Machine Learning Nanodegree
Capstone Proposal

Using Machine Learning to Predict Saudi Stock Performance

1st December 2019

Domain Background

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 on how interconnected the Saudi Market with forign markets and world events. Machine Learning can help us understand the weight of these factors to make better investments.

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 global market indexes among others to then try to predict how the TASI index will perform.

Dataset and Inputs

Data for various market indexes , currencies , Gold and Oil have been collected from publicly available sources (see reference) that spans ten years worth of data (01/01/2009 - 01/01/2019).

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

Sample

	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	..

Solution Statement

Since our problem involves data in a time series, I propose to use Long Short Term Memory LSTM because LSTM preserves memory for large dataset in a time series and would perform well for our stock data.

Benchmark Model

For benchmarking, we will try to predict the TASI index for 2019 data, an accuracy within 15% of the actual TASI index would be satisfactory from an investors point of view.

Evaluation Metrics

The evaluation metric would be the **prediction accuracy** of our model to predict the 2019 TASI index value.

Project Design

1. **Data Collection:** The first step would be data collection as data quality is key to make accurate predictions. A wide range of data has been collected representing major Saudi stocks, neighboring Gulf countries indexes, global market indexes, currencies , Gold and Oil prices for a total of 25 features overall.
2. **Data Processing:** The collected data is spread across individual CSV files, we need to first combine the data into one file and match the date to respective values. After our master CSV is assembled, the data has to be adjusted for Saudi Market weekend as trading is suspended. Finally we analyze our features and drop features that have low influence on the Saudi Market.

   The 10 year data will be split into three parts ,
   - 6 years for training data
   - 3 years for validation data
   - 1 year for benchmarking
3. **Train our Model:** We initialize and train our LSTM model with the training data
4. **Evaluating the Model:** We review our models results and tune it to improve prediction accuracy till we reach a satisfactory level.

Note: It is possible to improve our prediction accuracy by taking expert business advice on which 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.

References

The data has been collected from the below publicly available sources,

1. Investing.com (www.investing.com)
2. Yahoo Finance (https://finance.yahoo.com)
3. Quandl.com (www.quandl.com)