Data Challenge - Quantitative Analyst Dealstruck, Inc

Zongyan Wang

March 29, 2016

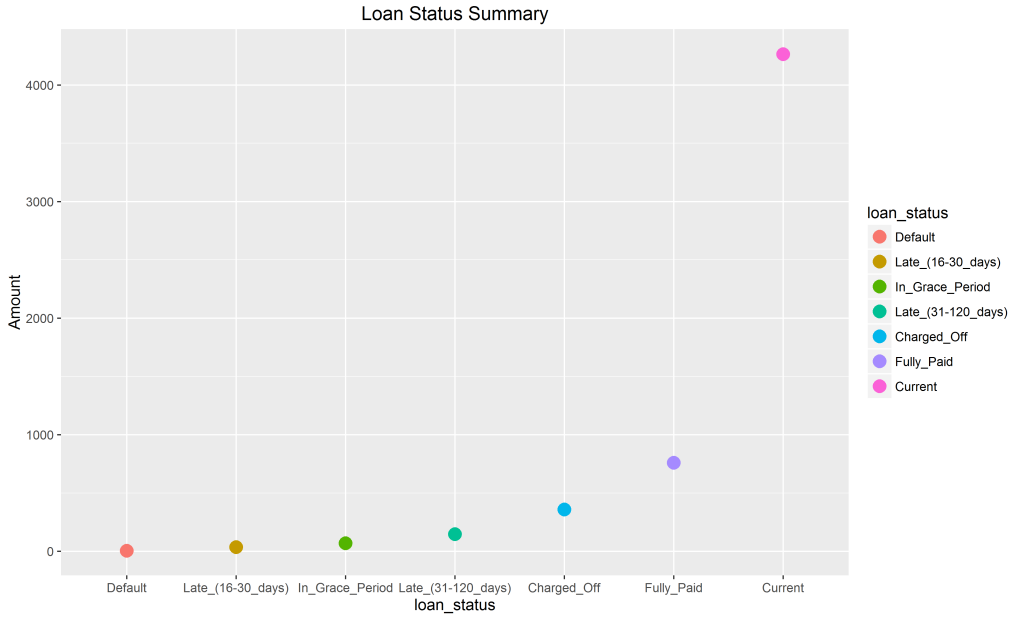
Abstract

Using the LC_biz_all.csv data, we are interested in the best predictors for delinquency events in small business loans, which could be measured by the variable "loan_status" in the data. The goal is not only prediction but also interpreting the model for important variables. We could fit a perfect model if we model "loan_status" with the rest of variables. However, if we delete the variables which are directly correlated to the delinquency events, there's no clear boundary for the levels of response variables. I have tried logistic regression, Kernel SVM, Re-sampling, ROC curve, changing cost function, EM algorithm for censoring data to solve this problem.

1 Summarize cleaning, exploratory analysis, visualizations

1.1 Data type overview

The data contain 5461 rows, and 56 variables. There are 10 variables include missing values, 14 variables are categorical variables, in which emp_titile, zip_code, earliest_cr_line has more 400 levels. loan_status is a categorical variable we are interested in, which has 7 levels and contains survival data.

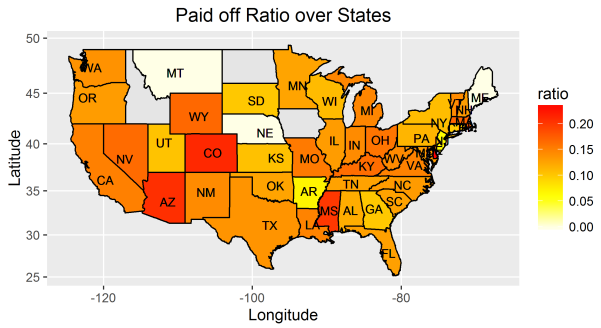


1.2 Data clean

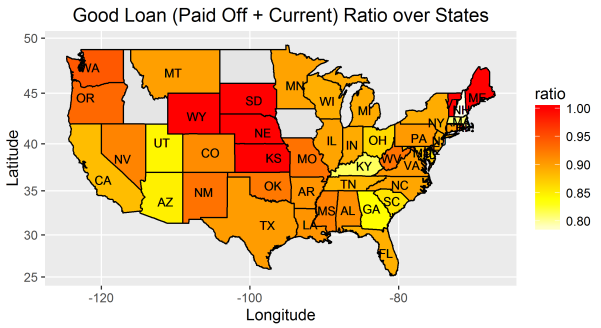
My data cleaning process includes transforming all the percentage to numbers, finding truncated data and missing data, and marking all the missing data to NA. I also replace all the blank space appearing in the data to "_", and transform variable earliest_cr_line from character to numeric value(amount of months). Those process help the model to read the numeric data correctly and locate the missing data.
I delete id, emp_title and zip_code since they have too many levels, which will increase the VC dimension dramatically.

1.3 Geographic information visualization

A visualization summary in terms of Geography.



(a) Fully Paid Ratio over US



(b) Good Loan Ratio over US

2 Analysis

2.1 Fit the loan status to all the rest of variables

I group the levels of response variables into 3 groups: "Current", "Fully Paid" and "Potential Delinquency"(Note: "Potential Delinquency includes the rest of 5 levels), We could use a logistic regression to distinguish each pair of those 3 groups.
To do so, I apply ANOVA on the pair of response variable and each numeric explanatory variable. I collect the variables with p-value lower than 0.1 and all the categorical variables. The important numeric variables are as follows:

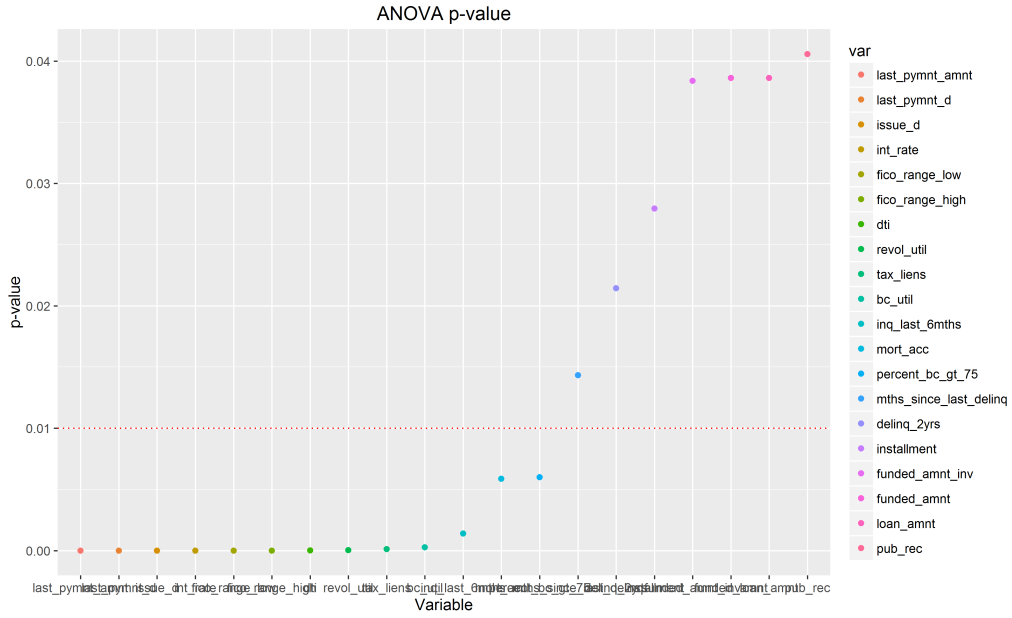


Figure 2: Numeric Variables with ANOVA p-value lower than 0.1

Fit logistic regression with those variables(important numeric variables and categorical variables), we get the result: (The result has been optimized with the use of ROC Curve)

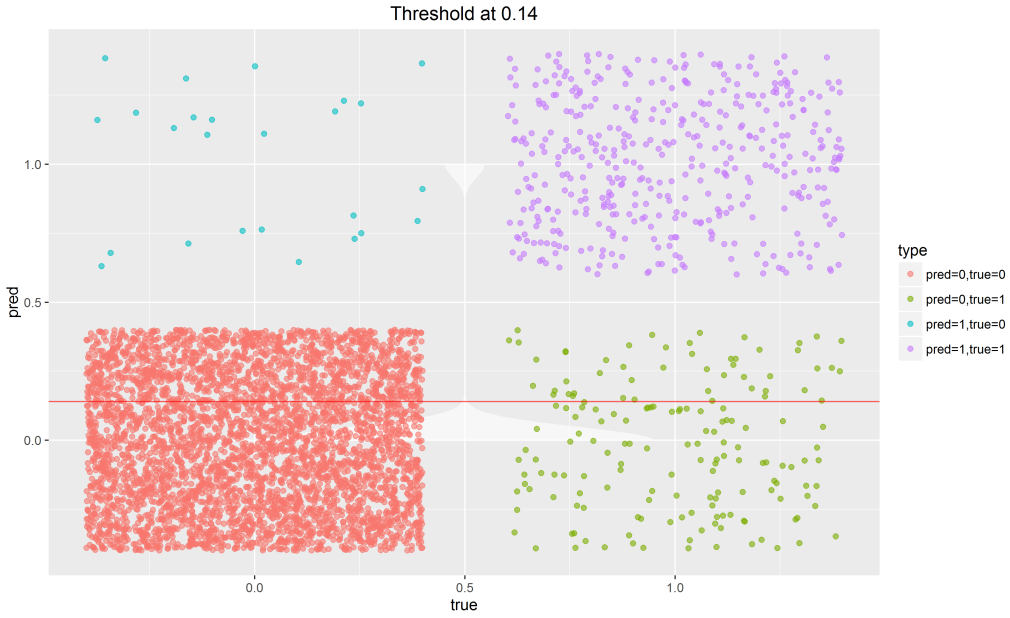


Figure 3: Comparison of prediction and true value for Current and Potential Delinquency

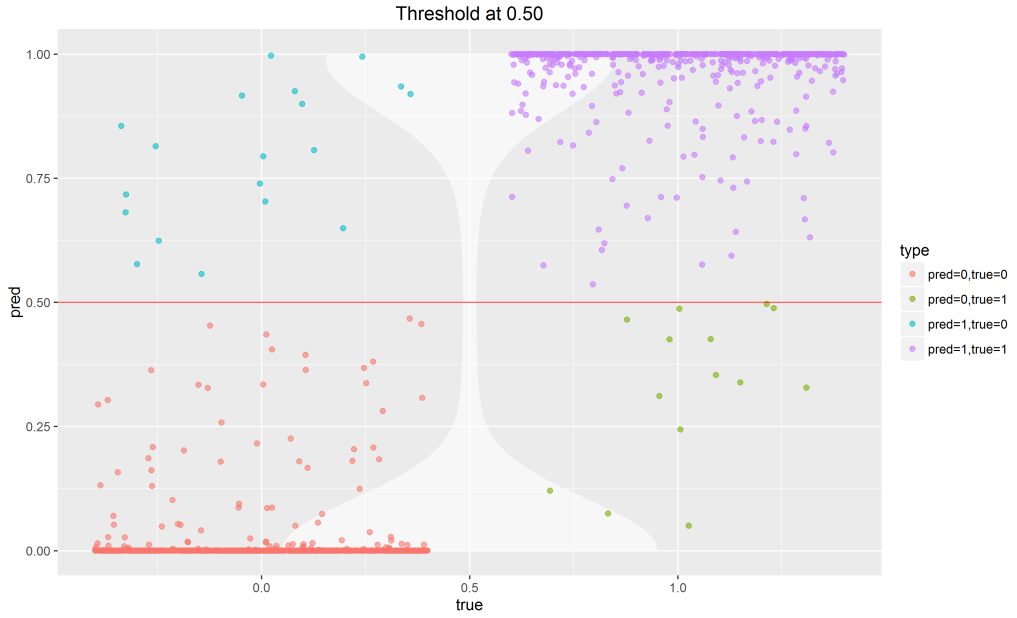


Figure 4: Comparison of prediction and true value for Fully Paid and Potential Delinquency

Both of them are very good fit.
We could also plot the normalized parameter estimated(parameter estimated/sd):

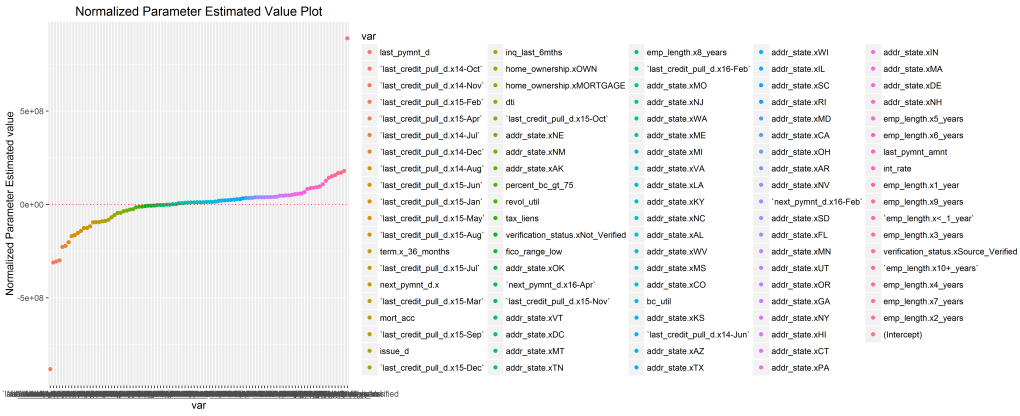


Figure 5: Normalized Parameter estimated for Current and Potential Delinquency

2.2 Fit the loan status to the variables interested

I notice variables "id", "emp_title", "last_pymnt_d", "next_pymnt_d", "zip_code", "issue_d", "last_credit_pull_d", "last_pymnt_amnt" are directly correlated to a delinquency events, and are not a characteristic of the candidate. Hence I would like to delete them and build the model again.
However, in this case, both of logistic regression and Gaussian kernel SVM perform poorly:

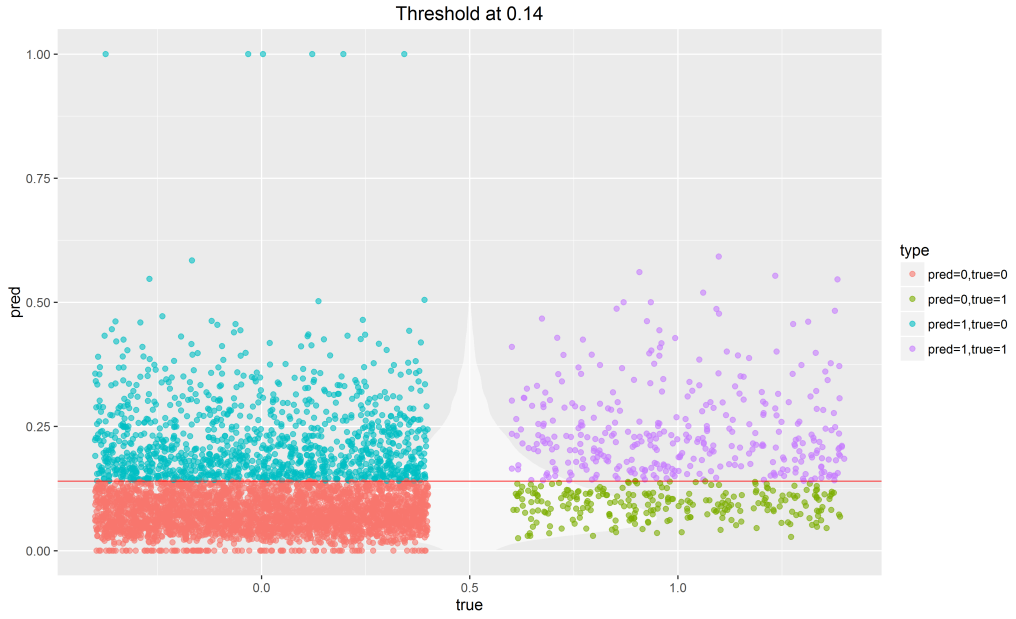


Figure 6: Comparison of prediction and true value for Current and Potential Delinquency(logistic regression)

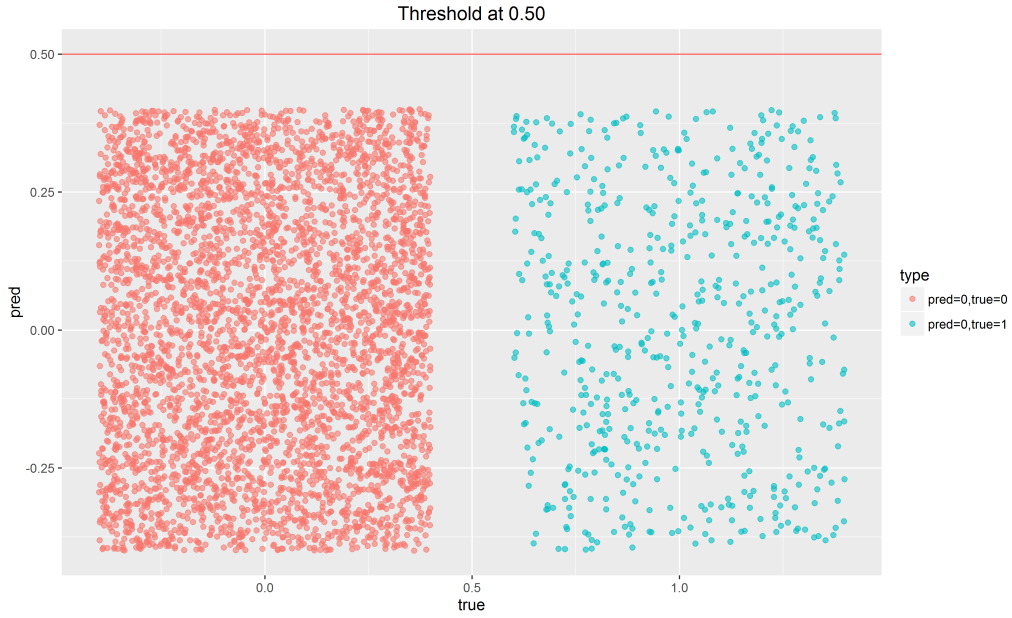


Figure 7: Comparison of prediction and true value for Cully Paid and Potential Delinquency(kernel)

2.3 Discussion to solve the problem

The goal is to distinguish the "Potential Delinquency" group from the rest of two groups. We meet the problem that after remove the directly relevant variables, the model doesn't perform good. To solve the problem, I have tried three ways(re-sampling, Change cost function/ROC curve, EM algorithm for censoring), however, none of them give a satisfied result. We could believe there will be no boundary(or pattern) between those 3 groups of data if we remove those directly relevant variables.

2.3.1 re-sampling

Both logistic regression and SVM are sensitive to the data size bias. In this data set, "Current" and "Fully Paid" dominate the data, so both of those methods will have a poorer performance. We could do bootstrap from the "Potential Delinquency" set to balance the data size of each group. However, this method doesn't give a better performance.

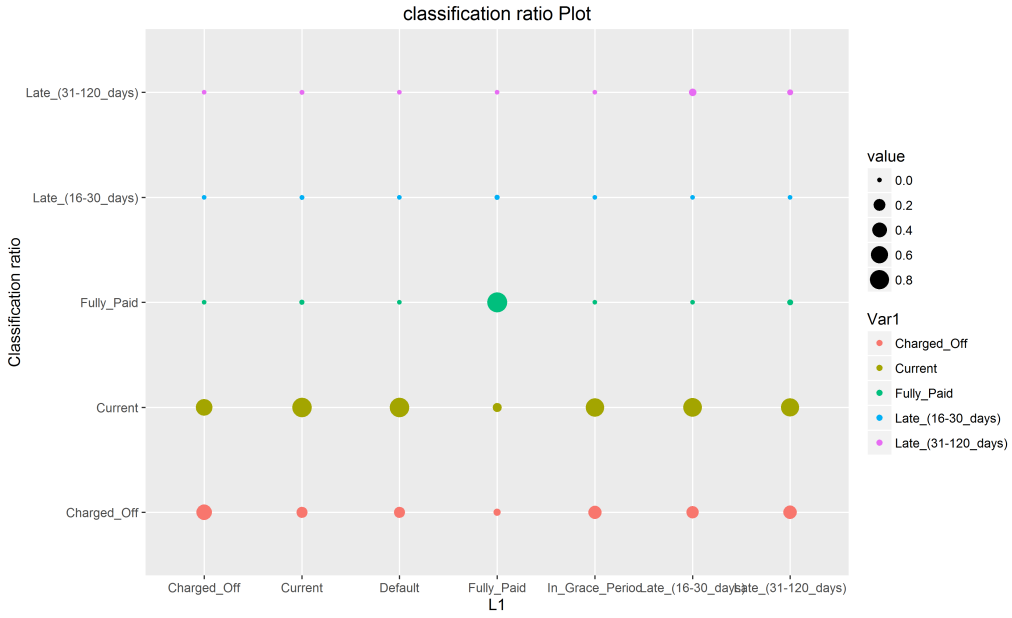


Figure 8: re-sample "Potential Delinquency" group before building the model, the result doesn't perform good

2.3.2 Change Cost function and ROC Curve

The next idea is instead of re-sampling from the group with small data set, we could change the cost function, or use ROC curve.

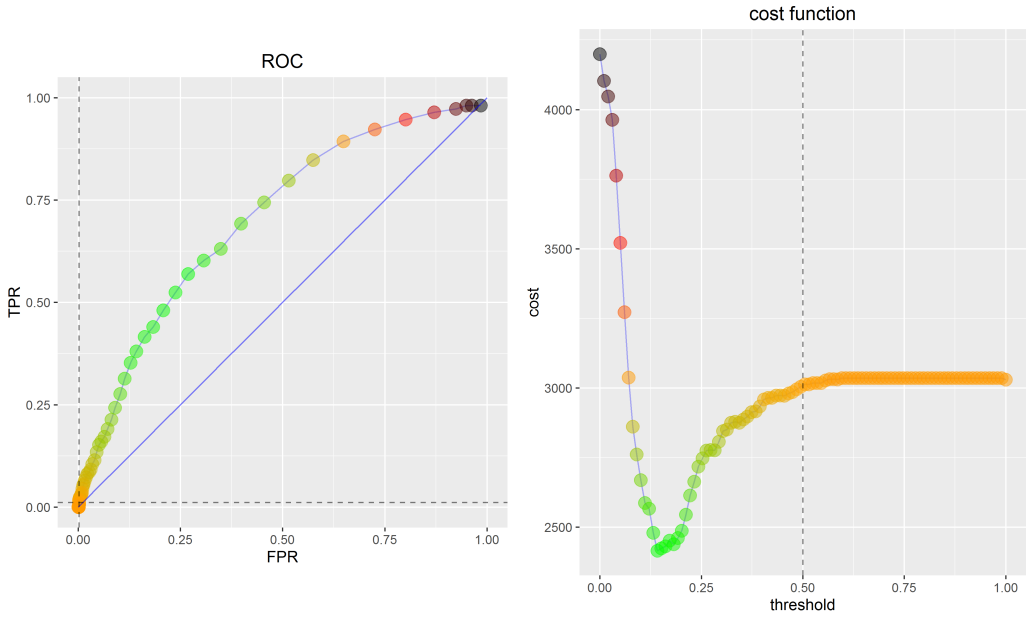


Figure 9: ROC Curve/Cost function(set cost Potential Delinquency 5 times as Current)

After we change the threshold to 0.14, which is the optimization point of the cost function, the result is shown at Figure 6.

2.3.3 EM algorithm for censoring data

Another idea is instead of using "loan status" and set the problem as a classification problem, we could use the length of loan payment. To do so, I transform issue_d and last_pymnt_d from character to numeric data, and use the difference of those two variables as our new response variable y_1 (time of "survival"). We need another response variable y_2 (hidden time of "survival"). A successful model will predict $y = y_1 + y_2$ (total "survival" time) longer than the "term"(36 months or 60 months) if "loan status" is "Potential Delinquency". Note y_2 is unknown but the expectation of y_2 could be estimated given the parameters and y_1. Assume the y_1 of "Charged Off" is 61 months (maximum of term plus 1 month), and y_2 of "Fully Paid" keep zero. I define the EM algorithm is shown as follows:

1. E-step: For a fixed θ^*, estimate $\hat{y}_2 = haty - y_1$ for "loan status" being

not "Fully Paid", if $\hat{y}_2 < 0$, then set $haty_2 = 1$,(need at least one more month to pay the loan). y_2 of "Fully Paid" keep to be 0.

2. M-Step: Fit the model logistic regression/Kernel SVM regression with y . and get the parameters estimated θ

There is no guarantee that this EM algorithm will convergence, so instead of giving a stopping rule, we do the EM 20 times. (The result shows the algorithm actually convergence):

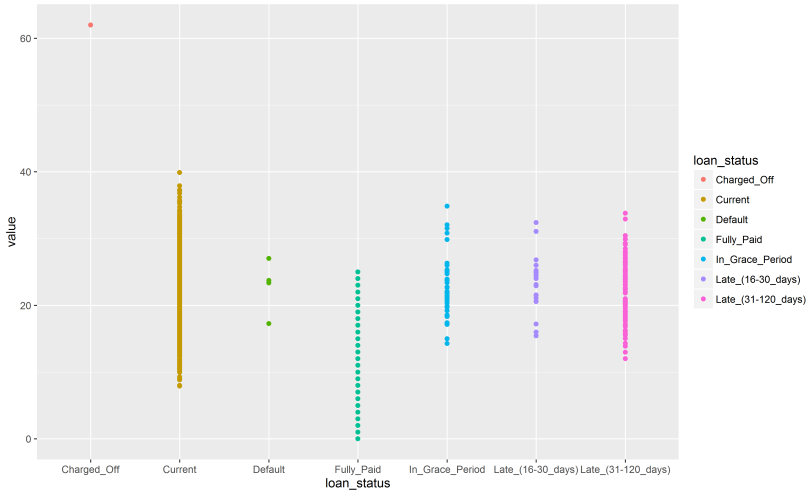


Figure 10: The total payment length estimated for "Current" and "Potential Delinquency Events"

The y_1 of "Charged Off" is set to be 61, and the EM estimate y_2 of "Charged Off" to be 1, so all of y of "Charged Off" is 62 in the plot. y of Fully Paid is equal to y_1 of "Fully Paid". The algorithm failed to distinguish "Potential Delinquency" from "Current".

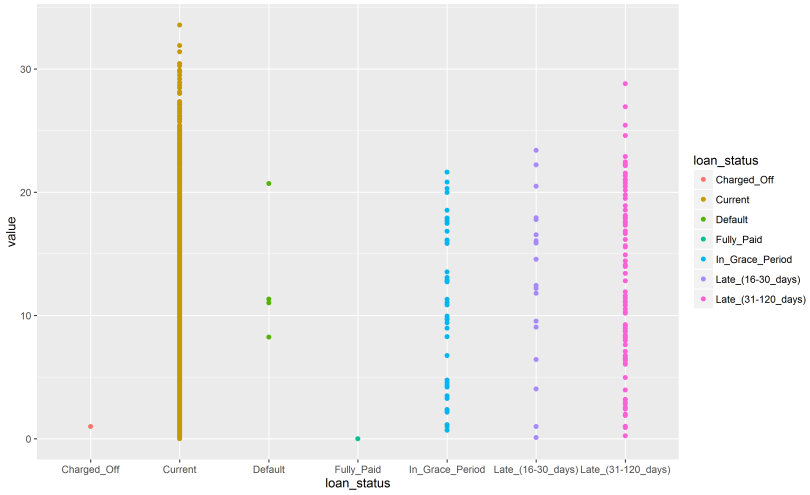


Figure 11: The "missing" part of payment length estimated for "Current" and "Potential Delinquency Events"

The algorithm failed to distinguish "Potential Delinquency" from "Current" in sense of the missing part of payment length.

3 Summary

It's very easy to build a model and distinguish the levels of "loan status" with all the variables. However, if we remove the variables correlated to the delinquency events and limit the variables to the variables correlated to candidates' characteristic, both logistic regression and kernel SVM fail to distinguish the potential delinquency events.
I have shown some popular methods nowadays trying to solve this problem, including re-sampling, ROC Curve, Changing cost function and even transform the problem from classification to regression(censoring data), and apply EM algorithm to estimate the length of payment. All of those methods failed to achieve our goal.
I would end here and give the statement that after removing the variables directly related to delinquency events, there probably be no clear pattern/boundary remained in the data to distinguish the delinquency events.