STOCK ANALYSIS AND PREDICTION

Sukh Gill

BRAINSTATION FINAL REPORT

Contents

Project Goal

The goal of this project is to analyze and predict stock prices and their movements. I was fascinated with an idea that a computer program in form of a machine learning algorithm can perform such complex and computational task on its own. But the main inspiration behind this project is to compare the performance of a machine generated predictions with that of stock analysts and hedge fund managers, as facts suggest that financial specialists are struggling to beat the market from past 10-15 years.

Why Oil & Gas Stock Prediction?

I have been working in financial industry from past four years in Alberta, Canada. As an oil-based economy Alberta has suffered most than any other province in Canada in past seven years. I decided to look at big five oil companies in North America to predict the future outcomes and see if Alberta will bank upon oil industry in future or there will be shift in trend. Stock prediction is one of quite common type of time-series analysis. But I believe most of the solutions do not take in account the financial jargons or technical analysis which makes the results quite unreliable.

Data Source

Yahoo finance is hub of all the stock prices. I was able to find stock data of all the companies for more than 25 years. For this project I chose a past five-year window to identify patterns. Following is the snapshot of **ExxonMobil** data acquired from Yahoo finance:

Date	Open	High	Low	Close	Adj Close
2016-05-09	88.6	88.76	87.23	88.57	69.38658
2016-05-10	88.9	90	88.82	89.99	70.49903
2016-05-11	89.16	89.68	88.72	88.81	70.15933
2016-05-12	89.8	89.93	89.15	89.67	70.83873
2016-05-13	89.28	89.66	88.51	88.66	70.04083
2016-05-16	89.2	89.67	89	89.57	70.75971
2016-05-17	89.27	89.91	89.16	89.53	70.72812
2016-05-18	89.8	90	88.65	89.35	70.58593
2016-05-19	88.61	90.16	88.45	90.11	71.18632
2016-05-20	90.39	90.4	89.38	89.74	70.89403
2016-05-23	89	89.9	88.82	89.6	70.78342
2016-05-24	89.89	90.4	89.55	89.67	70.83873
2016-05-25	89.85	90.37	89.59	90.26	71.30483
2016-05-26	90.25	90.46	89.27	89.8	70.94143

Date tells the date of record.

Open is the price of stock when market opens.

High is the highest stock value on that day.

Low is the lowest stock value on that day.

Close is the final value of stock at time of closing.

Adj Close is closing price adjusted after any dividend payable by companies.

EDA Process

This analysis required some technical financial indicators, i.e., calculations based on stock price and volume of stock traded. As the nature of stock price itself is random and sometimes volatile, it is a good idea to use technical indicators to predict at least which direction stocks are about to move. After considering group of indicators, I decide to include following four for my analysis:

- OBV (On-Balance Volume)
- MACD(Moving Average Convergence and Divergence)
- RSI (Relative Strength Index)
- ADX(Average Directional Movement Index)

The calculations for these indicators were done in excel file to permanently store them in csv file format for later use in jupyter notebooks. Following is the schema after including the technical indicators:

Date	Open	High	Low	Close	Adj Close	Volume	OBV	TR	+DM 1	-DM 1	TR14	DI 14 Diff	DI 14 Sum	DX	ADX	12_day_EMA	30_day_EMA	MACD	Change	Gain	Loss	Avg Gain	Avg Loss	RS	RSI
2016-05-09	101.62	101.8	99.08	100.35	80.79468	9342100	0																		
2016-05-10	100.64	101.29	100.41	101.27	81.5354	6698100	6698100												0.740722	0.740722	0				
2016-05-11	101.24	101.74	100.15	101.14	81.43075	6850700	-152600	1.589996	0.449997	0									-0.10465	0	0.104652				
2016-05-12	102.15	102.42	101.08	102.12	82.21977	6532400	6379800	1.339996	0.68	0									0.789024	0.789024	0				
2016-05-13	101.63	102.04	100.36	100.74	81.1087	5665200	714600	1.760002	0	0.720001									-1.11108	0	1.111076				
2016-05-16	101.87	102.6	101.4	102.29	82.35666	6779400	7494000	1.86	0.559997	0									1.247963	1.247963	0				
2016-05-17	101.07	102.18	100.34	100.75	81.97423	7028000	466000	1.950005	0	1.060006									-0.38243	0	0.382431				
2016-05-18	100.9	101.27	99.44	100.06	81.41281	5927500	-5461500	1.829995	0	0.899994									-0.56142	0	0.561417				
2016-05-19	99.26	100.08	98.52	99.85	81.24196	6241100	-1.2E+07	1.560005	0	0.920005									-0.17085	0	0.170852				
2016-05-20	100.3	100.4	99.49	99.79	81.19314	4707100	-1.6E+07	0.910004	0.32	0									-0.04882	0	0.048821				
2016-05-23	99.18	100.13	99.07	99.38	80.85954	4231200	-2.1E+07	1.059997	0	0.419998									-0.3336	0	0.333603				
2016-05-24	99.94	100.76	99.62	100.19	81.51859	4383300	-1.6E+07	1.380005	0.630005	0						100.6608			0.65905	0.65905	0				
2016-05-25	100.82	101.98	100.82	101.77	82.80413	5428200	-1.1E+07	1.790001	1.220001	0						100.7792			1.285546	1.285546	0				
2016-05-26	101.85	102.37	100.97	101.5	82.58445	4147200	-1.5E+07	1.400002	0.39	0						100.7983			-0.21968	0	0.219681				
2016-05-27	101.38	102.11	100.87	102.02	83.00754	5008900	-9967700	1.239998	0	0.099998						100.8717			0.423088	0.423088	0	0.367528	0.209467	1.75459	63.69696
2016-05-31	102.15	102.28	100.57	101	82.17764	8141800	-1.8E+07	1.709999	0	0.300003	21.38001	4.087799	43.84457	9.323388		100.7783			-0.8299	0	0.829903	0.314619	0.268745	1.170697	53.93184
2016-06-01	100.24	101.38	99.89	101.13	82.28342	4815200	-1.3E+07	1.489998	0	0.680001	21.28001	6.8267	42.541	16.04734		100.8108			0.105789	0.105789	0	0.322176	0.26127	1.233113	55.21946
2016-06-02	100.4	100.57	99.72	100.53	81.79524	6714100	-2E+07	1.409996	0	0.169998	21.35001	7.335284	36.56247	20.06233		100.6642			-0.48818	0	0.488182	0.265817	0.29614	0.897604	47.30197
2016-06-03	100.65	101.14	100	100.66	81.90102	4957800	-1.5E+07	1.139999	0.569999	0	20.73	4.668854	40.26945	11.59404		100.6567			0.105774	0.105774	0	0.273372	0.216778	1.26107	55.77316
2016-06-06	101.34	102.04	100.93	101.17	82.31505	5048200	-1.0E+07	1.270007	0.900002	0	20.25	3.145598	36.55504	8.584143		100.7402			0.414048	0.414048	0	0.313871	0.216778	0.985501	49.65251

As it is visible that number of columns increased almost three times after calculation of technical indicators. The Blue colored columns are the indicators used for analysis and the preceding yellow columns are used to calculate these.

EDA Steps in Jupyter notebook:

Null values: There are some null values in every dataset at beginning as some calculations required information of price for at least 14 preceding days. We will drop all rows with any null value in them.

Duplicates: There are no duplicate rows and columns found in the dataset.

Redundant columns: Some of the columns we will not use in our modelling, but these columns are useful for future purposes, so we will not drop them, instead we will create a new dataframe that will only have the columns we need.

	Date	Close	Volume	OBV	ADX	MACD	RSI
0	2016-06-20	102.610001	6627000	-728100	21.266787	0.585167	59.549267
1	2016-06-21	103.239998	4938700	4210600	21.120981	0.703833	61.814086
2	2016-06-22	102.290001	4866400	-655800	20.912818	0.763167	59.482833
3	2016-06-23	104.440002	5593600	4937800	20.341569	0.746500	66.725694
4	2016-06-24	101.900002	11481400	-6543600	21.149467	0.654667	52.738121

Now we have our dataset ready to run some models on it.

Modeling

Initially the idea was to use ARIMA and FB Prophet, but the power of neural network convinced me to use LSTM as well.

ARIMA: The results with ARIMA were quite unrealistic as the predictions were almost matching the actual price values as there was some data leakage.

FB Prophet: Prophet is FB's own time series model. This model takes in date and stock price. I got some decent results with Prophet but the customization is somewhat limited in this model. As Prophet calculate values on it is on and will create new columns based on that, as seen below:

	ds	trend	yhat_lower	yhat_upper	trend_lower	trend_upper	additive_terms	additive_terms_lower	additive_terms_upper	weekly	weekly_lower	weekly_upper
1325	2021-08-11	101.339464	91.602437	108.195297	98.732401	103.918911	-1.356407	-1.356407	-1.356407	0.284703	0.284703	0.284703
1326	2021-08-12	101.382019	91.060882	107.726826	98.757588	104.016276	-1.588505	-1.588505	-1.588505	0.246633	0.246633	0.246633
1327	2021-08-13	101.424574	91.523680	107.959323	98.755532	104.105024	-1.721304	-1.721304	-1.721304	0.298289	0.298289	0.298289
1328	2021-08-14	101.467129	89.498771	106.652475	98.715237	104.205943	-2.964914	-2.964914	-2.964914	-0.771704	-0.771704	-0.771704
1329	2021-08-15	101.509684	89.243162	106.668570	98.691689	104.305297	-3.126790	-3.126790	-3.126790	-0.771704	-0.771704	-0.771704

LSTM: Neural network are far more capable in performing complex computations and produce more accurate results. I ran a LSTM (Long-Short Term Memory) neural network which takes in account the actual price or any technical indicator to predict the stock prices. Following are the steps:

- Set a sequence of 60, which means training data is divided into data points which are sequence of 60 past values and use this data as input to first layer, this value can be stock price or any technical indicator such as RSI.
- There are three hidden layers used before reaching the output.
- Output will be stock price, i.e., a single value.
- Optimizer and loss metrics that are used are **Adam** and **RMSE**.
- Model is fitted in the end with epoch of 50 and batch value 32.
- Test data is created which is not seen by our model before.
- Predict the prices based on test data.

Following are results of our model on Chevron stock prices:

i.   First figure shows prediction based on RSI.
ii. Second figure prediction based on Stock Price itself
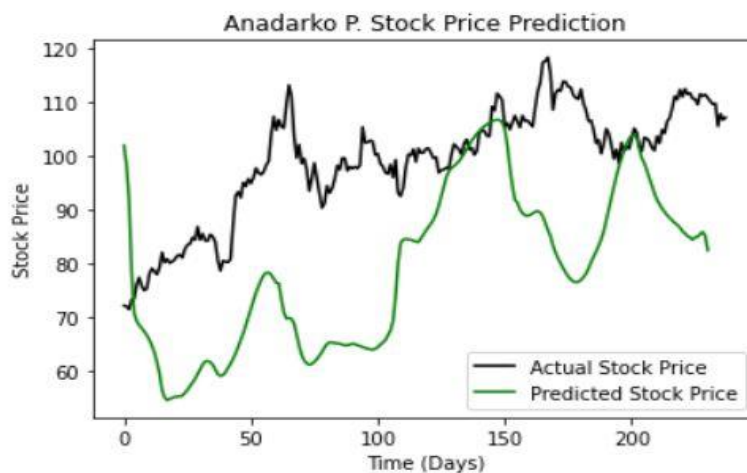
Fig i.

Fig ii.

CHEVRON Stock Price Prediction

Findings

Hypothesis: Goal was to use the historical data of stocks and predict stock prices and their direction with some accuracy.

Results: I anticipated that most predicted price will be way away from actual price. This was the case with some of the results but not all of them. Some results surprised me, especially when we train our model on MACD, a special type of moving average the prices were totally out of pattern with actual prices. This proves that based on our calculation some technical indicators are more accurate than others.

Best Performing Indicator: RSI (Relative Strength Index)

Conclusion

In the nutshell, this project shows the power of machine learning and deep learning to perform complex financial calculations. But the random nature of stock price remains a mystery as there are actually no definite pattern to it. We used technical indicators to predict the direction of stock, in which we were successful but not in all results. We can improve results by following steps:

- Increase the time of dataset, instead of 5 years we can train our data on 20 years of dataset.
- Using all other technical indicators and measuring their performances with each other. This will help in finding out the ones most useful in actual prediction.
- Lastly, we can go with reinforcement learning as an alternative to increase the learning capacity of our predicting algorithms.
- It will be a tough road, but I believe by the end of this decade we will see more automated trading systems than humans.