Data Mining Project Proposal - Prediction of Stock Prices Using Regression Models

Shimeng Lin 11/01/2017

I. Abstract

Stock prices indicate a company's financial health. A successful prediction of future price will lead to significant profits to stock market investors. In terms of data mining, this is a prediction task using regression techniques to predict future stock prices. My main goal is to select an optimal model that could give the most useful prediction of closing price among models of multiple linear regression, polynomial regression, support vector regression and random forest regression.

II. Data Description

Dataset containing historical stock prices of Edwards Lifesciences Corporation (EW) can be retrieved from Kaggle (*Source: https://www.kaggle.com/aumashe/stock-ew/data).* This dataset was downloaded from Yahoo Finance covering a time span from March 26th 2000 to September 9th 2017. Relevant features include date, open and close price, high and low price and trading volume. The closing price is the target variable to predict and other attributes including computed features such as return, rate of change, Williams %R or relative strength index could be explanatory variables. There are approximately 4,400 records in the Kaggle dataset. Most key features are numerical variables. In data preprocessing, I will identify and clean potential outliers and missing values to minimize their impacts on modeling.

III. Methodology

I will use Python and/or R to process the dataset and build the learning models. Stock prices could be non-linear and continuous, therefore, the regression techniques that I will consider applying are multiple linear regression, polynomial regression, support vector regression and random forest regression. In multiple linear regression models, I will identify the most influential factors that impact the closing price heavily using backward elimination method. An optimal polynomial degree will be decided in the polynomial regression model. In random forest regression model, the optimal number of trees in the forest will be selected as well.

IV. Evaluation

Multiple models will be built using different regression techniques as mentioned above. The dataset will be split into training and test sets. A training set will be used for building models and constructed by randomly selecting 80% of the records. The remaining data points will be allocated to the test set to provide a fair assessment of the models.

To evaluate the performance of the models, I will calculate confusion matrices for each of the regression models and the corresponding scalar measures such as accuracy, precision, recall and F-measure. Those indices could give a general idea about the performance of the models. After comparing the indices, the best model will be selected. Visualization techniques could be used to help compare the performance of the models. Lastly, the paper will discuss any potential overfitting or underfitting problems and other limitations. I will also explain the future research work for improvements.