Credit Score: Proposal of a Multi-Source Intelligent System for Predictive Credit Analysis

Bernardo Dirceu Tomasi[1], **João Mário Lopes Brezolin** [1]

[1]Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense – (IFSul)
CEP 99064-440, nº 150 – Passo Fundo – RS – Brasil

bernardodtomasi@gmail.com, joaobrezolin@ifsul.edu.br

Abstract. *The Brazilian economic scenario is characterized by high default rates and elevated interest rates. In this context, credit card debt stands out. This study proposes classification models and a scoring system to identify customers with a propensity for default based on behavioral analysis. The process involved the development of a credit card scoring system using three distinct datasets. The models achieved AUC-ROC of 88%, 84%, and 70%, respectively. The data obtained from the classification system were used to feed the scoring system, which performs a weighting between the probabilities of default from the different datasets. The results were deemed adequate, as the prediction models demonstrated the ability to distinguish between classes, and the dynamic and multifaceted scoring system proved effective in predicting customer behavior.*

Resumo. *O cenário econômico brasileiro é caracterizado pela alta inadimplência e taxas de juros elevadas. Nesse contexto destaca-se o endividamento por cartões de crédito. Este estudo propõe modelos de classificação e um sistema de pontuação para identificar clientes com propensão à inadimplência a partir de uma análise comportamental. O processo envolveu o desenvolvimento de um sistema de pontuação para cartões de crédito, usando três conjuntos de dados distintos. Os modelos obtiveram AUC-ROC de 88%, 84% e 70% respectivamente. Os dados obtidos pelo sistema de classificação foram utilizados para alimentar o sistema de pontuação que realiza a ponderação entre as probabilidades de inadimplência dos diferentes conjuntos de dados. Os resultados obtidos foram considerados adequados, uma vez que os modelos de previsão demonstraram capacidade de distinguir entre as classes, e o sistema de pontuação dinâmico e multifonte mostrou-se apto para realizar a previsão do comportamento dos clientes.*

1. Introdução

O cenário financeiro e econômico enfrenta desafios significativos, dentre os quais se ressalta a persistente alta taxa de juros [Feldmann 2023]. Essas taxas afetam a sociedade, influenciando empréstimos e investimentos. A alta taxa de juros desacelera a economia, reduzindo oferta e demanda de mercado devido ao aumento no custo de aquisição de bens e serviços [Zanatta and Nassif 2023]. Campos Neto ressalta a conexão entre as taxas de juros e a inadimplência de crédito, indicando que aumentos nas taxas estão ligados a esse fenômeno [Neto 2023]. Isso sublinha a necessidade de lidar eficazmente com a inadimplência, tendo o potencial de gerar impactos positivos na economia. Pesquisas

apontam para um crescimento expressivo da inadimplência no país. Conforme o SERASA (Serviços de Assessoria S.A) houve um aumento de 1,05 milhões de inadimplentes entre junho de 2023 e 2024, totalizando 72,5 milhões em junho de 2024 [SERASA 2024]. A pesquisa assinala que hoje mais de 33% da população está endividada e boa parte desse endividamento refere-se ao uso de cartão de crédito. Atualmente 4 em cada 10 brasileiros possuem dividas cartão de crédito [Nascimento 2023].

Nesse sentido, enfrentar o desafio da inadimplência é crucial, e a modelagem de crédito demonstra ser uma ferramenta eficaz para enfrentar esse desafio [Bernanke et al. 1991]. Atualmente, a modelagem de crédito está expandindo para além dos relatórios tradicionais e históricos de pagamento. Novas fronteiras incluem o comportamento dos usuários em aplicativos bancários, que oferecem percepções valiosas sobre o comportamento financeiro. Esses dados, combinados com técnicas de Aprendizado de Máquina (AM), têm potencial para aprimorar a modelagem de crédito, permitindo avaliações mais precisas e personalizadas [Zowasel 2023]. Este estudo propõe aplicar técnicas de modelagem de classificação de crédito para atender às necessidades práticas do cenário econômico brasileiro atual e explorar novas possibilidades na modelagem de crédito, com foco específico no mercado de cartão de crédito. Nessa proposta, foram desenvolvidos três modelos de previsão aplicados a conjuntos de dados distintos. Esses modelos foram utilizados para alimentar um sistema de pontuação ponderado de crédito (*score*) que foi utilizado para realizar a classificação dos clientes.

Dessa forma, o presente estudo se distingue por abordar a questão da modelagem de crédito por meio da integração de múltiplas fontes de dados, incluindo fontes alternativas de informação, como o comportamento do cliente. Ao adotar essa abordagem multifonte, o trabalho amplia a compreensão dos fatores que influenciam o risco de crédito.

O presente artigo está organizado como segue: a seção 2 apresenta uma revisão da literatura da área de modelagem de crédito; a seção 3 traz os materiais e métodos utilizados para desenvolver o estudo; a seção 4 apresenta os resultados obtidos e a seção 5 as considerações finais deste trabalho.

2. Aprendizado de Máquina e Modelagem de Crédito

A modelagem de classificação de crédito desempenha um papel crítico no contexto do setor financeiro, sendo essencial na análise e avaliação do risco de crédito associado a uma ampla variedade de participantes, incluindo tomadores de empréstimos, empresas e indivíduos [Brito and Assaf Neto 2008]. É amplamente reconhecido que uma gestão eficaz do risco de crédito é fundamental para a estabilidade e o sucesso das instituições financeiras, bem como para o sistema financeiro na totalidade [Bernanke et al. 1991]. Os modelos de *credit scoring*, também denominados modelos de classificação de risco de crédito, constituem ferramentas estatísticas empregadas para avaliar a capacidade de crédito de um indivíduo ou empresa, bem como determinar a probabilidade de inadimplência em suas obrigações de crédito [Johnson 2023]. O processo de avaliação realizado por um modelo de classificação de risco de crédito engloba diversos fatores, tais como histórico de pagamento, utilização de crédito, extensão do histórico de crédito, tipos de contas de crédito e consultas recentes de crédito.

Na última década observou-se um significativo progresso no uso de técnicas de Inteligência Artificial (IA) e em especial no uso do Aprendizado de Máquina (AM) em

diversas áreas e em especial na modelagem de crédito. A aplicação desses modelos no setor financeiro, combinada com novas técnicas de gerenciamento de banco de dados para análise e exploração do Big Data, resultou na consolidação dessas tecnologias como um padrão industrial [Silva Neto et al. 2020]. A pesquisa de [Nonato 2022] destaca a prevalência do uso de modelos matemáticos, particularmente a *Logistic Regression*, como um padrão na indústria financeira. No entanto, o mesmo estudo também aponta que os métodos de AM têm demonstrado maior precisão nos últimos anos ao lidar com dados econômicos. Na sua pesquisa, Nonato fez uma contribuição significativa ao destacar que o algoritmo *Random Forest* supera o *Logistic Regression* na tarefa de identificar os clientes mais propensos a liquidar suas dívidas. Além disso, o autor salienta que os modelos de pontuação de crédito ainda se baseiam tradicionalmente em dados históricos de crédito, informações sociodemográficas e registros de pagamentos, entre outros. Por outro lado, informações comportamentais obtidas de aplicativos móveis e dados transacionais detalhados ainda são vistas como fontes alternativas de informação.

A literatura apresenta diversos estudos que comparam e implementam modelos de AM na criação de *score* de crédito. Destaca-se o trabalho de Ge et al., que investigou a detecção de fraude em cartões de crédito utilizando modelos preditivos [Ge et al. 2020]. Neste estudo, os pesquisadores ampliaram o escopo ao empregar um conjunto de dados mais abrangente e robusto, testando algoritmos supervisionados de aprendizado de máquina, como *Logistic Regression*, SVM, *XGBoost* e *LightGBM*. Os resultados revelaram perspectivas importantes para a área de modelagem, destacando o desempenho superior do *LightGBM*, e o impacto positivo do uso de um conjunto de dados mais abrangente na qualidade das previsões. Por outro lado, trabalhos como o de Gahlaut et al. focaram na implementação de sistemas de pontuação de crédito, com ênfase em empréstimos [Gahlaut et al. 2017]. Esses pesquisadores realizaram uma série de testes utilizando algoritmos como *Decision Tree*, SVM, *Adaptive Boosting* (Bagging), *Logistic Regression*, *Random Forest* (RF) e *Neural Network*, visando a criação de um sistema de classificação bidimensional para categorizar os usuários como bons ou maus pagadores [Gahlaut et al. 2017]. Os resultados demonstraram sucesso na implementação de modelos eficazes de classificação de crédito com diferentes algoritmos de ML, tendo o RF o melhor desempenho.

No contexto da modelagem, observa-se uma progressiva consolidação bem-sucedida do uso de modelos matemáticos preditivos de AM para a geração de *scores* em diversas aplicações, incluindo o de cartão de crédito. Entretanto, observa-se que o algoritmo com melhor AUC-ROC de predição varia conforme o conjunto de dados utilizado. Nesse sentido, a contribuição deste estudo está na investigação de engenharia de recursos, na aplicação de ferramentas de modelagem e na realização de uma ampla variedade de testes com diferentes algoritmos de predição. Além disso, o estudo também contribui com a proposta de um sistema de pontuação abrangente e dinâmico, que considera aspectos comportamentais. Em vista disso, considera-se que a presente proposta representa uma fronteira ainda pouco explorada na modelagem de crédito.

3. Materiais e Métodos

Esta seção descreve a metodologia e os materiais empregados neste estudo, detalhando todas as etapas de seu desenvolvimento. Para o desenvolvimento deste estudo utilizou-se um conjunto de dados denominado "*Credit Card Customers and Fraud Risk*",

disponível na plataforma *Kaggle*, que foi criado por Haowen Wang [Wang 2023]. O conjunto divide-se em três grupos distintos: Tag (*User Tag Data*), Trd (*Transaction Behavior Data*) e Beh(*App Behavior Data*). O grupo Tag abrange 41 características relativas ao perfil financeiro e pessoal dos usuários. Esse conjunto compreende uma ampla variedade de dados, que vão desde informações financeiras, como a retenção de cartões de crédito, até detalhes pessoais, como idade, gênero, estado civil e nível educacional. O grupo Trd representa os dados de comportamento de transação. Esses dados compreendem as informações sobre as operações financeiras efetuadas pelos clientes no período de 60 dias anteriores à análise. Esse conjunto inclui dados como a direção das transações, o método de pagamento utilizado, bem como os códigos de classificação primária e secundária de receita e despesa. Além disso, são registrados o horário das transações e os valores envolvidos. O conjunto Trd abrange, no total, 8 características. Por fim, o terceiro grupo de dados denotado como Beh contém dados sobre o uso de ambientes digitais pelos clientes. Dentre os elementos desse conjunto, destaca-se o código da página visitada ou interagida pelo usuário, juntamente com a data e hora dessa interação.

3.1. Tecnologias

Para a execução deste estudo, foi utilizado o ambiente Jupyter Notebook, disponibilizado pelo software Anaconda. Foram empregadas diversas bibliotecas de código aberto do Python, conforme detalhado a seguir.

No pré-processamento dos dados, a biblioteca pandas foi empregada para a manipulação de *DataFrames*, facilitando a leitura, transformação e manipulação dos dados. A codificação de rótulos foi realizada com o LabelEncoder e a normalização dos valores foi feita utilizando o MinMaxScaler, ambos da sklearn.preprocessing. A seleção de características foi efetuada através do RFE e da mutual_info_classif da sklearn.feature_selection. O balanceamento de classes foi abordado com as técnicas de oversampling e undersampling fornecidas pelo SMOTE e NearMiss da imblearn.

Para a construção dos modelos, foram aplicados diversos métodos de predição. A sklearn forneceu ferramentas como LogisticRegression para regressão logística, RidgeClassifier para classificação com regularização Ridge, KNeighborsClassifier para classificação com K-vizinhos e DecisionTreeClassifier para árvores de decisão. Outros métodos incluem RandomForestClassifier, ExtraTreesClassifier, AdaBoostClassifier e BaggingClassifier. Também foram utilizados modelos de Naive Bayes e perceptron multicamadas, bem como técnicas de boosting como XGBClassifier, CatBoostClassifier e LGBMClassifier. A avaliação dos modelos foi realizada com a métrica roc_auc_score, da sklearn.metrics.

A visualização e análise dos dados foram conduzidas com as bibliotecas seaborn e matplotlib.pyplot, enquanto as operações numéricas e a álgebra linear foram realizadas com a biblioteca Numpy. Estas ferramentas foram selecionadas para suportar as diversas etapas do processo de análise e modelagem de dados.

3.2. Pré-processamento dos dados

Durante a exploração dos dados do conjunto de dados Tag, identificou-se a presença de dados ausentes em algumas variáveis. A variável *Método de pagamento do cartão de crédito* (atdd_type) apresentava 59% de dados ausentes, seguida pela variável *Grau acadêmico* (deg_cd) com 52%, *Nível educacional* (edu_deg_cd) com 31%, e *Nível*

acadêmico (acdm_deg_cd) com 0.002%. Na prática, as três últimas variáveis apresentam perspectivas semelhantes sobre os clientes, com informações quase redundantes. A variável mais específica é o deg_cd, com 13 categorias distintas, seguida pelo acdm_deg_cd com 8 categorias e, por fim, o edu_deg_cd com apenas 3 categorias. Diante desse cenário, optou-se excluir deg_cd e edu_deg_cd, mantendo apenas acdm_deg_cd, que ainda fornece uma perspectiva equilibrada das informações dos clientes. Para tratar os dados faltantes na variável *Nível acadêmico*, de natureza categórica, foi aplicada a técnica de imputação proporcional por amostragem aleatória [Kalton and Kish 1981], que amostra valores existentes mantendo a frequência das categorias originais, evitando viés. Como resultado, todos os dados faltantes foram preenchidos. A variável atdd_type foi removida devido ao risco associado ao preenchimento de seus valores ausentes. Testes com o algoritmo SAGAD [Silva et al. 2021] reduziram a taxa de dados ausentes de 59% para 23%, mas não conseguiram imputar os valores restantes, indicando que a imputação poderia gerar representações incorretas e consumir excessivos recursos computacionais. Os conjuntos de dados Trd e Beh não apresentavam dados ausentes e foram mantidos intactos.

Sucessivamente, durante a análise dos dados do conjunto de dados Tag, observou-se a oportunidade de combinar informações e criar uma nova representação dos dados. Isso foi realizado com as variáveis cur_debit_cnt (Número de cartões de débito) e cur_credit_cnt (Número de cartões de crédito), que foram somadas para criar a variável total_cards_cnt. Da mesma forma, as variáveis cur_debit_min_opn_dt_cnt e cur_credit_min_opn_dt_cnt, que representam os dias de posse de cartão de débito e crédito, foram somadas para gerar total_min_opn_dt_cnt. Adicionalmente, as variáveis l12_mon_fnd_buy_whl_tms (Número de compras de fundos nos últimos 12 meses), l12_mon_insu_buy_whl_tms (Número de compras de seguro nos últimos 12 meses) e l12_mon_gld_buy_whl_tms (Número de compras de ouro nos últimos 12 meses) foram somadas para criar total_l12_mon_buy_whl_tms.

Identificou-se também duas variáveis essenciais: dnl_mbl_bnk_ind e dnl_bind_cmb_lif_ind, que representam o download e o *bind* e o download e o *login*, respectivamente. Para simplificar as operações, decidiu-se criar a variável dnl, onde o valor é 1 se pelo menos uma das duas condições especificadas nas *features* citadas anteriormente for verdadeira, e 0 caso contrário. Por outro lado, os conjuntos de dados Trd e Beh, devido à sua baixa dimensionalidade, não passaram por nenhum processo de remoção ou fusão de dados. No entanto, ambas as bases tiveram suas variáveis de tempo, trx_time e time, respectivamente, convertidas para números inteiros. Ao concluir o processo de agrupamento nos conjuntos de dados, optou-se por empregar a técnica de *Label Encoding* para converter as colunas categóricas em valores numéricos em todos os 3 *data frames*. O Label Encoding atribui um número inteiro único a cada categoria presente nas colunas convertidas, viabilizando uma maior economia de memória [Kosaraju et al. 2023]. Pós a isso, os valores contidos no conjunto de dados foram normalizados para o intervalo entre 0 e 1, visando reduzir a sensibilidade a *outliers* [Ali et al. 2014].

Após a limpeza e transformação de dados, o conjunto Tag reduziu sua dimensionalidade de 41 para 33, ainda considerado elevado. Para lidar com isso e otimizar o poder de processamento, optou-se pela aplicação da técnica IGRF-RFE. Essa é uma técnica híbrida de seleção de características, unindo métodos de filtro e *wrapper*. Ele visa selecionar um subconjunto ideal de características a partir de grandes conjuntos de dados [Yin et al. 2023]. A técnica íntegra três métodos: *Information Gain* (IG), *Random Forest*

Importance (RFI) e *Recursive Feature Elimination* (RFE). A abordagem híbrida oferece benefícios significativos, combinando a eficiência dos métodos de filtro com a precisão dos métodos de *wrapper*. Além disso, consegue lidar com conjuntos de dados de alta dimensão, selecionando um subconjunto informativo de características [Yin et al. 2023]. Isso contribuiu para reduzir o risco de *overfitting* e melhorar a capacidade de generalização do modelo. Assim, a técnica foi utilizada no conjunto Tag para reduzir sua dimensionalidade de 33 para cerca de 15 variáveis, aproximadamente a metade do número inicial do processo de seleção.

3.3. Balanceamento dos Dados e Redução de Dimensionalidade

No conjunto de dados Tag, que contém cerca de 18 mil registros, optou-se por empregar a técnica SMOTE [Blagus and Lusa 2013] antes do processo de modelagem para balanceamento. Essa é uma técnica de aumento de dados que visa resolver o problema de desequilíbrio de classes. Nessa técnica geram-se novos exemplos sintéticos da classe minoritária por meio da interpolação entre instâncias pertencentes à mesma classe. Ao fazer isso, SMOTE aumenta a representação da classe minoritária, tornando-a mais balanceada em relação às outras classes. Nos *data frames* Trd e Beh, que contêm mais de 1 milhão e 900 mil registros, respectivamente, optou-se por empregar a técnica *NearMiss* para balanceamento dos conjuntos. Essa técnica baseia-se em subamostragem que seleciona exemplos da classe majoritária que estão "mais próximos" dos exemplos da classe minoritária, a fim de reduzir a disparidade entre as classes [Tanimoto et al. 2022]. Dessa forma, reduz-se o desequilíbrio entre as classes evitando distorções no modelo. Além disso, ao reduzir a quantidade de exemplos da classe majoritária, a técnica *NearMiss* melhorara a eficiência computacional ao reduzir o tempo de treinamento do modelo, uma vez que menos dados precisam ser processados [Tanimoto et al. 2022].

Ademais, após a aplicação do *NearMiss* no conjunto Beh, optou-se por utilizar a técnica de Análise Discriminante Linear (em sua sigla em inglês, LDA) para gerar um único componente. O LDA é uma técnica de redução de dimensionalidade que visa maximizar a separação entre as classes enquanto mantém a variância dentro de cada classe o mais baixa possível [Anowar et al. 2021]. A escolha do LDA após o balanceamento com NearMiss traz vantagens significativas para a modelagem. Ao reduzir a dimensionalidade do conjunto de dados para apenas um componente, o LDA simplifica a representação dos dados. Além disso, ao preservar a estrutura de classes no processo de redução de dimensionalidade, o LDA melhora a capacidade do modelo de discriminar entre as classes, resultando em melhores resultados de classificação [Anowar et al. 2021].

Os resultados obtidos com a aplicação do balanceamento dos dados estão detalhados na Tabela 1. Essa tabela permite uma comparação direta entre as proporções de clientes inadimplentes e não inadimplentes antes e após o balanceamento, evidenciando o impacto das técnicas empregadas na distribuição das classes, como esperado.

Tabela 1. Comparação da Proporção de Clientes Inadimplentes e Não Inadimplentes Antes e Após o Balanceamento

Conjunto de dados	Inadimplente	Não Inadimplente
Tag	22,4%	77,6%
Tag (Após Balanceamento)	50,0%	50,0%
TRD	15,7%	84,3%
TRD (Após Balanceamento)	50,0%	50,0%
Beh	15,1%	84,9%
Beh (Após Balanceamento)	50,0%	50,0%

3.4. Construção dos modelos

Para a construção dos modelos preditivos, decidiu-se inicialmente realizar um conjunto de testes, englobando diversos algoritmos de aprendizado de máquina supervisionado. Dentre eles, foram selecionados 13 algoritmos, abrangendo diferentes formas de classificação. A seleção dos algoritmos para os testes foi baseada naqueles previamente avaliados por pesquisadores em estudos relacionados (seção 2) e naqueles discutidos na literatura acadêmica [Ray 2019, Bonaccorso 2018]. Essa abordagem garantiu uma análise abrangente e representativa das técnicas de aprendizado de máquina disponíveis e sua aplicação no contexto investigado.

Ademais, a métrica escolhida para a avaliação dos modelos foi o AUC-ROC. Essa métrica avalia a capacidade do modelo de distinguir entre as classes positiva e negativa ao longo de todos os possíveis limiares de decisão. O AUC-ROC fornece uma medida geral da eficácia do modelo na separação das classes. A opção por essa métrica justifica-se pelo objetivo de integrar os modelos em um sistema mais complexo e probabilístico de pontuação de crédito. Um modelo com um AUC-ROC mais elevado indica uma capacidade superior de gerar probabilidades mais confiáveis, permitindo uma discriminação mais eficaz entre clientes inadimplentes e não inadimplentes.

4. Resultados e Discussões

Esta seção apresenta os resultados obtidos a partir dos testes realizados com diferentes modelos nos três conjuntos de dados, além de detalhar o modelo de pontuação de crédito proposto e a avaliação de sua implementação.

4.1. Testes dos Algoritmos

Para a condução dos experimentos, o parâmetro *Random State* foi configurado como 42 para todos os algoritmos, a fim de assegurar a consistência dos resultados. No entanto, para o algoritmo KNN, foram mantidas as configurações padrão, garantindo a replicabilidade e comparabilidade dos resultados obtidos. Os resultados dos testes dos modelos estão apresentados na Tabela 2.

Os testes revelaram que o modelo *Extra Trees Classifier* (ETC) apresentou o melhor desempenho para o conjunto de dados Tag, com um AUC-ROC de 88,26%. Para o conjunto Trd, o modelo Bagging demonstrou ser o mais eficaz, alcançando um AUC-ROC de 84,32%. Por outro lado, o ETC foi o modelo mais adequado para o conjunto Beh, obtendo um AUC-ROC de 70,65%. Os desempenhos do ETC nos conjuntos Tag e Beh, bem como do Bagging no conjunto Trd, foram considerados satisfatórios para o sistema de *Score*.

Tabela 2. Resultado dos Testes dos Algoritmos de ML

Modelo	Tag: AUC-ROC	Trd: AUC-ROC	Beh: AUC-ROC
AdaBoost	77,65%	72,49%	67,98%
Bagging	82,88%	**84,32%**	69,20%
Bernoulli NB	62,09%	63,94%	64,63%
CatBoost	85,47%	83,53%	68,12%
Decision Tree	77,99%	84,04%	69,62%
ETC	**88,26%**	83,83%	**70,65%**
KNN	73,38%	81,75%	68,05%
LightGBM	84,81%	83,25%	68,17%
Logistic Regression	61,51%	63,76%	64,81%
Neural Network	70,05%	69,74%	66,58%
Random Forest	85,74%	80,02%	69,73%
Ridge Classifier	61,46%	63,70%	64,63%
XGBoost	84,83%	83,37%	68,16%

4.2. Proposta de Score de Crédito

Este estudo propõe um sistema de avaliação de risco que utiliza várias fontes de dados. Conforme representado na Figura 1, o sistema é abastecido por três conjuntos de dados distintos que utiliza os modelos escolhidos na seção 3.3. O processo do *score* envolve dois fluxos de ação: primeiro, as informações são extraídas dos conjuntos de dados do banco sendo submetidas aos modelos de predição correspondentes, registrando-se o retorno desses modelos nos respectivos conjuntos como probabilidade de inadimplência; segundo, são extraídos o ID e a probabilidade desses conjuntos para efetuar o processamento e criar o *score*, registrando-o no banco de dados.

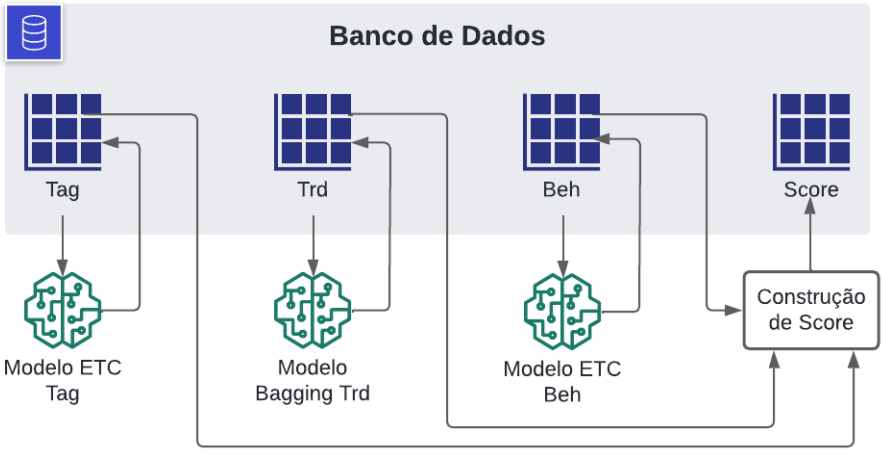


Figura 1. Processo de Construção de *Score*

A primeira etapa do segundo fluxo de informações consiste em identificar o ID obtido nos dados do conjunto Tag e, a partir desse, localizar os registros correspondentes desse cliente nos conjuntos de dados Trd e Beh. Os registros nesses dois últimos conjuntos

podem ser múltiplos, visto que um único cliente pode realizar diversas transações financeiras e acessos à aplicação da instituição bancária. Isso cria um desafio: se cada ação pode resultar em uma probabilidade de inadimplência diferente, como lidar com esses múltiplos retornos e obter uma probabilidade geral.

A área de Finanças Comportamentais atualmente reconhece que a independência ou dependência dos comportamentos financeiros de uma mesma pessoa pode variar de acordo com uma série de fatores, incluindo o planejamento das finanças pessoais [Piaia et al. 2008]. Assim, apresentando uma incerteza na natureza da correlação entre diferentes comportamentos. No presente estudo, optou-se por considerar os comportamentos e transações registrados nos conjuntos Beh e Trd como independentes entre si. Para resolver o desafio mencionado anteriormente, adotou-se a abordagem da probabilidade conjunta como a probabilidade geral.

Usou-se a soma dos logaritmos naturais (*S*) das probabilidades (*P*) para evitar problemas de *underflow* numérico (Equação 1). A soma dos logaritmos de um conjunto de números é igual ao logaritmo do produto desses números, equivalendo a multiplicar as probabilidades originais. Após somar os logaritmos, reverteu-se a operação do logaritmo elevando *e* a *S* para obter a probabilidade conjunta real (*PC*) (Equação 2).

$$S = \log(P_1) + \log(P_2) + \ldots + \log(P_n) \tag{1}$$

$$PC = e^S \tag{2}$$

Assim, os múltiplos retornos de cada conjunto de dados são submetidos a essa abordagem, resultando em uma probabilidade geral (*PC*) para Trd e outra para Beh. Ao final, para a criação do *score*, utilizou-se três probabilidades: a do conjunto Tag, a geral de Trd e a geral de Beh. É importante ressaltar que nenhuma probabilidade foi arredondada para 0 ou 1, preservando assim o retorno probabilístico dos modelos.

Para o desenvolvimento do sistema de pontuação, decidiu-se utilizar a métrica ponderada, na qual os pesos de cada probabilidade são definidos pelo AUC-ROC do modelo correspondente em relação ao todo. Desse modo, o *score* (*Sc*) é composto pela soma da probabilidade Tag (*Ptag*) ponderada pela probabilidade geral Trd (*Ptrd*) ponderada pela da probabilidade geral Beh (*Pbeh*). A Proposta de *score* está descrita na Equação 3, na qual *Atag*, *Atrd* e *Abeh* representam as AUC-ROC dos modelos ETC, Bagging e ETC aplicados aos conjuntos de dados Tag, Trd e Beh, respectivamente.

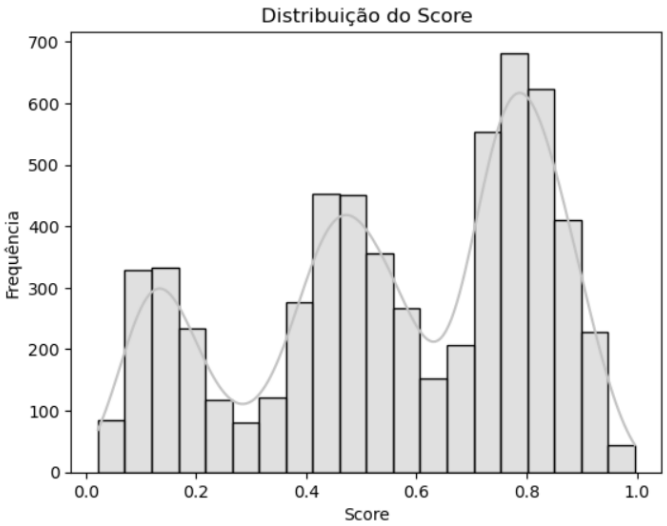
$$\begin{aligned} Sc &= PTag \times \tfrac{Atag+Atrd+Abeh}{Atag} \\ Sc &= Sc + PTrd \times \tfrac{Atag+Atrd+Abeh}{Atrd} \\ Sc &= Sc + PBeh \times \tfrac{Atag+Atrd+Abeh}{Abeh} \end{aligned} \tag{3}$$

No sistema de pontuação proposto, a relação entre a pontuação de crédito e a probabilidade de inadimplência é diretamente relacionada: quanto maior o *score*, maior a

probabilidade de o cliente ser inadimplente; e quanto menor o *score*, menor essa probabilidade. Portanto, uma pontuação considerada "alta" indica uma situação desfavorável, enquanto uma "baixa" é indicativo de uma condição positiva. É importante observar que o *score* varia numa escala de 0 a 1. Conforme já citado anteriormente, para implementar este sistema, foram utilizados conjuntos de dados de teste fornecidos por Wang [Wang 2023] no Kaggle, denominados df Tag, Trd e Beh. É importante ressaltar que esses conjuntos de dados não incluem indicadores prévios de inadimplência.

4.3. Teste de Implementação do Score

Ao implementar o sistema de pontuação proposto nos conjuntos de teste, gerou-se um novo conjunto de dados denominado *score*, que registrou o ID e o a pontuação de crédito dos clientes, como pode ser observado na Figura 2(A). Como resultado do processo de classificação, foram obtidos scores para 6000 clientes, variando de 0 a 1, cuja distribuição pode ser observada na Figura 2(B).



(A) Resultado do Score dos Clientes

	id	score
0	UC37930	0.132939
1	U5BE130	0.881200
2	UD025AE	0.117789
3	UC2D00D	0.082800
4	UAF705D	0.443600
...	...	...
5995	U71611F	0.892000
5996	U52CFE4	0.899200
5997	U6A5425	0.634903
5998	UB8871F	0.544415
5999	U5573FB	0.695716

6000 rows × 2 columns

(B) DF Score

Figura 2. Score - Teste de Implementação

Os resultados obtidos demonstraram que o estudo alcançou o seu objetivo, gerando um sistema de pontuação dinâmico a partir do uso de diferentes fontes de dados.

5. Conclusão e Trabalhos Futuros

O presente estudo propôs-se a desenvolver um sistema de pontuação de crédito utilizando técnicas de aprendizado de máquina, com o intuito de explorar as novas fronteiras da modelagem de crédito e, simultaneamente, oferecer uma possível solução para os desafios sociais atuais enfrentados pela sociedade brasileira. Os resultados obtidos ao longo da pesquisa destacam que as métricas alcançadas foram satisfatórias. Os modelos de previsão utilizados demonstraram altas taxas de AUC-ROC, enquanto as abordagens propostas para o desenvolvimento de um sistema de pontuação dinâmico e multifonte mostraram-se viáveis e capazes de produzir resultados positivos.

Entretanto, é importante realizar estudos adicionais, incluindo comparações com outras abordagens de modelagem de predição e pré-processamento de conjuntos de dados, a fim de validar ainda mais a eficácia dos modelos empregados neste estudo. Além disso, sugere-se conduzir testes e a implementação de diferentes sistemas de pontuação de crédito, visando comparar os resultados obtidos após a aplicação desses sistemas com os resultados obtidos neste estudo.

Referências

Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., and Faraj, R. H. (2014). Data normalization and standardization: a technical report. *Mach Learn Tech Rep*, 1(1):1–6.

Anowar, F., Sadaoui, S., and Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, 40:100378.

Bernanke, B. S., Lown, C. S., and Friedman, B. M. (1991). The credit crunch. *Brookings papers on economic activity*, 1991(2):205–247.

Blagus, R. and Lusa, L. (2013). Smote for high-dimensional class-imbalanced data. *BMC bioinformatics*, 14:1–16.

Bonaccorso, G. (2018). *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd.

Brito, G. A. S. and Assaf Neto, A. (2008). Modelo de classificação de risco de crédito de empresas. *Revista Contabilidade Finanças*, 19(46):18–29.

Feldmann, P. (2023). É a alta taxa de juros que prejudica o país. Jornal USP, 2023. Disponível em: https://jornal.usp.br/articulistas/paulo-feldmann/e-a-alta-taxa-de-juros-que-prejudica-o-pais/. Acesso em: 28 de Set. de 2023.

Gahlaut, A., Singh, P. K., et al. (2017). Prediction analysis of risky credit using data mining classification models. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

Ge, D., Gu, J., Chang, S., and Cai, J. (2020). Credit card fraud detection using lightgbm model. In *2020 International Conference on E-Commerce and Internet Technology (ECIT)*, pages 232–236.

Johnson, B. (2023). Credit scoring models 101: Types and examples for a stronger financial future. AVP, Global Enablement, 2023. Disponível em: https://www.highradius.com/resources/Blog/credit-scoring-models-types-and-examples/. Acesso em: 28 de Set. de 2023.

Kalton, G. and Kish, L. (1981). Two efficient random imputation procedures. In *Proceedings of the survey research methods section*, pages 146–151. American Statistical Association.

Kosaraju, N., Sankepally, S. R., and Mallikharjuna Rao, K. (2023). Categorical data: Need, encoding, selection of encoding method and its emergence in machine learning models—a practical review study on heart disease prediction dataset using pearson correlation. In *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 1*, pages 369–382. Springer.

Nascimento, A. (2023). Quatro em cada 10 brasileiros têm dívidas no rotativo do cartão de crédito. O Tempo, 2023. Disponível em: https://www.otempo.com.br/economia/quatro-em-cada-10-brasileiros-tem-dividas-no-rotativo-do-cartao-de-credito-1.2852910. Acesso em: 20 de Jun. de 2023.

Neto, R. C. (2023). Comissão de assuntos econômicos ouve presidente do banco central – 25/4/23. TV Senado, 2023. Disponível em: https://www.youtube.com/watch?v=TESK8O75Zpoab_channel=TVSenado. Acesso em: 25 de abril de 2023.

Nonato, C. T. (2022). Machine learning aplicado na concessão de crédito: estudo comparativo. Master's thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. Acesso em: 17 maio 2024.

Piaia, C. F. et al. (2008). Finanças pessoais e independência financeira: a educação e organização financeira como instrumentos de melhoria na vida das pessoas.

Ray, S. (2019). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 35–39. IEEE.

SERASA (2024). Mapa da inadimplência e negociação de dívidas no brasil. Disponível em: https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/. Acesso em: 6 de agosto de 2024.

Silva, H., Silva, R., and Porto, F. (2021). Sagad: Synthetic data generator for tabular datasets. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 1–12, Porto Alegre, RS, Brasil. SBC.

Silva Neto, V. J. d., Bonacelli, M. B. M., and Pacheco, C. A. (2020). O sistema tecnológico digital: inteligência artificial, computação em nuvem e big data. *Revista Brasileira de Inovação*, 19:e0200024.

Tanimoto, A., Yamada, S., Takenouchi, T., Sugiyama, M., and Kashima, H. (2022). Improving imbalanced classification using near-miss instances. *Expert Systems with Applications*, 201:117130.

Wang, H. (2023). Engenheiro de algoritmo. LinkedIn, 2023. Disponível em: https://www.linkedin.com/in/haowen-harvin-wang-77432a132/. Acesso em: 16 de Jun. de 2023.

Yin, Y., Jang-Jaccard, J., Xu, W., Singh, A., Zhu, J., Sabrina, F., and Kwak, J. (2023). Igrf-rfe: a hybrid feature selection method for mlp-based network intrusion detection on unsw-nb15 dataset. *Journal of Big data*, 10(1):15.

Zanatta, P. and Nassif, T. (2023). Queda de investimentos, menos crédito e desemprego: os efeitos negativos dos juros altos para a economia. CNN Brasil, 2023. Disponível em: https://www.cnnbrasil.com.br/economia/remedio-amargo-juros-altos-trazem-consequencias-negativas-para-a-economia-entenda/. Acesso em: 28 de Set. de 2023.

Zowasel (2023). The future of alternative credit scoring for digital banks. Medium, 2023. Disponível em: https://medium.com/zowasel/the-future-of-alternative-credit-scoring-for-digital-banks-749838ec66fd. Acesso em: 28 de Set. de 2023.