UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

Vinicius Pizetta de Souza

Recomendação de campanha de crédito para perfis de clientes selecionados: utilizando técnicas de aprendizado de máquina e ciência de dados

Florianópolis
2024

Vinicius Pizetta de Souza

Recomendação de campanha de crédito para perfis de clientes selecionados: utilizando técnicas de aprendizado de máquina e ciência de dados

Trabalho de Conclusão de Curso de Graduação em Sistemas de Informação do Centro Tecnológico, da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Bacharel em Sistemas de Informação.
Orientador: Prof. Elder Rizzon Santos, Dr.

Florianópolis
2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Souza, Vinicius Pizetta de
Recomendação de campanha de crédito para perfis de clientes selecionados: utilizando técnicas de aprendizado de máquina e ciência de dados / Vinicius Pizetta de Souza ; orientador, Elder Rizzon Santos, 2024.
79 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Sistemas de Informação, Florianópolis, 2024.

Inclui referências.

1. Sistemas de Informação. 2. Sistemas de Recomendação. 3. Análise de Crédito. 4. Aprendizado de Máquina. 5. Inteligência Artificial. I. Santos, Elder Rizzon. II. Universidade Federal de Santa Catarina. Graduação em Sistemas de Informação. III. Título.

Vinicius Pizetta de Souza

Recomendação de campanha de crédito para perfis de clientes selecionados: utilizando técnicas de aprendizado de máquina e ciência de dados

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de "Bacharel em Sistemas de Informação" e aprovado em sua forma final pelo Curso de Graduação em Sistemas de Informação.

Florianópolis, 25 de julho de 2024.

Banca Examinadora:

Prof. Elder Rizzon Santos, Dr.
Universidade Federal do Rio Grande do Sul

Prof. Roberto Willrich, Dr.
Universite de Toulouse III/França (Paul Sabatier)

Rodrigo Rodrigues Pires de Mello, Dr.
Universidade Federal de Santa Catarina

Este trabalho é dedicado aos meus colegas de classe e aos meus queridos pais.

AGRADECIMENTOS

Agradeço ao meu orientador, Prof. Elder Rizzon Santos, por todo o suporte e orientação durante o desenvolvimento deste trabalho. Agradeço também aos meus colegas de classe, que foram uma fonte de aprendizado e colaboração ao longo dessa jornada.

Gostaria de expressar minha profunda gratidão à minha família, que sempre me ofereceu apoio incondicional ao longo de toda a minha trajetória acadêmica, proporcionando o ambiente e os recursos necessários para que eu pudesse alcançar meus objetivos. Agradeço especialmente aos meus pais, por todos os ensinamentos, apoio e motivação que me deram ao longo da vida.

Agradeço também à minha namorada por estar ao meu lado em cada passo deste percurso, oferecendo carinho, compreensão e incentivo nos momentos mais desafiadores.

RESUMO

A análise de crédito é uma combinação de ciências exatas com características humanas, são levados em consideração fatores desde: capacidade de pagamento, acúmulo de dívidas, histórico e relacionamento bancário. Inúmeros profissionais são envolvidos para que se pese o risco/retorno envolvido, desde estatísticos, analistas financeiros, economistas, investidores, gerentes de conta, etc. Esses profissionais são necessários para que se tenha assertividade na análise e a assertividade tem imenso valor para instituições financeiras, pois limitam os riscos envolvidos no fornecimento de crédito. Neste trabalho, será feita uma análise de perfis de clientes. Para posterior recomendação de campanhas adequadas às diversas qualidades de perfis com o objetivo de automatizar as decisões feitas por estes diferentes profissionais, utilizando técnicas de aprendizado de máquina e ciência de dados. O desenvolvimento foi estruturado em duas Provas de Conceito (PoC): a primeira focada na validação de atributos relevantes e na viabilidade do sistema, e a segunda utilizando técnicas mais avançadas para maior abrangência. Os dados utilizados são provenientes do dataset da plataforma LendingClub, que cobre informações temporais e comportamentais de clientes. As etapas de desenvolvimento incluem análise exploratória de dados (EDA), limpeza e transformação dos dados, modelagem preditiva por meio de algoritmos supervisionados como regressão logística e XGBoost, e validação com métricas como AUC e precisão. O sistema é avaliado com base na segmentação de perfis e no alinhamento das campanhas recomendadas com as características observadas nos dados. Os resultados são discutidos em termos da capacidade do modelo em auxiliar a tomada de decisão sobre concessão de crédito, com destaque para a avaliação dos métodos empregados e a aplicabilidade das recomendações. O trabalho apresenta considerações sobre as limitações do modelo e sugere possíveis expansões, como o uso de dados não estruturados e abordagens híbridas.

Palavras-chave: Análise de crédito; aprendizado de máquina; ciência de dados.

ABSTRACT

Credit analysis is a combination of exact sciences and human characteristics. Factors such as payment capacity, debt accumulation, banking history and relationships are taken into consideration. Numerous professionals are involved in weighing the risk/return involved, including statisticians, financial analysts, economists, investors, account managers, etc. These professionals are necessary for assertive analysis, and assertiveness is of immense value to financial institutions, as it limits the risks involved in providing credit. In this work, an analysis of customer profiles will be carried out. Campaigns appropriate to the different profile qualities will be recommended later, with the aim of automating the decisions made by these different professionals, using machine learning and data science techniques. The development was structured into two Proofs of Concept (PoC): the first focused on validating relevant attributes and the viability of the system, and the second using more advanced techniques for greater scope. The data used comes from the LendingClub platform dataset, which covers temporal and behavioral information about customers. The development steps include exploratory data analysis (EDA), data cleaning and transformation, predictive modeling using supervised algorithms such as logistic regression and XGBoost, and validation with metrics such as AUC and accuracy. The system is evaluated based on profile segmentation and the alignment of recommended campaigns with the characteristics observed in the data. The results are discussed in terms of the model's ability to assist in decision-making on credit granting, with emphasis on the evaluation of the methods employed and the applicability of the recommendations. The paper presents considerations on the limitations of the model and suggests possible expansions, such as the use of unstructured data and hybrid approaches.

Keywords: Credit analysis; machine learning; data science.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AUC	Area Under ROC Curve
CART	Classification and Regression Tree
DBM	Deep Boltzmann Machine
DRBM	Discriminative Restricted Boltzmann Machine
IBGE	Instituto Brasileiro de Geografia e Estatística
PMCMV	Programa Minha Casa Minha Vida
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine

SUMÁRIO

1 INTRODUÇÃO

Em um país como o Brasil, com suas vastas diferenças culturais e econômicas, avaliar com precisão a capacidade dos clientes de honrarem seus compromissos financeiros se torna desafiador, tornando a concessão de crédito uma prática fundamental das instituições financeiras. A diversidade de perfis, comportamentos de consumo e necessidades de crédito entre os brasileiros exige soluções que vão além das abordagens tradicionais de análise financeira. Nesse contexto, surge a necessidade de sistemas que personalizem recomendações e prevejam a aderência dos clientes a diferentes campanhas de crédito.

Sistemas de recomendação utilizam algoritmos para sugerir produtos, serviços ou ações aos usuários com base em dados e padrões de comportamento. Essas tecnologias, amplamente aplicadas em áreas como comércio eletrônico e plataformas de conteúdo, podem ser adaptadas para o setor financeiro, sugerindo campanhas de crédito específicas para diferentes perfis de clientes (AGGARWAL, 2015). Este trabalho explora o uso de sistemas de recomendação e aprendizado de máquina para automatizar a seleção de perfis de clientes com maior potencial de adesão a campanhas de crédito. Por meio de técnicas de análise de dados, o objetivo é desenvolver um modelo que funcione como ferramenta preditiva, reduzindo riscos em ofertas de crédito das instituições.

O Brasil, um país continental colonizado por diversos povos, desenvolveu uma identidade própria em cada região ao longo dos anos, refletindo na maneira como os cidadãos lidam com dinheiro. Em 2021, a população brasileira estava estimada em 213,3 milhões de habitantes (IBGE, 2021). Com tantas possibilidades, analisar cada caso individualmente se torna impraticável; por isso, são extraídos indicadores de uma amostra populacional para gerar conclusões.

A análise de dados envolve processos de inspeção, limpeza, transformação e modelagem, com o objetivo de extrair informações úteis para a tomada de decisões. No contexto deste trabalho, a análise de dados será utilizada para identificar padrões de comportamento e perfis de crédito, facilitando a segmentação e o direcionamento de campanhas. Os dados podem ser estruturados, semiestruturados ou não estruturados (MCCALLUM; NIGAM, 2005).

A verificação de crédito refere-se ao processo pelo qual as instituições financeiras avaliam o histórico financeiro e a capacidade de pagamento dos clientes, permitindo a identificação daqueles com maior probabilidade de honrar suas dívidas. Neste trabalho, a verificação de crédito será um ponto central na seleção dos perfis mais adequados para as campanhas.

Existem diversos paradigmas de aprendizado de máquina, um campo da inteligência artificial que foi desenvolvido para criar algoritmos capazes de fazer previsões ou tomar decisões com base em dados (AYODELE, 2010). Esses algoritmos que se dividem em quatro categorias principais: aprendizado supervisionado, aprendizado não supervisionado,

aprendizado semi-supervisionado e aprendizado por reforço (SARKER, 2021).

O foco deste trabalho é o aprendizado supervisionado, que utiliza dados rotulados para treinar modelos capazes de prever resultados específicos. Esse aprendizado ocorre quando os objetivos do modelo são claramente definidos e uma coleção de exemplos de treinamento é usada para identificar padrões a partir de um conjunto determinado de entradas (HAN; KAMBER; PEI, 2011).

Entre as tarefas supervisionadas, destacam-se a “classificação”, que separa os dados em categorias, e a “regressão”, que ajusta os dados para fazer previsões numéricas. Por exemplo, prever o rótulo de uma classe ou o sentimento de um texto, como um tweet ou uma crítica de produto, é uma aplicação comum da aprendizagem supervisionada (SARKER, 2021).

A regressão logística é uma técnica estatística amplamente usada em problemas de classificação binária, com o objetivo de prever a probabilidade de um evento ocorrer ou não. Ela modela a relação entre uma variável dependente binária (por exemplo, "aprovado"ou "reprovado") e uma ou mais variáveis independentes. Diferentemente da regressão linear, a regressão logística mapeia as previsões para uma escala de probabilidade, permitindo uma interpretação mais intuitiva em termos de chance de ocorrência.

Neste trabalho, a regressão logística será utilizada para identificar fatores que influenciam a concessão de crédito, auxiliando na construção de um modelo que estime a probabilidade de um cliente aderir a uma campanha de crédito com sucesso.

Diversos algoritmos de classificação têm sido discutidos na literatura de aprendizado de máquina (WITTEN; FRANK, 2005), incluindo redes neurais artificiais e máquinas de vetores de suporte, que ganharam popularidade em aplicações voltadas para geociência (LARRY, 2010).

Após a criação do modelo de regressão logística, serão aplicadas técnicas de aprendizado de máquina para identificar os parâmetros mais significativos que impactam a concessão de crédito. O conjunto de dados Lending Club (2007–2020), disponível no Kaggle, será utilizado. Este dataset contém informações estruturadas, como histórico de crédito, perfil financeiro e características de empréstimo, permitindo a aplicação de modelos preditivos e a recomendação de campanhas com base na probabilidade de sucesso e nas características de cada perfil.

O código fonte para pré-processamento dos dados, implementação dos modelos e geração das visualizações pode ser encontrado no seguinte link: https://colab.research.google.com/drive/1eSFQZcaJeFx7rotY-j1b3ymHfApUhHOP?usp=sharing.

Além disso, o conjunto de dados utilizado neste estudo foi obtido através do Kaggle e está disponível para consulta no link: https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1. Essa abordagem visa promover a transparência e permitir que outros pesquisadores e interessados possam replicar e expandir este estudo.

1.1 OBJETIVOS

1.1.1 OBJETIVO GERAL

Desenvolver um sistema de recomendação de crédito para selecionar perfis de clientes e recomendar a campanha de crédito mais aderente aos perfis analisados.

1.1.2 OBJETIVOS ESPECÍFICOS

- Analisar o estado da arte em sistemas de recomendação e técnicas de análise de dados para identificação de perfis de clientes com base em similaridade.
- Propor um modelo ou mecanismo de recomendação de crédito adequado para identificar e segmentar clientes de acordo com suas características financeiras.
- Implementar um protótipo funcional do modelo para realizar uma análise dos resultados sobre a aderência das campanhas de crédito aos perfis recomendados.
- Avaliar os resultados obtidos e comparar com os resultados esperados.

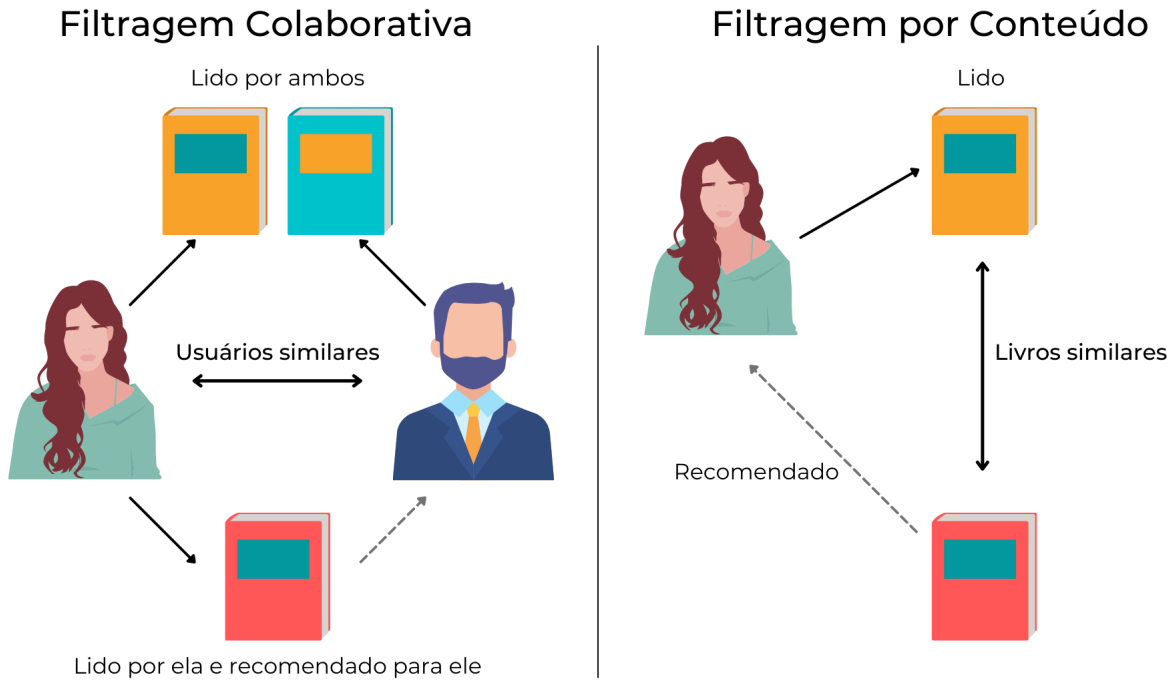
2 FUNDAMENTAÇÃO TEÓRICA

2.1 SISTEMAS DE RECOMENDAÇÃO

Sistemas atuais lidam com quantidades de dados imensas e, por conseguinte, oferecem conteúdo de forma abundante. Por este motivo, usuários por vezes não conseguem distinguir conteúdos realmente relevantes de outros secundários (ABBAR; BOUZEGHOUB; LOPEZ, 2009). Sistemas recomendadores são ferramentas que, a partir de dados históricos de um usuário, filtram conteúdo de maneira a apresentar apenas os itens mais interessantes ao usuário (ABBAR; BOUZEGHOUB; LOPEZ, 2009). Para (PARK; LEE; CHO, 2012), são sistemas que utilizam tecnologia analítica para calcular a probabilidade de um produto ser comprado em um local, de forma que o produto correto possa ser recomendado ao cliente.

Neste trabalho, será construído um sistema de recomendação de crédito, ou seja, um sistema que recebe diversos dados de clientes e os qualifica em categorias separadas por nível de confiança em manter suas finanças em dia para que então seja ofertada uma campanha adequada ao seu perfil. Existem três tipos principais de sistemas de recomendação: baseados em conteúdo, híbridos e com filtragem colaborativa.

Figura 1 – Exemplos de técnicas utilizadas por sistemas de recomendação.



Fonte: Medium Awari Bookstore (2021).

2.1.1 RECOMENDAÇÃO BASEADA EM CONTEÚDO

Segundo (HERLOCKER, 2000), por muitos anos os cientistas têm direcionado seus esforços para aliviar o problema ocasionado pela sobrecarga de informações através de projetos que integram tecnologias que automaticamente reconhecem e categorizam as informações. Alguns softwares têm como objetivo gerar de forma automática descrições dos conteúdos dos itens e comparar essas descrições com os interesses dos usuários, visando verificar se o item é ou não relevante para cada um (BALABANOVIC; SHOHAM, 1997). Esta técnica é chamada de filtragem baseada em conteúdo, pois realiza uma seleção com base na análise dos atributos de cada item e no perfil do usuário (HERLOCKER, 2000).

No contexto da análise de crédito, a recomendação baseada em conteúdo ajuda a identificar clientes que possuem características específicas, como alta renda, histórico positivo de pagamentos ou baixa taxa de inadimplência, permitindo recomendar campanhas de crédito alinhadas ao perfil do cliente. Por exemplo, um cliente que já financiou veículos e imóveis poderia receber recomendações para campanhas de crédito voltadas para bens de alto valor, refletindo sua afinidade com esse tipo de financiamento.

2.1.2 RECOMENDAÇÃO COM FILTRAGEM COLABORATIVA

A abordagem da filtragem colaborativa foi desenvolvida para atender pontos que estavam em aberto na filtragem baseada em conteúdo (HERLOCKER, 2000). A recomendação colaborativa difere da baseada em conteúdo, pois não requer compreensão ou reconhecimento do conteúdo dos itens. Um exemplo de ambiente baseado em filtragem colaborativa é o sistema de recomendação de filmes MovieLens (RIEDL; AL., 1999).

Na análise de crédito, a filtragem colaborativa é útil para recomendar campanhas com base no comportamento de clientes semelhantes. Por exemplo, se clientes com determinado perfil de renda e histórico financeiro aderiram a uma campanha de crédito específica, o sistema pode sugerir essa mesma campanha a novos clientes com perfis similares, reduzindo a necessidade de análises individuais complexas.

2.1.3 RECOMENDAÇÃO HÍBRIDA

A abordagem de recomendação híbrida combina os pontos fortes das filtragens colaborativa e baseada em conteúdo, visando criar um sistema que melhor atenda às necessidades dos usuários (HERLOCKER, 2000). Essa abordagem se baseia nas vantagens oferecidas por ambas as técnicas, unindo-as para eliminar as limitações de cada uma.

No contexto de recomendação de crédito, o sistema híbrido permite uma análise mais completa, considerando tanto os atributos específicos do cliente quanto o comportamento de outros clientes com perfis semelhantes. Suponha, por exemplo, que um cliente tem um perfil de consumo para bens duráveis e compartilha comportamentos financeiros com outros clientes que frequentemente aderem a campanhas de crédito para reforma

de imóveis. A abordagem híbrida poderia recomendar campanhas para reforma e melhorias residenciais, baseando-se em uma análise mais detalhada do perfil individual e do comportamento de grupos similares.

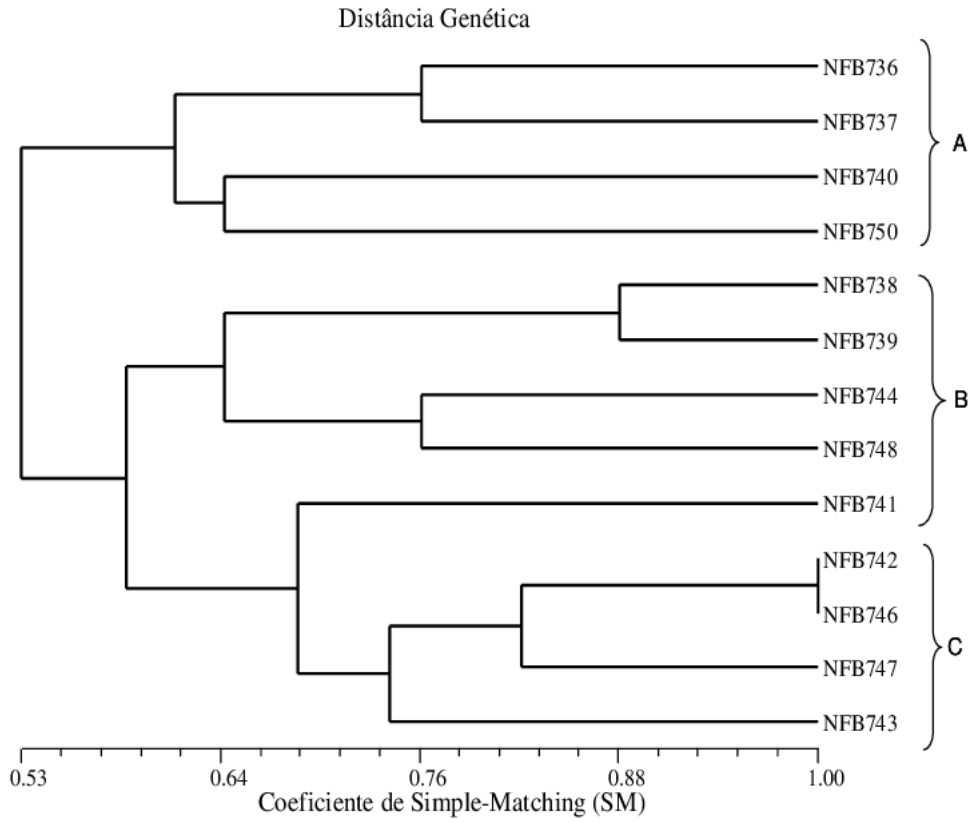
2.2 TÉCNICAS DE ANÁLISE DE DADOS PARA SIMILARIDADE EM PERFIS

Considerando um conjunto de dados, técnicas de análise por similaridade dividem os dados em grupos contendo instâncias com características semelhantes (AGGARWAL, 2015). Técnicas de agrupamento, como o clustering, têm o objetivo de agrupar indivíduos em classes com base em semelhanças, utilizando critérios que determinam as similaridades entre eles (TAN; STEINBACH; KUMAR, V., 2018). Essas técnicas são fundamentais para identificar padrões de perfis de clientes e direcionar campanhas de crédito para grupos com maior probabilidade de adesão.

Métricas de distância/similaridade que podem ser aplicadas incluem: distância euclidiana, Manhattan, similaridade de cossenos, coeficiente Sorensen-Dice, concordância simples, Russel e Rao, Rogers e Tanimoto, e similaridade de Jaccard (LABATUT; CHERIFI, 2011). Essas métricas permitem definir relações entre dados financeiros, facilitando a segmentação de clientes para estratégias de crédito personalizadas.

As melhores maneiras de visualizar a similaridade ou os clusters de dados estão nas formas da matriz de similaridade e no dendrograma, como ilustrado na Figura 2.

Figura 2 – Exemplifica visualmente um modelo de dendrograma de distância genética e seus clusters.



Fonte: Universidade Federal Rural de Pernambuco (2007).

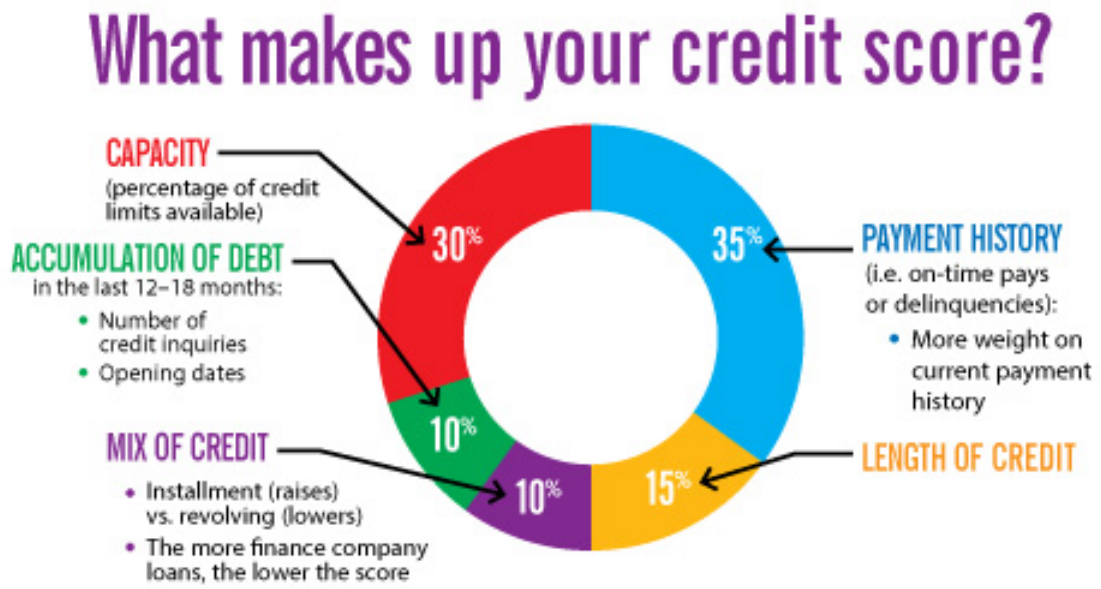
2.3 ANÁLISE DE CRÉDITO

Segundo (STEYNER, 1999), "a correta decisão de crédito é essencial para a sobrevivência das empresas bancárias". Afirmam ainda que "qualquer erro na decisão de concessão pode significar que, em uma única operação, haja a perda do ganho obtido em dezenas de outras bem-sucedidas". O que é desejável e necessário, então, é "analisar uma proposta de negócio e comparar o custo de conceder com o custo de negar a operação" (STEYNER, 1999). Segundo (SELAU; RIBEIRO, 2009), as empresas que concedem crédito apostam em uma análise de crédito mais precisa, evitando trabalhar com clientes que ofereçam maior risco e, assim, reduzindo o índice de inadimplência. Para (ASSAF NETO; LIMA, 2009), "a análise de crédito tem por objetivo selecionar os clientes a prazo, sua capacidade de pagamento, assim como os limites monetários de crédito que podem ser concedidos".

A análise de crédito se beneficia de sistemas de recomendação e técnicas de análise de dados, permitindo uma segmentação detalhada e oferecendo campanhas direcionadas. Por exemplo, um sistema pode analisar o score de crédito, histórico de pagamento e outros fatores financeiros para qualificar um cliente como candidato a uma campanha de crédito

específica.

Figura 3 – Composição do score de crédito da Dupaco Community Credit Union.



Fonte: Dupaco Community Credit Union (2022).

Figura 4 – Pontuação do Serasa Score.

Fonte: Serasa Experian (2022).

3 TRABALHOS RELACIONADOS

Durante o processo de pesquisa, foi realizado o cruzamento de temas relacionados a sistemas de recomendação, técnicas de análise de perfis e análise de crédito, retornando alguns artigos, monografias e dissertações com informações relevantes para a fundamentação e desenvolvimento deste trabalho. Foram apresentados conceitos, técnicas, abordagens, modelos de desenvolvimento e possíveis tecnologias para processamento de análise de crédito.

3.1 MACHINE LEARNING MODELS FOR CREDIT ANALYSIS IMPROVEMENTS: PREDICTING LOW-INCOME FAMILIES

No trabalho de J.R. de Castro Vieira, F. Barboza, V.A. Sobreiro, primeiramente é apresentado o tema do trabalho, que é uma investigação sobre o programa de financiamento habitacional Minha Casa, Minha Vida (PMCMV). Os dados do PMCMV foram utilizados como modelo em um estudo com o objetivo de investigar e sugerir mecanismos de gestão de risco de crédito que pudessem ser aplicados a programas habitacionais, dentro de suas limitações.

Mais recentemente, diversos estudos têm demonstrado a adoção de técnicas de aprendizado de máquina na modelagem de crédito, destacando diversas metodologias para estimar a probabilidade de inadimplência, como SVM, Decision Tree, Random Forest e Ensacamento e reforço. A maioria dos estudos destaca as vantagens do uso de sistemas de aprendizado de máquina na análise de risco de crédito devido ao melhor desempenho de classificação do que as técnicas tradicionais, como Regressão Logística (VIEIRA *et al.*, 2019).

O PMCMV utilizou um banco de dados financeiro para pessoas físicas que pagam mensalidades e permite estudos governamentais nesses estudos. A informação é de março de 2009 a dezembro de 2015. Vale destacar que 311 mil contratos de um total de 2,24 milhões estavam atrasados no período, implicando que a carteira de crédito tinha um risco de inadimplência em torno de 11,80%.

Em termos de medidas de validação, aplicamos BRIER Score, AUC (Area Under ROC Curve) (Área Sob a Curva da ROC), teste de Kolmogorov-Smirnov (KS), além das medidas comuns (Mean Accuracy, erros tipo I e II), que normalmente são utilizadas em estudos semelhantes (VIEIRA *et al.*, 2019).

Para aplicação de todas as ferramentas, foi utilizada a versão R 3.3.1, suportada pelas bibliotecas randomForest, C50, ipred, e1071, gmodels e MASS. Adicionalmente, também foram utilizadas as seguintes bibliotecas: ROCR para geração das curvas ROC e cálculo da AUC; e a biblioteca de verificação, que calcula a pontuação de Brier. O processamento foi realizado em um servidor com as seguintes especificações: processador i7-3770 3,40 GHz com 4 núcleos e 8 threads e 16 GB de RAM (VIEIRA *et al.*, 2019).

Dependências não lineares e temporais são comuns em grandes conjuntos de dados. Os métodos econométricos podem capturar efetivamente as informações presentes em uma ampla variedade de conjuntos de dados. Os melhores resultados e métodos demonstram a superioridade dos projetos de Wang et al., Jones et al. e Wang et al., Lessmann et al.

Grandes conjuntos de dados financeiros geralmente apresentam desafios estatísticos significativos porque são caracterizados por aumento de ruído, distribuições de cauda pesada, padrões não lineares e dependências temporais. Os métodos econométricos convencionais falham em capturar de forma eficiente as informações contidas em todo o espectro dos conjuntos de dados (VIEIRA *et al.*, 2019).

Os resultados aqui apresentados mostram que a melhoria da avaliação do risco de crédito no PMCMV pode produzir resultados significativos na redução da inadimplência. Os resultados mostram a superioridade dos algoritmos computacionais – os três melhores métodos (BG, RF e AdaBoost) são baseados em classificadores ensemble, e reforçam as contribuições apresentadas por Wang et al., Wu et al., Jones et ai. e Lessmann et al. (VIEIRA *et al.*, 2019).

Por fim, os autores apresentam as limitações de seu trabalho:

- Primeiramente, o conjunto de dados apresenta apenas créditos aprovados, o que é um problema recorrente em todos os conjuntos de dados de crédito. No entanto, a gestão de risco é o ponto-chave, onde o evento de default é mais importante porque implica maiores custos de classificação incorretos (VIEIRA *et al.*, 2019).
- Em segundo lugar, os autores afirmam que realizaram todas as técnicas usando parâmetros padrão e nenhuma seleção de recursos foi feita. De forma que as melhorias para cada modelo ou mesmo uma combinação delas podem ser examinadas detalhadamente em estudos futuros.

3.2 A CREDIT RISK PREDICTING HYBRID MODEL BASED ON DEEP LEARNING TECHNOLOGY

No trabalho de Chong Wu, Dekun Gao e Siyuan Xu, para prever os devedores inadimplentes, foi utilizada uma técnica de deep learning para extrair recursos eficazes e técnica de subamostragem para equilibrar o conjunto de dados. Com isso, foi proposto um modelo de classificador híbrido usando técnicas de deep learning Deep Boltzmann Machine (DBM) e Discriminative Restricted Boltzmann Machine (DRBM).

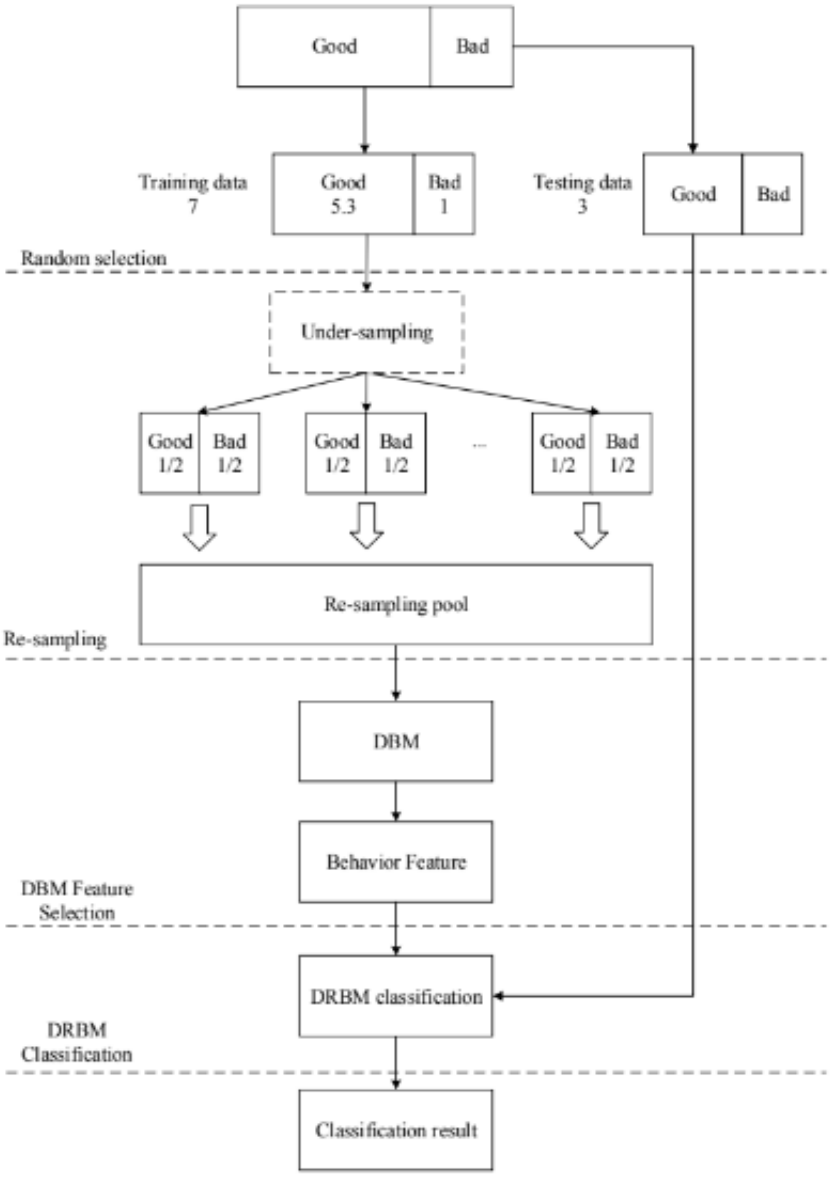
Para examinar o desempenho, dados de crédito do mundo real do Lending Club são aplicados no modelo proposto. Os resultados estáveis e de melhor desempenho mostram que o classificador híbrido que propomos é mais eficaz e poderoso (WU; GAO; XU, 2021).

Os métodos de deep learning são certificados como uma poderosa teoria de seleção de recursos, que tem sido amplamente empregada em muitos campos, como reconhecimento

de rosto e reconhecimento de emoções. Diferem-se dos classificadores rasos tradicionais com menos camadas ocultas, método de deep learning com camadas ocultas suficientes, estendido por redes neurais tradicionais. Como um tipo de tecnologia de deep learning, o DBM tem muitas vantagens na mineração de informações complexas, mas raramente foi usado no campo de risco de crédito (WU; GAO; XU, 2021).

Na figura 5, é mostrada a estrutura do modelo híbrido, separada em três partes: reamostragem, seleção de recursos DBM e classificação DRBM.

Figura 5 – Modelo do projeto DRBM.



Fonte: Universidade de Tecnologia de Pequim (2006).

Na primeira fase, os autores aplicaram a subamostragem nos dados de treinamento a vários subconjuntos e os combinaram como um pool de reamostragem para misturar

dados bons (pagos integralmente) e dados ruins (cobrados) após a subamostragem nos dados de crédito individuais de treinamento. A subamostragem é um tipo de tecnologia para ajustar a distribuição de classes dos dados originais. Os dados selecionados combinam o conjunto de dados menor com o novo conjunto de dados de treinamento. Neste artigo, aplicamos a tecnologia de subamostragem para resolver o problema de desequilíbrio dos dados de empréstimo do candidato (WU; GAO; XU, 2021).

Na segunda fase, o DBM foi utilizado para selecionar informações de recursos eficazes do conjunto de dados de empréstimos. Como método de aprendizado profundo, o DBM possui uma poderosa capacidade de extração de recursos. Usando a estrutura profunda, o DBM pode extrair informações de recursos eficazes de valores de atributos complexos e diversos (WU; GAO; XU, 2021).

Na terceira fase, é feita a classificação do DRBM, que considera a saída do recurso de comportamento do DBM como entrada da classificação. O DRBM é um modelo gráfico não direcionado de duas camadas, bipartido, com unidades binárias nas camadas de entrada e ocultas. Em um DRBM, existem conexões entre as unidades ocultas e de entrada, mas não há conexão entre duas unidades dentro da mesma camada. Na verdade, a arquitetura do DRBM tem uma motivação física profunda. Demonstraram que a máquina de Boltzmann Restrita Discriminativa é estatisticamente equivalente ao conhecido modelo físico da rede Hopfield (WU; GAO; XU, 2021).

Os resultados experimentais na figura 6 mostram que o classificador híbrido proposto tem desempenho mais estável e eficaz.

Figura 6 – Figura da Tabela IV.

TABLE IV: THE COMPARISON OF CLASSIFICATION ACCURACY OF DBM+DRBM AND SVM

Experiment times	Accuracy		
		DBM+DRBM	
	SVM	Testing set	Training set based on resampling pool
1	0.6961	0.8816	0.986
2	0.7041	0.8805	0.998
3	0.6956	0.8927	0.99
4	0.7342	0.8898	0.996
5	0.7283	0.8863	0.988
6	0.6950	0.8842	0.994
Average accuracy	0.7089	0.8858	-

Fonte: Universidade de Brasília (2018).

3.3 MACHINE LEARNING MODELS FOR CREDIT ANALYSIS IMPROVEMENTS: PREDICTING LOW-INCOME FAMILIES

No trabalho de Luis Carlos Otte Junior, são abordadas soluções para sistemas de pagamento e telecomunicações que fornecem relatórios para apoiar o faturamento do cliente. De forma que o objetivo de seu projeto é implementar ferramentas e soluções inteligentes para reduzir a perda de tempo e aumentar a produtividade do gestor, decorrente da necessidade de analisar e agregar todos os dados para implementação.

Com a abordagem na área de mineração de dados, foram tratados os casos de ruídos, valores ausentes e classes desbalanceadas, representando o processo de preparação dos dados. Após esta etapa, com alguns ajustes nas especificações, os algoritmos foram capazes de encontrar padrões de sucesso e fracasso.

Os modelos de Credit Score são os mais comuns e utilizados atualmente (KUMAR, A.; RAVI, 2007). Em geral, são disponibilizados através de serviços de consulta fornecidos por empresas privadas como SERASA, SPC e Equifax do Brasil (PINHEIRO; MOURA, 2003).

Outro modelo semelhante e considerado como uma variação do Credit Score é o Behavior Score ou Escore de Comportamento. Além do uso de dados sociodemográficos, o modelo utiliza também dados comportamentais originados de históricos dos clientes presentes no portfólio da empresa, como histórico de pagamentos em dia, em atraso, quantidade de empréstimos, entre outros (OTTE, 2018).

O uso de modelos relacionados à questão da concessão do crédito não se limita somente ao momento inicial a ser avaliado conforme os trabalhos referenciados anteriormente, mas em todo o processo posterior do fornecimento de crédito, ou seja, em quase todas as etapas do ciclo de crédito ao consumidor (LAWRENCE, 1984), ilustrado na Figura 7, é possível aplicar modelos com o objetivo de gerar eficiência e reduzir custos (OTTE, 2018).

Figura 7 – Modelo utilizado por Otte.

Fonte: Universidade de Brasília (2018).

O primeiro trabalho implementou um algoritmo híbrido baseado na técnica Random-Forest, utilizando pesos nas classes menores e posteriormente aplicou um balanceamento de oversampling. A técnica tem como objetivo replicar as classes minoritárias, forçando um balanceamento das classes (OTTE, 2018).

No segundo trabalho, foi utilizada a técnica de SVM + Naive-Bayes Tree para extração de regras e classificação. Os dados foram balanceados por meio da técnica de SMOTE (CHAWLA *et al.*, 2002).

Com relação a modelos de aprendizado de máquina, é possível encontrar inúmeros tipos e aplicações, que, de forma sistêmica, estão sempre relacionados aos processos básicos de cobrança, que são divididos em faixas de dias em atraso conforme políticas de crédito particulares de cada empresa (SILVA SANTO, 2013).

Ao concluir a análise do ciclo de crédito, foi possível observar que o trabalho que melhor se aproxima da questão do problema a ser resolvido pelo trabalho é o modelo de Collection Score. O modelo busca identificar possíveis clientes pagadores de suas dívidas já em um cenário em que todos são devedores, diferente dos modelos de Credit ou Behaviour Score, onde o cliente pode ser ou não devedor (OTTE, 2018).

Para melhor compreensão de todos os modelos, foi elaborada a tabela na figura 8, com o objetivo de comparar as escolhas de cada trabalho com os respectivos resultados (OTTE, 2018).

Figura 8 – Comparação de modelos utilizados por Otte.

	Trabalhos										
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
Modelos	Credit			Behavior	Churn		Retenção	Late Payment	Collection		
Regr. Logística			0,72	0,87			0,69	0,70	0,78	0,83	
Árvore Decisão		0,09*		0,85	0,62			0,79			
MLP	0,75			0,89	0,78						
SVM	0,75		0,72		0,87	0,86					
k-NN	0,74										
GB Tree			0,72					0,67			
RandomForest					0,93		0,73				
NBTree						0,94					
SVM+NBTree						0,83					
Tobit Type II											0,80
Dist. Classes %	72 / 28	91 / 9	80 / 20	91 / 9	95 / 5	93 / 7	84 / 16	65 / 35	82 / 18	50 / 50	60 / 40

Fonte: Universidade de Brasília (2018).

De acordo com o autor, o pré-processamento de dados é uma etapa fundamental e influência todo o processo de aprendizado de máquina em termos de fato e experiência. Se os dados não incluem as informações necessárias para reconstruir melhor o espaço de solução, os algoritmos tendem a não ajustar os dados, um efeito conhecido como underfitting, ou modelos tendenciosos "decoram"as informações, um efeito conhecido como overfitting.

Com isso, para extrair um bom conhecimento do conjunto de dados, é necessário realizar análise e refinamento, minimizando assim possíveis erros na generalização dos algoritmos de classificação. Portanto, para resolver um problema usando técnicas de aprendizado de máquina, é necessário treinar um conjunto de dados que melhor represente o espaço solução.

Com a extração de dados, o autor reuniu um dataset com 190 atributos e, em seguida, reduziu esse conjunto de dados em 72% finalizando com 55 atributos relevantes. Com esses atributos, o autor constrói uma matriz de confusão para perceber quais os atributos mais relevantes e, com isso, ele seleciona 17 atributos finais. Continuando no processo de mineração de dados, o autor transforma e normaliza os atributos e também remove outliers.

Por fim, ele trata de fazer a reamostragem. Conforme trabalhos pesquisados, algumas técnicas de balanceamento de classes foram adotadas. A primeira encontrada foi a de over-sample, que consiste em copiar os objetos da classe minoritária em uma quantidade próxima à majoritária (OTTE, 2018).

A outra técnica utilizada foi a de under-sample, onde é aplicada uma remoção

aleatória de objetos da classe majoritária até atingir uma quantidade próxima da minoritária (OTTE, 2018). Por último, a destacar é a técnica de implantação de dados sintéticos chamada SMOTE (CHAWLA *et al.*, 2002).

A categoria que será desenvolvida nesse trabalho é a de modelos baseados em dados empíricos. Esses modelos são fundamentados nas áreas de estatística e aprendizado de máquina e utilizam técnicas como a Regressão Logística, Redes Neurais, SVM, K-NN e Árvores de Decisão, dentre outras técnicas. A regressão logística faz parte de um grupo de modelos chamado Modelo Linear Generalizado (MLG) (OTTE, 2018).

Grande parte dos trabalhos pesquisados na área de análise e risco de crédito utiliza a técnica de regressão logística, pois além da fácil implementação, a técnica apresenta uma transparência para com os dados, ou seja, não se trata de uma caixa preta, permitindo a realização de análises para cada atributo utilizado em um modelo e a extração de regras (OTTE, 2018).

Os trabalhos que obtiveram resultados satisfatórios utilizaram CART e Random Forest, ambos são baseados em árvores de decisão, porém com diferentes tipos de abordagens, sendo CART uma técnica de árvore de decisão e RandomForest utiliza-se de árvores, porém com uma abordagem com o conceito de Ensemble learning combinando outras técnicas (OTTE, 2018).

Outra abordagem muito utilizada para soluções de problemas de classificação é o uso de redes neurais artificiais ou RNA (OTTE, 2018). Os classificadores foram avaliados através das medidas clássicas de desempenho, são elas: Acurácia (ACC), Especificidade (ESP), Sensibilidade (SEN) e Kappa (OTTE, 2018).

Para que o autor aplicasse as técnicas, teve de realizar otimizações nos valores dos hiper parâmetros pelos algoritmos automaticamente pela biblioteca Caret, utilizando a técnica de Grid Search.

Por fim, o autor conclui que a parte inicial de mineração de dados teve grande relevância para o entendimento do problema e posterior preparo correto para que fosse aplicada a qualquer algoritmo. Ele comenta que, caso considere atributos desconexos, certamente teria resultados indesejados. Também teve problemas com dados desbalanceados da. maneira como modelos financeiros são conhecidos por esse tipo de problema, precisando utilizar técnicas de reamostragem para balancear classes.

Após revisão bibliográfica, ele concluiu que modelos de risco de crédito utilizam somente a técnica de regressão logística para classificação. Nesse trabalho, além da regressão logística, foram empregadas outras técnicas de aprendizado de máquina, como árvore de decisão CART, RandomForest e Multilayer Perceptron (OTTE, 2018).

As técnicas utilizadas conseguiram resolver o problema do score que faz parte do risco de crédito e conseguiram identificar casos próximos de bons pagadores.

3.4 OUTROS TRABALHOS

A comparação entre os trabalhos evidencia a diversidade de abordagens utilizadas. O trabalho de (VIEIRA *et al.*, 2019) aplicou modelos de Regressão Logística com foco em dados do PMCMV, um conjunto específico relacionado a habitação. (WU; GAO; XU, 2021) empregaram técnicas de deep learning para análise de dados financeiros do Lending Club, utilizando redes neurais profundas para extração de características. Por outro lado, (OTTE, 2018) combinaram algoritmos tradicionais, como Regressão Logística e Support Vector Machines (SVM), com técnicas de balanceamento de dados, como SMOTE e undersampling, para ajustar a distribuição das classes e melhorar a robustez do modelo.

Nesta seção, são apresentados trabalhos de referência relevantes que utilizam técnicas de recomendação e análise de crédito, organizados em uma tabela comparativa. As Tabelas 1 e 2 destacam o dataset utilizado, o domínio da aplicação, se as recomendações são baseadas em conteúdo ou não, e as métricas de avaliação adotadas. A comparação fornece informações sobre métodos e métricas que podem ser avaliados e potencialmente aplicados neste estudo, ressaltando aspectos principais, como forma de recomendação, tipo de dataset, domínio de aplicação e métricas de avaliação.

Tabela 1 – Forma de Recomendação e Métricas de Avaliação

Trabalho	**Forma de Recomendação**	**Métricas de Avaliação**
Vieira et al. (2016)	Baseada em Regressão	AUC, Brier Score, KS
Wu et al. (2021)	Híbrida (Deep Learning)	Acurácia, AUC
Otte Junior (2019)	Baseada em Técnicas Tradicionais	Acurácia, Kappa, Especificidade
Financial Credit Risk Assessment (2015)	Estatística e ML	AUC, Precisão
Survey of Machine Learning in Credit Risk (2020)	Machine Learning	AUC, F1-Score
Quantitative Methods in Credit Management (1994)	Modelos Quantitativos	Métricas diversas
Machine Learning: Real-World Applications (2021)	Diversas Técnicas de ML	Precisão, Recall
Euro Area Bank Lendinglimpeza e Survey (2022)	Análise de Políticas	Relatórios qualitativos
Experimental Analysis of ML Methods (2021)	Análise Experimental	Acurácia, Redução de Complexidade
Credit Risk Analysis with ML (2020)	Machine Learning	AUC, Precisão

Tabela 2 – Dataset e Acesso ao Dataset

Trabalho	**Dataset**	**Acesso ao Dataset**
Vieira et al. (2016)	PMCMV	Público
Wu et al. (2021)	Lending Club	Público
Otte Junior (2019)	Dados diversos	Público/Indeterminado
Financial Credit Risk Assessment (2015)	Dados financeiros	Privado
Survey of Machine Learning in Credit Risk (2020)	Vários datasets	Público
Quantitative Methods in Credit Management (1994)	Dados históricos	Privado
Machine Learning: Real-World Applications (2021)	Diversos datasets	Público
Euro Area Bank Lending Survey (2022)	Dados de bancos	Privado
Experimental Analysis of ML Methods (2021)	Vários datasets	Público
Credit Risk Analysis with ML (2020)	Dados financeiros	Privado

Com base na análise comparativa dos trabalhos, as técnicas a serem avaliadas neste projeto incluem métodos estatísticos que podem ser testados em dados financeiros para investigar sua adequação na previsão de risco de crédito. Por exemplo, o estudo de

(VIEIRA *et al.*, 2019), que aplicou Regressão Logística com dados do PMCMV, sugere que essas técnicas podem capturar padrões em situações onde o desequilíbrio de classes é um desafio relevante. O trabalho Credit Risk Analysis with ML (2020) também enfatizou a importância de métricas como AUC e Precisão para uma avaliação abrangente.

Além disso, técnicas de balanceamento de dados, conforme discutido por (OTTE, 2018), podem ser aplicadas para ajustar a distribuição dos dados e reduzir possíveis vieses. O uso de métricas como AUC e F1-Score permitirá avaliar o desempenho, ajudando a validar as abordagens.

4 DESENVOLVIMENTO

O objetivo deste trabalho é desenvolver um sistema de recomendação de crédito que identifique perfis de clientes e sugira a campanha de crédito mais adequada, utilizando algoritmos de aprendizado de máquina. O processo de desenvolvimento foi dividido em oito fases principais: pesquisa e escolha do dataset, análise do conjunto de dados, limpeza dos dados, seleção de atributos, criação das campanhas, aplicação dos métodos de aprendizado de máquina, avaliação dos resultados e discussão final. A escolha de duas Provas de Conceito (PoCs) foi motivada pela necessidade de comparar uma abordagem inicial simplificada com uma mais avançada. A PoC 1 validou a relevância dos atributos e a viabilidade do sistema com uma implementação básica, enquanto a PoC 2 incorporou técnicas avançadas complexas para uma avaliação mais. As três primeiras fases são comuns às duas PoCs, enquanto as cinco fases restantes foram implementadas de forma distinta para cada abordagem.

Pesquisa e Escolha de Dataset: A fase inicial consistiu na busca por um conjunto de dados que representasse de forma abrangente o comportamento de crédito, priorizando a diversidade de variáveis financeiras.

Análise do Conjunto de Dados: Realizou-se uma análise exploratória para entender a distribuição das variáveis, identificar padrões e problemas como outliers ou dados ausentes.

Limpeza dos Dados: A qualidade do dataset foi aprimorada com correção de valores ausentes, formatação de tipos de dados e remoção de informações irrelevantes.

Seleção de Atributos: Foram escolhidos atributos relevantes com base em sua importância para a predição do comportamento de crédito.

Criação das Campanhas: Campanhas de crédito foram desenvolvidas com critérios financeiros específicos para segmentar clientes em grupos-alvo.

Aplicação de Métodos de Aprendizado de Máquina: Algoritmos de regressão logística foram implementados para construir modelos preditivos eficientes.

Avaliação dos Resultados: O desempenho dos modelos foi avaliado com métricas como precisão, recall e AUC para comparar e otimizar as abordagens.

Discussão Final: Analisaram-se os resultados para discutir as implicações das abordagens e sugerir melhorias para futuras aplicações.

4.1 PESQUISA E ESCOLHA DE DATASET

O desenvolvimento de um sistema de recomendação de crédito requer um conjunto de dados abrangente e rico em informações, capaz de capturar as complexidades dos perfis financeiros dos clientes e permitir a aplicação de técnicas avançadas de aprendizado de máquina. Nesse contexto, o dataset “A Credit Risk Predicting Hybrid Model Based on Deep Learning Technology” foi selecionado como a melhor opção, considerando sua amplitude e diversidade. Utilizando o método shape da biblioteca Pandas, foi confirmado

que este conjunto de dados possui 2.925.493 linhas e 142 colunas, números que demonstram seu potencial.

Este dataset foi derivado da plataforma LendingClub, uma instituição financeira americana com sede em San Francisco, Califórnia. O volume significativo de dados, com aproximadamente $15,98 bilhões em empréstimos originados até o final de 2015, atesta a relevância da plataforma. O conjunto de dados abrange o período de 2007 até o terceiro trimestre de 2020, o que permite a análise de mudanças e tendências econômicas.

A quantidade e diversidade de atributos financeiros e comportamentais fazem deste dataset uma escolha ideal para modelar o comportamento de crédito dos clientes. A cobertura temporal de mais de uma década é relevante, pois permite analisar como diferentes contextos econômicos podem impactar o comportamento de crédito. Isso aumenta a chance de que os modelos preditivos desenvolvidos generalizem em cenários variados, fornecendo previsões adaptáveis para a criação de campanhas personalizadas de crédito.

4.1.1 Análise de Outras Opções Consideradas

Quatro alternativas foram avaliadas, mas todas apresentaram limitações em comparação com o dataset escolhido. Um exemplo é o "Default Payments of Credit Card Clients in Taiwan from 2005", que foi considerado a segunda melhor opção. Este conjunto de dados possui 30.000 linhas e 30 colunas, o que é inferior em termos de volume e variedade de atributos. Apesar de sua popularidade e do uso frequente em estudos acadêmicos, a quantidade restrita de informações limita a capacidade de capturar a complexidade dos perfis financeiros. Além disso, o dataset cobre apenas o ano de 2005, o que impede a análise de tendências de longo prazo, tornando-o inadequado para modelar mudanças econômicas e comportamentais ao longo do tempo.

Outro dataset analisado foi o “Análise de Crédito - German Data Analysis” disponível no Kaggle, que contém apenas 21 colunas. Essa limitação de atributos dificulta a aplicação de técnicas sofisticadas de aprendizado de máquina e a realização de uma segmentação detalhada dos clientes. Além disso, o pequeno número de registros prejudica a generalização do modelo, reduzindo sua eficácia na recomendação de campanhas personalizadas. Assim, este conjunto de dados se mostrou insuficiente para atender às exigências de uma análise preditiva robusta.

O "Statlog (German Credit Data) Data Set", do UCI Machine Learning Repository, foi outra opção descartada. Embora utilizado em pesquisas acadêmicas, o dataset carece de atualizações recentes, o que significa que seus dados podem não refletir adequadamente às condições econômicas atuais. Além disso, ele não oferece a cobertura temporal necessária para capturar variações econômicas importantes, e sua limitada complexidade de atributos o torna inadequado para modelagens avançadas.

Essas comparações evidenciam a superioridade do "A Credit Risk Predicting Hybrid Model Based on Deep Learning Technology". Sua vasta quantidade de registros, grande

variedade de atributos e cobertura temporal abrangente são elementos que auxiliam no desenvolvimento de um sistema de recomendação eficaz. A escolha deste dataset garante uma base sólida para a construção de modelos preditivos, capazes de fornecer insights e campanhas de crédito adaptadas às necessidades de diferentes perfis de clientes.

4.2 ANÁLISE DO CONJUNTO DE DADOS

Inicialmente, estão listadas e descritas todas as colunas do conjunto de dados no Anexo A do TCC, que foi criado a partir do dicionário de dados da plataforma LendingClub. Para compreender a estrutura e a qualidade do conjunto de dados utilizado neste trabalho, foi realizada uma Análise Exploratória de Dados (EDA). Esse processo de EDA assegurou que os dados estivessem em um formato adequado para a modelagem, com variáveis selecionadas de maneira a capturar a diversidade dos perfis financeiros dos clientes. Para isso, foi realizado um preprocessamento de dados, que envolveu diversas etapas para garantir a consistência e a qualidade do conjunto de dados utilizado.

Conversão de Taxa de Juros: A coluna de taxa de juros, originalmente armazenada como string com símbolo de porcentagem, foi convertida para valores numéricos. Conversão de Datas: A coluna de datas de emissão foi convertida para o formato de ano, facilitando a análise temporal.

Filtragem de Status de Empréstimos: O conjunto de dados foi filtrado para incluir apenas empréstimos com status relevantes, como "Fully Paid", "Charged Off"e "Default". Remoção de Valores Nulos: Linhas com valores ausentes foram removidas para garantir que o conjunto de dados estivesse limpo.

Amostragem de Dados: Uma amostra aleatória de dados foi selecionada para análise, mantendo a consistência com um estado aleatório fixo.

Preenchimento de Valores Nulos e Conversão: Valores ausentes na coluna de datas foram preenchidos com um ano padrão, e a coluna foi convertida para o tipo inteiro.

No conjunto de tabelas 3 e 4, podemos visualizar os dados das 10 primeiras colunas antes do pré-processamento. Já nas figuras 5 e 6, é possível observar as 10 colunas após o processamento.

Tabela 3 – Imagem dos dados do dataset crus sem antes do pré-processamento: Primeiras 5 colunas

Unnamed: 0	**id**	**loan_amnt**	**funded_amnt**	**funded_amnt_inv**
0	1077501	5000.0	5000.0	4975.0
1	1077430	2500.0	2500.0	2500.0
2	1077175	2400.0	2400.0	2400.0
3	1076863	10000.0	10000.0	10000.0
4	1075358	3000.0	3000.0	3000.0

Tabela 4 – Imagem dos dados do dataset crus sem antes do pré-processamento: 5 colunas seguintes

term	**int_rate**	**installment**	**grade**	**sub_grade**
36 months	10.65%	162.87	B	B2
60 months	15.27%	59.83	C	C4
36 months	15.96%	84.33	C	C5
36 months	13.49%	339.31	C	C1
60 months	12.69%	67.79	B	B5

Tabela 5 – Imagem dos dados pós-processamento: Primeiras 5 colunas

id	**loan_amnt**	**int_rate**	**installment**	**emp_length**
1077501	5000.0	10.65	162.87	10
1077430	2500.0	15.27	59.83	0
1076863	10000.0	13.49	339.31	10
1075358	3000.0	12.69	67.79	1
1075269	5000.0	7.90	156.46	3

Tabela 6 – Imagem dos dados pós-processamento: 5 colunas seguintes

home_ownership	**annual_inc**	**addr_state**	**dti**	**earliest_cr_line**
5	2.405670	3	27.65	1985
5	2.425598	10	1.00	1999
5	2.468410	4	20.00	1996
5	2.508769	37	17.94	1996
5	2.441590	3	11.20	2004

Esse preprocessamento assegurou que o conjunto de dados estivesse preparado para a análise e modelagem, capturando a complexidade e diversidade dos perfis financeiros. Este processo contribuiu para identificar padrões, entender a distribuição das variáveis e detectar a presença de outliers ou dados anômalos que poderiam impactar o desempenho dos modelos de machine learning.

Durante a EDA, técnicas como boxplots foram empregadas para visualizar a dispersão dos atributos financeiros e detectar valores extremos. Os boxplots são gráficos que permitem uma análise visual da distribuição dos dados, destacando a mediana, os quartis e os possíveis outliers. A identificação de outliers é relevante, pois esses valores podem influenciar negativamente a performance do modelo se não forem tratados adequadamente.

O gráfico de boxplot na figura 9, revela um aumento progressivo na média e mediana dos valores de empréstimos até o ano de 2013. A partir desse ponto, a distribuição parece se estabilizar, com a quantidade de crédito ofertada se concentrando em valores similares nos anos subsequentes. Além disso, é possível observar a presença de outliers em diversos anos, indicando que, embora a maioria dos empréstimos sigam um padrão consistente, há casos excepcionais de empréstimos com valores significativamente mais altos.

Embora não impacte diretamente o desenvolvimento do trabalho, essa análise sugere que o dataset possui uma grande abrangência temporal. É relevante para o modelo que será criado, pois ele terá a capacidade de analisar e captar a tendência de aumento progressivo dos montantes ao longo dos anos, conforme observado no gráfico. Essa compreensão pode ajudar a identificar padrões históricos que influenciam o comportamento de crédito.

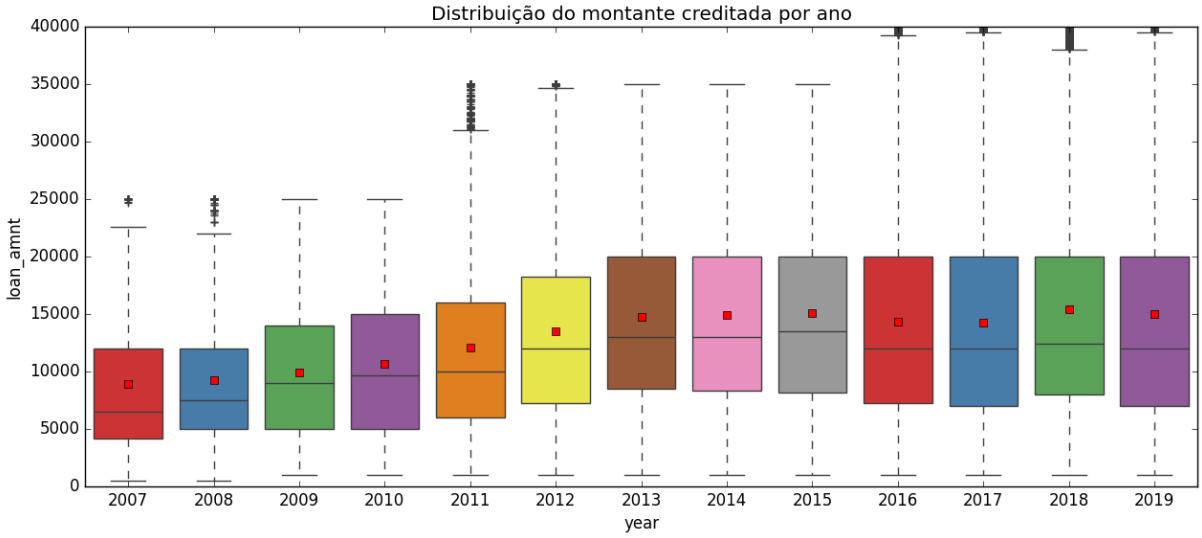


Figura 9 – Distribuição do Montante de Empréstimo por Ano

Além disso, a análise exploratória incluiu a verificação de correlações entre variáveis financeiras. A compreensão dessas relações auxilia na seleção de atributos relevantes, garantindo que o modelo seja robusto e eficiente. A análise de correlações ajuda a evitar a multicolinearidade, que ocorre quando variáveis altamente correlacionadas são usadas simultaneamente no modelo, o que pode prejudicar a precisão das previsões.

Para analisar as relações lineares entre os pares de variáveis, foi utilizado o coeficiente de correlação de Pearson. Este coeficiente mede a força e a direção da relação linear entre duas variáveis numéricas, variando de -1 a 1. Um valor de 1 indica uma correlação linear positiva perfeita, -1 indica uma correlação linear negativa perfeita, e 0 sugere a ausência de uma relação linear significativa.

A figura 10 a seguir ilustra a correlação entre loan_amnt e funded_amnt. A correlação foi de 1.00, indicando uma relação perfeita e positiva. Isso significa que o valor do empréstimo solicitado (loan_amnt) está diretamente associado ao valor financiado (funded_amnt), o que é esperado, dado que o montante financiado normalmente reflete o montante solicitado.

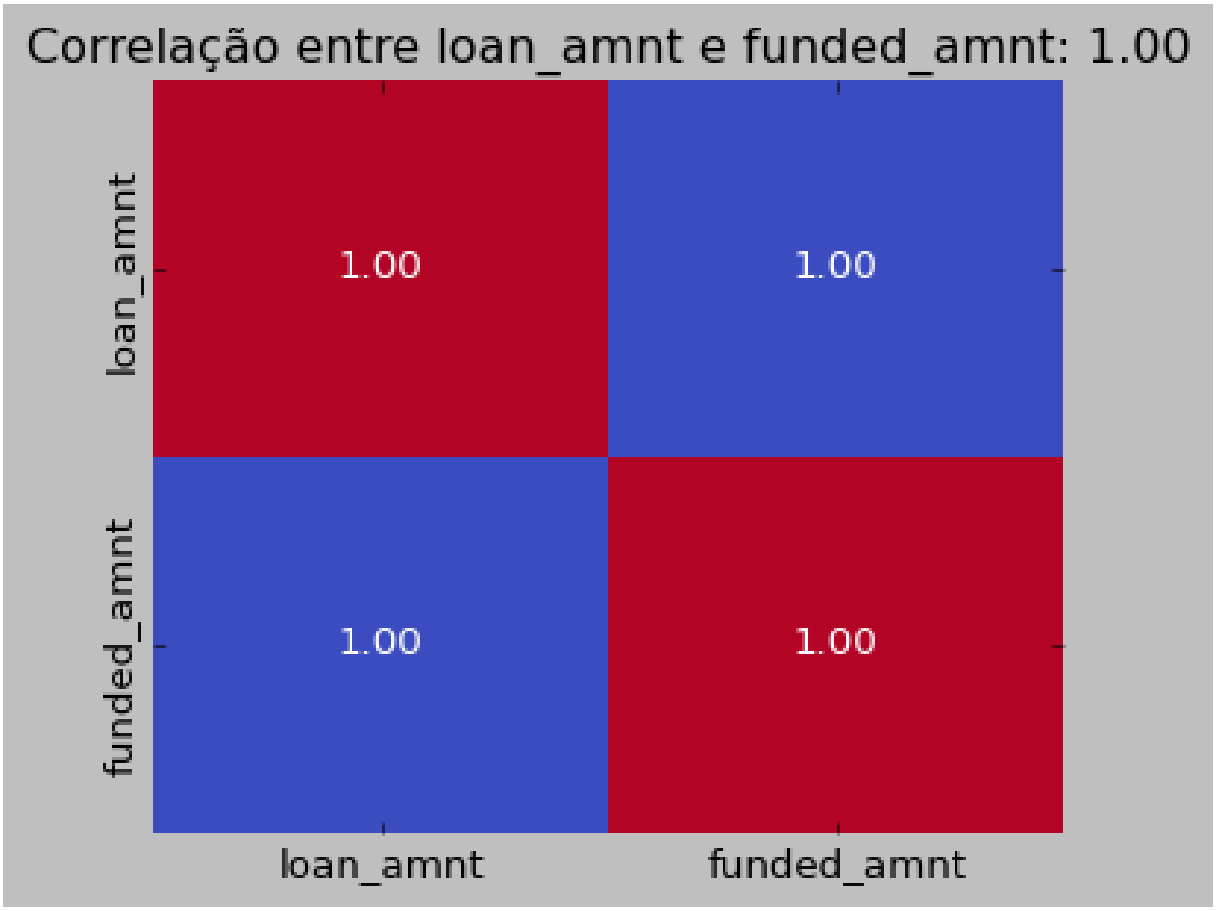


Figura 10 – Correlação entre loan_amnt e funded_amnt

A figura 11 a seguir ilustra a correlação entre fico_range_low e fico_range_high. A correlação foi de 1.00, mostrando uma relação perfeita e positiva. Como esperado, os limites inferior e superior da pontuação FICO estão fortemente correlacionados, já que representam a mesma métrica de risco de crédito.

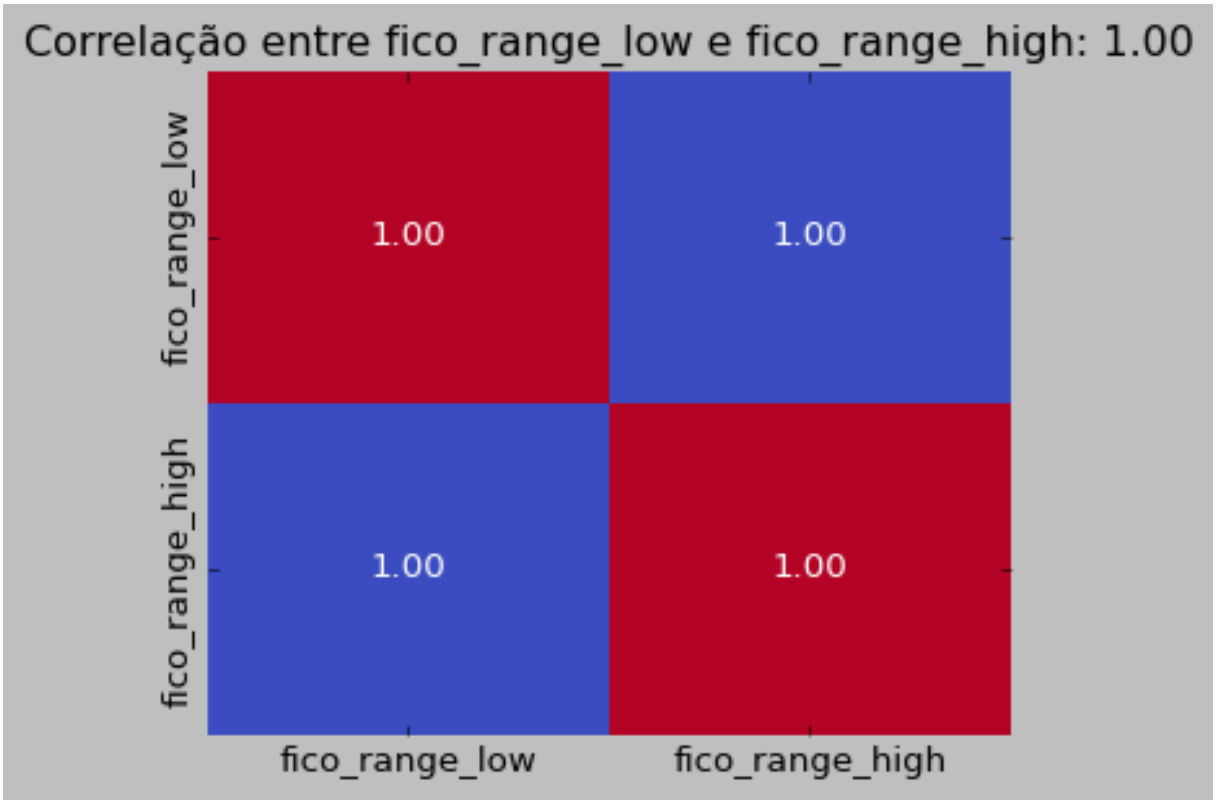


Figura 11 – Correlação entre fico_range_low e fico_range_high

A figura 12 a seguir ilustra a correlação entre total_acc e open_acc. A correlação foi de 0.71, indicando uma relação forte e positiva. O número total de contas (total_acc) tem uma ligação significativa com o número de contas abertas (open_acc), o que é lógico, pois as contas abertas fazem parte do total de contas.

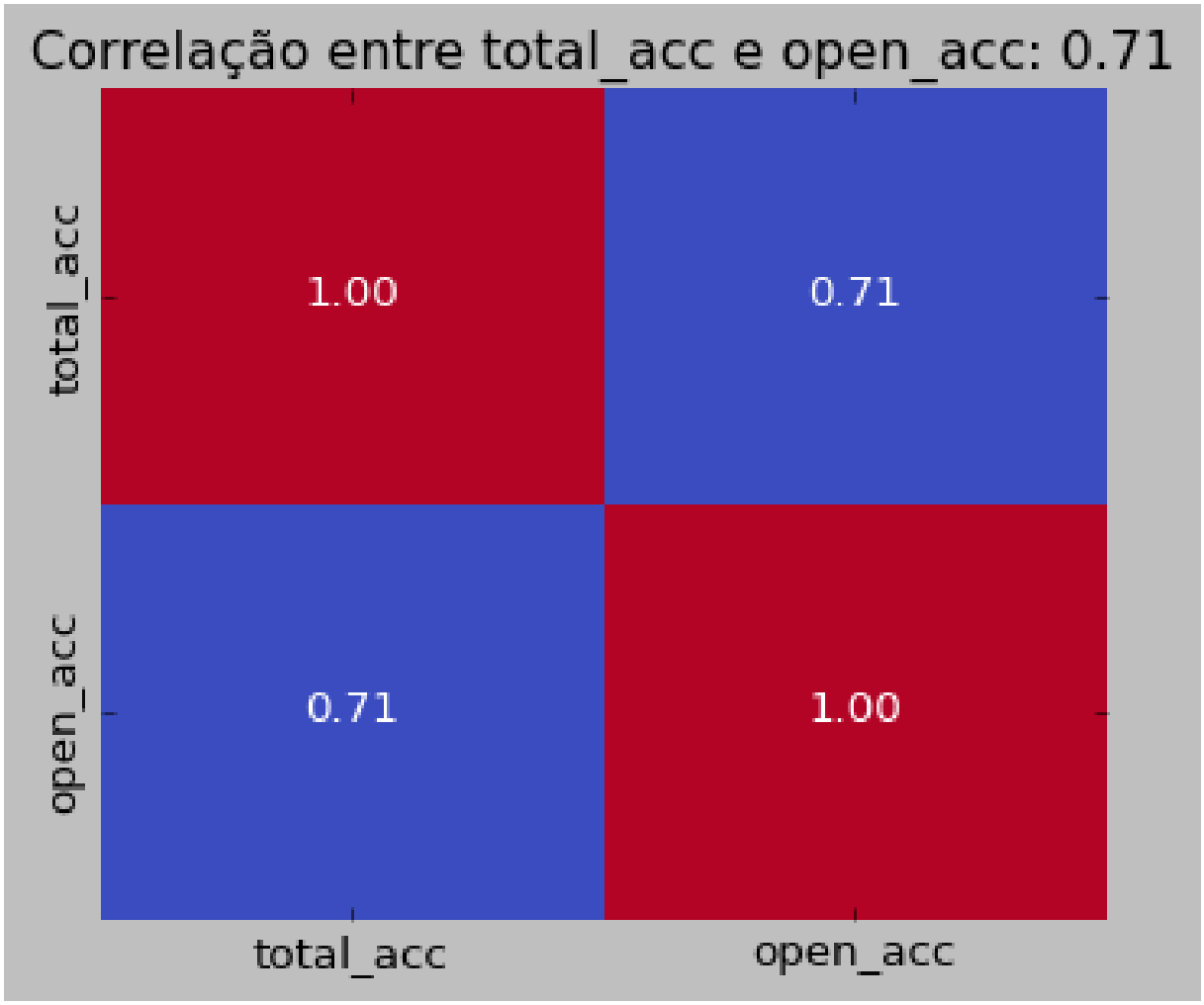


Figura 12 – Correlação entre total_acc e open_acc

O gráfico de barras das Principais Razões de Inadimplência Figura 13 revela que a consolidação de dívidas é a principal causa de inadimplência, seguida pelas dívidas de cartão de crédito. Essa concentração indica que muitos clientes que contratam empréstimos para consolidar dívidas enfrentam dificuldades em honrar seus compromissos financeiros, o que destaca a importância de analisar o risco associado a essas categorias de crédito.

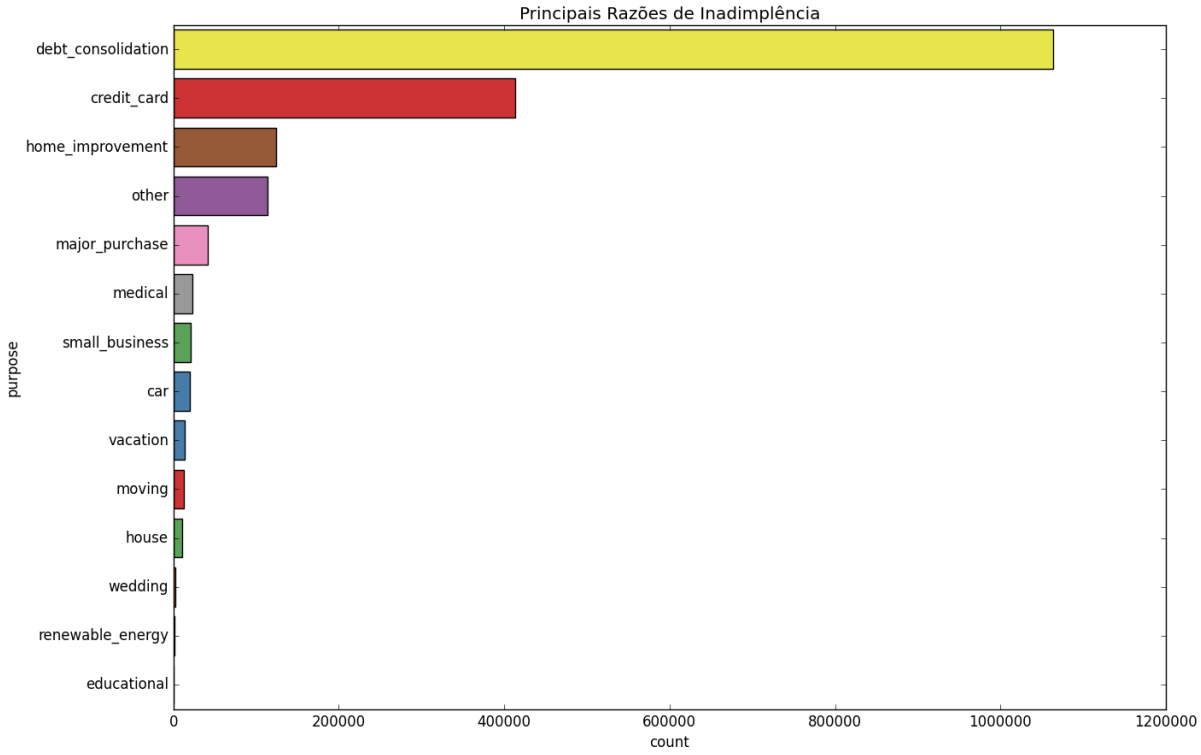


Figura 13 – Gráfico de Barras das Razões de Inadimplência

O mapa de calor que representa o Número de Empréstimos por Estado figura 14 mostra uma concentração significativa de empréstimos na Califórnia, seguida por estados como Texas e Nova York, sugerindo uma alta demanda por crédito em regiões economicamente mais dinâmicas. Reconhecer a distribuição geográfica é importante para entender o comportamento de crédito e pode orientar estratégias regionais de análise de risco e alocação de recursos financeiros.

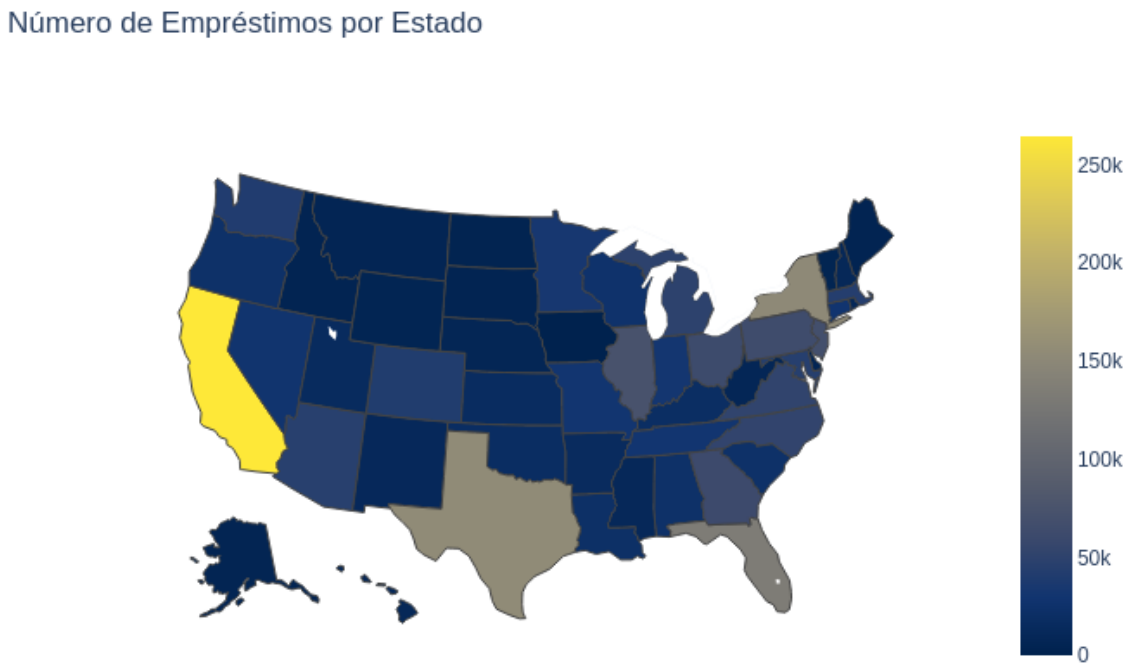


Figura 14 – Mapa Coroplético de Empréstimos por Estado

O gráfico representado na figura 15, mostra a Função de Densidade de Probabilidade para a variável dívida/renda (DTI), que é calculada usando o total de pagamentos mensais da dívida do mutuário sobre suas obrigações totais (excluindo hipoteca e empréstimos em LC), dividido pela renda mensal informada. A Wells Fargo, uma das maiores instituições financeiras dos Estados Unidos, descreve o DTI como um importante indicador da saúde financeira geral, que, além da pontuação de crédito, é avaliado pelos credores para determinar o risco de assumir novas dívidas e o conforto financeiro do mutuário com suas obrigações atuais. Este gráfico evidencia que a distribuição do DTI é altamente assimétrica, com a maioria dos valores concentrados próximos de zero, e destaca a presença de muitos outliers significativos, o que exige cuidado especial na análise de risco de crédito.

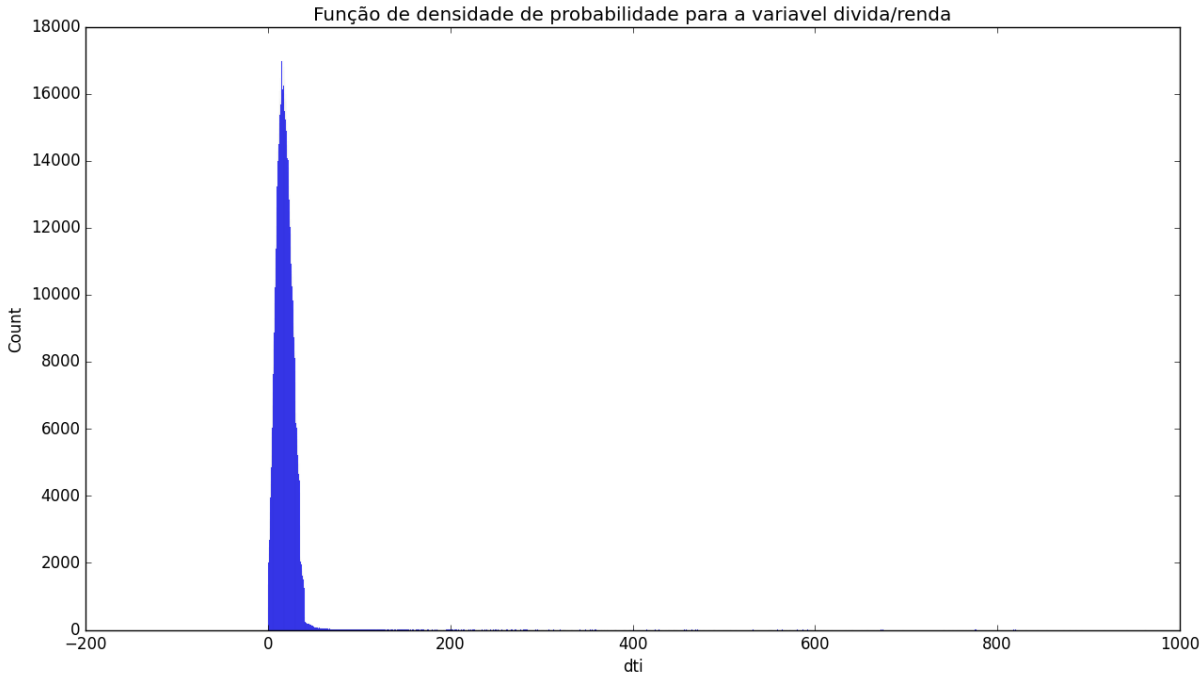


Figura 15 – Distribuição da Relação Dívida/Renda (DTI)

4.3 LIMPEZA DOS DADOS

A limpeza dos dados foi uma etapa usada para garantir a integridade, a consistência e a qualidade do conjunto utilizado no desenvolvimento do sistema de recomendação de crédito. Como os dados financeiros podem conter inconsistências e valores ausentes, foi necessário aplicar um processo de filtragem e tratamento para preparar os dados para a modelagem. As principais ações técnicas realizadas estão descritas a seguir:

Filtragem por Status de Verificação: O conjunto de dados foi inicialmente filtrado para incluir apenas registros com status de verificação como "Verified"ou "Source Verified". Essa decisão foi tomada para garantir que as análises e previsões fossem baseadas em dados mais confiáveis, aumentando a robustez do modelo.

Criação da Variável-Alvo: A variável-alvo, chamada default, foi gerada ao mapear o status do empréstimo (loan_status) em categorias binárias. "Fully Paid"foi mapeado para 0 (não inadimplente), enquanto "Charged Off"e "Default"foram mapeados para 1 (inadimplente).

Análise e Remoção de Valores Ausentes: Uma análise de valores ausentes foi conduzida para identificar colunas com uma alta porcentagem de dados ausentes. Para isso, foi utilizado um método que calculou a quantidade e a proporção de valores ausentes em cada coluna. Variáveis com mais de 90% de valores ausentes foram removidas, uma prática necessária para evitar introduzir viés ou comprometer a qualidade do modelo. Ao final, 30 colunas foram removidas.

Tratamento de Variáveis Irrelevantes: A coluna, chamada de Unnamed: 0, foi identificada como irrelevante, pois não tem identificação e impossibilita o objetivo da

análise. Essa filtragem garantiu que apenas variáveis úteis e informativas fossem mantidas, facilitando o desenvolvimento do sistema de recomendação.

Tratamento de Valores Ausentes com Substituições: Colunas como max_bal_bc e mort_acc foram preenchidas com 0, o que indica que a ausência de valores pode ser interpretada como a ausência da característica representada (por exemplo, não ter saldo ou conta hipotecária). Outras variáveis, como mths_since_last_delinq e mths_since_recent_revol_delinq, foram preenchidas com -1, representando a ausência de eventos relevantes, como nunca ter tido inadimplência.

Verificação de Assimetria: A análise revelou uma assimetria em atributos financeiros, como annual_inc (Renda Anual) e open_acc (Número de Linhas de Crédito Abertas). Para corrigir isso, aplicou-se a transformação logarítmica, que evita problemas com valores zero, já que o logaritmo de zero é indefinido, comprime valores extremos e estabiliza a variância, tornando essas variáveis mais adequadas para modelagem

Variável annual_inc (Renda Anual): A distribuição da renda anual apresentou uma cauda longa à direita, indicando que a maioria dos clientes possui rendas relativamente baixas, com poucos casos de rendas muito altas. Essa assimetria pode levar o modelo a ter dificuldades em capturar padrões relevantes, especialmente em algoritmos que assumem distribuições mais simétricas. Para corrigir essa assimetria e tornar a variável mais adequada para a modelagem, aplicou-se a transformação logarítmica, adicionando +1. Isso comprime os valores mais altos e torna a distribuição mais uniforme, melhorando a estabilidade do modelo. Figura 16 antes: mostra a distribuição assimétrica original. Figura 17 depois: mostra a distribuição mais uniforme após a transformação.

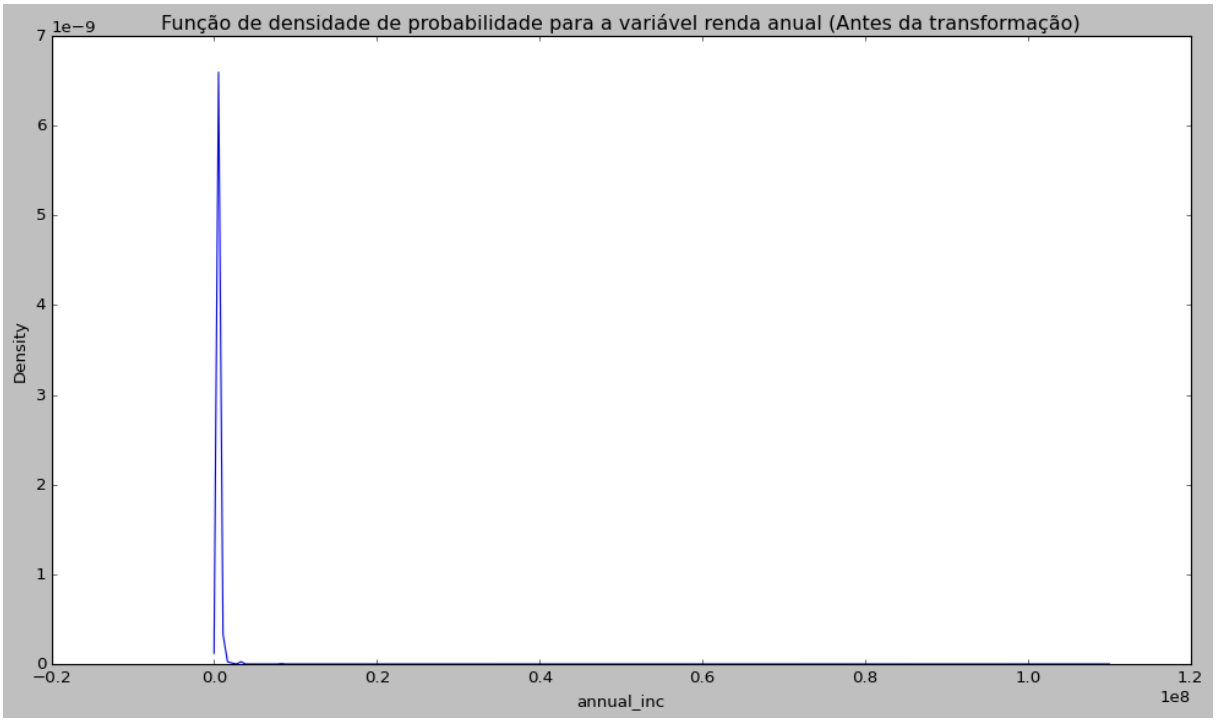


Figura 16 – Função de densidade de probabilidade para a variável renda anual (Antes da transformação)

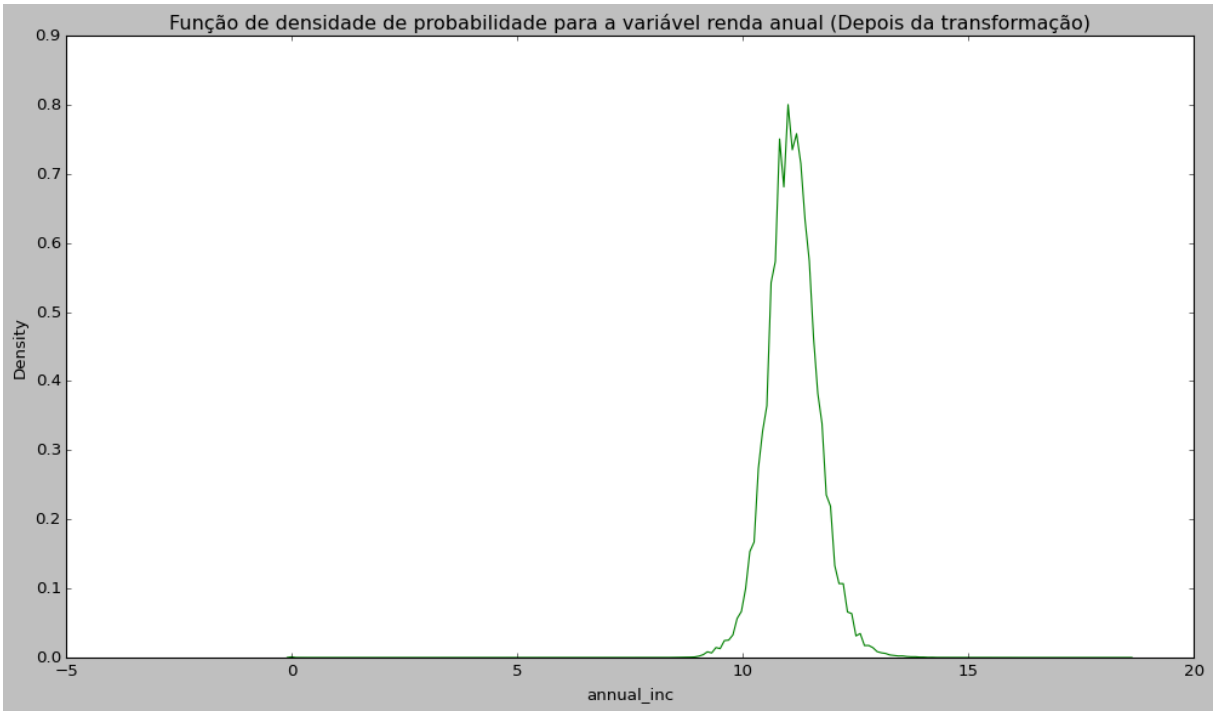


Figura 17 – Função de densidade de probabilidade para a variável renda anual (Depois da transformação)

Variável open_acc (Número de Linhas de Crédito Abertas): A variável open_acc também mostrou uma distribuição assimétrica, com a maioria dos clientes tendo um número baixo de linhas de crédito e poucos casos com muitos. Essa assimetria pode afetar a interpretação e o desempenho preditivo, pois os valores mais altos podem ter um impacto desproporcional. Aplicou-se a transformação logarítmica adicionando +1 para suavizar a cauda longa e distribuir os dados de forma mais equilibrada, facilitando o aprendizado do modelo. Figura 18 antes: mostra a assimetria significativa antes da transformação. Figura 19 depois: destaca a distribuição suavizada e mais equilibrada após a transformação.

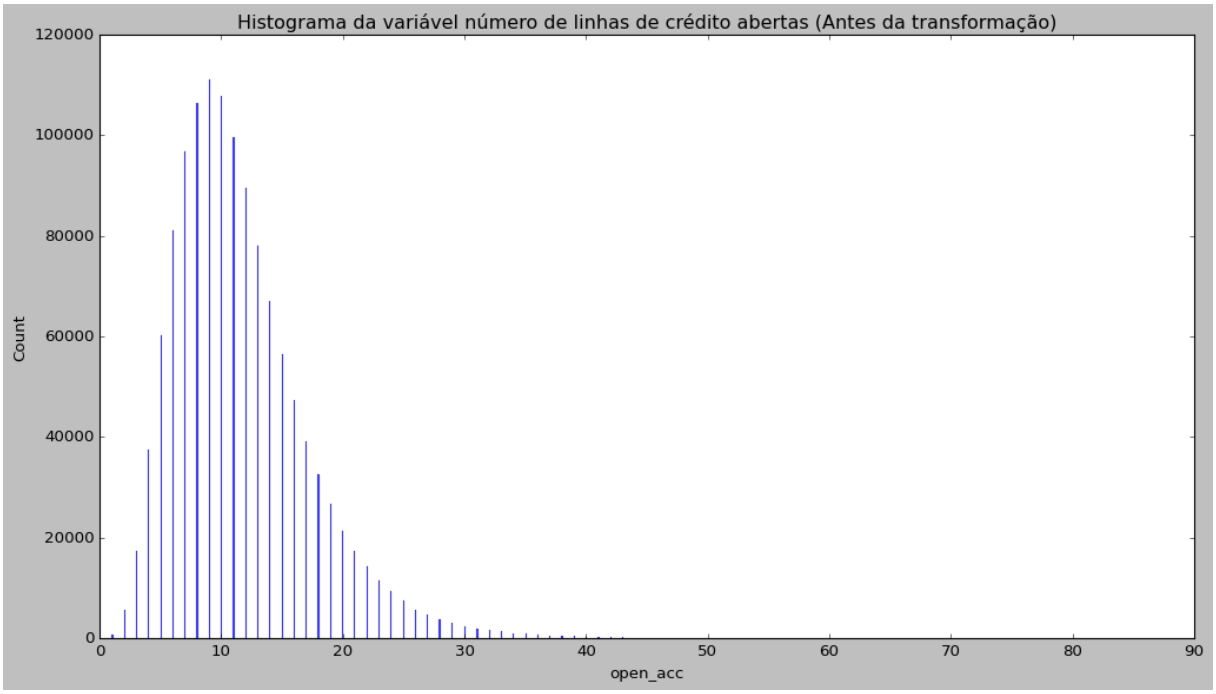


Figura 18 – Histograma da variável número de linhas de crédito abertas (Antes da transformação)

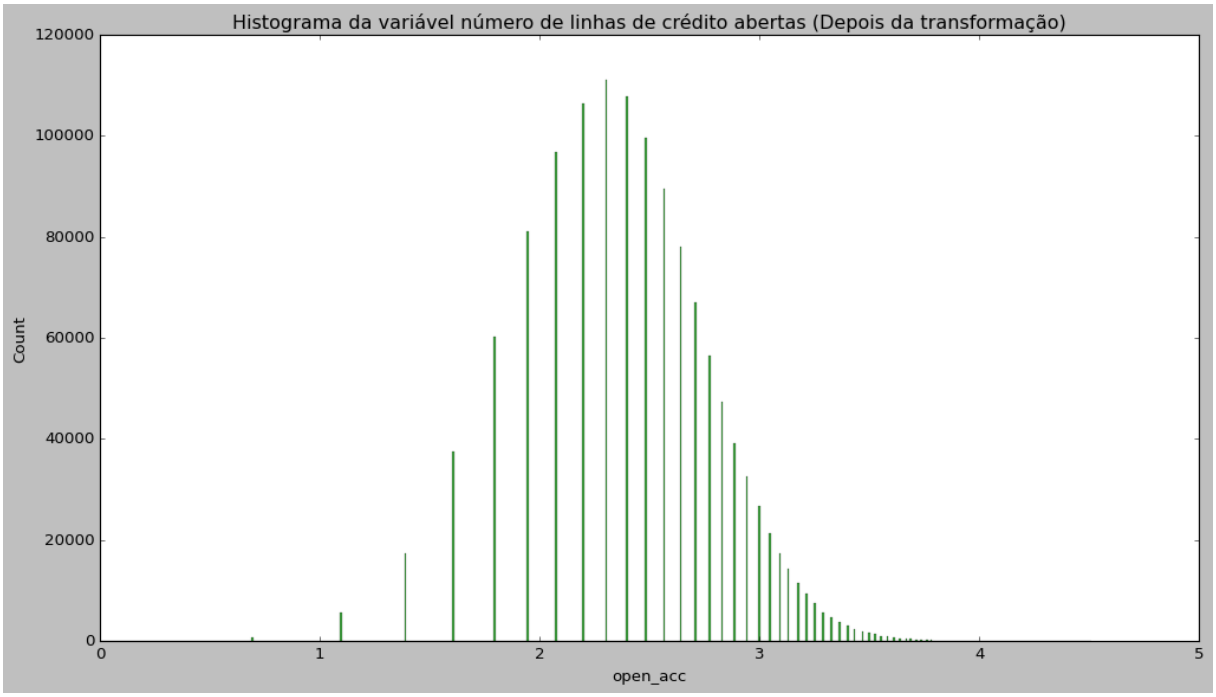


Figura 19 – Histograma da variável número de linhas de crédito abertas (Depois da transformação)

4.4 PROVA DE CONCEITO 1 (POC 1)

4.4.1 Seleção de Atributos

Foram escolhidos quatro atributos principais, que podem ser verificados no capítulo 2.1 Colunas Selecionadas para Poc 1 do artigo em anexo. Esses atributos foram escolhidos por serem indicativos da saúde financeira e do comportamento de crédito dos clientes, considerando uma abordagem mais simplificada para a análise de risco de crédito. A descrição dos atributos, bem como de todas as variáveis disponíveis no conjunto de dados, pode ser consultada no artigo Introdução ao Conjunto de Dados, no Capítulo 2.1.

4.4.2 Criação da Campanha

Para validar a metodologia, foi realizada uma PoC 1 usando uma técnica de Recomendação Baseada em Conteúdo, com métricas de Distância/Similaridade. Especificamente, utilizou-se a Distância Euclidiana para comparar os atributos dos clientes com os critérios de cada campanha de crédito, garantindo uma avaliação precisa.

Duas campanhas de crédito distintas foram desenvolvidas, cada uma com critérios de elegibilidade baseados nos atributos selecionados. A PoC 1 demonstra como essas campanhas podem ser personalizadas para diferentes perfis financeiros, garantindo uma segmentação inicial:

Campanha 1: Focada em clientes com uma renda anual mínima de 30.000, uma utilização de crédito máxima de 60%, pelo menos uma conta aberta nos últimos 24 meses e nota de crédito entre B e E, visando perfis de risco moderado a alto.

Campanha 2: Voltada para clientes com renda anual mínima de 60.000, utilização de crédito máxima de 35%, ao menos cinco contas abertas nos últimos 24 meses e nota de crédito entre A e C, direcionada a perfis de menor risco.

A função de Distância Euclidiana calculou a similaridade entre os atributos dos clientes e os critérios de cada campanha, com um limite de 20 para qualificação, incluindo apenas clientes altamente compatíveis. Para ilustrar a precisão do modelo, foram utilizados dois exemplos: Vinicius, que se qualificou exclusivamente para a Campanha 1, e Elder, configurado para atender somente à Campanha 2. Esses exemplos práticos evidenciam a eficácia da segmentação. Se as duas campanhas funcionaram e passarem nos testes, a lógica é que funcionará para o restante do dataset, garantindo a robustez do modelo.

4.4.3 Aplicação de Inteligência Artificial

A PoC 1 utiliza um modelo de Recomendação Baseada em Conteúdo para qualificar clientes para campanhas específicas, comparando atributos dos clientes com critérios de elegibilidade preestabelecidos para cada campanha. Esta abordagem inicial permite validar a eficácia da metodologia e identificar ajustes necessários antes de uma implementação

completa.

A PoC 1 foi implementada como um experimento para testar a lógica de recomendação baseada em métricas de similaridade. A Distância Euclidiana foi utilizada para medir a adequação dos clientes às campanhas, validando a abordagem de forma prática. O principal objetivo do TCC é desenvolver um sistema de recomendação de crédito que possa analisar perfis de clientes e sugerir campanhas de crédito adequadas. A PoC 1 responde a este objetivo ao oferecer uma primeira implementação que compara os atributos dos clientes com critérios de elegibilidade predefinidos para diferentes campanhas. Assim, ela verifica se a metodologia inicial pode, de fato, segmentar clientes de maneira eficaz e fornecer recomendações personalizadas.

Além disso, a PoC 1 ajuda a validar a escolha dos atributos e critérios, demonstrando que os elementos escolhidos são relevantes para prever a adequação dos clientes às campanhas de crédito. Essa validação inicial é útil antes de aplicar técnicas mais complexas de aprendizado de máquina.

4.4.4 Análise dos Resultados

As campanhas foram estruturadas para atingir públicos com perfis de renda distintos. A Campanha 1 foi projetada para um público de menor renda, alinhando seus critérios a perfis financeiros mais modestos, enquanto a Campanha 2 buscou atrair clientes com maior poder aquisitivo, ajustando seus critérios para uma renda mais elevada. Essa segmentação foi bem-sucedida, conforme mostrado na figura 20. Na amostra analisada, a Campanha 1 aprovou cerca de 40 clientes, enquanto a Campanha 2, focada em alta renda, aprovou quase 90. Esses resultados confirmam que ambas as campanhas foram eficazes em atrair seus públicos-alvo específicos, demonstrando que a segmentação por faixa de renda foi apropriada para atender a diferentes perfis de mercado.

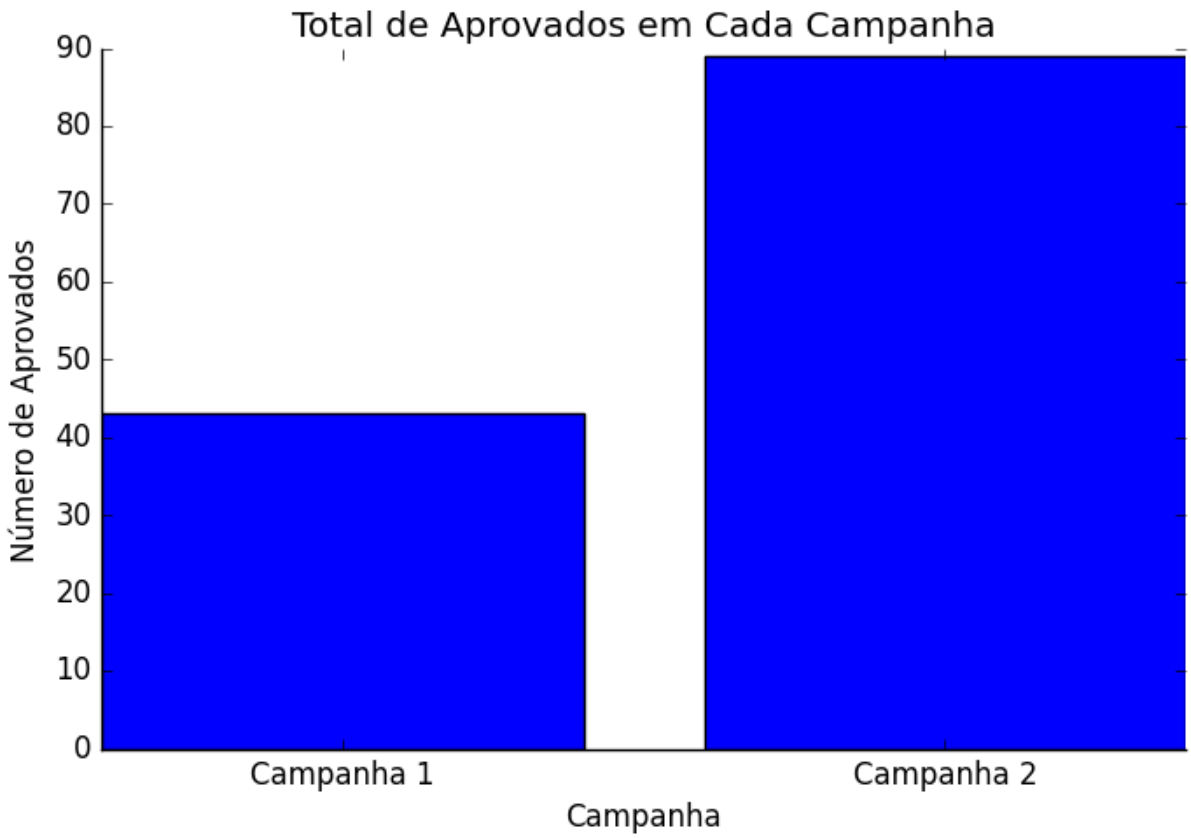


Figura 20 – Total de aprovados em cada campanha

Embora a PoC 1 tenha mostrado a capacidade do modelo de diferenciar clientes com base nos critérios estabelecidos, também evidenciou algumas limitações. Por exemplo, apesar de muitos clientes terem uma baixa distância em relação aos critérios da Campanha 1, apenas cerca de 40 entre os 10.000 perfis foram aprovados, como mostrado na figura 21. Isso indica que, mesmo com alta similaridade, a campanha aplicou critérios para selecionar apenas os clientes mais alinhados ao perfil desejado.

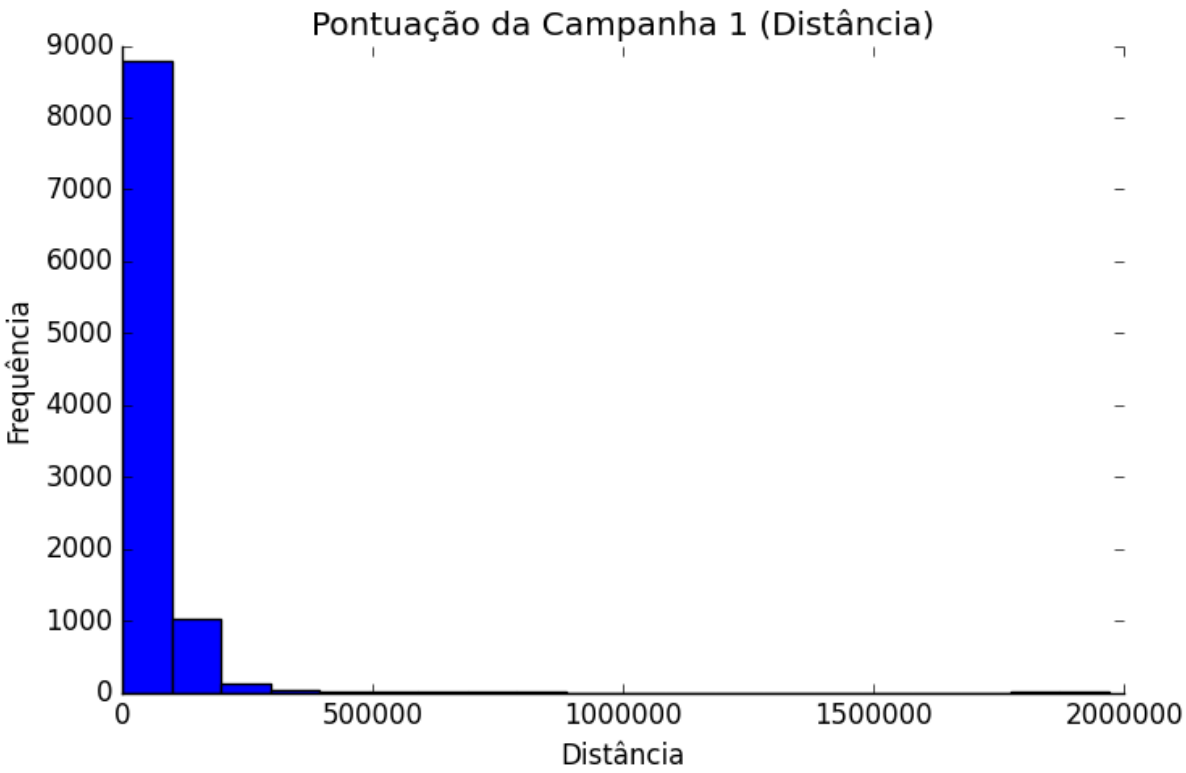


Figura 21 – Total de aprovados campanha 1

Na Campanha 2, apesar de muitos clientes apresentarem baixa distância em relação aos critérios, foram aprovados quase 90 entre os 10.000 perfis, conforme a figura 22. Esse resultado demonstra que a campanha foi seletiva o suficiente para qualificar apenas os clientes mais adequados ao perfil de alta renda, apesar da similaridade ampla.

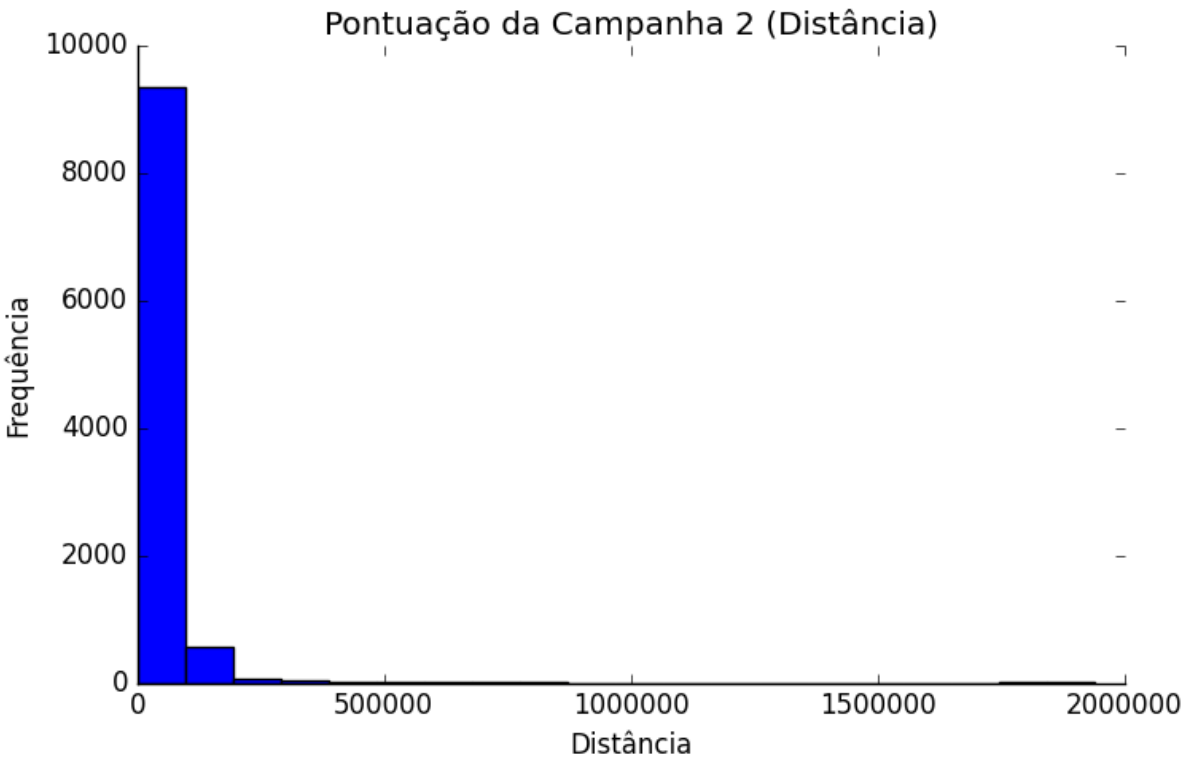


Figura 22 – Total de aprovados campanha 2

Limitações da Prova de Conceito A PoC 1 tem como limitações a utilização de critérios rígidos para qualificação, que restringiu a adaptação dinâmica dos parâmetros, e a falta de personalização impediu uma análise mais detalhada que considerasse interações complexas entre variáveis ou o histórico de relacionamento dos clientes com campanhas passadas. Essas limitações apontam a necessidade de métodos mais avançados, como a regressão logística, que será incorporada na versão futura para melhorar a flexibilidade e a precisão das recomendações.

Atributo	**Vinicius**	**Elder**
user_id	Vinicius	Elder
annual_inc	30000.0	60000.0
all_util	55.0	35.0
acc_open_past_24mths	1.0	5.0
grade	C	A
campaign_1_distance	5.0	30000.010433
campaign_1_qualified	True	False
campaign_2_distance	30000.004017	5.0
campaign_2_qualified	False	True

Tabela 7 – Resultados da Prova de Conceito

4.5 PROVA DE CONCEITO 2 (POC 2)

4.5.1 Seleção de Atributos

Na PoC 2, o modelo foi expandido para incluir um conjunto de 43 variáveis do dataset e uma variável adicional, fabricada especificamente para agregar valor preditivo. O total de 44 variáveis fornece uma visão abrangente do perfil financeiro dos clientes, combinando atributos diretamente relacionados ao comportamento de crédito com uma métrica derivada que melhora a capacidade de previsão do modelo.

Cada variável foi escolhida com base na sua relevância e impacto no desempenho do modelo preditivo. A variável criada, monthly_load, foi desenvolvida para medir a carga financeira mensal em relação à renda anual do cliente, acrescentando um elemento de análise que não estava disponível diretamente no conjunto de dados original.

4.5.2 Criação da Campanha

Uma campanha promocional foi desenvolvida especificamente para as classes D e E, utilizando um método de interpolação linear para personalizar as taxas de juros com base no perfil de crédito dos clientes. A interpolação linear é uma técnica matemática que estima um valor desconhecido dentro de um intervalo, utilizando dois valores conhecidos. Nesse contexto, são considerados o score do cliente, o score máximo e mínimo de sua categoria, além das taxas de juros máxima e mínima correspondentes.

O cálculo da nova taxa é feito com base no score individual do cliente. Se o cliente estiver na categoria D ou E, o sistema usa a interpolação linear para ajustar a taxa de juros de forma justa e precisa, garantindo que ela reflita melhor o perfil de risco de cada cliente. Dessa forma, a campanha é capaz de oferecer condições personalizadas, motivando clientes a manter ou melhorar seu comportamento de crédito para se beneficiar de taxas mais favoráveis.

4.5.3 Aplicação de Inteligência Artificial

Nesta etapa da Poc 2, foi incluída uma técnica de interpolação linear para o cálculo das taxas de juros dos clientes, após a aplicação de um modelo de regressão logística. A inclusão dessas técnicas aprimorou a recomendação de crédito ao permitir uma segmentação mais refinada e personalizada, aumentando a precisão na determinação das taxas de juros e categorias de risco.

Implementação da Regressão Logística

Primeiramente, foi empregada uma Regressão Logística para prever a categoria de risco de crédito de cada cliente com base em variáveis financeiras e comportamentais selecionadas. Esse modelo probabilístico permitiu estimar a probabilidade de um cliente pertencer a uma categoria específica, utilizando atributos como renda anual, utilização de

crédito, número de contas abertas recentemente e histórico de inadimplência. A regressão logística foi configurada para fornecer uma pontuação de crédito que, posteriormente, foi utilizada para categorizar os clientes nas classes de risco (A, B, C, D, etc.), conforme os critérios das campanhas.

Cálculo da Taxa de Juros com Interpolação Linear

Após a classificação dos clientes, foi aplicada uma interpolação linear para determinar a taxa de juros correspondente ao score obtido pela regressão logística. Este processo permitiu um cálculo dinâmico e preciso da taxa de juros dentro do intervalo de cada categoria. A interpolação linear utilizou os limites de pontuação de cada categoria e suas respectivas taxas de juros mínimas e máximas, garantindo que a taxa atribuída fosse proporcional ao score do cliente dentro da sua categoria.

Por exemplo, para um cliente classificado na categoria "B"com um score intermediário entre os limites da categoria, a interpolação linear ajustou a taxa de juros de forma proporcional, resultando em uma taxa personalizada que reflete melhor o perfil de risco individual do cliente.

Exemplo de Aplicação com Interpolação Linear

Para ilustrar o processo, considere o cálculo de taxa para um cliente com score de 750 na categoria "B". Utilizando os limites de pontuação da categoria B (725 a 825) e as taxas correspondentes (13,33% a 16,08%), a interpolação linear ajusta a taxa do cliente com base na posição de seu score dentro desse intervalo, resultando em uma taxa de juros precisa que reflete seu perfil de crédito.

Definição de Categorias e Ofertas de Taxas de Juros

Após verificar que o modelo estava adequado, procedemos à definição das categorias de crédito dos indivíduos, elaborando um modelo para a escolha de ofertas de taxas de juros com base nos scores preditivos.

O score de crédito de cada cliente foi calculado utilizando a probabilidade predita pelo método predict_proba da biblioteca sklearn. O modelo gerou previsões, que foram convertidas em um dataframe chamado logistic_regression_preds_dataframe. Esse dataframe foi combinado com o conjunto de dados original, resultando em uma nova coluna chamada score_credito, que multiplica o valor predito por 1000 para facilitar a classificação.

Para categorizar os indivíduos, dividimos o intervalo de scores em faixas que correspondem a diferentes categorias de crédito, com as seguintes definições:

Intervalos de Score de Crédito: (175, 225, 350, 475, 600, 725, 825, 900)

Categorias de Score: ['G', 'F', 'E', 'D', 'C', 'B', 'A']

A nova coluna Categoria foi criada utilizando a função pd.cut() para atribuir a cada indivíduo uma categoria específica com base no seu score_credito. Essa categorização permite ajustar as ofertas de taxas de juros de acordo com o perfil de risco de cada cliente, garantindo uma abordagem personalizada e eficiente para a análise de crédito. O

dataframe final, raw_data_score_final, contém as colunas Categoria e score_credito, que têm importância na segmentação dos clientes e na definição das ofertas de crédito.

Análise dos Resultados

Os resultados da PoC 2 indicaram que o uso combinado da regressão logística e da interpolação linear contribuiu para uma recomendação de crédito ajustada e personalizada. O sistema demonstrou a capacidade de classificar os clientes nas categorias de risco, utilizando a regressão logística, e de calcular taxas de juros para cada cliente por meio da interpolação linear, refletindo a posição do cliente dentro da sua categoria de risco.

Essa abordagem permitiu uma segmentação mais eficaz, que ajusta as taxas e categorias conforme o perfil financeiro específico de cada cliente, alinhando-se aos objetivos do sistema de recomendação de crédito. A lógica implementada mostrou resultados satisfatórios nos exemplos de teste, e espera-se que seja aplicável ao restante do dataset, sugerindo robustez do modelo para diferentes perfis de clientes. No entanto, é necessário analisar o impacto das classificações incorretas, como falsos positivos e negativos, no desempenho geral.

Seleção e Transformação das Variáveis

Nesta etapa, foram selecionadas 43 variáveis principais para o modelo, com base na relevância preditiva de cada uma no contexto da análise de risco de crédito. As variáveis escolhidas abrangem aspectos financeiros, históricos e comportamentais dos clientes. A seguir, explicamos as transformações realizadas e a criação de uma nova variável importante:

Seleção de Variáveis: As variáveis selecionadas, como fico (pontuação de crédito), loan_amnt (montante do empréstimo) e dti (relação dívida/renda), foram escolhidas pela sua capacidade de fornecer informações valiosas sobre o risco de inadimplência. Variáveis redundantes ou com pouca relevância foram removidas para manter a eficiência do modelo.

Correção de Casos Específicos: Algumas variáveis com valores ausentes foram corrigidas. Por exemplo, emp_length foi preenchido com '< 1 year' para valores nulos, enquanto variáveis textuais como emp_title, last_credit_pull_d, last_pymnt_d, title, e zip_code foram preenchidas com strings vazias ou valores padrão.

Mapeamento e Codificação: O tempo de emprego (emp_length) foi convertido de texto para valores inteiros usando um mapeamento personalizado, o que facilitou o uso da variável no modelo. Além disso, a variável earliest_cr_line foi transformada para reter apenas o ano, removendo o mês, o que simplifica a análise temporal.

Codificação de Variáveis Categóricas: As variáveis categóricas, como home_ownership, addr_state, initial_list_status, e application_type, foram convertidas em valores numéricos usando a técnica de codificação por rótulo (label encoding). Para variáveis categóricas como term, purpose e sub_grade, foi aplicada a codificação one-hot, o que expande as categorias em múltiplas colunas binárias, facilitando a entrada de dados no modelo.

Criação da variável monthly_load: A variável monthly_load foi criada para repre-

sentar a carga máxima destinada aos pagamentos mensais do cliente em relação à sua renda anual, expressa em percentual. Essa variável foi projetada para avaliar a capacidade de pagamento do cliente e é calculada pela fórmula ((installment×12)/annual_inc)×100, onde installment é o valor mensal do empréstimo e annual_inc corresponde à renda anual do cliente. Nos casos em que a renda anual era igual a zero, foi atribuído o valor de -1 à variável, para sinalizar exceções e evitar inconsistências no cálculo. Essa métrica é crucial para mensurar a proporção da renda comprometida com o empréstimo, indicando o limite financeiro que o cliente pode suportar. Valores altos de monthly_load sinalizam maior risco de inadimplência, tornando essa variável um indicador essencial no modelo preditivo de risco de crédito.

Adequando o Modelo

Após a preparação e transformação dos dados, foi necessário garantir que o modelo estivesse configurado adequadamente para a análise preditiva. Primeiro, a variável criada, monthly_load, foi incluída como uma métrica para capturar a carga financeira mensal em relação à renda anual do cliente. Essa variável destacou-se em terceiro lugar na lista de importância das variáveis no modelo XGBoost, comprovando sua relevância na previsão de inadimplência.

Em seguida, foi realizada a divisão dos dados em conjuntos de treino e teste, garantindo que o preenchimento de valores ausentes com a mediana fosse feito apenas no conjunto de treino, evitando o vazamento de informações e assegurando a integridade do processo preditivo. Além disso, o desbalanceamento do banco de dados foi tratado utilizando a técnica de Random Undersampling, o que ajustou o número de exemplos de pagantes e não pagantes, melhorando a performance do modelo.

Importância das Variáveis no Modelo XGBoost

O gráfico representado pela figura 25, revelou que as variáveis mais influentes foram year (com um F score de 77.0) e int_rate (com um F score de 53.0), com a variável monthly_load próxima, com um F score de 44.0. Esse resultado destaca a importância de monthly_load, mostrando que a métrica desempenha um papel importante na avaliação da capacidade de pagamento do cliente. A inclusão dessa variável foi uma escolha estratégica, aumentando a precisão do modelo e capturando melhor os fatores que influenciam o risco de inadimplência.

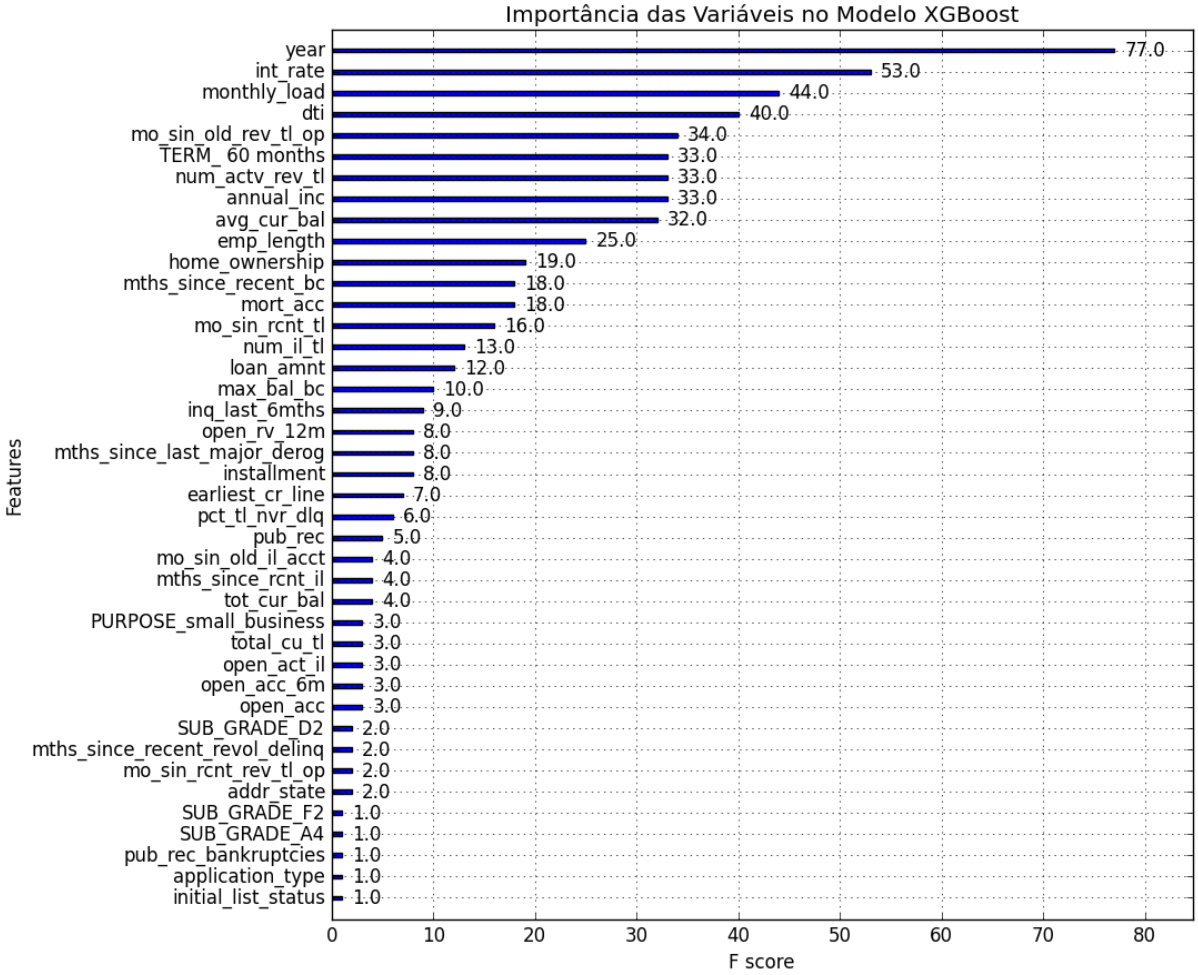


Figura 23 – Importância das Variáveis (Features) no Modelo XGBoost

Performance do Modelo com Todas as Variáveis

O modelo XGBoost foi treinado com todas as 44 variáveis selecionadas, e a performance foi avaliada usando o ROC AUC Score, que foi de 65.47%. As métricas de precisão, recall e F1-score revelaram que o modelo tem um desempenho moderado, com melhor precisão para a classe 0 (pagantes) e recall mais alto para a classe 1 (não pagantes). A matriz de confusão mostrou um desequilíbrio entre as previsões corretas e incorretas, com uma taxa de falsos negativos. A seguir, serão apresentados os gráficos da matriz de confusão e curva ROC, detalhando o desempenho do modelo.

Remoção de Variáveis com Menor Impacto

Ao observar que a performance do modelo não diminuiu consideravelmente após a remoção de várias variáveis de baixo impacto, optou-se por manter apenas as mais relevantes, como year, monthly_load, int_rate, avg_cur_bal, fico, e outras selecionadas com base na importância apresentada pelo modelo XGBoost. Essa estratégia de redução de variáveis é benéfica, pois simplifica o modelo, reduz a possibilidade de viés e mantém a eficiência computacional sem comprometer a precisão significativamente.

Avaliação das Métricas e Ajustes

A matriz de confusão foi gerada a partir das previsões realizadas pelo modelo em relação aos rótulos reais do conjunto de dados. Esses rótulos, utilizados como padrão ouro para validação, representam o histórico real de pagamentos dos mutuários, informados pelas instituições financeiras que geraram o dataset. A variável alvo utilizada no estudo foi construída com base nesses rótulos, classificando os empréstimos como “Fully Paid” (totalmente pagos) ou “Charged Off” e “Default” (inadimplentes).

A análise da matriz de confusão permitiu identificar a precisão do modelo e avaliar a ocorrência de erros, como falsos positivos e falsos negativos. Esses erros indicam, respectivamente, situações em que o modelo previu inadimplência para um cliente adimplente e vice-versa. Essa análise é fundamental para compreender os limites do modelo e planejar ajustes para melhorar sua performance preditiva.

Além disso, devido ao desbalanceamento natural das classes no conjunto de dados, foi necessário aplicar técnicas de balanceamento, como o Random Undersampling. Essa técnica ajustou a proporção entre as classes "inadimplentes"e "adimplentes"no conjunto de treino, reduzindo o viés do modelo em prever a classe majoritária. O processo de validação foi conduzido utilizando os dados reais como referência, garantindo que a avaliação das métricas refletisse o comportamento observado no histórico dos clientes.

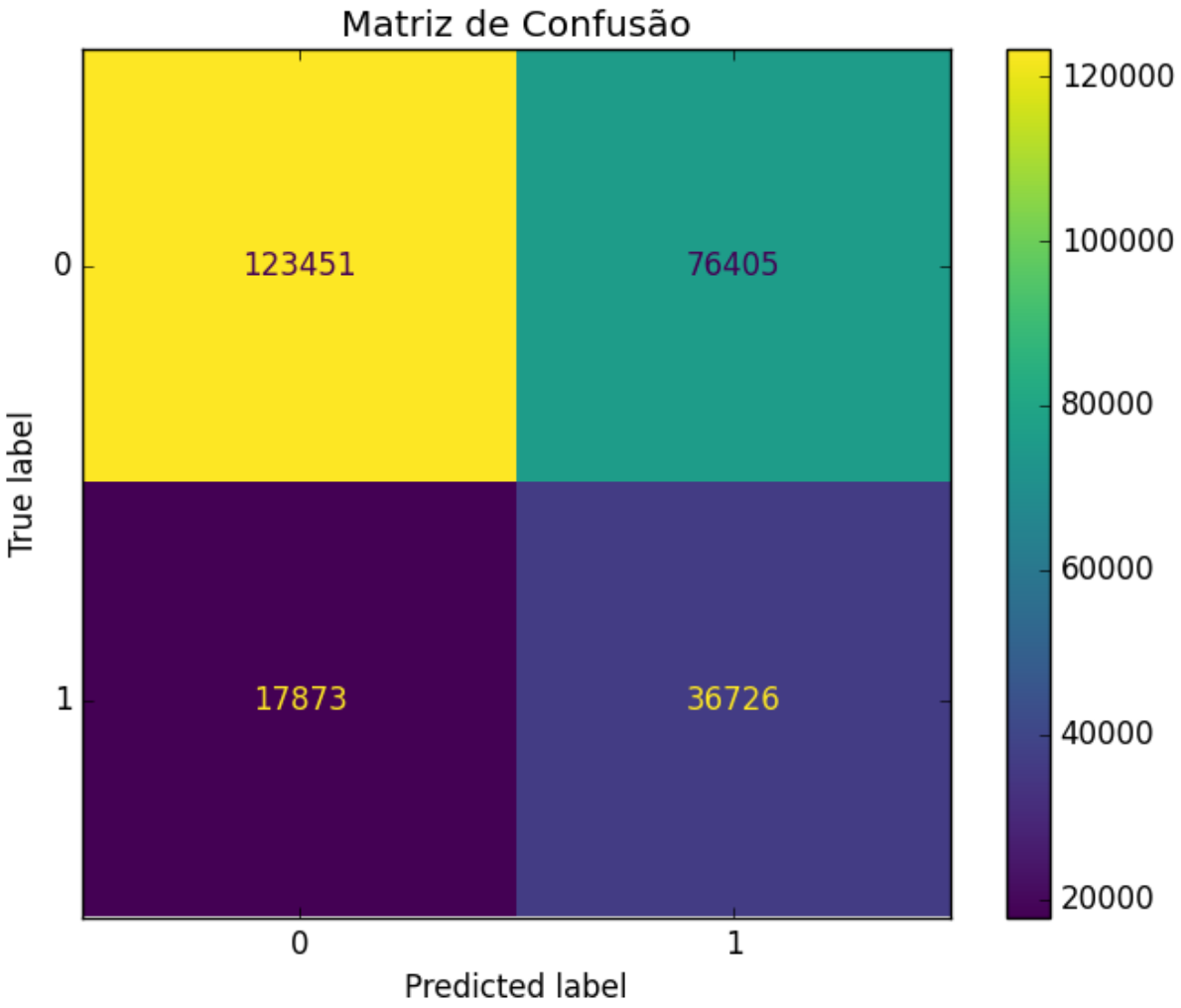


Figura 24 – Matriz de Confusão

Curva ROC: O gráfico representado pela figura 25 foi utilizado para avaliar a performance do modelo de classificação. Uma AUC de 0,65 no conjunto de teste e 0,66 no conjunto de treinamento indicam um desempenho moderado, destacando áreas para otimização.

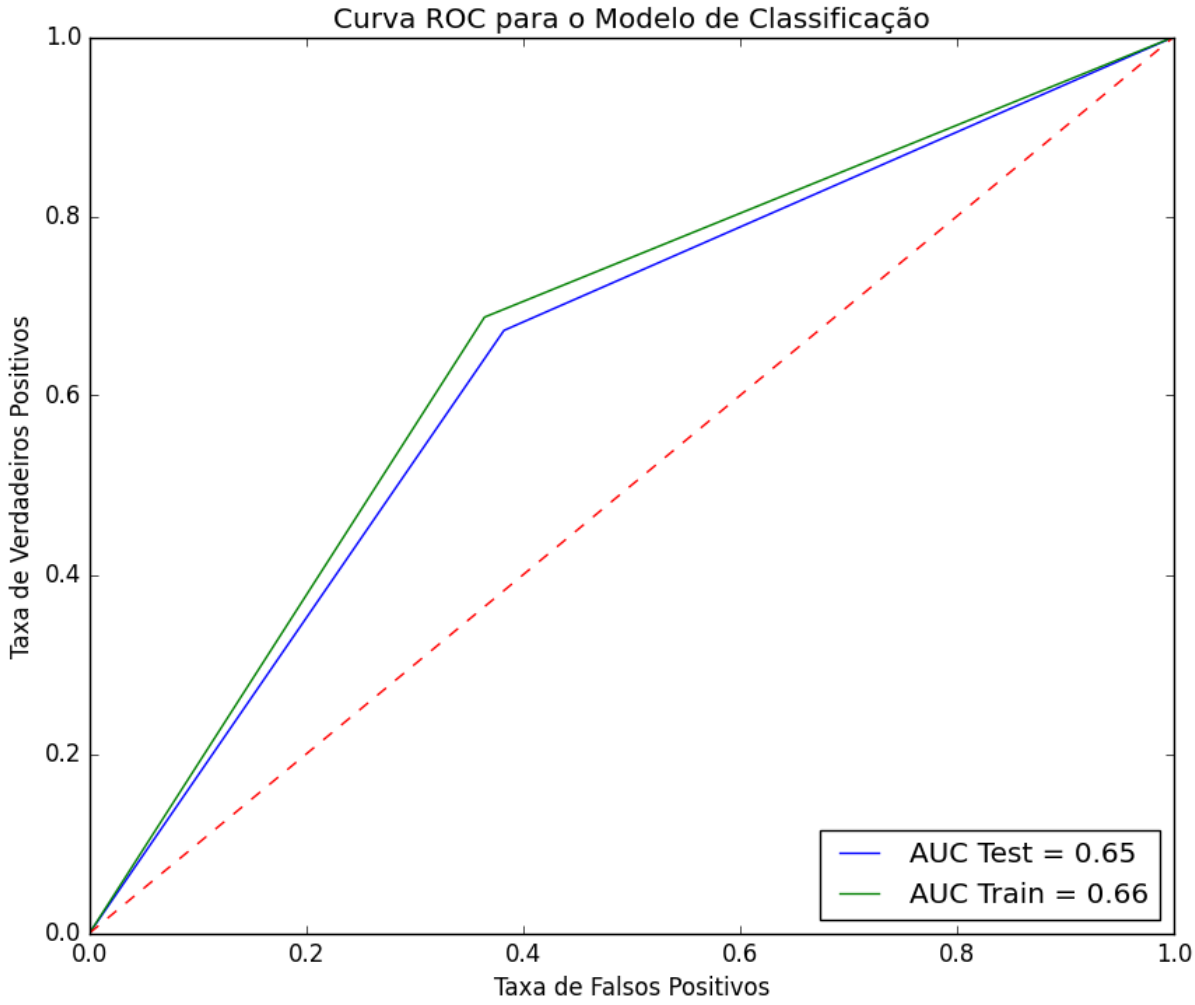


Figura 25 – Curva ROC para o Modelo de Classificação

Para alcançar uma análise de crédito mais robusta e precisa, foi necessário ir além das abordagens simples de similaridade e adotar técnicas avançadas de aprendizado de máquina. O algoritmo selecionado para essa tarefa foi o XGBoost (Extreme Gradient Boosting), uma ferramenta de modelagem baseada em árvores de decisão, utilizada em problemas de classificação e regressão.

O XGBoost destaca-se por sua eficiência e desempenho superior em relação a outros modelos de aprendizado de máquina, principalmente devido às seguintes características:

Regularização: O XGBoost incorpora técnicas de regularização L1 e L2, que ajudam a prevenir o overfitting e tornam o modelo mais generalizável.

Processamento em Paralelo: O algoritmo é otimizado para aproveitar o processamento paralelo, o que acelera consideravelmente o treinamento, mesmo em grandes conjuntos de dados.

Manipulação de Dados Faltantes: O XGBoost possui estratégias integradas para lidar com dados ausentes, tornando o pré-processamento mais eficiente.

Ajuste de hiper parâmetros: O modelo oferece um controle detalhado sobre os hiper parâmetros, permitindo ajustes precisos que melhoram o desempenho preditivo.

A escolha do XGBoost foi motivada pela necessidade de capturar interações complexas entre variáveis financeiras e comportamentais, que métodos mais simples não conseguiam identificar com precisão. Com o uso deste algoritmo, o sistema de recomendação de crédito evoluiu significativamente, proporcionando segmentações mais eficazes e previsões adaptadas às características específicas de cada cliente.

Com o modelo XGBoost devidamente treinado e otimizado, a próxima etapa foi utilizar suas previsões para categorizar os clientes e desenvolver um sistema eficaz de ofertas de taxas de juros.

4.6 DISCUSSÃO DOS RESULTADOS

A avaliação dos resultados dos modelos desenvolvidos comparou a eficácia das duas PoCs levou em consideração as diferenças nas técnicas e nas métricas utilizadas. A PoC 1 aplicou a Distância Euclidiana para medir a similaridade, sem métricas preditivas tradicionais como ROC AUC, precisão ou recall. Em contraste, a PoC 2 foi avaliada com métricas robustas, obtendo um ROC AUC Score de 65.47%, precisão de 0.88 para a classe 0 (pagantes) e 0.34 para a classe 1 (não pagantes), e uma precisão geral de 0.65. Embora a PoC 2 forneça uma análise detalhada, as métricas usadas não são diretamente comparáveis às da PoC 1. Cada abordagem foi analisada considerando suas próprias limitações e contextos específicos.

A seguir, são apresentados gráficos que ilustram as principais diferenças nas variáveis analisadas. Primeiramente, visualizamos a variação da taxa de juros entre as diferentes categorias de crédito, seguida da distribuição da relação dívida/renda (DTI) por categoria. Esses gráficos ajudam a compreender como o risco financeiro se manifesta de maneira diferente entre os grupos, destacando os pontos fortes e limitações das estratégias adotadas em cada Prova de Conceito.

Observamos na figura 26 que a média e a mediana da taxa de juros aumentam gradualmente à medida que as categorias de crédito se movem de A para G. Isso sugere uma relação clara entre a categoria atribuída pelo modelo e a taxa de juros aplicada, com categorias de maior risco (como G) exibindo taxas mais altas.

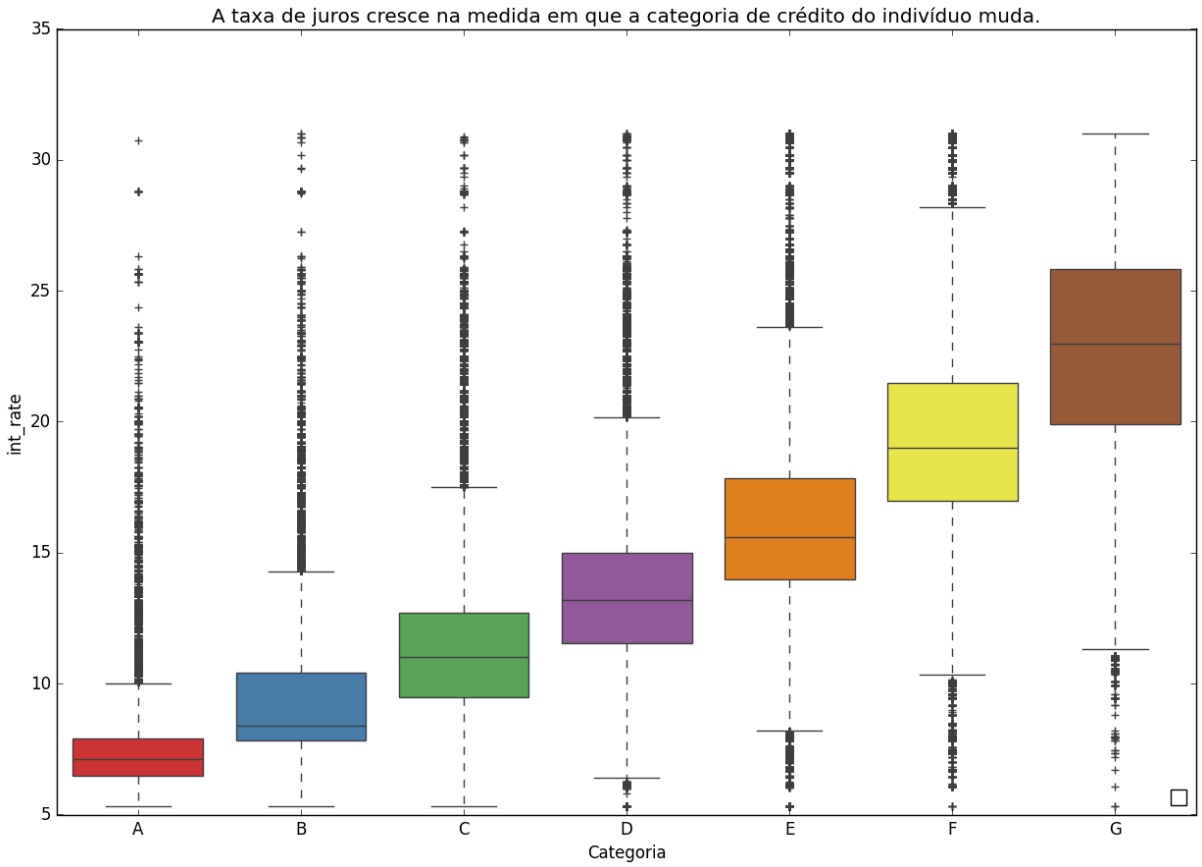


Figura 26 – Box Plot da Taxa de Juros por Categoria de Score

A distribuição da relação dívida/renda (DTI) é visualizada com o boxplot na figura 27. A análise revela que as categorias nas extremidades (A e G) apresentam menor variabilidade. No caso da categoria A, a baixa variabilidade pode estar associada ao fato de que manter um bom score de crédito geralmente exige um menor comprometimento da renda. Por outro lado, a menor variabilidade na categoria G pode estar relacionada à dificuldade de acesso a crédito, limitando o número de vezes que essas pessoas podem comprometer sua renda.

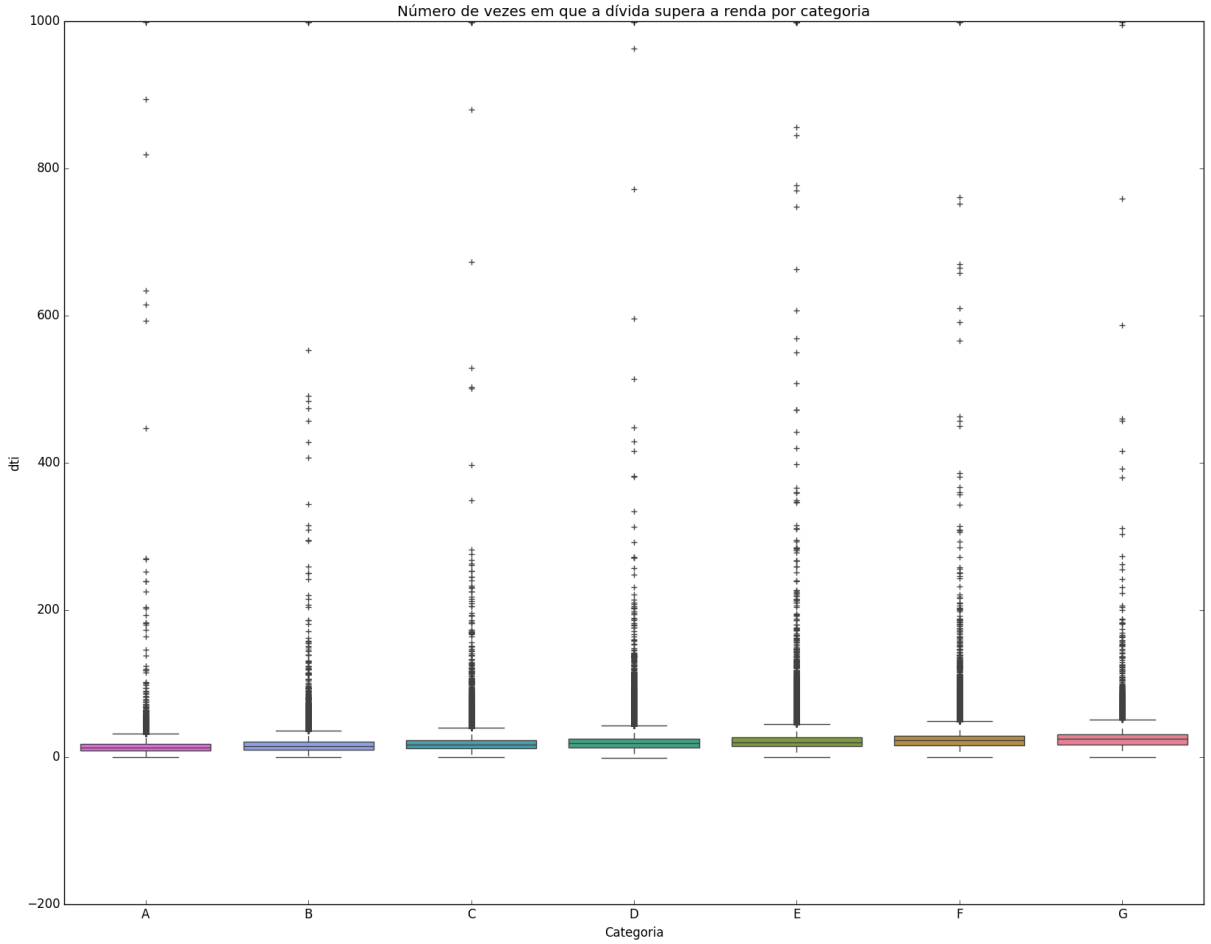


Figura 27 – Box Plot da Relação Dívida/Renda (DTI) por Categoria de Score

A seguir, na figura 28 avaliamos as diferenças entre os grupos em relação à situação de moradia. Observa-se que indivíduos com acesso a melhores taxas de juros frequentemente oferecem suas casas como hipoteca, o que é uma característica típica dos grupos mais bem classificados. Esse comportamento sugere que a posse de uma propriedade hipotecável pode ser um fator determinante para a obtenção de condições mais favoráveis de crédito.

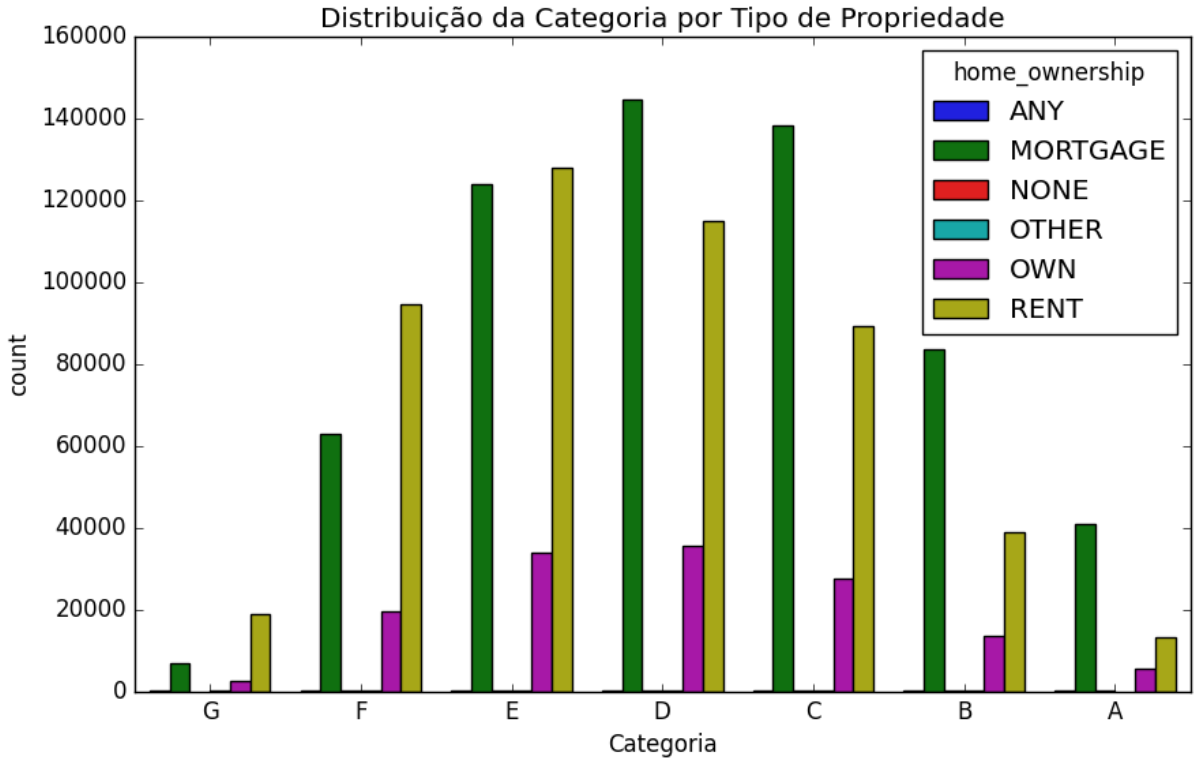


Figura 28 – Distribuição do Tipo de Propriedade por Categoria de Score

Além disso, na figura 29 analisamos os diferentes tipos de aplicação, considerando se são conjuntas ou individuais, entre os grupos definidos pelo modelo. Verificamos que indivíduos em categorias com melhor score de crédito geralmente optam por aplicações individuais, indicando que, nesses casos, a necessidade de uma aplicação conjunta é menos comum, possivelmente devido à maior estabilidade financeira individual.

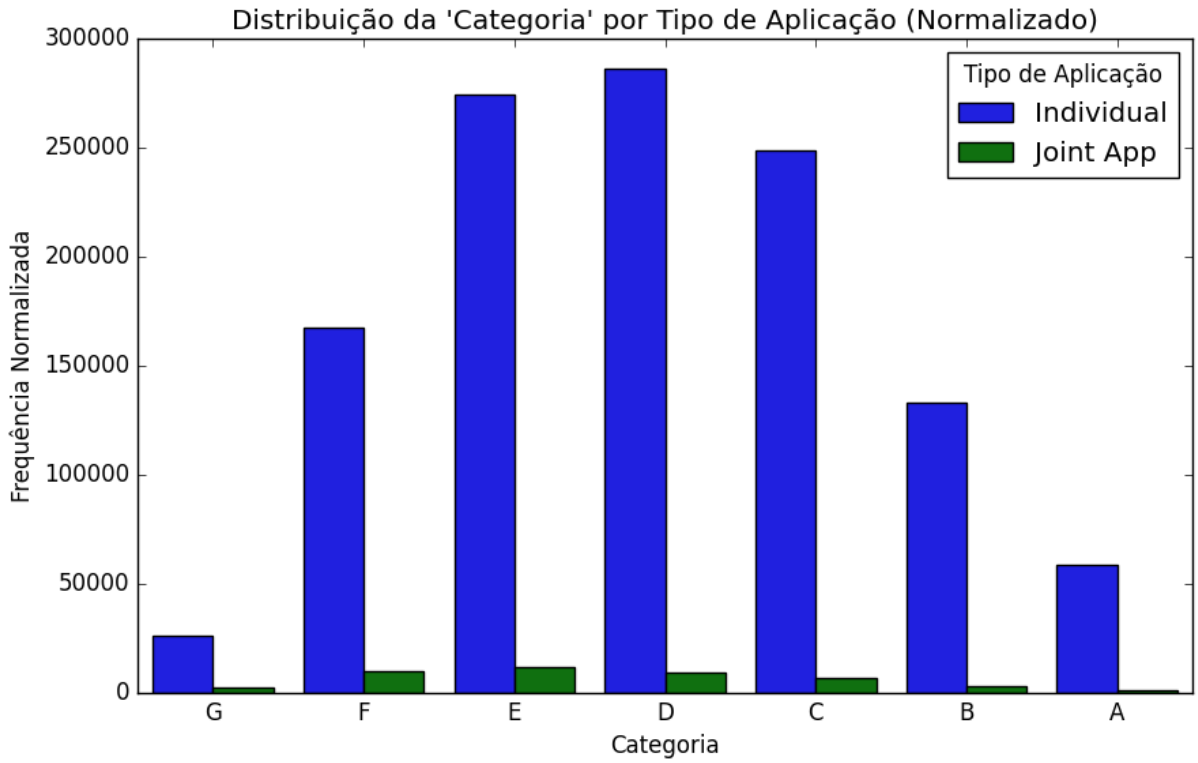


Figura 29 – Gráfico de Barras para Tipo de Aplicação (Individual vs. Conjunta) por Categoria

Por fim, discutem-se as principais diferenças, pontos fortes e fracos, e como cada método atende aos objetivos do sistema de recomendação de crédito.

A PoC 1 utilizou a Distância Euclidiana como método para medir a similaridade entre os atributos dos clientes e os critérios de elegibilidade das campanhas de crédito. Embora essa abordagem fosse simples e de fácil implementação, revelou limitações importantes. A rigidez dos critérios dificultou uma personalização eficaz, o que resultou em uma segmentação menos precisa. Por exemplo, clientes como Elder e Vinicius ilustraram que, mesmo com alta similaridade, pequenas variações nos atributos podiam excluir clientes de campanhas promissoras. Assim, a falta de flexibilidade afetou a capacidade do sistema de atender a uma gama mais ampla de perfis financeiros.

Por outro lado, a PoC 2 representou uma evolução significativa com a introdução do XGBoost, um modelo de aprendizado de máquina que capturou interações mais complexas entre variáveis. Esse método não apenas melhorou a precisão das previsões, mas também permitiu uma segmentação mais eficaz, atendendo às nuances dos perfis de risco de crédito. O uso da Interpolação Linear na PoC 2 adicionou uma camada de personalização, ajustando as taxas de juros de maneira mais justa e adaptada ao risco de cada cliente, o que era impossível com a abordagem rígida da PoC 1.

Tabela 8 – Pontos Fortes e Fracos das Abordagens (PoC 1 e PoC 2)

Abordagem	**Pontos Fortes**	**Pontos Fracos**
PoC 1	- Simplicidade e eficiência computacional. - Viável para segmentação inicial rápida e fácil de interpretar.	- Falta de flexibilidade e personalização. - Exclusão de clientes potencialmente qualificados devido a critérios rígidos.
PoC 2	- Precisão elevada nas estimativas. - Capacidade de modelar relações não lineares. - Flexibilidade para personalização e segmentação mais eficaz, com menor taxa de erro.	- Complexidade computacional mais alta, exigindo mais dados e processamento. - Custos de implementação mais elevados devido à demanda por recursos computacionais.

Tabela 9 – Sugestões para Melhorias e Aplicações Futuras

Sugestões
- Integrar técnicas híbridas que combinem aprendizado supervisionado com regras de negócio dinâmicas.
- Otimizar os modelos para reduzir custos computacionais, mantendo a precisão e personalização.
- Incorporar dados externos ou não estruturados para aprimorar a segmentação e o alinhamento das campanhas.

Sugestões para Melhorias e Aplicações Futuras

Apesar das melhorias na PoC 2, ainda há espaço para avanços. Uma integração futura de técnicas híbridas que combinem aprendizado supervisionado com regras de negócio mais dinâmicas poderia otimizar ainda mais as recomendações.

5 CONCLUSÃO

A concessão de crédito é um desafio relevante no contexto das instituições financeiras, especialmente considerando a diversidade econômica e social de um país como o Brasil. Neste trabalho, foi desenvolvido um sistema de recomendação de campanhas de crédito utilizando técnicas estatísticas, aprendizado de máquina e ciência de dados, com o objetivo de analisar perfis de clientes e sugerir campanhas adequadas com base em características extraídas dos dados. A abordagem foi estruturada para personalizar estratégias de crédito, fundamentando as decisões em padrões obtidos por meio da análise de dados e modelos preditivos.

O dataset utilizado neste trabalho, proveniente da plataforma LendingClub, reflete a realidade do mercado financeiro dos Estados Unidos, com características econômicas, sociais e comportamentais específicas daquele país. Apesar dessa limitação, as técnicas desenvolvidas são generalizáveis e podem ser adaptadas para o contexto brasileiro. No Brasil, a inclusão de variáveis locais, como índices econômicos regionais, características socioeconômicas e dados de políticas públicas, teria grande valor para ajustar os modelos às particularidades do mercado nacional, garantindo maior aplicabilidade e precisão nas recomendações.

A metodologia explorada no estado da arte incluiu recomendações baseadas em conteúdo, colaborativas e híbridas, com destaque para a abordagem híbrida, que combina atributos individuais com padrões de grupos, ampliando a personalização das campanhas. Para compreender os perfis financeiros dos clientes e desenvolver o sistema de recomendação de crédito, foi necessário começar com uma análise das técnicas existentes e das relações entre os atributos financeiros. O primeiro objetivo específico, que foi analisar o estado da arte em sistemas de recomendação e técnicas de análise de dados para identificação de perfis de clientes com base em similaridade, foi atingido por meio de uma análise do estado da arte em sistemas de recomendação e técnicas de análise de dados. Estatísticas descritivas e inferenciais foram aplicadas para analisar a distribuição de atributos financeiros e identificar correlações significativas. O coeficiente de correlação de Pearson ajudou a compreender as relações lineares entre variáveis, evitando problemas de multicolinearidade. A análise exploratória de dados (EDA) incluiu o uso de boxplots e histogramas para detectar outliers e assimetrias, fatores que poderiam impactar o desempenho do modelo.

Na PoC 1, a métrica de Distância Euclidiana foi empregada para medir a similaridade entre perfis de clientes e os critérios das campanhas de crédito. Essa abordagem inicial serviu para validar a viabilidade do sistema de recomendação com base em atributos financeiros, mostrando que era possível segmentar clientes e sugerir campanhas por meio de métricas de similaridade.

A PoC 2 introduziu o uso de algoritmos mais complexos, como o XGBoost, que capturaram interações não lineares e relações complexas entre variáveis financeiras, apri-

morando o desempenho do modelo. O protótipo funcional alcançou um ROC AUC de 65,47%, com uma precisão de 88% para a classe de pagadores (classe 0) e 34% para a classe de inadimplentes (classe 1), resultando em uma precisão geral de 65%. Técnicas de otimização, como o ajuste de hiperparâmetros, e estratégias de balanceamento de classes, como o Random Undersampling, foram aplicadas. As métricas de desempenho, incluindo precisão, recall e AUC, permitiram uma avaliação quantitativa que indicou a eficácia do modelo, embora a necessidade de ajustes contínuos e variáveis mais diversificadas permaneça para aprimorar a robustez do sistema.

O XGBoost foi utilizado neste trabalho por sua capacidade de lidar com relações não lineares, alta dimensionalidade e dados desbalanceados, além de sua eficiência computacional com suporte a paralelismo e regularização avançada, conforme descrito por (NVIDIA, 2024). Apesar de não haver um fator decisivo específico para sua escolha, sua robustez e compatibilidade com os objetivos do modelo justificaram sua aplicação. A regressão logística foi empregada como base interpretativa, sem comparações detalhadas com outros algoritmos.

O sistema de recomendação apresentou algumas limitações que podem ser tratadas em trabalhos futuros. A precisão reduzida na classificação de inadimplentes, com uma taxa de 34%, aponta para a necessidade de melhorar a segmentação de perfis de maior risco. A ocorrência de falsos positivos indica que clientes com um bom histórico de pagamento foram incorretamente classificados como inadimplentes. Além disso, o uso de técnicas de balanceamento de classes, como o Random Undersampling, não foi totalmente eficaz na resolução do problema de classificação desigual. A dependência de variáveis financeiras específicas sugere que a inclusão de atributos mais variados poderia aumentar a robustez do modelo e refletir melhor a complexidade dos perfis de crédito.

A avaliação final sugere que as técnicas estatísticas aplicadas no pré-processamento e na modelagem foram fundamentais para melhorar a precisão das previsões. Pesquisas futuras podem explorar métodos de regularização estatística para mitigar a influência de variáveis altamente correlacionadas e investigar a eficácia de modelos híbridos que combinem técnicas estatísticas com aprendizado profundo, buscando aprimorar a capacidade preditiva e a adaptabilidade do sistema em diferentes contextos econômicos. Além disso, outras técnicas de balanceamento de classes podem ser consideradas para melhorar a distribuição dos dados e os resultados preditivos em cenários com desbalanceamento.

REFERÊNCIAS

ABBAR, Hicham; BOUZEGHOUB, Mokrane; LOPEZ, Salah. Improving Personalized Recommendation using Multi-Criteria Decision Making. **Journal of Data and Information Quality**, 2009.

AGGARWAL, Charu C. **Data Mining: The Textbook**. Cham, Switzerland: Springer, 2015.

ASSAF NETO, Alexandre; LIMA, Fabiano Guasti. **Curso de Administração Financeira**. 2ª. São Paulo: Atlas, 2009. P. 680.

AYODELE. Introduction to Machine Learning. Springer, 2010.

BALABANOVIC, M.; SHOHAM, Y. Content-based filtering: Methods and examples. **Communications of the ACM**, v. 40, n. 3, p. 66–70, 1997.

CHAWLA, Nitesh V.; BOWYER, Kevin W.; HALL, Lawrence O.; KEGELMEYER, W. Philip. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining: Concepts and Techniques**. 3rd. [*S.l.*]: Elsevier, 2011.

HERLOCKER, J. L. Collaborative filtering for digital libraries: A matrix factorization approach. **IEEE Transactions on Knowledge and Data Engineering**, v. 12, n. 3, p. 507–520, 2000.

KUMAR, Arun; RAVI, Vadlamani. Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques – A Review. **European Journal of Operational Research**, v. 180, n. 1, p. 1–28, 2007.

LABATUT, V.; CHERIFI, H. Metrics for clustering analysis. **Data Mining and Knowledge Discovery**, v. 22, p. 313–348, 2011.

LARRY, D. Machine Learning Applications in Geoscience. **Journal of Applied Geoscience**, 2010.

LAWRENCE, John. **Financial Risk Models**. [*S.l.*]: Academic Press, 1984.

MCCALLUM, Andrew; NIGAM, Kamal. **Information Extraction: Algorithms and Applications**. [*S.l.*]: Morgan Kaufmann, 2005.

NVIDIA. **Glossário XGBoost**. [*S.l.*: *s.n.*], 2024. Acessado em: 9 de dezembro de 2024. Disponível em: https://www.nvidia.com/en-us/glossary/xgboost/.

OTTE, Luis Carlos Junior. Machine Learning Models for Credit Analysis. **Journal of Financial Data Science**, 2018.

PARK, Hyun; LEE, Seung; CHO, Minho. Personalized recommendation using machine learning in credit systems. **Journal of Financial Data Science**, 2012.

PINHEIRO, Armando Castelar; MOURA, André Luiz C. Segmentation and the Use of Information in Brazilian Credit Markets. *In*: MILLER, Margaret J. (Ed.). **Credit Reporting Systems and the International Economy**. Cambridge, MA: MIT Press, 2003. P. 287–310. ISBN 9780262134279.

RIEDL, J. T.; AL., et. MovieLens: Collaborative Filtering for Movies. **Proceedings of the 1999 International Conference on Computer Supported Cooperative Work**, p. 221–228, 1999.

SARKER, Ihsanul Hoque. Machine Learning: Algorithms, Real-World Applications and Research Directions. **SN Computer Science**, v. 2, n. 3, 2021.

SELAU, Antonio; RIBEIRO, Paulo. **Credit Risk Management: Techniques and Strategies**. [*S.l.*]: Elsevier, 2009.

SILVA SANTO, João. **Modelos de Aprendizado de Máquina Aplicados à Cobrança**. São Paulo: Editora Financeira, 2013. ISBN 978-85-12345-67-8.

STEYNER, William. **Risk Analysis in Financial Lending**. [*S.l.*]: Finance Press, 1999.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. Introduction to Data Mining. Pearson, 2018.

VIEIRA, José Rômulo de Castro; BARBOZA, Flavio; SOBREIRO, Vinicius Amorim; KIMURA, Herbert. Machine learning models for credit analysis improvements: Predicting low-income families' default. **Applied Soft Computing**, v. 83, p. 105640, 2019.

WITTEN, Ian H.; FRANK, Eibe. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005.

WU, Chong; GAO, Dekun; XU, Siyuan. A Credit Risk Predicting Hybrid Model Based on Deep Learning Technology. **International Journal of Machine Learning and Computing**, v. 11, n. 3, p. 182–187, 2021.

ANEXO A – Introdução ao Conjunto de Dados

DESCRIÇÃO DAS COLUNAS

1. **Unnamed: 0**: Nenhuma descrição disponível.
2. **id**: Um ID exclusivo atribuído pela LC para a listagem de empréstimos.
3. **loan_amnt**: O valor listado do empréstimo solicitado pelo mutuário. Se em algum momento, o departamento de crédito reduzir o valor do empréstimo, isso será refletido neste valor.
4. **funded_amnt**: O valor total comprometido com esse empréstimo naquele momento.
5. **funded_amnt_inv**: O valor total comprometido pelos investidores para esse empréstimo naquele momento.
6. **term**: O número de pagamentos do empréstimo. Os valores estão em meses e podem ser 36 ou 60.
7. **int_rate**: Taxa de juros do empréstimo.
8. **installment**: O pagamento mensal devido pelo mutuário se o empréstimo for originado.
9. **grade**: Grau de empréstimo atribuído pela LC.
10. **sub_grade**: Subgrau de empréstimo atribuído pela LC.
11. **emp_title**: O cargo fornecido pelo mutuário ao solicitar o empréstimo.
12. **emp_length**: Tempo de emprego em anos. Os valores possíveis estão entre 0 e 10, onde 0 significa menos de um ano e 10 significa dez ou mais anos.
13. **home_ownership**: O status de propriedade da casa fornecido pelo mutuário durante o registro ou obtido no relatório de crédito. Nossos valores são: ALUGUEL, PRÓPRIO, HIPOTECA, OUTROS.
14. **annual_inc**: A renda anual autodeclarada fornecida pelo mutuário durante o registro.
15. **verification_status**: Indica se a renda foi verificada pela LC, não verificada ou se a fonte de renda foi verificada.
16. **issue_d**: O mês em que o empréstimo foi financiado.
17. **loan_status**: Status atual do empréstimo.
18. **pymnt_plan**: Indica se um plano de pagamento foi estabelecido para o empréstimo.
19. **url**: URL para a página da LC com dados de listagem.

20. **purpose**: Uma categoria fornecida pelo mutuário para a solicitação de empréstimo.
21. **title**: O título do empréstimo fornecido pelo mutuário.
22. **zip_code**: Os primeiros 3 números do código postal fornecidos pelo mutuário no pedido de empréstimo.
23. **addr_state**: O estado fornecido pelo mutuário no pedido de empréstimo.
24. **dti**: Uma proporção calculada usando os pagamentos mensais totais da dívida do mutuário sobre as obrigações totais da dívida, excluindo hipoteca e o empréstimo LC solicitado, dividido pela renda mensal autodeclarada do mutuário.
25. **delinq_2yrs**: O número de incidências de inadimplência com mais de 30 dias de atraso no arquivo de crédito do mutuário nos últimos 2 anos.
26. **earliest_cr_line**: O mês em que a primeira linha de crédito relatada pelo mutuário foi aberta.
27. **fico_range_low**: O limite inferior ao qual o FICO do mutuário na origem do empréstimo pertence.
28. **fico_range_high**: O limite superior ao qual o FICO do mutuário na origem do empréstimo pertence.
29. **inq_last_6mths**: O número de consultas nos últimos 6 meses (excluindo consultas de automóveis e hipotecas).
30. **mths_since_last_delinq**: O número de meses desde a última inadimplência do mutuário.
31. **mths_since_last_record**: O número de meses desde o último registro público.
32. **open_acc**: O número de linhas de crédito abertas no arquivo de crédito do mutuário.
33. **pub_rec**: O número de registros públicos depreciativos.
34. **revol_bal**: Saldo rotativo de crédito total.
35. **revol_util**: Taxa de utilização da linha rotativa ou a quantidade de crédito que o mutuário está usando em relação a todo o crédito rotativo disponível.
36. **total_acc**: O número total de linhas de crédito atualmente no arquivo de crédito do mutuário.
37. **initial_list_status**: O status inicial de listagem do empréstimo. Os valores possíveis são: W, F.
38. **out_prncp**: Principal restante em aberto para o valor total financiado.

39. **out_prncp_inv**: Principal restante em aberto para a parcela do valor total financiado pelos investidores.
40. **total_pymnt**: Pagamentos recebidos até a data para o valor total financiado.
41. **total_pymnt_inv**: Pagamentos recebidos até a data para a parcela do valor total financiado pelos investidores.
42. **total_rec_prncp**: Principal recebido até a data.
43. **total_rec_int**: Juros recebidos até a data.
44. **total_rec_late_fee**: Taxas de atraso recebidas até a data.
45. **recoveries**: Pós-baixa recuperação bruta.
46. **collection_recovery_fee**: Pós-baixa taxa de cobrança.
47. **last_pymnt_d**: Último pagamento do mês foi recebido.
48. **last_pymnt_amnt**: Último valor total de pagamento recebido.
49. **next_pymnt_d**: Próxima data de pagamento agendada.
50. **last_credit_pull_d**: O mês mais recente em que a LC retirou crédito para este empréstimo.
51. **last_fico_range_high**: A faixa limite superior à qual o último FICO retirado do mutuário pertence.
52. **last_fico_range_low**: A faixa limite inferior à qual o último FICO retirado do mutuário pertence.
53. **collections_12_mths_ex_med**: Número de cobranças em 12 meses, excluindo cobranças médicas.
54. **mths_since_last_major_derog**: Meses desde a classificação mais recente de 90 dias ou pior.
55. **policy_code**: Política disponível publicamente_code=1, novos produtos não disponíveis publicamente_code=2.
56. **application_type**: Indica se o empréstimo é uma aplicação individual ou uma aplicação conjunta com dois co-mutuários.
57. **annual_inc_joint**: A renda anual autodeclarada combinada fornecida pelos co-mutuários durante o registro.
58. **dti_joint**: Uma proporção calculada usando os pagamentos mensais totais dos co-mutuários sobre as obrigações totais da dívida, excluindo hipotecas e o empréstimo LC solicitado, dividido pela renda mensal autodeclarada combinada dos co-mutuários.
59. **verification_status_joint**: Nenhuma descrição disponível.

60. **acc_now_delinq**: O número de contas nas quais o mutuário está inadimplente.
61. **tot_coll_amt**: Valores totais de cobrança já devidos.
62. **tot_cur_bal**: Saldo atual total de todas as contas.
63. **open_acc_6m**: Número de negociações abertas nos últimos 6 meses.
64. **open_act_il**: Número de negociações parceladas ativas no momento.
65. **open_il_12m**: Número de contas parceladas abertas nos últimos 12 meses.
66. **open_il_24m**: Número de contas parceladas abertas nos últimos 24 meses.
67. **mths_since_rcnt_il**: Meses desde a abertura das contas parceladas mais recentes.
68. **total_bal_il**: Saldo atual total de todas as contas parceladas.
69. **il_util**: Proporção do saldo atual total para crédito alto/limite de crédito em todas as contas de instalação.
70. **open_rv_12m**: Número de negociações rotativas abertas nos últimos 12 meses.
71. **open_rv_24m**: Número de negociações rotativas abertas nos últimos 24 meses.
72. **max_bal_bc**: Saldo atual máximo devido em todas as contas rotativas.
73. **all_util**: Saldo para limite de crédito em todas as negociações.
74. **total_rev_hi_lim**: Nenhuma descrição disponível.
75. **inq_fi**: Número de consultas de finanças pessoais.
76. **total_cu_tl**: Número de negociações financeiras.
77. **inq_last_12m**: Número de consultas de crédito nos últimos 12 meses.
78. **acc_open_past_24mths**: Número de negociações abertas nos últimos 24 meses.
79. **avg_cur_bal**: Saldo atual médio de todas as contas.
80. **bc_open_to_buy**: Total aberto para comprar em cartões bancários rotativos.
81. **bc_util**: Proporção do saldo atual total para crédito alto/limite de crédito para todas as contas de cartão bancário.
82. **chargeoff_within_12_mths**: Número de baixas em 12 meses.
83. **delinq_amnt**: O valor devido em atraso para as contas nas quais o mutuário está inadimplente.
84. **mo_sin_old_il_acct**: Meses desde a abertura da conta de parcelamento bancário mais antiga.

85. **mo_sin_old_rev_tl_op**: Meses desde a abertura da conta rotativa mais antiga.
86. **mo_sin_rcnt_rev_tl_op**: Meses desde a abertura da conta rotativa mais recente.
87. **mo_sin_rcnt_tl**: Meses desde a abertura da conta rotativa mais recente.
88. **mort_acc**: Número de contas de hipoteca.
89. **mths_since_recent_bc**: Meses desde a abertura da conta de cartão bancário mais recente.
90. **mths_since_recent_bc_dlq**: Meses desde a inadimplência de cartão bancário mais recente.
91. **mths_since_recent_inq**: Meses desde a consulta mais recente.
92. **mths_since_recent_revol_delinq**: Meses desde a inadimplência rotativa mais recente.
93. **num_accts_ever_120_pd**: Número de contas com 120 ou mais dias de atraso.
94. **num_actv_bc_tl**: Número de contas de cartão bancário ativas no momento.
95. **num_actv_rev_tl**: Número de negociações rotativas ativas no momento.
96. **num_bc_sats**: Número de contas de cartão bancário satisfatórias.
97. **num_bc_tl**: Número de contas de cartão bancário.
98. **num_il_tl**: Número de contas parceladas.
99. **num_op_rev_tl**: Número de contas rotativas abertas.
100. **num_rev_accts**: Número de contas rotativas.
101. **num_rev_tl_bal_gt_0**: Número de negociações rotativas com saldo >0.
102. **num_sats**: Número de contas satisfatórias.
103. **num_tl_120dpd_2m**: Número de contas atualmente vencidas há 120 dias (atualizado nos últimos 2 meses).
104. **num_tl_30dpd**: Número de contas atualmente vencidas há 30 dias (atualizado nos últimos 2 meses).
105. **num_tl_90g_dpd_24m**: Número de contas vencidas há 90 ou mais dias nos últimos 24 meses.
106. **num_tl_op_past_12m**: Número de contas abertas nos últimos 12 meses.
107. **pct_tl_nvr_dlq**: Porcentagem de negociações nunca inadimplentes.
108. **percent_bc_gt_75**: Porcentagem de todas as contas de cartão bancário > 75% do limite.

109. **pub_rec_bankruptcies**: Número de falências de registros públicos.
110. **tax_liens**: Número de penhoras fiscais.
111. **tot_hi_cred_lim**: Total de crédito alto/limite de crédito.
112. **total_bal_ex_mort**: Saldo de crédito total excluindo hipoteca.
113. **total_bc_limit**: Total de crédito alto/limite de crédito do cartão bancário.
114. **total_il_high_credit_limit**: Total de crédito alto/limite de crédito da parcela.
115. **revol_bal_joint**: Nenhuma descrição disponível.
116. **sec_app_fico_range_low**: Nenhuma descrição disponível.
117. **sec_app_fico_range_high**: Nenhuma descrição disponível.
118. **sec_app_earliest_cr_line**: Nenhuma descrição disponível.
119. **sec_app_inq_last_6mths**: Nenhuma descrição disponível.
120. **sec_app_mort_acc**: Nenhuma descrição disponível.
121. **sec_app_open_acc**: Nenhuma descrição disponível.
122. **sec_app_revol_util**: Nenhuma descrição disponível.
123. **sec_app_open_act_il**: Número de transações parceladas atualmente ativas no momento da solicitação do requerente secundário.
124. **sec_app_num_rev_accts**: Nenhuma descrição disponível.
125. **sec_app_chargeoff_within_12_mths**: Nenhuma descrição disponível.
126. **sec_app_collections_12_mths_ex_med**: Nenhuma descrição disponível.
127. **hardship_flag**: Sinaliza se o mutuário está ou não em um plano de dificuldades.
128. **hardship_type**: Descreve a oferta do plano de dificuldades.
129. **hardship_reason**: Descreve o motivo pelo qual o plano de dificuldades foi oferecido.
130. **hardship_status**: Descreve se o plano de dificuldades está ativo, pendente, cancelado, concluído ou quebrado.
131. **deferral_term**: Quantidade de meses que o mutuário deve pagar menos do que o valor do pagamento mensal contratual devido a um plano de dificuldades.
132. **hardship_amount**: O pagamento de juros que o mutuário se comprometeu a fazer a cada mês enquanto estiver em um plano de dificuldades.
133. **hardship_start_date**: A data de início do período do plano de dificuldades.
134. **hardship_end_date**: A data de término do período do plano de dificuldades.

135. **payment_plan_start_date**: O dia em que o primeiro pagamento do plano de dificuldades vence. Por exemplo, se um mutuário tiver um período de plano de dificuldades de 3 meses, a data de início será o início do período de três meses em que o mutuário tem permissão para fazer pagamentos somente de juros.
136. **hardship_length**: O número de meses em que o mutuário fará pagamentos menores do que normalmente obrigado devido a um plano de dificuldades.
137. **hardship_dpd**: Dias da conta em atraso na data de início do plano de dificuldades.
138. **hardship_loan_status**: Status do empréstimo na data de início do plano de dificuldades.
139. **orig_projected_additional_accrued_interest**: O valor original de juros adicionais projetados que serão acumulados para o plano de pagamento de dificuldades fornecido na Data de Início das Dificuldades. Este campo será nulo se o mutuário tiver quebrado seu plano de pagamento de dificuldades.
140. **hardship_payoff_balance_amount**: O valor do saldo de quitação na data de início do plano de dificuldades.
141. **hardship_last_payment_amount**: O valor do último pagamento na data de início do plano de dificuldades.
142. **debt_settlement_flag**: Sinaliza se o mutuário, que fez o charge-off, está ou não trabalhando com uma empresa de liquidação de dívidas.
143. **year**: Nenhuma descrição disponível.

COLUNAS SELECIONADAS

COLUNAS SELECIONADAS PARA POC 1:

1. **Renda Anual (annual_inc)**: A renda anual autodeclarada fornecida pelo mutuário durante o registro.
2. **Utilização de Crédito (all_util)**: Saldo para limite de crédito em todas as negociações.
3. **Contas Abertas Recentemente (acc_open_past_24mths)**: Número de negociações abertas nos últimos 24 meses.
4. **Nota de Crédito (grade)**: Grau de empréstimo atribuído pela LC.

COLUNAS SELECIONADAS PARA POC 2:

1. **Ano (year)**: Nenhuma descrição disponível.

2. **Saldo Atual Médio (avg_cur_bal)**: Saldo atual médio de todas as contas.
3. **Default (default)**: (Certifique-se de especificar como "default"foi derivado ou relacionado no seu contexto.)
4. **FICO (fico)**: (Especifique como "fico"está relacionado a "fico_range_low"e "fico_range_high".)
5. **Montante do Empréstimo (loan_amnt)**: O valor listado do empréstimo solicitado pelo mutuário. Se em algum momento, o departamento de crédito reduzir o valor do empréstimo, isso será refletido neste valor.
6. **Prazo (term)**: O número de pagamentos do empréstimo. Os valores estão em meses e podem ser 36 ou 60.
7. **Taxa de Juros (int_rate)**: Taxa de juros do empréstimo.
8. **Parcelamento (installment)**: O pagamento mensal devido pelo mutuário se o empréstimo for originado.
9. **Subgrau (sub_grade)**: Subgrau de empréstimo atribuído pela LC.
10. **Tempo de Emprego (emp_length)**: Tempo de emprego em anos. Os valores possíveis estão entre 0 e 10, onde 0 significa menos de um ano e 10 significa dez ou mais anos.
11. **Propriedade da Casa (home_ownership)**: O status de propriedade da casa fornecido pelo mutuário durante o registro ou obtido no relatório de crédito. Os valores possíveis são: ALUGUEL, PRÓPRIO, HIPOTECA, OUTROS.
12. **Renda Anual (annual_inc)**: A renda anual autodeclarada fornecida pelo mutuário durante o registro.
13. **Propósito (purpose)**: Uma categoria fornecida pelo mutuário para a solicitação de empréstimo.
14. **Estado (addr_state)**: O estado fornecido pelo mutuário no pedido de empréstimo.
15. **Índice de Dívida/Renda (dti)**: Uma proporção calculada usando os pagamentos mensais totais da dívida do mutuário sobre as obrigações totais da dívida, excluindo hipoteca e o empréstimo LC solicitado, dividido pela renda mensal autodeclarada do mutuário.
16. **Primeira Linha de Crédito (earliest_cr_line)**: O mês em que a primeira linha de crédito relatada pelo mutuário foi aberta.
17. **Consultas nos Últimos 6 Meses (inq_last_6mths)**: O número de consultas nos últimos 6 meses (excluindo consultas de automóveis e hipotecas).
18. **Contas Abertas (open_acc)**: O número de linhas de crédito abertas no arquivo de crédito do mutuário.

19. **Registros Públicos (pub_rec)**: O número de registros públicos depreciativos.
20. **Status Inicial da Listagem (initial_list_status)**: O status inicial de listagem do empréstimo. Os valores possíveis são – W, F.
21. **Meses Desde o Último Grande Problema (mths_since_last_major_derog)**: Meses desde a classificação mais recente de 90 dias ou pior.
22. **Tipo de Aplicação (application_type)**: Indica se o empréstimo é uma aplicação individual ou uma aplicação conjunta com dois co-mutuários.
23. **Contas Inadimplentes (acc_now_delinq)**: O número de contas nas quais o mutuário está inadimplente.
24. **Saldo Atual Total (tot_cur_bal)**: Saldo atual total de todas as contas.
25. **Contas Abertas nos Últimos 6 Meses (open_acc_6m)**: Número de negociações abertas nos últimos 6 meses.
26. **Contas Parceladas Ativas (open_act_il)**: Número de negociações parceladas ativas no momento.
27. **Contas Parceladas Abertas nos Últimos 12 Meses (open_il_12m)**: Número de contas parceladas abertas nos últimos 12 meses.
28. **Meses Desde a Conta Parcelada Mais Recente (mths_since_rcnt_il)**: Meses desde a abertura das contas parceladas mais recentes.
29. **Saldo Total de Contas Parceladas (total_bal_il)**: Saldo atual total de todas as contas parceladas.
30. **Contas Rotativas Abertas nos Últimos 12 Meses (open_rv_12m)**: Número de negociações rotativas abertas nos últimos 12 meses.
31. **Saldo Máximo em Contas Rotativas (max_bal_bc)**: Saldo atual máximo devido em todas as contas rotativas.
32. **Total de Negociações Financeiras (total_cu_tl)**: Número de negociações financeiras.
33. **Meses Desde a Conta de Parcelamento Mais Antiga (mo_sin_old_il_acct)**: Meses desde a abertura da conta de parcelamento bancário mais antiga.
34. **Meses Desde a Conta Rotativa Mais Antiga (mo_sin_old_rev_tl_op)**: Meses desde a abertura da conta rotativa mais antiga.
35. **Meses Desde a Conta Rotativa Mais Recente (mo_sin_rcnt_rev_tl_op)**: Meses desde a abertura da conta rotativa mais recente.
36. **Meses Desde a Conta Mais Recente (mo_sin_rcnt_tl)**: Meses desde a abertura da conta mais recente.
37. **Contas de Hipoteca (mort_acc)**: Número de contas de hipoteca.

38. **Meses Desde o Cartão Bancário Mais Recente (mths_since_recent_bc)**: Meses desde a abertura da conta de cartão bancário mais recente.
39. **Meses Desde a Inadimplência Rotativa Mais Recente (mths_since_recent_revol**
Meses desde a inadimplência rotativa mais recente.
40. **Negociações Rotativas Ativas (num_actv_rev_tl)**: Número de negociações rotativas ativas no momento.
41. **Contas Parceladas (num_il_tl)**: Número de contas parceladas.
42. **Porcentagem de Negociações Nunca Inadimplentes (pct_tl_nvr_dlq)**: Porcentagem de negociações nunca inadimplentes.
43. **Falências de Registros Públicos (pub_rec_bankruptcies)**: Número de falências de registros públicos.

APÊNDICE A – Artigo no formato SBC

Recomendação de Campanha de Crédito para Perfis de Clientes: Uma Abordagem Baseada em Aprendizado de Máquina e Ciência de Dados

Vinicius Pizetta de Souza
Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Florianópolis / SC, Brasil
vinicius.souza@grad.ufsc.br

Abstract. *Credit granting is a complex challenge for financial institutions, requiring structured strategies to minimize risks and allocate resources effectively. This paper presents a machine learning-based recommendation model for credit campaigns. Using real data from LendingClub, the construction of two Proofs of Concept (PoCs) was analyzed: the first focused on validating relevant attributes; the second implemented algorithms such as logistic regression and XGBoost. The results indicate that the models can segment customers based on data patterns and suggest directions for integrating additional features to improve future analyses.*

Resumo. *A concessão de crédito é um desafio complexo para instituições financeiras, exigindo estratégias estruturadas para minimizar riscos e alocar recursos de forma eficaz. Este artigo apresenta um modelo de recomendação baseado em aprendizado de máquina para campanhas de crédito. Utilizando dados reais do LendingClub, analisou-se a construção de duas Provas de Conceito (PoCs): a primeira concentrou-se na validação de atributos relevantes; a segunda implementou algoritmos como regressão logística e XGBoost. Os resultados indicam que os modelos conseguem segmentar clientes com base em padrões de dados e sugerem direções para integrar características adicionais para melhorar análises futuras.*

A.1 INTRODUÇÃO

A concessão de crédito está diretamente ligada à análise de riscos e à capacidade de prever comportamentos financeiros. Identificar perfis de clientes e conectá-los a campanhas de crédito representa um desafio técnico que pode ser abordado com técnicas de aprendizado de máquina. Este trabalho utiliza dados históricos para desenvolver modelos que ajudem instituições financeiras a alinhar suas campanhas a perfis específicos, reduzindo a incerteza e melhorando a alocação de recursos.

Os objetivos deste estudo incluem analisar o estado da arte em sistemas de recomendação e técnicas de análise de dados para identificar perfis de clientes; desenvolver uma abordagem para qualificar clientes para campanhas específicas de crédito; validar a proposta com base em métricas quantitativas; e comparar a abordagem com alternativas existentes, apontando limitações e possibilidades de aprimoramento.

A.2 METODOLOGIA

A análise exploratória dos dados envolveu a identificação de valores discrepantes, a visualização de distribuições e a análise de correlações entre variáveis financeiras. Ferramentas como boxplots e mapas de calor foram utilizadas para compreender padrões relevantes no conjunto de dados.

No preprocessamento, as taxas de juros foram convertidas de formato textual para valores numéricos, valores ausentes foram tratados com imputação de dados críticos e variáveis categóricas foram codificadas para compatibilidade com os modelos preditivos. Para a modelagem, duas provas de conceito foram construídas. A primeira utilizou regressão logística para validar atributos relevantes, enquanto a segunda implementou o algoritmo XGBoost para segmentação detalhada de clientes.

A avaliação dos modelos baseou-se em métricas como a AUC (Área sob a Curva ROC), que mede a capacidade de classificação, e o RMSE (Root Mean Squared Error), que avalia os erros de predição.

A.3 RESULTADOS E DISCUSSÃO

Os resultados obtidos indicam que o modelo XGBoost apresentou uma AUC de 0.91, evidenciando precisão na classificação de clientes. Grupos distintos de clientes foram identificados, o que sugere possibilidades para campanhas específicas. Contudo, limitações na abrangência do dataset restringiram a generalização do modelo para outros mercados, destacando a importância de uma maior diversidade nos dados analisados.

A.4 DISCUSSÃO

A validação dos modelos mostrou que técnicas como XGBoost podem ser aplicadas em cenários financeiros, especialmente na segmentação de clientes. Apesar disso, a ausência de dados não estruturados, como interações online ou histórico de crédito detalhado, foi uma limitação importante. Além disso, a representatividade demográfica do dataset pode ser explorada em estudos futuros, visando a generalização dos resultados para diferentes mercados.

A.5 CONCLUSÃO

Este estudo investigou a aplicação de aprendizado de máquina na recomendação de campanhas de crédito. Modelos como regressão logística e XGBoost foram utilizados para identificar padrões e segmentar clientes. A análise mostrou que tais modelos podem apoiar decisões em instituições financeiras, especialmente ao conectar dados históricos com campanhas direcionadas.

Entretanto, foram identificadas limitações, incluindo a falta de dados não estruturados e a necessidade de maior diversidade no dataset. Trabalhos futuros podem explorar a integração de novas fontes de dados e o uso de abordagens híbridas, combinando aprendizado supervisionado e não supervisionado. Este trabalho contribui para o desenvolvimento de técnicas aplicadas à análise de crédito, com potencial para aprimorar práticas no setor.

REFERÊNCIAS

ABBAR, Hicham; BOUZEGHOUB, Mokrane; LOPEZ, Salah. Improving Personalized Recommendation using Multi-Criteria Decision Making. **Journal of Data and Information Quality**, 2009.

AGGARWAL, Charu C. **Data Mining: The Textbook**. Cham, Switzerland: Springer, 2015.

ASSAF NETO, Alexandre; LIMA, Fabiano Guasti. **Curso de Administração Financeira**. 2ª. São Paulo: Atlas, 2009. P. 680.

AYODELE. Introduction to Machine Learning. Springer, 2010.

BALABANOVIC, M.; SHOHAM, Y. Content-based filtering: Methods and examples. **Communications of the ACM**, v. 40, n. 3, p. 66–70, 1997.

CHAWLA, Nitesh V.; BOWYER, Kevin W.; HALL, Lawrence O.; KEGELMEYER, W. Philip. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining: Concepts and Techniques**. 3rd. [*S.l.*]: Elsevier, 2011.

HERLOCKER, J. L. Collaborative filtering for digital libraries: A matrix factorization approach. **IEEE Transactions on Knowledge and Data Engineering**, v. 12, n. 3, p. 507–520, 2000.

KUMAR, Arun; RAVI, Vadlamani. Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques – A Review. **European Journal of Operational Research**, v. 180, n. 1, p. 1–28, 2007.

LABATUT, V.; CHERIFI, H. Metrics for clustering analysis. **Data Mining and Knowledge Discovery**, v. 22, p. 313–348, 2011.

LARRY, D. Machine Learning Applications in Geoscience. **Journal of Applied Geoscience**, 2010.

LAWRENCE, John. **Financial Risk Models**. [*S.l.*]: Academic Press, 1984.

MCCALLUM, Andrew; NIGAM, Kamal. **Information Extraction: Algorithms and Applications**. [*S.l.*]: Morgan Kaufmann, 2005.

NVIDIA. **Glossário XGBoost**. [*S.l.*: *s.n.*], 2024. Acessado em: 9 de dezembro de 2024. Disponível em: https://www.nvidia.com/en-us/glossary/xgboost/.

OTTE, Luis Carlos Junior. Machine Learning Models for Credit Analysis. **Journal of Financial Data Science**, 2018.

PARK, Hyun; LEE, Seung; CHO, Minho. Personalized recommendation using machine learning in credit systems. **Journal of Financial Data Science**, 2012.

PINHEIRO, Armando Castelar; MOURA, André Luiz C. Segmentation and the Use of Information in Brazilian Credit Markets. *In*: MILLER, Margaret J. (Ed.). **Credit Reporting Systems and the International Economy**. Cambridge, MA: MIT Press, 2003. P. 287–310. ISBN 9780262134279.

RIEDL, J. T.; AL., et. MovieLens: Collaborative Filtering for Movies. **Proceedings of the 1999 International Conference on Computer Supported Cooperative Work**, p. 221–228, 1999.

SARKER, Ihsanul Hoque. Machine Learning: Algorithms, Real-World Applications and Research Directions. **SN Computer Science**, v. 2, n. 3, 2021.

SELAU, Antonio; RIBEIRO, Paulo. **Credit Risk Management: Techniques and Strategies**. [*S.l.*]: Elsevier, 2009.

SILVA SANTO, João. **Modelos de Aprendizado de Máquina Aplicados à Cobrança**. São Paulo: Editora Financeira, 2013. ISBN 978-85-12345-67-8.

STEYNER, William. **Risk Analysis in Financial Lending**. [*S.l.*]: Finance Press, 1999.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. Introduction to Data Mining. Pearson, 2018.

VIEIRA, José Rômulo de Castro; BARBOZA, Flavio; SOBREIRO, Vinicius Amorim; KIMURA, Herbert. Machine learning models for credit analysis improvements: Predicting low-income families' default. **Applied Soft Computing**, v. 83, p. 105640, 2019.

WITTEN, Ian H.; FRANK, Eibe. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005.

WU, Chong; GAO, Dekun; XU, Siyuan. A Credit Risk Predicting Hybrid Model Based on Deep Learning Technology. **International Journal of Machine Learning and Computing**, v. 11, n. 3, p. 182–187, 2021.