CREDIT CARD DEFAULT PREDICTION

Debanjan Chakraborty

INEURON INTERNSHIP January 11 ,2024

1.INTRODUCTION

In situations where manageable debts, like credit cards, spiral out of control due to factors such as job loss or medical crises, predicting and preventing customer default becomes crucial for banks. This model uses demographic data (gender, age, marital status) and behavioral data (last payments, past transactions) to assess and predict the likelihood of a customer defaulting on their debts, helping banks mitigate risks associated with credit card delinquencies.

2.PROBLEM STATEMENT

The financial industry's remarkable advancement is accompanied by emerging trends in financial threats, particularly in assessing credit risk for commercial banks. One of the primary challenges faced by these banks is predicting the risk of credit default among their clients. The objective is to forecast the likelihood of credit default by analyzing the characteristics of credit card owners and their payment histories. This predictive modeling aims to enhance the banks' ability to anticipate and manage credit risks effectively.

3.DATASET INFORMATION

ID: ID of each client

LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary = credit)

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

EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

MARRIAGE: Marital status (1=married, 2=single, 3=others)

AGE: Age in years

PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)

PAY_2: Repayment status in August, 2005 (scale same as above)

PAY_3: Repayment status in July, 2005 (scale same as above)

PAY_4: Repayment status in June, 2005 (scale same as above)

PAY_5: Repayment status in May, 2005 (scale same as above)

PAY_6: Repayment status in April, 2005 (scale same as above)

BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)

BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)

BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)

BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)

BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)

BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)

PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)

PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)

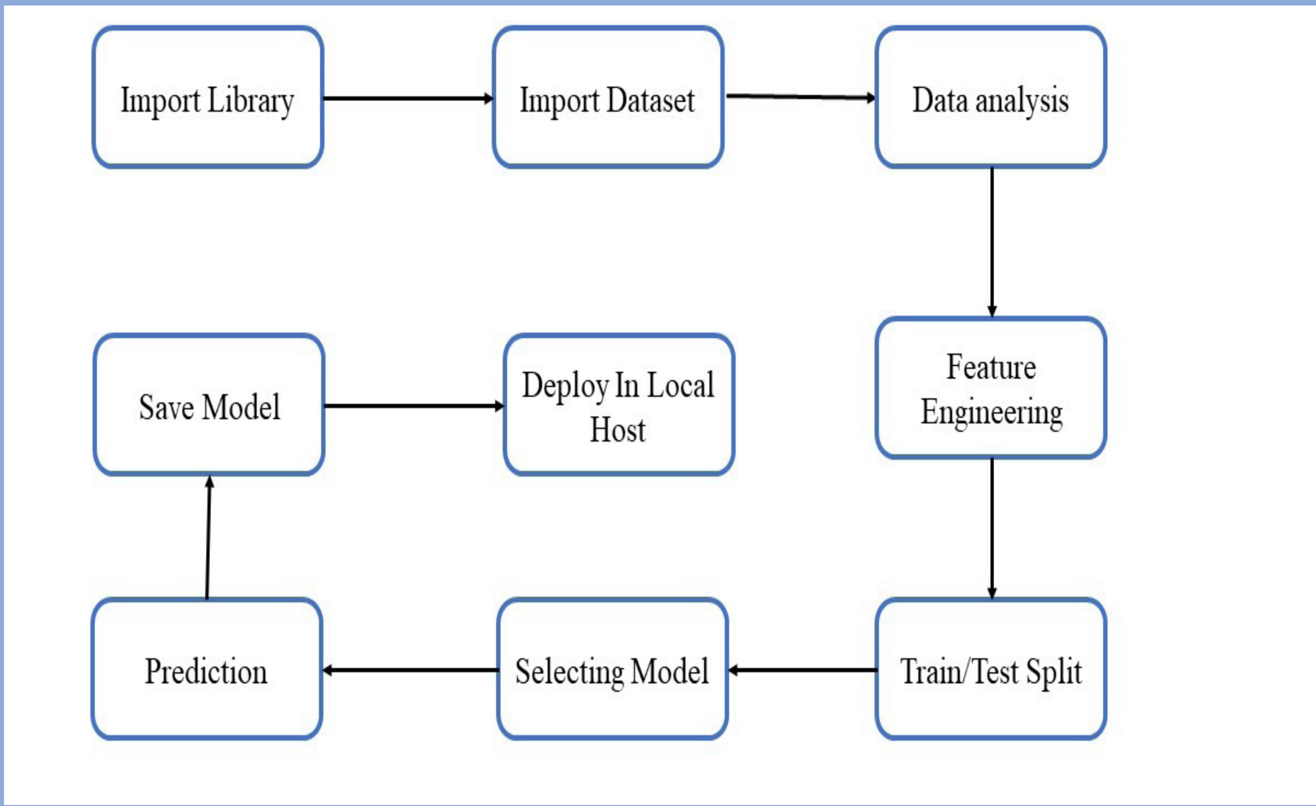
PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)

PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)

PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)

PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)

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


Import Library
Import Dataset
Data analysis
Feature Engineering
Train/Test Split
Selecting Model
Prediction
Save Model
Deploy In Local Host

4.Architecture Description.

4.1.Data Description:

The dataset was taken from Kaggle (URL:
https://www.kaggle.com/uciml/defaultof-

credit-card-clients-dataset), This dataset contains information on default

payments, demographic factors, credit data, history of payment, and bill statements

of credit card clients in Taiwan from April 2005 to September 2005.

4.2 Data Pre-processing.

This included importing of important libraries such as seaborn, matplotlib,
pandas etc. We imported the same dataset mentioned above from Kaggle.

4.3.Data Analysis

In our data analysis process, we diligently addressed null values, revamped column names, and utilized various visualization libraries like Seaborn, Matplotlib, and others. This strategic approach provided a comprehensive understanding of the data and highlighted the distribution of information within.

Given the absence of null values, we seamlessly transitioned into the visualization and analysis phase. For each distinct feature, a meticulous examination was conducted using visualizations. Key insights were systematically recorded, capturing crucial points that could significantly influence the final predictions. This thorough analysis not only ensures a robust understanding of the data but also sets the stage for informed and accurate predictions in subsequent stages of the project..

4.4. Feature Engineering

Merging 2 or mode columns to get indepth knowledge and information regarding
the data.

4.5.Train/Test Split

This library was imported from Sklearn to divide the final dataset into the ratio of
80-20%, where 80% of the data was used to train the model and the latter 20%
was used to predict the same.

4.6 Selecting Model

We tried and tested models such as RandomForest for the model and came up with the model with the best Performance.

4.7. Prediction

The Accuray of Random Forest was **82.06%** and the F1 score was **47.5%.**

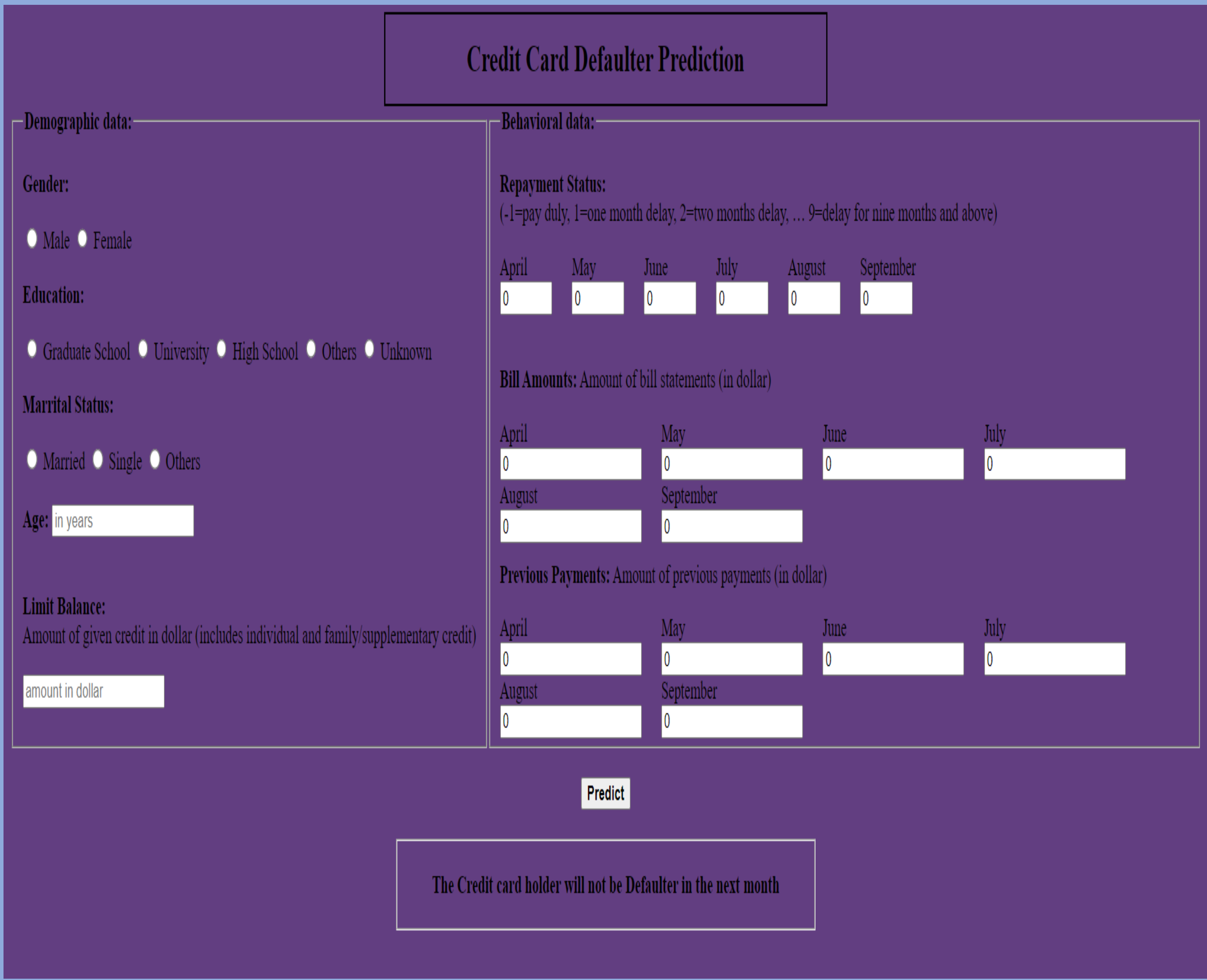
4.8 Save Model

Model was saved using the pickle library which saves the file in a binary mode.

4.9 Deploy in Local Host

Running on http://127.0.0.1:5000

The image in shown below,



We created a HTML template and deployed the model through Flask.