ANALYSIS OF FINANCIAL DATA USING DATA MINING TECHNIQUES

Sai Poojitha Battula   
*Computer Science*  
*California State University Northridge*Northridge, California  
sai-poojitha.battula.072@my.csun.eduGuruVishnu Bojanapureddy  
*Computer Science*  
*California State University Northridge*Northridge, California guruvishnu.bojanapureddy.642@my.csun.edu

Surya Teja Swamy  
*Computer Engineering*  
*California State University Northridge*Northridge, California  
surya-teja.swamy.388@my.csun.eduBhargavi Kothapeta  
*Computer Science*  
*California State University Northridge*Northridge, California  
bhargavi.kothapeta.967@my.csun.edu

*Abstract*— The main idea of this paper is to analyze the vast data sets and to provide insights on different financial investments, the dataset we are considering includes the different parameters related to financial investments for instance stocks, mutual funds, or bitcoins. The objective here is to provide insights, analyze the patterns, improve risk management, and predict future trends, also to improve decision-making. The scope of the paper is to use traditional data mining steps such as data pre-processing and data mining techniques such as regression analysis, ensemble learning, or neural networks. By training the financial data with different models, we can improve the accuracy of the existing system.

Keywords—Data mining, data pre-processing, neural networks, ensemble learning, regression analysis, prediction.

# Introduction

Data is a collection of distinct, objective facts about a process or an event that, by themselves, are not very useful and must be transformed into information. From straightforward numerical measures and text documents to more intricate information like location data, multimedia channels, and hypertext texts, we have been gathering a wide range of data.

Data mining is a trending and powerful field for performing advanced data analysis and it incorporates methods and tools from statistics and machine learning. Applications for advanced analytics and business intelligence make use of the information produced by this process, which comprises the examination of validated data.

The data mining techniques can be applied over a range of different sectors in real-time such as research, education, transportation, healthcare, finance and banking, business transactions, and much more.

To determine whether a business is steady and profitable enough to get capital investment, financial analysis of data is crucial. The balance sheet, cash flow statement, and income statement are where financial analysis concentrates their analysis.

Data mining methods have been applied to financial markets to uncover hidden patterns and forecast upcoming trends and behaviors. The mining of such data, especially high-frequency financial data, often calls for advanced statistical, mathematical, and artificial intelligence techniques.

# Related work

## Data Mining Models

There are a lot of data mining techniques/models that can be used to analyze and extract the data. So, it is necessary to understand each method and the efficiency of different techniques so that we can decide on the correct data mining method that can produce effective results with consistency.

In general, there are various data mining methods in use and the most popular ones are logistic regression, SVM, and Neural Networks [10]. There are a lot of research studies on these methods, there is a study conducted by Ken Ishibashi and his co-researchers [1], they did a comparison of different data mining models to select a model that will be good for financial data. Their study validated those techniques such as SVM and NN produced higher accuracy than the existing models. There is another review study made by Zhang. C [5], he along with his co-researchers studied 75 articles for a concept called decision fusion which is to fuse the outcomes of different resources.

Other studies that focus on NN techniques, N. Srinivasan, and C. Lakshmi [2], mentioned in their study that intermittent training of the data results in high accuracy in the prediction of stock values for the banking sector. There are different number of studies using neural networks are been presented in the reference articles [7][8][9] that are used not only to predict data trends but also regarding data fraud.

There is a study made by Bedour M. Alshammari and his co-researchers [11], on stock market prediction where they presented the data mining steps (from collection to prediction) clearly and reviewed and compared various techniques on stock data. Other studies that discussed the financial insights and recommendations [12][13] which are very near to our research.

# Methodology

Modern algorithms and approaches would likely be used by the machine learning (ML) system proposed for this project to analyze financial data and reach informed judgments. The proposed system would be created to improve the accuracy, dependability, and understandability of the results while addressing the system's flaws.

There are commonly two different learning methods used for ML modeling: supervised learning and unsupervised learning. Unsupervised learning applies to the model on raw data, which is frequently used to discover patterns or to cluster the data. Supervised learning applies to the labeling of the data, which is the main distinction between supervised and unsupervised learning. To categorize fresh data using these discovered patterns or to forecast outcomes, supervised machine learning will learn the relationship between input and output through labeled training data.

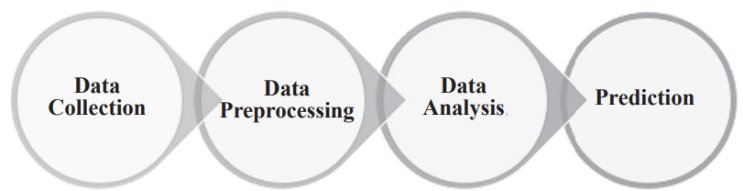


Fig. 1. Different stages of the system

## Data Pre-Processing

In any data mining project, the flow will start with data collection, we have collected the financial data which is used for stock value prediction. In the pre-processing stage, the collected data is processed and analyzed to verify whether the data is biased or not, duplication of data, null values, and outliers.

The dataset which we considered is of the size 851264 x 7, which includes different parameters related to the stocks, such as the name of the stock(company), high and low values of each stock, and open and closing values for the date. The data consists of various data types which include double, long, and object. From the pre-processing steps, we verified that the data is pure and doesn’t consist of any duplicated values and the outliers are removed and the null values are adjusted using statistical methods.

## Feature Selection

The dataset has different parameters that are mentioned in the earlier section, to train the models we require to select the feature or multiple features from the data.

Feature selection is an important step that will help the models to get better accuracy when proper features are used.

To estimate the proper features there are some techniques such as Principal Component Analysis (PCA) or Gene, but the overall features in our dataset are fewer in number we used the correlation technique to check the better features to select.

## Model Training

After the pre-processing of the data is completed, one of the main features will be selected for the prediction purpose. As we are using various NN and other data mining techniques we need to train all the models with the original data set, so we used the jupyter notebook powered by anaconda navigator as the software platform for this project, as the data is reasonable size there is not needed for any hardware assistance or any server assistance.

The presented approach would use prepared financial data to train ML models. Cross-validation methods may be used in this situation to enhance model performance and prevent overfitting. The financial data is divided into training and testing sets in a ratio of 80:20.

In this project we trained NN models such as Artificial Neural Networks (ANN) after going through various articles mentioned in the related work, we preferred to keep the number of layers to 6 [1][10] for the efficiency of the model.

Along with ANN, we also used some regression techniques, boosted tree classifier models, or ensemble models such as the extreme gradient boosting (XGBoost) classifier. Apart from the ensemble methods we used another popular modeling, which is linear regression or logistic regression for data prediction.

## Model Evaluation

After all the models are trained using the training set, the models are evaluated using the testing set, in the model evaluation we are analyzing the accuracy of each model and evaluation report which consists of the parameters such as precision, f1-score, and other related parameters which are useful to compare the models and analyze the hyperparameters used in the training of the models.

The accuracy of the models is analyzed for different values of the hyperparameters that are used in each model. The hyper-parameters values are finalized on the best accuracy of the model and some of the hyperparameters such as the number of clusters for k-means and the number of layers for ANN are finalized using a literature survey.

# Results

This project uses different models to predict the stock value based on the different values of a particular date, before training the model using the stock data, the data is cleaned up using pre-processing techniques and checked for null values and duplication of values. Here are some results from the pre-processing stage such as identified outliers, and correlation results.

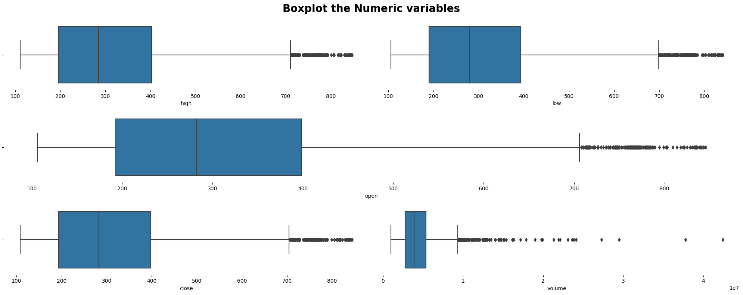


Fig. 2. Box Plot to identify the outliers

As it seems to have outliers in the data, but they aren’t considered as outliers, considering the stock data there might be some extreme values during the peak season.

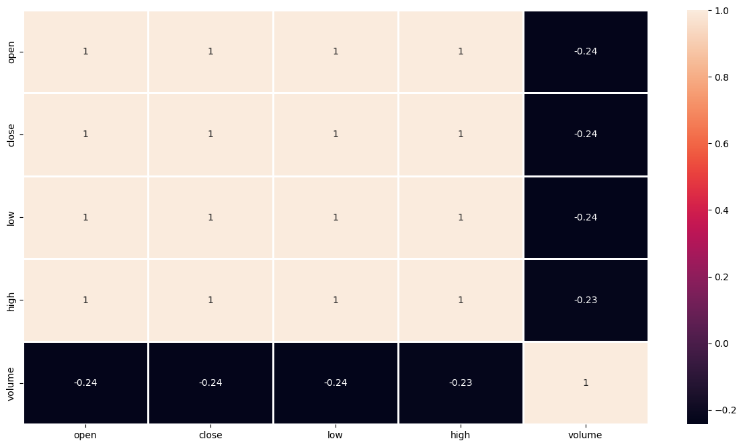


Fig. 3. Heatmap to display the correlation of the data

The evaluation of the linear regression model is shown in the plot below. The test data set is prepared for the model's evaluation, and the projected values are plotted against it. This model had a Mean Squared Error (MSE) and Mean Absolute Error (MAE) of 83, which is almost 99% accurate.

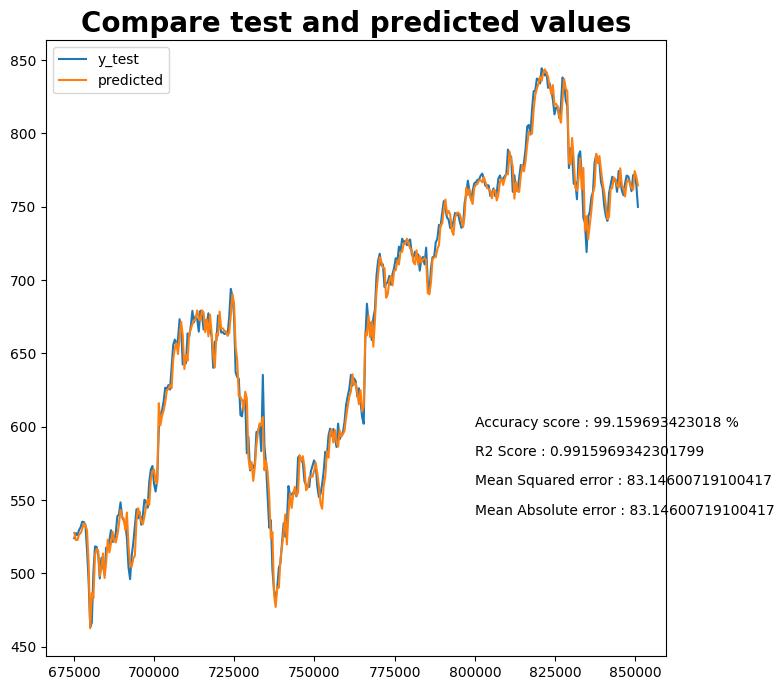


Fig. 4. Evaluation of Linear Regression Model

Other models are still in the testing stage such as ANN and other boosted tree classifiers. Currently, the methods aren’t providing better accuracy as expected. The comparison of prediction values against the testing data for ANN is provided below.

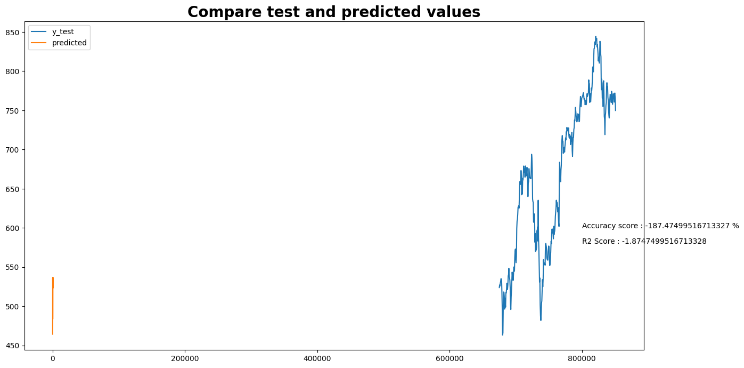


Fig. 5. Evaluation of XG boost model

As it is seen that the predicted values aren’t better than expected from the accuracy value it is clear that the models such as XG boost aren’t preferable even though they are better at classification of the data.

The ANN has the minimum loss compared to the linear regression. As the neural networks divide the data into batches for each batch of data the prediction loss is getting reduced, which helps in getting less MSE for overall data.

# Discussion and Limitations

The current data consists of open, close, high, and low stock values for each stock, it will help understand and predict of closing value for a particular day before the market gets closed. But to estimate a high or low or closing value it might require special data which includes the different factors which might influence the stock price of that particular company.

Our project is better than the current system which is available in the market, still, there are a few limitations that can be fixed to better the system. As we are dealing with larger datasets, we cannot predict the closing value of all company stocks at a time. If we are interested in a specific company, we can check for that stock value. Hence for further improvement, we are striving to modify the project in such a way it shall increase the accuracy at a time able to analyze and give better predictions.

# Conclusion

From this project, we are analyzing the stock data and can predict the closing value of the various company stocks. As it turns out that the stock dataset is a continuous and linearly dependent data the linear regression model can predict the values with higher accuracies even with outliers in the dataset. But by using the ANN model we can decrease the error. But the classification model which we used the XG Boost model isn’t predicting with good accuracy so we would like to conclude that the ensemble learning isn’t good for the prediction of stock value but regression and neural networks can be preferred as they are getting less error/high accuracy.

##### References

1. Ken Ishibashi, Takuya Iwasaki, Shota Otomasa and Katsutoshi Yada, “Model Selection for Financial Statement Analysis”, 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) Banff Center, Banff, Canada, October 5-8, 2017.
2. Dou Xuexia, “Application of data mining algorithms in the analysis of financial distress early warning model of listed company”, 2011 3rd International Conference on Computer Research and Development, 11-13 March 2011.
3. N. Srinivasan, C. Lakshmi, “Stock prediction and analysis using intermittent training data with artificial neural networks”, 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 17-18 March 2017.
4. Wang, S., Lin, C. J., Wu, C., & Chaovalitwongse, W. A. (2011). Early detection of numerical typing errors using data mining techniques. IEEE Transactions on systems, man, and cybernetics-part A: systems and humans, 41(6), 1199-1212.
5. Zhang, C., Sjarif, N. N., & Ibrahim, R. B. (2022). Decision Fusion for Stock Market Prediction: A Systematic Review. IEEE Access.
6. Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. Expert systems with applications, 32(4), 995-1003.
7. Thitimanan Damrongsakmethee and Victor-Emil Neagoe, “Data Mining and Machine Learning for Financial Analysis”, Indian Journal of Science and Technology, October 2017.
8. Suduan Chen, “Detection of fraudulent financial statements using the hybrid data mining approach”, SpringerPlus.
9. KK Tangod, GH Kulkarni, “Detection of Financial Statement Fraud using Data Mining Technique and Performance Analysis”, International Journal of Advanced Research in Computer and Communication Engineering, July 2015.
10. Haihua Liu, Shan Huang, Peng Wang and Zejun Li, “A review of data mining methods in financial markets”, AIMS Data Science in Finance and Economics Journal, 29 December 2021.
11. Bedour M. Alshammari, Fairouz Aldhmour, Zainab M. AlQenaei, Haidar Almohri, “Stock market prediction by applying big data mining”, Arab Gulf Journal of Scientific Research, 28 July 2022.
12. Elena Hernández-Nieves, Javier Parra-Domínguez, Pablo Chamoso, Sara Rodríguez-González and Juan M. Corchado, “A Data Mining and Analysis Platform for Investment Recommendations”, MDPI Article, 4 April 2021.
13. Ashvi Rebecka.K, Joy Jeffery Camillo, Volga.A Mohanan “Insight on Stock Market using Data Mining”, International Academic Conference on Applied Research in Engineering, Science & Technology, 14th – 15th September 2018.