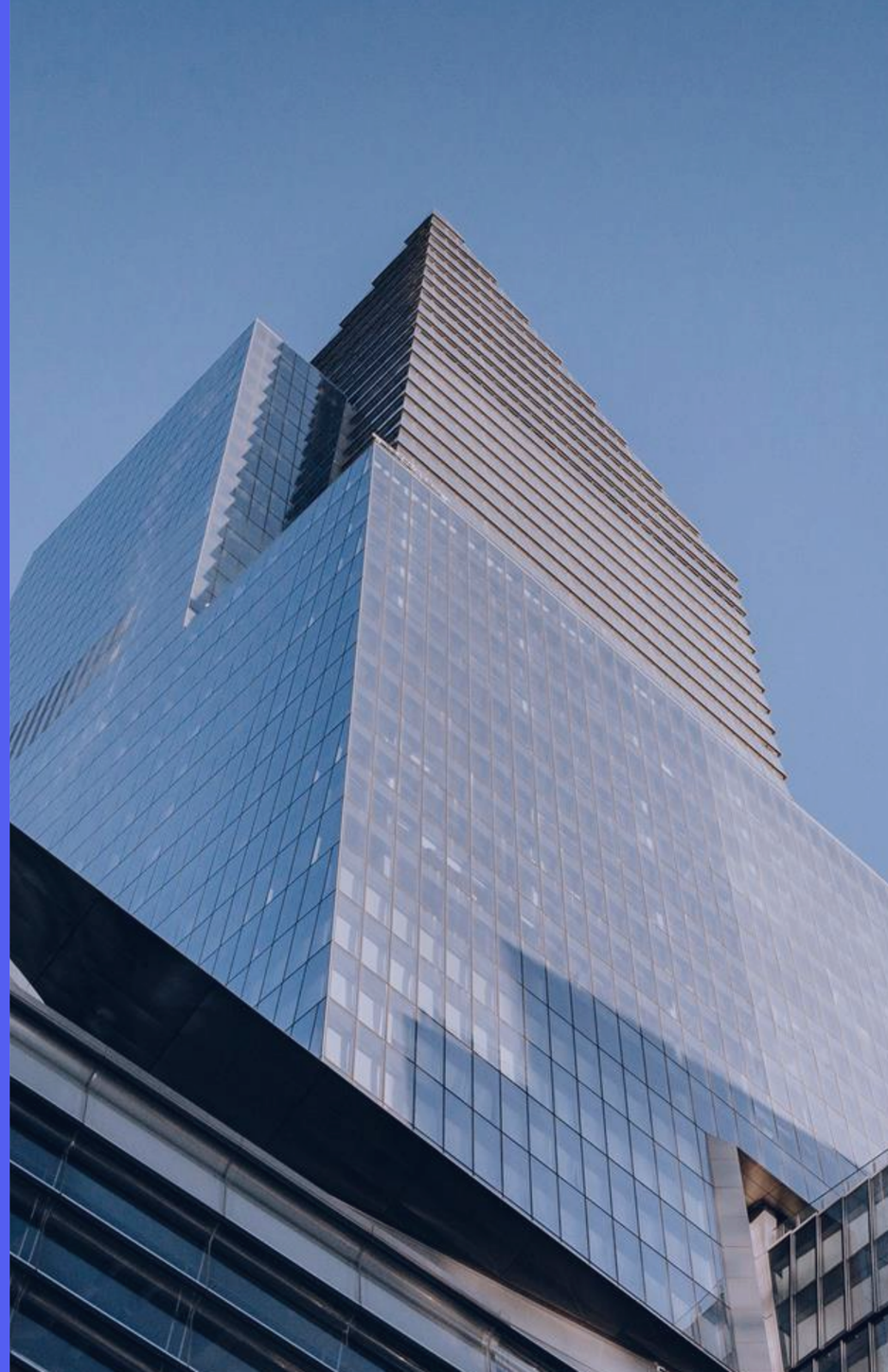
QuantiDados

PREVISÃO DE EVASÃO BANCÁRIA

INTEGRANTES DO GRUPO:

Eliamara Souza da Silva
Gustavo Rodrigues Melo
Lucas Feliciano da Silva
Roussian Di Ramos Alves Gaioso
Thais Moreira da Silva



DEFINIÇÃO DO PROBLEMA

As **instituições financeiras** são um dos principais **pilares da econômia** e do **mercado financeiro.**

Um dos principais produtos e serviços oferecidos por essas insituições é o **Cartão de Crédito.**

O estudo acerca da métrica que indica a **taxa de cancelamento (churn)** de cartões de crédito é fundamental para que a instituição **continue melhorando a qualidade dos seus serviços se mantendo no mercado.**

BASE DE DADOS

Dados

Amostra de dados bancários de cerca de 10 mil clientes

Formato

Dados tabulares disponivel como um arquivo csv

Features

21 features no total (categóricas e numéricas)

Não há dados duplicados

Não há dados faltantes

FEATURES

CATEGÓRICAS
(06 features)

- Gender
- Education Level
- Marital Status
- Income Category
- Card Category
- Attrition Flag

NUMÉRICAS
(15 features)

- CLIENTNUM
- Customer Age
- Dependent Count
- Months on book
- Total Relationship Count
- Months Inactive 12 Mon
- Contacts Count 12 Mon
- Credit Limit
- Total Revolving Bal
- AVG Open to By
- Total Amt Chng Q4 Q1
- Total Trans Amt
- Total Trans Ct
- Total Ct Chng Q4 Q1
- AVG Utilization Ratio

ANÁLISE EXPLORATÓRIA

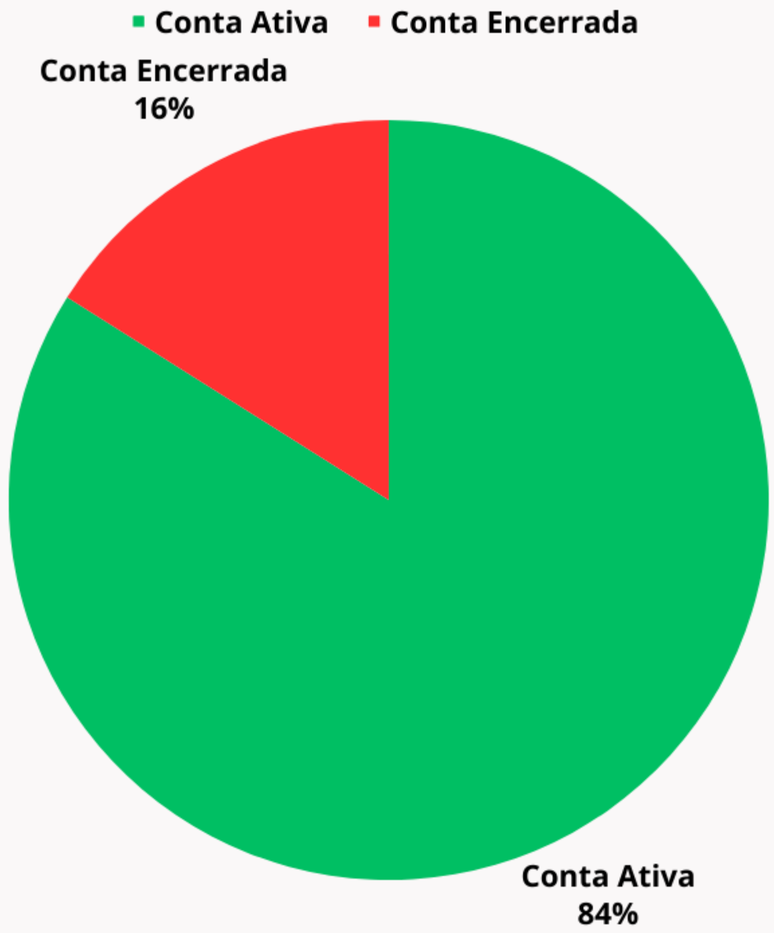
Target: Attrition Flag (binária)

- Conta Ativa = 0
- Conta Encerrada = 1

Problema de Classificação Binária

Dados desbalanceados:

- Contas Ativa = 84% (6.801 clientes)
- Contas Encerradas = 16% (1.300 clientes)
- **<u>Necessário balanceamento</u>**



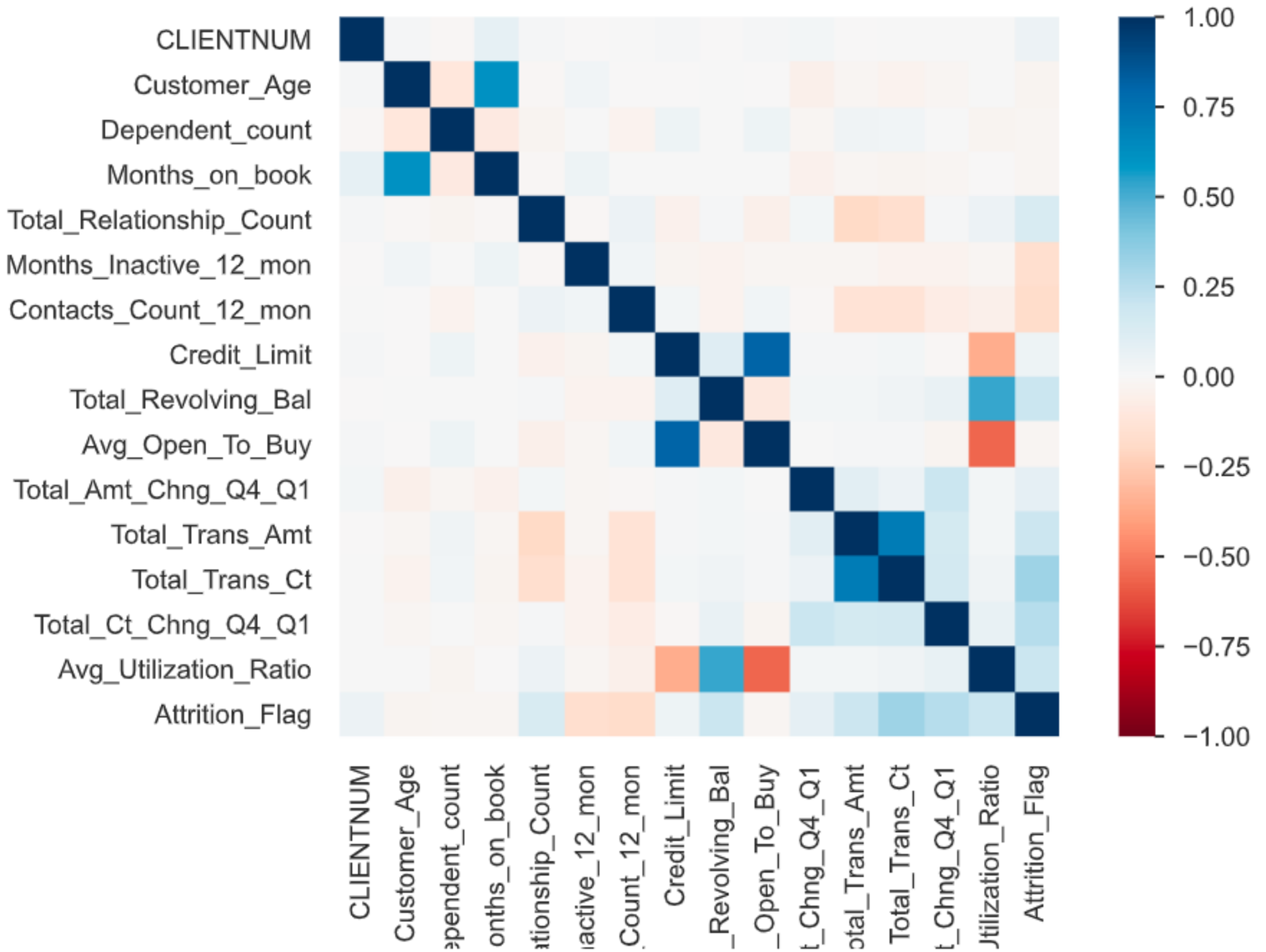
MATRIZ DE CORRELAÇÃO

Features Numéricas

Coeficiente de Correlação de Kendall

Altamente Correlacionadas:

- Customer_Age e Months_on_book
- Credit_Limit e Avg_Open_to_Buy
- Total_Revolving_Bal e Avg_Utilization_Ration
- Total_Trans_Amt e Total_Trans_Ct
- Credit_Limit e Avg_Utilization_Ratio
- Total_Revolving_Bal e Avg_Open_to_Buy



Distribuição dos salários por gênero

Distribuição dos clientes e as categorias de salários por gênero

ANÁLISE EXPLORATÓRIA

Salário x Gênero

Principais Insights:

- Diferença entre gênero nas categorias salariais
- Maior número de mulheres ganhando até $40K
- Maior numero de homens ganhando entre $60K-$80K e $80K-$120K

ANÁLISE EXPLORATÓRIA

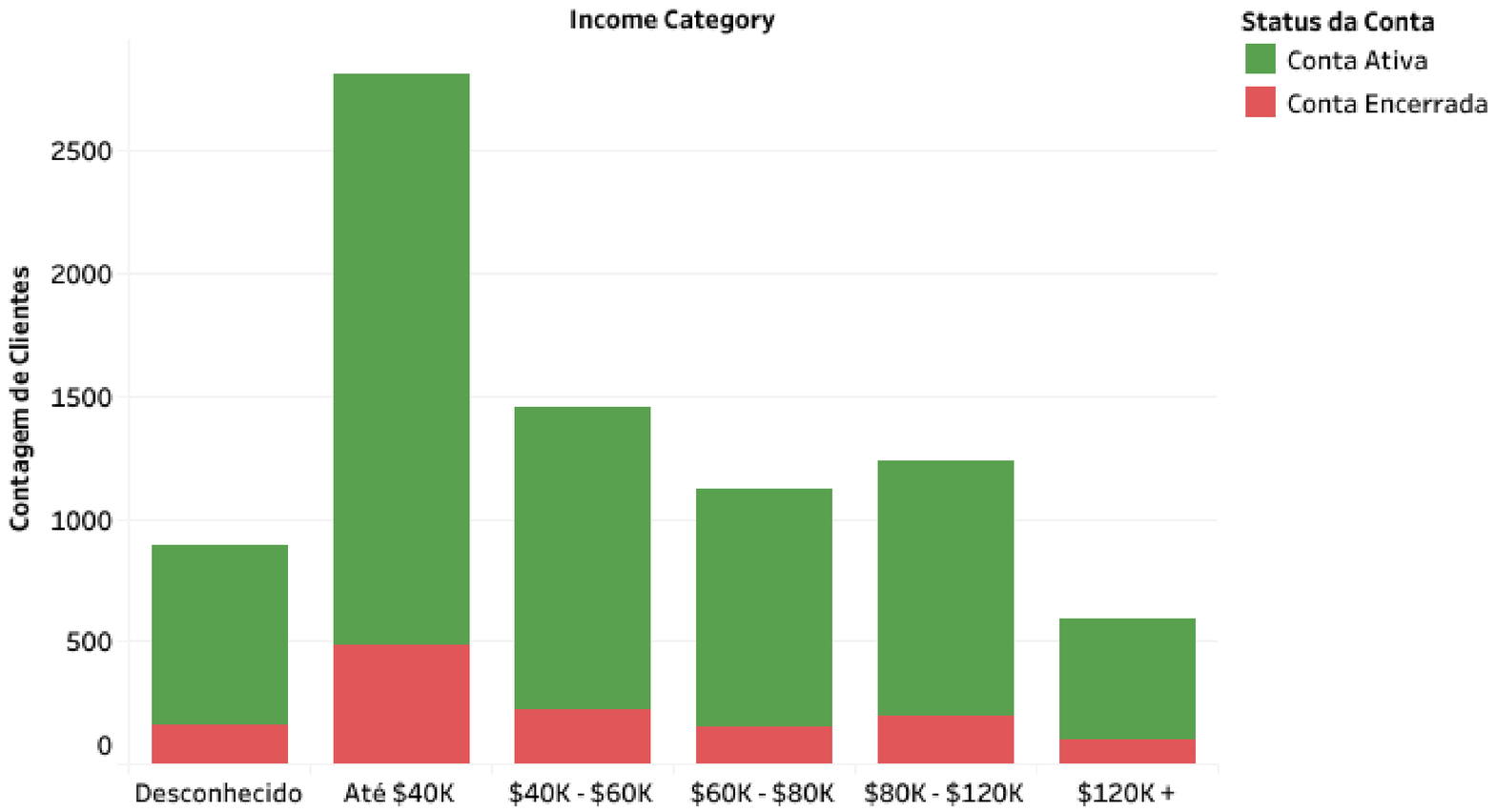
Salário X Status da Conta

Principais Insights:

- Dentre as contas encerradas, maior número entre clientes que ganham até $40K
- Soma dos clientes ativos que ganham acima de $40K ultrapassa aqueles que ganham até $40K.

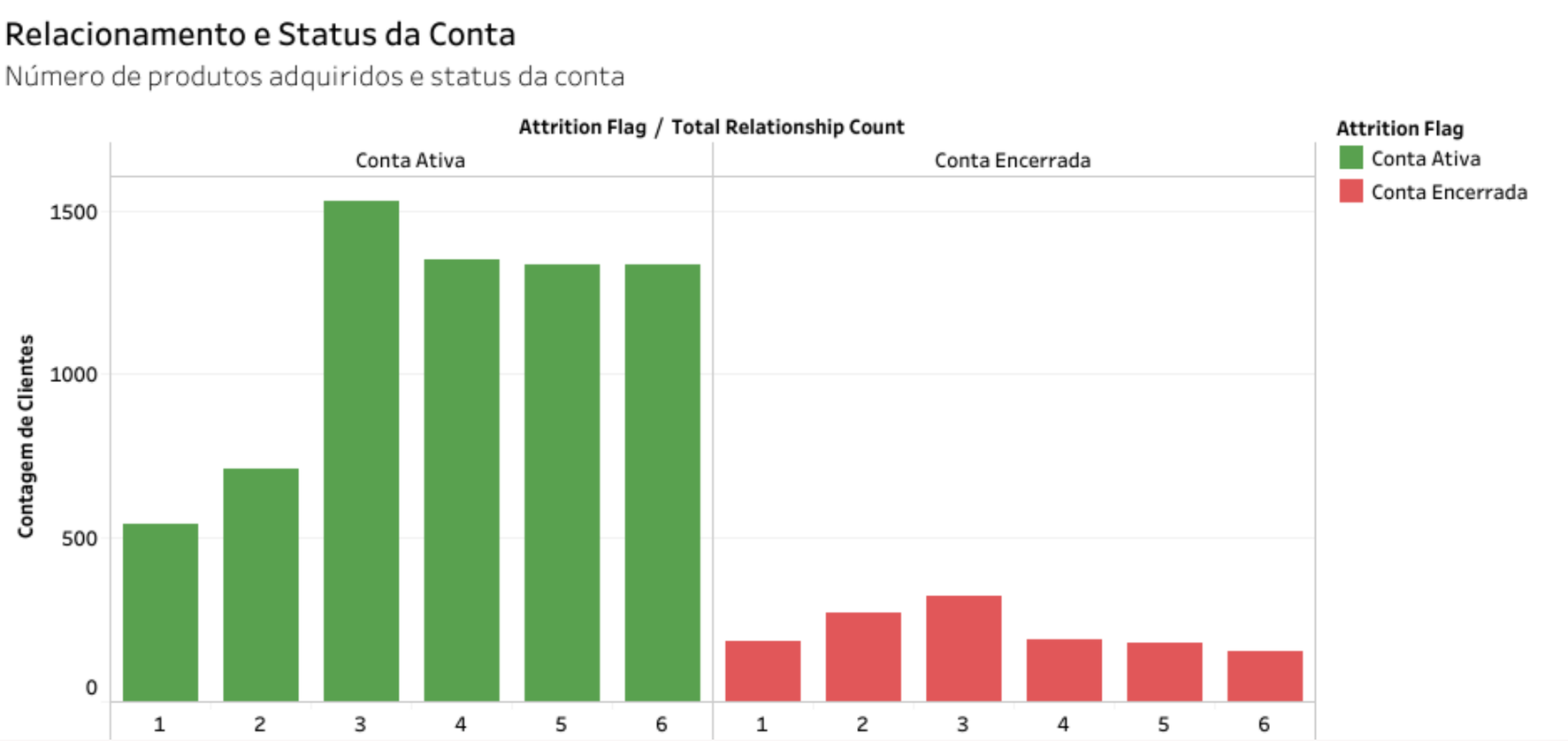
Distribuição do Salário e Status das Contas

Distribuição do salário dos clientes em relação a contas ativas e encerradas



ANÁLISE EXPLORATÓRIA

Produtos X Status da Conta



Principais Insights:

- O **maior número** de **contas ativas** são de clientes que consomem de **3 a 6 produtos** bancários.
- O **maior número** de **contas encerradas** são de clientes que consomem de **2 a 3 produtos** bancários.

PRÉ PROCESSAMENTO

#1

ONE HOT ENCODING
Variáveis Categóricas

#2

PADRONIZAÇÃO
Features Numéricas

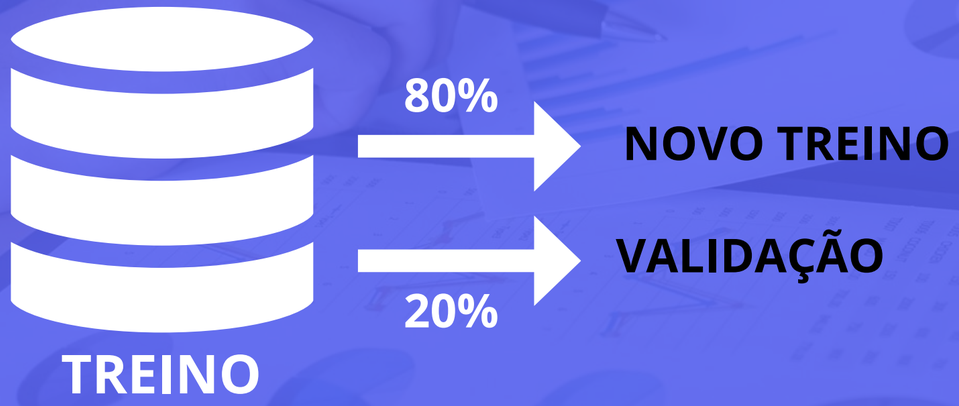
#3

TRATAMENTO
Target substituição
de texto por número

#4

EXCLUSÃO
Variável CLIENTNUM
(Identificador do Cliente)

PESOS E SEPARAÇÃO DOS CONJUNTOS



Ponderação de Pesos devido ao desbalanceamento das classes.

Pesos utilizados:

- Attrited Customer: 3.115769230769230 6
- Existing Customer: 0.595574180267607 7

MODELO BASELINE

MODELOS

- Regressão Logística
- Máquina de vetores-suporte para classificação
- Floresta Aleatória

PARÂMETROS

- Utilização dos parâmetros padrão.
- **class_weights** utilizado os pesos ponderados para variável target.
- **random_state** utilizado valor 0.

MÉTRICAS DE AVALIAÇÃO

Principal:

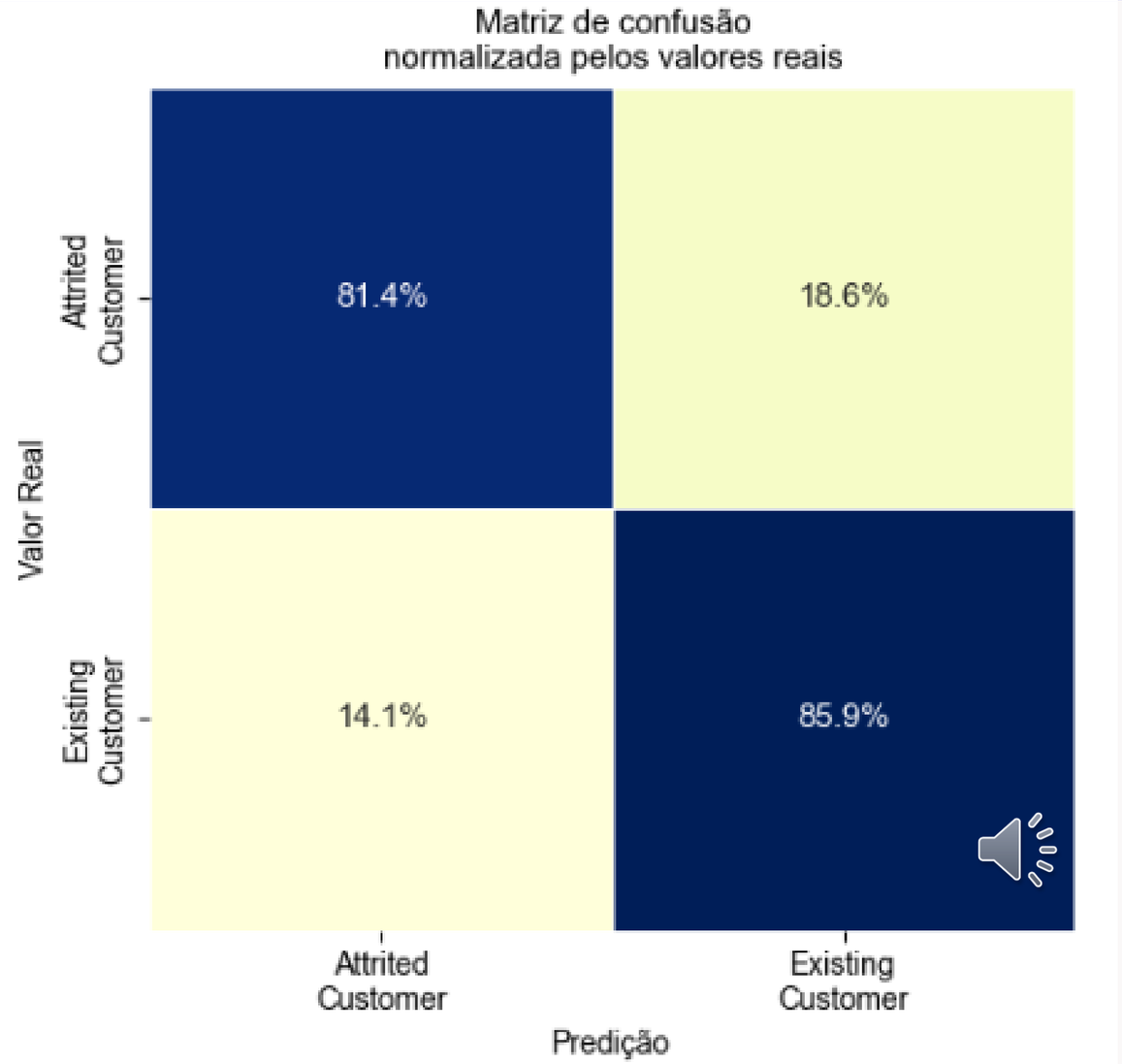
- Acurácia Balanceada

Outras:

- Acurácia
- F1 Score
- Matriz de Confusão
- RoC AuC

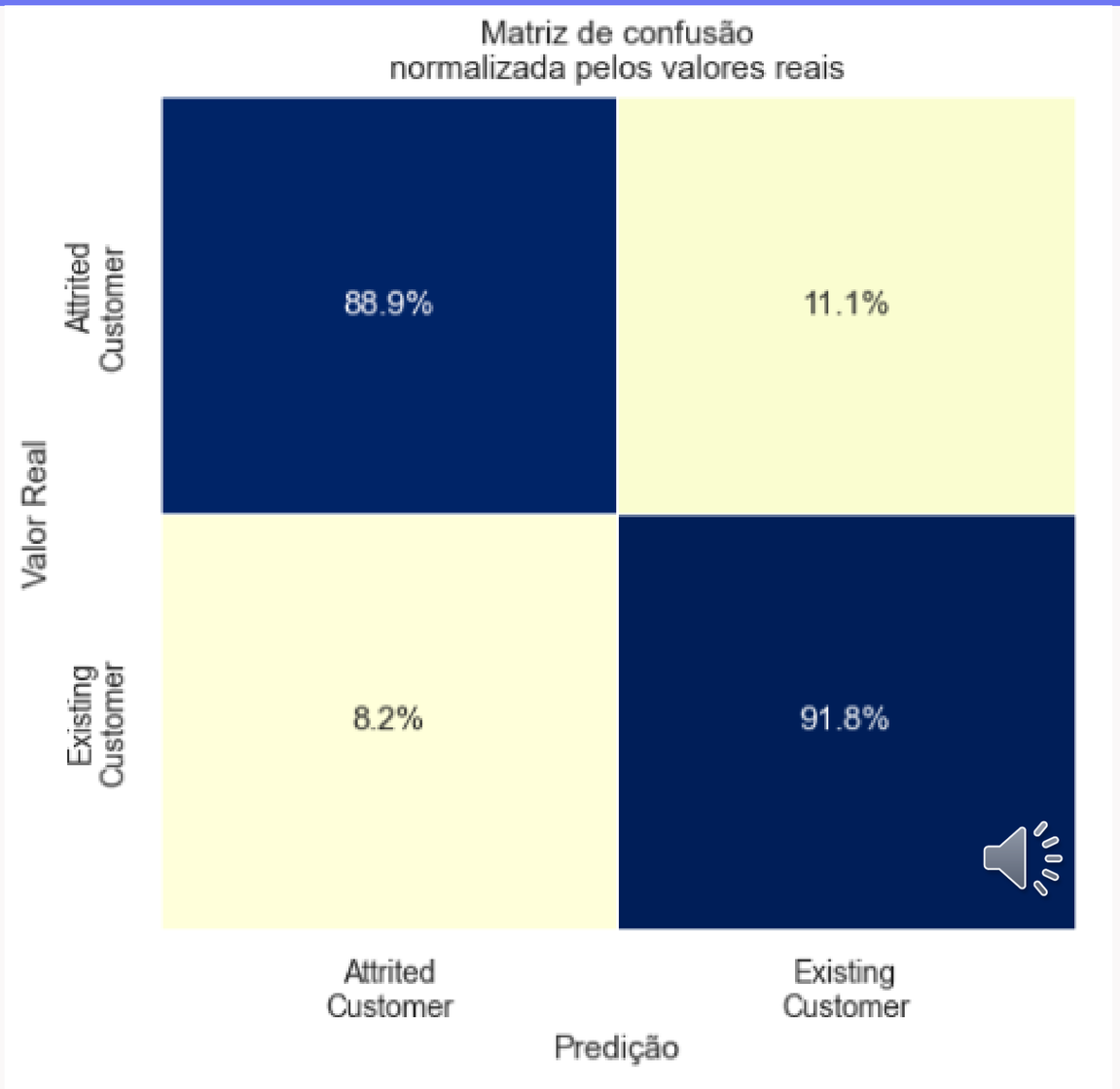
REGRESSÃO LOGÍSTICA

	VALIDAÇÃO
Acurácia Balanceada	0,8365
Acurácia	0,8519
F1 Score	0,9073
RoC AuC	0,8365



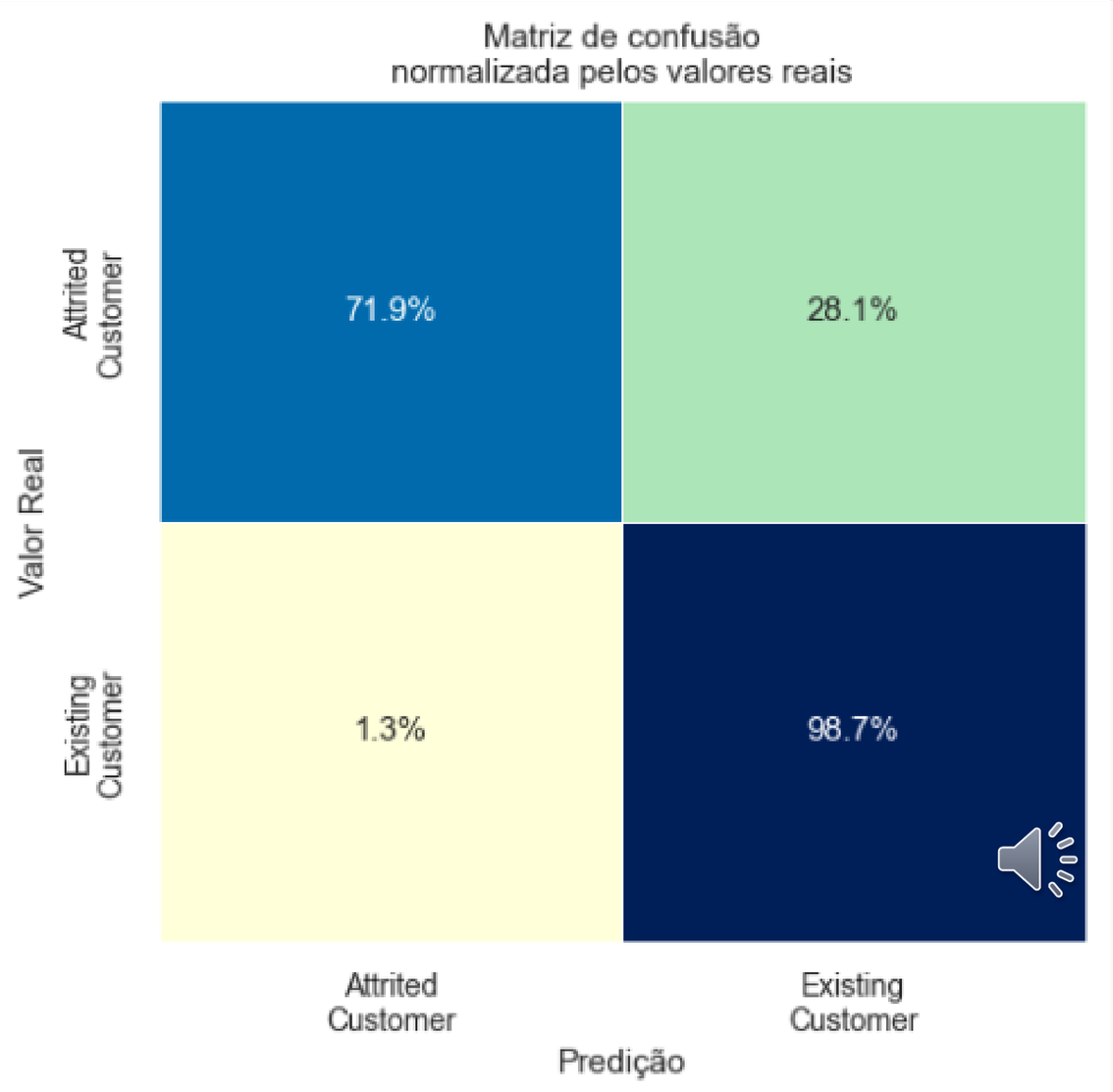
MÁQUINA DE VETOR-SUPORTE

	VALIDAÇÃO
Acurácia Balanceada	0,9037
Acurácia	0,9136
F1 Score	0,9472
RoC AuC	0,9037



FLORESTA ALEATÓRIA

	VALIDAÇÃO
Acurácia Balanceada	0,8531
Acurácia	0,9450
F1 Score	0,9680
RoC AuC	0,8531



BASELINES

Modelos	Acurácia Balanceada	Acurácia	F1 Score	RoU AuC
Regressão Logística	0,8365	0,8519	0,9073	0,8365
Máquina de Vetor-Suporte	0,9037	0,9136	0,9472	0,9037
Floresta Aleatória	0,8531	0,9450	0,9680	0,8531

- O modelo **SVM** acertou cerca de **89% da classe 0 “Attrited Customer”.**
- O modelo **Floresta Aleatória** acertou cerca de **99% da classe 1 “Existent Customer”.**
- O caminho escolhido foi testar **modelos mais robustos baseados em árvore de decisão.**

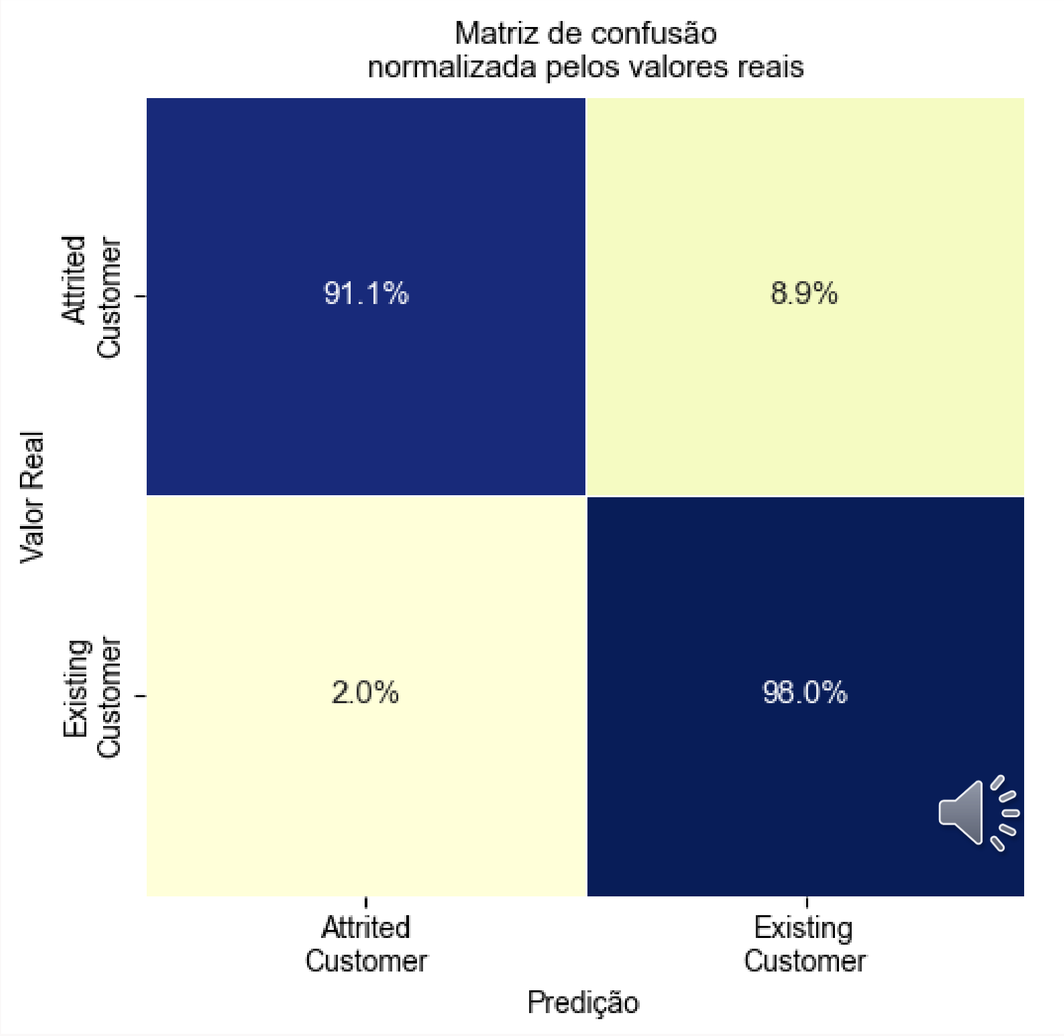
MODELOS TREINADOS

Modelos	Acurácia Balanceada	Acurácia	F1 Score	RoU AuC
SVM - Baseline	0,9037	0,9136	0,9472	0,9037
Regressão Logística + GridSearch	0,8497	0,8484	0,9037	0,9255
SVM + GridSearch	0,9029	0,9121	0,9460	0,9643
Floresta aleatória + GridSearch	0,9125	0,9315	0,9584	0,9726
XGBoost	0,9558	0,9635	0,9780	0,9558
CatBoost	0,9634	0,9699	0,9819	0,9634
Voting Ensemble	0,9429	0,9517	0,9708	0,9429
Stacking Ensemble	**0,9722**	**0,9659**	**0,9793**	**0,9722**

STACKING TODOS MODELOS

Modelo Escolhido

	SVM BASELINE	STACKING MODELO FINAL	Variação
Acurácia Balanceada	0,9148	0,9457	**+3,38%**
Acurácia	0,9171	0,9689	**+5,65%**
F1 Score	0,9489	0,9814	**+3,43%**
RoC AuC	0,9148	0,9457	**+3,38%**



QuantiDados

OBRIGADO!