CREDIT CARD DEFAULT

Feifei Yan

- Introduction
- Pre-processing
- Data visualization
- Data preparation
- Model selection
- Model performance

Problem

- Taiwan credit card crisis
- Predict customer’s default payment status for next month

Dataset

Data source: Machine Learning Repository **Time range**: from April,2005 to Sept,2005

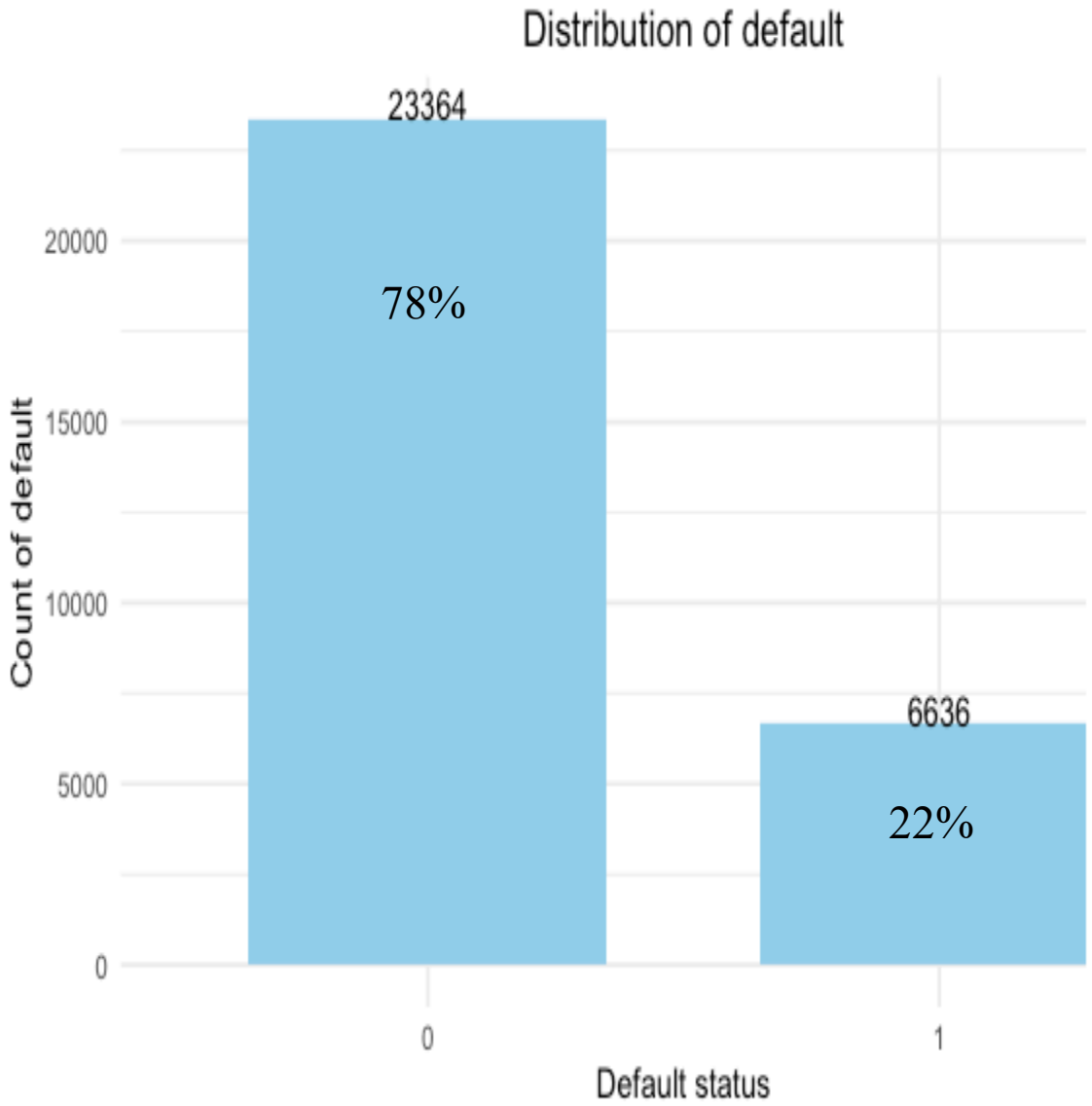
Attributes (24 columns, 30000 rows)

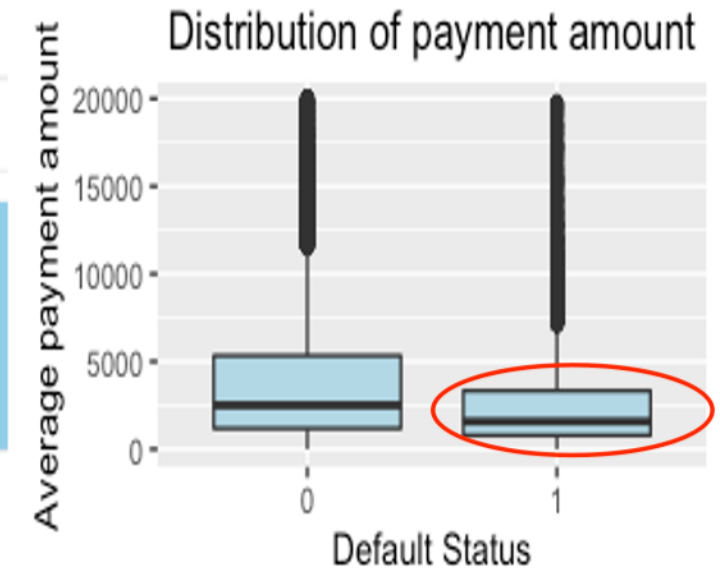
- default: default payment status (1=yes, 0=no)
- LIMIT_BAL: Amount of given credit
- SEX: Gender (1=male, 2=female)
- AGE: 21 to 79
- MARRIAGE: Marital status (0=unknown, 1=married, 2=single, 3=others)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 0,5,6=unknown)
- PAY_0, PAY_2,…PAY_6: History of past repayment status for each month (-2,0=unknown, -1=pay duly, 1=payment delay for one month, 2=payment delay for two months etc)
- BILL_AMT1, BILL_AMT2, …BILL_AMT6: Amount of bill for each month
- PAY_AMT1, PAY_AMT2, … PAY_AMT6: Amount of previous repayment for each month

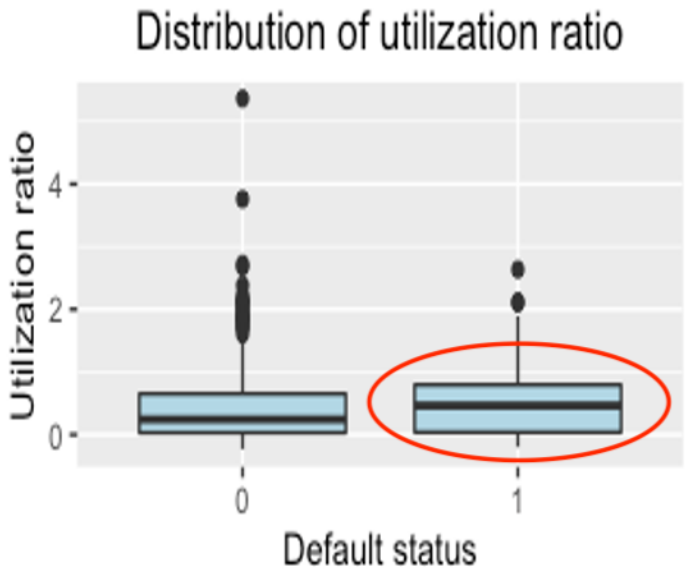
Data Pre-processing

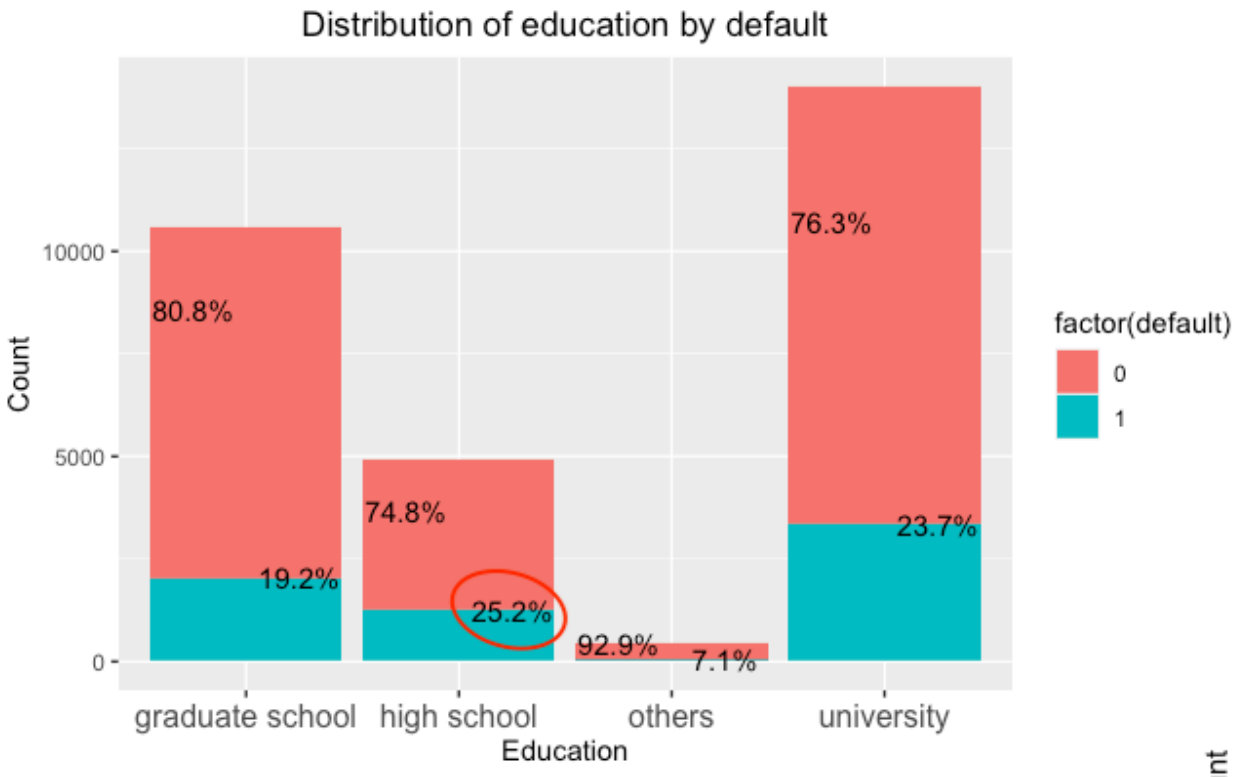
- Grouping: unknown values in education and marriage
- Create four new variables: number of missed payment, average bill amount, average repayment amount, utilization ratio

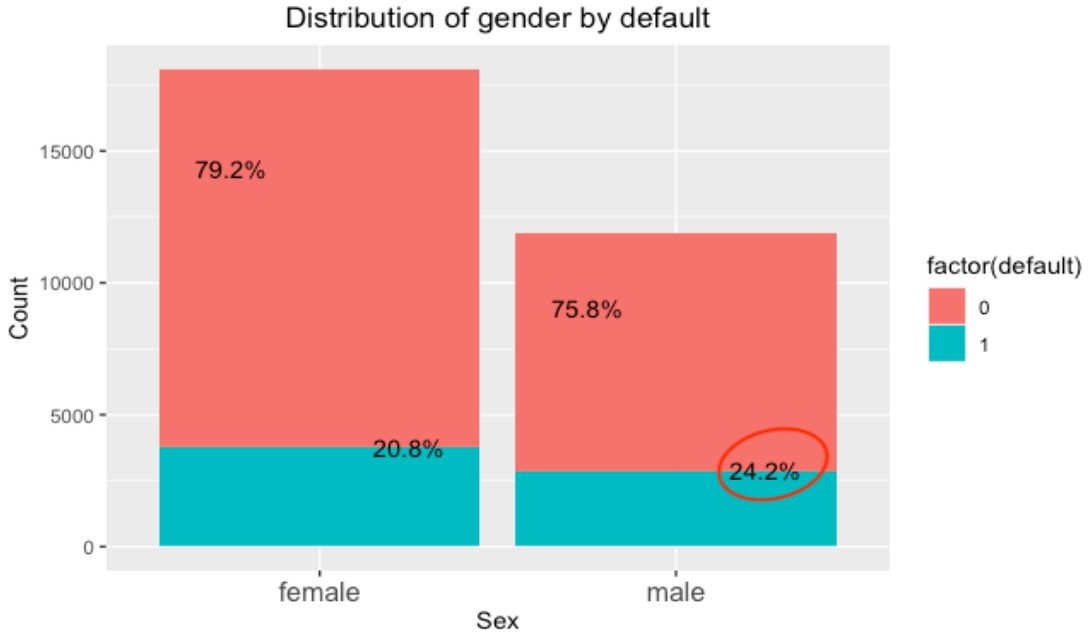
Exploratory Data Analysis

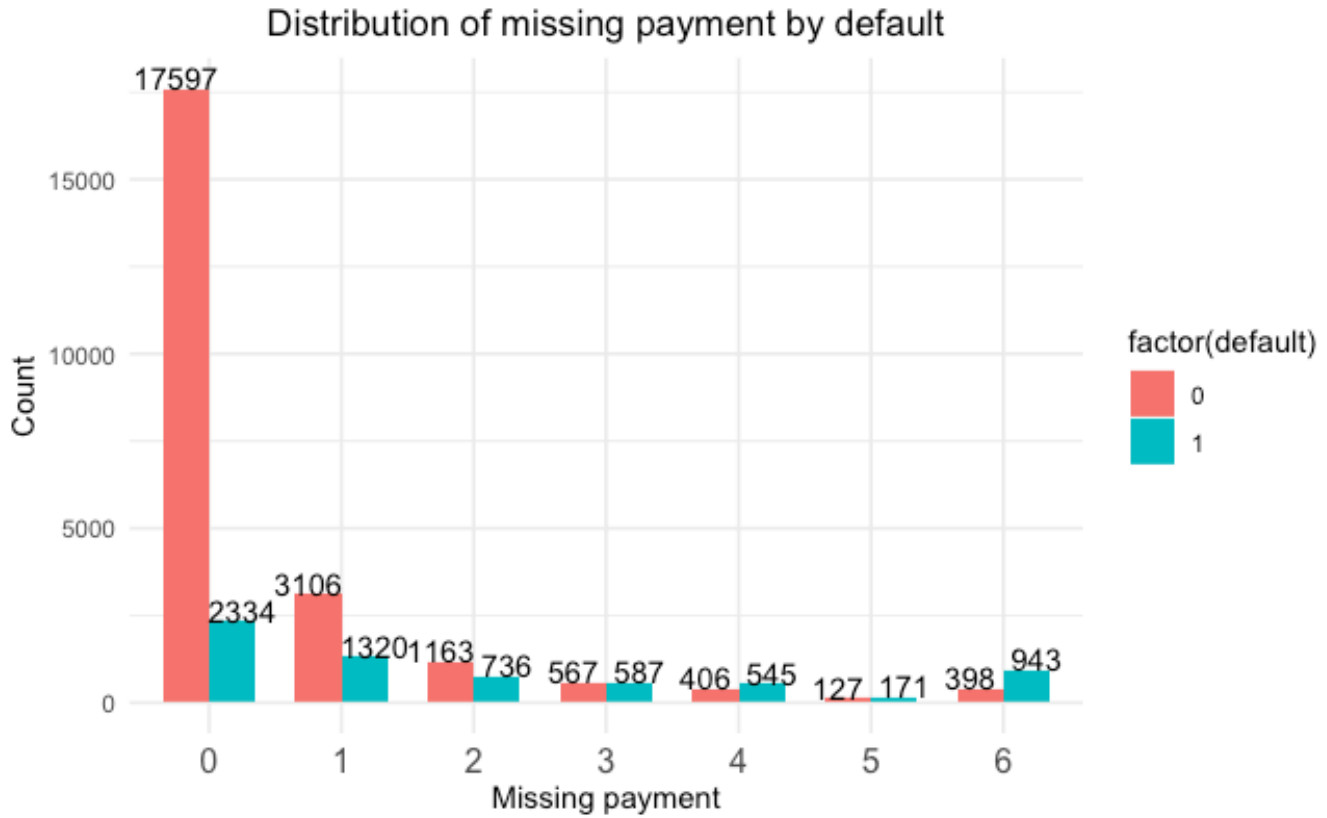







Distribution of education by default
Count
10000
5000
0
80.8%
19.2%
74.8%
25.2%
92.9%
7.1%
76.3%
23.7%
graduate school
high school
others
university
Education
factor(default)
0
1


Distribution of gender by default
Count
15000
10000
5000
0
79.2%
20.8%
75.8%
24.2%
female
male
Sex
factor(default)
0
1


Distribution of missing payment by default
Count
15000
10000
5000
0
17597
2334
3106
1320
1163
736
567
587
406
545
127
171
398
943
0
1
2
3
4
5
6
Missing payment
factor(default)
0
1

Data Preparation

No missing data/ No bad data points

Scale

Multicollinearity: Principle component analysis (PCA)

Model Selection

Random Forest

Naïve Bayes

XGBoost

1. Ensemble algorithm
2. Gradient boosting
3. Bagging vs boosting

Logistic Regression

KNN

Model Training

Tuning parameters: Random search

Stratified 5-fold cross-validation

Unbalance data: Oversampling/Under-sampling

Oversampling: randomly duplicates examples in the minority class (default)
Undersampling: randomly delete examples in the majority class (not default)

Model Performance

Algorithms	Accuracy	**Precision**	**Recall**	**F score**	**Prediction speed**	**Interpretability**
Random Forest	0.71	0.64	0.69	0.66	0.124	Low
XGBoost	0.75	0.66	0.69	0.67	0.032	Low
Naïve Bayes	0.76	0.66	0.68	0.67	1.614	High
Logistic Regression (under)	0.78	0.68	0.69	0.68	0.012	High
KNN	0.80	0.70	0.62	0.66	1.777	High

Conclusion & Limitation

Logistic regression using under-sampling is the best model

More data

THANK YOU