Assesment Report

on

“Predict Credit Card Fraud”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

By

Vibhor Gupta(202401100300277)

Under the supervision of

“Abhishek Shukla Sir”

KIET Group of Institutions, Ghaziabad

INTRODUCTION

The problem revolves around predicting whether a borrower will default on a loan based on various factors such as financial history, credit scores, etc. Loan default prediction is critical for financial institutions to minimize losses and make data-driven decisions.

In this task, we are using a dataset that includes various attributes of borrowers, including their financial history and other credit-related features. Our goal is to build a machine learning model to predict whether a borrower will default on a loan or not.

Key Concepts to Include:

- **Loan default and its impact on financial institutions**
- **Importance of credit scoring models in predicting loan default**
- **Machine learning models used in classification problems**

METHODOLOGY

To solve this problem, we used a classification approach. Here's how we approached the problem:

- **Data Preprocessing:**
  - Loaded the dataset from a CSV file.
  - Dropped the LoanID column as it was not useful for prediction.
  - One-hot encoded categorical features to make them suitable for machine learning models.
- **Feature Selection:**
  - The features used for prediction are financial information and credit scores, while the target variable is Default (whether the borrower defaults on the loan or not).
- **Model Selection:**
  - We selected a **Random Forest Classifier**, which is an ensemble learning method that works well for classification problems and is robust to overfitting.
- **Data Splitting and Scaling:**
  - The data was split into training and test sets (70% for training, 30% for testing).
  - Feature scaling was applied using StandardScaler to ensure that the model was not biased toward any particular feature due to scale differences.
- **Training the Model:**
  - We trained the Random Forest Classifier on the training data using the scaled features.

CODE

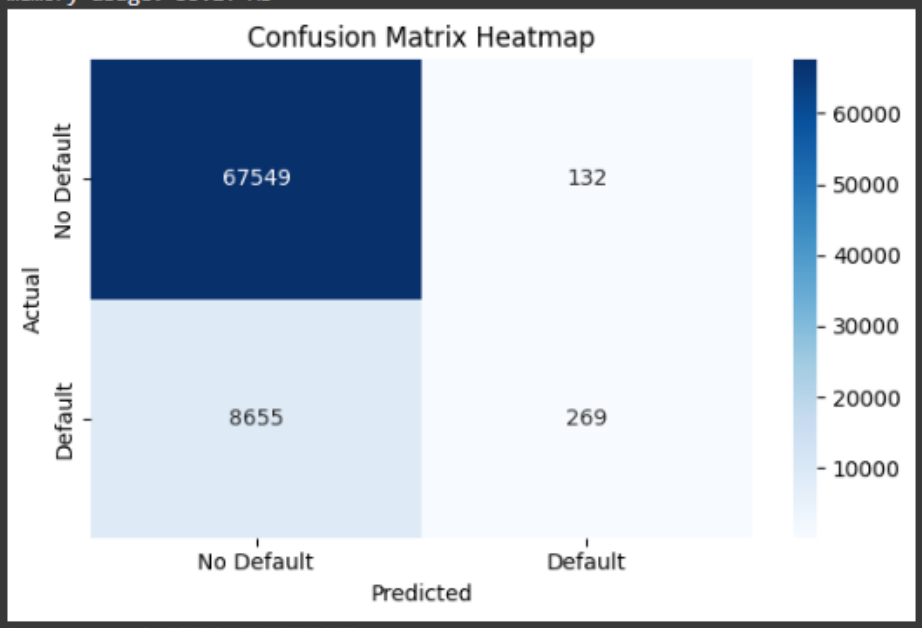
```
# Importing required libraries

import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import confusion_matrix, accuracy_score, precision_score,
recall_score

import seaborn as sns

import matplotlib.pyplot as plt


# Load the dataset

file_path = "1. Predict Loan Default.csv"

df = pd.read_csv(file_path)


# Display basic information and the first few rows

df.info(), df.head()


# Drop LoanID (not useful for prediction)

df = df.drop(columns=["LoanID"])


# One-hot encode categorical columns
```

```
df_encoded = pd.get_dummies(df, drop_first=True)


# Separate features and target
X = df_encoded.drop("Default", axis=1)
y = df_encoded["Default"]


# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)


# Feature scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)


# Train a classifier
clf = RandomForestClassifier(random_state=42)
clf.fit(X_train_scaled, y_train)


# Predictions
y_pred = clf.predict(X_test_scaled)
```

```
# Evaluation metrics
acc = accuracy_score(y_test, y_pred)
prec = precision_score(y_test, y_pred)
rec = recall_score(y_test, y_pred)

# Confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Plotting the heatmap
plt.figure(figsize=(6,4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['No Default', 'Default'], yticklabels=['No Default', 'Default'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix Heatmap')
plt.tight_layout()
plt.show()

# Print metrics
print("Accuracy: ", acc)
print("Precision: ", prec)
print("Recall: ", rec)
```

OUPUT

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 255347 entries, 0 to 255346
Data columns (total 18 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   LoanID          255347 non-null  object 
 1   Age             255347 non-null  int64  
 2   Income          255347 non-null  int64  
 3   LoanAmount      255347 non-null  int64  
 4   CreditScore     255347 non-null  int64  
 5   MonthsEmployed  255347 non-null  int64  
 6   NumCreditLines  255347 non-null  int64  
 7   InterestRate    255347 non-null  float64
 8   LoanTerm        255347 non-null  int64  
 9   DTIRatio        255347 non-null  float64
 10  Education       255347 non-null  object 
 11  EmploymentType  255347 non-null  object 
 12  MaritalStatus   255347 non-null  object 
 13  HasMortgage     255347 non-null  object 
 14  HasDependents   255347 non-null  object 
 15  LoanPurpose     255347 non-null  object 
 16  HasCoSigner     255347 non-null  object 
 17  Default         255347 non-null  int64  
dtypes: float64(2), int64(8), object(8)
memory usage: 35.1+ MB
```




```
Accuracy:  0.8852946935578617
Precision:  0.6708229426433915
Recall:  0.030143433437920215
```

REFERENCES

- Dataset: The dataset used for this project is titled "Predict Loan Default" and can be credited to the original source (if available).
- External Libraries:
    - pandas: For data manipulation.
    - scikit-learn: For machine learning model training and evaluation.
    - seaborn & matplotlib: For data visualization.