Machine Learning Engineer Nanodegree

Capstone Proposal

Jose' Teodoro da Silva
April 12st, 2018

Proposal: Trading cryptocurrencies with Machine Learning

Domain Background

Nos ultimos anos temos observado a explosao das criptomoedas, moedas virtuais construidas sobre complexos algoritmos criptograficos. Atualmente existem centenas de criptomoedas sendo negociadas pelo mundo todo. Essas negociacoes sao realizadas em casas de cambio denominadas de exchanges. As exchanges oferecem o servico de compra, venda e troca de criptomoedas, conversao dessas moedas para dolar ou mesmo real.
Apesar de toda atencao que o Bitcoin e o Ethereum vem recebendo, esse projeto tem por objetivo a negociacao de moedas nao tao conhecidas. Alem das moedas mais conhecidas como Bitcoin, Ethereum e Riple, temos outras menos conhecidas como 1ST, 2GIVE, ADA, BITB, DOGE, MANA, RDD, VOX, XDN, XLM e XVG. A vantagem dessas moedas menores e' seu preco baixo e possibilidade de grandes ganhos. Quando essas moedas menores aumentam de valor, esse aumento pode variar entre 30% a 250% em poucas horas. Assim como essas moedas aumentam o preco, elas tambem caem em questao de horas. Para referencias, vide o site https://coinmarketcap.com/ (https://coinmarketcap.com/) e o historico de crescimento das moedas citadas.

Problem Statement

Este projeto trata sobre a identificacao do momento para compra de uma dessas moedas antes ou no inicio de um aumento de precos acentuado. O momento em que a compra seja propricia para que realizar lucro revendendo aquela moeda em momento posterior. Neste sentido, o momento para venda, para realizacao do lucro, nao sera tratado nesse projeto.

A motivacao para esse projeto veio da observacao empirica de realizacao de `trading` durante seis meses e da observacao do avanco e retracao dos precos praticados no mercado para essas moedas. Apos observacao empirica, notei que muito da valorizacao dessas moedas se da' pelo movimento de manada das pessoas interessadas nelas e sobre a aceleracao/desaceleracao do preco praticado nas exchanges.
Seguindo o racional do aumento do preco devido a especulacao e interesse das pessoas pela moeda, o preco comeca a subir rapido e o numero de pessoas interessadas naquela moeda espeifica aumenta muito (tambem em questao de horas). Neste contexto, minha proposta e' criar um modelo predizer se uma moeda esta' entrando / entrou nesta fase de aumento vertiginoso baseado no aumento do volume de negociacoes e na aceleracao do seu preco em um curto periodo de tempo.

Datasets and Inputs

Existem varias exchanges em funcionamento. Dentre elas, existe uma que fornece dados em tempo real dos precos praticados e do livro de ordens e compras: Bittrex. O site dessa exchange e' https://bittrex.com/ (https://bittrex.com/). Na secao de APIs do site, podemos observar que eles disponibilizam informacoes sobre todas as moedas em tempo real. A partir disso, coletei informacoes sobre as seguintes moedas de novembro de 2017 ate janeiro de 2018: 1ST, 2GIVE, ADA, BITB, DOGE, MANA, RDD, VOX, XDN, XLM e XVG. Nesse periodo podemos observar aumentos e retracoes dos precos praticados para essas moedas e servira' como base para treinamento de nosso modelo.

Os dados brutos da serie historica de precos coletados a partir do site da Bittrex sao:

- MarketName: nome do mercado;
- TimeStamp: timestamp da amostra;
- Volume: valor total da moeda que esta circulando naquela exchange;
- Last: ultimo preco praticado;
- OpenSellOrders: quantidade de ofertas de venda (pessoas interessads em vender a moeda);
- OpenBuyOrders: quantidade de ofertas de compra (pessoas interessads em comprar a moeda);
- Bid: Maior valor numa oferta de compra;
- Ask: Menor valor numa oferta de venda.

Um exemplo do csv salvo com esses dados para a moeda RDD e':

```
MarketName;TimeStamp;Volume;Last;OpenSellOrders;OpenBuyOrders;Bid;Ask;
BTC-RDD;2017-12-
14T16:10:59.23;1097844228.19;0.000000090;10681;1881;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:11:13.12;1097844228.19;0.000000090;10681;1881;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:16:06.403;1102317759.04;0.000000100;10690;1882;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:21:21.377;1101497488.17;0.000000100;10684;1892;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:26:11.02;1105007036.0;0.000000090;10686;1892;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:31:26.527;1112006809.07;0.000000090;10694;1901;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:36:21.09;1111138944.99;0.000000090;10692;1906;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:41:12.44;1109766134.97;0.000000100;10676;1911;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:46:01.427;1104276654.69;0.000000100;10692;1906;0.000000090;0.000000100;
```

As coletas dos dados sobre as moedas foi realizada de 5 em 5 minutos, respeitando a disponibilidade da API do site que ficou indisponivel em alguns momentos. Essa indisponibilidade gerou alguns buracos nas amostras.

Esses dados caracterizam uma serie temporal do mercado e precos praticado para essas moeddas. Esta serie temporal sera transformada em um conjunto de amostras que sejam independentes entre si. Depois dessa transformacao, utilizarei esses pontos para o treinamento de um modelo de aprendizado supervisionado.

Essa transformacao dos dados sera' realizada do seguinte modo. Seja X^i um ponto qualquer da serie de dados brutos, conforme coletados do site da Bittrex.
Os dados utilizados para treinamento e teste (X"^i) serao gerados a partir desse X^i do seguinte modo:

MarketName;
Volume_d_s: X[Volume]^i / X[Volume]^i - 5
Volume_d_l: X[Volume]^i / X[Volume]^i - 10
Last_d_s: X[Last]^i / X[Last]^i - 5
Last_d_l: X[Last]^i / X[Last]^i - 10
OpenSellOrders_d_s: X[OpenSellOrders]^i / X[OpenSellOrders]^i - 5
OpenSellOrders_d_l: X[OpenSellOrders]^i / X[OpenSellOrders]^i - 10
OpenBuyOrders_d_s: X[OpenBuyOrders]^i / X[OpenBuyOrders]^i - 5
OpenBuyOrders_d_l: X[OpenBuyOrders]^i / X[OpenBuyOrders]^i - 10
Bid_d_s: X[Bid]^i / X[Bid]^i - 5
Bid_d_l: X[Bid]^i / X[Bid]^i - 10
Ask_d_s: X[Ask]^i / X[Ask]^i - 5
Ask_d_l: X[Ask]^i / X[Ask]^i - 10
Y: (X[Last]^i - 10 * 1.10) < X[Last]^i

Por exemplo, para os dados ficticios abaixo:

```
MarketName;TimeStamp;Volume;Last;OpenSellOrders;OpenBuyOrders;Bid;Ask;
BTC-RDD;2017-12-
14T16:10:59.23;1000;0.000000085;10681;1881;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:11:13.12;1010;0.000000090;10681;1881;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:16:06.403;1020;0.000000100;10690;1882;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:21:21.377;1030;0.000000100;10684;1892;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:26:11.02;1040;0.000000090;10686;1892;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:31:26.527;1050;0.000000090;10694;1901;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:36:21.09;1060;0.000000090;10692;1906;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:41:12.44;1080;0.000000100;10676;1911;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:46:01.427;1090;0.000000100;10692;1906;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:50:52.65;1100;0.000000090;10725;1914;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T16:56:18.077;1110;0.000000095;10730;1926;0.000000090;0.000000100;
BTC-RDD;2017-12-
14T17:01:22.033;1120;0.000000090;10733;1931;0.000000090;0.000000100;
```

Temos que o primeiro ponto gerado sera com o X^i partindo do dia 2017-12-14T16:56:18.077:

```
MarketName: BTC-RDD;
Volume_d_s = X[Volume]^i / X[Volume]^i -5 = 1110 / 1050
Volume_d_l =  X[Volume]^i / X[Volume]^i - 10 = 1110 / 1000
Last_d_s =  X[Last]^i / X[Last]^i - 5 = 0.000000095 / 0.000000085
Last_d_l =  X[Last]^i / X[Last]^i - 10 = 0.000000095 / 0.000000090
...
Y = ( (X[Last]^i - 10) * 1.10) < X[Last]^i = (0.000000085 * 1.10) < 0.000000095
= 'BUY'
```

Realizarei essas transformacoes para conseguir uma normalizacao dos dados mediante os diferentes valores das moedas e para transformar os dados da serie historica em amostras de dados passiveis de serem analisadas pelos classificadores. Para determinar o rotulo de crescimento acentuado ou nao, vou considerar que o preco precisa ser ao menos 10% maior que o preco no momento i-10. Se a diferenca for menor que o aumento de 10%, a resposta sera' SKIP, caso contrario sera' BUY.

Solution Statement

Vou analisar os dados transformados utilizando crossvalidation com os classificadores: DummyClassifier, Regressao Logistica, KNN e SVM. A performance desses classificadores sera verificada mediante a metrica F1 e o classificador mais performatico sera avaliado com um GridSearch para realizar o ajuste fino dos parametros.

Benchmark Model

O classificador `DummyClassifier` sera' utilizado, tao somente, como baseline para os demais classificadores. Esse classificador devolve um retorno aleatorio a cada chamada. Deste modo, podemos analisar se nossos classificadores sao no minimo melhores que o acaso.

Evaluation Metrics

Durante o crossvalidation sera utilizada apenas a metrica F1 para avaliacao dos resultados.
Utilizarei essa metrica por ser uma metrica que avalia tanto `recall` quanto `precision`.
Essa metrica tambem sera utilizada durante a busca em grade no `GridSearch` para realizar o ajuste fino dos parametros do classificador escolhido pelo melhor desempenho no `crossvalidation`.
No segundo momento, quando for avaliar a performance do classificador apos o ajuste fino, utilizarei a metrica F1 e a matrix de confusao para avaliar os resultados. Uma vez que e' mais interessante perder uma chance de compra do que comprar num momento unoportuno. Assim, nesse segundo momento, precisao e' mais importante que `recall`.

Project Design

O processo como um todo se divide nos passos de:

- recuperar os dados do site da Bittrex;
- transformar os dados de serie historia para amostras independentes;
- executar o `crossvalidation` para determinar qual classificador apresenta melhores resultados com os dado apresentados;
- no caso em que os classificadores nao se ajustem corretamente aos dados, sera necessaria a analise mais aprofundada da distribuicao dos dados para determinar se serao necessarias manipulacoes como criacao de atributos ou extracao de componentes

principais (PCA);
- executar o GridSearch para deteminar o ajuste fino dos parametros para o classficador escolhido;
- mensurar os resultados do classificar ajustado mediante a metrica F1 e o Precision para determinar a qualidade dos resultados;
- discutir as implicacoes, possibilidades e restricoes do modelo mediante o resultado alcancado.