# Project: Lending Club Data Analysis

By Manish Thapa

Jan 10, 2024

**Abstract**

In the following paper, we apply data analytic methods to predict loan status for the borrowers and lenders. Loans taken by the borrowers can be fully paid, default or charged off and these situations depend on hundreds of factors. In this paper we will primarily deal with few of those predictor variables and try to make prediction for loan status.

## 1 Introduction

Typically, commercial and non-commercial banks are the main lenders for loans. In contrast of those, Lending Club (LC) is a peer-to-peer online lending platform. It is the world's largest marketplace connecting borrowers and investors, where consumers and small business owners lower the cost of their credit and enjoy a better experience than traditional bank lending, and investors earn attractive risk-adjusted returns [1].

## 2 Methods

### 2.1 Data Source

We used Lending Club's data for this analysis. The data set is for the period from 2007 to 2011. There are more than 42000 observations and more than 100 variables. It is often very hard to run analysis with all the variables and observations. So, we cleaned this data and chose ten continuous variable and two categorical variables as our predictor variables and then we deleted other variables.

### 2.2 Measures

#### 2.2.1 Dependent Variable

Loan status is the dependent variable for our analysis which is a categorical variable that takes only two values: Fully paid or Charged Off. We recoded this variable as 1 for Fully paid and 0 for Charged Off.

#### 2.2.2A Independent Continuous Variables

#### 2.2.2A.1 Loan Amount

Loan amount is one of the most important continuous variables. It is denoted as loan_amnt in the data set. Lending Club enables borrowers to create unsecured personal loans between $1,000 and $40,000[4]. This is the listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value [2].

#### 2.2.2A.2 Funded Amount by the Borrowers

This is a continuous variable and denoted as funded_amnt_inv. This represents the total amount committed by investors for that loan at that point in time.

#### 2.2.2A.3 Interest Rate

Interest Rate on the loan denoted as int_rate in the data set. Normally, loans become more lucrative when interest rate is lower. So, we will be investigating whether loan status charged off situation has any association with higher interest rate or nor.

#### 2.2.2A.4 Installment Amount

The monthly payment owed by the borrower if the loan originates. This variable is denoted as installment in the data set.

#### 2.2.2A.5 Annual Income

This continuous variable is denoted as annual_inc. It is the annual income provided by the borrower during registration.

#### 2.2.2A.6 Debt to Income Ratio

This variable is denoted as dti in the data set, a ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.

#### 2.2.2A.7 Total Credit Revolving Balance

Total credit revolving balance denoted as. In credit card terms, a revolving balance is the portion of credit card spending that goes unpaid at the end of a billing cycle. The amount can vary, going up or down depending on the amount borrowed and the amount repaid. The amount of the charge for revolving a balance will depend on the size of the balance and the interest rate of the card. When the balance is paid off, the customer is no longer revolving the debt [3].

#### 2.2.2A.8 Number of Total Accounts

The total number of credit lines currently in the borrower's credit file. This variable is denoted as total_acc.

#### 2.2.2A.9 Total Payments by the Investors

Payments received to date for total amount funded. This variable is denoted as total_pymnt in the data set.

#### 2.2.2A.10 Last Payment Amount

Last total payment amount received when the loan status being fully paid or charged off. This variable is denoted as last_pymnt_amnt in the data set.

### 2.2.2B Independent Categorical Variables

#### 2.2.2B.1 Term

Term variable represents the duration of the loan. This categorical value takes only two values: 36 months and 60 months which were recoded as 1 and 0 respectively.

#### 2.2.2B.2 Home Ownership

Home ownership represents the residential situation of the borrowers. This categorical variable takes four values: Rent, Own, Mortgage, and Other which were recoded as 1, 2, 3, and 4 respectively.

## 3 Results

### 3.1 Descriptive Statistics

In a sample of 42536 observations the minimum loan amount was $500.00, and the maximum loan amount was $35000.00 with a mean of $11837.43 and standard deviation 7972.01 in the case where the loan was Charged Off. In contrast, the minimum and maximum loan amount was same in case where the loan was Fully Paid but the mean and standard deviation of the loan amount in this status was lower compared to the Charged Off status.

The minimum and maximum funded amounts by the investors are same in both charged off status and fully paid status. However, the mean and the standard deviation of the funded amount is higher for charged off status ($10307.15 & 7636.38) compared to fully paid status ($10110.15 & 7037.58)

The average interest rate seems higher for charged off status, 14% versus 12%.

All primary statistics such as mean, standard deviation, minimum, and maximum for installment variable was higher for charged off status versus fully paid status.

The mean annual income for fully paid status is $70164.39 with standard deviation 66283.92 whereas the mean annual income for charged off status was $63366.87 with standard deviation 49684.10. The same rule applies for maximum limit. Fully paid status group has the higher value for annual income ($6000000.00) compared to charged off status ($1250000.00).

The mean debt to income ration (dti) is higher for charged off status (14.05) compared to fully paid status (13.25). This is reasonable because, fully paid group has higher average income.

The mean and standard deviation of total credit revolving balance for the charged off status are $15194.46 and 27942.81 respectively which is higher compared to the mean and standard deviation of the fully paid status ( $14138.15 and 20783.18)

The mean and standard deviation of the total payments by the investors ($12239.24 & 9102.21) are higher for the fully paid status compared to the charged off status ($6112.44 & 6585.14)

More precise descriptive statistics of the continuous variables have been represented by the table below.

**Table1: Summary Statistics of the continuous predictive variables grouped by Loan Status**

| Loan Status | N Obs | Variable | Label | Mean | Std Dev | Minimum | Maximum | N | N Miss |
|---|---|---|---|---|---|---|---|---|---|
| Charged Off | 6431 | loan_amnt | Loan Amount | 11837.43 | 7971.07 | 500.00 | 35000.00 | 6431 | 0 |
| | | funded_amnt_inv | Funded Amount by the Investors | 10307.15 | 7636.38 | 0.00 | 35000.00 | 6431 | 0 |
| | | int_rate | Interest Rate | 0.14 | 0.04 | 0.05 | 0.24 | 6431 | 0 |
| | | installment | Installment Amount | 333.07 | 216.95 | 15.91 | 1305.19 | 6431 | 0 |
| | | annual_inc | Annual Income | 63366.87 | 49684.10 | 2000.00 | 1250000.00 | 6431 | 0 |
| | | dti | Debt to Income Ratio | 14.05 | 6.66 | 0.00 | 29.96 | 6431 | 0 |
| | | revol_bal | Total credit revolving balance | 15194.46 | 27942.81 | 0.00 | 1207359.00 | 6431 | 0 |
| | | total_acc | Number of total accounts | 21.54 | 11.80 | 1.00 | 74.00 | 6428 | 3 |
| | | total_pymnt_inv | Total Payments by the Investors | 6112.44 | 6585.14 | 0.00 | 55836.73 | 6431 | 0 |
| | | last_pymnt_amnt | Last Payment Amount | 324.31 | 522.43 | 0.00 | 12818.38 | 6431 | 0 |
| Fully Paid | 36104 | loan_amnt | Loan Amount | 10956.54 | 7298.75 | 500.00 | 35000.00 | 36104 | 0 |
| | | funded_amnt_inv | Funded Amount by the Investors | 10110.15 | 7037.58 | 0.00 | 35000.00 | 36104 | 0 |
| | | int_rate | Interest Rate | 0.12 | 0.04 | 0.05 | 0.25 | 36104 | 0 |
| | | installment | Installment Amount | 320.76 | 207.41 | 15.67 | 1295.21 | 36104 | 0 |
| | | annual_inc | Annual Income | 70164.39 | 66283.92 | 1896.00 | 6000000.00 | 36100 | 4 |
| | | dti | Debt to Income Ratio | 13.25 | 6.73 | 0.00 | 29.99 | 36104 | 0 |
| | | revol_bal | Total credit revolving balance | 14138.15 | 20783.18 | 0.00 | 952013.00 | 36104 | 0 |
| | | total_acc | Number of total accounts | 22.23 | 11.55 | 1.00 | 90.00 | 36078 | 26 |
| | | total_pymnt_inv | Total Payments by the Investors | 12239.24 | 9102.21 | 0.00 | 58563.68 | 36104 | 0 |
| | | last_pymnt_amnt | Last Payment Amount | 3020.96 | 4637.45 | 0.00 | 36115.20 | 36104 | 0 |

About 74.14% of loans were 36 months term of which 9.11% were charged off and 65.02% were fully paid and about 25.86% of the loans were of 60 months term of which 6.01% were charged off and 19.86% were fully paid. This information have been presented by table 2 and Figure 1.

| Table of loan_status by term | | | |
|---|---|---|---|
| | term(Term of the Loan payment) | | |
| loan_status(Loan Status) | 36 months | 60 months | Total |
| Charged Off | 3876<br>9.11 | 2555<br>6.01 | 6431<br>15.12 |
| Fully Paid | 27658<br>65.02 | 8446<br>19.86 | 36104<br>84.88 |
| Total | 31534<br>74.14 | 11001<br>25.86 | 42535<br>100.00 |

Table 2: Loan status by terms

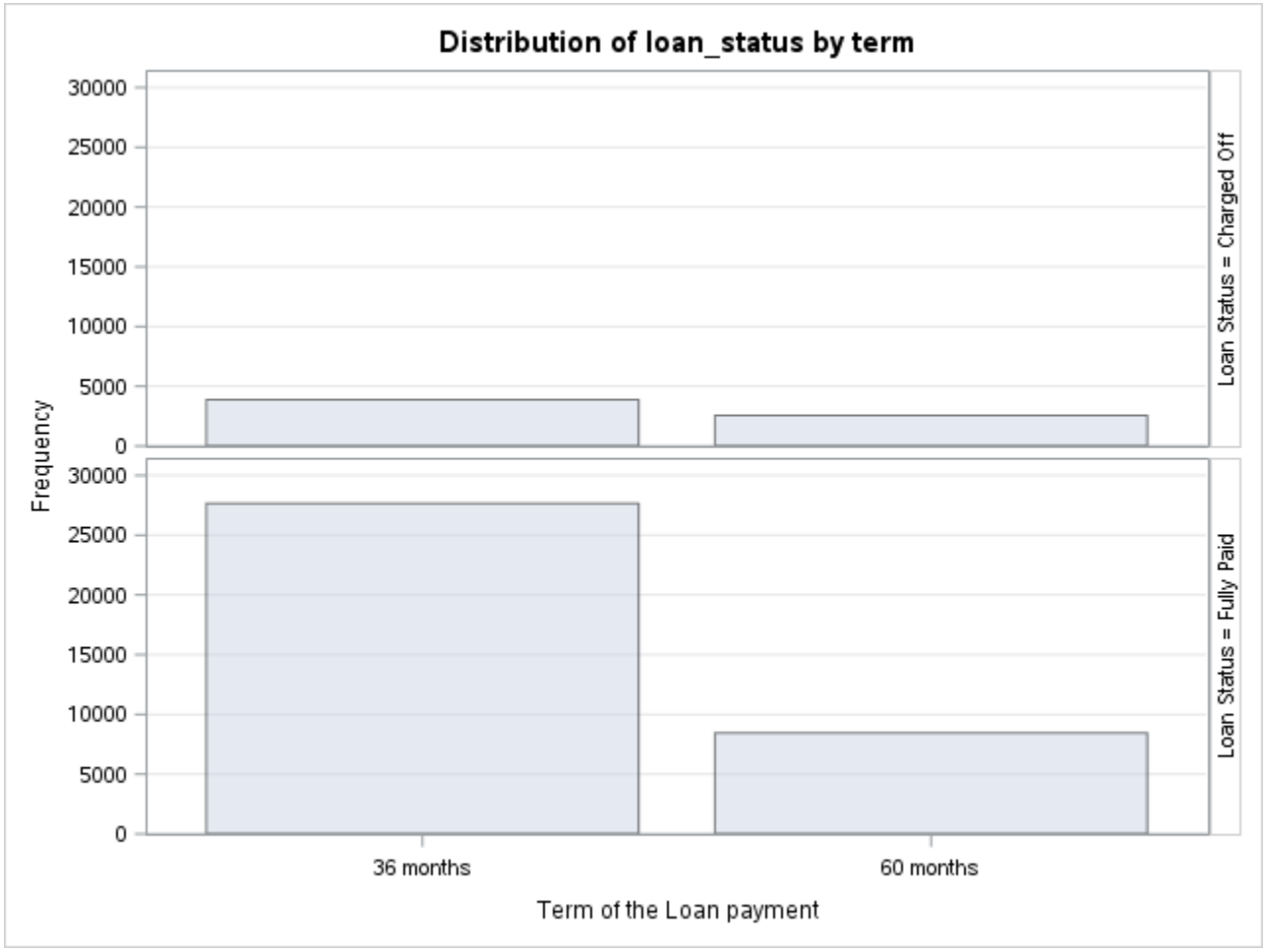


Figure 1: Loan status by terms

The highest proportion of the borrowers (47.45%) used live in rented housing of which 7.54% were charged off and 39.91% were fully paid. About 44.57% of the borrowers have had mortgage for home of which 6.35% where charged off and 38.23% fully paid. This information have been presented by the following table and figure.

| Table of loan_status by home_ownership | | | | | | |
|---|---|---|---|---|---|---|
| | home_ownership(Home Ownership) | | | | | |
| loan_status(Loan Status) | MORTGAGE | NONE | OTHER | OWN | RENT | Total |
| Charged Off | 2699<br>6.35 | 1<br>0.00 | 29<br>0.07 | 495<br>1.16 | 3207<br>7.54 | 6431<br>15.12 |
| Fully Paid | 16260<br>38.23 | 7<br>0.02 | 107<br>0.25 | 2756<br>6.48 | 16974<br>39.91 | 36104<br>84.88 |
| Total | 18959<br>44.57 | 8<br>0.02 | 136<br>0.32 | 3251<br>7.64 | 20181<br>47.45 | 42535<br>100.00 |

Table 3: Loan status by home ownership.

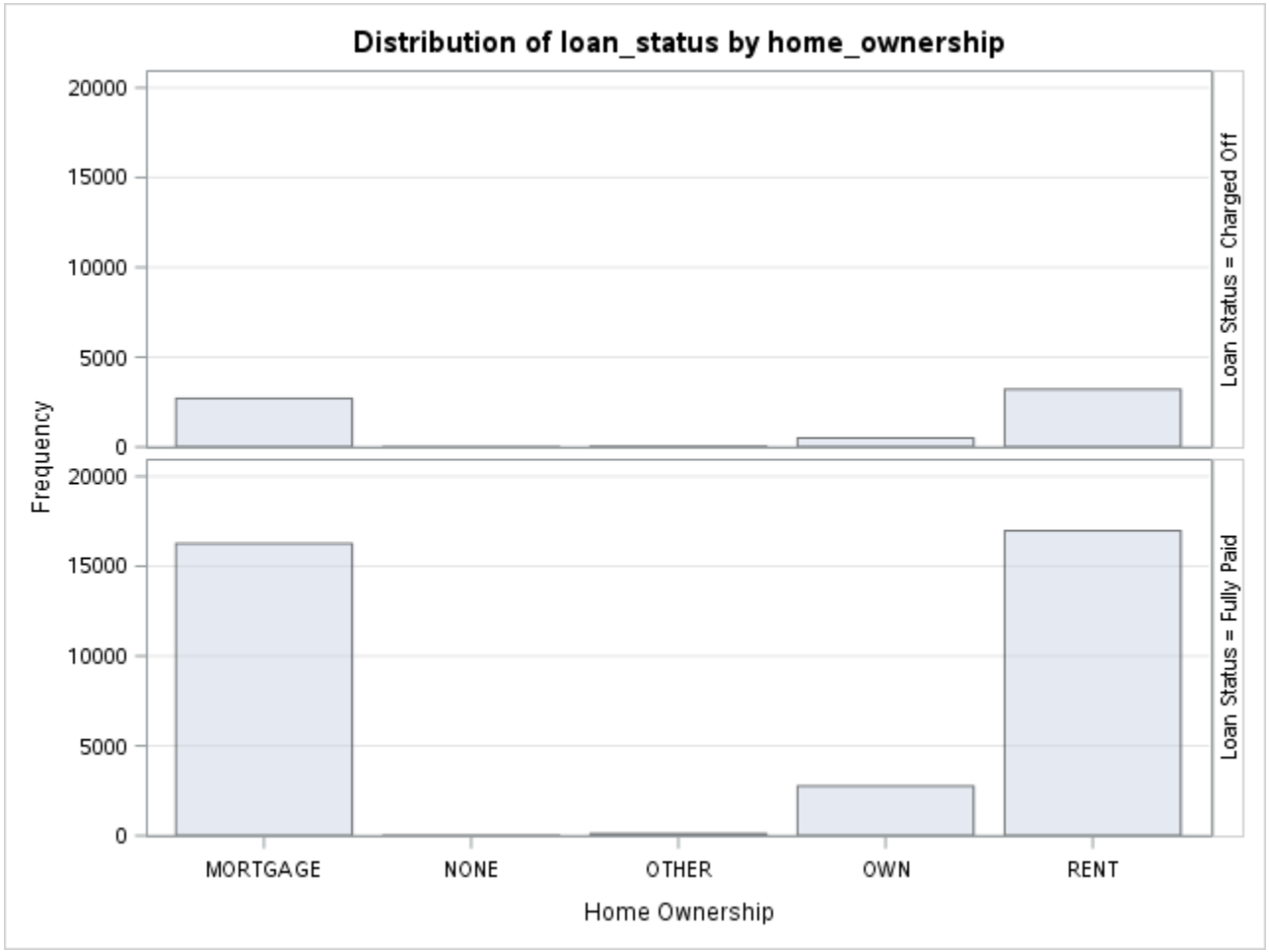


Figure 2: Loan status by home ownership.

## 3.2 Association of the predictive variables with the response variable

To check how Categorical predictors, associate with the response variable (loan_status) we ran a cross tabulation for analysis.

| Frequency<br>Expected<br>Cell Chi-Square<br>Row Pct<br>Col Pct | Table of loan_status by Term_in_months | | |
|---|---|---|---|
| | Term_in_months(Term_in_months) | | |
| loan_status(loan_status) | 36 | 60 | Total |
| Charged Off | 3876<br>4767.7<br>166.78<br>60.27<br>12.29 | 2555<br>1663.3<br>478.08<br>39.73<br>23.23 | 6431 |
| Fully Paid | 27658<br>26766<br>29.708<br>76.61<br>87.71 | 8446<br>9337.7<br>85.157<br>23.39<br>76.77 | 36104 |
| Total | 31534 | 11001 | 42535 |

Table 4

The above table is a crosstabulation of loan status by term in months. From the Pearson chi-square test of association, it is seen in the cell of Charged off at 60 months contributes the most measure of Chi-square statistic of value 478.08.

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 759.7234 | <.0001 |
| Likelihood Ratio Chi-Square | 1 | 705.7527 | <.0001 |
| Continuity Adj. Chi-Square | 1 | 758.8716 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 759.7055 | <.0001 |
| Phi Coefficient | | -0.1336 | |
| Contingency Coefficient | | 0.1325 | |
| Cramer's V | | -0.1336 | |

Table 5

Further analysis is observed in Table 5 of chi-square where the p-value less than 0.0001. With this result we reject the null hypothesis. The Cramer’s V value (-01336) also indicate that the association between Loan status and the term in months is weak.

| Frequency<br>Expected<br>Cell Chi-Square<br>Row Pct<br>Col Pct | Table of loan_status by home_ownership | | | | | |
|---|---|---|---|---|---|---|
| | home_ownership(home_ownership) | | | | | |
| loan_status(loan_status) | MORTGAGE | NONE | OTHER | OWN | RENT | Total |
| Charged Off | 2699<br>2866.5<br>9.7843<br>41.97<br>14.24 | 1<br>1.2095<br>0.0363<br>0.02<br>12.50 | 29<br>20.562<br>3.4624<br>0.45<br>21.32 | 495<br>491.53<br>0.0245<br>7.70<br>15.23 | 3207<br>3051.2<br>7.9524<br>49.87<br>15.89 | 6431 |
| Fully Paid | 16260<br>16093<br>1.7428<br>45.04<br>85.76 | 7<br>6.7905<br>0.0065<br>0.02<br>87.50 | 107<br>115.44<br>0.6167<br>0.30<br>78.68 | 2756<br>2759.5<br>0.0044<br>7.63<br>84.77 | 16974<br>17130<br>1.4165<br>47.01<br>84.11 | 36104 |
| Total | 18959 | 8 | 136 | 3251 | 20181 | 42535 |

Table 6

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 4 | 25.0469 | <.0001 |
| Likelihood Ratio Chi-Square | 4 | 24.7320 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 20.7080 | <.0001 |
| Phi Coefficient | | 0.0243 | |
| Contingency Coefficient | | 0.0243 | |
| Cramer's V | | 0.0243 | |

Table 7

A crosstabulation table for the home_onwership with loan status has been done in Table 6, the cell of Mortgage with Charged off shows the highest Chi-square value of 9.7843 relative to other cells. Table 7 support the assumption that there exist association between the variable since the p-value is less than significant level hence rejects null hypothesis. The Cramer's value (0.0243) shows a positive weak association.

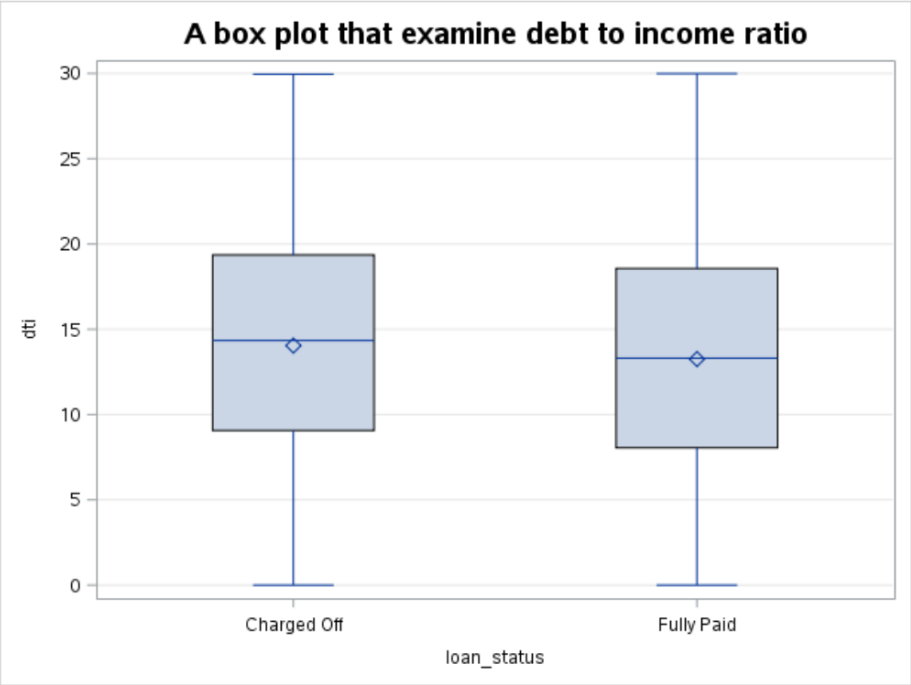


Figure 3: Loan status with dti

The Figure 3 above shows Loan status relate to debt income ratio, the 25$^{th}$ and 75$^{th}$ percentiles ranges from 8 and 19.5 and no outliers observed. The Fully Paid loan has a mean a little lower than Charged Off, this means that the higher the debt to income ratio the higher the probability of the loan being Charged Off.

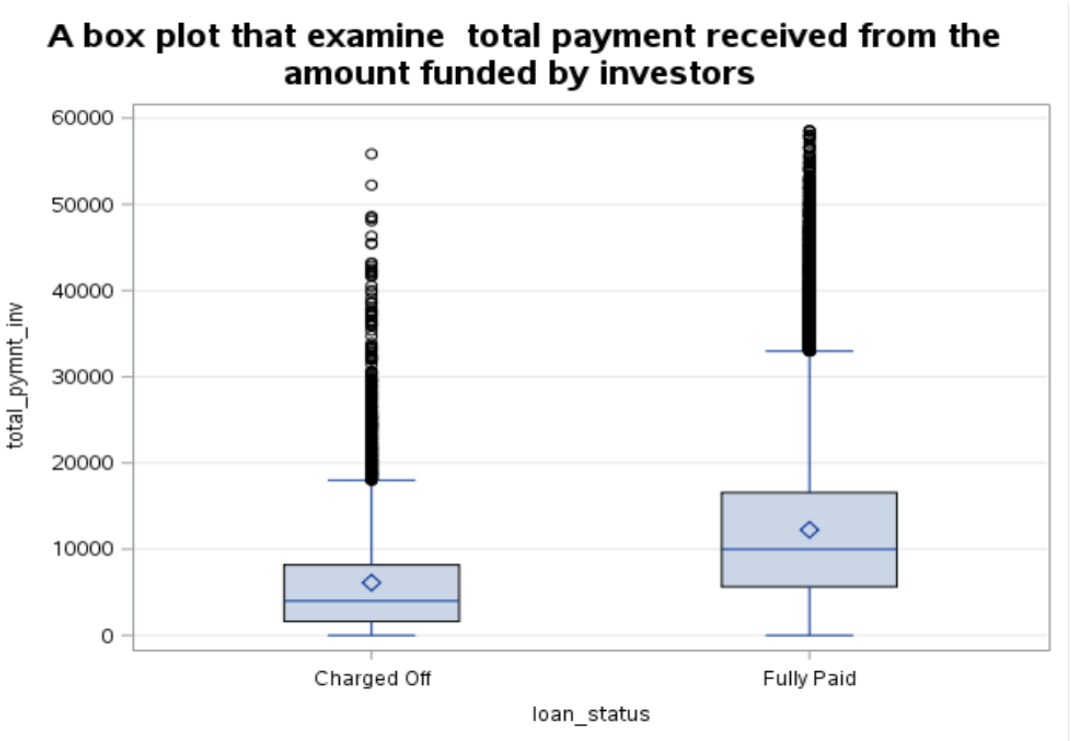


Figure 4: Loan status with total payment by the investors

The Charged Off boxplot lies as low as 0 and as high as $19000 with outliers as far as $55000. Most of the Charged off amount is about $5000. On the other hand, Fully Paid is between 0 and $35000 also with outlier of $60000. Most Fully Paid loan have total amount received of about $15000. This indicates that most of the total amount received was Fully Paid.

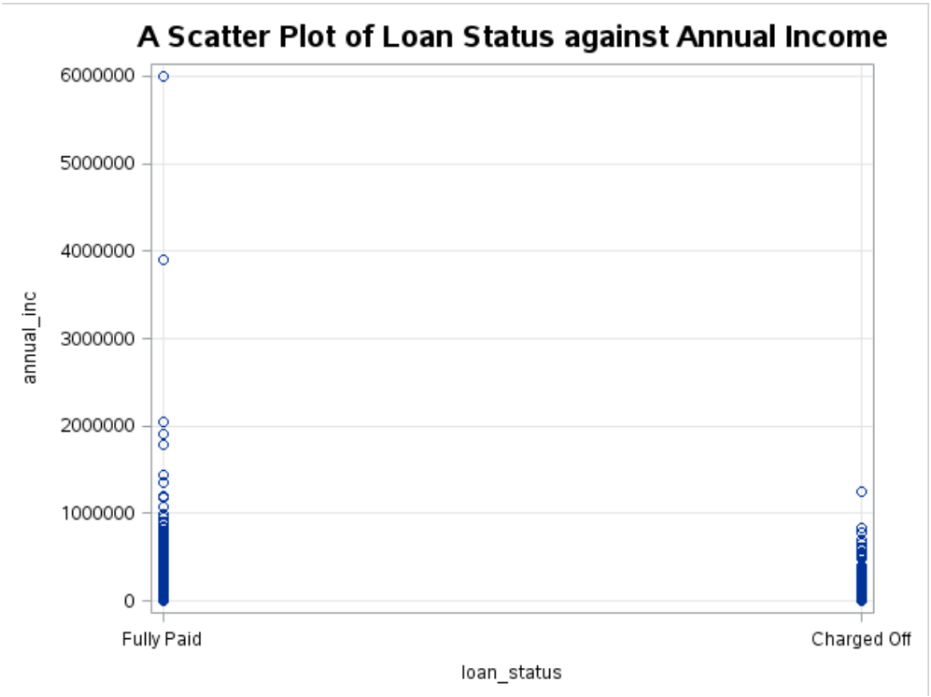


Figure 5: Loan Status with annual income

The figure above shows association between loan status and the annual income, the higher the annual income the higher the likelihood of the loan being fully paid. However, if the annual income is low, there is chances of the loan to be Charged Off.

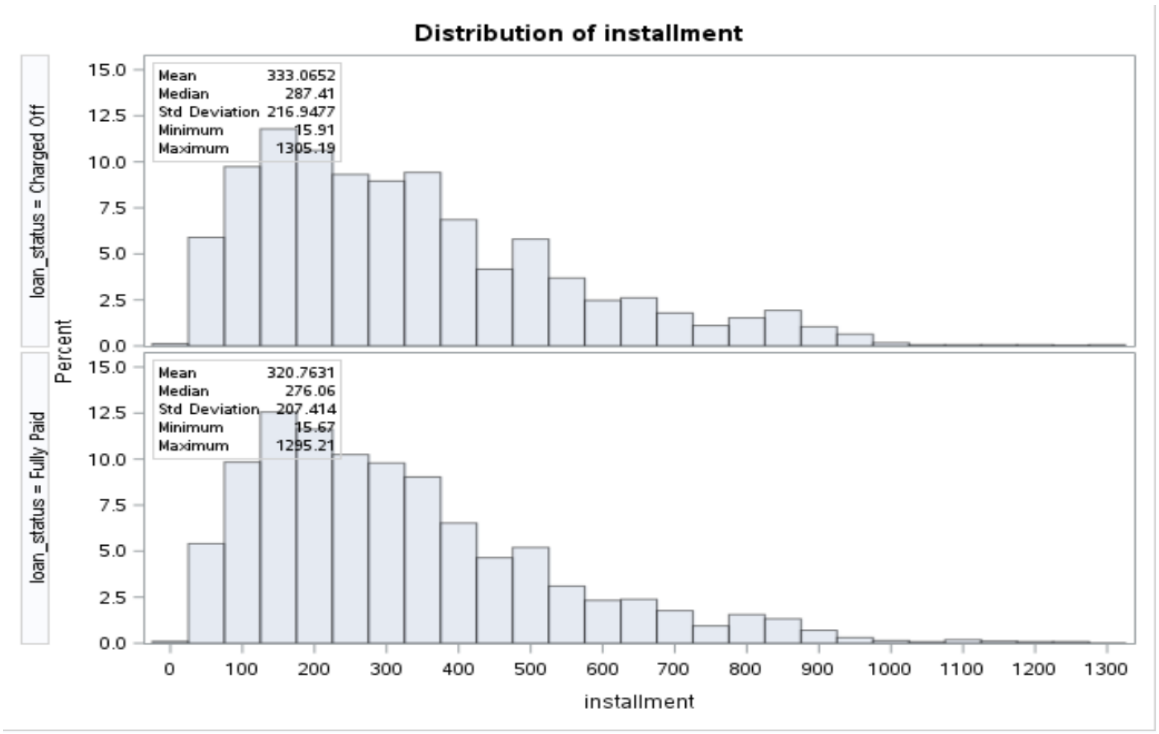


Figure 6:

The distribution of installment there certainly is weak association between Bonus and installment. The less the amount of installment, the more likely the loan is to be charged off.

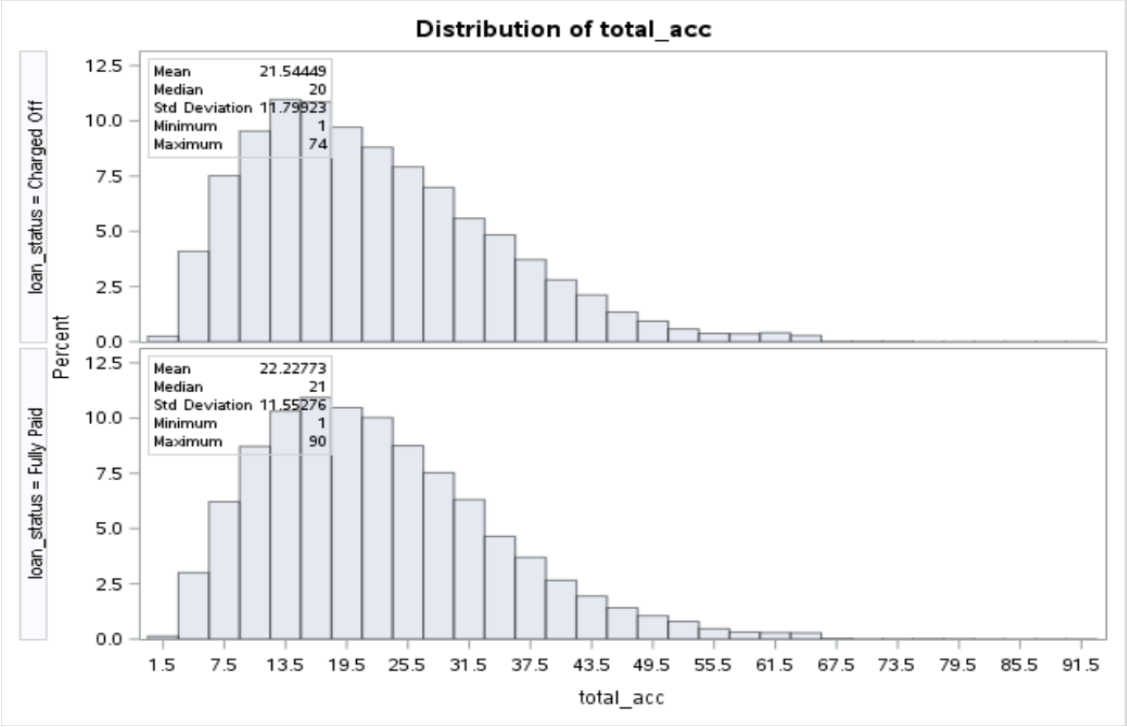


Figure 7

The distribution of total number of accounts show very minimal association with loan status, this means that the total number of accounts do not have significant effects on loan status.

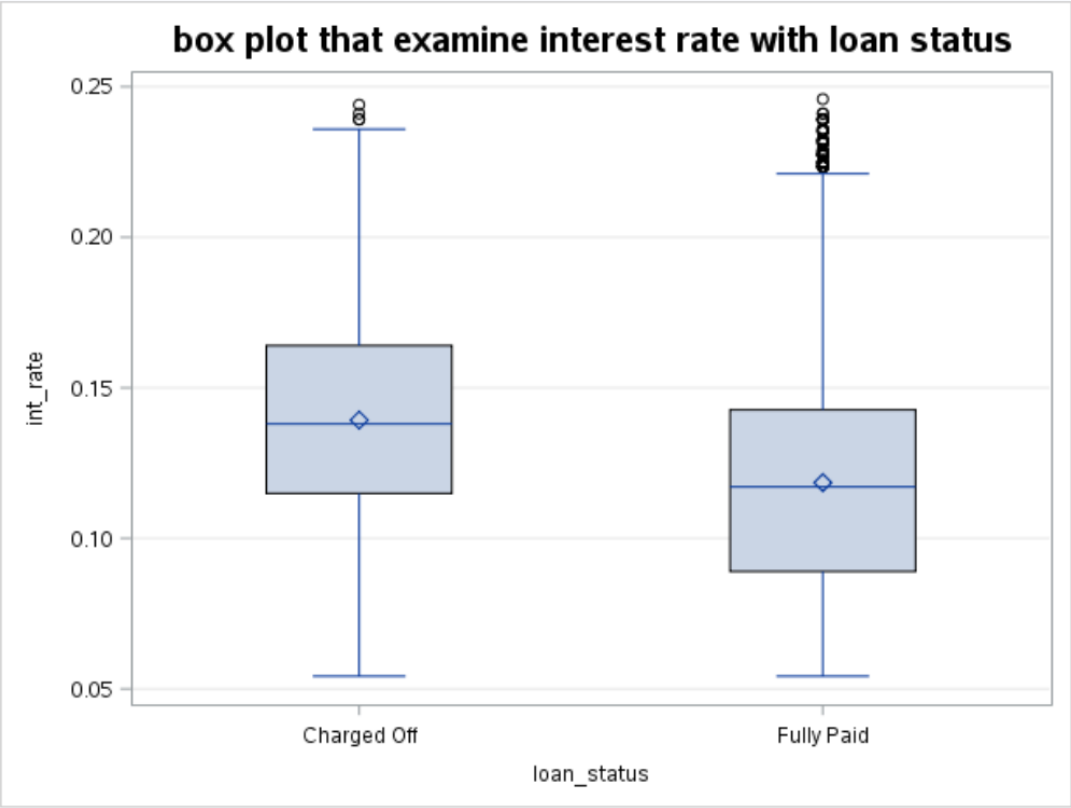


Figure 8

This box plot describes the upper and the lower quartile of loan status to range between 0.7 and 0.17. there is a correlation between the two variables in that most charged off loan has an interest rate of 0.13 or 13%. This indicates that the higher the rate the more the charged off loans.

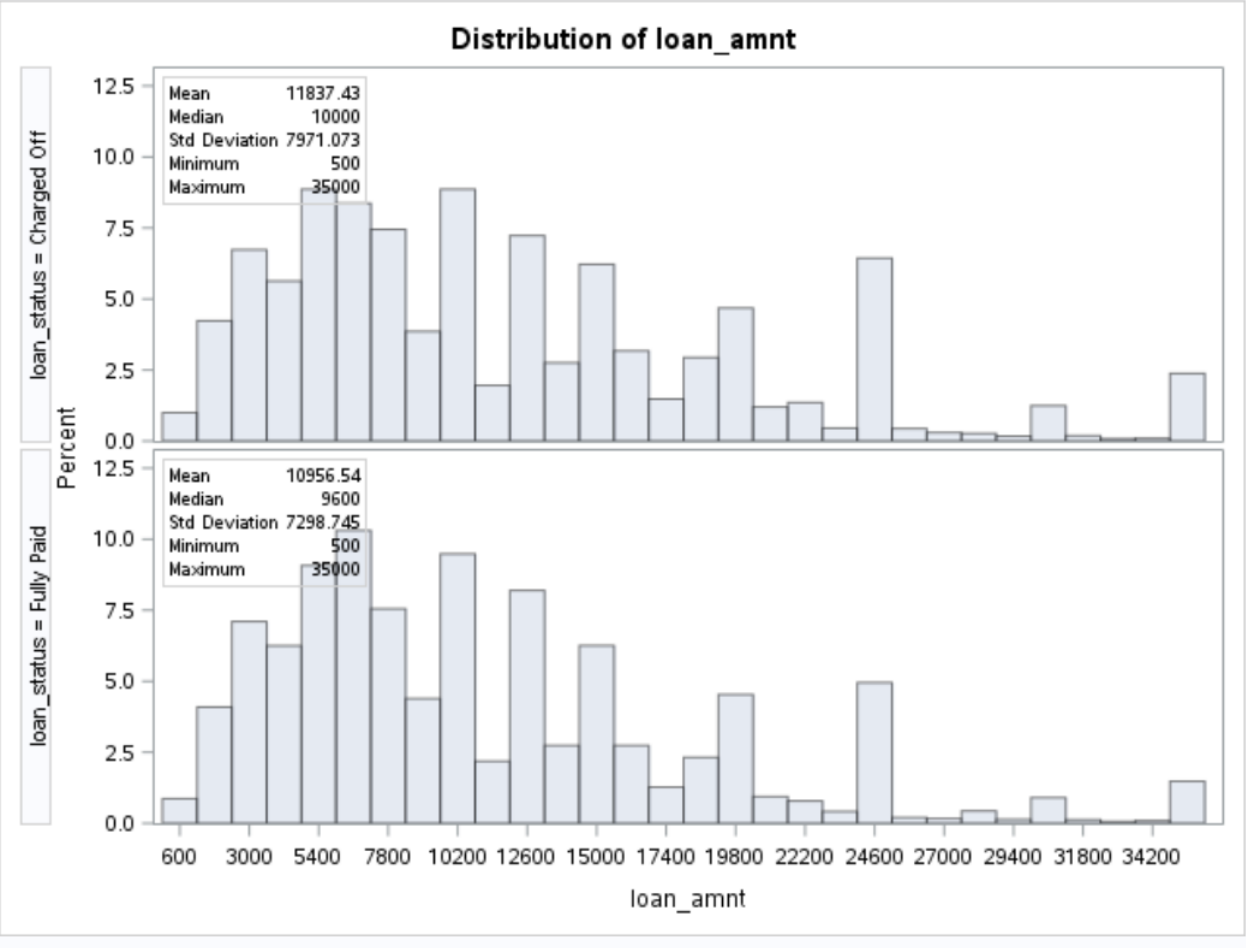


Figure 9

Distribution of loan amount certainly show an association with the loan status. The Fully Paid loan has lower mean value, lower standard deviation and lower median of 400 units less than Charged off. This shows that there is a relationship between loan status and loan amount. More of the loan amount is charged off than the fully paid.

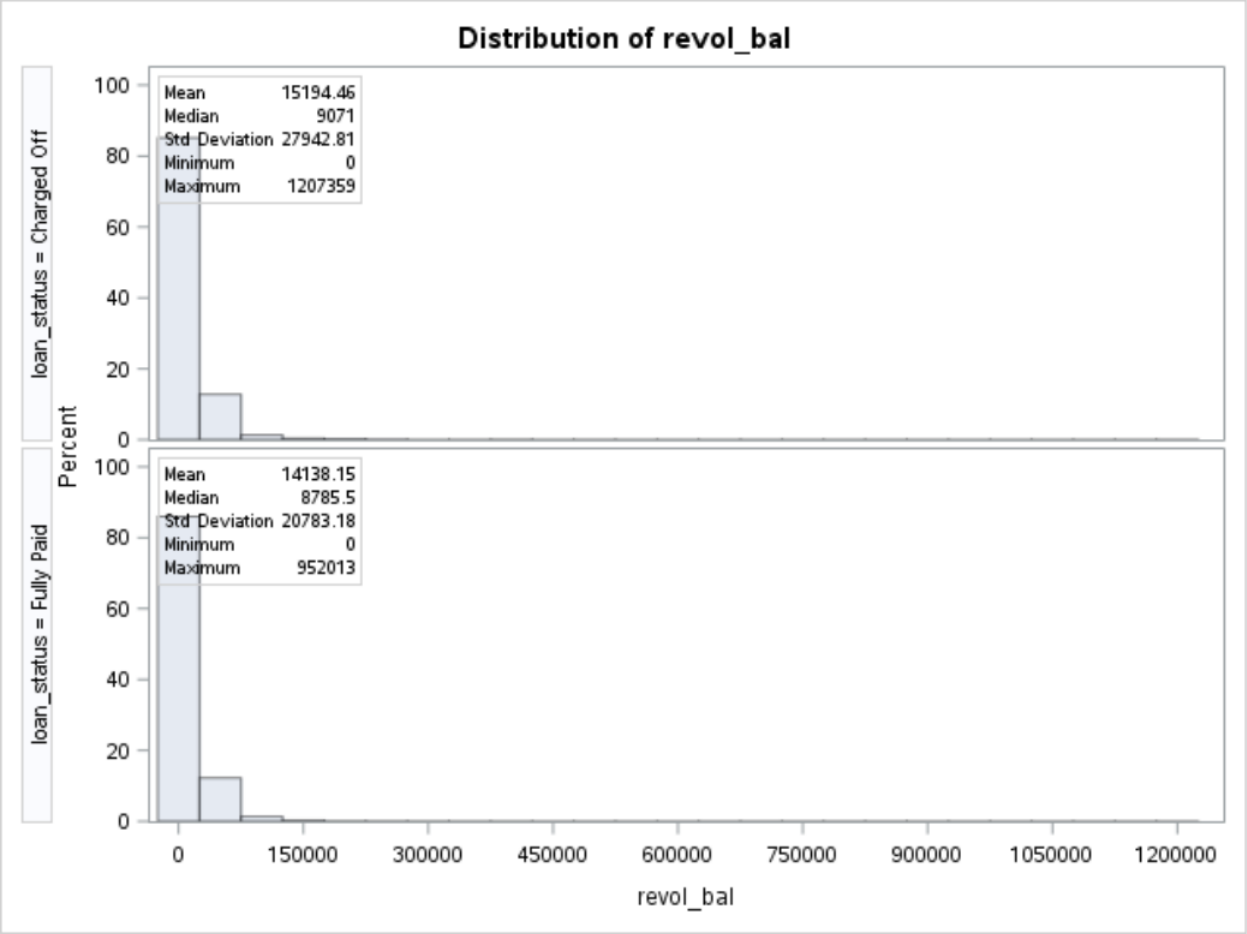


Figure 10

Distribution of revolving balance indicates a high standard deviation in charged off loan than in fully paid

There certainly appears to be an association between loan status and revolving balance. The larger the, revolving balance the more likely the loan is to be Charged Off. The median of charged off loan is almost larger 300 thousand more than Fully paid loan.

## 3.3 Bivariate Logistic Regression with training data

Initially we started with 12 predictor variables. To avoid complexity in the model we applied main effect bivariate logistic regression analysis with the adaptation of backward elimination method we get the following results. SAS studio has been used to generate the following results

| Summary of Backward Elimination | | | | | | |
|---|---|---|---|---|---|---|
| Step | Effect Removed | DF | Number In | Wald Chi-Square | Pr > ChiSq | Variable Label |
| 1 | Home_Ownership | 3 | 11 | 0.7231 | 0.8678 | |
| 2 | revol_bal | 1 | 10 | 0.9661 | 0.3257 | Revolving Balance |
| 3 | annual_inc | 1 | 9 | 1.4900 | 0.2222 | Annual Income |

Table 8: Insignificant variables for the model

Backward elimination method removed the effects home ownership, revolving balance, and annual income. So, these factors are not significant to calculate the probability of fully paid or charged off status for a customer. Table 8 lists those insignificant factors.

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| dti | 1 | 17.9452 | <.0001 |
| funded_amnt_inv | 1 | 1979.4726 | <.0001 |
| installment | 1 | 21.1960 | <.0001 |
| int_rate | 1 | 648.1114 | <.0001 |
| last_pymnt_amnt | 1 | 398.8428 | <.0001 |
| loan_amnt | 1 | 89.5191 | <.0001 |
| Term | 1 | 210.0046 | <.0001 |
| total_acc | 1 | 8.4520 | 0.0036 |
| total_pymnt_inv | 1 | 2178.7252 | <.0001 |

Table 9: List of significant factors for the model

The backward elimination method also provided the significant factors listed in table 9. We see that *Debt to Income Ratio, Funded Amount by the Investors, Installment Amount, Interest Rate, Last Payment Amount, Loan Amount, Term, Total Number of Accounts and Total Payment by the Investors* are the significant factors for the model to calculate the probability of the loan status.

Using the parameter estimates from table 10, we can construct a model to calculate the probability of a loan being fully paid.

$$\hat{p} = \frac{e^{5.8163-0.0222x_1-0.00236x_2+0.00320x_3-30.23x_4+0.00119x_5-0.00020x_6-1.75x_7+0.00976x_8+0.00220x_9}}{1+e^{5.8163-0.0222x_1-0.00236x_2+0.00320x_3-30.23x_4+0.00119x_5-0.00020x_6-1.75x_7+0.00976x_8+0.00220x_9}}$$

$$Where\ x_1 = Debt\ to\ Income\ Ratio,\quad x_2 = Funded\ Amount\ by\ Investors,\quad x_3 = Installment\ Amount$$

$$x_4 = Interest\ Rate\ ,\qquad x_5 = Last\ payment\ Amount,\qquad x_6 = Loan\ Amount,\qquad x_7 = \begin{cases} 0\ if\ term = 36\ months \\ 1\ if\ term = 60\ months \end{cases}$$

$$x_8 = Number\ of\ total\ accounts, \qquad x_9 = Total\ Pyament\ by\ the\ Investors$$

| Analysis of Maximum Likelihood Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | | 1 | 5.8163 | 0.1739 | 1118.3341 | <.0001 |
| dti | | 1 | -0.0222 | 0.00524 | 17.9452 | <.0001 |
| funded_amnt_inv | | 1 | -0.00236 | 0.000053 | 1979.4726 | <.0001 |
| installment | | 1 | 0.00320 | 0.000696 | 21.1960 | <.0001 |
| int_rate | | 1 | -30.2333 | 1.1876 | 648.1114 | <.0001 |
| last_pymnt_amnt | | 1 | 0.00119 | 0.000060 | 398.8428 | <.0001 |
| loan_amnt | | 1 | -0.00020 | 0.000021 | 89.5191 | <.0001 |
| Term | 1 | 1 | -1.7500 | 0.1208 | 210.0046 | <.0001 |
| total_acc | | 1 | 0.00976 | 0.00336 | 8.4520 | 0.0036 |
| total_pymnt_inv | | 1 | 0.00220 | 0.000047 | 2178.7252 | <.0001 |

Table 10: Parameter estimate and p-values of the significant factors

In this analysis we compared 231,909,384 pairs of combinations from response variable and predictor variables of which 97.7 are concordant. Also, the c-value is very close to 1. This information has been illustrated in table 11. Based on this evidence we can say that our model that we found is a good predictive model.

| Association of Predicted Probabilities and Observed Responses | | | |
|---|---|---|---|
| Percent Concordant | 98.0 | Somers' D | 0.959 |
| Percent Discordant | 2.0 | Gamma | 0.959 |
| Percent Tied | 0.0 | Tau-a | 0.247 |
| Pairs | 113776500 | c | 0.980 |

Table 11: Model effectiveness statistics

However, table 12 illustrates the goodness-of-fit of the test. The p-value is very significant even at less than 0.001. Which indicates that the model that we developed is not a good fit for predicting the probability of loan status.

| Hosmer and Lemeshow Goodness-of-Fit Test | | |
|---|---|---|
| Chi-Square | DF | Pr > ChiSq |
| 50099.7060 | 8 | <.0001 |

Table 12: Goodness of the fit of the model

The reasons behind this is the collinearity problem of the predictor variables. The predictor variables that we chose were affected by the collinearity problem that led the model to be unfit for the prediction.

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
| Intercept | Intercept | 1 | 1.10959 | 0.00602 | 184.30 | <.0001 | 0 |
| annual_inc | Annual Income | 1 | 7.351438E-8 | 2.159355E-8 | 3.40 | 0.0007 | 1.24224 |
| installment | Installment Amount | 1 | -0.00002710 | 0.00002587 | -1.05 | 0.2948 | 18.93924 |
| funded_amnt_inv | Funded Amount by the Investors | 1 | -0.00005951 | 6.577524E-7 | -90.47 | <.0001 | 14.26183 |
| int_rate | Interest Rate | 1 | -1.93431 | 0.04172 | -46.36 | <.0001 | 1.54758 |
| total_pymnt_inv | Total Payment by the Investors | 1 | 0.00005931 | 3.419371E-7 | 173.47 | <.0001 | 6.19283 |
| loan_amnt | Loan Amount | 1 | -0.00000862 | 7.29255E-7 | -11.83 | <.0001 | 18.94043 |
| revol_bal | Revolving Balance | 1 | -1.17016E-7 | 6.282475E-8 | -1.86 | 0.0625 | 1.24137 |
| total_acc | Number of Total Account | 1 | 0.00030125 | 0.00012437 | 2.42 | 0.0154 | 1.34796 |
| dti | Debt to Income Ratio | 1 | -0.00076495 | 0.00019999 | -3.83 | 0.0001 | 1.17329 |
| last_pymnt_amnt | Last Payment Amount | 1 | 0.00001717 | 3.226358E-7 | 53.23 | <.0001 | 1.29861 |
| Term | | 1 | -0.11446 | 0.00511 | -22.40 | <.0001 | 3.24717 |
| Home_Ownership | | 1 | -0.00200 | 0.00140 | -1.43 | 0.1517 | 1.18324 |

Table 13: VIF factors

## 3.4 Bivariate Logistic Regression with validation data

If we run the same analysis using the validation data we obtain the following results.

In this case, the backward selection method removed 5 insignificant variables which are given in table 14.

| Summary of Backward Elimination | | | | | | |
|---|---|---|---|---|---|---|
| Step | Effect Removed | DF | Number In | Wald Chi-Square | Pr > ChiSq | Variable Label |
| 1 | revol_bal | 1 | 11 | 0.0113 | 0.9155 | Revolving Balance |
| 2 | installment | 1 | 10 | 0.1449 | 0.7035 | Installment Amount |
| 3 | annual_inc | 1 | 9 | 0.4674 | 0.4942 | Annual Income |
| 4 | Home_Ownership | 3 | 8 | 5.0600 | 0.1675 | |
| 5 | total_acc | 1 | 7 | 1.9604 | 0.1615 | Number of Total Account |

Table 14: Insignificant factors

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| dti | 1 | 6.8035 | 0.0091 |
| funded_amnt_inv | 1 | 827.1147 | <.0001 |
| int_rate | 1 | 260.4085 | <.0001 |
| last_pymnt_amnt | 1 | 130.7098 | <.0001 |
| loan_amnt | 1 | 70.8307 | <.0001 |
| Term | 1 | 150.2239 | <.0001 |
| total_pymnt_inv | 1 | 928.9640 | <.0001 |

Table 15: Significant factors for the model when we select validation data

| Analysis of Maximum Likelihood Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | | 1 | 5.6959 | 0.2438 | 545.9305 | <.0001 |
| dti | | 1 | -0.0197 | 0.00755 | 6.8035 | 0.0091 |
| funded_amnt_inv | | 1 | -0.00226 | 0.000078 | 827.1147 | <.0001 |
| int_rate | | 1 | -27.7258 | 1.7181 | 260.4085 | <.0001 |
| last_pymnt_amnt | | 1 | 0.00125 | 0.000110 | 130.7098 | <.0001 |
| loan_amnt | | 1 | -0.00011 | 0.000013 | 70.8307 | <.0001 |
| Term | 1 | 1 | -1.8830 | 0.1536 | 150.2239 | <.0001 |
| total_pymnt_inv | | 1 | 0.00212 | 0.000069 | 928.9640 | <.0001 |

Table 16: Parameter estimate and p-value of the significant factors for the validation data set.

| Association of Predicted Probabilities and Observed Responses | | | |
|---|---|---|---|
| Percent Concordant | 97.7 | Somers' D | 0.955 |
| Percent Discordant | 2.3 | Gamma | 0.955 |
| Percent Tied | 0.0 | Tau-a | 0.244 |
| Pairs | 20811416 | c | 0.977 |

Table 17: Model effectiveness

## 4 Conclusion

The variables that we considered in this analysis are not good predictors for loan status prediction. Because there is a significant difference in the parameter estimates when we split the original data into training data and validation data. Also, there is collinearity problem among the variables that we selected. So, further investigation is required along with the selection of other significant predictors.