Capstone Project

Credit Card Default Prediction

By – Rohit Verma

Contents

Data pipelines

The data pipeline can be briefly summarized in the following five steps:

• **Data Preprocessing:** Here, we need to load the data and understand the features present in it.

• **Exploratory data analytics (EDA):** In this step, we need to perform univariate and bivariate analyses of the data, followed by feature transformations, if necessary.

• **Train/Test Split:** Train/test split, which you we can perform in order to check the performance of your models with unseen data. Here, for validation, we can use the k-fold cross-validation method.

•**Hyperparameter Tuning:** This is the final step at which we can try different models and fine-tune their hyperparameters until we get the desired level of performance on the given dataset.

• **Model Evaluation:** Evaluate the models using appropriate evaluation metrics. Note that since the data is imbalanced it is is more important to identify which are fraudulent transactions accurately than the non-fraudulent. Choose an appropriate evaluation metric which reflects this business goal.

Goal

- This project aims to apply different algorithms & techniques on Credit Card Fraud data set and compare the results.
- Measures used to compare those are Area Under the Curve, False (Alarm) Positive Rate and Recall (Detection Rate).
- Measure like Accuracy is not good because the data set is highly unbalanced.

Problems to Resolve

Problem Statement

- ML applications focused on credit score predicting.
- Relying on credit scores and credit history.
- Miss valuable customers with no credit history. I.e. immigrants.
- Regulatory constraints on banking industry forbids some ML algorithms.

Purpose of Project

- Conduct quantitative analysis on credit default risk by applying three interpretable machine learning models without utilizing credit score or credit history.

Who Should Care?

Credit Card Companies

Commercial Banks

Data Acquisition

Dataset

- Default Payments of Credit Card Clients in Taiwan from 2005
- Source: Public dataset from Kaggle.
- Original Source: UCI Machine Learning Repository*

Why This Dataset?

- Real credit card data
- Comprehensive and complete
- 30,000 customers
- Usage of 6 months
- Age from 20-79
- Demographic factors
- No credit score or credit history

Dataset

- The datasets contains transactions made by credit cards. This dataset presents transactions that occurred in two days, where we have **492** frauds out of **284,807** transactions. The dataset is highly unbalanced, the positive class (frauds) account for **0.172%** of all transaction.

- Due to confidentiality reasons, dataset available is not the original (raw) form but actually has been reduced using PCA. And the only features which have not been transformed with PCA are 'Time' and 'Amount'.

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
- Random Forest
- XGBoost

PART - 1

Data Preprocessing

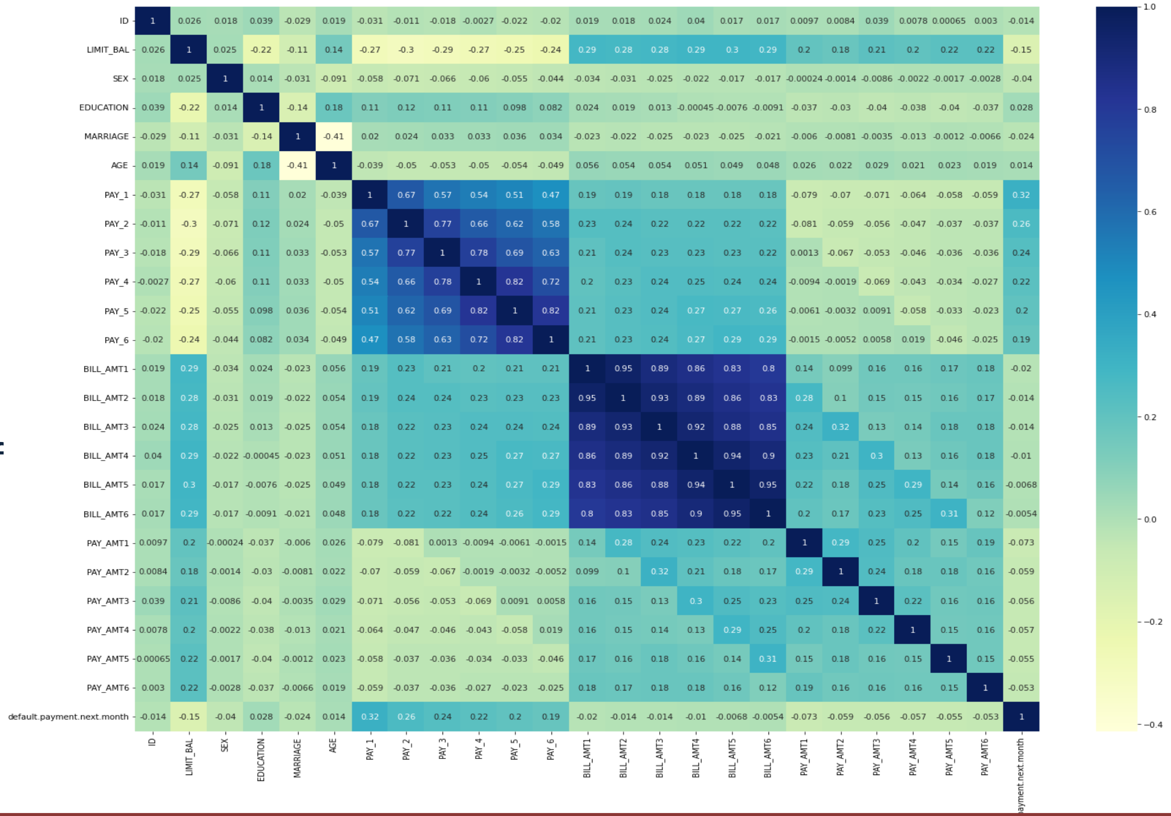
Correlation Analysis:

- It is essential to view attribute correlations to select the best features for modeling

- The visual to the right conveys the correlations of the remaining attributes

  Color = +/- Correlation

  Size = Intensity of Correlation



df.head()

- There are 5 rows and 25 columns

- Columns which are related with default transations are

1. Gender
2. Education
3. Marriage
4. Age
5. Credit_limit

[ ] df.head()

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3
0	1	20000	2	2	1	24	2	2	-1	-1	...	0	0	0	0	689	0
1	2	120000	2	2	2	26	-1	2	0	0	...	3272	3455	3261	0	1000	1000
2	3	90000	2	2	2	34	0	0	0	0	...	14331	14948	15549	1518	1500	1000
3	4	50000	2	2	1	37	0	0	0	0	...	28314	28959	29547	2000	2019	1200
4	5	50000	1	2	1	57	-1	0	-1	0	...	20940	19146	19131	2000	36681	10000

5 rows × 25 columns

Check Null values:

Null values

- We start our exploration by reducing the number of columns based on the previous criteria

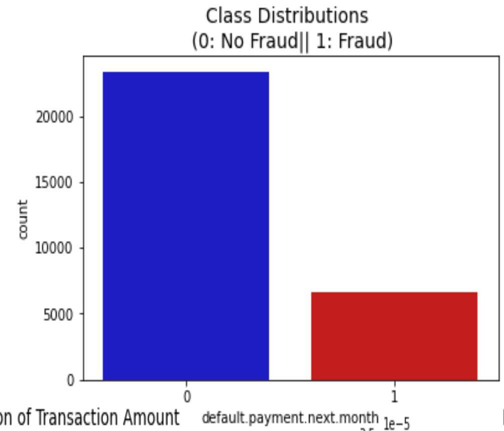
- Checking null or unnecessary values,we observed that there is no null values present in our dataset.

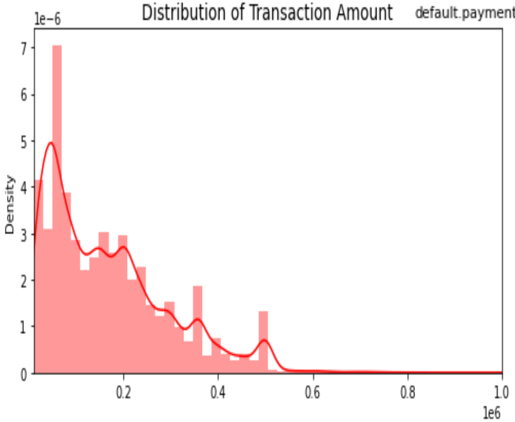
```
[ ] #check if any null values
    df.isnull().sum()

    ID                            0
    LIMIT_BAL                     0
    SEX                           0
    EDUCATION                     0
    MARRIAGE                      0
    AGE                           0
    PAY_1                         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
```

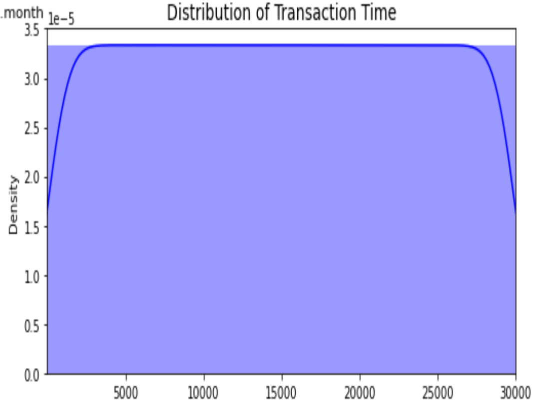
Fraud/Non Fraud:

- These graphs clearly shows that there are **23364** non fraud transactions and **6636** fraud trasactions.

- It's also shows that the overall default payments are **22.12%**





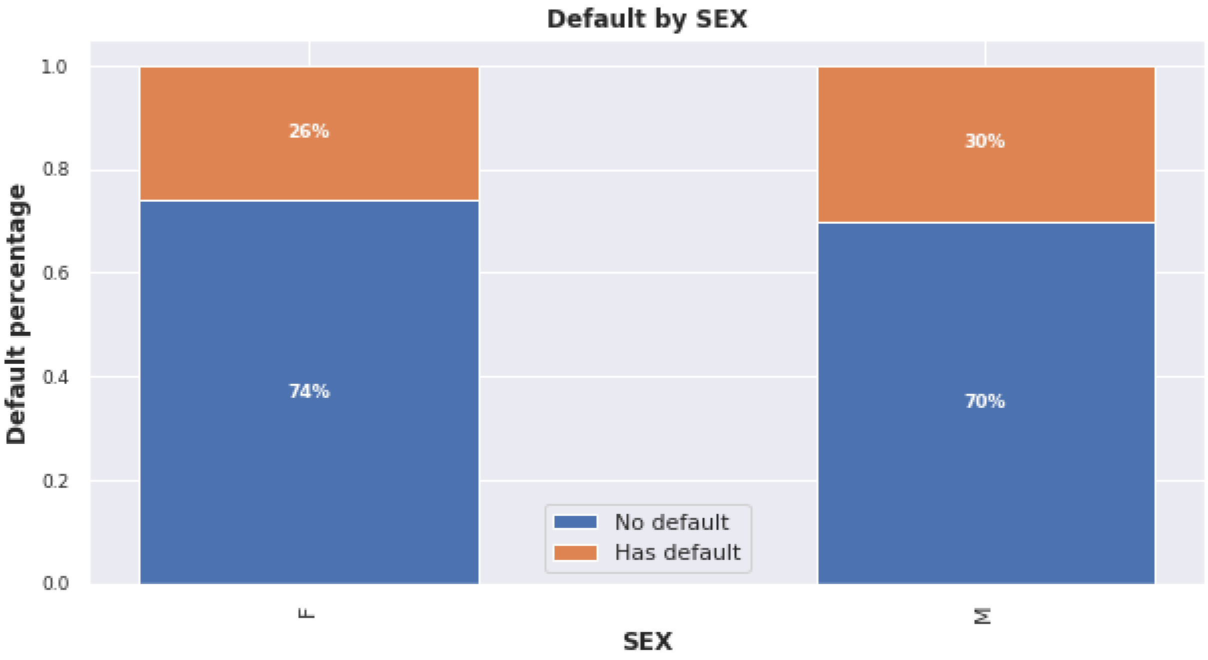


PART - 2

Exploratory Data Analysis

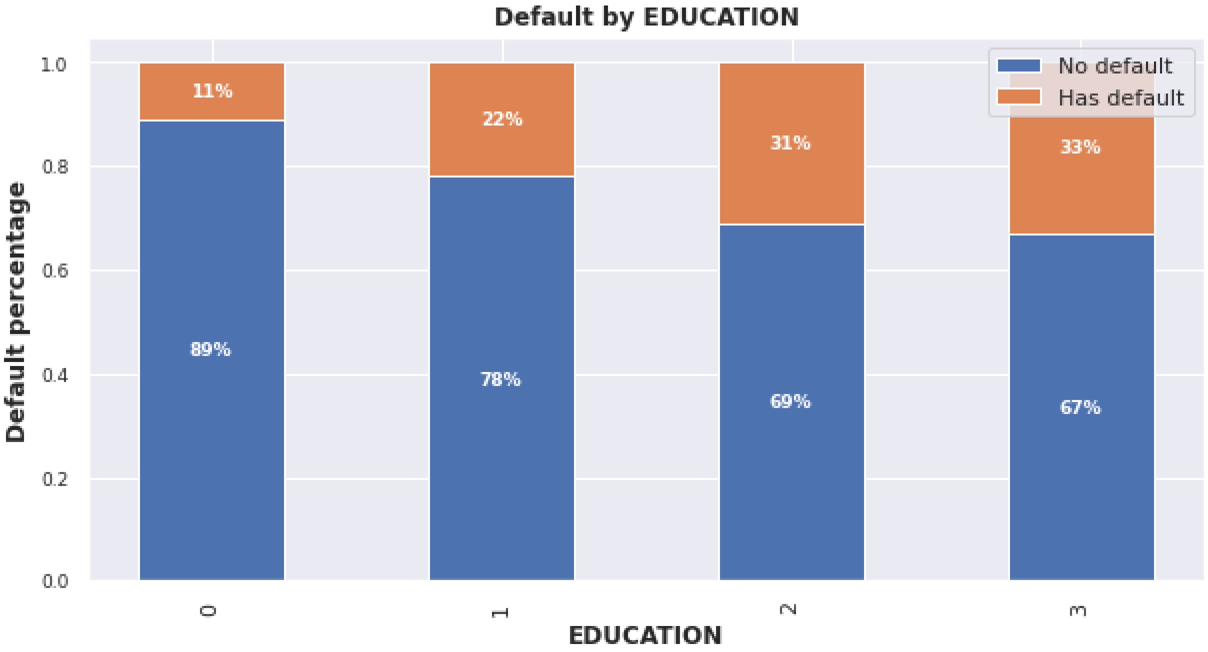
What demographic factors impact payment default risk?

Gender Variable



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

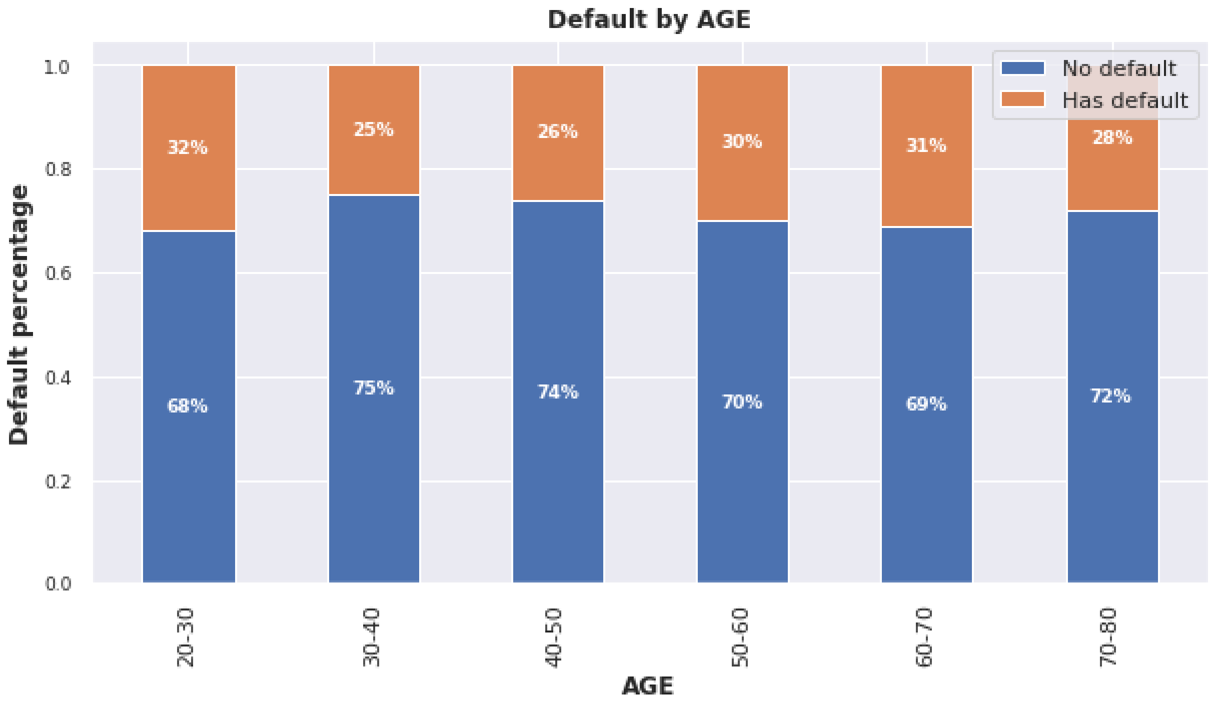
Education Variable



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

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

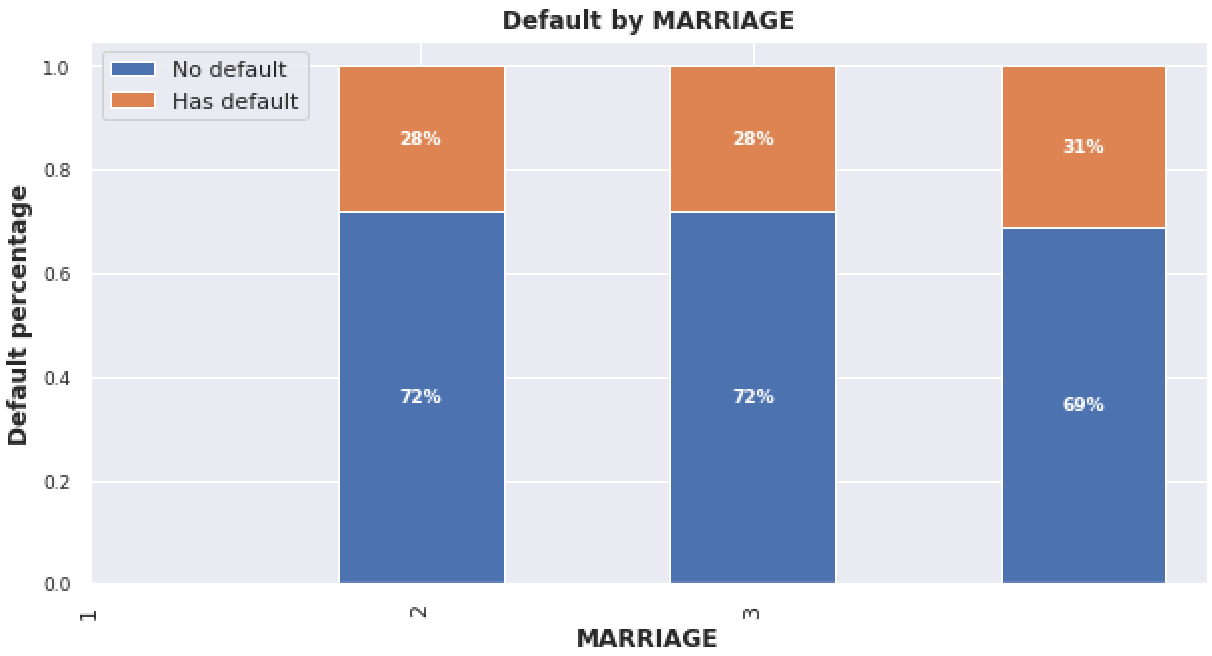
Age Variable



30-50:
Lowest risk

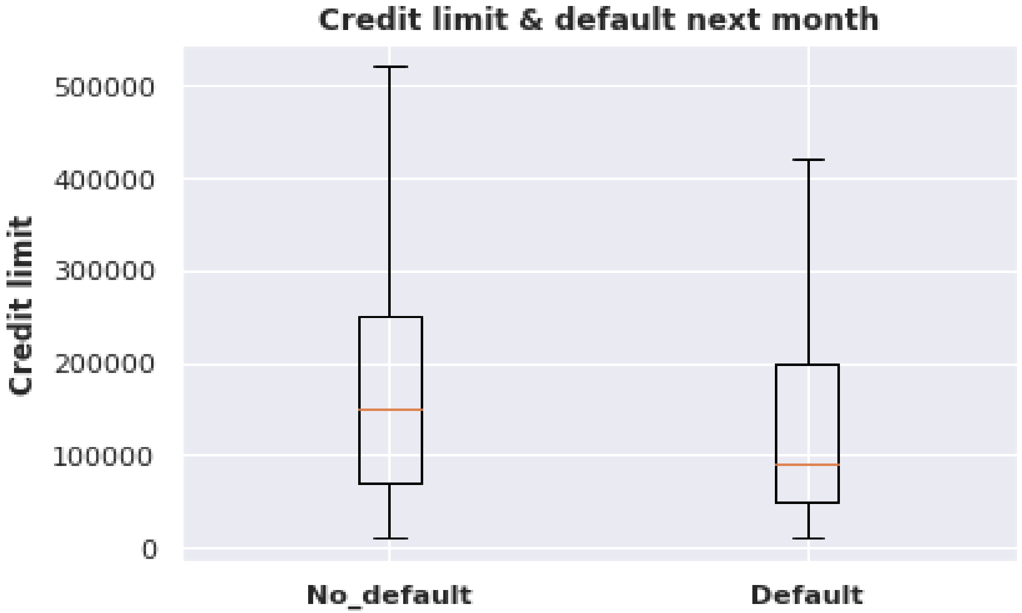
< 30 or < 50:
Risk increases

Marital Status Variable



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 payment risk are:
    - Education: Higher education is associated with lower default risk.
    - Age: Customers aged 30-50 have the lowest default risk.
    - Sex: Females have lower default risk than males in this dataset.
    - Credit limit: Higher credit limit is associated with lower default risk.

PART - 3

Predective Modeling

What precision and recall scores can the models achieve?

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 / 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

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	0.726	0.797
Random Forest	0.764	0.916
XGBoost	0.762	0.899

Hyperparameters Tuning

- **K-Fold Cross Validation** to get average performance on the folds.
- **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.

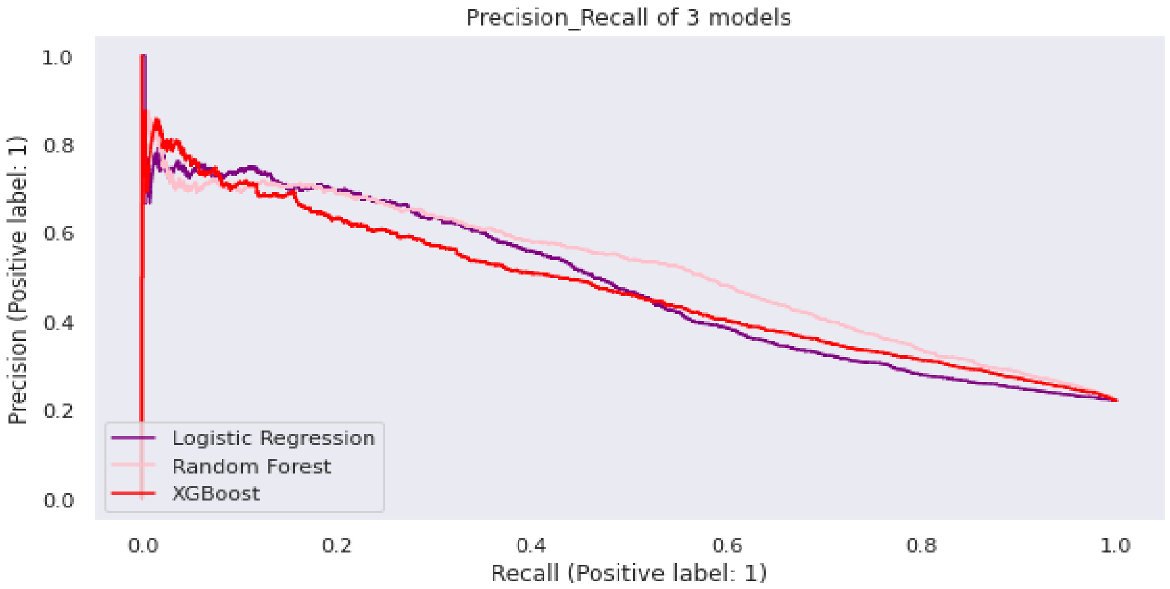
Model Comparisons

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

Models	Precision	Recall	F1 Score	Conclusion
Dummy Model	0.217	0.500	0.303	**Benchmark**
Logistic Regression	0.384	0.566	0.457	**Best recall**
Random Forest	0.513	0.514	0.514	**Best F1**
XGBoost	0.444	0.505	0.474	

Model Comparisons

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



Terminology:

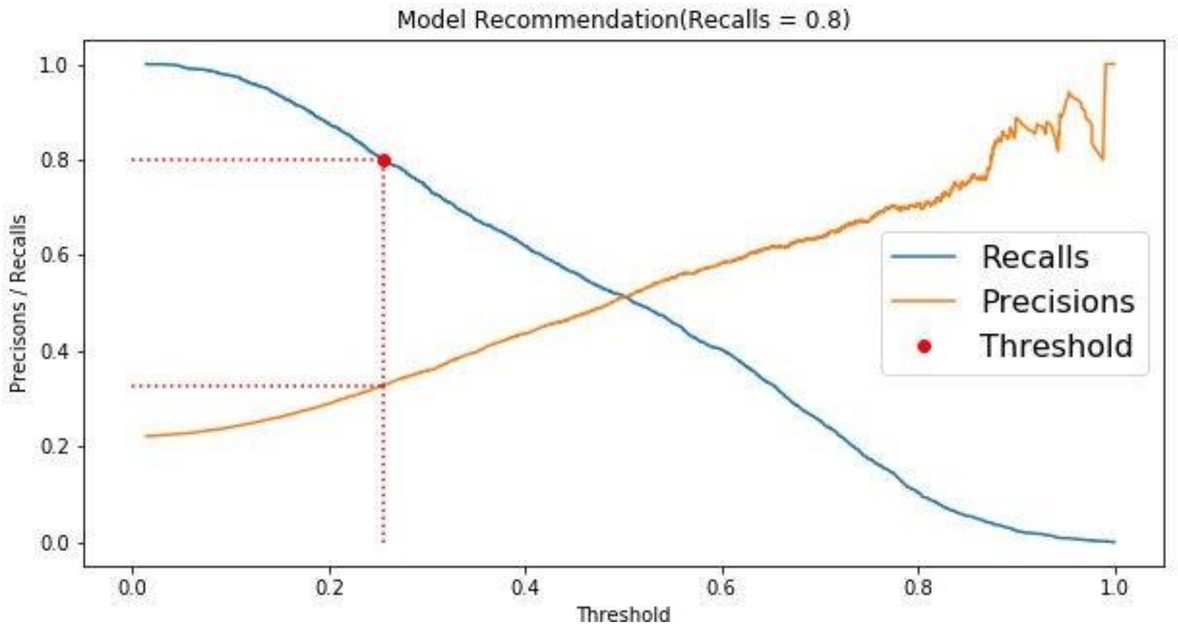
★ Recall: how many 1s are being identified?

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

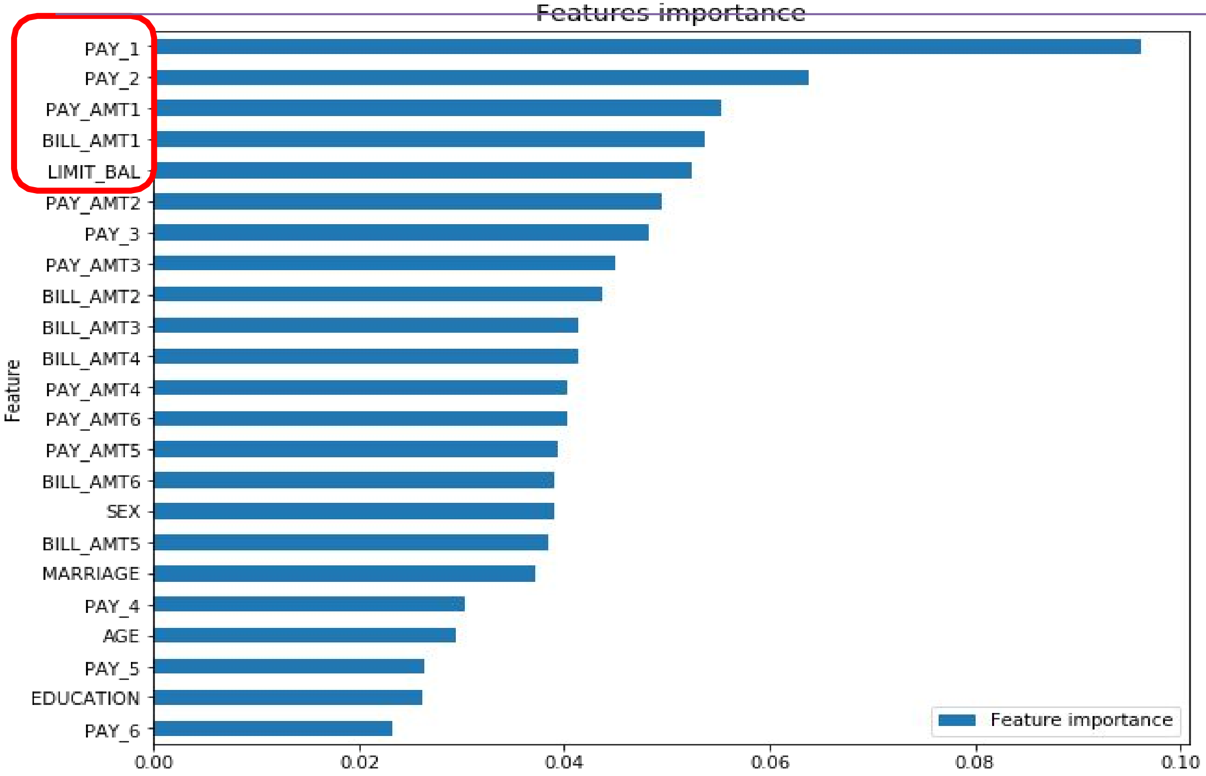
★ Precision and recall trade-off: high recall will cause low precision

Model Usage - Recommendation

- I.e. recall = 0.8. Threshold can be adjusted to reach higher recall.



Feature Importances



Best model Random Forest feature importances 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

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.
- Used only 30,000 records and not from US consumers.

Future Work

- Models are not exhaustive. Other models could perform better.
- Get more computational resources to tune XGBoost parameters.
- Acquire US customer data and more useful features.I.e.customer income.

Conclusions

- 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

Any Query?