

FIAP

NBA

ADELAIDE ALVES DE OLIVEIRA

PROFESSORA



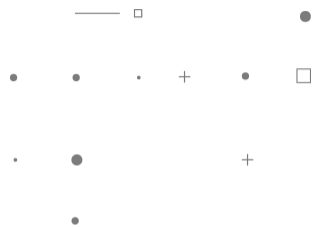
profadelaide.alves@fiap.com.br

Formação Acadêmica

- Bacharel em Estatística – UNICAMP
- Mestre em Ciências – FSP/USP

Atividades Profissionais

- Diretora Técnica Estatística da empresa **SD&W** - www.sdw.com.br
- Professora de Fundamentos Estatísticos, DataMining, Análise Preditiva e Machine Learning na FIAP dos cursos MBA: Big Data, Data Science, Business Intelligence & Analytics, Digital Data Marketing, IA & ML e Engenharia de Dados e nos Shift: People Analytics e Python Journey



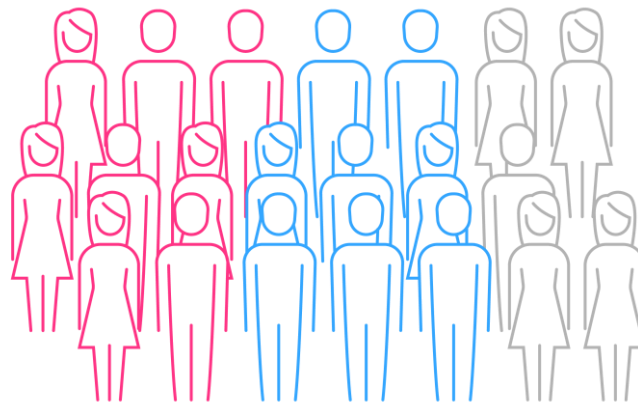
TÉCNICAS DE CLASSIFICAÇÃO



TÉCNICAS SUPERVISIONADAS

Métodos de Classificação

- Como os heavy users **se diferem** em seu perfil demográfico dos light users ?
- Quais são os clientes ativos que **se assemelham** aos clientes cancelados?
- Que **fatores ou atitudes** fazem com que os meus clientes **prefiram** o meu produto?
- Quais são as **características** que apresentam os clientes que compraram o produto de maior rentabilidade?



GRUPO A

GRUPO B

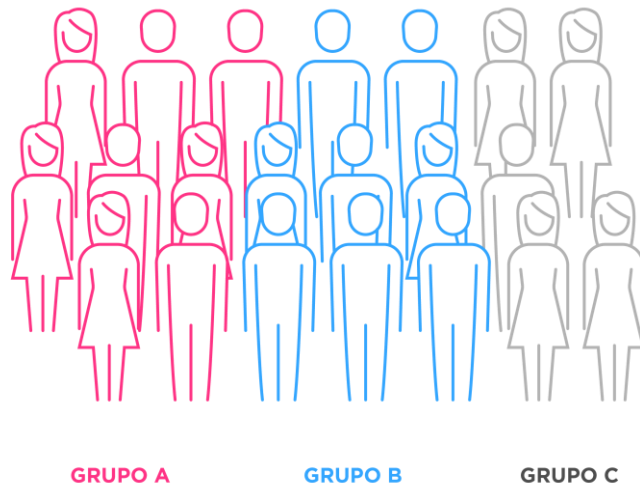
GRUPO C

Como separar grupos **previamente definidos**? Como definir critérios, funções das variáveis que discriminem os grupos?

TÉCNICAS SUPERVISIONADAS

Métodos de Classificação

- Dado um conjunto de treinamento onde cada registro contém um conjunto de atributos, e um dos atributos é a nossa variável de interesse é tipo categórica/classes.
- Encontrar um modelo para determinar o valor do atributo classe em função dos valores de outros atributos.
- Objetivo: definir a classe de novos registros, a classe deve ser atribuída o mais corretamente possível.



USO DOS MODELOS NO CICLO DO CLIENTE

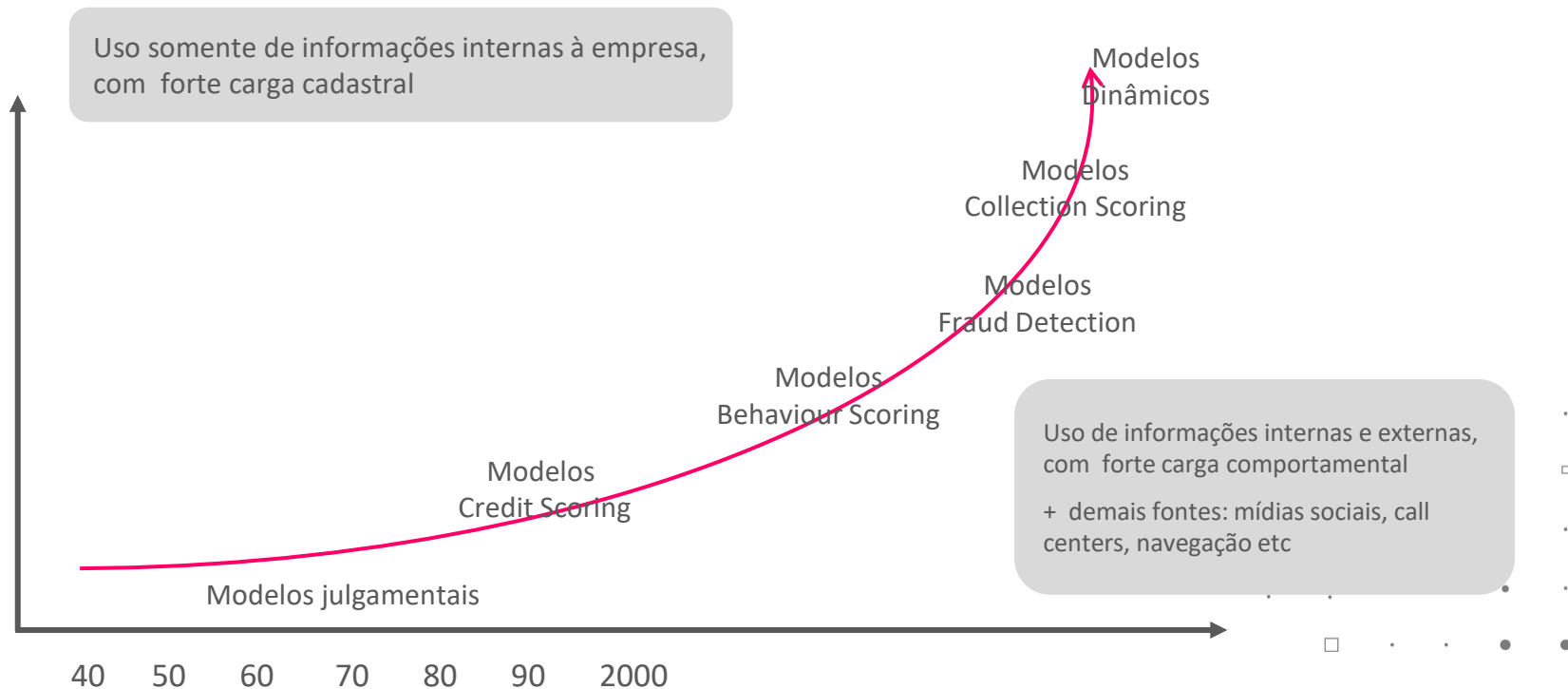


USO DOS MODELOS NO CICLO DO CLIENTE

MODELOS DE AQUISIÇÃO

- Adquirir **prospects** com os mesmos perfis dos bons clientes da empresa;
- Campanhas sobre os clientes da concorrência;
- Estimular os clientes à aquisição de novos produtos/serviços (*cross selling*).

EVOLUÇÃO DAS FERRAMENTAS DE GESTÃO DE RISCO



USO DOS MODELOS NO CICLO DO CLIENTE

- **Modelo de *Credit Scoring***
 - Considera informações/dados do contrato (tempo de relacionamento recente);
 - Probabilidade de o novo cliente vir a ser inadimplente.
- **Modelo de Inadimplência (*Behaviour Scoring*)**
 - Considera dados de utilização/comportamento dos clientes;
 - Probabilidade de o cliente vir a ser tornar um inadimplente.
- **Modelo de Cobrança (*Collection Scoring*)**
 - Considera dados de utilização dos clientes e do mercado;
 - Probabilidade de um cliente pagar.
- **Modelo de *Churn* e fraude/anomalias/abusos**
 - Considera dados de utilização dos clientes e do mercado;
 - Probabilidade de o cliente cancelar a “conta/serviço/produto”.

USO DOS MODELOS NO CICLO DO CLIENTE

MODELOS DE RETENÇÃO

Objetivo:

- Identificar na base de dados de clientes prováveis a cancelar o relacionamento com a empresa.
- Oferecer suporte a área de relacionamento e permitir que campanhas de fidelização sejam direcionadas a clientes com risco real de interromper o relacionamento com a instituição.

- Melhores resultados nas campanhas realizadas;
- Redução de custos de abordagens indesejadas;
- Satisfação dos clientes;
- Maior credibilidade.

USO DOS MODELOS NO CICLO DO CLIENTE

Mercado Financeiro - Exemplo

Objetivo:

- Identificar na base de dados correntistas prováveis a cancelar/inativar o relacionamento (conta corrente) com o banco;

Dimensões:

- Utilização: diretamente relacionadas à geração de receita de cada correntista (dados transacionais).

Exemplos: produto adquirido, quantidade de cheques emitidos, saldo médio, tempo de relacionamento, conta conjunta, etc.

- Demográficas: informações descritivas do cliente.

Exemplos: sexo, idade, endereço, profissão, estado civil, renda, etc

- Definição da janela de tempo de análise

- Planejamento amostral (técnicas estatísticas aliadas às restrições do Banco)

Benefícios:

- Realizar ações fidelizadoras sobre os correntistas propensos a cancelar/inativar sua conta corrente

TIPOS DE CLASSIFICADORES

Classificadores eager (espertos)

A partir da amostragem inicial (conjunto de treinamento), constroem um modelo de classificação capaz de classificar novos registros.

Uma vez pronto o modelo, o conjunto de treinamento não é mais utilizado na classificação de novos objetos (registros)

- Árvores e Regras de Decisão
- Redes Neurais
- Redes Bayesianas e Naïve Bayes
- Máquinas de Vetores de Suporte

Classificadores lazy (preguiçosos)

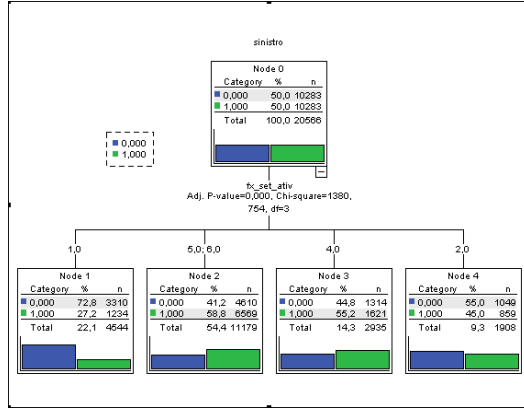
Cada novo registro é comparado com todo o conjunto de treinamento e é classificado segundo a classe do registro que é mais similar. Também conhecido como: Aprendizado baseado em exemplo (*Instance-based Learning*):

- Método kNN (k-nearest-neighbor)

Outros Métodos

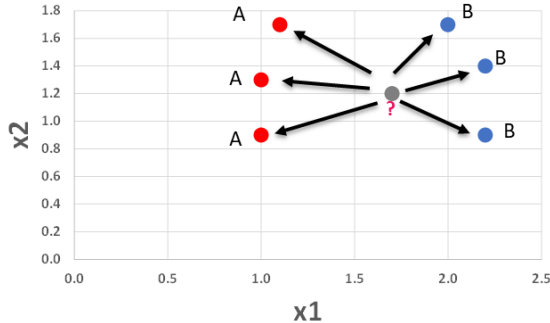
- Algoritmos Genéticos
- Conjuntos Fuzzy

Árvore de Decisão



K-NN K-Nearest Neighbors

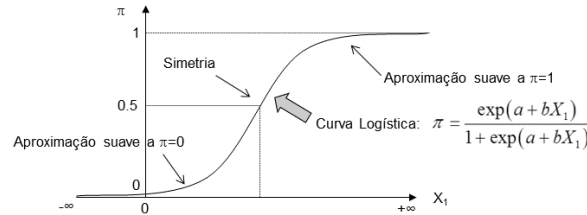
Qual a distância euclidiana entre os pontos?



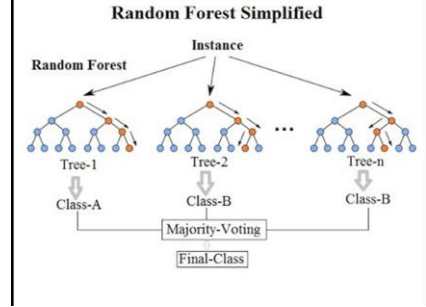
Técnicas de Classificação:

Regressão Logística

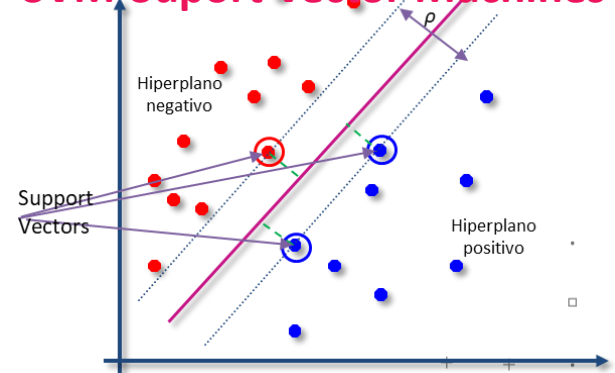
| variável | categoria | Coefficientes |
|----------------------------|-----------------------|---------------|
| fatura em atraso | até 3 dias | -1,276 |
| | 3 a 15 dias | -0,611 |
| | de 15 a 30 dias | 0,580 |
| | mais de 30 dias | 1,308 |
| Tempo de cliente | até 1 ano | 0,580 |
| | de 1 a 3 anos | 0,401 |
| | de 3 a 8 anos | -0,264 |
| | mais de 8 anos | -0,718 |
| valor da fatura | Até R\$250 | 0,262 |
| | R\$ 250 a R\$ 800 | 0,103 |
| | R\$ 800 a R\$ 1.499 | -0,105 |
| | Mais de R\$1.500 | -0,261 |
| % de gasto com alimentação | até 10% | 0,581 |
| | de 10% a 20% | 0,401 |
| | de 20% a 30% | -0,264 |
| | mais de 30% | -0,718 |
| Região de Risco | Região 4 | 1,067 |
| | Região 3 | 0,371 |
| | Região 2 | -0,368 |
| | Região 1 | -1,069 |
| renda mensal | Até R\$ 1.518 | 0,455 |
| | R\$ 1.519 a R\$ 3.000 | 0,080 |
| | R\$ 3.000 a R\$ 4.500 | -0,122 |
| | Mais de R\$ 4.500 | -0,413 |
| Constante | | 0,099 |



Random Forest

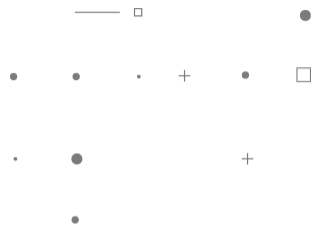


SVM-Suport Vector Machines



Naive Bayes

$$P(y | x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i | y)}{P(x_1, \dots, x_n)}$$



TÉCNICAS DE DISCRIMINAÇÃO

CLASSIFICAÇÃO

Descobertas Supervisionadas de Relações

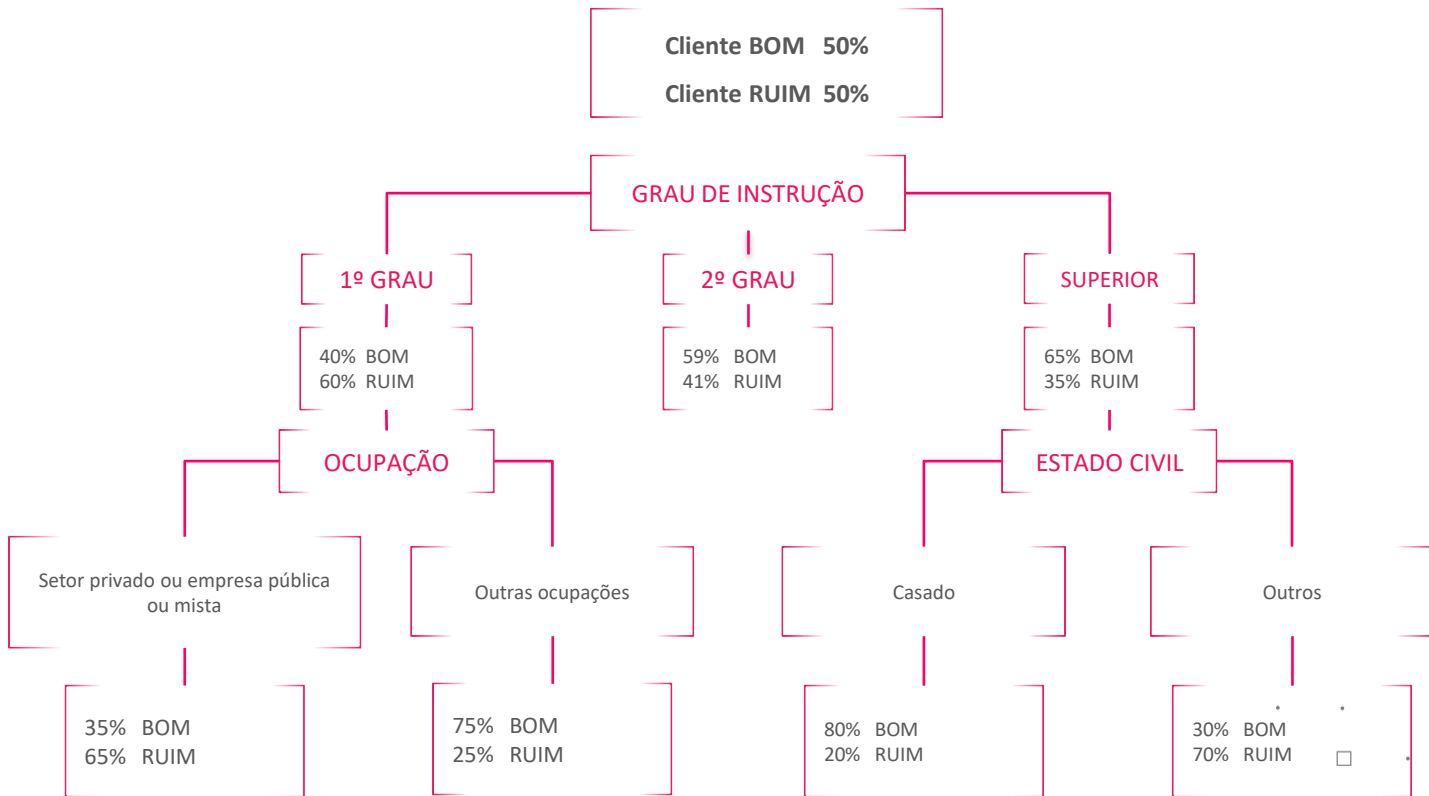


TÉCNICAS DE DISCRIMINAÇÃO

ÁRVORES DE DECISÃO

- Metodologia estatística de fácil interpretação e utilização.
 - São estruturas de dados compostas de um nó raiz e vários nós filhos, que por sua vez têm seus filhos também e se interligam por ramos, cada um representando uma regra. Os nós que não possuem filhos são chamados de nós folhas e os que têm são chamados de nós pais, ou de decisão.
- Têm como objetivo encontrar regras que discriminem dois grupos previamente conhecidos.
- Exemplo: Encontrar uma regra que trace perfil de pessoas mais propensas a aderir a um certo produto.

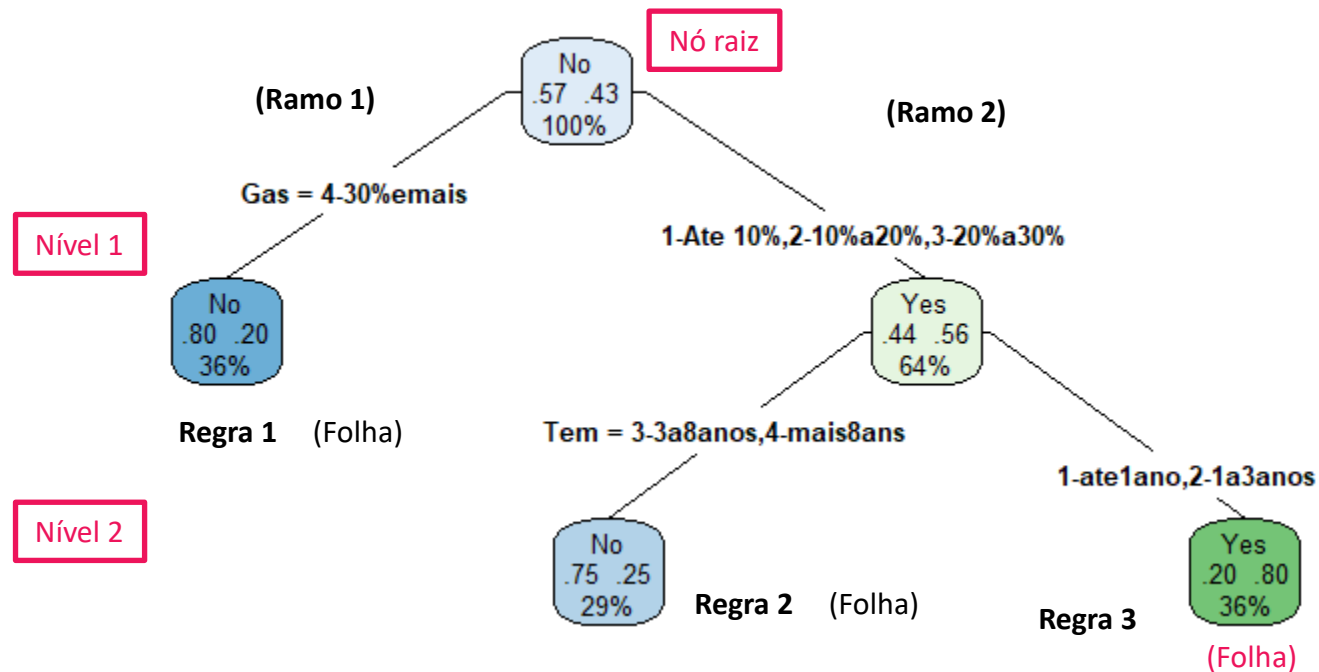
ÁRVORES DE DECISÃO – EXEMPLO



ÁRVORES DE DECISÃO – EXEMPLO

- Segmento: Área Financeira

A área de crédito deseja avaliar a propensão de um cliente tornar-se inadimplente.



ÁRVORES DE DECISÃO - PARÂMETROS

- Qual preditor e qual valor dividir os dados
- Profundidade e complexidade da árvore
- Resultado de cada folha

Tipos de Variáveis

- Todos os tipos de variáveis: Quantitativas e Categóricas (nominais ou ordinais)

ÁRVORES DE DECISÃO – ALGORITMOS

- Algoritmos utilizados para implementar uma árvore de decisão:

- CHAID: CHi-square Automatic Interaction Detector
- Algoritmo de Hunt-Szymanski (1976): diff (file comparison)
- 1R: regras usam só um atributo (Holte, 1993);
- ID3: binária categórica (Ross Quinlan, 1986);
- C4.5 similar a ID3 porém permite atributos numéricos;
- CART: Classification And Regression Trees (similar a C4.5 porém permite regressão);
- CART: C5.0: última versão que usa menos memória;

ÁRVORES DE DECISÃO – CRITÉRIOS

- Como selecionar os valores limites dos atributos para realizar a melhor partição do dados?
- A seleção dos “nós” a serem utilizados na árvore é baseada na Teoria da Informação de Shannon, mais especificamente nos conceitos de entropia e ganho de informação
- Existem varias métricas (chamadas de “medidas de impureza”):
 - Entropia;
 - Índice de Gini;
 - Erro de classificação;

Obs: CART usa Gini ID3 e C4.5 usa Entropia

No ScikitLearn existe um hiperparâmetro chamado “criterion” onde se pode definir o uso de Gini ou Entropia (default=gini)

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

ÁRVORES DE DECISÃO – CRITÉRIOS

$$\text{Entropy}(t) = - \sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t),$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2,$$

$$\text{Classification error}(t) = 1 - \max_i [p(i|t)],$$

Onde:

i : classe alvo entre “ c ” classes possíveis e,

t : nó da árvore,

$p(i|t)$: frequência que a classe i aparece dentro do nó.

ÁRVORES DE DECISÃO – CRITÉRIOS

- Independente da métrica de impureza, em todas as árvores o que se busca é o **Ganho de Informação**.

Ganho de informação: É a redução esperada da entropia ao utilizarmos um atributo na árvore

O ganho de informação é dado por:

$$\text{Ganho}(S, A) = \text{Entropia}(S) - \sum (|S_v| / |S|) * \text{Entropia}(S_v)$$

Onde:

Ganho (S, A) é o ganho do atributo A sobre o conjunto S

S_v = subconjunto de S para um valor do atributo A

$|S_v|$ = número de elementos de S_v

$|S|$ = número de elementos de S

EXEMPLO – MODELO DE PROPENSÃO

Segmento: Seguro Residencial

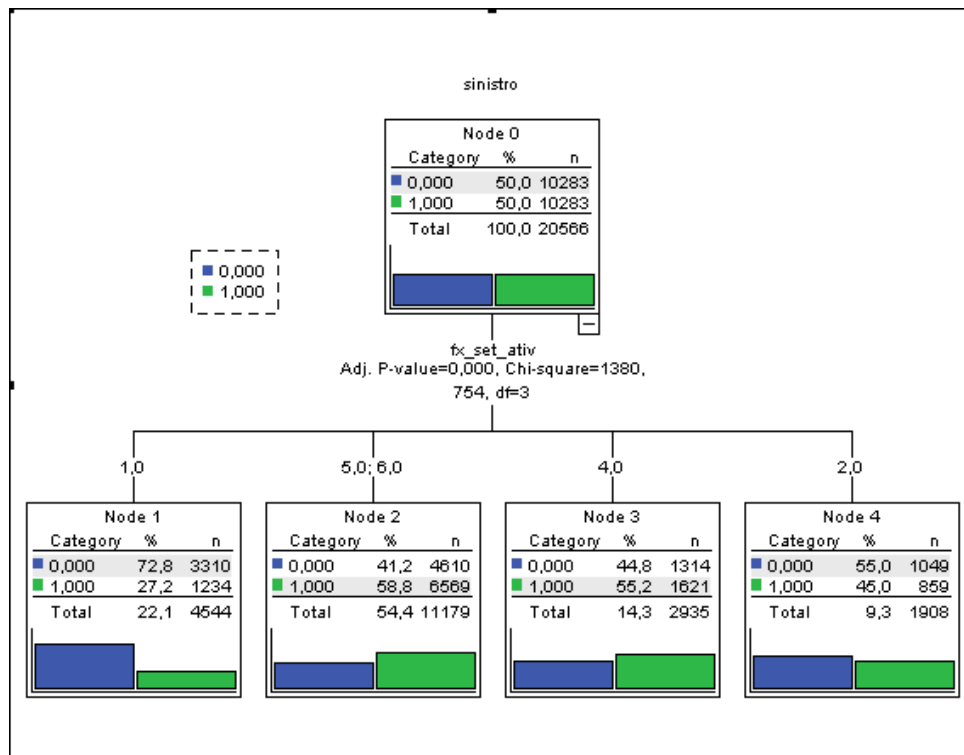
A área de Seguros deseja avaliar a propensão de um novo cliente sinistrar na apresentação de uma proposta.

Exemplo – Amostra de apólices

| <u>apolice</u> | <u>parcelas</u> | <u>qtde_cob</u> | <u>tpconstr</u> | <u>tipmora</u> | <u>clasmora</u> | <u>corretor</u> | <u>corrent</u> | <u>uf</u> | <u>set_ativ</u> | <u>Impseg(R\$)</u> | <u>sinistro</u> |
|----------------|-----------------|-----------------|-----------------|----------------|-----------------|-----------------|----------------|-----------|-----------------|--------------------|-----------------|
| 925578 | 6 | 9 | 6 | casa | moradia | 2 | N | MS | 90 | 100000 | 1 |
| 395699 | 1 | 9 | 6 | apto | moradia | 1 | S | ES | 26 | 30000 | 0 |
| 863771 | 11 | 9 | 6 | casa | moradia | 1 | S | SP | 24 | 200000 | 0 |
| 892165 | 11 | 9 | 6 | casa | moradia | 1 | S | MG | 27 | 30000 | 0 |
| 923092 | 1 | 9 | 6 | casa | veraneio | 2 | N | SP | 90 | 70000 | 0 |
| 1003098 | 4 | 9 | 6 | casa | veraneio | 1 | S | SP | 7 | 150000 | 1 |
| 955644 | 11 | 9 | 6 | casa | moradia | 1 | S | MG | 11 | 30000 | 1 |
| 987421 | 1 | 9 | 6 | casa | moradia | 2 | N | SP | 90 | 65000 | 1 |
| 744959 | 4 | 9 | 6 | casa | veraneio | 1 | S | RS | 18 | 70000 | 1 |
| 920814 | 11 | 9 | 6 | casa | moradia | 2 | S | SP | 90 | 100000 | 0 |
| 395550 | 2 | 9 | 6 | casa | moradia | 1 | S | ES | 26 | 20000 | 0 |
| 972615 | 6 | 9 | 6 | casa | veraneio | 2 | N | SP | 90 | 87500 | 1 |
| 958900 | 11 | 9 | 6 | casa | moradia | 1 | S | MG | 23 | 85000 | 1 |
| 911272 | 4 | 9 | 6 | casa | veraneio | 2 | N | SP | 90 | 150000 | 0 |
| 895508 | 11 | 9 | 6 | casa | moradia | 1 | S | MG | 33 | 50000 | 0 |
| 374234 | 1 | 9 | 6 | apto | moradia | 1 | N | DF | 6 | 30000 | 0 |
| 883254 | 11 | 9 | 6 | casa | moradia | 1 | S | SP | 24 | 100000 | 0 |
| 727885 | 3 | 9 | 6 | casa | moradia | 2 | S | RS | 90 | 180000 | 1 |
| 327315 | 11 | 9 | 6 | casa | moradia | 1 | S | BA | 21 | 20000 | 0 |
| 910241 | 11 | 9 | 6 | apto | moradia | 1 | S | SP | 49 | 50000 | 0 |
| 956554 | 10 | 9 | 6 | casa | moradia | 1 | S | MG | 27 | 70000 | 1 |
| 1000162 | 3 | 9 | 6 | casa | moradia | 2 | S | MS | 90 | 80000 | 1 |
| 920421 | 1 | 9 | 6 | casa | veraneio | 1 | S | SP | 1 | 40000 | 1 |

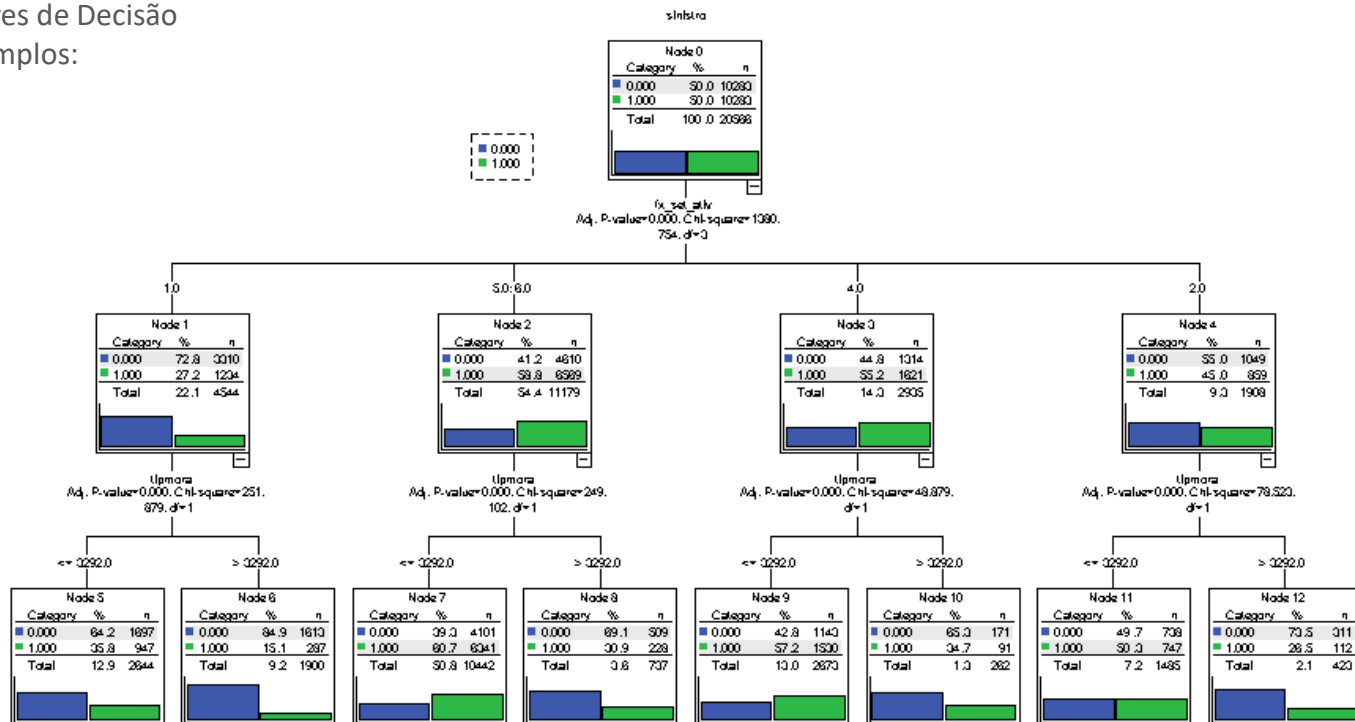
ÁRVORES DE DECISÃO – EXEMPLO

Árvores de Decisão
- exemplos:



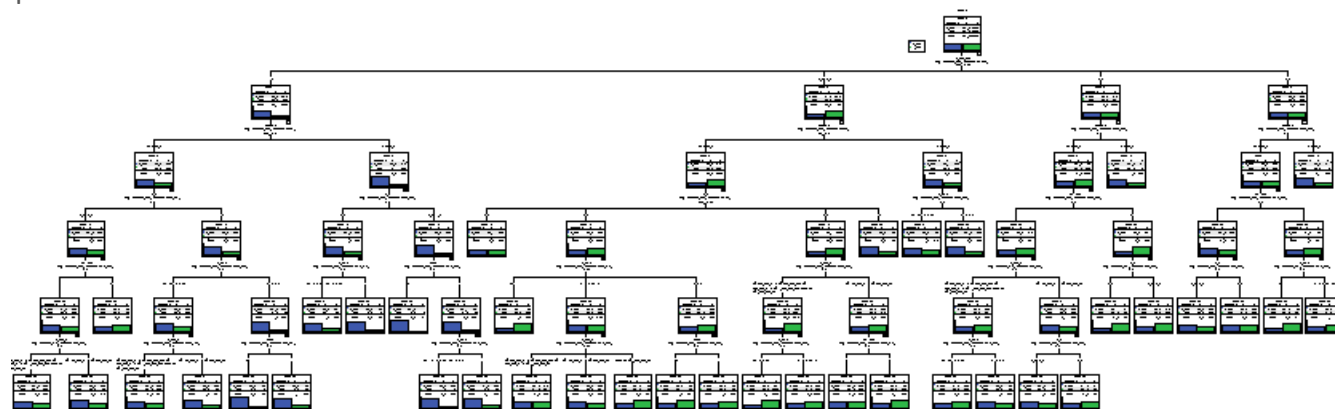
ÁRVORES DE DECISÃO – EXEMPLO

Árvores de Decisão
- exemplos:



ÁRVORES DE DECISÃO – EXEMPLO

Árvores de Decisão
- exemplos:



AVALIAÇÃO DO MODELO

Exemplo

| Classification | | | |
|--|-----------|-------|-----------------|
| Observed | Predicted | | |
| | 0 | 1 | Percent Correct |
| 0 | 5.137 | 999 | 83,7% |
| 1 | 1.208 | 4.412 | 78,5% |
| Overall Percentage | 81,0% | 81,5% | 81,2% |
| Growing Method: EXHAUSTIVE CHAID Dependent Variable: Resposta | | | |

- • + ● □

DEGRESS

Encontrar

TÉCNICA DE CLASSIFICAÇÃO

REGRESSÃO LOGÍSTICA

É uma técnica recomendada para situações em que a variável dependente é de natureza dicotômica ou binária.

Quanto às independentes, tanto podem ser categóricas ou não.

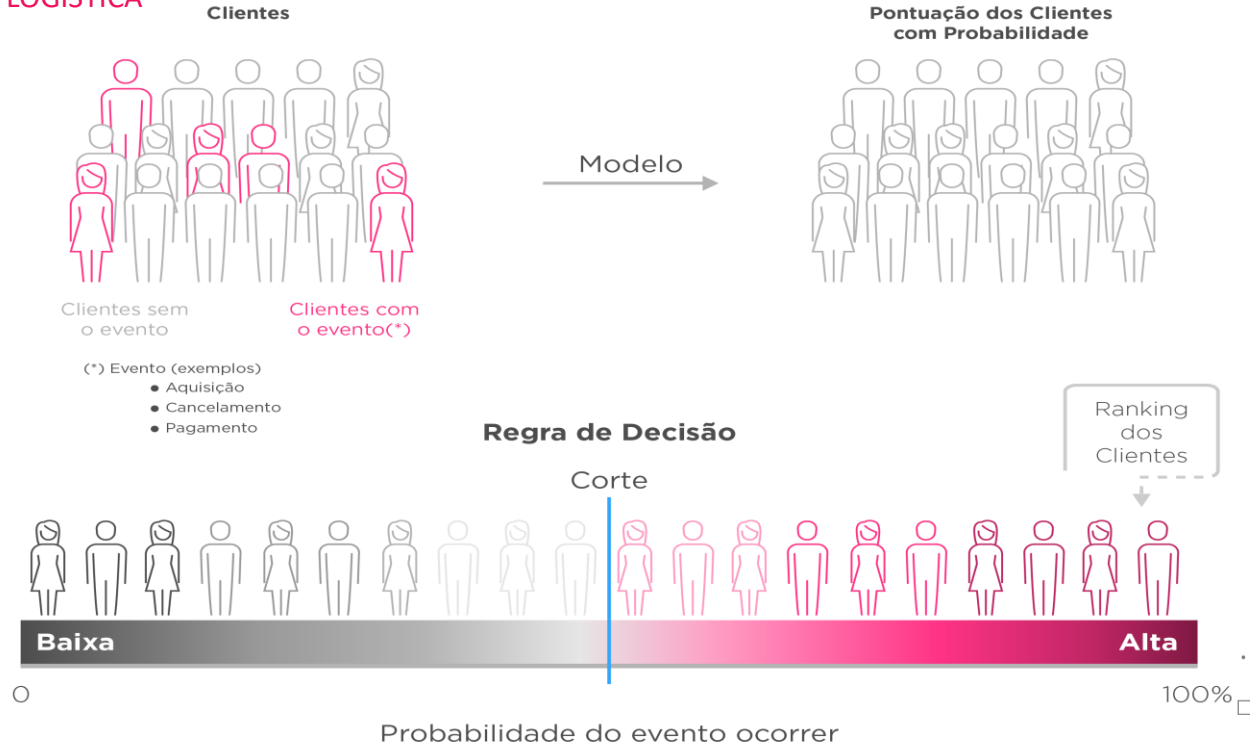
A regressão logística é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis preditoras.

Busca estimar a probabilidade da variável dependente assumir um determinado valor em função dos conhecidos de outras variáveis;

Os resultados da análise ficam contidos no intervalo de zero a um.

ANÁLISE DE REGRESSÃO LOGÍSTICA

REGRESSÃO LOGÍSTICA



ANÁLISE DE REGRESSÃO LOGÍSTICA

Probabilidade

Sendo Y: a resposta à preferência por um evento (sim ou não),

- a probabilidade de:
 - Preferência (*ou sucesso*) será p
 - Não-preferência (*de fracasso*) será (1-p)

“Chance de Ocorrência de um Evento”

- Chance = (probabilidade de sucesso) / (probabilidade de fracasso)

Exemplo, se a probabilidade de sucesso é 0,65:

→ : a chance é igual a: $p / (1-p) = p / q = 0,65 / 0,35 = 1,86$

ANÁLISE DE REGRESSÃO LOGÍSTICA

Exemplo: Preferência por canal de futebol

| Sexo | Prefere | Não prefere | Total |
|-----------|---------|-------------|-------|
| Masculino | 146 | 120 | 266 |
| Feminino | 110 | 124 | 234 |
| Total | 256 | 244 | 500 |

- **Chance** de preferir o canal de futebol entre **homens**:
 - $p_1 / (1-p_1) = (146/266) / (120/266) = 0,55 / 0,45 = 1,22$
- **Chance** de preferir o canal de futebol entre **mulheres**:
 - $p_2 / (1-p_2) = (110/234) / (124/234) = 0,47 / 0,53 = 0,89$
- **Razão de chances** de preferir canal de futebol **entre homens, em relação às mulheres**:
 - $[p_1/(1-p_1)] / [p_2/(1-p_2)] = 1,22 / 0,89 \doteq 1,37$

ANÁLISE DE REGRESSÃO LOGÍSTICA

Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada diretamente. No caso da variável dependente Y assumir apenas dois possíveis estados (1 ou 0) e haver um conjunto de p variáveis independentes X_1, X_2, \dots, X_p , o modelo de regressão logística pode ser escrito da seguinte forma:

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}}$$

onde, $g(x) = a + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$

Quando $g(x) \rightarrow -\infty$, então $P(Y=1) \rightarrow 0$
 Quando $g(x) \rightarrow +\infty$, então $P(Y=1) \rightarrow 1$

ANÁLISE DE REGRESSÃO LOGÍSTICA

Modelo de Regressão Logística

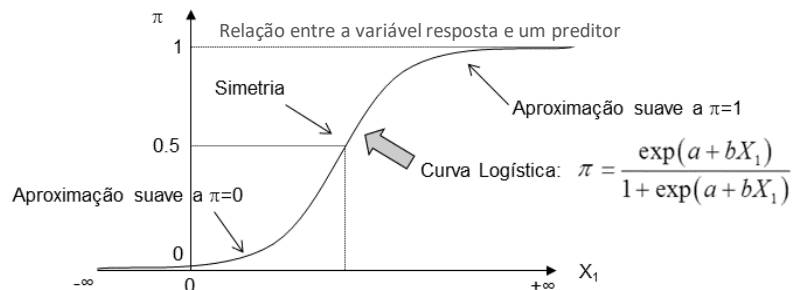
$$G = a + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$$

G: logit da resposta de preferência (sim) a :

Intersecção B_1, B_2, \dots, B_n : coeficientes logísticos

A função logística é dada pelo logito-inverso (anti-logit) que nos permite transformar o logito em probabilidade:

$$p = \frac{\exp(x)}{1 + \exp(x)}$$



ANÁLISE DE REGRESSÃO LOGÍSTICA

Método de Estimação dos Coeficientes

- Regressão Linear: Método dos Mínimos Quadrados
 - É o método que determina a linha reta mais apropriada, minimizando a soma dos quadrados das diferenças entre os valores estimados de Y por meio da reta de regressão e os valores observados de Y.
- Logística: Método da Máxima Verossimilhança (algoritmo iterativo)

Consiste em determinar uma função, denominada função de verossimilhança [$L(y, \vartheta)$], que é a função de probabilidade de ocorrência de um específico conjunto de dados e estimar os parâmetros que a maximizam.

ANÁLISE DE REGRESSÃO LOGÍSTICA

Método de Estimação dos Coeficientes

- Logística: Método da Máxima Verossimilhança (algoritmo iterativo)

Os coeficientes B_0, B_1, \dots, B_p são estimados a partir do conjunto dados, pelo método da máxima verossimilhança, em que encontra uma combinação de coeficientes que maximiza a probabilidade da amostra ter sido observada. Considerando uma certa combinação de coeficientes B_0, B_1, \dots, B_p e variando os valores de X .

Observa-se que a curva logística tem um comportamento probabilístico no formato da letra S, o que é uma característica da regressão logística. (Hosmer e Lemeshow ,1989)

ANÁLISE DE REGRESSÃO LOGÍSTICA

Classificação

Para utilizar o modelo de regressão logística para discriminação de dois grupos, a regra de classificação, **se a amostra for balanceada**, é a seguinte:

- se $P(Y=1) > 0,5$ então classifica-se $Y=1$
- se $P(Y=1) < 0,5$ então classifica-se $Y=0$

ANÁLISE DE REGRESSÃO LOGÍSTICA

Seleção Conjuntos de Atributos (Variáveis)

- Variáveis Discriminantes

Instrumento para selecionar variáveis (atributos) significativos

BACKWARD

FORWARD

STEPWISE

- **Backward Selection:** Procedimento constrói **adicionando todas as variáveis** e vai eliminando iterativamente uma a uma até que não haja mais variáveis.

- **Forward Selection:** Procedimento constrói iterativamente **adicionando variáveis uma a uma** até que não haja mais variáveis preditoras.

- **Stepwise:** Combinação de Forward Selection e Backward elimination. Procedimento constrói iterativamente uma sequência de modelos pela adição ou remoção de variáveis em cada etapa.

ANÁLISE DE REGRESSÃO LOGÍSTICA

Qualificação do Ajuste do Modelo

- Matriz de Classificação
- Estatística de Ajuste
- Verossimilhança : $-2 \log \text{Verossimilhança}$
- Significância do Modelo : Qui-quadrado (similar ao F regressão)
- Ganho no Modelo (significância)

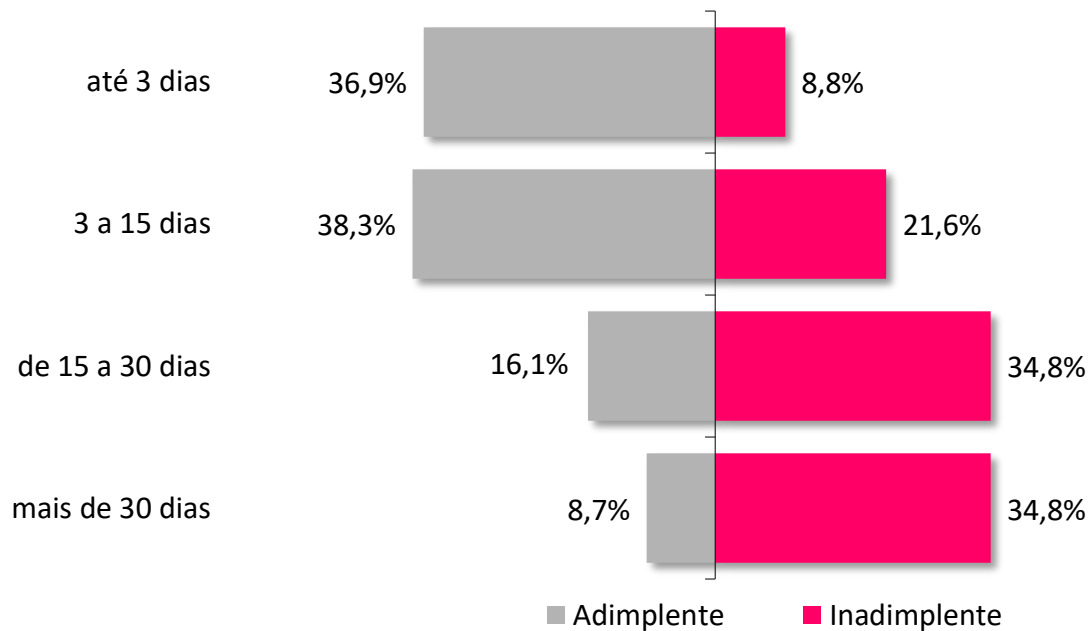
EXEMPLO – MODELO DE INADIMPLÊNCIA

Segmento: Cartões de Crédito

A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência.

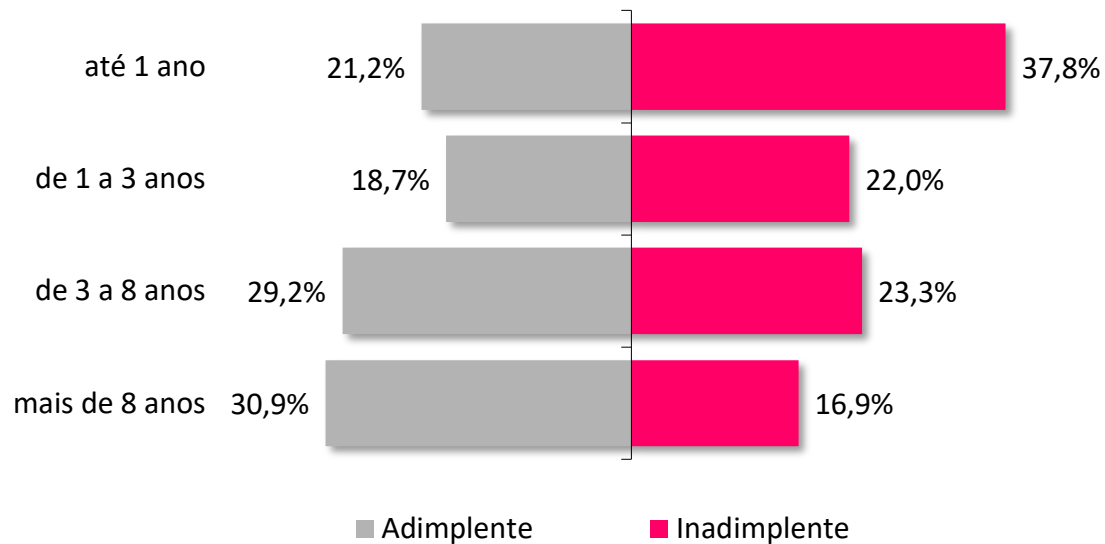
EXEMPLO – MODELO DE INADIMPLÊNCIA

Média de dias com pagamentos em atraso nos últimos 6 meses



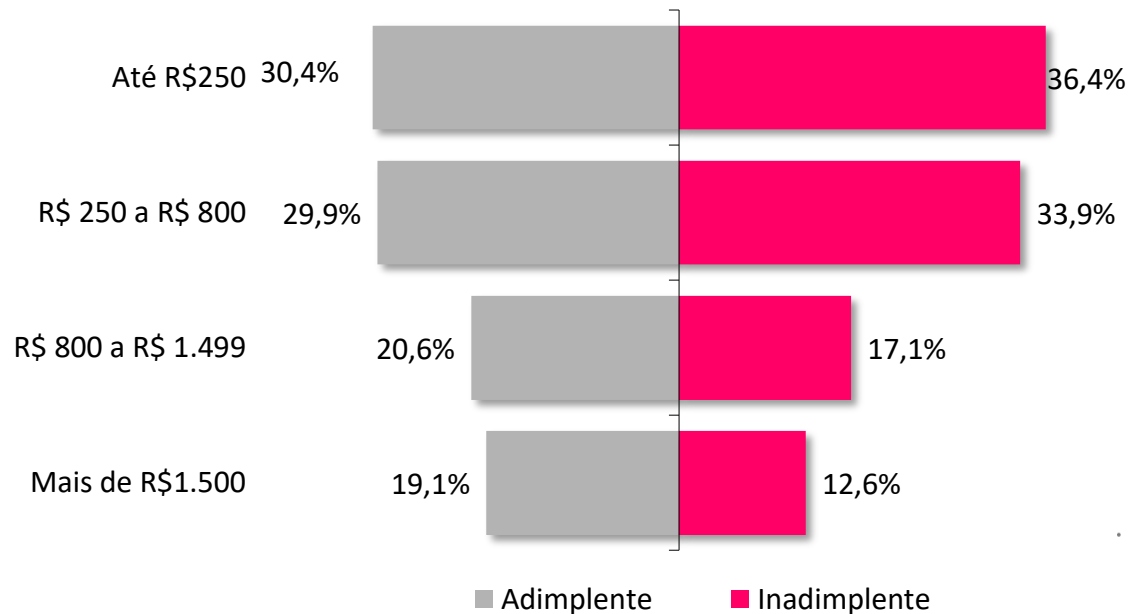
EXEMPLO – MODELO DE INADIMPLÊNCIA

Tempo de relacionamento em anos



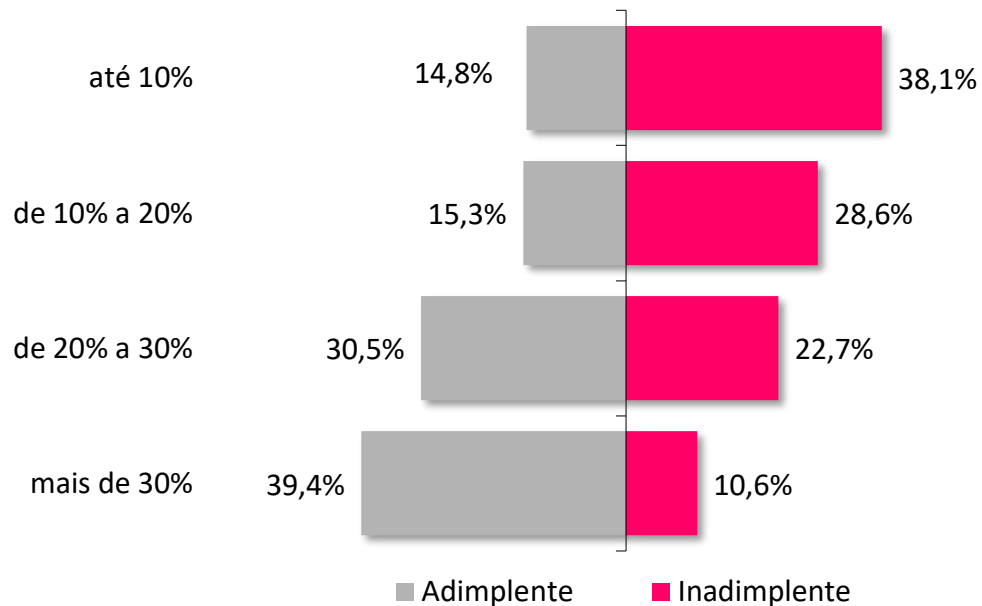
EXEMPLO – MODELO DE INADIMPLÊNCIA

Valor Médio da Fatura Mensal



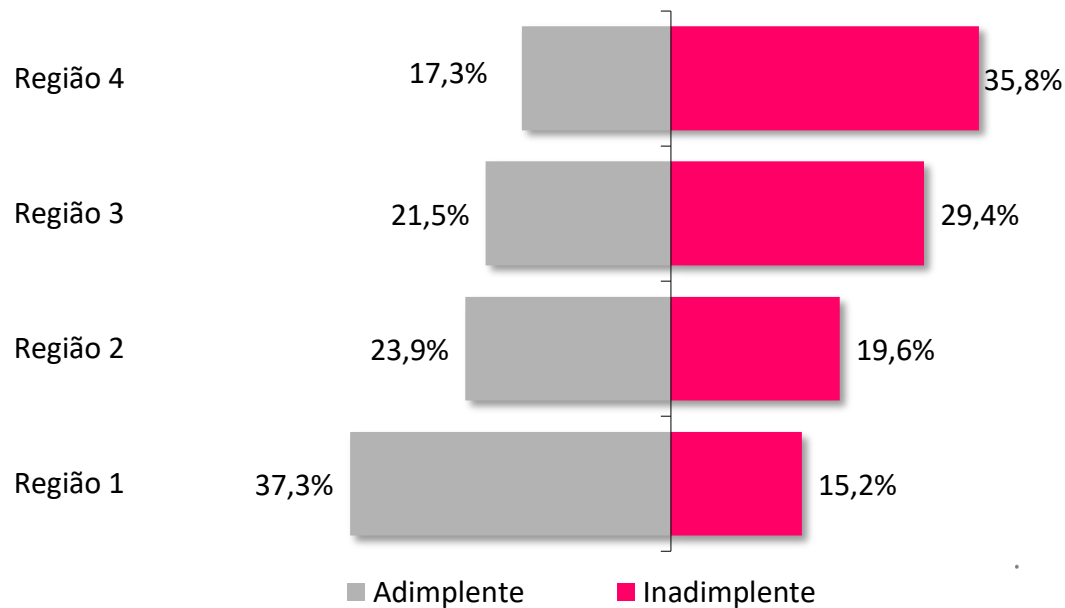
EXEMPLO – MODELO DE INADIMPLÊNCIA

Percentual dos gastos em alimentação



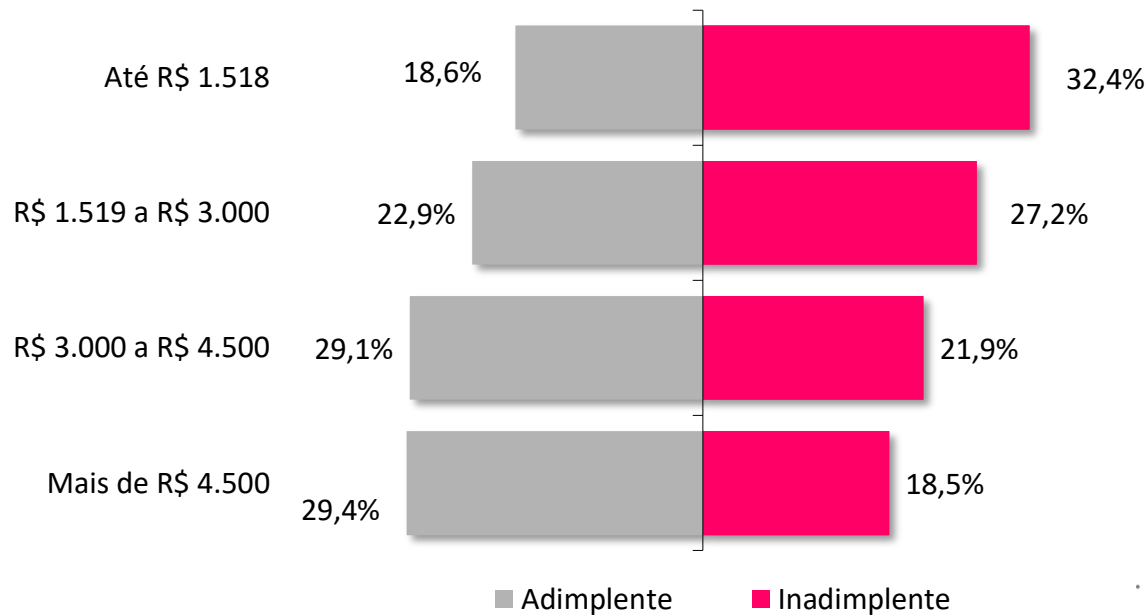
EXEMPLO – MODELO DE INADIMPLÊNCIA

Regiões de Risco



EXEMPLO – MODELO DE INADIMPLÊNCIA

Renda média mensal



EXEMPLO – MODELO DE INADIMPLÊNCIA

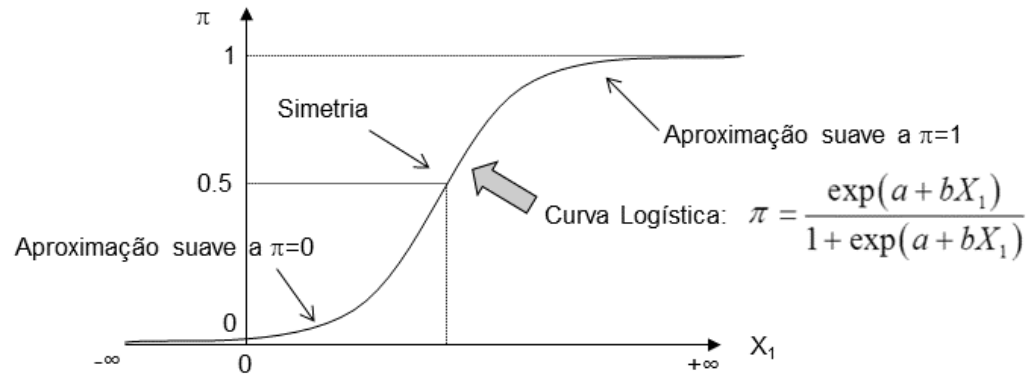
Tabela de Coeficientes do Modelo

| variável | categoria | Coeficientes |
|----------------------------|-----------------------|--------------|
| fatura em atraso | até 3 dias | -1,276 |
| | 3 a 15 dias | -0,611 |
| | de 15 a 30 dias | 0,580 |
| | mais de 30 dias | 1,308 |
| Tempo de cliente | até 1 ano | 0,580 |
| | de 1 a 3 anos | 0,401 |
| | de 3 a 8 anos | -0,264 |
| | mais de 8 anos | -0,718 |
| valor da fatura | Até R\$250 | 0,262 |
| | R\$ 250 a R\$ 800 | 0,103 |
| | R\$ 800 a R\$ 1.499 | -0,105 |
| | Mais de R\$1.500 | -0,261 |
| % de gasto com alimentação | até 10% | 0,581 |
| | de 10% a 20% | 0,401 |
| | de 20% a 30% | -0,264 |
| | mais de 30% | -0,718 |
| Região de Risco | Região 4 | 1,067 |
| | Região 3 | 0,371 |
| | Região 2 | -0,368 |
| | Região 1 | -1,069 |
| renda mensal | Até R\$ 1.518 | 0,455 |
| | R\$ 1.519 a R\$ 3.000 | 0,080 |
| | R\$ 3.000 a R\$ 4.500 | -0,122 |
| | Mais de R\$ 4.500 | -0,413 |
| Constante | | 0,099 |

EXEMPLO – MODELO DE INADIMPLÊNCIA

A partir da somatória dos pesos dos atributos de todas as variáveis (x) →

$$p = \frac{\exp(x)}{1 + \exp(x)}$$



EXEMPLO – MODELO DE INADIMPLÊNCIA

Modelo Logístico

Pesos definidos na modelagem

| | | | | |
|--------|------------------|----------------------------|-----------------|-------|
| -1,276 | Até 3 dias | Fatura em atraso | Mais de 30 dias | 1,308 |
| -0,718 | Mais de 8 anos | Tempo de Relacionamento | Até 1 ano | 0,580 |
| -0,261 | Mais de R\$1.500 | Valor da Fatura | Até R\$250 | 0,262 |
| -0,718 | Mais de 30% | % de gasto com alimentação | Até 10% | 0,580 |
| -1,069 | Região 1 | Região de Risco | Região 4 | 1,067 |
| -0,413 | Mais de R\$4.500 | Renda Mensal | Até R\$1.518 | 0,455 |
| 0,099 | | Constante | | 0,099 |
| 4% | Propensão | | | 98% |

AVALIAÇÃO DO MODELO

Exemplo

| Classification | | | |
|--|-----------|-------|-----------------|
| Observed | Predicted | | |
| | 0 | 1 | Percent Correct |
| 0 | 5.137 | 999 | 83,7% |
| 1 | 1.208 | 4.412 | 78,5% |
| Overall Percentage | 81,0% | 81,5% | 81,2% |
| Growing Method: EXHAUSTIVE CHAID Dependent Variable: Resposta | | | |

ANÁLISE DE REGRESSÃO LOGÍSTICA

Exemplo: Modelo Cross-Selling

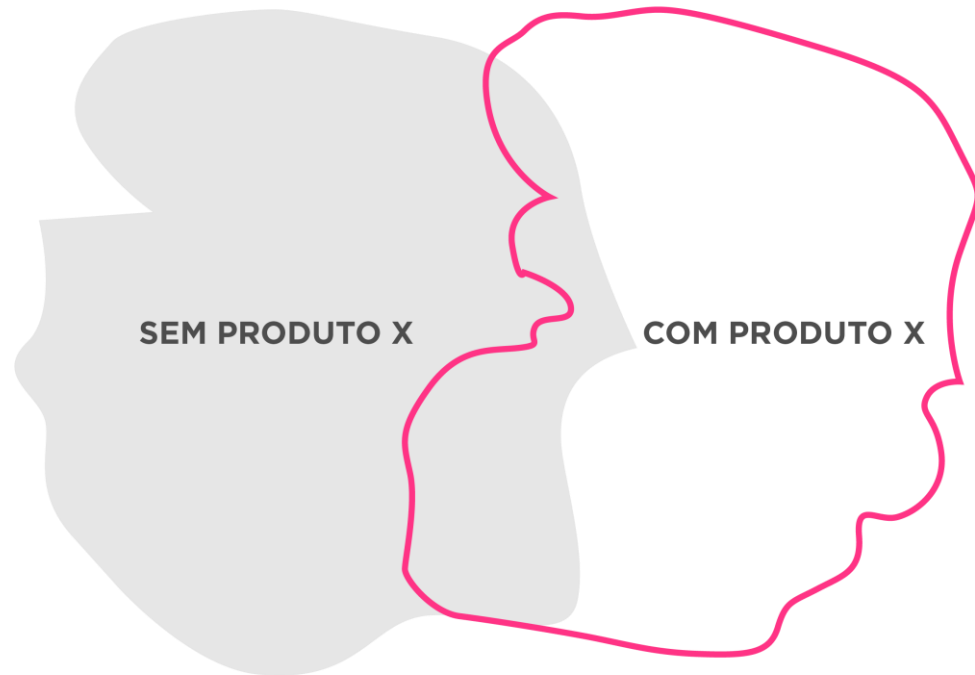
Propensão à Compra de um Produto

Objetivo:

Estabelecer público-alvo para a venda qualificada de um determinado Produto X, com uso dos mailing's internos do cliente, por meio do desenvolvimento de modelos preditivos.

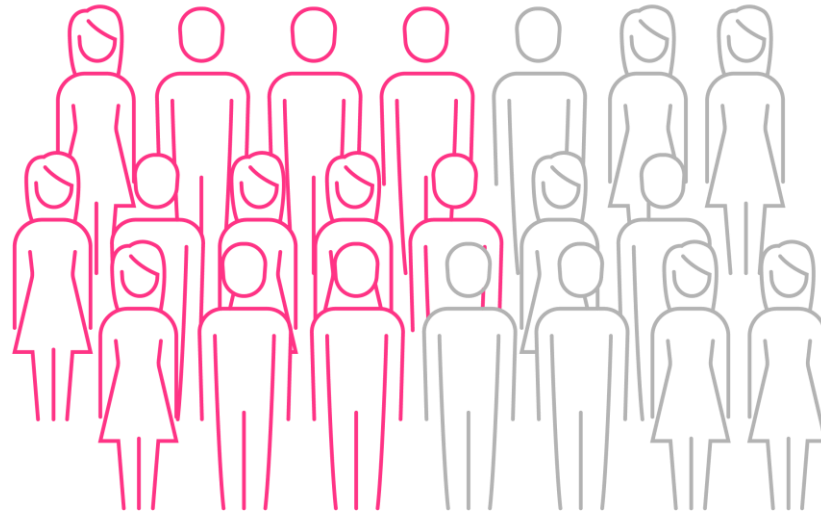
MODELOS CROSS SELLING

- Propensão de compra do Produto X



MODELOS CROSS SELLING

- Propensão de compra do Produto X



SEM PRODUTO X

COM PRODUTO X

MODELOS CROSS SELLING

Implementação

- Propensão de compra do Produto X

Algoritmo Matemático

Para associar uma probabilidade de compra de um produto X a cada cliente, os seguintes passos devem ser tomados:

1. Identificar as variáveis, associando os respectivos coeficientes;
2. Somar os coeficientes encontrados no item 1, juntamente com a constante do modelo determinando o valor de Y;
3. Efetuar a operação matemática que se segue, para determinação final do score.

$$\text{Probabilidade} = 100 \times e^{\frac{Y}{(1 + e)^Y}}$$

MODELOS CROSS SELLING

- Propensão de compra do Produto X

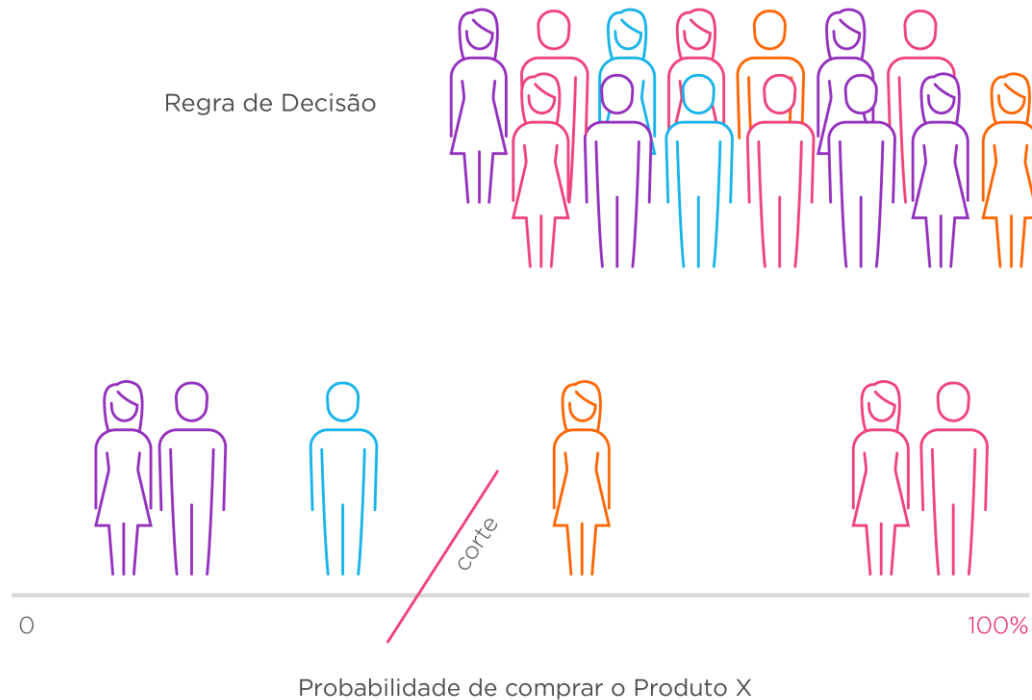
Regra de Decisão Estatística

Após associar a cada indivíduo sua probabilidade de compra do produto, deve-se submetê-la à Regra de Decisão, ou seja, se a probabilidade obtida for menor ou igual ao valor de corte* o assinante pertencerá ao grupo que não irá adquirir o produto, caso contrário, se essa probabilidade for maior que o valor de corte, ele pertencerá ao grupo que irá adquirir.

* valor de corte é o valor de probabilidade que define os grupos, segundo análise de acertos do modelo.

MODELOS CROSS SELLING

- Propensão de compra do Produto X



AVALIAÇÃO DO MODELO PREDITIVO

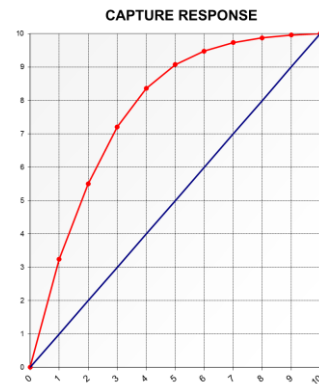
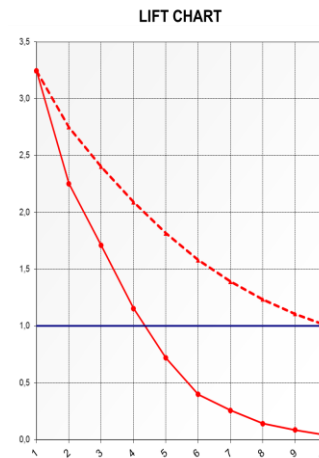
Exemplo:

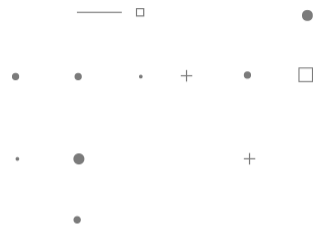
| Decil | Clientes Qtde | Base sem utilização de modelos | | | | | | |
|-------|----------------------|--------------------------------|-------|-------|------|-------------|---------|--------|
| | | Penetração | | | Lift | Lift Ac. | Capture | |
| | | Qtde | % | % Ac. | | | % | % Ac. |
| 1 | 5.300 | 612 | 11,5% | 11,5% | 1,00 | 1,00 | 10,0% | 10,0% |
| 2 | 5.300 | 612 | 11,5% | 11,5% | 1,00 | 1,00 | 10,0% | 20,0% |
| 3 | 5.300 | 612 | 11,5% | 11,5% | 1,00 | 1,00 | 10,0% | 30,0% |
| 4 | 5.300 | 612 | 11,5% | 11,5% | 1,00 | 1,00 | 10,0% | 40,0% |
| 5 | 5.300 | 612 | 11,5% | 11,5% | 1,00 | 1,00 | 10,0% | 50,0% |
| 6 | 5.300 | 612 | 11,5% | 11,5% | 1,00 | 1,00 | 10,0% | 60,0% |
| 7 | 5.300 | 612 | 11,5% | 11,5% | 1,00 | 1,00 | 10,0% | 70,0% |
| 8 | 5.300 | 612 | 11,5% | 11,5% | 1,00 | 1,00 | 10,0% | 80,0% |
| 9 | 5.300 | 612 | 11,5% | 11,5% | 1,00 | 1,00 | 10,0% | 90,0% |
| 10 | 5.300 | 612 | 11,5% | 11,5% | 1,00 | 1,00 | 10,0% | 100,0% |
| Total | 53.000 | 6.120 | 11,5% | | | | | |

AValiação DO MODELO PREDITIVO

Exemplo:

| Decil | Clientes | Aplicando Modelo | | | | | | |
|-------|----------|------------------|-------|-------|------|----------|---------|--------|
| | | Penetração | | | Lift | | Capture | |
| | Qtde | Qtde | % | % Ac. | Lift | Lift Ac. | % | % Ac. |
| 1 | 5.300 | 1.986 | 37,5% | 37,5% | 3,24 | 3,24 | 32,4% | 32,4% |
| 2 | 5.300 | 1.379 | 26,0% | 31,7% | 2,25 | 2,75 | 22,5% | 55,0% |
| 3 | 5.300 | 1.046 | 19,7% | 27,7% | 1,71 | 2,40 | 17,1% | 72,1% |
| 4 | 5.300 | 706 | 13,3% | 24,1% | 1,15 | 2,09 | 11,5% | 83,6% |
| 5 | 5.300 | 440 | 8,3% | 21,0% | 0,72 | 1,82 | 7,2% | 90,8% |
| 6 | 5.300 | 244 | 4,6% | 18,2% | 0,40 | 1,58 | 4,0% | 94,8% |
| 7 | 5.300 | 157 | 3,0% | 16,1% | 0,26 | 1,39 | 2,6% | 97,3% |
| 8 | 5.300 | 87 | 1,6% | 14,3% | 0,14 | 1,23 | 1,4% | 98,8% |
| 9 | 5.300 | 52 | 1,0% | 12,8% | 0,08 | 1,11 | 0,8% | 99,6% |
| 10 | 5.300 | 24 | 0,5% | 11,5% | 0,04 | 1,00 | 0,4% | 100,0% |
| Total | 53.000 | 6.120 | 11,5% | | | | | |





CLASSIFICADORES BAYESIANOS



CLASSIFICADORES BAYESIANOS

Classificadores Bayesianos são **modelos probabilísticos** que usam informação sobre a distribuição dos dados e a distribuição de probabilidade de pertencimento a cada classe.

CLASSIFICADORES BAYESIANOS

INFERÊNCIA BAYESIANA



PROBABILIDADES SUBJETIVAS

O fenômeno aleatório pode ser separado em etapas. A informação que ocorreu em uma determinada etapa pode influenciar nas probabilidades de ocorrências das etapas sucessivas.

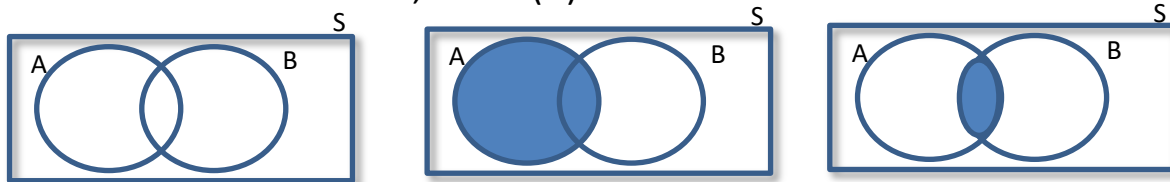
Definição:

Dados dois eventos A e B, a probabilidade condicional de A dado que ocorreu B é representado por $P(A/B)$ e dada por:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

PROBABILIDADE CONDICIONAL

Dados dois eventos A e B, com $P(A) \neq 0$:



$$\Rightarrow P(B | A) = \frac{P(A \cap B)}{P(A)}.$$

para o cálculo de $P(A \cap B)$:

$$P(A \cap B) = P(A) \cdot P(B | A)$$

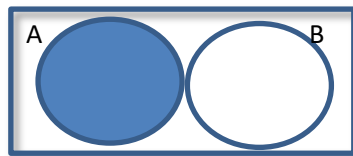
$$P(A \cap B) = P(B) \cdot P(A | B)$$

Teoria

PROBABILIDADE CONDICIONAL

Se os eventos forem independentes $P(A \text{ e } B) = P(A) * P(B)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B)}{P(B)} = P(A)$$



Eventos independentes:
são aqueles cuja probabilidade
de ocorrer não depende do
que aconteceu antes.

PROBABILIDADE CONDICIONAL

A técnica de Basket Analysis utiliza a probabilidade condicional para encontrar cestas de produtos. Algoritmo Apriori .

Um Exemplo de Sucesso!



- ✓ Descobriu-se que homens entre trinta e quarenta e cinco anos, que compram cervejas, nas sextas-feiras, após as dezesseis horas, também compram fraldas!
- ✓ Resultado: apenas mudando os produtos de lugar, colocando as fraldas ao lado de cervejas nos pontos de venda, obteve-se um aumento de mais de quarenta por cento nas vendas de fraldas.
- ✓ O que acha de possuir uma informação como essa?

A Wall-Mart soube tirar bom proveito dela!

- Support (frequência)

$$(A \cap B \Rightarrow C) = \%$$

- Confidence (probabilidade condicional)

$$(A \cap B \Rightarrow C) = \frac{P(A \cap B \cap C)}{P(A \cap B)}$$

- Lift(associação)

$$(A \& B \Rightarrow C) = \frac{P(A \cap B \cap C)}{P(A \cap B)P(C)}$$

TEOREMA DE BAYES

Sendo que:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{e} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Podemos isolar $P(A \cap B)$:

$$P(A \cap B) = P(A|B)P(B)$$

$$P(A \cap B) = P(B|A)P(A)$$

Assim chegamos no teorema de Bayes:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

TEOREMA DE BAYES

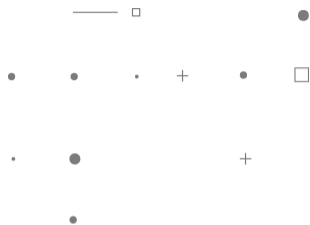
O Teorema de Bayes trata de problemas em que se deseja determinar a probabilidade de um evento ocorrer dada uma condição: (Probabilidade de ocorrer A, na condição de que B já tenha ocorrido – Probabilidade Condicional).

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Onde,

- $P(A)$ e $P(B)$ são as probabilidades a priori de A e B.
- $P(B|A)$ e $P(A|B)$ são as probabilidades posteriores de B condicional a A e de A condicional a B, respectivamente.

NAÏVE BAYES



USANDO TEOREMA DE BAYES

A técnica Naïve Bayes é um dos algoritmos de aprendizado supervisionado mais simples. Tem como premissa a suposição de independência entre as variáveis do problema (teorema de Bayes). Isto é, o classificador Naïve Bayes assume que o efeito de uma determinada variável é independente de outras variáveis. Mesmo que essas características dependam umas das outras, todas contribuem independentemente para a probabilidade e, é por isso que ele é conhecido como “naïve” (ingênuo).

USANDO TEOREMA DE BAYES

Considerando a premissa que os eventos são independentes:

A ocorrência do evento A em nada interfere na probabilidade de ocorrência do outro evento B.

A probabilidade de ambos ocorrerem é igual ao produto de suas probabilidades.

Depois de calcular a probabilidade a posteriori para várias classes diferentes, pode-se selecionar a classe com a maior probabilidade.

USANDO TEOREMA DE BAYES

Usando este teorema, o algoritmo de classificação Naïve Bayes calcula a probabilidade de cada classe para uma dada observação do conjunto de dados de análise como:

$$P(C = c_i | X_1 = x_1, \dots, X_p = x_p) = \frac{P(C = c_i) P(X_1 = x_1, \dots, X_p = x_p | C = c_i)}{P(X_1 = x_1, \dots, X_p = x_p)}$$

Em que $c_i (i=1,2)$ é a classe e x_1, \dots, x_p são os valores observados dos preditores (X_1, \dots, X_p) para uma dada observação do conjunto de dados analisado.

USANDO TEOREMA DE BAYES

Assim, assumindo que os valores dos preditores são independentes entre si, o numerador pode ser reduzido e o denominador pode ser ignorado, uma vez que, é o mesmo para todas as classes, não afetando os valores relativos de suas probabilidades. A probabilidade de uma observação pertencer à classe $C = c_i$ é proporcional (\propto) à expressão:

$$P(C = c_i | X_1 = x_1, \dots, X_p = x_p) \propto P(C = c_i) \prod P(X_i = x_i | C = c_i)$$

Produtório

USANDO TEOREMA DE BAYES

No caso de classificação binária, ao calcular $P(C = c_1 | X_1 = x_1, \dots, X_p = x_p)$ se tem a probabilidade de uma determinada observação ser da classe $C = c_1$ dado que possui as características contidas em X_1, \dots, X_p .

Ao calcular o complementar dessa probabilidade, ou seja, $1 - P(C = c_1 | X_1 = x_1, \dots, X_p = x_p)$

se tem a probabilidade de a observação ser classificada na outra classe $C = c_2$.

O algoritmo Naïve Bayes classifica a observação na classe que tem a maior probabilidade.

$MAP(y_i) = \max ((P(y_i) \prod P(x^j / y_i)))$; para cada classe

Este método é denominado estimativa por MAP (do inglês, *Maximun A Posteriori*).

USANDO TEOREMA DE BAYES

detecção de spam:

Dado uma mensagem que contém uma determinada palavra “X” queremos computar a probabilidade dela ser um spam.

A maioria das pessoas acostumadas a receber e-mails sabe que essa mensagem provavelmente é spam, no entanto, como podemos “indicar” ao “gerenciador” para uma tomada de decisão?

→ Sabemos que a palavra “X” está presente no e-mail.

Qual a chance deste e-mail ser um SPAM (S)? $P(S / X)$

$P(S)$ = Probabilidade de ter um SPAM → 10%

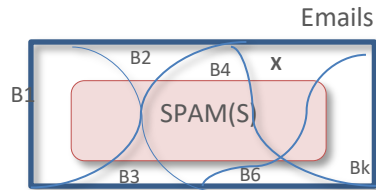
$P(NS)$ = Probabilidade de Não SPAM → 90%

$P(X/S)$ = Probabilidade de ter a palavra X aparecer em mensagens de SPAM's → 30%

$P(X/NS)$ = Probabilidade de ter a palavra X aparecer em mensagens sem SPAM's → 2%

$$\rightarrow P(S / X) = P(X / S) * P(S) / (P(X / S) * P(S) + P(X/NS) * P(NS))$$

$$\rightarrow P(S / X) = 30\% * 10\% / (30\% * 10\% + 2\% * 90\%) = 62,5\%$$



Exemplo: CLASSIFICADOR NAÏVE BAYES

Vamos acompanhar os cálculos seguindo um exemplo simples: a avaliação de concessão de um empréstimo (concessão: SIM ou NÃO). A tomada de decisão será em função das características do solicitante: seu salário, idade, local de moradia, histórico de empréstimos e transações anteriores. Mesmo que essas informações sejam interdependentes, eles ainda são considerados de forma independentes.

Exemplo: CLASSIFICADOR NAÏVE BAYES

- O que precisaremos é classificar se o solicitante de crédito vai ter ou não o crédito liberado, com base em suas características. Usaremos para esse exemplo como variáveis explicativas: estado civil, idade, se tem casa própria, situação de emprego. Variável Target: Crédito (SIM/NÃO).

| estadocivil | fx_etaria | casapropria | empregado | crédito |
|-------------|----------------|-------------|-----------|---------|
| SOLTEIRO | 26 a 34 anos | NAO | SIM | SIM |
| CASADO | 26 a 34 anos | SIM | NAO | SIM |
| DEMAIS | 26 a 34 anos | SIM | SIM | NAO |
| SOLTEIRO | 26 a 34 anos | SIM | SIM | SIM |
| CASADO | 35 e mais anos | NAO | NAO | SIM |
| ... | ... | ... | ... | ... |
| DEMAIS | 35 e mais anos | SIM | NAO | SIM |

Quadro 1 – Exemplo da base de dados

Exemplo: CLASSIFICADOR NAÏVE BAYES

Queremos calcular qual a probabilidade de ser liberado crédito para um solicitante dado que suas características são: seu estado civil é casado, sua idade é 30 anos, não tem cada própria e está empregado.

O classificador Naïve Bayes calcula a probabilidade de um evento nos seguintes passos:

- Passo 1: Calcular a probabilidade das classes (Probabilidade a priori):

$$P(\text{crédito}=\text{SIM}) = 11/20=0,55$$

$$P(\text{crédito}=\text{NAO}) = 9/20=0,45$$

Exemplo: CLASSIFICADOR NAÏVE BAYES

Passo 2: Criar a tabela de “Probabilidade” (likelihood) encontrando as probabilidades para cada atributo:

| | Crédito | | |
|--------------|---------|-----|--------------------------|
| estado civil | NAO | SIM | Cálculo de Probabilidade |
| casado | 2 | 5 | $7/20 = 0,35$ |
| demais | 3 | 2 | $5/20 = 0,25$ |
| solteiro | 4 | 4 | $8/20 = 0,40$ |

Quadro 2 – Cálculo de Probabilidade Estado Civil

| | Crédito | | |
|----------------|---------|-----|--------------------------|
| Faixa Etária | NAO | SIM | Cálculo de Probabilidade |
| 18 a 25 anos | 3 | 2 | $5/20 = 0,25$ |
| 26 a 34 anos | 4 | 4 | $8/20 = 0,40$ |
| 35 e mais anos | 2 | 5 | $7/20 = 0,35$ |

Quadro 3 – Cálculo de Probabilidade Faixa Etária

Exemplo: CLASSIFICADOR NAÏVE BAYES

- Passo 2: Criar a tabela de “Probabilidade” (likelihood) encontrando as probabilidades para cada atributo:

| | Crédito | | |
|--------------|---------|-----|--------------------------|
| Casa Própria | NAO | SIM | Cálculo de Probabilidade |
| NAO | 8 | 5 | $13/20 = 0,65$ |
| SIM | 1 | 6 | $7/20 = 0,35$ |

Quadro 4 – Cálculo de Probabilidade Ter Casa Própria

| | Crédito | | |
|-----------|---------|-----|--------------------------|
| Empregado | NAO | SIM | Cálculo de Probabilidade |
| NAO | 4 | 2 | $6/20 = 0,30$ |
| SIM | 5 | 9 | $14/20 = 0,70$ |

Quadro 5 – Cálculo de Probabilidade estar empregado

Exemplo: CLASSIFICADOR NAÏVE BAYES

- Passo 3: Agora, usamos a equação de Bayes para calcular a probabilidade posterior de cada classe. A classe com a maior probabilidade posterior é o resultado da previsão.

Calculando para o indivíduo com estas características: seu estado civil é casado, sua idade é 30 anos, não tem casa própria e está empregado.

→ $X = \{\text{estado civil} = \text{"casado"}; Fx_{\text{etaria}} = \text{"26 a 34 anos"}; \text{não tem casa própria e empregado} = \text{sim}\}$

Calculando na situação de Concessão de Crédito=SIM

Exemplo: CLASSIFICADOR NAÏVE BAYES

- Passo 3: Agora, usamos a equação de Bayes para calcular a probabilidade posterior de cada classe. A classe com a maior probabilidade posterior é o resultado da previsão.

Calculando para o indivíduo com estas características:

seu estado civil é casado, sua idade é 30 anos, não tem cada própria e está empregado.

- $P(\text{estado civil}=\text{"casado"} / \text{credito}=\text{SIM}) = 5/11$
- $P(\text{Fx_etaria}=\text{"26 a 34 anos"} / \text{credito}=\text{SIM}) = 4/11$
- $P(\text{ter cada própria}=\text{"Não"} / \text{credito}=\text{SIM}) = 5/11$
- $P(\text{empregado}=\text{"SIM"} / \text{credito}=\text{SIM}) = 9/11$

$P(\text{Sim} | \text{Casado}, 26\text{a}34\text{anos}, \text{NãoCasa}, \text{Empregado}) = ?$

⇒ $P(\text{Sim}/X) = P(\text{Casado}/\text{SIM})P(26\text{a}34\text{anos}/\text{SIM})P(\text{NãoCasa}/\text{SIM})P(\text{empregado}/\text{SIM}) = 0,0615$

Exemplo: CLASSIFICADOR NAÏVE BAYES

- Passo 3: Agora, usamos a equação de Bayes para calcular a probabilidade posterior de cada classe. A classe com a maior probabilidade posterior é o resultado da previsão.

Agora calculando na situação de Concessão de Crédito=NÃO

- $P(\text{estado civil}=\text{"casado"} / \text{credito}=\text{NAO}) = 2/9$
- $P(\text{Fx_etaria}=\text{"26 a 34 anos"} / \text{credito}=\text{NAO}) = 4/9$
- $P(\text{ter casa própria}=\text{"Não"} / \text{credito}=\text{NAO}) = 8/9$
- $P(\text{empregado}=\text{"SIM"} / \text{credito}=\text{NAO}) = 5/9$

⇒ $P(\text{Não} \mid \text{Casado}, 26\text{a}34\text{anos}, \text{NãoCasa}, \text{Empregado} = ?)$

⇒ $P(\text{Não} \mid \text{Casado}, 26\text{a}34\text{anos}, \text{NãoCasa}, \text{Empregado}) =$

⇒ $P(\text{Casado}/\text{NAO}) * P(26\text{a}34\text{anos}/\text{NAO}) * P(\text{NãoCasa}/\text{NAO}) * P(\text{empregado}/\text{NAO}) = 0,04877$

Exemplo: CLASSIFICADOR NAÏVE BAYES

A classe com a maior probabilidade posterior é o resultado da previsão.

$$\Rightarrow P(\text{Sim}/X) = P(\text{Casado}/\text{SIM})P(26\text{a}34\text{anos}/\text{SIM})P(\text{NãoCasa}/\text{SIM})P(\text{empregado}/\text{SIM}) = 0,0615$$

$$\Rightarrow P(\text{Casado}/\text{NAO}) * P(26\text{a}34\text{anos}/\text{NAO}) * P(\text{NãoCasa}/\text{NAO}) * P(\text{empregado}/\text{NAO}) = 0,04877$$

Como $P(\text{Sim}/X) > P(\text{Não}/X)$, classificamos esse solicitante com essas características como Crédito = SIM.

CLASSIFICADOR NAÏVE BAYES

Existem alguns tipos de modelo Naïve Bayes de acordo com as características das variáveis preditoras ou do tipo da variável target.

Gaussiana: Quando os atributos forem contínuos podemos optar por discretizar os dados, isto é, criar uma variável categorizada, ou **podemos considerar uma função de densidade de probabilidade no cálculo da probabilidade condicional** e, geralmente usa-se a Distribuição Normal. Se as funções contínuas não têm distribuição normal, devemos usar uma transformação ou métodos diferentes para convertê-las na distribuição normal.

CLASSIFICADOR NAÏVE BAYES

Bernoulli: O modelo binomial é útil se os **vetores são binários** (ou seja, zeros e uns). Uma aplicação seria de classificação de texto com um modelo conjunto de palavras onde 1 é quando a “palavra ocorre no documento” e 0, caso contrário, “palavra não ocorre. No caso das distribuições Bernoulli, cada variável só pode ter dois valores possíveis.

Multinomial: as variáveis independentes são discretas, ou melhor, é contagem do número de vezes que o desfecho x_i é observado durante as n tentativas. O classificador multinomial Naïve Bayes tem muitas aplicações, entre elas a classificação de informação textual. Por exemplo: documentos legais, documentos de políticas públicas, estratégias de mercado, avaliação de SPAMs entre outros.

CLASSIFICADOR NAÏVE BAYES

Aplicações do Algoritmo Naïve Bayes:

- Em aplicações na análise de crédito, diagnósticos médicos, busca por falhas em sistemas mecânicos, detecção de SPAMs entre outras aplicações.
- Aplicações na saúde, como sistemas que determinam se alguém tem uma doença ou não.
- Como classificador de SPAM ele analisa e-mails e tenta avaliar se ele é spam ou não (classes definidas) com base em suas informações e estrutura.
- Classificadores de sentimento. Nesses casos, eles analisam os textos e tentam identificar a emoção expressada, geralmente entre opções específicas como “neutro”, “positivo” ou “negativo”. Nas redes sociais é utilizado para identificar se o usuário está feliz ou triste ao publicar determinado texto.
- Em classificação de textos, têm boas taxas de sucesso em comparação com outros algoritmos.

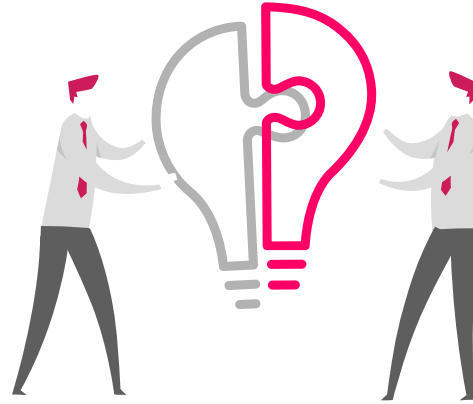
CLASSIFICADOR NAÏVE BAYES

Mais aplicações do Algoritmo Naïve Bayes:

- Sistema de Recomendação: utilizando filtragem colaborativa constroem um sistema de recomendação. Nesse caso, o objetivo é analisar certas pessoas e tentar sugerir algo que possa interessar a elas, seja conteúdo ou produtos.
- Usado para fazer previsões em tempo real: Por possuir uma velocidade relativamente alta e precisar apenas de poucos dados para realizar a classificação, o Naïve Bayes pode ser utilizado para previsões em tempo real.
- Pode ser utilizado para prever a probabilidade de múltiplas classes de variáveis.

EXERCITANDO

NAIVE BAYES



Adult



BIBLIOGRAFIA

- KUHN, M. / JOHNSON K. **Applied Predictive Modeling**, 1st ed. 2013, Corr. 2nd printing 2018 Edition
- LESKOVEC, RAJAMARAM, ULLMAN. **Mining of Massive Datasets**, 2014. <http://mmds.org>.
- HAIR, J.F. / ANDERSON, R.E. / TATHAN, R.L. / BLACK, W.C. **Análise multivariada de dados**, 2009
- TORGO, L. **Data Mining with R: Learning with Case Studies**, 2.a ed. Chapman and Hall/CRC , 2007
- MINGOTI, S.A.; **Análise de dados através de métodos de estatística multivariada**, UFMG, 2005
- CARVALHO, L.A.V., **Datamining – A mineração de dados no marketing, medicina, economia, engenharia e administração**. Rio de Janeiro: Editora Ciência Moderna, 2005.
- BERRY, M.J.A., LINOFF, G. **Data Mining Techniques For Marketing, Sales and Customer Support**. 3a. ed. New York: John Wiley & Sons, Inc., 2011.
- DUNHAM, M.H. **Data Mining - Introductory and Advanced Topics**. Prentice Hall, 2002.
- DINIZ, C.A.R. , NETO F.L. **Data Mining: Uma Introdução**. São Paulo: XIV Simpósio Nacional de Probabilidade e Estatística. IME-USP, 2000.

OBRIGADA!



/AdelaideAlves



profadelaide.alves@fiap.com.br

FIAP

Copyright © 2022 | Professor (a) Adelaide Alves de Oliveira

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente proibido sem consentimento formal, por escrito, do professor/autor.

FIAP