**Homework 4. 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. To use this dataset into the final model, the dataset has to be cleaned. Impute the appropriate columns and remove the unnecessary columns to improve the performance of the model and to get accurate results out of this model.

Table of Contents

[Question 1 4](#_Toc146468712)

[A. Step-by-step data cleaning 4](#_Toc146468713)

[B. Validation split. 5](#_Toc146468714)

[C. Calculations for Rate of Return (ROI) 5](#_Toc146468715)

[D. Pessimistic Model 5](#_Toc146468716)

[E. Optimistic Model 17](#_Toc146468717)

[F. Pessimistic Vs. Optimistic Final Model 29](#_Toc146468718)

# 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).

## Step-by-step data cleaning

1. Created the subset of the dataset keeping only Charged off and Paid off loans in loan\_status using Rows > Data Filter, and then, Tables > Subset. Saved it as a new dataset.
2. Changed the datatype of the variables next\_pymnt\_d, last\_credit\_pull\_d, last\_pymnt\_d from nominal categorical to continuous numeric variables with m-d-y format.
3. Revol\_util, term also had to be changed from nominal categorical to continuous numeric variable since they were in percentages.
4. Recoded and created new columns for emp\_length and grade to numbers and changed the datatype from nominal categorical to continuous numeric variables, keeping old columns as is.
5. Only 1 value from Home\_ownership had value Any. I changed it to Mortgage as Mortgage had 51% number of the data.
6. Recoded purpose and grouped all the data into other except Credit card, debt\_consolidation, and home\_improvement as these 3 variables had significant numbers of rows and named it purpose2.
7. Using conditional statement (Different\_Funded\_Amount), I checked whether the loan\_amnt requested and funded\_amnt had any difference. But both are the same. So, we can use any one of those in our model but not both as they could lead to target leakage.
8. I recoded the pub\_rec, named pub\_rec 2 to a new column, and grouped all 3+ public derogatory remarks to 3 to reduce the bin sizes. There were many small numbers of remarks after 3.
9. I assumed for the missing values in Last\_Credit\_Pull\_dt and Last Payment Date, I replaced the nulls with the issue date.
10. I created the Region column to split all the states into 4 regions, Northeast, North Central, South, and West.

## Validation split.

After the initial cleaning steps, filtering the data with only completed loans, replacing null values for Last Credit Pull Date and Last Payment Date with Issue Date. The validation split was done 50-50 with 888 as seed.

## Calculations for Rate of Return (ROI)

AS discussed in class, I calculated ROI by Optimistic and Pessimistic ways.

1. For the Optimistic way, I used the conditional If else statement. If total payment – funded amount is less than or equal to 0, then ((total payment – funded amount) / funded amount) \* (12 / term) else ((total payment – funded payment) / funded amount) \* 2 / actual loan length).
2. For pessimistic, I simply used ( (total payment – funded amount) / funded amount ) \* ( 12 / term ).
3. As I stated in section A, I also calculated the difference between the loan amount requested and the funded amount using an if-else conditional statement.

## Pessimistic Model

In this case, I’m modeling for both pessimistic and optimistic ROI. Firstly, I will look into the model for the Pessimistic Rate of Return. For the first iteration of the model, I took all the variables except title, emp\_title, loan\_amnt, Optimistic\_rate\_of\_return, instead of grade, I’m using grade\_num, and instead of addr\_state, I’m using addr\_state\_region.

|  |  |
| --- | --- |
| Figure 1. Pessimistic Model Iteration 1 | Figure 2. Pessimistic Model Iteration 2 |

Firstly, I removed sub\_grade as I already had gread\_num in the model.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated  Figure 3. Pessimistic Model Iteration 3 | Figure 4. Pessimistic Model Iteration 4 |

After removing the sub\_grade, I started removing the variables with the p-value less than 0.05. I removed recoveries, then acc\_now\_delinq.

|  |  |
| --- | --- |
| Figure 5. Pessimistic Model Iteration 5 | Figure 6. Pessimistic Model Iteration 6 |

I removed purpose 2from the model as it was the least significant variable in iteration 5 and I removed open\_acc from iteration 6 for te same reason. As I was removing these variables, the R2 and R2 Adjusted didn’t change significantly.

|  |  |
| --- | --- |
| Figure 7. Pessimistic Model Iteration 7 | Figure 8. Pessimistic Model Iteration 8 |

Then I removed annual\_inc and pub\_rec 2 from iteration 7 and 8 respectively.

|  |  |
| --- | --- |
| Figure 9. Pessimistic Model Iteration 9 | Figure 10. Pessimistic Model Iteration 10 |

Then I removed Revol\_util & Missing coded from iteration 9 and verification status from iteration 10.

|  |  |
| --- | --- |
| Figure 11. Pessimistic Model Iteration 11 | Figure 12. Pessimistic Model Iteration 12 |

Then I removed home\_ownership 2 and revol\_bal from the next 2 iterations. The R2and R2 Adjusted didn’t change much in all these iterations.

|  |  |
| --- | --- |
| Figure 13. Pessimistic Model Iteration 13 | Figure 14. Pessimistic Model Iteration 14 |

From model 13 I removed delinq\_2yrs and from iteration 14 I removed addr\_state\_region which was a variable I created to put all the states in their region. But the p-value for some of the states was higher than 0.05.

|  |  |
| --- | --- |
| Figure 15. Pessimistic Model Iteration 15 | Figure 16. Pessimistic Model Iteration 16 |

In iteration 15, is the iteration after removing addr\_state\_region. Here, The p-value was below 0.05 for all and I checked the VIF and the VIF for earliest\_credit\_line was the highest in all (10361203). So, I decided to remove it from the model. In iteration 16, the vif was highest for Last\_payment\_date.

|  |  |
| --- | --- |
| Figure 17. Pessimistic Model Iteration 17 | Figure 18. Pessimistic Model Iteration 18 |

From iteration 16, I removed the Last payment date, and the major high VIFs reduced drastically. The R2 and R2 Adj. didn’t change much. From iteration 17, I could see that the VIF for funded\_amnt was highest. So in iteration 18, I removed the funded amount. As I removed the funded\_amnt, the VIFs for others changed significantly.

|  |  |
| --- | --- |
| Figure 19. Pessimistic Model Iteration 19 | Figure 20. Pessimistic Model Iteration 20 |

From iteration 18, after removing funded\_amnt, the p-value for few variables again increased to more than 0.05. So, I removed the one with highest p-value and lowest log-worth (inq\_last\_6mths). From iteration 19, I removed issue\_date as it has the lowest log-worth and again the p-value was more than 0.05. After removing issue\_date, the p-value for grade\_num changed to 0.1147. So, I had to remove it in next iteration.

|  |  |
| --- | --- |
| Figure 21. Pessimistic Model Iteration 21 | Figure 22. Pessimistic Model Iteration 22 |

In iteration 21, the p-value for mths\_since\_last\_record Or Mean if Missing was at 0.0504. So, I decided to keep it as in the Effect Tests it was way under 0.05. The VIF was higher for installment and total\_pymnt. So, from iteration 21, I removed total\_payment as it had the highest VIF. Finally, I have a final model at iteration 22 for pessimistic rate of return. All the -p-values are VIFs are under the limit we set for the model. R2 atthe first model was 0.860297 and Adj R2 was 860203 and the RMSE was 0.028692. For the final model, R2 is at 0.84672 and Adj R2 is at 0.846696 and the RMSE was 0.028692.

## Optimistic Model

|  |  |
| --- | --- |
| Figure 23. Optimistic Model Iteration 1 | Figure 24. Optimistic Model Iteration 2 |

The first model had a lot of variables where p-value was greater than 0.05. I started by removing pymnt\_plan as there was no data in the column. Then from 2nd iteration, I removed acc\_now\_delinq as it had the highest p-value in the model.

|  |  |
| --- | --- |
| Figure 25. Optimistic Model Iteration 3 | Figure 26. Optimistic Model Iteration 4 |

The in 3rd iteration, I removed verification status as the p-value was 0.75905 and the log-worth was the least at 0.120. Then we got our 4th iteration. The VIF for earlies\_cr\_line was highest, so I removed it to check if it allows me to keep more variables in at the final model.

|  |  |
| --- | --- |
| Figure 27. Optimistic Model Iteration 5 | Figure 28. Optimistic Model Iteration 6 |

In 5th iteration, I removed Last\_payment\_date as it had the highest VIF. Then in 6th iteration I removed inq\_last\_6mths as it had the lowest log-worth and highest p-value at 0.80558. Which is significantly higher than 0.05.

|  |  |
| --- | --- |
| Figure 29. Optimistic Model Iteration 7 | Figure 30. Optimistic Model Iteration 8 |

From iteration 7, I removed open\_acc as it had the highest p-value of 0.58054 and lowest log-worth of 0.236. Then in iteration 8, I removed revol\_bal for the same reason.

|  |  |
| --- | --- |
| Figure 31. Optimistic Model Iteration 9 | Figure 32. Optimistic Model Iteration 10 |

From the 9th iteration, the home\_ownership\_2 had the highest p-value so I decided to remove it from the model. 10th iteration had the purpos\_2 variable with the lowerst log\_worth and the p-value was higher than 0.05. So, I removed it from the model.

|  |  |
| --- | --- |
| Figure 33. Optimistic Model Iteration 11 | Figure 34. Optimistic Model Iteration 12 |

From the iteration 11, I removed pub\_rec 2 as the VIF was higher than 4. Then from the model iteration 12, I removed length\_of\_credit as it had the highest p-value.

|  |  |
| --- | --- |
| Figure 35. Optimistic Model Iteration 13 | Figure 36. Optimistic Model Iteration 14 |

From iteration 13, I chose to remove annual\_inc as it had the lowest log-worth and highest p-value. In iteration 14, I had the revol\_util & MissingCoded variable which had the p-value of 0.0460. So, I removed that variable.

|  |  |
| --- | --- |
| Figure 37. Optimistic Model Iteration 15 | Figure 38. Optimistic Model Iteration 16 |

From iteration 15, I chose to remove the addr\_state\_region variable as the p-value of Northeast region was higher than expected 0.05. From the iteration 16, I removed mths\_since\_last\_delinq, since the p-value was higher for this variable.

|  |  |
| --- | --- |
| Figure 39. Optimistic Model Iteration 17 | Figure 40. Optimistic Model Iteration 18 |

I could see that in the iteration 17, the model still had a variable with p-value more than 0.05, mths\_since\_last\_record. So, I decided to remove the variable. From iteration 18, I started looking at VIF again. I removed Funded\_amnt from iteration 18 as the VIF was highest for that variable and it was more than the suggested VIF of 4.

|  |  |
| --- | --- |
| Figure 41. Optimistic Model Iteration 19 | Figure 42. Optimistic Model Iteration 20 |

After removing that variable from iteration 18, the p-value for issue\_d increased to 0.14168. So, I removed that and grade\_num variables from iteration 19. From iteration 20, I chose to remove the recovery as it had the highest VIF.

|  |  |
| --- | --- |
| Figure 43. Optimistic Model Iteration 21 | Figure 44. Optimistic Model Iteration 22 |

From iteration 21 we chose to remove total\_payment as the VIF was 13.94855 for this variable which is greater than the standard VIF of 4. After removing the last variable in iteration 21, the iteration 22 looked like an ideal model. All the VIFs are less than 4, all the p-values are less than 0.05.

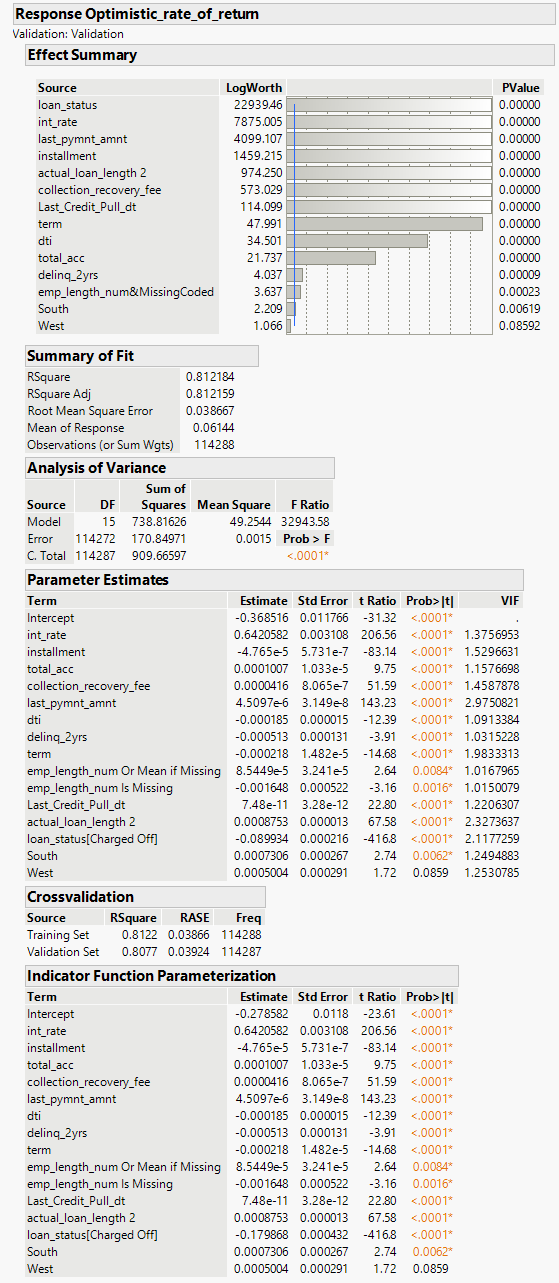


Figure 45. Optimistic Model (Final) Iteration 23

I wanted to check the relation of regions and optimistic rate of return. I added all 4 regions and finally concluded that South and West fit in with my model and certainly these regions have a significant impact on my model. Finally, the model I initially started with had the R2 of 0.840466 and in final model I have R2 of 0.812184. RMSE was 0.035641 and 0.038667 for old and final model respectively.

## Pessimistic Vs. Optimistic Final Model

|  |  |
| --- | --- |
| Figure 46. Pessimistic Final Model | Figure 47. Optimistic Final Model |

I added South region back to the Pessimistic Final Model and the model was fit in all standards (0.05 p-value and VIF 4). Other 3 regions were above 0.05 p-value for pessimistic. For the optimistic model, I got South and West both close to p-value or less than p-value. In the final models, R2 is 0.846731 and 0.812184 for pessimistic and optimistic respectively. Adjusted R2 is 0.846706 and 0.812159 for pessimistic and optimistic respectively. RMSE is 0.028101 and 0.038667 for pessimistic and optimistic respectively. The difference between the R2 and Adjusted R2 is very close to each other. For both models, frequencies are the same for testing and validation data as the data was split in 50-50. The difference between the R2 and RASE for both models for cross-validation in between Training and Validation sets are very close to each other for their model. i.e., For the Pessimistic model, the R2 is 0.9467 and 0.8465 for the training and validation data sets respectively. RASE is 0.02810 and 0.02813 for the training and validation data set respectively.

For the Optimistic model, the R2 is 0.8122 and 0.8077 for the training and validation data sets respectively. RASE is 0.03866 and 0.03924 for the training and validation data sets respectively.