****

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Programming for AI**

**ML Project**

Hamza Ahmad SP23-BAI-018

**Aksam Iftikhar**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Logistic Regression on Default of Credit Card Clients Dataset**

**1. Introduction**

**Dataset Description:**

* **Dataset Name:** Default of Credit Card Clients Dataset
* **Source:** UCI Machine Learning Repository
* **Key Features:**
  + Client demographic details (e.g., age, education level)
  + Credit and payment history
  + Bill statements and payment amounts
* **Target Variable:** default.payment.next.month (0 = No default, 1 = Default)
* **Size:** 30,000 records with 24 features
* **Missing Values:** No missing data was observed

**2. Methodology**

**Preprocessing Steps:**

* **Data Loading:**

The dataset was loaded using pandas.

* **Feature Scaling:**

The features were standardized using Standard Scaler to ensure uniform scaling.

* **Dimensionality Reduction:**

PCA (Principal Component Analysis) was applied to reduce the feature dimensions to 8 components, retaining most of the variance in the data.

* **Handling Class Imbalance:**

SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the dataset, ensuring equal representation of the target classes.

* **Train-Test Split:**

The data was split into training (80%) and testing (20%) subsets.

**Algorithm Applied:**

* **Logistic Regression:**

A statistical method to predict binary outcomes using a linear combination of features.

* **Optimization Techniques:**

**RandomizedSearchCV:**

**Performed hyperparameter tuning on the following parameters:**

Penalty type ('l1', 'l2')

Solver ('liblinear', 'saga')

Maximum iterations (100, 200, 300)

Conducted 5-fold cross-validation to select the best hyperparameters.

**3. Results**

**Best Hyperparameters:**

The best hyperparameters identified by RandomizedSearchCV:

* Penalty: l1
* Solver: saga
* Maximum iterations: 300

**Performance Metrics:**

* **Accuracy:** 0.64
* **Classification Report**:
  + Precision: 0.64
  + Recall: 0.65
  + F1-Score: 0.64

**Confusion Matrix:**

|  |  |  |
| --- | --- | --- |
|  | **Predicted: No Default** | **Predicted: Default** |
| **No Default** | 2986 | 1678 |
| **Default** | 1657 | 3025 |

**Execution Time:** 4.4 seconds

**Visualizations:**

1. **Class Distribution (Before and After SMOTE):**
   * Bar charts showed that SMOTE effectively balanced the target classes.
2. **Confusion Matrix:**
   * Visualized using Confusion Matrix Display.
3. **Performance Metrics:**
   * Bar chart displayed precision, recall, F1-score, and accuracy.

**4. Analysis**

**Insights:**

* The Logistic Regression model performed well in predicting credit card defaults, achieving an accuracy of 64%.
* SMOTE effectively balanced the dataset, improving the model's performance on the minority class.
* PCA helped reduce the feature space, improving computational efficiency without significant loss in predictive power.

**Challenges Faced:**

1. **Class Imbalance**:

The original dataset had a significant class imbalance, which was addressed using SMOTE.

1. **Hyperparameter Tuning**:

Selecting the optimal hyperparameters required computational resources

**Conclusion:**

The Logistic Regression model, optimized using RandomizedSearchCV and applied to the Default of Credit Card Clients Dataset, demonstrated robust performance. With an accuracy of 64% and balanced class predictions, the model is effective for credit risk prediction. Further enhancements can include experimenting with other algorithms such as Random Forest or XGBoost for potential performance gains.