Capstone Project-3

Credit Card Default Prediction

Individual Capstone Project

Sammed N. Majalekar

INTRODUCTION

Problem Statement- To predict whether the customer will make default on his/her credit card payment on next month.

Dataset-

We have a dataset having 30000 observations with 25 features including 1 target variable from Taiwan.

Steps Followed

1. 1.Understanding the Data
2. 2.Data Cleaning
3. 3.Data Visualization
4. 4.Feature Engineering
5. 5.Data Rescaling & Standardization
6. 6.Model Building
7. 7.Hyperparameter Tuning
8. 8.Conclusion

Understanding Data

Description of columns from dataset documentation

There are 25 variables:

- **ID:** ID of each client
- **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)

Understanding Data

- **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)

Target Variable

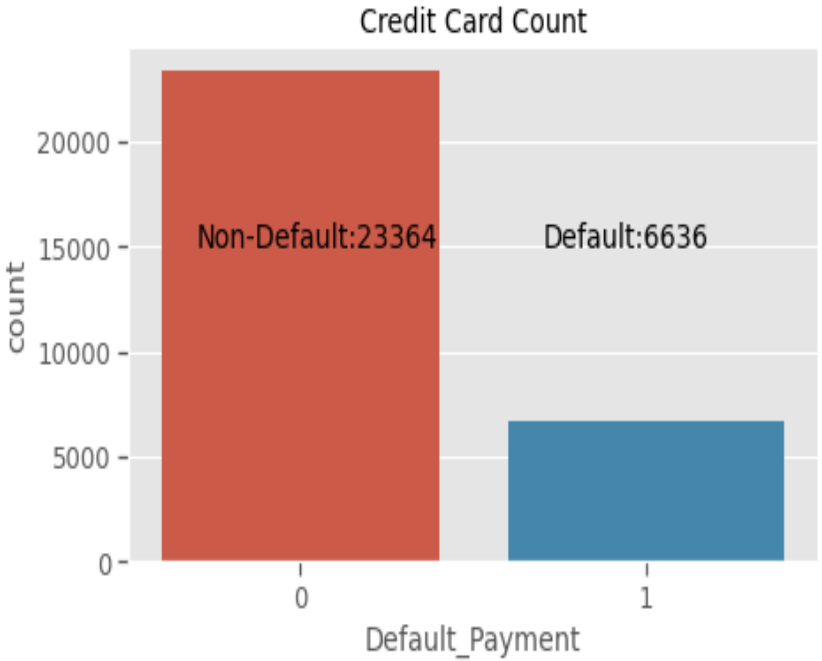
- **default.payment.next.month**: Default payment (1=yes, 0=no)

Data Cleaning

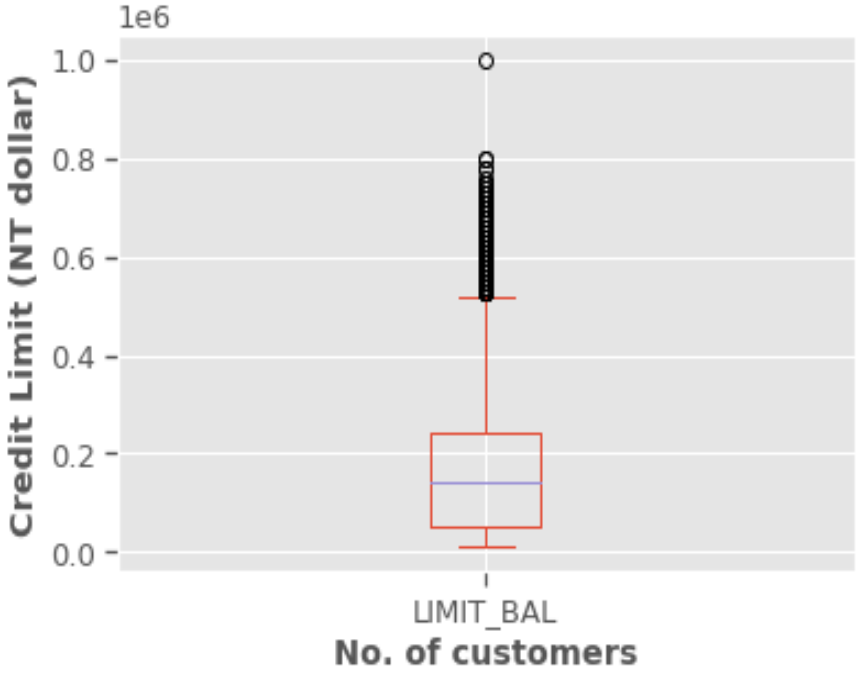
- Null values checking
- Duplicated Values checking
- Cleaning some categorical variables like Education, Gender ,Marital status, Age

Data Visualization

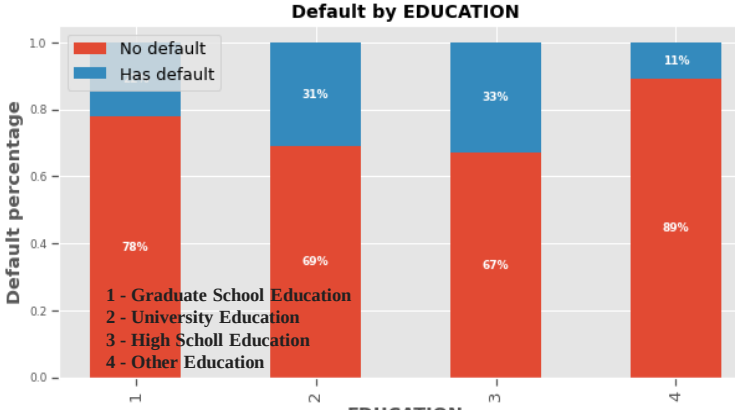
Default VS Non-Default



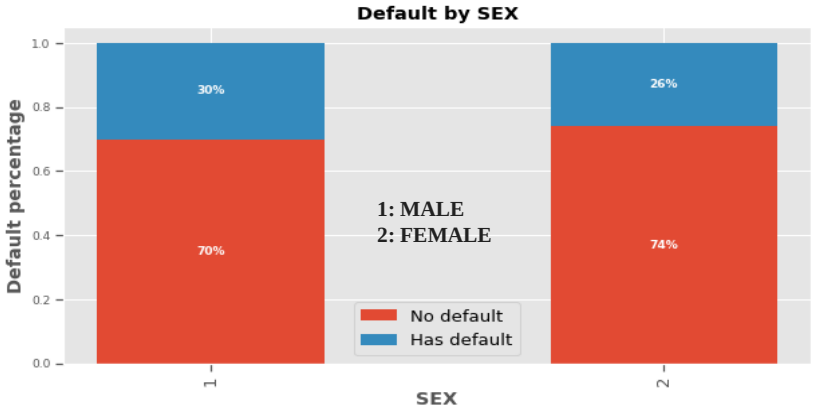
Credit Limit VS No. of customer



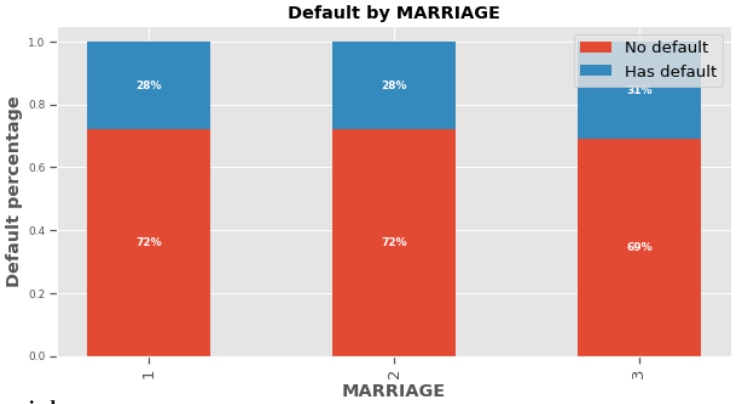
Default Percentage



Education

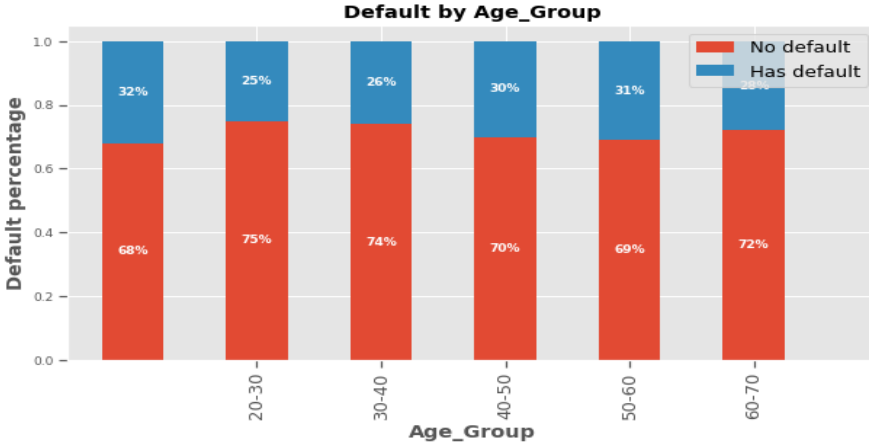


Gender



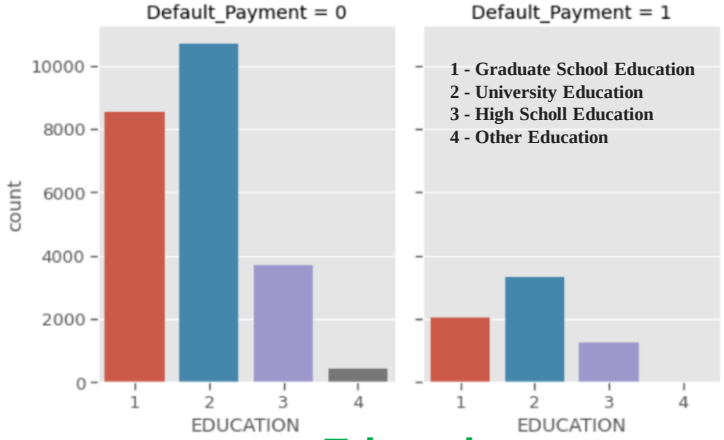
1 : Married
2 : Single
3 : Other Marital Status

Marital Status

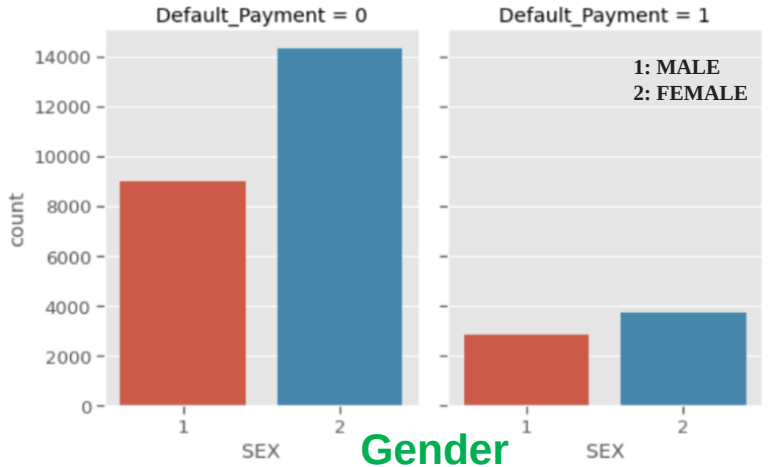


Age Group

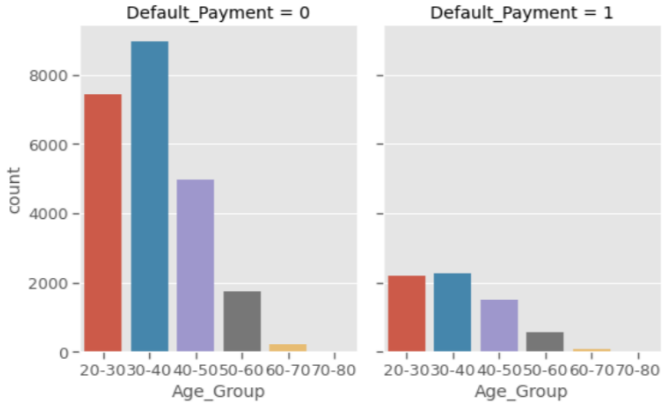
Default and Non-Default Counts



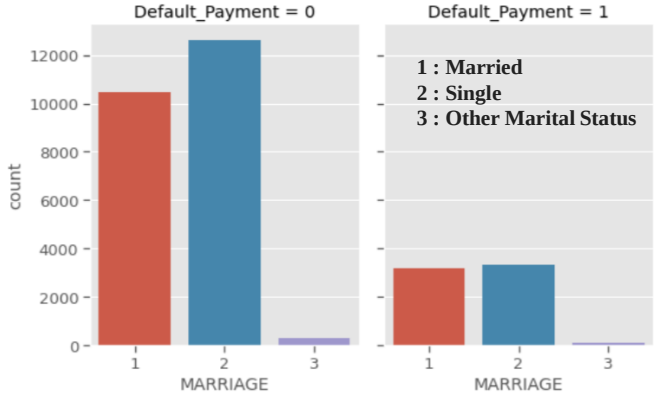
Education



Gender

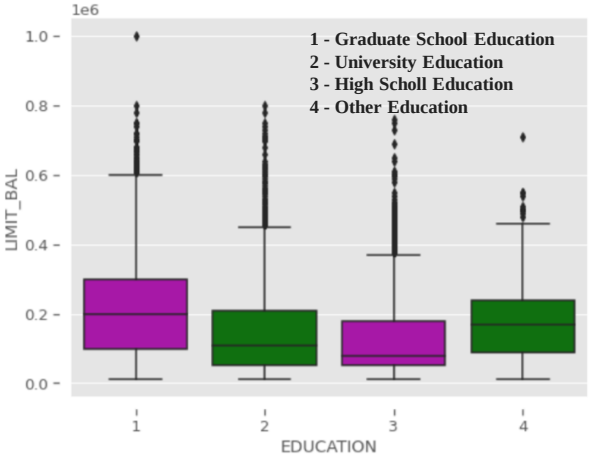


Age Group

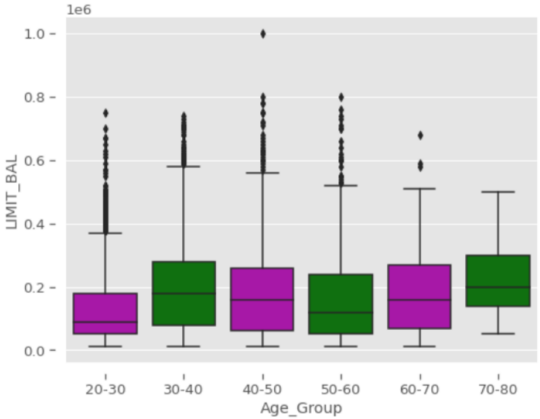


Marital Status

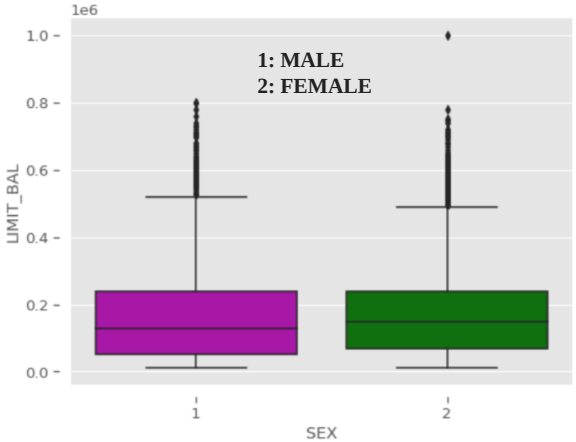
Credit Limit



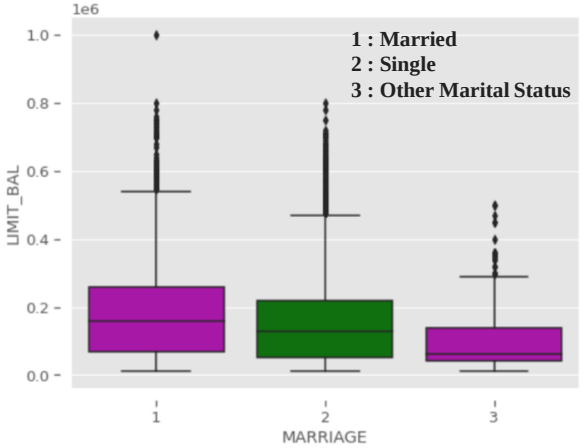
Education



Age Group

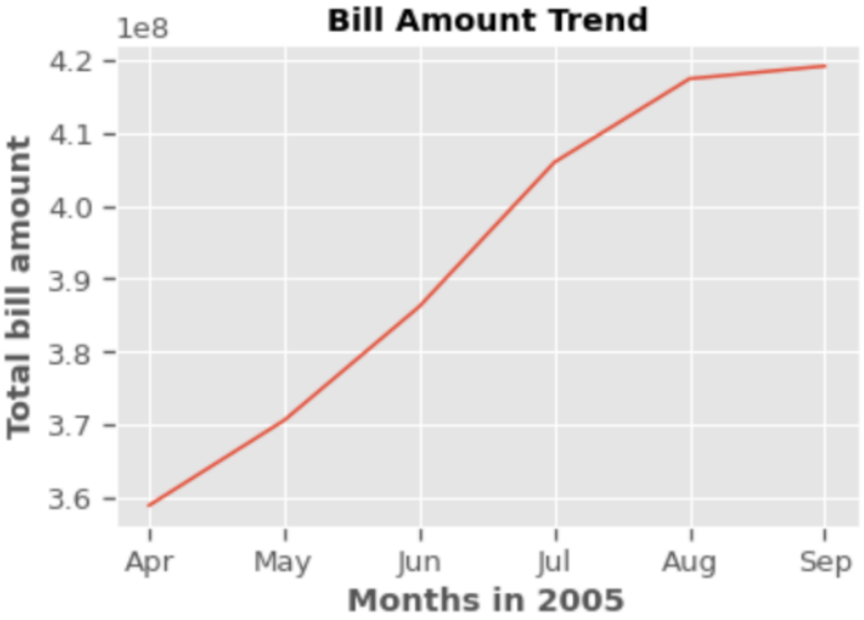


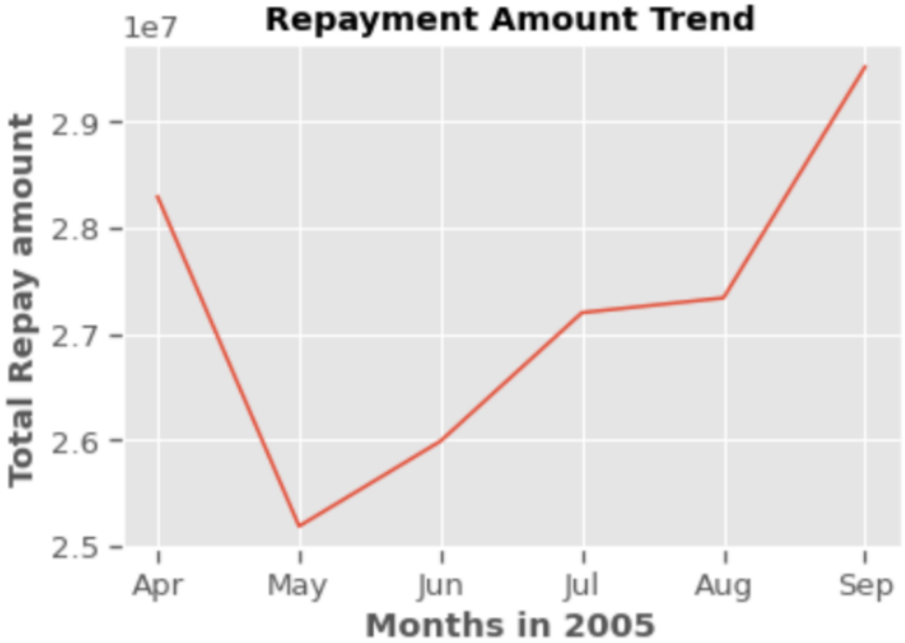
Gender



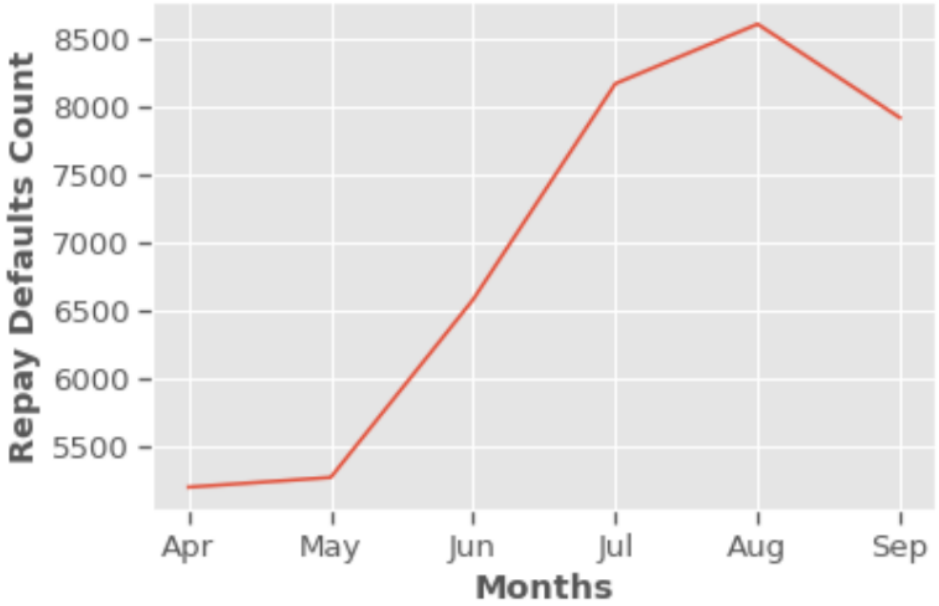
Marital Status

Bill and Repayment Amount Status In last 6 months



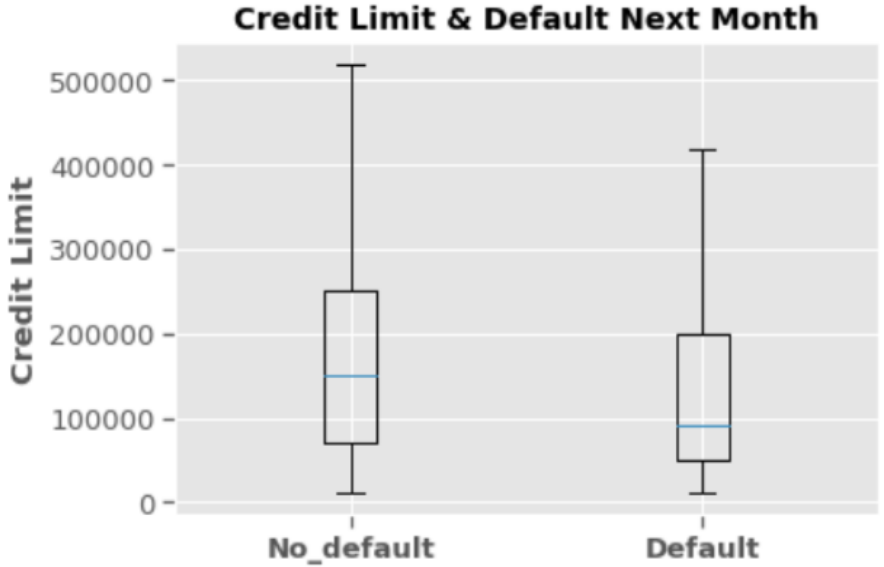


Repayment Status Trend in last 6 months



Here default payment(at least 1) has increased till August 2005 and then decreased a slight in September 2005.

Credit Limit VS Default Next Month



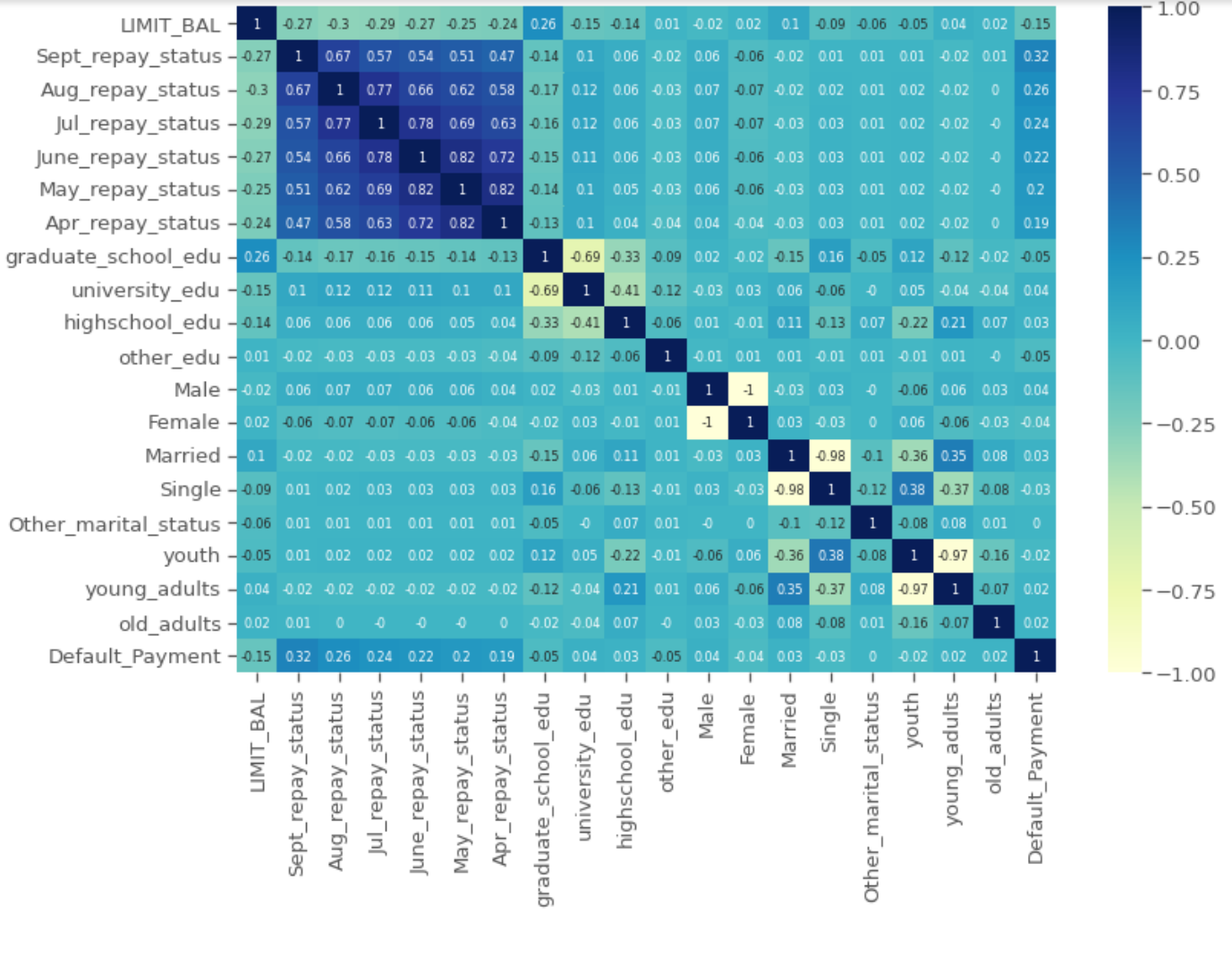
- **Here we clearly see that people having low credit limit tend to default the payment, because average credit limit is low for default payments than non-default payments.**

- **People with higher average credit limit are tending to repay the amount regularly.**

Correlation

- **Payment Defaulted or not is more correlated with repayment status in last 6 months than other features.**

- **Also Education, Marriage status and Age group have positively correlated.**



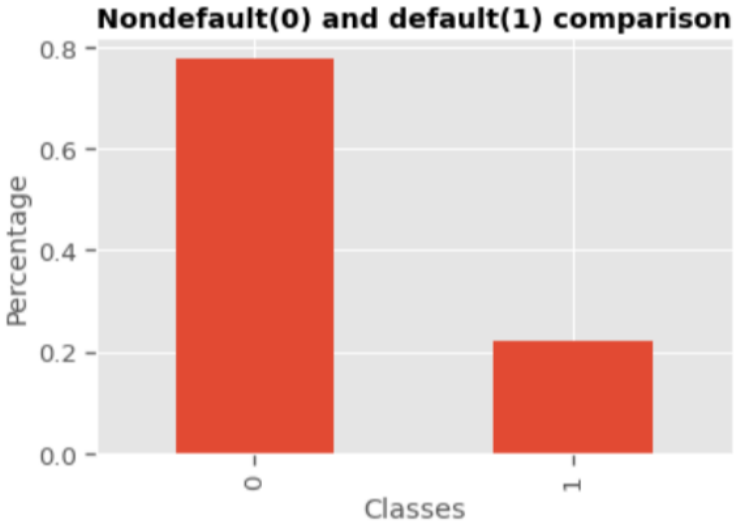
Feature Engineering

Encoding of Following Features:

1. Education
2. Gender
3. Marital Status
4. Age Group

Data Rescaling & Model Building

1. Checked for Class imbalance
2. Decided to use SMOTE (Synthetic Minority
Oversampling Technique)
3.Splitting data- Define a function to split data with and without SMOTE
4. Rescaling the Data with Standard scaler



Model Building- Logistic Regression

	Model	SMOTE	ROC_AUC Score
0	Logistic Regression	Without SMOTE	0.720545
1	Logistic Regression	With SMOTE	0.905483

Here ROC_AUC score on training data is much better using SMOTE, so we will SOMTE for tuning the model

Used Random Search Cross Validation For Hyper parameter Tuning.

```
print(lr_best.best_params_)
```

```
{'C': 1.645579535656045, 'penalty': 'l2'}
```

```
Logistic Regression Model on Test Data :
Precision:0.639
Recall:0.386
F1 score:0.482
```

Model Performance:

In this case of credit card default prediction, requires model with a high recall, but here in Logistic regression model recall is 0.386. So we will try another model called Random Forest.

Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

	Model	SMOTE	ROC_AUC Score
0	Random Forest	Without SMOTE	0.757105
1	Random Forest	With SMOTE	0.937544

Here also ROC_AUC score on training data is good with SMOTE, so we will use SMOTE.

Hyper parameter tuning

```
Fitting 3 folds for each of 81 candidates, totalling 243 fits
{'max_depth': 90,
 'min_samples_leaf': 3,
 'min_samples_split': 8,
 'n_estimators': 300}
```

Evaluation Metrics

```
Random Forest model on test data:
Precision:0.618
Recall:0.483
F1 score:0.542
```

XGBoost

ROC_AUC Score

	Model	SMOTE	ROC_AUC Score
0	XGBoost	Without SMOTE	0.777783
1	XGBoost	With SMOTE	0.920871

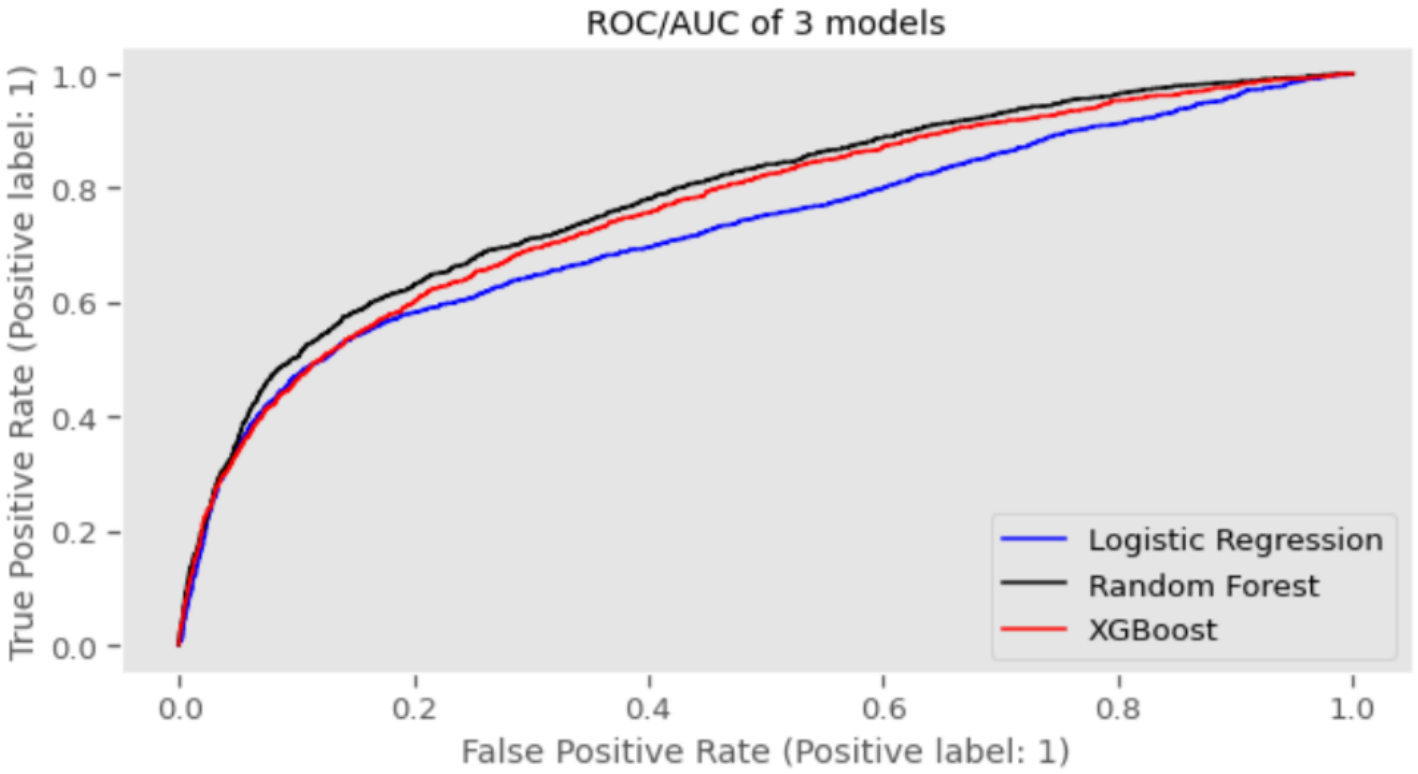
Hyper parameter Tuning

Fitting 3 folds for each of 100 candidates, totalling 300 fits
{'subsample': 0.6, 'scale_pos_weight': 3.5, 'n_estimators': 400, 'max_depth': 7, 'learning_rate': 0.1, 'gamma': 0.4, 'colsample_bytree': 0.7999999999999999}
0.9091361794042147

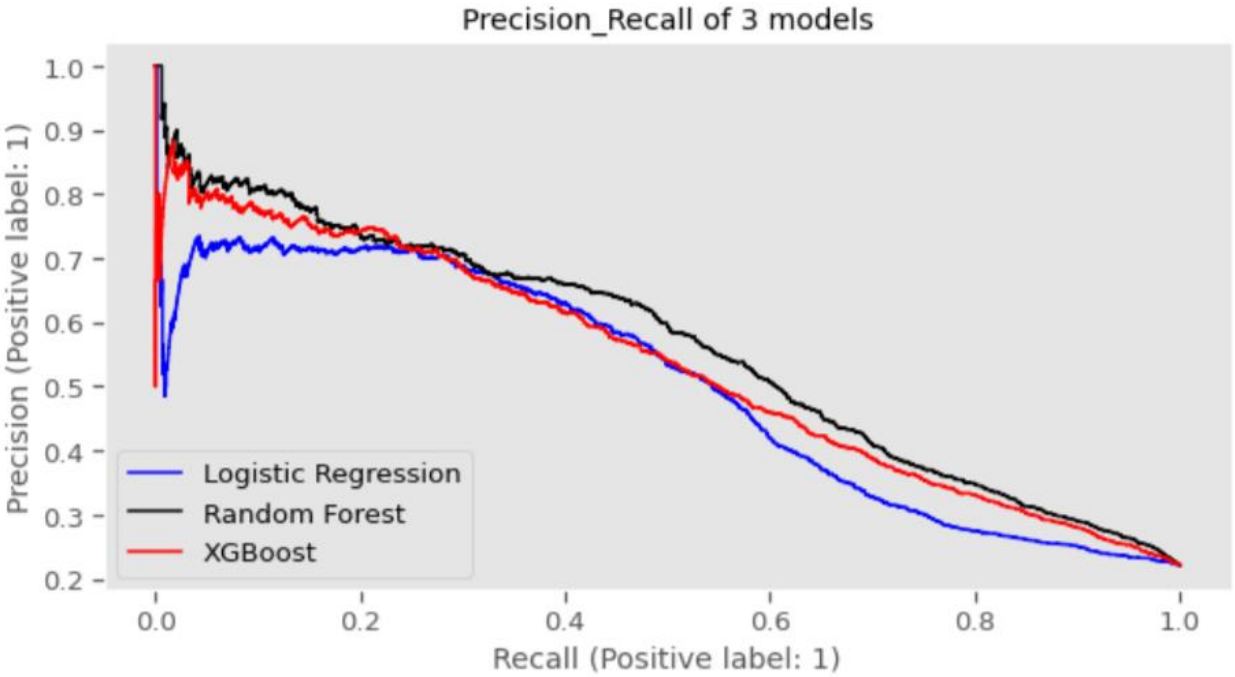
Confusion Metrics

```
XGBoost model on test data:
Precision:0.468
Recall:0.583
F1 score:0.519
```

ROC AUC Curve



Precision Recall Curve

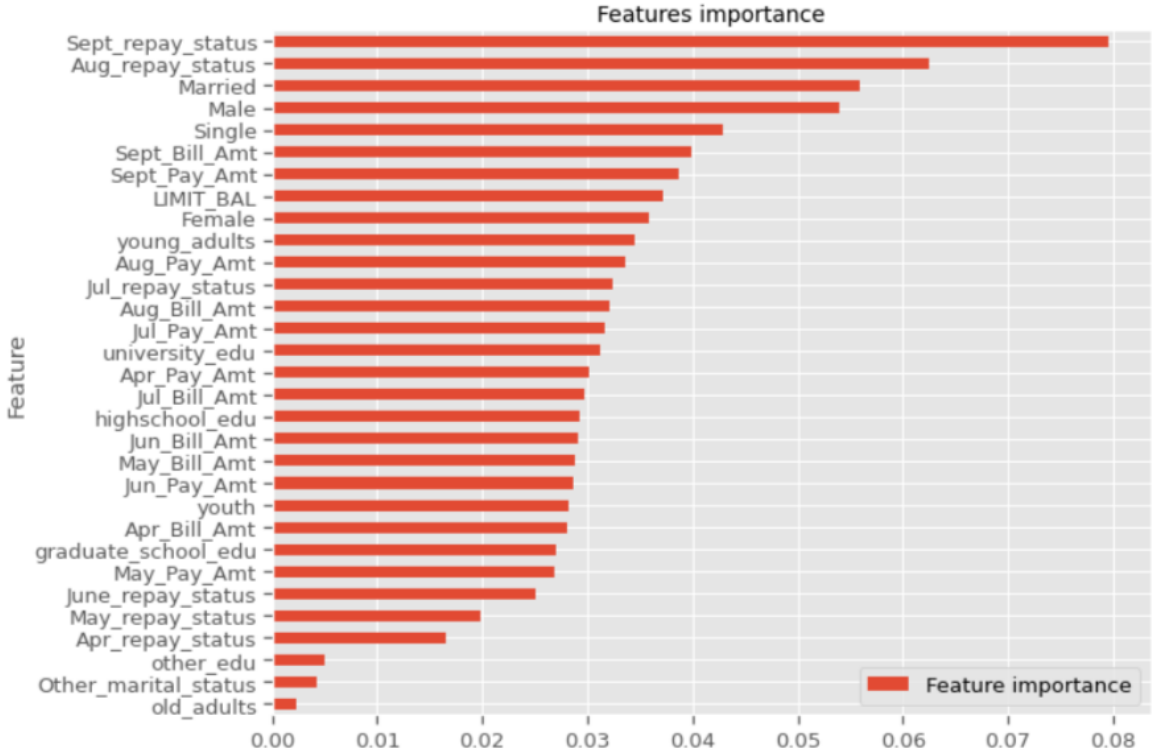


Confusion Metrics Comparison

	Precision	Recall	F1
Logistic Regression	0.639	0.386	0.482
Random Forest	0.681	0.483	0.542
XGBoost Classifier	0.468	0.583	0.513
Dummy Classifier	0.218	0.495	0.303

- Here our all of 3 models surpasses the dummy model.
- And we can see that Random Forest has a good Precision-Recall balance than other models, Random Forest will be our winner model.

Feature Importance



Here **'Sept_repay_status'** and **'Aug_repay_status'** has most important features while predicting the default, after that **'Married'** and **'Male‘** features are also having feature importance more than 0.5 score

Conclusion

Logistic Regression Model has highest precision but lowest recall, if our business cares more for precision then this model would be the winner.

XGBoost classifier has the highest recall but lowest precision.

Random Forest has a good precision and recall score that means it has a good balance of precision and recall which is higher F1, hence we will recommend the Random Forest Classifier Model.

From feature importance we get that last 2 months repayment status will helpful in deciding whether the customer will default or not.

Marital status and Gender are also fairly important features.

Challenges

- More computational power required
- Another model could perform better.
- Model can only be served as an aid in decision making instead of replacing human decision.
- Used only 30,000 records.

Thank You..!