Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas
Departamento de Estatística
Bacharelado em Estatística

Estudo de Desempenho de Modelos de IA aplicados ao churn de clientes

Erick Vinícius de Araújo Silva

Ouro Preto 25 Novembro de 2023

ERICK VINÍCIUS DE ARAÚJO SILVA

ESTUDO DE DESEMPENHO DE MODELOS DE IA APLICADOS AO CHURN DE CLIENTES

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Estatística do Universidade Federal de Ouro Preto - para a obtenção do título de Estatístico.

Orientador: Tiago Martins Pereira

Ouro Preto - MG
2023

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
COLEGIADO DO CURSO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Erick Vinícius de Araújo Silva

Estudo de desempenho de modelos de IA aplicados ao churn de clientes

Monografia apresentada ao Curso de Estatística da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovada em 22 de fevereiro de 2024

Membros da banca

Dr. Tiago Martins Pereira - Orientador (Universidade Federal de Ouro Preto)
Dra. Diana Campos de Oliveira - Membro (Universidade Federal de Ouro Preto)
Dr. Fernando Luiz Pereira de Oliveira - Membro (Universidade Federal de Ouro Preto)

Prof. Dr. Tiago Martins Pereira, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 22/02/2024

Documento assinado eletronicamente por **Tiago Martins Pereira**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 26/02/2024, às 10:21, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0672662** e o código CRC **DD7A52DC**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3559-1312 - www.ufop.br

AGRADECIMENTOS

Desejo expressar meu reconhecimento a todos os meus familiares, em especial, aos meus pais e meu irmão que são o alicerce mais sólido da minha vida em todos os momentos. Agradeço também aos meus professores e amigos que fiz durante essa árdua jornada que chamamos de graduação.

RESUMO

O presente trabalho discute o impacto do avanço tecnológico na sociedade, destacando a importância da Estatística no processamento e análise de dados gerados em grande volume. Enfatiza a relevância da classificação de dados, particularmente em problemas como a Detecção de Fraude, destacando o papel dos algoritmos de aprendizado de máquina nesse contexto. Um exemplo prático abordado é a aplicação desses modelos para prever o cancelamento de cartões de crédito, ressaltando a utilidade da análise de dados na identificação de padrões e na tomada de decisões preventivas. O problema de negócio central que visa ser estudado no artigo é citado inicialmente no portal Kaggle, uma plataforma relevante para cientistas de dados, que oferece competições que abordam desafios reais de negócios. O cenário proposto envolve um gerente de banco que busca compreender e prever a perda de clientes, utilizando técnicas de aprendizado de máquina em uma base de dados financeiros. O objetivo geral do trabalho é explorar, utilizar e comparar algoritmos de aprendizado de máquina, investigando os fatores que influenciam o cancelamento de cartões de crédito para auxiliar o banco na retenção de clientes e manutenção de sua competitividade no mercado.

Palavras-chave: Churn; Machine Learning; Estatística; IA.

ABSTRACT

The present work discusses the impact of technological advancement on society, emphasizing the importance of Statistics in processing and analyzing data generated in large volume. It highlights the relevance of data classification, particularly in issues such as Fraud Detection, emphasizing the role of machine learning algorithms in this context. A practical example addressed is the application of these models to predict credit card cancellations, emphasizing the utility of data analysis in pattern identification and preventive decision-making. The central business problem to be studied in the article is initially mentioned on the Kaggle platform, a relevant hub for data scientists that provides competitions addressing real business challenges. The proposed scenario involves a bank manager aiming to understand and predict customer loss, using machine learning techniques on financial data. The overall goal of the work is to explore, employ, and compare machine learning algorithms, investigating factors influencing credit card cancellations to assist the bank in customer retention and maintaining competitiveness in the market.

Keywords: Churn; Machine Learning; Estatística; IA.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

UFOP	Universidade Federal de Ouro Preto
DEEST	Departamento de Estatística

SUMÁRIO

1 INTRODUÇÃO

O avanço tecnológico tem o potencial de promover o desenvolvimento humano em diversos campos da nossa sociedade. Com esse avanço e o aumento exponencial no volume de dados e informação que são gerados dia após dia, surge a necessidade do desenvolvimento de técnicas e métodos capazes de processar e analisar esses dados com precisão e agilidade. A Estatística, como campo do conhecimento, ganha um papel de destaque nesse processo, uma vez que que estuda e desenvolve um grande volume de técnicas usadas no processo de descoberta da informação.

Dentro deste contexto, um dos pontos mais importantes em análise de dados são os problemas de classificação, que envolvem a categorização de dados em classes predefinidas. Formalmente, dado um conjunto de dados de treinamento composto por instâncias rotuladas, ou seja, um conjunto de dados onde cada indivíduo já possui a classe ou categoria correta associada, o objetivo é desenvolver um modelo capaz de generalizar para classificar novas instâncias não rotuladas. Técnicas estatísticas tradicionais como Regressão Logística e Análise discriminante podem ser utilizadas para este fim. Além dessas técnicas, alguns algoritmos de aprendizado de máquina como Redes Neurais, Máquinas de Vetores e Suporte, Árvores de decisão, dentre outros, também são utilizados com esta finalidade. De acordo com Borges (2020) os algoritmos de Machine Learning podem ser aplicados para resolução de diversos tipos de problemas, ou seja, são usados tanto para predizer um indicador num setor industrial quanto dentro da área da saúde, contribuindo assim para a eficácia dos processos, diagnósticos, tomadas de decisão entre outros. Esses modelos aprendem a relação entre os atributos e as classes a partir desse conjunto de dados e, em seguida, são capazes de prever a classe de novas instâncias que não foram vistas durante o treinamento. Devido a esse potencial, os modelos de classificação têm uma ampla gama de aplicações práticas em diversas áreas, como: Recomendação de Produtos; Reconhecimento de Imagens; Classificação de Textos; Detecção de Fraude entre vários outros.

Entre os exemplos citados, a Detecção de Fraude possui um grande espaço e importância dentre as aplicações práticas, pois o desenvolvimento desses modelos pode ajudar a identificar e prevenir transações fraudulentas em cartões de crédito e outras operações financeiras em todo o mundo. Eles funcionam analisando transações com base em diversos atributos, como o valor, a localização, a hora do dia, o tipo de transação, entre outros. Os modelos podem ser treinados em um conjunto de dados rotulados, que contém informações sobre transações fraudulentas e legítimas. Com base nesse conjunto de dados, o modelo aprende a identificar padrões que distinguem as transações fraudulentas das legítimas. Se uma transação se encaixa no padrão de fraude identificado pelo modelo, o sistema pode alertar o usuário ou funcionário responsável pela análise de risco permitindo que a transação seja investigada e possivelmente bloqueada.

1.1 DEFINIÇÃO DO PROBLEMA

Suponha que um gerente de negócios de um banco que oferece o cartão crédito como um de seus produtos está enfrentando um problema de perda de clientes. Eles precisam analisar os dados para descobrir o motivo por trás disso e aproveitar o mesmo para prever os clientes que provavelmente desistirão. Esse problema de classificação foi inicialmente proposto no repositório online Kaggle (KAGGLE, 2023).

O Kaggle é uma plataforma online de competições, onde os cientistas de dados podem acessar conjuntos de dados interessantes e desafiadores para construir e avaliar seus modelos de aprendizado de máquina, além de terem a oportunidade de colaborar com outros profissionais da área. O Kaggle foi fundado em 2010 e adquirido pelo Google em 2017. A plataforma é composta por uma comunidade global, que alimenta esse grande repositório em desafios de aprendizado de máquina, estatística e ciência de dados de maneira geral, para resolver problemas reais de negócios. As competições do Kaggle são realizadas em conjunto com empresas e organizações, que fornecem os conjuntos de dados e definem o problema a ser resolvido.

As competições do Kaggle são uma excelente maneira de se envolver em projetos de ciência de dados de alto nível e colaborar com outros cientistas de dados. Além disso, a plataforma oferece uma ampla variedade de recursos, incluindo tutoriais, conjuntos de dados, fóruns de discussão e ferramentas de visualização, que ajudam os cientistas de dados a melhorarem suas habilidades e conhecimentos.

A análise de dados pode ser uma ferramenta muito valiosa para a resolução do problema, já que permite ao gerente avaliar informações relevantes, como o perfil dos clientes que estão cancelando o cartão, os motivos que os levaram a tomar essa decisão, entre outros fatores. Além disso com base nesses dados, é possível desenvolver modelos preditivos que ajudem a prever quais clientes tem maior probabilidade de cancelar o cartão, permitindo que medidas sejam tomadas de forma preventiva. Segundo Katelaris e Themistocleous (2017), o churn ou desistência/cancelamento de clientes ocorre quando um cliente termina a relação com uma empresa para possivelmente iniciar um relacionamento com uma organização concorrente. Nas empresas, esse comportamento é medido pela taxa de churn. E ainda, Segundo Kotler (1999), conquistar um novo cliente custa entre 5 e 7 vezes mais do que manter um atual. Logo, investir na fidelização de clientes atuais é mais lucrativo do que investir na aquisição de novos clientes.

Dessa forma, a análise de dados pode ser uma estratégia importante para ajudar o banco a reter seus clientes e garantir a sua competitividade no mercado. Ao identificar os fatores que influenciam na perda de clientes e desenvolver soluções adequadas para lidar com essas questões, o banco poderá melhorar o seu relacionamento com seus clientes e oferecer um serviço cada vez mais adequado às suas necessidades.

1.2 *Objetivo geral*

O principal objetivo deste trabalho é o entendimento, a utilização e a comparação de algoritmos de aprendizado de máquina no contexto de análise de dados rotulados, utilizando para isso uma base de dados financeiros. Além disso, pretende-se investigar os fatores que influenciam no cancelamento de cartões de crédito.

1.3 Estrutura do trabalho

Este trabalho está dividido em cinco partes, iniciando com uma breve introdução sobre o tema abordado, apresentando a motivação e os principais aspectos para o desenvolvimento deste trabalho.

Na segunda parte, são apresentados a revisão dos conceitos teóricos necessários para a realização desse trabalho. Nesta etapa são explicadas algumas das diferenças existentes entre os métodos clássicos de aprendizado de máquina.

Na terceira parte do trabalho, foi apresentado a aplicação das técnicas para churn de clientes, como também a estruturação desses modelos e seus elementos.

Após isso, são analisados os resultados obtidos com a aplicação dos métodos propostos durante todo o trabalho. Também o conjunto experimental para a avaliação, apresentando o desempenho dos algoritmos. Por fim, são feitas as conclusões com base em cada resultado obtido e discutidas as possíveis melhorias que poderão ser feitas a partir de trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

2.1 Redes Neurais Artificiais

As redes neurais artificiais (RNAs) são a espinha dorsal da revolução atual em inteligência artificial. São sistemas computacionais inspirados no funcionamento do cérebro humano, capazes de aprender e tomar decisões. O sistema nervoso biológico apresenta uma variedade de arquiteturas complexas. Esses sistemas intricados são formados por células neurais, conhecidas como neurônios, cada uma desempenhando funções específicas. As células nervosas têm um corpo celular que contém dendritos e axônios, conforme ilustrado na Figura 1:

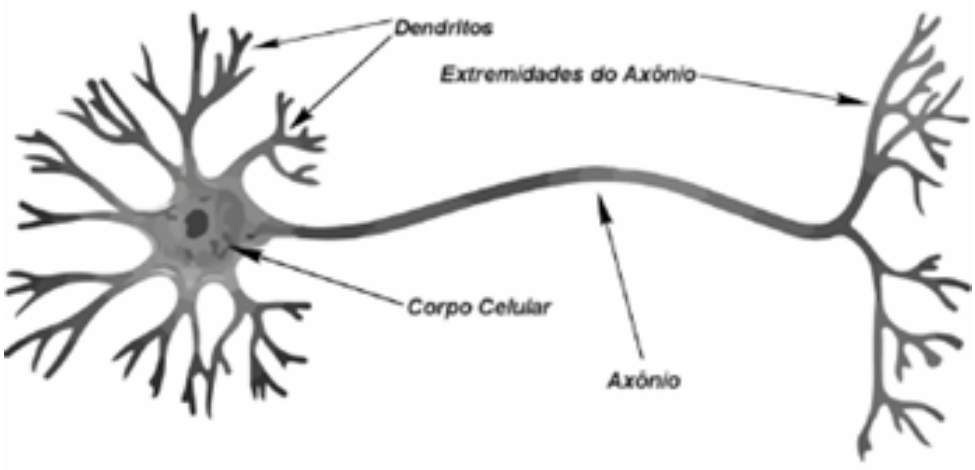

Figura 1 – Esboço de um neurônio biológico.

O corpo celular contém informações cruciais sobre as características da célula, além de um plasma que contém as substâncias moleculares necessárias para o funcionamento adequado da célula. A comunicação entre os neurônios ocorre por meio de impulsos que são captados pelos dendritos, responsáveis por receber e transmitir informações para o corpo celular através do axônio. O axônio, que se ramifica em colaterais, recebe sinais do corpo celular e os conduz até os dendritos, que por sua vez transmitem esses sinais para os dendritos de outros neurônios vizinhos por meio da sinapse, como representado na Figura 2:

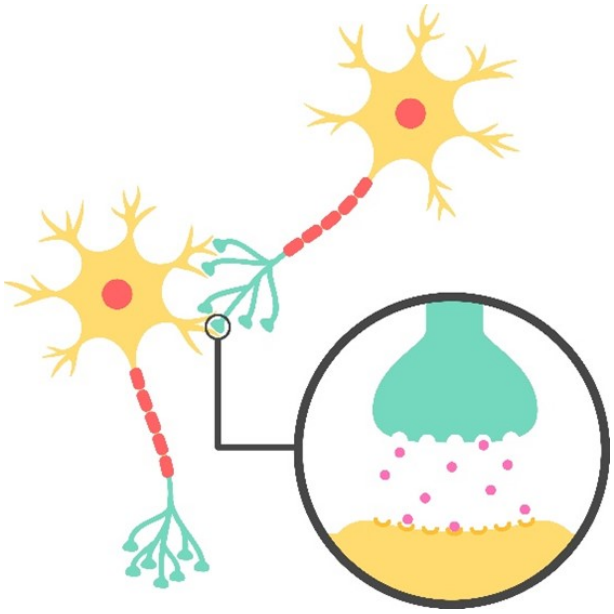

Figura 2 – Representação do mecanismo de comunicação entre neurônios biológicos.

A conexão entre as redes artificiais e biológicas reside na presença de axônios e dendritos, bem como na comunicação por meio de sinapses. Essa relação é representada na Figura 3, onde a letra x denota os sinais recebidos e os pesos sinápticos são simbolizados por w. Ambas as redes têm a capacidade de ajustar a amplitude das sinapses em várias camadas interconectadas.

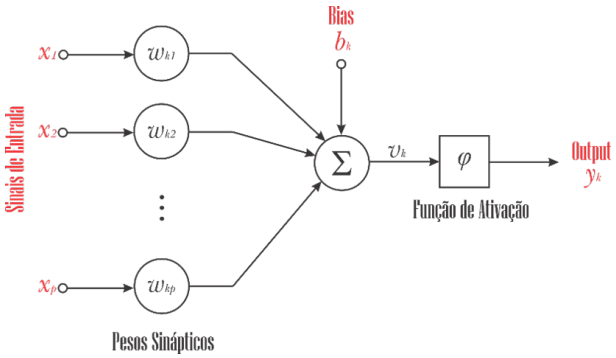

Figura 3 – Ilustração de um Neurônio Artificial. Adaptado (HAYKIN, 2001).

Em 1958, a Rede Neural Artificial Perceptron foi introduzida por Frank Rosenblatt (**??**), inspirado nos trabalhos de Walter Pitts e Warren Sturgis McCulloch (MCCULLOCH; PITTS, 1943). Esse modelo é um dos mais antigos e lida com um único neurônio, classificando o resultado de forma linear. Na Figura 3, o neurônio artificial é um Perceptron que recebe diversos valores de entradas $x_i, \quad i = 1 \ldots p$. Essas entradas multiplicam-se pelo peso da sinapse $w_{ki}, \quad i = 1 \ldots p$ e, no final, somam-se formando um conjunto de entrada

$$v = \sum_{k=1}^{n} w_k x_i + b_k$$

. Esse resultado passa por uma função de ativação linear e transmite a saída y. Quando o valor v exceder o limite da função de ativação, o neurônio será ativado e retornará um valor.

O parâmetro bias (viés), representado por b_k é um valor adicional adicionado a cada soma ponderada nas camadas intermediárias e na camada de saída. O objetivo do viés é permitir que a rede neural aprenda melhor e se ajuste a padrões mais complexos nos dados.

As funções de ativação em redes neurais são utilizadas para introduzir não linearidades nas saídas dos neurônios. Segundo os autores (HAYKIN, 2001), existem várias funções de ativação, onde cada uma delas tem características específicas. Algumas das funções de ativação comumente usadas são:

• Função Threshold: Útil em problemas onde se deseja atribuir uma saída binária, como 0 ou 1, com base em um limite. É definida da seguinte forma:

$$\varphi(x) = \begin{cases} 1 & \text{se } x \geq 0 \\ 0 & \text{se } x < 0 \end{cases} \tag{2.1}$$

• Função Sigmoide (Logística): A saída está no intervalo de 0 a 1. Essa função era popular em redes neurais antigas, mas atualmente é menos comum nas camadas intermediárias devido a problemas de desvanecimento de gradientes. É definida da seguinte forma:

$$\varphi(x) = \frac{1}{1+e^{-x}} \tag{2.2}$$

• Função Tangente Hiperbólica: Semelhante à sigmoide, mas com saída variando de -1 a 1. Também era mais comum em redes antigas e é usado ocasionalmente. É definida da seguinte forma:

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.3}$$

• Unidade Linear Retificada (ReLU): Torna-se zero para valores negativos e é linear para valores positivos. É uma escolha popular para camadas intermediárias devido à sua simplicidade e eficácia. É definida da seguinte forma:

$$\text{ReLU}(x) = \begin{cases} 0 & \text{se } x < 0 \\ x & \text{se } x \geq 0 \end{cases} \tag{2.4}$$

• Softmax: Usada na camada de saída para problemas de classificação multiclasse. Transforma os valores de saída em uma distribuição de probabilidade. É definida da seguinte forma

$$\sigma(x) = \frac{e^{x}}{\sum_{i=1}^{n} e^{x_i}} \tag{2.5}$$

Apesar do Perceptron demonstrar eficiência, a pesquisa de Minsky et .al (MINSKY; PAPERT, 1969) evidenciou que esse modelo não consegue resolver problemas nos quais as classes não podem ser separadas linearmente. Em outras palavras, para o Perceptron operar corretamente, é necessário que os exemplos a serem classificados estejam suficientemente distantes uns dos outros, assegurando que a superfície de separação seja um hiperplano (HAYKIN, 2001). Uma solução sofisticada para lidar com problemas que não são linearmente separáveis foi apresentada pelo algoritmo Backpropagation, proposto por Rumelhart e McClelland (RUMELHART; MCCLELLAND; GROUP, 1986). Esse algoritmo possibilita a adaptação dos pesos sinápticos em redes neurais com múltiplas camadas de neurônios totalmente conectados, conforme mostrado na Figura 4.

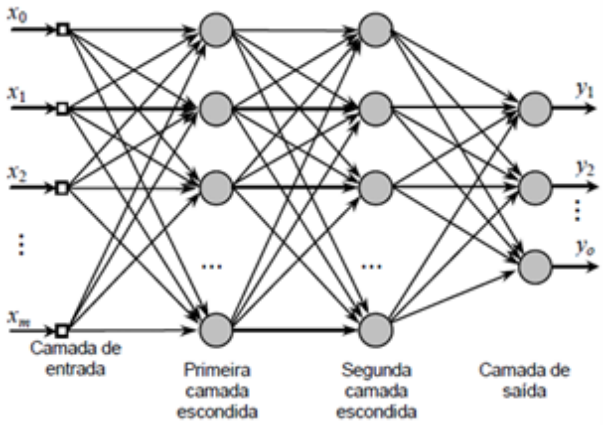

Figura 4 – Redes neurais MLP com duas camadas intermediárias.

O modelo MLP (Multi-Layer Perceptron) é composto de camadas de neurônios conectados que são treinados para aprender a relação entre um conjunto de entradas e uma saída correspondente. Essas camadas são geralmente compostas de uma camada de entrada, uma ou mais camadas intermediárias e uma camada de saída.

- Camada de Entrada: Nessa camada, cada neurônio representa um atributo ou característica dos dados de entrada. Os dados são normalmente normalizados para garantir que todas as características tenham o mesmo peso durante o processo de treinamento;
- Camadas Ocultas: Essas são camadas intermediárias entre a camada de entrada e a camada de saída. Cada neurônio nas camadas ocultas recebe os valores das camadas anteriores e aplica uma combinação linear ponderada desses valores, seguida de uma função de ativação. A função de ativação é importante para introduzir não linearidades no modelo, permitindo que ele capture relações complexas nos dados;
- Camada de Saída: Essa camada é responsável por gerar a saída do modelo. O número de neurônios na camada de saída dependerá do tipo de problema que estamos abordando. Por exemplo, em problemas de classificação binária, teremos um neurônio de saída, enquanto

em problemas de classificação multiclasse, o número de neurônios na camada de saída será igual ao número de classes.

O treinamento de uma RNA é feito através de um processo iterativo chamado de "back-propagation"(retropropagação), conforme mostrado na Figura 5. Ele consiste em dois principais passos: propagação direta (forward pass) e propagação reversa (backward pass).

- Propagação direta: Nessa etapa, os dados de treinamento são alimentados à RNA, e a saída é gerada pela rede seguindo o fluxo da camada de entrada para a camada de saída. Essa saída é comparada com as saídas esperadas (rótulos) dos dados de treinamento.
- Propagação reversa: Nessa etapa, os erros entre as saídas geradas e os rótulos são calculados. Em seguida, esses erros são propagados de volta pelas camadas da RNA, ajustando os pesos e viés de cada neurônio usando métodos de otimização, como Gradiente Descendente. O objetivo é minimizar a função de perda, que mede a diferença entre as saídas reais e as saídas previstas.

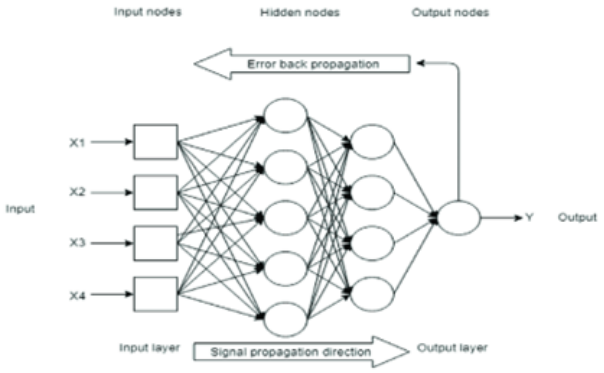

Figura 5 – Algoritmo backpropagation.

A aplicação de modelos de Redes Neurais Artificiais oferece diversas vantagens que podem ser resumidas de maneira abrangente. A flexibilidade é uma característica marcante, permitindo que as RNAs se adaptem a uma ampla gama de problemas ao aprender e mapear relações complexas nos dados. A representação hierárquica, potencializada pelas camadas ocultas, capacita as RNAs a aprenderem características abstratas e de alto nível, proporcionando uma compreensão mais profunda dos dados. Além disso, a capacidade de generalização é notável, permitindo que as RNAs apliquem padrões aprendidos em dados de treinamento para novos dados não previamente observados. O advento do Deep Learning e o aumento da capacidade computacional têm ampliado a aplicabilidade das RNAs, especialmente em grandes conjuntos de dados, onde esses modelos têm demonstrado um desempenho excepcional.

2.2 Algoritmos classificadores

2.2.1 SVM (Support Vector Machine)

O Support Vector Machine (SVM), em português Máquina de Vetores de Suporte, é um poderoso algoritmo de aprendizado de máquina utilizado para tarefas de classificação e regressão, elaborado com o estudo proposto por Boser (BOSER; GUYON; VAPNIK, 1992).

Em sua essência, o SVM tem como função identificar a fronteira de separação mais eficaz entre classes ou rótulos em um conjunto de dados linearmente separáveis. No contexto do SVM, as várias fronteiras de separação que conseguem dividir completamente as classes são denominadas hiperplanos. A Figura 6 ilustra essa ideia por meio de um exemplo básico de classificação binária.

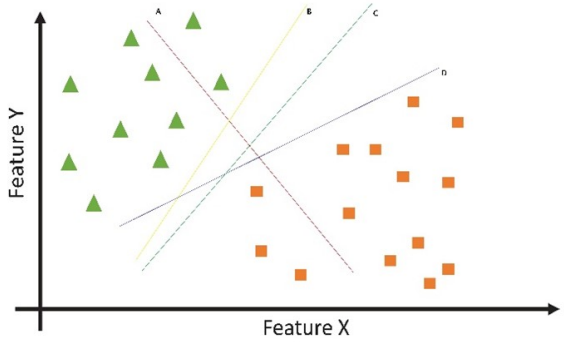

Figura 6 – Possíveis hiperplanos de separação para um conjunto de dados bidimensional e binário.

Desenvolvido para lidar com dados complexos, o SVM destaca-se pela capacidade de encontrar um hiperplano de separação ótimo em espaços multidimensionais. Sua abordagem única envolve a identificação de vetores de suporte, que são os pontos mais próximos do hiperplano de decisão, e a maximização da margem entre classes. A busca pelo hiperplano ótimo não é uma tarefa trivial, envolvendo multiplicadores de Lagrange, derivadas e manipulações matemáticas como problemas de otimização quadráticos.

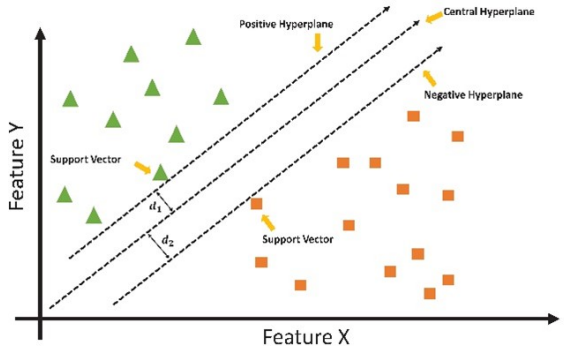

Figura 7 – Melhor hiperplano de separação, margem máxima e vetores de suporte.

Com a capacidade de lidar eficientemente com conjuntos de dados de alta dimensionalidade, o SVM é especialmente eficaz em cenários onde as relações entre as classes não são linearmente óbvias. A aplicação do truque do kernel permite a manipulação de dados não linearmente separáveis, ampliando significativamente o escopo de aplicação do SVM. Nela, o limite de decisão no espaço de entrada é representado por um hiperplano em dimensão superior no espaço (DRUCKER et al., 1996) (SARADHI V., 2005). Esse processo possibilita a identificação de um hiperplano capaz de separar as classes de maneira adequada, conforme mostrado na Figura 8. Ao retornar ao espaço original (de menor dimensão), torna-se visível uma fronteira de separação não linear.

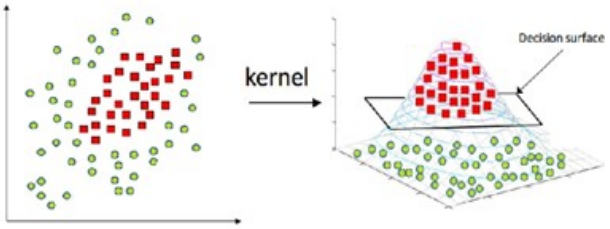

Figura 8 – Intuição por trás da aplicação do kernel trick.

Este modelo se destaca na classificação de dados distribuídos de maneira não regular, pois a separação não requer linearidade e pode variar entre os diferentes conjuntos de dados. O SVM é particularmente atrativo para iniciantes, uma vez que não exige um conhecimento aprofundado da base de dados para realizar previsões com alta precisão. Além disso, o SVM demonstra eficácia em espaços de alta dimensão, lidando bem com conjuntos de dados que possuem muitas características. Sua garantia de convergência para o melhor hiperplano possível é uma característica destacada, diferenciando-o de algoritmos que podem ficar presos em mínimos locais, como ocorre em redes neurais.

No entanto, os resultados do SVM são frequentemente desafiadores de interpretar, embora ainda possível. À medida que o tamanho do conjunto de dados aumenta, o tempo necessário para realizar os cálculos cresce rapidamente, comprometendo a interpretabilidade do modelo de forma ainda mais acentuada.

2.2.2 Árvores de Decisão

As árvores de decisão são modelos de aprendizado de máquina supervisionado que são frequentemente usados para problemas de classificação, embora também, possam ser usados para problemas de regressão. Esses modelos são construídos a partir de um conjunto de dados de treinamento, onde o objetivo é dividir o conjunto de dados em grupos homogêneos. O algoritmo de construção da árvore segue uma abordagem recursiva, onde em cada etapa, ele escolhe a melhor variável (ou atributo) para dividir os dados em subgrupos. Essa escolha é baseada em critérios como entropia, ganho de informação ou índice de Gini. O processo continua até que os

subgrupos sejam suficientemente homogêneos, ou até que algum critério de parada seja atingido. O resultado pode ser visto na Figura 9.

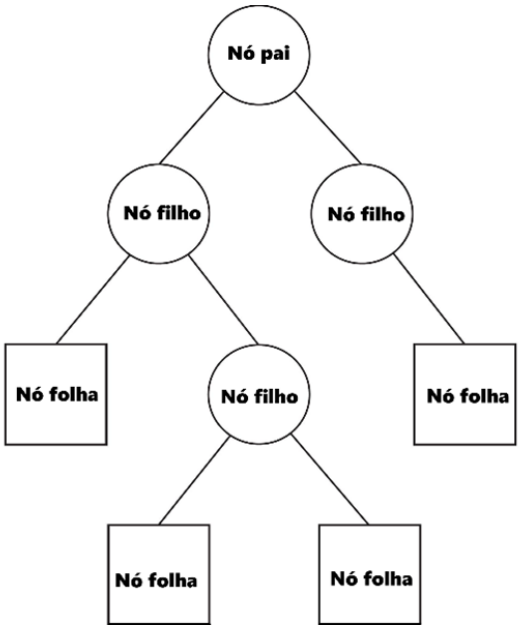

Figura 9 – Exemplo de árvore de decisão, mostrando os tipos de nós.

A entropia, conforme proposta por Shannon (SHANNON, 1948), é uma medida da incerteza ou desordem em um conjunto de dados. Quanto maior a entropia, maior a incerteza associada ao conjunto de dados. Em uma árvore de decisão, a entropia é usada para avaliar o quão homogêneo ou puro é um nó. Na construção de árvores de decisão, a entropia é calculada para cada nó que representa uma divisão nos dados. Seja p_i a proporção de instâncias da classe i em um nó, o cálculo da entropia (E(S)) é dado por:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{2.6}$$

onde c é o número de classes no conjunto de dados e $p_i \neq 0$. A entropia atinge o valor máximo quando as classes são equiprováveis, ou seja, quando todas as p_i são iguais. Ela é mínima quando o nó contém instâncias de apenas uma classe.

Durante esse processo, a entropia é utilizada para avaliar diferentes divisões nos dados. O objetivo é selecionar divisões que reduzam a entropia nos nós filhos, tornando o conjunto mais homogêneo. Essa redução na entropia é conhecida como ganho de informação.

O ganho de informação (G) é uma métrica crucial derivada da entropia. Para um nó pai S e seus nós filhos C1, C2, ..., Ck, o ganho de informação é calculado como:

$$G(S) = E(S) - \sum_{i=1}^{k} \frac{|C_i|}{|S|} E(C_i) \tag{2.7}$$

onde $G(S)$ é a entropia do nó pai, $|C_i|$ é o número de instâncias no nó filho C_i, e $|S|$ é o número total de instâncias no nó pai. A escolha da divisão é guiada pelo maior ganho de informação, indicando a divisão que melhor reduz a incerteza nos dados.

O índice de Gini é outro critério comum, particularmente utilizado no algoritmo CART (Classificação e Regressão por Árvore). É uma medida de quão frequentemente um elemento escolhido aleatoriamente seria incorretamente classificado se fosse rotulado de acordo com a distribuição das classes no nó.

$$G(S) = E(S) - \sum_{i=1}^{k} \frac{|C_i|}{|S|} E(C_i) \tag{2.8}$$

onde p_i é a proporção de instâncias da classe i no nó S. A divisão escolhida é aquela que minimiza a soma ponderada dos índices de Gini dos nós filhos.

Existem vários algoritmos para a construção de árvores de decisão em machine learning, cada um com suas características distintas.

• ID3 (Iterative Dichotomiser 3)

O ID3 foi um dos primeiros algoritmos desenvolvidos para construir árvores de decisão. Proposto por Ross Quinlan (QUINLAN, 1979)(QUINLAN, 1983), o ID3 utiliza o critério de ganho de informação para escolher as divisões nos dados. Ele é eficaz para problemas de classificação com atributos categóricos, mas pode não ser ideal para dados numéricos.

• C4.5

Uma evolução do ID3, o C4.5, também proposto por Ross Quinlan (QUINLAN, 2014), é um algoritmo mais robusto e flexível. Ele suporta tanto atributos categóricos quanto numéricos, além de lidar com dados ausentes. O critério utilizado é o ganho de informação, similar ao ID3, mas o C4.5 introduziu melhorias na forma como trata atributos contínuos.

• CART (Classificação e Regressão por Árvore)

O algoritmo CART foi desenvolvido por Breiman et al (BREIMAN, 2017) e é uma generalização do C4.5. Enquanto o C4.5 é voltado principalmente para classificação, o CART suporta tanto problemas de classificação quanto regressão. Para classificação, utiliza o índice de Gini como critério de divisão, enquanto para regressão, utiliza a redução de variância. O CART também introduziu o conceito de poda para evitar overfitting.

A escolha do algoritmo de árvore de decisão depende das características específicas do problema em questão, como a natureza dos dados, a presença de valores ausentes e a tarefa (classificação ou regressão). Cada algoritmo tem suas vantagens e limitações, e a seleção do mais adequado dependerá das características específicas do conjunto de dados e das metas do projeto.

Ao aplicar um modelo de Árvore de Decisão, espera-se encontrar várias características distintas. Primeiramente, destaca-se a interpretabilidade desse modelo, uma vez que as decisões

em cada nó são fundamentadas em testes lógicos simples aplicados aos atributos. Isso confere ao modelo uma notável capacidade de explicar de maneira transparente como as decisões são tomadas e quais fatores influenciam a classificação.

Além disso, as Árvores de Decisão apresentam uma habilidade inata para identificar relações não-lineares nos dados. Essa capacidade é particularmente valiosa ao lidar com dados faltantes, pois o modelo pode facilmente contornar os testes relacionados a atributos ausentes durante o processo de classificação, contribuindo para uma maior robustez.

Outro ponto relevante é a versatilidade desses modelos. As Árvores de Decisão podem ser empregadas tanto em problemas de classificação quanto em problemas de regressão. Essa característica as torna adequadas para uma variedade de tarefas relacionadas à modelagem de dados, proporcionando flexibilidade e adaptabilidade em diferentes contextos de aplicação.

2.2.3 Random Forest

Random Forest é uma técnica de aprendizado de máquina que opera como um ensemble de árvores de decisão. Desenvolvida para superar algumas das limitações associadas a árvores de decisão individuais, a Random Forest combina as previsões de várias árvores, proporcionando um modelo mais robusto e preciso. O primeiro algoritmo para florestas aleatórias foi criado em 1995 por Tin Kam Ho (HO, 1995).

O princípio de funcionamento da Random Forest envolve o conceito de Bootstrap Aggregating (Bagging). Isso inclui a criação de vários conjuntos de dados de treinamento por meio de amostragem com reposição (bootstrap), sendo cada conjunto utilizado para treinar uma árvore de decisão independente. Além disso, a técnica emprega Random Feature Selection, onde, em cada nó de decisão durante a construção da árvore, apenas um subconjunto aleatório de atributos é considerado para fazer a divisão. Essa seleção aleatória ajuda a decorrelacionar as árvores, aumentando a robustez do modelo.

Conforme pode ser visto na Figura 10, a fase de votação ou média é essencial na Random Forest. Para problemas de classificação, a técnica realiza uma votação entre as árvores para determinar a classe final. Já para problemas de regressão, as previsões das árvores são geralmente médias para obter a previsão final.

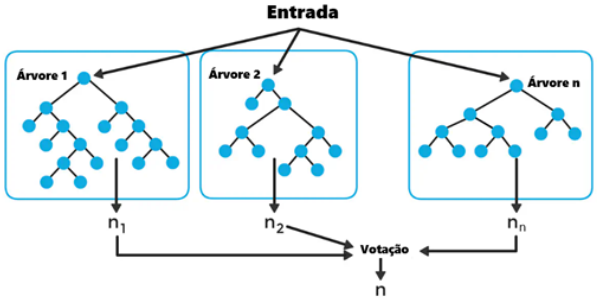

Figura 10 – Exemplo de random forest para problemas de classificação - a classificação final do algoritmo é obtida através da regra da maioria.

Entre as vantagens da Random Forest, destaca-se a redução de overfitting, alcançada pela utilização de várias árvores treinadas em subconjuntos diferentes de dados, aprimorando a generalização do modelo. Além disso, a técnica lida eficazmente com dados desbalanceados, pois combina previsões de várias árvores independentes, proporcionando robustez. A Random Forest é também robusta a outliers e ruídos nos dados, baseando suas decisões na maioria das previsões. Além disso, fornece uma medida de importância para cada atributo, facilitando a seleção de características relevantes. No entanto, a Random Forest enfrenta desafios, como a interpretabilidade reduzida devido à sua natureza ensemble. Além disso, o custo computacional pode ser significativo ao treinar várias árvores e realizar a combinação, especialmente para conjuntos de dados grandes.

A Random Forest encontra aplicação em uma variedade de cenários, incluindo classificação e regressão. Sua versatilidade estende-se à detecção de anomalias, sendo capaz de identificar padrões incomuns em dados. Adicionalmente, a importância de atributos calculada pela técnica pode ser utilizada na seleção de características. Em conclusão, a Random Forest emergiu como uma abordagem popular para a construção de modelos preditivos, proporcionando uma solução eficaz para desafios associados a árvores de decisão individuais. Sua capacidade de lidar com diversas complexidades, aliada à capacidade de fornecer uma visão da importância dos atributos, torna-a uma escolha valiosa em muitos cenários de aprendizado de máquina.

2.2.4 Boosting

Boosting (SCHAPIRE, 1990) é uma técnica de ensemble em machine learning projetada para melhorar o desempenho de modelos preditivos, focando em aprender com os erros dos modelos anteriores. Ao contrário do Random Forest, que cria várias árvores de decisão independentes, o Boosting constrói um conjunto sequencial de modelos, cada um corrigindo as deficiências do anterior.

No processo de Boosting, cada instância no conjunto de dados é inicialmente atribuída um peso igual. Durante o treinamento, o algoritmo dá mais peso às instâncias mal classificadas nos modelos anteriores, incentivando o modelo a corrigir esses erros. São utilizados modelos

fracos, geralmente árvores de decisão rasas, como classificadores base. Um modelo fraco é aquele que tem um desempenho apenas ligeiramente melhor do que o aleatório.

Cada modelo contribui para a previsão final com um peso proporcional à sua precisão, sendo modelos mais precisos mais influentes na previsão final. O processo de adaptação progressiva continua por várias iterações até que a melhoria na previsão atenda a um critério de parada, conforme ilustrado pela Figura 11.

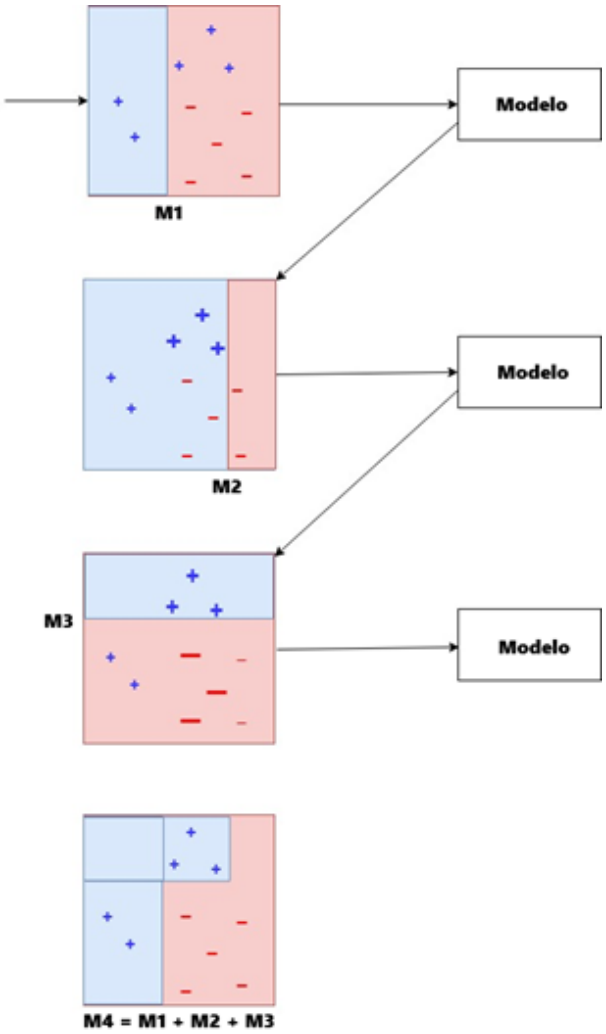

Figura 11 – Possíveis hiperplanos de separação para um conjunto de dados bidimensional e binário.

Alguns dos algoritmos de Boosting mais populares incluem o AdaBoost (Adaptive Boosting) (FREUND; SCHAPIRE, 1997), que ajusta os pesos das instâncias a cada iteração, focando nas instâncias classificadas incorretamente pelos modelos anteriores, e o Gradient Boosting (FRIEDMAN, 2001), um framework mais geral que inclui algoritmos como XGBoost (CHEN; GUESTRIN, 2016), LightGBM (KE et al., 2017) e CatBoost (PROKHORENKOVA et al., 2018). O Gradient Boosting treina modelos sequencialmente, ajustando os resíduos do modelo anterior.

Entre as vantagens do Boosting, destaca-se a melhoria incremental na precisão do modelo, a capacidade de lidar eficazmente com conjuntos de dados desbalanceados e a adaptação a dados complexos e não lineares. No entanto, o Boosting enfrenta desafios, sendo sensível a outliers e exigindo custos computacionais mais elevados, especialmente em conjuntos de dados grandes.

Esta técnica encontra aplicação em uma variedade de cenários, incluindo classificação, regressão, detecção de anomalias e problemas de ranking, destacando sua versatilidade e eficácia na construção de modelos preditivos mais robustos e precisos.

3 METODOLOGIA

3.1 Banco de dados

O conjunto de dados escolhido para esse trabalho, recebe o título de Credit Card Churn Prediction (SAN, 2022). A base de dados possui 10.126 observações e 20 variáveis, sendo elas:

- **Bandeira de Atrito** – Variável binária que define o estado do cliente. “Attrited Customer” para “Cliente Demitido” e “Existing Customer” para “Cliente Existente;

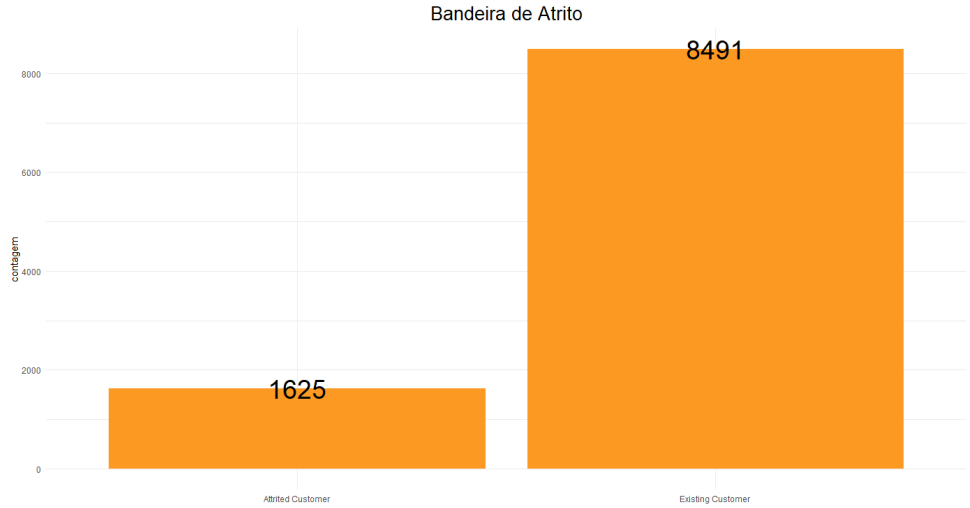

Figura 12 – Descrição gráfica da Variável “Bandeira de Atrito”.

- **Idade do Cliente** – Variável numérica que descreve a idade dos clientes observados. Variando de 26 até 73 anos, com média 46;

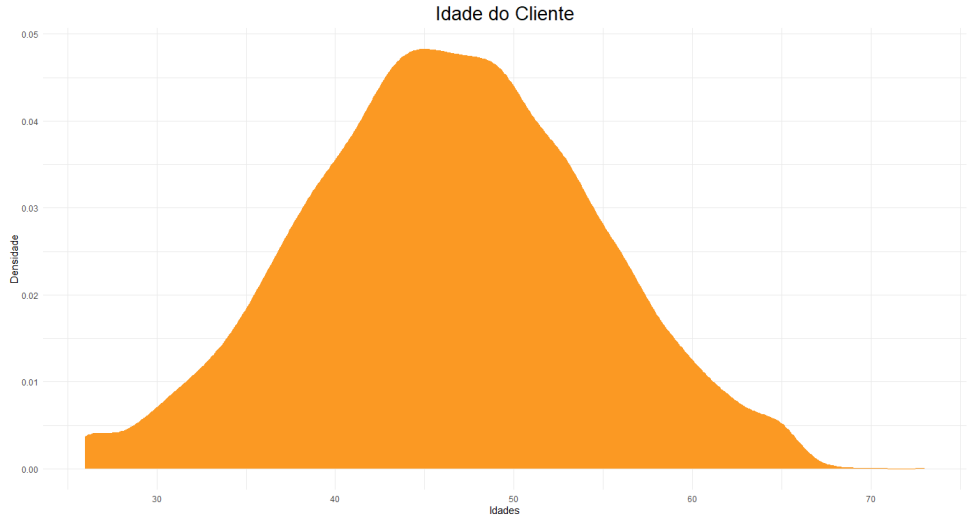

Figura 13 – Descrição gráfica da Variável “Idade”.

- **Gênero** – Variável que descreve o gênero dos clientes analisados, sendo “M” para Masculino e “F” para feminino.

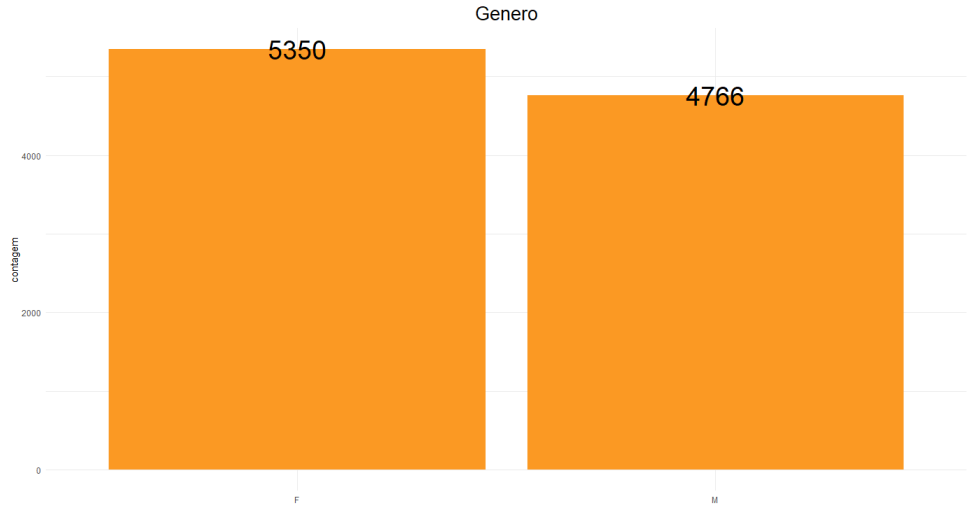

Figura 14 – Descrição gráfica da Variável “Gênero”.

- **Dependentes** – Variável numérica que descreve a quantidade de dependentes de cada cliente analisado. Variando de 0 a 5.

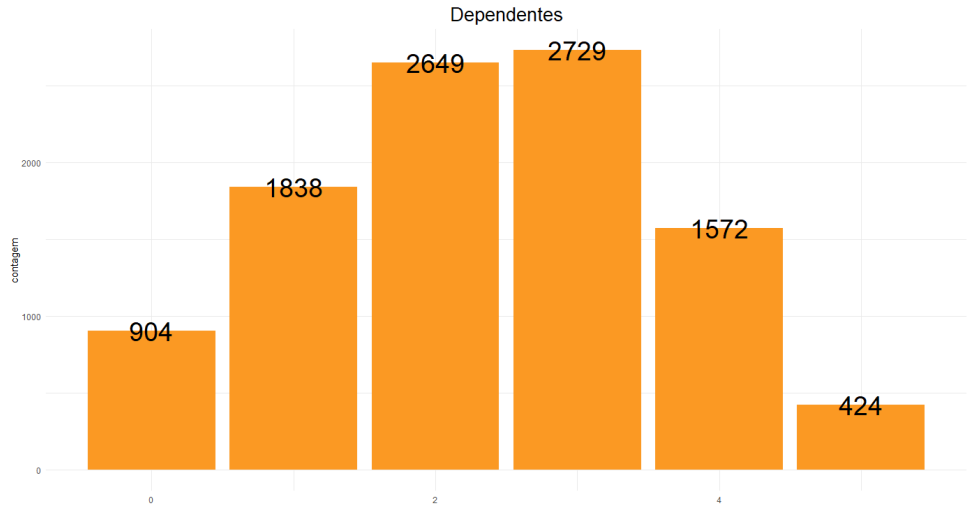

Figura 15 – Descrição gráfica da Variável “Dependentes”.

- **Escolaridade** – Variável categórica que descreve a escolaridade dos clientes analisados. Essa variável está dividida em 7 níveis, sendo eles: College (Formação profissional especializada); Doctorate (Doutorado); Graduate (Graduação); High School (Ensino Médio); Post-Graduate (Pós-Graduação); Uneducated (Analfabeto) e Unknown (Desconhecido).

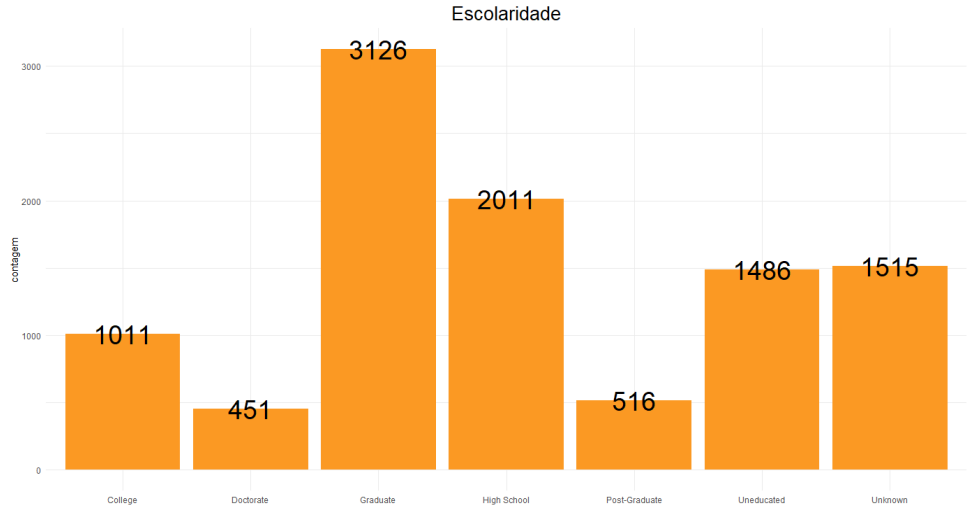

Figura 16 – Descrição gráfica da Variável "Escolaridade".

- **Estado Civil** – Variável categórica que descreve o estado civil dos clientes analisados. Essa variável possui os seguintes níveis: Divorced (Divorciado); Married (Casado); Single (Solteiro) e Unknown (Desconhecido).

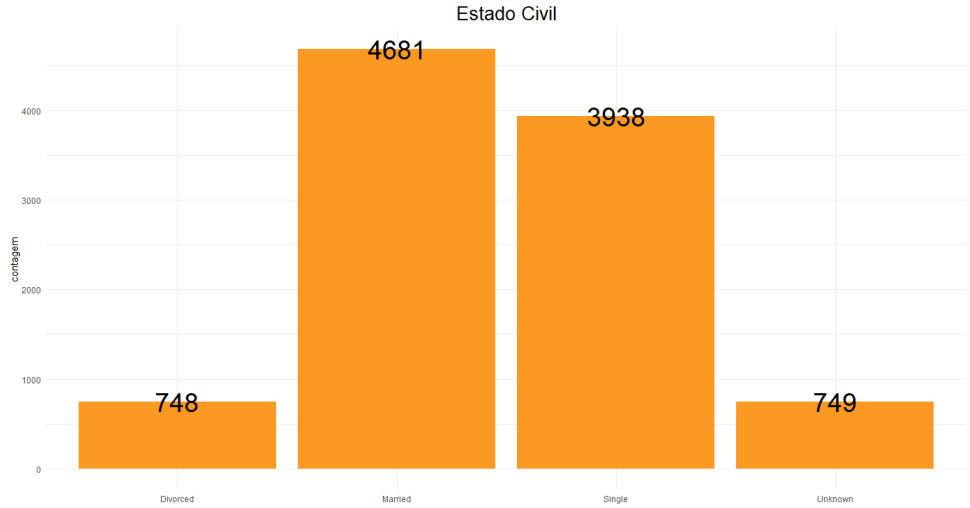

Figura 17 – Descrição gráfica da Variável "Estado Civil".

- **Categoria de Renda** - Variável categórica que classifica a renda anual dos clientes analisados.

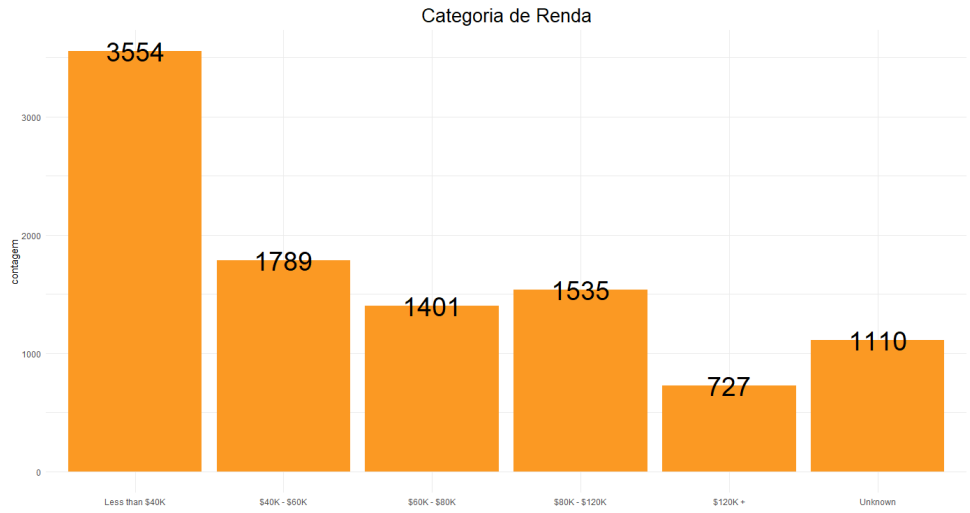

Figura 18 – Descrição gráfica da Variável “Categoria de Renda”.

- **Categoria do Cartão** - Variável categórica que descreve a linha do cartão usado pelo cliente.

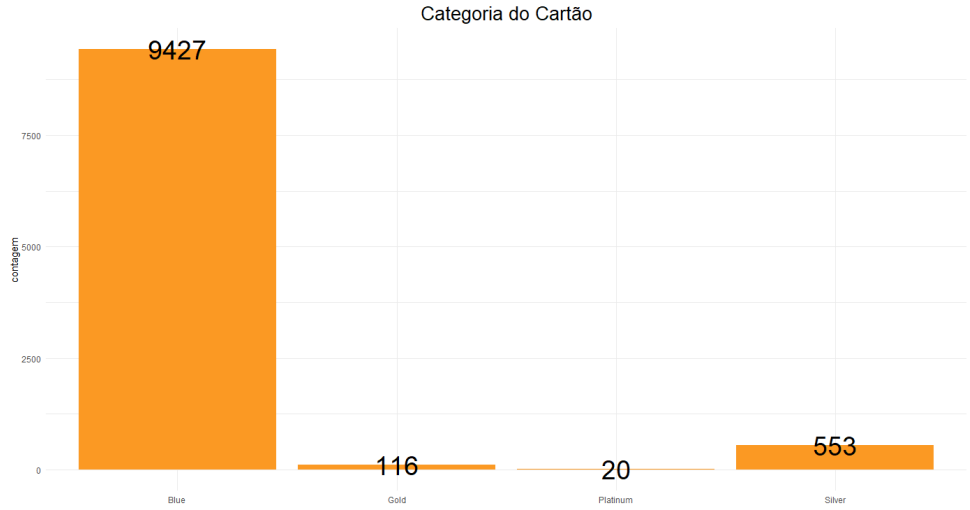

Figura 19 – Descrição gráfica da Variável “Categoria do Cartão”.

- **Meses no Livro** - Variável numérica que descreve o período de relacionamento com o banco em meses. Esse período vai de 13 até 56 meses, com média 36.

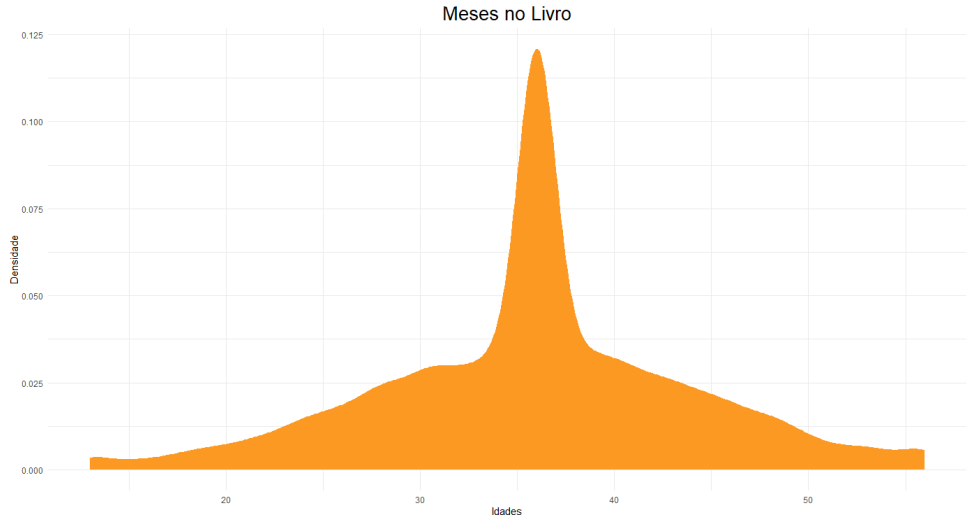

Figura 20 – Descrição gráfica da Variável “Meses no Livro”.

- **Contagem Total de Relacionamento** - Variável numérica que descreve o número de produtos detidos pelo cliente. Os valores variam de 1 a 6 produtos, com média 4.

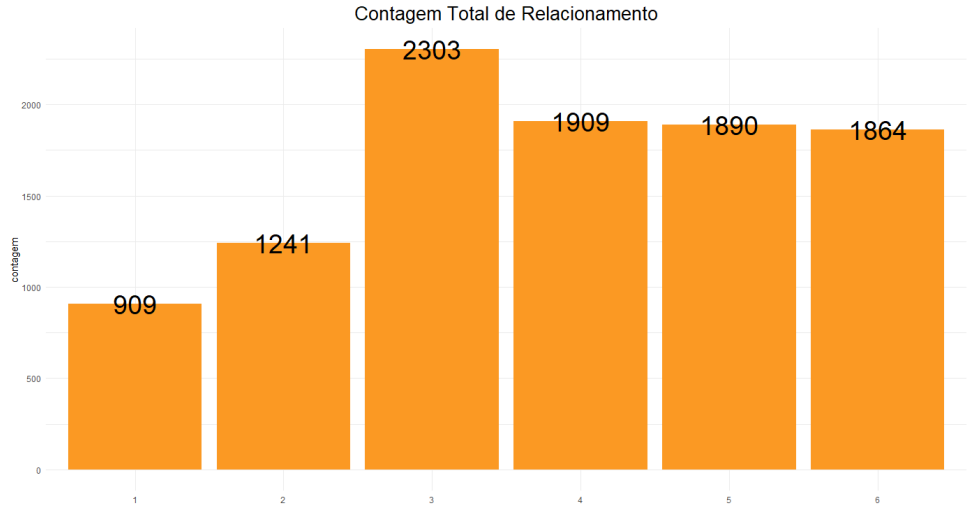

Figura 21 – Descrição gráfica da Variável “Contagem total de Relacionamento”.

- **Meses Inativos** - Variável que apresenta o número de meses inativos de um cliente no intervalo de 1 ano.

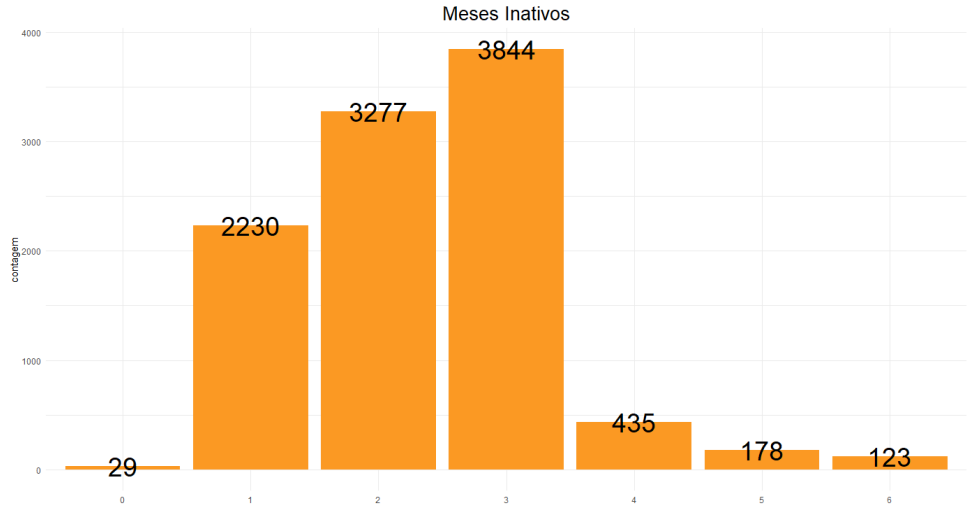

Figura 22 – Descrição gráfica da Variável “Meses Inativos”.

- **Contagem de Contatos** - Número de contatos nos últimos 12 meses.

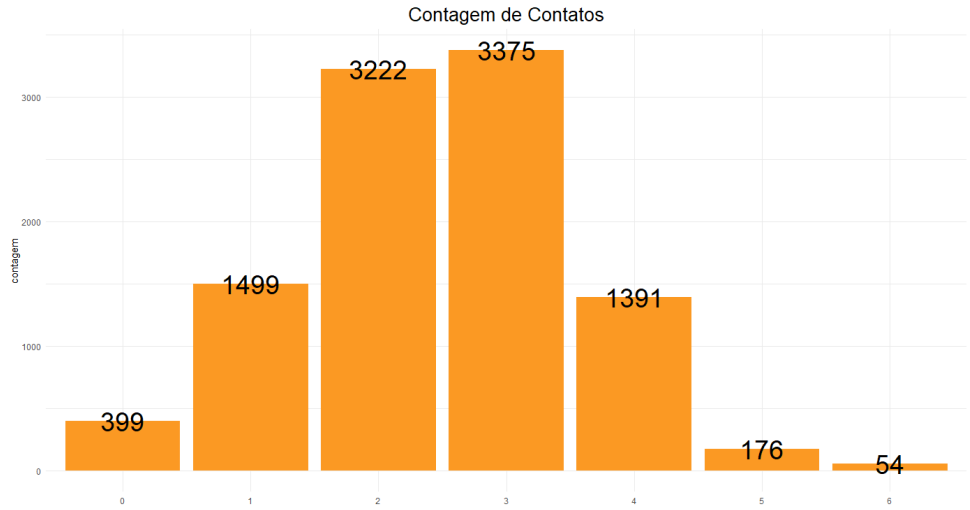

Figura 23 – Descrição gráfica da Variável “Contagem de contatos”.

- **Limite de Crédito** - Limite de crédito do cliente em questão. Esse valor varia de $1.438 até $34.516 com média $8.634,00.
- **Bola Giratória Total** - Saldo rotativo total no cartão de crédito. Esse valor varia de $ 0 até $ 2.517, com média $ 1.663,00.
- **Média Aberta Para Conta** - Linha de crédito aberta para compra. Média dos últimos 12 meses. Esse valor varia entre $3,00 e $ 34.516 com média $7.470,00.
- **Total-Amt-Chng-Q4-Q1** – Mudança no valor total das transações do cliente entre o quarto trimestre do ano anterior e o primeiro trimestre do ano atual. Esse valor varia entre 0 e 3,3970 com média 0,7599.

- **Total-Trans-Amt** – Valor total das transações realizadas pelo cliente nos últimos 12 meses. Esse valor varia de $510,00 a $18.484,00 com média $4.404,00.
- **Total-Trans-Ct** – Quantidade total das transações realizadas pelo cliente nos últimos 12 meses. Esse valor varia de 10 a 139 com média 64.86.
- **Total-Ct-Chng-Q4-Q1** – Representa a mudança na quantidade total de transações do cliente entre o quarto trimestre do ano anterior e o primeiro trimestre atual. Esse valor varia de 0 a 3.7140 com média 0.7122.
- **Avg-Utilization-Ratio** – É a proporção média do limite de crédito do cliente que foi usado os últimos 12 meses. Esse valor varia de 0 a 0,999 com média 0,275.

Variável	**Mínimo**	**Média**	**Máximo**
Limite de crédito	$1.438,30	$8634,00	$34.516,00
Bola giratória total	$0,00	$1.663,00	$2.517,00
Média aberta para conta	$3,00	$7.470,00	$34.516,00
Total-Trans-Amt	$540,00	$4.404,00	$18.484,00
Total-Amt-Chg-Q4-Q1	0	0,7599	3,3970
Total-trans-Ct	10	64,86	139
Total-Ct-Chng-Q4-Q1	0	0,7122	3,7140
Avg-Utilization-Ratio	0	0,275	0,999

3.2 Métricas para avaliação dos modelos

A avaliação de modelos de classificação é uma etapa crucial no desenvolvimento de sistemas de aprendizado de máquina, pois fornece insights sobre o desempenho do modelo e sua capacidade de generalização para dados não vistos. Diversas métricas e técnicas são utilizadas para avaliar a qualidade de modelos de classificação, cada uma fornecendo uma perspectiva diferente sobre seu desempenho.

Uma ferramenta fundamental para essa avaliação é a matriz de confusão, mostrada na Tabela 1, que resume o desempenho do modelo em termos de verdadeiros positivos (TP), falsos positivos (FP), verdadeiros negativos (TN) e falsos negativos (FN). Com base nesses valores, outras métricas podem ser calculadas.

Tabela 1 – Matriz de confusão.

		Classe Predita	
		Positiva	Negativa
Real	Positiva	VP	FN
	Negativa	FP	VN

3.2.1 Acurácia

A acurácia (ACC) é uma métrica fundamental na avaliação de modelos de classificação, proporcionando uma medida simples e intuitiva da precisão global do modelo. Essa métrica indica a proporção de predições corretas em relação ao total de predições feitas pelo modelo, sendo calculada pela fórmula:

$$ACC = \frac{VP + VN}{VP + FP + VN + FN} \tag{3.1}$$

Em contextos em que as classes no conjunto de dados são balanceadas, ou seja, têm uma distribuição relativamente igual, a acurácia fornece uma visão geral da capacidade do modelo em realizar previsões corretas.

A principal vantagem da acurácia reside em sua interpretação simples e ampla aplicabilidade a diferentes tipos de problemas de classificação. No entanto, é importante destacar suas limitações. Em conjuntos de dados desbalanceados, onde uma classe é muito mais prevalente que a outra, a acurácia pode ser enganosa, pois um modelo que simplesmente prevê a classe majoritária pode apresentar uma acurácia alta, mas não oferecer um desempenho útil. Nesse contexto, outras métricas como precisão, recall, F1-Score ou a curva ROC podem oferecer uma avaliação mais abrangente e informativa do desempenho do modelo.

3.2.2 Precisão

A precisão quantifica a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias que o modelo previu como positivas. A fórmula para calcular a precisão é dada por:

$$Preciso = \frac{TP}{TP + FP} \tag{3.2}$$

Essa métrica é particularmente relevante em situações em que os falsos positivos têm implicações significativas ou custos mais elevados do que os falsos negativos. Em cenários onde é crucial minimizar as previsões incorretas de instâncias negativas como positivas, a precisão desempenha um papel crucial. Uma das vantagens da precisão é que ela se concentra na qualidade das previsões positivas, proporcionando uma métrica clara e intuitiva. No entanto, é importante observar que a precisão não leva em conta os casos em que o modelo não consegue identificar instâncias positivas, ou seja, os falsos negativos.

A aplicação adequada da precisão ocorre em problemas nos quais evitar falsos positivos é crítico, e a ênfase está na qualidade das previsões positivas, não na cobertura total das instâncias positivas. Por outro lado, em situações em que os falsos negativos têm um impacto mais prejudicial

do que os falsos positivos, ou em conjuntos de dados desbalanceados, a precisão pode não fornecer uma visão completa do desempenho do modelo.

3.2.3 Sensibilidade

A sensibilidade, também conhecida como recall ou taxa de verdadeiros positivos, é uma métrica crucial na avaliação de modelos de classificação em análise de dados. Essa métrica quantifica a capacidade do modelo de identificar corretamente todas as instâncias positivas presentes no conjunto de dados, sendo calculada pela fórmula:

$$Sensibilidade = \frac{VP}{VP + FN} \tag{3.3}$$

A sensibilidade é particularmente relevante em situações em que os falsos negativos têm consequências significativas ou custos mais elevados do que os falsos positivos. Em cenários onde é crucial identificar corretamente todas as instâncias positivas e minimizar a ocorrência de casos em que o modelo deixa de reconhecer uma instância positiva, a sensibilidade desempenha um papel fundamental.

Entre as vantagens da sensibilidade está a sua ênfase na identificação correta de positivos verdadeiros, tornando-a valiosa em contextos onde a perda associada aos falsos negativos é mais crítica. Além disso, é uma métrica robusta em conjuntos de dados desbalanceados, nos quais uma classe é muito mais prevalente que a outra.

No entanto, é importante observar que a sensibilidade não leva em conta os casos em que o modelo erroneamente prevê instâncias negativas como positivas, ou seja, os falsos positivos, tornando-a adequada em problemas onde os falsos negativos têm um impacto mais prejudicial do que os falsos positivos e em situações onde a identificação correta de todas as instâncias positivas é crucial. Contudo, ao interpretar os resultados, é essencial considerar o contexto específico do problema e ponderar a importância relativa de evitar falsos negativos em comparação com falsos positivos. Em alguns casos, pode ser necessário combinar a sensibilidade com outras métricas, para uma avaliação abrangente do desempenho do modelo de classificação.

3.2.4 Especificidade

Enquanto a sensibilidade foca na capacidade do modelo de identificar corretamente todas as instâncias positivas, a especificidade está relacionada à capacidade de identificar corretamente todas as instâncias negativas no conjunto de dados. A fórmula para calcular a especificidade é dada por:

$$Especificidade = \frac{VN}{VN + FP} \tag{3.4}$$

A especificidade torna-se especialmente relevante em situações em que os falsos positivos têm consequências significativas ou custos mais elevados do que os falsos negativos. Ela é frequentemente usada em conjunto com a sensibilidade para obter uma avaliação mais equilibrada do desempenho do modelo em relação a ambas as classes.

As vantagens da especificidade incluem sua ênfase na identificação correta de instâncias negativas, sendo valiosa em cenários onde evitar falsos positivos é crucial. Além disso, sua utilização em conjunto com a sensibilidade fornece uma visão equilibrada do desempenho do modelo para ambas as classes. No entanto, assim como a sensibilidade, a especificidade não leva em conta os casos em que o modelo erroneamente prevê instâncias positivas como negativas, ou seja, os falsos negativos.

3.2.5 Taxa de Falso Positivo

A Taxa de Falso Positivo (TFP), também conhecida como taxa de alarme falso, é uma métrica importante em modelos de classificação, medindo a proporção de instâncias negativas erroneamente classificadas como positivas em relação ao total de instâncias negativas reais. A fórmula para calcular a Taxa de Falso Positivo é dada por:

$$TFP = \frac{FP}{FP + VN} \tag{3.5}$$

Esta métrica é particularmente relevante em situações em que os custos ou impactos associados aos falsos positivos são significativos ou indesejados. Por exemplo, em sistemas de detecção de fraudes, um falso positivo implica identificar erroneamente uma transação legítima como fraudulenta, podendo causar inconvenientes para o usuário.

A Taxa de Falso Positivo oferece uma medida específica para avaliar a frequência com que o modelo comete o erro de predizer incorretamente instâncias negativas como positivas. Seu uso é apropriado em contextos onde a minimização de falsos positivos é crucial, como em sistemas de segurança ou diagnósticos médicos.

3.2.6 F1-score

Trata-se de uma métrica que combina precisão e sensibilidade para fornecer uma medida única do desempenho de um modelo de classificação. Essa métrica é particularmente útil em situações onde há desbalanceamento entre as classes do conjunto de dados, ou seja, quando uma classe é muito mais prevalente que a outra. O F1-Score é calculado pela fórmula:

$$F1 - Score = 2 \times \frac{Preciso \times Recall}{Preciso + Recall} \tag{3.6}$$

As vantagens do F1-Score incluem seu papel de equilibrar precisão e sensibilidade, tornando-o adequado para cenários onde ambas as métricas são críticas e seu desequilíbrio pode levar a interpretações inadequadas do desempenho do modelo. Além disso, o F1-Score é robusto em situações de classes desbalanceadas, evitando interpretações enviesadas em direção à classe majoritária. No entanto, é importante notar que o F1-Score não distingue entre falsos negativos e falsos positivos separadamente. Em particular, é uma métrica recomendada em situações de desbalanceamento de classes, onde outras métricas podem gerar interpretações enviesadas.

3.2.7 Índice Kappa

O índice kappa, também conhecido como coeficiente kappa ou estatística kappa, é uma métrica particularmente útil quando se lida com problemas nos quais as classes podem não estar balanceadas.

Essencialmente, o kappa é uma medida de concordância entre as classificações previstas por um modelo e as classificações reais. Ele ajusta a taxa de concordância observada pela concordância que poderia ser esperada ao acaso. Dessa forma, o índice kappa leva em consideração a possibilidade de que o modelo possa acertar por acaso, proporcionando uma medida mais robusta do desempenho do modelo.

A fórmula para o índice kappa é dada por

$$Kappa = \frac{Po - Pe}{1 - Pe} \tag{3.7}$$

onde P_0 é a taxa de concordância observada entre as previsões do modelo e as verdadeiras classes, e Pe é a taxa de concordância esperada ao acaso, calculada com base nas proporções marginais de cada classe. O índice kappa varia de -1 a 1, onde:

• Kappa = 1 indica concordância perfeita entre as previsões do modelo e as classes reais. • Kappa = 0 indica concordância equivalente ao acaso. • Kappa = 1 indica discordância completa entre as previsões do modelo e as classes reais.

Geralmente, um valor de kappa superior a 0,6 é considerado razoável, indicando um nível substancial de concordância além do que seria esperado ao acaso. No entanto, a interpretação exata pode variar dependendo do contexto específico do problema.

O uso do índice kappa é especialmente valioso em tarefas de classificação com classes desbalanceadas, onde a acurácia por si só pode ser enganosa. Ao incorporar a concordância ao acaso, o kappa fornece uma medida mais confiável da qualidade do modelo.

3.2.8 Curva ROC

A Curva ROC (Receiver Operating Characteristic), ilustrada na Figura 24, é uma ferramenta gráfica usada para avaliar o desempenho de modelos de classificação binária em diferentes pontos de corte. Ela representa a taxa de verdadeiros positivos (sensibilidade) em relação à taxa de falsos positivos em vários níveis de limiar de decisão.

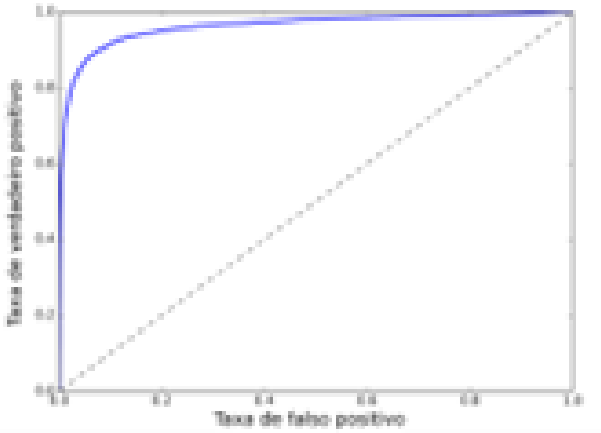

Figura 24 – Curva ROC.

A representação gráfica da Curva ROC é construída com o eixo x representando a Taxa de Falsos Positivos (TFP) e o eixo y representando a Taxa de Verdadeiros Positivos (TVP), que é uma designação alternativa para a sensibilidade. Cada ponto na curva corresponde a um limiar de decisão específico no modelo de classificação. O ponto ideal na curva ROC está no canto superior esquerdo, onde a sensibilidade é 1 (todos os verdadeiros positivos são capturados) e a taxa de falsos positivos é 0 (nenhum falso positivo).

A área sob a Curva ROC (AUC-ROC) é uma métrica resumida derivada da curva, indicando o desempenho global do modelo em discriminar entre as classes. Confome ilustrado na Figura 25, quanto maior a AUC-ROC, melhor é o desempenho do modelo. Se a AUC-ROC for 0.5, isso sugere que o modelo não é superior a um classificador aleatório.

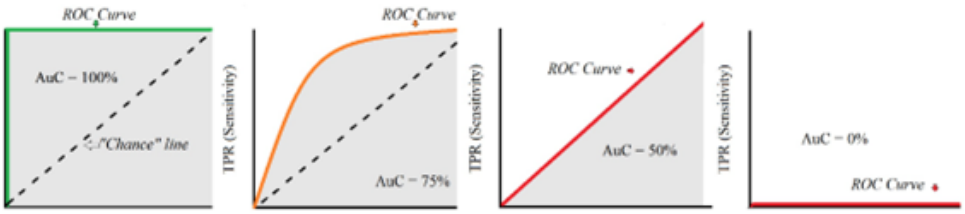

Figura 25 – Área sob a curva ROC em diferentes níveis.

A Curva ROC é vantajosa por oferecer uma visualização intuitiva do desempenho do modelo em diferentes limiares de decisão, sendo especialmente robusta em situações de desbalanceamento de classes. No entanto, é importante notar que a Curva ROC assume que os erros de classificação têm custos semelhantes, o que pode não ser verdade em todos os cenários.

Apesar de sua utilidade, a Curva ROC não aborda desequilíbrios internos nas classes e pode ser insensível a custos desiguais associados a falsos positivos e falsos negativos. Assim, a interpretação dos resultados deve levar em consideração o contexto específico do problema e as prioridades associadas a diferentes tipos de erros.

Essas medidas de avaliação são métricas que possuem um caráter global, sendo uma boa ferramenta para comparação equivalente entre os modelos aplicados. A escolha e interpretação dessas medidas são cruciais para uma análise completa e informada do desempenho dos modelos. Essas métricas não apenas informam sobre a qualidade das previsões, mas também fornecem insights valiosos para orientar decisões estratégicas na gestão de clientes e retenção.

3.2.9 Explicabilidade de modelos

O Índice SHAP (Shapley Additive exPlanations), desenvolvido por Lundberg et al (LUNDBERG; LEE, 2017) é uma técnica avançada no campo de explicabilidade de modelos de aprendizado de máquina, especialmente em contextos de classificação. Sua principal contribuição reside na capacidade de atribuir importância individual a cada variável em um modelo, permitindo uma interpretação mais detalhada e transparente das decisões preditivas.

Essa abordagem única é fundamentada na teoria dos jogos cooperativos, mais especificamente no valor de Shapley, que busca distribuir de maneira justa o ganho total de um jogo entre os jogadores. No contexto do SHAP, os "jogadores"são as variáveis preditoras do modelo.

O SHAP realiza uma decomposição aditiva para explicar a contribuição de cada variável. Isso significa que a contribuição de cada variável é descrita como a diferença entre o valor predito para a instância completa e o valor médio predito para todas as instâncias. Essa abordagem aditiva facilita a compreensão de como cada variável contribui para a mudança nas previsões. A Figura 26 ilustra o método, onde os atributos idade, gênero, pressão arterial e índice de massa corporal (IMC) têm valores de contribuições +0.4, -0.3, +0.1 e +0.1 para a saída apresentada.

Basicamente, os valores de Shapley calculam a importância de um recurso comparando o que um modelo prevê com e sem o recurso. No entanto, como a ordem na qual um modelo vê recursos pode afetar suas previsões, isso é feito em todas as ordens possíveis, para que os recursos sejam comparados de maneira justa. Uma das grandes vantagens do SHAP é sua capacidade de explicar tanto a importância global quanto local das variáveis. Em termos globais, é possível avaliar a influência de cada variável em todo o conjunto de dados, proporcionando uma visão abrangente das características mais relevantes. Em termos locais, o SHAP permite

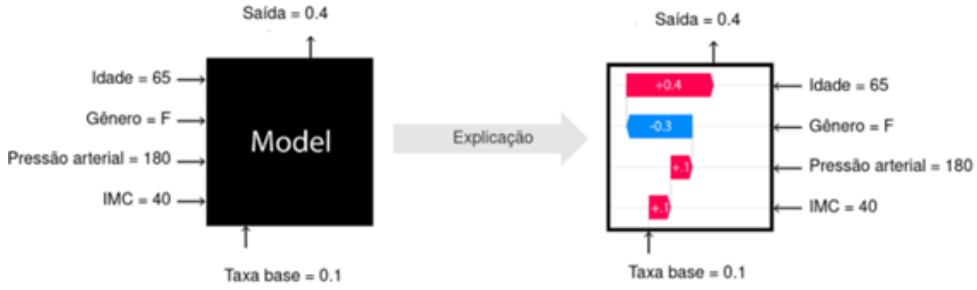

Figura 26 – Ilustração da explicação com SHAP. Adaptado (MITCHELL; FRANK; HOLMES, 2022).

entender como cada variável afeta a predição para uma instância específica, proporcionando uma interpretabilidade mais refinada e individualizada.

A visualização desse processo muitas vezes é realizada por meio de gráficos de valores SHAP, que destacam a contribuição de cada variável para cada predição. Essa representação gráfica não só facilita a interpretação, mas também é uma ferramenta eficaz para comunicar análises de importância a partes interessadas não técnicas. Além disso, o SHAP é uma técnica model-agnostic, o que significa que pode ser aplicado a diversos tipos de modelos de aprendizado de máquina, desde árvores de decisão até modelos lineares e redes neurais. Essa flexibilidade amplia a aplicabilidade do SHAP em uma variedade de cenários.

Neste contexto, o Índice SHAP destaca-se como uma ferramenta poderosa para a interpretabilidade de modelos de classificação, fornecendo uma compreensão detalhada e transparente das contribuições individuais das variáveis. Sua aplicação é valiosa em situações em que a explicação do modelo é crucial, seja para a tomada de decisões informadas ou para a comunicação eficaz com a equipe de negócios.

4 RESULTADOS

Definidas as medidas de avaliação obtidas na etapa de treinamento dos modelos, podemos agora analisar os primeiros resultados a partir do gráfico da Figura 27.

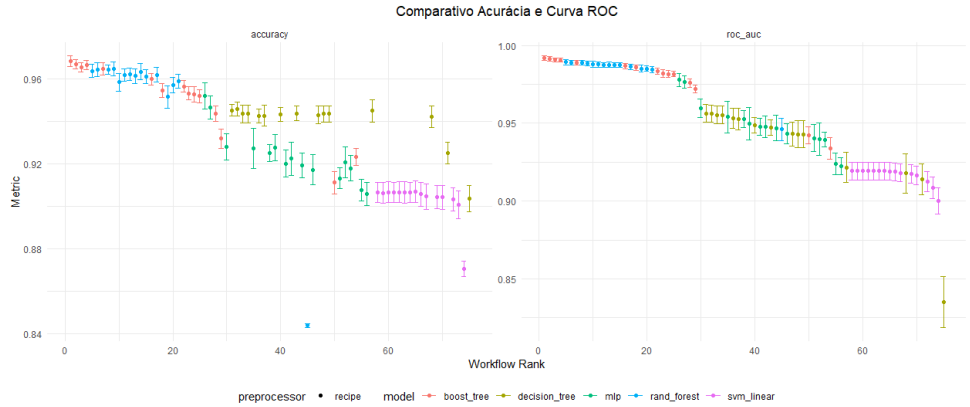

Figura 27 – Comparativo entre acurácia e área sob a curva roc obtidas através da especificação de cinco modelos de classificação.

A acurácia representa a proporção de previsões corretas em relação ao total de observações. Em termos simples, é a medida de quão preciso é o modelo em sua classificação geral. A curva ROC-AUC é uma representação gráfica dos verdadeiros positivos (sensibilidade) em relação à taxa de falsos positivos. No gráfico apresentado na Figura 28 podemos observar que os modelos Boosted (boost-tree) e Random Forest (rand-forest) obtiverem os melhores resultados, apresentando os mais altos índices em ambos os casos. É importante observar também que o modelo Randon Forest apresenta uma menor dispersão no gráfico, apresentando valores mais concentrados.

No gráfico da Figura 28 podemos ver a área sob a curva ROC (roc-auc) de maneira isolada o melhor índice para cada modelo:

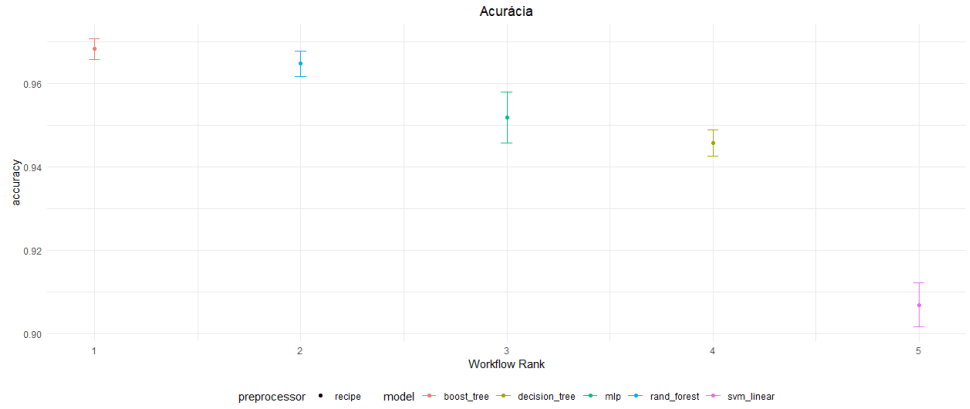

Figura 28 – Área sob a curva ROC obtida para o melhor modelo estimado em cada método proposto.

Pode ser observado que, de acordo com essa métrica (roc-auc), o modelo que melhor se ajustou ao conjunto de treinamento foi o Boosting, de forma que este modelo foi utilizado para o prosseguimento da análise, expondo-o ao conjunto de dados de testes. O conjunto de testes foi formado por um total de 2530 elementos amostrais selecionados aleatoriamente e não foi apresentado ao modelo durante a fase de treinamento. Neste conjunto de dados, aproximadamente 16% (407 observações) dos indivíduos pertencem à classe Attrited Customer (cliente desiste de usar o cartão de crédito) e aproximadamente 84% (2123 observações) pertencem à classe Existing Customer (cliente continua usando o cartão) da variável alvo Bandeira de Atrito. Após a aplicação do modelo vencedor na base de dados de teste, a seguinte matriz de confusão foi gerada e apresentada na Tabela 2:

Tabela 2 – Matriz de confusão oriunda da aplicação do melhor modelo (boosting) ao conjunto de testes.

		Classe Predita	
		Attrited Customer	Existing Customer
Real	Attrited customer	359	48
	Existing customer	30	2093

A partir da matriz de confusão apresentada na Tabela 1, as métricas de avaliação do modelo foram calculadas, considerando a classe “Attrited Customer” como classe positiva. Os resultados podem ser visualizados na Tabela 3:

Tabela 3 – Resultados da aplicação do algoritmo Boosting ao conjunto de dados de teste.

Métricas de Avaliação	**Estimativa**
Acurácia	0,969
Precisão	0,923
Snsibilidade	0,882
Especificidade	0,986
Taxa de Falso Positivo (TFP)	0,014
F1-Score	0,902
kappa	0,884

Podemos observar, em geral, um bom desempenho do algoritmo quando avaliado utilizando a base de dados teste. A acurácia do modelo, isto é, a proporção de predições corretas em relação ao total de predições feitas pelo modelo foi de 0,969, indicando que

O modelo classificou corretamente 96,9% das amostras, tanto para os clientes que desistem quanto para aqueles que não desistem do uso do cartão de crédito. Isso sugere uma boa capacidade geral do modelo em classificar os casos. A precisão do modelo, ou seja, a proporção de clientes corretamente classificadas como Attrited Customer em relação ao total de clientes que o modelo previu como Attrited Customer, foi de 0,923, indicando que o modelo conseguiu identificar corretamente 92,3% dos clientes classificados como possíveis desistentes do uso de cartão de crédito. Isso indica que a instituição financeira pode confiar nas previsões positivas do modelo ao

tomar decisões relacionadas à gestão de clientes ou campanhas direcionadas àqueles propensos a desistir do uso do cartão. Daqueles clientes que verdadeiramente deixaram de utilizar os serviços de cartão de crédito, o modelo conseguiu identificar corretamente 88,2% (sensibilidade = 0,882), indicando que o modelo é eficaz em detectar a maioria dos casos em que os clientes estão realmente desistindo do uso do cartão. Isso é importante porque, ao antecipar essas situações, as instituições financeiras podem tomar medidas proativas, como oferecer incentivos ou melhorar o atendimento ao cliente, para tentar reter esses clientes antes que eles abandonem completamente o uso do cartão. Já para os clientes que verdadeiramente possuem relação de crédito com o banco, o modelo consegue identificá-los com 98,6% de probabilidade. Uma alta especificidade significa que o modelo está se concentrando em clientes que têm uma probabilidade real de desistir, direcionando esforços de retenção de forma mais eficiente e direcionada. A taxa de alarme falso, ou seja, a probabilidade de o modelo classificar um cliente que utiliza o cartão de crédito normalmente e não tem a intenção de deixar de usar como um possível desistente (TFP) foi de apenas 1,4%, minimizando ações indesejadas, como oferecer incentivos ou promoções a clientes que não têm a intenção real de desistir do uso do cartão. Ao analisarmos a estimativa da métrica F1-score igual a 0,902 podemos afirmar que o modelo está alcançando um equilíbrio entre a identificação correta dos clientes que desistiram do cartão de crédito (positivos) e a minimização de falsos positivos (clientes erroneamente classificados como desistentes). Isso indica que o modelo é capaz de identificar clientes que estão propensos a desistir sem gerar muitos falsos positivos. Por fim, uma estimativa para o índice kappa igual a 0,884 indica uma concordância substancial entre as previsões do modelo e os casos reais de desistência.

Após a análise de todas as métricas de avaliação propostas, podemos sugerir que o modelo está fazendo previsões consistentes e confiáveis.

Partiremos agora para a análise da interpretabilidade do modelo. Modelos interpretáveis desempenham um papel crucial em problemas relacionados à desistência do uso de cartão de crédito. Compreender as decisões do modelo é essencial para confiar nas previsões e, além disso, é vital para comunicar eficazmente as razões por trás das previsões a todas as partes interessadas.

Ao interpretarmos corretamente as previsões de um modelo, podemos fornecer insights valiosos, identificando quais características específicas dos clientes têm uma influência significativa na predição de desistência. Isso permite que as instituições financeiras ou empresas ajam de maneira proativa para abordar as áreas específicas que podem estar contribuindo para a desistência. Além disso, a interpretabilidade do modelo ajuda a identificar as variáveis mais importantes na tomada de decisão. Saber quais fatores têm um peso significativo nas previsões possibilita a implementação de estratégias mais direcionadas e eficazes.

No gráfico apresentado pela Figura 29, podemos visualizar as variáveis por ordem de importância global:

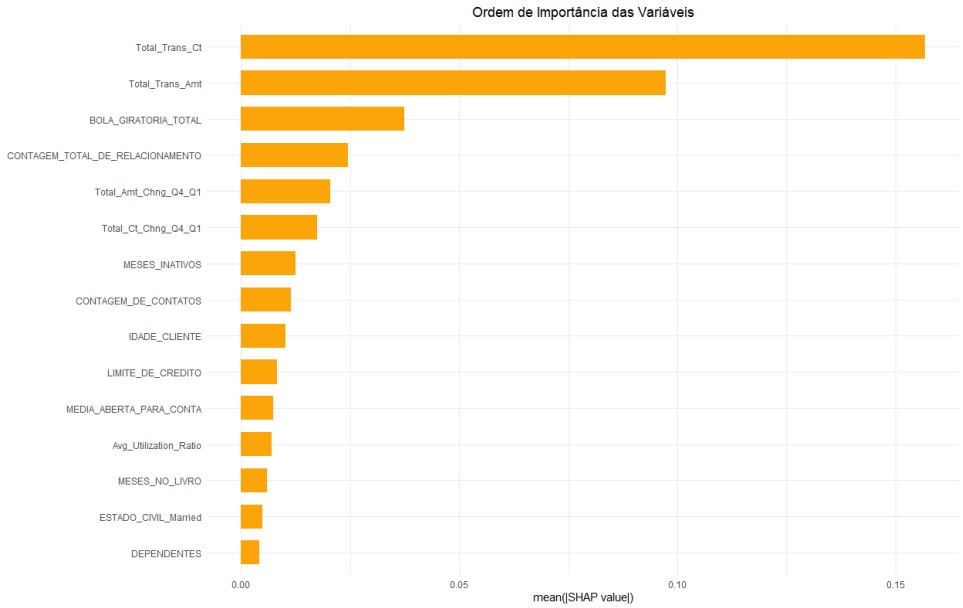

Figura 29 – Importância global das variáveis obtidas utilizando o índice Shap.

Note que, globalmente, a variável mais importante que o modelo utilizou para a classificação foi "Total-Trans-Ct" (que representa a quantidade de transações realizadas pelo cliente nos últimos 12 meses). Podemos observar que, quanto maior o valor do índice shap, maior será o impacto da variável na previsão da observação como Attrited Customer (cliente desiste de usar o cartão de crédito).

O gráfico apresentado na Figura 30 mostra a interação entre as duas variáveis mais importantes na esfera global: variável Total-Trans-Ct (quantidade de transações realizadas nos últimos 12 meses) e Total-Trans-Amt (valor total das transações nos últimos 12 meses) e seu efeito na variável Churn (Cancelamento do cartão de crédito):

Podemos observar que, à medida que o número de transações aumenta, a probabilidade de cancelamento diminui. Isso significa que clientes que realizam mais transações são menos propensos a cancelar seus cartões de crédito. Além disso, comportamento análogo pode ser observado em relação ao valor total das transações nos últimos 12 meses. Esses resultados são consistentes com a ideia de que clientes que estão mais satisfeitos com seus produtos e serviços são menos propensos a cancelar seus cartões de crédito. Clientes que gastam mais e realizam mais transações estão mais propensos a estar satisfeitos com seus cartões de crédito, pois estão usando-os com mais frequência.

A interpretabilidade local refere-se à capacidade de entender as decisões de um modelo em nível individual ou em relação a instâncias específicas de dados. Em problemas complexos, como previsão de desistência do uso de cartão de crédito, compreender por que um modelo faz uma determinada previsão para uma única observação pode ser tão importante quanto entender seu comportamento em escala global. Fornecer feedback explicativo sobre porque o modelo previu uma possível desistência pode aumentar a confiança dos clientes e contribuir para a

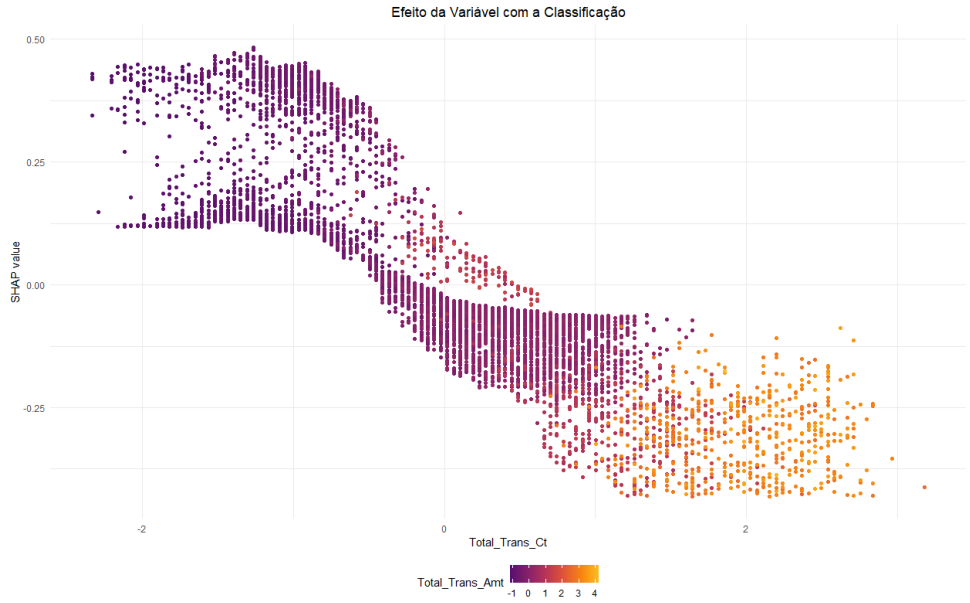

Figura 30 – Interação entre as variáveis Total-Trans-Ct e Total-Trans-Amt sobre a variável Churn.

retenção.

O gráfico apresentado na Figura 31 mostra a previsão para um único cliente escolhido aleatoriamente na base de testes. O valor de E(f(x)) = 0,128 corresponde ao valor base que seria previsto se nenhuma característica relevante fosse conhecida. Pode ser tratado como a previsão média do modelo na ausência de quaisquer recursos. Esse valor é o mesmo para todos os clientes da base de dados e representa o valor médio predito no conjunto de dados. Valores mais altos levam o modelo a prever 1 (classe positiva) e valores mais baixos levam o modelo a prever 0 (classe negativa). No nosso problema, na ausência de qualquer variável, o cliente seria classificado como sendo classe Existing Customer (cliente continua usando o cartão).

As variáveis que foram importantes para fazer a previsão para esta observação são mostrados em roxo e amarelo, com amarelo representando variáveis que aumentaram a pontuação do modelo, e roxo representando recursos que reduziram a pontuação. O tamanho desse impacto é representado pelo tamanho da barra. Os valores presentes dentro das barras são os valores do índice shap para cada variável correspondente e quantificam a magnitude do impacto de cada atributo na classificação final. O valor final estimado pelo modelo é dado pelo valor base somado com o total dos valores shap. Podemos notar que, para este cliente escolhido aleatoriamente foi classificado na classe negativa (f(x) = 0,0343), ou seja, um cliente que continua usando o cartão normalmente. A variável mais importante para essa classificação foi a variável Total-Trans-Ct, ou seja, a quantidade de transações realizadas nos últimos 12 meses.

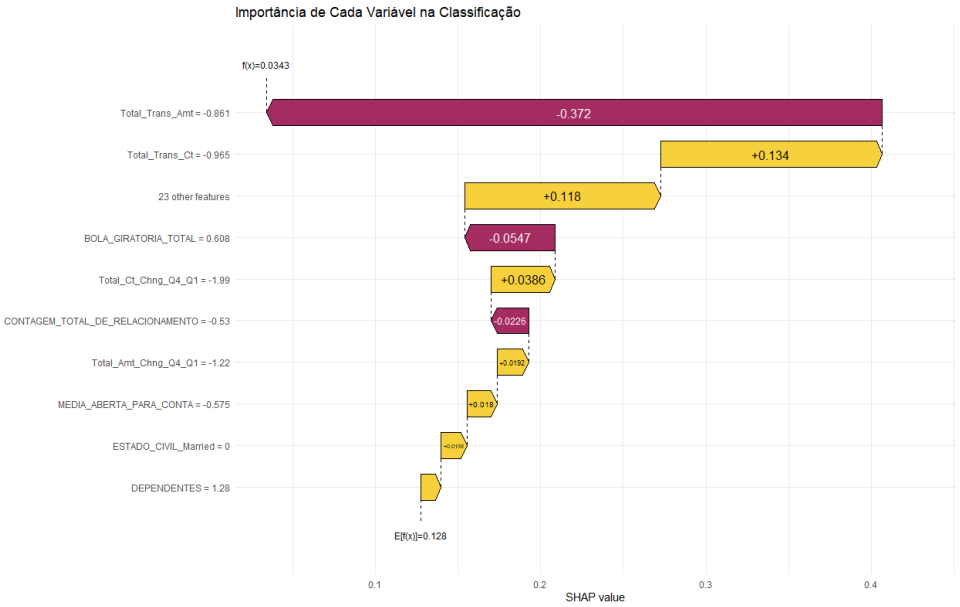

Figura 31 – Contribuição local das variáveis para a predição da classe de um cliente escolhido aleatoriamente.

5 CONCLUSÃO

O presente estudo teve como objetivo investigar os fatores que influenciam o cancelamento de cartões de crédito. Para isso, foi utilizado um conjunto de dados de cartão de crédito disponibilizado pelo Kaggle. Os resultados do estudo indicam que o gasto do cliente com o cartão de crédito e a atividade do cliente com o cartão são fatores importantes na previsão de cancelamento do cartão de crédito. Esses resultados são consistentes com a ideia de que clientes que estão mais satisfeitos com seus produtos e serviços são menos propensos a cancelar seus cartões de crédito. Clientes que gastam mais e realizam mais transações estão mais propensos a estar satisfeitos com seus cartões de crédito, pois estão usando-os com mais frequência.

As implicações desses resultados para a instituição financeira são significativas. Com esse estudo pode-se reduzir o churn de clientes do cartão de crédito focando em aumentar a satisfação do cliente. Isso pode ser feito oferecendo produtos e serviços de alta qualidade, resolvendo problemas rapidamente e fornecendo um bom atendimento e suporte ao cliente. Além disso, a instituição financeira pode usar as informações obtidas neste estudo para identificar clientes que estão em risco de cancelamento. Esses clientes podem ser contatados e oferecidos incentivos para permanecerem com a instituição financeira objeto do estudo.

Os resultados obtidos nesse trabalho são limitados pois têm como objeto central dados de uma única empresa, porém como são informações consistentes, as seguintes recomendações podem ser levadas em consideração para diminuir o churn de clientes nesse cenário:

- Oferecer recompensas aos clientes que gastam mais e fazem mais transações;
- Oferecer programas de fidelidade para clientes que usam seus cartões com frequência;
- Personalizar as ofertas para os clientes com base em seu histórico de compras;
- Investir em produtos de alta qualidade;
- Fornecer um bom atendimento ao cliente.

A implementação dessas recomendações pode ajudar essa instituição financeira a reduzir o churn de clientes de cartão de crédito e aumentar sua receita.

REFERÊNCIAS

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the fifth annual workshop on Computational learning theory**. [S.l.: s.n.], 1992. p. 144–152. Citado na página 22.

BREIMAN, L. **Classification and regression trees**. [S.l.]: Routledge, 2017. Citado na página 25.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 785–794. Citado na página 28.

DRUCKER, H. et al. Support vector regression machines. **Advances in neural information processing systems**, v. 9, 1996. Citado na página 23.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of computer and system sciences**, Elsevier, v. 55, n. 1, p. 119–139, 1997. Citado na página 28.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001. Citado na página 28.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2001. Citado 4 vezes nas páginas vi, 18, 19 e 20.

HO, T. K. Random decision forests. In: IEEE. **Proceedings of 3rd international conference on document analysis and recognition**. [S.l.], 1995. v. 1, p. 278–282. Citado na página 26.

KAGGLE. **Kaggle**. 2023. Disponível em: <https://www.kaggle.com/>. Citado na página 15.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in neural information processing systems**, v. 30, 2017. Citado na página 28.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017. Citado na página 42.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, p. 115–133, 1943. Citado na página 18.

MINSKY, M.; PAPERT, S. Perceptrons cambridge. **MA: MIT Press. zbMATH**, 1969. Citado na página 20.

MITCHELL, R.; FRANK, E.; HOLMES, G. Gputreeshap: massively parallel exact calculation of shap scores for tree ensembles. **PeerJ Computer Science**, PeerJ Inc., v. 8, p. e880, 2022. Citado 2 vezes nas páginas vi e 43.

PROKHORENKOVA, L. et al. Catboost: unbiased boosting with categorical features. **Advances in neural information processing systems**, v. 31, 2018. Citado na página 28.

QUINLAN, J. R. Discovering rules by induction from large collections of examples. **Expert systems in the micro electronics age**, Edinburgh University Press, 1979. Citado na página 25.

QUINLAN, J. R. Learning efficient classification procedures and their application to chess end games. In: **Machine learning**. [S.l.]: Elsevier, 1983. p. 463–482. Citado na página 25.

QUINLAN, J. R. **C4. 5: programs for machine learning**. [S.l.]: Elsevier, 2014. Citado na página 25.

RUMELHART, D. E.; MCCLELLAND, J. L.; GROUP, C. P. R. **Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations**. [S.l.]: MIT press, 1986. Citado na página 20.

SAN, A. **Credit Card Churn Prediction**. 2022. Disponível em: <https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn>. Citado na página 30.

SARADHI V., K. H. M. P. A. Decomposition method for support vector clustering. In: **In Proc. of the 2nd International Conference on Intelligent Sensing and Information Processing (ICISIP)**. [S.l.: s.n.], 2005. p. 268–271. Citado na página 23.

SCHAPIRE, R. E. The strength of weak learnability. **Machine learning**, Springer, v. 5, p. 197–227, 1990. Citado na página 27.

SHANNON, C. E. A mathematical theory of communication. **The Bell system technical journal**, Nokia Bell Labs, v. 27, n. 3, p. 379–423, 1948. Citado na página 24.