Analytics e Inteligência Artificial Data Science

Tema da aula

Tópicos de Modelagem

FIA BUSINESS SCHOOL

lab.data

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseadas em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

- Único curso de graduação em administração a receber as notas máximas
- A primeira escola brasileira a ser finalista da maior competição de MBA do mundo
- Única *Business School* brasileira a figurar no *ranking* LATAM
- Signatária do Pacto Global da ONU
- Membro fundador da ANAMBA - Associação Nacional MBAs
- Credenciada pela AMBA - *Association of MBAs*
- Credenciada ao *Executive MBA Council*
- Filiada a AACSB - *Association to Advance Collegiate Schools of Business*
- Filiada a EFMD - *European Foundation for Management Development*
- Referência em cursos de MBA nas principais mídias de circulação

BUSINESS SCHOOL lab.data

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.


Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil.

Os diretores foram professores de grandes especialistas do mercado.

+10 anos de atuação.

+9.000 alunos formados.

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria;
- Larga experiência de mercado na resolução de *cases;*
- Participação em congressos nacionais e internacionais;
- Professor assistente que acompanha o aluno durante todo o curso.

Estrutura

- 100% das aulas realizadas em laboratórios;
- Computadores para uso individual durante as aulas;
- 5 laboratórios de alta qualidade (investimento +R$2MM);
- 2 unidades próximas à estação de metrô (com estacionamento).

FIA BUSINESS SCHOOL lab.data

PROFA. DRA. ALESSANDRA DE ÁVILA MONTINI

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Tem muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e Inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em Estatística Aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Parecerista da FAPESP e colunista de grandes portais de tecnologia.

Currículo do Professor

PROF. ÂNGELO CHIODE, MSc

Bacharel, mestre e candidato ao PhD em Estatística (IME-USP), atua como professor de Estatística Aplicada para turmas de especialização, pós-graduação e MBA na FIA. Trabalha como consultor nas áreas de Analytics e Ciência de Dados há 13 anos, apoiando empresas na resolução de desafios de negócio nos contextos de finanças, adquirência, seguros, varejo, tecnologia, aviação, telecomunicações, entretenimento e saúde. Nos últimos 5 anos, tem atuado na gestão corporativa de times de Analytics, conduzindo projetos que envolviam análise estatística, modelagem preditiva e *machine learning*. É especializado em técnicas de visualização de dados e design da informação (Harvard) e foi indicado ao prêmio de Profissional do Ano na categoria Business Intelligence, em 2019, pela Associação Brasileira de Agentes Digitais (ABRADi).

Conteúdo Programático

DISCIPLINAS

- IA E TRANSFORMAÇÃO DIGITAL
- **ANALYTICS**
- INTELIGÊNCIA ARTIFICIAL: MACHINE LEARNING
- INTELIGÊNCIA ARTIFICIAL: DEEP LEARNING
- EMPREENDEDORISMO E INOVAÇÃO
- COMPORTAMENTO HUMANO E SOFT SKILLS

TEMAS: ANALYTICS E MACHINE LEARNING

- ANÁLISE EXPLORATÓRIA DE DADOS
- INFERÊNCIA ESTATÍSTICA
- TÉCNICAS DE PROJEÇÃO
- TÉCNICAS DE CLASSIFICAÇÃO
- **TÓPICOS DE MODELAGEM**
- TÉCNICAS DE SEGMENTAÇÃO
- TÓPICOS DE ANALYTICS
- MANIPULAÇÃO DE BASE DE DADOS
- AUTO ML

TEMAS: DEEP LEARNING

- REDES DENSAS
- REDES CONVOLUCIONAIS
- REDES RECORRENTES
- MODELOS GENERATIVOS

FERRAMENTAS

- **LINGUAGEM R**
- LINGUAGEM PYTHON
- DATABRICKS

FIA BUSINESS SCHOOL lab.data

Conteúdo da Aula

1. Introdução e Objetivo
2. Definição de Períodos
3. Categorização de Variáveis
4. Validação de Modelos

Referências Bibliográficas

1. Introdução e Objetivo

Introdução

Nas aulas anteriores, aprendemos a teoria a respeito de **modelos de regressão linear e logística**. Vimos que tais conceitos podem nos ajudar a solucionar inúmeros *cases* práticos, tais como:

- Existe associação entre o tempo de experiência dos vendedores de veículos de uma concessionária e o seu volume de vendas mensal?
- Existe associação entre a pluviometria de uma cidade em um mês de verão e a quantidade de casos de dengue no mês seguinte?
- Existe associação entre o limite de cartão de crédito definido historicamente para os clientes de um banco e o seu perfil demográfico e transacional?
- Existe associação entre o salário inicial definido pelos gestores de uma empresa para os analistas júniores e suas características durante o processo seletivo?
- Existe associação entre a utilização de um cupom de desconto fornecido em um varejo *online* e características sociodemográficas dos clientes?
- Existe associação entre a decisão de um investidor apostar em renda variável e o seu comportamento histórico em investimentos em renda fixa?
- Existe associação entre a inadimplência a um empréstimo e características comportamentais do tomador no mercado de crédito como um todo?

Objetivo

Nesta aula, vamos aprofundar um pouco mais o nosso entendimento e a aplicação de modelos de regressão, abordando alguns **tópicos adicionais** pertinentes:

- Como definir os **períodos históricos** para cálculo das variáveis explicativas?
- Como contornar certos problemas de **comportamento** associados às variáveis explicativas quantitativas (valores ausentes, *outliers* etc.)?
- Como garantir que o modelo tem o poder de predizer bem a variável resposta para **novas observações**?

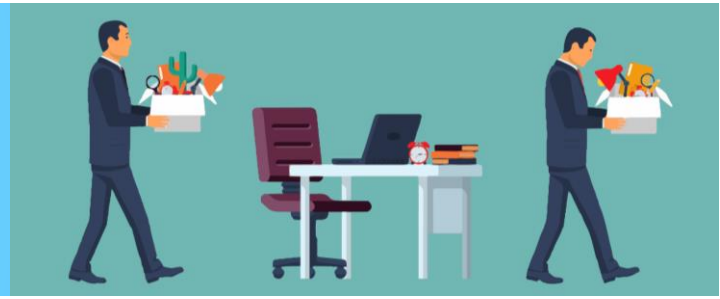
2. Definição de Períodos

Como Definir os Períodos de Cada Variável?

Durante a fase de planejamento de um modelo, deve-se definir as variáveis explicativas que serão testadas, bem como a variável resposta. Além disso, é necessário estabelecer os **períodos de referência** em que cada variável será calculada, com base no interesse de negócio.

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



Adaptado a partir de: https://www.kaggle.com/code/harikrishna9/why-employees-are-leaving/data

Variáveis disponíveis:

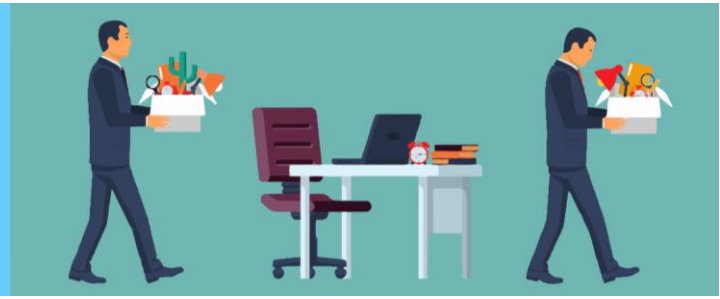
ult_avaliacao_clima	Nível de satisfação do funcionário, na pesquisa de clima do último trimestre
qtde_projetos_Xm	Quantidade de projetos desenvolvidos nos últimos **3 meses, 6 meses, 12 meses e 24 meses**
media_horas_trabalho_Xm	Quantidade média de horas trabalhadas por mês, nos últimos **3 meses, 6 meses e 12 meses**
tempo_empresa	Tempo na empresa, em anos
flag_promoção_Xm	Indica se recebeu promoção (1) ou não (0) nos últimos **3 meses, 6 meses e 12 meses e 'na vida'**
departamento	Área de atuação na empresa
patamar_salario	Patamar salarial (abaixo, próximo ou acima da média, para o cargo ocupado)
turnover	Indica se houve desligamento voluntário (1) ou não (0) nos 6 meses seguintes

Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

FIA BUSINESS SCHOOL lab.data

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



Temos dados disponíveis para **2 safras** (“fotografias”) distintas da base de dados de funcionários ativos.

Para cada safra, vamos denominar o mês de referência como **M0**:

- ✓ **Safra 1:** M0 = janeiro/2023 (último dia do mês);
- ✓ **Safra 2:** M0 = julho/2023 (último dia do mês).

Podemos denominar os meses anteriores à referência como **M-1, M-2, ...** e os meses posteriores como **M+1, M+2, ...**:

- ✓ **Safra 1:** M-1 = dezembro/2022; M-2 = novembro/2022; ... **/** M+1 = fevereiro/2023; M+2 = março/2023; ...
- ✓ **Safra 2:** M-1 = junho/2023; M-2 = maio/2023; ... **/** M+1 = agosto/2023; M+2 = setembro/2023; ...

Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

FIA BUSINESS SCHOOL lab.data

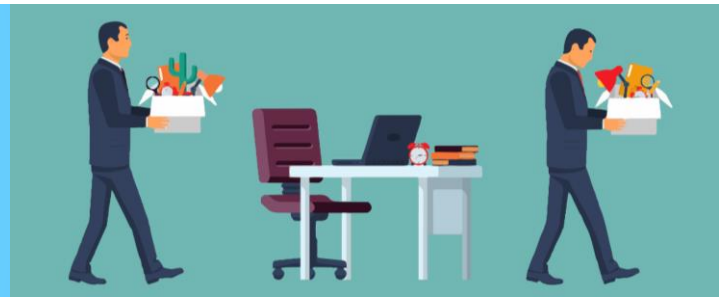
Previsão *versus* Histórico

Quando o modelo tem objetivo **preditivo**, tal como no exemplo de *turnover* de funcionários, é fundamental que a variável resposta seja calculada em um período **posterior e distinto** do período de cálculo das variáveis explicativas. Dessa forma, garantimos que o modelo possa ser aplicado para predizer o que ocorrerá no futuro, em vista do que já se conhece do passado.

FIA BUSINESS SCHOOL lab.data

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



	Histórico						Previsão					
	...	M-4	M-3	M-2	M-1	M0	M+1	M+2	M+3	M+4	M+5	M+6
Safra 1	...	set/22	out/22	nov/22	dez/22	**jan/23**	fev/23	mar/23	abr/23	mai/23	jun/23	jul/23
Safra 2	...	mar/23	abr/23	mai/23	jun/23	**jul/23**	ago/23	set/23	out/23	nov/23	dez/23	jan/24

Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

FIA BUSINESS SCHOOL lab.data

Modelos de Diagnóstico

Por outro lado, há situações em que o modelo cumpre apenas um papel **diagnóstico** em vez de preditivo, de forma a descrever as relações observadas entre variável resposta e as explicativas no histórico e gerar conhecimento acerca do fenômeno de interesse. Neste caso, **não há um período de previsão**.

Exemplo: Predição de preços de imóveis.

Variável	Período de cálculo
Idade do imóvel, em anos	**M0**
Distância do imóvel até a estação de metrô mais próxima, em km	**M0**
Quantidade de comércios próximos ao imóvel	**M0**
Preço do imóvel, em milhares de reais/m²	**M0**

Neste problema, modelamos o **preço atual** de um imóvel em vista de outras **características atuais** do mesmo imóvel. Não geramos um modelo que visa predizer algo para o futuro (ex.: qual será o preço do imóvel daqui a 1 ano), mas sim uma ferramenta diagnóstica a partir da qual o preço dos imóveis possa ser explicado a partir de outras características observadas **ao mesmo tempo**.

Seleção de Períodos por Poder Preditivo

A definição do período histórico para cálculo de cada variável explicativa pode ser realizada com base no **maior poder preditivo** demonstrado por elas em diferentes janelas temporais. Ou seja, pode-se escolher o período que **maximiza a relação** entre cada variável explicativa e a variável resposta.

Information Value (IV)

O **valor da informação**, ou ***information value* (IV)**, é um índice que mensura o grau de associação entre duas variáveis qualitativas.

No contexto da regressão logística, o IV é calculado entre as variáveis explicativas e a variável resposta binária, na fase preliminar de análise exploratória, a fim de identificar quais variáveis explicativas têm **maior potencial de discriminância** no modelo.

Information Value (IV)

Para calcular o índice IV de uma variável, ela deve ser subdividida em **categorias**. Caso ela seja quantitativa, esse processo é realizado automaticamente nos *softwares*.

A fórmula do índice IV é dada por:

$$IV = \sum log\left(\frac{\%\, Y = 1}{\%\, Y = 0}\right) \cdot [(\%\, Y = 1) - (\%\, Y = 0)]$$

onde:

- $\%\, Y = 0$ representa a porcentagem de concentração de **valores zero** na variável resposta em determinada categoria, em relação ao total de zeros.
- $\%\, Y = 1$ representa a porcentagem de concentração de **valores um** na variável resposta em determinada categoria, em relação ao total de zeros.

A somatória é realizada sobre as categorias da variável explicativa.

Information Value (IV)

Quanto **maior** o IV de uma variável explicativa, **maior** o grau de associação entre ela e a variável resposta.

A tabela a seguir propõe uma interpretação acerca do grau de associação para diferentes patamares de IV:

Valor	Associação
$IV \geq 0{,}50$	Excelente
$0{,}30 \leq IV < 0{,}50$	Forte
$0{,}10 \leq IV < 0{,}30$	Média
$0{,}02 \leq IV < 0{,}10$	Fraca
$IV < 0{,}02$	Muito fraca

Information Value (IV)

Exemplo de variável qualitativa com associação **forte** a partir do IV:

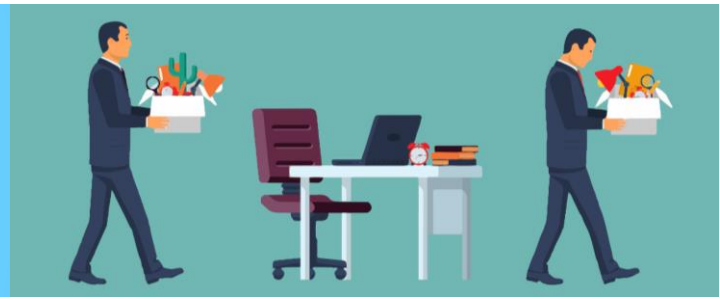
Quantidade de atrasos anteriores do cliente	Cliente pagou em dia a dívida atual?					WOE	IV
	Não (Y = 0)		Sim (Y = 1)		Total		
Nenhum atraso	100	10%	400	40%	500	0,60	0,18
1 atraso	200	20%	300	30%	500	0,18	0,02
2 atrasos	300	30%	200	20%	500	-0,18	0,02
3+ atrasos	400	40%	100	10%	500	-0,60	0,18
Total	**1.000**	**100%**	**1.000**	**100%**	**2.000**	-	**0,40**

Exemplo de variável qualitativa com associação **muito fraca** a partir do IV:

Idade do cliente	Cliente pagou em dia a dívida atual?					WOE	IV
	Não (Y = 0)		Sim (Y = 1)		Total		
18 a 30 anos	400	40%	380	38%	780	-0,02	0,0004
31 a 45 anos	300	30%	320	32%	620	0,03	0,0006
46 a 60 anos	200	20%	170	17%	370	-0,07	0,0021
61+ anos	100	10%	130	13%	230	0,11	0,0034
Total	**1.000**	**100%**	**1.000**	**100%**	**2.000**	-	**0,007**

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



Para as variáveis que possuem diferentes janelas temporais de cálculo, podemos selecionar o período que fornece **maior *information value* (IV)** em relação à variável resposta.

qtde_projetos_*X*m	
Período	**IV**
3 meses	0,003
6 meses	0,008
12 meses	0,003
24 meses	**0,010**

Referência: safra de janeiro/2023.

media_horas_trabalho_*X*m	
Período	**IV**
3 meses	0,006
6 meses	**0,015**
12 meses	0,011

Referência: safra de janeiro/2023.

flag_promoção_*X*m	
Período	**IV**
3 meses	**0,026**
6 meses	0,00004
12 meses	0,00006
'na vida'	0,008

Referência: safra de janeiro/2023.

Caso se tratasse de uma **regressão linear**, poderíamos substituir o IV pelo **coeficiente de correlação linear**.

Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

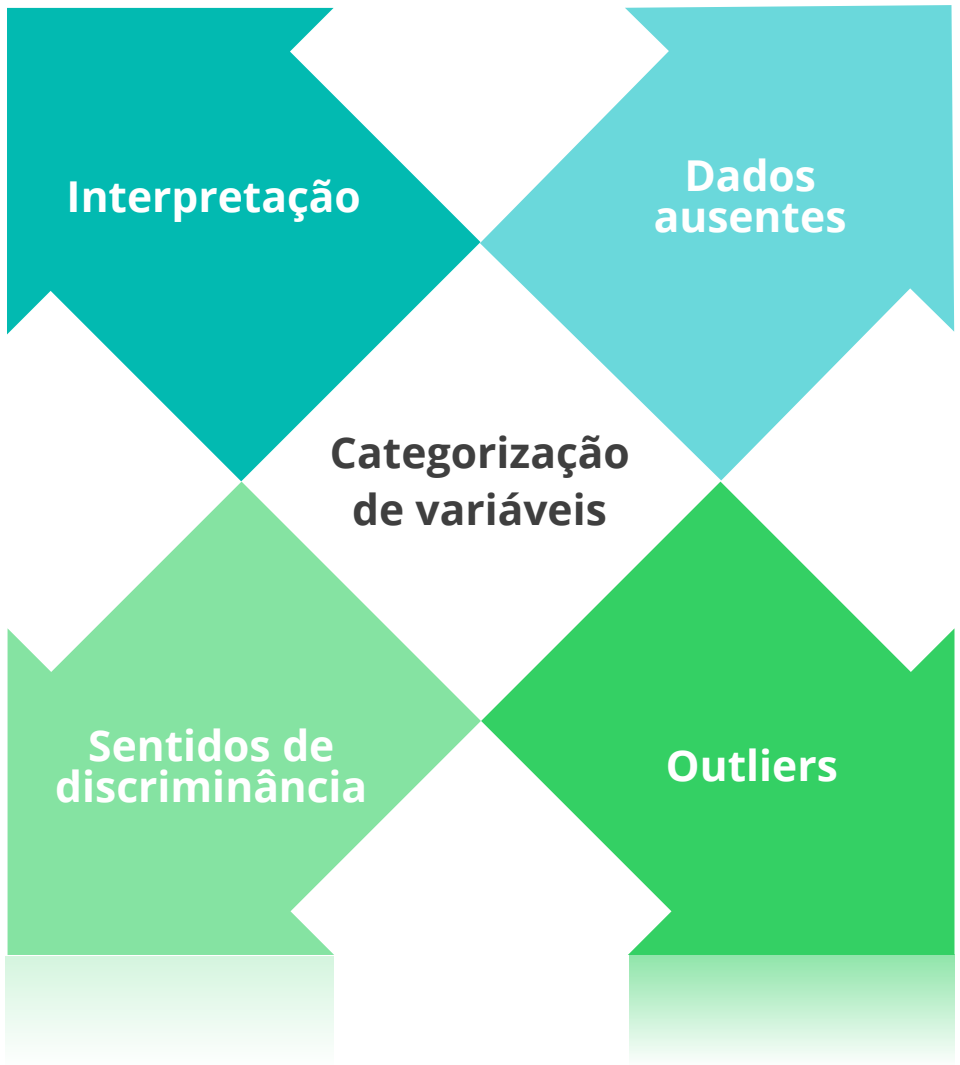
FIA BUSINESS SCHOOL lab.data

3. Categorização de Variáveis

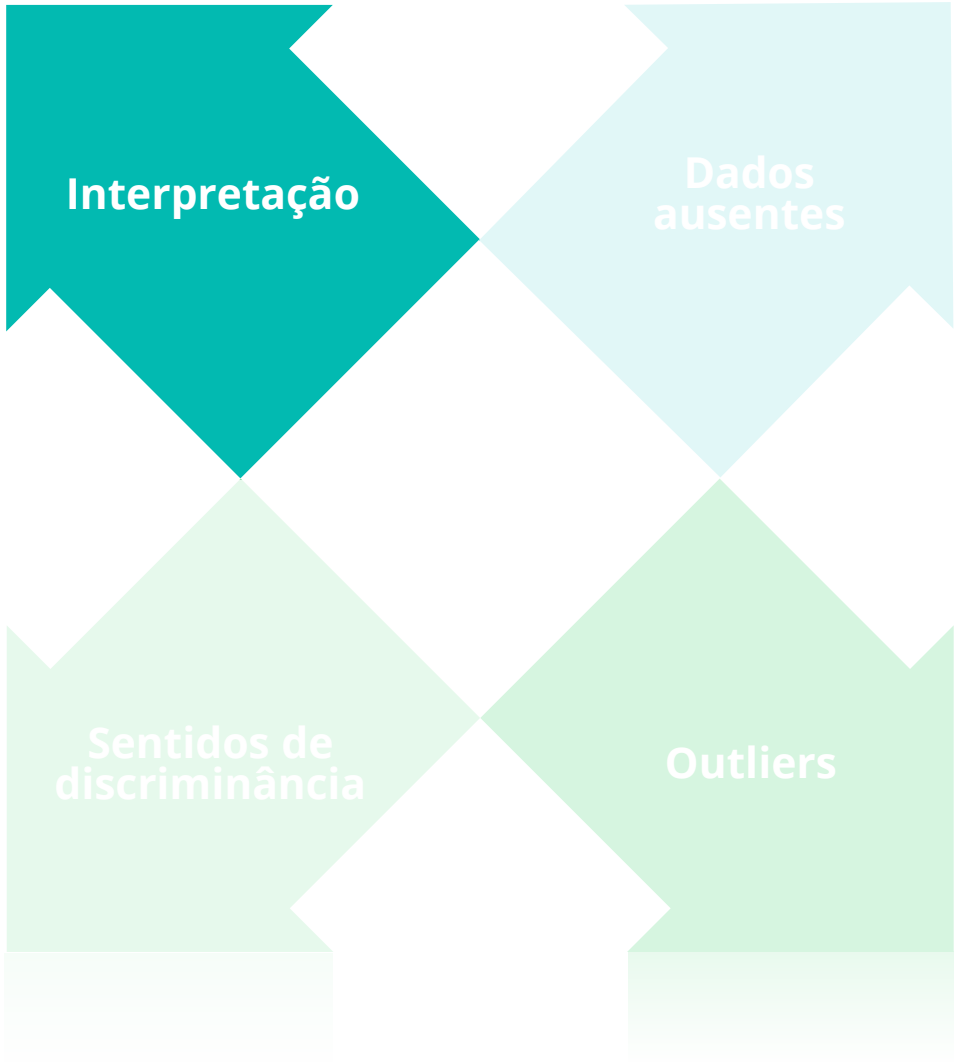
Por Que Categorizar uma Variável Quantitativa?

Com frequência, é comum que se realize um processo de **categorização prévia** das variáveis quantitativas candidatas a compor modelos de regressão. Essa categorização pode trazer algumas vantagens para a qualidade da modelagem e interpretação dos resultados.

Por Que Categorizar uma Variável Quantitativa?



Por Que Categorizar uma Variável Quantitativa?



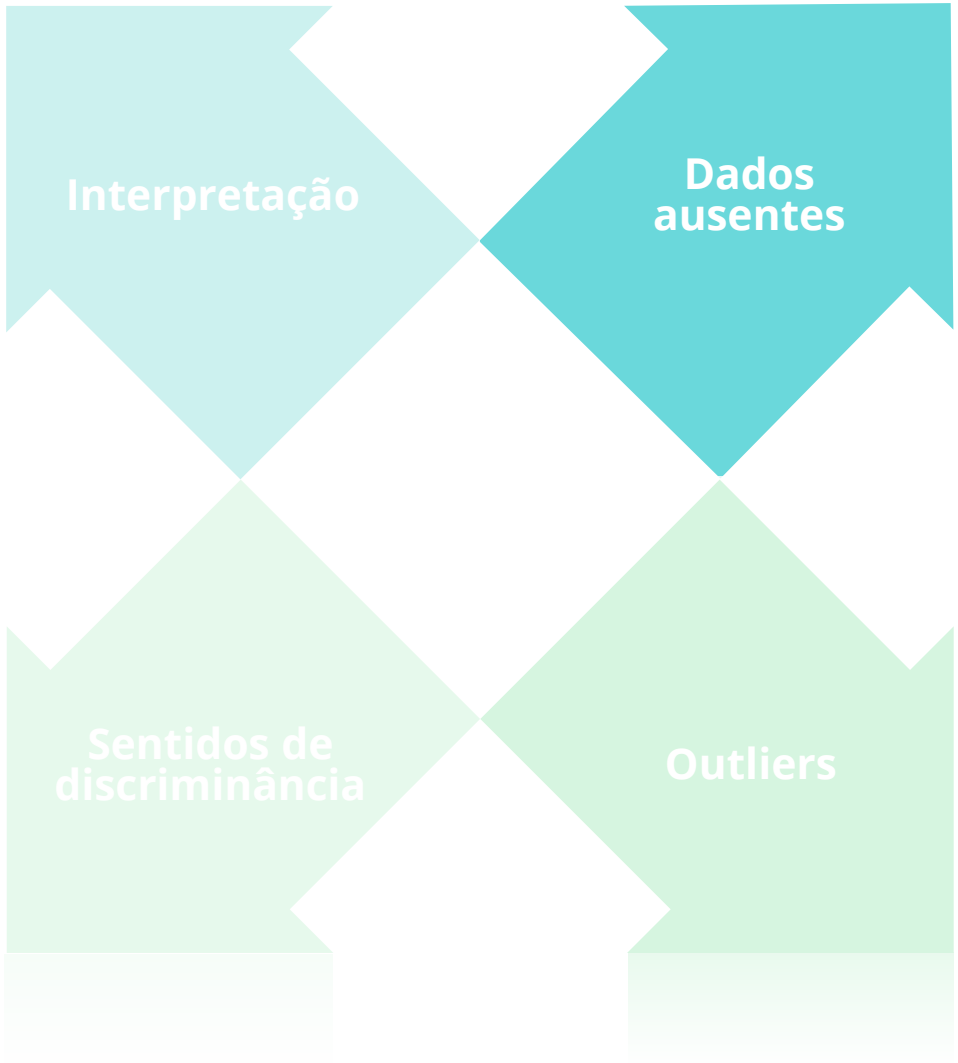
Em algumas situações, a interpretação dos **parâmetros estimados** por modelos de regressão é facilitada quando as variáveis quantitativas são categorizadas.

Exemplo na regressão linear:

- **Sem categorização**: *A cada 1 unidade de incremento na variável explicativa X, incrementa-se, em média, β unidades na resposta Y, mantidas fixas as demais variáveis.*
- **Com categorização**: *Ao passar de A para B na variável explicativa X, incrementa-se, em média, β unidades na resposta Y, mantidas fixas as demais variáveis.*

FIA BUSINESS SCHOOL lab.data

Por Que Categorizar uma Variável Quantitativa?

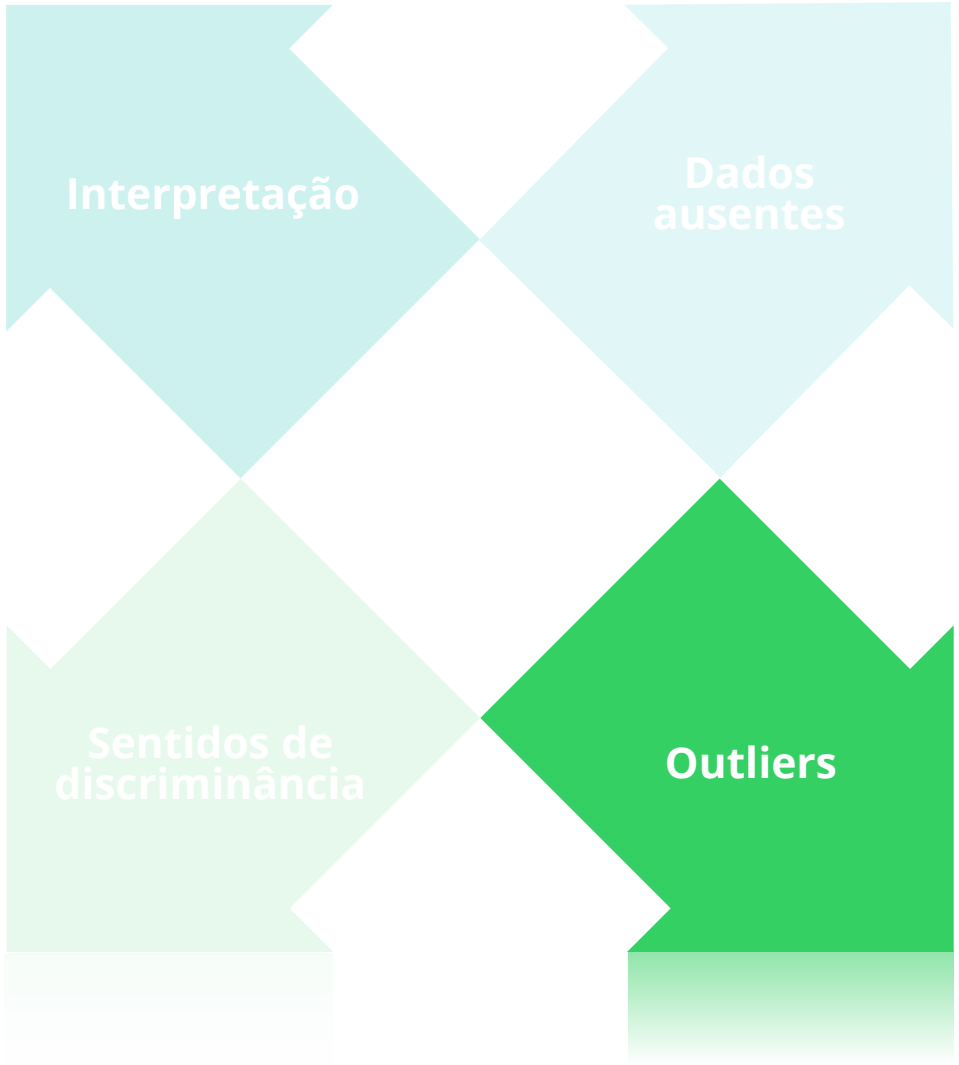


Infelizmente, os modelos de regressão (linear ou logística) não conseguem lidar com **valores ausentes** nas variáveis. Quaisquer observações (linhas) que apresentem ausência de informação para ao menos 1 variável (coluna) são automaticamente **desconsideradas** da modelagem.

Já com a categorização, as observações com valores ausentes são representadas por meio de uma categoria apropriada, tal como *"Sem informação"*.

Note que a ausência de informação pode possuir caráter **explicativo** acerca da variável resposta. Exemplo: clientes com tempo desde a última reclamação = *ausente* são menos propensos ao cancelamento de um serviço.

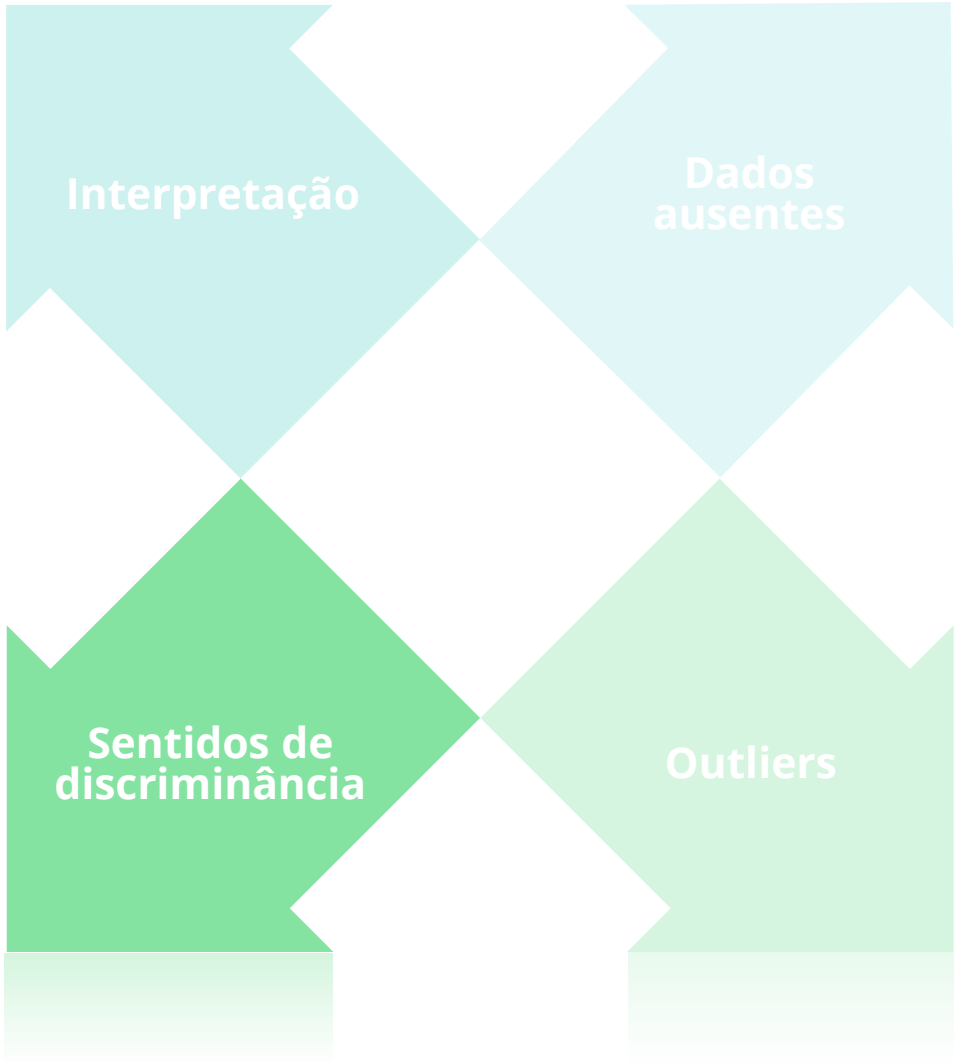
Por Que Categorizar uma Variável Quantitativa?



A ocorrência de **valores atípicos** (*outliers*) nas variáveis explicativas quantitativas pode afetar as estimativas dos parâmetros, tanto na regressão linear quanto na regressão logística.

Já quando as variáveis são categorizadas, os *outliers* são naturalmente englobados na categoria **inferior** (com valores mais baixos) ou **superior** (com valores mais altos). Dessa forma, não exercem influência exacerbada no processo de estimação.

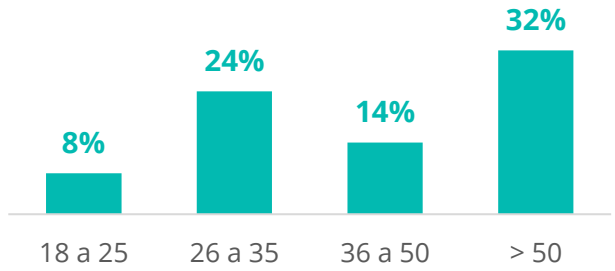
Por Que Categorizar uma Variável Quantitativa?



Por fim, a estimação de um parâmetro β único para uma variável explicativa quantitativa pode ser algo problemático, especialmente em casos de **oscilação no padrão de discriminância** desta variável ao longo de sua escala.

Exemplo na regressão logística:

% de cancelamento de seguro, por faixa etária

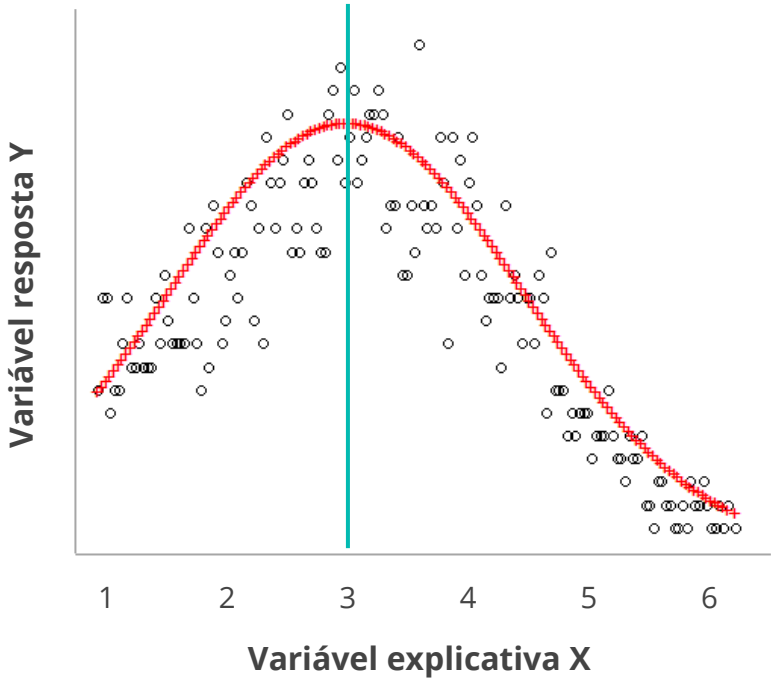


Não há uma tendência única de crescimento/decrescimento da propensão ao cancelamento à medida que a idade aumenta; o padrão é de **alternância**.

Como Categorizar as Variáveis Quantitativas?

No caso da **regressão linear**, em que a variável resposta Y é quantitativa, uma **inspeção gráfica** costuma ser suficiente para avaliar quais categorias devem ser formadas na variável explicativa X.

Exemplo:



Categorização sugerida:

Categoria 1: $X \leq 3$

Categoria 2: $X > 3$

Como Categorizar as Variáveis Quantitativas?

Já no caso da **regressão logística**, é possível dividir a variável explicativa X com base em grupos de mesmo tamanho (percentis) e examinar as **variações na taxa de resposta** por categoria.

Exemplo:

3 faixas da variável X	Taxa de resposta (Y = 1)
Faixa (tercil) 1	32%
Faixa (tercil) 2	14%
Faixa (tercil) 3	9%

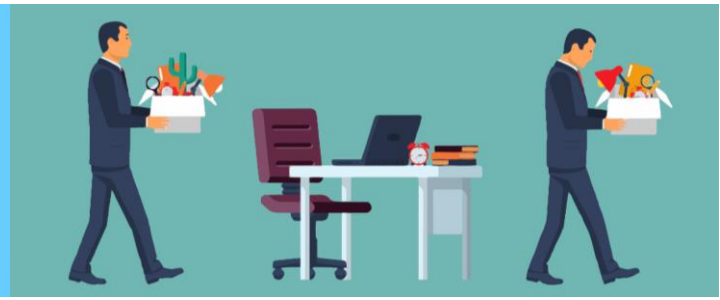
4 faixas da variável X	Taxa de resposta (Y = 1)
Faixa (quartil) 1	35%
Faixa (quartil) 2	25%
Faixa (quartil) 3	13%
Faixa (quartil) 4	8%

5 faixas da variável X	Taxa de resposta (Y = 1)
Faixa (quintil) 1	36%
Faixa (quintil) 2	**28%**
Faixa (quintil) 3	**26%**
Faixa (quintil) 4	14%
Faixa (quintil) 5	7%

Categorização sugerida: em quatro faixas (quartis), pois ao dividir em cinco faixas, começamos a ter categorias com taxas de resposta semelhantes.

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



Realizando uma análise exploratória breve da safra de janeiro/2023, vemos que a variável ***ult_avaliação_clima*** é a única que possui valores ausentes. Logo, precisamos categorizá-la, a fim de que os funcionários que não responderam a última pesquisa de clima também sejam considerados no modelo.

```
ult_avaliacao_clima
Min.   :0.0000
1st Qu.:0.3900
Median :0.6900
Mean   :0.6218
3rd Qu.:0.8800
Max.   :1.0000
NA's   :217
```

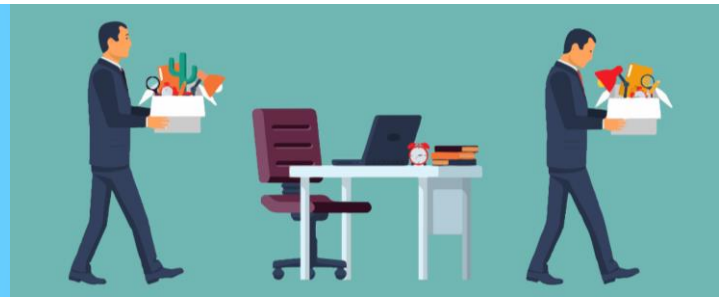
217 valores ausentes
(5,6% dos registros da safra de jan/23)

Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

FIA BUSINESS SCHOOL lab.data

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



Avaliando cenários de categorização da variável ***ult_avaliação_clima*** em 3, 4, 5 e 6 faixas (além dos valores ausentes), a divisão em **4 faixas** parece ser suficiente.

	0	1
[0,0.51]	0.7821138	0.2178862
(0.51,0.83]	0.8221504	0.1778496
(0.83,1]	0.8717300	0.1282700
<NA>	0.8248848	0.1751152

	0	1
[0,0.39]	0.7744035	0.2255965
(0.39,0.69]	0.8110497	0.1889503
(0.69,0.88]	0.8426230	0.1573770
(0.88,1]	0.8714286	0.1285714
<NA>	0.8248848	0.1751152

	0	1
[0,0.31]	0.7720000	0.2280000
(0.31,0.59]	0.7991632	0.2008368
(0.59,0.78]	0.8147651	0.1852349
(0.78,0.92]	0.8777633	0.1222367
(0.92,1]	0.8614009	0.1385991
<NA>	0.8248848	0.1751152

	0	1
[0,0.245]	0.7635468	0.2364532
(0.245,0.51]	0.8003221	0.1996779
(0.51,0.69]	0.8140704	0.1859296
(0.69,0.83]	0.8296875	0.1703125
(0.83,0.93]	0.8780069	0.1219931
(0.93,1]	0.8656716	0.1343284
<NA>	0.8248848	0.1751152

Clientes que não responderam a última pesquisa de clima possuem risco intermediário de *turnover*

Cenários com 5 e 6 faixas apresentam categorias com taxas de resposta muito semelhantes

Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

FIA BUSINESS SCHOOL lab.data

4. Validação de Modelos

Averiguando a Estabilidade do Modelo

Ao desenvolver um modelo preditivo baseado numa **amostra** dos indivíduos de interesse, duas questões essenciais que devemos averiguar são:

1. A qualidade do modelo se manteria **estável** ao generalizá-lo para outras observações, no **mesmo período** de tempo?
2. A qualidade do modelo se manteria **estável** ao generalizá-lo para futuras observações, em **novos períodos** de tempo?

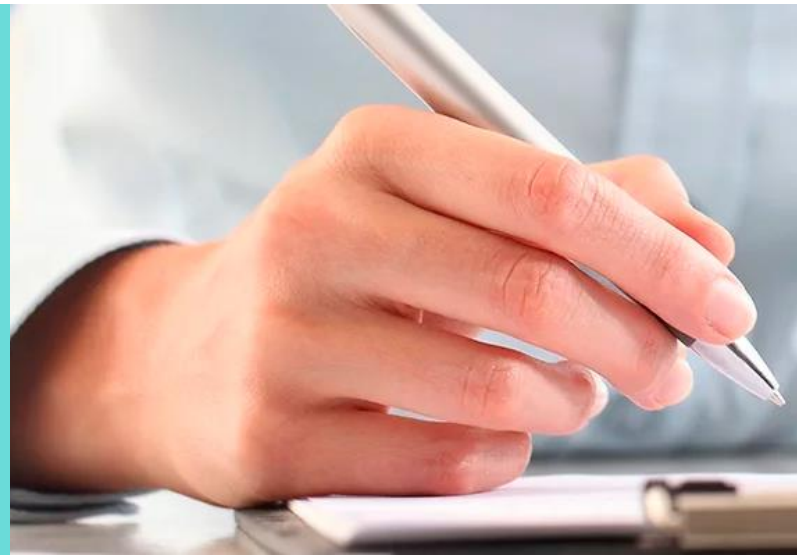
Averiguando a Estabilidade do Modelo

Ao desenvolver um modelo preditivo baseado numa **amostra** dos indivíduos

VAMOS COMEÇAR POR ESTE PROCESSO

ue devemos averiguar são:

1. A qualidade do modelo se manteria **estável** ao generalizá-lo para outras observações, no **mesmo período** de tempo?
2. A qualidade do modelo se manteria **estável** ao generalizá-lo para futuras observações, em **novos períodos** de tempo?

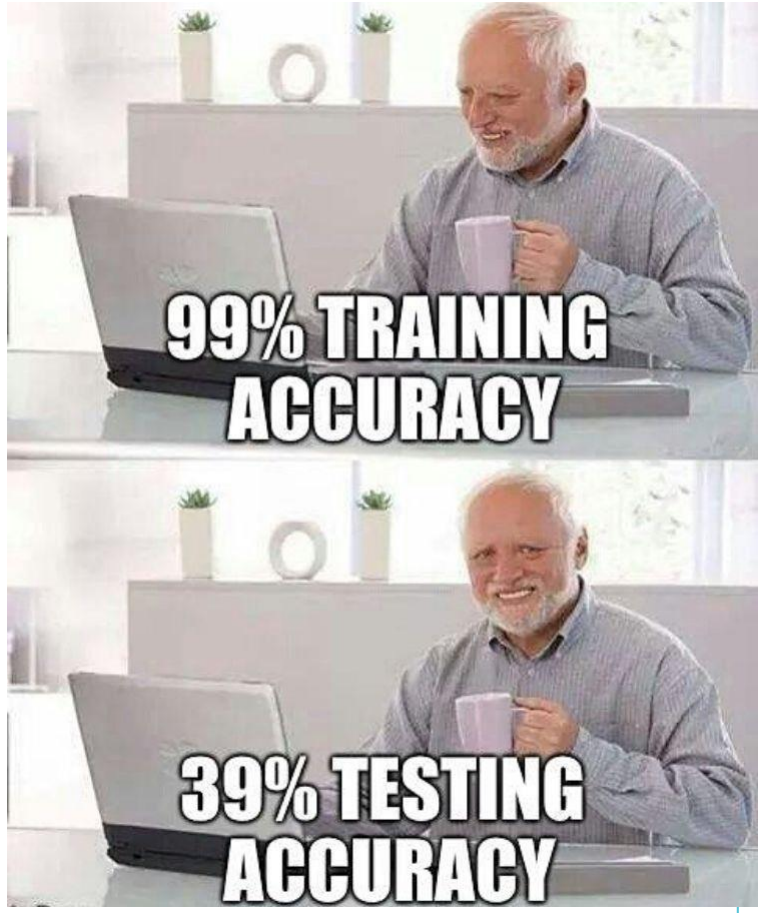


Problema de Superajuste

Quando um modelo de regressão fica extremamente **bem ajustado** (ou **superajustado**) aos dados da amostra, sua capacidade de extrapolação para novas observações da população torna-se **pobre**.

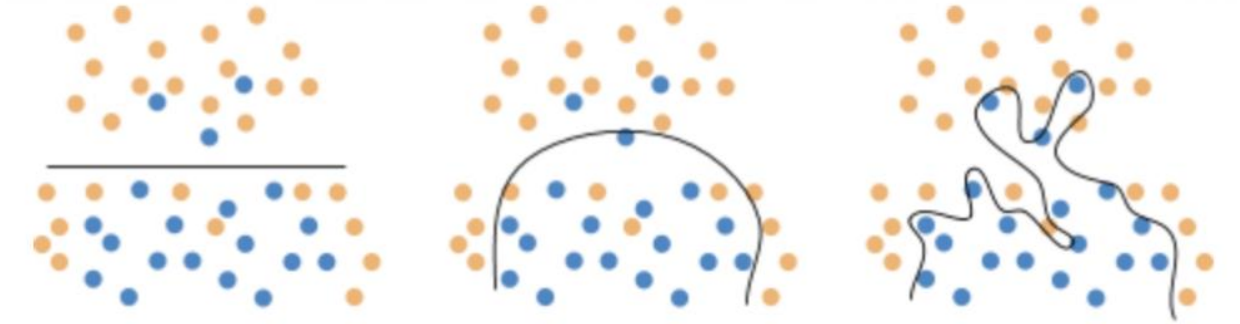
Nessas situações, o alto desempenho obtido na amostra é **artificial**, pois, na realidade, o algoritmo não conseguiu capturar padrões genéricos acerca do fenômeno de interesse, mas sim, padrões **extremamente específicos** dos dados disponíveis.

Superajuste: Memes



Superajuste: Representação Visual

Meta: discriminar pontos a partir de suas cores.



Subajuste *(underfitting)*	**Modelo ideal**	**Superajuste** *(overfitting)*

Conjuntos de Treino e Teste

Estratégia: deixar previamente reservada uma **parte da base amostral** para **testarmos o modelo**.

Amostra COMPLETA
(100% das observações)

Conjunto de TREINO
(70 a 80% das observações)

Conjunto de TESTE
(20 a 30% das observações)

Desenvolvemos as regras do modelo neste conjunto e verificamos o seu desempenho.

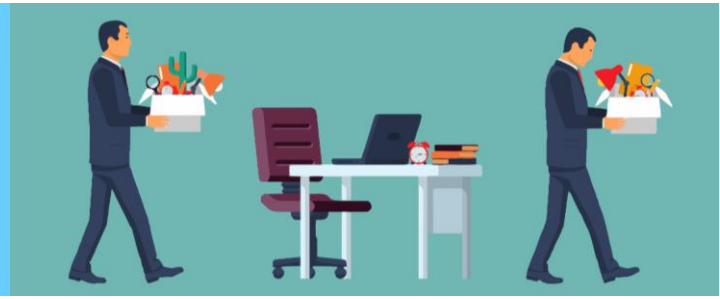
Aplicamos neste conjunto o modelo desenvolvido no treino, para verificar se o desempenho se mantém estável; se houver queda expressiva, há indício de superajuste.

Considerações:

- Os conjuntos de treino e teste devem ser definidos de forma **aleatória** (com base num sorteio), a fim de serem comparáveis e evitar quaisquer vieses.
- O desempenho do algoritmo no conjunto de **teste** é a **referência principal** para julgar a qualidade do algoritmo, pois mensura a capacidade real que o algoritmo tem de extrapolar as predições para novos dados.
- Em bases de dados pequenas, a divisão de conjuntos de treino e teste pode ser inviável.

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



A base de dados referente a janeiro/2023 foi adotada para construção do modelo, e foi subdividida aleatoriamente entre **treino** e **teste**:

- **2.708** observações para treino (**70%**)
- **1.161** observações para teste (**30%**)

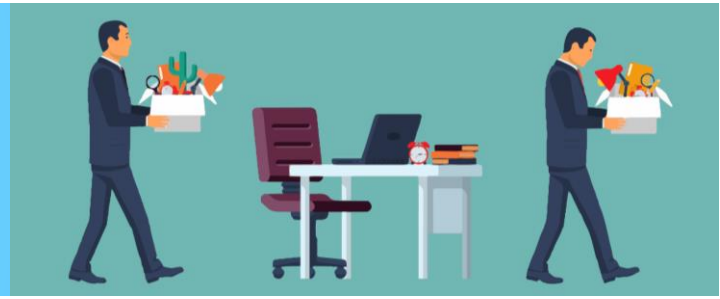
A subdivisão entre os conjuntos foi realizada fixando a semente aleatória 12345 no R. A semente tem o papel de controlar os processos que envolvem aleatoriedade, garantindo **reprodutibilidade** (ou seja, obtenção dos mesmos resultados) quando o código é executado em diferentes momentos ou diferentes computadores.

Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

FIA BUSINESS SCHOOL lab.data

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



Seguindo o processo de seleção de variáveis via *stepwise backward*, as variáveis explicativas estatisticamente significativas para predizer a variável resposta no conjunto de treino, com 95% de confiança, são:

- *departamento*
- *patamar_salario*
- *tempo_empresa*
- *ult_avaliacao_clima_**cat*** (categorizada por conta dos *missing values*)
- *flag_promocao_3m*

Não há indício de colinearidade entre as variáveis explicativas, visto que os índices VIF obtidos são próximos de 1.

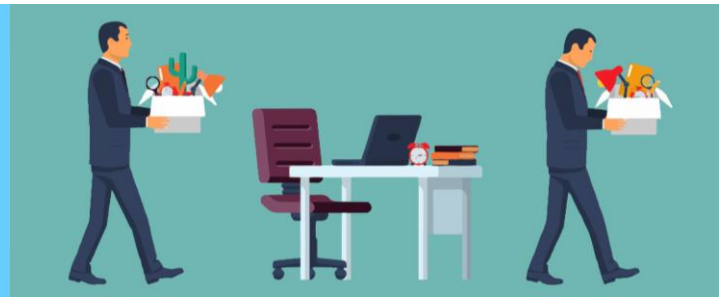
Obs.: para *dummies* referentes a uma mesma variável qualitativa, é natural que os VIF possam ser mais altos, tal como ocorre para a variável *departamento* neste *case*.

Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

FIA BUSINESS SCHOOL lab.data

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



DESEMPENHO DO MODELO: JAN/23

Ponto de corte = **17,9%** (proporção de indivíduos com *turnover* no conjunto de treino)

Dados	Acurácia	Sensibilidade	Especificidade	KS	AUC
Treino	**63,1%**	**61,2%**	**71,9%**	**33,7**	**71,8**
Teste	**60,6%**	**59,3%**	**67,0%**	**28,2**	**68,4**

Não houve queda expressiva de desempenho no conjunto de teste, o que indica que o modelo **não está superajustado**.

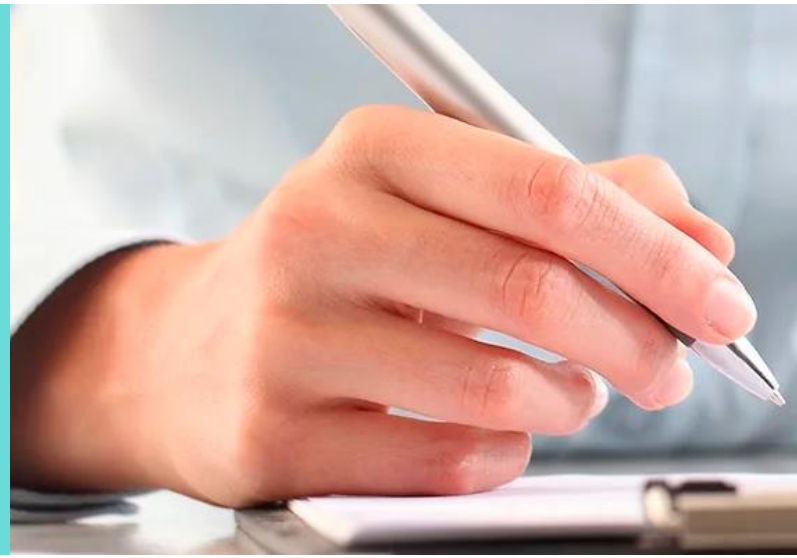
Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

FIA BUSINESS SCHOOL lab.data

Averiguando a Estabilidade do Modelo

Ao desenvolver um modelo preditivo baseado numa **amostra** dos indivíduos de interesse, duas questões essenciais que devemos averiguar são:

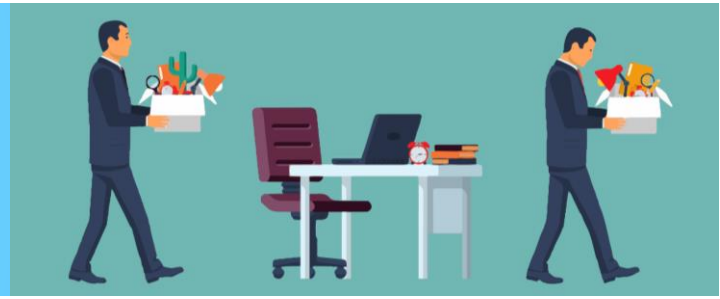
1. A qualidade do modelo se manteria estável ao generalizá-lo para outras observações, no mesmo período de tempo?
2. A qualidade do modelo se manteria **estável** ao generalizá-lo para futuras observações, em **novos períodos** de tempo?



FIA BUSINESS SCHOOL lab.data

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



Para avaliar a estabilidade preditiva do modelo **ao longo do tempo**, vamos simplesmente aplicá-lo na próxima fotografia disponível, referente aos funcionários ativos na safra de **julho/23**.

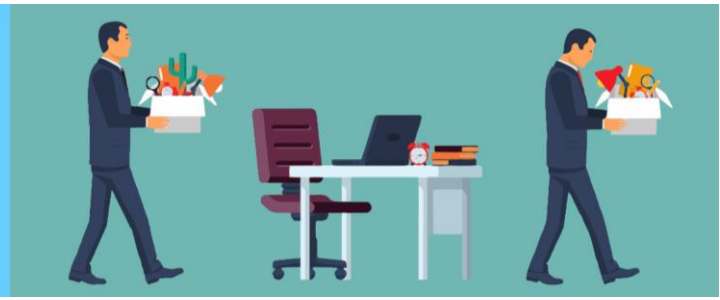
Este processo é chamado de **validação *out-of-time*** do modelo.

Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

FIA BUSINESS SCHOOL lab.data

Case: *Turnover* de Funcionários

A área de Recursos Humanos de uma grande empresa deseja entender quais perfis de funcionários são mais propensos a **solicitarem desligamento** *(turnover)*.
A empresa detém dados de 3.869 funcionários que estavam ativos em janeiro/2023; e 4.486 funcionários que estavam ativos em julho/2023.



DESEMPENHO DO MODELO: JUL/23

Ponto de corte = **17,9%** (proporção de indivíduos com *turnover* no conjunto de treino)

Dados	Acurácia	Sensibilidade	Especificidade	KS	AUC
Teste: jan/23	**60,6%**	**59,3%**	**67,0%**	**28,2**	**68,4**
Validação: jul/23	**61,5%**	**59,8%**	**68,8%**	**29,0**	**69,1**

O desempenho do modelo de manteve **estável** na safra de validação *out-of-time*, tal como desejado.

Arquivos: Turnover_Funcionarios_Jan_23 e ..._Jul_23 (.txt)

FIA BUSINESS SCHOOL lab.data

Referências Bibliográficas

TÓPICOS DE MODELAGEM

- James, G. *An Introduction to Statistical Learning - With Applications in R*. 2ª edição. Springer, 2021.
- Morettin, P. A., Singer, J. M. *Estatística e Ciência de Dados*. 1ª edição. LTC, 2022.

lab.data

http://labdata.fia.com.br
Instagram: @labdatafia
Facebook: @LabdataFIA