Credit Card Default Model

WENYING WU

Aim

To build a model that can predict whether a customer will default on credit card payment base on a given data set.

Data category given over 23101 customer:

ID, Credit Limit, Gender, Education Level, Marriage Status, Age,

Repayment Status over past 6 months,

Bill amount over past 6 months,

Payment amount over past 6 months,

Default payment next month

Approach Taken

Understand the data set – Exploratory data analysis

Prepare the data set for modelling – Make assumptions base on data understanding

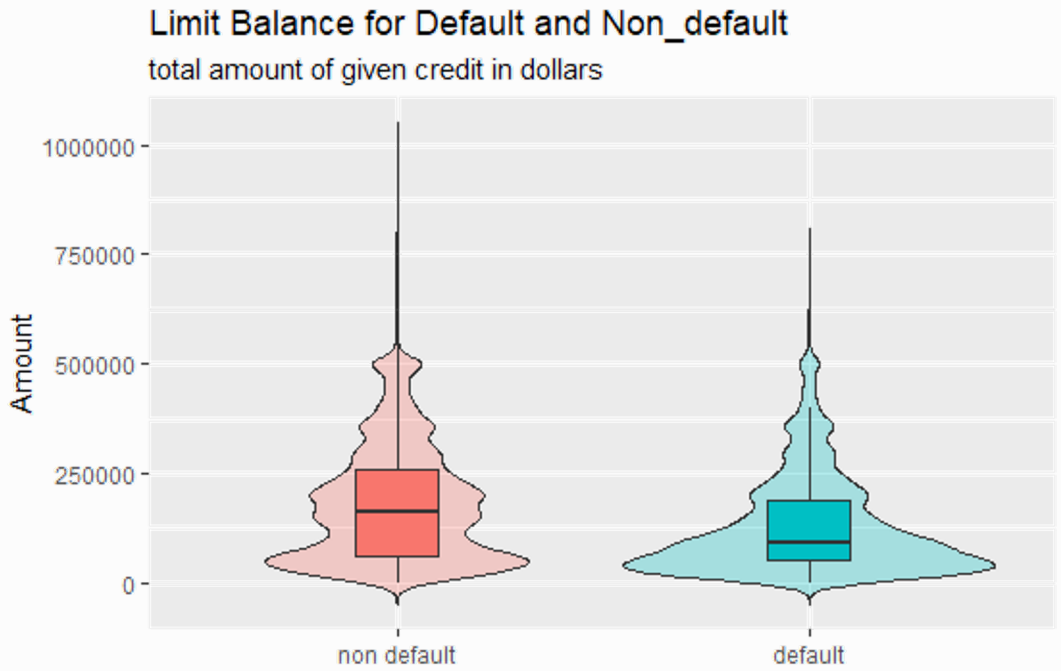
Set model evaluation metric – Recall in this case

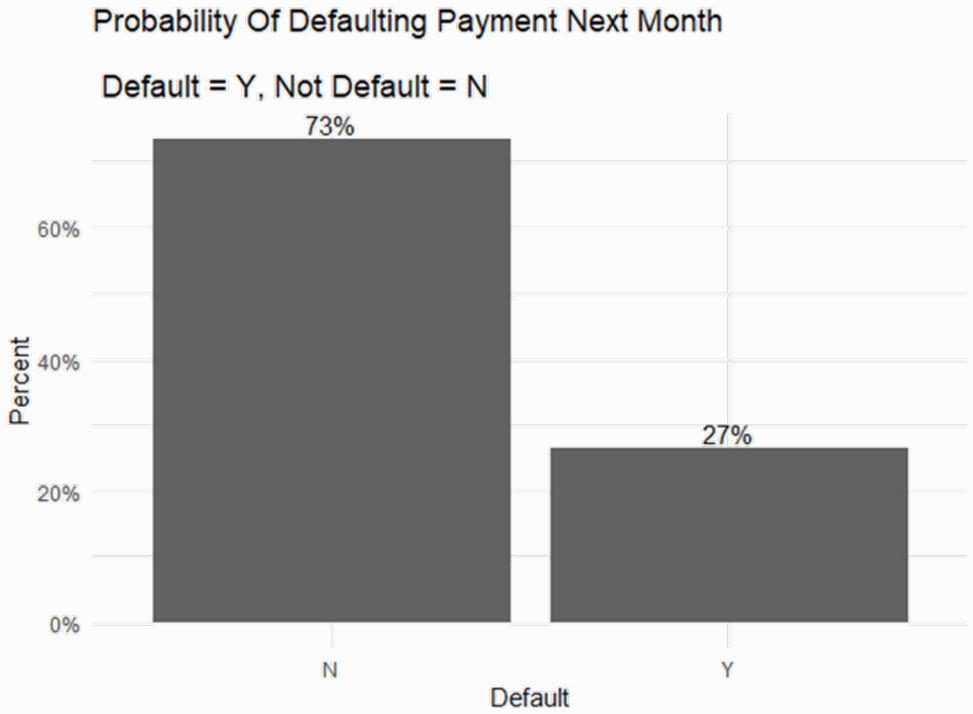
Train various model for selection – Linear logistic models & Tree Based models

Exploratory Data Analysis (EDA)

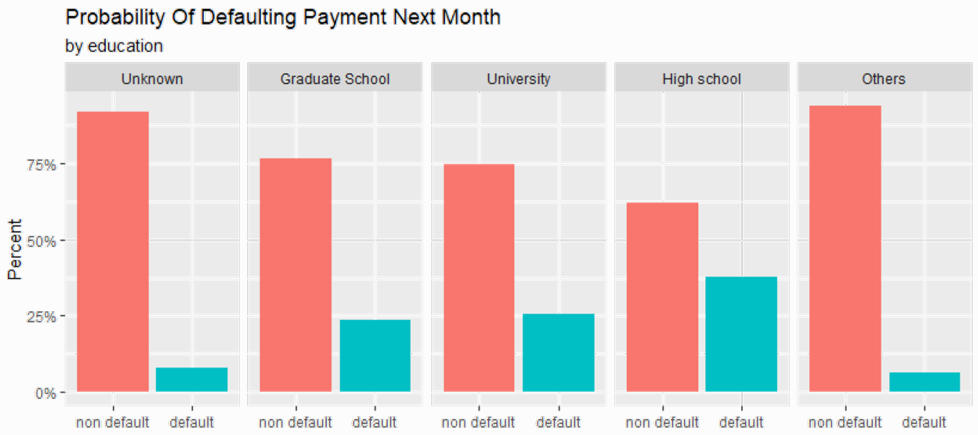
No.	EDA Finding
1	Default only made up 27% of given data set.
2	Female clients have lower chance to default in general.
3	Higher education level customer generally has lower chance to default.
4	Married and Other marriage status clients are more likely to default in general.
5	The age groups that have biggest chance to default are 55-64, more than 65 and under 25 years old.
6	Customer with lower credit limit are more likely to default.

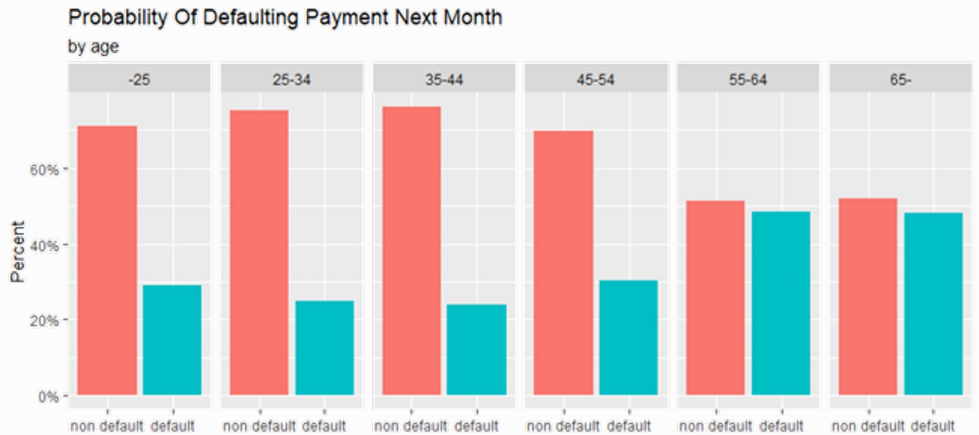
EDA - GRAPHS

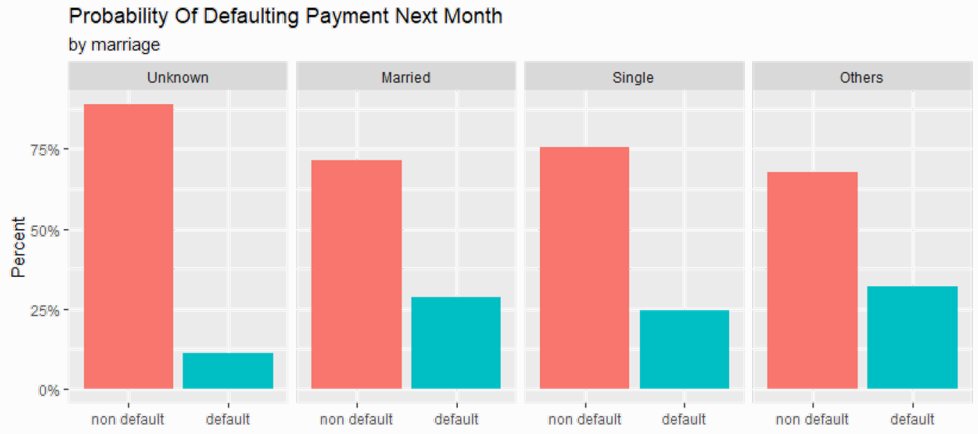

Limit Balance for Default and Non_default
total amount of given credit in dollars
Amount
1000000
750000
500000
250000
0
non default
default

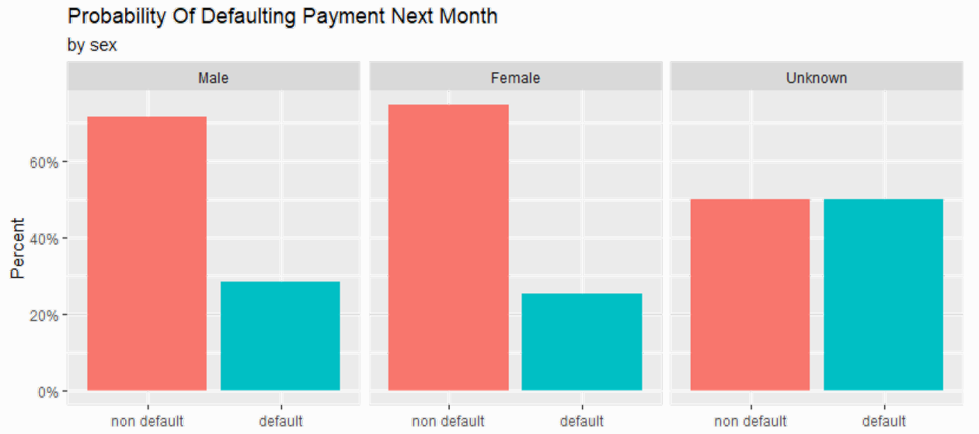

Probability Of Defaulting Payment Next Month
Default = Y, Not Default = N
Percent
73%
27%
60%
40%
20%
0%
N
Y
Default

EDA - GRAPHS


Probability Of Defaulting Payment Next Month
by education
Unknown
Graduate School
University
High school
Others
Percent
75%
50%
25%
0%
non default
default


Probability Of Defaulting Payment Next Month
by age
-25
25-34
35-44
45-54
55-64
65-
Percent
60%
40%
20%
0%
non default
default


Probability Of Defaulting Payment Next Month
by marriage
Unknown
Married
Single
Others
Percent
75%
50%
25%
0%
non default
default


Probability Of Defaulting Payment Next Month
by sex
Male
Female
Unknown
Percent
60%
40%
20%
0%
non default
default

Assumptions

All Payment Amount data are invalid

> These data are either 0 or same number across each month in given data set. And the numbers are much higher than bill amount even credit limit, which does not make sense.

All other data are valid

> Except for obvious input values like ‘RYDE’ in Sex

Model Evaluation Metric - Recall

Recall formula:

$$Recall = TP/(TP+FN)$$

In this case, TP means defaulter predicted to be defaulter, FN means defaulter predicted to be non-defaulter.

Reason for choosing recall:

Bank will not want any defaults not being identified to avoid any monetary loss. Therefore, we want the lowest FN and the highest Recall.

Result Obtained

Model Name	Recall (0 - 1)
Linear Logistic Model	0.2533
GBM	0.4260
Random Forest	0.4488
XGBoost	0.5046

XGBoost Model is recommended

Potential Ethical Issues

Discrimination Concern – Age, Gender and Education and potentially more

Most model's predicting power improve with access to more data and data type. Discrimination concern might extend to race, ethnicity and disability.

Misclassification Result

1. Defaulter classified as non-defaulter – Loss of avenue.

2. Non-defaulter classified as defaulter – Unfair to misclassified customer.

Model Interpretation

If a sophisticated model is chosen, the interpretation process might be a frustration to both financial advisor and customer, might even lead to loss of customer.

Data Leaking

The potential of data leaking exists in data collection, process and storage stages