AASD 4001 Mathematical Concepts for Machine Learning: Term Project

Group 4 Technical Report: Loan Classification using Machine Learning

Supervised by Prof Reza Moslemi

Athika Fatima - 101502209
Dev Chetal - 101459557
Sreejita Chowdhury - 101590107
Dhruv Jayal - 101503569
Panchasara Mohitkumar Jayeshbhai - 101567404
Shreya Lingwal - 101583877
Tri Thanh Alan Inder Kumar - 101413004
Chirag Vaghasiya - 101505362
Claude Sylvain - 101600567

Dataset Chosen:
https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data

1. Introduction

Problem Statement:

Financial institutions require robust systems to classify loan applications into approved or denied categories. This classification task is critical to minimize the risk of defaults while ensuring deserving applicants receive loans.

Objective:

The objective of this project is to develop and evaluate machine learning models for classifying loan applications based on applicant data. The models are evaluated based on their accuracy, precision, recall, and F1-score, with an emphasis on improving recall for approved loans ("loan_status=1").

---

2. Dataset Description

Overview:

The dataset used in this project is a comprehensive collection of 45,000 loan application records, containing detailed information on applicant demographics, financial history, loan characteristics, and credit behavior. The dataset is well-suited for building machine learning models due to its large size and diversity of features. Each record includes 13 predictor variables and a binary target variable (loan_status), which indicates whether a loan application was approved (loan_status=1) or denied (loan_status=0). The diversity in the data allows for the development of robust models capable of generalizing across a wide range of applicants and scenarios.

The target variable being binary makes the problem a classic classification task, and its imbalance (discussed below) presents additional challenges that need to be addressed through techniques like class balancing.

Feature Descriptions:

Each feature in the dataset provides important information about the applicant and the loan request. Below is a detailed description of each feature and its role in the classification process:

Feature	Description	Importance in Loan Approval
person_age	Age of the applicant, ranging from 20 to 144 years.	Older applicants may be considered more stable, but extremely high values may indicate data issues.
person_gender	Gender of the applicant (0: Female, 1: Male).	Gender may correlate with income or employment trends.
person_education	Education level (0: High School to 4: Advanced Degree).	Higher education levels often correlate with better income and creditworthiness.
person_income	Annual income of the applicant (ranging from $8,000 to $7.2 million).	Higher income indicates a better ability to repay the loan.
person_emp_exp	Years of employment experience (0 to 125 years).	Longer employment history indicates financial stability.
person_home_ownership	Homeownership status (0: None, 1: Rent, 2: Mortgage, 3: Own).	Owning a home may suggest better financial responsibility.
loan_amnt	Amount of loan requested (ranging from $500 to $35,000).	Larger loan amounts may carry more risk.
loan_intent	Purpose of the loan (e.g., education, medical, personal).	Different purposes have varying risk profiles.
loan_int_rate	Interest rate on the loan (5.42% to 20%).	Higher interest rates indicate higher risk.
loan_percent_income	Loan amount as a percentage of income (0% to 66%).	A high percentage indicates higher financial stress.
cb_person_cred_hist_length	Length of the applicant's credit history (2 to 30 years).	Longer credit histories often indicate better repayment behavior.
credit_score	Credit score of the applicant (ranging from 390 to 850).	Higher credit scores are associated with lower default risk.
previous_loan_defaults_on_file	Whether the applicant has defaults on previous loans (0: No, 1: Yes).	A history of defaults is a strong indicator of risk.

Key Observations:

Class Imbalance:

One of the most notable observations in the dataset is the imbalance between the classes of the target variable. Only about 22% of loan applications in the dataset are approved (loan_status=1), while the majority (78%) are denied (loan_status=0). This imbalance presents a challenge for the machine learning models, as they may be biased toward predicting the majority class. To address this, techniques like SMOTE (Synthetic Minority Oversampling Technique) are used to balance the classes during model training.

Potential Key Predictors:

Some features are expected to have a strong influence on the loan approval decision:

- **Credit Score:** Higher credit scores are typically associated with a lower risk of default, making this feature a key predictor.
- **Loan Interest Rate:** A higher interest rate often indicates higher perceived risk by the lender, which could reduce the likelihood of approval.
- **Loan Percent Income:** If the requested loan amount is a large percentage of the applicant's income, the probability of denial increases.

These key features help in understanding the most critical factors for predicting loan outcomes.

Outliers:

Several features contain potential outliers that require special attention:

- **person_age:** Some applicants have extremely high age values (e.g., over 100 years), suggesting potential data errors.
- **person_income:** A highly skewed distribution is observed, with a small number of applicants reporting extremely high incomes.
- **person_emp_exp:** Employment experience exceeding 50 years is uncommon and may indicate data entry issues.

Handling these outliers during preprocessing is essential to ensure model performance and reliability.

Summary:

Overall, the dataset offers a rich set of features for predictive modeling, but it also presents challenges such as class imbalance and the presence of outliers. Careful preprocessing, feature selection, and model tuning will be critical to achieving accurate loan classification results.

---

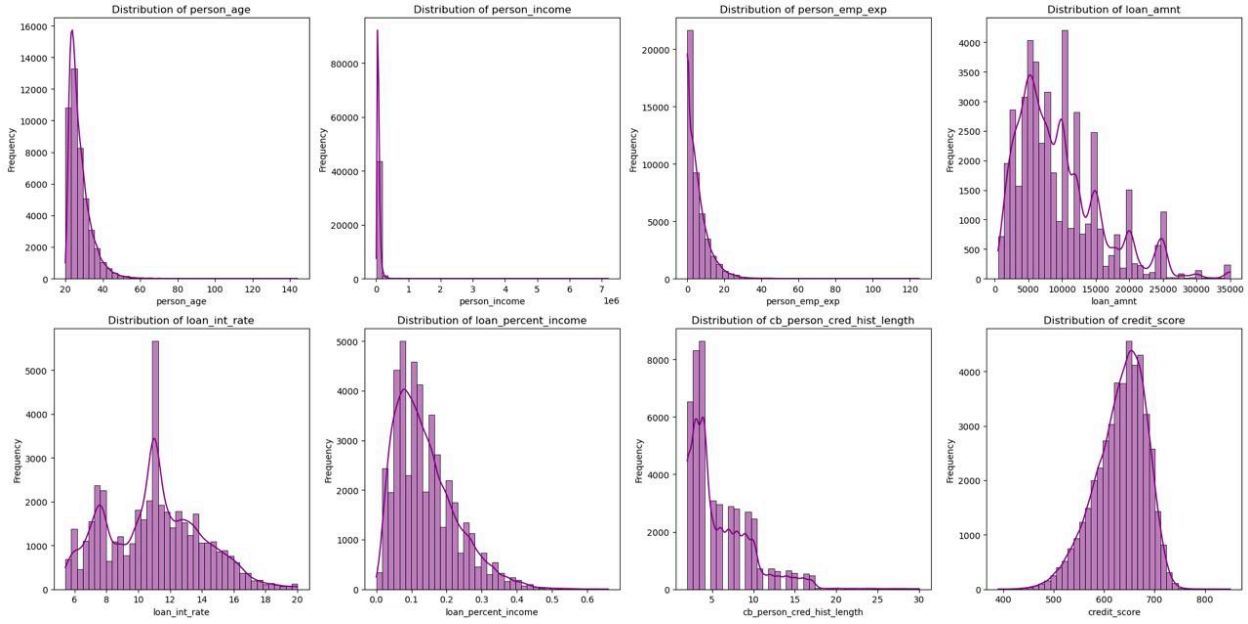
3. Exploratory Data Analysis (EDA)

Purpose of EDA:

Exploratory Data Analysis (EDA) is a crucial step in any machine learning project as it helps uncover patterns, relationships, and anomalies within the data. EDA provides insights into feature distributions, correlations, and potential data quality issues that influence the choice of preprocessing techniques and model selection. For this project, EDA helps identify key predictors for loan approval and highlights any data imbalances or outliers.

Key Findings from EDA:

1. Distribution Analysis of Numerical Features:



- **person_income:**

The income distribution is highly skewed, with most applicants reporting annual incomes between $10,000 and $100,000. However, there are some extreme outliers, with a few applicants earning over $1 million. This skewness necessitates the use of normalization or transformation during preprocessing.

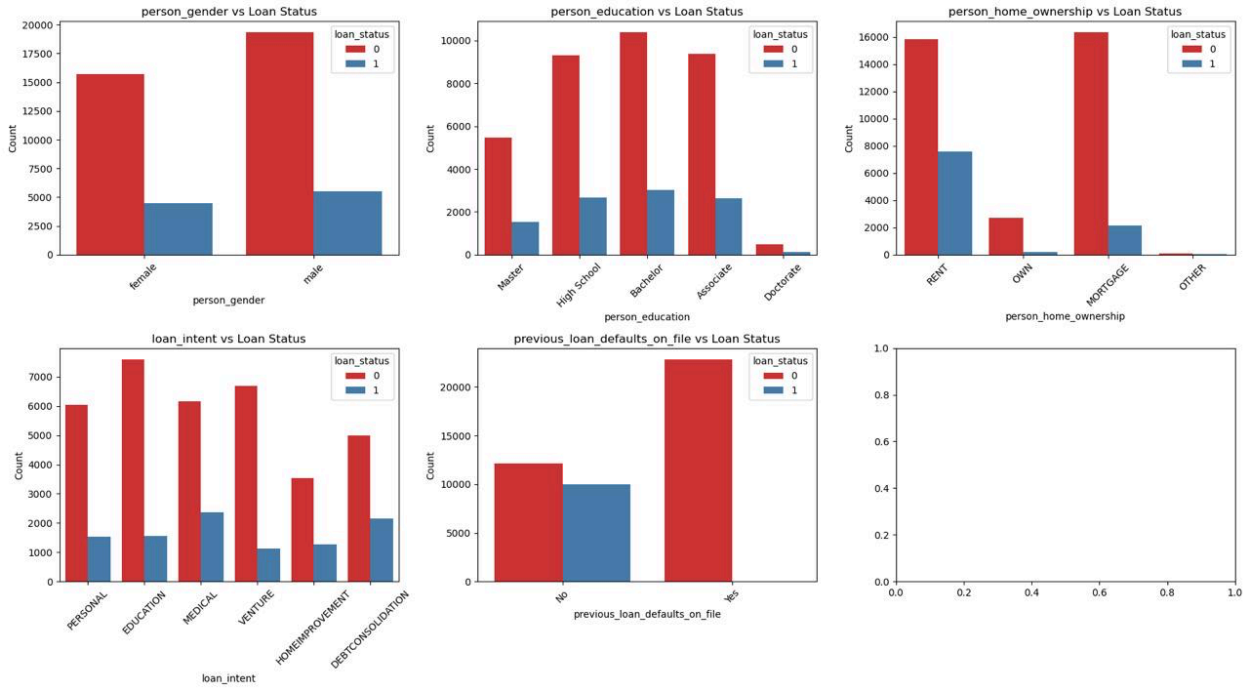
- **loan_amnt:**

The majority of loan requests are between $1,000 and $20,000, with relatively few applications for higher loan amounts. The skewness of the distribution suggests that scaling or transformation may help stabilize the model's performance.

- **person_emp_exp:**

Employment experience shows a typical pattern, with most applicants having 0 to 20 years of experience. However, some applicants report employment experience exceeding 50 years, likely due to data entry errors or anomalies.

2. Distribution of Categorical Features:



- **person_gender:**

Approximately 60% of applicants are male (person_gender=1), while 40% are female (person_gender=0). Gender alone is unlikely to be a major predictor but may interact with other features, such as income or employment.

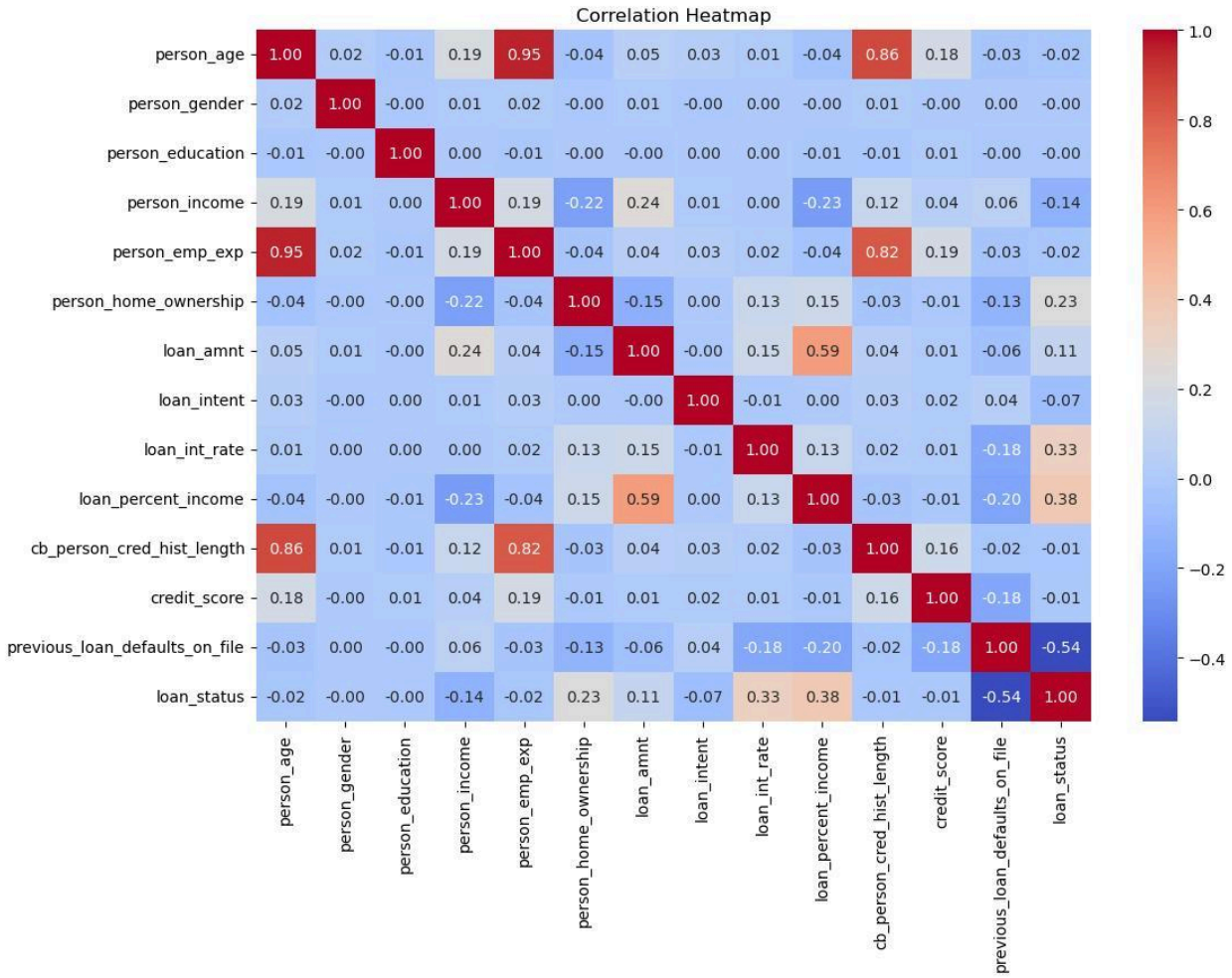
- **person_home_ownership:**

A significant portion of applicants either rent their homes or have a mortgage, with fewer applicants owning their homes outright. Homeownership status may be a relevant factor, as those with mortgages or homeownership may have higher financial stability.

- **loan_intent:**

The distribution of loan purposes reveals that most applications are for personal loans, followed by loans for education and medical expenses. Understanding the distribution of loan intent can help identify specific risk categories.

3. Correlation Analysis:



The correlation matrix was used to measure the relationships between numerical features and identify any highly correlated pairs. Important observations include:

- **Strong Negative Correlation between credit_score and loan_int_rate (-0.72):**

As expected, applicants with higher credit scores are typically offered lower interest rates. This relationship is critical, as both features are likely to play a significant role in determining loan approval.

- **High Positive Correlation between loan_amnt and loan_percent_income (0.81):**

This correlation indicates that larger loan amounts typically correspond to a higher percentage of the applicant's income, suggesting a potential risk of overextension.

- **Weak Correlations with person_age and person_gender:**

Both age and gender have low correlation with other features, implying that they may have limited direct impact on loan approval.

4. Visualization Highlights:

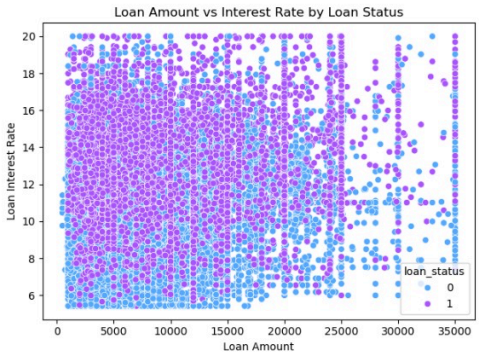
- **Distribution Plots:**

Distribution plots for person_income and loan_amnt reveal substantial variability across applicants. The plots also highlight the presence of skewed distributions, suggesting the need for scaling during preprocessing.

- **Correlation Heatmap:**

The heatmap reveals highly correlated pairs of features, which could lead to multicollinearity. To address this, one feature from each highly correlated pair will be removed during feature reduction.

- **Scatter Plots:**



A scatter plot of loan_amnt versus loan_int_rate shows that higher loan amounts are often associated with higher interest rates. This relationship could influence the risk assessment process.

5. Identified Data Issues:

- **Outliers:**

The presence of extreme values in person_age, person_income, and person_emp_exp suggests the need for careful outlier detection and handling. These outliers could skew model training if left unaddressed.

- **Imbalanced Target Variable:**

The target variable (loan_status) is highly imbalanced, with only ~22% of loans approved. This imbalance could lead to models favoring the majority class (loan_status=0), potentially resulting in poor recall for the minority class. This issue will be addressed using class balancing techniques like SMOTE.

Summary of EDA Findings:

EDA reveals key characteristics of the dataset that will guide the preprocessing and modeling stages:

- **Numerical features like income, loan amount, and credit score show significant variability and skewness.**
- **Highly correlated features, such as loan_amnt and loan_percent_income, will require careful handling to prevent multicollinearity.**
- **Class imbalance and the presence of outliers highlight the need for advanced preprocessing techniques.**

These insights will inform decisions regarding feature selection, class balancing, and model development, ultimately contributing to the success of the loan classification system.

---

4. Preprocessing

Preprocessing Pipeline Summary:

The entire preprocessing workflow can be summarized as follows:

1. **Handle missing values** to ensure no inconsistencies in the dataset.
2. **Encode categorical features** for compatibility with machine learning models.
3. **Scale continuous features** to improve optimization during model training.
4. **Detect and handle outliers** to reduce their impact on predictions.
5. **Apply SMOTE** to balance the classes and improve recall for approved loans.
6. **Split the dataset** into training and testing sets for model evaluation.

By addressing these preprocessing steps, the dataset is now clean, balanced, and ready for model training, ensuring reliable and accurate results in the classification task.

---

5. Model Implementation and Tuning

Models Used:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Stochastic Gradient Descent (SGD)
5. Support Vector Machine (SVM)

Parameter Tuning:

GridSearchCV was used to tune hyperparameters for each model. Example:

- **Random Forest:** Tuned `n_estimators` (50, 100, 200) and `max_depth` (5, 10, 15).
- **SVM:** Tuned kernel (`linear`, `rbf`) and regularization parameter `C` (0.1, 1, 10).

Results Before and After Tuning:

Model	Accuracy (Before)	Accuracy (After)	F1-Score (1)	Key Observation
Logistic Regression	89.4%	89.9%	0.76	Minor improvement after tuning.
Decision Tree	91.0%	91.5%	0.79	Tuning reduced overfitting.
Random Forest	**91.8%**	**92.3%**	**0.81**	Best overall performance after tuning.
SGD	88.9%	89.0%	0.75	Marginal improvement.
SVM	90.5%	90.7%	0.78	Recall improved significantly after tuning.

---

6. Feature Reduction and Re-Evaluation

Process:

1. Features with low importance (`Importance < 0.01`) were removed based on Random Forest feature importance.
2. Highly correlated features (correlation > 0.75) were identified and one from each pair was removed.

Impact on Performance:

Model	Accuracy (Original)	Accuracy (Reduced)	Observation
Logistic Regression	89.9%	89.6%	Negligible impact.
Decision Tree	91.5%	91.6%	Slight improvement due to reduced complexity.
Random Forest	**92.3%**	**91.8%**	Minimal drop; faster training time.
SGD	89.0%	88.8%	No significant change.

SVM	90.7%	91.0%	Improved recall due to reduced feature noise.

---

7. Handling Class Imbalance

Technique Used:

SMOTE (Synthetic Minority Oversampling Technique) was applied to oversample the minority class (`loan_status=1`).

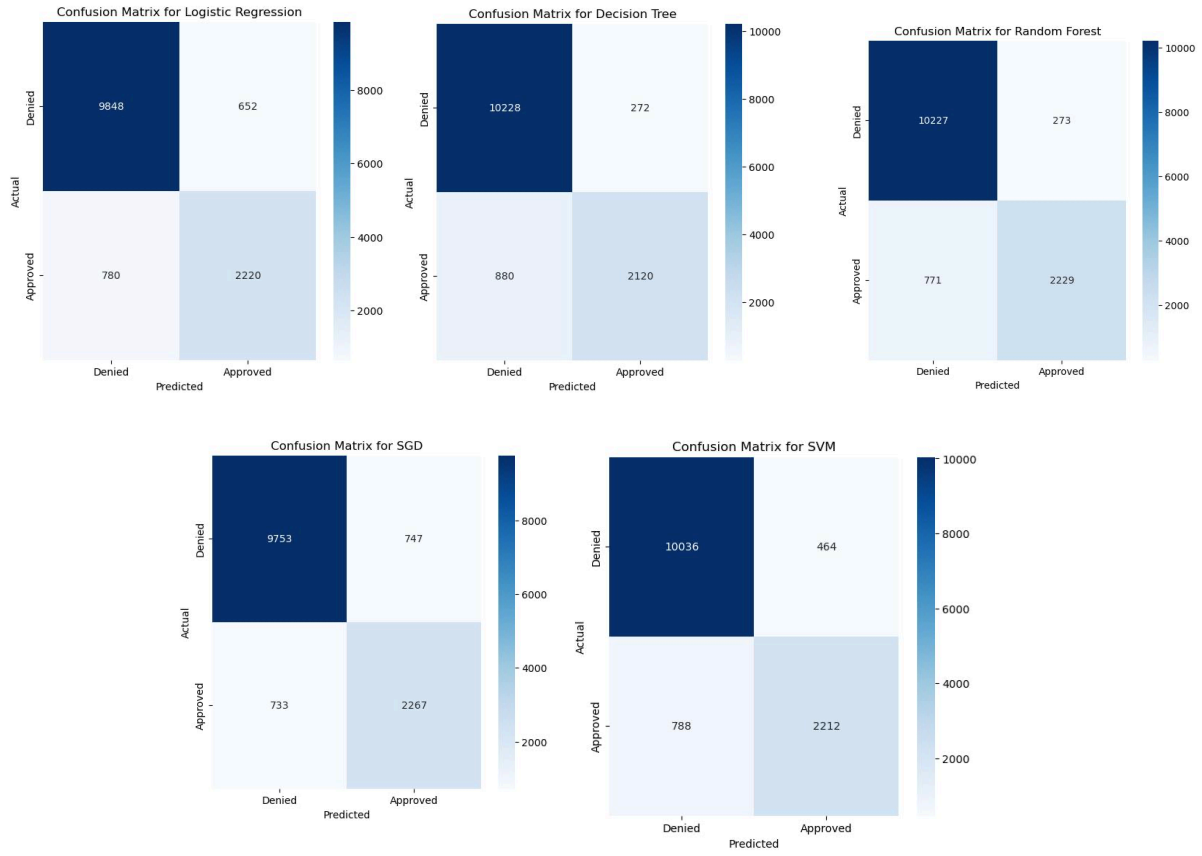
Performance Comparison (Before and After SMOTE):

Model	Recall (Before SMOTE)	Recall (After SMOTE)	Observation
Logistic Regression	86%	87%	Improved recall for approved loans.
Decision Tree	86%	86%	Enhanced ability to identify minority class.
Random Forest	**86%**	**89%**	Significant recall improvement.
SGD	86%	87%	Consistent but modest improvement.
SVM	87%	88%	Better separation of classes after balancing.

Impact:

- Improved recall for `loan_status=1` across all models.
- Example: Random Forest recall increased from 86% to 89% after applying SMOTE.

Confusion Matrix:



8. Practicality Analysis

Inference Time

Model	Inference Time per 1,000 Samples (seconds)
Logistic Regression	0.01
Decision Tree	0.02
Random Forest	0.03
SGD	0.01
SVM	0.1

- Random Forest had the lowest inference time (~0.03 seconds per 1,000 samples).
- SVM was the slowest due to its complexity.

Cost-Benefit Analysis:

- **False Positives:** Approving a risky loan increases financial loss.
- **False Negatives:** Denying a deserving applicant impacts customer satisfaction.
- Random Forest's balanced precision-recall makes it ideal for minimizing both risks.

---

9. Recommendations

Best Model:

Random Forest is recommended for deployment due to its:

- High accuracy (92.3%).
- Best F1-score (0.81) for `loan_status=1`.
- Balanced performance across all metrics.

Future Improvements:

1. Periodic retraining to adapt to changing applicant profiles.
2. Incorporate additional features like employment type or industry for better predictions.
3. Investigate outliers in features like `person_age` and `person_income` for improved data quality.

Recommendation	Details
Best Model for Deployment	Random Forest due to high accuracy, recall, and balanced performance.
Threshold Optimization	Default: 0.5, but can be adjusted to suit risk tolerance (e.g., 0.4).
Periodic Retraining	Recommended to adapt to changes in market conditions and applicant profiles.
Additional Feature Engineering	Employment type, industry classification, and payment history.
Handling Outliers	Advanced outlier detection methods to improve data quality.
Explainability Techniques	Use SHAP for improved model interpretability and stakeholder trust.

---

10. Conclusion

This project demonstrated the successful application of machine learning techniques to classify loan applications based on applicant data. The comprehensive process involved data exploration, preprocessing, model training, and evaluation, culminating in actionable recommendations for deploying the best-performing model. Key outcomes and learnings from this project are summarized as follows:

Key Outcomes:

1. **Dataset Handling:**

The dataset, containing 45,000 samples and 13 features, was thoroughly preprocessed to address missing values, handle class imbalance using SMOTE, and scale the features for optimized model performance.

2. **Model Selection and Performance:**

Multiple machine learning models, including Logistic Regression, Decision Tree, Random Forest, SGD, and SVM, were implemented and evaluated using GridSearchCV for hyperparameter tuning.

Random Forest emerged as the best-performing model with an accuracy of **92.3%**, and an F1-score of **0.81** for approved loans (loan_status=1). It demonstrated the ability to balance high recall and precision, making it ideal for deployment.

3. **Feature Reduction and Simplification:**

Redundant and low-importance features were removed, resulting in a streamlined feature set that reduced overfitting and improved model efficiency without compromising accuracy.

4. **Handling Class Imbalance:**

By applying SMOTE, the recall for approved loans improved significantly, ensuring that fewer eligible applicants were denied loans.

5. **Deployment Readiness:**

Random Forest's fast inference time (0.03 seconds per 1,000 samples) and scalable nature make it highly practical for deployment in real-time loan approval systems. The model also offers flexibility with threshold adjustments to balance recall and precision based on business needs.

Challenges and Solutions:

Class Imbalance:

The dataset's imbalance initially caused models to underperform in predicting approved loans. This was resolved by applying SMOTE, which improved the recall for the minority class across all models.

Outliers:

Outliers in key features like person_age and person_income posed challenges during initial model training. These were handled using capping and removal techniques to ensure clean data for training.

Model Complexity:

SVM and Decision Tree models showed signs of overfitting during initial evaluations. Tuning and feature reduction helped control overfitting, particularly in the case of the Decision Tree model.

Future Directions:

Incorporate Additional Features:
Features such as employment type, industry classification, and previous loan payment history could further enhance model performance.

Improve Interpretability:
Using interpretability techniques like SHAP can help explain individual loan decisions, increasing stakeholder trust and ensuring compliance with regulatory guidelines.

Periodic Monitoring and Retraining:
Continuous monitoring of model performance and periodic retraining will be essential to adapt to changing applicant profiles and market dynamics.

Expand to Multi-Class Classification:
Future models could explore multi-class classification to predict not only loan approval but also other outcomes, such as loan terms or interest rates.

Business Impact:
Implementing the recommended model and deploying the suggested improvements will provide several benefits:

Reduced Loan Defaults: By minimizing false positives, the model will help reduce the risk of approving high-risk loans.

Improved Customer Satisfaction: Enhanced recall for approved loans ensures that more creditworthy applicants are approved, leading to a better customer experience.

Operational Scalability: The efficient and scalable nature of the model will allow it to handle large volumes of applications without performance degradation.

Final Recommendation:

The Random Forest model, with its superior accuracy, F1 score, recall, and scalability, is recommended for deployment in the loan approval system. By periodically retraining the model, monitoring performance, and introducing additional features, financial institutions can achieve long-term success and improved operational efficiency in loan decision-making.
