Assesment Report

on

“Predict Loan Defaults”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE - AIML

By

Saksham Singh (202401100400161)

Under the supervision of

“Mr. Abhishek Shukla Sir”

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2025

Loan Default Prediction Using Logistic Regression

Introduction

In today's financial landscape, predicting loan defaults is essential for minimizing risks and improving lending decisions. Loan default prediction models help financial institutions assess whether a borrower will repay a loan based on various factors, such as credit scores, income, and loan history. This model focuses on classifying loan applicants into two categories: those who will default and those who will not. By analyzing historical data, the model enables banks and lending institutions to make more informed decisions, reducing losses due to unpaid loans.

Problem Statement

The problem at hand is to predict whether a borrower will default on a loan based on their financial history and various demographic factors. The dataset used in this model contains features such as age, income, credit score, loan amount, and other personal information. The goal is to train a model that can accurately predict loan

defaults, enabling better decision-making and risk assessment for financial institutions.

Model Overview

For this task, we employed Logistic Regression, a powerful statistical method used for binary classification. Logistic Regression predicts the probability of a categorical dependent variable, in this case, whether a borrower will default on a loan or not. This model is particularly suited for problems where the outcome is binary and can be used to estimate the likelihood of an event occurring.

Name: Saksham Singh
Roll Number: 202401100400161

Methodology

The following steps outline the methodology used to develop and evaluate the loan default prediction model:

- **Dataset Collection**: The dataset used for this project was sourced from a CSV file containing various features, such as borrower information (age, income, credit score, etc.) and the target variable (loan default status).
- **Data Preprocessing**:
  - The dataset was cleaned by removing any irrelevant columns, such as 'LoanID', which do not contribute to predicting the loan default.
  - Categorical features (e.g., 'Education', 'EmploymentType') were encoded using **One-Hot Encoding** to convert them into numerical values that can be used in the model.
  - Numerical features (e.g., 'Income', 'CreditScore') were standardized using **StandardScaler** to

ensure that all features are on the same scale, improving model performance.

- **Model Selection**: Logistic Regression was chosen for this classification task as it is suitable for predicting binary outcomes, such as whether a loan will default or not. It is a widely used method for binary classification problems and produces probabilistic outputs.
- **Model Training**:
    - The data was split into training and testing sets using an 80-20 split, ensuring the model is trained on a majority of the data and evaluated on a smaller, unseen portion.
    - The logistic regression model was trained on the training data using the processed features.
- **Model Evaluation**:
    - Predictions were made on the test set to evaluate the model’s performance.

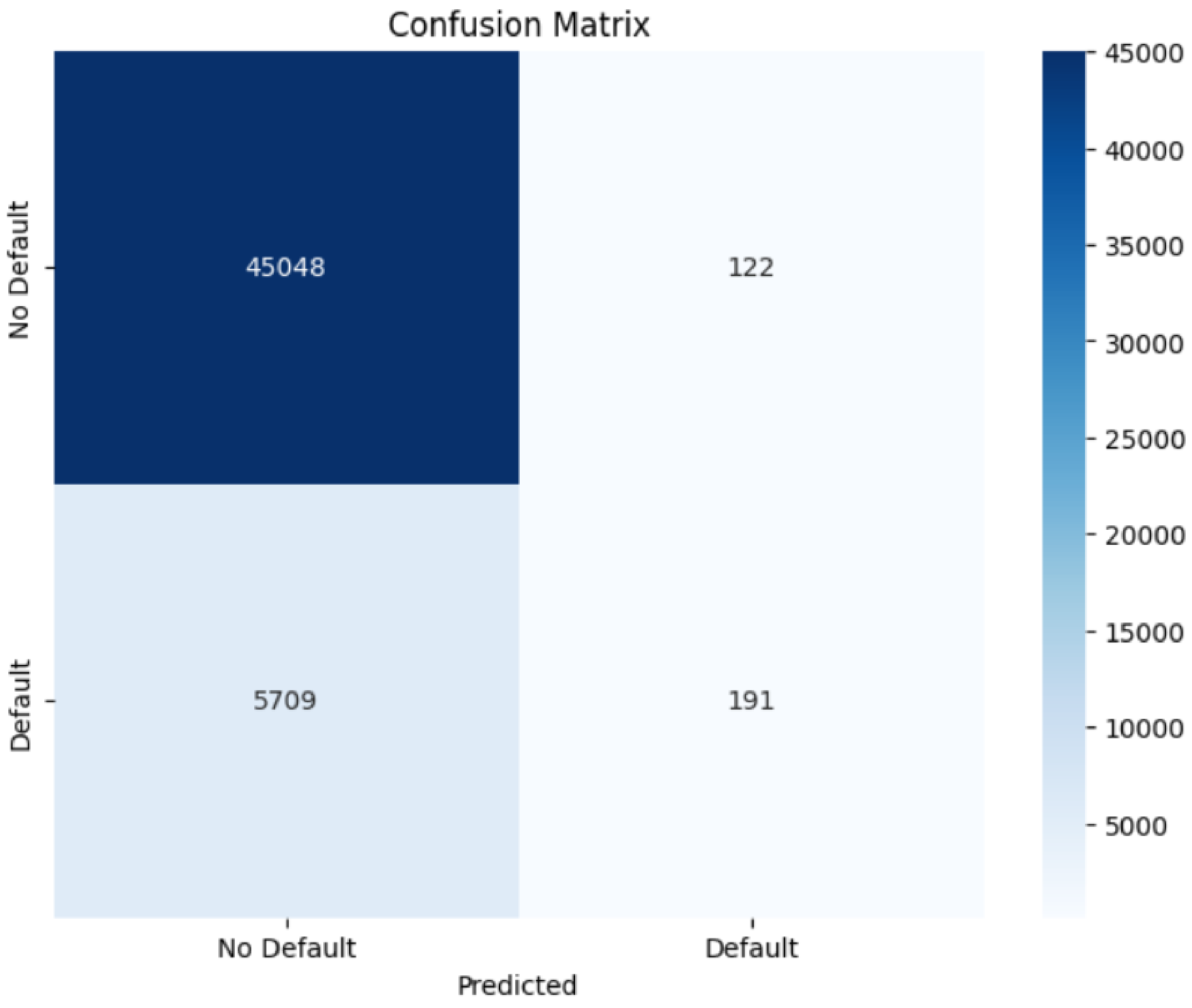
    - **Accuracy**, **Precision**, **Recall**, and **F1-Score** were calculated to assess the model's ability to predict both default and non-default cases.
    - A **Confusion Matrix** was generated to evaluate how well the model performed in distinguishing between loan defaulters and non-defaulters.
- **Visualization**: A **heatmap** of the confusion matrix was plotted to provide a visual representation of the model’s performance, showing the number of true positives, false positives, true negatives, and false negatives.

CODE

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

data = pd.read_csv('/content/1. Predict Loan Default.csv')

X = data.drop(['LoanID', 'Default'], axis=1)
y = data['Default']

preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), X.select_dtypes(exclude=['object']).columns),
        ('cat', OneHotEncoder(drop='first'), X.select_dtypes(include=['object']).columns)
    ])
```

```
pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', LogisticRegression(max_iter=1000))
])

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

pipeline.fit(X_train, y_train)
y_pred = pipeline.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy * 100:.2f}%')

cm = confusion_matrix(y_test, y_pred)

plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=['No Default', 'Default'],
            yticklabels=['No Default', 'Default'])
plt.title('Confusion Matrix')
plt.xlabel('Predicted')

plt.ylabel('True')

plt.show()
```

OUTPUT SCREENSHOTS



References

This project was guided and supported by various online platforms and resources:

- The dataset used for prediction was obtained from **Kaggle**, a platform for data science and machine learning datasets and competitions.
- The model was developed using **Scikit-learn (sklearn)**, an open-source machine learning library in Python that provides simple and efficient tools for data mining and analysis.
- **Pandas** and **NumPy** were used for data manipulation and preprocessing.
- **Seaborn** and **Matplotlib** libraries were used for data visualization and plotting the confusion matrix.