# Analytics e Inteligência Artificial Data Science

Tema da aula
**Regressão Logística**


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



FIA 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
2. Objetivo
3. Regressão Logística Simples
   - Função Logística
   - Probabilidade vs. Chance
4. Regressão Logística Múltipla
5. Análise de Desempenho
   - Acurácia
   - Sensibilidade e Especificidade
   - Estatística KS
   - Área Abaixo da Curva ROC (AUC)
6. *Cases* Adicionais
   - *Credit Score*

Referências Bibliográficas

# 1. Introdução



# *Case:* Fraude em Cartão de Crédito



**Exemplo:**
Identificar a probabilidade de uma transação de cartão de crédito representar uma fraude, com base em suas características: diferença entre o valor da transação e o valor médio usual das transações do cliente, realização em horário atípico, realização em localização atípica, tempo de casa do cliente etc.

**Aplicação:**
Cartão de crédito



FIA BUSINESS SCHOOL lab.data

# *Case:* Sinistro em Seguro



**Exemplo:**
Identificar a probabilidade de um indivíduo sofrer um sinistro, com base em características demográficas e em seu estilo de vida.

**Aplicação:**
Seguradoras

# *Case:* Doenças Cardíacas



**Exemplo:**
Identificar a probabilidade de um paciente apresentar problemas cardíacos, de acordo com seu hábito de vida: quantidade de horas de sono, quantidade de refeições diárias, frequência de consumo de frituras e doces, frequência de exercícios físicos, valor de colesterol total, valor de triglicérides etc.

**Aplicação:**
Área médica



FIA BUSINESS SCHOOL lab.data

# *Case: Downgrade* de Plano



**Exemplo:**
Identificar os clientes com maior propensão a um *downgrade* de seu plano de telefonia, com base em seu perfil transacional recente: quantidade de ligações realizadas, volume de dados consumidos (em GB), quantidade de reclamações na central de atendimento etc.

**Aplicação:**
Telecomunicações

**Exemplo:**
Identificar a probabilidade de um futuro cliente de uma instituição financeira se tornar inadimplente ao adquirir um crédito pessoal *(credit score).*

**Aplicação:**
Segmento bancário

# 2. Objetivo



# Objetivo



Ao contrário da regressão linear, em que o objetivo era predizer uma variável resposta quantitativa, o modelo de **regressão logística** almeja predizer um **evento binário**, a partir de um conjunto de variáveis explicativas.

De forma genérica, um evento binário costuma ser representado por meio de uma variável que assume os seguintes valores:

- Assume **1**, quando ocorre o evento de interesse;
- Assume **0**, caso contrário.

Nesta aula, vamos estudar a teoria por trás da regressão logística e empregá-la em alguns *cases* práticos.

# 3. Regressão Logística Simples



Tal como a regressão linear simples, a regressão logística simples envolve apenas **uma variável explicativa**. Novamente, nossa notação será:

- $\boldsymbol{Y}$: variável resposta, de natureza binária ($Y = 0$ ou $Y = 1$).
- $\boldsymbol{X}$: variável explicativa, que pode ser tanto quantitativa quanto qualitativa.

# Estratégia



Em vez de modelarmos a variável resposta $Y$ de forma direta, o artifício utilizado pelo modelo de regressão logística será o de modelar a **probabilidade** de ocorrência do evento, que varia em uma escala quantitativa.

- $\boldsymbol{P(Y = 1)}$, ou simplesmente $\boldsymbol{p}$, é a probabilidade que $Y$ seja igual a 1, ou seja, a probabilidade de ocorrência do evento. Varia entre 0 (ou 0%) e 1 (ou 100%).

---

Sob a ótica da modelagem estatística, assumimos que o valor da probabilidade $\boldsymbol{p}$ não é o mesmo para todas as observações. Em vez disso, ele varia a depender do valor da **variável explicativa** $X$.

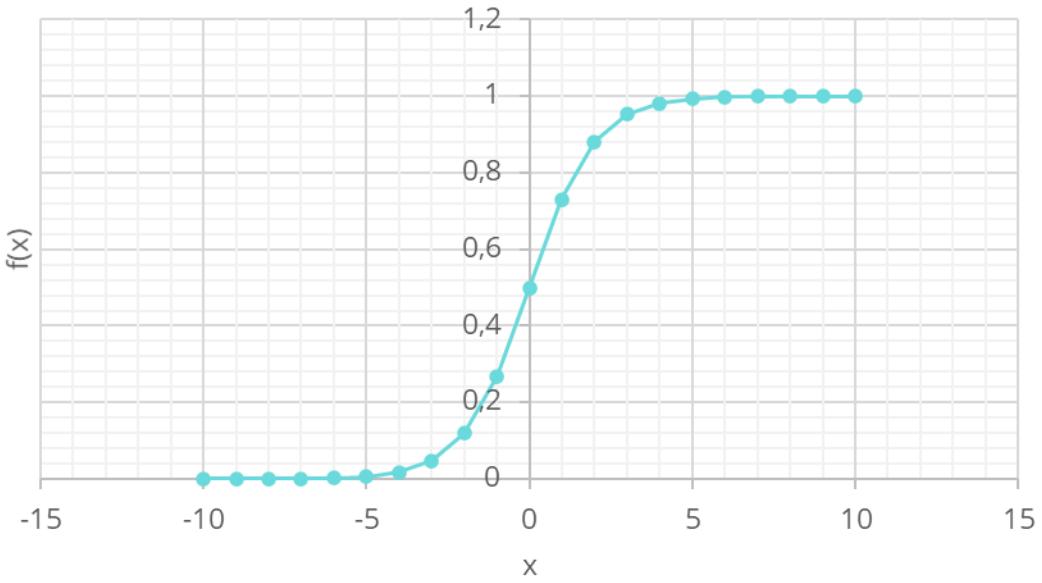
*Exemplo:* no *case* de fraude em cartão de crédito, quanto **maior** a diferença positiva entre o valor de uma transação específica e o valor usual das transações do cliente ($X$), **maior** a probabilidade ($p$) de que se trate de uma fraude.

# Função Logística



A **função logística** é uma função matemática que recebe um valor numérico qualquer e retorna valores no intervalo de 0 a 1.

O gráfico abaixo ilustra o formato da **função logística**. Em vez de ser uma reta, trata-se de uma curva em S.




Fórmula da **função logística**:

$$f(X) = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

- $\boldsymbol{\beta}$ é um valor fixado qualquer
- $\boldsymbol{e}$ é um número convencional na matemática (tal como o número *pi*), chamado **número de Euler**; corresponde aproximadamente a 2,718.



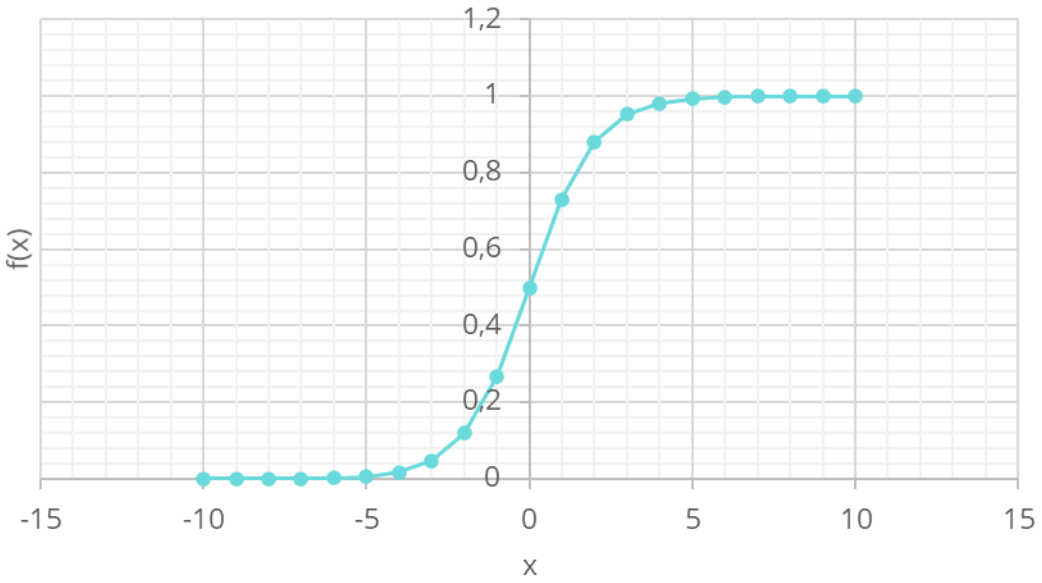
FIA BUSINESS SCHOOL lab.data

# Função Logística



A **função logística** é uma função matemática que recebe um valor numérico qualquer e retorna valores no intervalo de 0 a 1.

O gráfico abaixo ilustra o formato da **função logística**. Em vez de ser uma reta, trata-se de uma curva em S.




Fórmula da **função logística**:

$$f(X) = \frac{e^{0,8 \cdot X}}{1 + e^{0,8 \cdot X}}$$

Quando o coeficiente $\beta$ é **positivo**, o valor de $f(X)$ **aumenta** à medida que $X$ aumenta.

A figura ao lado representa o caso em que $\beta = 0,8$.



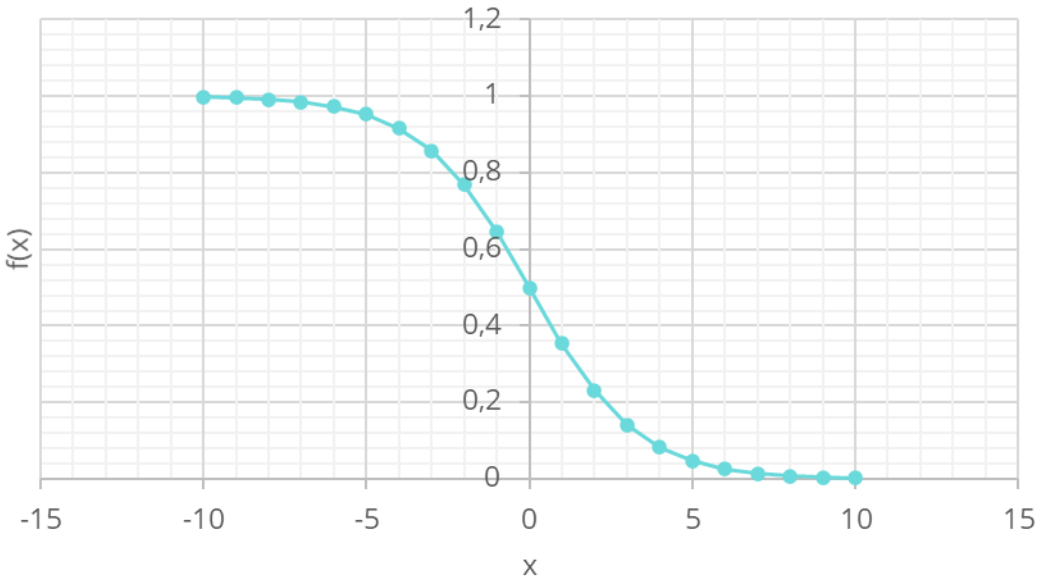
FIA BUSINESS SCHOOL lab.data

# Função Logística



A **função logística** é uma função matemática que recebe um valor numérico qualquer e retorna valores no intervalo de 0 a 1.

O gráfico abaixo ilustra o formato da **função logística**. Em vez de ser uma reta, trata-se de uma curva em S.




Fórmula da **função logística**:

$$f(X) = \frac{e^{-0,6 \cdot X}}{1 + e^{-0,6 \cdot X}}$$

Quando o coeficiente $\beta$ é **negativo**, o valor de $f(X)$ **diminui** à medida que $X$ aumenta.

A figura ao lado representa o caso em que $\beta = -0,6$.

# Modelo de Regressão Logística Simples



A função logística será utilizada para obtermos a **probabilidade** ($p$) associada ao evento de natureza binária ($Y$). Com isso, podemos definir formalmente a equação do **modelo de regressão logística**.

$$p = P(Y = 1) = f(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

A única diferença entre a equação acima e a fórmula usual da função logística é a inclusão de um termo independente $\beta_0$ nos expoentes, chamado **intercepto**. Dessa forma, os parâmetros do modelo são:

- $\boldsymbol{\beta_0}$, que define o patamar natural em torno do qual oscila o valor $p$.
- $\boldsymbol{\beta_1}$, que define o grau de associação (não linear) entre a variável explicativa $X$ e a probabilidade $p$.

  Note que quando $\beta_1 = 0$, o valor de $p$ depende apenas do intercepto: $p = e^{\beta_0}/(1 + e^{\beta_0})$.

# Modelo de Regressão Logística Simples



A função logística será utilizada para obtermos a **probabilidade** ($p$) associada ao evento de natureza binária ($Y$). Com isso, podemos definir formalmente a equação do **modelo de regressão logística**.

$$p = P(Y = 1) = f(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Realizando algumas transformações na equação acima, é possível mostrar a sua **equivalência** com a seguinte equação alternativa:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

# Modelo de Regressão Logística Simples



A função logística será utilizada para obtermos a **probabilidade** ($p$) associada ao evento de natureza binária ($Y$). Com isso, podemos definir formalmente a equação do **modelo de regressão logística**.

$$p = P(Y = 1) = f(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Realizando algumas transformações na equação acima, é possível mostrar a sua **equivalência** com a seguinte equação alternativa:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Esta equação mostra uma íntima relação entre a **regressão logística** e a **regressão linear**.



FIA BUSINESS SCHOOL lab.data

# Probabilidade *vs.* Chance



Podemos definir um conceito alternativo ao de probabilidade, que corresponde à **chance** de ocorrência de um evento binário.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Na equação acima, o termo $\frac{p}{1-p}$ representa a **chance** associada ao evento $Y = 1$.

*Exemplo: case* de fraude em cartão de crédito.

- Suponha que $p = 0{,}10$, ou seja, existe 10% de probabilidade de uma transação representar fraude. Então, a **chance** de fraude é dada por:

$$\frac{p}{1-p} = \frac{0{,}1}{0{,}9} = \frac{1}{9}$$

- Ou seja, para cada **1 fraude**, existem **9 não fraudes**. Costuma-se ler como *"chance de 1 pra 9".*

# Probabilidade *vs.* Chance



Podemos definir um conceito alternativo ao de probabilidade, que corresponde à **chance** de ocorrência de um evento binário.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Na equação acima, o termo $\frac{p}{1-p}$ representa a **chance** associada ao evento $Y = 1$.

*Exemplo: case* de fraude em cartão de crédito.

- Num exemplo extremo oposto, suponha que $p = 0{,}90$, ou seja, existe 90% de probabilidade de uma transação representar fraude. Então, a **chance** de fraude é dada por:

$$\frac{p}{1-p} = \frac{0{,}9}{0{,}1} = 9$$

- Ou seja, para cada **9 fraudes**, existiria **1 não fraude**. Costuma-se ler como *"chance de 9 pra 1".*

# Interpretação do Modelo



Podemos definir um conceito alternativo ao de probabilidade, que corresponde à **chance** de ocorrência de um evento binário.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Podemos dizer que, na regressão logística simples, o **logaritmo da chance de ocorrência do evento** é modelada de acordo com uma **regressão linear**.

Dessa forma, a interpretação dos parâmetros pode ser redefinida como:

- A cada incremento de 1 unidade no valor da variável $X$, **aumenta-se**, em média, $\boldsymbol{\beta_1}$ unidades no **logaritmo da chance do evento**. Ou seja, o impacto de $X$ sobre o logaritmo da chance é linear (aditivo).
- Caso a variável $X$ não afete de forma estatisticamente significativa a probabilidade $p$ (ou seja, $\beta_1 = 0$), então, $\boldsymbol{\beta_0}$ corresponde ao **logaritmo da chance do evento**, que passa a ser um valor fixo.

# Interpretação do Modelo



Podemos definir um conceito alternativo ao de probabilidade, que corresponde à **chance** de ocorrência de um evento binário.

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 X} = e^{\beta_0} e^{\beta_1 X}$$

Eliminando o logaritmo da equação anterior e deixando-a escrita em termos da **chance**, chegamos a esta outra equação equivalente, que sugere uma interpretação adicional para o parâmetro $\beta_1$.

- A cada incremento de 1 unidade no valor da variável $X$, a **chance do evento** é **multiplicada**, em média, por $e^{\beta_1}$ unidades.

Ou seja, o impacto de $X$ sobre a chance não é linear (aditivo), mas sim **multiplicativo**.

# 4. Regressão Logística Múltipla

# Modelo de Regressão Logística Múltipla

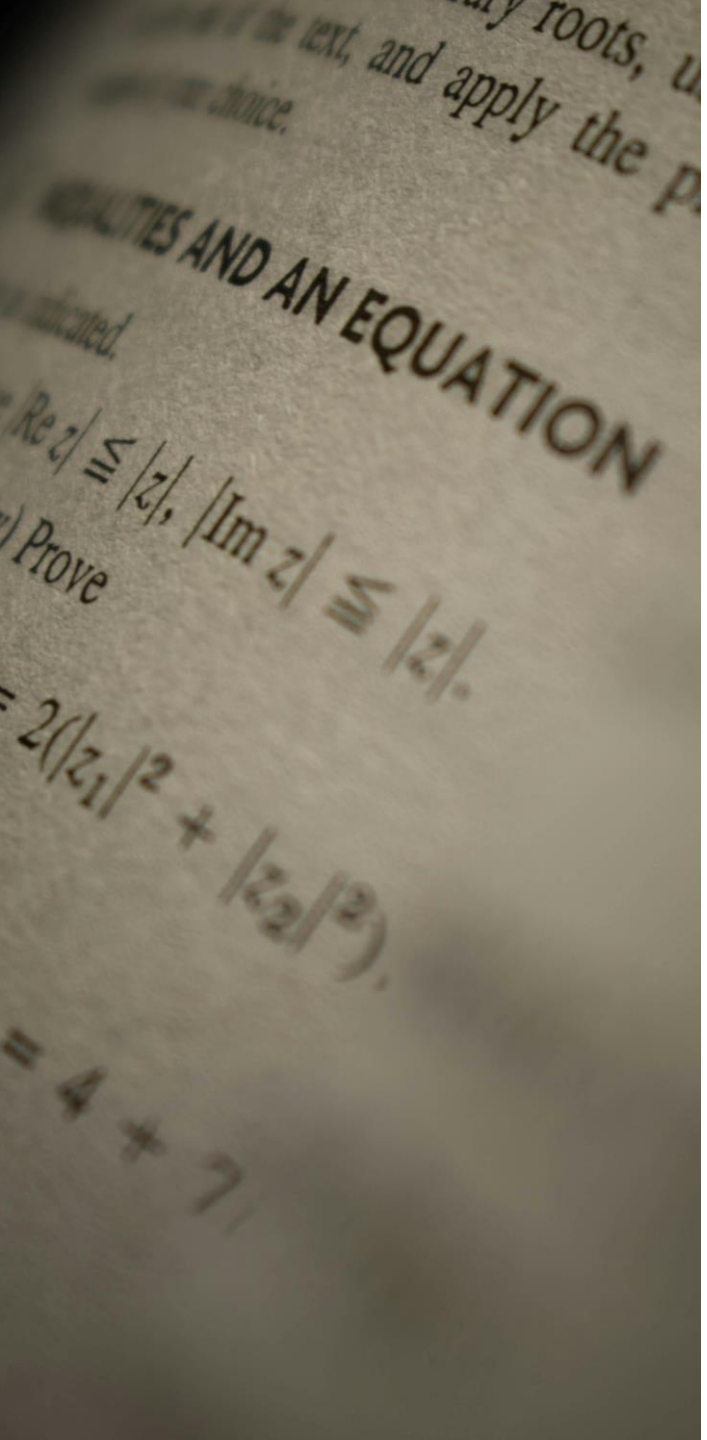


De forma análoga à regressão logística simples, a probabilidade ($p$) no caso **múltiplo,** considerando $k$ variáveis explicativas, pode ser calculada somando novos termos aos expoentes da função logística.

$$p = P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}$$

Realizando algumas transformações na equação acima, é possível mostrar a sua **equivalência** com a seguinte equação alternativa:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$




# Ajuste do Modelo



As expressões que vimos anteriormente correspondem ao **modelo teórico** de regressão logística, no qual os parâmetros $\beta_0, \beta_1, \ldots, \beta_k$ são assumidos **populacionais** e **desconhecidos**.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ \ldots \ + \beta_k X_k$$

Já o **modelo ajustado** a partir de dados **amostrais** nos leva a obter estimativas dos parâmetros, que chamaremos de $b_0, b_1, \ldots, b_k$. Consequentemente, teremos valores **preditos** da probabilidade $p$ para cada indivíduo, que serão denotados como $\hat{p}$.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1 X_1 + \ \ldots \ + b_k X_k$$

# Variáveis Qualitativas



Tal como na regressão linear, as **variáveis qualitativas** são representadas por meio de variáveis *dummy* associadas a cada categoria (exceto uma, que é fixada como referência).

# *Case:* Compra de Perfumes



Um varejista de perfumes estabeleceu uma ação de *marketing* que consiste em fornecer um **cupom** de 20% de desconto para clientes que **baixam o aplicativo** em seu celular, mas não realizam sua primeira compra em até 7 dias após o *download*. Mediante esse estímulo, alguns clientes utilizam o cupom nos 30 dias seguintes (período de validade), e outros, naturalmente, não utilizam. A empresa deseja compreender qual o **perfil dos clientes que utilizam o cupom de desconto**.

**LINHA DO TEMPO DOS CLIENTES SOB INVESTIGAÇÃO**

**D0**

*Download* do aplicativo

**D7**

Concessão do cupom, caso não tenha comprado

PERÍODO PARA AVALIAÇÃO DA VARIÁVEL RESPOSTA (UTILIZAÇÃO DO CUPOM: SIM OU NÃO)

**D37**

Término da validade do cupom

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



Um varejista de perfumes estabeleceu uma ação de *marketing* que consiste em fornecer um **cupom** de 20% de desconto para clientes que **baixam o aplicativo** em seu celular, mas não realizam sua primeira compra em até 7 dias após o *download*. Mediante esse estímulo, alguns clientes utilizam o cupom nos 30 dias seguintes (período de validade), e outros, naturalmente, não utilizam. A empresa deseja compreender qual o **perfil dos clientes que utilizam o cupom de desconto**.

| Variável | Descrição |
|---|---|
| ID_Cliente | Código de identificação do cliente |
| Genero | Gênero do cliente: feminino, masculino |
| Idade | Idade do cliente, em anos |
| Cidade | Cidade de residência do cliente: Belo Horizonte, Curitiba, Rio de Janeiro, São Paulo |
| Sistema_Operac | Sistema operacional do aparelho no qual o cliente baixou o aplicativo: Android, iOS, outros |
| Utilizou_Cupom | Indicação se o cliente utilizou o cupom de desconto (1) ou não (0) nos 30 dias seguintes à concessão |

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



Um varejista de perfumes estabeleceu uma ação de *marketing* que consiste em fornecer um **cupom** de 20% de desconto para clientes que **baixam o aplicativo** em seu celular, mas não realizam sua primeira compra em até 7 dias após o *download*. Mediante esse estímulo, alguns clientes utilizam o cupom nos 30 dias seguintes (período de validade), e outros, naturalmente, não utilizam. A empresa deseja compreender qual o **perfil dos clientes que utilizam o cupom de desconto**.

| ID_Cliente | Genero | Idade | Cidade | Sistema_Operac | Utilizou_Cupom |
|---|---|---|---|---|---|
| 00001 | Feminino | 35 | Sao_Paulo | iOS | 0 |
| 00002 | Feminino | 35 | Sao_Paulo | Android | 0 |
| 00003 | Feminino | 38 | Sao_Paulo | Android | 0 |
| 00004 | Feminino | 41 | Sao_Paulo | Outros | 0 |
| 00005 | Masculino | 46 | Sao_Paulo | Android | 0 |
| 00006 | Masculino | 42 | Sao_Paulo | iOS | 0 |
| 00007 | Feminino | 38 | Rio_Janeiro | Android | 0 |
| 00008 | Feminino | 39 | Sao_Paulo | iOS | 1 |
| ... | ... | ... | ... | ... | ... |

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 1:** com todas as variáveis explicativas

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade + Sistema_Operac,
    family = binomial(link = "logit"), data = dados_perfume)

Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)        -5.588703   0.289323 -19.316 < 0.0000000000000002 ***
GeneroMasculino    -1.127865   0.123387  -9.141 < 0.0000000000000002 ***
Idade               0.087793   0.006238  14.074 < 0.0000000000000002 ***
CidadeCuritiba     -0.068981   0.197877  -0.349              0.72739
CidadeRio_Janeiro  -0.431917   0.134592  -3.209              0.00133 **
CidadeSao_Paulo    -0.956902   0.132815  -7.205    0.000000000000581 ***
Sistema_OperacioS   0.005315   0.092899   0.057              0.95438
Sistema_OperacOutros  0.232414   0.137429   1.691              0.09081 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 1:** com todas as variáveis explicativas

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade + Sistema_Operac,
    family = binomial(link = "logit"), data = dados_perfume)

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)
(Intercept)       -5.588703   0.289323 -19.316 < 0.0000000000000002 ***
GeneroMasculino   -1.127865   0.123387  -9.141 < 0.0000000000000002 ***
Idade              0.087793   0.006238  14.074 < 0.0000000000000002 ***
CidadeCuritiba    -0.068981   0.197877  -0.349              0.72739
CidadeRio_Janeiro -0.431917   0.134592  -3.209              0.00133 **
CidadeSao_Paulo   -0.956902   0.132815  -7.205    0.000000000000581 ***
Sistema_OperacioS  0.005315   0.092899   0.057              0.95438
Sistema_OperacOutros  0.232414   0.137429   1.691              0.09081 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

O modelo de regressão logística é ajustado no R por meio da função **glm** (*generalized linear models*)

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 1:** com todas as variáveis explicativas

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade + Sistema_Operac,
    family = binomial(link = "logit"), data = dados_perfume)

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)
(Intercept)       -5.588703   0.289323 -19.316 < 0.0000000000000002 ***
GeneroMasculino   -1.127865   0.123387  -9.141 < 0.0000000000000002 ***
Idade              0.087793   0.006238  14.074 < 0.0000000000000002 ***
CidadeCuritiba    -0.068981   0.197877  -0.349              0.72739
CidadeRio_Janeiro -0.431917   0.134592  -3.209              0.00133 **
CidadeSao_Paulo   -0.956902   0.132815  -7.205    0.000000000000581 ***
Sistema_OperacioS  0.005315   0.092899   0.057              0.95438
Sistema_OperacOutros  0.232414   0.137429   1.691              0.09081 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Os parâmetros do modelo são estimados usando o **método de máxima verossimilhança**.

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 1:** com todas as variáveis explicativas

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade + Sistema_Operac,
    family = binomial(link = "logit"), data = dados_perfume)

Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)        -5.588703   0.289323 -19.316 < 0.0000000000000002 ***
GeneroMasculino    -1.127865   0.123387  -9.141 < 0.0000000000000002 ***
Idade               0.087793   0.006238  14.074 < 0.0000000000000002 ***
CidadeCuritiba     -0.068981   0.197877  -0.349              0.72739
CidadeRio_Janeiro  -0.431917   0.134592  -3.209              0.00133 **
CidadeSao_Paulo    -0.956902   0.132815  -7.205    0.000000000000581 ***
Sistema_OperacioS   0.005315   0.092899   0.057              0.95438
Sistema_OperacOutros  0.232414   0.137429   1.691              0.09081 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Para cada variável qualitativa, a **primeira categoria**, em ordem alfabética, é omitida; para as demais categorias, os parâmetros são estimados.

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 1:** com todas as variáveis explicativas

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade + Sistema_Operac,
    family = binomial(link = "logit"), data = dados_perfume)

Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)        -5.588703   0.289323 -19.316 < 0.0000000000000002 ***
GeneroMasculino    -1.127865   0.123387  -9.141 < 0.0000000000000002 ***
Idade               0.087793   0.006238  14.074 < 0.0000000000000002 ***
CidadeCuritiba     -0.068981   0.197877  -0.349              0.72739
CidadeRio_Janeiro  -0.431917   0.134592  -3.209              0.00133 **
CidadeSao_Paulo    -0.956902   0.132815  -7.205    0.000000000000581 ***
Sistema_OperacioS   0.005315   0.092899   0.057              0.95438
Sistema_OperacOutros  0.232414   0.137429   1.691              0.09081 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Os **desvios padrão** das estimativas também são apresentados.

Arquivo: Compra_Perfumes (.txt)



# *Case:* Compra de Perfumes



## **Modelo 1:** com todas as variáveis explicativas

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade + Sistema_Operac,
    family = binomial(link = "logit"), data = dados_perfume)

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)
(Intercept)       -5.588703   0.289323 -19.316 < 0.0000000000000002 ***
GeneroMasculino   -1.127865   0.123387  -9.141 < 0.0000000000000002 ***
Idade              0.087793   0.006238  14.074 < 0.0000000000000002 ***
CidadeCuritiba    -0.068981   0.197877  -0.349              0.72739
CidadeRio_Janeiro -0.431917   0.134592  -3.209              0.00133 **
CidadeSao_Paulo   -0.956902   0.132815  -7.205    0.000000000000581 ***
Sistema_OperacioS  0.005315   0.092899   0.057              0.95438
Sistema_OperacOutros 0.232414   0.137429   1.691              0.09081 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A partir das estimativas e seus desvios padrão, podemos obter os **intervalos de 95% de confiança** para cada parâmetro $\beta_i$:

$$IC(\boldsymbol{\beta_i}; 95\%) = [\, \boldsymbol{b_i} \pm 1{,}96 \cdot \boldsymbol{DP(b_i)} \,]$$

($i$ varia de 1 até $k$)

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 1:** com todas as variáveis explicativas

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade + Sistema_Operac,
    family = binomial(link = "logit"), data = dados_perfume)

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)
(Intercept)       -5.588703   0.289323 -19.316 < 0.0000000000000002 ***
GeneroMasculino   -1.127865   0.123387  -9.141 < 0.0000000000000002 ***
Idade              0.087793   0.006238  14.074 < 0.0000000000000002 ***
CidadeCuritiba    -0.068981   0.197877  -0.349              0.72739
CidadeRio_Janeiro -0.431917   0.134592  -3.209              0.00133 **
CidadeSao_Paulo   -0.956902   0.132815  -7.205     0.000000000000581 ***
Sistema_OperacioS  0.005315   0.092899   0.057              0.95438
Sistema_OperacOutros  0.232414   0.137429   1.691              0.09081 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tal como na regressão linear, os ***p*-valores** correspondem a testes envolvendo as seguintes hipóteses, para cada um dos parâmetros $\beta_i$ do modelo populacional:

$$\begin{cases} H\text{: } \beta_i = 0 \\ A\text{: } \beta_i \neq 0 \end{cases}$$

($i$ varia de 1 até $k$)

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 1:** com todas as variáveis explicativas

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade + Sistema_Operac,
    family = binomial(link = "logit"), data = dados_perfume)

Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)        -5.588703   0.289323 -19.316 < 0.0000000000000002 ***
GeneroMasculino    -1.127865   0.123387  -9.141 < 0.0000000000000002 ***
Idade               0.087793   0.006238  14.074 < 0.0000000000000002 ***
CidadeCuritiba     -0.068981   0.197877  -0.349              0.72739
CidadeRio_Janeiro  -0.431917   0.134592  -3.209              0.00133 **
CidadeSao_Paulo    -0.956902   0.132815  -7.205     0.000000000000581 ***
Sistema_OperacioS   0.005315   0.092899   0.057              0.95438
Sistema_OperacOutros  0.232414   0.137429   1.691              0.09081 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Adotando 95% de confiança, o **sistema operacional** não é um fator significativo para explicar a utilização do cupom, pois os ***p*-valores** envolvidos são **maiores que 0,05**.

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 2:** sem a variável *Sistema_Operac*

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade, family = binomial(link = "logit"),
    data = dados_perfume)

Coefficients:
                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)       -5.538180   0.280799 -19.723 < 0.0000000000000002 ***
GeneroMasculino   -1.130207   0.123361  -9.162 < 0.0000000000000002 ***
Idade              0.087654   0.006232  14.065 < 0.0000000000000002 ***
CidadeCuritiba    -0.113573   0.184014  -0.617              0.53710
CidadeRio_Janeiro -0.449736   0.131397  -3.423              0.00062 ***
CidadeSao_Paulo   -0.975022   0.129970  -7.502   0.0000000000000629 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Retirando a variável *Sistema_Operac*, não há mais variáveis com todos os *p*-valores acima de 0,05.

Logo, chegamos ao **modelo final**.

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 2:** sem a variável *Sistema_Operac*

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade, family = binomial(link = "logit"),
    data = dados_perfume)

Coefficients:
                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)     -5.538180   0.280799 -19.723 < 0.0000000000000002 ***
GeneroMasculino -1.130207   0.123361  -9.162 < 0.0000000000000002 ***
Idade            0.087654   0.006232  14.065 < 0.0000000000000002 ***
CidadeCuritiba  -0.113573   0.184014  -0.617              0.53710
CidadeRio_Janeiro -0.449736   0.131397  -3.423              0.00062 ***
CidadeSao_Paulo -0.975022   0.129970  -7.502   0.0000000000000629 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sem diferença em relação a BH

Com diferença em relação a BH

Como o *p*-valor associado a **Curitiba** é **maior que 0,05**, isso indica que os clientes de BH (referência) e Curitiba **não possuem diferença significativa** no que diz respeito à probabilidade de utilização do cupom.

Porém, a variável *Cidade* continua sendo relevante no modelo, pois ao menos um *p*-valor é menor que 0,05.

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 2:** sem a variável *Sistema_Operac*

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade, family = binomial(link = "logit"),
    data = dados_perfume)

Coefficients:
                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)      -5.538180   0.280799 -19.723 < 0.0000000000000002 ***
GeneroMasculino  -1.130207   0.123361  -9.162 < 0.0000000000000002 ***
Idade             0.087654   0.006232  14.065 < 0.0000000000000002 ***
CidadeCuritiba   -0.113573   0.184014  -0.617              0.53710
CidadeRio_Janeiro -0.449736   0.131397  -3.423              0.00062 ***
CidadeSao_Paulo  -0.975022   0.129970  -7.502   0.0000000000000629 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sem diferença em relação a BH

Com diferença em relação a BH

Em situações como esta, temos duas alternativas:

(1) **Finalizar** o modelo tal como está;

OU

(2) **Agrupar** BH e Curitiba em uma mesma categoria e **reajustar** o modelo.

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 2:** sem a variável *Sistema_Operac*

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade, family = binomial(link = "logit"),
    data = dados_perfume)

Coefficients:
                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)      -5.538180   0.280799 -19.723 < 0.0000000000000002 ***
GeneroMasculino  -1.130207   0.123361  -9.162 < 0.0000000000000002 ***
Idade             0.087654   0.006232  14.065 < 0.0000000000000002 ***
CidadeCuritiba   -0.113573   0.184014  -0.617              0.53710
CidadeRio_Janeiro -0.449736   0.131397  -3.423              0.00062 ***
CidadeSao_Paulo  -0.975022   0.129970  -7.502   0.0000000000000629 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Interpretação**

Clientes do gênero **masculino** apresentam **menor chance** de utilizar o cupom do que clientes do gênero feminino.

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 2:** sem a variável *Sistema_Operac*

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade, family = binomial(link = "logit"),
    data = dados_perfume)

Coefficients:
                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)      -5.538180   0.280799 -19.723 < 0.0000000000000002 ***
GeneroMasculino  -1.130207   0.123361  -9.162 < 0.0000000000000002 ***
Idade             0.087654   0.006232  14.065 < 0.0000000000000002 ***
CidadeCuritiba   -0.113573   0.184014  -0.617              0.53710
CidadeRio_Janeiro -0.449736   0.131397  -3.423              0.00062 ***
CidadeSao_Paulo  -0.975022   0.129970  -7.502   0.0000000000000629 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Interpretação**

Quanto maior a **idade** do cliente, **maior a chance** de ele utilizar o cupom.

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



## **Modelo 2:** sem a variável *Sistema_Operac*

```
Call:
glm(formula = Utilizou_Cupom ~ Genero + Idade + Cidade, family = binomial(link = "logit"),
    data = dados_perfume)

Coefficients:
                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)     -5.538180   0.280799 -19.723 < 0.0000000000000002 ***
GeneroMasculino -1.130207   0.123361  -9.162 < 0.0000000000000002 ***
Idade            0.087654   0.006232  14.065 < 0.0000000000000002 ***
CidadeCuritiba  -0.113573   0.184014  -0.617              0.53710
CidadeRio_Janeiro -0.449736   0.131397  -3.423              0.00062 ***
CidadeSao_Paulo -0.975022   0.129970  -7.502   0.0000000000000629 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Interpretação**

Clientes do **Rio de Janeiro** apresentam **menor chance** de utilizar o cupom do que clientes de Belo Horizonte e Curitiba; a chance é ainda menor em **São Paulo**.

Arquivo: Compra_Perfumes (.txt)

# Avaliação de Colinearidade



Tal como na regressão linear, é possível que haja problema de **colinearidade** na regressão logística.

Para assegurar que não há colinearidade, podemos examinar a **estatística VIF**, almejando observar valores **mais próximos de 1** do que de 2 ou mais.

No *case* de compra de perfumes, os VIF das variáveis explicativas, obtidos no R a partir da função **vif** do pacote **car** são:

- ***Gênero:*** *1,002*
- ***Idade:*** *1,002*
- ***Cidade:*** *1,003*

Portanto, não há indícios de colinearidade.

Arquivo: Compra_Perfumes (.txt)

# Equação do Modelo Final



Podemos calcular a **probabilidade predita** (ou **estimada**) de que cada cliente utilize o cupom, dadas as suas características, substituindo os valores de $b_0, b_1, \ldots, b_k$ na equação obtida:

$$\hat{p} = \frac{e^{b_0 + b_1 \cdot Genero(masculino) + b_2 \cdot Idade + b_3 \cdot Cidade(CUR) + b_4 \cdot Cidade(RJ) + b_5 \cdot Cidade(SP)}}{1 + e^{b_0 + b_1 \cdot Genero(masculino) + b_2 \cdot Idade + b_3 \cdot Cidade(CUR) + b_4 \cdot Cidade(RJ) + b_5 \cdot Cidade(SP)}}$$

Substituindo os valores estimados dos parâmetros, com arredondamento de 2 casas decimais:

$$\hat{p} = \frac{e^{-5{,}54 - 1{,}13 \cdot Genero(masculino) + 0{,}09 \cdot Idade - 0{,}11 \cdot Cidade(CUR) - 0{,}45 \cdot Cidade(RJ) - 0{,}98 \cdot Cidade(SP)}}{1 + e^{-5{,}54 - 1{,}13 \cdot Genero(masculino) + 0{,}09 \cdot Idade - 0{,}11 \cdot Cidade(CUR) - 0{,}45 \cdot Cidade(RJ) - 0{,}98 \cdot Cidade(SP)}}$$

Arquivo: Compra_Perfumes (.txt)

# Equação do Modelo Final



Podemos calcular a **probabilidade predita** (ou **estimada**) de que cada cliente utilize o cupom, dadas as suas características, substituindo os valores de $b_0, b_1, \ldots, b_k$ na equação obtida:

$$\hat{p} = \frac{e^{b_0 + b_1 \cdot Genero(masculino) + b_2 \cdot Idade + b_3 \cdot Cidade(CUR) + b_4 \cdot Cidade(RJ) + b_5 \cdot Cidade(SP)}}{1 + e^{b_0 + b_1 \cdot Genero(masculino) + b_2 \cdot Idade + b_3 \cdot Cidade(CUR) + b_4 \cdot Cidade(RJ) + b_5 \cdot Cidade(SP)}}$$

*Exemplo:* para uma cliente do gênero **feminino**, com **50 anos de idade** e que mora na cidade de **SP**, teríamos:

$$\hat{p} = \frac{e^{-5{,}54 - 1{,}13 \cdot 0 + 0{,}09 \cdot 50 - 0{,}11 \cdot 0 - 0{,}45 \cdot 0 - 0{,}98 \cdot 1}}{1 + e^{-5{,}54 - 1{,}13 \cdot 0 + 0{,}09 \cdot 50 - 0{,}11 \cdot 0 - 0{,}45 \cdot 0 - 0{,}98 \cdot 1}} = 0{,}106 = 10{,}6\%$$

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# Equação do Modelo Final



Naturalmente, não precisamos realizar o cálculo das probabilidades preditas de forma manual.
Ele pode ser realizado, em linguagem R, a partir do comando:

***predict (nome_do_modelo, nome_da_base, type = "response")***

# 5. Análise de Desempenho



# Ponto de Corte



Após obter as probabilidades preditas por meio do ajuste do modelo, precisamos fornecer uma predição para a **variável resposta original** ($Y$), de natureza binária, que é o nosso interesse original.

Ou seja: para quais observações o modelo prevê que o evento **ocorre** ($Y = 1$) e para quais o modelo prevê que o evento **não ocorre** ($Y = 0$)?

Responder esta questão corresponde a estabelecer um **ponto de corte** sobre as probabilidades preditas pelo modelo. Ou seja, para cada observação, toma-se a seguinte decisão:

- Caso a probabilidade predita ($\hat{p}$) seja **alta**, isto é, **acima** de um ponto de corte, predizemos que $Y = 1$.
- Caso a probabilidade predita ($\hat{p}$) seja **baixa**, isto é, **abaixo** de um ponto de corte, predizemos que $Y = 0$.

**Regra geral:** Um bom valor de ponto de corte consiste na **proporção geral** de observações da base em que $Y = 1$.

*Pode-se adotar outros pontos de corte, a depender do interesse de negócio, como discutiremos adiante.*

*Exemplo:* Para a tomada de decisão de negócio no *case* de compra de perfumes, queremos classificar cada futuro cliente como um potencial **utilizador de cupom** ($Y = 1$) ou **não utilizador de cupom** ($Y = 0$), a depender do seu perfil. Neste caso, um ponto de corte inicial razoável seria de **5,2%**, pois esta é a proporção de indivíduos que utilizaram cupom na amostra.

A **tabela de classificação** apresenta o cruzamento da variável resposta observada ($Y$) com a variável resposta predita ($\hat{Y}$), sendo esta predição proveniente do **modelo**, bem como do **ponto de corte**.

Um bom modelo apresenta alto volume de observações na **diagonal principal** da tabela, ou seja, muitas observações em que o valor predito é igual ao valor observado ($\hat{Y} = Y$).

**Tabela de classificação**, avaliada em um determinado ponto de corte:

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | $VN$ | $FP$ | $\boldsymbol{VN + FP}$ |
| | **1** | $FN$ | $VP$ | $\boldsymbol{FN + VP}$ |
| | **Total** | $\boldsymbol{VN + FN}$ | $\boldsymbol{FP + VP}$ | $\boldsymbol{VN + FN + FP + VP}$ |

Nomenclatura:

- $VP$ = *verdadeiro positivo*
- $VN$ = *verdadeiro negativo*
- $FP$ = *falso positivo*
- $FN$ = *falso negativo*



FIA BUSINESS SCHOOL lab.data

# Acurácia



A **acurácia** corresponde ao percentual geral de **registros classificados corretamente**, englobando tanto registros com $Y = 1$ quanto $Y = 0$.

$$Acur = \frac{VP + VN}{VP + VN + FP + FN}$$

**Tabela de classificação**, avaliada em um determinado ponto de corte:

| | | Variável resposta **predita** 0 | | 1 | | Total |
|---|---|---|---|---|---|---|
| Variável resposta **observada** | **0** | $VN$ | + | $FP$ | | $\boldsymbol{VN + FP}$ |
| | **1** | $FN$ | | $VP$ | / | $\boldsymbol{FN + VP}$ |
| | **Total** | $\boldsymbol{VN + FN}$ | | $\boldsymbol{FP + VP}$ | | $\boldsymbol{VN + FN + FP + VP}$ |

Nomenclatura:

- $VP$ = *verdadeiro positivo*
- $VN$ = *verdadeiro negativo*
- $FP$ = *falso positivo*
- $FN$ = *falso negativo*



FIA BUSINESS SCHOOL lab.data

# Sensibilidade



A **sensibilidade** corresponde ao percentual geral de **registros classificados corretamente**, englobando apenas registros com $Y = 1$.

$$Sensib = \frac{VP}{VP + FN}$$

**Tabela de classificação**, avaliada em um determinado ponto de corte:

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | $VN$ | $FP$ | $\boldsymbol{VN + FP}$ |
| | **1** | $FN$ | $VP$ | $\boldsymbol{FN + VP}$ |
| | **Total** | $\boldsymbol{VN + FN}$ | $\boldsymbol{FP + VP}$ | $\boldsymbol{VN + FN + FP + VP}$ |

Nomenclatura:

- $VP$ = *verdadeiro positivo*
- $VN$ = *verdadeiro negativo*
- $FP$ = *falso positivo*
- $FN$ = *falso negativo*

# Especificidade



A **especificidade** corresponde ao percentual geral de **registros classificados corretamente**, englobando apenas registros com $Y = 0$.

$$Especif = \frac{VN}{VN + FP}$$

**Tabela de classificação**, avaliada em um determinado ponto de corte:

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | $VN$ | $FP$ | $\boldsymbol{VN + FP}$ |
| | **1** | $FN$ | $VP$ | $\boldsymbol{FN + VP}$ |
| | **Total** | $\boldsymbol{VN + FN}$ | $\boldsymbol{FP + VP}$ | $\boldsymbol{VN + FN + FP + VP}$ |

Nomenclatura:

- $VP$ = *verdadeiro positivo*
- $VN$ = *verdadeiro negativo*
- $FP$ = *falso positivo*
- $FN$ = *falso negativo*



FIA BUSINESS SCHOOL lab.data


# Interpretação das Medidas de Desempenho



Os índices de **acurácia**, **sensibilidade** e **especificidade** variam de 0 a 1 (ou de 0% a 100%). Sua interpretação é subjetiva, ou seja, depende do que se considera uma taxa de acerto relevante do ponto de vista do contexto.

De qualquer forma, alguns patamares de referência para julgar a **discriminância** do modelo (capacidade de classificar zeros e uns na resposta) são:

| Valor | Discriminância |
|---|---|
| Abaixo de 50% | Nenhuma (pior que o aleatório) |
| De 50% a 60% | Fraca |
| De 60% a 70% | Satisfatória |
| De 70% a 80% | Boa |
| Acima de 80% | Muito boa |

# *Case:* Compra de Perfumes



Análise de desempenho no *case* de **compra de perfumes**.

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | 7.790 | 4.190 | **11.980** |
| | **1** | 218 | 442 | **660** |
| | **Total** | **8.008** | **4.632** | **12.640** |

- **442** clientes que utilizaram o cupom de desconto foram classificados **corretamente**.
- **218** clientes que utilizaram o cupom de desconto foram classificados **incorretamente**.

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Compra de Perfumes



Análise de desempenho no *case* de **compra de perfumes**.

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | 7.790 | 4.190 | **11.980** |
| | **1** | 218 | 442 | **660** |
| | **Total** | **8.008** | **4.632** | **12.640** |

- **7.790** clientes que não utilizaram o cupom de desconto foram classificados **corretamente**.
- **4.190** clientes que não utilizaram o cupom de desconto foram classificados **incorretamente**.

Arquivo: Compra_Perfumes (.txt)

# *Case:* Compra de Perfumes



Análise de desempenho no *case* de **compra de perfumes**.

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | 7.790 | 4.190 | **11.980** |
| | **1** | 218 | 442 | **660** |
| | **Total** | **8.008** | **4.632** | **12.640** |

**Acurácia**

$$Acur = \frac{7.790 + 442}{12.640} = 65{,}1\%$$

**Sensibilidade**

$$Sensib = \frac{442}{660} = 67{,}0\%$$

**Especificidade**

$$Especif = \frac{7.790}{11.980} = 65{,}0\%$$

**Interpretações**

- ***Acurácia***: A cada **100 clientes**, o modelo identifica corretamente quem utiliza ou não o cupom para **65** deles.
- ***Sensibilidade***: A cada **100 clientes** que **utilizam** o cupom, o modelo identifica corretamente **67** deles.
- ***Especificidade***: A cada **100 clientes** que **não utilizam** o cupom, o modelo identifica corretamente **65** deles.

Arquivo: Compra_Perfumes (.txt)



FIA BUSINESS SCHOOL lab.data

# Estatística KS

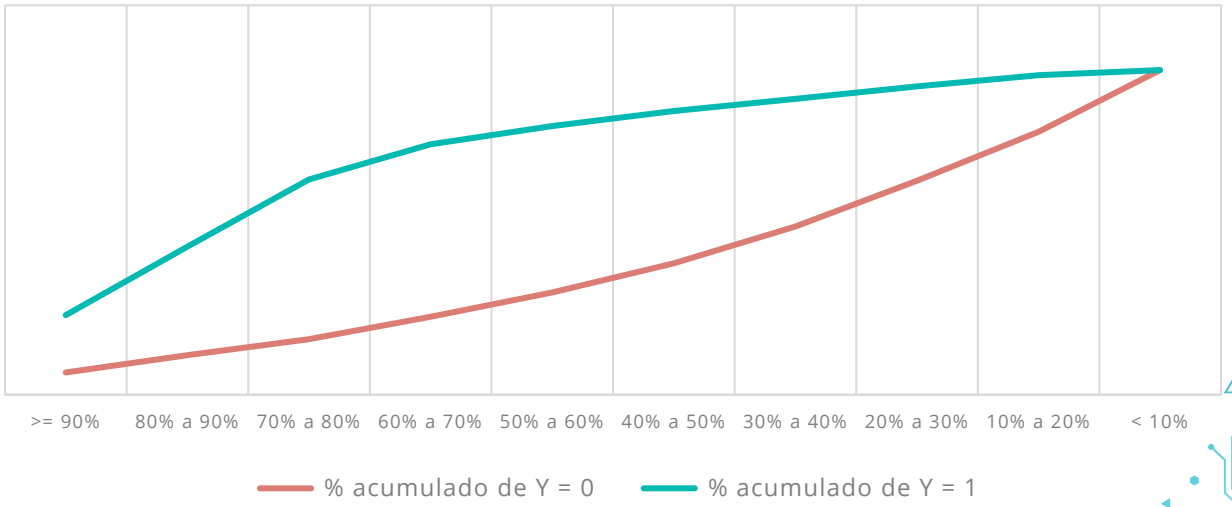


A **Estatística de Kolmogorov-Smirnov** (KS) avalia a capacidade de o modelo **segregar** os 0's e 1's da variável resposta ($Y$) por meio das probabilidades $\hat{p}$. Em outras palavras, é a capacidade de o modelo estimar probabilidades $\hat{p}$ altas quando $Y = 1$, e probabilidades $\hat{p}$ baixas quando $Y = 0$.

Isso é mensurado a partir do grau máximo de **separação** entre as curvas de frequências relativas (%) acumuladas de **dois grupos**: observações com $Y = 0$ *versus* observações com $Y = 1$. Essas curvas são construídas em relação aos percentis das probabilidades preditas ($\hat{p}$).

| Percentis de $\hat{p}$ | % Y = 0 | % Acum. Y = 0 | % Y = 1 | % Acum. Y = 1 |
|---|---|---|---|---|
| >= 90% | 6,8% | 6,8% | 24,5% | 24,5% |
| 80% a 90% | 5,4% | 12,2% | 21,1% | 45,6% |
| 70% a 80% | 4,9% | 17,1% | 20,6% | 66,3% |
| 60% a 70% | 6,9% | 24,0% | 10,9% | 77,2% |
| 50% a 60% | 7,5% | 31,5% | 5,6% | 82,8% |
| 40% a 50% | 8,9% | 40,5% | 4,6% | 87,4% |
| 30% a 40% | 11,3% | 51,8% | 3,8% | 91,2% |
| 20% a 30% | 14,2% | 65,9% | 3,8% | 95,0% |
| 10% a 20% | 15,1% | 81,0% | 3,5% | 98,5% |
| < 10% | 19,0% | 100,0% | 1,5% | 100,0% |

% acumulado de *Y* = 0 e *Y* = 1
*versus* percentis da probabilidade predita



# Estatística KS



A **Estatística de Kolmogorov-Smirnov** (KS) avalia a capacidade de o modelo **segregar** os 0's e 1's da variável resposta ($Y$) por meio das probabilidades $\hat{p}$. Em outras palavras, é a capacidade de o modelo estimar probabilidades $\hat{p}$ altas quando $Y = 1$, e probabilidades $\hat{p}$ baixas quando $Y = 0$.

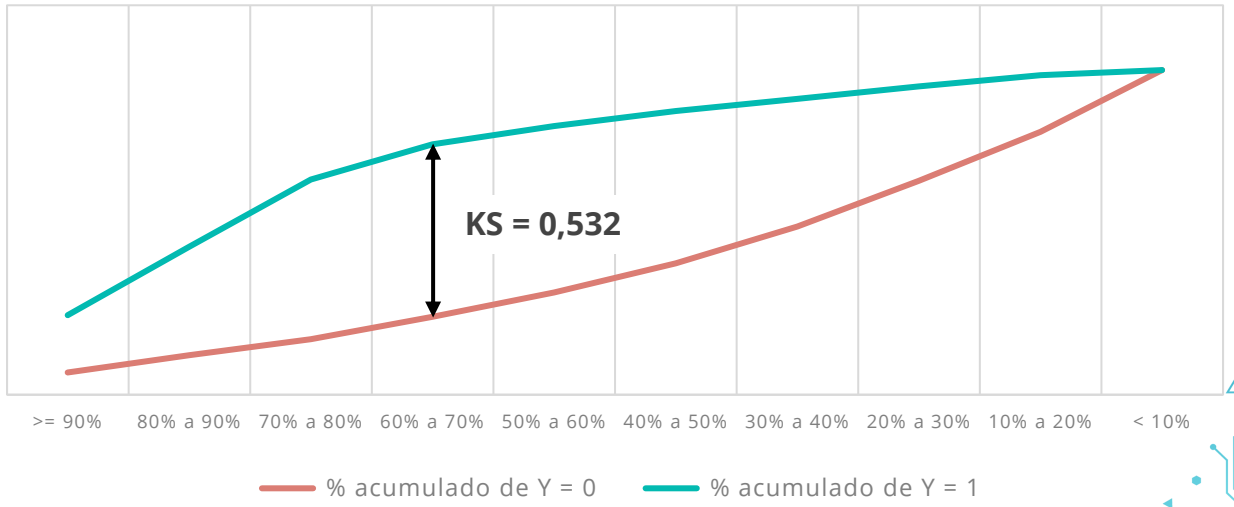
Isso é mensurado a partir do grau máximo de **separação** entre as curvas de frequências relativas (%) acumuladas de **dois grupos**: observações com $Y = 0$ *versus* observações com $Y = 1$. Essas curvas são construídas em relação aos percentis das probabilidades preditas ($\hat{p}$).

| Percentis de $\hat{p}$ | % Y = 0 | % Acum. Y = 0 | % Y = 1 | % Acum. Y = 1 |
|---|---|---|---|---|
| >= 90% | 6,8% | 6,8% | 24,5% | 24,5% |
| 80% a 90% | 5,4% | 12,2% | 21,1% | 45,6% |
| 70% a 80% | 4,9% | 17,1% | 20,6% | 66,3% |
| 60% a 70% | 6,9% | 24,0% | 10,9% | 77,2% |
| 50% a 60% | 7,5% | 31,5% | 5,6% | 82,8% |
| 40% a 50% | 8,9% | 40,4% | 4,7% | 87,5% |
| 30% a 40% | 11,3% | 51,8% | 3,8% | 91,2% |
| 20% a 30% | 14,2% | 65,9% | 3,8% | 95,0% |
| 10% a 20% | 15,1% | 81,0% | 3,5% | 98,5% |
| < 10% | 19,0% | 100,0% | 1,5% | 100,0% |

**Máxima separação entre as curvas**

**KS = 0,772 – 0,240 = 0,532**

% acumulado de $Y = 0$ e $Y = 1$
*versus* percentis da probabilidade predita






FIA BUSINESS SCHOOL lab.data

# Estatística KS



O valor do KS varia numa escala de **-1** (total discriminância errada de 0's e 1's) a **1** (total discriminância correta de 0's e 1's). Também é comum multiplicá-lo por 100, reportando-o numa escala de **-100** a **100**.

| Estatística KS | Poder de discriminância |
|---|---|
| < 0 | Nenhum (pior que o aleatório) |
| Entre 0 e 0,2 | Baixo |
| Entre 0,2 e 0,3 | Aceitável |
| Entre 0,3 e 0,4 | Bom |
| Entre 0,4 e 0,5 | Muito bom |
| Entre 0,5 e 0,6 | Excelente |
| > 0,6 | Excelente, mas suspeito |

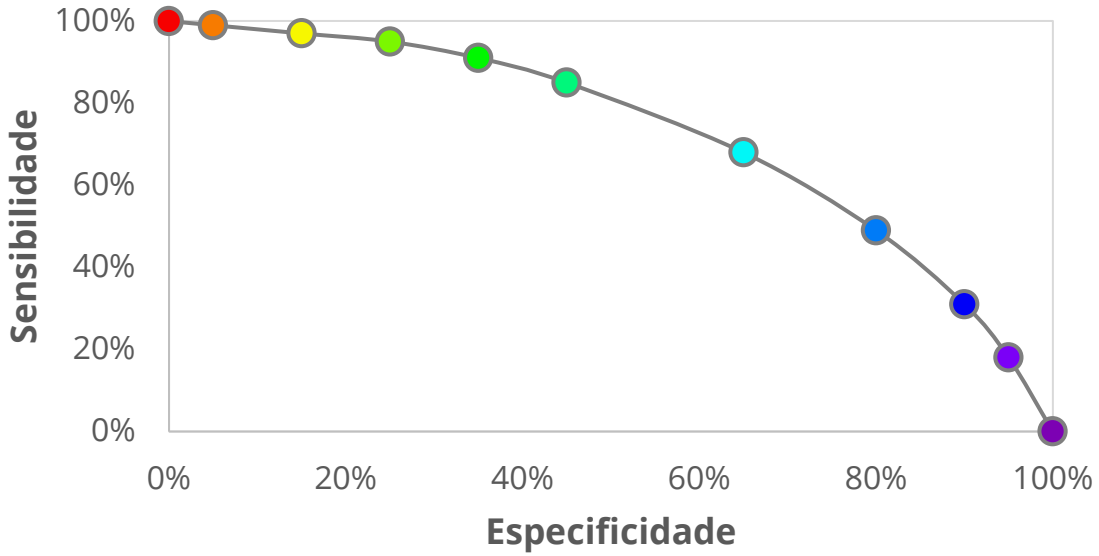
- Valores de KS **acima de 0,6**, apesar de serem possíveis de se observar na prática, indicam um poder de discriminância atipicamente alto. Nesses casos, é importante checar a consistência das variáveis, especialmente se alguma delas está relacionada de forma determinística com a variável resposta.

# Curva ROC

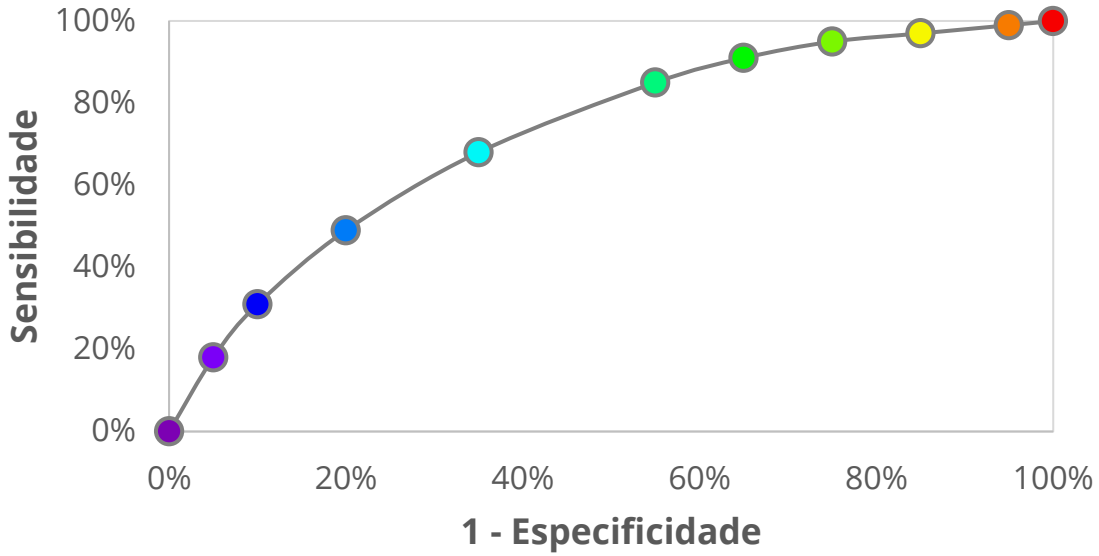


A análise ROC (*Receiver Operating Curve*) foi desenvolvida entre 1950 e 1960 para avaliar a detecção de sinais em radar e na psicologia sensorial. Particularmente, a **curva ROC** permite avaliar a variação da **sensibilidade** *versus* **especificidade** para diferentes valores de ponto de corte, em um modelo de regressão logística.

Comportamento típico da sensibilidade *versus* especificidade




Comportamento típico da sensibilidade *versus* 1-especificidade **(curva ROC)**




Ponto de corte

0 1



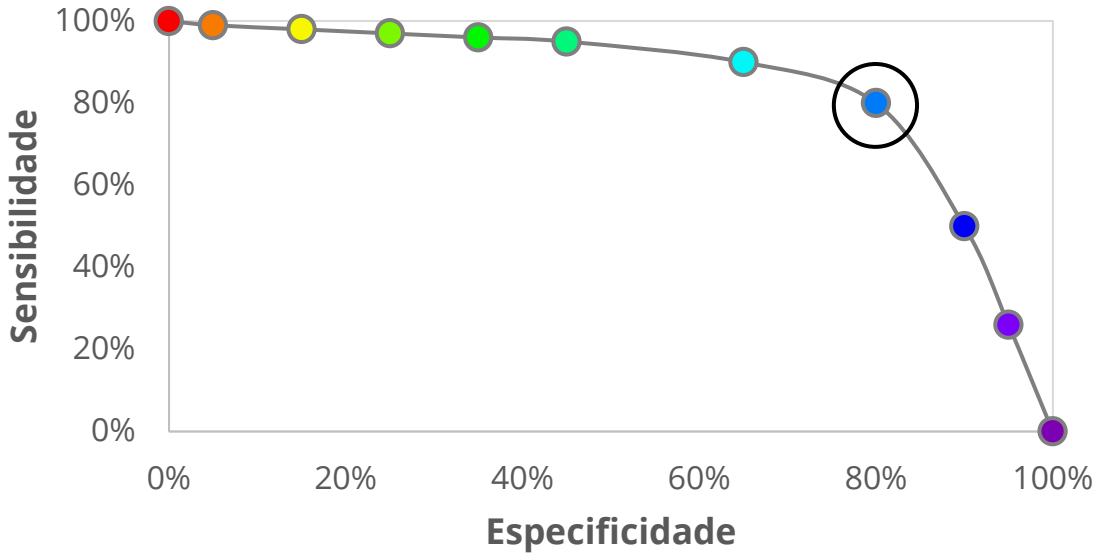
FIA BUSINESS SCHOOL lab.data

# Curva ROC

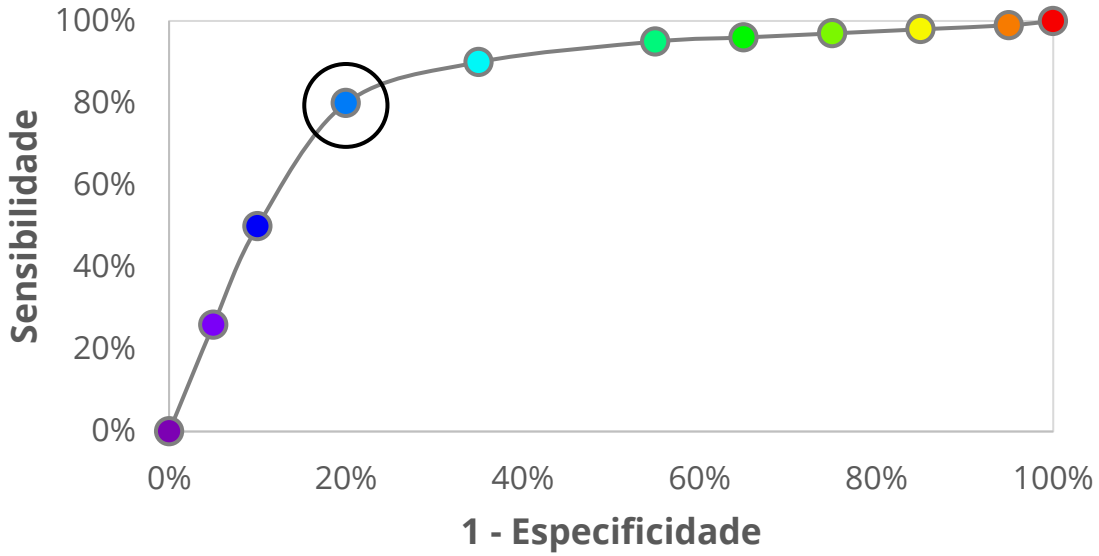


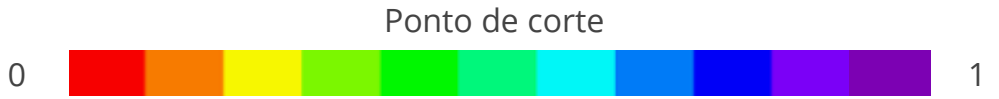
Quando o modelo possui **alto poder de discriminância** de 0's e 1's, existem pontos de corte que propiciam altos níveis de sensibilidade e de especificidade, concomitantemente. Isso leva a uma curva ROC com **concavidade mais acentuada**.

Comportamento da sensibilidade *versus* especificidade para um **ótimo modelo**




Comportamento da sensibilidade *versus* 1-especificidade para um **ótimo modelo**





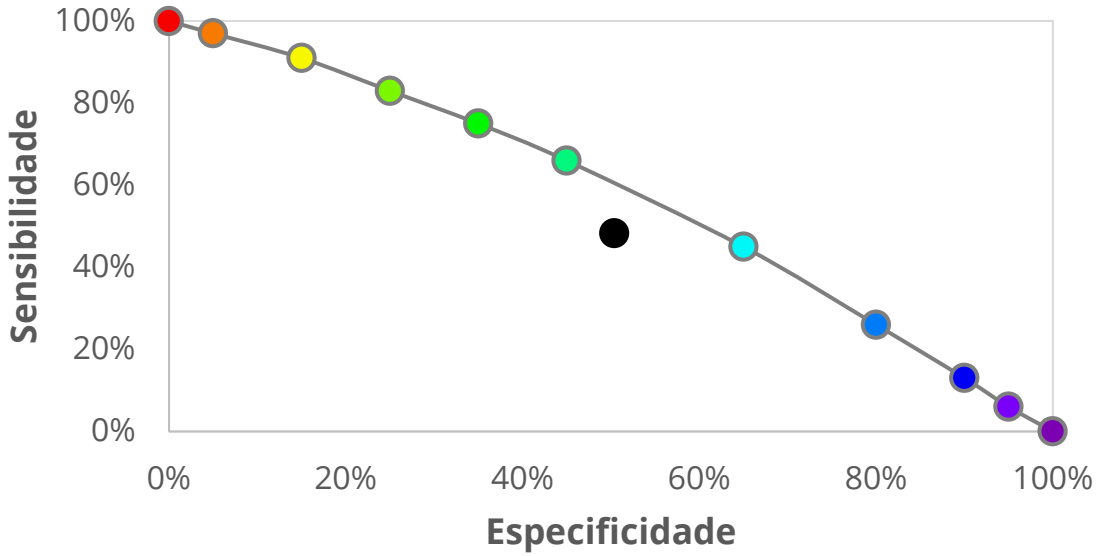

FIA BUSINESS SCHOOL lab.data

# Curva ROC

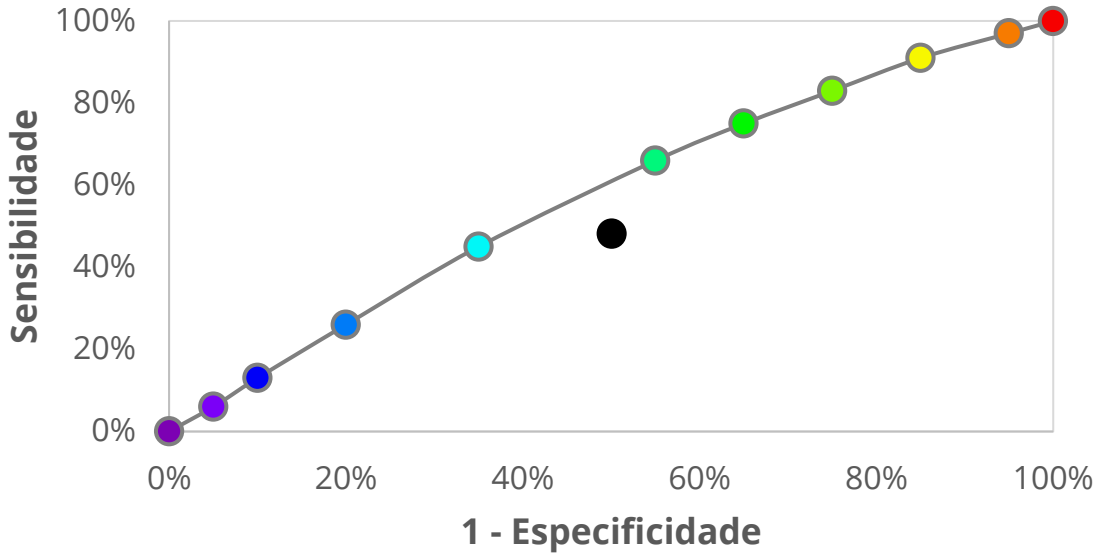


Já quando o modelo possui **baixo poder de discriminância** de 0's e 1's, o ponto de maior balanceamento entre sensibilidade e especificidade costuma estar próximo de 50% / 50%, equivalente a uma classificação aleatória. Isso leva a uma curva ROC com **concavidade menos acentuada**.

Comportamento da sensibilidade *versus* especificidade para um **modelo ruim**




Comportamento da sensibilidade *versus* 1-especificidade para um **modelo ruim**




Ponto de corte

0 1



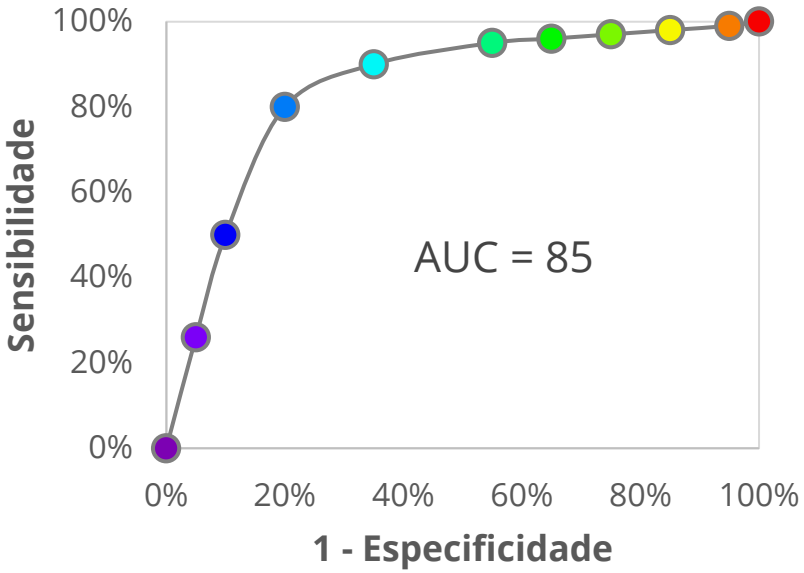
FIA BUSINESS SCHOOL lab.data

# Área Abaixo da Curva ROC (AUC)

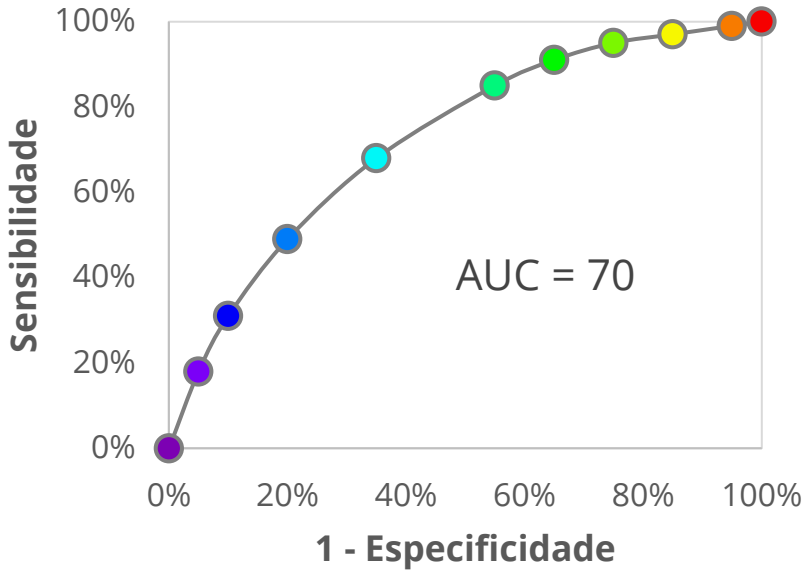


O valor da **área abaixo da curva ROC**, abreviado como **AUC**, é outro indicador da qualidade de classificação do modelo, associado a diferentes pontos de corte possíveis. Nesse sentido, é uma medida mais abrangente que o KS, que considera apenas o máximo distanciamento entre as curvas, ou seja, um único ponto de corte.

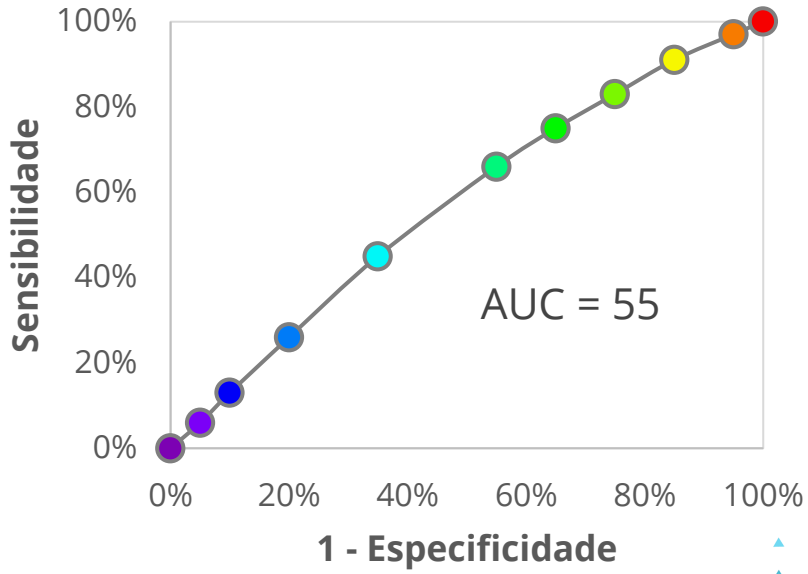
Exemplo de modelo **ótimo**




Exemplo de modelo **bom**




Exemplo de modelo **ruim**




Ponto de corte




FIA BUSINESS SCHOOL lab.data

# Área Abaixo da Curva ROC (AUC)



O valor do AUC varia numa escala de **0** (total discriminância errada de 0's e 1's) a **1** (total discriminância correta de 0's e 1's). Também é comum multiplicá-lo por 100, reportando-o numa escala de **0** a **100**.

| Área abaixo da curva (AUC) | Poder de discriminância |
|---|---|
| < 0,5 | Nenhum (pior que o aleatório) |
| 0,5 a 0,6 | Baixo |
| 0,6 a 0,7 | Aceitável/Bom |
| 0,7 a 0,8 | Muito bom |
| 0,8 a 0,9 | Excelente |
| > 0,9 | Excelente, mas suspeito |

- Valores de AUC **acima de 0,9**, apesar de serem possíveis de se observar na prática, indicam um poder de discriminância atipicamente alto. Nesses casos, é importante checar a consistência das variáveis, especialmente se alguma delas está relacionada de forma determinística com a variável resposta.

# *Case:* Compra de Perfumes



Análise de desempenho no *case* de **compra de perfumes**.

Utilizando as funções do pacote **ROCR** do R, chegamos aos seguintes resultados:

- **KS = 0,32**
- **AUC = 0,71**

A análise conjunta dos dois índices aponta que o modelo apresenta **bom poder de discriminância** dos indivíduos que utilizam e que não utilizam o cupom.

Arquivo: Compra_Perfumes (.txt)

# 6. *Cases* Adicionais



# *Case:* Investimento em Ações



Um banco de investimentos deseja estimar a probabilidade de seus clientes tomarem a decisão de investir pela 1ª vez em **renda variável** nos próximos 6 meses, dado que já possuem algum tipo de investimento em realizado em fundos de **renda fixa**.

Para isso, examinaram a foto histórica dos clientes, retroagida em 6 meses, com a seguinte variável resposta: *Investiu_Variavel_6M* = 1 (se investiu pela 1ª vez em renda variável nos 6 meses seguintes) ou = 0 (se não investiu em renda variável nos 6 meses seguintes).

| Variável | Descrição |
|---|---|
| Investimento_Fixa | Valor (R$) investido em fundos de renda fixa como CDB, LCI/LCA e CRI/CRA, nos últimos 12 meses |
| Investimento_Tesouro | Valor (R$) investido no Tesouro Direto, nos últimos 12 meses |
| Investimento_Poupanca | Valor (R$) investido em poupança, nos últimos 12 meses |
| Rendimento_Liquido_12M | Valor líquido total de rendimentos associados aos investimentos de renda fixa, nos últimos 12 meses |
| Saldo_Conta | Saldo (R$) disponível na conta para novos investimentos |
| Investiu_Variavel_6M | Indicação se o cliente investiu (1) ou não (0) em renda variável pela 1ª vez, nos próximos 6 meses |

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



Um banco de investimentos deseja estimar a probabilidade de seus clientes tomarem a decisão de investir pela 1ª vez em **renda variável** nos próximos 6 meses, dado que já possuem algum tipo de investimento em realizado em fundos de **renda fixa**.

Para isso, examinaram a foto histórica dos clientes, retroagida em 6 meses, com a seguinte variável resposta: *Investiu_Variavel_6M* = 1 (se investiu pela 1ª vez em renda variável nos 6 meses seguintes) ou = 0 (se não investiu em renda variável nos 6 meses seguintes).

A base contém dados de 7.817 clientes. As primeiras linhas são apresentadas a seguir.

| ID_Cliente | Investimento_Fixa | Investimento_Tesouro | Investimento_Poupanca | Rendimento_Liq_12M | Saldo_Conta | Investiu_Variavel_6M |
|---|---|---|---|---|---|---|
| #0001 | R$ 0,00 | R$ 5.434,42 | R$ 0,00 | R$ 229,52 | R$ 0,00 | 0 |
| #0002 | R$ 8.240,53 | R$ 4.066,68 | R$ 0,00 | R$ 429,71 | R$ 6.145,54 | 0 |
| #0003 | R$ 3.843,34 | R$ 4.757,95 | R$ 0,00 | R$ 65,21 | R$ 0,00 | 0 |
| #0004 | R$ 2.297,05 | R$ 4.201,29 | R$ 0,00 | R$ 308,23 | R$ 0,00 | 0 |
| #0005 | R$ 2.477,00 | R$ 4.146,93 | R$ 0,00 | R$ 156,40 | R$ 1.328,92 | 0 |
| #0006 | R$ 0,00 | R$ 1.224,36 | R$ 0,00 | R$ 44,74 | R$ 1.252,03 | 0 |
| #0007 | R$ 7.581,91 | R$ 6.226,31 | R$ 0,00 | R$ 36,74 | R$ 0,00 | 0 |
| ... | ... | ... | ... | ... | ... | ... |

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



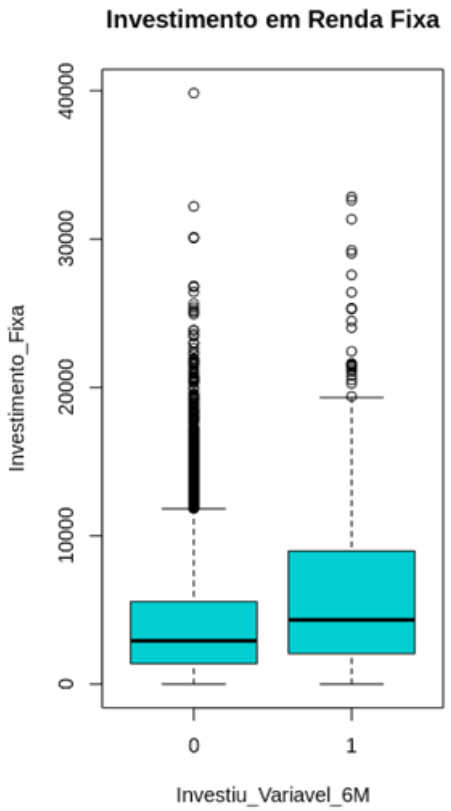
Um banco de investimentos deseja estimar a probabilidade de seus clientes tomarem a decisão de investir pela 1ª vez em **renda variável** nos próximos 6 meses, dado que já possuem algum tipo de investimento em realizado em fundos de **renda fixa**.

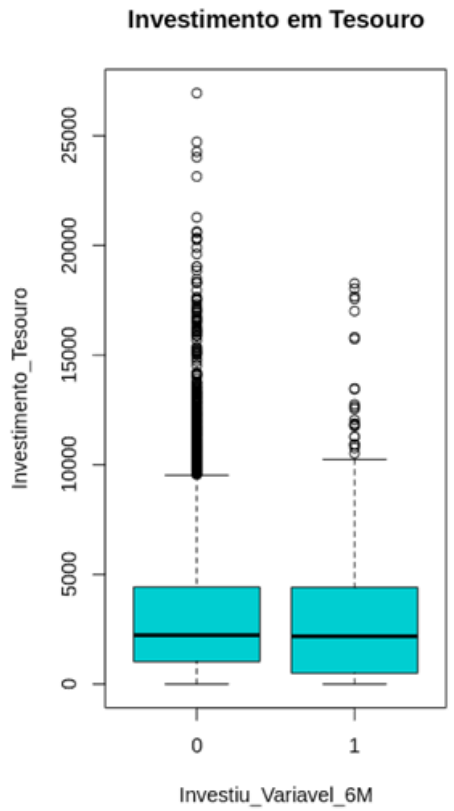
Para isso, examinaram a foto histórica dos clientes, retroagida em 6 meses, com a seguinte variável resposta: *Investiu_Variavel_6M* = 1 (se investiu pela 1ª vez em renda variável nos 6 meses seguintes) ou = 0 (se não investiu em renda variável nos 6 meses seguintes).

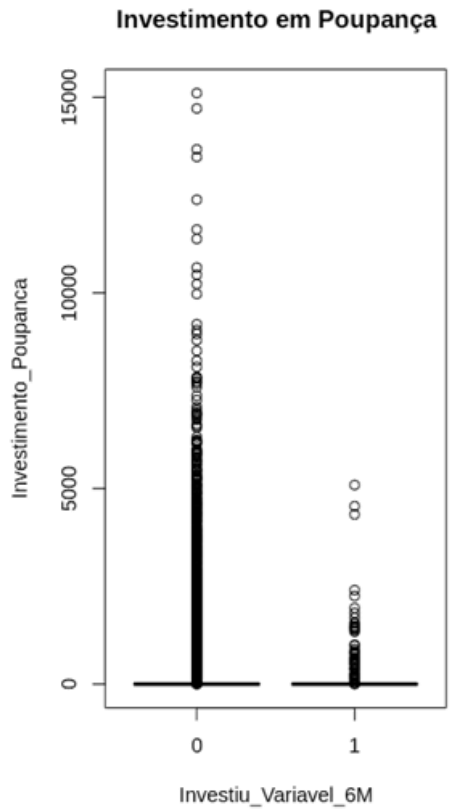
(a) Faça uma breve análise exploratória da base de dados.
(b) Faça a análise bivariada de cada variável explicativa *versus* variável resposta. Quais aspectos parecem estar associados à decisão de investir em renda variável nos próximos 6 meses?
(c) Construa um modelo de regressão logística múltipla, selecionando variáveis estatisticamente significativas com 95% de confiança e atentando-se a colinearidade. Quais aspectos estão associados à decisão de investir em renda variável nos próximos 6 meses? Interprete as estimativas dos parâmetros no modelo final.
(d) Escreva a equação estimada do modelo final.
(e) Obtenha a tabela de classificação e as medidas de desempenho. Como você avalia a qualidade do modelo?
(f) Estime a probabilidade de investir em renda variável para um investidor que tem 6.000 reais investidos em renda fixa, não tem investimento em poupança, e tem 3.000 reais de saldo em conta.

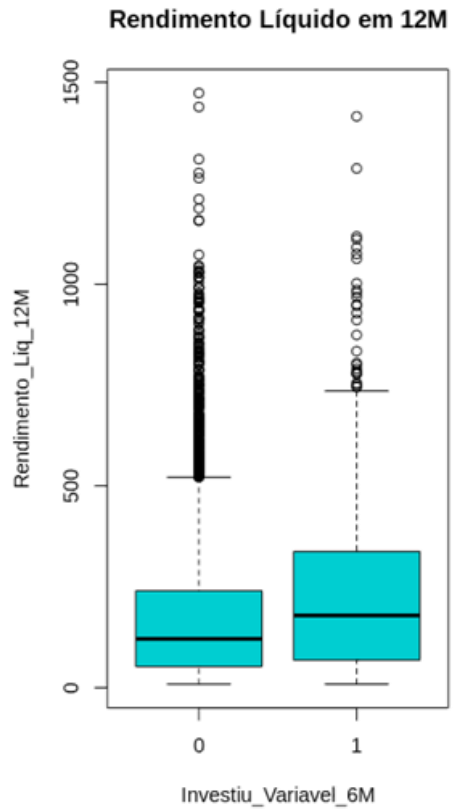
Arquivo: Investimento_Acoes (.txt)

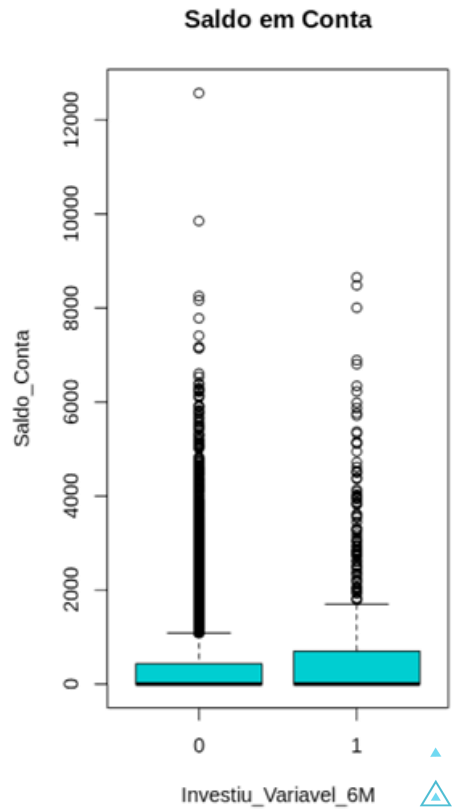
## Análise bivariada




Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



## **Modelo 1:** com todas as variáveis explicativas

```
Call:
glm(formula = Investiu_Variavel_6M ~ Investimento_Fixa + Investimento_Tesouro +
    Investimento_Poupanca + Rendimento_Liq_12M + Saldo_Conta,
    family = binomial(link = "logit"), data = dados_investimento)

Coefficients:
                      Estimate  Std. Error z value            Pr(>|z|)
(Intercept)         -3.03374236  0.08096865 -37.468 < 0.0000000000000002 ***
Investimento_Fixa    0.00008317  0.00001087   7.648   0.0000000000000204 ***
Investimento_Tesouro -0.00001037  0.00001534  -0.676              0.49906
Investimento_Poupanca -0.00052821  0.00010053  -5.254   0.0000001486182413 ***
Rendimento_Liq_12M    0.00078165  0.00028365   2.756              0.00586 **
Saldo_Conta           0.00019513  0.00003278   5.953   0.0000000026284680 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A variável **Investimento_Tesouro** apresenta *p*-valor acima de 5%. Logo, não é uma variável estatisticamente significativa para explicar a decisão de investir em renda variável, e pode ser retirada do modelo.

Arquivo: Investimento_Acoes (.txt)

# *Case:* Investimento em Ações



## **Modelo 2:** sem a variável *Investimento_Tesouro*

```
Call:
glm(formula = Investiu_Variavel_6M ~ Investimento_Fixa + Investimento_Poupanca +
    Rendimento_Liq_12M + Saldo_Conta, family = binomial(link = "logit"),
    data = dados_investimento)

Coefficients:
                      Estimate  Std. Error z value             Pr(>|z|)
(Intercept)         -3.05918437  0.07194886 -42.519 < 0.0000000000000002 ***
Investimento_Fixa    0.00008578  0.00001011   8.484 < 0.0000000000000002 ***
Investimento_Poupanca -0.00052722  0.00010055  -5.243          0.00000015775 ***
Rendimento_Liq_12M    0.00068468  0.00024342   2.813                0.00491 **
Saldo_Conta           0.00019556  0.00003278   5.966          0.00000000244 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Agora, **todas as variáveis** apresentam *p*-valor abaixo de 5%. Logo, são variáveis estatisticamente significativas para explicar a decisão de investir em renda variável.

Arquivo: Investimento_Acoes (.txt)



Os VIF das variáveis explicativas, obtidos no R a partir da função **vif** do pacote **car**, são:

- ***Investimento_Fixa:*** *1,535*
- ***Investimento_Poupanca:*** *1,001*
- ***Rendimento_Liq_12M:*** *1,537*
- ***Saldo_Conta:*** *1,002*

Portanto, **há indícios de colinearidade** entre as variáveis *Investimento_Fixa* e *Rendimento_Liq_12M*.

A seguir, vamos avaliar o desempenho do modelo com as 4 variáveis explicativas. Em seguida, reajustaremos o modelo, retirando uma das variáveis envolvidas na colinearidade.

# *Case:* Investimento em Ações



Análise de desempenho no *case* de **investimento em ações** *(modelo 2).*

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | 4.909 | 2.300 | **7.209** |
| | **1** | 277 | 331 | **608** |
| | **Total** | **5.186** | **2.631** | **7.817** |

- **331** clientes que investiram em renda variável foram classificados **corretamente**.
- **277** clientes que investiram em renda variável foram classificados **incorretamente**.

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



Análise de desempenho no *case* de **investimento em ações** *(modelo 2).*

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | 4.909 | 2.300 | **7.209** |
| | **1** | 277 | 331 | **608** |
| | **Total** | **5.186** | **2.631** | **7.817** |

- **4.909** clientes que não investiram em renda variável foram classificados **corretamente**.
- **2.300** clientes que não investiram em renda variável foram classificados **incorretamente**.

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações

Análise de desempenho no *case* de **investimento em ações** *(modelo 2).*

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | 4.909 | 2.300 | **7.209** |
| | **1** | 277 | 331 | **608** |
| | **Total** | **5.186** | **2.631** | **7.817** |

**Acurácia**

$$Acur = \frac{4.909 + 331}{7.817} = 67{,}0\%$$

**Sensibilidade**

$$Sensib = \frac{331}{608} = 54{,}4\%$$

**Especificidade**

$$Especif = \frac{4.909}{7.209} = 68{,}1\%$$

**Interpretações**

- ***Acurácia***: A cada **100 clientes**, o modelo identifica corretamente quem investe ou não em ações para **67** deles.
- ***Sensibilidade***: A cada **100 clientes** que **investem** em ações, o modelo identifica corretamente **54** deles.
- ***Especificidade***: A cada **100 clientes** que **não investem** em ações, o modelo identifica corretamente **68** deles.

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



## Análise de desempenho no *case* de **investimento em ações** *(modelo 2).*

Utilizando as funções do pacote **ROCR** do R, chegamos aos seguintes resultados:

- **KS =** 0,25
- **AUC =** 0,66

A análise conjunta dos dois índices aponta que o modelo apresenta **poder de discriminância aceitável** dos indivíduos que investem e que não investem em ações nos próximos 6 meses.

Arquivo: Investimento_Acoes (.txt)

# *Case:* Investimento em Ações



## **Modelo 3:** sem a variável *Rendimento_Liq_12M*

```
Call:
glm(formula = Investiu_Variavel_6M ~ Investimento_Fixa + Investimento_Poupanca +
    Saldo_Conta, family = binomial(link = "logit"), data = dados_investimento)

Coefficients:
                      Estimate   Std. Error z value             Pr(>|z|)
(Intercept)         -2.999416204  0.068369380 -43.871 < 0.0000000000000002 ***
Investimento_Fixa    0.000102193  0.000008159  12.526 < 0.0000000000000002 ***
Investimento_Poupanca -0.000523673  0.000100636  -5.204          0.00000019541 ***
Saldo_Conta          0.000193633  0.000032805   5.903          0.00000000358 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Novamente, **todas as variáveis** apresentam *p*-valor abaixo de 5%. Logo, são variáveis estatisticamente significativas para explicar a decisão de investir em renda variável.

Arquivo: Investimento_Acoes (.txt)

# *Case:* Investimento em Ações



## **Modelo 3:** sem a variável *Rendimento_Liq_12M*

```
Call:
glm(formula = Investiu_Variavel_6M ~ Investimento_Fixa + Investimento_Poupanca +
    Saldo_Conta, family = binomial(link = "logit"), data = dados_investimento)

Coefficients:
                      Estimate   Std. Error z value             Pr(>|z|)
(Intercept)         -2.999416204  0.068369380 -43.871 < 0.0000000000000002 ***
Investimento_Fixa    0.000102193  0.000008159  12.526 < 0.0000000000000002 ***
Investimento_Poupanca -0.000523673  0.000100636  -5.204          0.00000019541 ***
Saldo_Conta          0.000193633  0.000032805   5.903          0.00000000358 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Interpretação**

Clientes com maiores valores investidos em **renda fixa** nos últimos 12 meses apresentam **maior chance** de investir em renda variável nos próximos 6 meses.

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



## **Modelo 3:** sem a variável *Rendimento_Liq_12M*

```
Call:
glm(formula = Investiu_Variavel_6M ~ Investimento_Fixa + Investimento_Poupanca +
    Saldo_Conta, family = binomial(link = "logit"), data = dados_investimento)

Coefficients:
                      Estimate   Std. Error z value             Pr(>|z|)
(Intercept)         -2.999416204  0.068369380 -43.871 < 0.0000000000000002 ***
Investimento_Fixa    0.000102193  0.000008159  12.526 < 0.0000000000000002 ***
Investimento_Poupanca -0.000523673  0.000100636  -5.204          0.00000019541 ***
Saldo_Conta          0.000193633  0.000032805   5.903          0.00000000358 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Interpretação**

Clientes com maiores valores investidos em **poupança** nos últimos 12 meses apresentam **menor chance** de investir em renda variável nos próximos 6 meses.

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



## **Modelo 3:** sem a variável *Rendimento_Liq_12M*

```
Call:
glm(formula = Investiu_Variavel_6M ~ Investimento_Fixa + Investimento_Poupanca +
    Saldo_Conta, family = binomial(link = "logit"), data = dados_investimento)

Coefficients:
                     Estimate   Std. Error z value             Pr(>|z|)
(Intercept)       -2.999416204  0.068369380 -43.871 < 0.0000000000000002 ***
Investimento_Fixa  0.000102193  0.000008159  12.526 < 0.0000000000000002 ***
Investimento_Poupanca -0.000523673  0.000100636  -5.204          0.00000019541 ***
Saldo_Conta        0.000193633  0.000032805   5.903          0.00000000358 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Interpretação**

Clientes com maiores **saldos disponíveis** para novos investimentos apresentam **maior chance** de investir em renda variável nos próximos 6 meses.

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



Houve redução expressiva do VIF da variável *Investimento_Fixa*, após a retirada do *Rendimento_Liq_12M*:

- ***Investimento_Fixa:* *1,001***
- ***Investimento_Poupanca:*** *1,001*
- ***Saldo_Conta:*** *1,002*

Portanto, **não há mais indícios de colinearidade**.

Arquivo: Investimento_Acoes (.txt)

# *Case:* Investimento em Ações



Análise de desempenho no *case* de **investimento em ações** *(modelo 3).*

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | 4.874 | 2.335 | **7.209** |
| | **1** | 278 | 330 | **608** |
| | **Total** | **5.152** | **2.665** | **7.817** |

- **330** clientes que investiram em renda variável foram classificados **corretamente**.
- **278** clientes que investiram em renda variável foram classificados **incorretamente**.

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



Análise de desempenho no *case* de **investimento em ações** *(modelo 3).*

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | 4.874 | 2.335 | **7.209** |
| | **1** | 278 | 330 | **608** |
| | **Total** | **5.152** | **2.665** | **7.817** |

- **4.874** clientes que não investiram em renda variável foram classificados **corretamente**.
- **2.335** clientes que não investiram em renda variável foram classificados **incorretamente**.

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



Análise de desempenho no *case* de **investimento em ações** *(modelo 3).*

| | | Variável resposta **predita** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Total** |
| Variável resposta **observada** | **0** | 4.874 | 2.335 | **7.209** |
| | **1** | 278 | 330 | **608** |
| | **Total** | **5.152** | **2.665** | **7.817** |

**Acurácia**

$$Acur = \frac{4.874 + 330}{7.817} = 66{,}6\%$$

(antes: 67,0%)

**Sensibilidade**

$$Sensib = \frac{330}{608} = 54{,}3\%$$

(antes: 54,4%)

**Especificidade**

$$Especif = \frac{4.874}{7.209} = 67{,}6\%$$

(antes: 68,1%)

**Interpretações**

- ***Acurácia***: A cada **100 clientes**, o modelo identifica corretamente quem investe ou não em ações para **67** deles.
- ***Sensibilidade***: A cada **100 clientes** que **investem** em ações, o modelo identifica corretamente **54** deles.
- ***Especificidade***: A cada **100 clientes** que **não investem** em ações, o modelo identifica corretamente **68** deles.

Arquivo: Investimento_Acoes (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Investimento em Ações



## Análise de desempenho no *case* de **investimento em ações** *(modelo 3).*

Utilizando as funções do pacote **ROCR** do R, chegamos aos seguintes resultados:

- **KS =** 0,24 (antes: 0,25)
- **AUC =** 0,66 (antes: 0,66)

A análise conjunta dos dois índices aponta que o modelo apresenta **poder de discriminância aceitável/bom** dos indivíduos que investem e que não investem em ações nos próximos 6 meses.

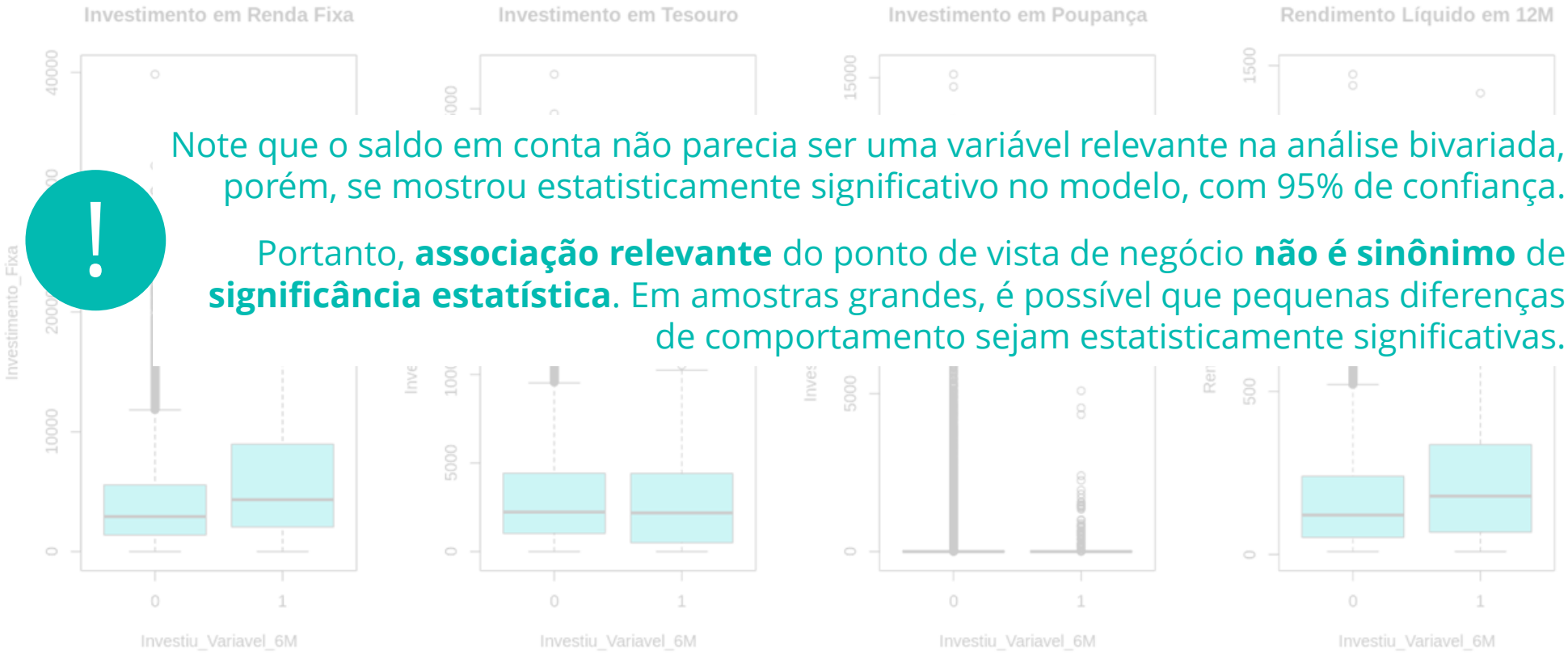
**Em resumo, não houve redução relevante de desempenho do modelo após a retirada da variável *Rendimento_Liq_12M*.**

Arquivo: Investimento_Acoes (.txt)



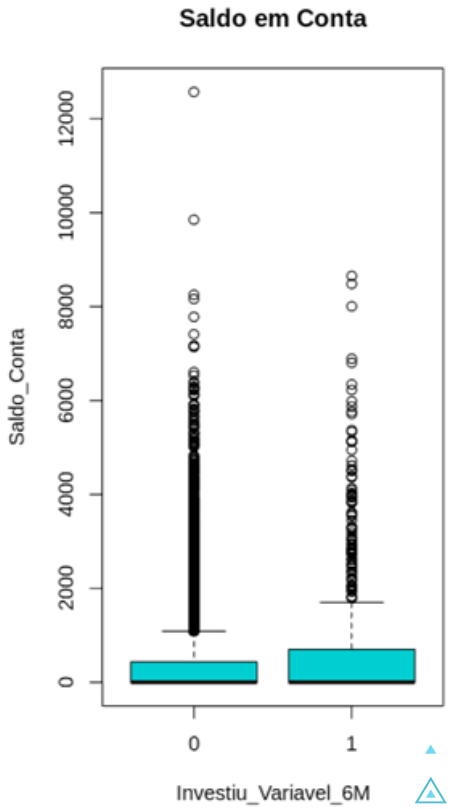
FIA BUSINESS SCHOOL lab.data

## Análise bivariada




Note que o saldo em conta não parecia ser uma variável relevante na análise bivariada, porém, se mostrou estatisticamente significativo no modelo, com 95% de confiança.

Portanto, **associação relevante** do ponto de vista de negócio **não é sinônimo** de **significância estatística**. Em amostras grandes, é possível que pequenas diferenças de comportamento sejam estatisticamente significativas.




Arquivo: Investimento_Acoes (.txt)

# Aplicação: *Credit Score*



O ***credit score*** é uma aplicação muito comum do modelo de regressão logística na área financeira. Seu objetivo consiste em predizer a **probabilidade** *(score)* de um indivíduo ser adimplente em seus pagamentos.

Modelos de *credit score* permitem que as instituições avaliem o **risco de crédito** de forma mais eficaz e tomem melhores decisões sobre a concessão em diferentes contextos, tais como:

- Empréstimos (para pessoas físicas ou jurídicas)
- Cartões de crédito
- Cheque especial
- Financiamentos (veículos, imóveis)
- Hipotecas

Outros objetivos comuns na área de crédito são:

- ***Collection score****:* probabilidade de recuperar dívidas de um cliente inadimplente.
- ***Behaviour score****:* probabilidade de um cliente manter padrões de pagamento positivos, baseando-se em seu comportamento histórico.

FIA BUSINESS SCHOOL lab.data

# Aplicação: *Credit Score*



A variável resposta em modelos de ***credit score*** costuma ser definida como:

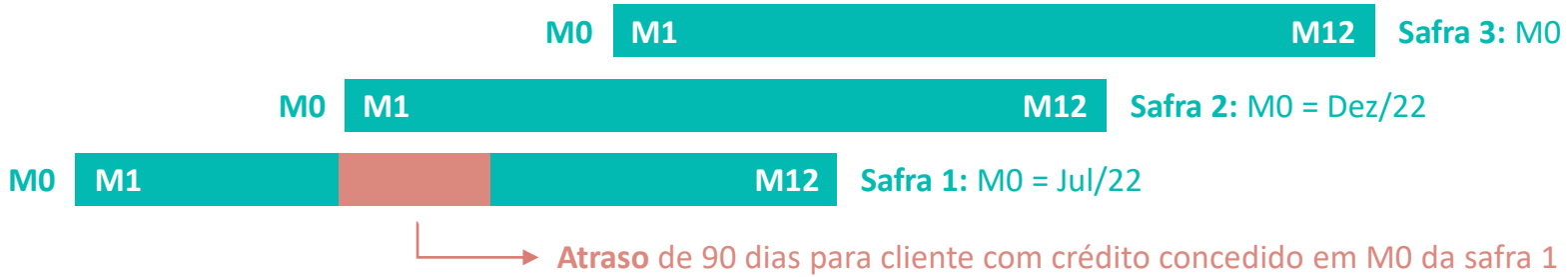
- $Y = 1$: cliente ***bom***, que apresentou **menos** de $A$ dias de atraso ao longo de $B$ meses
- $Y = 0$: cliente ***mau***, que apresentou $A$ dias de atraso **ou mais** ao longo de $B$ meses

Valores comuns para $A$: **30**, **60**, **90** dias de atraso.

Valores comuns para $B$: **3**, **6**, **12** meses.

Em modelos de *credit score*, é recomendável considerar clientes provenientes de diversas **safras de observação**, para capturar mudanças de comportamento ao longo do tempo e efeitos sazonais. Por outro lado, não é apropriado analisar dados muito antigos, que podem não refletir as dinâmicas das políticas de crédito mais recentes.

*Exemplo: Atraso de 90 dias ou mais em 12 meses, entre clientes de 3 safras "empilhadas".*




FIA BUSINESS SCHOOL lab.data

# *Case: Credit Score*



A área de risco de crédito de um banco precisa desenvolver um modelo para atribuir uma **pontuação de crédito** *(score)* para novos indivíduos solicitantes de **empréstimos pessoais**, com base em informações históricas que denotam sua capacidade de pagamento fora da instituição. Para isso, consideraram uma base de dados com 3 safras empilhadas de clientes que tiveram empréstimo aprovado.

| Variável | Descrição |
|---|---|
| ID | Código identificador do cliente |
| SAFRA | Safra de seleção |
| IDADE | Idade, em anos |
| RENDA_MEDIA_MENSAL | Renda média mensal nos últimos 12 meses, em R$ |
| TOTAL_INVESTIMENTOS | Valor total em investimentos que possui em outras instituições, em R$ |
| QTDE_CONSULTAS_CREDITO_12M | Quantidade de consultas de créditos realizadas em seu nome, nos últimos 12 meses |
| QTDE_CARTOES | Quantidade de cartões de crédito que possui em outras instituições |

*(continua no próximo slide)*

Arquivo: Credit_Score (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case: Credit Score*



A área de risco de crédito de um banco precisa desenvolver um modelo para atribuir uma **pontuação de crédito** *(score)* para novos indivíduos solicitantes de **empréstimos pessoais**, com base em informações históricas que denotam sua capacidade de pagamento fora da instituição. Para isso, consideraram uma base de dados com 3 safras empilhadas de clientes que tiveram empréstimo aprovado.

| Variável | Descrição |
|---|---|
| QTDE_EMPRESTIMOS_12M | Quantidade de contratações de empréstimos em outras instituições, nos últimos 12 meses |
| QTDE_CHEQUE_ESPECIAL_12M | Quantidade de contratações de cheque especial em outras instituições, nos últimos 12 meses |
| QTDE_PGTOS_EM_ATRASO_12M | Quantidade de parcelas pagas em atraso em outras instituições, nos últimos 12 meses |
| TOTAL_DIAS_ATRASO_12M | Total de dias de atraso de pagamento de parcelas em outras instituições, nos últimos 12 meses |
| FLAG_PGTO_PARCIAL_12M | Indicação de se realizou ou não algum pagamento parcial de parcelas, nos últimos 12 meses |
| VALOR_PGTOS_12M | Valor total em pagamentos de parcelas em outras instituições, nos últimos 12 meses |
| PERC_MEDIO_LIMITE_TOMADO_12M | Percentual médio mensal de limite comprometido em cartões de crédito, nos últimos 12 meses |
| RESPOSTA_MAU_BOM | Indicação de se foi *mau* (0) ou *bom* (1), ou seja, se apresentou ou não 30 dias ou mais de atraso de pagamento de parcelas do empréstimo nos 12 meses seguintes |

Arquivo: Credit_Score (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case: Credit Score*



A área de risco de crédito de um banco precisa desenvolver um modelo para atribuir uma **pontuação de crédito** *(score)* para novos indivíduos solicitantes de **empréstimos pessoais**, com base em informações históricas que denotam sua capacidade de pagamento fora da instituição. Para isso, consideraram uma base de dados com 3 safras empilhadas de clientes que tiveram empréstimo aprovado.

(a) Faça uma breve análise exploratória da base de dados.
(b) Faça a análise bivariada de cada variável explicativa *versus* variável resposta. Quais aspectos parecem estar associados à definição de cliente *bom* adotada pelo banco?
(c) Construa um modelo de regressão logística múltipla, selecionando variáveis estatisticamente significativas com 95% de confiança e atentando-se a colinearidade. Quais aspectos estão associados à definição de cliente *bom*? Interprete as estimativas dos parâmetros no modelo final.
(d) Escreva a equação estimada do modelo final.
(e) Obtenha a tabela de classificação e as medidas de desempenho. Como você avalia a qualidade do modelo?
(f) Calcule a probabilidade de *bom* para um cliente que, nos últimos 12 meses, realizou uma solicitação de empréstimo, não usou cheque especial, não teve atrasos e pagou 10.000 reais em parcelas.

Arquivo: Credit_Score (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Cancelamento em Telecom



O diretor de *marketing* de uma empresa de telefonia móvel deseja criar um modelo para calcular a probabilidade de que cada cliente **cancele o serviço nos próximos 6 meses**, a fim de realizar ações de retenção ativa. A base de dados representa uma amostra histórica recente (retroagida no tempo) com 10.000 clientes.

| Variável | Descrição |
|---|---|
| ID_CLIENTE | Código identificador do cliente |
| SCORE_CREDITO | Pontuação de *bureau* de crédito referente ao potencial de adimplência (quanto maior, menor o risco) |
| GENERO | Gênero |
| IDADE | Idade, em anos |
| TEMPO_RELACIONAMENTO | Tempo de relacionamento, em anos |
| TIPO_PLANO | Tipo de plano: pré-pago ou pós-pago |
| RENDA | Renda mensal declarada |
| CANCELOU | Indicação de se cancelou o serviço (1) ou não (0), nos 6 meses seguintes |

Arquivo: Cancelamento_Telecom (.txt)



FIA BUSINESS SCHOOL lab.data

# *Case:* Cancelamento em Telecom



O diretor de *marketing* de uma empresa de telefonia móvel deseja criar um modelo para calcular a probabilidade de que cada cliente **cancele o serviço nos próximos 6 meses**, a fim de realizar ações de retenção ativa. A base de dados representa uma amostra histórica recente (retroagida no tempo) com 10.000 clientes.

(a) Faça uma breve análise exploratória da base de dados.
(b) Faça a análise bivariada de cada variável explicativa *versus* variável resposta. Quais aspectos parecem estar associados ao cancelamento do serviço nos próximos 6 meses?
(c) Construa um modelo de regressão logística múltipla, selecionando variáveis estatisticamente significativas com 95% de confiança e atentando-se a colinearidade. Quais aspectos estão associados ao cancelamento do serviço nos próximos 6 meses? Interprete as estimativas dos parâmetros no modelo final.
(d) Escreva a equação estimada do modelo final.
(e) Obtenha a tabela de classificação e as medidas de desempenho. Como você avalia a qualidade do modelo?
(f) Estime a probabilidade de cancelamento do serviço para um cliente que possui plano pré-pago, renda mensal de 5.000 reais e 30 anos de idade.

Arquivo: Cancelamento_Telecom (.txt)



FIA BUSINESS SCHOOL lab.data

# Referências Bibliográficas

REGRESSÃO LOGÍSTICA




- Agresti, A. *Categorical Data Analysis*. 2ª edição. Wiley, 2002.
- Conover, W. J. *Practical Nonparametric Statistics*. Wiley, 1999.
- Hosmer, D. W. e Lemeshow, S. *Applied Logistic Regression*, 2ª ed. New York: Wiley, 2000.
- James, G. *An Introduction to Statistical Learning - With Applications in R*. 2ª edição. Springer, 2021.


lab.data

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