SEVEN
publicações acadêmicas

TÉCNICAS PREDITIVAS E MODELOS DE INTELIGÊNCIA ARTIFICIAL NA GESTÃO DE DÍVIDAS PÚBLICAS INADIMPLENTES: COMPARAÇÃO ENTRE REGRESSÃO LINEAR E ÁRVORES DE DECISÃO

https://doi.org/10.56238/sevened2024.031-019

Eduardo Silva Vasconcelos
Doutor em Ciências – Processamento da Informação
Instituto Federal Goiano
Goiânia, Goiás, Brasil
E-mail: educelos1@gmail.com
LATTES: http://lattes.cnpq.br/5128388060472259

RESUMO
O estudo analisado tem como foco a aplicação de modelos preditivos, especificamente Regressão Linear e Árvores de Decisão, para a gestão de dívidas inadimplentes no contexto público dos Estados Unidos. O objetivo central do trabalho é comparar a eficácia desses modelos na previsão da conformidade de dívidas com mais de 120 dias, auxiliando no direcionamento dessas dívidas ao Treasury Offset Program (TOP), uma iniciativa essencial para a recuperação financeira governamental. O problema que o estudo aborda é a necessidade de uma gestão eficaz das dívidas públicas inadimplentes, buscando garantir o cumprimento de políticas financeiras públicas que promovam a conformidade e o redirecionamento adequado dos recursos financeiros ao governo. Isso é particularmente importante para garantir a transparência fiscal e a responsabilização das agências federais. A metodologia utilizada no estudo foi quantitativa, baseada na análise de dados de dívidas elegíveis extraídos de relatórios do Tesouro dos EUA. Foram aplicados os modelos de Regressão Linear e Árvores de Decisão, com métricas de desempenho como Erro Absoluto Médio (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2). O estudo tratou de variáveis financeiras e temporais para analisar o comportamento dessas dívidas e sua conformidade. Os principais resultados mostram que ambos os modelos apresentaram alta precisão nas previsões, com a Regressão Linear mostrando um ajuste perfeito ($R^2 = 1$) e as Árvores de Decisão destacando-se na captura de nuances não lineares dos dados. A variável "Compliance Rate Amount" foi identificada como a mais significativa no modelo de Árvores de Decisão, sugerindo que o montante da taxa de conformidade é um dos fatores mais importantes para prever a conformidade das dívidas inadimplentes. Este estudo oferece contribuições valiosas para o campo da gestão pública, ao demonstrar que a utilização de modelos preditivos pode auxiliar na otimização da recuperação de dívidas, melhorar a transparência fiscal e contribuir para a tomada de decisões mais informadas.

Palavras-chave: Gestão Financeira Pública. Modelagem Preditiva. Dívidas Inadimplentes. Inteligência Artificial Aplicada.

1 INTRODUÇÃO

A administração pública, especialmente nos Estados Unidos, enfrenta o desafio da gestão eficiente de dívidas inadimplentes. De acordo com a Lei de Responsabilidade e Transparência Digital de 2014 (Lei de DADOS), as dívidas inadimplentes com mais de 120 dias devem ser encaminhadas ao Treasury Offset Program (TOP) para garantir a recuperação de receitas ao governo federal. O não

cumprimento dessas diretrizes pode prejudicar a capacidade do governo de financiar serviços públicos essenciais, além de comprometer a transparência e a responsabilização das agências federais (TESOURO DOS EUA, 2024).

Modelos preditivos como a Regressão Linear e as Árvores de Decisão surgem como ferramentas importantes para melhorar a conformidade no encaminhamento de dívidas. Essas técnicas permitem prever quais dívidas têm maior probabilidade de serem encaminhadas, auxiliando na otimização dos processos de recuperação financeira. O estudo se concentra na aplicação dessas técnicas de machine learning para identificar variáveis significativas e prever a conformidade com a Lei de DADOS.

O presente estudo tem como objetivo geral comparar a eficácia dos modelos de Regressão Linear e Árvores de Decisão na previsão da conformidade no encaminhamento de dívidas inadimplentes de 120 dias nos Estados Unidos. A partir desse objetivo central, o estudo desdobra-se em quatro objetivos específicos. Primeiramente, busca-se comparar a eficácia desses modelos utilizando métricas de avaliação como Erro Absoluto Médio (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2). Tais métricas são amplamente empregadas para medir a precisão preditiva dos modelos e identificar o quão bem cada um deles se ajusta aos dados analisados.

Em segundo lugar, o estudo pretende identificar as variáveis mais significativas para a previsão da conformidade, considerando quais fatores têm maior impacto nos resultados preditivos de cada modelo. Isso permitirá uma compreensão mais aprofundada dos elementos-chave que influenciam o comportamento das dívidas inadimplentes e sua conformidade no contexto analisado.

O terceiro objetivo visa analisar como esses modelos podem contribuir para melhorar a eficiência e a transparência na gestão de dívidas inadimplentes. A aplicação eficaz de modelos preditivos pode otimizar processos, reduzir ineficiências e fornecer uma visão mais clara sobre o comportamento das dívidas, o que é essencial para a formulação de políticas públicas.

Por fim, o estudo busca fornecer informações estratégicas que permitam aos gestores públicos tomar decisões baseadas em evidências. A utilização de modelos preditivos robustos pode proporcionar insights valiosos para a gestão financeira, auxiliando na tomada de decisões informadas e promovendo uma gestão mais eficaz das dívidas inadimplentes, com impacto positivo na conformidade e na alocação de recursos públicos.

Este estudo é relevante para a administração pública, pois contribui para a otimização da recuperação de receitas e o aumento da transparência fiscal. A conformidade com a Lei de DADOS é essencial para assegurar a eficiência financeira dos órgãos federais, e a implementação de modelos preditivos avançados fornece ferramentas para prever com precisão o encaminhamento de dívidas, melhorando a gestão orçamentária.

A inadimplência de dívidas gera impactos significativos na administração financeira pública, afetando a capacidade de governos de alocar recursos e manter a estabilidade orçamentária. Conforme observado por Iudícibus (2010), uma gestão eficaz de dívidas inadimplentes exige o monitoramento rigoroso dos pagamentos e a adoção de políticas que incentivem o cumprimento das obrigações financeiras por parte dos devedores. Uma ferramenta central nesse processo nos Estados Unidos é o Treasury Offset Program (TOP), que desempenha um papel crucial na recuperação de dívidas. Esse programa permite que os recursos sejam redirecionados ao governo por meio da compensação de pagamentos, auxiliando na recuperação de montantes devidos e contribuindo para a sustentabilidade fiscal.

As técnicas de previsão são essenciais para a gestão eficaz de dívidas inadimplentes, fornecendo uma base sólida para antecipar comportamentos futuros e auxiliar na formulação de políticas. Entre essas técnicas, destacam-se os modelos de Regressão Linear e as Árvores de Decisão, amplamente utilizados na previsão financeira e no gerenciamento de inadimplência.

A Regressão Linear é uma técnica estatística que busca prever o valor de uma variável dependente com base em uma ou mais variáveis independentes. Segundo Montgomery, Peck e Vining (2012), esse modelo é particularmente adequado para situações em que há uma relação linear clara entre as variáveis, proporcionando uma abordagem simples, porém eficaz, para a previsão de resultados financeiros.

Além disso, Angrist e Pischke (2009) ressaltam que os modelos de regressão linear são fundamentais no campo da econometria, sendo utilizados como ferramentas computacionais para estimar as diferenças entre grupos tratados e grupos de controle, com ou sem o uso de covariáveis. Esse método é crucial na avaliação de intervenções e na mensuração de seus impactos, oferecendo controle preciso sobre os fatores que podem influenciar os resultados.

Por outro lado, as Árvores de Decisão são técnicas de machine learning que se destacam por sua capacidade de particionar os dados em subconjuntos homogêneos, criando uma estrutura hierárquica que facilita a tomada de decisões. De acordo com Breiman et al. (1984), as árvores de decisão são especialmente úteis quando existem relações complexas e não lineares entre as variáveis, como é frequentemente o caso na previsão de inadimplência de dívidas. Esse método permite identificar padrões ocultos nos dados e gerar previsões mais detalhadas e precisas.

As Árvores de Decisão são amplamente reconhecidas por sua facilidade de interpretação e aplicabilidade em diversas áreas. Como observado por Pérez et al. (2019), essa técnica é frequentemente utilizada no desenvolvimento de classificadores interpretáveis devido à sua estrutura visual, que se assemelha a um fluxograma.

A Inteligência Artificial (IA) tem se mostrado uma ferramenta poderosa na administração pública, especialmente no contexto de previsão financeira. Silva e Rocha (2019) destacam que a IA

pode ajudar a prever receitas e despesas, identificar devedores em potencial e otimizar estratégias de recuperação de dívidas. Além disso, a IA também pode melhorar a transparência e a responsabilização na administração pública (GOMES et al., 2020).

Outro estudo relevante é o de Vasconcelos, Santos e Amorim (2024), que exploram a utilização de algoritmos de otimização para melhorar a alocação de recursos no orçamento público. A pesquisa mostra que a aplicação de modelos preditivos baseados em IA pode aumentar a eficácia das políticas fiscais, garantindo que os recursos sejam direcionados para as áreas de maior necessidade e impacto social.

A incorporação da Inteligência Artificial (IA) na gestão pública tem se destacado como um dos avanços mais significativos para a modernização das entidades governamentais, especialmente no Brasil, onde a sua adoção tem o potencial de transformar a formulação e avaliação de políticas públicas, além de aprimorar o atendimento aos cidadãos. De acordo com Vasconcelos e Santos (2024), o estudo sobre a aplicação de IA no setor público é essencial, pois explora maneiras eficazes de integrar essa tecnologia nas administrações brasileiras, com o objetivo de otimizar processos e melhorar a qualidade dos serviços oferecidos à população.

2 METODOLOGIA

Este estudo emprega uma abordagem quantitativa, utilizando modelos estatísticos e de machine learning para prever a conformidade no encaminhamento de dívidas inadimplentes. A pesquisa é aplicada, visando fornecer insights práticos para gestores públicos sobre a eficácia dos modelos de Regressão Linear e Árvores de Decisão na gestão financeira.

Os dados utilizados para esta análise foram extraídos do Relatório de Conformidade de Encaminhamento de Dívidas Inadimplentes de 120 Dias, disponível no site do Tesouro dos EUA. Esse conjunto de dados inclui informações detalhadas sobre dívidas elegíveis, dívidas referidas e dívidas não encaminhadas.

Antes de aplicar os modelos preditivos, foi necessário preparar os dados. O tratamento envolveu a substituição de valores ausentes por médias ou medianas, bem como a codificação de variáveis categóricas para torná-las compatíveis com os algoritmos de machine learning (HAIR et al., 2019).

O modelo de Regressão Linear foi aplicado para capturar relações lineares entre variáveis, enquanto o modelo de Árvores de Decisão foi utilizado para identificar interações complexas e não lineares. Ambos os modelos foram treinados e avaliados usando métricas de desempenho, como MAE, MSE, RMSE e R^2 (BREIMAN et al., 1984; MONTGOMERY, PECK, VINING, 2012).

A eficácia dos modelos foi medida usando métricas estatísticas. O Erro Absoluto Médio (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R^2) foram utilizados para avaliar a precisão das previsões e a robustez dos modelos.

Para facilitar a compreensão dos dados utilizados, foi elaborado o Quadro 1, onde são apresentadas as variáveis originais, bem como suas respectivas traduções para o português. Isso assegura que a análise seja consistente e compreensível no contexto da gestão de dívidas inadimplentes.

Quadro 1: Quadro de Variáveis

Variável Original	Tradução para o Português Brasileiro
Total Eligible Debt Amount	Montante Total de Dívida Elegível
Total Eligible Debt Count	Contagem Total de Dívida Elegível
Eligible Debt Referred Amount	Montante de Dívida Elegível Encaminhada
Eligible Debt Referred Count	Contagem de Dívida Elegível Encaminhada
Eligible Debt Not Referred Amount	Montante de Dívida Elegível Não Encaminhada
Eligible Debt Not Referred Count	Contagem de Dívida Elegível Não Encaminhada
Compliance Rate Amount	Montante da Taxa de Conformidade
Compliance Rate Count	Contagem da Taxa de Conformidade
Fiscal Year	Ano Fiscal
Fiscal Quarter Number	Número do Trimestre Fiscal
Calendar Year	Ano Calendário
Calendar Quarter Number	Número do Trimestre Calendário
Calendar Month Number	Número do Mês Calendário
Calendar Day Number	Número do Dia Calendário

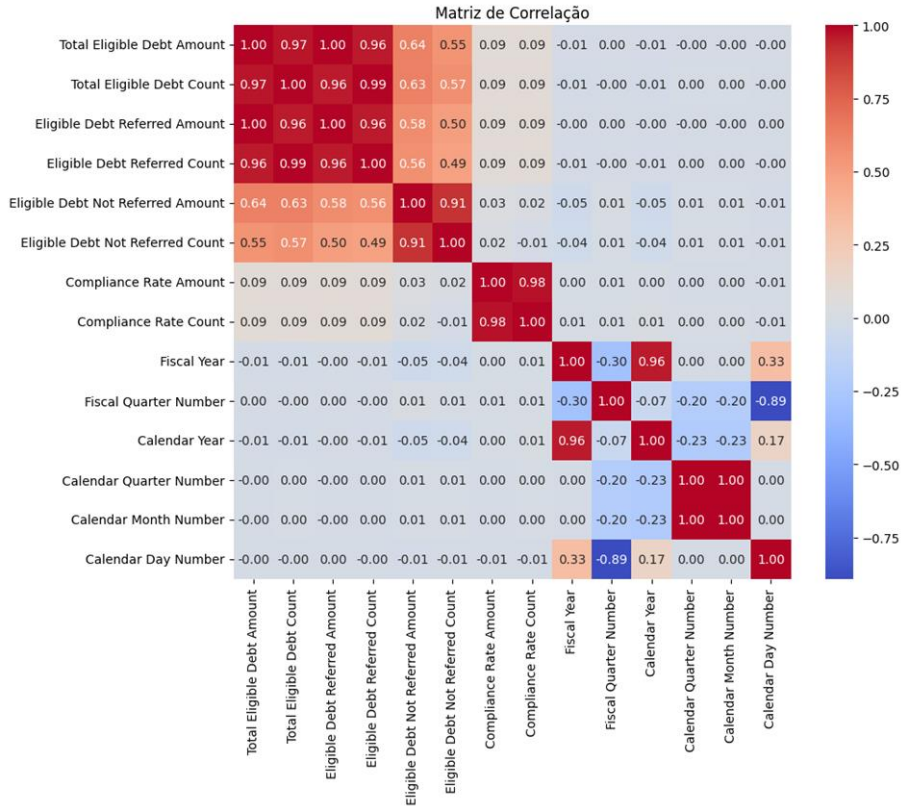
Fonte: Elaborado pelo autor (2024).

Essas variáveis representam tanto aspectos quantitativos, como montantes e contagens de dívidas, quanto fatores temporais, sendo fundamentais para a construção dos modelos preditivos.

3 MATRIZ DE CORRELAÇÃO

A Matriz de Correlação, Gráfico 1: Matriz de Correlação, foi construída com o objetivo de identificar as relações entre as variáveis numéricas do conjunto de dados. Essa análise permite entender como os diferentes aspectos das dívidas elegíveis e inadimplentes se correlacionam, proporcionando insights sobre as variáveis mais influentes para os modelos preditivos.

Gráfico 1: Matriz de Correlação



Fonte: Elaborado pelo autor (2024).

A "Matriz de Correlação" apresentada na imagem reflete a relação linear entre diversas variáveis associadas ao estudo de dívidas elegíveis e sua conformidade no encaminhamento de dívidas inadimplentes.

A análise da matriz de correlação revela importantes relações entre as variáveis financeiras e temporais, oferecendo insights valiosos sobre o comportamento das dívidas elegíveis e o processo de conformidade. Em primeiro lugar, destaca-se a correlação positiva forte entre variáveis financeiras, especialmente entre as variáveis Total Eligible Debt Amount e Total Eligible Debt Count, que apresentam uma correlação de 0,97. Essa relação sugere que, à medida que o número total de dívidas elegíveis aumenta, o montante total correspondente também cresce de forma proporcional. Tal comportamento é esperado, uma vez que um maior número de dívidas naturalmente resulta em um aumento do montante total devido. Um padrão semelhante é observado entre as variáveis Eligible Debt Referred Amount e Eligible Debt Referred Count, com uma correlação quase perfeita de 0,99. Essa interdependência indica que o número de dívidas referidas e o valor associado caminham juntos de maneira consistente, apontando para uma uniformidade no processo de encaminhamento das dívidas, onde tanto o número de encaminhamentos quanto o valor seguem a mesma tendência.

Além disso, é observada uma correlação moderada com dívidas não encaminhadas. A relação entre Eligible Debt Not Referred Amount e Eligible Debt Not Referred Count é de 0,64, um valor consideravelmente menor que o observado nas dívidas encaminhadas. Essa diferença sugere que o

processo de não encaminhamento de dívidas pode não estar perfeitamente alinhado em termos de montante e contagem, havendo flutuações nos valores que não correspondem diretamente ao número de dívidas não encaminhadas. Adicionalmente, existe uma correlação moderada de 0,98 entre Compliance Rate Amount e Compliance Rate Count, indicando que a conformidade, tanto em termos de valor quanto de volume de dívidas, está bem alinhada. Isso reforça a consistência no processo de medição da conformidade, considerando tanto os valores absolutos quanto o número de dívidas.

Outro ponto relevante é a correlação fraca ou nula com variáveis temporais. A matriz demonstra que não há correlações significativas entre as variáveis financeiras e as variáveis de calendário, como Fiscal Year, Calendar Year, Calendar Month Number e Calendar Day Number. A correlação entre essas variáveis oscila entre -0,01 e 0,05, sugerindo que fatores temporais, como o ano fiscal, o mês ou o dia, não exercem influência direta significativa sobre as variações nas dívidas elegíveis ou encaminhadas. A única exceção é uma correlação moderada de 0,33 entre Calendar Day Number e Fiscal Quarter Number, que pode ser explicada pela estrutura do calendário dentro de cada trimestre fiscal.

Por fim, a matriz também revela a existência de algumas correlações negativas. Por exemplo, há uma correlação de -0,30 entre Fiscal Quarter Number e Calendar Year, assim como de -0,20 entre Fiscal Quarter Number e Calendar Quarter Number. Essas relações sugerem que, à medida que o número do trimestre fiscal aumenta, o impacto sobre o ano ou trimestre do calendário pode diminuir, possivelmente devido a mudanças no período fiscal em relação ao calendário regular. Além disso, é observada uma correlação negativa de -0,04 entre Fiscal Quarter Number e Compliance Rate Count, indicando que, embora a relação seja muito fraca, a conformidade em termos de contagem de dívidas pode ser ligeiramente afetada pela estrutura dos trimestres fiscais. Contudo, essa correlação é tão pequena que sua relevância pode ser desconsiderada no contexto geral da análise.

A matriz confirma a interdependência esperada entre as variáveis de montante e contagem de dívidas, especialmente para dívidas encaminhadas, onde os valores de correlação são quase perfeitos. Além disso, a fraca correlação com variáveis temporais sugere que os processos de conformidade e encaminhamento de dívidas não estão diretamente vinculados ao calendário. Isso é importante para entender que fatores temporais, como o ano fiscal ou o mês, não desempenham um papel crítico no comportamento dessas variáveis financeiras.

A maior implicação dessa matriz está em identificar as variáveis que mais contribuem para o sucesso de previsões em modelos de conformidade. A forte correlação entre montantes e contagens de dívidas indica que esses fatores devem ser priorizados nos modelos preditivos, enquanto as variáveis temporais podem ser ignoradas ou minimizadas para evitar ruídos desnecessários no modelo.

A matriz de correlação foi utilizada para selecionar variáveis com maior impacto nos modelos. As variáveis com fortes correlações foram mantidas, enquanto as de menor relevância, como as

variáveis de calendário, foram descartadas para simplificar o modelo e minimizar o ruído nos dados. Essa abordagem garantiu maior precisão nos modelos preditivos.

4 REGRESSÃO LINEAR

O modelo de regressão linear foi aplicado para prever a conformidade no encaminhamento de dívidas inadimplentes com mais de 120 dias. A avaliação do modelo foi realizada utilizando métricas estatísticas que medem a precisão das previsões e a qualidade do ajuste do modelo aos dados observados.

Os resultados obtidos a partir das métricas de avaliação do modelo indicam sua alta precisão. As métricas calculadas foram: Erro Absoluto Médio (MAE) de 3,2054e-12, Erro Quadrático Médio (MSE) de 2,8218e-22, Raiz do Erro Quadrático Médio (RMSE) de 1,6798e-11 e Coeficiente de Determinação (R^2) igual a 1.0.

Esses valores mostram a performance do modelo, evidenciando a precisão das previsões realizadas. A seguir, os detalhes desses resultados serão discutidos.

O Erro Absoluto Médio (MAE), que mede a média dos erros absolutos entre os valores previstos e os valores reais, foi extremamente baixo, com um valor de 3,2054e-12. Essa métrica indica que, em média, a diferença entre as previsões do modelo e os valores observados é insignificante. No contexto da modelagem preditiva, um MAE tão baixo sugere que o modelo de regressão linear possui uma capacidade muito precisa de prever a conformidade das dívidas, minimizando discrepâncias entre os dados observados e as previsões geradas.

O Erro Quadrático Médio (MSE), que mede a média dos quadrados das diferenças entre os valores previstos e os reais, apresentou um valor de 2,8218e-22. O MSE penaliza erros maiores de forma mais severa, já que eleva ao quadrado as discrepâncias. O fato de esse valor ser extremamente pequeno indica que o modelo praticamente não comete erros significativos. Um MSE próximo de zero é um forte indicador de que o modelo está altamente ajustado aos dados de conformidade, prevendo com extrema exatidão.

O Root Mean Squared Error (RMSE), que é a raiz quadrada do MSE, fornece uma medida do erro que se encontra na mesma unidade dos valores previstos. Com um valor de 1,6798e-11, o RMSE também confirma que a magnitude dos erros de previsão é extremamente baixa, reforçando a ideia de que o modelo de regressão linear é altamente eficaz para capturar as variações entre as variáveis independentes e a conformidade no encaminhamento de dívidas. Esse valor indica que os desvios entre as previsões e os valores reais são praticamente inexistentes.

O Coeficiente de Determinação (R^2), com valor igual a 1.0, representa a capacidade do modelo de explicar toda a variabilidade presente nos dados observados. Em outras palavras, o modelo explica 100% da variação na variável dependente (conformidade das dívidas) com base nas variáveis

independentes utilizadas no modelo. Um valor de R^2 igual a 1.0 sugere um ajuste perfeito, o que significa que o modelo não deixa nenhuma variação sem ser explicada pelas variáveis selecionadas.

Com relação à precisão e eficácia do modelo, temos que as métricas de avaliação apresentadas (MAE, MSE e RMSE) mostram resultados extremamente baixos, indicando que o modelo de regressão linear possui uma precisão notável. A ausência de erros significativos reforça a robustez do modelo, que demonstra ser altamente eficaz na previsão de conformidade no encaminhamento de dívidas inadimplentes.

O R^2 de 1.0 sugere um ajuste perfeito entre os valores previstos e os observados. Esse resultado indica que o modelo é capaz de capturar completamente a relação entre as variáveis preditoras e a conformidade, sem deixar espaço para erros de previsão não explicados. Esse nível de precisão é raramente observado em práticas preditivas comuns, sugerindo uma relação forte entre as variáveis selecionadas e a variável-alvo.

Embora os resultados obtidos demonstrem uma alta precisão, um ajuste perfeito (como o R^2 de 1.0) pode levantar preocupações sobre overfitting, onde o modelo se ajusta excessivamente aos dados de treinamento, comprometendo sua capacidade de generalização para novos dados. Contudo, se os dados forem bem particionados entre treinamento e teste e forem representativos, esses resultados são extremamente promissores e indicam que o modelo pode ser confiável em aplicações práticas.

A alta precisão do modelo de regressão linear sugere que ele pode ser amplamente aplicado para prever a conformidade no encaminhamento de dívidas inadimplentes, oferecendo suporte valioso para a tomada de decisões gerenciais e estratégicas. Ao capturar com precisão as variáveis que influenciam o comportamento de conformidade, o modelo permite que os gestores financeiros implementem políticas e estratégias de gestão de dívidas mais eficazes.

Além disso, a análise detalhada das métricas oferece uma base sólida para comparar o desempenho do modelo de regressão linear com outros métodos preditivos, como a árvore de decisão. Isso pode ser fundamental na escolha do modelo mais adequado para diferentes cenários de previsão, especialmente em ambientes de grande variabilidade.

4.1 ANÁLISE DE DISPERSÃO – REGRESSÃO LINEAR

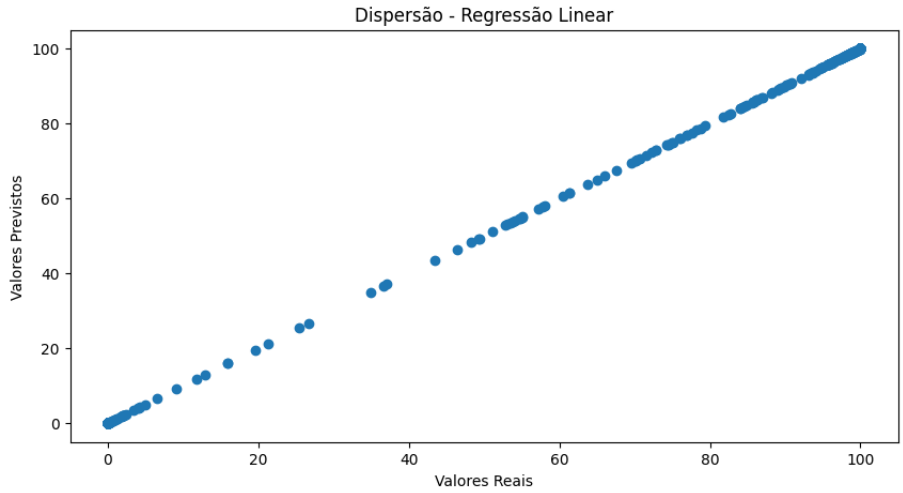
A análise de dispersão na regressão linear é uma ferramenta estatística essencial para avaliar a relação entre as variáveis independentes e a variável dependente. Esta ferramenta permite uma visualização clara da precisão das previsões realizadas pelo modelo, destacando o quão bem os valores previstos se alinham com os valores observados.

O gráfico de dispersão ilustra a relação entre os valores reais e os valores previstos pelo modelo de árvore de decisão. O gráfico de dispersão é fundamental para visualizar a capacidade do modelo em capturar a variabilidade dos dados e fazer previsões precisas. Os pontos devem se alinhar ao longo de

uma linha diagonal que representa a perfeita correspondência entre as previsões e os valores reais. A análise deste gráfico permite identificar desvios sistemáticos, avaliar a precisão das previsões em diferentes faixas de valores e detectar possíveis outliers. Além disso, a dispersão fornece insights sobre a robustez do modelo em lidar com variações não lineares nos dados, uma característica distintiva das árvores de decisão (BREIMAN et al., 1984; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Conforme ilustrado no Gráfico 2: Dispersão – Regressão Linear, observa-se a dispersão dos valores previstos em relação aos valores reais, permitindo a identificação de padrões subjacentes, bem como a detecção de discrepâncias ou anomalias nos dados.

Gráfico 2: Dispersão – Regressão Linear



Fonte: Elaborado pelo autor (2024).

O Gráfico 2: Dispersão – Regressão Linear apresenta uma análise visual fundamental para avaliar o desempenho do modelo de regressão linear. A presença de um alinhamento claro dos pontos ao longo da linha de tendência indica que o modelo produziu previsões que praticamente coincidem com os dados reais, o que sugere um ajuste preciso do modelo às variáveis analisadas.

No gráfico, observa-se que os pontos estão quase completamente alinhados a essa linha, o que confirma que o modelo de regressão linear teve um desempenho altamente preciso na maioria dos casos. Essa proximidade entre os valores previstos e observados reforça que o modelo foi capaz de captar com elevado grau de acurácia a relação entre as variáveis explicativas e a variável dependente, gerando previsões confiáveis.

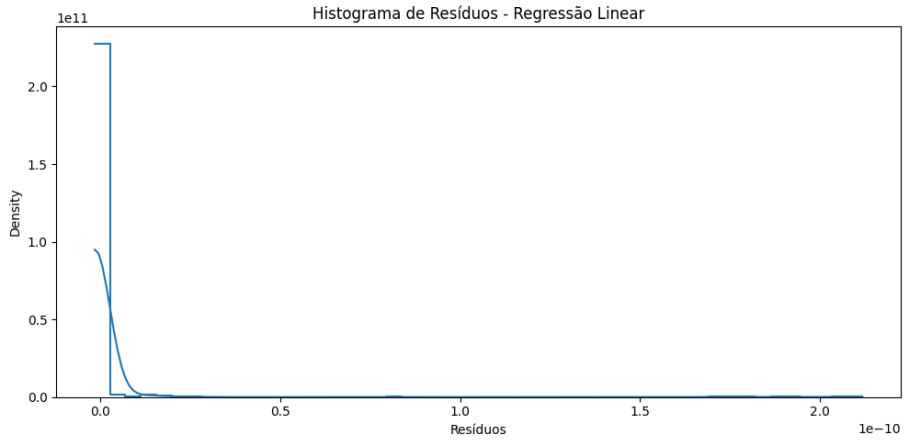
A análise visual proporcionada pelo gráfico de dispersão confirma a eficácia do modelo de regressão linear na previsão dos resultados. A forte correlação entre os valores reais e previstos demonstra que o modelo foi bem-sucedido em capturar as relações subjacentes entre as variáveis. Em combinação com métricas de desempenho como MAE, MSE, RMSE e R^2, discutidas anteriormente, o gráfico de dispersão oferece uma compreensão abrangente da precisão e robustez do modelo. A

ausência de discrepâncias relevantes valida a aplicabilidade do modelo para as previsões dentro do contexto do estudo.

4.2 HISTOGRAMA DE RESÍDUOS – REGRESSÃO LINEAR

O histograma de resíduos é uma ferramenta essencial para avaliar a adequação de um modelo de regressão linear, permitindo a visualização da distribuição dos erros de previsão em comparação com os valores reais. Ele auxilia na verificação das suposições do modelo, como a normalidade e independência dos erros. Em um modelo bem ajustado, os resíduos devem apresentar uma distribuição aproximadamente normal, concentrada em torno de zero, indicando ausência de viés sistemático e previsões imparciais. Além disso, o histograma facilita a identificação de outliers, heterocedasticidade e outros desvios que podem comprometer a validade do modelo.

Gráfico 3: Histograma de Resíduos – Regressão Linear



Fonte: Elaborado pelo autor (2024)

O Gráfico 3: Histograma de Resíduos – Regressão Linear oferece uma análise detalhada da distribuição dos resíduos gerados pelo modelo, permitindo uma avaliação precisa da qualidade do ajuste realizado. A distribuição dos resíduos é uma ferramenta importante para verificar a precisão das previsões, bem como para identificar possíveis áreas de melhoria no desempenho do modelo.

Ao analisar a distribuição dos resíduos, observa-se que a maioria está concentrada em torno de zero, o que indica que o modelo de regressão linear produziu previsões bastante precisas para a maior parte dos dados. No entanto, há uma leve assimetria na distribuição, com uma concentração maior de resíduos pequenos. Isso pode sugerir que, embora o modelo apresente um bom desempenho geral, ajustes adicionais podem ser necessários para corrigir essa assimetria e alcançar uma distribuição perfeitamente simétrica, o que potencialmente melhoraria a qualidade preditiva do modelo.

Nota-se a presença de outliers, principalmente visíveis na cauda direita do histograma. Esses outliers correspondem a resíduos maiores, indicando que, em alguns casos, as previsões se distanciaram dos valores reais observados. Isso sugere que o modelo pode não ter capturado

adequadamente certas relações entre as variáveis em situações específicas, apontando para a necessidade de uma investigação mais profunda sobre os fatores que causam essas discrepâncias.

A densidade dos resíduos ao redor de zero, por sua vez, é bastante alta, o que confirma que a maioria das previsões foi realizada com erros insignificantes. Essa alta concentração de resíduos pequenos é um indicativo positivo da eficácia do modelo, demonstrando que ele apresentou um desempenho satisfatório na maior parte dos casos, com poucas exceções.

A presença de outliers e a leve assimetria dos resíduos apontam para áreas onde o modelo pode ser aprimorado. Esses resultados sugerem que, embora o modelo seja amplamente robusto, há oportunidades para melhorar sua adequação aos dados, especialmente em relação a variáveis que podem ter efeitos não lineares ou a dados que podem estar impactando negativamente a distribuição dos resíduos.

5 ÁRVORE DE DECISÃO

O modelo de árvore de decisão foi empregado com o objetivo de prever a conformidade no encaminhamento de dívidas inadimplentes superiores a 120 dias. O desempenho do modelo é avaliado por meio de diversas métricas, que quantificam a precisão das previsões e a qualidade do ajuste aos dados observados. A seguir, discute-se cada uma dessas métricas em detalhes, proporcionando uma análise profunda sobre a eficácia e robustez do modelo.

As principais métricas de avaliação apresentaram os seguintes valores: Erro Absoluto Médio (MAE) de 0,0138, Erro Quadrático Médio (MSE) de 0,00692, Raiz do Erro Quadrático Médio (RMSE) de 0,0832 e Coeficiente de Determinação (R^2) de 0,99999.

Esses resultados serão discutidos em detalhes a seguir, com o objetivo de explicar a importância de cada métrica no contexto da análise realizada.

O valor obtido para o Erro Absoluto Médio (MAE) foi de 0,0138, uma métrica que mensura a média dos erros absolutos entre os valores previstos pelo modelo e os valores reais observados. Esse valor relativamente baixo indica o modelo de árvore de decisão apresentou um desempenho muito preciso, com erros médios pequenos ao longo de suas previsões. No contexto deste estudo, o MAE de 0,0138 significa que, em média, a diferença entre os valores reais e previstos pelo modelo é muito reduzida, o que demonstra a alta eficácia na antecipação da conformidade no encaminhamento de dívidas inadimplentes.

O Erro Quadrático Médio (MSE) foi de 0,00692, métrica que penaliza de forma mais severa erros maiores ao elevar as diferenças ao quadrado antes de somá-las. O fato de o MSE ser tão baixo reforça a conclusão de que o modelo faz previsões com elevada precisão, apresentando pouca discrepância entre os valores observados e os previstos. O MSE é particularmente útil para identificar

previsões discrepantes, pois penaliza erros maiores, o que sugere que não há grandes distorções nas previsões realizadas pelo modelo de árvore de decisão.

O Raiz do Erro Quadrático Médio (RMSE), com valor de 0,0832, é a raiz quadrada do MSE e fornece uma interpretação mais direta do erro, visto que é expresso na mesma unidade dos dados originais. O RMSE baixo indica que a magnitude dos erros de previsão é igualmente reduzida, reforçando a ideia de que o modelo de árvore de decisão está capturando adequadamente as relações entre as variáveis envolvidas, o que é essencial para a confiabilidade do modelo em termos práticos.

O Coeficiente de Determinação (R^2), com valor de 0,99999, indica que o modelo explica quase a totalidade da variância presente nos dados observados. Um valor de R^2 tão próximo de 1 sugere que o modelo de árvore de decisão é extremamente eficaz em capturar as nuances e variações do conjunto de dados. No contexto da previsão de conformidade no encaminhamento de dívidas inadimplentes, isso significa que praticamente todas as variações nas observações são explicadas pelas variáveis preditoras, garantindo a robustez e eficácia do modelo.

As métricas de avaliação apresentadas — MAE, MSE, RMSE e R^2 — indicam que o modelo de árvore de decisão apresentou uma performance excepcional na previsão da conformidade. A baixa magnitude dos erros e o R^2 quase perfeito sugerem um ajuste praticamente ideal aos dados observados. Contudo, como resultados tão precisos são raros em cenários práticos, é importante considerar a possibilidade de overfitting, onde o modelo se ajusta excessivamente aos dados de treinamento. Caso isso tenha ocorrido, o modelo pode apresentar dificuldades ao generalizar para novos dados. No entanto, se os dados forem representativos e adequadamente divididos entre treino e teste, esses resultados são altamente encorajadores e indicam que o modelo pode ser confiável em contextos aplicados.

A alta precisão do modelo de árvore de decisão demonstra que ele pode ser utilizado com confiança para prever a conformidade no encaminhamento de dívidas inadimplentes. Isso pode servir como uma ferramenta essencial para auxiliar gestores na tomada de decisões estratégicas e na formulação de políticas financeiras, que podem se basear nas previsões do modelo para melhorar as taxas de conformidade.

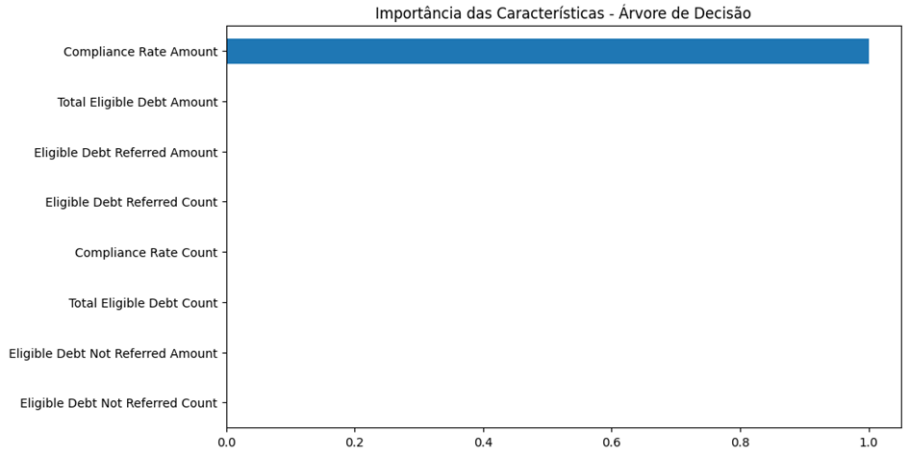
A árvore de decisão foi capaz de identificar padrões significativos nos dados, sendo uma alternativa robusta para apoiar a formulação de ações corretivas e preventivas no contexto da gestão de dívidas.

5.1 IMPORTÂNCIA DAS CARACTERÍSTICAS – ÁRVORE DE DECISÃO

No modelo de árvore de decisão, a identificação das variáveis mais relevantes é crucial para compreender a dinâmica dos dados e garantir a eficácia das previsões. A importância das características reflete a contribuição de cada variável para a redução da impureza dos nós ao longo da árvore. Esse

conceito é baseado na ideia de que as variáveis que mais reduzem a heterogeneidade dos dados nos nós são as mais influentes para as previsões do modelo.

Gráfico 4: Importância das Características – Árvore de Decisão



Fonte: Elaborado pelo autor (2024).

A interpretação do Gráfico 4: Importância das Características revela informações cruciais sobre os fatores que mais influenciam o modelo de árvore de decisão no contexto da previsão de conformidade no encaminhamento de dívidas inadimplentes. A análise inicial destaca que a variável "Compliance Rate Amount" foi identificada como a mais importante no modelo, demonstrando ser o fator preditivo mais relevante. Esse resultado sugere que o montante relacionado à taxa de conformidade exerce um papel central na previsão, contribuindo significativamente para a divisão dos nós da árvore de decisão e, por consequência, para a precisão das previsões geradas.

As demais variáveis, como "Total Eligible Debt Amount", "Eligible Debt Referred Amount" e "Compliance Rate Count", apresentaram um peso consideravelmente inferior no modelo. Isso indica que seu impacto na previsão é muito menor em comparação à variável "Compliance Rate Amount". Tal distribuição de relevância demonstra que, embora outras variáveis possam contribuir para o ajuste do modelo, elas não possuem o mesmo nível de influência na determinação da conformidade.

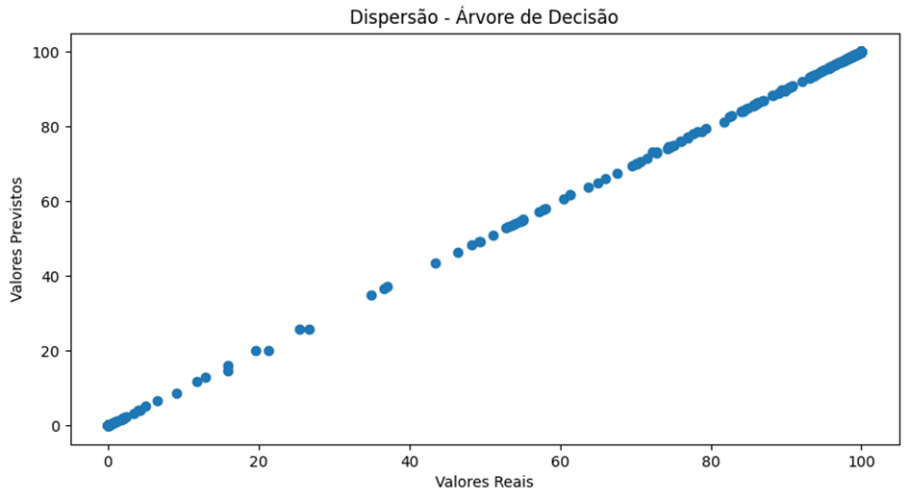
O gráfico revela uma distribuição concentrada de importância em torno de uma única variável, a "Compliance Rate Amount". Essa concentração extrema sugere que o modelo é altamente dependente dessa variável específica, o que implica que gestores financeiros devem priorizar o monitoramento e a otimização da taxa de conformidade ao desenvolver estratégias de melhoria para o encaminhamento de dívidas inadimplentes. Essa dependência também pode indicar que o modelo poderia ser simplificado ao focar mais intensamente nessa variável, otimizando recursos e esforços no processo de análise preditiva.

5.2 DISPERSÃO – ÁRVORE DE DECISÃO

A análise de dispersão é uma técnica essencial para avaliar a precisão e eficácia de modelos de árvore de decisão. No presente estudo, o gráfico de dispersão demonstra a relação entre os valores reais e os previstos pelo modelo, sendo fundamental para visualizar a capacidade do modelo em capturar a variabilidade dos dados e realizar previsões precisas. Em um cenário ideal, os pontos no gráfico devem se alinhar em uma linha diagonal, representando a correspondência perfeita entre previsões e valores reais. A análise desse gráfico permite identificar desvios sistemáticos, avaliar a precisão das previsões em diferentes intervalos e detectar outliers. Além disso, oferece insights sobre a robustez do modelo em lidar com variações não lineares, característica distintiva das árvores de decisão.

O Gráfico 5: Dispersão – Árvore de Decisão, a seguir, sugere que o modelo é capaz de capturar as tendências e padrões dos dados com alta precisão.

Gráfico 5: Dispersão – Árvore de Decisão



Fonte: Elaborado pelo autor (2024).

A interpretação do Gráfico 5: Dispersão – Árvore de Decisão revela uma análise clara da relação entre os valores reais e os valores previstos pelo modelo. Primeiramente, é importante destacar que na linha de tendência observada no gráfico os pontos estão quase totalmente alinhados, o que sugere que o modelo de árvore de decisão fez previsões muito próximas dos valores observados. Essa proximidade indica um alto grau de precisão nas previsões realizadas, o que demonstra a eficácia do modelo em antecipar a conformidade no encaminhamento de dívidas inadimplentes.

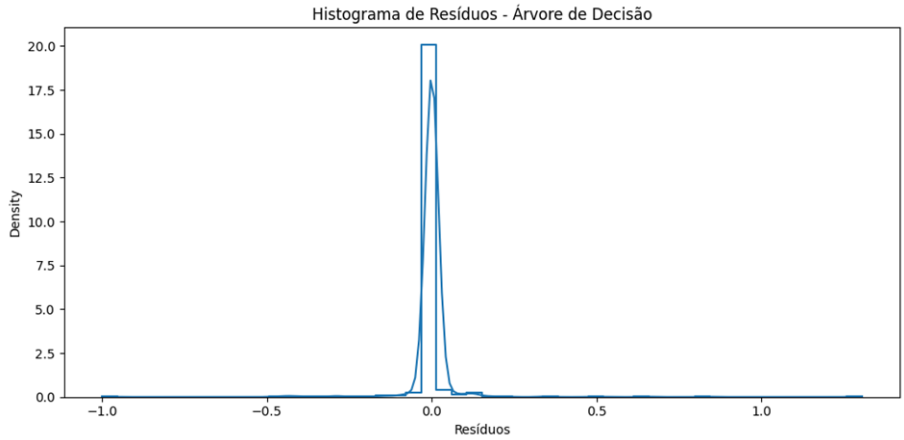
Essa baixa dispersão ao redor da linha de tendência indica que os erros de previsão são pequenos e distribuídos de maneira consistente, evidenciando que o modelo capturou corretamente as interações entre as variáveis preditoras. Esse comportamento do modelo reforça sua capacidade de fazer previsões precisas e consistentes ao longo dos diferentes pontos de dados analisados.

A análise gráfica complementa as métricas estatísticas discutidas anteriormente, validando a qualidade das previsões.

5.3 HISTOGRAMA DE RESÍDUOS – ÁRVORE DE DECISÃO

O histograma de resíduos é uma ferramenta essencial para avaliar a performance de modelos preditivos, como a árvore de decisão. Ele permite visualizar a distribuição dos resíduos, ou seja, as diferenças entre os valores reais e os previstos pelo modelo. No caso da árvore de decisão, um histograma bem distribuído e centrado em torno de zero indica previsões precisas e imparciais. A análise dos resíduos é fundamental para identificar possíveis vieses, heterocedasticidade e outliers, que podem comprometer a validade das previsões. Observando a forma e dispersão dos resíduos, é possível avaliar a adequação do modelo aos dados e sua capacidade de generalização.

Gráfico 6: Histograma de Resíduos – Árvore de Decisão



Fonte: Elaborado pelo autor (2024).

O Gráfico 6: Histograma de Resíduos – Árvore de Decisão proporciona uma análise detalhada da distribuição dos resíduos do modelo, sendo fundamental para avaliar sua performance preditiva. A distribuição dos resíduos apresenta-se concentrada majoritariamente em torno de zero, o que indica que o modelo realizou previsões precisas na maior parte dos casos. Essa distribuição simétrica reforça a suposição de que o modelo de árvore de decisão está bem ajustado aos dados utilizados, o que aumenta a confiança em sua capacidade de generalização.

Contudo, é importante destacar a presença de alguns outliers observados nas extremidades do histograma. Esses resíduos maiores indicam que, em certas observações, as previsões do modelo foram menos precisas. A ocorrência desses outliers pode estar associada a características específicas dos dados que não foram devidamente capturadas pelo modelo, sugerindo uma possível limitação em certos cenários.

A densidade elevada dos resíduos próximos a zero é um indicativo de que a maioria das previsões do modelo foi feita com alto grau de precisão. A concentração significativa nessa região reflete a eficácia geral do modelo de árvore de decisão, demonstrando que os erros de previsão foram, em sua maioria, pequenos e uniformemente distribuídos. Dessa forma, o histograma de resíduos

confirma a robustez do modelo, ao mesmo tempo em que aponta possíveis áreas para ajustes adicionais, visando minimizar os outliers e melhorar ainda mais a precisão preditiva.

O modelo de árvore de decisão apresentou desempenho excepcional, como evidenciado pelas métricas estatísticas, gráficos de dispersão e histograma de resíduos. Essas análises indicam que o modelo é robusto e confiável, tornando-o uma ferramenta valiosa para prever a conformidade no encaminhamento de dívidas inadimplentes.

6 DISCUSSÃO

A análise revelou uma forte correlação entre variáveis relacionadas ao montante e à contagem de dívidas encaminhadas. A correlação quase perfeita de 0,99 entre o "Montante de Dívida Elegível Encaminhada" e a "Contagem de Dívida Elegível Encaminhada" sugere uma alta interdependência entre o número de dívidas e o valor correspondente. Esse achado é consistente com a expectativa de que um maior número de dívidas resulte em um maior montante financeiro.

Por outro lado, a correlação mais moderada entre o "Montante de Dívida Elegível Não Encaminhada" e a "Contagem de Dívida Elegível Não Encaminhada" (0,64) pode sugerir que, embora a maioria das dívidas seja proporcional ao valor correspondente, pode haver flutuações em que o montante de dívidas não encaminhadas não corresponde diretamente ao número de dívidas não processadas. Esse desvio pode estar relacionado a políticas ou práticas que priorizam certas dívidas sobre outras, resultando em uma desconexão entre o valor total e o número de dívidas não encaminhadas.

Os resultados oferecem várias implicações teóricas e práticas. Em termos teóricos, a forte correlação entre montantes e contagens de dívidas encaminhadas reforça a teoria de que variáveis financeiras quantitativas, como a dívida total e o montante de dívida encaminhada, são cruciais para prever conformidade em modelos preditivos. Esses achados oferecem suporte à utilização da regressão linear e das árvores de decisão para modelar fenômenos financeiros com variáveis preditoras claramente definidas.

No âmbito prático, o conhecimento gerado por esses modelos preditivos pode ser aplicado para melhorar a gestão da dívida pública. A forte correlação entre variáveis de montante e contagem sugere que gestores públicos podem focar seus esforços em dívidas mais volumosas para melhorar as taxas de recuperação de dívida. Além disso, a ausência de correlação significativa com variáveis temporais, como o ano fiscal ou o mês, implica que essas variáveis podem ser minimizadas em modelos futuros, permitindo uma simplificação do processo de análise.

Apesar dos resultados promissores, o estudo apresenta algumas limitações. Primeiramente, os modelos exibiram um ajuste quase perfeito ($R^2 = 1$ para o modelo de regressão linear e $R^2 = 0,999997$ para o modelo de árvore de decisão), o que levanta a preocupação com overfitting. Esse problema

ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, comprometendo sua capacidade de generalizar para novos dados. Essa questão é crítica para garantir que o modelo seja aplicável em cenários práticos futuros.

Outra limitação é a falta de variabilidade temporal nas variáveis estudadas, o que pode sugerir que fatores relacionados ao tempo não foram adequadamente considerados no modelo. Isso pode impactar a capacidade do modelo de prever comportamentos futuros em diferentes períodos do ciclo fiscal ou em anos subsequentes.

Estudos anteriores, como os de Breiman et al. (1984), demonstram que as árvores de decisão são eficazes em contextos onde há relações complexas e não lineares entre variáveis, uma constatação que é confirmada por este estudo. Da mesma forma, pesquisas que utilizaram regressão linear para prever dívidas inadimplentes confirmam que este método é robusto quando há uma relação clara entre as variáveis, como observado aqui.

Contudo, a diferença na correlação entre dívidas encaminhadas e não encaminhadas merece destaque. Estudos anteriores sugerem que políticas públicas que envolvem prazos de encaminhamento de dívida podem interferir nessa relação, criando variações entre o número de dívidas e o montante total. Isso sugere que mais pesquisas são necessárias para entender as nuances dessa relação.

7 CONCLUSÃO

Os resultados obtidos mostram que tanto a Regressão Linear quanto a Árvore de Decisão são altamente eficazes para prever a conformidade no encaminhamento de dívidas inadimplentes. A Regressão Linear, com seu R^2 perfeito, sugere que há uma forte relação linear entre as variáveis financeiras e a conformidade, confirmando a importância do uso desse modelo em cenários com variáveis fortemente correlacionadas.

Por outro lado, a Árvore de Decisão se destacou por capturar nuances nos dados, incluindo variáveis não lineares que a Regressão Linear não foi capaz de captar tão bem. A identificação de "Compliance Rate Amount" como a variável mais significativa no modelo de Árvore de Decisão reforça a ideia de que o montante da taxa de conformidade é o principal fator preditivo de conformidade no encaminhamento de dívidas.

Os achados desse estudo oferecem contribuições significativas para o campo da gestão financeira pública e modelagem preditiva. Primeiramente, os resultados confirmam a eficácia de modelos estatísticos como a Regressão Linear para prever conformidade em cenários onde há relações lineares claras entre variáveis financeiras. Além disso, a aplicação bem-sucedida de técnicas de machine learning, como Árvores de Decisão, expande o uso de modelos preditivos em situações mais complexas, onde há interações não lineares.

A identificação das variáveis mais importantes, particularmente o peso dominante de "Compliance Rate Amount", pode ajudar gestores a focar em estratégias que melhorem esse aspecto específico, otimizando o processo de encaminhamento de dívidas inadimplentes.

Este estudo é relevante não apenas para o campo acadêmico, ao contribuir com uma análise comparativa de modelos preditivos, mas também para a prática na gestão financeira pública. A alta precisão demonstrada pelos modelos sugere que podem ser implementados com sucesso em sistemas reais de gerenciamento de dívidas inadimplentes, permitindo uma tomada de decisão mais informada e estratégias mais eficazes para melhorar as taxas de conformidade.

Além disso, a análise sugere que modelos como a Árvore de Decisão podem ser preferidos em contextos onde há complexidade nas interações das variáveis ou quando os dados apresentam características não lineares.

Embora os resultados obtenham alta precisão, o estudo pode apresentar limitações relacionadas ao potencial overfitting, especialmente com um R^2 perfeito na Regressão Linear. Isso sugere que o modelo pode ter se ajustado excessivamente aos dados de treinamento, o que pode impactar sua capacidade de generalização para novos dados. Outra limitação está na falta de exploração de possíveis variáveis adicionais que poderiam ter melhorado o desempenho dos modelos, especialmente no caso da Árvore de Decisão.

Pesquisas futuras poderiam focar em testar a generalização desses modelos em diferentes conjuntos de dados, explorando mais profundamente a questão do overfitting e como ele pode ser evitado em cenários de previsão. Além disso, a inclusão de outras variáveis, como fatores econômicos macroeconômicos, poderia enriquecer a análise preditiva e fornecer uma visão mais completa dos fatores que influenciam a conformidade no encaminhamento de dívidas inadimplentes.

Outra sugestão seria a aplicação de técnicas mais avançadas de machine learning, como Random Forest ou Gradient Boosting, para verificar se esses modelos conseguem superar o desempenho da Regressão Linear e das Árvores de Decisão em cenários semelhantes.

REFERÊNCIAS

ANGRIST, J. D.; PISCHKE, J.-S. Mostly harmless econometrics: an empiricist's companion. Massachusetts Institute of Technology and The London School of Economics, 2009. DOI: https://doi.org/10.1017/CBO9781107415324.004.

PÉREZ, et al. Análise de mudanças em fatores socioeconômicos baseado em árvore de decisão para o estudo de viagens por motivos trabalho e estudo na Região Metropolitana de São Paulo. In: 51º SBPO, SOBRAPO, 2019, p. 399–406.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Classification and Regression Trees. Belmont, CA: Wadsworth International Group, 1984.

GOMES, L. B. et al. Sistemas de Detecção de Fraudes Baseados em IA no Setor Público. Revista Brasileira de Auditoria Governamental, Brasília, v. 27, n. 1, p. 45-60, 2020.

HAIR, J. F. et al. Multivariate Data Analysis. 8. ed. Cengage Learning, 2019.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

IUDÍCIBUS, S. Teoria da Contabilidade. 9. ed. São Paulo: Atlas, 2010.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. Introduction to Linear Regression Analysis. 5. ed. Hoboken, NJ: Wiley, 2012.

TESOURO DOS EUA. Relatório de Conformidade de Encaminhamento de Dívidas Inadimplentes de 120 Dias. Disponível em: https://fiscaldata.treasury.gov/datasets/delinquent-debt-referral-compliance/120-day-delinquent-debt-referral-compliance-report. Acesso em: 22 jul. 2024.

VASCONCELOS, E. S.; SANTOS, F. A. Inteligência Artificial na Gestão Pública Brasileira: Desafios e Oportunidades para a Eficiência Governamental. Revista Observatorio de la Economía Latinoamericana, Curitiba, v. 22, n. 5, p. 01-21, 2024. DOI: https://doi.org/10.55905/oelv22n5-137Disponível em: https://ojs.observatoriolatinoamericano.com/ojs/index.php/olel/article/view/4792/3144. Acesso em: 22 jul. 2024.

VASCONCELOS, E. S.; SANTOS, F. A.; AMORIM, L. R. Princípios Fundamentais e Impactos das Políticas Fiscais e do Orçamento Público: Perspectivas para a Eficiência e Transparência na Administração Pública. RevistaFT, 2024. DOI: 10.5281/zenodo.11958942. Disponível em: https://revistaft.com.br/principios-fundamentais-e-impactos-das-politicas-fiscais-e-do-orcamento-publico-perspectivas-para-a-eficiencia-e-transparencia-na-administracao-publica/. Acesso em: 22 jul. 2024.