Credit Card Default Prediction

Predicting whether a customer will default on his/her credit card

Team Members

Siddhi Thakur
Mohammad Jibran

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.

Content

- Data Summary
- Exploratory Data Analysis
- Predictive Modelling
    - Modeling Overview
    - Correct Imbalanced Classes
    - Hyperparameters Tuning
    - Model Comparisons
    - Model – Recommendation
- Conclusions

Data Summary

- This dataset contains information on default payments, demographic factors, credit limit, history of payments, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. It includes 30,000 rows and 25 columns, and there is no credit score or credit history information.

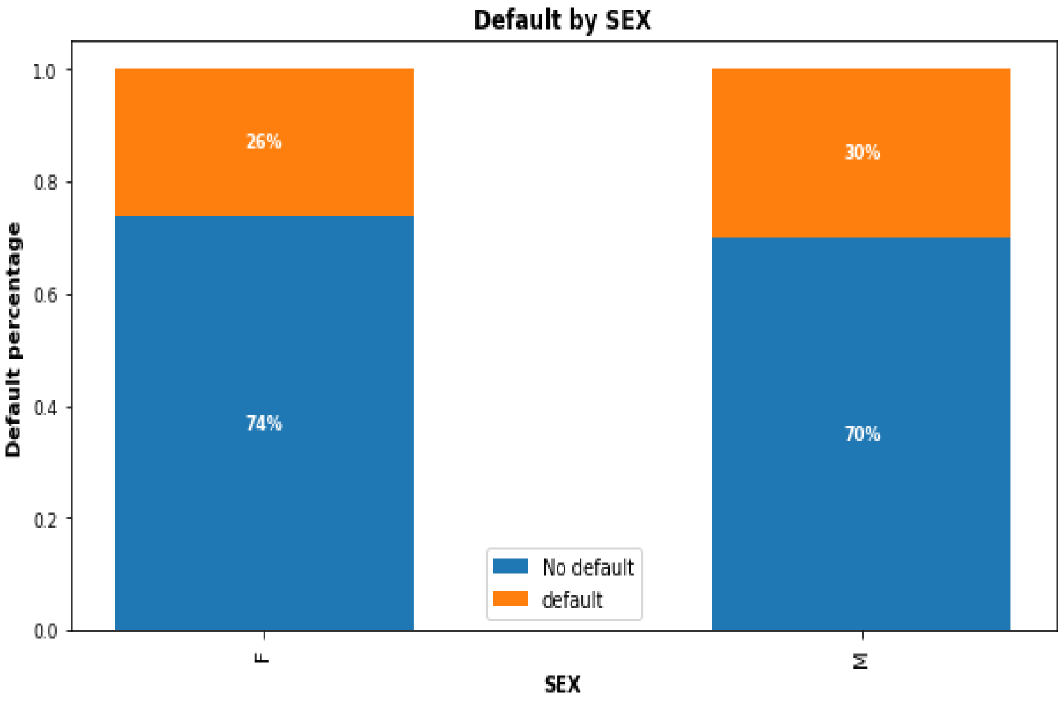
- Overall, the dataset is very clean, but there are several undocumented column values. As a result, most of the data wrangling effort was spent on searching information and interpreting the columns.

```
df.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
```

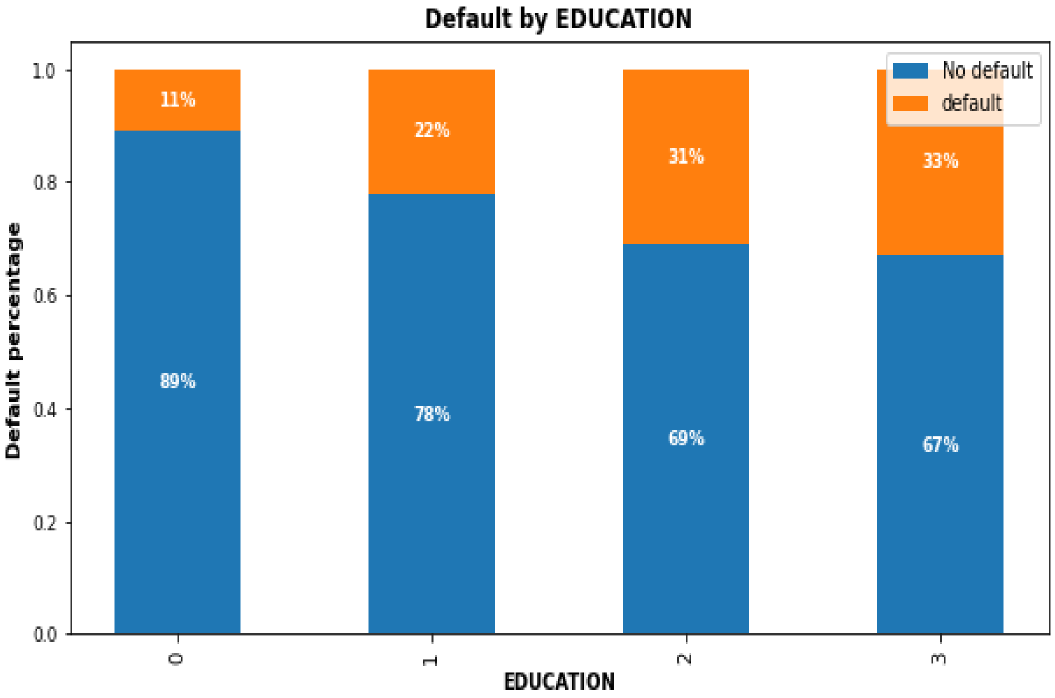
Exploratory Data Analysis

Which sex group tends to have more delayed payments?



- 30% of males and 26% of females have payment default.
- The difference is not significant.

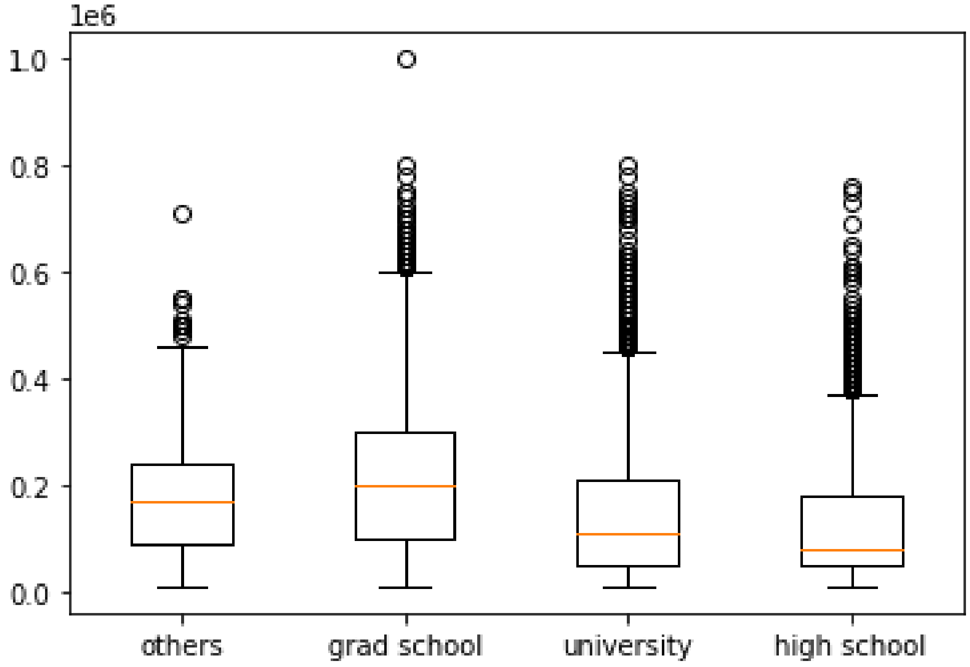
Did customers with higher education have less delayed payment?



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

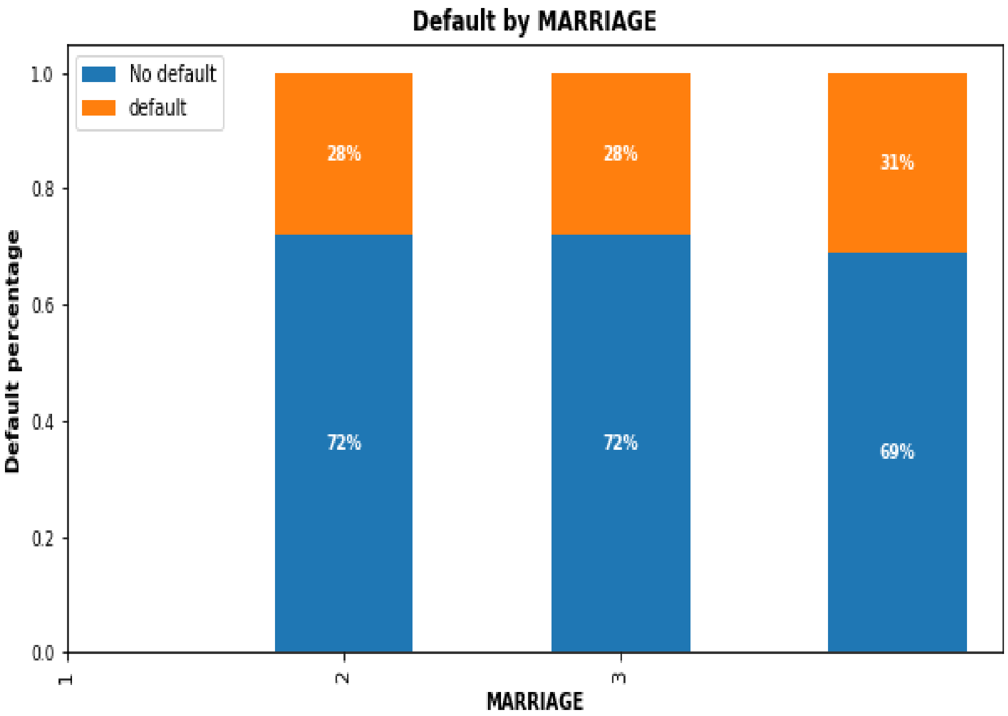
- **Notice there is an education group "others" which appears to have the least default payment, but this group only has 468 (or 1.56%) customers, and we don't know what consists of this group.**

Did customers with a high education level get higher credit limits?



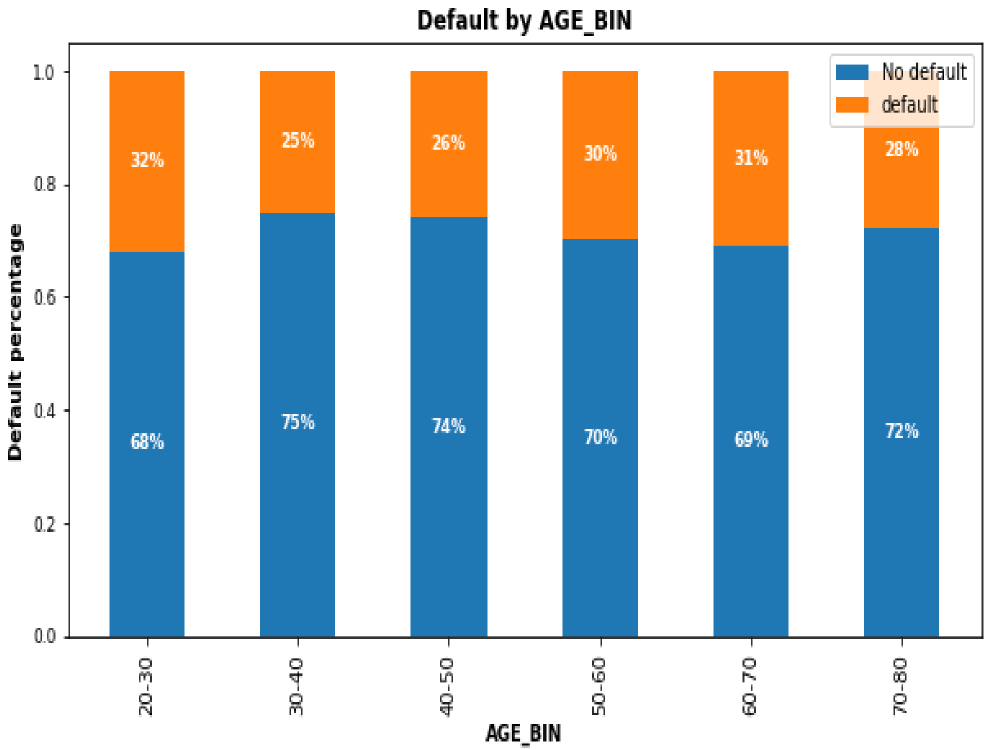
- Customers with grad school education have the highest 25% percentile, highest median, highest 75th percentile and highest maximum numbers, which suggests customers with higher education levels do get higher credit limits.

Does marital status have anything to do with default risk?



- No significant correlations of default risk and marital status.

Do younger people tend to miss the payment deadline?



- The default probability increases for customers younger than 30.

- Customers aged between 30 and 50 have the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all have higher delayed payment rates.

Predictive Modeling

Modeling Overview

- **Define Problem –** Supervised learning / Binary Classification
- **Imbalanced Classes –** 78% non-default vs 22% default
- **Models used –** Logistic Regression
  Random Forest
  XGBoost

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	0.766	0.919
XGBoost	0.762	0.892

Hyperparameters Tuning

- **Randomized Search** on **Logistic Regression** since C has large search space.
- **Grid Search** on **Random Forest** on limited parameters combinations
- **Randomized Search** on **XGBoost** because multiple hyperparameters to tune.

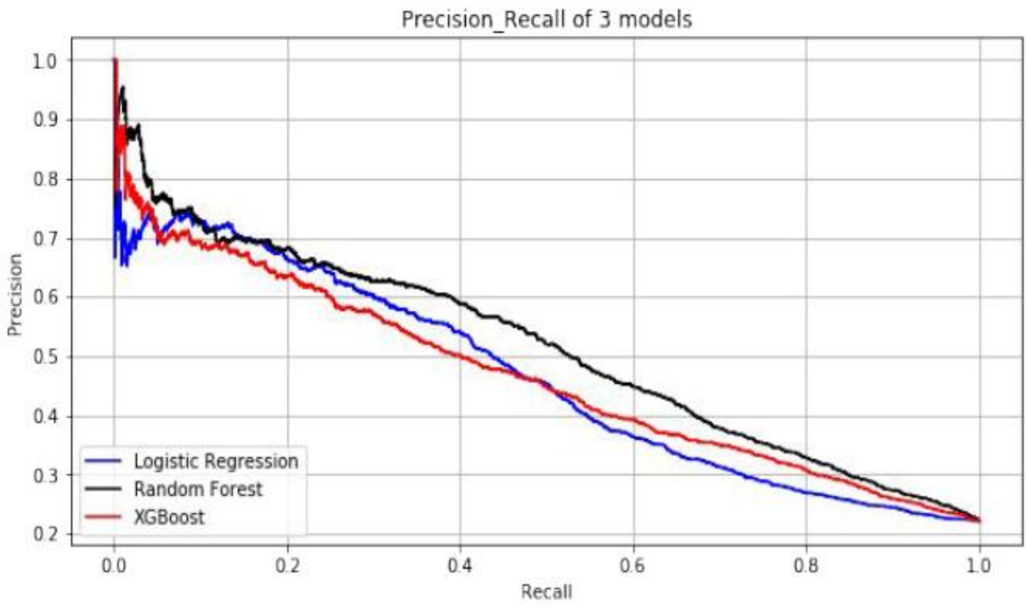
Model Comparisons

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

Models	Precision	Recall	F-1 Score	Conclusion
Dummy Model	0.217	0.500	0.303	Benchmark
Logistic Regression	0.379	0.561	0.453	Best recall
Random Forest	0.527	0.505	0.516	Best F1
XGBoost	0.444	0.501	0.474	

Model Comparisons

- Compare within 3 models
- **Random Forest** (black line) has the best Precision_Recall score



Terminology

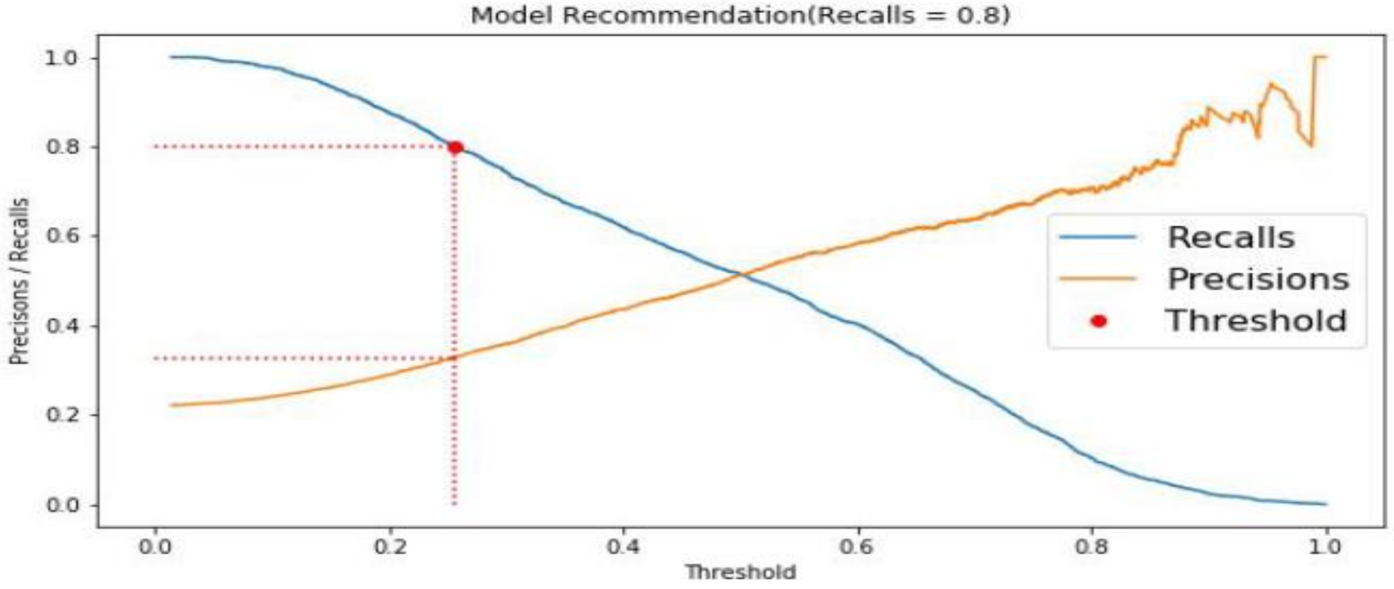
- ❖ **Recall -** How many +ve are being identified?
- ❖ **Precision** - Among all the +ve results, how many are truly +ve?

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

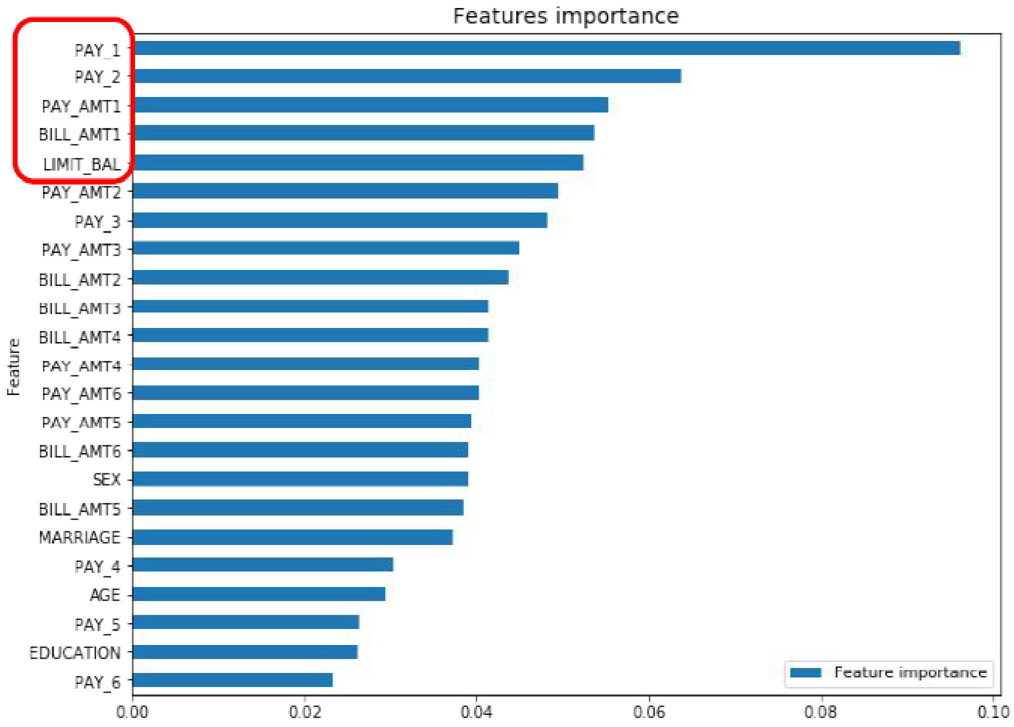
Model - Recommendation

- **Recommended Recall = 0.8**

  Threshold can be adjusted to reach higher recall.



Feature Importance



- **Best model Random Forest feature importance plot –**

- **PAY_1:** most recent month's 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

Conclusions

- Logistic Regression model has the highest recall but the lowest precision, if the business cares recall the most, then this model is the best candidate.
- If the balance of recall and precision is the most important metric, then Random Forest is the ideal model. Since Random Forest has slightly lower recall but much higher precision than Logistic Regression, I would recommend Random Forest.
- Recent 2 payment status and credit limit are the strongest default predictors.
- Random Forest has the best precision and recall balance.
- Higher recall can be achieved if low precision is acceptable.

Thank You