Capstone Project - 3

CREDIT CARD DEFAULT PREDICTION

Team Members

Harisha Chennozwala

Niharika Soni

Satya Prakash

APPROACH OVERVIEW

AI

Introduction

What is Credit Card Default and the objectives of this project

Problem & Objective

- Problem statement
- Problem Objective

Data Acquisition

Dataset

- Why this Dataset?
- Dataset features

Data Cleaning

Understand and Clean

- Finding information on undocumented column values.
- Clean data to get it ready for analysis.

Exploratory Data Analysis

Graphical & Statistical:

- Exam data with visualizations.
- Verify findings with statistical test.

Machine Learning Models

- Predictive Modeling
- Hyperparameter Tuning
- Model Performance

Metrics

- ROC_AUC Curves
- Precision_Recall curve

Models

- Comparing Models with dummy classifier
- Model Recommendation

Feature Importance

- Best model feature importances plot.

Conclusion

- Limitations
- Future work
- Suggestions

Credit Card Default is a complex phenomenon involving many factors beyond the scope of the present research. The variables which we have examined here capture some key behaviors and provide the issuer a better understanding of current and potential customers, specifically which would inform their strategy in the new market.

PROBLEM DESCRIPTION

This Project is aimed at predicting the case of customers default payments in Taiwan from 2005.

Quantitative analysis on credit default risk

To evaluate which customers will default on their credit card payments.

OBJECTIVE

CLASSIFIER MODEL

NO DEFAULT — More consistency to loaning process

DEFAULT — Telltale signs of a potential defaulter

DATA DESCRIPTION

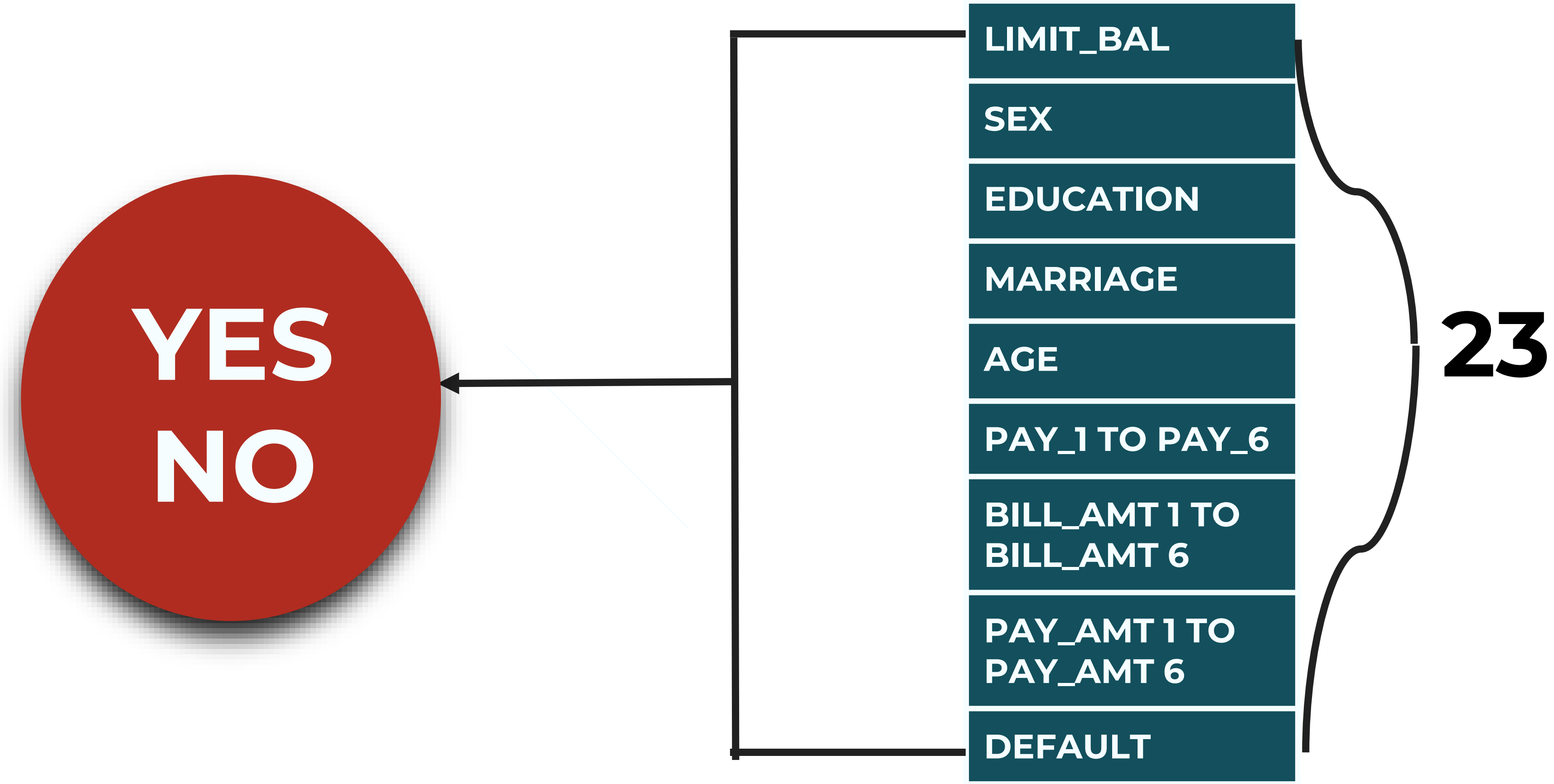
30,000 CLIENTS

23 VARIABLES

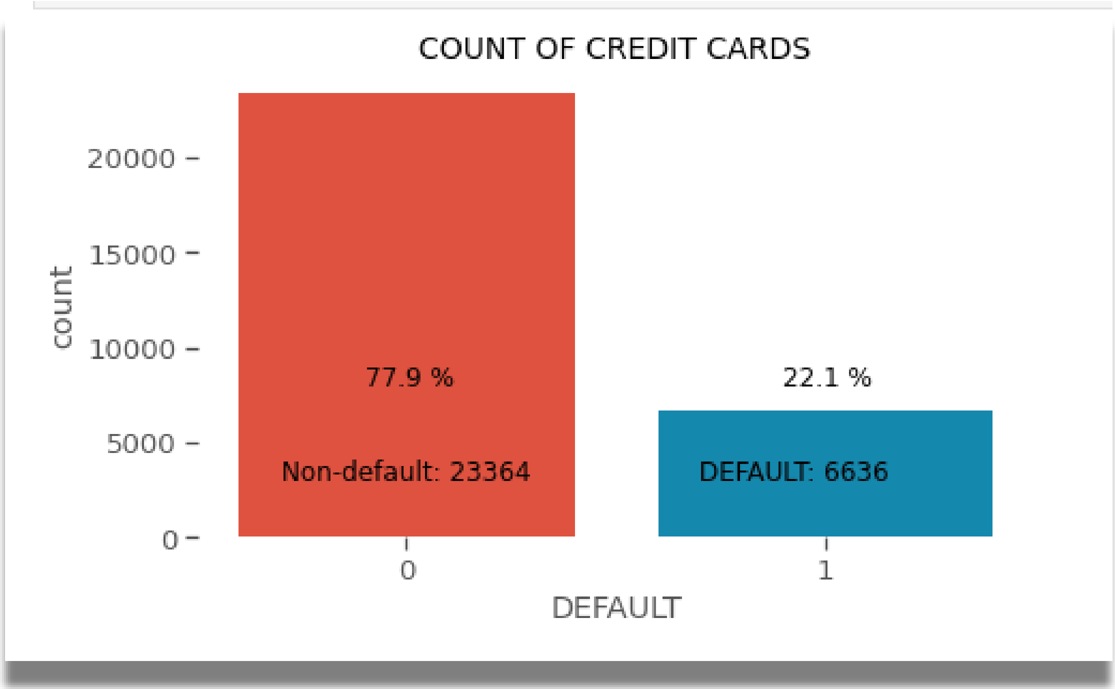
- AGE (20-79)
- GENDER
- MARRIAGE
- EDUCATION

- April to September, 2005
- Bill Statements
- Payment Amount
- Repayment Status

DATASET FEATURES



TARGET SKEW



- **Class Imbalance**
- **22% Defaulters**
- **Taken into account**

DATA CLEANING

- Converting the column names to proper names
- Renaming column PAY_0 to PAY_1 and default.payment.next.month as DEFAULT
- Converting the data type of all the columns to integer.
- There is no missing data in the entire dataset.
- Overall, the dataset is very clean, but there are several undocumented column values. As a result, most of the data wrangling effort was spent on searching information and interpreting the columns.

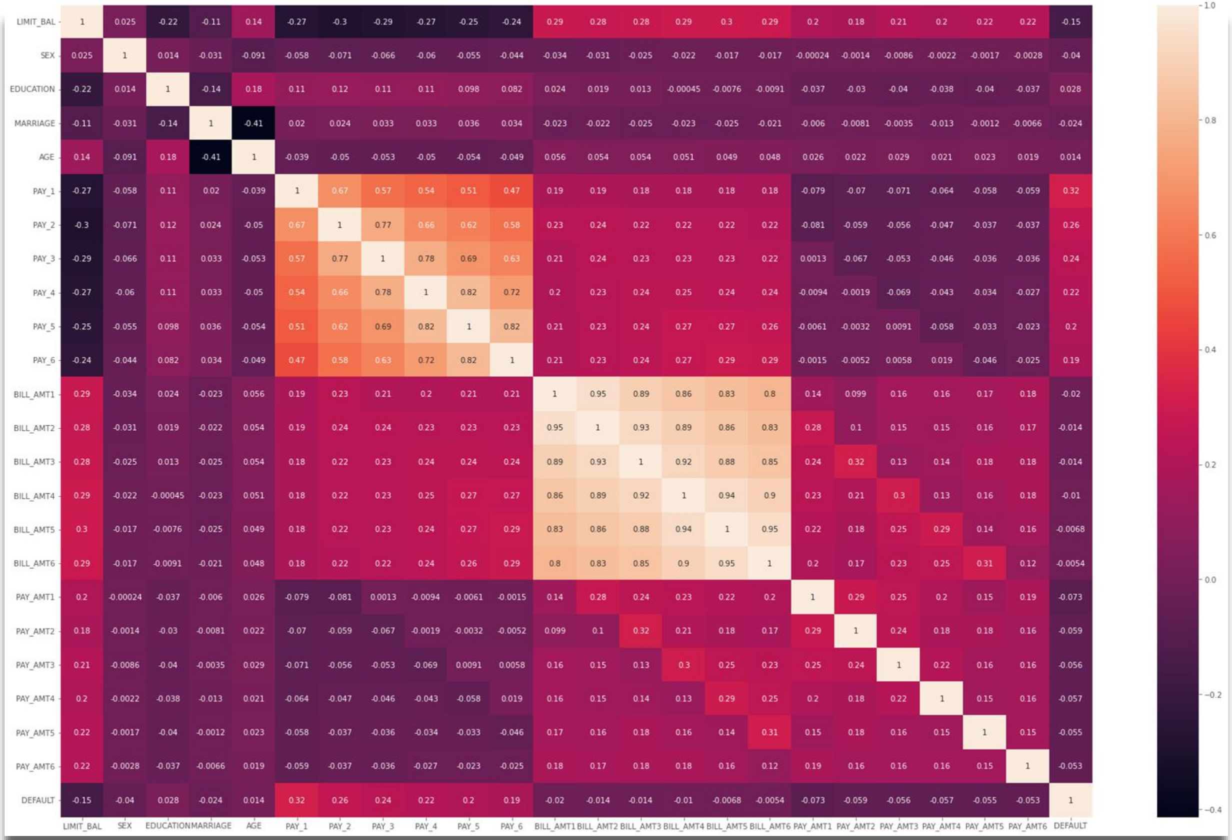
EXPLORATORY DATA ANALYSIS

What demographic factors impact payment default risk?

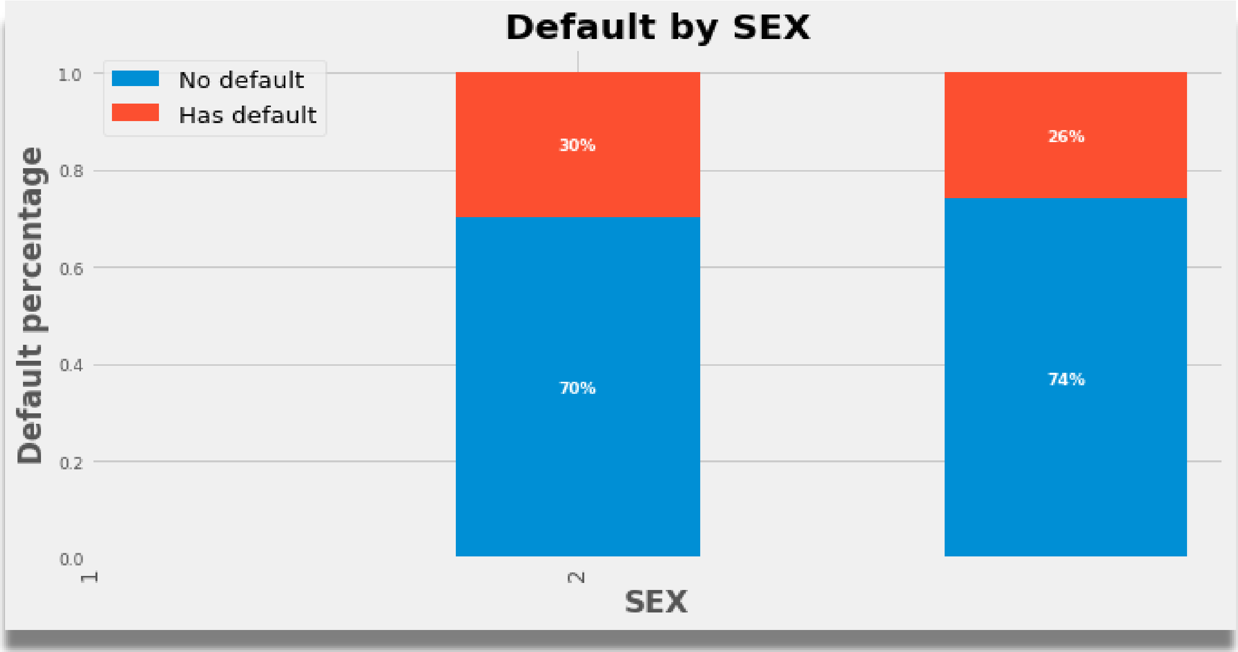
CHECKING THE CORRELATION OF DEFAULT VARIABLE WITH OTHER NUMERIC VARIABLES

PAY_1 to PAY_6 are highly correlated with our dependent variable 'DEFAULT'

Distribution of credit limit amounts. The three largest credit limit amount groups are $50k, $20k, and $30k, respectively.



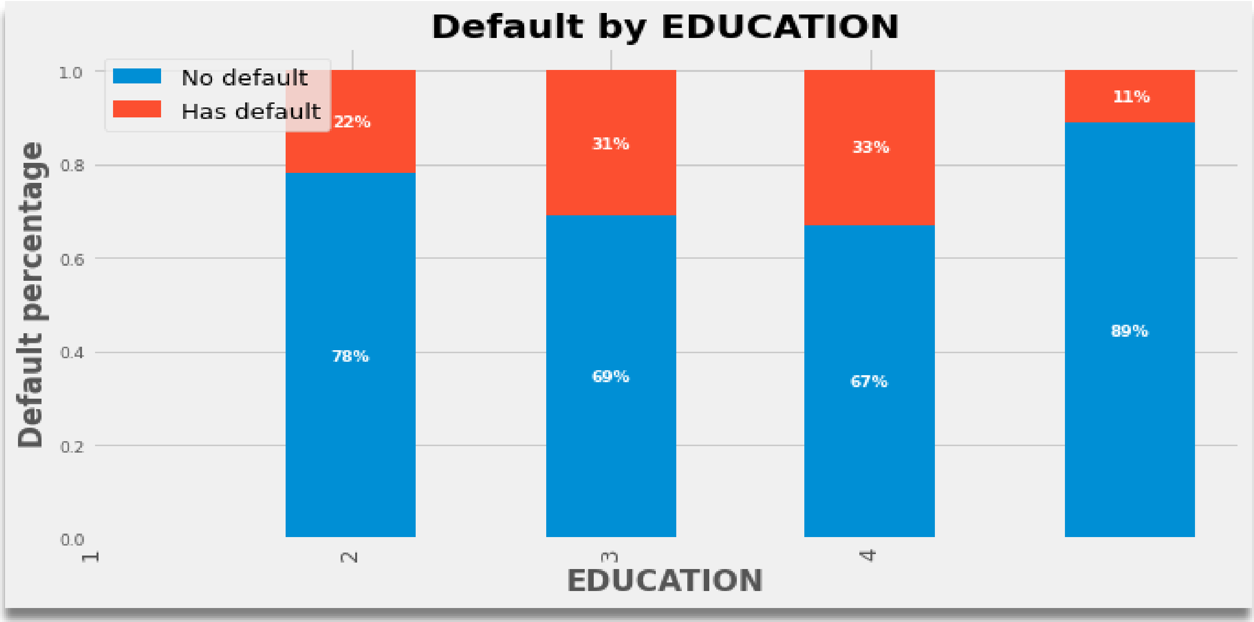
GENDER VARIABLE



30% of MALES and

26% of FEMALES have payment default.

EDUCATION VARIABLE

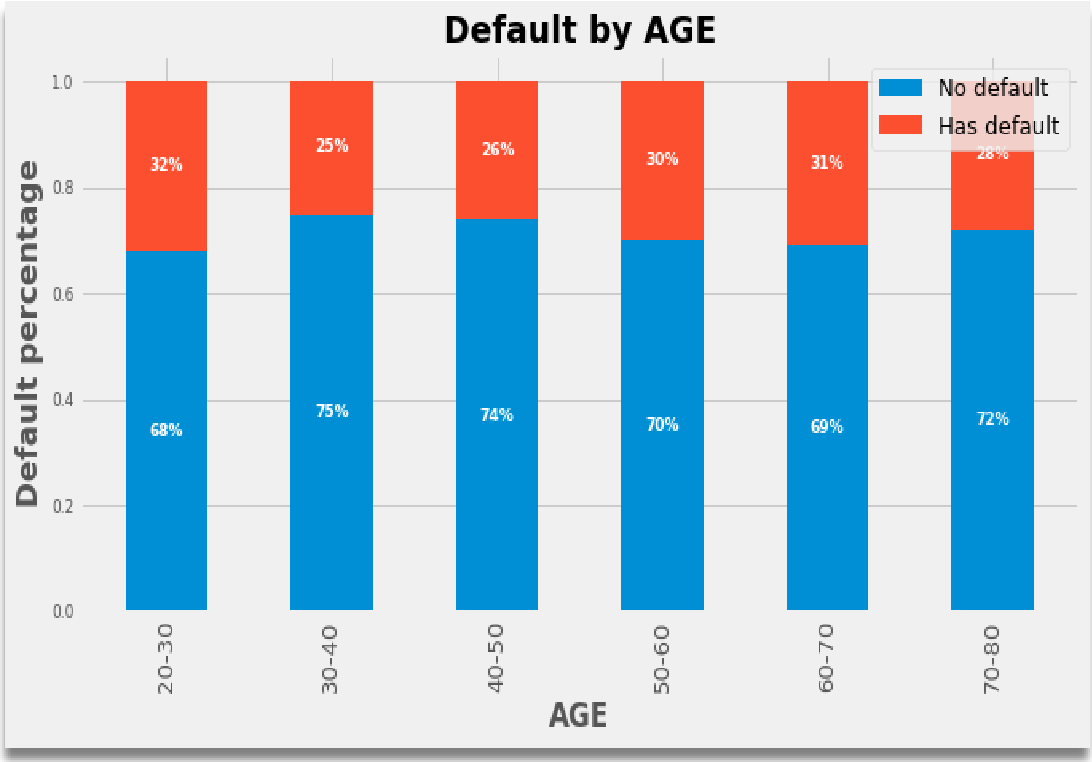


Higher education level, **lower** default risk.

1- graduate school, 2 - university, 3 - high school, 4 - Others

Others-4' only consists 1.56% of total customers even if they appear to have the least default.

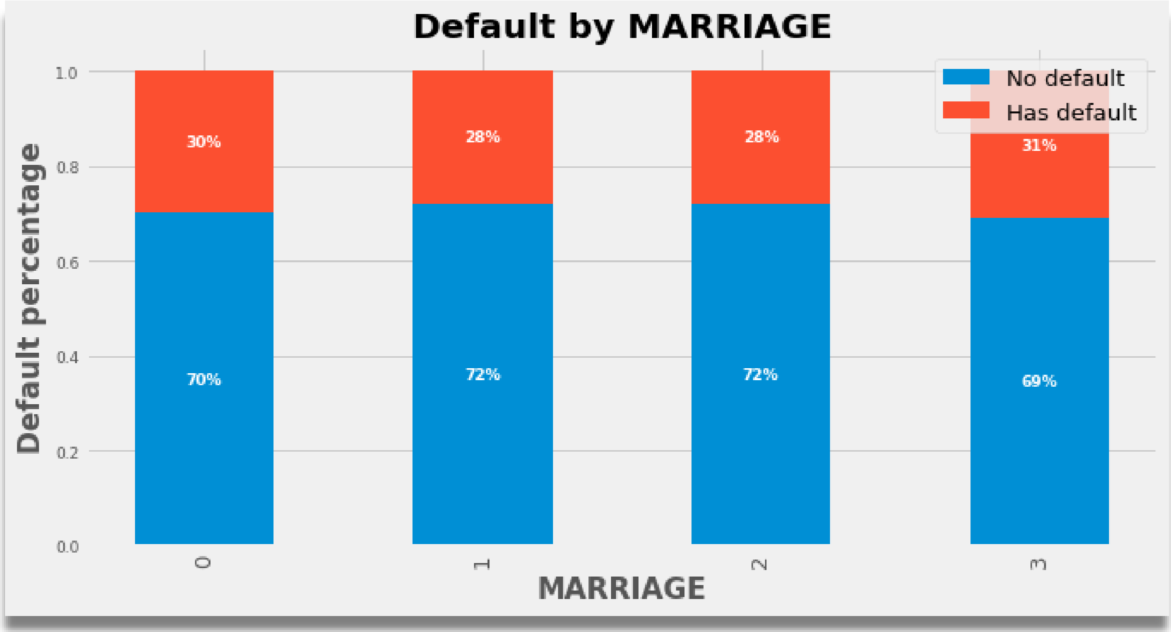
AGE VARIABLE



30-50: Lowest risk

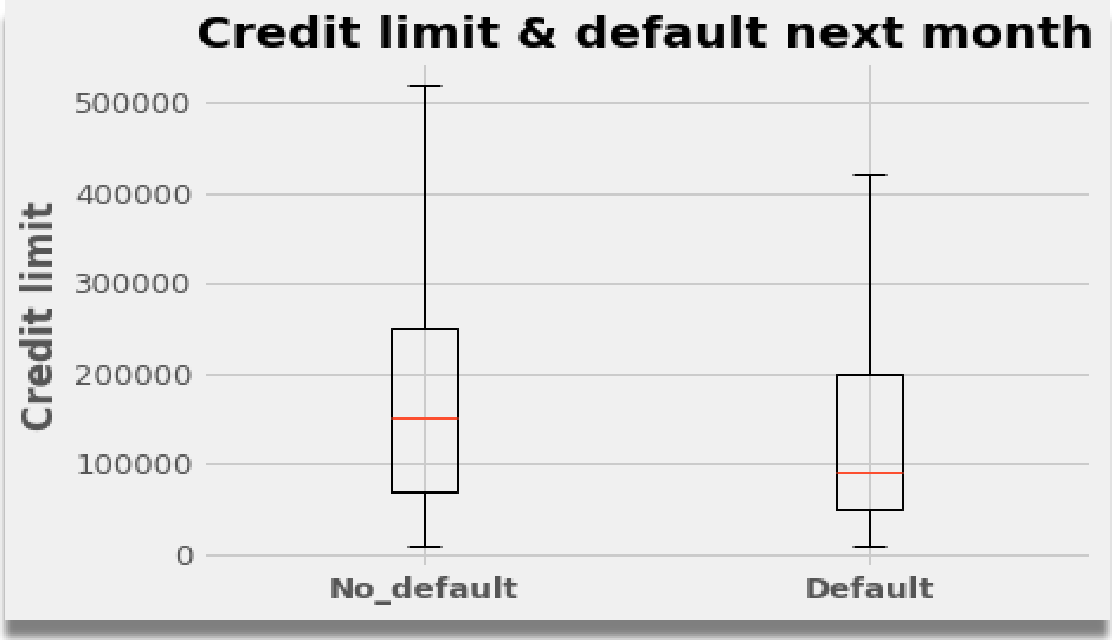
<30 or >50: Risk increases

MARITAL STATUS VARIABLE



There is no significant correlations of **default risk** and **marital status**.

CREDIT LIMIT VARIABLE



Higher
credit limits,

Lower
default risk.

EDA SUMMARY:

Demographic factors that impact default risk are:

1. **Education**: Higher education is associated with lower default risk.
2. **Age**: Customers aged 30-50 have the lowest default risk.
3. **Sex**: Females have lower default risk than males in this dataset.
4. **Marriage**: There appears to be no correlation between default payment and marital status.
5. **Credit limit**: Higher credit limit is associated with lower default risk.

PREDICTIVE MODELING

What precision and recall scores can the models achieve?

MODELING OVERVIEW:

1	DEFINE PROBLEM	Supervised learning/ binary classification
2	IMBALANCED CLASSES	78% non-default vs 22% default.
3	TOOLS USED	Scikit learn Library and imblearn
4	MODELS APPLIED	• Logistic Regression • Random Forest • 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 with the three models.

CORRECT IMBALANCED CLASSES

- Fit every model without and with SMOTE oversampling for comparison.
- Training AUC scores improved significantly with SMOTE.

MODELS	AUC without SMOTE	AUC with SMOTE
Logistic Regression	0.725	0.797
Random Forest	0.765	0.920
XGBoost	0.781	0.860

HYPERPARAMETERS TUNING

- **K-fold Cross Validation** to get average performance on the folds.
- **Randomised Search** on Logistic Regression since C has large search space.
- **Grid Search** on Random Forest on limited parameters combinations.
- **Randomised Search** on XGBoost because multiple hyperparameters to tune.

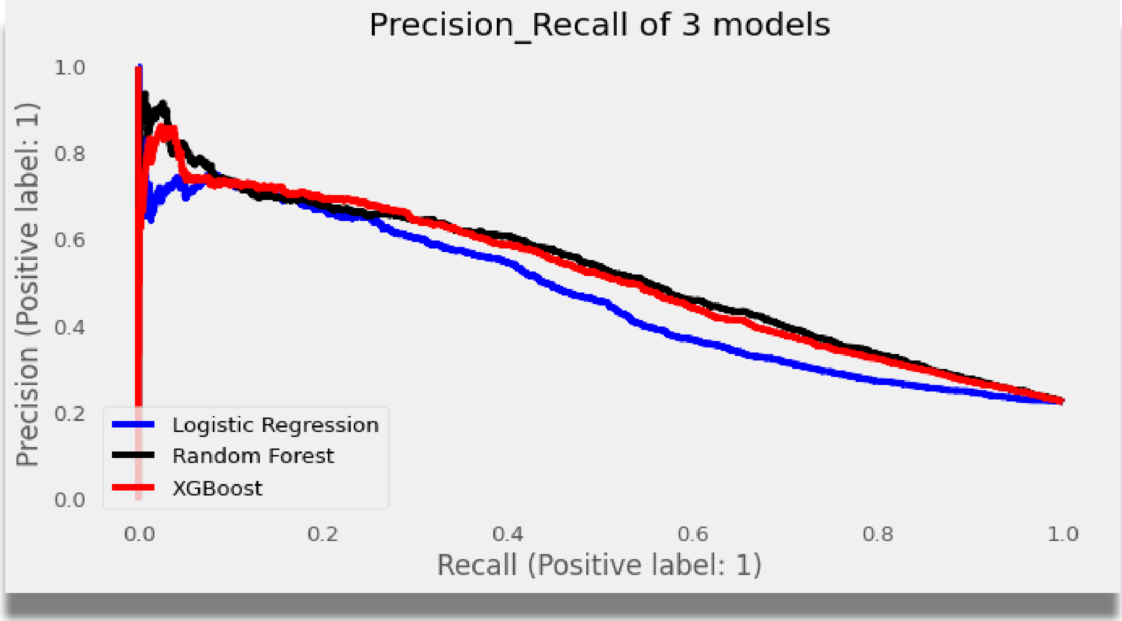
MODEL COMPARISONS

- Compare the models to Scikit-learn's dummy classifier.
- All models performed better than dummy model.

Models	Precision	Recall	F1 Score	Conclusion
Dummy Model	0.216	0.482	0.298	**Benchmark**
Logistic Regression	0.387	0.567	0.460	**Best Recall**
Random Forest	0.496	0.555	0.524	**Best F1**
XGBOOST	0.496	0.530	0.513	

MODEL COMPARISONS

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



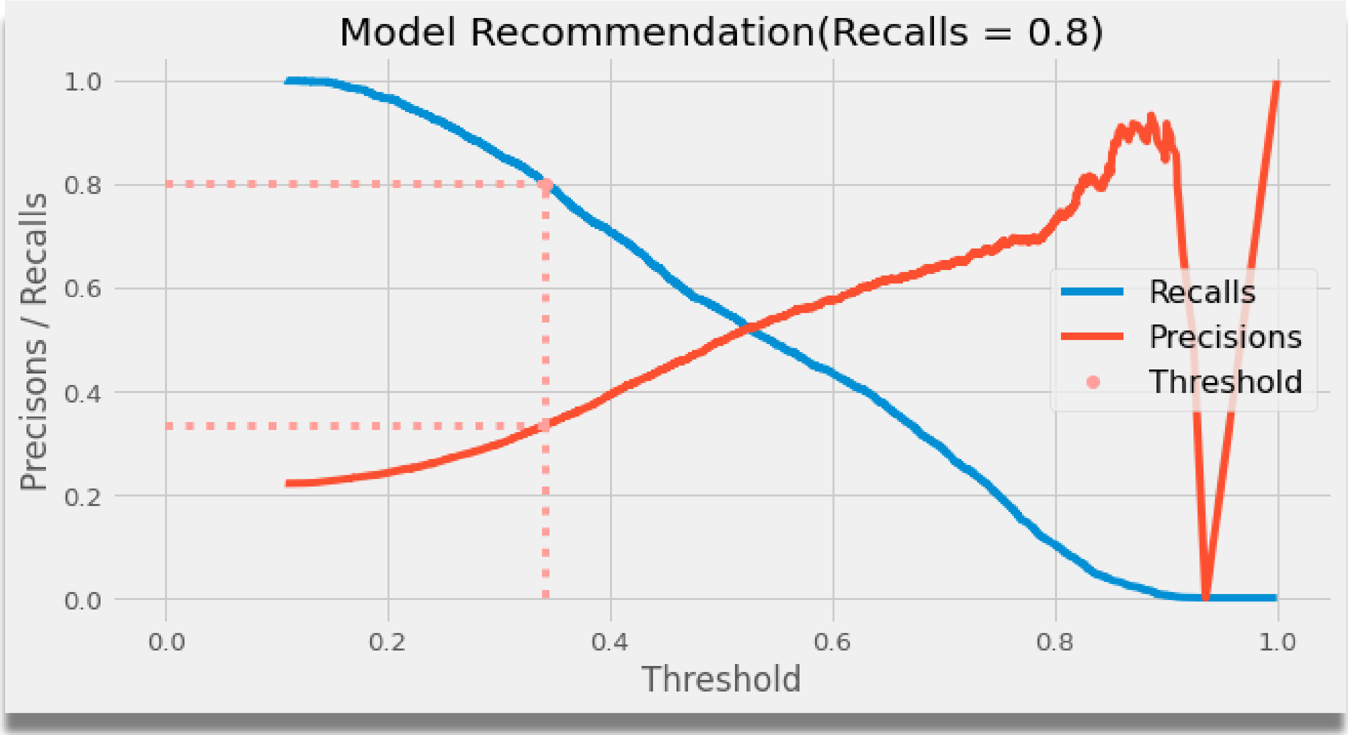
Terminology:

★ **Recall**: How many 1s are being defined?

★ **Precision**: Among all the 1s that are flagged, how many are truly 1s?

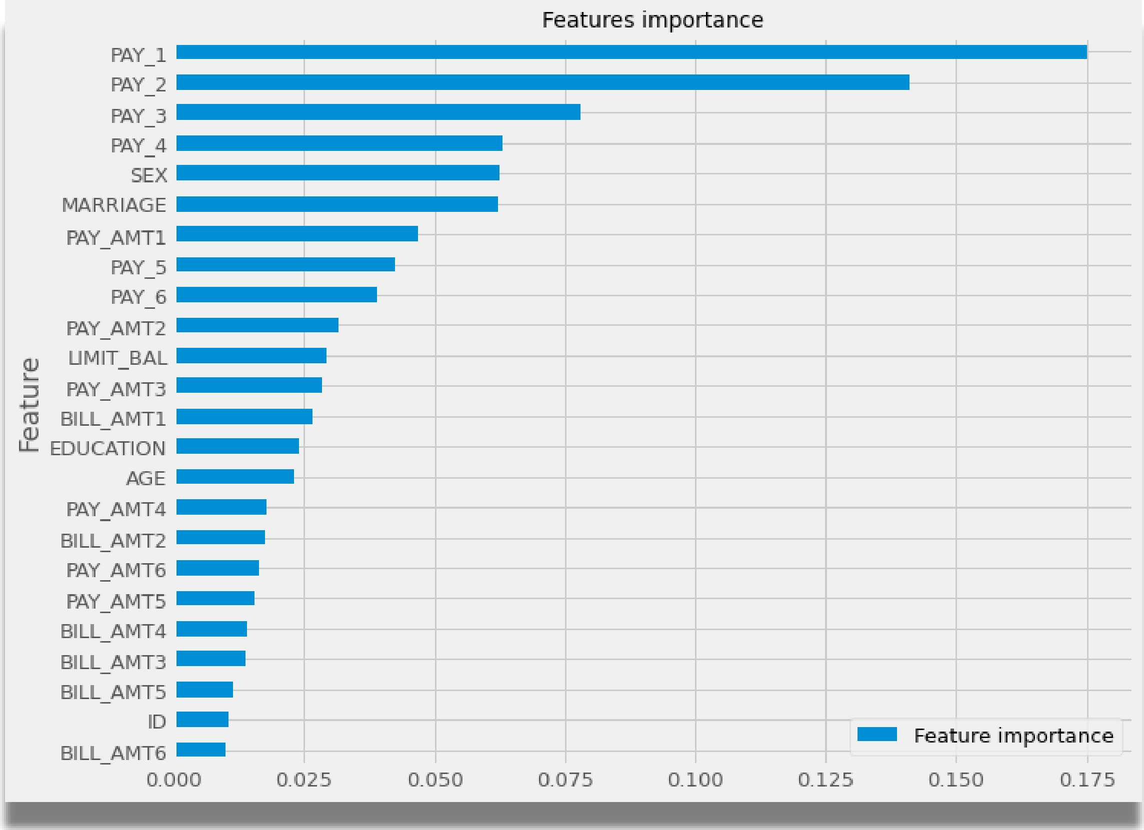
★ **Precision** and recall trade-off: high recall will cause low precision.

MODEL USAGE-RECOMMENDATION



Recall = 0.8. Threshold can be adjusted to reach higher recall.

FEATURE IMPORTANCES



Best model Random Forest feature importances plot.

PAY_1: most recent months payment status.
PAY_2: the month prior to current month's payment status.
BILL_AMT1: most recent month's bill amount.
LIMIT_BAL: credit limit

LIMITATIONS & FUTURE WORK

LIMITATIONS

- Best model Random Forest can only detect 51% of default.
- Model can only be served as an aid in decision making instead of replacing human decision.

FUTURE WORK

- Models are not exhaustive. Other models could perform better.
- Get more computational resources to tune XGBoost parameters.
- Incorporate datasets from different countries.

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.

THANK YOU