Stock Market Forecasting

OBJECTIVE

The analysis in this report focused on deep learning models for sequential data such as stock price. Stock market, as one of financial market, is highly volatile with characteristics of high risk and high return which attracts majority of investors. In the past, most of investors relied on traditional statistical methods that might not be able to detect fluctuations due to hidden important information. However, with advanced progress of science and technology, implementation of machine learning applications rises in economics field, especially investment, in recent years. This increases prediction accuracy with many optimized stock forecasting methods developed. Hence, it could increase of investor's change to earn higher profits through trading and investment.

DATASET

The data used in this analysis was sourced from Yahoo Finance that contains daily historical data from earlier 2020 until mid-2023. Data for Apple Inc. stock quote, shortened as **AAPL**, was downloaded as a DataFrame, which provides **Date**, **Open** (daily opening stock price), **High** (daily maximum stock price), **Low** (daily minimum stock price), **Close** (daily closing stock price), and **Volume** (daily total number of shares traded). Below is the table containing the mentioned data.

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits	Date
2020-01-02	72.246685	73.310001	71.990615	73.249031	135480400	0.0	0.0	18263.0
2020-01-03	72.468610	73.305113	72.310090	72.536896	146322800	0.0	0.0	18264.0
2020-01-06	71.649191	73.153920	71.395555	73.114899	118387200	0.0	0.0	18267.0
2020-01-07	73.124647	73.383158	72.549096	72.771027	108872000	0.0	0.0	18268.0
2020-01-08	72.471046	74.246484	72.471046	73.941635	132079200	0.0	0.0	18269.0
...	...	...	...	...	...	...	...	...
2023-06-26	186.578051	187.796407	184.980202	185.020157	48088700	0.0	0.0	19534.0
2023-06-27	185.639310	188.135938	185.419605	187.806381	50730800	0.0	0.0	19535.0
2023-06-28	187.676554	189.643899	187.347013	188.994781	51216800	0.0	0.0	19536.0
2023-06-29	188.825013	189.813684	188.685203	189.334320	46347300	0.0	0.0	19537.0
2023-06-30	191.371579	194.217727	191.002068	193.708420	85069600	0.0	0.0	19538.0

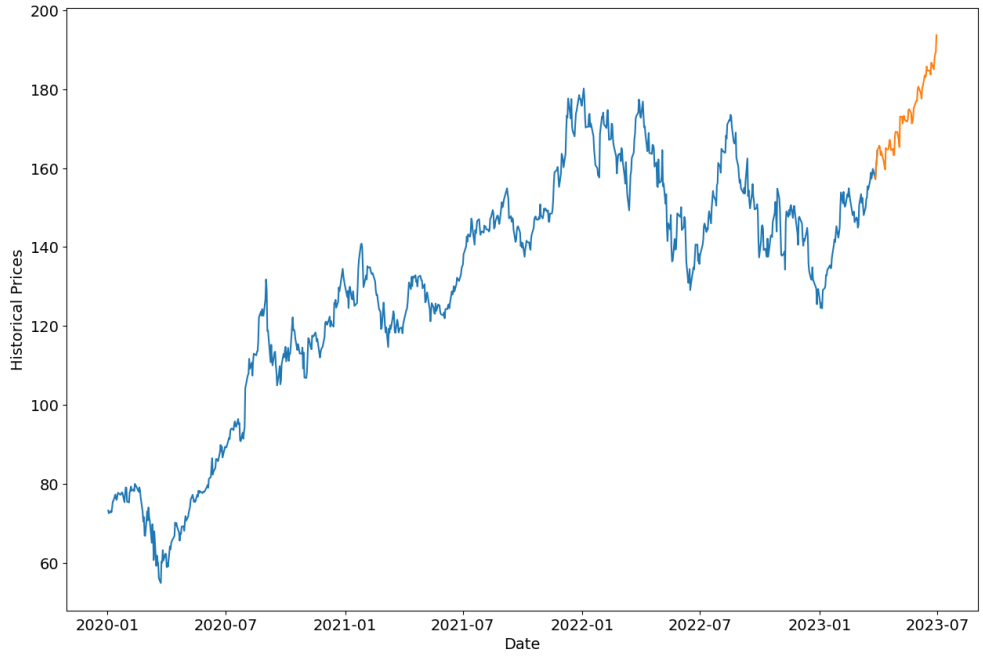
For forecasting purpose, we can use all or only one feature that are provided in the table above. To be clear, we can investigate the following table which contains statistical detail of the mentioned data.

	Open	High	Low	Close	Volume	Dividends	Stock Splits	Date
count	880.000000	880.000000	880.000000	880.000000	8.800000e+02	880.000000	880.000000	880.000000
mean	133.068657	134.725735	131.536134	133.210532	1.051757e+08	0.003469	0.004545	18900.031818
std	31.082760	31.296271	30.887492	31.113751	5.509420e+07	0.027347	0.134840	368.427435
min	55.755904	55.858576	51.974144	54.848965	3.519590e+07	0.000000	0.000000	18263.000000
25%	118.204979	119.695351	116.842187	118.360489	7.045718e+07	0.000000	0.000000	18578.750000
50%	140.465762	142.221930	138.043387	140.520241	8.987460e+07	0.000000	0.000000	18899.500000
75%	154.291590	156.512512	152.641743	154.892838	1.214590e+08	0.000000	0.000000	19219.250000
max	191.371579	194.217727	191.002068	193.708420	4.265100e+08	0.240000	4.000000	19538.000000

In overall, we can see that Open, High, Low, and Close stock price values are close to each other, whether in the max or min values. Even though there is huge gap between max and min values, the mean values are close to the max values which indicates that the values range is not too spread out, which also can be seen by standard deviation values. Hence, we can use only one feature for this analysis. We choose Close stock price for this occasion. The closing price and its trading volume is visualized in the following figure.

FEATURE ENGINEERING

Before evaluating process begin, I split data into train and test set. For this analysis, I set train dataset from January 1, 2020 until March 31, 2023 (818 trading days) and tes dataset from April 1, 2023 onward, which accumulates the second quarter of the year of trading days. The visualisation is provided next.



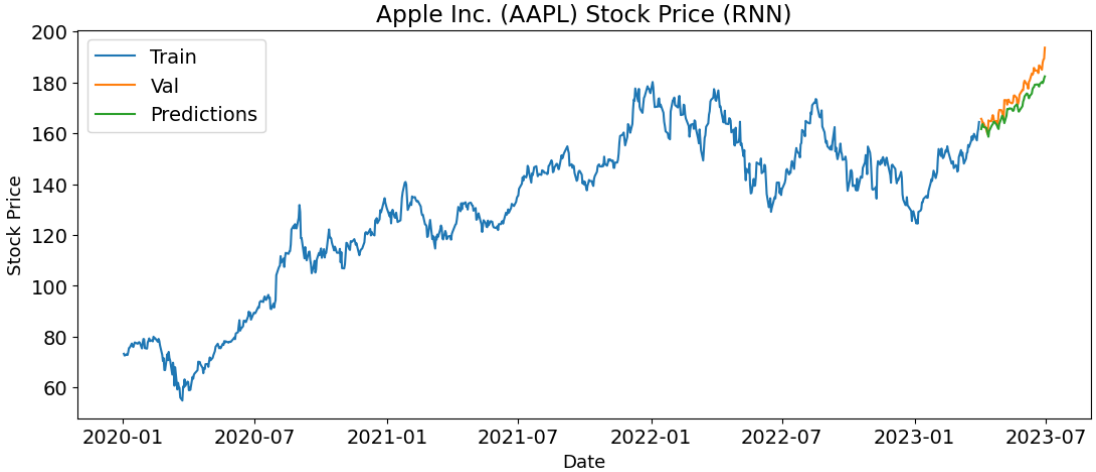
Then, the train and test dataset are converted into array. Next step, the datasets are scaling down using MinMaxScaler available under ScikitLearn. Dataset values will be in between 0 and 1 after scaling. From train dataset, we create x_train and y_train. I set steps equal to 5 which means that the 5-day pattern of trading days in a week and the value after that will be recorded as independent and dependent variables respectively. I also build x_test and y_test from test dataset which will be used for the model's evaluation. I will predict the values based on the x_test and then compare them to the original y_test value.

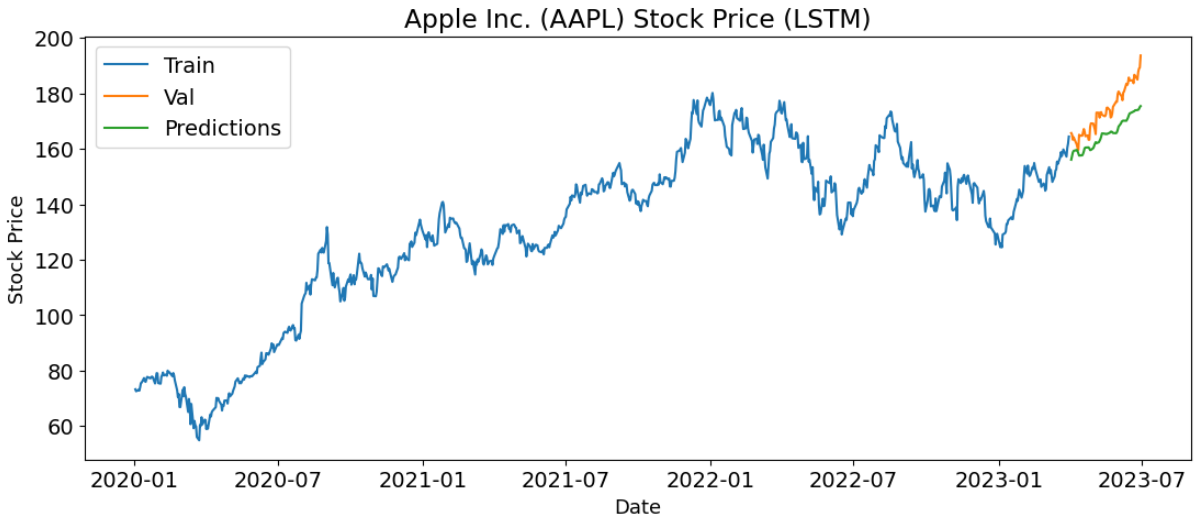
MODEL TRAINING AND FINDING SUMMARY

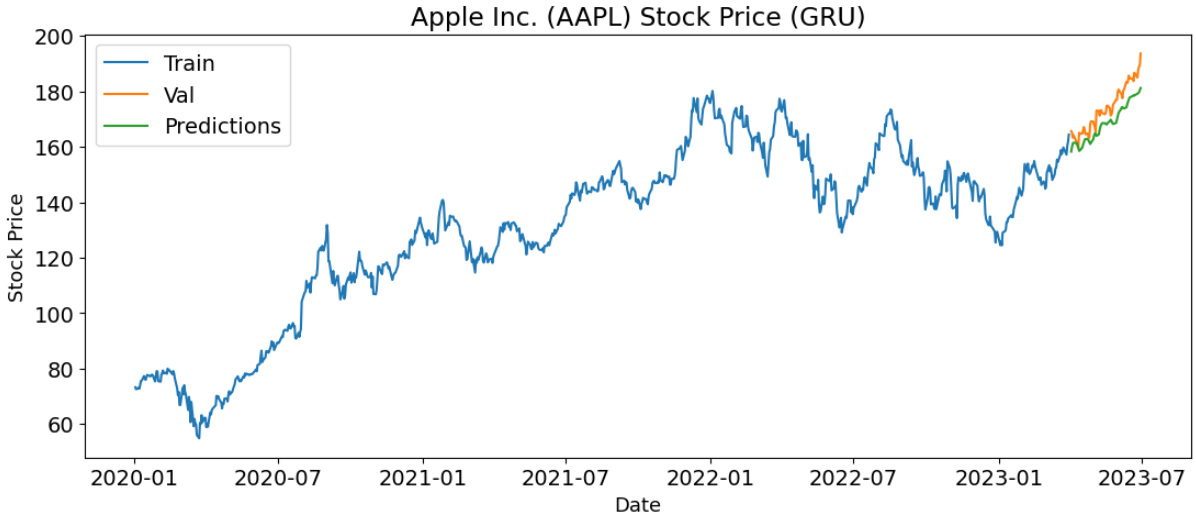
We choose three models known to be used for sequential data for this analysis; RNN, LSTM, and GRU. Basically, the hyperparameters that are set to build the three models are almost identical. For compilation, I have used Adam and for loss function Mean Squared Error. To investigate which is the best model, I used RMSE values as the comparison.

Model	RMSE Value
SimpleRNN	11.43951958782879
LSTM	11.72281350220954
GRU	12.809379624859286

From the table above, we can conclude that simpleRNN is the best model to be used to forecast the future values for this dataset, as it has the lowest value of RMSE.







I found that for less than 1000 data points, RNN become the most fitted model to be used to predict the future values, despite this model suffers from vanishing gradient problem which hinders the model from using long term information. For dataset consisted of thousand data points, LSTM is found to more effective, and sometime, if the data is large enough but less expressive, GRU can be the most fitted model for forecasting.

For the next analysis, we can use more data points, such as include some years data pre-pandemic. Also, we can have experiment in training modela with different values of hypermeters for each model to find a or some better model or models to represent this kind of data in order to produce forecasting tool for investors and traders.