CREDIT CARD DEFAULT PREDICTION

Low-Level Design (LLD)

Author: Ritam Rakshit

Date: April 23, 2024

iNeuron

1. Problem Statement

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faced by commercial banks is the risk prediction of credit clients. The goal is to predict the probability of credit default based on the credit card owner's characteristics and payment history.

2. Dataset Information

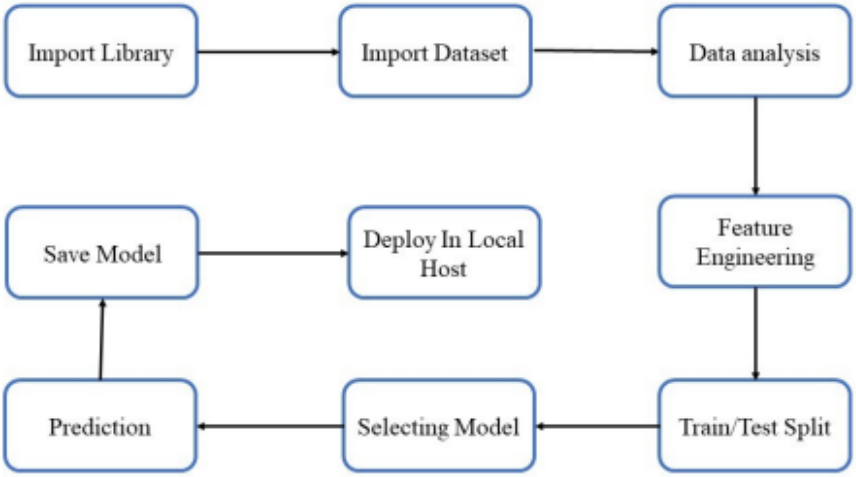
This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Content

There are 25 variables:

- 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)



3. Architecture Description.

3.1 Data Description

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

3.2 Data Preprocessing

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

3.3 Data Analysis

Here we handled the null values, changed the column names, and plotted multiple graphs in Seaborn, matplotlib and other visualization libraries for proper understanding of the data and the distribution of information in the same. As there were no null values in the data, we proceeded with the visualization and analysis. For each specific feature, we analysed the data using visualization and jotted down the important key points which can impact the final predictions.

3.4 Data Ingestion

Here we divided the data into 3 CSV files, raw.csv, train.csv and test.csv. 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.

3.5 Data Transformation (Feature Engineering)

We performed scaling and encoding using Scikitlearn. First, we divided the data into two categories, categorical data and numerical data. Then I used the fit-transform method. Here we also read that train and test CSV and changed them into arrays. Then saved it in preprocessor.pkl file for further processing.

3.6 Model Trainer

Here we train and select the best machine-learning model for predicting credit card defaults based on the provided data. We tried and tested multiple models such as LogisticRegression, Support Vector Classifier, KNeighborsClassifier, RandomForestClassifier, GaussianNB, AdaBoostClassifier, and GradientBoostingClassifier for the model and came up with the model with the best performance, i.e the GradientBoostingClassifier.

3.7 Prediction

The Accuracy of GradientBoostingClassifier was 82.22 and F1 score was 47.50

3.8 Save model

The model was saved using the pickle library

3.9 Deployment

We created HTML templates and the deployed model through Flask. Then created a docker image of it, deployed it in AWS ecr and used Apprunner .

Here are the images of our application-