# Auto Loan Credit Decisioning Model

A K M Rokonuzzaman Sonet

May 21, 2025

**Abstract**

This report presents the development and evaluation of a predictive model to support credit approval decisions for auto loan applications. Using a dataset of historical applications and outcomes, we trained and compared multiple models including Logistic Regression, Decision Trees, and Random Forests.

We performed extensive exploratory data analysis (EDA), handled missing and imbalanced data, and engineered a fair and interpretable machine learning pipeline. Logistic Regression with class-weight balancing was selected for its strong AUC-ROC, high recall, and explainability.

The model was further assessed for fairness and interpretability using demographic analysis and LIME explanations. Results showed no significant evidence of gender or racial bias in the model's predicted approvals, supporting its readiness for responsible deployment.

## 1 Introduction

In the modern consumer credit market, automated decision-making models play a crucial role in determining the approval of credit products, such as auto loans. These models must not only accurately assess the risk of default but also operate fairly across demographic groups to ensure compliance with regulatory guidelines and support equitable access to credit.

This project focuses on developing a credit decisioning model for auto loan applications using historical applicant data. Each applicant record includes demographic information, credit behavior indicators, trade activity, and loan application details. The primary goal is to identify applicants with good credit quality—those who are unlikely to default—so that the institution can make informed and responsible lending decisions.

The target variable in this task is `bad_flag`, which serves as a proxy for recent loan performance:

- `bad_flag`= 1: Good credit quality (never delinquent or only minor delinquencies)

- `bad_flag` = 0: Bad credit quality (serious delinquencies, charge-offs, or defaults)

An additional variable, `aprv_flag`, indicates whether the application was approved (1) or not (0). While this variable is not used for model training, it helps assess fairness and alignment between historical approvals and modeled predictions.

This report begins with an exploratory analysis of the data to understand patterns, distributions, and potential issues in the dataset. Modeling and evaluation will follow in subsequent sections.

# 2 Exploratory Data Analysis (EDA)

The primary goal of this section is to understand the data structure, assess data quality, investigate relationships between predictors and the target variable (`bad_flag`), and prepare the data for predictive modeling.

## 2.1 Dataset Description

- Number of training records $\approx 21000$
- Number of testing records $\approx 5400$
- Number of features: 41
- Target variable: `bad_flag` (0 = bad credit, 1 = good credit)
- Secondary variable: `aprv_flag` (loan approval status)

Features include both numerical and categorical variables such as FICO score, loan-to-value ratio, credit utilization metrics, delinquency indicators, and demographic information.

## 2.2 Missing Value Analysis

A missing value analysis of the data sets revealed that 33 of 42 predictor variables exhibit missing values, with 9 features showing a missing rate exceeding 50%. These columns of high absence were consistently found in both the training and the test sets (see Figure 1).

To evaluate their modeling relevance, we examined the correlation between each of these 9 features and the target variable `bad_flag`. We observed that the features with the highest missingness generally exhibited weak correlations, both positive and negative, with the target variable (see Table 1). None of them demonstrated sufficient predictive signal to justify retention, so we removed them from the modeling pipeline.

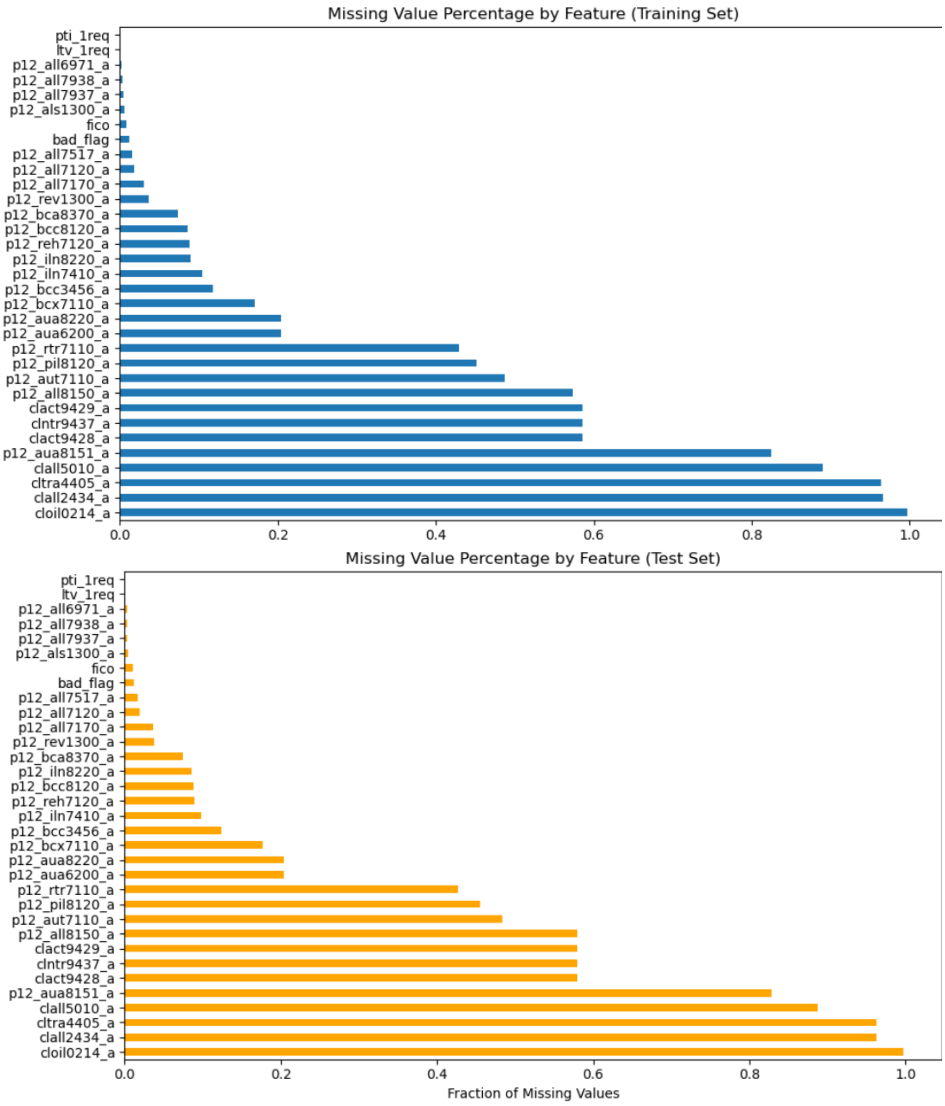



Figure 1: Missing value percentages by feature for train and test data

After removing the 9 features that exhibited both high missingness and weak correlation with the target variable, we turned our attention to the remaining variables that contain a lower proportion of missing values. For these features, we adopted an imputation strategy based on the type and distribution of the data. To guide our imputation strategy, we first examined the number of unique values for each feature in the training and test datasets (see Table 2). This helped distinguish between continuous variables and those that are likely categorical or frequency-based. For instance, features such as `p12_all6250_a`, `p12_aua6200_a`, `p12_all2427_a`, `p12_alm6200_a`, and `p12_all6971_a` contain only a handful of distinct values. These are indicative of status codes or categorical counts rather than continuous measures. As an example, the feature `p12_all6250_a` represents 'worst ever status on a trade in the first 12 months' and includes values such as 1, 30, 60, 90, 120 and 400—making it appropriate for mode-based imputation. In

contrast, features like `fico`, `amtfinanced_1req`, and `ltv_1req` show a much broader range of values and are considered continuous. These were imputed using the median, a strategy chosen to reduce the impact of skewed distributions and outliers. This approach ensures that the imputation process respects the semantic nature of each variable while minimizing distortion in downstream modeling. The imputed values are used consistently across both training and test sets to maintain alignment.

| Feature | Train Missing % | Test Missing % | Correlation with bad_flag |
|---|---|---|---|
| cloil0214_a | 99.639 | 99.667 | -0.164 |
| clall2434_a | 96.672 | 96.296 | -0.008 |
| cltra4405_a | 96.357 | 96.241 | -0.002 |
| clall5010_a | 88.938 | 88.759 | 0.014 |
| p12_aua8151_a | 82.449 | 82.796 | -0.069 |
| clntr9437_a | 58.567 | 57.926 | 0.117 |
| clact9429_a | 58.567 | 57.926 | 0.109 |
| clact9428_a | 58.567 | 57.926 | 0.114 |
| p12_all8150_a | 57.387 | 57.907 | -0.089 |
| p12_aut7110_a | 48.764 | 48.370 | 0.057 |
| p12_pil8120_a | 45.122 | 45.519 | -0.058 |
| p12_rtr7110_a | 42.928 | 42.667 | 0.076 |
| p12_aua6200_a | 20.420 | 20.370 | 0.079 |
| p12_aua8220_a | 20.420 | 20.370 | -0.087 |
| p12_bcx7110_a | 17.074 | 17.741 | 0.103 |
| p12_bcc3456_a | 11.825 | 12.426 | -0.102 |
| p12_iln7410_a | 10.377 | 9.778 | 0.066 |
| p12_iln8220_a | 8.933 | 8.519 | -0.062 |
| p12_reh7120_a | 8.849 | 8.963 | 0.071 |
| p12_bcc8120_a | 8.581 | 8.759 | -0.038 |
| p12_bca8370_a | 7.336 | 7.444 | -0.098 |
| p12_rev1300_a | 3.684 | 3.722 | -0.096 |
| p12_all7170_a | 3.069 | 3.593 | 0.083 |
| p12_all7120_a | 1.759 | 1.944 | 0.049 |
| p12_all7517_a | 1.592 | 1.722 | 0.099 |
| bad_flag | 1.194 | 1.185 | 1.000 |
| fico | 0.810 | 1.056 | -0.197 |
| p12_als1300_a | 0.532 | 0.444 | -0.097 |
| p12_all7937_a | 0.389 | 0.352 | -0.139 |
| p12_all7938_a | 0.315 | 0.278 | -0.142 |
| p12_all6971_a | 0.204 | 0.278 | 0.085 |
| ltv_1req | 0.023 | 0.037 | 0.082 |
| pti_1req | 0.014 | 0.037 | 0.038 |

Table 1: Missing Value Percentages and Correlations with Target Variable

### 2.3 Target Variable Distribution

The target variable `bad_flag` is a binary indicator representing the credit quality of an auto loan applicant, where 1 and 0 indicate **Good Credit Quality** and **Poor or Bad Credit Quality** respectively. An analysis of its distribution reveals a significant class imbalance (Figure 2) , with the majority of records labeled as poor credit quality.

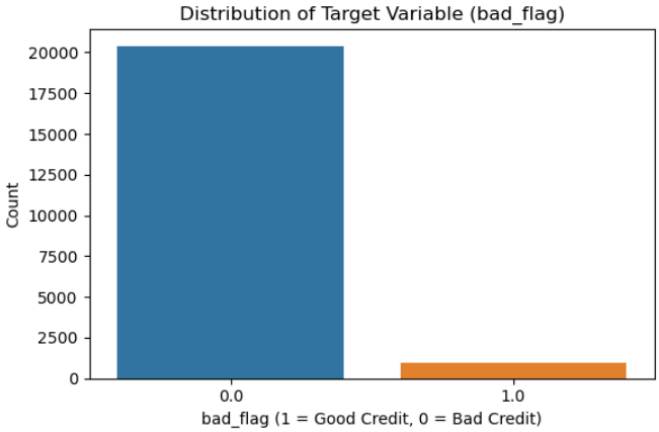



Figure 2: Distribution of Target Variable (`bad_flag`)

Specifically, over 95% of applicants in the dataset are classified as Poor Credit Quality (`bad_flag` = 0), while fewer than 5% fall into the Good Credit Quality (`bad_flag` = 1) category. Also, there are approximately 1.2% missing values in the target variable column.

To appropriately address the modeling task given the substantial class imbalance in the target variable, we adopt a structured three-stage approach.

- First, during preprocessing, we handle missing data with care. In particular, we exclude rows with missing values in the target variable (`bad_flag`) rather than imputing them, as any synthetic filling could introduce bias and degrade model validity.
- Second, in the modeling stage, we explicitly account for the imbalance between the majority (bad credit) and minority (good credit) classes. We experiment with two complementary strategies:($i$) training on the imbalanced data using algorithms with built-in support for class weighting, and ($ii$) applying resampling techniques—either through undersampling the majority class or oversampling the minority class (e.g., SMOTE).
- Finally, we evaluate model performance using metrics appropriate for imbalanced classification. Rather than relying on overall accuracy—which can be misleading—we focus on metrics such as the Area Under the ROC Curve (AUC-ROC), $F_1$-score, precision, recall, and the full classification report. These provide a more comprehensive and balanced assessment of the model's ability to distinguish between good and bad credit applicants.

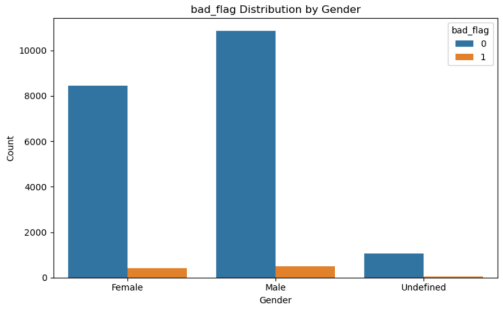
### 2.4 Categorical Feature Analysis

To understand how categorical features relate to credit quality, we analyzed the distribution of the target variable `bad_flag` across key categorical predictors, including Gender, Race, `collateral_dlrinput_newused_1req` (Vehicle Type), and `aprv_flag` (approval status). Prior to plotting, we removed rows with missing values in either `bad_flag` or `aprv_flag` to ensure consistency. Figure 3 illustrates how credit quality varies across demographic and loan-approval related attributes.

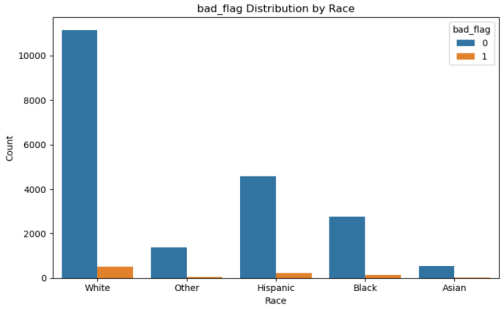
- Gender: Both male and female applicants show a dominant number of bad_flag = 0 (bad credit). However, the proportions of good vs bad credit are fairly consistent across genders, suggesting no strong gender bias in credit quality distribution.
- Race: White and Hispanic applicants make up the largest racial groups in the dataset. However, across all races, a majority of applicants are classified as bad credit (bad_flag = 0), indicating a skew in the dataset toward applicants with recent delinquency or severe credit events.
- Vehicle Type: A significantly higher number of applications are for used vehicles, and most of these applicants are labeled as bad credit. The same trend exists, albeit with lower volume, among new vehicle applicants. This may reflect a riskier profile among used car borrowers or just higher representation in the dataset.
- Approval Status: The approval decisions do not perfectly align with credit quality. Surprisingly, among the approved applicants (aprv_flag = 1), a significant number are labeled as bad credit (bad_flag = 0), indicating potential approval of high-risk applicants. Conversely, some good credit applicants (bad_flag = 1) have also been rejected, suggesting that approval decisions are influenced by factors beyond the recent loan performance summarized in bad_flag. These may include policy constraints, external scores, application errors, or manual underwriting criteria.

### 2.5 Frequency-Based Feature Analysis

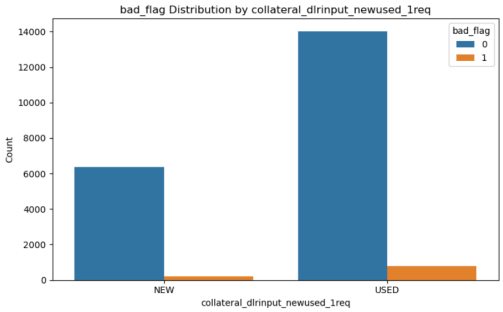
We analyzed five frequency-based credit indicators reflecting the worst delinquency status or trade activity in various time windows (see Figure 4). These features include: Worst ever status on a trade in the first 12 months, Worst ever status on auto loans, Number of trades with delinquencies in the past 12 months, Worst ever status on a trade in the past 12 months, Worst ever trade status on trade including non-medical collections and indeterminates. In each case, a value of 1 represents the least severe status (typically current or never delinquent), while higher values (e.g., 30, 60, 90, 400) correspond to more severe delinquency flags or charge-offs. In all features, the lowest credit severity values (like 0 or 1) are not necessarily dominated by good credit applicants. In fact, the proportion of good credit increases slightly in more severe categories like 60, 90, or even 120, though the total number of observations is much lower. This



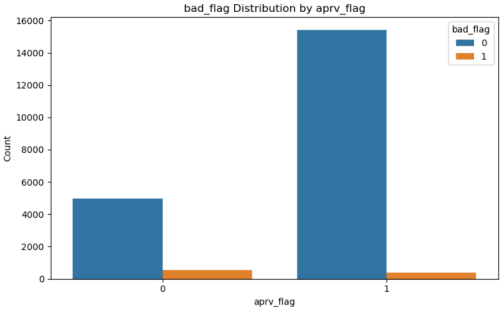

(a) By Gender




(b) By Race




(c) By Vehicle Type




(d) By Approval Flag

Figure 3: Distribution of the target variable across selected categorical features

confirms that while these features are predictive of credit quality, they are affected by class imbalance and must be interpreted in relative terms (e.g., percentages) — not raw counts. Percentages of good credit for each feature value are shown in Table 5, 3, 6, 7, and 4.

## 2.6 Continuous Feature Analysis

To further understand behavioral differences between applicants with good and bad credit profiles, we conducted a detailed analysis of several continuous variables. These features reflect core aspects of an individual's financial standing—such as credit scores, loan utilization, and historical borrowing patterns.

Before visualization, outliers were removed from each feature using the Interquartile Range (IQR) method. This ensures that extreme values do not distort the distribution plots and that comparisons between classes remain meaningful. Each feature was visualized using KDE histograms after removing outliers (see Figure 5). The key insights from this bi-variate analysis with the target variable are as follows:

- Creditworthiness Indicators:
  - FICO Credit Score: Good credit applicants (bad_flag = 1) are clearly skewed toward higher FICO values (e.g., above 700), while bad credit applicants peak

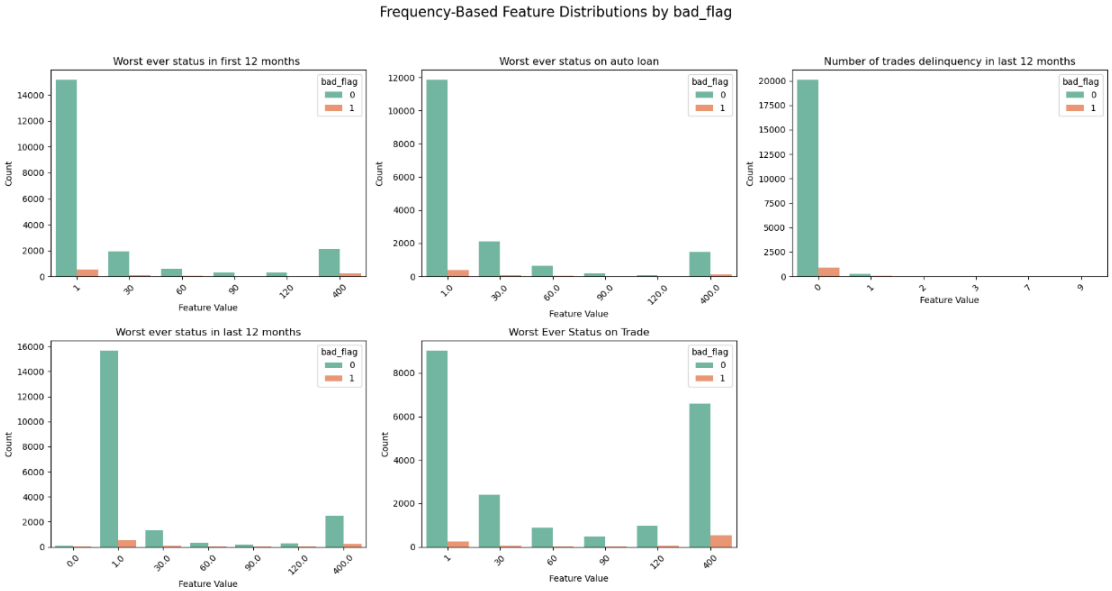



Figure 4: Frequency-based features and target variable

much lower. This confirms FICO score as a powerful predictor — higher FICO strongly correlates with better credit behavior.

- Payment-to-Income Ratio (PTI): Applicants with good credit typically have lower PTI, suggesting they manage debt obligations more effectively relative to their income. Higher PTI is more common among bad credit cases, consistent with financial stress and repayment risk.

- Loan-to-Value Ratio (LTV): Bad credit applicants show a noticeable concentration at higher LTV values, reflecting higher risk-taking or lack of equity. Lower LTVs are more prevalent among good credit applicants, aligning with more secure loan structures.

- Loan Amount and Utilization:

  - Amount Financed (Requested): Distributions for both groups overlap significantly. However, bad credit applicants show a slight peak at lower financing amounts, which may reflect lower eligibility or conservative offers from lenders.

  - Max Single Balance-to-Credit Amount Ratio: This ratio shows a skewed distribution where higher utilization is more frequent in bad credit applicants, suggesting overextension or underpayment.

  - Overall Balance to Credit Amount Ratio: Similar to the above, bad credit applicants tend to have higher balances relative to their credit limits, which may indicate chronic underpayment or financial stress.

- Account and Trade Activity:
  - Percentage of Trades Reported in Last 24 Months: Higher trade activity appears more balanced across both groups, but good credit applicants lean slightly toward higher recent reporting, suggesting more active or transparent borrowing behavior.
  - Total Number of Open Revolving Bankcard Trades: Bad credit applicants tend to cluster at higher values, possibly due to overextension or multiple low-limit cards. Good credit cases are more balanced.
  - Average Life Span of Bankcard: Good credit applicants have bankcards with longer average lifespan, supporting the idea that account age contributes to credit stability.
  - Months Since Oldest Auto Loan: Longer auto loan history is more typical among good credit applicants, indicating longevity and reliability in major loan categories.

The distributions reveal strong behavioral patterns: FICO score, PTI, and LTV are top indicators of credit quality, while utilization ratios and account age also show useful but subtler separation. These features are essential for modeling and risk assessment.

# 3 Data Preprocessing

Prior to model training, the dataset underwent a comprehensive preprocessing pipeline to ensure consistency, minimize bias, and handle missing values appropriately. The goal was to prepare the data for Logistic Regression while maintaining statistical integrity, especially under class imbalance.

## 3.1 Target Variable Handling

The target variable, `bad_flag`, was used to indicate recent credit performance (1 = good credit, 0 = bad credit). Any records with missing values in `bad_flag` were removed entirely, as imputing a target variable could introduce serious modeling bias. This ensured that only applicants with observed outcomes were used in training.

## 3.2 Feature Cleaning and Missing Value Imputation

Features with more than 50% missing data were dropped to avoid unreliable imputation and to reduce noise. The remaining missing values were handled based on variable type:

- **Continuous numerical features** (e.g., `fico`, `ltv_1req`, `pti_1req`) were imputed using the **median** value of the column.

- **Categorical and status-based features** (e.g., `Gender`, `Race`, delinquency flags like `p12_all6250_a`) were imputed using the **mode** (most frequent) value.

### 3.3 Encoding and Scaling

Categorical variables such as `Gender`, `Race`, and `collateral_dlrinput_newused_1req` were label encoded into numeric format using `LabelEncoder`, as required by scikit-learn's logistic regression implementation.

All continuous numerical features were standardized using `StandardScaler`, transforming them to have zero mean and unit variance. This step is essential for logistic regression, which is sensitive to feature scales.

### 3.4 Train/Validation Split

The processed dataset was split into 80% training and 20% validation subsets using stratified sampling based on `bad_flag`.

## 4 Logistic Regression Modeling and Evaluation

To establish a baseline for credit risk classification, we trained a Logistic Regression (LR) model using the training dataset. The target variable, `bad_flag`, indicates recent credit performance, with 1 representing good credit and 0 representing bad credit. Due to the highly imbalanced nature of the data (less than 5% good credit cases), we implemented two strategies:

1. Logistic Regression with `class_weight='balanced'` to address imbalance during optimization.
2. Logistic Regression trained on over-sampled data using SMOTE (Synthetic Minority Over-sampling Technique).

### 4.1 Baseline Model (Without Resampling)

The baseline model achieved a validation accuracy of 69%, and an AUC-ROC of 0.796. It successfully identified 75% of good credit cases (recall), although with low precision (10%). This tradeoff is typical in imbalanced classification where capturing the minority class is prioritized.

Table 8: Confusion Matrix – Logistic Regression (No Resampling)

| | **Predicted 0** | **Predicted 1** |
|---|---|---|
| **Actual 0** | 2813 | 1265 |
| **Actual 1** | 48 | 144 |

### 4.2 SMOTE Model (With Resampling)

The model trained on SMOTE-enhanced data achieved a slightly better accuracy of 72%, and recall for good credit remained high at 71%. However, precision for good credit predictions remained low (11%), and AUC-ROC decreased slightly to 0.779, suggesting less confident separation between classes.

Table 9: Confusion Matrix – Logistic Regression (With SMOTE)

| | **Predicted 0** | **Predicted 1** |
|---|---|---|
| **Actual 0** | 2925 | 1153 |
| **Actual 1** | 55 | 137 |

### 4.3 Top Predictive Features

Below are the 10 most influential features ranked by absolute coefficient value from the baseline model:

Table 10: Top 10 Predictive Features: Logistic Regression (With and Without SMOTE)

| **Feature** | **Without SMOTE (Coeff.)** | **With SMOTE (Coeff.)** |
|---|---|---|
| fico | -1.183 | -0.978 |
| Gender | -0.960 | -0.164 |
| ltv_1req | +0.568 | +0.491 |
| p12_als1300_a | -0.257 | -0.158 |
| p12_pil8120_a | -0.241 | -0.127 |
| p12_rev1300_a | -0.235 | -0.176 |
| p12_aut7110_a | +0.203 | +0.154 |
| p12_aua0300_a | -0.164 | — |
| p12_all7517_a | +0.139 | +0.202 |
| p12_bcc3456_a | -0.136 | — |
| collateral_dlrinput_newused_1req | — | +0.514 |

### 4.4 ROC Curve Comparison

Figure 6 shows the ROC curves for both models. The baseline model demonstrates slightly better ranking ability than the SMOTE-enhanced version, with higher AUC-ROC.

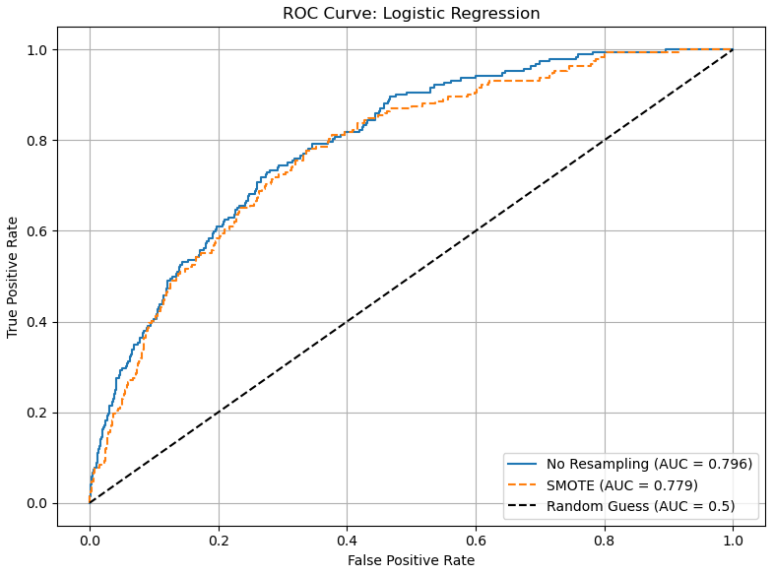



Figure 6: ROC Curve for Logistic Regression Models

# 5 Machine Learning Model Development and Evaluation

To extend our analysis beyond Logistic Regression, we developed two machine learning models: a **Decision Tree** and a **Random Forest**. These tree-based models can capture non-linear relationships and interactions between features, which may improve predictive performance. We evaluated both models under two training regimes: (i) using the imbalanced training dataset with class-weight balancing, and (ii) applying SMOTE (Synthetic Minority Oversampling Technique) to rebalance the training data.

## 5.1 Baseline Models (Without Resampling)

**Decision Tree:** Trained with `class_weight='balanced'`, the decision tree achieved a validation accuracy of 92%. It identified only 14% of actual good credit cases (recall), resulting in an $F_1$-score of 0.14 for class 1. The AUC-ROC was 0.550, indicating weak ranking ability.

Table 11: Confusion Matrix – Decision Tree (No Resampling)

| | **Predicted 0** | **Predicted 1** |
|---|---|---|
| **Actual 0** | 3914 | 164 |
| **Actual 1** | 165 | 27 |

**Random Forest:** The random forest model with 100 trees predicted all samples as class 0, failing to detect any good credit cases (recall = 0%). Despite a high accuracy

of 95%, the model was effectively non-functional for the minority class. However, it achieved a relatively high AUC-ROC of 0.785, indicating good probabilistic ranking.

Table 12: Confusion Matrix – Random Forest (No Resampling)

| | **Predicted 0** | **Predicted 1** |
|---|---|---|
| **Actual 0** | 4077 | 1 |
| **Actual 1** | 192 | 0 |

### 5.2 SMOTE-Based Models (Balanced Training)

**Decision Tree with SMOTE:** The SMOTE-enhanced decision tree showed a slight improvement in recall for class 1 (19%), but at the cost of precision (10%). The overall accuracy decreased to 88%, and AUC-ROC remained weak at 0.552. The model struggled to generalize from oversampled data.

Table 13: Confusion Matrix – Decision Tree (With SMOTE)

| | **Predicted 0** | **Predicted 1** |
|---|---|---|
| **Actual 0** | 3739 | 339 |
| **Actual 1** | 156 | 36 |

**Random Forest with SMOTE:** This model achieved the best class 1 precision (18%) among all models so far, with 16 out of 192 good credit cases correctly identified (recall = 8%). AUC-ROC was 0.768, comparable to the baseline random forest. While precision improved, recall remained low, indicating that additional tuning or threshold adjustment is needed.

Table 14: Confusion Matrix – Random Forest (With SMOTE)

| | **Predicted 0** | **Predicted 1** |
|---|---|---|
| **Actual 0** | 4006 | 72 |
| **Actual 1** | 176 | 16 |

## 5.3 ROC Curve Comparison

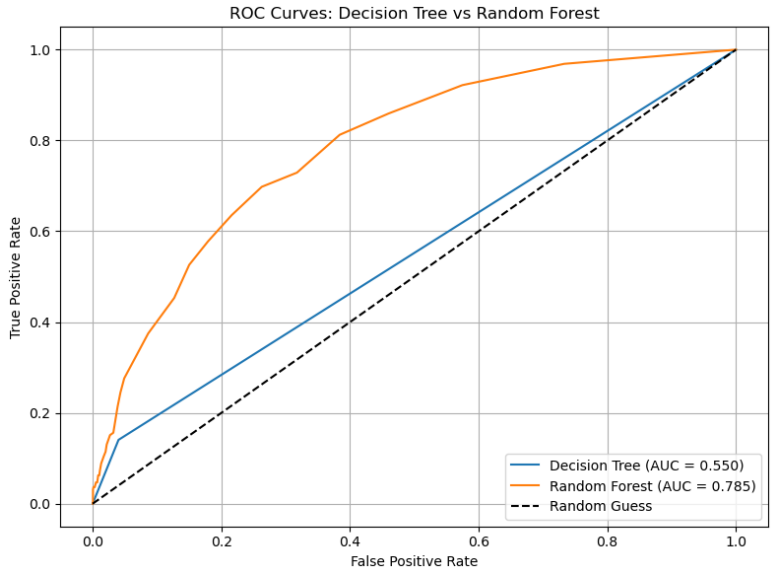



Figure 7: ROC Curves: Decision Tree and Random Forest without SMOTE

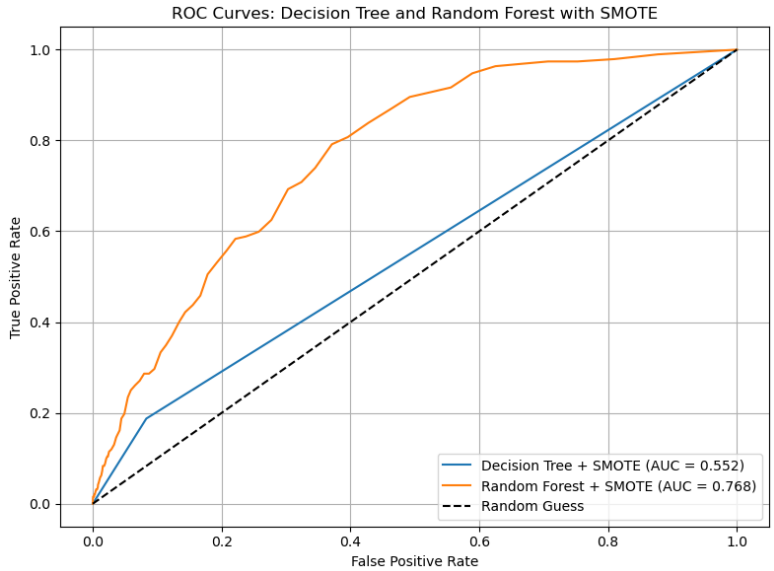



Figure 8: ROC Curves: Decision Tree and Random Forest with SMOTE

## 5.4 ROC Curve Comparison

Figure 7 and Figure 8 compare the classification performance of Decision Tree and Random Forest models with and without SMOTE. Each plot shows the model's ability to distinguish between good and bad credit applicants using predicted probabilities.

**Without SMOTE:** The Random Forest model significantly outperforms the Decision Tree, achieving an AUC-ROC of 0.785 versus 0.550. However, the Random Forest fails to recall any true positives due to the default threshold of 0.5 and extreme class imbalance. The Decision Tree makes some positive predictions, but with limited effectiveness.

**With SMOTE:** Both models show modest improvements in minority class detection. The Decision Tree improves slightly (AUC = 0.552), while the Random Forest maintains a strong AUC (0.768), despite some tradeoff in recall and precision. These results demonstrate that SMOTE increases the model's exposure to positive cases, which leads to better ranking performance and higher true positive rates at certain thresholds.

Random Forest consistently demonstrates stronger probability separation regardless of sampling strategy. However, SMOTE helps both models assign more meaningful probabilities to the positive class. Threshold tuning may further improve recall for business deployment.

# 6 Model Comparison and Business Recommendation

This section compares the performance of the classification models developed in previous stages, with a focus on supporting business decisions in loan approvals. Evaluation is based on predictive power, class imbalance handling, and practical interpretability.

## 6.1 Model Comparison Summary

We evaluated six models across key metrics:

- **Logistic Regression (with and without SMOTE)**
- **Decision Tree (with and without SMOTE)**
- **Random Forest (with and without SMOTE)**

The following table summarizes their validation performance:

Table 15: Model Comparison Summary

| Model | Accuracy | Recall (Class 1) | Precision (Class 1) | AUC-ROC |
|---|---|---|---|---|
| Logistic Regression (no SMOTE) | 69% | **75%** | 10% | **0.796** |
| Logistic Regression (SMOTE) | 72% | 71% | 11% | 0.779 |
| Decision Tree (no SMOTE) | 92% | 14% | 14% | 0.550 |
| Decision Tree (SMOTE) | 88% | **19%** | 10% | 0.552 |
| Random Forest (no SMOTE) | **95%** | 0% | 0% | 0.785 |
| Random Forest (SMOTE) | 94% | 8% | **18%** | 0.768 |

### 6.2 Interpretation

**Predictive Strength:** Logistic Regression (no SMOTE) achieved the best AUC (0.796) and the highest recall (75%) for the minority class. This model is better at identifying applicants with good credit, which aligns with the goal of reducing false rejections.

Random Forest with SMOTE had the highest precision for class 1 (18%), meaning when it predicted someone as a good credit risk, it was more likely correct. However, it missed most good applicants due to low recall (8%).

Decision Trees underperformed in both variants, with low AUC and inconsistent class 1 predictions.

**Practical Considerations:** Logistic Regression is simple, interpretable, and well-understood by business stakeholders. It assigns meaningful probabilities to each applicant, making it suitable for ranking and threshold tuning.

Random Forests are more complex and less transparent but offer stronger probability separation. They are better suited for risk-based scoring than binary classification.

### 6.3 Business Recommendation

We recommend using the **Logistic Regression model trained with class-weight balancing (without SMOTE)** as the primary model to assist in loan approval decisions.

- It achieves the highest recall and AUC, making it more inclusive of good applicants.
- The model is transparent and easily explainable.
- It allows threshold customization based on business risk tolerance.

In deployment, the model's predicted probabilities can be used to:

- **Rank applicants** by creditworthiness.
- **Set approval cutoffs** based on portfolio risk appetite.
- **Support analyst review** of borderline cases.

## 7 Fairness and Interpretability

In this section, we examine the transparency and fairness of our model to ensure its suitability for real-world loan decisioning. We address three business-critical questions: (i) how to explain a rejection to an applicant, (ii) whether our model exhibits gender bias, and (iii) whether there is racial disparity in predicted approvals.

## 7.1 Explaining Rejections to Customers

We selected a Logistic Regression model for its simplicity and interpretability. Its output can be explained in terms of the weighted contribution of each input feature to the final prediction. For instance, a low FICO score or a high loan-to-value ratio may negatively influence the outcome.

To enhance user-facing explanations, we use LIME (Local Interpretable Model-Agnostic Explanations), which produces feature-level attributions for individual decisions.

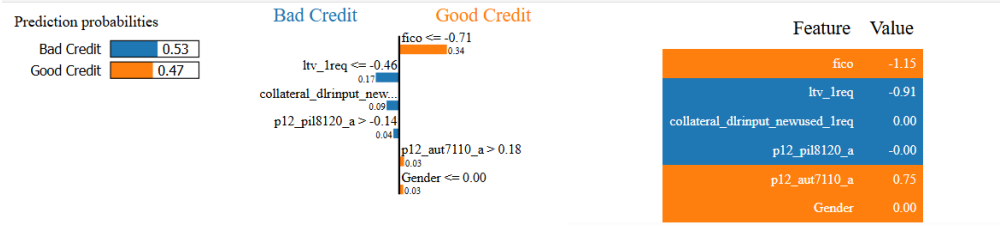



Figure 9: LIME explanation for a specific applicant

As shown in Figure 9, the features contributing most to the prediction were a low FICO score and a high LTV ratio. These insights make model decisions explainable to both internal stakeholders and external customers.

## 7.2 Fairness Analysis by Gender

We evaluated predicted approval rates across gender groups using the evaluation dataset and our trained model. Approval was defined as a prediction of good creditworthiness (`bad_flag` = 1).

Table 16: Approval Rate by Gender (Model Predictions)

| Gender | Approval Rate (%) |
|---|---|
| Female | 35.5 |
| Male | 32.2 |
| Undefined | 30.1 |

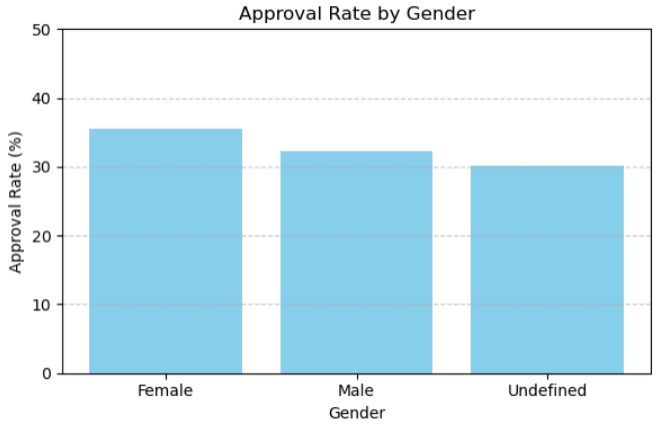



Figure 10: Model-predicted approval rate by gender

Approval rates are similar between males and females, suggesting no major gender-based disparity in model predictions.

### 7.3 Fairness Analysis by Race

We also evaluated the predicted approval rates across race categories:

Table 17: Approval Rate by Race (Model Predictions)

| **Race** | **Approval Rate (%)** |
|---|---|
| Asian | 29.5 |
| Black | 36.2 |
| Hispanic | 34.9 |
| Other | 37.0 |
| White | 31.9 |

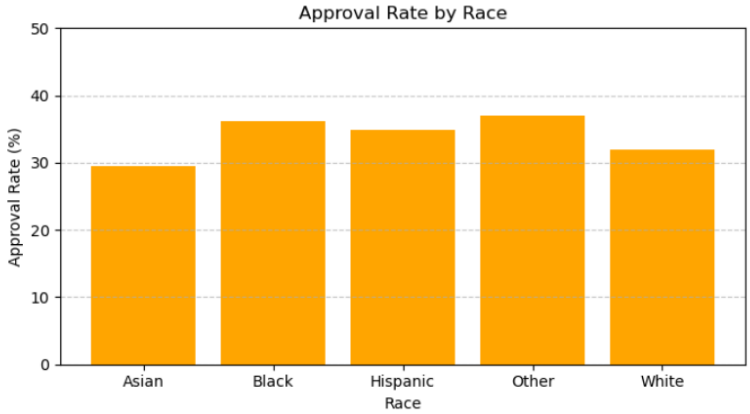



Figure 11: Model-predicted approval rate by race

Approval rates vary modestly across race groups, with no evidence of systematic bias against any one category. Thus, our model does not discriminate against customers with racial backgrounds.

# 8 Conclusion

In this project, we developed a predictive model for assisting loan approval decisions using historical application data. After evaluating multiple models, we selected a Logistic Regression classifier with class-weight balancing due to its strong recall for identifying good credit applicants, highest AUC-ROC, and ease of explanation. We used LIME to further enhance individual-level interpretability. Fairness analysis showed that the model does not significantly disadvantage applicants based on gender or race. Approval rates across these groups were consistent, and local explanations supported transparency in model decisions.

We recommend this model as a baseline for assisting auto loan approvals in a responsible and auditable manner. Future work may explore advanced calibration methods, human-in-the-loop decision systems, and continuous fairness monitoring.

| Feature | Unique Values (Train) | Unique values (Test) |
|---|---|---|
| amtfinanced_1req | 16961 | 5665 |
| ltv_1req | 7876 | 3763 |
| pti_1req | 2123 | 1633 |
| clall5010_a | 1071 | 286 |
| fico | 446 | 402 |
| p12_iln8220_a | 378 | 313 |
| p12_bca8370_a | 351 | 266 |
| p12_bcc8120_a | 330 | 237 |
| p12_all7120_a | 285 | 237 |
| p12_pil8120_a | 254 | 210 |
| p12_all8370_a | 248 | 189 |
| p12_aua8220_a | 237 | 201 |
| p12_reh7120_a | 169 | 135 |
| p12_bcx7110_a | 135 | 119 |
| p12_rtr7110_a | 131 | 112 |
| p12_aut7110_a | 109 | 105 |
| p12_all7170_a | 101 | 94 |
| p12_all7937_a | 96 | 86 |
| p12_all7938_a | 96 | 89 |
| p12_iln7410_a | 94 | 88 |
| p12_all7517_a | 86 | 71 |
| p12_all8150_a | 83 | 83 |
| p12_aua8151_a | 83 | 84 |
| p12_als1300_a | 82 | 73 |
| clact9429_a | 74 | 49 |
| p12_rev1300_a | 62 | 54 |
| clntr9437_a | 44 | 28 |
| p12_aua0300_a | 37 | 28 |
| clact9428_a | 36 | 28 |
| p12_bcc3456_a | 18 | 16 |
| p12_all6971_a | 7 | 7 |
| p12_aua6200_a | 6 | 6 |
| p12_alm6200_a | 6 | 6 |
| p12_all2427_a | 6 | 5 |
| p12_all6250_a | 6 | 6 |
| Race | 5 | 5 |
| cltra4405_a | 3 | 2 |
| Gender | 3 | 3 |
| aprv_flag | 2 | 2 |
| collateral_dlrinput_newused_1req | 2 | 2 |
| clall2434_a | 2 | 3 |
| cloil0214_a | 2 | 2 |
| bad_flag | 2 | 2 |

Table 2: Number of unique values by feature

Table 3: Distribution of Good Credit Proportions by Feature Value for `p12_all6250_a` (Worst Ever Status in First 12 Months)

| | Feature Value | Total Count | Percent Good Credit |
|---|---|---|---|
| 0 | 1.00 | 15697.00 | 3.36 |
| 1 | 30.00 | 2034.00 | 4.47 |
| 2 | 60.00 | 657.00 | 8.68 |
| 3 | 90.00 | 307.00 | 9.45 |
| 4 | 120.00 | 321.00 | 8.72 |
| 5 | 400.00 | 2332.00 | 9.65 |

Table 4: Distribution of Good Credit Proportions by Feature Value for `p12_aua6200_a` (Worst Ever Status on Auto Loan)

| | Feature Value | Total Count | Percent Good Credit |
|---|---|---|---|
| 0 | 1.00 | 12227.00 | 3.09 |
| 1 | 30.00 | 2196.00 | 4.14 |
| 2 | 60.00 | 690.00 | 7.54 |
| 3 | 90.00 | 194.00 | 8.76 |
| 4 | 120.00 | 81.00 | 11.11 |
| 5 | 400.00 | 1585.00 | 8.26 |

Table 5: Distribution of Good Credit Proportions by Feature Value for `p12_all2427_a` (Number of Trades Delinquency in Last 12 Months)

| | Feature Value | Total Count | Percent Good Credit |
|---|---|---|---|
| 0 | 0.00 | 20987.00 | 4.25 |
| 1 | 1.00 | 330.00 | 17.27 |
| 2 | 2.00 | 24.00 | 33.33 |
| 3 | 3.00 | 5.00 | 20.00 |
| 4 | 7.00 | 1.00 | 0.00 |
| 5 | 9.00 | 1.00 | 0.00 |

Table 6: Distribution of Good Credit Proportions by Feature Value for `p1_all697_a` (Worst Ever Status in Last 12 Months)

| | Feature Value | Total Count | Percent Good Credit |
|---|---|---|---|
| 0 | 0.00 | 65.00 | 12.31 |
| 1 | 1.00 | 16225.00 | 3.37 |
| 2 | 30.00 | 1389.00 | 5.83 |
| 3 | 60.00 | 392.00 | 9.95 |
| 4 | 90.00 | 178.00 | 6.18 |
| 5 | 120.00 | 328.00 | 10.37 |
| 6 | 400.00 | 2728.00 | 8.61 |

Table 7: Distribution of Good Credit Proportions by Feature Value for `p12_alm6200_a` (Worst Ever Status on Trade)

| | Feature Value | Total Count | Percent Good Credit |
|---|---|---|---|
| 0 | 1.00 | 9287.00 | 2.61 |
| 1 | 30.00 | 2472.00 | 2.43 |
| 2 | 60.00 | 906.00 | 3.53 |
| 3 | 90.00 | 511.00 | 5.09 |
| 4 | 120.00 | 1017.00 | 4.92 |
| 5 | 400.00 | 7155.00 | 7.66 |

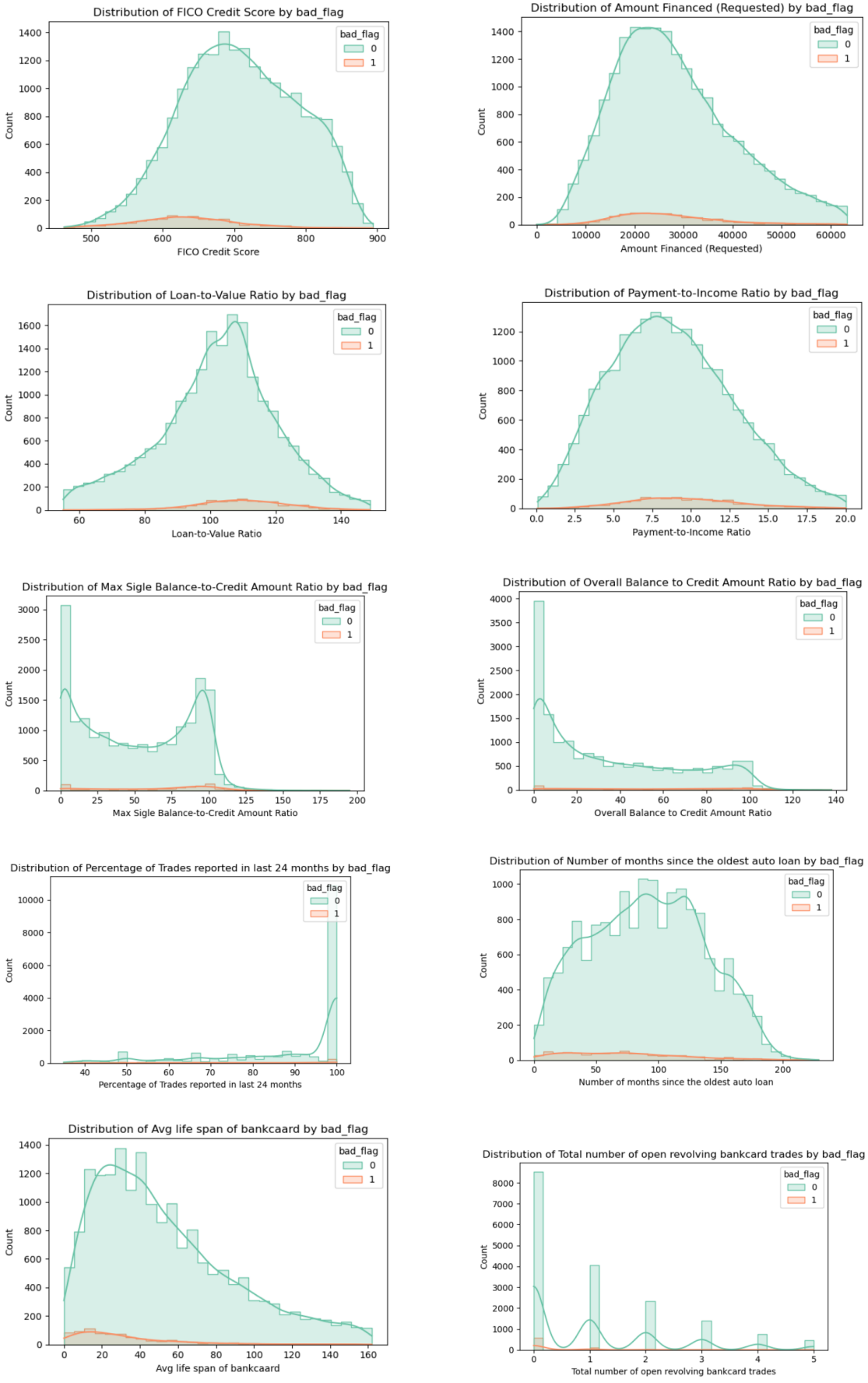



Figure 5: Bi-variate analysis of selected continuous features with target variable