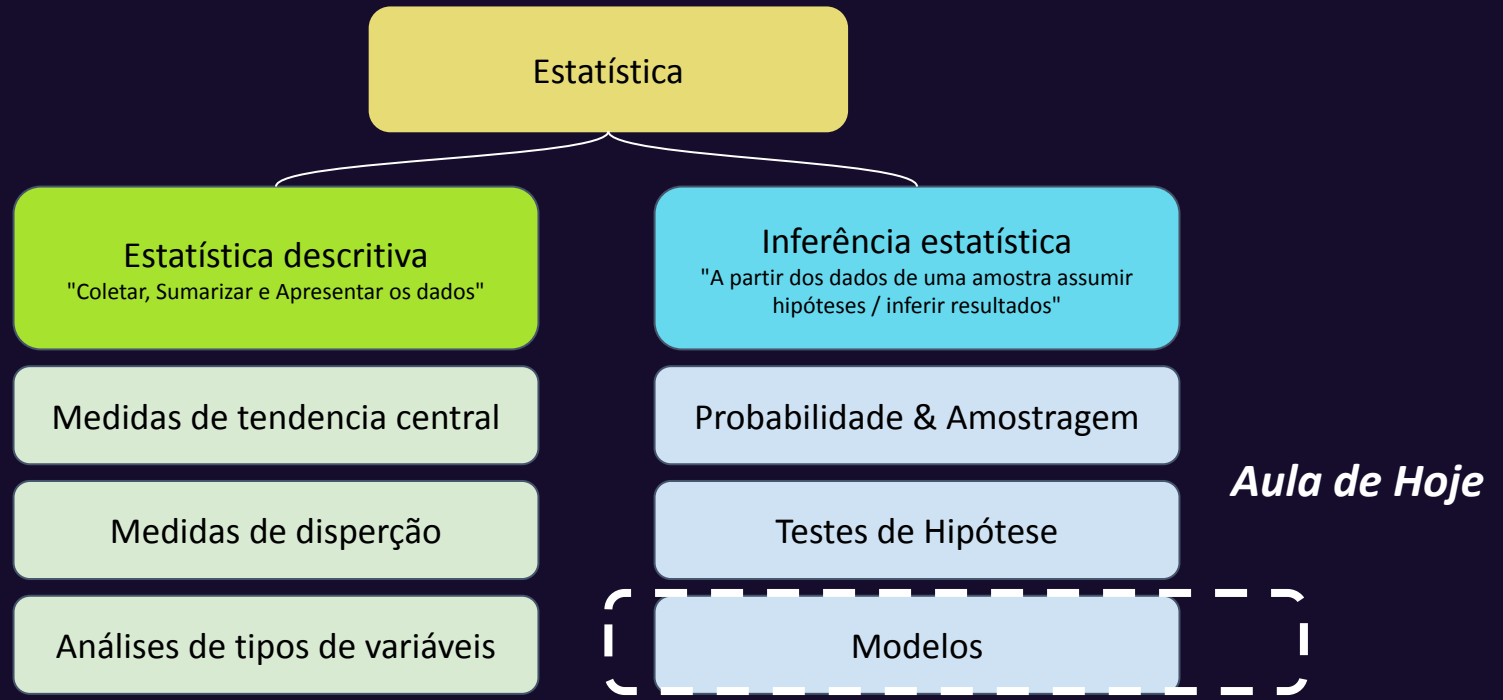
Bootcamp Data Analytics

Modelos de Classificação

WOMAKERS CODE

Onde estamos no Mapa da Estatística: Modelos de Classificação

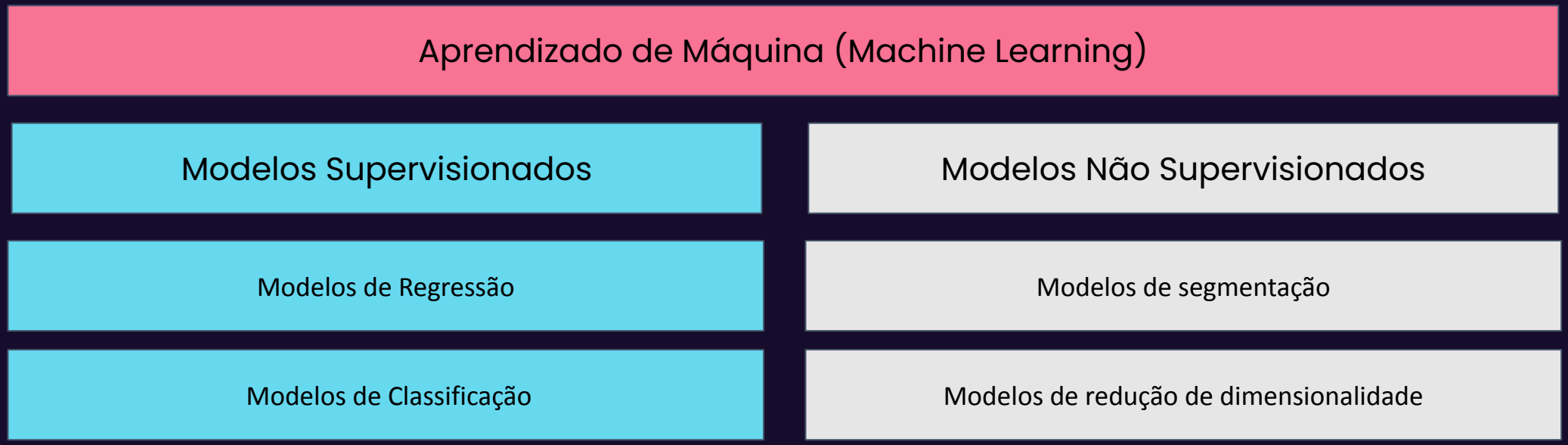


Onde estavamos..

Aprendizado de Máquina (Machine Learning)

Modelos Supervisionados	Modelos Não Supervisionados
São modelos em que temos uma variável de interesse que desejamos **prever seu valor**. Essa variável de interesse, pode ser chamada de variável target ou variável dependente. Ex: preço, renda...	São modelos utilizados na detecção de padrões. Nesse caso o próprio algoritmo encontra os padrões e não temos uma variável target.

Supervisionados vs Não Supervisionados



Os modelos supervisionados podem ser divididos em:

1. modelos de regressão nos quais a variável target, que queremos prever sao contínuas. Exemplo: renda, preço de um imóvel.
2. modelos de classificação nos quais a variável target, que queremos prever são binárias (podem ser chamadas também de variáveis dummy), assumindo valores de 0 ou 1. Exemplo: fraude (1) ou não fraude (0), inadimplência (1) ou adimplência

No universo dos modelos supervisionados

Aprendizado de Máquina (Machine Learning)

Modelos Supervisionados	Modelos Não Supervisionados
Modelos de Regressão	Modelos de segmentação
Modelos de Classificação	Modelos de redução de dimensionalidade

Na aula de hoje iremos introduzir os modelos de classificação. focando em um tipo de modelo específico a regressão logística.

Estatística : Classificação

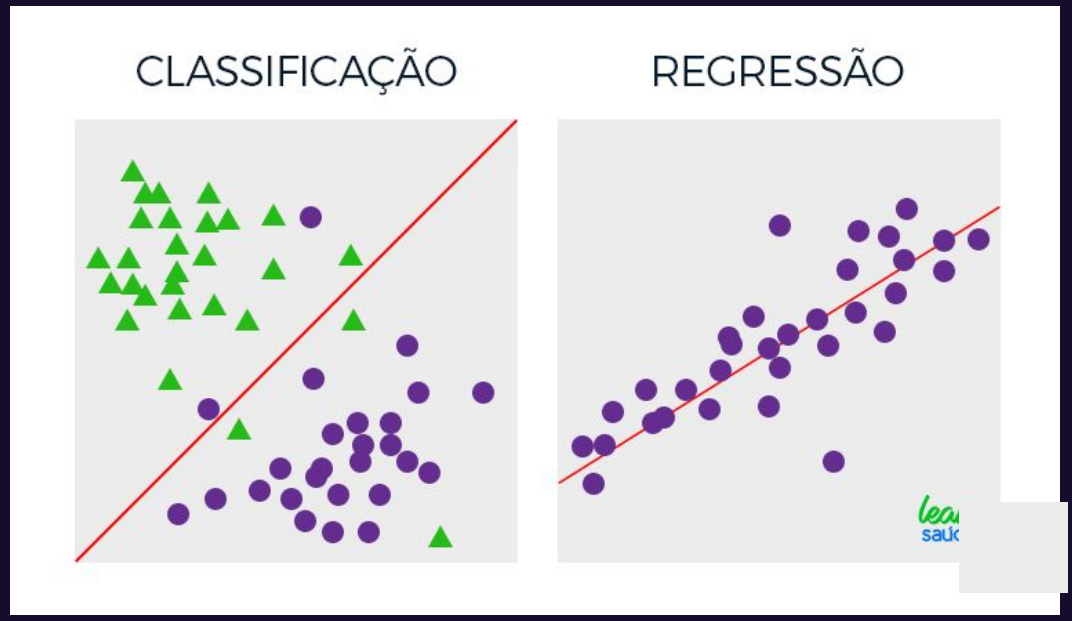
Introdução

WOMAKERS CODE

Classificação vs Regressão

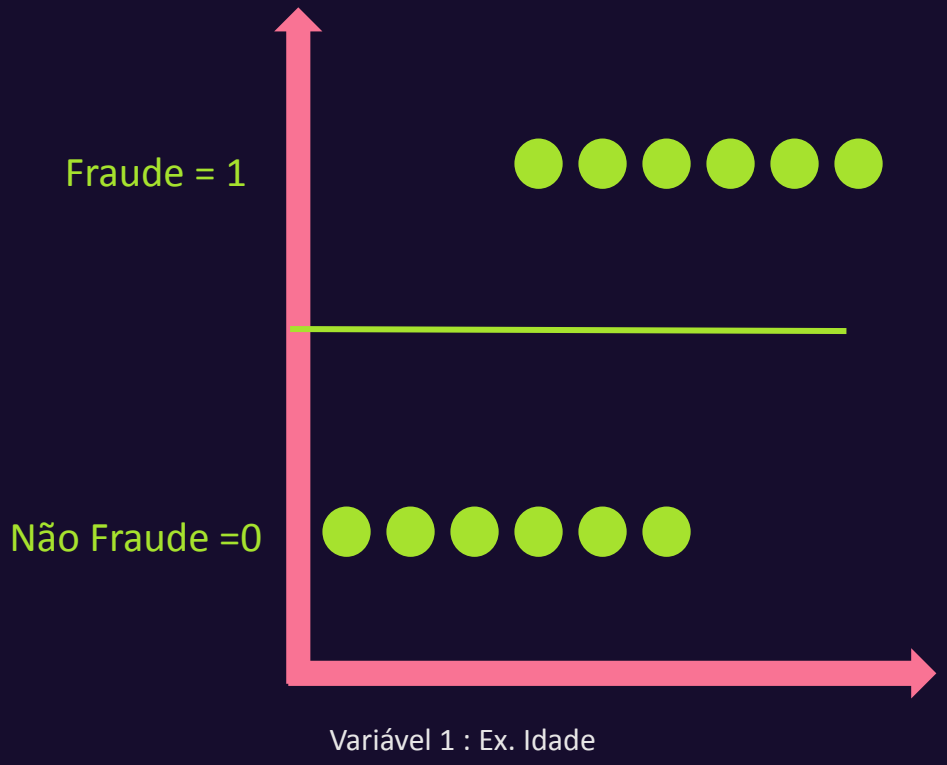
Diferente de um modelo de regressão em que queremos prever o valor de uma variável numérica como renda, valor do imóvel, etc nos modelos de classificação queremos classificar uma entidade, por exemplo:

- Classificar se uma transação é fraude ou não fraude
- Classificar se o indivíduo irá pagar o empréstimo ou não
- Classificar se o indivíduo tem uma comorbidade ou não



Modelos de Classificação

Na prática modelos de classificação são responsáveis por reconhecer e agrupar objetos com base em categorias pré-definidas.



Em todo modelo de classificação temos na base de dados que vamos analisar uma variável chamada target. Essa variável target no caso do modelo de classificação conterá a marcação do grupo/ subpopulação. Exemplo: Fraude

Nos modelos de classificação queremos entender quais variáveis estão relacionadas a uma maior probabilidade de ocorrência de uma categoria. ex: idade.

A marcação geralmente segue o padrão de dummy. sendo no caso ao lado Fraude = 1 ; Não Fraude = 0

Modelos de Classificação

Variáveis Explicativas

- Cor
- Calorias
- Tamanho embalagem
- Tempo Validade

+

"Label"

Vegetal

Não Vegetal

Fit Modelo

Salva o modelo em um arquivo binário (.pkl)

Novo Input

- Cor: Marrom
- Calorias: 400kcal
- Tamanho embalagem: 250
- Tempo Validade: 3 meses

+

Aplica Modelo

Previsão

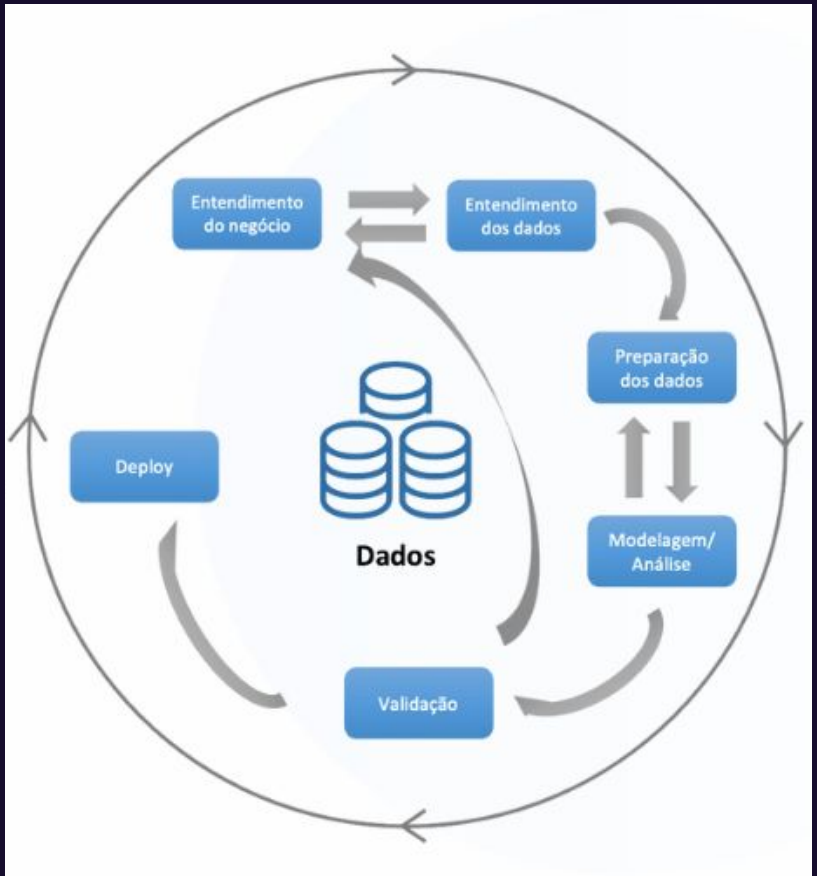
Não Vegetal

Probabilidade: 70%

O Passo a passo da modelagem

Metodologia CRISP-DM

1. Entendimento do negócio: Verificamos o que queremos classificar ; Qual a target?
2. Entendimento dos dados: Fazemos uma análise exploratória
3. Preparação dos dados: Nessa etapa vamos
    a. Separar os dados entre treino e teste
    b. Analisar dados faltantes e outliers
4. Modelagem (Fit do modelo):Nessa etapa vamos fazer o fit do modelo na base de treino. de modo que ele vai aprender os padrões nessa base.
5. Análise de Performance do modelo (validacao): Vamos aplicar o modelo construído na base de treino na base de teste e ver a performance, se classificou corretamente ou não.
6. Deploy: Vamos "salvar" esse modelo para qualquer nova entidade com as variáveis utilizadas ter uma classificação



Preparação dos dados: Treino e Teste

Os modelos de classificação precisam aprender os padrões dos dados para classificar então um dos requisitos necessários para a modelagem será separar a sua base de dados em treino e teste.

- **Random Split:** Podemos escolher uma separação aleatória: 80% da base será destinada ao treino e 20% teste
- **Out of time:** Se na base de dados contiver uma dimensão temporal podemos escolher uma separação temporal : Meses 1,2,3,4,5,6,7,8,9 destinados ao treino e meses 10, 11, 12 ao teste
- **Out of sample:** Se na base de dados contiver uma dimensão " id" podemos escolher uma separação por id. Treino contém 80% dos "ids" unicos e 20% dos "ids" unicos no teste.

Qual escolher? A escolha dependerá do caso de uso.

Se a variável de interesse tem um efeito temporal podemos separar por tempo exemplo: Fraudes. No caso de uma fraude a dimensão temporal é muito importante pois novos padrões de fraude surgem a cada dia.

Preparação dos dados: Analise de valores nulos e outliers.

Posterior à separação de treino e teste é necessário analisarmos variáveis que possuem valores nulos e outliers.

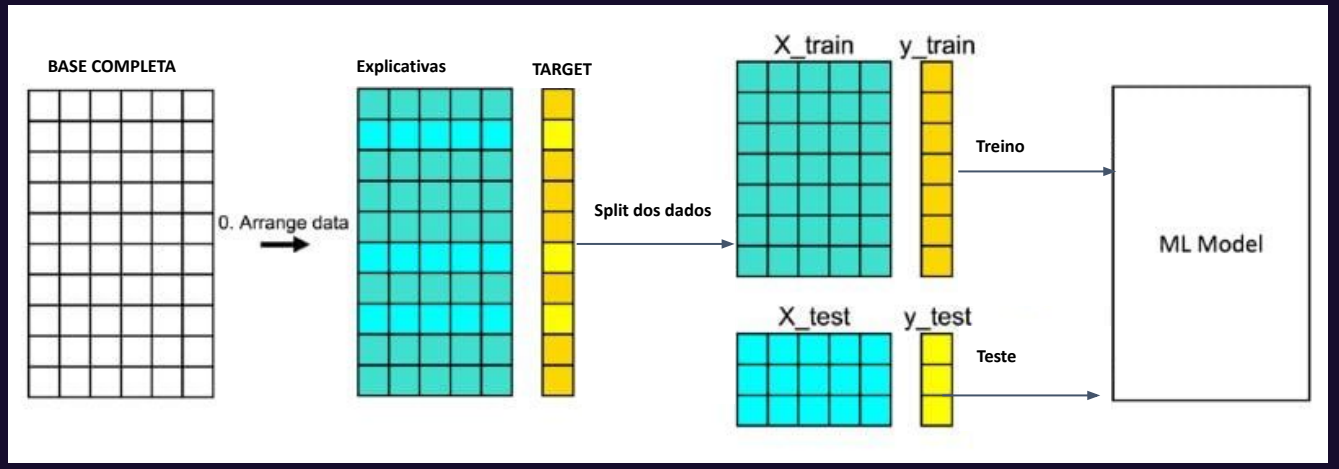
Muitos modelos nao "sao suportados" para lidarem com valores nulos, então nesse caso os valores podem ser substituídos pela média ; mediana ; outro valor à critério da analista.

Um ponto de atenção é que tais substituições devem ocorrer após a separação do treino e teste, pois caso contrário pode haver o que chamamos de vazamento de dados de teste no treino.

A substituição de valores pela média total poderia fazer com que valores discrepantes na base de teste fossem captados já na base de treino, o que leva ao overfit.

Modelagem: Os modelos de classificação

- Os modelos de classificação vão ser fitados na base de treino conforme a imagem abaixo
- São inputs dos modelos: X_train, y_train
- Sendo:
  X_train : o data frame de treino com todas as variáveis explicativas da target no formato numérico e limpo.
  y_train: a variável target, no caso como dummy para a base de treino.
  X_test : o data frame de teste com todas as variáveis explicativas da target no formato numérico e limpo.
  y_test: a variável target, no caso como dummy para a base de teste



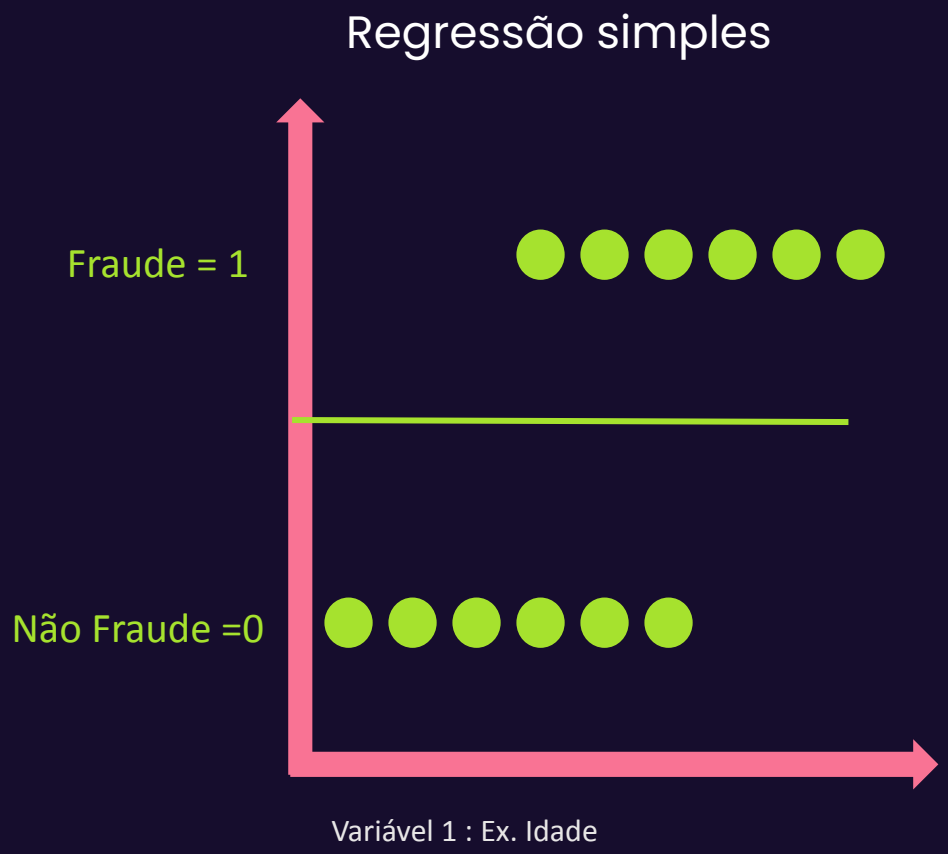
Estatística : Classificação

Modelagem: A Regressão Logística

WOMAKERS CODE

O problema da regressão simples

A Regressão simples não resolve o problema da classificação



Por estarmos com variáveis dummy 0 ou 1, note que o eixo y passa a ser uma probabilidade do evento fraude.

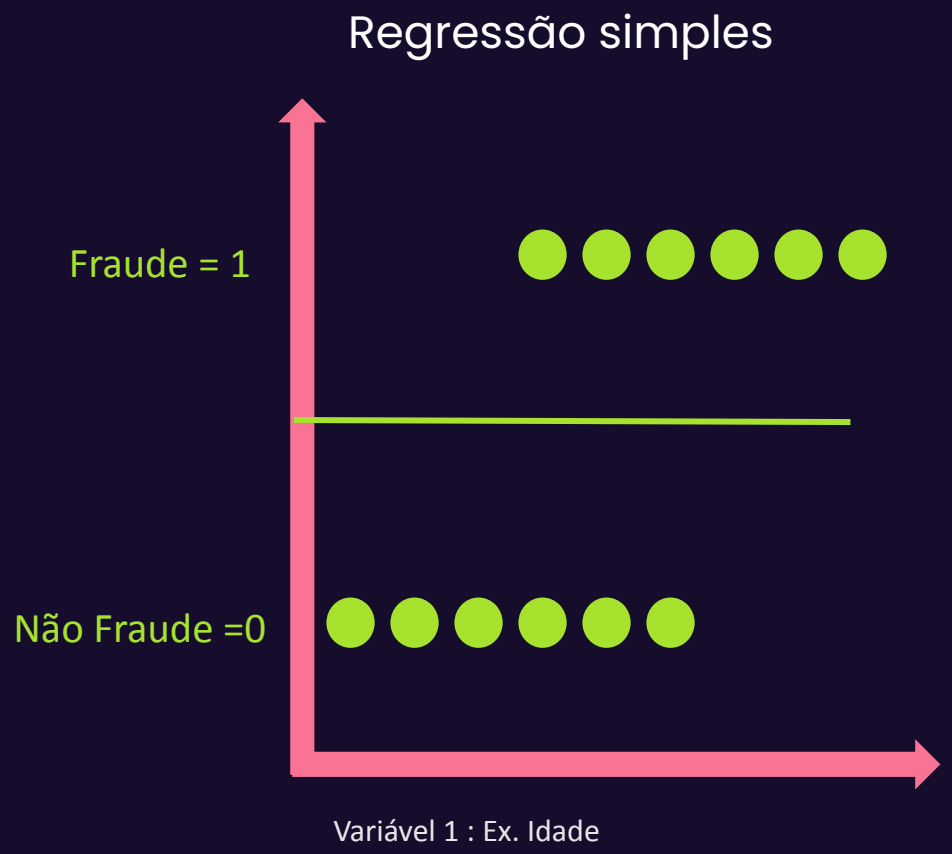
se fitássemos uma regressão simples, para qualquer idade o indivíduo teria probabilidade 0.5 de fraude.

Se selecionarmos um indivíduo com idade alta,
pela regressao simples a probabilidade de ser fraude seria 0.5 se fosse baixa também seria 0.5

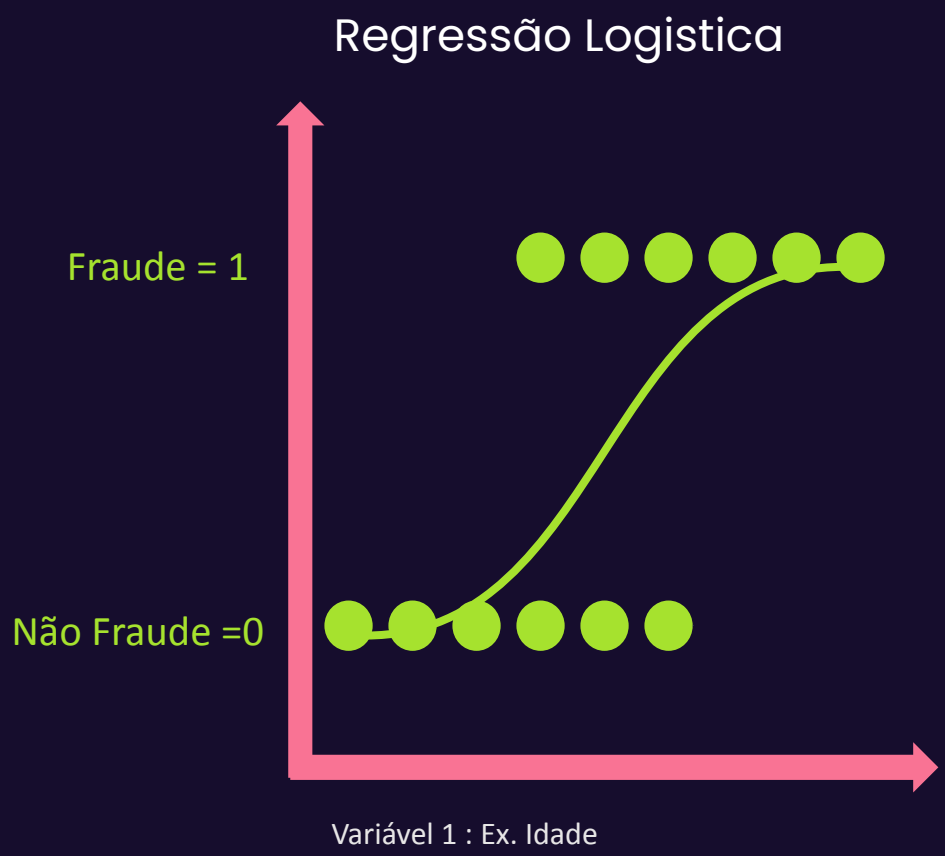
... a regressão simples não parece resolver nosso problema

O problema da regressão simples

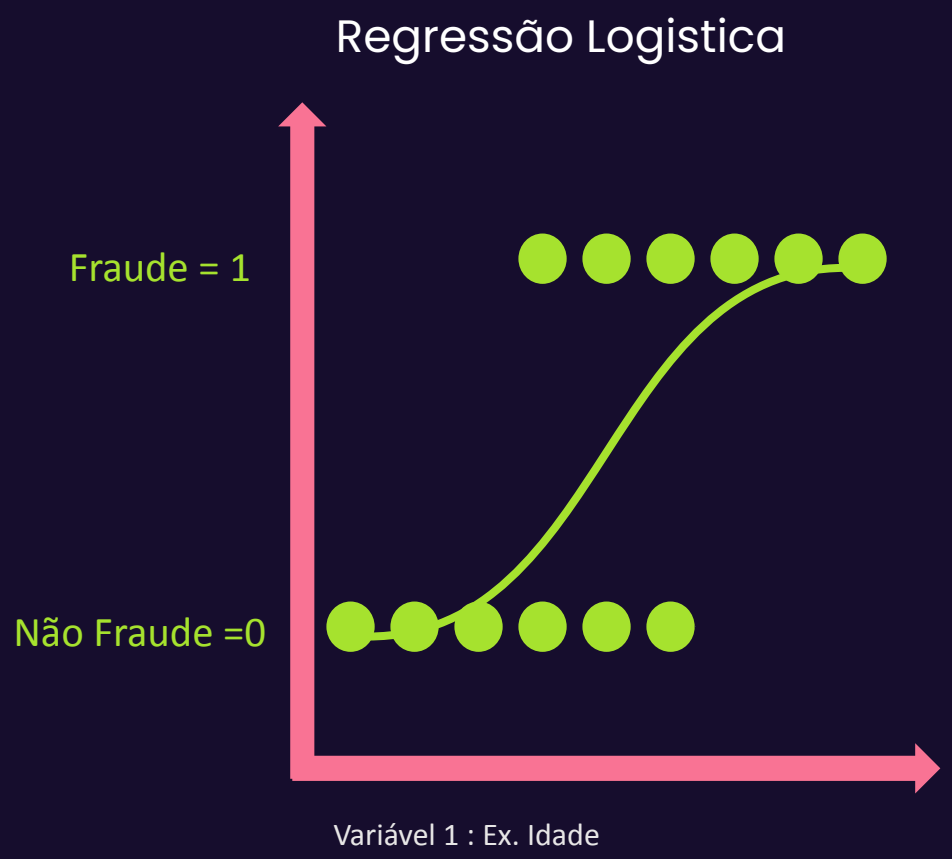
A regressão logística resolve esse problema fitando uma curva em formato de S. Essa curva segue uma função chamada: função logistica... por isso o nome do modelo



X



O problema da regressão simples

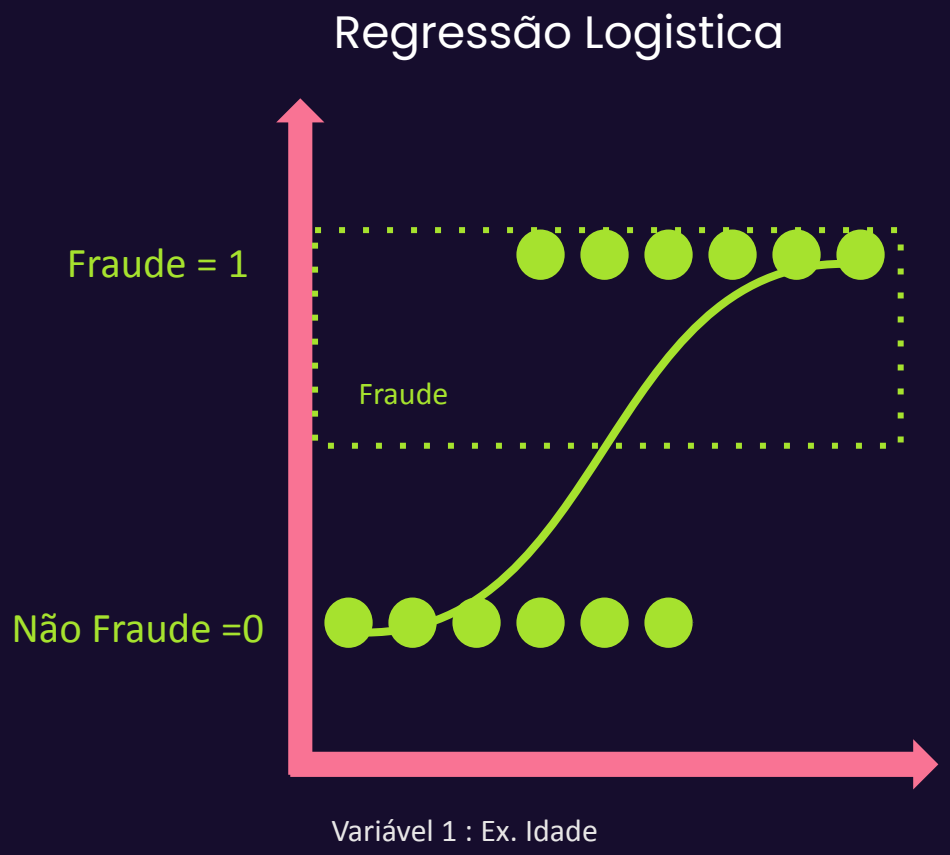


Diferentemente da regressão simples se selecionarmos um indivíduo com idade alta,
pela regressão logística a probabilidade de ser fraude seria 1

se fosse selecionado um indivíduo com baixa idade a probabilidade seria próxima à 0.

... a função logística se adequa muito mais aos dados.

Como funciona a regressão logística?



Na prática, se a probabilidade for maior que 0.5, podemos classificar como fraude ; se for menor que 0.5 podemos classificar como não fraude.

A regressão logística não tem o conceito de "resíduo" como na regressão simples, por isso não usamos a minimização dos quadrados dos resíduos para fitar a curva. ela utiliza um método chamado Máxima verossimilhança.

A função logística é dada por:

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

sendo g(x) linear , g(x) = a1 + b1*X1 + b2*X2 + b3*X3

Fit e interpretação

O fit do modelo utiliza os inputs X_train e y_train e pode ser feito utilizando como base o pacote sklearn.

O Output do modelo assim como no modelo de regressão é uma tabela, como a abaixo: em que analisaremos os p-valores e coeficientes.

No exemplo ao lado:

- a variável target admit indica se o aluno passou ou não no processo seletivo.
- as variáveis explicativas são notas do aluno em provas diferentes.

A interpretação dos coeficientes é um pouco diferente no caso. para ver o efeito de X1 em y é necessário aplicar a fórmula:
A nota X1 aumenta em (exp^coeficiente) a probabilidade de admissão

Logit Regression Results

Dep. Variable:	admit	No. Observations:	400
Model:	Logit	Df Residuals:	394
Method:	MLE	Df Model:	5
Date:	Fri, 30 Mar 2018	Pseudo R-squ.:	0.08292
Time:	12:02:40	Log-Likelihood:	-229.26
converged:	True	LL-Null:	-249.99
		LLR p-value:	7.578e-08

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.9900	1.140	-3.500	0.000	-6.224	-1.756
C(rank)[T.2.0]	-0.6754	0.316	-2.134	0.033	-1.296	-0.055
C(rank)[T.3.0]	-1.3402	0.345	-3.881	0.000	-2.017	-0.663
C(rank)[T.4.0]	-1.5515	0.418	-3.713	0.000	-2.370	-0.733
gre	0.0023	0.001	2.070	0.038	0.000	0.004
gpa	0.8040	0.332	2.423	0.015	0.154	1.454

Estatística : Classificação

Métricas de Performance

WOMAKERS CODE

Como saber se o modelo está performando bem?

Existem diferentes métricas de performance de um modelo de classificação. Diferentemente do modelo de regressão linear em que tínhamos o R-quadrado, nos modelos de classificação temos múltiplas métricas de performance a serem analisadas de acordo com o problema que queremos resolver.

A base da maioria das métricas de formance é o que chamamos de matriz de confusão.A matriz de confusão é construída com base nos dados preditos pelo modelo na base de teste vs a classificacao original.

		Classe Esperada	
		Fraude	Não Fraude
Classe Prevista	Fraude	VP (verdadeiro positivo)	FP (Falso Positivo)
	Não Fraude	FN (Falso Negativo)	VN (Verdadeiro Negativo)

A Matriz de confusão: Métricas

		Classe Esperada	
		Fraude	Não Fraude
Classe Prevista	Fraude	VP (verdadeiro positivo)	FP (Falso Positivo)
	Não Fraude	FN (Falso Negativo)	VN (Verdadeiro Negativo)

1. **Acurácia:** Verifica o total de acertos dividido pelo total de observações

   VP + VN / (VP + FP + FN + VN)

2. **Precisão:** Verifica o total de acertos positivos dadas as observações previstas positivas. (Hit rate)

   VP / (VP + FP)

3. **Recall:** Verifica o total de acertos positivos dado o que realmente era positivo (Catch rate)

   VP / (VP + FN)

Exemplos & Métricas escolhidas

Escolhemos métricas como a acurácia quando:

- O objetivo de negócio é o acerto independente da classe.
- quando temos classes balanceadas (proporção de 1ns e 0s na base é semelhante)
- Exemplo: Prever se é cenoura ou laranja.

Escolhemos métricas como precisão e recall quanto:

- O objetivo de negócio é o acerto da classe positiva.
- Quando temos classes naturalmente desbalanceadas (ex: proporção de 1ns é menor)
- Exemplo: Detectar fraude (VP é mais importante do que VN) ; Classificação de inadimplência.

Atenção ao utilizar acurácia em bases desbalanceadas!!! Ela pode mascarar o problema que queremos resolver. Podemos ter valores de acurácia alta, mas isso porque a quantidade de informações da classe negativa é muito maior e o modelo acaba classificar somente como essa classe.

Estatística : Classificação

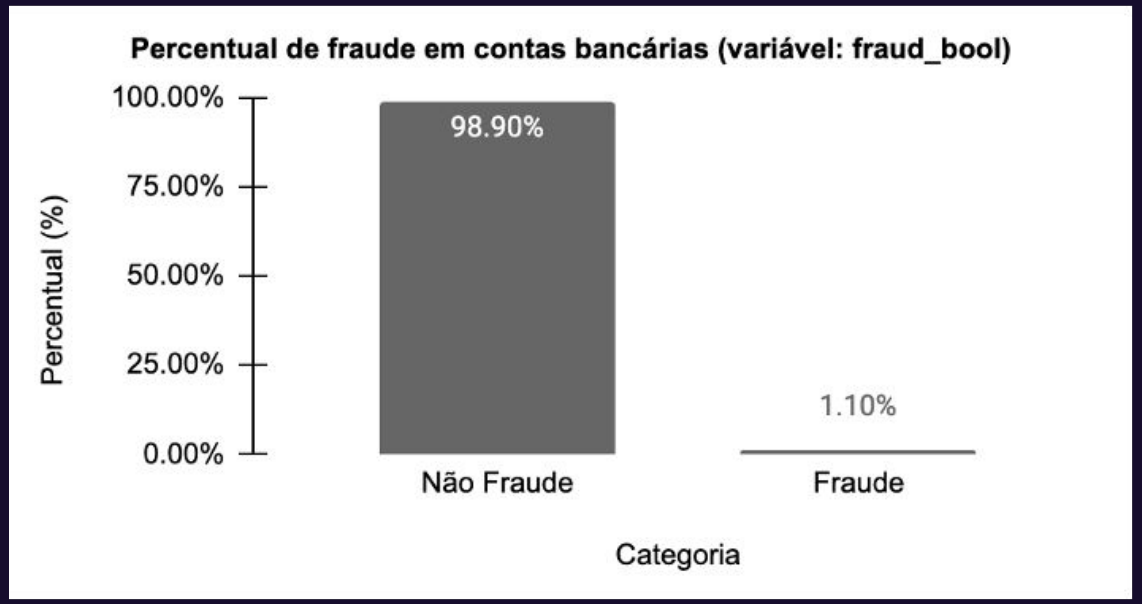
Cuidados em Modelos de Classificação

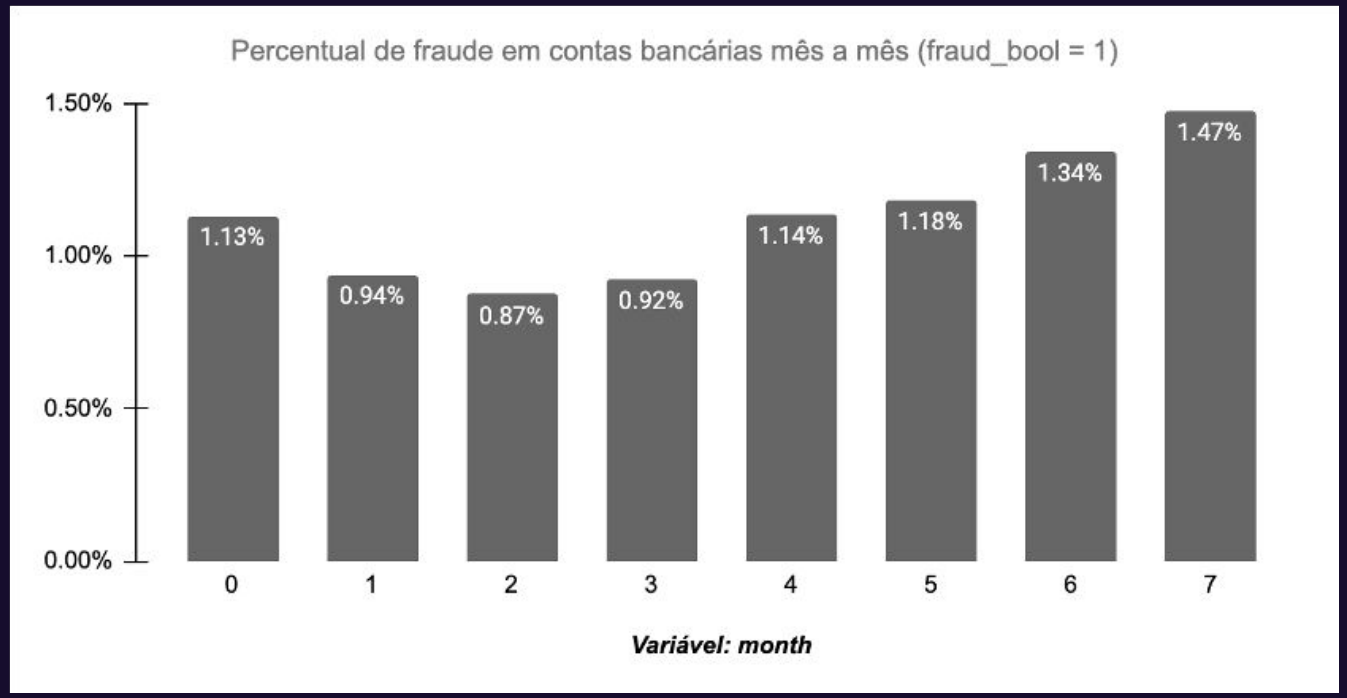
WOMAKERS CODE

1. O Desbalanceamento

Mencionamos que o desbalanceamento ocorre quando temos uma proporção muito maior de uma classe. Como podemos detectar desbalanceamento?

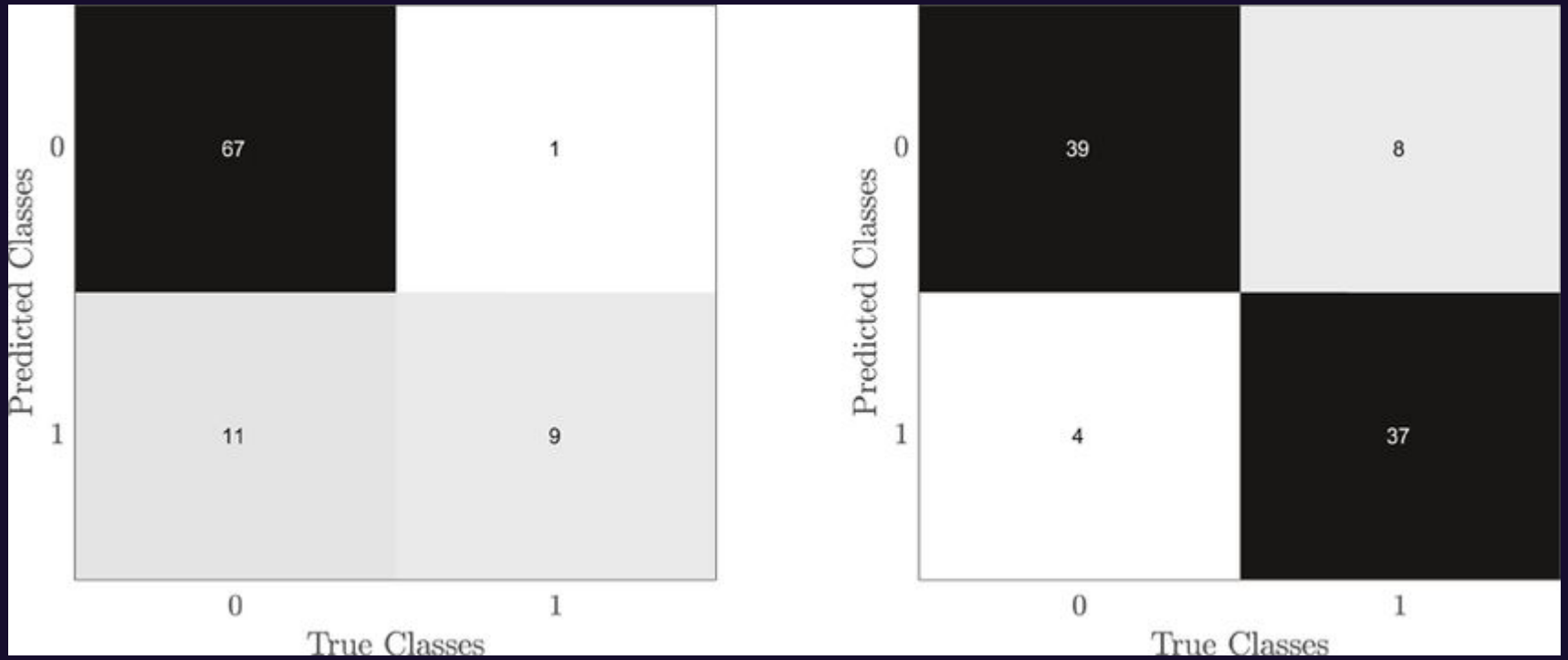
Gráficos com as contagens ou percentuais da quantidade de cada categoria ou ainda percentuais ao longo do tempo





1. O Desbalanceamento

O que acontece com a acurácia no caso de uma base desbalanceada?
A matriz da esquerda contém os dados desbalanceados a da direita balanceados. Vemos que quando os dados estão desbalanceados o modelo acaba só aprendendo os padrões de uma das classes

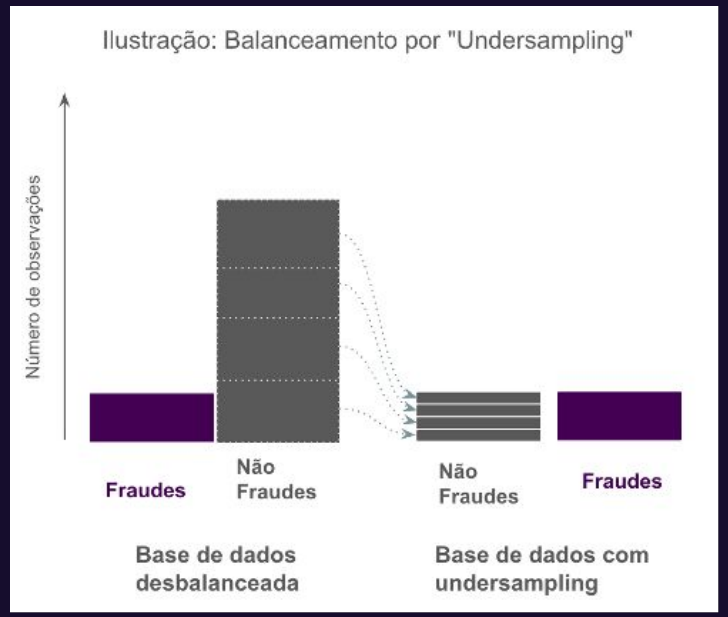


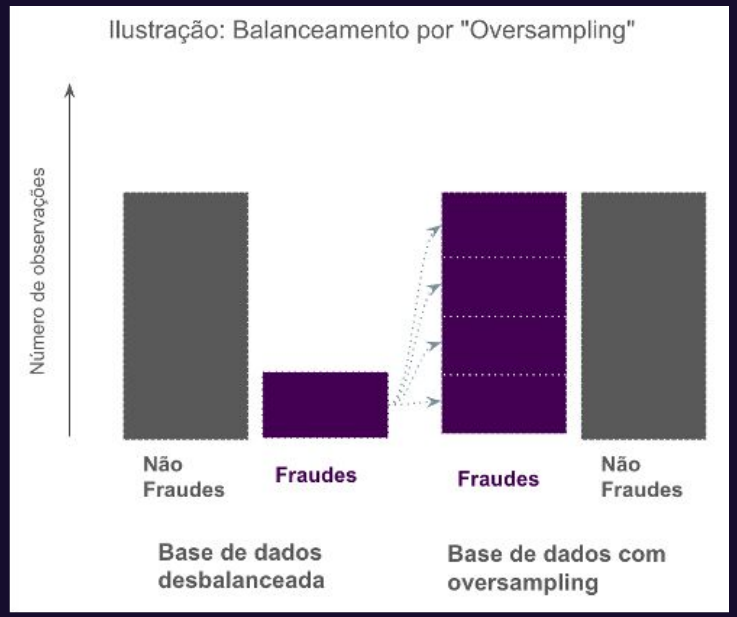
Oversampling vs Undersampling

Para corrigir o desbalanceamento existem diversas técnicas possíveis. Duas delas são: Undersampling e Oversampling.

No Undersampling diminuimos a volumetria total da base de modo que a quantidade de dados da classe majoritária (não fraude) fique igual a da minoritária. A seleção é feita aleatoriamente.

No oversampling repetimos o número de observações da classe minoritária várias vezes até completarmos a mesma quantidade de observações da classe majoritária



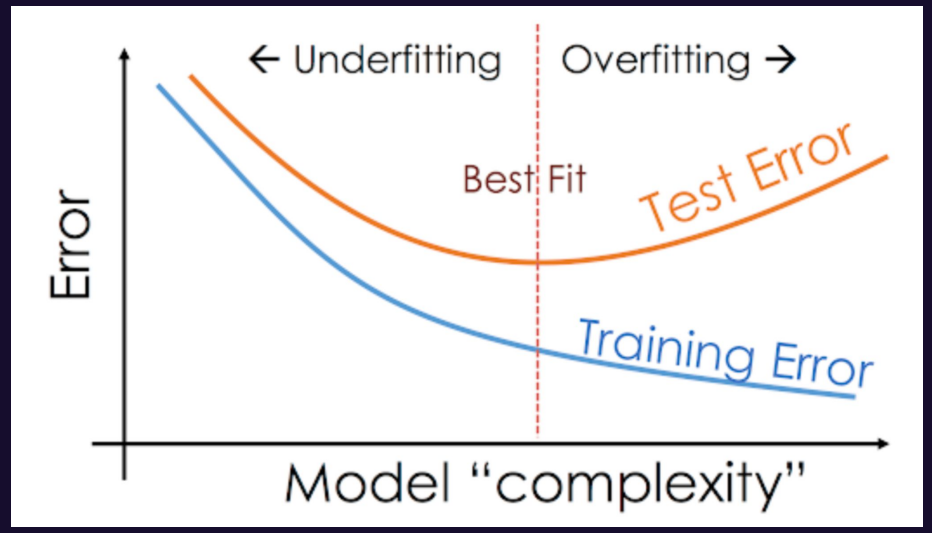


2. Overfit

Assim como nos modelos de regressão outro cuidado que temos ter nos modelos de classificação é com o overfit

O overfit ocorre quando adicionamos muita complexidade ao modelo, ou ainda quando trabalhamos com amostras pequenas. Quando o overfit acontece dizemos que o modelo perdeu capacidade de generalização.

Vemos que o modelo está com overfit quando ao aplicarmos a matriz de confusão na base de teste ou em uma nova base os valores de acurácia, precisão e recall decaem significativamente, do que foi visto anteriormente.



Vamos Praticar em Python!