T.C BAHÇEŞEHİR UNIVERSITY

FORECASTING

PROFIT &LOSS PRICES AND PROFITABILITY OF COMPANIES

ALKAN ULUTAŞ

Computer Engineering Master Program, 2018

Primary Supervisor: Asst. Prof. C. Okan Şakar

# CONTENTS

[CONTENTS 2](#_Toc503395151)

[ÖZET 3](#_Toc503395152)

[ABSTRACT 4](#_Toc503395153)

[1 INTRODUCTION 5](#_Toc503395154)

[2 LITERATURE REVIEW 6](#_Toc503395155)

[3 DEFINITION OF BALANCE SHEET 7](#_Toc503395156)

[4 DEFINITION OF ALGORITHMS 8](#_Toc503395157)

[4.1 LINEAR REGRESSION 8](#_Toc503395158)

[4.2 DECISION TREE 8](#_Toc503395159)

[4.3 RANDOM FOREST 8](#_Toc503395160)

[4.4 SUPPORT VECTOR MACHINE 9](#_Toc503395161)

[4.5 MULTILAYER PERCEPTRON 9](#_Toc503395162)

[5 EXPERIMENT AND CONCLUSION 11](#_Toc503395163)

[5.1 DATASET INTRODUCTION 11](#_Toc503395164)

[5.1.1 DATASET COLLECTION 11](#_Toc503395165)

[5.2 EXPERIMENTAL SETUP 14](#_Toc503395166)

[5.3 DEFINITION of METRICS 14](#_Toc503395167)

[5.4 RESULTS OF EXPERIMENTS 15](#_Toc503395168)

[5.4.1 EXPERIMENT 1 15](#_Toc503395169)

[5.4.2 EXPERIMENT 2 16](#_Toc503395170)

[6 CONCLUSION 18](#_Toc503395171)

[7 ACKNOWLEDGMENT 19](#_Toc503395172)

[8 REFERENCES 20](#_Toc503395173)

# LIST OF FIGURES

4.4 Support Vectors

4.5: Multilayer Perceptron

# ÖZET

Makine Öğrenme Algoritmalarını uygulamak; finansal işleri daha kolay ve güvenilir hale getirerek, çok sayıda sektör için başarılı adımlar atılmasına sebebiyet vermektedir. Bu algoritmalar şirket sahiplerinin veya yatırımcıların daima kâr elde etme isteklerine yardımcı olarak, geleceği tahminleyen günümüz ihtiyaçlarının başında gelmektedir. Makine öğrenimi geleceği tahminleme noktasında devreye girerek; şirket sahiplerine ve yatırımcılara algoritma ve modelleri ile kolaylık sağlayarak yardımcı olur. Finansal sektörün doğası gereği ekonomiyi etkileyen ve bilinmeyen çok fazla faktörlerden dolayı şirkete ait kar veya zararı tahmin etmek oldukça güçtür.

Bu çalışmada ilk aşama olarak; şirketlere ait bilanço verileri kullanılarak gelecek döneme ait kar ve zarar tutarı regresyon algoritmasıyla tahmin edilmeye çalışılacaktır. Aynı veri seti kullanılarak yapılacak olan bir sonraki çalışmada ise gelecek bilanço döneminin karlı ya da zararlı olup / olmayacağı sınıflandırma algoritmalarıyla tahmin edilecektir. Bu tahminlemelerde Linear Regression, Simple Linear Regression, Random Forest, Decision Table, J48, Support Vector Machine, Multilayer Perceptron algoritmaları kullanılacaktır.

Veri setinden de bahsetmek gerekirse Borsa İstanbul’da yer alan imalat sanayi şirketlerinin geçmiş bilanço dönemlerine ait tüm finansal kalemlerin yer aldığı bir veri setidir. Veri seti Foreks Bilgi İletişim Hizmetleri A.Ş Foreks tarafında sunulmuş olup, çalışmanın ilerleyen aşamalarında veri setiyle ilgili detaylı bilgi paylaşılmaktadır.

# ABSTRACT

Applying Machine Learning Algorithms are making financial works easier, more reliable and getting more success for lots of sectors. One of the needs at these sectors is predicting profit/loss of firms for their surviving since owners or investors always want to make profit. Machine Learning takes place at this point and provides advantage to investors and owners by its algorithms and models since it is so hard to predict profit or loss which is very dynamic and susceptible to quick changes because of the underlying nature of the financial domain and unknown factors which effects economy (like Election Results, Rumors etc.)[1].

In this study; first step is predicting the net profit/loss price of firms and second one is forecasting the economic situation of firms in next financial terms will be profitable or not profitable for owners and investors by using different classification methods which are Linear Regression, Simple Linear Regression, Random Forest, Decision Table, J48, Support Vector Machine, Multilayer Perceptron. On the other hand, if mentioning the dataset; it is a data set in which all the financial items belonging to the past balance sheet periods of the manufacturing industry companies belonging to the Istanbul Stock Exchange are included.

**Keywords-**Data Mining, Machine Learning, Profit/loss prediction, Prediction, Forecasting balance sheet, Financial forecast, Linear regression, Simple linear regression, Random forest, J48, Multilayer perceptron, Support vector machine, Decision table.

# INTRODUCTION

Nowadays, everything is shifting from manual to automated systems. The aim of this study is to predict the profit and loss prices so as to provides near truth data and help owners and investors for their strategies. The common method for predicting profit or loss of firms depends on experienced management teams of firms that making prediction to firm will gain money or not. But this method is risky and depends on non-financial professionals so thus leading to bankruptcy of firms, loss of the investments and future of firms.

Because of all these reasons there is a need for an automated system. Machine Learning algorithms could be used to help owners and investors to survive their firms in reliable methodology and strategy. Also, the new system will be cost and time efficient.

This paper is about predicting the profit/loss price of firms and it provides information about the profitability status of the next terms for investors and owners by using balance sheets of firms which have fully detailed features about balance sheet items.

The research efficiently analyses previous balance sheet details to predict the future for profit or loss price, also forecasting the profitability status of next terms which is based on types of regression models and classification algorithms as Linear Regression, Simple Linear Regression, Random Forest, Decision Table, J48, Support Vector Machine, Multilayer Perceptron.

The study consists of 320 attributes and 923 instances about financial terms, and an application programming interface (API) for its integration which is gotten from Foreks Bilgi İletişim Hizmetleri A.Ş.to forecast net profit/lost price.

The study also would be very beneficial for economic forecasters and prediction market since it is hard to find good system which makes good predictions about profit or loss price of firms and provide almost as real values.

# LITERATURE REVIEW

Many sources include models that make forecasts on financial tables. Corporate financing estimators focus on estimating new funds or sources of external financing for large firms with historical financial data. Both approaches are well-suited for the analysis of long-term capital requirements of established firms [5].

Why working with forecasted financial statements is necessary? Forecasting financial statements is not optional for the management because it can provide a guide to the future performance of the firm [2].

For instance, the business environment in New Zealand and the special characteristic of small and medium-sized enterprise (SME) are facing in terms of budgeting and the importance of accurate budgets [3]. Neural Network, LibSVM, SMO, Multilayer Perceptron, RBF Network, Naïve Bayes, AD Tree, BF Tree, IBk models are applied on dataset. In conclusion, SME owners with limited accounting and financial knowledge could create budgets without the hassle of managing error-prone spreadsheets [3].

Another related study is about how well could bank balances be predicted? Traditionally these models have been applied to predictions of stock market data. Support Vector Machines applied on financial dataset for this study and it has shown that it is possible to create accurate predictions of bank balances through the use of support vector machines [4].

# DEFINITION OF BALANCE SHEET

The Balance Sheet is a financial statement showing the entity's assets at a certain date and the sources of those assets. The balance sheet is one of the most important statements of the accounting system. It is extremely useful in terms of balance sheet operations.

The assets of the balance sheet are divided into two main groups as "Current assets" and "Fixed assets". Current assets consist of sub-accounts such as bank accounts, receivables to be collected in a short time and product stocks of the company. Note that such assets have the ability to cash in or swiftly move into cash, and we can call returning assets "liquid assets.

Fixed assets include immovable (tangible fixed assets) such as partnerships (financial fixed assets), trademarks, patents and license rights (intangible fixed assets), possessed buildings or machines, ownership of investments and long-term receivables as well as assets that are not readily convertible into cash. The sum of current assets and fixed assets is called "total assets".

An entity's asset structure may vary depending on the sector in which it is active, for example, assets held by industrial companies may be proportionately higher in the total assets, while those operating in the service sector may be expected to have a higher proportion of current assets. Generally, the structure of the asset is questioned in terms of "liquidity". For example; there is a risk that the liquid assets of enterprises with short-term liabilities are not sufficient.

Liability is composed of "Short-term liabilities", "Long-term liabilities" and "Equity" account groups. In summary, short-term liabilities consist of loans and trade payables that the operator must pay in a short period of less than 1 year and long-term liabilities consist of financial and commercial liabilities to be payable after 1 year. For example; an entity that will receive an investment loan with a maturity of 10 years will pay a short-term portion in the next year and the remaining portion is a long-term financial obligation.  
Equity, or, in other words, equity, is the accumulation of profits from capital and past years that it has put into operation in proportion to the partnerships of its business partners. In this case; it is seen that assets (assets) are funded by financial and commercial debts and equity (passive) when considering the balance sheet structure. It is generally welcomed that the self capital is strong, that is, the debts on the passive side of the balance are lower than the equity capital. The balance sheet can be expressed in the simplest form by the following formula:

*assets = debts + capital*

The balance sheet is periodically presented as "interim" and "year-end". 3rd, 6th and 9th months of the year, the balance sheet issued at the 12th month is called "year end balance sheet". From an analytical point of view, the calculations are tracked for interim periods or are used in comparison with the balance sheet periods in previous years.

In this study, we will use some features of balance sheet to predict net profit/loss prices and next term profitability of companies, in next sections more details will be given about financial dataset.

# DEFINITION OF ALGORITHMS

## LINEAR REGRESSION

Linear Regression is one of the simplest supervised learning algorithms. We assume that we have a linear relationship between the Y variable and our estimator variables X1, X2, ..., Xn. This method seems extremely simple, but it is very useful both conceptually and practically.

## DECISION TREE

The Decision Tree learning method is one of machine learning topics. There are applications in the literature such as a classification tree or a regression tree that can be considered as sub-methods of decision tree learning. In decision tree learning, a tree structure is created, class labels at the level of the leaves of the tree, and the processes on the properties with the leaves that go to these leaves and from the beginning. During decision tree learning, learned information is modeled on a tree. All the interior nodes of this tree represent the inputs. [6]

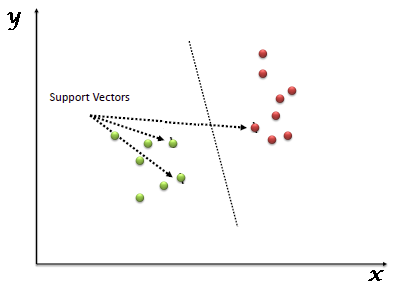
In decision tree learning, when a tree is learned, the cluster on which it is trained is divided into subclasses according to various properties, which is repeated recursively and continues until there is no effect on the prediction of the repetition. This process is called recursive partitioning. In general, the way data comes in during data mining is as follows:(x,Y) = (x1,x2,x3,x4,…, Y). According to this notation, the values ​​from X1 to Xn are the inputs to the system, while the value of Y is the desired value to be obtained as the output of the system. [6]

## RANDOM FOREST

Random Forest is one of the most used methods in data mining. When a data is to be classified according to the decision tree, a known data set of class labels is needed. By applying decision-making steps on the data set, a large number of recorded data are divided into groups of few as trees and leaves. When each division operation is performed, they become similar to each other in terms of their characteristics.

## SUPPORT VECTOR MACHINE

A Support Vector Machine (SVM) is a supervised machine learning algorithm. It is generally used for both classification and regression purposes that hyperplane best divides a dataset into two classes, as shown in the image below.

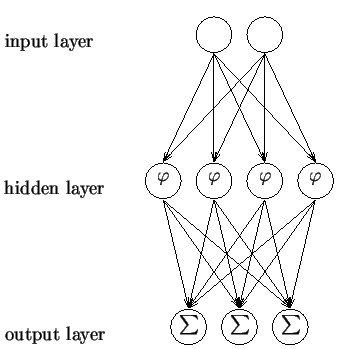


*Figure 4.4: Suppor Vectors*

Hyperplane is drawn randomly and an error is calculated according to the distances of the erroneous data points in the region of the separation. A new line is drawn in small steps by shifting the first line or by changing the slope, learning is made according to the variation in the line and the most suitable separation line is created. The best line of separation is the line that maximizes the distance between these points by dividing the points close to each other.

## MULTILAYER PERCEPTRON

An MLP is a network of simple *neurons* called *perceptrons*. The basic concept of a single perceptron was introduced by Rosenblatt in 1958. Perceptron, although extremely limited, is the oldest one of the neuron networks. Perceptron, a neuron cell based on the principle of producing an output by taking a lot of input. Perceptron could be used in problems divided into two parts by a linear function. Multilayer neuron network architecture is shown below:



*Figure 4.5: Multilayer Perceptron*

**Input Layer:** It takes the inputs from the outside world and sends them to the middle layer. There is no information processing in this layer. Every incoming information goes to the next layer as it arrives. Each process element has only one input and one output. That is, each process element in the input layer is connected to all of the process elements in the next layer.

**Hidden Layer:** The hidden layers process the information coming from the input layer and send it to the next layer. A multi-layered network may have multiple hidden layers and multiple process elements in each layer.

**Output Layer:** It processes the information from the hidden layer and sends out the output to the outside world.

# EXPERIMENT AND CONCLUSION

## DATASET INTRODUCTION

### DATASET COLLECTION

Balance Sheets of companies are gotten from Foreks Bilgi İletişim Hizmetleri A.Ş who serves dataset to us by providing application programming interface (API) as each balance sheet period. Dataset is prepared for only manufacturing industry sector which has the highest number of company and is combined four balance sheet periods as attributes for predicting profit and loss price of next period. For instance; finding the profit and loss price of first balance sheet period of 2014 (2014-1) by using four balance sheet period of 2013 (2013-1, 2013-2, 2013-3, 2013-4)

Balance Sheets belong to years between 2013 and 2017 which include 320attributes (80 x 4 periods) and 923 instances about financial terms.

Dataset features are shown below:

|  |  |  |
| --- | --- | --- |
| Feature Number | Feature Description | Data Type |
| feature 1 | working assets | numeric |
| feature 2 | cash and cash equivalents | numeric |
| feature 3 | trade receivables | numeric |
| feature 4 | trade receivables from affiliates | numeric |
| feature 5 | trade receivables from nonaffiliated | numeric |
| feature 6 | receivables from financial activities | numeric |
| feature 7 | receivables from non-related parties of financial sector operations | numeric |
| feature 8 | other receivables | numeric |
| feature 9 | other receivables from related parties | numeric |
| feature 10 | other receivables from unrelated parties | numeric |
| feature 11 | derivative financial instruments | numeric |
| feature 12 | Stocks | numeric |
| feature 13 | biological asset | numeric |
| feature 14 | prepaid expenses | numeric |
| feature 15 | assets related to current period tax | numeric |
| feature 16 | other current assets | numeric |
| feature 17 | subtotal | numeric |
| feature 18 | assets classified as held for sale | numeric |
| feature 19 | real assets | numeric |
| feature 20 | financial investments | numeric |
| feature 21 | other receivables | numeric |
| feature 22 | other receivables from related parties | numeric |
| feature 23 | other receivables from unrelated parties | numeric |
| feature 24 | investments valued by equity method | numeric |
| feature 25 | tangible assets | numeric |
| feature 26 | intangible assets | numeric |
| feature 27 | special assessment | numeric |
| feature 28 | other intangible assets | numeric |
| feature 29 | prepaid expenses | numeric |
| feature 30 | deferred tax assets | numeric |
| feature 31 | other fixed assets | numeric |
| feature 32 | total assets | numeric |
| feature 33 | Resources | numeric |
| feature 34 | short-term liabilities | numeric |
| feature 35 | short-dated loan | numeric |
| feature 36 | short-term portions of long-term borrowings | numeric |
| feature 37 | other financial liabilities | numeric |
| feature 38 | trade payables | numeric |
| feature 39 | trade payables to related parties | numeric |
| feature 40 | trade payables to unrelated parties | numeric |
| feature 41 | liabilities under employee benefits | numeric |
| feature 42 | other liabilities | numeric |
| feature 43 | other liabilities to related parties | numeric |
| feature 44 | other liabilities to unrelated parties | numeric |
| feature 45 | derivative financial instruments | numeric |
| feature 46 | government incentives and grants | numeric |
| feature 47 | deferred incomes | numeric |
| feature 48 | current income tax liability | numeric |
| feature 49 | short-term provisions | numeric |
| feature 50 | short-term provisions related to employee benefits | numeric |
| feature 51 | other short-term provisions | numeric |
| feature 52 | other short term liabilities | numeric |
| feature 53 | Subtotal | numeric |
| feature 54 | liabilities of asset groups classified as sales | numeric |
| feature 55 | long-term liabilities | numeric |
| feature 56 | long-term borrowings | numeric |
| feature 57 | long-term provisions | numeric |
| feature 58 | long-term provisions for employee benefits | numeric |
| feature 59 | other long-term provisions | numeric |
| feature 60 | liabilities related to current period tax | numeric |
| feature 61 | deferred tax liabilities | numeric |
| feature 62 | other long-term liabilities | numeric |
| feature 63 | owner’s equity | numeric |
| feature 64 | equity of the parent company | numeric |
| feature 65 | paid capital | numeric |
| feature 66 | capital adjustment differences | numeric |
| feature 67 | recovered shares | numeric |
| feature 68 | premiums and discounts on shares | numeric |
| feature 69 | other comprehensive income and expenses that will not be reclassified to profit or loss | numeric |
| feature 70 | revaluation and measurement gain / losses | numeric |
| feature 71 | other earnings / losses | numeric |
| feature 72 | other comprehensive income and expenses that will be reclassified to profit or loss | numeric |
| feature 73 | foreign currency translation differences | numeric |
| feature 74 | hedonic gain / losses | numeric |
| feature 75 | revaluation and classification gains / losses | numeric |
| feature 76 | other earnings / losses | numeric |
| feature 77 | restricted reserves | numeric |
| feature 78 | profits / losses of previous years | numeric |
| feature 79 | net period profit / loss | numeric |
| feature 80 | total resources | Numeric |

Table 1: Attribute Details

## EXPERIMENTAL SETUP

WEKA is an open-source data mining application developed by University of Waikato [3].In first experiment (E1); Linear Regression, Simple Linear Regression, Decision Table, Random Forest classification algorithms are applied on dataset by cross validation (10-Folds)for predicting profit and loss price of next period. In second experiment (E2); Random Forest, J48, Support Vector Machine, Multilayer Perceptron are applied on dataset by cross validation (10-Folds) for predicting next period will be more profitable or not profitable.

## DEFINITION of METRICS

The formulas used during the study are shown below:

***Accuracy****= (Number of Correctly Classified Examples) / (Number of Examples)*

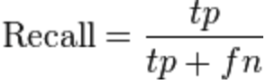


*Figure 5.3.1 Accuracy Formula*

***Precision****= (Number of True Positives) / (Total Number of Predicted Positives)*

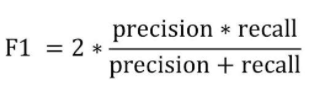


*Figure 5.3.2 Precision Formula*

***Recall****= (Number of True Positives) / (Actual Number of Positives)*

*Figure 5.3.3 Recall Formula*

***F-measure****= 2\*Recall\*Precision / (Recall + Precision)*



*Figure 5.3.4 F-measure Formula*

*****MCC****= ((True Positive \* True Negative) – (False Positive\*False Negative)) / square root ((True Positive + False Positive) \* (True Positive + False Negative) \* (True Negative + False Positive) \* (True Negative + False Negative))*

*Figure 5.3.5 MCC Formula*

## RESULTS OF EXPERIMENTS

### EXPERIMENT 1

Purpose of the E1 is predicting profit and loss price of next period, so that Simple Linear Regression, Linear Regression, Decision Table and Random Forest are applied for 320 attributes and 923 instances by cross validation (10-Folds).The classification models and results are saved on Table 2.

The correlation coefficient is a value used to represent the relationship between variables. The variable takes a value between -1 and 1 when calculated using the following formula. Negative values ​​indicate a negative relationship, while positive values ​​indicate a positive relationship. When the values ​​are 1 or -1, there is a perfect relationship. When the values ​​approach 0, the relationship between the variables decreases.

The correlation coefficient, mean absolute error, root mean squared error and relative absolute error and root relative squared error values of each classification models are shown Table 2.

As you see on Table 2; Simple Linear Regression has the highest Correlation Coefficient value, so that represents the relationship between variables seems better than other models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Simple Linear Regression | Linear Regression | Decision Table | Random Forest |
| Correlation coefficient | 0.9319 | 0.6296 | 0.8031 | 0.8909 |
| Mean absolute error | 117196962604 | 222367692091 | 172294399183 | 124157156523 |
| Root mean squared error | 291690805639 | 781839916386 | 481031357384 | 365814029529 |
| Relative absolute error | 36.005 % | 68.3153 % | 52.9319 % | 38.1433 % |
| Root relative squared error | 36.2412 % | 97.14 % | 59.7659 % | 45.4507 % |

Table 2: Result of Experiment 1

### EXPERIMENT 2

Purpose of the E2 is forecasting the economic situation of firms in next financial terms will be profitable or not profitable for owners and investors. Random Forest, J48, Support Vector Machine and Multilayer Perceptron are applied for 320 attributes and 923 instances by cross validation (10-Folds). The classification models and results are saved on Table 2.

The accuracy, F-measure, MCC and precision and recall values of each classification models are shown Table 3.

Support Vector Machine (Linear) has the highest accuracy value, since number of correctly classified examples is very high.

In addition, Precision value is also very high because of number of True Positives is pretty high. For Recall, we could also make the same comment.

Accuracy is pretty high and F-measure, Precision and Recall values are so close to each other, so that represents applying Support Vector Machine (Linear) will provide better results for our experiments.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Random Forest | J48 | SVM  Linear | SVM  RBF | Multilayer Perceptron |
| Accuracy | 63.1636 | 82.8819 | 87.2156 | 59.0466 | 58.5049 |
| F-Measure | 0.713 | 0.853 | 0.892 | 0.743 | 0.722 |
| MCC | 0.213 | 0.648 | 0.735 | 0 | 0.042 |
| Precision | 0.660 | 0.863 | 0.887 | 0.590 | 0.597 |
| Recall | 0.774 | 0.844 | 0.897 | 1 | 0.912 |

Table 3: Result of Experiment 2

# CONCLUSION

In this study; with Support Vector Machine and Linear Regression algorithms for forecasting of profit/loss price and next term profitability are addressed by using Balance Sheets dataset of firms, since dataset is linear and suitable for Support Vector Machine and Linear Regression algorithms.

The four financial terms of firms are combined for modeling to predict each next term of profit/loss price and next term profitability of firms.

As a result, by adding more feature and financial term to dataset; time series forecasting would be good study.

# ACKNOWLEDGMENT

I must acknowledge all the support given to me by the staff from Foreks Bilgi İletişim Hizmetleri A.Ş., especially Business Analyst Sevde Kuş. Lastly, special thanks to my supervisor Asst. Prof. Cemal Okan Şakar and Bahçeşehir University for their supports.

# REFERENCES

[1] Stock Trend Prediction by Using K-Means and Apriori All Algorithm for Sequential Chart Pattern Mining, KUO-PING WU, YUNG-PIAO WU AND HAHN-MING LEE, 2014

[2] Forecasting Financial Statements with No plugs and No Circularity, Ignacio Vélez–Pareja, 2009

[3] Determining the Accuracy of Budgets A Machine Learning Application for Budget Change Pattern Recognition, Kai Leung Yip, 2012

[4] Ruey S. Tsay. Analysis of Financial Time Series, Second Edition. JohnWiley & Sons, Inc., 2005

[5] Forecasting Financial Statements Using Risk Management Associates Industry Data, Terrance Jalbert, University of Hawaii at Hilo, James E. Briley, Northeastern State University, Mercedes Jalbert, The Institute for Business and Finance Research, 2012

[6] Şadi Evren ŞEKER, “Karar Ağaçları (Decision Tree)” <http://bilgisayarkavramlari.sadievrenseker.com/> Apr 7, 2012