Previsão de Evasão Bancária *- Grupo Quantidados*

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

Introdução

As instituições financeiras fazem parte do dia a dia das pessoas. Um dos principais pilares do sistema econômico em que estamos inseridos são as instituições financeiras, ou popularmente conhecidas como Bancos. Além disso, também é um dos pilares do mercado financeiro.

O sistema econômico atual já passou por diversas fases, a mais recente e transformadora pode ser entendida como o Capitalismo Informacional, em que a principal fonte de capital se tornou as informações, os dados. Sendo assim, para se manterem no mercado e entregarem um serviço de qualidade, as instituições financeiras têm investido cada vez mais em estudos e pesquisas a respeito do comportamento de seus clientes.

Um dos produtos mais consumidos dessas instituições é o Cartão de Crédito. Portanto, o objetivo deste trabalho é analisar e predizer o comportamento dos clientes de um banco em relação às taxas de cancelamento do cartão de crédito.

Análise Exploratória

Os dados analisados consistem em dados bancários de cerca de 10 mil clientes de uma instituição financeira. Esses dados estão estruturados de forma tabular em um arquivo csv. Nessa base encontramos 21 features entre categóricas e numéricas. Além disso, não há dados duplicados ou faltantes na base.

Analisando as 21 features apresentadas, temos 06 categóricas e 15 numéricas. Nas features categóricas nós temos informações qualitativas. Já nas features numéricas nós temos informações quantitativas.

Variáveis Categóricas:

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

Variáveis Numéricas:

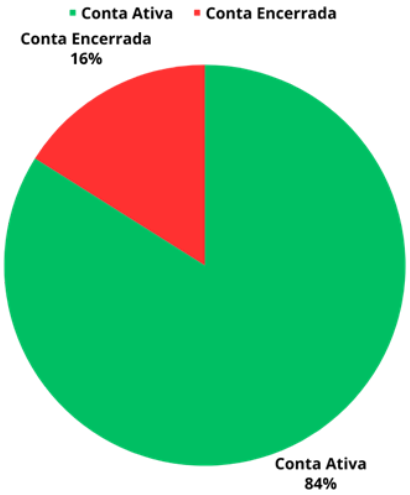
- CLIENTNUM
- Customer Age

- 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

A nossa variável alvo é a Attrition Flag, ela é uma variável binária, em que o valor 1 representa as contas encerradas e o valor 0 representa as contas ativas. Sendo assim, o nosso problema é um problema de classificação binária. O Gráfico 1 nos mostra a distribuição da variável Attrition Flag na base de dados. Em verde estão as contas ativas, já em vermelho estão as contas encerradas. Pode-se notar que há 84% de clientes ativos e somente 16% de contas encerradas. Sendo essa a nossa variável alvo, podemos concluir que a base de dados está desbalanceada.

Sendo assim, esse desbalanceamento vai se refletir ao longo da base de dados refletindo na análise comportamental dos clientes, uma vez que há um número maior de informações comportamentais a respeito de clientes que possuem contas ativas, do que aqueles que encerraram suas contas.

Gráfico 1 - Atividade do Cliente na Instituição

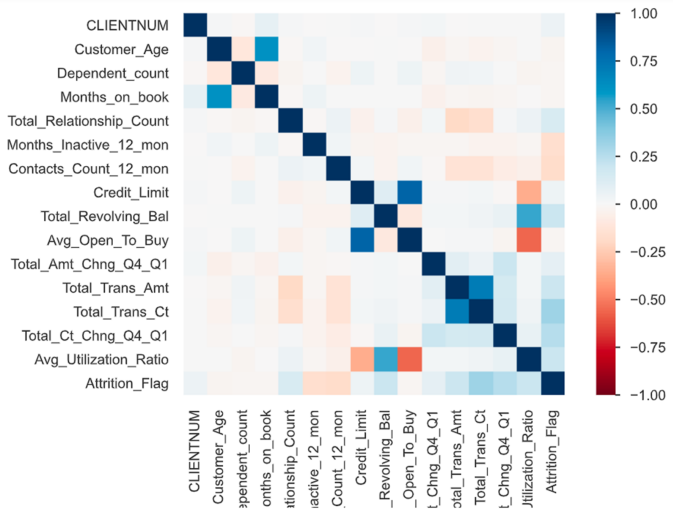


Para entender o comportamento da base de dados, foi calculada uma matriz de correlação entre as variáveis numéricas. Ela foi calculada através do coeficiente de Kendall uma vez que as variáveis existentes não possuem uma distribuição normal. A matriz se encontra no Gráfico 2. As cores se identificam da seguinte forma: quanto mais vermelho menos correlacionada e quanto mais azul, mais correlacionada.

Portanto, as variáveis mais correlacionadas são:

- 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

Gráfico 2 - Matriz de Correlação Variáveis Numéricas

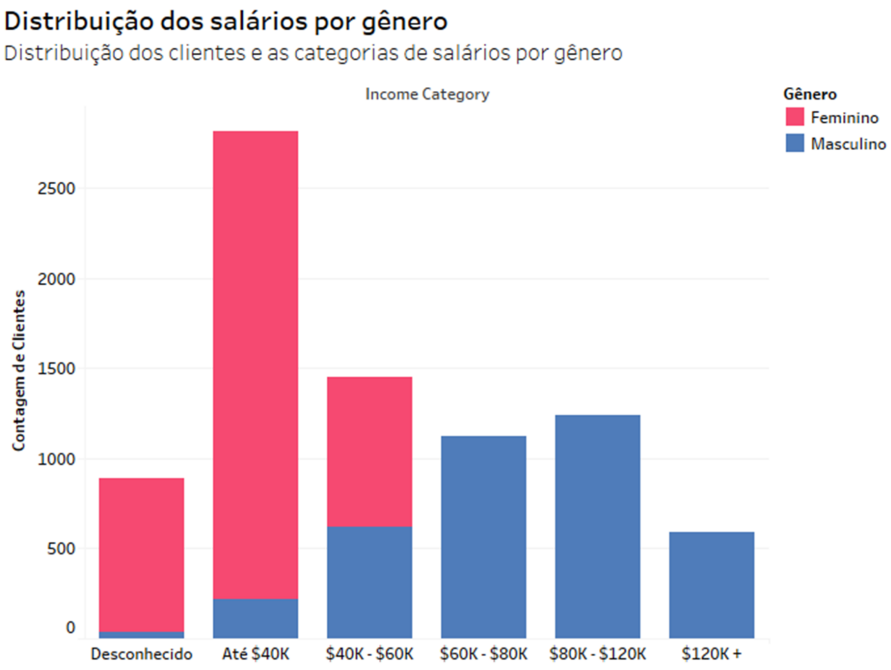


Seguindo com a análise comportamental da base de dados, foi possível extrair alguns insights relacionados às variáveis categóricas. São eles o comportamento entre salário e gênero, salário e status da conta e consumo de produtos bancários e status da conta. No Gráfico 3 é apresentado a distribuição do salário anual em dólares por gênero, a cor rosa representa o gênero feminino e a cor azul representa o gênero

masculino. Os salários estão organizados em ordem crescente. As mulheres se enquadram somente em 3 categorias salariais, sendo elas, “Desconhecido”, “Até $40K” e de” $40K a $60K”.

Em contrapartida, além de apresentar as mesmas categorias salariais que as mulheres, os homens ainda possuem outras três categorias de salários maiores, sendo elas, “de $60K a $80K”, “de $80K a $120K” e “mais de $120k”. Vale destacar que a maior concentração salarial entre as mulheres está na categoria de salário “Até $40K” enquanto os homens estão mais concentrados na categoria de “$80K a $120K”. O comportamento da base de dados reflete o comportamento da sociedade atual em que temos de forma recorrente homens ganhando salários maiores do que as mulheres.

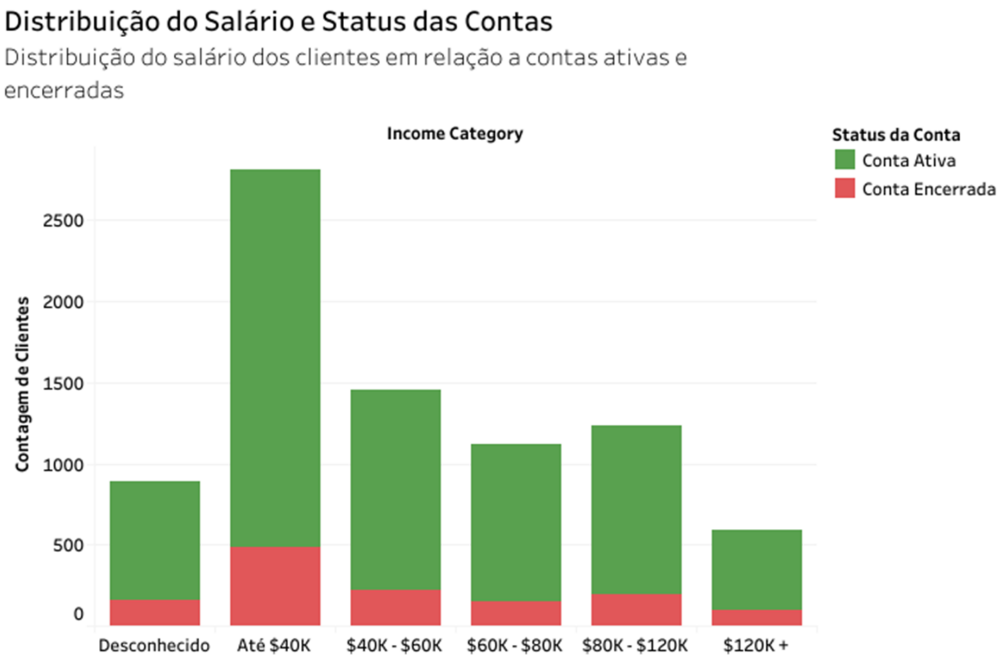
Gráfico 3 - Distribuição do salário por gênero



O segundo insight proposto diz respeito à distribuição do salário e o status da conta. No Gráfico 4, em vermelho temos as contas encerradas e em verde temos as contas ativas. Da mesma forma, as categorias de salário estão organizadas de forma crescente. É possível identificar que dentre as contas encerradas, temos uma concentração maior de clientes que ganham “até $40K”. Porém, também temos muitos

clientes ativos que ganham “até 40 mil dólares”, mas também podemos identificar um número considerável de clientes ativos nas categorias salariais maiores.

Gráfico 4 - Distribuição do Salário e o Status das Contas

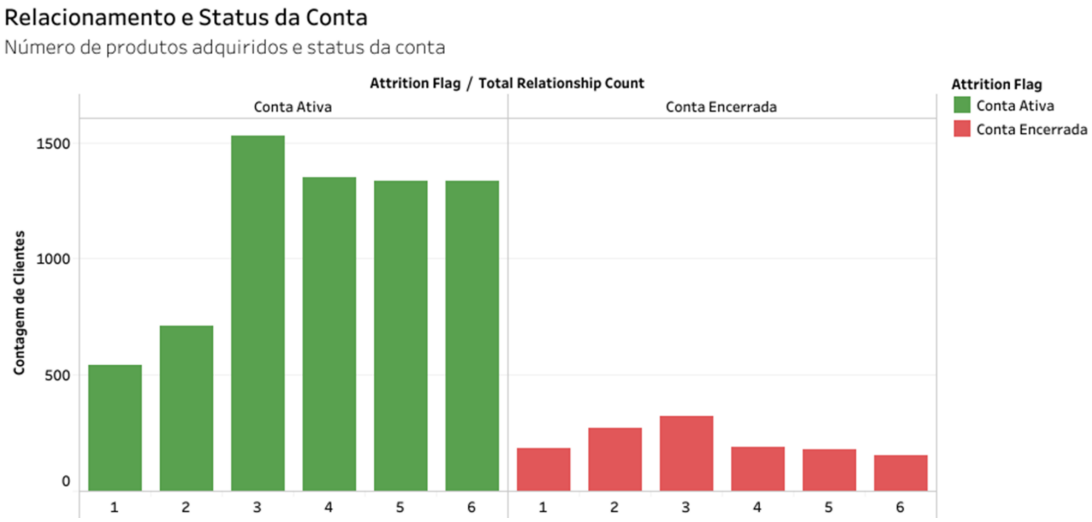


Por fim, o último insight trazido, apresentado no Gráfico 5 diz respeito ao número de produtos adquiridos e o status da conta. A esquerda em verde temos as contas ativas e as barras estão ordenadas em número de produtos adquiridos em ordem crescente.

Já à direita em vermelho temos as contas encerradas e da mesma forma o número de produtos adquiridos em ordem crescente. O maior número de contas ativas é de clientes que consomem de 3 a 6 produtos bancários. Já o maior número de contas encerradas é de clientes que consomem de 2 a 3 produtos bancários.

Pode-se dizer que aqueles que consomem mais produtos bancários se mantêm ativos no banco, enquanto aqueles que consomem menos produtos têm maior possibilidade de encerrarem suas contas.

Gráfico 5 - Relacionamento e o Status da Conta



Pré-Processamento

O pré-processamento do conjunto de dados contou com as seguintes etapas:

- One hot encoding das variáveis categóricas, onde categorias vira uma nova coluna onde a coluna da categoria correspondente se torna 1 e as outras colunas 0;
- Padronização das variáveis numéricas, onde a distribuição resultante tem média igual a 0 e variância unitária;
- Substituição do valor textual da variável target por número correspondente a classe;
- Exclusão da variável *CLIENT_NUM*, uma vez que essa variável era um identificador de cada um dos exemplos, e, portanto, não foi considerada relevante para o treinamento dos modelos.

Durante a etapa de análise exploratória foi verificado que as classes estavam desbalanceadas, portanto, utilizamos pesos para cada uma das classes que seria o imputado em cada um dos modelos estudados, os valores desses se encontram na tabela abaixo.

Classe	Número	Peso
Attrited Customer	0	3.1157692307692306
Existing Customer	1	0.5955741802676077

Também fizemos a separação do conjunto de treino em dois novos conjuntos, treino e validação em uma razão de 4 para 1 respectivamente. Em alguns modelos não utilizamos essa separação, mas utilizamos a validação cruzada com 5 e 3 dobras.

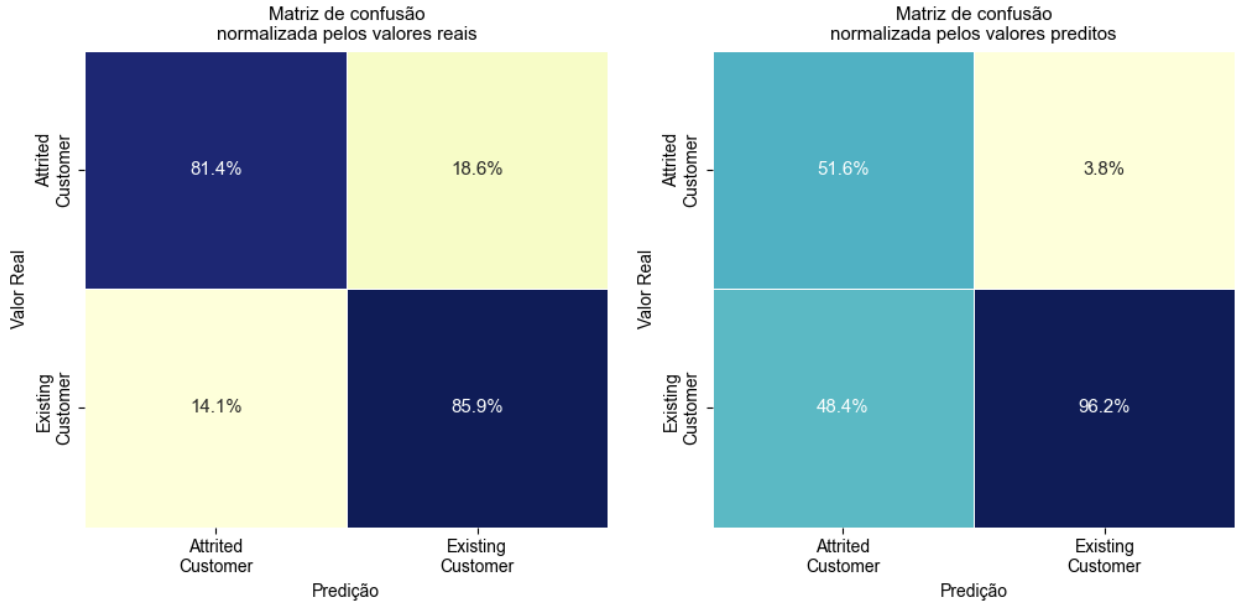
Baselines

Tomamos como baseline alguns modelos simples sem alteração dos hiperparâmetros. Os únicos hiperparâmetros definidos foram os pesos das classes (*class_weight*, ou *sample_weight*, a depender do modelo) e *random_state* (ou seed, a depender do modelo), este último parâmetro foi utilizado para manter os experimentos reprodutíveis por outras pessoas.

Os três modelos escolhidos como baseline foram: modelo de regressão logística; modelo de máquina de vetores suporte para classificação; e modelo de floresta aleatória. Para avaliar os modelos baselines utilizamos o conjunto de validação, utilizando a acurácia balanceada como principal métrica de avaliação entre os modelos, também observamos outras métricas como acurácia, score f-1, área sob a curva RoC x AuC e matriz de confusão.

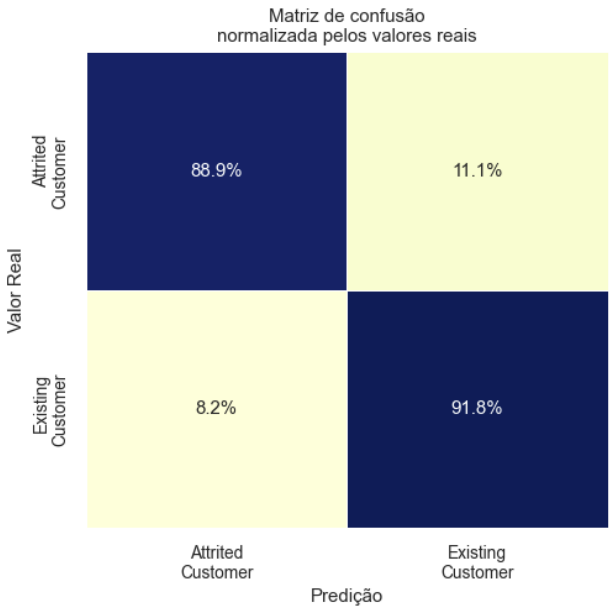
Baseline - Regressão Logística

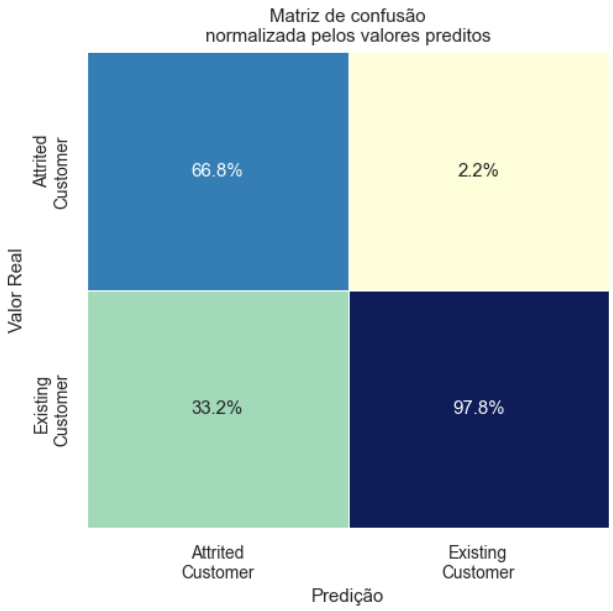
O modelo de regressão logística apresentou um resultado satisfatório para ambas as classes, a taxa de positivos reais foi igual a 85,9% e a taxa de negativos reais foi de 81,4% sobre o teste sobre o conjunto de validação. A acurácia balanceada foi igual a 83,66%. Segue abaixo as matrizes de confusão.



Baseline - Máquina de Vetor -Suporte

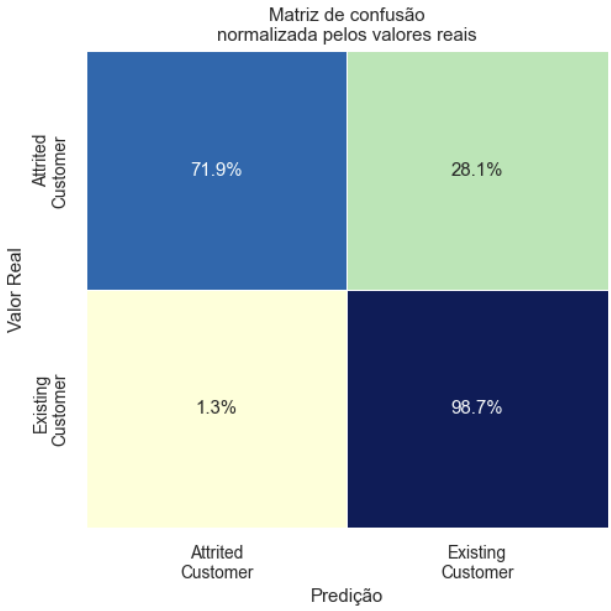
Já o modelo de máquina de vetor suporte apresentou um resultado ainda melhor do que o modelo de regressão logística com uma taxa de positivos reais igual a 91,8% e uma taxa de negativos reais igual à 88,9%. A acurácia balanceada foi igual a 90,37%. Segue abaixo as matrizes de confusão.

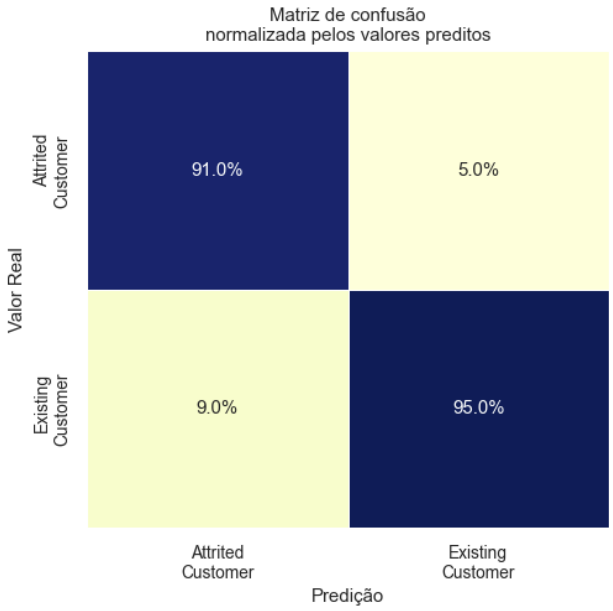




Baseline - Floresta Aleatória

O modelo de floresta aleatória também apresentou um resultado satisfatório com um taxa de positivos reais igual a 98,7% e uma taxa de negativos reais igual a 71,9%. A acurácia balanceada desse modelo foi igual a 85,31% portanto podemos considerar este modelo como melhor que o de regressão logística mas pior que o de máquina de vetor suporte. Vale notar que a taxa de negativos reais foi a menor, e a taxa de positivos reais a maior entre todos os modelos. Segue abaixo as matrizes de confusão.





Grid Search e outros modelos

Após treinamento e avaliação dos modelos baseline, fizemos uma busca em grade dos hiperparâmetros de cada um dos modelos. Durante essa busca, a avaliação das métricas para encontrar o melhor modelo foi feita utilizando validação cruzada com 5 dobras.

Regressão Logística + Grid Search

Os hiperparâmetros buscados foram:

- *C*: valor inverso da força de regularização;
- *penalty*: norma da penalidade;
- *max_iter*: número máximo de iterações para o modelo convergir;
- *tol*: tolerância para o critério de parada;
- *l1_ratio*: quando penalidade elasticnet é a combinação entre norma l1 e l2.

O melhor modelo encontrado teve os seguintes valores de hiperparâmetros: C = 0,1; penalty = l2; max_iter = 1000; tol = 0,0001. A acurácia balanceada foi igual a 84,97% +- 1,04%.

Máquina de Vetor-suporte + Grid Search

Os hiperparâmetros buscados foram:

- *C*: valor inverso da força de regularização;
- *kernel*: função kernel;
- *max_iter*: número máximo de iterações para o modelo convergir;
- *tol*: tolerância para o critério de parada;

O melhor modelo encontrado teve os seguintes valores de hiperparâmetros: C = 1; kernel = rbf; max_iter = 10000; tol = 0,001. A acurácia balanceada foi igual a 90,29% +- 1,26%.

Floresta Aleatória + Grid Search

Os hiperparâmetros buscados foram:

- *n_estimators*: número de árvores;
- *criterion*: função que mede a qualidade da separação dos nós da árvore;
- *max_depth*: máxima profundidade das árvores;
- *max_features*: número máximo de features das árvores;

O melhor modelo encontrado teve os seguintes valores de hiperparâmetros: n_estimators = 50; criterion = entropy; max_depth = 7; max_features = sqrt. A acurácia balanceada foi igual a 91,25% +- 0,40%.

Ensembles

Após obter os melhores modelos através de busca em grade, utilizamos ensembles dos melhores modelos, os ensembles utilizados foram ensemble de votação e ensemble de stacking.

No ensemble de votação estudado, cada predição dos modelos conta como um voto e a predição final é decidida por voto majoritário. A acurácia balanceada desse ensemble foi igual à 91,54% +- 0,60%.

No ensemble de stacking estudado, os modelos escolhidos pertencem à primeira camada de decisão, as predições dessa primeira camada serão utilizadas como features da segunda camada, também é feita a separação do conjunto de treino em uma estratégia semelhante a validação cruzada, onde a parte separada para treino é usada como treino dos modelos da primeira camada e a parte separada para validação usada como treino da segunda camada, a fim de evitar overfitting do ensemble. A acurácia balanceada desse ensemble foi igual à 94,24% +- 0,21%.

Outros Modelos

Também testamos outros modelos, XGBoost e CatBoost, ambos ensembles de algoritmos de árvore de decisão que utilizam gradient boosting, durante o estudo desses modelos fizemos a busca em grade dos hiperparâmetros.

XGBoost

Os hiperparâmetros buscados foram:

- *n_estimators*: número de árvores;
- *subsample*: fração aleatória de exemplos das amostras utilizadas em cada árvore;
- *min_child_weight*: peso mínimo necessário para criar novos nós;
- *gamma*: parâmetro de regularização usado para controlar a divisão dos nós;
- *colsample_bytree*: fração de features usadas em cada árvore;
- max_depth: máxima profundidade das árvores.

O melhor modelo encontrado teve os seguintes valores de hiperparâmetros: n_estimators = 250; subsample = 0,9; min_child_weight = 2; gamma = 3; colsample_bytree = 1; max_depth = 3. A acurácia balanceada foi igual a 95,58% +- 0,42%.

Cat Boost

Os hiperparâmetros buscados foram:

- max_depth: máxima profundidade das árvores;
- *learning_rate*: taxa de aprendizado;

- Iterations: número máximo de iterações;

O melhor modelo encontrado teve os seguintes valores de hiperparâmetros: max_depth = 5; learning_rate = 0,05; iterations = 300. A acurácia balanceada foi igual a 96,34% +- 0,27%.

Testando novamente ensembles

Testamos novamente os ensembles de votação e stacking dessa vez utilizando os melhores modelos encontrados por busca em grade de hiperparâmetros para todos os 5 modelos estudados (fora os ensembles anteriores). A acurácia balanceada do novo ensemble de votação foi igual à 94,29% +- 0,97%. A acurácia balanceada do novo ensemble de stacking foi igual à 97,22% +- 1,72%.

Comparando os modelos estudados

Abaixo segue a tabela com as métricas de todos os modelos testados, sobre o conjunto de validação (ou métrica da validação cruzada):

Modelo	Acurácia Balanceada	Acurácia	Score F1	RoC AuC	Ranking
Baseline: Regressão Logística	0,8366	0,8519	0,9073	0,8366	12
Baseline: SVM	0,9037	0,9136	0,9472	0,9037	7
Baseline: Floresta aleatória	0,8531	0,9451	0,9681	0,8531	10
Regressão Logística + GridSearch	0,8497 ± 0,0104	0,8484 ± 0,0088	0,9037 ± 0,0062	0,9255 ± 0,0034	11
SVM + GridSearch	0,9029 ± 0,0126	0,9121 ± 0,0064	0,9460 ± 0,0040	0,9643 ± 0,0052	8
Floresta aleatória + GridSearch	0,9125 ± 0,0040	0,9315 ± 0,0048	0,9584 ± 0,0031	0,9726 ± 0,0042	5
1º Voting Ensemble	0,9052 ± 0,0136	0,9154 ± 0,0060	0,9481 ± 0,0037	0,9052 ± 0,0136	6
1º Stacking Ensemble	0,8851 ± 0,0057	0,9424 ± 0,0021	0,9658 ± 0,0014	0,8851 ± 0,0057	9
XGBoost + GridSearch	0,9558 ± 0,0042	0,9635 ± 0,0034	0,9780 ± 0,0021	0,9558 ± 0,0042	3
CatBoost + GridSearch	0,9634 ± 0,0027	0,9699 ± 0,0031	0,9819 ± 0,0019	0,9634 ± 0,0027	2
2º Voting Ensemble	0,9429 ± 0,0097	0,9517 ± 0,0082	0,9708 ± 0,0051	0,9429 ± 0,0097	4
2º Stacking Ensemble	0,9722 ± 0,0172	0,9659 ± 0,0115	0,9793 ± 0,0070	0,9722 ± 0,0172	1

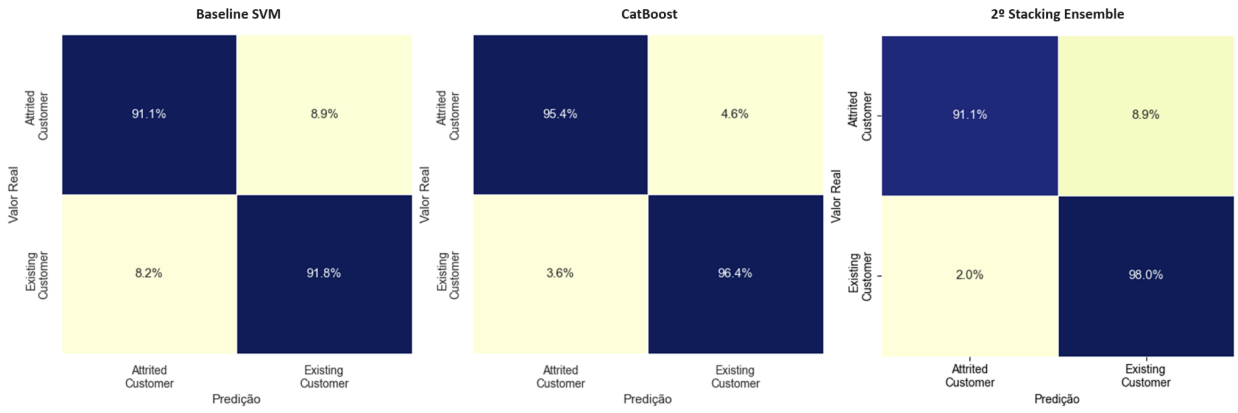
Entre todos os modelos estudados, o melhor modelo encontrado foi o 2º Ensemble de stacking. Vale notar que os modelos XGBoost, CatBoost foram os 2º e 3º melhores modelos, podemos considerar que houve um empate técnico entre CatBoost e o 2º ensemble de stacking devido ao desvio padrão de 1,72% do ensemble, lembrando que os valores indicados são a média sobre a validação cruzada mais o desvio padrão.

Comparando os 2 melhores modelos e o melhor baseline

Devido ao empate técnico entre o modelo CatBoost e o 2º Stacking Ensemble, iremos comparar o resultado desses modelos com o melhor modelo baseline sobre o conjunto de teste. Na imagem a seguir vemos a matriz de confusão normalizada pelos valores reais e as métricas de avaliação dos 3 modelos para o conjunto de teste.

Ao analisar a acurácia balanceada, nota-se que o modelo CatBoost apresentou uma performance melhor que os outros modelos, e ao observar a matriz de confusão

podemos considerá-lo como um modelo mais justo que o 2º Stacking Ensemble devido ao equilíbrio entre as taxas de positivos reais e de negativos reais. Além da acurácia balanceada, algo que corrobora para considerar a performance do modelo CatBoost como a melhor é o fato da taxa de negativos reais ser a maior dentre todos os modelos, como essa corresponde ao acerto do modelo quanto ao Churn do cliente ela acaba sendo mais interessante do que a taxa de positivos reais.



Modelo	Acurácia Balanceada	Acurácia	Score F1	RoC AuC
Baseline: SVM	0,9148	0,9171	0,9489	0,9148
CatBoost + GridSearch	0,9591	0,9625	0,9773	0,9591
2º Stacking Ensemble	0,9457	0,9689	0,9814	0,9457

Melhorias

Algumas melhorias e novos modelos poderiam ter sido estudadas nesse projeto, como por exemplo o estudo e aplicação de modelos de deep learning para dados tabulares. Outros estudos também poderiam ter sido feitos, como por exemplo explicabilidade e interpertabilidade dos modelos.

Outras melhorias relacionadas ao código do pré-processamento de dados também poderiam ter sido feitas, como por exemplo configurar o processamento de dados para considerar a variável target igual a 1 para Attrited customer e 0 para Existing customer a fim de tornar mais claro os resultados do modelo e transformar a métrica de acurácia em algo relacionado aos clientes que cancelaram o cartão.