IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 5, Ser. I (Sep. –Oct. 2022), PP 03-08
www.iosrjournals.org

Stock Price Trend Forecasting Using Machine Learning

VISHALI M
Department of Computer Science & Engineering, Sri Venkateswara Institute of Science and Technology, Tiruvallur

Abstract*— :*
Generally, predicting how the stock market will perform is one of the most difficult things to do. It can be described as one of the most critical process to predict that. This is a very complex task and has uncertainties. To prevent this problem in One of the most interesting (or perhaps most profitable) time series data using machine learning techniques. Hence, stock price prediction has become an important research area. The aim is to predict machine learning based techniques for stock price prediction results in best accuracy. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variant analysis, bi-variant and multi-variant analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. To propose a machine learning-based method to accurately predict the stock price Index value by prediction results in the form of stock price increase or stable state best accuracy from comparing supervise classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given transport traffic department dataset with evaluation of GUI based user interface stock price prediction by attributes. Dataset with evaluation classification report, identify the confusion matrix and to categorizing data from priority and the result shows that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with precision, Recall and F1 Score.

Index Terms*— Stock Price Forecasting, Machine Learning, Supervised Machine Learning Technique, Industrialistic Future Prediction.*

Date of Submission: 01-09-2022 Date of Acceptance: 12-09-2022

I. INTRODUCTION

The goal is to develop a machine learning model for real-time air quality forecasting, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm. Monitoring and preserving stock price has become one of the most essential activities in many industrial. The We aim to predict the daily adjusted closing prices of Vanguard Total Stock Market. The model will be trained using the train set, model hyper parameters will be tuned using the validation set, and finally the performance of the model will be reported using the test set. Below plot shows the adjusted closing price split up into the respective train, validation and test sets. The scope of this project is to investigate a dataset of stock price prediction using machine learning technique. We try to reduce this risk factor behind predicting from stock price to safe industrial economy so as to save lots of industrial economy efforts and assets and to predict stock price is decreased or stayed as the same

II. RELATED WORK

Lobna Nassar, Ifeanyi Emmanuel Okwuchi, Muhammad Saad - Deep Learning Based Approach for Fresh Produce Market Price Prediction.

The ARIMA model has the highest mean absolute percentage error (MAPE) hence the lowest performance compared to the conventional ML models. In addition, among the conventional the Gradient Boosting (GB) is the best due to having the least MAPE error. Finally, the performance of LSTM simple DL model is higher than all of the tested conventional ML models for two FP (Watermelon and Bok Choy). This is due to having less market for these two FP which leaves us with data that is closer in nature to time series. It is also found that the best performing model according to the aggregated measure is the compound DL model, ATTCNN-LSTM, which outperforms the ML and simple DL models in accuracy of price prediction especially after adding the attention.

Xie Chen, Deepu Rajan, Chai Quek - A Deep Hybrid Fuzzy Neural Hammerstein-Wiener Network for Stock Price Prediction

A deep hybrid fuzzy neural Hammerstein-Wiener model (FNHW), is proposed in this paper. The implication and inference of a neuro-fuzzy is based on the fuzzy rule base that has been formed during training. It requires the training data to be able to adequately represent entire system behaviors. However, the test data may vary with distribution shift in time series domain. Furthers, the training data may be derived from steady state while the test data which is in the form of dynamically changing represented by drastic data shift under certain scenario such as financial crisis. The soundness of rule base inference from neuro-fuzzy system on the steady-state data is achieved as well as inheriting the good approximation accuracy and excellent asymptotic tracking advantages of Hammerstein-Wiener model on the dynamically changing data. The effectiveness of proposed model is evaluated on two financial stock price prediction datasets. A deep hybrid fuzzy neural Hammerstein-Wiener network for stock price prediction by combining the benefits of neural fuzzy system and Hammerstein-Wiener model to handle steady-state and dynamically changing data correspondingly. Asymptotic tracking ability of Hammerstein-Wiener model to handle dynamically changing data. We performed the experiments on two different financial stock price prediction datasets and showed that the prediction performance has been significantly improved using our model when compared to other state-of-art neuro-fuzzy systems.

Rubi Gupta, Min Chen "Sentiment Analysis for Stock Price Prediction"

Analysis on StockTwits data and to understand the impact of sentiments on stock price movements. They plan to further improve the work in the following areas. First, in this work, we use two types of sentiments: bullish (positive) and bearish (negative). Adding neutral sentiment might reduce noise and potentially enhance accuracy of the work. Second, our analysis is limited to five companies. An expansion to broader set of companies or all StockTwits data might yield more insights into the data, leading to more effective application in stock price prediction. They are use the optional sentiment labels provided by StockTwits users as the ground truth data for model training. Sentiment information is used in addition to the past stock time series data to improve the accuracy of stock price movement prediction. The effectiveness of the proposed work on stock price prediction is demonstrated through engine to apply some operation on packet if packet is corrupted. Sometimes they also generate alert if any anomalies found in the packet. Basically it matches the pattern of whole string so, by changing the sequence or by adding some extra value intruder can fool the IDS but pre-processor re-arranges the string and IDS can detect the string. Pre-processor does one very important task i.e. defragmentation. Because sometimes intruder break the signature into two parts and send them in two packets so, before checking the signature both packet should be defragmented and only then signature can be found and this is done by pre-processor. The Detection Engine Its main work is to find out intrusion activity exits in packet with the help of rules and if found then apply appropriate rule otherwise it drops the packet. It takes different time to respond different packet and also depends upon the power of machine and number of rules defines in the system. experiments on five companies. Stock Twits is a relatively new micro blogging website, which is becoming increasingly popular for users to share their discussions and sentiments about stocks and financial markets. Provided a reasonable evidence that sentiments data has a positive impact on the accuracy of stock price change prediction.

Jiannan Chen, Junping Du, Feifei Kou "Prediction of Financial Big Data Stock Trends Based on Attention Mechanism"

Stock trend prediction has always been the focus of research in the field of financial big data. Stock data is complex nonlinear data, while stock price is changing over time. Based on the characteristics of stock data, this paper proposes a financial big data Stock Trend Prediction Algorithm based on attention mechanism (STPA). We adopt Bidirectional Gated Recurrent Unit (BGRU) and attention mechanism to capture the long-term dependence of data on time. Reduction algorithm based on the attention mechanism (STPA) proposed the entire algorithm is divided into three layers. That is, the stock price change trend vector representation layer, the BGRU feature extraction layer, and the stock price change trend prediction attention mechanism layer.STPA method to predict the change trend of financial big data stocks. STPA uses the Bidirectional Gated Recurrent Unit model and introduces attention mechanism technology.STPA method performs better than the current mainstream algorithms in predicting stock changes in the financial stock price data set, which demonstrates the effectiveness of the proposed method. From the experimental results, STPA method performs better than the current mainstream algorithms in predicting stock changes in the financial stock price data set, which demonstrates the effectiveness of the proposed method.

Rahma Firsty Fitriyana, Brady Rikumahu, AndryAlamsyah "Principal Component Analysis to Determine Main Factors Stock Price of Consumer Goods Industry"

Stock price is the important factor in achieving the profit in stock investment, and the prediction is usually done by relating the price of a stock to factors that influence it. The problem is, there are a large number of variables that can be used to predict the stock prices so it is difficult for a potential investor to choose which variables should be used in predicting the stock prices. This research used the Principal Component Analysis as the dimension reduction method to form major components that influence the stock prices without losing the information and uses data from five companies. Analysis method can be used to find the main determinants of stock prices by adding new variables to get more accurate results such as macroeconomic factors and adding other financial ratios, because there are many variables affecting stock prices, including macroeconomic factors that did not included in this research. Next research could include those factors to see their impact on stock prices.

III. PROPOSED METHODOLOGY

Exploratory Data Analysis of Air Quality Prediction. Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

ARCHITECTURE DIAGRAM

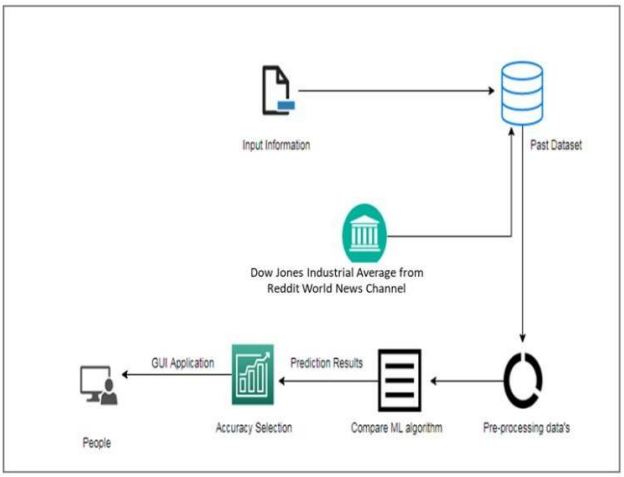

Fig. 1 Architecture of Stock Price Prediction Mechanism

At the beginning, we consider the whole training set as the root.Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.On the basis of attribute values records are distributed recursively.

We use statistical methods for ordering attributes as root or internal node. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed.

This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept.A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

1. LIST OF MODULES

2 1.1 Data Validation process.
3 1.2 Exploration data analysis of visualization

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters.:

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process.

1.2 EXPLORATION DATA ANALYSIS OF VISUALIZATION.

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots.
How to chart time series data with line plots and categorical quantities with bar charts.
How to summarize data distributions with histograms and box plots. How to summarize the relationship between variables with scatter plots.

1.3 ACCURACY RESULTS OF LOGISTIC REGRESSION AND DECISION TREE ALGORITHMS.
In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness— In the example below 2 different algorithms are compared:Logistic Regression and Decision tree

1.3.1 TRAINING THE DATASET:
The first line imports iris data set which is already predefined in sklearn module and raw data set is basically a table which contains information about various varieties.
For example, to import any algorithm and train_test_split class from sklearn and numpy module for use in this program.
To encapsulate load_data() method in data_dataset variable. Further divide the dataset into training data and test data using train_test_split method. The X prefix in variable denotes the feature values and y prefix denotes target values.

1.3.2 TESTING THE DATASET
Now, the dimensions of new features in a numpy array called ‘n’ and it want to predict the species of this features and to do using the predict method which takes this array as input and spits out predicted target value as output.

So, the predicted target value comes out to be 0. Finally to find the test score which is the ratio of no. of predictions found correct and total predictions made and finding accuracy score method which basically compares the actual values of the test set with the predicted values.

1.4 ACCURACY RESULTS OF RANDOM FOREST AND SVM ALGORITHM.

In this module we are going to compare the accuracy of Random forest and SVM.The accuracy results of each algorithm is compared.

10 1.5 GUI BASED PREDICTION RESULTS OF STOCK WILL RISE OR NOT.

Tkinter is a python library for developing GUI (Graphical User Interfaces). We use the tkinter library for creating an application of UI (User Interface), to create windows and all other graphical user interface .The inputs will be given and the output will be predicted whether the stock will rise or not.

False Positives (FP): a person who will pay predicted as defaulter. when actual class is no and predicted class is yes. e.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

1.5.1 COMPARING ALGORITHM WITH PREDICTION IN THE FORM OF BEST ACCURACY RESULT

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit- learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

1.5.2 PREDICTION RESULT BY ACCURACY

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

True Positive Rate(TPR) = TP / (TP + FN)

False Positive rate(FPR) = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non- defaulters.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Precision = TP / (TP + FP)

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict). Recall= TP/ (TP+FN). Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and

false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

IV. CONCLUSION

We also observed that the choice of the indicator function can dramatically improve/reduce the accuracy of the prediction system. Also a particular Machine Learning Algorithm might be better suited to a particular type of stock, say Technology Stocks, whereas the same algorithm might give lower accuracies while predicting some other types of Stocks.

REFERENCES

[1]. Lobna Nassar, " Integrated Long-term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market", DOI 10.1109/ACCESS.2020.2969293, IEEE Access.
[2]. DeepuRajan, "A Deep Hybrid Fuzzy Neural Hammerstein-Wiener Network for Stock Price Prediction ", Xie Chen, DeepuRajan, Chai Quek School of Computer Science and Engineering Nanyang Technological University 50 Nanyang Avenue, Singapore.
[3]. Rubi Gupta, "Deep Learning Based Approach for Fresh Produce Market Price Prediction", Electrical and Computer Engineering Department University of Waterloo Ontario, Canada.
[4]. JiannanChen, "Prediction of Stock Prices using Machine Learning (Regression Classification) Algorithms ", 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020
[5]. Rahma Firsty Fitriyana, "CUDA parallel computing framework for stock market prediction using K-means clustering", Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020) IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9
[6]. Feifei Kou "Principal Component Analysis to Determine Main Factors Stock Price of Consumer Goods Industry", 2020 International Conference on Data Science
[7]. ZheXue"Prediction of Financial Big Data Stock Trends Based on Attention Mechanism", 2020 IEEE International Conference on Knowledge Graph (ICKG)

VISHALI M. "Stock Price Trend Forecasting Using Machine Learning." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 24(5), 2022, pp. 03-08.