­Trust Rating Logic (Continued)

## Data:

The data is from 2011-07-1 to 2011-12-31. There are about 644616 loans and the number of unique customer id is about 209043, so average customer applied 3 loans. However, we did not want to look at each customer’s behaviour affect our analysis, so I remove the customer whose have committed more than 3 loans with us. The criteria I used to remove those redundant customers is based on their application day of second loan. We will keep the data with the closest application day of second loan as data may shift over the time. However, this may lead to biased sample.

After remove some data with missing value, there are 209039 records remaining, those form our data set.

## Features:

In addition to the features we extract from database, I also generate some new features based on those available features. There are as follows:

**AmountOrigDiff**<-SecondLoanAmountOriginal -FirstLoanAmountOriginal;

**ScoreDiff**<-SecondMixedScore - FirstMixedScore;

**utilization1**<-FirstLoanAmountAgreed/FirstTrustRating;

**utilization2**<-SecondLoanAmountAgreed/SecondTrustRating;

**utilDiff**<-utilization2 - utilization1;

**IncomeDiff**<-SecondAnnualIncome - FirstAnnualIncome;

**utilRatio**<-utilization2/utilization1;

**trratio**<-SecondTrustRating/FirstTrustRating;

**utitrratio**<-utilRatio/trratio;

**logutitrratio**<-log(utitrratio);

**FirstLoanIncome**<-FirstLoanAmountOriginal/(FirstAnnualIncome/12;

**SecondLoanIncome**<-SecondLoanAmountOriginal/(SecondAnnualIncome/12);

## Data Analysis:

First of all, I look at the correlation between the ThirdBad(which is our target variable regarding to the arrear45 of the 3rd loan).

|  |  |
| --- | --- |
| Features | ThirdBad |
| trratio | 0.143327437 |
| SecondDaysSinceLastPayOff | 0.192283574 |
| FirstDaysSinceLastProposal | 0.014239612 |
| utilization2 | 0.177176136 |
| ScoreDiff | 0.039581859 |
| SecondDaysSinceLastProposal | 0.083722571 |
| utilization1 | 0.11117426 |
| AmountAgreeDiff | 0.099965761 |
| utilDiff | 0.074491548 |
| AmountOrigDiff | 0.101861475 |
| SecondDelinquency | 0.051321347 |
| FirstDelinquency | 0.036579505 |
| SecondLoanTerm | 0.217151376 |
| logutitrratio | 0.050502792 |
| SecondTotalExt | 0.085577062 |
| utilRatio | 0.037741304 |
| utitrratio | 0.030808262 |
| FirstTotalExt | 0.057150413 |
| IncomeDiff | 0.002533562 |
| FirstAnnualIncome | 0.022267267 |
| SecondAnnualIncome | -0.02154682 |
| FirstLoanTerm | 0.144315086 |
| SecondTotalTop | 0.083445673 |
| SecondLoanIncome | 0.162488337 |
| FirstTotalTop | 0.040963846 |
| FirstLoanIncome | 0.107635759 |
| SecondLoanAmountAgreed | 0.130766342 |
| SecondLoanAmountOriginal | 0.131970262 |
| SecondMixedScore | 0.328719764 |
| FirstLoanAmountAgreed | 0.069672619 |
| FirstLoanAmountOriginal | 0.070134362 |
| FirstMixedScore | 0.277383603 |
| SecondTrustRating | 0.004800172 |
| FirstTrustRating | 0.041830106 |
| FirstLoanNumber | 0.094276883 |
| SecondLoanNumber | 0.094276883 |

Table 1 The Correlation between the Arrear45 and Features

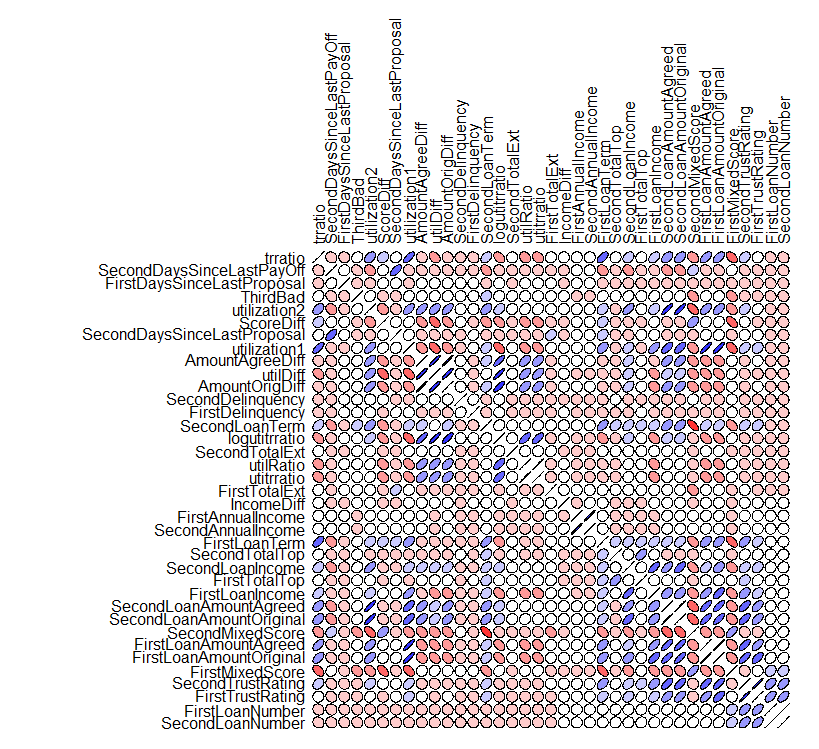
The figure 1 represents the similar information, but includes correlation within all the features.

Figure 1 the Correlations

To gain insight of the features with high correlation, we produce the following figure:

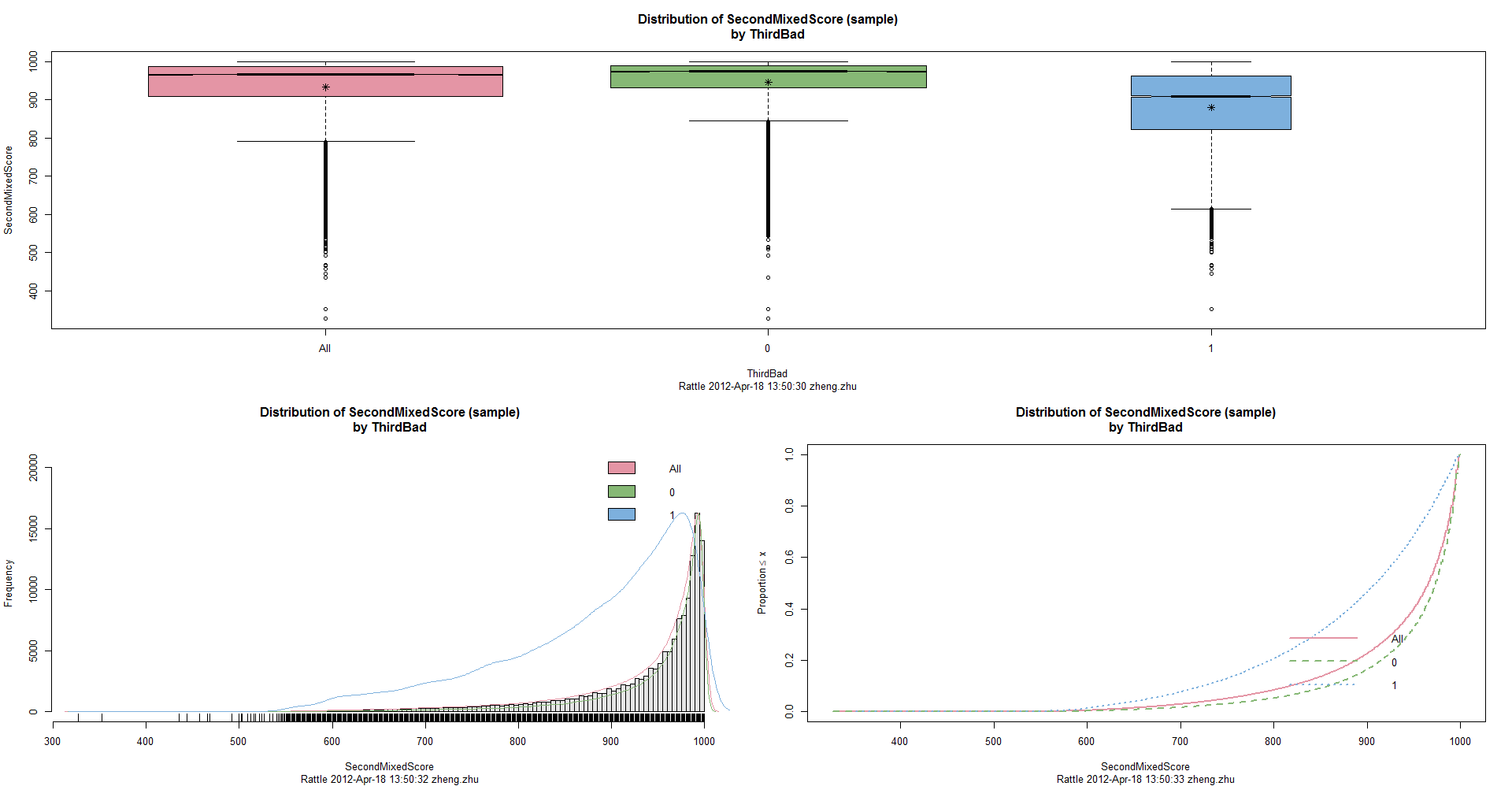


Figure 2 Second Mixed Score

From Figure 2, we can see the difference of distribution of Second Mixed Score regarding to the arrears45.

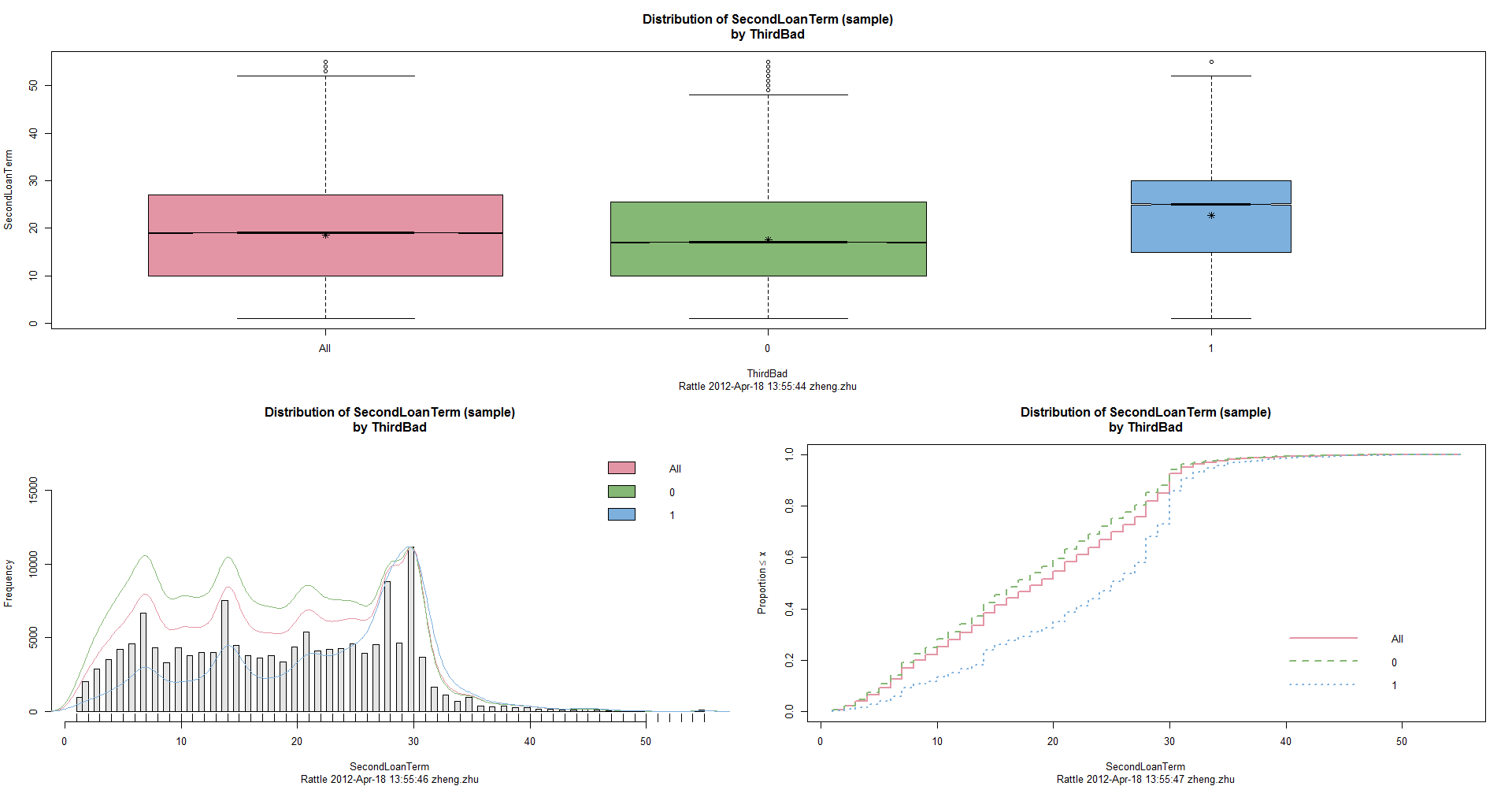


Figure 3 Second loan term

Figure 3 show the second loan term based on arrears45,

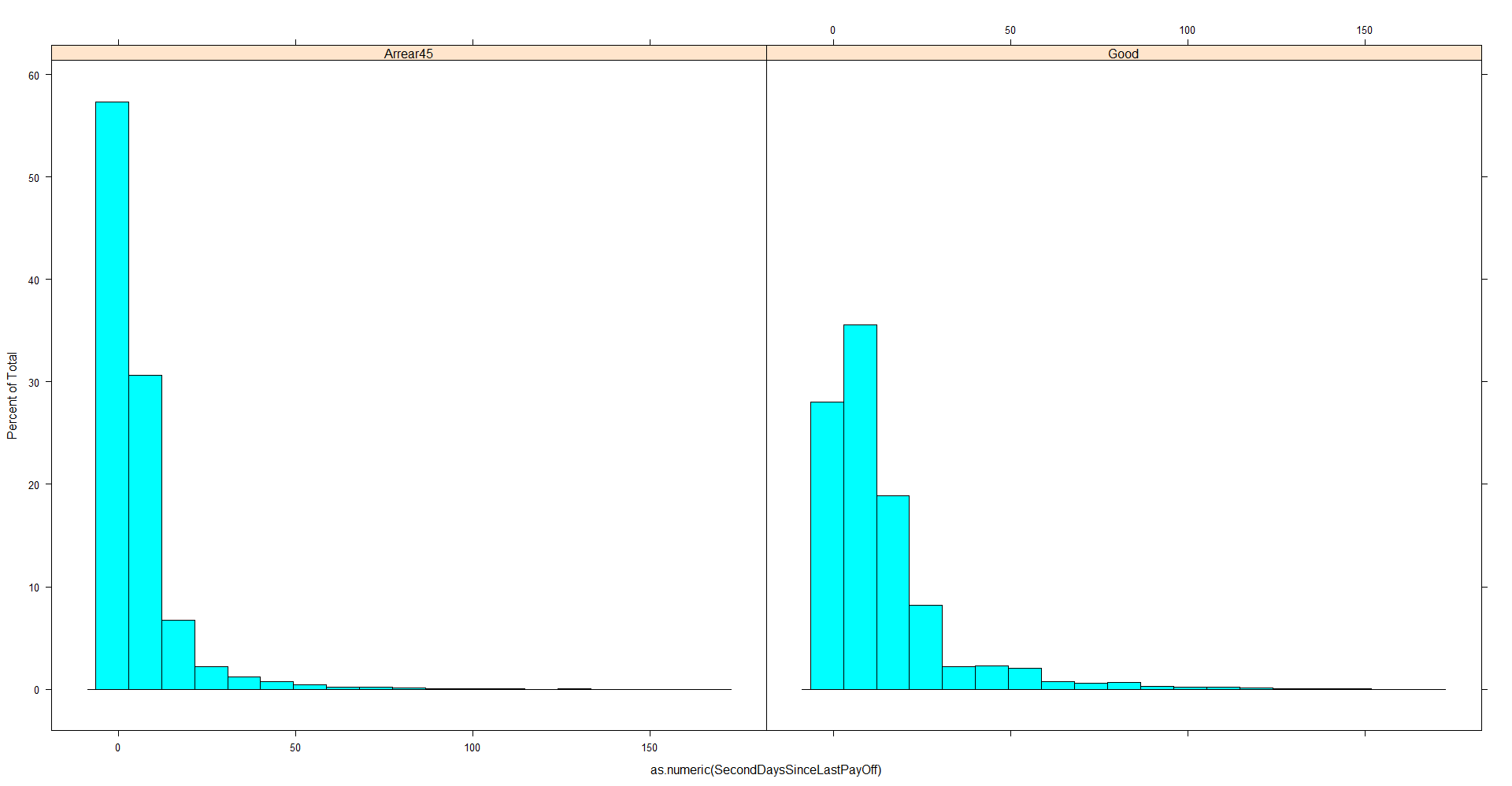


Figure 4 Histogram of Second Days Since Last Pay Off

Again, the histogram of Days Since Last Payoff is different for Arrears45.

Secondly, we will investigate the relation between the arrears45 and some combined features (i.e., the difference between 2 loans).

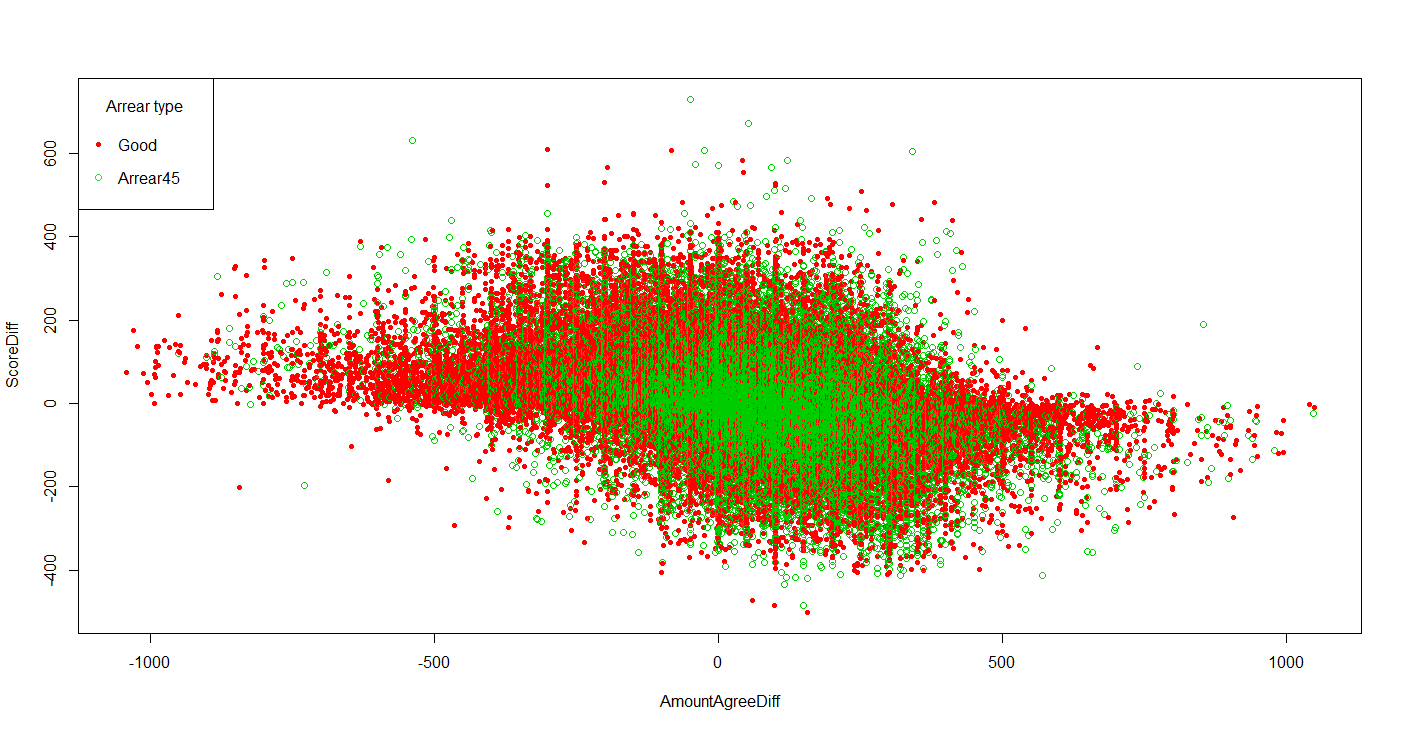


Figure 5 Amount Difference vs Score Difference

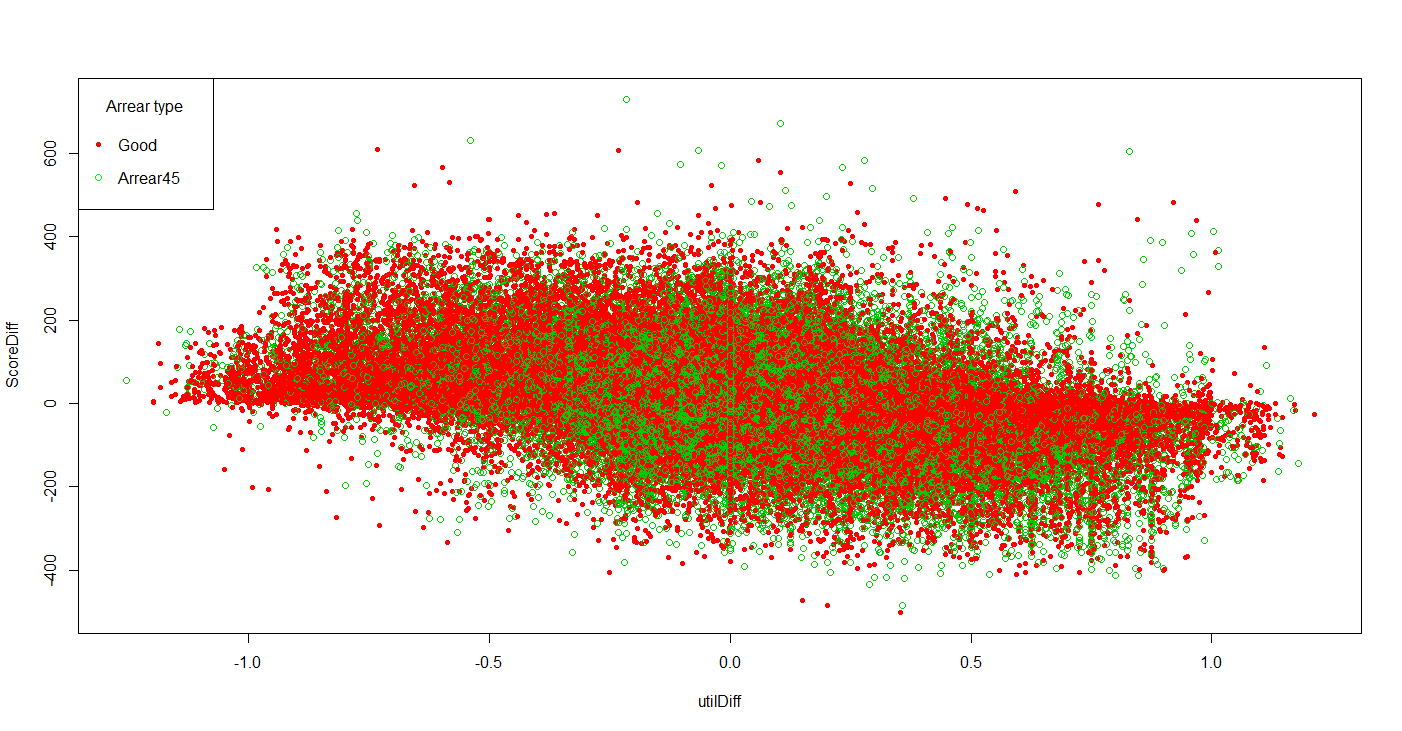
Figure 5 shows the scatter plot of amount agree difference of previous two loans vs mixed score difference of previous two loans, Majority of the green dots (Arrears45) are located on right of amount agree difference 0.

Figure 6 Utilization Difference vs Score Difference

Figure 6 does not display any patterns significantly.

Next we look at the utitrratio (which defines in features section) with mixed score difference. Notice that here we look at the log utitrratio because a large proportion of utitrratio is within the range [0, 1], using log function will be easy for visualization purpose.

Figure 7 shows the scatter plot of the two variables we mention above. We can see that upper figure with pink dots (0, denote a good loan) has more data than bottom figure (1, denote arrears45) for x-axis of less than -2 and greater than 2.

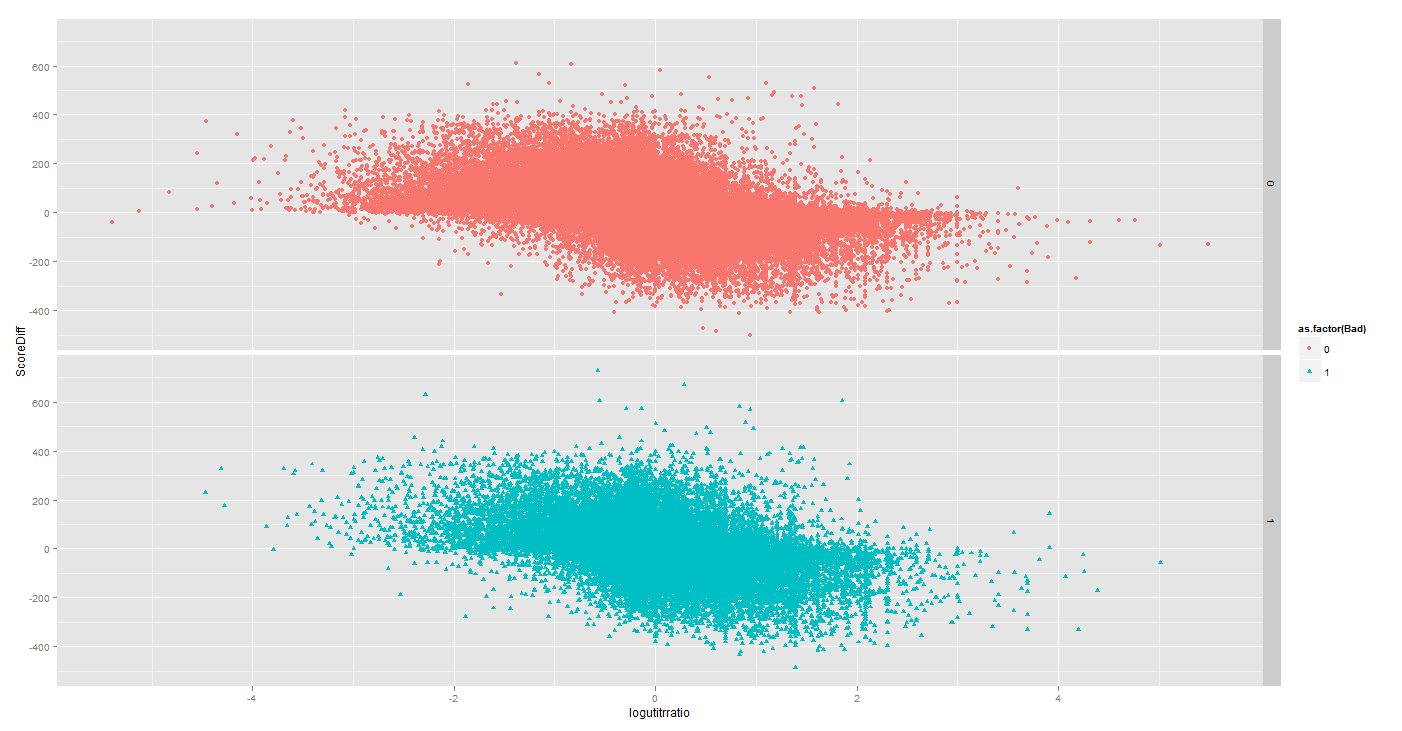


Figure 7 Log of Utilization trust rating ratio vs Score DIfference

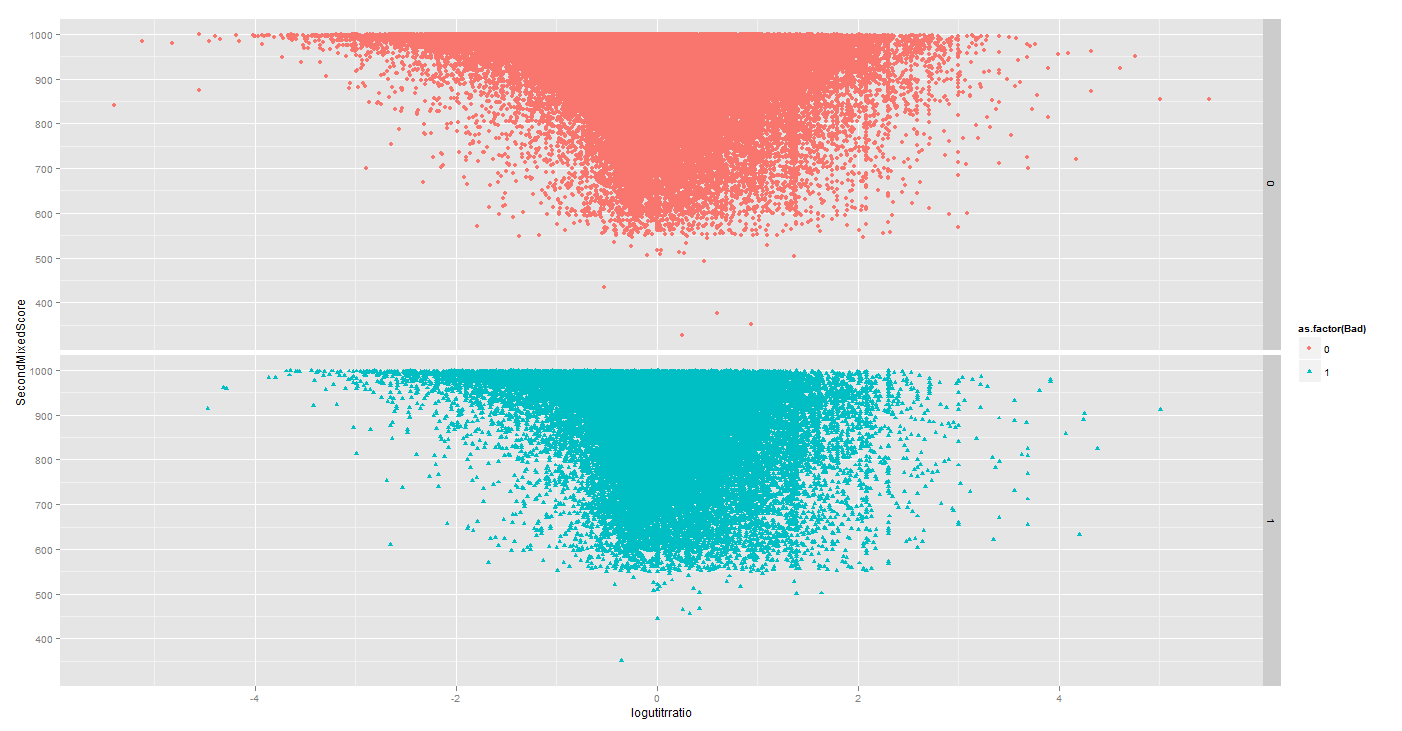


Figure 8 Log of Utilizatio trust rating ratio vs Second Mixed Score

Figure 8 represent the relation between Log of Utilizatio trust rating ratio and Second Mixed Score.

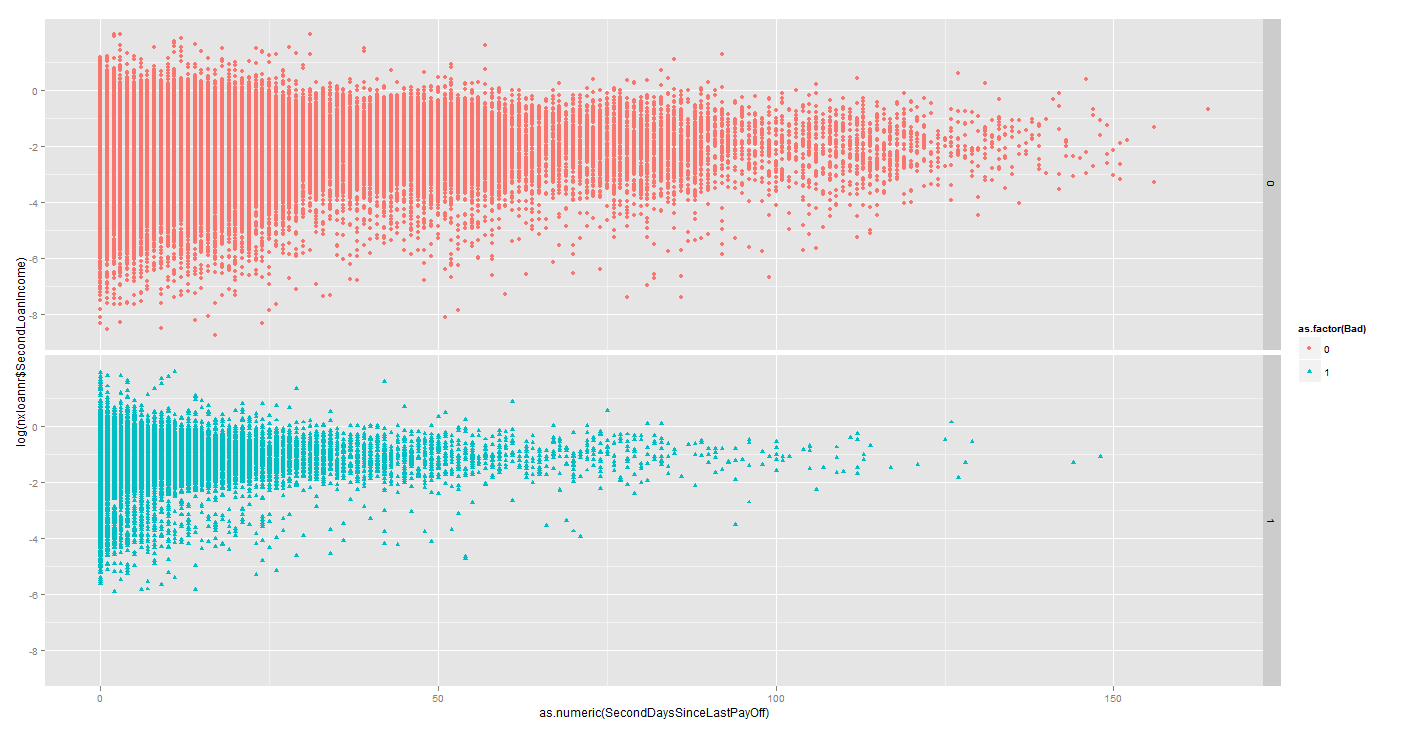


Figure 9 Second Days Since Last PayOff vs log Second Loan Amount/Income ratio

Figure 9 shows some interesting phenomena about Second Days Since Last Payoff and log(Second Loan Amount/Income). We can see that there are some difference between the good loan (Bad = 0) and arrears45 (Bad=1).­­­ We can see that for arrears45, majority of them are at y>-2, while for good loan, the case is opposite. Moreover, when the Days Since Last PayOff is greater than 100, there are more good loans.

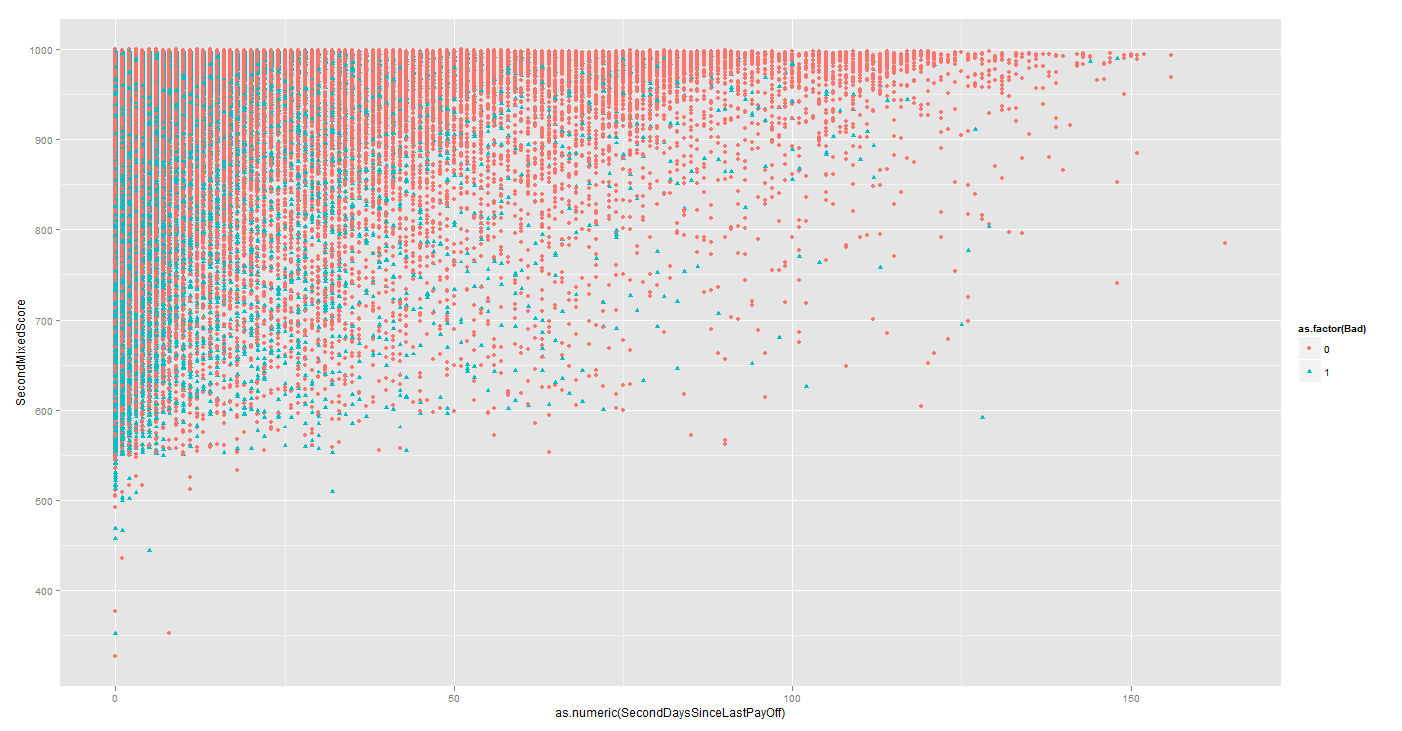


Figure 10 Second Days Since Last PayOff vs Second Mixed Score (Combined)

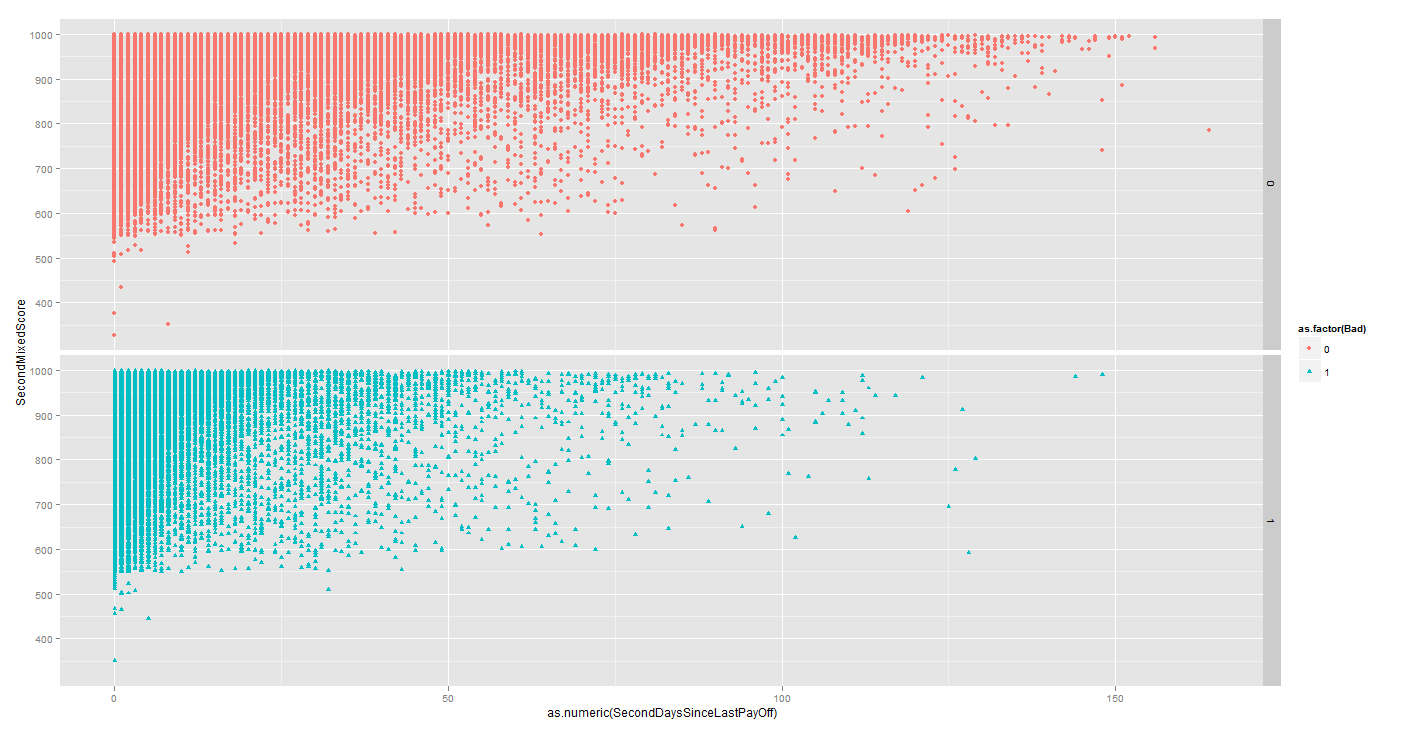


Figure 11 Figure 10 Second Days Since Last PayOff vs Second Mixed Score (Seperated)

Figure 10 and 11 shows the Second Days Since Last PayOff vs Second Mixed Score, however, we did not detect any pattern regarding to Second Mixed Score.

## Models:

Using random Forest with 150 trees and 6 variables.

OOB estimate of error rate: 18.78%

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | class.error |
| 0 | 101715 | 4569 | 0.0429886 |
| 1 | 20361 | 6124 | 0.7687748 |

Summary of the Ada Boost model:

Call:

ada(ThirdBad ~ ., data = crs$dataset[crs$train, c(crs$input,

crs$target)], control = rpart.control(maxdepth = 30, cp = 0.01,

minsplit = 20, xval = 10), iter = 50)

Loss: exponential Method: discrete Iteration: 50

Final Confusion Matrix for Data:

Final Prediction

True value 0 1

0 112794 3590

1 24473 5470

Train Error: 0.192

Out-Of-Bag Error: 0.202 iteration= 6

Additional Estimates of number of iterations:

train.err1 train.kap1

50 1

Variables actually used in tree construction:

[1] "AmountOrigDiff" "FirstAnnualIncome" "FirstDaysSinceLastProposal"

[4] "FirstDelinquency" "FirstMixedScore" "FirstTrustRating"

[7] "ScoreDiff" "SecondDaysSinceLastPayOff" "SecondDaysSinceLastProposal"

[10] "SecondDelinquency" "SecondLoanTerm" "SecondMixedScore"

[13] "SecondTotalTop" "SecondTrustRating" "trratio"

[16] "utilDiff" "utilization2"

Frequency of variables actually used:

SecondMixedScore SecondDaysSinceLastPayOff FirstTrustRating

40 38 17

SecondDelinquency FirstMixedScore SecondDaysSinceLastProposal

15 13 7

FirstAnnualIncome FirstDelinquency SecondLoanTerm

5 5 5

SecondTotalTop SecondTrustRating trratio

5 5 4

FirstDaysSinceLastProposal utilization2 AmountOrigDiff

2 2 1

ScoreDiff utilDiff

1 1

Time taken: 24.52 mins

## Analysis of Current Trust Rating Logic:

One comment from Anna is to slow down the speed of trust rating. One question will be what the speed of trust rating look likes and which part of speed we should consider to slow down.

To answer this question, we looked at the L0 loan happened between 2011-01-01 to 2011-07-01,there are 178909 records, amongst those customers who committed L0 loan, we find one subset of them who have finish more than 30 loans with us. There are 523 Customers in total left. And those customers form our dataset. And we looked at the first 30 loans from those customers.

First of all, we look at the trust rating for each different loan number. We looked at the mean level and median level since we need aggregate all the trust rating together.

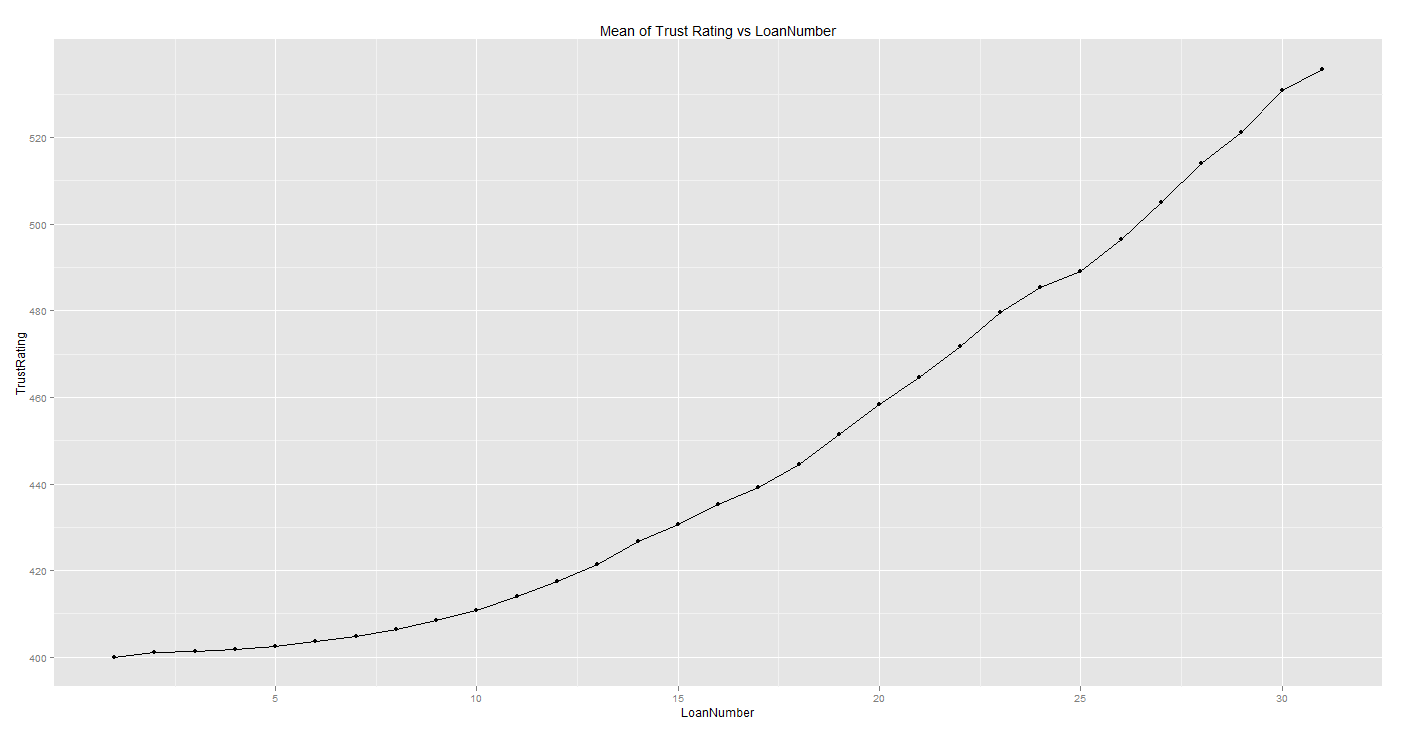


Figure 12 Mean of Trust Rating Update with More than 30 Loans

Note that those customers are good customers as they stay with us and do business with us. We can still see that for the first 15 loans, the increment is very slow. After that, the velocity increases to linear. However, mean may be distorted by the outlier. So we also look at the median of the trust rating update. We can see that the velocity of trust rating is very slow in this case.

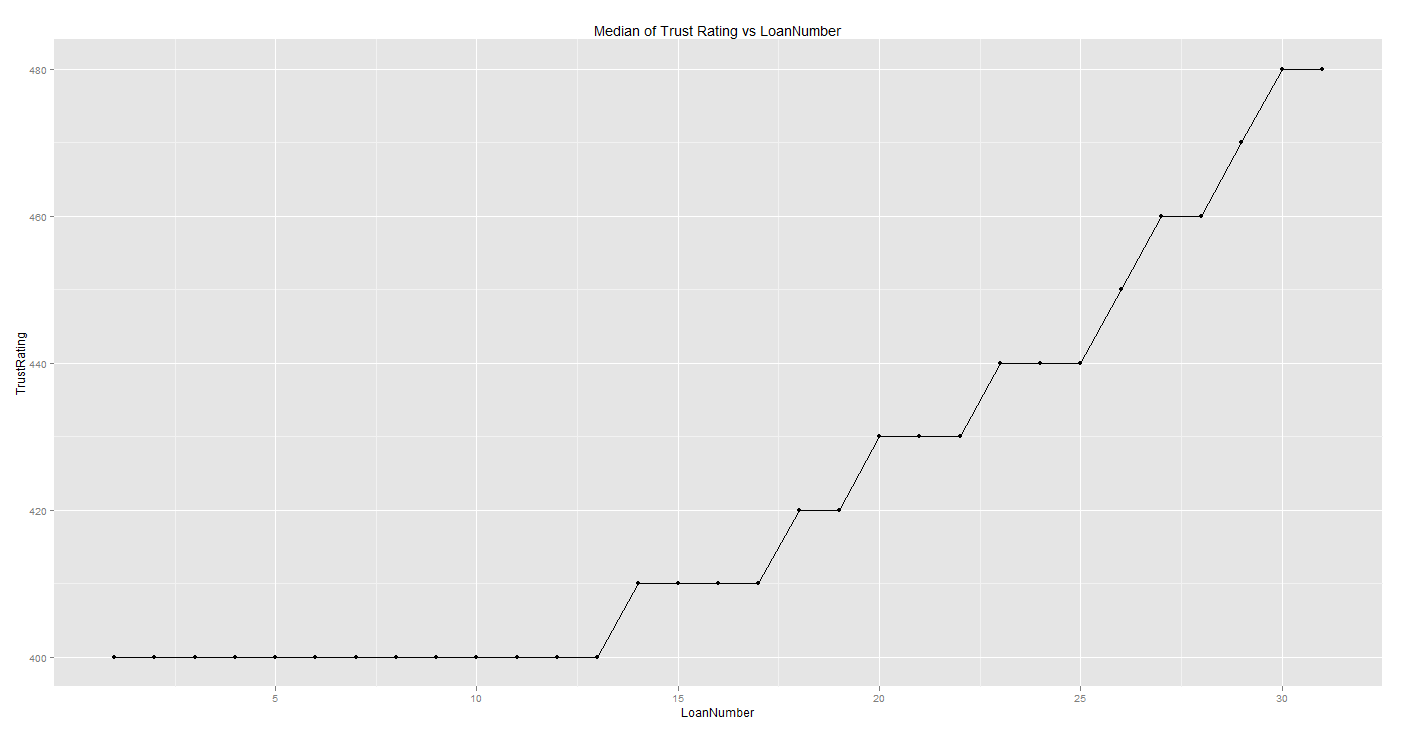


Figure 13 Median of Trust Rating Update with more than 30 loans

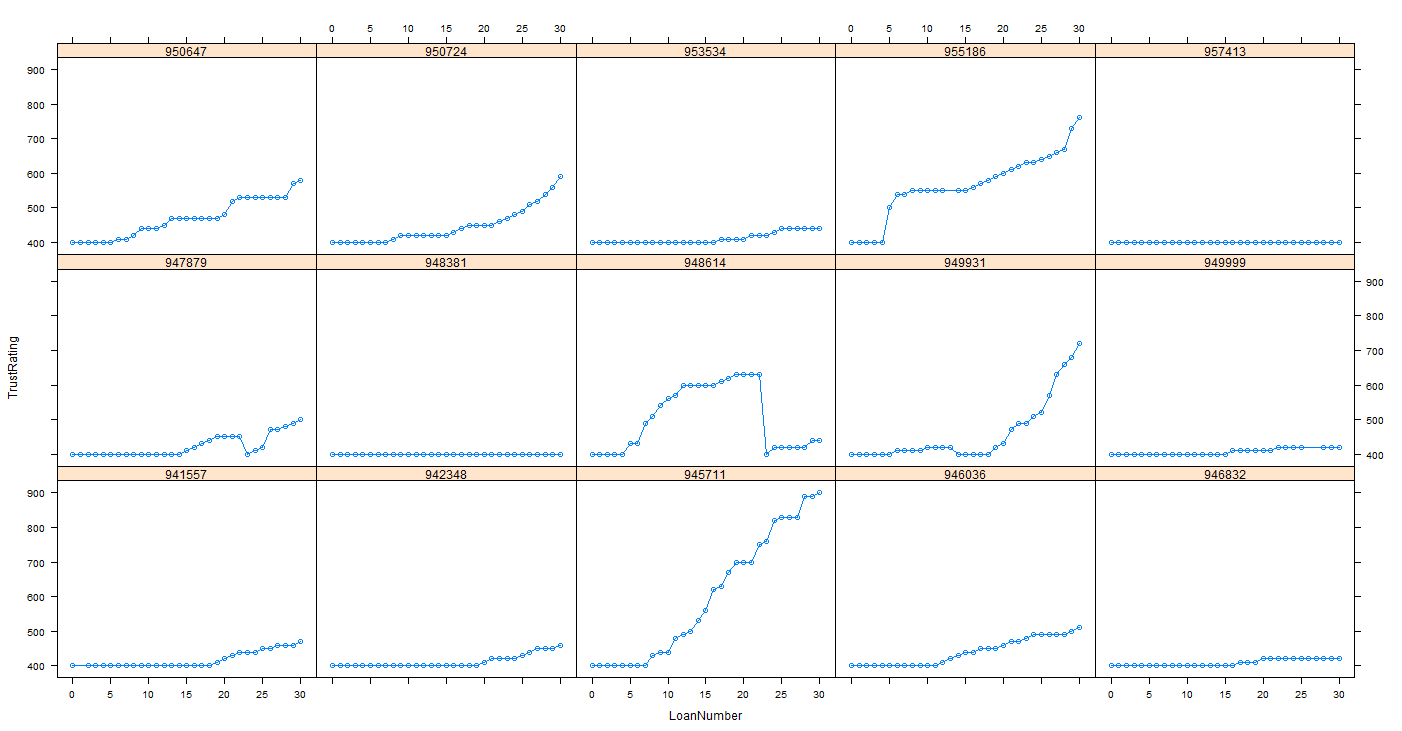


Figure 14 Trust Rating Path with more than 30 loans

According to the current trust rating logic, the increase of the trust rating will be a quarter of trust rating given the mixed score is higher than a certain threshold. We can estimate the upper bound of the number of loan for one good customer reach the maximum trust rating (1000).

**400X1.25^n=1000**, **n =** 4.106284, therefore the best case for one good customer reach trust rating 1000 is 5 ignoring the interest factor. If we reduce the 1.25 to 1.1, then the customer reach trust rating will be 10 if we ignore the interest.

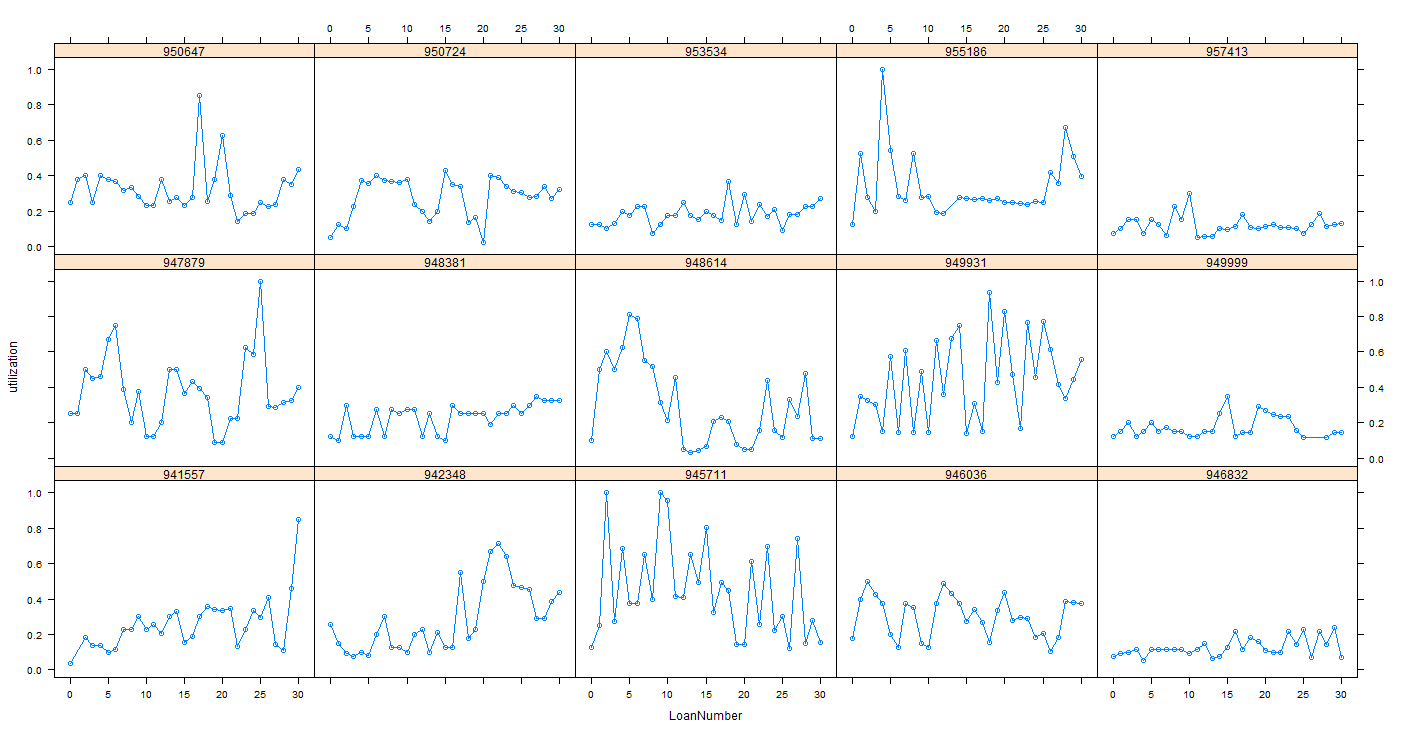


Figure 15 Utilization Along Each Application

We also look at the customer with 30 loans who reach 1000 trust rating and their trust rating trend and utilization trend is shown below:

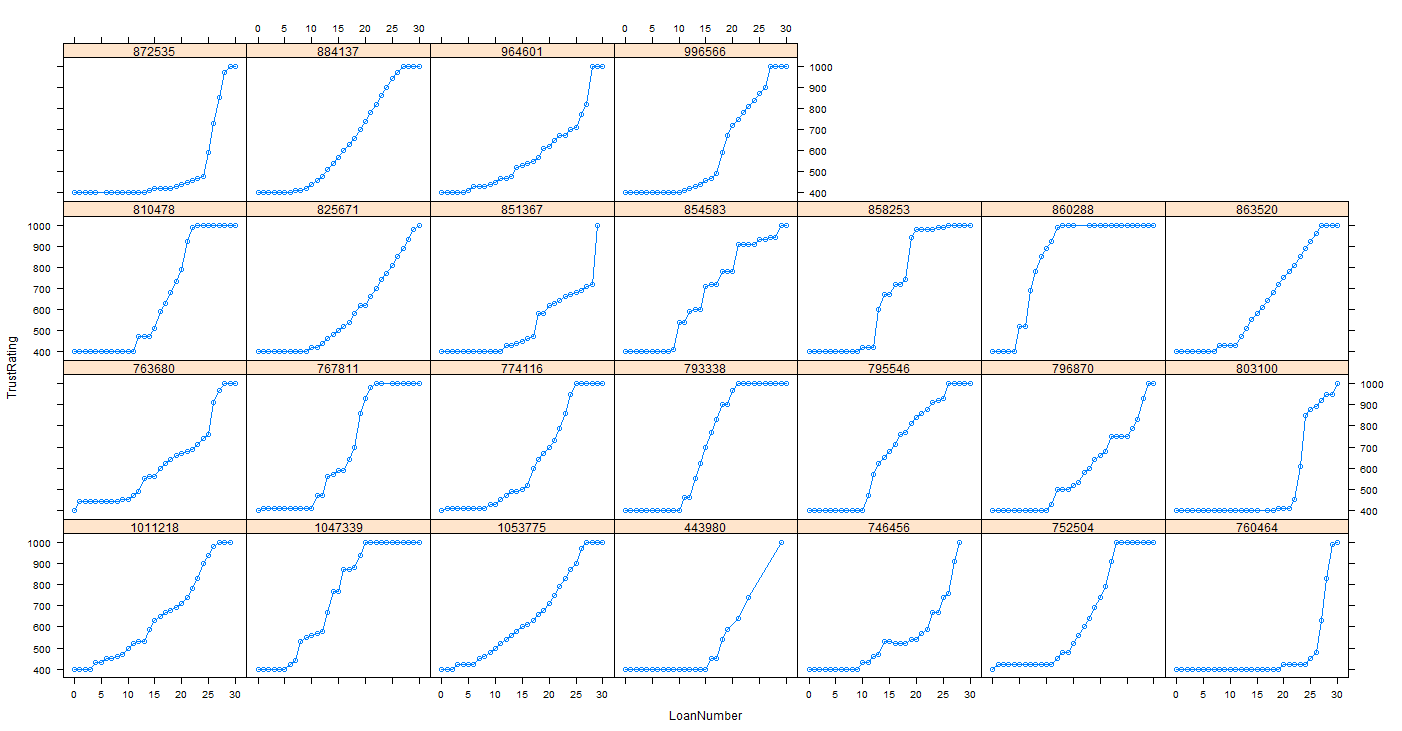


Figure 16 Trust Rating Path with more than 30 loans (reach 1000 TR)

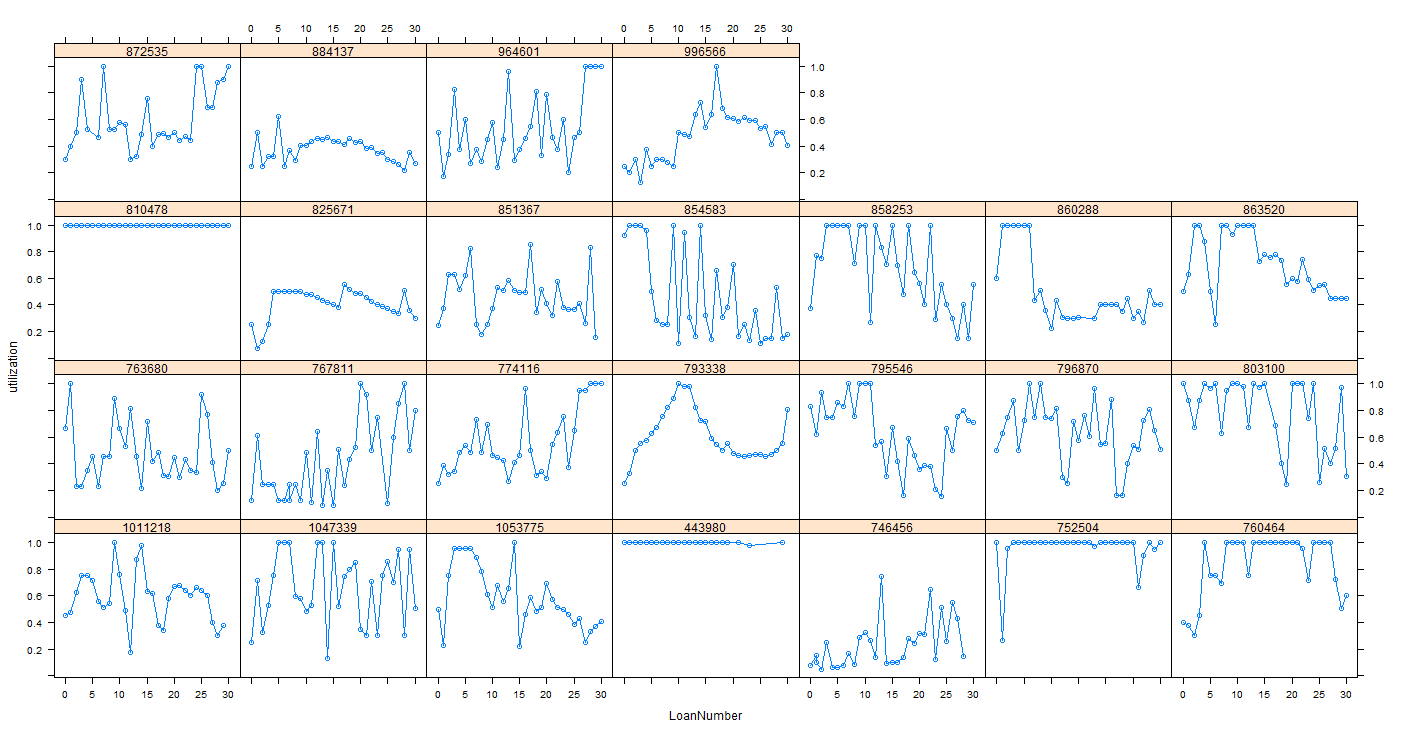


Figure 17 Utilization Along Each Application (reach 1000 TR)

To look at how long it will take the customer to reach the 1000 trust rating on overall data set rather than those who have committed 30 loans with us.

The data I use is from 2011-01-01 to 2011-07-01 as before, in this case, I only look for the case whose trust rating of next loan is 1000 and trust rating of the previous is less than 1000. This gives us 521 customers in total. Note here we only consider those customers who have eventually reach 1000 trust rating and the difference from the previous data is previous data I only look at the customers who committed more than 30 loans with us, ignoring whether or not they will reach 1000 trust rating.

The mean of the loan number reach 1000 trust rating from that dataset is 12.48 and the standard deviation of that is 5.58. The min of the loan number reach 1000 trust rating is 5 and the max is 69. The mode and median are both 12.

Combine the previous data, we can say it is very likely that customers committed more loans with us usually relatively slow velocity in terms of trust rating compare to those customers who reach 1000 trust rating.

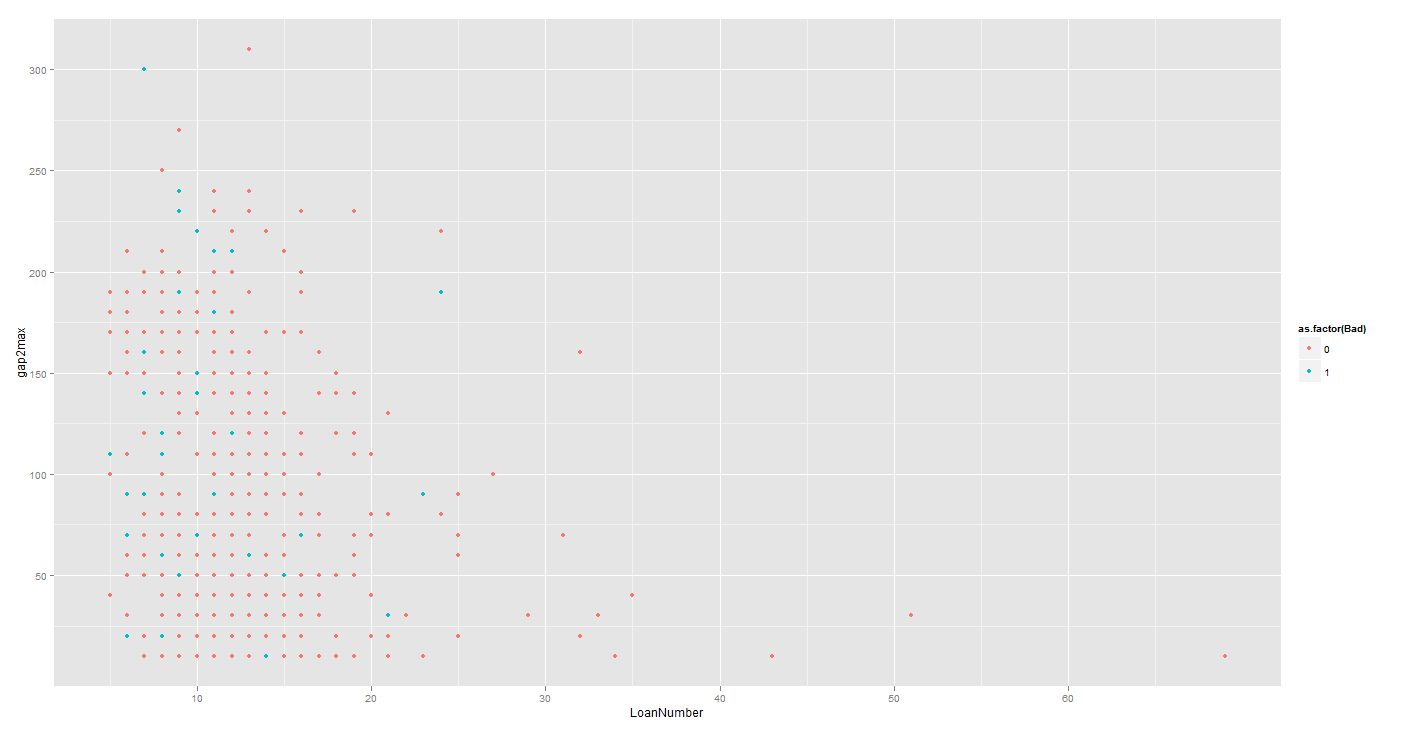


Figure 18 Loan Number to Reach 1000 and Gap to 1000

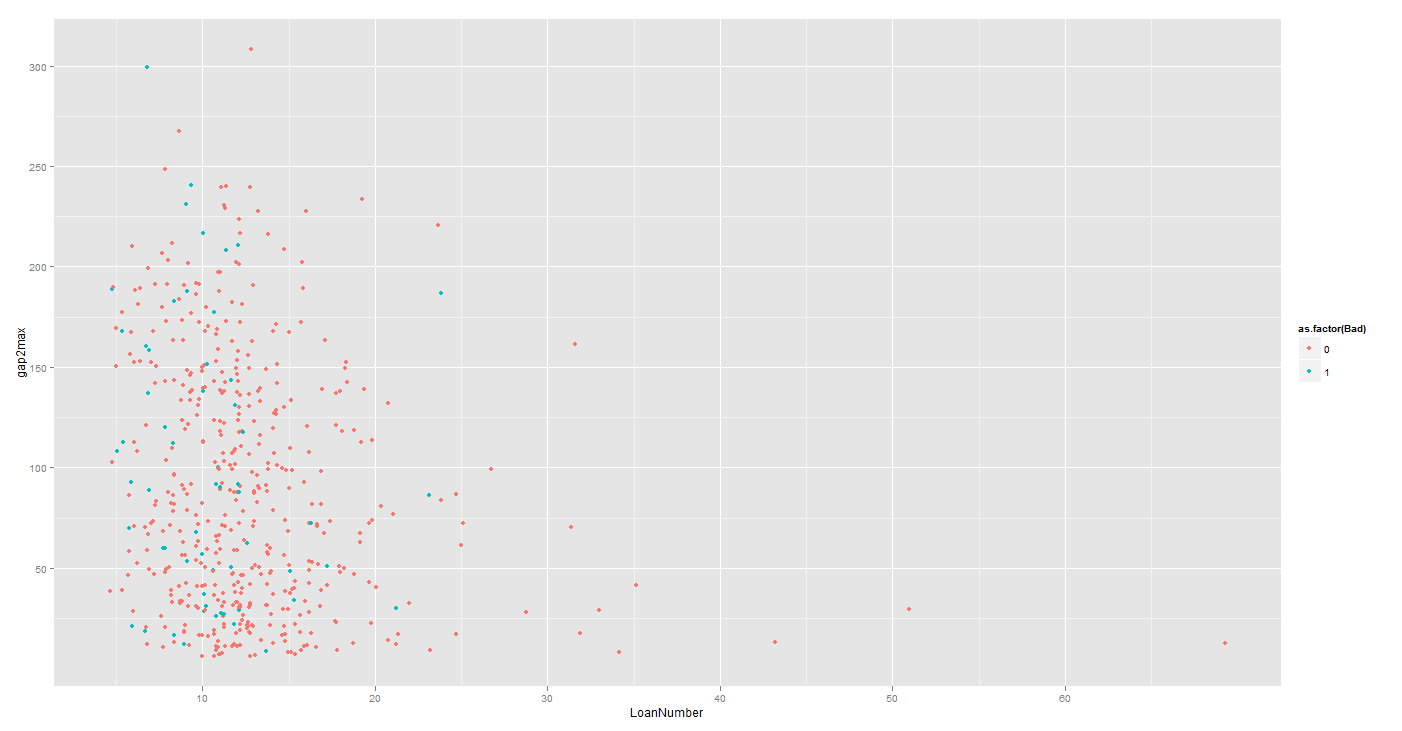


Figure 19 Loan Number to Reach 1000 and Gap to 1000 (Jitter)

How do those customers reach 1000 trust rating so quickly? I take one subset of these customers and look at the trust rating path as Figure 20 shows:

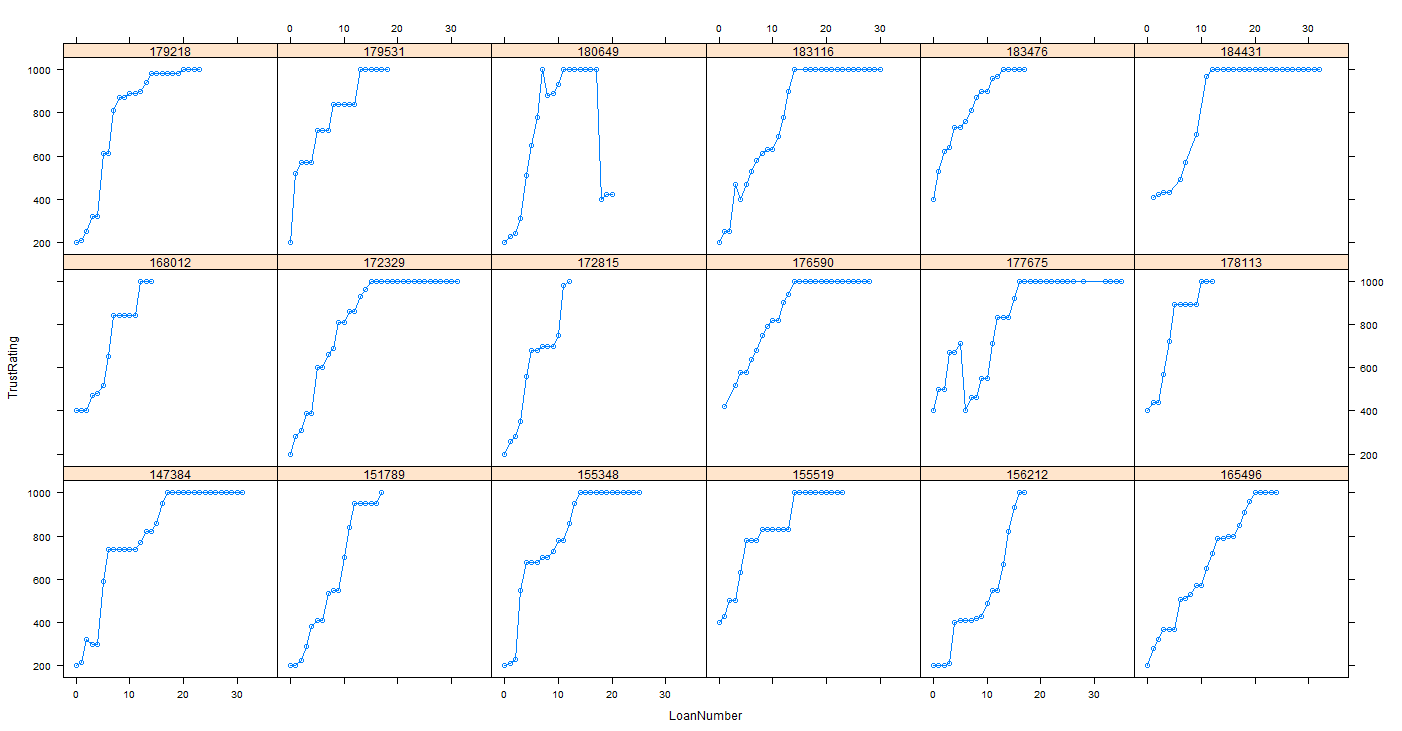
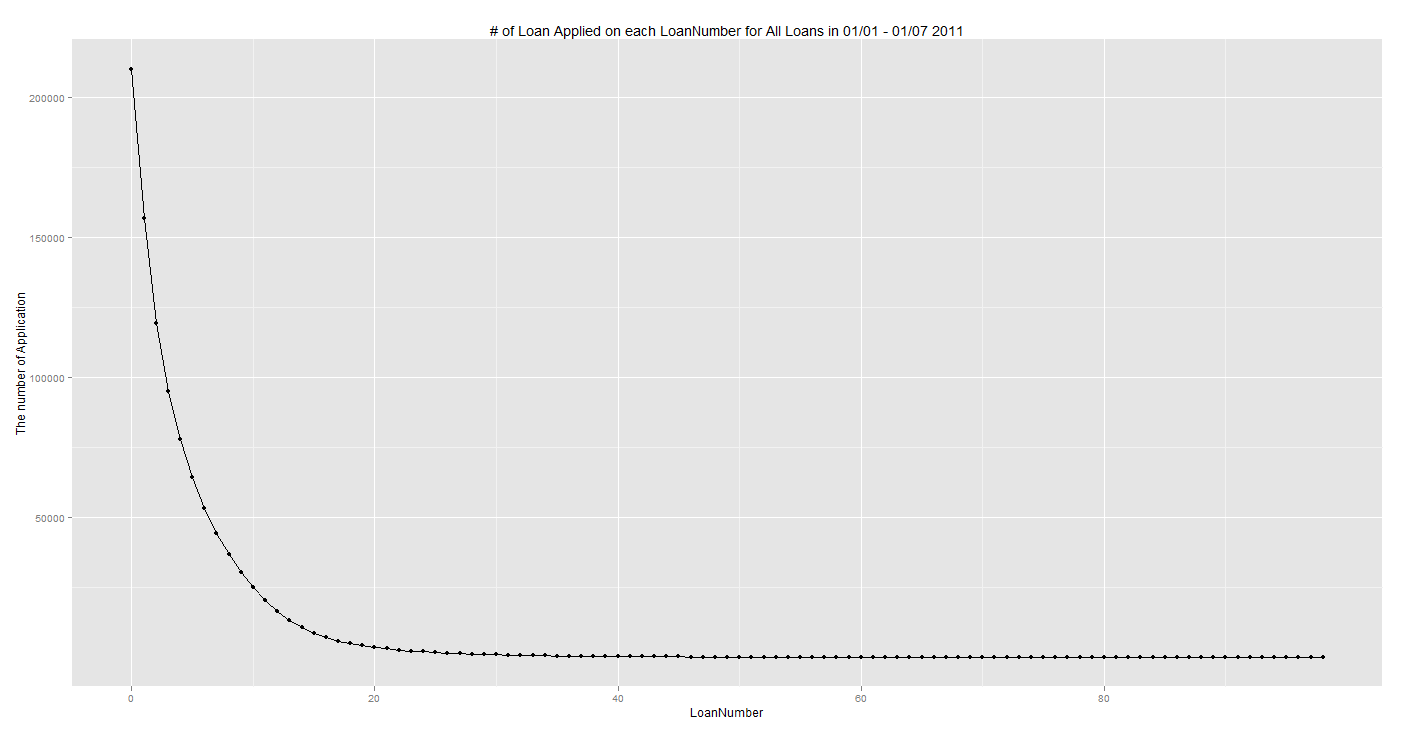


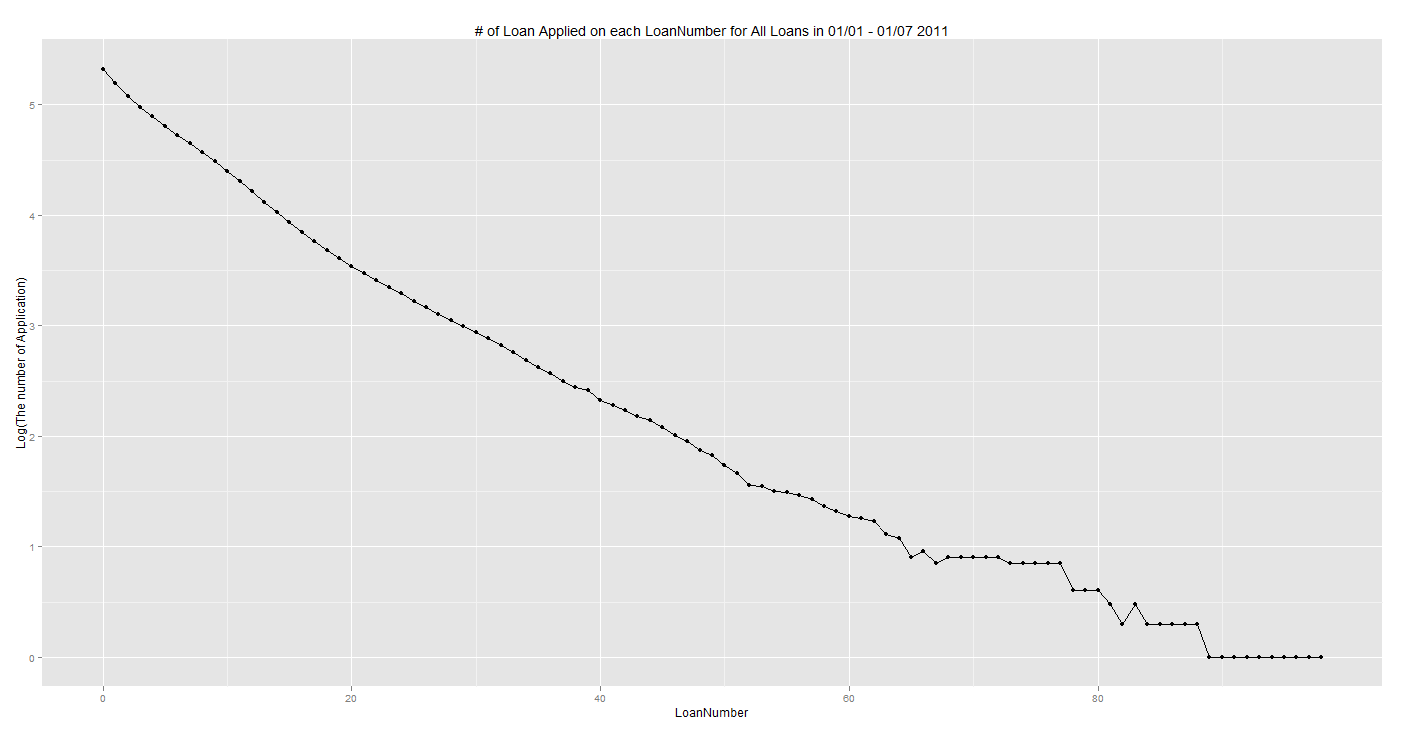
Figure 20 the path to 1000 Trust Rating

This leads us to another question, how fast the velocity of trust rating in general, which includes all customers no matter how many loans they are with us and how the trust rating they are at last. Note the number of applications will be one non-increasing function since the customers can only carry on with us at Ln or drop out from us.

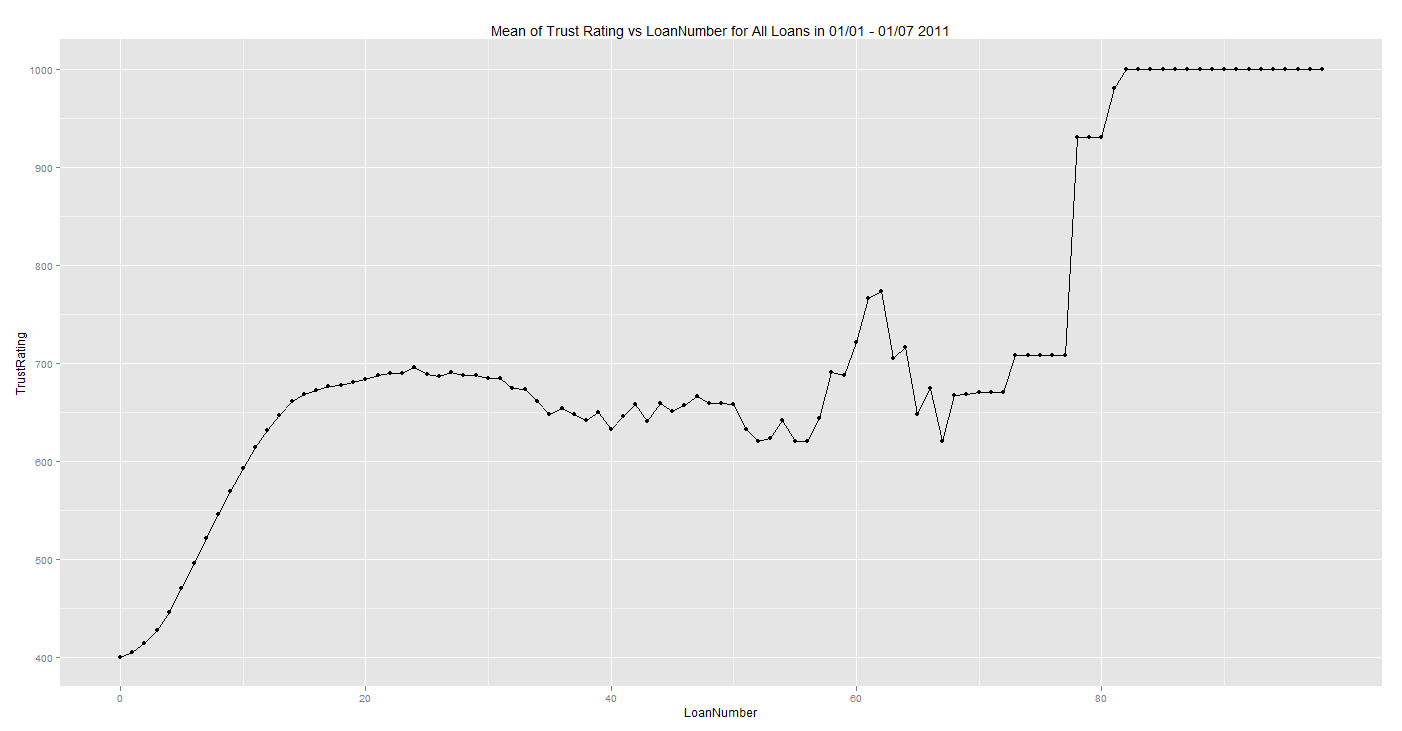
The number of loan applied on each loan number for all 2011-01-01 to 2011-07-01 is shown below:



To make the distribution more clear, we take the log on Y with base 10 and plot the following figure.



In comparison, the average trust ratings on each loan are shown in below figure.



From the above analysis, we can see that if the customer reach 1000, it will be very likely to be there within 12 loans. Of course they are good customer in most cases. If the customers have more loans(i.e.,Ln>30) with us, then the velocity of trust rating is usually slower even than overall trust rating velocity(overall trust rating is define as the average trust rating for Ln customer in one period, the higher Ln, the less customers). Both more loans and quick 1000 are good in general, but quick 1000 are more risky from my perspective (maybe not).