CSE 4554

TEAM:
Azmayen Fayek Sabil - 190042122
Tasfia Barshat - 190042126
S.A.M Sajratul Yeaken Mollah Prangon - 190042132

Machine Learning Project
Stock Price Prediction

Introduction

Prediction of future stock prices has been an area that attracted many of us over a long period of time. While some believe that it is impossible to predict stock prices accurately, there are formal propositions demonstrating that with the choice of appropriate variable and suitable modeling, it is possible to predict the future stock prices and stock price movement patterns, with a fairly high level of accuracy. Researchers have also worked on technical analysis of stocks with a goal of identifying patterns in the stock price movements using advanced data mining techniques.

Problem statement

The goal of this project is to collect the stock price of SP500 from the NYSE of the United States over a long period of 95 years. We hypothesize that it is possible for a machine learning model to learn from the features of the past movement patterns of dailySP500 index values, and these learned features can be effectively exploited in accurately forecasting the future index values of the SP500 series. In the present work, we follow four different approaches in building long and short-term memory (LSTM) network-based models in order to augment the predictive power of our forecasting models.

Application domain

Challenges

First challenge was to collect the suitable dataset that contains a large number of samples. After that we had to do some preprocessing of the dataset to make it more meaningful and helpful to the model. Third challenge was to find that sweet spot where the model will perform the best.

Contribution

This project has the scope to contribute to the stock price prediction scenario.

Background Study

Before starting off the project, we dived into the study of Stock exchange and related research papers. In addition to this, we looked into previously done work on stock price prediction using various models. Through performing extensive exploration, it was understood that time series models perform better with LSTM and ARIMA.

Overview of the project and related terms

This project tries to predict the future value of SP500 stock. We will for now try to predict if tomorrow's closing value is greater or less than today's closing value.

- Closing value
- Target value

Data collection and feature analysis

The dataset for this project was collected from yahoo finance.

Description of the models

We used four algorithms here.

- Randomforest classifier
- Linear regression
- Decision tree
- LSTM
- ARIMA (in progress)

Implementation

Overview of the experiment

Our goal is to predict the future stock prices and see which models perform better to predict the prices.

Feature engineering

First we decreased the size of the dataset. As it was from 1927-2022 and data between 1927 and 1990 showed very linear behavior. So we thought it wouldn't help much in terms of training the models. Then what we did was;

- Dataset size decrement
- Deleting samples containing null values
- Introducing some new features from the existing ones

Training and test set generation

First we trained and tested the dataset by splitting them.

- For randomforest it was last 100 samples for testing and rest of the samples for training
- We used backtesting method to train the model more efficiently and trained and tested randomforest again
- Then we created some interesting features from the existing ones and retrained and retested using randomforest.
- For the linear regression model the split was 75-25.
- For the LSTM model the split was 95-5.

Running the classifier

Result Analysis

Overview

We tried different models with different sets of datasets or different approaches.

Generation of confusion matrix

–hasn't been generated yet

Result analysis

- Precision Score for RandomForest Classifier : 41.33
- Precision Score for RandomForest Classifier after backtesting : 53.04
- Precision Score for RandomForest Classifier after feature engineering : 56.89

- R2 score for Linear Regression : 18.85
- For the general decision tree classifier, it didn't work that well
- RMSE for LSTM for 1 epoch : 70.59

Conclusion

This project proposes RFC,LSTM built to forecast future values for SP500 assets, the result of our model has shown some promising results. The testing results confirm that our model is capable of tracing the evolution of closing prices for SP500s. For our future work we will try to find the best sets for data length and number of training epochs that better suit our assets and maximize our predictions accuracy.

References

1. Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models Sidra Mehtab , Jaydip Sen and Abhishek Dutta
2. Social Media and Forecasting Stock Price Change Joseph Coelho Dawson D'Almeida Scott Coyne Nathan Gilkerson Katelyn Mills

3. Stock Market Prediction Using LSTM Recurrent Neural Network Adil MOGHAR, Mhamed HAMICHE