

MBA Business Analytics e Big Data

Análise Preditiva

Prof. Dr. João Rafael Dias

1º semestre - 2020

Aprendizagem supervisionada
Regressão e classificação
Formas de treino e validação
Bias-variance trade-off
Avaliação e comparação de modelos
Prática no RStudio

Estrutura de uma árvore de decisão
Intuição
Particionamento dos nós na regressão
e classificação
Poda da árvores vs *overfitting*

Introdução e motivações
Feature engineering
Tratamento de variáveis
Transformação de variáveis
Arquivos de trabalho
Prática no RStudio

Regressão linear múltipla
Coeficiente de determinação
Regressão logística
Odds e log odds
Comparação entre as regressões
Multicolinearidade
Seleção de variáveis *step-wise*
Prática no RStudio

Modelos de *ensemble*
Bootstrap
Random forest
Adaptive boosting
Prática no RStudio

OBJETIVOS E FOCO

- Capacitar os alunos a construir modelos preditivos e, avaliar e comparar diferentes tipos de algoritmos usados em modelagem
- Aplicações em problemas de regressão e classificação binária
- Foco em aprendizagem supervisionada
- Primeiro contato com *machine learning*
- *Problem-based approach*

PROGRAMA

- Princípios de modelagem preditiva
- *Feature engineering* nas variáveis
- Introdução às técnicas supervisionadas
- Avaliação e comparação de modelos preditivos
- Regressão linear múltipla
- Regressão logística
- Árvores de decisão
- *Random Forest*
- *Boosting (adaboost)*

AVALIAÇÃO

- Trabalho individual (peso **70%**): case aplicado
- Trabalho em grupo (peso **30%**): projeto
- Datas de entrega: a definir

METODOLOGIA

- Aulas expositivas, exercícios práticos e discussão de cases
- Ferramentas: RStudio e MS Excel

MATERIAIS DE REFERÊNCIA

- GARETH, J., T. HASTIE e R. TIBSHIRANI. **An Introduction to Statistical Learning: With Applications in R**. Springer, 2014.
- KUHN, M e K. JOHNSON. **Applied Predictive Modeling**. Springer, 2013.
- SIEGEL, E.. **Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie or Die**. Willey, 2017.
- TOWARDS DATA SCIENCE: <https://towardsdatascience.com/>
- MACHINE LEARNING MASTERY: <https://machinelearningmastery.com/start-here/>
- ANALYTICS VIDHYA: <https://www.analyticsvidhya.com/>
- SICSÚ, A. L.. **Credit Scoring: Desenvolvimento, Implantação e Acompanhamento**. Blucher, 2010;

Getting Started



George Box

(1919 – 2013)

“Essentially, all models are wrong,
but **some are useful**”

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?"

Ciclo de vida em modelagem preditiva

Para começar...

- Modelos são uma simplificação ou aproximação da realidade; podem ser tratados como idealizações de diferentes tipos de processos e sistemas dinâmicos

Realidade



Modelagem

Princípios de modelagem preditiva

Princípios de modelagem preditiva

Em quais contexto usamos?

- Modelagem preditiva é uma ferramenta, baseada em dados, que nos ajuda a responder problemas de negócio do dia-a-dia auxiliando na tomada de decisão
- Ela sempre tem associada uma pergunta, e para cada uma existe um tipo específico de ferramenta que abrange conjuntos de técnicas

O que você quer fazer?



Previsão entre categorias

Isto é A ou B? Isto é A ou B ou C ou D?

Previsão de valores

Qual é o valor? Quanto é?

Classificar imagens

O que essa imagem representa?

Extrair informação de texto

Que informações existem nesse texto?

Descobrir estruturas

Como isso está organizado?

Gerar recomendações

No que eles estarão interessados?

Encontrar ocorrências não usuais

Isso é atípico ou estranho?

Nosso foco

- Hoje é muito difícil pensar em onde o conceito de *predictive analytics* não é utilizado...



Bancos

- Detecção de clientes arriscados
- Recuperação de dívidas
- Oferta de produtos de créditos
- Precificação
- Fraudes



Marketing

- Propensão à compras
- Propensão à contratação de serviços
- Análise de sentimento
- Segmentação de perfis



Seguros

- Acionamento do seguro
- Precificação
- Acidentes e sinistros
- Óbitos
- Perfis de público
- Fraudes

- Hoje é muito difícil pensar em onde o conceito de *predictive analytics* não é utilizado...



Saúde

- Propensão à óbito
- Comorbidades
- Desenvolvimento de doenças
- Detecção de câncer
- Re-entrada em hospital
- No-show



RH

- Seleção de currículos
- Propensão de atrito e demissão
- Performance
- Absenteísmo
- Processos trabalhistas

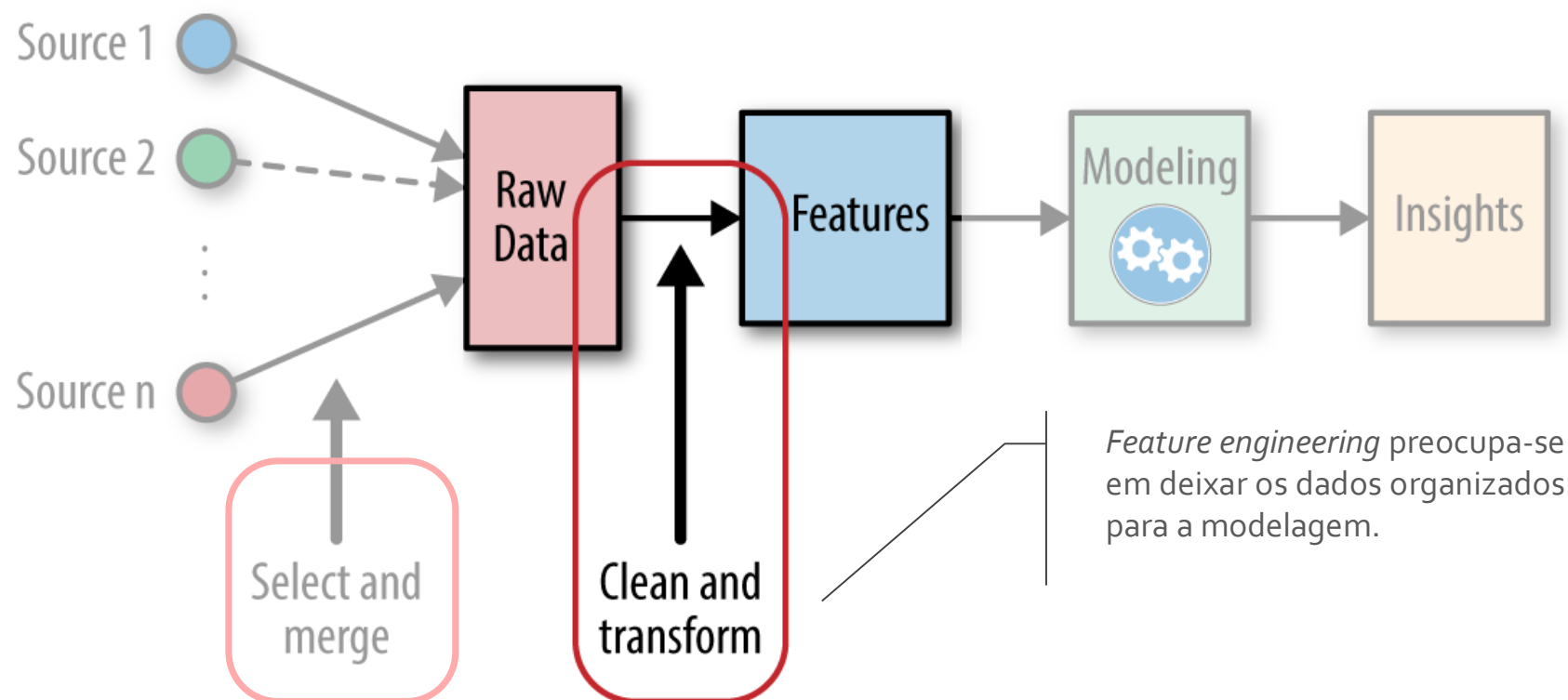


Telecons

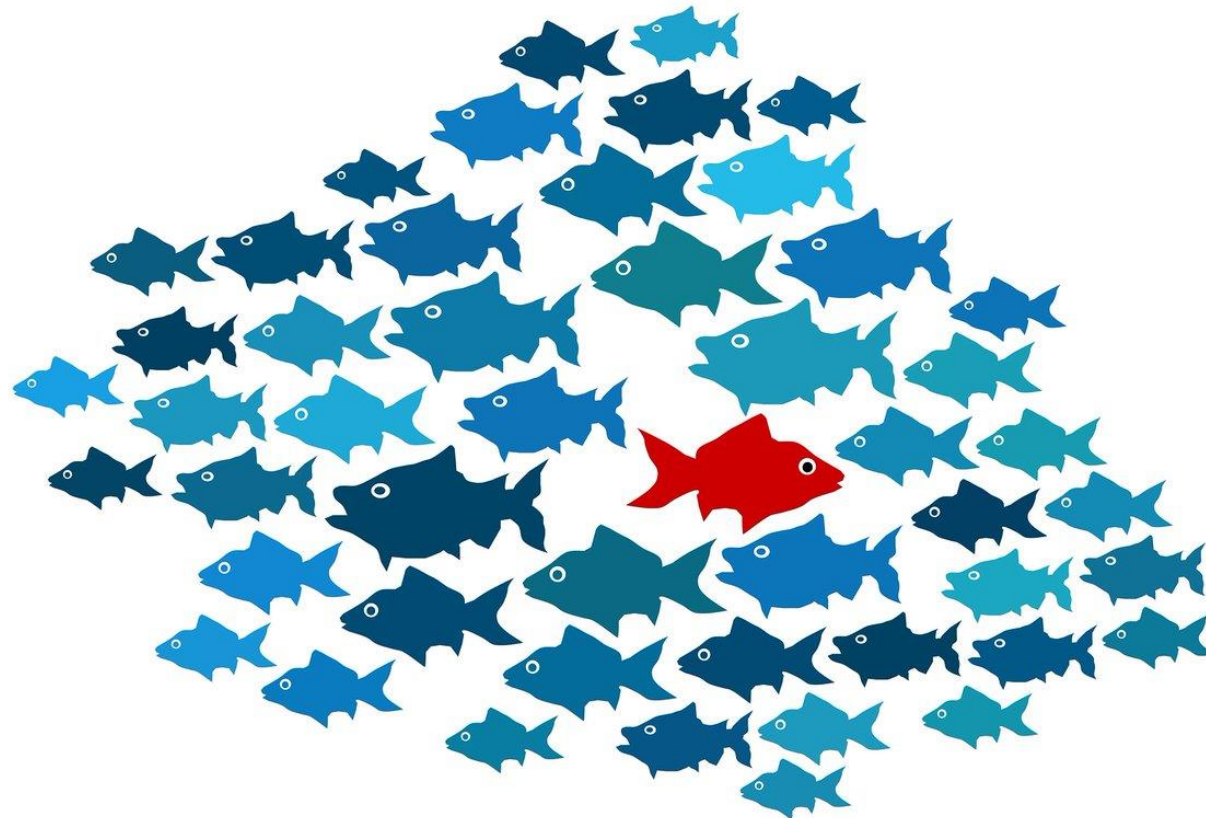
- *Churning* de contratos
- Adesão de serviços
- Análise de sentimento
- Detecção de falhas de cobertura
- Precificação

Feature engineering

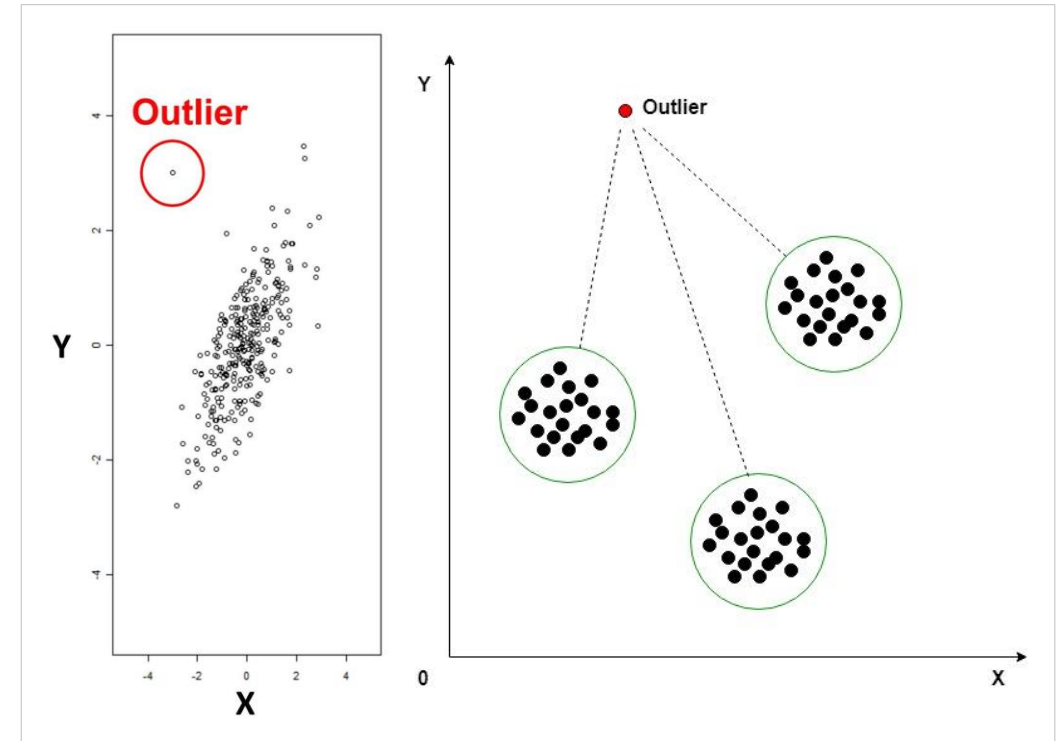
- No processo de *feature engineering* procuramos tratar e transformar as variáveis que irão ser usadas no treino dos algoritmos
- É nesse momento que são identificados e tratados dados aberrantes (*outliers*) e dados em brancos ou omissos (*missing values*)
- Conseguimos criar variáveis, fazer transformações ou até mesmo reduzir a dimensionalidade do *dataframe*



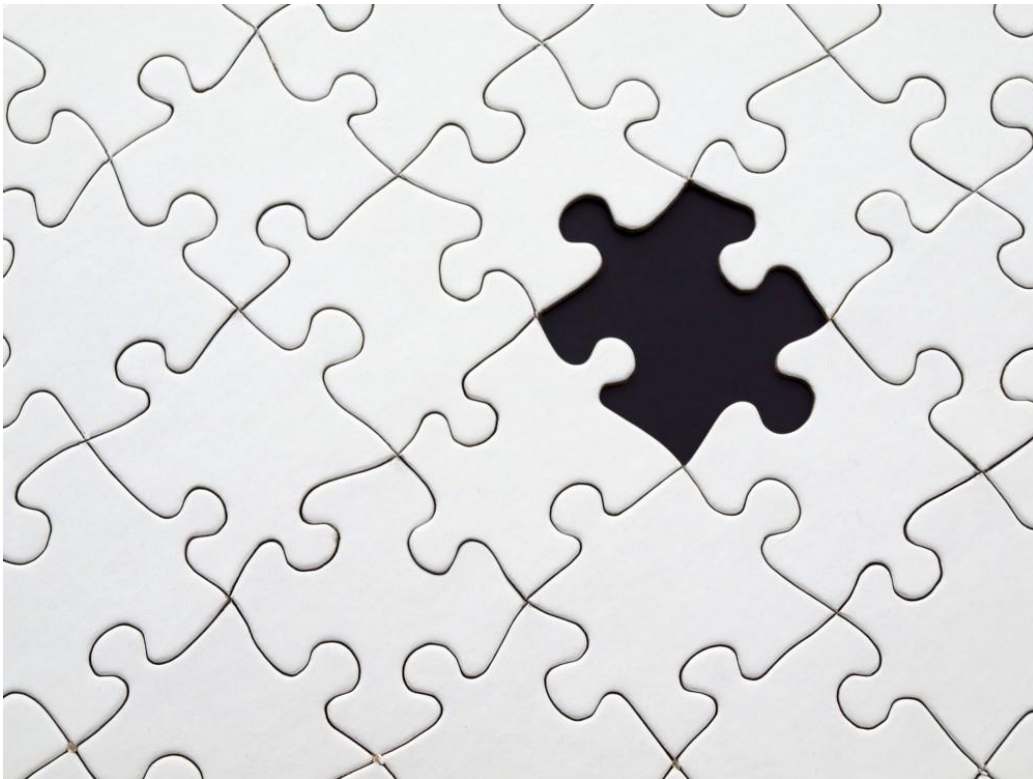
- São dados que encontram-se distantes dos demais pontos da amostra. Não significa que sejam errados ou falsos, apenas que são discrepantes
- Eles desviam dos padrões gerais da variável na amostra



Exemplo



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



Exemplo

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

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

- São variáveis binárias ou indicadoras muito usadas em regressão linear múltipla, mas também são usadas em outros tipos de algoritmos.
- Com elas são criadas variáveis independentes a partir de uma variável categórica, indicando a presença ou não presença de uma determinada categoria (são excludentes)
- R consegue tratar variáveis categóricas transformando automaticamente em *dummies*, já o Python precisa tratar antes.

DUMMY VARIABLES

With five categories (c), you can create four (c – 1) **dummy variables**.
The omitted category (here **WHITE**) is known as the “**referent**” category.
Each of the four dummy categories is compared to the referent.

Which categories were labeled as 1, 2, 3, 4, and 5 is totally arbitrary. Numerical scores have no meaning in this context (no reason to think of a 5 as being “higher” on some property than a 4, or a 4 as higher than a 3, etc.).

Original Variable Race-Ethnicity	Black	Hispanic	Asian	Other
White/European-American (1)	0	0	0	0
Black/African-American (2)	1	0	0	0
Hispanic/Latino (3)	0	1	0	0
Asian-American (4)	0	0	1	0
Other (5)	0	0	0	1

Trabalhamos sempre com $k-1$ *dummies*, onde k é o número de categorias da variável qualitativa

Correção de assimetria

- Correção / redução de assimetria
 - Há testes de hipótese que assumem normalidade (simetria)
 - importante para PCA (assimetria influencia nas correlações)
 - Diminui o efeito de *outliers* devido a normalização das magnitudes

- Algumas recomendações:

Assimetria positiva moderada: \sqrt{x}

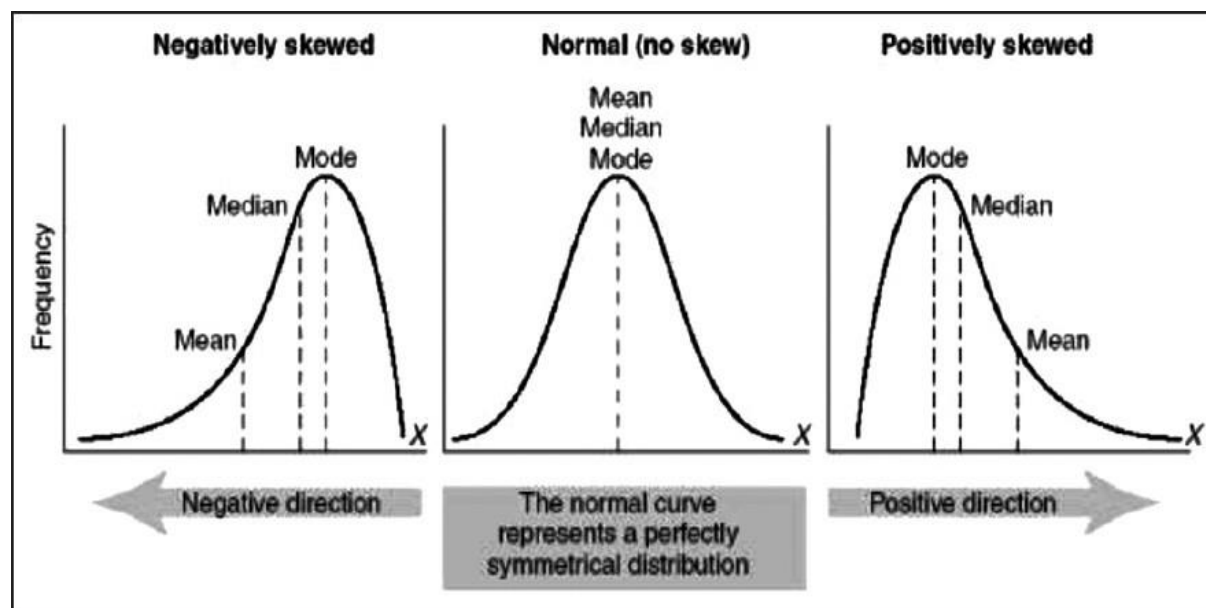
Assimetria positiva pronunciada: $\log_{10} x$

Assimetria positiva pronunciada (com zero): $\log_{10}(x + C)$

Assimetria negativa moderada: $\sqrt{(K - x)}$

Assimetria negativa pronunciada: $\log_{10}(K - x)$

onde C e K são constantes



<https://www.fromthegenesis.com/skewness/>

Introdução à aprendizagem supervisionada

Introdução à aprendizagem supervisionada

Overview

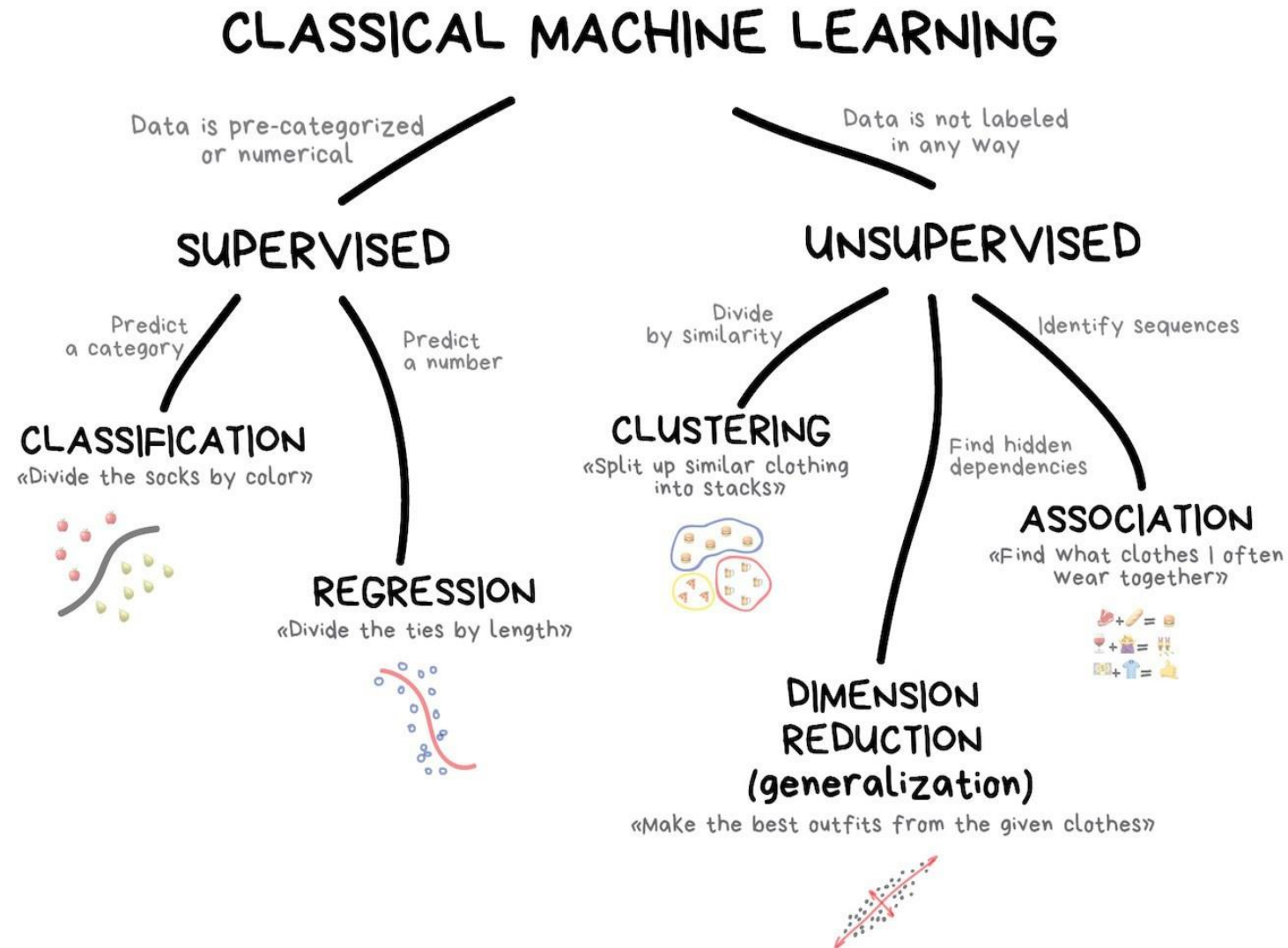
- Aprendizagem supervisionada é um tipo de técnica em *machine learning* na qual são inseridos além dos dados uma variável auxiliar que representa a resposta de interesse (*target*)
- A partir dos dados que são “rotulados” o algoritmo aprende padrões e conseguimos usá-lo para fazer previsões
- Eles podem ser divididos em dois grupos de tarefas distintas (*tasks*)

Regressão

Quando a variável resposta é um **número**

Classificação

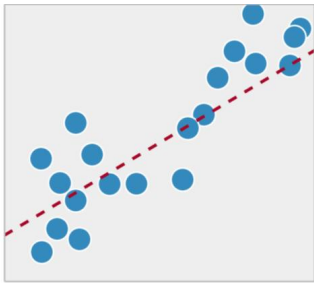
Quando a variável resposta é uma **categoria** (ou várias)



Introdução à aprendizagem supervisionada

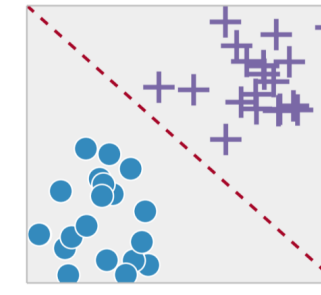
Regressão vs Classificação

- Exemplo de problemas



Regressão

- Preço de imóvel
- Salários
- Vendas
- Demandas
- Previsão de preço de ação
- Temperatura e chuva
- Diagnósticos médicos



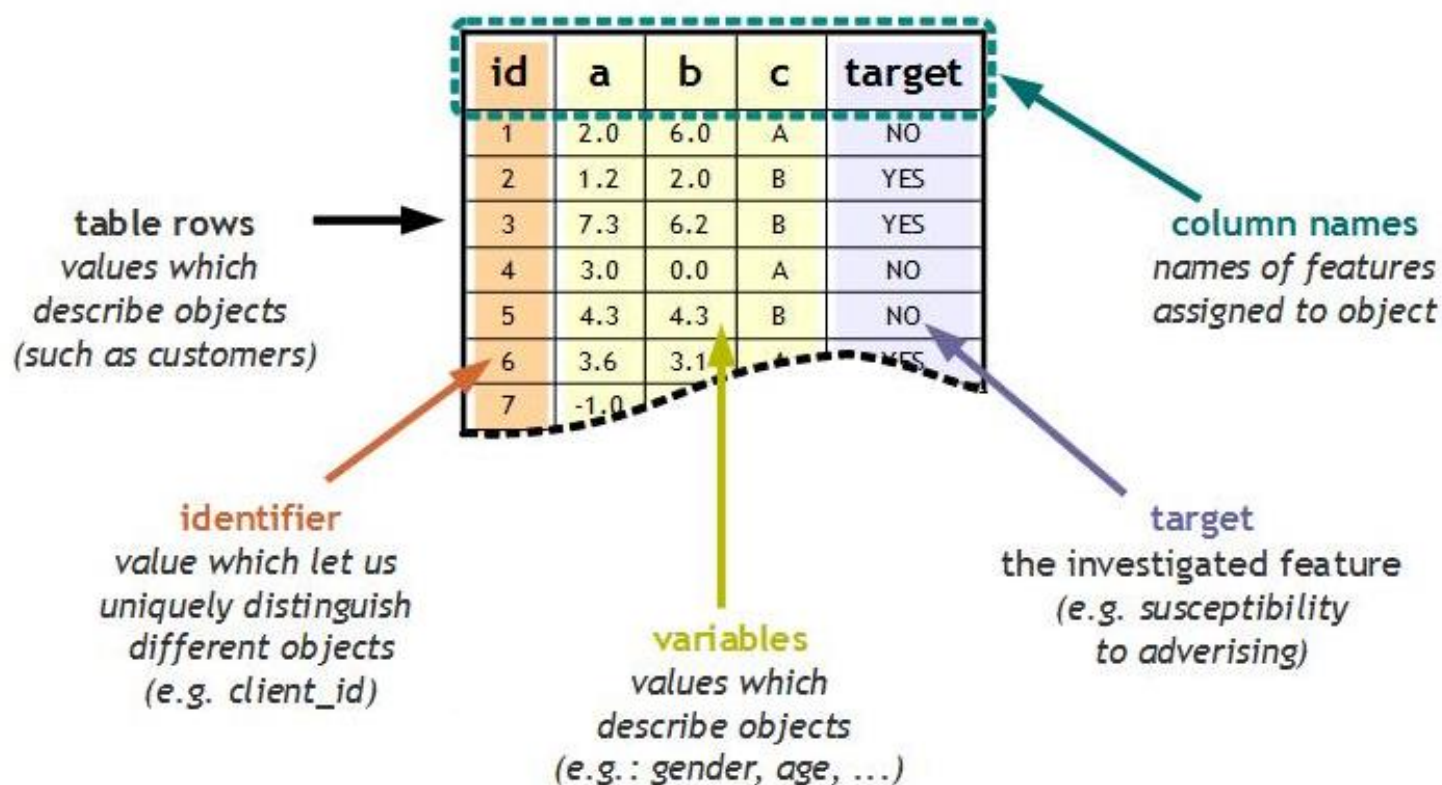
Classificação

- | | |
|-----------------------------|---------------------------------|
| • Risco de crédito | • Análise de sentimentos |
| • Cancelamento de contratos | • Detecção de idioma |
| • Atrito com a empresa | • Classificação de imagens |
| • Fraudes | • Intensidade de algum fenômeno |
| • Filtros de previsão | |
| • Propensão | |
| • Diagnósticos médicos | |

Fonte de dados

- Independente de ser um problema de regressão ou classificação utilizamos um conjunto de dados para o treino do algoritmo que possua um “supervisor” ou professor que dá as respostas para a máquina
- A base de dados possui o formato tabular onde temos os seguintes elementos

Exemplo para um problema de classificação...



The diagram shows a table with 7 rows and 5 columns. The columns are labeled 'id', 'a', 'b', 'c', and 'target'. The rows contain numerical and categorical data. Annotations with arrows point to specific parts of the table: 'table rows values which describe objects (such as customers)' points to the entire table; 'column names names of features assigned to object' points to the header row; 'identifier value which let us uniquely distinguish different objects (e.g. client_id)' points to the 'id' column; 'variables values which describe objects (e.g.: gender, age, ...)' points to columns 'a' and 'b'; and 'target the investigated feature (e.g. susceptibility to advertising)' points to the 'target' column.

id	a	b	c	target
1	2.0	6.0	A	NO
2	1.2	2.0	B	YES
3	7.3	6.2	B	YES
4	3.0	0.0	A	NO
5	4.3	4.3	B	NO
6	3.6	3.1	YES	
7	-1.0			

Papeis das variáveis

Variável independente

Variável que descreve as propriedades de um objeto as quais são aprendidas e relacionadas pelo algoritmo (chamadas de *feature*, atributos ou variável explicativa)

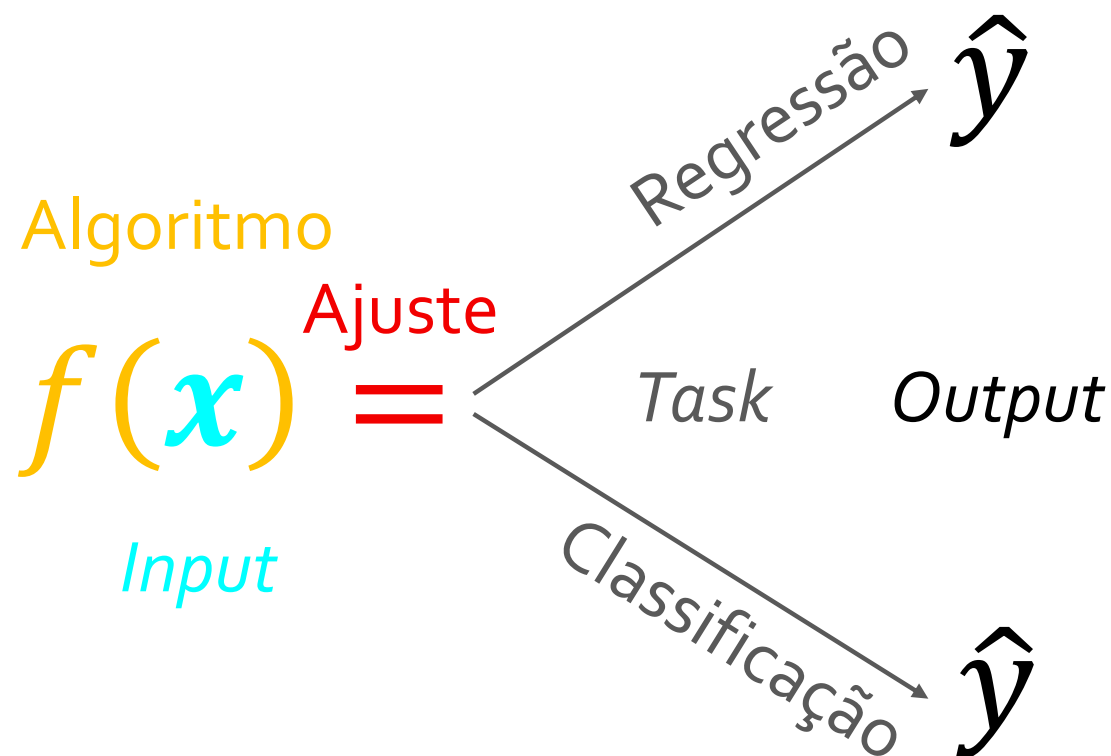
Variável dependente

Variável que descreve o valor ou característica que queremos que o algoritmo aprenda (chamadas de *target*, variável resposta)

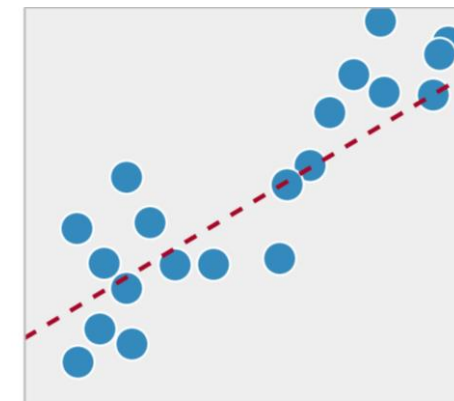
Introdução à aprendizagem supervisionada

Um pouco mais sobre a saída do modelo

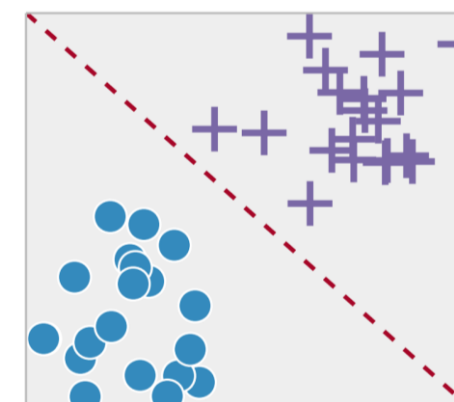
- A saída do modelo é tratada de forma diferente dependendo da tarefa ser de regressão ou classificação



Como saída retorna um **valor** que possui a **mesma unidade de medida** da variável resposta que foi utilizada para o treino do algoritmo. É confrontada diretamente com y



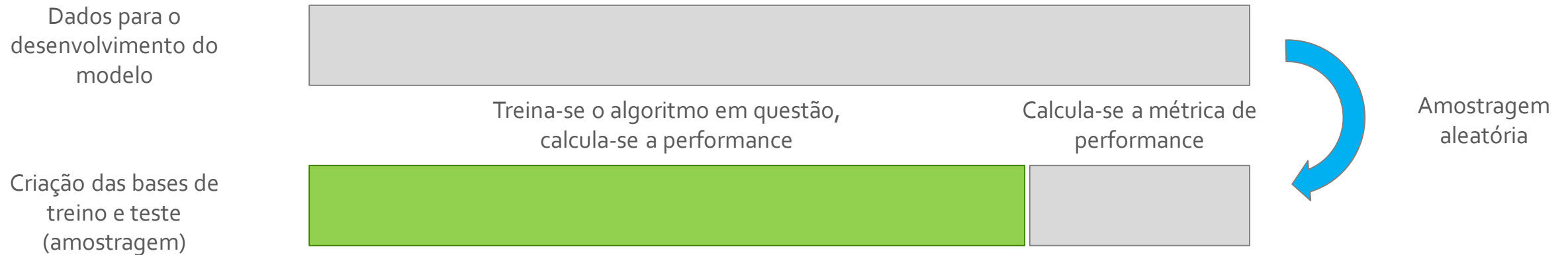
Como saída (maioria dos algoritmos) retorna um valor que pode ser interpretado como a **probabilidade** de se pertencer a uma categoria específica. A partir dela **constrói-se o label** específico



Introdução à aprendizagem supervisionada

Estratégia *hold-out*

- No processo de modelagem precisamos construir o conjunto de dados do público o qual deseja-se desenvolver o modelo.
- Num contexto de aprendizagem supervisionada, usualmente particionamos a amostra em duas partes mutuamente exclusivas (não necessariamente de tamanhos iguais):



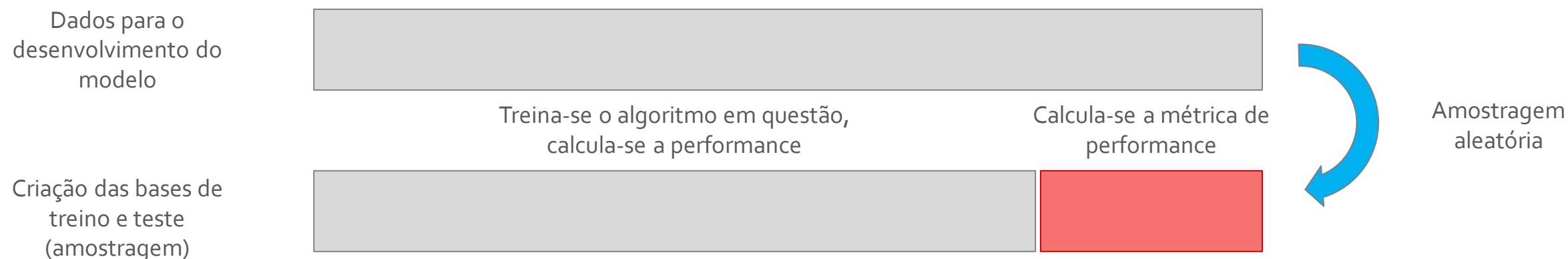
Amostra de treino

Aplicar o algoritmo e construir as regras de predição, sendo normalmente a maior parte da amostra. Essa parte dos dados é denominada de **amostra de desenvolvimento**, treino ou aprendizado (*learning* ou *training set*)

Introdução à aprendizagem supervisionada

Estratégia *hold-out*

- No processo de modelagem precisamos construir o conjunto de dados do público o qual deseja-se desenvolver o modelo.
- Num contexto de aprendizagem supervisionada, usualmente particionamos a amostra em duas partes mutuamente exclusivas (não necessariamente de tamanhos iguais):



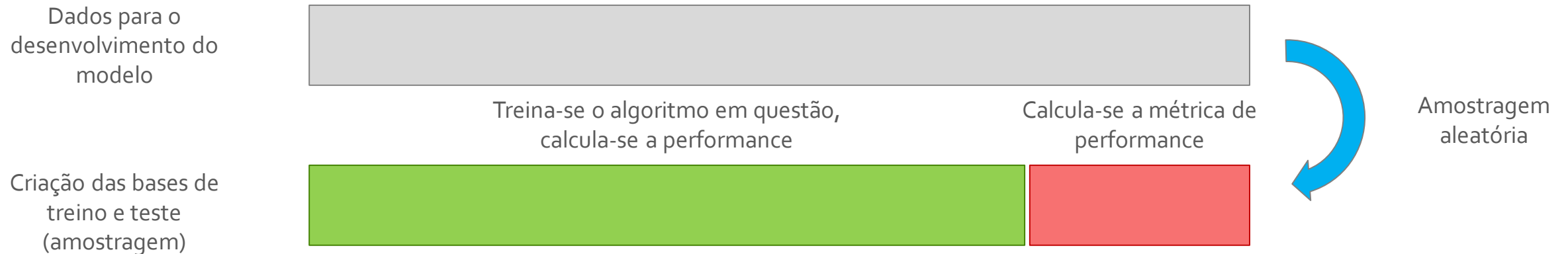
Amostra de teste

Amostra usada para estimar a capacidade de previsão do modelo e a qualidade do ajuste, denominada **amostra teste** (*test set*).

Introdução à aprendizagem supervisionada

Estratégia *hold-out*

- No processo de modelagem precisamos construir o conjunto de dados do público o qual deseja-se desenvolver o modelo.
- Num contexto de aprendizagem supervisionada, usualmente particionamos a amostra em duas partes mutuamente exclusivas (não necessariamente de tamanhos iguais):



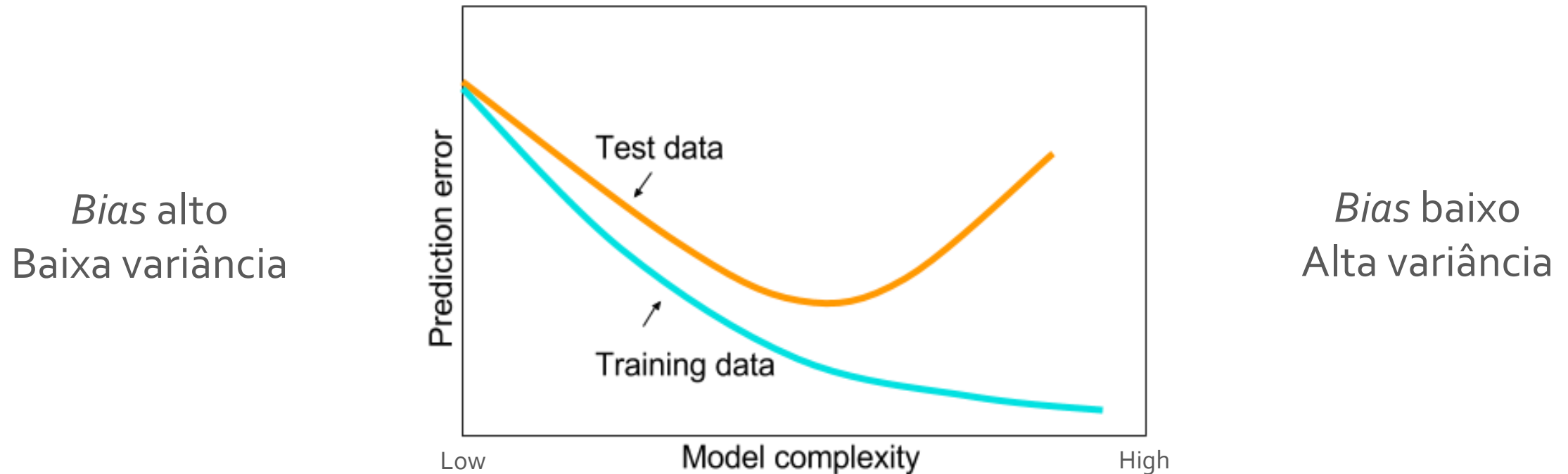
Intuição

Construir modelos fazendo-os aprender com os dados de uma amostra e em seguida testar seu grau de previsão com registros que não foram usados previamente no aprendizado dos algoritmos (poder de generalização)

Introdução à aprendizagem supervisionada

Poder de generalização e fontes de erro

- Como mencionado, particionamos a amostra em diferentes partes com o intuito de treinar o modelo, fazer previsões e validar essas previsões
- **Duas coisas podem acontecer:** nós super-ajustamos o modelo ou sub-ajustamos o modelo (*overfit* e *underfit*)
- Não queremos nenhuma das duas situações pois ambas levam a um modelo com **baixa acurácia** (grau de acerto) e **baixo poder de generalização** (não podemos generalizar as previsões para nenhum outro dado).

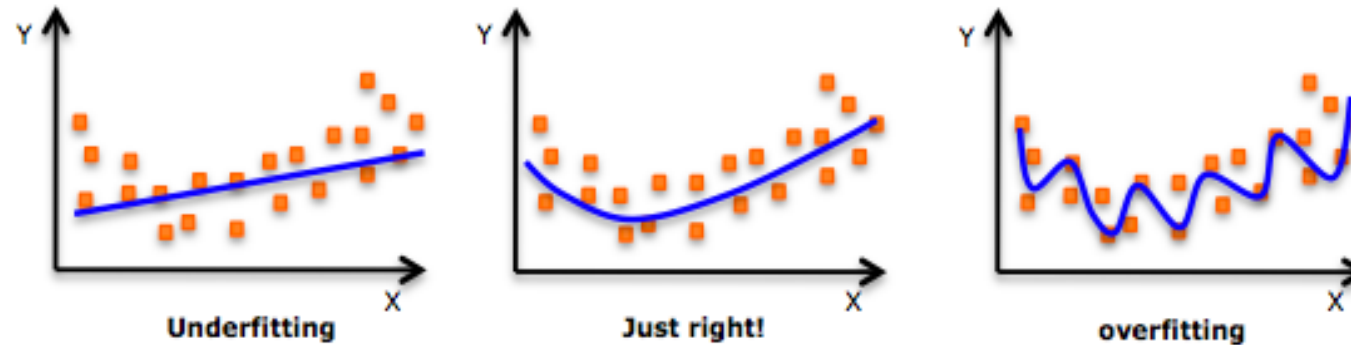


Introdução à aprendizagem supervisionada

Overfitting vs Underfitting

Exemplos

Regressão

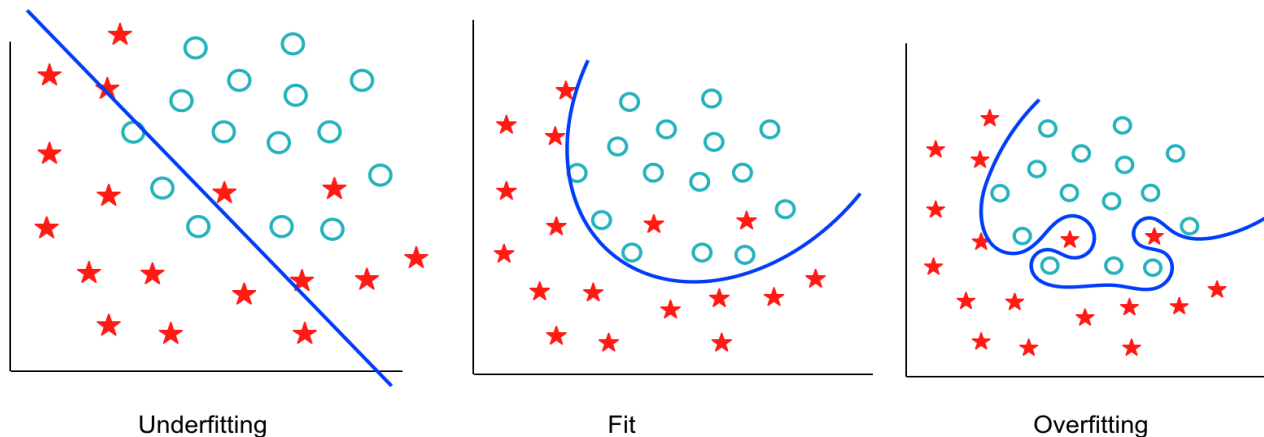


Previsor **pouco** flexível
Baixa complexidade
Alto viés
Baixa Variância

<https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76>

Previsor **muito** flexível
Alta complexidade
Baixo viés
Alta Variância

Classificação



<https://www.quora.com/Whats-the-difference-between-overfitting-and-underfitting>

Avaliação de modelos preditivos

Métricas de avaliação

- Uma das partes mais importantes em *machine learning* é aprofundar-se na avaliação dos modelos e nas métricas de performance. Mas escolher a métrica certa é crucial nessa etapa
- Aqui antes de colocar em produção é necessário avaliar o poder preditivo do ajuste encontrado. Existe poder de generalização?
- As diferentes tarefas de regressão e classificação pelas suas naturezas serem distintas em termos de resposta do modelo possuem métricas diferentes

Regressão

Mean Squared Error (MSE)
Root Mean Squared Error (RMSE)
Mean Absolute Error (MAE)
R2 ou coeficiente de determinação
R2 ajustado

Classificação

Confusion matrix
Taxa de erro e acurácia
Sensitividade/Especificidade
Precision/Recall/F1 score
AUC/ROC
Gini/KS/Log-loss



Existem muitas métricas, veremos apenas uma parte delas...

Mean Squared Error (MSE)

O erro médio quadrático é uma das métricas mais comuns em regressão. Ela é simplesmente a média das diferenças ao quadrado entre o valor *target* y e o valor previsto pelo modelo. Pela formulação, o MSE pode ser otimizado melhor do que outras métricas. Quanto menor melhor.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE)

Métrica similar ao MSE, mas aqui é aplicada a raiz quadrada sobre a média dos erros quadráticos. Ela é muito utilizada por possuir a mesma unidade da variável *target* y .

RMSE essencialmente mostra qual é o desvio médio dos valores previstos em relação ao *target*. Quanto menor melhor.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Mean Absolute Error (MAE)

É a média da diferença absoluta entre os valores previstos e os valores do *target*. O MAE é mais robusto à *outliers* e não penaliza os erros de forma tão extrema quanto o MSE.

Não é muito adequada para aplicações nas quais se quer prestar mais atenção aos erros extremos. Quanto menor melhor

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

R² e R² ajustado

Essas métricas indicam a proporção da variância (incerteza) na variável *y* que é estatisticamente explicada pela regressão. Ela pode ser usada como uma métrica da qualidade do ajuste linear. O R² ajustado leva em consideração o número de variáveis utilizadas na regressão. Quanto mais próxima de 1 melhor

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Confusion matrix (matriz classificação)

Obtendo a classificação de cada observação na amostra pelo modelo, pode-se confrontar com a classificação real de cada indivíduo em uma tabela cruzada.

Com isso é possível determinar quantas observações foram classificadas corretamente e quantas foram classificadas incorretamente.

A forma usada para estabelecer a matriz de classificação é determinar um ponto de corte c (cutoff) na probabilidade gerada pelo modelo, onde (para classificação binária)

$$\hat{y} = \begin{cases} \text{nao evento (0,-)} & \text{se prob} < c \\ \text{evento (1,+)} & \text{se prob} \geq c \end{cases}$$

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)



Confusion matrix (matriz classificação)

Taxa de erro

$$error = \frac{FP + FN}{(TP + TN + FP + FN)}$$

taxa de erro, contabilizando o total de casos incorretamente classificado

Acurácia

$$accuracy = 1 - error$$

taxa de acerto, contabilizando o total de casos corretamente classificados

True positive rate

$$TPR = \frac{TP}{(TP + FN)}$$

sensitividade, indica a proporção de casos positivos corretamente classificados (também denominado de *Recall*)

False positive rate

$$TNR = \frac{TN}{(TN + FP)}$$

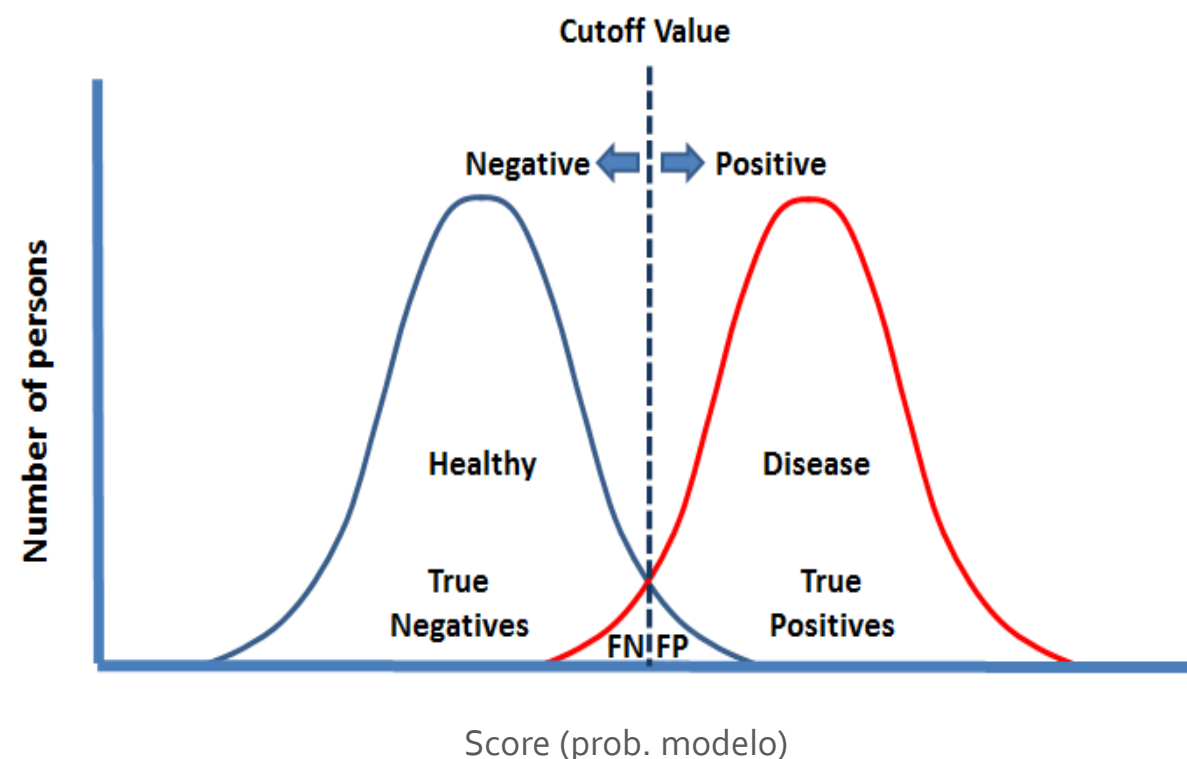
especificidade, indica a proporção de casos negativos corretamente classificados.

Curva ROC e AUC

O ROC (Receiver Operating Characteristics) é muito utilizada para comparar diferentes modelos. A curva ROC foi desenvolvida pela primeira vez por engenheiros elétricos e engenheiros de radar durante a Segunda Guerra Mundial para detectar objetos inimigos em campos de batalha. Ela é baseada em:

Sensitividade: capacidade de classificar corretamente o rótulo de classe alvo (evento positivo) - TPR

Especificidade: capacidade de classificar corretamente o rótulo de classe não alvo (evento negativo) – TNR

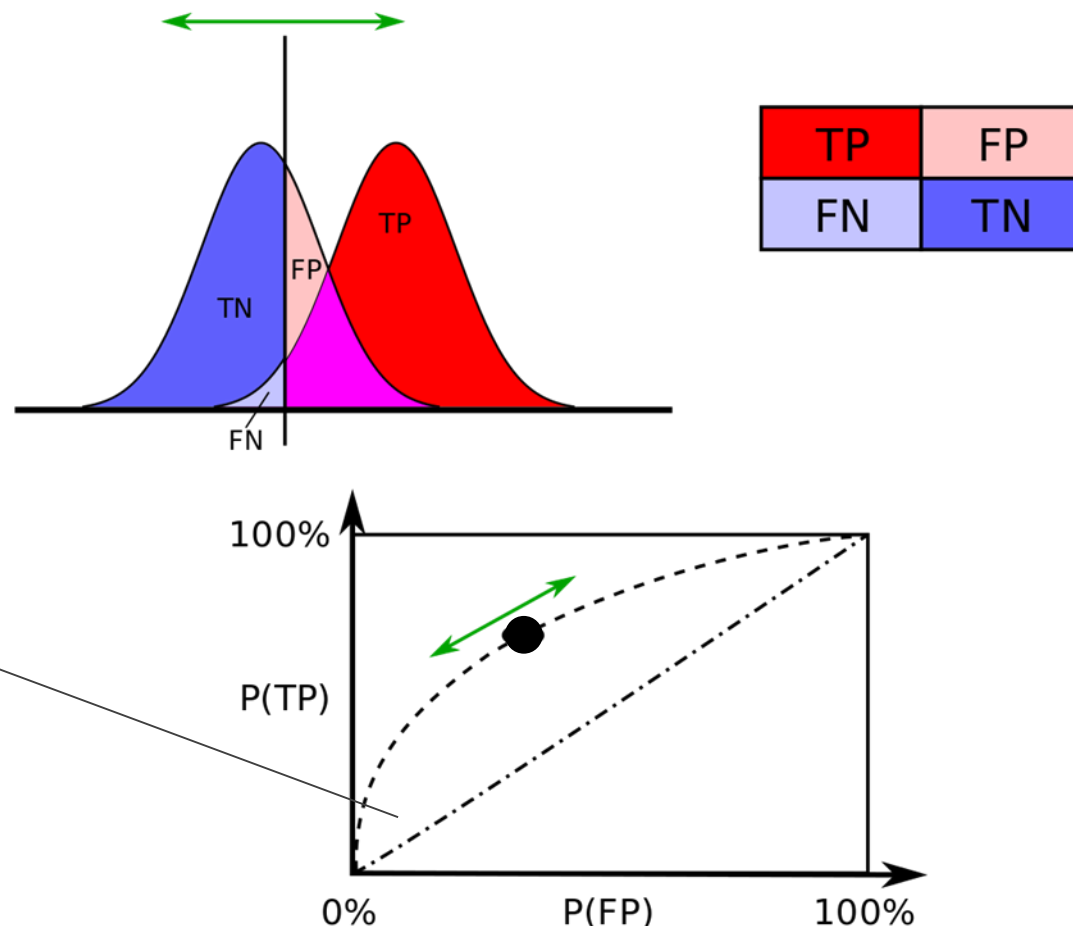


Curva ROC e AUC

A curva ROC é construída variando o ponto de corte ao longo do score do modelo afim de se obter as diferentes classificações e consequentemente o TPR e TNR.

Para cada valor de cutoff os valores das métricas TPR e TNR são diretamente calculados e plotados num gráfico (aqui TNR é plotado como $1 - \text{TNR} = \text{Falsos alarmes positivos}$)

A quantidade expressa pela área sob a curva é o AUC (ou AUROC). Também pode ser pensada como a probabilidade do modelo ranquear um observação positiva do que uma negativa. Quanto maior melhor



https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Curva ROC e a separação entre dois grupos

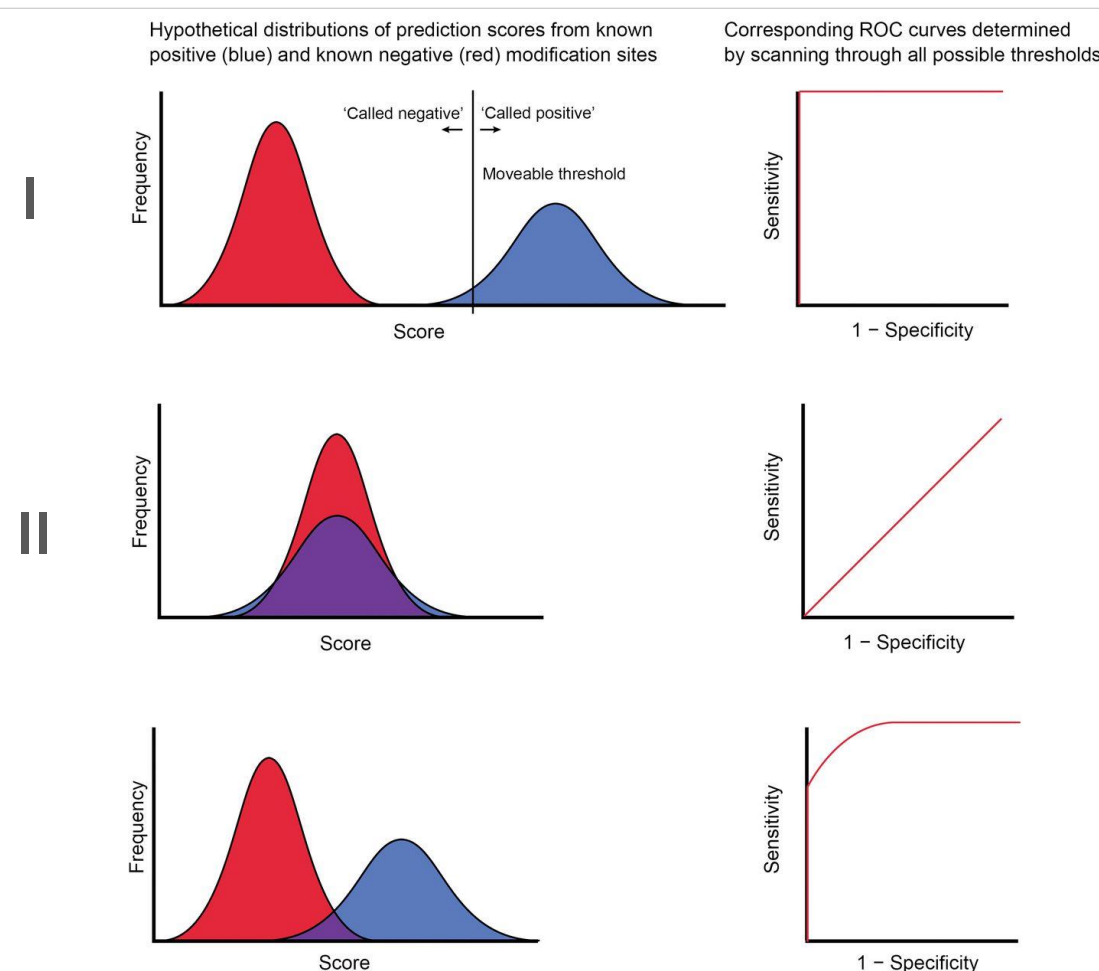
De forma geral, quanto mais afastada for a curva da diagonal melhor é a qualidade de discriminação do modelo.

Na figura ao lado:

I
Modelo perfeito:

II
Modelo inútil: (similar a à classificação randômica)

Caso dois modelos sejam comparados, o melhor é o que possui a curva mais afastada da diagonal



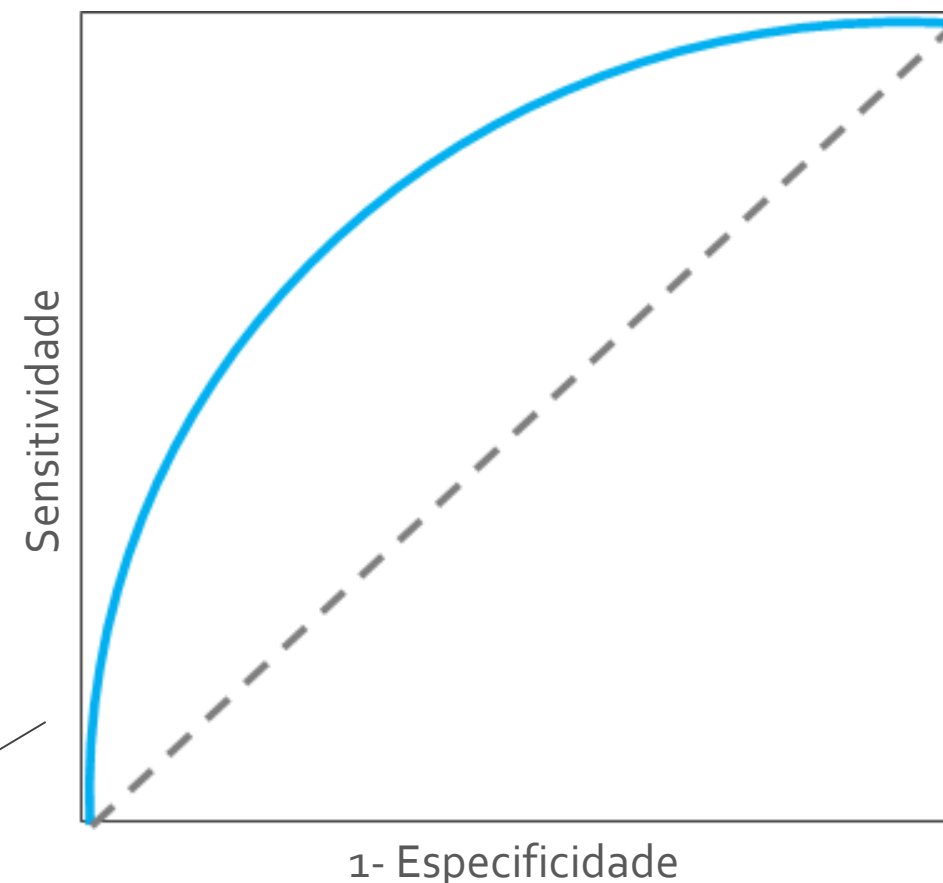
Coeficiente de Gini

O conceito foi criado por um estatístico italiano (Corrado Gini, 1912) originalmente para medir desigualdade de renda numa nação. Em modelagem preditiva, o coeficiente de Gini possui conceito similar à curva ROC, e é definido usando a área sobre a curva ROC (AUC) e a da classificação randômica.

É dado por

$$Gini = 2 \times AUC - 1$$

A área entre a linha azul e pontilhada é o Gini



Arquivos de trabalho

- **Task:** Regressão
- Representa um problema de marketing para previsão de vendas com dados de clientes
- 1.000 observações e 8 variáveis

	AGE	GENDER	MARITAL	INCOME	CHILDREN	HISTORY	CATALOGS	AMOUNT
1	Old	Female	Single	47500	0	High	6	755
2	Middle	Male	Single	63600	0	High	6	1318
3	Young	Female	Single	13500	0	Low	18	296
4	Middle	Male	Married	85600	1	High	18	2436
5	Middle	Female	Single	68400	0	High	12	1304
6	Young	Male	Married	30400	0	Low	6	495
7	Middle	Female	Single	48100	0	Medium	12	782
8	Middle	Male	Single	68400	0	High	18	1155
9	Middle	Female	Married	51900	3	Low	6	158
10	Old	Male	Married	80700	0	None	18	3034
11	Young	Male	Married	43700	1	None	12	927
12	Middle	Male	Married	111800	3	High	18	2065
13	Middle	Female	Married	44100	1	Medium	24	704
14	Middle	Male	Married	111400	0	High	12	2136
15	Old	Female	Married	110000	0	High	24	5564

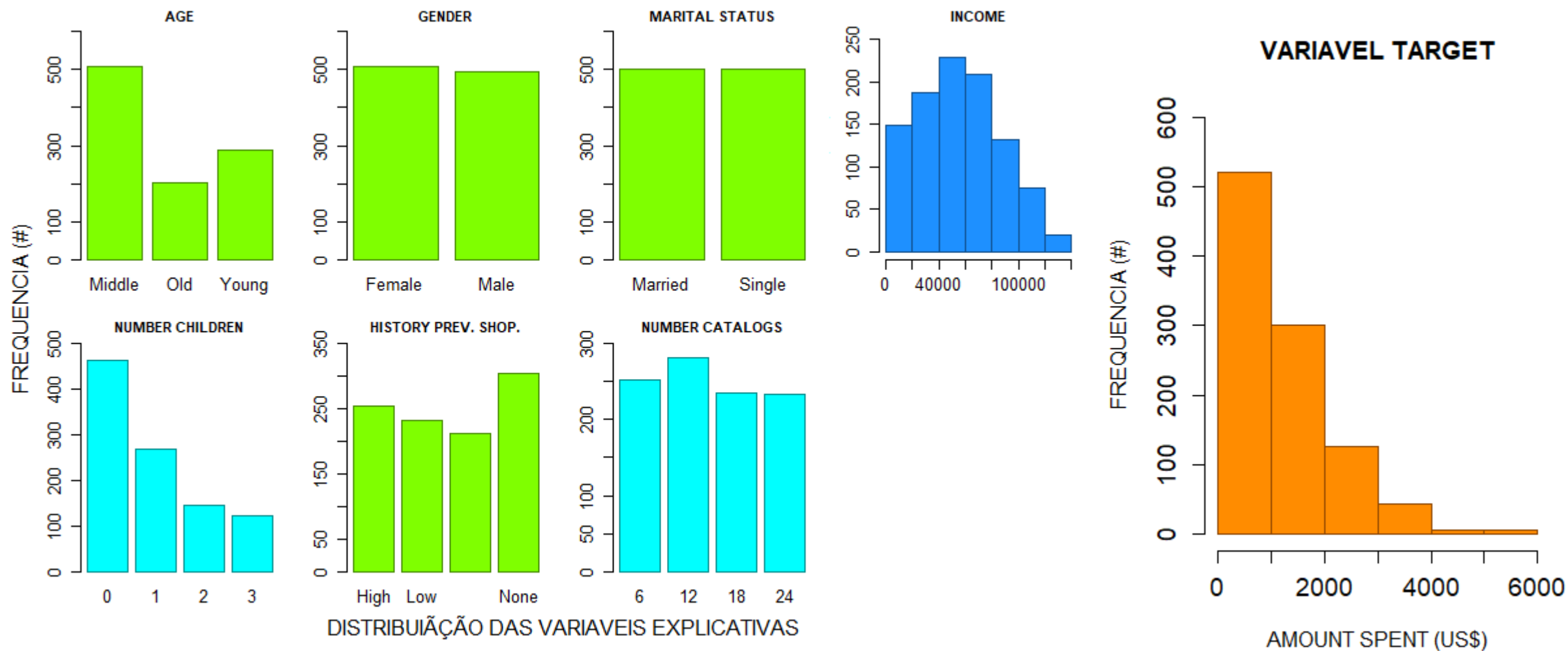
Metadados

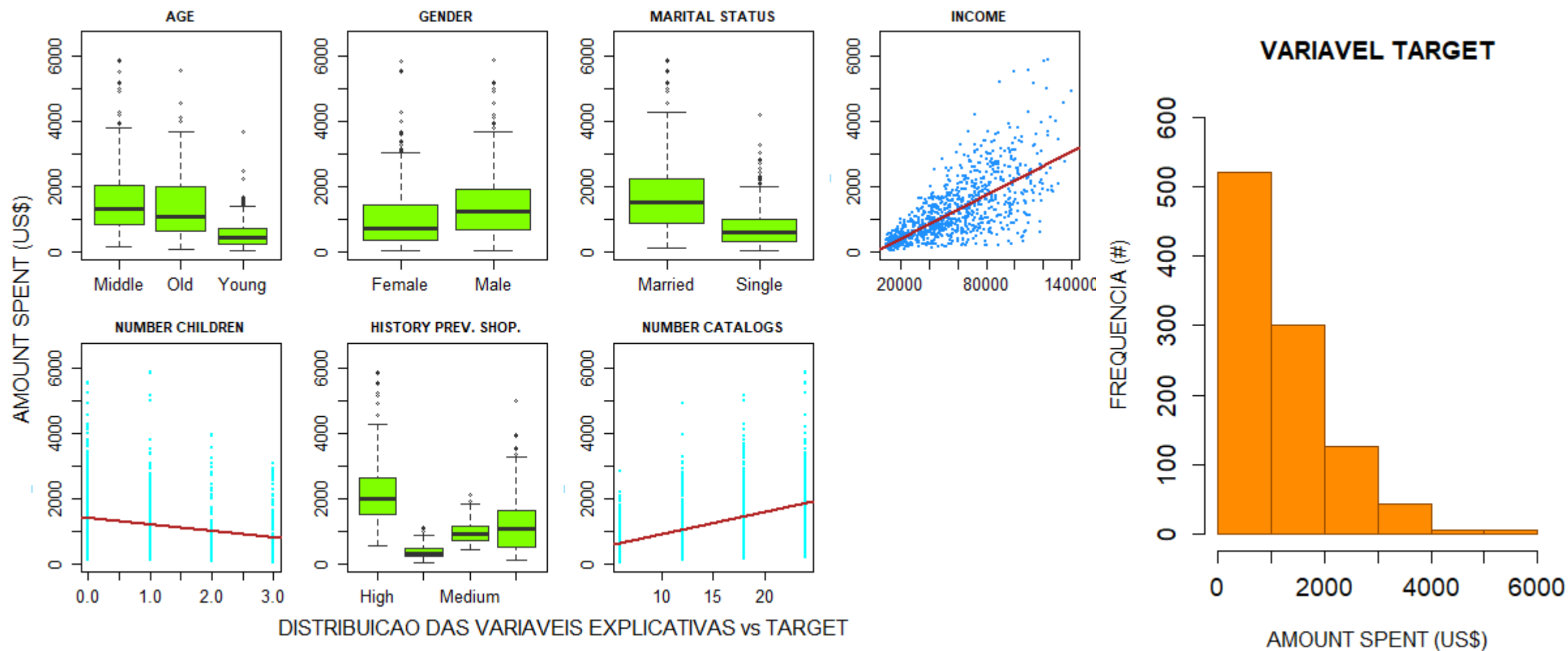
AGE: Faixa de idade
GENDER: Sexo
MARITAL: Tipo de estado civil
INCOME: Renda anual do cliente
CHILDREN: Número de filhos
HISTORY: Tipo de histórico de compras do cliente
CATALOS: Número de catálogos enviados
AMOUNT: Valor total gasto pelo cliente



Arquivos de trabalho

Análise univariada I





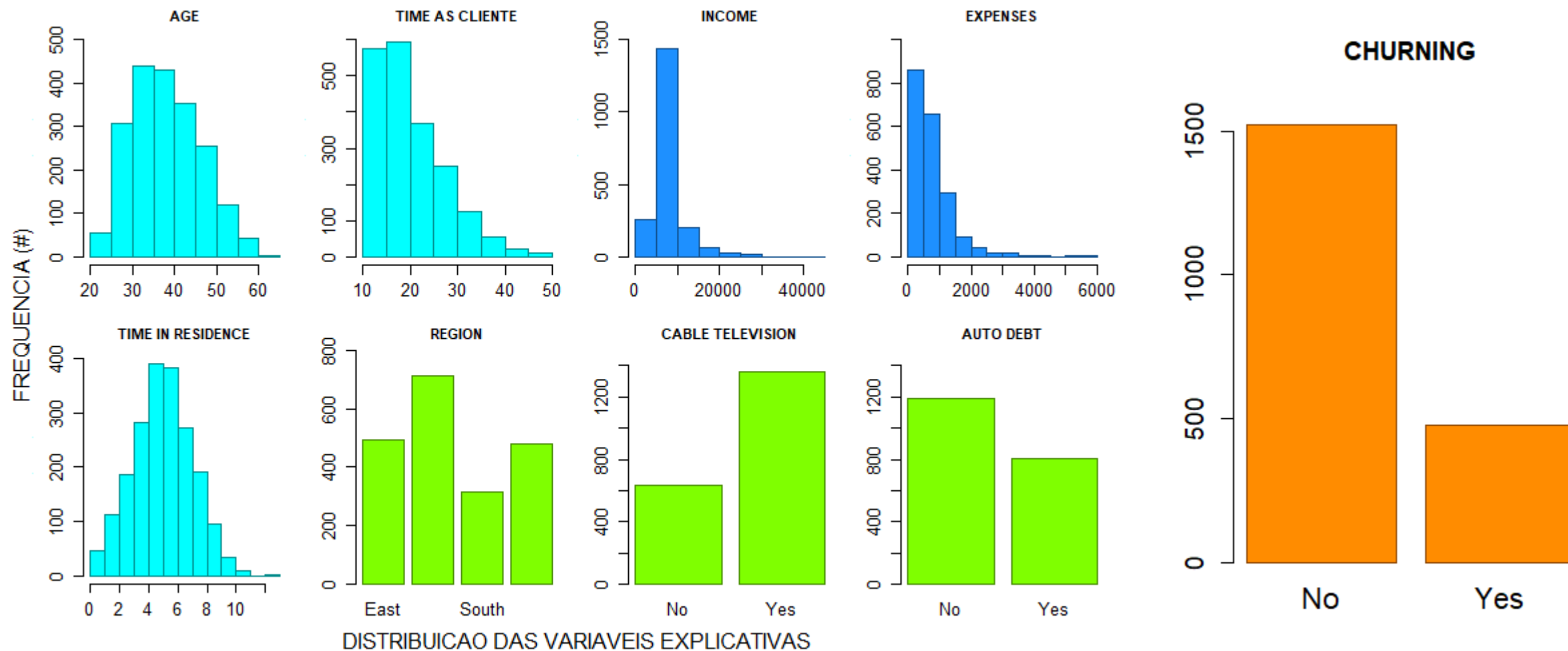
- **Task:** Classificação
- Representa um problema de previsão de churning de clientes em uma telecomm
- 2.000 observações e 9 variáveis

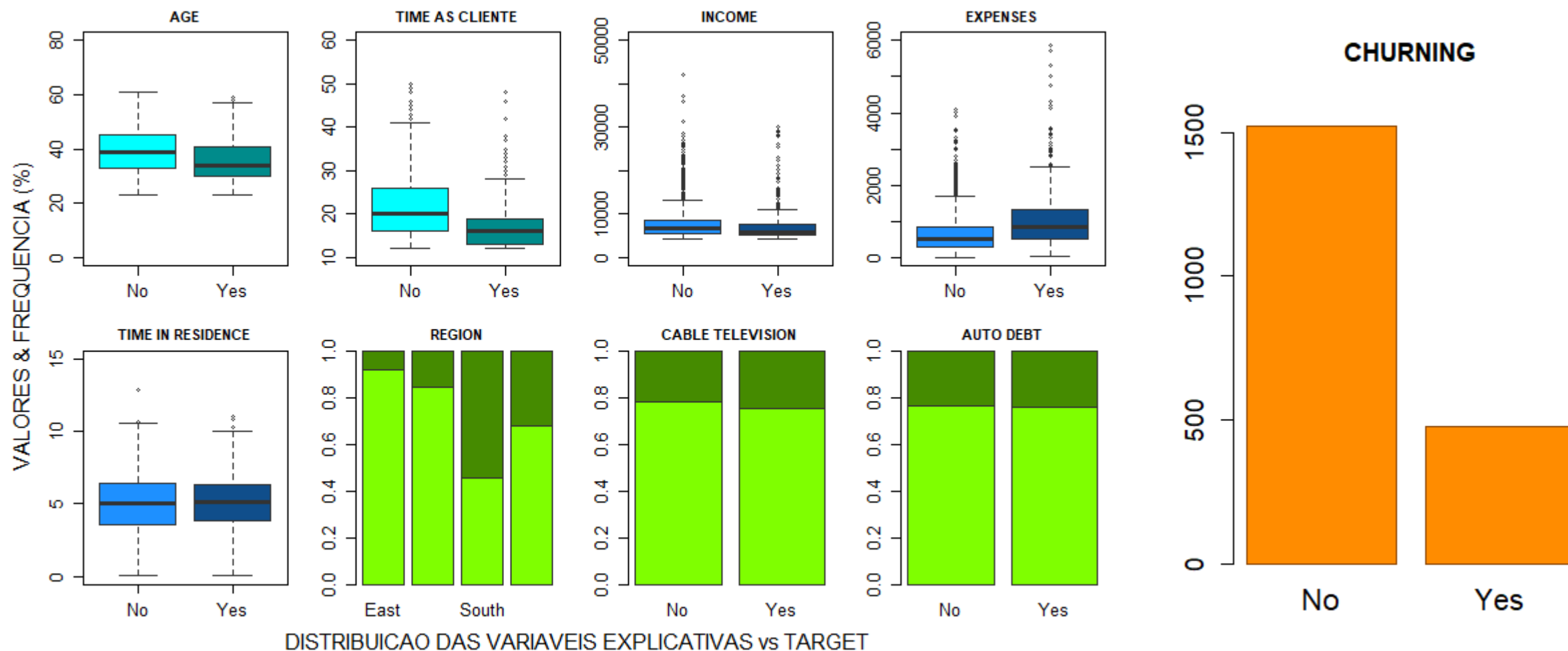
	AGE	CLITIME	INCOME	EXPENSES	RESIDTIME	REGION	CABLETV	AUTODEBT	CHURN
1	51	26	5320	543	7.3	North	Yes	No	No
2	36	16	5620	482	4.5	North	Yes	No	No
3	35	15	4860	593	4.8	North	No	No	No
4	40	22	6590	1184	6.2	East	Yes	No	No
5	52	30	6370	634	2.2	North	No	No	No
6	38	16	6120	146	5.6	West	Yes	No	No
7	27	18	5600	1273	4.8	West	No	Yes	No
8	45	29	10080	717	4.1	North	Yes	No	No
9	35	12	4720	446	6.9	North	No	No	No
10	30	21	5840	184	7.2	North	Yes	Yes	No
11	44	34	13550	1367	4.9	North	Yes	Yes	No
12	30	23	5750	856	8.1	South	No	No	No
13	39	20	6870	592	6.1	West	No	Yes	No
14	39	21	6880	593	2.8	North	No	No	No
15	45	18	7160	285	8.0	North	Yes	Yes	No

Metadados

AGE: Idade
CLITIME: Tempo como cliente
INCOME: Renda mensal
EXPENSES: Total de despesas mensais
REGION: Região onde mora
CABLETV: Indicador se possui TV à cabo
AUTODEBT: Indicador se possui débito automat.
CHURN: Indicador de *churning*







Prática no RStudio

...foco de hoje

- ***Feature engineering* nas bases de dados**

Preparação (simplificada) das bases de dados para o treino de algoritmos focando em alguns elementos discutidos em aula

- **Avaliando o desempenho de modelos**

Avaliando e comparando a capacidade preditiva de modelos de regressão e classificação para pacientes com diabetes

Análises dos *outputs* dos modelos

