Capstone Project

Credit Card Default Prediction

Presented By:

Aman Verma

Problem Statement:

This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the K-S chart to evaluate which customers will default on their credit card payments

Business Objective: The goal is to predict the likelihood of customers defaulting on their credit card payments in Taiwan. This prediction is essential for risk management purposes.

Significance: By accurately predicting payment default, businesses can proactively identify customers at risk and implement appropriate measures to minimize financial losses and improve customer retention.

Data Overview

Dataset Description: The project utilized a comprehensive dataset containing information on credit card customers. The dataset includes various attributes such as demographic details, payment history, credit limit, and bill amounts for multiple months.

The 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. It consists of 30,000 rows and 25 columns. I observed that there are no duplicate values and misssing values/null values in the dataset. There are 24 Independent features and Default payment next month is the target variable. All independent features are of Int data type. There are categorical variables like Sex, Education and Marriage. Remaining all independent variables are Numerical.

Imbalanced Data: One important characteristic of the dataset is its class imbalance, with a majority of non-defaulters and a minority of defaulters. This required to address the imbalance during the modeling process.

There are 25 variables:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit
- SEX: Gender 1=male, 2=female
- EDUCATION: 1=graduate school, 2=university, 3=high school, 0, 4, 5, 6=others)
- MARRIAGE: Marital status 1=married, 2=single, 3=divorce, 0=others
- AGE: Age in years

History of Past Payments

- PAY_0: Repayment status in September, 2005
- -2: No consumption -1: Paid in full 0: The use of revolving credit 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)

Amount of bill Statement

- BILL_AMT1: Amount of bill statement in September, 2005
- BILL_AMT2: Amount of bill statement in August, 2005
- BILL_AMT3: Amount of bill statement in July, 2005
- BILL_AMT4: Amount of bill statement in June, 2005
- BILL_AMT5: Amount of bill statement in May, 2005
- BILL_AMT6: Amount of bill statement in April, 2005

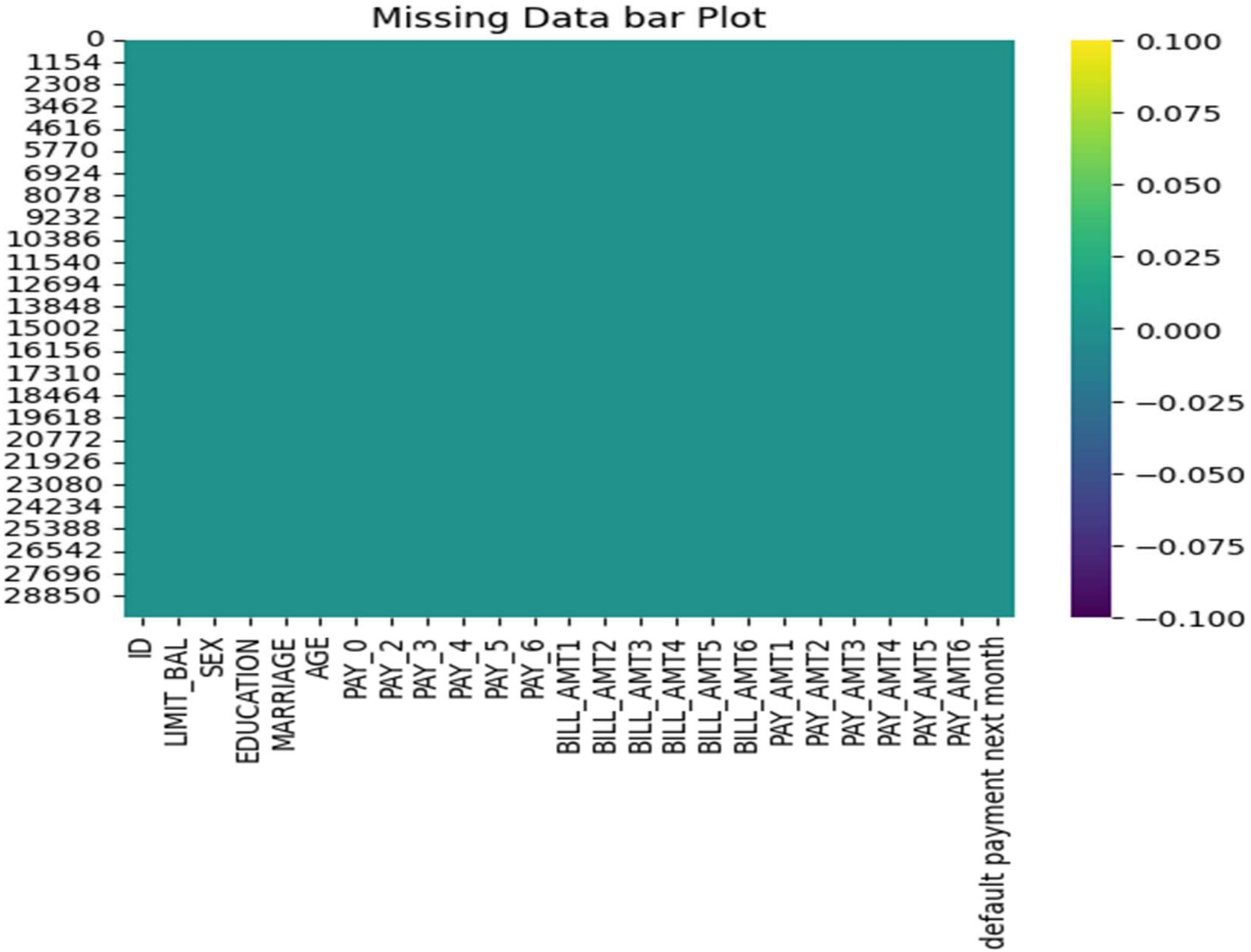
Amount of Previous Payments -Previous amount Paid

- PAY_AMT1: Amount of previous payment in September, 2005
- PAY_AMT2: Amount of previous payment in August, 2005
- PAY_AMT3: Amount of previous payment in July, 2005
- PAY_AMT4: Amount of previous payment in June, 2005
- PAY_AMT5: Amount of previous payment in May, 2005
- PAY_AMT6: Amount of previous payment in April, 2005

default payment next month: Default payment
1=yes, 0=no

Missing Values/Null Values

There are no misssing values/null values in the dataset.

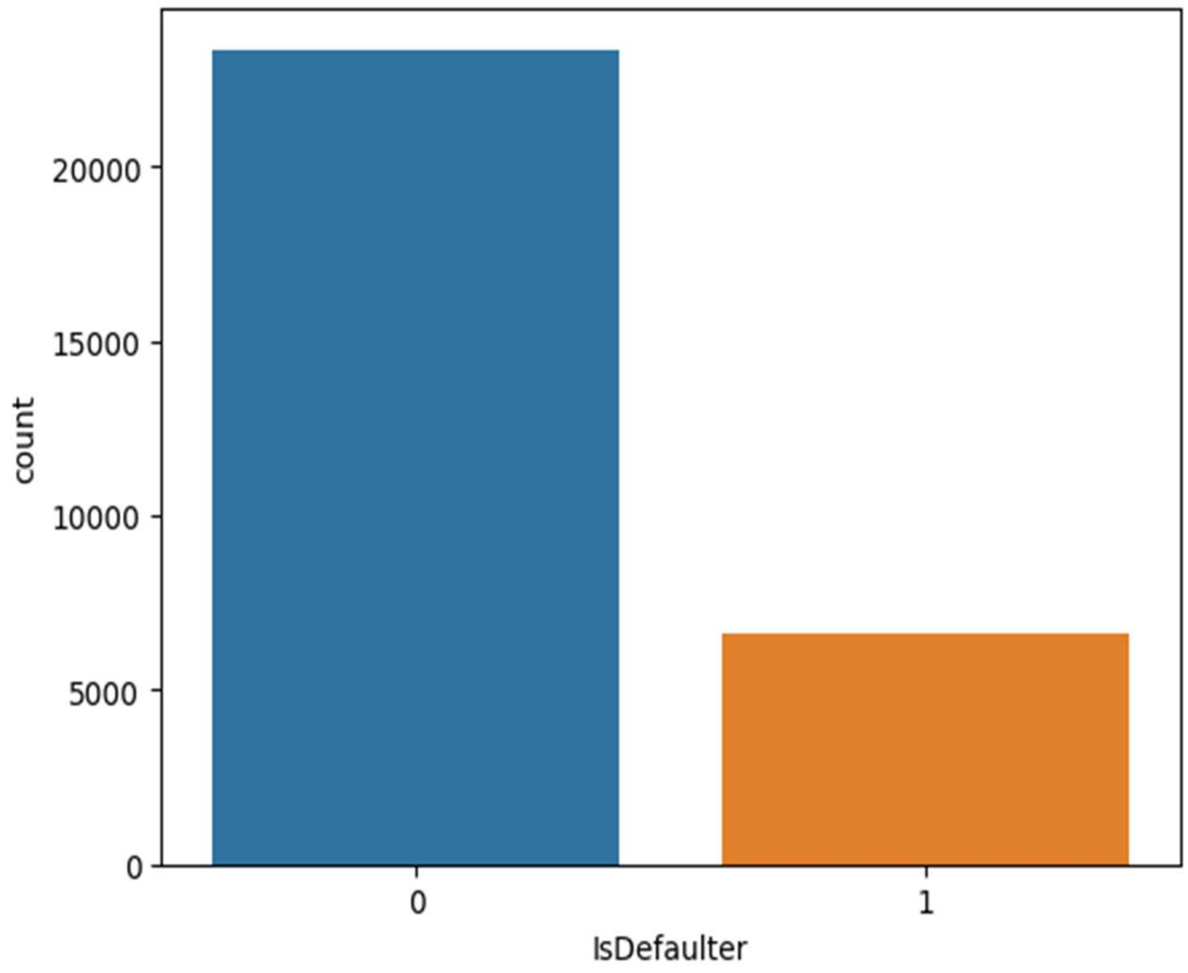


Exploratory Data Analysis

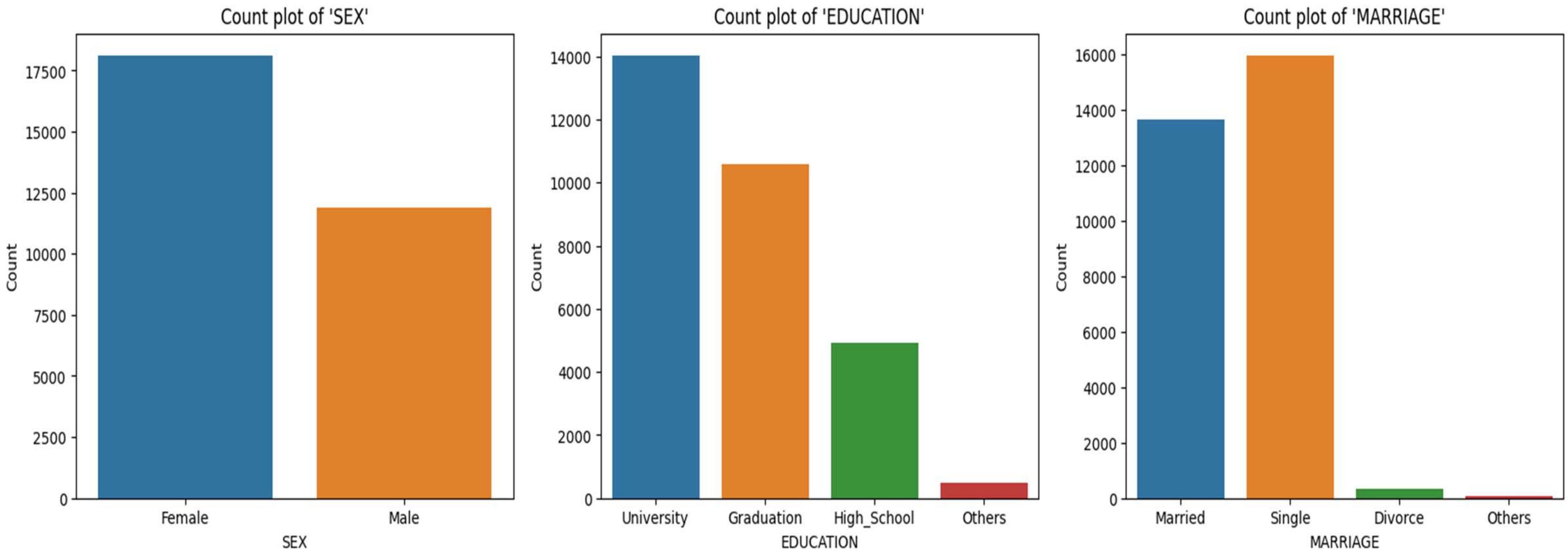
What demographic factors impact payment default risk?

Univariate Analyses

- Both the classes are not in proportion.
- Which means that the dataset is imbalanced.
- Data balancing is required.

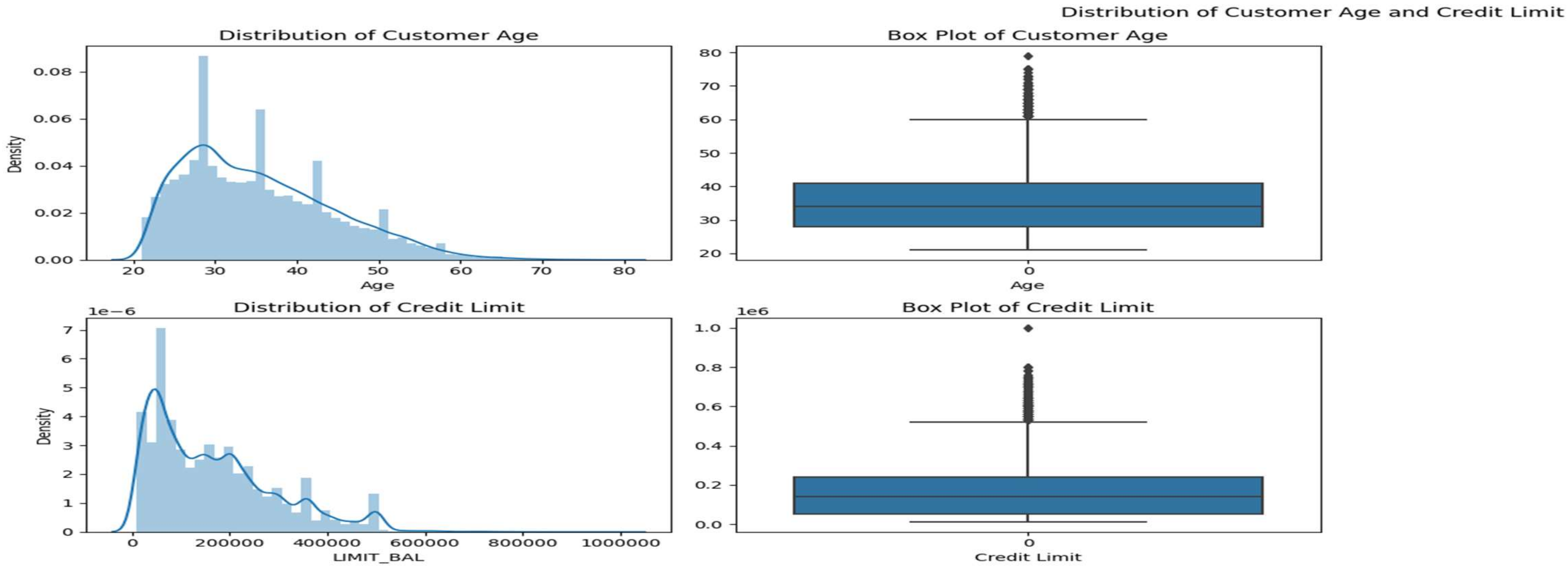


Exploring Categorical Variables



1.Females are utilizing credit card facility more than males
2.Education level as university and graduation are more than other types.
3.Divorced people and others are less likely to utilize or using credit cards

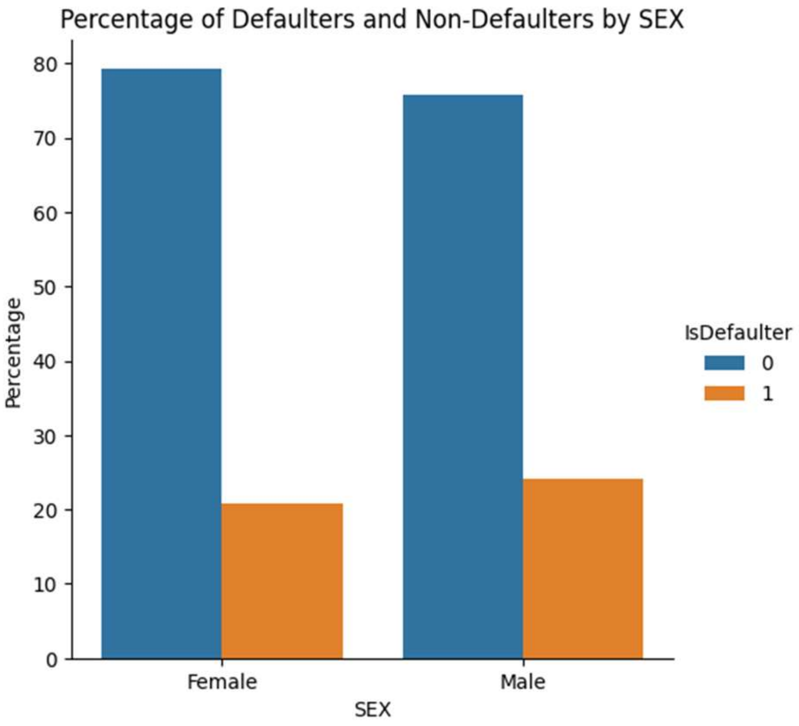
Visualizing Age and Credit Limit Distributions

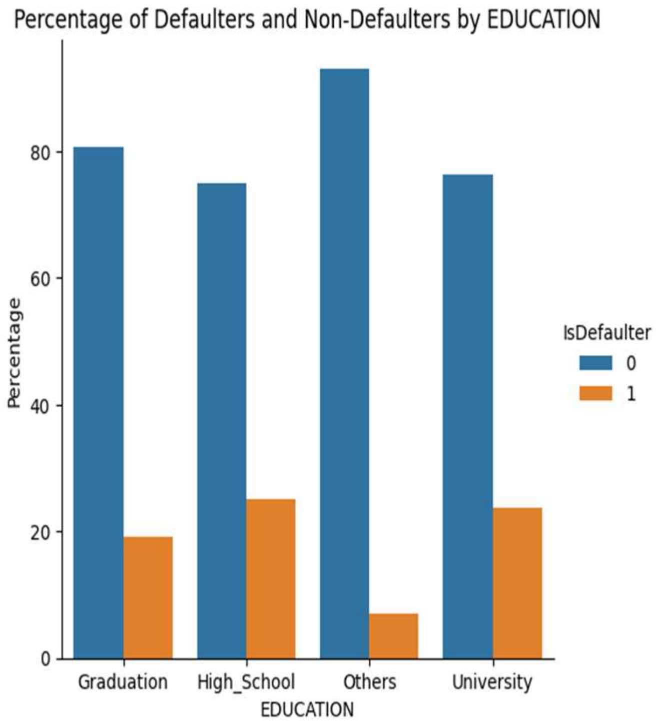


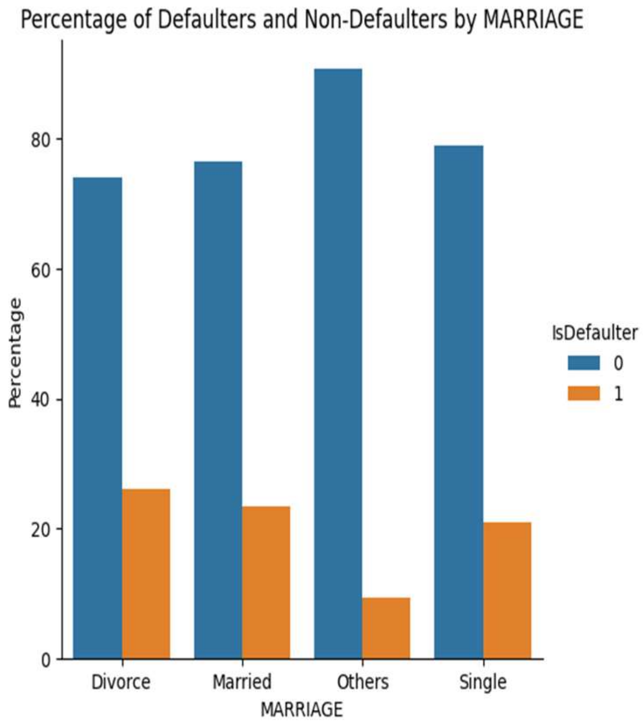
For 'AGE': The distribution plot shows the age distribution of customers, indicating the most common age range and any potential skewness in the data. The box plot provides information about the median, quartiles, and potential outliers, which may be useful for identifying age-related patterns in credit defaulters.

For 'LIMIT_BAL': The box plot helps identify any potential outliers and the spread of credit limits, which can help in identifying customers with higher or lower credit limits.

Visualizing Default Rates by Category

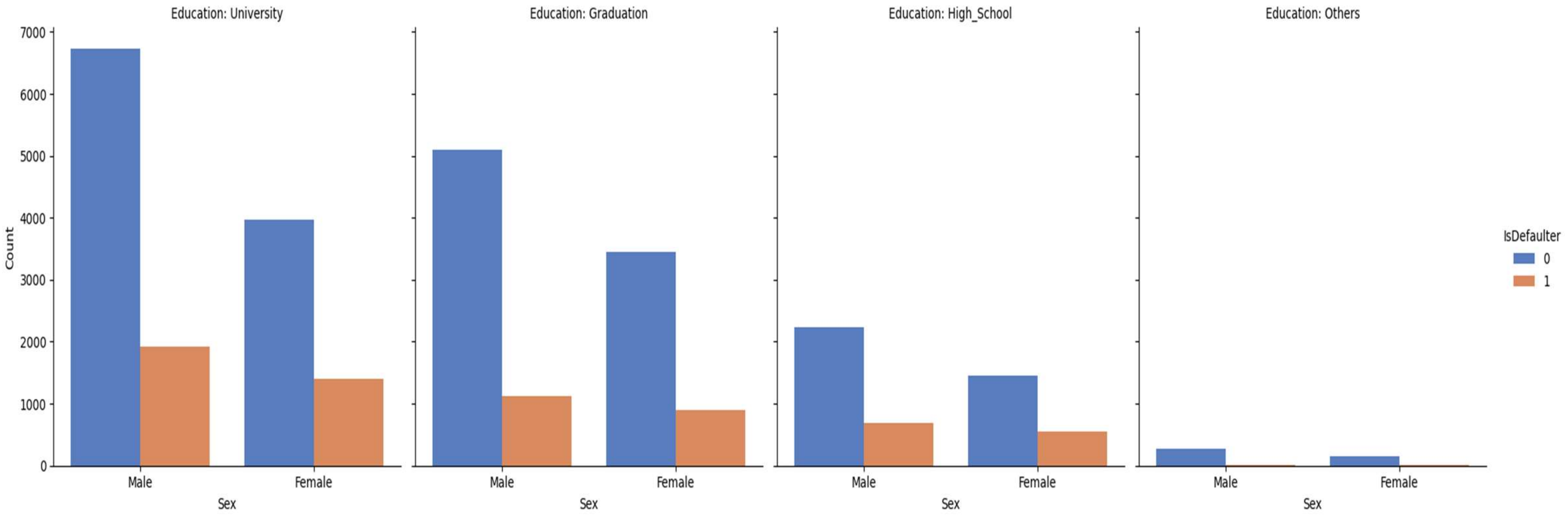

Percentage of Defaulters and Non-Defaulters by SEX


Percentage of Defaulters and Non-Defaulters by EDUCATION


Percentage of Defaulters and Non-Defaulters by MARRIAGE

Gender-wise, males exhibit a higher default rate compared to females.
Education level plays a role, with higher default rates observed among high school and university students.
Marital status influences default behavior, with married and divorced individuals being more prone to default.

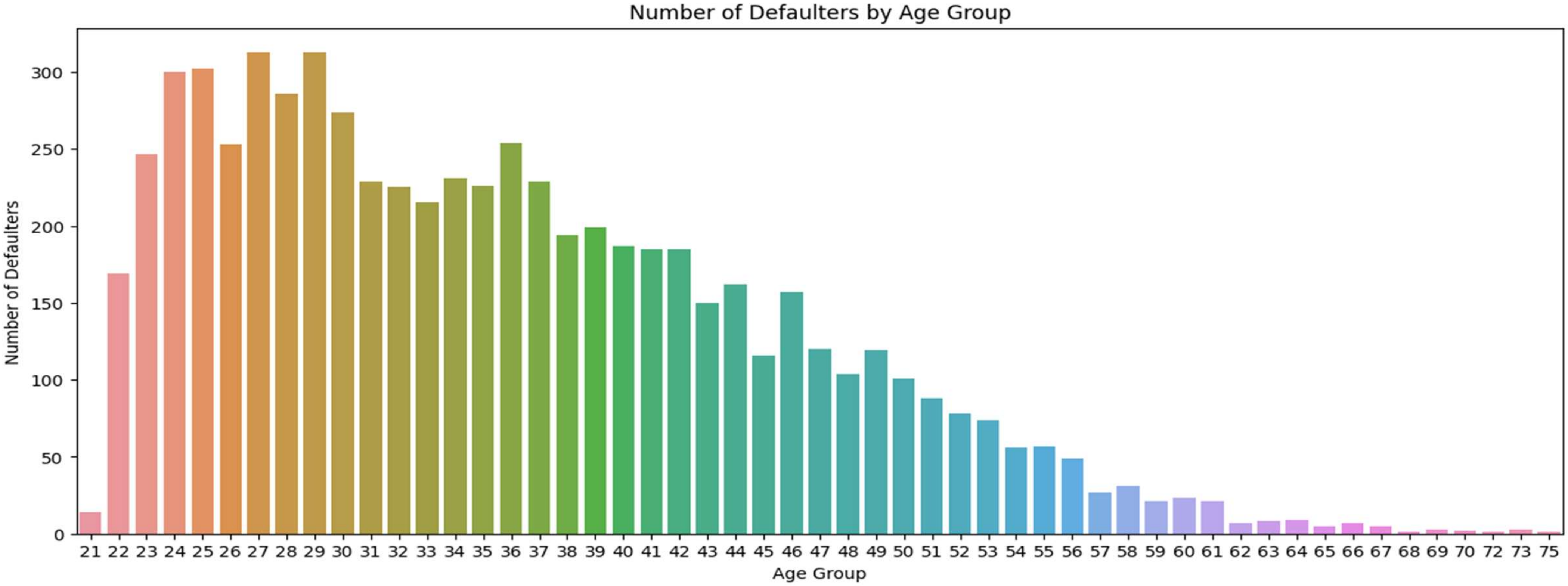
Exploring Interactions between Education Levels, Gender, and Defaults



we can observe which combinations of 'SEX' and 'EDUCATION' have a higher or lower proportion of defaulters, allowing us to identify potential patterns or trends in credit default behavior among different groups.

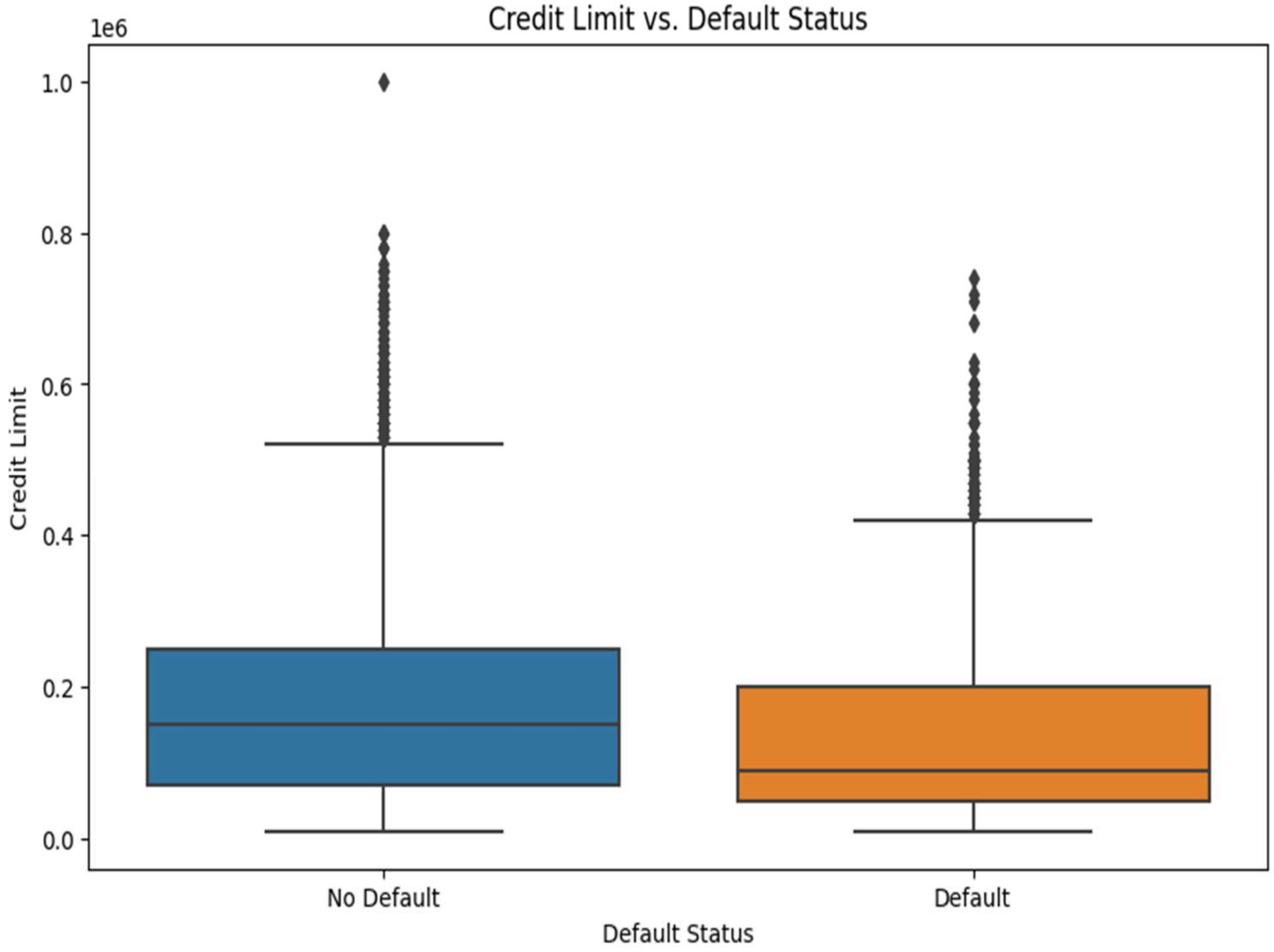
For a particular combination of 'SEX' and 'EDUCATION' like university and males which has a significantly higher proportion of defaulters

Analyzing Credit Defaulters by Age

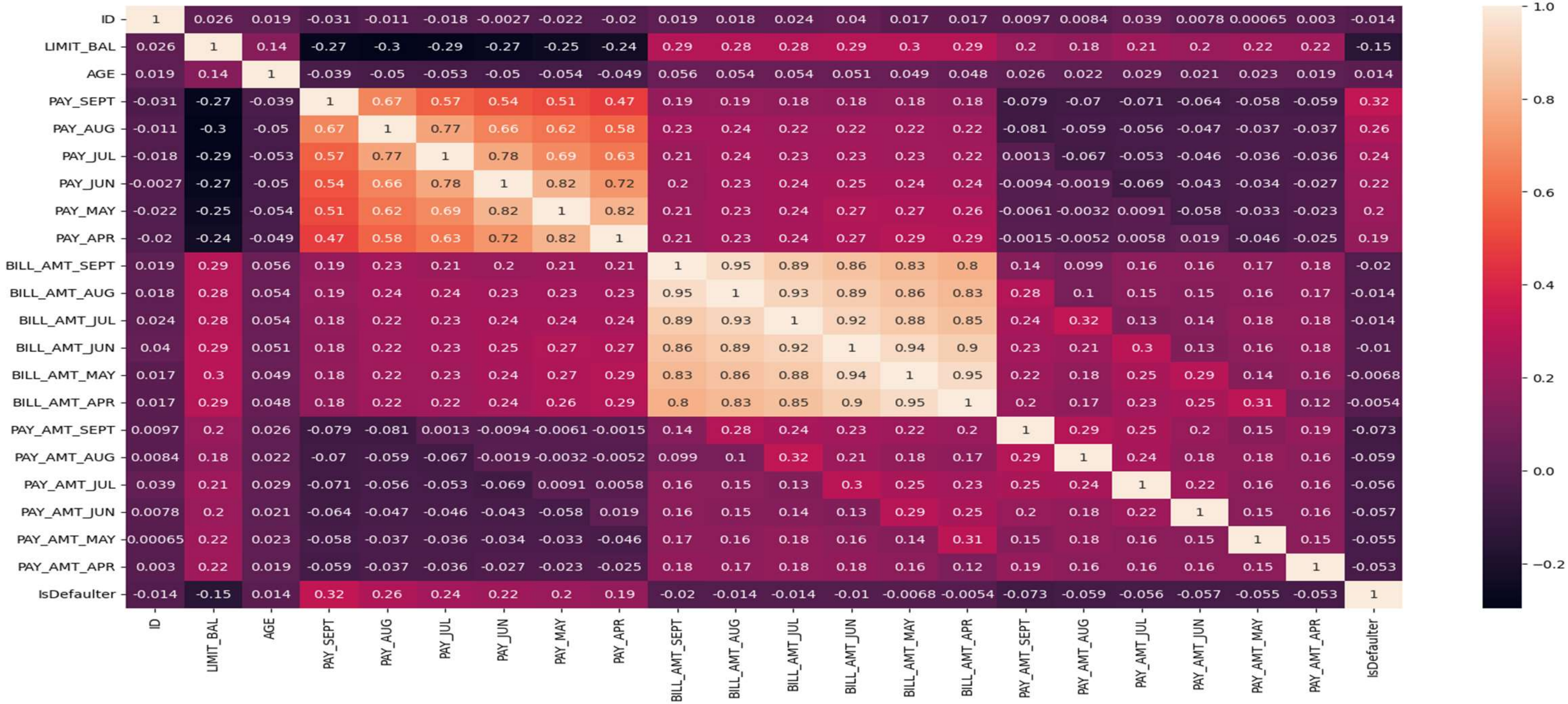


More number of defaulters lies in age 20-42

Credit Limit Impact on Default Status



- Higher credit limits,
- lower default risk.



	ID	LIMIT_BAL	AGE	PAY_SEPT	PAY_AUG	PAY_JUL	PAY_JUN	PAY_MAY	PAY_APR	BILL_AMT_SEPT	BILL_AMT_AUG	BILL_AMT_JUL	BILL_AMT_JUN	BILL_AMT_MAY	BILL_AMT_APR	PAY_AMT_SEPT	PAY_AMT_AUG	PAY_AMT_JUL	PAY_AMT_JUN	PAY_AMT_MAY	PAY_AMT_APR	IsDefaulter
ID	1	0.026	0.019	-0.031	-0.011	-0.018	-0.0027	-0.022	-0.02	0.019	0.018	0.024	0.04	0.017	0.017	0.0097	0.0084	0.039	0.0078	0.00065	0.003	-0.014
LIMIT_BAL	0.026	1	0.14	-0.27	-0.3	-0.29	-0.27	-0.25	-0.24	0.29	0.28	0.28	0.29	0.3	0.29	0.2	0.18	0.21	0.2	0.22	0.22	-0.15
AGE	0.019	0.14	1	-0.039	-0.05	-0.053	-0.05	-0.054	-0.049	0.056	0.054	0.054	0.051	0.049	0.048	0.026	0.022	0.029	0.021	0.023	0.019	0.014
PAY_SEPT	-0.031	-0.27	-0.039	1	0.67	0.57	0.54	0.51	0.47	0.19	0.19	0.18	0.18	0.18	0.18	-0.079	-0.07	-0.071	-0.064	-0.058	-0.059	0.32
PAY_AUG	-0.011	-0.3	-0.05	0.67	1	0.77	0.66	0.62	0.58	0.23	0.24	0.22	0.22	0.22	0.22	-0.081	-0.059	-0.056	-0.047	-0.037	-0.037	0.26
PAY_JUL	-0.018	-0.29	-0.053	0.57	0.77	1	0.78	0.69	0.63	0.21	0.24	0.23	0.23	0.23	0.22	0.0013	-0.067	-0.053	-0.046	-0.036	-0.036	0.24
PAY_JUN	-0.0027	-0.27	-0.05	0.54	0.66	0.78	1	0.82	0.72	0.2	0.23	0.24	0.25	0.24	0.24	-0.0094	-0.0019	-0.069	-0.043	-0.034	-0.027	0.22
PAY_MAY	-0.022	-0.25	-0.054	0.51	0.62	0.69	0.82	1	0.82	0.21	0.23	0.24	0.27	0.27	0.26	-0.0061	-0.0032	0.0091	-0.058	-0.033	-0.023	0.2
PAY_APR	-0.02	-0.24	-0.049	0.47	0.58	0.63	0.72	0.82	1	0.21	0.23	0.24	0.27	0.29	0.29	-0.0015	-0.0052	0.0058	0.019	-0.046	-0.025	0.19
BILL_AMT_SEPT	0.019	0.29	0.056	0.19	0.23	0.21	0.2	0.21	0.21	1	0.95	0.89	0.86	0.83	0.8	0.14	0.099	0.16	0.16	0.17	0.18	-0.02
BILL_AMT_AUG	0.018	0.28	0.054	0.19	0.24	0.24	0.23	0.23	0.23	0.95	1	0.93	0.89	0.86	0.83	0.28	0.1	0.15	0.15	0.16	0.17	-0.014
BILL_AMT_JUL	0.024	0.28	0.054	0.18	0.22	0.23	0.24	0.24	0.24	0.89	0.93	1	0.92	0.88	0.85	0.24	0.32	0.13	0.14	0.18	0.18	-0.014
BILL_AMT_JUN	0.04	0.29	0.051	0.18	0.22	0.23	0.25	0.27	0.27	0.86	0.89	0.92	1	0.94	0.9	0.23	0.21	0.3	0.13	0.16	0.18	-0.01
BILL_AMT_MAY	0.017	0.3	0.049	0.18	0.22	0.23	0.24	0.27	0.29	0.83	0.86	0.88	0.94	1	0.95	0.22	0.18	0.25	0.29	0.14	0.16	-0.0068
BILL_AMT_APR	0.017	0.29	0.048	0.18	0.22	0.22	0.24	0.26	0.29	0.8	0.83	0.85	0.9	0.95	1	0.2	0.17	0.23	0.25	0.31	0.12	-0.0054
PAY_AMT_SEPT	0.0097	0.2	0.026	-0.079	-0.081	0.0013	-0.0094	-0.0061	-0.0015	0.14	0.28	0.24	0.23	0.22	0.2	1	0.29	0.25	0.2	0.15	0.19	-0.073
PAY_AMT_AUG	0.0084	0.18	0.022	-0.07	-0.059	-0.067	-0.0019	-0.0032	-0.0052	0.099	0.1	0.32	0.21	0.18	0.17	0.29	1	0.24	0.18	0.18	0.16	-0.059
PAY_AMT_JUL	0.039	0.21	0.029	-0.071	-0.056	-0.053	-0.069	0.0091	0.0058	0.16	0.15	0.13	0.3	0.25	0.23	0.25	0.24	1	0.22	0.16	0.16	-0.056
PAY_AMT_JUN	0.0078	0.2	0.021	-0.064	-0.047	-0.046	-0.043	-0.058	0.019	0.16	0.15	0.14	0.13	0.29	0.25	0.2	0.18	0.22	1	0.15	0.16	-0.057
PAY_AMT_MAY	0.00065	0.22	0.023	-0.058	-0.037	-0.036	-0.034	-0.033	-0.046	0.17	0.16	0.18	0.16	0.14	0.31	0.15	0.18	0.16	0.15	1	0.15	-0.055
PAY_AMT_APR	0.003	0.22	0.019	-0.059	-0.037	-0.036	-0.027	-0.023	-0.025	0.18	0.17	0.18	0.18	0.16	0.12	0.19	0.16	0.16	0.16	0.15	1	-0.053
IsDefaulter	-0.014	-0.15	0.014	0.32	0.26	0.24	0.22	0.2	0.19	-0.02	-0.014	-0.014	-0.01	-0.0068	-0.0054	-0.073	-0.059	-0.056	-0.057	-0.055	-0.053	1

The repayment status variables (PAY_SEPT to PAY_APR) are positively correlated with each other, indicating that if a customer was delayed in one month, they were likely to be delayed in subsequent months as well.

The amount of bill statements (BILL_AMT_SEPT to BILL_AMT_APR) are positively correlated with each other, suggesting that customers with higher bill amounts in one month tend to have higher bill amounts in other months too.

EDA Summary

- Demographic factorsthat impact default risk are:
- Education: Higher education is associated with lower default risk.
- Age: Customers aged 20 -40 have high default risk.
- Sex: Females have lower default risk than males in this dataset.
- Credit limit: Higher credit limit is associated with lower default risk.

Predictive Modeling

What recall, f1 score and ks scores can the models achieve?

Modeling Overview

Define Problem:	Supervised learning / binary classification
Imbalanced Classes:	78% non-default vs. 22% default
Models Applied:	Logistic Regression / Random Forest /SVM/ XGBoost

Modeling Steps

Data Preprocessing

- Feature selection
- Feature engineering
- SMOTE oversampling
- Train-test data splitting (75%/25%)
- Training data rescaling

Fitting and Tuning

- Start with default model parameters
- Hyperparameters tuning
- Measure recall, f1score, precision, roc_auc_score and ks statistic on training data and test data

Model Evaluation

- Models testing
- Recall, f1 score, ks statistic score comparison.
- Compare within the 4 models.(along with tuned models)

- Outlier Treatment: The outlier treatment technique used is capping, which involves setting the extreme values (outliers) to specific percentiles of the data. Specifically, the 10th percentile and 90th percentile are used as lower bound and upper bound to cap the values.

- Categorical Encoding: I did Binning for AGE variable and Label Encoding. Each age group is now represented by an integer from 0 to 5. Performed one-hot encoding on the SEX, EDUCATION, and MARRIAGE columns using the pd.get_dummies() function.

- Feature Generation: BILL_AMT_SEPT, BILL_AMT_AUG, BILL_AMT_JUL, BILL_AMT_JUN, BILL_AMT_MAY, BILL_AMT_APR are highly correlated to each other. So, I am creating a new feature which average of bill amounts.

- Feature Selection: After dealing with correlation, multi-collinearity and seeing the feature importance using Random Forest and decision tree classiffiers, These are my important features 'LIMIT_BAL', 'AGE', 'PAY_SEPT', 'PAY_AUG', 'PAY_JUL', 'PAY_JUN', 'PAY_MAY', 'PAY_APR', 'PAY_AMT_SEPT', 'PAY_AMT_AUG', 'PAY_AMT_JUL', 'PAY_AMT_JUN', 'PAY_AMT_MAY', 'PAY_AMT_APR', 'SEX_Female', 'SEX_Male', 'EDUCATION_Graduation', 'EDUCATION_High_School', 'EDUCATION_University', 'MARRIAGE_Divorce', 'MARRIAGE_Married', 'MARRIAGE_Single', 'AVG_BILL_AMT

Before SMOTE : 0 :23364, 1:6636

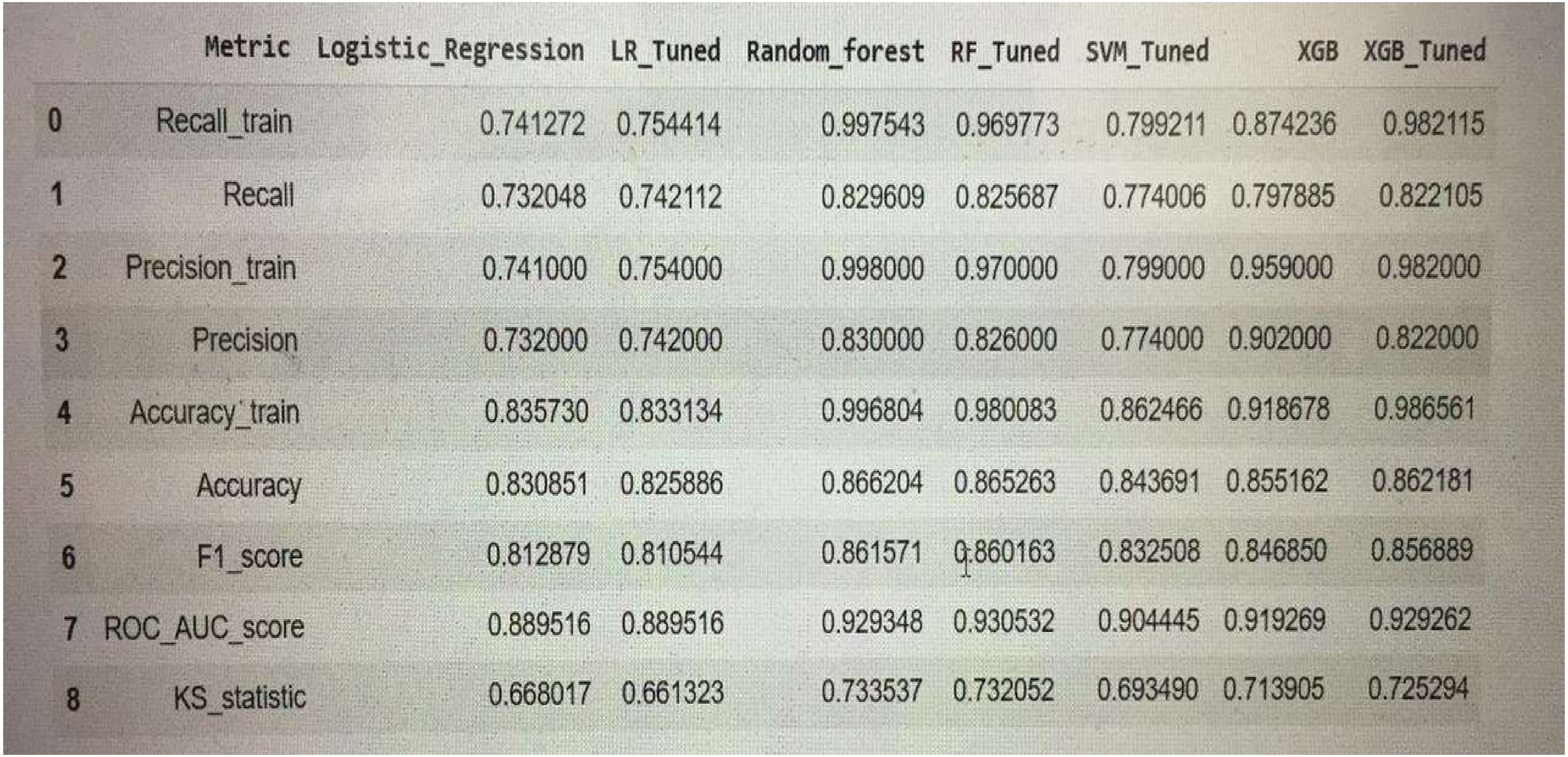
After SMOTE: 0: 23364, 1:23364

Now shape of the dataset is (46728, 23)

Hyperparameters Tuning

- Randomized Search on Random Forest ,XGB and SVM
- Grid Search on Logistic Regression on limited parameters Combinations.

Model Comparisons



	Metric	Logistic_Regression	LR_Tuned	Random_forest	RF_Tuned	SVM_Tuned	XGB	XGB_Tuned
0	Recall_train	0.741272	0.754414	0.997543	0.969773	0.799211	0.874236	0.982115
1	Recall	0.732048	0.742112	0.829609	0.825687	0.774006	0.797885	0.822105
2	Precision_train	0.741000	0.754000	0.998000	0.970000	0.799000	0.959000	0.982000
3	Precision	0.732000	0.742000	0.830000	0.826000	0.774000	0.902000	0.822000
4	Accuracy_train	0.835730	0.833134	0.996804	0.980083	0.862466	0.918678	0.986561
5	Accuracy	0.830851	0.825886	0.866204	0.865263	0.843691	0.855162	0.862181
6	F1_score	0.812879	0.810544	0.861571	0.860163	0.832508	0.846850	0.856889
7	ROC_AUC_score	0.889516	0.889516	0.929348	0.930532	0.904445	0.919269	0.929262
8	KS_statistic	0.668017	0.661323	0.733537	0.732052	0.693490	0.713905	0.725294

- I finalized Tuned Random forest Model to be my best model based on the evaluation metrics recall, F1 score, KS statistic.

- Even though XGB tuned model is giving high recall, f1 score and Ks, by observing the metrics on train data and test data, the model is likely overfitting. There is too much difference in train and test, model is almost learning everything from train data. For this reason, i disregard this model.

- Tuned Random forest is performing well on the chosen metrics than remaining models Logistic Regression and SVM.

Conclusion-:

- The main goal of the project was to create a machine learning model that predicts credit card payment defaults for the next month. After performing data visualization, I identified several important features related to defaulters. During feature engineering, I added a new feature representing the average bill amount over the past six months and removed individual variables. Additionally, since the data was heavily imbalanced, I applied the SMOTE technique to balance the dataset.

- Four models were developed and evaluated: Logistic Regression, Random Forest, SVM, and XGBoost Classifier. The evaluation metrics used were recall, f1 score, and KS statistic, focusing on the objective of predicting defaulters rather than simply classifying defaulters and non-defaulters.

- The optimized (tuned) Random Forest Classifier model showed promising results with a recall of 83%, f1 score of 86%, and KS statistic of 74%. These results were consistent on both the train and test datasets. However, the tuned XGBoost model exhibited overfitting, while SVM and Logistic Regression yielded lower scores compared to Random Forest. XGBoost performed well in all evaluation metrics except for recall, where it achieved a score of approximately 79%.

- The important features that played a crucial role in prediction were identified as 'LIMIT_BAL', 'AGE', 'PAY_SEPT', 'PAY_AUG', 'PAY_JUL', 'PAY_JUN', 'PAY_MAY', 'PAY_AMT_SEPT', 'PAY_AMT_AUG', 'PAY_AMT_JUN', 'PAY_AMT_MAY', 'PAY_AMT_APR', 'SEX_Female', 'SEX_Male', 'EDUCATION_Graduation', 'EDUCATION_High_School', 'EDUCATION_University', 'MARRIAGE_Married', 'MARRIAGE_Single', and 'AVG_BILL_AMT'.

- It is worth noting that if we had access to additional information such as customer income or annual spending, it could have greatly improved our model's ability to estimate whether a customer defaults or not.

- Based on the evaluation and considering the business objective, the final model chosen is the tuned Random Forest model due to its high recall. However, if the business places a higher emphasis on precision and minimizing the misclassification of non-defaulters as defaulters, I would recommend the XGBoost model, which demonstrated high precision and f1 score. Given the dataset and features at hand, our model performs well on all data points. With such high recall, we can confidently deploy this model for further predictive tasks using future data.

Thank you