Udacity Machine Learning Nanodegree Capstone Project Proposal

Domain Background

I work in an investment firm that manages investment portfolios for institutions and high net-worth individuals. The company has a research department that chooses equity and fixed-income investments for the portfolios of our clients.

I would like to identify economic indicators that drive stock market returns and develop an algorithm based on the indicators to predict market returns to help my company make money. I will be using the S&P 500 as the targeted stock market to research as it is the largest stock market in the world and has the most widely available information.

Problem Statement

I would like to develop an regression algorithm that can predict stock market returns, specifically the S&P 500 index as it is the largest stock market with the most available information of all stock markets.

This problem has previously been researched and the following are academic papers relating to this problem:

1. http://cs229.stanford.edu/proj2017/final-reports/5234854.pdf
2. https://arxiv.org/pdf/1603.00751.pdf
3. https://etd.auburn.edu/xmlui/bitstream/handle/10415/5652/Application%20of%20machine%20learning%20techniques%20for%20stock%20market%20prediction.pdf?sequence=2&isAllowed=y

Datasets and Inputs

To solve this problem, I will use the economic indicators in this Kaggle dataset.

https://www.kaggle.com/robertnolan/economic-indicators

The dataset will form the input of this problem and the output of this problem is an algorithm that has predictive power for the S&P 500 using economic indicators.

There are 14 features in the dataset including the S&P 500 level. The dataset spans from January 2007 to September 2017 with monthly data points, resulting in 129 records.

The market return will be derived in the preprocessing step from the difference between the previous month to the next month and will be expressed in percentage terms.

Solution Statement

The algorithm developed should have a high predictive accuracy used in the validation data set with high r^2 score measured in the validation step. The algorithm I would like to try are linear regression, polynomial regression and multivariate regression.

Benchmark Model

The benchmark model used will a simple linear regression to get the baseline score.

Evaluation Metrics

Evaluation metrics that will be used to evaluate the success of my model are 1) mean squared error, 2) mean absolute error, 3) explained variance score and 4) r^2 score.

http://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics

Project Design

I will run the dataset against multiple supervised learning models including linear regression, polynomial regression, and multivariate regression to decide which model performs the best measure by the evaluation metrics.

These are the following steps I will use in my project

1. Preparation – I will first examine the three academic papers earlier to prepare the guide me through this project
2. Explore the data – I will examine the data to explain each of the predictive variables, and apply feature importance to extract out the most relevant and predictive variables. I will also talk about the targeted variable, market return
3. Prepare the data – I will preprocess the data to examine outliers, normalize numerical features, transform skewed features, and split the data. In this step, I will also preprocess the data to obtain the monthly change in each of the features instead of absolute number. The date range that will be used for my testing is the last 2 years, approximately 20% of the entire dataset, from October 2015 to September 2017. I will use all the data before October 2015 as my training set.

4. Evaluate the model performances – I will apply the data different algorithms to the dataset and evaluate how the models performance based on the evaluation metrics
5. Conclusion – I will make a recommendation as to the best model and summarize the project