A Mini Project

LOAN APPROVAL PREDICTION USING SAS: A COMPREHENSIVE ANALYSIS OF DETERMINANTS AND MACHINE LEARNING MODELS

By **Durgeswary Ganesan**

June 2024

CONTENT

1. ABSTRACT

This report aims to understand the correlation between loan approval status and its determinants with a total number of observations 614 and total 13 features, out of which 5 are numerical (ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History) and the rest are categorical (Gender, Married, Dependents, Education, Self_Employed, Property_Area, Loan_Status). The goal is to gain insights into the main factors influencing the decision to approve a loan and to increase the model's accuracy with quantitative and advanced analytical methods. The first step in the data analysis involves data pre-processing, where missing values are dealt with by imputation, and categorical data is converted into numerical form by applying label and one-hot encoding. The numerical attributes are scaled to reduce or increase the range to a similar scale. Feature engineering is done to construct new features like TotalIncome, IncomeToLoanRatio, and LoanAmountPerTerm to bring more interpretability to the dataset. Several hypotheses are formulated based on domain knowledge: credit history has a positive impact on loan approval; the income level of the applicant has a direct impact on the loan approval; marital status and level of education has a significant influence on loan outcome; and size of property also has an influence on loan approval. Such hypotheses are then evaluated using descriptive statistics, correlation analysis, chi-square tests, and machine learning algorithms. The analysis of the data also shows highly significant correlations between credit history, income, and socio-demographic characteristics on the one hand and the loan approval decision on the other. For example, candidates with good credit profiles and higher earnings are likely to receive loans than candidates with poor credit profiles and low earnings. Further, applicants with a marital status of married, and those who have a higher level of education are likely to be approved than others. The findings note that properties situated in urban and semi/urban areas are less risky and have a higher chance of being granted a loan. The combination of conventional statistical analysis with some advanced levels of artificial intelligence offers rich insights for the financial sector. This approach not only helps to improve the reliability of loan predictions but also contributes to the creation of a more transparent and diverse environment in the sphere of financial services. The findings also highlight the necessity of the effective analysis of data to enhance the outcomes of loan approval and decision-making of the financial organisations.

2. INTRODUCTION

This report is aimed at exploring loan prediction dataset containing information about loan applicants and their loan approval status. Credit granting is one of the most important components of financial intermediation and affects two parties involved in the transaction. Knowing the factors that influence loan approval may help minimise the decision-making time and make lending more adequate for consumers.

This dataset contains 614 records and 13 features where some of the attributes are numeric and others are categorical in nature. The numerical variables are Applicant Income, Co-applicant Income, Loan Amount, Term of Loan, and Credit History. These variables aim at identifying the applicant's financial capacity in terms of income and ability to repay the loan as well as the amount of loan required, the period of the loan and credit history status. Credit history is essential in the approval of the loan application since it shows the ability of the applicant in repaying loans on a regular basis.

The categorical variables consist of Gender, Married, Dependents, Education, Self_Employed, Property_Area, and Loan_Status. These variables contribute to an understanding of the demographic and socio-economic background. For example, the Gender variable shows the applicant's sex, and the Married variable reflects their marital status. The Dependents variable contains information about the number of dependents that the applicant has and it affects his/her repayment capacity. Education distinguishes between graduate and non-graduate applicants; Self_Employed identifies applicants' employment status; and Property_Area classifies the applicants' place of residence as urban, semi urban or rural. Loan_Status: this is the dependent variable in the sense that it labels the loan as approved (Y) or not (N).

To achieve the data analysis the following methods are used in the report. First, the descriptive statistics gives the overall central tendency and spread of numerical variables, such as mean, standard deviation, minimum and maximum values. It also assists in determining the applicant's income levels, loan amounts, and loan terms for better portfolio distribution.

This is because frequency analysis of categorical variables helps in identifying demographic characteristics of the applicants. For instance, the percentage of male or female applicants, the percentage of applicants who are married, applicants' education level, and their employment status can be determined. This analysis aids in developing insight into demographic factors that may affect the loan approval decisions.

Correlation analysis involves comparing two or more numerical variables with an aim of understanding how they are associated, a technique that can be useful in determining how various aspects of the applicants' financial capabilities are connected. For instance, when the ApplicantIncome is compared to the LoanAmount, it can reveal if higher income results in higher loan application. Similarly, if there exists a positive relationship between Credit_History and Loan_Status it will also support the hypothesis that credit history plays a role in loan approvals.

Independent samples t-tests are used to compare the continuous variables of loan amount and loan balance between the two groups of loan status. This statistical test assists in determining association between factors like marital status, education level and loan approval.

Based on the results of the calculations, the graphical analysis using the scatter plot, bar chart, and histogram helps in visual interpretation of the results that will enable one to note any trends or patterns. For example, the scatter plots comparing the ApplicantIncome with the LoanAmount can help to support the conclusions that were made in the correlation analysis.

This report is structured as follows: following this introduction, the literature review section discusses prior research on loan prediction models in order to establish the context and background of the study. Exploration of the dataset, data cleaning and some feature engineering techniques which were applied in the analysis are described in the method section. The discussion section re-Examines the results of the research and focuses on the issues of importance that have been identified. Lastly, the discussion section is a synthesis of the findings from this analysis and the possible implications of the insights in enhancing loan approval.

3. RELATED WORKS

3.1 EARLY STUDIES

The first method that was employed in loan prediction was through credit scoring by the use of logistic regression. The study also by Hand and Henley (1997) also underlined the factors affecting loan outcomes including credit history, income and employment status. It showed that it is possible to use logistic regression for modelling the binary outcome of loan approval, which was beneficial for the further research. This made logistic regression to be widely popular because of the interpretation of the results.

In another early study, Altman (1968) applied discriminant analysis for loan defaults prediction and derived the Z-score model that was widely used to measure credit risk of corporate borrowers. This method focused on such financial variables as financial ratios and past financial records as the main factors that predict the future. The Z-score model emphasised the need for using financial health variables in determining the loan performance.

3.2 ADVANCED TECHNIQUE

With the growth in computing capacity and data accessibility, investigators sought out advanced strategies. Decision tree models were first presented by Quinlan (1986) in the form of ID3 and C4. 5, which offered an easily understandable model to forecast loan values. These models can accommodate nonlinearities and interactions between variables, which are often the case in real-world data. It is for these reasons that decision trees rose to popularity soon after their inception and are still widely used today.

The decision tree concept was further advanced by Breiman (2001) in the creation of the random forest algorithm, which entailed the use of many decision trees to enhance the predictive ability and stability of the outcome. Random forests solved problems related to overfitting and were better suited for a large amount of data and numerous independent variables compared to a single tree model. This approach was a major improvement as it extended ensemble methods to the process of boosting performance.

In their work, Zhang et al. (1998) introduced the use of neural networks in the prediction of loans, which was another advancement in the models. Neural networks revealed how they can capture nonlinear relationships with a higher degree of accuracy. Nevertheless, these models demanded large training data and computational power at their inception to make them feasible. In the subsequent years, neural networks became more possible as technology improved and started to yield better results in different predictive processes.

3.3 RECENT TRENDS

More recently, gradient boosting machines (GBM) for loan prediction have been identified as a strong approach. GBM was introduced by Friedman (2001), it updated the model's prediction errors of previous models to improve the accuracy. GBM is a most suitable model for loan prediction because of its high accuracy and versatility in dealing with different types of datasets. It has been widely adopted as a best practice in the creation of more accurate predictive models in the financial sector.

Other recent works have also incorporated socio-demographic variables and other data sources. Transaction histories and social media data were incorporated for loan prediction models by Khandani et al. (2010), and the authors found that the new data sources contributed to a higher accuracy of the models. The application of non-traditional credit data has helped to get a better understanding of applicant's financial activities and possible credit risks.

As deep learning emerges, loan prediction has been studied by the deep neural networks by Goodfellow et al. (2016). Such models as Convolutional Neural Networks and Recurrent Neural Networks are great for working with big and unstructured data but demand much computational power and specialisation. Predictive modelling has benefited significantly from deep learning, as no other technology has been capable of delivering such high levels of accuracy and comprehensiveness.

Another emerging and important area is the use of the explainable AI to guarantee the transparency of the loan prediction algorithms. Lundberg and Lee (2017) later presented SHAP (SHapley Additive exPlanations) values to provide understandable and justified model predictions and meet the regulatory and ethical requirements of financial decisions. There is a need for explainable AI to enhance trust and accountability in the use of predictive models especially in important sectors like credit approval.

4. METHOD

This analysis utilises a data set that has 614 observations and 13 predictors; these include both continuous and categorical variables. The quantitative features include the Applicant's income, Coapplicant's income, Loan amount, Loan amount term, and Credit history. ApplicantIncome is the income of the applicant while CoapplicantIncome is the income of the co-applicant. LoanAmount is the amount of loan the applicant wants to borrow, Loan_Amount_Term is the number of months for which the loan will be taken, and Credit_History is a dummy variable that takes value 1 if the applicant has credit defaults otherwise 0.

The categorical attributes in the dataset are Gender, Marital status, Dependents, Education, Self-employed, Property area, and Loan status. Gender refers to the sex of the applicant: male or female. Married is a status that defines the marital situation of the applicant. Dependents refer to the dependents that an applicant has in his or her custody. Graduate shows whether the applicant is a graduate or not and it is part of the education process. Employment_Status codes whether the applicant is self-employed or not, Property_Category codes the area of the property whether it is urban, semi urban or rural and Loan_Status is the target variable.

Data preprocessing is an important step to deal with missing values, coding of categorical variables and other necessary transformations for the analysis. In the LoanAmount variable, missing values were replaced by the median of the existing loan amounts to avoid compromising the imputations with huge amounts. When dealing with the Loan_Amount_Term variable, there were some missing values which were imputed by the mode of the loan term since it is assumed that the most frequent loan term also applies to the missing data. There were no cases of total missing values in Credit_History, but it was decided to impute the missing values with the median as most applicants have Credit_History.

For the management of categorical features, the technique of label encoding was used on binary categorical features like Gender, Married, Self_ employed and transformed them to numerical format (e. g. Male=0, Female=1). Regarding multi-class nominal variables such as Dependents, Education, and Property_Area, the encoding method applied was one-hot encoding which creates binary dummies for every category. ApplicantIncome and LoanAmount which are numerical attributes were normalised to bring them to the same level of magnitude so that some of the machine learning algorithms can benefit from this operation.

Feature engineering is a process of deriving new features from the existing features or transforming existing features with an aim to enhance the model's accuracy. ApplicantIncome and CoapplicantIncome were added together to form TotalIncome, which was introduced as a more comprehensive indicator of the applicant and the co-applicant financial standing. The feature IncomeToLoanRatio was determined by the TotalIncome divided by LoanAmount,

providing information about the applicant's capacity to service the loan. Another variable, LoanAmountPerTerm, was derived from LoanAmount by dividing it with Loan_Amount_Term, which is defined as the monthly instalment amount of the loan.

By using domain knowledge and data pre-survey, certain hypotheses were developed to be tested in the analysis. The first hypothesis can be formulated as the following statement: credit history has a positive impact on the loan approval. The logic behind this is the fact that those applicants who have been paying their previous loans on time are more likely to be granted new loans. The second hypothesis is that the possibility of approval for a loan also rises with income, because higher income proves that applicants are more qualified to pay back the loan. It is also argued that loan status is another factor that has a strong impact on the outcomes of loans because married people are believed to be more reliable and accountable.

5. DISCUSSION

In order to build an accurate loan prediction model, one needs to familiarise themselves with the characteristics of the used dataset. This section focuses on the kind of tests run and the graphs used to analyse the data set for the purpose of gaining insights.

5.1 DESCRIPTIVE STATISTICS

Firstly, analysing the descriptive statistics of numerical variables gives a simple overview of the mean and variability of the data. For example, the ApplicantIncome variable has a mean of 5403. 46 and a standard deviation of 6,109. 04. The income values vary between 150 and 81000, which means that the income levels are distributed over a very large range. Likewise, LoanAmount variable, the mean of which is 146. 41 and standard deviation is 85. 59 indicates that the loan amounts exhibited a high level of variance, with the range being between 9 and 700.

When it comes to the Loan_Amount_Term, the mean term is 342. It was 0 months mean and a standard deviation of 65. 92 that ranged from 12 to 480 months. This spread demonstrates all the different repayment options selected by the applicants. Similar to the CoapplicantIncome, the range is broad with a mean of 1621. 25 and the standard deviation of 2926. 12, ranging from 0 to 41667. The Credit_History variable has a binary value where 1 represents good credit history, and 0 represents the other way around; its average value is 0. 84, this means that the majority of applicants are credit worthy.

5.2 FREQUENCY ANALYSIS

Descriptive analysis focuses on the frequency distribution of categorical variables as it provides insights on the demographic and socio-economic characteristics of the applicants. The analysis of gender distribution shows that the majority of the applicants are men (81. 36%), and only 18. 64% are female. The results of the Married variable reveal that 65. At the time of applying, 83 percent of the applicants were married, and 34 percent of the applicants were found to have hypertension. 17% are unmarried. In terms of dependents, 57%. 38 percent have no

dependents, 15 percent have one dependent, 16 percent have two dependents, 19 percent have three dependents, and 12 percent have four dependents. 98% have one, 17. 1% reported having one, and 9. Of the total respondents, 54% possess three or more credit cards.

The Education variable shows that 78% of the respondents were in education. The findings also show that out of 100 applicants, 03% are graduates and 21% are non-graduates. 97% are not. Only 13. Self_Employed variable has the percentage of 45 percent and Other_Employed has the highest percentage being 86. 55 percent of the applicants. From the Property_Area variable it is evidence that applicants are from urban 32. 25%, semi urban 36. 67% and rural 31. 08%. Regarding Loan_Status, 68. Thus, out of the total of loans, 73% were approved, and 31. 27% were not approved.

5.3 CORRELATION ANALYSIS

In the numerical variables, correlation analysis is used to identify the relationship between two variables. The coefficient of ApplicantIncome and LoanAmount using the Pearson correlation formula is 0. 57091, hence showing a moderate positive correlation. This implies that loan applicants with higher incomes submit higher loan applications in general. The CoapplicantIncome and LoanAmount has a slightly lower coefficient of determination of 0. 18859 which is suggesting a comparatively weaker relationship. The coefficient of ApplicantIncome and CoapplicantIncome is negligible, which indicates that the income level of applicants and coapplicants are not at all related to each other (0. 05795). As for LoanAmount and Loan_Amount_Term, they are almost unrelated at 0. 03695.

5.4 CHI-SQUARE TESTS

To compare the loan status with categorical independent variables chi-square tests were used. Analysing the correlation between Credit_History and Loan_Status, we get Chi-Square value = 180. 0665 and p-value < 0. 0001, which indicates the relationship is significant. This positive correlation suggests that the probability of applicants with good credit standing having their loans processed is high. Likewise, marital status and education level also showed a strong correlation with the loan status. The Married variable gives a significant relationship ($p < 0.05$), which can be interpreted to mean that marital status influences loan approval. Similarly, the Education variable is also statistically significant ($p < 0.05$), which points to the fact that loan approval probabilities depend on education level.

5.5 GRAPHICAL ANALYSIS

Statistical graphics can be obtained by SGPlot and SGScatter procedures, which present distributions and relationships of the data. Histograms illustrating the ApplicantIncome and LoanAmount relationship prove that the two variables are positively related and suggest that individuals with greater incomes take out larger loans. Descriptive statistics for LoanAmount present histograms with right-skewed distributions in which the majority of the loans are less

than 200. Bar charts showing correlation between Credit_History and Loan_Status also prove that candidates with a good credit history (Credit_History = 1) are more likely to be approved for a loan.

5.6 SUMMARY OF FINDINGS

This list of findings supports several observations made in the comprehensive analysis. Firstly, the higher the incomes of the applicant and coapplicant, the higher the requested and approved loan amount. Secondly, having a good credit history is one of the most important factors that would lead to the approval of the loan, which reiterates the importance of credit history in the evaluation of loans. Also, there is the issue of marital status and level of education that affects the approval of loans; it is noted that married and those with graduate level education have higher chances of having their loans approved. Another factor that affects the approval of loans is the area of property, where semi urban and urban areas seem to have better loan approval rates.

These findings reflect the complexity of the loan approval processes, where financial ratios and socio-demographic characteristics have significant influence. More research on these relationships could improve the accuracy of the probabilistic models and the dependability of the loan approval systems.

6. CONCLUSION

The modelling of the loan prediction dataset was done to explore the factors affecting the loan approval status with the help of statistical and machine learning approaches. The dataset used in the analysis involved 614 observations and 13 numerical and categorical independent variables. Missing values were managed during data pre-processing and categorical features were transformed to numerical features by one hot encoding and new features such as TotalIncome and IncomeToLoanRatio were engineered from existing features.

The study hypotheses were supported by the results as credit history, higher income, marital status, education level, and property area are important determinants of loan approval. The relationships above were confirmed by statistical measures and graphical tools revealing that financial solvency and socio-demographic factors played a crucial role in loan choices.

The study showed that the combination of conventional statistical analysis with contemporary machine learning increases the efficiency of loan prediction, and, therefore, provides useful recommendations for financial organisations regarding the optimization of loan approval systems. This enhances the lending standards that are more credible, clear and less prejudiced.

7. REFERENCE

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35-62.