Prediction of Loan Status using Logistic Regression Model

Subha Raut

Aniket Singh

Subham Singh

Youngstown State University

Dr. Andy Chang

SAS Programming and Data Analytics

1. Introduction

This paper is about predictive modeling for loan data. The model used in this paper is Logistic Regression. When a customer borrows money from the bank, they borrow the money for a specific time. The banks need to determine that the borrower can pay the loan or not before issuing the loan. As we know, in the US, we have many different credit scores. The most widely used would be the FICO score. This system is excellent to know about the customer because a right FICO score customer would potentially be a good borrower. Theoretically, that sounds true, but a lousy FICO score borrower might not be a bad borrower, and similarly, a good FICO score borrower might not be a good borrower either. There are many factors behind being an unsuccessful borrower and a good borrower than just a FICO score. So, instead of relying on FICO scores, banks need to be smart about the financial decisions they make. Situations like this are where predictive modeling comes into play. Predictive modeling can help determine if the borrower can pay the balance or not. The paper's predictive modeling shows the ultimate result that if the account is going to be delinquent or not.

2. Methods/Data Preparation

The loan data used for this model was obtained from the blackboard. As this is a part of our SAS data analytics course, the data has been provided by Dr. Andy Chang. The data provided is in excel format (xlsx). After importing the data to the SAS studio, we created training and validation data using stratified sampling. As for the training data, we choose to have 67% of the data for the training data and the remaining 33% as validation data. The data contains 141 columns and 20,000 rows. Before the data could be used, we noticed that 76 columns had no data; since the columns did not have any information, we choose to remove them from the dataset. Each of the rows describes some information about the loan information and other

financial information that might help the model. The model needs to predict if the account will be delinquent or not at the end. So, the only variable that we need to predict is 'Loan_status' that had two possible outcomes, either "Delinquent" or "Not Delinquent." The target variable would be 'Loan_status'.

In predictive modelling, it is important to prepare the input variables since the data is more likely to be unmanaged/unstructured. There are several problems that needs to be resolved in a data for predictive modelling. The most common problems are missing values, high number of levels in categorical variables, redundancy, multicollinearity. We have used different techniques to handle these problems, which we have described below:

2.1. Handling missing values:

In order to deal with the missing values in numerical predictor variables, we imputed the missing values with the median. For categorical inputs, we introduced new category in place of the missing values. Since the imputation of values by using the unconditional mean or median of the variable does not take into account the relationship with other inputs, there could be a significant misrepresentation of the data. Therefore, it would have been more efficient if we would have used cluster imputation to deal with the missing values. However, due to the time constraint for this project, we are not using this technique to deal with the missing values.

2.2. Dealing with categorical inputs:

Too many levels of a categorical variable are one of the most frequently occurring problems in predictive modeling. When the nominal variables have few numbers of levels, we can use dummy coding to include nominal inputs in the logistic regression model. However, variables such as zip code can have too many levels. For instance, in our dataset, the variable zip code had around 800 levels. The number of dummy variables for the zip code variable would be one less

than the number of levels in zip code. It would lead to the problem of high dimensionality which will ultimately cause the over-fitting of the data. Thus, the model will not be able to generalize properly to a new dataset. Moreover, inclusion of categorical variables with too many inputs can lead to the problem of quasi-complete separation. The problem of quasi-complete separation occurs when a level of the categorical variable has a target response of either 0% or 100% and can cause the interpretation of the regression model. We solved these problems by collapsing the levels based on the reduction in the chi-square test of association between the categorical input and the response variable. However, we did not include few variables such as zip code for our final model because chi-square test did not seem to be a valid test as 59% of the cells had expected counts less than 5 for cross-tabulation of the variables zip code and loan status. We could have merged the categories in zip code variable further or smarter variables could have been created by linking the given data to different datasets for better result.

2.3.Reducing redundancy:

Datasets with too many input variables can cause different problems in development of predictive models. For instance, the probability of redundant variables in a dataset with more input variables is high. In high dimensional datasets, the association between the inputs makes it difficult to identify the best predictor variables. So, we reduced the redundancy by clustering numeric variables in the training data set. Finally, we used criteria such as 1-Rsquare statistic, relationship between predictor variable and the target, and subject matter knowledge to choose a variable from each cluster.

Cluster	Variable	1-Rsquare Ratio	Variable Label
Cluster 1	loan_amnt	0.0929	loan_amnt
	funded_amnt	0.0690	funded_amnt
	funded_amnt_inv	0.0726	funded_amnt_inv
	installment	0.1505	installment
	total_pymnt	0.0711	total_pymnt
	total_pymnt_inv	0.0799	total_pymnt_inv
	total_rec_prncp	0.1869	total_rec_prncp
	total_rec_int	0.4600	total_rec_int
Cluster 2	recoveries	0.1019	recoveries
	collection_recovery_fee	0.1039	collection_recovery_fee
Cluster 3	pub_rec	0.0620	pub_rec
	pub_rec_bankruptcies	0.0594	pub_rec_bankruptcies
Cluster 4	open_acc	0.1747	open_acc
	total_acc	0.1845	total_acc
Cluster 5	delinq_2yrs	0.2393	delinq_2yrs
	mths_since_last_delinq	0.2341	mths_since_last_delinq
Cluster 6	dti	0.4117	dti
	revol_util	0.4816	revol_util
Cluster 7	settlement_term	0.0000	settlement_term
Cluster 8	int_rate	0.0000	int_rate
Cluster 9	mths_since_last_record	0.0000	mths_since_last_record
Cluster 10	inq_last_6mths	0.0000	inq_last_6mths
Cluster 11	total_rec_late_fee	0.0000	total_rec_late_fee
Cluster 12	annual_inc	0.3550	annual_inc
	revol_bal	0.3431	revol_bal
Cluster 13	settlement_amount	0.0000	settlement_amount
Cluster 14	settlement_percentage	0.0000	settlement_percentage
Cluster 15	last_pymnt_amnt	0.0000	last_pymnt_amnt

Figure 1:Reducing redundancy by clustering the variables.

2.4 Variable screening

It is significant that inputs should have a linear relationship with the target in logistic regression. If the linear relationship is violated, the model will not generalize well in other datasets. Thus, it was important to detect the nonlinear relationships. We detected the nonlinear associations by using a variable screening method. The variable screening method compares the ranks of the Spearman correlation statistic with the ranks of Hoeffding's D statistic. Variables with low rank of Spearman and high rank of Hoeffding have week association with the target. Therefore, we eliminated these types of variables from the model.

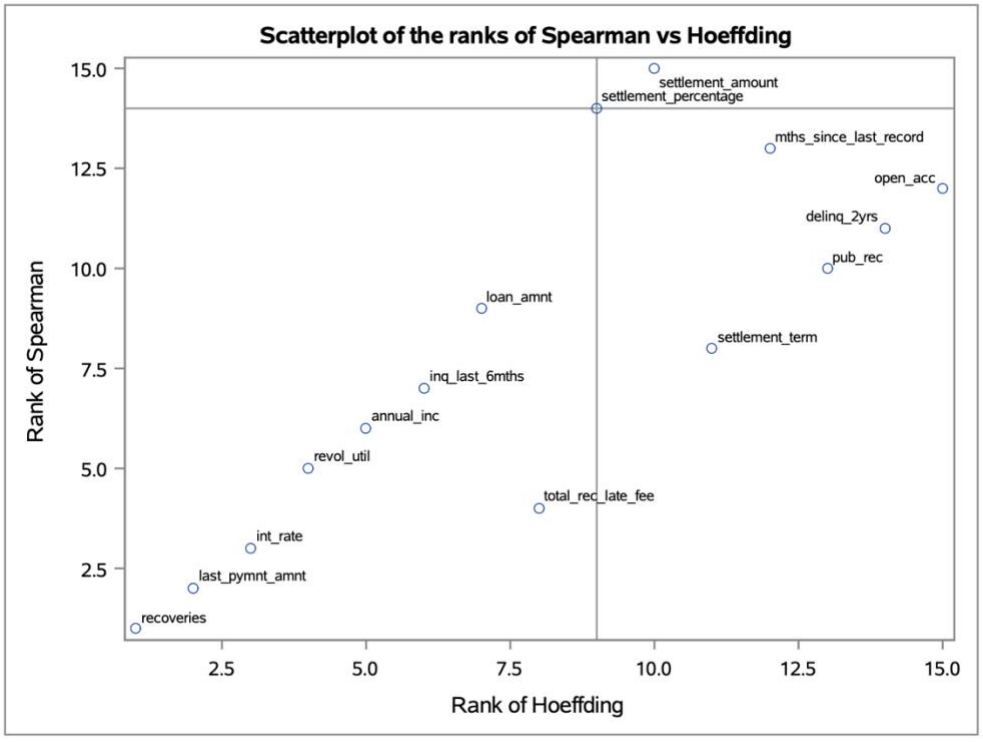

Figure 2: Scatterplot of the ranks of Spearman vs Hoeffding

2.5 Sequential selection of variables:

After variable screening, we used backward elimination for subset selection to select the most predictive variables for our model. The variables that were found to the best predictors with their meanings are as follows:

- 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.
- int_rate: Interest rate on the loan.
- inq_last_6mths: The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
- total_rec_last_fee: Late fees received to date.
- annual_inc: The self-reported annual income provided by the borrower during registration.
- last_pymnt_amnt: Last total payment amount received.
- purpose: A category provided by the borrower for the loan request.
- addr_state: The state provided by the borrower in the loan application
- term: The number of payments on the loan. Values are in months and can be either 36 or 60.

3. Analysis/Results

3.1. The Logistic Regression Model.

After finding the best nine predictors for predicting the response variable, binary logistic regression was used to find the probability estimation formula for a loan delinquency case.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.8970	0.2598	11.9175	0.0006
loan_amnt		1	-0.00006	5.045E-6	164.7513	<.0001
int_rate		1	-12.6651	0.8570	218.4207	<.0001
inq_last_6mths		1	-0.1512	0.0284	28.4062	<.0001
total_rec_late_fee		1	-0.0295	0.00402	53.9911	<.0001
annual_inc		1	7.945E-6	9.186E-7	74.8147	<.0001
last_pymnt_amnt		1	0.00226	0.000116	381.0931	<.0001
purpose	1	1	0.6476	0.1285	25.3786	<.0001
purpose	2	1	0.3753	0.1218	9.4945	0.0021
addr_state	1	1	0.8992	0.1954	21.1807	<.0001
addr_state	2	1	0.7423	0.1899	15.2816	<.0001
addr_state	3	1	0.5676	0.1905	8.8753	0.0029
addr_state	4	1	0.4857	0.1948	6.2204	0.0126
term	36 months	1	0.5400	0.0737	53.6575	<.0001

Figure 3:Analoysis of the maximum likelihood estimates from the logistic regression equation

The logistic regression equation is:

P (Fully Paid) =

$$\frac{1}{1+e^{-(0.8970-0.00006a-12.6651b-0.1512c-0.0295d+7.945ze-6e+0.00226f+0.6476g+0.3753h+0.8992i+0.7423j+0.5676k+0.4857l+0.5400m}}$$

Where,

a=loan_amnt,

b=int_rate,

c=inq_last_6mths,

d= total_rec_late_fee,

e=annual_inc,

f=last_pymnt_amnt,

g=purpose(1),

h=purpose(2),

i=addr_state(1),

j=addr_state(2),

k=addr_state(3),

l=addr_state(4),

m=term(36 months)

3.2. Interpretation of Odds Ratio

From figure 4, we can see the odds ratio estimates and profile-likelihood confidence intervals for different predictor variables. Odds ratio are an important tool for analysis since odds ratio greater than 1 indicates a positive association and odds ratio less than 1 indicates a negative association. For instance, if there is a $1000 increase in loan amount, the odds of a customer to not pay the full loan amount increase by a factor of 1.067.

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
loan_amnt	1000.0	0.937	0.928	0.947
int_rate	1.0000	<0.001	<0.001	<0.001
inq_last_6mths	1.0000	0.860	0.813	0.909
total_rec_late_fee	1.0000	0.971	0.963	0.978
annual_inc	1000.0	1.008	1.006	1.010
last_pymnt_amnt	1.0000	1.002	1.002	1.002
purpose 1 vs 3	1.0000	1.911	1.484	2.456
purpose 2 vs 3	1.0000	1.455	1.145	1.846
addr_state 1 vs 5	1.0000	2.458	1.671	3.597
addr_state 2 vs 5	1.0000	2.101	1.443	3.040
addr_state 3 vs 5	1.0000	1.764	1.210	2.556
addr_state 4 vs 5	1.0000	1.625	1.106	2.375
term 36 months vs 60 months	1.0000	1.716	1.485	1.982

Figure 4:Estimates of odd ratio and profile-likelihood confidence intervals

3.3.Model Performance

To do an honest assessment of how well our model performs on different sample of data, we used different statistical tests. At first, to test the predictive accuracy of our statistical model, we used the concordance statistic. The higher the c-statistic, the better the model can discriminate between subjects who do experience the outcome of interest and subjects who do not. From figure 5, we can see that the value of c-statistic for our model is 0.889 which is very close to 1. However, the concordance statistic is not enough to test the predictive accuracy of our statistical model since this statistic is derived using our training dataset. Therefore, the validation data was also used to test the predictive power of the model. Since the prediction of a logistic regression model is a logit or probability, in order to use it to classify the positive and negative response, we

need to define a cutoff value. To determine a cutoff value, ROC curve was used to visualize and quantify the tradeoff we're making between the two measures. After, visualizing the ROC curve, we decided to use 0.7 as a cutoff value to classify between the two responses of loan status variable.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	88.9	Somers' D	0.778
Percent Discordant	11.1	Gamma	0.778
Percent Tied	0.0	Tau-a	0.201
Pairs	23023286	c	0.889

Figure 5: Association of predicted probabilities and observed responses.

We used classification table as another method to test the predictive accuracy of our logistic regression model. In this table the observed values for the dependent outcome and the predicted values (at a user defined cut-off value, in our case p=0.7) from the validation data set are cross classified. From figure 6, we can see that the correct rate of classification is 84.9. This test also indicates that our model is able to generalize well.

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.700	10047	1279	759	1250	84.9	88.9	62.8	93.0	50.6

Figure 6: Sensitivity and Specificity analysis.

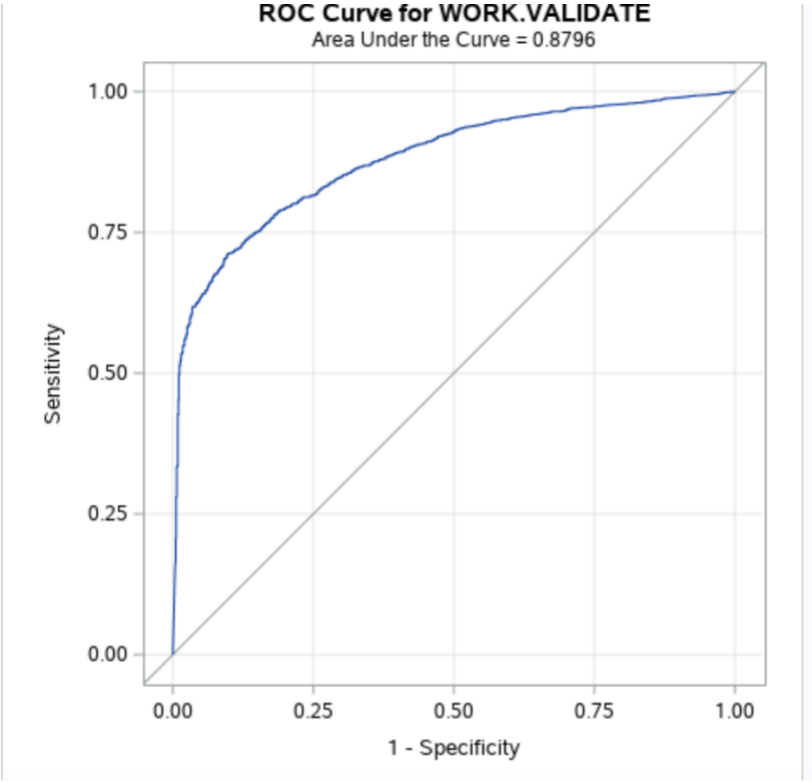

Figure7: ROC curve for validation data.

4. Conclusion:

We solved the problem of predicting loan delinquency status by building a binary logistic regression model using a data with different features. We partitioned our dataset in two groups: training and validation. The model which was built using the training data was able to generalize well in the validation data set. We used other statistical tools to measure the model performance like K-S statistic and we found that the model can be further improved by including different other variables like description. From procedures like text-mining we could have found a lot of details about a customer who has applied for loan.

The final summary of our project is described below:

➔ The logistic regression equation is:

P (Fully Paid) =

$$\frac{1}{1+e^{-(0.8970-0.00006a-12.6651b-0.1512c-0.0295d+7.945ze-6e+0.00226f+0.6476g+0.3753h+0.8992i+0.7423j+0.5676k+0.4857l+0.5400m}}$$

- ➔ The correct rate of classification is 84.9.
- ➔ The probability of non-delinquency for a customer with case like 20000 of our data is 0.336005. Since the p-value is less than 0.7, we do not approve the loan for cases like 20000.

Appendix

Task performed by each group member are described below:

1. **Subha Raut:** Preparation of data using Excel and SAS.
2. **Aniket Singh:** Second stage preparation of data using SAS. Detection of non-linearities by variable screening method.
3. **Subham Singh:** Measured the model performance.