Optimal Portfolio Construction in Large Cap US Stocks using Machine Learning

Yuezhong Chen, Sheil Sarda

Motivation: *(Briefly introduce the problem you are planning to work on)*

Finance has been seeking to meet the needs of today's globalized modern economies as the stock markets become more and more well-developed after 21 centuries. In order to conduct activities in this complex and changing environment, introducing the Machine Learning Methods to deal with large number of datasets of historical stock prices start to be necessary in selecting the most profitable stocks. We are trying to figure out an effective way in picking the stocks with the comparative high rate of return under the analysis of their previous closed prices. The characters of a growing stock consist of high return, relative fast growth rate and low risk of investment. So, we will also introduce some technical indicators in measuring those corresponding characters.

Data set: *(Briefly describe your proposed dataset (what is n, p, sets of features, ...) Provide a link to the data.)*

https://www.nasdaq.com/market-activity/quotes/historical

Related Work: *(Include at least one citation}(does not have to be a publication, can be a website or blog post) and one paragraph description of prior work related to your project.)*

Literature Review

Harry Markowitz, P. T. (n.d.). *Mean-variance analysis in portfolio choice and capital markets.*

Nandita Dwivedi, M. R. (n.d.). *Predicting Stable Portfolios Using Machine Learning*. Retrieved from https://medium.com/sfu-cspmp/predicting-stable-portfolios-using-machine-learning-f2e27d6dbbec

Tadlaoui, G. (2018). *Intelligent Portfolio Constreuction: Machine-Learning enabled Mean-Variance Optimization.* Retrieved from https://www.imperial.ac.uk/media/imperial-college/faculty-of-natural-sciences/department-of-mathematics/math-finance/Ghali_Tadlaoui_01427211.pdf

Van-Dai Ta, C.-M. L. (2020). Portfolio Optimization-Based Stock Prediction Using Long-Short Term Memory Network in Quantitiative Trading. *Applied Sciences*. Retrieved from https://doi.org/10.3390/app10020437

XingYu Fu, J. D. (2018). A Machine Learning Framework for Stock Selection. Retrieved from https://arxiv.org/pdf/1806.01743.pdf

Problem Formulation: *(Describe how you will frame your problem as a machine learning task.)*

The process of selecting a portfolio is composed of two stages: the first to analyze the historical data and build an idea on the behavior of assets in the future, and the second one uses these insights to build the portfolio. (Markowitz). Our work attempts to combine both Machine Learning and Investment Strategies by using ML to

predict the stock direction in the first phase of the portfolio construction. We aim at comparing the performances of a portfolio constructed with the classic structure with one derived from a machine-learning enabled version.

We choose to work only on US Large Cap stocks for generalization purposes. We aim at forecasting expected returns for a set of stocks. This is done in two steps: we first use a supervised learning algorithm to forecast the direction of the stock, we then forecast the amplitude of the move to capture the volatility of the returns.

Methods:

In order to achieve the optimum expected returns, finance data is a time series data, we first measure technical indicators from historical stock data and produce features and goals from historical stock data.
The features for linear models, xgboost models, random forest and neural network models are pre-processed and prepared. Then to forecast the future price of stocks in the US Large Cap markets, linear models, decision trees, random forests and neural networks are used. Direction of the stock is measured using the supervised classification approach. The logistic regression is used as the baseline model, which is compared with the more complex models e.g. xgboost, random forest and neural network classifiers.

Supervised classification methods:

1. Logistic classifier – The purpose of binary logistic regression is to train a classifier which can make a binary judgement on a new input observation. The sigmoid classifier that assists in making this decision is used.
2. Xgboost classifier - It is based on ensemble learning, which is a subset of Machine learning algorithm that train and predict at once to generate a single better performance with many models. It is based on three approach bagging, stacking and boosting.
3. Random forest classifier - A random forest is a method that suits different sub-samples of the dataset with a number of decision tree classifiers and uses averages to improve predictive accuracy and over-fitting power.
4. Neural Network Classifier - Neural network is an algorithm of supervised learning that learns a function by training on a dataset, with "n" as a number of dimensions as input and "d" as the number of output dimensions. It is distinct from logistic regression, in that there can be one or more non-linear layers, called hidden layers, between the input and the output layer.

Regression methods:

1. Linear Regression – It finds the linear regression line that predicts the expected return of a particular stock. It is a linear approach to modelling the relationship between one or more explanatory variables and a scalar response i.e. dependent variable.
2. Random forest Regression - It is a non-linear approach to modelling the relationship between one or more explanatory variables and a scalar response i.e. dependent variable. It fits a number of decision tree classifiers and averages them to predictive or forecast values.
3. Neural Network Regression – It is a non-linear function approximator for regression, it can be trained provided a set of features.

Evaluation:

The classification models are evaluated using classification accuracy as the evaluation metric. However for evaluating the regression models the following metrics are being used:

1. Mean absolute error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - ypred_i|$$

Where y_i, is the ith observation and ypred is the ith predicted return.

2. Mean Squared error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - ypred_i)^2$$

3. Root Mean Squared error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - ypred_i)^2}$$

Project plan: *(Provide a rough timeline of your project work schedule, including which team members are responsible for what portions of the project.)*

Important dates	**Deliverables**	**Sheil / Yuezhong**
11/16/20	Project Proposal Due	
11/23/20	Aggregate data	Yuezhong
11/27/2020	Pre-process data	Sheil
12/2/2020	Fit regression models (Linear, Random Forest, etc.)	Yuezhong + Sheil
12/4/2020	Fit supervised classification models (Logistic, XGBoost, etc.)	Yuezhong + Sheil
12/9/20	Benchmark model performance and finish project report	Yuezhong + Sheil