New Mexico State University

Applied Machine Learning

(C S - 487)

Stock Market Prediction System

Project Report

Team Members:

Sikta Das

Nathan Gonzales

Spring 2020

PROJECT PROPOSAL : Create a Stock Market Prediction System in Python

The Stock Market can be very unpredictable. It is one of the most well-known infrastructures through which anyone can potentially make a fortune. Computer scientists, specifically data science people, find it very interesting to be able to predict the unknown. Machine Learning has the potential to help predict the financial forecast of the stock market. This is our first experience with Machine Learning, and we think this would be a fun project to work on.

PROBLEM:

Recently, we have noticed that stocks related to travel such as different airlines and cruises have been falling due to all sorts of travel bans because of the COVID-19. We plan on looking into the airline stock data and predicting a model. Although, it will probably take time for everything to get back to normal, the model would give an idea as to how long it will take for the stocks to get back up, once everything is back to normal.

We aim to predict the daily adjusted closing prices of Airline stocks, using data from the previous N days (ie. forecast horizon=1). We will use five years of historical prices for each of the airlines.

SOLUTION:

We will split the data into three parts:

- Training Data - 60%
- Testing Data - 20%
- Validation Data - 20%

We will do this each of the three datasets and we use the validation data to tune the model parameters. We will train the data using the training data and evaluate the performance of the model using the testing data.

We will use the following methods for defining and creating our model:

- Last Value
- Moving Average
- Long Short Term Memory

To evaluate the effectiveness of our methods, we will use the root mean square error and mean absolute percentage error metrics. For both metrics, the lower the value, the better the prediction.

DESCRIPTION OF DATASETS:

We downloaded daily data of three airlines - Southwest Airlines (LUV.csv), American Airlines (AAL.csv) and United (UAL.csv) - for the past 5 years from finance.yahoo.com.

Each of our three datasets has the following attributes:

Attribute	Description	Type
Date	Trading Date	Object
Open	Price of a security at the beginning of the trading day	Float
High	Highest price in a given period of time	Float
Low	Lowest price in a given period of time	Float
Close	Price of a security at the end of the trading day	Float
Adj Close	Adjusted closing price amends a stock's closing price to accurately reflect that stock's value after accounting for any corporate actions.	Float
Volume	Number of shares that changed hands during a given day.	Integer

LUV.csv (Southwest Airlines):

Number of Rows	1259
Number of Columns	7

	Open	High	Low	Close	Adj Close	Volume
count	1259.000000	1259.000000	1259.000000	1259.000000	1259.000000	1.259000e+03
mean	49.987927	50.560826	49.391755	49.972105	48.605800	5.802736e+06
std	8.058949	8.058744	8.092287	8.063319	8.179054	3.250383e+06
min	31.690001	32.889999	29.150000	30.860001	30.832319	1.098400e+06
25%	42.810000	43.320000	42.225001	42.820000	40.910856	3.847450e+06
50%	51.840000	52.400002	51.330002	51.880001	50.840229	5.120600e+06
75%	55.929998	56.545000	55.355000	55.935000	54.771265	6.945550e+06
max	66.500000	66.989998	65.980003	66.290001	64.447029	4.382490e+07

AAL.csv (American Airlines):

Number of Rows	1258
Number of Columns	7

	Open	High	Low	Close	Adj Close	Volume
count	1258.000000	1258.000000	1258.000000	1258.000000	1258.000000	1.258000e+03
mean	39.641089	40.215692	39.043545	39.604189	38.474277	8.711933e+06
std	8.245643	8.269731	8.232948	8.259598	7.819786	8.050554e+06
min	10.650000	11.360000	10.010000	10.250000	10.250000	1.352400e+06
25%	33.292500	33.835000	32.632500	33.214999	32.604170	5.249350e+06
50%	40.254999	40.860001	39.660000	40.215001	38.726505	6.960400e+06
75%	46.027499	46.557501	45.434999	45.965001	44.528633	9.603475e+06
max	58.790001	59.080002	57.799999	58.470001	56.988728	1.015531e+08

UAL.csv (United Airlines):

Number of Rows	1256
Number of Columns	7

	Open	High	Low	Close	Adj Close	Volume
count	1256.000000	1256.000000	1256.000000	1256.000000	1256.000000	1.256000e+03
mean	69.338113	70.221584	68.412962	69.279888	69.279888	4.607454e+06
std	14.620562	14.623152	14.650414	14.636181	14.636181	3.833029e+06
min	21.340000	23.990000	17.799999	21.280001	21.280001	7.711000e+05
25%	57.770000	58.464999	56.895002	57.654999	57.654999	2.673400e+06
50%	69.215000	69.915001	68.125000	69.220001	69.220001	3.795800e+06
75%	82.439999	83.587502	81.387499	82.384998	82.384998	5.219300e+06
max	97.669998	97.849998	95.959999	96.699997	96.699997	6.782620e+07

RESULT ANALYSIS

Dataset: Southwest Airlines

	Method	RMSE	MAPE(%)
0	Last Value	1.125	1.470
1	Moving Average	1.237	1.644
2	LSTM	1.365	1.813

1. Last Value - On the test set, the RMSE is 1.125 and MAPE is 1.470%.
2. Moving Average Implementation:
   - We will use N_opt=2 in this work since our aim here is to use moving average
   - On the test set, the RMSE is 1.237 and MAPE is 1.644% using N_opt=2
3. Long Short Term Memory - On the test set, the RMSE is 1.365 and MAPE is 1.813%.

Dataset: American Airlines

	Method	RMSE	MAPE(%)
0	Last Value	0.836	2.423
1	Moving Average	0.991	2.981
2	LSTM	1.261	2.611

1. Last Value - On the test set, the RMSE is 0.836 and MAPE is 2.423%
2. Moving Average Implementation:
   - We will use N_opt=2 in this work since our aim here is to use moving average
   - On the test set, the RMSE is 0.991 and MAPE is 2.981% using N_opt=2
3. Long Short Term Memory - On the test set, the RMSE is1.261 and MAPE is 2.611%

Dataset: United Airlines

	Method	RMSE	MAPE(%)
0	Last Value	1.964	1.982
1	Moving Average	2.248	2.981
2	LSTM	1.946	1.776

1. Last Value - On the test set, the RMSE is 1.964 and MAPE is 1.982%
2. Moving Average Implementation:
   - We will use N_opt=2 in this work since our aim here is to use moving average
   - On the test set, the RMSE is 2.248 and MAPE is 2.981% using N_opt=2
3. Long Short Term Memory - On the test set, the RMSE is 1.946 and MAPE is 1.776%

LIMITATIONS

- I am using time series, yet splitting it up into training and test sets randomly. I was not sure how or on what basis I could split it.
- I was not able to implement Extreme Gradient Boosting since for some reason I was not able to install the xgboost package onto my mac.