Capstone Proposal: Machine Learning Engineer Nanodegree
Stock Price Indicator

Abhishek Kumar
October 3rd, 2020

Domain Background

Investment firms, hedge funds and even individuals have been using financial models to better understand market behavior and make profitable investments and trades. A wealth of information is available in the form of historical stock prices and company performance data, suitable for machine learning algorithms to process.

In this project, we will build stock price indicator which will take input of last 3 years and try to predict stock price for Next day, after 7 days and after 14 days.

Problem Statement

The primary goal of this project is to take last 3 years of Open, High, Low, Close, etc. data for specific stock and try to predict following:

a) Stock price for next day
b) Stock price for 7 days from today
c) Stock price for 14 days from today

Datasets and Inputs

We will use data from Yahoo! Finance by directly downloading .csv files. I have already downloaded IRCTC data.

In this dataset of IRCTC, we will have last 5yr's data which has 1231 rows and 7 columns (Date and Numerical)

Solution Statement

For this problem, there are multiple methods we can try. I will start with:

- Regression model (Linear regression, XGBoost, KNN)
- Auto ARIMA
  - ARIMA models take into account the past values to predict the future values. There are three important parameters in ARIMA (p, q d). Parameter tuning for ARIMA consumes a lot of time. So, we will use auto ARIMA which automatically selects the best combination of (p, q, d) that provides the least error
- Long Short Term Memory
  - LSTM is a deep learning technique, we will use two layers of LSTM modules, and a dropout layer in-between to avoid over-fitting.

Benchmark model

We will use LSTM as benchmark model and try to achieve similar accuracy using other simplistic mode.

Evaluation Metrics

We will use the root mean square error (RMSE) as evaluation metrics. RMSE is measure of accuracy between actual value and prediction. We will use it to compare above mentioned methods. The lower the value, the better prediction.

Project Design

We will be using the following step as project design:

1. Import necessary libraries (pandas, NumPy, matplotlib, MinMaxScaler, sklearn, GridSearchCV, auto_arima, keras, Sequential, Dense, Dropout, LSTM) and datasets (IRCTC.csv)
2. Explore and pre-process the data
    a. Remove rows having null values
    b. Create extra features like Is_month_start, Is_month_end, Is_Quater_start, Is_quater_end
    c. We will first split data mentioned in below step then scale our data using MinMaxScaler
3. Split data into Train, validate and test
    a. Test Set (Now, Now-1 months)
    b. Validation set (Now-1 months, Now-2 months)
    c. Train set (Now-2 months, Now-2years 2 months)
4. We will multiple predictor model like XGBoost, KNN, LightGBM
5. We will evaluate the model on validation set by comparing rmse value of each build
6. For hyperparameter tuning for the model GridSearchCV will be used

(Repeat step 4-6) until we are satisfied with results

7. Test the model on testing set

References

Original repo for Project - GitHub:
https://docs.google.com/document/d/1ycGeb1QYKATG6jvz74SAMqxrlek9Ed4RYrzWNhWS-0Q/pub

Model Benchmarking:
https://machinelearningmastery.com/how-to-know-if-your-machine-learning-model-has-good-performance/