**Título do Projeto**

#### **Otávio Fiori Maginador**

Universidade Federal do ABC (UFABC)

otavio.fiori@aluno.ufabc.edu.br

**Abstract –** This article contains a Data Analysis about the public dataset BankChurners. It will compare different classification models and data preparation trying to classify if a client will leave his bank. After comparing the models, it will be analysed the best decision threshold to classify the client.

**Keywords –** Bank Churners. Unbalanced. Classification. AUC Score. F1-Score.

# 1. Introdução

É bastante comum clientes abandonarem uma instituição financeira mesmo após anos de relacionamento ainda que diversas ações de produtos e marketing sejam realizadas, fenômeno o qual se denomina *churn*. No setor bancário americano, por exemplo, estima-se que 30% da base de clientes seja vulnerável à migração [1] e algumas variáveis podem ajudar a prever tal vulnerabilidade, como características do indivíduo, atributos da organização e do produto e fatores externos (como variáveis macroeconômicas e culturais). Aliado a isso é geralmente mais caro conquistar novos clientes do que manter os já existentes. Em muitos setores, o custo de aquisição de novos clientes pode ser cinco vezes superior ao seu custo de retenção.

A maioria dos estudos necessários para evitar este fenômeno são comumente realizados pelas equipes de CRM (*Customer Relationship Management*), que é principalmente responsável por gerir toda a comunicação e tratamento de uma instituição com seus clientes, assim como garantir a satisfação deles com os produtos vendidos pela empresa. Dentre seus principais trabalhos está a modelagem de *churn*, onde a equipe tenta prever quais são os clientes mais propensos a deixá-la em breve.

Infelizmente, no Brasil, ainda são escassos os artigos científicos que abordam conceitos de data mining e a questão do abandono de clientes, por isso este artigo tem os objetivos: i) modelar a probabilidade de clientes deixarem a empresa utilizando dados do histórico de relacionamento e uso dos produtos pelo cliente, ii) descrever as possíveis variáveis que influenciam o abandono/permanência do cliente, iii) verificar e comparar as métricas necessárias para validação do modelo.

# 2. Dados

Para estudo e escrita deste artigo foi utilizado o banco de dados público *Credit Cards Customers*, em que reúne informações transacionais de cartões de crédito, de relacionamento da instituição com o cliente e algumas características individuais dele.

## 2.1 Variáveis preditoras

Dentre todas as variáveis do *dataset* temos as contínuas e as categóricas. As variáveis contínuas são aquelas com características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores fracionais fazem sentido, que neste *dataset* são:

* *Customer\_Age*: idade do cliente;
* *Dependent\_count*: número de dependentes do cliente;
* *Months\_on\_book*: tempo de relacionamento com o banco;
* *Total\_Relationship\_Count*: quantidade de produtos contratados no banco;
* *Months\_Inactive\_12\_mon*: meses inativos nos últimos 12 meses;
* *Contacts\_Count\_12\_mon*: quantidade de contatos com o banco nos últimos 12 meses;
* *Credit\_Limit*: limite de crédito no cartão de crédito;
* *Total\_Revolving\_Bal*: saldo rotativo no cartão de crédito;
* *Avg\_Open\_To\_Buy*: média da linha de crédito para compra dos últimos 12 meses;
* *Total\_Amt\_Chng\_Q4\_Q1*: alteração no valor de transação do quarto trimestre dividido pelo primeiro trimestre;
* *Total\_Trans\_Amt*: valor total transacionado nos últimos 12 meses;
* *Total\_Trans\_Ct*: quantidade total de transações nos últimos 12 meses;
* *Total\_Ct\_Chng\_Q4\_Q1*: alteração na contagem de transações do quarto trimestre sobre o primeiro trimestre;
* *Avg\_Utilization\_Ratio*: taxa média de utilização do cartão.

Nesta primeira parte é possível notar que são utilizadas muitas variáveis relacionadas ao uso do cartão de crédito em si, principalmente o que chamamos de variáveis janeladas, que são aquelas que fazem cálculos entre janelas de tempo específicas, como a *Total\_Ct\_Chng\_Q4\_Q1*. Este tipo de variável é muito utilizado pois o seu cálculo gera uma informação muito relevante indicando se o cliente está aumentando ou diminuindo o volume de transações no último período.

Uma variável categórica é uma variável que pode assumir apenas um número limitado, e geralmente fixo, de valores possíveis, atribuindo cada indivíduo ou outra unidade de observação a um determinado grupo ou categoria nominal com base em alguma propriedade qualitativa. Neste *dataset* temos 5 variáveis deste tipo:

* *Gender*: o gênero do cliente;
* *Education\_Level*: nível educacional;
* *Marital\_Status*: estado civil;
* *Income\_Category*: categoria do salário;
* *Card\_Category*: categoria do cartão.

Nestas variáveis é possível notar dados que qualificam o cliente, como a categoria do cartão e salário, informações muito utilizadas também para traçar o perfil de cliente da empresa.

## 2.2 Variável resposta

A variável resposta de um modelo é aquela à qual desejamos prever os resultados e neste *dataset* será utilizada a variável ‘*Attrition\_Flag*’ que pode assumir os valores ‘*Existing Customer*’ ou ‘*Attrited Customer*’, sendo a segunda a identificação do *churn*.

|  |  |  |
| --- | --- | --- |
| Attrition\_Flag | Quantidade de Clientes | Percentual |
| Attrited Customer | 1.627 | 16,07% |
| Existing Customer | 8.500 | 83,93% |
| Total | 10.127 | 100,00% |

**Tabela 1. Distribuição do *churn*.**

Como pode ser visto na tabela cima apenas 16,1% dos clientes apresentaram *churn* no *dataset*, o que significa que temos dados desbalanceados e para isso serão necessárias algumas métricas além de acurácia para a validação do modelo.

# 3. Métodos e Métricas

Nesta sessão serão apresentados os algoritmos utilizados no artigo (Regressão Logística, Naive Bayes e Árvores de Decisão) assim como as métricas necessárias para a validação dos modelos: acurácia, recall e curva ROC (*Receiver Operating Characteristic Curve*).

## 3.1 Regressão Logística

É uma técnica recomendada para situações em que a variável dependente é de natureza dicotômica ou binária. Quanto às independentes, tanto podem ser categóricas ou não. A regressão logística é um recurso que permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias.

Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada diretamente. No caso da variável dependente Y assumir apenas dois possíveis estados (1 ou 0) e haver um conjunto de p variáveis independentes X1, X2, ..., Xp, o modelo de regressão logística pode ser escrito da seguinte forma:

(1) onde, (2).

Esta técnica assume as premissas de que todas as variáveis contínuas seguem uma distribuição normal, ou seja, onde média é igual a 0 e desvio padrão é igual a 1, assim como que todas as variáveis são independentes. Estas premissas fazem com que sejam necessários alguns pré-processamentos nas variáveis preditoras.

## 3.2 Naive Bayes

O classificador Naive Bayes é um algoritmo que se baseia nas descobertas de Thomas Bayes para realizar predições em aprendizagem de máquina. O termo “*naive*” (ingênuo) diz respeito à forma como o algoritmo analisa as características de uma base de dados: ele assume que as variáveis são independentes entre si.

 O seu funcionamento pode ser facilmente descrito em termos estatísticos: ele calcula probabilidades (3) para cada uma das K possíveis saídas da classe resposta, representadas por um vetor (4) codificando cada uma das *n* variáveis. O grande problema da fórmula acima é que se o número de variáveis *n* for muito grande ou se um recurso assumir muitos valores, basear esse modelo em tabelas de probabilidade é inviável, por isso é utilizada a probabilidade condicional:

(5).

Por haver a premissa de que as variáveis são independentes entre si é necessário, assim como na regressão logística, aplicar pré-processamento aos dados.

## 3.3 Árvores de Decisão

Uma árvore de decisão é um algoritmo de aprendizado de máquina supervisionado que é utilizado para classificação e para regressão. Assim como um fluxograma, a árvore de decisão estabelece nós (*decision nodes*) que se relacionam entre si por uma hierarquia. Existe o nó-raiz (*root node*), que é o mais importante, e os nós-folha (*leaf nodes*), que são os resultados. No contexto de *machine learning*, a raiz é um dos atributos da base de dados e o nó-folha é a classe ou o valor que será gerado como resposta.

Na ligação entre nós, temosregras de *if-else* e ao chegar em um nó A, o algoritmo se pergunta acerca de uma regra, uma condição, como “a característica X do registro analisado é menor do que 15?”. Se for menor, então ele vai para um lado da árvore, se for maior, então ele vai para outro. No próximo nó, segue a mesma lógica.

Para fazer a procura do dado que maximiza o acerto do algoritmo no próximo passo é utilizado o algoritmo *Greedy Search* ou busca gananciosa. Ele cria partições nos dados e com base na melhoria de alguma medida estatística (definido pelo pesquisador) ele escolhe a variável do próximo nó.

## 3.4 Acurácia e Recall

Para entender melhor cada métrica, primeiro é necessário entender alguns conceitos.

Uma matriz de confusão é uma tabela que indica os erros e acertos de um modelo, comparando com o resultado esperado.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Detectado | |
|  |  | Sim | Não |
| Real | Sim | VP | FN |
| Não | FP | VN |

**Tabela 2. Matriz de Confusão.**

Na tabela anterior definimos Verdadeiro Positivo como VP, Falso Negativo como FN, Falso Positivo como FP e Verdadeiro Negativo como VP.

A acurácia é o quanto o modelo acertou, ou seja (6) e o *recall* é a relação de quantos positivos reais o modelo acertou, ou seja (7).

A acurácia é mais utilizada em casos em que a variável resposta está balanceada, já o *recall* é mais utilizado em *datasets* desbalanceados em que o Falso Negativo é mais prejudicial que o Falso Positivo e, no caso do *churn*, é pior para uma empresa prever que o cliente não irá abandoná-la e ele de fato a deixar do que o contrário.

## 3.5 Curva ROC

A curva ROC é uma representação gráfica que ilustra o desempenho (ou performance) de um sistema classificador binário à medida que o seu limiar de discriminação varia. A curva ROC é também conhecida como curva de característica de operação relativa, porque o seu critério de mudança é resultado da operação de duas características (VP e FP).

A curva ROC é obtido pela representação da razão (8) versus a razão (9), para vários valores do limiar de classificação. O RVP é também conhecido como sensibilidade (ou taxa de verdadeiros positivos), e RFP com taxa de falsos positivos. Além de sua representação visual é utilizada também o cálculo da área abaixo da curva (AUC – *area under the curve*) que varia entre 0 e 1.

# 4. Desenvolvimento do Modelo

Nesta sessão serão apresentadas as variáveis preditoras e as etapas de pré-processamento utilizadas assim como o desenvolvimento do modelo e seus resultados.

## 4.1 Pré-processamento

Como abordado nos itens 3.1 e 3.2, os algoritmos de Regressão Logística e Naive Bayes requerem um determinado cuidado com os dados que serão apresentados aos modelos, por isso são realizadas as análises de distribuição das variáveis contínuas e suas correlações.

Ao se fazer a análise de distribuição das variáveis contínuas foi verificado que a *Months Inactive 12 mon, Credit Limit, Total Amt Chng Q4 Q1, Total Ct Chng Q4 Q1* e *Avg Utilization Ratio* tem distribuições bastante diferentes da normal.

Assim como também podemos perceber que há uma correlação alta entre algumas variáveis pela tabela abaixo.

|  |  |  |
| --- | --- | --- |
| Variável 1 | Variável 2 | Correlação de Pearson |
| Avg\_Open\_To\_Buy | Credit\_Limit | 1 |
| Total\_Trans\_Ct | Total\_Trans\_Amt | 0,81 |
| Months\_on\_Book | Customer\_Age | 0,79 |

**Tabela 3. Correlação de Pearson.**

Dada a análise as variáveis contínuas mencionadas foram normalizadas e as variáveis com alta correlação removidas do *dataset*.

## 4.2 Resultados

Feito o pré-processamento dos dados foram treinados os 3 modelos citados anteriormente após fazer a busca de melhores hiper parâmetros utilizando a validação cruzada com 5 *folds*.

Os melhores hiper parâmetros da regressão logística foram 0,5 para o inverso da força de regularização e penalidade l2, para o Naive Bayes foi utilizado como número de Laplace e para a árvore de decisão o gini foi utilizado como critério de decisão, profundidade máxima de 7 e o mínimo de amostras na folha foi de 0,05.

Os resultados foram os apresentados abaixo.

|  |  |  |  |
| --- | --- | --- | --- |
| Algoritmo | Acurácia | Recall | AUC Score |
| Regressão Logística | 0,91 | 0,57 | 0,91 |
| Naive Bayes | 0,89 | 0,66 | 0,87 |
| Árvore de Decisão | 0,90 | 0,56 | 0,93 |

**Tabela 4. Métricas dos modelos.**

Para este caso em que o *churn* ocorre apenas com poucos clientes é possível notar que a acurácia dos modelos foram muito altas, porém utilizá-la pode ser um grande erro de interpretação, pois ao verificar a tabela 1 podemos entender que caso o modelo classificasse todos os clientes como *Existing Customers* a acurácia já seria de 83,93%, por isso as métricas de *recall* e *AUC Score* são importantes.

É possível perceber que a Árvore de Decisão e a Regressão Logística tiveram uma pontuação maior na *AUC Score* e isso mostra que ambas são melhores que o Naive Bayes para distinguir uma classe da outra o que permite que posteriormente sejam feitos cortes nas probabilidades, ou aplicar *thresholds*, para que as previsões acertem mais clientes que deram *churn*.

Em contrapartida, o Naive Bayes apesar de apresentar uma *AUC Score* mais baixa ainda assim apresentou um bom resultado além de ter um *recall* mais expressivo que os demais. Isso permite que as previsões realizadas pelo modelo priorizem o erro de classificar que um cliente irá deixar a instituição e o mesmo não deixar do que prever que o cliente não irá deixar a empresa e no final ocorrer o *churn*.

# 5. Conclusões

Considerando que as perdas de lucro devido ao abandono de clientes podem ser significativas, a previsão de *churn* representa uma informação muito relevante aos gestores. Clientes com indícios de *churn* podem ser alvos de ações de marketing direcionado, público de novos produtos para aumentar a retenção e até mesmo de descontos nos produtos que mais os interessam.

No caso de auxiliar a empresa em tomadas de decisões e de ser mais ativo ao indicar um possível *churn* o algoritmo do Naive Bayes se mostrou mais interessante pois obteve um melhor equilíbrio entre as métricas apuradas.

As variáveis presentes neste estudo mostraram um ótimo poder de discriminação de abandono do cliente podendo assim auxiliar e criar *insights* relevantes para que empresas apliquem as técnicas aqui destacadas e evoluam cada vez mais em aprendizado de máquina no Brasil.

# Referências

[1] ACCENTURE. The new customer imperative: retaining and acquiring customers in a changed banking landscape. The Point, v. 9, n. 4, p. 1-8, 2009.

[2] Regressão Logística. Disponível em: < https://edisciplinas.usp.br/pluginfile.php/3769787/mod\_resource/content/1/09\_RegressaoLogistica.pdf>. Acesso em: 04 de maio de 2023.

[3] FILHO, Mario. Precisão, Recall e F1 Score Em Machine Learning. Disponível em: < https://mariofilho.com/precisao-recall-e-f1-score-em-machine-learning/>. Acesso em: 04 de maio de 2023.