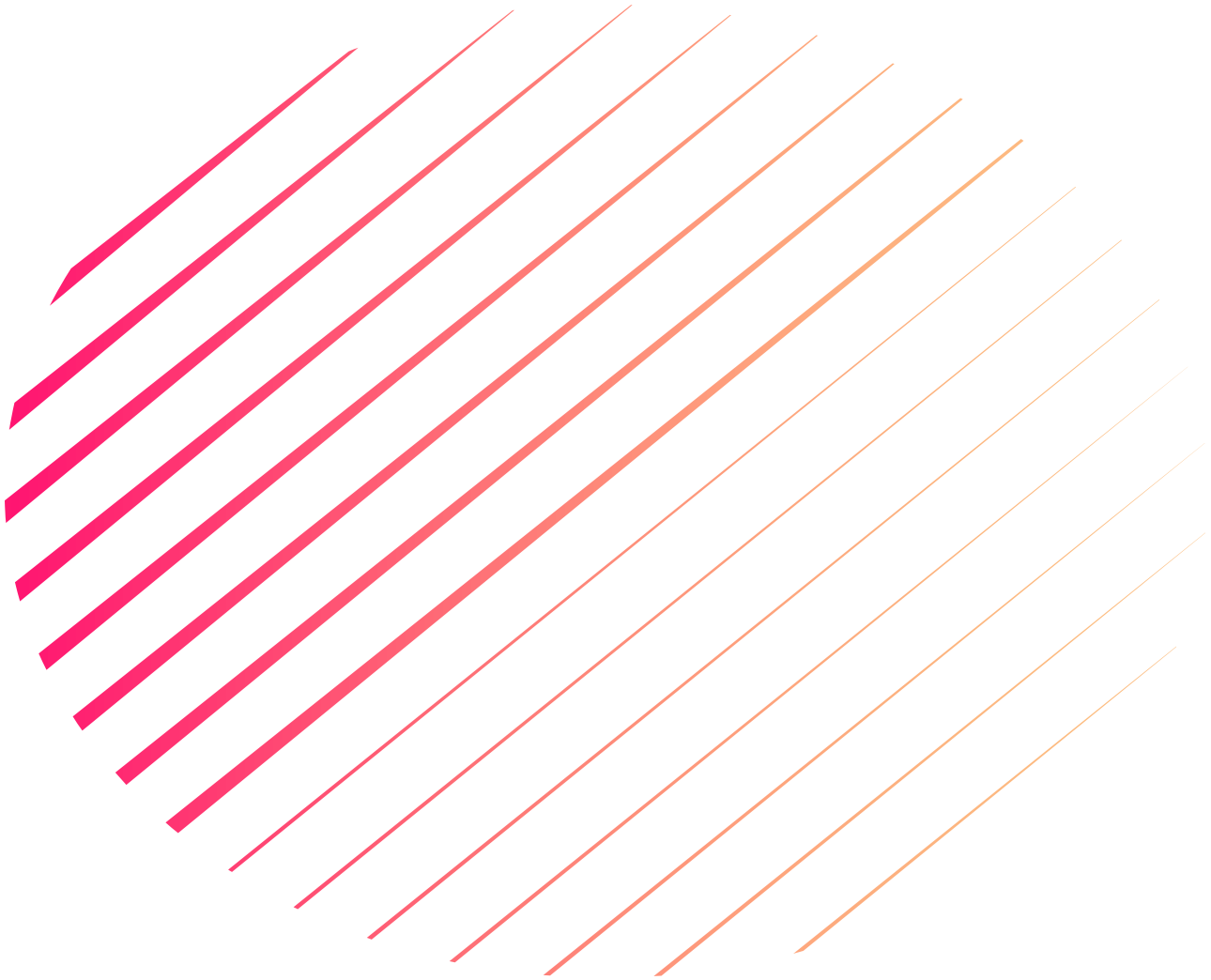
Credit Card Default Project

High Level Document (HLD)

HARSHA LOURDU M



Abstract

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 credit card owner's characteristics and payment history.

Use of various classification models like Decision Tree, Random Forest, XGBoost, MLPClassifier was done. XGBoost was selected as best model from above. The XGBoost model provided 82.63% accuracy in training. Testing accuracy for model was 82.3%. 80% data was used for training and 20% data was used for testing.

After providing various details the model will predict whether customer will default next month or not.

1. Introduction

1.1. Why this High-Level Design Document?

The Purpose of this High-Level Design Document is to add necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help and detect contradictions prior to coding and can be used as a reference manual for how modules interact at higher level.

The HLD will:

1.1.1. Present all of design aspects and define them in detail.
1.1.2. Describe the user interface being implemented.
1.1.3. Describe software interfaces.
1.1.4. Include Design features and architecture of the project.

1.2. Scope

The HLD document presents the entire structure of the project in parts, such as the data ingestion, data pre-processing, solution development, and the deployment part along with their respective architectures. This uses non-technical to mild technical terms which should be understandable to the administrators of the system.

1.3. Definitions

Term	Description
IDE	Integrated Development Environment
EDA	Exploratory Data Analysis
XGBoost	Extreme Gradient Boost Algorithm
ML	Machine Learning
MLP Classifier	Multi-Layer Perceptron Classifier
KNN	K-Nearest Neighbours

VS Code	Visual Studio Code

2. General Description

2.1. Product Perspective

This Credit Card Default Prediction is Machine Learning model based on XGBoost algorithm. Which helps us to predict whether customer will default next month payment or not. The app will also predict the probability of default.

2.2. Problem Statement

Financial threats are displaying a trend in 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.

2.3. Proposed Solution

The Solution here is a web application which takes the details of customer and those details will betaken by ML model in backend, which will predict whether customer will default or not. This app willalso predict probability of default on front end page of user.

2.4. Technical Requirements

I used python version 3.10.7 with some important libraries to develop a machine learning model, which accurately predicts Yes/No and probability of credit card default.

Then, the model is used as a back-end software for a front-end web application which can be usedby the users.

Front end is designed using Dash Library for this web app.

2.5. Dataset

For training and testing the model, I used the public data set available in Kaggle, “Default of Credit Card Clients Dataset” by UCI.

URL - https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

Variable Name	Measurement Unit	Description
LIMIT_BAL	NT Dollar	Amount of given credit

SEX	Integer	Gender (1=male, 2=female)
EDUCATION	Integer	1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown
MARRIAGE	Integer	Marital status (1=married, 2=single, 3=others)
AGE	Years	Age of the person in years
PAY_0-6	Integer	Repayment status for various months
BILL_AMT1-6	NT Dollar	Amount of billed statements for various months
PAY_AMT1-6	NT Dollar	Amount of Previous payments done
default.payment.next.month	Binary	Will Customer default? Yes/No

2.6. Tools and technologies used

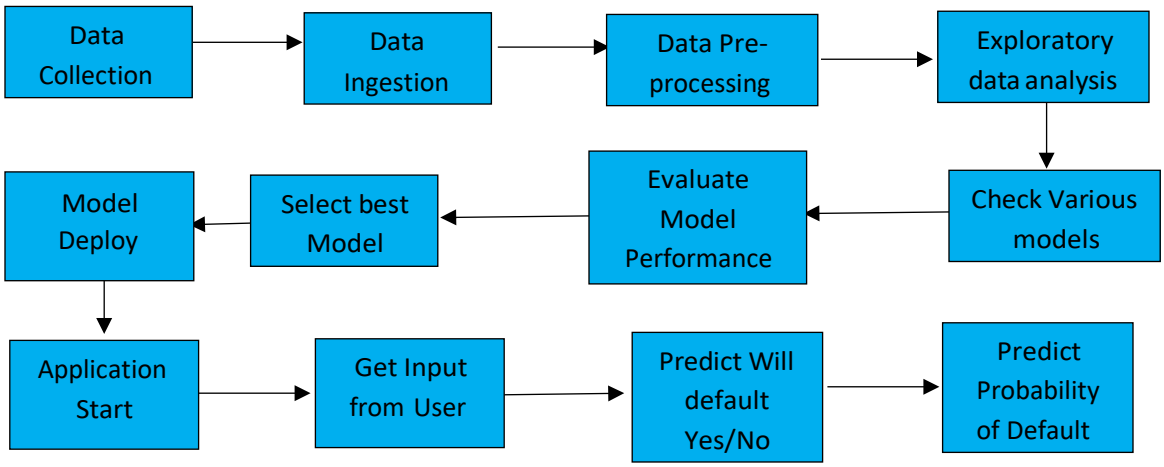
- Visual Studio Code (VS Code) was used as IDE (Integrated Development Environment). Everything including Jupyter notebooks, python apps, Git add, commit was done using VS studio Code.
- Jupyter Notebooks in VS Code were used for Exploratory Data analysis. Various Models like Decision Tree Classifier, Random Forest, XGBoost, KNN Classifier, MLP Classifier were tested in Jupyter Notebooks inside VS Code.
- VS Code was used to build .py files in which logging was also done. Other libraries used were

Numpy, pandas, Matplotlib, Seaborn for further exploration.

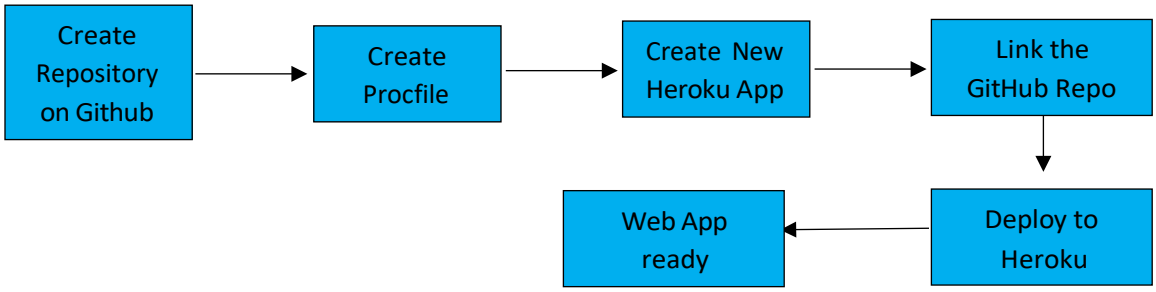
- Dash Library was used to build the front end of the web application and interaction withbackend was also done using dash call-back decorator.
- Statsmodels library is used for checking relationship with target variable.
- GitHub is used as Version Management System.

3. Design Details

3.1. Process Flow



3.2. Deployment Process



3.3. Event Log

In this project, I used the "logging" library in both the development and deployment stages, which keeps

logging the events at every step into the ".log" files. One of the advantages of event logging is, it makes debugging much easier, we can directly go to that specific line of code, which has errors.

3.4. Performance

The ML-based Probability of Default Predictor application is used for predicting the Probability Of Default based on various attributes of the customer. So, it should be as accurate as possible, so thatit will not mislead the bank authorities. Model retraining is very important to keep it relevant in order to keep the model dynamic to changing times and customer behaviour.

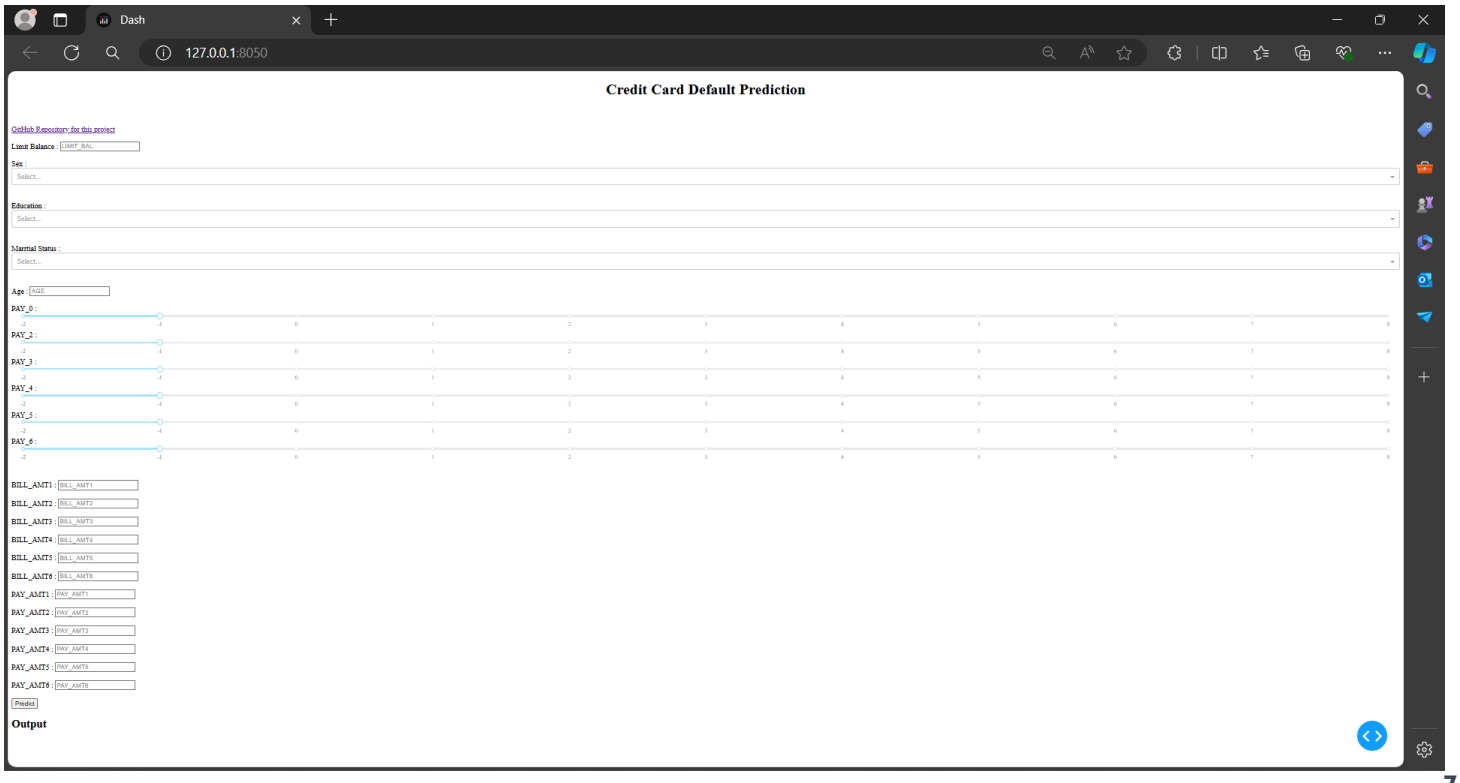
3.5. Reusability

The code written and the components used can be reused without any problem. Model Pickle files are also created for every model tested.

3.6. Application Compatibility

The different components or modules of this project use python version 3.10.7 as their interface between them. Each component has its own task to perform and it is the job of the python version to ensure proper transfer of the information.

3.7. User Interface

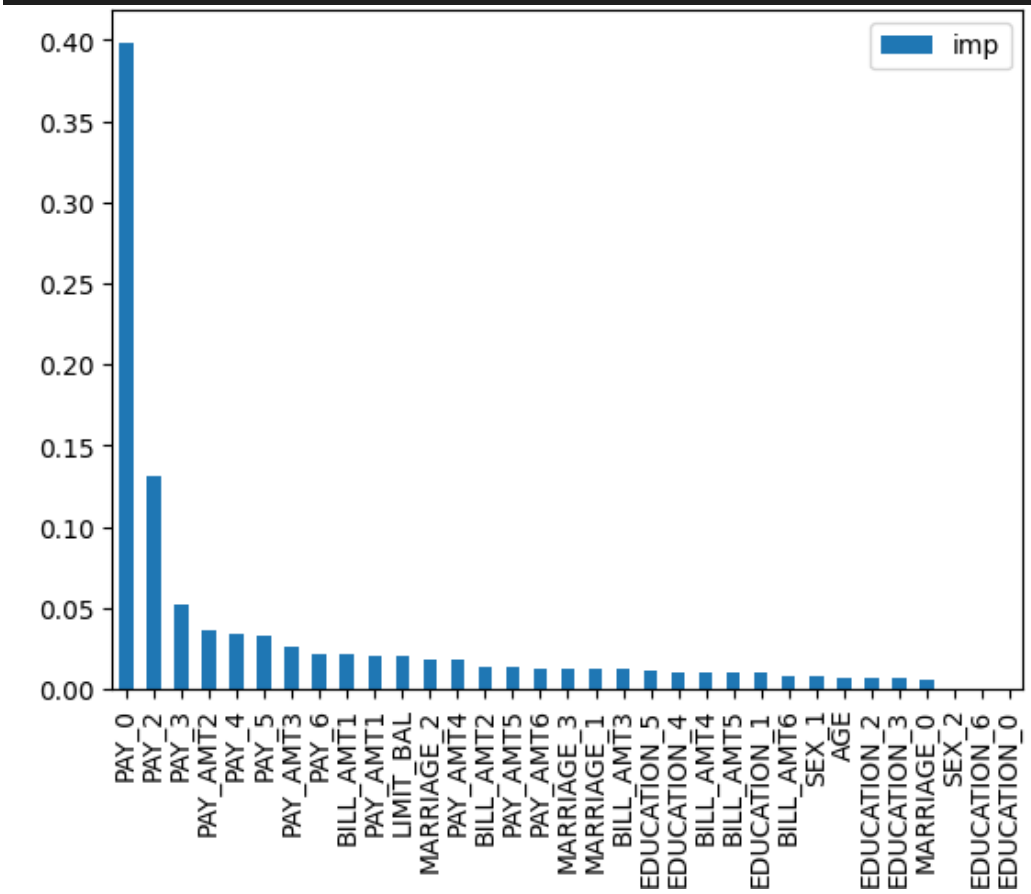


4. XGBoost Model Classification report

The XGBoost model used has 82.63% accuracy in testing and 82.3% accuracy in testing. Testing data is 20% of overall data.

Below is Classification report for selected XGBoost model.

	precision	recall	f1-score	support
0	0.84	0.96	0.89	4654
1	0.71	0.36	0.48	1346
accuracy			0.82	6000
macro avg	0.77	0.66	0.69	6000
weighted avg	0.81	0.82	0.80	6000



Below is feature importance for XGBoost model

5. Conclusion

Credit Card Default Predictor is used to predict that customer will default or not and also the probability of default given various attributes of the customer using with the help of ML and Data Science techniques in order to reduce bad debts of the bank.