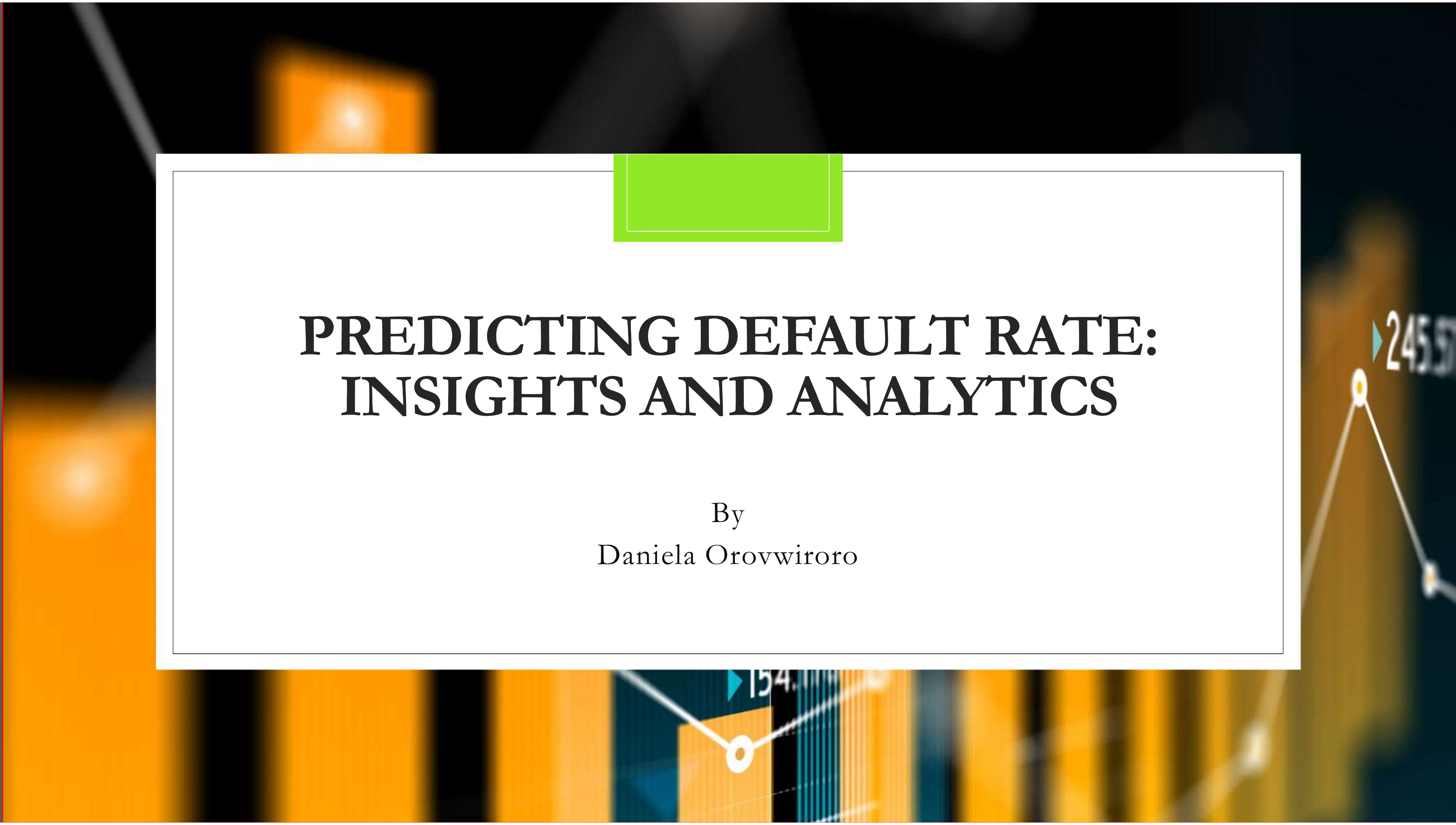


PREDICTING DEFAULT RATE: INSIGHTS AND ANALYTICS

By

Daniela Orovwiroro

Content

1 Introduction

2 Data Cleansing

3 Exploratory Data Analysis

4 Feature Engineering

5 Data Modeling

6 Business Insights

01 Introduction

Key Problems And Issues

- Some consumers issued credit by the company have a history of payment defaults.
- Data on customers and their default status have been collected and this requires data processing
- There is risk exposure of some customers using their credit line beyond their repayment capabilities which would translate to high debt accumulation.
- Hence, there is need to identify and predict risky and non-risky customers and identify the potential of the customers to repay the debt.

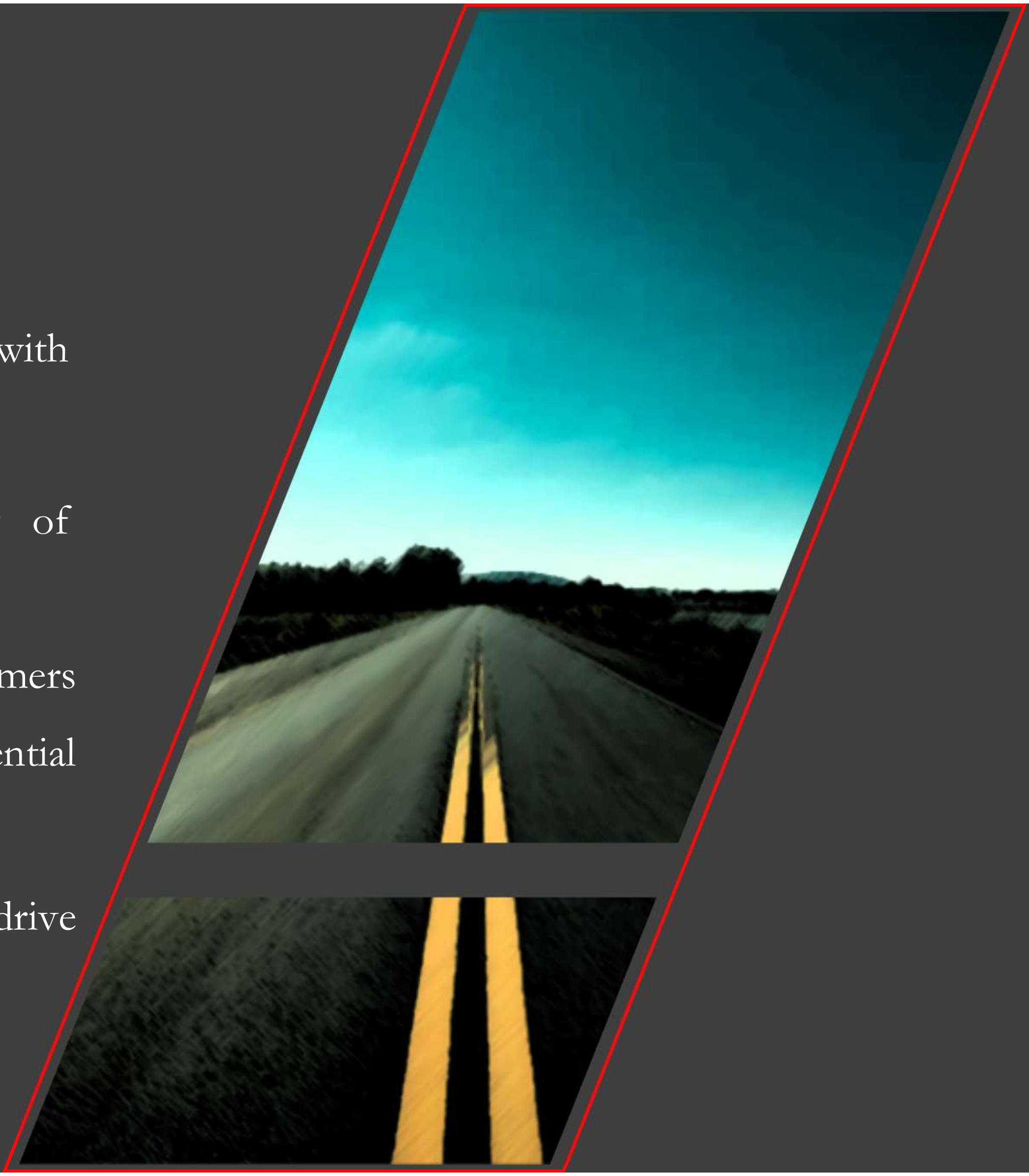
The Goal

1. Summarize key drivers and their relationship with default rate (Y).

2. Build models that predict the probability of customers default

3. Identify attributes of potential non-risky customers who have a high probability of settling their potential credit liability

4. Derive insights from the data that would drive business growth



02 Data Cleansing

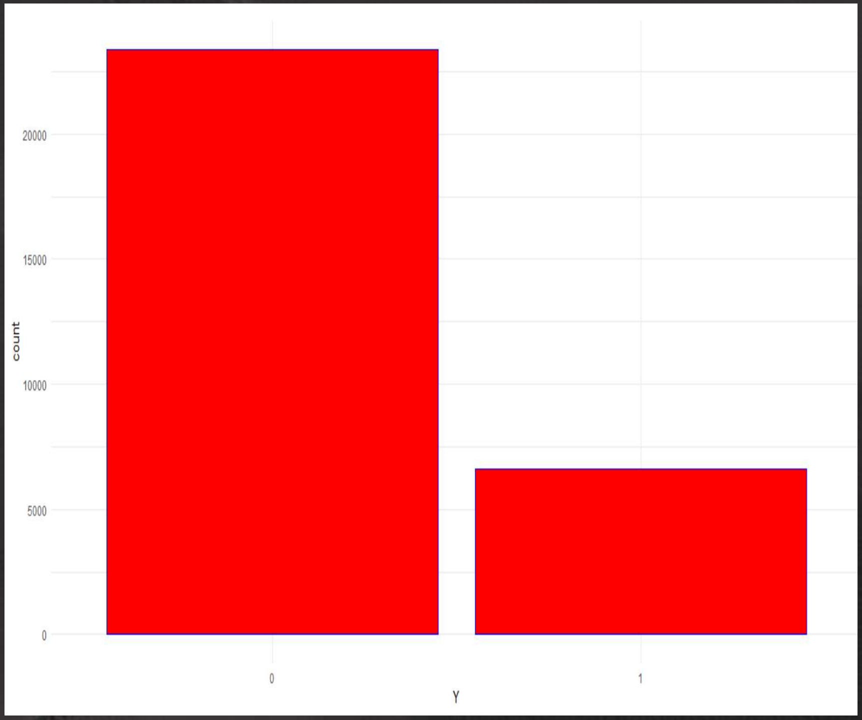
Overview of the Dataset

Independent variables

Credit limit
• Gender
• Age
• Marital status
• Level of education
• History of their past repayments made (April to September) (RepayS_5 to RepayS_0)
• Amount of bill statement (BillS_5 to BillS_0)
• Amount of previous payment (PrePay_5 to PrePay_5)

Dependent variable

• Default - (Yes = 1, No = 0)



Data cleansing is important because it involves the preparing of data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This improves the data quality and translates to overall productivity.

Load the csv file

R software

Data cleansing

1. The dataset has 24 variables and 30000 rows

```
# Examine the imported credit dataset.
dim(credit)
```

```
## [1] 30000    24
```

2. Renamed the headers of the dataset to get more insights on the dataset.

```
# Rename the headers of the dataset
credit= rename(credit, Credit_Amount = X1, Gender = X2, Education = X3, Marital_status = X4,
               Age = X5,RepayS_0 = X6,RepayS_1 = X7, RepayS_2 = X8, RepayS_3 = X9,
               RepayS_4 = X10, RepayS_5 = X11, BillS_0 = X12, BillS_1 = X13,
               BillS_2 = X14, BillS_3 = X15,
               BillS_4 = X16, BillS_5 = X17,
               PrePay_0 = X18, PrePay_1 = X19,
              PrePay_2 = X20, PrePay_3 = X21,
                PrePay_4 = X22, PrePay_5 = X23,
               Default= Y)
# Review the first 6 rows of credit dataset.
head(credit)
```

```
##    Credit_Amount Gender Education Marital_status Age RepayS_0 RepayS_1
## 1          20000      2         2              1  24        2        2
## 2         120000      2         2              2  26       -1        2
## 3          90000      2         2              2  34        0        0
## 4          50000      2         2              1  37        0        0
## 5          50000      1         2              1  57       -1        0
## 6          50000      1         1              2  37        0        0
##    RepayS_2 RepayS_3 RepayS_4 RepayS_5 BillS_0 BillS_1 BillS_2 BillS_3
## 1        -1       -1       -2       -2    3913    3102     689       0
## 2         0        0        0        2    2682    1725    2682    3272
## 3         0        0        0        0   29239   14027   13559   14331
## 4         0        0        0        0   46990   48233   49291   28314
## 5        -1        0        0        0    8617    5670   35835   20940
## 6         0        0        0        0   64400   57069   57608   19394
```

3. Checked for missing values

```
# We check if the dataset has any missing values by checking rows of data
credit[!complete.cases(credit),]
```

```
##  [1] Credit_Amount  Gender         Education      Marital_status
##  [5] Age            RepayS_0       RepayS_1       RepayS_2
##  [9] RepayS_3       RepayS_4       RepayS_5       BillS_0
## [13] BillS_1        BillS_2        BillS_3        BillS_4
## [17] BillS_5        PrePay_0       PrePay_1       PrePay_2
## [21] PrePay_3       PrePay_4       PrePay_5       Default
## <0 rows> (or 0-length row.names)
```

4. Checked and removed duplicates

```
#Let's check if there are duplicated data
dup_rows <- duplicated(credit)
dup_rows_num <- sum(dup_rows)
dup_rows_num
```

```
## [1] 35
```

```
#here we remove the duplicate datapoints
credit = credit %>% distinct()
```

03

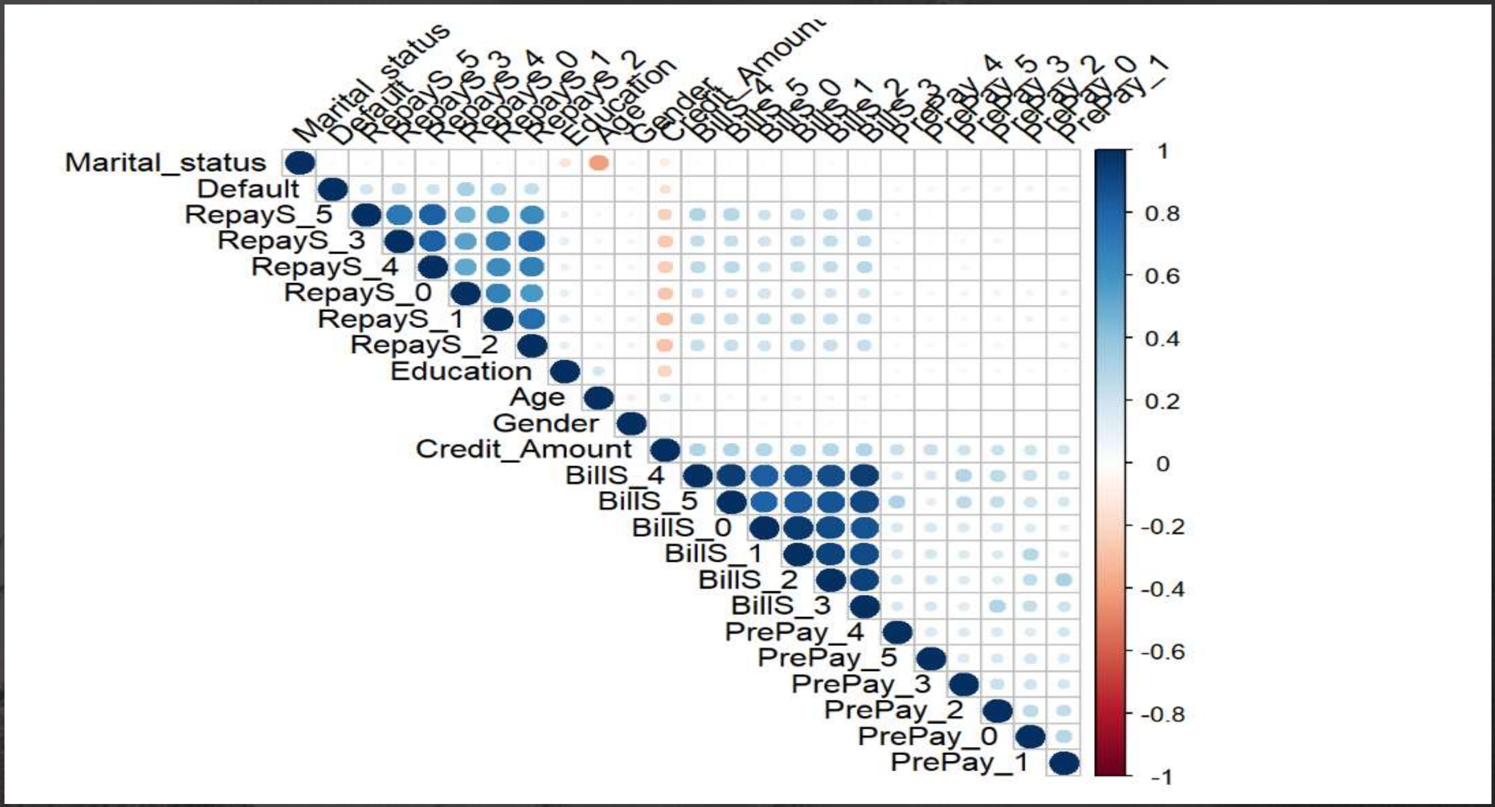
Exploratory Data Analysis

Quantitative description of the variables in the dataset

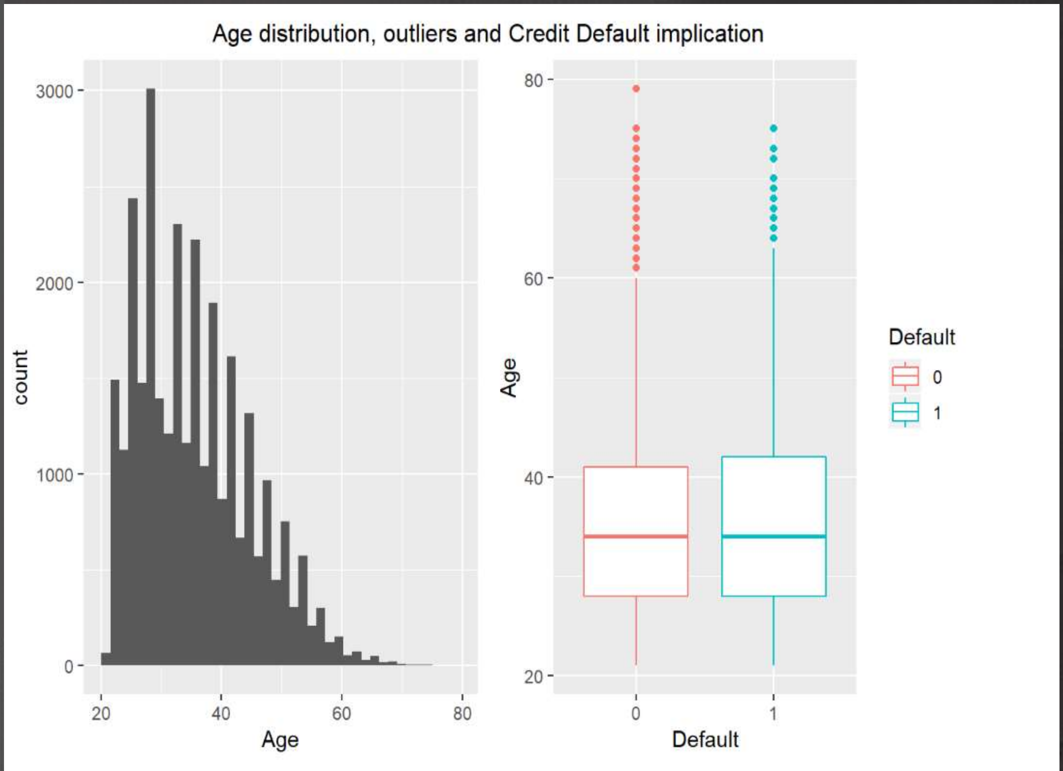
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Credit_Amount	1	29965	167442.01	129760.14	140000	151551.23	133434.00	10000	1000000	990000	0.99	0.54	749.61
Gender	2	29965	1.60	0.49	2	1.63	0.00	1	2	1	-0.42	-1.82	0.00
Education	3	29965	1.85	0.79	2	1.78	1.48	0	6	6	0.97	2.08	0.00
Marital_status	4	29965	1.55	0.52	2	1.55	0.00	0	3	3	-0.02	-1.36	0.00
Age	5	29965	35.49	9.22	34	34.69	8.90	21	79	58	0.73	0.04	0.05
RepayS_0	6	29965	-0.02	1.12	0	-0.06	1.48	-2	8	10	0.73	2.73	0.01
RepayS_1	7	29965	-0.13	1.20	0	-0.20	0.00	-2	8	10	0.79	1.58	0.01
RepayS_2	8	29965	-0.16	1.20	0	-0.23	0.00	-2	8	10	0.84	2.09	0.01
RepayS_3	9	29965	-0.22	1.17	0	-0.30	0.00	-2	8	10	1.00	3.51	0.01
RepayS_4	10	29965	-0.26	1.13	0	-0.36	0.00	-2	8	10	1.01	4.00	0.01
RepayS_5	11	29965	-0.29	1.15	0	-0.39	0.00	-2	8	10	0.95	3.44	0.01
BillS_0	12	29965	51283.01	73658.13	22438	35422.71	32382.95	-165580	964511	1130091	2.66	9.79	425.51
BillS_1	13	29965	49236.37	71195.57	21295	33896.87	30993.75	-69777	983931	1053708	2.70	10.29	411.29
BillS_2	14	29965	47067.92	69371.35	20135	32122.39	29273.94	-157264	1664089	1821353	3.09	19.77	400.75
BillS_3	15	29965	43313.33	64353.51	19081	29265.52	27696.45	-170000	891586	1061586	2.82	11.30	371.76
BillS_4	16	29965	40358.33	60817.13	18130	26970.42	26262.78	-81334	927171	1008505	2.87	12.29	351.33
BillS_5	17	29965	38917.01	59574.15	17124	25773.58	24919.54	-339603	961664	1301267	2.84	12.26	344.15
PrePay_0	18	29965	5670.10	16571.85	2102	3002.19	2859.94	0	873552	873552	14.66	414.76	95.73
PrePay_1	19	29965	5927.98	23053.46	2010	2881.28	2950.37	0	1684259	1684259	30.44	1639.54	133.18
PrePay_2	20	29965	5231.69	17616.36	1804	2473.24	2662.75	0	896040	896040	17.21	563.61	101.77
PrePay_3	21	29965	4831.62	15674.46	1500	2203.19	2223.90	0	621000	621000	12.90	276.98	90.55
PrePay_4	22	29965	4804.90	15286.37	1500	2206.14	2223.90	0	426529	426529	11.12	179.83	88.31
PrePay_5	23	29965	5221.50	17786.98	1500	2169.25	2223.90	0	528666	528666	10.63	166.94	102.75
Default	24	29965	0.22	0.42	0	0.15	0.00	0	1	1	1.34	-0.20	0.00

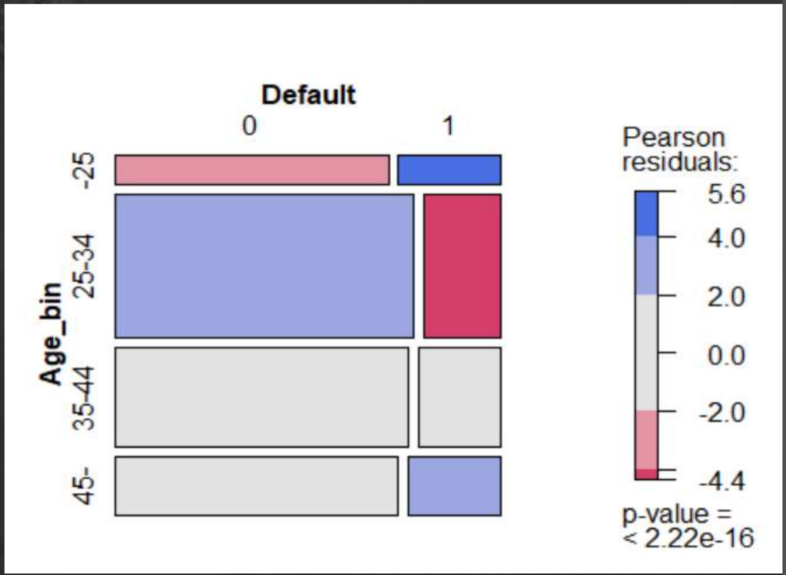
Descriptive statistics

Correlation plot between the predictors and the target variable



Impact of Independent variables on the dependent variable

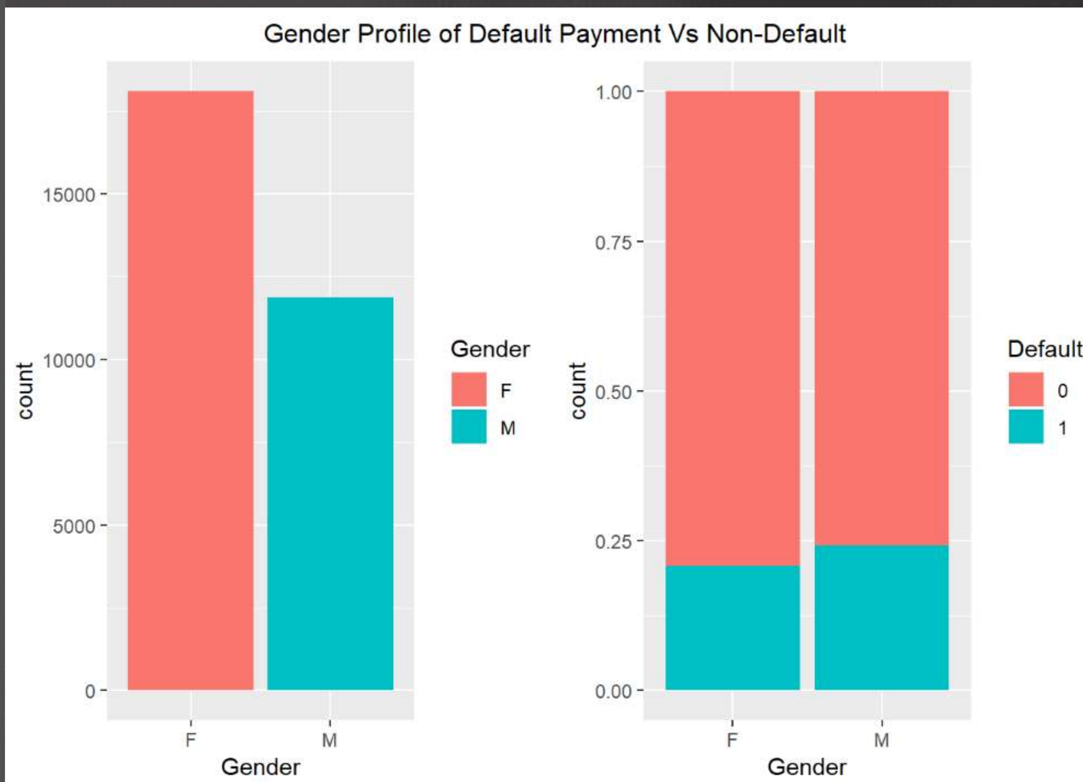


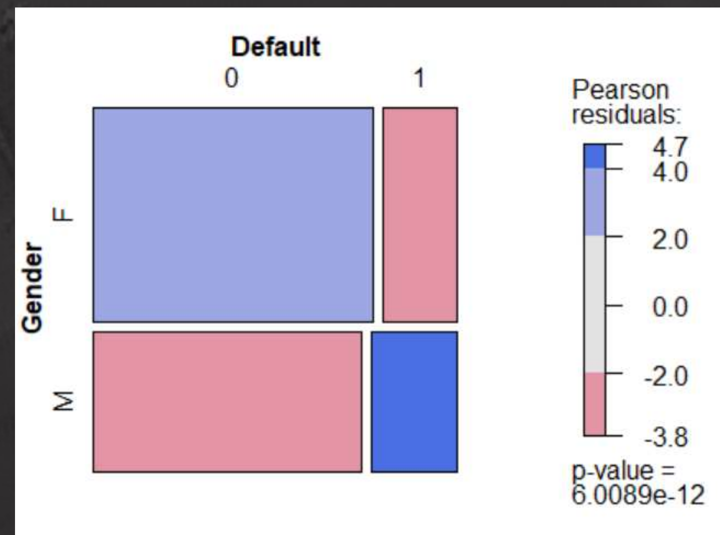


Age Distribution

Customers under age 25 have the highest possibility of defaulting and age group 25 – 34 have the smallest possibility of defaulting.

Impact of Independent variables on the dependent variable

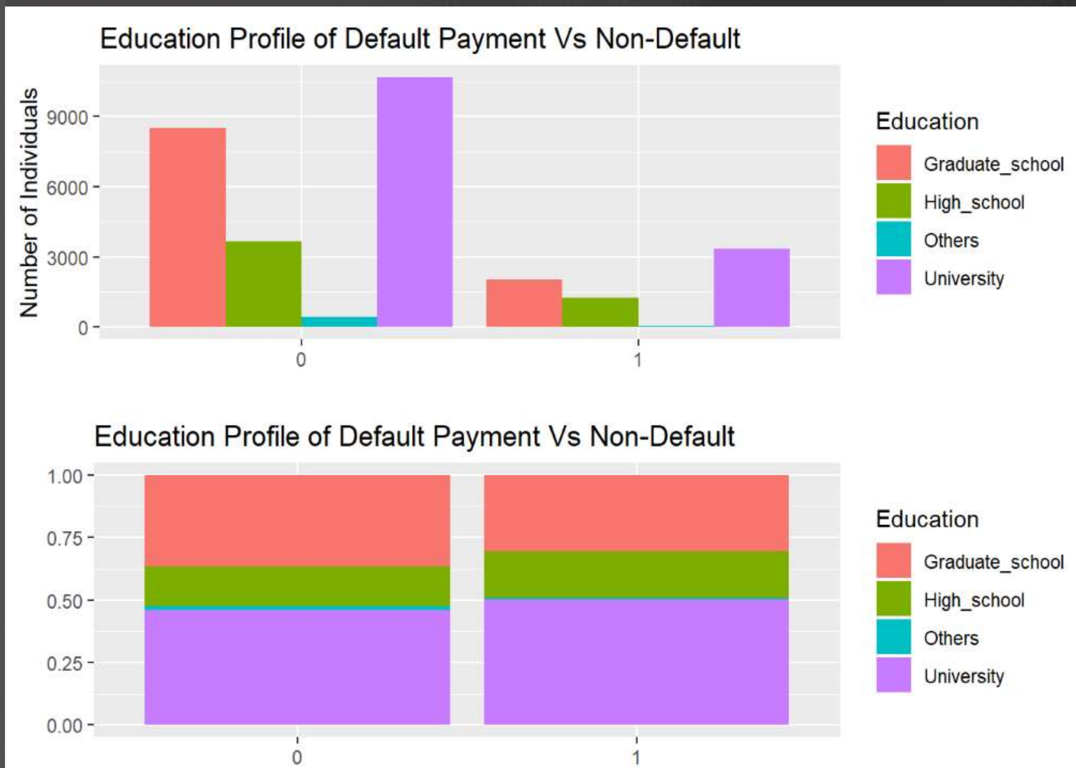


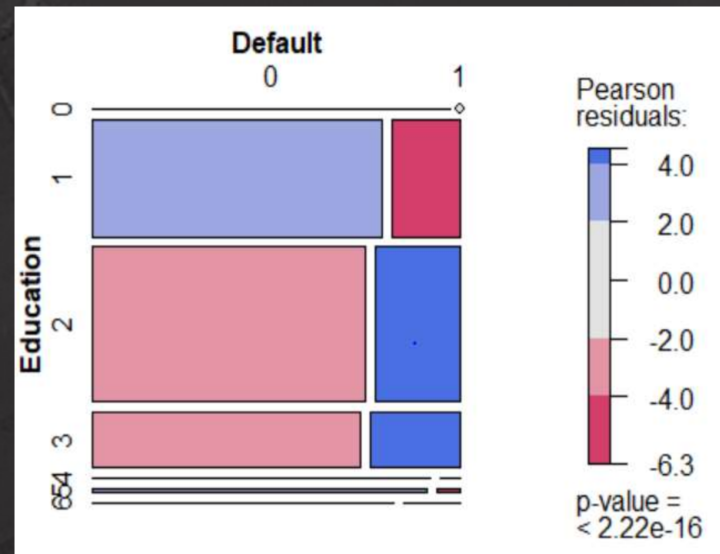


Gender Distribution

Male customers have a higher probability of defaulting compared to the female customers.

Impact of Independent variables on the dependent variable



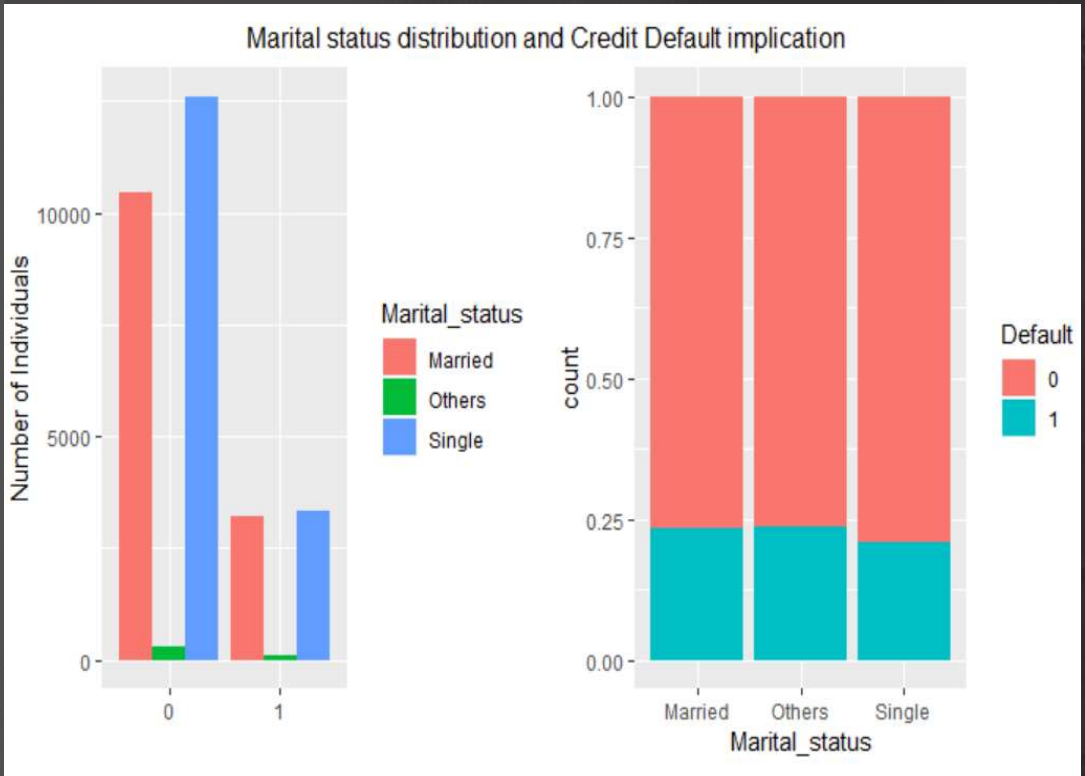


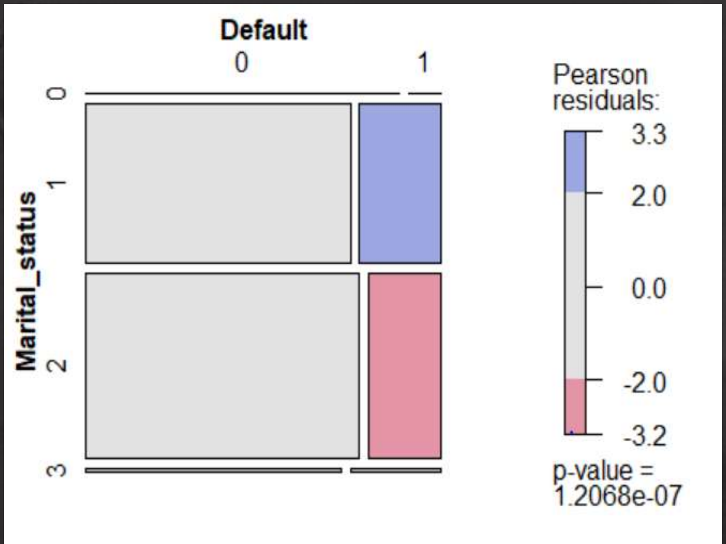
Education Distribution

There is a correlation between education and default.

Customers with higher degrees have a lower probability of defaulting compared to customers with other degrees.

Impact of Independent variables on the dependent variable

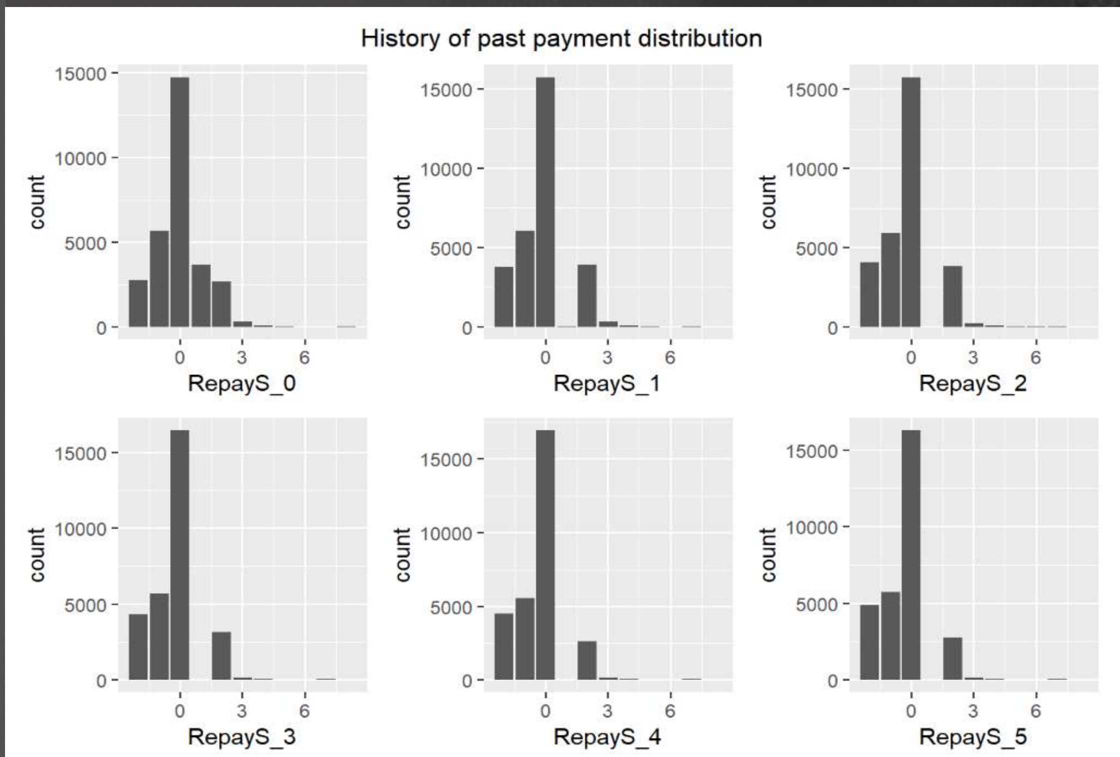


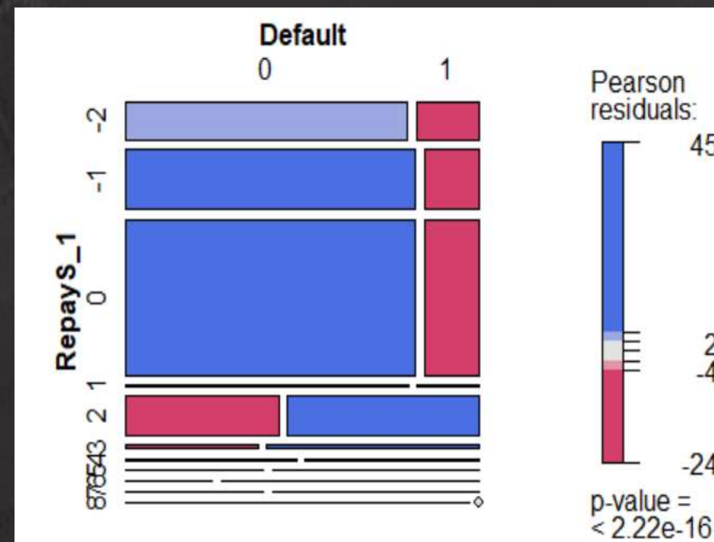


Marital status Distribution

Married customers have a higher probability of defaulting when compared to the single customers.

Summary of predictors' impact to dependent variable





Historical past repayments

Customers who have delayed payment of at least 1 month in any of the previous months, have an increased chance of default.

04 Feature Engineering

Feature Engineering of variables

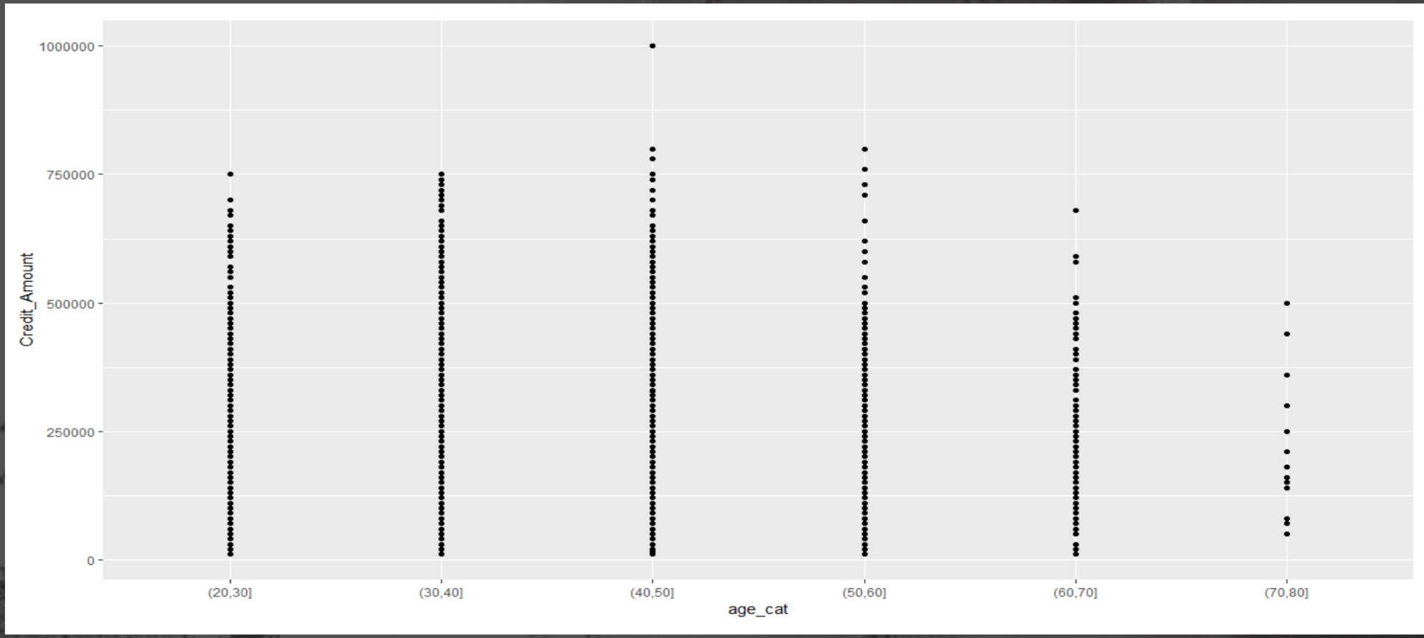
1. Analyze the classes of the education

```
## # A tibble: 8 x 3
##   Education      Default      n
##   <chr>          <chr>    <dbl>
## 1 Graduate_school 0       0.808
## 2 High_school     0       0.748
## 3 Others          0       0.929
## 4 University      0       0.763
## 5 Graduate_school 1       0.192
## 6 High_school     1       0.252
## 7 Others          1       0.0705
## 8 University      1       0.237
```

Customers with a high school degree and customers with a university degree have a higher probability of defaulting compared to customers who have graduate school degree and others.

Feature Engineering of variables

2. Create age buckets

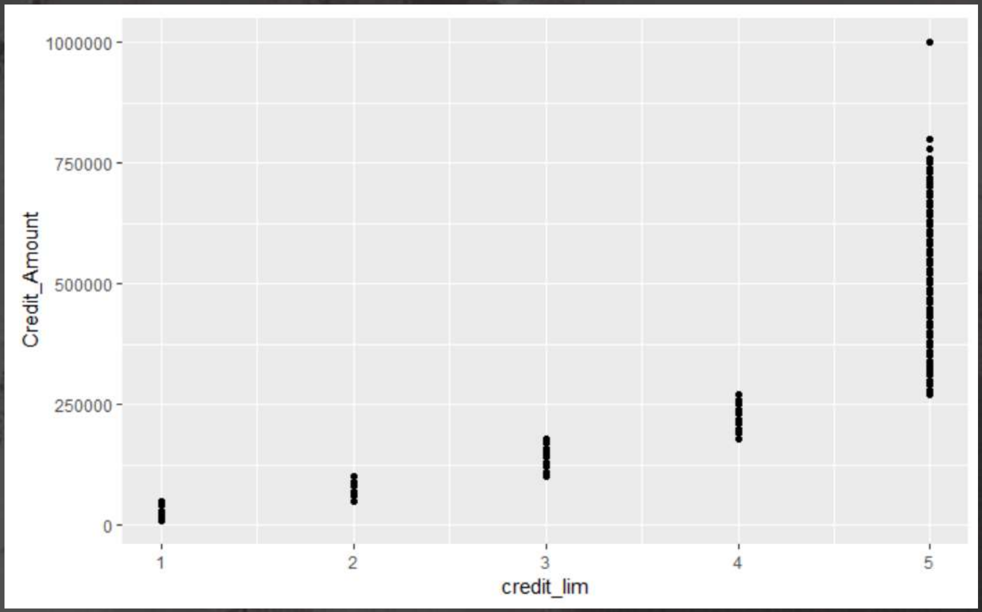


Credit limit for various age groups differs. Customers less than 50 years have higher credit limit as compared to the older customers

Feature Engineering of variables

3. Create buckets for credit limit

credit_lim <int>	min(Credit_Amount) <int>	max(Credit_Amount) <int>
1	10000	50000
2	50000	100000
3	100000	180000
4	180000	270000
5	270000	1000000



Created a new column based on credit limit and converted the values to multiple binarized vectors

Feature Engineering of variables

4. One-Hot Encoding

We converted categorical variables into a form of binary variable for each unique integer value where we applied algorithms to do predictions.

Gender	Education	Marital_status
F	University	Married
F	University	Single
F	University	Single
F	University	Married
M	University	Married
M	Graduate_school	Single
M	Graduate_school	Single
F	University	Single
F	High_school	Married

Gender_F	Gender_M	Education_Graduate_school	Education_High_school	Education_Others	Education_University	Marital_status_Married	Marital_status_Others	Marital_status_Single
1	0	0	0	0	1	1	0	0
1	0	0	0	0	1	0	0	1
1	0	0	0	0	1	0	0	1
1	0	0	0	0	1	1	0	0
0	1	0	0	0	1	1	0	0
0	1	1	0	0	0	0	0	1
0	1	1	0	0	0	0	0	1
1	0	0	0	0	1	0	0	1
1	0	0	1	0	0	1	0	0

Feature Engineering of variables

Other new features introduced into the dataset includes:

Amount owed = (Cumulative sum of Bill statement - Cumulative sum of payment amount) for each customer

Average amount owed over a 6-month period = (Amount owed /6) for each customer

Balance to limit ratio = round(Average amount owed over a 6-month period /credit limit, 3)

05 Data Modeling

Data Modeling

The business problem in this case is classified under supervised machine learning. Here we have historical data with independent variables (x) and a dependent variable (Y) and we want to use an algorithm to learn the mapping function from the input to the output and train a model to predict the dependent variable (Y).

Confusion Matrix

True Positive (TP)	A customer who is a defaulter and predicted by the model as a defaulter
True Negative (TN)	A customer who is a non-defaulter and predicted by the model as non-defaulter.
False Positive (FP)	A customer who is predicted by the model as a defaulter is a non-defaulter.
False Negative (FN)	A customer who is predicted as a non-defaulter is a defaulter.

	Non-Defaulter (predicted) - 0	Defaulter (predicted) – 1
Non-Defaulter (actual) – 0	**TN**	**FP**
Defaulter (actual) – 1	**FN**	**TP**

Data Modeling

1. Logistic regression model:

It is used for classification tasks which uses a linear equation with independent predictors to predict a value.

```
          Reference
Prediction    0    1
         0 4426  838
         1  241  488

               Accuracy : 0.82
                 95% CI : (0.81, 0.8296)
    No Information Rate : 0.7787
    P-Value [Acc > NIR] : 1.991e-15

                  Kappa : 0.3772
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9484
            Specificity : 0.3680
         Pos Pred Value : 0.8408
         Neg Pred Value : 0.6694
             Prevalence : 0.7787
         Detection Rate : 0.7385
   Detection Prevalence : 0.8784
      Balanced Accuracy : 0.6582

       'Positive' Class : 0
```

2. Naïve Bayes Classifier:

This classifier is based on Bayes theorem.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3108  382
         1 1559  944

               Accuracy : 0.6761
                 95% CI : (0.6641, 0.688)
    No Information Rate : 0.7787
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2868
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6660
            Specificity : 0.7119
         Pos Pred Value : 0.8905
         Neg Pred Value : 0.3771
             Prevalence : 0.7787
         Detection Rate : 0.5186
   Detection Prevalence : 0.5823
      Balanced Accuracy : 0.6889

       'Positive' Class : 0
```

Data Modeling

3. Stochastic Gradient Boosting :

This model is used for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4422  832
         1  245  494

               Accuracy : 0.8203
                 95% CI : (0.8103, 0.8299)
    No Information Rate : 0.7787
    P-Value [Acc > NIR] : 1.179e-15

                  Kappa : 0.3803
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9475
            Specificity : 0.3725
         Pos Pred Value : 0.8416
         Neg Pred Value : 0.6685
             Prevalence : 0.7787
         Detection Rate : 0.7379
   Detection Prevalence : 0.8767
      Balanced Accuracy : 0.6600

       'Positive' Class : 0
```

4. Linear Discriminant Analysis:

This is a classification technique which takes labels into consideration. The goal of Linear Discriminant Analysis is to project the features in higher dimension space onto a lower dimensional space.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4400  267
         1  809  517

               Accuracy : 0.8205
                 95% CI : (0.8105, 0.8301)
    No Information Rate : 0.8692
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3897
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8447
            Specificity : 0.6594
         Pos Pred Value : 0.9428
         Neg Pred Value : 0.3899
             Prevalence : 0.8692
         Detection Rate : 0.7342
   Detection Prevalence : 0.7787
      Balanced Accuracy : 0.7521

       'Positive' Class : 0
```

Data Modeling

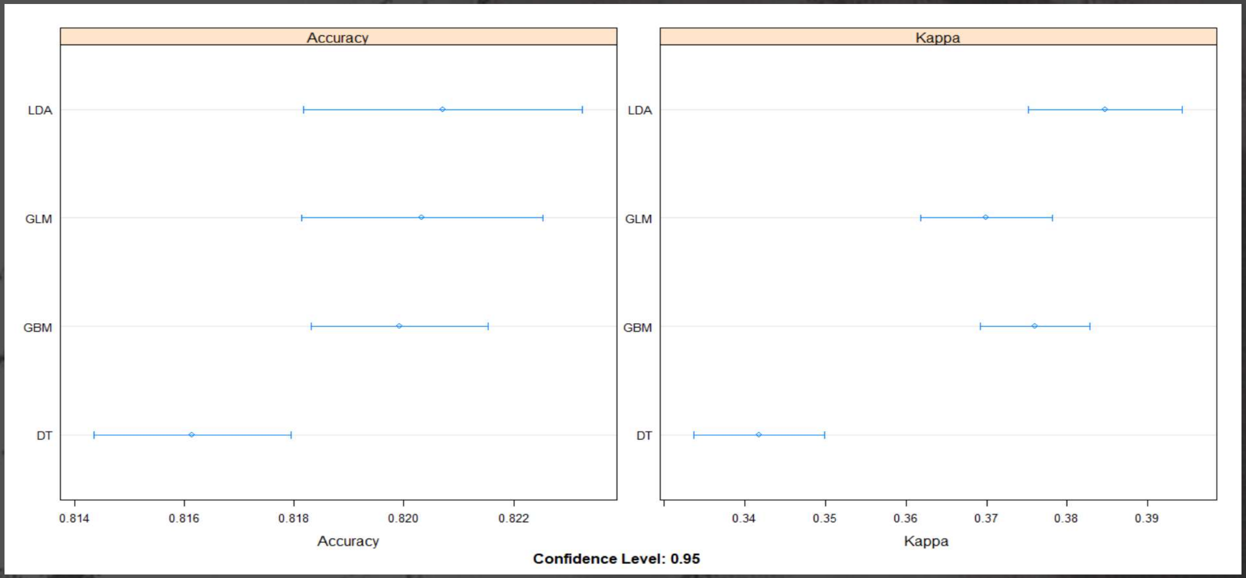
5. Decision Tree:

Decision Trees are broadly used supervised models for classification and regression tasks. A decision tree can be used to visually and explicitly represent decisions and decision making.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4478  189
         1  890  436

               Accuracy : 0.82
                 95% CI : (0.81, 0.8296)
    No Information Rate : 0.8957
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3556
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8342
            Specificity : 0.6976
         Pos Pred Value : 0.9595
         Neg Pred Value : 0.3288
             Prevalence : 0.8957
         Detection Rate : 0.7472
   Detection Prevalence : 0.7787
      Balanced Accuracy : 0.7659

       'Positive' Class : 0
```

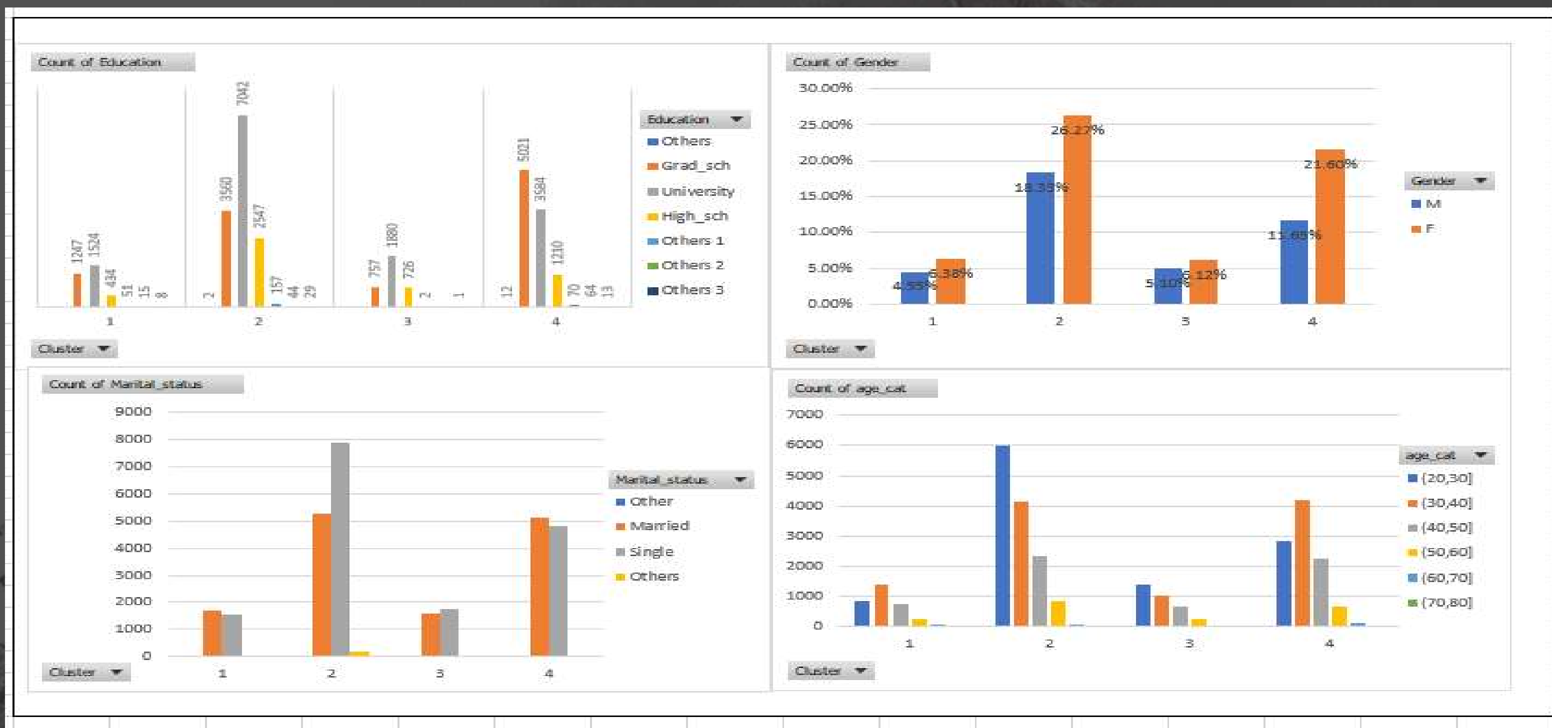
Model Selection - Gradient Boosting

After using various machine learning algorithms the best performing model was chosen based on minimum false negative value and accuracy.

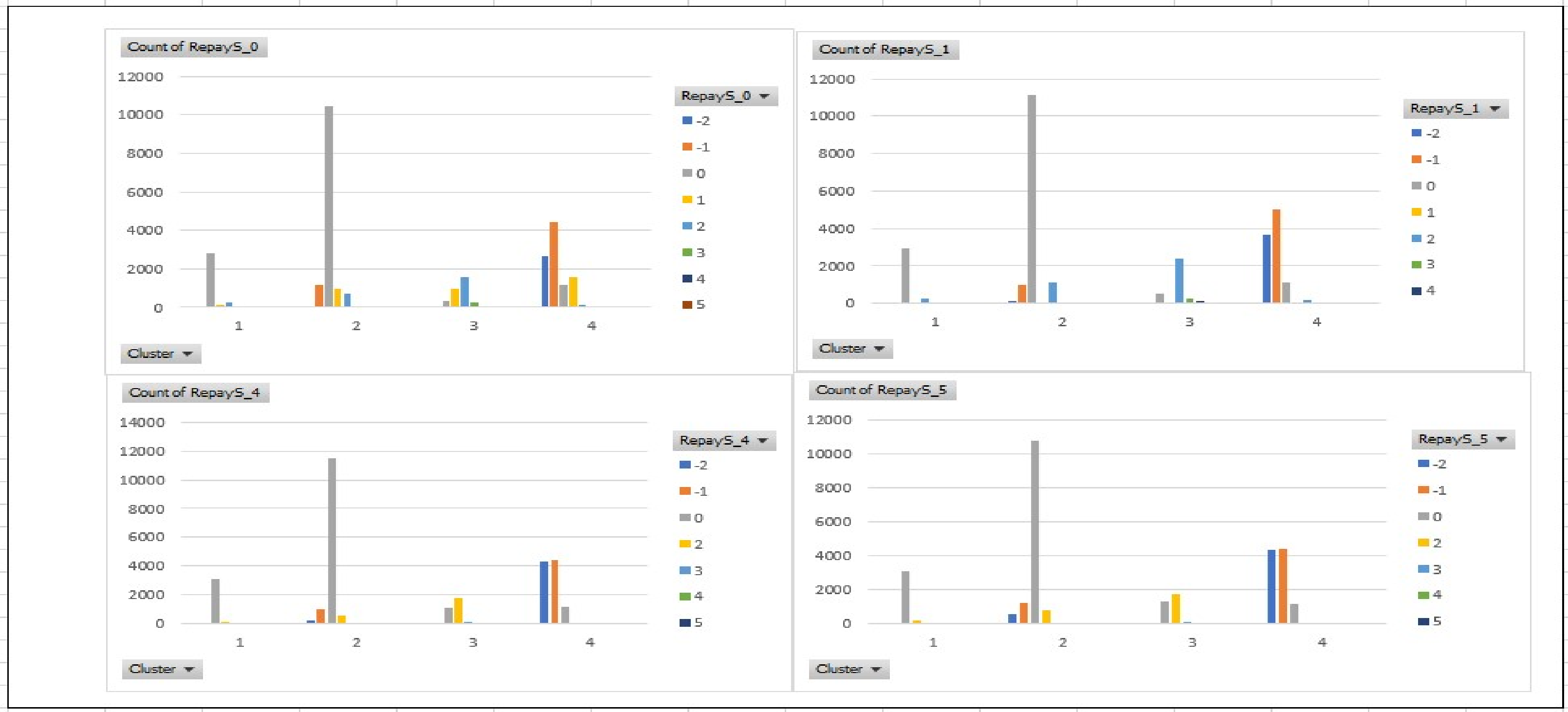


06 Business Insights

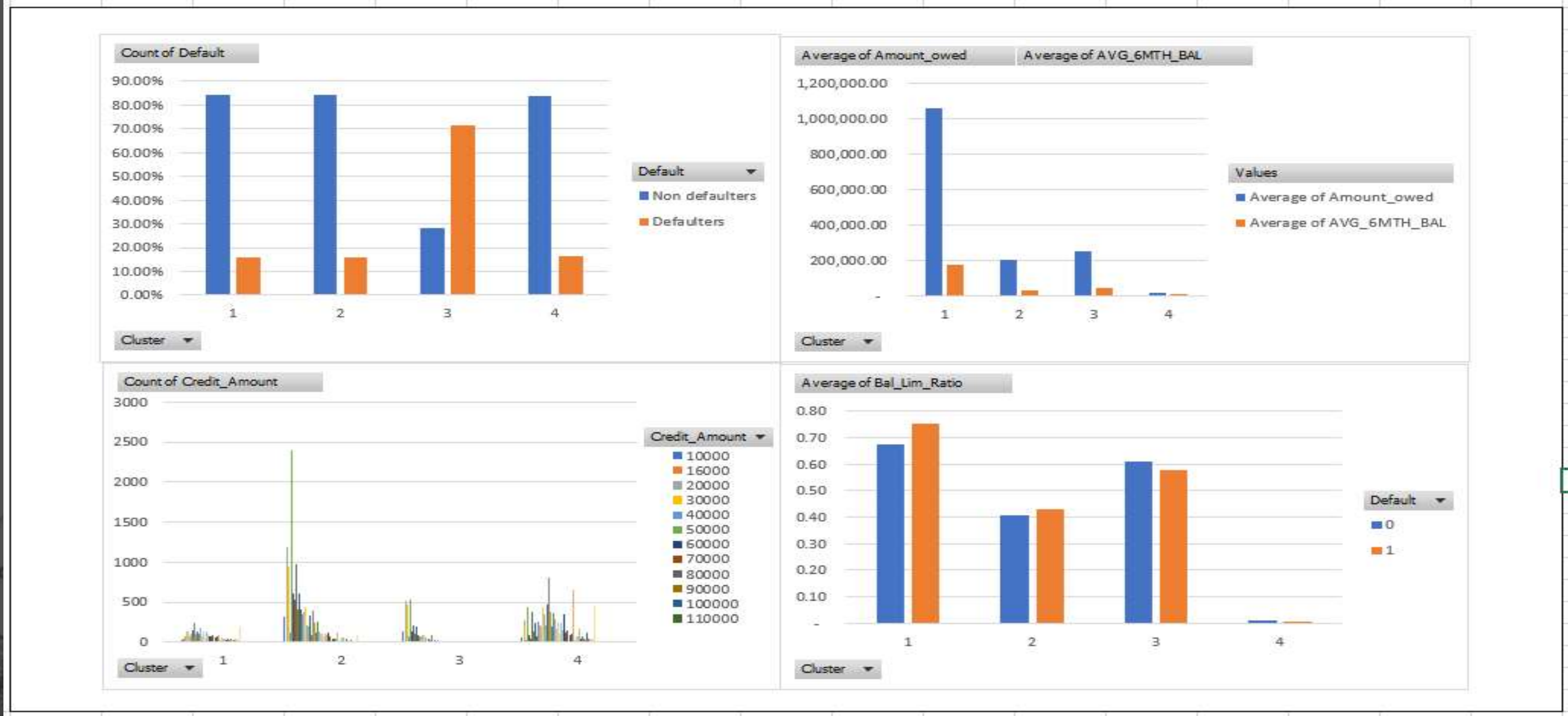
Segmentation of customers - Demographic profile



Segmentation of customers - Past monthly payment



Segmentation of customers – Default rate



Business insights

Group 1 – Middle age low default customers
- Balanced mix of married and single customers with average age of 37 years
- Pay their credit liability when due
- Customers in this group have high credit liability

Group 2 – Young university graduates low default customers
- Mostly single female customers with average age of 34 years
- Pay their credit liability when due and have a low default rate
- Customers in this segment have a university degree and higher degrees

Group 4 – Married Higher degree low debt customers
- Balanced mix of married and single customers with average age of 37 years
- Pay their credit liability early
- Customers in this segment have low credit liability
- Customers in this segment have a university degree and higher degrees

Group 3 – Late paying high default customers
- Single and married customers with average age of 35 years
- Highest percentage of defaulters above 70%
- Have an average of 2 months payment delay and have a high default rate

Recommendations

Group 2 – Young university graduates low default customers

• The company can increase the credit limit of customers in this segment if the customer is interested

• Offer reward packages and loyalty programs for customers in this segment

Group 1 – Middle age low default customers

•The company can reduce the credit limit of customers in this segment

• The company can send regular mails and offer incentives and coupons to encourage the customers reduce their outstanding liability

Group 3 – Late paying high default customers

• The company can send regular remainders to customers in this segment bimonthly on their outstanding liability

• The company can reduce the credit limit of customers in this segment

Group 4 – Higher degree low debt customers

• Advertise and target more customers with similar demographics to customers in this segment

• Offer reward packages and loyalty programs for customers in this segment

• The company can increase the credit limit of customers in this segment if the customer is interested

Conclusion

I performed data cleansing, exploration and visualization of the dataset to identify key drivers and their relationship with default rate (Y).

Trained and tested 5 machine learning models decision tree, Gradient boosting, Naïve Bayes, Linear Discriminant Analysis and logistic regression models to predict the customers' probability of default.

Based on the business insights I have given recommendations to improve the business.


Thank you