**PW SKILLS: FULL STACK DATA SCIENCE PRO**

***Submitted by: POONAM SAINI***

**LOW level design (LLD)**

****

**Credit Card Default Prediction**

**INTRODUCTION**

In today's increasingly digital financial landscape, managing credit risk is more important than ever. A credit card default occurs when a cardholder doesn't make the agreed payments on their credit card debt. This can occur due to various reasons, such as unemployment, medical emergencies, or financial difficulties. Defaulting on credit card payments can have serious consequences, including harm to credit scores, accumulating debt from interest and penalties, and the possibility of legal action by creditors.

Predicting credit card defaults is vital as it helps financial institutions identify at-risk customers early. By analyzing data like demographics, payment history, and spending behavior, predictive models can estimate the likelihood of defaults. This allows banks to take steps like offering financial guidance, adjusting credit limits, or restructuring payment plans to lower the risk of default and minimize financial losses.

**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.

In today's increasingly digital financial landscape, managing credit risk is more important than ever. A credit card default occurs when a cardholder doesn't make the agreed payments on their credit card debt. This can occur due to various reasons, such as unemployment, medical emergencies, or financial difficulties. Defaulting on credit card payments can have serious consequences, including harm to credit scores, accumulating debt from interest and penalties, and the possibility of legal action by creditors.

Predicting credit card defaults is vital as it helps financial institutions identify at-risk customers early. By analyzing data like demographics, payment history, and spending behavior, predictive models can estimate the likelihood of defaults. This allows banks to take steps like offering financial guidance, adjusting credit limits, or restructuring payment plans to lower the risk of default and minimize financial losses.

**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.

**Data Insights**

This dataset includes details on default payments, demographic characteristics, credit information, payment history, and bill statements for credit card clients in Taiwan, covering the period from April 2005 to September 2005.

**Data Variables Overview**

This dataset consists of 25 variables, described as follows:

**ID**: Unique identifier for each client.

**LIMIT\_BAL**: Total credit amount provided in NT dollars (includes both individual and family/supplementary credit).

**SEX:** Gender of the client (1 = male, 2 = female).

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

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

**AGE:** Client's age in years.

**PAY\_1:** Repayment status for September 2005 (-1 = pay duly, 1 = payment delayed by one month, 2 = payment delayed by two months, and so on up to 9 = payment delayed by nine months or more).

**PAY\_2:** Repayment status for August 2005 (same scale as above).

**PAY\_3:** Repayment status for July 2005 (same scale as above).

**PAY\_4:** Repayment status for June 2005 (same scale as above).

**PAY\_5:** Repayment status for May 2005 (same scale as above).

**PAY\_6:** Repayment status for April 2005 (same scale as above).

**BILL\_AMT1:** Amount on the bill statement for September 2005 (in NT dollars).

**BILL\_AMT2:** Amount on the bill statement for August 2005 (in NT dollars).

**BILL\_AMT3:** Amount on the bill statement for July 2005 (in NT dollars).

**BILL\_AMT4:** Amount on the bill statement for June 2005 (in NT dollars).

**BILL\_AMT5:** Amount on the bill statement for May 2005 (in NT dollars).

**BILL\_AMT6:** Amount on the bill statement for April 2005 (in NT dollars).

**PAY\_AMT1:** Amount of payment made in September 2005 (in NT dollars).

**PAY\_AMT2:** Amount of payment made in August 2005 (in NT dollars).

**PAY\_AMT3:** Amount of payment made in July 2005 (in NT dollars).

**PAY\_AMT4**: Amount of payment made in June 2005 (in NT dollars).

**PAY\_AMT5:** Amount of payment made in May 2005 (in NT dollars).

**PAY\_AMT6:** Amount of payment made in April 2005 (in NT dollars).

**default.payment.next.month:** Indicates whether the client defaulted on payment in the next month (1 = yes, 0 = no).

**Architecture Description:**

**Dataset overview**

The dataset used in this project was sourced from Kaggle (URL). It contains detailed information on credit card clients in Taiwan, covering the period from April 2005 to September 2005. The dataset includes variables related to default payments, demographic factors, credit data, payment history, and bill statements. These factors provide a comprehensive view of the clients' credit behavior, which is essential for predicting the likelihood of default.

**Data Preprocessing**

The data preprocessing stage involved importing essential libraries such as Seaborn, Matplotlib, and Pandas. These libraries were crucial for data manipulation, visualization, and analysis. We then imported the dataset from Kaggle, as mentioned earlier, and proceeded to clean, transform, and prepare the data for model building.

**Data Analysis**

In the data analysis phase, we undertook the following steps:

* **Null Value Handling**: We assessed the dataset for null values, but since there were none, we moved forward without needing to address missing data.
* **Column Renaming**: To enhance clarity, we renamed columns, making the dataset more intuitive and easier to work with.
* **Visualization:** We utilized Seaborn, Matplotlib, and other visualization libraries to plot multiple graphs, providing a clear understanding of the data distribution and relationships between features.
* **Feature Analysis:** Each feature was thoroughly analyzed using visualizations to identify key points that could influence the final predictions. These insights were documented to guide the modeling process.

This thorough data analysis was crucial for understanding the dataset's structure and for informing subsequent modeling decisions.

**Data Transformation and Feature Engineering**

In the data transformation phase, we focused on feature engineering to prepare the data for model training:

* **Data Categorization:** We divided the data into two categories: categorical and numerical.
* **Scaling and Encoding:** Using Scikit-learn, we applied scaling to the numerical data and encoding to the categorical data. The fit-transform method was employed to standardize these transformations.
* **Data Conversion:** The processed data was then converted into arrays, facilitating easier handling during model training.
* **Saving Processed Data:** The transformed data, including both the train and test datasets, was saved in an IPython Notebook (.ipynb) file for further processing.

These steps ensured that the data was appropriately transformed, allowing the model to accurately interpret and learn from the features during training.

**Model Training and Selection**

In this phase, we focused on training and selecting the best machine learning model for predicting credit card defaults:

* **Model Training:** We experimented with various machine learning algorithms to identify the most effective model. The models tested included:
  + Logistic Regression
  + Support Vector Classifier (SVC)
  + K-Nearest Neighbors Classifier (KNeighborsClassifier)
  + Random Forest Classifier (RandomForestClassifier)
  + Gaussian Naive Bayes (GaussianNB)
  + AdaBoost Classifier
  + Gradient Boosting Classifier (GradientBoostingClassifier)
* **Model Selection:** After evaluating the performance of each model, the Gradient Boosting Classifier emerged as the best-performing model based on key performance metrics.

This process ensured that the most accurate and reliable model was selected for predicting credit card defaults, providing a robust solution to the problem

**Prediction**

The Gradient Boosting Classifier, selected as the best model, achieved an accuracy of 82.22% and an accuracy score of 69% on the test data. These metrics indicate the model's effectiveness in predicting credit card defaults.

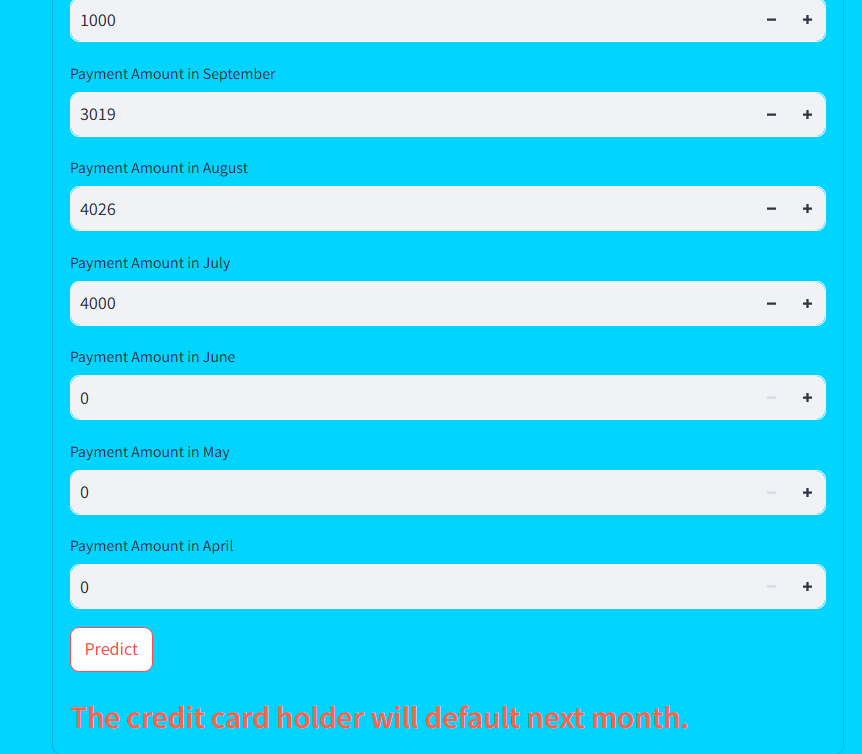
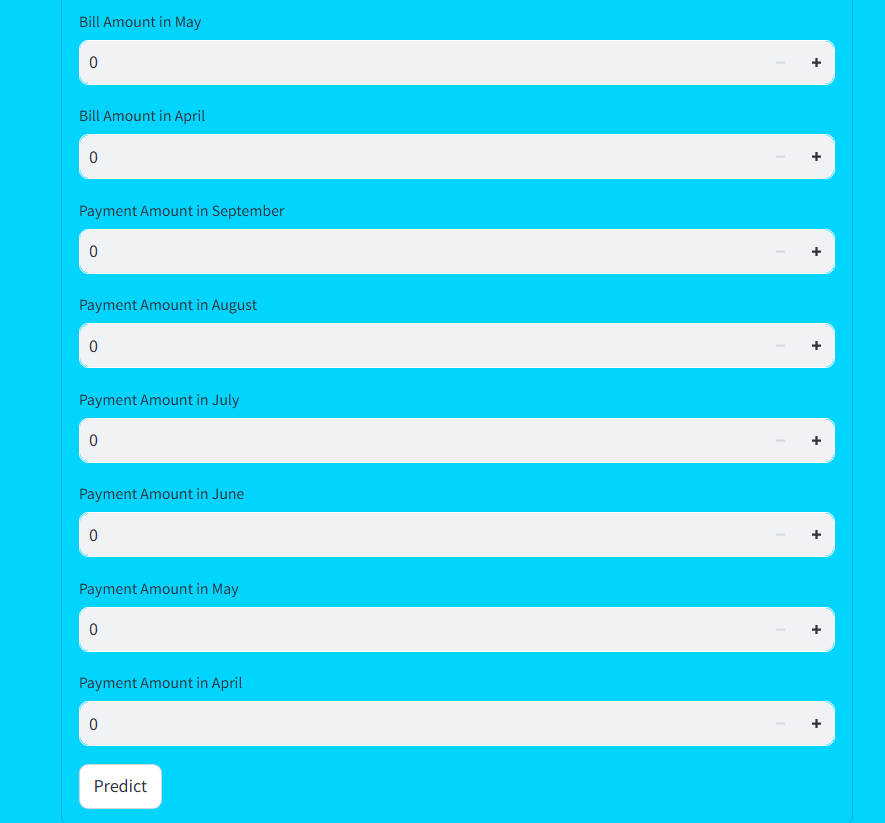
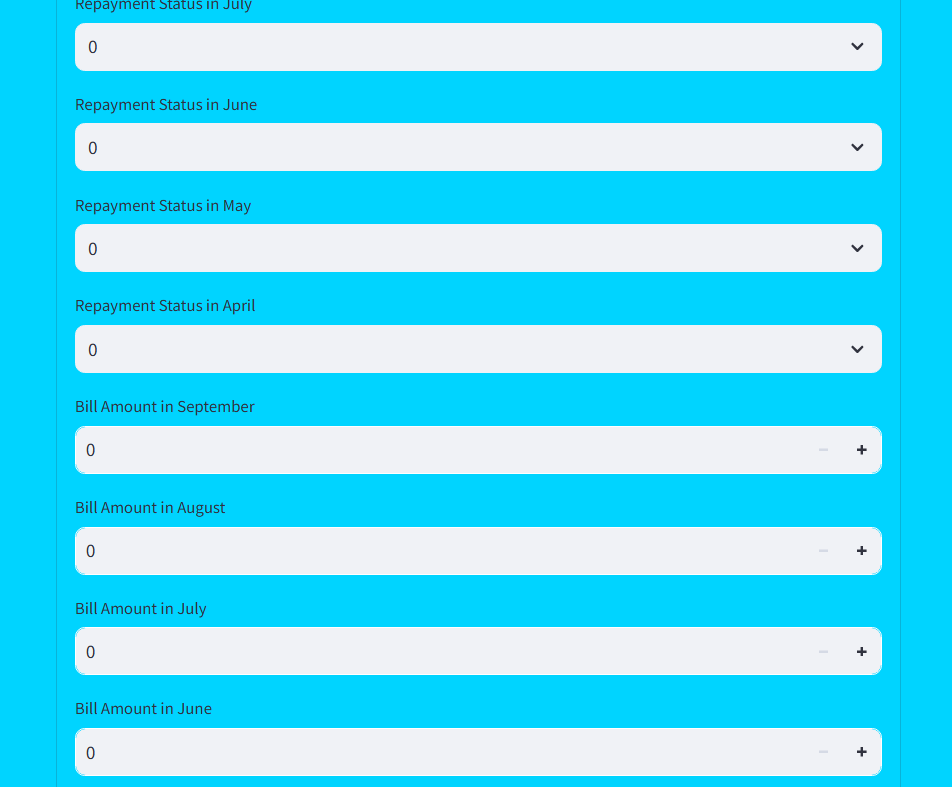
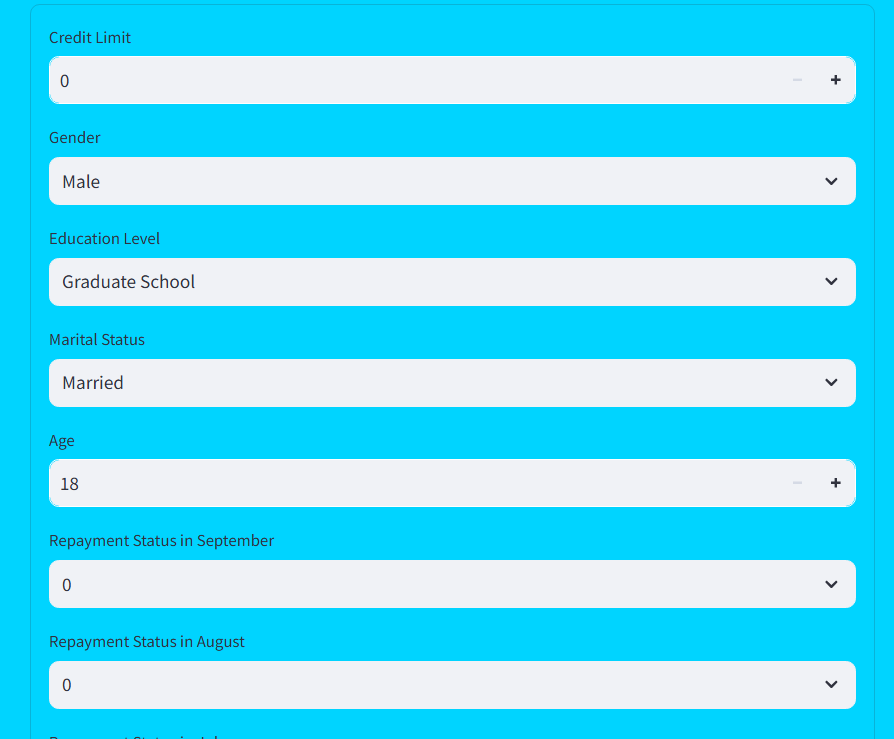
**Model Saving**

The trained model was saved using the pickle library, allowing for easy storage and later use in deployment.

**Deployment**

The saved model was deployed using Streamlit, enabling an interactive web application where users can input data and receive predictions. Below are images of the deployed application:

****

****