MSiA401

Team: Group 6

Team Members:

Shreyashi Ganguly

Daniel Halteh

Shuyan Li

Zach Zhu

OPTIMIZE FUTURE

PROMOTION

FOR A RETAIL COMPANY

2019 Fall MSiA 401 Project 1

Contents

C.1 Project 1

- **Business Situation**: A retail company sells upscale clothing on its website and via catalogs, which help drive customers to the website. All customers were sent a catalog mailing in early fall 2012 and purchases made by them during fall 2012 were recorded. There is one row for each customer. The `targdol` is the response variable, which is the purchase amount during fall 2012; `targdol` $= 0$ indicates that the customer did not make a purchase. Other variables are potential predictor variables which give information about the customer as of the time of the mailing. We want to build a predictive model for responders to a similar promotion in future and how much they will buy. The purpose of the model is primarily prediction but it is also of interest to learn which are the most important predictors. The model is also intended to choose a subset of customers to be targeted in future promotions.

- **Data**: The data are in the file `Project 1 data.csv`. There are a total of 101,532 customers, of whom 9571 (9.43%) are respondents. The data are randomly divided into a training set with 50418 observations and the remaining 51,114 into a test set. (`train` is the indicator variable for training or test set: `train` =1 for the training set, `train` =0 for the test set). The definitions of the variables are as follows.

Variable	Description
targdol	dollar purchase resulting from catalog mailing
datead6	date added to file
datelp6	date of last purchase
lpuryear	latest purchase year
slstyr	sales ($) this year
slslyr	sales ($) last year
sls2ago	sales ($) 2 years ago
sls3ago	sales ($) 3 years ago
slshist	LTD dollars
ordtyr	number of orders this year
ordlyr	number of orders last year
ord2ago	number of orders 2 years ago
ord3ago	number of orders 3 years ago
ordhist	LTD orders
falord	LTD fall orders
sprord	LTD spring orders
train	training/test set indicator (1 = training, 0 = test)

LTD means "life-to-date," i.e., the time since the customer purchased for the first time.

- **Hints**:

  1. Some errors in the data are as follows. If you run a histogram or frequency distribution of the `datelp6` variable among only those with `targdol` > 0 you will see that, for the most part, `datelp6` equals one of two distinct dates in the calendar year. It is as if the data are binned into six-month bins. There are other

Executive Summary

This project aims to identify a potential pool of customers to be targeted for a catalog marketing campaign, by predicting their expected purchase amount. Since the typical response rate on such campaigns are found to be low, the problem statement is tackled via a two-model approach – first to classify customers as responders and non-responders, second to estimate dollar amount of purchases made by responders.

From our analysis, we observed that recency of purchase and level of engagement (increase or decrease in recent orders as compared to prior orders) came out to be the most significant predictors for identifying potential responders to the campaign. Besides this, a seasonal trend is response was also observed, meaning a customer who typically shops in Fall is more likely to respond to a Fall campaign than a customer who shops more in Spring. As for predicting the expected purchase amount, average amount spent in the past was a major predictor.

The top 1000 customers identified by the combination of our models resulted in a final pay-off of $24,742, which, though not perfect, amounted to 60% of the actual purchases made by these customers. Also, the total sales from the predicted pool of top 1000 customers amounted to 34.08% of the total sales from the actual top 1000 buyers.

Introduction

In the contemporary world, the mailing catalog is one of the most effective methods to promote sales, besides online marketing. A retail company, selling upscale clothing on its website, utilizes both methods to promote sales. As part of one such campaign, all their customers were sent a catalog mailing in early fall 2012 and purchases made by them during fall 2012 were recorded. The firm is now interested in building a model for predicting responders to a similar promotion in future together with their expected purchase amount. Since the response rate is around 9.43%, it is of interest to identify the most important predictors as well, so that a potential target pool can be identified.

Since only a few customers respond to such events, the expected purchase amounts cannot be modelled directly. Hence the project is split into two components,

1. The first component classifies the customers into potential responders or non-responders. A binary logistic regression was used to model this probability
2. The second component is to predict the expected purchase amount for a customer, given that he/she responds. A linear regression model was trained on *only* the responders to estimate the dollar amount of future purchases

The company shared response records for 101,532 customers of whom 9,571 are respondents. We divided the given records into training and test set in a roughly 50:50 ratio, with 50,418 records in the training sample, and 51,115 records in the test sample. The response variable

for the project is *'targdol'* which is the dollar amount of purchase made during fall 2012. Besides the features we engineered for the project, the dataset had 14 predictors related to recent purchase frequency and amounts.

Our analysis contains three parts: data preprocessing, model fitting and model validation. The dataset shared with us had some inconsistencies, which were corrected as part of the data preprocessing exercise. Also, exploratory data analysis was carried out to better understand the predictors and their relationship with each other. At this stage several features were created using the raw predictors. The clean data was then split into training and test samples as mentioned above and two separate models were trained. For the classification problem a binary logistic regression model was fit on the training data considering *'targdol > 0'* as success. The accuracy of the model prediction was assessed using Correct Classification Rate (CCR), F1-score and AUC metrics. For the multiple linear regression model, a logarithmic transformation of *targdol* was taken to make the distribution symmetric. Adjusted R^2 and AIC were used to identify the best model and Mean Square Prediction Error was calculated to assess accuracy of prediction. The predictions from both the models were multiplied to give the final estimated dollar purchase, in line with the following formula,

$$E(y) = E(y|y > 0) * P(y > 0), \qquad where\ y = targdol$$

The cumulative predictions were compared against the actual *targdol* values to assess cumulative accuracy. MSPE was used to compute the accuracy of our predictions from a

statistical standpoint. To validate the models from the purview of financial gains, we devised two metrices – ‘capture rate’ and ‘pay-off’ rate. ‘Capture rate’ is defined as the total purchase amount of the predicted top 1000 customers compared to the total purchase amount of the actual top 1000 customers. Our models captured 34.08% of the total purchase of the actual top 1000 customers. ‘Pay-off rate’ is defined as the proportion of actual purchases predicted by the model among the top 1000 predicted buyers. Our combined models capture 60.37% of the maximum pay-off

Data Cleaning and Exploratory Data Analysis

Before fitting the classification and multiple regression models, we first explored the dataset to get an overall understanding of variables as well as the entire quality of the dataset. According to our exploration, we reformatted the variables, resolved the inconsistencies and null values, and created new features so as to better prepare the dataset for further model fitting and analysis.

First, we checked all types of variables to see if there are exist inappropriate data types. We found for sales or number of orders, the original data used ‘num’ or ‘int’ data types, which is consistent and ideal. However, it used ‘chr’, which makes further reference difficult. Therefore, we converted ‘chr’ to formatted ‘Date’, so that we can directly refer to the year, month or day in feature engineering and model fitting. We also found that, for ‘lpuryear’, the

original column only contained the last digit of the year, which may confuse. Thus, we add ‘200’ if the original value is larger than or equal to 3 or ‘201’ if the original value is less than 3. We used ‘200’ and ‘201’ as the first three digits because, from our observation, the earliest date of last purchase was ‘2003’ and the latest of last purchase was ‘2012’.

After checking data types, we started handling missing values and inconsistencies. First of all, we found there were 834 records that "date added" after "date last purchase", which is counterintuition because the purchased could only be recorded after customers were added into a file. Thus, we imputed ‘datead6’ with ‘datelp6’ for those records.

Second, we discovered 728 records that had NA values in ‘lpuryear’. Accordingly, we imputed ‘lpuryear’ through ‘datelp6’ because the latest purchase year is exactly the year in ‘datelp6’. After imputation, we needed to keep consistent with ‘lpuryear’ and ‘datelp6’. Therefore, we compared the updated ‘lpuryear’ with ‘datelp6’ and updated both by the larger value.

We also looked at reasonableness between the sales for each year and the number of orders each year. We knew that if there were orders in the year, sales mush be larger than 0, and vice versa. Thus, we checked each pair in ‘slstyr’ and ‘ordtyr’, ‘slslyr’ and ‘ordlyr’, ‘sls2ago’ and ‘ord2ago’, ‘sls3ago’ and ‘ord3ago’, and ‘slshist’ and ‘ordhist’. For every pair, We did not find any cases that the number of orders is 0 but the sales is larger than 0, but we

discovered many cases in each pair that sales are 0 but the number of orders is not 0. To resolve those inconsistencies, we imputed the number of orders with sales, such as 0.

Another inconsistency existed between the sum of 'falord' and 'sprord' and 'ordhist'. Following the hints in the textbook, we looked that for the most part, 'datelp6' equals one of two distinct dates in the calendar year and the data are binned into six-month bins. To solve it, we compared the sum of 'falord' and 'sprord' and 'ordhist'. If we found 'ordhist' is less than the sum of 'falord' and 'sprord', we updated 'ordhist'; if 'ordhist' is larger than the sum of 'falord' and 'sprord' and the month in 'datelp6' is less than 7, we chose to update 'sprord' using 'ordhist' – 'falord'; last, if 'ordhist' is larger than the sum of 'falord' and 'sprord' and the month in 'datelp6' is greater than or equal to 7, we chose to update 'falord' using 'ordhist' – 'sprord'. After running those commands sequentially, we checked and kept the consistency between the sum of 'falord' and 'sprord' and 'ordhist'.

After removing missing values and resolve errors, we created eight new features that we thought should be useful in modeling, such as (please refer to Appendix 1 for the details). First and foremost, we added an id column to identify each customer uniquely.

Outlier:

Since the data is imbalanced with only 9.43% being responders, we can't use normal outlier detection methods such as box plot to eliminate all outliers which might remove a significant portion of the responders. Thus, we looked at the plots of the 21 features we

planned to fit in our model (see in Appendix 1) and set two criteria based on two of the interaction variables: year_btwn and falord. We manually removed observations where:

1. years between last purchase and 2012 exceeded 30 years
2. LTD fall orders exceeded 80

Overall, we only removed 21 observations which account for an extremely small portion of the total data. The scatterplots for the variables both before and after outlier removal are illustrated in Appendix 2. In addition, we used Cook's Distance to identify a few more outliers in our multiple linear regression model.

Model Fitting

Classification Model

Firstly, we transformed the response variable *targdol* into a binary variable and created a new column called *responder*. This binary variable has a value of 1 if *targdol* > 0 and a value of 0 if *targdol* = 0. A total of 21 variables (14 provided in the data and 8 engineered) were considered as contenders for the model. To assess their bivariate relation to the predictors Information Value was computed. The results are documented in Appendix 3 Part I. As expected, recency and bulk of past purchases came out to be the most significant predictors. Also, the bivariate relationship was assessed between the log-odds and the predictor variables to verify their linear relationship. Most of the feautures showed very less

variability owing to which their trend could not be assessed adequately. The variables regarding recency of purchase depicted the clearest linear trends.

To obtain our baseline model, a stepwise logistic regression was run with the criteria of minimizing AIC using *glm()*. This baseline model as presented in Appendix 3 Part II had 15 predictors. Some of the predictors like *'ord3ago'* had p-values > 0.05, others showed counter-intuitive signs (*'ordhist'*, *'slshist'* etc.). Also, the VIFs were very high (>10). Due to these concerns, several iterations were carried out to find the most optimum model, which satisfied the following criteria,

1. All predictors significant – p-values < 0.05
2. All predictors showing intuitive coefficient sign
3. VIFs < 10 for all features

The final model, as presented in Appendix 3 Part III, has 8 features from the initial set of 20 features, each with p-values < 10^{-3}, and VIFs < 5. The residual deviance is 26939 and AIC of 26957. This model was then assessed for its performance on both the train and test samples. In-spite of the imbalance in the classes, the model achieves a F1-score of 74.2 and AUC of 79.2 on the test sample. A threshold probability of 0.1 was used to compute the correct classification rate of 69.9%. The results are depicted in Appendix 3 Part IV.

Multiple Regression Model

Before fitting our Multiple Regression Model, we first needed to subset our training dataset such that it only included the observations who were classified as responders. This follows under the conditional probability rules, and more importantly factors out for the heavily uneven classes.

The first step when building our Multiple Linear Regression Model involves fitting our predictor "targdol" on all the predictors at hand. Through this approach, we call this full model our base model upon which we will soon initiate our variable selection process.

After fitting the initial model, our focus turned to the dependent variable "targdol", which we then intuitively chose to run our log transformation. This choice seemed necessary as our Normal Q-Q plot was not distributed normally (PLOT 1), and further, our residual plot had an increasing conic shape, indicating to us that the variance was not constant (PLOT 2).

After log-transforming our predictor variable, we noticed a slight jump in our adjusted R-squared value. Further, when checking our assumptions by plotting our graphs, we notice significant improvements in both our Q-Q plot and the residual plots. First, when looking at the updated Q-Q plot (PLOT 3), we can see that the plot approximates the Normal plot much more closely than the pre-log-transformation model. The residuals plot (PLOT 4), in a similar fashion, improves drastically—we can see that from the plot that there is no longer a conic shape, indicating to us that our assumption of constant variance is followed properly.

Our next step is to perform forward stepwise selection to determine the best subset of our models. To accomplish this, we first fit a Multiple Regression Model on the null model, which is the model with no predictors included. We then iteratively perform forward stepwise subset selection using the log-transformed model from earlier as our scope parameter. From this, we can see that four variables (slstyr_3ago, ord3ago, lpurseason, and shopper) have all been dropped from the model, thus giving us a minimum AIC criterion value of -2933.4.

We next look at our VIF's to keep track of any instances of multicollinearity. We see that one of the predictors, ordhist, has a high VIF, so we remove it. Please not that while a majority of the variation in ordhist alone seemed to be explained by the remaining independent variables, we will use ordhist later in the model selection process to add to our features.

After the log-transformation, forward stepwise subset selection, and VIF selection processes, we arrive at an adjusted R^2 of approximately 0.1153.

The next step in our model selection process includes adding interaction variables to the list of predictors. While this could have been done arbitrarily, i.e. creating an interaction term for every possible combination of predictors, we chose to rely on business-related intuition and domain knowledge to create interaction terms that best followed logically. We also chose to isolate this step for interaction variables associated with the sales histories of the model, while continuing with other predictors in later steps. Once again, we use forward

stepwise subset selection to help reduce our model and choose the subset of predictors that results in the lowest AIC criterion value. After selecting the best subset of predictors, we then once again run our Multiple Linear Regression Model on said subset, which results in an increased adjusted R-squared value of 0.1256.

Now we continue with our variable creation portion by adding interaction terms that extend beyond sales history. Specifically, we looking at the order history and sales history variables and decided to mutate them to create a new variable that is the quotient of the two. Our intuition here was that while we realized the variables themselves might be highly significant, their ratio may help provide more information in the context of interactions as well. We called this new imputed value as "average_amt", which represents the average amount for a given sale.

Another important step at this model development stage was to log transform the predictors that involved both sales amounts and order counts, two variables that are historically log-transformed in many business-driven prediction settings. We see that these transformations further improve the residuals plot (PLOT 5) by flattening out the variance curve and spreading out our residuals. Interestingly, this now brings our adjusted R-squared value up to 0.1502.

We lastly run one more iteration of forward stepwise subset selection to help check the significance added by each of our recently created predictors. Our final model, corresponding

the subset with the lowest AIC of 3125.6, gives us an adjusted R-squared value of 0.1505.

When looking at the high VIF values associated with our interaction variables and their co-dependents, it is important to note that these signs of multicollinearity arise primarily due to the repetition of the original predictors themselves. Given that our interaction terms here are highly significant, statistician Paul Allison states that in the presence of terms that are products of one another (i.e. interaction terms), high VIF values are not cause for alarm (Allison, Paul). As such, we chose to leave in said terms and proceed with next steps, regardless of the multicollinearity presence.

Model Validation

To assess the models we fit, first the individual model predictions were validated in isolation against the actual values in the test sample. For the logistic regression model, metrics like CCR, F-1 score and AUC were computed. For the linear regression model, MSPE was used to compute accuracy.

The prediction from both the models were combined as per the formula given below to get the predicted *'targdol'*,

$$E(y) = E(y|y > 0) * P(y > 0),$$

$$where, P(y > 0) = prediction\ from\ logistic\ model,$$

$$E(y|y > 0) = prediction\ from\ the\ linear\ model$$

The expected dollar amount obtained above was then compared against the actual *'targdol'* values to assess the cumulative accuracy of the model. The accuracy was measured using Mean Square Prediction Error. To assess the business impact of the project, the customers were ranked in decreasing order of their predicted purchase values. The top 1000 customers were identified as the target pool. The sum of the expected dollar amount for this pool was compared against the sum of their actual purchases to provide a 'pay-off' rate for the campaign. Also, 'capture rate' was computed to assess the overlap between the top 1000 pool identified by our model and the actual high purchase customers. For both the sets, sum of their actual purchase amount was compared.

The results for the final Logistic Regression model are as follows,

Sample	**Correct Classification Rate**	**F1 Score**	**Area Under ROC**
Train	69.94%	73.37%	78.67%
Test	69.74%	74.25%	79.19%

The results for the final Linear Regression model are as follows,

Sample	**MSPE**
Train	4803.93
Test	4712.59

The final results from combining the two models above are as follows,

Sample	MSPE	Pay-off Rate	Capture Rate
Test	405.55	60.37%	34.08%

Conclusion

The final logistic regression model includes 7 variables which were *year_btwn*, *ordhist*, *sprord*, *shopper*, *yob*, *slstyr_lyr, slstyr*. The following four predictors are most significant with relatively lowest p-values below 2e-16: *year_btwn, sprord, yob, shopper*. Hence, we suggest that these predictors can be utilized to identify potential pool of target customers.

The final multiple linear regression model includes 13 predictors which were *log(slshist + e) , ordhist, log(slstyr + e), sprord, falord, year_btwn, log(sls2ago + e), log(slslyr + e), log(ord2ago + e), log(ordlyr + e), log(average_amt + e), log(slstyr + e):log(sls2ago + e), year_btwn:log(average_amt + e)*. The following three predictors are most significant with relatively lowest p-values below 2e-16: *falord, sprord log(sls2ago + e)* and *log(average_amt + e)*. Thus, we suggest that these four predictors have the strongest power for predicting the amount of purchase for responding customers.

Overall, the target pool of top 1000 customers identified by our models, churned total sales amounting to $24,742, which when compared to the $40,984 actually spent by these customers amounts to a 60.37% of pay-off rate. Additionally, the total purchase amount of the

predicted top 1000 customers when compared with the total purchase amount of the actual top 1000 customers, gave a capture rate of 34.08%. The logistic and linear regression models presented here, provide multiple significant predictors that can be used to optimize future promotion. From our analysis, we could identify few key categories of information, which if captured, can lead to optimum target marketing – namely, the recency of purchases made by the customer, how engaged the customer has been with the firm, i.e., have his purchases declined over time or increased, and lastly is there a seasonal pattern to the customer's shopping behavior.

While the model we built were the best given the data provided, additional predictors, if available, could lead to definite improvement in performance. Some of the key features which could enhance the models are – demographic information about the customers (age, gender, education, occupation, location, etc.), an estimate of the customer's income, information about competition. In conclusion, our logistic and regression models identified key predictors that made intuitive sense and provided adequate performance in the validation data. We are confident that our models can be utilized to identify optimum pool of target customers in order to maximize return on investment for marketing campaigns.

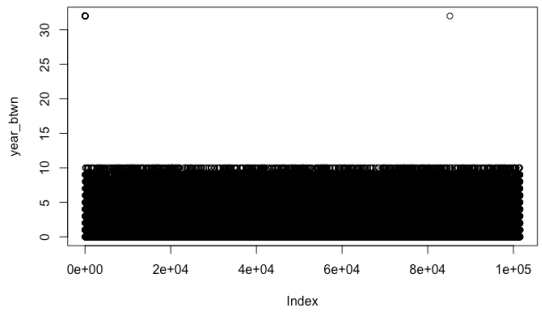
Appendix

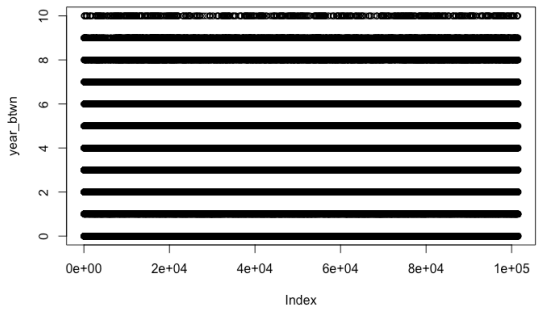
Appendix 1: Engineered Features

Feature Name	Information Category	Rationale
lpurseason	Seasonality of purchase	If last purchase was made in Spring or Fall
year_btwn	Recency of activity	Number of years between 2012 and last purchase year
yob	Engagement of customer	Number of years on book as on 2012
shopper	Seasonality of purchase	Does customer majorly shop in Fall or Spring or Tied?
slstyr_lyr	Engagement of customer	Ratio of sales this year to last year
slstyr_2ago	Engagement of customer	Ratio of sales this year to two years ago
slstyr_3ago	Engagement of customer	Ratio of sales this year to three years ago
average_amt	Engagement of customer	Average amount spent by customer in all prior orders

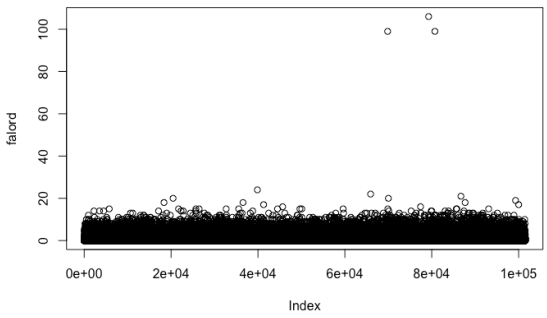
Appendix 2: Outliers

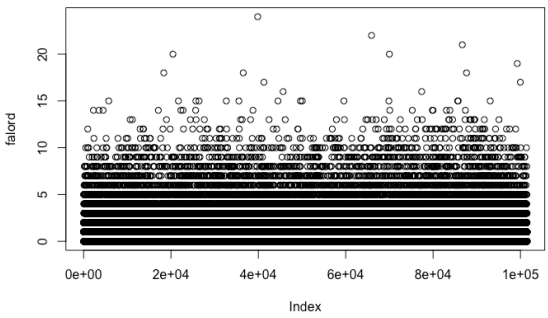
years between last purchase and 2012 exceeded 30 years





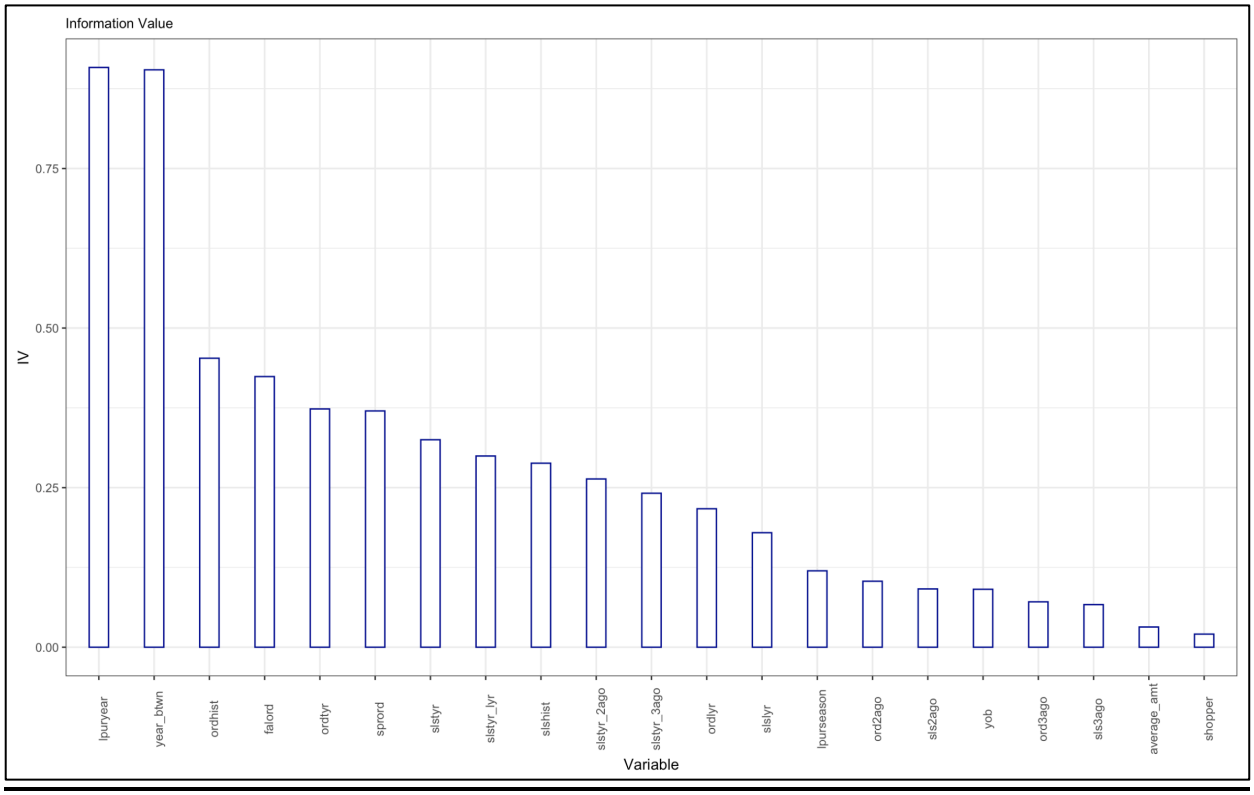
LTD fall orders exceeded 80





Appendix 3: Logistic Regression

Part I – Information Value



Part II – Baseline Model – from stepwise regression

```
Call:
glm(formula = responder ~ year_btwn + falord + as.factor(lpurseason) +
    ordhist + sprord + as.factor(shopper) + yob + ordtyr + ordlyr +
    ord3ago + ord2ago + slstyr_lyr + slstyr + slstyr_3ago + slshist,
    family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0106  -0.4451  -0.3217  -0.1930   4.0448

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.2600654  0.0630147 -35.866  < 2e-16 ***
year_btwn             -0.3577555  0.0143664 -24.902  < 2e-16 ***
falord                 1.6491801  0.0511216  32.260  < 2e-16 ***
as.factor(lpurseason)S 0.3419904  0.0472664   7.235 4.64e-13 ***
ordhist               -1.4024944  0.0494735 -28.348  < 2e-16 ***
sprord                 1.4910418  0.0507561  29.377  < 2e-16 ***
as.factor(shopper)S   -0.3797917  0.0681807  -5.570 2.54e-08 ***
as.factor(shopper)T   -0.0501633  0.0609050  -0.824 0.410149
yob                    0.0318959  0.0046239   6.898 5.27e-12 ***
ordtyr                 0.1383287  0.0412423   3.354 0.000796 ***
ordlyr                 0.0894119  0.0311633   2.869 0.004116 **
ord3ago                0.0608084  0.0330457   1.840 0.065749 .
ord2ago                0.0688398  0.0293393   2.346 0.018959 *
slstyr_lyr            -0.0023083  0.0007962  -2.899 0.003742 **
slstyr                 0.0042767  0.0010400   4.112 3.92e-05 ***
slstyr_3ago           -0.0024456  0.0009102  -2.687 0.007211 **
slshist               -0.0003399  0.0001567  -2.169 0.030074 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31905  on 50407  degrees of freedom
Residual deviance: 25330  on 50391  degrees of freedom
AIC: 25364
```

```
Number of Fisher Scoring iterations: 6

                          GVIF Df GVIF^(1/(2*Df))
year_btwn             2.097000  1        1.448102
falord               33.729481  1        5.807709
as.factor(lpurseason) 1.937258  1        1.391854
ordhist              55.215743  1        7.430730
sprord               14.040205  1        3.747026
as.factor(shopper)    3.208550  2        1.338373
yob                   1.895128  1        1.376636
ordtyr                2.673184  1        1.634988
ordlyr                1.653255  1        1.285790
ord3ago               1.543057  1        1.242198
ord2ago               1.371797  1        1.171237
slstyr_lyr            3.197818  1        1.788244
slstyr                8.013048  1        2.830733
slstyr_3ago           5.033237  1        2.243488
slshist               2.334018  1        1.527749
```

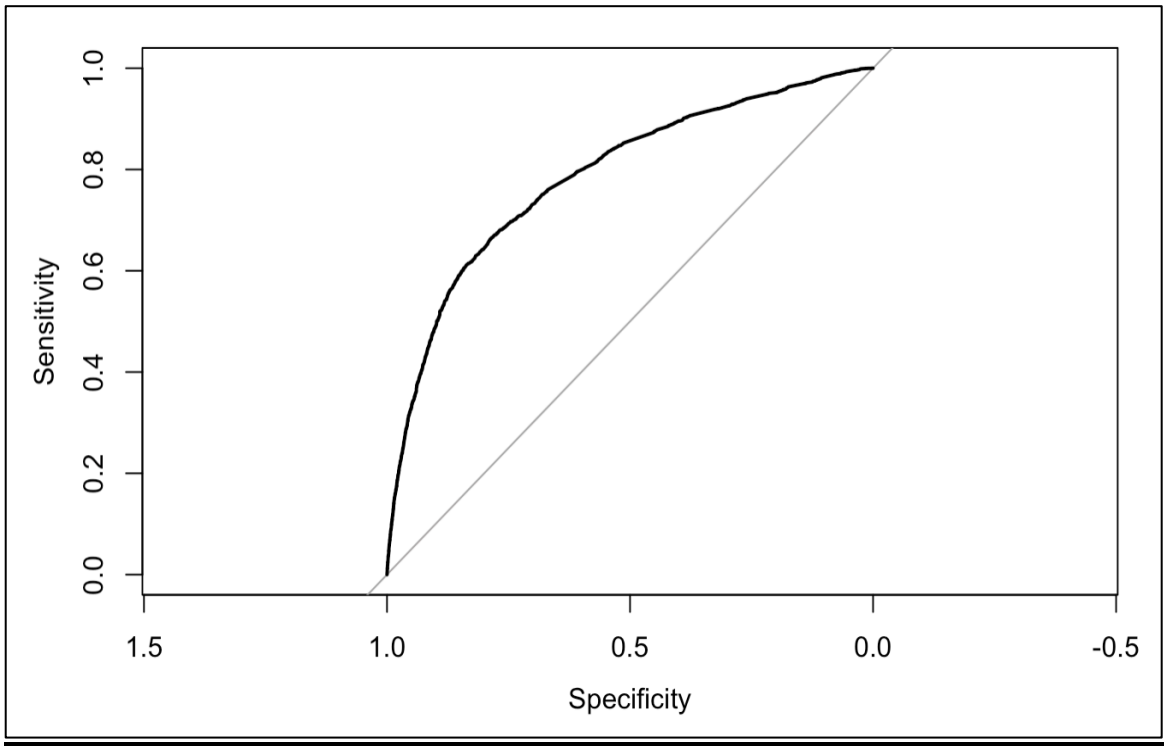
Part III – Final Logistic Model

```
Call:
glm(formula = responder ~ year_btwn + ordhist + sprord + as.factor(shopper) +
    yob + slstyr_lyr + slstyr, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1340  -0.4861  -0.3352  -0.1750   3.4917

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.4299214  0.0432812 -33.038  < 2e-16 ***
year_btwn           -0.5033374  0.0122538 -41.076  < 2e-16 ***
ordhist              0.0373650  0.0120024   3.113  0.00185 **
sprord               0.2873397  0.0252783  11.367  < 2e-16 ***
as.factor(shopper)S -0.8595761  0.0586857 -14.647  < 2e-16 ***
as.factor(shopper)T -0.2331406  0.0528733  -4.409 1.04e-05 ***
yob                  0.0401194  0.0040757   9.843  < 2e-16 ***
slstyr_lyr          -0.0041005  0.0007086  -5.787 7.17e-09 ***
slstyr               0.0028029  0.0005927   4.729 2.25e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31905  on 50407  degrees of freedom
Residual deviance: 26939  on 50399  degrees of freedom
AIC: 26957

Number of Fisher Scoring iterations: 6

                       GVIF Df GVIF^(1/(2*Df))
year_btwn          1.373179  1        1.171827
ordhist            3.955549  1        1.988856
sprord             4.127747  1        2.031686
as.factor(shopper) 2.425728  2        1.247988
yob                1.674480  1        1.294017
slstyr_lyr         2.554607  1        1.598314
slstyr             2.686575  1        1.639077
```

Part IV – Accuracy of Logistic Model

Training Data

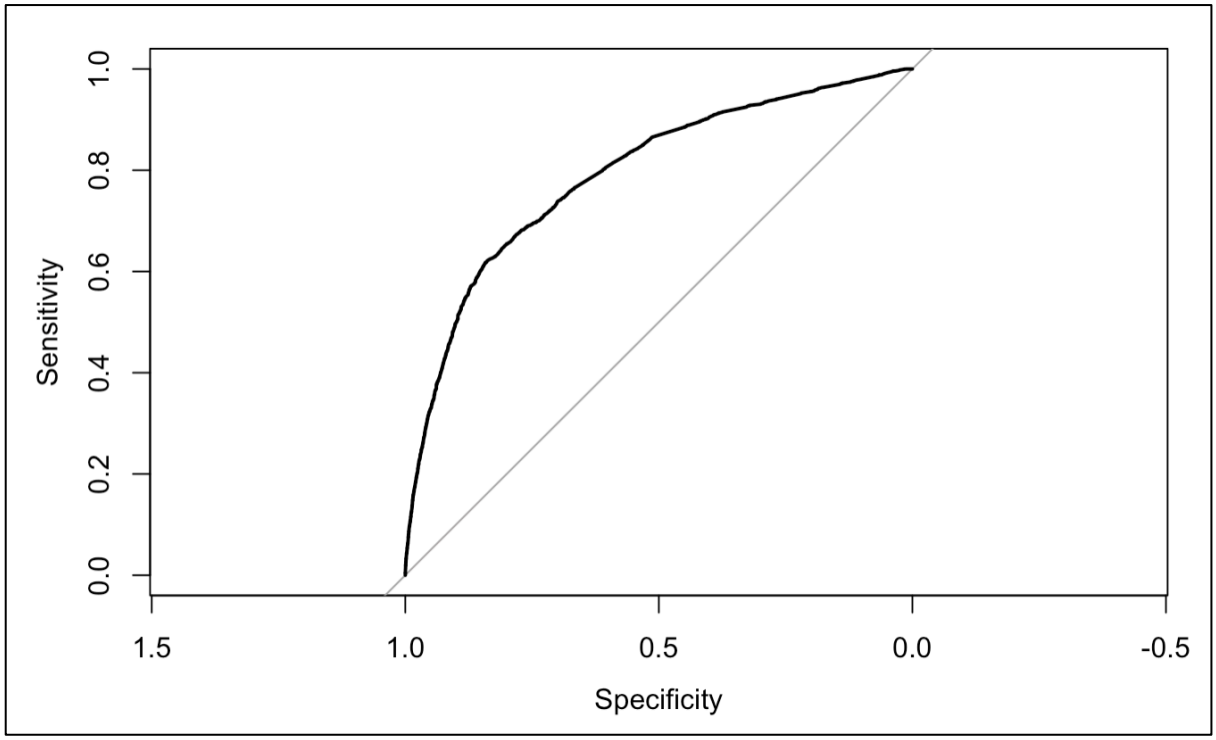
```
    FALSE  TRUE
  0 31701 13862
  1  1290  3555
[1] "CCR: 0.699413"
[1] "Sensitivity: 0.733746"
[1] "Specificity: 0.695762"
[1] "F1: 0.733746"
```



AUC = 78.67

Test Data

```
    FALSE  TRUE
  0 32135 14242
  1  1217  3509
[1] "CCR: 0.697493"
[1] "Sensitivity: 0.742488"
[1] "Specificity: 0.692908"
[1] "F1: 0.742488"
```



AUC = 79.19

Appendix 4: Linear Regression

Model 1: Base Model

```
Call:
lm(formula = log(targdol + 1) ~ slstyr + slslyr + sls2ago + sls3ago +
    slshist + ordtyr + ordlyr + ord2ago + ord3ago + ordhist +
    falord + sprord + lpurseason + year_btwn + yob + shopper +
    slstyr_lyr + slstyr_2ago + slstyr_3ago + slstyr, data = lm_train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1997 -0.5244 -0.0232  0.4787  3.8378

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.4268024  0.0359797  95.243  < 2e-16 ***
slstyr       0.0025726  0.0005972   4.308 1.68e-05 ***
```

```
slslyr       0.0011043  0.0003526   3.132  0.00175 **
sls2ago      0.0016305  0.0004041   4.035 5.55e-05 ***
sls3ago     -0.0002994  0.0002695  -1.111  0.26662
slshist      0.0009850  0.0001320   7.463 9.98e-14 ***
ordtyr      -0.0348885  0.0222380  -1.569  0.11675
ordlyr      -0.0548253  0.0222270  -2.467  0.01367 *
ord2ago     -0.0553321  0.0244886  -2.260  0.02390 *
ord3ago     -0.0070151  0.0234651  -0.299  0.76498
ordhist     -0.1486415  0.0165311  -8.992  < 2e-16 ***
falord       0.1288772  0.0166076   7.760 1.03e-14 ***
sprord       0.1212084  0.0167781   7.224 5.83e-13 ***
lpurseasonS -0.0030354  0.0329037  -0.092  0.92650
year_btwn    0.0251653  0.0079849   3.152  0.00163 **
yob         -0.0053968  0.0026828  -2.012  0.04432 *
shopperS     0.0087895  0.0419016   0.210  0.83386
shopperT    -0.0021971  0.0367642  -0.060  0.95235
slstyr_lyr  -0.0019160  0.0004870  -3.934 8.47e-05 ***
slstyr_2ago  0.0009517  0.0005686   1.674  0.09426 .
slstyr_3ago -0.0003153  0.0005682  -0.555  0.57901
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7379 on 4824 degrees of freedom
Multiple R-squared:  0.1181,   Adjusted R-squared:  0.1145
F-statistic: 32.32 on 20 and 4824 DF,  p-value: < 2.2e-16
```

VIF

```
                GVIF Df GVIF^(1/(2*Df))
slstyr      8.582908  1        2.929660
slslyr      2.466429  1        1.570487
sls2ago     2.947271  1        1.716762
sls3ago     2.326126  1        1.525164
slshist     6.931850  1        2.632841
ordtyr      2.340037  1        1.529718
ordlyr      2.519855  1        1.587405
ord2ago     2.917432  1        1.708049
```

```
ord3ago      2.316404  1        1.521974
ordhist     27.675993  1        5.260798
falord      15.286367  1        3.909778
sprord       6.450446  1        2.539773
lpurseason   2.407390  1        1.551577
year_btwn    1.874912  1        1.369274
yob          1.831786  1        1.353435
shopper      2.939022  2        1.309335
slstyr_lyr   4.056285  1        2.014022
slstyr_2ago  5.838101  1        2.416216
slstyr_3ago  5.926515  1        2.434444
```

Model 2: Keeping interactions that were significant

```
Call:
lm(formula = log(targdol + 1) ~ slshist + ordhist + slstyr +
    slstyr_lyr + sprord + falord + year_btwn + sls2ago + slslyr +
    ord2ago + ordlyr + yob + sls3ago + sls2ago:slslyr + slstyr:sls2ago
+ slslyr:sls3ago, data = lm_train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0518 -0.5235 -0.0290  0.4701  3.7883

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.373e+00  2.510e-02 134.393  < 2e-16 ***
slshist         1.068e-03  1.197e-04   8.921  < 2e-16 ***
ordhist        -1.596e-01  1.503e-02 -10.617  < 2e-16 ***
slstyr          3.065e-03  4.145e-04   7.395 1.65e-13 ***
slstyr_lyr     -1.929e-03  4.515e-04  -4.273 1.97e-05 ***
sprord          1.290e-01  1.429e-02   9.030  < 2e-16 ***
falord          1.316e-01  1.545e-02   8.517  < 2e-16 ***
year_btwn       3.197e-02  6.905e-03   4.630 3.75e-06 ***
sls2ago         3.247e-03  4.907e-04   6.617 4.07e-11 ***
slslyr          2.013e-03  3.725e-04   5.403 6.85e-08 ***
ord2ago        -9.343e-02  2.446e-02  -3.819 0.000136 ***
```

```
ordlyr          -5.778e-02  2.142e-02  -2.697 0.007017 **
yob             -5.100e-03  2.593e-03  -1.967 0.049286 *
sls3ago          6.326e-04  3.349e-04   1.889 0.059001 .
sls2ago:slslyr -1.311e-05  2.390e-06  -5.486 4.33e-08 ***
slstyr:sls2ago -7.180e-06  1.913e-06  -3.753 0.000176 ***
slslyr:sls3ago -9.438e-06  2.448e-06  -3.856 0.000117 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7333 on 4828 degrees of freedom
Multiple R-squared:  0.1285,       Adjusted R-squared:  0.1256
F-statistic: 44.48 on 16 and 4828 DF,  p-value: < 2.2e-16
```

VIF

slshist	ordhist	slstyr	slstyr_lyr	sprord
5.775292	23.171150	4.186855	3.530314	4.738713
falord	year_btwn	sls2ago	slslyr	ord2ago
13.395837	1.419913	4.400438	2.787710	2.948258
ordlyr	yob	sls3ago	sls2ago:slslyr	slstyr:sls2ago
2.369983	1.733352	3.637821	1.890025	1.715154
slslyr:sls3ago				
3.330835				

Model 3: Add average_amt variable and logs to predictors, as necessary

```
Call:
lm(formula = log(targdol + 1) ~ log(slshist + e) + ordhist +
    log(slstyr + e) + slstyr_lyr + sprord + falord + year_btwn +
    log(sls2ago + e) + log(slslyr + e) + log(ord2ago + e) + log(ordlyr +
    e) + yob + log(sls3ago + e) + log(average_amt + e) + log(sls2ago +
    e):log(slslyr + e) + log(slstyr + e):log(sls2ago + e) + log(slslyr +
    e):log(sls3ago + e) + log(average_amt + e):year_btwn, data = lm_train)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5274 -0.5207 -0.0366  0.4660  3.5778

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1.367e+00  1.093e-01  12.515  < 2e-16 ***
log(slshist + e)                 -1.313e-01  3.544e-02  -3.705 0.000214 ***
ordhist                          -8.686e-02  1.581e-02  -5.494 4.12e-08 ***
log(slstyr + e)                   1.973e-03  2.428e-03   0.812 0.416610
slstyr_lyr                       -7.358e-05  3.010e-04  -0.244 0.806915
sprord                            1.271e-01  1.455e-02   8.734  < 2e-16 ***
falord                            1.414e-01  1.525e-02   9.273  < 2e-16 ***
year_btwn                         9.109e-02  3.066e-02   2.971 0.002980 **
log(sls2ago + e)                  1.859e-01  2.519e-02   7.381 1.84e-13 ***
log(slslyr + e)                   7.504e-02  2.421e-02   3.099 0.001954 **
log(ord2ago + e)                 -2.605e-01  3.458e-02  -7.534 5.83e-14 ***
log(ordlyr + e)                  -1.029e-01  3.298e-02  -3.121 0.001812 **

yob                              -3.388e-03  2.744e-03  -1.234 0.217130
log(sls3ago + e)                  4.530e-04  2.385e-03   0.190 0.849361
log(average_amt + e)              4.689e-01  4.485e-02  10.454  < 2e-16 ***
log(sls2ago + e):log(slslyr + e) -2.661e-04  2.824e-04  -0.942 0.346241
log(slstyr + e):log(sls2ago + e) -6.716e-04  2.789e-04  -2.408 0.016059 *
log(slslyr + e):log(sls3ago + e) -1.233e-04  2.919e-04  -0.422 0.672711
year_btwn:log(average_amt + e)   -1.763e-02  8.681e-03  -2.031 0.042333 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7229 on 4817 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared:  0.1533,    Adjusted R-squared:  0.1502
F-statistic: 48.46 on 18 and 4817 DF,  p-value: < 2.2e-16
```

VIF

```
                log(slshist + e)                          ordhist
                       13.434686                        26.315932
                log(slstyr + e)                       slstyr_lyr
                        2.259606                         1.614068
                          sprord                           falord
                        5.053808                        13.412030
                       year_btwn                log(sls2ago + e)
                       28.745046                       230.033639
                log(slslyr + e)                log(ord2ago + e)
                      221.262921                       230.882052
                log(ordlyr + e)                              yob
                      218.703321                         1.994837
                log(sls3ago + e)            log(average_amt + e)
                        1.950289                         8.338922
log(sls2ago + e):log(slslyr + e) log(slstyr + e):log(sls2ago + e)
                        2.089072                         1.991126
log(slslyr + e):log(sls3ago + e)   year_btwn:log(average_amt + e)
                        2.294655                        29.260631
```

Stepwise to get the final model

```
Start:  AIC=-3118.87
log(targdol + 1) ~ log(slshist + e) + ordhist + log(slstyr +
    e) + slstyr_lyr + sprord + falord + year_btwn + log(sls2ago +
    e) + log(slslyr + e) + log(ord2ago + e) + log(ordlyr + e) +
    yob + log(sls3ago + e) + log(average_amt + e) + log(sls2ago +
    e):log(slslyr + e) + log(slstyr + e):log(sls2ago + e) + log(slslyr
+
    e):log(sls3ago + e) + log(average_amt + e):year_btwn

                                  Df Sum of Sq    RSS     AIC
- slstyr_lyr                       1     0.031 2517.6 -3120.8
- log(slslyr + e):log(sls3ago + e)  1     0.093 2517.7 -3120.7
- log(sls2ago + e):log(slslyr + e)  1     0.464 2518.1 -3120.0
- yob                              1     0.796 2518.4 -3119.3
<none>                                       2517.6 -3118.9
- year_btwn:log(average_amt + e)    1     2.155 2519.7 -3116.7
- log(slstyr + e):log(sls2ago + e)  1     3.032 2520.6 -3115.1
- log(ordlyr + e)                  1     5.091 2522.7 -3111.1
```

```
- log(slshist + e)                   1     7.173 2524.8 -3107.1
- ordhist                            1    15.777 2533.4 -3090.7
- log(ord2ago + e)                   1    29.668 2547.3 -3064.2
- sprord                             1    39.866 2557.4 -3044.9
- falord                             1    44.940 2562.5 -3035.3

Step:  AIC=-3120.81
log(targdol + 1) ~ log(slshist + e) + ordhist + log(slstyr +
    e) + sprord + falord + year_btwn + log(sls2ago + e) + log(slslyr +
    e) + log(ord2ago + e) + log(ordlyr + e) + yob + log(sls3ago +
    e) + log(average_amt + e) + log(sls2ago + e):log(slslyr +
    e) + log(slstyr + e):log(sls2ago + e) + log(slslyr + e):log(sls3ago
+
    e) + year_btwn:log(average_amt + e)

                                  Df Sum of Sq    RSS     AIC
- log(slslyr + e):log(sls3ago + e)  1     0.094 2517.7 -3122.6
- log(sls2ago + e):log(slslyr + e)  1     0.468 2518.1 -3121.9
- yob                               1     0.784 2518.4 -3121.3
<none>                                        2517.6 -3120.8
+ slstyr_lyr                       1     0.031 2517.6 -3118.9
- year_btwn:log(average_amt + e)    1     2.142 2519.8 -3118.7
- log(slstyr + e):log(sls2ago + e)  1     3.027 2520.7 -3117.0
- log(ordlyr + e)                   1     5.475 2523.1 -3112.3
- log(slshist + e)                  1     7.187 2524.8 -3109.0
- ordhist                           1    15.826 2533.4 -3092.5
- log(ord2ago + e)                  1    30.221 2547.8 -3065.1
- sprord                            1    40.014 2557.6 -3046.6
- falord                            1    45.063 2562.7 -3037.0

Step:  AIC=-3122.63
log(targdol + 1) ~ log(slshist + e) + ordhist + log(slstyr +
    e) + sprord + falord + year_btwn + log(sls2ago + e) + log(slslyr +
    e) + log(ord2ago + e) + log(ordlyr + e) + yob + log(sls3ago +
    e) + log(average_amt + e) + log(sls2ago + e):log(slslyr +
    e) + log(slstyr + e):log(sls2ago + e) + year_btwn:log(average_amt
+
    e)
```

```
                                    Df Sum of Sq    RSS     AIC
- log(sls3ago + e)                   1     0.102 2517.8 -3124.4
- log(sls2ago + e):log(slslyr + e)   1     0.588 2518.3 -3123.5
- yob                                1     0.808 2518.5 -3123.1
<none>                                           2517.7 -3122.6
+ log(slslyr + e):log(sls3ago + e)   1     0.094 2517.6 -3120.8
+ slstyr_lyr                         1     0.032 2517.7 -3120.7
- year_btwn:log(average_amt + e)     1     2.116 2519.8 -3120.6
- log(slstyr + e):log(sls2ago + e)   1     2.952 2520.7 -3119.0
- log(ordlyr + e)                    1     5.430 2523.1 -3114.2
- log(slshist + e)                   1     7.100 2524.8 -3111.0
- ordhist                            1    16.203 2533.9 -3093.6
- log(ord2ago + e)                   1    30.155 2547.9 -3067.1
- sprord                             1    40.084 2557.8 -3048.2
- falord                             1    45.007 2562.7 -3038.9

Step:  AIC=-3124.44
log(targdol + 1) ~ log(slshist + e) + ordhist + log(slstyr +
    e) + sprord + falord + year_btwn + log(sls2ago + e) + log(slslyr +
    e) + log(ord2ago + e) + log(ordlyr + e) + yob + log(average_amt +
    e) + log(sls2ago + e):log(slslyr + e) + log(slstyr + e):log(sls2ago
+
    e) + year_btwn:log(average_amt + e)

                                    Df Sum of Sq    RSS     AIC
- log(sls2ago + e):log(slslyr + e)   1     0.548 2518.4 -3125.4
- yob                                1     0.884 2518.7 -3124.7
<none>                                           2517.8 -3124.4
+ log(sls3ago + e)                   1     0.102 2517.7 -3122.6
+ slstyr_lyr                         1     0.038 2517.8 -3122.5
- year_btwn:log(average_amt + e)     1     2.104 2519.9 -3122.4
- log(slstyr + e):log(sls2ago + e)   1     2.857 2520.7 -3120.9
- log(ordlyr + e)                    1     5.460 2523.3 -3116.0
- log(slshist + e)                   1     7.121 2524.9 -3112.8
- ordhist                            1    16.102 2533.9 -3095.6
- log(ord2ago + e)                   1    30.182 2548.0 -3068.8
- sprord                             1    40.082 2557.9 -3050.1
```

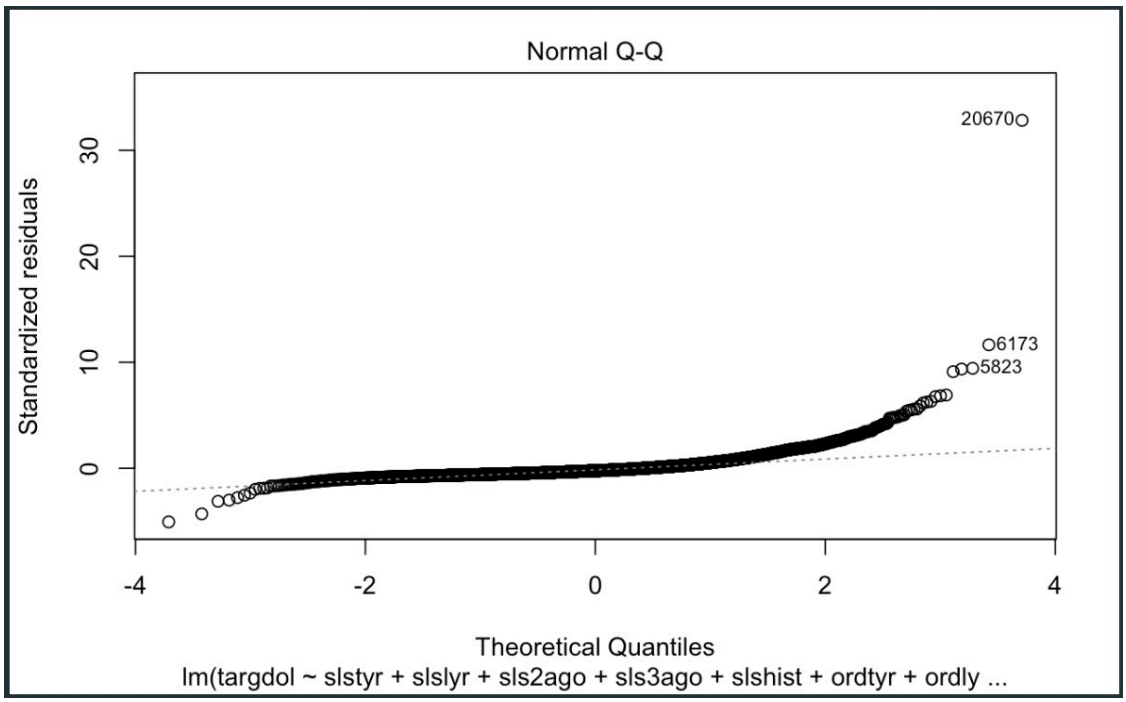
```
- falord                                1    44.976 2562.8 -3040.8

Step:  AIC=-3125.38
log(targdol + 1) ~ log(slshist + e) + ordhist + log(slstyr +
    e) + sprord + falord + year_btwn + log(sls2ago + e) + log(slslyr +
    e) + log(ord2ago + e) + log(ordlyr + e) + yob + log(average_amt +
    e) + log(slstyr + e):log(sls2ago + e) + year_btwn:log(average_amt
+ e)

                                   Df Sum of Sq    RSS     AIC
- yob                               1     0.927 2519.3 -3125.6
<none>                                         2518.4 -3125.4
+ log(sls2ago + e):log(slslyr + e)  1     0.548 2517.8 -3124.4
+ log(sls3ago + e)                  1     0.063 2518.3 -3123.5
+ slstyr_lyr                        1     0.042 2518.3 -3123.5
- year_btwn:log(average_amt + e)    1     2.105 2520.5 -3123.3
- log(slstyr + e):log(sls2ago + e)  1     2.947 2521.3 -3121.7
- log(ordlyr + e)                   1     5.391 2523.8 -3117.0
- log(slslyr + e)                   1     5.613 2524.0 -3116.6
- log(slshist + e)                  1     6.801 2525.2 -3114.3
- ordhist                           1    16.415 2534.8 -3096.0
- log(ord2ago + e)                  1    29.983 2548.3 -3070.2
- sprord                            1    39.820 2558.2 -3051.5
- falord                            1    44.647 2563.0 -3042.4

Step:  AIC=-3125.6
log(targdol + 1) ~ log(slshist + e) + ordhist + log(slstyr +
    e) + sprord + falord + year_btwn + log(sls2ago + e) + log(slslyr +
    e) + log(ord2ago + e) + log(ordlyr + e) + log(average_amt +
    e) + log(slstyr + e):log(sls2ago + e) + year_btwn:log(average_amt
+
    e)

                                   Df Sum of Sq    RSS     AIC
<none>                                         2519.3 -3125.6
+ yob                               1     0.927 2518.4 -3125.4
+ log(sls2ago + e):log(slslyr + e)  1     0.592 2518.7 -3124.7
+ log(sls3ago + e)                  1     0.124 2519.2 -3123.8
```

```
+ slstyr_lyr                          1     0.029 2519.3 -3123.7
- year_btwn:log(average_amt + e)      1     2.097 2521.4 -3123.6
- log(slstyr + e):log(sls2ago + e)    1     3.277 2522.6 -3121.3
- log(ordlyr + e)                     1     5.202 2524.5 -3117.6
- log(slslyr + e)                     1     5.494 2524.8 -3117.1
- log(slshist + e)                    1    10.062 2529.4 -3108.3
- ordhist                             1    16.451 2535.7 -3096.1
- log(ord2ago + e)                    1    29.805 2549.1 -3070.7
- sprord                              1    40.490 2559.8 -3050.5
- falord                              1    44.025 2563.3 -3043.8

Call:
lm(formula = log(targdol + 1) ~ log(slshist + e) + ordhist +
    log(slstyr + e) + sprord + falord + year_btwn + log(sls2ago +
    e) + log(slslyr + e) + log(ord2ago + e) + log(ordlyr + e) +
    log(average_amt + e) + log(slstyr + e):log(sls2ago + e) +
    year_btwn:log(average_amt + e), data = lm_train)

Coefficients:
                      (Intercept)                 log(slshist + e)
                        1.3729305                      -0.1382198
                          ordhist                  log(slstyr + e)
                       -0.0879392                       0.0016146
                           sprord                          falord
                        0.1271182                       0.1394937
                        year_btwn                 log(sls2ago + e)
                        0.0862302                       0.1862698
                 log(slslyr + e)                 log(ord2ago + e)
                        0.0763973                      -0.2593866
                 log(ordlyr + e)              log(average_amt + e)
                       -0.1016961                       0.4755233
log(slstyr + e):log(sls2ago + e)    year_btwn:log(average_amt + e)
                       -0.0006853                      -0.0169909
```

Final Model:

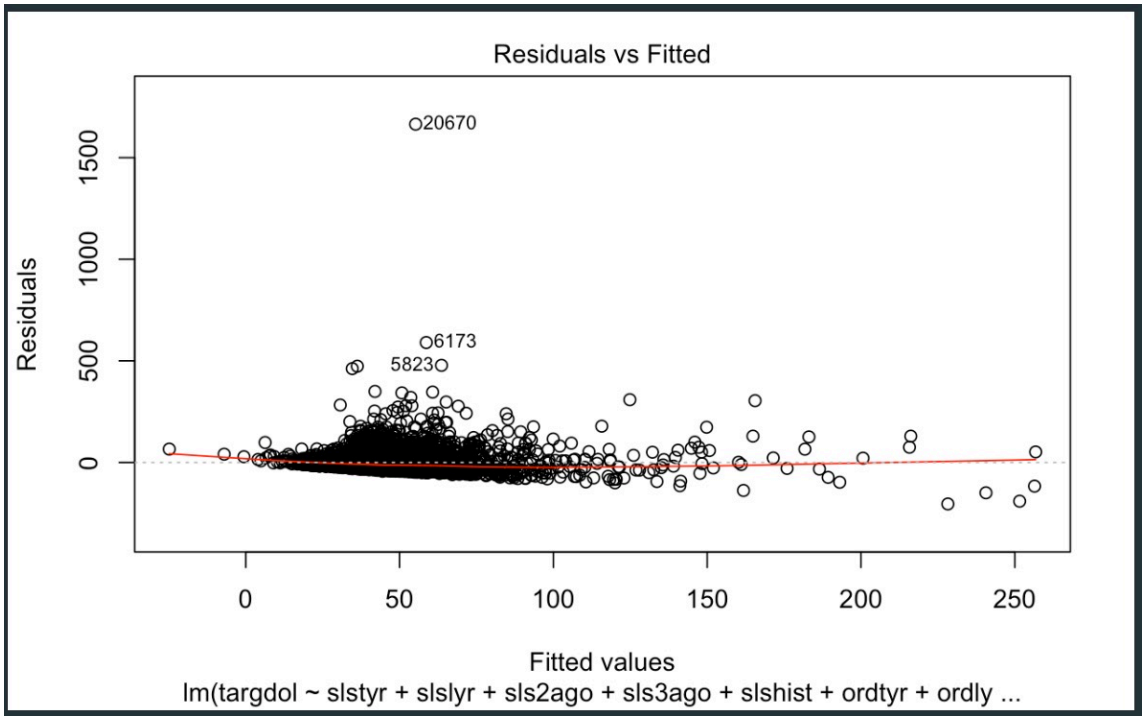
```
Call:
lm(formula = log(targdol + 1) ~ log(slshist + e) + ordhist + 
    log(slstyr + e) + sprord + falord + year_btwn + log(sls2ago + 
    e) + log(slslyr + e) + log(ord2ago + e) + log(ordlyr + e) + 
    log(average_amt + e) + log(slstyr + e):log(sls2ago + e) + 
    year_btwn:log(average_amt + e), data = lm_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5309 -0.5199 -0.0340  0.4674  3.5855 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       1.3729305  0.1056126  13.000  < 2e-16 ***
log(slshist + e)                 -0.1382198  0.0314953  -4.389 1.17e-05 ***
ordhist                          -0.0879392  0.0156717  -5.611 2.12e-08 ***
log(slstyr + e)                   0.0016146  0.0022647   0.713  0.47591    
sprord                            0.1271182  0.0144397   8.803  < 2e-16 ***
falord                            0.1394937  0.0151961   9.180  < 2e-16 ***
year_btwn                         0.0862302  0.0300663   2.868  0.00415 ** 
log(sls2ago + e)                  0.1862698  0.0250009   7.451 1.10e-13 ***
log(slslyr + e)                   0.0763973  0.0235592   3.243  0.00119 ** 
log(ord2ago + e)                 -0.2593866  0.0343420  -7.553 5.06e-14 ***
log(ordlyr + e)                  -0.1016961  0.0322280  -3.156  0.00161 ** 
log(average_amt + e)              0.4755233  0.0404029  11.770  < 2e-16 ***
log(slstyr + e):log(sls2ago + e) -0.0006853  0.0002736  -2.504  0.01230 *  
year_btwn:log(average_amt + e)   -0.0169909  0.0084810  -2.003  0.04519 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7228 on 4822 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared:  0.1527,    Adjusted R-squared:  0.1505 
F-statistic: 66.87 on 13 and 4822 DF,  p-value: < 2.2e-16
```

PLOT 1



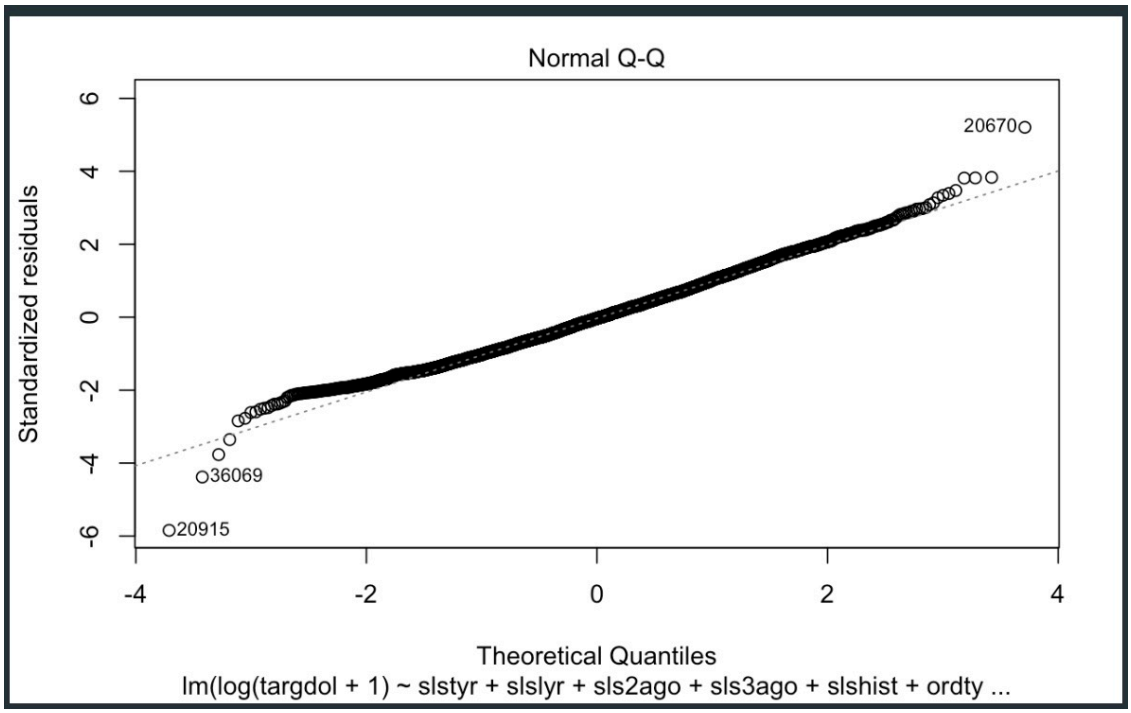
*Note the normality assumption here is a bit shaky towards the tails

PLOT 2



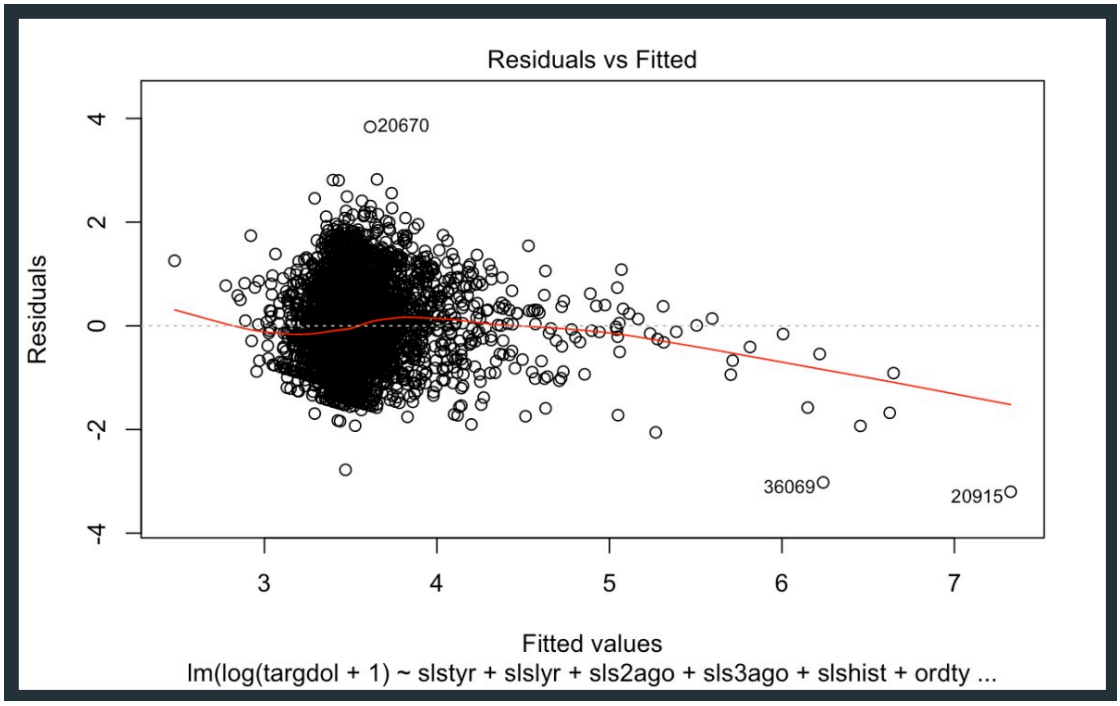
*Note the increasing conic shape of the residuals

PLOT 3



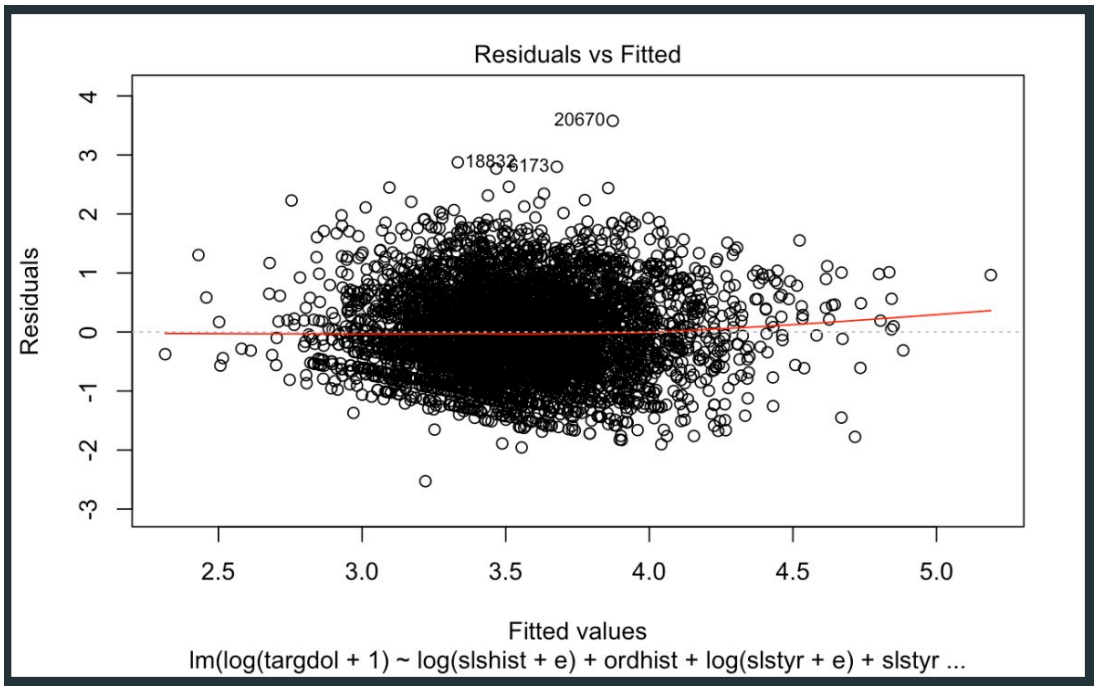
*Standardized residuals closely resemble those of the Normal Distribution

PLOT 4



*Residual plot here no longer resembles the funnel-shape

PLOT 5



*Final residual plot appears to fully satisfy the constant variance requirement

Appendix 5: References

Allison, Paul. "When Can You Safely Ignore Multicollinearity? | Statistical Horizons". *Statisticalhorizons.Com*, 2019, https://statisticalhorizons.com/multicollinearity. Accessed 2 Dec 2019.

Tamhane, Ajit. *Predictive Analytics: Parametric Models For Regression And Classification Using R*. 1st ed., Wiley-Interscience, 2019, pp. 275-279.