International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Review Article)

The usage of statistical analysis to predict credit card acceptance and income level

Khalid Adnan Ali *, Sara Hussien AbuIktish and Isra Mohammad Hasan

Department of Computer Science and Engineering, American University of Sharjah, Sharjah P.O. Box 26666, United Arab Emirates.

International Journal of Science and Research Archive, 2025, 16(01), 1630-1637

Publication history: Received on 11 June 2025; revised on 18 July 2025; accepted on 21 July 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.1.2171

Abstract

This paper presents a statistical analysis of variables within a credit card approval dataset to predict approval decisions and income level of applicants. Descriptive analysis was performed on continuous variables and inferential statistical techniques, including confidence intervals, t-tests, chi-square tests, and ANOVA, were employed to predict credit card approval. Using a binary logistic regression model, a misclassification rate of 0.48% was achieved. As for predicting income level, Various models were developed and evaluated to identify the most effective approach. Findings indicate that Model 5, which incorporated data centering and adjusted reference levels for categorical variables, demonstrated superior performance by effectively mitigating multicollinearity, though it still had a low R^2 of 20.8%, MAPE of 32.3% and RAE or 0.86.

Keywords: Statistical Analysis; ANOVA; Regression; Multicollinearity

1. Introduction

In the contemporary financial landscape, marked by increasing digitalization, financial institutions are progressively relying on data-driven models to evaluate credit card applications. The accurate assessment of an individual's creditworthiness is paramount for mitigating risks within these institutions. Consequently, the development and utilization of predictive models capable of evaluating applicants based on diverse variables have become critically important.

This study investigates the relationship between various applicant characteristics and credit card acceptance, with the primary objective of predicting both credit card approvals and the income level of the applicant. The initial phase of our research involved a comprehensive analysis of the dataset through visual and descriptive methods to understand its distribution and inherent characteristics. Subsequently, an inferential phase explored statistical relationships among variables using techniques such as Two-Sample T-Tests, Chi-Squared Tests, and ANOVA. Following the data analysis, we constructed predictive models, the performance of which was tested using appropriate metrics: misclassification rate for binary models, R2, Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE) for continuous models. According to the performance of the model, remedial measures were taken to enhance performance.

Unlike many machine learning approaches that often operate as black boxes, this study emphasizes transparency through the use of interpretable statistical models. This approach balances statistical rigor with practical applicability, ensuring that the findings are relevant for both academic research and real-world credit assessment practices. Furthermore, this study provides insights into the limitations of using raw data for financial predictions and underscores the importance of considering interactions between variables and meticulous feature selection in building robust predictive tools.

* Corresponding author: Khalid Adnan Ali

2. Literature Review

Credit scoring and the prediction of credit card acceptance have been long-standing areas of research in finance and statistics. Early approaches primarily relied on expert systems and traditional statistical methods such as logistic regression and discriminant analysis [1]. These methods provided transparent and interpretable models, which were crucial for regulatory compliance and understanding the underlying factors influencing credit decisions.

For instance, a study by Hand and Henley [1] provides a comprehensive overview of statistical methods used in consumer credit scoring, highlighting the importance of robust statistical techniques in assessing credit risk. They emphasize that while complex models might offer marginal improvements in predictive accuracy, the interpretability of simpler statistical models often outweighs these benefits, especially in regulated environments where explaining decisions to applicants is necessary.

This paper builds upon existing literature by focusing on the application of interpretable statistical models to predict credit card acceptance. While acknowledging the advancements in machine learning, we aim to demonstrate that rigorous statistical analysis, coupled with appropriate remedial measures, can yield effective and transparent predictive models. Our work contributes to the ongoing discussion on balancing predictive accuracy with model interpretability in the context of financial decision-making.

3. Methodology

3.1. Dataset Description

The dataset utilized in this study, titled "Credit Card Approval Prediction," was sourced from Kaggle. Its primary objective is to predict credit card approval based on a diverse set of applicant characteristics, referred to as predictors. The dataset comprises 20 distinct predictors, categorized as either continuous or categorical variables.

3.2. Data Preprocessing

Prior to analysis and model building, the raw dataset underwent several preprocessing steps to ensure data quality and suitability for statistical modeling:

- Encoding: Non-binary categorical variables were transformed using label encoding. Binary variables were retained in their original format.
- Missing Values: Any instances of missing or anomalous values within the dataset were identified and removed.
- Feature Removal: Predictors exhibiting zero variance (e.g., variables where all observations had the same value, such as owning a phone) were excluded from the dataset as they provide no discriminatory information.
- Standardization: Continuous variables were centered around their respective means during the application of certain regression models. This centering procedure was primarily implemented to mitigate issues related to multicollinearity among predictors.

Following these preprocessing steps, the final dataset consisted of 12,000 rows, each containing 19 variables which include the predictors and the response variables.

4. Statistical Analysis

4.1. Exploratory Data Analysis (EDA)

4.1.1. Continuous Variable

To assess the distributional properties of the continuous predictors, histograms were plotted for each variable, often overlaid with a fitted normal distribution curve. This visual assessment was crucial for identifying skewness, kurtosis, and the overall shape of the data, which are important considerations for selecting appropriate statistical models and transformations.

Total_Income, Total_Good_Debt, Years_of_Working, and Total_Bad_Debt consistently displayed right-skewed distributions. This indicates that the majority of values are concentrated at the lower end, with long tails extending towards higher values. These variables also exhibited moderate-high kurtosis, suggesting that their distributions are more peaked and possess heavier tails than a normal distribution. Such characteristics often point to the presence of outliers or extreme values, which can significantly influence model performance, particularly in linear regression and ANOVA.

In contrast, Applicant Age appeared to be approximately normally distributed, showing moderate skewness and a platykurtic shape (flatter than a normal distribution). This suggests that Applicant_Age has lighter tails and fewer extreme values, making it a more stable variable for use in parametric statistical tests.

Understanding these distributional properties guided our subsequent modeling decisions. For instance, variables with pronounced skewness were considered for transformations (e.g., Box-Cox transformation) to better meet the assumptions of normality required by certain statistical methods like regression or ANOVA. Early identification of outliers and non-normality also informed decisions regarding remedial measures or the selection of alternative modeling approaches. The corresponding histograms for these variables are presented in Figures 1–5, the findings are also summarized in Table 1.

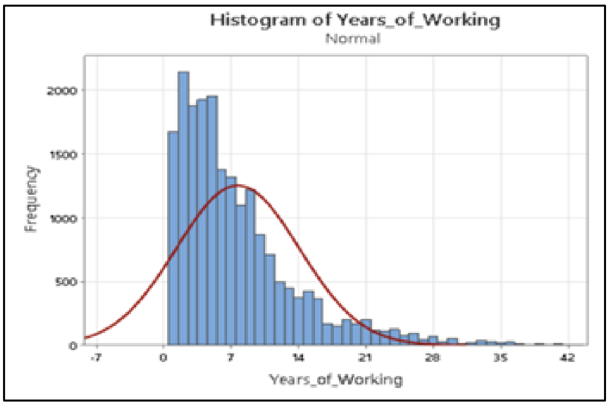

Figure 1 Histogram of Years_of_Working

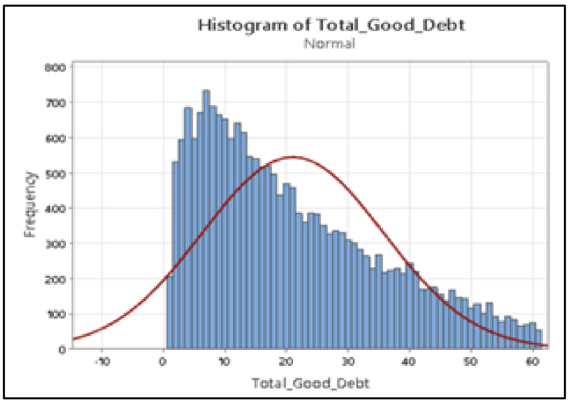

Figure 2 Histogram of Total_Good_Debt

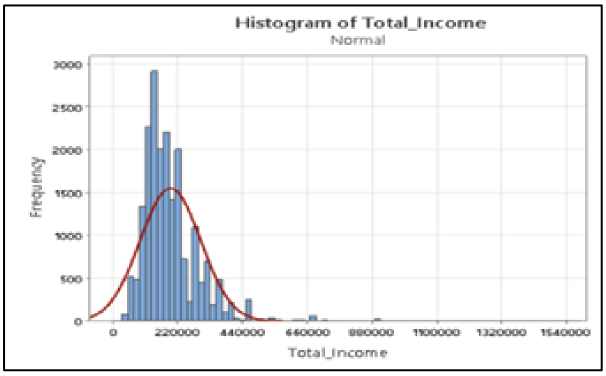

Figure 3 Histogram of Total_Income

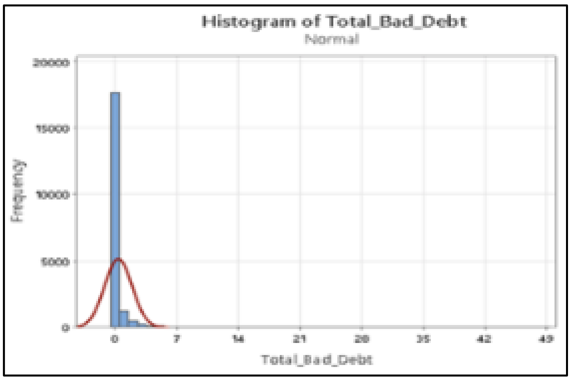

Figure 4 Histogram of Total_Bad_Debt

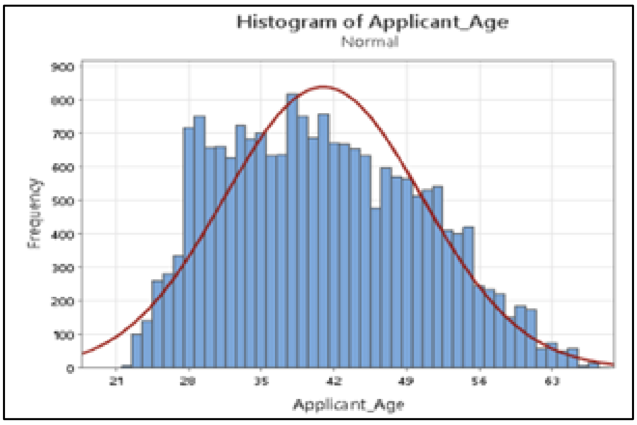

Figure 5 Histogram of Applicant_Age

Table 1 Summary of continuous variables EDA

Variable	Skewness	Kurtosis
Total_Income	Right	High
Total_Good_Debt	Right	High
Years_of_Working	Right	High
Total_Bad_Debt	Right	Moderate
Applicant_Age	Approx. Normal	Low

4.1.2. Categorical Variables

Categorical variables were visualized using bar charts to examine their distributions and facilitate comparisons across different factor levels. These charts provided valuable insights into the frequency of each category within the dataset. For the Family Status variable, a clear concentration was observed in category 2, corresponding to married applicants. In the Income Type variable, categories 3, 4, and 5 showed very low representation, while Income Type 1 (commercial associate) and Type 2 (working) were the most frequent, with Type 2 approximately double that of Type 1. This imbalance in representation was noted as a potential factor influencing model performance. Regarding Education Type, the most common categories were Type 2 (Higher Education) and Type 5 (Secondary Education), indicating that a significant portion of applicants had attained at least a secondary level of education. Gender distribution appeared relatively balanced, with a slight numerical advantage for female applicants. For Total Children, the bar chart indicated an expected decreasing trend in counts as the number of children increased, with most applicants having zero or one child. Job Title displayed a diverse distribution, with Job Title 4 (laborers, sales, security) having the highest frequency and Job Title 6 (cooks, cleaners) the lowest.

4.2. Statistical Inference

To further investigate relationships between variables and support our modeling efforts, various hypothesis tests were conducted.

4.2.1. Chi-Squared Test

Chi-squared tests were employed to assess the existence of correlations between pairs of categorical variables. This analysis was crucial for identifying potential multicollinearity issues within the models. Significant relationships were identified between Housing Type and Total Children, and between Job Title and Education Type. For both pairs, the p-value for the hypothesis test was 0, which is significantly less than the alpha level of 0.05. This led to the rejection of the null hypothesis and hence confirming a statistically significant relationship between these categorical variables. These findings are important for understanding the underlying structure of the data and for guiding feature selection in model development.

4.2.2. One-Way ANOVA

One-Way Analysis of Variance (ANOVA) was conducted to examine the impact of categorical variables with Total income as the response variable.

Gender

Hypothesis Test Method:

- H_0: $\mu_1 = \mu_2$ (The means of Total Income for different genders are equal)
- H_a: At least one of the means is different from the others

The null hypothesis was rejected. This indicates that at least one of the gender categories has a significantly different mean Total Income.

Job_Title

Hypothesis Test Method:

- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ (The means of Total Income for different job titles are equal)
- H_a: At least one of the means is different from the others

Like the gender analysis, the p-value for Job_Title was less than 0.05, leading to the rejection of the null hypothesis. This suggests that at least one job title category has a significantly different mean Total Income.

4.2.3. Two-Way ANOVA

Two-Way ANOVA was performed to investigate the effects of two categorical factors and their influence on the total income as the response variable.

Total_Income versus Income_Type & Job_Title

The p-values for both Income_Type and Job_Title were less than 0.05, indicating that both factors are statistically significant in influencing Total Income. Furthermore, the Variance Inflation Factor (VIF) for most job titles and Income Type 4 were close to 1, suggesting low multicollinearity. However, Income Types 1, 2, and 3 exhibited high VIF values, indicating significant multicollinearity within these categories.

Total_Income versus Total_Children & Family_Status

The p-value for Family_Status was less than 0.05, confirming its statistical significance. Conversely, the p-value for Total_Children was greater than 0.05, suggesting it was not a significant factor. Most VIF values for family statuses and Total_Children categories 3 and 4 were below 5, indicating low multicollinearity. However, Total_Children categories 0, 1, and 2 showed high VIFs, pointing to multicollinearity issues.

Total_Income versus Job_Title & Education_Type

Both Job_Title and Education_Type had p-values less than 0.05, indicating their statistical significance. While most VIF values for job titles were close to 1, Education Type displayed very high VIFs, signifying substantial multicollinearity. This finding suggested that Education Type might need to be removed or transformed in subsequent modeling to address multicollinearity.

The results of the inferential statistics are summarized in Table 2.

Table 2 Summary of Inferential Statistics

Test	Variables	Outcome	Notes
Chi-Square	Housing Type vs Total Children	Significant relationship	Multicollinear candidates
Chi-Square	Job Title vs Education Type	Significant relationship	Multicollinear candidates
One-Way ANOVA	Total Income and Gender	Significantly different means	Residuals non-normal
One-Way ANOVA	total Income and Job Title	Significantly different means	Non-constant variance
Two-Way ANOVA	Income Type & Job Title	Both statistically significant	high VIF for Income Types 1–3
Two-Way ANOVA	Total Children & Family Status	Status statistically significant	High VIF for Total Children Types 0-2
Two-Way ANOVA	Job Title & Education Type	Both statistically significant	High VIF for Education Type

5. Modelling and Remedial Measures

This section details the development of predictive models, and the remedial measures applied to address identified issues such as multicollinearity and improve model performance. Initial visual analysis through matrix plots indicated that while most variables exhibited random relationships, some linear relationships were present, such as between Total Children and Total Family Members, and Applicant Age and Years of Working. These relationships are intuitively expected, as working years typically increase with age, and family size correlates with the number of children. These observations, coupled with the findings from the Chi-square and ANOVA tests, confirm the existence of significant relationships among predictors.

5.1. Binary Logistic Regression Models

Given the presence of 'Status' as an important categorical variable, binary logistic regression models were developed to predict this response variable.

5.1.1. Model 1 (Binary Logistic Regression - Full Predictor Set)

The first binary logistic regression model incorporated all available predictors. Analysis of this model revealed that a majority of the predictors were statistically insignificant at a 10% significance level, as indicated by p-values exceeding 0.1 and confidence intervals containing 1. Notable exceptions were the continuous variables and Total_Bad_Debt , and the categorical variable Job_Title 5 , Total_Good_Debt, which showed low p-values. The model summary indicated relatively high values for AIC (96.68) and BIC (397.21), suggesting considerable prediction error and lower likelihood of accurately predicting the response.

5.1.2. Model 2 (Binary Logistic Regression - Reduced Predictor Set)

For the second binary logistic model, a remedial measure was implemented: the removal of predictors identified as insignificant in Model 1. This refinement resulted in a significant reduction in both AIC (55.67) and BIC (87.30). The lower AIC and BIC values in Model 2 indicate its superiority over Model 1, suggesting reduced prediction error and an increased likelihood of correctly predicting the response variable.

5.2. Multiple Regression Models for Continuous Response (Total Income)

Multiple regression models were developed to predict the continuous response variable, Total Income.

5.2.1. Model 1 (Multiple Regression - Full Predictor Set)

The initial multiple regression model included all 17 available predictors against Total Income as the response. The Analysis of Variance (ANOVA) table indicated that the predictor was owned_Realty insignificant at a 5% significance level (p-value = 0.081), while other predictors were significant. The model also suffered from a lack of fit (p- value = 0), implying that it inadequately described the relationship between predictors and the response, possibly due to missing significant variables. High MSE and SS values further highlighted the model's erroneous nature. The Coefficients table revealed that most predictors had low VIF values (between 1-2), suggesting minimal multicollinearity. However, Education_Type and Housing_Type did exhibit significant VIF values. This was consistent with earlier Chi-square test findings, which showed relationships between the two pairs Education/Job and Housing/Children. The R^2 value of 18% was very low, further explaining the lack of fit and suggesting a weak correlation between predictors and the response. Residual plots confirmed these issues: the normal probability plot showed a lack of normality and linearity, and the versus fits plot indicated non-constant variance (fanning out of residuals). While independence of residuals seemed validated, the assumption of a zero mean might have been violated.

5.2.2. Model 2 (Multiple Regression - Remedial Procedures)

Building on the insights from Model 1, Model 2 incorporated multiple remedial procedures. Firstly, the owned_Realty predictor was eliminated due to its insignificance. Secondly, a Box-Cox transformation was applied to improve the relationship between the response variable and predictors. A log transformation (lambda = 0) was chosen after comparing it with a square root transformation (lambda = 0.5), as it yielded better results.

This attempt proved successful, leading to improvements in R-squared and residual plots. The R-squared increased to 20.76%, indicating a better, albeit still low, correlation. The normal probability plot showed significant improvement, with more points aligning with the line of best fit, suggesting a more normally distributed model. The histogram also displayed a clear bell-shaped curve. Linearity, zero mean, and independence of residuals assumptions were largely validated. However, the constant variance assumption remained slightly violated, though with clear improvement compared to Model 1.

Despite these improvements, Model 2 still suffered from high VIFs, lack of fit, and insignificance of some predictors. Specifically, Total_Children became insignificant (p-value = 0.388), further supported by its consistently low F-value.

5.2.3. Model 3 (Multiple Regression - Further Remedial Measures)

Model 3 continued with remedial measures, eliminating the Total_Children predictor due to its confirmed insignificance. Additionally, interaction terms were explored to enhance the correlation between the response and predictors. An initial attempt with interactions between Job_Title and Education_Type showed minimal effect. Subsequently, an interaction between Applicant_Age and Total_Family_Members was introduced, based on observed relationships in scatter plots.

Model 3 demonstrated modest improvements over Model 2. The R-squared increased slightly to 21.01%, and minor positive changes were observed in the residual plots, particularly in the tails of the normal probability plot and histogram, indicating a slightly more normally distributed model. The versus fits plot showed a slight decrease in the fanning out of residuals, suggesting the variance of residuals was closer to constant. The MSE and SS values significantly decreased, indicating a reduction in average error. However, multicollinearity, as suggested by high VIF values, remained a primary concern.

5.2.4. Model 4 (Multicollinearity Attempt - Predictor Removal):

Following the application of Box-Cox transformation, removal of insignificant predictors, and inclusion of interaction terms, efforts shifted to addressing the persistent multicollinearity indicated by high VIFs. The first approach to mitigate multicollinearity involved removing predictors that showed strong relationships with others, as identified by Chi-square tests. Specifically, Job_Title was eliminated due to its correlation with Education_Type . However, this resulted in a negligible decrease in VIF and a significant drop in R-squared to 17.73%. This indicated that Model 4 was worse than previous models, as it still suffered from multicollinearity and exhibited a weaker correlation.

5.2.5. Model 5 (Multiple Regression - Chosen Model)

Model 5 adopted a different strategy to address multicollinearity. Continuous predictors were centered around their means, and for categorical variables with high VIFs, their reference levels were changed. This approach successfully mitigated multicollinearity, with almost all VIF values falling between 1 and 2, and only two categorical variable levels slightly exceeding 2. While the R-squared value slightly decreased compared to Model 3, it remained higher than Model 2. The minor decrease in R-squared (0.23%) was deemed acceptable given the significant reduction in multicollinearity, which would have had a more detrimental impact on the model. Other statistical aspects of the model remained consistent with Model 3. Consequently, Model 5 was selected as the final multiple regression model, offering the best balance for accurately predicting the Total Income response variable.

6. Model Evaluation

6.1. Binary Logistic Model Evaluation

To assess the effectiveness of our binary logistic model, we predicted the credit card approval status using a testing dataset and compared these predictions against the actual values. The confusion matrix provided a detailed comparison between the actual and predicted outcomes. For instance, 5002 instances where the actual status was 1 were correctly predicted as 1 by our model. Only 24 instances where the actual status was 0 were incorrectly predicted as 1.

The model had an exceptionally low misclassification rate of 0.48%, indicating its high performance. Furthermore, the measures of association table, obtained during model creation, provided additional confirmation of the model's excellence. A 100% concordant percentage was achieved, meaning that for all 1,940,485 pairs, the model's prediction matched the actual dataset's response. This high concordant percentage strongly supports the conclusion that our binary logistic model is highly effective in predicting the credit card approval status.

6.2. Multiple Regression Model Evaluation

Evaluating the multiple regression model, which uses a continuous response variable (Total Income), required a different approach than the binary logistic model. Instead of misclassification rates, we focused on determining the proximity of our predictions to the actual data. This was achieved by feeding the predictors from the testing dataset into the chosen Model 5 to obtain predictions for Total Income.

We utilized four primary error metrics:

- Mean Squared Error (MSE): A very large MSE value was obtained which was expected given the low R-squared value and the previously identified lack of fit in earlier models.
- Root Mean Squared Error (RMSE): Similarly, RMSE also yielded a high value, further indicating the model's limitations in accurately predicting the continuous response.
- Mean Absolute Percentage Error (MAPE): The MAPE was calculated at 32.3%. This signifies that, on average, the difference between the predicted and actual income is 32.3% of the actual value.
- Relative Absolute Error (RAE): The RAE, which compares our model's forecast error to that of a simplistic or naive model, yielded a value of 0.86, indicating that our improved model performs better than a simplistic model

To summarize, our multiple regression model, despite remedial measures, did not perform exceptionally well in accurately predicting the continuous response variable (Total Income). This limitation may be attributed to the absence of other significant variables crucial to better performance, which would also explain the persistently low R-squared values and the model's inability to precisely predict Total Income.

7. Conclusion

This study undertook a comprehensive statistical analysis to predict credit card acceptance, leveraging a dataset from Kaggle. Our methodology encompassed a thorough exploratory data analysis, hypothesis testing using t-tests, Chi-squared tests, and ANOVA, followed by the development and evaluation of both binary logistic regression and multiple linear regression models. The iterative process of model refinement, including the application of remedial measures such as Box-Cox transformation, elimination of insignificant predictors, and strategic handling of multicollinearity through data centering and reference level adjustments, was central to our approach. The binary logistic regression model demonstrated exceptional performance in predicting credit card approval status, achieving a remarkably low misclassification rate of 0.48% and a 100% concordant percentage. This indicates a highly accurate and reliable model for binary classification tasks related to credit card acceptance.

Conversely, the multiple linear regression model, aimed at predicting Total Income, faced greater challenges. Despite significant efforts in model optimization, including addressing multicollinearity, the model exhibited high MSE, RMSE, and MAPE values. While it outperformed a simplistic model (RAE of 0.86), its predictive accuracy for the continuous response variable remained limited. This suggests that the dataset, as provided, may lack certain critical variables necessary for a more robust prediction of Total Income, a limitation reflected in the persistently low R-squared scores.

In summary, this research highlights the efficacy of interpretable statistical models in predicting credit card approval, offering a transparent alternative to opaque machine learning approaches. It also underscores the importance of rigorous data preprocessing and iterative model refinement in addressing statistical challenges such as multicollinearity and non-normality. While the binary classification model proved highly successful, the continuous prediction model revealed inherent data limitations.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Hand, D. J., & Henley, W. E. (1997). Statistical methodology for credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523-541.