Machine Learning Engineer Nanodegree

Capstone Proposal

Chunyan Lei

May 20th, 2020

Domain Background

Quant trading is currently widely used in the finance industry. Among the quantitative strategies in the secondary market, the multi-factor strategy can be described as the earliest created but also one of the most changing investment strategies. A multi-factor model is a financial model that employs multiple factors in its calculations to explain market phenomena and/or equilibrium asset prices. It can be used to analyze the returns of individual securities but also of entire portfolios. Therefore, it is a good model for stock selection and risk prediction, which means its importance in trading and investment.

The traditional multi-factor linear regression model is not stable enough sometimes. However, machine learning has a good performance in prediction and classification. The traditional multi-factor linear regression model also proves that multiple factors are indeed related to stock prices. This relationship is naturally suitable for applying machine learning to stock selection and timing, which makes the possibility of obtaining excess returns higher.

For the factor model, there are many researches of it. Below are some links of the researching papers on it. They mostly only implemented linear regression model. We will try some other machine learning models.

https://www.tandfonline.com/doi/abs/10.1080/13518470110071137

http://cdmd.cnki.com.cn/Article/CDMD-10056-1018059319.htm

https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.1996.tb05202.x

Problem Statement

This is a regression problem. In this project, we aim to explain the stock price with factors using machine learning algorithms according to the following steps.

1. Download data using Quandl package in Python
2. Calculate factors, fundamental and technical factors included
3. Normalize the data
4. Split the dataset into train, validation, test dataset.
5. Implement models
6. Evaluate models

Datasets and Inputs

We collect the daily data about each stock that Dow Jones index measures from May 1^{st} 2018 to May 1^{st} 2020 using Quandl Package, including P/E, P/B, P/S, RoA current ratio and the closing price, then we will calculate some technical factors including MACD, RSI, KDJ, OBV and VR.

Solution Statement

We will analyse all the stocks that Dow Jones index measures. For each stock, we first download and prepare the data as stated above. Then split the dataset into approximately 5:1:1, respectively train, validation and test set. Finally run three machine learning regressors including Supported Vector Machine Regressor, Stochastic Gradient Decent Regressor and Random Forest Regressor, and evaluate the model by calculating MSE. A good regression model explains the most part of the stock price by which we can estimate the stock price with the features we calculate, that is to say it gives us a criterion to select stocks.

Benchmark Model

We will use the Stochastic Gradient Decent Regressor as the benchmark model because it's a simple linear model.

Evaluation Metrics

We will use Mean Squared Error (MSE) as evaluation metrics, which measures the average of the squares of the errors, that is, the average squared difference between the estimated values and the actual value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where Y is the vector of observed values of the variable being predicted, with $\hat{Y}$ being the predicted values.

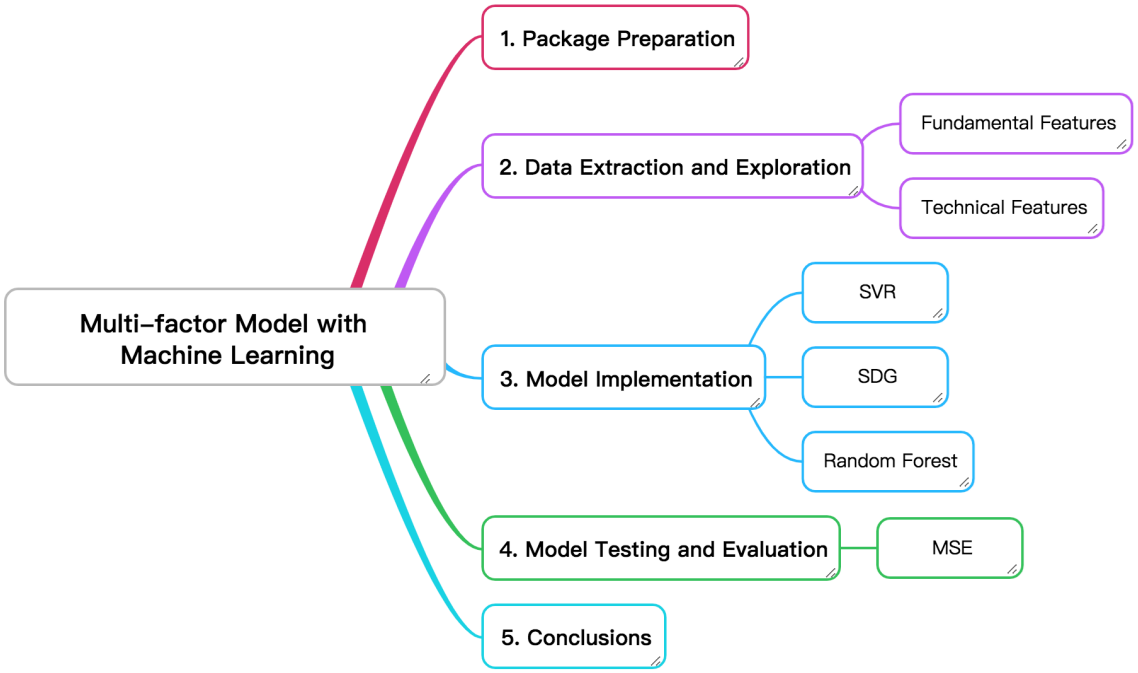
Project Design

Programming language: Python 3.6

Library: Pandas, Numpy, Scikit-Learn, Matplotlib, Quandl, Datatime

Algorithm: Supported Vector Machine Regressor, Stochastic Gradient Decent Regressor and Random Forest Regressor

Workflow:


Multi-factor Model with Machine Learning
1. Package Preparation
2. Data Extraction and Exploration
Fundamental Features
Technical Features
3. Model Implementation
SVR
SDG
Random Forest
4. Model Testing and Evaluation
MSE
5. Conclusions