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 (New York Stock Exchange) 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 daily SP500 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 five different approaches to modeling the stock price prediction. Moreover, in building long and short-term memory (LSTM) network-based models are used in order to augment the predictive power of our forecasting models.

Application domain

Stock market price prediction is a part of the Financial domain which can be used to correctly predict or forecast the next stock prices based on previous data.

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 the best fit where the model will perform the best.

Contribution

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

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 like stock prices 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. The Standard and Poor's 500, or simply the S&P 500, is a stock market index tracking the stock performance of 500 large companies listed on stock exchanges in the United States.

SP500 (standard and poor's 500) dataset:

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits
1927-12-30 00:00:00-05:00	17.660000	17.660000	17.660000	17.660000	0	0.0	0.0
1928-01-03 00:00:00-05:00	17.760000	17.760000	17.760000	17.760000	0	0.0	0.0
1928-01-04 00:00:00-05:00	17.719999	17.719999	17.719999	17.719999	0	0.0	0.0
1928-01-05 00:00:00-05:00	17.549999	17.549999	17.549999	17.549999	0	0.0	0.0
1928-01-06 00:00:00-05:00	17.660000	17.660000	17.660000	17.660000	0	0.0	0.0
...	...	...	...	...	...	...	...
2022-12-16 00:00:00-05:00	3890.909912	3890.909912	3827.909912	3852.360107	7493660000	0.0	0.0
2022-12-19 00:00:00-05:00	3853.790039	3854.860107	3800.040039	3817.659912	3969610000	0.0	0.0
2022-12-20 00:00:00-05:00	3810.469971	3838.239990	3795.620117	3821.620117	3985370000	0.0	0.0
2022-12-21 00:00:00-05:00	3839.489990	3889.820068	3839.489990	3878.439941	3775200000	0.0	0.0
2022-12-22 00:00:00-05:00	3853.260010	3853.260010	3819.550049	3820.750000	276590687	0.0	0.0

23860 rows × 7 columns

Description of the models

We used five models in our project to compare the result from different models using the same dataset.

- Linear regression
- Decision tree
- Randomforest classifier
- LSTM
- ARIMA

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

Decreasing the size of the dataset was the first step. As the data was from the timeline 1927-2022 and data between 1927 and 1990 showed very linear behavior, so it would not help much in terms of training the models. After that, as part of the feature engineering process, the following steps were conducted:

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

Training and test set generation

Splitting the dataset into training and testing sets.

- For the Random Forest classifier, the last 100 samples were used for testing and rest of the samples for training.
- We used the Backtesting method to train the model more efficiently as well as trained and tested Random Forest classifier again after Backtesting.
- After that, we created some interesting features from the existing ones and retrained and re-tested using Random Forest classifier.

- For the linear regression model the split was 75-25 which means 75% data were used for training while the remaining 25% data for testing the model.
- For the LSTM model the split was 95-5 which means 95% data were used for training while the remaining 5% data for testing the model.
- For the ARIMA model, the split was 90-10 which means 90% data were used for training while the remaining 10% data for testing the model.

Running the classifier

- In case of Linear Regression, we had to choose the correct features for the model. First we fed the model all the features except Target as that was the purpose for the model to predict. And second we fed the model just the tomorrow and close to predict the price.
- In case of Decision Tree, we did almost the same as the linear regression. First we fed the model all the features except Target as that was the purpose for the model to predict. And second we fed the model just the tomorrow and close to predict the price.
- In case of Random Forest Classifier, we tried three different scenarios for the model to train. First was to give all the features to the model to predict the price. Then we introduced a function called backtesting, what it does is it first takes n-k samples to train and k samples to test. Here n is the total number of days and k kind of represents step size. Another approach was that we generated n-th days prior closing value of the stock. Using those extra features we again trained the model.
- In the case of LSTM, we trained the model using only the close feature and predicted the target value. Then we took a step ahead and used the close and close prior to 5 days to the current date to train the model.
- In case of ARIMA, first the autoARIMA feature was used to extract the best fitted value for p,d and q in order to build the model. After that, we build the ARIMA model using the best fitted value which was (0,1,2) for our used dataset.

Result Analysis

Overview

We used five different models approaches applied on the same dataset to predict the output and performed a comparative analysis on the result.

Result analysis

Model Name	Conditions	Batch Size	Epochs	Precision Score	RMSE	MAPE
Linear Regression	Initial Dataset	N/A	N/A	33.66	N/A	N/A
	Used close and tomorrow features	N/A	N/A	33.73	N/A	N/A
Decision Tree	Initial Dataset	N/A	N/A	20.2995	N/A	N/A
	Used close and tomorrow features	N/A	N/A	48.543	N/A	N/A
Random Forest Classifier	Initial Dataset	N/A	N/A	38.7	N/A	N/A
	After backtesting	N/A	N/A	53.02	N/A	N/A
	After improving the dataset	N/A	N/A	56.89	N/A	N/A
LSTM	Used Close feature	1	1	N/A	96.18	
	Used Close feature and Close feature 5 days prior to the current date	16	1	N/A	76.02	1.67
		16	3	N/A	71	1.51
		16	5	N/A	48.22	1.01
ARIMA		N/A	N/A	N/A	27.78	2.8

After analyzing the results of different models we came across these few suggestions that could have helped us to improve the result of our models.

1. We can introduce new columns like inflation rate.
2. Natural language processing: by seeing different news websites and analyzing different behavior of different companies that can help the models to predict stock price of a company.
3. Stock prices are driven by a number of factors like industry performance, company news and performance, investor confidence, micro and macro economic factors like employment rates, wage rates etc.
4. We can also explore the relationship between StockTwits and stock market closing prices through URL mining and sentiment analysis to check whether this information can enhance stock price prediction accuracy.
5. One approach can be like this where a model is required to make a one-week prediction, and the actual data for that week is used in the model for making the forecast for the next week. This is both realistic and practical, as in most of the real-world applications, forecast horizons longer than one week are not used.

Conclusion

This project proposes LSTM and ARIMA models built to forecast future values for SP500 assets, the conclusion displaying our model has shown some promising results. The testing results confirm that these two models are capable of tracing the evolution of closing prices for SP500s. For our future work, we plan 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