STAT 448 Final Project: Predicting Home Loan Approval Using Socio-Economic Factors

Group 18

Team Members:
Mauro Canchola (EDA, Modeling)
Steven Chen (Inference, Diagnostics)
Aditya Rajan (Interpretation, Writing)

Fall 2025

Contents

Abstract

This project examines socio-economic and financial factors associated with home loan approval using data from a Kaggle home loan application dataset. Descriptive analysis, hypothesis testing, and logistic regression modeling are used to identify key predictors of loan approval. Credit history emerges as the dominant factor, with additional contributions from property area and marital status. The final model demonstrates good predictive performance and provides interpretable insights relevant to automated lending decisions.

1 Introduction

The process of evaluating home loan applications is a central task for financial institutions, as it requires balancing access to credit with the management of default risk. Traditionally, loan approval decisions have relied on a combination of applicant demographics, financial characteristics, and credit history. With the increasing availability of structured application data, statistical modeling provides a systematic framework for understanding which factors most strongly influence loan approval and for supporting automated decision-making.

In this project, we analyze a home loan application dataset obtained from the Kaggle repository to identify socio-economic and financial predictors associated with loan approval. The primary objective is to model the probability that a loan application is approved as a function of applicant characteristics, including income, credit history, marital status, education level, self-employment status, and property location. In addition to developing a predictive model, we aim to provide interpretable results that highlight the relative importance of these predictors in the loan approval process.

Our analysis proceeds in several stages. First, we conduct a descriptive and exploratory analysis to summarize the distribution of applicant characteristics and loan outcomes. Next, we perform formal statistical tests to assess associations between loan approval status and both categorical and continuous predictors. Finally, we fit and evaluate a logistic regression model using variable selection and diagnostic checks to identify the most influential factors and assess predictive performance. The findings of this study provide insight into how lending decisions are reflected in the data and demonstrate the use of statistical modeling for binary classification problems in a real-world financial context.

2 Data Description

The data used in this study come from the *Home Loan Approval* dataset obtained from Kaggle and contain information on 480 loan applicants. Each observation corresponds to a single loan application and includes demographic, socio-economic, and financial characteristics, along with the final loan approval decision.

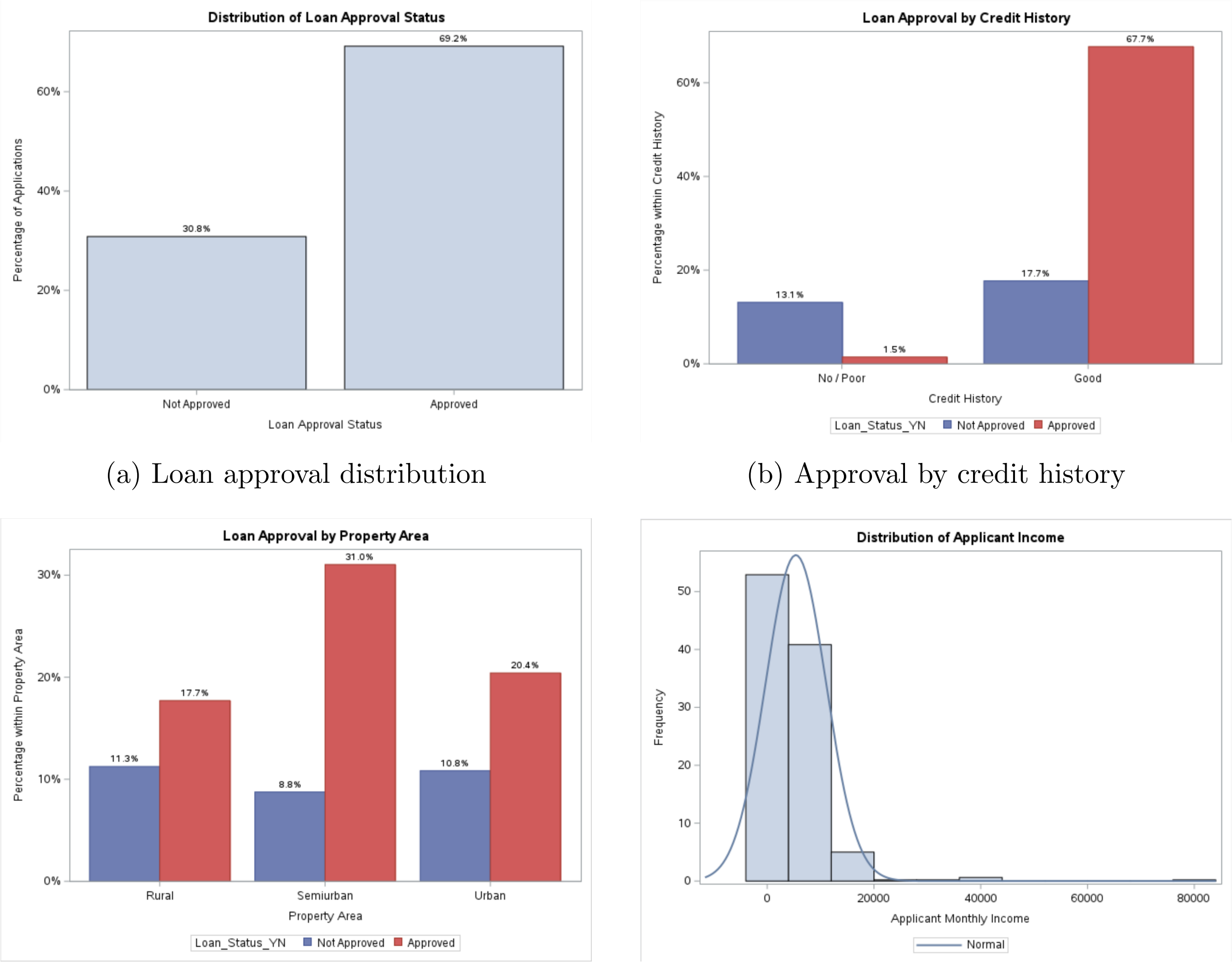
The response variable, `Loan_Status`, indicates whether an application was approved (`Yes`) or not approved (`No`). Predictor variables include applicant demographics (gender, marital status, number of dependents, education level, and self-employment status), financial characteristics (applicant income, co-applicant income, loan amount, and loan term), credit history, and property area (rural, semi-urban, or urban).

Basic data preprocessing was performed to ensure consistency and reliable inference. Categorical variables were recoded to standardize labeling, and rare or missing categories with very small counts were removed to avoid sparse cells in inferential analyses. A total household income variable was constructed by summing applicant and co-applicant income; however, this variable was excluded from regression modeling due to strong collinearity with its component variables.

Overall, approximately 69% of loan applications were approved, indicating moderate class imbalance. The dataset provides a realistic setting for examining factors associated with loan approval and for developing a binary classification model.

3 Exploratory Data Analysis

3.1 Loan Approval Distribution



(a) Loan approval distribution

(b) Approval by credit history

(c) Approval by property area

(d) Distribution of applicant income

Figure 1: Exploratory data analysis of loan approval outcomes and key predictors.

Figure 1 summarizes key patterns in the loan approval data. Approximately 69% of applications were approved, indicating a moderate class imbalance (Figure 1a). Approval rates vary sharply by credit history, with applicants possessing a positive credit record approved at substantially higher rates (Figure 1b). Property area also shows meaningful variation, with higher approval rates in semi-urban and urban locations relative to rural areas (Figure 1c). Other categorical variables, including gender, education level, self-employment status, and number of dependents, display relatively similar approval proportions across categories (not shown).

3.2 Continuous Predictors

Applicant income is strongly right-skewed, while co-applicant income is zero for many observations (Figure 1d). Loan amounts are more symmetrically distributed with a small number of large outliers, and loan terms show little variability, clustering at 360 months. Boxplots comparing approved and non-approved applications show substantial overlap in income and loan amount, with similar medians and spreads across groups. These patterns suggest that continuous financial variables alone are unlikely to strongly differentiate loan approval outcomes.

4 Preliminary Association Analysis

4.1 Categorical Predictors

Table 1: Chi-square tests of association between categorical predictors and loan approval.

Predictor	χ^2	df	p-value
Credit History	134.5	1	< 0.0001
Property Area	12.4	2	0.0022
Marital Status	6.1	1	0.0139

4.2 Continuous Predictors

Table 2: Association tests for continuous predictors by loan approval status.

Variable	t-test p-value	Wilcoxon p-value
Applicant Income	0.346	0.424
Co-applicant Income	0.284	0.242
Total Income	0.172	0.341
Loan Amount	0.116	0.215
Loan Amount Term	0.865	0.425

Formal statistical tests were conducted to assess associations between loan approval status and both categorical and continuous predictors. For categorical variables, chi-square tests indicate that credit history is strongly associated with loan approval ($p < 0.0001$), with applicants possessing a positive credit record far more likely to be approved. Property area and marital status also show statistically significant associations with loan approval (p = 0.002 and p = 0.014, respectively), while gender, education level, self-employment status, and number of dependents do not exhibit significant relationships.

For continuous predictors, two-sample t-tests and Wilcoxon rank-sum tests were used to compare approved and non-approved applications. None of the continuous variables, including applicant income, co-applicant income, total income, loan amount, or loan term, show statistically significant differences between groups at the 0.05 level. Results are consistent across parametric and nonparametric tests, suggesting that departures from normality do not materially affect the conclusions.

Overall, the association analysis highlights credit history as the dominant factor related to loan approval, with secondary contributions from property area and marital status. In contrast, continuous financial variables show limited discriminatory power when considered individually. These findings motivate the use of multivariable logistic regression to evaluate the joint effects of predictors and to quantify their relative importance in determining loan approval outcomes.

5 Logistic Regression Modeling

Logistic regression was used to model the probability of loan approval as a function of applicant characteristics. This approach is appropriate given the binary response variable and allows for interpretable assessment of multiple predictors simultaneously.

5.1 Collinearity Assessment

Collinearity among continuous predictors was evaluated using correlation analysis, variance inflation factors (VIFs), and condition indices. Total household income was found to be highly collinear with applicant income and co-applicant income, as it is a linear combination of these variables. This was reflected in large VIF values and extreme condition indices when total income was included, leading to unstable parameter estimates. Consequently, total income was excluded from subsequent models to mitigate multicollinearity.

5.2 Model Selection

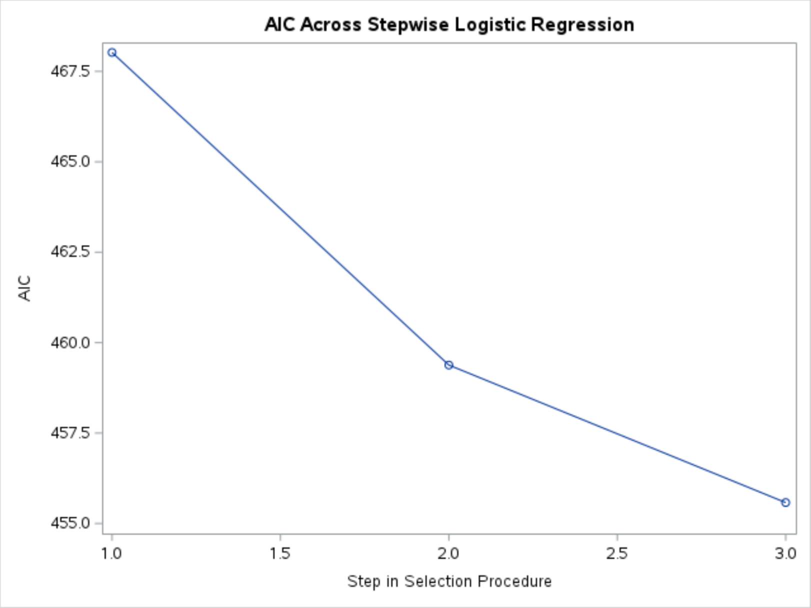


Figure 2: Change in AIC across the stepwise logistic regression procedure. The largest reduction occurs when credit history enters the model, with subsequent steps providing smaller improvements.

Table 3: Summary of stepwise variable selection for the logistic regression model.

Step	Predictor Entered	p-value
1	Credit History	< 0.0001
2	Property Area	0.0011
3	Marital Status	0.0201

Model selection was performed using stepwise logistic regression implemented in `PROC LOGISTIC`. Candidate predictors included demographic, socio-economic, and financial variables identified in earlier analyses. Figure 2 displays the change in AIC across the selection procedure, with the largest reduction occurring when credit history entered the model. Subsequent additions yielded smaller improvements, indicating diminishing returns.

The final model retained three categorical predictors: credit history, property area, and marital status (Table 3). Continuous predictors did not enter the model, consistent with their lack of significant univariate associations. Overall, the selected model achieves a favorable balance between parsimony and predictive performance.

5.3 Final Model Interpretation

Table 4: Odds ratios and 95% confidence intervals for the final logistic regression model.

Predictor	Odds Ratio	95% CI	p-value
Credit History (Good vs Poor)	43.84	(19.08, 100.73)	< 0.001
Semi-Urban vs Rural	2.77	(1.57, 4.89)	< 0.001
Married vs Not Married	1.72	(1.09, 2.74)	0.019

Table 4 presents odds ratios and 95% confidence intervals for the final logistic regression model. Credit history is by far the strongest predictor of loan approval. Applicants with a positive credit history have odds of approval approximately 44 times higher than those without a credit history, holding other factors constant. This result underscores the dominant role of creditworthiness in lending decisions.

Property area also has a meaningful effect on loan approval. Applicants residing in semi-urban areas have approximately 2.8 times higher odds of approval compared to those in rural areas, while no statistically significant difference is observed between urban and rural locations. Marital status shows a smaller but statistically significant effect, with married applicants having roughly 1.7 times higher odds of approval than unmarried applicants.

Overall, the final model suggests that loan approval decisions in this dataset are driven primarily by credit history, with secondary contributions from location and marital status. Financial variables such as income and loan amount do not appear to influence approval once credit history is taken into account.

6 Model Performance and Classification

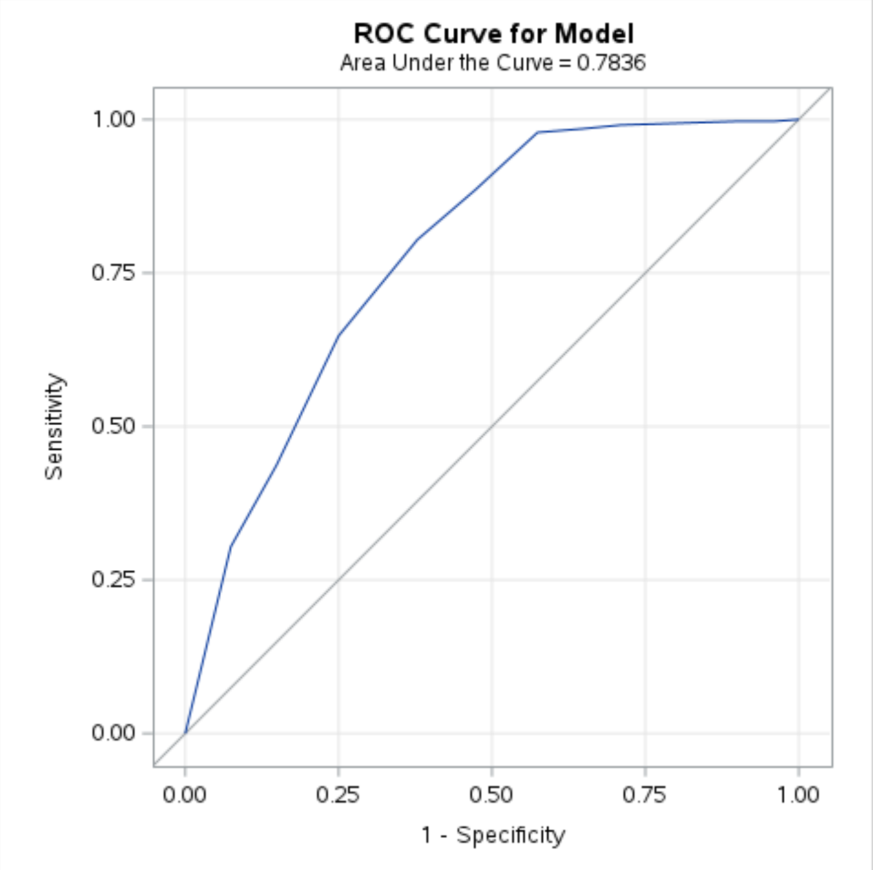


Figure 3: ROC curve for the final logistic regression model (AUC = 0.7836).

The predictive performance of the final logistic regression model was evaluated using both discrimination and classification metrics. Discrimination was assessed using the receiver operating characteristic (ROC) curve, which yielded an area under the curve (AUC) of approximately 0.78. This indicates good ability to distinguish between approved and non-approved loan applications and represents a substantial improvement over random classification.

Classification performance was examined by converting predicted probabilities into approval decisions using standard probability cutoffs. At a cutoff of 0.5, the model achieved an overall accuracy of $\approx 81\%$ and very high sensitivity ($\approx 98\%$), indicating that most approved loans are correctly identified. However, specificity was lower ($\approx 43\%$), reflecting a tendency of the model to favor approval in borderline cases. This behavior is consistent with the class imbalance observed in the response variable.

Overall, the final model demonstrates strong performance driven primarily by credit history, with classification characteristics that can be adjusted with probability cutoff to reflect different lending risk preferences.

7 Conclusions and Limitations

Conclusions

- Credit history is the dominant factor associated with loan approval, with applicants possessing a positive credit record substantially more likely to be approved.
- Property area and marital status contribute additional explanatory power, with higher approval odds observed for semi-urban applicants and married individuals.
- Continuous financial variables, including applicant income, co-applicant income, loan amount, and loan term, show limited discriminatory ability when considered individually.
- The final logistic regression model demonstrates good predictive performance, achieving an AUC of approximately 0.78.
- Classification results show very high sensitivity and lower specificity, indicating a tendency to favor approval decisions that can be adjusted through probability cutoff selection.

Limitations

- The analysis is based on a single dataset with a moderate sample size, which may limit generalizability to other lending populations.
- Class imbalance in the response variable influences classification behavior and contributes to lower specificity.
- Important predictors such as detailed credit scores, employment stability, and debt-to-income ratios were not available in the dataset.
- Model performance was evaluated on the same data used for fitting, and results may be optimistic without external validation.

References

[1] Kaggle. (2019). Home loan approval prediction dataset. https://www.kaggle.com/code/unmoved/classify-home-loan-approval

[2] Kim J. H. (2019). Multicollinearity and misleading statistical results. Korean journal of anesthesiology, 72(6), 558–569. https://doi.org/10.4097/kja.19087

[3] SAS Institute Inc. (2020). Collinearity in regression: The COLLIN option in PROC REG. SAS IML Blog. https://blogs.sas.com/content/iml/2020/01/23/collinearity-regression-collin-option.html

A Model Diagnostics

Table 5: Top 10 observations by Cook's distance for the final logistic regression model.

Obs	Status	$\hat{p}$	Pearson	Deviance	Leverage	Cook's D
1	Yes	0.045	4.603	2.490	0.0098	0.212
2	Yes	0.078	3.435	2.258	0.0149	0.182
3	Yes	0.111	2.837	2.099	0.0204	0.171
4	Yes	0.111	2.837	2.099	0.0204	0.171
5	Yes	0.095	3.084	2.169	0.0173	0.170
6	Yes	0.182	2.117	1.845	0.0277	0.131
7	Yes	0.182	2.117	1.845	0.0277	0.131
8	No	0.894	-2.899	-2.117	0.0058	0.049
9	No	0.894	-2.899	-2.117	0.0058	0.049
10	No	0.894	-2.899	-2.117	0.0058	0.049

Table 5 lists the observations with the largest Cook's distance values. The most influential cases correspond to applications whose observed outcomes conflict with the model's fitted probabilities (e.g., approved cases with low predicted approval probability and denied cases with high predicted approval probability). Overall, influence values are limited in magnitude and do not suggest that model results are driven by a single extreme observation.

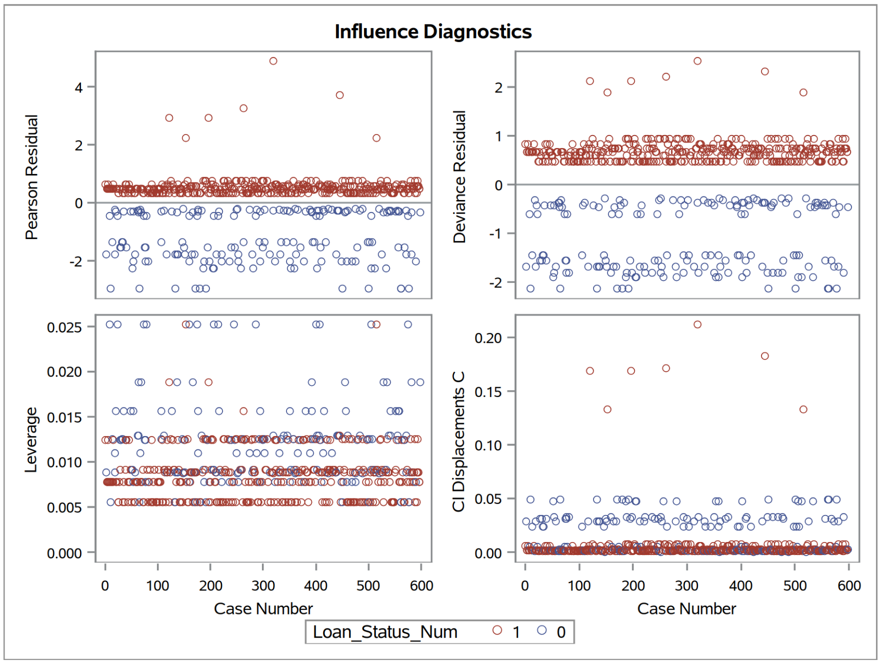


Figure 4: Influence diagnostics.

Figure A1 displays influence diagnostics for the final logistic regression model. Pearson and deviance residuals are generally centered around zero, with a small number of observations exhibiting larger residuals corresponding to outcomes that conflict with the model's fitted probabilities. Leverage values are uniformly low, indicating no observations with extreme predictor configurations. Cook's distance values are modest in magnitude and do not suggest that model estimates are driven by a small number of influential cases. Overall, the diagnostics support the stability and adequacy of the fitted model.

Overall goodness-of-fit was additionally assessed using `PROC GENMOD`, which produced fit statistics consistent with the `PROC LOGISTIC` results and provided no evidence of lack of fit.