Project Title: Loan Approval Prediction Using Machine Learning

Project Overview:

This project aims to predict loan approval status (approved or rejected) using a set of demographics, financial, and loan-related features. The dataset includes 45,000 records and 14 variables, representing a synthetic version of a Credit Risk dataset. The goal is to build a reliable classification model that can distinguish between approved and rejected loans based on the provided features. The project involves data exploration, preprocessing, feature engineering, and model evaluation of various classification algorithms.

Project Description:

The dataset used in this project contains information on individuals applying for loans, including demographic details (age, income, gender), loan-related attributes (loan amount, interest rate, loan intent), and financial status (credit score, homeownership). Using this data, the objective is to predict whether a loan will be approved or rejected. The project follows a structured approach, beginning with an in-depth Exploratory Data Analysis (EDA) to uncover patterns and relationships, followed by preprocessing to prepare the data for machine learning models, and ending with the training and evaluation of several classification algorithms.

Key Steps and Analysis:

1. Data Loading and Preliminary Exploration

- **Objective**: Ensure data quality by identifying and addressing missing or anomalous values.
- **Actions**:
    - Filtered out unrealistic ages (>100) and income outliers.
    - Ensured no missing values were present.
    - Data summary confirmed the dataset had a balanced mix of numeric and categorical features.

2. Exploratory Data Analysis (EDA)

- **Objective**: Understand the structure of the dataset and identify relationships and trends that inform model development.
- **Key Findings**:
    - **Categorical Variables**: Gender and homeownership were key influencers of loan approval outcomes.
    - **Numerical Variables**: Income, loan amount, and credit score had significant impacts on loan decisions.
    - **Interest Rates**: Defaulted loans had higher interest rates, particularly in loan intents like debt consolidation and venture loans.
    - **Age Analysis**: The highest loan application rates were seen in the 20-30 age group.

3. Data Preprocessing

- **Objective**: Prepare the data for machine learning models by encoding categorical features, standardizing or normalizing data as necessary, and splitting it into training and testing sets.
- **Actions**:
    - Applied one-hot encoding to categorical features like gender, homeownership, and loan intent.
    - Normalized data for KNN, while standardizing it for models like Logistic Regression and Random Forest.
    - Split the dataset into 80% training and 20% testing data, with cross-validation for robust model performance evaluation.

4. Model Selection and Evaluation

- **Objective**: Build classification models and evaluate their performance using appropriate metrics (accuracy, precision, recall, F1-score, ROC-AUC).
- **Models Tested**:
    - **Logistic Regression**: Baseline model for binary classification.
    - **Random Forest**: Ensemble model known for capturing complex feature interactions.
    - **K-Nearest Neighbors (KNN)**: Distance-based algorithm sensitive to feature scale.
    - **Naive Bayes**: Probabilistic model assuming feature independence.
- **Performance Summary**:
    - **Random Forest**: Best performer with 92.69% accuracy, 0.8223 F1-score, and 0.9735 ROC-AUC.
    - **Logistic Regression**: Strong baseline with a high ROC-AUC (0.9541), suitable for interpretability.
    - **Naive Bayes**: High recall (97.61%) but struggled with precision, making it less suitable for balanced performance.
    - **KNN**: Performed moderately well but had lower recall, struggling to identify all rejected loans.

5. Key Insights

- **Best Model**: Random Forest was the most balanced and accurate, making it the best candidate for deployment.
- **Model Suitability**: Logistic Regression is ideal when model interpretability is critical, while Random Forest excels in scenarios requiring high predictive power.
- **Preprocessing Importance**: Standardizing and normalizing features significantly impacted the models' performance, particularly for KNN and Random Forest.

6. Conclusion

- The project demonstrates that ensemble methods like Random Forest are superior for loan approval prediction, balancing accuracy and model interpretability.
- Logistic Regression provides an excellent baseline, particularly when understanding the relationships between features is necessary.
- The feature engineering and preprocessing steps laid a strong foundation for building robust models that perform well in real-world applications.

Part 1: Data Loading and Preliminary Exploration

1. Introduction

The initial step of the project involved importing, exploring, and cleaning the dataset to prepare it for further analysis. The primary goal was to ensure the data's quality and consistency, addressing any anomalies or outliers that could potentially skew the modeling process.

2. Libraries Used

To begin, the following Python libraries were imported:

- **Pandas**: For data manipulation and analysis.
- **NumPy**: For numerical operations.
- **Matplotlib and Seaborn**: For visualizing distributions and identifying patterns in the dataset.

3. Data Loading

The dataset, which contains 45,000 records and 14 variables, was loaded into a Pandas DataFrame. An initial examination was conducted using the info() and describe() functions to understand the structure, data types, and statistical summaries of the features.

4. Data Quality and Anomaly Detection

During this step, a thorough examination of the dataset revealed some inconsistencies and anomalies:

- **Age Anomalies**: A small subset of entries included individuals with an age greater than 100 years, which is unrealistic in the context of financial risk assessment.
- **Income Outliers**: Some records contained annual income values exceeding $5 million, which is implausibly high for typical loan applicants.

5. Data Cleaning

To address these issues:

- **Age Filtering**: The dataset was filtered to include only individuals between 20 and 70 years old, representing a more realistic range for financial activity.
- **Retained Realistic Income**: Extremely high-income entries were excluded to focus on a plausible dataset for analysis.

6. Missing Data Analysis

The dataset was checked for missing values across all features using the isnull().sum() function. No missing values were detected, ensuring completeness of the data.

7. Outcome of Preliminary Exploration

At the end of this phase:

- The dataset was cleaned to remove illogical entries, improving its overall quality.
- The age range of individuals was restricted to 20–70 years to ensure relevance and accuracy.
- There were no missing values, allowing us to proceed to the next stage without imputation.

By addressing anomalies and verifying the data's integrity, a strong foundation was established for exploratory data analysis and predictive modeling in subsequent parts of the project.

Part 2: Exploratory Data Analysis (EDA)

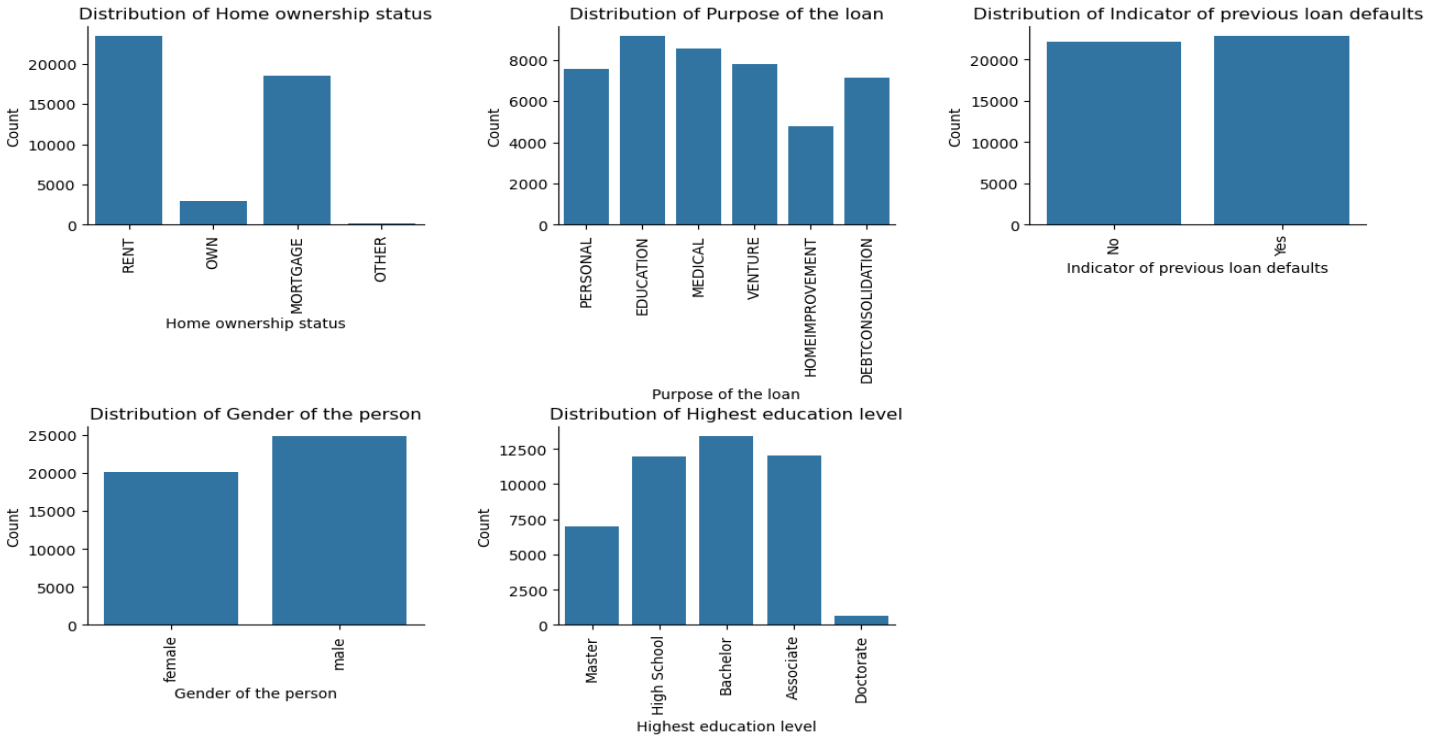
1. Objectives

The primary aim of this phase was to understand the dataset's structure and key features, identify potential relationships and trends, and derive insights that would inform the predictive modeling process. The EDA included both univariate and multivariate analyses, as well as feature-specific explorations.

2. Univariate Analysis

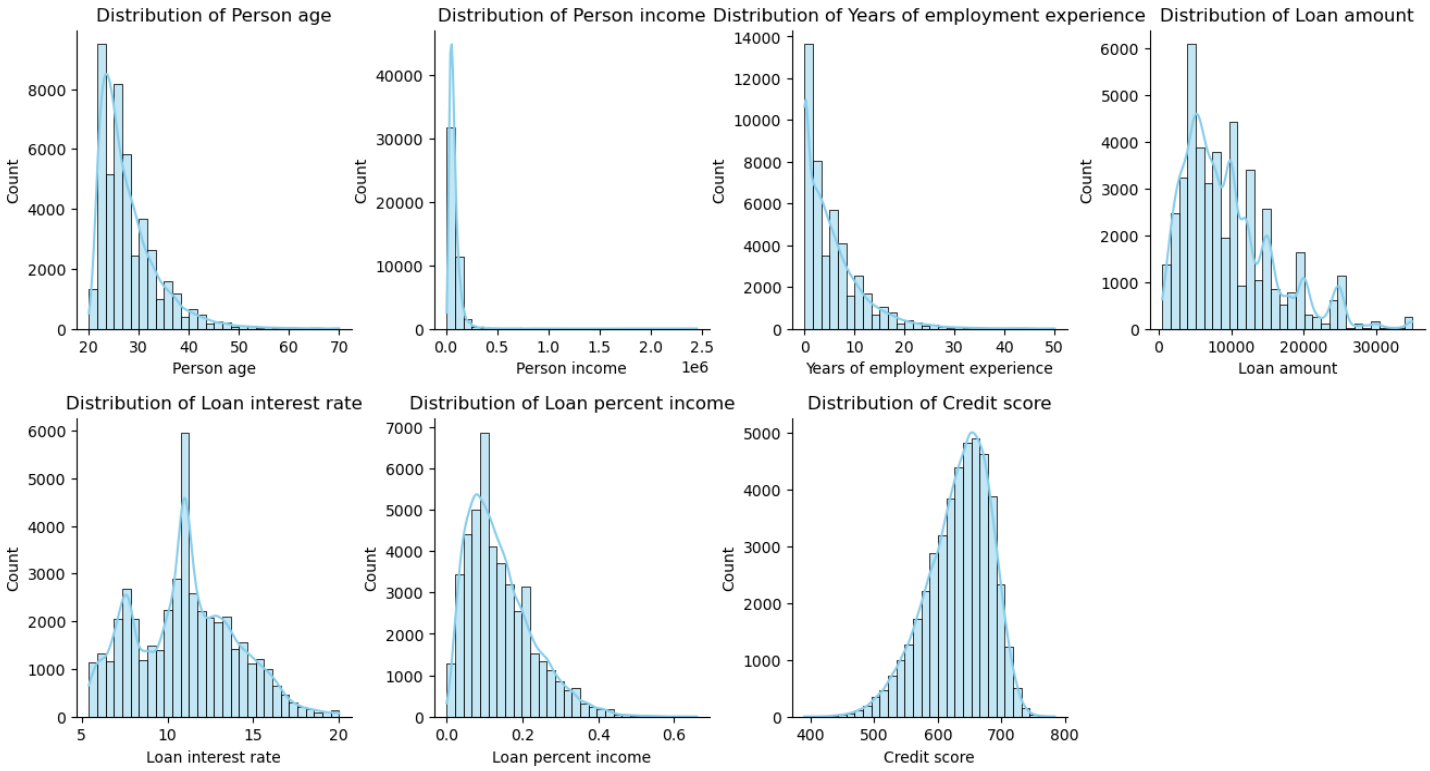
a. Categorical Variables

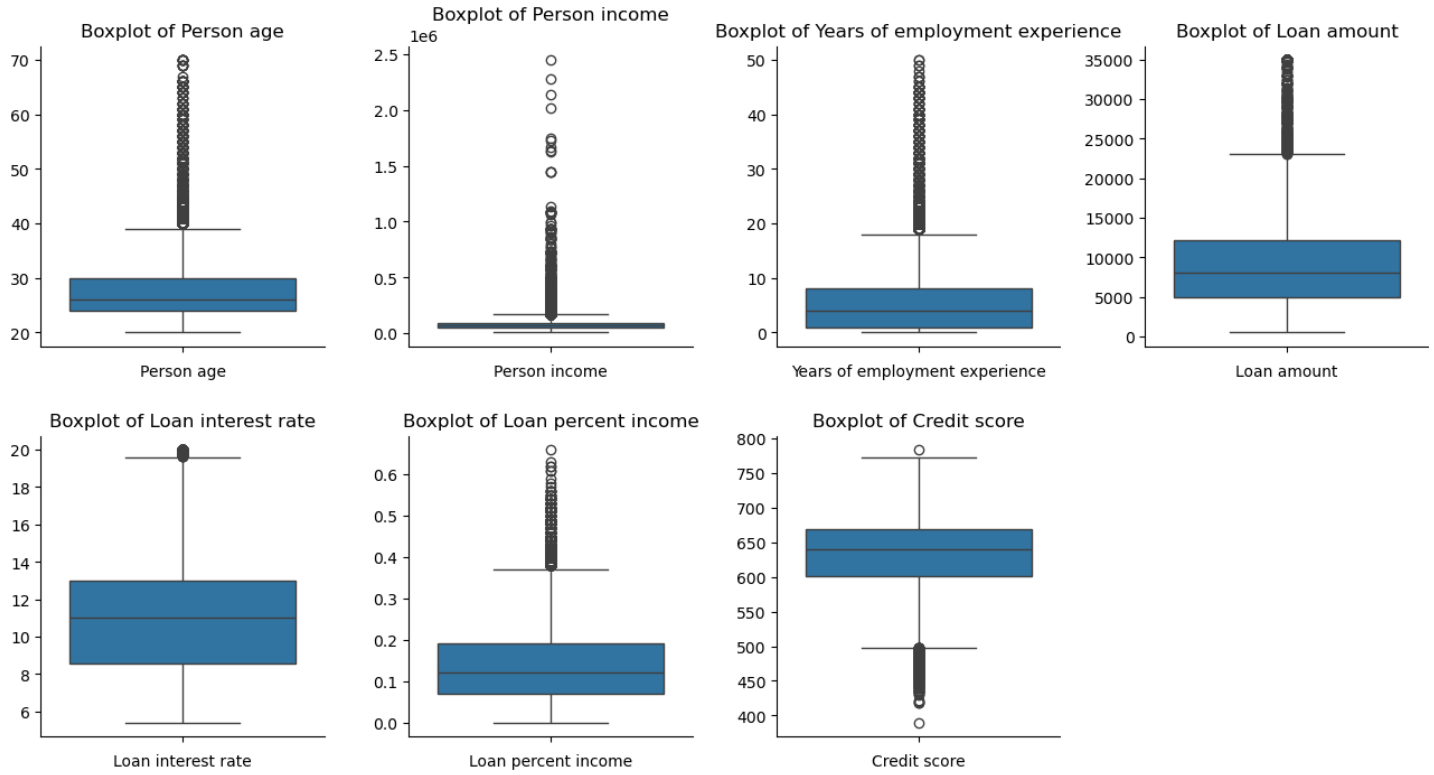
- **Methodology**: Checked the value counts for each categorical column and visualized their distributions using bar plots.
- **Findings**:
    - **Gender**: A balanced distribution (~50% male and 50% female).
    - **Indicator of Previous Loans**: Approximately equal proportions of defaults and non-defaults.
    - **Education Level**: High school, bachelor's, and associate degrees were the most common, while master's and doctorate levels were the least common.
    - **Home Ownership**: Rent and mortgage were the dominant categories, with "own" and "other" being less frequent.
    - **Loan Intent**: Distribution across purposes (education, medical, personal, etc.) appeared normal, with no category dominating significantly.



b. Numerical Variables

- **Methodology**: Plotted histograms and boxplots for numeric columns to examine distributions and detect outliers.
- **Findings**:
  - **Distribution**: Most numeric variables exhibited normal distributions, with slight skewness observed in person_income and person_age.
  - **Outliers**: Detected in all numeric columns using boxplots, but they were retained to preserve data integrity and account for potential future instances.





3. Multivariate Analysis

a. Pairwise Relationships

- **Methodology**: Created a pairplot with loan_status as the hue to explore relationships between numeric variables.
- **Findings**: Patterns indicated clusters of approved and rejected loans based on key features like income, loan amount, and credit score.

b. Correlation Analysis

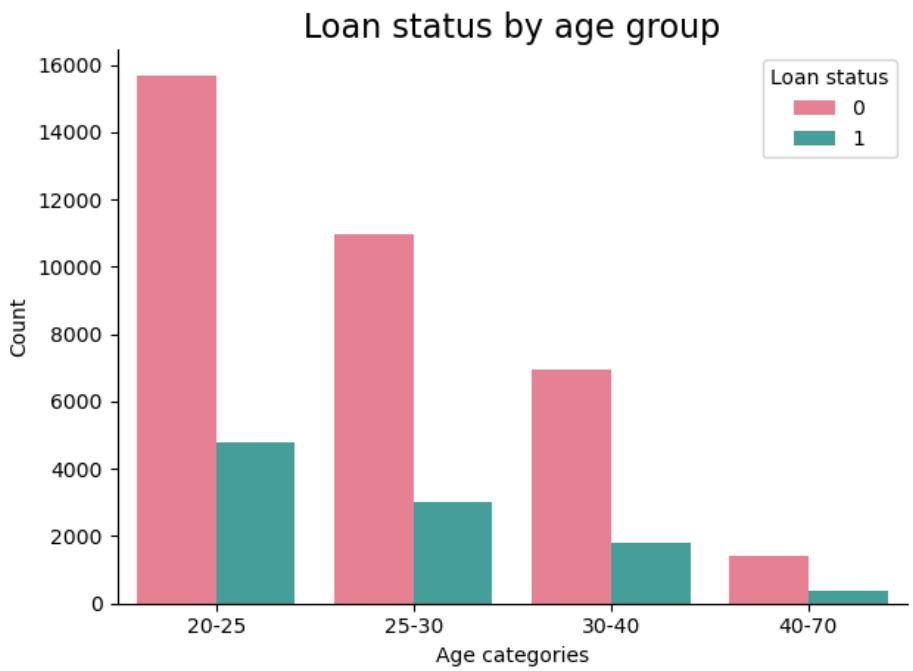
- **Methodology**: Computed a correlation matrix to identify relationships between features.
- **Key Correlations**:
    - Positive correlations:
        - loan_int_rate (interest rate) and loan_status (0.33).
        - loan_percent_income (loan as a percentage of income) and loan_status (0.38).
    - Negative correlation:
        - person_income and loan_status (-0.17).

4. Feature-Specific Analysis

a. Age of the Person

- **Methodology**: Binned person_age into 4 groups: 20–25, 25–30, 30–40, and 40–70. Visualized using a grouped bar chart.
- **Findings**:

- Age 20–30 had the highest number of loan applications.

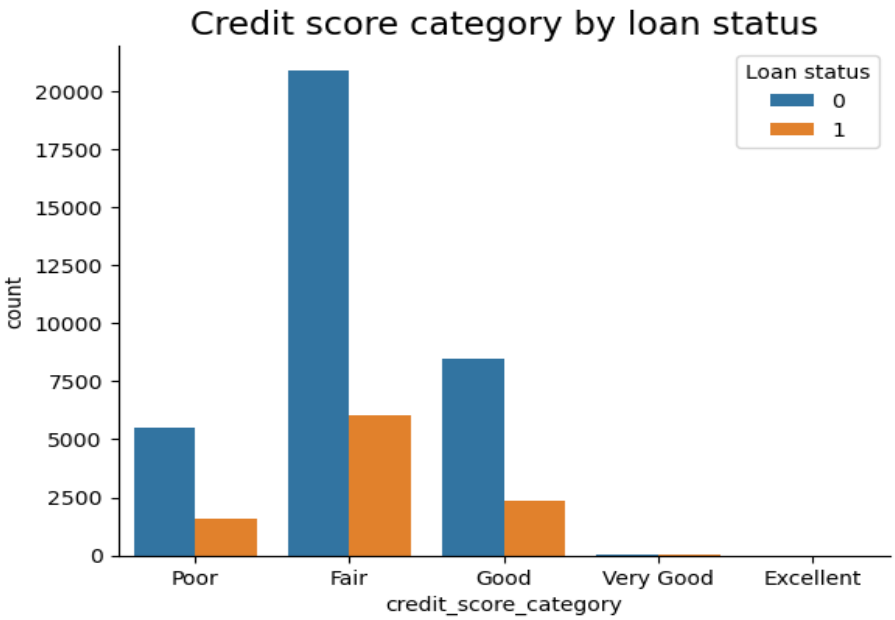


b. Affordability Ratios

- **Methodology**: Examined loan_percent_income to understand loan affordability.
- **Findings**: Higher loan percentages relative to income correlated with increased rejection rates.

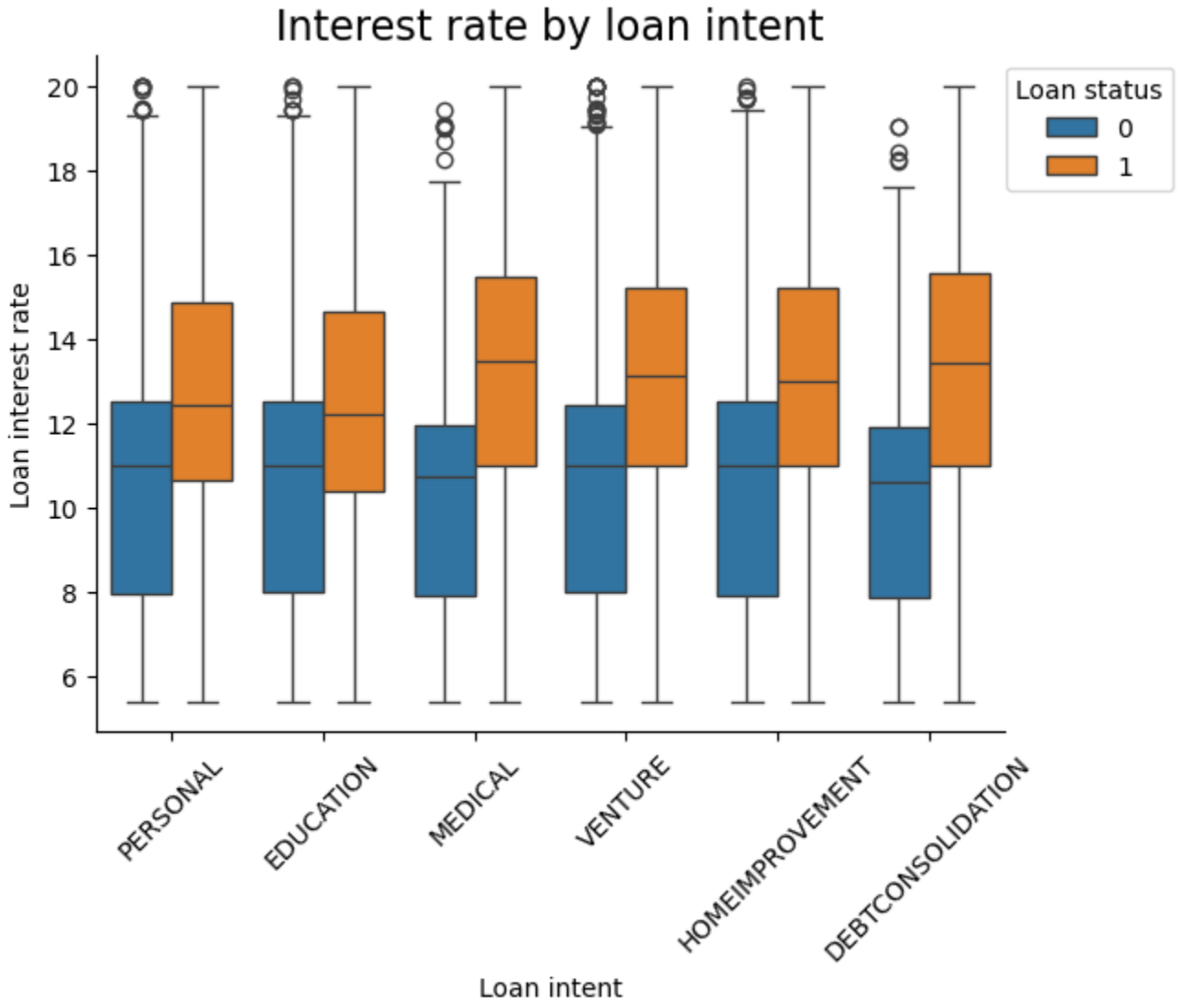
c. Credit Score

- **Methodology**: Categorized credit scores into five groups: Poor, Fair, Good, Very Good, and Excellent.
- **Findings**:
  - Applicants with “Fair” credit scores requested the most loans, followed by “Good” and “Poor.”
  - “Very Good” and “Excellent” categories had relatively fewer applicants.



d. Interest Rates by Loan Intent

- **Methodology**: Used boxplots to compare interest rates for approved and rejected loans across different intents.
- **Key Insights**:
    1. **Higher Interest Rates for Defaulted Loans**:
        - Defaulted loans generally had higher median interest rates across most categories (e.g., Venture and Debt Consolidation).
    2. **Variation Across Intents**:
        - Categories like Personal, Venture, and Home Improvement showed larger variations in interest rates, while Education and Medical loans had smaller spreads.
    3. **Outliers**:
        - Prominent in categories such as Debt Consolidation, indicating special cases.
    4. **Education Loans**:
        - Median interest rates were similar for both approved and rejected loans, suggesting interest rates might not significantly impact repayment in this category.



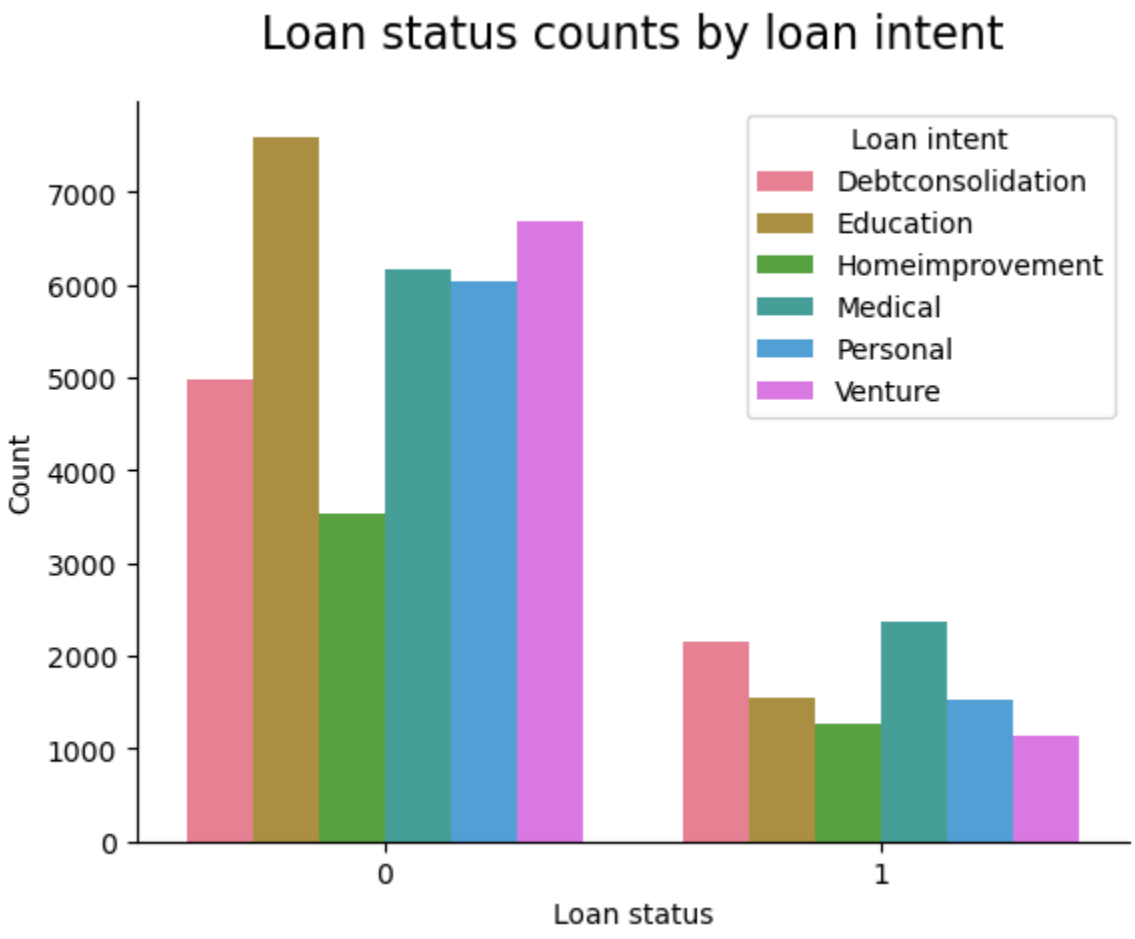
5. Loan Status Comparisons

a. Loan Approval and Rejection by Gender

- **Findings**:
    - Male applicants tended to request more loans and had higher rejection and approval rates compared to females.

b. Loan Status by Loan Intent

- **Findings**:
    - **Rejections**:
        - Education loans faced the highest rejection rates, followed by medical, personal, and venture loans.
        - Home improvement loans had the lowest rejection rates.
    - **Approvals**:
        - Medical loans were the most approved, followed by debt consolidation and education loans.



c. Home Ownership vs. Loan Status

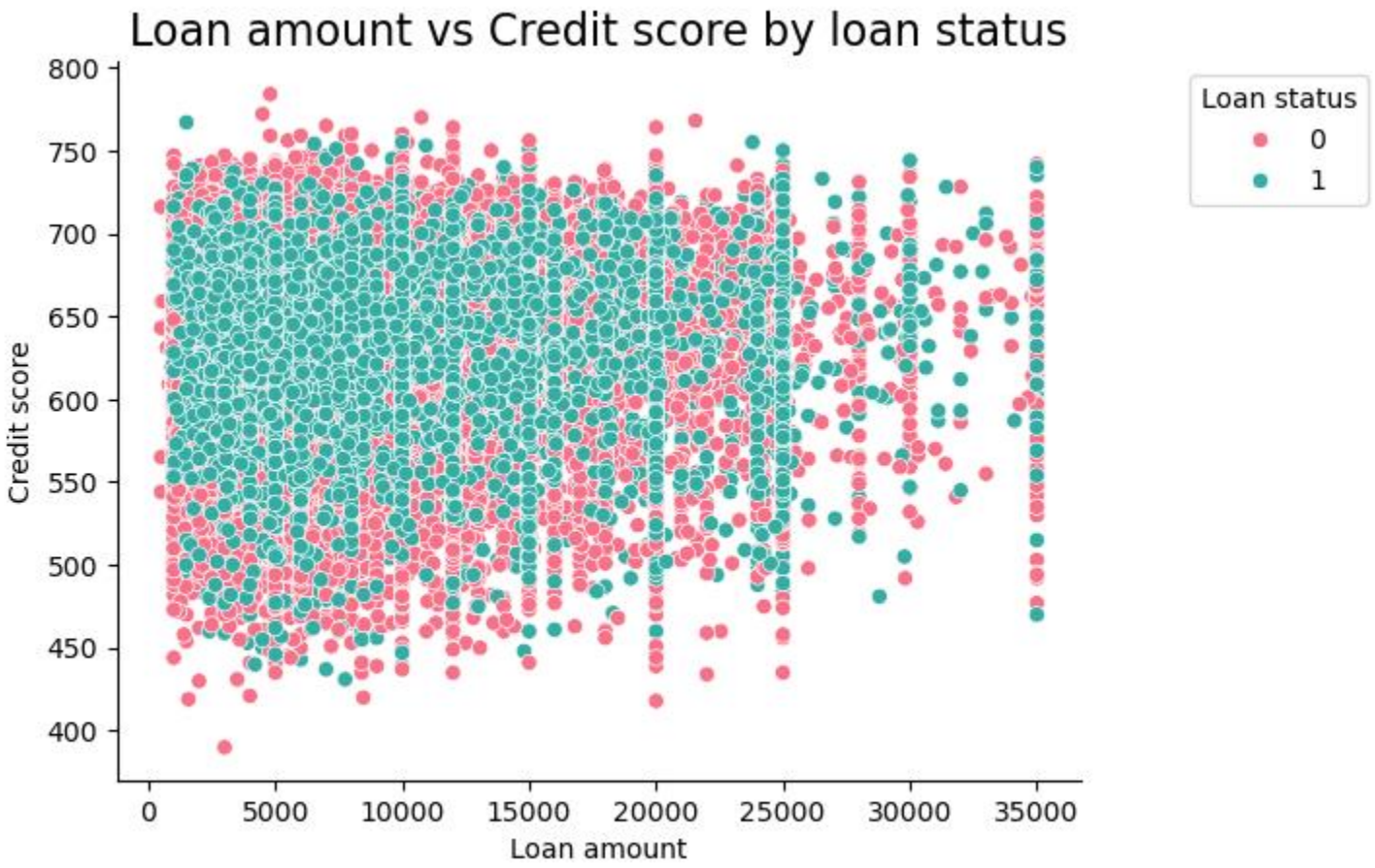
- **Findings**:
    - Renters had the highest number of approved loan requests.
    - Mortgage holders faced the most rejections, followed by renters and owners.

d. Income by Loan Status

- **Methodology**: Boxplots were used to analyze income distribution.
- **Findings**:
    - Rejected applications had significant outliers, with incomes above $500,000 often rejected.
    - Approved applications were concentrated between $30,000 and $500,000.

e. Loan Amount vs. Credit Score by Loan Status

- **Findings**:
    - Approved loans were clustered between credit scores of **550–700** and loan amounts ranging from **$1,000–$25,000**.



6. Summary of EDA

The EDA revealed several patterns and trends:

1. **Categorical Variables**: Gender, loan intent, homeownership, and credit score categories strongly influence loan outcomes.
2. **Numeric Variables**: Income, loan amount, credit score, and affordability ratios provide critical insights into approval and rejection trends.
3. **Interest Rates**: Higher interest rates correlate with increased rejection risks for specific loan intents.

4. **Outliers**: Retaining outliers ensured data integrity while uncovering anomalies in income and loan amount distributions.

These findings provide a solid foundation for feature engineering and model development in the next phase of the project.

Part 3: Modeling

1. Objective

The objective was to build and evaluate classification models to predict the loan approval status (loan_status). Given the binary nature of the target variable (approval or rejection), classification algorithms were selected.

2. Preprocessing

a. Feature Encoding

- All categorical variables were converted to numerical representations using one-hot encoding, ensuring the data was suitable for machine learning algorithms.
- Based on the insights from EDA, all features (both numeric and categorical) were retained as they contributed to predicting loan outcomes.

b. Normalization and Standardization

- **Standardization**: Applied to features for **Logistic Regression**, **Random Forest**, and **Naive Bayes**, as these models benefit from standardized inputs due to their reliance on distances, weights, or probabilities.
- **Normalization**: Used for **K-Nearest Neighbors** (KNN) since it is a distance-based algorithm and is highly sensitive to the scale of features.

c. Data Splitting

- The dataset was split into **80% training data** and **20% testing data**.
- Cross-validation was performed on the training data to ensure the models generalized well. The test set was reserved for final evaluation.

d. Metrics for Evaluation

To comprehensively evaluate model performance, the following metrics were used:

- **Accuracy**: Overall correctness of the predictions.
- **Precision**: The ability to identify only the relevant cases (approved loans).
- **Recall**: The ability to identify all relevant cases.
- **F1-Score**: A balance between precision and recall.
- **ROC-AUC**: The area under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between classes.

3. Models Tested

a. Logistic Regression

A baseline linear model for binary classification.

b. Random Forest

An ensemble method known for its robustness and ability to handle feature interactions.

c. K-Nearest Neighbors (KNN)

A distance-based algorithm that benefits from normalized features.

d. Naive Bayes

A probabilistic model assuming feature independence, suitable for categorical data.

4. Results

The following table summarizes the performance of each model:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic regression	0.8958	0.7773	0.7441	0.7602	0.9532
Random forest	**0.9256**	**0.8959**	0.7522	**0.8177**	**0.9721**
KNN	0.8663	0.7315	0.6288	0.6762	0.9025
Naïve Bayes	0.7480	0.4678	**0.9768**	0.6326	0.9390

After selecting **Random Forest** as the best-performing model based on cross-validation results, it was further validated on the reserved **test data**. The test results are as follows:

Metric	Test performance
Accuracy	0.928
Precision	0.8969
Recall	0.7657
F1-Score	0.8261
ROC-AUC	0.9755

5. Key Insights

1. **Best Model**:
    - **Random Forest** achieved the highest overall performance with an accuracy of **92.69%**, an F1-score of **0.8223**, and an ROC-AUC of **0.9735**. It demonstrated a strong ability to balance precision and recall while distinguishing between loan approvals and rejections.
    - Random Forest continued to outperform other models when evaluated on test data. It achieved an **accuracy of 92.80%**, an **F1-score of 0.8261**, and an **ROC-AUC of 0.9755**.
2. **Validation Performance**:
    - The test performance metrics aligned well with the cross-validation results, indicating that the model generalized effectively to unseen data.

3. **Normalization vs. Standardization**:
    - The choice of normalization for KNN and standardization for other models was crucial for optimal performance.
    - Normalization improved KNN's ability to compute accurate distances between data points.
    - Standardization ensured that Logistic Regression, Random Forest, and Naive Bayes operated effectively without being skewed by varying feature scales.
4. **Trade-offs**:
    - **Logistic Regression** performed well as a baseline model with a high ROC-AUC of **0.9541**, making it a viable option for interpretability-focused use cases.
    - **KNN** showed moderate performance but struggled with recall, indicating difficulty in identifying all rejected loans.
    - **Naive Bayes** excelled in recall (**97.61%**), making it effective for identifying rejected loans but at the cost of precision and overall accuracy.
5. **Model Suitability**:
    - **Random Forest** is recommended for deployment due to its superior performance on both cross-validation and test data, making it the most reliable choice for predicting loan approvals.
    - **Logistic Regression** could be considered for scenarios requiring simpler models or interpretability.

6. Conclusion

The modeling phase demonstrated that ensemble methods like Random Forest outperform simpler models in this dataset. However, trade-offs exist between precision, recall, and interpretability, depending on the specific use case. The normalization and standardization steps ensured that all models operated effectively. The insights from EDA and model evaluation indicate that the features effectively capture the patterns underlying loan approval decisions, making the model robust and reliable for prediction.