Statistical Prediction for Loan Status

Trelan Hakeem-Barron

Statistics 5811: Spring 2019

Professor: Andy Chang

**Executive Summary**

**Introduction**

In this study we are going to explore various factors related to customers’ loan status and find a good predictive model for customers’ loan status, such as good or bad loan.

**Methods**

In this study we looked at data from the following Lending Club data site. The first data set is for the period from 2007 to 2011. This data shows multiple variables used to build our final predictive model. With a good predictive model, we can find out the factors that are related to customers’ loan status.

**Results**

From our results at the end of the study, it seems that different variable created different models. The variables are good factors for finding out the customers loan status overall.

**Conclusion**

From the results off of this study alone, we were able to determine that the variables were good for our predictive model. By using our predictive model, we found out that our predictive model holds and the variables are good predicting factors in finding out if it is a good or bad loan.

**Introduction:**

In today's society we have a major increase in in overall loans. This includes student loans, refinancing, and credit card payoffs. Currently, California has the most loan issuance. The loans can be good or bad. Is the loan good or bad? This is what we shall figure out in our experiment. We will be building a predictive model to discover the factors of a good or bad loan.

**Methods**

This is data that is being used from the Lending Club. The interest in this data is loan\_status. The research we are doing is to help further the study of the factors and contributions of a good or bad loan status. The data set is from the period from 2007-2011. This data contains multiple variables that will be used to create our predictive model. With a good predictive model, we can find out the factors that are related to customers’ loan status.

**Analysis:**

***Descriptive Statistics:***

Below you can see the results for the loan status predictions. Based off the predictions, we are able to start building our model.

**Figure 1: Parameter Estimatses**

| **Parameter Estimates** | | | | |
| --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **t Value** |
| **Intercept** | 1 | 0.687171 | 0.016249 | 42.29 |
| **loan\_amnt** | 1 | 0.000047941 | 0.000002330 | 20.57 |
| **total\_pymnt** | 1 | -0.000029747 | 0.000001305 | -22.79 |
| **open\_acc** | 1 | -0.008880 | 0.001360 | -6.53 |
| **annual\_inc** | 1 | -0.000000781 | 0.000000135 | -5.80 |
| **installment** | 1 | -0.000591 | 0.000085585 | -6.91 |

| **Stop Details** | | | | |
| --- | --- | --- | --- | --- |
| **Candidate For** | **Effect** | **Candidate SBC** |  | **Compare SBC** |
| **Entry** | delinq\_amnt | -7159.4840 | > | -7167.4513 |
| **Removal** | annual\_inc | -7142.2892 | > | -7167.4513 |

In Figure 1, we see our parameter estimates for our variables. Some of our variables were removed dur to being obsolete in our model. Some of those variables were delinq\_amnt and pub\_rec.

**Figure 2: Stepwise Selection Survey**

| **Stepwise Selection Summary** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Effect Entered** | **Effect Removed** | **Number Effects In** | **SBC** | **ASE** | **Validation ASE** | **Test ASE** |
| **\* Optimal Value of Criterion** | | | | | | | |
| **0** | **Intercept** |  | 1 | -6270.2035 | 0.2224 | 0.2155 | 0.2164 |
| **1** | **loan\_amnt** |  | 2 | -6410.4068 | 0.2147 | 0.2085 | 0.2086 |
| **2** | **total\_pymnt** |  | 3 | -7040.3934 | 0.1842 | 0.1837 | 0.1916 |
| **3** | **open\_acc** |  | 4 | -7097.1526 | 0.1814 | 0.1835 | 0.1896 |
| **4** | **installment** |  | 5 | -7142.2892 | 0.1791 | 0.1793 | 0.1841 |
| **5** | **annual\_inc** |  | 6 | -7167.4513\* | 0.1777 | 0.1778\* | 0.1839 |

Looking at figure 2, we see the variables that are significant. I used a stepwise selection to do this. Everything looks good overall.

**Figure 3: Selected Model for Loan Status**

**Selected Model**

**The selected model is the model at the last step (Step 5).**

|  |  |
| --- | --- |
| **Effects:** | Intercept loan\_amnt total\_pymnt open\_acc annual\_inc installment |

In figure 3, we have our predictor model for Loan Status. This can be used to see if it is a good or bad loan.

**Summary:**

From the results of this predictive model, we can see that the overall best predictors for loan status is loan amount, total payment, open account, annual income, and installment. We can use this selected model in the future to determine which are good or bad loans.

**Code:**

data donepaper;

set work.finalstatsdata;

overall\_status=.;

if loan\_status='Fully Paid' then overal\_status=0;

if loan\_status = 'Does not me' then overall\_status=0;

if loan\_status= 'Charged Off' then overall\_status = 1;

run;

proc glmselect data=work.donepaper;

model overall\_status = loan\_amnt dti total\_pymnt open\_acc total\_acc pub\_rec annual\_inc delinq\_amnt installment;

partition fraction(test=0.25 validate=0.25);

store out=work.realdata;

run;