Stock Price Trend Forecasting Using Supervised Machine-Learning Methods.

Pruthul Raj
IIITM,GWALIOR

***Abstract*—The project's goal is to investigate a variety of forecasting methodologies for predicting future stock returns based on previous returns and numerical news indicators in order to build a portfolio of numerous stocks to diversify risk. We accomplish this by understanding seemingly chaotic market data using supervised learning algorithms for stock price forecasting.**

Keywords—Stock-Market, Regression, Normalization, Gradient Boosting Regressor , Bagging Regressor, Random Forest Regressor, Adaboost Regressor ,K Neighbour Regressor , Root Mean Squared Error (RMSE) , R-Squared Value(r^2 value).

I. Introduction

Stock markets offer opportunities for investors to benefit while also posing risks. Scholars and skilled investors have conducted extensive research into stock market timing strategies, developing ideas and hypotheses in the process. The use of regression techniques as a predictive analytic on stock price patterns is investigated in this paper. The stock market fluctuates wildly, and there are a plethora of complex financial metrics to keep track of. However, technological advancements offer an opportunity to profit consistently from the stock market, as well as assisting experts in identifying the most useful metrics for better forecasting. The ability to forecast market value is critical in order to maximise stochastic benefit. The ability to forecast market value is critical for maximising benefit from stock option purchases while minimising risk.

II. Literature Review

Since the inception of financial markets, a great deal of research has gone into creating models that can forecast stock price movements. The efficient market hypothesis and the random walk principle are two well-known models.

The random walk hypothesis was the subject of Louis Bachelier's PhD dissertation, titled "The Theory of Speculation". The hypothesis argues that stock market price evolves according to random walk and that the market stock price cannot be predicted. The hypothesis is in line with the efficient-market hypothesis [1]. Fundamental analysis and technical analysis are the two most common schools of thinking among professional investors.

The fundamental analysis method identifies promising stocks by examining their fundamental characteristics. The intrinsic values of businesses are studied using statistics from financial reports such as the balance sheet, cash book, and profit and loss statement [2]. Operating efficiency, corporate valuation, growth equilibrium, financial leverage, and corporate liquidity are all examples of financial ratio statistics[4]. Technical analysis approach identifies chart patterns based on a company's historical share price. This approach does not gain insight into the business side of a company; it assumes the available public information does not offer a competitive trading advantage. This technique predicts trends in advance through chart patterns. Financial ratio statistics that include operating performance, corporate valuation, growth equilibrium, financial leverage and corporate liquidity form the basis of fundamental attributes [3].

Technical analysis approach identifies chart patterns based on a company's historical share price. This approach does not gain insight into the business side of a company; it assumes the available public information does not offer a competitive trading advantage. This technique predicts trends in advance through chart patterns [4].

The second method of machine learning is to look for a potentially linear or non-linear relationship that exists with enough indicators . Artificial intelligence includes a branch called machine learning. This method identifies trends in training datasets and develops its own principles, which are then applied to forecasting in testing datasets .

Machine learning methods include regression techniques. Legendre published the least squares method, which was the first type of regression, in 1805. Gauss published the Gauss-Markov theorem in 1821, which is a further extension of the principle of least squares. Francis Galton coined the word "regression" to describe a biological phenomenon in the nineteenth century. The study of stock market forecasting strategies has progressed into the world of technology. One of the most widely used methods is machine learning.

Common regression analysis involves inputs of numerical data which may consist of infinite or a wide range of values. Using the fundamental analysis method, we begin this study by collecting numerical data in real-valued format. The numerical values are then converted into ordinal values using a different transformation procedure. Only a set of categorical enumerated values makes up the ordinal values.The ordinal values contain only a range of categorical enumerated values. The relationships between the dependent and the independent ordinal variables are correlated based on the enumerated values.

III.OBJECTIVE & METHODOLOGY

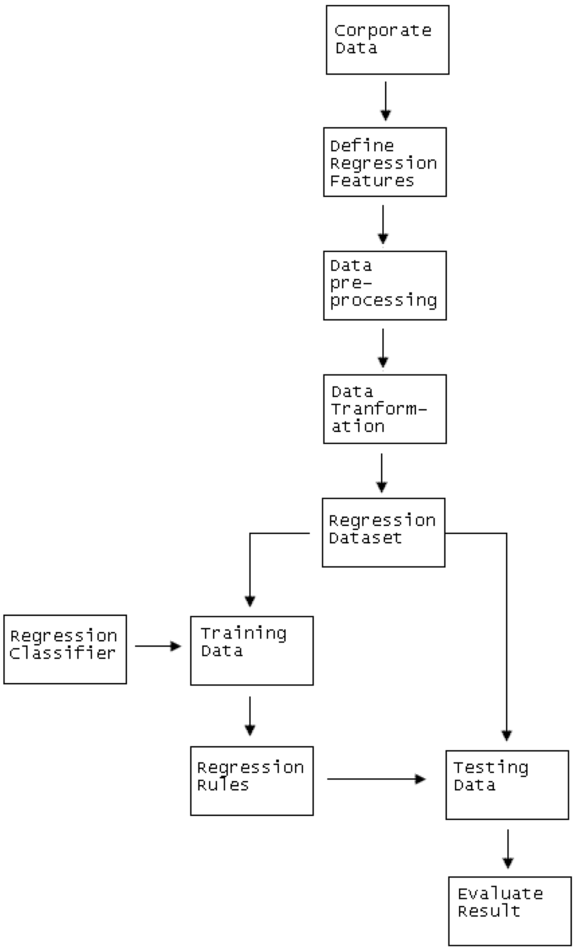

Fig. 1.Methodology

A. Data Preprocessing

The pre-processing stage involves

- Data discretization: Part of data reduction but with particular importance, especially for numerical data
- Data transformation: Normalization.
- Data Cleaning: Fill in missing values.
- Data Integration: Integration of data files.

To analyse, the data set is divided into training and testing sets after it has been converted into clean data. The training values are used as the most recent values in this case. Testing data accounts for 5-10% of the overall dataset.

B. Feature Selection and Feature Generation

We created new features from the base features which provided better insights of the data like 50 day moving average, previous day difference, etc. To prune out less useful features, in Feature Selection, we select features according to the k highest scores, with the help of an linear model for testing the effect of a single regressor, sequentially for many regressors. We used the Select K-Best Algorithm, with f regression as the scorer for evaluation.

C. Training and testing

During the data training stage, one after another each regression classifier was used as predictive analytic on the dataset. A percentage split specifies a regression classifier to split the dataset into training data and testing data proportionally. Training data provides learning process for each classifier to formulate its own regression rules. The regression rule was used on the testing data for predictions of future stock price trends. The test result was then evaluated.

IV. RESULTS & DISCUSSION

For analyzing the efficiency of the system we are used the Root Mean Square Error(RMSE) and r2 score value.

A. Root Mean Squared Error (RMSE)

The square root of the mean/average of the square of all of the error. The use of RMSE is very common and it makes an elegant general purpose error metric for numerical predictions. Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors.

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Fig. 2. RMSE Value calculation

B. R-Squared Value(r^2 value)

The R2 value can be anywhere between 0 and 1, and the higher it is, the more accurate the regression model is, since the linear regression model can explain more uncertainty. The proportionate amount of variance in the response variable explained by the independent variables is indicated by the R2 value.

The R-squared statistic indicates how close the data is to the fitted regression line. For multiple regression, it's also known as the coefficient of determination or the coefficient of multiple determination.

RESULT TABLE

Algorithm	RMSE	R^2 Value
Random Regressor	1.4325434e-07	0.956669
Bagging Regressor	1.329966e-07	0.959771
Adaboost Regressor	2.9882972e-07	0.909611
K -Neighbours Regressor	0.00039015	-117.01176
Gradient-Boosting Regressor	1.274547e-07	0.961448

V. CONCLUSIONS

According to the data, the Gradient Boosting Regressor consistently outperforms the others. Bagging Regressor, Random Forest Regressor, Adaboost Regressor, and K Neighbour Regressor come next. The Bagging Regressor is found to work well because Bagging (Bootstrap sampling) is based on the fact that combining several separate base learners reduces the error significantly.

Therefore, we want to produce as many independent base learners as possible. Each base learner is generated by sampling the original data set with replacement. From the results, it is safe to say that additional hidden layer(s) improve upon the score of the models. Random Forest is an extension of bagging where the major difference is the incorporation of randomized feature selection.

Regression techniques may benefit from the use of a particular data type, such as converting real numbers into categorical ordinal data. The outcomes are favourable when less structured data are transformed into more structured data in ordinal form. Since there are so many different data types, more studies can be done to compare the effects of transforming different data types in regression techniques for stock market trend prediction.

VI. REFERENCES

[1] Lo, A.W. and Mackinlay, A.C. A Non-Random Walk Down Wall Street 5th Ed. Princeton University Press, 2002.

[2] Graham, Benjamin; Dodd, David (December 10, 2004). Security Analysis. McGraw-Hill. ISBN 978-0071448208.

[3] Walsh, Ciaran (2003) Key Management Ratios, Third Edition, Prentice Hall.

[4] Kirkpatrick and Dahlquist. Technical Analysis: The Complete Resource for Financial Market Technicians. Financial Times Press, 2006, page 3. ISBN 0-13-153113-1.