A Comparative Study of Classification Models for Predicting Credit Card Approval

Divya Kothari
Student ID: 240485889

1 Introduction and Problem Statement

This project focuses on predicting whether a credit card application will be approved or rejected by analyzing applicant data. Features such as age, income, credit score, and other demographic or financial variables are studied to make informed predictions. This task is important for making quick and reliable financial decisions.

In this report, we carefully examine the dataset and apply several machine learning models, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Random Forests. All models were built using NumPy and evaluated using key metrics like accuracy, precision, recall, and F1-score to determine their effectiveness.

2 Analysis of the Dataset

2.1 Dataset Overview

2.1.1 Dataset Description

The credit card dataset includes 15 features and one target variable (Rejected) with values:
0: Application approved.
1: Application rejected.

Input Variables

- **Ind_ID:** Unique identifier for clients.
- **Gender:** Gender of the client.
- **Car_owner:** Whether the client owns a car or not.
- **Property_owner:** Whether the client owns property or not.
- **Children:** Number of children.
- **Annual_income:** Annual income of the client.
- **Type_Income:** Type of income source (e.g., working, state servant).

- **Education:** Education level.
- **Marital_status:** Marital status of the client.
- **Housing_type:** Housing style (e.g., apartment, co-op).
- **Birthday_count:** Days since birth (negative values indicate days before today).
- **Employed_days:** Days since the start of employment (negative values indicate days before today).
- **Type_Occupation:** Occupation type (e.g., laborer, IT staff).
- **Family_Members:** Number of family members.

Target Variable **Rejected:** Binary target variable where:
0: Application approved.
1: Application rejected.

Feature Types

- **Categorical:** Gender, Car_owner, Property_owner, Type_Income, Education, Marital_status, Housing_type, Type_Occupation.
- **Numerical:** Children, Annual_income, Age, Employment (years), Family_Members.

2.2 Initial Exploration

The dataset was first examined using `df.head()` and `df.describe()` to gain an overview of the features and their statistical summaries. `df.info()` revealed missing values and inconsistent naming conventions (e.g., `Propert_owner` corrected to `Property_owner`).

Missing Value Imputation

- **Gender:** Filled with mode.
- **Annual_income and Birthday_count:** Filled with median.
- **Type_Occupation:** Filled with mode.

2.3 Cleaning and Transformations

- Missing values handled using mode (categorical) and median (numerical).
- Features `Birthday_count` and `Employed_days` converted to Age and Employment (years) by taking the absolute values and scaling appropriately to improve interpretability.
- Categorical variables were mapped to numerical values for analysis.

Target Variable Distribution The dataset is imbalanced, with 88.7% approved applications (0) and 11.3% rejected (1):
Approved (0): 1373 samples.
Rejected (1): 175 samples.

2.4 Analysis of Features

2.4.1 Categorical Features

Categorical features were analyzed through bar plots to understand the distribution of categories and their relationship with `Rejected`.

2.4.2 Numerical Features

Numerical features were analyzed through histograms and boxplots for patterns, outliers, and trends.

2.4.3 Correlation Analysis

A heatmap was created to study the correlation among features and with the target variable. Key observations:

- **Children** and **Family_Members** had a high correlation (0.89).
- Correlation with `Rejected`:
  - **Children:** -0.022.
  - **Annual_income:** 0.024.
  - **Family_Members:** -0.031.
  - **Age:** 0.044.
  - **Employment (years):** -0.097.
- Negative correlations between:
  - **Age and Children:** -0.28.
  - **Age and Family_Members:** -0.27.

Key Visual Insights

- **Boxplots by Rejected Status:**
  - Employment (years): Approved clients had higher mean values and longer tails.
  - Annual_income: Outliers were present for approved clients.
- **Stacked Bar Charts:**
  - Education: Higher rejection rates for `Lower Secondary`.
  - Type_Income: State servants had the lowest rejection proportion.
  - Housing_type: `Co-op apartments` had the highest rejection proportion.
  - Type_Occupation: IT staff faced the highest rejection rates.

3 Data Preparation for Modeling

- **Mapping Categorical Features:** Converted categorical features to numerical values.
- **Normalization:** Scaled numerical features to a range of [0, 1].
- **Balancing the Target Variable:** Applied undersampling for class 0 and oversampling for class 1 to address the imbalance.

Major Inferences

- **Employment Stability:** Shorter employment durations strongly indicate higher rejection likelihood.
- **Income Type:** State servants and property owners are less likely to face rejection.
- **Demographics:** Larger families and higher education levels are associated with lower rejection probabilities.
- **Housing Type:** Clients living in `Co-op apartments` experienced the highest rejection rates.

4 Methods

4.1 Logistic Regression and Performance Evaluation

- Feature standardization to normalize input data.
- Implementation of logistic regression using the sigmoid activation function.
- Optimization of weights through gradient descent.
- Performance evaluation using metrics such as accuracy, F1 score, and confusion matrix.
- Plotting ROC curves to analyze classification thresholds.
- K-Fold cross-validation for robust evaluation.

4.2 Mathematical Formulations

1. Standardization

$$X_{\text{standardized}} = \frac{X - \mu}{\sigma + \epsilon}$$

where μ is the mean, σ the standard deviation, and $\epsilon = 10^{-8}$ prevents division by zero.

2. Sigmoid Function

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad z = Xw$$

3. Cost Function (Binary Cross-Entropy)

$$J(w) = -\frac{1}{m} \sum_{i=1}^{m} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]$$

4. Gradient Descent

$$w \leftarrow w - \eta \cdot \nabla J(w)$$

5. Classification Metrics

- Accuracy: $\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$
- F1 Score: $\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$