**LAWRENCE TECHNOLOGICAL UNIVERSITY**

**DEPARTMENT OF COMPUTER SCIENCE**

**MACHINE LEARNING ASSIGMDENT -4(FINAL REPORT)**

“Credit Card Default Prediction”

**Data set information:**

Predictive Model to Identify Default Risk

Introduction & Domain Knowledge

Objective: Predict whether a credit card holder will default on payment in the next month.

Importance: Helps financial institutions manage risk and optimize lending.

Dataset: UCI Default of Credit Card Clients (30,000 records of Taiwanese credit card holders).

Key Features: Includes demographic data, financial behavior, and payment history.

**Dataset Analysis**

Total Data: 30,000 rows, 25 columns.

Categories:

Demographic: Gender, Age, Education, Marital Status.

Financial: Credit Limit, Payment History, Bill Amounts, Payment Amounts.

Target Variable: Default payment next month (1 = Default, 0 = No Default).

Feature Analysis & Selection

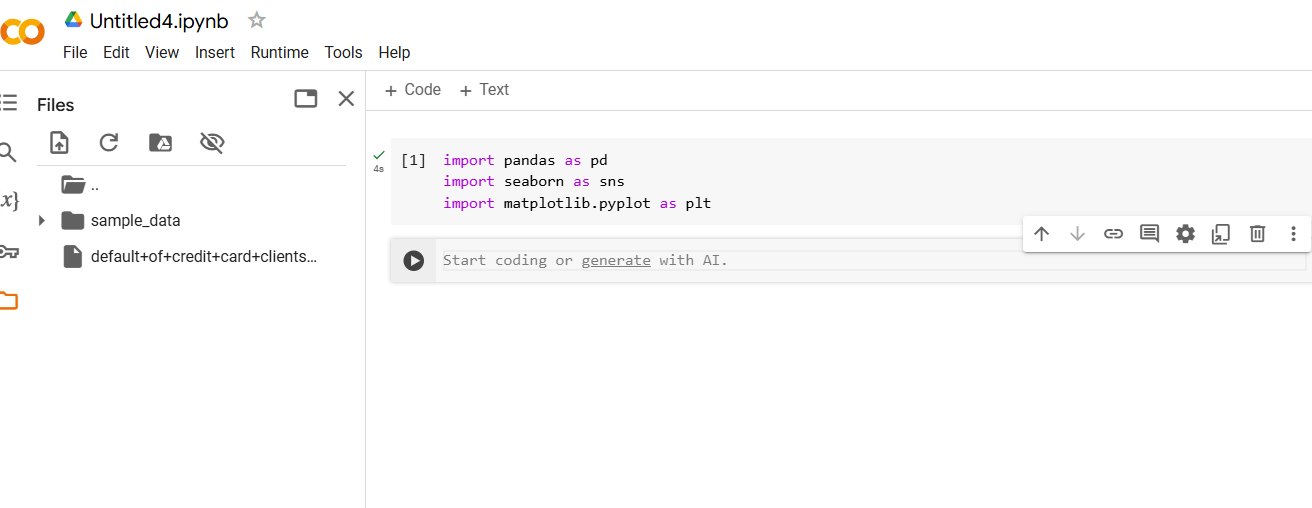
Demographic Features: Weak correlation but potentially valuable in combination.

Payment History (PAY\_0 to PAY\_6): Strong correlation with default.

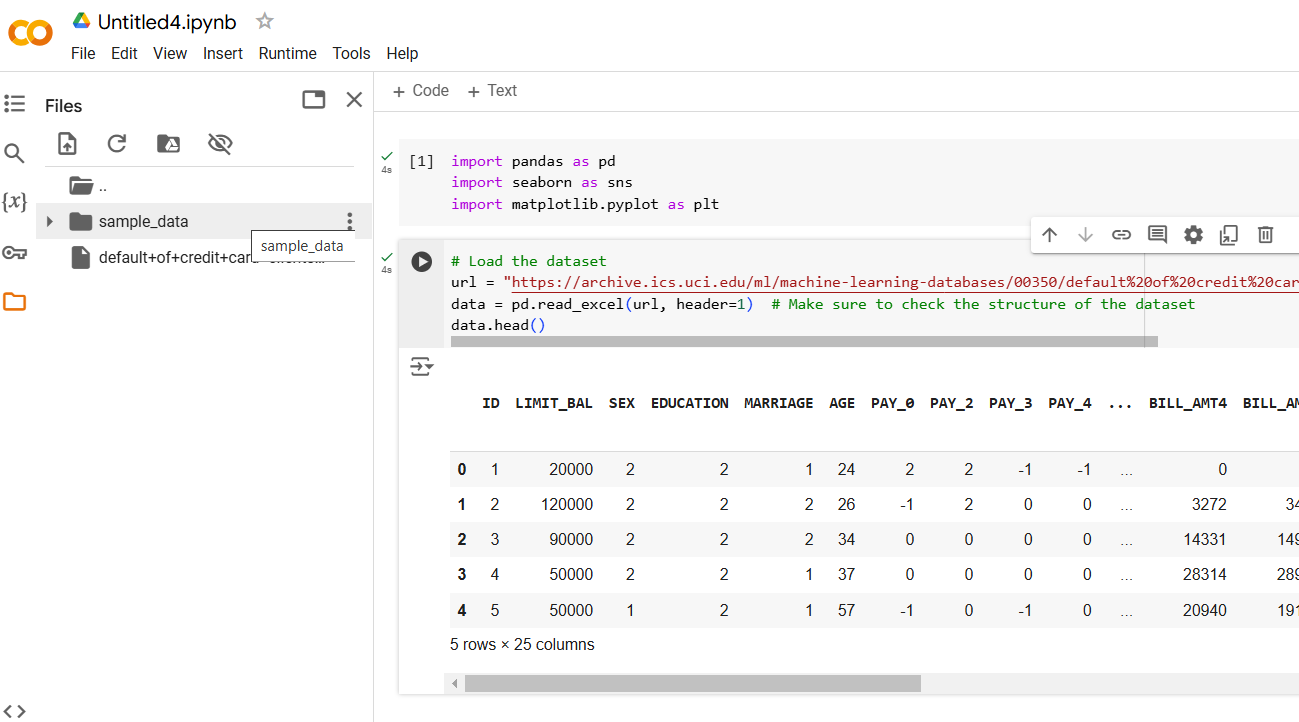
Billing and Payment Amounts: Moderate predictive power; provides insights into financial discipline. Conclusion: Payment history and payment amounts are critical predictors.

* DETALIED STEPS

IMPORTING USEFUL LIBRARIES

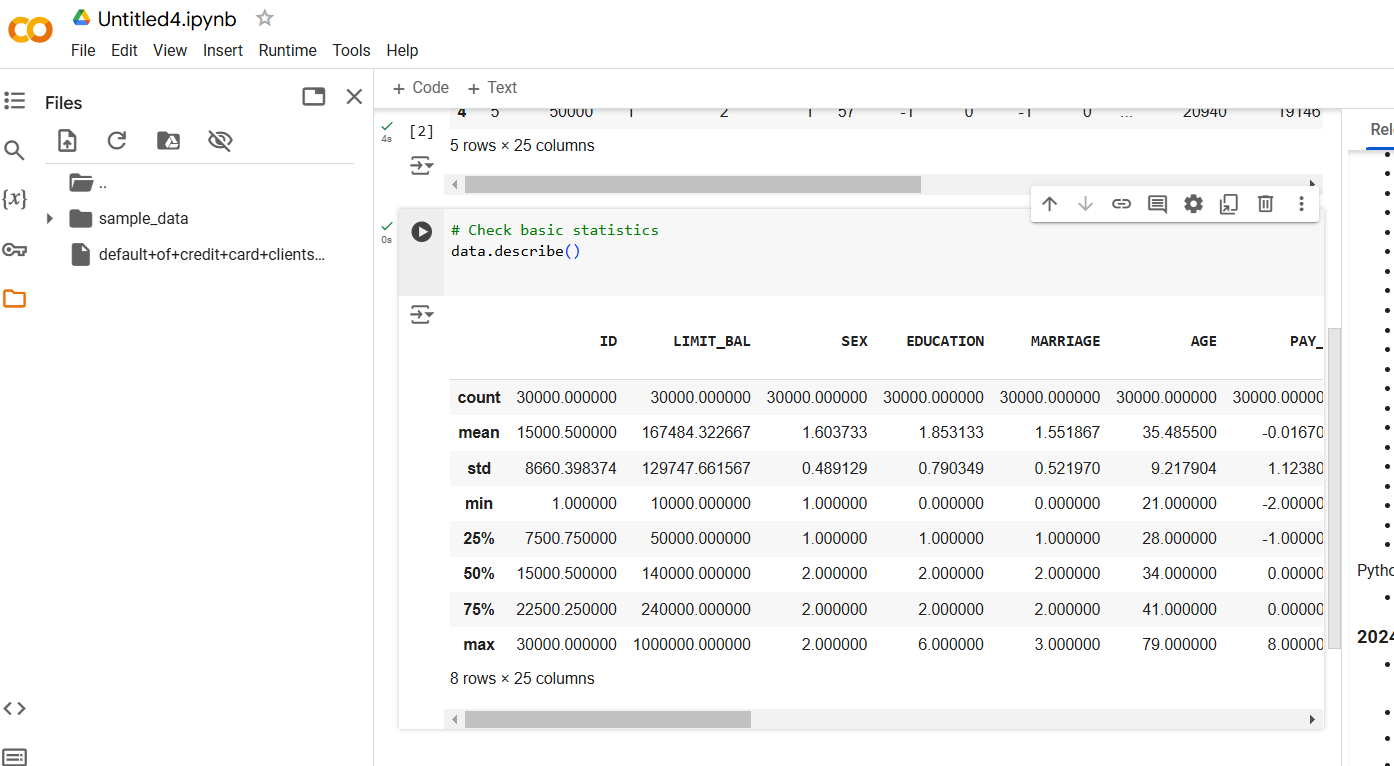


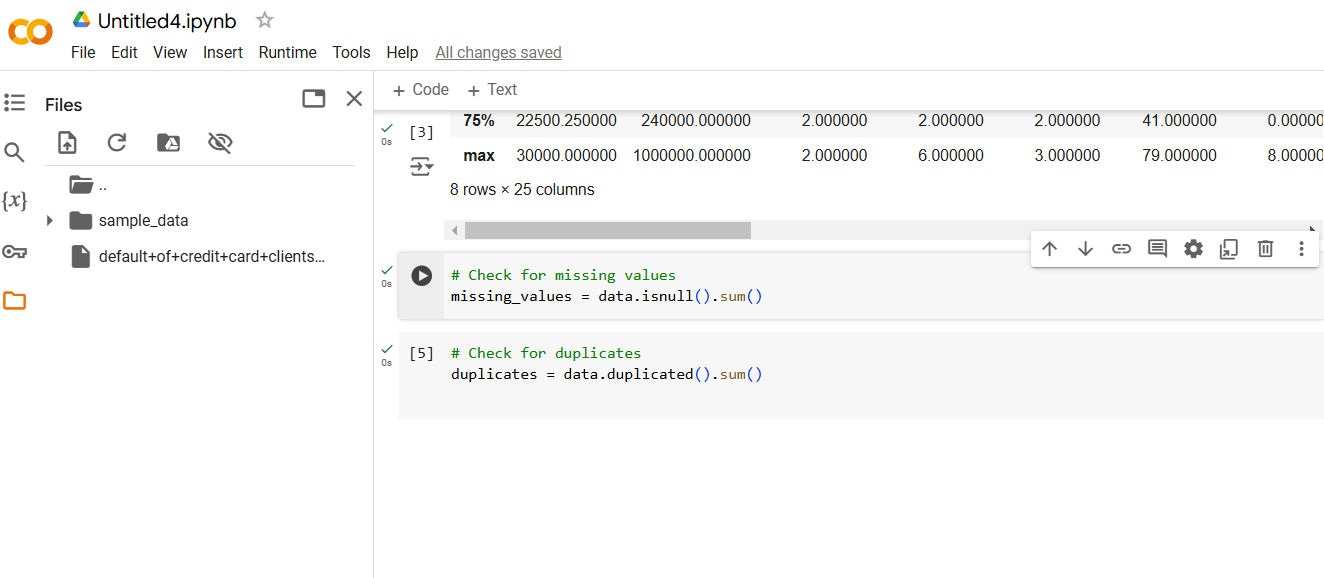
LOAD THE DATASET



DATASET FIRST VIEW

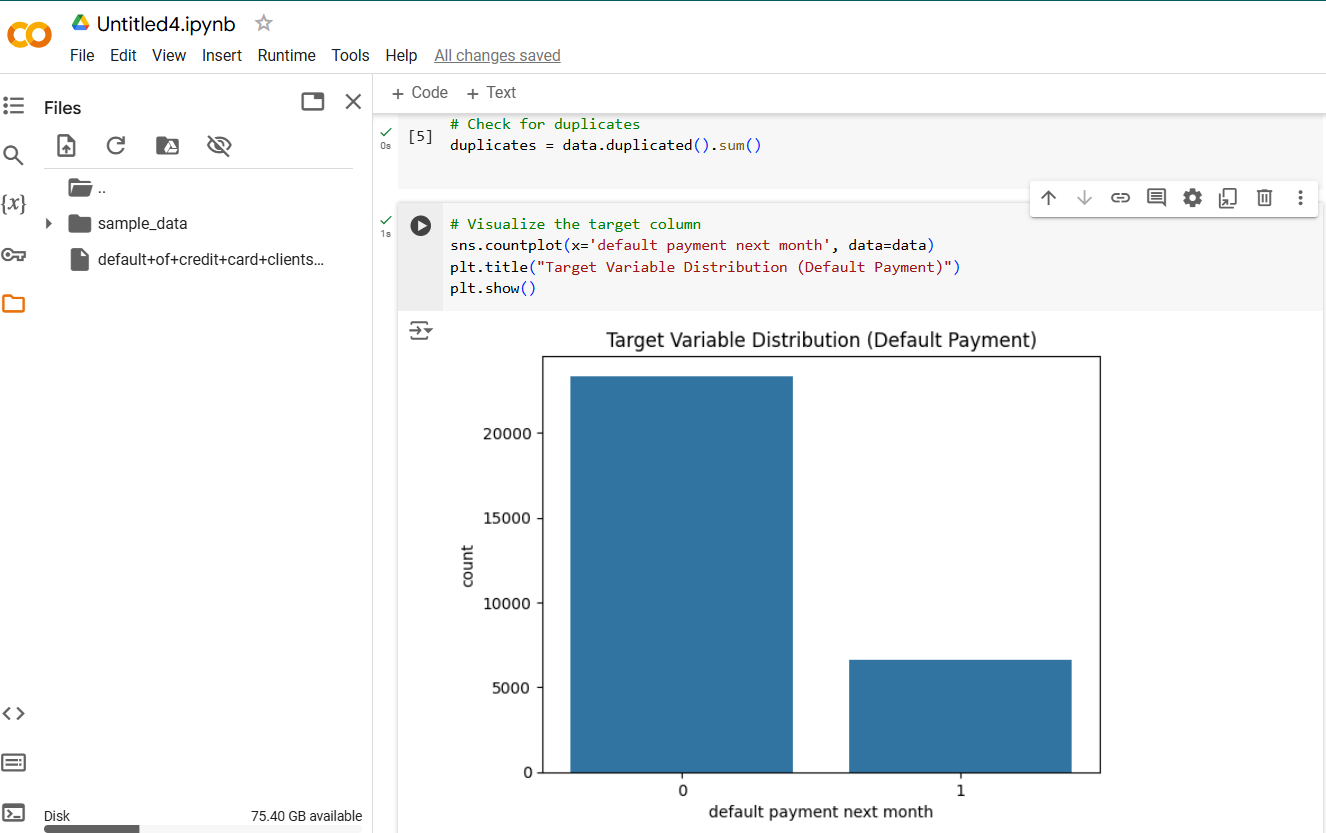
VARIABLES DESCRIPTION



CHECK FOR MISSING VALUES

CHECKING FOR DUPLICATIONS

Data Visualization - Target Variable Distribution



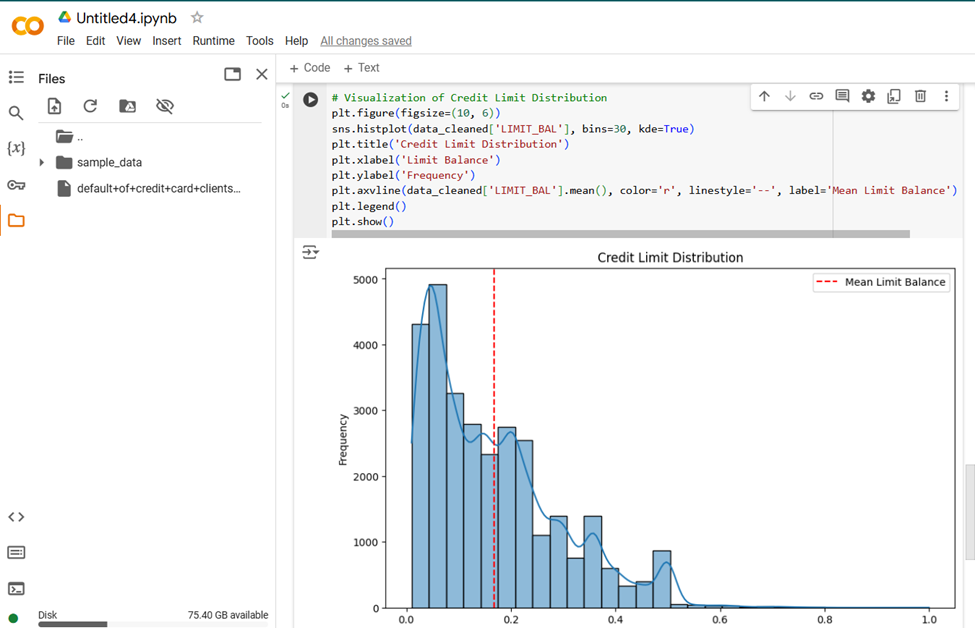
Target Imbalance:

Non-Defaults: ~22,000

Defaults: ~5,000

Challenge: Imbalanced data can skew predictions.

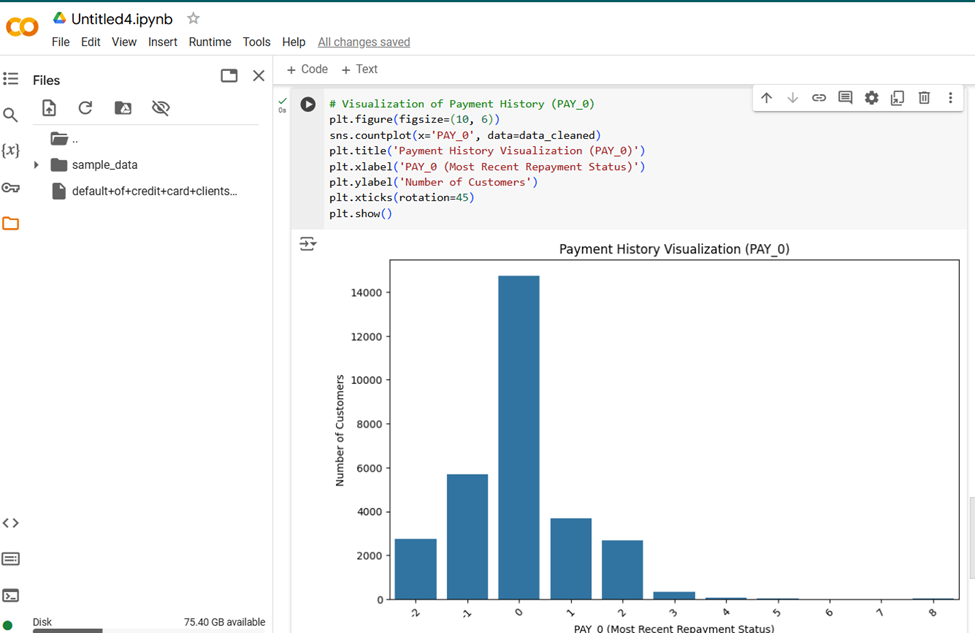
Data Visualization - Credit Limit Distribution



Observation: Most customers have a credit limit below $200,000.

Mean Limit Balance: Slightly above $200,000.

Data Visualization - Payment History

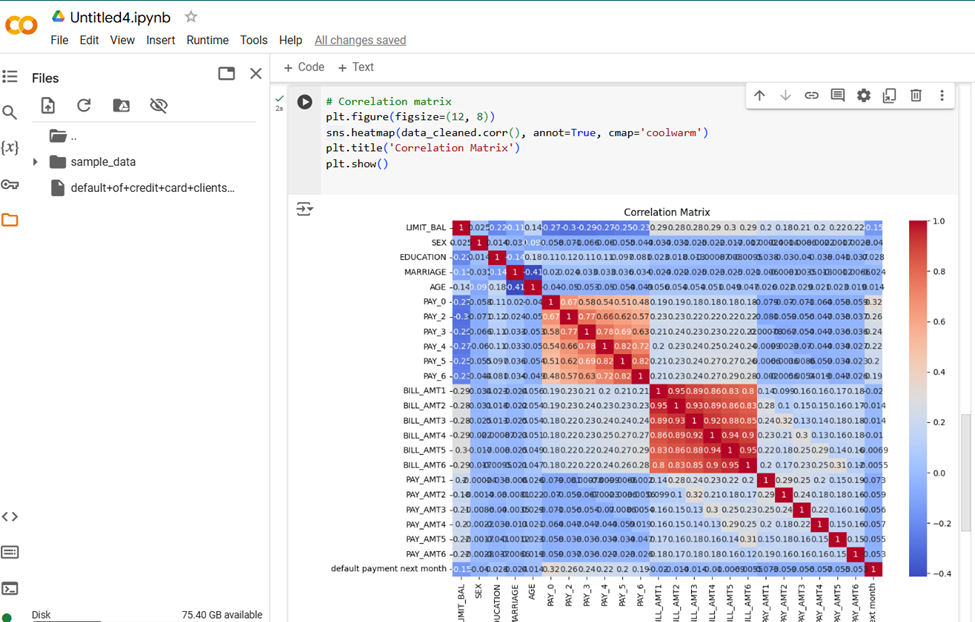


Observation: Majority on-time payments, with fewer delayed payments.

PAY\_0: Indicates repayment status (recent).

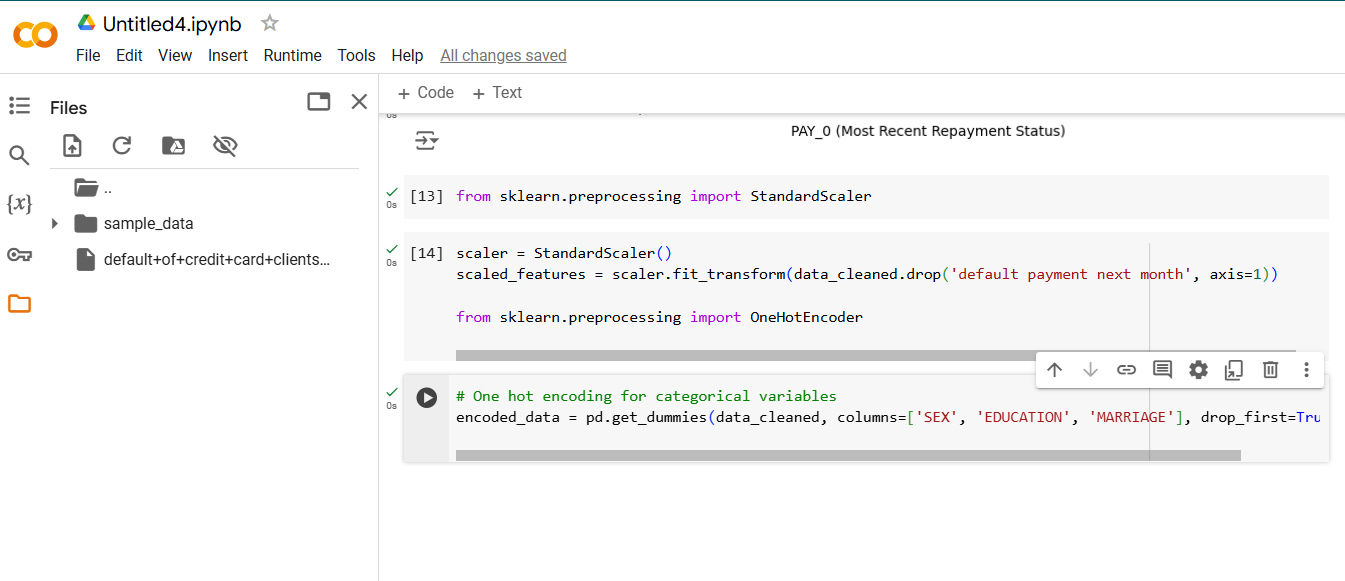
Visual: (Include the countplot for PAY\_0)

Correlation Matrix



Key Correlations:PAY\_0 to PAY\_6 positively correlate with defaulSt.BILL\_AMT1-6 have lower correlations.

Conclusion: Past payment behavior is a strong predictor.

Data Cleaning & Preprocessing

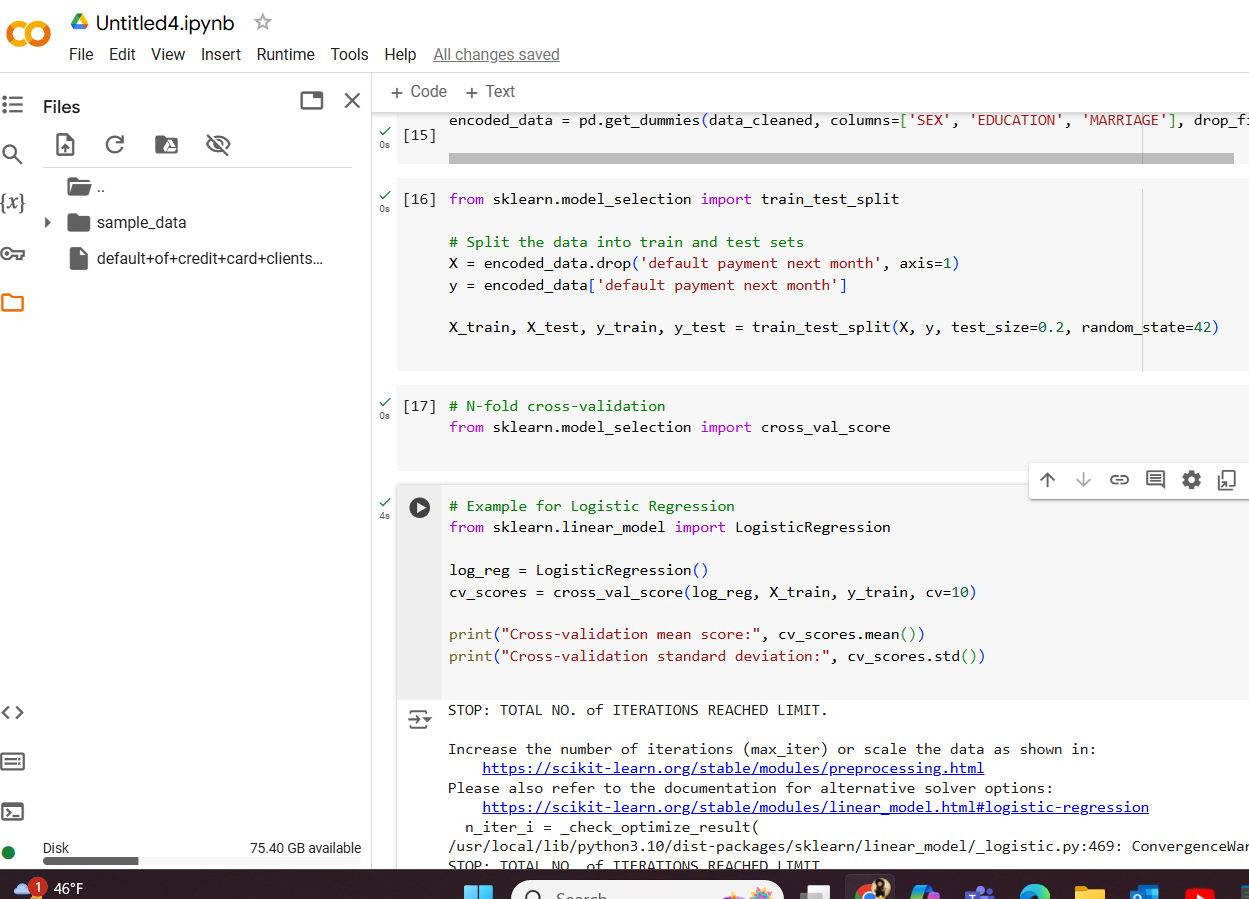
Duplicates: Found and removed

Irrelevant Features: Dropped 'ID' column.

Encoding: Categorical variables like SEX, EDUCATION, and MARRIAGE encoded with one-hot encoding.

Scaling: Applied StandardScaler to LIMIT\_BAL, BILL\_AMT, PAY\_AMT columns.

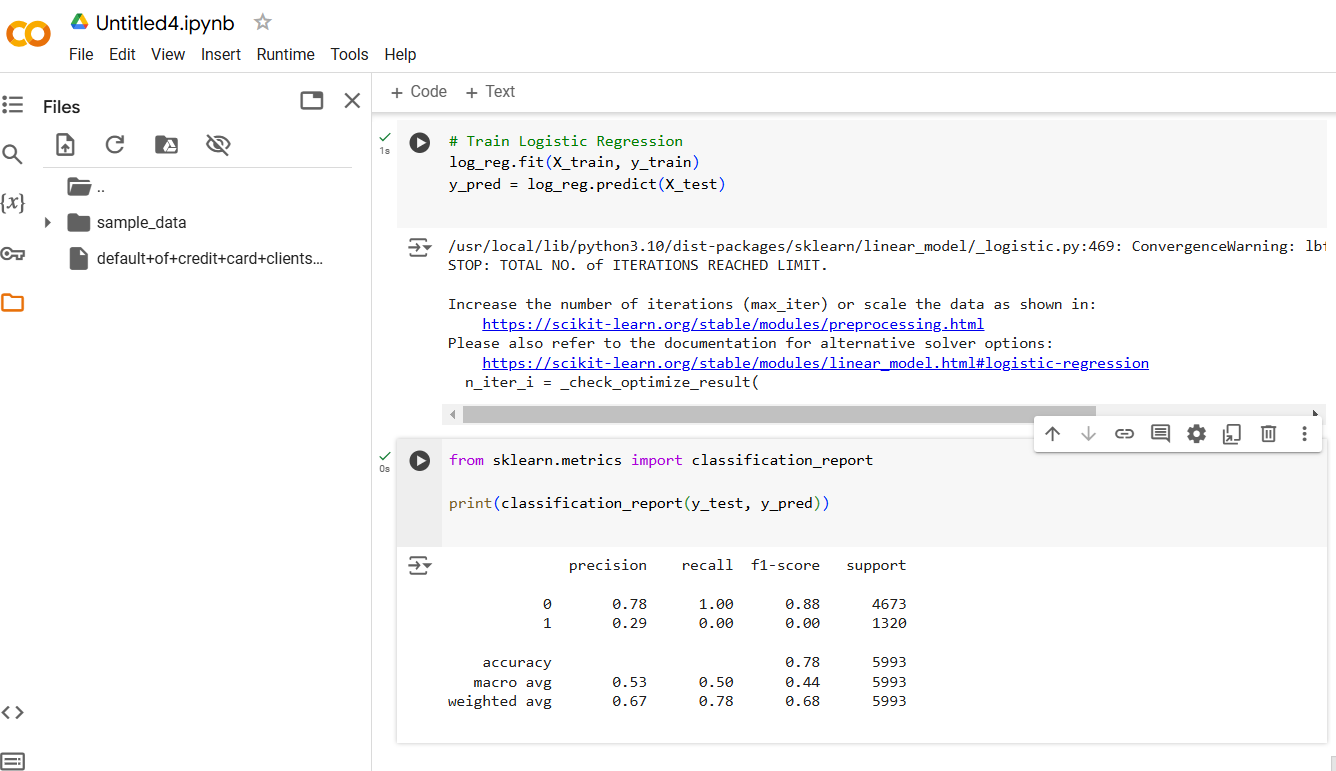
Model Building & Evaluation Plans



Logistic Regression:

Cross-Validation Score: Mean and Standard Deviation.

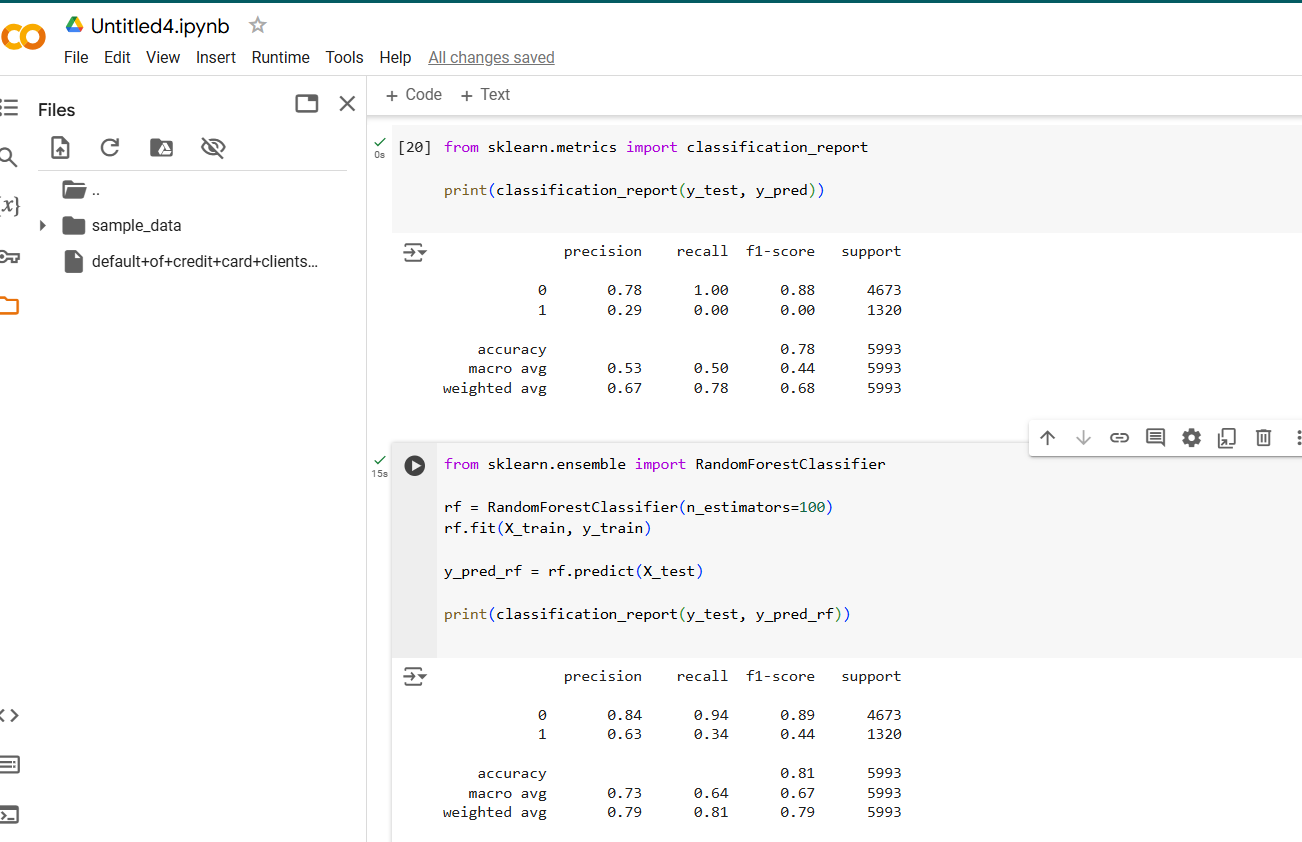
Logistic Regression & Random Forest Results



Logistic Regression:

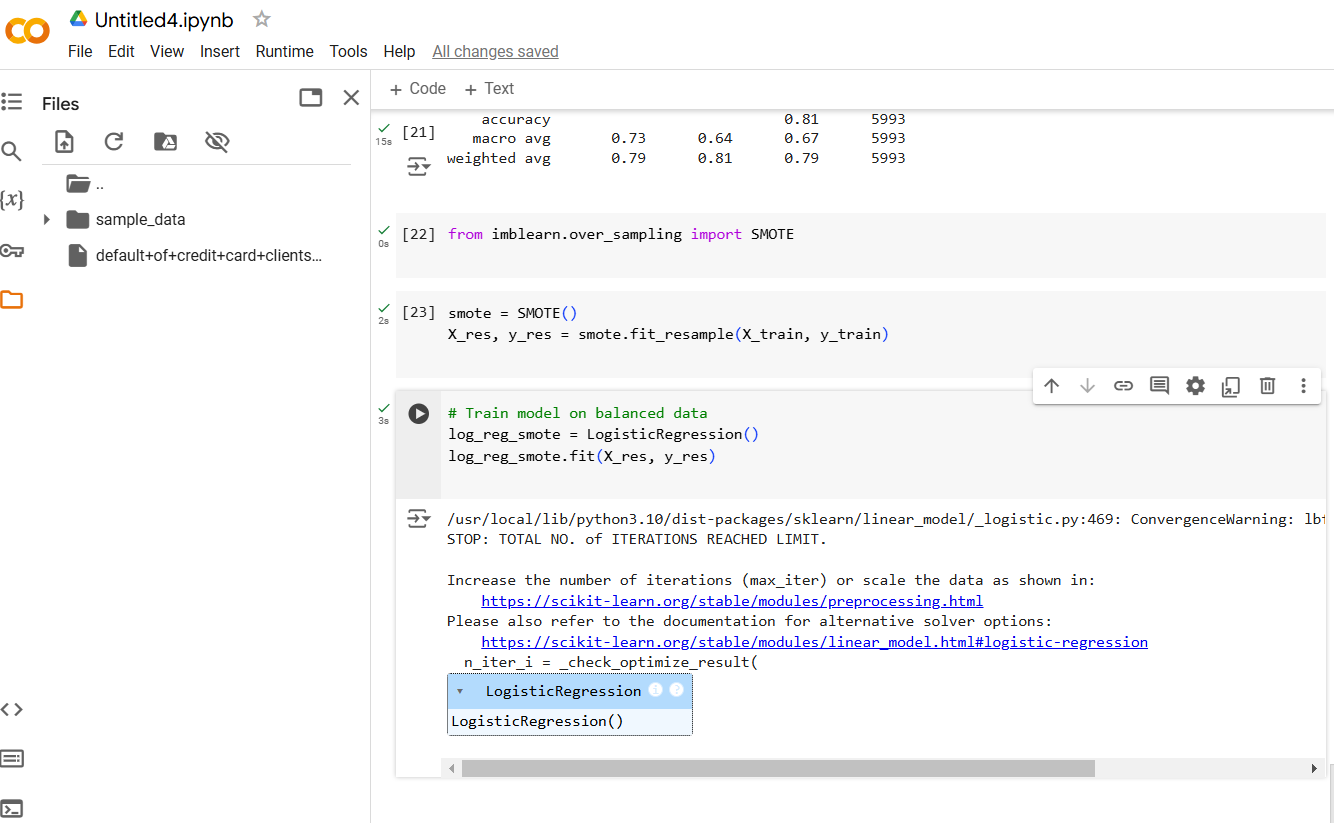
Cross-Validation Score: Mean and Standard Deviation.

Classification Report: Precision, Recall, F1-score.



Random Forest:

Classification Report: Precision, Recall, F1-score.Initial observations and performance comparison.

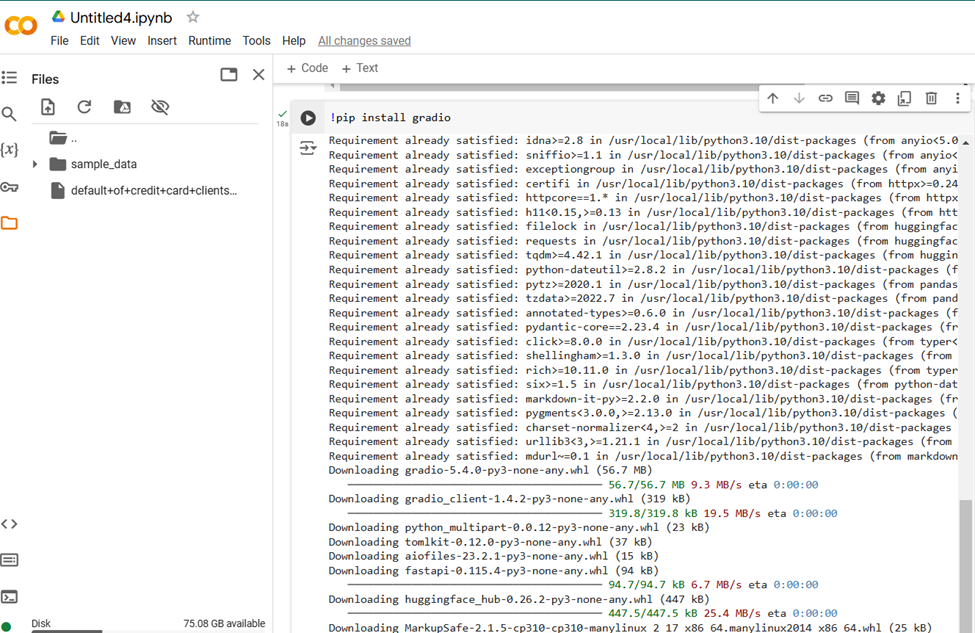


Balancing the Dataset: SMOTE to address class imbalance.

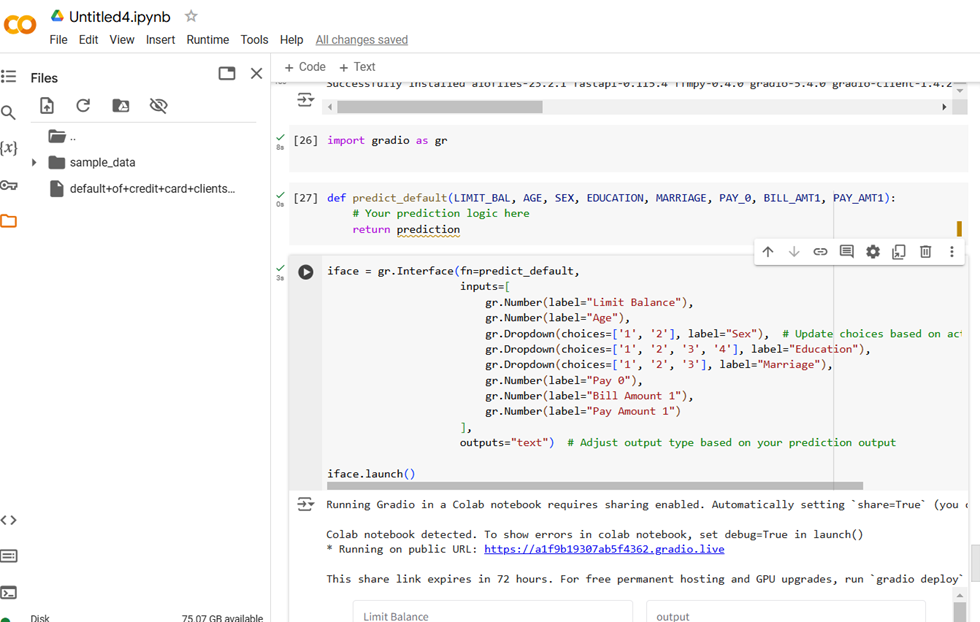
Model Selection: Logistic Regression (interpretability) and Random Forest (complex relationships).

Evaluation Metrics: Precision, Recall, F1-score, ROC-AUC to assess model performance.

INSTALLING GRADIO



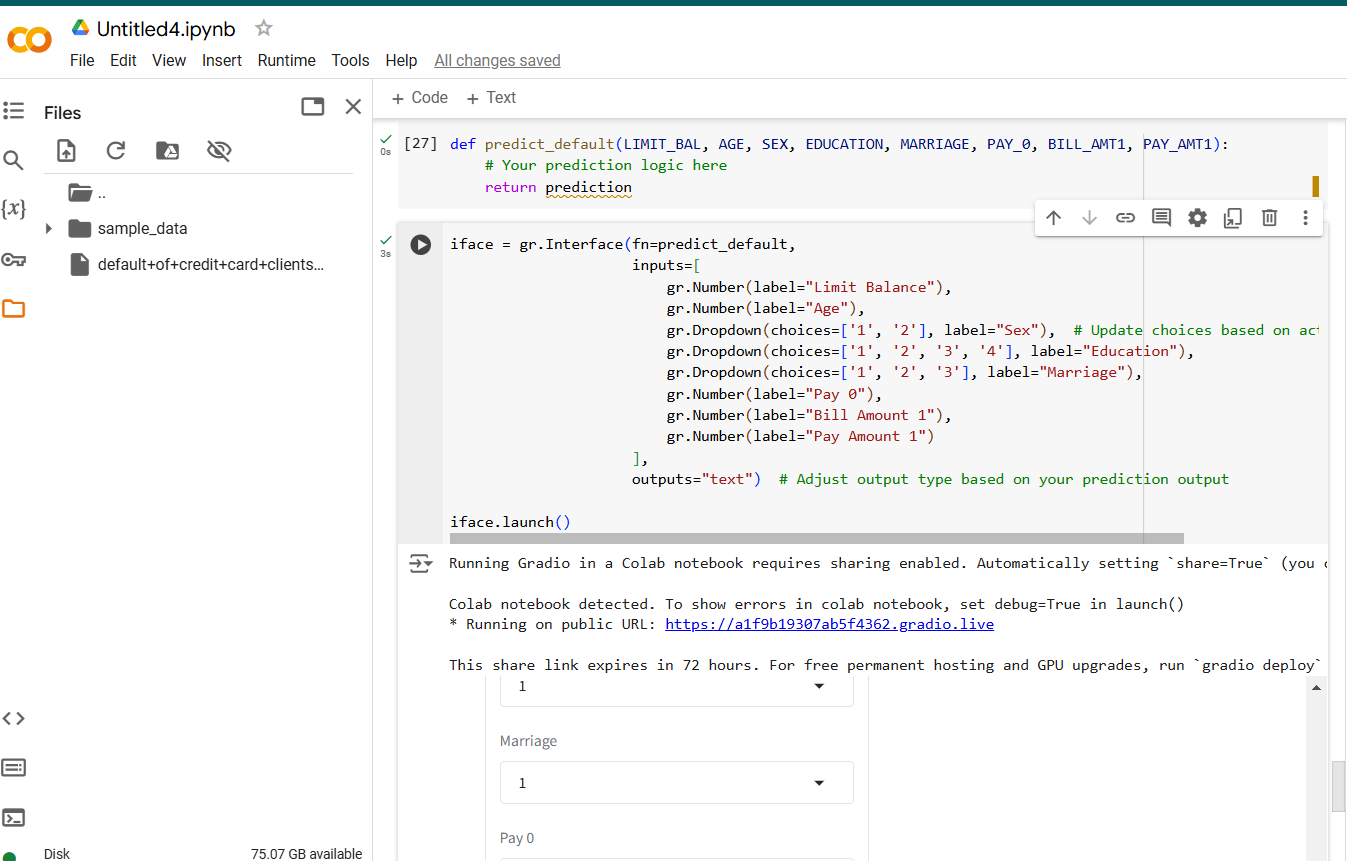
DEPLOYMENT PLAN

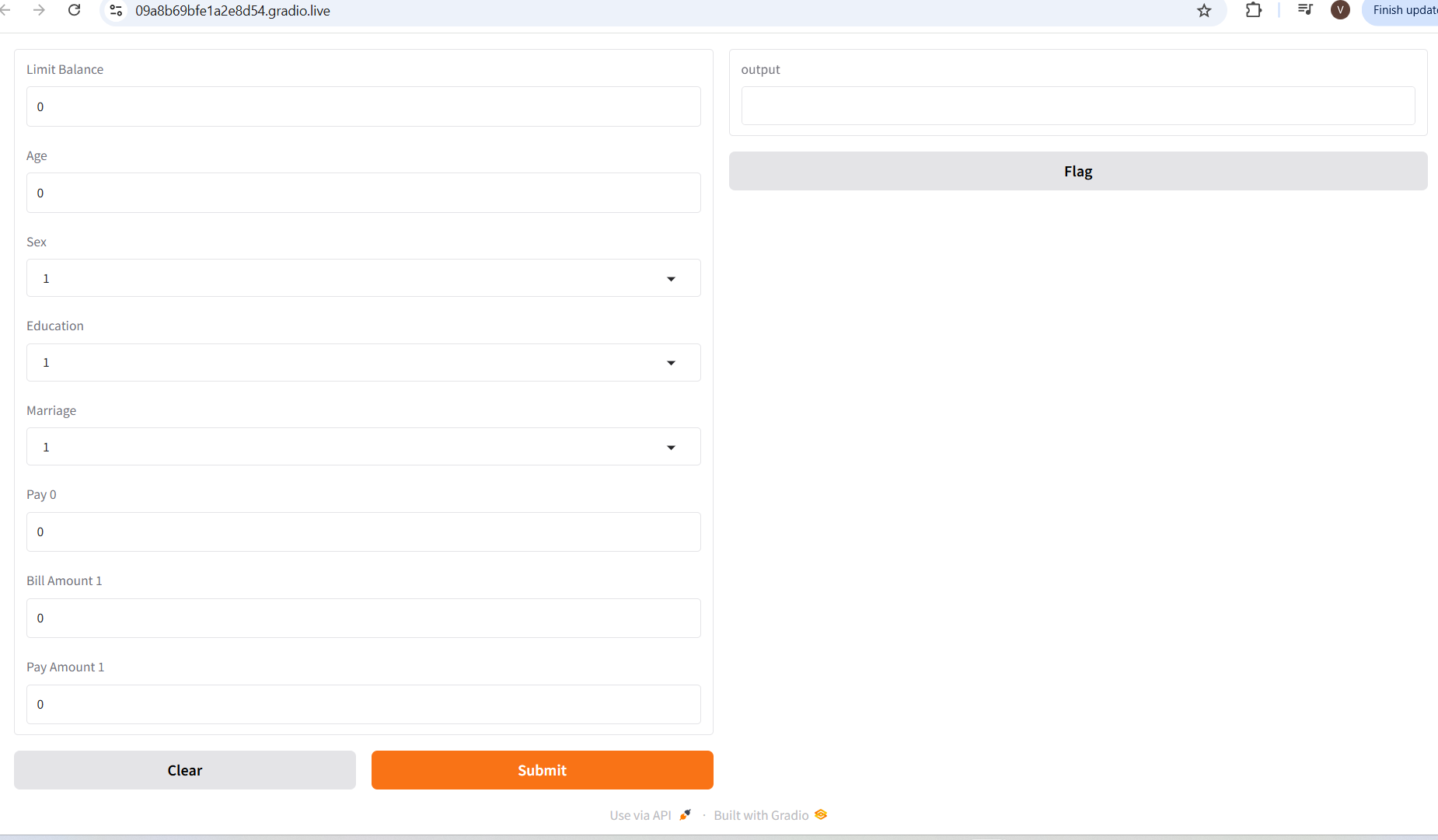


Interface: Gradio for model deployment.

User Inputs:Credit Limit, Age, Sex, Education, Marital Status, Recent Payment Status, Bill Amount, Payment Amount.

Output: Prediction of default status.





Conclusion

Summary: Explored and preprocessed data, identified key features, balanced the dataset, and built predictive models.

Next Steps: Refine models, test additional algorithms, optimize deployment for practical use.