Credit Risk Data Analysis

2STAT4.5 – PROJECT DISSERTATION

1)Amartya Banerjee

Roll No.: 96/STA-190005

2) Sunetra Mukherjee

Roll No.: 96/STA – 190048

(Department of Statistics, University of Kalyani, Nadia)

Under the Guidance of Dr. Kiranmoy Chattopadhyay

(Assistant Professor, Department of Statistics , Bidhannagar Govt. College)

Credit Risk Data Analysis

1)Introduction

One of the leading banks would like to predict whether a customer is eligible to apply for loan or not. This model also called as PD Models (Probability of Default).

Credit scoring is perhaps one of the most "classic" applications for predictive modelling, to predict whether or not credit extended to an applicant will likely result in profit or losses for the lending institution. There are many variations and complexities regarding how exactly credit is extended to individuals, businesses, and other organizations for various purposes (purchasing equipment, real estate, consumer items, and so on), and using various methods of credit (credit card, loan, delayed payment plan). But in all cases, a lender provides money to an individual or institution, and expects to be paid back in time with interest commensurate with the risk of default. Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit.

These techniques determine who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrowers to the lenders. Further, they help to assess the risk in lending. Credit scoring is a dependable assessment of a person's credit worthiness since it is based on actual data. A lender commonly makes two types of decisions: first, whether to grant credit to a new applicant, and second, how to deal with existing applicants, including whether to increase their credit limits. In both cases, whatever the techniques used, it is critical that there is a large sample of previous customers with their application details, behavioural patterns, and subsequent credit history available. Most of the techniques use this sample to identify the connection between the characteristics of the consumers (annual income, age, number of years in employment with their current employer, etc.) and their subsequent history.

Typical application areas in the consumer market include: credit cards, auto loans, home mortgages, home equity loans, mail catalogue orders, and a wide variety of personal loan products. Here we mainly have concentrated on loans provided by bank.

In our study and analysis, we have used and fitted various models using logistic regression method and finally tried to draw a conclusion regarding the loan status which is a categorical variable.

2)Summary of the Dataset: -

2.1: Data Structure:

We have 2 data sets to run the analysis. The 1st one, train data set contains 13 columns and 614 rows and the 2nd one is test data set with 13 Columns and 366 rows. The columns are respectively Loan Id (i.e., the ID of the customer which is unique), Gender (Male or Female), Dependence (i.e., how many members are dependent on the applicant), Education (Graduate or Not Graduate), Self Employed (yes or no), Applicants Income (in Rupees), Co-applicants' income (in rupees if any), Loan Amount, Loan Amount Term, Credit History, Property Area (Rural or Urban), and finally the Loan Status (Yes or No). Here, we combined our data and then split the data in order to have more data to train data and predict on the remaining data.

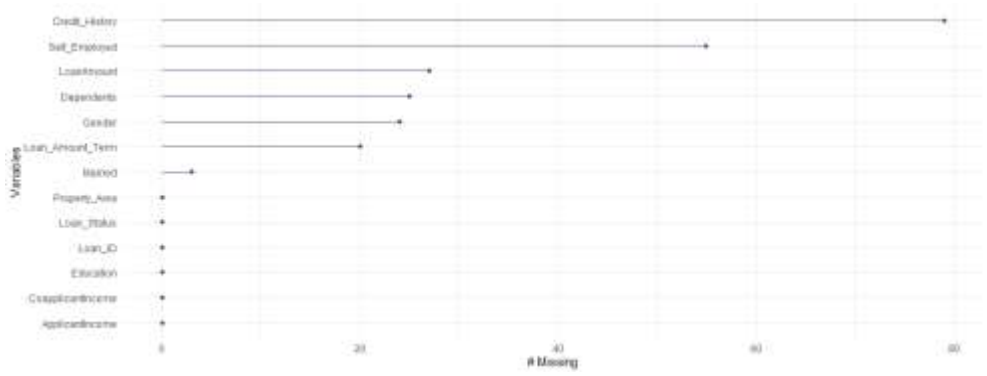
The data set:

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
LP001034	Male	No	1	Not Graduate	No	3596	0	100	240		Urban	Y

2.2: Data Overview:

To detect if there is any missing value firstly, we have run the data to observe the columns data. Once we can detect the columns with maximum number of missing values we can observe and decide things more accurately for our further analysis. First, we will plot a graph using nanir package in R. Figure 1 shows total number of missing values per columns. It's clear that credit history, Self Employed, Loan amount, dependents, Genders, Loan Amount term and Married columns have certain missing values.

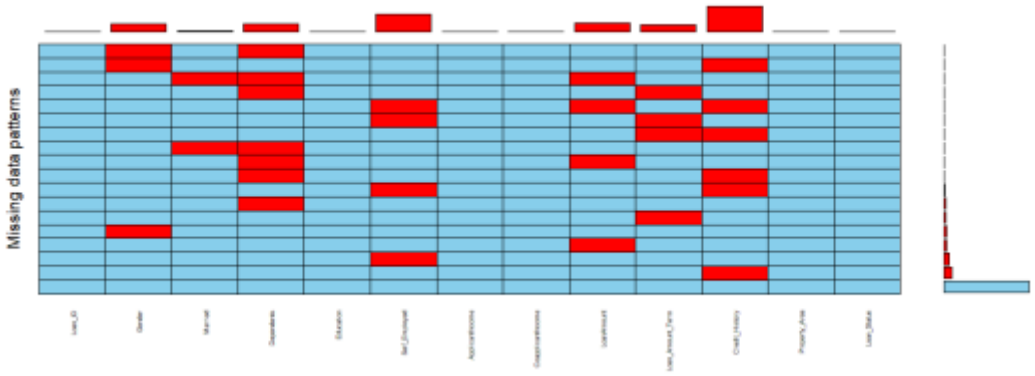
Figure 1: Total Number of missing values in each independent variable



Now to understand clearly the density of missing values per columns and per row we need deeper analysis which cannot be determined through above graph. So, Aggregation Plot is used indicating missing values as

red. Now the bar plots above indicating total number of missing values per columns while the histograms on the right indicate the density of missing value occurrences.

Figure 2: Aggregation Plot showing the pattern of the missing values of independent variable



As we can in the Figure 2, there is no similar pattern in missing values by row wise and column wise. So, for analysis purpose we removed the rows containing missing values and then carry-on further analysis.

3)Analysis

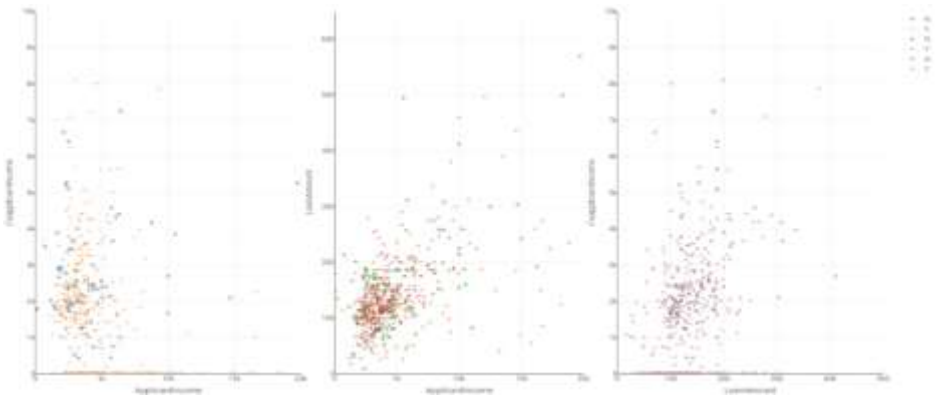
Part I: Studying the data based on different plots and charts and trying to understand relationships between different variables

a) Scatter Plot:

For our dataset we will look into the linear relationship between continuous independent variables. In this data set we have three continuous variable Co-applicant Income, Applicant Income and Loan Amount. We will verify whether there is any linear relationship between these continuous variables.

In Figure 3, left most graph showing Co-applicant Income VS Applicant Income, middle graph showing Applicant Income VS Loan Amount and right most graph showing Co-applicant Income VS Loan Amount plot. Loan Status (Yes or No) have been shown in graphs by different colours of plot. Which colour indicating which status is shown in the legend attached to the plot.

Figure 3: Three different scatter diagram with respect to their variables



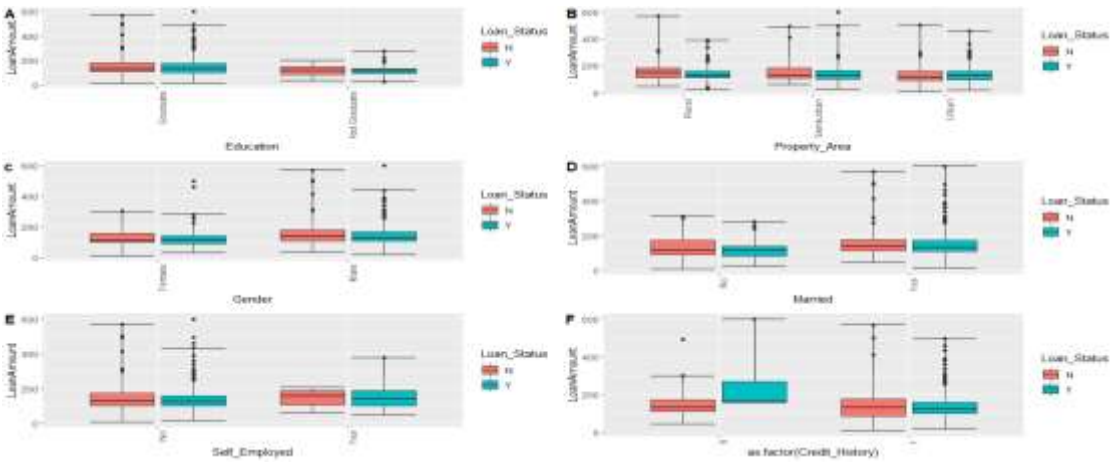
Here we see that in each scatter plot, the responses of Yes and No mixed altogether. So, nothing could be said about relationships on these three graphs. So, we could say that there is **may not be any corelation** between Applicant Income, Co applicant Income and Loan Amount with respect to Loan Status.

b) Box Plot:

A boxplot is a standardized way of displaying the distribution of data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

For our data, we used this visualization to look for the outliers present in the data

Figure 4: Six boxplots for various independent variables

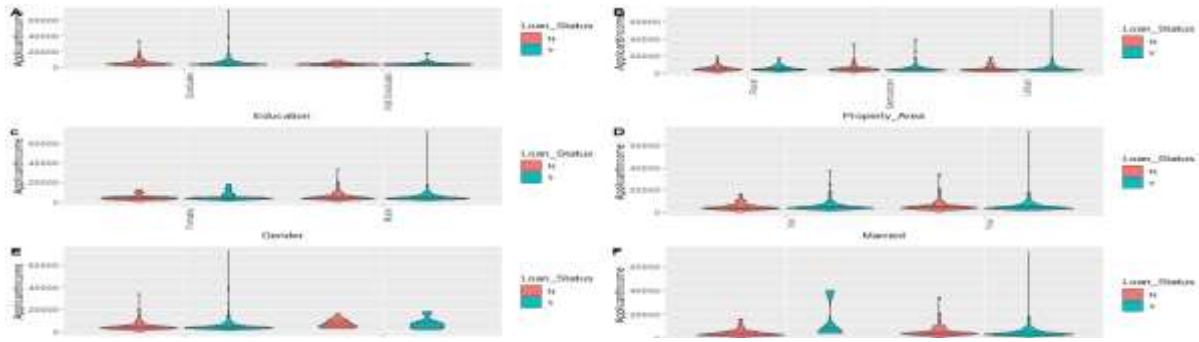


As we can see that there is not much difference of positions of box plots except for Credit History in their Marital Status (For Non-Married Case), So we can say that there may be some effect of marital status with loan status.

c) Violin Plot:

In general, violin plots are a method of plotting numeric data and can be considered a combination of the box plot with a kernel density plot. In the violin plot, we can find the same information as in the box plots: median (a white dot on the violin plot), interquartile range (the black bar in the centre of violin), the lower/upper adjacent values. These values can be used in a simple outlier detection technique (Tukey's fences) — observations lying outside of these "fences" can be considered outliers. The advantage of the violin plot over the box plot is that aside from showing the above-mentioned statistics, it also shows the entire distribution of the data. We do this especially when dealing with multimodal data, i.e., a distribution with more than one peak.

Figure 5: Six violin plots for various independent variables



In these figures since all plot have good number of outliers as we can see that black lines are extended above. We used this violin to detect variability in the data and but we cannot find any significant difference as the violins in any plot showing any good width which actually indicates variance. Therefore, we cannot find any conclusion in violin diagrams that could have helped in further proceeding.

d) Mosaic Plot:

A **mosaic plot** (also known as a **Marimekko diagram**) is a graphical method for visualizing data from two or more qualitative variables. It is the multidimensional extension of spine plots, which graphically display the same information for only one variable. It gives an overview of the data and makes it possible to recognize relationships between different variables.

In our data the categorical variables are first put in order. Then, each variable is assigned to an axis. We have plotted the loan stats in YES or NO with every constraint and tried to understand which factor has more effect in approval of a Loan Status.

Other than Mosaic Plot, we have also calculated Mosaic Matrix for testing of Association of two independent variable. So,

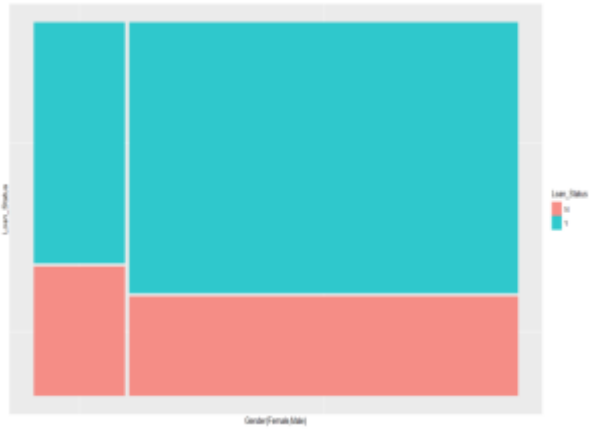
Null Hypothesis: There is no association between two variables

Alternative Hypothesis: There is association between two variables

Here, we describe our mosaic plot and matrix so that it can be understandable for the figures. For Mosaic Plot, Gender is plotted in X axis and divided into levels and each of these cases, how many values of Loan Status (Y axis) of "YES" and "NO" are proportionally with respect to each case of Gender. As for Mosaic Matrix, for each cell, there are data written in different colours. Each colour has different meaning which are one by one. Data written in numeric values in **black** colour means **Observed Value,** numeric values in **crimson** colour means **Expected value,** percentage values in **violet** colour means percentage data within Loan Status category and percentage values is **green colour** means percentage data within Gender category.

The Mosaic Plot for loan status against different factors is as follows: -

Figure 6: Mosaic Plot and Mosaic Matrix for Gender Vs Loan Status: -

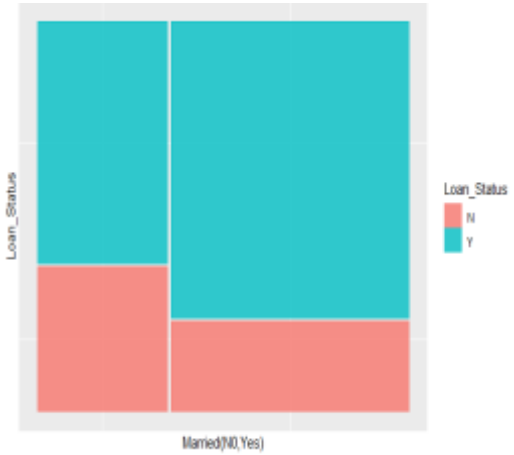


Gender	*Loan_Status* N	Y	***Total***
Female	38	71	109
	31	78	109
	34.9 %	65.1 %	100 %
	23.3 %	17.2 %	18.9 %
Male	125	342	467
	132	335	467
	26.8 %	73.2 %	100 %
	76.7 %	82.8 %	81.1 %
Total	163	413	576
	163	413	576
	28.3 %	71.7 %	100 %
	100 %	100 %	100 %

χ^2=2.470 · df=1 · φ=0.070 · p=0.116

From the p value which is calculated from Chi Square Distribution, the value is greater than 0.05, so there is *no significance association* between Gender and Loan Status.

Figure 7: Mosaic Plot and Mosaic Matrix for Marital Status Vs Loan Status: -

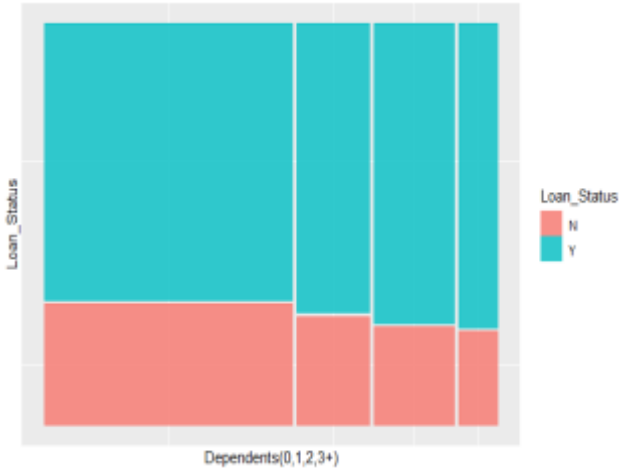


Married	*Loan_Status* N	Y	***Total***
No	76	127	203
	57	146	203
	37.4 %	62.6 %	100 %
	46.6 %	30.8 %	35.2 %
Yes	87	286	373
	106	267	373
	23.3 %	76.7 %	100 %
	53.4 %	69.2 %	64.8 %
Total	163	413	576
	163	413	576
	28.3 %	71.7 %	100 %
	100 %	100 %	100 %

χ^2=12.220 · df=1 · φ=0.150 · p=0.000

There is *significance association* between Martial Status vs Loan Status (p value is less than 0.05)

Figure 8: Mosaic Plot and Mosaic Matrix for Dependents Vs Loan Status: -

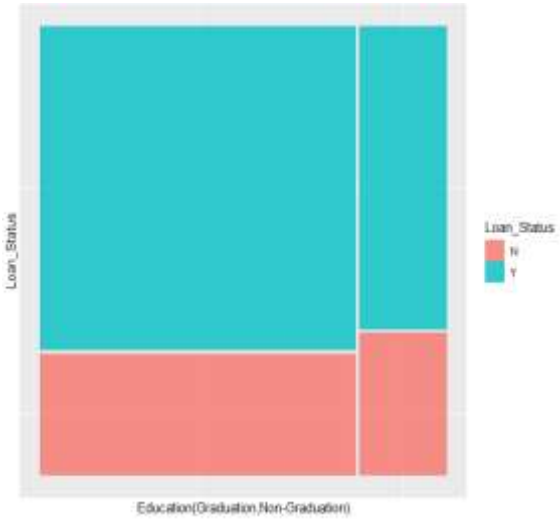


clents	Loan_Status N	Loan_Status Y	Total
	99	226	325
	92	233	325
	30.5 %	69.5 %	100 %
	60.7 %	54.7 %	56.4 %
	26	69	95
	27	68	95
	27.4 %	72.6 %	100 %
	16 %	16.7 %	16.5 %
	26	79	105
	30	75	105
	24.8 %	75.2 %	100 %
	16 %	19.1 %	18.2 %
	12	39	51
	14	37	51
	23.5 %	76.5 %	100 %
	7.4 %	9.4 %	8.9 %
	163	413	576
	163	413	576
	28.3 %	71.7 %	100 %
	100 %	100 %	100 %

09 · df=3 · Cramer's V=0.039 · p=0.571

There is *no significance association* between Dependents vs Loan Status (p value is greater than 0.05)

Figure 9: Mosaic Plot and Mosaic Matrix for Education Vs Loan Status: -

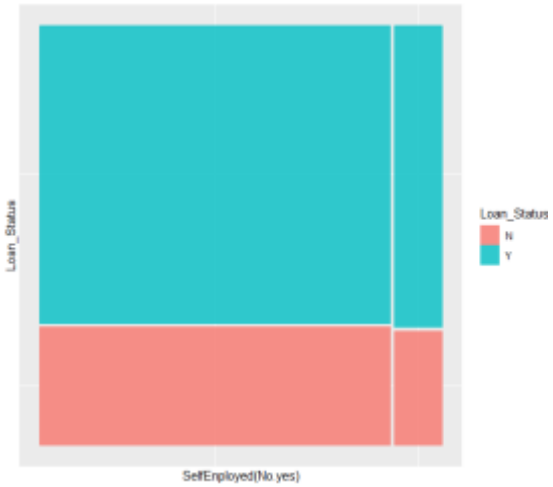


There is *no significance association* between Education vs Loan Status (p value is greater than 0.05)

Education	Loan_Status N	Loan_Status Y	Total
Graduate	123	328	451
	128	323	451
	27.3 %	72.7 %	100 %
	75.5 %	79.4 %	78.3 %
Not Graduate	40	85	125
	35	90	125
	32 %	68 %	100 %
	24.5 %	20.6 %	21.7 %
Total	163	413	576
	163	413	576
	28.3 %	71.7 %	100 %
	100 %	100 %	100 %

$\chi^2=0.858 \cdot df=1 \cdot \varphi=0.043 \cdot p=0.354$

Figure 10: Mosaic Plot and Mosaic Matrix for Self Employed Vs Loan Status: -

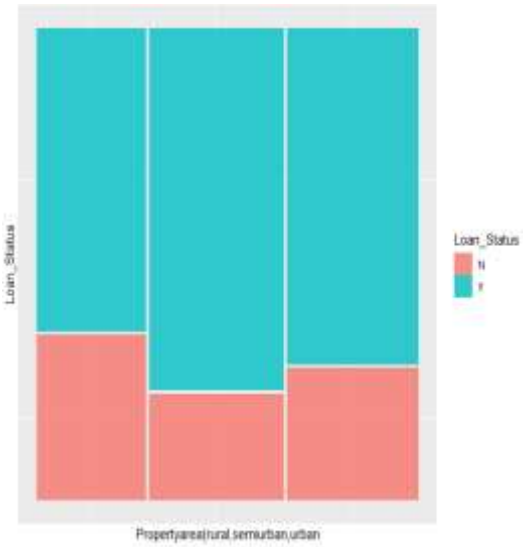


Self_Employed	*Loan_Status* N	Y	***Total***
No	144	363	507
	143	364	507
	28.4 %	71.6 %	100 %
	88.3 %	87.9 %	88 %
Yes	19	50	69
	20	49	69
	27.5 %	72.5 %	100 %
	11.7 %	12.1 %	12 %
Total	163	413	576
	163	413	576
	28.3 %	71.7 %	100 %
	100 %	100 %	100 %

χ^2=0.000 · df=1 · φ=0.006 · p=0.994

There is *no significance association* between Self Employed vs Loan Status (p value is greater than 0.05)

Figure 11: Mosaic Plot and Mosaic Matrix for Property Area Vs Loan Status: -

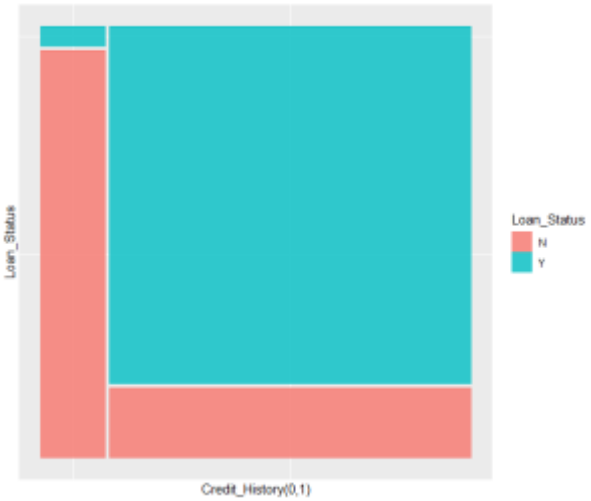


Property_Area	*Loan_Status* N	Y	***Total***
Rural	59	108	167
	47	120	167
	35.3 %	64.7 %	100 %
	36.2 %	26.2 %	29 %
Semiurban	47	160	207
	59	148	207
	22.7 %	77.3 %	100 %
	28.8 %	38.7 %	35.9 %
Urban	57	145	202
	57	145	202
	28.2 %	71.8 %	100 %
	35 %	35.1 %	35.1 %
Total	163	413	576
	163	413	576
	28.3 %	71.7 %	100 %
	100 %	100 %	100 %

χ^2=7.261 · df=2 · Cramer's V=0.112 · p=0.027

There is *significance association* between Property Area vs Loan Status (p value is lesser than 0.05)

Figure 12: Mosaic Plot and Mosaic Matrix for Credit



Credit_History	*Loan_Status* N	Y	***Total***
0	83	4	87
	25	62	87
	95.4 %	4.6 %	100 %
	50.9 %	1 %	15.1 %
1	80	409	489
	138	351	489
	16.4 %	83.6 %	100 %
	49.1 %	99 %	84.9 %
Total	163	413	576
	163	413	576
	28.3 %	71.7 %	100 %
	100 %	100 %	100 %

$\chi^2=223.543 \cdot df=1 \cdot \varphi=0.628 \cdot p=0.000$

History Vs Loan Status: -

There is *significance association* between Credit History vs Loan Status (p value is lesser than 0.05)

Part II: Fitting Appropriate Model

As we see that the dependent variable of our data is Categorical Variable which has only Two levels. (Loan Status -> Yes and No). For this, general linear model will not be applicable here as the predicted value will cluster into two classes which is not proper. We need a model such that it will help us to decide whether we should give loan or not. For this we will apply **Logistic Regression** model. To apply this logistic regression model on our data, we will use R software and its package. But before that we should know what is logistic regression and why it will help on our data in detail.

a) Logistic Regression:

In the linear regression model $X\beta + \varepsilon$, there are two types of variables – explanatory variables $X_1, X_2, \ldots X_k$and study variable y. These variables can be measured on a continuous scale as well as like an indicator variable.
When the explanatory variables are qualitative, then their values are expressed as indicator variables, and then dummy variable models are used.
When the study variable is a qualitative variable, then its values can be expressed using an indicator variable taking only two possible values 0 and 1. In such a case, the logistic regression is used. For example, y can denote the values like success or failure, yes or no, like or dislike, which can be denoted by two values 0 and 1.

The log-odds:

$$\frac{p}{1-p} = b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$$

By simple algebraic manipulation (and dividing numerator and denominator by $b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$ *), the probability that* $Y = 1$ *is*

$$p = \frac{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k} + 1} = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$
$$= S_b(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots . + \beta_k x_k)$$

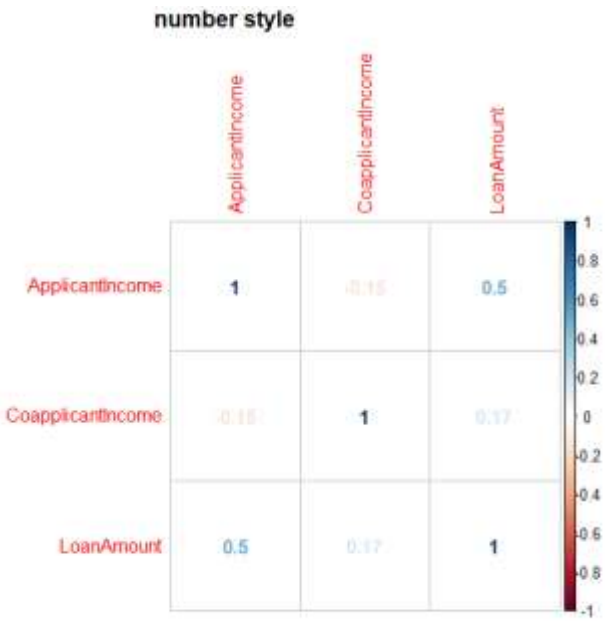
Where S_b is the sigmoid function with base b. The above formula shows that once β_i are fixed, we can easily compute either the log-odds that $Y = 1$ for a given observation, or the probability that $Y = 1$ for a given observation. The main use-case of a logistic model is to be given an observation $(x_1, x_2, \ldots, x_k)$, and estimate the probability p that $Y = 1$. In most applications, the base b of the logarithm is usually taken to be e. However, in some cases it can be easier to communicate results by working in base 2, or base 10.

Now for our data, we have check that whether our data, independent variable are correlated or not (multicollinearity present or not). For this we compute correlation matrix for our numerical independent variables.

b) Correlation Matrix:

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all possible pairs of values in a table. Now for our data, we have check that whether our data, independent variable are correlated or not (multicollinearity present or not). For this we compute correlation matrix for our numerical independent variables.

Figure 13: Correlation Matrix for our Continuous independent variable



As we can see that the correlation is quite low between numerical independent variables, we can assume that there is **no collinearity** present in our numerical independent variable.

c) Choosing our optimal model:

To choose our optimal model, we at first create our model taking all independent variables (both categorical and numerical) and compute their AIC value. After that we remove one by one those variables whose does not that have that much influence our predicting loan status. By doing this process, after certain point of time, we

will get our lowest value possible of AIC and after that AIC value will increase. We will simply choose those AIC value models those have lower values. Before doing that, let us know brief explanation AIC value.

Step 1: AIC:

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

AIC is founded on information theory. When a statistical model is used to represent the process that generated the data, the representation will almost never be exact; so, some information will be lost by using the model to represent the process. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of under fitting.

The Akaike information criterion is named after the Japanese statistician Hirotugu Akaike, who formulated it. It now forms the basis of a paradigm for the foundations of statistics and is also widely used for statistical inference.

In time of analysis, we started with a Logistic Regression model having all independent variable. In that model, we calculated p value of each independent variable on whether they are significant variable or not. We also calculated AIC value of the model. Now, our goal is to reduce AIC as much as possible. So, for the better of the model, we will eliminate those independent variables which are not that much important. So, we will eliminate that independent variable which has the most p value as the more p value of an independent variable, the less significant to the model. By this way, we remove one variable and again fit a model. After that, we do the same things again and again until we see a case where removing an independent variable will increase the AIC value instead of decreasing. In that situation, we will choose that model which has lowest AIC value. In this way, we will get the best Logistic regression Model with lowest AIC value possible.

After doing such step-by-step analysis, the best model we get is the model where we have independent variables Martial Status, Credit Status ad Property Area. Here, we name this model as “Model 1”. The details of that model are given below:

```
Call:
glm(formula = Loan_Status ~ Married + Credit_History + Property_Area,
    family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2622  -0.2284   0.5186   0.6591   2.4261

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.1555     0.5900  -7.043 1.88e-12 ***
MarriedYes            0.7446     0.2461   3.026 0.002481 **
Credit_History1       4.8271     0.5382   8.969  < 2e-16 ***
Property_AreaSemiurban   1.0619     0.3103   3.422 0.000621 ***
Property_AreaUrban       0.5220     0.2841   1.837 0.066160 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 686.31  on 575  degrees of freedom
Residual deviance: 447.22  on 571  degrees of freedom
AIC: 457.22

Number of Fisher Scoring iterations: 5
```

Also, for in case of error, we also take our 2nd best model which is selected by AIC values (in short, 2nd lowest AIC valued model) and will do comparison with our best model chosen by AIC (There does not have that much of difference our best and 2nd best model in AIC values). So, our 2nd model has independent variables are Martial Status, Credit History, Property Area Status, Loan Amount (we name this Model as "Model 2"). The details of that model are given below:

```
Call:
glm(formula = Loan_Status ~ Married + Credit_History + Property_Area +
    LoanAmount, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3326  -0.2214   0.4960   0.6240   2.5623

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -3.879159   0.627896  -6.178 6.49e-10 ***
MarriedYes              0.810017   0.253445   3.196 0.001393 **
Credit_History1         4.813429   0.538199   8.944  < 2e-16 ***
Property_AreaSemiurban  1.051221   0.310978   3.380 0.000724 ***
Property_AreaUrban      0.491303   0.285447   1.721 0.085219 .
LoanAmount             -0.002044   0.001693  -1.207 0.227299
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 686.31  on 575  degrees of freedom
Residual deviance: 445.81  on 570  degrees of freedom
AIC: 457.81

Number of Fisher Scoring iterations: 5
```

Now logically, we believe that giving someone depends on how much he income. The more s/he can income, it has to chance to her/his loan, for this, s/he has better probability to get her/his loan. Although our data does not say so. For this we excluded Applicant Income variable. But due to our belief, we will add Applicant Income variable with our AIC optimized model and do further analysis. If further analysis does not give satisfying results, we will reject our self believe model. So, after adding our Applicant Income variable in our AIC optimized model, the details of that model (we name this Model as "Model 3") are given below with it's AIC value.

```
Call:
glm(formula = Loan_Status ~ Married + Credit_History + Property_Area +
    ApplicantIncome, family = binomial(), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3041  -0.2277   0.5167   0.6548   2.4271

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -4.173e+00  6.050e-01  -6.897 5.29e-12 ***
MarriedYes              7.438e-01  2.462e-01   3.021 0.002516 **
Credit_History1         4.827e+00  5.383e-01   8.967  < 2e-16 ***
Property_AreaSemiurban  1.062e+00  3.103e-01   3.422 0.000621 ***
Property_AreaUrban      5.221e-01  2.842e-01   1.837 0.066144 .
ApplicantIncome         3.601e-06  2.720e-05   0.132 0.894700
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 686.31  on 575  degrees of freedom
Residual deviance: 447.20  on 570  degrees of freedom
AIC: 459.2

Number of Fisher Scoring iterations: 5

> |
```

Step 2: Check for Overdispersion:

When we apply logistic regression in data, then we have to check that whether the observed variance is larger than the expected from the logistic model. If the dispersion is higher than expected, then overdispersion exists. It is a situation where the residual deviance of the model is large relative to the residual degrees of freedom. If Overdispersion exists, then it indicates that the model does not fit the data well. The explanatory variables may not describe the model. If there exists overdispersion, one potential solution Beta-Binomial family and Quasi- Likelihood method. One thumb rule to detect overdispersion exists or not is

$$\frac{Residual\ Deviance}{d.f._{residual}} > 1.5$$

If the ratio value is greater than 1.5, then overdispersion exists, otherwise it is not.

In our data

Model	Over dispersion ratio
Model1	0.7919
Model2	0.7936
Model3	0.7943

For each of the model we can see that the value of the ratios is less than 1.5. Clearly overdispersion does not exists.

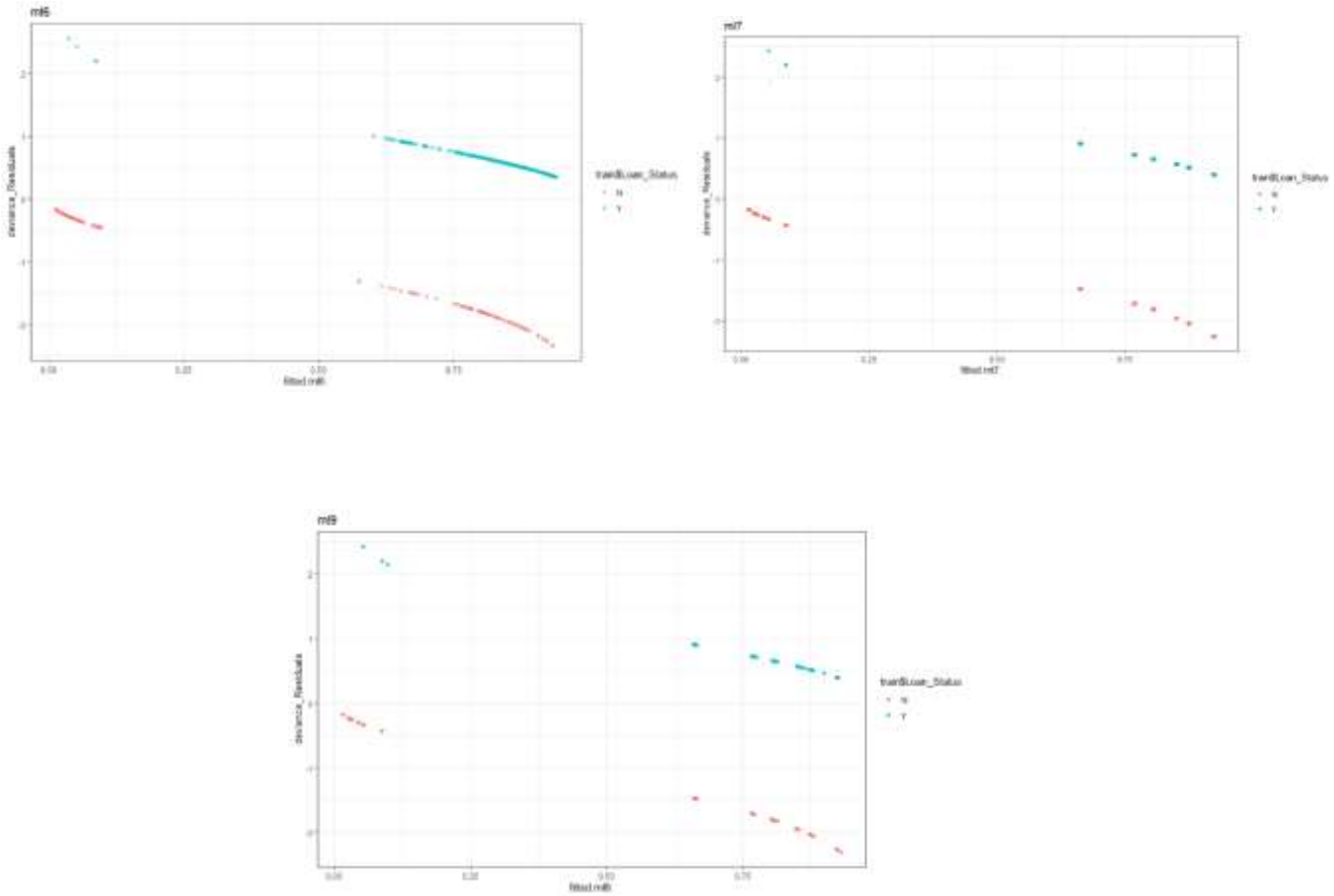
Step 3: Evaluation of fitting of Logistic Regression Model:

Here to see that whether our fitting of Logistic Regression in our data correct or not, we will observe two different plots.

a) Deviance Residual vs Fitted Plot:

With the help of this plot, we will check that whether our choose of our model is correct or not. To describe that whether our model fitting is correct or not, we, at first, we show our results of our three different models. The result of Model 1, 2 and 3 are given below respectively. (In coding, in our model naming, ml6 means Model 1, ml7 means Model 2, ml9 means Model 3)

Figure 14: Deviance residual vs Fitted Plot for our three different models



Now as we can see that all three graphs are similar. All three graphs similar pattern with blue and red colour points have significant gap between them and those different coloured points form curvy line with space between them.

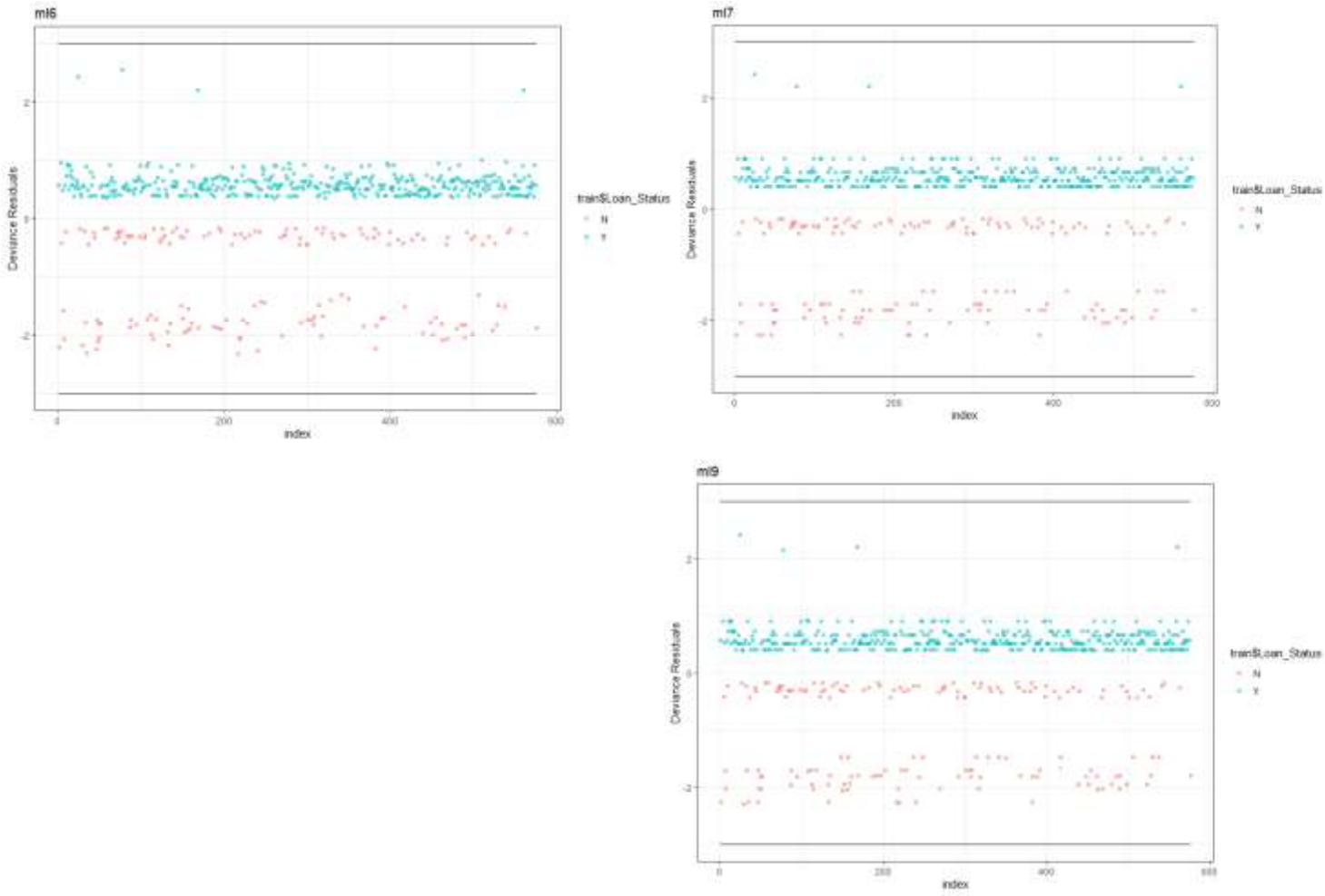
Now, we have to remember that our independent variable is categorical variable with two levels. On this kind of data, whether we can apply Negative Binomial or Poisson distribution or Logistic Regression. Now if our Residual vs Fitted Plot showed scattered points plotted over our plot then we should decide that we have to

plot Negative Binomial or Poisson and if our plot shows points are clustered (Kind of discrete) on our plot we should apply Logistic regression. Since we see kind of discrete pattern in our plot, our decision to fit Logistic Regression Model is correct.

b) Deviance Residual Graph:

With the help of this graph, we have to see whether our error lies outside the significant range. Also, we have to see that whether our error is mixed up or not, i.e., the decision of to give loan or not are mixed up or not. Here, we are showing graphs for three different models (Model 1 ,2 and 3 respectively) to check our previously described points

Figure 15: Deviance residual for our three different models



As we can see that for our 3 different models, the error lies between -3 to +3 value. Also, Loan Status decisions are not mixed each other. So, we can use our Logistic Regression Models.

Now we will do some further analysis about which models are good and which model is bad in the sense that which we work as they have better chance to give more accurate Loan Status

Step 3) Further Analysis using Standardized Pearsonian Test and Hosmer Lemeshow Test:

a) Standardized Pearsonian Test:

In this analysis, we will use our 3 models which are described before and do further analysis. In Standardized Pearsonian Test, we will calculate Chi Square Test statistic value for our three models. After that we will check that how much is there difference from the critical value at 5% level of significance and that value which closest from critical value. Here, the degrees of freedom of Chi Square distribution are 571 (For Model 1 and Model 3) and 572 (For Model 2). So critical values at 5% level of significance are 627.6989 and 628.7476. Before we do our analysis, we will give brief explanation about Standardized Pearsonian Test

Explanation:

With ungrouped data, the formula for the classic Pearson chi-square test is:

$$X^2 = \sum \frac{(yi - \hat{\pi}i)2}{(\hat{\pi}i(1-\hat{\pi}i))}$$

Where y_i is the dependent variable with values of 0 or 1
As we've just discussed, the problem with the classic Pearson GOF test is that it does not have a chi-square distribution when the data are not grouped. But Osius and Rojek (1992) showed that $X2$ has an asymptotic normal
distribution with a mean and standard deviation that they derived. Subtracting the mean and dividing by the Standard deviation yields a test statistic that has approximately a standard normal distribution under the null hypothesis. McCullagh (1985) derived a different mean and standard deviation after conditioning on the vector of estimated regression coefficients. In practice, these two versions of the standardized Pearson are nearly identical, especially in larger samples. Farrington (1996) also proposed a modified X^2 test, but his test does not work when there is only one case per profile. For the remainder of this paper, I shall refer to the standardized Pearson test as simply the Pearson test.

Here the Chi Square Test Statistic values of our models are given below

	Model 1	Model 2	Model 3
Test Statistic Value	552.4793	544.9928	546.3952

As we can that there is not that much of difference between Chi Square of Test Statistic values for three models. So cannot conclude anything using Standardized Pearsonian test

After applying Standardised Pearsonian Test, we will apply Hosmer Lemeshow test. Here, for three different models, we will apply 3 different times (one, using out fitted grouped in 8 sections, another time with 16 sections and another time 32 sections), In different analysis, we will compare their p values and compare that with each other. AS we will increase number of groups in our test, it is expected to also their p value. A short note on Hosmer Lemeshow test is given and after the analysis of our data is given after that.

b) Hosmer Lemeshow Test:

The **Hosmer Lemeshow test** is a statistical test for goodness of fit for logistic regression models. It is used frequently in risk prediction models. The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. The Hosmer Lemeshow test specifically

identifies subgroups as the deciles of fitted risk values. Models for which expected and observed event rates in subgroups are similar are called well calibrated. The Hosmer Lemeshow test can determine if the differences between observed and expected proportions are significant, indicating model lack of fit.

So, our null hypothesis is the model is good fit of Logistic Regression Model and alternative hypothesis is the model is not a good fit of Logistic Regression Model for at least one group

The results of our models of 8 groupings are given below

	Model 1	Model 2	Model 3
Test Statistic Value	4.6804	1.6957	6.9544
Degrees of freedom	6	6	6
p-value	0.5854	0.9455	0.3251

The results of our models of 16 groupings are given below:

	Model 1	Model 2	Model 3
Test Statistic Value	6.7322	5.0577	17.285
Degrees of freedom	14	14	14
p-value	0.9445	0.985	0.2413

The results of our models of 32 groupings are given below

	Model 1	Model 2	Model 3
Test Statistic Value	20.826	6.38	30.362
Degrees of freedom	30	30	30
p-value	0.8933	1	0.4472

As we can see that for Model 1 and Model 2, p value is high which indicates the model is good fit of Logistic Regression. On the other hand, for Model 3, the p value is comparatively low from Model 1 and Model 2 in cases of groupings. So, we exclude Model 3 from our analysis and do prediction for only Model 1 and Model 2

Part III: Prediction

From the previous section applying different kind of goodness of fit test we get two models. Now to get the prediction performances using test dataset, we can use different methods. Some of them are mentioned here: **a)** Confusion Matrix **b)** Classification Error. **c)** Roc Curve and AUC measure

a) Confusion Matrix:

Since our data has categorical dependent variable, we will do our several analyses with the help of Confusion Matrix. The results of our Confusion Matrix will be given but before that we will give some brief information what is Confusion Matrix and what type of results it gave us and its importance.

A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

Let's start with an example confusion matrix for a binary classifier (though it can easily be extended to the case of more than two classes):

Example confusion matrix for a binary classifier:

	Predicted 'NO'	Predicted 'YES'
Actual 'NO'	50	10
Actual 'YES'	5	100

- *There are two possible predicted classes: "yes" and "no". If we were predicting the presence of a disease, for example, "yes" would mean they have the disease, and "no" would mean they don't have the disease.*
- *The classifier made a total of 165 predictions (e.g., 165 patients were being tested for the presence of that disease).*
- *Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times.*
- *In reality, 105 patients in the sample have the disease, and 60 patients do not.*

- ***True Positives (TP):*** *These are cases in which we predicted yes (they have the disease), and they do have the disease. (i.e., in our problem TP=100)*
- ***True Negatives (TN):*** *We predicted no, and they don't have the disease.*
  (i.e., in our problem TN=50)
- ***False Positives (FP):*** *We predicted yes, but they don't actually have the disease. (Also known as a "Type I error."). (i.e., in our problem FP=10)*
- ***False Negatives (FN):*** *We predicted no, but they actually do have the disease. (Also known as a "Type II error."). (i.e., in our problem FN=100)*
- **Accuracy***: How often it predicts correct values (TP+TN)/total = (100+50)/165 = 0.9*
- **Sensitivity:** Number of positive values are correctly predicted (TP/P) = (100/105) = 0.95
- **Specificity:** Number of negative values are correctly predicted (TN/N) = (50/60) =0.83
- **Balanced Accuracy:** *This value can be used for evaluating how good a binary classifier is. Especially when the classes are imbalanced i.e., one of two classes appears a lot more often than other.*
  The Balanced accuracy = (Sensitivity + Specificity)/2

Here we give results of Confusion Matrix of our Model 1 and Model 2 respectively.

Model 1:

	Prediction	
Actual Value	NO	YES
NO	26	4
YES	19	144

Some necessary information of this Confusion matrix is given below

Accuracy	Sensitivity	Specificity	Balanced Accuracy
0.8808	0.9730	0.5778	0.7754

Model 2:

	Prediction	
Actual Value	NO	YES
NO	25	4
YES	20	144

Some necessary information of this Confusion matrix is given below

Accuracy	Sensitivity	Specificity	Balanced Accuracy
0.8756	0.9730	0.5556	0.7643

Here, we see that Model 1 has better accuracy whereas Model 2 has better balanced accuracy.

b) Classification Error:

The classification error E_i depends on the number of samples incorrectly classified during prediction (false positive and false negative) and evaluated by the formula

$$E_i=(f/n) *100$$

Where f is the number of sample cases incorrectly classified, and n is the total sample. Here the model which have lower Classification Error has better fitting. We are putting results of Model 1 and Model 2 and then compare and decide which Model is better one

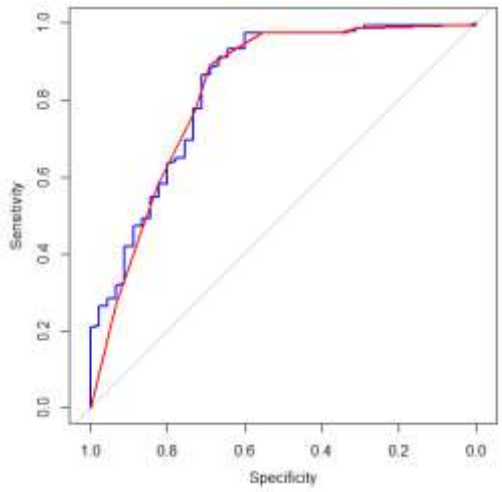
	Model 1	Model 2
Classification Error	0.1243523	0.119171

As we can see that Model 1 has lower classification error than Model 2. So, with the help of Classification error, we can say that Model 1 is better model than Model 2

c) ROC Curve:

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. AUC (Area Under Curve) is a measure of area under the curve. Here, we give graphs of ROC Curve of two different models and later we compare it on the same graph

Figure 16: ROC Curve for our two different models and plotting them in same graph



	Model 1 (Blue)	Model 2 (Red)
AUC	0.832	0.825

Here, we see that there is no that much of difference but Model 1 is still better than Model 2 as Model 1 has better AUC value. On the other hand, we see that the curve Model 1 is kind of Step Lines where Model 2 is a continuous curve. This is due to Model 1 has no continuous independent variable but Model 2 continuous independent variable (Loan Amount)

4)Summary and Discussion

Every banking system has some way to identify suitable candidates for approval of credit cards among many applications. Banks approve credit cards to the corresponding applicant based on different parameters which will have significant importance to validate the credibility of application. Banks collect different information when applicants apply for credit cards. Based on those information decisions are made about the approval of credit card to the corresponding applicants. Here in this project, our goal is to determine the information that are important to decide whether the applicant can get approval of credit card or not. This project is carried out to select such information on which bank should priorities while approving the credit card.

In analysis, at first missing data is checked and removed. After those various graphical representations are implemented to understand various inherent features in the data. These statistical plots reveal different types of association and trend of various features present in the data. It can be observed that there is no interdependency between the continuous variables. Moreover, some observations can be made from the graphs that applicants credit history status impact current loan approval significantly. We find the best logistic regression model for approving status of credit card based on available data in terms of AIC values and other statistical hypothesis testing. From this analysis, it can be noticed that 'marital status', 'loan amount' are the important information to decide whether an applicant can get credit card or not. Thereafter we checked that whether our fit is actually good or not. Later best two models are chosen and predicted with those two models and at the end, we compare their AUC (area under curve) value (from ROC curve), classification error, accuracy and decided that prediction power of both the models are same.

At the end, we can say that we chose are quite good. Based on our test data we have checked that there are 88% and 87.7% chance that these two models can correctly classify whether some applicant's loan status will be approved or not. From this analysis we can also see that there are other variables playing more important role than 'income' to determine applicant's credit card approval status. As from our finding, income is not very important factor to approve credit card to someone.

5) Acknowledgement

It has been a great opportunity to gain lots of experience in Logistic Regression followed by machine learning. We would like to express our special thanks of gratitude to our project supervisor **Dr. Kiranmoy Chatterjee** (Assistant Professor, Dept. of Statistics, Bidhannagar College, Kolkata) *who gave the golden opportunity to do this wonderful project.*

5) References

- https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
- https://www.statmethods.net/advstats/glm.html
- https://datasciencetexts.com/subjects/logistic_regression.html
- https://www.routledge.com/Logistic-Regression-Models/Hilbe/p/book/9781138106710