Machine Learning Nanodegree

Capstone Proposal

Kundan kumar

March 26, 2018

1 Domain Background

The domain background of this project is used to create machine learning model which can predict the stock price accurately. The model will understand the stock price of the company wisely and will able to predict future value of the companies stock.Financial data are very important to understand the company progress. It will help us to understand will the company will progress or not and helpful in investing money in the companies stocks.

2 Problem Statement

We will predict the stock prices of the companies over the years.Based on the prices we can make a decision whether we need to invest on the stock and which companies is more safer to invest. We will predict the stock prices using RNN network(lstm).

3 Datasets and Inputs

The dataset is taken from kaggle and it's contains 4 files which will help in making predictions:

1. prices .csv contains
   - Raw and daily prices of the stock
   - Data contains from 2010 to 2016.
   - Approx 140 stock splits with time

2. Prices-split-adjusted.csv contains

   - Same as prices.csv but with more adjustments for splits

3. Securities.csv

   - Describes company division on sectors

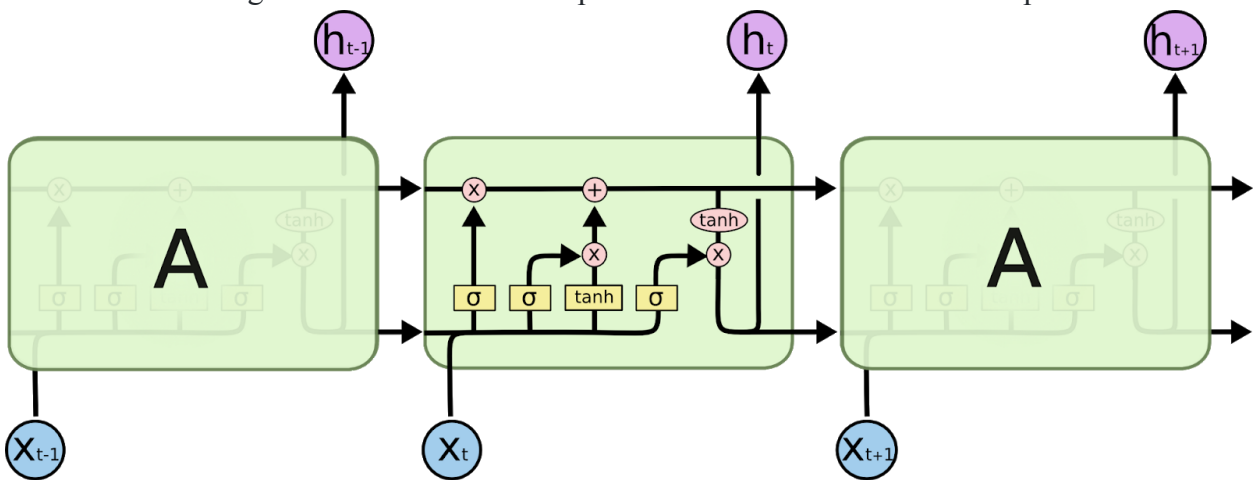
4. Fundamentals.csv

   - Extracted from annual SEC between 2012 to 2016.

The datasets will be extracted from Kaggle here.

4 Solution Statement

The common approach to such problems to use simple regression. Moreover we need to understand which features we need to consider the considers for making the models .We will make the simple feed forward network and check how model is performing and then we will switch the RNN to get the better model. Simple LSTM model for stock market prediction.



Thanks to (http://colah.github.io/posts/2015-08-Understanding-LSTMs/)

5 Benchmark Model

This is kaggle data , so that best kaggle score will be benchmark for test set.We will find the Means squared error, lower the mean Squared error better will be the model.I will also try to use feedforward network and regression model and compare with the lstm model. We will run the regression classifier to get the base MSE. After that we will compare our model and check it beat it and by how much. We will take the best model which satisfy the requirement.

6 Evaluation Metrics

There are several way to predict the model .Some of the common evaluation metrics are :

a. Mean Squared Error
b. R2 Score
c. Mean Absolute Error

We will be using the Means squared error for the evaluation of our models.

7 Project Design

a. Data Preprocessing: We will perform the normalization and scaling on the datasets. Also we will divide the datasets in training, testing and validation sets.
b. Feature Scaling : We will find the relevant features which can be used for making a model.
c. Model Selections: We will perform experiment on various algorithms to find the best algorithm for this case.
d. Model Tuning: We will tune the algorithm to increase the performance and also check whether by increasing the performance may not cause overfitting.
e. Testing : We will test our model by giving testing datasets to know how well model is performing.
f. Visualization: we will visualize the outcome and decide how well the companies is performing.

Tools and Libraries Used:

a. Python & Jupyter Notebook
b. Numpy and Pandas scikitlearn,seaborn,matplotlib
c. Tensorflow,Keras.

Other libraries will be added as per the requirements

References

[1]Kaggle,"New York Stock Exchange": https://www.kaggle.com/dgawlik/nyse

[2]Time Series Analysis:
https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/

[3]RNN : https://en.wikipedia.org/wiki/Recurrent_neural_network
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[4] http://colah.github.io/posts/2015-08-Understanding-LSTMs/