Machine Learning Engineer Nanodegree

Capstone Project : Using Machine Learning to Predict Saudi Stocks

Ageel Assif December 21th, 2019

I. Definition

Project Overview

Saudi Arabia supplies the world's fuel by producing over 10 million barrels of oil per day, this makes it one of the wealthiest countries in the Middle East and a very fertile ground for investment. I have been investing in the Saudi Stock market (TASI index) for a long time and I am always astonished at how interconnected the Saudi Market with forign markets and world political events. The markets are all interconnected as the recent US & Chain trade war demonstrated. When the US slowed investments in China, the chinese demand for Oil was also reduced, affecting the global Oil prices and circling back to Saudi economy and adversely affecting the Saudi Market.

In this project I hope to use Machine Learning to understand how the Saudi Market is integrated with the world and be able to read the top influencers to make better investments choices. Thankfully the market data is freely available online, I have selected to use (Investing.com,Yahoo Finance and Quandl.com) for the project.

Problem Statement

In this project , I plan to apply Machine Learning to discover how the Saudi Stock Market (TASI) is influenced by a wide range of data from Gold , Oil and other global market indexes to then try to predict how the TASI index will perform.

To accomplish this we do the following:

1. Gather 10 year data for the Saudi Stock Index (TASI) and a multitude of other indices and other possible influences like the price of Oil and Gold. This data is freely available from the following resources [Investing.com, Yahoo Finance & Quandl.com]
2. Merge the data into one large dataset , clean then normalize it
3. Analyze the data and select the best features
4. Train a linear regression model as baseline
5. Train a LSTM model and benchmark it against our baseline
6. Predict 2019 TASI index and compare it with the actual index

The resulting model is expected to be able to predict how the TASI index will perform.

Metrics

Since we aim for Stock Predictions , the metric for our model will be prediction accuracy and we will measure it using Root Mean Square Error.

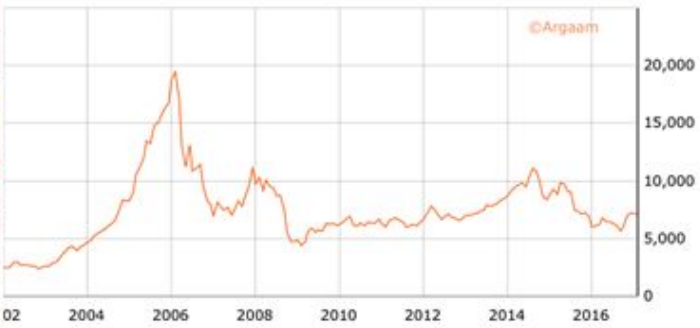
Root Mean Square Error (**RMSE**) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; **RMSE** is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

https://www.statisticshowto.datasciencecentral.com/rmse/

II. Analysis

Data Exploration

Data for various market indexes , currencies , Gold and Oil have been collected from publicly available sources that spans years of data from 2006 - 2019, however I have decided to limit this data to the last ten years of data (01/01/2009 - 01/01/2019) to exclude the TASI crash of 2006. That year would prove to be an outlier since the index rose to a ridiculous historical max (20,000) and dropped hard to where it stabilized till today between 6000-9000.



Saudi Crash of 2006,

- Chart source: https://www.wamda.com/2017/02/saudi-arabia-vcs-opinion
- Image source: http://saudigazette.com.sa/article/523713

Below table lists our initially collected data as 25 CSV files,

Name	Description	Date	Source
TASI	Saudi Stock Index	1/1/2006 - 11/28/2019	investing.com
Saudi SABIC	Local Market - SABIC Leading Chemicals Company in Saudi	1/1/2006 - 11/29/2019	investing.com

(Dropped) Saudi AlAhly Bank	Local Market - Leading Saudi Bank, dropped because it was only listed from 2014 onwards and considered an outlier in our dataset	11/13/2014 - 12/01/2019	investing.com
Saudi AlRajhi Bank	Local Market - Leading Saudi Bank	1/1/2006 - 12/01/2019	investing.com
Saudi Telecom STC	Local Market -The #1 Telecom company in Saudi Arabia	1/1/2006 - 12/01/2019	investing.com
Saudi Electric SCECO	Local Market -The only Power company in Saudi Arabia	1/1/2006 - 12/01/2019	investing.com
Oil	OPEC Crude Oil price	1/2/2003 - 11/28/2019	quandl.com
Gold	Gold prices	1/2/2006 - 11/29/2019	investing.com
Shanghai SSEC	Global Market - Chines Shanghai composite index	1/4/2006 - 11/29/2019	investing.com
Dow Jones	Global Market – the US Dow Jones index	1/3/2006 - 11/29/2019	finance.yahoo.com
Dollar	The US Dollar	1/3/2006 - 11/29/2019	investing.com
NYSE	Global Market – the US New York Stock Exchange	1/3/2006 - 11/29/2019	finance.yahoo.com
NASDAQ	Global Market – The US NASDAQ	1/3/2006 - 11/29/2019	finance.yahoo.com
Nikkei	Global Market – The Japanese Nikkei	1/4/2006 - 11/29/2019	investing.com

(Dropped) SAR vs Dollar	The Saudi Riyal vs the US Dollar This was dropped due to very minimal change over time as the Riyal is pegged to the US Dollar	1/2/2006 - 11/29/2019	investing.com
Karachi	Global Market – The Pakistani Karachi index	1/2/2006 - 11/29/2019	investing.com
Tel Aviv	Global Market – The Israeli Tel Aviv index	1/1/2006 - 11/29/2019	investing.com
FTSE	Global Market – the British Finical Time Stock Exchange	1/3/2006 - 11/29/2019	investing.com
German DAX	Global Market – the German Dax index	1/2/2006 - 11/29/2019	investing.com
EURO vs Dollar	The Euro vs the US Dollar	1/2/2006 - 11/29/2019	investing.com
Dubai Fi Market DFM	Regional Market – the Dubai Finance Market	8/3/2007 - 11/28/2019	investing.com
Gulf bank of Kuwait	Regional Market – the Gulf bank of Kuwait	1/2/2006 - 11/29/2019	investing.com
Egypt EGX	Regional Market – The Egyptian market	1/3/2006 - 12/01/2019	investing.com
Russia RTS	Global Market – The Russian RTS index	1/10/2006 - 11/29/2019	investing.com
Turky BIST	Global Market – The Trukish BIST index	1/2/2006 - 11/29/2019	investing.com

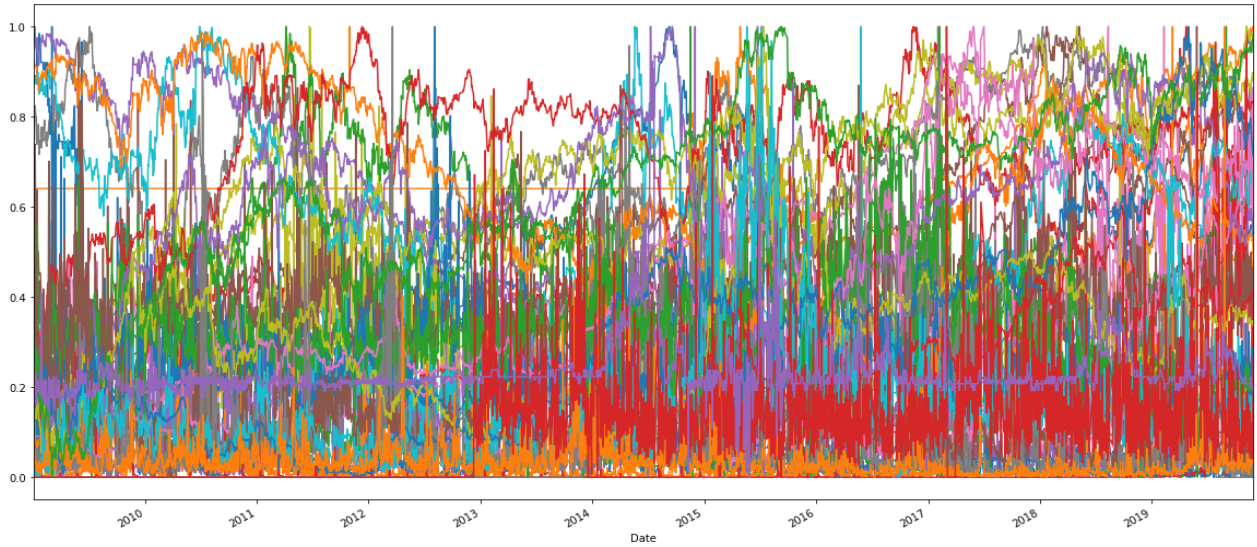
- After further data review, we find that ALAHLY bank (one of the leading banks in Saudi Arabia) only had data from 2014 which makes it an outlier unfortunately and we drop it from our dataset.
- We also drop the USD vs SAR , the Saudi Riyal is pegged to the Dollar and barely changes and has no impact on our prediction
- Out of the available features for each stock , we dropped the Low, High, Closing and Change columns as they are not as important as the actual price and traded volume.

Our data will look something like the below table having all the features neatly aligned with the TASI data.

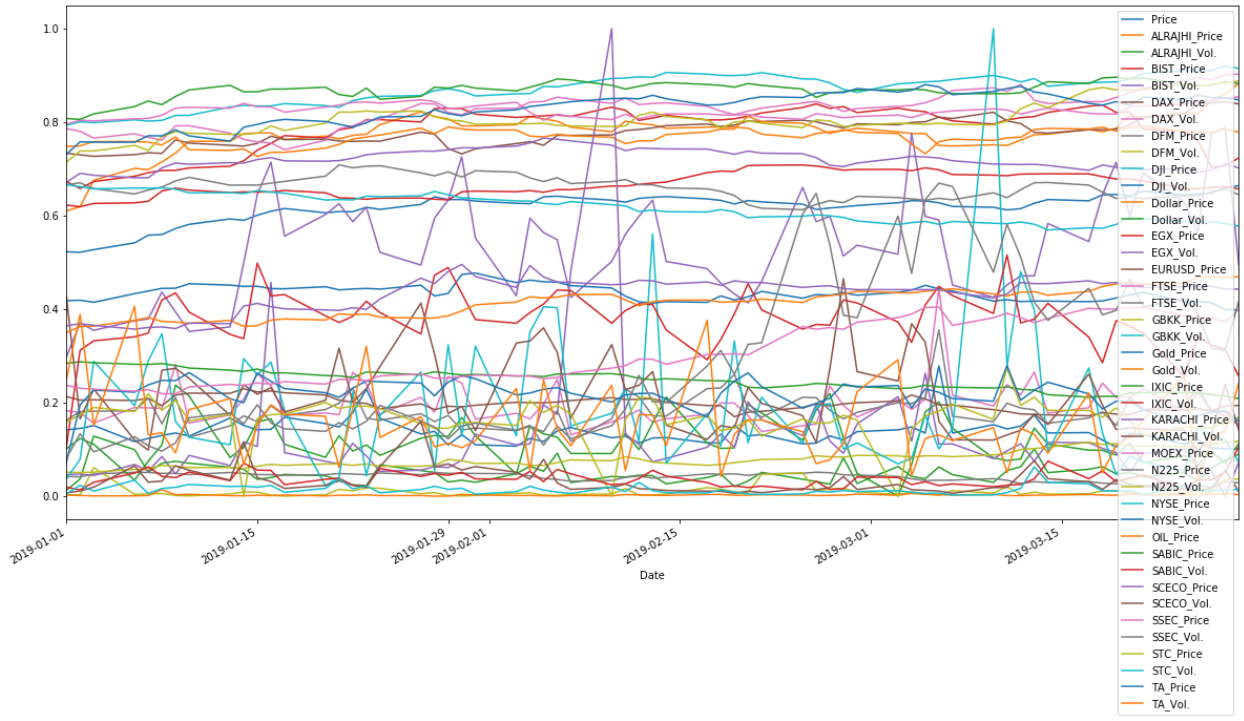
	TASI closing	The Dollar	Dow Jones	Gold	Oil	21 Others...
01/01/2009	8123	98.17	8992.25	878.8	40.44	..
01/02/2009	7803	98.238	9027.129	857.2	43.98	..
01/03/2009	7950	98.176	8954.5	865.4	40.44	..

Exploratory Visualization

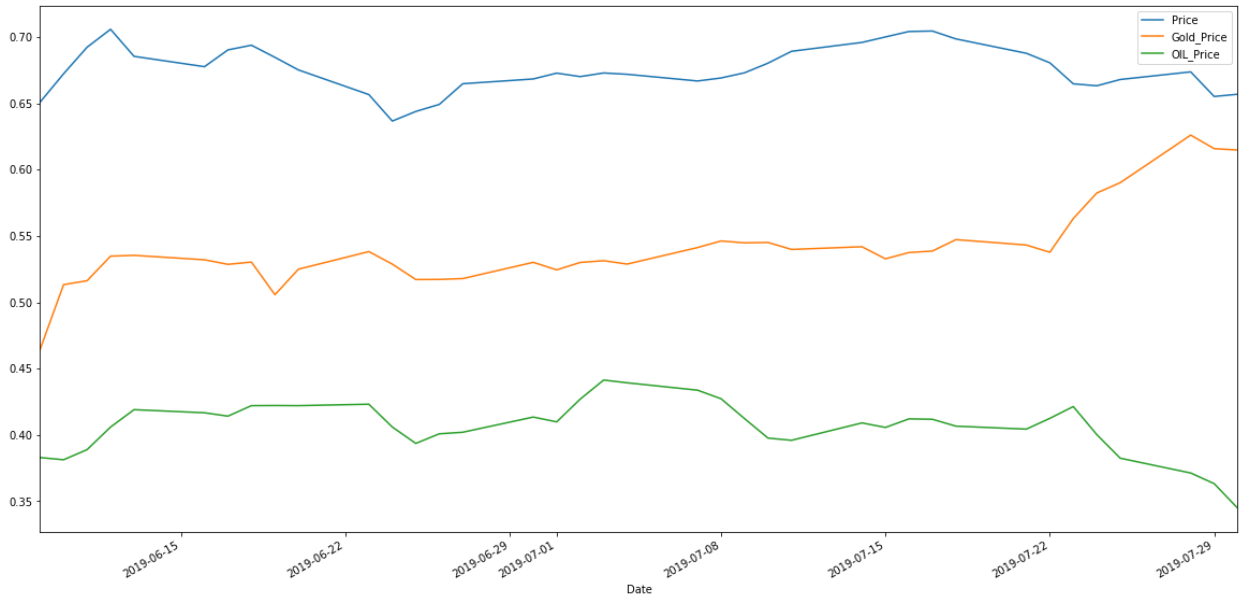
If we plot all the data we get an unreadable Rembrandt painting as it hosts too much data over ten years.



As we zoom in we can slowly start to see the data, below is zoomed to one quarter [Q1 2019] , the graph starts to be readable.



We can also take a closer look at TASI verses select features, in this example we compare TASI vs Gold vs Oil , we can see that Gold has a bigger effect on the TASI index than Oil , which is surprising since Oil is the main product of Saudi Arabia.



Algorithms and Techniques

We are in pursuit of predicting stock prices spanning a long period of time (time series) where the prediction doesn't always depend on nearest prediction but rather taking into account the full breadth of the data. For this reason I have chosen the Long Short Term Memory (**LSTM**) model which is a type of Recurrent Neural Network (RNN) and it is expected to perform well for time series problems.

Benchmark

For benchmarking we will compare the accuracy from a regression model that will simply try to predict the stock prices based on a linear pattern. We will then compare the linear predictions against the LSTM model and see how big of an improvement is realized using LSTM.

III. Methodology

Data Preprocessing

The data will be formatted in one dataset, where each row will correspond to a day and the respective features values.

1. Combine the data files into one file

   Merging the data files wasn't an easy task, the Saudi market closes on Saudi weekend (Friday - Saturday , note below image where the data skips two days) and thus we have to merge the data and align the dates properly.

Date	Price	Open	High	Low	Vol.	Change %
28-Nov-19	7,859.06	7,853.08	7,889.10	7,823.28	82.00M	0.08%
27-Nov-19	7,853.08	7,877.83	7,927.60	7,825.72	90.16M	-0.31%
26-Nov-19	7,877.83	8,013.69	8,024.69	7,877.83	190.10M	-1.70%
25-Nov-19	8,013.69	7,999.57	8,026.62	7,944.93	104.95M	0.18%
24-Nov-19	7,999.57	8,062.61	8,084.89	7,999.57	82.46M	-0.78%
21-Nov-19	8,062.61	8,054.06	8,062.61	8,030.91	177.28M	0.11%
20-Nov-19	8,054.06	8,045.32	8,054.49	8,018.21	105.73M	0.11%
19-Nov-19	8,045.32	8,000.33	8,045.32	7,993.01	120.00M	0.56%
18-Nov-19	8,000.33	7,927.95	8,000.33	7,933.83	110.37M	0.91%

   TASI.csv file sample

   We start with the TASI file as a base then we loop on the remaining files and we use ***Pandas concat*** function to automate the date alignment and produce one large file with all the prices aligned to the respective TASI date.

2. Drop extra columns , [Open, High , Low and Change %] they add little value compared to main Price and Volume columns.
3. Unify the columns to have the same title (Vol. = Volume ; Price = Opening)
4. Fill the null values with ***interpolate*** function
5. The volume volume gives us an indication of stock and market growth and I decided to keep it in our dataset. However the values are written in financial shorthand instead of numbers, so instead of 1,000,000 it abbreviated as 1 M. A Python function was written to convert this column to a number.

A	B	C	D	E	F	G	H	I	J
Date	Index	ALRAJHI_Price	ALRAJHI_V	BIST_Price	BIST_Vol.	DAX_Price	DAX_Vol.	DFM_Price	DFM_Vol.
11/28/2019	7859.06	63.2	4360000	106903.68	1770000000	13236.38	71750000	0.925	26650000
11/27/2019	7853.08	62.2	5260000	107126.18	1440000000	13245.58	36770000	0.883	5720000
11/26/2019	7877.83	61.8	6980000	105844.17	1750000000	13287.07	67300000	0.878	12560000
11/25/2019	8013.69	62.4	10940000	105983.14	2020000000	13236.42	89580000	0.843	1770000
11/24/2019	7999.57	64.5	6730000	105382.47	2240000000	13246.45	54920000	0.815	938710
11/21/2019	8062.61	64.2	4730000	106588.41	1560000000	13163.88	74720000	0.834	759760
11/20/2019	8054.06	64.9	6790000	106805.2	1770000000	13137.7	74520000	0.835	1160000
11/19/2019	8045.32	64.1	6240000	106785.08	2200000000	13158.14	68740000	0.83	2910000
11/18/2019	8000.33	64.6	7460000	107528.68	1930000000	13221.12	68660000	0.827	952090

The full code for above data processing can be found in the **mixer.py** in the **data** directory along with the raw data files used in this project.

Implementation

After the data is fully processed we split the data into 9 years for training (2009-2018) and one year for testing (2019).

We then train the two models below,

- SKLearn Linear Regression model ,
    - default parameters
- Keras LSTM model , the sequential model will have the below parameters
    - Default activation: hyperbolic tangent (tanh).
    - Input_shape: (1, total_featuers)
    - Loss function mean_squared_error
    - epochs=100

    The results for both models will be evaluated based on Root Mean Squared Error and we finally plot the predictions of both models against the actual values for 2019.

Refinement

Initial run with only 7 years (2009-2016) of training data and try to predict 2019

For Linear Regression	**0.48 RMSE**
For LSTM	**0.44 RMSE** with Epochs set to 10
For LSTM	**0.46 RMSE** with Epochs set to 100

One of the factors for this result was that there was a black out period for two years (2017-2018) which adversely affected our models prediction. As we can see in the below group , the prediction baseline doesn't even align, which suggests that for time series predictions it's better to have a continuous timeline.

The results can be improved by predicting 2017 instead of jumping two years to 2019 or If we expand the training data to 9 years. If we take the latter approach, with 9 years of data , we can observe the below results:

For Linear Regression	**0.12 RMSE**
For LSTM	**0.06 RMSE** with Epochs set to 10
For LSTM	**0.04 RMSE** with Epochs set to 100

This is a big improvement on our initial run.

IV. Results

Model Evaluation and Validation

LSTM Predictions vs Actual Test data (TASI for 2019) **Test Score: 0.04 RMSE**

The resulting prediction for 2019 is pretty good from the point of view of an investor and the graph closely aligns with actual values.

Justification

The regression score of **0.12 RMSE** isn't bad and as an investor I can still use the result graph as an indicator for the market direction rather than value. The LSTM model was a big improvement scoring **0.04 RMSE** , the graph closely matches the actual market graph for 2019 for both value and direction and can be used as an investment tool to aid decision making.

Sklearn Regression Predictions vs Actual Test data (TASI for 2019) **Test Score: 0.12 RMSE**

LSTM Predictions vs Actual Test data (TASI for 2019) **Test Score: 0.04 RMSE**

V. Conclusion

Free-Form Visualization

The predictions graph follows the actual index very closely especially the actual spikes and drops, while there is a variation in the price as a value the more we move in time. however as an investor I can use the predictions graph to aid my investment decision of when to expect a market dip and sell my stocks and when to hold on as the market is expected to recover.

- Green - SKLearn Linear Regression Prediction
- Blue - Actual 2019
- Red - LSTM Prediction

Reflection

In this project have predicted the Saudi Stock Market TASI index for 2019, we accomplished this by following the below steps

1. Analyze possible market influencers
2. Gather daily stock prices from online sources such as investing.com, Yahoo and Quandl.
3. Merge the data with respect to TASI operation
4. Remove outliers and interpolate missing data
5. Visualize the data for further insight
6. Select appropriate model for time series prediction, in our case LSTM
7. Train and evaluate the model
8. Predict 2019 index
9. Plot the results showing actual index vs our LSTM prediction

The results are very satisfactory as our predictions nicely warp around the actual index however the results start to diverge as time goes by.

Improvement

It is possible to improve our prediction accuracy by taking expert business advice on which important business factors to include in our dataset. Further improvement is also possible by utilizing paid services which can give us more granular data to as fine as one minute time series for each feature.

We can also benefit from combining out market numbers with market sentiment. A sentiment overview from global politics and investor confidence can be analyzed from news websites and social media to add another dimension towards our model's accuracy.

Furthermore the landscape is evermore changing and our model needs to be retrained with the introduction of new factors. For example the introduction of the Saudi Aramco IPO was a very disruptive factor, the stock has attracted more investors (local and foreing) and changed the dynamics of the Saudi Market, our model will need to be dynamic and updated frequently to remain relevant.