PREDICTIVE TECHNIQUES AND ARTIFICIAL INTELLIGENCE MODELS IN THE MANAGEMENT OF DEFAULTED PUBLIC DEBTS: COMPARISON BETWEEN LINEAR REGRESSION AND DECISION TREES

Eduardo Silva Vasconcelos[1]

ABSTRACT

The analyzed study focuses on the application of predictive models, specifically Linear Regression and Decision Trees, for the management of delinquent debts in the public context of the United States. The main objective of the work is to compare the effectiveness of these models in predicting the compliance of debts older than 120 days, helping to direct these debts to the Treasury Offset Program (TOP), an essential initiative for government financial recovery. The problem that the study addresses is the need for effective management of delinquent public debts, seeking to ensure compliance with public financial policies that promote compliance and the proper redirection of financial resources to the government. This is particularly important to ensure fiscal transparency and accountability of federal agencies. The methodology used in the study was quantitative, based on the analysis of eligible debt data extracted from U.S. Treasury reports. The Linear Regression and Decision Trees models were applied, with performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2). The study dealt with financial and temporal variables to analyze the behavior of these debts and their compliance. The main results show that both models showed high accuracy in the predictions, with the Linear Regression showing a perfect fit ($R^2 = 1$) and the Decision Trees excelling in capturing nonlinear nuances of the data. The "Compliance Rate Amount" variable was identified as the most significant in the Decision Tree model, suggesting that the amount of the compliance rate is one of the most important factors to predict the compliance of delinquent debts. This study offers valuable contributions to the field of public management, by demonstrating that the use of predictive models can help optimize debt recovery, improve fiscal transparency, and contribute to more informed decision-making.

Keywords: Public Financial Management, Predictive Modeling, Delinquent Debts, Applied Artificial Intelligence.

---

[1] Doctor of Science – Information Processing
CV: http://lattes.cnpq.br/5128388060472259
Instituto Federal Goiano
Goiânia, Goiás, Brazil
E-mail: educelos1@gmail.com

INTRODUCTION

Public administration, especially in the United States, faces the challenge of efficiently managing delinquent debts. Under the Digital Accountability and Transparency Act of 2014 (DATA Act), delinquent debts older than 120 days must be forwarded to the Treasury Offset Program (TOP) to ensure the recovery of revenues to the federal government. Failure to comply with these guidelines can undermine the government's ability to fund essential public services, as well as compromise the transparency and accountability of federal agencies (US TREASURY, 2024).

Predictive models such as Linear Regression and Decision Trees emerge as important tools to improve compliance in debt routing. These techniques make it possible to predict which debts are most likely to be forwarded, helping to optimize financial recovery processes. The study focuses on the application of these machine learning techniques to identify significant variables and predict compliance with the DATA Act.

The present study aims to compare the effectiveness of Linear Regression and Decision Tree models in predicting compliance in the routing of 120-day delinquent debts in the United States. From this central objective, the study unfolds into four specific objectives. First, it seeks to compare the effectiveness of these models using evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2). Such metrics are widely employed to measure the predictive accuracy of models and identify how well each of them fits the analyzed data.

Secondly, the study aims to identify the most significant variables for predicting compliance, considering which factors have the greatest impact on the predictive results of each model. This will allow for a deeper understanding of the key elements that influence the behavior of non-performing debts and their compliance in the context analyzed.

The third objective aims to analyze how these models can contribute to improving efficiency and transparency in the management of delinquent debts. The effective application of predictive models can optimize processes, reduce inefficiencies, and provide a clearer view of debt behavior, which is essential for public policy formulation.

Finally, the study seeks to provide strategic information that allows public managers to make evidence-based decisions. The use of robust predictive models can provide valuable insights for financial management, aiding in informed decision-making and

promoting more effective management of delinquent debts, with a positive impact on compliance and the allocation of public resources.

This study is relevant for public administration, as it contributes to the optimization of revenue recovery and the increase of fiscal transparency. Compliance with the DATA Act is essential to ensure the financial efficiency of federal agencies, and the implementation of advanced predictive models provides tools to accurately predict debt routing, improving budget management.

Debt defaults have significant impacts on public financial management, affecting the ability of governments to allocate resources and maintain budget stability. As noted by Iudícibus (2010), effective management of delinquent debts requires strict monitoring of payments and the adoption of policies that encourage debtors to comply with financial obligations. A central tool in this process in the United States is the Treasury Offset Program (TOP), which plays a crucial role in debt recovery. This program allows resources to be redirected to the government through the clearing of payments, assisting in the recovery of amounts due and contributing to fiscal sustainability.

Forecasting techniques are essential for effective management of non-performing debt, providing a solid foundation for anticipating future behaviors and assisting in policymaking. Among these techniques, Linear Regression models and Decision Trees stand out, widely used in financial forecasting and delinquency management.

Linear Regression is a statistical technique that seeks to predict the value of a dependent variable based on one or more independent variables. According to Montgomery, Peck and Vining (2012), this model is particularly suitable for situations in which there is a clear linear relationship between variables, providing a simple but effective approach to forecasting financial results.

In addition, Angrist and Pischke (2009) emphasize that linear regression models are fundamental in the field of econometrics, being used as computational tools to estimate the differences between treated groups and control groups, with or without the use of covariates. This method is crucial in evaluating interventions and measuring their impacts, offering precise control over the factors that can influence the results.

On the other hand, Decision Trees are machine learning techniques that stand out for their ability to partition data into homogeneous subsets, creating a hierarchical structure that facilitates decision-making. According to Breiman et al. (1984), decision trees are especially useful when there are complex and nonlinear relationships between variables, as

is often the case in predicting debt defaults. This method allows you to identify hidden patterns in the data and generate more detailed and accurate forecasts.

Decision Trees are widely recognized for their ease of interpretation and applicability in several areas. As noted by Pérez et al. (2019), this technique is often used in the development of interpretable classifiers due to its visual structure, which resembles a flowchart.

Artificial Intelligence (AI) has proven to be a powerful tool in public administration, especially in the context of financial forecasting. Silva and Rocha (2019) highlight that AI can help predict income and expenses, identify potential debtors, and optimize debt recovery strategies. In addition, AI can also improve transparency and accountability in public administration (GOMES et al., 2020).

Another relevant study is that of Vasconcelos, Santos, and Amorim (2024), who explore the use of optimization algorithms to improve the allocation of resources in the public budget. Research shows that applying AI-based predictive models can increase the effectiveness of tax policies by ensuring that resources are directed to the areas of greatest need and social impact.

The incorporation of Artificial Intelligence (AI) in public management has stood out as one of the most significant advances for the modernization of government entities, especially in Brazil, where its adoption has the potential to transform the formulation and evaluation of public policies, in addition to improving service to citizens. According to Vasconcelos and Santos (2024), the study on the application of AI in the public sector is essential, as it explores effective ways to integrate this technology into Brazilian administrations, with the aim of optimizing processes and improving the quality of services offered to the population.

METHODOLOGY

This study employs a quantitative approach, using statistical and machine learning models to predict compliance in the routing of delinquent debts. The research is applied, aiming to provide practical insights for public managers on the effectiveness of Linear Regression and Decision Tree models in financial management.

The data used for this analysis was extracted from the 120-Day Delinquent Debt Forwarding Compliance Report, available on the U.S. Treasury website. That dataset includes detailed information on eligible debts, referred debts and non-forwarded debts.

Before applying the predictive models, it was necessary to prepare the data. The treatment involved replacing missing values with means or medians, as well as coding categorical variables to make them compatible with machine learning algorithms (HAIR et al., 2019).

The Linear Regression model was applied to capture linear relationships between variables, while the Decision Tree model was used to identify complex and nonlinear interactions. Both models were trained and evaluated using performance metrics such as MAE, MSE, RMSE, and R² (BREIMAN et al., 1984; MONTGOMERY, PECK, VINING, 2012).

The effectiveness of the models was measured using statistical metrics. The Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and the Coefficient of Determination (R²) were used to evaluate the accuracy of the predictions and the robustness of the models.

To facilitate the understanding of the data used, Chart 1 was elaborated, where the original variables are presented, as well as their respective translations into Portuguese. This ensures that the analysis is consistent and understandable in the context of bad debt management.

Table 1: Table of Variables

Original Variable	Translation into Brazilian Portuguese
Total Eligible Debt Amount	Total Eligible Debt Amount
Total Eligible Debt Count	Total Eligible Debt Count
Eligible Debt Referred Amount	Amount of Eligible Debt Forwarded
Eligible Debt Referred Count	Eligible Debt Count Forwarded
Eligible Debt Not Referred	Amount of Eligible Debt Not Forwarded
Eligible Debt Not Referred Count	Count of Eligible Debt Not Forwarded
Compliance Rate Amount	Amount of the Compliance Fee
Compliance Rate Count	Compliance Rate Count
Fiscal Year	Fiscal Year
Fiscal Quarter Number	Fiscal Quarter Number
Calendar Year	Calendar Year
Calendar Quarter Number	Calendar Quarter Number
Calendar Month Number	Calendar Month Number
Calendar Day Number	Calendar Day Number

Source: Prepared by the author (2024).

These variables represent both quantitative aspects, such as amounts and counts of debts, and temporal factors, being fundamental for the construction of predictive models.

CORRELATION MATRIX

The Correlation Matrix, Graph 1: Correlation Matrix, was constructed with the objective of identifying the relationships between the numerical variables of the data set. This analysis allows us to understand how the different aspects of eligible and delinquent debts correlate, providing insights into the most influential variables for predictive models.

Graph 1: Correlation Matrix
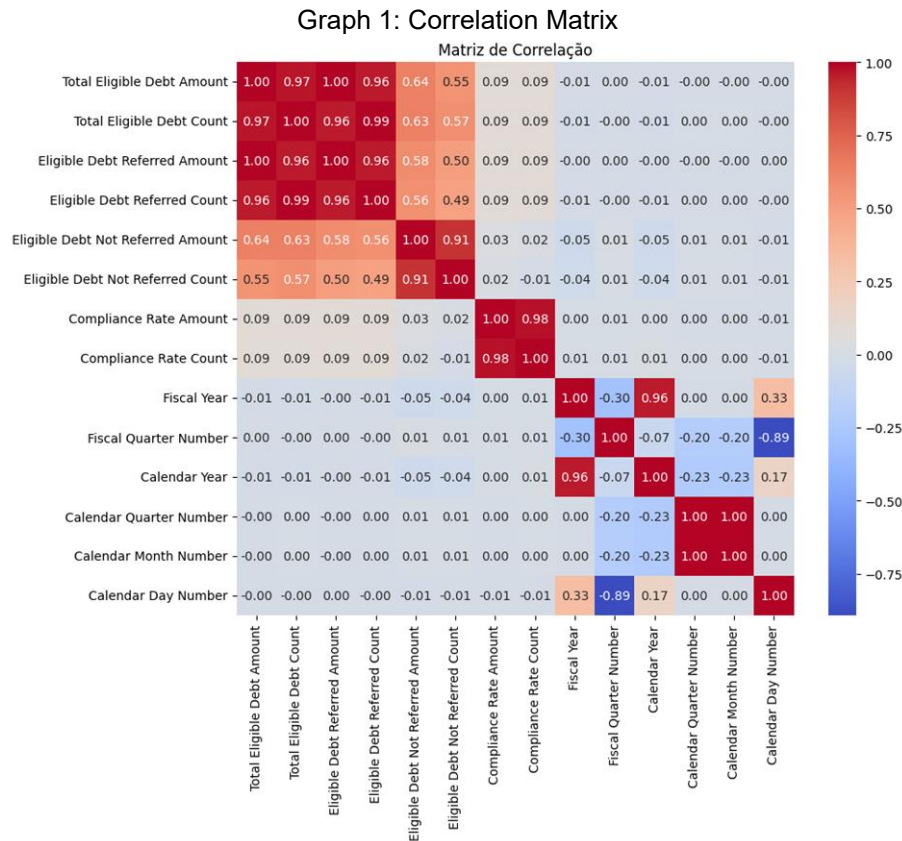
Source: Prepared by the author (2024).

The "Correlation Matrix" shown in the image reflects the linear relationship between several variables associated with the study of eligible debts and their compliance in the forwarding of delinquent debts.

Analysis of the correlation matrix reveals important relationships between financial and temporal variables, offering valuable insights into the behavior of eligible debts and the compliance process. First, the strong positive correlation between financial variables stands out, especially between the variables Total Eligible Debt Amount and Total Eligible Debt Count, which have a correlation of 0.97. This ratio suggests that as the total number of eligible debts increases, the corresponding total amount also grows proportionately. Such behavior is to be expected, since a higher number of debts naturally results in an increase in the total amount owed. A similar pattern is observed between the Eligible Debt Referred

Amount and Eligible Debt Referred Count variables, with a near-perfect correlation of 0.99. This interdependence indicates that the number of referred debts and the associated value go together consistently, pointing to a uniformity in the debt forwarding process, where both the number of referrals and the value follow the same trend.

In addition, a moderate correlation is observed with non-forwarded debts. The ratio between Eligible Debt Not Referred Amount and Eligible Debt Not Referred Count is 0.64, a value considerably lower than that observed in forwarded debts. This difference suggests that the process of non-routing of debts may not be perfectly aligned in terms of amount and count, and there are fluctuations in the amounts that do not directly correspond to the number of unforwarded debts. Additionally, there is a moderate correlation of 0.98 between Compliance Rate Amount and Compliance Rate Count, indicating that compliance, both in terms of amount and volume of debt, is well aligned. This reinforces consistency in the compliance measurement process, considering both absolute amounts and the number of debts.

Another relevant point is the weak or no correlation with temporal variables. The matrix shows that there are no significant correlations between financial variables and calendar variables, such as Fiscal Year, Calendar Year, Calendar Month Number, and Calendar Day Number. The correlation between these variables ranges from -0.01 to 0.05, suggesting that temporal factors, such as the fiscal year, month, or day, do not exert a significant direct influence on variations in eligible or forwarded debts. The only exception is a moderate correlation of 0.33 between Calendar Day Number and Fiscal Quarter Number, which can be explained by the structure of the calendar within each fiscal quarter.

Finally, the matrix also reveals the existence of some negative correlations. For example, there is a correlation of -0.30 between Fiscal Quarter Number and Calendar Year, as well as -0.20 between Fiscal Quarter Number and Calendar Quarter Number. These relationships suggest that as the fiscal quarter number increases, the impact on the calendar year or quarter may diminish, possibly due to changes in the fiscal period from the regular calendar. In addition, a negative correlation of -0.04 is observed between Fiscal Quarter Number and Compliance Rate Count, indicating that, although the relationship is very weak, compliance in terms of debt count may be slightly affected by the structure of fiscal quarters. However, this correlation is so small that its relevance can be disregarded in the general context of the analysis.

The matrix confirms the expected interdependence between the amount and count variables of debts, especially for forwarded debts, where the correlation values are almost perfect. In addition, the weak correlation with temporal variables suggests that the compliance and debt routing processes are not directly linked to the calendar. This is important to understand that temporal factors, such as the fiscal year or the month, do not play a critical role in the behavior of these financial variables.

The biggest implication of this matrix is to identify the variables that most contribute to the success of predictions in compliance models. The strong correlation between amounts and debt counts indicates that these factors should be prioritized in predictive models, while temporal variables can be ignored or minimized to avoid unnecessary noise in the model.

The correlation matrix was used to select variables with the greatest impact on the models. Variables with strong correlations were retained, while those with less relevance, such as calendar variables, were discarded to simplify the model and minimize noise in the data. This approach ensured greater accuracy in the predictive models.

LINEAR REGRESSION

The linear regression model was applied to predict compliance in the routing of delinquent debts older than 120 days. The evaluation of the model was performed using statistical metrics that measure the accuracy of the predictions and the quality of the model's fit to the observed data.

The results obtained from the model's evaluation metrics indicate its high accuracy. The calculated metrics were: Mean Absolute Error (MAE) of 3.2054e-12, Mean Squared Error (MSE) of 2.8218e-22, Root Mean SquareError (RMSE) of 1.6798e-11 and Coefficient of Determination (R^2) equal to 1.0.

These values show the performance of the model, evidencing the accuracy of the predictions made. The details of these results will be discussed below.

The Mean Absolute Error (MAE), which measures the average of the absolute errors between the predicted values and the actual values, was extremely low, with a value of 3.2054e-12. This metric indicates that, on average, the difference between the model's predictions and the observed values is negligible. In the context of predictive modeling, such a low MAE suggests that the linear regression model has a very accurate ability to

predict debt compliance while minimizing discrepancies between the observed data and the generated forecasts.

The Mean Squared Error (MSE), which measures the mean of squares of the differences between the predicted and actual values, presented a value of 2.8218e-22. The MSE penalizes larger errors more severely, as it squares the discrepancies. The fact that this value is extremely small indicates that the model makes virtually no significant errors. An MSE close to zero is a strong indicator that the model is highly adjusted to compliance data, predicting with extreme accuracy.

The Root Mean Squared Error (RMSE), which is the square root of the MSE, provides a measure of the error that is in the same unit as the predicted values. With a value of 1.6798e-11, the RMSE also confirms that the magnitude of the prediction errors is extremely low, reinforcing the idea that the linear regression model is highly effective in capturing the variations between the independent variables and compliance in debt routing. This value indicates that the deviations between the forecasts and the actual values are practically non-existent.

The Coefficient of Determination (R^2), with a value equal to 1.0, represents the model's ability to explain all the variability present in the observed data. In other words, the model explains 100% of the variation in the dependent variable (debt compliance) based on the independent variables used in the model. A value of R^2 of 1.0 suggests a perfect fit, which means that the model leaves no variation unexplained by the selected variables.

Regarding the accuracy and effectiveness of the model, we have that the evaluation metrics presented (MAE, MSE, and RMSE) show extremely low results, indicating that the linear regression model has remarkable accuracy. The absence of significant errors reinforces the robustness of the model, which proves to be highly effective in predicting compliance in the routing of delinquent debts.

The R^2 of 1.0 suggests a perfect fit between the predicted and observed values. This result indicates that the model is able to fully capture the relationship between predictor variables and compliance, leaving no room for unexplained prediction errors. This level of accuracy is rarely observed in common predictive practices, suggesting a strong relationship between the selected variables and the target variable.

Although the results obtained demonstrate a high accuracy, a perfect fit (such as the R^2 of 1.0) can raise concerns about overfitting, where the model fits excessively to the training data, compromising its ability to generalize to new data. However, if the data are

well partitioned between training and testing and are representative, these results are extremely promising and indicate that the model can be reliable in practical applications.

The high accuracy of the linear regression model suggests that it can be widely applied to predict compliance in the routing of delinquent debts, offering valuable support for managerial and strategic decision-making. By accurately capturing the variables that influence compliance behavior, the model enables money managers to implement more effective debt management policies and strategies.

In addition, the detailed analysis of the metrics provides a solid basis for comparing the performance of the linear regression model with other predictive methods, such as the decision tree. This can be critical in choosing the most suitable model for different forecasting scenarios, especially in environments of high variability.
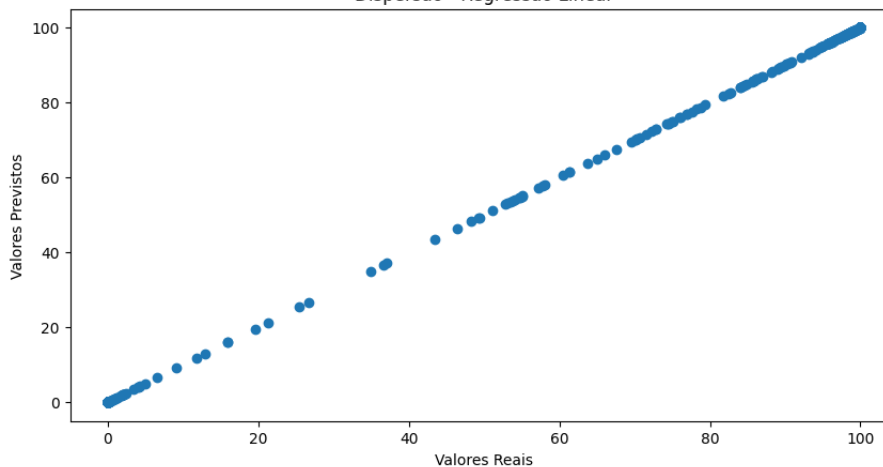
DISPERSION ANALYSIS – LINEAR REGRESSION

Dispersion analysis in linear regression is an essential statistical tool to evaluate the relationship between the independent variables and the dependent variable. This tool allows for a clear visualization of the accuracy of the predictions made by the model, highlighting how well the predicted values align with the observed values.

The scatter plot illustrates the relationship between the actual values and the values predicted by the decision tree model. The scatter plot is critical to visualizing the model's ability to capture data variability and make accurate predictions. The points should line up along a diagonal line that represents the perfect match between the predictions and the actual values. The analysis of this graph allows you to identify systematic deviations, evaluate the accuracy of forecasts in different ranges of values and detect possible outliers. In addition, dispersion provides insights into the robustness of the model in dealing with nonlinear variations in data, a distinguishing feature of decision trees (BREIMAN et al., 1984; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

As illustrated in Graph 2: Dispersion – Linear Regression, the dispersion of the predicted values in relation to the actual values is observed, allowing the identification of underlying patterns, as well as the detection of discrepancies or anomalies in the data.

Graph 2: Dispersion – Linear Regression

Source: Prepared by the author (2024).

Graph 2: Dispersion – Linear Regression presents a fundamental visual analysis to evaluate the performance of the linear regression model. The presence of a clear alignment of the points along the trend line indicates that the model has produced predictions that practically coincide with the actual data, which suggests a precise fit of the model to the analyzed variables.

In the graph, it is observed that the points are almost completely aligned to this line, which confirms that the linear regression model performed highly accurately in most cases. This proximity between the predicted and observed values reinforces that the model was able to capture with a high degree of accuracy the relationship between the explanatory variables and the dependent variable, generating reliable predictions.

The visual analysis provided by the scatter plot confirms the effectiveness of the linear regression model in predicting the results. The strong correlation between the actual and predicted values demonstrates that the model was successful in capturing the underlying relationships between the variables. In combination with performance metrics such as MAE, MSE, RMSE, and R^2, discussed earlier, the scatter plot offers a comprehensive understanding of the model's accuracy and robustness. The absence of relevant discrepancies validates the applicability of the model to the predictions within the context of the study.

RESIDUAL HISTOGRAM – LINEAR REGRESSION

The residual histogram is an essential tool to assess the adequacy of a linear regression model, allowing the visualization of the distribution of prediction errors in

comparison with the actual values. It assists in verifying the assumptions of the model, such as normality and independence from errors. In a well-fitted model, the residuals should present an approximately normal distribution, concentrated around zero, indicating the absence of systematic bias and unbiased forecasts. In addition, the histogram facilitates the identification of outliers, heteroscedasticity, and other deviations that can compromise the validity of the model.

Graph 3: Residual Histogram – Linear Regression

Source: Prepared by the author (2024)
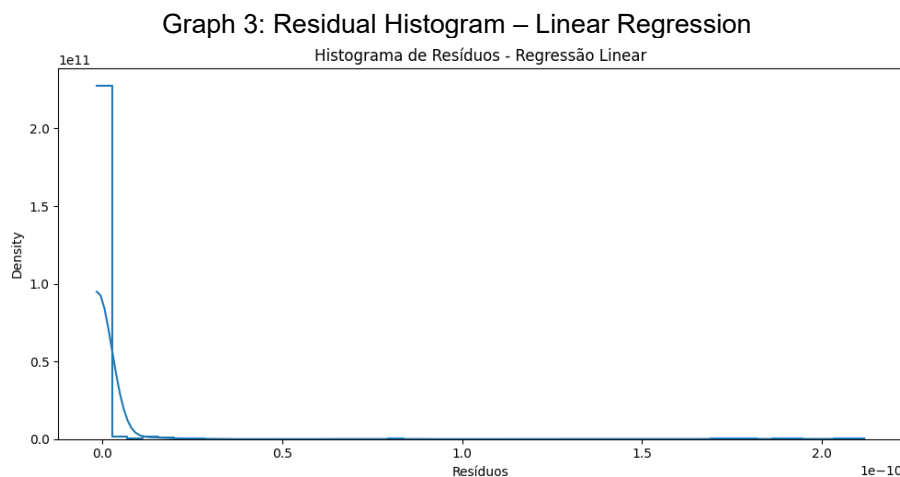
Graph 3: Residual Histogram – Linear Regression offers a detailed analysis of the distribution of the residuals generated by the model, allowing an accurate assessment of the quality of the adjustment made. The distribution of the residuals is an important tool to verify the accuracy of the forecasts, as well as to identify possible areas of improvement in the model's performance.

When analyzing the distribution of the residuals, it is observed that most of them are concentrated around zero, which indicates that the linear regression model produced very accurate predictions for most of the data. However, there is a slight asymmetry in the distribution, with a higher concentration of small waste. This may suggest that while the model performs well overall, additional adjustments may be needed to correct this asymmetry and achieve a perfectly symmetric distribution, which would potentially improve the predictive quality of the model.

Note the presence of outliers, mainly visible on the right tail of the histogram. These outliers correspond to larger residuals, indicating that, in some cases, the forecasts have distanced themselves from the actual observed values. This suggests that the model may

not have adequately captured certain relationships between variables in specific situations, pointing to the need for further investigation into the factors that cause these discrepancies.

The density of the residuals around zero, in turn, is quite high, which confirms that most of the predictions were made with negligible errors. This high concentration of small residues is a positive indicator of the model's effectiveness, demonstrating that it presented a satisfactory performance in most cases, with few exceptions.

The presence of outliers and the slight asymmetry of the residues point to areas where the model can be improved. These results suggest that, although the model is broadly robust, there are opportunities to improve its adequacy to the data, especially in relation to variables that may have nonlinear effects or data that may be negatively impacting the distribution of the waste.

DECISION TREE

The decision tree model was used with the objective of predicting compliance in the forwarding of delinquent debts longer than 120 days. The performance of the model is evaluated through several metrics, which quantify the accuracy of the predictions and the quality of the fit to the observed data. Below, each of these metrics is discussed in detail, providing an in-depth analysis of the effectiveness and robustness of the model.

The main evaluation metrics presented the following values: Mean Absolute Error (MAE) of 0.0138, Mean Squared Error (MSE) of 0.00692, Root Mean Squared Error (RMSE) of 0.0832 and Coefficient of Determination (R^2) of 0.99999.

These results will be discussed in detail below, with the aim of explaining the importance of each metric in the context of the analysis performed.

The value obtained for the Mean Absolute Error (MAE) was 0.0138, a metric that measures the average of the absolute errors between the values predicted by the model and the actual values observed. This relatively low value indicates the decision tree model performed very accurately, with small average errors throughout its predictions. In the context of this study, the MAE of 0.0138 means that, on average, the difference between the actual and predicted values by the model is very small, which demonstrates the high effectiveness in anticipating compliance in the routing of delinquent debts.

The Mean Squared Error (MSE) was 0.00692, a metric that penalizes larger errors more severely by squaring the differences before adding them up. The fact that the MSE is so low reinforces the conclusion that the model makes predictions with high accuracy, with

little discrepancy between the observed and predicted values. The MSE is particularly useful for identifying outlier predictions because it penalizes larger errors, which suggests that there are no major distortions in the predictions made by the decision tree model.

The Root Mean Square Error (RMSE), with a value of 0.0832, is the square root of the MSE and provides a more direct interpretation of the error, since it is expressed in the same unit as the original data. The low RMSE indicates that the magnitude of the prediction errors is similarly reduced, reinforcing the idea that the decision tree model is adequately capturing the relationships between the variables involved, which is essential for the reliability of the model in practical terms.

The Coefficient of Determination (R^2), with a value of 0.99999, indicates that the model explains almost all the variance present in the observed data. A value of R^2 so close to 1 suggests that the decision tree model is extremely effective at capturing the nuances and variations of the data set. In the context of predicting compliance in the forwarding of delinquent debts, this means that virtually all variations in observations are explained by the predictor variables, ensuring the robustness and effectiveness of the model.

The evaluation metrics presented—MAE, MSE, RMSE, and R^2—indicate that the decision tree model performed exceptionally well in predicting compliance. The low magnitude of the errors and the almost perfect R^2 suggest a practically ideal fit to the observed data. However, because such accurate results are rare in practical scenarios, it is important to consider the possibility of overfitting, where the model fits excessively to the training data. If this has occurred, the model may present difficulties when generalizing to new data. However, if the data are representative and appropriately split between training and testing, these results are highly encouraging and indicate that the model can be trusted in applied contexts.

The decision tree model's high accuracy demonstrates that it can be used with confidence to predict compliance in the routing of delinquent debt. This can serve as an essential tool to assist managers in making strategic decisions and formulating financial policies, which can be based on the model's predictions to improve compliance rates.

The decision tree was able to identify significant patterns in the data, being a robust alternative to support the formulation of corrective and preventive actions in the context of debt management.

IMPORTANCE OF CHARACTERISTICS – DECISION TREE

In the decision tree model, identifying the most relevant variables is crucial to understanding the dynamics of the data and ensuring the effectiveness of the predictions. The importance of the characteristics reflects the contribution of each variable to the reduction of the impurity of the nodes along the tree. This concept is based on the idea that the variables that most reduce the heterogeneity of the data in the nodes are the most influential for the model's predictions.

Graph 4: Importance of Traits – Decision Tree
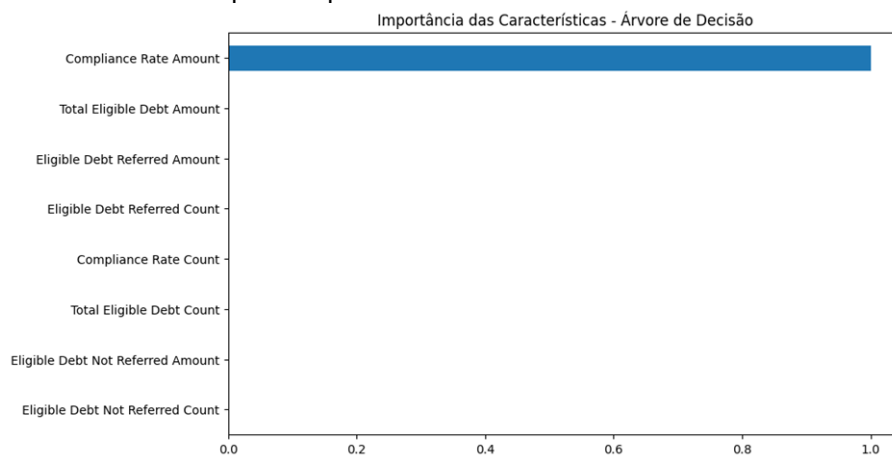
Source: Prepared by the author (2024).

The interpretation of Graph 4: Importance of Characteristics reveals crucial information about the factors that most influence the decision tree model in the context of predicting compliance in the routing of delinquent debts. The initial analysis highlights that the variable "Compliance Rate Amount" was identified as the most important in the model, proving to be the most relevant predictive factor. This result suggests that the amount related to the compliance fee plays a central role in the forecast, contributing significantly to the division of the nodes of the decision tree and, consequently, to the accuracy of the generated forecasts.

The other variables, such as "Total Eligible Debt Amount", "Eligible Debt Referred Amount" and "Compliance Rate Count", had a considerably lower weight in the model. This indicates that its impact on the forecast is much smaller compared to the "Compliance Rate Amount" variable. Such distribution of relevance demonstrates that, although other variables can contribute to the adjustment of the model, they do not have the same level of influence in the determination of compliance.

The graph reveals a concentrated distribution of importance around a single variable, the "Compliance Rate Amount". This extreme concentration suggests that the model is highly dependent on this specific variable, implying that financial managers should prioritize monitoring and optimizing the compliance rate when developing improvement strategies for the routing of delinquent debts. This dependence may also indicate that the model could be simplified by focusing more intensely on this variable, optimizing resources and efforts in the predictive analysis process.

SCATTER – DECISION TREE

Scatter analysis is an essential technique for evaluating the accuracy and effectiveness of decision tree models. In the present study, the scatter plot demonstrates the relationship between the actual values and those predicted by the model, being essential to visualize the model's ability to capture the variability of the data and make accurate predictions. In an ideal scenario, the points on the graph should line up in a diagonal line, representing the perfect match between predictions and actual values. The analysis of this graph allows you to identify systematic deviations, evaluate the accuracy of forecasts at different intervals, and detect outliers. In addition, it offers insights into the robustness of the model in dealing with nonlinear variations, a distinctive feature of decision trees.

Graph 5: Dispersion – Decision Tree, below, suggests that the model is capable of capturing trends and patterns in the data with high accuracy.

Graph 5: Dispersion – Decision Tree
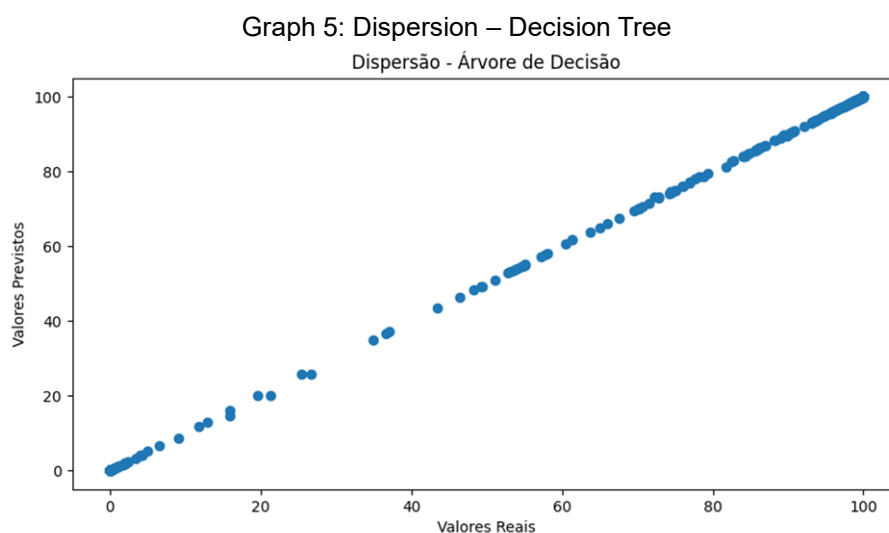
Source: Prepared by the author (2024).

The interpretation of Graph 5: Dispersion – Decision Tree reveals a clear analysis of the relationship between the actual values and the values predicted by the model. First, it is important to highlight that in the trend line observed in the graph the points are almost completely aligned, which suggests that the decision tree model made predictions very close to the observed values. This proximity indicates a high degree of accuracy in the forecasts made, which demonstrates the effectiveness of the model in anticipating compliance in the forwarding of delinquent debts.

This low dispersion around the trend line indicates that the prediction errors are small and consistently distributed, showing that the model correctly captured the interactions between the predictor variables. This model behavior reinforces its ability to make accurate and consistent predictions across the different data points analyzed.

Graphical analysis complements the statistical metrics discussed earlier, validating the quality of the forecasts.

RESIDUALS HISTOGRAM – DECISION TREE

The residual histogram is an essential tool for evaluating the performance of predictive models, such as the decision tree. It allows you to visualize the distribution of waste, that is, the differences between the actual values and those predicted by the model. In the case of the decision tree, a histogram that is well distributed and centered around zero indicates accurate and unbiased predictions. The analysis of the residuals is essential to identify possible biases, heteroscedasticity and outliers, which can compromise the validity of the forecasts. By observing the shape and dispersion of the residuals, it is possible to evaluate the adequacy of the model to the data and its ability to generalize.

Graph 6: Residual Histogram – Decision Tree
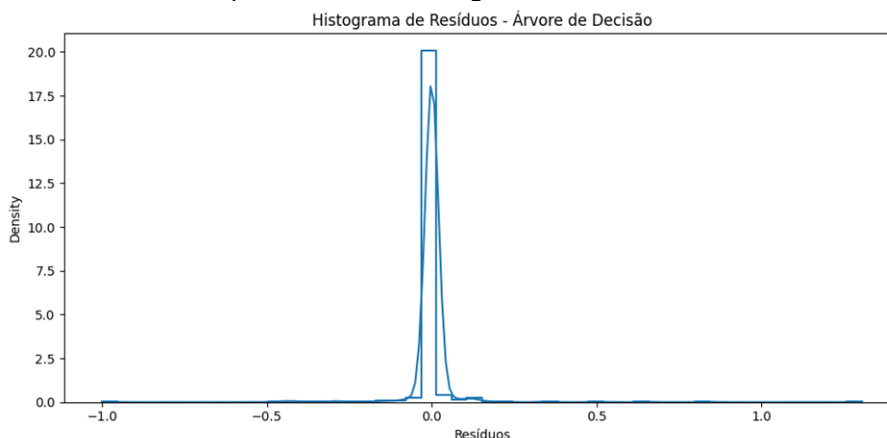
Source: Prepared by the author (2024).

Graph 6: Residual Histogram – Decision Tree provides a detailed analysis of the distribution of the model's residuals, being essential to evaluate its predictive performance. The distribution of waste is mostly concentrated around zero, which indicates that the model made accurate predictions in most cases. This symmetric distribution reinforces the assumption that the decision tree model is well-tuned to the data used, which increases confidence in its ability to generalize.

However, it is important to highlight the presence of some outliers observed at the ends of the histogram. These larger residuals indicate that, in certain observations, the model's predictions were less accurate. The occurrence of these outliers may be associated with specific characteristics of the data that were not properly captured by the model, suggesting a possible limitation in certain scenarios.

The high density of near-zero waste is an indication that most of the model's predictions were made with a high degree of accuracy. The significant concentration in this region reflects the overall effectiveness of the decision tree model, demonstrating that the prediction errors were mostly small and evenly distributed. In this way, the residual histogram confirms the robustness of the model, while pointing out possible areas for further adjustments, aiming to minimize outliers and further improve predictive accuracy.

The decision tree model showed exceptional performance, as evidenced by the statistical metrics, scatter plots, and residual histogram. These analyses indicate that the model is robust and reliable, making it a valuable tool for predicting compliance in the routing of delinquent debts.

DISCUSSION

The analysis revealed a strong correlation between variables related to the amount and count of debts forwarded. The near-perfect correlation of 0.99 between the "Amount of Eligible Debt Forwarded" and the "Count of Eligible Debt Forwarded" suggests a high interdependence between the number of debts and the corresponding amount. This finding is consistent with the expectation that a greater number of debts will result in a greater financial amount.

On the other hand, the more moderate correlation between the "Amount of Eligible Debt Not Forwarded" and the "Count of Eligible Debt Not Forwarded" (0.64) may suggest that while most debts are proportional to the corresponding amount, there may be

fluctuations where the amount of unforwarded debts does not directly correspond to the number of unprocessed debts. This deviation may be related to policies or practices that prioritize certain debts over others, resulting in a disconnect between the total amount and the number of unforwarded debts.

The results offer several theoretical and practical implications. In theoretical terms, the strong correlation between amounts and counts of forwarded debts reinforces the theory that quantitative financial variables, such as total debt and the amount of forwarded debt, are crucial for predicting compliance in predictive models. These findings support the use of linear regression and decision trees to model financial phenomena with clearly defined predictor variables.

On a practical level, the knowledge generated by these predictive models can be applied to improve public debt management. The strong correlation between amount and count variables suggests that public managers can focus their efforts on larger debts to improve debt recovery rates. In addition, the absence of significant correlation with temporal variables, such as the fiscal year or the month, implies that these variables can be minimized in future models, allowing a simplification of the analysis process.

Despite the promising results, the study has some limitations. First, the models exhibited a near-perfect fit ($R^2 = 1$ for the linear regression model and $R^2 = 0.999997$ for the decision tree model), which raises the concern of overfitting. This problem occurs when the model fits too much to the training data, compromising its ability to generalize to new data. This question is critical to ensure that the model is applicable in future practical scenarios.

Another limitation is the lack of temporal variability in the variables studied, which may suggest that time-related factors were not adequately considered in the model. This can impact the model's ability to predict future behaviors at different periods of the fiscal cycle or in subsequent years.

Previous studies, such as those by Breiman et al. (1984), demonstrate that decision trees are effective in contexts where there are complex and non-linear relationships between variables, a finding that is confirmed by this study. Similarly, research that has used linear regression to predict delinquent debts confirms that this method is robust when there is a clear relationship between variables, as noted here.

However, the difference in the correlation between forwarded and non-forwarded debts deserves to be highlighted. Previous studies suggest that public policies that involve debt forwarding deadlines can interfere with this relationship, creating variations between

the number of debts and the total amount. This suggests that more research is needed to understand the nuances of this relationship.

CONCLUSION

The results obtained show that both Linear Regression and Decision Tree are highly effective in predicting compliance in the routing of delinquent debts. Linear Regression, with its perfect R², suggests that there is a strong linear relationship between financial variables and compliance, confirming the importance of using this model in scenarios with strongly correlated variables.

On the other hand, Decision Tree stood out for capturing nuances in the data, including nonlinear variables that Linear Regression was not able to capture as well. The identification of "Compliance Rate Amount" as the most significant variable in the Decision Tree model reinforces the idea that the amount of the compliance rate is the main predictive factor of compliance in debt routing.

The findings of this study offer significant contributions to the field of public financial management and predictive modeling. First, the results confirm the effectiveness of statistical models such as Linear Regression to predict compliance in scenarios where there are clear linear relationships between financial variables. In addition, the successful application of machine learning techniques, such as Decision Trees, expands the use of predictive models in more complex situations, where there are nonlinear interactions.

Identifying the most important variables, particularly the dominant weight of "Compliance Rate Amount", can help managers focus on strategies that improve this specific aspect, optimizing the process of forwarding delinquent debts.

This study is relevant not only to the academic field, by contributing to a comparative analysis of predictive models, but also to the practice in public financial management. The high accuracy demonstrated by the models suggests that they can be successfully implemented in real non-performing debt management systems, allowing for more informed decision-making and more effective strategies to improve compliance rates.

In addition, the analysis suggests that models such as the Decision Tree may be preferred in contexts where there is complexity in the interactions of variables or when the data have nonlinear characteristics.

Although the results obtain high accuracy, the study may have limitations related to the potential for overfitting, especially with a perfect R² in Linear Regression. This suggests

that the model may have over-adjusted to the training data, which may impact its ability to generalize to new data. Another limitation is the lack of exploration of possible additional variables that could have improved the performance of the models, especially in the case of the Decision Tree.

Future research could focus on testing the generalizability of these models across different datasets, further exploring the issue of overfitting and how it can be avoided in prediction scenarios. In addition, the inclusion of other variables, such as macroeconomic economic factors, could enrich predictive analytics and provide a more complete picture of the factors that influence compliance in the routing of non-performing debts.

Another suggestion would be to apply more advanced machine learning techniques, such as Random Forest or Gradient Boosting, to see if these models can outperform Linear Regression and Decision Trees in similar scenarios.

REFERENCES

1. Angrist, J. D., & Pischke, J.-S. (2009). Mostly harmless econometrics: An empiricist's companion. Massachusetts Institute of Technology and The London School of Economics. https://doi.org/10.1017/CBO9781107415324.004

2. Pérez, et al. (2019). 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 (pp. 399–406).

3. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth International Group.

4. Gomes, L. B., et al. (2020). Sistemas de detecção de fraudes baseados em IA no setor público. Revista Brasileira de Auditoria Governamental, 27(1), 45–60.

5. Hair, J. F., et al. (2019). Multivariate data analysis (8th ed.). Cengage Learning.

6. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.

7. Iudícibus, S. (2010). Teoria da contabilidade (9th ed.). São Paulo: Atlas.

8. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (5th ed.). Hoboken, NJ: Wiley.

9. Tesouro dos EUA. (2024). 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.

10. Vasconcelos, E. S., & Santos, F. A. (2024). Inteligência artificial na gestão pública brasileira: Desafios e oportunidades para a eficiência governamental. Revista Observatorio de la Economía Latinoamericana, 22(5), 01–21. https://doi.org/10.55905/oelv22n5-137. Disponível em: https://ojs.observatoriolatinoamericano.com/ojs/index.php/olel/article/view/4792/3144. Acesso em: 22 jul. 2024.

11. Vasconcelos, E. S., Santos, F. A., & Amorim, L. R. (2024). Princípios fundamentais e impactos das políticas fiscais e do orçamento público: Perspectivas para a eficiência e transparência na administração pública. RevistaFT. https://doi.org/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.