# Earnings Prediction Using Artificial Intelligence – Support Vector Machines (SVM)

# INTRODUCTION

The ability to predict earnings based on past performance has been recognized as a measure of earnings quality (Penman and Zhang 2002) and earnings announcement may provide only a modest amount of new information to the share market (Ball and Shivakumar 2008) it has been shown that investors over rely on old earnings performance when predicting future earnings performance (Bloomfield et al. 2003). These issues highlight the necessity to develop a tool to better predict future earnings.

Many research papers have concentrated on the importance of earnings announcements and forecasts in the determination of investment decisions. While earlier research has only been able to show relatively low informativeness of earnings (Ball and Brown 1968; Beaver, 1968; Foster, Olsen and Shevlin 1984; and Bernard and Thomas 1990) later studies were able to show the incremental information content of specific components of the financial statements. For example, information for future earnings and cash flows (Finger 1994); prediction of sign changes in the future earnings using various income statement and balance sheet components (Ou 1990; Ou and Penman 1989); current earnings and current price as predictors of future earnings (Shroff 1999); higher information content in bad-news periods than in good-new periods (Roychowdhury and Sletten 2012); and the ability of disaggregated earnings data to predict next period's earnings in the banking industry (Alam and Brown 2006) are just a few of the examples.

Ou and Penman (1989) were the first researchers to focus on the usefulness of accounting information to predict the direction of movement of earnings relative to trend adjusted current earnings. The study is important because it evaluates whether accounting information can consequently be used as the basis for profitable investment strategy. Given investors' reliance on earnings this could be a valuable tool for a profitable investment strategy. The authors found that financial statement analysis can provide a measure to indicate future earnings, which in turn is may be used as a successful investment strategy. However, evidence from subsequent studies (Holthausen and Larker 1992; Bernard et al. 1997; Stober 1992; Setiono and Strong 1998; and Bird, Gerlach and Hall 2001) have been mixed.

Artificial intelligence tools offer a possibility to create more sophisticated and precise models for complex and computationally demanding decision processes, as well as performing analysis and prediction processes faster and more effectively (Danėnas 2013). Support Vector Machines (SVM) is one of these methods widely applied as an effective solution to many various pattern recognition, classification, regression and forecasting problems. SVM has also been applied to financial forecasting, although mainly in the credit risk field. SVM technique has proven itself as an effective solution in the credit risk field with results comparable to or better than most of the other artificial intelligence techniques such as Neural Networks (Danenas and Garsva 2011).

The aim of this paper is to propose a structure for earnings prediction which implements advanced technologies and techniques such as SVM and utilizes the eXtensive Business Reporting Language (XBRL).

The paper is organized as follows, the second section reviews academic literature examining research conducted using SVM and on the validity of XBRL data and its limitations. The third section outlines the decision model process employed, the data used and the results of the model. The last section concludes the paper.

# ACADEMIC RESEARCH

In this section will be presented a review of relevant literature on two issues: SVM methodology and its implementation in financial analysis and XBRL data, validity and limitations.

## SVM Method and its Implementation in Financial Analysis

The ability of accounting information to predict the direction of movement earnings has been studied in accounting literature using many methods. The linear traditional statistical modeling techniques have commonly been used in most cases. The foundation paper in this area (Ou and Penman 1989), which was cited 124 times (according to PROQUEST), used multivariate regression analysis. Following articles used similar statistical methods with varying results (Holthausen and Larcker 1992; Bernard et al. 1997; Stober 1992; Setiono and Strong 1998; Bird et al. 2001).

When linear approximation is not valid the accuracy of traditional statistical modeling significantly decreases (Etemadi et al. 2015). A starting point for using non-linear models to predict earnings was when research found that there might be a non-linear relationship between some accounting variables and future earnings (Abarbanell and Bushee 1997).

Financial time series data is characterized by noise, non-stationary, chaos and high degree of uncertainty. Using machine learnings algorithms for these type of tasks is widely gaining popularity because of the ability of artificial intelligence techniques to map non-linear data (Chandwani and Saluja 2014).

A review of research using financial information to forecast EPS, examined 16 articles published between 1990 and 2015. The independent variables included historical EPS, financial ratio information and general variables (stock price, macroeconomic variables etc.). The review comparing different forecasting techniques, traditional statistical models and artificial intelligence techniques, found that artificial intelligence techniques achieve better forecasting accuracy compared to customary statistical based methods (Rajakumar and Ramya 2017).

A sample of 283 firms U.S. quarterly data between the years 1992-2002 was used to assess the efficacy of NN models, based on financial ratios, in forecasting EPS. Results indicated that NN approach presented superior forecast accuracy over linear ARIMA models. The results were even more pronounced when additional ratios were added, and industry specific models were introduced (Zhang et al. 2004).

Using the same models NN models (as Zhang et al. 2004) the issue was further examined using a sample of 723 Chinese firms quarterly data between the first quarter of 1999 and the third quarter of 2008. The results confirmed the forecast accuracy of NN models, and the improved forecasting accuracy by adding financial ratios (Cao and Gan 2010).

Qiu et al. (2014) used SVM to examine the ability of each of the predictors, EPS and stock return, from one year to predict company performance, measured as the change in each of the predictors, in the following year. The sample consisted of manufacturing industry firms (SIC code 2000-3999) in the United states from 1997 to 2003. Their results showed that the SVM model, for each of the predictors outperforms the majority vote baseline (declare all firms as average preforming, 50% accuracy) in three out of the 6 years examined. The model also, outperforms analysts' forecasts in predicting cumulative return. However, analysts offer improvement, of up to 20% in predicting EPS. They attribute the inability of the SVM model to predict EPS to the fact that analysts' have access to more information. It is interesting to note that when examining the predictive ability of the SVM model over time it performed best for the year 2000 which was considered most volatile and the hardest year to predict for analysts. A model where the training data included all the years previous to the test year, and not just one year previous, showed similar patterns.

While there is little research on earnings forecast using machine learning, there is much more research on stock price predictions using these methods. Neural Networks (NN), which implements the empirical risk minimization principal, is one of the most widely accepted machine learnings techniques for stock price forecasting, its crucial drawback is the over-fitting problem leading to a poor level of generalization. SVM has been gaining increasing popularity in this area as it is said to improve the generalization property of NN; many papers affirm that SVM is a superior technique as SVM decreases the level of risk in information data and leads to a higher degree of accuracy by using a structural method (Sap and Awan 2005).

Although most of the studies, utilizing machine learning, used technical indicators to predict stock price, there is some research which used financial ratios as predictors. Wu and Xu (2006) found that financial ratios can be used to predict stock prices with the aid of NN and rough set theory. The study examined annual data of companies traded on the Chinese stock exchange and found that the NN methodology improves when feature selection through rough set theory, is conducted.

Han and Chen (2007) used 3 financial ratios (Earnings per share, Book Value per share and Net Profit Growth rate) to identify stock with outstanding growth (identified by experts). Their model, using SVM on Chinese publicly traded companies, was able to achieve an accuracy level of 75%-86%.

Data mining methods were used in the modeling of 44 financial ratios in an attempt to predict stock price, the model achieved an 80% accuracy level (Barak and Modarres 2015), and a model using SVM with 14 financial ratios to predict stock price, achieved an accuracy of 77-85% (Huang 2012). Both models improved their efficiency by using feature selection, which is considered a very crucial aspect in financial information modeling (Barak and Modarres 2015; Huang 2012; Tsai and Hsiao 2010).

Raposo and Cruz (2002) used 9 financial ratios to model Brazilian textile companies and predict stock prices. Their results, applying NN, showed an accuracy of 70% which increased to 75% when PCA feature selection was introduced.

The performance accuracy of models based on SVM was found to be one of the top machine learning techniques in an examination of articles published from 2000 to 2015. Of the 30 articles, identified as relevant, 12 used fundamental information, mainly financial ratios, to predict stock price. The prediction accuracy of the SVM models was 96.5%, higher than that of Decision Tree, NN and Bayesian methods (Kamley et al. 2016).

From this review it is evident that there is only limited research in earnings prediction using machine learning techniques, however models utilizing machine learning outperform traditional statistical methods. In general, prediction models, using financial information improve when using machine learning in general, and SVM in particular, with feature selection methods.

## XBRL

XBRL (eXtensible Business Reporting Language) is a freely available and global standard for exchanging business information. XBRL allows the expression of [semantic meaning](https://en.wikipedia.org/wiki/Semantics#Computer_science) commonly required in [business reporting](https://en.wikipedia.org/wiki/Business_reporting). One use of XBRL is to define and exchange financial information, such as financial statements.

The SEC has created the XBRL U.S. GAAP Financial Reporting Taxonomy. This taxonomy is a collection of accounting data concepts and rules that enables companies to present their financial reports electronically. The SEC's deployment was launched in 2008 in phases, and all public U.S. GAAP companies were required to file their financial reports using the XBRL reporting technology starting from June 15, 2011.

XBRL has several advantages over COMPUSTAT, which has been a popular source of financial information for both academics and practitioners. Among XBRL data advantages are the fact that it is freely available while COMPUSTAT is costly. XBRL filings also have a time advantage, it takes an average of 14 weekdays from the time a company files with the SEC for that data to appear in COMPUSTAT (D’Souza et al. 2010), while XBRL data is published concurrently with the related PDF versions, and is immediately available. In addition, the reliability of COMPUSTAT has also been questioned, prior studies have shown that COMPUSTAT data may differ from the original corporate financial (Miguel 1977; Kinney and Swanson 1993; Tallapally et al. 2011) and data found in other accounting databases (Rosenberg and Houglet 1974; Yang et al. 2003).

The quality of the newly mandated SEC XBRL data, used in presenting past earnings performance, is a key factor for the success of its use and implementation for both academics and practitioners. Quality of the data provided by XBRL filings has been measured in several ways, among them: the number of errors in the computation of the filings (Debreceny et al. 2010; Williams 2015; Chychyla and Kogan 2015); in comparison with other sources of financial data (Boritz and No 2013); in assessing irregularities in accounting data (Henselmann et al. 2015); and in its ability to replicate prior research, that relied on private vendor databases (such as COMPUSTAT), (Baranes and Palas 2016; Williams 2015).

XBRL data has been shown to improve analyst forecast accuracy (Liu and O’Farrell 2013), and to provide a simple measure for identifying firms suspected of managing earnings (Henselmann et al. 2015)

Other studies found that XBRL is a useful tool not only for investors but for other financial decisions, such as loan decisions regarding loan size and interest rates (Kaya and Pronobis 2016).

.The aim of the SEC XBRL mandate is to decrease information asymmetry by improving the information processing capability of regulatory filings, however, early research has found inconsistencies (Boritz and No 2008), errors (Debreceny et al. 2010), or unnecessary extensions (Debreceny et al. 2011) in the XBRL filings. There were also found to be inconsistencies with other data aggregators (Boritz and No 2013; Chychyla and Kogan 2015). However, research findings are that the number of errors per filing is significantly decreasing as more quarters pass and when companies file more times (Du et al. 2011).

While this suggests that filers learn from their experience and therefore future filings will improve, a significant number of required accounting elements for financial statement analysis, are still expected to be missing from current XBRL filings.

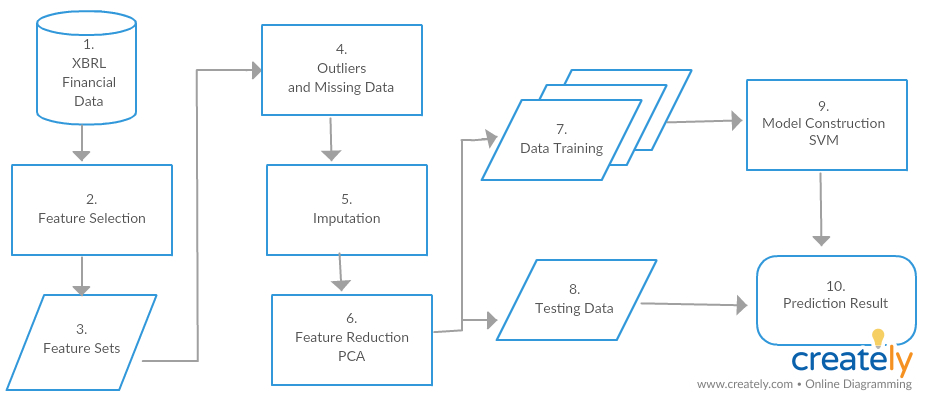
There is little research on using XBRL data with artificial intelligence, however the research seems to be promising. A financial crisis prediction model, based on SVM with XBRL financial data, was introduced by (Lin et al. 2008), their findings indicate that the suggested combination, of SVM technique and XBRL data provides a more accurate prediction ability than previous attempts. A structure for decision support system for credit risk valuation which implements advanced techniques such as SVM and XBRL was suggested by (Danenas and Garsva 2011), the structure was implemented with positive results (Danėnas 2013).

The research therefore suggest that using XBRL data with artificial intelligence techniques may be able to provide researchers and investors with a valid tool for decision making.

# MODEL, DATA AND RESUTLS

The objective of this study is to investigate the ability of financial information, in XBRL format, to predict future earnings using advanced artificial intelligence techniques, specifically SVM. A model was developed to code the financial ratios and streamline the data analysis for subsequent prediction, then develop a prediction model.

The model is presented in Figure 1.



The details of the model's stages are as follows:

1. **XBRL financial data** - Using the NASDAQ company list (<http://www.nasdaq.com/screening/company-list.aspx>) all 6,784 tickers listed on all of the three major US stock exchanges (AMEX, NASDAQ, and NYSE) were found.

The quarterly financial data was obtained using XBRL Analyst (created by FinDynamics); an Excel plugin that allows users to access the company’s XBRL tagged data from its XBRL SEC filing via the XBRL US database. Using this software not only allows for easy access and analysis of the data but also for the calculation of any missing balances. For example, the balance reported in each XBRL filing for total liabilities is not available on the original XBRL filing but is extracted and calculated on the XBRL Analyst. The data obtained was from quarterly filings from Q1/ 2012 to Q3/2017 (23 quarters).

1. **Feature selection** - the process of selecting a subset of relevant features to be used in the model construction, was used to create the financial ratios.

Of the 6,784 tickers 2,675 tickers were removed. The reasons for removal: there wasn't any data reported in XBRL format, tickers for non-common stocks, tickers for companies with IPO's between 2012 and 2017, and tickers for companies with more than one ticker (the same CIK).

The final sample included 4,109 companies (61% of all tickers listed) that were publicly traded on Q3/2017. These findings are compatible with previous studies where the final sample included 59.2% of listed companies (296) (Williams 2015), and 68.6% of listed companies (343) (Baranes and Palas 2016) of the total population of S&P 500 companies. Table 1 lists descriptive data for these companies.

**Table 1 - Descriptive Data for the Study Sample**

1. **Feature Sets** - In the attempt to duplicate the Ou and Penman (1989) study as closely as possible 58 variables were extracted from the XBRL filing data (Appendix 1). It should be noted that some of the variables had to be calculated from the original filing, whereas some variables were already calculated as part of the XBRL Analyst tool. This database contained 90,113 records.
2. **Outliers and missing data** - Additional records were removed in three stages. In the first stage, every company that had more than 35% of the variables missing (more thatn 20 variables) was removed, in the second stage, every variable which had more than 15% missing data points was eliminated.

Once these two stages were implemented a third stage, the removal of outliers was implemented. Removal of outliers is important because it can drastically bias/change the fit estimates and predictions. Although the Interquartile range (IQR) method (Barbato et al. 2011) is a popular method, it could not be used in this case since the data is asymmetrical. That is why the simple method of eliminating the top 2% and the bottom 2% was chosen, any data value beyond these limits was recognized as an outlier and eliminated.

Once all stages have been implemented 64,837 records remained, 72% of the original observations (90,113 records).

1. **Imputation** – previous research found that the use of XBRL data in financial analysis maybe incomplete because the data is not available (Williams 2015; Chychyla and Kogan 2015). An accounting element may not be extractable from an XBRL company filing due to several reasons, among them: the preparer erroneously did not tag the accounting element, the preparer used the wrong tag for an accounting element, or the SEC’s protocol for the preparation of XBRL company filings set forth in the EDGAR Filer Manual did not permit or require a tag.

To overcome this problem, of complex incomplete data, multiple imputation is the best method to be employed (Rubin 1996). There are several approaches for imputing multivariate data, Multivariate Imputation by Chained Equations (MICE) is considered to be a better alternative in cases where no suitable multivariate distribution can be found. MICE specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. Starting from an initial imputation, MICE draws imputations by iterating over the conditional densities.

For the purpose of this study the package of MICE in R was implemented, the package provides five iterations for implementation, all iterations were used in the current analysis, providing five different data sets.

Table 2 presents changes from the original data (data) to that of the data with imputation (full data). The number of observations increased by about 10%, however this small change allowed for the most important change, and that is the number of companies that are were utilized by the models, which increased by an average of 144%. This means that more than twice as many companies may be examined by the models and used in the classification for prediction purposes.

**Table 2 – Changes in Data due to Imputation**

1. **Feature Reduction** – because not all of the chosen ratios are informative and can provide high discrimination power feature reduction can be used to filter out redundant and/or irrelevant features resulting in more representative features for better prediction performance (Tsai 2009). Feature reduction, as the preprocessing step, is one of the most important steps in data mining process (Tsai and Hsiao 2010) and is the first most important step in developing an SVM forecaster (Cao et al. 2006). In the model all available variables can be used as inputs of SVM, but irrelevant features or correlated features could deteriorate the generalization performance of SVM. There are many methods of feature reduction, among them: stepwise regression, factor analysis, genetic algorithms, decision trees and autoencoding. For the current study the feature reduction method of principal component analysis (PCA) was chosen. PCA has demonstrated the ability to improve the generalization performance of SVM (Cao et al. 2006). In models using financial information, for example bankruptcy predictions, applying PCA increased the performance of prediction (Tsai 2009).

PCA is a multivariate statistical technique. It aims at reducing the dimensionality of a database with a large number of interrelated variables. In particular, it extracts a small set of factors or components that are constituted of highly correlated elements, while retaining their original characters. After performing PCA, the uncorrelated variables which are called components, will replace the original variables. The total variability of a dataset produced by the complete set of m variables can often be accounted for primarily by a smaller set of k components of these variables (kbm). Therefore, the new dataset consists of n records on k components rather than n records on m variables as the original one. Specifically, eigenvalues and eigenvectors of the principal components are computed in order to find a linear combination of the original variables that makes the greatest variance. The first principal component accounts for as much of the variability in the data, and the second principal component accounts for the remaining variability and so on. Particularly, the level of the variability for each feature lies in the range [0,1], in which the feature with 1 represents the highest variability. Therefore, if we need the components (i.e. features) which can explain 90% (i.e. 0.9) of the variability, features with 90% of the variability or higher can be selected (Jolliffe 2002).

In the current study the variables (vectors) are the 58 financial ratios of all the companies remaining in the database for all quarters. Each variable is examined based on its ability to provide information regarding the total variability of the dataset. The first variable (the first principal component) provides the most information, the next variable is analyzed using the remaining variables and provides a lower level of variability, the remaining variables provide decreasing information levels. The eigenvalues represent a measure of the variance explained by the principal component. It is common practice to use the Kaiser criterion, discard all variables with an eigenvalue smaller than 1 (Costello and Osborne 2005).

Each principal component is a linear combination of the complete dataset, however each variable in the original set has a different coefficient which rates its weight in calculating the principal component, variables with a coefficient lower than 0.3 were discarded. These discarded variables are used in the computerized analysis, however they are eliminated in examination of the principal components.

The remaining principal components are the explanatory variables. Table 3 presents those principal components that were common for all of the five datasets (provided in the imputation stage), for each industry.

**Table 3 – Principal Components**

As can be seen from table 3, the variables with the most explanatory value for most of the industries (7 out of 8), Change in Total Assets, Change in Total Revenue, and Long-Term Debt to Equity, represent the change in the company's ability to create revenues, change in the company's ability to create profits and change in the company's risk level (solvency), respectively. Of the next explanatory value (6 out of 8 industries), all the variables represent profitability except for one (current ratio which represents liquidity).

In total 49 variables were found to be explanatory variables for all the industries, for each industry between 15 and 27 variables were used to created the model, on average 21.5 explanatory variables for each industry.

1. **Data training** – the purpose of data training is to label the variables to be implemented in the model, in this case to label the dependent and independent variables. The independent variables are the financial ratios for year Xi the dependent variable is the change (+/-) in earnings for the year Xi+1, above the drift. The drift term was estimated as the mean earnings per share change over the four prior quarters to the estimated quarter (Ou and Penman 1989).
2. **Testing data** – once the model is created, using the data training set, it is used to examine its accuracy on the testing data set. The independent variable of the testing data is Q3/2017.
3. **Model construction** – the purpose of the research is to construct a model which will predict the change in earnings one period ahead, after removal of the earnings drift. The model will present the change in earnings in the form increase (+) or decrease (-).

There are many statistical techniques that may be used to create the model. SVM is a popular tool in time series forecasting, due to its generalization performance ability, the absence of local minima and the sparse representation of solution (Cao et al. 2006). Unlike most of the traditional methods which implement the Empirical Risk Minimization Principal, SVM seeks to minimize an upper bound of the generalization error rather than minimize the training error (Vapnik 2013). It is a supervised artificial intelligence algorithm which is mostly used in classification problems. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, a classification is performed by finding the hyper-plane that differentiate the two classes very well. Support Vectors are simply the co-ordinates of individual observation.

Kernel methods are a class of algorithms for [pattern analysis](https://en.wikipedia.org/wiki/Pattern_analysis). The linear function and the radial basis function (RBF) are the two popular kernel functions suggested for SVM classifiers. As opposed to the linear kernel, the RBF kernel nonlinearly maps the samples into the high-dimensional space, which makes it feasible for nonlinear problems. One of the challenging problems using RBF kernel to build SVM model is the selection of parameter values for (C, σ) in order to ensure satisfactory prediction performance. The RBF model may yield poor performance if these parameters are not carefully chosen (Li et al. 2015).

The current research implements the SVM model in R version 3.4.1 using library e1071 with RBF kernel.

1. **Prediction Results** - the analysis creates 5 models for the 5 data sets (from the 5 different versions of imputation). Each model predicts for each company whether there will be an increase in profitability in the next quarter or a decrease. The prediction is than compared to the actual performance of the company in Q3/2017. For each company if 4 or more of the data sets have the same prediction, that is the prediction that is chosen. Those companies where only 3, or less, of the datasets have the same predictions are ignored in the accuracy test (Ou and Penman, 1989, used a probability of <40% and >60% to eliminate uncertain predictions).

The results of the model are presented in Table 4.

**Table 4 – Accuracy of Prediction of SVM model**.

As can be seen from Table 4 the models were able to provide an accuracy of prediction between 57.1% and 77.1% with an average of 66.9%. These results are compatible with models based on multiple regression (Ou and Penman 1989; Holthausen and Larcker 1992) and data from COMPUSTAT, that ranged around 66% accuracy.

# CONCLUSIONS

The focus of this study has been to implement artificial intelligence techniques, specifically SVM on XBRL data and create a model to predict earnings movement.

The findings of the study suggest that artificial intelligence can be used to predict one quarter ahead earnings based on XBRL data. The results of the model, which had an accuracy rate of up to 77%, depending on the industry, are compatible with previous research, which was based on traditional statistical methods (multiple regression) models and COMPUSTAT data found the same results (Ou and Penman 1989; Holthausen and Larcker 1992).

The current model, based on SVM and XBRL data has several advantages over the traditional statistical models based on COMPUSTAT: XBRL data provides more timely information and therefore the model may be updated every quarter as information is published, SVM has been shown to be a more effective when non-linear data is used, such as financial information.

This study contributes to previous research by introducing artificial intelligence techniques which may be used in conjunction with XBRL company filings. The attempt of the study is to examine the validity of using artificial intelligence in financial prediction models.

The main limitation of this study is the relatively short time period data (from 2012) of the SEC XBRL mandate. The short time period not only limits the amount of data available but may also cause other problems such as inconsistencies, errors, or unnecessary extensions in the XBRL filings (Debreceny et al. 2011; Du et al. 2013). However, given that there are indications that XBRL quality increases over time (Du et al. 2013), the methodology may be tested again in the future.

There are several possible extensions of this study among them examining other artificial intelligence techniques for both feature reduction and model construction. The passage of time, which will allow higher quality filings, will also enhance the use of the XBRL data.

1. BIBLIOGRAPHY

Abarbanell, J., and B. J. Bushee. 1997. Fundamental analysis, future EPS, and stock prices. *Journal of Accounting Research* 35 (1): 1–24.

Alam, P., and C. a. Brown. 2006. Disaggregated earnings and the prediction of ROE and stock prices: a case of the banking industry. *Review of Accounting and Finance* 5 (4): 443–463.

Ball, R., and L. Shivakumar. 2008. How much new information is there in earnings? *Journal of Accounting Research* 46 (5): 975–1016.

Barak, S., and M. Modarres. 2015. Developing an approach to evaluate stocks by forecasting effective features with data mining methods. *Expert Systems with Applications* 42 (3): 1325–1339.

Baranes, A., and R. Palas. 2016. The Prediction of Earnings Movements Using Accounting Data: Using XBRL. *International Journal of Accounting Research* 4 (2): 1–7.

Barbato, G., E. M. Barini, G. Genta, and R. Levi. 2011. Features and performance of some outlier detection methods. *Journal of Applied Statistics* 38 (10): 2133–2149.

Bernard, V., J. Thomas, and J. Wahlen. 1997. Accounting-Based Stock Price Anomalies: Separating Market Inefficiencies from Risk. *Contemporary Accounting Research* 14 (2): 89–136.

Bird, R., R. Gerlach, and A. D. Hall. 2001. “The prediction of earnings movements using accounting data: An update and extension of Ou and Penman” -- a response. *Journal of Asset Management* 2 (2): 180–195.

Bloomfield, R. J., R. Libby, and M. W. Nelson. 2003. Do Investors Overrely on Old Elements of the Earnings Time Series? *Contemporary Accounting Research* 20 (1): 1–31.

Boritz, J. E., and W. G. No. 2008. SEC’s XBRL voluntary program on EDGAR: A case for quality assurance. *Current* 2 (2): A36–A50.

———. 2013. The Quality of Interactive Data: XBRL versus Compustat, Yahoo Finance, and Google Finance: 0–51.

Cao, L. J., Z. Jingqing, C. Zongwu, and L. K. Guan. 2006. an Empirical Study of Dimensionality Reduction in Support Vector Machine ∗. *Neural Network World*.

Cao, Q., and Q. Gan. 2010. Assessing model efficacy in forecasting EPS of Chinese firms using fundamental accounting variables: a comparative study. *International Journal of Society Systems Science* 2 (3): 207–225.

Chandwani, D., and M. S. Saluja. 2014. Stock Direction Forecasting Techniques : An Empirical Study Combining Machine Learning System with Market Indicators in the Indian Context. *International Journal of Computer Applications* 92 (11): 8–17.

Chychyla, R., and A. Kogan. 2015. Using XBRL to Conduct a Large-Scale Study of Discrepancies between the Accounting Numbers in Compustat and SEC 10-K Filings. *Journal of Information Systems* 29 (1): 37–72.

Costello, A. B., and J. W. Osborne. 2005. Best Practices in Exploratory Factor Analysis : Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Education* 10: 1–9.

D’Souza, J. M., K. Ramesh, and M. Shen. 2010. The interdependence between institutional ownership and information dissemination by data aggregators. *Accounting Review*.

Danėnas, P. 2013. Support Vector Machines Based Classifiers in Intelligent Decision Support System for Credit Risk Evaluation. VILNIUS UNIVERSITY.

Danenas, P., and G. Garsva. 2011. SVM AND XBRL BASED DECISION SUPPORT SYSTEM FOR CREDIT RISK EVALUATION Paulius Danenas , Gintautas Garsva. *17th International Conference on Information and Software Technologies (IT 2011), Technologija, Kaunas, Lithuania* (January 2011): 190–198.

Debreceny, R., S. Farewell, M. Piechocki, C. Felden, and A. Gräning. 2010. Does it add up? Early evidence on the data quality of XBRL filings to the SEC. *Journal of Accounting and Public Policy* 29 (3): 296–306.

Debreceny, R. S., S. M. Farewell, M. Piechocki, C. Felden, A. Gräning, and A. D’Eri. 2011. Flex or break? Extensions in XBRL disclosures to the SEC. *Accounting Horizons* 25 (4): 631–657.

Du, H., M. a. Vasarhelyi, and X. Zheng. 2013. XBRL Mandate: Thousands of Filing Errors and So What? *Journal of Information Systems* 27 (1): 61–78.

Du, H., M. A. Vasarhelyi, and X. Zheng. 2011. XBRL Mandate: Thousands of Filing Errors and So What?

Etemadi, H., A. Ahmadpour, and S. M. Moshashaei. 2015. Earnings Per Share Forecast Using Extracted Rules from Trained Neural Network by Genetic Algorithm. *Computational Economics* 46 (1): 55–63.

Finger, C. a. 1994. The ability of earnings to predict future earnigns and cash flow. *The Journal of Accounting Research* 32 (2): 210–223.

Han, S., and R.-C. Chen. 2007. Using SVM with Financial Statement Analysis for Prediction of Stocks. *communications of the IIMA* 7 (4): 63–72.

Henselmann, K., D. Ditter, and E. Scherr. 2015. Irregularities in Accounting Numbers and Earnings Management – A Novel Approach Based on SEC XBRL Filings. *Journal of Emerging Technologies in Accounting* 12 (1): 117–151.

Holthausen, R. W., and D. F. Larcker. 1992. The prediction of stock returns using financial statement information. *Journal of Accounting and Economics* 15 (2–3): 373–411.

Huang, C.-F. 2012. A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing* 12 (2): 807–818.

Jolliffe, I. T. 2002. *Principal Component Analysis*. 2nd editio. New York: Springer.

Kamley, S., S. Jaloree, and R. S. Thakur. 2016. Performance forecasting of share market using machine learning techniques: A review. *International Journal of Electrical and Computer Engineering* 6 (6): 3196–3204.

Kaya, D., and P. Pronobis. 2016. The benefits of structured data across the information supply chain: Initial evidence on XBRL adoption and loan contracting of private firms. *Journal of Accounting and Public Policy* 35 (4).

Kinney, M. R., and E. P. Swanson. 1993. The Accuracy and Adequacy of Tax Data in COMPUSTAT. *Journal of the American Taxation Association* 15 (1): 121.

Li, C. K., D. Liang, F. Lin, and K. L. Chen. 2015. The application of corporate governance indicators with xbrl technology to financial crisis prediction. *Emerging Markets Finance and Trade* 51 (0): S58–S72.

Lin, F., D. Liang, and S.-J. Chiu. 2008. The Study of a Financial Crisis Prediction Model based on XBRL. In *2008 Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, 147–153.

Liu, C., and G. O’Farrell. 2013. The role of accounting values in the relation between XBRL and forecast accuracy. *International Journal of Accounting and Information Management* 21 (4): 297–313.

Miguel, J. G. S. 1977. The Reliability of R&D Data in COMPUSTAT and 10-K Reports. *The Accounting Review* 52 (3): 638–641.

Ou, J. a. 1990. The Information Content of Nonearnings Accounting Numbers as Earnings Predictors. *Journal of Accounting Research* 28 (1): 144–163.

Ou, J. A., and S. H. Penman. 1989. Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics* 11 (4): 295–329.

Penman, S. H., and X. J. Zhang. 2002. Accounting conservatism, the quality of earnings, and stock returns. *Accounting Review* 77 (2): 237–264.

Qiu, X. Y., P. Srinivasaan, and Y. Hu. 2014. Supervised Learning Models to Predict Firm Performance With Annual Reports: An Empirical Study. *Journal of the Association for Information Science and Technology* 65 (2): 400–413.

Rajakumar, M. P., and J. Ramya. 2017. A Comparison of Intelligent Soft Computing Techniques for Forecasting Earnings Per Share. *international journal of pure and applied mathematics* 114 (9): 167–177.

Raposo, R. D. C. T., and A. J. D. O. Cruz. 2002. Stock Market Prediction Based on Fundamentalist Analysis with Fuzzy- Neural Networks 2 Input Data and Sector Choice 3 The Choice of Economic Indicators. In *Proc. of the 3rd WSEAS Int. Conf. on Neural Networks and Applications (NNA’02)*.

Rosenberg, B., and M. Houglet. 1974. ERROR RATES IN CRSP AND COMPUSTAT DATA BASES AND THEIR IMPLICATIONS. *The Journal of Finance* 29 (4): 1303–1310.

Roychowdhury, S., and E. Sletten. 2012. Voluntary disclosure incentives and earnings informativeness. *Accounting Review* 87 (5): 1679–1708.

Rubin, D. B. 1996. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*.

Sap, M., and A. M. Awan. 2005. Stock Market Prediction using Support Vector Machines 17(2), 27-35. *Jurnal Teknologi Maklumat* 17 (2): 27–35.

Setiono, B., and N. Strong. 1998. Predicting stock returns using financial statement information. *Journal of Business Finance and Accounting* 25 (5–6): 631–657.

Shroff, P. K. 1999. The variability of earnings and non-earnings information and earnings prediction. *Journal of Business Finance and Accounting* 26 (7–8): 863–882.

Stober, T. L. 1992. Summary financial statement measures and analysts’ forecasts of earnings. *Journal of Accounting and Economics* 15 (2–3): 347–372.

Tallapally, P., M. S. Luehlfing, and M. Motha. 2011. The Partnership Of EDGAR Online And XBRL - Should Compustat Care? *The Review of Business Information Systems* 15: 39–46.

Tsai, C. F. 2009. Feature selection in bankruptcy prediction. *Knowledge-Based Systems* 22 (2): 120–127.

Tsai, C. F., and Y. C. Hsiao. 2010. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems* 50 (1): 258–269.

Vapnik, V. 2013. *The nature of statistical learning theory*. Springer science & business media.

Williams, K. L. 2015. *The Prediction of Future Earnings Using Financial Statement Information: Are XBRL Company Filings up to the Task?*

Wu, W., and J. Xu. 2006. Fundamental analysis of stock price by artificial neural networks model based on rough set theory. *World Journal of Modeling an Simulation* 2 (1): 36–44.

Yang, D. C., M. a. Vasarhelyi, and C. Liu. 2003. A note on the using of accounting databases. *Industrial Management & Data Systems* 103 (3): 204–210.

Zhang, W., W. Zhang, Q. Cao, Q. Cao, M. J. Schniederjans, and M. J. Schniederjans. 2004. Neural network earnings per share forecasting models: A comparative analysis of alternative methods. *Decision Sciences* 35 (2): 205–237.

**Table 1 - Descriptive Data for the Study Sample**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | N | Frequency | Percent |
| Stock Exchange | AMEX | 4,109 | 201 | 4.88% |
| NASDAQ | 4,109 | 2254 | 54.85% |
| NYSE | 4,109 | 1654 | 40.27% |
| Size (Revenues) | < $10,000,000 | 4.109 | 317 | 7.7% |
| $10,000,000- $100,000,000 | 4.109 | 799 | 19.4% |
| $100,000,000-$500,000,000 | 4.109 | 979 | 23.8% |
| $500,000,000-$1,000,000,000 | 4.109 | 500 | 12.2% |
| $1,000,000,000-$10,000,000,000 | 4.109 | 1188 | 28.9% |
| $10,000,000,000-$100,000,000,000 | 4.109 | 318 | 7.7% |
| >$100,000,000,000 | 4.109 | 7 | 0.2% |
| Industry (SIC Code) | Agriculture, Forestry and Fishing (01-09) | 4.109 | 13 | 0.3% |
| Mining (10-14) | 4.109 | 192 | 4.7% |
| Construction (15-17) | 4.109 | 51 | 1.2% |
| Manufacturing (20-39) | 4.109 | 1597 | 38.9% |
| Transportation, Communications, Electric, Gas and Sanitary Services (40-49) | 4.109 | 343 | 8.3% |
| Wholesale Trade (50-51) | 4.109 | 109 | 2.7% |
| Retail Trade (52-59) | 4.109 | 222 | 5.4% |
| Real Estate (60-67) | 4.109 | 958 | 23.3% |
| Services (70-89) | 4.109 | 624 | 15.2% |
| Public Administration (91-99) | 4.109 | 0 | 0.0% |

**Table 2 – Changes in Data due to Imputation**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Total Sample | Mining (10-14) | Construction (15-17) | Manufacturing (20-39) | Transportation, Communications, Electric, Gas and Sanitary Services (40-49) | Wholesale Trade  (50-51) | Retail Trade (52-59) | Finance, Insurance and Real Estate (60-67) | Services (70-89) |
| # companies before imputation |  |  |  |  |  |  |  |  |  |
| # companies after imputation |  |  |  |  |  |  |  |  |  |
| % Increase |  |  |  |  |  |  |  |  |  |
| # variables before imputation |  |  |  |  |  |  |  |  |  |
| # variables after imputation |  |  |  |  |  |  |  |  |  |
| % Increase |  |  |  |  |  |  |  |  |  |
| # observations before imputation |  |  |  |  |  |  |  |  |  |
| # observations after imputation |  |  |  |  |  |  |  |  |  |
| % Increase |  |  |  |  |  |  |  |  |  |

**Table 3 – Principal Components**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Explanatory Variables | # Industries the PCA appears | Mining (10-14) | Construction (15-17) | Manufacturing (20-39) | Transportation, Communications, Electric, Gas and Sanitary Services (40-49) | Wholesale Trade  (50-51) | Retail Trade (52-59) | Finance, Insurance and Real Estate (60-67) | Services (70-89) |
| ΔTotal Assets | 7 | X | X | X | X | X | - | X | X |
| ΔTotal Revenue | 7 | X | X | X | X | X | - | X | X |
| Long-Term Debt/Equity | 7 | - | X | X | X | X | X | X | X |
| ΔCapital Expenditures/total assets | 6 | X | X | - | X | X | X | X | - |
| ΔSales/Total Assest | 6 | X | X | X | X | - | - | X | X |
| Current Ratio | 6 | X | - | X | X | X | X | - | X |
| ROA | 6 | X | X | X | X | - | - | X | X |
| ROE | 6 | X | X | X | X | - | - | X | X |
| Pre taxes income/Sales | 6 | X | X | X | X | - | - | X | X |
| Net Profit Margin | 6 | X | - | X | X | X | X | - | X |
| Operating Income to Total assets | 5 | X | - | X | X | - | X | X | - |
| EBITDA Margin Ratio | 5 | X | X | - | X | X | - | X | - |
| ΔCurrent Ratio | 5 | X | X | X | - | - | - | X | X |
| ΔWorking capital | 5 | X | - | - | X | X | X | X | - |
| ΔLong-Term Debt/Equity | 5 | X | - | X | X | X | - | - | X |
| ΔWorking capital to total assets | 4 | X | X | - | - | - | - | X | X |
| Account Receivable Turnover | 4 | X | - | X | X | - | - | - | X |
| Quick Ratio | 4 | X | - | - | X | X | - | - | X |
| Sales/Total Assest | 4 | X | - | X | - | - | X | - | X |
| Sales to total working capital | 4 | - | - | X | X | - | X | - | X |
| Sales to Fixed assets | 4 | - | - | X | - | X | X | - | X |
| Working capital to total assets | 4 | X | - | - | X | - | X | - | X |
| Net Income over OCF | 4 | X | - | X | X | - | - | - | X |
| ΔTotal Long-Term Debt | 4 | X | - | X | X | - | - | - | X |
| ΔDepreciation over Plant | 4 | X | X | - | - | X | X | - | - |
| ΔSales to total working capital | 4 | X | X | - | - | - | - | X | X |
| Days sales in Accounting Recv. | 3 | X | - | - | X | - | - | - | X |
| Depreciation over Plant | 3 | X | - | - | - | X | - | - | X |
| Equity/Fixed assets | 3 | - | - | X | X | - | X | - | - |
| Sales to total cash | 3 | - | X | - | X | - | - | X | - |
| ΔEquity/Fixed assets | 3 | - | - | X | - | X | X | - | - |
| ΔPre taxes income/Sales | 3 | X | X | - | - | - | - | - | X |
| ΔNet Profit Margin | 3 | X | X | - | - | - | - | X | - |
| ΔOperating Income to Total assets | 3 | - | X | - | - | - | - | X | X |
| Gross Profit Margin | 2 | - | - | X | - | - | - | - | X |
| ΔQuick Ratio | 2 | - | X | - | - | X | - | - | - |
| ΔROE | 2 | - | - | - | - | X | X | - | - |
| ΔDays sales in Accounting Recv. | 2 | - | X | - | - | - | - | X | - |
| Inventory to total assets | 1 | - | - | X | - | - | - | - | - |
| Δinventory | 1 | - | - | - | - | X | - | - | - |
| ΔInventory to total assets | 1 | - | - | - | - | - | X | - | - |
| ΔEBITDA Margin Ratio | 1 | - | - | - | - | X | - | - | - |
| Inventory Turnover | 1 | - | - | - | X | - | - | - | - |
| Sales to total Inventory | 1 | - | - | - | - | X | - | - | - |
| Cash From Operations (CFO) to Total Debt | 1 | - | - | - | - | X | - | - | - |
| ΔInventory Turnover | 1 | - | - | - | - | X | - | - | - |
|  |  | 27 | 19 | 22 | 24 | 21 | 15 | 18 | 26 |

**Table 4 – Accuracy Results**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Average | Mining (10-14) | Construction (15-17) | Manufacturing (20-39) | Transportation, Communications, Electric, Gas and Sanitary Services (40-49) | Wholesale Trade  (50-51) | Retail Trade (52-59) | Finance, Insurance and Real Estate (60-67) | Services (70-89) |
| Model Accuracy | 66.91% | 70.87% | 77.14% | 57.90% | 75.30% | 57.14% | 75.00% | 60.53% | 61.37% |

**Appendix 1 – Variables**

|  |  |
| --- | --- |
|  | **Variables** |
| 1 | Account Receivable Turnover |
| 2 | Current Ratio |
| 3 | Quick Ratio |
| 4 | Inventory Turnover |
| 5 | Total Debt To Equity |
| 6 | ROA |
| 7 | ROE |
| 8 | Gross Profit Margin |
| 9 | Days sales in Accounting Recv. |
| 10 | Inventory to total assets |
| 11 | Depreciation over Plant |
| 12 | Long-Term Debt/Equity |
| 13 | Equity/Fixed assets |
| 14 | Times Interest Earned |
| 15 | Sales/Total Assest |
| 16 | Pre taxes income/Sales |
| 17 | Net Profit Margin |
| 18 | Sales to total cash |
| 19 | Sales to total Inventory |
| 20 | Sales to total working capital |
| 21 | Sales to Fixed assets |
| 22 | Working capital to total assets |
| 23 | Operating Income to Total assets |
| 24 | EBITDA Margin Ratio |
| 25 | Cash From Operations (CFO) to Total Debt |
| 26 | Payment Of Dividends as % of OCF |
| 27 | Net Income over OCF |
| 28 | ΔDepreciation (&Amortization), IS |
| 29 | Δinventory |
| 30 | ΔResearch & Development Expense |
| 31 | ΔTotal Assets |
| 32 | ΔTotal Long-Term Debt |
| 33 | ΔTotal Revenue |
| 34 | ΔCurrent Ratio |
| 35 | ΔQuick Ratio |
| 36 | ΔInventory Turnover |
| 37 | ΔDividends per share |
| 38 | ΔTotal Debt To Equity |
| 39 | ΔROE |
| 40 | ΔGross Profit Margin |
| 41 | ΔWorking capital |
| 42 | ΔDays sales in Accounting Recv. |
| 43 | ΔInventory to total assets |
| 44 | ΔDepreciation over Plant |
| 45 | ΔCapital Expenditures/total assets |
| 46 | ΔLong-Term Debt/Equity |
| 47 | ΔEquity/Fixed assets |
| 48 | ΔTimes Interest Earned |
| 49 | ΔSales/Total Assest |
| 50 | ΔPre taxes income/Sales |
| 51 | ΔNet Profit Margin |
| 52 | ΔSales to total Inventory |
| 53 | ΔSales to total working capital |
| 54 | ΔResearch & Development Expense to Sales |
| 55 | ΔWorking capital to total assets |
| 56 | ΔOperating Income to Total assets |
| 57 | ΔEBITDA Margin Ratio |
| 58 | LΔCapital Expenditures/total assets |