MESTRADO PROFISSIONAL

EM ECONOMIA

UTILIZAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA EM DUAS DAS MAIORES INSTITUIÇÕES BANCÁRIAS BRASILEIRAS PARA CÁLCULO DE RISCO DE CRÉDITO PARA O CRÉDITO DIRETO AO CONSUMIDOR

ULYSSES ARAUJO BISPO

Brasília-DF, 2024

ULYSSES ARAUJO BISPO

UTILIZAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA EM DUAS DAS MAIORES INSTITUIÇÕES BANCÁRIAS BRASILEIRAS PARA CÁLCULO DE RISCO DE CRÉDITO PARA O CRÉDITO DIRETO AO CONSUMIDOR

Dissertação apresentada ao Programa de Pós Graduação em Economia, do Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa, como requisito parcial para obtenção do grau de Mestre.

Orientador
Professor Doutor Mathias Schneid Tessmann.

Brasília-DF 2024

ULYSSES ARAUJO BISPO

UTILIZAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA EM DUAS DAS MAIORES INSTITUIÇÕES BANCÁRIAS BRASILEIRAS PARA CÁLCULO DE RISCO DE CRÉDITO PARA O CRÉDITO DIRETO AO CONSUMIDOR

Dissertação apresentada ao Programa de Pós Graduação em Economia, do Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa, como requisito parcial para obtenção do grau de Mestre.

Aprovado em 04 / 12 / 2024

Banca Examinadora

Prof. Dr. Mathias Schneid Tessmann - Orientador

Prof. Dr. Sergio Jurandyr Machado

Prof. Dr. Daniel Tavares de Castro

Código de catalogação na publicação – CIP

Cutter Bispo, Ulysses Araujo

Utilização de modelos de aprendizado de máquina em duas das maiores instituições bancárias brasileiras para cálculo de risco de crédito para o crédito direto ao consumidor / Ulysses Araujo Bispo. — Brasília: Instituto Brasileiro Ensino, Desenvolvimento e Pesquisa, 2024

55 f. :

Orientador: Prof. Dr. Mathias Schneid Tessmann

Dissertação (Mestrado Profissional em Economia) — Instituto Brasileiro Ensino, Desenvolvimento e Pesquisa – IDP, 2025.

1. Análise de riscos. 2. Crédito direto ao consumidor. 3. Crédito bancário. 4. Aprendizado de máquina. I.Título

CDD 330

Elaborada pela Biblioteca Ministro Moreira Alves

AGRADECIMENTOS

Agradeço a Deus e a minha família, Jorge, Eliseth, Mateus e Luana em especial a meu pai que durante tempos difíceis foi o suporte da nossa família. Agradeço também a minha amiga Marilia e ao meu trabalho especial Julcermir, Cazza e Eduardo Agra que que tiveram paciência comigo e me ensinaram muito e por fim e não menos importante ao professor Mathias que me respondia prontamente ainda que seja de madrugada.

RESUMO

Este trabalho busca avaliar qual dos métodos de aprendizado de máquina apresenta maior eficiência na análise de risco de crédito. Para isso, foram considerados os modelos Suport Vector Machine (SVM), Deep Learning com Perceptron Multicamadas (MLP), Gradient Boosting e Decision Tree, utilizando métricas como acurácia, precisão, recall, F1-Score, AUC-ROC e Validação Cruzada para compará-los. Além disso, foi utilizada a Logistic Regression, um dos modelos mais aplicados no mercado bancário, como referência para demonstrar que todos os métodos de aprendizado de máquina analisados oferecem um desempenho superior ao da regressão Logistica. Foram utilizados dados de duas das maiores instituições financeiras brasileiras, referidas como Banco A e Banco B, com bases de clientes analisadas ao longo de 12 meses para identificar padrões de inadimplência. Os resultados revelam que a aplicação de modelos de aprendizado de máquina proporciona uma ferramenta mais robusta para a análise de risco de crédito no setor financeiro, especialmente em empréstimos de crédito direto ao consumidor (CDC).

Palavras-chave: Análise de Risco de Crédito, Modelos de Aprendizado de Máquina, Deep Learning, Multilayer Perceptron.
Classificação JEL: C45, C52, G21, G32

ABSTRACT

This paper aims to evaluate which machine learning methods demonstrate the highest efficiency in credit risk analysis. For this purpose, the models Support Vector Machine (SVM), Deep Learning with Multilayer Perceptron (MLP), Gradient Boosting, and Decision Tree were considered, utilizing metrics such as accuracy, precision, recall, F1-Score, AUC-ROC, and Cross-Validation to compare them. Additionally, Logistic Regression, one of the most applied models in the banking market, was used as a reference to demonstrate that all analyzed machine learning methods offer superior performance compared to Logistic Regression. Data from two of the largest Brazilian financial institutions, referred to as Bank A and Bank B, were used, with customer bases analyzed over 12 months to identify default patterns. The results reveal that the application of machine learning models provides a more robust tool for credit risk analysis in the financial sector, especially in direct consumer credit (CDC) loans.

Keywords: Credit Risk Analysis, Machine Learning Models, Deep Learning, Multilayer Perceptron.

LISTA DE ABREVIATURAS E SIGLAS

AUC-ROC	Área Sob a Curva Característica de Operação do Receptor
MLP	Perceptron Multicamadas (Multilayer Perceptron)
SVM	Máquina de Suporte Vetorial (Support Vector Machine)
BACEN	Banco Central do Brasil
JEL	Classificação do Journal of Economic Literature
RBF	Radial Basis Function (Função de Base Radial)
ReLU	Rectified Linear Unit (Unidade Linear Retificada)
L1/L2	Regularização L1/L2 (Lasso/Ridge)
XGBoost	Extreme Gradient Boosting
RNA	Redes Neurais Artificiais
CART	Classification and Regression Tree (Árvore de Classificação e Regressão)
AM	Aprendizado de Máquina
Big Data	Grande Volume de Dados
OECD	Organização para a Cooperação e Desenvolvimento Econômico (Organization for Economic Co-operation and Development)
COVID-19	Coronavirus Disease 2019

LISTA DE ILUSTRAÇÕES

Figura 1
Correlação entre as variáveis
..... 25
Figura 2
Correlação entre as variáveis
..... 29
Figura 4
Gráfico regressão Logística
..... 32

LISTA DE TABELAS

Tabela 1
Variáveis transformadas em categorias
..... 24
Tabela 2
Estatísticas descritivas
..... 26
Tabela 3
Conversão de Variáveis Transformadas em Categóricas
..... 28
Tabela 4
Estatísticas descritivas
..... 30
Tabela 5
Resultados do Banco A
..... 45
Tabela 6
Resultados do Banco B
..... 46

SUMÁRIO

1. INTRODUÇÃO 13

2. FUNDAMENTAÇÃO TEÓRICA 17

3. METODOLOGIA 23

3.1 DADOS 23

3.1.2 BANCO A 23

3.1.3 BANCO B 27

3.2 LOGISTIC REGRESSION 31

3.3 DECISION TREE 32

3.4 GRADIENT BOOSTING 35

3.5 SUPPORT VECTOR MACHINES 36

3.6 DEEP LEARNING O MULTILAYER PERCEPTRON (MLP) 38

4. RESULTADOS 44

5. CONCLUSÃO 49

REFERÊNCIAS 51


1

1 INTRODUÇÃO

A crescente integração econômica e financeira em nível mundial, impulsionada pelo avanço das tecnologias da informação, resultou em um ambiente de constante mudança, frequentemente abalado por crises econômicas, financeiras e, mais recentemente, sanitárias, como a pandemia de COVID-19 (Coelho et al., 2021). Estes eventos têm impactos significativos sobre a sociedade, especialmente no que tange à economia e ao mercado financeiro, influenciando a capacidade de pagamento de empresas e pessoas físicas. A instabilidade provocada por essas crises leva tanto pessoas físicas quanto empresas a enfrentarem dificuldades financeiras, que podem culminar em situações extremas de falência. Isso desestimula as vendas a crédito pelas empresas e aumenta o risco e o custo do crédito no mercado financeiro (Coelho et al., 2021).

Nesse contexto de instabilidade econômica, o gerenciamento eficaz do risco de crédito torna-se essencial para as instituições financeiras. Segundo Schrickel (2000), risco de crédito é "o ato de disponibilização de recursos próprios, ou de terceiros, em troca de uma remuneração futura em um prazo previamente estipulado". Ou seja, envolve a possibilidade de o tomador de empréstimo não cumprir com suas obrigações financeiras, impactando negativamente a instituição que concedeu o crédito. A importância do tema risco de crédito é amplificada pela necessidade de instituições financeiras se manterem solventes e capazes de suportar choques econômicos. Modelos de risco de crédito eficazes são essenciais para garantir que os bancos possam emprestar com segurança, contribuindo assim para uma economia mais sustentável (OECD, 2021).

A introdução do Big Data no gerenciamento de risco de crédito alterou fundamentalmente a forma como as instituições financeiras avaliam e gerenciam esse risco (Guo et al., 2016; Leng et al., 2017). Tradicionalmente, o gerenciamento de risco de crédito dependia fortemente de dados históricos e do julgamento subjetivo de especialistas para prever a probabilidade de inadimplência dos tomadores de empréstimos (Kang & Ausloos, 2017). Essa abordagem convencional, embora eficaz até certo ponto, era limitada pela disponibilidade e profundidade dos dados, o que frequentemente

resultava em avaliações de risco menos precisas (Butaru et al., 2016). Com o advento das tecnologias de Big Data, as instituições financeiras passaram a incorporar grandes quantidades de dados estruturados e não estruturados de diversas fontes, como atividades de mídia social, histórico de transações e indicadores econômicos em tempo real (Bi & Liang, 2022).

Essa integração de diversos tipos de dados possibilita uma abordagem mais abrangente e dinâmica da avaliação de risco, melhorando significativamente a precisão e o poder preditivo dos modelos de risco de crédito (Bi & Liang, 2022). Além disso, o Big Data permite a análise de grandes volumes de dados em tempo real, fornecendo às instituições financeiras insights acionáveis que podem ser usados para gerenciar e mitigar riscos de forma proativa (Islam, 2024; Maraj et al., 2024). Dessa forma, a aplicação do Big Data no gerenciamento do risco de crédito não é apenas uma melhoria incremental, mas representa um afastamento significativo dos métodos tradicionais, oferecendo uma estrutura mais robusta para entender e administrar o risco em um ambiente econômico volátil (Rahman et al., 2024).

Apesar dos benefícios evidentes da incorporação do Big Data ao gerenciamento do risco de crédito, a transição enfrenta desafios significativos (Rahman et al., 2024). Um dos principais problemas é a qualidade e a integração dos dados. O grande volume de informações geradas de várias fontes pode ser avassalador, e garantir que esses dados sejam precisos, relevantes e oportunos é fundamental para o sucesso de qualquer iniciativa de Big Data (Younus et al., 2024).

Diante disso, o objetivo deste trabalho é verificar qual dos métodos de aprendizado de máquina apresenta maior eficiência na análise de risco de crédito. Foram considerados os modelos Decision Trees (árvores de decisão), Support Vector Machines (Máquinas de Suporte Vetorial), Gradient Boosting (Gradient Boosting) e por último Artificial Neural Networks focado em Deep Learning (Redes Neurais Artificial focada em Aprendizado Profundo), utilizando métricas como acurácia, precisão, recall, F1-Score, AUC-ROC e Validação Cruzada para compará-los. Além disso, utilizou-se a Regressão Logística, um dos modelos mais aplicados no mercado bancário, como referência para demonstrar que todos os métodos de aprendizado de máquina analisados oferecem um desempenho superior. Para realizar esta análise, foram utilizadas bases de dados de duas das maiores

instituições financeiras do Brasil que, por questões de sigilo, serão tratadas como Banco A e Banco B. Quanto à temporalidade das bases, ambas são compostas por uma safra de clientes que foi analisada durante 12 meses e, ao final do período, foi gerada uma variável target do tipo booleano em que 1 representa que o cliente foi inadimplente ao longo dos 12 meses e 0 representa que foi adimplente.

Este trabalho está dividido em mais três partes: o referencial teórico, que visa entender a importância da análise de crédito e quais são os principais modelos de aprendizado de máquina abordados na literatura atual; a metodologia, que explica como os modelos foram construídos, seus resultados e as métricas de significância e ajuste utilizadas para medir se os modelos conseguem se ajustar bem aos cenários; e, por úlimo, a conclusão, que, à luz dos resultados obtidos, explora por que o modelo de Deep Learning with Multilayer Perceptron (MLP), foi o escolhido como vencedor diante dos cenários dos dois bancos analisados para a linha de Crédito Direto ao Consumidor (CDC).


2

2 FUNDAMENTAÇÃO TEÓRICA

Historicamente, a avaliação de risco de crédito era realizada por meio de métodos subjetivos, como o modelo dos "5 Cs" — Caráter, Capacidade, Condição, Capital e Colateral —, que dependia fortemente da experiência do gestor e da análise humana. Esses cinco fatores formavam uma estrutura qualitativa essencial para decisões de crédito, sendo aplicados de maneira subjetiva para avaliar o perfil de risco dos clientes. No entanto, com o avanço da tecnologia e o aumento da disponibilidade de dados, a gestão de risco de crédito passou a incorporar técnicas de Machine Learning, ou machine learning (AM), permitindo uma análise mais robusta e precisa. Essas novas abordagens substituem a subjetividade dos 5 Cs, possibilitando o uso de algoritmos que identificam padrões complexos em grandes volumes de dados e geram previsões mais acuradas sobre a probabilidade de inadimplência. Os modelos de Machine Learning conseguem integrar dados estruturados e não estruturados, como informações transacionais e comportamentais, que complementam a análise tradicional.

Os modelos de Machine Learning permitem uma avaliação dinâmica e adaptativa, ajustando as previsões em tempo real conforme novos dados são incorporados, reduzindo a necessidade de análises qualitativas isoladas e aumentando a capacidade de escalabilidade, já que as instituições financeiras podem processar volumes massivos de dados de forma eficiente. Apesar dos desafios com a interpretabilidade dos modelos mais complexos, os avanços em Machine Learning proporcionam uma análise de risco de crédito que alia precisão e profundidade, transformando a estrutura original dos 5 Cs em um processo automatizado, adequado ao cenário financeiro contemporâneo (Montevechi et al., 2024).

Nos últimos anos, os modelos de machine learning e os algoritmos ensemble (uma técnica em machine learning que combina múltiplos modelos para melhorar o desempenho e a precisão das previsões), como o Gradient Boosting, permitem capturar padrões ocultos em grandes volumes de dados, eliminando ainda mais a necessidade de julgamentos subjetivos. Suhadolnik et al. (2023) reforçam essa ideia, mostrando que o uso de algoritmos avançados,

como XGBoost e deep neural networks, promove uma análise quantitativa mais precisa, baseada puramente em dados. Zhang e Yu (2024) acrescentam que a diversidade de dados disponível na era do Big Data possibilita uma modelagem detalhada que afasta a análise de crédito de métodos baseados em intuição e aproxima a decisão de crédito de um processo mais imparcial e objetivo.

Com a chegada da era do Big Data e o aumento das capacidades computacionais, modelos baseados em algoritmos de Machine Learning, such as decision trees and the Gradient Boosting, começaram a se destacar na análise de crédito (Lessmann et al., 2015). Essas técnicas permitiram o processamento de grandes volumes de dados, oferecendo maior precisão em previsões e possibilitando a criação de perfis detalhados para cada cliente (Markov et al., 2022). Segundo Zhong et al. (2014), deep learning artificiais (RNA) e máquinas de vetores de suporte (SVM) destacam-se como as abordagens de machine learning (AM) mais empregadas. Em contraste com as técnicas estatísticas tradicionais, os métodos de machine learning dispensam o conhecimento prévio das relações entre as variáveis de entrada e saída.

Com o advento de sistemas como o Open Banking, há maior acesso a dados financeiros e comportamentais dos consumidores, o que permite a criação de perfis de crédito mais completos e a utilização de scores comportamentais, ou behavior scores, que refinam ainda mais as análises tradicionais (Vicente, 2020). Estes scores consideram o comportamento de consumo, padrões de pagamento e até dados externos, o que potencializa a capacidade preditiva dos modelos de risco (Bravo et al., 2023). O behavioral scoring permite personalizar a avaliação do crédito com base em dados comportamentais específicos, uma abordagem cada vez mais adotada em modelos de machine learning. Por exemplo, Wang et al. (2018) desenvolveram um modelo que utiliza o behavioral scoring para calcular a probabilidade de inadimplência em plataformas de empréstimos peer-to-peer, utilizando dados históricos de comportamento do usuário, o que melhora a precisão preditiva em comparação com métodos puramente financeiros.

No contexto dos modelos de Machine Learning, os métodos baseados em ensembles, como o Gradient Boosting e as florestas aleatórias, demonstraram ser particularmente eficazes. Esses modelos combinam múltiplas árvores de decisão para melhorar a robustez e a

precisão das previsões, compensando a falta de desempenho de um único modelo (Chopra e Bhilare, 2018). Estudos recentes apontam que o Gradient Boosting, em particular, é superior na previsão de inadimplência em conjuntos de dados desequilibrados, um cenário comum no setor financeiro (Zhang e Yu, 2024). Pesquisas recentes têm indicado que a combinação de classificadores, ou ensemble, apresenta desempenho superior ao das técnicas de inteligência artificial isoladas (Wang et al., 2018).

Atualmente, a avaliação de risco de crédito se baseia em algoritmos de aprendizado supervisionado e não supervisionado, explorando dados tanto estruturados quanto não estruturados. Modelos como o support vector machines (SVM) e multilayer perceptron (MLP) têm sido amplamente aplicados na predição de inadimplência, destacando-se pela capacidade de capturar relações complexas e não lineares nos dados (Dastile et al., 2020). Em contrapartida, logistic regression continua popular devido à sua simplicidade e interpretabilidade, especialmente importante em instituições financeiras que precisam justificar decisões de crédito para reguladores (Florez-Lopez e Ramon-Jeronimo, 2015). Enquanto os métodos estatísticos tendem a ajustar os dados ao modelo, necessitando que os pesquisadores definam as estruturas do modelo (Huang et al., 2004), as técnicas de machine learning extraem automaticamente conhecimento dos dados, gerando modelos complexos que se ajustam melhor a esses dados.

Os avanços em machine learning também impulsionaram o uso de deep learning, como o Multilayer Perceptron (MLP), na análise de crédito. Apesar de sua complexidade e custo computacional, essas redes são capazes de capturar padrões mais profundos e sutis nos dados, aprimorando a detecção de risco de crédito em contextos onde os dados são ricos e variados (Montevechi et al., 2024). Entretanto, um desafio contínuo é a "caixa-preta" que esses modelos representam, dificultando a transparência e a explicação dos resultados para fins regulatórios (Jovanovic et al. 2024). De acordo com Farias e Silva (2023), tanto os modelos econométricos/estatísticos quanto os modelos de machine learning são altamente eficazes, pois eliminam o subjetivismo humano e maximizam os lucros devido à maior assertividade em emprestar para as pessoas que terão maior probabilidade de pagamento. Contudo, os modelos de machine learning têm se destacado, pois conseguem lidar melhor com a não linearidade dos dados do que os modelos estatísticos.

Embora existam diversos métodos estatísticos e de inteligência artificial disponíveis, ainda não há consenso sobre a melhor estratégia de análise financeira. Montevechi et al. (2024) destacam que, apesar dos avanços em técnicas de Machine Learning, desafios como a interpretabilidade dos modelos complexos persistem, indicando que a escolha da abordagem ideal depende do contexto específico e das necessidades da instituição financeira. Segundo Farias e Silva (2023), Addo, Guégan e Hassani (2018), o mercado é dinâmico e evolui a passos largos, e a comparação entre os modelos de risco de crédito baseados em técnicas estatisticas e de machine learning que utilizam o sistema de Credit Scoring é uma evolução natural. Pois, em vista dos diferentes tipos de cenários e variáveis, cada modelo pode se ajustar melhor ou pior a um determinado problema de análise de crédito, sendo fundamental saber quais são os melhores modelos ou aqueles que melhor se adaptam ao contexto desejado, melhorando ainda mais a gestão dos recursos pelas instituições financeiras.

Neste contexto, métodos de aprendizado supervisionado, como o Gradient Boosting, têm se mostrado particularmente eficientes na criação de "credit scores" preditivos para o setor bancário. Esses scores não apenas preveem a inadimplência, mas também ajudam a definir estratégias de marketing e relacionamento com o cliente, otimizando o valor do cliente para a instituição (Drobetz et al., 2021).

Este trabalho se aprofunda na eficiência desses modelos, especialmente em continuação ao estudo "Mensuração da Eficiência de Modelos de Análise de Crédito Segundo uma Ótica Behaviorista" da Universidade de Brasília (UNB) (Bispo 2015), no qual foi explorada uma análise focada em regressão logística. Agora, com mais robustez, pretende-se realizar uma análise focada em modelos de machine learning que serão listados abaixo. O trabalho buscou inspiração nos estudos de Addo, p.; Guégan, d.; Hassani, b. (2018). Credit Risk Analysis Using Machine and Deep Learning Models e no trabalho de Farias, I.; Silva, M. (2023). Ciência de Dados no Mercado de Crédito: Estratégias para Mitigação de Riscos e Otimização de Decisões com Modelagem Preditiva, que exploraram modelos de Deep Learning and machine learning para previsão de inadimplência de empréstimos, oferecendo novas perspectivas e potencializando as ferramentas de análise de crédito. A grande riqueza deste trabalho concentra-se no fato de possuir bases reais de duas das maiores instituições financeiras do Brasil e, principalmente, no estudo das técnicas mais referenciadas e

usadas nas instituições financeiras para prever risco de crédito de pessoa física no Brasil.


3

3 METODOLOGIA

3.1 DADOS

Antes das bases serem usadas para treinar os modelos elas precisam ser tratadas, em vista disso, foi realizado a limpeza das bases de cada Banco que será detalhada nos tópicos seguintes. É importante ressaltar que as bases são observações do comportamento dos clientes ao longo de 12 meses, tanto Banco A quanto Banco B, e ao final do período foi marcado se o cliente entrou em inadimplência ou não, bem como a posição da variável no momento da marcação, por exemplo a variável do banco AChequesDevolvidosSemestre que mede se o cliente teve cheque devolvido mais de 2 vezes ao longo desses 12 meses. Depois das bases limpas e tratadas, conforme a especificidade de cada uma, as bases foram separadas em Base de treino (usada para treinar os modelos) e base de teste (usada para testar o modelo quanto a sua assertividade por meio da técnica de validação cruzada).

3.1.2 BANCO A

Em relação à base de dados do Banco A, foi disponibilizada uma amostra de 147.000 clientes, dos quais 78% eram adimplentes e 22% inadimplentes, coletados ao longo de 12 meses. A base de dados continha 52 variáveis cadastrais, as quais estão descritas e detalhadas no apêndice. A escolha dos clientes, adimplentes e inadimplentes, foi feita aleatoriamente a partir da base de dados cadastrais do banco, definindo-se como inadimplentes aqueles com dívidas vencidas há mais de noventa dias, emitentes de cheques sem fundos ou com restrições em órgãos de proteção ao crédito.

Foram realizadas cinco etapas de pré-processamento de dados, detalhadas a seguir, na "Base do Banco A", para preparar os dados com o objetivo de classificar clientes como bons ou maus pagadores, tendo a variável " IndicadorInadimplencia " como alvo. É importante ressaltar que buscou-se criar uma base que atendesse aos pré-requisitos de todos os modelos, para que eles recebessem bases o mais semelhantes possível, ou seja, todos os cinco modelos que serão testados para o Banco A receberão bases semelhantes, de forma que a comparação dos

resultados dos modelos seja atribuída mais ao modelo em si do que à diferença das bases. Um exemplo, a fim de ilustrar melhor a ideia, é a variável "Sexo" do Banco A, que estava em caractere: "F" para feminino e "M" para masculino. Esses valores foram substituídos por variáveis categóricas: 0 para feminino e 1 para masculino, pois os modelos de support vector machines and logistic regression lidam melhor com a categorização dos campos do que com caracteres.

Na primeira etapa, transformaram-se as variáveis categóricas utilizando codificação one-hot. Segundo Quan e Sun (2024) One-hot encoding é uma técnica comum em ciência de dados para transformar variáveis categóricas em um formato adequado para algoritmos, criando colunas binárias onde “1” indica presença e “0” ausência, especialmente útil quando não há ordem entre categorias. A Tabela 1 apresenta as variáveis transformadas.

Tabela 1 – Variáveis transformadas em categorias

Variáveis Transformadas em Categóricas	
Indicador de default	Indicador de anotação em ser
Natureza da ocupação principal	Nível de instrução
Indica estado civil	Indicador de cheque especial no semestre
Tipo de Benefício	Indicador de conta corrente no semestre
Sexo	

Fonte: Elaborado pelo autor.

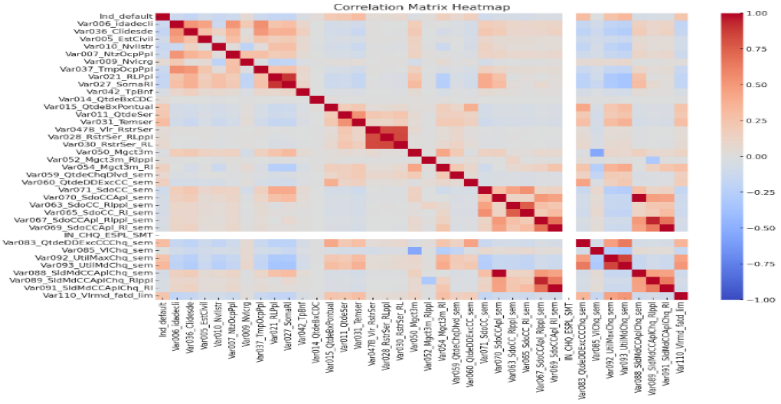
As demais variáveis foram convertidas para o tipo numérico adequado aos modelos. Na Segunda etapa, trataram-se os valores ausentes e outliers. Os valores numéricos faltantes foram substituídos pela mediana de cada variável; tal estratégia foi empregada, pois a mediana não é afetada por valores extremos. Além disso, buscou-se evitar a perda de dados da amostra; entretanto, em alguns casos isso não foi possível, como no caso dos valores nulos nas variáveis categóricas, que precisaram ser removidos. Para lidar com outliers, aplicou-se a winsorização, limitando-os aos 1° e 99° percentis. A escolha pelo método de winsorização foi fundamentada nos trabalhos de Martin et al (2024), que analisaram a winsorização em comparação com outras técnicas de controle de outliers, destacando sua eficiência em

suavizar valores extremos sem descartá-los, o que preserva a integridade dos dados para modelos preditivos.

Os dados numéricos foram normalizados utilizando a técnica de escalonamento Min-Max, trazendo-os para a mesma escala. Vale ressaltar que o método Min-Max foi escolhido porque reduz cada valor para um intervalo entre 0 e 1, com base nos valores mínimo e máximo originais, sendo útil quando se deseja manter a forma da distribuição original e os outliers são relevantes. Segundo Farias e Silva (2023), a normalização por esse método garante que variáveis em diferentes escalas contribuam igualmente para a análise. Apesar da aparente contradição de, nos outliers, retirarmos os valores extremos e, na normalização, tentarmos manter o máximo número de dados, tal estratégia foi proposital, pois o banco A é um banco internacional e muito grande, o que faz com que os 1% mais ricos da base tenham valores tão extremos quando comparados aos demais da base.

Na terceira etapa, realizou-se uma análise de correlação para identificar e remover variáveis com alta correlação. Foi calculada a matriz de correlação entre as variáveis, identificando os pares com correlação acima de um limiar estabelecido (0,9). A Figura 1 apresenta a correlação entre as variáveis.

Figura 1 – Correlação entre as variáveis



Fonte: Elaborado pelo autor.

As variáveis que apresentavam alta correlação foram removidas para reduzir a multicolinearidade no modelo. Na quarta etapa, abordou-se o balanceamento de classes aplicando técnicas de oversampling. Esta técnica foi escolhida com base nos trabalhos dos autores Steve, Olusegun e Paul (2024). Segundo esses autores, o oversampling é uma estratégia eficaz para aumentar a representatividade da classe minoritária, permitindo ao modelo de

machine learning uma melhor performance ao lidar com o desbalanceamento de classe, que é justamente o caso deste trabalho, pois identificou-se um desbalanceamento entre as classes da variável alvo indicador, em que 70% eram adimplentes e 30% inadimplentes. Desta forma, o método aumentou a representação da classe minoritária, evitando a perda de dados que ocorreria com a subamostragem da classe majoritária.

Por fim, na quinta etapa, prepararam-se os dados para o modelo. Foram separadas as features (variáveis independentes) e a variável alvo (IndicadorInadimplencia), dividindo o conjunto de dados em conjuntos de treino e teste para a validação do modelo. As estatísticas descritivas das variáveis escolhidas do Banco A estão na Tabela 2.

Tabela 2 – Estatísticas descritivas

Variável	Média	Desvio Padrão	Mínimo	Máximo	Tipo de Variável
Atividade econômica do empregador	721,24	219,94	0	1046	Numérica
Cliente desde (em anos)	10,15	7,96	1,1	86,7	Numérica
Idade do cliente	46,57	16,46	18,1	115	Numérica
Indicador de anotação no Serasa	0,18	0,38	0	1	Categórica
Indicador de anotação no Serasa	0,18	0,38	0	1	Categórica
Indicador de aplicação no semestre	1	0	1	1	Categórica
Indicador de cheque especial no semestre	0,87	0,34	0	1	Categórica
Indicador de inadimplência	0,1	0,29	0	1	Categórica
Média da margem de contribuição	100,75	210,33	-290,23	18736,24	Numérica
Natureza da ocupação principal	5,07	3,52	1	77	Categórica
Nível de instrução	3,6	2,01	0	9	Categórica

Quantidade de anotações baixadas por CDC	0	0,02	0	4	Numérica
Anotações Baixadas exceto Serasa	0,18	0,74	0	27	Numérica
Quantidade de restrições no Serasa	0,63	2,57	0	128	Numérica
Renda líquida principal	3017,78	134946,02	0	51731884,44	Numérica
Saldo médio mais aplicações menos utilizações em cheque especial dividido pela renda líquida principal	1	0	1	1	Numérica
Tempo de ocupação principal	12,5	11,19	0,1	88,5	Numérica
Tipo de benefício	2,01	0,84	1	7	Categórica
Utilização máxima de cheque especial no semestre	1,16	4,88	0	800,48	Numérica
Valor de restrição no Serasa	2660,58	509711,04	0	195382082,7	Numérica
Sexo	0,55	0,5	0	1	Categórica

Fonte: Elaborado pelo autor

3.1.3 BANCO B

Para o Banco B, utilizou-se a mesma estratégia de cinco etapas. É importante ressaltar que, para o Banco B, o critério para definir clientes inadimplentes foi o mesmo utilizado para o Banco A; contudo, a variável target aqui é o Tipo_cliente. Quanto à temporalidade, a base também foi extraída conforme o padrão do Banco A, para que as bases sejam similares entre si nesse aspecto. A base do Banco B é composta por 36 variáveis cadastrais, vale ressaltar também que diferente do Banco A a base do Banco B já veio dividida em 50% dos dados para adimplentes e 50% dos dados sendo inadimplentes.

A base do Banco B também seguiu o critério de marcação dos dados do Banco A, em que a variável recebeu o valor após 12 meses de observação. Um exemplo disso é o Tipo_cliente, que só após 12 meses foi marcado como adimplente ou inadimplente. É importante ressaltar que cada banco, de acordo com sua realidade, escolhe as variáveis a serem utilizadas. Buscou-se criar uma base que atendesse aos pré-

requisitos de todos os modelos, para que estes recebessem bases o mais semelhantes possível, assim como foi feito para o Banco A.

Na primeira etapa, procedeu-se à transformação de variáveis categóricas utilizando codificação one-hot. As variáveis convertidas foram:

Tabela 3 – Conversão de Variáveis Transformadas em Categóricas
Transformação em categoria
Conta salário na instituição
Restrição resolvida ou não na Serasa nos últimos 3 anos
Escolaridade
Sexo
Tipo de residência

Fonte: Elaborado pelo autor.

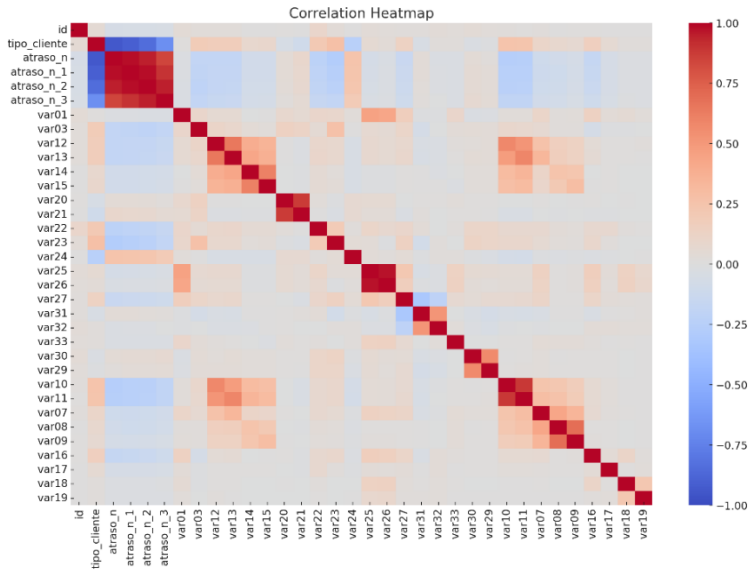
Posteriormente, diversas variáveis foram transformadas para o tipo numérico, visando adequá-las aos modelos. Na segunda etapa, trataram-se os valores ausentes e outliers. Os valores numéricos faltantes foram substituídos pela mediana de cada variável, enquanto as linhas com valores nulos nas variáveis categóricas foram removidas. Em seguida, aplicou-se a winsorização para limitar os outliers aos percentis de 1% e 99%, reduzindo o impacto de valores extremos. Os dados numéricos foram normalizados utilizando a técnica de escalonamento Min-Max, assegurando que todas as variáveis estivessem na mesma escala.

Na terceira etapa, realizou-se uma análise de correlação para identificar e remover variáveis altamente correlacionadas. Calculou-se a matriz de correlação entre as variáveis numéricas, identificando pares com correlação acima de 0,9. As variáveis que apresentavam alta correlação foram removidas para reduzir a multicolinearidade no modelo.

Os modelos de aprendizado de máquina considerados incluem artificial Deep Learning, Support vector machines (máquinas de vetores de suporte - SVM), gradient boosting (Gradient Boosting), Decision trees (Arvores de decisão), Logistic Regression (Regressão Logística) e

neural networks (redes neurais). Cada um desses modelos foi treinado e validado para avaliar o desempenho na previsão do problema proposto.

Figura 2 – Correlação entre as variáveis



Fonte: Elaborado pelo autor.

Na quarta etapa, abordou-se o balanceamento de classes, aplicando-se técnicas de reamostragem da classe minoritária. Identificou-se o desbalanceamento entre as classes da variável alvo "TipoCliente" e aumentou-se a representação da classe minoritária por meio de técnicas de Oversampling, evitando a perda de dados que ocorreria com a Subamostragem da classe majoritária. Essa estratégia foi importante, pois, apesar de a base estar, a princípio, balanceada, no processo de limpeza da base houve perda de informações, deixando a base com 50% de adimplentes e 45% de inadimplentes.

Por fim, na quinta etapa, prepararam-se os dados para o modelo. Separaram-se as variáveis independentes (features) e a variável alvo (TipoCliente), dividindo o conjunto de dados em conjuntos de treino e teste. As variáveis escolhidas, após a limpeza dos dados, para compor a base que treinará os modelos do Banco B estão listadas abaixo:

Tabela 4 – Estatísticas descritivas					
Variável	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**	**Tipo de Variável**
Classificação do cliente segundo definição	0,5	0,5	0	1	Categorica
Quantidade de dias em atraso no contrato de Crédito Direto ao Consumidor	71	69,09	0	210	Numérica
Percentual de utilização do crédito rotativo	30,36	40,57	-1	1.584,00	Numérica
Multa no cartão de crédito	1,67	16,64	-1	1.308,30	Numérica
Percentual de utilização do cartão de crédito	11,96	31,71	-1	830	Numérica
Percentual de parcelamento da fatura (var12)	8,17	24,07	0	100	Numérica
Saldo na conta corrente	179,22	1.048,06	-1	66.029,83	Numérica
Saldo na poupança	343,94	3.824,09	-1	294.767,09	Numérica
Valor de aplicações	71,77	3.400,00	0	539.292,50	Numérica
Cheques sem fundo (motivo 11)	1,19	5,08	-1	252	Numérica
Cheques sem fundo (motivo 12)	0,53	2,65	-1	117	Numérica
Conta salário na instituição	0,35	0,48	0	1	Categorica
Restrição resolvida ou não na SERASA	0,4	0,49	0	1	Categorica
Renda líquida	2.271,11	4.163,42	2	360.000,00	Numérica
Escolaridade	4,04	1,53	1	10	Categorica
Sexo	0,59	0,49	0	1	Categorica
Tipo de residência	1,47	1,16	1	7	Categorica

Fonte: Elaborado pelo autor.

Dentro desse contexto, exploraram-se mais detalhadamente modelos analíticos como Logistic Regression, Decision Tree, Gradient Boosting, Support Vector Machines, e Deep Learning. Esses modelos são fundamentais não só pela sua capacidade de processar grandes

volumes de dados, mas também por oferecerem insights mais precisos e detalhados sobre o comportamento financeiro dos clientes, abrindo caminho para discussões mais aprofundadas sobre suas aplicações práticas e teóricas no ambiente financeiro contemporâneo, no cenário de análise de crédito.

3.2 LOGISTIC REGRESSION

Segundo Dastile et al., (2020) a logistic regression é um método estatístico amplamente utilizado para avaliar a capacidade de crédito dos mutuários devido à sua simplicidade e transparência nas previsões. Ela permite estimar a probabilidade de um evento ocorrer, como a inadimplência de um empréstimo, com base em variáveis explanatórias que podem ser tanto contínuas quanto categóricas. Essas variáveis incluem, por exemplo, características do solicitante como renda, idade e histórico de crédito. A logistic regression é capaz de modelar a relação não linear entre essas variáveis e o resultado, garantindo que as previsões estejam limitadas entre 0 e 1, utilizando uma função sigmoide que cria uma curva em formato de "S" (Dastile et al., 2020).

Segundo Corrar (2007), a logistic regression se caracteriza como uma técnica estatística diferente dos outros modelos, pois permite estimar a probabilidade de ocorrência de determinado evento em face a um conjunto de variáveis explanatórias, ou seja, para retornar valores em forma de probabilidade, o modelo se vale das variáveis explanatórias Xi (renda, sexo e restrições, por exemplo) como variáveis de entrada do modelo para o cálculo da probabilidade. De acordo com Chopra e Bhilare (2018), esse modelo foi tradicionalmente utilizado em análises de crédito para distinguir entre bons e maus pagadores, fornecendo uma probabilidade de inadimplência que auxilia no processo de tomada de decisão dos bancos. A principal vantagem da logistic regression reside na sua capacidade de gerar resultados que são interpretáveis pelos gestores, o que é crucial para justificar a concessão ou negação de crédito, cumprindo obrigações de supervisão bancária em diversos países (Chopra e Bhilare, 2018).

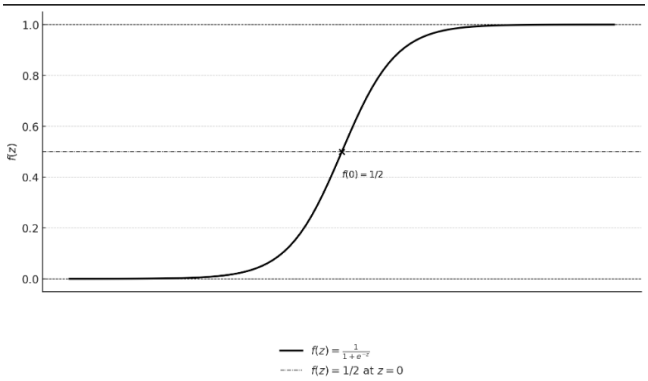
Hosmer e Lemeshow (1989) apontam que a técnica de logistic regression segundo uma ótica de análise de crédito caracteriza-se por descrever a relação entre várias variáveis explanatórias independentes Xi (sendo elas discretas ou contínuas) e uma variável dependente dicotômica f(Z), representando a presença de inadimplência (1) ou

ausência de inadimplência (0). Além disso, Wang et al. (2018) destacam que, embora modelos mais sofisticados de machine learning tenham sido introduzidos no campo do credit scoring, a logistic regression ainda é preferida em muitos casos devido à sua facilidade de implementação e interpretabilidade dos resultados. Matematicamente ela pode ser descrita da seguinte forma:

$$p = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^{k} \beta_i X_i)}} \quad (1)$$

onde a probabilidade do evento ocorrer, representada por p é modelada considerando diferentes componentes. O termo α refere-se ao intercepto ou constante do modelo, enquanto os β_i representam os coeficientes das variáveis independentes Xi. A expressão $\sum_{i=1}^{k} \beta_i X_i$ descreve a soma ponderada das variáveis independentes X, cada uma multiplicada por seu respectivo coeficienteβ_i , contribuindo para a estimativa da probabilidade.

Figura 4 – Gráfico regressão Logística



Fonte: Elaborado pelo autor.

3.3 DECISION TREE

As Decision Trees são amplamente utilizadas para a avaliação de risco de crédito devido à sua robustez e transparência. Esses modelos permitem criar regras claras e compreensíveis, tornando-se essenciais na análise de crédito. Desenvolvidas inicialmente por Quinlan (1986), as Decision Trees dividem os dados em subconjuntos homogêneos por meio de decisões sequenciais, facilitando a interpretação e a comunicação com reguladores financeiros, conforme discutido por Montevechi et al. (2024) e Zhang & Yu (2023).

No contexto de risco de crédito, as Decision Trees categorizam clientes em grupos de maior ou menor risco, baseando-se em atributos

como renda, histórico de empréstimos e relacionamento com a instituição financeira. Segundo Dastile et al. (2020), essa segmentação facilita a identificação de padrões que indicam a probabilidade de inadimplência, tornando o processo decisório mais eficaz e transparente.

O Information Gain é uma métrica importante para a construção das árvores, pois avalia a capacidade de um atributo em reduzir a incerteza nos dados, destacando os pontos de divisão mais informativos (Quinlan, 1986). Matematicamente, o Information Gain pode ser expresso como a diferença entre a entropia do conjunto de dados antes e depois da divisão, ou seja:

$$Ganho\_de\_Informação(A) = Entropia(S) - \sum_{i=1}^{n} \left(\frac{|S_i|}{|S|} Entropia(S_i) \right) \quad (2)$$

onde Entropia(S) representa a medida de incerteza do conjunto de dados |S| e |Si| são os subconjuntos resultantes da divisão com base no atributo A. Quanto maior o ganho de informação, maior será a redução na incerteza, tornando o atributo A mais adequado para a divisão dos dados.

Por outro lado, o Índice Gini, proposto por Breiman et al. (1984), quantifica a impureza de um nó em uma Árvore de Decisão, medindo o grau de mistura das classes. Quanto menor o valor do Índice Gini, mais homogêneo é o grupo formado, tornando o modelo mais eficiente na separação dos solicitantes de crédito em grupos de risco. Matematicamente, o Índice Gini pode ser expresso como:

$$Gini(S) = 1 - \sum_{x \in X} p(x)^2 \quad (3)$$

onde p(x) é a proporção de cada classe no conjunto S. Essa fórmula calcula a probabilidade de uma amostra ser incorretamente classificada se for rotulada aleatoriamente de acordo com a distribuição das classes. Estudos como o de Dastile et al. (2020) demonstram que o Índice Gini é mais eficaz para modelos de classificação de risco de crédito, especialmente em situações de desbalanceamento de classes, comuns no contexto de análise de crédito. Segundo Zhang et al. (2024), o Índice Gini também apresenta melhor desempenho ao ser aplicado em conjunto com métodos como o Random Forest, garantindo maior

precisão e robustez na classificação de inadimplentes. Além disso, Montevechi et al. (2024) ressaltam que o Índice Gini permite uma identificação mais consistente dos atributos relevantes, o que é essencial para a eficácia dos modelos de crédito.

Modelos como o CART (Classification and Regression Tree) e Florestas Aleatórias (Random Forest) são amplamente utilizados. Para o trabalho em questão, foi utilizado o modelo de Random Forest, que combina diversas Árvores de Decisão para formar um modelo mais robusto. Matematicamente, o Random Forest pode ser descrito como um conjunto de NÁrvores de Decisão, onde cada árvore Ti é treinada em um subconjunto de dados amostrado aleatoriamente com reposição (técnica de bootstrap). Para realizar a classificação, cada árvore gera uma predição $hi(x)$, e a predição final é obtida pela votação majoritária das N árvores:

$$H(x) = majority\_vote\{h_i(x)\}_{i=1}^{N} \quad (4)$$

onde *H(x)* é o resultado da classificação do modelo de Random Forest. Essa abordagem permite reduzir o viés e a variância, tornando o modelo mais eficaz para problemas complexos, como a classificação de inadimplentes, destacando-se por sua alta acurácia em problemas de classificação. Segundo Dastile et al. (2020), modelos de Random Forest superam os modelos simples de Árvores de Decisão e outros classificadores em termos de precisão ao lidar com grandes volumes de dados desbalanceados, o que é comum em problemas de análise de crédito. A combinação de diversas Árvores de Decisão em uma única estrutura permite ao Random Forest aumentar a robustez e reduzir o risco de overfitting (quando o modelo se ajusta excessivamente aos dados de treinamento, perdendo generalidade), conforme também observado por Zhang et al. (2024).

As Árvores de Decisão são preferidas no credit scoring devido à sua transparência e facilidade de validação, características essenciais para cumprir as regulamentações financeiras. Segundo Hand e Henley (1997), elas permitem que analistas e reguladores compreendam claramente as razões por trás das previsões do modelo, o que é fundamental para justificar decisões de crédito. Zhang et al. (2024) reforçam que, além da capacidade preditiva, os modelos devem ser compreensíveis para gestores e reguladores.

Para evitar o overfitting as Decision Trees podem ser podadas. Esposito et al. (1997) discutem estratégias de poda que eliminam ramos

com pouca contribuição ao poder preditivo, melhorando a generalização do modelo. Montevechi et al. (2024) também destacam a importância da poda para evitar sobreajuste e melhorar a capacidade de generalização do modelo. Essa adaptabilidade, aliada à possibilidade de interpretação transparente, solidifica a posição das Decision Trees como uma das principais ferramentas na análise de risco de crédito moderna.

3.4 GRADIENT BOOSTING.

O O Modelo de Aumento de Gradiente, conhecido como Gradient Boosting, foi originalmente introduzido por Friedman (2001) e continua sendo amplamente utilizado em uma variedade de problemas preditivos complexos, incluindo a análise de risco de crédito. Esse modelo é valorizado por sua capacidade de combinar previsões incrementais de modelos simples, geralmente árvores de decisão, de forma que os erros dos modelos anteriores sejam continuamente minimizados (Dastile et al., 2020).

Entre os modelos mais utilizados para risco de crédito, destacam-se tanto métodos estatísticos tradicionais quanto aqueles baseados em aprendizado de máquina. A regressão logística, apesar de sua simplicidade e ampla adoção, muitas vezes é superada em termos de desempenho por modelos mais avançados, como Gradient Boosting e XGBoost, que possuem uma maior capacidade de captura de padrões complexos (Xia et al., 2020). A seleção de parâmetros no Gradient Boosting, como a taxa de aprendizado e o número de árvores, é fundamental para evitar o overfitting e garantir um bom desempenho em novos dados. Estudos sugerem que abordagens híbridas e de ensemble, que combinam a robustez de modelos complexos com a simplicidade de métodos mais tradicionais, são o futuro da modelagem de risco de crédito (Zhang & Yu, 2024).

No contexto de credit scoring, o Gradient Boosting se destaca por identificar os fatores que contribuem para o risco de crédito de um indivíduo, corrigindo iterativamente os erros dos modelos anteriores e tornando a previsão mais robusta. A técnica constrói um modelo aditivo de forma sequencial: a cada passo, uma nova árvore é adicionada ao modelo, buscando minimizar os resíduos dos modelos anteriores. Esse processo de ajuste iterativo é conhecido como descida de gradiente e tem se mostrado altamente eficaz em lidar com a natureza complexa

dos dados financeiros, frequentemente caracterizados por interações não lineares e padrões complexos (Zhang et al., 2024).

Matematicamente, o Gradient Boosting funciona da seguinte maneira: em cada etapa, o modelo é atualizado adicionando um termo, onde é uma árvore de decisão ajustada aos resíduos do modelo anterior, e é a taxa de aprendizado que controla a contribuição de cada árvore no modelo final. A equação que representa essa atualização é:

$$[F_{t+1}(x) = F_t(x) + \gamma_t h_t(x)] \quad (5)$$

Essa abordagem é eficaz para minimizar diferentes tipos de funções de perda, como o erro quadrático para regressão e a perda logarítmica para classificação. Estudos recentes mostram que técnicas como XGBoost e Gradient Boosting têm obtido melhor desempenho em termos de métricas como AUC e precisão, em comparação com modelos mais simples, como a regressão logística (Montevechi et al., 2024).

Os modelos de Gradient Boosting são aplicados para prever a probabilidade de inadimplência, classificando os clientes em grupos de risco. A flexibilidade em ajustar árvores de decisão para lidar com resíduos específicos em cada iteração torna o Gradient Boosting particularmente robusto em cenários financeiros, o que explica sua popularidade crescente entre instituições financeiras que buscam não apenas precisão, mas também interpretabilidade em suas previsões de risco de crédito (Suhadolnik et al., 2023).

3.5 SUPPORT VECTOR MACHINES

O Support Vector Machine (SVM) é um método de machine learning amplamente utilizado para classificação e regressão, incluindo aplicações na análise de risco de crédito. A SVM funciona identificando um hiperplano ótimo que maximiza a margem de separação entre diferentes classes de dados. Em outras palavras, ela estabelece uma fronteira clara entre as classes, tornando-a eficaz para distinguir entre clientes de alto e baixo risco de crédito. No contexto do risco de crédito, isso significa separar grupos como inadimplentes e não inadimplentes com base em atributos individuais dos clientes (Montevechi et al., 2024; Suhadolnik et al., 2023).

O hiperplano é uma superfície de decisão que separa os pontos de diferentes classes no espaço de atributos. A SVM busca construir

esse hiperplano de forma a maximizar a margem, ou seja, a distância entre o hiperplano e os pontos de dados mais próximos de ambas as classes, conhecidos como vetores de suporte. Dessa forma, a SVM garante uma separação eficiente entre os grupos, aumentando a precisão do modelo.

No entanto, em muitos casos práticos, os dados não são perfeitamente separáveis linearmente. Para lidar com isso, a SVM introduz o parâmetro de regularização C, que controla o equilíbrio entre maximizar a margem e minimizar o erro de classificação. O parâmetro C penaliza erros de classificação no conjunto de treinamento: valores altos de C dão maior penalidade aos erros, levando a uma margem menor e possível overfitting; valores baixos de C permitem uma margem maior e mais erros, podendo resultar em underfitting (acontece quando o modelo de machine learning não consegue capturar adequadamente os padrões nos dados). Assim, C influencia diretamente a flexibilidade do modelo e sua capacidade de generalização.

$$[\min\left(\frac{1}{2}|\omega|^2 + C\sum\xi_i\right) \quad \text{sujeito a} \quad y_i(\omega \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0] \qquad (6)$$

Em que ω é o vetor normal ao hiperplano, x_i são os vetores de entrada, y_i são as etiquetas de classe dos vetores de entrada, e b é o termo de viés do hiperplano e ξ_i são as variáveis de folga que permitem a violação da margem.

Essa separação é essencial para garantir que os grupos (como inadimplentes e não inadimplentes) sejam classificados de forma eficiente, aumentando a precisão do modelo.

Para lidar com problemas de classificação não linear, a Support Vector Machine (SVM) utiliza o truque do kernel, uma técnica que transforma o espaço de entrada em um espaço de maior dimensão, onde a separação linear pode ser realizada de forma mais eficaz. Esse mapeamento é feito sem a necessidade de calcular explicitamente a transformação, o que torna o processo mais eficiente. O truque do kernel é especialmente relevante na análise de risco de crédito, pois os dados de clientes geralmente apresentam relações complexas e não lineares, que são melhor capturadas ao transformar o espaço de representação dos dados (Addo et al., 2018; Suhadolnik et al., 2023).

Existem diferentes tipos de kernels que podem ser usados em SVM, como o kernel linear, polinomial e o radial basis function (RBF), sendo este último um dos mais comuns devido à sua capacidade de lidar com relações complexas entre variáveis. A escolha do kernel depende da natureza dos dados e da complexidade do problema a ser resolvido.

Segundo Suhadolnik et al. (2023), na literatura, o kernel mais frequentemente utilizado para problemas de risco de crédito é o kernel RBF, devido à sua flexibilidade e à capacidade de capturar relações não lineares de maneira eficiente. O kernel RBF tem mostrado resultados consistentes em estudos empíricos, destacando-se pela sua eficácia na previsão de inadimplência, e é por isso que o kernel RBF será utilizado neste trabalho.

Para exemplificar melhor matematicamente, o kernel RBF pode ser definido da seguinte forma na figura abaixo.

$$K(x, x') = \exp(-\gamma \|x - x'\|^2) \quad (7)$$

Em que x e x' são vetores de características, já o parâmetro γ (gama) é crucial no kernel RBF, pois define o alcance da influência de um único ponto de dado. Valores altos de γ resultam em uma influência próxima, tornando o modelo mais complexo e potencialmente sujeito a overfitting. Valores baixos de γ expandem a influência dos pontos, levando a uma fronteira de decisão mais suave e generalista, o que pode causar underfitting. Assim, γ afeta diretamente a complexidade da fronteira de decisão e a capacidade do modelo em capturar padrões não lineares nos dados.

A escolha do valor de γ afeta a complexidade do modelo: valores altos de γ resultam em maior complexidade, enquanto valores baixos geram uma superfície de decisão mais suave, em outras palavras, a fronteira de decisão criada é menos complexa e tende a ser mais linear.

3.6 DEEP LEARNING O MULTILAYER PERCEPTRON (MLP)

O Multilayer Perceptron (MLP) é um modelo de artificial neural networks que consiste em a feedforward neural network composta por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída, onde os neurônios são interconectados através de pesos sinápticos. O conceito de MLP é uma extensão do perceptron simples, introduzido por Rosenblatt em 1958 e posteriormente aprimorado para

lidar com problemas de não linearidade (Zhang et al., 2024). O MLP tornou-se particularmente eficaz na modelagem de relações complexas entre variáveis devido à sua capacidade de aprender representações profundas e hierárquicas dos dados, sendo amplamente utilizado em aplicações financeiras, como a avaliação de risco de crédito (Addo et al., 2018).

No contexto da avaliação de risco de crédito, o MLP é utilizado para prever a probabilidade de inadimplência de clientes, um desafio relevante para instituições financeiras que buscam otimizar a alocação de recursos e minimizar prejuízos decorrentes de não pagamento. Segundo Zhang et al. (2024), a capacidade do MLP de aprender padrões complexos o torna adequado para tarefas de classificação, como o credit scoring, onde os clientes são categorizados como adimplentes ou inadimplentes com base em variáveis financeiras e de comportamento. O aprendizado ocorre através do ajuste dos pesos da rede, utilizando algoritmos de retropropagação para minimizar o erro entre as previsões e os valores reais (Montevechi et al., 2024).

A aplicação do MLP em risco de crédito segue um processo que envolve etapas como o pré-processamento dos dados, seleção de variáveis e treinamento do modelo. A escolha dos hiperparâmetros é um fator essencial para o bom desempenho do MLP. De acordo com Addo et al. (2018), essa escolha é feita através de uma abordagem de otimização, que inclui técnicas como a grade de hiperparâmetros e o uso de critérios de parada antecipada para evitar o excesso de treinamento e garantir melhor generalização dos resultados em novos dados. Os hiperparâmetros incluem a taxa de desistência, funções de ativação, funções de regularização L1 e L2, e o número de camadas ocultas e neurônios em cada camada (Montevechi et al., 2024).

No MLP, a arquitetura feedforward é fundamental para sua estrutura e desempenho. O termo "feedforward" refere-se à arquitetura do MLP, onde as informações fluem em uma única direção, da camada de entrada até a camada de saída, sem ciclos ou retroalimentação. Essa característica permite que o MLP realize transformações lineares e não lineares sobre os dados de entrada, capturando padrões complexos úteis na avaliação de risco de crédito (Addo et al., 2018). De acordo com Montevechi et al. (2024), o uso de funções de ativação como a ReLU (Rectified Linear Unit) e a Sigmoid contribui para que o MLP consiga mapear relações não lineares, garantindo uma maior precisão nas previsões.

Além disso, é comum utilizar penalidades de regularização L1 e L2 para evitar o sobreajuste. A regularização L1 (ou Lasso) incentiva a rede a aprender pesos esparsos, eliminando conexões irrelevantes e tornando o modelo mais simples e interpretável (Zhang et al., 2024). Já a regularização L2 (ou Ridge) favorece pesos menores em geral, ajudando a manter todas as conexões, mas com valores reduzidos, o que evita que algum peso domine o processo de aprendizado (Chopra e Bhilare, 2018). Essas técnicas de regularização são fundamentais para promover um ajuste mais balanceado e uma maior capacidade de generalização do modelo.

Segundo Addo et al. (2018), os modelos de redes neurais artificiais, incluindo o MLP, são eficazes na análise de risco de crédito devido à sua habilidade de lidar com grandes volumes de dados e variáveis complexas. O cálculo do risco de crédito por meio do MLP envolve a atribuição de pesos às conexões entre os neurônios, ajustados durante o treinamento para minimizar a função de perda. Segundo Suhadolnik et al. (2023), o uso de técnicas de aprendizado supervisionado permite que o modelo aprenda a partir de exemplos rotulados, em outras palavras, o aprendizado supervisionado, neste trabalho, ocorre a partir das variáveis alvo que classificam o cliente como adimplente ou inadimplente desenvolvendo a capacidade de prever a probabilidade de inadimplência com alta precisão. Esse processo de aprendizado supervisionado é essencial para garantir que o MLP consiga identificar padrões que correlacionem características dos clientes com o risco de inadimplência.

O estudo de Addo et al. (2018) aponta que os modelos de MLP apresentam bom desempenho em tarefas de classificação de risco de crédito. Segundo Addo et al. (2018), a validação cruzada e o uso de métricas como a AUC (Área Sob a Curva ROC) são fundamentais para garantir que o modelo esteja funcionando conforme esperado e que suas previsões sejam confiáveis. Em vista das análises propostas por Addo et al. (2018), este trabalho seguiu a mesma abordagem para treinar e validar os modelos, testando diferentes configurações, como variações no número de neurônios e ajustes nos parâmetros L1 e L2, até identificar o modelo de melhor desempenho, cujos resultados serão apresentados na seção de resultados.

No contexto de um MLP com três camadas ocultas, as fórmulas para a camada de entrada, a segunda camada oculta (d3) e a camada de saída são as seguintes:

- Camada de Entrada:

- Para o banco A (21 variáveis de entrada): $(x_i = \text{input}_i)(i = 1,2,\ldots,21)$

- Para o banco B (18 variáveis de entrada): $(x_i = \text{input}_i)(i = 1,2,\ldots,18)$

onde:

(x_i) é o valor de entrada do i-ésimo neurônio da camada de entrada. (input_i) é o valor da i-ésima variável de entrada (característica do cliente ou do empréstimo).

- Segunda Camada Oculta (d3):

$$[z_k = \text{ReLU}\left(\sum_j v_{kj} \cdot h_j + c_k\right)] \quad (8)$$

onde: (z_k) é o valor de saída do k-ésimo neurônio da camada d3. $(\text{ReLU}(x) = \max(0, x))$ é a função de ativação ReLU. (h_j) é o valor de saída do j-ésimo neurônio da camada anterior (d2). (v_{kj}) é o peso sináptico entre o j-ésimo neurônio da camada d2 e o k-ésimo neurônio da camada d3. (c_k) é o termo de bias do k-ésimo neurônio da camada d3.

- Camada de Saída:

$$[y = \text{sigmoid}\left(\sum_k u_k \cdot z_k + d\right)] \quad (9)$$

Onde: (y) é a saída do modelo, representando a probabilidade de inadimplência. A $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ é a função de ativação sigmoid, que garante que a saída esteja entre 0 e 1. Z_k é o valor de saída do k-ésimo neurônio da camada d3. U_ké o peso sináptico entre o k-ésimo neurônio da camada d3 e o neurônio de saída. $Ja\ o\ d$ é o termo de bias do neurônio de saída. É importante ressaltar o bias funciona de maneira similar ao intercepto em uma equação de regressão linear, permitindo que o modelo se desloque no espaço de busca e não fique restrito a passar pela origem (ou seja, quando todas as entradas são zero, a saída não precisa ser zero).

Apesar da vasta literatura existente, não há um padrão ouro para o treinamento de modelos de crédito baseados em redes neurais. No

entanto, é crucial considerar os riscos de overfitting e perda de acurácia, buscando um equilíbrio entre a complexidade do modelo e seu desempenho em dados não vistos. Para garantir a eficácia e eficiência do modelo proposto, realizamos uma série de testes, detalhados nas seções seguintes, que visam avaliar sua capacidade de generalização e robustez frente a diferentes cenários (Zhang et al., 2024).


4

4 RESULTADOS

A eficiência dos modelos empregados é fundamental para assegurar a previsão adequada da inadimplência em cenários de crédito. Para isso, é essencial abordar e mitigar problemas como a multicolinearidade. Os modelos preditivos utilizados incluem: Decision Tree (Random Forest), Gradient Boosting Model, Support Vector Machine (SVM) e Multilayer Perceptron (MLP), cada um com suas características específicas, que foram ajustadas para garantir melhor desempenho e evitar problemas como o overfitting.

Ao explorar os parâmetros dos modelos, a Random Forest utilizou múltiplas árvores de decisão para aumentar a precisão e controlar o overfitting. Para garantir o desempenho, foram ajustados o número de árvores e a profundidade máxima, visando um balanceamento entre viés e variância do modelo. O ajuste do número de árvores e da profundidade máxima das árvores foram fatores cruciais para o desempenho. No Banco A, a acurácia foi de 79,38%, enquanto no Banco B foi de 86,32%, demonstrando que o ajuste dos hiperparâmetros, como o número de estimadores, contribuiu para uma melhora significativa na performance.

No Modelo de Aumento de Gradiente, a correção progressiva dos erros dos modelos anteriores, minimizando as perdas por meio de técnicas de descida de gradiente, foi um ponto essencial. Para garantir o desempenho, foram ajustados a taxa de aprendizado e o número de estimadores, buscando otimizar a convergência e evitar o overfitting. O ajuste da taxa de aprendizado e do número de estimadores permitiu atingir uma acurácia de 79,74% no Banco A e 87,54% no Banco B. A melhoria progressiva da precisão ao ajustar esses parâmetros demonstra a robustez do modelo, embora ele possa demandar um tempo maior de treinamento devido à complexidade dos ajustes.

Para as Support Vector Machines (SVM), foram utilizados o kernel radial (RBF), o parâmetro de regularização C e o parâmetro gama. A escolha do kernel radial permitiu lidar com a não linearidade dos dados, enquanto o ajuste dos parâmetros C e gama visou garantir um equilíbrio entre a margem de separação e o erro de classificação, aumentando a capacidade do modelo de generalizar em diferentes

conjuntos de dados. Esses parâmetros influenciam diretamente a flexibilidade e a capacidade do modelo em separar classes de forma eficaz. No Banco A, a acurácia alcançada foi de 78,58%, enquanto no Banco B foi de 76,91%, indicando que o ajuste adequado dos parâmetros foi essencial para melhorar a capacidade de generalização do modelo em diferentes conjuntos de dados. Esses resultados reforçam a importância de uma calibração cuidadosa dos hiperparâmetros para maximizar o desempenho do SVM.

No Deep Learning model, implementou-se um MLP com três camadas ocultas, sendo que as camadas de entrada do Banco A continham 54 variáveis e as camadas de entrada do Banco B continham 49 variáveis, utilizando a ReLU activation function. Além disso, aplicaram-se L1 and L2 regularization techniques durante o treinamento para evitar overfitting e melhorar a capacidade de generalização do modelo. O MLP apresentou a maior acurácia nos dois bancos, sendo 84,45% para o Banco A e 94,00% para o Banco B, demonstrando uma excelente capacidade de modelagem de relações complexas entre as variáveis de entrada, embora com um custo computacional mais elevado. Esses resultados evidenciam que o MLP é altamente eficaz na captura de padrões não lineares, proporcionando um desempenho superior na classificação de inadimplência.

A Logistic Regression foi utilizada como um modelo de referência, por ser um método simples e amplamente aceito. No entanto, sua acurácia foi de 79% no Banco A e 77,96% no Banco B. Apesar de sua interpretação ser simples e direta, a Regressão Logística apresentou limitações para capturar a complexidade dos dados, especialmente em comparação com os modelos mais avançados de machine learning.

Tabela 5 – Resultados do Banco A

Modelo	Acurácia	Precisão	Recall	F1-Score	AUC-ROC	Valid. Cruzada
Floresta Aleatória	79.38%	78.66%	80.32%	79.49%	87.16%	86.65%
Aumento de Gradiente	79.74%	79.15%	80.44%	79.79%	87.90%	87.40%
Máquina de Suporte Vetorial	78.58%	78.23%	78.88%	78.55%	85.00%	85.56%

Deep Learning	84,45%	81,03%	91,01%	85.00%	85,00%	86,55
Regressão logística	79%	79%	79%	79%	79%	78%

Fonte: Elaborada pelo autor.

Tabela 6 – Resultados do Banco B

Modelo	Acurácia	Precisão	Recall	F1-Score	AUC-ROC	Valid. Cruzada
Floresta Aleatória	86.32%	87.18%	85.09%	86.12%	93.36%	93.18%
Aumento de Gradiente	87.54%	87.44%	87.62%	87.53%	94.10%	93.99%
Máquina de Suporte Vetorial	76.91%	75.88%	78.76%	77.29%	88.45%	88.02%
Deep Learning	94.00%	93,86%	93,84%	0,9344	94,00%	93,95%
Regressão logistica	78%	79%	77%	78%	78%	78%

Fonte: Elaborada pelo autor.

Os resultados mostram que todos os modelos de machine learning superaram a regressão logística em termos de acurácia, precisão e recall. A random forest, com acurácia de 79,38% no Banco A e 86,32% no Banco B, e o gradient boosting, com acurácia de 79,74% no Banco A e 87,54% no Banco B, apresentaram melhorias substanciais devido à sua capacidade de manipular dados complexos e ajustar hiperparâmetros críticos. O MLP, com acurácia de 84,45% no Banco A e 94,00% no Banco B, destacou-se em ambos os bancos, sendo a melhor escolha para lidar com relações não lineares, embora seu custo de treinamento seja maior. Já a regressão logística, embora simples e fácil de interpretar, mostrou-se limitada para capturar a complexidade dos dados, tornando-se menos competitiva em cenários que exigem alta precisão preditiva.

Por fim, o modelo de deep learning apresentou o melhor desempenho geral, destacando-se pela alta acurácia e um excelente equilíbrio entre precisão e recall. Em ambos os bancos, o deep learning mostrou-se a melhor escolha, evidenciando sua capacidade de capturar relações complexas nos dados e oferecer um desempenho superior em comparação aos outros modelos, especialmente nos cenários de classificação de riscos de crédito, onde a precisão na

identificação de riscos é crucial para minimizar perdas financeiras e garantir a sustentabilidade do negócio.


5

5 CONCLUSÃO

O presente trabalho buscou avaliar e comparar qual modelo de aprendizado de máquina apresenta maior eficiência na análise de risco de crédito. Para isso, foram considerados os modelos Support Vector Machine (SVM), Deep Learning com Perceptron Multicamadas (MLP), Gradient Boosting e Decision Tree, utilizando métricas como acurácia, precisão, recall, F1-Score, AUC-ROC e Validação Cruzada para compará-los. Adicionalmente, utilizou-se a Regressão Logística, um dos modelos mais aplicados no mercado bancário, como referência para demonstrar que todos os métodos de aprendizado de máquina analisados oferecem um desempenho superior.

Os resultados indicam que os modelos de aprendizado de máquina superam significativamente a Regressão Logística em termos de desempenho na análise de risco de crédito. Especificamente, entre os modelos de aprendizado de máquina, o Deep Learning com Perceptron Multicamadas (MLP) apresentou o melhor desempenho, alcançando maiores índices de acurácia e métricas associadas. Isso confirma que os modelos de aprendizado de máquina não apenas são mais eficazes que métodos estatísticos tradicionais, como também que o MLP se destaca entre eles na predição de inadimplência.

Esses achados são úteis para instituições financeiras porque demonstram que a implementação de modelos avançados de aprendizado de máquina, especialmente o Deep Learning, pode aprimorar a precisão na avaliação de risco de crédito. Isso contribui para uma melhor tomada de decisão, redução de perdas por inadimplência e maior eficiência no gerenciamento de risco.

Como sugestão para pesquisas futuras, recomenda-se explorar abordagens híbridas que combinem a robustez dos modelos de aprendizado de máquina com técnicas que promovam maior interpretabilidade. Isso atenderia às exigências regulatórias e facilitaria a adoção dessas tecnologias pelas instituições financeiras, potencializando os benefícios observados neste estudo.

REFERÊNCIAS

REFERÊNCIAS

REFERÊNCIAS

ADDO, P. M.; GUEGAN, D.; HASSANI, B. Credit risk analysis using machine and deep learning models. Risks, v. 6, n. 2, p. 38, 2018.

BI, W.; LIANG, Y. Risk assessment of operator's big data internet of things credit financial management based on machine learning. Mobile Information Systems, v. 2022, p. 1-11, 2022.

BRAVO, C.; CALABRESE, R.; LESSMANN, S. Credit risk and artificial intelligence: on the need for convergent regulation. SSRN, 2023.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Classification and regression trees. Monterey: Wadsworth, 1984.

BUTARU, Florin; CHEN, Qi; CLARK, Bryan J.; DAS, Sanjiv; LO, Andrew W.; SIDDIQUE, Arnab R. Risk and risk management in the credit card industry. Journal of Banking & Finance, v. 72, n. 11, p. 218-239, 2016.

CHOPRA, A.; BHILARE, P. Application of ensemble models in credit scoring models. Business Perspectives and Research, v. 6, n. 2, p. 129–141, 2018.

COELHO, F. F.; AMORIM, D. P. de L. Analisando métodos de machine learning e avaliação do risco de crédito. Revista Gestão & Tecnologia, 2021.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. S. Análise multivariada para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas, 2007.

DASTILE, X.; CELIK, T.; POTSANE, M. Statistical and machine learning models in credit scoring: a systematic literature survey. Applied Soft Computing, v. 91, Article 106263, 2020.

DROBETZ, W.; HOLLSTEIN, F.; OTTO, T.; PROKOPCZUK, M. Estimating stock market betas via machine learning. SSRN, 2021.

ESPOSITO, F.; MALERBA, D.; SEMERARO, G. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 19, n. 5, p. 476-491, 1997.

FARIAS, I.; SILVA, M. Ciência de dados no mercado de crédito: estratégias para mitigação de riscos e otimização de decisões com modelagem preditiva. Informática & Negócios, v. 2, n. 1, p. 45-60, 2023. Disponivel em: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ric.cps.sp.gov.b

r/bitstream/123456789/19765/3/informaticanegocios_2023_2_isaacdasilvafarias_cienciadedadosnomercadodecreditoestrategiaspara.pdf

FLOREZ-LOPEZ, R.; RAMON-JERONIMO, J. M. Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. Expert Systems with Applications, v. 42, n. 13, p. 5738-5751, 2015.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. Annals of Statistics, v. 29, n. 5, p. 1189-1232, 2001.

GUO, Y.; ZHOU, W.; LUO, C.; LIU, C.; XIONG, H. Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research, v. 249, n. 2, p. 417-426, 2016.

HAND, D. J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A, v. 160, n. 3, p. 523-541, 1997.

HOSMER, D. W.; LEMESHOW, S. Applied logistic regression. New York: Wiley, 1989.

HUANG, J.; CHEN, H.; HSU, C. J.; CHEN, W. H.; WU, S. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems, v. 37, n. 4, p. 543-558, 2004.

JOY, Z. H.; ISLAM, S.; RAHAMAN, M. A.; HAQUE, M. N. Advanced cybersecurity protocols for securing data management systems in industrial and healthcare environments. Global Mainstream Journal, v. 10, n. 4, p. 1950021, 2024.

JOVANOVIC, Z.; HOU, Z.; BISWAS, K.; MUTHUKKUMARASAMY, V. Robust integration of blockchain and explainable federated learning for automated credit scoring. Computer Networks, v. 243, p. 110303, 2024.

KANG, M.; AUSLOOS, M. Um estudo de problema inverso: credit risk ratings as a determinant of corporate governance and capital structure in emerging markets: evidence from Chinese listed companies. Economies, v. 5, n. 4, art. 47, 2017.

LENG, A.; XING, G.; FAN, W. Credit risk transfer in SME loan guarantee networks. Journal of Systems Science and Complexity, v. 30, n. 5, p. 1084-1096, 2017.

MARAJ, M. A. H. S. I.; MAHMUD, N. U. U. Information systems in health management: innovations and challenges in the digital era. International Journal of Health and Medical, v. 1, n. 2, p. 14-25, 2024.

MARKOV, A.; SELEZNYOVA, Z.; LAPSHIN, V. Credit scoring methods: Latest trends and points to consider. The Journal of Finance and Data Science, v. 8, p. 180-201, 2022.

MARTIN, R. D.; GUERARD, J. B.; XIA, D. Z. Resurrecting Earnings-to-Price with Robust Control for Outliers. [s.l.]: SSRN, 2024.

MONTEVECHI, A. A.; MIRANDA, R. de C.; MEDEIROS, A. L.; MONTEVECHI, J. A. B. Advancing credit risk modelling with machine learning: a comprehensive review of the state-of-the-art. Engineering Applications of Artificial Intelligence, v. 137, p. 109082, 2024.

OECD. Economic Outlook. Paris: OECD Publishing, 2021. Disponivel: https://www.oecd-ilibrary.org/economics/oecd-economic-outlook/volume-2021/issue-1_edfbca02-en acessado em 20/11/2024.

QUAN, J.; SUN, X. Credit risk assessment using the factorization machine model with feature interactions. Humanities and Social Sciences Communications, 2024.

QUINLAN, J. R. Induction of decision trees. Machine Learning, v. 1, n. 1, p. 81-106, 1986.

RAHMAN, M. M.; ISLAM, S.; KAMRUZZAMAN, M.; JOY, Z. H. Advanced query optimization in SQL databases for real-time big data analytics. Revista Acadêmica de Administração de Empresas, Inovação e Sustentabilidade, v. 4, n. 3, p. 1-14, 2024.

SCHRICKEL, W. K. Análise de crédito: concessão e gerência de empréstimos. 5. ed. São Paulo: Atlas, 2000.

STEVE, M. N.; OLUSEGUN, J.; PAUL, H. Addressing class imbalance with synthetic data generation. 2024.

SUHADOLNIK, Nicolas; UEYAMA, Jo; SILVA, Sergio Da. Machine learning for enhanced credit risk assessment: An empirical approach. Journal of Risk and Financial Management, v. 16, n. 12, art. 496, 2023.

BISPO, Ulysses Araujo. Aplicação do Modelo de Regressão Logística na Análise do Risco de Crédito de duas Instituições Bancárias. 2015. Trabalho de Graduação – Universidade de Brasília, Brasília, 2015.

VICENTE, J. Fintech disruption in Brazil: a study on the impact of open banking and instant payments in the Brazilian financial landscape. Social Impact Research Experience, n. 86, 2020.

WANG, Zhao; JIANG, Cuiqing; DING, Yong; LYU, Xiaozhong; LIU, Yao. A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. Electronic Commerce Research and Applications, [s.l.], v. 27, p. 74–82, 2018.

XIA, Y.; LIU, Y.; LIU, N. Gradient boosting for credit scoring: a survey and new insights. Information Fusion, v. 64, p. 149-161, 2020.

ZHANG, L.; YU, Q.; ZHOU, B.; ZHANG, Y.; HU, Z. Incorporating Feature Interactions and Contrastive Learning for Credit Prediction. IEEE Access, 2023.

YOUNUS, M.; HOSSEN, S.; ISLAM, M. M. Advanced business analytics in textile & fashion industries: driving innovation and sustainable growth. International Journal of Management Information Systems and Data Science, v. 1, n. 2, p. 37-47, 2024.

ZHANG, Xiaoming; YU, Lean. Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods. Expert Systems With Applications, v. 237, p. 121484, 2024

ZHONG, H.; MIAO, C.; SHEN, Z.; FENG, Y. Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings. Neurocomputing, v. 128, p. 285-295, 2014.

DROBETZ, W.; HOLLSTEIN, F.; OTTO, T.; PROKOPCZUK, M. Estimating stock market betas via machine learning. SSRN, 2023.


idp
A ESCOLHA QUE
TRANSFORMA
O SEU CONHECIMENTO