

Capstone Project-3

Supervised ML – Classification

Credit Card Default Prediction

Presented by:

Vishal Raul

Content

1. Introduction
2. Defining Problem Statement
3. Data Summary
4. Approach Overview
5. Exploratory Data Analysis(EDA)
6. Modelling Overview
7. Feature Importance
8. ML Model-Classification
9. Conclusion



Introduction

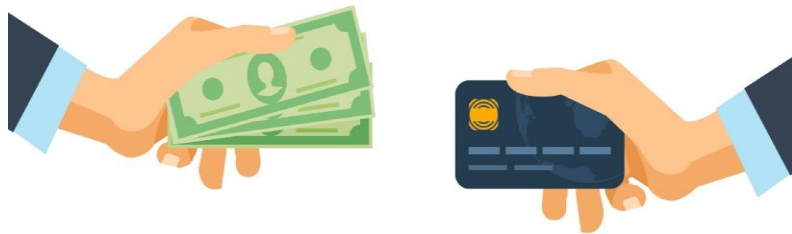


Credit risk plays a major role in the banking industry business. Bank's main activities involve granting loan, credit card, investment, mortgage, and others. Credit card has been one of the most booming financial services by banks over the past years. However, with the growing number of credit card users, banks have been facing an escalating credit card default rate.

In today's world credit cards have become a lifeline to a lot of people so banks provide us with credit cards. Now we know the most common issue there is in providing these kind of deals are people not being able to pay the bills. These people are what we call "defaulters".

Defining Problem Statement

- Our task is to build a model that predict the case of customers default payments in Taiwan.
- Conduct quantitative analysis on credit default risk by applying three interpretable machine learning models without utilizing credit score or credit history.



Data Summary:



Data Set Name :- default of credit card clients.csv

Statistics –

- Rows - 30000
- Features - 25

Columns:-

'ID', '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_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6', 'default payment next month'.

Data fields -



X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X2: Gender

X3: Education X4: Marital status X5: Age (year).

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005

X6 : the repayment status in September, 2005;

X7 : the repayment status in August, 2005; . . .;

X11 : the repayment status in April,

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .;

X17 : amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar).

X18 : amount paid in September, 2005;

X19 : amount paid in August, 2005; . . .;

X23 : amount paid in April, 2005.

Loading the Dataset

```
# loading the dataset
data = pd.read_csv('/content/drive/MyDrive/Capstone Project-03/default of credit card clients.csv')
data.head()
```

	ID	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_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
0	1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
1	2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
2	3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
3	4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
4	5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0

Attribute Information:- Dtypes and Null values

#Information of the dataset

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 30000 entries, 0 to 29999
```

```
Data columns (total 25 columns):
```

#	Column	Non-Null Count	Dtype
0	ID	30000 non-null	int64
1	LIMIT_BAL	30000 non-null	int64
2	SEX	30000 non-null	int64
3	EDUCATION	30000 non-null	int64
4	MARRIAGE	30000 non-null	int64
5	AGE	30000 non-null	int64
6	PAY_0	30000 non-null	int64
7	PAY_2	30000 non-null	int64
8	PAY_3	30000 non-null	int64
9	PAY_4	30000 non-null	int64
10	PAY_5	30000 non-null	int64
11	PAY_6	30000 non-null	int64
12	BILL_AMT1	30000 non-null	int64
13	BILL_AMT2	30000 non-null	int64
14	BILL_AMT3	30000 non-null	int64
15	BILL_AMT4	30000 non-null	int64
16	BILL_AMT5	30000 non-null	int64
17	BILL_AMT6	30000 non-null	int64
18	PAY_AMT1	30000 non-null	int64
19	PAY_AMT2	30000 non-null	int64
20	PAY_AMT3	30000 non-null	int64
21	PAY_AMT4	30000 non-null	int64
22	PAY_AMT5	30000 non-null	int64
23	PAY_AMT6	30000 non-null	int64
24	default payment next month	30000 non-null	int64

```
dtypes: int64(25)
```

```
memory usage: 5.7 MB
```

#Attribute function null values

```
data.isna().sum()
```

ID	0
LIMIT_BAL	0
SEX	0
EDUCATION	0
MARRIAGE	0
AGE	0
PAY_0	0
PAY_2	0
PAY_3	0
PAY_4	0
PAY_5	0
PAY_6	0
BILL_AMT1	0
BILL_AMT2	0
BILL_AMT3	0
BILL_AMT4	0
BILL_AMT5	0
BILL_AMT6	0
PAY_AMT1	0
PAY_AMT2	0
PAY_AMT3	0
PAY_AMT4	0
PAY_AMT5	0
PAY_AMT6	0
default payment next month	0

dtype: int64

Approach Overview

Data Cleaning

Understand and Clean

- Find information on undocumented columns values
- Clean data to get it ready for analysis

Data Exploration

Graphical and Statistical

- Exam data with visualization
- Verify findings with statistical tests

Predictive Modeling

Machine Learning

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost

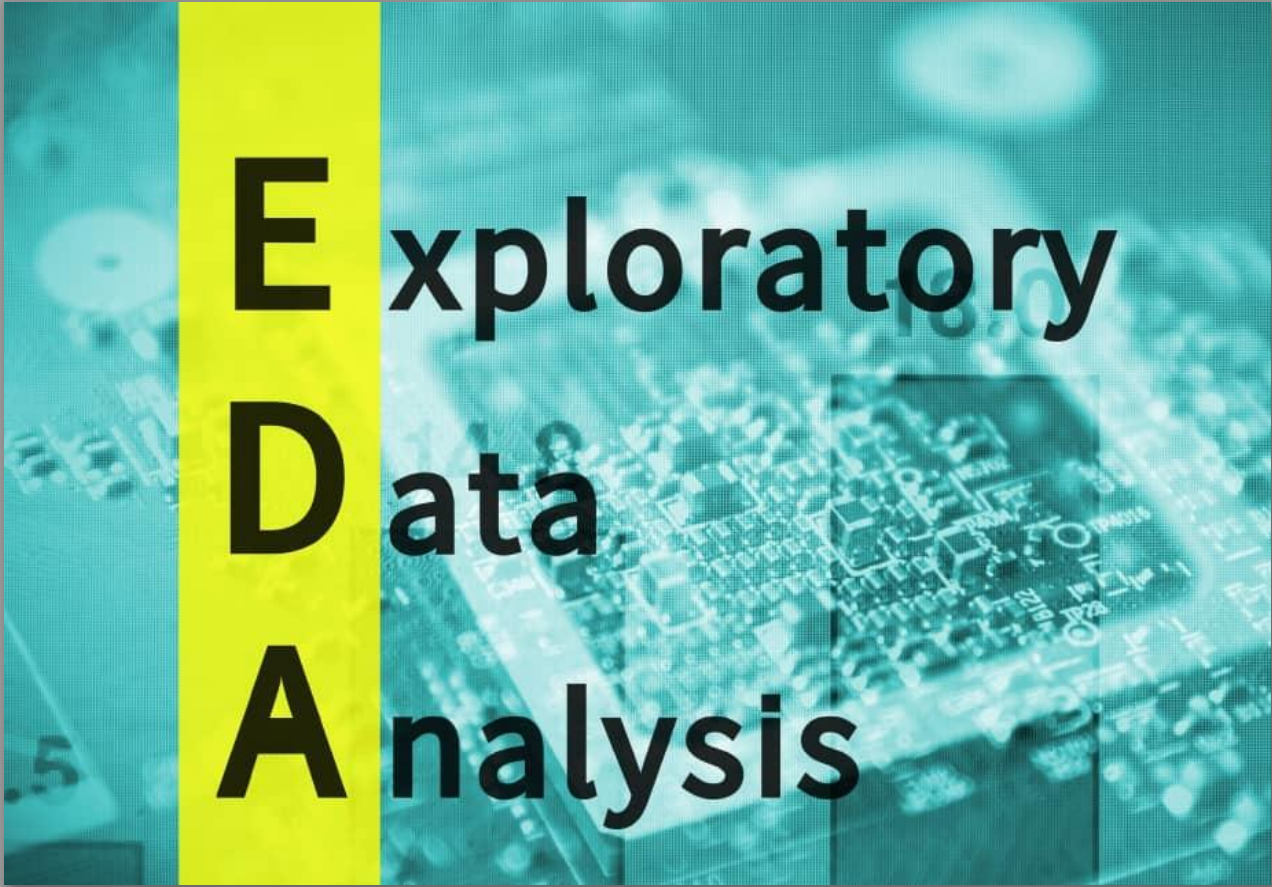
Descriptive Statistics



#Descriptive statistics
data.describe()

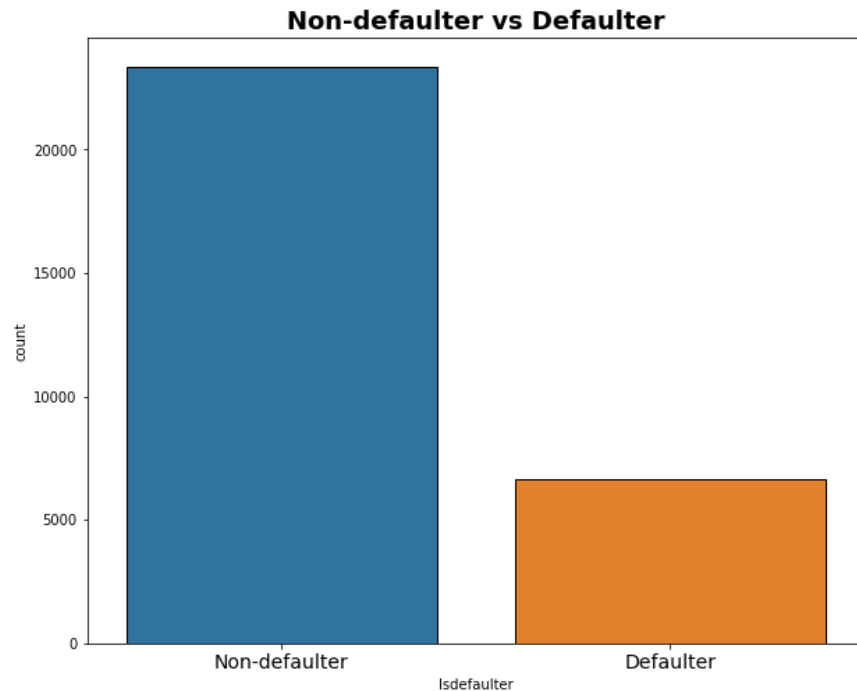
	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2
count	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000
mean	15000.500000	167484.322667	1.603733	1.853133	1.551867	35.485500	-0.016700	-0.133767	-0.166200	-0.220667	-0.266200	-0.291100	51223.330900	49179.075167
std	8660.398374	129747.661567	0.489129	0.790349	0.521970	9.217904	1.123802	1.197186	1.196868	1.169139	1.133187	1.149988	73635.860576	71173.768783
min	1.000000	10000.000000	1.000000	0.000000	0.000000	21.000000	-2.000000	-2.000000	-2.000000	-2.000000	-2.000000	-2.000000	-165580.000000	-69777.000000
25%	7500.750000	50000.000000	1.000000	1.000000	1.000000	28.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	3558.750000	2984.750000
50%	15000.500000	140000.000000	2.000000	2.000000	2.000000	34.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	22381.500000	21200.000000
75%	22500.250000	240000.000000	2.000000	2.000000	2.000000	41.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	67091.000000	64006.250000
max	30000.000000	1000000.000000	2.000000	6.000000	3.000000	79.000000	8.000000	8.000000	8.000000	8.000000	8.000000	8.000000	964511.000000	983931.000000

BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
3.000000e+04	30000.000000	30000.000000	30000.000000	30000.000000	3.000000e+04	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000
4.701315e+04	43262.948967	40311.400967	38871.760400	5663.580500	5.921163e+03	5225.68150	4826.076867	4799.387633	5215.502567	0.221200
6.934939e+04	64332.856134	60797.155770	59554.107537	16563.280354	2.304087e+04	17606.96147	15666.159744	15278.305679	17777.465775	0.415062
-1.572640e+05	-170000.000000	-81334.000000	-339603.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000
2.666250e+03	2326.750000	1763.000000	1256.000000	1000.000000	8.330000e+02	390.00000	296.000000	252.500000	117.750000	0.000000
2.008850e+04	19052.000000	18104.500000	17071.000000	2100.000000	2.009000e+03	1800.00000	1500.000000	1500.000000	1500.000000	0.000000
6.016475e+04	54506.000000	50190.500000	49198.250000	5006.000000	5.000000e+03	4505.00000	4013.250000	4031.500000	4000.000000	0.000000
1.664089e+06	891586.000000	927171.000000	961664.000000	873552.000000	1.684259e+06	896040.00000	621000.000000	426529.000000	528666.000000	1.000000



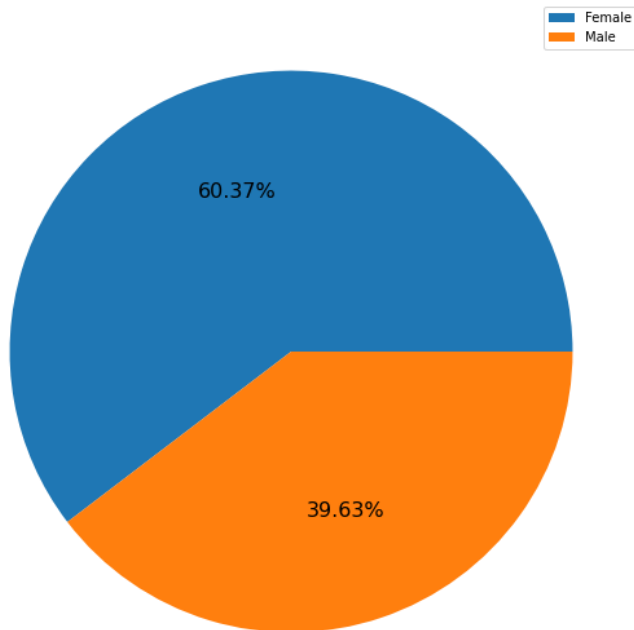
Exploratory **D**ata **A**nalysis

Isdefaulter

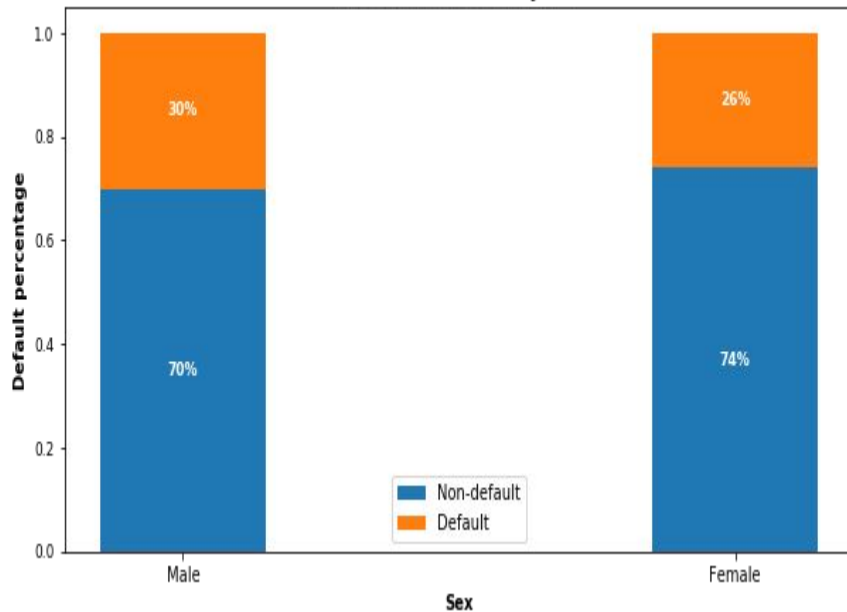


- From above bar graph, there are only 22% defaulter.

Percentage of Male and female

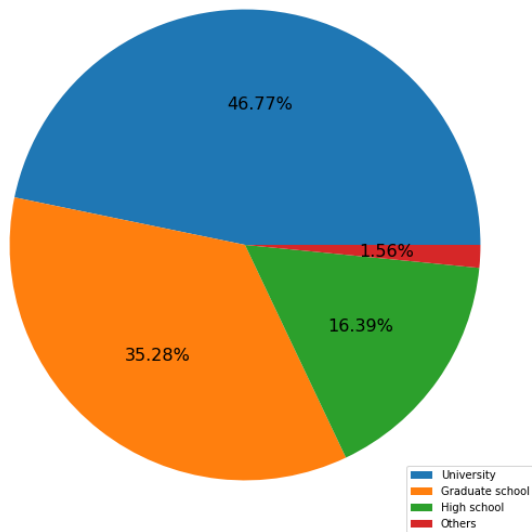


Percent of default by sex

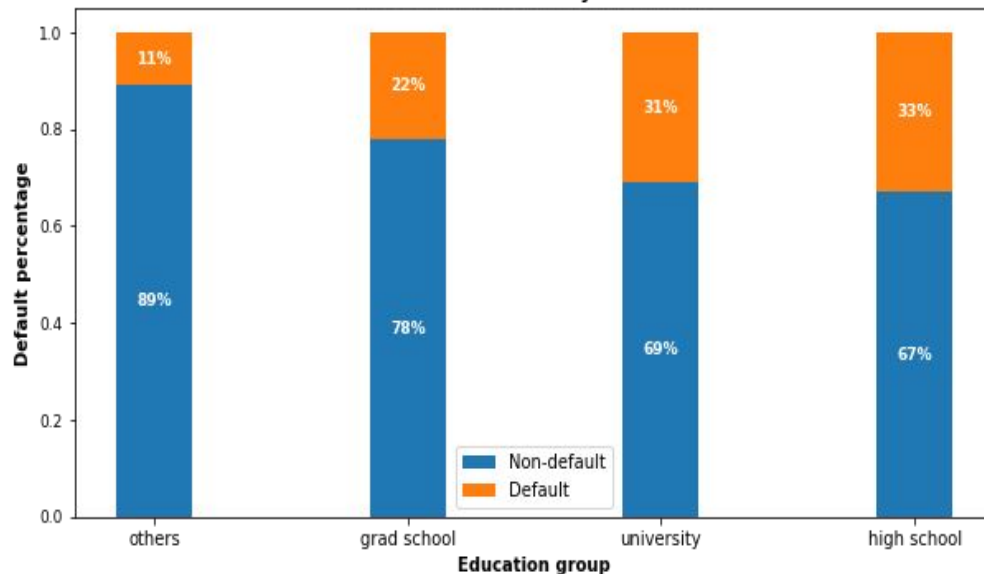


- **30% of males and 26% of females have payment default.**

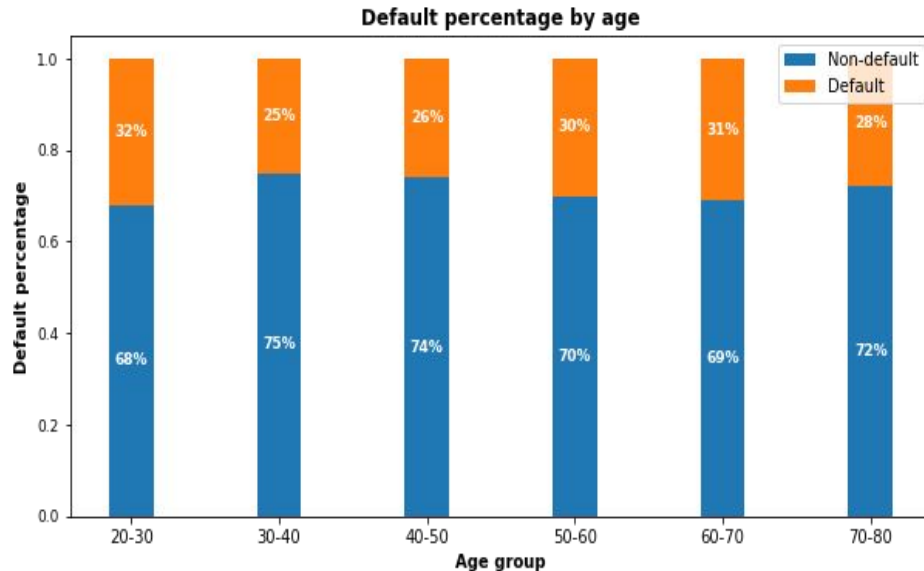
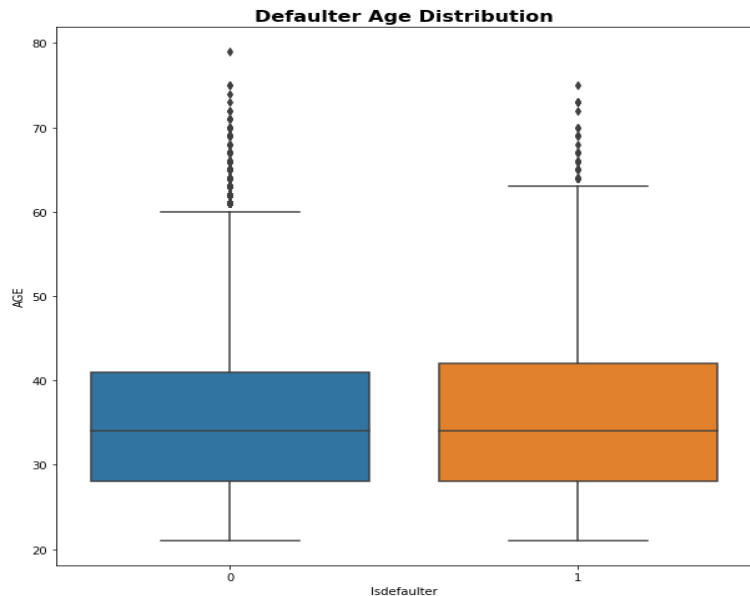
Percentage of default by education



Percent of default by education



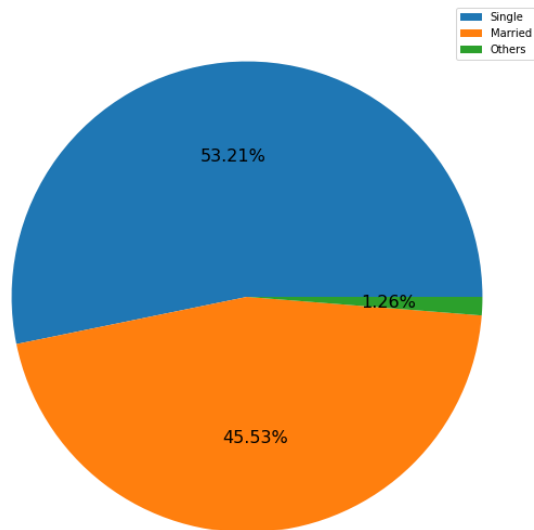
- Maximum number of defaulter are educated people (University and Graduate school)



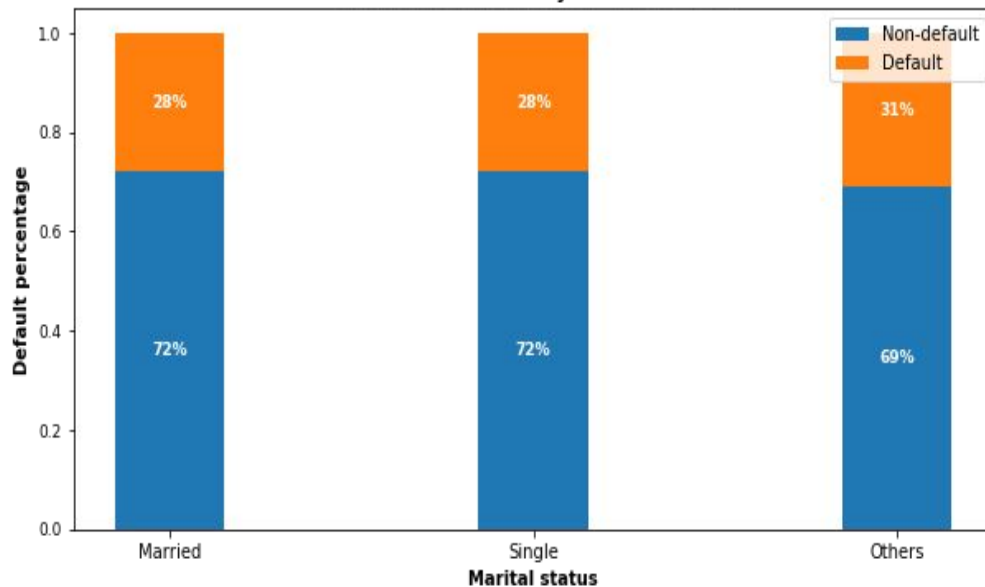
Age :-30-50: Lowest risk **Age:-< 30 or >50:** Risk increases

- **Customers aged 30-50 have the lowest default risk.**

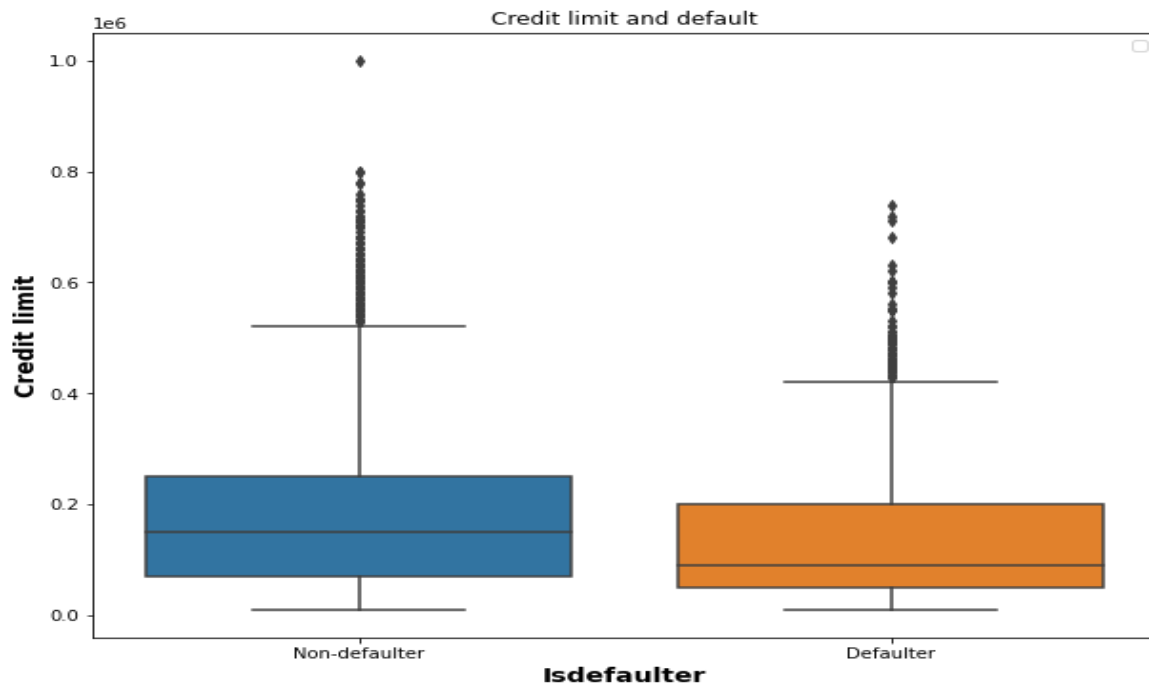
Education of Defaulter



Percent of default by marital status

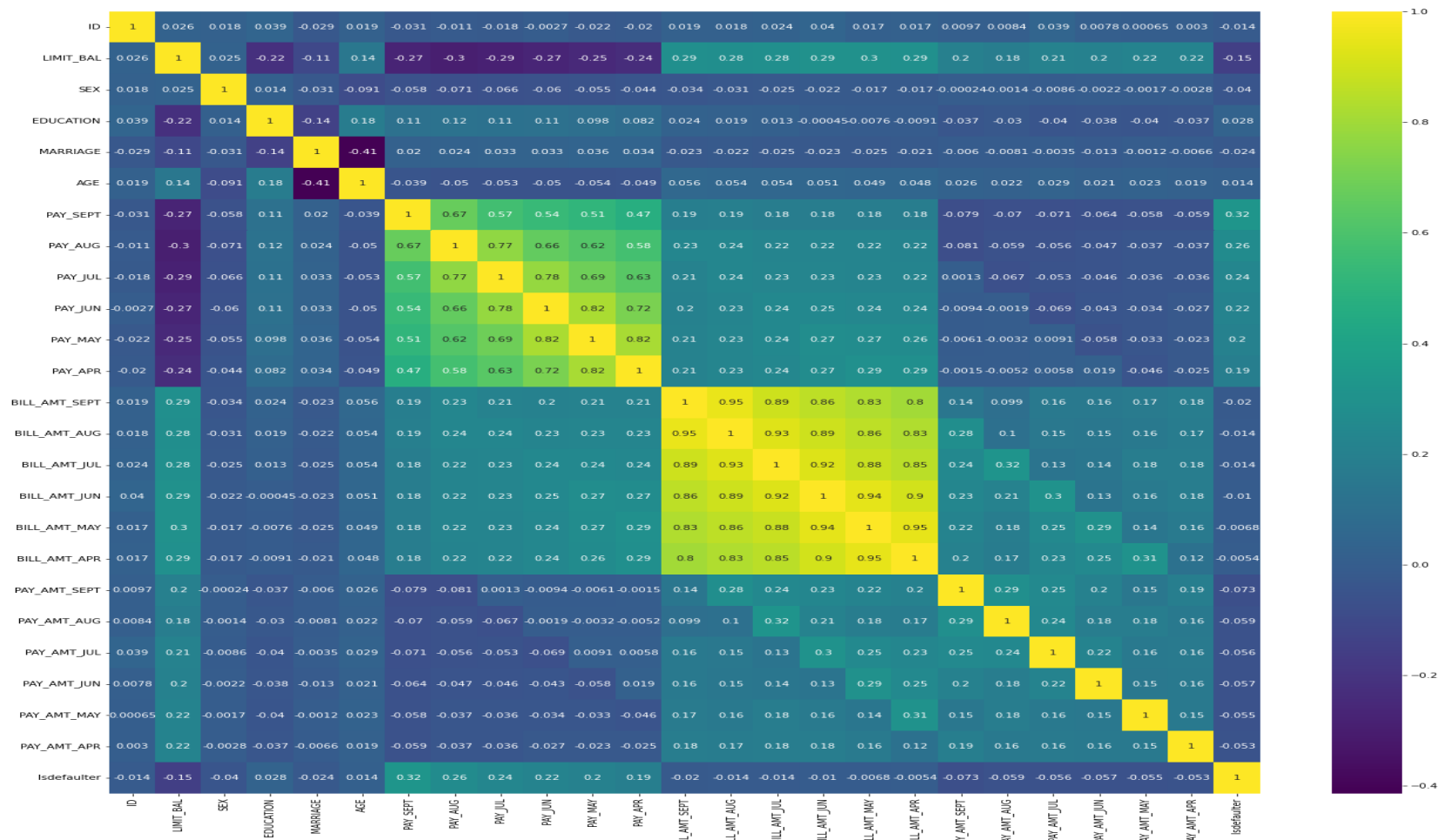


- No significant correlations of default risk and marital status



- Higher credit limit is associated with lower default risk.

Extracting the correlation of the dataset using Heatmap



Modeling Overview

Define Problem:

Supervised learning / binary classification

Imbalanced Classes:

78% non-default vs. 22% default

Tools Used:

Scikit learn library and imblearn

Models Applied:

Logistic Regression / Random Forest / Decision Tree / XGBoost

Modeling Steps

Data Preprocessing

- Feature selection
- Feature engineering
- Train-test data splitting (70%-30%)
- Training data rescaling
- SMOTE oversampling

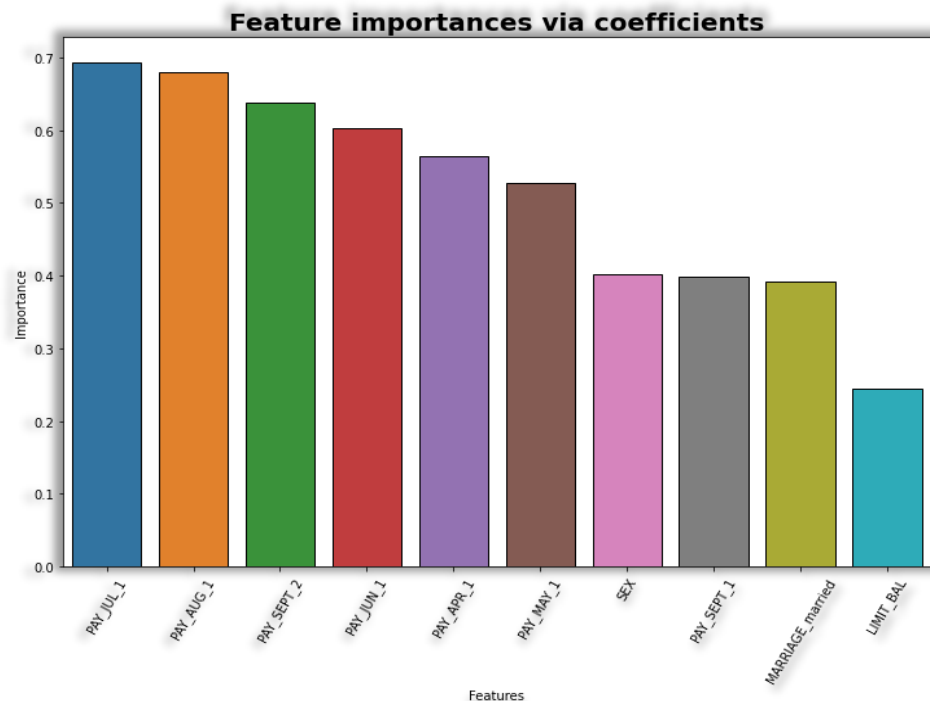
Fitting and Tuning

- Start with default model parameters
- Hyperparameters tuning
- Measure ROC-AUC on training data

Model Evaluation

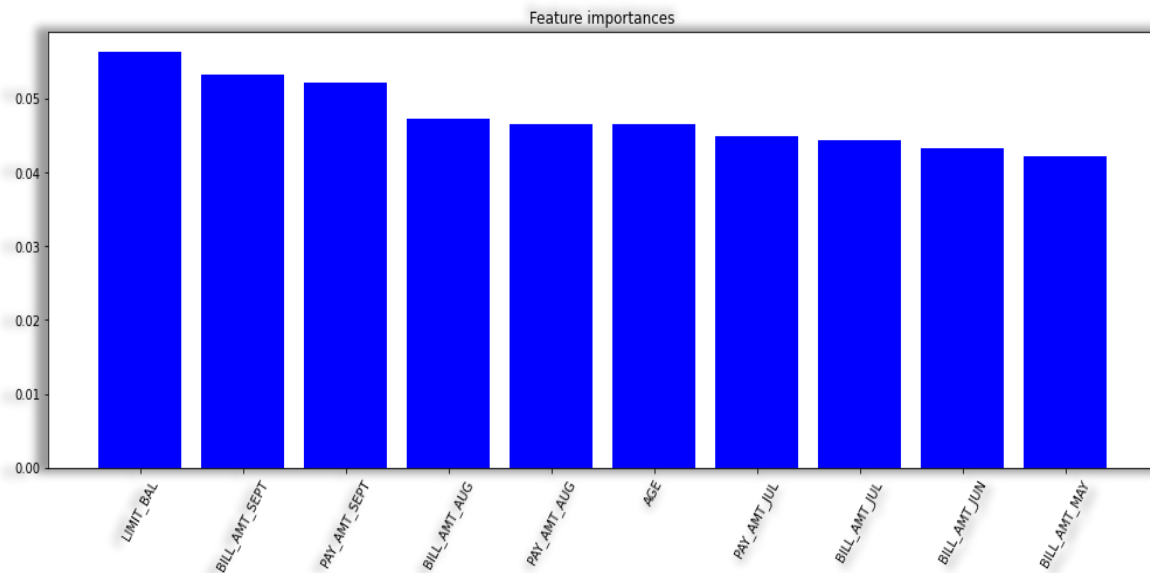
- Models testing
- Precision - Recall score
- Compare with sklearn dummy classifier
- Compare within the 3 models

Logistic Regression



The accuracy on test data is 0.7544047364291319
The precision on test data is 0.6939649022685119
The recall on test data is 0.789354105809802
The f1 on test data is 0.7385923620074407
The roc_score on test data is 0.7581745710415945

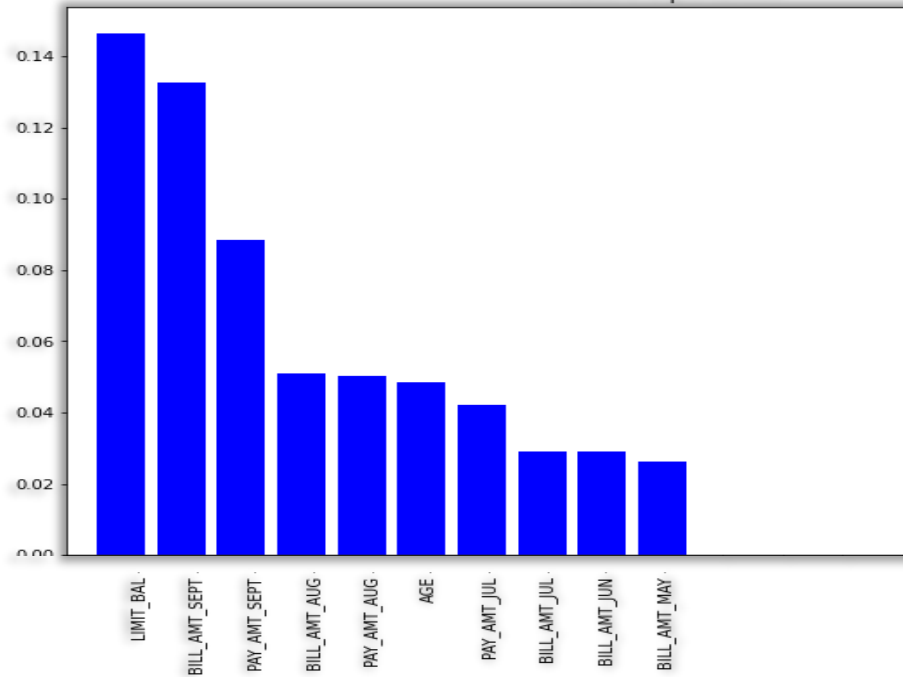
Random Forest



The accuracy on test data is 0.8332792944685818
The precision on test data is 0.8025940337224384
The recall on test data is 0.8550504352632307
The f1 on test data is 0.8279922392453335
The roc_score on test data is 0.834538902871107

XGB CLF

Feature importances



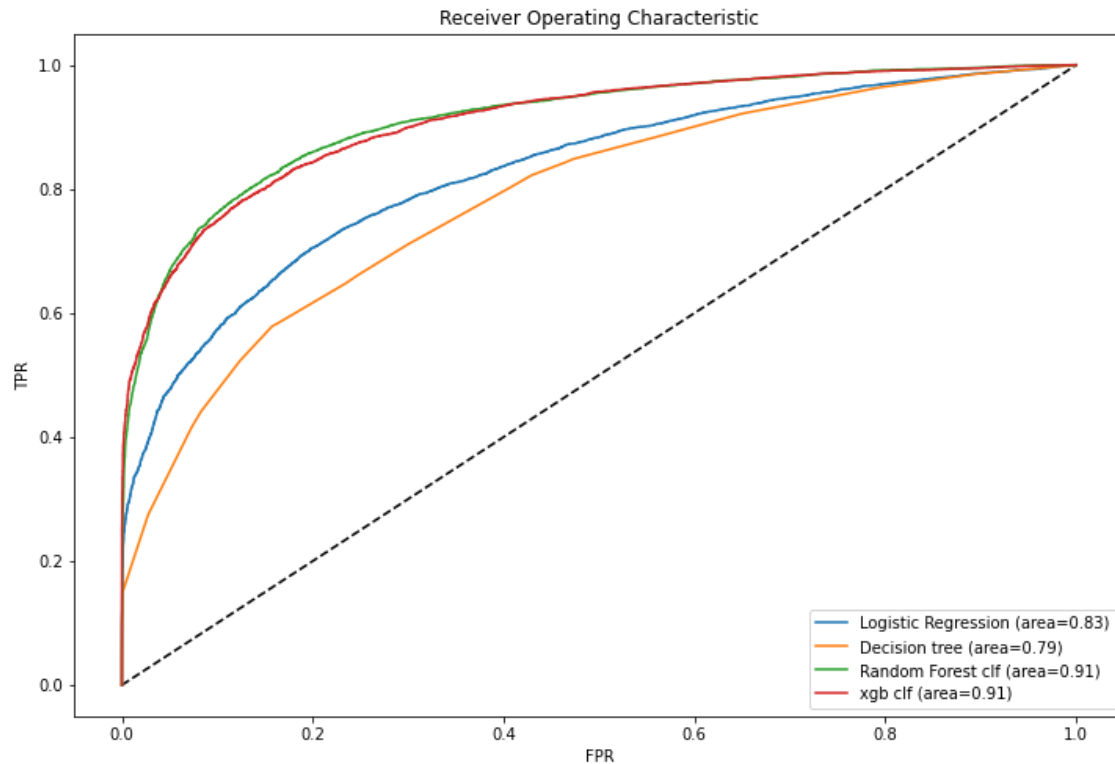
The accuracy on test data is 0.8265352441475909
The precision on test data is 0.7888456549935149
The recall on test data is 0.8531350820591949
The f1 on test data is 0.8197317878563246
The roc_score on train data is 0.828400633166597

Model Comparison

	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.755022	0.754405	0.693965	0.789354	0.738592
1	Decision Tree Clf	0.710136	0.710136	0.581842	0.782624	0.667460
2	Random Forest CLf	0.999521	0.833279	0.802594	0.855050	0.827992
3	Xgboost Clf	0.914843	0.826535	0.788846	0.853135	0.819732

- **Observations shows us that Decision Tree classifier isn't performing well Approach , others all three model are perform better.**

Model Comparison



- Compare within 3 models.
- Random Forest (black line) has the best precision_recall score.

Conclusion

- Recent 2 payment status and credit limit are the strongest default predictors.
- Dormant customers can also have default risk.
- Random Forest has the best precision and recall balance.
- Higher recall can be achieved if low precision is acceptable.
- Model can be served as an aid to human decision.
- Suggest output probabilities rather than predictions.
- Model can be improved with more data and computational resources.

References

1. <https://www.researchgate.net/publication/344914401>
2. <https://www.kaggle.com/>
3. <https://www.analyticsvidhya.com/>
4. <https://learn.almabetter.com/>