**Predicting Credit Card Default Using Machine Learning**

A Comparative Study of Classification Algorithms

**1. Overview of the Problem**

Credit card default occurs when a borrower fails to make the required payments on their credit card debt. This poses a significant risk to financial institutions, potentially leading to substantial losses and affecting overall credit market stability. With the increasing volume of credit card users globally, accurately predicting default behavior has become a critical challenge in financial risk management.

Machine learning offers powerful tools for analyzing complex patterns in customer data, enabling institutions to identify high-risk individuals before defaults occur. This report explores the application of various supervised learning algorithms to predict credit card default using historical customer data.

1. 1 Importance and Relevance of the Problem

Financial Impact: Credit card defaults contribute to billions in losses annually for banks and lenders. Early detection can reduce risk exposure and improve portfolio health.

Customer Profiling: Understanding default patterns helps institutions tailor credit offerings and improve customer segmentation

1. 2 Objectives of the Study

1. To apply and compare multiple machine learning algorithms for predicting credit card default.

2. To evaluate model performance using appropriate classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

3. To identify the most effective algorithm(s) for this task based on empirical results

**2. Dataset Description**

The dataset focuses on predicting whether a credit card client will default on their payment in the next month. It was originally collected by researchers in Taiwan and includes demographic, financial, and behavioral data for 30,000 clients

Source of the Dataset: UCI Machine Learning Repository

<https://archive-beta.ics.uci.edu/dataset/350/default+of+credit+card+clients>

The dataset contains 23 features plus one target variable. Here are the main categories:

**Demographic Features:**

LIMIT\_BAL: Amount of credit given (NT dollars)

SEX: Gender (1 = male, 2 = female)

EDUCATION: Education level (1 = graduate school, 2 = university, etc.)

MARRIAGE: Marital status (1 = married, 2 = single, etc.)

AGE: Age in years

**Payment History:**

PAY\_0 to PAY\_6: Repayment status from April to September (e.g., -1 = pay duly, 1 = payment delay for one month, etc.)

**Bill Statement & Payment Amounts:**

BILL\_AMT1 to BILL\_AMT6: Amount of bill statements from April to September

PAY\_AMT1 to PAY\_AMT6: Amount paid in the respective months

Target Variable:

default.payment.next.month: Binary (1 = default, 0 = no default)

**Assumptions and Limitations**

Geographic Bias: Data is based solely on clients in Taiwan, which may limit generalizability to other populations

Time-Specific: The data reflects a specific time period and economic context

No Missing Values: According to the UCI repository, the dataset does not contain missing values

**3. Methodology (Algorithms)**

This study implements and compares five supervised machine learning algorithms for credit card default prediction. Each algorithm represents a different learning paradigm, enabling comprehensive evaluation of linear vs. non-linear, parametric vs. non-parametric, and individual vs. ensemble approaches.

**4. Evaluation Metrics**

**5. Results & Discussions**

**5.1 Overall Performance Comparison**

**- Main metrics table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
| Logistic Regression |  |  |  |  |  |
| SVM |  |  |  |  |  |
| Random Forest |  |  |  |  |  |
| Decision Tree |  |  |  |  |  |
| KNN |  |  |  |  |  |

**- Bar charts for key metrics**

**- ROC curves overlay**

**5.2 Algorithm-Specific Analysis**

**3.2.1 Random Forest**

**- Hyperparameter tuning results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| n\_estimators | max\_depth | min\_samples\_split | Accuracy | F1-Score | ROC-AUC |
| 100 | **10** | **2** |  |  |  |
| 200 | **15** | **5** |  |  |  |
| 300 | **20** | **10** |  |  |  |

**- Coefficient analysis**

**3.2.2 SVM**

**- Kernel comparison**

**- C and gamma tuning**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| kernel | C | gamma | Accuracy | F1-Score | ROC-AUC |
| 100 | **10** | **2** |  |  |  |
| 200 | **15** | **5** |  |  |  |
| 300 | **20** | **10** |  |  |  |

**5.3 Impact of Data Preprocessing**

**3.3.1 Scaling Effects**

**3.3.2 Class Balancing Techniques**

**3.3.3 PCA Dimensionality Reduction**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Preprocessing | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
| Logistic Regression | **Scaled + Balanced** |  |  |  |  |  |
| Logistic Regression | **Scaled + Unbalanced** |  |  |  |  |  |
| Logistic Regression | **Scaled + Balanced + PCA** |  |  |  |  |  |
| Logistic Regression | **Scaled + Balanced + PCA** |  |  |  |  |  |
| SVM |  |  |  |  |  |  |

**5.4 Comparative Analysis**

**- Best performing algorithm**

**- Trade-offs (accuracy vs. time)**

**- Robustness across preprocessing methods**

**5.5 Limitations**

**- Dataset limitations**

**- Computational constraints**

**- Algorithm-specific challenges**

**5.6 Future Work**

**- Ensemble methods**

**- Deep learning approaches**

**- Feature engineering improvements**

**6. Conclusion**

**7. References**