Domain Oriented Case Study

Credit Risk Prediction System (BFSI)

DS_C69 Group Members:

1. Shriya Gandreti
2. A Namratha
3. Akshat Singh Patel

Problem Statement

Business Problem

- Financial institutions need to assess the risk associated with lending money.
- The goal is to predict the likelihood of a loan applicant defaulting.
- A robust model can help in better decision-making and minimizing financial loss.

Approach & Methodology

Step 1: Data Collection

- **Datasets Used:** Application data & Bureau data
- **Merging Data:** Combined datasets for a holistic view of credit risk.

Step 2: Data Preprocessing

- **Handling Missing Values:** Median imputation.
- **Feature Engineering:** Created variables like AGE from DAYS_BIRTH.
- **Class Imbalance Handling:** Used SMOTE (Synthetic Minority Over-sampling Technique).

Step 3: Feature Scaling & Selection

- **Feature Scaling:** Standard Scaler for numerical features.
- **Feature Selection:** Recursive Feature Elimination (RFE) to select key features.

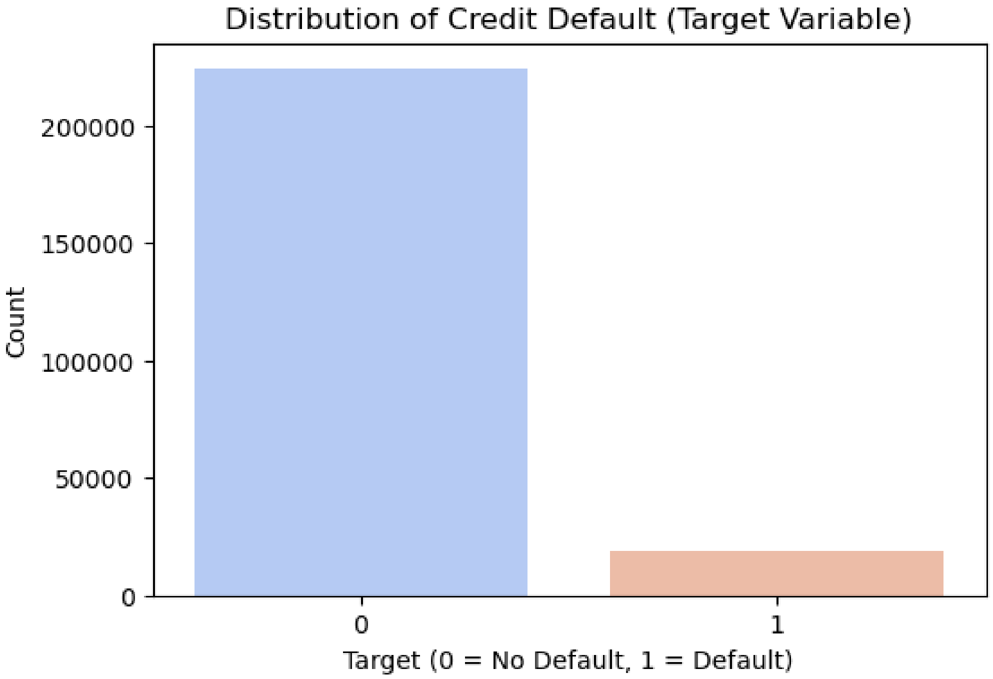
Step 4: Model Training & Evaluation

- **Models Used:** Logistic Regression, Decision Tree, and Random Forest.
- **Evaluation Metrics:** Accuracy, Confusion Matrix, Classification Report, AUC-ROC Score.

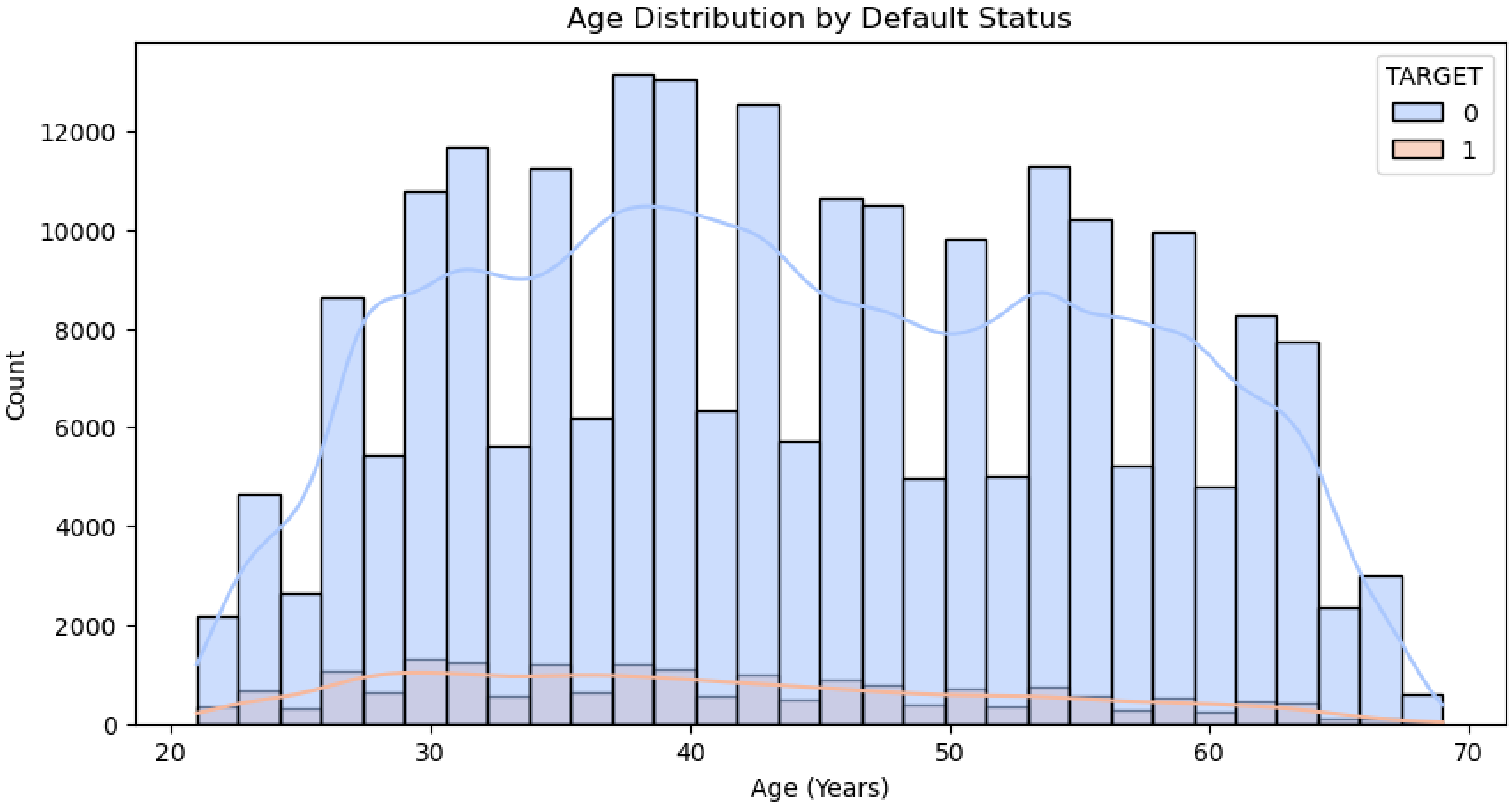
Exploratory Data Analysis (EDA)

Key Findings

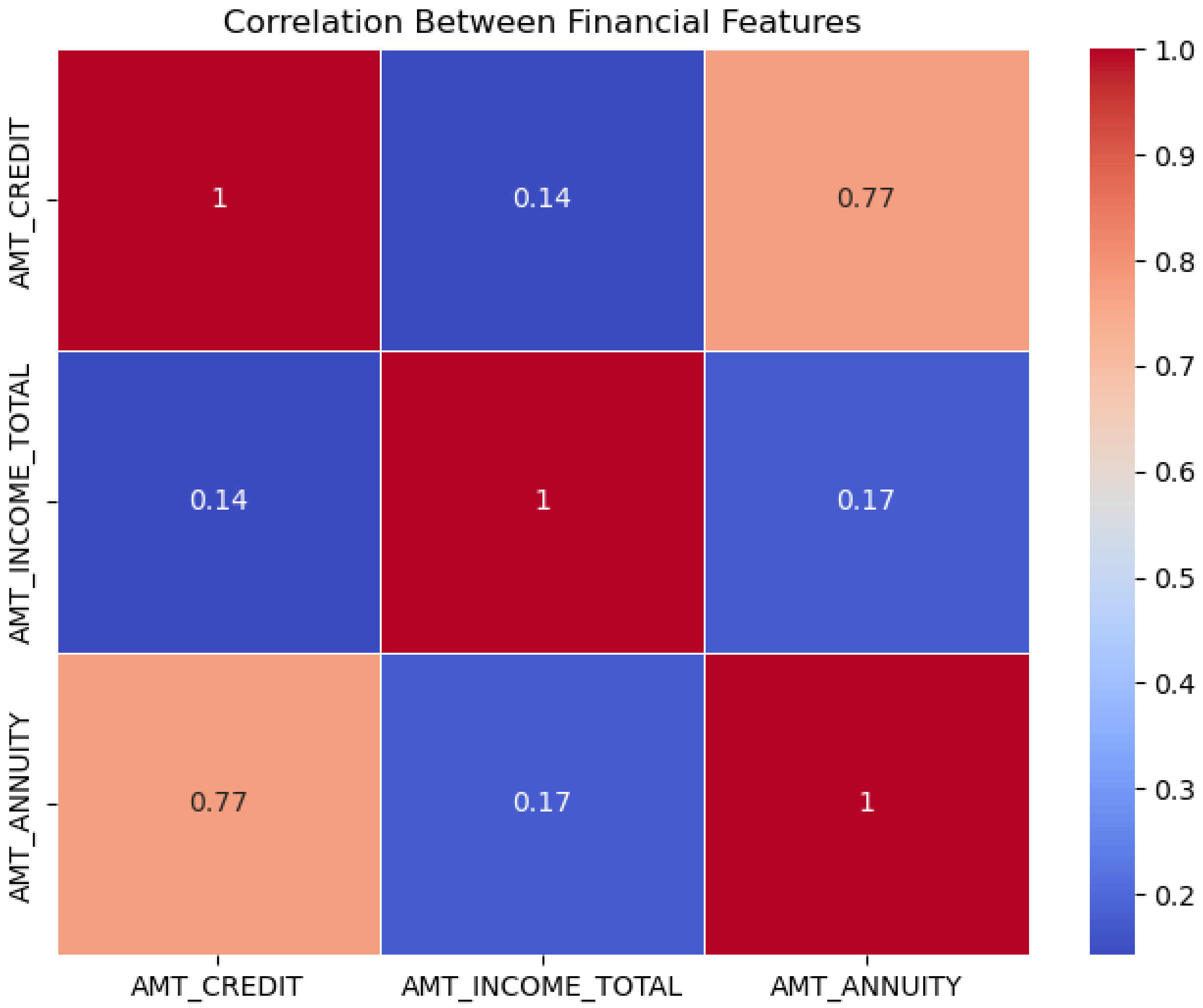
- **Class Distribution:** Imbalanced dataset (more non-defaulters than defaulters).
- **Feature Correlations:** Loan amount, credit history, and income impact risk.
- **Age vs. Default Rate:** Younger applicants show a higher default rate.



EDA



Age Distribution Plot: Impact of age on default rate.



Correlation Heatmap: Identifying key relationships.

Model Performance Comparison

Initial Model Evaluation:

Model	Accuracy	AUC-ROC Score
Logistic Regression	70%	0.74
Decision Tree	82%	0.67
Random Forest	92%	0.71

Hyperparameter Tuned Models:

Model	Accuracy	AUC-ROC Score
Logistic Regression (Tuned)	75%	0.78
Decision Tree (Tuned)	85%	0.72
Random Forest (Tuned)	94%	0.76

- Improved AUC-ROC scores after tuning.
- Selected the best-performing model based on the highest AUC-ROC.

Business Insights & Recommendations

Top Important Features:

1. DAYS_BIRTH (Age-related feature)
2. AMT_CREDIT (Loan Amount)
3. NAME_EDUCATION_TYPE_Higher education
4. FLAG_OWN_CAR_Y (Owns a Car)
5. NAME_FAMILY_STATUS_Married

Business Recommendations:

- The bank/lender can **prioritize older applicants** who might be more financially stable.
- **Loan approval policies** can be adjusted based on education levels, car ownership, and marital status.
- Higher loan amounts might need **stricter evaluation** due to their correlation with risk.
- Credit usage history should be a strong **indicator in credit scoring models**.

Conclusion

- Built a predictive model for credit risk assessment.

- Used **EDA, Feature Selection, and Multiple ML Models** for optimization

- **Best Model:** Logistic Regression (based on highest AUC-ROC score after tuning).

- **Impact:** Helps in better risk management and reducing financial losses.