Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

DCC057 Mineração de Dados

Trabalho Prático 4 — Proposta

Alexander Thomas Mol Holmquist

19 de março de 2021

1 Introdução

A predição de preços e a sistematização deste processo é de caráter essencial para o sucesso de um investidor. Neste contexto, contar com uma ferramenta automatizada pode ser muito útil, pelas informações que ela oferece. A automatização de investimentos já se consolidou como uma alternativa que atrai muitos clientes, ao ponto de existirem empresas dedicadas a este propósito, como [1, 2]. Isto é feito através de nada menos que mineração de dados.

A bolsa de Nova York, sendo a maior bolsa financeira do mundo, é de muita importância para o mercado mundial. O presente trabalho busca analisar um conjunto de dados desta bolsa [3] coletado entre 2010 e 2016 [4], gerando um preditor de preços. Especificamente, os objetivos de negócio são:

1. **Objetivo principal:** Produzir um preditor robusto para o preço de ações na bolsa de Nova York.
2. **Objetivo secundário:** Analisar a validade dos modelos: reversão à média e fator momento estatisticamente, para a bolsa de Nova York.

Note que não é pretendida a generalização do modelo para bolsas financeira que não a de Nova York. Aos objetivos de negócio acima, correspondem os seguintes objetivos de mineração de dados:

1. **Objetivo principal:** construir um modelo de regressão com pontuação R2 maior que 95% para o preço.
2. **Objetivo secundário:** Verificar se os modelos mencionados atingem erro quadrado médio menor para o conjunto de teste, quando comparados com o modelo de regressão do projeto. Então analisar a esta diferença do ponto de vista estatístico, especialmente em termo da média e variância dos erros.

2 Fonte de dados

Somente uma fonte de dados será utilizada neste projeto. Ela pode ser encontrada no endereço [3]. O tamanho total dos arquivos fornecidos, todos no formato CSV, é cerca de 105.8 megabytes. Este projeto, contudo, provavelmente se restringirá ao arquivo ”prices.csv”, de tamanho aproximado 51.7 megabytes. Ele contém as colunas:

- **date:** data a que a entrada se refere.
- **symbol:** o símbolo (abreviatura) correspondente ao nome da empresa referenciada pelo registro.
- **open:** preço de uma ação ao abrir a bolsa no dia.
- **close:** preço de uma ação ao fechar a bolsa no dia.
- **low:** preço mínimo de uma ação da empresa no dia.
- **high:** preço máximo de uma ação da empresa no dia.
- **volume:** quantidade de ações negociadas no dia.

3 Plano de projeto

Este projeto será guiado pelos princípios estabelecidos pelo CRISP-DM. Com base nas fases 2 a 6 do CRISP-DM, segue o plano inicial do projeto:

3.1 Carregamento e descrição dos dados

Decidiu-se integrar as tarefas de carregamento e descrição dos dados em uma só fase. Isto difere do último projeto realizado para esta disciplina, que distinguiu entre as duas.

- **Passo 1:** embarcar os dados em um ambiente da linguagem Python, provavelmente um *Jupyter notebook*. Estas duas serão as ferramentas fundamentais utilizadas por este projeto, juntamente com outras tecnologias que se mostrarem necessárias. Talvez seja necessário utilizar servidores remotos, como aqueles disponibilizados pela ferramenta *Google colab* [5].

  Se for preciso realizar alguma transformação nos dados para realizar a tarefa, deve-se registrar no relatório.

- **Passo 2:** verificar a qualidade dos dados, especialmente em respeito a valores faltantes, mas pode incluir outras observações. Anotar a frequência dos valores faltantes e propor solução. Se não puder esperar até a etapa de preparação de dados, lidar com eles já neste passo.

- **Passo 3:** listar estatísticas básicas de cada atributo, como variância, curtose e assimetria.

- **Passo 4:** produzir visualizações interessantes dos dados, como histogramas ou gráficos que mostrem a correlação de dois atributos.

- **Passo 5:** produzir representações de dimensão reduzida dos dados, se possível. Gerar gráficos explicativos. Uma possível saída deste passo é a descoberta do quanto um subconjunto de atributos contribui para a variância do conjunto como um todo.

- **Passo 6:** explorar os dados. Este passo tem como objetivo dar liberdade para caminhos não cobertos nos passos anteriores.

Saídas:

- Relatório de carregamento e descrição dos dados. Deve incluir reflexões induzidas pela exploração realizada, assim como qualquer outra consideração interessante.

- Código Python para carregamento e descrição básicos dos dados

Risco: pode ser que se descubra que os dados não são válidos para atingir os objetivos do projeto. Neste caso, deve-se buscar outro banco de dados.

3.2 Preparação dos dados e início da modelagem

Esta fase procura integrar as fases 3 e 4 do projeto passado – trabalho prático 3 desta disciplina. Isto porque se percebeu que é muito difícil saber exatamente qual formato os modelos requerem, sem ao menos começar a montá-los, e sem experiência técnica prévia.

- **Passo 1:** solucionar problemas de qualidade de dados, procurando seguir as propostas do relatório de carregamento e descrição de dados.
- **Passo 2:** compilar uma lista de modelos compatíveis com as observações feitas até o momento. Comparar analiticamente estes modelos, deixando na lista somente modelos que se confundirem em termos de qualidade esperada. Esta classificação deve seguir conhecimentos de mineração de dados e experiências passadas, tanto do autor como de grupos externos, com o problema em questão.
- **Passo 3:** construir um ambiente de testes para tais modelos, se possível. Excluir da lista os modelos para os quais a construção do ambiente de teste se mostrar impraticável ou desnecessária. Para cada modelo diferente, um processo de normalização deve ser aplicado. Portanto, o código desses processos deve ser armazenado.

Saídas:

- Relatório de preparação dos dados e dos modelos. De preferência deve conter as razões da escolha ou não de cada modelo considerado.
- Ambiente básico para aplicação e teste do(s) modelo(s) escolhido(s). Código Python relacionado.

3.3 Calibração dos modelos e retrospectiva

Decidiu-se separar esta seção com o propósito único de garantir que o modelo escolhido é o melhor que se consegue para o contexto. Note que esta fase reúne tarefas das fases 4 e 5 do CRISP-DM.

- **Passo 1:** aplicar todos os modelos que restaram como candidatos da fase passada.
- **Passo 2:** calibrar os parâmetros dos modelos da melhor forma possível, de acordo com os resultados obtidos no conjunto de validação. Construir tabela com métricas sumarizando os resultados. Escolher o melhor modelo.
- **Passo 3:** refletir no processo de mineração de dados até o momento. Garantir que nenhum aspecto foi ignorado, e que o modelo em mãos foi elegido de forma correta. A partir daqui, é assumido que o modelo é de fato o melhor que se consegue.

Saídas:

- Relatório de calibração de modelos e retrospectiva.
- Modelo calibrado do projeto.

Risco: no processo de modelagem, há o perigo de se esquecer os objetivos do projeto, e portanto dispender esforços desnecessários.

3.4 Avaliação

Aqui, os resultados obtidos são avaliados em vista dos objetivos de negócio.

- **Passo 1:** argumentar como os resultados obtidos (neste ponto considerados como o melhor possível) atendem aos objetivos iniciais de negócio. Se não atendem, explicar por quê isso foi inevitável.
- **Passo 2:** produzir relatório de mineração de dados (leva em conta aspectos de todo o projeto).

Saída: relatório de mineração de dados.

Referências

[1] https://www.betterment.com/. Accessado em 19 de março de 2021.

[2] https://www.wealthfront.com/. Acessado em 19 de março de 2021.

[3] https://www.kaggle.com/dgawlik/nyse. Acessado em 19 de março de 2021.

[4] https://www.nyse.com/markets/reports. Acessado em 19 de março de 2021.

[5] https://colab.research.google.com/. Acessado em 19 de março de 2021.