**Introduction**

In a loan, the borrower initially borrows a sum of money(principal) and is obligated to pay an equal amount of money along with interest to the lender at a later time.

Problem statement: When a loan is given out by the lender, it also causes a credit risk. A credit risk is the risk of default on a loan that may arise from a debtor failing to make necessary payments. The risk to the lender comprises of lost principal and interest, disorder in cash flows, and an increase in collection charges.

To determine the credit risk accompanying a loan portfolio, it would be worthwhile to forecast whether a borrower will default. Consequently, actions can be taken to mitigate possible loss to the creditor.

In this project, Probability of Default model has been developed with the help of past credit information about existing borrowers and their behavior with the bank to calculate and measure the risk of each borrower.

Probability of default is a monetary term labeling the likelihood of a default over a certain time horizon. It offers an estimate of the likelihood that a debtor will be unable to meet his/her loan obligations. PD is used in an assortment of credit analyses and risk management background. The probability of default is an estimation of the chances that the default event will happen.

Based on the credit risk the lender can take the following steps for high-risk applicants:

* Lessening credit/services if the risk level is too high
* Allocating a lower starting credit limit on a credit card or line of credit
* Charging greater interest rate on a loan

**About the data**

Our dataset was sourced from [*https://www.lendingclub.com/info/download-data.action*](https://www.lendingclub.com/info/download-data.action). This is Lending Club’s complete loan data for all loans issued through 2007 - 2011. Lending Club is the world’s largest online credit marketplace which facilitates in personal loans, business loans, and financing for elective medical procedures. Investors provide the capital to enable many of the loans in exchange for earning interest. Lending club operates fully online with no branch infrastructure.

This dataset contains a total of 42538 rows and 115 columns. Out of these 115 columns over 68 columns were missing values. The includes the data regarding the current loan status and latest payment information.

**Attributes**

The attributes in our dataset describe loan details. We’ve index values like Member ID, ID, Employer Name, Title, State Address, City Address, Loan Listed Date, Loan Issued Date and Purpose of the loan. These index values were removed as they do not contain information related to the prediction.

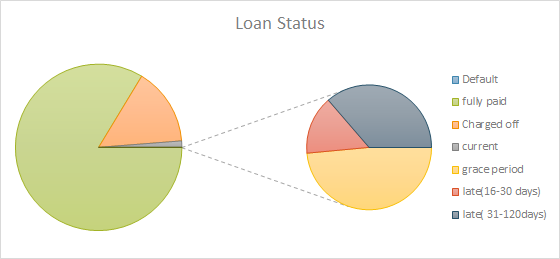
The table below gives descriptions of some of the important attributes.

|  |  |
| --- | --- |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| loan\_amnt | 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. |
| total\_pymnt | Payments received to date for total amount funded |
| Fico\_range\_high and fico\_range\_low | The upper and lower boundary range of the borrower’s FICO at loan origination belongs to. |
| Loan\_status | Current status of the loan |

The values of Loan\_status is as follows:

* Current: The loan is yet to be fully paid but the obligor is paying as per schedule and no delay has been observed
* Fully Paid: The loan has been completely paid by the user without any hiccups on the installments
* Grace Period: A period of 0-15 days wherein the obligor has dues and has not paid up his installment
* Late (16-30 days): A period of 15 days after the grace period where the obligor is put into the watch list and is marked as delinquent in the first bucket
* Late (31-120 days): A period of 1-4 months after grace period for the obligor where he/she has failed to pay after soft collection strategies. Requires the company to pay heed to such accounts to avoid write-offs
* Default: After 4 months of non-payment of dues, the company has marked the obligor as defaulter and his credit score is drastically affected. Loan company is unde
* Charged-Off: The loan has been written off after default by the Loan Company that it cannot be recovered and shown as a loss in the financial statements of the creditorr severe stress and the loan is difficult to be re-collected

The distribution of the different Loan\_status values is illustrated in the below pie chart.

****

**The process**

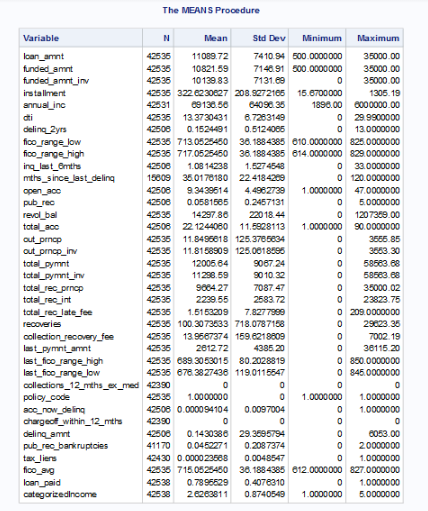
**Data Cleaning**

**Means Procedure**

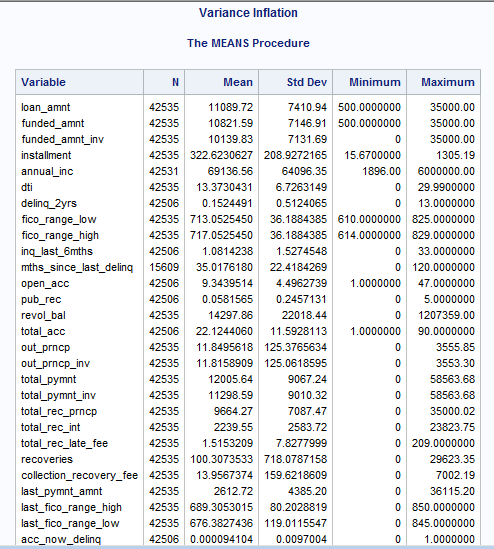
Means Procedure is used for analysing the numeric data.

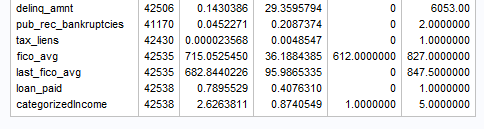
The below table provides the Mean Value, Standard Deviation, Minimum value and the Maximum Value of all the variables.

The Standard Deviation of some of the variables is zero. Which means, that there is no variation in the variable and it would not help in predicting the dependent variable. So, we remove these variables.

****

After removing the variables having Standard Deviation zero, we get the following variables that might help in predicting the loan\_paid variable that is our dependent variable. We get the mean value for loan\_paid as 0.7895, which determines that , the loan is paid 78% of the times.

****

****

The dependent variable loan\_paid is created from the loan\_status variable. The loan\_status variable indicates the status of the payment of the loan.

The “Fully Paid", "Current" or "Grace Period status is set as 1 while the rest are set as 0.

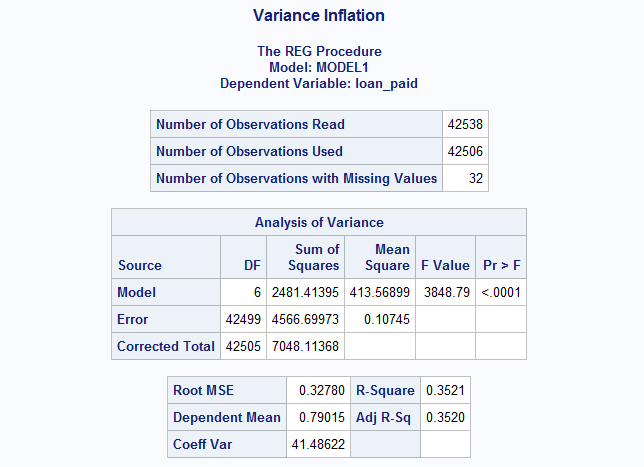
One more categorical variable is created called the “categorizedIncome” which categorizes the annual income of the people into 5 categories. The major part of the data falls into 4000 to 1 lakh annual income range, so this range is categorized into 1,2,3 while the remaining data is categorized into 4 and 5.

The fico\_avg and the last\_fico\_avg variables are calculated and created using the current and last FICO values.

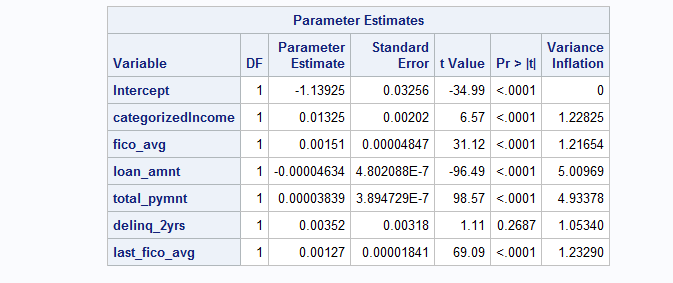
We use all this data to carry out the further analysis.

**Variance Inflation(VIF):**

Variance Inflation(VIF) helps in determining the multi-collinearity among the variables. That is, which variable is correlated to which other variables. The VIF values should be less than 5 or 10 for a clean dataset.

****

We can see that R-squared value is 0.3521 that is 35% of the times the dependent variable can be predicted correctly with the help of the independent variables.

****

Except the delinq\_2yrs, the Pr-value of all the variables is less than 5% which depicts that all these variables are significant.

We remove the variables having high VIF value.

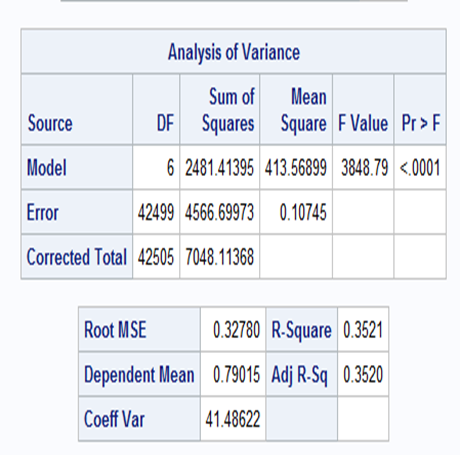
From above we can see that the VIF value for the variables is less than 10 which tells us that the variables are not highly correlated to each-other.

So, these variables now can be used for predicting the dependent variable (loan\_paid).

**Principle Component Analysis**

**Multiple regression**

Multiple regression is a type of regression model used when we want to predict the value of the variable based on the value of two or more variables. The variables we want to predict is called the dependent variable. The variables we are using to predict the value  of the dependent variables are called the independent variables. In our model we have considered dependent variable as loan\_ paid  which has only two category 0 and 1(It’s 1 if the customer pays the loan else it’s 0). And we have used income of the customer , FICW avg, loan amount taken by the customer, total payment and deling of the customer of 2 years, how many account does the customer has opened and dti are the independent variables.

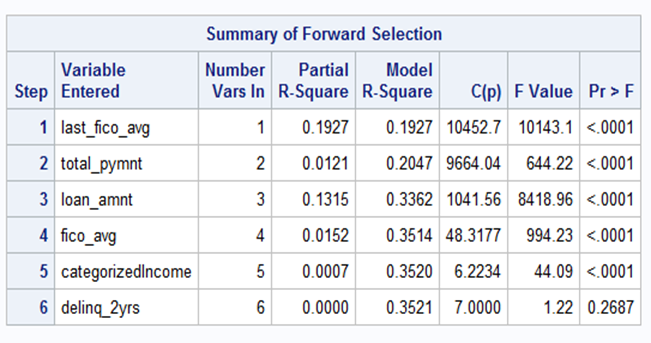


The model which we obtain by running the PROC Reg is significant, as we can see in the above diagram the P-value is more than 0.05. The R-square value tells how well the model fit the data. The R-square value is 0.3521, that is we explain the 35% of the original variability and the left 65% is the residual variability.

**MODEL-SELECTION METHODS**

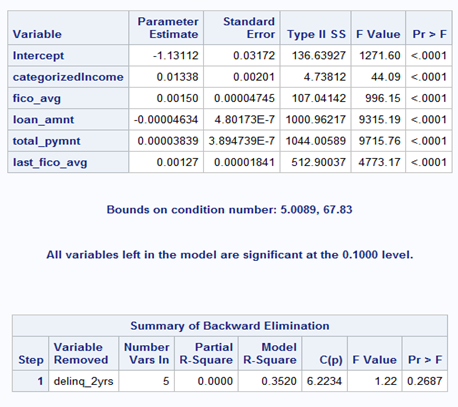
**i. Forward Selection**

The forward-selection technique begins with no variables in the model. For each of the independent variables, the FORWARD method calculates https://lh3.googleusercontent.com/Uz7RiBWM73TogUQLhCAP-UB9CfuGLbpMC4QA4k9xAxQpHRH7fX1foLF7UKDtZ6g0inwVdzCsTKnqY5-FQzridbtyVRvl8CpylB_f1bG9lVd7tLFAXnVLqlUePQuGtQ3FV7ojV-8Astatistics that reflect the variable’s contribution to the model if it is included. The https://lh3.googleusercontent.com/cWUrdw30OIvC5VNbhierXFfsZ_sDC5UOrj3UXZRwo1fSW6XFKCh2OA1APZXoXkv17c5TRtl1h3A5lk9s3VI7CSDqKfEuf_eWa5d7WuhpIH3CU_QpWe7McaXSUtVuXj8oABjv3WCF-values for these https://lh4.googleusercontent.com/kTC6vud0rpoYM2l5WUtcQM6QaH-USwSo5ggtlPIVUzL_1mU_eN_aOmM9W34BpwhZSqi34ZYeEpWaeNLAaN6hWu04vrfdXCejjyNWFe9OeyfTebAbw2Owk04sScUWjgX0W4ha-i63statistics are compared to the SLENTRY= ( 0.50 if the SLENTRY= option is omitted). If no https://lh6.googleusercontent.com/BPplWlmSuU4wF-POjJyHGCj9uozBYvsCMV6c43MxQf7Te7JvcEt_eE84hKeBCdqqjSMgcBDqwg7KiXT2_2aFyoa1qL7Slk9p2N0DXFcGr4nSn7KpegYvdnPIFelP2PqOBwfRAozjstatistic has a significance level greater than the SLENTRY= value, the FORWARD selection stops. Otherwise, the FORWARD method adds the variable that has the largest https://lh6.googleusercontent.com/IKgxZzL0nIXoQG4DiAtJjlYnHdFrvC7gBWPPFfOrBe9jTwv428-ltJJ2HRHXw2HiQwSZSjMsLtiq2lHb6wS2p9wGNtlFRT3lJVkbispnhMxQUh2G1qkWIkrHRUrYzQDC2OZk3a5ostatistic to the model. The FORWARD method then calculates https://lh6.googleusercontent.com/_MpBkrQY2WNOYanF-8M_it5yyb3SaUer2lfC7eulVInfAbhY9XVCq8BEdPbE5gyWQJI-pOeTBjXWQiT4r-MXtdPiM4054cHVIVMefm3BG51R-v5JME6-tcKtcxs0J46Smyyern6Nstatistics again for the variables still remaining outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant https://lh5.googleusercontent.com/8MLo0ZECESyWJegSLwOMEXmb_qKu2cbygigZ3JpqVgpuNtX1HvHVuwFFOYbleELyfszDYxGMbIQltMPnVYDKbqIX2HxSTVlFWwpWFySWwsMH6MqoMOgpSj5KDtANmOEnwuKOj8Zastatistic. Once a variable is in the model, it stays.



**ii. Backward Elimination**

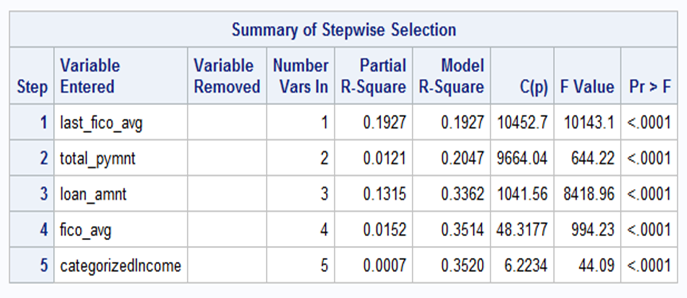
The backward elimination technique begins by calculating https://lh3.googleusercontent.com/Li5IQvK8JhN0S5c-Lp_A0vi85xA1FasBiO8TyGqoCAWE1Yty7C48dLV1UoDbsCnSYKOSx6JliDwxn3vHScaekC2mIyGov2oiwkkgBni7st7_DlAobTMXaKtUzmQE1WZKIy7XuoXpstatistics for a model which includes all of the independent variables. Then the variables are deleted from the model one by one until all the variables remaining in the model produce https://lh4.googleusercontent.com/IUCttQrv0L37invSc_Hy_3EH8EMI8xIoDkQKSi-DPKk2LoN1U5R5YYDhZWsgPdyeo25lcQt7LDBsSpab57qVu-Xs-MXl6YG2tAdJDGm54ZxUTGF731lniEGlzTOWG2pxNIUFzFFBstatistics significant at the SLSTAY= level  ( at the 0.10 level if the SLSTAY= option is omitted). At each step, the variable showing the smallest contribution to the model is deleted.



As we can see the above  model is what we obtain after the backward selection process. The variable deling\_2yrs is been removed from the model. Since it has the least contribution to the model and the its  F statistical is not significant at 0.10 level.

**iii. Stepwise selection**

The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay there. As in the forward-selection method, variables are added one by one to the model, and the https://lh6.googleusercontent.com/nY1wTQ5XYP8oSbQC3HXaR3WzJrIaElzl62YON5dJkOc_YkGu7iRuHWHDWgK4IYR8xn7aT19Il4_Nz-8Q2ASNI7iIIo6XvHrORYkaPt_udkjPQXATQV149YAGkZwqfP5NfN4-Qq88statistic for a variable to be added must be significant at the SLENTRY= level. After a variable is added, however, the stepwise method looks at all the variables already included in the model and deletes any variable that does not produce an https://lh4.googleusercontent.com/1BMeFxIIT4zC7yd84KAmuFaPZDxeP0l57RoPJRj4mTuBmLlgoUMjhZViJQ64WumHAUUZFaQ56mx-Kp8UmrTV9rowRKUPFJ6xoWat0sFTOg1H8WIKrygoGUZEVqDclGtdfMfH8b9mstatistic significant at the SLSTAY= level. Only after this check is made and the necessary deletions are accomplished can another variable be added to the model. The stepwise process ends when none of the variables outside the model has an https://lh3.googleusercontent.com/e1pb6so32dr1mSW6XSguQB4Y3SlRznzHJl2FeYaXBnZd4fHCLrrDsEj4tdFzOXtGNqWQTxGJ04M145zCl2rIFu_8bgsAapb4nShlfHPeG0oUmuEQRab9bKUNwNNy_hhNgPTofMI7statistic significant at the SLENTRY= level and every variable in the model is significant at the SLSTAY= level, or when the variable to be added to the model is the one just deleted from it.

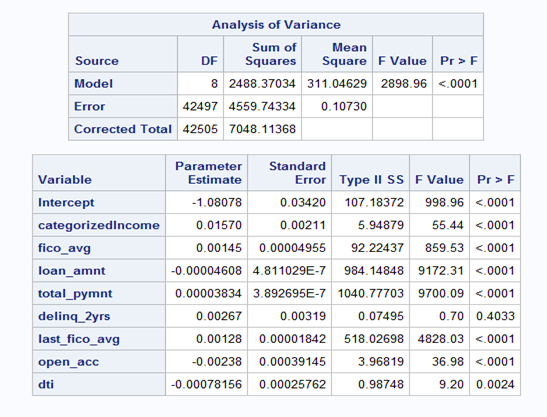


**iv. Subset Procedure**

We have used the **MAXR** selection procedure to find the best model which fits the data.The maximum https://lh4.googleusercontent.com/jN6TA0dAW9qOi7fv953zsRvECHBce02JSGsAuPzYYGCFjYUm--RxnVtAXbmZAAxlFaCfDKKuPx0nMOw8rroERRJM9mg2485OnAW-iVK7Z9WfeqMwc1MmdplvTFWiToXSbqHHxfFiimprovement technique does not settle on a single model. Instead, it tries to find the "best" one-variable model, the "best" two-variable model, and so forth, although it is not guaranteed to find the model with the largest https://lh5.googleusercontent.com/LxhR1UD4Ec2jFdq7-5YBCFp1ACPfWYpB7aGmf0EzKWLiLGdJYeyvmaCRJ2H8Brt8sYGSN2G1IDXTZH7oXyTq7PtWk7Zr7RQR2QkfhDi8VbgMtIQPr8eQ-NbIQJUUkzUlysYhNHl_for each size.

The MAXR method begins by finding the one-variable model producing the highest https://lh4.googleusercontent.com/D5f957ZSlW__jcGuM-EtmDZs-P8IfG916xgqBq_IWHHM4g_QnBvm73GGbLD8lUWqP2zqPp2aWrjgB2Udo6w7ESBRBkcqMZ5pHuMp4Xj-kSOcqfjlUgpcWzdHzSNM2aGWSF5wxQdN. Then another variable, the one that yields the greatest increase in https://lh3.googleusercontent.com/bsBtd-lukk9OQwwqtBUgRw3Hca3vAaRMQVi-4a0An5PfqPNMmF4LggphQa_ZT50OhpM1_vq8UIPfmrgpeUu81b1UxfTsR_T94zyg1eh14F0EmlHhK4hM32ehk_zm3uJcGdVs6xSz, is added. Once the two-variable model is obtained, each of the variables in the model is compared to each variable not in the model. For each comparison, the MAXR method determines if removing one variable and replacing it with the other variable increases https://lh3.googleusercontent.com/mS70ata3aIUn0vEF3tfglT8YguZmKsPsx2yiqSSV0A71VdtlCgIUFAkyqfHWBj6mwNoVVpwLGTDqsPVkfrdFzfPpNr1-yjiMhpTidlT6_qwmDgm7kcp8g-RQLBdBHGfUJ2lBhdi5. After comparing all possible switches, the MAXR method makes the switch that produces the largest increase in https://lh3.googleusercontent.com/w07fV9q7swk6RGItVIcrW3BTu12YCQVFlKu9i8Xyz7rvho80za1eLKZxQTHRXKo3sPGqDw-QiRu-z0Q5_d3DieYqMjpJP_iyWC1grJxV4ZR0XhrAI1AAv0AGZrULbziEh-JSVOZL. Comparisons begin again, and the process continues until the MAXR method finds that no switch could increase https://lh6.googleusercontent.com/C1AWcl9e98CQPS8SIyiE-dmLD-fSJ2b18mzjhoS9ftyeXZHhsfLj6jaOhehoReIGoFIjaBwTrT6Vn7BmCqe8jvkh03YmNIqgddPSiZAiXnPUT-TsmjjRm9Hgp0P5uRaOEjgCNvno. Thus, the two-variable model achieved is considered the "best" two-variable model the technique can find. Another variable is then added to the model, and the comparing-and-switching process is repeated to find the "best" three-variable model, and so forth.

The difference between the STEPWISE method and the MAXR method is that all switches are evaluated before any switch is made in the MAXR method. In the STEPWISE method, the "worst" variable might be removed without considering what adding the "best" remaining variable might accomplish. The MAXR method might require much more computer time than the STEPWISE method.



When we applied MAXR selection on our data set the variable Last\_fico\_avg entered the table in step 1 and in step 3 the Last\_fico\_avg variable is replaced with loan\_amnt. But the MAXR selection will again check all the variable and verify whether it is significant and add the variable to the model. So, the Last\_fico\_avg is added again to model at step 4. This is the main advantage of the MAXR selection process.

**Logistic Regression**

Logistic Regression is a type of regression model used for categorical dependent variable. It is typical when the categorical variable is binary in nature. Here our variable is *loan\_paid*(1 if the loan is paid, else it is 0). Logistic regression is used mostly in academic and industrial areas. The most commonly used area is in financial firms for the credit score analysis. We have taken the loan data to predict the credit risk based on a few factors such as the income of the customer, fico average, loan amount, the total payment by the customer to the bank, delinquent amount in the past 2 years and the last fico average. We will be predicting the probability if the bank is risking or not. Logistic regression is a popular modelling technique to find this.

**First model**

Dependent Variables : *loan\_paid*

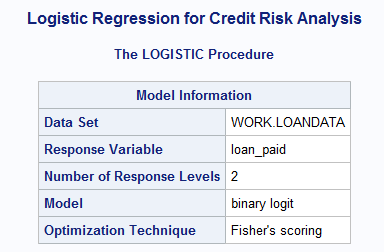
Independent Variables : *categorizedIncome, fico\_avg, loan\_amnt, total\_pymnt, delinq\_2yrs, last\_fico\_avg*

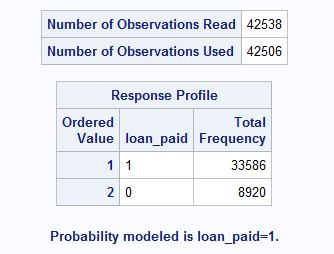
**Code :**

*title "Logistic Regression for Credit Risk Analysis";   
proc logistic data= loandata descending;  
class categorizedIncome/ param=ref ;  
model loan\_paid = categorizedIncome fico\_avg loan\_amnt total\_pymnt delinq\_2yrs last\_fico\_avg;  
run;*

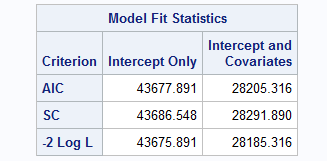
The logistic data model is set to descending because by default the logistic model takes ascending and the value when it is ascending is *loan\_data* = 0. We want to predict *loan\_data* = 1. Our categorical variable is categorizedIncome. This takes all the categories into consideration. There will be 4 separate estimates else the values for *categorizedIncome* will be taken as continuous is the class is not set. The left hand side is the dependent variable *loan\_paid* and the right hand side are the independent variables *categorizedIncome fico\_avg loan\_amnt total\_pymnt delinq\_2yrs last\_fico\_avg.*

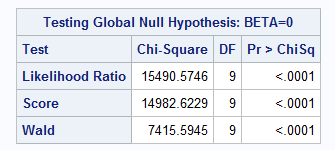
**Output for the first model :**

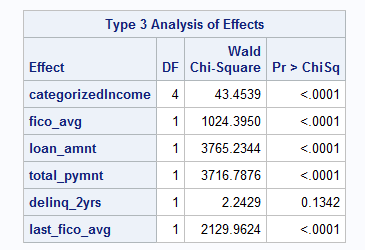
****

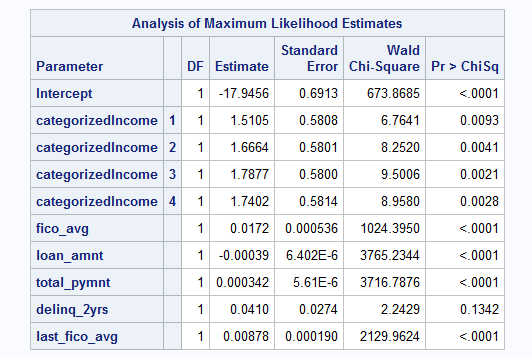
****

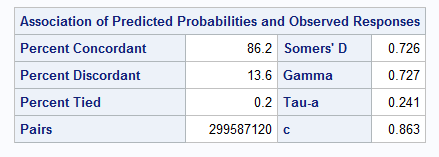
****

****

****

****

****

****

In the first model, the number of observations are 42538 and the number of observations used are 42506 because of the missing data points in the dependent and independent variable. Hence, the whole row is deleted. The response profile signifies the frequency of the dependent variable i.e *loan\_paid*. There is a good distribution for the probability to be 1. The following steps can be continued as the model convergence status is satisfied and the result will be optimal.

The model fit statistics decides if a model is fit  based on the AIC, SC and -2 Log L values. The lesser the value, the better the model. These values are determined by the independent variables. If more and more independent variables are used, the AIC , SC and -2 Log L values will be penalized.

In the test for Global Null Hypothesis, all the Beta parameters are set to 0. The alternate to the null hypothesis is that at least one of the Beta parameter is non-zero. It means that the model is able to get a good result than just a random guess. Which signifies that the independent variables are of some good. If the Pr > ChiSq value is not significant (i.e. > 0.05) then the model has to be reworked. In our case, all the three values are significant and hence we can proceed.

In the Analysis of Maximum Likelihood Estimates the intercept value is -17.9456 and the slopes are mentioned below in the table.  The Pr > ChiSq values need to be significant for the model to be a best fit. But the delinq\_2yrs value is not less than 0.05 and hence this value has to be eliminated and the model has to be reworked to obtain a better result. Which is done below.

**Second model**

Dependent Variables : *loan\_paid*

Independent Variables : *categorizedIncome, fico\_avg, loan\_amnt, total\_pymnt, last\_fico\_avg*

**Code :**

*title "Logistic Regression for Credit Risk Analysis without delinq\_2yrs";*

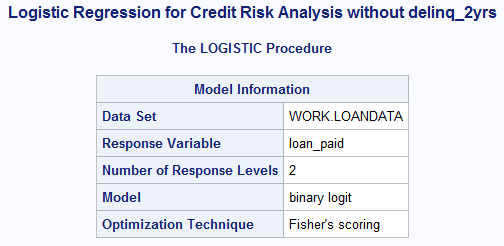
*proc logistic data= loandata descending;*

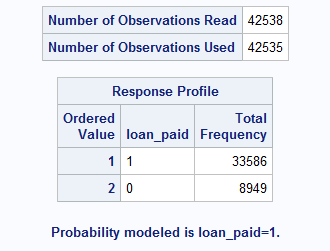
*class categorizedIncome/ param=ref ;*

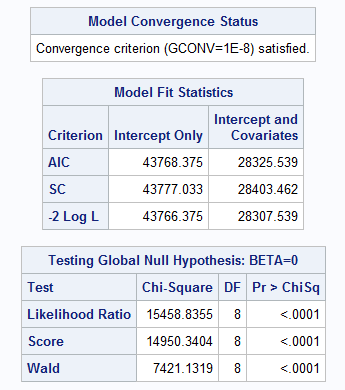
*model loan\_paid = categorizedIncome fico\_avg loan\_amnt total\_pymnt last\_fico\_avg;*

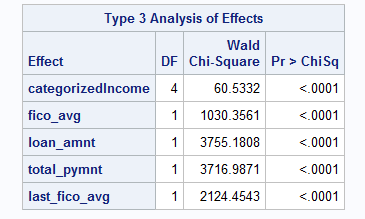
*run;*

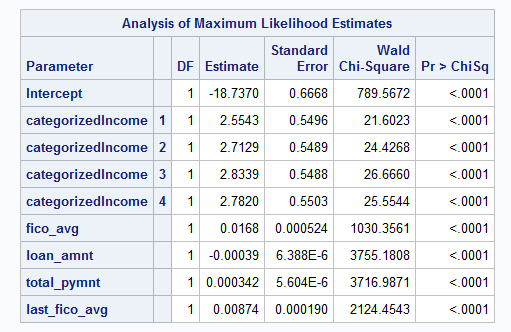
**Output for the Second model:**

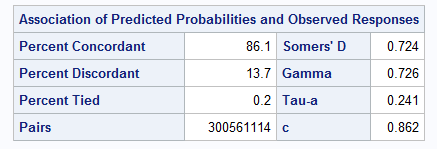
****

****

****

****

****

****

In the Analysis of Maximum Likelihood Estimates the intercept value is -17.9456 and the slopes are mentioned below in the table.  The Pr > ChiSq values need to be significant for the model to be a best fit. Here, all the values are less than 0.05 which means that they are significant and this is a better model than the first one.

The association of Predicted Probabilities and observed responses the Percent Concordant is 86.1 percent. The higher the value, the better the model. The Concordant-statistics value is 0.862. This value is greater than 0.5 and almost close to 1. Hence, it is a good model.

**Conclusion**

Multiple iterations of variable selection and model specifications are required to select a best model. Data Collection methods must be robust and verified by the loan company before using them for modelling.

**References**

https://www.lendingclub.com/info/download-data.action