Revolutionizing Loan Decisions with Machine Learning

Presented by: Group 4

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

Table of Contents

Challenges
in Traditional Loan Classification

- High Volume of Applications
- Imbalanced Dataset Challenge
- Need for Scalability
- Risk of Defaults
- Complexity of Relationships
- Deep Learning

How AI Can Help?

The primary goal of leveraging AI in loan classification is to revolutionize the decision-making process by making it faster, more accurate, and less biased.

01

Automate and enhance decision-making

02

Utilize historical data for predictive analysis

03

Reduce bias and improve accuracy

04

Increase loan approval speed and efficiency

Dataset Features

Detailed Overview of Dataset Characteristics

1 Dataset Size

The dataset comprises a total of 45,000 samples, providing a robust base for analysis.

2 Target Variable

The primary target variable is loan_status, indicating whether a loan is approved (1) or denied (0)

3 Demographic Features

Demographics include person_age, person_gender, and person_education, crucial for understanding borrower profiles.

4 Financial Stability Indicators

Financial stability is assessed through person_income, person_home_ownership, and credit_score.

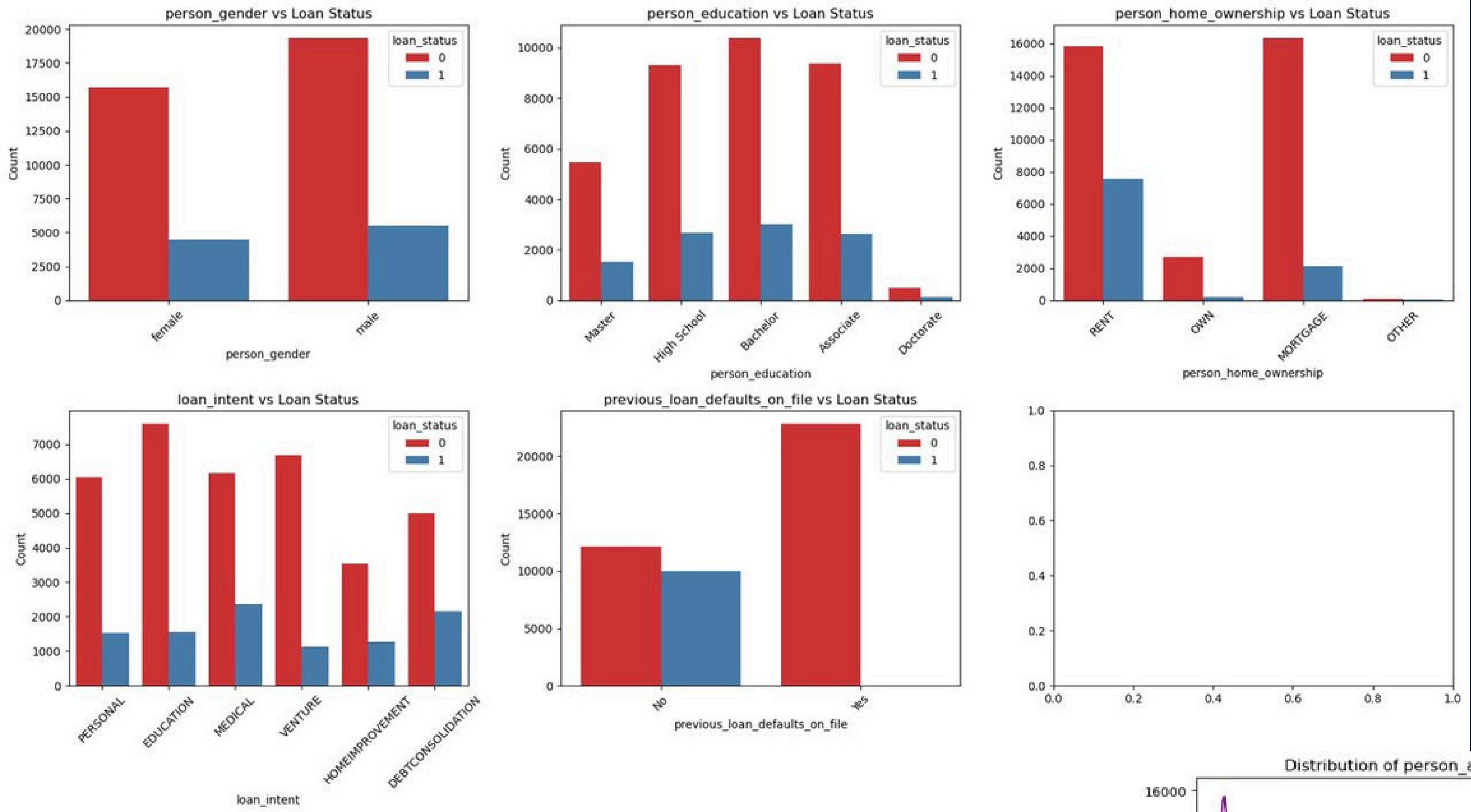
5 Behavioral History

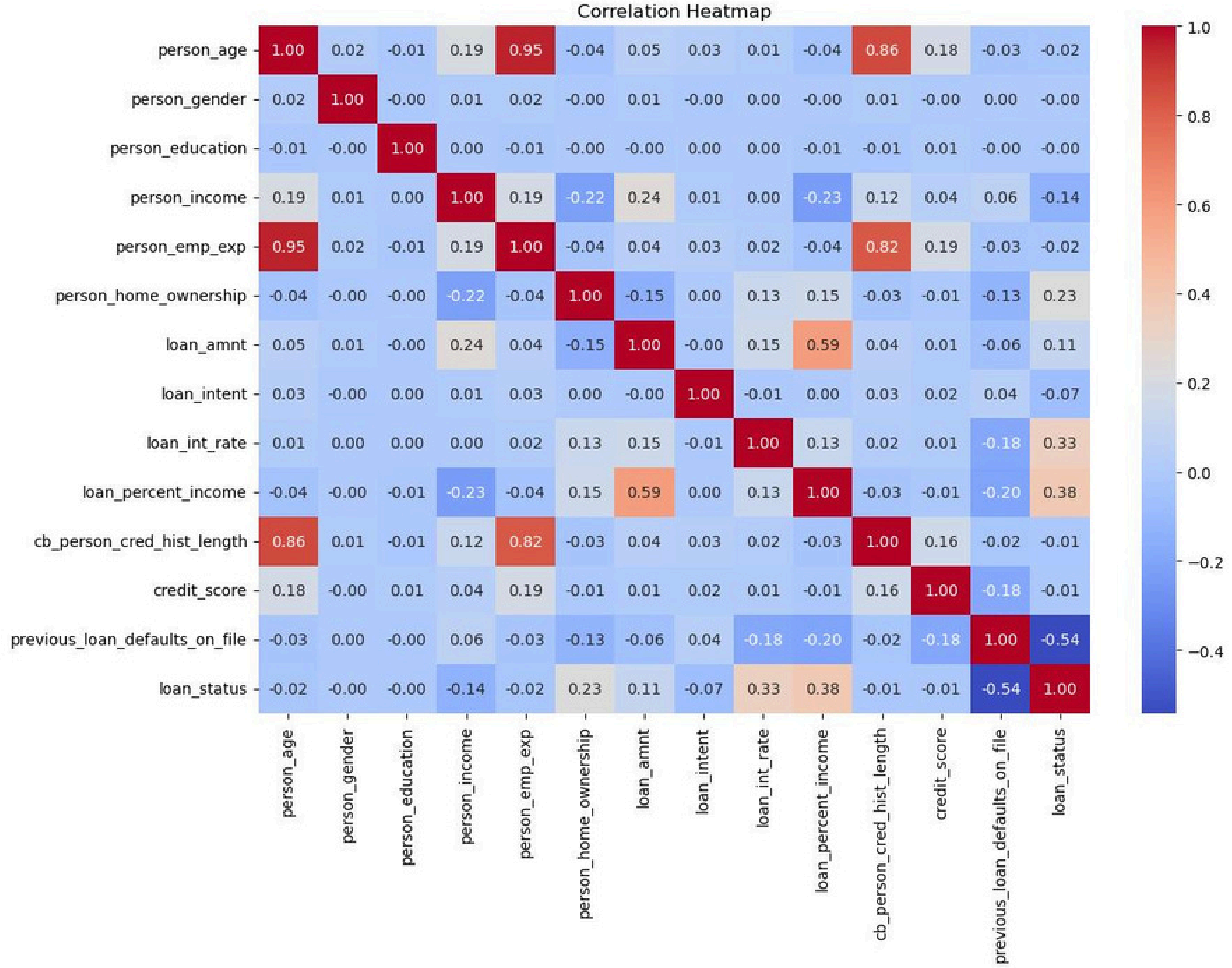
Behavioral features include cb_person_cred_hist_length and previous_loan_defaults_on_file to assess credit behavior.

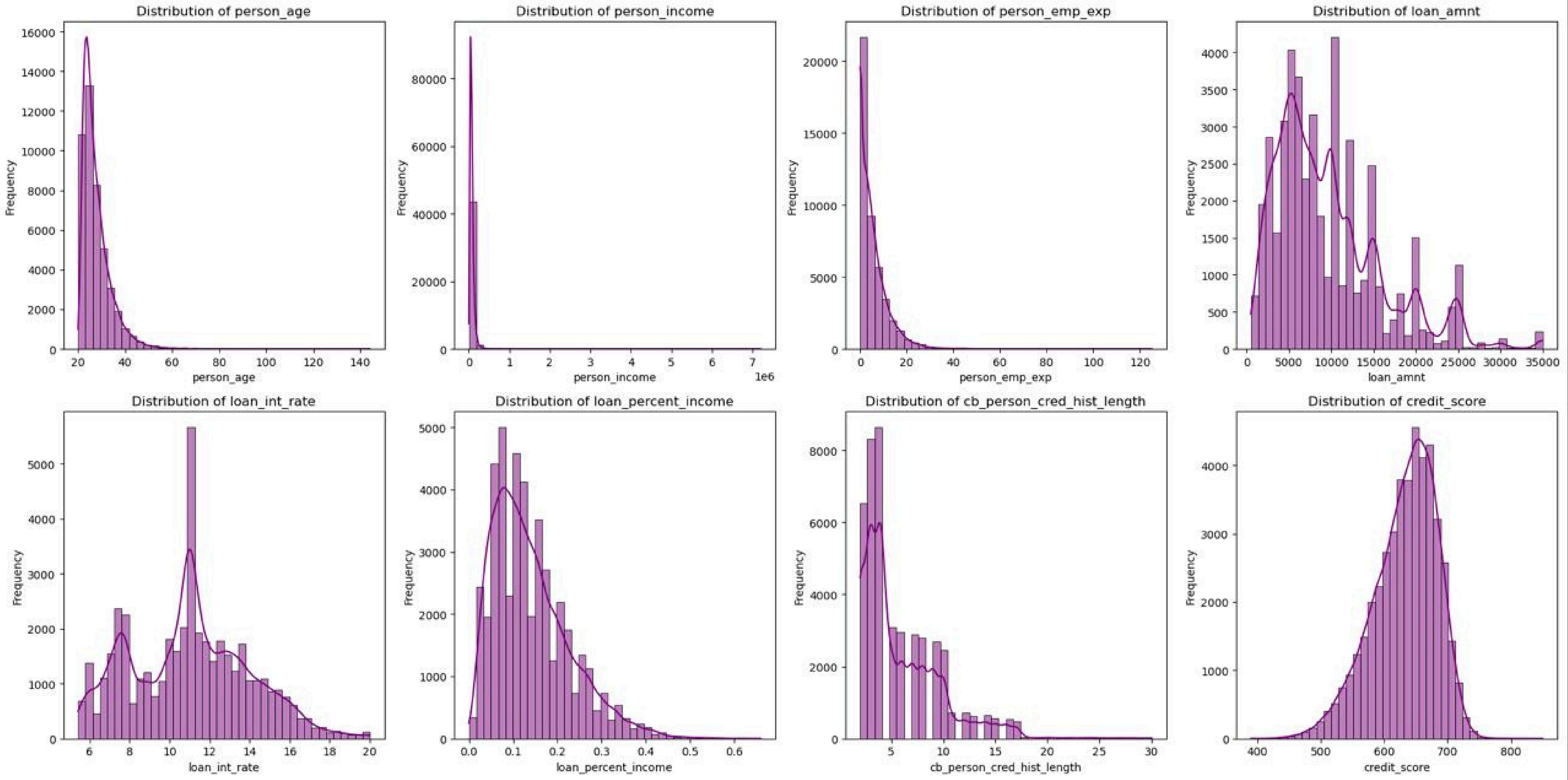
6 Loan Details

Loan specifics such as loan_amnt, loan_int_rate, and loan_percent_income are included to evaluate loan conditions.

Graphical Representation







Key Observations
on Loan Approvals

Imbalanced Dataset
Only 22% of loans were approved, indicating a significant imbalance.

Strong Predictive Features
Certain features significantly influence loan approval rates.

Financial Burden Indicator
Higher loan percent of income correlates with greater financial burden.

Credit Score Correlation
Higher credit scores are closely linked to increased likelihood of loan approval.

Interest Rate Impact
Loan interest rate has an inverse relationship with loan status (-0.72).

Outlier Detection
Outliers found in person age, income, and employment experience.

Correlation Insights
High correlation (0.81) identified between loan amount and loan percent of income.

Decoding Machine Learning Approach

1 Logistic Regression

- Accuracy Achieved: 89.9%
- Key Strength: High interpretability and ease of implementation.
- Challenge: Struggles to capture non-linear relationships in complex datasets.

2 Decision Trees

- Accuracy Achieved: 91.5%
- Key Strength: Ability to handle non-linear relationships & provide intuitive visualizations.
- Challenge: Prone to overfitting without parameter tuning.

3 Random Forests

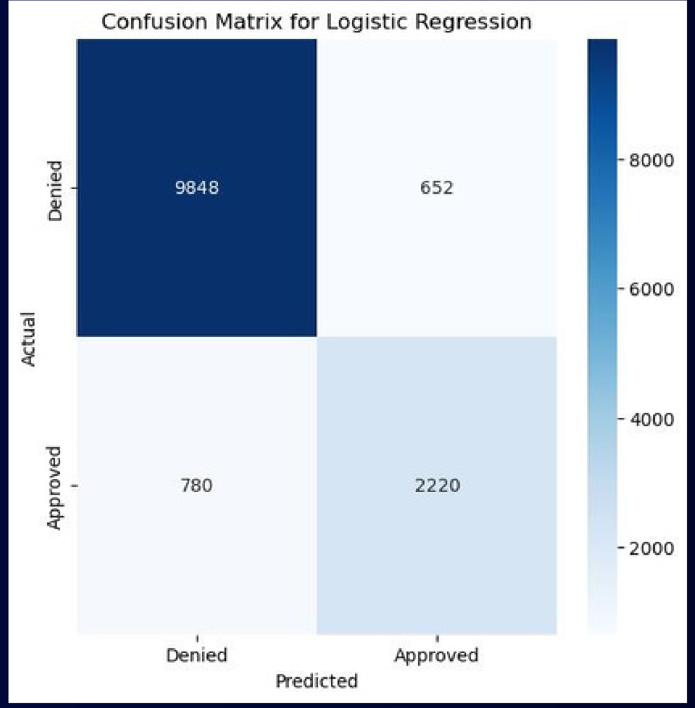
- Accuracy Achieved: 92.3% (Best-performing model in this project).
- Key Strength: High accuracy and resistance to overfitting.
- Feature Importance Insight: credit_score was the most influential feature, with an importance score of 0.24.

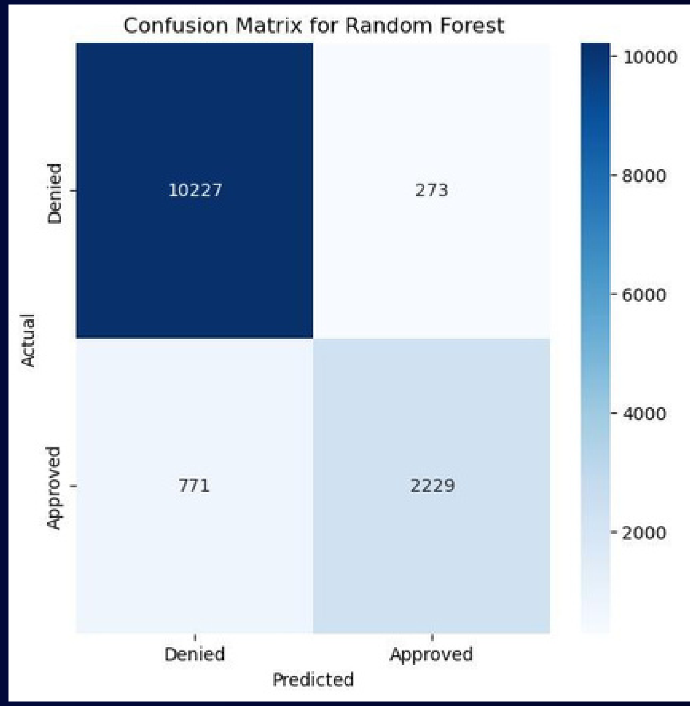
4 Stochastic Gradient Descent (SGD)

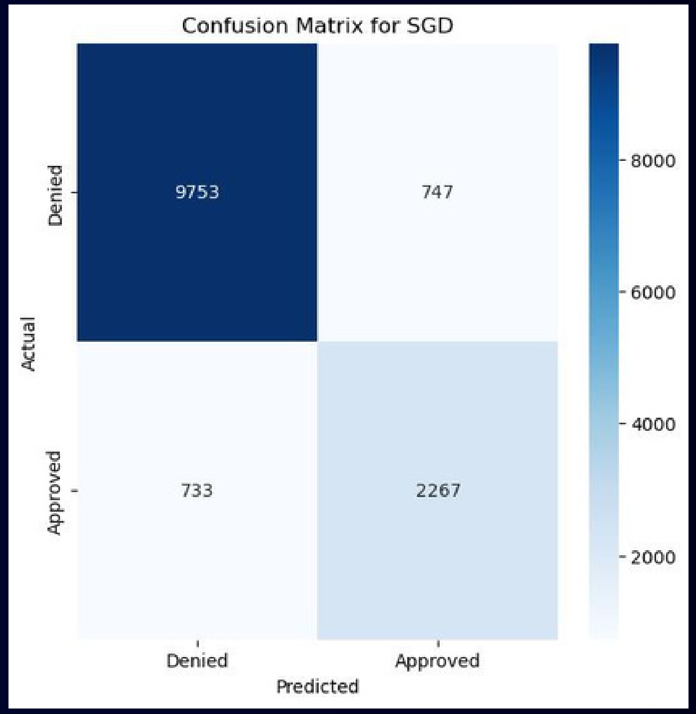
- Accuracy Achieved: 89.0%
- Key Strength: Fast training on high-dimensional data.
- Challenge: Sensitive to hyperparameter tuning (e.g., learning rate).

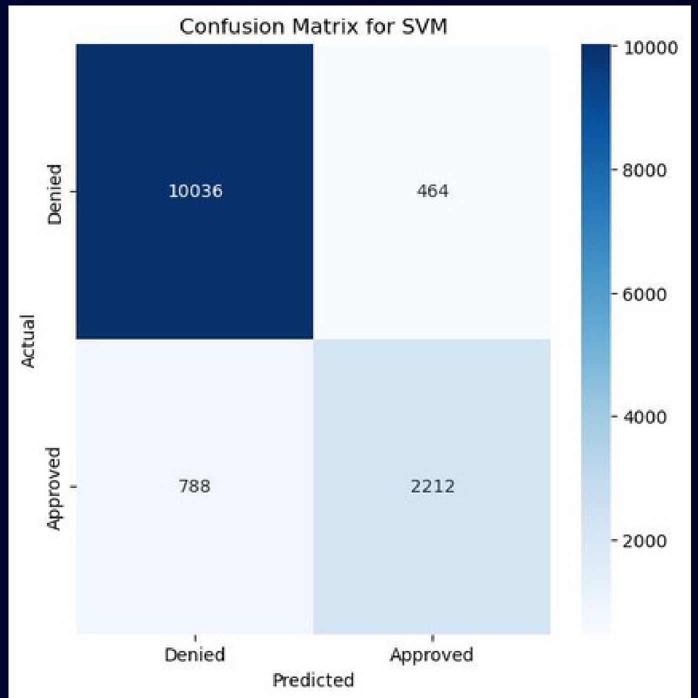
5 Support Vector Machine (SVM)

- Accuracy Achieved: 90.7%
- Key Strength: Effective in datasets with complex boundaries.
- Challenge: Computationally expensive for large datasets, with slower inference times (˜0.3 seconds per 1,000 samples).





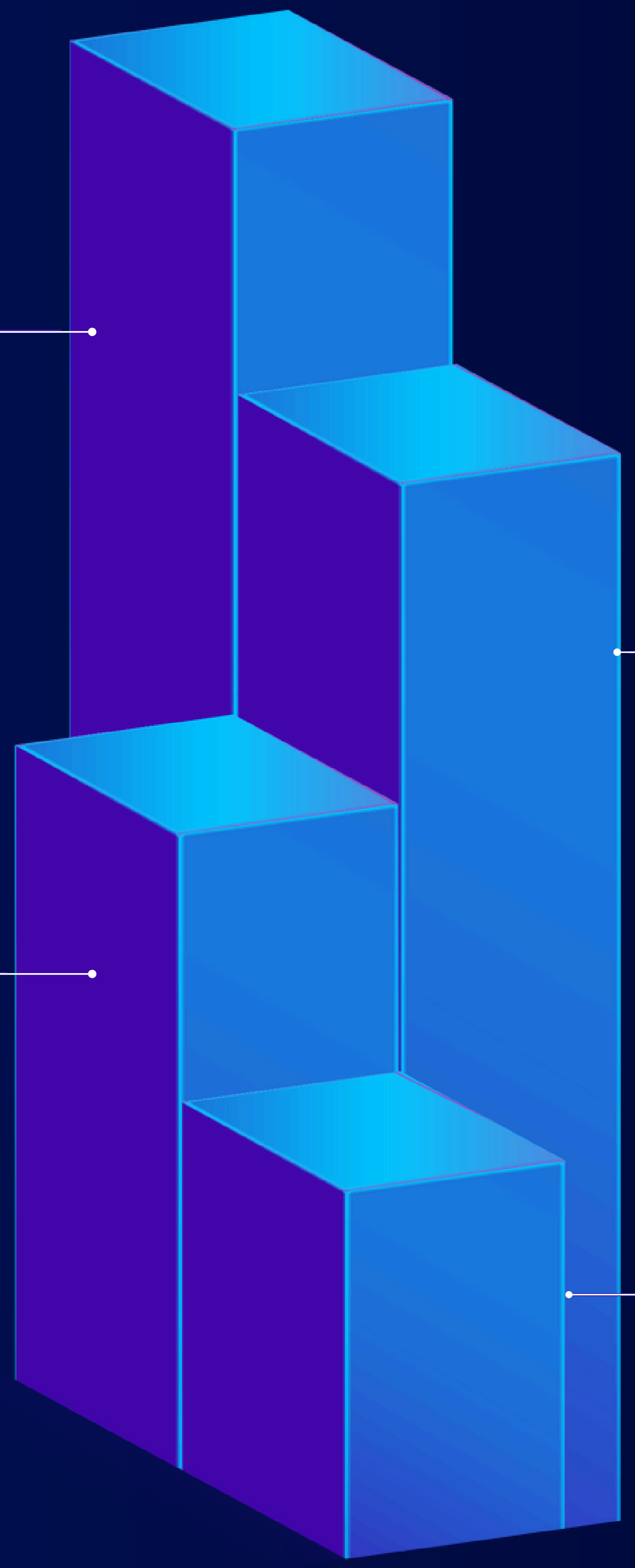




Preprocessing Steps for AI Models

Handling Categorical Data

Encoded features like person_home_ownership and loan_intent using LabelEncoder for numerical processing.



Normalization of Numerical Features

Scaled person_income and loan_amnt with StandardScaler for uniform feature scaling.

Class Imbalance Handling

Addressed dataset imbalance (22% loans approved) using SMOTE to generate synthetic minority samples.

Data Splitting

Divided dataset into 70% for training and 30% for testing, ensuring stratification by loan_status.

Hyperparameter Tuning Techniques

Enhancing Model Performance through Tuning

Objective of Hyperparameter Tuning

Optimize model performance by fine-tuning various parameters.

GridSearchCV Methodology

Utilizes 5-fold cross-validation to ensure robust model evaluation.

Tuning Parameters for Models

Different models have specific parameters to tune for optimal performance.

Logistic Regression Tuning

Key parameter: C (regularization strength) to prevent overfitting.

Decision Tree Tuning

Parameters tuned include max_depth and min_samples_split for tree complexity.

Random Forest Optimization

Focus on n_estimators and max_depth for better ensemble performance.

SVM Parameter Tuning

Tune the kernel type (linear, rbf) and C value for regularization.

Positive Outcomes of Tuning

Accuracy, recall, and F1-scores improved across all models post-tuning.

Best Performing Model

Random Forest achieved the highest accuracy at 92.3% and F1-score of 0.81.

Feature Reduction, Impact Analysis & SMOTE

1

Feature Selection Process

Removed features with low importance (< 0.01) and high correlation (> 0.75).

2

Random Forest Accuracy Change

Accuracy slightly decreased from 92.3% to 91.8% after feature reduction.

3

Improvements in Decision Tree

Decision Tree and SVM models improved due to reduced noise from feature reduction.

4

Interpretability Gains

Feature reduction led to improved interpretability and faster model inference speed.

5

Model Performance Comparison

Logistic Regression accuracy: 89.9%, Decision Tree: 91.5%, SVM: 90.7%.

6

Best Performing Model

Random Forest remained the best model with highest accuracy and F1-score.

1

Issue of Class Imbalance

The minority class (loan_status=1) constituted only 22% of the dataset, leading to potential model bias.

2

Implementing SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) was utilized to generate synthetic samples for the minority class, enhancing dataset balance.

3

Significant Impact on Recall

Post-implementation, recall for loan_status=1 saw a notable increase across all models, indicating improved prediction performance.

4

Example of Recall Improvement

For instance, Random Forest model recall improved from 74% in the imbalanced dataset to 79% after balancing, showcasing SMOTE's effectiveness.

Insights
on Solution Practicality

Deployment and Practicality
Understanding how models perform in real-world scenarios is crucial.

Cost-Benefit Analysis
Analyzing costs linked to false positives and negatives helps in decision-making.

False Negatives
Missing out on good applicants can lead to significant missed opportunities.

Inference Time
Random Forest operates at 0.03 seconds for 1,000 samples, SVM is slower.

False Positives
Approving bad loans results in financial losses, impacting overall profitability.

Model Recommendation
Random Forest is recommended for its high accuracy and quick inference time.

Challenges and Future Enchancements

Exploring the obstacles and potential improvements in loan assessments

1. **Outliers in Key Features**

   Identifying outliers in features like person_age and person_income can skew results.

2. **Limited Applicant Information**

   Insufficient details about the applicant's profession or industry can hinder accurate assessments.

3. **Incorporating Additional Features**

   Future models should include employment type and geographic data for better prediction accuracy.

4. **Utilizing Explainability Tools**

   Implementing SHAP or LIME can enhance understanding of model predictions and decisions.

5. **Periodic Retraining of Models**

   Regularly retraining models will ensure adaptability to evolving loan applicant profiles and trends.

Key Conclusions and Insights

Overview of Machine Learning Implementation Results

1. **Implementation of ML Models**
   Successfully implemented machine learning models for efficient loan classification.

2. **Performance of Random Forest**
   Random Forest outperformed other models with an accuracy of 92.3% and an F1-score of 0.81.

3. **Recommendations for Future Use**
   Provided recommendations for deployment and future improvements in the loan classification process.

THANK YOU!

ANY and ALL questions are welcome! :)