Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

Henrique Valaski de Mello
Willian Meira Schlichta

Análise Comparativa de Modelos de Scoring

Curitiba
2024

Henrique Valaski de Mello
Willian Meira Schlichta

Análise Comparativa de Modelos de Scoring

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Graduação em Estatística da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientador(a): Fernando Lucambio Pérez

Curitiba
2024

Dedicamos este trabalho à Gabriel Borges Schlichta e a família Valaski de Mello.

Agradecimentos

Chegar ao fim deste Trabalho de Conclusão de Curso (TCC) em Estatística foi uma jornada e tanto, e temos muitas pessoas a agradecer por todo o apoio e ajuda ao longo do caminho.

Primeiro, agradecemos a Deus por nos dar saúde, força e determinação para enfrentar os desafios dessa caminhada acadêmica. Sua presença foi essencial para superarmos os momentos difíceis.

Às nossas famílias, que sempre estiveram ao nosso lado, oferecendo suporte emocional e muito incentivo. Um agradecimento especial aos nossos pais, que confiaram em nós e nos motivaram constantemente. Também não podemos esquecer dos nossos amigos, que entenderam nossas ausências e nos apoiaram quando mais precisávamos.

Queremos agradecer especialmente ao nosso orientador, Professor Doutor Fernando Lucambio Perez, pela dedicação, paciência e orientação. Suas sugestões e críticas construtivas foram essenciais para o desenvolvimento deste trabalho.

Aos nossos colegas de curso, com quem compartilhamos essa jornada, dividindo conhecimentos, experiências e muitos momentos inesquecíveis. A amizade e a colaboração de todos vocês foram fundamentais para nosso crescimento acadêmico e pessoal. Também gostaríamos de agradecer aos demais professores do curso de Estatística da UFPR, que nos ensinaram e inspiraram ao longo de toda a graduação. Seus ensinamentos foram cruciais para a realização deste TCC.

Agradecemos à Universidade Federal do Paraná (UFPR) e ao Departamento de Estatística pela estrutura e recursos que foram essenciais para a realização deste trabalho. Também agradecemos aos funcionários administrativos e bibliotecários pela eficiência e disposição em ajudar sempre que precisávamos.

Aos participantes da pesquisa e às instituições que forneceram os dados utilizados neste estudo, muito obrigado pela colaboração e confiança. Sem vocês, este trabalho não teria sido possível.

Por fim, agradecemos a todos que, de alguma forma, contribuíram para a realização deste TCC. A cada um de vocês, nossa mais profunda gratidão.

Henrique Valaski de Mello e Willian Meira Schlichta

"Sem dados, você é apenas mais uma pessoa com uma opinião."
W. Edwards Deming.

Resumo

Este estudo analisou a eficácia de três métodos distintos de modelagem estatística — regressão logística, *XGBoost* e *Random Forest* — em um conjunto real de dados de crédito para determinar qual técnica apresenta o melhor desempenho na previsão de inadimplência. A metodologia aplicada consistiu em uma análise exploratória das variáveis, uma fase de seleção de variáveis, e a subsequente aplicação dos modelos estatísticos. Esta abordagem permitiu uma compreensão profunda tanto do comportamento da base de dados quanto das variáveis que influenciam a inadimplência neste caso.

Os resultados indicaram que o modelo *XGBoost* superou os demais devido à sua capacidade de manejar grandes volumes de dados e sua eficácia em capturar não-linearidades e complexidades do conjunto de dados. Em contraste, a regressão logística, apesar de ser uma técnica mais tradicional, mostrou robustez e relevância, e o modelo Random Forest, embora menos eficiente que os outros modelos, ainda se mostrou competente.

Este trabalho não apenas foca na aplicação prática dessas técnicas de modelagem no setor financeiro, mas também sublinha a importância de uma análise rigorosa de dados antes da modelagem. A análise e seleção cuidadosa de variáveis são essenciais para desenvolver modelos de previsão de crédito mais precisos e confiáveis. Embora o *XGBoost* tenha apresentado a melhor performance geral, os resultados reforçam a validade de múltiplas abordagens dependendo das necessidades específicas e do contexto de aplicação.

Palavras-chave: Inadimplência de crédito. Regressão logística. XGBoost. Random Forest. Modelagem estatística. Análise de dados. Setor bancário.

Sumário

1 Introdução

Em um mundo cada vez mais regido por dados, a capacidade de avaliar e classificar a probabilidade de um indivíduo ou entidade cumprir suas obrigações financeiras tornou-se um pilar crucial nas decisões financeiras. Esta avaliação, comumente conhecida como "scoring", é a espinha dorsal de muitas instituições financeiras, ditando aprovações de crédito, determinando taxas de juros e influenciando estratégias de gerenciamento de risco. Dada a sua importância, a precisão e robustez desses modelos de score são vitais, não apenas para a saúde financeira das instituições, mas também para a economia em geral.

Historicamente, as técnicas de modelagem de score se baseavam em abordagens estatísticas tradicionais, como evidenciado no artigo "A Survey of Credit and Behavioural Scoring; Forecasting financial risk of lending to consumers" (THOMAS, 2000). No entanto, a paisagem da ciência de dados tem sofrido uma transformação rápida nas últimas décadas. Com a ascensão da aprendizagem de máquina e a disponibilidade de poder computacional sem precedentes, surgiram novas técnicas promissoras que desafiam os métodos tradicionais.

Métodos como regressão linear e regressão logística, citados no artigo, representam o legado de técnicas amplamente empregadas e testadas. No entanto, na era atual da ciência de dados, abordagens como *Random Forest* e *XGBoost* têm emergido como potências, sendo aplicadas em uma variedade de domínios devido à sua capacidade de lidar com grandes volumes de dados, complexidade e não-linearidades.

No entanto, conforme destacado no artigo, surge uma interrogação fundamental: "Se diferentes métodos produzem resultados semelhantes, qual método deve ser preferido?" Esta questão sugere que a escolha de um método não deve ser baseada apenas em métricas de desempenho, mas também em considerações práticas como facilidade de implementação, custo, interpretabilidade e robustez.

O objetivo é não apenas comparar as técnicas tradicionais com as abordagens mais contemporâneas em termos de precisão e desempenho, mas também considerar fatores intrínsecos e extrínsecos que possam influenciar a escolha de um método em detrimento de outro. A relevância deste estudo não se limita apenas à academia, mas também tem implicações práticas significativas para o setor financeiro.

2 Material e Métodos

2.1 Descrição das Variáveis Avaliadas

Neste estudo, foram avaliadas diversas variáveis de uma base de dados reais provenientes de uma instituição financeira, com o objetivo de prever o a *target* para avaliação de crédito. As variáveis analisadas incluem:

- **Safra**: Variável categórica que indica o período de entrada do cliente na base de dados.
- **Idade_Bem**: Variável categórica que representa a idade do veículo financiado.
- **Perc_prestacao**: Variável numérica que indica o percentual da prestação em relação ao valor total da dívida.
- **Qtd_Prestacoes**: Variável numérica que representa a quantidade de parcelas do financiamento.
- **Vlr_Mercado_Veiculo**: Variável numérica que indica o valor de mercado do veículo no ato do financiamento.
- **Perc_Entrada_Veiculo**: Variável numérica que representa o percentual do valor de entrada em relação ao valor do veículo.
- **Idade_Cliente**: Variável numérica que indica a idade do cliente no ato do financiamento.
- **Taxista**: Variável categórica que indica se o cliente que solicitou o financiamento é taxista (sim ou não).
- **Autonomo**: Variável categórica que indica se o cliente que solicitou o financiamento é autônomo (sim ou não).
- **Cod_Serasa_Titular**: Variável categórica que representa a classificação do cliente de acordo com o Serasa.
- **Cod_Serasa_Conjuge**: Variável categórica que representa a classificação do cônjuge do cliente de acordo com o Serasa.
- **Possui_Veiculo**: Variável categórica que indica se o cliente possui outro veículo (sim ou não).
- **Marca_Veiculo**: Variável categórica que representa a marca do veículo a ser financiado.
- **UF_Concessionaria**: Variável categórica que indica a Unidade Federativa da concessionária onde o financiamento foi solicitado.
- **Modelo**: Variável categórica que representa o modelo do veículo a ser financiado.
- **Percentual_Renda_Comprometida**: Variável numérica que indica o percentual da renda do cliente comprometida com o valor de cada parcela.
- **Regiao_UF**: Variável categórica que representa a região da moradia do cliente.

- **MOB4**: Variável categórica que indica se o cliente teve atraso maior que 90 dias num período de 120 dias (sim ou não).

2.2 Modelos Utilizados

2.2.1 Regressão Logística

A regressão logística, é um método estatístico utilizado para analisar um conjunto de dados no qual existem uma ou mais variáveis independentes que determinam um resultado binário. O resultado, ou variável resposta, é dicotômico, significando que ele possui apenas dois possíveis valores, que geralmente são representados como 0 e 1. Esta técnica é amplamente utilizada em campos como medicina, biologia, marketing, ciências sociais, e mais, especialmente para prever a presença ou ausência de uma característica ou resultado com base nos valores de outras variáveis.

A distinção fundamental entre regressão logística e regressão linear reside no tipo de variável dependente que eles visam prever. Enquanto a regressão linear é aplicada a variáveis contínuas, a regressão logística é usada para variáveis categóricas binárias. Na prática, isso significa que a regressão logística é ideal para situações onde se deseja modelar a probabilidade de um evento ocorrer, como por exemplo, a probabilidade de um paciente ter uma determinada doença, baseando-se em diversas variáveis independentes como idade, sexo, indicadores de saúde, entre outros.

O modelo de regressão logística é fundamentado na função logística, também conhecida como função sigmoidal. A função logística transforma a combinação linear das variáveis independentes (denotadas por x_i) e os coeficientes do modelo (denotados por β) em uma probabilidade (π_i). Esta probabilidade fica sempre entre 0 e 1, e é calculada pela fórmula:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}$$

Além disso, o modelo de regressão logística pode ser expresso em termos de *odds ratio*, através da seguinte relação logarítmica:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

Nesse contexto, os coeficientes β representam o impacto logarítmico na razão de chances (*odds*) do evento de interesse ocorrer para um aumento unitário na variável independente correspondente. Por meio da estimativa desses coeficientes, o modelo busca maximizar a verossimilhança dos dados observados, proporcionando uma base quantitativa para a previsão de eventos.

Vale ressaltar que, se a variável resposta Y possuir mais de duas categorias, a abordagem padrão da regressão logística binária não será adequada. Nesse caso, variantes

da regressão logística, como a multinomial ou ordinal, podem ser consideradas para tratar de respostas multicategóricas.

A Regressão Logística é frequentemente aplicada em vários campos devido ao seu amplo leque de vantagens e sua capacidade de interpretação e análise. As principais vantagens incluem:

- **Predição de probabilidade**: A modelagem direta da probabilidade de ocorrência de eventos torna este modelo ideal para previsões probabilísticas, além de simples classificações.
- **Interpretação clara dos coeficientes**: Na regressão logística, os coeficientes explicam a variação nas chances de um evento ocorrer com base em cada unidade de variação das variáveis explicativas.
- **Aplicabilidade em estudos de caso-controle**: Este modelo é particularmente útil para analisar a relação entre exposições e desfechos em estudos de caso e controle.
- **Flexibilidade de modelagem**: Aceita variáveis categóricas como preditores, proporcionando versatilidade em diferentes tipos de análise de dados.
- **Facilidade de implementação**: Oferece uma programação menos complexa em comparação com outros modelos estatísticos, facilitando sua aplicação.

No entanto, existem desvantagens e limitações:

- **Necessidade de grandes amostras**: Para ser eficaz, necessita de grandes conjuntos de dados para estimativas precisas.
- **Problemas com desequilíbrio de classes**: A performance pode cair quando há uma disparidade significativa entre as classes da variável dependente.
- **Sensibilidade a outliers**: Valores anômalos podem influenciar negativamente as estimativas dos coeficientes e a interpretação dos resultados.
- **Limitação a respostas binárias**: É adequado somente para variáveis dependentes binárias, sendo inaplicável para respostas contínuas ou multiclasse.

2.2.2 Árvore de Decisão

Para podermos adentrar aos próximos dois algoritmos que apresentaremos, precisamos falar sobre o conceito fundamental que são a base destes métodos, que é *Arvore de Decisão*.

Uma árvore de decisão é um algoritmo de aprendizado supervisionado usado para classificação, bem como em problemas de regressão. Trata-se de um fluxograma com estrutura de uma árvore "invertida", utilizado para prever resultados a partir de divisões baseadas em atributos. Esse processo de subdivisões, com base num teste de valor para cada atributo do conjunto de dados, é chamado particionamento recursivo.

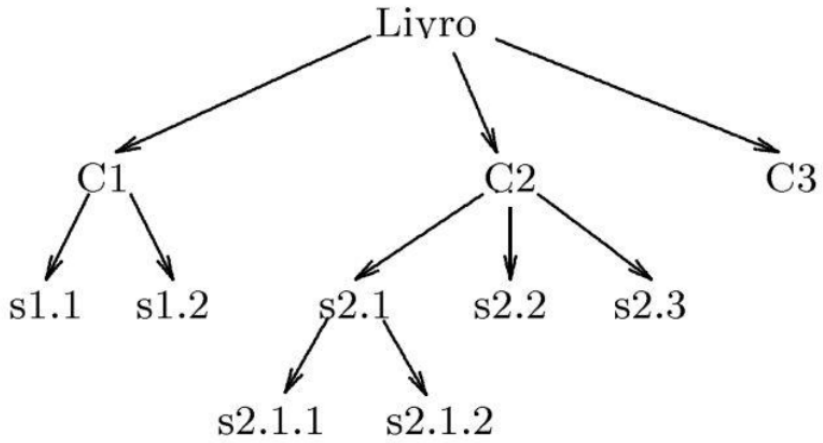

Figura 1 – Representação do sumário de um livro em forma de árvore.

A árvore de decisão é uma coleção de elementos chamados nós, dentre os quais um é distinguido como o nó-raiz, que impõe uma estrutura hierárquica sobre todos os outros nós. Formalmente, temos o sequinte:

a. um único *nó* é um arvore, sendo que este *nó* também é a raiz da árvore;
b. suponha que t seja um *nó* T_1, T_2,...,T_k sejam árvores com raízes t_1, t_2,...,t_k, respectivamente. Podemos construir uma nova árvore transformando t no *pai* dos *nós* t_1, t_2,...,t_k. Nessa árvore, t será a raiz e T_1, T_2,...,T_k serão as subárvores ou ramos da raiz. Os *nós* t_1, t_2,...,t_k são chamados *filhos* do *nó* t.

A figura 1 ilustra o sumário de um livro representado através de uma árvore. A raiz, chamada *Livro*, possui três subárvores cujas raízes correspondem aos capitulos $C1$, $C2$ e $C3$, os quais são *nós-filhos* de *Livro*. Se t_1, t_2,...,t_k é uma sequência de *nós* em uma árvore tais que t_i é o pai de t_{i+1} para $1 \leq i \leq k$, então esta sequencia é denominada um caminho do *nó* t_1 até o *nó* t_k. No exemplo da *Figura 16*, a sequência $C2$, $s2.1$, $s2.1.2$ é um caminho do capítulo $C2$ até a seção $s2.1.2$. Todos os *nós* que não possuem *filhos* são chamados de *nós-terminais* ou *folhas*, tais como: $C3$ e $s2.3$. Já os *nós* que contém *filhos* são chamados de *nós internos* ou *não terminais*, tais como: $C1$ e $s2.1$.

2.2.3 XGBoost

O XGBoost (*Extreme Gradient Boosting*) é uma metodologia de aprendizado de máquina, concebida especificamente para endereçar problemas supervisonados, tais como classificação e regressão. Este algoritmo foi desenvolvido e apresentado ao mundo acadêmico e profissional por Tianqi Chen e Carlos Guestrin no ano de 2016 (CHEN TIANQI; GUESTRIN, 2016), marcando um passo significativo na evolução dos sistemas de inteligência artificial.

O *XGBoost* se insere na categoria dos algoritmos de *gradient boosting*, que são técnicas que visam criar um modelo preditivo por meio da combinação de diversos modelos

mais simples e menos eficientes. A principal estratégia utilizada pelo *XGBoost* envolve a implementação sequencial de árvores de decisão, onde cada nova árvore é construída para corrigir os erros gerados pelas árvores anteriores, num processo iterativo e aditivo. Esse mecanismo garante a melhoria contínua do desempenho do modelo até atingir um nível de precisão considerável.

Apesar de sua eficácia e sofisticação, o *XGBoost* apresenta um grau de complexidade que pode tornar a interpretação dos modelos gerados um desafio, especialmente quando se trata de entender as relações não lineares entre as variáveis preditoras e a variável alvo.

O modelo *XGBoost* é notável pela sua capacidade de construir progressivamente um modelo robusto por meio da adição sequencial de árvores de classificação e regressão ($CART$). A mecânica central do algoritmo reside na formulação aditiva, onde o valor predito, denotado por $\hat{y}_i$, é calculado como a soma dos resultados de K árvores distintas, cada uma representada pela função f_k, que por sua vez pertence a um espaço funcional $\mathcal{F}$ abrangendo todas as possíveis árvores de decisão.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

No núcleo do procedimento do *XGBoost* está o conceito de otimização sequencial. O algoritmo inicia com uma predição básica e a partir daí, calcula os resíduos com base na discrepância entre os valores preditos e observados. Utilizando esses resíduos, uma árvore de decisão é então formulada. Cada folha dentro desta árvore é designada com uma pontuação de similaridade, que serve para avaliar a homogeneidade dos dados contidos dentro dela. Além disso, cada possível divisão de dados na árvore é avaliada com base em um ganho de similaridade, facilitando a seleção da melhor maneira de bifurcar os dados em cada nó.

Este procedimento não se detém após a construção de uma única árvore; pelo contrário, ele segue de forma iterativa, com cada nova árvore sendo moldada pelos resíduos resultantes da árvore anterior. Este processo iterativo e acumulativo permite que o modelo refine continuamente suas previsões, com o objetivo de reduzir os erros residuais de maneira significativa ou até atingir um determinado número de iterações. Dessa maneira, cada árvore subsequente é informada e ajustada pelas anteriores, o que significa que cada nova adição é uma resposta direta aos erros passados, promovendo um aprendizado cumulativo e dinâmico dentro do modelo.

Além disso, uma parte integral da estrutura do *XGBoost* é a sua função objetivo, que consiste de dois componentes principais: a função de perda $l(y_i, \hat{y}_i^{(t)})$ e o termo de regularização $\omega(f_i)$. A função de perda avalia o quão bem o modelo está performando, isto é, medindo a discrepância entre os valores reais y_i e os valores preditos $\hat{y}_i^{(t)}$. Em paralelo, o termo de regularização contribui para a estrutura do modelo, impondo penalidades para a complexidade das árvores e ajudando a diminuir o problema de *overfitting*. Este equilíbrio

entre ajuste do modelo e controle de complexidade é crucial para garantir que o *XGBoost* não apenas alcance uma alta precisão preditiva, mas também mantenha uma generalização adequada a novos dados.

$$\text{obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \omega(f_i)$$

Os hiperparâmetros são elementos usados na configuração de modelos de aprendizado de máquina como o *XGBoost*, permitindo o ajuste do modelo para obter a melhor performance possível enquanto evitam o *overfitting*. Aqui estão alguns dos hiperparâmetros utilizados no *XGBoost*, cada um com seu respectivo impacto no desempenho e complexidade do modelo:

- **gamma (min_split_loss)**: Valor mínimo de perda necessário para fazer uma divisão adicional em um nó da árvore. Ajustar esse parâmetro ajuda a controlar a complexidade do modelo.
- **colsample_bytree, colsample_bylevel, colsample_bynode**: Esses parâmetros controlam a fração de colunas a serem amostradas para construir cada árvore, a cada nível ou em cada divisão.
- **Profundidade da Árvore (max_depth)**: Este hiperparâmetro limita a profundidade máxima de cada árvore de decisão construída. Uma profundidade maior permite que o modelo capture interações mais complexas entre as características, mas também aumenta o risco de aprender demais dos dados de treinamento, resultando em *overfitting.* O ajuste desse parâmetro é importante para encontrar o equilíbrio certo entre bias e variância.
- **Taxa de Aprendizado (learning_rate/eta)**: Também conhecida como taxa de encolhimento, esse parâmetro modula a contribuição de cada árvore no resultado final. Uma taxa de aprendizado menor pode tornar o modelo mais robusto para overfitting, pois cada árvore adicionada ajusta-se de forma mais sutil aos erros residuais. No entanto, isso geralmente requer um número maior de árvores para atingir uma performance de modelo ótima.
- **Mínimo de Amostras por Folha (min_child_weight)**: Este parâmetro define o número mínimo de amostras que devem estar presentes em uma folha (nó final) da árvore. Valores maiores impedem que o modelo crie folhas com pouquíssimas amostras, o que pode ser benéfico para prevenir overfitting, mantendo a estrutura das árvores mais simples e mais generalizada.
- **Subamostragem (subsample)**: Este parâmetro determina a fração das observações (linhas dos dados) a serem aleatoriamente amostradas para cada árvore. A

subamostragem pode ajudar a tornar o modelo mais diversificado e menos propenso a overfitting, já que cada árvore é construída a partir de uma amostra diferente de dados. No entanto, uma taxa muito baixa pode levar a uma performance inadequada do modelo, pois cada árvore terá menos dados para aprender.

2.2.4 Random Forest

As Random Forests são algoritmos de aprendizado supervisionado, construídos pela combinação de algoritmos mais básicos, mas que os tornam mais robustos. Tal característica está na capacidade de combinar um grande número de árvores de decisão, com diferentes respostas, em um único resultado. Em problemas de classificação, o resultado mais votado ou que mais se repete, será o escolhido; no caso dos problemas de regressão, é calculada a média dos valores fornecidos pelos modelos para obtenção do resultado final.

O objetivo principal das Random Forest é agrupar conjuntos de preditores não necessariamente ótimos, ao invés de buscar otimizar cada preditor de uma única vez, como acontece nos modelos CART (*Classification and Regression Trees*). Como as arvores individuais são perturbadas aleatoriamente, a floresta se beneficia de uma exploração mais extensa do espaço total de atributos, permitindo obter um melhor desempenho preditivo.

Cada árvore dentro da Random Forest é construida, basicamente, da seguinte maneira:

- A partir de um conjunto de treino com N observações, é feita uma amostragem *bootstrap*, com reposição, também de N observações. Essa amostra será o conjunto de treino para o crescimento de uma árvore de decisão;
- Supondo M atributos de entrada, um número $m \ll M$ é especificado, de forma que em cada *nó*, m variáveis sejam selecionadas de M, de forma aleatória, e a melhor divisão entre m será utilizada em cada *nó*;
- Cada árvore individual é construída na maior extensão possível, ou seja, não há poda.

Os algoritmos *Random Forests* funcionam de forma bastante eficiente quando aplicados a grandes volumes de dados, podendo lidar com centenas de atributos de entrada. Além disso, apresentam alta precisão quando comparados a outros modelos. Também fornecem estimativas de quais atributos são mais importantes na regressão ou classificação e geram estimativas internas de erro, a medida em que se avança na sua construção.

Quando em um determinado conjunto de treinamento são definidas as reamostragens do tipo *bootstrap*, para construção de cada árvore de decisão, cerca de um terço destas observações são deixadas de fora de cada amostra. Esses registros são chamados *Out-of-Bag* e são usados internamente pela *Random Forest*, para obter uma estimativa

não viesada do erro de classificação, à medida em que outras árvores são adicionadas ao modelo.

Para estimar o grau de importância de cada atributo do conjunto de dados, o algoritmo de *Random Forest* realiza o seguinte processo:

- Após cada árvore ser construída na floresta é realizada uma validação interna com as observações *out-of-bag*, calculando-se a proporção de instâncias *out-of-bag* que foram classificadas corretamente;
- Em seguida, permuta-se os valores de uma das variáveis m, de forma aleatória, dentro das instâncias *out-of-bag*, e outra validação interna é realizada pelo modelo;
- Por fim, é calculada a diferença entre a proporção de instâncias *out-of-bag* classificadas incorretamente, após permutação da variável m, e a proporção de instâncias classificadas incorretamente dos dados *out-of-bag* intocados. A média dessa diferença nos fornece a pontuação de importância da variavel m.

A taxa de erro das *Random Forests* depende, principalmente, de dois fatores:

- da correlação entre duas árvores quaisquer do modelo, onde o aumento da correlação aumenta o erro;
- e da força individual de cada árvore, ou seja, uma árvore com baixa taxa de erro é um classificador forte e, assim, aumentar a força de árvores individuais diminui a taxa de erro como um todo.

Para um melhor ajuste do modelo, especificamos alguns hiperparêmtros do *Random Forest*:

- *ntree*: Determina o número de árvores na floresta. Um número maior de árvores aumenta a precisão do modelo mas também o custo computacional.
- *mtry*: Especifica o número de variáveis a serem consideradas aleatoriamente em cada divisão do nó. Influencia diretamente a diversidade entre as árvores e pode impactar tanto a precisão quanto o risco de overfitting.
- *node_size*: Define o tamanho mínimo de um nó terminal. Aumentar esse valor ajuda a prevenir o overfitting, assegurando que cada nó final tenha um número suficiente de observações.
- *replace*: Indica se a amostragem das observações para a construção das árvores é feita com substituição. Usar substituição aumenta a variância entre as árvores, o que pode ser benéfico para a generalização do modelo.

2.3 Ferramentas Utilizadas

O estudo foi realizado utilizando o software estatístico R (versão 4.4.0) e o ambiente de desenvolvimento RStudio (versão 2023.12.1). Para a implementação dos modelos e análise dos dados, foram utilizados os seguintes pacotes do R:

- **ggplot2**: Utilizado para visualização de dados. (WICKHAM, 2016)
- **dplyr** e **tidyr**: Utilizados para a manipulação de dados. (WICKHAM et al., 2023) e (WICKHAM; VAUGHAN; GIRLICH, 2024)
- **ROSE**: Utilizado para tratamento de desbalanceamento nas classes. (LUNARDON; MENARDI; TORELLI, 2014)
- **corrplot**: Utilizado para a criação de gráficos de correlação. (WEI; SIMKO, 2021)
- **MASS**: Utilizado para funções estatísticas e de modelagem. (VENABLES; RIPLEY, 2002)
- **randomForest**: Utilizado para a implementação do modelo Random Forest. (LIAW; WIENER, 2002)
- **xgboost**: Utilizado para a implementação do modelo de XGBoost. (CHEN et al., 2024)
- **caret**: Utilizado para a criação de pipelines de pré-processamento e treinamento de modelos. (Kuhn; Max, 2008)
- **xtable**: Usado para criação de tabelas de contingência (DAHL et al., 2019)

2.4 Procedimentos

1. **Pré-processamento dos Dados**:
    - Os dados foram importados e limpos utilizando os pacotes `tidyr`, `dplyr` e `janitor`.
    - Variáveis categóricas foram codificadas e valores ausentes foram tratados adequadamente.
    - A análise exploratória dos dados foi realizada com `xtable` e visualizada com `ggplot2`.
    - Identificação e exclusão de valores extremos (*outliers*) que poderiam afetar a análise.
    - Os registros com valores ausentes (*NA's*) foram removidos do conjunto de dados.
    - Utilização de técnicas para lidar com o desbalanceamento das classes, garantindo uma representação adequada de cada categoria.
2. **Treinamento e Avaliação dos Modelos**:

- Os dados foram divididos em conjuntos de treinamento (70% da base) e teste (30% da base).
- Modelos de regressão logística, *Random Forest* e *XGBoost* foram treinados utilizando o pacote `caret`.
- A avaliação dos modelos foi realizada com métricas apropriadas, como acurácia, sensibilidade, especificidade, coeficiente KS (*Kolmogorov e Smirnov*), gráfico de distribuição cumulativa em uma base *out of time* (dados fora das safras utilizadas para treinamento e teste).

3 Resultados e Discussão

3.1 Análise Descritiva

Iniciamos a análise com um resumo estatístico gerado pela saída do *summary* da base de dados, o que facilita a compreensão da distribuição dos dados e auxilia na identificação de padrões ou anomalias significativas.

```
##      SAFRA         IDADE_BEM        PERC_PREST        QTDE_PREST
##  201803 : 11488   NOVO :373394   Min.   :0.00000   Min.   : 6.00
##  201907 : 11335   USADO:105259   1st Qu.:0.02651   1st Qu.:36.00
##  201905 : 11071                  Median :0.03049   Median :48.00
##  201903 : 11049                  Mean   :0.03478   Mean   :43.58
##  201811 : 10979                  3rd Qu.:0.03780   3rd Qu.:60.00
##  201910 : 10954                  Max.   :0.17622   Max.   :72.00
##  (Other):411777
```

```
##   VLR_MER_VEI      PERC_ENT_VEI       IDADE       TAXISTAS
##  Min.   :  2020   Min.   : 0.00   Min.   :18.00   Não:466658
##  1st Qu.: 42521   1st Qu.:24.52   1st Qu.:38.00   Sim: 11995
##  Median : 62525   Median :42.42   Median :47.00
##  Mean   : 69837   Mean   :41.58   Mean   :47.83
##  3rd Qu.: 86644   3rd Qu.:60.00   3rd Qu.:58.00
##  Max.   :343700   Max.   :98.37   Max.   :90.00
```

```
##  AUTONOMOS     COD_SER_TIT      COD_SER_CONJ      VLR_RENDA
##  Não:457736   N      :427751          :398012   Min.   :        450
##  Sim: 20917   3      : 18460   N      : 69833   1st Qu.:       4800
##               2      : 13388   3      :  4426   Median :       6500
##               4      : 10548   4      :  2665   Mean   :      80305
##               0      :  8193   2      :  2496   3rd Qu.:      10000
##               Y      :   173   0      :  1172   Max.   :24348006000
##               (Other):   140   (Other):    49
```

```
##   POSSUI_VEIC                MARCA              UF
##  Não    : 57223   MARCA_B     :123314   SP     :130131
##  Não Inf: 11829   OUTRAS MARCAS: 59238   SC     : 41365
##  Sim    :409601   MARCA_A     :296101   RJ     : 39118
##                                         PR     : 34785
##                                         RS     : 30928
##                                         PE     : 25253
```

```
##                                          (Other):177073
##            MODELO
##  MOD_0       :127956
##  MOD_1       : 71269
##  OUTRAS MARCAS: 59238
##  MOD_3       : 40938
##  MOD_4       : 39378
##  MOD_5       : 35254
##  (Other)     :104620

##    COMP_RENDA              REGIAO                MOB4
##  Min.   : 0.00   Centro Oeste  : 49048   Adimplente  :476610
##  1st Qu.:12.20   Nordeste      : 96791   Inadimplente:  2043
##  Median :17.85   Norte         : 26299
##  Mean   :17.61   Sao Paulo     :130131
##  3rd Qu.:23.37   Sudeste sem SP: 69306
##  Max.   :30.00   Sul           :107078
```

3.1.1 Variáveis Numéricas

Em seguida, apresentaremos gráficos de densidade das variáveis numéricas em relação à variável resposta, MOB4. Estes gráficos nos permitirão visualizar as distribuições e explorar possíveis relações entre as variáveis numéricas e o comportamento da variável resposta.

A figura 2 analisa a distribuição da variável "PERC_PRESTACAO", que refere-se ao percentual do valor da prestação em relação ao valor do veículo, para os grupos de inadimplentes e adimplentes da variável resposta MOB4. Observa-se que os inadimplentes têm uma distribuição mais concentrada, com um pico notável em torno de 2.5%, indicando uniformidade nos valores da prestação como percentual do valor do veículo. Já os adimplentes apresentam uma distribuição mais ampla e vários picos, sugerindo uma maior diversidade nos valores das prestações.

A figura 3 apresentada analisa a distribuição da variável "Quantidade de prestações" para os grupos de inadimplentes e adimplentes. Para os inadimplentes, observa-se uma distribuição que possui três picos significativos, sugerindo pontos específicos onde a quantidade de prestações é mais comum entre os que falharam em manter seus pagamentos em dia. Os picos ocorrem em 48 e 60 prestações. No caso dos adimplentes, a distribuição também apresenta picos em 24, 36, 48 e 60 prestações.

A figura 4 apresenta a distribuição do valor de mercado dos veículos, em reais, para os grupos "Inadimplente" e "Adimplente". Observa-se que os inadimplentes têm uma distribuição mais concentrada, com um pico predominante em torno de R$ 50.000,

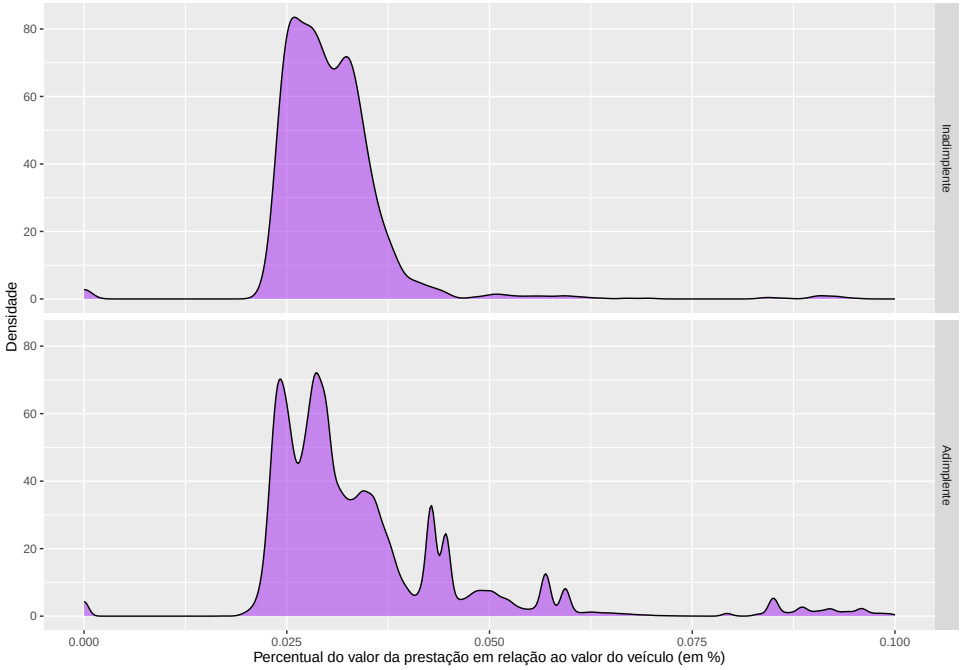

Figura 2 – Distribuição do percentual de prestação pela variável resposta.

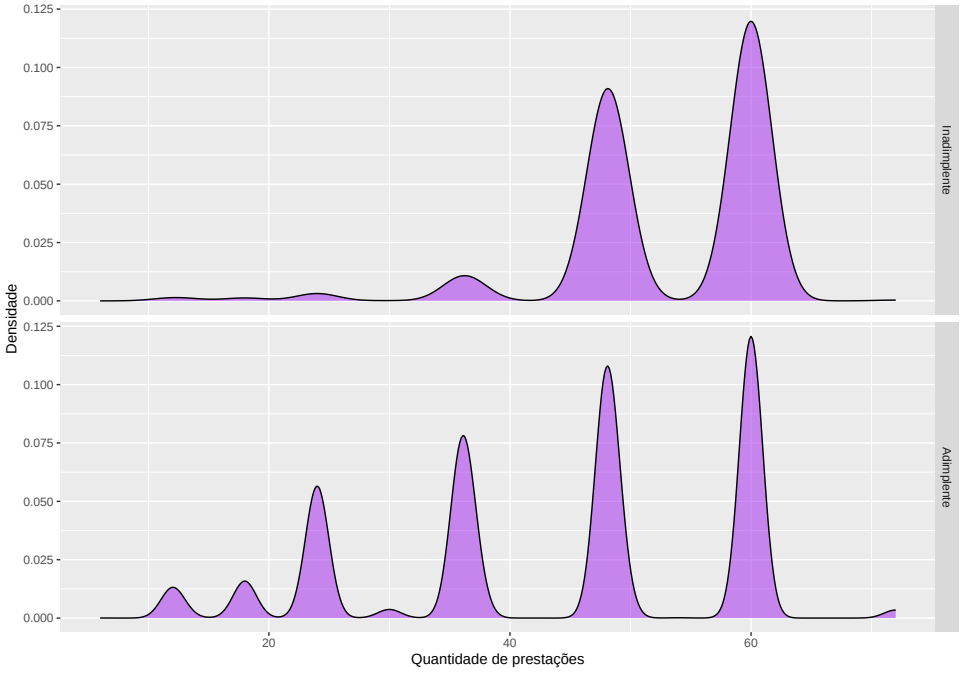

Figura 3 – Distribuição do número de prestações pela variavel resposta.

indicando que a maioria possui veículos de menor valor. Já os adimplentes apresentam uma distribuição mais variada, com picos notáveis em R$ 40.000, R$ 70.000 e R$ 85.000, refletindo uma maior diversidade nos valores de seus veículos.

A figura 5 ilustra a distribuição do “Percentual de Entrada do Veículo” para os grupos “Adimplente” e “Inadimplente”. Observamos que os inadimplentes apresentam uma curva com um pico significativo em torno de 25% de entrada, com densidade mais alta e decaindo progressivamente à medida que o percentual de entrada aumenta. Para os

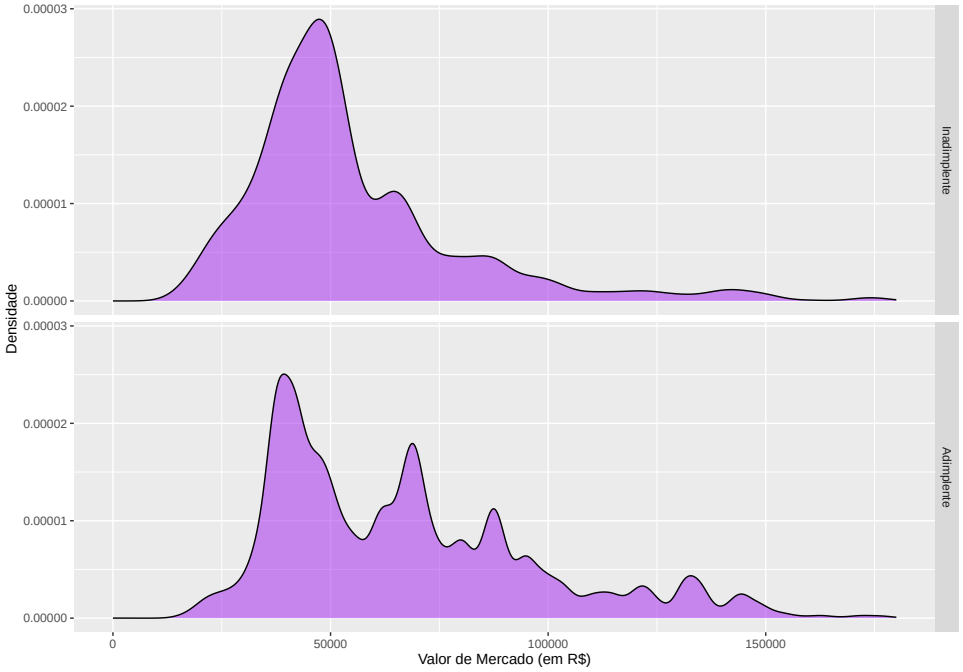

Figura 4 – Distribuição do valor de mercado do veículo pela variavel resposta.

adimplentes, a distribuição é mais variada, com múltiplos picos localizados em torno de 10%, 30%, 50% e 75%, indicando que há uma diversidade maior nos valores percentuais de entrada para a compra de veículos.

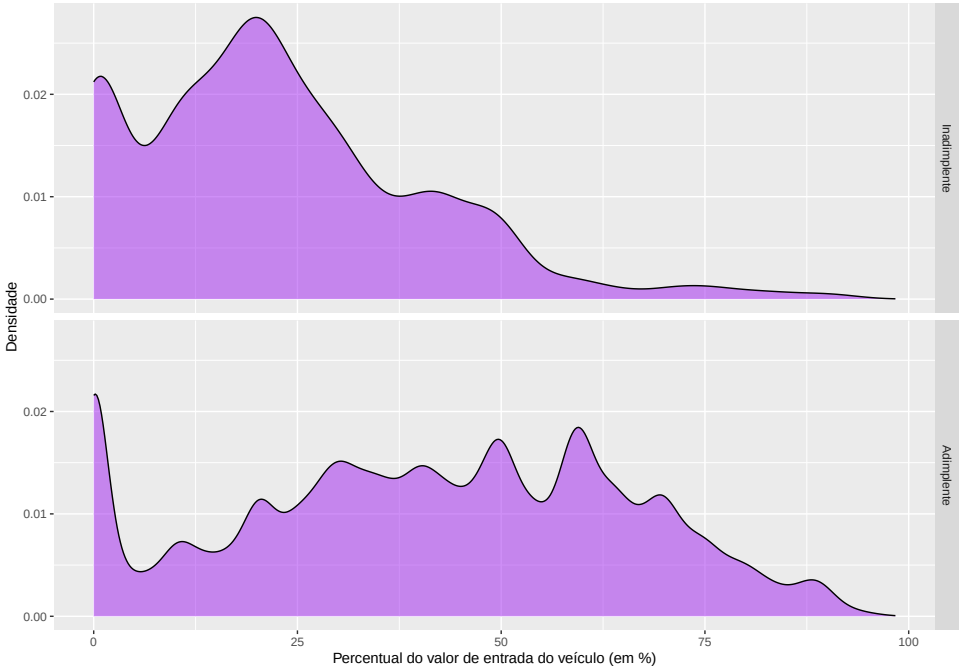

Figura 5 – Distribuição do percentual de entrada do veículo pela variavel resposta.

A figura 6 ilustra a distribuição de idade para os grupos “Inadimplente” e “Adimplente”. No gráfico dos inadimplentes, a densidade aumenta progressivamente até cerca de 40 anos, onde atinge um pico, e então declina gradualmente, mantendo-se mais constante

após os 60 anos. Este padrão sugere que a maior parte dos inadimplentes está concentrada na faixa dos 30 aos 50 anos. Para os adimplentes, a distribuição também mostra um aumento até um pico por volta dos 40 anos, mas a curva é mais ampla, indicando uma distribuição mais uniforme de idades até cerca de 60 anos, após o que a densidade diminui mais acentuadamente.

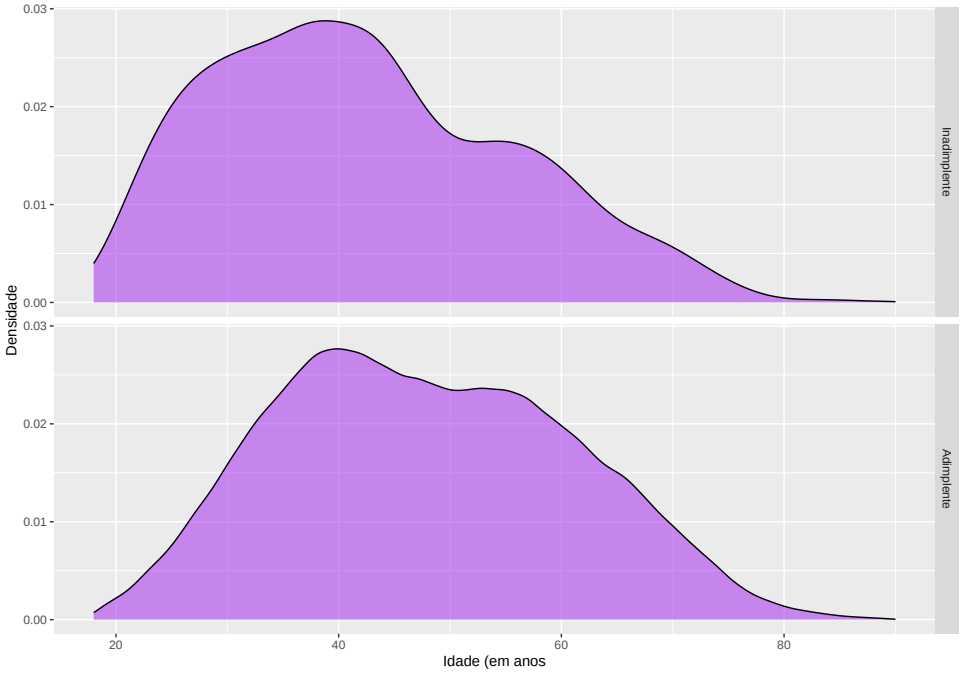

Figura 6 – Distribuição da idade do cliente pela variavel resposta.

A figura 7 exibe a distribuição da renda, em reais, para os grupos “Inadimplente” e “Adimplente”. No gráfico dos inadimplentes, observa-se uma distribuição com um pico acentuado em torno de R$ 5.000, após o qual a densidade cai drasticamente e se mantém relativamente baixa para valores mais altos de renda. Isso indica que a maioria dos inadimplentes possui uma renda concentrada em torno desse valor. Para os adimplentes, a distribuição é mais fragmentada com vários picos menores, o maior deles em torno de 5.000 reais, seguido de picos menores espalhados entre 10.000 a 30.000 reais. Isso sugere que os adimplentes possuem uma variedade maior de níveis de renda, embora ainda haja uma concentração significativa em faixas de renda mais baixa.

A figura 8 mostra a distribuição do percentual de renda comprometida entre os grupos “Inadimplente” e “Adimplente”. Para os inadimplentes, a distribuição é concentrada perto de 25%. Isso indica que os inadimplentes geralmente têm uma parcela mais alta de renda comprometida. Por outro lado, a distribuição entre os adimplentes é bem mais dispersa. Esta dispersão sugere que adimplentes podem ter uma proporção variável de renda comprometida, mas geralmente permanece dentro de um limite razoável.

A figura 9 é uma matriz de correlação que exibe as relações entre diversas variáveis financeiras e demográficas. Cada célula da matriz mostra o coeficiente de correlação entre duas variáveis, representado tanto numericamente quanto visualmente por meio

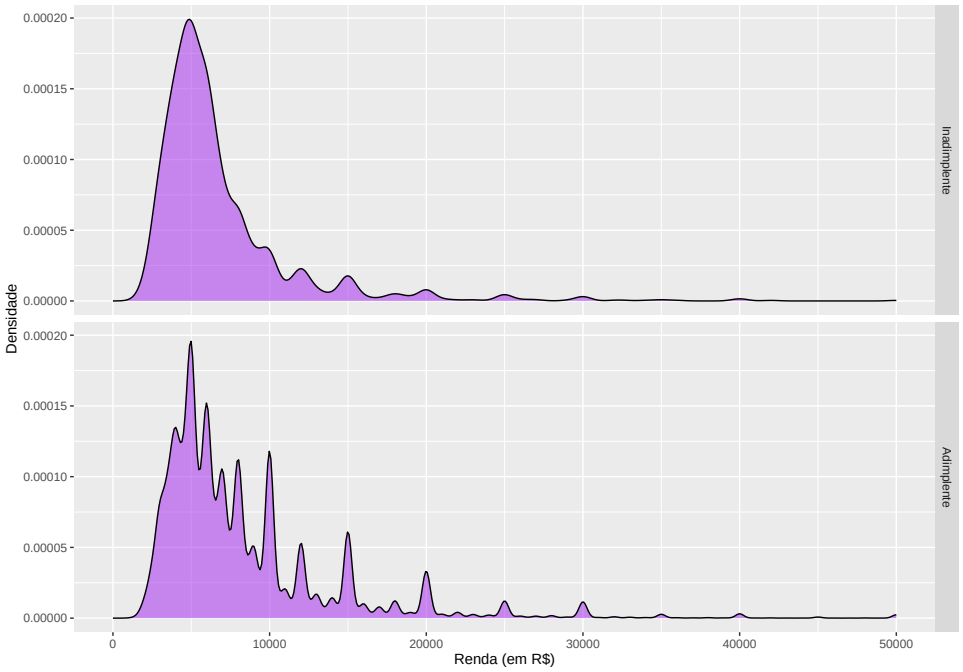

Figura 7 – Distribuição da renda pela variavel resposta.

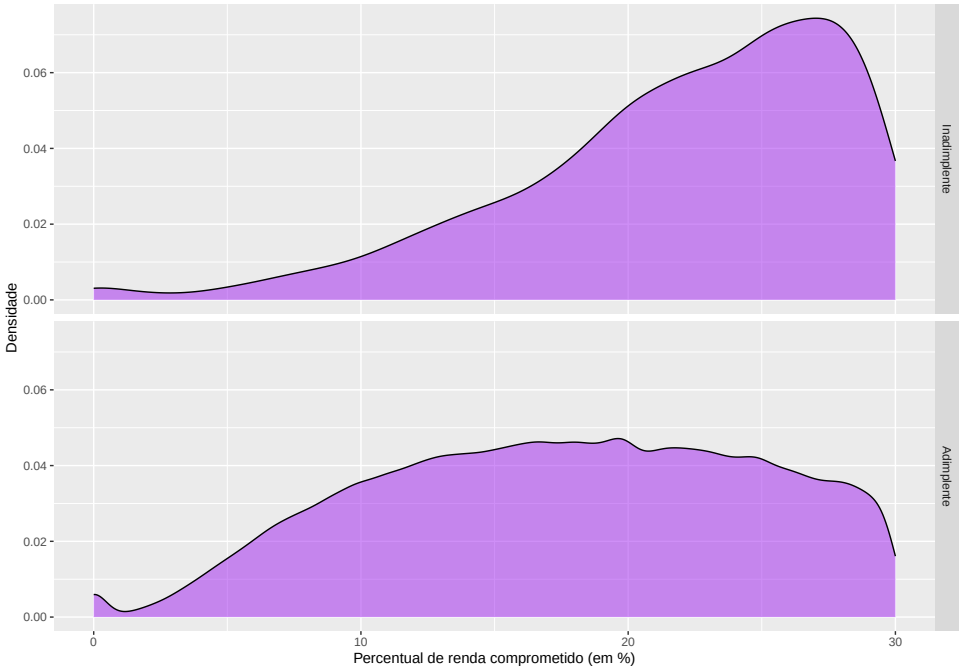

Figura 8 – Distribuição do percentual de rendacoprometida pela variavel resposta.

da coloração e do gráfico de dispersão para correlações significativas. As variáveis que possuem correlações muito altas, serão retiradas da análise.

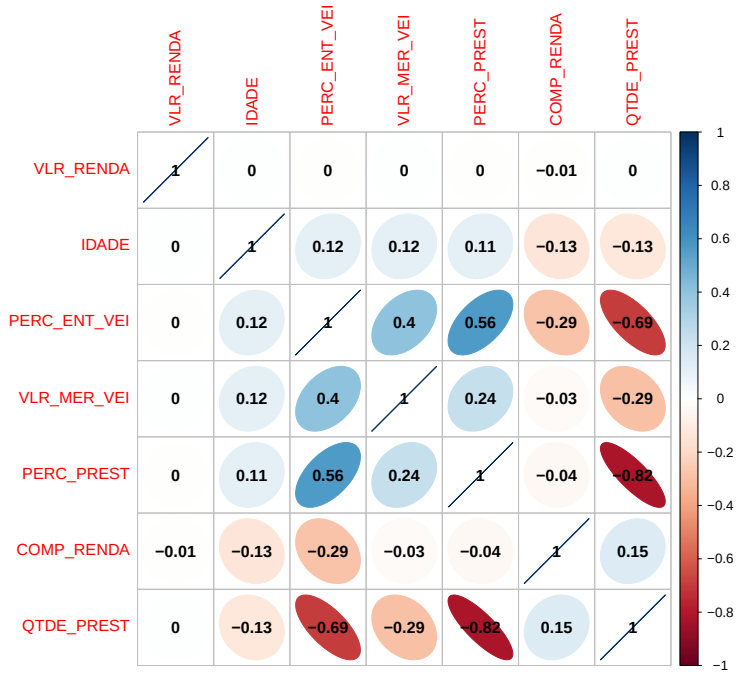

Figura 9 – Matriz de correlação entre as variáveis numéricas.

3.1.2 Variáveis Categóricas

A tabela 1 mostra dados sobre inadimplência relacionada à idade do veículo, diferenciando veículos novos e usados. Observa-se que veículos usados têm uma taxa de inadimplência mais alta (0.74%) em comparação com veículos novos (0.34%). Além disso, a quantidade de veículos novos adimplentes supera significativamente a de usados (372.134 contra 104.476), indicando que mais veículos novos estão sendo financiados.

Tabela 1 – Inadimplencia por Idade do veículo.

Idade	MOB0	MOB1	Tx_Inadimp
USADO	104476	783	0.74
NOVO	372134	1260	0.34

A tabela 2 apresenta dados sobre inadimplência segmentada pela ocupação dos indivíduos como taxistas. Observa-se que a inadimplência entre os não taxistas e taxistas é quase idêntica, com taxistas apresentando uma leve vantagem, com 0.42% contra 0.43% dos não taxistas. Embora a proporção de taxistas inadimplentes seja levemente menor, o número absoluto de não taxistas envolvidos é substancialmente maior, com 464.665 adimplentes contra 11.945 adimplentes que são taxistas.

Tabela 2 – Inadimplência por Taxistas.

Taxista	MOB0	MOB1	Tx_Inadimp
Não	464665	1993	0.43
Sim	11945	50	0.42

A Tabela 3 mostra a inadimplência dividida entre trabalhadores autônomos e não autônomos. Enquanto os não autônomos têm uma taxa de inadimplência de 0.41%, os autônomos apresentam uma taxa significativamente maior, de 0.78%. Embora o número absoluto de autônomos adimplentes (20.754) seja menor em comparação aos não autônomos (455.856), a proporção de inadimplentes entre os autônomos é quase o dobro da observada entre os não autônomos.

Tabela 3 – Inadimplência por Autonomos

Autonomo	MOB0	MOB1	Tx_Inadimp
Sim	20754	163	0.78
Não	455856	1880	0.41

A Tabela 4 destaca a inadimplência associada a diferentes códigos Serasa do titular. Observa-se uma variação significativa nas taxas de inadimplência, que variam de 0% para os códigos C, P, W e Z, até 1.29% para o código 0, que apresenta a maior taxa de inadimplência. O código N, que compreende a maioria dos titulares com 426,099 casos, tem uma baixa taxa de inadimplência de 0.39%.

Tabela 4 – Inadimplência por Código Serasa Titular.

Cod Serasa Titular	MOB0	MOB1	Tx_Inadimp
0	8087	106	1.29
3	18306	154	0.83
2	13309	79	0.59
Y	172	1	0.58
4	10497	51	0.48
N	426099	1652	0.39
C	9	0	0.00
P	5	0	0.00
W	74	0	0.00
Z	52	0	0.00

A tabela 5 fornece indícios sobre como os códigos de Serasa atribuídos aos cônjuges podem estar associados à inadimplência dos titulares de crédito. Nota-se que cônjuges com o código ‘0’ têm a maior percentagem de inadimplência dos titulares associados, com 0.49%. Isso sugere que, quando o cônjuge tem um código ‘0’, existe uma maior probabilidade de inadimplência por parte do titular. Em contraste, códigos como ‘C’, ‘P’, ‘W’, ‘Y’, e ‘Z’, mostram zero incidência de inadimplência dos titulares.

A tabela 6 mostra a inadimplência baseada na posse de veículo. Observa-se que entre os que não possuem veículo, a percentagem de inadimplência é significativamente maior, com 0,97%. Isso contrasta com os que possuem veículo ou não têm essa informação disponível, onde as taxas de inadimplência são consideravelmente menores, em 0,35% e 0,30%, respectivamente.

Tabela 5 – Inadimplência por Código Serasa Conjuge

Cod Serasa Conjuge	MOB0	MOB1	Tx_Inadimp
	396049	1963	0.49
2	2487	9	0.36
3	4416	10	0.23
4	2659	6	0.22
0	1171	1	0.08
N	69779	54	0.08
C	1	0	0.00
P	1	0	0.00
W	19	0	0.00
Y	23	0	0.00
Z	5	0	0.00

Tabela 6 – Inadimplência por Quem Possui Veiculo.

Possui Veículo	MOB0	MOB1	Tx_Inadimp
Não	56669	554	0.97
Sim	408148	1453	0.35
Não Inf	11793	36	0.30

A tabela 7 apresenta dados de inadimplência segmentados por marca de veículo tratada, evidenciando diferenças notáveis entre as marcas. Veículos da MARCA_B apresentam a menor taxa de inadimplência, com apenas 0,17%. Em contraste, veículos classificados como “Outras Marcas” têm a maior taxa de inadimplência, de 0,95%. A MARCA_A possui uma taxa intermediária de inadimplência, de 0,43%, posicionando-se entre as outras duas categorias em termos de risco financeiro associado.

Tabela 7 – Inadimplência por Marca do Veículo

Marca Veículo	MOB0	MOB1	Tx_Inadimp
OUTRAS	58672	566	0.95
MARCA_A	294835	1266	0.43
MARCA_B	123103	211	0.17

A tabela 8 exibe as taxas de inadimplência por unidade federativa (UF) onde as concessionárias estão localizadas, mostrando diferenças significativas entre os estados. A Bahia (BA) registra a maior taxa de inadimplência, com 1,12%. Por outro lado, Mato Grosso do Sul (MS) e Rio Grande do Norte (RN) apresentam as menores taxas, com 0,15%. Estados como São Paulo (SP) e Mato Grosso (MT) também mostram baixas taxas de inadimplência, em torno de 0,17%.

A tabela 9 destaca a inadimplência por modelo de veículo, mostrando variações entre os modelos. Os cinco veículos com as maiores taxas de inadimplência variam de 2,02% a 1,09%, mas têm pouca representatividade numérica, correspondendo a menos de 0,5% do total da base. Os veículos categorizados como “OUTROS”, que incluem veículos

Tabela 8 – Inadimplência por UF da Concessionária.

UF	MOB0	MOB1	Tx_Inadimp
RR	1395	16	1.13
BA	23568	267	1.12
SC	41029	336	0.81
DF	20028	150	0.74
PE	25067	186	0.74
PI	5319	36	0.67
GO	14921	100	0.67
PB	7285	47	0.64
CE	11687	74	0.63
RO	2702	17	0.62
SE	4813	30	0.62
AL	4113	25	0.60
MA	9642	49	0.51
PA	12214	52	0.42
PR	34653	132	0.38
AP	1829	6	0.33
RS	30837	91	0.29
MG	22461	57	0.25
ES	7652	18	0.23
TO	2639	6	0.23
AC	459	1	0.22
RJ	39038	80	0.20
AM	4953	10	0.20
SP	129903	228	0.17
MT	7714	13	0.17
RN	4576	7	0.15
MS	6113	9	0.15

usados e de diversas marcas, registram uma taxa de inadimplência de 0,95% e representam 12,4% do total da base. O modelo MOD_10 se destaca pela sua presença na base, com 26,7% de todos os financiamentos e uma taxa de inadimplência de 0,61%. Alguns modelos apresentam 0% de inadimplência, embora os números absolutos sejam muito pequenos, representando menos de 0,1% do total da base.

A tabela 10 mostra a inadimplência por região no Brasil, destacando que a região Nordeste apresenta o maior percentual de inadimplência, com 0.74%. Em contraste, a região de São Paulo tem a menor taxa de inadimplência, com apenas 0.17%. As demais regiões, incluindo Centro-Oeste, Norte, Sudeste sem SP e Sul, mostram taxas intermediárias de inadimplência, variando de 0.22% a 0.56%.

Esta análise descritiva das variáveis, tanto as numéricas quanto categóricas, ofereceu uma visão ampla do comportamento da base de dados. Variáveis que não serão utilizadas diretamente nos modelos também fornecem informações valiosas, que podem ser empregadas em outras tomadas de decisão. Esse é o motivo pelo qual foi realizado uma

Tabela 9 – Inadimplência por Modelo do Veículo

Modelo Veículo	MOB0	MOB1	Tx_Inadimp
MOD_24	194	4	2.02
MOD_17	148	2	1.33
MOD_03	1243	16	1.27
MOD_07	80	1	1.24
MOD_23	91	1	1.09
OUTROS	58672	566	0.95
MOD_14	12568	80	0.63
MOD_13	325	2	0.61
MOD_10	127179	777	0.61
MOD_19	40734	204	0.50
MOD_05	840	4	0.47
MOD_06	7960	28	0.35
MOD_15	8619	25	0.29
MOD_26	30087	72	0.24
MOD_18	18571	44	0.24
MOD_22	16994	40	0.23
MOD_21	4579	9	0.20
MOD_16	1816	3	0.16
MOD_04	39316	62	0.16
MOD_09	71199	70	0.10
MOD_02	35221	33	0.09
MOD_01	15	0	0.00
MOD_08	63	0	0.00
MOD_11	43	0	0.00
MOD_12	19	0	0.00
MOD_20	5	0	0.00
MOD_25	26	0	0.00
MOD_27	3	0	0.00

Tabela 10 – Inadimplência por Região.

Região	MOB0	MOB1	Tx_Inadimp
Nordeste	96070	721	0.74
Centro Oeste	48776	272	0.56
Sul	106519	559	0.52
Norte	26191	108	0.41
Sudeste sem SP	69151	155	0.22
Sao Paulo	129903	228	0.17

análise tão minuciosa de todas as variáveis antes de iniciar o processo de modelagem.

3.2 Seleção de Variáveis

Como parte do desenvolvimento dos modelos, a primeira etapa envolve a seleção de variáveis, processo para o qual empregaremos a análise de significância por meio da

regressão logística. Este método foi escolhido devido à sua eficácia em identificar variáveis relevantes, um passo importante na construção de modelos preditivos. Após essa etapa inicial de seleção, prosseguiremos com técnicas mais avançadas, como *Random Forest* e *XGBoost*.

Antes da criação dos modelos preditivos, é realizado um tratamento dos dados, que inclui a aplicação da técnica de *downsampling*. O *downsampling* é adotado porque modelos de regressão logística tendem a performar melhor quando as classes do conjunto de dados são equilibradas. Esta técnica ajuda a reduzir o desequilíbrio entre as classes, diminuindo a prevalência da classe majoritária para se equiparar à minoritária, o que pode melhorar a generalização do modelo ao reduzir o viés causado por desequilíbrios.

Após a aplicação do *downsampling*, procede-se com a análise do *valor-p* das variáveis por meio da regressão logística. Esta análise é usada para identificar quais variáveis possuem influência significativa no modelo, permitindo assim uma seleção de características relevantes para a predição. Variáveis com *valores-p* baixos são indicativas de uma associação estatisticamente significativa com a variável resposta, e, portanto, são mantidas no modelo final.

Os tratamentos aplicados às variáveis antes da modelagem foram focados principalmente na remoção de *outliers* e na eliminação de variáveis com alta correlação. A remoção de *outliers* é feita para evitar distorções nos resultados do modelo, já que esses valores extremos podem influenciar de maneira desproporcional as previsões. Por outro lado, a eliminação de variáveis com alta correlação visa reduzir a multicolinearidade, garantindo que o modelo não inclua informações redundantes que poderiam comprometer a interpretação dos efeitos de cada variável.

Com uma compreensão inicial sobre o comportamento das variáveis e após a aplicação dos devidos tratamentos, estamos prontos para construir o modelo logístico e identificar as variáveis significativas. Para assegurar a confiabilidade dos resultados, o modelo será repetido 20.000 vezes. Essa abordagem é necessária devido ao uso do *downsampling*, que envolve a seleção de uma amostra aleatória dos casos onde a variável resposta é mais frequente. Ao repetir o processo várias vezes e registrar os resultados de todos os modelos, aumentamos significativamente a confiança de que os resultados obtidos são representativos da base de dados como um todo. Este método nos permite minimizar os efeitos do acaso na seleção da amostra, garantindo assim uma maior estabilidade e precisão nas estimativas do modelo.

As tabelas 11, 12, 13 e 14 apresentam as médias dos *valores-p* obtidos ao longo das 20 mil repetições do modelo. Com base nesta tabela, selecionamos as variáveis significativas: ‘PERC_PRESTACAO’, ‘AT013_PERCENTUAL_ENTRADA_VEICU’, ‘AT014_IDADE’, ‘Autonomos’, ‘SERASA_TITULAR’ e ‘comp_renda’. A escolha dessas variáveis é corroborada tanto pela análise descritiva quanto pelos resultados da simulação, indicando uma coerência entre a observação inicial do comportamento das variáveis e

Tabela 11 – Significância das variáveis 1 de 4

Variable	MeanPrZ	Significativas
(Intercept)	0.088240	Não
PERC_PRESTACAO	0.000008	Sim
AT013_PERCENTUAL_ENTRADA_VEICU	0.000000	Sim
AT014_IDADE	0.000000	Sim
TAXISTASS	0.026113	Sim
AutonomosS	0.000333	Sim
AT014_COD_SERASA_TITULAR2	0.006380	Sim
AT014_COD_SERASA_TITULAR3	0.079054	Não
AT014_COD_SERASA_TITULAR4	0.009415	Sim
AT014_COD_SERASA_TITULARN	0.000000	Sim
AT014_COD_SERASA_TITULARY	0.096116	Não
AT014_COD_SERASA_CONJUGE0	0.007655	Sim
AT014_COD_SERASA_CONJUGE2	0.093846	Não

Tabela 12 – Significância das variáveis 2 de 4

Variable	MeanPrZ	Significativas
AT014_COD_SERASA_CONJUGE3	0.098054	Não
AT014_COD_SERASA_CONJUGE4	0.007897	Sim
AT014_COD_SERASA_CONJUGEN	0.000000	Sim
AT014_IND_POSSUI_VEICN	0.000001	Sim
AT014_IND_POSSUI_VEICS	0.006001	Sim
MARCA_TRATADAOUTRAS MARCAS	0.000000	Sim
MARCA_TRATADA_A	0.060676	Não
UF_ConcessionariaAL	0.089943	Não
UF_ConcessionariaAM	0.024681	Sim
UF_ConcessionariaAP	0.066690	Não
UF_ConcessionariaBA	0.047147	Sim
UF_ConcessionariaCE	0.068038	Não
UF_ConcessionariaDF	0.064881	Não

os resultados estatísticos significativos obtidos. Esta congruência reforça a validade das variáveis escolhidas para a modelagem final, assegurando que o modelo logístico reflete adequadamente as tendências e características da base de dados.

Tabela 13 – Significância das variáveis 3 de 4

Variable	MeanPrZ	Significativas
UF_ConcessionariaES	0.089933	Não
UF_ConcessionariaGO	0.060988	Não
UF_ConcessionariaMA	0.084294	Não
UF_ConcessionariaMG	0.072157	Não
UF_ConcessionariaMS	0.047834	Sim
UF_ConcessionariaMT	0.029847	Sim
UF_ConcessionariaPA	0.060361	Não
UF_ConcessionariaPB	0.071145	Não
UF_ConcessionariaPE	0.090534	Não
UF_ConcessionariaPI	0.063710	Não
UF_ConcessionariaPR	0.075971	Não
UF_ConcessionariaRJ	0.052125	Não
UF_ConcessionariaRN	0.050387	Não

Tabela 14 – Significância das variáveis 4 de 4

Variable	MeanPrZ	Significativas
UF_ConcessionariaRO	0.098181	Não
UF_ConcessionariaRR	0.050668	Não
UF_ConcessionariaRS	0.062406	Não
UF_ConcessionariaSC	0.080961	Não
UF_ConcessionariaSE	0.097410	Não
UF_ConcessionariaSP	0.030747	Sim
UF_ConcessionariaTO	0.046079	Sim
comp_renda	0.000000	Sim

3.3 Modelos Testados

3.3.1 Regressão Logística

Para a realização do modelo de regressão logística, foi utilizada uma base de dados contendo informações até maio de 2023, na qual foi realizada a remoção de *outliers* e aplicado o *downsampling* para equilibrar as classes, e ultilizando as variaveis previamente selecionadas. A robustez do modelo foi testada utilizando uma base de dados *"out of time"*, que compreende dados não incluídos no treinamento, especificamente das safras de junho de 2023 em diante. Esta mesma técnica, com algumas poucas alterações que serão mencionadas, foi utilizada para todos os outros modelos. Esta abordagem permite avaliar a capacidade do modelo de generalizar e performar bem em novos conjuntos de dados, garantindo que as previsões sejam confiáveis e aplicáveis a situações reais fora do contexto do conjunto de dados de treinamento.

O resumo do modelo de regressão logística apresentado revela que as variáveis

PERC_PRESTACAO, *AT013_PERCENTUAL_ENTRADA_VEICU*, *AT014_IDADE*, *Autonomos* e *SERASA_TITULAR* são estatisticamente significativas, com *valores-p* extremamente baixos, indicando forte associação com a variável resposta *MOB4*. Notavelmente, a variável *AT013_PERCENTUAL_ENTRADA_VEICU* tem o maior impacto negativo, como evidenciado pelo seu *valor-z* significativamente alto. O modelo ajustado exibe uma boa adequação, com uma diferença notável entre a *deviance* nula e a residual e um *AIC* de 4023.5, sugerindo que o modelo captura bem a variabilidade dos dados com as variáveis selecionadas.

```
##
## Call:
## glm(formula = MOB4 ~ ., family = binomial(), data = dados_treino)
##
## Coefficients:
##                                Estimate Std. Error z value
## (Intercept)                     3.95463    0.37865  10.444
## PERC_PRESTACAO                  0.39806    0.10756   3.701
## AT013_PERCENTUAL_ENTRADA_VEICU -0.93902    0.05733 -16.378
## AT014_IDADE                    -0.37347    0.03918  -9.532
## AutonomosS                      0.55158    0.16202   3.404
## comp_renda                    144.86475   14.69982   9.855
## SERASA_TITULAR1                -0.99030    0.11522  -8.595
##                                           Pr(>|z|)
## (Intercept)                    < 0.0000000000000002 ***
## PERC_PRESTACAO                              0.000215 ***
## AT013_PERCENTUAL_ENTRADA_VEICU < 0.0000000000000002 ***
## AT014_IDADE                    < 0.0000000000000002 ***
## AutonomosS                                  0.000663 ***
## comp_renda                     < 0.0000000000000002 ***
## SERASA_TITULAR1                < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4827.1  on 3481  degrees of freedom
## Residual deviance: 4009.5  on 3475  degrees of freedom
## AIC: 4023.5
##
## Number of Fisher Scoring iterations: 4
```

Para analisar a performance do modelo, utilizou-se o gráfico de distribuição cumulativa, que é mostrado na figura 10, que é eficaz para visualizar como diferentes pontos de corte do modelo afetam a captura do *target* e a proporção da população necessária para alcançar esses resultados. Este gráfico ilustra a porcentagem do target que o modelo consegue distinguir em vários pontos de corte, bem como a porcentagem da população que é requerida para chegar a esse resultado.

Por exemplo, no ponto de corte de 0.5, o modelo consegue capturar 82.5% dos casos com $MOB4 = 1$, utilizando apenas 33% da população. Isso indica que o modelo é bastante eficiente, pois com um terço da população, ele consegue identificar a maioria dos casos positivos. A linha vermelha ("Propostas") e a linha azul ("*Target*") no gráfico ajudam a visualizar esta eficiência em diferentes níveis de corte, facilitando a compreensão de como ajustar o modelo para maximizar sua precisão.

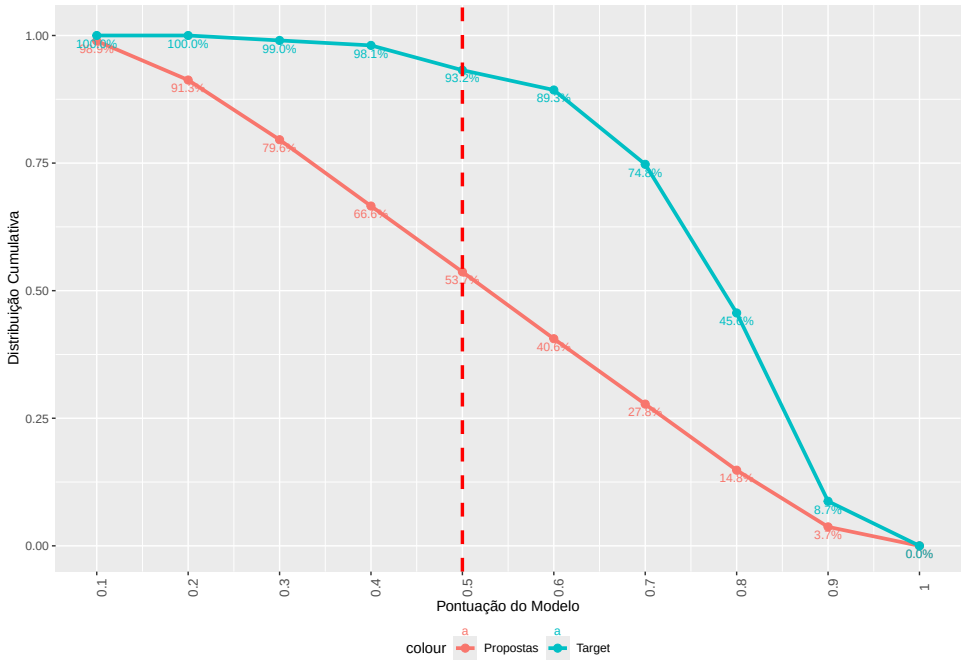

Figura 10 – Distribuição cumulativa da regressão logística.

Para avaliar a eficácia do modelo ao longo do tempo, foram utilizado o *Coeficiente KS (Kolmogorov-Smirnov)*. Esta métrica mede a maior distância vertical entre as curvas cumulativas de frequência das classes positivas e negativas, oferecendo uma medida direta da capacidade do modelo de separar essas classes. Valores de *KS* variam de 0 a 1, com valores maiores indicando melhor capacidade de discriminação. Um valor de *KS* acima de 0.2 é geralmente aceitável, acima de 0.3 é considerado bom, e acima de 0.5 indica uma excelente capacidade discriminatória do modelo. Conforme o gráfico da figura 11, notamos que o modelo apresentou ótimos valores, reforçando sua performance.

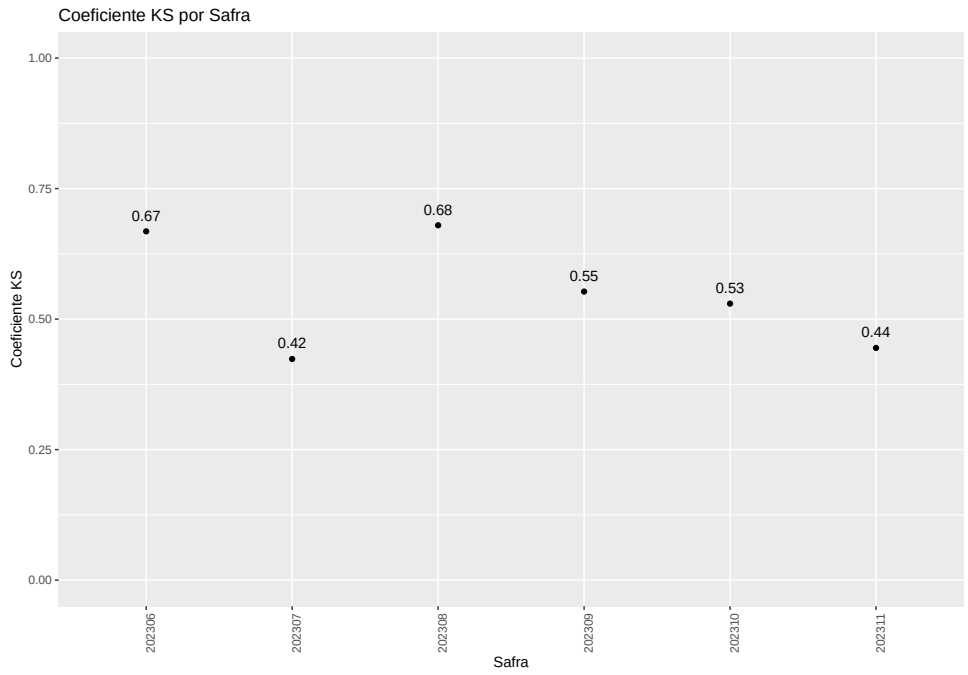

Figura 11 – Coeficiente KS por safra *out of time* do modelo logístico.

3.3.2 XGBoost

Para implementar o modelo *XGBoost*, o processo adotado foi semelhante ao utilizado na regressão logística, porém com algumas diferenças específicas relacionadas à natureza do modelo e aos tratamentos de dados necessários para otimizar sua performance. Uma das principais alterações foi a escolha do *upsampling* em vez do *downsampling*. Diferente do *downsampling*, que reduz o número de observações da classe majoritária, o *upsampling* replica aleatoriamente os dados da classe minoritária até que as duas classes tenham um número semelhante de observações. Esta abordagem é útil para modelos como o *XGBoost*, que podem ser sensíveis a desequilíbrios de classe e beneficiam-se de um maior volume de dados para aprender padrões mais complexos sem perder informações importantes.

Outra diferença no *XGBoost* é que é necessario a seleção e otimização de parâmetros, e para identificar os melhores valores desses parâmetros, foi realizada uma simulação. Esta simulação realizada para otimizar os parâmetros do modelo *XGBoost* envolve uma técnica conhecida como busca em grade (*grid search*). Essa abordagem é usada para explorar várias combinações de parâmetros a fim de identificar a configuração que maximiza a performance do modelo em termos de uma métrica específica, neste caso, a *sensibilidade*.

A ideia geral do *grid search* é criar um espaço multidimensional de parâmetros, onde cada dimensão representa um parâmetro específico do modelo que influencia seu comportamento e performance. Para cada ponto neste espaço, uma combinação específica dos valores dos parâmetros, o modelo é treinado e avaliado contra um conjunto de validação.

A métrica de performance calculada para cada combinação de parâmetros permite uma comparação objetiva entre diferentes configurações.

Ao final do processo de simulação, a combinação de parâmetros que resulta na melhor métrica de performance é selecionada como a configuração ideal. Este método é particularmente útil para modelos como o *XGBoost*, onde a interação entre os parâmetros pode ser não-linear e o impacto de cada parâmetro na performance do modelo pode ser difícil de prever intuitivamente. O *grid search* assegura que a seleção dos parâmetros seja abrangente e baseada em evidências empíricas, maximizando as chances de alcançar um modelo eficaz e otimizado.

O resultado do modelo é apresentado na figura 12 com o gráfico de distribuição cumulativa e na figura 13 que mostra o coeficiente *KS* nas safras "*out of time*".

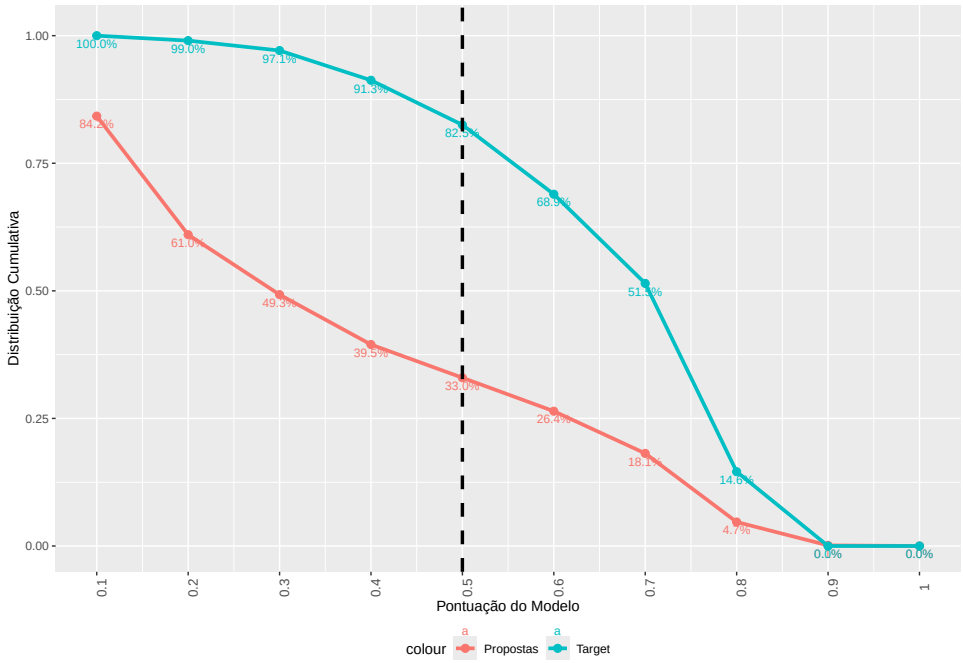

Figura 12 – Distribuição cumulativa do modelo *XGBoost*.

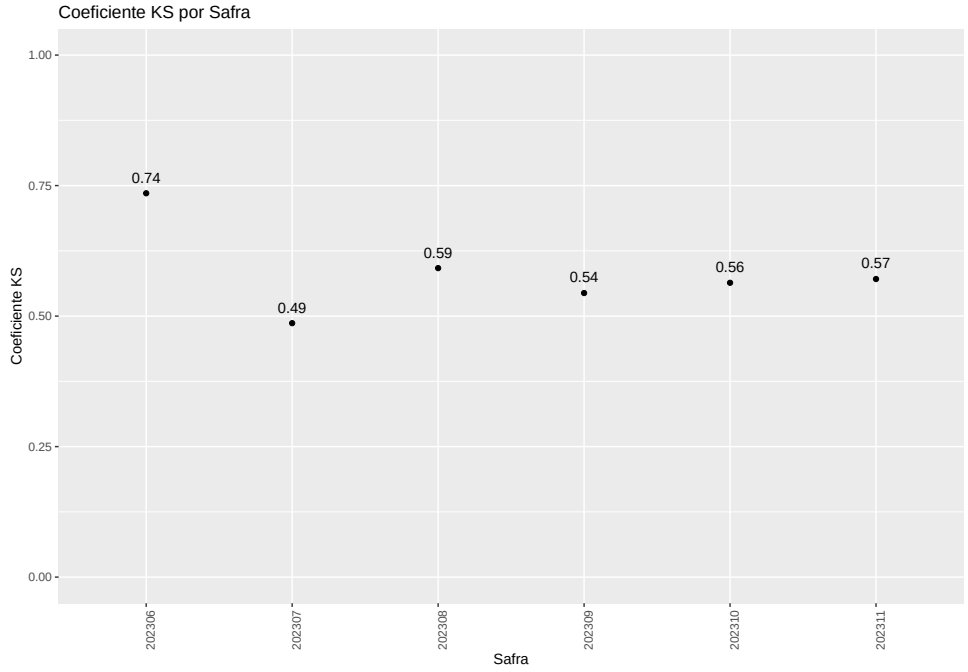

Figura 13 – Coeficiente KS por safra *out of time* do modelo *XGBoost*.

3.3.3 RANDOM FOREST

No caso do modelo de *Random Forest*, o processo de otimização de parâmetros também empregou a técnica de *grid search* similar ao utilizado para o *XGBoost*. No entanto, uma diferença foi a escolha de implementar o *downsampling* ao invés do *upsampling*. O *downsampling* foi bastante eficaz no caso do *Random Forest*. Essa abordagem não apenas melhorou a performance do modelo ao lidar com um conjunto de dados mais balanceado, mas também resultou em um tempo de treinamento significativamente menor, tornando o processo mais eficiente, uma vez que o custo computacional do algoritmo *Random Forest* é bem elevado. Os parametros ultilizados nesse caso foram *ntree*, *mtry* e *nodesize*.

O resultado do modelo é apresentado na figura 14 com o gráfico de distribuição cumulativa e na figura 15 que mostra o coeficiente *KS* nas safras "*out of time*".

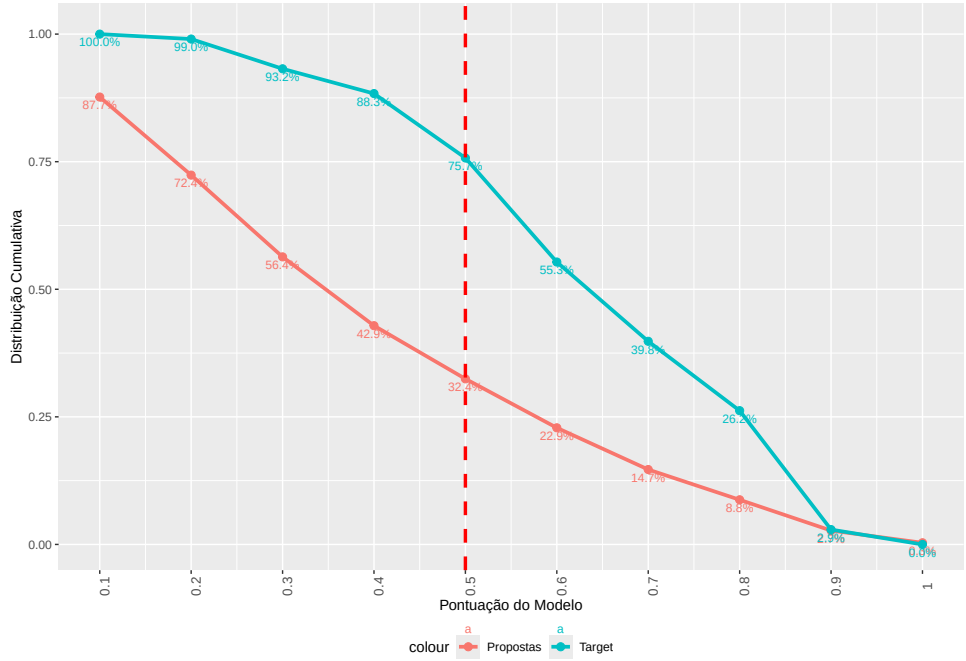

Figura 14 – Distribuição cumulativa do modelo *Random Forest*.

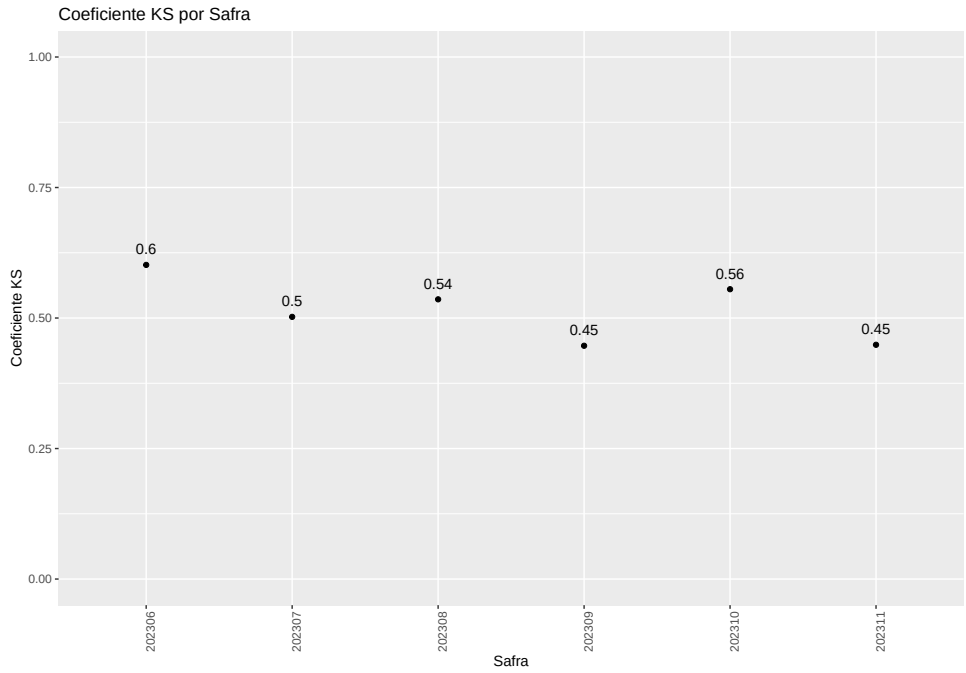

Figura 15 – Coeficiente KS por safra *out of time* do modelo *Random Forest*.

3.4 Comparações entre os Modelos

Os resultados obtidos revelam que o modelo *XGBoost* exibiu o desempenho mais destacado. Para ilustrar essa comparação, empregamos o gráfico de distribuição cumulativa. No contexto de modelagem de crédito, um limiar aceitável de sacrifício de propostas é de aproximadamente 15%. Portanto, esse valor foi selecionado como referência no nosso gráfico.

Na análise da regressão logística, ao sacrificar cerca de 15% das propostas, capturamos aproximadamente 46% da variável alvo. Em contraste, com o modelo *XGBoost*, o mesmo percentual de propostas resulta na captura de cerca de 50% da variável alvo. Já o modelo *Random Forest* apresenta uma performance inferior, alcançando aproximadamente 40% da variável alvo com os mesmos 15% de propostas "sacrificadas".

Portanto, com base nos critérios adotados, o *XGBoost* demonstrou ser superior, seguido pela *Regressão Logística* e, por último, o *Random Forest*.

4 Considerações Finais

O objetivo deste estudo foi avaliar a aplicabilidade e eficácia de três técnicas distintas de modelagem estatística em um conjunto real de dados de crédito, com o propósito de identificar qual apresentava a melhor performance. A metodologia adotada envolveu três fases: análise exploratória de dados, seleção de variáveis e aplicação de modelos estatísticos.

A fase de análise exploratória foi feita com muito cuidado, com um exame detalhado para cada variável disponível. Este processo é fundamental não apenas para compreender o comportamento da base de dados, mas também para extrair informações relevantes que podem ser úteis, mesmo quando as variáveis não são diretamente utilizadas na modelagem final. No setor bancário, entender as nuances e influências das variáveis é quase tão vital quanto construir um modelo robusto.

A seleção de variáveis representou o segmento mais desafiador e crucial da modelagem. Durante essa etapa, a regressão logística demonstrou ser uma ferramenta valiosa, reafirmando sua relevância, mesmo sendo uma técnica tradicional. Foram empregadas múltiplas metodologias para assegurar a significância das variáveis escolhidas. Combinada com a análise descritiva, esta abordagem proporcionou uma sólida confiança nas decisões tomadas sobre as variáveis a serem incluídas nos modelos finais.

O modelo *XGBoost* destacou-se com o melhor desempenho, o que era esperado dado o seu estado contemporâneo e sua popularidade na modelagem de crédito. Em seguida, a regressão logística demonstrou sua robustez, reforçando sua adequação neste contexto. O modelo *Random Forest*, apesar de não ter alcançado um desempenho tão expressivo, ainda assim se mostrou competente.

No universo financeiro, cada potencial inadimplemento pode representar custos significativos, tornando crucial a escolha de modelos com maior assertividade, mesmo que o incremento em precisão seja modesto. No entanto, isso não elimina a relevância de explorar outros métodos. O processo empregado neste estudo provou ser robusto e abrangente, oferecendo uma estrutura que pode ser facilmente adaptada a outros contextos analíticos.

Embora um modelo tenha se destacado sobre os demais, todos provaram sua utilidade, especialmente a regressão logística, que mostrou fortes indicativos de sua capacidade em contextos semelhantes. Este estudo incorporou integralmente o conhecimento adquirido ao longo do curso universitário em modelagem estatística, apresentando uma metodologia eficaz para abordar problemas similares. Assim, mesmo que o *XGBoost* tenha sido o mais assertivo, a experiência com os outros modelos foi igualmente valiosa e instrutiva.

Referências

CHEN, T. et al. *xgboost: Extreme Gradient Boosting*. [S.l.], 2024. R package version 1.7.7.1. Disponível em: <https://CRAN.R-project.org/package=xgboost>.

CHEN TIANQI; GUESTRIN, C. *XGBoost: A Scalable Tree Boosting System*. [S.l.], 2016. R package version 1.7.7.1. Disponível em: <https://xgboost.readthedocs.io/en/stable/>.

DAHL, D. B. et al. *xtable: Export Tables to LaTeX or HTML*. [S.l.], 2019. R package version 1.8-4. Disponível em: <https://CRAN.R-project.org/package=xtable>.

Kuhn; Max. Building predictive models in r using the caret package. *Journal of Statistical Software*, v. 28, n. 5, p. 1–26, 2008. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v028i05>.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002. Disponível em: <https://CRAN.R-project.org/doc/Rnews/>.

LUNARDON, N.; MENARDI, G.; TORELLI, N. ROSE: a Package for Binary Imbalanced Learning. *R Journal*, v. 6, n. 1, p. 82–92, 2014.

THOMAS, L. C. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, v. 16, n. 2, p. 149–172, 2000. ISSN 0169-2070. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169207000000340>.

VENABLES, W. N.; RIPLEY, B. D. *Modern Applied Statistics with S*. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: <https://www.stats.ox.ac.uk/pub/MASS4/>.

WEI, T.; SIMKO, V. *R package 'corrplot': Visualization of a Correlation Matrix*. [S.l.], 2021. (Version 0.92). Disponível em: <https://github.com/taiyun/corrplot>.

WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: <https://ggplot2.tidyverse.org>.

WICKHAM, H. et al. *dplyr: A Grammar of Data Manipulation*. [S.l.], 2023. R package version 1.1.4. Disponível em: <https://CRAN.R-project.org/package=dplyr>.

WICKHAM, H.; VAUGHAN, D.; GIRLICH, M. *tidyr: Tidy Messy Data*. [S.l.], 2024. R package version 1.3.1. Disponível em: <https://CRAN.R-project.org/package=tidyr>.