Predictive Modeling on Sales Catalog Responses

MSiA 401 Project Report

Anisha Dubhashi, Joe Gilbert, Zili Li, Vincent Wang

I. Executive Summary

We performed an analysis of data on customers of an upscale clothing retail company, aiming to find the customers who would be expected to spend the most with the retail company in response to receiving a catalog from the company in the mail. To do this, we constructed two models: a classification model to predict each customer's probability of placing an order and a regression model to predict each ordering customer's money spent. By combining these, we could identify customers with highest expected sales to target with the catalog.

After building and refining the models, we found some key predictors to be time since a customer had last ordered and the customer's spending in each of the last few years, for predicting that customer's probability of ordering and expected number of dollars spent if they did order, respectively.

Additionally, we were able to predict a set of top 1000 customers who proved to yield a payoff of $44,844.91, about 37.3% of the maximum theoretical payoff from a set of that size. From another perspective, this dollar amount is 19.8% of the total money spent by all the customers in the test set ($226,456.30), but was achieved by targeting only ~2% of those customers. That ratio demonstrates the effectiveness of our models for predicting high-spending customers and giving the retail company the ability to more efficiently target customers to send the catalog.

II. Introduction

In order to maximize profits, it is in any retailer's interest to target advertisements towards customers who are most likely to spend the most money with them as a response to the advertisement. The retailer in this case could send catalogs to every customer in its records, but we know that a relatively small percentage of recipients will make a purchase. Thus, it would be beneficial to identify this subset and send catalogs to only them, and in doing so avoid the waste of sending catalogs to customers who will not make any purchase anyways. Accordingly, the goal of the analysis performed here is to find those customers who are most likely to make purchases, and among those, the ones who are most likely to spend more, in order to allow such targeting advertising.

The data available to us for the analysis come from the retail company of interest, and thus include information about all customers' previous purchases. Additionally, we have the amount each customer spent in response to this catalog, denoted as TARGDOL, which provides us with a response variable to predict. As approximately 90% of the customers in the data set have TARGDOL = 0, we built two different models -- one to classify whether customers will respond to the catalog by making a purchase, and another that predicts, given that a customer responds, how much they will spend. By combining these models, we derived an estimation of expected sales:

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

After specifying our models with the training set, we validated them on the test set with the goal of finding the model whose 1000 top predicted customers had the highest total dollars spent. The model that maximized this total is the one we chose as our final model.

Before fitting the model, we hypothesized what information would contribute the most to the estimation of TARGDOL. Customers who had ordered consistently and recently seemed likely to order again, and customers with high life-to-date sales seemed likely to order more dollars' worth of merchandise. These data are captured by several of the predictor variables available to us, as well as some new ones that we would introduce via interaction between the given predictors. In the following sections, we will discuss the methods used to clean the data and remove inconsistencies between the predictor variables, as well as the steps taken to create relevant new variables to introduce to the model (some but not all of which proved to be significant in the final model). Then we will address the creation of the logistic model to predict whether recipients of the catalog would make a purchase and the multiple linear regression to predict how many dollars each responding customer would spend. Finally, we will discuss the validation of the logistic and linear models on the test data and how we chose our final model with the highest payoff.

III. Model Fitting

A. Data Cleaning

Before beginning any model-fitting, we performed some exploratory data analysis on the data available. This gave us some sense for the distribution within and relationship between

each of the variables. It also revealed places where predictor variables conveying overlapping information disagreed with one another. We spent some time working with these predictors with the goal of creating one universal truth that would hold across all predictors for each customer.

First and foremost among these inconsistencies was information about when each customer's most recent purchase was. Among the predictor variables, we had several places that should have given us the same information along these lines, but in many cases didn't. The predictor DATELP6 gave a calendar date of the customer's last purchase, and the predictor LPURYEAR gave the year of the customer's last purchase if it had happened since 2002 (N/A otherwise). The year from DATELP6 should have agreed with LPURYEAR, but in 7,772 cases (7.7% of the time), it did not (not including the N/A values in LPURYEAR). Additionally, we had a suite of variables indicating sales each year for the last 4 years - SLSTYR, SLSLYR, SLS2AGO, and SLS3AGO; and a corresponding suite with number of orders each year for the last 4 years - ORDTYR, ORDLYR, ORD2AGO, and ORD3AGO. Non-zero values should have been consistent between these 2 sets (for instance, a customer with SLSLYR > 0 should also have ORDLYR > 0, as one can not be responsible for sales without having made at least one order), and the information conveyed by these variables should also have been consistent with the year of last purchase indicated by DATELP6 and LPURYEAR (for instance, a customer with last purchase in 2012 should have ORDTYR > 0, and a customer with last purchase in 2005 should have 0 for ORDTYR, ORDLYR, ORD2AGO, and ORD3AGO).

Our approach for correcting the inconsistencies between the various indicators of the customer's last purchase was to look at all of them and go with the one indicating the most

recent date, then impute on the others accordingly. The intuition behind this approach was that it was likely that the reason for many of the inconsistencies was that one variable had been updated while the others had not. Thus, the most recent suggested last purchase date is most likely to be the truth. We also noticed when examining the SLS and ORD sets of predictors that what they considered to be this year versus last year was not the calendar year, but along the lines of July of one year to June of the next. We had to consider this, what we called the "fiscal year", in order to use the SLS and ORD variables for comparison to and validation of DATELP6 and LPURYEAR.

Besides the information about recency of last purchase, we saw inconsistencies between the variables FALORD, SPRORD, and ORDHIST. FALORD and SPRORD should sum to ORDHIST, but in 8,792 cases (8.7% of the time) this was not the case. Following the same rationale as above stating that one variable may have been updated while the other was not, we fixed this consistency in the following way: When FALORD + SPRORD > ORDHIST, we updated ORDHIST to be equal to the sum. When FALORD + SPRORD < ORDHIST, we used the season of the most recent order as indicated by DATELP6 to decide where to add the difference.

B. New Variable Creation

Since the dates we were provided (DATELP6, LPURYEAR, and dates from the various ORDYR variables) had some discrepancies as discussed above, we wanted to create one standardized variable to represent recency. We decided to create the variable YEARS_SINCE_PURCHASE based on the cleaned date variables in order to exclude DATELP6 and LPURYEAR from the regression models. YEARS_SINCE_PURCHASE represents the number of

years since each customers last purchase from the year 2012. Next, we created both RECENTSEASON and LPMONTH to represent the season (Spring or Fall) and the month of the customers last purchase, respectively.

In addition to the new variables related to the time of each customer's most recent purchase, we wanted to create some interaction variables to capture each customer's consistency of ordering in recent years. We used interactions between various combinations of ORDTYR, ORDLYR, ORD2AGO, and ORD3AGO to create a categorical variable describing levels of consistency each customer had in ordering over the last several years (Table 1).

All of the above was done before attempting to fit any models. As we went deeper into creating and refining the models, we wound up creating more interaction variables to ensure that our models would be able to capture any relationships between multiple predictor variables and TARGDOL. Those additional interactions between predictors will be addressed in the sections below (see Table 2 for all created variables considered).

C. Logistic Regression for Classification

Our next step after creating new variables was to fit multiple logistic regression models in order to classify customers into groups of respondents and non-respondents to the catalog. We began with fitting the full logistic regression model to establish a baseline. After looking at the deviance and AIC results, we decided to look at possible variable transformations to see if we could get better results. We performed some exploratory analyses to understand the relationships between the proportions of people responding to the catalog and the various predictors. We created plots of logistic transforms of proportions responding versus numerical

predictors (Figure 1). We saw that a few relationships were nonlinear; therefore, we transformed the predictors SLSHIST, ORD2AGO, SPRORD, and YEARS_SINCE_PURCHASE with square root transformations, and the predictor ORDTYR with squared transformation.

After transforming some predictors, we need to exclude some variables from the model for multicollinearity reasons as well as logical reasons. We decided not to include DATEAD6 because we did not think that the date the customer was added to the file was a logical predictor of responding to the catalog. Also, the regression models can not include all three variables FALORD, SPRORD, and ORDHIST since FALORD and SPRORD sum to ORDHIST, so we excluded ORDHIST from the model. We chose to exclude ORDHIST from the model because we had previously created a transformation of SPRORD.

After excluding the appropriate predictors and including new predictors, we refit the full logistic regression model, obtaining better results. We then performed a stepwise regression, which excluded additional predictors, resulting in a partial logistic regression model (R Output 1). We also experimented with manually excluding predictors from our stepwise regression that caused multicollinearity. Although there were a couple predictors with VIFs higher than 10, most due to the transformations we created, we decided not to drop the predictors because they were all statistically significant. We used the model from the stepwise regression as our partial model.

In order to compare the full and the partial model, we looked at a couple different diagnostic tests. First, we looked at the AIC for each model: 24,332 for full and 24,297 for partial. Since we want to minimize AIC, it is a good result that the partial AIC is slightly slower than the full AIC. Next, we ran a Chi-squared deviance test (R Output 2) to compare the partial

model to the full model. The test was not significant, indicating that the partial model is not significantly worse than the full model. We also produced ROC curves (Figure 2) to assess the performance of the partial model for varying cutoff probabilities. We quantified the discriminating power of the partial model by calculating the AUC of the ROC curves. The AUCs were the same for both partial and full (AUC = 0.8274), which agrees with the deviance test result. Since the partial model produces similar results to the full model while dropping unnecessary predictors, we use the partial model as our final logistic regression model.

D. Multiple Linear Regression

The first step we took for performing a linear regression of TARGDOL against the predictor variables was to create a subset of the training data consisting only of rows for which TARGDOL > 0. We continued to work just with this subset for the rest of the process of fitting the multiple regression.

Next we created a correlation matrix for TARGDOL and all of the numeric predictor variables, in order to get a sense for which predictors to expect to be most significant as well as to check for instances of multicollinearity to be cautious of. As would be expected, we saw high correlations between the interaction variables and the original variables that contributed to them. Also as expected, we saw high correlation between ORDTYR, ORDLYR, ORD2AGO, and ORD3AGO and their respective SLSTYR, SLSLYR, SLS2AGO, and SLS3AGO. Naturally, the number of orders a customer made in a year would be correlated with the number of dollars they spent in that same year. However, because we imagined these predictors might contribute to the model in different ways, we did not remove them at this point, but rather made a note to look out for them when we checked for multicollinearity after fitting the model. Unfortunately, we

did not see any predictors having high correlation with TARGDOL. Some had higher than others (for instance, SLSHIST had .23 correlation with TARGDOL), so we had some expectation of what might turn out to be significant in a linear model.

After this exploratory analysis, we built an initial linear model of TARGDOL against all of the variables in the final data set (see "New Variable Creation" above). This model had a low R^2 and many insignificant variables, though the ones that were significant matched with our expectations based on correlation. More importantly, this initial model allowed us to create a normal plot (Figure 3) and residual plot (Figure 4) to check the assumptions of normality and homoscedasticity and determine whether a transformation of the response variable was needed. Based on the shape of the normal plot, a log transformation of the response variable was chosen. A second model identical to the first besides this transformation yielded a slightly higher R^2 and a much more linear normal plot (Figure 5). Repeating this with a square root transformation of the response variable even further increased the R^2 but did not solve the normality problem.

Moving forward with specification of the model, we created added variable plots for each of the predictor variables being used (Figure 6). We were hoping for this to reveal any needed transformations of the predictor variables (e.g. square, square-root, log, etc.) but based on the plots none of these transformations would be helpful.

Next, we wanted to ensure that any possibly significant interactions would have a chance of being captured by the model, so we created a number of interaction variables between the SLSTYR, SLSLYR, SLS2AGO, and SLS3AGO variables. While adding these predictors to the untransformed, log-transformed, and square root-transformed models increased the R^2

and the adjusted R^2 in all three cases (additionally, ANOVA tests on all three corresponding model pairs showed that in each case the model including the new predictors was significantly different than the original model), many of the new predictors were not statistically significant. Therefore, we used stepwise regression to derive partial models containing fewer predictors, all of which would ideally be statistically significant. Just as in the creation of the logistic model above, we converted the categorical variable indicating recent consistency of ordering into an array of dummy variables (leaving the reference category out to avoid perfect multicollinearity) so that the stepwise regression would not remove the entire category if only some of the categories were not deemed significant predictors.

Stepwise regressions on the untransformed, log-transformed, and square root-transformed response variables yield three new partial models (R Output 3), each of which was shown by an Extra-SS ANOVA test to not be significantly worse than their corresponding full model (R Output 4). Additionally, in each case, almost all predictor variables were shown to be highly statistically significant.

Next, we checked the normality and homoscedasticity assumptions and addressed outliers and influential variables in these models. The qq-plot of untransformed model still indicated long-tailed data. The homoscedasticity assumption seemed to be satisfied except for a few outliers. For the log-transformed model, the normality assumption was satisfied and fitted value plot looked similar to the untransformed. For the root-transformed model, qq-plot indicated right-skewed data and the fitted value plot was pretty close to the other two. These results are consistent with the checks for normality and homoscedasticity performed before adding more interactions and performing stepwise regression.

At this stage, we checked for multicollinearity in each of the models by looking at the variance inflation factor for each predictor. In the untransformed and log-transformed models, all VIFs were below 10, so multicollinearity was not an issue. In the square root-transformed model, one interaction variable had a VIF of approximately 13. Since this was just 1 variable and the VIF was not much higher than 10, this was deemed to not be a serious issue. As it turned out, the square root-transformed model would not be our best model when it came to validation so the issue was irrelevant.

By checking standardized residuals, we identified outliers in each of the models. For the untransformed model, there were 66 outliers with standardized residual greater than 3. We decided to remove the top 3 outliers which were obs. 41618, 12517, 11788. After removing the outliers, the R^2 increased to 0.1456. For the log-transformed model, there were 16 outliers. After removing the top 3 outliers, 41618, 97678, and 12517, the R^2 increased to 0.1265. For the square root-transformed model, there were 55 outliers. After removing the top 3 outliers, 41618, 12517, and 37929, the R^2 increased to 0.1421. We also used Cook's distance to detect influential observations in our models. Using the Criterion $D_i > 4/(n-p-1)$, there were about 200 influential observations in each model (untransformed, log-transformed, and square root-transformed). We considered that removing all the outliers were unreasonable, because these observations with higher TARGDOL are exactly the type of cases we are hoping to predict. However, for better model fitness we decided to remove the most problematic three influential observations from the models. We removed 48690, 72035, and 100006 from untransformed model, and 48690, 72035, and 72673 from log-transformed model, and removed 48690, 72035, and 38551 from square root-transformed model**.** After removing these outliers, we refitted our

models. The R^2 of the log-transformed and square root-transformed models improved slightly, but the R^2 of the untransformed model actually got worse, so we decided not to remove any influential observations from the untransformed model. We choose the final linear model based on the expected payoffs below.

IV. Model Validation

After fitting models with the training set, we validated their performance via statistical and financial criteria on the test set. The statistical criterion used was the Mean Square Prediction Error of the multiple linear regression, to check how well the linear regression predicted the sales generated from each customer who did respond to the catalog. The financial criterion used was the sum of the actual TARGDOL values for the 1000 customers that our logistic and linear models conjunctively predicted to be the top 1000 spenders.

From a statistical standpoint, we evaluated the multiple linear regression models we had created based on their Mean Square Prediction Error. The model with the lowest value here was the partial model yielded by a stepwise regression that had no transformation on the response variable and had three outliers removed, but no influential observations, which had MSPE = 2530.261. The second best model was the untransformed model and no outliers were removed, which had MSPE = 2538.889. Log-transformed and square root-transformed models had higher MSPE, which were 3018.689 and 2726.8 respectively.

To validate on the financial criterion, we combined our models to generate an expected value of TARGDOL for all customers. The formula used to derive this is the predicted probability

of making an order from the logistic model multiplied by the predicted TARGDOL given that the customer makes an order from the multiple linear model. This is alternatively represented as:

$$E(TARGDOL) = E(TARGDOL|TARGDOL>0)*P(TARGDOL>0)$$

From this formula, we had an expected value of TARGDOL for each customer in the test set. By sorting according to this value and taking the actual TARGDOL of the 1000 customers predicted by our combined model to spend the most, we found the payoff for the partial model that did not use a transformation on TARGDOL and did not remove influential observations (because this was the model that performed the best against the statistical criterion above). This payoff was $44,479.10.

Because we had fit multiple models during the model fitting stage, we also tried using them to compute payoff to see if perhaps, a model that performed less well in various statistical criteria such as R^2 or MSPE might still yield a higher payoff. The log-transformed model had a slightly higher payoff, at $44,587.38. The square root-transformed model yielded slightly more, $44,864.21, but both of these cases also had a worse MSPE. Interestingly, testing the untransformed model built without removal of any variables yielded the highest payoff of all, $44,844.91. Additionally, this finding reinforced our reasoning that removing outliers and influential observations may be damaging, as it is exactly these extraordinary customers we hope to predict. Evidently, even removing just three outliers from the training set used to build the model results in reduced accuracy of predicting high-spending customers. Although it had some problems with normality, this untransformed partial model built using all observations from the training set had the nearly the lowest MSPE and its payoff was the highest of all models tested.

V. Conclusions

Although our final model only yielded a payoff of $44,844.91 from the top 1000 "targeted" customers (maximum payoff from 1000 customers was approximately $120,000, so we only captured 37.3% of maximum potential payoff), randomly selecting a set of customers to target did much worse. Across 10 trials, the payoff from a random set of 1000 customers was on average, only ~$4300. Therefore, while the model is far from perfect, it allows a retailer to much more effectively target customers to send the catalog to.

As for which predictors turned out to be significant in the logistic classification and multiple linear regression prediction models, we saw overlap but also some variables that contributed significantly to one model but not the other. Notably, all of the information on recent sales volume per customer was highly significant in predicting the amount spent, but this was not the case in predicting whether or not they would respond to the catalog. On the other hand, information on time since last purchase was in general significant for predicting whether the customer would respond, but not for predicting amount spent. Another interesting result in the final logistic model was a highly significant and relatively large-in-magnitude negative coefficient on a predictor we had created indicating whether the customer had already spent more this year than they did last year. We can interpret this as evidence that customers are on a "budget", so to speak, so if they have already surpassed last year's spending in this year, they will not make another order this year.

We also had a few hypotheses regarding unavailable information that we thought might have made a significant contribution to the model. For one, data on the income level of

customers likely would have helped predict both whether a customer would order as well as how many dollars they would spend. Additionally, other demographic information such as geographic data (state of residence, rural vs. urban area, etc.) and gender may have also contributed to one or both models. One might argue that if a relationship between these missing variables and TARGDOL does exist, it would be captured by each customer's historical orders. However, including the missing variables directly in the model might help build a more accurate model. Additionally, things like income level and place of residence can change from year to year.

Furthermore, we are told that the retail company makes sales via its website. If there is data recorded about customers' browsing habits (for instance, if they had viewed or even bookmarked web pages for particular products on the website) and it could be tied to the other data we have about the same customer, it could be very helpful in predicting both whether they will order and how much they will spend.

In conclusion, our final logistic and multiple linear regression models allow us to predict which customers are most likely to place large orders from the catalog, and would thus be the most rewarding to target with marketing campaigns, including the mailing of the catalog itself. This insight is of high value to the retail company, because it allows them to greater improve their return on investment in regard to the dispersion of promotional materials.

VI. Appendix

Tables

Table 1: Consistency Interaction Categories

Category	**Definition (mutually exclusive)**	**Final Models Included In**
Consistent 1	Ordered this year, last year, 2 years ago, and 3 years ago	Linear
Consistent 2	Ordered last year, 2 years ago, and 3 years	
Consistent 3	Ordered this year, last year, and 2 years ago	Logistic, Linear
Consistent 4	Ordered this year, last year, and 3 years ago	Logistic
Consistent 5	Ordered this year, 2 years ago, and 3 years ago	Linear
Consistent 6	Ordered this year and last year	
Consistent 7	Ordered last year and 2 years ago	Logistic
Consistent 8	Ordered 2 years ago and 3 years ago	
Consistent 9	Ordered this year and 2 years ago	Logistic, Linear
Consistent 10	Ordered last year and 3 years ago	Logistic
Consistent 11	Ordered this year and 3 years ago	Logistic
Reference	Ordered 1 or fewer times in past 4 years	

Table 2: All Created Variables

Variable	**Definition**	**Final Models Included In**
ConsistencyCategory	See Table 1	See Table 1
years_since_purchase	Number of years between determined year of last purchase and 2012	Logistic
recentseason1	0 for if last purchase occurred in spring, 1 if last purchase occurred in fall	Logistic, Linear
lpmonth	Month of last purchase based on DATELP6	Logistic
slshist_0.5	Square root of SLSHIST	Logistic
ordytr_2	$ORDTYR^2$	Logistic
sprord_0.5	Square root of SPRORD	Logistic
years_since_purchase_0.5	Square root of years_since_purchase	Logistic
slstyr_slslyr	slstyr*slslyr	Linear
slslyr_sls2ago	slslyr*sls2ago	Linear
sls2ago_sls3ago	sls2ago*sls3ago	Linear
slstyr_slslyr_sls2ago	slstyr*slslyr*sls2ago	Linear
slstyr_sls2ago	slstyr*sls2ago	Linear
slstyr_sls3ago	slstyr*sls3ago	Logistic
slstyr_slshist	slstyr*slshist	Logistic
slslyr_slshist	slslyr*slshist	Linear
sls2ago_slshist	sls2ago*slshist	Logistic, Linear
slstyr_slslyr_sls2ago_sls3ago	slstyr*slslyr*sls2ago*sls3ago	Logistic
slstyr_greater_slslyrTRUE	1 if slstyr > slslyr, 0 otherwise	Logistic, Linear
slslyr_greater_sls2agoTRUE	1 if slstyr > sls2ago, 0 otherwise	Linear

Figures

Figure 1: Plots of Logistic Transforms versus Numerical Predictors

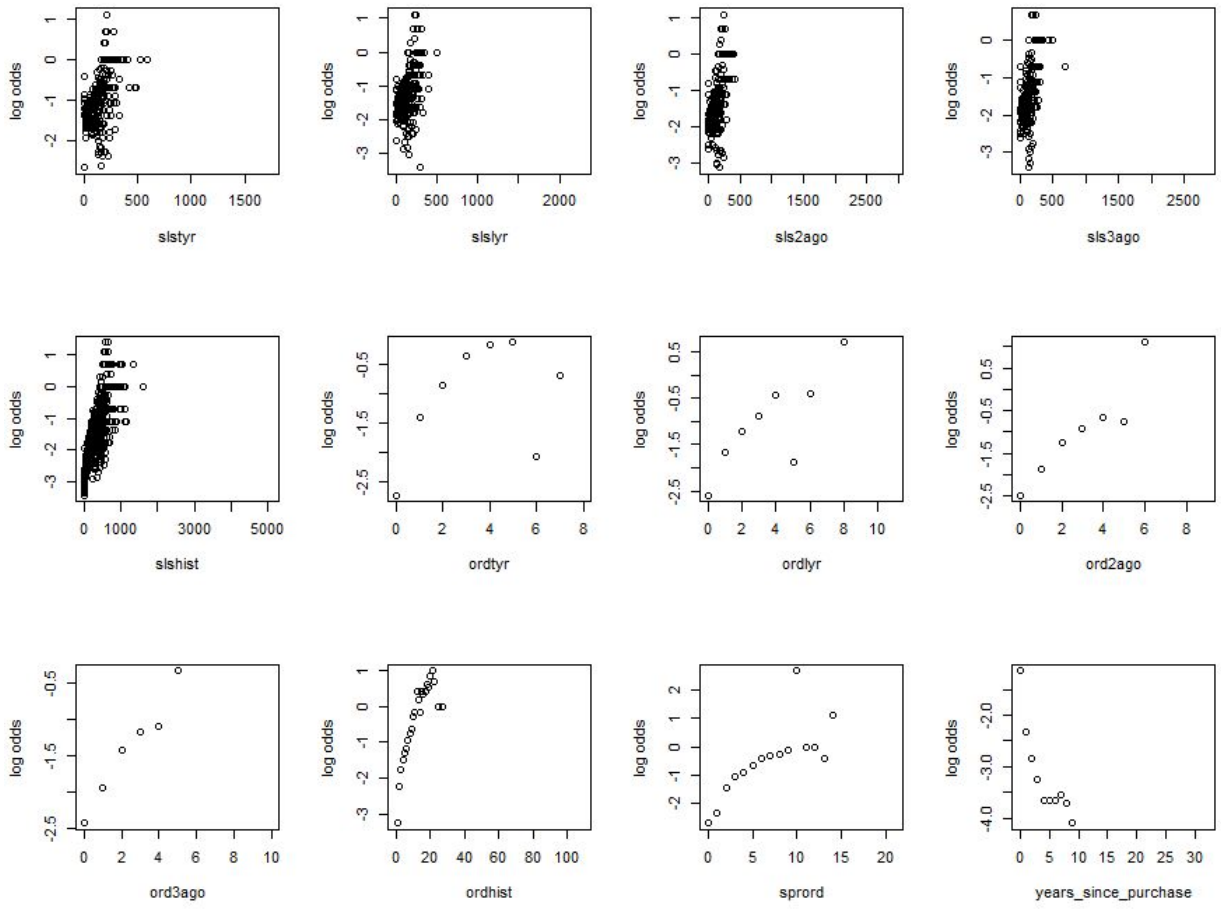


Figure 2: ROC Curve for Partial Logistic Regression Model

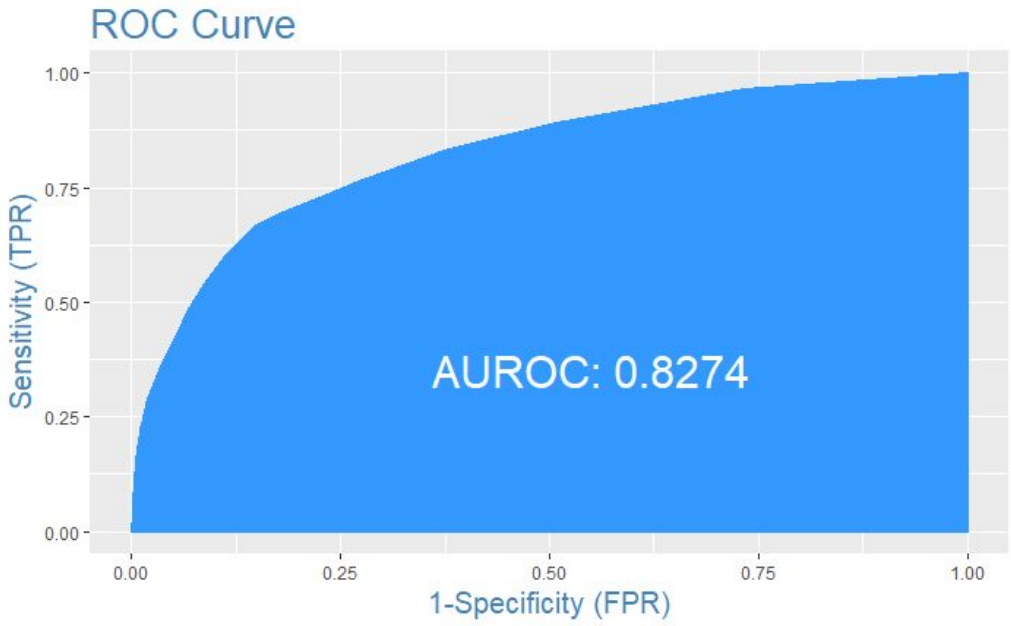


Figure 3: Normal Q-Q plot of Initial Model

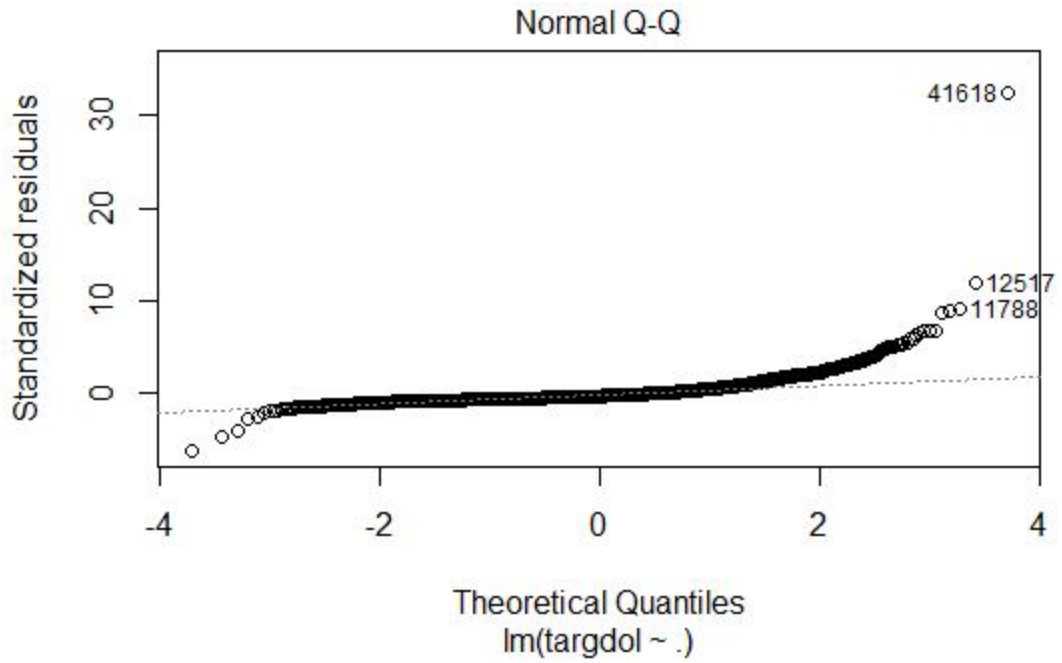


Figure 4: Residual vs. Fitted plot of Initial Model

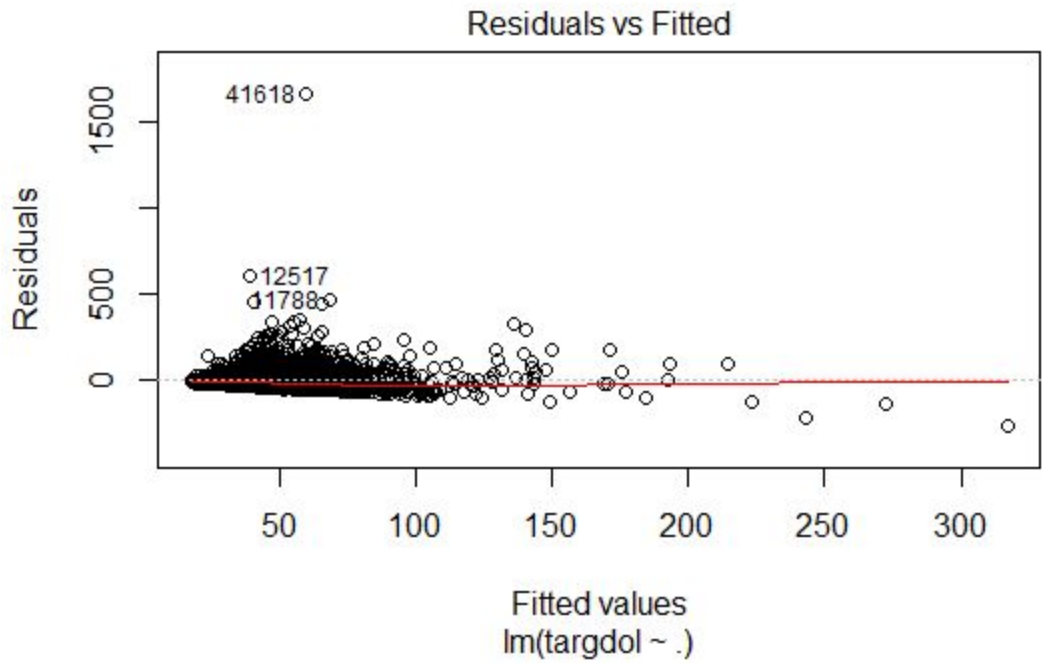


Figure 5: Normal Q-Q plot of Log-Transformed Model

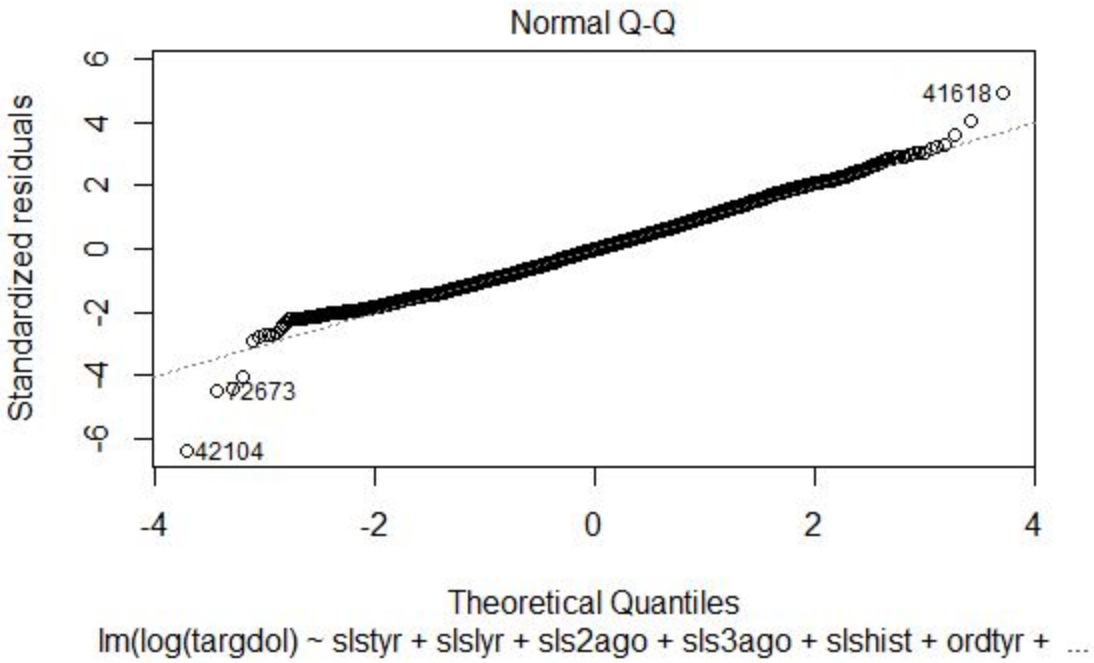
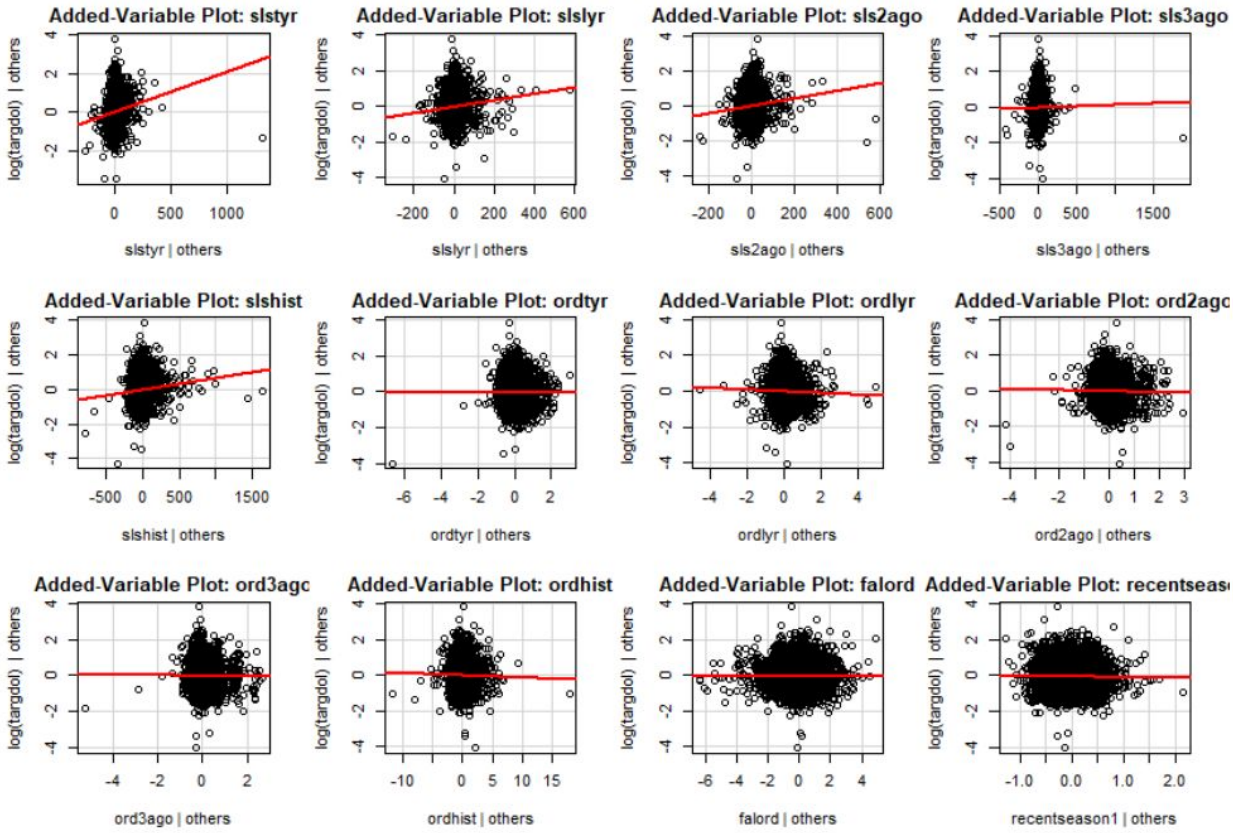


Figure 6: Added-Variable plot of Log-Transformed Model



R Output

R Output 1: Final Logistic Regression Model: Partial Model using Stepwise Regression

```
Call:
glm(formula = targdol ~ sls3ago + slshist + ordtyr + ordlyr +
    falord + sprord + recentseason + years_since_purchase + lpmonth +
    slshist_0.5 + ordtyr_2 + sprord_0.5 + years_since_purchase_0.5 +
    slstyr_slslyr_sls2ago_sls3ago + slstyr_sls3ago + slstyr_slshist +
    sls2ago_slshist + slstyr_greater_slslyr + consistent3 + consistent4 +
    consistent7 + consistent9 + consistent10 + consistent11,
    family = binomial(link = "logit"), data = train_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.6238  -0.4080  -0.2657  -0.1791   2.9691


Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -9.678e-01  1.149e-01  -8.424  < 2e-16 ***
sls3ago                        1.942e-03  4.858e-04   3.996 6.43e-05 ***
slshist                       -3.758e-03  4.333e-04  -8.671  < 2e-16 ***
ordtyr                        -1.170e+00  1.100e-01 -10.630  < 2e-16 ***
ordlyr                        -1.588e-01  3.375e-02  -4.706 2.53e-06 ***
falord                         3.413e-01  1.241e-02  27.510  < 2e-16 ***
sprord                         2.794e-01  3.539e-02   7.894 2.93e-15 ***
recentseason1                 -4.304e+00  1.303e-01 -33.041  < 2e-16 ***
years_since_purchase           6.118e-01  2.758e-02  22.186  < 2e-16 ***
lpmonth                        4.421e-01  1.505e-02  29.369  < 2e-16 ***
slshist_0.5                    7.320e-02  1.093e-02   6.697 2.13e-11 ***
ordtyr_2                       2.288e-01  2.967e-02   7.713 1.23e-14 ***
sprord_0.5                    -5.158e-01  7.661e-02  -6.733 1.66e-11 ***
years_since_purchase_0.5      -3.253e+00  9.372e-02 -34.705  < 2e-16 ***
slstyr_slslyr_sls2ago_sls3ago  7.122e-10  4.901e-10   1.453 0.146169
slstyr_sls3ago                 9.190e-06  5.129e-06   1.792 0.073202 .
slstyr_slshist                 1.764e-06  4.323e-07   4.081 4.49e-05 ***
sls2ago_slshist                1.919e-06  8.374e-07   2.291 0.021959 *
slstyr_greater_slslyrTRUE     -9.346e-01  6.933e-02 -13.481  < 2e-16 ***
consistent3                   -1.826e-01  1.108e-01  -1.648 0.099272 .
consistent4                   -5.048e-01  1.396e-01  -3.616 0.000299 ***
consistent7                    1.578e-01  9.152e-02   1.724 0.084691 .
consistent9                   -1.908e-01  9.855e-02  -1.936 0.052858 .
consistent10                   1.695e-01  1.058e-01   1.602 0.109263
consistent11                  -4.668e-01  1.176e-01  -3.970 7.18e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31907  on 50417  degrees of freedom
Residual deviance: 24247  on 50393  degrees of freedom
AIC: 24297
```

R Output 2: Analysis of Deviance Between Full Logistic Regression Model and Partial Logistic Regression Model

```
Analysis of Deviance Table

Model 1: targdol ~ slstyr + slslyr + sls2ago + sls3ago + slshist + ordtyr +
    ordlyr + ord2ago + ord3ago + falord + sprord + recentseason +
    years_since_purchase + lpmonth + slshist_0.5 + ordtyr_2 +
    ord2ago_0.5 + sprord_0.5 + years_since_purchase_0.5 + slstyr_slslyr +
    slslyr_sls2ago + sls2ago_sls3ago + slstyr_slslyr_sls2ago +
    slstyr_slslyr_sls2ago_sls3ago + slstyr_sls2ago + slstyr_sls3ago +
    slslyr_sls3ago + slslyr_sls2ago_sls3ago + slstyr_slshist +
    slslyr_slshist + sls2ago_slshist + sls3ago_slshist + slstyr_greater_slslyr +
    slslyr_greater_sls2ago + sls2ago_greater_sls3ago + consistent1 +
    consistent2 + consistent3 + consistent4 + consistent5 + consistent6 +
    consistent7 + consistent8 + consistent9 + consistent10 +
    consistent11
Model 2: targdol ~ sls3ago + slshist + ordtyr + ordlyr + falord + sprord +
    recentseason + years_since_purchase + lpmonth + slshist_0.5 +
    ordtyr_2 + sprord_0.5 + years_since_purchase_0.5 + slstyr_slslyr_sls2ago_sls3ago +
    slstyr_sls3ago + slstyr_slshist + sls2ago_slshist + slstyr_greater_slslyr +
    consistent3 + consistent4 + consistent7 + consistent9 + consistent10 +
    consistent11
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1     50371      24238
2     50393      24247 -22  -9.3889   0.9911
```

R Output 3: Final Multiple Linear Regression Model:

```
Residuals:
    Min      1Q  Median      3Q     Max
-217.92  -24.61  -11.51    9.53 1658.84

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.318e+01  1.746e+00  24.730  < 2e-16 ***
slstyr                      2.372e-01  3.157e-02   7.514 6.79e-14 ***
slslyr                      1.719e-01  3.583e-02   4.798 1.65e-06 ***
sls2ago                     1.794e-01  3.740e-02   4.796 1.66e-06 ***
sls3ago                     7.608e-02  2.909e-02   2.615  0.00896 **
slshist                     3.641e-02  7.572e-03   4.808 1.57e-06 ***
ordtyr                     -3.186e+00  1.703e+00  -1.871  0.06140 .
ordlyr                     -3.533e+00  1.777e+00  -1.989  0.04680 *
ord2ago                    -3.783e+00  1.742e+00  -2.171  0.02999 *
ord3ago                    -4.677e+00  1.568e+00  -2.984  0.00286 **
recentseason1              -3.854e+00  1.508e+00  -2.555  0.01065 *
slstyr_slslyr               5.992e-04  3.230e-04   1.855  0.06368 .
slslyr_sls2ago             -1.925e-03  2.826e-04  -6.812 1.08e-11 ***
sls2ago_sls3ago            -5.160e-04  1.285e-04  -4.015 6.04e-05 ***
slstyr_slslyr_sls2ago       1.094e-05  2.188e-06   5.003 5.84e-07 ***
slstyr_sls2ago             -8.791e-04  2.142e-04  -4.105 4.11e-05 ***
slstyr_sls3ago              3.069e-04  2.122e-04   1.446  0.14820
slstyr_slshist             -1.461e-04  2.528e-05  -5.779 7.98e-09 ***
slslyr_slshist             -1.235e-04  5.875e-05  -2.103  0.03553 *
sls2ago_slshist             2.509e-04  5.803e-05   4.324 1.56e-05 ***
slstyr_greater_slslyrTRUE  -3.827e+00  2.510e+00  -1.525  0.12737
slslyr_greater_sls2agoTRUE -4.430e+00  2.588e+00  -1.712  0.08698 .
consistent1                -1.319e+01  3.678e+00  -3.585  0.00034 ***
consistent3                -9.087e+00  4.477e+00  -2.030  0.04245 *
consistent5                -1.561e+01  4.885e+00  -3.195  0.00141 **
consistent9                -9.213e+00  4.602e+00  -2.002  0.04537 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.66 on 4819 degrees of freedom
Multiple R-squared:  0.1142,    Adjusted R-squared:  0.1096
F-statistic: 24.85 on 25 and 4819 DF,  p-value: < 2.2e-16
```

R Output 4: ANOVA Between Full Linear Regression Model and Partial Linear Regression Model

```
Analysis of Variance Table

Model 1: targdol ~ slstyr + slslyr + sls2ago + sls3ago + slshist + ordtyr +
    ordlyr + ord2ago + ord3ago + ordhist + falord + recentseason +
    years_since_purchase + slstyr_slslyr + slslyr_sls2ago + sls2ago_sls3ago +
    slstyr_slslyr_sls2ago + slstyr_slslyr_sls2ago_sls3ago + slstyr_sls2ago +
    slstyr_sls3ago + slslyr_sls3ago + slslyr_sls2ago_sls3ago +
    slstyr_slshist + slslyr_slshist + sls2ago_slshist + sls3ago_slshist +
    slstyr_greater_slslyr + slslyr_greater_sls2ago + sls2ago_greater_sls3ago +
    consistent1 + consistent10 + consistent11 + consistent2 +
    consistent3 + consistent4 + consistent5 + consistent6 + consistent7 +
    consistent8 + consistent9
Model 2: targdol ~ slstyr + slslyr + sls2ago + sls3ago + slshist + ordtyr +
    ordlyr + ord2ago + ord3ago + recentseason + slstyr_slslyr +
    slslyr_sls2ago + sls2ago_sls3ago + slstyr_slslyr_sls2ago +
    slstyr_sls2ago + slstyr_sls3ago + slstyr_slshist + slslyr_slshist +
    sls2ago_slshist + slstyr_greater_slslyr + slslyr_greater_sls2ago +
    consistent1 + consistent3 + consistent5 + consistent9
  Res.Df      RSS  Df Sum of Sq      F Pr(>F)
1   4804 12345047
2   4819 12369495 -15    -24448 0.6342  0.849
```