**Homework 5. Lending Dataset Modeling**

Author: Rohit V. Akole

UConn ID: 3133252

University of Connecticut

OPIM5604 Predictive Modeling

Professor: Pankaj Prakash

Abstract

This homework is about the Lending Dataset based on year 2014 data. I used the previously cleaned dataset to calculate the logistic regression of Loan Status. I also calculated the Confusion Matrix, Odds Ratio, Precision, recall, and accuracy.

Table of Contents

[Question 1 4](#_Toc147019538)

[A. Nominal Logistic Fit for Loan\_Status 4](#_Toc147019539)

[B. Odds Ratios for the Final Model (without Indicator columns) 9](#_Toc147019540)

[C. Gains and Lift Charts for Final Model 10](#_Toc147019541)

[D. Confusion Matrix at 50% Threshold vs 32.93% (10% of highest prob[charged-off] of validation 11](#_Toc147019542)

[For 50% Cut Off Value 11](#_Toc147019543)

[For 32.93% Cut Off Value 11](#_Toc147019544)

# Question 1

Q1. From the lending\_2014\_ver1\_raw.jmp dataset, build two Multiple Linear Regression Models to predict the rate of return for the loans in the dataset. The first outcome variable to be modeled will be based on method M1 (the pessimistic method) and the second outcome variable to be modeled will be based method M2 (the optimistic method).

## Nominal Logistic Fit for Loan\_Status

|  |  |
| --- | --- |
| Figure 1. Logistic Regression Iteration 1 | Figure 2. Logistic Regression Iteration 2 |

In the above model iterations, Iteration 1 includes all the variables available at the time before the loan was issued to mitigate the target leakage. In iteration 1, we can see that the total account had the highest p-value in Effect Wald Test. Even though the home\_ownership\_2 [OWN] had the highest p-value of 0.8459 in parameter estimates, the overall variable in Effect Test has the p-value of less than 0.05. Overall, the home\_ownership\_2 is a significant variable. Hence, I decided to keep it and remove total\_acc instead. After removing total\_acc, All my variables have p-value less than 0.05.

In iteration 2, revol\_util has the highest p-value in Effect test. So I removed revol\_util from the model and emp\_length\_num Or Mean if Missing has the p-value higher than 0.05. As discussed, if the p-value of Is Missing is higher, then it is acceptable to keep the variable in the model. But here, the p-value is higher for the emp\_length\_num Or Mean if Missing at 0.2389. To deal with the null values of emp\_length\_num, I took the mean of the column and decided to group null values with the mean of the column which was 6. So, in 3rd iteration I replaced all the null values for emp\_length\_num and created a column emp\_length\_num2. It didn’t seem like a good idea to group them with a column with value 10 even though we had the largest frequency of data in that column, but that was because I grouped all the 10 and above years into 10 to reduce the bin size. So, going with the mean to replace the null values seemed like a good idea.

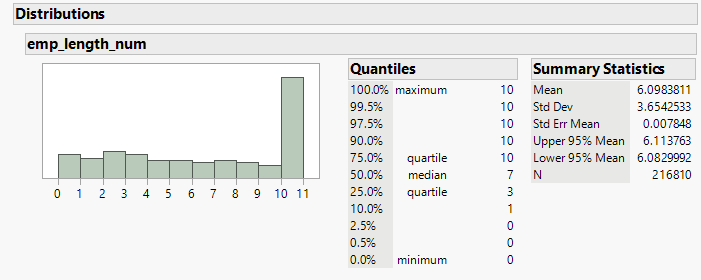


Figure 3. Distribution of emp\_length\_num column

|  |  |
| --- | --- |
| Figure 4. Logistic Regression Iteration 3 | Figure 5. Logistic Regression Iteration 4 |

In iteration 3, after changing the emp\_length\_num with emp\_length\_num2, the p-value of the variable changed to 0.1457, which was still higher than 0.05. So, I removed that variable. After removing the variable emp\_length\_num2, I got the final model, where all the p-values were less than 0.05 in Effect Wald Tests. With Indicator Function Parameterization, all else being equal, I could see that the home\_ownership\_2 [OWN] and addr\_state\_region [Northeast] have a p-value less than 0.05.

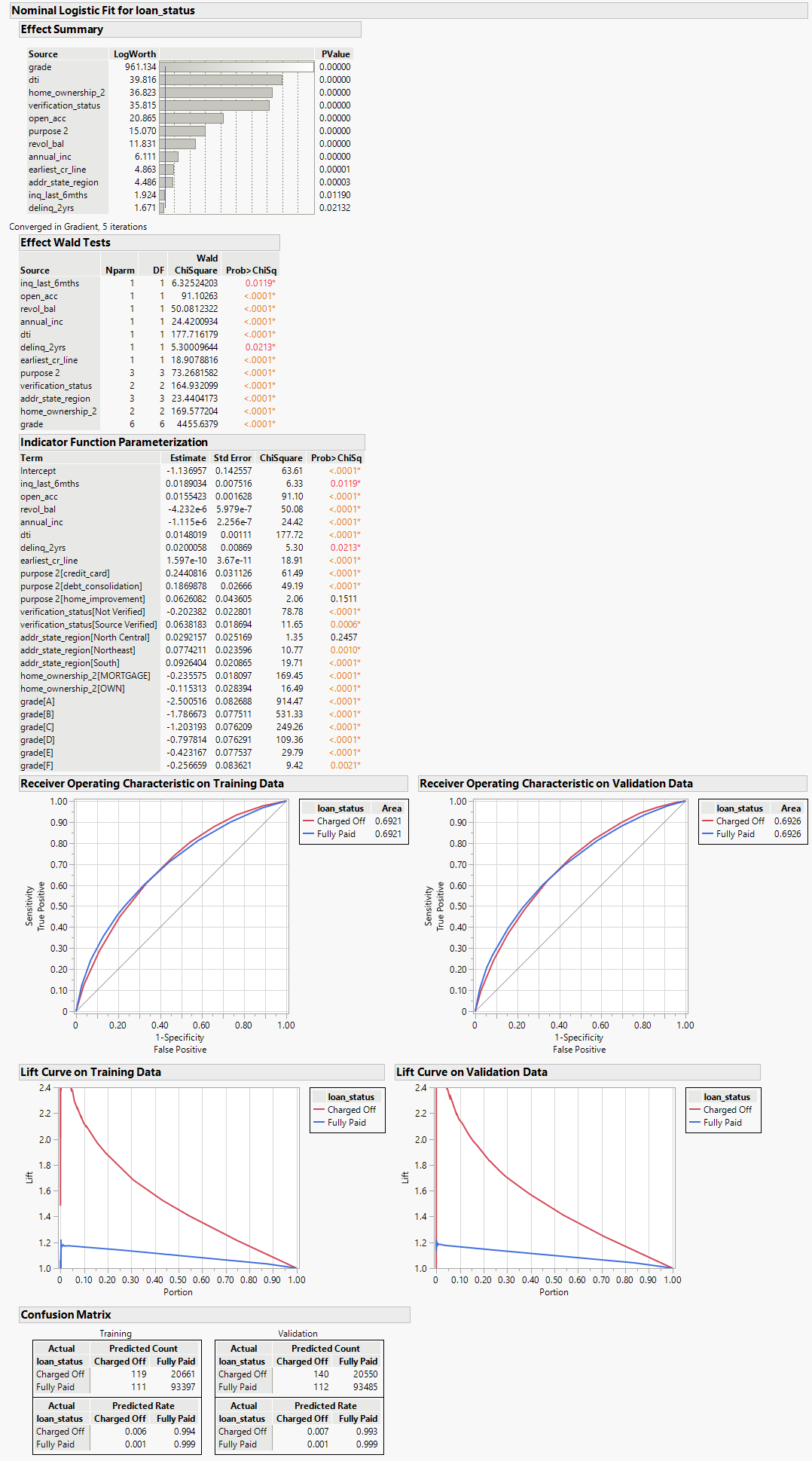


Figure 5. Final Model without Indicator Columns

In the confusion matrix, I have true positive Charged off 119, false positive 111 and True negative I have 93397 and false negative I have 20661 for Training data set. For the Validation dataset, I have 140, 112, 93485, and 20550 respectively. For the lift curve, large number of data for charged off is in first 20%.



Figure 6. Final Model with Indicator Columns

Both models without Indicator columns and with Indicator columns are closely like each other. The percentage distribution of the Confusion matrix is the same for both. Under ROC, the Area is 0.06921 for without Indicator, and 0.6920 for with Indicator columns model.

## Odds Ratios for the Final Model (without Indicator columns)

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated  Figure 7. Logistic Regression Odds Ratio Part 1 | Figure 8. Logistic Regression Odds Ratio Part 2 |

1. For the odds in Purpose 2, Credit card over Other have 27.64% higher chances of loan to charged off. That’s the highest ratio in Purpose 2 odds ratio. For the same variable, others have 21.66% lower odds of Charged off.
2. For the Verification Status, Source verified have 30.49% higher chance of charges off as compared to not verified. Not verified has 23.37% lower chances of the loan getting charged off.
3. For State Regions, there are a lot of Odds ratios that are overlapping 1 in Lower and Upper 95%.
4. People who are on Rent have 26.56% higher odds of loan getting charged off as compared to Mortgage. People on mortgage has 20.99% lower chances of loan getting in charged off vs rent.
5. In Grade, G over A grade has the worst odds ratio. G grade has 12 times more odds of loan getting charged off as compared to A grade. Similarly, A grade loans have 91.8% less odds of loan getting charged off as compared to G grade.

## Gains and Lift Charts for Final Model

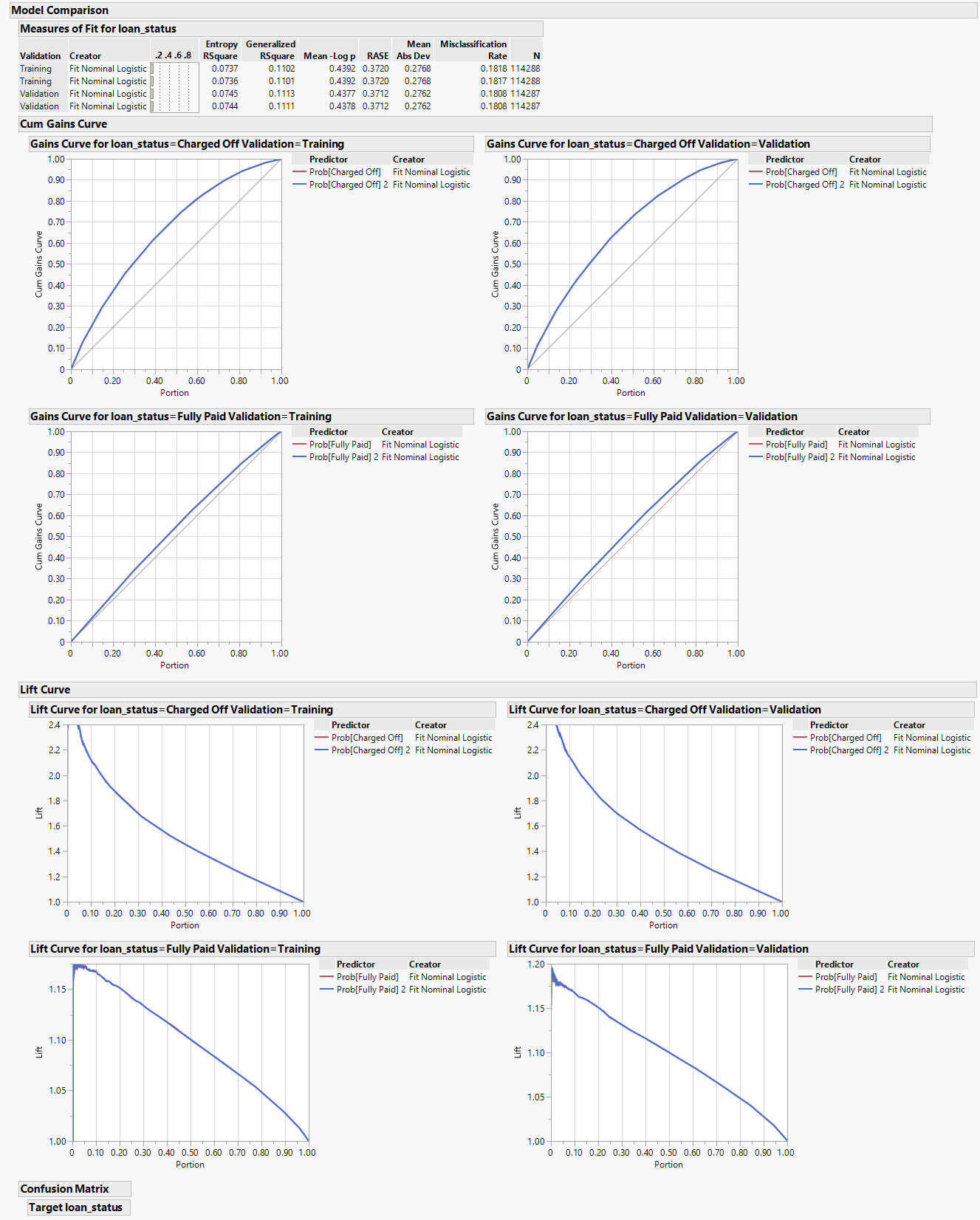


Figure 9. Cum Gains Curve and Lift Chart with both final model comparison

I compared both models with and without indicator columns using model comparison. Results are very close for both models. As we can analyze from the Gains Curve for charged off, the trend is similar for training and validation data. For 10% of population, my model is capturing 20% defaults. For 20% of the population, it’s a little less than 40% defaults. In 30% of the population, we have captured 50% of the charged-off accounts and so on. On the lift curve, we can see the slope is downwards for charged-off accounts. At the 10% of the population, the probability of charged off is 2.1, for 20% of the sample size, the probability of charged off is 1.9.

## Confusion Matrix at 50% Threshold vs 32.93% (10% of highest prob[charged-off] of validation

|  |  |
| --- | --- |
| Figure 10. Confusion Matrix 50% Probability | Figure 11. Confusion Matrix 32.93% Probability |

I calculated the confusion matrix for the 50% default probability threshold as well and I calculated the same by sorting the validation data and taking the value from Prob [Charged off] of 11428 row number as it was the row of 10% sample size. I got a probability of 32.93% from that row. So, I calculated the confusion matrix for the ideal scenario as well.

### For 50% Cut Off Value

Accuracy for 50% = (TP+TN) / total = (259+186882) / 228575 = 0.8187 = 81.87%

Recall for 50% = TP / actual events = 259 / 41470 = 0.0062 = 0.62%

Precision for 50% = TP / predicted yes = 259 / 482 = 0.5373 = 53.73%

### For 32.93% Cut Off Value

Accuracy for 32.93% = (TP+TN) / total = (8786+173238) / 228575 = 0.7963 = 79.63%

Recall for 32.93% = TP / actual events = 8786 / 41470 = 0.2119 = 21.19%

Precision for 0.3293 = TP / Predicted yes = 8786 / 22653 = 0.3878 = 38.78%

For the lending case, as an investor, Recall is important as it tells me the proportion of actual charged-off loans my model can identify. It is significantly better as I reduced the cut-off value based on my validation data. Also, precision is important as I want my model to predict the charged-off loans correctly and put them in the correct bins. In this case, Accuracy will be less significant as most of the loans are getting fully paid as compared to charged-off loans.