Machine Learning Engineer Nanodegree

Capstone Proposal

Sai

March 25, 2018

Proposal

Domain Background

Cryptocurrency is a digital currency in which encryption techniques are used to regulate the generation of units of currency and verify the transfer of funds, operating independently of a central bank. There is lot of buzz about cryptocurrency and everyday in the news we hear volatility in the prices of cryptocurrencies such as Bitcoin.

There is already some research done in the area. Issac Madan, Shourya Saluja and Aojia Zhao from Stanford university has published research paper on automated bitcoin trading engine using Binomial GLM, SVM and Random forest which is detailed in the following link.

https://pdfs.semanticscholar.org/e065/3631b4a476abf5276a264f6bbff40b132061.pdf

Problem Statement

The extremely nonlinear nature of the crypto market data makes it very difficult to design a system that can predict the future direction of the crypto prices in Bitcoin with sufficient accuracy. Goal of the project is to predict price movements(that's either up or down) of cryptocurrencies such as Bitcoin. As problem involves predicting and classifying Bitcoin prices to rise or fall this is a binary classification problem.

Datasets and Inputs

Bitcoin is actual implementation of decentralization issued under the consent of participants not the central bank. Hence variants like purchasing power or interest rate parity does not impact Bitcoin. Bitcoin price is largly impacted by demand and supply.16.5 million Bitcoins are mined since it was created and the Bitcoins are capped at 21 millions part of the reason for scarcity and rush.

Hence to predict Bitcoin price we need to predict Demand and Supply. Blockchain is the technology used to create Bitcoin."Block Chain information" along with "Bitcoin" trading information includes features that can determine demand and supply and hence the price.

Therefore, the project uses historic blockchain information and Bitcoin trading information to predict future Bitcoin prices.

This project uses data from https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory which is publicly available. The data consist of "Bitcoin_Dataset" which consist of historic Blockchain information and Bitcoin prices for last 4 years.

Bitcoin dataset consist of several important blockchain information that might help us understand rise and fall continuum , most importantly, Total bitcoins, Trade volume, block size, avg block size, Orphaned blocks, transactions, hashrate, Mining difficulty and Miners revenue.

We have a balanced dataset with around 47% of records with rising prices and 43% of records with falling prices and 10% of records with no price change.

Solution Statement

The main goal of speculating a cryptocurrency price is to make profitable trades. To make profitable trades one doesn't need to know accurate value of the cryptocurrency price, but one merely needs to predict whether prices rise or fall. In these lines we will train our models to classify subsequent day's closing price to be likely higher and lower than the last, based on past 10 days price data.

The solution tests the data on selected features using the models learnt in this course(Logistic regression, SVM and Random Forest) to pick a best model.

Benchmark Model

We will start with building model that would get results better than Naive prediction. We will derive Naive prediction based on a Naive model that always estimates prices to go up.

Evaluation Metrics

The solution will be evaluated based on Accuracy, Precision and time taken to run. Accuracy will help us evaluate how often model makes right prediction and is derived by calculating number of correct predictions by total number of predictions
Precision tells us what proportion of predictions made toward rise in bitcoin prices resulted in rise. It is derived by calculating the ratio of true positives to all positives. Precision is an important metric in Bitcoin prediction as knowing false positives is important in that false positives will result in losses in the trading.

Both Accuracy and prediction will be measured against the time to choose a mode that performs relatively faster at higher prediction and accuracy.

Project Design

Data Exploration: Data Exploration will involve exploring the data through visualizations and code to understand how each feature is related to the others. We will observe a statistical description of the dataset, consider the relevance of each feature, and select a few sample data points from the dataset which we will track through the course of this project.

Feature Relevance and selection: Feature relevance will involve identifying relevance of features and selecting the features by determining their correlation coefficient and building scatter matrix.

Feature Scaling: In feature scaling data is evaluated for normal distribution. In case the data is not normally distributed, most likely with financial data, non linear scaling is applied by applying natural logarithm.

Outlier Detection: Detecting outliers is very important as the outliers can skew the results. Outlier steps are calculated based on 1.5 times the interquartile range. Any data point outside the outlier step will be identified as outlier.

Feature Transformation: Feature transformation involves using Principal component analysis(PCA) to draw conclusions about the underlying structure of the data. Dimensionality analysis is performed to identify least number of dimensions needed to explain most of the variance in the dataset.

Naive predictor performance: We will build a naive predictor that will always predicts Bitcoin prices to go up and will derive accracty for this model. We will later evaluate the model we derive against Naive accuracy.

Creating a training and predicting pipeline: We will be training and testing the models on time series.At the initial training step the machine is learned with N(train) training data, and the prediction performance is measured using N(test) test data. Next, after t’−t time from time t, the machine is trained using again the N(train) data from time t’ to update old learning data, and the performance of N(test) test data is thereafter measured. The machine is trained through the entire range in this way and average accuracy is derived from all the time series.

Model Evaluation: Model performance of (logistic regression,SVM, Random forest) are evaluated against the time and accuracy.

Optimization: Based on the accuracy on training and testing dataset and the time they took to run, models will be optimized using various optimization techniques until desired accuracy is above naive prediction.