MBA Business Analytics e Big Data

Análise Preditiva

Prof. Dr. João Rafael Dias

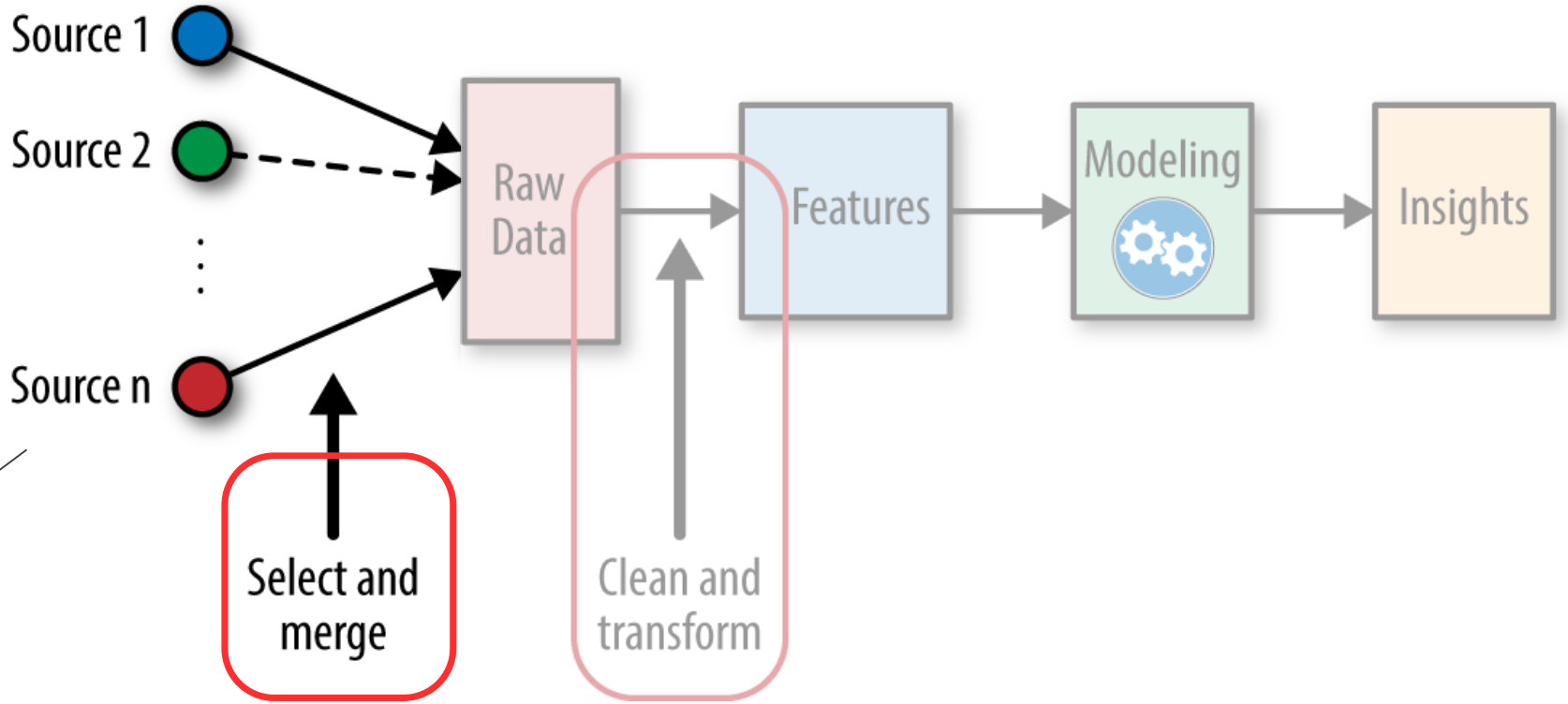
1º semestre - 2020

Feature engineering

Prof. Dr. João Rafael Dias Pinto

Feature engineering
Dados como matéria prima

- São a matéria prima de qualquer projeto que envolva modelagem e *analytics*
- Podem ser estruturados e não estruturados (mais fácil VS mais comum) e estarem em diferentes formatos (texto, tabular, imagens, vídeos etc.)
- A qualidade dos dados definirá a qualidade do modelo a ser treinado e utilizado
- Pensar nas diversas fontes que existem, e como elas estão relacionadas com a variável alvo



Que variáveis deveriam ser utilizadas?

https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/ch01.html

Feature engineering

Dados como matéria prima

- São a matéria prima de qualquer projeto que envolva modelagem e *analytics*
- Podem ser estruturados e não estruturados (mais fácil VS mais comum) e estarem em diferentes formatos (texto, tabular, imagens, vídeos etc.)
- A qualidade dos dados definirá a qualidade do modelo a ser treinado e utilizado
- Pensar nas diversas fontes que existem, e como elas estão relacionadas com a variável alvo

Variáveis cadastrais

Variáveis transacionais

Variáveis que indiquem comportamentos

Variáveis sócio-demográficas

Variáveis com geolocalização

Variáveis de texto

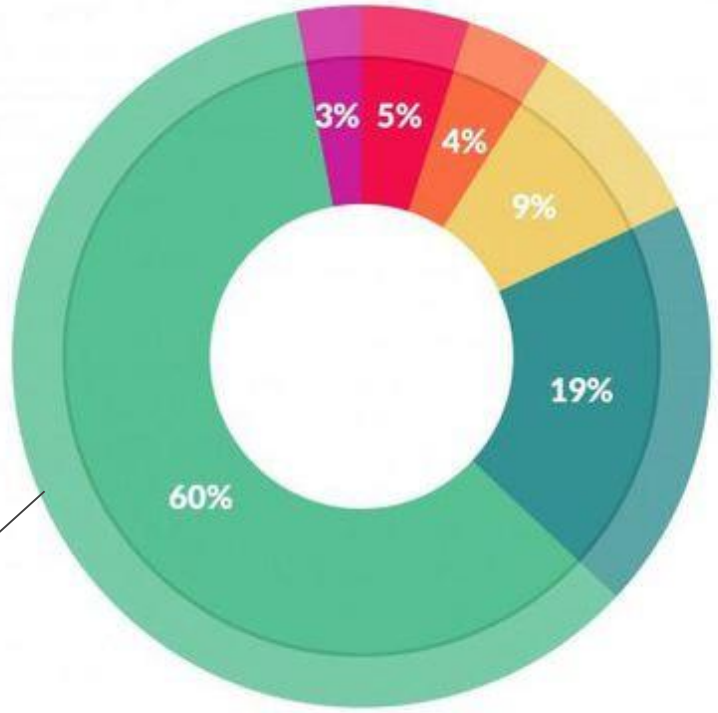
Variáveis de redes sociais, imagens etc

+ exploradas

- exploradas

Feature engineering
Dados como matéria prima

- São a matéria prima de qualquer projeto que envolva modelagem e *analytics*
- Podem ser estruturados e não estruturados (mais fácil VS mais comum) e estarem em diferentes formatos (texto, tabular, imagens, vídeos etc.)
- A qualidade dos dados definirá a qualidade do modelo a ser treinado e utilizado
- Pensar nas diversas fontes que existem, e como elas estão relacionadas com a variável alvo



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Manipulação e organização dos dados custa muito tempo!!

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#3cafeccc6f63

Feature engineering

Feature engineering + análise exploratória de dados

- O processo de *feature engineering* e análise exploratória de dados andam lado à lado nessa etapa
- Precisamos fazer conjuntamente esses dois processos para garantir o devido tratamento da base de dados e aprimorar as análises e a modelagem

Relembrando...

A análise exploratória busca:

- Entender o **comportamento** das variáveis pela suas distribuições
- Identificar **mudanças de padrão** no tempo das variáveis
- Estabelecer a **relação** entre as variáveis
- Estabelecer a **relação** das **variáveis** com a variável **resposta** (alvo do modelo)

Análise Univariada

Análise Bivariada

Feature engineering

Feature engineering + análise exploratória de dados

- O processo de *feature engineering* e análise exploratória de dados andam lado à lado nessa etapa
- Precisamos fazer conjuntamente esses dois processos para garantir o devido tratamento da base de dados e aprimorar as análises e a modelagem

Relembrando...

A análise exploratória busca:

- Entender o **comportamento** das variáveis pela suas distribuições
- Identificar **mudanças de padrão** no tempo das variáveis
- Estabelecer a **relação** entre as variáveis
- Estabelecer a **relação** das **variáveis** com a variável **resposta** (alvo do modelo)

Analisa cada variável individualmente sem verificar relações entre outras variáveis.

Analisa a relação entre duas variáveis na base de dados. Geralmente, o foco é dado na relação entre as variáveis previsoras e a variável resposta

Feature engineering

Feature engineering + análise exploratória de dados

- O processo de *feature engineering* e análise exploratória de dados andam lado à lado nessa etapa
- Precisamos fazer conjuntamente esses dois processos para garantir o devido tratamento da base de dados e aprimorar as análises e a modelagem

Relembrando...

A análise exploratória busca:

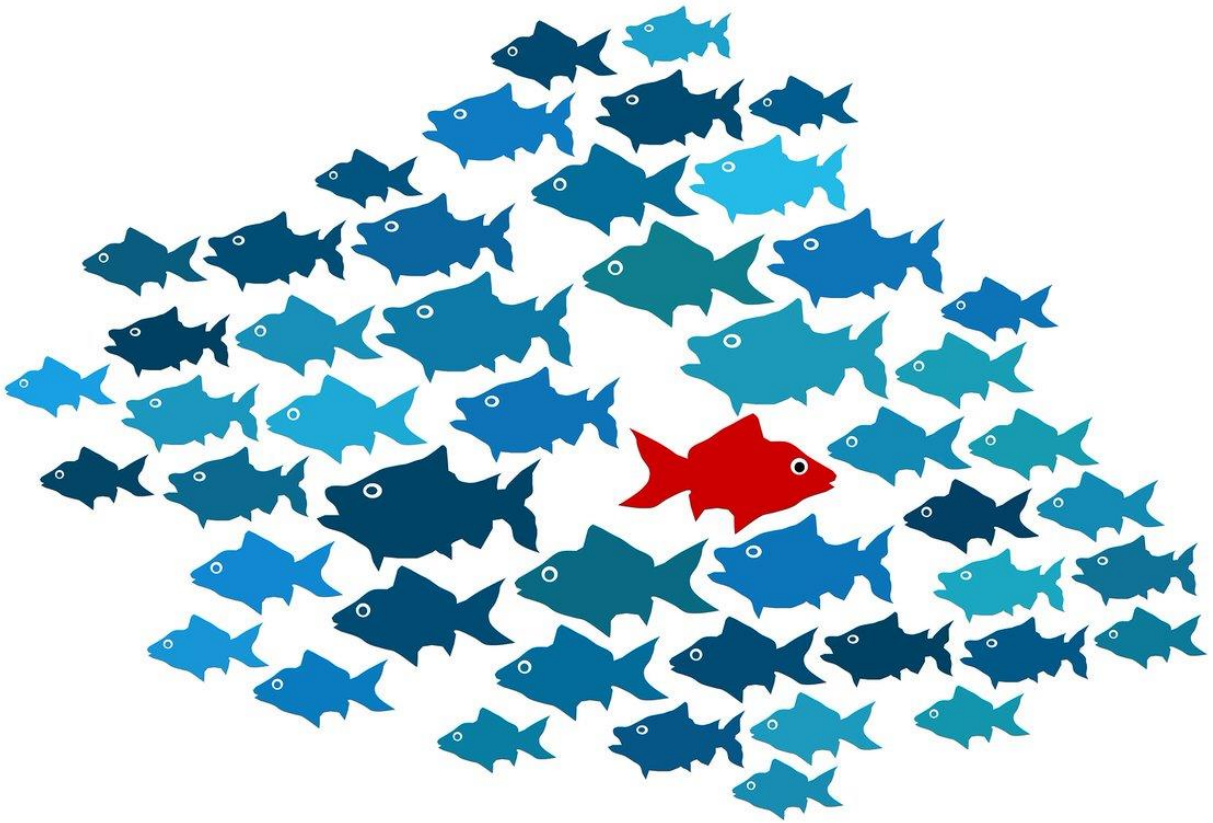
- Entender o **comportamento** das variáveis pela suas distribuições
- Identificar **mudanças de padrão** no tempo das variáveis
- Estabelecer a **relação** entre as variáveis
- Estabelecer a **relação** das **variáveis** com a variável **resposta** (alvo do modelo)

...ferramentas de visualização dos dados

Feature engineering

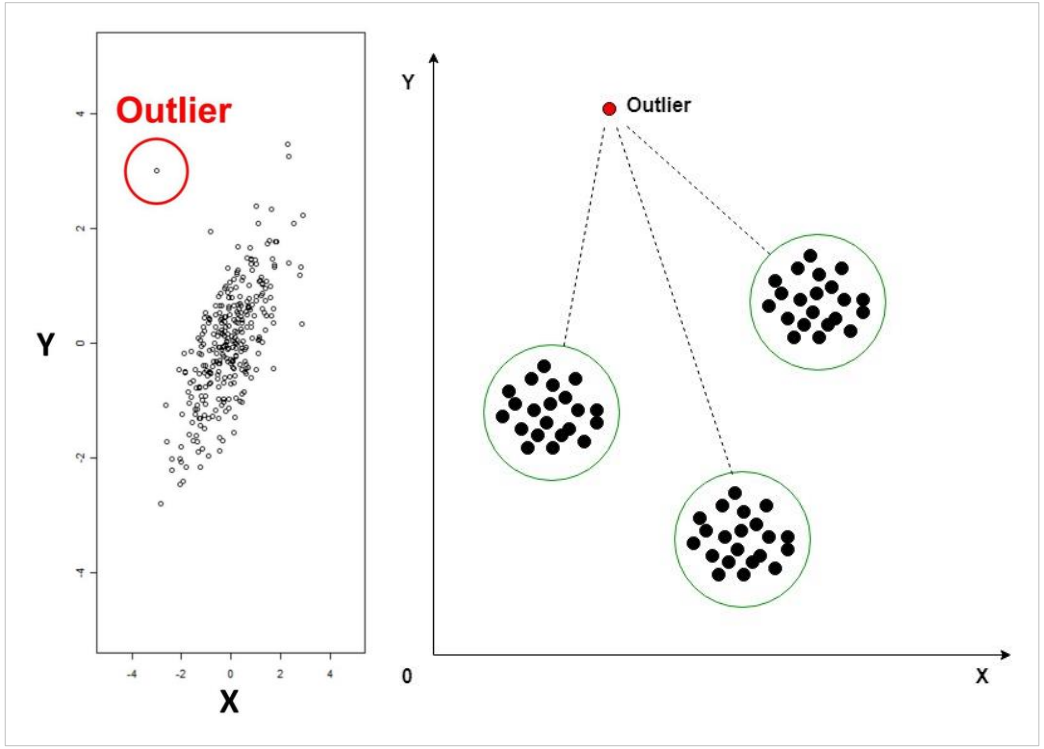
Outliers

- São dados que encontram-se distantes dos demais pontos da amostra. Não significa que sejam errados ou falsos, apenas que são discrepantes
- Eles desviam dos padrões gerais da variável na amostra



https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561

Exemplo



https://towardsdatascience.com/how-to-use-machine-learning-for-anomaly-detection-and-condition-monitoring-6742f82900d7

Feature engineering

Ocorrência de *outliers*

- **Erros na imputação dos valores**
  - indivíduo informou a renda anual no campo de renda mensal
  - indivíduo mentiu sobre o faturamento da empresa para tentar conseguir mais crédito
  - erro de medidas (pés – centímetros, etc)
- **Erros de codificação / digitação**
  - idade do motorista = 4 anos
  - altura 17,8 m (ao invés de 1,78 m)
- **Erros de amostragem**
  - indivíduos de outra população foi selecionado por engano. Por exemplo uma amostra de pessoas apenas com ensino médio completo incluiu por engano ou falha pessoas com pós-graduação
- **Resultado atípico justificável**
  - aluno de graduação com 70 anos
  - dados do salário dos funcionários incluindo o do CEO
- **Problemas sistêmicos**
- **Outros**

Feature engineering

Identificação de *outliers*

- **Com diagramas de boxplot**

  - Cuidado pois eles pressupõem normalidade dos dados (o que não é real na maioria das vezes)

  - Em distribuições assimétricas somente se o ponto estiver muito afastado

- **Pontos que se afastam mais de 3 desvios padrão em relação à média**

  - Q1 – 1,5*IQR ou Q3 + 1,5*IQR,

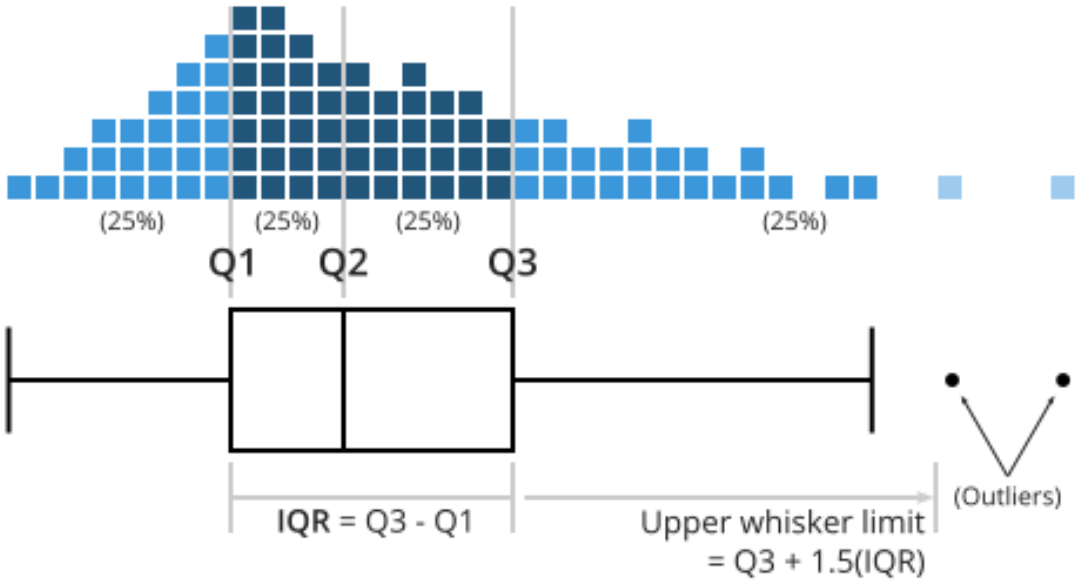
com IQR = Q3 - Q1

  - Pode não funcionar se a variável for fortemente assimétrica

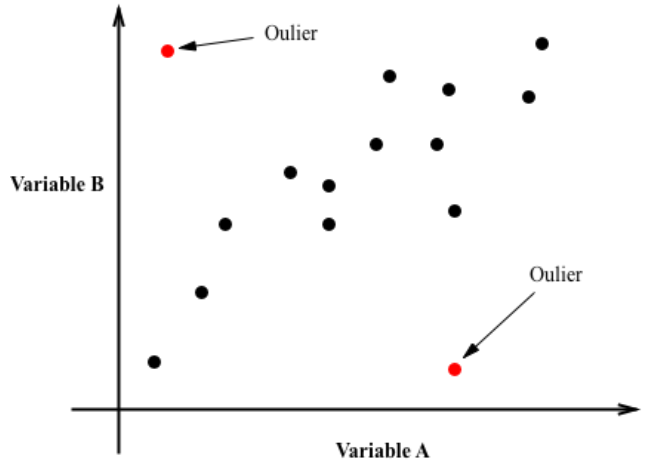
- **Análise visual dos extremos**

  - Subjetivo

  - Ajuda se conhecer o contexto do problema (*problem domain*)



https://chartio.com/learn/charts/box-plot-complete-guide/



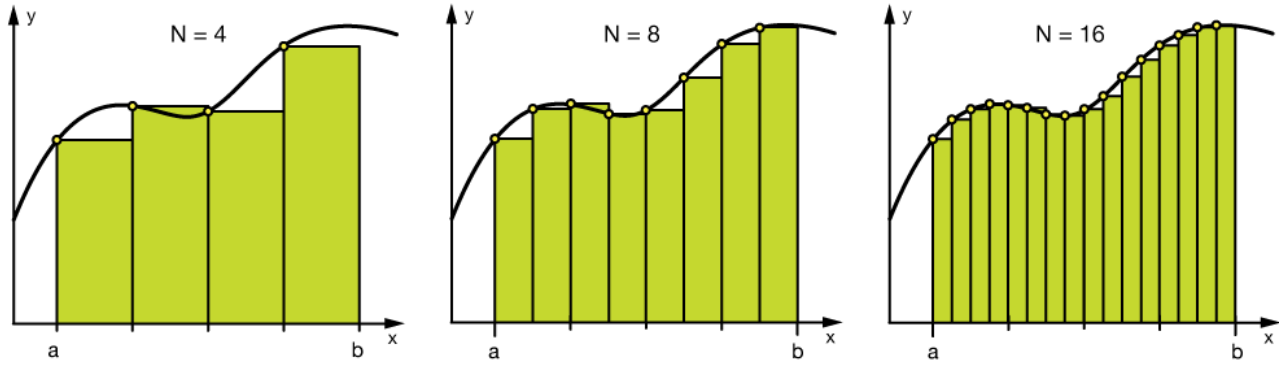
https://www.kaggle.com/nowke9/statistics-3-bivariate-data

Feature engineering

Como proceder com *outliers*

- **Grande número de outliers**
  - Pode ser problema de amostragem
  - Precisa investigar antes de tomar qualquer decisão
- **Grande base de dados com poucos outliers**
  - Remoção completa daquele registro
  - Faz a remoção cujo casos forem superior / inferior a algum determinado valor (julgamento do *expert*), ou removem-se todos os registros
- **Imputação de valores**
  - Utilizar o P1% ou P99% no lugar do outlier, ou P2% ou P98% dependendo do número de outliers.
- **Transformação de variáveis contínuas**
  - Discretização é um jeito de camuflar outliers (*binning*)

Essa transformação consegue "eliminar" os outliers



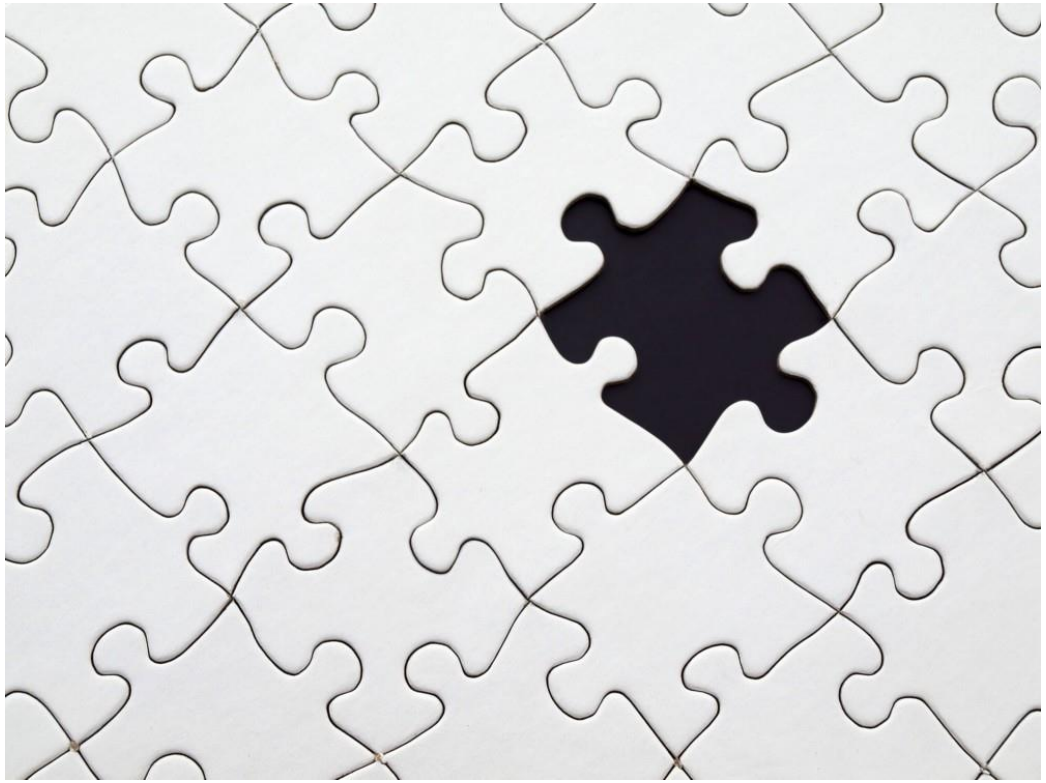
Perceba que podemos "quebrar" uma variável continua em diversas faixas

https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

Feature engineering

Missing values

- No mundo real, há algumas observações onde um elemento particular é ausente. E isso pode estar relacionado a diversos fatores como, por exemplo, dados com problema, falhas ao carregar a informação ou extração incompleta
- É um dos maiores desafios no processo de construção de um modelo preditivo



Exemplo

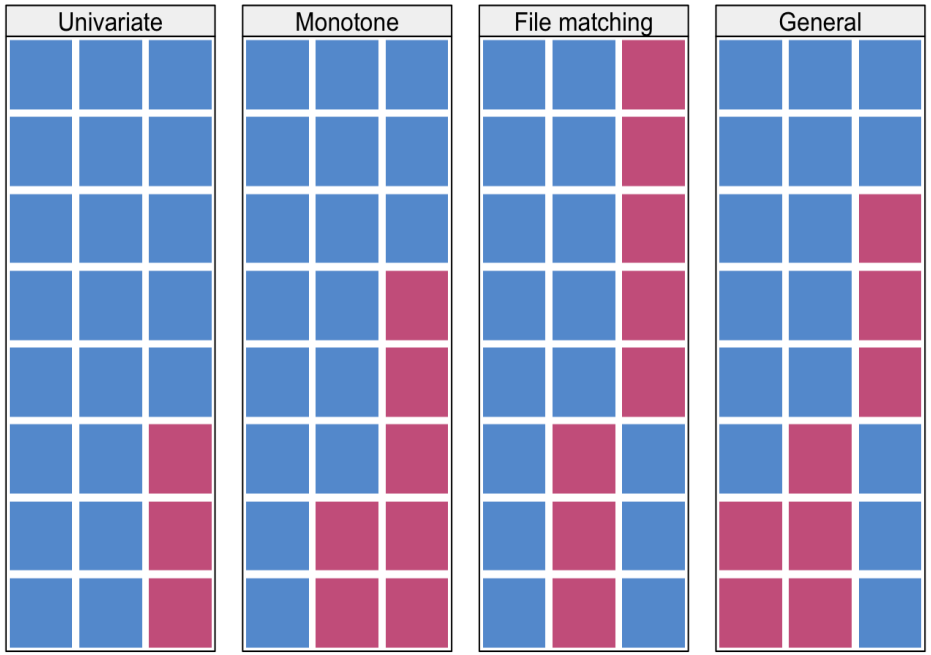
ID	Color	Weight	Broken	Class
1	Black	80	Yes	1
2	Yellow	100	No	2
3	Yellow	120	Yes	2
4	Blue	90	No	2
5	Blue	85	No	2
6	**?**	60	No	1
7	Yellow	100	**?**	2
8	**?**	40	**?**	1

https://www.researchgate.net/publication/280097054_An_Evolutionary_Missing_Data_Imputation_Method_for_Pattern_Classification

Feature engineering

Ocorrência de *missing values*

- **Por que estão ocorrendo?** (pergunta fundamental no início da investigação)
  - Valores não previstos no momento do fornecimento dos dados
  - Banco de dados formado a partir informações distintas
  - Há campos que não são comuns
  - Falha na coleta dos dados
  - *Missing* estrutural (não se aplica: tempo de emprego de profissional autônomo, por exemplo)



https://stefvanbuuren.name/fimd/missing-data-pattern.html

- **Padrões de ocorrência:**
  - Aleatória (MAR – *missing at random*)
  - Não aleatório (MNAR – *missing not at random*)
    - Relação com a variável alvo (*missing* ocorrem em maior proporção para clientes inadimplentes)
    - Relação com outras variáveis (Idade não informada mais frequente por mulheres que homens)
    - Relação com a própria variável previsora (pessoas com salários maiores tendem a não informar os valores)

Feature engineering

Como proceder com *missing values*

- Não existe uma forma única de se lidar com *missing values*, depende de cada situação
- Independentemente da metodologia, todas possuem prós e contras
- **Alternativas**

- Substituir por novas observações: em geral não é viável e não conduz a boas amostras

- Exclusão de casos com *missing values:* pode ter perda de informação e reduzir o tamanho da amostra para modelagem

- Exclusão de variáveis com um percentual grande valores *missings* (~60% - 70%, discutível!)

Todo o registro é eliminado se uma das variáveis apresentar MV

int_rate	grade	emp_length
6.62	A	1
11.71	B	6
11.71	B	NA
11.71	B	3
15.96	C	1
16.29	D	0
15.27	C	4

Só a variável que possuir o MV é eliminada da modelagem

int_rate	grade	emp_length
6.62	A	1
11.71	B	6
11.71	B	NA
11.71	B	3
15.96	C	1
16.29	D	0
15.27	C	4

- Imputar valores: recomendado para variáveis quantitativa

- Categorizar variáveis quantitativas e criar uma categoria "*Missing*"

Feature engineering

Como proceder com *missing values*

- **Imputar "valor lógico"** (no caso de ocorrência não aleatória)
  - O que é lógico? Possui um alto risco de afetar os resultados do modelo.
- **Imputar média/ mediana/ moda**
  - Possui alto risco principalmente se forem muitos casos
  - Reduz a variabilidade natural dos dados
  - Optar usar a média/ mediana/ moda de observações "similares" (usar técnica de clusterização)
- **Previsão a partir de outras variáveis**
  - Só é possível se a variável com MV puder ser explicada pelas outras variáveis (regressão, árvore, etc.)
  - Tal procedimento pode gerar um viés pois os valores imputados podem ser bem "mais comportados" que os valores reais
- **Para se aprofundar:**

https://www.r-bloggers.com/missing-value-treatment/
https://stefvanbuuren.name/fimd/
https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17
https://www.kdnuggets.com/2017/09/missing-data-imputation-using-r.html

Feature engineering

Geração de novas variáveis

- Esta etapa é muito importante para conseguir extrair mais informação dos dados que possuímos.
- Precisa-se de um conhecimento do negócio grande para obter variáveis com grande poder explicativo
- No processo de geração de novas variáveis pode-se combinar variáveis quantitativas (por exemplo na forma de uma razão)

VALOR EM ATRASO/FATURAMENTO
VALOR DA PARCELA/RENDA MENSAL
RENDA/NÚMERO DE DEPENDENTES

- Existem também interações entre variáveis qualitativas e quantitativas

IDADE + TIPO DO PLANO DE TELEFONIA
ESCOLARIDADE + TIPO DE ESCOLA (Publica ou Particular)

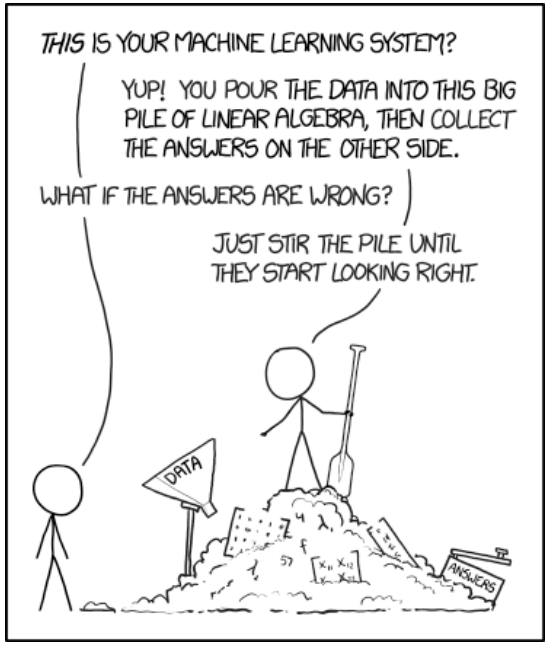
- Toda a criação de variáveis como as destacadas acima dependem muito da relação com a variável resposta usada para treinar o algoritmo, por isso é necessário sempre criar *boxplots*, *scatterplots* e tabelas de contingência.

Feature engineering

Transformação de variáveis

- Permitem extrair mais informação dos dados e auxiliam bastante caso existam restrições de uso para algum algoritmo específico (i.e normalidade das distribuições)
- Permitem comparar a importância/impacto de informações medidas em escalas distintas em regressão
- Podem evitar problemas decorrente do processamento de dados (instabilidade numérica devido à variáveis com magnitudes muito diferentes)
- São requisitos obrigatórios para algoritmos que usam conceito de distância (PCA, clusterização, knn, etc) e que usam algum método de otimização (SVM, redes neurais, etc)
- Podem atenuar o efeito de *outliers*

- **Tipos de transformações:**
  - Criação de novas variáveis
  - Centralização (*centering*)
  - Escalonamento (*scaling*)
  - Padronização (*standardization* | *centering* + *scaling*)
  - Correção de assimetria
  - Discretização de variáveis quantitativas
  - Quantificação de variáveis categóricas



https://xkcd.com/1838/

Feature engineering

Padronização

- A centralização (*centering*) considera a diferença entre os valores e a média da variável
- O escolanamento (*scaling*) equaliza as variâncias das diferentes variáveis (em geral iguais à 1)
- Existem algumas padronizações que utilizam os valores mínimos e máximos (menos frequentes em *machine learning*)