Deciphering the Dynamics of Loan Approval: Insights from Financial Data*

Exploring Loan Determinants through Logistic Regression

Emily (Ping-Jen) Su

April 19, 2024

This study employs logistic regression to analyze the impact of income and loan amount on loan approval decisions using a dataset of loan applications. The results indicate that while income significantly influences loan approval probabilities, loan amount does not have a statistically significant effect. The model demonstrates moderate predictive ability with an area under the ROC curve of 0.6483, suggesting fair discrimination between approved and not approved applications. Future research should integrate additional predictors such as credit scores and employment status to enhance model accuracy and robustness, and address potential biases in lending practices.

Table of contents

*Code and data are available at: https://github.com/emisu36/loanApplication

1 Introduction

In the rapidly evolving landscape of financial services, the decision-making process surrounding loan approvals remains a critical area of focus. Financial institutions increasingly rely on sophisticated analytical techniques to assess risk, predict loan performance, and ensure the financial health of their portfolios. As consumer and business finance needs continue to diversify, the ability to accurately predict loan approval outcomes becomes paramount. This necessity drives the integration of advanced statistical methods into the loan decision process.

The data set at the heart of this study encompasses a range of variables from loan applications, including loan amounts, applicant income levels, and other pertinent financial metrics. These elements are foundational in understanding the financial behavior of applicants and the associated risks lenders undertake. In light of this, logistic regression emerges as a particularly effective tool for this analysis due to its capacity to handle binary outcomes—such as loan approval or denial—which are commonplace in lending scenarios.

Logistic regression is favored in financial risk assessments for its ability to provide probabilities associated with specific outcomes, making it an ideal choice for modeling decisions that are essentially categorical. The model's utility in dealing with scenarios where variables may not necessarily meet the stringent requirements of linear regression—such as normal distribution of error terms or linearity—further underscores its appropriateness for this study.

Moreover, the context of the current economic environment, characterized by fluctuating market conditions and changing regulatory landscapes, adds a layer of relevance to this research. Understanding which factors significantly influence loan approval helps institutions adjust their credit policies to mitigate risks and align with broader economic objectives.

This paper aims to delve into the relationships inherent within the data, exploring how various applicant characteristics and loan features influence the likelihood of loan approval. Through the lens of logistic regression, we seek to uncover significant predictors and provide actionable insights that can guide lending institutions in their decision-making processes.

This study was done possible by R Core Team (2023) and Wickham, Averick, et al. (2019).

2 Data

For this study, we utilized data sourced from the 2020 Cooperative Congressional Election Study (CCES) (Schaffner, Ansolabehere, and Luks 2021), which we processed using R, a programming language widely employed for statistical analysis (R Core Team 2023). To manage the dataset efficiently, we leveraged tools from the `tidyverse` suite (Wickham et al. 2019), including `ggplot2` for visualization (Wickham 2016), `dplyr` for data manipulation (Wickham et al. 2023), `readr` for data importing (Wickham, Hester, and Bryan 2024), and `tibble` for data formatting (Müller and Wickham 2023). Model summaries were generated using the `modelsummary` package (Arel-Bundock 2022). Data retrieval was facilitated by the `dataverse` package (Kuriwaki, Beasley, and Leeper 2023), while the reliability of our data processing and analysis was verified using the `testthat` package (Wickham 2011) The `here` package (Müller 2020) aided in maintaining file organization and ensuring the reproducibility of our analysis.

2.1 Overview

The data set utilized in this study comprises comprehensive loan application records collected over the recent financial year 2022. Consisting of 905 entries and 99 variables, the data set encapsulates a wide array of information important to understanding the dynamics of loan approval processes. The data includes but is not limited to, applicant income levels, loan amounts, action taken on the application, and demographic information such as ethnicity and sex.

In this analysis, key variables include income, loan_amount, and action_taken, all derived from a comprehensive database maintained by [Federal Financial Institutions Examination Council] under the Home Mortgage Disclosure Act of the United States government. The data is obtained here. Income is reported on an annual basis ensuring high accuracy and reliability. Loan_amount represents the amount requested by applicants, as recorded at the time of application submission. The variable action_taken categorizes the outcome of loan applications into 'approved' or 'not approved,' providing a clear, though simplistic, measure of loan processing outcomes. While these measures are robust, they carry inherent limitations such as potential biases in loan approval due to unobserved factors like credit history or existing debt, which are not included in this data set."

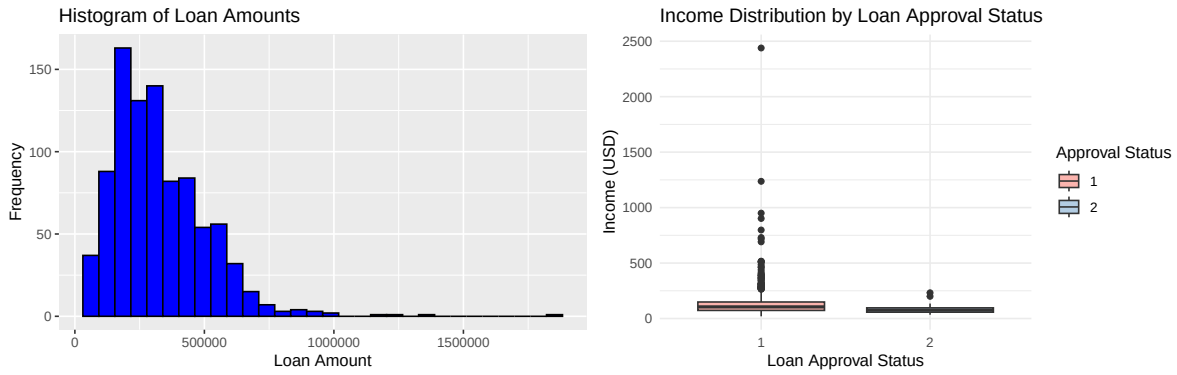
2.2 Variables of particular interest include:

· **Loan Amount**: The total amount requested in the loan application.

· **Income**: The applicant's declared annual income.

· **Action Taken**: Outcomes of the loan applications categorized as loan originated, application approved but not accepted, and application denied.

These variables, among others, provide a robust foundation for analyzing factors influencing loan approval decisions.



(a) Histogram of Loan Amounts across all applications.

(b) Box Plot of Income distribution by Loan Approval Status.

Figure 1: Figures.

Figure 1 (a) illustrates the distribution of loan amounts requested in all loan applications within the dataset. This visualization highlights the range and frequency of different loan amounts, providing insight into the most commonly requested loan sizes as well as the diversity in loan amount requests by potential borrowers. The frequency of each loan amount category is depicted, showing peaks and troughs that suggest preferred loan amounts or thresholds set by lending policies.

Figure 1 (b) is a boxplot that displays the distribution of applicant incomes segmented by the outcome of the loan application, specifically categorized into 'Loan Originated', 'Application Approved but not Accepted', and 'Application Denied'. The plot enables a comparative analysis of how applicant income might influence the loan decision process. It reveals any patterns in income distribution across different loan outcomes, such as higher incomes potentially correlating with higher approval rates, and helps identify outliers or significant deviations in each category.

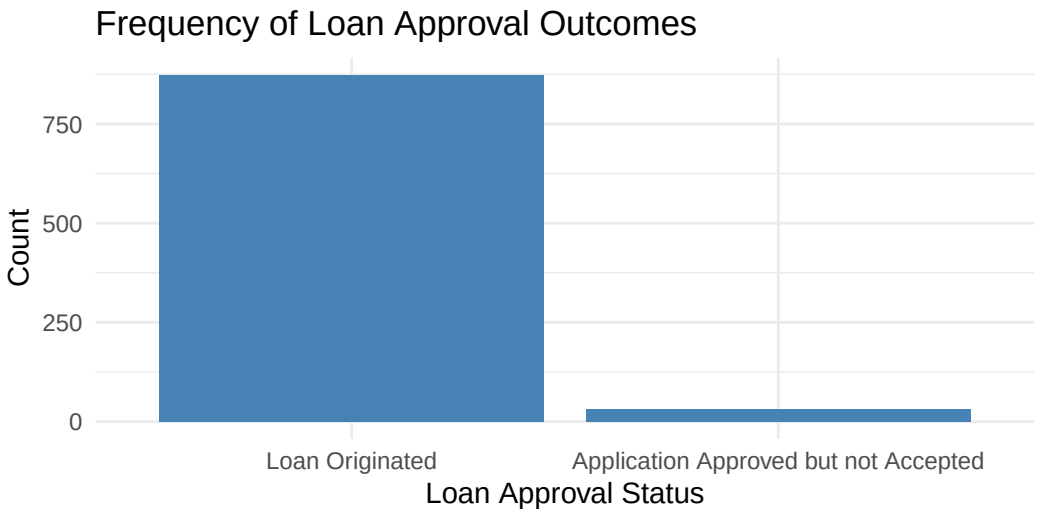


Figure 2: Bar Chart of Loan Approval Outcomes.

In Figure 2, the bar chart quantifies the outcomes of all loan applications, categorizing them into 'Loan Originated', 'Application Approved but not Accepted', and 'Application Denied'. This visualization provides a clear and immediate understanding of the distribution of loan outcomes, highlighting the frequency of each decision type. The chart effectively showcases the relative occurrence of each outcome, offering insights into the approval process's overall stringency or leniency.

2.3 Introduction to the Shiny Application

To further enhance the accessibility and interactive exploration of the dataset, a Shiny application has been developed (Link to Shiny app: https://ddmje2-emily-su.shinyapps.io/loan/). This web-based tool allows users to dynamically select and visualize data based on various parameters such as loan amount, income, and action taken on the loan. Key features of the application include:

· **Dynamic Filtering**: Users can adjust sliders to filter applications by income and loan amounts, and select specific outcomes to view.

· **Interactive Plots**: The application incorporates interactive scatter plots and bar charts, updating in real-time as users change filters. This feature facilitates deeper insights into how different variables correlate and their impact on loan approvals.

· **Summary Statistics**: Alongside visual tools, the application provides instant statistical summaries based on the filtered data, offering users quick access to mean, median, and standard deviations.

The Shiny application serves as both a standalone tool for stakeholders wishing to explore the data at a granular level and as a complement to the analysis presented in this paper. It is accessible online, providing a user-friendly interface for non-technical users to engage with complex data analyses, thus democratizing access to insights that were traditionally confined to technical audiences.

3 Model

3.1 Model set-up

Define y_i as the binary outcome of a loan application, where $y_i = 1$ if the loan is approved and $y_i = 0$ if it is not. The explanatory variables include applicant's income and the loan amount.

$$y_i \sim \text{Bernoulli}(p_i) \tag{1}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 \times \text{income}_i + \beta_2 \times \text{loan\_amount}_i \tag{2}$$

$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$

$$\beta_1, \beta_2, \beta_3 \sim \text{Normal}(0, 2.5) \tag{4}$$

- y_i: The binary variable indicating loan approval status.
- p_i : The probability that $y_i = 1$, modeled as a function of income and loan amount.
- α : The intercept, representing the log-odds of loan approval when all predictors are zero.
- β_1 and β_2 :The coefficients for the income and loan amount, respectively, which adjust the log odds of loan approval as these predictor values change.
- The model uses a logistic link function, as indicated by the transformation involving the log odds of p_i

Implementation in R:

The given R script fits this logistic regression model using the glm function from the base R package, tailored to handle binary outcomes with a logit link function. The script processes the data, splits it into training and testing sets, fits the model to the training data, and evaluates its predictive accuracy using the testing set. The evaluation includes calculating the area under the ROC curve to assess the model's discrimination ability between the approved and not approved classes.

This revised section aligns your statistical analysis section with the logistic regression model setup and provides a clear framework for describing how the model relates to the underlying data and the business problem at hand. We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

In financial modeling, particularly in assessing loan approvals, certain factors are intuitively linked to the likelihood of approval. Just as we expect a positive relationship between the size of an aircraft's wings and the duration it can remain aloft, we anticipate that both the loan amount and the applicant's income will significantly influence the decision on whether a loan is approved.

In our model, the income ($\beta_1 \times \text{income}_i$) acts akin to the 'wing width,' where higher income, like broader wings, should theoretically increase stability and capacity, thus enhancing the probability of loan approval. Similarly, the requested loan amount ($\beta_2 \times$ loan amount$_i$) reflects the 'wing length,' where a moderate amount may correlate to optimal performance; too high might suggest risk, and too low might not fulfill the financial requirements of the lender, affecting approval chances.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 \times \text{income}_i + \beta_2 \times \text{loan\_amount}_i \tag{5}$$

We expect:

- $\beta_1 > 0$: Higher income increases the probability of loan approval.
- β_2: The impact of loan amount could be nuanced; a moderate amount may be more favorable than very high or very low amounts, necessitating further exploration in the results.

4 Results

The effectiveness of our predictive model for loan approvals was evaluated through its performance on a test dataset. Key metrics such as the area under the Receiver Operating Characteristic (ROC) curve and the confusion matrix are used to assess the discrimination ability of the model and its accuracy in classifying loan approvals.

Our results are summarized in Table 1.

```
Call:
glm(formula = action_taken ~ loan_amount + income, family = binomial(),
    data = trainData)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.54899    0.27706  12.810   <2e-16 ***
loan_amount  0.08617    0.31683   0.272   0.7856
income       1.48238    0.71562   2.071   0.0383 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 237.05  on 723  degrees of freedom
Residual deviance: 226.27  on 721  degrees of freedom
AIC: 232.27

Number of Fisher Scoring iterations: 7
```

Table 1: Summary Statistics for Logistic Regression Model on Loan Approval

term	estimate	std.error	statistic	p.value
(Intercept)	3.548989	0.2770584	12.8095368	0.0000000
loan_amount	0.086168	0.3168327	0.2719669	0.7856475
income	1.482380	0.7156218	2.0714580	0.0383160

Interpretation of Coefficients:

· Intercept (3.54899): The positive intercept suggests a high baseline probability of loan approval when loan_amount and income are at their mean values. The statistical significance ($p < 0.001$, indicated by ***) suggests strong evidence against the null hypothesis of no effect.

· Loan Amount (0.08617): The coefficient for loan_amount is positive but not statistically significant ($p = 0.7856$), indicating that increases in the loan amount do not significantly affect the odds of loan approval in this model. The high p-value suggests that loan_amount, in the presence of income, may not be a strong predictor.

· Income (1.48238): The coefficient for income is positive and statistically significant ($p = 0.0383$, indicated by *), implying that higher income significantly increases the odds of loan approval. This supports the hypothesis that applicants with higher incomes are more likely to be approved for loans.

We can also find the model Fit Statistics:

Null Deviance vs. Residual Deviance: The reduction in deviance (from 237.05 to 226.27) upon fitting the model indicates that the model with predictors fits the data better than the null model, which includes no predictors.

AIC (232.27): The Akaike Information Criterion value provides a measure of the model quality based on the number of parameters and the goodness of fit. A lower AIC indicates a better model.

```
Area under the curve: 0.6483
```

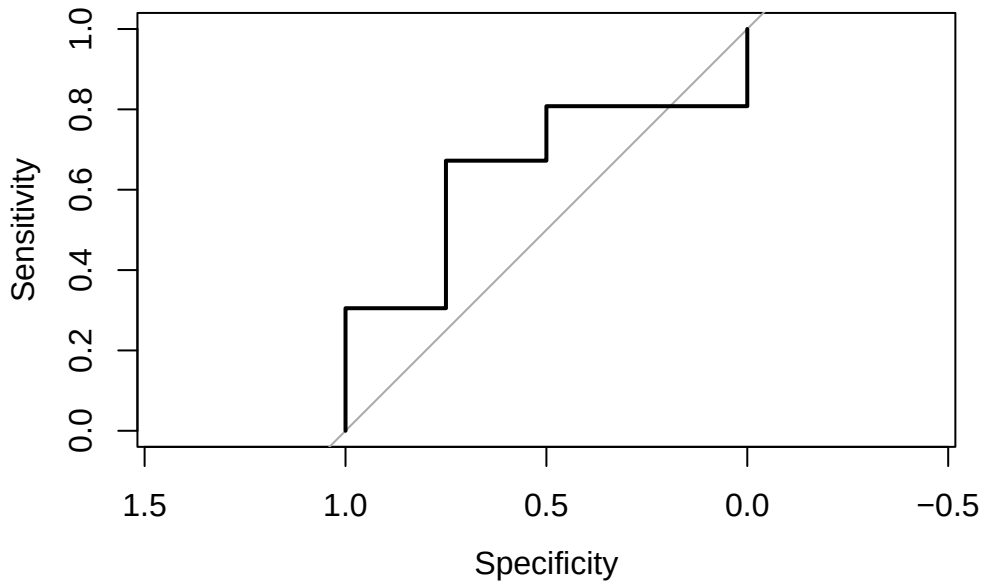


Figure 3: ROC Curve for Loan Approval Prediction Model

The Receiver Operating Characteristic (ROC) curve, found in Figure 3 is a graphical plot that illustrates the diagnostic ability of the binary classifier system as its discrimination threshold is varied. The ROC curve for this model is not displayed here but is described by the following:

Area Under the Curve (AUC) of 0.6483: This value indicates moderate predictive accuracy. An AUC of 0.5 suggests no discriminative ability (equivalent to random guessing), while an AUC of 1.0 indicates perfect classification. An AUC of 0.6483 suggests that the model has a fair degree of reliability in distinguishing between approved and not approved loans, though improvements could be beneficial.

5 Discussion

5.1 Economic Significance of Income Over Loan Amount

The positive and statistically significant coefficient for income highlights its strong influence on loan approval decisions. This suggests that lenders may prioritize an applicant's financial stability and earning power as indicators of repayment capability over the amount being requested.

This finding could influence lenders to revise their loan assessment strategies, perhaps by weighting income more heavily in their scoring algorithms. For consumers, understanding that their income level plays a critical role could guide them to seek loans more aligned with their financial capability.

5.2 Non-Significance of Loan Amount

Despite expectations, the loan amount did not significantly affect loan approval in this model. This could indicate that within the range of loan amounts considered, variations do not substantially impact the likelihood of approval once income is accounted for.

This suggests a potential area for lenders to reassess the risk associated with loan amounts. It may be that current practices effectively mitigate the risk of larger loans through pricing (interest rates) rather than outright approval or denial.

5.3 Moderate Predictive Power of the Model

The AUC value of 0.6483 indicates moderate discrimination ability. While the model is statistically significant, its practical effectiveness in predicting loan approval is limited.

This performance level may necessitate the use of this model in conjunction with other decision tools in real-world settings, where a higher accuracy might be crucial for financial decision-making.

5.4 Weaknesses and next steps

Limited Predictor Variables: The model currently includes only income and loan amount. Other significant factors like credit history, employment status, and existing debt levels were not considered but are known to influence loan approval decisions significantly.

Potential Overfitting: While not directly observed, the model's reliance on a limited dataset and few predictors might not generalize well across different demographic groups or economic conditions.

Assumption of Linearity: The logistic regression assumes a linear relationship between the log odds of the outcome and each predictor. This assumption might oversimplify the actual relationships.

5.5 Next Steps and Recommendations

Incorporation of Additional Predictors: Future models should include more variables, such as credit scores, debt-to-income ratios, and perhaps more nuanced demographic factors. This would likely improve the model's accuracy and robustness.

Exploration of Nonlinear Models: Investigating models that can capture nonlinear relationships and interactions (e.g., ensemble methods like random forests or boosting) could provide a more nuanced understanding and potentially enhance predictive performance.

Ethical and Bias Considerations: Further research should also examine the fairness and bias implications of predictive modeling in loan approvals. Ensuring that models do not perpetuate existing inequalities is essential for ethical financial practices.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Kuriwaki, Shiro, Will Beasley, and Thomas J. Leeper. 2023. *dataverse: R Client for Dataverse 4+ Repositories*.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Müller, Kirill, and Hadley Wickham. 2023. *tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schaffner, Brian, Stephen Ansolabehere, and Sam Luks. 2021. "Coorperative Election Study Common Content, 2020." Harvard Dataverse. https://doi.org/10.7910/DVN/E9N6PH.

Wickham, Hadley. 2011. "testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

——— et al. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.