ISSN: 2317-0840

Machine Learning: Predizendo Risco de Crédito

Rafael Almeida Pereira Melo [1†], Paulo Henrique Sales Guimarães[2], Marcel Irving Pereira Melo[3]

[1]*Mestrando em Estatística e Experimentação Agropecuária, Departamento de Estatística, Universidade Federal de Lavras (UFLA)*
[2]*Departamento de Estatística, Universidade Federal de Lavras (UFLA)*
[3]*Egresso no curso de Estatística e Experimentação Agropecuária, Departamento de Estatística, Universidade Federal de Lavras (UFLA)*

Resumo: *O propósito deste artigo é empregar técnicas de Machine Learning para desenvolver modelos de classificação aplicados a dados de risco de crédito. Para atingir esse objetivo, utilizou-se o software R, implementando diversas metodologias, incluindo Regressão Logística, Bagging, Random Forest e Árvores de Decisão. Além disso, as técnicas de ensemble foram aplicadas para combinar esses modelos, buscando melhorar a precisão das previsões. A métrica escolhida para a avaliação comparativa dos modelos foi a Acurácia e AUC (área abaixo da curva). Este estudo busca explorar a eficácia dessas abordagens em prever e classificar riscos de crédito, contribuindo para uma compreensão mais aprofundada e uma aplicação prática de técnicas de aprendizado de máquina no contexto financeiro.*

Palavras-chave: *Aprendizado de Máquina; Risco de crédito; Mercado Financeiro; Inadimplência; Modelos de Classificação.*

Machine Learning: Predicting Credit Risk

Abstract: *The purpose of this article is to employ Machine Learning techniques to develop classification models applied to credit risk data. To achieve this goal, the R software was utilized, implementing various methodologies, including Logistic Regression, Bagging, Random Forest, and Decision Trees. Additionally, ensemble techniques were applied to combine these models, aiming to improve the accuracy of the predictions. The metric chosen for the comparative evaluation of the models was Accuracy and AUC (area under the curve). This study seeks to explore the effectiveness of these approaches in predicting and classifying credit risks, contributing to a deeper understanding and practical application of machine learning techniques in the financial context.*

Keywords: *Machine Learning; Credit Risk; Financial Market; Default; Classification Models.*

[†]Autor correspondente: ralmeidamelo23@gmail.com

Manuscrito recebido em: 07/06/2024
Manuscrito revisado em: 01/10/2024
Manuscrito aceito em: 03/10/2024

Introdução

Em um mundo no qual o crédito se tornou essencial para o desenvolvimento individual e empresarial, a previsão precisa do risco de inadimplência assume um papel crucial para a estabilidade do sistema financeiro. A pandemia de COVID-19 trouxe consigo um aumento significativo na inadimplência. Esse cenário levou os bancos a adotarem uma postura mais cautelosa em relação ao risco, resultando em restrições de liquidez (BENTO, 2023). Diante desse cenário desafiador, torna-se urgente a busca por ferramentas mais precisas e eficientes para prever o risco de crédito e garantir a saúde do sistema financeiro.

As instituições financeiras se deparam com um enorme volume de dados heterogêneos, tornando a análise manual e a identificação de padrões complexos extremamente desafiadoras. A previsão precisa do risco de crédito exige a consideração de uma multiplicidade de fatores inter-relacionados, como histórico de crédito, renda, comportamento de pagamento e características socioeconômicas. Os métodos tradicionais de análise de risco, muitas vezes baseados em regras rígidas e modelos estatísticos simples, não se mostraram suficientemente eficazes para capturar a complexa dinâmica do risco de crédito no cenário atual.

Uma abordagem adequada para fazer previsões inclui os métodos conhecidos como Machine Learning. Os modelos de regressão e classificação baseados em Machine Learning podem lidar com os problemas mencionados dos modelos tradicionais e são poderosas ferramentas quando se busca alta acurácia (LOPES, 2018).

O Machine Learning oferece uma abordagem inovadora e poderosa para análise de risco de crédito, permitindo a identificação de padrões complexos e sutis nos dados que passariam despercebidos por métodos tradicionais. Modelos de Machine Learning são flexíveis e adaptáveis, capazes de aprender com novos dados e se ajustar a mudanças nas condições do mercado, garantindo a precisão das previsões ao longo do tempo. Além disso, o Machine Learning se destaca pela capacidade de processar grandes volumes de dados heterogêneos, tornando-o ideal para lidar com a avalanche de informações geradas pelas instituições financeiras. Essa característica permite a utilização de conjuntos de dados mais abrangentes e ricos em informações, o que contribui para a maior precisão e confiabilidade das previsões de risco de crédito (SHI et al, 2022).

Neste estudo, exploramos o potencial do Machine Learning para prever o risco de crédito, utilizando diferentes técnicas e modelos para analisar um conjunto de dados abrangente de solicitações de empréstimo.

Para avaliar a probabilidade de inadimplência, diversos dados são considerados, como idade, renda, histórico de pagamentos e fatores econômicos. Com a crescente complexidade do setor financeiro, métodos avançados tornam-se necessários para prever e administrar o risco de crédito. Técnicas de *Machine Learning* são ferramentas poderosas para analisar grandes conjuntos de dados e identificar padrões complexos.

Objetivos

Este trabalho visa explorar e avaliar a eficácia de diferentes técnicas de *Machine Learning* na modelagem e classificação de riscos de crédito. Serão utilizadas metodologias como Regressão Logística, *Bagging*, *Random Forest* e Árvores de Decisão, além de técnicas de ensemble, como o *majority voting*, que combinam esses modelos para melhorar a precisão das previsões, usando o software R. Os objetivos específicos incluem: realizar uma análise exploratória dos dados de risco de crédito, implementar modelos de *Machine Learning* e o *majority voting*, e comparar as metodologias empregadas.

Metodologia

Os dados analisados no artigo foram obtidos do *Kaggle*[ii], uma plataforma que oferece uma ampla variedade de conjuntos de dados abrangendo diversos temas.

Antes da análise, os dados foram pré-processados para garantir sua qualidade e confiabilidade. Valores ausentes foram imputados utilizando a técnica de média por grupo, considerando o grupo de referência mais adequado para cada variável. A escolha dessa técnica se deu pela sua robustez à presença de outliers e pela capacidade de preservar as características distributivas dos dados. No entanto, é importante reconhecer que a imputação por média pode introduzir um viés nos dados, especialmente em casos com alta proporção de valores ausentes.

Além disso, os dados foram padronizados utilizando a técnica de normalização Z-score, que converte cada variável para uma distribuição normal com média zero e desvio padrão unitário. A escolha dessa técnica se deu pela sua simplicidade e pela capacidade de preservar as relações entre as variáveis. A padronização dos dados foi importante para melhorar a convergência do algoritmo de aprendizado de máquina utilizado e facilitar a comparação entre as variáveis, que possuem escalas diferentes.

O conjunto de dados inicial continha 12 variáveis. Após análise exploratória, verificou-se que 4 dessas variáveis eram altamente correlacionadas e não contribuíam significativamente para o poder preditivo do modelo. As variáveis removidas foram: propriedade da casa, intenção do empréstimo, classificação do empréstimo e status do empréstimo. A remoção dessas variáveis redundantes reduziu a multicolinearidade no modelo, facilitou a interpretação dos resultados e contribuiu para a melhora do desempenho do modelo de aprendizado de máquina.

O conjunto de dados utilizado consiste em 32.581 observações e 8 variáveis, fornecendo informações detalhadas sobre solicitações de empréstimos e características pessoais dos requerentes. As variáveis são: **idade, renda, tempo de emprego, valor do empréstimo financiado, taxa de juros, endividamento, tempo de relacacionamento, cb person default on file (que indica se a pessoa é ou não inadimplente)**.

Para análise, foi utilizado o software R, no qual foi aplicado aos dados as metologias de Regressão Logística, Bagging, Random Forest e Árvores de Decisão.

Os dados foram divididos em 75% para treino e 25% para teste. Essa divisão é comum no aprendizado de máquina e fornece um tamanho de treino adequado para o modelo aprender os padrões dos dados, enquanto reserva uma quantidade suficiente de dados para avaliar seu desempenho em um conjunto independente.

No entanto, é importante ressaltar que a proporção ideal entre dados de treino e teste pode variar de acordo com o tamanho do conjunto de dados original, a complexidade do modelo e a sensibilidade desejada na avaliação. Em casos de conjuntos de dados menores, proporções alternativas como 60%/40% ou 80%/20% podem ser mais adequadas. Técnicas como validação cruzada também podem ser utilizadas para avaliar o desempenho do modelo em diferentes divisões de dados, fornecendo uma estimativa mais precisa da generalização do modelo.

Para avaliação dos modelos, foram utilizadas as métricas: Acurácia e Curva ROC (Área abaixo da curva).

Regressão Logística

Considere novamente o conjunto de dados de risco de crédito, onde a resposta inadimplência cai em uma das duas categorias, Sim ou Não. Em vez de modelar essa resposta Y diretamente, a regressão logística modela a probabilidade de que Y pertença a uma categoria específica.

[ii] `https://www.kaggle.com/datasets/laotse/credit-risk-dataset`

Para os dados de risco de crédito, a regressão logística modela a probabilidade de inadimplência utilizando todas as variáveis disponíveis. Por exemplo, a probabilidade de inadimplência dada a dívida pode ser escrita como

$$\begin{aligned}&\text{Pr(inadimplência = Sim—idade, renda, tempo de emprego,}\\&\text{valor do empréstimo financiado, taxa de juros, endividamento,}\\&\text{tempo de relacionamento).}\end{aligned} \tag{1}$$

Os valores de Pr(inadimplência = Sim—variáveis), que abreviamos como p(variáveis), variarão entre 0 e 1. Então, para qualquer conjunto dado de valores das variáveis, uma previsão pode ser feita para inadimplência. Por exemplo, pode-se prever inadimplência = Sim para qualquer indivíduo para quem p(variáveis) > 0.5. Este ponto de decisão é chamado de cutoff. No caso específico deste estudo, foi utilizado um cutoff de 23%, ou seja, previu-se inadimplência = Sim para qualquer indivíduo para quem p(variáveis) > 0.23. Esse ponto de decisão, ou cutoff, é escolhido para balancear a sensibilidade e especificidade do modelo de acordo com os objetivos do estudo.

Alternativamente, se uma empresa deseja ser conservadora na previsão de indivíduos que estão em risco de inadimplência, pode optar por usar um limiar ainda mais baixo.

Modelo Logístico

Para modelar $p(X)$, a probabilidade de um evento ocorrer dado um conjunto de variáveis explicativas X, utilizamos uma função que produza resultados entre 0 e 1 para todos os valores de X (JAMES, 2023). Na regressão logística, essa função é a função logística:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \tag{2}$$

Na equação em (2), β_0 e β_1 são os coeficientes do modelo, representando o intercepto e a inclinação da relação entre as variáveis explicativas e a probabilidade do evento de interesse, como a inadimplência. Esses coeficientes são estimados a partir dos dados de treinamento usando o método de máxima verossimilhança, permitindo que o modelo se ajuste aos dados e faça previsões sobre a probabilidade de inadimplência com base nas características dos clientes.

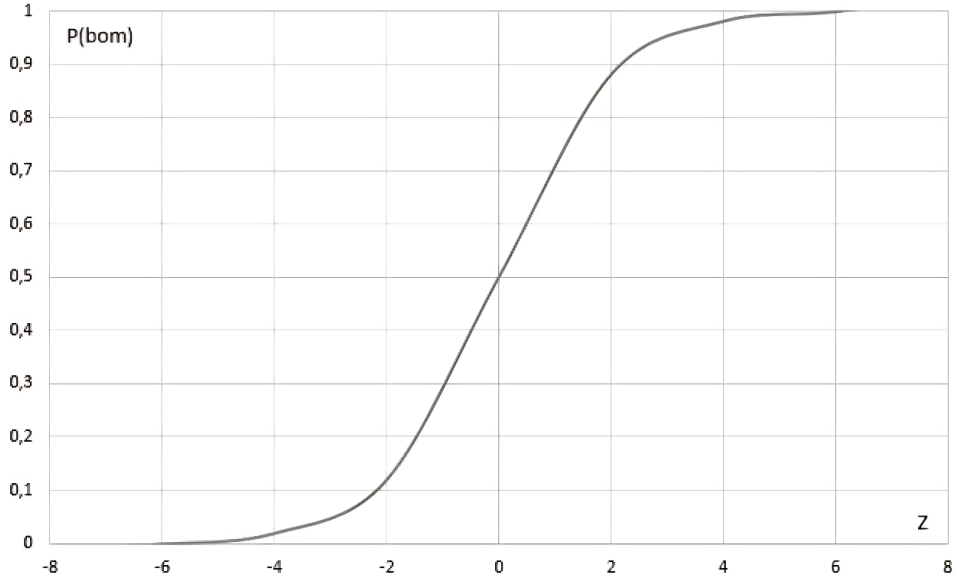
A função logística gera uma curva em forma de S, o que significa que, independentemente do valor de X, obteremos uma previsão sensata. Para valores baixos das variáveis, a previsão da probabilidade de inadimplência se aproxima de zero, mas nunca é inferior a zero. Da mesma forma, para valores altos das variáveis, a previsão se aproxima de um, mas nunca ultrapassa um.

Regressão Logística com Stepwise

A Regressão Logística com *Stepwise* para modelos de classificação é uma abordagem que visa selecionar as variáveis mais relevantes para prever a classe de um conjunto de dados. Essa técnica é semelhante à regressão logística com *stepwise* para modelos de regressão (Figura 1), mas é aplicada em problemas de classificação, onde a variável dependente é categórica.

No método *stepwise forward*, o algoritmo começa com um modelo vazio e, a cada passo, adiciona a variável que mais melhora a capacidade de classificação do modelo. Isso é feito com base em critérios como o valor-p, o AIC (Akaike Information Criterion) ou o BIC (Bayesian Information Criterion). O processo continua até que não haja mais variáveis que possam melhorar o modelo.

Figure 1: Logistics Curve.



Source: Mourão (2022).

No método *stepwise backward*, todas as variáveis são incluídas inicialmente no modelo e, a cada passo, a variável que menos contribui para a classificação é removida. O processo continua até que a remoção de qualquer variável piore significativamente a capacidade de classificação do modelo.

Imagine que, ao aplicar a regressão logística *stepwise forward* para prever a inadimplência de clientes, o algoritmo selecione as variáveis "renda", "idade"e "histórico de crédito". Isso significa que essas variáveis são as mais relevantes para prever a inadimplência nesse conjunto de dados. As instituições financeiras podem utilizar essa informação para direcionar seus recursos de forma mais eficiente, focando em clientes com menor risco de inadimplência.

A regressão logística *stepwise* para modelos de classificação é útil para automatizar a seleção de variáveis em problemas de classificação, tornando o processo mais objetivo e menos suscetível a viéses. No entanto, assim como na regressão logística *stepwise* para modelos de regressão, é importante usar essa técnica com cautela e considerar a validade dos critérios de seleção de variáveis utilizados.

Árvores de Decisão para classificação

Uma árvore de classificação é muito semelhante a uma árvore de regressão, exceto que ela é usada para prever uma resposta qualitativa em vez de quantitativa. Para uma árvore de regressão, a resposta prevista para uma observação é a média das respostas das observações de treinamento que pertencem ao mesmo nó terminal. Em contraste, para uma árvore de classificação, prevemos que cada observação pertença à classe mais comum das observações de treinamento na região à qual ela pertence. Na interpretação dos resultados de uma árvore de classificação, estamos interessados não apenas na previsão da classe correspondente a um nó terminal específico, mas também nas proporções de classe entre as observações de treinamento que caem nessa região. (JAMES et al, 2023).

A tarefa de desenvolver uma árvore de classificação é semelhante à tarefa de desenvolver uma árvore de regressão. Assim como no caso da regressão, foi utilizado a divisão binária recursiva para desenvolver uma árvore de classificação. No entanto, no caso da classificação, o RSS (soma dos quadrados dos resíduos) não pode ser usado como critério para fazer as divisões

binárias. Uma alternativa natural ao RSS é a taxa de erro de classificação. Como planejamos atribuir uma observação em uma determinada região à classe mais comum das observações de treinamento nessa região, a taxa de erro de classificação é simplesmente a fração das observações de treinamento nessa região que não pertencem à classe mais comum:

$$E = 1 - \max_k(\hat{p}_{mk}) \tag{3}$$

Índice de Gini e Entropia em Árvores de Classificação

O índice de Gini e a entropia são medidas utilizadas na construção de árvores de classificação para avaliar a qualidade de uma divisão em um nó. A classificação baseada em erro não é tão sensível quanto essas medidas, por isso são preferíveis na prática.

O índice de Gini é uma medida da impureza do nó e é definido como:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4}$$

em que $\hat{p}_{mk}$ é a proporção de observações de treinamento na região m que são da classe k. Um valor baixo de Gini indica que um nó contém predominantemente observações de uma única classe.

A entropia é outra medida de impureza, dada por:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk}) \tag{5}$$

Assim como o índice de Gini, a entropia é próxima de zero se as proporções $\hat{p}_{mk}$ são todas próximas de zero ou um, o que indica um nó puro.

Tanto o índice de Gini quanto a entropia são mais sensíveis à pureza do nó do que a taxa de erro de classificação, sendo geralmente preferidos na construção da árvore. No entanto, a taxa de erro de classificação pode ser mais adequada durante a poda da árvore se a precisão da previsão da árvore podada final for o objetivo.

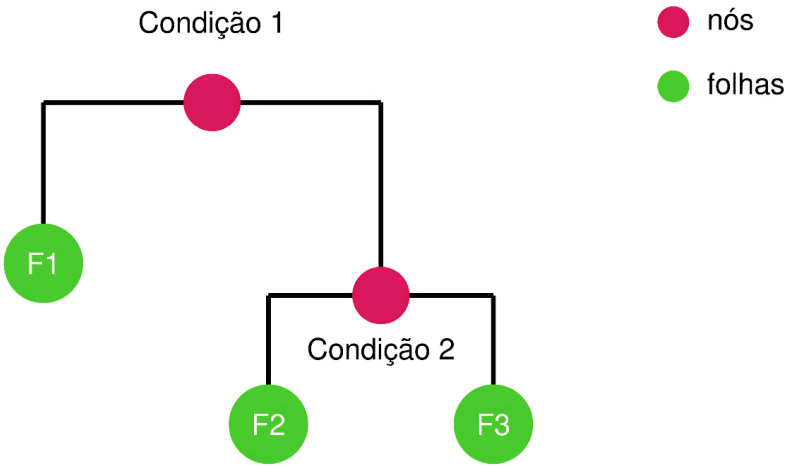
Divisão de Variáveis Qualitativas em Árvores de Classificação

Em árvores de classificação, é possível dividir nós com base em variáveis qualitativas. Um nó pode ser dividido atribuindo alguns dos valores qualitativos a um ramo e os restantes a outro ramo.

Uma característica importante das árvores de classificação é a pureza dos nós. A divisão em um nó é feita para aumentar a pureza, ou seja, para garantir que um nó contenha predominantemente observações de uma única classe. Isso é importante porque nos permite ter mais confiança nas previsões feitas pelas árvores de classificação.

Para prever uma nova observação utilizando uma árvore de decisão, o processo é o seguinte: começamos pela raiz e verificamos se a condição no nó atual é atendida. Se sim, seguimos para o nó filho à esquerda; caso contrário, seguimos para o nó filho à direita (Figura 2). Este processo é repetido até alcançarmos uma folha da árvore.

Figure 2: Decision Tree Example.



Source: Izbicki *et al.* (2020).

Imagine que queremos construir uma árvore de decisão para prever a inadimplência de clientes, utilizando as variáveis "renda", "idade"e "histórico de crédito". A árvore pode começar dividindo os clientes em dois grupos: aqueles com renda acima de um determinado valor e aqueles com renda abaixo desse valor. Em seguida, cada grupo pode ser dividido novamente com base em outros critérios, como idade ou histórico de crédito. O objetivo é construir uma árvore que divida os clientes em grupos cada vez mais homogêneos em termos de risco de inadimplência.

Bagging

O *bootstrap* é uma ideia poderosa utilizada em muitas situações onde é difícil calcular o desvio padrão diretamente. No contexto de métodos de aprendizado estatístico como árvores de decisão, o bootstrap pode melhorar significativamente a precisão (JAMES et al, 2023)

Árvores de decisão frequentemente sofrem de alta variância. Isso significa que ao dividir os dados de treinamento em duas partes aleatórias e ajustar uma árvore de decisão a ambas, os resultados podem variar consideravelmente. Em contraste, procedimentos com baixa variância produzem resultados consistentes em diferentes conjuntos de dados; por exemplo, a regressão linear tende a ter baixa variância quando a razão n para p é moderadamente grande.

O *bagging* é um procedimento geral para reduzir a variância de um método de aprendizado estatístico. Consiste em obter muitos conjuntos de treinamento da população, construir um modelo de previsão separado usando cada conjunto e, em seguida, média das previsões resultantes. Isso reduz a variância, resultando em um modelo de aprendizado estável. Por exemplo, podemos calcular $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_B(x)$ usando B conjuntos de treinamento separados e média deles para obter um único modelo de baixa variância:

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x). \tag{6}$$

Embora o *bagging* possa melhorar as previsões para muitos métodos de regressão, é especialmente útil para árvores de decisão. No *bagging* de árvores de regressão, construímos B árvores de regressão usando B conjuntos de treinamento bootstrap e médiamos as previsões resultantes. Isso reduz a variância e melhora a precisão da previsão.

Para problemas de classificação, o *bagging* pode ser estendido de forma simples. Registramos a classe prevista por cada uma das B árvores para uma dada observação de teste e fazemos uma votação majoritária para determinar a classe final prevista.

O bagging é uma técnica eficaz para melhorar a precisão dos modelos de árvores de decisão, tornando-os mais estáveis e robustos.

Imagine que queremos construir um modelo de árvore de decisão para prever a inadimplência de clientes. Podemos utilizar o *bagging* para reduzir a variância do modelo, construindo 100 árvores de decisão a partir de diferentes conjuntos de treinamento bootstrap dos dados de clientes. Em seguida, podemos combinar as previsões das 100 árvores para obter um único modelo com maior precisão e robustez.

Random Forest

O método *Random Forest* representa um avanço em relação às árvores de decisão agregadas (*bagged trees*), introduzindo um ajuste aleatório que torna as árvores mais independentes entre si. Assim como no *bagging*, construímos uma floresta de árvores de decisão a partir de amostras de treinamento *bootstrap*. Porém, ao construir essas árvores de decisão, em cada divisão, selecionamos um subconjunto aleatório de m preditores como candidatos à divisão, escolhendo apenas um desses m preditores. O valor típico de m é em torno de $\sqrt{p}$, onde p é o número total de preditores.

Ao construir uma floresta aleatória, o algoritmo não considera a maioria dos preditores disponíveis em cada divisão na árvore. Se um preditor muito forte estiver presente no conjunto de dados, a maioria ou todas as árvores na coleção agregada utilizarão esse preditor forte na divisão superior. Consequentemente, as previsões das árvores agregadas serão altamente correlacionadas. Infelizmente, a média de muitas quantidades altamente correlacionadas não leva a uma redução tão significativa na variância quanto a média de muitas quantidades não correlacionadas. Em particular, isso significa que o bagging não reduzirá substancialmente a variância em relação a uma única árvore nesse cenário.

A técnica *Random Forest* supera esse problema ao forçar cada divisão a considerar apenas um subconjunto dos preditores. Assim, em média, aproximadamente $(p - m)/p$ das divisões não considerarão o preditor forte, permitindo que outros preditores tenham mais influência. Podemos pensar nesse processo como decorrelacionando as árvores, tornando a média das árvores resultantes menos variável e, portanto, mais confiável. A principal diferença entre *bagging* e florestas aleatórias é a escolha do tamanho do subconjunto de preditores m. Por exemplo, se uma floresta aleatória for construída usando $m = p$, então ela se reduz a simplesmente *bagging*. As florestas aleatórias usando $m \approx \sqrt{p}$ levam a uma redução no erro de teste em relação ao bagging.

O uso de um pequeno valor de m na construção de uma floresta aleatória geralmente é útil quando temos um grande número de preditores correlacionados. Podemos aplicar florestas aleatórias a conjuntos de dados de alta dimensão, como aqueles com muitos genes expressos em diferentes condições biológicas, onde a correlação entre os genes pode ser significativa. A aplicação cuidadosa dessas técnicas pode resultar em melhorias significativas na capacidade de prever resultados complexos com base em grandes conjuntos de dados (JAMES et al, 2023).

Imagine que queremos construir um modelo de Floresta Aleatória para prever a inadimplência de clientes. Podemos utilizar a técnica para reduzir a correlação entre as árvores e aumentar a confiabilidade das previsões, selecionando aleatoriamente um subconjunto de variáveis, como renda, idade e histórico de crédito, em cada divisão da árvore. Essa abordagem pode levar a um modelo mais preciso e robusto para prever a inadimplência de clientes.

Ensemble de Modelos - Majority Voting

Uma técnica amplamente utilizada para melhorar a precisão preditiva em problemas de classificação é o ensemble de modelos, que combina diferentes algoritmos para gerar uma predição mais robusta. No presente estudo, foi aplicada a técnica de Majority Voting, na qual

as predições de múltiplos modelos são combinadas para formar a predição final. Essa abordagem visa melhorar a capacidade preditiva ao aproveitar as características distintas de cada modelo.

O processo de Majority Voting é realizado ao reunir as predições binárias (0 ou 1) de cada um desses modelos. Para cada instância, a predição final é determinada pela "votação da maioria", ou seja, a classe que é mais frequentemente predita pelos modelos participantes é escolhida como a predição final. Formalmente, isso significa que se mais da metade dos modelos classificarem uma instância como 1 *(default)*, essa será a predição final. Caso contrário, será classificada como 0 *(não-default)*.

A principal vantagem dessa abordagem é que o ensemble pode reduzir o risco de overfitting de um modelo individual, uma vez que os erros específicos de cada modelo tendem a ser compensados por outros modelos. O ensemble captura diferentes padrões presentes nos dados que cada modelo isolado pode não detectar, tornando as predições mais robustas e generalizáveis.

No estudo, o ensemble foi implementado utilizando as predições de quatro modelos: regressão logística, bagging, random forest e árvore de decisão. As probabilidades preditivas de cada modelo foram obtidas, e o corte de 0,23 foi escolhido para a regressão logística, conforme ajustado anteriormente, para otimizar o equilíbrio entre sensibilidade e especificidade. A implementação do ensemble resultou em uma melhora na acurácia geral do modelo, refletindo a robustez da técnica ao combinar os diferentes pontos fortes de cada algoritmo.

Acurácia

A acurácia é uma métrica que avalia o quão preciso um modelo é em relação aos dados em geral (COELHO, 2021). Ela é obtida pela divisão do número de unidades corretamente classificadas, verdadeiros positivos (TP) e verdadeiros negativos (TN), pelo número total de previsões feitas pelo classificador, levando em consideração os falsos positivos (FP) e falsos negativos (FN).

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Acurácia Balanceada

A acurácia balanceada considera o desempenho do modelo em cada classe individualmente e calcula a média entre elas. Isso é particularmente importante em situações de desequilíbrio de classes, como neste estudo, onde o número de adimplentes é significativamente maior que o de inadimplentes. Em casos de desbalanceamento de dados, a acurácia simples pode ser enganosa, já que o modelo tende a favorecer a classe majoritária (adimplentes), resultando em uma avaliação distorcida do desempenho. A acurácia balanceada, por outro lado, oferece uma medida mais equitativa, avaliando a performance em ambas as classes de forma justa e proporcionando uma visão mais completa da capacidade preditiva dos modelos utilizados (BRODERSEN *et al.*, 2010).

Curva ROC

A curva ROC é uma abordagem gráfica na qual o critério de avaliação obedece às faixas de avaliação de desempenho do modelo, dadas pelo valor da área abaixo da curva (AUC), sendo:

- Valores menores ou iguais a 0,5 não há discriminação no modelo;
- Valores entre 0,5 e 0,7 têm baixa discriminação;
- Valores entre 0,7 e 0,8 têm discriminação aceitável;

- Valores entre 0,8 e 0,9 a discriminação é excelente; e
- Valores acima de 0,9 a discriminação é excepcional (Favero, 2009, citado em Bordin, 2022).

Resultados

Foram obtidos os seguintes resultados de acurácia e AUC dos modelos utilizados (Tabela 1):

Table 1: Accuracy, Balanced Accuracy and AUC Results of the Models Used.

Método	Acurácia	Acurácia Balanceada	AUC
Regressão Logística	0.8201	0.7986	0.8623
Bagging	0.8289	0.7018	0.8868
Random Forest	0.8254	0.6947	0.8802
Árvores de Decisão	0.8254	0.6973	0.8331
Ensemble Majority Voting	0.8243	0.7653	0.8841

Source: from the authors (2024).

Dentre os modelos testados, o *bagging* apresentou a melhor performance em termos de AUC, com um valor de 0.8868, indicando excelente discriminação entre as classes de adimplentes e inadimplentes. No entanto, ao observar a acurácia balanceada de 0.7018, percebemos que o *bagging* teve dificuldades em lidar com o desbalanceamento do conjunto de dados, sendo menos eficaz na previsão da classe minoritária (inadimplentes), o que é evidenciado pela diferença significativa entre a acurácia simples (0.8289) e a balanceada.

A acurácia balanceada oferece uma visão mais realista do desempenho do modelo em situações desbalanceadas. Modelos como o *random forest* e a árvore de decisão apresentaram valores de acurácia balanceada ainda menores (0.6947 e 0.6973, respectivamente), apesar de terem acurácias globais de 0.8254, sugerindo que esses modelos também favoreceram a classe majoritária (adimplentes).

Por outro lado, a *regressão logística*, embora seja um dos modelos mais tradicionais, demonstrou um desempenho notável. Com uma acurácia de 0.8201 e uma acurácia balanceada de 0.7986, a *regressão logística* conseguiu manter uma boa performance, mesmo lidando com a desproporção entre as classes. Isso indica que, apesar de sua simplicidade, a *regressão logística* pode ser uma escolha eficaz em cenários de classificação, sendo capaz de capturar relações significativas entre as variáveis preditivas e a variável resposta.

Além disso, a capacidade da *regressão logística* de fornecer probabilidades associadas a cada classe facilita a interpretação dos resultados, o que a torna uma ferramenta valiosa para análises de risco de crédito. Essa eficiência e interpretabilidade reforçam a ideia de que modelos mais simples e tradicionais podem ser tão eficazes quanto abordagens mais complexas, especialmente quando combinados com métricas adequadas de avaliação, como a acurácia balanceada.

O ensemble com *majority voting* alcançou uma acurácia balanceada de 0.7653, sendo o mais equilibrado entre as classes, embora sua acurácia simples (0.8243) não tenha sido a mais alta. Isso indica que o ensemble conseguiu um desempenho mais robusto, garantindo maior equidade na classificação das duas classes, o que o torna uma escolha mais apropriada para esse problema desbalanceado.

Portanto, embora o *bagging* tenha a melhor AUC, a *regressão logística* se destaca por sua eficiência e interpretabilidade, enquanto o ensemble de *majority voting* oferece uma abordagem mais robusta para lidar com o desbalanceamento dos dados. Essa análise sugere que a acurácia balanceada, em conjunto com outras métricas, oferece uma avaliação mais justa do desempenho do modelo em contextos de dados desbalanceados e deveria ser considerada para a seleção do modelo mais adequado.

Conclusão

Em geral, modelos de Machine Learning oferecem uma abordagem promissora e precisa para prever a inadimplência, fornecendo insights valiosos para instituições financeiras na gestão de riscos e na tomada de decisões. Os resultados do estudo demonstram o potencial do Machine Learning como uma ferramenta poderosa para prever o risco de crédito, onde a Regresão Logística, apesar de ser um modelo mais tradicional, apresentou uma boa performance com uma acurácia balanceada de 0.7986, evidenciando sua eficácia na identificação de inadimplentes, mesmo em um cenário desbalanceado.

Embora o *bagging* tenha apresentado a melhor AUC e acurácia geral, a acurácia balanceada se destacou como uma métrica crucial para avaliar o desempenho dos modelos, especialmente em conjuntos de dados desbalanceados. A acurácia balanceada permite uma visão mais justa da capacidade preditiva, assegurando que ambas as classes (adimplentes e inadimplentes) sejam consideradas na avaliação do modelo.

A capacidade de identificar com maior precisão clientes com maior probabilidade de inadimplência pode auxiliar na tomada de decisões mais assertivas na concessão de crédito e na mitigação de riscos. Assim, o estudo não apenas valida o uso de técnicas de Machine Learning, mas também destaca a importância de considerar a acurácia balanceada como uma métrica fundamental para otimizar as operações das instituições financeiras e reduzir perdas.

Agradecimentos

Este trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Brasil. Agradecemos pelo financiamento concedido através da bolsa de mestrado.

Referências

COELHO, F. F.; AMORIM, D. P. de L.; CAMARGOS, M. A. de. Analisando métodos de machine learning e avaliação do risco de crédito. *Revista Gestão & Tecnologia*, v. 21, n. 1, p. 89–116, 2021.

MOURÃO, V. D. G.; CAJUEIRO, D. O. Estudo comparativo entre técnicas de machine learning para classificação do tomador PJ – MPE (Micro e Pequenas Empresas) [Dissertação de Mestrado, Universidade de Brasília].

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística* [livro eletrônico]. São Carlos, SP: Rafael Izbicki, 2020.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer, 2023. ISBN 978-3-030-70396-1.

LOPES, L. P. Predicting the price of Brazilian Natural coffee using statistical machine learning models. *Sigmae*, v. 7, n. 1, p. 1–16, 2018. Disponível em: `https://publicacoes.unifal-mg.edu.br/revistas/index.php/sigmae/article/view/699`.

BENTO, C. E. A. Inadimplência em tempos de pandemia: uma análise do crédito imobiliário. Dissertação de mestrado profissional, Fundação Getulio Vargas, Escola de Políticas Públicas e Governo, 2023.

BORDIN, I. T.; ROSSONI, D. F. Fatores associados à percepção de prejuízo na aprendizagem de estudantes de universidades públicas brasileiras durante a pandemia de Covid-19. *Sigmae, Programa de Pós-Graduação em Bioestatística, Universidade Estadual de Maringá (UEM)*, 2022.

SHI, S.; TSE, R.; LUO, W.; D'ADDONA, S.; PAU, G. Machine learning-driven credit risk: a systemic review, 2022.

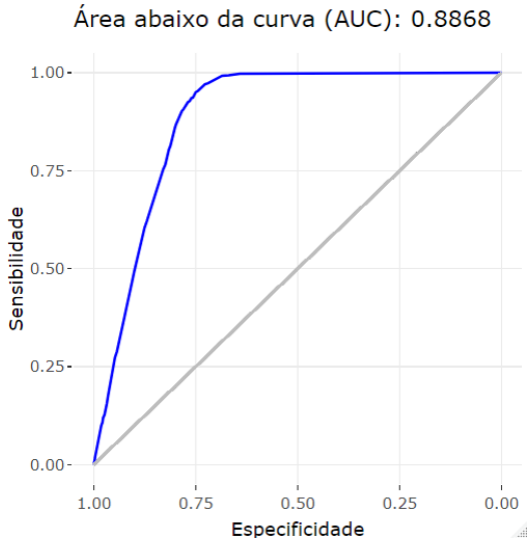
BRODERSEN, K. H.; ONG, C. S.; STEPHAN, K. E.; BUHMANN, J. M. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, p. 3121-3124, IEEE, 2010.

Apêndice

Os dados analisados neste artigo, bem como o script utilizado para obter os resultados, estão disponíveis no repositório GitHub: https://github.com/meloRAP/creditRisk. Este repositório contém todos os detalhes necessários para a replicação dos resultados, incluindo o pré-processamento dos dados, a implementação dos modelos de machine learning, e o código utilizado para gerar as tabelas e figuras apresentadas no artigo.

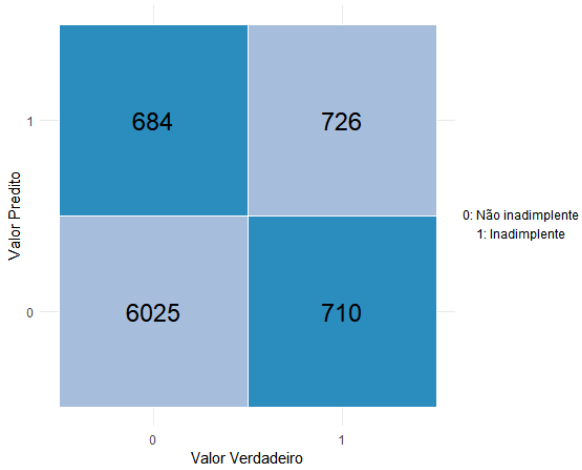
Também são apresentados resultados complementares do desempenho do modelo de Bagging por meio da curva ROC e da matriz de confusão, ilustradas nas Figuras 4 e 5, respectivamente.

Figure 4: ROC curve - Bagging.



Source: from the authors (2024).

Figura 5: Confusion Matrix - Bagging.



Source: from the authors (2024).