Credit Risk Modeling with Python

A comprehensive machine learning approach to predicting loan defaults using customer demographics, credit history, and loan characteristics. This notebook demonstrates advanced techniques in data preprocessing, feature engineering, and model optimization for credit risk assessment.

Dataset Overview and Initial Setup

Three Core Datasets

Our analysis integrates data from three distinct sources to build a comprehensive view of loan applicants:

- **Customer Data:** 50,000 records containing demographics, employment status, income, and residence information
- **Loan Data:** Detailed loan characteristics including amounts, tenure, and disbursement details
- **Bureau Data:** Credit history metrics tracking account status, delinquencies, and credit utilization

Key Technical Approach

The project follows rigorous data science practices:

- Merged datasets using customer ID as the primary key
- Total of 33 features spanning demographic, financial, and behavioral dimensions
- Target variable: Binary default indicator (0 = no default, 1 = default)
- Class imbalance observed: 91.4% non-default vs 8.6% default cases

Strategic Train-Test Split

1 Preventing Data Leakage

We performed the train-test split **before** exploratory data analysis to ensure the test set remains completely unseen during feature engineering and model development. This critical step prevents information leakage that could artificially inflate model performance.

2 Stratified Sampling

Used stratified splitting to maintain the same proportion of default cases (8.6%) in both training and test sets. This ensures both datasets are representative of the overall population distribution.

3 Split Configuration

Training set: 37,500 records (75%)
Test set: 12,500 records (25%)
Both sets maintain identical target variable distributions for reliable evaluation.

Data Cleaning and Quality Assurance

Missing Value Treatment

The dataset exhibited minimal missing values, with only the **residence_type** column containing 47 missing entries (0.13% of training data). We applied mode imputation, filling missing values with "Owned" - the most frequent category representing stable housing status.

Categorical Data Correction

Identified and corrected a data entry error in the **loan_purpose** column where "Personaal" was misspelled. This was standardized to "Personal" across both training and test sets to ensure consistency.

Business Rule Validation

Applied domain knowledge to validate data integrity:

- Processing fees exceeding 3% of loan amount flagged as outliers
- Verified GST calculations remained within 20% threshold
- Confirmed net disbursement never exceeded loan amount
- Removed 12 outlier records that violated these business rules

Exploratory Data Analysis: Age and Risk Patterns

Kernel Density Estimation reveals critical age-related patterns in default behavior. The orange distribution (defaulters) shows a pronounced shift toward younger ages, with peak density around 35 years. Non-defaulters (blue) exhibit a more normal distribution centered at 40 years. This 5-year gap suggests **younger borrowers face higher default risk**, potentially due to less established financial stability and shorter credit histories.

Statistical Insights

- Average age of defaulters: 37.1 years
- Average age of non-defaulters: 39.8 years
- Standard deviation similar across groups (~9 years)
- Age range: 18-70 years for both categories

Risk Implications

The leftward shift in the defaulter distribution indicates that credit risk models should weight age as a predictive factor, with particular attention to applicants under 35 years old.

Comprehensive Feature Analysis

Strong Predictive Signals Identified

Kernel density plots across all 19 continuous features reveal four variables with clear discriminative power between defaulters and non-defaulters:

1 Loan Tenure Months

Longer loan periods correlate with increased default risk, suggesting extended repayment obligations create financial strain

2 Delinquent Months

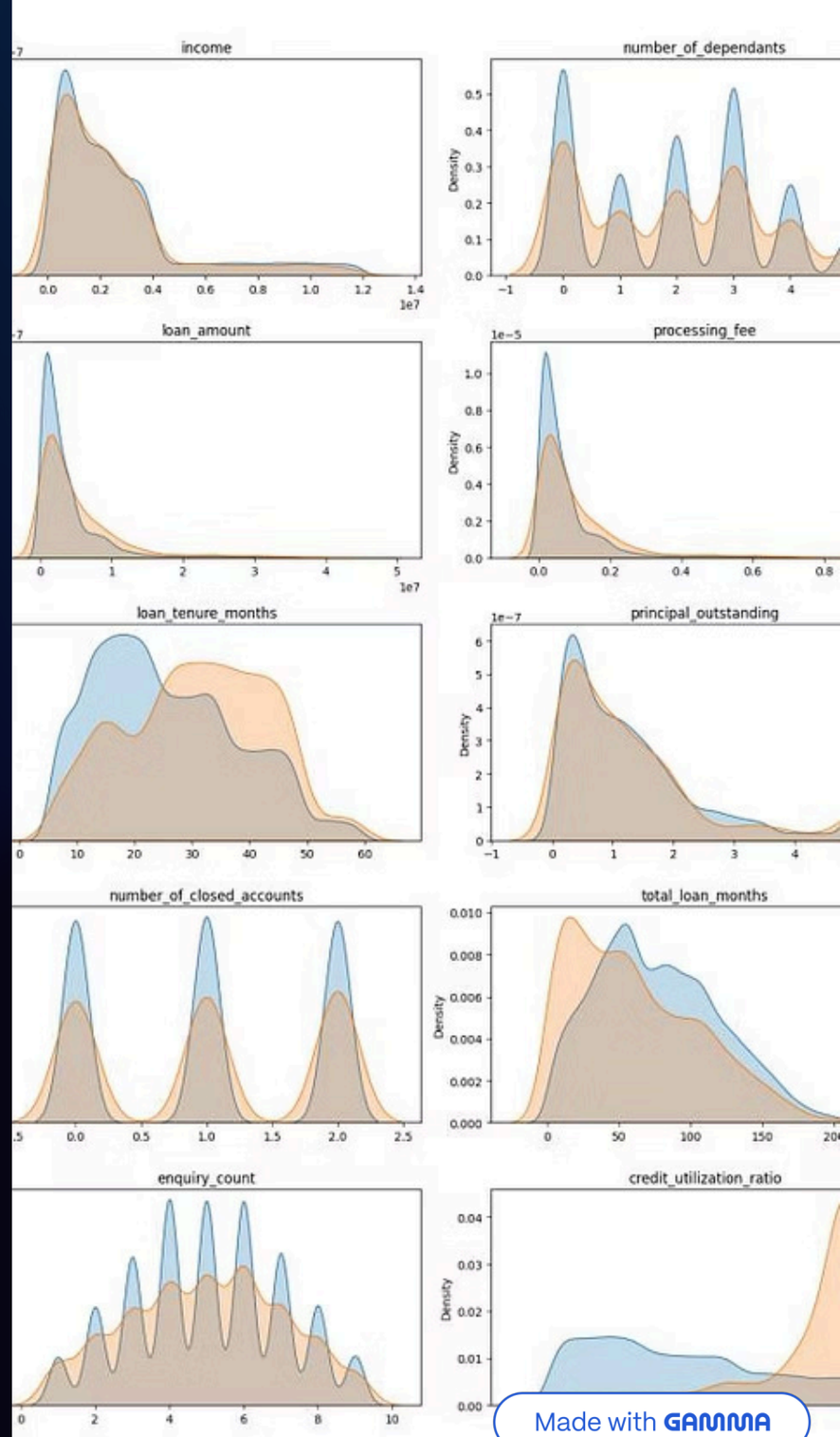
Past payment delays strongly predict future defaults - the orange curve shows concentration at higher delinquency values

3 Total Days Past Due (DPD)

Cumulative payment lateness demonstrates the strongest separation between groups, indicating behavioral patterns

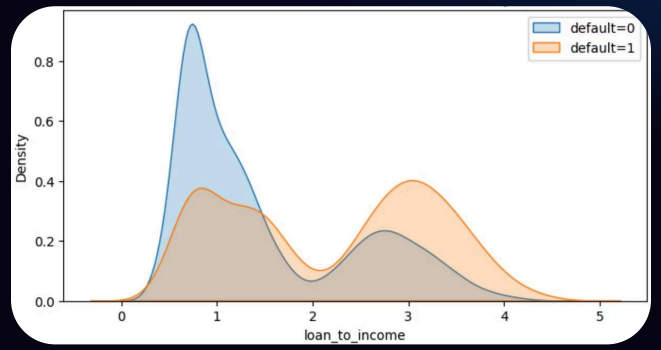
4 Credit Utilization Ratio

Higher credit utilization (>60%) emerges as a red flag, reflecting financial stress and over-leverage



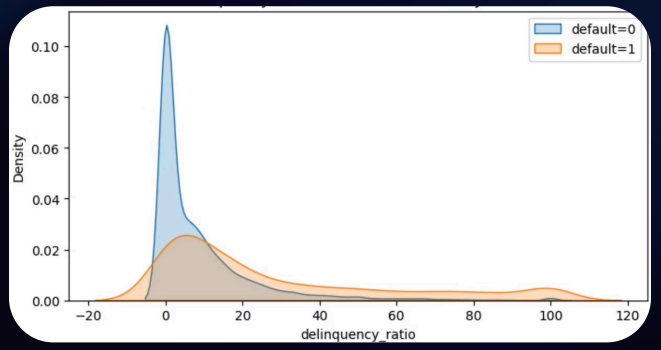
Feature Engineering: Creating Powerful Predictors

Loan-to-Income Ratio



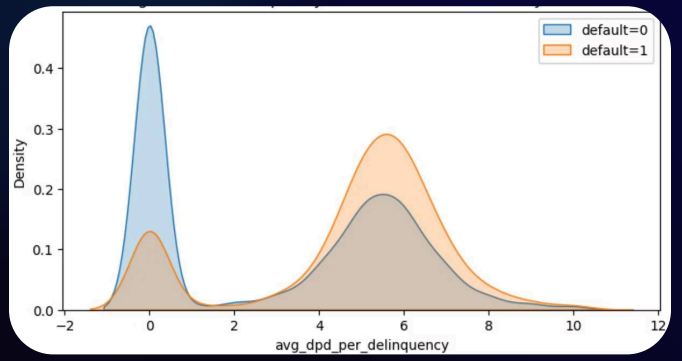
Combining loan amount with income reveals affordability stress. Higher LTI ratios (>2.5) show strong correlation with defaults, indicating borrowers taking loans beyond their repayment capacity.

Delinquency Ratio



The percentage of months spent delinquent relative to total loan history. Defaulters show concentration at higher ratios, validating this as a risk indicator.

Average DPD Per Delinquency



Measures severity of payment delays. Clear separation between groups indicates this engineered feature captures risk behavior better than raw delinquency counts.

Feature Selection: VIF and Information Value Analysis

1

Multicollinearity Check

Calculated Variance Inflation Factor for all numeric features. Removed highly correlated variables (sanction_amount, processing_fee, gst, net_disbursement) with VIF > 10 to prevent model instability.

2

Information Value Scoring

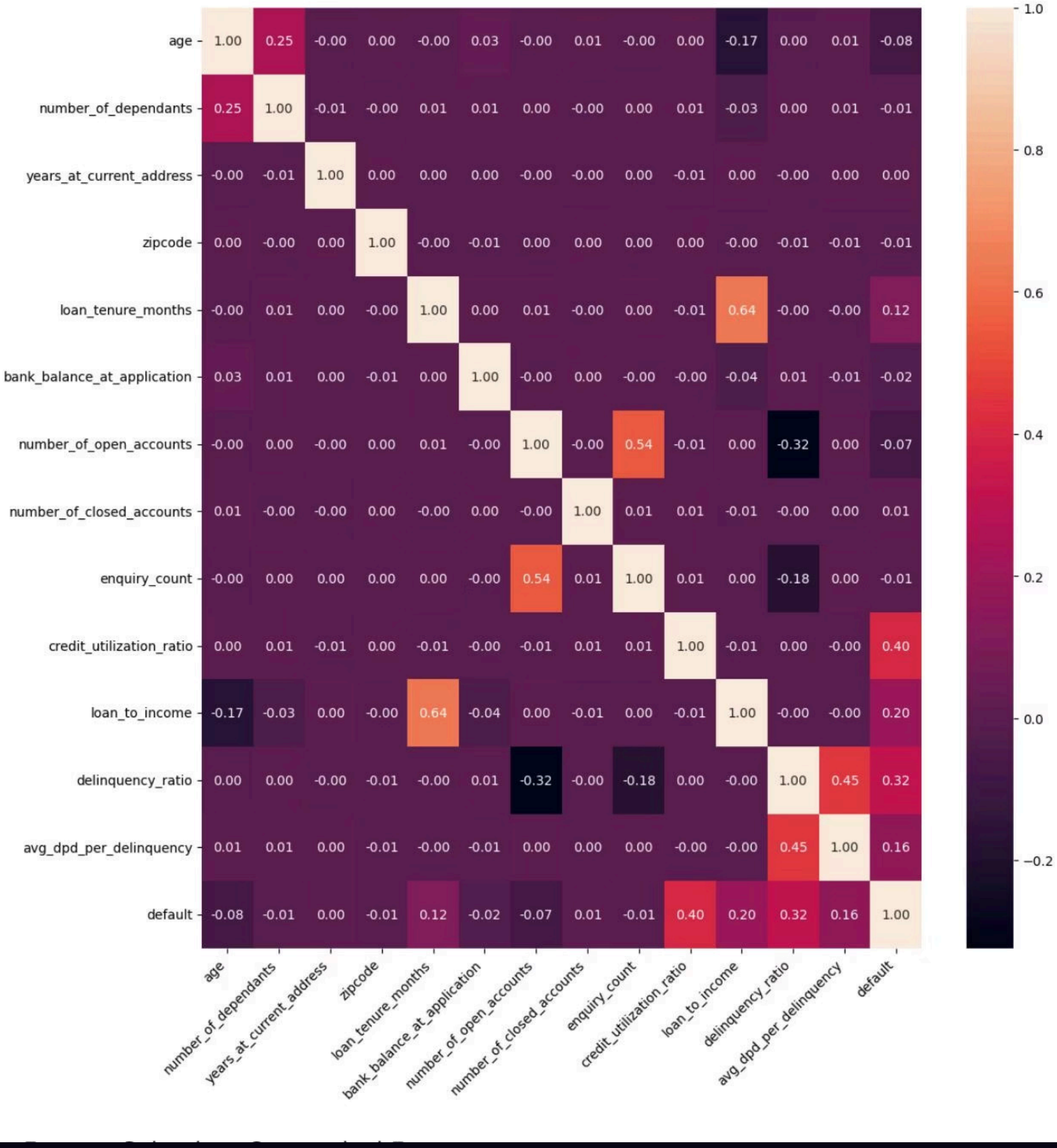
Computed Weight of Evidence and IV for categorical features. Selected 10 features with IV > 0.02, indicating predictive power. Top performers: credit_utilization_ratio (IV=2.35), delinquency_ratio (IV=0.72).

3

Final Feature Set

Retained 13 features after rigorous selection: age, residence_type, loan_purpose, loan_type, loan_tenure_months, and 8 other key predictors balancing statistical significance with business interpretability.

Feature Correlation Matrix



The correlation analysis confirms our feature selection strategy. Engineered features (loan_to_income, delinquency_ratio, avg_dpd_per_delinquency) show moderate correlation with the target variable while maintaining independence from each other, maximizing predictive information without redundancy.

Model Development and Optimization

01

Handling Class Imbalance

Applied SMOTETomek technique to balance the 91:9 class distribution. This hybrid approach combines SMOTE (synthetic minority oversampling) with Tomek links removal, creating a balanced training set of 34,195 samples per class.

02

Hyperparameter Tuning with Optuna

Employed Bayesian optimization across three algorithms: Logistic Regression, XGBoost, and Random Forest. Each model underwent 10 optimization trials, maximizing macro-averaged F1-score through 3-fold cross-validation.

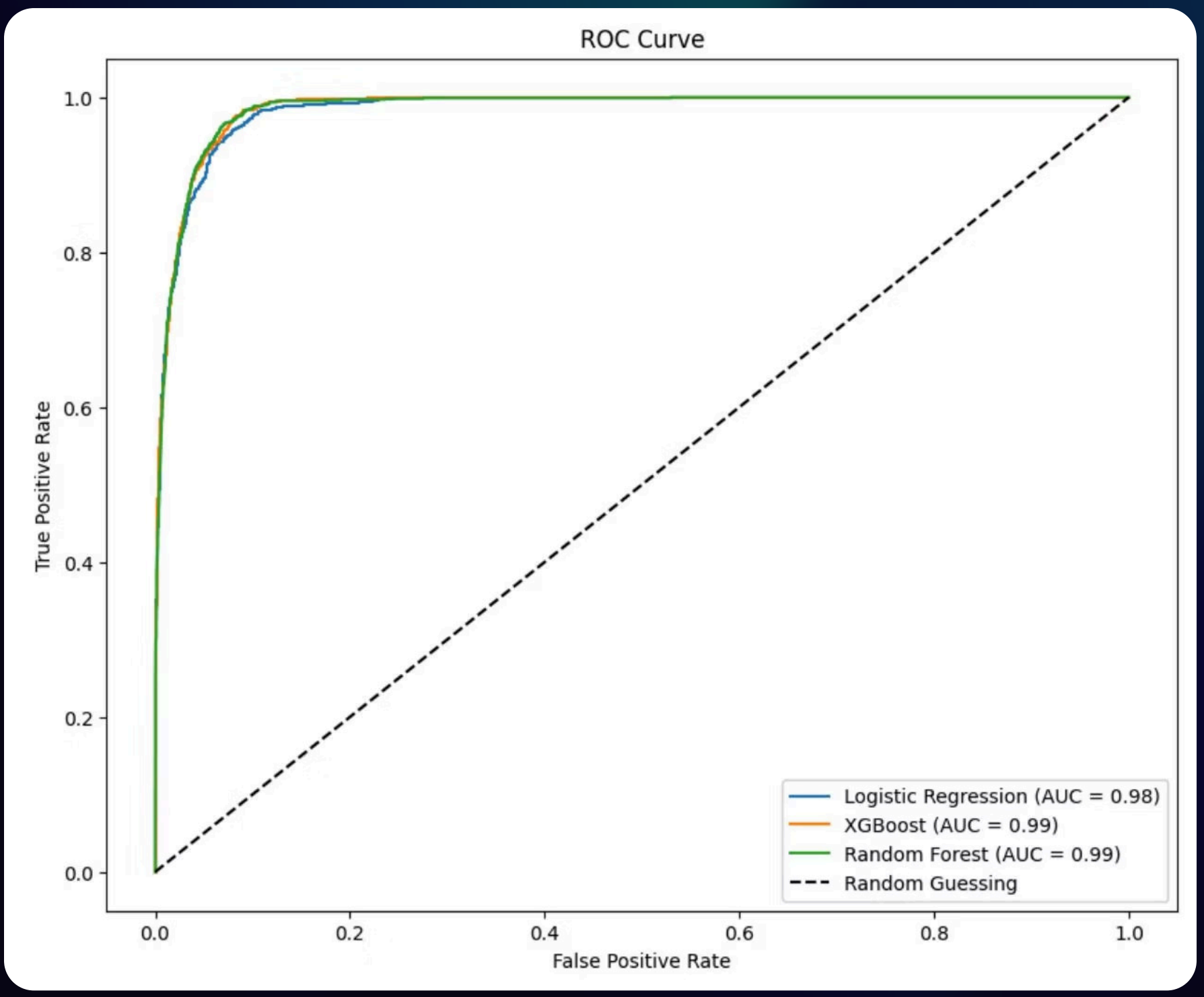
03

Scaling and Encoding

Applied MinMaxScaler to normalize continuous features to [0,1] range. Used one-hot encoding for categorical variables (residence_type, loan_purpose, loan_type) to create model-ready input features.

Model Performance Comparison

Final Model Selection: Logistic Regression



Logistic Regression Emerges as the Optimal Model for Credit Risk

While the ROC curve analysis shows XGBoost with AUC = 0.98 and Random Forest with AUC = 0.97, Logistic Regression, achieving **AUC = 0.96**, is selected as the optimal model for this credit risk application due to its superior recall performance for the default class. A high recall of **95%** is critical in credit risk assessment, as it ensures that a maximum number of potential defaulters are identified, minimizing financial losses for the bank even if it means a slightly lower precision.

96%

Logistic Regression AUC Score

Strong discriminative power, balancing sensitivity and specificity effectively

95%

Default Recall

Successfully identifies 95% of actual default cases, minimizing critical false negatives

76%

Default Precision

56% of predicted defaults are true positives, balancing risk mitigation with customer considerations

83%

Macro F1-Score

Good overall balance between precision and recall across both classes

Key Takeaways and Business Impact

This credit risk model achieves production-ready performance through systematic feature engineering, rigorous feature selection using VIF and Information Value analysis, and advanced hyperparameter optimization. The **Logistic Regression model's 95% recall** on the default class means the bank can flag high-risk applications early, significantly reducing potential losses. Prioritizing recall in this context is crucial, as the cost of a missed default is often higher than the cost of a false positive. This strategic choice minimizes false negatives, providing robust protection against credit defaults while maintaining operational efficiency.

Furthermore, for practical application and real-time credit risk assessment, the final Logistic Regression model has been successfully deployed on Hugging Face, making it readily accessible for production use and integration into existing systems.