Coordenação:

Prof. Dr. Adolpho Walter Pimazzi Canton

Profa. Dra. Alessandra de Ávila Montini

Big Data

Tema da Aula: Estudo de caso

Prof. Anderson França

Maio de 2018

Estudo de Caso - Intermediário

Exploratória - Regressão Linear

The Boston Housing Dataset

O conjunto de dados contém informações coletadas pelo Serviço de Censo dos EUA sobre habitação na área de Boston.

Os dados foram originalmente publicados por Harrison, D. e Rubinfeld, D.L. "*Hedonic prices and the demand for clean air*", J. Environ. Economics & Management, volume 5, 81-102, 1978. Link para a publicação

O trabalho investiga os problemas metodológicos associados ao uso de dados do mercado habitacional para medir a disposição de pagar por ar limpo.

Dataset

O Dataset disponibilizado contém 506 observações com 14 variáveis, sendo:

1. **CRIM -** taxa de criminalidade per capita por cidade
2. **ZN -** proporção de terrenos residenciais destinados a lotes com mais de 25.000 pés quadrados (25.000 pés² = 2.322,576 mts²)
3. **INDUS -** proporção de hectares comerciais não varejistas por cidade.
4. **CHAS -** Variável dummy de Charles River (1 se o setor delimita rio; 0 caso contrário)
5. **NOX -** concentração de óxidos nítricos (partes por 10 milhões)
6. **RM -** número médio de quartos por habitação
7. **AGE -** proporção de unidades ocupadas pelo proprietário construídas antes de 1940
8. **DIS -** distâncias ponderadas para cinco centros de emprego de Boston
9. **RAD -** índice de acessibilidade para rodovias radiais
10. **TAX -** taxa de imposto sobre propriedades de valor integral por US $ 10.000
11. **PTRATIO -** proporção aluno-professor por cidade
12. **B - 1000(Bk - 0.63)^2** onde Bk é a proporção de negros por cidade
13. **LSTAT -** % *lower status* da população
14. **MEDV -** Valor mediano de residências ocupadas pelo proprietário em US $ 1.000

Utilizar a folha de exercícios Boston Housing

Estudo de Caso - Intermediário

Cluster K-means e Hierárquico

Dataset - USArrests

A base de dados contém estatísticas, em detenções por 100.000 residentes por agressão, homicídio e violação em cada um dos 50 estados dos EUA em 1973. Também é dada a percentagem da população que vive em áreas urbanas.

O estudo foi feito por McNeil e pode ser encontrado em https://books.google.ch/books?id=zl9qAAAAMAAJ&pg=PA20.

Estudo de Caso - Difícil

Exploratória - Data Wrangling - Cluster

Uma empresa brasileira - líder em seu segmento, coletou dados das campanhas de marketing de pesquisa durante 6 meses entre 2015 e 2016.

A empresa busca entender o comportamento de suas campanhas, similaridade e dissimilaridade para melhorar o processo de otimização de cada grupo de campanhas.

Para isso, abriu processo de seleção para que diversas consultorias apresentassem suas análises iniciais e a empresa com melhor desempenho irá assumir todos os projetos da empresa.

Dataset

Download

```
dados                                  334874 obs. of 10 variables
 DIA : chr "01/08/2015" "01/08/2015" "01/08/2015" "01/08/2015" ...
 ID CAMPANHA : chr "CX75X" "CXU44" "CCU64" "CC474" ...
 EXIBICOES : int 21 3 361 224 2 2 12 30 31 26 ...
 INTERAÇÕES : int 0 0 8 9 0 0 1 4 2 0 ...
 COMPRAS : int 0 0 0 0 0 0 0 0 0 0 ...
 CUSTODEMIDIA : num 0 0 14.4 11.2 0 ...
 POTENCIAL_EXIBICAO : num 23.37 8.61 3964.69 2238.6 2.46 ...
 EXIBICAOPERDIDAPORQUALIDADE: num 2.46 6.15 1607.28 2014.74 1.23 ...
 EXIBICAOPERDIDAPORORCAMENTO: num 0 0 1997 0 0 ...
 Campaign_name : chr "Produto630" "Produto824" "Categoria165" "Produto1354"
```

Dataset

O Dataset disponibilizado contém 334.874 observações com 10 variáveis, sendo:

Dia (char): O dia em que a campanha foi veiculada
ID Campanha (char): Identificação da campanha
EXIBIÇÕES (int): Quantidade de vezes que uma pessoa visualizou uma campanha
INTERAÇÕES (int): Quantidade de vezes que uma pessoa interagiu com uma campanha
COMPRAS (int): Compras realizadas após a interação com a campanha
CUSTODEMIDIA (num): Custo da campanha
POTENCIAL_EXIBIÇÃO (num): O quanto a campanha poderia ter sido visualizada
EXIBIÇÃOPERDIDAPORQUALIDADE (num): Quanto uma campanha não exibiu por falta de relevância no conteúdo
EXIBIÇÃOPERDIDAPORORCAMENTO (num): Quanto uma campanha não exibiu por falta de orçamento
Campaign_name (char): Identificação da campanha

Métricas Relativas

Existem algumas métricas que são de extrema importância para se avaliar o desempenho das campanhas que não estão no dataset. São elas:

CTR: a relação entre interação e exibição (interação/exibição)
CPI: custo médio de cada interação interação (custo/interação)
CR: é a relação entre os clientes que clicaram na campanha e finalizaram uma compra (compra/interação)

Outras informações

- A empresa não fez exigência de ferramentas específicas.
- O resultado de cada etapa e a justificativa do uso da técnica deve ser descrito no documento
- O documento deve ser entregue em formato apresentação ou pdf contendo os resultados

Dicas

Exploratória: é fundamental conhecer a base de dados. Não economizem tempo analisando todas as variáveis e questionando a relação com o negócio.
Cluster: Existem muitas campanhas, portanto, não se esqueçam do cluster.
Exploratória: Não se esqueçam de explorar seus grupos e entender pontos como similaridade e dissimilaridade, tenham em mente que o objetivo é otimizar esforços.
Futuras análises: Existem alguma relação entre as variáveis? É possível utilizar algum outro modelo?
Próximos passos: Deixe suas recomendações.

Anderson França
Email: anderson.frca@gmail.com
LinkedIn: /andersonfranca1/