UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Consumer Lending Business: Revisão de técnicas de Machine Learning

Ícaro Almeida Aguiar

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Ícaro Almeida Aguiar

Consumer Lending Business: Revisão de técnicas de Machine Learning

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Concentration area: Artificial Intelligence and Big Data

Orientador: Prof. Dr. Marcelo Manzato

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m	Aguiar, Icaro Almeida Consumer Lending Business: Revisão de técnicas de Machine Learning / Ícaro Almeida Aguiar ; Manzato, Marcerlo. – São Carlos, 2024. 81 p. : il. (algumas color.) ; 30 cm. Monograph (MBA in Artificial Intelligence and Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024. 1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. MANZATO, M.. II. Título.

Ícaro Almeida Aguiar

Consumer Lending Business: Revisão de técnicas de Machine Learning

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Área de concentração: Inteligência Artificial e Big Data

Original version

São Carlos

2024

AGRADECIMENTOS

Gostaria de agradecer inicialmente ao meus pais, André e Kátia, por terem me dado condições e apoio para realizar a graduação na USP, e desde cedo terem me incentivado a pesquisa. Sem sombra de dúvidas, a minha busca pela especialização de IA e Big Data e a execução deste trabalho, são apenas um reflexo do que aprendi em casa e pelo seu amor ao conhecimento e a pesquisa.

Quero agradecer também ao meu orientador, Marcelo Manzato, pelo aprendizado, disponibilidade e paciência ao longo da orientação deste trabalho.

Por fim, agradeço a minha namorada, Maria Eduarda Cavalcante, pelo apoio e compreensão durante a escrita deste trabalho.

RESUMO

AGUIAR, I. A. **Consumer Lending Business: Revisão de técnicas de Machine Learning**. 2024. 81p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Este trabalho examina o uso de vários modelos de *Machine Learning* (ML) e *Deep Learning* (DL) para avaliar sua eficácia na previsão de *scores* de crédito, utilizando um conjunto de dados público baseado na população alemã. A pesquisa aborda lacunas na literatura existente ao analisar o impacto de diferentes hiperparâmetros e características no desempenho dos modelos, focando em métricas como AUC, acurácia e MAPE. Também discute a importância da seleção de características para evitar vazamento de dados nos modelos de *scoring* de crédito, garantindo a aplicabilidade dos modelos em cenários do mundo real.

A análise utiliza o conjunto de dados de crédito alemão, que, apesar de seu uso comum em estudos anteriores, oferece potencial adicional para explorar o ajuste fino de modelos de *Machine Learning*. Este estudo avalia cuidadosamente a relevância de várias características no conjunto de dados, especialmente aquelas que podem não estar disponíveis no momento da avaliação de crédito, como o valor do crédito. Ao excluir tais características do treinamento do modelo, a pesquisa busca criar modelos que sejam mais práticos para aplicações do mundo real, onde nem todas as informações podem estar disponíveis antecipadamente.

Os resultados sugerem que, enquanto estudos anteriores incluíram todas as características disponíveis, o que potencialmente leva a vieses, a abordagem deste estudo de seleção cuidadosa de características e ajuste de hiperparâmetros produz modelos que apresentam bom desempenho em métricas tradicionais e são mais aplicáveis a situações do mundo real. Esta pesquisa contribui para o desenvolvimento de sistemas de *scoring* de crédito mais confiáveis e oferece insights sobre a aplicação eficaz de técnicas avançadas de *Machine Learning* e *Deep Learning* em contextos financeiros.

Palavras-chave: Machine Learning. Deep Learning. Credit Scoring. Feature Selection. Hyperparameter Tuning. German Credit Dataset. AUC. Accuracy. MAPE. Data Leakage. Modelos financeiros. Aplicações reais.

ABSTRACT

AGUIAR, I. A. **Consumer Lending Business: Machine Learning techniques review**. 2024. 81p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

This study examines the use of various Machine Learning (ML) and Deep Learning (DL) models to evaluate their effectiveness in predicting credit scores using a public dataset based on the German population. The research addresses gaps in existing literature by analyzing the impact of different hyperparameters and features on model performance, focusing on metrics such as AUC, accuracy, and MAPE. It also discusses the importance of feature selection to prevent data leakage in credit scoring models, ensuring the models' applicability in real-world scenarios.

The analysis uses the German credit dataset, which, despite its common usage in past studies, offers further potential for exploring the fine-tuning of ML models. This study carefully evaluates the relevance of various features in the dataset, especially those that may not be available at the time of credit scoring, like the credit amount. By excluding such features from the model training, the research aims to create models that are more practical for real-world applications where all information may not be available upfront.

The results suggest that while previous studies have included all available features, potentially leading to biases, this study's approach of careful feature selection and hyperparameter tuning produces models that perform well on traditional metrics and are more applicable to real-world situations. This research contributes to the development of more reliable credit scoring systems and offers insights into the effective application of advanced ML and DL techniques in financial contexts.

Keywords: Machine Learning. Deep Learning. Credit Scoring. Feature Selection. Hyperparameter Tuning. German Credit Dataset. AUC. Accuracy. MAPE. Data Leakage. Financial Models. Real-World Applications.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e Motivação

A utilização de score de crédito para consentimento de empréstimos começa ao redor de 1950 (THOMAS; CROOK; EDELMAN, 2017), com as primeiras aplicações sendo baseadas em frameworks generalistas, tais como os 5C's:

- *Character: do you know the person or their family?*
- *Capital: how much is being asked for?*
- *Collateral: what is the borrower willing to put up from their resources?*
- *Capacity: what is their repayment ability?*
- *Condition: what are the conditions in the market?*

Com base nesses frameworks, análises manuais eram performadas e avaliadas individualmente. No entanto, essas abordagens geram gargalos para instituições financeiras, não havendo ganhos de escala e gerando uma má experiência ao consumidor. Assim, surgiu o *scorecard*, que permitiu avaliar a nota de um cliente com base em diferentes métricas, como informações demográficas, número de dependentes, tempo no emprego atual e informações vindas de *bureaus* locais (e.g. Serasa e Boa Vista) (DASTILE; CELIK; POTSANE, 2020).

Além disso, *scores* de crédito permitem instituições financeiras terem maior controle e previsibilidade dos seus portfólios, fazendo com que o risco possa ser melhor controlado e, como consequência, permita um custo menor (i.e. taxas de juros menores) (LAWRENCE; SOLOMON, 2013). Também é importante salientar que o conceito de *score* de crédito vem se popularizando cada vez mais no Brasil, principalmente com a lei do cadastro positivo (Brasil, 2011). Atualmente, diferentes empresas provêem serviços de monitoramento desses dados (e.g. Serasa e SPC) e *fintechs* que disponibilizam essa informação para clientes poderem melhorar seu perfil de crédito, tendo explicações sobre mudanças no crédito concedido (e.g. PicPay, Mercado Pago e Nubank).

Olhando mais para o histórico dessa métrica, em 1980 a utilização de técnicas de aprendizado de máquina propiciaram a transformação do cálculo de risco de *default*, fazendo com que instituições financeiras conseguissem desenvolver estratégias de larga escala e possibilitaram também a redução do custo de crédito (PROVOST; FAWCETT, 2013).

Com o avanço da tecnologia e melhor capacidade de processamento de dados nas decadas de 1990 até 2010 (SILVA, 2023), houve um avanço na utilização de modelos de

aprendizado de máquina no contexto de crédito, principalmente devido a sua natureza não-linear (WANG; XU; PUSATLI, 2015).

Essas técnicas estatísticas, de acordo com (DASTILE; CELIK; POTSANE, 2020), são *Linear Discriminant Analysis* (LDA), *Logistic Regression* (LR) e Naïve Bayes (NB), e para aprendizado de máquina incluiriam *k-Nearest Neighbor* (k-NN), *Decision Trees* (DTs), *Support Vector Machines* (SVMs), *Artificial Neural Networks* (ANNs), *Random Forests* (RFs), *Boosting*, *Extreme Gradient Boost* (XGBoost), *Bagging*, *Restricted Boltzmann Machines* (RBMs), *Deep Multi-Layer Perceptron* (DMLP), *Convolutional Neural Networks* (CNNs) e *Deep Belief Neural Networks* (DBNs), sendo uma lista não exaustiva.

Para a aplicações reais, a grande gama de diferentes técnicas de aprendizado de máquina e métodos estatísticos acaba não sendo aplicado por instituições financeiras em seus modelos de crédito devido a questões de performance, escalabilidade e custo, que podem restringir a utilização de alguns modelos citados anteriormente (HUYEN, 2022).

Dessa forma, a motivação deste trabalho é avaliar diferentes técnicas de *Machine Learning* em um *dataset* público com base nos dados de crédito de um banco Alemão na década de 1990 (HOFMANN, 1994), avaliando o resultado dos modelos por meio dos princípios de *data mining*, bem como sua explicabilidade.

1.2 Objetivos

Os principais objetivos deste trabalho de conclusão de curso são:

1. Elencar os principais modelos utilizados para *consumer lending business* na literatura
2. Explorar a performance dos diferentes modelos de aprendizado de máquina em um dataset público
3. Discutir os resultados obtidos e vantagens/desvantagens dos modelos utilizados

1.3 Organização do texto

Este trabalho é dividido nas seguintes seções:

- Fundamentação teórica: São abordados os diferentes métodos de aprendizado de máquina que serão utilizados no trabalho e contexto sobre a aplicação do problema
- Trabalhos relacionados: É feita uma revisão da literatura utilizando artigos de referência, discussões conceituais acerca do tema de crédito e comparação de técnicas tradicionais e novos métodos propostos
- Proposta e desenvolvimento: Exposição da motivação do trabalho e metodologia utilizada para seu desenvolvimento

- Avaliação experimental: Exposição do conjunto de dados, configurações utilizadas para o experimento, por fim os resultados e discussões acerca do tema

2 FUNDAMENTAÇÃO TEÓRICA

Durante o capítulo serão apresentados e discutidos diferentes técnicas de processamento de dados, modelos de *Machine Learning* citados na literatura para problemas de crédito e *score* de risco. Além disso, também será apresentada uma breve introdução a conceitos de crédito e maior elucidação sobre o problema.

2.1 Técnicas de balanceamento de dados: SMOTE

SMOTE, que significa *Synthetic Minority Over-sampling Technique*, é uma técnica amplamente utilizada para lidar com problemas de aprendizado de máquina em conjuntos de dados desequilibrados. Em cenários onde a classe de interesse é significativamente menor que outras, como em detecção de fraudes ou diagnósticos médicos, a distribuição desigual pode levar a modelos enviesados, que tendem a favorecer a classe majoritária. Para mitigar esse problema, o SMOTE é aplicado para aumentar o número de exemplos da classe minoritária, garantindo que o modelo tenha mais oportunidades para aprender e reconhecer padrões associados a essa classe (CHAWLA *et al.*, 2002).

O principal objetivo do SMOTE é melhorar a sensibilidade do modelo em relação à classe minoritária, equilibrando a distribuição das classes sem simplesmente replicar os exemplos existentes. Diferente de métodos tradicionais de *oversampling*, que duplicam exemplos da classe minoritária, o SMOTE gera novos exemplos sintéticos. Isso permite que o modelo construa uma compreensão mais robusta da classe minoritária, o que é essencial para aplicações onde a precisão na detecção de casos raros é crucial.

O funcionamento do SMOTE baseia-se na geração de exemplos sintéticos ao longo das linhas entre os exemplos reais da classe minoritária e seus vizinhos mais próximos. Ao invés de concentrar o aprendizado em pontos isolados, o SMOTE expande a distribuição da classe minoritária no espaço de características, permitindo ao modelo aprender padrões mais generalizáveis e menos suscetíveis a *overfitting*. Esse processo resulta em uma melhor performance do modelo, especialmente em contextos onde a detecção precisa de eventos raros é crítica.

2.2 Modelos de Machine Learning

2.2.1 Logistic Regression

A regressão logística é um método supervisionado utilizado de forma frequente para estimar a probabilidade de uma instância pertencer a uma determinada classe. Se a probabilidade estimada for maior do que o limite (usualmente de 50%), o modelo prediz que uma instância pertence a uma classe ou outra (GéRON, 2021).

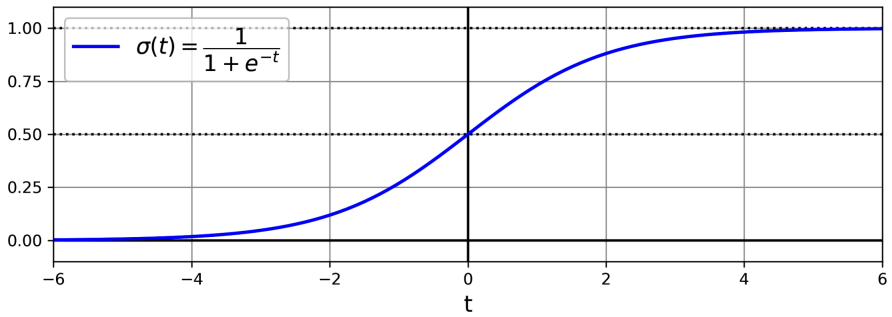

Figura 1 – Função logística

Assim como a regressão linear, o modelo de regressão logística calcula a soma ponderada dos seus inputs mais um termo independente. Mas, ao invés do seu resultado ser direto, o resultado é submetido a uma função logística (ou sigmoide), conforme descrito abaixo:

$$h_\theta(x) = \sigma(\theta^x) \tag{2.1}$$

Onde, θ^x é aplicado em:

$$\sigma(t) = \frac{1}{1+e^{-t}} \tag{2.2}$$

O comportamento da função logística pode ser descrito na figura 1, que possui forma de "S"transformando qualquer valor real em um valor compreendido entre 0 e 1. Essa transformação é crucial quando queremos modelar probabilidades, que, por definição, devem estar dentro desse intervalo.

Na regressão logística, usamos essa função para converter as previsões do modelo em probabilidades. Mais especificamente, a regressão logística é usada para modelar a probabilidade de um determinado evento ocorrer. Os parâmetros do modelo são ajustados para maximizar a probabilidade de que as previsões do modelo coincidam com os resultados observados nos dados de treinamento.

2.2.1.1 Vantagens

1. Interpretabilidade: A regressão logística fornece coeficientes que representam a contribuição de cada variável independente para a probabilidade do resultado binário. Isso facilita a interpretação do efeito de cada variável no resultado e ajuda a entender a relação entre as variáveis independentes e dependentes.

2. Computacionalmente Eficiente: Comparada com modelos mais complexos, como redes neurais, a regressão logística é computacionalmente eficiente e rápida para

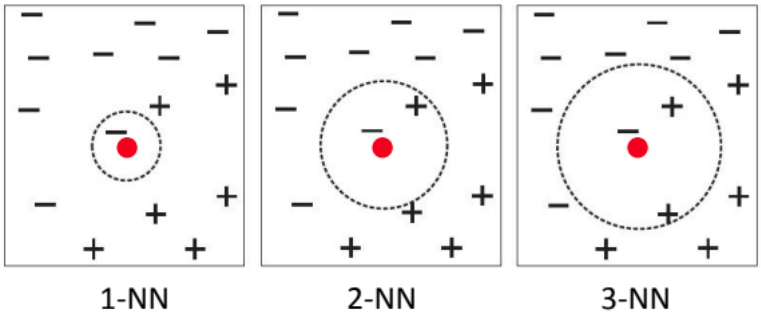

Figura 2 – Exemplificação do kNN

treinar, especialmente em grandes conjuntos de dados. Isso a torna uma escolha prática para problemas onde o tempo de computação é uma consideração importante.

2.2.1.2 Desvantagens

1. Limitação na Complexidade do Modelo: A regressão logística é linear por natureza, o que significa que só pode modelar relações lineares entre as variáveis independentes e a variável dependente. Isso pode limitar sua capacidade de capturar relações complexas ou não lineares nos dados.

2.2.2 k Neareast Neighbors

O método k-Nearest Neighbors (kNN) é um algoritmo simples e eficaz para tarefas de classificação e regressão, que atribui a uma amostra a classe mais comum entre seus 'k' vizinhos mais próximos no espaço de características.

2.2.2.1 Como a Escolha de 'k' Afeta o Desempenho do Algoritmo

A escolha do valor de 'k' (o número de vizinhos considerados) é crucial para o seu desempenho, podendo ser visualizado na figura 2.

2.2.2.2 Valor de 'k' Muito Pequeno (k = 1, 2, 3...)

Sensibilidade ao Ruído: Com valores pequenos de 'k', o modelo tende a ser muito sensível ao ruído e a *outliers*. Isso acontece porque, com poucos vizinhos, a decisão é fortemente influenciada por qualquer ponto atípico que esteja próximo do ponto a ser classificado.

Overfitting: Um 'k' pequeno pode levar o modelo a superajustar-se aos dados de treinamento, capturando nuances desnecessárias e irrelevantes dos dados, o que pode piorar o desempenho em novos dados (generalização).

Exemplo prático: Imagine um problema onde estamos tentando classificar se uma fruta é uma maçã ou uma laranja. Se 'k' for 1, e há uma única maçã vermelha perto de

uma área dominada por laranjas, o kNN pode classificar incorretamente uma nova fruta como maçã, simplesmente porque ela está mais próxima daquela maçã outlier.

2.2.2.3 Valor de ‘k‘ Muito Grande (k = 20, 30, ...)

Perda de Detalhes Locais: Com um valor grande de ‘k‘, o modelo começa a perder detalhes locais, porque ele está considerando muitos vizinhos, inclusive aqueles que podem estar muito distantes do ponto em questão.

Underfitting: Um valor de ‘k‘ grande pode levar a um modelo que é muito genérico, não capturando adequadamente as fronteiras entre diferentes classes. Isso resulta em um modelo que pode ter dificuldades em diferenciar corretamente entre classes próximas.

Exemplo prático: Suponha que estamos tentando classificar e-mails como spam ou não spam. Se ‘k‘ for muito grande, o algoritmo pode considerar muitos e-mails irrelevantes ao determinar a classificação de um novo e-mail, levando a uma classificação imprecisa, talvez classificando um e-mail legítimo como spam porque a maioria dos seus vizinhos são spam, mesmo que eles estejam longe no espaço das características.

2.2.2.4 Vantagens

- Simples para entendimento e implementação

2.2.2.5 Desvantagens

- O valor de k é determinado experimentalmente
- Custo computacional alto, que pode ser mitigado por técnicas de paralelismo e indexação para acelerar o cálculo de distâncias, tais como KD-Tree

2.2.3 SVC

O Support Vector Classifier (SVC), derivado das Support Vector Machines (SVMs), é uma ferramenta de aprendizado supervisionado para classificação. Ele identifica um hiperplano de separação ótimo para maximizar a margem entre classes, oferecendo robustez em conjuntos de dados complexos. Esse hiperplano ajuda a melhorar a generalização do modelo, tornando-o mais robusto em novos dados.

Quando os dados não são linearmente separáveis no espaço original, o SVC utiliza técnicas de kernels para transformar os dados em um espaço de maior dimensão, onde a separação entre as classes se torna mais viável. Os kernels funcionam como funções que mapeiam os dados de entrada para esse novo espaço de maneira implícita, sem a necessidade de realizar a transformação diretamente. Isso significa que, em vez de calcular todas as novas coordenadas em um espaço de alta dimensão (o que pode ser computacionalmente

inviável), o kernel calcula diretamente os produtos internos entre os pontos nesse espaço transformado.

Essa abordagem permite ao SVC encontrar um hiperplano linear no novo espaço de alta dimensão que corresponde a uma fronteira de decisão não linear no espaço original. Dessa forma, o SVC pode lidar com problemas de classificação complexos, onde as classes não podem ser separadas por uma linha reta ou plano simples.

O uso de kernels não apenas facilita a separação das classes, mas também melhora a eficiência do algoritmo, já que a transformação dos dados é feita de maneira eficiente, economizando tempo e recursos computacionais.

2.2.3.1 Vantagens:

1. Eficiência em Espaços de Alta Dimensão: O SVC é eficaz em espaços de alta dimensionalidade, o que significa que pode lidar bem com conjuntos de dados com muitas características, como em problemas de processamento de linguagem natural ou visão computacional.

2. Robustez com Dados Não Linearmente Separáveis: O SVC pode ser adaptado para lidar com dados que não são linearmente separáveis, utilizando funções de kernel para mapear os dados em espaços de características mais complexos, onde a separação linear pode ser alcançada.

2.2.3.2 Desvantagens:

1. Exigência Computacional: O treinamento do SVC pode ser computacionalmente exigente, especialmente em grandes conjuntos de dados, devido à necessidade de resolver um problema de otimização complexo para encontrar o hiperplano de separação ótimo.

2. Interpretabilidade Complexa: Interpretar os resultados do SVC pode ser desafiador, especialmente quando são aplicadas transformações de kernel para mapear os dados em espaços de características de alta dimensão. Isso pode dificultar a compreensão das relações entre as características e as classes.

2.2.4 Ensemble Learning

Procura melhorar a acurácia combinando predições de múltiplos estimadores. A utilização de múltiplos classificadores gera novas classes e essas podem ser utilizadas para gerar outras novas ou combinadas para gerar uma final. Podendo ser combinadas de forma hierárquica, paralela ou sequencial. O valor final pode ser obtido por meio de combinação de classe majoritária, média ou algoritmo combinador. (MARCACINI, 2023)

2.2.4.1 Vantagens

1. Melhoria na Precisão: Uma das principais vantagens do ensemble learning é que ele geralmente produz modelos mais precisos do que qualquer modelo individual no ensemble. Ao combinar múltiplos modelos que podem capturar diferentes aspectos dos dados ou diferentes fontes de variabilidade, o ensemble pode reduzir o erro de generalização e melhorar a precisão das previsões.

2. Robustez e Estabilidade: Ensemble learning tende a produzir modelos mais robustos e estáveis. Ao usar múltiplos modelos e/ou algoritmos diferentes, o ensemble pode lidar melhor com ruído nos dados, outliers e outros problemas que podem afetar negativamente o desempenho de um único modelo. Isso resulta em uma generalização mais robusta para novos dados.

3. Flexibilidade e Adaptabilidade: O ensemble learning é altamente flexível e pode ser aplicado a uma ampla variedade de problemas de aprendizado de máquina. Pode-se combinar diferentes tipos de modelos (por exemplo, árvores de decisão, redes neurais, SVMs) ou ajustar diferentes parâmetros para cada modelo no ensemble. Isso permite que o ensemble se adapte melhor às características específicas do problema e dos dados.

2.2.4.2 Desvantagens

1. Complexidade e Custos Computacionais: Ensemble learning pode ser mais complexo computacionalmente e exigir mais recursos computacionais do que a construção de um único modelo. Treinar e manter múltiplos modelos pode aumentar significativamente os custos computacionais, especialmente para grandes conjuntos de dados ou ensembles com muitos modelos.

2. Interpretabilidade Reduzida: Enquanto um único modelo pode ser mais facilmente interpretado e explicado, especialmente em termos de como as características afetam as previsões, ensembles podem ser mais difíceis de interpretar. A combinação de vários modelos pode obscurecer a relação entre as características e as previsões, tornando mais difícil entender como o modelo está tomando suas decisões.

3. Sensibilidade a Overfitting: Embora ensemble learning possa ajudar a reduzir o overfitting em comparação com modelos individuais, ele ainda pode ser sensível a overfitting, especialmente se os modelos no ensemble forem muito complexos ou altamente correlacionados. Se os modelos individuais no ensemble estiverem superajustados aos mesmos padrões nos dados de treinamento, o ensemble pode não generalizar bem para novos dados, resultando em um desempenho pior do que o esperado.

2.2.4.3 Random Forest

O método Random Forest é um comitê de classificadores de *decision trees*, que possuem predições combinadadas usualmente treinadas via *bagging* (GéRON, 2021). Cada árvore é induzida usando um subconjunto aleatório de atributos preditivos usados na escolha do atributo para cada nó, possui profundidade nos ramos e com classificação ocorre por voto majoritário (ROMERO, 2023).

2.2.4.4 Boosting

Boosting se refere a qualquer método de comitê de classificadores que pode combinar diversos modelos com aprendizado mais fraco e transformá-los em um mais forte. A ideia principal é treinar os modelos de forma sequencial, onde os sucessores corrigem os predecessores. Existem diferentes tipos de métodos que utilizam esta técnica, a seguir os principais e mais utilizados atualmente são destacados (DASTILE; CELIK; POTSANE, 2020).

2.2.4.4.1 Gradient Boost

Método que utiliza *decision trees* rasas e fracas com cada árvore aprendendo e melhorando em relação a anterior. Por conta dessa arquitetura sequencial e restrições em paralelismo, acaba necessitando de muitas árvores (> 1.000), gerando custo computacional mais alto (ROMERO, 2023).

2.2.4.4.2 Extreme Gradient Boosting (XGBoost)

Framework otimizado e distribuído baseado no *Gradient Boost*, projetado para ser altamente eficiente, flexível e portátil (ROMERO, 2023). Sua capacidade de paralelização faz com que seja mais escalavel e possa resolver problemas com bilhões de exemplos. É um dos modelos mais utilizados atualmente, sendo bastante recorrente em competições como Kaggle/KDD Cup (CHEN; GUESTRIN, 2016). O crescimento da sua árvore é, usualmente, *level-wise*, completando um nível por vez.

2.2.4.4.3 Light Gradient Boosting Machine (LightGBM)

Framework baseado também no *Gradient Boost*, que utiliza algoritmos de aprendizado de árvores de decisão, sendo projetado pela Microsoft. Assim como o XGBoost possui desempenho mais eficiente e flexível do que o seu predecessor. Crescimento dos ramos é feito de forma *leaf-wise*, dividingo um nó folha por vez.

2.2.4.4.4 XGBoost vs Light GBM

O LightGBM e o XGBoost são ambos frameworks de gradient boosting comumente usados para tarefas de aprendizado supervisionado, principalmente em problemas de classificação e regressão. As principais diferenças entre eles são:

1. Velocidade e Eficiência:
    - O LightGBM é geralmente mais rápido e mais eficiente em termos de memória do que o XGBoost. Ele alcança isso usando um algoritmo baseado em histograma para a construção da árvore, o que reduz o uso de memória e permite a computação paralela e distribuída.
    - O XGBoost, por outro lado, usa um algoritmo pré-classificado, que pode ser mais lento e mais intensivo em termos de memória, especialmente para conjuntos de dados grandes.
2. Manuseio de Características Categóricas:
    - O LightGBM suporta nativamente características categóricas, o que significa que você não precisa codificá-las com one-hot encoding. Ele usa uma técnica chamada "Gradient-based One-Side Sampling"(GOSS) para lidar eficientemente com características categóricas.
    - O XGBoost requer que as características categóricas sejam codificadas com one-hot encoding antes do treinamento, o que pode levar a um aumento no uso de memória e sobrecarga computacional, especialmente para variáveis categóricas de alta cardinalidade.
3. Estratégia de Crescimento da Árvore:
    - O LightGBM usa uma estratégia de crescimento de folhas, onde ele cresce a árvore nó a nó, selecionando a folha com a maior perda de delta para crescer em cada iteração. Isso pode levar ao overfitting se não for adequadamente regularizado, mas geralmente resulta em uma convergência mais rápida.
    - O XGBoost, por outro lado, usa uma estratégia de crescimento por nível, onde ele cresce a árvore nível por nível. Esta abordagem tende a ser mais conservadora e menos suscetível ao overfitting, mas pode ser mais lenta do que o LightGBM, especialmente para árvores profundas.
4. Técnicas de Regularização:
    - Tanto o LightGBM quanto o XGBoost suportam várias técnicas de regularização, como regularização L1 e L2 (para controlar a complexidade do modelo),

max depth (para controlar a profundidade da árvore), *min child weight* (para controlar o número mínimo de instâncias necessárias em cada nó filho) e gamma (para controlar a redução mínima de perda necessária para fazer uma partição adicional).

- O LightGBM introduz técnicas adicionais de regularização, como "Gradient-based One-Side Sampling"(GOSS) e "Exclusive Feature Bundling"(EFB) para melhorar a eficiência e reduzir o overfitting.

2.3 Técnicas de interpratibilidade dos modelos: SHAP

As SHapley Additive exPlanations (SHAP) são uma abordagem baseada na teoria dos jogos para explicar o resultado de modelos de aprendizado de máquina (LUNDBERG; LEE, 2017b). Baseia-se no conceito de valor de Shapley da teoria dos jogos cooperativos, que distribui de forma justa os ganhos totais entre os jogadores com base em suas contribuições individuais. No contexto do aprendizado de máquina, cada característica de um modelo é considerada um jogador em um jogo, e o objetivo é atribuir de forma justa a predição às diferentes características. O SHAP fornece uma medida unificada de importância das características calculando a contribuição média de cada característica para a predição em todas as possíveis combinações de características.

O método SHAP garante que os valores de importância das características que ele fornece sejam consistentes e somem a diferença entre a predição e a predição média no conjunto de dados. Essa consistência é uma vantagem significativa em relação a outros métodos de importância de características, pois garante que adicionar uma característica que sempre aumenta a predição terá um valor SHAP positivo. Além disso, os valores SHAP são aditivos, o que significa que a soma dos valores SHAP para todas as características é igual ao resultado do modelo, permitindo uma interpretação direta da contribuição de cada característica para a predição.

Uma das forças do SHAP é sua aplicabilidade a vários tipos de modelos, incluindo os complexos como *ensemble tree models* e redes neurais. Ao decompor uma predição em contribuições individuais de cada característica, os valores SHAP fornecem insights claros e detalhados sobre como o modelo toma decisões. Essa transparência é crucial para validação de modelos, depuração e construção de confiança com as partes interessadas, pois ajuda a identificar quais características estão impulsionando as predições e se essas contribuições estão alinhadas com o conhecimento do domínio e as expectativas (LUNDBERG; LEE, 2017b).

2.4 Consumer Lending Business

O mercado de crédito foi fundamental para evolução da economia, possibilitando o seu desenvolvimento, tendo uma grande adoção em diferentes países e rápido crescimento

nos últimos anos (LAWRENCE; SOLOMON, 2013). No entanto, para que o cliente final tenha acesso ao crédito algumas perguntas devem ser respondidas: A quem devo conceder crédito? Quanto devo conceder?

Este é um problema que possui diferentes soluções e técnicas para ser resolvido e vêm sofrendo mudanças e melhorias constantes. Como discutido nas seções anteriores, diferentes modelos e técnicas estatísticas podem ser utilizadas, principalmente para terem um ganho de escala na aplicação de políticas de crédito (LAWRENCE; SOLOMON, 2013). O primeiro grande caso de sucesso foi o banco americano *Capital One* na decáda 90, o qual utilizava técnicas estatísticas para gerar seus *scores* de crédito (PROVOST; FAWCETT, 2013).

Usualmente, o primeiro passo para responder a essa pergunta é por meio de um *score* de crédito, que surge da transformação de dados relevantes (i.e. *features*) em uma variável categórica (e.g. bom/mal pagador) e/ou númerica (e.g. probabilidade de não pagamento). O primeiro é mais comum para datasets públicos, como descrito em (HOFMANN, 1994), enquanto segundo é mais utilizado em aplicações práticas de instituições financeiras. Há ainda a possibilidade de a partir de uma variável númerica, serem criados diferentes intervalos de valores, culminando em uma variável categórica. Uma das vantagens dessa abordagem é a segmentação de clientes em grupos, minimizando eventuais problemas com a acurácia do modelo.

A partir do *score*, outros dados importantes para a concessão de crédito podem ser calculados, como:

1. Gastos para os próximos meses
2. Perdas esperadas dado não pagamento
3. Valor de crédito concedido
4. Gasto esperado dado aumento do crédito concedido

A lista acima não é exaustiva, mas traz um panorama de diferentes métricas que podem ser utilizadas e dependem do *score* de crédito. Tanto a métrica de *score*, quanto as outras métricas podem ser calculadas por meio de modelos de aprendizado de máquina. Para instituições financeiras as métricas de risco de crédito possuem importância vital, pois além de influenciarem na tomada de decisão sobre o consentimento de crédito para um cliente, também são utilizadas para realizar as provisões esperadas por perdas que se valem desses *scores* para realizarem o cálculo da provisão, as quais devem ser feitas de forma mensal e obrigatória no Brasil (Banco Central do Brasil, 1999).

3 TRABALHOS RELACIONADOS

Para avaliação dos trabalhos relacionados, bem como seus objetivos, resultados e lacunas serão analisados diferentes referências divididas em três categorias diferentes.

3.1 Revisão da literatura

Nesta seção serão explorados dois artigos principais que fazem uma revisão sistemática da literatura acerca do tema

3.1.1 Classification methods applied to credit scoring: Systematic review and overall comparison

O artigo de LOUZADA; ARA; FERNANDES tem como objetivo apresentar uma revisão sistemática abrangente da teoria e aplicação de técnicas de classificação binária para análise financeira de *credit scoring*. Os principais objetivos do artigo são analisar minuciosamente a literatura existente sobre *credit scoring*, cobrindo mais de 20 anos de pesquisa, de 1992 a 2015, e fornecer um estudo de simulação experimental primário sob nove metodologias gerais. A metodologia de revisão sistemática utilizada nesta pesquisa é sistemática, qualitativa e quantitativa, incorporando 187 artigos recuperados das bases de dados ScienceDirect, Engineering Information, Reaxys e Scopus. O artigo empregou um procedimento bem definido para selecionar e classificar os artigos, incluindo categorias instrumentais para classificação.

O estudo utilizou um processo rigoroso para coleta e análise de dados, levando em consideração o ano de publicação, título do periódico, coautores e um esquema conceitual baseado em 13 questões para entender a aplicação histórica das técnicas de *credit scoring*. O artigo cobriu extensivamente os principais objetivos dos artigos revisados, suas peculiaridades dos artigos de *credit scoring*, seus métodos de classificação, tipos de conjuntos de dados e outros critérios relevantes para validação de métodos de *Machine Learning*, como abordagens de validação e critérios de custo de classificação incorreta.

Os artigos revisados abrangeram uma variedade de conjuntos de dados, sendo a maioria privados devido a preocupações de confidencialidade, embora conjuntos de dados públicos também fossem usados com menos frequência, possuindo uma mistura de variáveis contínuas e discretas. Conjuntos de dados de referência, como os conjuntos de dados de crédito australiano (QUINLAN,) e alemão (HOFMANN, 1994), foram comumente usados, destacando seu papel como referências padrão devido à dificuldade de acessar conjuntos de dados diversificados e abrangentes. Olhando mais especificamente para a etapa de validação dos modelos nos artigos, os conjuntos de dados citados anteriormente chegam a representar 45% dos 187 artigos explorados (LOUZADA; ARA; FERNANDES, 2016).

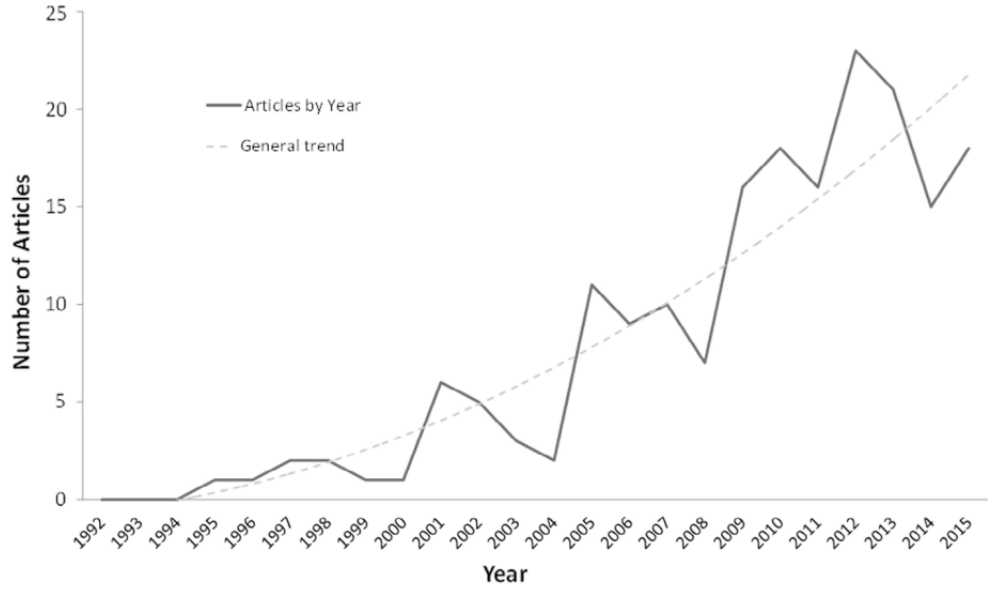

Figura 3 – Número de artigos relacionados a *credit score* por ano elaborado por (LOUZADA; ARA; FERNANDES, 2016).

Os principais achados da pesquisa incluem uma tendência crescente no número de artigos publicados na área de *credit scoring*, como destacado na figura 3, uma preferência por propor novos métodos para classificação de crédito, técnicas híbridas se destacando como as mais comuns e uma diminuição na ênfase em comparar técnicas tradicionais nos períodos recentes.

Além disso, observou-se que a maioria dos artigos não emprega simulações exaustivas (e.g. Monte Carlo), e a imputação de dados ausentes é utilizada com frequência na análise de *credit scoring*, relatando diferentes técnicas para mitigar estes problemas, tais como: remoção das instâncias nulas e substituição de valores nulos. O artigo também descobriu uma alta frequência de procedimentos de *feature selection* e observou uma mudança no desempenho preditivo dos métodos quando dados desbalanceados são encontrados. Adicionalmente, a pesquisa notou que a Regressão logística (23.4%) e redes neurais (21.0%) são as técnicas mais comumente utilizadas em estudos comparativos.

Por fim, o estudo determinou que a máquina de vetores de suporte (SVM) era um método de alto desempenho preditivo e baixo esforço computacional em comparação com outros métodos, sendo mais resiliente a *datasets* desbalanceados.

3.1.2 Statistical and machine learning models in credit scoring: A systematic literature survey

O artigo desenvolvido por DASTILE; CELIK; POTSANE traça a história da análise de crédito desde os anos 1950, começando com o método dos 5C's que avaliava caráter, capital, colateral, capacidade e condições de mercado. A ineficiência deste método

levou ao desenvolvimento dos *scorecards* de crédito, que usam dados de aplicação e de *bureau* para atribuir pontuações aos mutuários com base no histórico de pagamento. A regressão logística foi tradicionalmente usada para os *scorecards*, mas modelos sofisticados de aprendizado de máquina (ML) agora oferecem maior precisão. Os autores visam a revisar modelos estatísticos e de *Machine Learning* usados na análise de crédito de 2010 a 2018 e propor um framework para análise de crédito.

Uma pesquisa sistemática de literatura é empregada para revisar técnicas estatísticas e de ML na análise de crédito, abordando limitações e propondo um framework. Técnicas como Análise Discriminante Linear (LDA), Regressão Logística (LR), k-Nearest Neighbor (k-NN), Árvores de Decisão (DTs), Máquinas de Vetores de Suporte (SVMs), Redes Neurais Artificiais (ANNs), *Random Forests* (RFs) e modelos de aprendizado profundo são examinadas. O estudo incluiu 74 estudos primários de 2010 a 2018, selecionados de bases de dados como Google Scholar, Science Direct, IEEEXplore, ACM e Springer-Link. A pesquisa utiliza meta-análise com tabelas de resumo, gráficos de pizza e histogramas, focando no uso de conjuntos de dados, transparência dos modelos e comparação da literatura.

Os autores revisam técnicas de *Feature Selection* (FS) e *Feature Engineering* (FE) na análise de crédito. FS busca envolve selecionar um subconjunto de características usando métodos de filtro, wrapper e *embedded*. FE cria novas características a partir das existentes, utilizando técnicas como Análise de Componentes Principais (PCA) e Autoencoders. A distinção entre FS e FE é crucial; FS seleciona características originais enquanto FE cria novas. Diferentes artigos destacam que remover características redundantes pode melhorar o desempenho, mas FS nem sempre melhora a previsão (DASTILE; CELIK; POTSANE, 2020).

Técnicas para aumentar a transparência dos modelos na análise de crédito, como NeuroRule e Trepan, são detalhadas. NeuroRule usa uma rede neural para extrair regras compreensíveis, tornando decisões complexas e transparentes. Trepan induz árvores de decisão para aproximar previsões de classificadores, fornecendo uma estrutura interpretável. Essas técnicas visam a tornar os modelos de *Machine Learning* mais transparentes, crucial na análise de crédito, onde entender os motivos da rejeição de um empréstimo é essencial. A implementação garante que os modelos estejam alinhados com diretrizes éticas e padrões regulatórios.

Várias limitações das técnicas estatísticas como LDA, Naïve Bayes e LR são discutidas, destacando suas suposições. LDA assume distribuição normal multivariada, o que é violado por características categóricas. LR assume uma relação linear, nem sempre válida. Técnicas não paramétricas como k-Nearest Neighbor e árvores de decisão, e o desafio de interpretabilidade das ANNs, também são abordados. Os autores enfatizam a necessidade de processos de análise de crédito transparentes e sugerem alternativas para lidar com essas limitações, como a análise de sobrevivência para técnicas prospectivas.

Tendências emergentes em modelos de ML para análise de crédito são destacadas, notando a preferência regulatória por modelos transparentes. Métodos como extração de regras (Neurorule, Trepan) e SHAP (SHapley Additive exPlanations) fornecem explicações de modelos. Um estudo empírico usando Modelos Aditivos Generalizados com interações de características mostra que modelos interpretáveis podem igualar modelos de *Machine Learning* de alto desempenho.

3.2 Discussões concetuais acerca do tema

Nesta seção serão abordados trabalhos que discutem mais sobre o mercado de crédito e desafios enfrentados para a implementação de técnicas de *Machine Learning* e estatísticas, sendo utilizado três referências extraídas do trabalho desenvolvido por LOUZADA; ARA; FERNANDES:

- *Not If but When Will Borrowers Default* BANASIK; CROOK; THOMAS
- *'Lending by numbers': credit scoring and the constitution of risk within American consumer credit* HAND
- *Modelling consumer credit risk* MARRON

3.2.1 Evolução do Mercado de Crédito

Os três textos exploram coletivamente a evolução e as metodologias de avaliação de risco de crédito ao consumidor, destacando mudanças significativas de abordagens tradicionais baseadas em julgamento para modelos estatísticos sofisticados. O foco de BANASIK; CROOK; THOMAS está no papel crescente da inteligência artificial e do aprendizado de máquina na previsão de inadimplências de empréstimos, mostrando como essas tecnologias avançadas podem melhorar a precisão e a eficiência das avaliações de risco em comparação com os métodos tradicionais. De maneira semelhante, HAND analisa a transição do setor bancário de varejo do Reino Unido para sistemas automatizados de *credit scoring*, enfatizando os benefícios dos modelos estatísticos em relação ao julgamento humano.

MARRON examina o panorama do crédito ao consumidor nos EUA, traçando a progressão histórica de avaliações qualitativas para práticas de gestão burocrática e baseada na população, impulsionadas pelos princípios econômicos keynesianos e novas formas de consumo. Juntos, esses textos sublinham o papel fundamental dos avanços tecnológicos e dos marcos legislativos na formação das metodologias modernas de avaliação de risco de crédito.

Interessante notar que apesar de tratarem de diferentes países os três textos trouxeram uma tendência clara no início do século XX: A utilização de métodos de

aprendizado de máquina para trazer escalabilidade ao mercado de crédito. Em paralelo, um grande exemplo disso é a companhia Capital One, que bebeu nessa fonte e se utilizou dos *scores* de créditos e outras técnicas de aprendizado estatístico para trazer escalabilidade e custos reduzidos para sua operação bancária (PROVOST; FAWCETT, 2013).

3.2.2 Desafios enfrentados

Um tema chave nos textos são os desafios e limitações inerentes aos modelos de pontuação de crédito. BANASIK; CROOK; THOMAS discute os potenciais vieses e preocupações éticas associados aos modelos de IA e aprendizado de máquina, destacando a necessidade de transparência e justiça nesses sistemas avançados. HAND aborda questões como a precisão preditiva, a qualidade dos dados e os vieses nos dados históricos, sugerindo que abordagens centradas no cliente e modelos holísticos poderiam capturar melhor as relações complexas.

MARRON também destaca os riscos metodológicos, processuais e temporais associados aos modelos estatísticos, defendendo a necessidade de refinamento contínuo para mitigar esses desafios (MARRON, 2007). Tanto BANASIK; CROOK; THOMAS quanto MARRON reconhecem as implicações sociopolíticas mais amplas dos sistemas de pontuação de crédito, enfatizando a necessidade de supervisão regulatória para garantir práticas de empréstimo justas e equitativas.

Apesar dos textos serem datados de algumas decádas atrás, este é um tema de grande relevância atualmente, passando a ser vigiado de forma mais incisiva a partir de legislações dedicadas à privacidade dos dados (MATTIUZO, 2023). No Brasil, os dados utilizados para construção de *score* de crédito passaram a ser regulamentadas apenas em 2019, devido a um complemento a lei do cadastro positivo, estabelecendo que análise de risco de crédito não poderiam utilizar dados relacionados à origem social e étnica, à saúde, à informação genética, ao sexo e às convicções políticas, religiosas e filosóficas (Brasil, 2011).

3.2.3 Implicações Sociopolíticas e Econômicas

O conceito de colonização do risco e suas implicações são minuciosamente examinados por MARRON, que discute as extensões temporais, espaciais e funcionais da avaliação de risco na indústria de crédito ao consumidor. Isso é contrastado com BANASIK; CROOK; THOMAS que destaca a natureza dinâmica dos modelos de avaliação de risco impulsionados por IA, que continuamente aprendem e se adaptam a novos dados, potencialmente revolucionando a gestão do risco de crédito.

HAND foca nas ferramentas preditivas e nas medidas de desempenho críticas para a avaliação da credibilidade no mercado bancário de varejo do Reino Unido. MARRON investiga ainda a transformação da percepção de risco de uma potencial perda para uma

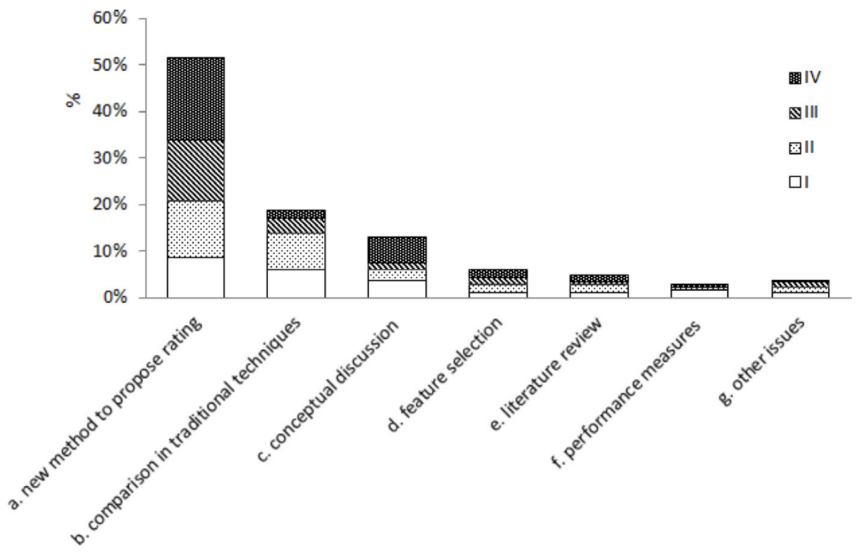

Figura 4 – Diferentes tipos de categorias de artigos produzidos relacionados *credit score*elaborado por (LOUZADA; ARA; FERNANDES, 2016).

empresa lucrativa, como visto no desenvolvimento de modelos de pontuação de lucro e precificação baseada em risco.

Essa mudança ilustra a natureza evolutiva da gestão de risco de crédito, onde os credores se adaptam às mudanças do mercado ao abraçar clientes de maior risco para potencial lucratividade, um tema ressonante nos três textos.

3.3 Comparação de técnicas tradicionais e novos métodos propostos em Aprendizado de máquina

A utilização de novos métodos para aplicações de *credit score* e comparação de técnicas tradicionais é um dos principais campos de pesquisa da área e responsável por cerca de 70% dos trabalhos avaliados por LOUZADA; ARA; FERNANDES, mais detalhes na figura 4.

Dessa forma, para tornar a discussão de novos métodos mais objetiva, dois artigos foram selecionados para trazer métodos inovadores e que possuíam relação com o trabalho proposto. Estes artigos são:

- A data driven ensemble classifier for credit scoring analysis. HSIEH; HUNG
- Benchmarking state-of-the-art classification algorithms for credit scoring BAESENS T VAN GESTEL; VANTHIENEN

O primeiro artigo propõe uma abordagem inovadora para a análise de crédito, utilizando um sistema de classificação em conjunto que combina várias técnicas de classificação para melhorar a precisão e o desempenho da generalização (HSIEH; HUNG,

2010). Ele introduz uma etapa de pré-processamento de classificação por classe para criar clusters homogêneos, o que aumenta a eficiência dos classificadores em conjunto. Este método é validado usando um conjunto de dados reais de crédito alemão (HOFMANN, 1994), demonstrando uma melhoria significativa na precisão de 78,46% para 89,16% após o pré-processamento. O classificador em conjunto, construído com um sistema de votação ponderada por confiança, supera os classificadores individuais, oferecendo uma solução abrangente para aplicações práticas de análise de crédito.

Em contraste, o segundo artigo concentra-se em uma avaliação abrangente de várias técnicas de classificação para modelos de análise de crédito, com o objetivo de distinguir com precisão bons solicitantes de empréstimo dos maus (BAESENS T VAN GESTEL; VANTHIENEN, 2003). O estudo inclui técnicas estatísticas tradicionais, modelos não paramétricos e redes neurais. Entre os métodos avaliados estão a regressão logística (LOG), análise discriminante linear (LDA) e quadrática (QDA), programação linear (LP), k-vizinhos mais próximos (KNN), máquinas de vetor de suporte (SVM), máquinas de vetor de suporte de mínimos quadrados (LS-SVM), redes neurais (NN), e classificadores Bayesianos, incluindo o ingênuo de Bayes e o ingênuo de Bayes aumentado por árvore (TAN). Os parâmetros dos modelos SVM e LS-SVM foram ajustados utilizando uma busca em grade, enquanto para as redes neurais foi usado o framework de evidência Bayesiana. As árvores de decisão foram ajustadas utilizando algoritmos de poda e discretização específicos.

Os resultados do estudo revelam que tanto classificadores não lineares como o RBF LS-SVM e redes neurais (NN), quanto classificadores lineares mais simples, como a regressão logística (LOG) e a análise discriminante linear (LDA), apresentaram desempenho muito bom (BAESENS T VAN GESTEL; VANTHIENEN, 2003). Especificamente, o RBF LS-SVM obteve a melhor média de classificação para a medida de porcentagem de observações corretamente classificadas (PCC), com outros métodos como LOG, LP, Lin LS-SVM, NN, C4.5 dis e KNN100 apresentando desempenhos estatisticamente similares. No que diz respeito à AUC, o classificador NN teve a melhor média de classificação, com LDA, LOG, RBF LS-SVM, Lin LS-SVM e TAN não sendo estatisticamente inferiores a ele. Esses resultados indicam que, apesar da complexidade adicional dos classificadores não lineares, as técnicas lineares mais simples ainda são altamente eficazes para muitos conjuntos de dados de análise de crédito, que são tipicamente fracamente não lineares.

Além de destacar a competitividade entre os diferentes métodos de classificação, o estudo também sublinha a importância de utilizar métricas de desempenho apropriadas para avaliar os classificadores (BAESENS T VAN GESTEL; VANTHIENEN, 2003). A pesquisa enfatiza que a porcentagem de observações corretamente classificadas (PCC) pode não ser suficiente por si só, especialmente em casos onde os custos de erros de classificação variam. Assim, a AUC é sugerida como uma métrica complementar, fornecendo uma

visão mais robusta do desempenho do classificador independentemente da distribuição das classes. A análise dos resultados através dessas métricas revelou que, embora classificadores complexos como RBF LS-SVM e redes neurais ofereçam excelente desempenho, métodos mais simples como LOG e LDA são também altamente eficazes, oferecendo uma alternativa prática e competitiva para a análise de crédito em ambientes reais.

Outra diferença notável está nos critérios de avaliação e nos conjuntos de dados utilizados. O primeiro artigo usa um único conjunto de dados do Repositório de Aprendizado de Máquina da UCI (HOFMANN, 1994), focando em melhorar a precisão e a interpretabilidade do seu classificador em conjunto proposto (HSIEH; HUNG, 2010). Em contraste, o segundo artigo emprega uma gama mais ampla de conjuntos de dados de várias fontes, incluindo instituições financeiras de múltiplos países, e avalia o desempenho dos classificadores usando tanto PCC quanto AUC (BAESENS T VAN GESTEL; VANTHIENEN, 2003). Essa abordagem extensa de avaliação comparativa fornece uma compreensão mais generalizada do desempenho dos classificadores em diferentes cenários de análise de crédito, destacando a competitividade e o potencial de técnicas de classificação simples e complexas.

3.4 Foundation models e o futuro no mercado de crédito

Até então diferentes trabalhos foram discutidos que se valiam de técnicas de *Machine Learning* e *Deep Learning* tradicionais (e.g. NN). No entanto, é cada vez mais comum a utilização de modelos com grande escala de parâmetros, que são baseados em *deep neural networks* e aprendizado supervisionado (BOMMASANI *et al.*, 2022).

Os *Foundation Models* representam uma evolução significativa no desenvolvimento da IA, marcando uma mudança em relação aos modelos tradicionais de *Machine Learning* e *Deep Learning*. Esses modelos surgiram da crescente necessidade de gerenciar e processar grandes e variadas quantidades de dados de maneira eficiente. Diferentemente de seus predecessores, os modelos fundacionais são projetados para serem adaptáveis a uma ampla gama de tarefas sem a necessidade de re-treinamento extensivo, como ilustrado na figura 5. Essa adaptabilidade contrasta fortemente com os modelos de ML e DL anteriores, que muitas vezes eram limitados pela necessidade de dados específicos de domínio e engenharia de características para resolução de problemas.

Pensando para o contexto de crédito isso abre diferentes possibilidades, tais como:

- Modelos de risco que permitem avaliar o comportamento do cliente com a instituição financeira
- Modelos de renda do cliente
- Previsão de *churn*

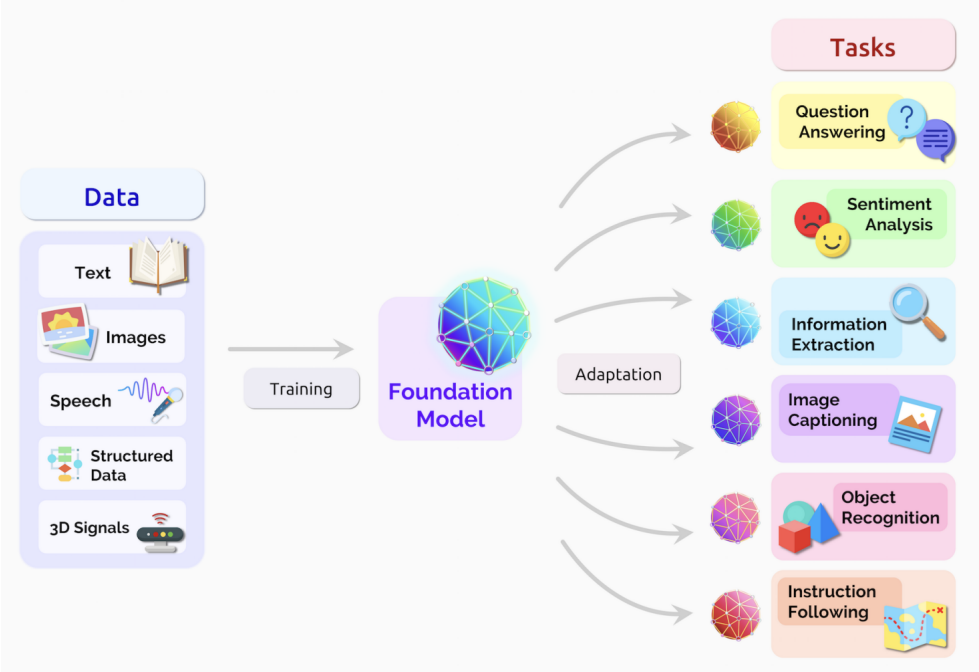

Figura 5 – Ilustração sobre o funcionamento do *Foundation Model*, sendo capaz de centralizar as informações de diferentes formatos e ser adaptado para diversas aplicações BOMMASANI *et al.*

- Modelos de conversão

A lista acima não é exaustiva, mas consegue ilustrar a grande gama de possibilidades que esses modelos podem trazer para o mercado de crédito. O tema é ainda pouco explorado para literatura com o viés do mercado de crédito, sendo mais discutido para aplicações de linguagem, medicina e robótica (BOMMASANI *et al.*, 2022). No entanto, algumas empresas já fornecem serviços relacionados a *Foundation Models* no mercado brasileiro e mundial:

- HyperPlane (adquirida pelo Nubank em 2024): Startup de inteligência artificial que se destaca na modelagem de decisões de crédito, superando algoritmos tradicionais através do uso de dados não estruturados e redes neurais avançadas. Antes da aquisição a empresa possui quatro produtos principais, sendo eles: preditor de renda, predição de evasão de clientes, conversão e risco de crédito.

- Avra: Startup que utiliza inteligência artificial para ajudar empresas a tomar decisões mais assertivas sobre contratos com PMEs, analisando o risco de crédito e prevendo o *lifetime value* de cada cliente. A startup oferece quatro produtos principais, incluindo um score de crédito para PMEs, predição de LTV, segmentação de audiência e venda de inteligência de modelo, diferenciando-se no mercado por sua abordagem única e especializada.

4 PROPOSTA E DESENVOLVIMENTO

4.1 Motivação e Proposta

Este trabalho visa explorar diferentes modelos de *Machine Learning* e suas aplicações em um *dataset* público baseado na população Alemã (HOFMANN, 1994), observando o impacto de diferentes dos modelos (e.g. LightGBM, XGBoost) em diferentes métricas de avaliação de classificadores para cálculo do score de crédito (i.e. AUC e Acuracidade) e avaliação de regressores para valor do crédito (i.e. MAPE).

Apesar do *dataset* mencionado anteriormente ter sido explorado em diferentes trabalhos e contextos como discutido no capítulo anterior, ainda há uma lacuna na exploração mais profunda dos hiperparâmetros. Além disso, muitos dos trabalhos utilizam todas as *features* disponibilizadas no *dataset*, porém nem todas elas estão disponíveis para a resolução de um problema de *score* de crédito, indicando também *data leakage* (HUYEN, 2022). Por exemplo, uma dos atributos presentes na tabela é *Credit amount*, que consiste no valor de crédito concedido ao cliente. Para o contexto do problema de nota de crédito, esse é um tipo de informação obtida somente após o cálculo do *score* de crédito, pois é necessário primeiro entender o nível de risco do cliente, para então o designá-lo para a melhor politíca ou rejeitar a sua requisição. Sendo assim, caso seja utilizado para o cálculo do mesmo, fará com que o resultado seja enviesado e, principalmente, não possua bom desempenho quando aplicado a situações reais.

Por fim, também será realizada uma discussão de forma mais abrangente sobre o ciclo de implementação dos modelos utilizados (i.e. MLOps), entendendo os custos e requisitos de aplicação para modelos de *Machine Learning* e *Deep Learning*.

4.2 Metodologia

4.2.1 Extração dos dados

Como mencionado anteriormente, o *dataset* escolhido para execução desse trabalho será baseado na população Alemã da década de 90 (HOFMANN, 1994), o mesmo será extraido do repositório da UC Irvine.

4.2.2 Pré-processamento e exploração dos dados

Afim de realizar melhorias na interpretabilidade dos resultados e bem como seu processamento por parte dos modelos, será necessário realizar diferentes sub-etapas de pré-processamento descritas abaixo:

1. Renomeação das *features* com base na documentação proposta no repositório (e.g. "Attribute1": "Status of existing checking account")

2. Verificação de dados faltantes

3. Separação em *features* catégoricas e numéricas para posterior pré-processamento, resultando na tabela 1

4. *Mapping* das *features* categóricas e seus valores para maior interpretabilidade dos resultados (e.g. Question: "Foreign worker", Answer: A201: yes, A202: no)

Feature Type	**Features**
Numérica	Duration in month, Credit amount, Installment rate in percentage of disposable income, Present residence since, Age in years, Number of existing credits at this bank, Number of people being liable to provide maintenance for
Categórica	Status of existing checking account, Credit history, Purpose, Savings account/bonds, Present employment since, Personal status and sex, Other debtors / guarantors, Property, Other installment plans, Housing, Job, Telephone, Foreign worker

Tabela 1 – Features numéricas e categóricas

Após essa primeira etapa do pré-processamento dos dados será feita a análise exploratória para entender a distribuição dos atributos e correlações, principalmente para melhor entendimento das técnicas de *encoding* categóricas que serão utilizadas. Com esse entendimento será feito um segundo pré-processamento, já voltado para o *input* nos modelos.

1. Aplicação de *Label Encoder* para variáveis ordinais

2. Aplicação de *OneHot Encoder* para variáveis não ordinais

3. Aplicação de SMOTE para balanceamento de classes no *dataset*

4.2.3 Modelagem e refinamento dos modelos

4.2.3.1 Modelos de Classificação

Como elucidado anteriormente, será feito a avaliação dos modelos no contexto do *score* de crédito (i.e. se um indivíduo é bom ou mal pagador). Os modelos abaixo serão avaliados e os que possuírem melhor resultado serão utilizados para um refinamento a posteriori. As métricas utilizadas para essa avaliação são AUC e acuracidade. Os modelos utilizados são:

- LogisticRegression
- GaussianNB
- KNN
- LinearSVC
- SVC
- Random Forest
- GBM
- XGBoost
- LightGBM

Os modelos acima foram escolhidos por serem utilizados extensivamente na literatura, conforme discutido no capítulo anterior. Para o refinamento dos valores, bem como entendimento da performance e impacto das variáveis dos modelos será feito um *Grid Search* nos modelos que possuírem melhor desempenho.

4.2.3.2 Modelos de Regressão

Para avaliação da performance do modelo para o cálculo do limite de crédito concedido, o mesmo *dataset* será utilizado, incluindo a variável target da sub-seção anterior. Para essa aplicação os seguintes modelos serão utilizados:

- Decision Tree
- Random Forest
- GBM
- XGBoost
- LightGBM

Similar aos modelos de classificação, o *Grid Search* será utilizado para refinamento dos modelos que obtiveram melhor performance anteriormente.

4.2.3.3 Interpretabilidade

Afim de melhorar a interpretabilidade e entender o impacto *features* específicas no desempenho dos modelos, a biblioteca SHAP será utilizada (LUNDBERG; LEE, 2017a) e seus resultados serão discutidos a fim de checar se de fato os resultados encontrados fazem sentido perante ao problema proposto.

5 AVALIAÇÃO EXPERIMENTAL

5.1 Conjuntos de Dados

Como mencionado no capítulo de metodologia, o conjunto de dados de *scores* de crédito Alemão se encontra originalmente codificado, onde o nome dos atributos e instâncias eram de difícil compreensão para entendimento do problema. Após um primeiro pré-processamento, é possível realizar uma análise exploratória e ter um conhecimento melhor acerca dos dados.

No total o *dataset* contém 1.000 registros, os quais são divididos nas seguintes *features* e *target*:

- Attribute 1: Status of existing checking account
- Attribute 2: Duration in month
- Attribute 3: Credit history
- Attribute 4: Purpose
- **Attribute 5: Credit amount**
- Attribute 6: Savings account/bonds
- Attribute 7: Present employment since
- Attribute 8: Installment rate in percentage of disposable income
- Attribute 9: Personal status and sex
- Attribute 10: Other debtors / guarantors
- Attribute 11: Present residence since
- Attribute 12: Property
- Attribute 13: Age in years
- Attribute 14: Other installment plans
- Attribute 15: Housing
- Attribute 16: Number of existing credits at this bank
- Attribute 17: Job

- Attribute 18: Number of people being liable to provide maintenance for
- Attribute 19: Telephone
- Attribute 20: Foreign worker
- **Target1: Risk**

Para resolução do problema de classificação, a variável de Risco será escolhida como target, enquanto que para a resolução do problema de regressão a variável *Credit amount* será utilizada como target.

Importante mencionar que para o problema de classificação a *feature* de *Credit amount* não foi utilizada para evitar *data leakage* (HUYEN, 2022). Os valores de crédito disponíveis para o cliente são obtidos a partir do momento em que a política de crédito possui um *score* de risco do cliente. Assim, essa é uma informação que não estaria disponível para novos clientes, impossibilitando o cálculo da métrica de risco.

5.1.1 Exploração dos dados

Após o mapeamento dos dados foi possível realizar a exploração das diferentes variáveis e entender melhor o conjunto de dados utilizados. Seguem abaixo as principais análises exploratórias realizadas, priorizando a ótica de risco, crédito disponível e sua correlação com os atributos dos conjuntos de dados.

5.1.1.1 Impacto da propriedade nos valores de crédito e avaliação de risco

Como pode ser observado em 6, os clientes que possuem propriedade, carro ou outros ativos (que não sejam ativos financeiros) possuem menor principal disponível (i.e. crédito disponível) enquanto os clientes, nestas categorias, de menor risco possuem na média um valor de crédito menor que os clientes de maior risco. Além disso, é possível notar uma grande quantidade de *outliers* para os clientes de baixo risco nestas categorias, mostrando que há uma grande variabilidade nos valores de crédito concedidos a esses clientes.

Olhando para a categoria dos clientes que não possuem propriedade ou não se sabe, é possível notar que o grupo possui maiores níveis de crédito quando comparado às outras categorias, tanto para clientes de baixo e alto risco.

De forma geral, era esperado que clientes com ativos reais (e.g. real estate e outras categorias) possuíssem um acesso a valores de crédito mais significativos que o grupo que não possui nenhuma propriedade, dado sua possibilidade de ter ativos para o pagamento. No entanto, devido a alta presença de *outliers* para clientes de baixo risco nas categorias de ativos reais, também é possível inferir que há uma alta variabilidade de clientes nessas categorias. Essa alta variabilidade também pode ser interpretada como uma oportunidade

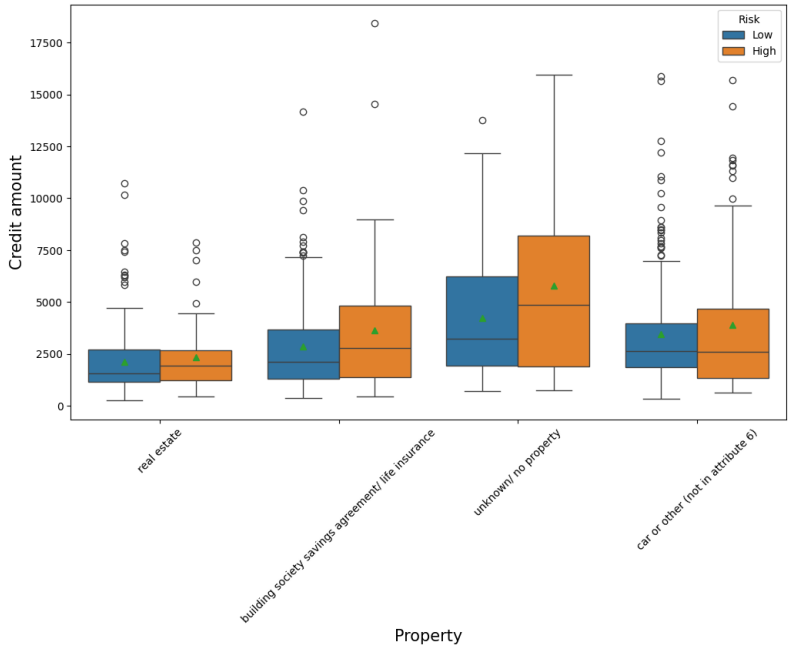

Figura 6 – Boxplot com os atributos de crédito disponível e tipo de propriedade, segmentados por risco

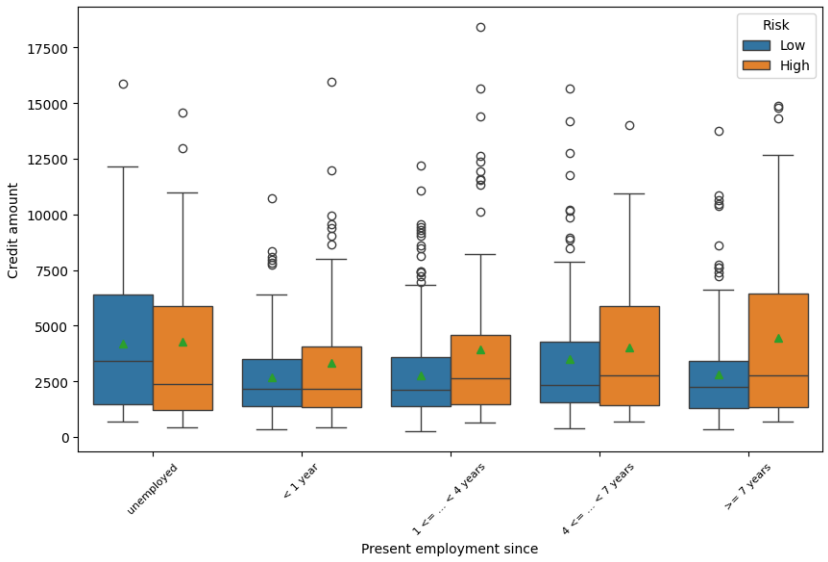

Figura 7 – Boxplot com os atributos de crédito disponível e tipo de tempo de emprego, segmentados por risco

para a métrica de risco, que poderia possuir mais subcategorias (i.e. maior segmentação comparada ao risco baixo/alto).

5.1.1.2 Impacto do tempo de emprego nos valores de crédito e avaliação de risco

De acordo com a figura 7, é possível notar que clientes que estão desempregados possuem acesso a valores de crédito similares a clientes que possuem longo histórico em seus empregos. Além disso, para categoria de clientes desempregados, a média de crédito

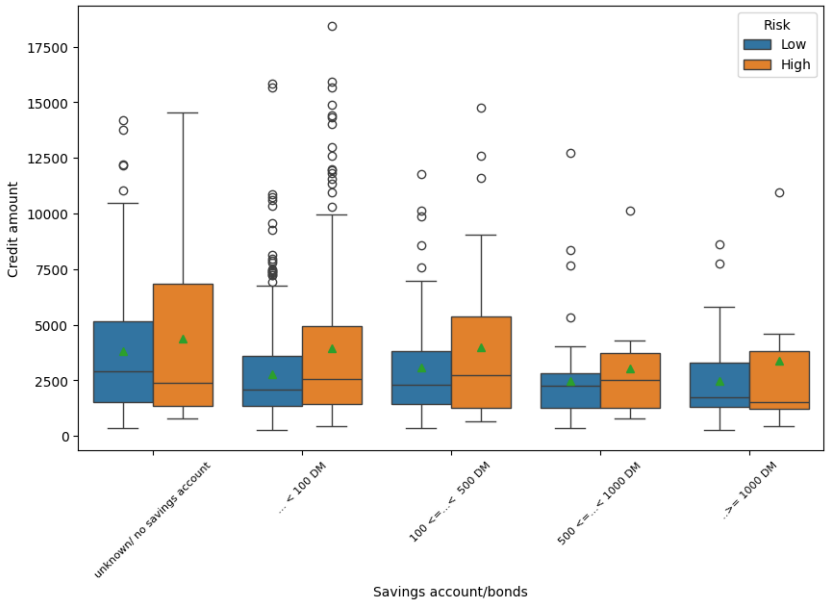

Figura 8 – Boxplot com os atributos de crédito disponível e investimentos, segmentados por risco. Valores ordenados por Deutsche Mark (DM), moeda utilizada em 1994.

consentido é similar entre as diferentes categorias de risco, mostrando pouca diferenciação entre os grupos de risco para essa categoria.

Para as demais categorias, é possível notar que o valor de crédito consentido é maior para aqueles clientes de alto risco, enquanto os de baixo risco possuem alta variabilidade dos dados, tendo portanto grande número de *outliers*. Conclusão similar a seção anterior, indicando que a métrica de risco possa não estar agrupando de forma efetiva o conjunto de dados.

5.1.1.3 Impacto de investimentos nos valores de crédito e avaliação de risco

Para todas as diferentes categorias, os clientes com maior risco possuem, na média, valores de crédito maiores do que os clientes de baixo risco. Em específico, os clientes que não possuem investimento ou sua posição é desconhecida, possuem maiores valores de crédito. Apenas algumas categorias possuem alta variabilidade, estando concentradas nos clientes que possuem investimentos até 100 DM.

5.1.1.4 Impacto do histórico de dívidas nos valores de crédito e avaliação de risco

As diferentes categorias de histórico de dívida mostram que na média, para clientes de baixo risco, o grupo de clientes que possui bom histórico de crédito (i.e. "no credits taken/all credits paid back dully"). No entanto, não há um padrão linear para as diferentes categorias e o mesmo ocorre para clientes de alto risco.

Outro ponto que é possível notar, para clientes que já possuem dívidas que foram

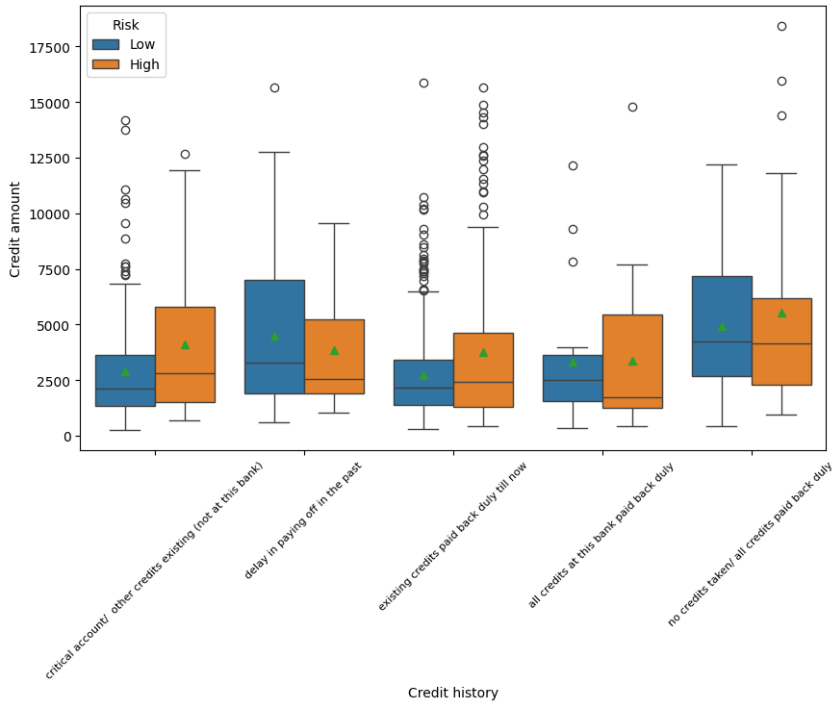

Figura 9 – Boxplot com os atributos de crédito disponível e histórico de dívida, segmentados por risco.

pagas até então, há uma grande variabilidade dos dados. Assim como as seções anteriores, isso pode indicar que há uma oportunidade de melhoria no *score* de risco, que poderia possuir um melhor agrupamento dos clientes.

5.1.1.5 Balanceamento de classes

Risk	**Count**
Low	700
High	300

Tabela 2 – Distribution of Risk Categories

Para o problema de classificação é extremamente importante entender a separação das classes e a quantidade de instâncias que pertencem a cada uma. Por meio da tabela 4, é possível notar que há uma desproporção dos dados, o que pode prejudicar a performance do modelo. Para mitigar esse problema, será utilizado a técnica SMOTE, conforme discutido no capítulo anterior.

5.2 Configuração Experimental

5.2.1 Preparação dos dados para aplicação do modelo

5.2.1.1 Transformação dos dados

Após a exploração dos dados e entendimento do contexto por trás das variáveis, a normalização e *encoder* foram aplicados aos atributos. As seguintes regras foram utilizadas:

1. Variáveis numéricas: *Standard Scaler*
2. Variáveis categóricas:
   a) Atributos exceto por *Personal status* e *Other installment plans*: *Label Encoding*
   b) *Personal status* e *Other installment plans*: *One-Hot Encoding*

Para os atributos exceto por *Personal status* e *Other installment plans*, é necessário a utilização de *Label Encoding* devido a relação hierárquica que os dados possuem. Por exemplo: Para o problema de *score* de risco e valor de crédito concedido, um cliente desempregado necessariamente está ranqueado numa posição inferior a um cliente com muitos anos empregado. Enquanto que *Personal status* não possui nenhuma relação hierárquica.

5.2.1.2 Balanceamento de classes e divisão treino/teste

Para aplicação do SMOTE as seguintes configurações foram utilizadas:

Parameter	**Value**
seed	42
sampling_strategy	'auto'
k_neighbors	5
random_state	42

Tabela 3 – SMOTE Parâmetros

Ao fim da aplicação do SMOTE, a distribuição final resultou em:

Risk	**Count**
High	700
Low	700

Tabela 4 – Distribuicão de Risco após aplicação do SMOTE

A aplicação do SMOTE nos modelos utilizados neste trabalho geraram ganhos de performance relevantes para modelos baseados em Árvores de decisão, chegando a ter uma melhoria de 12% na acurácia e 30% no AUC, estando alinhado com relatos da literatura (LOUZADA; ARA; FERNANDES, 2016).

Para a separação dos dados em grupos de teste e treino, foi utilizada a técnica de *hold out* com uma proporção de 20% para o agrupamento de teste.

5.3 Resultados e Discussões

5.3.1 Classificação: Score de risco

5.3.1.1 Resultados dos modelos

Com base nos modelos descritos na seção de metodologia para classificação, o resultado abaixo foi encontrado para acurácia dos modelos:

Modelo	**Acurácia (%)**	**AUC (%)**
LogisticRegression	75.71%	75.68%
GaussianNB	73.57%	73.05%
KNN	71.79%	71.08%
LinearSVC	75.71%	75.75%
SVC	75.71%	75.45%
CART	72.14%	72.10%
Random Forest	85.43%	85.50%
GBM	81.79%	81.48%
XGBoost	86.79%	86.93%
LightGBM	85.71%	85.78%

Tabela 5 – Acurácia e AUC Modelos de Classificação de Risco

O modelo XGBoost foi o que possuiu melhor acurácia comparado ao outros modelos e em linha com os resultados reportados para o conjunto de dados em (HOFMANN, 1994), como pode ser observado na tabela 6.

Considerando que o conjunto de dados inicial possuía 70% de clientes com bom risco de crédito, isso implica que um modelo com bom desempenho precisaria, no mínimo, ter 70% de acurácia. Sendo assim, os modelos de Regressão Logistica, GaussianNB, KNN e LinearSVC performaram pouco comparado a esta referência, enquanto modelos baseados em árvores de decisão performaram bem (na média 15 p.p. acima dos 70%).

Model Name	**Acc Lower Bound (%)**	**Acc Upper Bound (%)**
Logistic Regression	70.00%	80.80%
Neural Network (MLP)	58.40%	70.40%
Random Forest	72.80%	83.20%
SVC	64.80%	76.00%
XGBoost	69.20%	80.00%

Tabela 6 – Benchmark retirado de (HOFMANN, 1994)

Com base nos resultados obtidos e os valores de referência, é possível inferir que modelos baseados em árvores de decisão possuem melhor acurácia, enquanto modelos mais complexos como Redes Neurais, têm baixo desempenho. Isso pode ser atribuido ao baixo número de instâncias do conjunto de dados Alemão (1.000 instâncias), fazendo com que o modelo não consiga ter um bom aprendizado e capacidade de generalização (ALZUBAIDI L. *et al.*, 2023).

O trabalho desenvolvido por HAMORI *et al.* possui resultados semelhantes ao resultados encontrados anteriormente, onde o autor aponta ainda que modelos de redes neurais trazem uma camada adicional de complexidade devido ao refinamento dos hiperparâmetros ser mais complexo que comparado às técnicas de boost (i.e. função de ativação, número de camadas intermediárias e inclusão de *dropout*) (HAMORI *et al.*, 2018).

Além dos pontos trazidos por HAMORI *et al.* acerca de modelos de redes neurais, também é importante destacar que para uma aplicação real estes tipos de modelo também exigem uma operacionalização complexa, devido a necessidade de GPUs para computação dos dados, enquanto modelos de árvore de decisão podem utilizar apenas CPUs (HUYEN, 2022).

Tipo da máquina	**CPU Cores**	**Memória (GB)**	**GPUs**	**Preço por hora (US$/h)**
t4g.large	2	8	0	0.0672
t4g.xlarge	4	16	0	0.1344
t4g.2xlarge	8	32	0	0.2688
g4dn.xlarge	4	16	1	0.526
g4dn.2xlarge	8	32	1	0.752
g4dn.4xlarge	16	64	1	1.204

Tabela 7 – Comparação de preços por hora da AWS, extraídos do (AWS, a) e (AWS, b)

Para efeitos de comparação, as máquinas que possuem GPU e mesmas configurações de CPU e Memória são em média 236% mais caras.

Apesar dos pontos trazidos anteriormente, a aplicação de redes neurais para *scores* de risco está se tornando cada vez mais comum e com um conjunto de dados extenso pode ter acurácia e robustez melhores do que técnicas tradicionais (ORESKI; ORESKI; ORESKI, 2012). Isso, associado com os preços mais altos das GPUs, reforçam o ponto que nem todas as aplicações (i.e. problema a ser resolvido e conjunto de dados disponíveis) permitem a utilização de redes neurais.

5.3.1.2 Refinamento por meio do GridSearch

A fim de ter um melhor entendimento do modelo XGBoost e realizar uma otimização no seu desempenho, o GridSearch foi utilizado contendo os parâmetros da tabela 8. Ao total 324 candidatos foram testados com 20 *folds*.

Hiperparâmetros	**Valores**	**Melhores Valores**	**Valores padrões**
`n_estimators`	50, 100, 150	150	100
`subsample`	0.6, 1.0, 1.5	0.6	1.0
`max_depth`	4, 6, 8, 10	6	6
`learning_rate`	0.1, 0.01, 0.05	0.1	0.3
`booster`	`gbtree`, `gblinear`, `dart`	`dart`	`gbtree`

Tabela 8 – Hiperparâmetros utilizados no GridSearch, Melhores Valores, e Valores Padrões para XGBoost

Com base nos melhores parâmetros encontrados descritos em 8, os resultados da tabela 9 foram encontrados.

Métrica	**Valores**
Accuracy	0.8643
Precision	0.8642
Recall	0.8643
F1 Score	0.8642
Cohen Kappa Score	0.7272
Matthews Corrcoef	0.7273
ROC AUC Score	0.8633

Tabela 9 – Métricas de performance do modelo

O modelo demonstra um desempenho robusto, alcançando uma acurácia de 86.43% e valores bem equilibrados de precisão, recall e F1, todos em torno de 86.42%. Esses métricas indicam que o modelo é eficaz na classificação correta tanto de instâncias positivas quanto negativas, minimizando o risco de classificações incorretas. A alta precisão e recall sugerem que o modelo é confiável na identificação de casos positivos, mantendo uma baixa taxa de falsos positivos.

O ROC AUC Score de 86.33% mostra que o modelo tem uma forte capacidade de distinguir entre classes em diferentes configurações de limiar. Coletivamente, essas métricas sugerem que o modelo não é apenas preciso, mas também consistente e robusto em suas capacidades preditivas.

Comparando os dados obtidos na tabela 5 e 9, nota-se que o modelo otimizado via GridSearch teve performance ligeiramente pior que o modelo inicial. A principal hipótese para isso é o baixo número de instâncias, o que faz com que o modelo não consiga melhorar a sua capacidade de generalização e naturalmente estar em torno do sua acurácia máxima

Figura 10 – Gráfico de SHAP para melhor modelo de classificação encontrado.

global. Isso fica mais evidente quando olhamos para o método de validação cruzada utilizado pelo GridSearch, o *K-fold*, que possui performance similar ao modelo anterior, que utilizou *holdout*, uma técnica que pode ter enviesamento dos dados.

5.3.1.3 Explicabilidade do modelo

O gráfico de SHAP, figura 10, fornece uma visualização detalhada de como várias *features* influenciam as previsões do modelo, ilustrando tanto a magnitude quanto a direção do impacto de cada característica. O gradiente de cores ajuda a entender se valores mais altos ou mais baixos de uma característica contribuem positiva ou negativamente para a previsão, oferecendo uma visão aprofundada do processo de tomada de decisão do modelo. Além disso, os principais atributos são ordenados de forma descendente, onde "Status of Existing Checking Account"é o mais relevante para o modelo. Para o problema de classificação, clientes de baixo risco possuem valores iguais a 1 e alto risco são 0.

A "Status of Existing Checking Account"destaca-se como a característica mais influente nas previsões do modelo. Valores altos dessa característica tendem a aumentar a previsão, enquanto valores mais baixos têm o efeito oposto. Da mesma forma, "Duration in Month"é outra característica importante, onde durações mais longas geralmente têm um impacto negativo na previsão, aumentando o risco, enquanto durações mais curtas afetam positivamente, diminuindo o risco. E por fim, "Purpose", cujos valores mais altos tendem a afetar positivamente o resultado, diminuindo o risco do cliente.

Figura 11 – Gráfico de SHAP para Status of Existing Checking Account.

5.3.1.3.1 Status of Existing Checking Account

Para melhor visualização da figura 11, considerar a seguinte ordem dos dados:

1. No checking account
2. < 0 DM
3. >= 200 DM
4. 0 <= ... < 200 DM

Com base no gráfico, podemos observar que para clientes que não possuem conta e/ou possuem saldo devedor os valores de SHAP tendem a decrescer a predição (i.e. aumentar o Risco). Portanto, para esses clientes que não possuem saldo bancário e/ou são devedores, o seu risco é maior, o que está em linha com o ponto de vista de negócios.

Para clientes que possuem uma quantia entre 0 e 200 DM, há um aumento da predição, indicando que o grupo tende a ser de baixo risco, pois possui capacidade reservas financeiras que poderiam ser utilizadas eventualmente para pagar o empréstimo.

Por último, temos o grupo com valores acima ou iguais a 200 DM, que também tem uma tendência a ter menor risco, no entanto, menor que o grupo de 0 e 200 DM. Do ponto de vista de negócios, este é um grupo que deveria ter o menor risco, dado que o principal ponto avaliado aqui é a capacidade financeira de pagamento do empréstimo a ser tomado. Porém, a observação indica o contrário e isso pode ser atribuído ao baixo número de dados para esse grupo. Como pode ser visto na figura 11, este grupo é o que possui menor quantidade de *data points*.

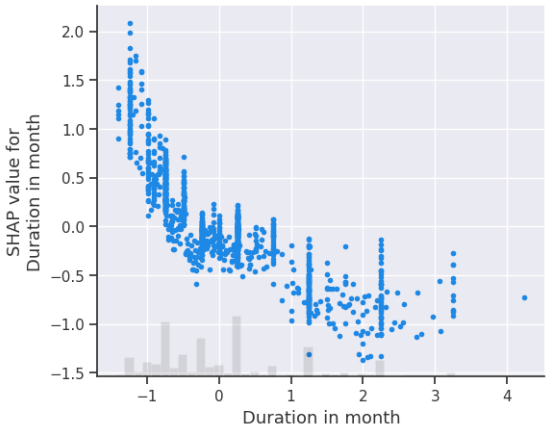

Figura 12 – Gráfico de SHAP para Duração dos empréstimos.

5.3.1.3.2 Duration in month

Analisando a figura 12 podemos perceber que quanto maior a duração do empréstimo, maior tende a ser o risco (i.e. Risco alto = 0 e Risco baixo = 1). Essa dinâmica está em linha com o observado na literatura, que aponta que *durations* mais longas tendem a expor as instituições financeiras a maiores problemas e perdas para empréstimo sem garantia (JIMENEZ; SAURINA, 2003), principalmente em cenário de preços maiores (i.e. taxas de juros).

Embora o conjunto de dados não inclua a informação específica da taxa aplicada a cada cliente, podemos estimar a taxa de juros utilizada em 1994 com base na taxa de juros real do país, que era de 9,1% ao ano (World Bank, 2024b), e na inflação de 2,7% ao ano (World Bank, 2024a), resultando em uma taxa nominal de aproximadamente 12% ao ano. Além disso, considerando o *spread* médio da época, que variava entre 2% e 6% (World Bank, 2024c), aplicado à taxa de juros básica, podemos inferir que esse cenário contribui para o comportamento observado no *dataset*, onde há uma predominância de empréstimos de curto prazo, possivelmente devido às taxas mais elevadas que poderiam ser repassadas aos clientes.

Além disso, é possível observar que a distribuição possui inclinação positiva, possuindo mais dados para empréstimos de curto prazo e poucos casos em que são concedidos empréstimos mais longos. Isso também pode fazer com que a amostra seja enviesada, fazendo com que o modelo consiga realizar uma separação clara entre os diferentes grupos de risco.

Olhando mais a fundo para a distribuição dos dados, figura 13, nota-se que a média e mediana da duration é mais alta quanto maior for o risco, o que explica o comportamento encontrado na figura 12. Além disso, também há desbalanceamento dentro das diferentes razões dos empréstimos, o que apesar de estar em linha com o apontado pela literatura, também faz com que a métrica de *Duration* seja enviesada pelos clientes de alto risco.

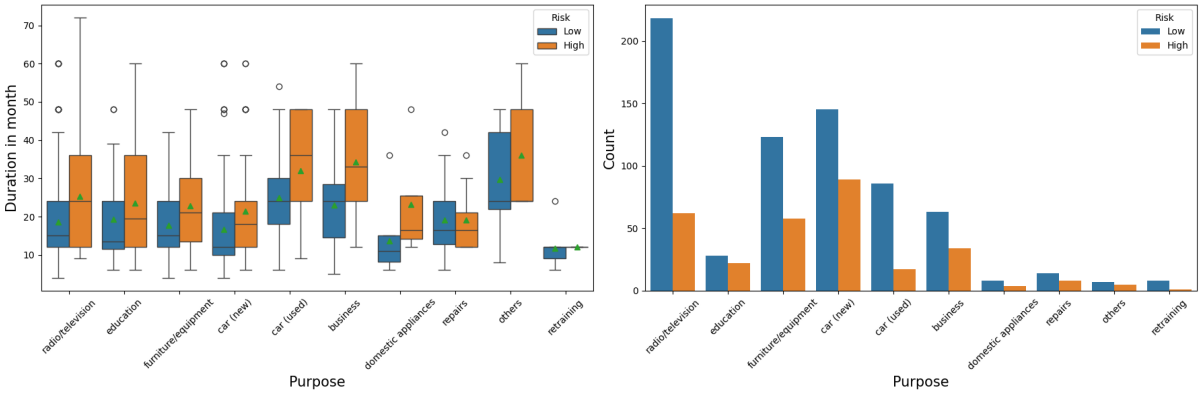

Figura 13 – Gráfico de boxplot para Duration, Purpose e Risco.

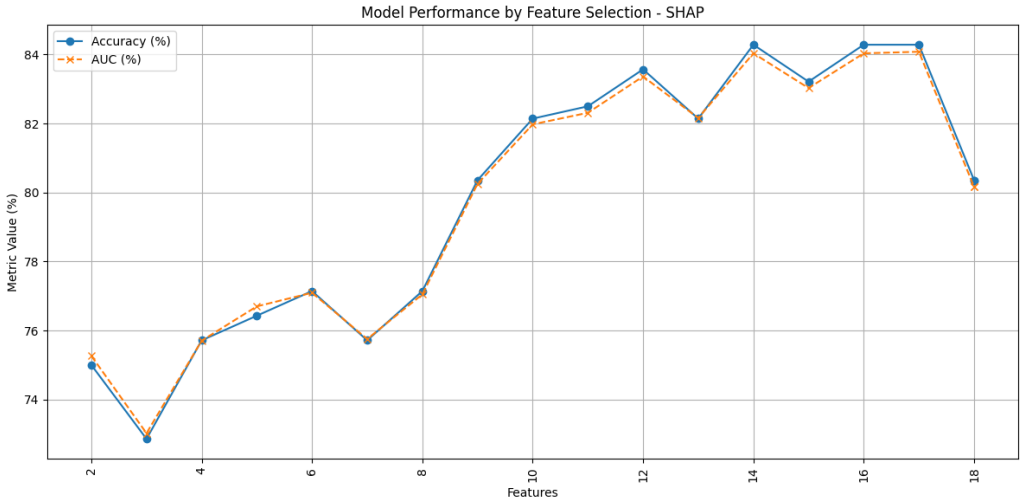

Figura 14 – Evolução da acurácia e AUC com o número de *features* utilizados por meio de SHAP.

Portanto, todo cliente que possuir maior duração no seu empréstimo, já terá o seu *score* de Risco apontado como alto.

5.3.1.4 Feature Selection

Como discutido nas seções anteriores, *feature selection* pode influenciar positivamente no resultado dos modelos de ML e é utilizado amplamente no contexto de *scores* de crédito (DASTILE; CELIK; POTSANE, 2020). Nesta seção serão analisados diferentes métodos para seleção de atributos e seus resultados serão discutidos.

5.3.1.4.1 Seleção baseada em SHAP

Neste caso, os atributos são filtrados de acordo com a sua relevância SHAP ao modelo, figura 10, onde o filtro começa com as 2 principais *features*: *Status of existing checking account* e *Duration in month*.

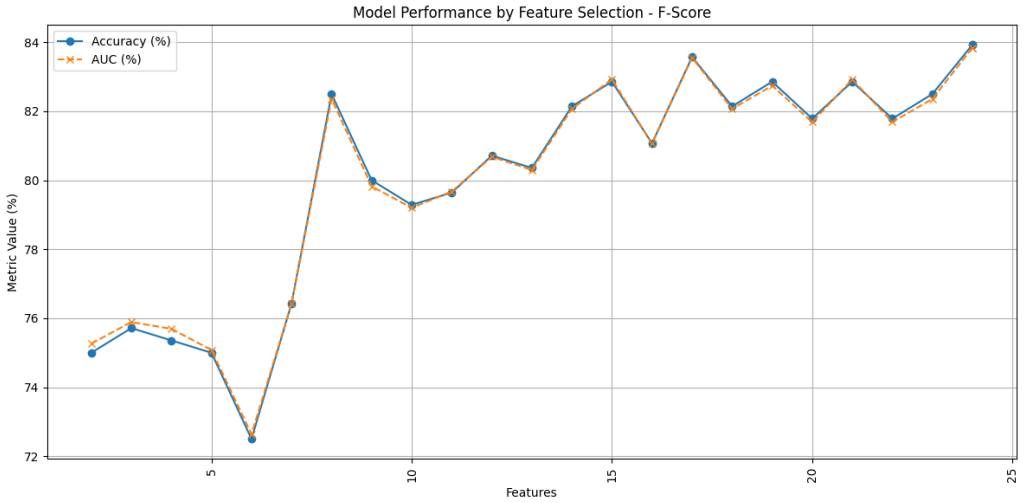

Figura 15 – Evolução da acurácia e AUC com o número de *features* utilizados por meio de F-Score.

Nota-se na figura 20 que o modelo de classificação tem um aumento na sua performance à medida que a quantidade de atributos vai aumentando, tendo pouca variabilidade a partir das 10 *features* mais importantes.

5.3.1.4.2 Seleção baseada em F-Score

Utilizando métodos mais tradicionais como o F-Score, foi possível chegar na seguinte ordem de importância para cada atributo:

Feature	**F-Score**
Status of existing checking account	265,84
Duration in month	73,22
Savings account/bonds	62,09
Sex_0	58,07
Telephone	38,16
Plan_0	33,92
Age in years	30,91
Installment rate in percentage of disposable income	20,59
Purpose	17,45
Job	14,18
Foreign worker	10,18
Sex_1	10,15
Number of existing credits at this bank	10,13
Other features	53,49

Tabela 10 – F-Score para diferentes atributos

Baseado nisso, foi possível realizar a mesma análise desenvolvida anteriormente, possibilitando entender o impacto da quantidade de atributos no resultado do modelo.

É possível notar que a ordem de importância designada pelo método F-Score trouxe grande variabilidade nos resultados, passando a ter uma estabilidade apenas após o 14 atributo. Isso evidencia também que a quantidade de *features* selecionadas possui relevância e análises como essa podem trazer melhores entendimentos sobre a dinâmica do modelo. No entanto, também é necessário ponderar que o *dataset* utilizado neste trabalho é de baixa dimensão (i.e. 25 atributos para 1.400 instâncias pós SMOTE), o que pode tornar este tipo de análise proibitiva para conjuntos de dados de alta dimensão como estudos de genoma, NLP e outras aplicações.

Outro ponto que também pode ser observado é que a discussão sobre seleção de atributos se entrelaça com a discussão de explicabilidade do modelo, pois as *features* mais importantes para o desempenho do modelo devem ser preservadas. Assim, é possível notar semelhanças entre alguns atributos elencados pelo F-Score, tabela 10 e outros pelo SHAP, figura 10.

5.3.2 Regressão: Valor de crédito concedido

5.3.2.1 Resultados dos modelos

O objetivo da regressão é encontrar o valor de crédito concedido para cada cliente, com base no conjunto de dados e no *score* de risco.

Com base nos modelos descritos na seção de metodologia para regressão, o resultado abaixo foi encontrado para o erro (MAPE) dos modelos:

Modelo	**MAPE (%)**
Decision Tree	11.83%
Random Forest	8.27%
GBM	5.75%
XGBoost	6.59%
LightGBM	4.99%

Tabela 11 – MAPE dos modelos de Regressão de crédito concedido

Os modelos que tiveram melhor desempenho de forma geral foram os baseados em gradiente. Em específico o LightGBM possui melhor desempenho, tendo erro 0.7 p.p. menor que o GBM.

Dado a baixa performance obtida com GridSearch na classificação, ela não foi utilizada para o problema de regressão.

5.3.2.2 Explicabilidade dos modelos

Para melhor compreensão de como o modelo se comporta, foi realizada a análise de SHAP na figura 16.

As principais *features* do modelo são:

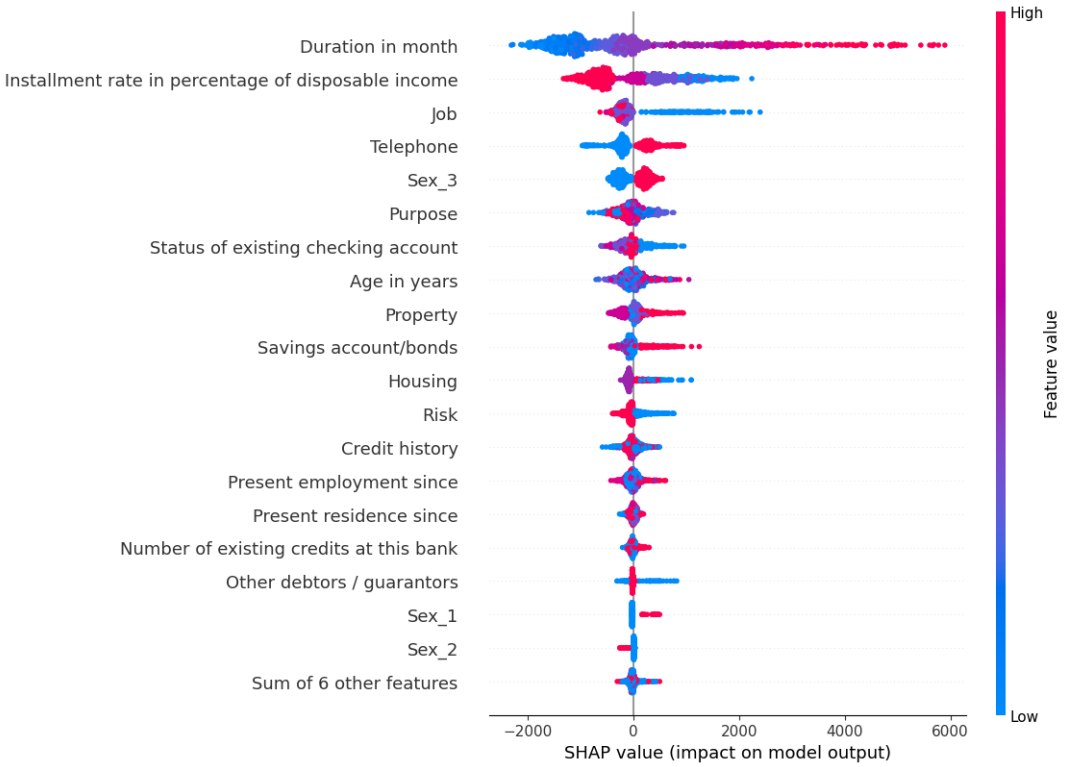

Figura 16 – Gráfico de SHAP para o modelo de regressão LightGBM

1. *Duration in month*: Quanto maior for o seu valor em meses, maior será o valor do crédito concedido

2. *Installment rate in percentage of disposable income*: Quanto menor for o percentual, maior será o crédito concedido

3. *Job*: Para empregos mais qualificados, o crédito concedido é maior

Interessante notar que a métrica de Risco é, apenas, a 12 variável mais importante para o modelo, reforçando o argumento discutido na seção anterior: O *score* de risco do conjunto de dados é pouco efetivo para agrupar os diferentes clientes e possui pouca influência para a decisão da concessão de crédito. Isso vai de encontro ao observado na literatura, que reforça a métrica de risco como um pilar importante para tomada de decisão de crédito (LAWRENCE; SOLOMON, 2013).

Podemos observar na figura 17 o impacto dos atributos discutidos anteriormente no valor de crédito emprestado para um cliente específico. O objetivo desse gráfico é mensurar o quanto cada *feature* impacta o valor final de crédito concedido a partir de um ponto médio (nesse caso DM 3.360,62). Dessa forma, podemos perceber que a duração do empréstimo (valor normalizado para o conjunto de dados) contribui significativamente para o aumento do crédito, seguido do Telefone (i.e. cliente ter um telefone cadastrado) e parcialmente compensado pelo percentual da parcela em relação a sua renda.

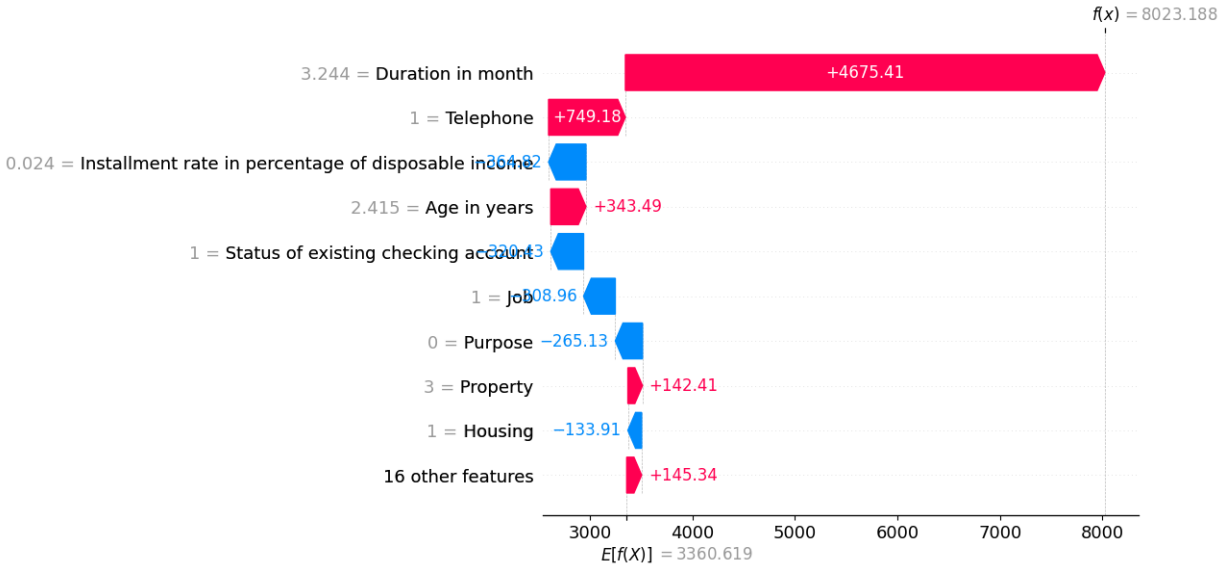

Figura 17 – Gráfico de cascata para valor de crédito concedido e influência dos diferentes atributos

Para entender mais a fundo o impacto da duração do empréstimo no valor do crédito concedido, precisamos entender melhor a distribuição no conjunto de dados. Na figura 18 podemos visualizar isso em diferentes agrupamentos de *duration*.

À medida que a duração do empréstimo dos grupos aumenta, podemos perceber um aumento na média e mediana dos valores de crédito concecido. Somado a isso, há uma maior concentração de instâncias nos agrupamentos de 10 a 30 meses, o que corroboram para o enviesamento do modelo nessa direção.

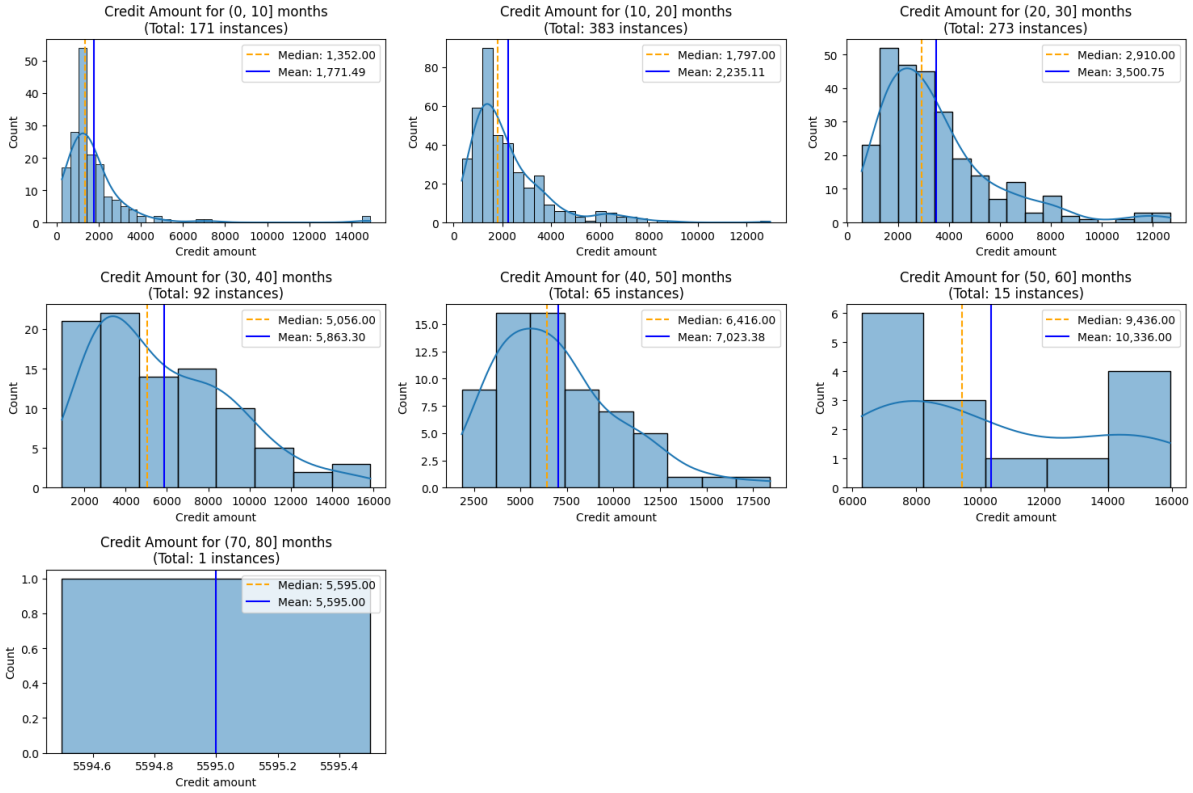

Figura 18 – Histogramas do valor de crédito concedido agrupados por duração do empréstimo

A figura 18 também mostra que a quantidade de instâncias para durações mais longas (>40 meses) é significativamente mais baixa que o restante. Isso contribui para que o modelo tenha o viés de durações mais longas gerarem um empréstimo maior, devido a amostra limitada. Um exemplo seria o grupo (70,80] que possui um único cliente, cujo empréstimo é menor que a média do grupo (30,40].

5.3.2.3 Feature Selection

Seguindo os mesmos passos da seção anterior, o impacto da seleção de *features* será análisado para regressão.

5.3.2.3.1 Seleção baseada em SHAP

Neste caso, os atributos são filtrados de acordo com a sua relevância SHAP ao modelo, figura 19, onde o filtro começa com as 2 principais *features*: *Duration in month* e *Installment rate in percentage of disposable income*.

Nota-se na figura 19 que o modelo de regressão tem uma variação grande no seu desempenho a depender da quantidade de atributos utilizados. O modelo possui sua melhor performance com 3 *features* e possui sua pior performance com 10 atributos.

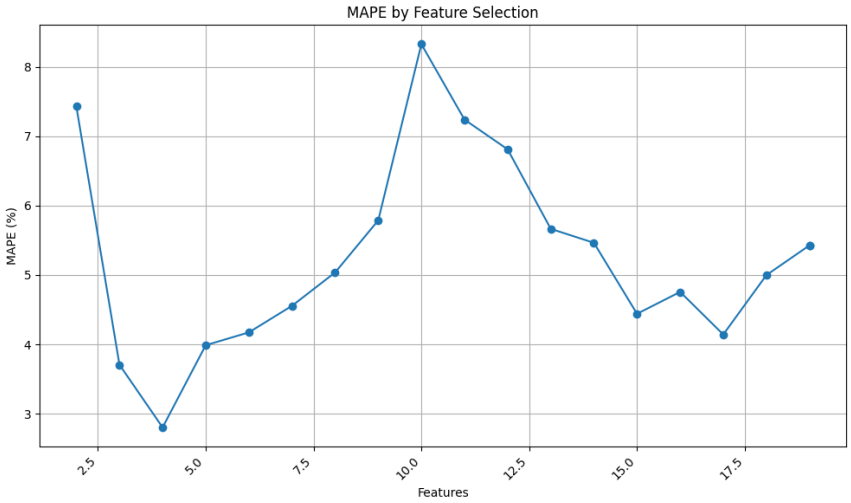

Figura 19 – Evolução do MAPE com o número de *features* utilizados por meio de SHAP.

5.3.2.3.2 Seleção baseada em F-Score

Utilizando métodos mais tradicionais como o F-Score, foi possível chegar à ordem de importância para cada atributo, que pode ser visualizada na tabela 12.

Feature	**F-Score**
Duration in month	2.56
Plan_2	1.84
Number of existing credits at this bank	1.45
Telephone	1.41
Plan_1	1.24
Housing	1.23
Plan_0	1.23
Present employment since	1.18
Risk	1.17
Credit history	1.16
Savings account/bonds	1.16
Sex_1	1.14
Number of people being liable to provide maintenance for	1.12
Installment rate in percentage of disposable income	1.10
Status of existing checking account	1.09
Property	1.03
Other debtors / guarantors	1.00
Other features	11.85

Tabela 12 – F-Scores para problema de regressão e diferentes atributos

É possível notar a grande diferença entre a ordenação dos atributos F-Score, tabela 12, e SHAP, figura 16, com apenas alguns atributos em comuns na lista das *features* mais relevantes. Ainda é possível notar que a métrica de Risco também possui pouco impacto

Figura 20 – Evolução do MAPE com o número de *features* utilizados por meio de F-Score.

no modelo, reforçando as discussões feitas anteriormente com base nos resultados obtidos pelo SHAP.

A evolução da quantidade de *features* mostra que os 10 primeiros atributos mais relevantes possuem o melhor desempenho e a seguir o desempenho do modelo se deteriora estabilizando a partir dos 20 atributos. Comparado a análise de SHAP é possível notar que os principais atributos elencados pelo F-Score não possuem impacto relevante no resultado modelo, pois este necessita de uma quantidade maior para que ter MAPEs menores.

6 CONCLUSÕES

O trabalho teve como objetivo explorar e avaliar a eficácia de diversos modelos de machine learning no setor de crédito, especialmente na avaliação de risco de crédito e na determinação de valores de empréstimo. A investigação baseou-se na análise de conjuntos de dados públicos, mais em específico o conjunto de dados de crédito alemão dos anos 1990. A motivação deste trabalho gira em torno da necessidade de modelos serem precisos e escaláveis nas instituições financeiras para mitigar riscos e aprimorar os processos de tomada de decisão.

Alinhado com os objetivos da monografia, cada modelo foi testado e comparado usando métricas de desempenho chave, incluindo acurácia, *Area Under Curve* (AUC) e Erro Absoluto Médio Percentual (MAPE). Os resultados revelaram que, embora modelos tradicionais como a Regressão Logística fornecessem uma base sólida, modelos mais sofisticados, como Random Forest e XGBoost, demonstraram desempenho superior para em problemas de classificação tendo melhor acurácia preditiva e manejo de relações não lineares nos dados. Para problemas de regressão, o modelo LightGBM também mostrou resultados promissores quando comparado a outros modelos tradicionais e sofisticados como XGBoost.

A incorporação de valores SHAP (SHapley Additive exPlanations) adicionou uma camada essencial de interpretabilidade às previsões dos modelos. Essa abordagem destacou a importância de características específicas, como o status da conta corrente existente e a duração dos empréstimos, na influência do risco de crédito e a duração dos empréstimos para definição do valor do empréstimo. Ao fornecer *insights* sobre a contribuição de cada característica, a análise SHAP não só aumentou a transparência dos modelos, mas também ofereceu exemplos claros para entender a veracidade dos resultados encontrados. Além disso, o estudo ressalta a necessidade de uma seleção criteriosa de características para evitar *data leakage*, o que poderia comprometer a integridade e confiabilidade dos modelos.

Durante o trabalho foi possível notar que a métrica de risco disponibilizada no conjunto de dados possuía pouco poder preditivo sobre as variáveis de valor de crédito concedido, além de possuir limitação na separação da população. Isso acaba evidenciando uma limitação do trabalho, dado que os resultados obtidos são dependentes da qualidade dos dados para treinamento dos modelos.

Como trabalhos futuros, a exploração de um novo *score* de risco para melhor segmentação do conjunto de dados poderia trazer melhores *insights* sobre o problema. Além disso, a utilização de outros conjuntos de dados com quantidade maior de instâncias também seria uma boa oportunidade para aplicar as metodologias desenvolvidas nesta

monografia e explorar outros tipos de modelos não abordados, como *Deep Learning*.

REFERÊNCIAS

ALZUBAIDI L., B. J. A.-S. A. *et al.* A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. **Journal of Big Data**, v. 10, n. 46, 2023. Available at: <https://doi.org/10.1186/s40537-023-00727-2>.

AWS, A. W. S. **AWS EC2 Instance Types - G4**. <https://aws.amazon.com/ec2/instance-types/g4/>. Acessado em: 2024-08-17.

AWS, A. W. S. **AWS EC2 Pricing**. <https://aws.amazon.com/ec2/pricing/on-demand/>. Acessado em: 2024-08-17.

BAESENS T VAN GESTEL, S. V. M. S. J. S. B.; VANTHIENEN, J. Benchmarking state-of-the-art classification algorithms for credit scoring. **Journal of the Operational Research Society**, Taylor & Francis, v. 54, n. 6, p. 627–635, 2003. Available at: <https://doi.org/10.1057/palgrave.jors.2601545>.

BANASIK, J.; CROOK, J. N.; THOMAS, L. C. Not if but when will borrowers default. **The Journal of the Operational Research Society**, Palgrave Macmillan Journals, v. 50, n. 12, p. 1185–1190, 1999. ISSN 01605682, 14769360. Available at: <http://www.jstor.org/stable/3010627>.

Banco Central do Brasil. **Resolução N 2682, de 21 de Dezembro de 1999**. 1999. <https://www.bcb.gov.br/pre/normativos/res/1999/pdf/res_2682_v2_L.pdf>. Accessed: 2024-06-06.

BOMMASANI, R. *et al.* **On the Opportunities and Risks of Foundation Models**. 2022. Available at: <https://arxiv.org/abs/2108.07258>.

Brasil. **Lei No 12.414, de 09 de Junho de 2011**. 2011. <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12414.htm>. Accessed: 2024-06-06.

CHAWLA, N. V. *et al.* Smote: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, AI Access Foundation, v. 16, p. 321–357, jun. 2002. ISSN 1076-9757. Available at: <http://dx.doi.org/10.1613/jair.953>.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. *In*: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Available at: <https://doi.org/10.1145/2939672.2939785>.

DASTILE, X.; CELIK, T.; POTSANE, M. Statistical and machine learning models in credit scoring: A systematic literature survey. **Applied Soft Computing**, v. 91, p. 106263, 2020. ISSN 1568-4946. Available at: <https://www.sciencedirect.com/science/article/pii/S1568494620302039>.

GéRON, A. **Hands-on machine learning with scikit-learn and tensorflow: Concepts, tools, and techniques to build Intelligent Systems**. [*S.l.: s.n.*]: O'Reilly, 2021.

HAMORI, S. *et al.* Ensemble learning or deep learning? application to default risk analysis. **Journal of Risk and Financial Management**, v. 11, n. 1, 2018. ISSN 1911-8074. Available at: <https://www.mdpi.com/1911-8074/11/1/12>.

HAND, D. J. Modelling consumer credit risk. **IMA Journal of Management Mathematics**, v. 12, n. 2, p. 139–155, 10 2001. ISSN 1471-678X. Available at: <https://doi.org/10.1093/imaman/12.2.139>.

HOFMANN, H. **Statlog (German Credit Data)**. 1994. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

HSIEH, N.-C.; HUNG, L.-P. A data driven ensemble classifier for credit scoring analysis. **Expert Systems with Applications**, v. 37, n. 1, p. 534–545, 2010. ISSN 0957-4174. Available at: <https://www.sciencedirect.com/science/article/pii/S0957417409004771>.

HUYEN, C. **Designing Machine Learning Systems**. O'Reilly Media, 2022. ISBN 9781098107918. Available at: <https://books.google.com.br/books?id=EThwEAAAQBAJ>.

JIMENEZ, G.; SAURINA, J. **Loan characteristics and credit risk**. [*S.l.*], 2003. Available at: <https://ideas.repec.org/p/fip/fedhpr/857.html>.

LAWRENCE, D.; SOLOMON, A. **Managing a Consumer Lending Business: 2nd Edition**. Solomon Lawrence Partners, 2013. ISBN 9780971753730. Available at: <https://books.google.com.br/books?id=eWd_NAEACAAJ>.

LOUZADA, F.; ARA, A.; FERNANDES, G. B. Classification methods applied to credit scoring: Systematic review and overall comparison. **Surveys in Operations Research and Management Science**, v. 21, n. 2, p. 117–134, 2016. ISSN 1876-7354. Available at: <https://www.sciencedirect.com/science/article/pii/S1876735416300101>.

LUNDBERG, S.; LEE, S.-I. **A Unified Approach to Interpreting Model Predictions**. 2017.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *In*: GUYON, I. *et al.* (ed.). **Advances in Neural Information Processing Systems 30**. Curran Associates, Inc., 2017. p. 4765–4774. Available at: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.

MARCACINI, R. **MBA IA e Big Data, ICMC, Notas de Aula**. 2023.

MARRON, D. 'lending by numbers': credit scoring and the constitution of risk within american consumer credit. **Economy and Society**, Routledge, v. 36, n. 1, p. 103–133, 2007. Available at: <https://doi.org/10.1080/03085140601089846>.

MATTIUZO, M. **MBA IA e Big Data, ICMC, MBA IA e Big Data, ICMC, Notas de Aula**. 2023.

ORESKI, S.; ORESKI, D.; ORESKI, G. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. **Expert systems with applications**, Elsevier, v. 39, n. 16, p. 12605–12617, 2012.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking**. O'Reilly Media, 2013. ISBN 9781449374297. Available at: <https://books.google.com.br/books?id=EZAtAAAAQBAJ>.

QUINLAN, R. **Statlog (Australian Credit Approval)**. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C59012.

ROMERO, R. **MBA IA e Big Data, ICMC, Notas de Aula**. 2023.

SILVA, D. F. **Lecture notes in Redes Neurais e Deep Learning - Fundamentos**. [*S.l.: s.n.*]: ICMC - USP, 2023.

THOMAS, L.; CROOK, J.; EDELMAN, D. **Credit scoring and its applications**. [*S.l.: s.n.*]: SIAM, 2017.

WANG, X.; XU, M.; PUSATLI, Ö. T. A survey of applying machine learning techniques for credit rating: Existing models and open issues. *In*: ARIK, S. *et al.* (ed.). **Neural Information Processing**. Cham: Springer International Publishing, 2015. p. 122–132. ISBN 978-3-319-26535-3.

World Bank. **Germany Inflation Rate (CPI)**. 2024. Accessed: 2024-08-28. Available at: <https://www.macrotrends.net/global-metrics/countries/DEU/germany/inflation-rate-cpi>.

World Bank. **Germany Real Interest Rate**. 2024. Accessed: 2024-08-28. Available at: <https://ycharts.com/indicators/germany_real_interest_rate>.

World Bank. **Interest Rate Spread (Lending Rate Minus Deposit Rate)**. 2024. Accessed: 2024-08-28. Available at: <https://data.worldbank.org/indicator/FR.INR.LNDP?end=2023&start=1967&view=map&year=1994>.

APÊNDICES

APÊNDICE A – APÊNDICE 1: DESCRIÇÃO DOS ATRIBUTOS CATEGÓRICOS DO CONJUNTO DE DADOS

Abaixo uma descrição dos atributos categóricos:

A.1 Status of existing checking account

Este atributo descreve o status da conta corrente existente do solicitante, refletindo sua estabilidade financeira.

- 0 DM
- 0 <= ... < 200 DM
- = 200 DM / salário por pelo menos 1 ano
- sem conta corrente

A.2 Credit history

Este atributo descreve o histórico de crédito do solicitante, indicando sua confiabilidade no pagamento de empréstimos e créditos anteriores.

- nenhum crédito tomado/ todos os créditos pagos pontualmente
- todos os créditos neste banco pagos pontualmente
- créditos existentes pagos pontualmente até agora
- atraso no pagamento no passado
- conta crítica/ outros créditos existentes (não neste banco)

A.3 Purpose

Este atributo especifica a finalidade para a qual o solicitante está solicitando crédito, como a compra de um carro ou financiamento de educação.

- carro (novo)
- carro (usado)
- móveis/equipamentos
- rádio/televisão

- eletrodomésticos
- reformas
- educação
- (férias - não existe?)
- requalificação
- negócios
- outros

A.4 Savings account/bonds

Este atributo detalha o montante de poupança ou títulos que o solicitante possui, indicando suas reservas financeiras.

- ... < 100 DM
- 100 <= ... < 500 DM
- 500 <= ... < 1000 DM
- ... >= 1000 DM
- desconhecido/ sem conta poupança

A.5 Present employment since

Este atributo indica há quanto tempo o solicitante está empregado em seu trabalho atual, refletindo a estabilidade no emprego.

- desempregado
- < 1 ano
- 1 <= ... < 4 anos
- 4 <= ... < 7 anos
- >= 7 anos

A.6 Personal status and sex

Este atributo captura o estado civil e o sexo do solicitante, o que pode influenciar a avaliação de risco de crédito.

- masculino: divorciado/separado
- feminino: divorciada/separada/casada
- masculino: solteiro
- masculino: casado/viúvo
- feminino: solteira

A.7 Other debtors/guarantors

Este atributo identifica se há co-solicitantes ou garantes associados à solicitação de crédito.

- nenhum
- co-solicitante
- garantidor

A.8 Property

Este atributo lista o tipo de propriedade possuída pelo solicitante, que pode ser usada como garantia para o empréstimo.

- imóveis
- acordo de poupança/seguro de vida
- carro ou outro (não na conta poupança)
- desconhecido/ sem propriedade

A.9 Other installment plans

Este atributo detalha quaisquer outros planos de parcelamento que o solicitante possa ter, como com bancos ou lojas.

- banco
- lojas
- nenhum

A.10 Housing

Este atributo descreve a situação de moradia atual do solicitante, se ele aluga, possui ou mora gratuitamente.

- aluguel
- próprio
- gratuito

A.11 Job

Este atributo categoriza o emprego do solicitante, variando de desempregado a altamente qualificado, indicando a estabilidade de renda.

- desempregado/ não qualificado - não residente
- não qualificado - residente
- empregado qualificado / oficial
- gestão/ autônomo/ empregado altamente qualificado/ oficial

A.12 Telephone

Este atributo indica se o solicitante possui um telefone registrado em seu nome, o que pode ser um método de contato.

- nenhum
- sim, registrado em nome do cliente

A.13 Foreigner worker

Este atributo indica se o solicitante é um trabalhador estrangeiro, o que pode afetar a avaliação de risco de crédito.

- sim
- não

A.14 Risk

Este atributo classifica o risco de crédito geral do solicitante como baixo ou alto, com base em seu perfil financeiro.

- Baixo
- Alto