UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Ciência e Tecnologia
Câmpus de Sorocaba

LEONARDO ADERALDO VARGAS

ANÁLISE DE RISCO DE CRÉDITO DIRECIONADA POR MODELAGEM MATEMÁTICA E APRENDIZADO DE MÁQUINA

Sorocaba - SP
2024

LEONARDO ADERALDO VARGAS

ANÁLISE DE RISCO DE CRÉDITO DIRECIONADA POR MODELAGEM MATEMÁTICA E APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso apresentado ao Instituto de Ciência e Tecnologia de Sorocaba, Universidade Estadual Paulista (UNESP), como parte dos requisitos para a obtenção do grau de Bacharel em Engenharia de Controle e Automação.

Orientador: Prof. Dr. Leopoldo André Lusquino Filho

Coorientador: Prof. Dr. Galdenoro Botura Junior

Sorocaba - SP
2024

V297a

Vargas, Leonardo Aderaldo

Análise de risco de crédito direcionada por modelagem matemática e aprendizado de máquina / Leonardo Aderaldo Vargas. -- Sorocaba, 2024

84 p.

Trabalho de conclusão de curso (Bacharelado - Engenharia de Controle e Automação) - Universidade Estadual Paulista (UNESP), Instituto de Ciência e Tecnologia, Sorocaba

Orientador: Leopoldo André Lusquino Filho

Coorientador: Galdenoro Botura Junior

1. Finanças. 2. Aprendizado do computador. 3. Administração de riscos. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Universidade Estadual Paulista (UNESP), Instituto de Ciência e Tecnologia, Sorocaba. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

ANÁLISE DE RISCO DE CRÉDITO DIRECIONADA POR MODELAGEM MATEMÁTICA E APRENDIZADO DE MÁQUINA

LEONARDO ADERALDO VARGAS

ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO PARTE DO REQUISITO PARA A OBTENÇÃO DO GRAU DE **BACHAREL EM ENGENHARIA DE CONTROLE E AUTOMAÇÃO**

Profº. Dr. Everson Martins

Coordenador

BANCA EXAMINADORA:

Prof. Dr. Leopoldo André Lusquino Filho

Orientador/UNESP-Campus de Sorocaba

Prof. Dr. Luis Armando de Oro Arenas

UNESP-Campus de Sorocaba

Benedito Faustinoni

Santander-Membro Externo

Março de 2024

AGRADECIMENTOS

Primeiramente agradeço a Deus, pois sempre recorri a Ele em minhas orações pedindo para que meus caminhos permanecessem abertos e iluminados.

Agradeço à minha família, especialmente minha mãe Marcia, meu irmão Gustavo e minhas tias Eulalia e Maria Aparecida, os quais foram responsáveis pela melhor educação, alicerce, amor e incentivos que uma pessoa pode ter.

Agradeço à minha namorada Patrícia, a qual sempre acreditou em mim e me apoiou em todas as circunstâncias.

Agradeço a todos meus amigos da faculdade, os quais foram e ainda são grandes companheiros no quesito acadêmico, profissional e pessoal.

Agradeço a todos meus colegas do Banco Santander, empresa a qual eu atuo, por todos os ensinamentos adquiridos no cotidiano e por me mostrarem que este trabalho possui um grande impacto positivo para a sociedade.

Agradeço aos professores Galdenoro Botura Junior e Leopoldo André Dutra Lusquino Filho pelo interesse no tema, pelas excelentes orientações durante a execução e por acreditarem que eu pudesse fazer um bom trabalho.

Agradeço à faculdade Universidade Estadual Júlio de Mesquita Filho (UNESP), por proporcionar-me uma formação acadêmica excepcional a qual é um grande motivo de orgulho para mim.

RESUMO

Ao longo dos anos, o Brasil aumenta a demanda por produtos de crédito, todavia, essa situação é acompanhada pelo acréscimo significativo da quantidade de pessoas que não honram seus compromissos financeiros. Essa relação entre concessões e perdas destaca a importância do desenvolvimento de abordagens capazes de maximizar a assertividade e minimizar os riscos, visto que decisões ruins podem acabar endividando o cliente e acabando com a saúde financeira da empresa. Neste âmbito, este trabalho propõe melhorar a concessão de crédito de uma instituição financeira através da substituição da política de crédito vigente por um modelo matemático de Machine Learning. Devido à falta de disponibilidade de dados reais, optou-se por utilizar uma base de dados fictícia do Kaggle a fim de simular um ambiente realista. Visando combater a inadimplência e erros operacionais, o software desenvolvido na linguagem Python tem como intuito receber uma base de dados de clientes novos, aplicar a modelagem e classificar quais clientes podem ou não receber o crédito. Durante a pesquisa, explicou-se a fundamentação teórica da análise de crédito e do aprendizado de máquina, proporcionando base técnica para o correto entendimento dos processos. Através da implementação do modelo, obteve-se ganho de 17% de Recall, 5% de AUC, 9% de KS e 6,88% de ROCP, representando um aumento de quase R$ 100 milhões de lucro e permitindo a melhor classificação de perfil de risco do cliente.

Palavras-chave: risco de crédito; análise de dados; modelagem matemática; aprendizado de máquina; gestão de riscos.

ABSTRACT

Over the years, Brazil increases the demand for credit products, however, this situation is accompanied by a significant increase in the number of people who do not honor their financial commitments. This relationship between concessions and losses highlights the importance of developing approaches capable of maximizing assertiveness and minimizing risks, since bad decisions can end up indebting the customer and ending the financial health of the company. In this context, this work proposes to improve the credit granting of a financial institution by replacing the current credit policy by a mathematical model of Machine Learning. Due to the lack of real data availability, it was decided to use a fictitious Kaggle database in order to simulate a realistic environment. In order to combat default and operational errors, the software developed in the Python language is intended to receive a database of new customers, apply the modeling and classify which customers can or can not receive credit. During the research, the theoretical foundation of credit analysis and machine learning was explained, providing a technical basis for the correct understanding of the processes. Through the implementation of the model, it was obtained a gain of 17% of Recall, 5% of AUC, 9% of KS and 6.88% of ROCP, representing an increase of almost R$ 100 million of profit and allowing the best classification of customer risk profile.

Keywords: credit risk; data analysis; mathematical modeling; machine learning; risk management.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

SPC Brasil	Serviço de Proteção ao Crédito
BACEN	Banco Central do Brasil
ML	Machine Learning
PD	Probability of Default
EAD	Exposure at Default
LGD	Loss Given Default
EL	Expected Loss
ROCP	Return on Credit Portfolio
MC	Matriz de Confusão
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
FN	Falso Negativo
FP	Falso Positivo
AUC	Área Sob a Curva
KS	Teste de Hipótese de Kolmogorov-Smirnov
CRISP-DM	Cross Industry Standard Process for Data Mining

SUMÁRIO

1 INTRODUÇÃO

1.1 Considerações iniciais

Crédito é uma palavra derivada do latim, significando “confiança”. Ele é uma espécie de empréstimo solicitado por um cliente a alguma instituição financeira visando antecipar algum tipo de gasto, auxiliando o solicitante quando o mesmo não possui o capital.

Implantado durante a Revolução Industrial, o crédito possibilitou que pessoas planejassem investimentos e abrissem seu próprio negócio, empresas adquirissem novas tecnologias e aumentassem sua produção e as instituições financeiras obtivessem lucro e capital de giro para ampliar seu patrimônio. Sebben (2020) afirma que essa situação resultou em inúmeras ascensões sociais e movimentação da economia, portanto, tornou-se uma operação extremamente importante para o desenvolvimento da sociedade em geral.

Pelo fato de ser um empréstimo, a concessão de crédito é realizada sob condições de incerteza, logo, ao antecipar recursos, a instituição naturalmente insere-se em um ambiente sujeito a perdas, muitas vezes por situações ao acaso que acontecem com os tomadores, mas também por conta de pessoas fraudadoras e má intencionadas que solicitam crédito já sabendo que não o pagarão. Por tratar-se de uma incerteza, esse processo está diretamente ligado a riscos, logo, o credor necessita de algumas garantias as quais serão protocoladas após uma análise criteriosa sobre diversas informações a respeito do cliente.

O risco de crédito nada mais é do que a probabilidade de perda financeira decorrente do não cumprimento de obrigações de pagamento por parte do solicitante, logo, uma boa análise baseia-se em determinadas metodologias a fim de garantir maior confiabilidade e segurança, resultando na criação de uma “nota de cliente” e classificando-o como bom ou mau pagador.

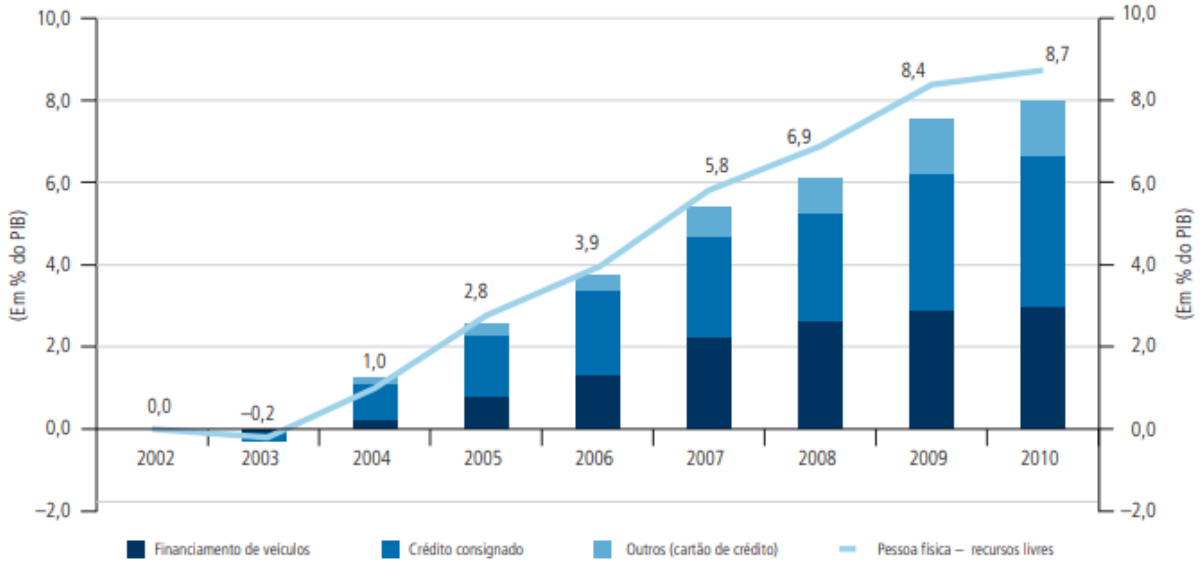
Tal análise fundamenta-se em inúmeras variáveis compostas de dados internos e externos sobre o cliente, pois é de suma importância definir corretamente o perfil do solicitante a fim de melhorar a gestão, mitigar os riscos, atender às necessidades do cliente e garantir que a instituição financeira receba o pagamento para continuar existindo. Um processo de concessão de crédito mal feito pode endividar um cliente e prejudicar a saúde financeira da empresa, causando severos danos a ambos os lados.

Como envolve diversas variáveis e diversos clientes, as bases de dados costumam ser bastante extensas e analisá-las a olho nu torna-se praticamente impossível. Nesse âmbito, a utilização de softwares baseados em técnicas matemáticas e estatísticas tornaram-se fundamentais, pois permitem a classificação em larga escala, de forma automática e segura, criando-se assim o que chamamos de “Modelos de Risco de Crédito”, os quais, segundo Sicsú (2010), possuem o objetivo de prever, na data de decisão do crédito, a probabilidade da concessão tornar-se uma perda para o credor.

1.2 Contexto e justificativa

A demanda por produtos de crédito aumentou no Brasil com o passar dos anos. Na Figura 1, Mora (2015), em conjunto com o Instituto de Pesquisa Econômica Aplicada, expressa essa descoberta através da representatividade da concessão de crédito de financiamento de veículos, crédito consignado, cartão de crédito e outros recursos em relação ao PIB nacional.

Figura 1 – Representatividade de Produtos de Crédito em relação ao PIB

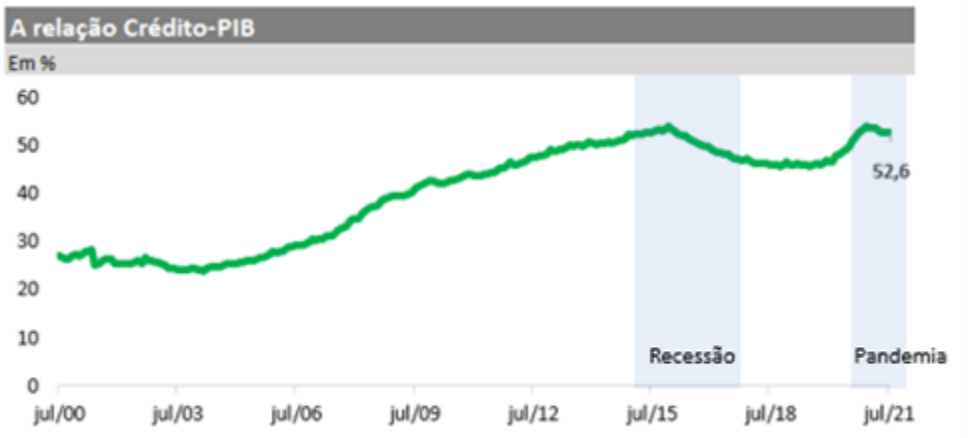


Fonte: Mora (2015, p.22).

Em anos recentes, Sfeir (2021) afirma que de acordo com o relatório “Estatísticas Monetárias e de Crédito” de agosto de 2021, o saldo total das operações de crédito com o Sistema Financeiro Nacional foi de quatro trilhões e trezentos bilhões de reais, sendo representado na figura 2. Além disso, mais do que crescimento, também houve mudança no

perfil do crédito concedido, com ampliação da participação de recursos destinados a pessoas físicas e aumento do fornecimento por bancos privados.

Figura 2 – Evolução do Crédito como proporção do PIB



Fonte: Sfeir (2021).

Mesmo com a crescente demanda por produtos de crédito, o Brasil conta com uma das maiores inadimplências mundiais. Lima (2008) diz que a história mostra que mesmo bons clientes têm potencial de não honrarem suas obrigações financeiras. Segundo levantamentos do “Mapa da Inadimplência e Negociação de Dívidas no Brasil”, relatório mensal publicado pelo Serasa Limpa Nome, a figura 3 mostra que em novembro de 2023 o Brasil alcançou em torno de 71 milhões de pessoas devedoras.

Figura 3 – Evolução da Inadimplência em 2022 e 2023

Fonte: Serasa (2023).

Ao não pagarem suas dívidas, as instituições financeiras deixam de receber o pagamento e os clientes tendem a endividarem-se cada vez mais, prejudicando todo o ecossistema econômico. Nesse âmbito, bancos e fintechs naturalmente enrijecem seu processo de concessão de crédito e elevam as taxas de juros, prejudicando os bons pagadores de concretizarem muitos sonhos. Empresas deixam de existir, negócios deixam de vender e muitas pessoas não conseguem financiar seu imóvel ou carro após longos anos de trabalho árduo. Como forma de auxiliar o cenário, as instituições financeiras reinventam-se a todo momento, seja atualizando as políticas ou buscando novas metodologias, resultando em fortes cobranças sobre os analistas de crédito, os quais passam boa parte do tempo analisando informações minuciosamente de forma manual e gastando esforço em partes que poderiam ser automatizadas. Visando acabar com este tipo de problema, a proposta é criar um software matemático capaz de classificar corretamente os clientes de alto risco de forma automática e segura, baseando-se exclusivamente em dados.

Ao longo da pesquisa, apresentou-se uma introdução conceitual sobre risco de crédito e técnicas de Machine Learning para familiarização do leitor a respeito do tema. Além disso, desenvolveu-se uma política de crédito para ser o benchmarking do modelo desenvolvido de forma que, ao final do trabalho, espera-se que os resultados obtidos provem que o método quantitativo possui bastante potencial para ser utilizado em análises de risco de crédito.

2 LEVANTAMENTO BIBLIOGRÁFICO

Para correta compreensão do trabalho, é necessário o conhecimento de determinados conteúdos, portanto, apresentou-se alguns fundamentos a respeito de Risco de Crédito e Aprendizado de Máquina a fim de nivelar, ao menos conceitualmente, algumas teorias importantes.

Além disso, o intuito desta pesquisa é o desenvolvimento de um software de Aprendizado de Máquina voltado para a correta classificação de pessoas físicas, portanto, embora algumas metodologias sejam análogas para ambos os casos, conceitos exclusivos de classificação de pessoas jurídicas (empresas) não serão abordados.

2.1 Crédito e seus princípios

Embora sejam palavras semelhantes, “risco” e “incerteza”, no contexto de crédito, diferem um pouco. Risco é uma probabilidade em função de algo eventual e incerto, portanto, não depende de nenhuma das partes. A incerteza ocorre quando a instituição financeira não possui dados sobre o cliente, portanto, torna-se impossível avaliá-lo.

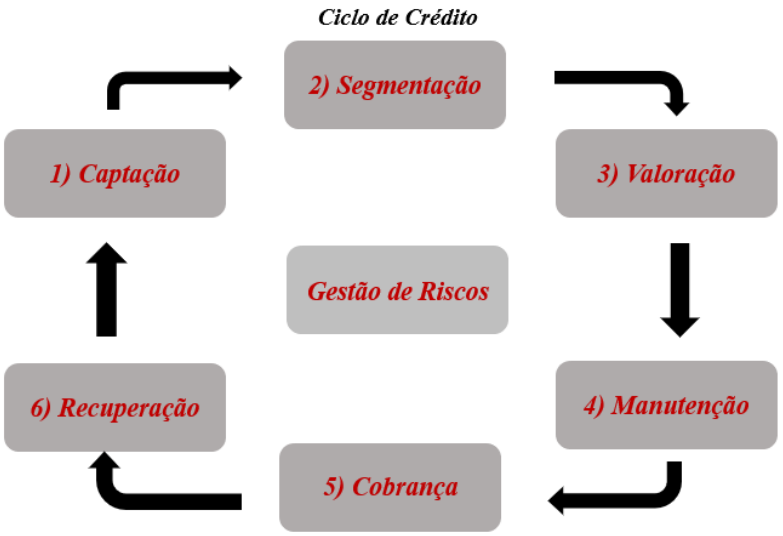
Como dito anteriormente, o risco de crédito nada mais é do que a probabilidade de perda financeira decorrente do não cumprimento de obrigações de pagamento por parte do solicitante. Por tratar-se de uma “operação de confiança”, toda vez que há uma antecipação de recursos há chances da não recuperação do valor e é justamente este risco que o credor aceita passar visto que será recompensado futuramente através dos juros.

Embora o pagamento de juros seja rentável ao banco, deseja-se evitar clientes completamente inadimplentes, pois eles oferecem problemas de rentabilidade e jamais pagarão suas dívidas. Somado ao fato de que existe instituições nacionais que balizam a taxa de juros e decidem quando a mesma aumenta ou diminui, o cenário de crédito requer mudanças constantes. Dada a situação, o objetivo da análise de risco de crédito é justamente descobrir quem são os bons e maus pagadores, reduzindo o volume de crédito concedido a pessoas que não poderão honrá-lo ou concedendo volume aos clientes adimplentes.

De acordo com Tchilian (2022), a concessão de crédito aliada a uma boa gestão de riscos representa uma das principais fontes de renda para uma instituição financeira, portanto, para facilitar o ecossistema, criou-se um fluxo chamado Ciclo de Crédito, o qual consiste em seis etapas: Captação, Segmentação, Valoração, Manutenção, Cobrança e Recuperação.

A figura 4 mostra um resumo sobre cada uma das etapas:

Figura 4 – Fluxograma do Ciclo de Crédito



Fonte: Autoria Própria.

Captação **–** A captação é o ponto de partida do ciclo, onde a instituição financeira busca atrair novos clientes. Isso é feito por meio de estratégias de marketing e vendas, identificando produtos adequados ao público-alvo e direcionando esforços para conquistar novos negócios. A captação pode ser ativa, realizada pelo time comercial da instituição, ou passiva, quando os próprios clientes buscam a instituição.

Segmentação **–** Dado que cada cliente possui diferentes necessidades e perfis financeiros, a segmentação é fundamental. Ela envolve a divisão dos clientes em grupos com características similares, permitindo a oferta de produtos e serviços específicos para atender às necessidades de cada segmento. Isso melhora a eficácia das ofertas e aumenta as chances de satisfação do cliente.

Valoração **–** A etapa de valoração é onde as metodologias de risco de crédito entram em ação. Esses métodos avaliam os clientes com base em diversas variáveis, a fim de classificá-los como bons ou maus pagadores. Além disso, há políticas de crédito, as quais são regras definidas para estabelecer limites e critérios para a concessão de crédito, garantindo consistência nas decisões.

Manutenção – A manutenção refere-se à gestão contínua dos relacionamentos com os clientes. Isso envolve o atendimento às necessidades dos clientes ao longo do tempo e o incentivo à fidelidade. Clientes antigos são valorizados, uma vez que sua permanência tende a ser benéfica tanto em termos de rentabilidade quanto de redução do risco de crédito.

Cobrança e Recuperação – Esta etapa lida com situações de inadimplência. As políticas de cobrança são implementadas para lidar com clientes que não conseguem cumprir suas obrigações de pagamento. Isso pode envolver medidas como a aplicação de juros, multas ou a renegociação da dívida. O objetivo é minimizar as perdas e manter um equilíbrio entre os lucros da instituição e a recuperação dos valores em débito.

2.2 O processo de decisão

Na história do sistema financeiro, pelo fato das decisões de empréstimos serem abrangentes e pautarem-se em inúmeras informações, um dos marcos mais significativos foi a introdução dos *"5 C's do Crédito".* De acordo com Sebben (2020), em conjunto, os *5 C's do Crédito* norteiam todo o processo de concessão de crédito criando os principais fatores da análise de risco e auxiliam a expor a probabilidade de um solicitante honrar ou não o pagamento dos empréstimos, sendo fundamentais durante todo o Ciclo de Crédito para que a instituição financeira minimize as perdas e maximize os resultados.

A seguir detalhou-se cada um desses critérios:

Caráter – Sendo o elemento básico para decisões de crédito, este critério avalia características pessoais e profissionais do cliente, como sua reputação em termos de integridade e honestidade. Informações como histórico de pagamentos em dia, dívidas em atraso, antecedentes criminais, informações de *bureus*, entre outras questões fazem parte da avaliação do comportamento e imagem do indivíduo.

Capacidade – Refere-se à validação sobre as condições do tomador pagar suas dívidas, avaliando questões como renda e suas fontes, conhecimento técnico, área de atuação profissional, despesas e outras dívidas existentes, além de adequar os valores das prestações e prazos respeitando as limitações do cliente a fim de não o endividar.

Colateral – É a garantia do pagamento do empréstimo a qual o credor pode recorrer em casos de inadimplência do solicitante, portanto, são bens de valor como casas, carros,

entre outros, avaliando-se o valor desses bens e sua liquidez a fim de decidir se é suficiente para cobrir o valor do empréstimo.

Condições **–** Indica as condições referentes ao contexto econômico no qual o empréstimo será realizado, avaliando as características socioeconômicas do tomador e do mercado nacional, como estabilidade econômica da região, as taxas de juros e as condições do mercado a fim de definir se o momento para concessões é propício.

Capital **–** Apresenta uma análise interna sobre as finanças da instituição a fim de garantir que ela possui o dinheiro solicitado pelo cliente, avaliando questões como patrimônio líquido, balanço financeiro, projeções de rentabilidade, bens e fluxo financeiro interno.

Por meio de uma análise minuciosa das variáveis referentes aos 5C's do Crédito, objetiva-se identificar padrões através de dados históricos capazes de identificar bons e maus pagadores. A partir dessas descobertas, propõem-se a criação de uma política de crédito composta de cortes estratégicos nas variáveis identificadas como mais significativas.

Essa metodologia serviu por muito tempo como método de avaliação da capacidade de um cliente obter crédito, sendo um conjunto de critérios utilizados pelos credores para balizar a concessão e entender se o mesmo é elegível ou não ao empréstimo solicitado. Mesmo sendo comprovadamente eficaz, à medida que a complexidade das transações financeiras e a quantidade de dados disponíveis aumentaram ao longo dos anos, tornou-se improvável a manutenção de técnicas manuais. Nesse contexto, a introdução de modelos estatísticos como metodologia para a concessão de crédito foi amplamente aceita pelas empresas, pois eles fornecem objetividade e precisão na avaliação do risco de crédito de um cliente.

Os modelos capazes de discriminar bons e maus pagadores são denominados de *Credit Scoring.* Pautando-se em variáveis cadastrais, informações de mercado externas e variáveis comportamentais, o intuito dessa abordagem é receber dados de entrada, entender a interação dos clientes em relação aos pagamentos e retornar a probabilidade do cliente tornar-se inadimplente baseando-se em teorias matemáticas.

A probabilidade retornada pelo modelo é muitas vezes expressa como uma pontuação de risco e transformada em uma espécie de ranking, sendo uma medida chave que orienta decisões importantes. Na avaliação de risco de crédito, os principais indicadores que

desempenham papéis críticos na quantificação do risco envolvido em empréstimos e operações de crédito são a PD, LGD e EAD.

Probability of Default – A Probabilidade de Inadimplência (PD) é um dos pilares centrais da análise de risco de crédito. Calculada através de modelos estatísticos avançados, essa métrica estima a chance de um cliente não honrar suas obrigações de pagamento.

$$\boldsymbol{PD = \frac{\sum Número\ de\ Inadimplências}{\sum Número\ Total\ de\ Operações}}$$

Loss Given Default – A Perda em Caso de Inadimplência (LGD) refere-se ao percentual da exposição que se espera ser perdida em caso de inadimplência. Essa métrica considera o percentual de recuperação esperado em caso de inadimplência.

$$\boldsymbol{LGD = \sum (1 - Taxa\ de\ Recuperação\ Esperada)}$$

Exposure at Default – A Exposição em Caso de Inadimplência (EAD) quantifica o valor de exposição sujeito a ser perdido durante uma concessão de crédito em caso de inadimplência do cliente. Essa métrica calcula o valor real esperado caso o cliente não pague o empréstimo.

$$\boldsymbol{EAD = \sum Valor\ de\ Exposição\ x\ LGD}$$

Perda Esperada – A Perda Esperada (EL) é o valor esperado de perdas que uma instituição financeira espera incorrer em suas exposições de crédito. Essa é a principal métrica de risco de crédito, pois representa uma estimativa das perdas prováveis com base na probabilidade de inadimplência, na recuperação esperada e no valor de exposição.

$$\boldsymbol{EL = \sum PD\ x\ LGD\ x\ EAD}$$

Segundo Lima (2008), os termos probabilidade de inadimplência (PD), perda dada a inadimplência (LGD) e exposição a inadimplência (EAD) começaram a dar um contorno mais

técnico aos cálculos necessários para a fixação do capital regulatório. Os valores eram estipulados, na grande maioria dos casos, em arbitrários e conservadores. Isso requeria um capital regulatório maior do que o necessário para fazer frente à perda. A abordagem avançada reconhecia esse conservadorismo e passou a permitir que cada instituição pudesse desenvolver modelos internos de fixação dos valores de PD, LGD e EAD. O grande benefício seria medir de forma mais adequada o risco de crédito de sua carteira e assim manter um capital regulatório adequado a essa exigência. Durante este processo, o maior desafio é atender as margens de risco que a instituição está disposta a correr de forma rentável e lucrativa, ao mesmo tempo que proporciona bons produtos e serviços ao cliente.

Para exemplificar, pode-se pensar no financiamento de um veículo. Supondo que o veículo esteja avaliado em R$100.000,00; o financiamento esteja atrelado à garantia de que em caso de inadimplência a instituição recuperará o carro e, consequentemente, 60% do valor inicial; sabe-se que a EAD é de R$60.000,00 e a LGD de 60%. Pensando em um cliente com PD de 10%, a EL pode ser definida como

$$\boldsymbol{EL = \ PD\ x\ LGD\ x\ EAD = \ 0.1\ x\ 0.6\ x\ R\$60.000,00 = R\$\ 3.600,00}$$

portanto, para um veículo avaliado em R$100.000,00, a perda esperada seria de R$3.600,00 para clientes com essa probabilidade de inadimplência. Dessa forma, percebe-se a importância da existência de um modelo matemático bem calibrado para cálculo de PD.

2.3 Modelos de Aprendizado de Máquina para Classificação

Aprendizado de Máquina é uma subárea da Ciência da Computação responsável pela confecção de algoritmos pautados em técnicas matemáticas e estatísticas, sendo um dos tópicos mais relevantes da área de Inteligência Artificial. O destaque do Machine Learning deve-se ao fato de que ele é capaz de reconhecer padrões complexos através de dados de entrada e então tomar novas decisões baseadas no aprendizado anterior. Essa característica é extremamente relevante, pois ela permite a antecipação de diversos cenários e possibilita a otimização de inúmeros processos. Pelo grande poder preditivo e capacidade de generalização, este tipo de modelagem proporciona a automação de processos de forma segura. Em virtude de as decisões serem tomadas automaticamente por algoritmos matemáticos continuamente, um bom modelo será robusto ao ser apresentado a dados novos,

denotando assim sua importância para ambientes dinâmicos e complexos. Sendo assim, nota-se que técnicas de Aprendizado de Máquina são muito importantes e possibilitam o avanço de muitos segmentos da sociedade.

Kumar (2022) afirma que “Inferência” e “Predição” são conceitos muito relevantes em questões de modelagem, todavia, embora tenham intersecções, são propostas diferentes. Do ponto de vista matemático, a inferência concentra-se em extrair conclusões sobre uma população através de uma amostra. Com um viés muito mais próximo da Estatística Clássica, ela é o processo de avaliar a relação entre a variável dependente e as variáveis independentes, ligando-se a questões como “Quais preditores estão associados com a variável alvo?”, “Como a mudança dos preditores influência na magnitude da variável alvo?”. Durante questões inferências, costuma-se utilizar testes de hipótese e intervalos de confiança para rejeitar ou não determinadas hipóteses. Por sua vez, a predição simplesmente pauta-se em determinados dados históricos a fim de realizar previsões futuras. Em suma, um modelo matemático é treinado para entender o comportamento médio de um sistema com informações passadas e prever o futuro com a maior assertividade possível. Embora não tenha um caráter estritamente estatístico, ainda sim modelos de predição são construídos com base em técnicas matemáticas e estatísticas avançadas.

Modelos de Classificação estão contidos no Aprendizado de Máquina Supervisionado. Segundo Géron (2019), no Aprendizado Supervisionado, os dados são apresentados ao algoritmo com os dados de entrada acompanhados dos resultados, chamados de rótulos. A partir dos rótulos, o modelo é treinado e estima uma função matemática capaz de classificar novas amostras.

Sendo assim, o principal objetivo desses modelos é utilizar dados históricos e previamente rotulados para construir uma representação matemática dos padrões presentes na amostra de treinamento. Essa representação é capturada na forma de um modelo capaz de generalizar e prever, para novas instâncias nunca vistas anteriormente, a qual classe elas pertencem.

Essa classe nada mais é do que a representação de uma probabilidade. Isso significa que, após passar pela equação, a nova instância terá uma determinada probabilidade de pertencer a classe negativa e outra probabilidade de pertencer a classe positiva. No contexto

de um modelo de crédito, esse novo elemento terá uma probabilidade estimada de ser qualificado para o empréstimo e uma probabilidade estimada de não ser qualificado.

Modelos Lineares de Regressão

Modelos de Regressão são uma ferramenta muito importante na modelagem estatística. Segundo Morettin e Singer (2022), uma regressão é uma técnica para modelar a relação entre variáveis independentes e uma variável dependente a fim de estimar o valor esperado de uma variável resposta. Os dois modelos lineares mais famosos desse tipo são a regressão linear (voltada para inferir uma variável dependente contínua) e a regressão logística (voltada para inferir uma variável dependente qualitativa).

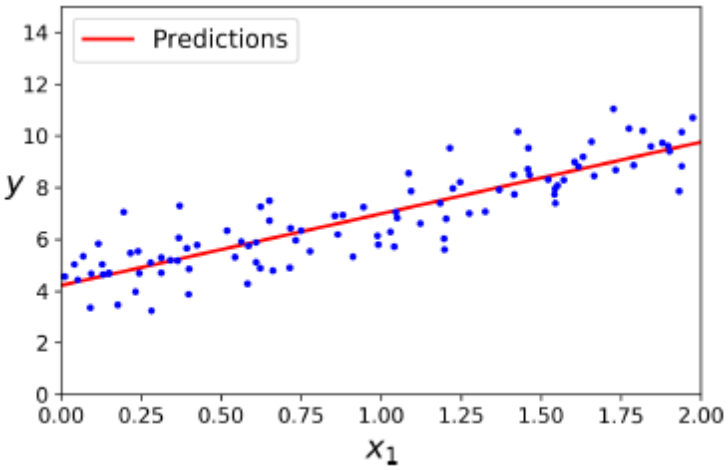
Em geral, a equação para um modelo de regressão linear pode ser definida como:

$$\boldsymbol{f(x) = \beta 0 + \sum_{i=1}^{N} BiXi + \sum_{i=1}^{N} \varepsilon i,}$$

em que $\beta 0$ é o intercepto, βi o coeficiente angular, *xi* são as variáveis utilizadas, εi corresponde ao erro aleatório os quais representam desvios entre os valores observados e preditos.

Bruce e Bruce (2019) afirmam que o intuito desse modelo é estimar os coeficientes que ajustam-se à amostra de treino, entendam os padrões e tornem-se uma equação representativa do fenômeno desejado. Em outras palavras, “estimar o quanto Y mudará quando X mudar em determinada quantidade”. Na figura 5, os pontos azuis são cada uma das amostras de uma base de dados e a reta em vermelho é representa a equação do modelo de regressão linear.

Figura 5 – Modelo de Regressão Linear



Fonte: Géron (2019, p.92).

Como a regressão linear possui variável dependente contínua, ela não pode ser utilizada para prever a classe de um indivíduo. Nesse âmbito, os estatísticos criaram a regressão logística, a qual é uma extensão da regressão linear, adaptada para lidar com problemas em que a variável de dependente é categórica e sua distribuição é uma distribuição de Bernoulli, como é o caso de eventos binários.

No contexto de modelagem de risco de crédito, a variável dependente é uma variável binária que indica se um indivíduo é considerado um bom ou mau pagador. Dessa forma, ao definir-se a classe positiva como 1 e a negativa como 0, a variável dependente será a probabilidade da instância de pertencer a classe 1.

Para que isso ocorre, necessita-se de uma forma capaz de transformar uma variável dependente contínua contida no intervalo [-∞, ∞] para uma variável categórica binária contida no intervalo [0,1]. A função responsável por transformar valores contínuos em termos de probabilidade é a Função Sigmóide, a qual é uma transformação não-linear. Dessa forma, após a aplicação da Sigmóide na equação a Regressão Linear, tem-se a equação da regressão logística:

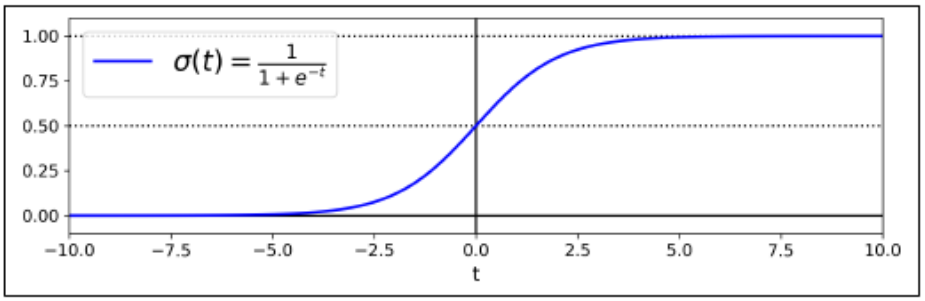
$$f(x) = \frac{1}{1 + e^{-(\beta 0 + \sum_{i=1}^{N} BiXi + \sum_{i=1}^{N} \varepsilon i)}}$$

sendo assim possível estabelecer uma probabilidade de corte para a classe de interesse acima da qual considera-se um registro como pertence àquela classe, ou, em outras palavras, para definir os limites de separação entre um bom e mau pagador.

Nesse contexto, há também os ajustes em relação aos parâmetros. Conforme Morettin e Singer (2022), os parâmetros são ajustados pelo método da máxima verossimilhança, o qual possui o objetivo de encontrar os coeficientes do modelo que possuem a maior probabilidade de obter uma distribuição análoga a dos dados reais.

A figura 6 traz a curva de uma regressão logística. Os coeficientes estimados representam a razão entre a probabilidade de sucesso e a probabilidade de falha da variável dependente para cada unidade de mudança nas variáveis independentes. Em outras palavras, eles representam a contribuição relativa de cada variável na previsão da probabilidade da classe e indicam como elas afetam as chances de um indivíduo ser um mau pagador.

Figura 6 – Modelo de Regressão Logística



Fonte: Géron (2019, p.114).

Modelos Bayesianos

Morettin e Bussab (2017) expressam que a probabilidade condicional é um conceito importante na estatística o qual permite lidar com incertezas em situações de eventos dependentes. Nesse âmbito, o Teorema de Bayes é um mecanismo o qual descreve a forma de atualizar a probabilidade de uma hipótese com base em novas evidências. Sua equação é definida como:

$$\boldsymbol{P(A|B) = \frac{P(B|A)P(B)}{P(B)}}$$

sendo P(A) a probabilidade do evento A, P(B) a probabilidade do evento B, P(B|A) a probabilidade a priori e P(A|B) a probabilidade a posteriori.

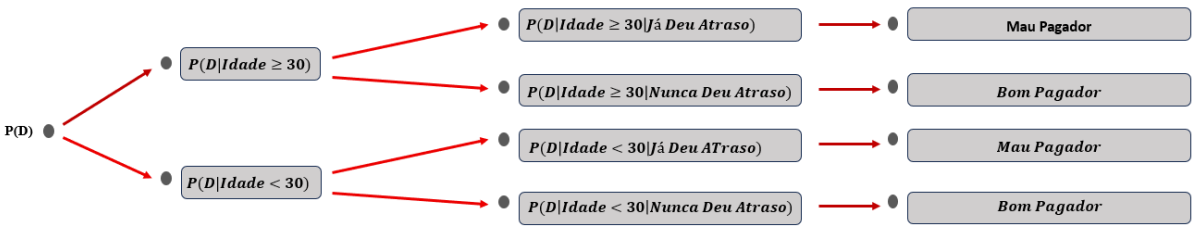
A partir do Teorema de Bayes, pode-se modelar um fenômeno dada as variáveis, criando-se o Naive Bayes. Bruce e Bruce (*2019*) descrevem o Naive Bayes como um algoritmo o qual pauta-se na probabilidade de observação de valores preditores, dado um resultado, para estimar a probabilidade de observar o resultado dado um conjunto de preditores. Em outras palavras, o algoritmo cria tabelas de contingência e calcula as probabilidades de cada classe para cada uma das variáveis e, a partir dos resultados, compreende se uma nova instância pertencerá a classe 1 ou 0.

Embora seja uma abordagem muito interessante, este modelo assume que as variáveis são independentes entre si, todavia, é muito raro encontrar um fenômeno de independência como este na prática. Sua equação é expressa como:

$$\boldsymbol{f(x) = \frac{P(x1, x2, \dots, xn|y)P(y)}{P(x1, x2, \dots xn)}}$$

sendo $P(x1, x2, \dots, xn|y)$ a probabilidade dos atributos $[x1, x2, \dots, xn]$ ocorrerem dada a classe y, $P(y)$ a probabilidade marginal da classe y e $f(x)$ a probabilidade condicional da nova instância ser um bom ou mau pagador dado um conjunto de atributos $[x1, x2, \dots, xn]$. A figura 7 representa o funcionamento do Naive Bayes para as variáveis "idade" e "se já deu atraso".

Figura 7 – Modelo de Naive Bayes exemplificado para uma Target (PD) e 2 Variáveis (Idade e Se Já Deu Atraso)



Fonte: Autoria Própria.

Modelos Baseados em Distância

Bruce e Bruce (*2019*) afirmam que modelos baseados em Distância destacam-se por sua simplicidade conceitual e abrangência. No ramo de classificação, o mais famoso é o KNN (K-Nearest Neighbors).

Como mostrado na figura 8, o KNN compara a nova instância com os K elementos mais próximos baseados nas variáveis utilizadas. Essa comparação é realizada via cálculos geométricos de distância e, a depender da distância escolhida, os resultados podem ser distintos. Dentre as distâncias mais famosas estão a Euclidiana, Manhattan e Similaridade de Cossenos.

A distância Euclidiana é a fórmula clássica para cálculos de distância em um espaço tridimensional. Através das coordenadas, ela mede a distância linear direta entre dois pontos no espaço. Além disso, ela é recomendada para variável contínuas.

$$\boldsymbol{f(x) = \sqrt{(xi - xn)^2 + (yi - yn)^2 + (zi - zn)^2}}$$

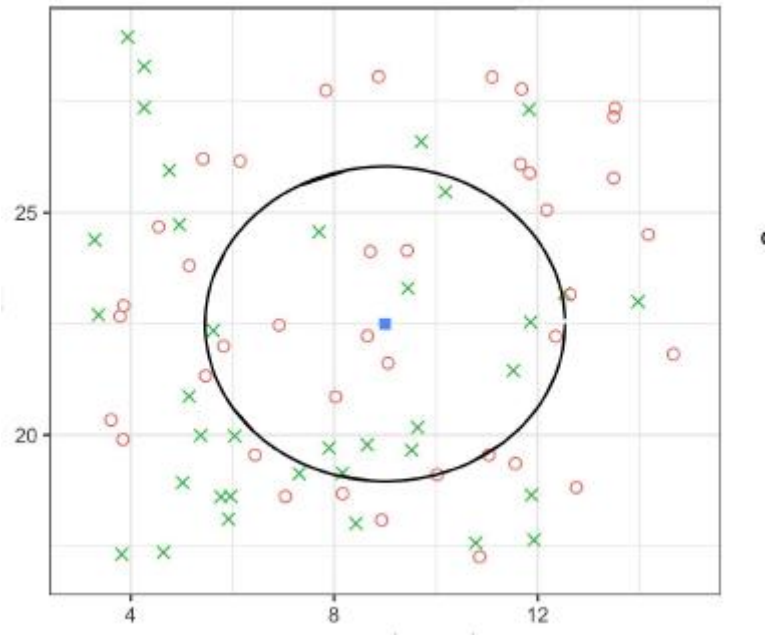
A distância de Manhattan é similar a Euclidiana, todavia, neste caso ela torna-se interessante em situações as quais a distância linear direta entre dois pontos não é permitida ou não atende a problema em questão. Dessa forma, a distância é calculada apenas em relação a ângulos retos. Assim como a Euclidiana, ela também é recomendada para variável contínuas.

$$\boldsymbol{f(x) = |xi - xn| + |yi - yn| + |zi - zn|}$$

Finalmente, a similaridade de cossenos mede a similaridade direcional entre dois vetores em um espaço tridimensional. Os cálculos são realizados através do produto escalar e da norma dos vetores e o resultado final varia de -1 a 1, ao passo que quanto maior o valor, maior a similaridade entre os vetores. Recomenda-se este tipo de distância tanto para variáveis contínuas quanto qualitativas.

$$\boldsymbol{f(x) = \frac{[xi, yi, zi].[xn, yn, zn]}{||[xi, yi, zi]||.||[xn, yn, zn]||}}$$

Figura 8 – Modelo de KNN para Classificação



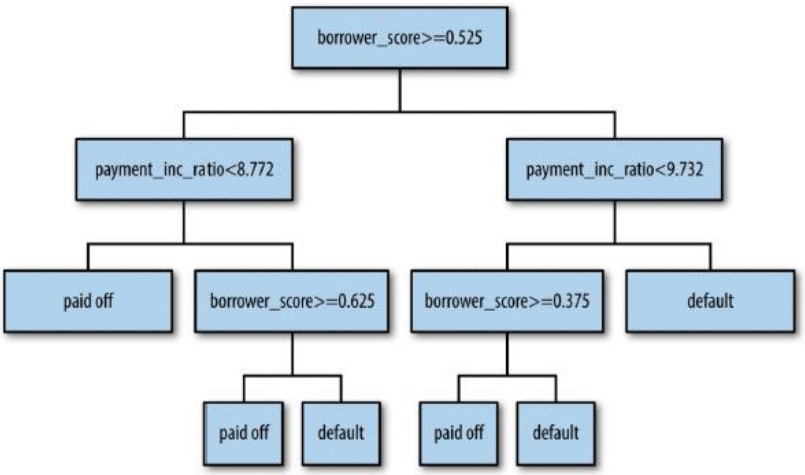
Fonte: Bruce e Bruce (2019, p.219).

Modelos de Árvore

Árvores de Decisão são modelos amplamente utilizados, sendo talvez os mais famosos. Conhecidas pelo seu alto poder preditivo e fácil entendimento, esses algoritmos tornaram-se extremamente relevantes tanto na literatura quanto nas mais variadas aplicações. Inspiradas na estrutura de uma árvore, são modelos baseados em regras sequenciais de fácil interpretabilidade.

Morettin e Singer (2022) relatam que uma Árvore de Decisão é uma espécie de fluxograma no qual as observações percorrem uma série de condições determinadas pelas variáveis do modelo a fim de resultar em uma decisão final. A partir da figura 9, nota-se que a árvore parte de um nó raiz, passando gradualmente pelos nós filhos de tal forma que escolhe-se o atributo mais informativo para realizar a divisão em cada etapa. Ao final do processo, encontra-se uma folha representativa da classe à qual a instância pertence.

Figura 9 – Modelo de Árvore de Decisão para Classificação



Fonte: Bruce e Bruce (2019, p.228).

A escolha da variável de maior relevância é feita recursivamente através de medidas de impureza as quais objetivam produzir subconjuntos mais puros. Quanto mais puro um nó, maior seu poder decisório. Nesse âmbito, as medidas mais famosas para analisar impureza são o Índice Gini e a Entropia.

Segundo Géron (2019), o Índice Gini mede a impureza de um nó com base na probabilidade de classificações incorretas ao selecionar dois elementos da amostra e atribuí-los a diferentes classes, portanto, quanto maior seu valor, mais impuro está o nó e, consequentemente, menor seu poder preditivo. Dessa forma, na construção de uma árvore, o algoritmo testará todas as variáveis e escolherá aquela à qual resulta no menor valor de Gini como nó raiz. Sua equação pode ser definida como:

$$\boldsymbol{f(x) = 1 - \sum_{i=1}^{N} p_i{}^2}$$

sendo p_i a proporção de observações da classe i.

Em relação a Entropia, Géron (2019) expressa que ela mede a pureza dos dados em determinado nó da árvore. Pautando-se na teoria da informação, ela reflete a incerteza em uma distribuição de probabilidade. Dessa forma, na construção de uma árvore, o algoritmo também testará todas as variáveis e escolherá aquela à qual resulta no menor valor de entropia como nó raiz. À medida que a árvore se aprofunda, espera-se que os nós filhos possuam valores de entropia maiores que o nó pai. Sua equação pode ser definida como:

$$f(x) = -\sum_{i=1}^{N} p_i \log_2(p_i)$$

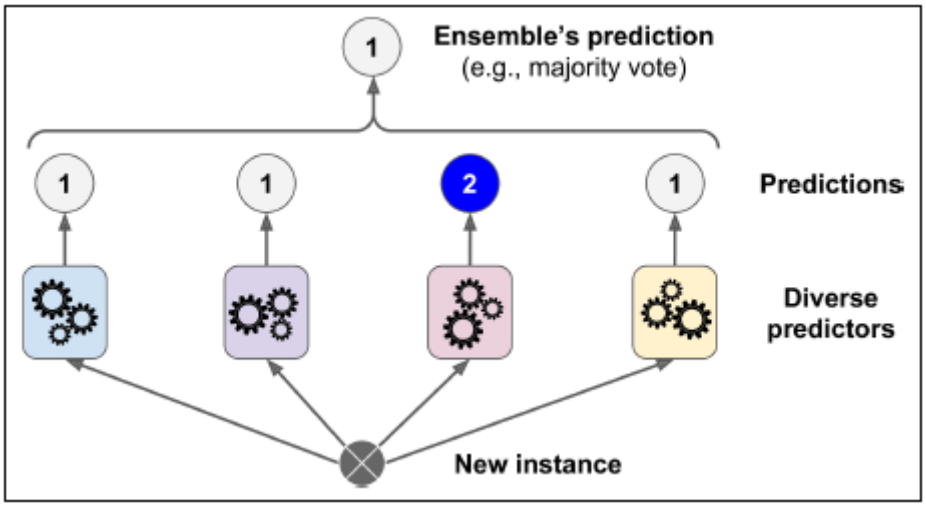
sendo p_i a proporção de observações da classe i.

A principal diferença entre essas medidas está na maneira de avaliação da impureza. Géron (2019) declara que enquanto o Índice Gini concentra-se nas classificações incorretas, a Entropia baseia-se na incerteza da distribuição de probabilidade. Em termos práticos, ambas geralmente levam à mesma decisão, todavia, em certos casos há leves divergências dependendo da problemática.

Modelos de Ensemble

O avanço do poder computacional proporcionou o desenvolvimento de técnicas mais robustas as quais foram responsáveis por grandes resultados, sendo a mais famosa denominada como Ensemble. Conforme a figura 10, métodos Ensemble são algoritmos capazes de combinar diversos preditores fracos a fim de criar um preditor forte e robusto, assegura Géron (2019). Há três tipos de métodos de Ensemble: Bagging, Boosting e Stacking. Para este trabalho, abordou-se os dois primeiros.

Figura 10 – Representação de um Ensemble para Classificação

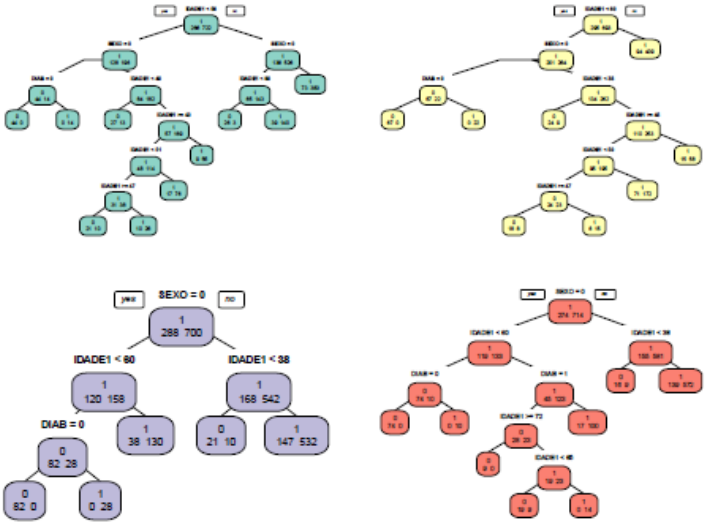


Fonte: Géron (2019, p.148).

Morettin e Singer (2022) alegam que a técnica de Bagging é um método para gerar múltiplas versões de um preditor a partir de vários conjuntos de treinamento a fim de diminuir a variância desse modelo e proporcionar predições mais fidedignas. O principal algoritmo

para esta metodologia é a Random Forest, a qual é criada a partir da combinação de diversas Árvores de Decisão. Cada árvore da Random Forest é criada através de amostragens aleatórias e com reposição de linhas e colunas da base de dados de tal forma que todos os elementos possuem a mesma probabilidade de serem selecionadas. Em casos de classificação, a classe final será àquela com maior ocorrência entre os diversos preditores. A figura 11 mostra a combinação de Árvores de Decisão durante a criação de uma Random Forest.

Figura 11 – Modelo de Random Forest para Classificação



Fonte: Morettin e Singer (2022, p.329).

O Boosting é uma técnica amplamente reconhecida e altamente poderosa no campo de aprendizado de máquina. Ao contrário do Bagging, o Boosting, como explicado por Morettin e Singer (2022), envolve a geração sequencial de árvores de decisão com base na atualização de pesos para cada elemento no conjunto de treinamento. Como mostrado na figura 12, o processo inicia-se com a criação de uma árvore de decisão inicial, seguida pelo cálculo dos resíduos iniciais que representam a diferença entre as probabilidades a priori e a probabilidade a posteriori.

Os exemplos classificados incorretamente recebem pesos mais altos, tornando-os mais influentes nas iterações subsequentes. Essas árvores, construídas sequencialmente, visam corrigir as classificações errôneas e aprimorar a precisão do modelo. O Gradient Boosting é o exemplo mais conhecido dessa técnica e é notável por seu uso de métodos de otimização

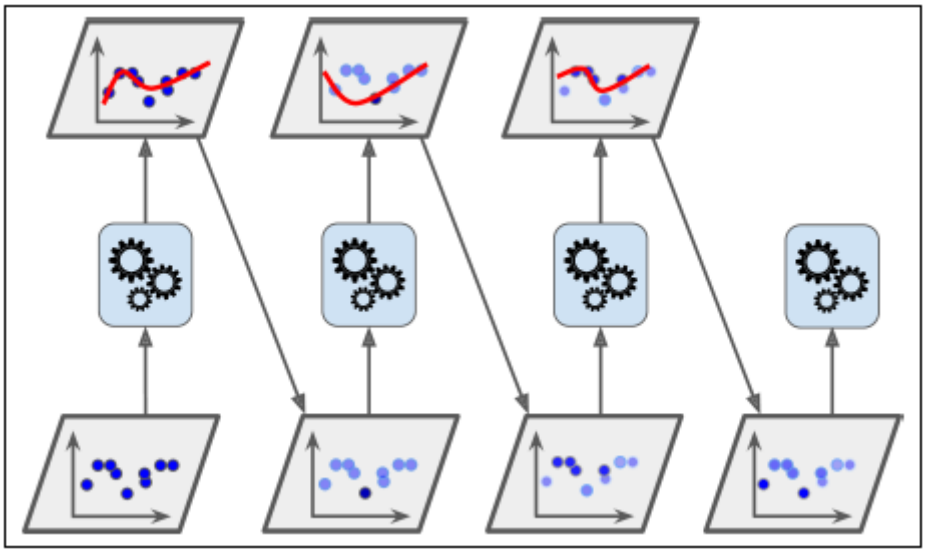
iterativa, conferindo-lhe uma capacidade preditiva fantástica. O objetivo principal do Boosting é reduzir o viés do modelo por meio de correções progressivas dos erros de previsão, como descrito por Géron (2019).

O cálculo dos resíduos citados acima é realizado a partir de uma função de custo denominada LogLoss, a qual é representada pela equação a seguir:

$$L(y_t, y_p) = -\frac{1}{N}\sum_{i=1}^{N}[y_t \log y_p + (1 - y_t)\log(1 - y_p)]$$

onde y_t corresponde ao valor real, o y_p ao valor predito e o N o número de amostras. Após o somatório, função calcula a média dos erros a fim de dimensionar de maneira geral a performance do modelo. Caso o valor real seja 1 e o valor predito 1, isso significa que estamos acertando a maioria dos registros, portanto, a função de custo tenderá a 0, logo, o modelo pode ser considerado bom. Caso o valor real seja 1 e o valor predito 0, a função de custo tenderá ao infinito, indicando que o modelo necessita de ajustes.

Figura 12 – Modelo de Gradient Boosting para Classificação



Fonte: Géron (2019, p.155).

2.4 Métricas de Avaliação

Como todo modelo de Aprendizado de Máquina está relacionado a probabilidades, inevitavelmente ele comentará erros, portanto, é de suma importância compreender sua performance. Para entender seu desempenho, abordou-se as principais metodologias e métricas de problemas de Classificação para concessão de crédito, sendo elas: acurácia, precisão, sensibilidade, AUC, KS, Hold-Out e Validação Cruzada.

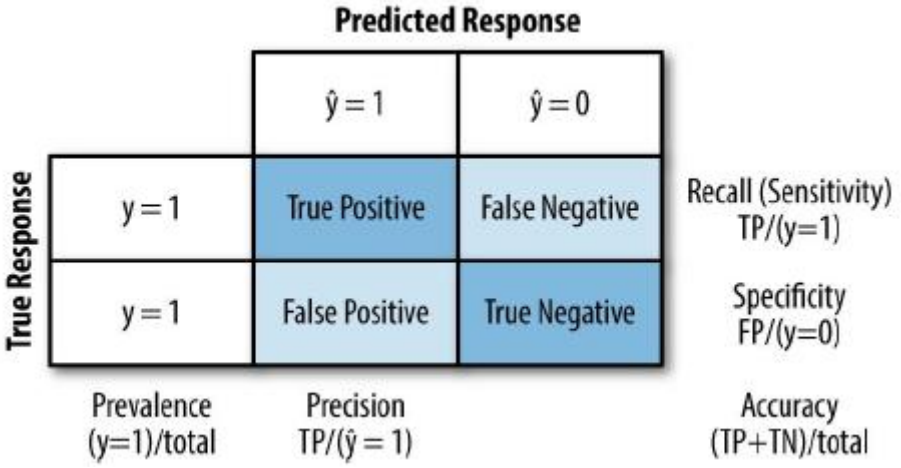
Matriz de Confusão

Conforme Morettin e Bussab (2017), Inferência Estatística é o processo de fazer afirmações sobre as características de uma população com base em informações dadas por amostras. Ao submetermos uma amostra a um modelo probabilístico, qualquer que seja a decisão tomada, existem dois tipos de erro: Erro Tipo I e Erro Tipo II. O primeiro ocorre ao rejeitar-se a hipótese nula quando ela é verdadeira e o segundo ocorre ao não se rejeitar a hipótese nula quando ela é falsa.

Sicsú (2010) amplia essa perspectiva ao contextualizar esses conceitos em problemas de concessão de crédito. No cenário financeiro, o Erro Tipo I ocorre quando recusa-se uma operação que seria lucrativa para o credor caso acontecesse e o Erro Tipo II quando aprova-se uma operação a qual dará prejuízo à instituição. Para representar essas situações e quantificar a performance de um modelo de classificação, os estatísticos desenvolveram a Matriz de Confusão.

De acordo com Bruce e Bruce (2019), uma Matriz de Confusão é uma matriz quadrada utilizada para comparar os valores preditos do modelo com os valores reais. Representada na figura 13, sua diagonal é composta pelos acertos do modelo e os demais valores são erros cometidos. Durante a classificação de um elemento, há quatro situações possíveis, sendo elas Verdadeiro Negativo, Verdadeiro Positivo, Falso Negativo e Falso Positivos.

Figura 13 – Matriz de Confusão para Classes Binárias



Fonte: Bruce e Bruce (2019, p.201).

Um Verdadeiro Negativo (VN) refere-se aos casos em que o modelo previu corretamente uma instância como pertencente à classe negativa e de fato ela pertence à classe negativa, ou seja, o cliente é adimplente e o modelo afirma que ele pagará o empréstimo.

Um Verdadeiro Positivo (VP) refere-se aos casos em que o modelo previu corretamente uma instância como pertencente à classe positiva e de fato ela pertence à classe positiva, ou seja, o cliente é inadimplente e o modelo afirma que ele não pagará o empréstimo.

Um Falso Negativo (FN) refere-se aos casos em que o modelo previu incorretamente uma instância como pertencente à classe negativa mas ela pertence à classe positiva, ou seja, o cliente é inadimplente e o modelo afirma que ele pagará o empréstimo.

Um Falso Positivo (FP) refere-se aos casos em que o modelo previu incorretamente uma instância como pertencente à classe positiva mas ela pertence à classe negativa, ou seja, o cliente é adimplente e o modelo afirma que ele não pagará o empréstimo.

A seguir, conforme Bruce e Bruce (2019), Sicsú (2010) e Géron (2019), explicou-se as principais métricas de avaliação para modelos de classificação:

Acurácia

A acurácia é uma métrica simples a qual quantifica a proporção de previsões corretas feita pelo modelo em relação ao total de previsões. Embora indique a performance geral do modelo, em problemas de classes desbalanceadas ela não performa bem, pois pode ser enganosa visto que nesses casos uma classe é muito mais comum que a outra.

$$\boldsymbol{Acurácia} = \frac{\boldsymbol{VN + VP}}{\boldsymbol{VN + VP + FN + FP}}$$

Precisão

A precisão quantifica a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias classificadas como positivas pelo modelo. Em suma, ela representa a capacidade de um modelo em prever corretamente a classe positiva, portanto, bons valores de precisão significam a diminuição do Erro Tipo I.

$$\boldsymbol{Precisão} = \frac{\boldsymbol{VP}}{\boldsymbol{VP + FP}}$$

Sensibilidade

A sensibilidade quantifica a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias que realmente são positivas na amostra. Em suma, ela representa a capacidade de um modelo em capturar a classe positiva, portanto, bons valores de sensibilidade significam a diminuição do Erro Tipo II.

$$\boldsymbol{Sensibilidade} = \frac{\boldsymbol{VP}}{\boldsymbol{VP + FN}}$$

F1-Score

O f1-score representa a média harmônica entre a precisão e a sensibilidade. Ele é uma métrica valiosa para problemáticas de dados desbalanceados, ou seja, quando há muito mais amostras de determinada classe em relação a outra. Em suma, ela representa o equilíbrio entre a capacidade do modelo de prever a classe positiva e de capturar a classe positiva, portanto, bons valores de f1-score significam um bom desempenho do modelo de classificação.

$$\boldsymbol{F1 - Score} = \frac{\boldsymbol{2(precisão\ x\ sensibilidade)}}{\boldsymbol{precisão + sensibilidade}}$$

Curva ROC e AUC

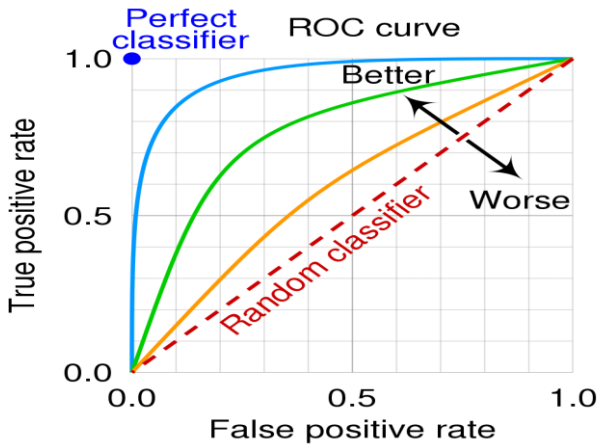
Após prévio entendimento da precisão e sensibilidade, nota-se que há uma relação entre elas, ou seja, inferir a classe positiva impõe que mais amostras da classe negativa serão classificadas como positivas e vice-versa. Dada a situação, o classificador ideal atenderia a classificação de ambas as classes corretamente. A métrica responsável por avaliar essa troca denomina-se “Curva de Característica Operatória Receptora” (Curva ROC).

A curva ROC é uma representação gráfica da taxa de VP em função da taxa de FP para diferentes pontos de corte em um modelo de classificação. A partir dessa curva criada, pode-se calcular a área sob a curva (AUC), a qual é uma métrica contida no intervalo [0,1], onde um valor maior indica melhor desempenho do modelo.

Como expresso na figura 14, a reta pontilhada corresponde ao desempenho de um classificador aleatório, portanto, uma AUC abaixo da curva linear automaticamente invalidaria o modelo. Embora relativo, Sicsú (2010) conta que muitos especialistas categorizam um modelo como “bom/ideal” quando este possui AUC igual ou superior a 0.7.

$$\boldsymbol{Taxa\ de\ Verdadeiros\ Positivos} = \frac{\boldsymbol{VP}}{\boldsymbol{VP + FN}};\ \boldsymbol{Taxa\ de\ Falsos\ Positivos} = \frac{\boldsymbol{FP}}{\boldsymbol{VN + FP}}$$

Figura 14 – Curva ROC e AUC



Fonte: Receiver [...] (2024).

Teste de Kolmogorov-Smirnov (KS)

Outra métrica muito famosa para modelos de crédito é o teste de hipótese de Kolmogorov-Smirnov (KS). Este teste estatístico é empregado para avaliar a capacidade

discriminativa de um modelo, medindo a diferença acumulada entre as distribuições de probabilidade das classes de bons e maus clientes.

Sicsú (2010) afirma que o valor do KS é calculado como a maior distância entre as curvas de distribuição de probabilidade acumulada, podendo variar no intervalo [0, 1]. Quanto mais próximo de 1, mais evidente a separação entre as duas classes, indicando melhor poder de discriminação do modelo. Pode-se definir sua equação como

$$F_{bons}(k) = \frac{\textbf{número de bons com Score} \leq K}{\textbf{número de bons}}$$

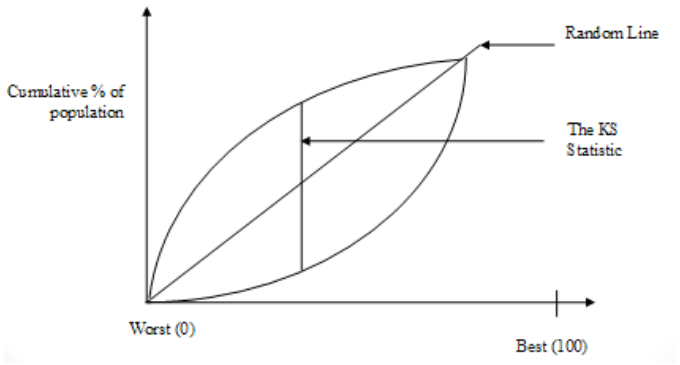
$$F_{maus}(k) = \frac{\textbf{número de maus com Score} \leq K}{\textbf{número de maus}}$$

$$KS = Max[F_{bons}(k) - F_{maus}(k)]$$

em que *k* varre o conjunto de possíveis valores do score. A figura 15 representa com clareza a diferença entre as distribuições.

Assim como a Curva ROC, há valores predefinidos de KS capazes de avaliar a eficácia de um modelo. Novamente, Sicsú (2010) apresenta uma tabela composta de pontuações na qual considera-se um modelo "bom/ideal" caso possua KS igual ou superior a 0.3.

Figura 15 – Kolgomorov-Smirnov (KS)



Fonte: Bhalla (2016).

3 DESENVOLVIMENTO

3.1 Produto

O trabalho “Análise de Risco de Crédito direcionada por Modelagem Matemática e Aprendizado de Máquina” apresenta uma introdução ao mundo do crédito auxiliado por técnicas de Matemática Aplicada, Estatística e Machine Learning para a classificação de clientes inadimplentes, a fim de auxiliar a tomada de decisão. Direcionada pelas técnicas e ferramentas de Ciência de Dados, o software desenvolvido visa aplicar o conhecimento assimilado durante o curso de Engenharia de Controle e Automação em processos de análise de risco, sendo responsável pela concessão de crédito para pessoas físicas de forma rápida e automática, pautando-se exclusivamente em conceitos matemáticos e estatísticos a fim de maximizar a confiança e minimizar o risco do credor.

3.2 Requisitos

Para assegurar a eficiência do projeto, é imprescindível que o software execute a concessão de crédito de forma totalmente automatizada, utilizando métodos avançados de aprendizado de máquina. A capacidade de classificar com precisão os bons e maus pagadores possibilitará uma definição mais acurada do perfil de risco do cliente, de modo que isso resultará em uma maior oferta de crédito para clientes com alta probabilidade de adimplência e reduzirá a perda esperada no portfólio.

Para comprovar sua aplicabilidade, objetiva-se conquistar resultados superiores as de uma política de crédito convencional baseada em variáveis majoritariamente consolidadas no mercado de crédito. Os resultados serão expressos em função das métricas de avaliação, sendo elas a acurácia, precisão, sensibilidade, f1-score, AUC, KS e ROCP, bem como a criação de um *ranking* para definir o correto perfil de risco do cliente.

Sendo assim, com base nos requisitos citados, deseja-se atender às seguintes etapas:

- Definição clara do Bom e Mau Pagador
- Análise e Seleção de Variáveis
- Criação da Política de Crédito e dos Modelos de Aprendizado de Máquina
- Comparação de Resultados entre ambas metodologias através de métricas de Machine Learning e Risco de Crédito

- Escolha da melhor metodologia para entrar em vigência

3.3 Materiais

Python

Conforme Levada (2021), a necessidade de lidar com grandes volumes de dados pressupõe, de maneira mandatória, a utilização de linguagens de programação. Como há muitas opções, deve-se escolher uma que atenda melhor as necessidades do problema. Pela versatilidade, facilidade e extensa documentação, optou-se pelo Python, visto que ela é a linguagem de programação mais utilizada mundialmente.

Criada em 1991 como uma linguagem de alto nível e orientada a objetos, ela prova diariamente sua robustez e importância. Combinando uma sintaxe bem definida e legível, ela possui inúmeros Frameworks de matemática, estatística e computação científica, portanto, fornece base completa para trabalhar com aprendizado de máquina.

Além disso, ela é utilizada em muitas outras áreas da computação, facilitando a integração de softwares e profissionalizando todo o ecossistema. Por fim, ela conta com a maior comunidade do mundo, logo, correção de bugs e inconsistências tornam-se naturalmente mais rápidas. Dessa forma, utilizou-se o ecossistema Python para a implementação dos códigos responsáveis pelas análises estatísticas, modelos matemáticos, otimizações, métricas de avaliação e políticas de crédito.

GIT

Trabalhos que envolvem software sempre estão sujeitos a mudanças, visto que os produtos e serviços são muito dinâmicos e a sociedade está em constante evolução, portanto, é muito comum que o código-fonte sofra alterações com o passar do tempo. Nesse âmbito, uma técnica muito importante são os sistemas de controle de versão, os quais são responsáveis pelo registro de todas as modificações de código. A partir deles, pode-se criar versões de um software, possibilitando flutuar em estágios antigos e novos sem comprometer o projeto.

O GIT é o sistema de controle de versão mais famoso mundialmente. Baseando-se em repositórios, ele permite que as versões de um software fiquem salvas. Além disso, ele fornece diversos ganhos como proteção criptográfica, alto desempenho e a possibilidade dos

desenvolvedores programarem ao mesmo tempo, tornando-se assim fundamental para o cotidiano de um profissional de tecnologia. Para este trabalho, utilizou-se o GIT para salvar as versões do código-fonte no servidor do GitHub (serviço gratuito para gerenciar repositórios) através de um repositório de código aberto a fim de incentivar a colaboração no meio acadêmico e permitindo que o público interessado possa consultar o projeto.

Kaggle

A plataforma Kaggle é a maior comunidade de cientistas, engenheiros e analistas de dados do mundo. Nela, encontram-se inúmeras bases de dados fictícias e reais destinadas a profissionais que desejam treinar suas habilidades técnicas e analíticas. Ela também é responsável por realizar competições, as quais são desafios impostos por empresas do mundo todo com uma premiação em dinheiro para quem solucionar o problema.

Pelo fato de risco de crédito envolver diversas informações sensíveis, utilizar uma base de dados verdadeira é praticamente impossível sem passar pelos processos burocráticos de confiabilidade, dessa forma, optou-se por uma amostra de colunas e registros da base de dados do Kaggle (Lending Club Loan Data). A problemática descreve a situação de uma instituição financeira denominada Lending Club, a qual é uma empresa norte-americana responsável por operar uma plataforma online de empréstimos. O intuito da empresa é contar com o capital de investidores para conceder crédito a pessoas que procuram empréstimos. Após a concessão, espera-se que no futuro os investidores recebam o capital somado aos juros, simulando um ambiente real de uma instituição financeira. Os dados foram devidamente anonimizados e alterados a fim de garantir a segurança e a confiabilidade das informações sensíveis, transformando-se em um conjunto fictício.

Ambiente de Desenvolvimento

Desenvolveu-se o software em um ambiente local utilizando um computador com especificações robustas. O sistema conta com 32GB de RAM, uma Placa de Vídeo RTX 2060 Super com 8GB de RAM dedicada, um Processador Ryzen 5600x com 6 núcleos de processamento, e um SSD de 2 TB de armazenamento. Essas configurações foram essenciais para garantir um ambiente de desenvolvimento ágil e eficiente, permitindo a manipulação de grandes volumes de dados e processamento intensivo necessários para o treinamento dos modelos de aprendizado de máquina e demais testes do software de forma eficaz.

3.4 Metodologia

De acordo com Lima (2021), ao longo dos anos, a capacidade computacional aumentou exponencialmente, portanto, uma quantidade astronômica de dados passou a ser gerada diariamente. Para utilizar todo o potencial de seus dados, demandou-se a criação de uma metodologia para projetos de Ciência de Dados, assim dando início ao CRISP-DM.

CRISP-DM é um processo de Mineração de Dados criado em 1996. Sua principal função é dividir projetos complexos de Ciência de Dados em partes menores, facilitando a execução das tarefas e o entendimento dos pontos por parte de pessoas não-técnicas. Pautando-se na constante evolução, essa metodologia permite que todas as etapas sejam revisitadas ao longo do projeto, corrigindo falhas e agilizando as entregas. A seguir há uma descrição das etapas:

Entendimento do Negócio - Inicialmente, leva-se em consideração todo o contexto do negócio e da empresa e define-se o objetivo do projeto, identificando as necessidades e alinhando as expectativas. Esta provavelmente é a parte mais importante do CRISP-DM, pois um problema mal definido condena as demais etapas a caminhos distintos do esperado.

Compreensão dos Dados – Posteriormente, realiza-se uma análise exploratória dos dados a fim de compreender as informações presentes em cada uma das variáveis, bem como suas características estatísticas a fim de inferir seu comportamento. Além disso, nesta etapa também são aplicados testes de hipótese para levantar possíveis insights visando enriquecer a modelagem.

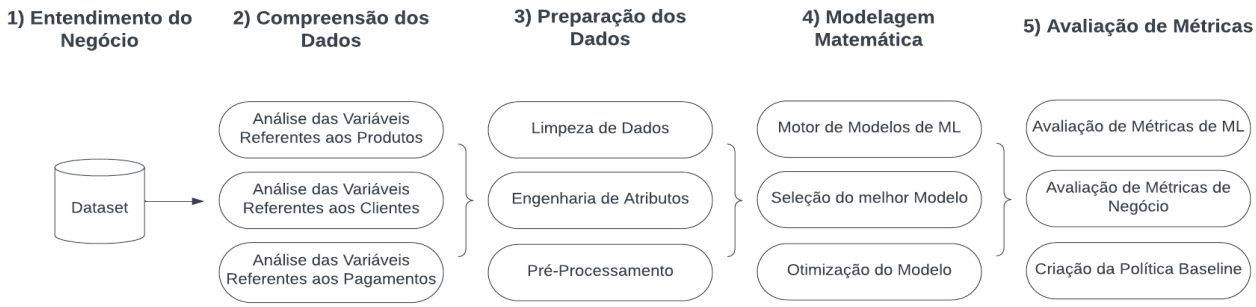
Preparação dos Dados – Esta etapa é responsável pelo tratamento e limpeza de dados, combinação de variáveis a fim de criar novos atributos e transformação dos dados no formato correto através de algumas transformações matemáticas e estatísticas para auxiliar a modelagem. Caso os dados estejam em formatos diferentes, alguns modelos de aprendizado de máquina funcionarão da forma errada, prejudicando o reconhecimento de padrões esperado.

Modelagem – A parte da modelagem consiste na aplicação dos métodos quantitativos definidos para solucionar o problema em questão. Durante o processo, os melhores modelos são testados e otimizados para que consigam realizar a tarefa corretamente. A má escolha de um modelo pode comprometer o projeto, portanto, esta etapa é fundamental.

Avaliação – Nesta fase avalia-se a performance do modelo escolhido. Inicialmente define-se as métricas mais importantes para o problema em questão e, caso apresente bons resultados durante a avaliação, constata-se que o modelo está apto para ser levado adiante.

Como problemas de análise de crédito tendem a ser complexos, extensos e muito importantes para uma instituição, aplicou-se a metodologia CRISP-DM como direcionamento para o problema de concessão de crédito proposto a fim de facilitar o andamento do projeto como um todo. Dessa forma, o fluxo *end-to-end* do projeto é expresso na figura 16:

Figura 16 – Descrição detalhada das Etapas do CRISP-DM



Fonte: Autoria Própria.

3.5 Definição da Target

Inicialmente, definiu-se o que deveria ser considerado um mau pagador. Esta categorização é de suma importância, pois ela representa o evento que o modelo tentará estimar, portanto, critérios bem definidos são cruciais para o correto aprendizado do modelo. Definições ambíguas ou mal construídas impedem a correta interpretação das variáveis preditoras, prejudicando drasticamente os resultados retornados pelo modelo.

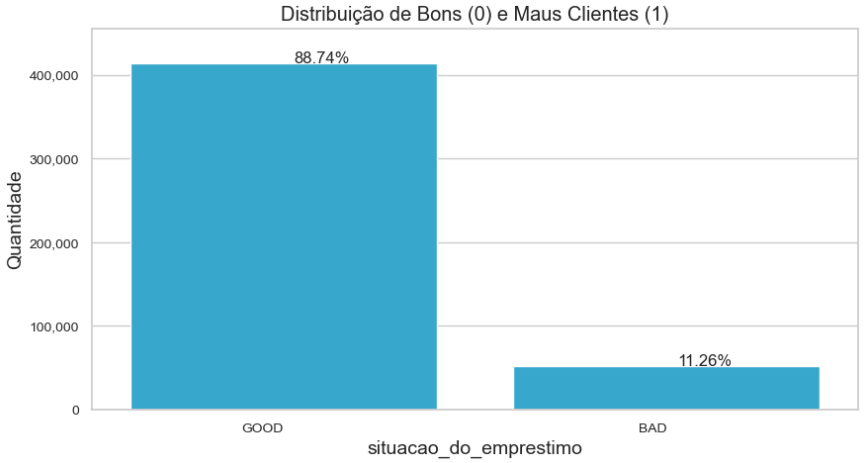
Outro grande problema de não se ter clareza sobre o que é um mau pagador consiste na má interpretação das análises. Dependendo dos critérios, determinados resultados anteriormente positivos podem facilmente ser interpretados como negativos. Para este projeto, decidiu-se que um mau pagador atenderia aos seguintes critérios:

- Estar em um processo de cobrança (isso significa que o cliente atualmente é inadimplente e já foi encaminhado para a cobrança)

- Estar inadimplente (recentemente tornou-se inadimplente e, em caso de manutenção desse status, em breve será encaminhado para a cobrança)
- Posse de 1 ou mais contas em outras instituições em estado de inadimplência (está inadimplente em outras instituições financeiras)
- Não atende ao CMA (o cliente não atende aos critérios mínimos de aprovação do empréstimo. Ex: menor de idade)

Com as regras estabelecidas, escolheu-se o número “0” para representar o cliente “Bom” e o número “1” para representar o cliente “Mau”, visto que o intuito do projeto é definir com precisão quem são os “Maus Pagadores” e assim evitar ao máximo casos de inadimplência. A partir desta definição, a partir da figura 17 pode-se visualizar a distribuição de “Bons e Maus Pagadores”, revelando a existência de um problema de classes desbalanceadas. Tal disparidade é importante e deve-se levar em consideração durante o processo de modelagem a fim de aplicar técnicas específicas para lidar com este tipo de caso.

Figura 17 – Distribuição de Bons e Maus Pagadores



Fonte: Autoria Própria.

3.6 Métodos de Avaliação

Como deseja-se prever o mau pagador, as principais métricas técnicas de serem avaliadas são o *Recall*, a *AUC* e o *KS*. A primeira é importante pois fornece a correta identificação dos verdadeiros maus pagadores; a segunda avalia o desempenho do modelo ao longo da faixa de limiares de decisão, sendo assim, ela fornece a capacidade global do modelo em prever todas as instâncias corretamente; a terceira avalia a diferença entre as distribuições

da classe negativa e positiva, dessa forma, ela mostra o quão bem o modelo separa os bons pagadores dos maus pagadores.

Em relação as métricas de negócio, definiu-se uma metodologia capaz de estimar o impacto financeiro de cada tipo de erro e acerto. Um Verdadeiro Negativo (VN) corresponde ao cliente adimplente classificado corretamente, dessa forma, representa um lucro do valor de exposição somado aos juros; um FN corresponde ao cliente inadimplente classificado incorretamente, sendo assim, representa a perda do valor de exposição; um FP corresponde ao cliente adimplente classificado incorretamente, portanto, não tem ganhos nem perdas; e um VP é o cliente inadimplente classificado corretamente, logo, também não tem ganhos nem perdas. A partir dessa regra, criou-se a equação para cálculo do retorno financeiro, bem como a estimativa de lucro em relação ao valor total de exposição (ROCP):

$$\begin{bmatrix} Qt\ Verdadeiro\ Negativo & Qt\ Falso\ Positivo \\ Qt\ Falso\ Negativo & Qt\ Verdadeiro\ Positivo \end{bmatrix} . \begin{bmatrix} Exposição + Juros & 0 \\ Exposição & 0 \end{bmatrix}$$

$$Retorno\ Financeiro = Qt\ VN * (Exposição + Juros) - Qt\ FN * Exposição$$

$$Return\ on\ Credit\ Portfolio\ = \frac{Retorno\ Financeiro}{Valor\ Total\ de\ Exposição}\ x\ 100\%$$

3.7 Análise de Variáveis

Definir o perfil de risco de um cliente costuma ser bastante desafiador, portanto, tarefas desse tipo envolvem fatores interconectados e de difícil percepção. Nesse âmbito, o ponto de partida para entender o risco de crédito das operações consiste na seleção criteriosa das melhores variáveis, pois elas carregam as informações necessárias para a criação de uma abordagem capaz de identificar o bom e mau pagador.

Posteriormente, procurou-se entender exatamente o significado de cada variável. Tal definição é importante pois a correta compreensão de cada informação é crucial para comprovar a consistência da análise. Essa clareza contribui com a interpretação e com o extermínio de ambiguidades, permitindo que o conceito seja devidamente entendido e as análises sejam esclarecedoras. Na seção de *Apêndice* encontram-se as definições de cada uma das variáveis utilizadas no projeto.

Durante o processo de análise de variáveis, deve-se levar em consideração tanto o pensamento conceitual quanto o pensamento quantitativo. Ao identificar uma variável com potencial, objetiva-se entender se sua relação com a inadimplência conceitualmente faz sentido e, então, provar a hipótese com alguma abordagem consolidada na literatura. A combinação desses caminhos é interessante por duas justificativas: a primeira é que se ambas chegarem ao mesmo resultado, a probabilidade da hipótese ser verdadeira torna-se maior; a segunda é que a validação através de uma metodologia quantitativa impede a tomada de decisão exclusivamente criada através de vieses previamente estabelecidos.

Como o intuito deste trabalho é realizar uma análise de risco de crédito direcionada por métodos quantitativos, optou-se pelo foco majoritário em análises estatísticas, sendo as principais o Weight of Evidence (WOE) e a observação das distribuições de probabilidade das variáveis.

Weight of Evidence é uma medida estatística muito consolidada em risco de crédito, pois é eficaz e fácil de ser entendida. Basicamente, o WOE avalia a força da associação de uma classe com a variável alvo. A partir da criação de uma tabela de contingência, realiza-se o cálculo da taxa de eventos positivos e negativos para então determinar o quanto a classe associa-se com o evento de interesse. Pode-se definir sua equação como

$$WOE = \ln\left[\frac{P(c|bom)}{P(c|mau)}\right]$$

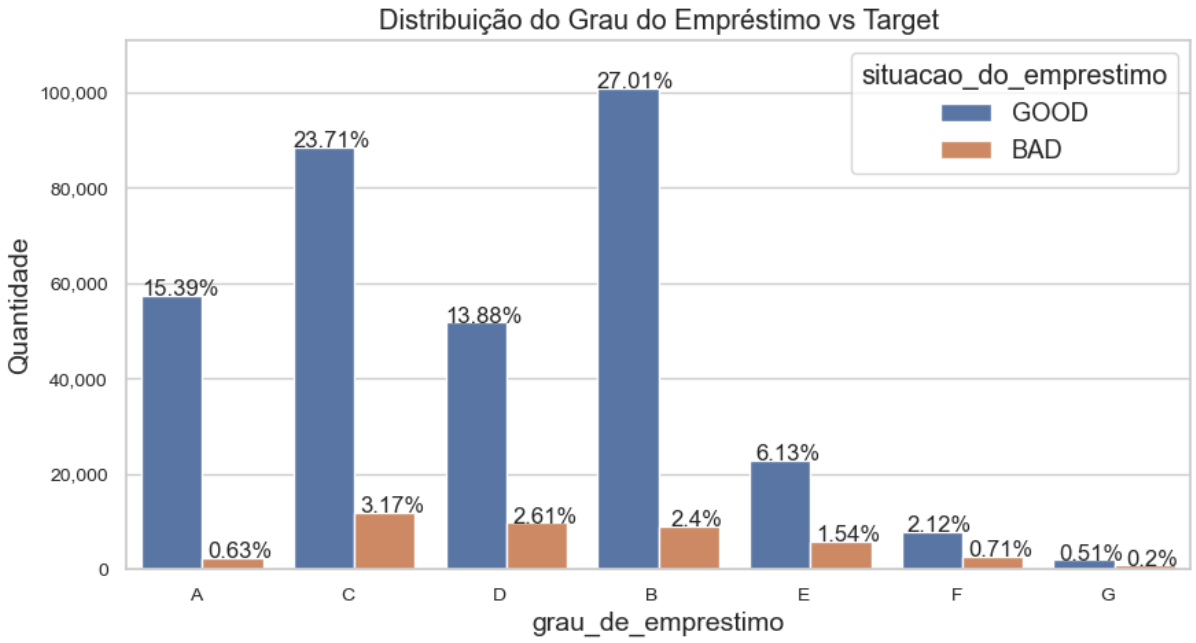
onde *ln* é o logaritmo neperiano, *c* é categoria de determinada variável preditora, P(c|bom) e P(c|mau) correspondem a frequência de bons e maus clientes para essa categoria respectivamente. Valores positivos de WOE significam que a categoria está associada a classe positiva, ao passo que valores negativos mostram associação com a classe negativa.

Embora robusta, este tipo de abordagem funciona apenas para variáveis categóricas e, na grande maioria dos casos, existem variáveis contínuas igualmente importantes no processo

de decisão. Neste caso, pode-se aplicar métodos de discretização para então, através de gráficos e análises visuais, compreender se essa variável possui relação com o evento de interesse. Uma metodologia famosa consiste na criação de decis que, ao serem analisados em conjunto com o evento de interesse, provam a importância dessa variável para compreensão do perfil de risco do cliente. Para exemplificar, decidiu-se trazer dois exemplos para representar os casos citados. A análise completa das demais variáveis segue a mesma abordagem e pode ser encontrada no código fonte disponibilizado no repositório do *GitHub*.

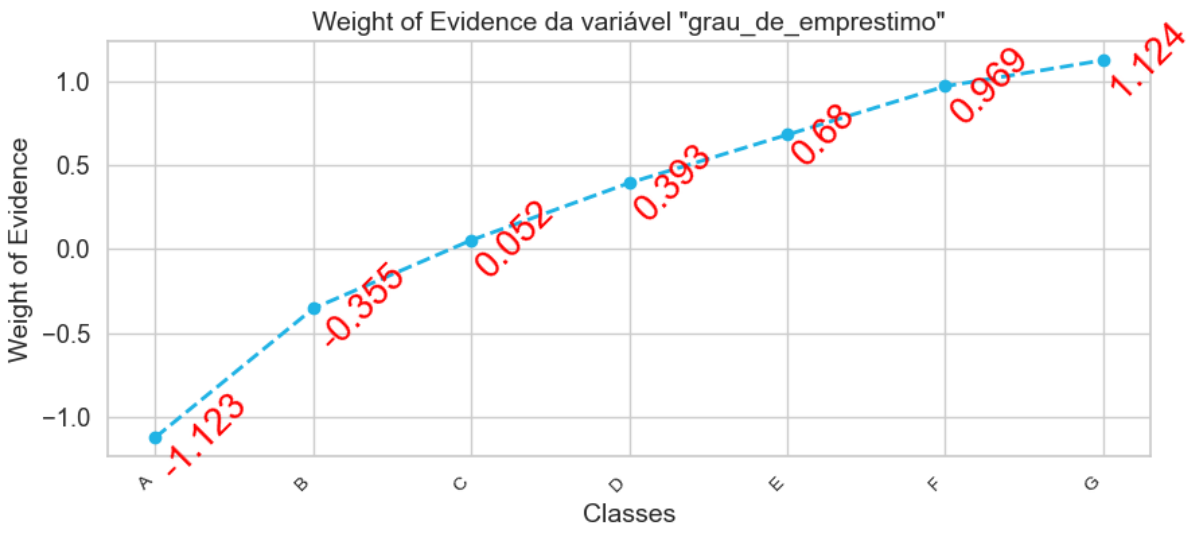
A primeira variável exemplo denomina-se *Grau do Empréstimo*. Embora a competição não traga exatamente o que significa cada uma das classes, sabe-se que ela representa um ranking e corresponde ao nível de empréstimo solicitado pelo cliente, de forma que a classe "A" seria empréstimos mais robustos e "G" empréstimos mais comuns. Como exposto nas figuras 18 e 19, através da análise de WOE nota-se que as classes "A" e "B", por assumirem valores negativos, associam-se com a classe 0, logo, estão majoritariamente relacionadas com bons pagadores, ao passo que as demais assumem valores positivos e relacionam-se de forma mais expressiva com maus pagadores.

Figura 18 – Distribuição de Bons e Maus Pagadores vs Grau de Empréstimo



Fonte: Autoria Própria.

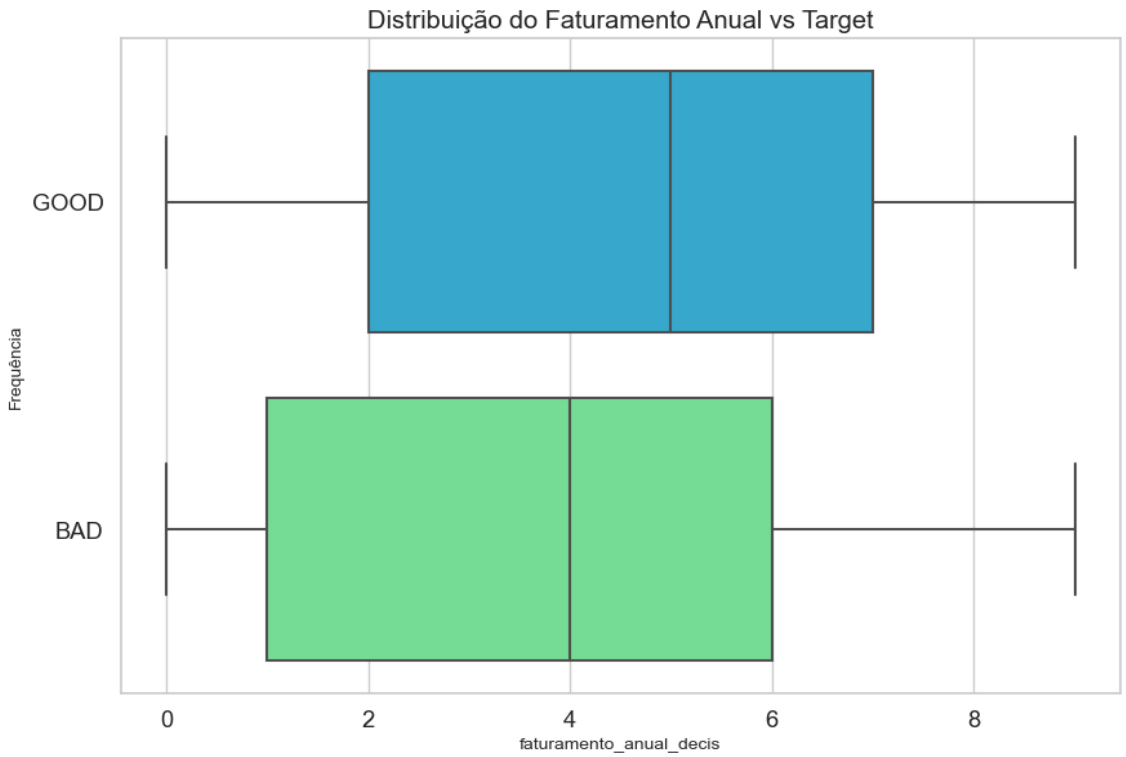
Figura 19 – Weight of Evidence da Variável Grau do Empréstimo



Fonte: Autoria Própria.

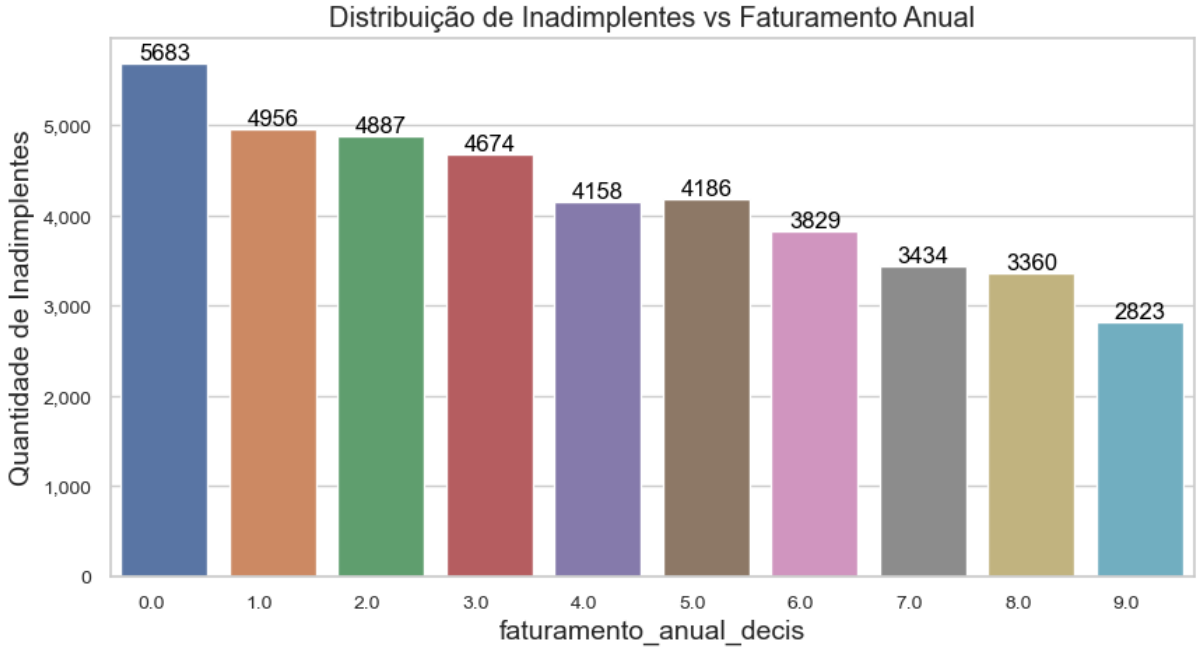
O segundo exemplo consiste na variável *Faturamento Anual*, a qual representa o rendimento anual do cliente. Por tratar-se de uma variável contínua, necessita-se de outra abordagem para a realização da análise. Dada a situação, optou-se por transformar a variável em decis e analisar a quantidade de inadimplentes em cada faixa. Por construção, a criação de decis garante que cada faixa possui a mesma quantidade de pessoas, logo, ao cruzarmos essas faixas com a quantidade de inadimplentes, pode-se entender se o módulo da variável possui relação com o evento de interesse. No caso do *Faturamento Anual*, quanto maior o decil, maior é seu faturamento, ou seja, mais dinheiro o cliente recebe. Por ter mais dinheiro, a tendência é que este cliente possua maior oportunidade de arcar com seus compromissos financeiros e, portanto, pagar seu empréstimo. Pelas figuras 20 e 21 nota-se que esse fenômeno de fato acontece na prática, pois conforme o decil aumenta, a quantidade de inadimplentes diminui.

Figura 20 – Boxplot da Variável Faturamento Anual



Fonte: Autoria Própria.

Figura 21 – Distribuição de Decis da Variável Faturamento Anual vs Maus Pagadores



Fonte: Autoria Própria.

Conforme mencionado anteriormente, objetiva-se conquistar a convergência entre as análises conceituais e quantitativas. Nos dois exemplos houve o acordo entre os resultados, contudo, é comum que algumas variáveis não sigam exatamente a mesma linha de raciocínio e, nestes casos, deve-se sempre dar peso maior para a segunda abordagem.

3.8 Construção da Política de Crédito

De acordo com a Serasa, a política de crédito é um documento que indica regras e critérios responsáveis por direcionar a empresa durante a tomada de decisão em uma concessão, portanto, sua criação representa uma etapa fundamental durante uma análise de risco. Para este trabalho, optou-se pela formulação de uma política tradicional novamente baseada em regras conceituais e quantitativas já consolidadas na literatura. O intuito é que ela seja um *baseline* capaz de ser comparada aos modelos de Machine Learning propostos, destacando ganhos ou perdas de cada uma das abordagens.

Em metodologias de modelagem matemática, grande parte dos processos ocorrem de forma automática, logo, a seleção de variáveis pode ser menos rígida visto que os cálculos geralmente são realizados por softwares. Pelo fato de tratar-se de uma abordagem menos sofisticada e mais manual quando comparada a esses modelos, a criação de uma política demanda uma seleção de variáveis mais criteriosa, sendo assim, objetiva-se construir uma política baseada no menor número de variáveis ao mesmo tempo que mantém-se o poder de discriminação. Para esta etapa, utilizou-se o *Information Value* (IV), o conceitual dos *5C's do Crédito* e a análise de distribuição de probabilidade.

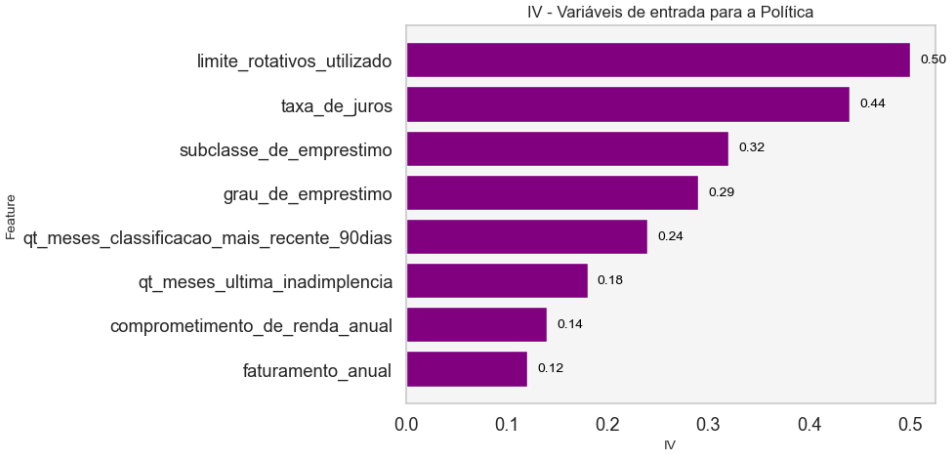
Conforme abordado por Sicsú (2010), a estatística IV permite avaliar o potencial discriminador de uma variável. Dada uma variável categórica aleatória, pode-se atribuir a cada categoria um WOE, logo, ao combinar a diferença das probabilidades condicionais das categorias para o evento positivo e negativo com esses WOE's, consegue-se obter o IV da variável. O IV é definido pela seguinte equação

$$IV = \sum \big(P(c|bom) - P(c|mau)\big) x \ln\left[\frac{P(c|bom)}{P(c|mau)}\right]$$

e, quanto maior o IV, mais informativa é a característica. Como exposto na figura 22, calculou-se o IV de cada uma das variáveis, entretanto, escolheu-se levar adiante apenas as

que tiveram valores superiores a 0.1 pois, segundo Sicsú (2010), resultados superiores a 0.1 sugerem que a informação é realmente útil.

Figura 22 – Information Value (IV)



Fonte: Autoria Própria.

Posteriormente, segmentou-se cada uma das informações através da análise dos *5C's de Crédito* a fim de compreender se estas encaixavam-se nos conceitos propostos. Pela tabela 1, todas as variáveis encaixaram-se em algum dos segmentos, contudo, para a parte de *Capital*, não se encontrou nenhuma informação presente na base de dados, pois esta característica diz respeito às finanças da instituição como patrimônio, balanço e demais projeções financeiras.

Tabela 1 – Segmentação das variáveis através dos 5C's do Crédito

5C's do Crédito	***Variáveis***
Caráter	*qt_meses_classificacao_mais_recente_90dias* *qt_meses_ultima_inadimplencia*
Capacidade	*limite_rotativos_utilizado* *comprometimento_de_renda_anual* *faturamento_anual*
Colateral	*grau_de_emprestimo* *subclasse_de_emprestimo*
Condições	*taxa_de_juros*
Capital	-

Fonte: Autoria Própria.

Finalmente, realizou-se a análise de distribuição de probabilidade para cada uma das variáveis. Como mencionado anteriormente, essa análise serve para comprovar a existência de uma relação bem comportada entre a característica e a variável resposta. Para variáveis contínuas, decidiu-se aplicar a transformação em decil, ao passo que as categóricas não necessitam de transformações, como mostrado nas tabelas 2, 3, 4 e 5.

Tabela 2 – Distribuição de Probabilidade para Faturamento Anual, Comprometimento de Renda e Taxa de Juros

Decil	*% BAD Faturamento Anual*	*% BAD Comprometimento de Renda*	*%BAD Taxa de Juros*
0	*14.6%*	*8.1%*	*3.5%*
1	*13.8%*	*8.3%*	*5.3%*
2	*13%*	*8.8%*	*7.5%*
3	*12%*	*9.4%*	*8.2%*
4	*11.8%*	*9.8%*	*10%*
5	*11%*	*10.2%*	*11.3%*
6	*10.3%*	*11.4%*	*12.7%*
7	*9.3%*	*13.2%*	*14.9%*
8	*8.6%*	*14.9%*	*18.2%*
9	*7.9%*	*18.4%*	*22.2%*

Fonte: Autoria Própria.

Tabela 3 – Distribuição de Probabilidade para Limite de Rotativos Utilizado, Qt Meses desde a última inadimplência e Qt Meses Classificação mais recente em 90d

Decil	*% BAD Limite de Rotativos Utilizado*	*% BAD Qt Meses desde última inadimp.*	*%BAD Qt Meses desde classif. mais recente 90d*
0	*11.8%*	*11.7%*	*11.5%*
1	*11.4%*	*10.8%*	*11%*
2	*11.8%*	*10.8%*	*10.6%*
3	*11.6%*	*10.5%*	*10.2%*

4	*11.8%*	*10.4%*	*10.6%*
5	*11.4%*	*10.3%*	*9.7%*
6	*11.7%*	*10.5%*	*9.9%*
7	*11.2%*	*10.7%*	*9.5%*
8	*10.4%*	*10.6%*	*10.7%*
9	*9.4%*	*11.3%*	*11.2%*

Fonte: Autoria Própria.

Tabela 4 – Distribuição de Probabilidade para Classe do Produto

Classe de Produto	*% BAD Classe de Produto*
A	*4%*
B	*8.2%*
C	*11.8%*
D	*15.8%*
E	*20%*
F	*25.1%*
G	*28.1%*

Fonte: Autoria Própria.

Tabela 5 – Distribuição de Probabilidade para Subclasse do Produto

Subclasse do Produto	***%BAD Subclasse do Produto***
A1	*1.9%*
A2	*2.9%*
A3	*3.5%*
A4	*4.6%*
A5	*5.2%*

B1	*6.2%*
B2	*7.2%*
B3	*8.1%*
B4	*9.1%*
B5	*10%*
C1	*10.7%*
C2	*11%*
C3	*11.9%*
C4	*12.3%*
C5	*13.5%*
D1	*14.7%*
D2	*15.4%*
D3	*15.4%*
D4	*16.8%*
D5	*17.5%*
E1	*18.4%*
E2	*19.4%*
E3	*20%*
E4	*21.3%*
E5	*22.3%*
F1	*22.3%*
F2	*24.7%*
F3	*25.7%*
F4	*26%*
F5	*26.4%*
G1	*26.8%*

G2	*28.6%*
G3	*29.3%*
G4	*29.5%*
G5	*30%*

Fonte: Autoria Própria.

Em risco de crédito, costuma-se buscar relações que possuam ordenação em relação a PD, portanto, manteve-se apenas as informações as quais atendiam esse requisito, sendo elas o *faturamento anual*, *comprometimento de renda, taxa de juros, classe do produto e subclasse do produto.* Como a política tem o intuito de ser uma metodologia para concessão de crédito, ela deve ser capaz de separar os clientes bons dos maus. Para isso, através da combinação das cinco variáveis mencionadas, criou-se uma regra para definição da PD do cliente. A equação da regra é dada por

$$PD_{Política} = PD_{Faturamento\ Anual} + PD_{Compromet.\ Renda} + PD_{Taxa\ de\ Juros} + PD_{Classe\ Produto} + PD_{Subclasse\ Produto}$$

sendo cada PD_n a probabilidade de inadimplência de cada característica na qual o cliente se encontra.

Para exemplificar, pode-se pensar num cliente de faturamento anual limitado, alto comprometimento de renda, submetido a taxas de juros elevadas e contratando um produto de classe e subclasse menos expressivos. Pelos dados mostrados anteriormente, percebe-se que esse cliente se encontra em grupos de alta inadimplência, logo, aproxima-se de um perfil mais arriscado. Em outro cenário, um cliente com alto faturamento anual, baixo comprometimento de renda, taxa de juros menor e que esteja contratando um produto de maior prestígio aproxima-se de um cliente menos arriscado. Dessa forma, a regra desenvolvida visa justamente encontrar o perfil de risco ideal de cada cliente e determinar se a concessão de crédito pode acontecer ou não.

3.9 Construção dos Modelos

A construção dos modelos de Machine Learning representa o núcleo de uma análise de risco de crédito direcionada por modelagem matemática. Nesta etapa, explorou-se diversas

técnicas robustas de modelagem a fim de obter-se a melhor configuração possível para o modelo. Tal configuração é responsável pela maximização do poder preditivo e capacidade de discriminação, ou seja, ela auxilia na conquista do melhor resultado possível para o problema proposto.

Sabe-se que, embora os modelos de Aprendizado de Máquina sejam criados com base em técnicas matemáticas e estatísticas, sua implementação final se dá na forma de software. Isso implica que todos os processos subjacentes ao aprendizado, desde a aquisição e o pré-processamento dos dados até a construção e o treinamento do modelo, devem ser traduzidos para linguagem de máquina para que o computador possa executar essas tarefas de forma precisa e eficaz. Infelizmente, a realidade dos dados do mundo real é frequentemente complexa e desafiadora. Erros humanos, falhas operacionais e outras fontes de ruído podem introduzir uma alta frequência de dados que não se conformam às expectativas ou são divergentes do comportamento ideal. Nesse contexto, a qualidade dos dados é frequentemente discutível, o que pode prejudicar a eficácia do modelo de Aprendizado de Máquina. Sendo assim, entra em cena o pré-processamento de dados, uma etapa crítica na preparação dos dados para a modelagem.

O principal objetivo do pré-processamento é melhorar a qualidade e a utilidade dos dados, tornando-os mais adequados para o treinamento de modelos de Aprendizado de Máquina. Isso envolve uma série de técnicas e transformações aplicadas aos dados brutos, como limpeza para corrigir erros, tratamento de valores ausentes, normalização e seleção de características relevantes. Uma base de dados de alta qualidade e bem tratada é fundamental para garantir que o modelo aprenda com precisão os padrões presentes nos dados e, por sua vez, faça previsões precisas em novos exemplos. Dessa forma, abordou-se três técnicas de pré-processamento: *Target Encoder, Min-Max Scaler e Simple Imputer*.

Encoding é uma técnica muito útil quando se lida com variáveis categóricas. Sua função é a aplicação de um processo de discretização a fim de transformar dados categóricos em dados discretos ou, em outras palavras, transformar classes em números. Nesse âmbito, uma técnica robusta para essa tarefa é o *Target Encoder*. Ele é um *encoder* o qual transforma variáveis categóricas em variáveis discretas ou contínuas de forma inteligente. Ao invés de criar uma coluna para cada categoria, ele mantém a coluna de variável categórica de forma a substituir cada categoria por um valor numérico específico. Neste caso, geralmente utiliza-se alguma variável contínua agrupada pela categórica. Esta abordagem é interessante pois ela

reduz a dimensionalidade de forma eficiente, possibilitando a criação de uma base de dados menos esparsa, facilitando os cálculos matemáticos do modelo e simplificando o poder computacional demandado.

Neste projeto, aplicou-se o *Target Encoder* em todas as variáveis categóricas com o auxílio da PD calculada para cada classe. Exemplificou-se o processo através da tabela 6 simulando a transformação de uma variável categórica denominada como "Quantidade de Anos no Mesmo Emprego", na qual transformou-se quatro categorias em quatro valores, criando uma representação compacta e eficaz das informações categóricas para uso em modelos de Aprendizado de Máquina.

Tabela 6 – Aplicação do Target Encoder

Quantidade de Anos no Mesmo Emprego	***% BAD Quantidade de Anos no Mesmo Emprego***
Até 3 Anos	*11.76%*
Até 6 Anos	*11.6%*
Até 9 Anos	*11.54%*
10 anos ou +	*10.39%*

Fonte: Autoria Própria.

Em diversas bases de dados é comum encontrar variáveis com valores ausentes. Ignorar essas informações e excluí-las da análise não é uma boa prática visto que resultam na perda de dados importantes, piorando a performance do modelo. A imputação de valores faltantes tornou-se essencial para modelos de Aprendizado de Máquina justamente para permitir que essas informações sejam levadas em consideração pelo algoritmo. Embora fundamental, a imputação deve ser realizada da forma correta.

Nesse âmbito, uma das técnicas mais consolidadas na literatura denomina-se *Simple Imputer*. O *Simple Imputer* é um método de preenchimento de dados faltantes o qual permite de forma automática que o elemento faltante seja substituído pela média/mediana (em caso de variáveis quantitativas) ou moda (em caso de variáveis qualitativas). Essa abordagem é interessante pois traz facilidade na hora de tratar inúmeras variáveis de maneira automática. Como realizou-se previamente a transformação das variáveis categóricas em discretas, optou-se por preencher os valores faltantes com a mediana de cada variável.

Durante a etapa de treinamento, determinados modelos entendem a importância das variáveis de forma diferente devido a escalas de magnitude distintas. Variáveis em unidades

maiores tendem a influenciarem majoritariamente o modelo, criando assim um algoritmo incompatível com a verdadeira situação. Técnicas de escalonamento são utilizadas em Aprendizado de Máquina para ajustar as escalas das variáveis e, assim, garantir que todas as variáveis tenham o peso ideal no processo de treinamento do modelo. Dessa forma, pode-se comparar variáveis de forma justa e significativa, inserindo confiabilidade no resultado do modelo.

Uma renomada metodologia de escalonamento é o *Min-Max Scaler*. Essa técnica é popularmente conhecida por normalização e redimensiona as variáveis de um conjunto de dados para um intervalo específico contido em [0,1]. Sua notoriedade deriva da manutenção da estrutura original dos dados mesmo após o redimensionamento e da preservação de outliers, garantindo o caráter informativo. Sua equação pode ser definida por:

$$f(x) = \frac{x - \min(x)}{\max(x) - \min(x)}$$

sendo x a variável. Para todas as variáveis presentes na base de dados, aplicou-se o *Min-Max-Scaler* a fim de garantir que todas as informações estivessem na mesma escala e, portanto, o mesmo peso durante o aprendizado dos modelos.

Engenharia de atributos é o processo de criação de novas variáveis a partir das variáveis iniciais de uma base de dados, bem como a seleção das melhores. Seu objetivo é melhorar a qualidade dos dados e fornecer informações as quais sejam ainda mais relevantes que os dados brutos. Esta abordagem permite que o modelo entenda melhor os padrões e ajuste-se melhor ao desafio proposto, ao mesmo tempo que reduz a dimensionalidade dos dados e a complexidade do modelo, diminuindo a demanda por poder computacional. Dessa forma, para a engenharia de atributos aplicou-se três técnicas consolidadas, sendo elas o *Variance Threshold*, *Mutual Information* e *Feature Importance*.

Bruce e Bruce (2019) afirmam que a variância de uma amostra é uma medida de variabilidade a qual indica a distância entre os valores da média aritmética ou, em outras palavras, mostra o quão dispersos estão os dados. Quanto maior a variância, mais variados estão os dados. Modelos de Aprendizado de Máquina beneficiam-se de variáveis com variâncias significativas, pois variáveis constantes simplesmente não agregam informações relevantes para o modelo. Um método muito famoso para esta tarefa é o *Variance Threshold*, o qual elimina variáveis abaixo de um limiar pré-definido de variância. A ideia é que recursos

com baixa variação. Dessa forma, além de reduzir a dimensionalidade da base de dados, o modelo será alimentado apenas com variáveis com possibilidades reais de agregarem positivamente.

Para a análise variáveis categóricas, um dos métodos mais robustos é o *Mutual Information*. Esta técnica é uma medida estatística que quantifica a dependência entre duas variáveis aleatórias, sendo um excelente método para entender se a variável resposta possui dependência com a variável de entrada. Basicamente, calcula-se a probabilidade de duas classes ocorrem juntas e a probabilidade de ocorrem separadas, as quais posteriormente são utilizadas para calcular a entropia e finalmente entender o quanto a variável de entrada fornece de informação para predizer a variável resposta. Valores mais altos significam maior dependência entre a variável de entrada e a variável resposta e sua equação pode ser definida como

$$Mutual\ Information = \sum\sum p_{(x,y)} log\left[\frac{p_{(x,y)}}{p(x)p(y)}\right]$$

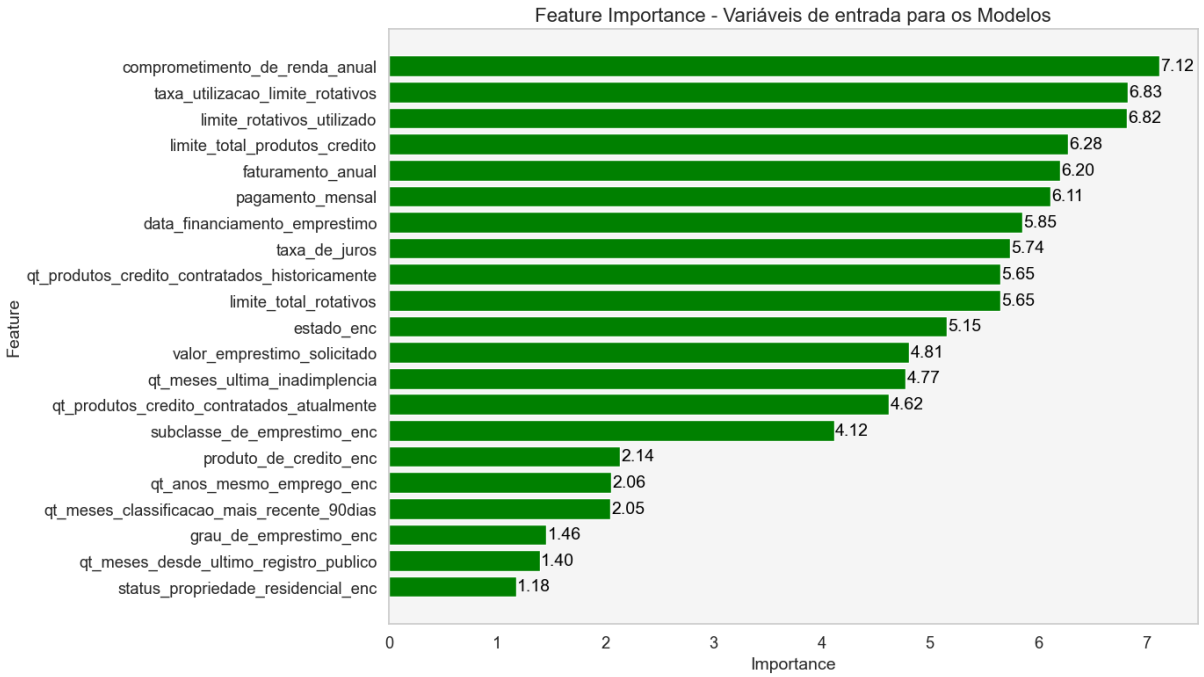
sendo $p_{(x,y)}$ a probabilidade de ocorrem juntas, p(x) a probabilidade de ocorrer o evento x e p(y) a probabilidade de ocorrer o evento y.

Modelos de Árvore possuem um processo embutido muito interessante denominado Feature Importance. Essa técnica é uma medida a qual avalia o grau de contribuição de cada variável de entrada no desempenho de um modelo de Aprendizado de Máquina. Ela fornece um ranking o qual indica a relevância de cada característica em relação à variável de resposta do modelo. A pontuação do ranking geralmente é definida pela função de custo associada ao modelo, sendo as mais famosas o Índice Gini e a Entropia, abordados anteriormente. Resumidamente, o modelo calcula um desses indicadores e entende a proporção de importância daquela variável frente as demais. Esse método é interessante pois pode-se escolher as *N* melhores variáveis do modelo ou, em necessidades mais abrangentes, escolher àquelas as quais possuam importância maior que um limiar pré-definido. Além de robusta, o Feature Importance possui fácil interpretação, portanto, é extremamente disseminado em casos em que a explicação da decisão do modelo é questionada.

Estas técnicas são fundamentais para redução de dimensionalidade ao passo que removem informações irrelevantes e permitem que o modelo foque apenas em características importantes. Decidiu-se eliminar variáveis com variância e informação mútua igual a zero,

assim como aquelas com *feature importance* menor que um. A escolha desses valores fundamenta-se na carência de variabilidade para aquelas com variância igual a zero, as com informação mútua igual a zero não oferecerem nenhum ganho de informação, e *feature importance* menor que 1 significa que tal característica praticamente não auxilia o modelo classificar a amostra corretamente. A figura 23 traz as variáveis de entrada para os modelos de Machine Learning:

Figura 23 – Variáveis de Entrada para os Modelos de Machine Learning



Fonte: Autoria Própria.

Posteriormente, para a criação do motor de modelos, testou-se cinco métodos, sendo eles a Regressão Logística, o Naive Bayes, o KNN Classifier, a Random Forest e o XGBoost. Estes correspondem, respectivamente, aos modelos lineares de regressão, modelos bayesianos, modelos baseados em distância, modelos de baggin e modelos de boosting. A diversidade de modelos torna-se estratégica uma vez que cada um possui particularidades, logo, oferecem vantagens e desvantagens específicas e interessantes de serem discutidas.

A Regressão Logística é o método mais famoso em análises de risco de crédito. Sua notoriedade ocorre por ser um método paramétrico, facilitando a compreensão da contribuição de cada variável de entrada em relação a variável resposta. Esta característica é muito bem vista, pois ela permite que pessoas não técnicas compreendam o motivo da aprovação ou negação do crédito de forma mais fácil.

O Naive Bayes carrega decisões baseadas em probabilidades condicionais, algo também importante no cenário de crédito. Durante a concessão, muitos eventos possuem relação e complementam-se, logo, abordagens bayesianas agregam informação e conseguem até mesmo explicar algumas decisões.

O KNN é capaz de classificar uma nova instância com base em amostras semelhanças, portanto, em situações em que existem muitos perfis de clientes distintos, esta abordagem ganha bastante força. Além disso, por tratar-se de um modelo baseado em distâncias, o KNN ainda consegue lidar com relações complexas entre as variáveis e captar efeitos interessantes.

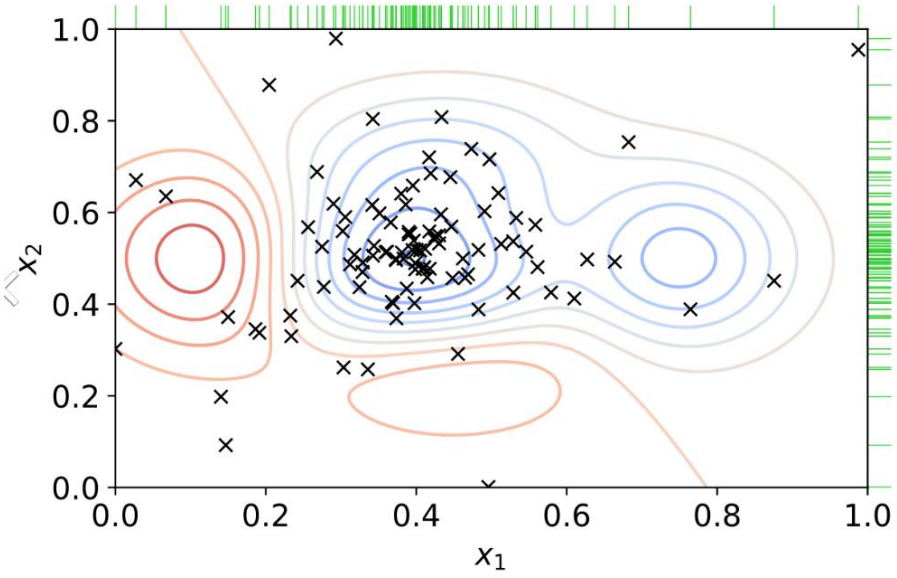
A Random Forest é um método de combinação de diversas árvores para criar um classificador final composto do voto majoritário dentre todas. Por ser um método de *Bagging*, a Random Forest seleciona X amostras com N colunas para criar suas predições, logo, essa abordagem é muito robusta e estável, além de ser capaz de também lidar com relações complexas entre as variáveis. Um de seus diferenciais é, por ser um modelo de árvore, conceitualmente o processo de decisão é criado através de regras, assemelhando-se com uma política de crédito. Essa característica permite alta explicabilidade, logo, pessoas não técnicas conseguem entender facilmente o processo de decisão.

Por sua vez, o XGBoost é também um método de combinação de árvores, todavia, o classificador final é composto de apenas uma árvore extremamente bem treinada. Por tratar-se de um método de *Boosting*, essa abordagem visa a criação de árvores sequenciais nas quais permitem a propagação e a correção de erros, resultando em uma árvore final muito robusta. Além de todos os benefícios de modelos de árvore anteriormente citados, o grande diferencial do XGBoost é sua capacidade de predição, a qual é consideravelmente forte dada a propagação e correção de erros.

Geralmente, o primeiro modelo treinado não possui a melhor performance possível. Isso ocorre pois costuma-se decidir o melhor modelo dentre vários testados e, então, otimizá-lo para alcançar resultados ainda melhores. Para alcança-los, há diversos métodos, sendo o principal conhecido como otimização de hiperparâmetros. Hiperparâmetros são configurações definidas antes do treinamento de um modelo. Eles representam características construtivas, como o número de vizinhos mais próximos para um KNN, o número de profundidade de uma Árvore de Decisão ou o número de neurônios de uma Rede Neural, por exemplo. A otimização de hiperparâmetros visa achar a melhor configuração possível a fim de aperfeiçoar o modelo, conquistando os melhores resultados possíveis.

O *Bayes Search* é uma técnica de otimização de hiperparâmetros cuja utiliza abordagem bayesiana para encontrar a melhor combinação possível. Resumidamente, sua função é construir um modelo probabilístico que relaciona os hiperparâmetros do modelo com a métrica de avaliação escolhida. Dado um conjunto inicial de hiperparâmetros, testa-se de maneira iterativa a combinação dos valores desse conjunto e avalia-se a métrica de avaliação para cada combinação. O algoritmo possui caráter adaptativo, portanto, a atualização ocorre apenas para hiperparâmetros inseridos nos espaços de busca mais promissoras. A combinação de hiperparâmetros escolhida será aquela à qual apresente a maior probabilidade de retornar as melhores métricas de avaliação ou, em outras palavras, a que proporciona ao modelo a melhor performance dentre todas as combinações testadas. Além de eficiente, essa técnica pauta-se em teorias consolidadas e de fácil interpretação, portanto, empregou-se o *Bayes Search* como método de otimização de hiperparâmetros. O processo descrito pode ser visualizado através da figura 24.

Figura 24 – Bayes Search



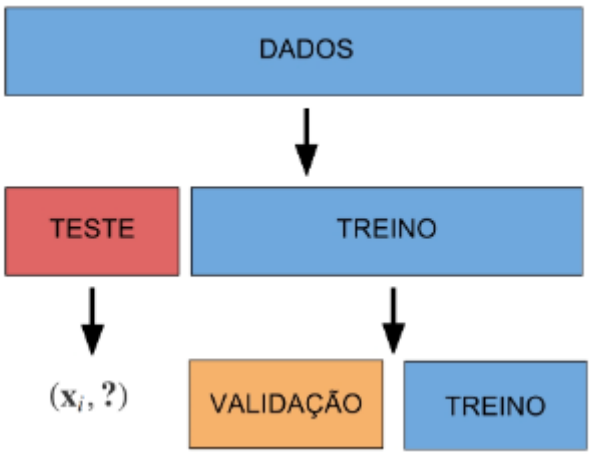
Fonte: Hyperparameter [...] (2024)

4 RESULTADOS

Ao desenvolver-se modelos de Aprendizado de Máquina, deve-se garantir que eles possuam boa capacidade de generalização para dados não vistos. Nesse âmbito, para simular seu desempenho em dados novos, há duas técnicas fundamentais: Holdout e Cross Validation.

A figura 25 mostra a técnica de Holdout, o qual é um método que consiste em separar, de forma aleatória, uma parcela dos dados para treinamento e o restante para teste. Os dados de treinamento servem para que o modelo monte uma equação matemática capaz de ajustar-se aos dados e entender os padrões contidos na amostra, ao passo que os dados de teste simulam dados novos aos quais serão submetidos ao modelo. Como baseia-se na aleatoriedade, esta técnica está sujeita ao viés de seleção de amostra, logo, o modelo pode capturar padrões existentes apenas na amostra treino e performar de maneira não satisfatória nos dados de teste, resultando em uma alta variância.

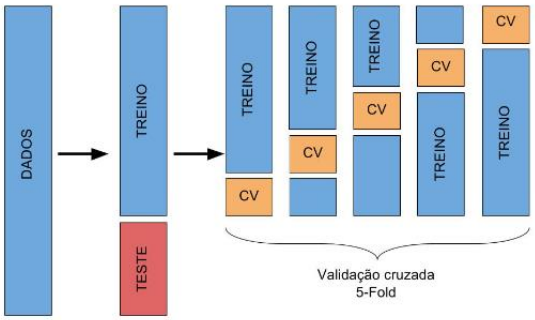
Figura 25 – Holdout



Fonte: Universidade Federal do Paraná ([ca. 2018])

No Cross Validação, a amostra é dividida aleatoriamente e separada em K grupos. De forma repetitiva, o mesmo modelo é treinado *k* vezes, sendo *K-1* grupos utilizados para treino e um grupo para teste. Iterativamente, a amostra de teste muda e o resultado final é uma média aritmética das métricas de todos os treinamentos, portanto, ao final do processo, o modelo consegue reduzir bastante a variância e fornecer um resultado mais robusto e fidedigno da performance do modelo, como exposto pela figura 26.

Figura 26 – Cross Validation



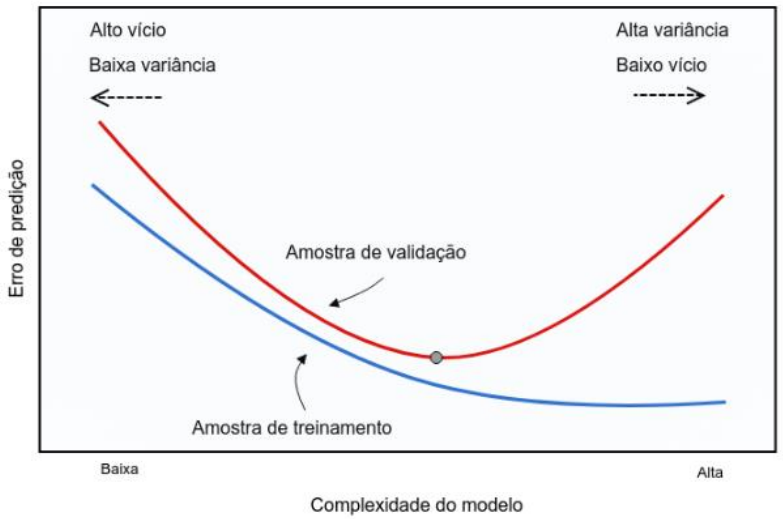
Fonte: Universidade Federal do Paraná ([ca. 2018])

Após a obtenção das métricas de treino e teste, pode-se compará-las a fim de analisar se o modelo possui generalização. Conforme Géron (2019), um modelo possui boa generalização quando as métricas de treino e teste são satisfatórias e possuem valores relativamente próximos. Embora o conceito seja simples, nem sempre os modelos alcançam bons resultados e há disparidade entre os cenários de treino e teste.

Denomina-se Overfitting o cenário o qual o modelo possui bons resultados em dados de treino, mas performa abaixo do esperado em dados de teste. Em suma, isso significa que o modelo foi capaz de criar uma equação matemática tão complexa ao ponto de assimilar até mesmo os ruídos da amostra de treino.

O cenário de Underfitting é o inverso, logo, o modelo possui maus resultados tanto em dados de treino quanto de teste. Em suma, o modelo não conseguiu de criar uma equação calibrada a qual fosse representativa e capaz de captar os padrões na amostra de treino. Em casos de Overfitting, Géron (2019) afirma que a variância é alta, visto que esse algoritmo está extremamente sujeito aos dados de entrada; em casos de Underfitting, o viés é alto, pois o modelo não consegue detectar o padrão de nossos dados. Existem diversas técnicas para tratar ambos os casos, entretanto, espera-se que o modelo tenha uma troca justa entre ambos e consiga performar bem de forma generalizada, como demonstrado pela figura 27.

Figura 27 – Trade-Off Viés x Variância



Fonte: Universidade Federal do Paraná ([ca. 2018])

4.1 Resultados da Política de Crédito

A partir da tabela 7, encontram-se os resultados da política de crédito:

Tabela 7 – Resultados da Política de Crédito

Acuracia	***Precision***	***Recall***	***F1-Score***	***AUC***	***KS***	***Etapa***
0.72	*0.19*	*0.44*	*0.26*	*0.66*	*0.23*	*Treino*
0.72	*0.19*	*0.44*	*0.26*	*0.67*	*0.24*	*Teste*

Fonte: Autoria Própria.

Aparentemente, pelos resultados acima, nota-se que a política de crédito é promissora. Ela consegue separar bem a classe positiva da negativa e possui boa assertividade do verdadeiro mau pagador. Tais valores mostram que essa política possui regras robustas, portanto, percebe-se a capacidade de captação dos principais perfis de risco do portfólio.

Mesmo com resultados promissores, nota-se que ainda há oportunidade de melhorias. Como a decisão é fruto de regras determinísticas, essa abordagem penaliza excessivamente clientes de perfil de risco mais alto e favorece clientes de perfil de risco mais baixo. Embora faça sentido, existem clientes que fogem a regra, portanto, todos que seguem esse comportamento provavelmente serão classificados incorretamente.

Como forma de mitigar, pode-se criar políticas auxiliares para tratar casos específicos, como por exemplo uma política específica para clientes que estão exclusivamente atrás de determinado produto de crédito ou buscam a contratação de determinada classe de empréstimo. A combinação de diversas políticas teria potencial de aumentar a performance, todavia, a grande quantidade de regras dificultaria o entendimento de muitas aprovações e reprovações, comprometendo a transparência dos resultados.

A elaboração de políticas específicas para cada público cria um labirinto de requisitos complexos de serem decifrados. Em casos pontuais nos quais requerem análises, os analistas gastariam horas criando regras reversas, aumentando a propensão a erros e transformando um processo destinado a facilitar a concessão em um obstáculo para a compreensão das decisões.

4.2 Resultados dos Modelos de Machine Learning

Os resultados do motor de modelos nas visões de treino, teste e validação cruzada são expostos pela tabela 8. Nota-se que a Regressão Logística, a Random Forest e o XGBoost tiveram bom desempenho, com destaque para o último, o qual foi o vencedor. Em relação ao Naive Bayes e o KNN, ambos sofreram perdas consideráveis de resultado, logo, são modelos inválidos e não deveriam ser levados adiante.

Tabela 8 – Resultado do Motor de Modelos

Acuracia	***Precision***	***Recall***	***F1-Score***	***AUC***	***KS***	***Etapa***	***Classificador***
0.74	*0.20*	*0.43*	*0.27*	*0.67*	*0.25*	*Treino*	
0.74	*0.20*	*0.43*	*0.27*	*0.68*	*0.25*	*Teste*	***Reg. Logística***
0.74	*0.20*	*0.43*	*0.27*	*0.67*	*0.25*	*V Cruz*	
0.84	*0.24*	*0.18*	*0.21*	*0.66*	*0.24*	*Treino*	
0.84	*0.24*	*0.18*	*0.20*	*0.66*	*0.25*	*Teste*	***Naive Bayes***
0.84	*0.24*	*0.18*	*0.21*	*0.66*	*0.24*	*V Cruz*	
0.90	*0.69*	*0.24*	*0.36*	*0.91*	*0.79*	*Treino*	
0.86	*0.15*	*0.05*	*0.08*	*0.52*	*0.04*	*Teste*	***KNN***
0.86	*0.14*	*0.05*	*0.07*	*0.52*	*0.04*	*V Cruz*	
0.74	*0.21*	*0.49*	*0.30*	*0.70*	*0.29*	*Treino*	
0.74	*0.21*	*0.49*	*0.30*	*0.70*	*0.29*	*Teste*	***Random Forest***
0.74	*0.21*	*0.49*	*0.30*	*0.70*	*0.29*	*V Cruz*	

0.74	*0.22*	*0.51*	*0.31*	*0.71*	*0.30*	*Treino*	
0.74	*0.22*	*0.51*	*0.31*	*0.71*	*0.30*	*Teste*	***XGBoost***
0.74	*0.22*	*0.50*	*0.30*	*0.71*	*0.30*	*V Cruz*	

Fonte: Autoria Própria.

Em relação a Regressão Logística, por ser um modelo linear, nota-se que ele não conseguiu captar toda a complexidade dos dados. Mesmo assim, a conquista de resultados promissores denota o quão útil pode ser um modelo simples.

Quanto ao Naive Bayes, embora o cálculo das probabilidades condicionais seja interessante, uma das premissas deste algoritmo é a independência dos eventos, portanto, algumas relações entre variáveis ficam mascaradas e impedem que o modelo entenda toda a informação dos dados disponíveis. Essa situação fica clara nos resultados acima, pois o recall inferior em relação aos concorrentes mostra que esse modelo acerta pouco o verdadeiro mau pagador.

Percebe-se que o KNN foi o pior modelo, pois ele sofre de Overfitting. Isso significa que durante a etapa de treinamento, o algoritmo ajustou-se tanto à amostra de treino que, ao ser apresentado a amostras novas, ele é incapaz de realizar uma classificação justa.

A Random Forest teve excelentes resultados, provando o quão estável e forte é essa abordagem. Além de separar muito bem a classe positiva e a negativa dado os valores de AUC e KS, pelos valores de recall este modelo consegue acertar consideravelmente bem o verdadeiro mau pagador.

O XGBoost, pelos resultados acima, provou ser o melhor modelo dentre os testados. Superando os demais em todos os cenários, este algoritmo consolida-se como o mais robusto e eficaz, oferecendo o melhor desempenho possível. Dessa forma, optou-se por leva-lo adiante durante o processo de otimização.

Finalmente, adotou-se o algoritmo Bayes Search como otimizador do modelo vencedor entre os citados. O intuito desta etapa é melhorar ainda mais o desempenho do modelo através da melhor configuração possível de hiperparâmetros. A principal vantagem dessa metodologia é a sua eficácia, visto que ele explora apenas regiões promissoras durante sua busca em espaço, garantindo bons resultados em pouco tempo. Os resultados do modelo XGBoost otimizado encontram-se na tabela 9:

Tabela 9 – Resultado do XGBoost otimizado com Bayes Search

Acuracia	***Precision***	***Recall***	***FF1-Score***	***AUC***	***KS***	***Etapa***	***Classificador***
0.69	*0.21*	*0.64*	*0.32*	*0.74*	*0.34*	*Treino*	*Bayes Search + XGBoost*
0.69	*0.20*	*0.61*	*0.31*	*0.72*	*0.32*	*Teste*	*Bayes Search + XGBoost*
0.69	*0.20*	*0.60*	*0.30*	*0.71*	*0.30*	*V Cruz*	*Bayes Search + XGBoost*

Fonte: Autoria Própria.

Pela a tabela, embora tenha sofrido pequena perda de precision, observa-se melhora expressiva de recall, AUC e KS. Como as métricas de maior importância melhoraram, assim como os valores de treino, teste e validação cruzada são próximas, pode-se considerá-lo um modelo totalmente estável e capaz de realizar a separação entre bons e maus pagadores.

Diferentemente das políticas de crédito, modelos de Machine Learning absorvem toda a complexidade do problema de modo a combinar todas as variáveis de maneira automática. Essa combinação, muitas vezes não intuitivas para seres humanos, permite que o método quantitativo reconheça padrões que geralmente são perdidos em soluções mais simples. Além disso, como o modelo ganhador foi um método Boosting baseado em Árvore de Decisão, a interpretabilidade da árvore torna-se mais fácil devido a ferramentas como Feature Importance.

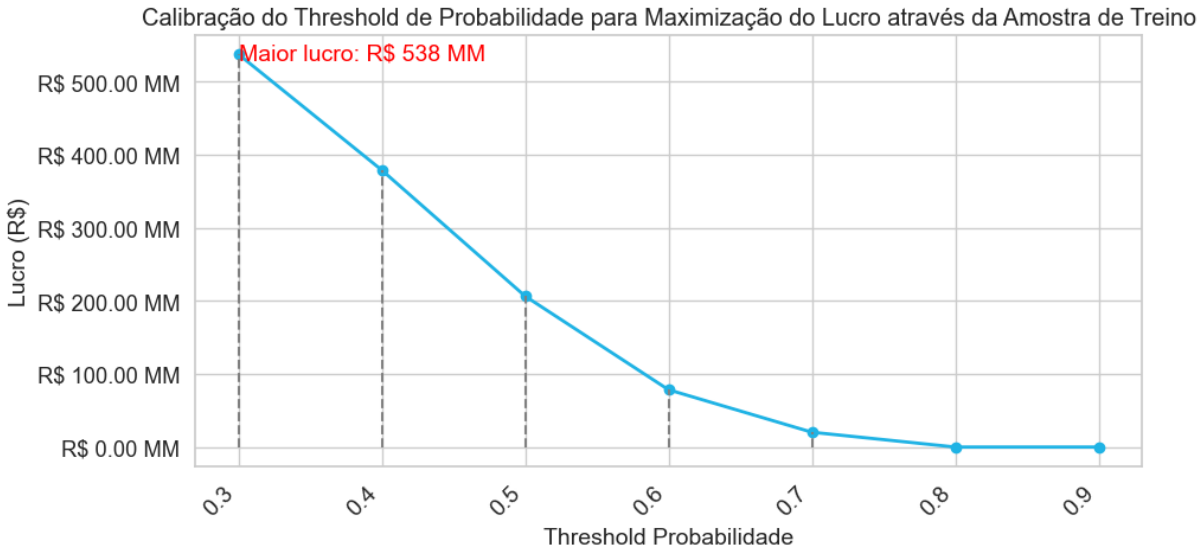
Como pontos de melhoria, em casos de disponibilidade de maior poder computacional, optar por melhorias na parte de pré-processamento como o KNN Imputer, o qual preenche valores ausentes de acordo com a similaridade dos N vizinhos mais próximos; e a substituição do Bayes Search pelo o algoritmo Hyperopt, uma vez que o Hyperopt incorpora o Bayes Search junto de mecanismos como o early stopping, o qual é um método que finaliza a otimização assim que a mesma deixar de apresentar melhorias, garantindo a obtenção da melhor combinação de hiperparâmetros possível e otimizando ainda mais os resultados.

4.3 Definição do Threshold de Probabilidade

Embora as métricas sejam satisfatórias tanto para a política quanto para o modelo, é importante pontuar que a probabilidade retornada por ambas metodologias pode não ser a melhor possível. Essa decisão deve ser guiada pelo impacto financeiro, de modo que o limiar

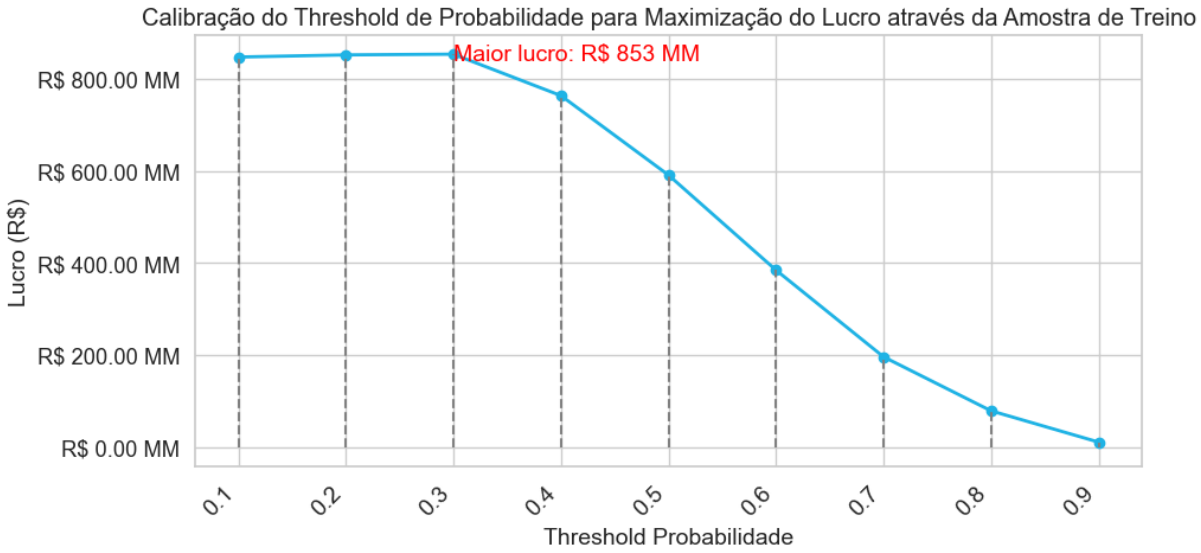
de decisão responsável por conceder ou não o crédito seja aquele que resulte no maior lucro possível. Para decidir o melhor ponto de corte, como mostrado nas figuras 28 e 29, criou-se uma abordagem iterativa responsável por demonstrar que clientes com probabilidade de serem bons pagadores igual ou inferior a 0.3 não estão deveriam receber crédito para ambas metodologias.

Figura 28 – Calibração do Threshold de Probabilidade para a Política de Crédito



Fonte: Autoria Própria.

Figura 29 – Calibração do Threshold de Probabilidade para o Modelo de Machine Learning



Fonte: Autoria Própria.

4.4 Abordagem Tradicional vs Abordagem por Modelagem Matemática e Aprendizado de Máquina

Mesmo que a política e o modelo sejam abordagens diferentes, o intuito de ambas é identificar corretamente os bons e maus pagadores. Esta etapa oferece uma visão holística a respeito das vantagens e desvantagens de cada uma delas, portanto, buscou-se quantificar o impacto de ambas as metodologias na concessão de crédito. Sendo assim, objetiva-se compreender qual delas proporciona maior eficiência através de resultados expressos em métricas de retorno financeiro, risco de crédito e Machine Learning.

Os resultados da tabela 10 evidenciam que em todos os cenários o modelo superou a política. Valores superiores de precision, recall, f1-score, AUC e KS demonstram que o método matemático foi capaz de identificar melhor os bons e maus pagadores. Notavelmente, o modelo alcançou um lucro adicional de quase R$ 100 milhões, representando um ganho de 6.68% de ROCP em comparação com a política.

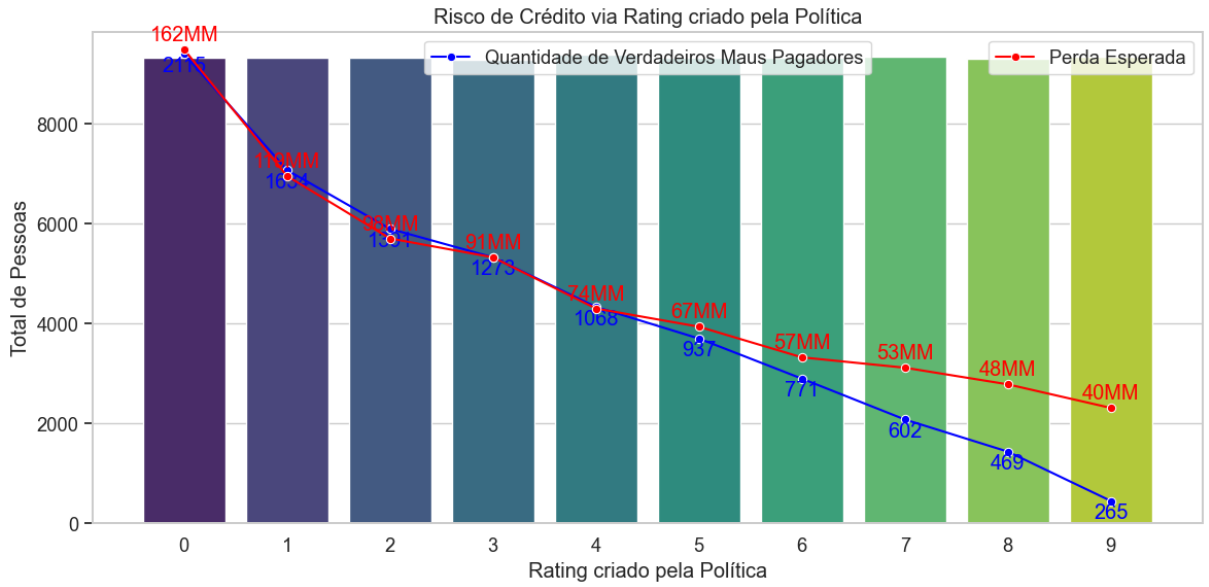
Tabela 10 – Comparação Política vs Modelo de Aprendizado de Máquina

Etapa	***Método***	***Total de Exposição***	***Retorno Financeiro***	***ROCP***	***Precision***	***Recall***	***F1-Score***	***AUC***	***KS***
Amostra Final	*Política*	*R$1.335 B*	*R$ 118MM*	*8.87%*	*0.19*	*0.44*	*0.26*	*0.67*	*0.23*
Amostra Final	*Modelo*	*R$1.335 B*	*R$ 210MM*	*15.75%*	*0.20*	*0.61*	*0.31*	*0.72*	*0.32*

Fonte: Autoria Própria.

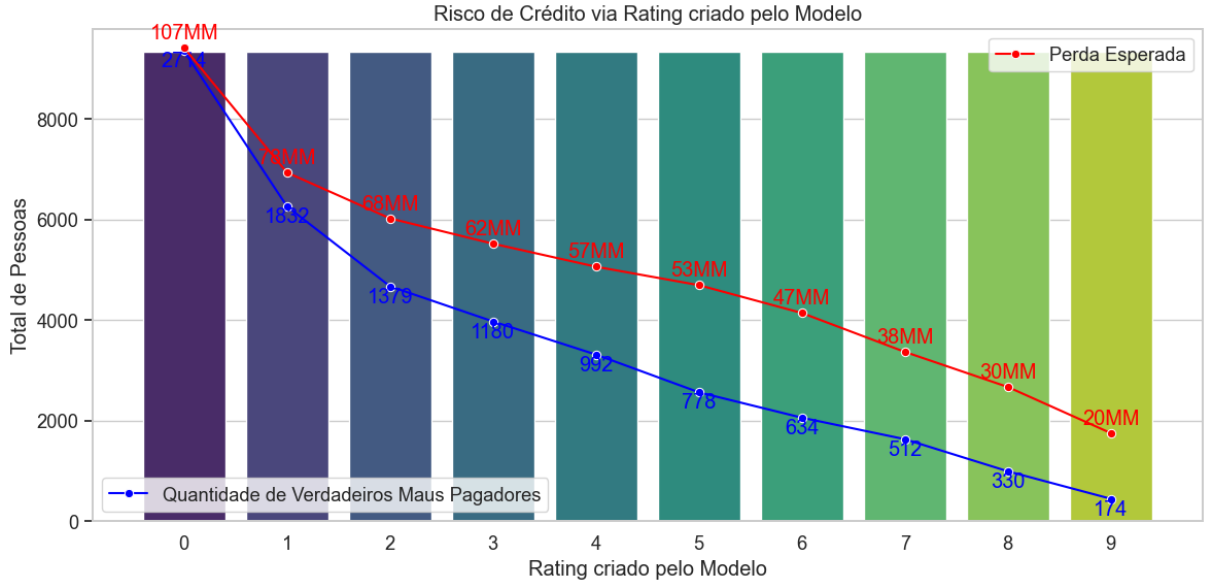
Para identificar diferentes perfis de risco, criou-se um sistema de pontuação (Rating) baseado na PD de cada cliente, como mostrado pelas figuras 30 e 31 e pela tabela 11. Esse sistema afirma que quanto maior o Rating, maior a probabilidade de adimplência, dessa forma, nota-se que o modelo possui ordenação superior quando comparado a política. Em todos os Ratings, o modelo incorre em perdas financeiras menores. Além disso, Ratings mais baixos concentram mais clientes inadimplentes, ao passo que Ratings mais altos agrupam mais clientes adimplentes, evidenciando ganhos expressivos na discriminação dos perfis de risco do portfólio.

Figura 30 – Risco de Crédito x Rating - Política



Fonte: Autoria Própria.

Figura 31 – Risco de Crédito x Rating - Modelo



Fonte: Autoria Própria.

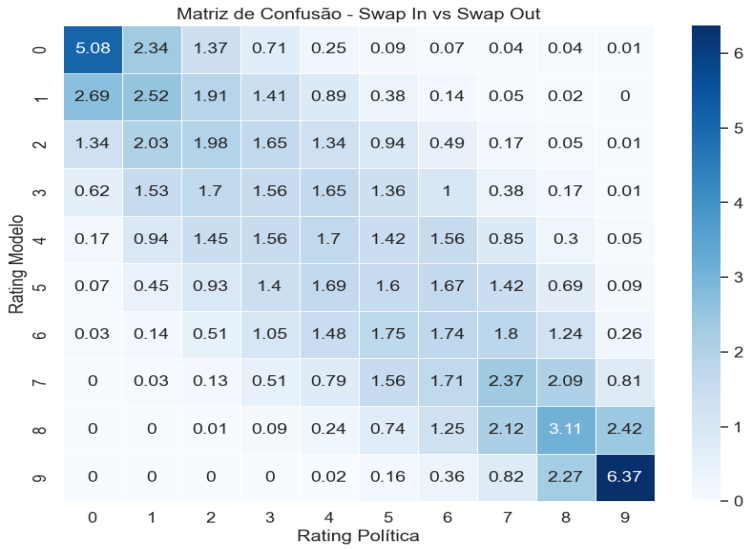
Tabela 11 – Ordenação de Bad para ambas metodologias

Rating	***BAD Política***	***BAD Modelo***
0	*22.69%*	*29.10%*
1	*17.53%*	*19.64%*
2	*14.92%*	*14.79%*
3	*13.73%*	*12.65%*
4	*11.39%*	*10.64%*
5	*10.05%*	*8.34%*
6	*8.28%*	*6.80%*
7	*6.44%*	*5.49%*
8	*5.04%*	*3.54%*
9	*2.84%*	*1.87%*

Fonte: Autoria Própria.

A fim de compreender como o modelo altera o perfil dos clientes em comparação ao Rating criado pela política, criou-se um *Swap In – Swap Out*. Essa análise identifica quantas instâncias o modelo inferiu um Rating diferente, de modo a entender a performance de quem aumentou, diminuiu ou permaneceu constante. Pelos valores expressos na figura 32 e na tabela 12, nota-se que o público de *upgrade* é excelente dado o ROCP e a taxa de inadimplência, ao passo que o público de *downgrade* não apresenta performance tão interessante e com maior índice de *default*, sendo este mais um indicativo de que a abordagem matemática é superior a política.

Figura 32 – Swap In – Swap Out



Fonte: Autoria Própria.

Tabela 12 – Swap In – Swap Out Resumido

Swap In - Swap Out	***Quantidade de Pessoas***	***ROCP Segmentado***	***%BAD***
Upgrade	*36.38%*	*23.46%*	*8.54%*
Manutenção	*28.02%*	*13.22%*	*11.64%*
Downgrade	*35.61%*	*7.59%*	*13.81%*

Fonte: Autoria Própria.

Além disso, segmentando a análise de Swap In – Swap Out para a visão de produtos, percebe-se pelas tabelas 13 e 14 que em quase todos os casos o modelo sugeriu Upgrade para clientes com alto retorno financeiro e Downgrade para clientes de baixa rentabilidade. Nos poucos casos de ROCP negativo, nota-se que em todos o modelo aplicou Manutenção ou Downgrade, sendo este mais um sinal positivo da aplicação do método matemático.

Tabela 13 – Swap In – Swap Out Segmentado I

Destino	***Swap In - Swap Out***	***Quantidade de Pessoas***	***ROCP Segmentado***
Reforma da Casa	*Upgrade*	*38%*	*24.86%*
	Manutenção	*27.82%*	*13.27%*
	Downgrade	*33.38%*	*8.2%*
Cartão de Crédito	*Upgrade*	*35%*	*22.16%*
	Manutenção	*31.75%*	*13.13%*
	Downgrade	*33.15%*	*9.31%*

Consolidação de Dívidas	*Upgrade*	*37.27%*	*23.88%*
	Manutenção	*27.38%*	*13.54%*
	Downgrade	*35.35%*	*7.89%*
Férias	*Upgrade*	*45.76%*	*12.74%*
	Manutenção	*18.95%*	*21.63%*
	Downgrade	*35.28%*	*4.2%*
Casa	*Upgrade*	*27.56%*	*30.66%*
	Manutenção	*26.41%*	*11.15%*
	Downgrade	*46.01%*	*6.62%*
Médicos	*Upgrade*	*44.87%*	*18.45%*
	Manutenção	*20.36%*	*16.51%*
	Downgrade	*34.75%*	*2.22%*

Fonte: Autoria Própria.

Tabela 14 – Swap In – Swap Out Segmentado II

Destino	***Swap In - Swap Out***	***Quantidade de Pessoas***	***ROCP Segmentado***
Pequeno Negócio	*Upgrade*	*26.62%*	*21.16%*
	Manutenção	*24.04%*	*4.69%*
	Downgrade	*49.34%*	*-2.83%*
Carro	*Upgrade*	*23.69%*	*30.98%*
	Manutenção	*47.95%*	*8.09%*
	Downgrade	*28.36%*	*8.77%*
Outro	*Upgrade*	*39.14%*	*23.81%*
	Manutenção	*23.68%*	*13.76%*
	Downgrade	*37.18%*	*3.16%*
Compra Importante	*Upgrade*	*29.33%*	*21.28%*
	Manutenção	*30.94%*	*10.63%*
	Downgrade	*39.73%*	*5.85%*
Casamento	*Upgrade*	*6.58%*	*20.61%*
	Manutenção	*22.59%*	*19.67%*
	Downgrade	*70.83%*	*5.39%*
Mudança	*Upgrade*	*36.71%*	*20.83%*
	Manutenção	*23.59%*	*14.01%*
	Downgrade	*39.7%*	*2.82%*

Educação	*Upgrade*	*1.2%*	*29.93%*
	Manutenção	*6.74%*	*-63.09%*
	Downgrade	*92.13%*	*-15.57%*
Energia Renovável	*Upgrade*	*44.29%*	*23.05%*
	Manutenção	*12.86%*	*-8.55%*
	Downgrade	*44.29%*	*-3.65%*

Fonte: Autoria Própria.

4.5 Dados Reais vs Dados de Simulação

A revisão da literatura proporciona maior embasamento teórico, logo, destaca-se como um meio estratégico de aprofundamento sobre o tema e identificação de melhorias. Além de insights valiosos, deseja-se comparar os resultados obtidos em dados reais com os valores resultantes da simulação criada neste trabalho. A comparação entre dados reais e dados de simulação é fundamental para o entendimento das diferenças entre a cenários empíricos e de teste. Nesta etapa, objetiva-se apresentar o Estado da Arte através de desafios similares ao deste trabalho. Como forma de direcionar o estudo e inferir as nuances entre ambientes diferentes, comparou-se este trabalho com os resultados de duas pesquisas sobre modelagem de risco de crédito a partir de dados reais.

Na dissertação de Castro (2022), explorou-se a comparação entre diversas metodologias de Machine Learning aplicadas à gestão de risco de crédito através de uma amostra de dados da empresa Serasa Experian. Na problemática, considerou-se um cliente como inadimplente aquele que possua ao menos uma dívida não paga após um ano a data de concessão. Finalmente, testou-se os modelos de Regressão Logística, SVM, Random Forest e Gradient Boosting, de modo que as principais métricas avaliadas foram o KS e a AUC. Como a pesquisa envolve pessoas físicas e jurídicas, optou-se por explorar apenas os resultados referentes ao primeiro caso. Dessa forma, verificou-se a utilização de 281 variáveis de entrada para todos modelos, destacando a performance do Gradient Boosting como o modelo de maior desempenho, registrando um AUC de 79,6% e KS de 45%.

Almeida (2021) pautou-se na aplicação do modelo de Redes Neurais *long short-term memory* (LSTM) em uma amostra de dados da instituição Sicoob. Neste trabalho, definiu-se que o cliente seria classificado como inadimplente caso tivesse atraso superior a 90 dias após

12 meses em relação a data de concessão do crédito. Observou-se o uso de apenas 7 variáveis de entrada, identificando um AUC de 78,25% e KS de 42%.

A tabela 15 consolida os resultados da comparação entre os três projetos:

Tabela 15 – Dados Reais vs Dados de Simulação

Abordagem	***AUC***	***KS***
XGBoost Dados Reais	*79.6%*	*45%*
LSTM Dados Reais	*78.25%*	*42%*
XGBoost Dados Simulação	*72%*	*32%*

Fonte: Autoria Própria.

Em relação ao trabalho de Castro (2022), nota-se que a grande diferença quanto a este projeto está no número de variáveis. Enquanto o primeiro possui 281, o segundo apresenta 21. Embora maior quantidade de variáveis não signifique necessariamente melhor desempenho do modelo, significa uma exposição mais abrangente a diferentes informações. Dessa forma, a amplitude de variáveis fornece mais informação sobre diversos eventos e o método de seleção pode escolher as melhores dentre elas. Isso não apenas enriquece o processo de decisão como contempla maior capacidade de generalização e aprendizado do modelo.

Comparado a dissertação de Almeida (2021), a primeira discrepância ocorre durante a montagem da *Target*, visto que o autor utilizou uma amostra balanceada entre adimplentes e inadimplentes durante o treinamento. Apesar de não parecer significativo e ser uma condição extremamente rara em risco de crédito, o fato da amostra estar balanceada auxilia consideravelmente o modelo a aprender o fenômeno, pois durante o treinamento há mais instâncias do evento de interesse para que o algoritmo compreenda o comportamento. No presente projeto, como historicamente há diferença significativa da quantidade de adimplentes e inadimplentes, optou-se por realizar o treinamento com uma amostra desbalanceada de modo que o modelo pudesse aprender o comportamento real.

A segunda diferença está contida no algoritmo utilizado, pois Almeida (2021) aplicou um modelo extremamente potente de Redes Neurais Artificiais Recorrentes (RNN's) conhecido como LSTM. RNN é uma arquitetura projetada para processas dados sequenciais de modo que sua principal característica é a capacidade de ter conexões retroativas, portanto, as informações são mantidas em um tipo de "memória" e utilizada em estágios futuros para melhorar a predição. Nesse âmbito, o LSTM é um modelo RNN especializado para lidar com

desafios associados às dependências temporais de modo que a retenção da “memória” ocorre por períodos estendidos e é feita pelos *gates* (portões). Como risco de crédito é muito relacionado às condições econômicas dos clientes e as mesmas mudam ao longo do tempo, este modelo tem potencial para alcançar bons resultados, todavia, modelos de Redes Neurais em geral são complexos e difíceis de serem interpretados, tornando-se um desafio para o entendimento de pessoas não técnicas.

5 CONCLUSÃO

Dado que o principal objetivo de uma análise de risco de crédito é separar os bons e maus pagadores de modo a maximizar o retorno financeiro, concluiu-se que esta pesquisa representou um avanço ao comparar a abordagem tradicional com a abordagem direcionada por modelagem matemática e aprendizado de máquina.

Os resultados demonstraram que técnicas de Machine Learning, embora complexas, são capazes de proporcionar ganhos financeiros ao mesmo tempo que são uma automação do processo de concessão. Pelo fato de o método quantitativo captar relações complexas entre as variáveis de maneira automática, os analistas de crédito podem otimizar o tempo gasto na avaliação e melhorar a eficiência dos projetos.

Além da otimização do lucro e da discriminação, a abordagem matemática apresenta maior estabilidade justamente por basear-se em mais variáveis. Essa característica permite que a concessão seja baseada em diversos fatores, proporcionando maior abrangência de decisão e, consequentemente, auxiliando na segmentação de perfis de risco.

Embora promissora, destacam-se algumas ressalvas a respeito das limitações deste estudo. Pela ausência de grande poder computacional, a aplicação de técnicas mais avançadas de pré-processamento não foi possível. Além disso, todo o processo de análise baseou-se em indivíduos tomadores, portanto, instâncias as quais um dia já foram aprovadas. Isso incorre que amostras já negadas desde o início foram excluídas, resultando na criação de um possível viés. Conforme explicado por Sicsú (2010), como um modelo de Credit Scoring destina-se a avaliar todos os proponentes potenciais, ele deve basear-se nos bons e maus clientes de mercado e não apenas nos bons e maus clientes anteriormente aprovados pelo credor. Para trabalhos futuros, recomenda-se a ampliação ou troca da base de dados de modo que a nova população tenha exemplos de todos os casos citados, permitindo assim a expansão da pesquisa a qual envolve também a inferência de negados.

Finalmente, é altamente recomendável implementar metodologias de monitoramento contínuo do modelo. Para isso, o desenvolvimento de um ambiente de engenharia de Machine Learning torna-se crucial para uma eficaz gestão dos riscos. Esse ambiente possibilita a monitoração recorrente dos resultados do modelo, assegurando sua estabilidade e confiabilidade ao longo do tempo. Além disso, caso ocorram alterações nos resultados, esse ambiente permite o retreinamento automático do modelo, garantindo sua adaptação à

população em constante mudança. Dessa forma, o modelo se mantém aplicável e eficaz, refletindo as necessidades e características da população atual.

Embora a problemática inicial seja relacionada ao risco de crédito, a mesma metodologia pode ser facilmente adaptada e aplicada em outros contextos. Um exemplo prático disso é a otimização da venda de um determinado produto. Ao definir regras determinísticas que identifiquem clientes com alta propensão de compra e clientes com baixa propensão, é possível criar um modelo de classificação capaz de identificar pessoas com maior probabilidade de adquirir o produto.

Assim como foi feito com o rating para risco de crédito, pode-se desenvolver um rating de propensão à compra com base na probabilidade calculada pelo modelo de classificação. Esse rating pode então ser usado para segmentar os clientes e direcionar esforços e investimentos em campanhas específicas apenas para aqueles com alta propensão a comprar. Essa abordagem direcionada resulta em um uso mais eficiente dos recursos e uma maior assertividade na venda do produto para o público-alvo mais adequado.

REFERÊNCIAS

ABREU, Mariana da Conceição Ferreira. **Modelos de avaliação de risco de crédito**. Dissertação (Mestrado em Economia na especialização de Economia Financeira) – Universidade de Coimbra, Coimbra, 2020.

ALMEIDA, Gustavo Durães. **Modelagem de risco de crédito via LSTM**. Dissertação (Mestrado em Estatística) – Universidade de Brasília, Brasília, 2021.

ARAÚJO, Elaine Aparecida; MONTREUIL CARMONA, Charles Ulises de. **Desenvolvimento de modelos credit scoring com abordagem de regressão logística para a gestão da inadimplência de uma instituição de microcrédito**. Contabilidade Vista & Revista, Minas Gerais, vol. 18, n. 3, p. 107 – 131, set. 2007.

ARAÚJO, João Paulo Bezerra. **Interpretabilidade de modelos de machine learning: aplicação no mercado de crédito**. Trabalho de Conclusão de Curso (Bacharel em Engenharia Elétrica) – Universidade Federal do Ceará, Fortaleza, 2020.

BHALLA, Deepanshu. **SAS: calculating ks statistics**. [S. I.]: Listen DATA, 2016. Disponível em: https://www.listendata.com/2016/01/sas-calculating-ks-test.html. Acesso em: 22 jul. 2023.

BORIN, Edson. **Capacitação profissional em tecnologias de inteligência artificial**. Instituto de Computação, Universidade Estadual de Campinas, Campinas, 2023.

BRUCE, Peter; BRUCE Andrew. **Practical statistics for data scientists**. 1. ed. Rio de Janeiro: Alta Books, 2019.

CAMARGO, Bruna Emy. **Número de inadimplentes volta a crescer e chega a 65 milhões de pessoas em janeiro**. São Paulo: Estadão, 2023. Disponível em: https://www.estadao.com.br/economia/numero-inadimplentes-cresce-65-milhoes-pessoas-janeiro/#:~:text=Quatro%20em%20cada%20dez%20brasileiros,m%C3%AAs%20do%20ano%2C%20segundo%20pesquisa&text=O%20n%C3%BAmero%20de%20inadimplentes%20no,rela%C3%A7%C3%A3o%20a%20dezembro%20de%202022. Acesso em: 19 mar. 2023.

CASTRO, Jane Simões de. **Estudo comparativo entre metodologias de aprendizado de máquina e híbridas aplicadas a risco de crédito**. Dissertação (Mestrado em Administração) – FECAP, São Paulo, 2022.

ESTATÍSTICAS monetárias e de crédito. [S.I.]: Banco Central do Brasil, [ca, 2023]. Disponível em: https://www.bcb.gov.br/estatisticas/estatisticasmonetariascredito. Acesso em: 17 dez. 2023.

FORTI, Melissa. **Técnicas de machine learning aplicadas na recuperação de crédito do mercado brasileiro**. Dissertação (Mestrado em Economia) – Escola de Economia de São Paulo, Fundação Getúlio Vargas, São Paulo, 2018.

GÉRON, Aurélien. **Hands-on machine learning with scikit-learn, keras and tensorflor: concepts, tools and techniques to build intelligent systems**. 2. ed. Rio de Janeiro: Alta Books, 2019.

GIMENES, Pedro. **Política de crédito: o que é, quais são as fases e a importância para o seu negócio**. [S.I.]: Boa Vista, 2022. Disponível em: https://www.boavistaservicos.com.br/blog/destaque/o-que-e-politica-de-credito/. Acesso em: 25 jan. 2024.

GUIMARÃES XAVIER, Caroline. **Risco na análise de crédito**. Trabalho de Conclusão de Curso (Bacharel em Ciências Contábeis) – Departamento de Ciências Contábeis, Universidade Federal de Santa Catarina, Florianópolis, 2011.

HARRISON, Matt. **Machine learning pocket reference**. 1. ed. São Paulo: Novatec, 2020.

HYPERPARAMETER optimization. In: WIKIPÉDIA: a enciclopédia livre. [São Francisco, CA: Wikimedia Foundation, 2024]. Disponível em: https://en.wikipedia.org/wiki/Hyperparameter_optimization. Acesso em: 31 jan. 2024.

CHAIA, Alexandre Jorge. **Modelos de gestão do risco de crédito e sua aplicabilidade ao mercado brasileiro**. Dissertação (Mestrado em Administração) – Departamento de Administração, Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2003.

KUMAR, Ajitesh. **Machine learning: inference & prediction difference**. [S. I.]: Analytics Yogi, 2023. Disponível em: https://vitalflux.com/machine-learning-inference-prediction-difference/#:~:text=Prediction%20is%20the%20process%20of,the%20predictor%20and%20response%20variables. Acesso em: 23 ago. 2023.

LEVADA, Alexandre Luis Magalhães. **Programação científica com Python**. Departamento de Computação, Centro de Ciências Exatas e Tecnologia, Universidade Federal de São Carlos, São Carlos, 2021.

LIMA, Jorge Cláudio Cavalcante de Oliveira. **A importância de conhecer a perda esperada para fins de gerenciamento do risco de crédito**. Revista do BNDES, Rio de Janeiro, V.15, N.30, P.271-302, 2008. Acesso em: 27 ago. 2023.

LIMA, Rafaela Somavila. **Criação de projeto de ciência de dados utilizando a metodologia CRISP-DM em conformidade com a LGPD**. Trabalho de Conclusão de Curso (Especialização em Ciência de Dados e Suas Aplicações) – Departamento Acadêmico de Informática, Universidade Tecnológica Federal do Paraná, Paraná, 2021.

LUKOSIUNAS, Andreza. **Aplicação de técnicas de machine learning em modelos de escore de crédito**. Dissertação (Mestrado em Economia) – INSPER, São Paulo, 2018.

MONTOYA, Anna; ODINTSOV, Kirill; KOTEK, Martin. **HOME CREDIT DEFAULT RISK**. Kaggle Competition. Disponível em: https://www.kaggle.com/competitions/home-credit-default-risk/overview. Acesso em: 10 janeiro 2023.

MORETTIN, Pedro A.; BUSSAB, Wilton De O. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.

MORETTIN, Pedro A.; SINGER, Julio M. **Estatística e Ciência de Dados**. 1. ed. São Paulo: LTC, 2022.

MORA, Mônica. **A evolução do crédito no Brasil entre 2003 e 2010**. Rio de Janeiro: Instituto de Pesquisa Econômica Aplicada, 2015. Disponível em: https://repositorio.ipea.gov.br/bitstream/11058/3537/1/td2022.pdf. Acesso em: 17 dez. 2023.

PARK, Sung. **Understand and use a business credit risk score**. [S.I]: Experian, 2020. Disponível em: https://blogbr.clear.sale/ciclo-de-credito. Acesso em: 15 jul. 2023.

PEREIRA, Pedro Miguel Pinhal. **Análise de risco de crédito usando algoritmos de machine learning**. Dissertação (Mestrado em Matemática Financeira) – Departamento de Matemática, Universidade de Lisboa, Lisboa, 2020.

RECEIVER operating characteristic. In: WIKIPÉDIA: a enciclopédia livre. [São Francisco, CA: Wikimedia Foundation, 2024]. Disponível em: https://en.wikipedia.org/wiki/Receiver_operating_characteristic. Acesso em: 31 jan. 2024.

SANTOS, Patrick Ferreira dos. **Uso de técnicas de machine learning para análise de risco de crédito**. Dissertação (Mestrado Profissional em Economia) – Departamento de Economia,

Faculdade de Administração Contabilidade e Economia, Universidade de Brasília, Brasília, 2022.

SEBBEN, Renivaldo José. **Análise de risco de crédito e cobrança: como conceder crédito com segurança e recuperar créditos inadimplentes**. 1. ed. São Paulo: Novatec Editora Ltda, 2020.

SELAU, Lisiane Priscila Roldão. **Modelagem para concessão de crédito a pessoas físicas em empresas comerciais: da decisão binária para a decisão monetária**. Tese (Doutorado em Administração) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

SERASA. **Descubra aqui o que é a política de crédito e como elaborar uma**. [S.I.]: Serasa 2022. Disponível em: https://www.serasaexperian.com.br/blog-pme/descubra-aqui-o-que-e-a-politica-de-credito-e-como-elaborar-uma/. Acesso em: 25 jan. 2024.

SERASA. **Mapa da inadimplência e negociação de dívidas no Brasil: dezembro 2023**. [S.I.]: Serasa, 2023. Disponível em: https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/. Acesso em: 19 dez. 2023.

SERASA. **Política de crédito: veja tudo o que você precisa saber sobre o assunto**. [S.I.]: Serasa, 2021. Disponível em: https://www.serasaexperian.com.br/conteudos/credito/politica-de-credito-veja-tudo-o-que-voce-precisa-saber-sobre-o-assunto/. Acesso em: 19 jan. 2024.

SFEIR, Elias. **A relação crédito-pib no Brasil: histórico e comparação internacional**. [S. l.]: ANBC, 2021. Disponível em: https://anbc.org.br/a-relacao-credito-pib-no-brasil/. Acesso em: 17 dez. 2023.

SHELCI SILVA, Juelline. **Gerenciamento integrado de riscos: modelos de predição de risco de crédito em machine learning para a identificação de ativos problemáticos em uma instituição financeira**. Dissertação (Mestrado Profissional em Economia) – Departamento de Economia, Faculdade de Administração Contabilidade e Economia, Universidade de Brasília, Brasília, 2022.

SICSÚ, Abraham Laredo. **Credit Scoring: desenvolvimento, implantação e acompanhamento**. 1. ed. São Paulo: Blucher, 2010.

SILVA, Daniel de Oliveira Silva. **Otimização de hiperparâmetros de algoritmos de machine learning aplicado no contexto de análise de risco de crédito**. Trabalho de

Conclusão de Curso (Especialização em Ciência de Dados) – Universidade Tecnológica Federal do Paraná, Dois Vizinhos, 2022.

TCHILIAN, Felipe. **Ciclo de crédito: entenda e otimize a jornada do cliente**. [S.I]: ClearSale, 2022. Disponível em: https://blogbr.clear.sale/ciclo-de-credito. Acesso em: 22 mar. 2023.

UNIVESTIDADE FEDERAL DO PARANÁ. Laboratório de Estatística e Geoinformação. **Métodos de reamostragem**. [S. l.]: UFPR, [ca. 2018]. Disponível em: http://cursos.leg.ufpr.br/ML4all/apoio/reamostragem.html. Acesso em: 22 ago. 2023.

APÊNDICE A – VARIÁVEIS DISPONÍVEIS NA BASE DE DADOS

Variável	Descrição
Situação_do_emprestimo	*Situação atual do empréstimo, sendo possível classificar o cliente em "Bom" ou "Ruim"*
Qt_parcelas	*Quantidade de parcelas escolhidas para o pagamento da dívida*
Grau_de_emprestimo	*Grau de empréstimo atribuído*
Subclasse_de_emprestimo	*Subclasse de empréstimo atribuído*
Produto_de_credito	*Produto de crédito contratado pelo cliente*
Valor_emprestimo_solicitado	*Valor do empréstimo solicitado pelo cliente*
Taxa_de_juros	*Taxa de juros do empréstimo*
Data_financiamento_emprestimo	*Data em que o cliente pretende pagar o empréstimo financiado*
Produto_disponível_publicamente	*Flag que indica se o produto está disponível publicamente ou não*
Plano_de_pagamento	*Flag que indica se algum plano de pagamento foi implementado para o empréstimo*
Tipo_de_concessao_do_credor	*Status da listagem inicial do empréstimo*
Pagamento_mensal	*Valor da parcela acrescida de juros*
Cargo_cliente	*Cargo fornecido pelo cliente*
Qt_anos_mesmo_emprego	*Duração de tempo em que o cliente está no mesmo emprego*
Status_propriedade_residencial	*Flag que indica o status da propriedade residencial do cliente*
Renda_comprovada	*Flag que indica se a renda foi comprovada ou não*
Inadimplência_vencida_30dias	*Número de incidências de inadimplência vencidas há mais de 30 dias nos últimos 2 anos*
Faturamento_anual	*Rentabilidade anual declarada pelo cliente*
Comprometimento_de_renda_anual	*Porcentagem da renda anual comprometida*
Estado	*Estado do cliente*
Limite_total_produtos_credito	*Limite total considerando todos os produtos de crédito*
Limite_total_rotativos	*Limite total de rotativos*
Limite_rotativos_utilizado	*Valor do Limite de rotativos utilizado*

Taxa_utilizacao_limite_rotativos	***Percentual de uso do limite de rotativos***
Qt_produtos_credito_contratados_atualmente	***Quantidade de produtos de crédito que o cliente tem atualmente contratado***
Qt_produtos_credito_contratados_historicamente	***Número total de produtos de crédito que o cliente contratou em seu histórico***
Registros_publicos_depreciativos	***Número de registros públicos depreciativos (registros criminais, processos judiciais, demissões por má conduta, etc...)***
Consulta_credito_6meses	***Número de consultas nos últimos 6 meses***
Data_contratacao_primeiro_produto_credito_line	***Data em que o primeiro produto de crédito foi contratado pelo cliente***
Qt_meses_desde_ultimo_registro_publico	***Número de meses desde o último registro público***
Qt_meses_classificacao_mais_recente_90d	***Meses desde a classificação mais recente de 90 dias***
Qt_meses_ultima_inadimplencia	***Meses desde a última inadimplência do cliente***