Done By- Ovi Jadhav & Swara Vedak

STOCK PREDICTION - REPORT

Title Report

Introduction

This is the prediction about stock market and we can solve it by using classification algorithm. So, in this research we use linear regression and Decision tree regression as a classification for prediction stock market. STOCK MARKET is an ancient technique where people can easily do the trade stock and they can either lose or profit. People who have a misconception about the stock market are that they think it looks like gambling because they just know about profit or loss but nothing else.

So the lacking of proper information and analyzing capability people think like that but the revolution of data science, big data, and awareness of the people are getting proper guidelines about the future and passed trading information. The stock market is related to the individual and national economy, so connection to the ethnic economy and it's vital for a nation because of the impact of each country's GDP.

A company sells their stock by the stock exchange then, they listed the price that is called IPO or initial public offering, and then the people buy that share from brokerages. Brokerages work like an intermediate medium between seller and buyer but for that brokerages charge the amount for doing this job [1]. It can be a bank, any small company including licenses, and so on.

Awareness is the main part of the stock market nowadays because needs to have a solid idea about trending stock rate, past and future also. This awareness could understand the upcoming rate of stock and be analyze the risk. Now a day, all of these could possible by the invention of data science, machine learning, and big data. So we can simply say that machine learning and data science is the best

gift for the stock market and another relevant field. The objective of this paper is to get a better decision using two supervised regression machine learning algorithms and the use of statistic formula gives us better accuracy of stock price predict. Actually, would discuss two regression models using different size of the dataset and getting the performance in a different aspect. So, the main concern about this paper is the machine learning algorithm, statistic, and a graphical view of data from a different outcome. Finally, compare the perform**ance based on the size of the dataset and algorithmically.**

ABSTRACT-

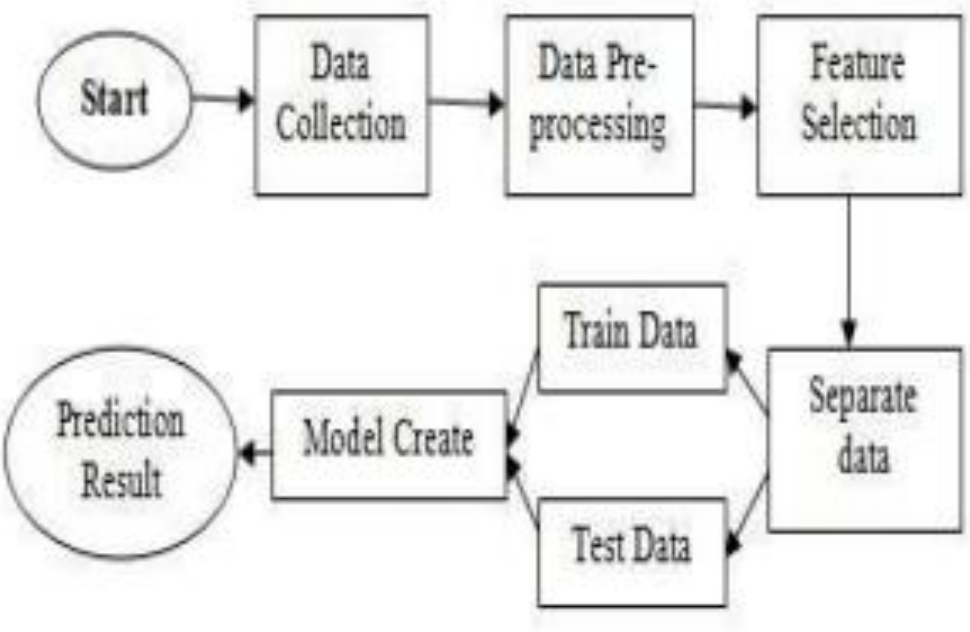
In business, the Stock market is a more perplexing and sophisticated way to do business. Every business owner wants to reduce the risk and make an immense profit using an effective way. The bank sector, brokerage corporations, small ownerships, all depends on this very body to earn profit and reduce risks. However, using the machine learning algorithm of this paper to predict the future stock price and shuffle by using subsist algorithms and open source libraries to assist in inventing this unsure format of business to a bit more predictable.

The proposed system of this paper works in two methods – Linear Regression and Decision Tree Regression. Two models like Linear Regression and Decision Tree Regression are applied for different sizes of a dataset for revealing the stock price forecast prediction accuracy. Moreover, the authors of this paper have revealed some development that could be the club to acquire better validity in these approaches.

Keywords —Data Analysis, Linear Regression, Decision Tree Regressor, Big Data, Stock Analysis, Supervised Machine Learning

PROPOSED METHODOLOGY:

The methodology of a paper is a vital and key feature because this stage discusses the whole process of functional activity. The machine learning algorithm mainly followed the four steps of methodology for analyzing the data and predicting the outcome and taking a decision..



A. Dataset which is provided by our professor is about the following year stock analysis and we need to predict the stock for the next day. So using few machine learning models we aim to predict the future/next day value. The models we are using to predict the future/next day values are 1. Linear Regression 2. Decision Tree Regression Model.

B. Data Cleaning and Preprocessing During this period, we checked whether data have a null value and unknown types of data that have been cleaned and filled up using a statistical formula. Like, where we founded a null value in a correspondent attribute then I checked types of values either discrete or classifier value. If, its classification value then calculating the median else calculated mean value and put it on the null places.

C. Separate Data into Train and Test Dataset Prepared data separated into two part train and test the ratio of 80% and 20% respectively. The percentage of trains and tests would impact the accuracy of predicting the result. For three sizes of a dataset, we have applied the same ratio for the train and test. At this stage what ration you want to choose for the train and test dataset it's up to you but if you take more train dataset compare to test then accuracy would be better. The general ratio for test and train dataset is 80% and 20% respectively

D. Train Data Fit in Model Choose an algorithm is vital for getting better predictions and selecting a proper algorithm based on the dataset. For this dataset, I selected Linear Regression and Decision Tree Regression to calculating the prediction of the diverse size of the dataset and train the model using the training dataset.

E. Test the Data The test data set is 20% of the total and this learning approach applied the test dataset and gets the result then evaluates with the actual output.

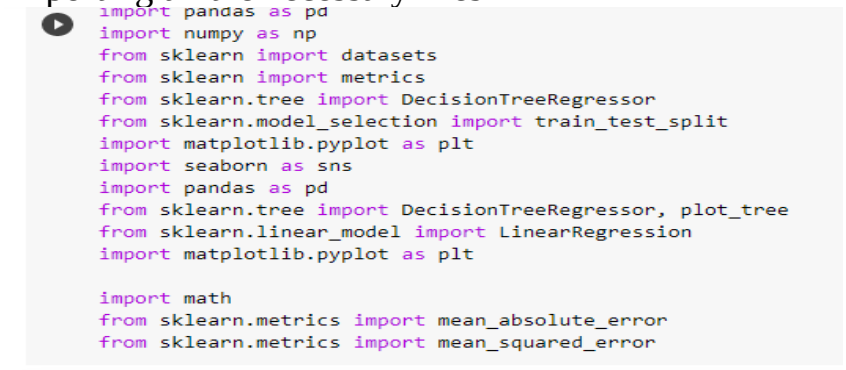
Model we have used:

Linear regression is a commonly used statistical technique in stock analytics because it can help us understand the relationship between different variables that can impact stock prices, such as company financials, market indices, interest rates, and geopolitical events. Linear regression can also help us make predictions about future stock prices based on historical data and other relevant factors.

One of the key benefits of using linear regression in stock analytics is that it can help us identify trends and patterns in stock price movements. This can be useful in developing trading strategies and making investment decisions.

Steps we have used to predict the data

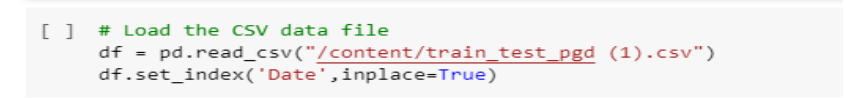
1. Importing all the necessary files



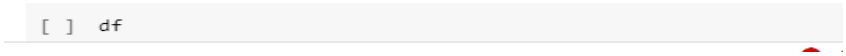
```
import pandas as pd
import numpy as np
from sklearn import datasets
from sklearn import metrics
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
from sklearn.tree import DecisionTreeRegressor, plot_tree
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt

import math
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
```

```
New Section
```



```
# Load the CSV data file
df = pd.read_csv("/content/train_test_pgd (1).csv")
df.set_index('Date',inplace=True)
```



```
df
```

```
[15] # Create a linear regression model and fit it to the training data
     model = LinearRegression()
     model.fit(Xtrain, ytrain)
```

```
LinearRegression()
```

```
[16] model.score(Xtrain,ytrain)
```

```
0.9930794311983154
```

```
[17] ypred = model.predict(X_test)
     df_pred = pd.DataFrame(y_test.values, columns=['Actual'], index=y_test.index)
     df_pred['Predicted'] = ypred
```

```
# Predict the Close value for the test data
y_pred = model.predict(X_test)
```

```
[ ] df['Forecast']= df['Close'].shift(-1)
```

```
[ ] df.tail()
```

```
df.shape
```

```
[ ] df.isnull()
```

```
[ ] df= df.drop('2022-12-30')
```

```
[ ] df.tail()
```

```
[ ] X= df[['Open','Low','High']]
    y= df['Forecast']
```

```
[ ] Xtrain=X[:600]

    ytrain=y[:600]

    X_test=X[600:]
    y_test=y[600:]
```

```
[19] # Print the accuracy score
     accuracy = model.score(X_test, y_test)
     print("Accuracy:", accuracy)
```

```
Accuracy: 0.980143992960772
```

```
[20] df_pred["Difference"]=(df_pred["Actual"]-df_pred["Predicted"])*100/df_pred["Actual"]
```

```
df_pred.tail()
```

Date	Actual	Predicted	Difference
2022-12-23	893.200012	882.574784	1.189569
2022-12-26	900.650024	888.118795	1.391354
2022-12-27	898.950012	894.656046	0.477665
2022-12-28	908.049988	903.183756	0.535899
2022-12-29	890.849976	908.372505	-1.966945

2. Decision Tree Model

   Decision Tree Model is a machine learning algorithm that is used for both classification and regression tasks. It is a tree-like model where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label or a numerical value.

   Decision tree model is widely used in various fields, including finance, marketing, healthcare, and many more. It is a powerful tool that can help us make decisions based on data, and it is often used in combination with other machine learning algorithms to improve their performance.

```
[ ]  # Load the dataset
     dt = pd.read_csv('/content/train_test_pgd (1).csv')
     dt.set_index('Date',inplace=True)
     dt
```

```
[ ]  # Separate features and target variable
     dt['Forecast']= dt['Close'].shift(-1)
```

```
[ ]  dt
```

```
     dt.tail()
```

```
     dt.shape
```

```
[ ]  dt.isnull()
```

```
[ ]  dt= dt.drop('2022-12-30')
```

```
[ ] dt= dt.drop('2022-12-30')
```

```
[ ] dt.tail()
```

```
[ ] #high-low open-close
    X= dt[['Open','High','Low']]
    Y= dt['Forecast']
```

```
[ ] # Split the data into training and testing sets
    X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.20,random_state = False)
```

```
model = DecisionTreeRegressor()
model.fit(X_train, Y_train)
```

```
DecisionTreeRegressor
DecisionTreeRegressor()
```

```
[ ] y_pred = model.predict(X_test)
    metrics.r2_score(Y_test,y_pred)
    #R2 score shows much of the variation of a dependent variable is explained by an independent variable in a regression model.
```

```
0.990247998781668
```

```
[ ] model
```

```
DecisionTreeRegressor
DecisionTreeRegressor()
```

```
dt_pred = pd.DataFrame(Y_test.values, columns=['Actual'], index=Y_test.index)
dt_pred['Predicted'] = y_pred
```

```
[ ]
```

```
[ ] rmse = math.sqrt(mean_squared_error(Y_test, y_pred))
    rmse
```


```
[ ] rmse = math.sqrt(mean_squared_error(Y_test, y_pred))
    rmse
```


```
17.14721767445323
```

```
[ ] mae = mean_absolute_error(Y_test, y_pred)
    mae
```

```
12.579665934244794
```

LIMITATIONS OF THE MODEL USED:

1. Linear Regression Model:

Linear regression is a commonly used statistical technique in stock analysis, but it has several limitations that should be considered when using it to analyze stocks.

Linear Assumptions: One of the main assumptions of linear regression is that the relationship between the independent variable(s) and the dependent variable is linear. However, the relationship between stock prices and the variables that affect them may not be linear. For example, there may be nonlinear relationships between a company's earnings and its stock price due to market sentiment and investor behavior.

Multicollinearity: Another issue that may arise in stock analysis is multicollinearity, which occurs when two or more independent variables are highly correlated. This can lead to difficulties in interpreting the coefficients of the regression equation and can result in inaccurate predictions.

Outliers: Linear regression is also sensitive to outliers, which are extreme values that can skew the results of the analysis. In stock analysis, outliers may be caused by unexpected events such as market shocks, earnings surprises, or geopolitical events.

Stationarity: The assumption of stationarity in linear regression implies that the statistical properties of the data do not change over time. However, stock prices are known to be nonstationary, with trends and patterns that can change over time due to various economic, political, and social factors.

Overfitting: Overfitting occurs when a regression model is overly complex and is trained to fit the noise in the data rather than the underlying relationships. This can lead to poor performance when the model is used to make predictions on new data.

Data Quality: Finally, the quality of the data used to train the linear regression model can significantly impact its accuracy and usefulness. If the data is incomplete, inconsistent, or inaccurate, the model's predictions may be unreliable.

It's important to keep these limitations in mind when using linear regression in stock analysis and to consider alternative techniques such as nonlinear regression, time-series analysis, and machine learning algorithms when appropriate.

2. Decision Tree Model:

Decision tree models have become increasingly popular in stock analysis due to their ability to handle nonlinear relationships and interactions between variables. However, like any other model, decision trees have limitations that should be considered when using them for stock analysis.

Overfitting: Decision trees are prone to overfitting, which occurs when a model is too complex and captures noise in the data instead of the underlying relationships. This can result in a model that performs well on the training data but poorly on new data.

Instability: Decision trees are sensitive to small changes in the data, which can lead to instability and unreliable predictions. This can be especially problematic in stock analysis, where unexpected events can have a significant impact on the market.

Interpretability: While decision trees can be useful in making predictions, they can be difficult to interpret. This is because decision trees can become quite complex, with many branches and nodes. This can make it challenging to understand how the model arrived at a particular prediction.

Variable selection bias: Decision trees can be influenced by the variables used to build them. If certain variables are included or excluded, this can affect the accuracy of the model. Therefore, it's important to carefully select the variables used in the model to ensure that they are relevant and meaningful for stock analysis.

Limited extrapolation ability: Decision trees are limited in their ability to extrapolate beyond the range of the data used to train them. This means that

predictions may be less reliable when applied to data outside of the range of the training data.

Handling of missing values: Decision trees can struggle with missing values in the data. This is because decision trees require complete data to make predictions. Therefore, imputation of missing values is often required before using decision tree models.

Overall, decision tree models can be a powerful tool in stock analysis. However, it's important to be aware of their limitations and to carefully select the variables used in the model to ensure that the predictions are reliable and interpretable.

Conclusion:

Machine learning has some great application and still, now it's a very popular tool and it's also depending much on data even it has evolved the future into a neural network and deep learning. All about this paper is stock market price predict using machine learning. There are various ways to implement the stock price prediction but applied only two regression algorithm.
The main aim of this paper is to use supervised machine learning algorithm Linear Regression and Decision Tree Regression have been applied for stock price prediction. The result reveals that Linear Regression is given better accuracy for both small and big datasets. On the other hand, Decision Tree Regression expresses the poor prediction price based on the size of the dataset.

Resources Used

1. Google [www.google.com]
2. Google. Scholar [www.google.scolar.com]
3. Kaggle [www.kaggle.com]

Thank You,
Swara Vedak-10
Ovi Jadhav-14