Capstone Proposal :
Build a Stock Price Indicator
Shu Huang
Mar, 2018

Domain Background

Machine learning for trading is important in stock market. It becomes very useful tools for decision making involved with investment. Although the market is complicated and influenced by many factors and research of machine learning algorithm has been applied to stock market, Kim et al, 2003 has conducted research on prediction of the direction of stock on daily time series using SVM, Tsai and Wang, 2009 research on forecasting of stock price using ensemble learners. In this capstone project, I chose to investigate prediction of stock price using machine learning models.

Problem Statement

The problem I defined to solve is to predict a stock price of 7 days later. I will solve it as a regression problem. The input is one of historical data sets (for example : google's) available from Yahoo Finance, which contains adj-closing price and other important information over the past years and. The reason to choose a short period of time other than a long time later (for example one month later) is that price can be complicated in the long run, any incident can have either a positive or negative on market. The predicted price is numeric value and thus quantifiable and can be measured by metrics. The approach to solve this price prediction problem is also reproducible, though I might just choose one or two stocks as example in this project. It will be generalize to apply to any stock on market.

Datasets and Inputs

The dataset is downloaded from yahoo finance :
https://finance.yahoo.com/quote/GOOG?p=GOOG
Google's historical dataset from 2006 - 2016. There are 2769 rows of data in total. The outcome would be numerical value, which is prediction of stock's close price at 7 days later.
The data is in CSV data format, there are 7 columns in the data set, including :

Date: trading date of the stock
Open : opening price of the stock
High : highest price of the stock in the trading day
Low : lowest price of the stock in the day
Close : closing price of the stock

Adj Close : adjusted price of the stock, it would be different from close price is splitting or dividends happened.

The reason to choose this stock is the historical data is complete and no missing data for a long period of time. Other stocks such as apple, amazon can also be used to solve the problem. I am trying to find a more general solution for price prediction. So the stock selection is kind of random as long as it has a complete dataset.

Solution Statement

I will solve this problem as a regression problem. I am trying to apply machine learning models to training dataset, starting from linear regression model to complex models (KNN, Decision Tree, Random Forest and so on).
First of all , I will fit the data using the models directly without any refinement to get a sense of how appropriate of these models to solve this particular problem.
Second, I will refinement models which do not perform well.
Third, I also will investigate different training size on the performance of prediction. because, for example, decision tree regressor may favor small sample size.
The output should be a model with a good metric score.

Benchmark Model

The benchmark model would be a KNN model , which is trained on first 1000 rows and prediction of test data (1000 :1252 rows).

Evaluation Metrics

I am trying to use mean_absolut_error and R2 score to measure performance of the models. The reason to use them because I approach this problem as a regression problem and not a classification problem. Other metrics such as explained_variance is very similar to R2 score

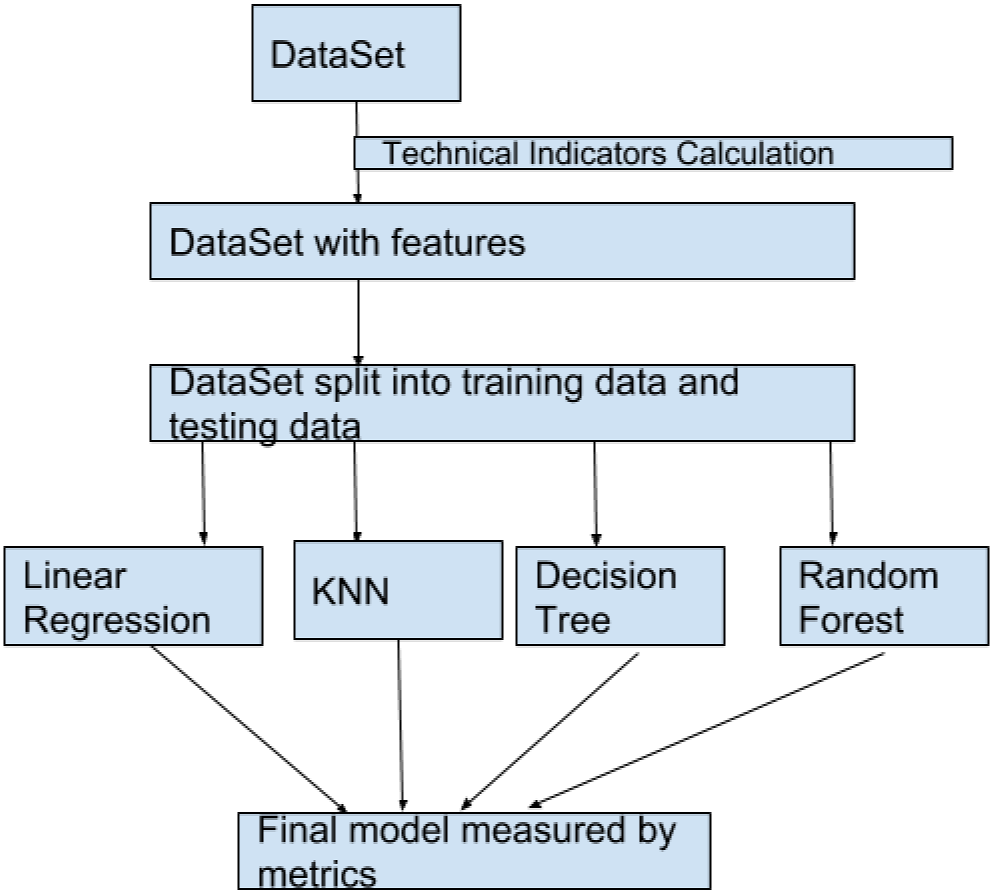
1. mean_absolute_error : mean of the difference between predicted value to the true value
2. R2_score from sklearn.metrics to measure the performance of each model .

R squared measures how close the data is fitted to the regression line. It is coefficient of determination and always between 0 and 1. 0 indicates the model explains none of the variance of data around mean and always predict the y value disregarding the input features and 1 indicates the model explains all the variance of data around mean. In general , the higher r2 coefficient, the better the model fits the data.

Project Design

The project design is pretty straightforward

- preprocess dataset, drop missing data, change data format if necessary
- add more features(technical indicators) for prediction
- once data is ready for training, split manually in time order. Different training dataset is split and test dataset is right after training dataset's date. For example (first 800 rows, the following 252 rows as test dataset)
- fit data using regression models and get initial solution
- refine models by tuning parameters and get refined results and compare with initial solution
- choose a final solution which beat the benchmark and also most generalized model with best performance if there is one.


DataSet
Technical Indicators Calculation
DataSet with features
DataSet split into training data and testing data
Linear Regression
KNN
Decision Tree
Random Forest
Final model measured by metrics

Reference:

1. Kim, K-j., 2003. Financial time series forecasting using support vector machines. Neurocomputing, 55(1), pp. 307-319
2. Tsai, C. & Wang, S., 2009. Stock price forecasting by hybrid machine learning techniques. Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009