Machine learning for detecting stock prices on NUMERAI dataset

ISSM-581(R): Research Methods III
Spring 2021

Vishal Palakonda (vpalakon@student.concordia.ab.ca)

Sri Krishna Kiriti Palacharla(spalacha@student.concordia.ab.ca)

Uttam Kumar Malla(umalla@student.concordia.ab.ca)

Research Project
Submitted to the Faculty of Graduate Studies
Concordia University of Edmonton

In Partial Fulfilment of the
Requirements of ISSM-581 course

Concordia University of Edmonton
FACULTY OF GRADUTE STUDIES
Edmonton, Alberta

Advisor: Dr Sergey Butakov (sergey.butakov@concordia.ab.ca)
Department of Information Systems Security Management
Concordia University of Edmonton,
Edmonton T5B 4E4, Alberta, Canada

Machine learning for detecting stock prices on NUMERAI dataset

Vishal Palakonda

Sri Krishna Kiriti Palacharla

Uttam Kumar Malla

Approved:

Sergey Butakov [Original Approval on File]

Sergey Butakov Date: June 23, 2021

Primary Supervisor

Patrick Kamau [Original Approval on File]

Patrick Kamau, PhD, MCIC, PChem. Date: June 23, 2021

Dean, Faculty of Graduate Studies

Table of Contents

List of Tables

List of Figures

Machine learning for detecting stock prices on NUMERAI dataset

Vishal Palakonda
Student ID: 143476
vpalakon@student.concordia.ab.ca

Uttam Kumar Malla
Student ID: 144312
umalla@student.concordia.ab.ca

Sri Krishna Kiriti Palacharla
Student ID: 143869
spalacha@student.concordia.ab.ca

ISSM581(R) Research Project, Spring 2021

Advisor: Dr. Sergey Butakov
sergey.butakov@concordia.ab.ca

Department of Information Systems Security Management
Concordia University of Edmonton,
Edmonton T5B 4E4, Alberta, Canada

Abstract- **Machine learning (ML) has numerous applications, one being able to predict or forecast the time series. Financial world was always interested in the stock market forecast, and ML tools has their niche in trading. Researchers across various fields have tried creating a fault-proof method in predicting the stock prices. In this paper, NUMERAI dataset has been used to create a prediction model using common ML models such as Decision Trees, Artificial Neural Networks, Random Forest, etc. the prediction datasets obtained using these models will be uploaded to the NUMERAI tournament. This paper provides insights on the models built and discusses the metrics associated with the results.**

Keywords- Machine Learning, NUMERAI dataset, Decision trees, Random Forest, Artificial Neural Network.

I. Introduction

Prediction of the stock market prices leads to profits and helps in changing decision-making policies. So, predicting the stock market makes it an important part of the shareholders' investment process. Predicting the stock market becomes a challenge because of the blaring, chaotic and non-stationery data and thus makes it difficult for investors to invest their money in the stock market for making profits. Many good techniques have come up for predicting the trends of the stock market.

The methodologies like Fuzzy classifier, Support vector machine (SVM) regressor, Bayesian model, Artificial neural networks regressor (ANN), etc. are being used to predict the stock market. The most used technique in attaining effective stock market predictions are fuzzy-based technique and ANN [1]. The current stock market still has a lot of limitations even after there has been an extensive research in predicting the trends of the stock market. Predicting makes it difficult because of the number of features present that makes it hard to quantify their effects on the stock prices. NUMERAI dataset is the real time financial data that helps in understanding the real-world machine learning problems better and exploring the technologies related to the field. The outcome derived by building an efficient model and using it to predict the stock prices' signals in NUMERAI dataset helps in exploring the shortcomings of the traditional machine learning tools and the new advanced tools.

The stock market data is always prone to non-economic factors such as political decisions and natural disasters which make the data unpredictable and noisy. The stock data's unpredictability is also due to the incomplete information from past behavior to enable capturing the relation between the past and future prices of the stock. The incomplete data of the stock prices are considered as noisy characteristics, which makes it a challenge to maximize gains and mitigate risks [2].

The stock prices need to be considered as a dynamic and susceptible to sudden changes due to the fundamental nature of the monetary sphere and because of the mixture of both known and unknown parameters. The main objective of this project is to learn and apply a machine learning model on the NUMERAI dataset [3], which is the stock market prices and their history and to come up with a model that is efficient in predicting the stock prices in the future.

II. Literature Review

Machine learning has been used to predict the stock prices since its rise. The most common models of machine learning used in predicting stock market prices are:

- Artificial Neural Network (ANN)
- Random Forest
- Decision Trees
- Gradient regression boosting
- Linear Regression

Neural Network Model: Neural network is a prediction method that uses mathematical models which are simpler. These mathematical models allow response variable and its predictors to have a complex linear relationship [4].

The neural network architecture is a network of neurons that are organized as layers. The predictors that are also known as inputs act as a bottom layer and forecasts known as outputs act as top layer. There may also contain some hidden layers in between also called as intermediate layer. The simplest neural network has no intermediate layers or hidden layers, so it is simply called linear regression[4].

The coefficients that are attached to the inputs are called weights. The forecast or prediction can be obtained by linear combinations of all the inputs. The weights are selected using learning algorithm in the network framework that can minimise the mean square error which is a cost function.

Once the hidden layers are added, the network becomes non-linear.

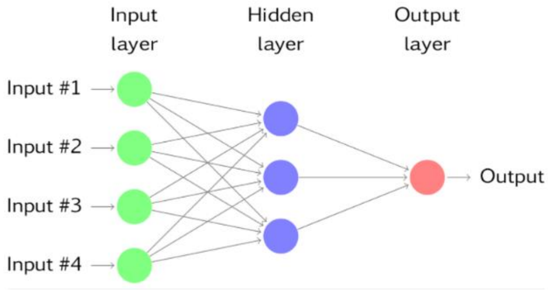

Fig 1. The above diagram shows four inputs or predictor and a single hidden layer with three neurons [4].

Decision Tree Network: Decision trees are supervised learning method used both for regression methods and classification models. The goal is to learn simple decision rules that can be inferred from the features and to create a model that can predict the target value [5].

The decision tree can be applied to regression problems using the DecisionTreeRegressor class.

Advantages of Decision trees are:

- Very simple to interpret and understand.
- Little preparation of data is enough.
- The cost associated with the decision tree is logarithmic to number of tree's data points.
- Can handle multiple output problems [5].

Disadvantages of Decision Trees:

- Minor variations in data can generate a completely different tree.
- The learners can create an over-complex trees.
- The outputs or predictions are never continuous or smooth but are a constant approximation [5].

Random Forest: This combines different algorithms for the purpose of classification and is called as ensemble classifier. Multiple decision trees are created on a random subset of data and their total votes are added to decide the class, or weights are assigned to individual trees contribution [6]. In other words, Random Forest constructs and groups multiple decision trees together to achieve more stable and accurate predictions.

Gradient regression boosting: The method which converts the weak learners to a strong learner is called boosting [7]. Each tree is known as a fit on the version that is modified from the original dataset. the AdaBoost algorithm trains a decision tree, and each observation is assigned to an equal weight. Once the first tree is evaluated, the weights of the observations that are difficult to classify are increased. And the weights of the observations were lowered which are easy to classify. Using the weighted data, the second tree is derived. The main idea here is to improve the model using the predictions generated by the first tree. The new model ten becomes Tree 1 + Tree 2 [7]. The classification error is computed from the two-tree ensemble model and a new third tree is grown to predict the new results from the revised residuals. This process is repeated for a specific number of times. The observations are classified using the new trees that were not classified by trees before[7]. The final prediction of this model is the weighted sum of all the predictions made by the models that came before.

Linear Regression: It is a supervised machine learning model whose outcome or predictions are continuous and has a constant slope. The outcomes with continuous range are predicted using this model. This model is used to predict continuous ranges rather than classifying into categories. The two types of linear regression are: simple regression, multi-variable regression.

Voting Regression: The concept of Voting Regressor is combining different machine learning regressions, in this case Gradient Regression Boosting, Random Forest Regressor, Decision Tree Regressor and Linear Regression to return the averages of the predicted values. This kind of regressor is used to balance out any weaknesses in the individually well performing models.

Even though there are many methodologies and models that are used to predict the stock prices, some models prove to be more efficient than others. According to the results presented by M. Kumar and M. Thenmozhi [8] support vector machines performed better than neural network and random forest by 5.51% and 1.04% respectively predicting the movement of direction of NIFTY index. And linear regression gave low mean squared error when predicting the EMA pattern for Vatsal H. Shah [3].

Some of the common papers have been reviewed on what stock price datasets have been used and what kind of models were prepared to predict the stock prices. The results or outcomes are also reviewed.

Auth or	Stock	Technique used	Results
[3]	YHOO, GOOGL	Decision Stump, Linear Regression, SVM	Mae: 46.665 Rmse: 57.819 Rae: 46.8% Rse: 50.9%
[9]	NASDAQ, DJIA, S&P 500	Trend Prediction, Regression, Multiclass Classification	MS: 55.341%
[10]	CNX Nifty and S&P BSE SENSEX	Support Vector Regression (SVR), Artificial Neural Network (ANN), Random Forest (RF), fusion prediction models	MS: n/a
[11]	NASDAQ, DJIA, S&P 500, Nikkei 225, HSI, FTSE100, DAX, ASX	Multiple Additive Regression Trees and SVM	MS: n/a
[12]	Nke, GS, JNJ, PFE and JPM	Random Forest Regression; Artificial Neural Network	Rmse: 1.10, Mbe: -0.0521
[13]	SSEC	Artificial neural network, Deep neural network, Recursive feature elimination	Ms: n/a
[14]	S&P CNX NIFTY	Support vector machine; Random Forest; Forecasting	MS: n/a

Table 1: Review of common papers which have been used to determine stock prices.

III. Experiment

3.1 Methodology

The NUMERAI data is available at the competition web site [15]. It consists of three parts. NUMERAI training dataset for the training of the models. NUMERAI tournament dataset which contains financial data where you use abstract financial data to train machine learning models to forecast the stock market. NUMERAI example predictions dataset contains all the example predictions for the tournament dataset.

The NUMERAI dataset is collected by the organization by converting the stock prices into signals and figuring out if the signal produced has any predictive value after it is neutralized with all the other signals [15]. All the data NUMERAI has on their website was sent by companies like PayPal who created a strong signal and send their signals in a small csv file every week to the NUMERAI organization. The organization that sends out this small csv file to NUMERAI receive large payouts when it is submitted to the NUMERAI'S API each week [15].

Data Handling: Normalizing the relevant data and making it suitable for the models that are about to explore and handle the missing data. As it is stocks data every feature will play a crucial role in the end result.

Model Building: Various Machine learning models will be used in order to get the desirable model. The expectation is to build the best model that give the most desirable results.

Model Verification: The accuracy of the model will be calculated using the R Squared value.

Dataset: The latest dataset of NUMERAI consists of 1665889 rows and 314 columns. It was downloaded on June 11th, 2021 at 18:00 UTC. The columns in this dataset are id, era, featured and target [15].

- Id: it corresponds to the stock at that particular time.
- Era: it is that particular time.
- Features: it describes the quantitative attributes of the NUMERAI stock at that particular time.
- Target: it indicates a measure of performance of four weeks into the future.

3.2 Models

Neural Network Model: The number of hidden layers used in building the model in predicting the dataset is four. The keras api layers are used as the hidden layer.

The Keras api used to get the Mean Square Error is a higher level api of TensorFlow. By using the multiple Keras layers and training the model, the MSE of the dataset is 5.374%. The custom loss function has been created and the value obtained is 40.5465%.

The prediction dataset has also been derived using the keras api layers in the neural network architecture.

Linear Regression: The multi-variable regression is used in predicting the stock price. The R squared value or coefficient of determination obtained is 0.2748%.

The intercept value obtained for Linear Regression is 48.015%.

Decision Tree Network: By using the decision tree regression model, the mean square error of the NUMERAI dataset obtained is 10.5548%. And the maximum error obtained is 1.0.

The important features from all the variables that are present has been derives as histogram below. Using the make regression function imported from sklearn.datasets, 1000 samples and 10 features were considered to get the feature importance plot.

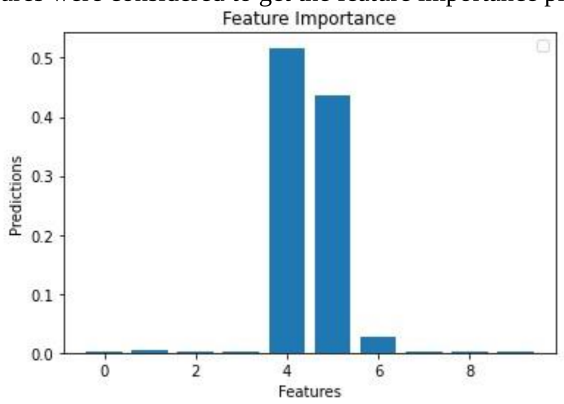

Fig 2. Importance of the attributes

.

Random Forest: The data has been pre-processed and VotingRegressor, RandomForestRegressor are imported from sklearn.linear_model. The data has been fitted for both X and Y and the predictions are made.

Voting Regression: A meta model has been created using Voting Regressor to get the best model among the others by using the estimator function and fitting the base regressors by average all the predictions into a single prediction.

Gradient regression boosting: By using the Gradient regression boosting model, the mean square error of the NUMERAI dataset obtained is 4.97%.

The blue line (training error) in the graph decreases rapidly at first and slows down gradually but decreases as more trees are added to the ensemble model. The red line (testing set error) decreases in the beginning rapidly but reaches the minimum early and starts increasing.

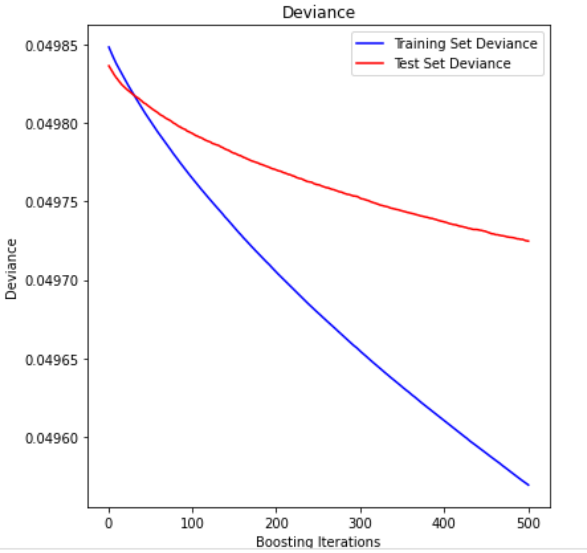

Fig 3. The above figure shows the deviance in the training dataset and testing dataset.

3.3 Results

Metrics of comparison

Metrics	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Neural Network	0.16873852	0.05208748	0.2282268
Decision Tree	0.23965993	0.1058474	0.325342
Random Forest	0.154066	0.049888	0.223357
Voting Regression	0.17904	0.0542749	0.232969
Linear Regression	0.154313	0.049959	0.2235

Table 2: Comparison of metrics of different models performed on the NUMERAI dataset.

Prediction datasets derived:

Model	Prediction Dataset		
Neural Network Model		Id	Target
	0	n0003aa52cab36c2	0.457355
	1	n000920ed083903r	0.525007
	1731014	nffc3c5ab0235de20	0.349899
	1731015	nffe7141f0da4d819	0.355876
Random Forest		Id	Target
	0	n0003aa52cab36c2	0.5150252451
	1	n000920ed083903r	0.5125545454
	1731014	nffc3c5ab0235de20	0.5000541515
	1731015	nffe7141f0da4d819	0.4975225854
Decision Tree		Id	Target
	0	n0003aa52cab36c2	0.752101145
	1	n000920ed083903r	0.752541544
	1731014	nffc3c5ab0235de20	0.258484515
	1731015	nffe7141f0da4d819	0.252568415
Voting Regression		Id	Target
	0	n0003aa52cab36c2	0.552275
	1	n000920ed083903r	0.371187
	1731014	nffc3c5ab0235de20	0.651444
	1731015	nffe7141f0da4d819	0.5851262

Table 3: prediction datasets derived from the models built on the NUMERAI dataset.

Voting Regressor: The concept of Voting Regressor is combining different machine learning regressions, in this case Gradient Regression Boosting, Random Forest Regressor, Decision Tree Regressor and Linear Regression to return the averages of the predicted values. This kind of regressor is used to balance out any weaknesses in the individually well performing models.

An average of 20 samples were randomly taken from the dataset to create a plot for all the regressors, in this case Linear Regressor, Gradient Boosting Regression, Random Forest Regreesor, Decision Tree Regressor, Neural Network (Keras layer Regressor) and Voting Regressor.

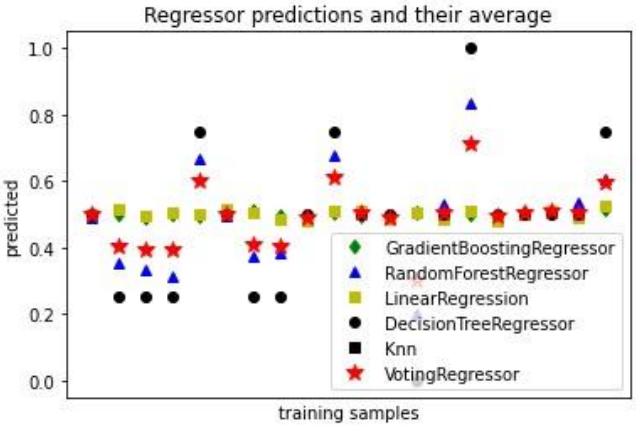

Fig 4. The above figure shows the regression predictions and their averages.

The plot for Voting Regression overlaps with Neural Network Model and Linear Regression model for most of the points which indicates that the Neural Network Model and Linear Regression is the best among the others. And there are some outliers that are close to the Random Forest Regressor.

Even though the papers that are reviewed used different stocks in order to create a good model, some of the models built are the same. The results that are obtained varies between the models created and the papers reviewed. The mean absolute error and root mean squared error are way better than the results produced by V. H. Shah [3] but the root mean squared error obtained by M. Vijha, A. Kumar, V. Tikkiwal and D. Chandola [12], is better when compared with the results derived using the Random Forest model.

Some stock prices are easy to predict than the others because there are many unknown factors involved in stock prices.

IV Conclusion

The research was carried out to create a model to predict the stock prices of the signals collected by the NUMERAI. Different regression models like Linear Regressor, Decision Tree Regressor, Neural Network Regressor, Gradient Regression Boosting and Random Forest Regressor are created, and the prediction datasets have been derived. To evaluate the performance of the models, metrics like Mean Squared error, Mean Absolute error and Root mean squared error are used to compare the models. The best results are produced by the Neural Network Model with Mse, Mae and RMse as 5.2087%, 16.87% and 22.822% respectively. The models that are created and their prediction datasets are uploaded to the NUMERAI tournament which will be evaluated and compared with the actual real values.

NUMERAI provides datasets that are completely pre-processed so there was no need to do any pre-processing. The simple features like era and id were removed since they do not have any importance. The models can be further improved to reduce the error ratio. The updated NUMERAI datasets can be found at [15].

The models have been submitted for a competition and awaiting the results.

References

[1] G. Dattatray and Kumar, "Systematic analysis and review of stock market prediction techniques," vol. 34, 2019.

[2] M. Walid, "Challenges of Stock Prediction," p. 19, 2020.

[3] V. H. Shah , "Machine Learning Techniques for Stock Prediction".

[4] Hyndman, Rob; Athanasopoulos, George, in *Forecasting: Principles and Practice*.

[5] Scikit Learn, "Decision Trees," 2018.

[6] Cuelogic Technologies, "Evaluation of Machine Learning Algorithms for Intrusion Detection system," 2019.

[7] Singh, Harshdeep, "Understanding Gradient Boosting Machines," 2018.

[8] M. Kumar and M. Thenmozhi, "FORECASTING STOCK INDEX MOVEMENT: A COMPARISION OF SUPPORT VECTOR MACHINES AND RANDOM FOREST," 2006.

[9] S. Shen and H. Jiang, *Stock Market Forecasting Using Machine Learning Algorithms,* p. 5, 2018.

[10] J. Patel, S. Shah, P. Thakkar and K. Kotecha, "Predicting stock market index using fusion of machine learning techniques," *Expert Systems with Applications,* vol. 42, no. 4, pp. 2162-2172, 2015.

[11] S. Shen, H. Jiang and T. Zhang, "Stock Market Forecasting Using Machine Learning Algorithms," *Department of Electrical Engineering, Stanford University, Stanford, CA,* pp. 1-5, 2012.

[12] M. Vijha, A. Kumar, V. Tikkiwal and D. Chandola, "Stock Closing Price Prediction using Machine Learning Techniques".

[13] J. Shen and M. O. Shafq, "Short-term stock market price trend," 2020.

[14] M. Kumar and M. Thenmozhi, "FORECASTING STOCK INDEX MOVEMENT: A COMPARISION OF SUPPORT VECTOR MACHINES AND RANDOM FOREST," 2006.

[15] NUMERAI, 2021.