Introdução à Ciência de Dados

Long Assignment

Churn Data

Inês Silva
Maria Miguel Ribeiro
Renatha Vieira

CRISP-DM

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

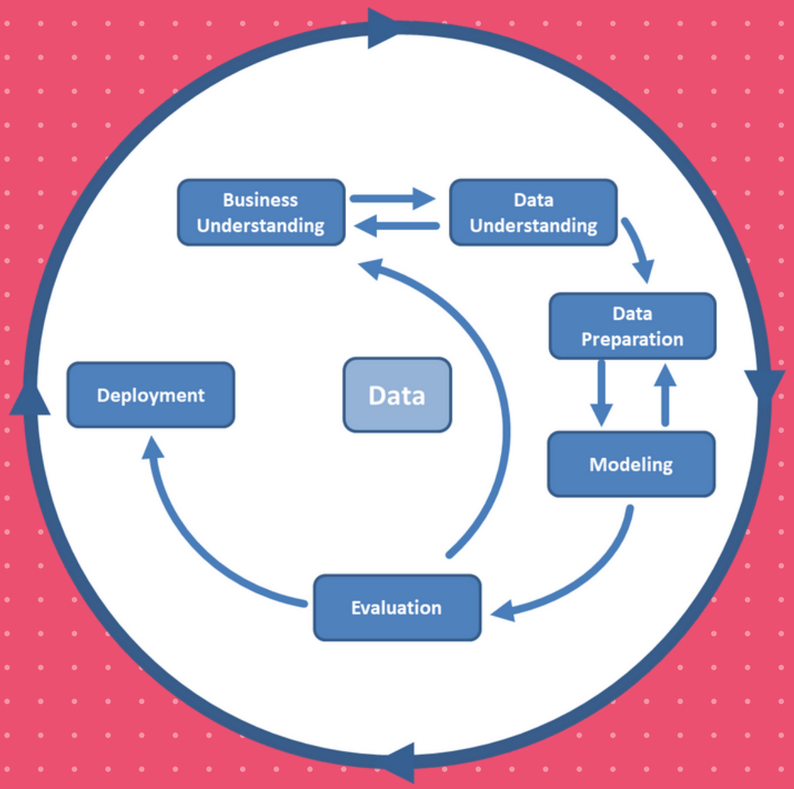


imagem: https://towardsdatascience.com/using-crisp-dm-to-grow-as-data-scientist-a07ce3fd9d56, disponível em 03/01/2023

BUSINESS UNDERSTANDING

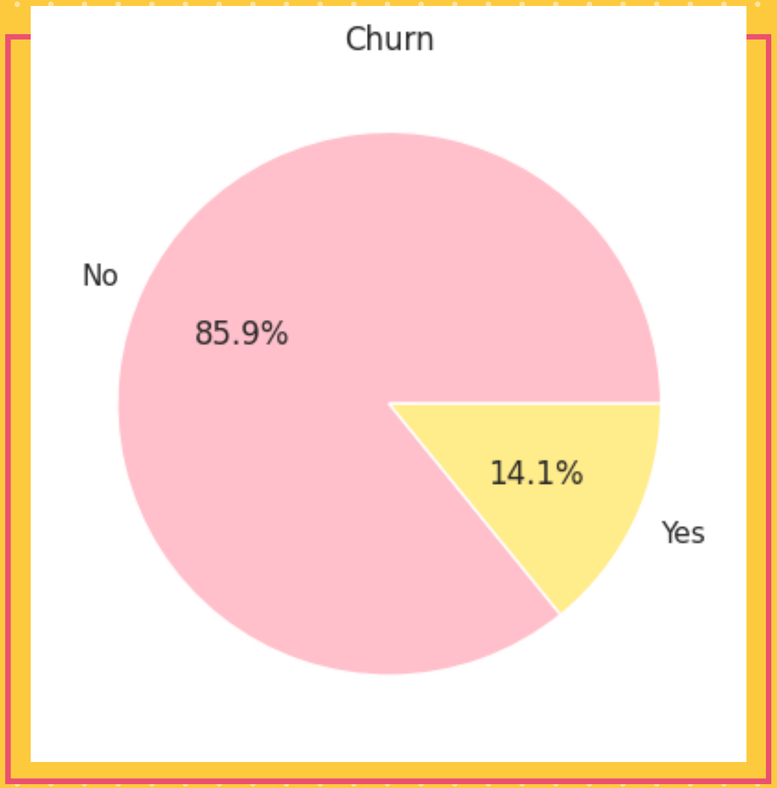
O objetivo deste projeto é analisar, processar os dados e desenvolver modelos preditivos para uma previsão precisa de desistência. Isto significa que o objetivo empresarial, neste caso, é prever corretamente se um cliente irá ou não desistir da empresa.

DATA UNDERSTANDING

Exploração do conjunto de dados de modo a visualizar as diferentes variáveis e compreender possíveis padrões relevantes, com o intuito de extrair perceções (insights).

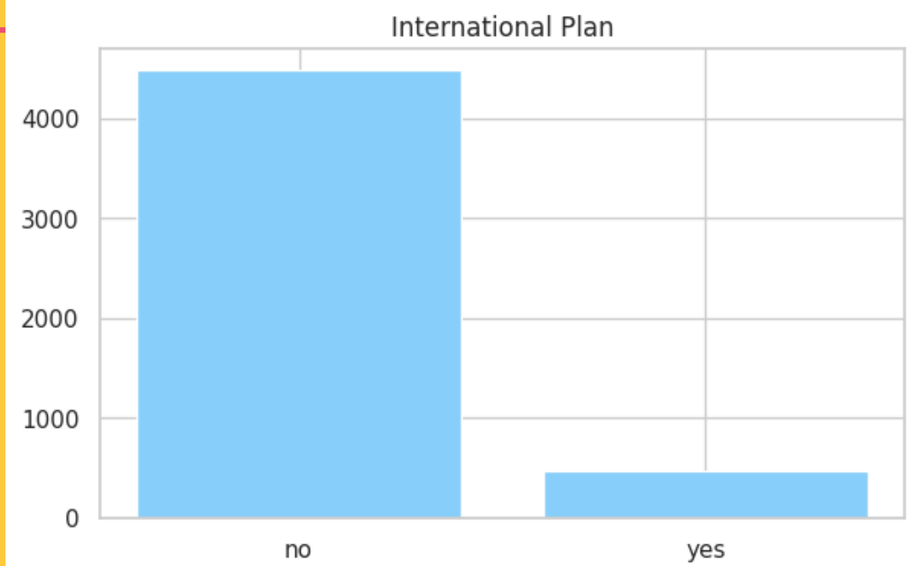
Data Understanding

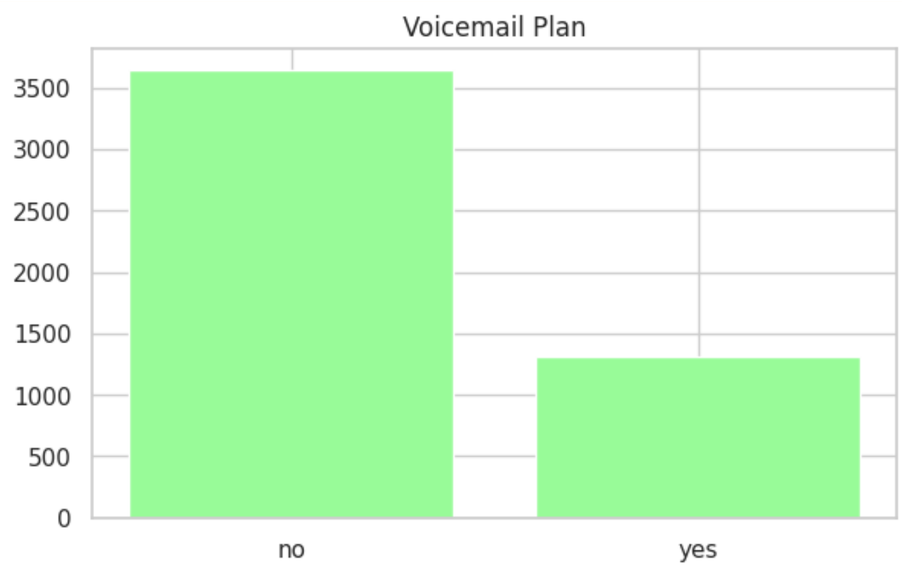
TARGET



Data Understanding

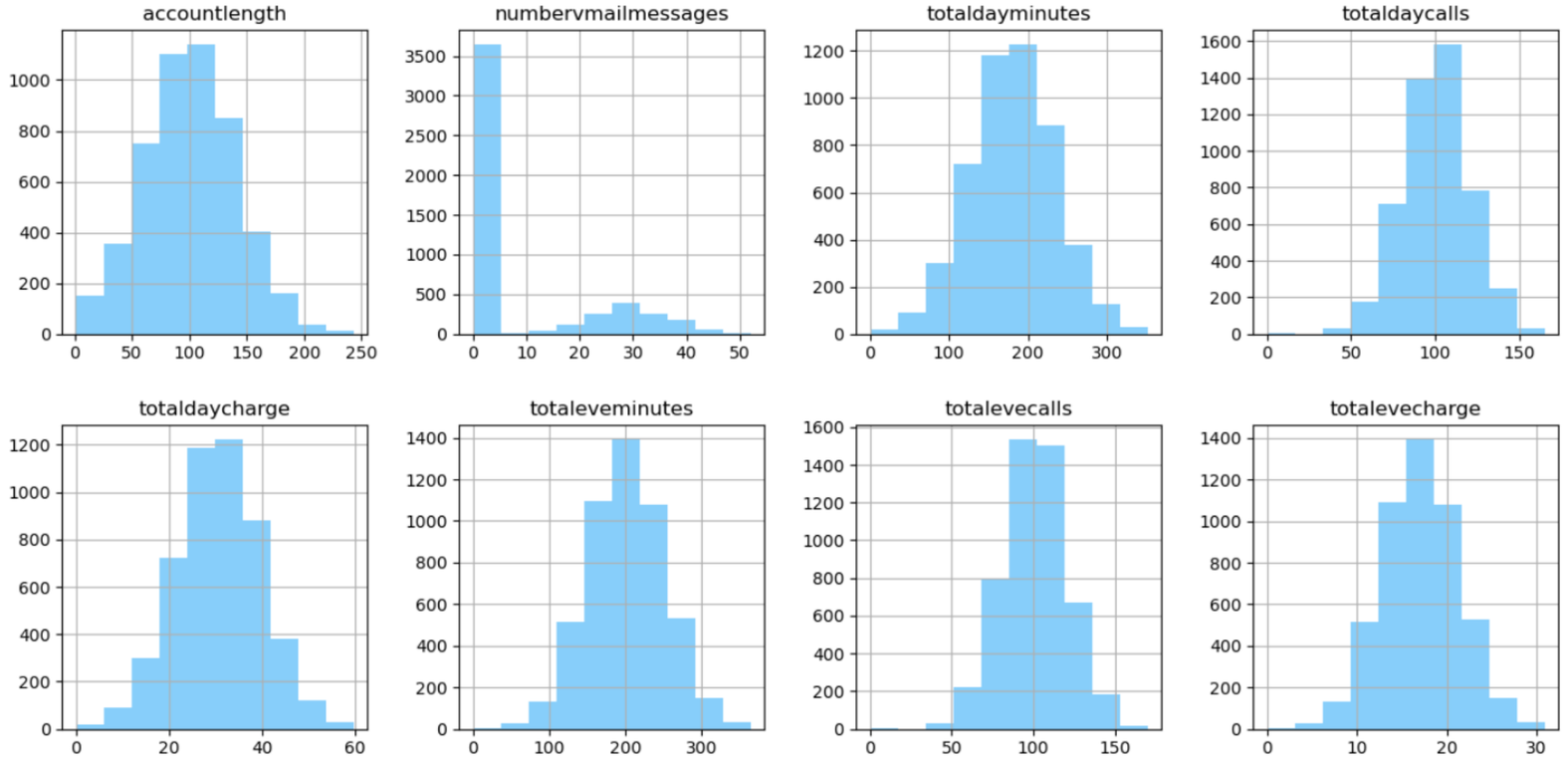
VARIÁVEIS NOMINAIS





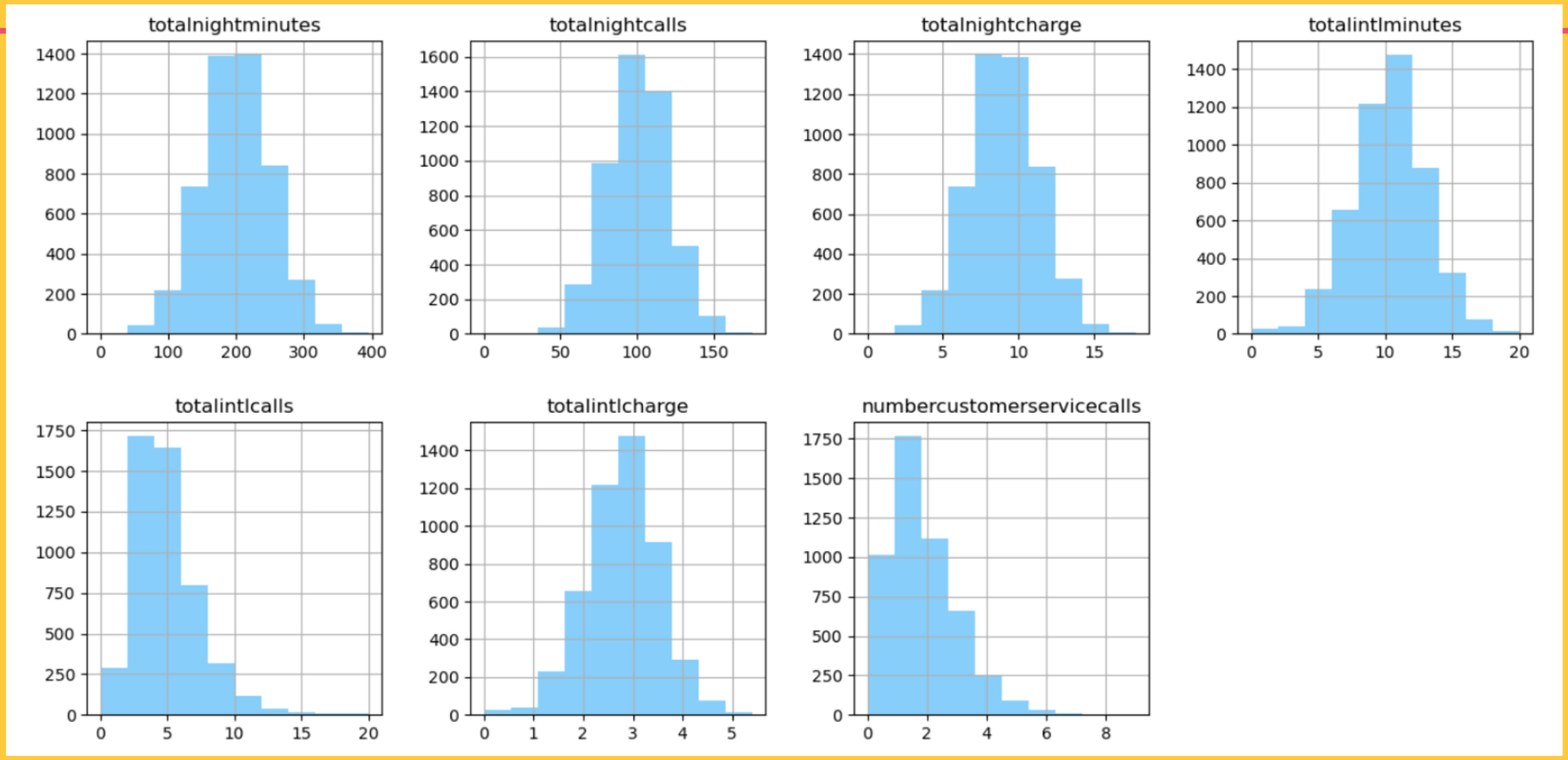
Data Understanding

VARIÁVEIS NUMÉRICAS



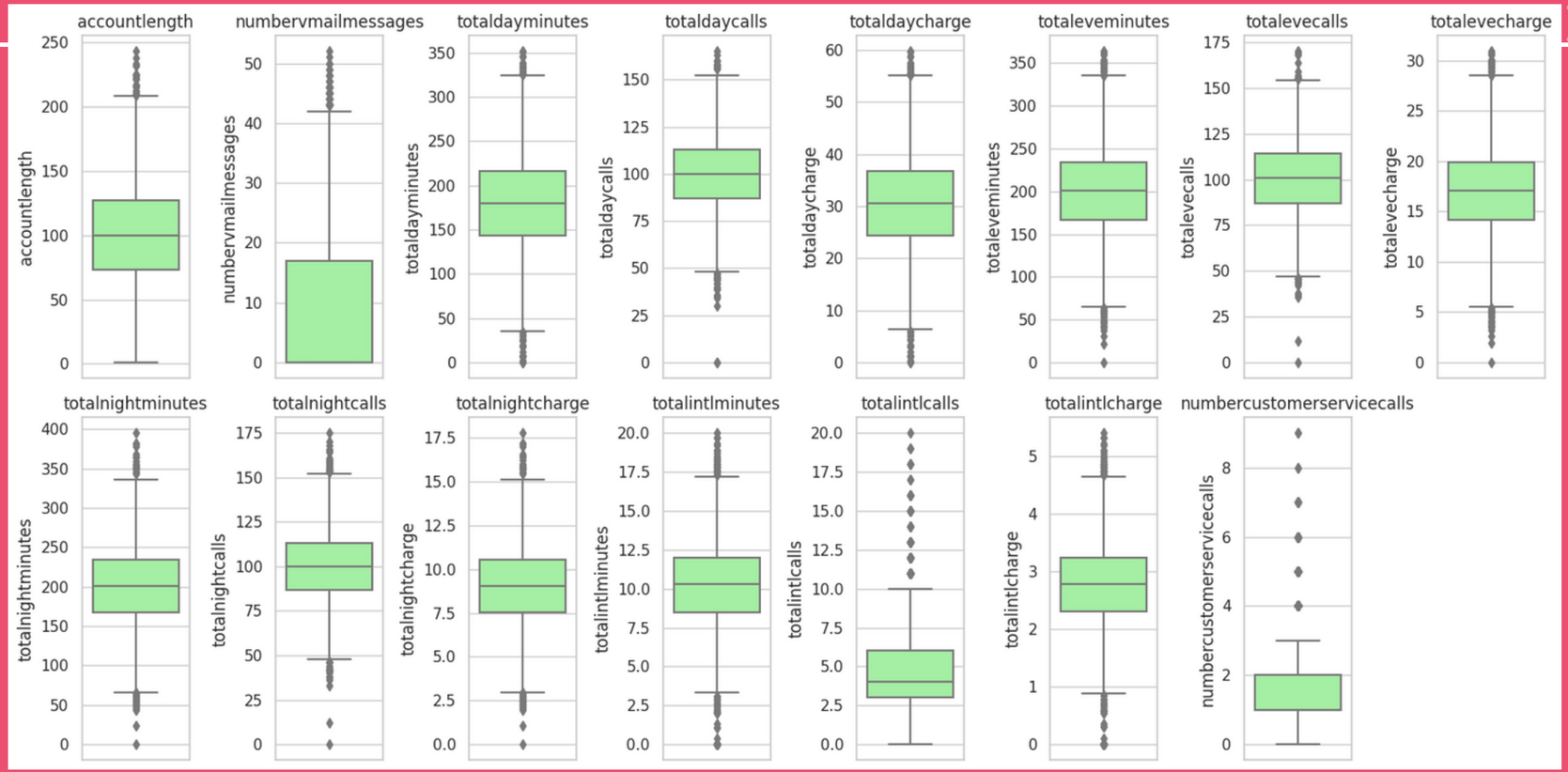
Data Understanding

VARIÁVEIS NUMÉRICAS



Data Understanding

OUTLIERS



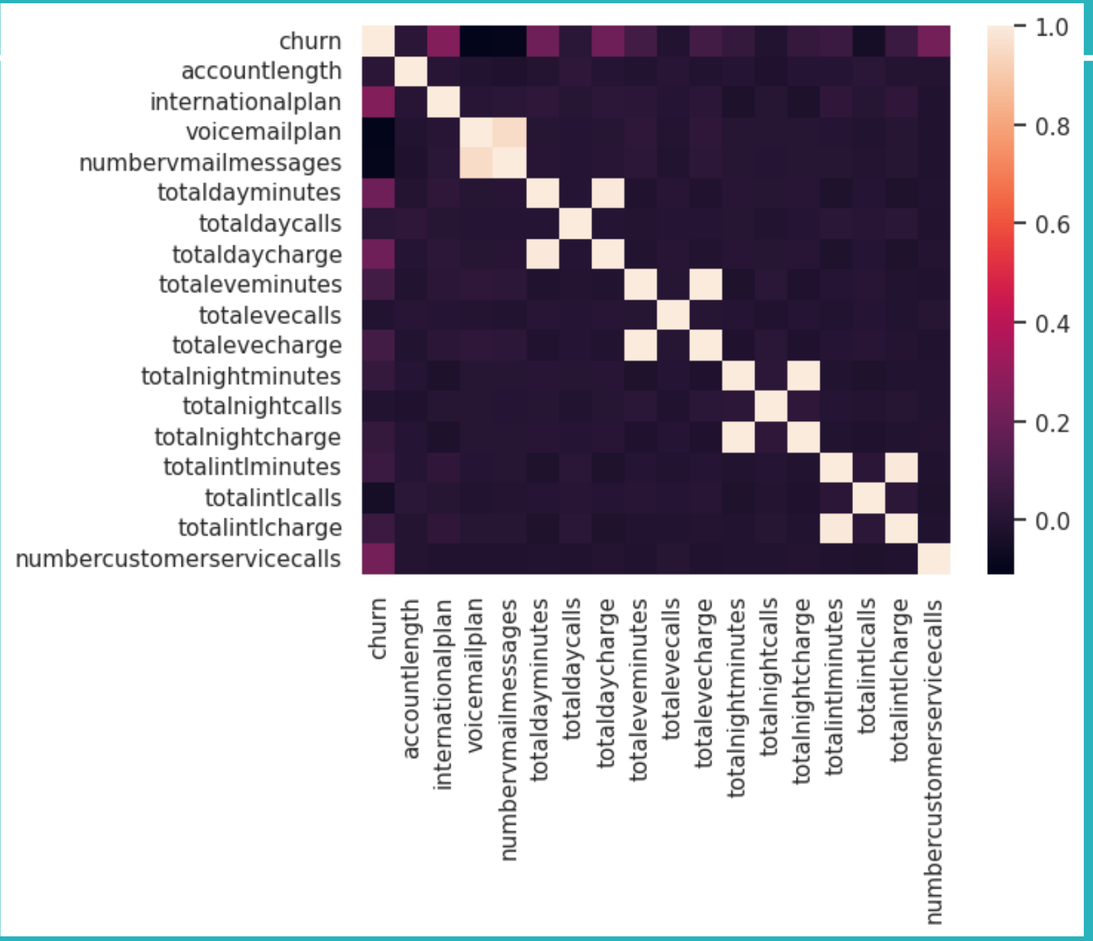
Data Understanding

OUTLIERS

	Quantidade de Outliers
accountlength	24
numbervmailmessages	60
totaldayminutes	35
totaldaycalls	35
totaldaycharge	36
totaleveminutes	42
totalevecalls	27
totalevecharge	42
totalnightminutes	39
totalnightcalls	43
totalnightcharge	39
totalintlminutes	71
totalintlcalls	114
totalintlcharge	70
numbercustomerservicecalls	392

	Coluna	Outlier	Distancia	Limite
0	accountlength	243.0	35.00000	Superior
1	numbervmailmessages	52.0	9.50000	Superior
2	totaldayminutes	0.0	34.95000	Inferior
3	totaldaycalls	0.0	48.00000	Inferior
4	totaldaycharge	0.0	5.99500	Inferior
5	totaleveminutes	0.0	64.91250	Inferior
6	totalevecalls	0.0	46.50000	Inferior
7	totalevecharge	0.0	5.47875	Inferior
8	totalnightminutes	0.0	65.45000	Inferior
9	totalnightcalls	0.0	48.00000	Inferior
10	totalnightcharge	0.0	2.93500	Inferior
11	totalintlminutes	0.0	3.25000	Inferior
12	totalintlcalls	20.0	9.50000	Superior
13	totalintlcharge	0.0	0.89000	Inferior
14	numbercustomerservicecalls	9.0	5.50000	Superior

Data Understanding



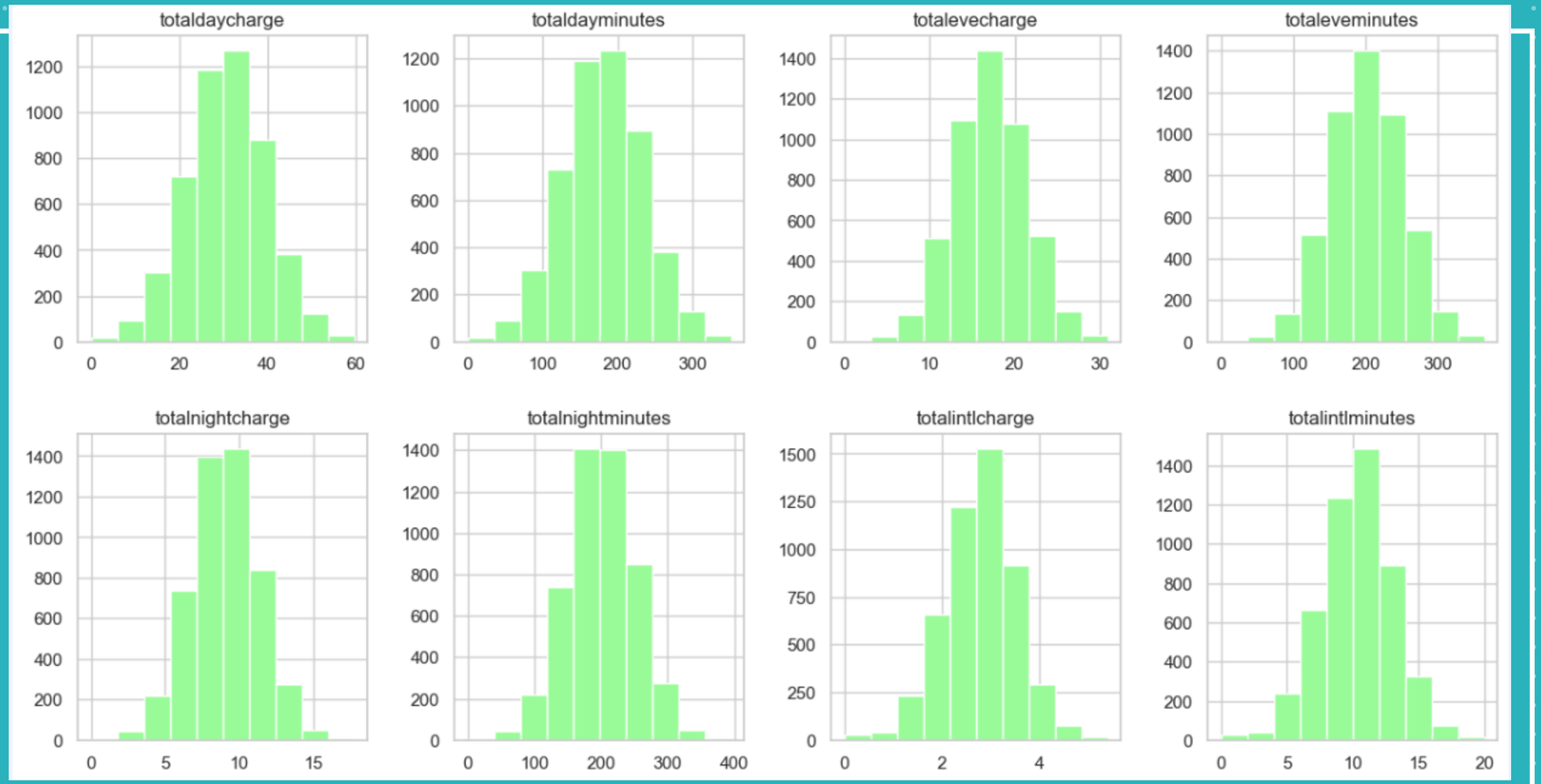
CORRELACÃO

Verificou-se uma grande correlação entre as seguintes variáveis:

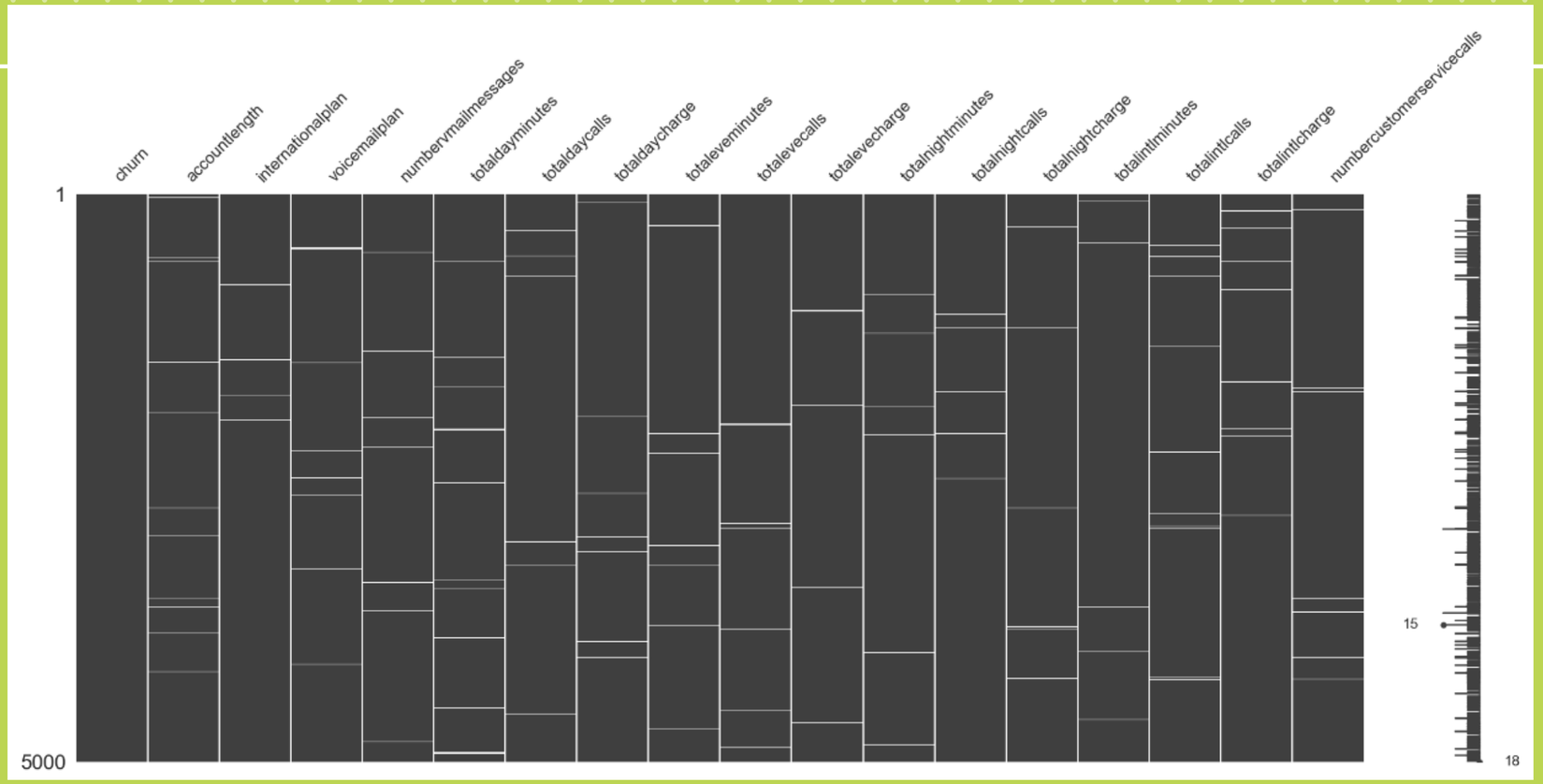
- 'totaldaycharge' e 'totaldayminutes';
- 'totalevecharge' e 'totaleveminutes';
- 'totalnightcharge' e 'totalnightminutes';
- 'totalintlcharge' e 'totalintlminutes'.

'totaldaycharge' ≅ 0.17 * 'totaldayminutes',
'totalevecharge' ≅ 0.085 * 'totaleveminutes',
'totalnightcharge' ≅ 0.045 * 'totalnightminutes',
'totalintlcharge' ≅ 0.27 * 'totalintlminutes'.

Data Understanding



Data Understanding

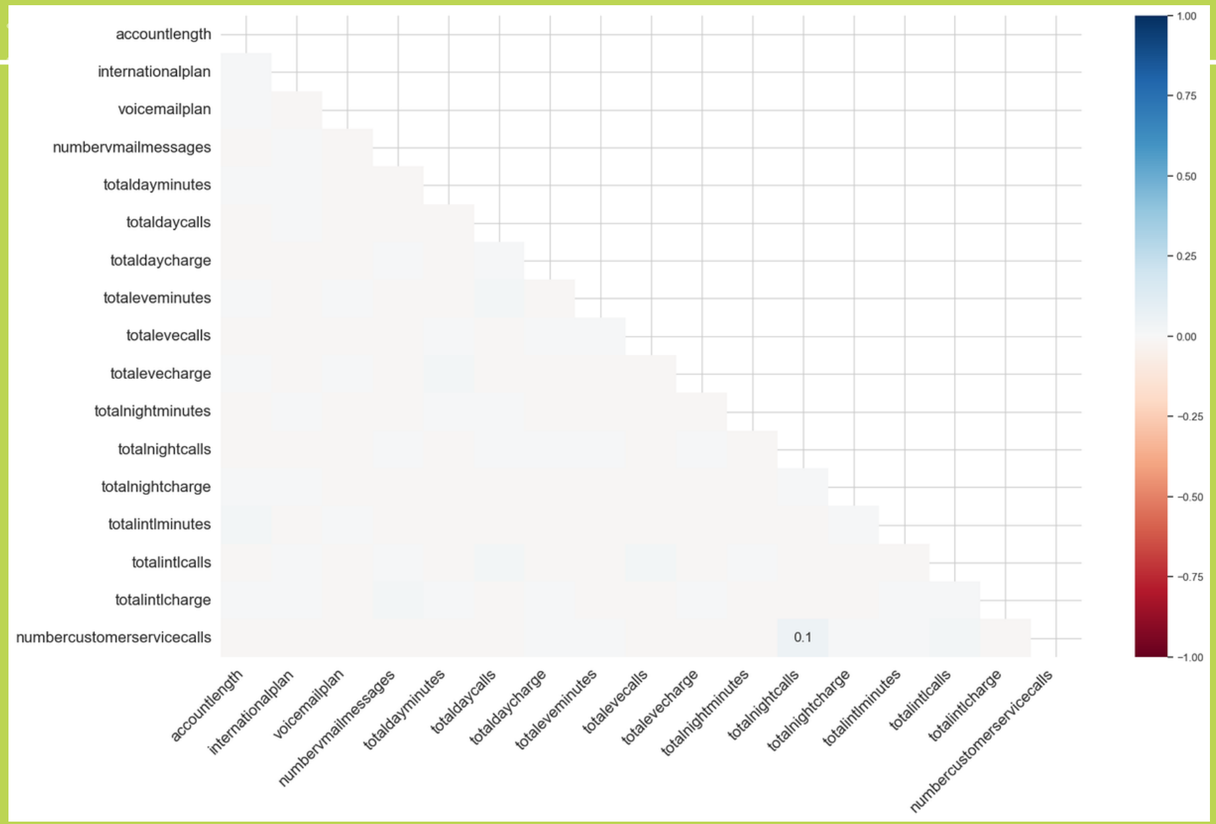


MISSING VALUES

Uso do pacote Missingno

Análise da correlação entre os missing values

Data Understanding



CORRELACÃO

Os missing values são completamente aleatórios (MCAR). São dados que não possuem nenhuma dependência em relação a dados observados ou não observados.

Data Pre-Processing

Imputação de Valores

Variáveis Correlacionadas

Para preencher os missing values de 'totaldayminutes', 'totaleveminutes', 'totalnightminutes' e 'totalintlminutes' utilizamos a seguinte fórmula:

MINUTES = CHARGE / TAXA POR MINUTO

Missing Values Restantes

Como os missing values não apresentam grande correlação entre si, o mais recomendado para a sua imputação são os valores de tendência central, nomeadamente, a mediana.

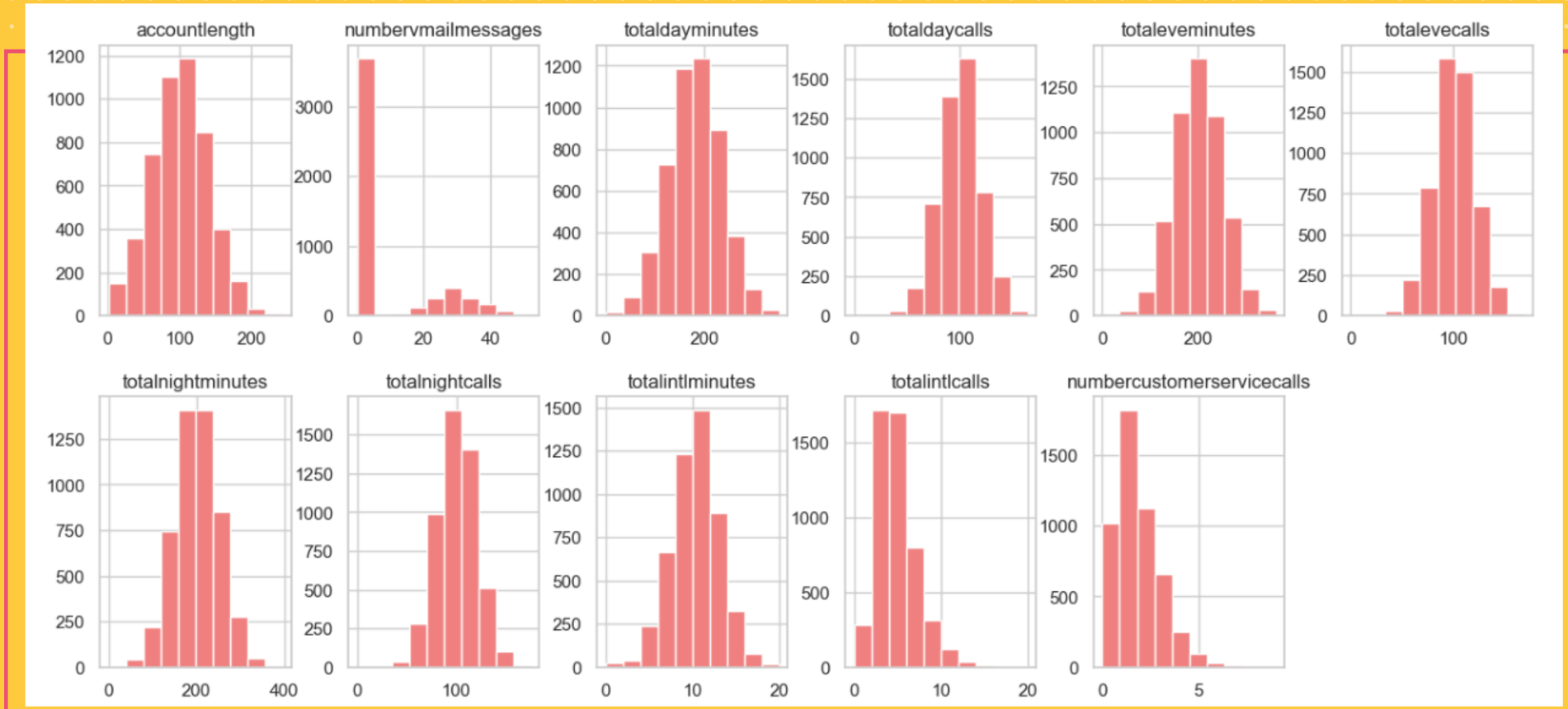
Data Pre-Processing

Exclusão de Varáveis

Variáveis Correlacionadas

Como as variáveis 'totaldaycharge', 'totalevecharge', 'totalnightcharge' e 'totalintlcharge' derivam das variáveis que contabilizam o total de minutos para cada categoria (dia, noite, total e internacional) estas podem ser excluídas.

Data Pre-Processing



Modelação

Dados Desequilibrados

Como o dataset é desiquilibrado, favorece usuários que não abandonam a companhia telefónica.

Dados Equilibrados

Utilização do método SMOTE:
As técnicas de SMOTE são especialmente desenhadas para tratar de datasets desiquilibrados, gerando dados sintéticos para a classe minoritária.

Dados Desequilibrados

Ensemble AdaBoost

Mean Accuracy: 0.8754
Mean Precision: 0.7603
Mean Recall: 0.6415

Decision Tree

Mean Accuracy: 0.9122
Mean Precision: 0.8681
Mean Recall: 0.7412

Support Vector Machine

Mean Accuracy: 0.8578
Mean Precision: 0.4289
Mean Recall: 0.5000

Naive Bayes

Mean Accuracy: 0.8732
Mean Precision: 0.7646
Mean Recall: 0.6054

K-nearest Neighbors

Mean Accuracy: 0.8874
Mean Precision: 0.8307
Mean Recall: 0.6378

Neural Network Classifier

Mean Accuracy: 0.8642
Mean Precision: 0.7136
Mean Recall: 0.6228

Dados Equilibrados

Ensemble AdaBoost

Mean Accuracy: 0.8753
Mean Precision: 0.7597
Mean Recall: 0.6452

Decision Trees

Mean Accuracy: 0.9145
Mean Precision: 0.8798
Mean Recall: 0.7427

Support Vector Machine

Mean Accuracy: 0.8580
Mean Precision: 0.4290
Mean Recall: 0.5000

Naive Bayes

Mean Accuracy: 0.8727
Mean Precision: 0.7717
Mean Recall: 0.6041

K-nearest Neighbours

Mean Accuracy: 0.8863
Mean Precision: 0.8292
Mean Recall: 0.6363

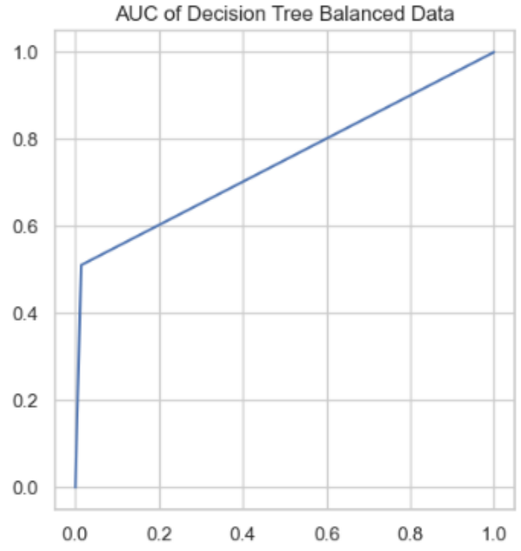
Neural Network Classifier

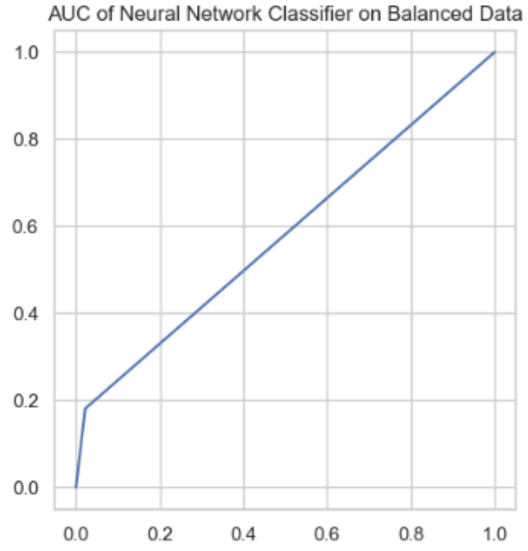
Mean Accuracy: 0.9205
Mean Precision: 0.8726
Mean Recall: 0.7752

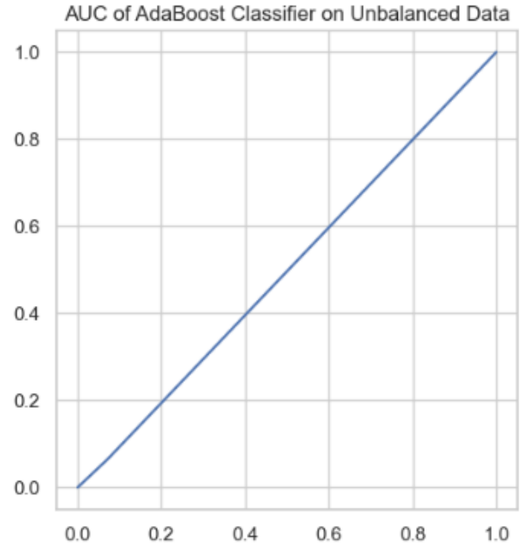
Avaliação e escolha do modelo

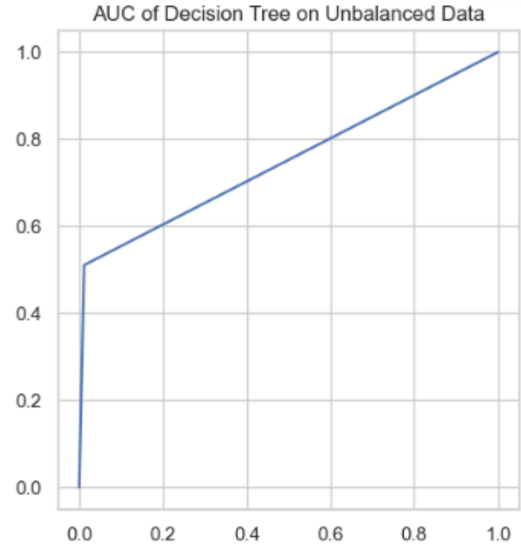
O recall foi a métrica utilizada para escolher dentre os modelos testados o que melhor se adequa ao nossso problema.
A ênfase do problema está na minimização de falsos negativos, ou seja, minimizar o errar na previsão ao prever que o cliente irá ficar quando na realidade sai.

Avaliação Final - AUC









Escolha do Modelo

Desequilibrados

Decision Trees

Mean Recall: 0.7412

AUC score: 0.7478463222453396

Neural Networks

Mean Recall: 0.6228

AUC score: 0.5795588284743938

Equilibrados

Decision Trees

Mean Recall: 0.7427

AUC score: 0.7495884825240852

AdaBoost

Mean Recall: 0.6452

AUC score: 0.6551274659714738

Conclusão

O melhor modelo para este problema de negócio em específico é o Decision Tree, uma vez que apresentou um melhor AUC e recalll tanto para o dataset equilibrado como para o desequilibrado.

Obrigada!

Inês Silva, Maria Miguel Ribeiro & Renatha Vieira