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

Leonardo Aderaldo Vargas, Prof. Dr. Leopoldo André Lusquino Filho
Prof. Dr. Galdenoro Botura Junior
UNESP – Univ Estadual Paulista, Instituto de Ciência e Tecnologia, Engenharia de Controle e Automação, Sorocaba SP
E-mails: leonardo.vargas@unesp.br; galdenoro.botura@unesp.br , leopoldo.lusquino@unesp.br

Abstract

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. After the change, a gain of 17% of Recall, 5% of AUC, 9% of KS and 6.88% of ROCP was obtained, representing an increase of almost R$ 100 million of profit and allowing the best classification of the client's risk profile.

Resumo

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. Após a mudança, 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.

1. Introdução

O grande desafio da análise de risco de crédito reside em distinguir os clientes confiáveis e aqueles propensos a inadimplência de modo a maximizar a rentabilidade. Nesse âmbito, classificações incorretas resultam em perdas financeiras significativas para o credor, além de aumentar o endividamento do cliente, prejudicando todo o ecossistema econômico.

O avanço computacional permitiu a aplicação de novas metodologias de análise, portanto, o presente projeto apresenta a comparação entre uma política de crédito e um modelo de Machine Learning capazes de classificar o perfil de risco dos clientes, de modo a compreender se a abordagem matemática é superior a abordagem tradicional.

Como dados de crédito são sensíveis, optou-se pela utilização de dados anonimizados da empresa *Lending Club*, a qual é uma empresa norte-americana responsável por operar uma plataforma online de empréstimos para pessoas físicas.

2. Conceituação e fundamentação teórica

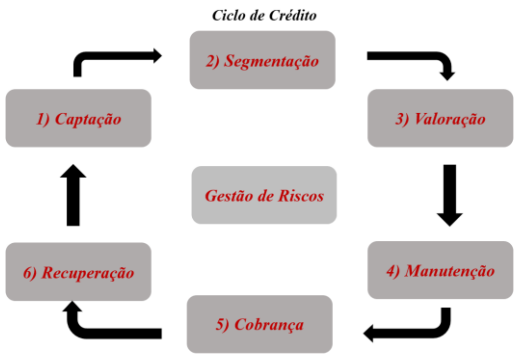
2.1 Crédito e seus princípios

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. 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.

Figura 1 – Fluxograma do Ciclo de Crédito



Fonte: *Autoria Própria.*

A etapa de valoração é onde as metodologias de risco de crédito entram em ação e avaliam os clientes com base em diversas variáveis a fim de classificá-los como bons ou maus pagadores, logo, são o foco deste trabalho.

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.

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.

Capacidade – Refere-se à validação sobre as condições do tomador pagar suas dívidas, 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 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, 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

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.

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*. Seu objetivo é determinar a probabilidade do cliente tornar-se inadimplente baseando-se em teorias matemáticas.

2.3 Modelos de Aprendizado de Máquina para Classificação

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 identificar a classe de novas amostras.

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 – A regressão logística é uma extensão da regressão linear capaz de transformar a varável resposta contínua em uma variável categórica. 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.

Modelos Bayesianos – O Teorema de Bayes é um mecanismo o qual descreve a forma de atualizar a probabilidade de uma hipótese com base em novas evidências. A partir do Teorema de Bayes, pode-se modelar um fenômeno dada as variáveis, criando-se o Naive Bayes, o qual é 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.

Modelos Baseados em Distâncias – O KNN compara a nova instância com os K elementos mais próximos baseados nas variáveis utilizadas. Essa comparação é realizada via cálculos de distância e, a depender da distância escolhida, os resultados podem ser distintos.

Modelos de Árvore – 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. Parte-se 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.

Modelos Ensemble – Métodos Ensemble são algoritmos que combinam diversos preditores fracos de forma a criar um preditor forte e robusto. Neste trabalho, abordou-se os métodos de Bagging e Boosting. Morettin e Singer (2021) 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. o Boosting envolve a geração sequencial de árvores de decisão com base na atualização de pesos para cada elemento no conjunto de treinamento. 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. 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

2.4 Métricas de Avaliação

Uma Matriz de Confusão é uma matriz quadrada utilizada para comparar os valores preditos do modelo com os valores reais. Durante a classificação de um elemento, há quatro situações possíveis, sendo elas Verdadeiro Negativo (VN), Verdadeiro Positivo (VP), Falso Negativo (FN) e Falso Positivos (FP). Pode-se avaliar a quantidade de cada um desses indicadores e, consequentemente, a performance de um modelo de classificação a partir das seguintes métricas:

Precision - A precisão quantifica a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias classificadas como positivas pelo modelo.

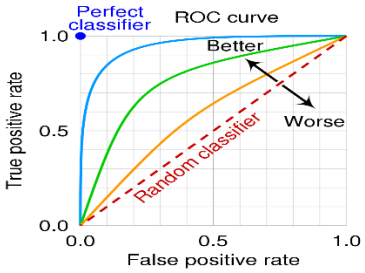
$$Precisão = \frac{VP}{VP+FP} \quad \text{(I)}$$

Recall – O recall quantifica a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias que realmente são positivas na amostra.

$$Recall = \frac{VP}{VP+FN} \quad \text{(II)}$$

AUC – 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), onde um valor maior indica melhor desempenho do modelo. O intuito da AUC é avaliar o quão bem o modelo classifica ambas as classes.

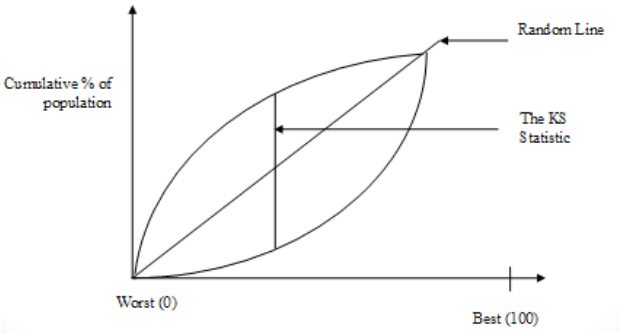
Figura 2 – Curva ROC e AUC



Fonte: *Receiver operating characteristic. Disponível em: https://en.wikipedia.org/wiki/Receiver_operating_characteristic*

KS – Laredo (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.

Figura 3 – Kolgomorov-Smirnov (KS)



Fonte: *SAS: Calculating KS Statistics, Listen Data. https://www.listendata.com/2016/01/sas-calculating-ks-test.html*

3. Desenvolvimento

3.1 Produto

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 Materiais

Na etapa de materiais, utilizou-se a linguagem de programação Python, o sistema de controle de versão GIT e o repositório de bases de dados Kaggle.

3.3 Metodologia

De acordo com Rafaela 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.

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.

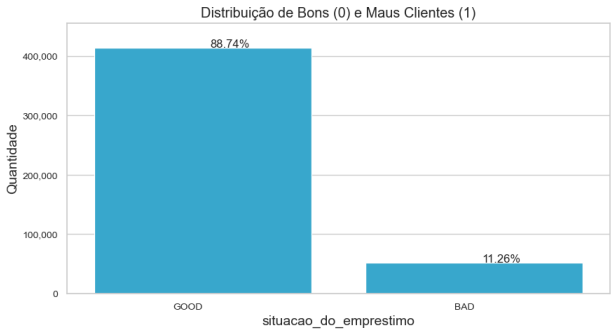
3.4 Definição da Target

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, pode-se visualizar a distribuição de "Bons e Maus Pagadores", revelando a existência de um problema de classes desbalanceadas.

Figura 4 – Distribuição de Bons e Maus Pagadores



Fonte: *Autoria Própria.*

3.5 Método de Avaliação

Como deseja-se prever o mau pagador, as principais métricas técnicas de serem avaliadas são o *Recall*, a *AUC* e o *KS*. Em relação as métricas de negócio, definiu-se uma metodologia capaz de estimar o impacto financeiro de cada tipo de erro e acerto. Um 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):

$$\begin{bmatrix} VN & FP \\ FN & VP \end{bmatrix} \cdot \begin{bmatrix} Exposição + Juros & 0 \\ Exposição & 0 \end{bmatrix} \quad \text{(III)}$$

$$RF = VN * (Exposição + Juros) - Qt\ FN * Exposição\ (IV)$$

$$ROCP = \frac{Retorno\ Financeiro}{Valor\ Total\ de\ Exposição}\ x\ 100\%\ \text{(V)}$$

3.6 Análise de Variáveis

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. 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

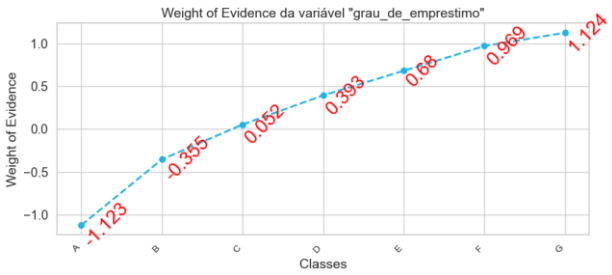
força da associação de uma classe com a variável alvo. Valores positivos de WOE significam que a categoria está associada a classe positiva, ao passo que valores negativos mostram associação com a classe negativa. O WOE é definido pela seguinte equação:

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

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.

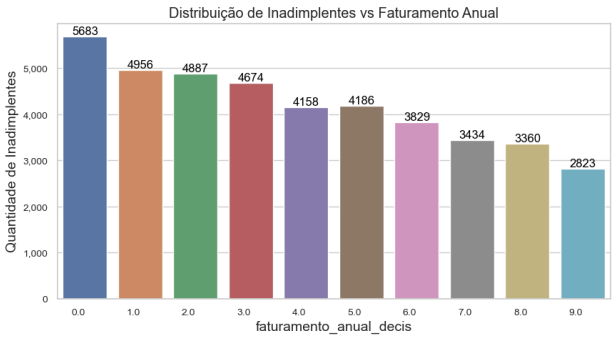
Para exemplificar, decidiu-se trazer dois exemplos para representar os casos citados. Para o primeiro, nota-se que as classes "A" e "B", por assumirem valores negativos, associam-se com a classe 0, logo, estão majoritariamente relacionadas com bons pagadores, ao passo que as demais assumem valores positivos e relacionam-se de forma mais expressiva com maus pagadores. Para o segundo, 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.

Figura 5 – Weight of Evidence da Variável Grau do Empréstimo



Fonte: *Autoria Própria.*

Figura 6 – Boxplot da Variável Faturamento Anual



Fonte: *Autoria Própria.*

3.7 Criação 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.

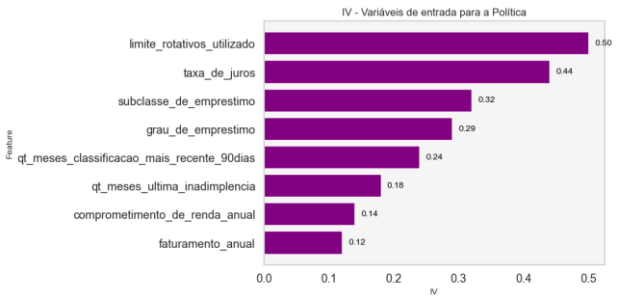
Pelo fato de tratar-se de uma abordagem menos sofisticada e mais manual quando comparada a esses modelos, a criação de uma política demanda uma seleção de variáveis mais criteriosa, sendo assim, objetiva-se construir uma política baseada no menor número de variáveis ao mesmo tempo que mantém-se o poder 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 Laredo (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:

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

Dessa forma, escolheu-se levar adiante apenas as que tiveram valores superiores a 0.1 pois, segundo Laredo (2010), resultados superiores a 0.1 sugerem que a informação é realmente útil.

Figura 7 – 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. 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.

Finalmente, realizou-se a análise de distribuição de probabilidade para cada uma das variáveis a fim de comprovar a existência de uma relação bem comportada entre a característica e a variável resposta de modo a manter apenas as variáveis com ordenação evidente.

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

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

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

3.8 Criação 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.

Infelizmente, a realidade dos dados do mundo real é frequentemente complexa e desafiadora. 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. 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. Neste projeto, aplicou-se o *Target Encoder* em todas as variáveis categóricas com o auxílio da PD calculada para cada classe.

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 justamente para permitir que essas informações sejam levadas em consideração pelo algoritmo. 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).

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 para ajustar as escalas das variáveis e para ajustar as escalas das variáveis e, assim, garantir que todas as variáveis tenham o peso ideal no processo de treinamento. Sendo assim, escolheu-se o *Min-Max Scaler*, a qual é uma técnica capaz de redimensionar as variáveis de um conjunto de dados para um intervalo específico contido em [0,1]. Sua equação é definida por:

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

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. Para esta etapa, aplicou-se três técnicas consolidadas, sendo elas o *Variance Threshold*, *Mutual Information* e *Feature Importance*.

O *Variance Threshold* 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.

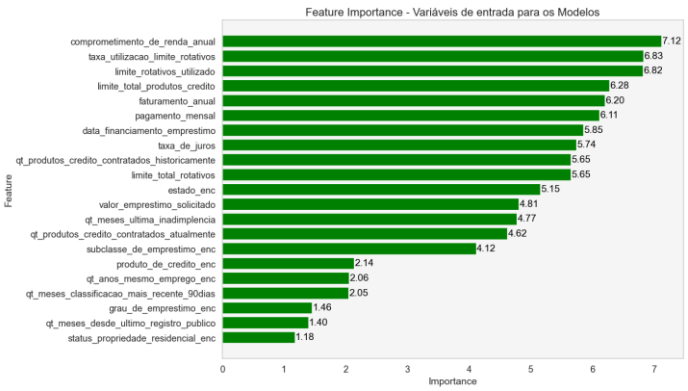
Para a análise variáveis categóricas, um dos métodos mais robustos é o *Mutual Information*. Esta técnica é uma medida estatística que quantifica a dependência entre duas variáveis aleatórias, sendo um excelente método para entender se a variável resposta possui dependência com a variável de entrada. Sua equação é definida como:

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

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.

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 seguir encontram-se as variáveis de entrada para os modelos de Machine Learning:

Figura 8 – 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.

Finalmente, como o primeiro modelo treinado costuma não possuir a melhor performance possível, utilizou-se o algoritmo de Bayes Search para a otimização de hiperparâmetros. 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.

4. Resultados obtidos

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.

O modelo vencedor dentre os testados foi o XGBoost, portanto, optou-se por levá-lo para a otimização de hiperparâmetros e compará-lo à política de crédito desenvolvida.

Os resultados 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 1 –Política vs Modelo de Aprendizado de Máquina

Método	*Política*	*XGboost + Bayes Search*
Exposição	*R$1.335 B*	*R$1.335 B*
Retorno	*R$118MM*	*R$210MM*
ROCP	*8.87%*	*15.75%*
Precision	*0.19*	*0.20*
Recall	*0.44*	*0.61*
F1-Score	*0.26*	*0.31*
AUC	*0.67*	*0.72*
KS	*0.23*	*0.32*

Fonte: *Autoria Própria.*

Para identificar diferentes perfis de risco, criou-se um sistema de pontuação (Rating) baseado na PD de cada cliente. 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.

Tabela 2 – Risco de Crédito x Rating - Política

Rating	*Perda Esperada (R$)*	*Qt Maus Pagadores*
Rating 0	*162 MM*	*2115*
Rating 1	*119 MM*	*1601*
Rating 2	*98 MM*	*1301*
Rating 3	*91 MM*	*1273*
Rating 4	*74 MM*	*1068*
Rating 5	*67 MM*	*937*
Rating 6	*57 MM*	*771*
Rating 7	*53 MM*	*602*
Rating 8	*48 MM*	*469*
Rating 9	*40 MM*	*265*

Fonte: *Autoria Própria.*

Tabela 3 – Risco de Crédito x Rating - Modelo

Rating	*Perda Esperada (R$)*	*Qt Maus Pagadores*
Rating 0	*107 MM*	*2714*
Rating 1	*78 MM*	*1832*
Rating 2	*68 MM*	*1379*
Rating 3	*62 MM*	*1180*
Rating 4	*57 MM*	*992*
Rating 5	*53 MM*	*778*
Rating 6	*47 MM*	*634*
Rating 7	*38 MM*	*512*
Rating 8	*30 MM*	*330*
Rating 9	*20 MM*	*174*

Fonte: *Autoria Própria.*

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 Laredo (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, como a população costuma estar em constante mudança econômica e social, sugere-se também a implementação de metodologias para retreino do modelo. Nesse âmbito, o desenvolvimento de um ambiente de engenharia de Machine Learning seria crucial para a correta gestão de riscos, pois ele permitiria a recorrente adaptação do modelo de forma totalmente automática, de modo que o mesmo sempre fosse aplicável para a população atual.

6. Referências

[1] LAREDO SICSÚ, Abraham. **CREDIT SCORING: DESENVOLVIMENTO, IMPLANTAÇÃO E ACOMPANHAMENTO**. São Paulo: Blucher, 2010. Disponível em: <https://www.blucher.com.br/credit-scoring_9788521205333>. Acesso em: 22 janeiro 2023.

[2] SEBBEN, Renivaldo José. **ANÁLISE DE RISCO DE CRÉDITO E COBRANÇA: COMO CONCEDER CRÉDITO COM SEGURANÇA E RECUPERAR CRÉDITOS INADIMPLENTES**. Novatec Editora Ltda, 2020. Disponível em: <https://www.amazon.com.br/gp/product/8575228269/ref=as_li_tl?ie=UTF8&camp=1789&creative=9325&creativeASIN=8575228269&linkCode=as2&tag=novatec03-20 >. Acesso em 12 de março 2023.

[3] GUIMARÃES XAVIER, Caroline. **RISCO NA ANÁLISE DE CRÉDITO**. Tese (Bacharel em Ciências Contábeis) – Departamento de Ciências Contábeis, Universidade Federal de Santa Catarina. Florianópolis, p.70, 2011. Acesso em: 15 fevereiro 2023.

[4] JORGE CHAIA, Alexandre. **MODELOS DE GESTÃO DO RISCO DE CRÉDITO E SUA APLICABILIDADE AO MERCADO BRASILEIRO**. Tese (Mestrado em Administração) – Departamento de Administração, Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo. São Paulo, p.126, 2003. Acesso em: 15 fevereiro 2023.

[5] 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. Acesso em: 20 fevereiro 2023.

[6] 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**. Tese (Mestrado Profissional em Economia) – Departamento de Economia, Faculdade de Administração Contabilidade e Economia, Universidade de Brasília. Brasília, p.74, 2022. Acesso em: 20 fevereiro 2023.

[7] FORTI, Melissa. **TÉCNICAS DE MACHINE LEARNING APLICADAS NA RECUPERAÇÃO DE CRÉDITO DO MERCADO BRASILEIRO**. Tese (Mestrado em Economia) – Escola de Economia de São Paulo, Fundação Getulio Vargas. São Paulo, 2018. Acesso em: 21 fevereiro 2023.

[8] SANTOS, Patrick Ferreira dos. **USO DE TÉCNICAS DE MACHINE LEARNING PARA ANÁLISE DE RISCO DE CRÉDITO**. Tese (Mestrado Profissional em Economia) – Departamento de Economia, Faculdade de Administração Contabilidade e Economia, Universidade de Brasília. Brasília, p.57, 2022. Acesso em: 19 fevereiro 2023.

[9] ARAÚJO, João Paulo Bezerra de. **INTERPRETABILIDADE DE MODELOS DE MACHINE LEARNING: APLICAÇÃO NO MERCADO DE CRÉDITO**. Tese (Bacharel em Engenharia Elétrica) – Universidade Federal do Ceará. Fortaleza, p.73, 2020. Acesso em: 19 de fevereiro 2023.

[10] 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.

[11] MORETTIN, Pedro A.; SINGER, Julio M. **ESTATÍSTICA E CIÊNCIA DE DADOS**. LTC, 2022. Disponível em: < https://www.grupogen.com.br/livro-estatistica-e-ciencia-de-dados-pedro-alberto-morettin-e-julio-da-motta-singer-editora-ltc-9788521638162>. Acesso em: 23 fevereiro 2023.

[12] GÉRON, Aurélien. **HANDS-ON MACHINE LEARNING WITH SCIKIT-LEARN, KERAS AND TENSORFLOR: CONCEPTS, TOOLS AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS**. O'Reilly, 2019. Disponível em: < https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ >. Acesso em: 25 Janeiro 2023.

[13] HARRISON, Matt. **MACHINE LEARNING POCKET REFERENCE**. O'Really, 2019. Disponível em: < https://www.oreilly.com/library/view/machine-learning-pocket/9781492047537/ >. Acesso em: 27 Janeiro 2023.

[14] MORETTIN, Pedro A.; BUSSAB, Wilton De O. **ESTATÍSTICA BÁSICA**. São Paulo: Saraiva, 2017. Disponível em: < https://www.saraiva.com.br/estatistica-basica---morettin---saraiva-21397/p >. Acesso em: 10 janeiro 2023.

[15] CAMARGO, Bruna Emy. **NÚMERO DE INADIMPLENTES VOLTA A CRESCER E CHEGA A 65 MILHÕES DE PESSOAS EM JANEIRO**. Estadão.com.br, São Paulo, 16 de fevereiro de 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%20

de%20inadimplentes%20no,rela%C3%A7%C3%A3o%20a%20dezembro%20de%202022. >. Acesso em: 19 de março de 2023.
[16] **MAPA DA INADEIMPLÊNCIA E NEGOCIAÇÃO DE DÍVIDAS NO BRASIL**. Serasa, São Paulo, janeiro de 2023. Disponível em: <https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/>. Acesso em: 19 de março de 2023.
[17] LIMA, Rafaela Somavila. **CRIAÇÃO DE PROJETO DE CIÊNCIA DE DADOS UTILIZANDO A METODOLOGIA CRISP-DM EM CONFORMIDADE COM A LGPD**. Tese (Especialização em Ciência de Dados e Suas Aplicações) – Departamento Acadêmico de Informática, Universidade Tecnológica Federal do Paraná. p.35, 2021. Acesso em: 21 de março de 2023.
[18] LEVADA, Alexandre Luis Magalhães. **PROGRAMAÇÃO CIENTÍFICA COM PYTHON**. Departamento de Computação, Centro de Ciências Exatas e Tecnologia, Universidade Federal de São Carlos. p.107, 2021. Acesso em: 22 de março de 2023.
[19] TCHILIAN, Felipe. **CICLO DE CRÉDITO: ENTENDA E OTIMIZE A JORNADA DO CLIENTE**. ClearSale, 2022. Disponível em: <https://blogbr.clear.sale/ciclo-de-credito>. Acesso em 22 de março de 2023.
[20] BRUCE, Peter; BRUCE Andrew. **PRACTICAL STATISTICS FOR DATA SCIENTISTS**. O'Really, 2017. Disponível em: < https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/>. Acesso em: 27 de abril de 2023.
[21] **MÉTODOS DE REAMOSTRAGEM**. Laboratório de Estatística e Geoinformação, Universidade Federal do Paraná. Disponível em: <http://cursos.leg.ufpr.br/ML4all/apoio/reamostragem.html>. Acesso em: 22 de agosto de 2023.
[22] PARK, Sung. **UNDERSTAND AND USE A BUSINESS CREDIT RISK SCORE**. Experian, 2020. Disponível em: <https://blogbr.clear.sale/ciclo-de-credito>. Acesso em 15 de julho de 2023.
[23] Bhalla, Deepanshu. **SAS: CALCULATING KS STATISTICS**. Listen DATA, 2016. Disponível em: < https://www.listendata.com/2016/01/sas-calculating-ks-test.html>. Acesso em 22 de julho de 2023.
[24] Kumar, Ajitesh. **MACHINE LEARNING: INFERENCE & PREDICTION DIFFERENCE**. Data Analytics, 2022. Disponível em: < https://vitalflux.com/machine-learning-inference-prediction-difference/#:~:text=Prediction%20is%20the%20process%20of,the%20predictor%20and%20response%20variables.>. Acesso em 23 de agosto de 2023.
[25] PEREIRA, Pedro Miguel Pinhal. **ANÁLISE DE RISCO DE CRÉDITO USANDO ALGORITMOS DE MACHINE LEARNING**. Tese (Mestrado em Matemática Financeira) – Departamento de Matemática, Universidade de Lisboa. 2020. Acesso em: 23 de agosto 2023.
[26] ABREU, Mariana da Conceição Ferreira. **MODELOS DE AVALIAÇÃO DE RISCO DE CRÉDITO**. Tese (Mestrado em Economia na especialização de Economia Financeira) – Universidade de Coimbra. 2020. Acesso em: 23 de setembro 2023.
[27] SILVA, Daniel de Oliveira Silva. **OTIMIZAÇÃO DE HIPERPARÂMETROS DE ALGORITMOS DE MACHINE LEARNING APLICADO NO CONTEXTO DE ANÁLISE DE RISCO DE CRÉDITO**. Trabalho (Especialização em Ciência de Dados) –Universidade Tecnológica Federal do Paraná. 2022. Acesso em: 14 de setembro 2023.
[28] OLIVEIRA LIMA, Jorge Cláudio Cavalcante de. **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 de agosto 2023.
[29] 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. 2012. Acesso em: 30 de agosto 2023.
[30] BORIN, Edson. **CAPACITAÇÃO PROFISSIONAL EM TECNOLOGIAS DE INTELIGÊNCIA ARTIFICIAL**. Instituto de Computação, Universidade Estadual de Campinas. 2023. Acesso em: 30 de agosto 2023.
[31] **ESTATÍSTICAS MONETÁRIAS E DE CRÉDITO**. Banco Central do Brasil, São Paulo. Disponível em: <https://www.bcb.gov.br/estatisticas/estatisticasmonetariascredito>. Acesso em: 17 de dezembro de 2023.
[32] Mora, Mônica. **A EVOLUÇÃO DO CRÉDITO NO BRASIL ENTRE 2003 E 2010**. Instituto de Pesquisa Econômica Aplicada. Rio de Janeiro, 2015. Disponível em: <https://repositorio.ipea.gov.br/bitstream/11058/3537/1/td2022.pdf>. Acesso em: 17 de dezembro de 2023.
[33] Sfeir, Elias. **A RELAÇÃO CRÉDIT0-PIB NO BRASIL: HISTÓRICO E COMPARAÇÃO INTERNACIONAL**. Brasil, 2021. Disponível em: <https://www.linkedin.com/pulse/rela%C3%A7%C3%A3o-cr%C3%A9dito-pib-brasil-hist%C3%B3rico-e-compara%C3%A7%C3%A3o-elias-

sfeir/?originalSubdomain=pt>. Acesso em: 17 de dezembro de 2023.
[34] **MAPA DA INADIMPLÊNCIA E NEGOCIAÇÃO DE DÍVIDAS DO BRASIL**. Serasa Limpa Nome, Brasil, novembro 2023. Disponível em: <https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/>. Acesso em: 17 de dezembro de 2023.
[35] **POLÍTICA DE CRÉDITO: VEJA TUDO O QUE VOCÊ PRECISA SABER SOBRE O ASSUNTO**. Serasa Experian, Brasil, novembro 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 de janeiro de 2024.

[36] **DESCUBRA AQUI O QUE É A POLÍTICA DE CRÉDITO E COMO ELABORAR UMA**. Serasa Experian, Brasil, julho 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 de janeiro de 2024.
[37] **POLÍTICA DE CRÉDITO: O QUE É, QUAIS SÃO AS FASES E A IMPORTÂNCIA PARA O SEU NEGÓCIO**. Boa Vista, Brasil, novembro 2022. Disponível em: <https://www.boavistaservicos.com.br/blog/destaque/o-que-e-politica-de-credito/>. Acesso em: 25 de janeiro de 2024.
[38] Castro, Jane Simões de. **ESTUDO COMPARATIVO ENTRE METODOLOGIAS DE APRENDIZADO DE MÁQUINA E HÍBRIDAS APLICADAS A RISCO DE CRÉDITO**. Tese (Mestrado em Administração) – FECAP, Brasil, 2022. Disponível em: < http://tede.fecap.br:8080/handle/123456789/818>. Acesso em: 31 de janeiro de 2024.
[39] Lukosiunas, Andreza. **APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING EM MODELOS DE ESCORE DE CRÉDITO**. Tese (Mestrado em Economia) – INSPER, Brasil, 2018. Disponível em: <https://repositorio.insper.edu.br/handle/11224/2573>. Acesso em: 31 de janeiro de 2024.
[40] Silva, Danjiel de Oliveira. **OTIMIZAÇÃO DE HIPER-PARÂMETROS DE ALGORITMOS DE MACHINE LEARNING APLICADO NO CONTEXTO DE ANÁLISE DE RISCO DE CRÉDITO**. Trabalho (Especialização em Ciência de Dados) – Universidade Tecnológica Federal do Paraná, Brasil, 2022. Disponível em: < http://repositorio.utfpr.edu.br/jspui/handle/1/31719>. Acesso em: 31 de janeiro de 2024.
[41] 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, Brasil, 2020. Disponível em: <http://repositorio.ufc.br/handle/riufc/61901>. Acesso em: 31 de janeiro de 2024.
[42] Almeida, Gustavo Durães. **MODELAGEM DE RISCO DE CRÉDITO VIA LSTM**. Tese (Mestrado em Estatística) – Universidade de Brasília, Brasil, 2021. Disponível em: <http://repositorio2.unb.br/jspui/handle/10482/42580>. Acesso em: 31 de janeiro de 2024.
[43] **RECEIVER OPERATING CHARACTERISTIC**. Disponível em: < https://en.wikipedia.org/wiki/Receiver_operating_characteristic>. Acesso em: 31 de janeiro de 2024.