LOW LEVEL DESIGN (LLD)

CREDIT CARD DEFAULT PREDICTION

Written By / Author	Nishika Patel Vaibhav Joshi
Document Version	LLD-V2.0
Last Revised Date	04/07/2023

Document Version Control

Date Issued	Version	Description	Author
03/07/2023	LLD-V1.0	First Version of Complete LLD	NISHIKA PATEL
04/07/2023	LLD-V2.0	Final Version of Complete LLD	VAIBHAV JOSHI

Contents

Abstract

At times, even a seemingly manageable debt, such as credit cards, can spiral out of control. Loss of job, medical emergency, or business failure are all events that can take a toll on your finances. Credit card debt is often the first to become overwhelming due to the hefty finance charges (compounded on daily balances) and other penalties. Many of us can relate to this, having missed credit card payments once or twice due to forgotten due dates or cash flow problems. But what happens when this continues for months? How can you predict whether a customer will be a defaulter in the coming months? To reduce the risk for banks, this model has been developed.

A little development inside the precision of sorting out extreme danger credits might need to save you misfortunes of more than $eight billion. Developing consumer spending patterns to limit risk exposures in this area is becoming increasingly important due to the risks associated with this large portion of the economy. For this to be a potential choice, the expectations need to be modestly exact. A solid rendition isn't least difficult a helpful gadget for the loaning foundations to decide using a credit card score applications, but it might furthermore help the clients to be conscious of the ways of behaving which can hurt their FICO rating scores. Utilizing economic statistics is the primary motivation behind risk prediction, For instance, corporation transactional statistics, alternate statistics, and patron transactions, amongst others, to forecast the patron's enterprise overall performance or individual credit score card statistics and to lessen bogs and vulnerabilities. Several risk prediction models are entirely dependent on statistical methods, such as logistic regression, Random Forest, *Gradient Boost* and Decision making.

The goal of credit default prediction is to enable financial entities determine whether or not to lend to a customer. The resulting check is frequently a threshold cost that allows decision-makers to make the financing choice. The popular form is based on economic ratios, profit accounts, and stability sheet statistics.

1. Introduction

This document will be used for documenting Low-level designs of project.

1.1 Why this Low-Level design document?

The purpose of this LLD or a Low-Level Design (LLD) document is to give the internal logical design of the actual program code for Swiggy Data Analysis project. LLD describes the class diagrams with the methods and relations between classes and program specs. It describes the modules so that the programmer can directly code the program from the document. This document is intended for both the stakeholders and the developers of this project and will be proposed to the higher management for its approval.

The main objective of the project is to analyse the various aspects with different use caseswhich covers many aspects of Swiggy Food Delivery Service. It helps in not only understanding the meaningful relationships between attributes but it also allows us to do our own research and come-up with our findings.

1.2 Scope

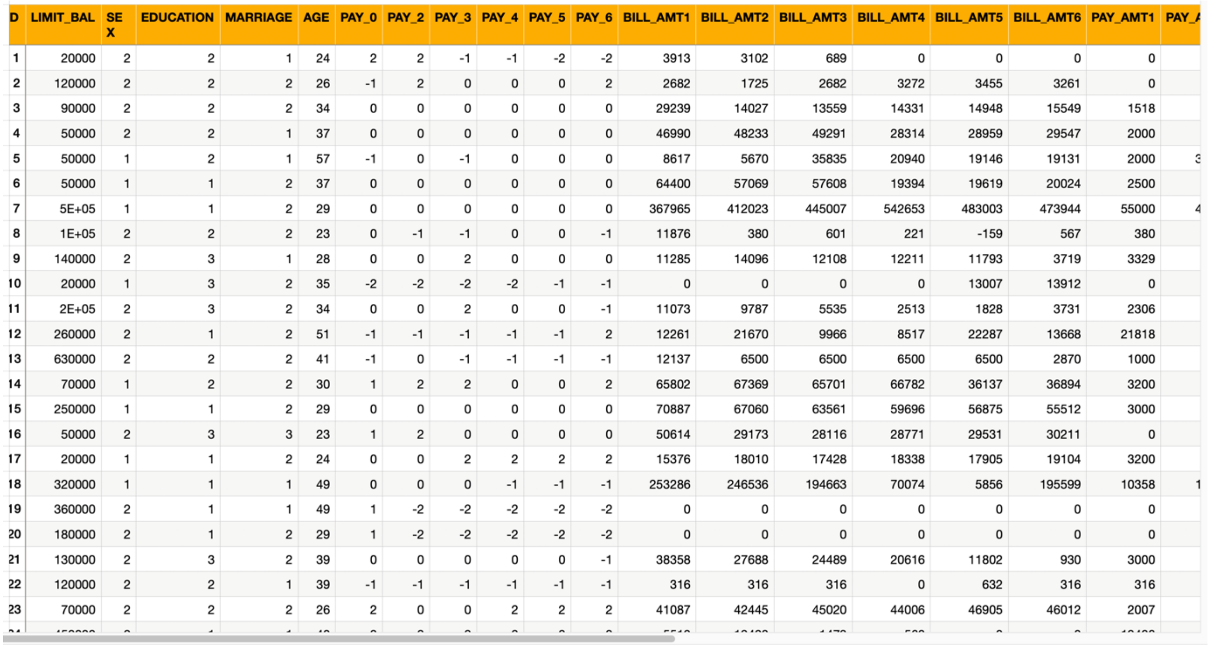
Low-level design (LLD) is a component-level design process that follows a step-by step refinement process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work. This study demonstrates the how different analysis help out to make better business decisions and help analyse customer trends and satisfaction, which can lead to new and better products and services.

1.3 Constraints

The analysis must be user friendly, code must be neat & clean, EDA must be automated as much as possible because it will save huge amount of time. Moreover, users should not be required to have any of the coding knowledge as the insights they are looking for are mentioned in-detail with respective visuals.

2. Technical Specifications

2.1 Credit Card Default Predication Dataset –



D	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_A
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	3
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	
7	5E+05	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	4
8	1E+05	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	
9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	
11	2E+05	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	
12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	
13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	
14	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	
15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	
16	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	
17	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	
18	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	1
19	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	
20	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	
21	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930	3000	
22	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316	316	
23	70000	2	2	2	26	2	0	0	2	2	2	41087	42445	45020	44006	46905	46012	2007	

2.2 Credit Card Default Predication Dataset Overview

The Listings dataset consists of a table with 30000 records and 25 features. There are a total 0% of records having Missing Values. In short, there are no Missing Values present in the dataset.

Name	Description
ID	Name of the Shop/Restaurants
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)

2.3 DATASET DESCRITION:

The dataset was taken from Kaggle. 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. Architecture

3.1 Data Analysis:

We used the Seaborn and Matplotlib libraries to create different graphs to better comprehend the data and information distribution. We proceeded with the visualisation and analysis because there were no null values in the data.

We analysed the data utilising visualisation for each unique feature and noted the significant key elements that could influence the final forecasts.

3.2 Exploratory Data Analysis (EDA) –

- "Exploratory Data Analysis" (EDA) is a "Data Exploration" step in the Data Analysis Process, where a number of techniques are used to better understand the dataset being used.
- Understanding the Dataset can refer to a number of things including but not limited to…
- Extracting Important "Variables".
- Identifying "Outliers", "Missing Values", or "Human Error".
- Understanding the Relationships between variables.
- Ultimately, maximizing our insights of a dataset and minimizing potential "Error" that may occur later in the process.

- In other words, it will gives you a better Understanding of the "Variables" and the "Relationships" between them.

- Here, we make use of dataprep module to automate our EDA process.

- It provides the following information:

- **Overview**: detect the types of columns in a DataFrame.

- **Variables**: variable type, unique values, distinct count, missing values Quartile statistics like minimum value, Q1, median, Q3, maximum, range, interquartile range Descriptive statistics like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness.

- **Correlations**: highlighting of highly correlated variables, Spearman, Pearson and Kendall matrices

- **Missing Values**: Bar Chart, Heatmap and spectrum of missing values.

3.3 Data Pre-processing:

Important libraries like Seaborn, Matplotlib, Pandas, and others had to be imported as part of this. The identical dataset from Kaggle that was stated above was imported. Additionally, we looked to see whether any columns had a standard deviation of zero. Additionally, the pre-processing checks for the presence of null values; the dataset has no null values, therefore we continue using it in its current state.

3.4 Train Test Split:

This library was imported from sklearn in order to split the final dataset into two halves, with 70% of the data being used to train the model and the remaining 30% being used to predict the same.

3.5 Model Selection:

Our datasets will be trained using four different algorithms: XGBoost Classifier, Random Forest, Decision Tree and Logistic Regression. Any model we employed is saved in the model's comparison table.

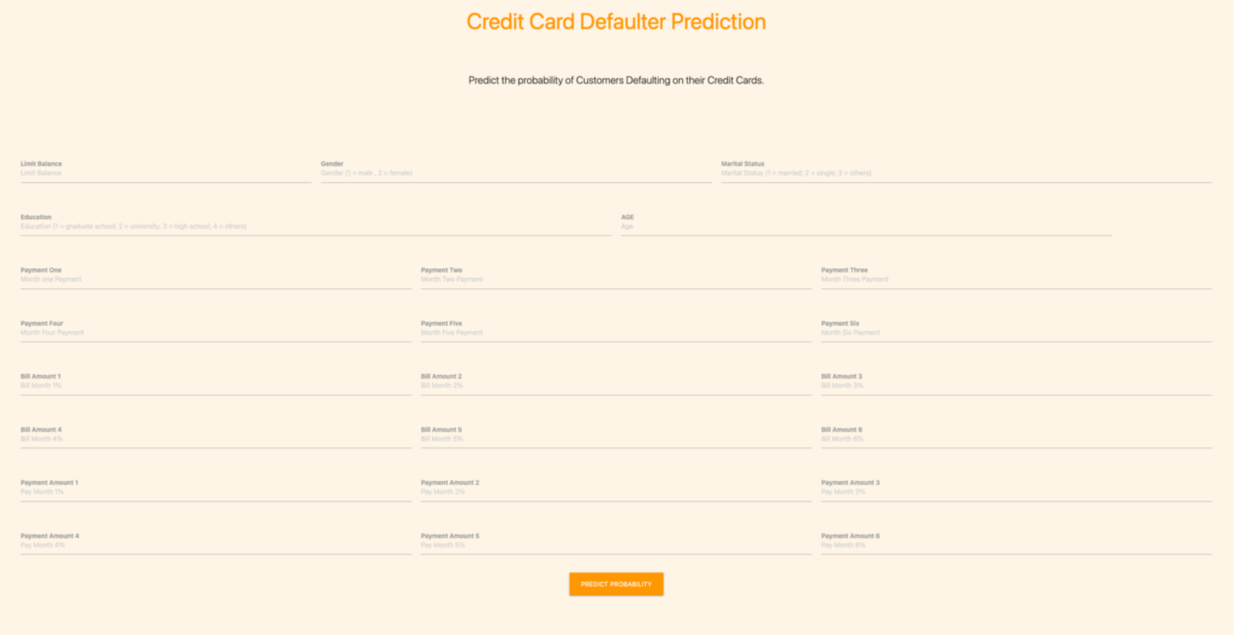
3.6 Prediction:

We used the user's input to inform our predictions. Then, after loading the relevant model that had been previously saved during training, we made a prediction. The pickle library, which saves the model in binary mode, was used to save the model.

3.7 Deploy:

This one has been set up on a local host, where a web page has been developed to receive input from the user and provide a prediction.

Here is a picture of the same thing.



Credit Card Defaulter Prediction

Predict the probability of Customers Defaulting on their Credit Cards.

Limit Balance
Limit Balance

Gender
Gender (1 = male , 2 = female)

Marital Status
Marital Status (1 = married; 2 = single; 3 = others)

Education
Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)

AGE
Age

Payment One
Month one Payment

Payment Two
Month Two Payment

Payment Three
Month Three Payment

Payment Four
Month Four Payment

Payment Five
Month Five Payment

Payment Six
Month Six Payment

Bill Amount 1
Bill Month 1%

Bill Amount 2
Bill Month 2%

Bill Amount 3
Bill Month 3%

Bill Amount 4
Bill Month 4%

Bill Amount 5
Bill Month 5%

Bill Amount 6
Bill Month 6%

Payment Amount 1
Pay Month 1%

Payment Amount 2
Pay Month 2%

Payment Amount 3
Pay Month 3%

Payment Amount 4
Pay Month 4%

Payment Amount 5
Pay Month 5%

Payment Amount 6
Pay Month 6%

4. Technology Stack

Data Manipulation Library	Pandas
Library use for Build Whole model	Sklearn, Pickle,Imblearn, seaborn xgboost
EDA	dataprep
Dataset	.CSV Format
IDE	Jupyter Notebook