Machine Learning Engineer Nanodegree

Capstone Proposal

Santhosh Kumar Rathode
June 26, 2020

Stock Price Prediction using Machine Learning

Domain Background

In Today's market investing in a stock is a biggest challenge to the investors. Stocks prices are unpredictable, it depends on various factors such as social media, political, financial, crisis, global relationship, demand, president and many more. Buying a stock means buying a part of the company. Making money or losing money depends on stock price up and down, so predicting price of the stock is the biggest achievement for the investors.

Number of companies from Small to large businesses are trading publicly on the world stock market with billions of investors. Platform is so simple that anyone can choose any stock exchange and invest with min amount. In order to get more profit out of stock market one should play smartly with good strategy. That can be achieved by building best Machine Learning models to predict the future trend by using past historical data and event driven indicators.

Herbart Alexander Simon says that "Machine learning is concerned with computer programs that automatically improve their performance through experience." The field of prediction is primarily focused on minimizing the errors and giving more accurate predictions.

Motivation

As like everyone, I also started investing in stock market and still I am not clear that how stock market is going up and down. Once I took this Machine Learning Engineer program I started analyzing data, models and methodologies, then I am able to understand. For this Nanodegree project I am choosing the project of my interest, so I can show more interest towards my ML program.

Problem Statement

Prediction stock market is not that easy and at the same time it's not that tough. Stock market is a complex and forecasting is characterized by data intensity, events, hidden relationships and nature of the market. Predicting fluctuating price with time in stock market is quite difficult.

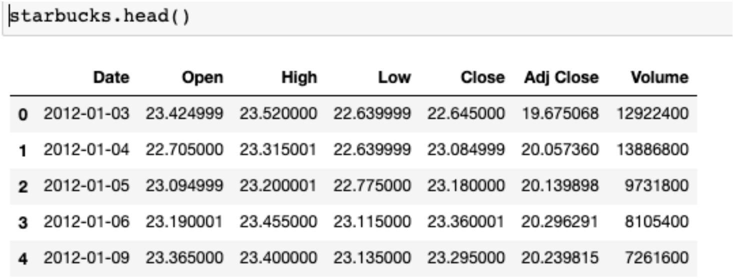
Prediction in investment and trading is ongoing research, as part of this Nanodegree capstone project I am going to predict the stock price by using well-known machine learning models. I will be considering past 8 years of time-series historical data for selected stock.

Datasets and Inputs

Stock market historical data can be found from many places, but I choose to download from https://finance.yahoo.com/. There are many web scraping API's available to get the data based on selected dates. I downloaded CSV files, I am going to analyze 8 years of historical data from 2011-12-30 to 2019-12-30.

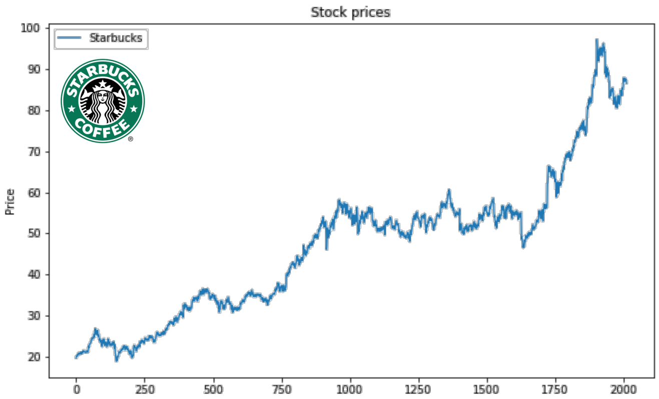
For the stock prediction I am selecting two stocks from two different sectors one from Coffee house and another from Telecommunications.

1. Starbucks (SBUX)
   Sample data:

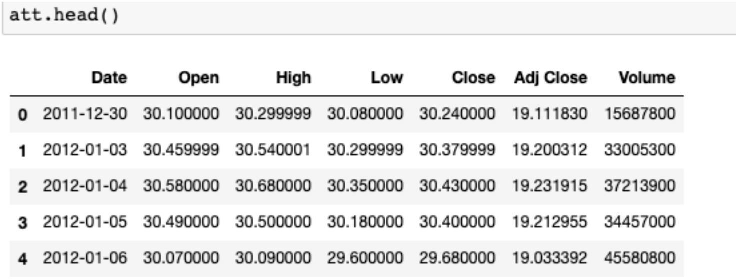


```
starbucks.head()
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2012-01-03	23.424999	23.520000	22.639999	22.645000	19.675068	12922400
1	2012-01-04	22.705000	23.315001	22.639999	23.084999	20.057360	13886800
2	2012-01-05	23.094999	23.200001	22.775000	23.180000	20.139898	9731800
3	2012-01-06	23.190001	23.455000	23.115000	23.360001	20.296291	8105400
4	2012-01-09	23.365000	23.400000	23.135000	23.295000	20.239815	7261600

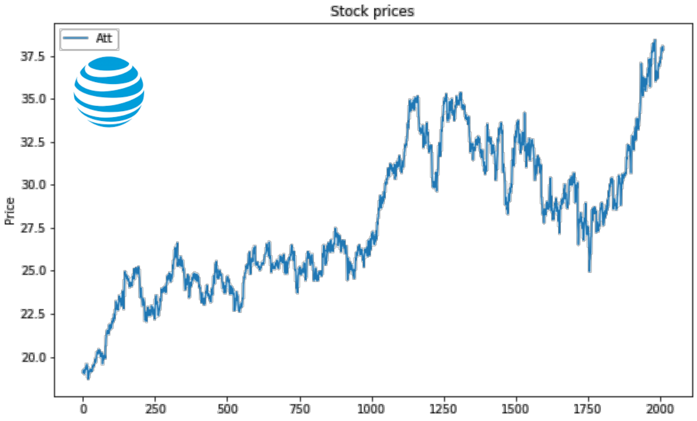


2. ATT (T)
   Sample data:



```
att.head()
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2011-12-30	30.100000	30.299999	30.080000	30.240000	19.111830	15687800
1	2012-01-03	30.459999	30.540001	30.299999	30.379999	19.200312	33005300
2	2012-01-04	30.580000	30.680000	30.350000	30.430000	19.231915	37213900
3	2012-01-05	30.490000	30.500000	30.180000	30.400000	19.212955	34457000
4	2012-01-06	30.070000	30.090000	29.600000	29.680000	19.033392	45580800



This time series data will be split in training, testing and validation set.

Dataset Characteristics

Columns	**Description**
Date	Calendar date of the trading day
Open	Opining price of the trading day
High	Highest price of the stock traded during the day
Low	Lowest price of the stock traded during the day
Close	Closing price of the trading day
Adj Close	Adjusted closing price of the trading day
Volume	Number of shares traded in exchange during the the day

Solution Statement

There are number of algorithms/models to predict time-series based problems. In this project I don't want to jump to top model with all possible indicators to predict stocks. As a beginner I am going to complete project by applying simple supervised learning methods, which I learned from this Nanodegree course. This time-series based project is the regression problem and proposed solution involved applying Long Short-Term Memory (LSTM) from recurrent neural network architecture. Decision Tree Regression model for the benchmarking.

For the model I am going to use 8 years of historical data with few technical indicators such as moving average convergence-divergence (MACD), relative strength index (RSI) and simple moving average (SMA).
I will compare both the benchmark and solution models and provide more insights.

Optional: As of now I am thinking of these models to implement for this project but once I start implementing I might think of better models and better technical indicators to use for this project.

Benchmark Model

After going through lot of materials I have decided to use Decision Tree Regressor as my benchmark model with default hyperparameters. At the end I will compare the metrics with solution models.

Evaluation Metrics:

I am going to use two most popular metrics methods.

1. Root Mean Square Error (RMSE)
2. R-Squared (R2)

- RMSE is popular and it is mostly used for numerical predictions. It compares a predicted value and known value.

$$RMSE = \sqrt{(f-o)^2}$$

  f = forecast (expected values or unknown results)
  o = observed values (known results)

- R2 is a statistical measurement of how close the data are to the fitted in regression line. It means it can provide how much variation in the dependent variable can be explained by the variation in the independent variables.

I will apply both the metrics to the benchmark model and solution models and I will decide the model which has lower RMSE and higher R2 score.

Project Design

Project will be completed on the Jupyter Notebook with python language. I will be importing libraries project such as pandas, numpy, scikit-learn and as required. I will be presenting project code and more insight in final report.

Work flow involves:

- Data collection
- Data exploration
- Data preprocessing
- Data visualization
- Feature selection
- Normalization
- Split the data into Train, Test and Validation sets
- Implementing model prediction and evaluation
- Finally, compare the result with benchmark model.

References

https://www.diva-portal.org/smash/get/diva2:1019373/FULLTEXT01.pdf
https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/
https://machinelearningmastery.com/multi-step-time-series-forecasting-long-short-term-memory-networks-python/
https://en.wikipedia.org/wiki/Stock_market
https://finance.yahoo.com/
https://blog.quantinsti.com/machine-learning-trading-predict-stock-prices-regression/
https://www.youtube.com/watch?v=JuLCL3wCEAk&feature=youtu.be