Case study

Credit Default

Vamsi Grandhi

Contents

[**Business Objective** 2](#_Toc476219433)

[**Data** 2](#_Toc476219434)

[**Exploratory analysis:** 2](#_Toc476219435)

[**Data Cleansing (part a)** 2](#_Toc476219436)

[**Data Imputation for categorical & nominal variables** 2](#_Toc476219437)

[**Relationship between variables** 3](#_Toc476219438)

[**Relationship between gender & education** 4](#_Toc476219439)

[**Relationship between gender & default payment status** 4](#_Toc476219440)

[**Relationship between gender, education & default payment status** 5](#_Toc476219441)

[**Relationship between sex, marital status & Balance limit** 5](#_Toc476219442)

[**Relationship between sex, marital status & default payment status** 5](#_Toc476219443)

[**Relationship between sex, education & Balance limit** 6](#_Toc476219444)

[**Feature Engineering** 6](#_Toc476219445)

[**Working state** 7](#_Toc476219446)

[**Relationship between working state, education & Balance limit** 7](#_Toc476219447)

[**Average Repayment** 7](#_Toc476219448)

[**Average Bill amount** 7](#_Toc476219449)

[**Bill amount to balance limit ratio** 7](#_Toc476219450)

[**Amount owed** 7](#_Toc476219451)

[**Repayment ratio** 7](#_Toc476219452)

[**Age bin** 7](#_Toc476219453)

[**No of missing payments** 8](#_Toc476219454)

[**Relationship between age bin, education & balance limit** 8](#_Toc476219455)

[**Data cleansing part b** 8](#_Toc476219456)

[**Outlier detection with bill amount to balance limit ratio** 8](#_Toc476219457)

[**Outlier detection with age group, repayment to bill amount ratio, working state & education** 9](#_Toc476219458)

[**Modelling** 9](#_Toc476219459)

[**Data Cleansing** 9](#_Toc476219460)

[**Data partitioning** 9](#_Toc476219461)

[**Data Balancing** 9](#_Toc476219462)

[**Data Modelling** 10](#_Toc476219463)

# **Business Objective**

1. Exploratory analysis on credit default data set
2. Model to identify the default status for next month based on historical data

## **Data**

The given data for **30,000** individuals has their transaction history for 6 months (April to September), demographics, education, and payment history. The given data has following columns.

X1: Amount of the given credit (NT dollar)

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year)

X6 - X11: History of past payment.

X12-X17: Amount of bill statement (NT dollar)

X18-X23: Amount of previous payment (NT dollar)

### **Exploratory analysis:**

### **Data Cleansing (part a)**

All the variables in the given data are continuous and they are converted into categorical, nominal & continuous variables based on basic knowledge.

The given data has been segregated as **14** continuous variables and **9** categorical variables

|  |  |
| --- | --- |
| **Continuous Variables** | **Categorical & Nominal Variables** |
| 1. X1: Amount of the given credit card  2. X5: age  3. X6-X11: History of payment  4. X12-X17: Amount of bill statement  5. X18-X23: Amount of previous payment | 1. X2: Gender  2. X3: Education  3. X4: Marital Status |

**Missing data**: There is **no** missing data in either of the continuous or categorical variables.

**Data Imputation**: The following procedures has been adopted to handle unmentioned levels of categorical variables

**Credit default proportion:** only **22%** of given records are default and other are non-defaulters indicating level of imbalance.

### **Data Imputation for categorical & nominal variables**

**Gender**: No imputation has been done on gender variable and there are **more females credit card holders** compared to males.

|  |  |
| --- | --- |
| Gender | No of credit card holders |
| female | 18112 |
| male | 11888 |

**Education:** The data has more levels of education such as 0,4,5,6 apart from mentioned ones i.e; high school, graduate school, university & others. The levels **0,4,5,6** along with category **“other”** are mentioned as **“unknown”**. There are more number **of university graduated** credit card holders compared to others

**Data before imputation:**

|  |  |
| --- | --- |
| Education level | No of credit card holders |
| University | 14030 |
| Graduate school | 10585 |
| High school | 4917 |
| 5 | 280 |
| others | 123 |
| 6 | 51 |
| 0 | 14 |

**Data after Imputation:**

|  |  |
| --- | --- |
| Education level | No of credit card holders |
| University | 14030 |
| Graduate school | 10585 |
| High school | 4917 |
| Unknown | 468 |

**Marital Status:** The data has an extra level of marital status “0” apart from mentioned ones i.e; single, married & others. The levels **0,** along with category **“other”** are mentioned as **“others”**. There are more **single** credit card holders compared to married & others

**Data before imputation:**

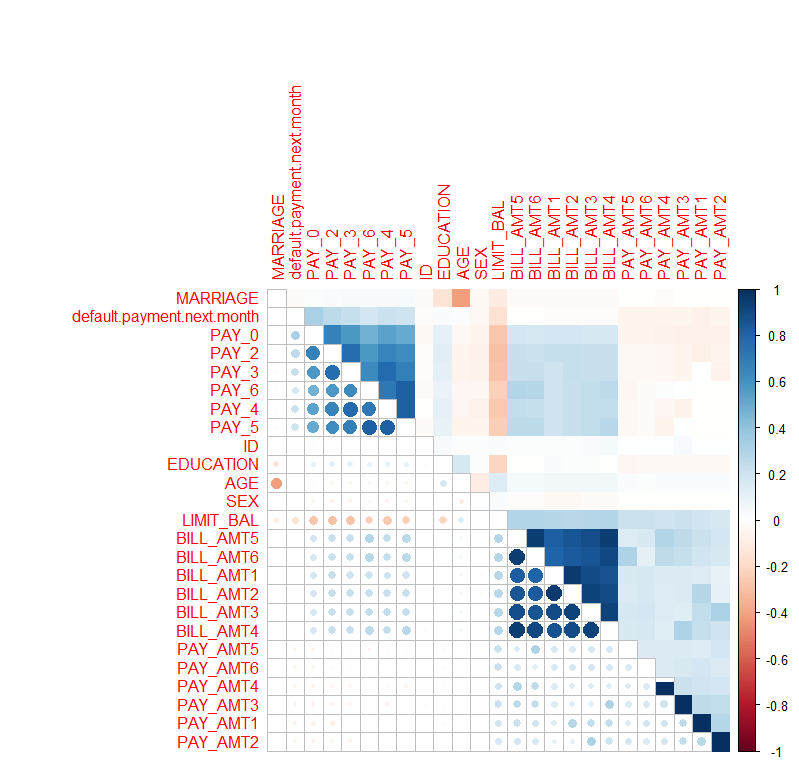
|  |  |
| --- | --- |
| Marital Status | No of Credit card holders |
| single | 15964 |
| married | 13659 |
| others | 323 |
| 0 | 54 |

**Data after imputation:**

|  |  |
| --- | --- |
| Marital Status | No of Credit card holders |
| single | 15964 |
| married | 13659 |
| others | 377 |

## **Relationship between variables**

The relationship between variables is described in the following section. A mixed correlation plot between variables is given below. It is quite evident from the plot that payment made by individual is **very likely** correlated with billing amount i.e; **an individual pay very less than the billing amount.** The correlation plot is given below.



### **Relationship between gender & education**

Female credit card holders are highly educated compared to male

|  |  |  |
| --- | --- | --- |
| Education | Female | Male |
| University | 8656 | 5374 |
| Graduate School | 6231 | 4354 |
| High School | 2927 | 1990 |
| Unkown | 298 | 170 |
| Total | 18112 | 11888 |

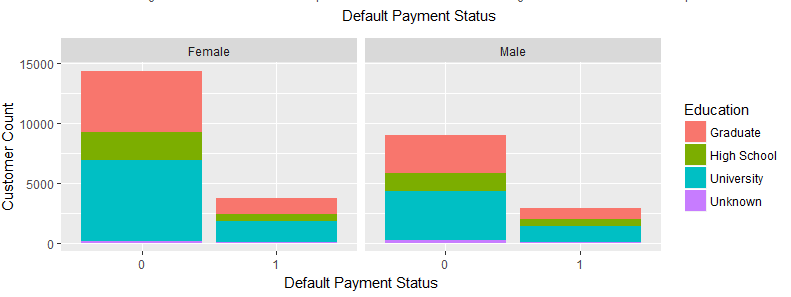
### **Relationship between gender & default payment status**

**24%** of males are credit card defaulters against their counterpart’s i.e: females who constitute for **20%**

|  |  |  |  |
| --- | --- | --- | --- |
| Gender | Credit Card default status (0) | Credit Card default status (1) | Proportion of credit card defaulters |
| female | 14349 | 3763 | 20% |
| male | 9015 | 2873 | **24%** |

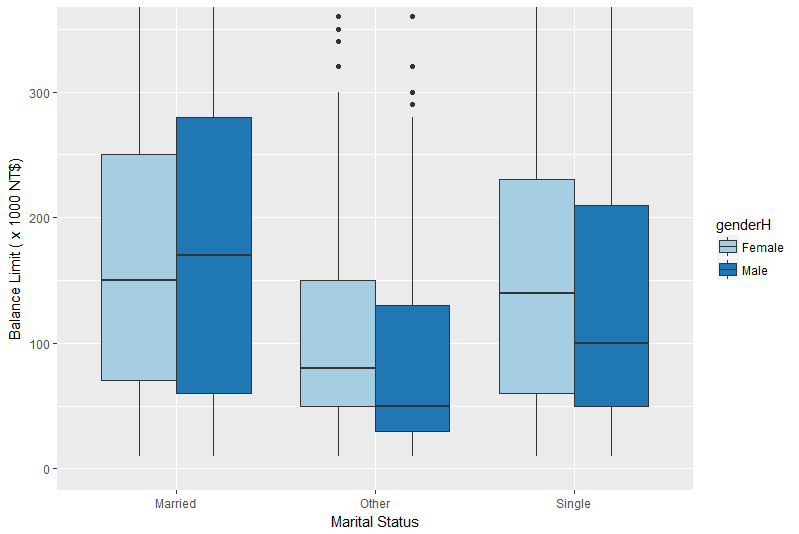
### **Relationship between gender, education & default payment status**

There are number of male & female **university graduate** credit card defaulters compared to other educational backgrounds



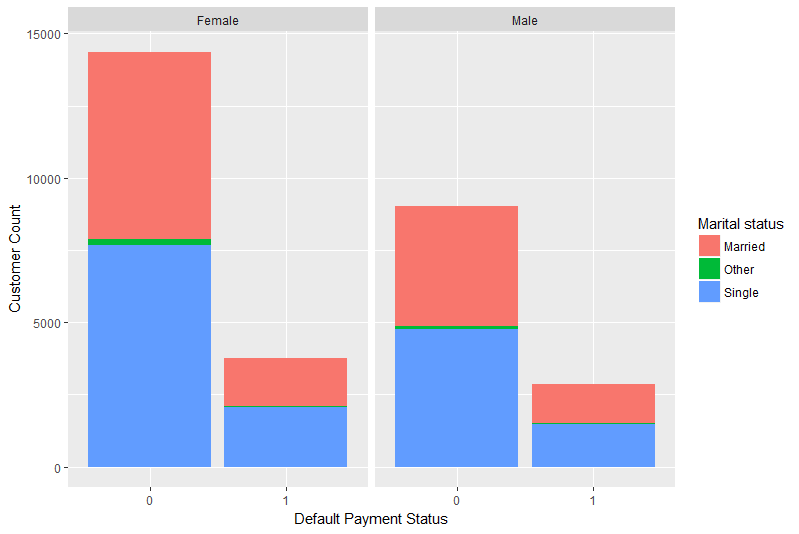
### **Relationship between sex, marital status & Balance limit**

Married males & females have higher balance limit compared to single and other males & females and their variance is also high compared to other groups. Below is the figure which illustrates this phenomenon.



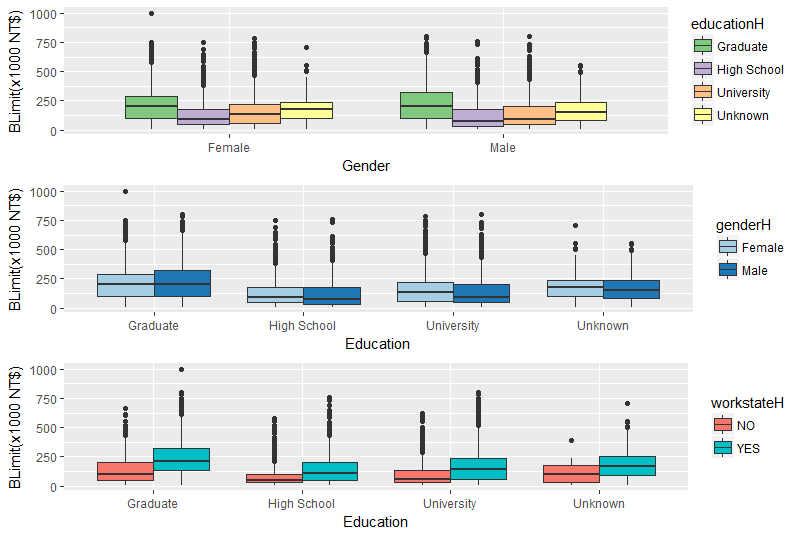
### **Relationship between sex, marital status & default payment status**

Single females & males are tend to be more default compared to married males & females. Below is the figure which illustrates this phenomenon.



### **Relationship between sex, education & Balance limit**

Both male &female **graduates** followed by **university** male & female students have higher balance limit compared to high school & unknown education level. The below given images explain the phenomenon



## **Feature Engineering**

Some variables have been generated out of the data to understand more about the patterns of default & for building the model in predicting credit default. The feature engineered variables are as follows:

1. Working state
2. Average\_repayment
3. Average\_bill\_amount
4. Bill amount to balance limit ratio
5. Amount owed
6. Repayment ratio
7. Age\_bin
8. No of missing payments

### **Working state**

An individual is considered to be working if sum of all the **payment history status is <=0** i.e; he/she paid either due payment or full payment. The calculation is described as follows:

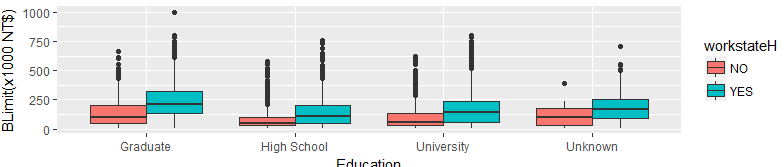
Working status - yes if sum (pay\_0 + pay\_2 + pay\_3 + pay\_4 + pay\_5 + pay\_6) < 0

Working status - NO if sum (pay\_0 + pay\_2 + pay\_3 + pay\_4 + pay\_5 + pay\_6) > 0

### **Relationship between working state, education & Balance limit**

The following inferences can be made from the below figure:

1. Credit cards have been issued to working people
2. Balance limit of working graduates, high school pass outs compared to non-working individuals.



### **Average Repayment**

It’s an average repayment made by an individual for the 6 months (April – September) in the given data.

Average repayment = Average (Previous payments (X18-X23))

### **Average Bill amount**

It’s an average bill amount made by an individual for the 6 months (April – September) in the given data.

Average repayment = Average (Bill statement (X12-X17))

### **Bill amount to balance limit ratio**

The bill amount to balance limit ratio for an individual is calculate as

Bill amount to balance limit ratio = Average bill amount / Balance limit

### **Amount owed**

The remaining amount (billing amount – paid amount) that is owed by an individual to banks. It is calculated as **amount\_owed = sum(bill statement(X12-X17)) – sum(previous payments(X18-X23))**

### **Repayment ratio**

It is calculated as ratio of average repayment made to average billing amount for an individual. It is calculated as **repayment ratio = average (previous payments(X18-X23)) / average (bill statement(X12-X17))**

### **Age bin**

A nominal variable has been created by binning continuous age variable ranging from **0-80** into eight buckets.

|  |  |
| --- | --- |
| Age Group | Bin number |
| 0-10 | 1 |
| 11-20 | 2 |
| 21-30 | 3 |
| 31-40 | 4 |
| 41-50 | 5 |
| 51-60 | 6 |
| 61-70 | 7 |
| 71-80 | 8 |

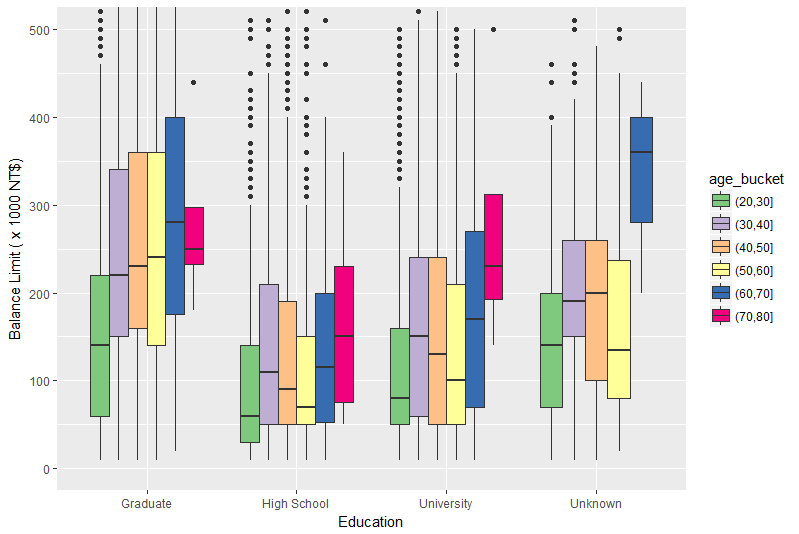
### **No of missing payments**

If payment status of any individual for the six months (april – September) is greater than equal to 1 i.e; payment delayed by one month is considered as missing payment

**No of missing payments = countif ( pay\_0:pay\_6,”>0”)**

### **Relationship between age bin, education & balance limit**

It is quite evident that university graduate of all age groups have higher balance limit compared to other educational backgrounds.



## **Data cleansing part b**

### **Outlier detection with bill amount to balance limit ratio**

Some unusual billing amount patterns have been detected for individuals with bill amount to balance limit ratio **greater than 1.25.**

Their monthly bills increased suddenly **twice or three** from the previous month and far greater than balance limit and their repayment to bill ratio is of **average 0.08** and there are in total of **79 candidates**.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Bill\_amt1 | Bill\_amt2 | Bill\_amt3 | Bill\_amt4 | Bill\_amt5 | Bill\_amt6 | Repayment/bill ratio | Bill amount\_balance limit ratio |
| 673 | 99568 | 32326 | 31840 | 37075 | 37662 | 36904 | 0.276 | 1.52 |
| 921 | 471814 | 478380 | 395612 | 386295 | 356206 | 352257 | 0.033 | 1.69 |
| 971 | 42784 | 41009 | 44267 | 47149 | 48497 | 14774 | 0.104 | 1.98 |
| 1678 | 90231 | 90647 | 92309 | 93880 | 99418 | 101392 | 0.042 | 1.89 |

Out of **79 candidates 57 are non-defaulters** which is illogical and they are removed from the data considering them as outliers.

Out of **57 individuals 20 are non-working** and they have an average billing amount of **10, 4987 USD** and have an average repayment ratio of **0.08**

### **Outlier detection with age group, repayment to bill amount ratio, working state & education**

**62** Individuals within age group of 20-30 & non –working high school graduates with an average repayment ratio less than 0.04 are **not considered as defaulters**. With intuition these records have been removed from data considering them as outliers.

## **Modelling**

As mentioned earlier that the data being highly unbalanced the data has been balanced using **smote technique** ,**under & over sampling techniques** and the same data has been used for model building **using naïve bayes model, & XG Boost.**

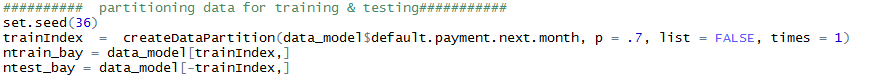
### **Data Cleansing**

The records which have Inf, NaN/ NA repayment ratio & bill amount to bill limit ratio have been removed prior to model building.



### **Data partitioning**

70% of data has been used for training and 30% for model testing for naïve bayes & Xg boost algorithms

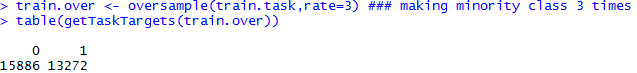


### **Data Balancing**

The training data has been balanced using the following techniques

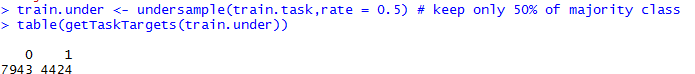
**Oversampling technique:**

This technique is used to balance data by replicating the minority class by **3 times**



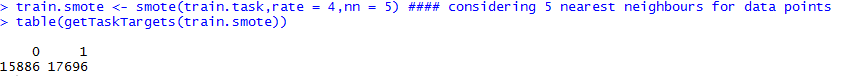
**Under sampling technique:**

This technique is used to balance data by randomly selecting **50% of majority class.**



**Smote technique:**

This technique is used to balance data by replicating the minority class by 4 times and considering 5 nearest neighboring data points.



### **Data Modelling**

The unbalanced, over sampled, under sampled & smote data have been used for building XG boost & navie bayes model.

**Steps in building the model**

The unbalanced, over sampled, under sampled & smote training data has been used to build the above mentioned models using **5 fold cross validation technique** and found that **unbalanced data** has **higher accuracy and sensitivity**. **Sensitivity** is considered as performance metric as it’s important to predict the exact number of defaulters with high accuracy.

**Testing the models:**

The models built through unbalanced data has been used to test the data and the results are given in the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | naïve bayes | | XG Boost | |
| Metrics | Accuracy % | Sensitivity % | Accuracy % | Sensitivity % |
| Unbalanced  (train.task) | **77.55** | **48.65** | **82.43** | **68.19** |
| Oversample (train.over) | 65.75 | 35.44 | 79.17 | 55.18 |
| under sample (train.under) | 60.41 | 32.18 | 76.99 | 47.4 |
| Smote (train.smote) | 42.36 | 25.84 | 80.74 | 58.31 |

From the above table it’s quite evident that XG boost is giving better results on unbalanced dataset compared to navie bayes algorithm.