	0	0	0	0	0	0	0.2	0.1746	0	0
Min	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Max	90	1798697	37	67	1033382	1034278	3622	2.8464	2.8232	2.8669	4.4286	5.5226	0.3545	0.3706
Standard Deviation	9.4263	223403.948	2.788	5.5856	62582.4073	133205.116	684.6199	0.8555	0.5503	0.755	0.3577	0.6386	0.1321	0.1367
Unique Values	29	96	14	21	42	61	57	37	15	28	128	247	162	125
Coefficient of Variance	2.85394653	3.453198538	4.2701792	3.84444903	6.89087805	4.43905161	2.63747994	2.91085403	4.83143108	3.41011743	1.34221388	1.6669277	1.12617221	1.34812623

Order Flow Analytics

Statistics	Average quantity of arrived buy orders	Average quantity of arrived sell orders	Volatility of arrived buy order quantities	Volatility of arrived sell order quantities
Mean	5067.5631	7518.3734	14306.3067	19174.3217
Median	0	0	0	0
Min	0	0	0	0
Max	92593.7143	78516.9667	288134.0683	189744.6592
Standard Deviation	9801.5632	13103.0973	30664.7461	37300.6011
Unique Values	143	143	143	143
Coefficient of Variance	1.934176843	1.742810127	2.143442521	1.945341362

Volume Weighted Average Price Analytics

Statistics	VWAP of trades	VWAP of all trades	VWAP of buyer-initiated trades	VWAP of seller-initiated trades
Mean	0.549	0.1968	0	0.1063
Median	0	0	0	0
Min	0	0	0	0
Max	2.86	2.8443	0	2.8596
Standard Deviation	1.1101	0.7152	0	0.5341
Unique Values	68	25	1	14
Coefficient of Variance	2.022040073	3.634146341	0	5.024459078

Buyer Seller Analytics

Statistics	Number of buyer-initiated trades	Number of seller-initiated trades	Quantity of buyer-initiated trades	Quantity of seller initiated trades	Buyer / seller ratio 1	Buyer / seller ratio 2	Cumulative buyer / seller ratio 1	Cumulative buyer / seller ratio 2
Mean	1.5824	1.5118	18432.1471	19562.8265	0.4254	74.8016	0.1885	0.1682
Median	0	0	0	0	0	0	0	0
Min	0	0	0	0	0	0	0	0
Max	133	136	1569911	1611612	19	16640	2.5625	5.0309
Standard Deviation	8.9147	11.915	116577.325	149154.5974	1.5931	925.0719	0.4816	0.4863
Unique Values	19	15	50	39	30	64	52	48
Coefficient of Variance	5.633657735	7.88133351	6.324674188	7.624388909	3.744945933	12.36700686	2.554907162	2.891200951

Garanti Bankasi Order Arrival Analytics

Statistics	Number of arrived orders	Cumulative number of arrived orders	Quantity of arrived orders	Cumulative quantity of arrived orders	Number of arrived buy orders	Number of arrived sell orders	Quantity of arrived buy orders	Quantity of arrived sell orders
Mean	31.2596	4797.1121	387678.4838	40440871.62	10.4484	12.8673	84433.5752	214766.6785
Median	21	0	134898	0	0	4	0	9779
Min	0	0	0	0	0	0	0	0
Max	329	20626	32234019	211979447	103	260	713960	31773429
Standard Deviation	38.5288	6102.004	1807191.368	56172194.55	16.3316	21.9384	150075.1521	1754529.981
Unique Values	89	167	263	168	54	57	163	184
Coefficient of Variance	1.232542963	1.272016137	4.661572524	1.388995645	1.563071858	1.704973071	1.777434531	8.169470207

Order Cancellation Analytics

Statistics	Number of cancelled orders	Quantity of cancelled orders	Number of cancelled buy orders	Number of cancelled sell orders	Quantity of cancelled buy orders	Quantity of cancelled sell orders	Cumulative number of cancelled orders	VWAP of cancelled orders	VWAP of cancelled buy orders	VWAP of cancelled sell orders	Cancel / order ratio 1	Cancel / order ratio 2	Cumulative cancel / order ratio 1	Cumulative cancel / order ratio 2
Mean	10.2478	174775.9499	2.9292	3.944	21092.6608	130848.879	1023.5074	1.8149	1.1735	1.2129	0.3609	0.5577	0.2237	0.1917
Median	0	0	0	0	0	0	0	0	0	0	0.4255	0.3762	0.2362	0
Min	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Max	244	31584568	51	204	882999	31301160	9718	10.9407	10.9012	10.9856	2.1111	25.5333	0.5174	0.7773
Standard Deviation	21.6548	1732251.769	7.7967	13.3875	70708.6429	1714709.01	2356.4742	4.0431	3.3578	3.4189	0.2755	1.6932	0.2219	0.2373
Unique Values	56	134	29	33	78	91	75	58	38	39	181	271	178	147
Coefficient of Variance	2.11311696	9.911270798	2.66171651	3.39439655	3.35228654	13.104499	2.30235189	2.22772605	2.86135492	2.81878143	0.76336935	3.03604088	0.99195351	1.23787167

Order Flow Analytics

Statistics	Average quantity of arrived buy orders	Average quantity of arrived sell orders	Volatility of arrived buy order quantities	Volatility of arrived sell order quantities
Mean	3338.317	5058.8873	6900.6033	10469.4168
Median	0	4147.646	0	6050.7447
Min	0	0	0	0
Max	15972.9151	88290.6701	46571.2187	146741.0768
Standard Deviation	3775.9609	8934.0415	8771.3947	18618.3784
Unique Values	167	190	167	190
Coefficient of Variance	1.131097167	1.766009197	1.271105484	1.778358695

Volume Weighted Average Price Analytics

Statistics	VWAP of trades	VWAP of all trades	VWAP of buyer-initiated trades	VWAP of seller-initiated trades
Mean	2.6118	1.2092	0	1.2701
Median	0	0	0	0
Min	0	0	0	0
Max	10.9751	10.9402	0	10.9595
Standard Deviation	4.6312	3.4086	0	3.478
Unique Values	83	39	1	41
Coefficient of Variance	1.773183245	2.818888521	0	2.738367058

Buyer Seller Analytics

Statistics	Number of buyer-initiated trades	Number of seller-initiated trades	Quantity of buyer-initiated trades	Quantity of seller-initiated trades	Buyer / seller ratio 1	Buyer / seller ratio 2	Cumulative buyer / seller ratio 1	Cumulative buyer / seller ratio 2
Mean	1.9912	2.2507	8460.7286	9860.9322	0.5792	153.2132	0.2345	0.2111
Median	0	0	0	0	0	0	0	0
Min	0	0	0	0	0	0	0	0
Max	41	100	254127	489134	25	22696.8	2.9505	2.6296
Standard Deviation	5.9542	8.5675	29916.4856	42860.2573	1.945	1380.0697	0.5338	0.501
Unique Values	25	29	57	57	69	93	63	57
Coefficient of Variance	2.990257131	3.806593504	3.535923088	4.346471148	3.35808011	9.007511755	2.276332623	2.373282804

THY A.O Order Arrival Analytics

Statistics	Number of arrived orders	Cumulative number of arrived orders	Quantity of arrived orders	Cumulative quantity of arrived orders	Number of arrived buy orders	Number of arrived sell orders	Quantity of arrived buy orders	Quantity of arrived sell orders
Mean	59.1294	12103.7941	379383.5529	589172.22.73	27.7735	22.3206	142177.4971	169769.1441
Median	49	10573	5	454075.31	18.5	13	67599.5	47186
Min	0	0	0	0	0	0	0	0
Max	392	38637	9	220326.884	268	144	177971.2	505525.0
Standard Deviation	61.9478	11971.8472	568012.5905	608048.54.94	37.3216	29.9609	215693.3185	371709.7054
Unique Values	125	214	290	214	85	80	231	213
Coefficient of Variance	1.04766	0.98909	1.49719	1.03203	1.34378	1.34229	1.51707	2.18950
	4952	8716	8775	8717	4543	8146	0724	0968

Order Cancellation Analytics

Statistics	Number of cancelled orders	Quantity of cancelled orders	Number of cancelled buy orders	Number of cancelled sell orders	Quantity of cancelled buy orders	Quantity of cancelled sell orders	Cumulative number of cancelled orders	VWAP of cancelled orders	VWAP of cancelled buy orders	VWAP of cancelled sell orders	Cancel / order ratio 1	Cancel / order ratio 2	Cumulative cancel / order ratio 1	Cumulative cancel / order ratio 2
Mean	14.3794	90081.2441	5.9794	4.9912	34531.5412	32042.3412	2011.6706	4.0049	2.6569	2.0619	0.3097	0.4185	0.2055	0.218
Median	0	0	0	0	0	0	0	0	0	0	0.3496	0.3964	0.2768	0.3156
Min	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Max	179	1014512	118	61	676881	768404	14416	16.4451	16.3584	16.5316	1.0952	3.7714	0.3753	0.5203
Standard Deviation	23.7746	157258.9288	13.2719	10.5843	79050.2468	82777.5761	3738.8463	7.0025	5.9926	5.4275	0.1919	0.3445	0.1576	0.2072
Unique Values	67	163	46	40	109	103	102	85	57	44	235	297	223	183
Coefficient of Variance	1.65337914	1.745745525	2.21960397	2.12059224	2.28921861	2.58338102	1.85857779	1.74848311	2.25548572	2.63228091	0.6196319	0.82317802	0.76690998	0.95045872

Order Flow Analytics

Statistics	Average quantity of arrived buy orders	Average quantity of arrived sell orders	Volatility of arrived buy order quantities	Volatility of arrived sell order quantities
Mean	2936.9426	3234.2856	6828.9184	6814.3562
Median	2985.6835	3019.566	5595.2463	5221.2974
Min	0	0	0	0
Max	18543.5	54770.6486	55537.9071	96898.0803
Standard Deviation	2478.5503	4734.3501	7228.2244	10327.0138
Unique Values	245	188	245	188
Coefficient of Variance	0.843921941	1.463800878	1.058472803	1.515479012

Volume Weighted Average Price Analytics

Statistics	VWAP of trades	VWAP of all trades	VWAP of buyer-initiated trades	VWAP of seller-initiated trades
Mean	5.7764	2.6687	0	2.4847
Median	0	0	0	0
Min	0	0	0	0
Max	16.5171	16.4681	0	16.4849
Standard Deviation	7.7834	6.0193	0	5.8567
Unique Values	122	57	1	53
Coefficient of Variance	1.347448238	2.255517668	0	2.357105486

Buyer Seller Analytics

Statistics	Number of buyer-initiated trades	Number of seller-initiated trades	Quantity of buyer-initiated trades	Quantity of seller initiated trades	Buyer / seller ratio 1	Buyer / seller ratio 2	Cumulative buyer / seller ratio 1	Cumulative buyer / seller ratio 2
Mean	9.2294	5.4824	25557.2588	11270.5	1.1244	124.6001	0.407	0.3627
Median	0	0	0	0	0	0.045	0	0
Min	0	0	0	0	0	0	0	0
Max	115	124	995637	431752	16	34809.5	2.059	2.5155
Standard Deviation	18.7661	15.4147	80365.6445	43894.2559	2.2933	1898.5025	0.6102	0.6512
Unique Values	57	39	123	84	141	173	110	86
Coefficient of Variance	2.033295772	2.811670072	3.144533032	3.894614782	2.039576663	15.23676546	1.499262899	1.795423215

Aselsan Order Arrival Analytics

Statistics	Number of arrived orders	Cumulative number of arrived orders	Quantity of arrived orders	Cumulative quantity of arrived orders	Number of arrived buy orders	Number of arrived sell orders	Quantity of arrived buy orders	Quantity of arrived sell orders
Mean	68.8059	17600.6324	60167.7706	13129710.67	30.4324	25.2529	25770.2706	25261.0559
Median	53	15844	34795.5	11581653	20	13	9963	8680.5
Min	0	0	0	0	0	0	0	0
Max	717	52972	723287	41618977	398	319	367594	383203
Standard Deviation	89.351	15980.726	87574.02	12376816.35	46.7562	43.9293	51467.9168	45171.89
Unique Values	124	228	297	228	88	75	239	211
Coefficient of Variance	1.298595033	0.907963171	1.455497173	0.942657204	1.536395421	1.739574465	1.997181853	1.788202765

Order Cancellation Analytics

Statistics	Number of cancelled orders	Quantity of cancelled orders	Number of cancelled buy orders	Number of cancelled sell orders	Quantity of cancelled buy orders	Quantity of cancelled sell orders	Cumulative number of cancelled orders	VWAP of cancelled orders	VWAP of cancelled buy orders	VWAP of cancelled sell orders	Cancel / order ratio 1	Cancel / order ratio 2	Cumulative cancel / order ratio 1	Cumulative cancel / order ratio 2
Mean	11.2382	10533.25	3.6118	4.2647	2857.6824	4337.6765	1524.3559	7.2515	3.3834	5.3204	0.2232	0.3233	0.1382	0.1234
Median	0	0	0	0	0	0	0	0	0	0	0.2372	0.2489	0.1843	0.1709
Min	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Max	197	178709	159	89	144480	133900	12607	32.9885	32.7371	33.3074	0.6316	3.1979	0.2502	0.3125
Standard Deviation	21.2989	21652.6641	12.41	10.7665	11555.2361	12686.8536	3062.2328	13.5362	9.8477	12.1295	0.1308	0.3519	0.0998	0.1223
Unique Values	53	157	31	34	86	98	90	77	37	56	224	305	232	177
Coefficient of Variance	1.89522343	2.055648931	3.43595991	2.5245621	4.04356905	2.92480401	2.00886998	1.86667586	2.91059289	2.27980979	0.58602151	1.08846273	0.72214182	0.9910859

Order Flow Analytics

Statistics	Average quantity of arrived buy orders	Average quantity of arrived sell orders	Volatility of arrived buy order quantities	Volatility of arrived sell order quantities
Mean	465.6616	514.8496	1484.0059	1305.8295
Median	416.7998	550.1843	1237.3684	1148.7264
Min	0	0	0	0
Max	3364.274	1823.2843	12767.2345	7269.5898
Standard Deviation	480.9087	473.5978	1669.3511	1333.0167
Unique Values	235	212	235	212
Coefficient of Variance	1.032742876	0.919876018	1.124895191	1.02081987

Volume Weighted Average Price Analytics

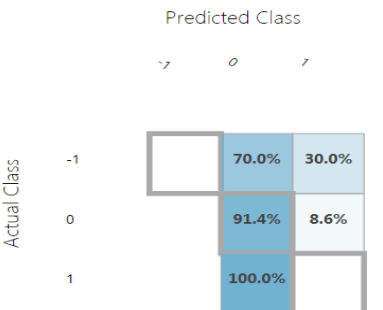
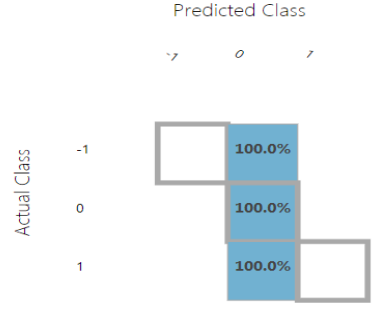
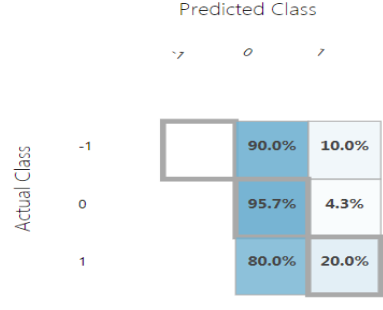
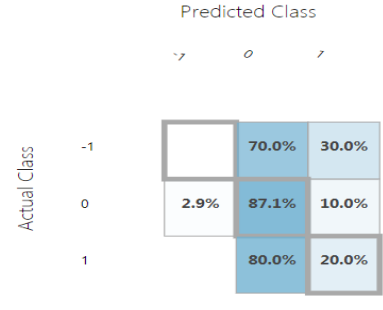
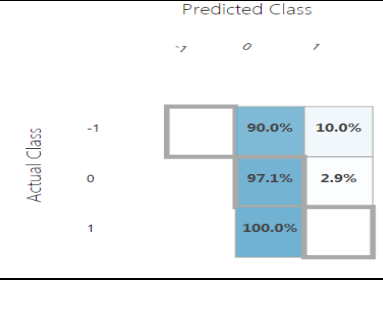
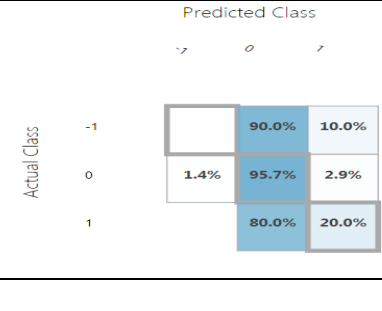
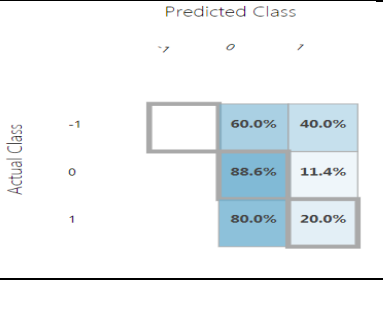
Statistics	VWAP of trades	VWAP of all trades	VWAP of buyer-initiated trades	VWAP of seller-initiated trades
Mean	16.7543	8.6346	0	9.004
Median	32.098	0	0	0
Min	0	0	0	0
Max	33.1604	33.03	0	33.1489
Standard Deviation	16.2939	14.4127	0	14.5884
Unique Values	176	91	1	95
Coefficient of Variance	0.972520487	1.669179811	0	1.620213239

Buyer Seller Analytics

Statistics	Number of buyer-initiated trades	Number of seller-initiated trades	Quantity of buyer-initiated trades	Quantity of seller initiated trades	Buyer / seller ratio 1	Buyer / seller ratio 2	Cumulative buyer / seller ratio 1	Cumulative buyer / seller ratio 2
Mean	15.0765	11.2853	6535.2294	4392.5735	1.0724	7.47	0.4069	0.3925
Median	0	0	0	0	0.0417	0.0429	0	0
Min	0	0	0	0	0	0	0	0
Max	612	333	235416	127198	21	1792	1.9982	1.6202
Standard Deviation	48.3327	35.6824	22672.2915	14379.7377	2.4245	97.7685	0.5635	0.5595
Unique Values	57	55	126	97	151	178	125	118
Coefficient of Variance	3.205830266	3.161847713	3.469241875	3.273647601	2.260816859	13.08815261	1.384861145	1.425477707

APPENDIX-C: ALGORITHM RESULTS

Multi Class Supervised Classification Algorithm Results for THY

	Decision Tree	Logistic Regression	Neural Network	Decision Jungle
	PCA & Default Parameters			
Average accuracy	0.835294	0.882353	0.866667	0.819608
Confusion Matrix	Predicted Class ↘ 0 ↗ 	Predicted Class ↘ 0 ↗ 	Predicted Class ↘ 0 ↗ 	Predicted Class ↘ 0 ↗ 
	PCA & Cross Validation			
Overall accuracy	0.829412	0.838235	0.835294	0.838235294
	PCA & Tune Model Parameters			
Average accuracy	0.866667	0.866667		0.827451
Confusion Matrix	Predicted Class ↘ 0 ↗ 	Predicted Class ↘ 0 ↗ 		Predicted Class ↘ 0 ↗ 

Multi Class Supervised Classification Algorithm Results for Yapi Kredi

	Decision Tree	Logistic Regression	Neural Network	Decision Jungle																																																
PCA & Default Parameters																																																				
Average accuracy	0.615686	0.694118	0.678431	0.6																																																
Confusion Matrix	<p style="text-align: center;">Predicted Class</p> <p style="text-align: center;">↘ 0 ↗</p> <table border="1"> <tr> <td>Actual Class -1</td> <td>11.1%</td> <td>83.3%</td> <td>5.6%</td> </tr> <tr> <td>Actual Class 0</td> <td>23.9%</td> <td>56.5%</td> <td>19.6%</td> </tr> <tr> <td>Actual Class 1</td> <td>4.8%</td> <td>57.1%</td> <td>38.1%</td> </tr> </table>	Actual Class -1	11.1%	83.3%	5.6%	Actual Class 0	23.9%	56.5%	19.6%	Actual Class 1	4.8%	57.1%	38.1%	<p style="text-align: center;">Predicted Class</p> <p style="text-align: center;">↘ 0 ↗</p> <table border="1"> <tr> <td>Actual Class -1</td> <td></td> <td>100.0%</td> <td></td> </tr> <tr> <td>Actual Class 0</td> <td></td> <td>100.0%</td> <td></td> </tr> <tr> <td>Actual Class 1</td> <td></td> <td>100.0%</td> <td></td> </tr> </table>	Actual Class -1		100.0%		Actual Class 0		100.0%		Actual Class 1		100.0%		<p style="text-align: center;">Predicted Class</p> <p style="text-align: center;">↘ 0 ↗</p> <table border="1"> <tr> <td>Actual Class -1</td> <td></td> <td>100.0%</td> <td></td> </tr> <tr> <td>Actual Class 0</td> <td>4.3%</td> <td>95.7%</td> <td></td> </tr> <tr> <td>Actual Class 1</td> <td>14.3%</td> <td>85.7%</td> <td></td> </tr> </table>	Actual Class -1		100.0%		Actual Class 0	4.3%	95.7%		Actual Class 1	14.3%	85.7%		<p style="text-align: center;">Predicted Class</p> <p style="text-align: center;">↘ 0 ↗</p> <table border="1"> <tr> <td>Actual Class -1</td> <td>5.6%</td> <td>83.3%</td> <td>11.1%</td> </tr> <tr> <td>Actual Class 0</td> <td>13.0%</td> <td>52.2%</td> <td>34.8%</td> </tr> <tr> <td>Actual Class 1</td> <td>4.8%</td> <td>52.4%</td> <td>42.9%</td> </tr> </table>	Actual Class -1	5.6%	83.3%	11.1%	Actual Class 0	13.0%	52.2%	34.8%	Actual Class 1	4.8%	52.4%	42.9%
Actual Class -1	11.1%	83.3%	5.6%																																																	
Actual Class 0	23.9%	56.5%	19.6%																																																	
Actual Class 1	4.8%	57.1%	38.1%																																																	
Actual Class -1		100.0%																																																		
Actual Class 0		100.0%																																																		
Actual Class 1		100.0%																																																		
Actual Class -1		100.0%																																																		
Actual Class 0	4.3%	95.7%																																																		
Actual Class 1	14.3%	85.7%																																																		
Actual Class -1	5.6%	83.3%	11.1%																																																	
Actual Class 0	13.0%	52.2%	34.8%																																																	
Actual Class 1	4.8%	52.4%	42.9%																																																	
PCA & Cross Validation																																																				
Overall accuracy	0.402941176	0.435294	0.473529	0.491176																																																
PCA & Tune Model Parameters																																																				
Average accuracy	0.694118	0.694118		0.694118																																																
Confusion Matrix	<p style="text-align: center;">Predicted Class</p> <p style="text-align: center;">↘ 0 ↗</p> <table border="1"> <tr> <td>Actual Class -1</td> <td></td> <td>100.0%</td> <td></td> </tr> <tr> <td>Actual Class 0</td> <td></td> <td>97.8%</td> <td>2.2%</td> </tr> <tr> <td>Actual Class 1</td> <td></td> <td>95.2%</td> <td>4.8%</td> </tr> </table>	Actual Class -1		100.0%		Actual Class 0		97.8%	2.2%	Actual Class 1		95.2%	4.8%	<p style="text-align: center;">Predicted Class</p> <p style="text-align: center;">↘ 0 ↗</p> <table border="1"> <tr> <td>Actual Class -1</td> <td></td> <td>100.0%</td> <td></td> </tr> <tr> <td>Actual Class 0</td> <td></td> <td>100.0%</td> <td></td> </tr> <tr> <td>Actual Class 1</td> <td></td> <td>100.0%</td> <td></td> </tr> </table>	Actual Class -1		100.0%		Actual Class 0		100.0%		Actual Class 1		100.0%			<p style="text-align: center;">Predicted Class</p> <p style="text-align: center;">↘ 0 ↗</p> <table border="1"> <tr> <td>Actual Class -1</td> <td></td> <td>100.0%</td> <td></td> </tr> <tr> <td>Actual Class 0</td> <td></td> <td>100.0%</td> <td></td> </tr> <tr> <td>Actual Class 1</td> <td></td> <td>100.0%</td> <td></td> </tr> </table>	Actual Class -1		100.0%		Actual Class 0		100.0%		Actual Class 1		100.0%													
Actual Class -1		100.0%																																																		
Actual Class 0		97.8%	2.2%																																																	
Actual Class 1		95.2%	4.8%																																																	
Actual Class -1		100.0%																																																		
Actual Class 0		100.0%																																																		
Actual Class 1		100.0%																																																		
Actual Class -1		100.0%																																																		
Actual Class 0		100.0%																																																		
Actual Class 1		100.0%																																																		

Multi Class Supervised Classification Algorithm Results for Garanti

	Decision Tree	Logistic Regression	Neural Network	Decision Jungle
PCA & Default Parameters				
Average accuracy	0.858824	0.87451	0.866667	0.858824
Confusion Matrix	<p style="text-align: center;">Predicted Class -1 0 1</p>	<p style="text-align: center;">Predicted Class -1 0 1</p>	<p style="text-align: center;">Predicted Class -1 0 1</p>	<p style="text-align: center;">Predicted Class -1 0 1</p>
PCA & Cross Validation				
Overall accuracy	0.794117647	0.832353	0.835294	0.8
PCA & Tune Model Parameters				
Average accuracy	0.87451	0.866667		0.87451
Confusion Matrix	<p style="text-align: center;">Predicted Class -1 0 1</p>	<p style="text-align: center;">Predicted Class -1 0 1</p>		<p style="text-align: center;">Predicted Class -1 0 1</p>

Multi Class Supervised Classification Algorithm Results for Aselsan

	Decision Tree	Logistic Regression	Neural Network	Decision Jungle
PCA & Default Parameters				
Average accuracy	0.945098	0.937255	0.960784	0.929412
Confusion Matrix	<p style="text-align: center;">Predicted Class ↘ 0 ↗</p>	<p style="text-align: center;">Predicted Class ↘ 0 ↗</p>	<p style="text-align: center;">Predicted Class ↘ 0 ↗</p>	<p style="text-align: center;">Predicted Class ↘ 0 ↗</p>
PCA & Cross Validation				
Overall accuracy	0.923529412	0.923529	0.923529	0.926471
PCA & Tune Model Parameters				
Average accuracy	0.960784	0.960784		0.960784
Confusion Matrix	<p style="text-align: center;">Predicted Class ↘ 0 ↗</p>	<p style="text-align: center;">Predicted Class ↘ 0 ↗</p>		<p style="text-align: center;">Predicted Class ↘ 0 ↗</p>

REFERENCES

- Admati, A. R. (1988). A theory of intraday patterns: Volume and price variability. *The Review of Financial Studies*, 1(1), 3-40.
- Barclay, M. J. (1993). Stealth trading and volatility: Which trades move prices? *Journal of Financial Economics*, 34(3), 281-305.
- Bloomberg. (2019, 07 07). BIST 100 Stock Prices.
- Borsa İstanbul. (2019, 07 07). *Equity Market Data Analytics*. Retrieved from Borsa İstanbul: <https://www.borsaistanbul.com/en/data/data-dissemination/equity-market-data-analytics>
- Borsa İstanbul. (2019, 07 07). *List of Equity Market Data Analytics*. Retrieved from Borsa İstanbul: <https://www.borsaistanbul.com/docs/default-source/veriler/list-of-the-equity-market-data-analytics.xlsx?sfvrsn=2>
- Foster, F. D. (1993). Variations in trading volume, return volatility, and trading costs: Evidence on recent price formation models. . *The Journal of Finance*, 48(1), 187-211.
- Gunduz, H., Yaslan, Y., & Cataltepe, Z. (2017). Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. *Knowledge-Based Systems*, 138-148.
- Hautsch, N. (2001). Modelling intraday trading activity using Box-Cox ACD models. Available at SSRN 289643.
- Ibrahim, A. H. (2014, October). *Data Normalization and Standardization for Neural Networks Output Classification*. Retrieved from <https://ahmedhanibrahim.wordpress.com>: <https://ahmedhanibrahim.wordpress.com/2014/10/10/data-normalization-and-standardization-for-neural-networks-output-classification/>
- Investopedia. (2019, 07 07). *Technical Analysis*. Retrieved from Investopedia: <https://www.investopedia.com/terms/t/technicalanalysis.asp>
- Investopedia. (2019, 07 07). *World High Frequency Algorithmic Trading*. Retrieved from Investopedia: <https://www.investopedia.com/articles/investing/091615/world-high-frequency-algorithmic-trading.asp>
- Mittermayer, M. A. (2004). Forecasting intraday stock price trends with text mining techniques. In *37th Annual Hawaii International Conference on System Sciences*, Proceedings of the (pp. 10-pp). IEEE.

Navlani, A. (2018, September 7). *Understanding Logistic Regression in Python*. Retrieved from Data Camp: <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>

Navlani, A. (2018, May 16). *Understanding Random Forests Classifiers in Python*. Retrieved from Data Camp: <https://www.datacamp.com/community/tutorials/random-forests-classifier-python>

Navlani, A. (2019, January 18). *Neural Network Models in R*. Retrieved from Data Camp: <https://www.datacamp.com/community/tutorials/neural-network-models-r>

Shang, H. (2017). Forecasting intraday S&P 500 index returns: A functional time series approach. *Journal of Forecasting* , 741-755.

Sharma, A. (2019, September 06). *Principal Component Analysis (PCA) in Python*. Retrieved from Data Camp: <https://www.datacamp.com/community/tutorials/principal-component-analysis-in-python>

Shotton, J., Sharp, T., Nowozin, P. K., Winn, J., & Criminisi, A. (2013). Decision jungles: Compact and rich models for classification (. *In Advances in Neural Information Processing Systems*, pp. 234-242.

Vella, V. &. (2014). Enhancing risk-adjusted performance of stock market intraday trading with neuro-fuzzy systems. *Neurocomputing*, 141, 170-187.

Vella, V. (2014). Enhancing risk-adjusted performance of stock market intraday trading with neuro-fuzzy systems . *Neurocomputing*, 141, 170-187.

Yingsaeree, C. (2012). *Algorithmic trading: Model of execution probability and order placement strategy*. (Doctoral dissertation, UCL (University College London)).