Analysis of Factors Predictive of Loan Default Among Individual Borrowers

Yiyang Chen

2024.4.7

Abstract

This study investigates the predictive factors of loan default among individual borrowers using a dataset of 100,000 applicant details. By employing logistic regression analysis and various model diagnostics and selection techniques, we identify key predictors influencing the likelihood of loan default. The findings offer insights for financial institutions to enhance their risk management strategies.

Introduction

In the face of dynamic financial environments, precise loan default prediction remains pivotal. This study builds on the predictive analytics groundwork laid by Tham et al. (2023), Madaan et al. (2021), and Lai (2020), integrating logistic regression with machine learning techniques to examine borrower characteristics' influence on default risk. Our findings add to the current literature by demonstrating the importance of socioeconomic factors in determining loan defaults, while also matching complicated statistical approaches with real-world applicability.

2. Methods

2.1 Choice of Methods

To model the binary outcome of loan default risk, a logistic regression framework was utilized. Logistic regression is a subset of generalized linear models (GLM) appropriate for dichotomous response variables. The model assumes the log odds of the probability of default, which is mathematically represented as:

$$\log\left(\tfrac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

where:

- π represents the probability of loan default,
- Xi represents the predictor variables,
- βi represents the coefficients.

Given the nature of the dataset, variable selection was pivotal to refine the model and mitigate potential multicollinearity.

2.2 Variable Selection: AIC, BIC, and LASSO

Three variable selection methods were employed to determine the most significant predictors:

- **AIC (Akaike Information Criterion)**: A method favoring model parsimony, which helps prevent overfitting by penalizing excessive complexity. A stepwise algorithm was utilized to minimize the AIC value, adding or removing variables iteratively.
- **BIC (Bayesian Information Criterion)**: Like AIC but with a stronger penalty term for the number of parameters in the model, BIC favors simpler models, particularly in larger datasets.
- **LASSO (Least Absolute Shrinkage and Selection Operator)**: A regularization technique that performs variable selection and regularization to enhance the prediction accuracy and interpretability of the model. It shrinks some coefficients toward zero, effectively performing variable selection.

2.3 Model Violations/Diagnostics

Model Assumption:

- Linearity of the relationship between independent variables and the log odds of the outcome.
- Independence of observations.
- Absence of multicollinearity.
- Adequate sample size.
- No influential outliers.
- Correct specification of the model

Model assumptions for the logistic regression were scrutinized using a suite of diagnostic tools. This included examining the normal Q-Q plots to detect significant deviations from normality, indicative of outliers or influential observations. To evaluate the model's predictive capability and how well the probabilities aligned with the observed outcomes, calibration plots and ROC curves were utilized.

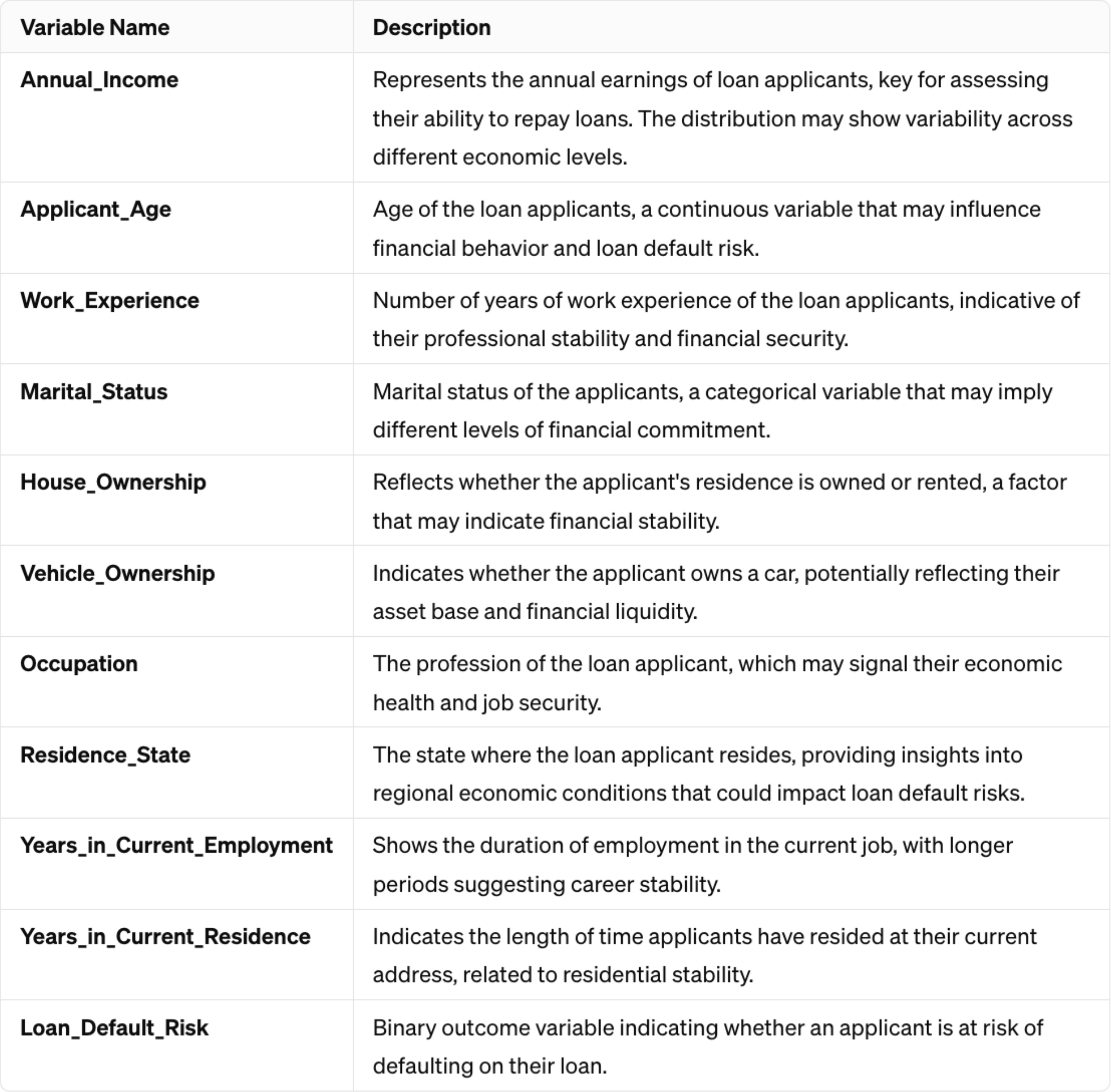
To further assess the influence of individual data points, DFBETAs were calculated for each predictor. DFBETAs are a measure of how much each point's presence or absence influences the model's estimated coefficients. By examining the DFBETAs plots for each predictor, we could determine if any data points were unduly influencing the model's predictions. Specifically, points with large absolute DFBETA values suggest that those observations have a substantial impact on the parameter estimates and may be considered outliers.

Deviance residuals were also plotted against predictors to check for any apparent patterns that might suggest model misspecification, non-linearity, or interaction effects that were not included in the model. Given the diagnostic plots' suggestions, transformations, including logarithmic and polynomial, were applied to certain predictors to enhance linearity in the logit. The models were then reassessed, with the AUC serving as the performance metric to ensure the adjustments improved model accuracy.

3. Data Description and Summaries

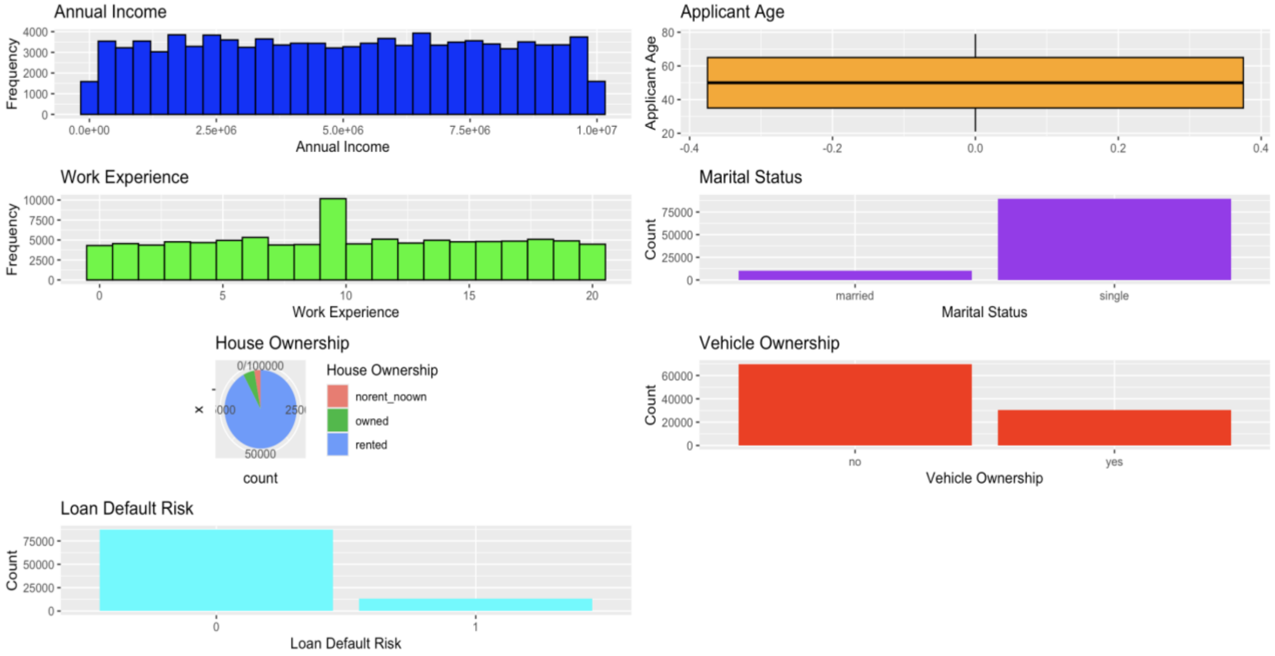
The dataset provides a comprehensive overview of loan applicants within India, offering a rich array of variables relevant for evaluating loan default risk.

Table 1: Summary of Variables



Variable Name	Description
Annual_Income	Represents the annual earnings of loan applicants, key for assessing their ability to repay loans. The distribution may show variability across different economic levels.
Applicant_Age	Age of the loan applicants, a continuous variable that may influence financial behavior and loan default risk.
Work_Experience	Number of years of work experience of the loan applicants, indicative of their professional stability and financial security.
Marital_Status	Marital status of the applicants, a categorical variable that may imply different levels of financial commitment.
House_Ownership	Reflects whether the applicant's residence is owned or rented, a factor that may indicate financial stability.
Vehicle_Ownership	Indicates whether the applicant owns a car, potentially reflecting their asset base and financial liquidity.
Occupation	The profession of the loan applicant, which may signal their economic health and job security.
Residence_State	The state where the loan applicant resides, providing insights into regional economic conditions that could impact loan default risks.
Years_in_Current_Employment	Shows the duration of employment in the current job, with longer periods suggesting career stability.
Years_in_Current_Residence	Indicates the length of time applicants have resided at their current address, related to residential stability.
Loan_Default_Risk	Binary outcome variable indicating whether an applicant is at risk of defaulting on their loan.

Table 2: Summary of Key Variables



The distribution of annual income reveals a predominance of lower to middle-income levels among the applicants, with work experience showing a moderate to high range across the dataset, indicative of career maturity. Applicant age is broadly distributed, with a concentration in the middle-age category, and marital status trends towards a higher proportion of single individuals. Home and vehicle ownership patterns suggest a mix of both assets and non-assets holders, providing a glimpse into the applicants' economic standings.

4. Results

4.1 Final Model Development

The process to find the final logistic regression model started with the inclusion of all potential predictors in our generalized linear model.
Stepwise Selection: AIC drops Annual-Income and Years_in_Current_Residence and BIC drops Occupation.
LASSO Regularization: The LASSO method further sculpted the predictor space by applying a penalty to the coefficients, effectively shrinking less relevant predictor impacts to zero. OccupationConsultant, OccupationStatistician, Residence_StateMizoram are dropped by LASSO
Model Comparisons: Each candidate model, derived from the AIC, BIC, and LASSO techniques, was subjected to a performance comparison using the Area Under the Curve (AUC) from the Receiver Operating Characteristic (ROC) analysis. The AIC-based model showcased the largest AUC(0.61)(Figure 3) compared with 0.59 of BIC-based model(figure 4) and 0.60 of LASSO based model.
Diagnostic Checks: Ensuring model reliability, diagnostic assessments were carried out for AIC-based model. DFBETAs for each predictor provided insights into individual data points'

influence on coefficient estimates(Figure1). Deviance residuals were also examined for any systematic deviations that could hint at model misspecification.(Table 4)

Predictor Transformations: Variables exhibiting influential data points or non-linear patterns in the diagnostic plots underwent transformations. Both logarithmic and polynomial transformations were explored to enhance model fit and AUC. It turned out that polynomial transformation has better performance.

Outlier Management: Outliers, as indicated by the diagnostic checks, were identified and removed to refine the model's predictive acumen.

Final Model Validation: The culmination of these steps led to the final model. Calibration plots and ROC curves validated the final model's performance. A calibration plot depicted the agreement between predicted probabilities and observed outcomes, while the ROC curve substantiated the model's classification effectiveness.

Figure 1: DFBETAs for Significant Predictors

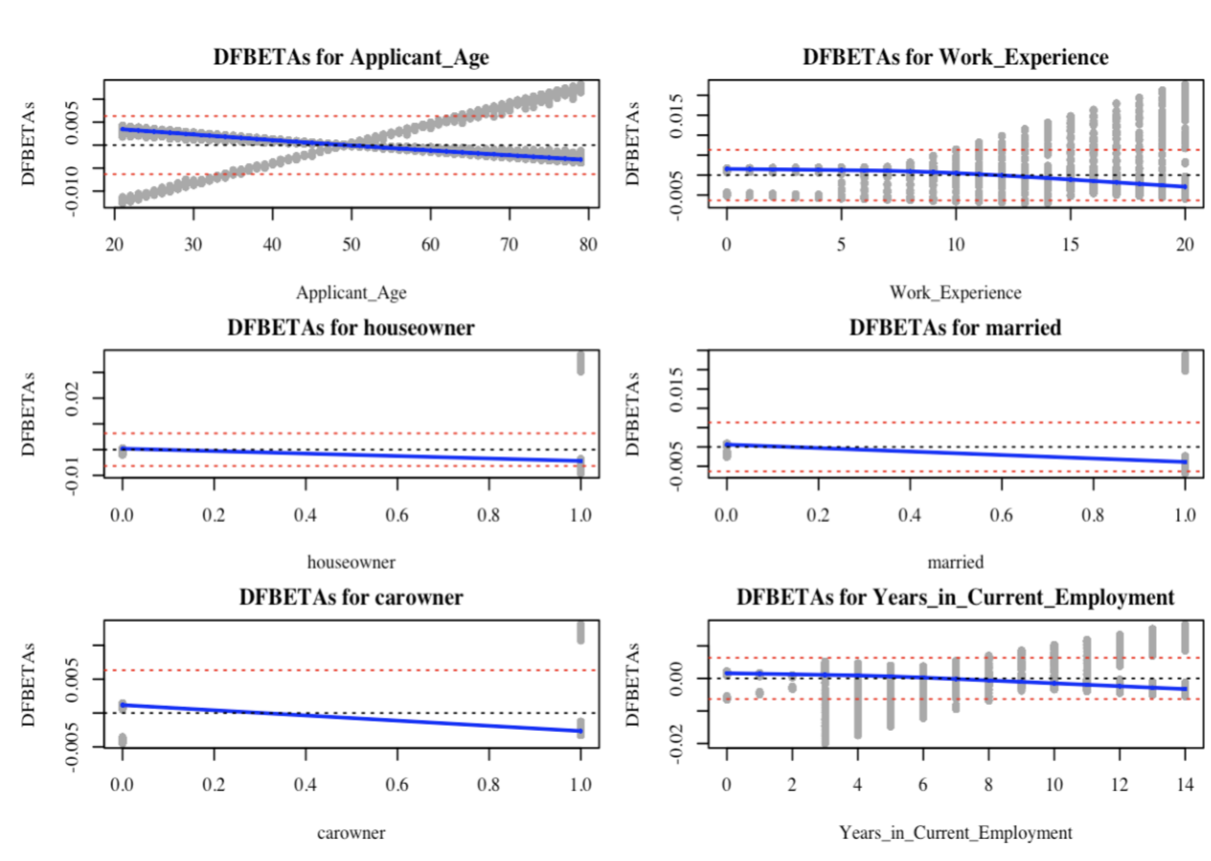
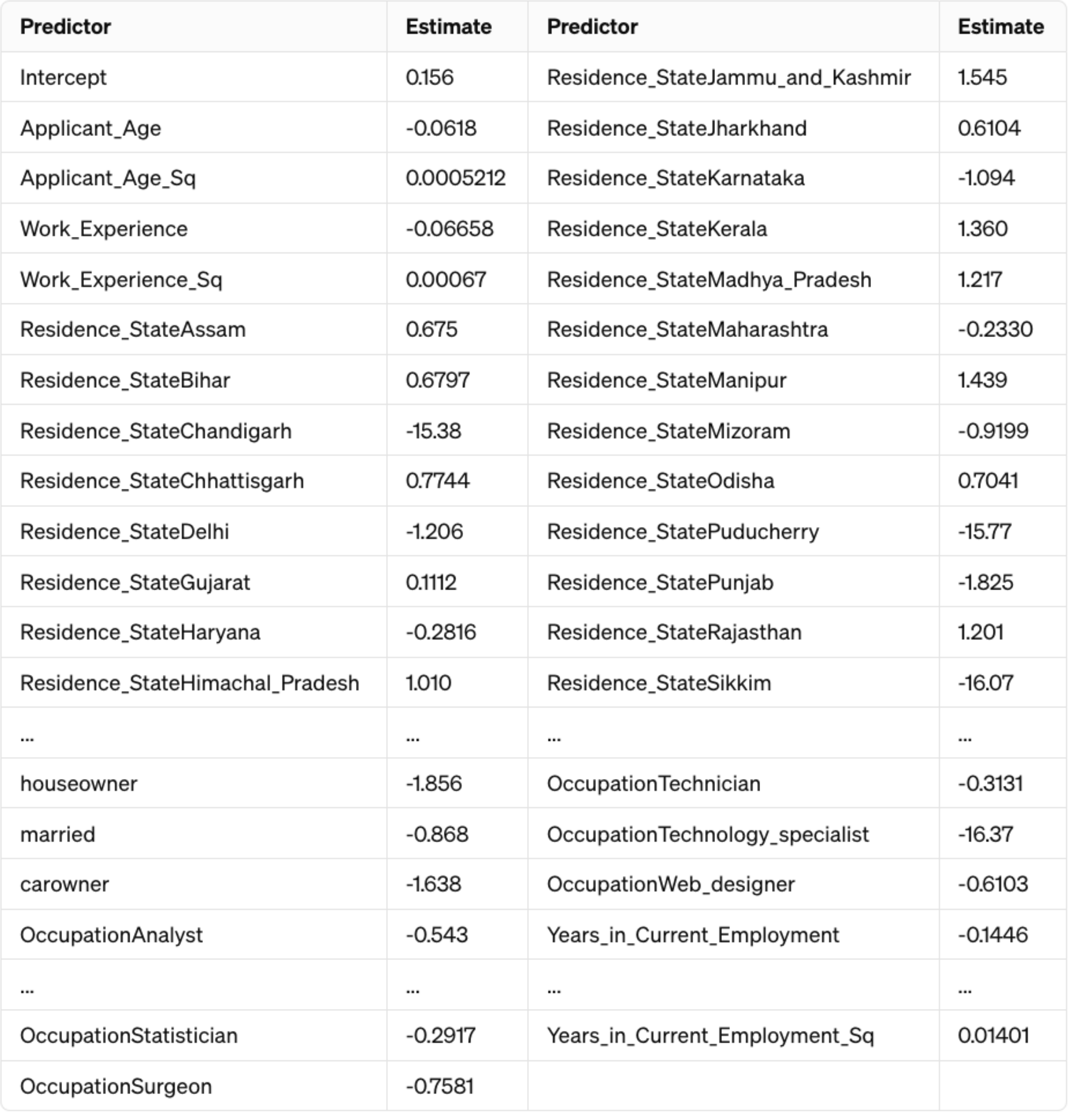


Table 3: Coefficients of GLM



Predictor	Estimate	Predictor	Estimate
Intercept	0.156	Residence_StateJammu_and_Kashmir	1.545
Applicant_Age	-0.0618	Residence_StateJharkhand	0.6104
Applicant_Age_Sq	0.0005212	Residence_StateKarnataka	-1.094
Work_Experience	-0.06658	Residence_StateKerala	1.360
Work_Experience_Sq	0.00067	Residence_StateMadhya_Pradesh	1.217
Residence_StateAssam	0.675	Residence_StateMaharashtra	-0.2330
Residence_StateBihar	0.6797	Residence_StateManipur	1.439
Residence_StateChandigarh	-15.38	Residence_StateMizoram	-0.9199
Residence_StateChhattisgarh	0.7744	Residence_StateOdisha	0.7041
Residence_StateDelhi	-1.206	Residence_StatePuducherry	-15.77
Residence_StateGujarat	0.1112	Residence_StatePunjab	-1.825
Residence_StateHaryana	-0.2816	Residence_StateRajasthan	1.201
Residence_StateHimachal_Pradesh	1.010	Residence_StateSikkim	-16.07
...	...	...	...
houseowner	-1.856	OccupationTechnician	-0.3131
married	-0.868	OccupationTechnology_specialist	-16.37
carowner	-1.638	OccupationWeb_designer	-0.6103
OccupationAnalyst	-0.543	Years_in_Current_Employment	-0.1446
...	...	...	...
OccupationStatistician	-0.2917	Years_in_Current_Employment_Sq	0.01401
OccupationSurgeon	-0.7581		

Note: Table truncated for brevity.

4.2 Goodness of Final Model

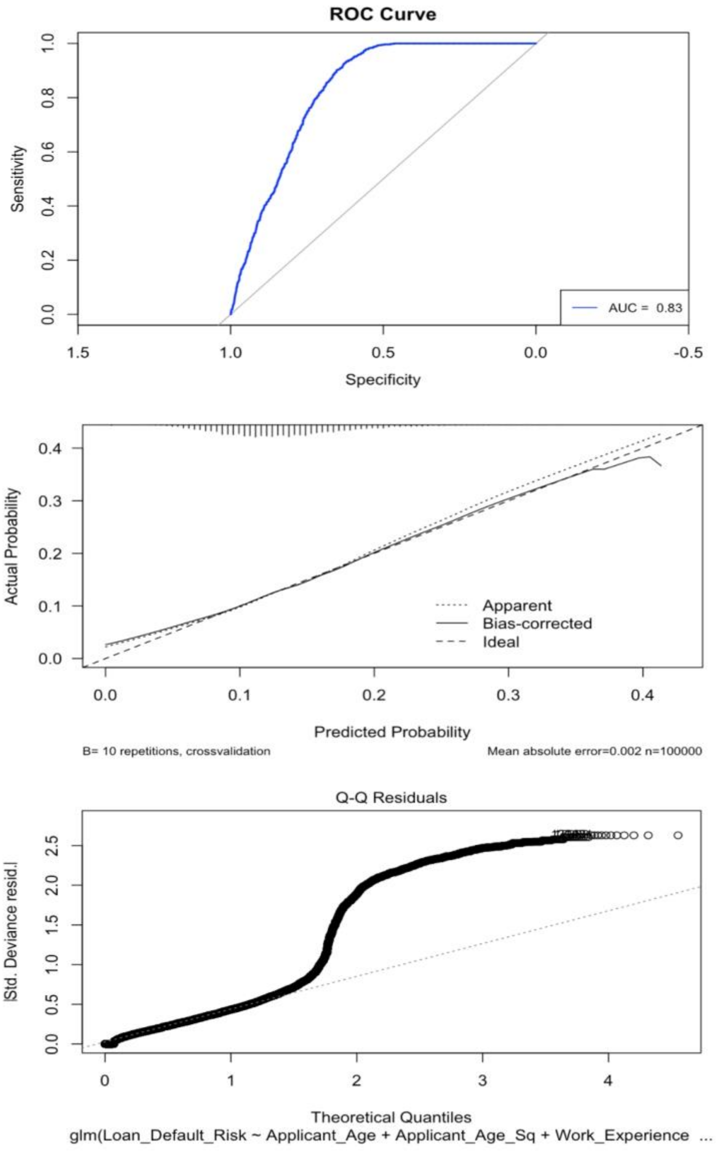
The model's goodness-of-fit was evaluated using several diagnostic plots. The Receiver Operating Characteristic (ROC) curve (Figure 2) exhibited an Area Under Curve (AUC) of 0.83, indicative of the model's strong capability to distinguish between events of loan default and non-default. This value suggests that the model has good discriminative power, although it is not perfect.

The calibration plot (Figure 2) demonstrates excellent agreement between the predicted probabilities and the actual outcomes. The alignment of the bias-corrected calibration line with the ideal line reveals that the model's predictions are accurate and well-calibrated, with a minimal mean absolute error, highlighting the model's precision.

The Q-Q plot (Figure 2) is used to assess the normality of the residuals, a critical assumption in many statistical models. For logistic regression, the expected distribution of residuals is not normal since the response is binary; hence, some deviation from the theoretical quantiles is expected.

Overall, while the model displays a strong predictive ability as shown by the AUC, the calibration plot assures us that the probability predictions are reliable.

Figure 2 goodness of fitting plots



5. Discussion

Our final model reveals the intricate dynamics of loan default risk, with the age factor exhibiting an inverted U-shaped relationship, suggesting initial high risk that decreases and later subtly rises with aging. These non-linear influences suggesting a life-cycle effect on default risk, initially high for young borrowers with limited credit history and then decreasing with financial maturity, before rising slightly as retirement approaches.

Occupational categories yield rich insights: for instance, the heightened default risk in volatile professions like Petroleum Engineers reflects the critical impact of job security on financial reliability. Analysts, on the other hand, are more likely to be proficient in finance, demonstrating the protective effect of financial knowledge. Marital status and asset ownership emerge as important stabilizers, lowering default risks and emphasizing the significance for economic anchoring.

5.1 Limitation

While the model's predictive strength is affirmed by an AUC of 0.83, the Q-Q plot points to outliers, signaling that the model may not fully capture the complexities of certain atypical borrowers or unaccounted-for economic phenomena. This limitation, while not undermining the model’s overall utility, suggests a scope for incorporating additional data, such as credit history or economic indicators, to enhance its predictive accuracy. However, Occupation variable, with its many categories, presents a limitation by complicating the model and risking overfitting, which can impact interpretability and model performance. We could consider grouping similar occupations. This strategy would consolidate the numerous categories into broader clusters, thereby reducing the model's complexity and enhancing its interpretability without significantly compromising the predictive power.

References

1. Tham, A. W., Kakamu, K., & Liu, S. (2023). Bayesian Statistics for Loan Default. Journal of Risk and Financial Management, 16(3), 203. https://doi.org/10.3390/jrfm16030203
2. Madaan, M., Kumar, A., Keshri, C., Jain, R., & Naqarath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012042). IOP Publishing.
3. Lai, L. (2020, August). Loan default prediction with machine learning techniques. In 2020 International Conference on Computer Communication and Network Security (CCNS) (pp. 5-9). IEEE.
4. Yaminh. (n.d.). Applicant Details for Loan Approve [Data set]. Kaggle. Retrieved [April 7, 2024], from https://www.kaggle.com/datasets/yaminh/applicant-details-for-loan-approve

Appendices

Table 4: Deviance residuals of predicators

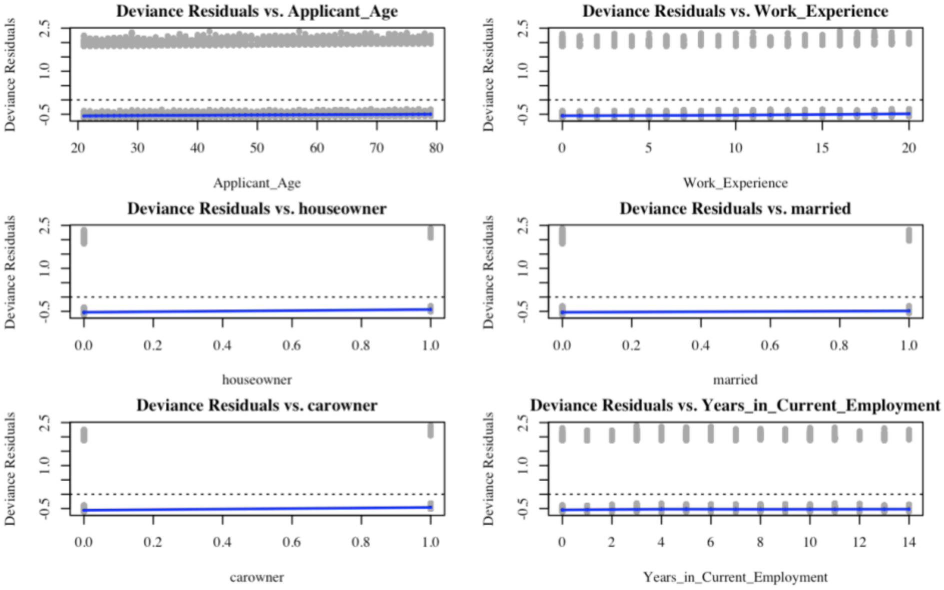


Figure 3: ROC Curve of initial AIC-based model

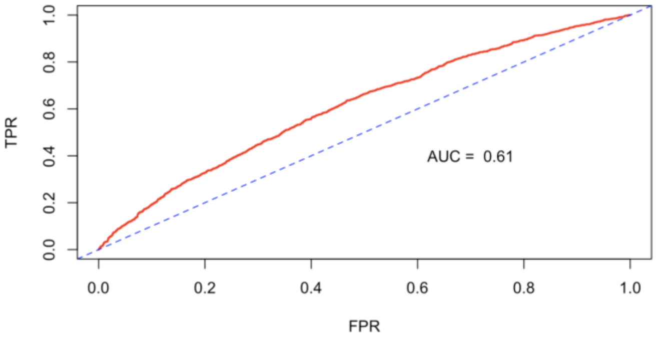
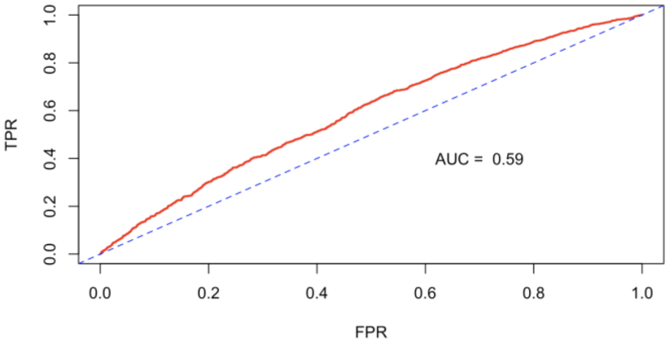


Figure 4: ROC Curve of BIC-based model


TPR
AUC = 0.59
FPR