Predição de Evasão Bancária

Grupo QuantiDados

Definição do problema

As instituições financeiras são um dos pilares do mercado financeiro, oferecendo diversas linhas de crédito a seus clientes. Dentro a grande gama de produtos que as instituições oferecem, uma das mais comuns e mais utilizadas é o cartão de crédito.

Para que as instituições consigam sua sustentabilidade no mercado, é necessário que estas invistam cada vez mais em estudos de métricas que auxiliem na continuidade dos bons resultados das instituições. Com isso, o estudo acerca da métrica que indica a taxa de cancelamento (churn) de cartões de crédito é fundamental para que a instituição continue melhorando cada vez mais a qualidade dos seus serviços, elevando o grau de satisfação dos seus clientes, reduzindo a evasão dos mesmos e, consequentemente, melhorando ou aumentando o seu faturamento.

Desta forma a Predição de Evasão Bancária do produto cartão de crédito pode ser encarada como um desafio de classificação, onde os modelos são treinados para distinguir dentre os clientes propensos a abandonar seus cartões e aqueles mais propensos a permanecer. Desta maneira, a instituição poderá prever e atuar de forma proativa, reduzindo suas taxas e melhorando seus resultados.

Base de Dados

Características

A base de dados para o estudo foi disponibilizada no formato “valores separados por vírgula” (.csv), dividida entre base de treinamento e de testes, juntas elas contém cerca de 10.000 amostras e 21 variáveis. Nesta estão contidas informações de clientes de uma instituição financeira, tais como: dados sociodemográficos, relacionamento com a instituição, limite do cartão de crédito, dentre outras.

Após a importação dos dados, observou-se que dentre as 21 variáveis, 6 são qualitativas (categóricas) e 15 quantitativas (numéricas). A partir disso, foi verificado se haviam valores nulos, duplicados ou inconsistentes. Como resultado, não foi encontrado nenhum aspectos que prejudicasse a análise.

Base de Dados

Variáveis

Das 21 variáveis presente nos conjuntos de dados, foram selecionadas 19 features:

Categóricas (6):

- Gender
- Education Level
- Marital Status
- Income Category
- Card Category
- Attrition Flag (**target**)

Numéricas (15):

- CLIENTNUM (**desconsiderada**)
- 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

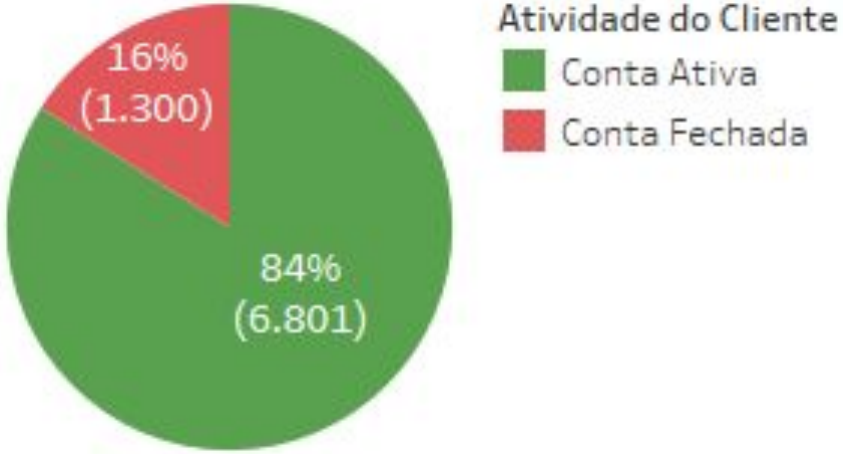
Análise Exploratória

Classes

TARGET: Attrition Flag

84% de Clientes Ativos

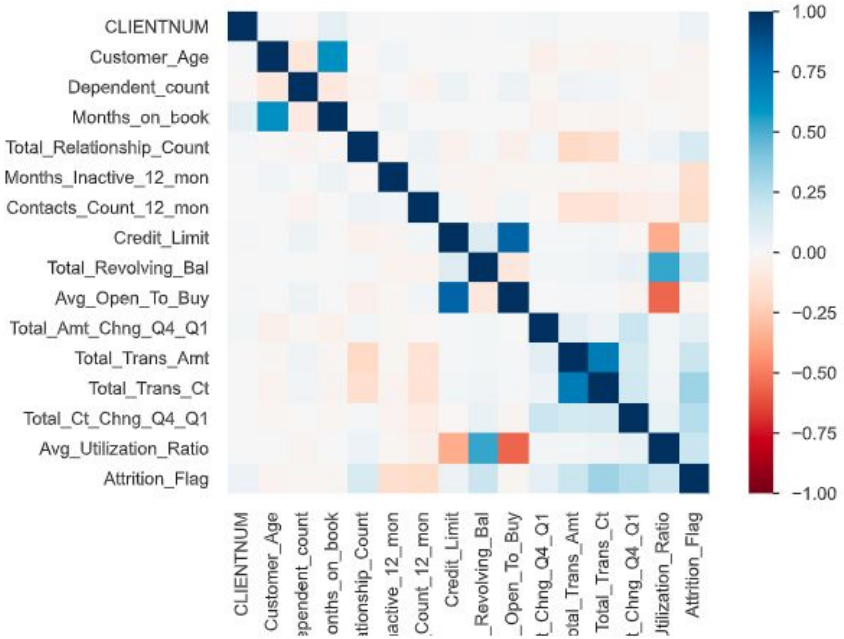
16% de Contas Encerradas



Com o auxílio do gráfico acima, verificamos que os dados não estão balanceados.

Análise Exploratória

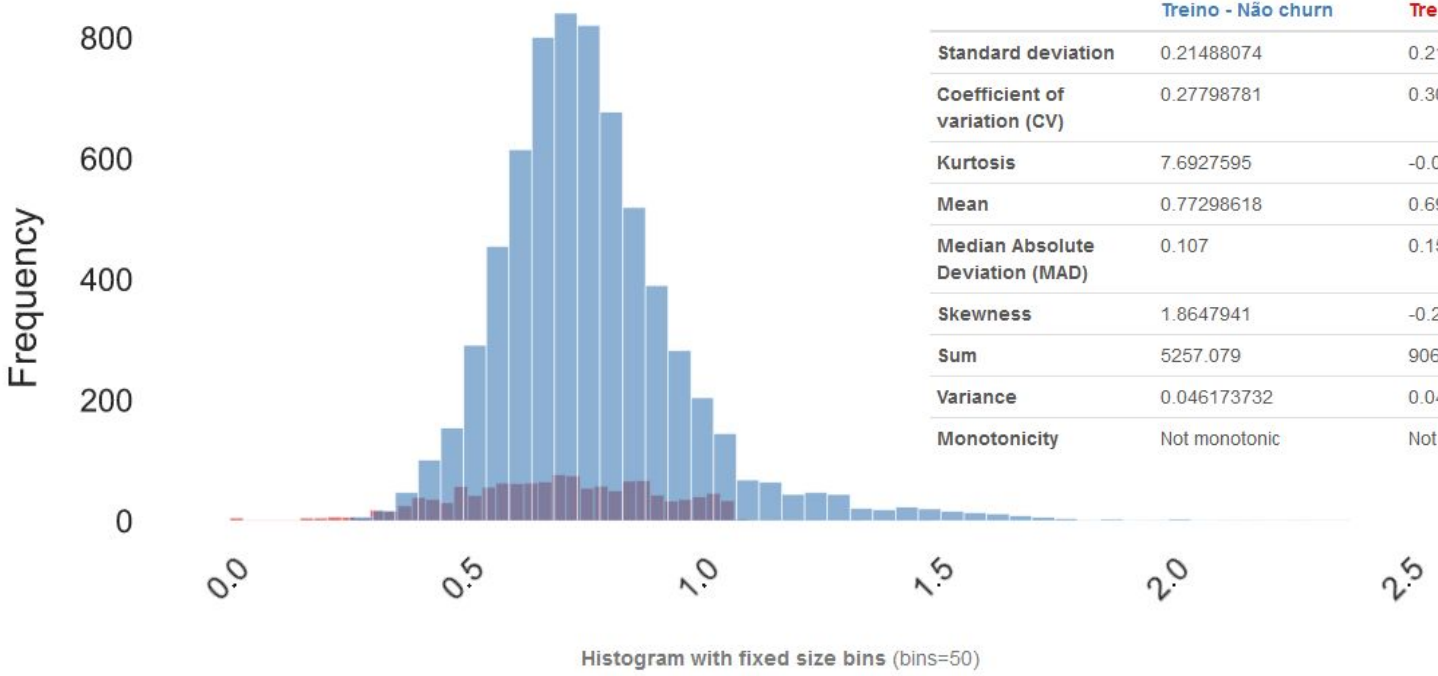
Matriz de Correlação



- Coeficiente de correlação Kendall (dados não paramétricos).

Análise Exploratória

Distribuição: Total Amt Chng Q4 Q1

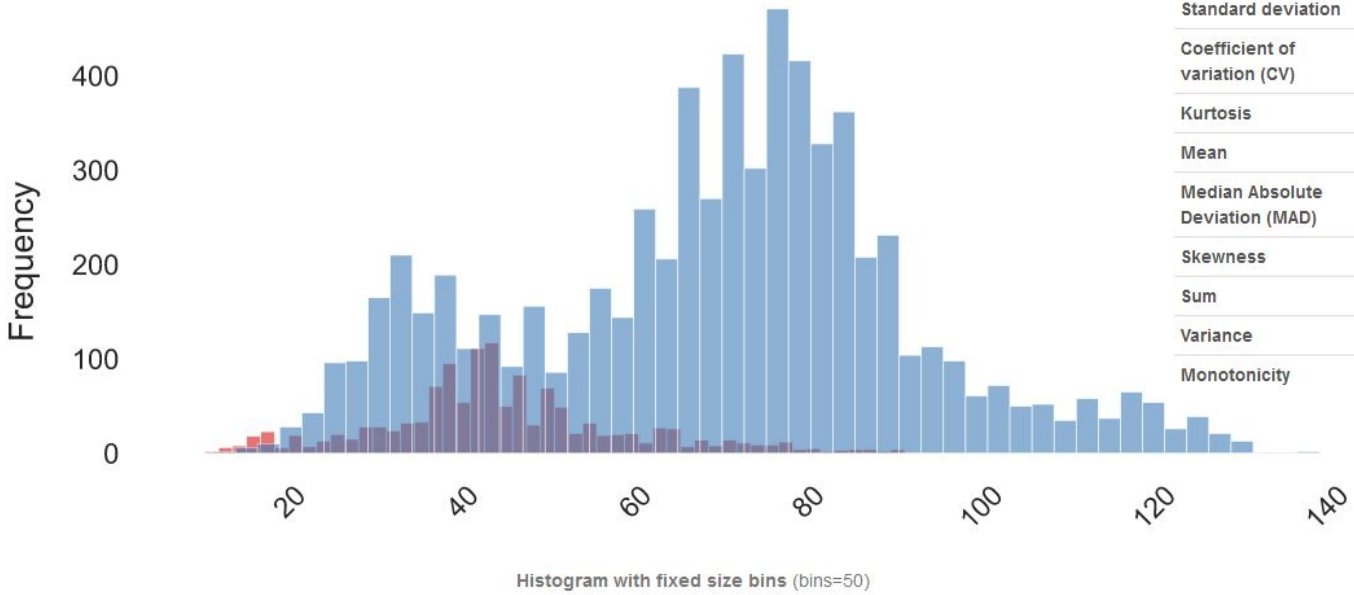


Descriptive statistics

	Análise Exploratória - Treino - Não churn	Análise Exploratória - Treino - Churn
Standard deviation	0.21488074	0.21489572
Coefficient of variation (CV)	0.27798781	0.30826998
Kurtosis	7.6927595	-0.039245567
Mean	0.77298618	0.69710231
Median Absolute Deviation (MAD)	0.107	0.154
Skewness	1.8647941	-0.20367439
Sum	5257.079	906.233
Variance	0.046173732	0.046180169
Monotonicity	Not monotonic	Not monotonic

Análise Exploratória

Distribuição: Total Trans Amt

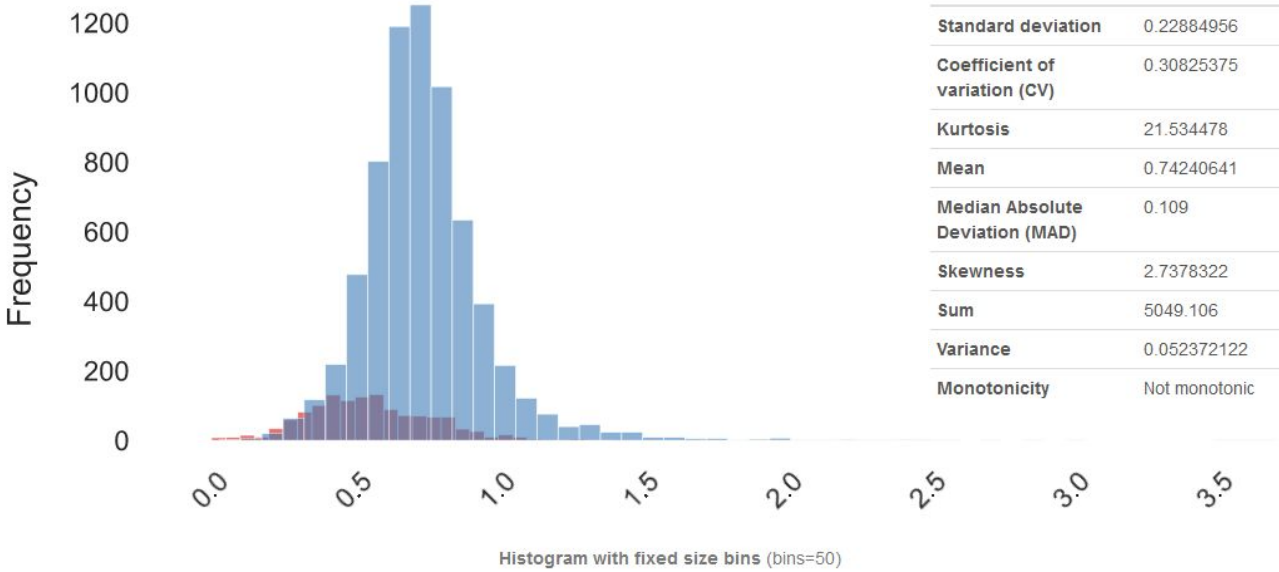


Descriptive statistics

	Análise Exploratória - Treino - Não churn	Análise Exploratória - Treino - Churn
Standard deviation	22.932922	14.3452
Coefficient of variation (CV)	0.33319406	0.32307889
Kurtosis	-0.19829295	0.63344033
Mean	68.827525	44.401538
Median Absolute Deviation (MAD)	13	7
Skewness	-0.0036227414	0.45684978
Sum	468096	57722
Variance	525.91892	205.78476
Monotonicity	Not monotonic	Not monotonic

Análise Exploratória

Distribuição: Total Ct Chng Q4 Q1



Descriptive statistics

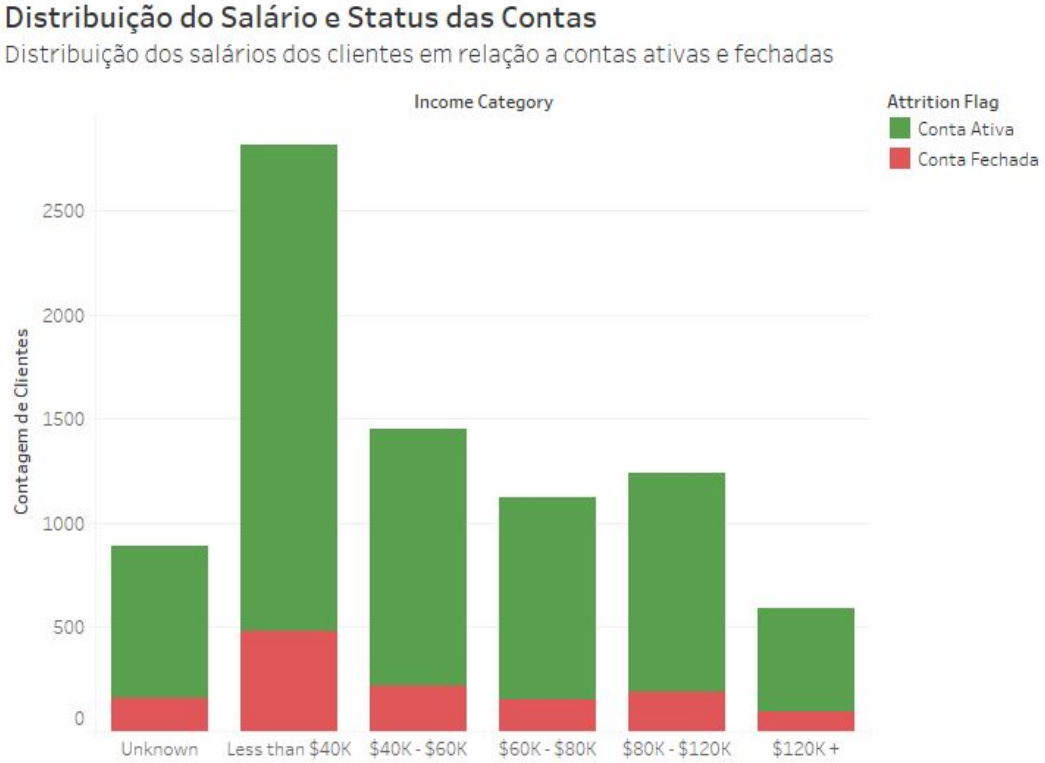
	Análise Exploratória - Treino - Não churn	Análise Exploratória - Treino - Churn
Standard deviation	0.22884956	0.23057677
Coefficient of variation (CV)	0.30825375	0.41618448
Kurtosis	21.534478	6.1824817
Mean	0.74240641	0.55402538
Median Absolute Deviation (MAD)	0.109	0.142
Skewness	2.7378322	1.1594444
Sum	5049.106	720.233
Variance	0.052372122	0.053165646
Monotonicity	Not monotonic	Not monotonic

Análise Exploratória

Distribuição dos salários por gênero

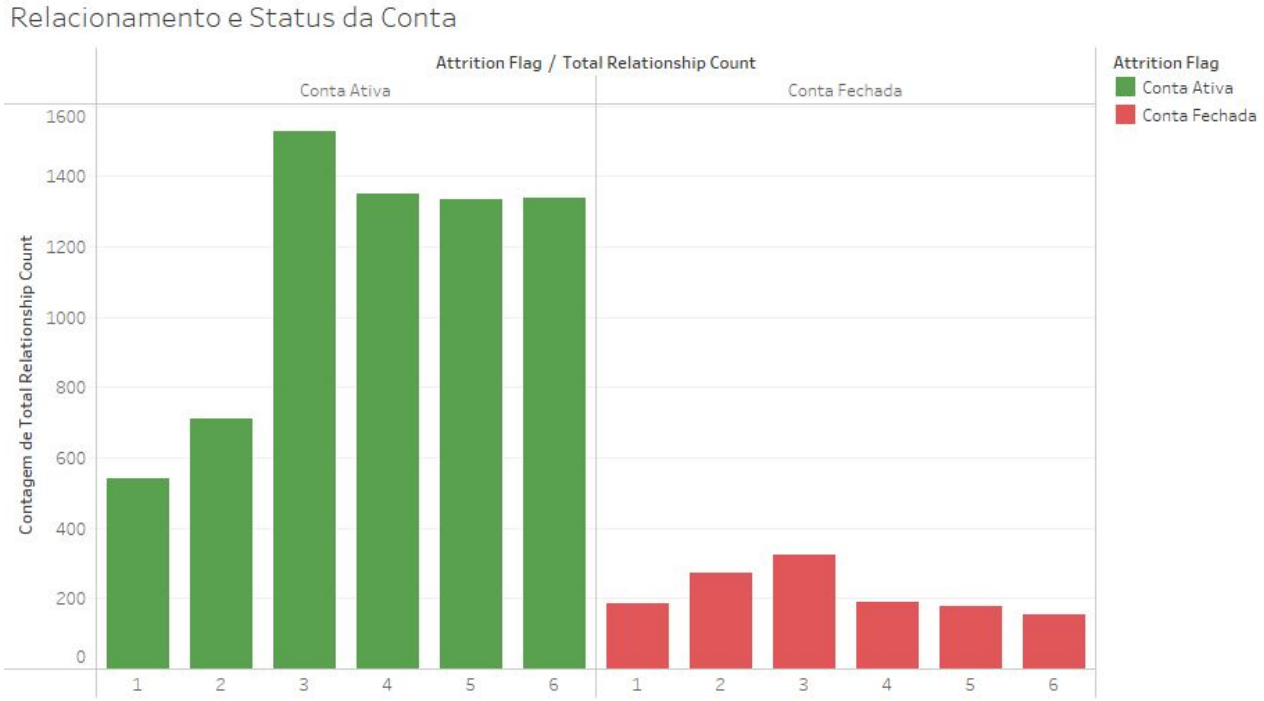
Análise Exploratória

Distribuição do salário e status das contas



Análise Exploratória

Relacionamento e status da conta



Pré Processamento

Features

As features e variável target passaram por alguns tratamentos antes de serem utilizadas nos modelos:

- Features Categóricas: One Hot Encoding
- Features Numéricas: Padronização
- Variável Target: Substituição do texto das classes por variável numérica
- Exclusão da variável *CLIENTNUM*

Pré Processamento

Pesos e separação dos conjuntos

O conjunto de dados de treino foi separado em dois conjuntos: treino (80%) e validação (20%). Também foram calculados os pesos de cada uma das classes, uma vez que as classes estão desbalanceadas, utilizamos a ponderação por pesos, para melhorar a performance ao treinar os modelos.

Os seguintes pesos foram utilizados:

- Attrited Customer: 3.1157692307692306
- Existing Customer: 0.5955741802676077

Baseline

Modelos

- Foram escolhidos 3 modelos como baseline para esse problema:
  - Regressão Logística
  - Máquina de vetores-suporte para classificação
  - Floresta Aleatória
- Foram utilizados os valores padrão para cada um dos modelos baseline, com exceção do parâmetro class_weights, responsável por fazer o ponderamento das classes, e do parâmetro random_state, usado para tornar os experimentos reprodutíveis

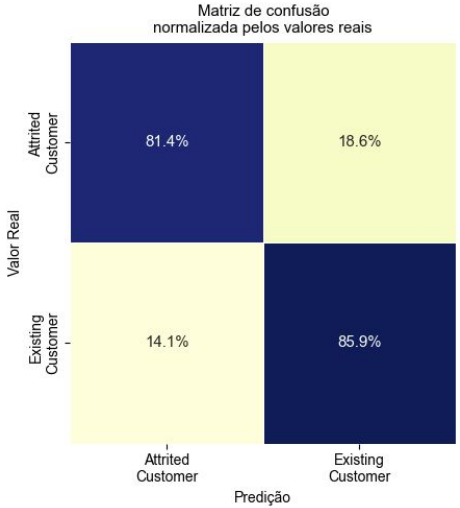
Baseline

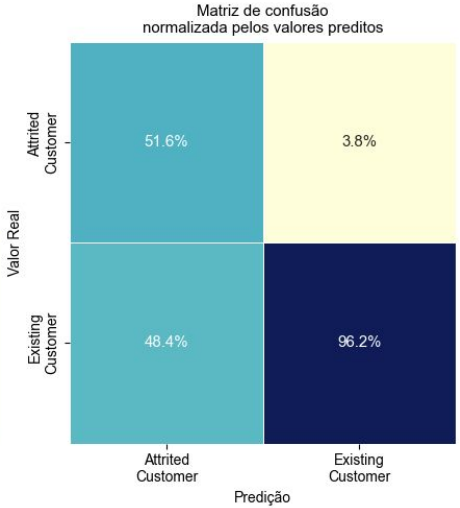
Regressão Logística

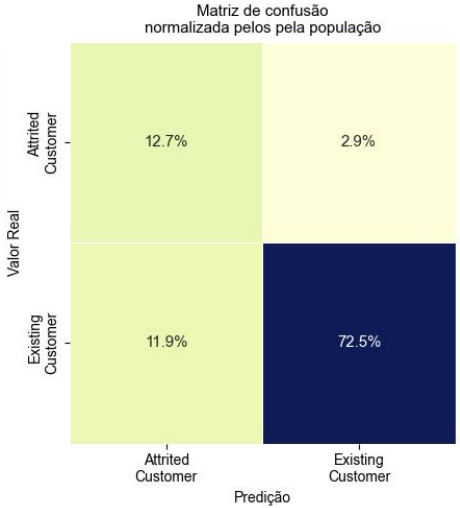
Modelo	Acurácia	Acurácia Balanceada	F1	RoC AuC
Regressão Logística	0,8519	0,8365	0,9073	0,8365

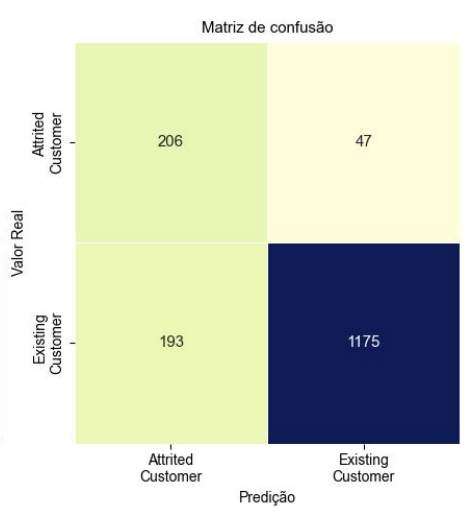
Baseline

Regressão Logística









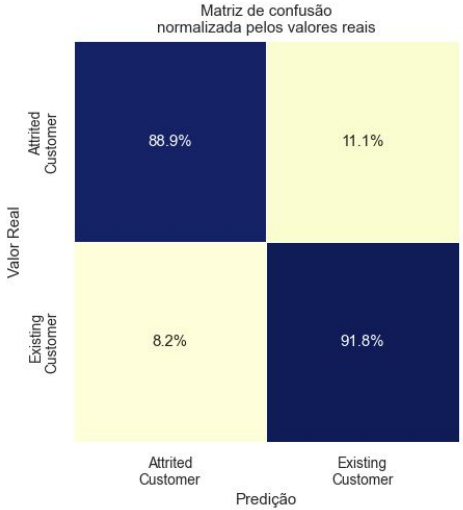
Baseline

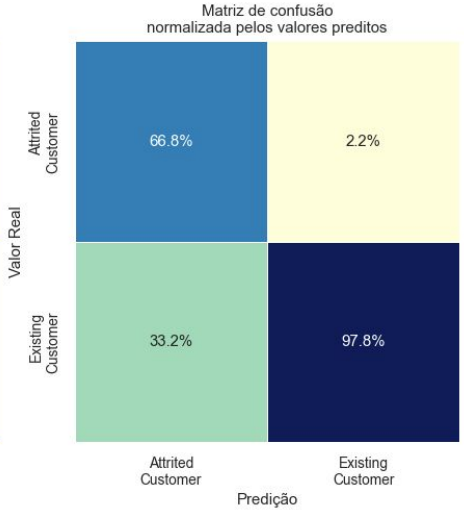
Máquina de vetor-suporte

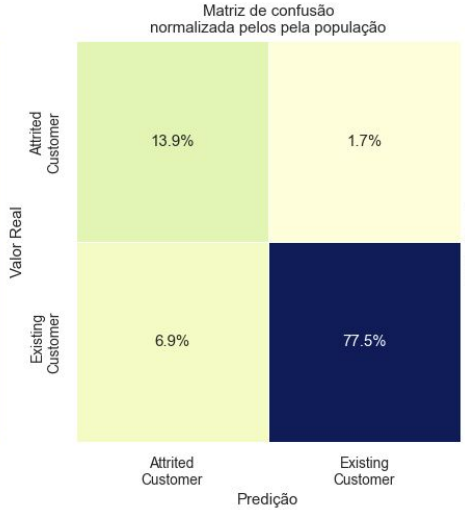
Modelo	Acurácia	Acurácia Balanceada	F1	RoC AuC
Regressão Logística	0,8519	0,8365	0,9073	0,8365
Máquina de vetor-suporte	0,9136	0,9037	0,9472	0,9037

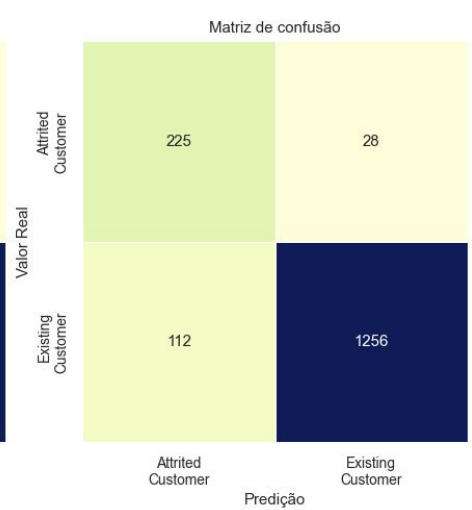
Baseline

Máquina de vetor-suporte









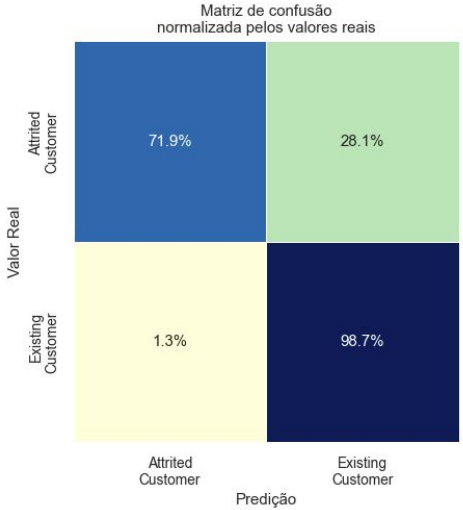
Baseline

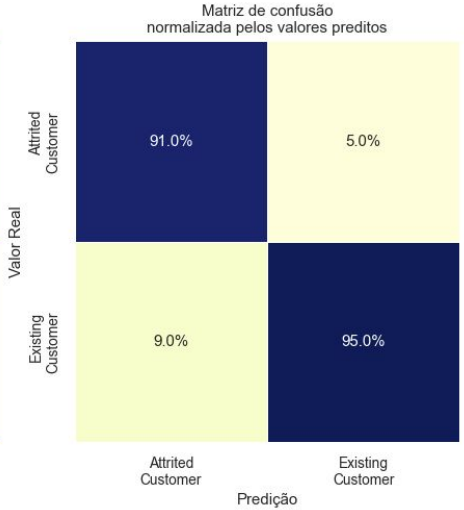
Floresta Aleatória

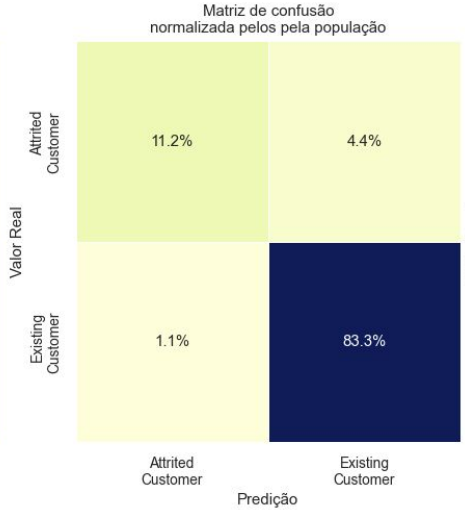
Modelo	Acurácia	Acurácia Balanceada	F1	RoC AuC
Regressão Logística	0,8519	0,8365	0,9073	0,8365
Máquina de vetor-suporte	0,9136	0,9037	0,9472	0,9037
Floresta Aleatória	0,9450	0,8531	0,9680	0,8531

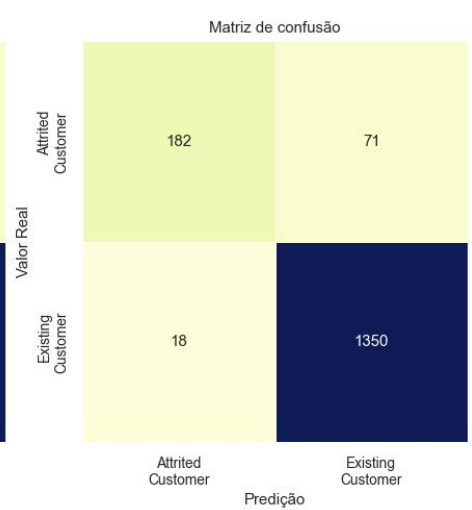
Baseline

Floresta Aleatória









Próximos Passos

- Busca de hiperparâmetros para cada um dos modelos Baseline através de Gridsearch e/ou Randomsearch
- Implementação de outras técnicas para tratar desbalanceamento
- Seleção do melhor modelo
- Avaliação do melhor modelo sobre o dataset de teste