|  |
| --- |
| LSI FACO score reverse engineering |
| (Linear Regression) |
| Wang Peiyu*EMSE-6765* *Data Analysis* *For Engineers* *And Scientists*December 4, 2018 |

# Abstract

In order to help Loan Startup Incorporated (LSI) to increase their profitability, improve customer satisfaction and find opportunity to expand into the market with an “improved and enhanced Credit Score metric”, This paper analyzed loan application data to try to reverse engineer FACO score.

This paper includes the background introduction of LSI and the problem, the description of the data, model development, model validation, model improvement and forecasting with the regression model. It also provides conclusion and insights or concerns regarding the model developed.

# Introduction

Loan Startup Incorporated (LSI) is a startup company that provides crowdsourced loans to qualified individuals through an online portal, and to increase revenue, improve customer satisfaction and expand into related markets, LSI wants to develop an improved and enhanced Credit Score metric and develop a linear regression model based on Loan Regression Data, using the Log of FACO score, the model should be consist of the best predictors to represent the characteristics or behaviors of FACO score.

# Data and Methods

## Data Description

The original data consist of 14 columns of categories and 1299 rows of data. Among them, there are continuous data including annual\_inc, delinq\_2yrs, dti, emp\_length, inq\_last\_6mths, int\_rate, loan\_amnt, open\_acc, revol\_bal, revol\_util, total\_acc and faco, and also categorical data term and home\_ownership that would require special processing when analyzing.

## Approach

Because term is text data, so before analyzing, the corresponding data need to be transformed into something that can be analyzed. According to the Data Directory, term is either 36 months or 60 months, so another column called term\_36 is added to represent the term data. It has the categorical data type, so when it is 1, it means that the term for that specific row of loan is 36 months, otherwise, it is 60 months. As for another column of text data, home\_ownership, because there are three categories of home ownership for all the data, mortgage, rent or own. Therefore, two more columns of categorical data are added to represent the home\_ownership so that it can also be included in the analysis. One of the categorical data added is called home\_ownership\_rent, and it can be either 1 or 0. When home\_ownership\_rent is 1, it means that this specific row of data of loan is for the renting type. When it is 0, it means that this specific row of data of loan is either for mortgage or owning, and we need the second column of categorical data, home\_ownership\_own to help determine if it is for mortgage or owning. When home\_ownership\_own is 1, it means that this specific row of data of loan is for the owning type, otherwise, when both home\_ownership\_rent and home\_ownership\_own are 0, it means that this specific row of data of loan is for the mortgage type.

# Model Development

## Variable Selection

Since this linear regression is for reverse engineering FACO score, FACO score is the one that we are the most interested in, therefore, it is the dependent variable. However, with the help of the hint in the conversations and the normal distribution graph of FACO score, shown as Figure 1, we can see that the distribution of original FACO data is not very normal. In order to normalize it, we graphed log(FACO), and as the Figure 2 shows, log(FACO) has a more normalized distribution than FACO. Therefore, we use log(FACO) as the dependent variable instead of FACO.

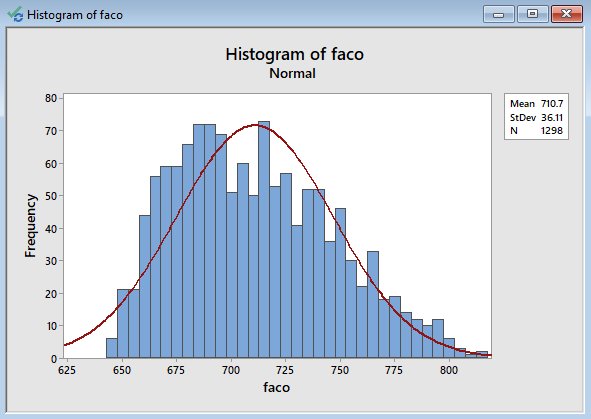


Figure 1: Distribution of FACO Data

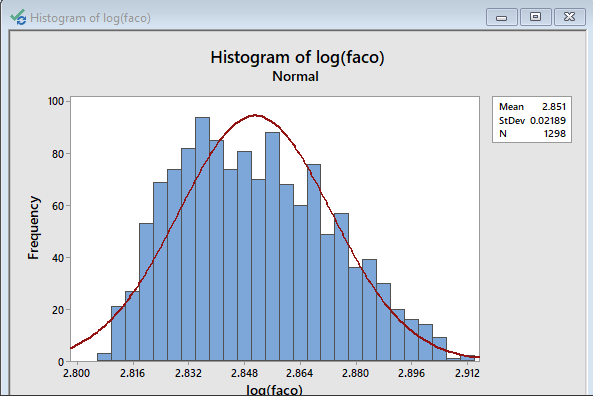


Figure 2: Distribution of Log10(FACO) Data

However, from the probability plot of faco, log(faco) and log(log(faco)), shown as figure 3, the P-values of all 3 are less than 0.005 which means that the data is abnormal, therefore a subset of data needs to be selected.

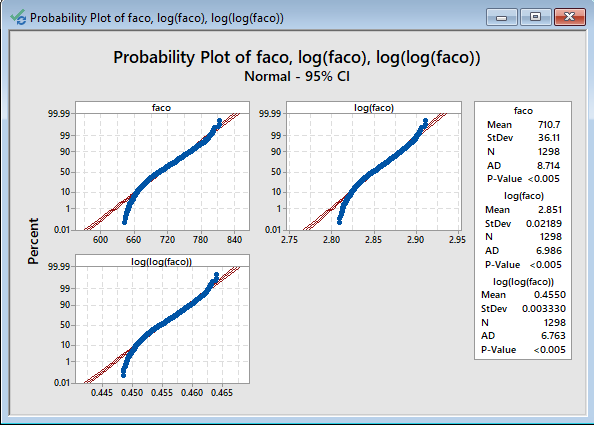


Figure 3: probability plot

Therefore multiple combination of subset has been tested and here are some among them that satisfy the normal distribution requirement and still have a reasonable amount of data:

1. 60-month mortgage
2. Loan amount greater than 17000

I think analyzing the FACO scores of the loans which are greater than 17000 is more meaningful, since in that case we can see what decides the FACO score when the person is borrowing a larger amount of money. Therefore I chose the sub dataset where the loan amount is greater than 17000 for my analysis.

Shown as Figure 4, we can see that the P-value of the subset selected is 0.171, greater than 0.05, therefore we have evidence that the data from the subset is normally distributed.

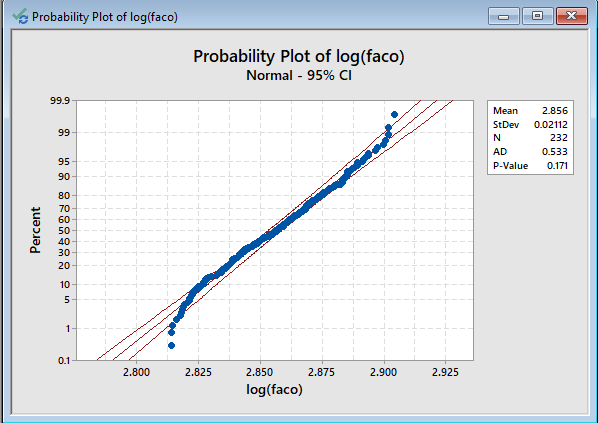


Figure 4: probability plot of the subset selected

### Correlations

In order to select the best independent variables or predictors that could best characterize the behavior of the dependent variable, we need to run the correlation and see which variables have stronger relations with the dependent variable selected, log(FACO). As Figure 4 shows, the variables that has the strongest relation with log(FACO) are int\_rate, revol\_util, delinq\_2yrs, revol\_bol and dti, there are more variables like home\_ownership\_rent, dti and total\_acc which can be included in the model as well depending on how strong the relations with the dependent variable should be.

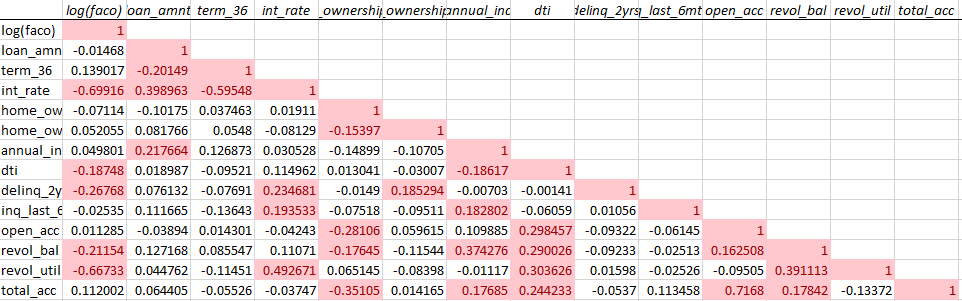


Figure 4: Data Correlation Analysis

### Parsimony Analysis

In order to decide which variables should be included in the model, we have to run the linear regression analysis and compare the adjusted **R2** to make the decision. In the first model, I included all the 13 variables in the model, and the adjusted **R2** is 76.47%, shown as Figure 5.

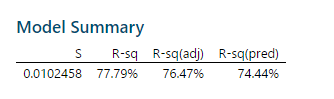


Figure 5: **R2 Adj. of model including all variables**

Then I tried to run the linear regression analysis with variables int\_rate, dti, delinq\_2yrs, revol\_bal, revol\_util, total\_acc and home\_ownership\_rent, and the adjusted **R2** is 64.03%, as shown in Figure 6.

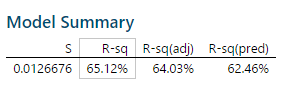
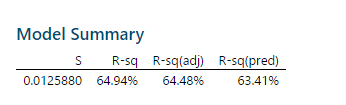


Figure 6: **R2 Adj. of model including 7 variables**

Then I ran the linear regression analysis with variables int\_rate, delinq\_2yrs and revol\_util, and the adjusted **R2** is 64.48% as shown in Figure 6 after getting rid of 4 variables. Therefore, the variables chosen are variables int\_rate, delinq\_2yrs and revol\_util.



**Figure 7: R2 Adj. of model including 3 variables**

### Variables Selected:

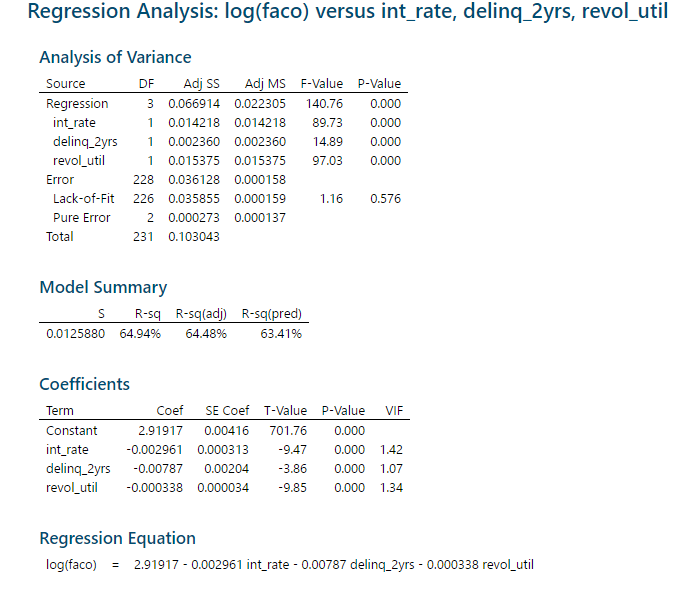
int\_rate, delinq\_2yrs and revol\_util

## Initial Model: Diagnosis

### Analysis or Residuals

The regression equation is log(faco)=2.91917-0.002961\*int\_rate-0.00787\*delinq\_2yrs-0.0000338\*revol\_util

And the initial analysis of the intimal model is shown below:



*Figure 8: Initial Model-analysis*

|  |  |
| --- | --- |
| What we can know from the initial model analysis:   * The variance inflation factors (VIF) for every independent variable is a little above 1 so it’s safe to say that there’s little multicollinearity and the coefficients are relatively stable. * The Adjusted R-Squred=64.48% is relatively high, but not significantly high, therefore we’d like a higher adjusted R-Square value * The critical F value is significantly high * The statistical significance, or P-value is significantly low * The individual P-values for the independent variables are also significantly low, so we wouldn’t want to discard any of these variables from the regression model   Figure 9 shows the Residual Plots for log(faco),  Figure 9 : Initial Model-Residuals  Overall, the residual plot seems to support the assumption of normality for residuals, even though there are deviation and outlier. The deviation and outlier don’t invalidate the normality of the residuals, and there are only several outliers from the normal probability plot, which indicates goodness of fit for normality test.  Besides what has been mentioned above, there’s some heteroscedasticity in the plot of the residual versus fitted values for log(FACO). Therefore some relationship can be observed between the residuals and the dependent variable and the explanatory variables, and we might need to reduce the heteroscedasticity with some data transformations. |  |

## Model Improvement: Outlier Analysis

From the Probability Plot, we can see that there are several outliers, but most of the data are within the normal range, which means that it might need outlier review or removal.

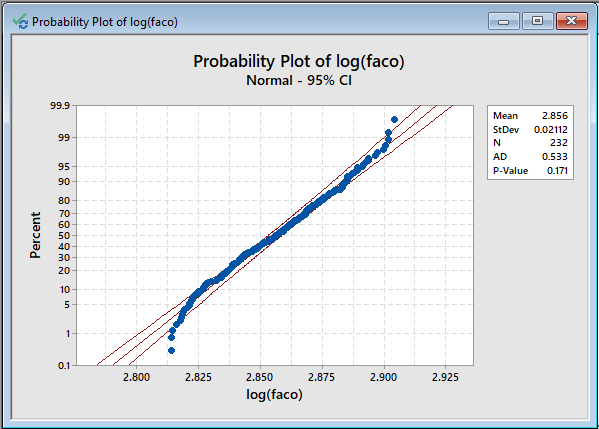


Figure 10: Probability Plot of Log(FACO)

## Final Model: Diagnosis

Initial analysis of the regression model log(faco)=2.91917-0.002961\*int\_rate-0.00787\*delinq\_2yrs-0.0000338\*revol\_util indicates that the model is within reason.

The analysis of the residuals versus fitted values indicates that the majority of the values fall within expected thresholds. The minority that doesn’t fall within expected thresholds doesn’t invalidate the model.

To determine whether we can improve the current model, we take a look at the interactions between the independent variables and the dependent variable. And we see that the behaviors of the independent variables delinq\_2yrs and int\_rate are close, shown in Figure 11, therefore, I decided to add another independent variable delinq\_2yrs\*int\_rate and see if the newly added variable will improve the model significantly.

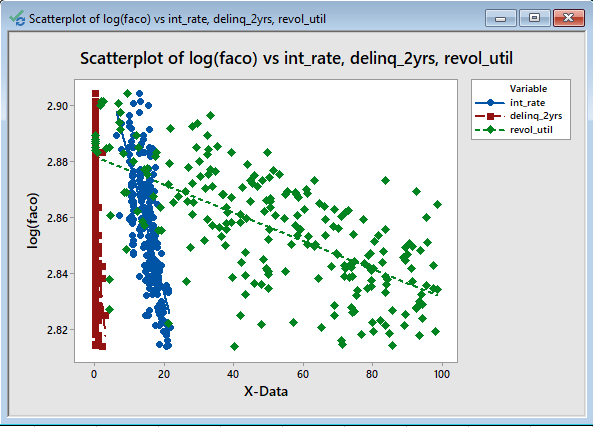


Figure 11: Scatter Plot of Log(FACO) vs int\_rate,delinq\_1yrs,revol\_util

After running the regression model with the newly added variable delinq\_2yrs\*int\_rate, and as shown in Figure 12, the adjusted R-square value is not significantly improved, and it doesn’t justify the adding of the new variable. Therefore, I think the initial model works the best so far

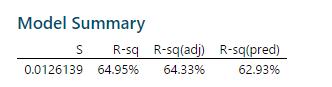


Figure 12: Adjusted R-sq of the new model

## Forecasting Dependent Variable Values

I tried to use the model to try to forecast the FACO value to see how well the model can do. I chose a set of independent variables, int\_rate=10.74, delinq\_2yrs=0, revol\_util=66.2, and put them in the regression model log(faco)=2.91917-0.002961\*int\_rate-0.00787\*delinq\_2yrs-0.0000338\*revol\_util, and the result we get is log(faco)=2.91917-0.002961\*10.74-0.00787\*0-0.0000338\*66.2=2.885131. Compare the result we get from the regression model to the data we have on book, 2.85319848132242, we can see that they are not the same, but they are pretty close.

Using the same way, we apply all the values of the independent variables to the model to see how close the predictions are to the actual data. Shown as Figure 13. We can see that the forecast is doing a relatively good job.

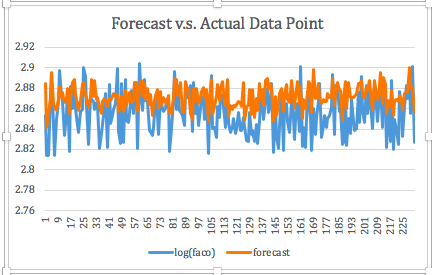


Figure 13: Forecast V.S. Actual Data

Provided values for each of the independent variables:

Int\_rate=10.74,

Delinq\_2yrs=0,

Revol\_util=66.2

XT0= (10.74, 0, 66.2)

bhatT=(2.91917,0.002961,0.00787, 0.0000338)

yhat= 2.885131

95% Prediction Interval:

XT0bhat=2.885131

Df error=228

T228, .975=1.970423

Standard deviation= 0.02112

Log(FACO)LB=2.853161

Log(FACO)UB=2.858626

FACO-LB=713.1177

FACO-UB=722.1471

Log(FACO) is not within the 95 % prediction interval

FACO is not within the 95% prediction interval

# Conclusions

The prediction interval for FACO score is very narrow, it’s actually indicating that int\_rate, delinq\_2yrs and revol\_util all have small deviations. Using the model to predict, the FACO score generated is 2.885131, while the actual FACO data for that dataset is 2.853198, which is smaller than the predicted score. Which indicates that the model is usable, but when LSI uses this regression model, they need to take the fact that the FACO score predicted is likely to be lower or higher than the actual FACO score.