International Institute of
Information Technology, Bangalore

Project Strategy Document
DS 707 Data Analytics

Exploratory Analytics and Classification

Akanksha Dwivedi - MT2016006
Hitesha Mukherjee - MS2016007
Nayna Jain - MS2017003
Tarini Chandrashekhar - MT2016144

Instructors :
Prof. Ramanathan Chandrashekhar
Prof. Uttam Kumar

November 5, 2017

Contents

1 Data Exploration

1.1 Introduction

Multivariate time series (MTS) data sets are common in many multimedia, medical, process industry and financial applications such as gesture recognition, video sequence matching, EEG/ECG data analysis or prediction of abnormal situation or trend of stock price. MTS data sets are high dimensional as they consist of a series of observations of many variables (multidimendsional variable) at a time. For analysis of MTS data in order to extract knowledge, a compact representation is needed. For feature subset selection for MTS data sets, popular techniques for machine learning or pattern recognition problems are modified.

Any data mining or pattern recognition task such as knowledge/rule extraction, clustering or classification of data is preceeded by data preprocessing. Preprocessing of data is the process in which redundant or irrelevant information from the data is removed while the most discriminatory information is retained to represent the data in a compact manner. This preprocessing stage is often known as feature extraction or feature subset selection. The next step for classification or clustering is to design a similarity measure for identifying similar time series to make clusters or classes or to extract rules.

1.1.1 Time Series Classification

Our data is based on mining Bitcoin and Etherium crypto currencies. Basically our data is a Historical Timeseries data.It has wide variety of features. Time series classification is to build a classification model based on labelled time series and then use the model to predict the label of unlabelled time series. The way for time series classification with R is to extract and build features from time series data first, and then apply existing classification techniques, such as SVM, k-NN, neural networks, regression and decision trees, to the feature set.

1.2 Selecting Appropriate Classification Technique

1.2.1 Supervised versus Unsupervised learning

This is one of the most fundamental distinctions between learning methods. Supervised learning involves developing descriptions from pre-classified set of training examples, where the classifications are assigned by an expert in the problem domain. The aim is to produce descriptions that will accurately classify unseen test examples. In unsupervised learning, no prior classification is provided, and it is up to the learning scheme itself to generate one based on its

analysis of the training data.

We have used Supervised Learning Model for classification of our dataset.In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

1.3 Build Classification Model Parameter Setting

1.3.1 Support Vector Machine for Classification

We have considered the Bitcoin Dataset which has 24 features or attributes in it. We have extracted 17 important features and build a subset of the data. We have further classified our data into training and test data. 70 percentage of data is classified as Training and the rest as testing data. We have used SVM Algorithmn to build the model based on the trainng data and predicted the Market_Price based on Model built and Test Data

1.3.2 Random Forest

Cryptocurrency data is similar to stock analysis data. As discussed in paper by Luckyson, Snehasu, Sudeepa, Random Forest has been used to predict the stock prices. Random Forest is an ensemble learning method for classification and regression by considering lot of decision trees at training time and specifying the class based on the mode of the classes as identified by different trees. Here the Random Forest is applied on bitcoin_dataset to predict the bitcoin market price.

1.3.3 Linear Regression

Linear Regression is used for predictive analysis. In this case, there is a response variable whose outcome has to be predicted based on the input variables which are also called as dependent variables. Linear Regression is used with continuous type of data. We have used Linear Regression to predict the market cap based on Open Price. The mean square error which we got was 0.2127196982. It seems to be doing average estimation

Chart below shows the predicted vs actual value.

1.4 Visualizing Using Tableau

1.5 Assessing the Classification Model Built

References

[1] Predicting the direction of stock market prices using random forest. Luckyson Khaidem Snehanshu Saha Sudeepa Roy Dey. khaidem90@gmail.com snehanshusaha@pes.edu sudeepar@pes.edu

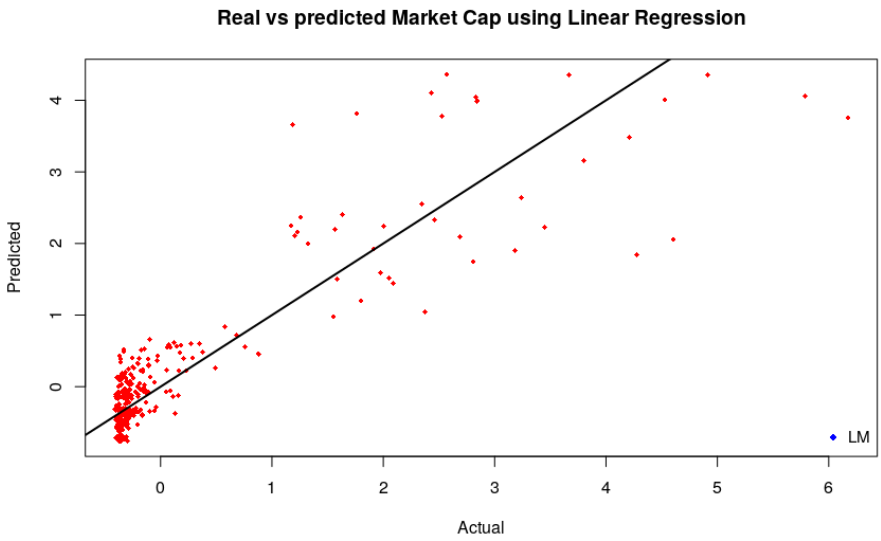

Figure 1: Bitcoin Market Cap Prediction based on Open Price