Machine Learning Nanodegree

Capstone Proposal

Kundan kumar

June 25, 2018

1 Domain Background

The stock market is a model which is used to understand the financial behavior of the markets as Economic data is significant to follow the company progress. It helps us to understand whether the company will progress or not and also helpful in investing money in the companies' stocks. It has a huge impact on the economy. Investors lose or gain depends on the fall or rise of Share market price. Investors can make a rough guess by analyzing the stock data, study company history, industry trends, etc. The stock market is highly volatile, and it is difficult to guess the prices of the stocks, but by the recent advancement in the technology, it can make a better prediction and gives a better idea how and where to invest so that we can make an optimal profit with minimal risk. Machine learning models are much efficient for analyzing and predicting the stock prices. 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 the future value of the company's stock. This project uses deep learning for the stock price prediction like Recurrent neural networks (RNN) which is the dominant model for sequential data. A sliding window approach used for predicting future prices. This project uses Long-Short Term Memory (LSTM) deep learning model to predict stock prices. Recurrent neural networks (RNNs) are useful for time series data. Also, this project uses Keras library for building an LSTM to predict stock prices using historical adjusted closing price.

2 Problem Statement

In this project, we are going to predict the future closing value of a given stock over a period (next day) in the future. This project uses LSTM model (Long Short Term Memory networks) to predict the adjusted closing price of the google stock based on the historical datasets. Here, we will be going to predict the next day price of the “GOOGL” stock from New York Stock Exchange Data-Sets.

Main Goals of this Project: -

a. Exploratory analysis on the prices of stocks.
b. Implement Linear regression model and find it’s model accuracy.
c. Implement LSTM model from Keras library.
d. Compare the accuracy results of models.

3 Datasets and Inputs

The dataset imported from Kaggle, and it contains four 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

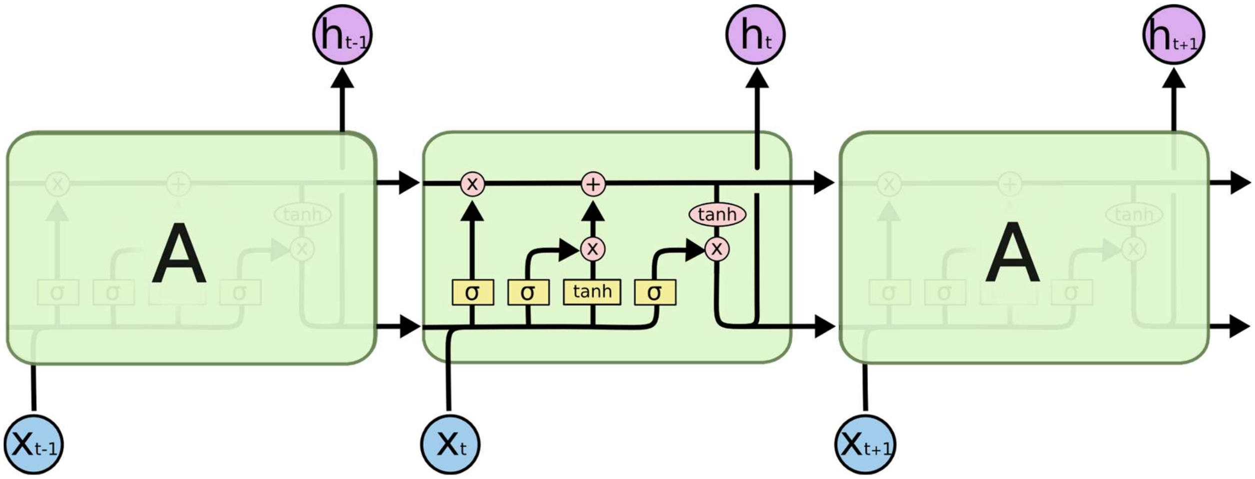
   - Extracted from annual SEC between 2012 to 2016.

The datasets will be extracted from Kaggle here.

From the New York data sets, we are using the daily prices of Google from January 2010 to Dec 2017. These data points are time series data. The goal of the project to predict the next day adjusted closing price after training. There is 501 stock symbols in the datasets. Out of which we are using google stock(GOOGL) for this project. The stock symbol is categorical variable while open, close, low high, and volume is numerical variable.

4 Solution Statement

The standard approach to such problems to use simple regression and check how the model is performing and then we will switch the RNN to get the better model LSTM model for stock market prediction. LSTM model is capable of learning from time series data. Using a Keras implementation of the Tensor Flow library, the solution will utilize an LSTM model and supported by Pandas Data Frame library for convenient time series data schema. The measures of performance based on the predicted stock ticker price in comparison to both the actual cost and the benchmark model's anticipated price. Past adjusted closing stock prices will be features for training the model. The expected output will the predicted next day price of the stock. We are going to predict for the Google (GOOGL) stock from the Newyork datasets. Probably we are using training window around 2000.



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 the benchmark for the test set. We will find the Mean squared error, lower the mean Squared error better will be the model. Regression model will be used as a benchmark 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 by how much it beats. We will take the best model which satisfy the requirement.

6 Evaluation Metrics

There are several ways to predict the model. Some of the standard evaluation metrics are:

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

Mean squared error are used for the evaluation of our models.

7 Project Design

This project will be implemented through the Keras/Tensor Flow library using LSTM Neural Networks.

a. Data Preprocessing: We will perform the normalization and be scaling on the datasets. There will be 80/20 split on training and test data across the models.We will be using time series cross(**sklearn.model_selection.TimeSeriesSplit**)validator for train/test split. These methods return it returns first k folds as train set and the (k+1)th fold as the test set in the kth split of cross-validation. Moreover, I am also exploring the rolling window technique to train/test split.
b. Feature Scaling: We will find the relevant features which can be used for making a model.
c. Model Selections: We will experiment with various algorithms (linear regression, neural net, etc.) 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 improving the performance may not cause overfitting.
e. Testing: We will test our model by giving testing datasets to know how well a model is performing.
f. Visualization: we will visualize the outcome and decide how well the companies are performing

Tools and Libraries Used:

a. Python & Jupyter Notebook
b. Numpy and Pandas scikit-learn, seaborn, matplotlib
c. TensorFlow, Keras.

Other libraries will be used 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/

[5]https://www.researchgate.net/publication/321503983_Stock_price_prediction_using_LSTM_RNN_and_CNN-sliding_window_model

[6] https://www.ijsr.net/archive/v6i4/ART20172755.pdf

[7] https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877