Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Marcelo Ferraz de Oliveira

11 de julho de 2018

Proposta

Histórico do assunto

O projeto tratará sobre análise de transações da bolsa de valores.

Em contraste com métodos tradicionais de análise, que normalmente buscam predizer o preço futuro baseado apenas na série histórica de preços, pretendo realizar uma classificação supervisionada baseada na variação dos preços, volumes e participantes nos negócios realizados Este tipo de análise é conhecida como *Tape Reading*.

A análise em questão foi o motivo do meu ingresso neste Nanodegree e, com os conhecimentos adquiridos até o momento, acredito que serei capaz de obter um resultado útil.

Descrição do problema

A análise denominada *Tape Reading* preconiza que as oportunidades de negócio devem ser identificadas através da análise do histórico de negócios realizados e do *livro de ofertas*, levando em consideração as características e intensidades que os negócios são realizados.

Este tipo de análise foi a utilizada por antigos operadores de *pregão presencial* - operadores de ativos financeiros de épocas anteriores à negociação eletrônica atual - e hoje é usada por operadores especializados e alguns poucos robôs de negociação automática, devido à sua maior complexidade e maior variedade de atributos analisados.

O problema, então, é tentar classificar dados anteriores que levem a operações lucrativas da mesma forma que é feita pelas pessoas, através de algoritmos de aprendizagem.

Conjuntos de dados e entradas

O conjunto de dados é formado por dados históricos tick-a-tick (negócio a negócio) da ação PETR4, dos pregões realizados entre os dias 01/02/2018 e 30/06/2018. Cada dia de negociação se encontra em um arquivo diferente. A quantidade de registros é variável, girando em torno de 50.000 a 100.000 entradas/dia. Cada registro representa um negócio realizado, contendo os atributos:

- ativo: código B3 do ativo;
- id negócio: identificação única do negócio;
- Tempo: hora da realização do negócio, com precisão de segundos;
- Preço: valor que a transação foi realizada;
- Volume: quantidade de ativos negociados na transação;
- comprador: código da corretora que o comprador utilizou para realizar o negócio;
- vendedor: código da corretora que o vendedor utilizou para realizar o negócio;
- direção: informação sobre quem tomou a iniciativa do negócio - comprador ou vendedor;
- direto: indica se o negócio foi ou não direto - tipo de negócio fechado diretamente entre participantes, de uma mesma corretora, que apenas é registrado posteriormente pela B3 e não gera os mesmos impactos - a princípio - que um negócio convencional.
- tempo_msc: hora da realização do negócio, em formato Unix Epoch, com precisão de microssegundos.
- pcompra1 a pcompra5; pvenda1 a pvenda5: informações sobre o valor dos 5 melhores preços de compra e de venda disponíveis no livro de ofertas;
- compra1 a compra5; venda1 a venda5: informações sobre a quantidade de ativos ofertados nos 5 melhores preços de compra e de venda;

Este conjunto possui todas as informações possíveis que a B3 disponibiliza sobre os negócios realizados, sendo, portanto, suficiente para a análise em questão.

Como se trata de um projeto de classificação

O conjunto de dados foi elaborado através da análise do fluxo de transações completo obtido por mim através de contrato de acesso às cotações da B3, através do Distribuidor Cedro Finances, na categoria de *Usuário não profissional,* que proíbe o uso e distribuição de dados apenas para fins comerciais, sendo de uso livre para uso e análise próprios.

A política de distribuição comercial da B3 está disponível em: http://www.bmfbovespa.com.br/pt_br/servicos/market-data/distribuidores/politica-comercial-e-contratos/

O contrato do serviço de cotações da Cedro Finances está disponível em: http://files.cedrotech.com/Juridico/CONTRATOPADRAO.pdf

Descrição da solução

Para conseguir identificar oportunidades de negócio através dos dados, será necessário realizar o treinamento de um modelo de classificação supervisionada, para verificar se a transação:

- provocará elevação de preços
- provocará redução de preços
- não provocará alteração de preços

Esta verificação será feita através da criação de uma variável-alvo baseada na variável original do conjunto de dados “Preço”: através da informação da variação do valor da variável, será definido um novo atributo, assumindo um dos três valores possíveis: elevou, reduziu ou não alterou o preço.

Com o modelo treinado, deverá ser possível prever se um conjunto de transações poderá provocar um movimento nos preços.

Modelo de referência (benchmark)

O modelo de referência a ser utilizado será a predição aleatória (random guessing). Será gerada uma sequência aleatória de resultados, que será comparada com a variável-alvo da mesma forma que o modelo de previsão.

Métricas de avaliação

Para a verificação inicial do modelo, como se trata de um problema de classificação, será utilizado o escore F1, descrito pela fórmula:

$$F1 = \frac{2 * precision * recall}{precision - recall}$$

Design do projeto

O primeiro passo será a exclusão de registros desnecessários para a análise. Um exemplo são os registros de *negócios diretos* e de *leilão* - negócios realizados no início e no fim do pregão de negociação e que seguem uma lógica completamente diferentes dos negócios convencionais. Também será analisada a integridade e completude de cada arquivo de dados, com a consequente exclusão de registros incompletos ou danificados.

Devido ao grande volume de dados, a próxima etapa será a realização algum tipo de agrupamento de negócios, para o problema ser computacionalmente analisável. A agregação óbvia é por janela de tempo, transformando os atributos de cada

transação em atributos de variação ou agregação no período de tempo determinado. Com o agrupamento, será possível unir todas as bases de dados em uma só.

Com os dados agregados, será necessária a transformação do conjunto de dados - até então uma série temporal - em um conjunto não temporal, possibilitando uma classificação supervisionada. Para isso, serão geradas variáveis de valores anteriores (*lag variables*) de determinados parâmetros, de modo que cada um dos registros sejam temporalmente independentes dos demais.

Seguindo o mesmo procedimento, a variável-alvo será obtida através de uma variável posterior (lead variable): o preço do próximo registro. Para efeitos de classificação, através da informação da variação do valor da variável posterior com a atual, esta assumirá um dos três valores possíveis: elevou, reduziu ou não alterou.

Com a base de dados transformada, segundo os procedimentos acima, será possível finalmente realizar a análise dos dados e o treinamento do modelo.

Será feita uma análise exploratória dos dados, para verificar possíveis correlações entre os atributos, com objetivo principal de realizar uma redução de dimensionalidade, devido ao grande número de atributos gerados por variáveis de valores anteriores.

Ademais, será feita a divisão da base de treinamento e teste, sendo a base de teste o período final da base de dados completa pois, apesar da

Por fim, serão criadas bases de dados de treinamento, validação cruzada e teste e treinados modelos de aprendizagem supervisionada.

Referências

https://www.tororadar.com.br/blog/o-que-e-tape-reading-analise-de-fluxo-de-ordens

https://scalpertrader.com.br/lendo-mercado-tape-reading/

https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/