Predicting Customer Responses to Retail Promotions

Meenakshi Kommineni, Neelam Modi, Mark Noll

Executive Summary

In this project, we used data containing information about customers' historical purchase patterns to predict customer purchases patterns in the future. The first model (logistic regression) predicts the likelihood of a customer making a purchase in the Fall 2012 timeframe. The second model (multiple linear regression) predicts each customer's expected Fall 2012 purchase amount, conditional on them making a purchase.

The qualitative insights from both models largely confirmed our expectations. The most important predictors of whether a customer made a purchase were order quantities in prior years, while the most important predictors of purchase amount were the customers' purchase amounts in prior years, as well as their consistency of purchases. Nonetheless, the predictive power of the joint model was limited. While the classification model yielded a high proportion of correct classifications, the multiple regression model struggled to capture the substantial variation in purchase amounts. Combined, the two models tended to underestimate the total purchase amount in Fall 2012 by roughly 20%. Finally, our analysis excluded customers whose purchase amounts were above $300, compared to a median purchase amount of $32. Special attention should be paid to these customers going forward given their potentially large purchase amounts and impact on total sales.

1. Introduction

To help drive customers to their website, an upscale clothing retail company distributed a catalog mailing in early Fall 2012 to their historical customer-base and recorded any purchases made by them that season. They also maintained a database consisting of the customers' historical sales and order information as of the time of the mailing. The retail company contacted us to see if we could use this data to predict how customers will respond to any future promotions they offer to help identify which customers they should target in the future.

The analytical goal of this project was to identify data points that serve as good predictors of a customers' likelihood of making a purchase and their purchase amount. Our *a priori* hypotheses would be that historical purchases and order quantities could serve as good predictors of future purchase behavior, at least on average, and that more recent behavior would better predict purchases than less recent behavior. To accomplish this goal, we used a three-step approach: (1) Data Cleaning & Exploratory Data Analysis, (2) Model Fitting, and (3) Model Validation. In step (1), we analyzed the given data through various diagnostic tests, addressed any inconsistencies that we observed, and created new variables in the form of interactions, transformations, and indicators. In step (2), we fit the data to two models: a classification model and a multiple linear regression model, which we then combined and validated in step (3) to obtain one final predictive model that we recommend the retail company use alongside their future promotions.

2. Data Cleaning & Exploratory Data Analysis

Cleaning and preparing the dataset was an essential first step before fitting any models. Note, in this section, we commonly use the shorthand names for each of the predictor variables; Table 2a serves as a glossary for what these names refer to. First, we discuss data inconsistencies and outliers. Out of approximately 101,500 customers, there were 17 customers that did not have any sales or order data so they were removed from the dataset. Next, 645 customers had inconsistencies between sales and order data in a given year (i.e., one was greater than zero, while the other was not), so these customers were dropped from the dataset. We also observed that the dataset consisted of several extremely high spenders. While the median purchase amount for customers that made a purchase was roughly $32, and the 99th percentile was $240, there were 52 customers who spent over $300 in Fall 2012 (Figure 2a depicts the purchase amounts of

the roughly 9,500 customers that made a purchase, while Figure 2b shows a histogram). We excluded these customers from the dataset as outliers.

Next, we discuss new variables created. We noticed that the *ordhist* data did not match the summed *falord* and *sprord* data for 2,685 customers. This is most likely due to the retail company not updating all three columns consistently. Because all 2,685 observations in the *ordhist* column were less than the sum of the *falord* and *sprord* columns, we decided to remove *ordhist* and instead create an *ordhist_new* variable which sums *falord* and *sprord*. We created indicator variables for all order and sales quantity variables for the past four years. For example, *ordtyr_bin* equaled 1 if a customer had an *ordtyr* value greater than zero, and 0 otherwise. To capture consistency of customers' recent purchasing behavior, we also created *ordconsistency* and *slsconsistency* interaction variables by summing the order and sales indicator variables, respectively, for the past four years (the two variables are equivalent and are thus never used in the same model; see Figure 2b and Figure 2c). Next, we created sales per order variables for each of the past four years and life-to-date (LTD). The null values in these new variables were imputed with 0. To recency of customer purchases, we created new historical order and sales variables (titled *ordhist3yr* and *ordhist4yr* for example) which focus only on the past four years (rather than the existing LTD variables) as summations of the *ordtyr, ordlyr, ord2ago,* and *ord3ago* data points. Averages of these recent historical order and sales variables were also created (titled *ordhist3yr_avg* and *ordhist4yr_avg* for example). Because the retail company conducted their promotion in the Fall of 2012, we were interested in understanding if Fall order histories played a bigger role in predicting sales than Spring, therefore we created a variable that is the ratio of Fall to Spring LTD order histories titled *falordshare*. Lastly, we noticed that the data in the *lpuryear* and *datelp6* columns did not match in many instances (i.e., the date of last purchase was often stamped after the year listed in the last purchase year column), thus we

created a new variable *lpuryear_new* which reflected the year of the date in the *datelp6* column. A histogram plot of *datelp6* (see Figure 2d) revealed that the data were not distributed evenly across months and were instead concentrated in months March and September each year. Moreover, in about 800 observations in the data, customers were added to the database after their last purchase date, which is unreasonable. There were also instances where sales were made but the *datelp6* and *lpuryear* columns were not updated. Thus, we decided not to use any information contained in *datelp6* or *lpuryear* as part of our model fitting. We instead captured recency of customer purchases via the order/sales history variables as previously described.

Finally, we discuss transformations made to the data to symmetrize the distributions. First, we visualized the distribution of our response variable (excluding those who spent $0) via a histogram (see Figure 2e). This showed that the data were heavily skewed to the right, even after removing the customers who spent over $300, thus we log-transformed *targdol* to obtain *targdol_log*, where targol is the natural log of (*targdol* + 1). This resulted in a more symmetric distribution as shown in Figure 2f. We similarly log-transformed all our sales predictor variables.

3. Model Fitting

We used a two-step model fitting approach involving a classification model (Section 3.1) followed by a multiple regression model (Section 3.2). For customers $i = 1, \ldots, n$, the classification model was used to calculate the probability of a customer making a purchase, i.e., $Pr(y_i > 0)$, and the multiple regression model was used to calculate the expected amount a customer will purchase, assuming that customer makes a purchase, i.e., $E(y_i \mid y_i > 0)$. By multiplying the outputs of both models, we obtained the expected purchase amount $E(y_i) = E(y_i \mid y_i > 0)Pr(y_i > 0) + E(y_i \mid y_i = 0)Pr(y_i = 0) = E(y_i \mid y_i > 0)Pr(y_i > 0)$. Note that we relied on the training dataset (specified by the indicator variable *train* = 1) for all model fitting, and the training dataset for all model validation.

3.1. Classification Model

We used logistic regression to model the response probabilities, $Pr(y > 0)$. First, we initialized the data set by binarizing the response variable, *targdol_log*. The new response variable was *targdol_bin* which equaled 1 if a customer made a purchase (hereafter referred to as a "respondent") and 0 otherwise ("nonrespondent"). Next, we observed that the percentage of respondents was very low (only 9.53% of all customers in the training dataset). To compensate for this, we oversampled respondents by simply duplicating by a factor of m=4, which resulted in a ~30% response rate, which we deemed acceptable for fitting a logistics regression model.

Next, we explored the correlations among each of the predictors that we defined in our Data Cleaning and Exploratory Data Analysis section. Figure 3.1a shows a correlation plot of all predictors. Clearly, there was high correlation between all variables which had binarized counterparts (e.g., *slstyr_log* and *slstyr_bin*). There was also high correlation between sales quantity variables and order quantity variables (e.g., *slstyr_log* and *ordtyr*), sales recency variables (e.g., *slshist4yr* and *slshist3yr*), order recency variables (e.g., *ordhist4yr* and *ordhist3yr)*, and the consistency variables (e.g., *slsconsistency* and *ordconsistency*). Some of these high correlations were indicative of a multicollinearity problem, which we checked by running a logistic regression on all of the predictor variables. We excluded variables that exhibited perfect multicollinearity with other variables and re-fit the model, then calculated the variance inflation factors (VIFs) and observed that many of them were greater than 10 (see Figure 3.1b for the output). To alleviate the strong multicollinearity among predictors, we decided to run the rest of our models on two different subsets of the predictor variables: one which included a majority of sales-related variables and very few order-related variables and the second vice versa. Figure 3.1c shows the updated correlation plots of these subsets.

This still left a large number of predictor variables in each subset of the data, and thus the next step was to apply more refined model fitting and variable selection methods. Our first model (titled "lr_model1") was obtained using backward stepwise regression, which algorithmically eliminated all non-significant predictors (see Table 3.1a for the output). We checked VIFs to ensure there was no multicollinearity and we calculated the area under the ROC curve (AUC) associated with this model and obtained a value of 0.7699 which we deemed to be satisfactory. Our second and third models (titled "lr_model2" and "lr_model3" respectively) were obtained using best subset regression and the C_p criterion (see Table 3.1a for the respective outputs). Again, we checked the VIFs and calculated the AUCs to both be approximately equal to 0.7670. Next, we used Lasso regression to see if any additional predictors could be dropped, but we obtained the exact same output achieved by stepwise regression. As such, we omit the Lasso models from our report to avoid redundancy. Lastly, we ran two simpler logistic regression models based solely on intuition, that are based off of the prior models but exclude certain redundant variables. We say "simple" to refer to the fact that we only included seven predictor variables in each model. The first model (titled "lr_model4") involved only order-related variables and the second (titled "lr_model5" involved primarily sales-related variables. The VIFs showed no multicollinearity and the AUCs (~0.767 each) were similar to the AUCs of the best subset regression models, which had a higher number of predictors.

In total, we ended up with five candidate logistic regression models, all summarized in Table 3.1a. Note that AUC was highest for the stepwise regression model, however this model also included the most predictors thus making interpretation more difficult than the other candidate models. Looking at the coefficients, we observed that the estimates largely agree in terms of magnitude and directionality across all five models. Looking at the coefficient estimate individually, they aligned with our expectations. Customers who had recently ordered from the

company (as captured by *ordtyr*) were more likely to have made a purchase in Fall 2012. Customers who had ordered last year, two years ago, or three years ago were less likely to have made a purchase in Fall 2012. Quantity of Fall/Spring orders and order consistency also had a positive impact on likelihood of making a purchase. Overall, the models favored order-related variables as predictors of whether a customer will purchase, instead of sales-related variables.

3.2. Multiple Regression Model

The goal of the multiple linear regression model was to predict the purchase amount in dollars of each customer, conditional on the customer having made a purchase. Recall that 90% of customers did not make a purchase at all in Fall 2012. This means that all regression models were fit using the approximately 4,800 customers in the training dataset that had a positive purchase amount, as measured by *targdol_log*, the dependent variable for these models.

As in the classification model development process, we first removed multicollinearity in the dataset, keeping variables related to historical purchase amounts (as opposed to order quantities) because our goal was to predict purchase amounts. As discussed above, the models were only fit on customers whose purchase amounts were below $300. This excludes a total of 52 customers in the original data with purchase amounts greater than $300. The range of purchases was from $1 to $300, with a median value of $33 in the training dataset. To begin, we fit a model that included all predictor variables that were not perfectly collinear. The R^2 for this model was 0.14, which provides an upper bound for the R^2 of all other models. We note that this R^2 value implies that only 14% of the variation in the purchase amount could be explained by predictor variables in our dataset, which was fairly low. As discussed later in this paper, this affects the total predictive power of any multiple linear regression model for purchase amount.

Our first non-trivial model was developed using an exhaustive search algorithm using the "regsubsets" command in R. The algorithm identified the best set of predictors, as indicated by

the residual sum of squares (RSS), for different model sizes ranging from one to twenty predictors. For each quantity of predictors, the algorithm selected the model that minimized the RSS. It also returned values of the Adjusted R-squared (R^2_{Adj}) and Bayes Information Criterion (BIC), shown in Figure 3.2a. Since the ultimate goal of modeling was both prediction and interpretability of the model, we selected the model that minimized the BIC, which had eight terms in total, over models with a slightly higher R^2_{Adj} but more model terms. Models with more terms also contained very high levels of multicollinearity, as measured by VIFs.

To develop additional candidate models for prediction, we proceeded to conduct Lasso regression; however, the optimal value of λ selected by cross-validation was zero, implying that no variables were to be excluded except those that were not statistically significant. This resulted in a model with thirteen terms in total, after dropping non-significant variables. To produce another model with fewer predictors, we also fit a Lasso model with a λ value of 0.01, which also produced relatively low mean-squared error (MSE) as shown in Figure 3.2b (note that $log(0.01)$ equals approximately -4.60). This produced a model with nine significant predictors. Lastly, the final algorithmic method that we employed for model selection was forward stepwise regression. This produced the same model as in the Lasso regression with λ equal to zero, therefore we omitted this model from further discussion.

These three algorithm-based models are summarized in the first three columns of Table 3.2a. We note several important observations at this point. First, all models included *slstyr_log*, *slslyr_log*, *sls2ago_log*, *sls3ago_log*, and *slshist_log*. Moreover, the sign and the relative magnitude of these variables' coefficients was consistent with our prior expectations. In these models, the coefficient on *slstyr_log* was approximately 0.19, which implies that all else equal, a 100% increase in purchases in the most recently measured sales year was associated with a 19% increase in purchases of Fall 2012. The slightly lower coefficients on *slslyr_log*, *sls2ago_log*,

and *sls3ago_log* are also consistent with our expectations. The final notable finding was that the coefficient on *slsconsistency* was highly significant and negative. This implies that, conditional on making a purchase and holding all other variables constant, purchase amounts are lower for customers that have bought more consistently in past years.

Next, with these three models already developed, we tested a number of models with fewer predictors. We did this in part because some of the algorithm-based models exhibited fairly high multicollinearity and were less interpretable than a simpler model, and also because we wanted to examine the predictive power of variables not included in these algorithm-based models. The first model we fit, which we refer to as the "Simple" model, removed *ord2ago* and *ord3ago* from the "Exhaustive Search Model", resulting in a model with only seven predictors, each of which had an intuitive coefficient and whose $R^2_{Adj}$was nearly as high as the algorithm-based models. This model is presented in the final column of Table 3.2a.

The additional models we fit included models that: (i) added *falord, sprord,* and *falordshare* as predictors; (ii) removed *slsconsistency* from the models, given its negative sign; (iii) added interaction terms between *slstyear* and *slslyear*, *falordshare*, and other variables; and (iv) used only *slshist4yr_avg*, the four year average purchase amount, as a predictor. However, all of these latter models had R^2_{Adj} values less than 0.06, and were thus not reported in the appendix tables. The failure of these models to produce better models carries several implications. First, because the variables *falord, sprord,* and *falordshare* did not add predictive power, there was no seasonality effect associated with sales. Second, *slsconsistency* was an important predictor of sales, despite its negative coefficient. Third, no interaction terms coefficients were not found to improve model fit relative to the models above. Fourth, the more complicated regression models were better at predicting purchase amounts than a basic heuristic

represented by using the average observed purchase amount over the past four years, as represented by *slshist4yr_avg*.

Based on the above findings, we selected two models as candidates to evaluate jointly with the classification model. First, we selected the "Exhaustive Search: Lowest BIC" model, hereafter titled "mr_model1". Even though the more complicated Lasso models had slightly higher R^2_{Adj} values, they were afflicted with multiple variables with VIFs greater than 10 and were not as interpretable. Our second selected model was the "Simple" model, hereafter titled "mr_model2". This model had an R^2_{Adj} very similar to those of more complicated models but had the additional advantage of being more readily interpretable. Table 3.2a provides a summary of the four multiple regression models that were discussed at length in this section (i.e., the three algorithmic ones and the simple one) including our two chosen candidate models.

3.2.1 Model Diagnostics for "Simple" Multiple Regression Model

We conclude this section with a discussion of several basic multiple linear regression diagnostic checks. To prevent redundancy, we only include the results associated with "mr_model2" here (and correspondingly, in the Appendix figures) because we obtained very similar results for "mr_model1". Our model diagnoses focused primarily on the analysis of residuals. In general, regression models require four conditions to hold: (i) normality of residuals; (ii) constant variance (homoscedasticity) of residuals; (iii) independence of residuals, and (iv) no outlier or overly influential observations.

To check if these requirements were met by our models, we plotted several plots involving residuals (see Figures 3.2c-3.2f). Figure 3.2c shows the normal QQ-plot associated with "mr_model2". The residuals were almost normally distributed, with some slight heavy-tailedness at both extremes. Figure 3.2d shows the residuals vs fitted plot. Note that all variables are on the logarithmic scale, and must be exponentiated to return to the purchase amount in

dollars. The residuals vs. fitted plot in Figure 3.2d showed some heteroscedasticity among residuals, as there are relatively few values in the bottom left corner of the graph. However, this was largely a function of having zero-truncated data in our dependent variable *targdol_log*, revealed by the straight lines in the residuals. Apart from this, the residuals appeared to have roughly constant variance. The initial log transformation helped to ensure this. Figure 3.2e shows the same model fitted to the original values of *targdol*, which clearly violates our regression assumptions. Finally, we checked for outlier and/or influential observations using Cook's distance and hat values in Figure 3.2f. There were no unduly influential observations in the plot, and none of the observations had Cook's distance or leverage values exceeding the thresholds described in the textbook (Tamhane, 2020). Outliers were already partly addressed with our decision to exclude customers with purchase amounts above $300. Finally, it is worth noting that the median absolute value residual was approximately 0.50 on the training data, which implied that for the median customer with a $33 order, our predicted value would differ by $13-$22 depending on the sign of the error. Unfortunately, this difference was fairly large for any individual customer. Regardless of which candidate model we used, differences of this magnitude between the fitted and actual values were largely unavoidable because the large variations in the data could not be explained by any of our predictor variables. Nonetheless, the model still has predictive power and produces mostly intuitive results. For "mr_model2", we presented prediction results for two different customers. One was assumed to have never made any purchases, while the other was assumed to have made $40 purchases for each of the past four years. Our model suggested that the former's expected purchase amount conditional on making a purchase was $26 (below the median), while the latter's expected purchase amount was $35. For a third customer who was assumed to have made $100 purchases in each of the past four years, the expected purchase amount was $70. In addition, the R^2_{Adj} of "mr_model2" was roughly

twice that of a model that predicts purchase amounts based on *slshist4yr_avg* (the average purchase amount over the past four years) alone.

4. Model Validation

Once we had several logistic regression candidate models and multiple regression candidate models, we proceeded to evaluate the overall model fit. Before calculating the expected customer purchase amount, $E[y]$, we performed some preliminaries. First, since we oversampled the respondents in the training dataset used to obtain the logistic regression models, we acknowledged that any response probabilities that we estimate using these models will be biased upwards. Therefore, we applied the formula given in Appendix C of the textbook to "undo" the oversampling procedure and obtain the bias-adjusted estimated response probabilities (Tamhane, 2020). These probabilities, denoted $Pr(y_i > 0)$ for customers $i = 1, \ldots, n$, were calculated by applying each logistic regression candidate model to the test dataset. After obtaining these probabilities, we performed our first model validation by calculating the maximum correct classification rate (CCR) associated with each model. We did this via a numerical search to determine the optimal cutoff probability which yields the highest CCR value (see Table 4a for the results). We observed that all models correctly classified over 91% of the customers, when the cutoff probability was around 0.65. Though our objective was not to perform hard classification in this project, these results gave us confidence in our candidate models' ability to produce reliable soft classifications as well. Next, we applied each of the multiple regression candidate models to the test dataset to obtain the predicted purchase amount of each customer, denoted $E(y_i \mid y_i > 0)$ for customers $i = 1, \ldots, n$.

Finally, by multiplying the predicted purchase amount and the predicted probability together, we predicted the expected purchase amount of each customer $i = 1, \ldots, n$, as follows:$E(y_i) = E(y_i \mid y_i > 0)Pr(y_i > 0)$. Note that because we applied a log-transformation to

the response variable initially, we performed an inverse log transformation to obtain the dollar-value purchase amount predictions. In other words, for each predicted purchase amount value, we calculated: $y_{pred} \leftarrow e^{y_{pred}} - 1$. Because we had five logistic regression candidate models and two multiple regression candidate models, we had a total of ten different predictions for the expected purchase amount of each customer. To make a recommendation to the retail company about which of these ten models they should use to predict promotional responses going forward, we used three different criteria which span statistical, financial, and classification metrics.

Criteria 1: *Mean-Squared Prediction Error (MSE) and* R^2. We compared the predicted purchase amount values with the actual purchase amount values given in the test dataset and obtained the MSE and R^2 values outlined in Table 4b. All ten models perform similarly in terms of both metrics, with an MSE around 0.97 and an R^2 of 0.10. The corresponding residual plots are also provided in Figure 4a. The best model according to this criterion, i.e., the one which minimizes MSE and maximizes R^2, was the one using multiple regression model 1 and logistic regression model 3 (titled "mr_model1 & lr_model3" in Table 4b). Note that these models include fewer predictors than the other candidate models and still fared better statistically.

Criteria 2: *Ability to Predict Purchase Amount from the Top 1,000 Customers.* We sorted the customers in the test dataset based on their predicted purchase amount (highest to lowest) and calculated the total purchase amount of the top 1,000 predicted highest spenders. Then, we sorted the same customers based on their actual purchase amount (again, highest to lowest) and calculated the total purchase amount of the top 1,000 known highest spenders. By dividing these two sums, we obtained the proportion of the top purchase amount that our models captured (see Table 4c for the results). The best model according to this criterion was the one using multiple regression model 2 and logistic regression model 1 (titled "mr_model2 & lr_model1" in Table

4c). In this case, the models predicted a combined purchase amount from the top 1,000 customers to be $1,862.69 when the actual was $2,239.79.

Criteria 3: *Ability to Identify the Top 1,000 Customers*. We defined this criteria because the final model that we choose is also intended to identify a subset of customers to be targeted in future promotions. Similar to Criteria 2, we first sorted the customers in the test dataset based on their predicted purchase amount (highest to lowest) and then based on their actual purchase amount. We compared the customers included in the top 1,000 predicted list to those included in the top 1,000 actual list. Table 4d contains the percentage of correctly identified customers predicted by each model. All models perform similarly, with about 17% overlap of predicted vs. actual top customers. The best model according to the criterion was the one using multiple regression model 1 and logistic regression model 1 (titled "mr_model1 & lr_model1" in Table 4d). In this case, 178 of the top 1,000 customers were accurately identified.

In summary, looking at all three of these criteria together, we chose the model consisting of the multiple regression model 2 and the logistic regression model 3 (jointly referred to as "mr_model2 & lr_model3" in the aforementioned data tables) because it performs above average across all three criteria and it does not include too many predictors thus making it easily interpretable and more easily employable for the retail company (i.e., they do not have to acquire too much customer information to sustain this model). A summary of this model was provided in Table 4e, including a list of the top 10 customers that the company should target in future promotions as identified by this model.

5. Conclusions

In summary, this project provided insights about customer behavior in response to a promotional offer in the retail industry. From the classification model, we observed that order history and order consistency tend to be the best predictors of whether a customer will make a

purchase or not. Our modeling indicated that customers who ordered in recent years were more likely to make a future purchase while the opposite was true for customers who ordered in less recent years. From the multiple regression model, we discovered that recent purchase amounts were the best predictors of future purchases. Our modeling indicated that a 100% increase in purchases in the past four years implies an increase in future purchases of between 15-20%, with purchases in the most immediate prior year having slightly more predictive power than less recent purchases. We also learned from our multiple regression model that customers who consistently made purchases in the past tended to have lower future purchase amounts, all else equal; however, our classification model showed that these customers are more likely to make a purchase, meaning that they are still high-value customers.

Together, the classification model proved to be a strong predictor of whether a given customer will make a purchase, but the multiple regression model proved to be an imperfect predictor of the purchase amounts of these customers, especially for those who were high spenders relative to the median purchase of approximately $32. It is likely that predictions could be improved if additional demographic data about customers were collected, as well as any data related to online marketing interactions. Finally, this analysis focused exclusively on customers with purchases that were below $300 in the Fall 2012 campaign. We would recommend targeting customers with a history of spending at or above this level in the future, given both their high purchase probability and their high levels of historical purchases.

References

Tamhane, A. C. (2020). *Predictive Analytics: Parametric Models for Regression and Classification Using R*. John Wiley & Sons.

Appendix

Tables & Figures

Predictor Variable Name	Description
ordtyr	Number of orders this year[1]
ordlyr	Number of orders last year
ord2ago	Number of orders 2 years ago
ord3ago	Number of orders 3 years ago
falord	Lifetime-to-date (LTD) fall orders
sprord	LTD spring orders
falordshare	LTD share of fall orders [*falord* / (*falord* + *sprord)*]
ordhist_new	LTD order numbers, equal to sum of *falord* and *sprord* (the existing variable *ordhist* was dropped)
ordhist4yr	Cumulative 4 year order quantity
ordhist3yr	Cumulative 3 year order quantity
ordhist2yr	Cumulative 2 year order quantity
ordconsistency & *slsconsistency*	Count of years the customer has made a purchase in the past 4 years; values are integers from 0-4
ordhist4yr_avg	Average order quantity per year over last 4 years
ordhist3yr_avg	Average order quantity per year over last 3 years
ordhist2yr_avg	Average order quantity per year over last 2 years
ordtyr_bin	Indicator of orders this year (1 if *ordtyr* > 0, 0 otherwise)
ordlyr_bin	Indicator of orders last year (1 if *ordlyr* > 0, 0 otherwise)
ord2ago_bin	Indicator of orders 2 years ago (1 if *ord2ago* > 0, 0 otherwise)
ord3ago_bin	Indicator of orders 3 years ago (1 if *ord3ago* > 0, 0 otherwise)
slstyr_log	Log-transformed sales quantity (i.e. purchase amount) this year, given by formula ln(*slstyr* + 1)

[1] Note: We determined that the company used a calendar year of September 1 to August 31, though we did not use periods of time less granular than one year in our models because of data inconsistency as described in Section 2.

slslyr_log	Log-transformed sales quantity this year
sls2ago_log	Log-transformed sales quantity 2 years
sls3ago_log	Log-transformed sales quantity 3 years ago
slshist_log	Log-transformed LTD sales history
slshisr4yr_log	Log-transformed cumulative sales quantity for past 4 years
slshist3yr_log	Log-transformed cumulative sales quantity for past 3 years
slshist2yr_log	Log-transformed cumulative sales quantity for past 4 years
slstyr_bin	Indicator of sales this year (1 if *slstyr* > 0, 0 otherwise)
slslyr_bin	Indicator of sales last year (1 if *slslyr* > 0, 0 otherwise)
sls2ago_bin	Indicator of sales 2 years ago (1 if *sls2ago* > 0, 0 otherwise)
sls3ago_bin	Indicator of sales 3 years ago(1 if *sls3ago* > 0, 0 otherwise)
slsordhist	LTD sales per order
slsordtyr	Sales per order this year
slsordlyr	Sales per order last year
slsord2ago	Sales per order 2 years ago
slsord3ago	Sales per order 3 years ago
slshist4yr_avg	Average sales quantity per year over last 4 years
slshist3yr_avg	Average sales quantity per year over last 3 years
slshist2yr_avg	Average sales quantity per year over last 2 years
train	Training/ test set indicator (1= training, 0 = test)

Table 2a. Glossary of all predictor variables, after data cleaning and exploratory data analysis

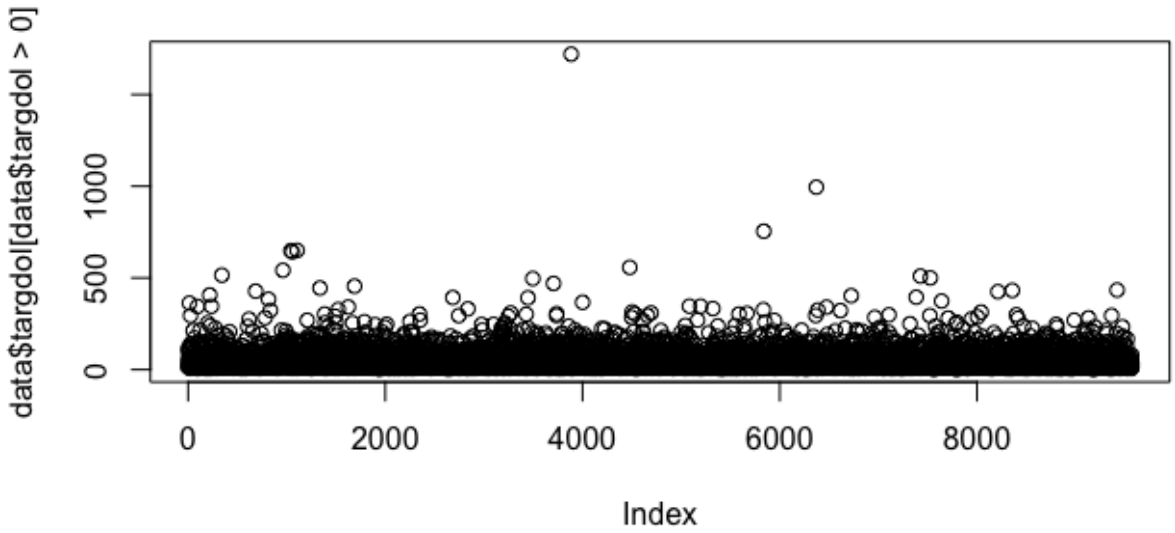

Figure 2a. Scatterplot (one dot per customer) of non-zero purchase amounts in the full dataset. The 99th percentile purchase amount was $240. In our final dataset, we dropped fifty two customers whose purchase amount exceeded $300.

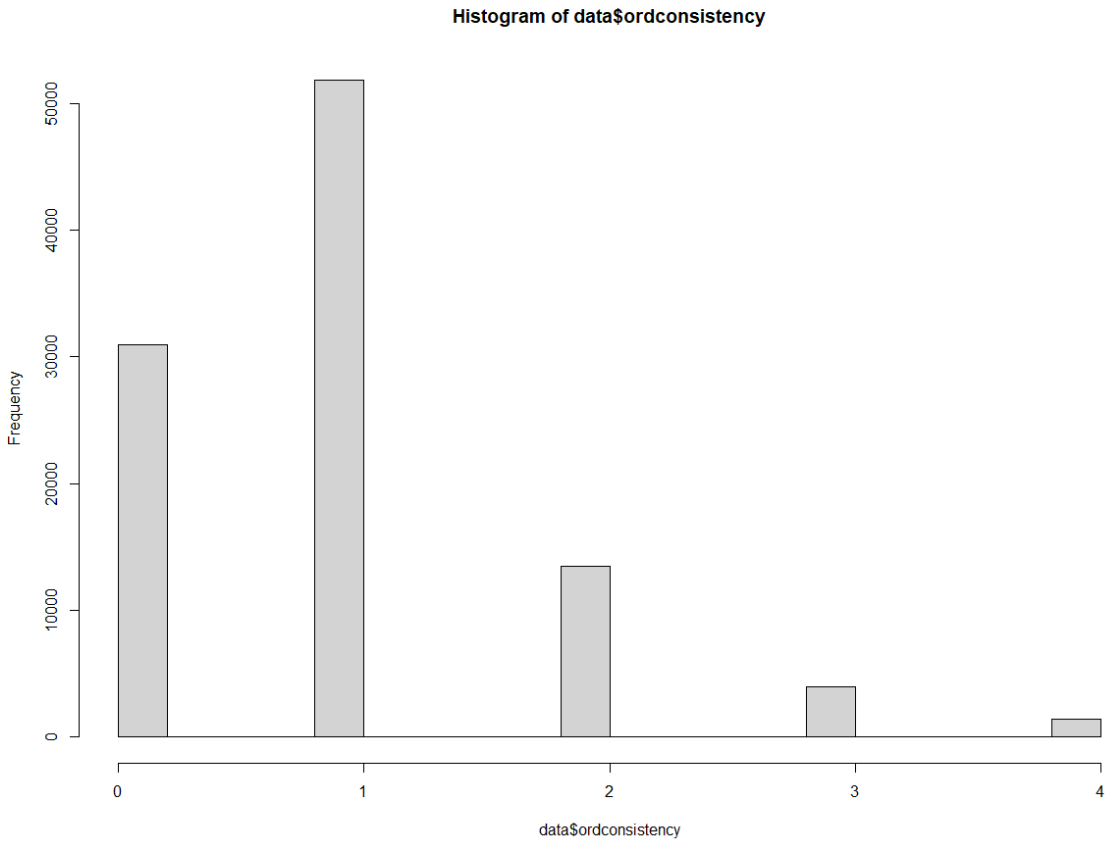

Figure 2b. Histogram of order consistency

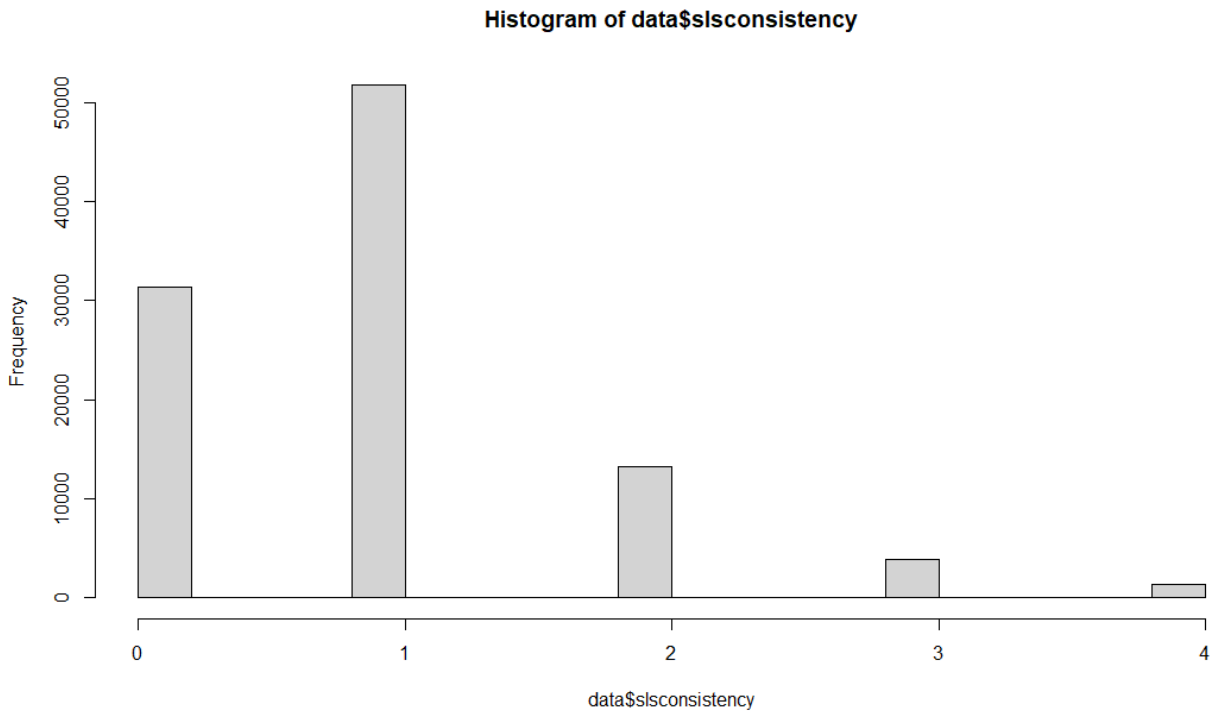

Figure 2c. Histogram of sales consistency

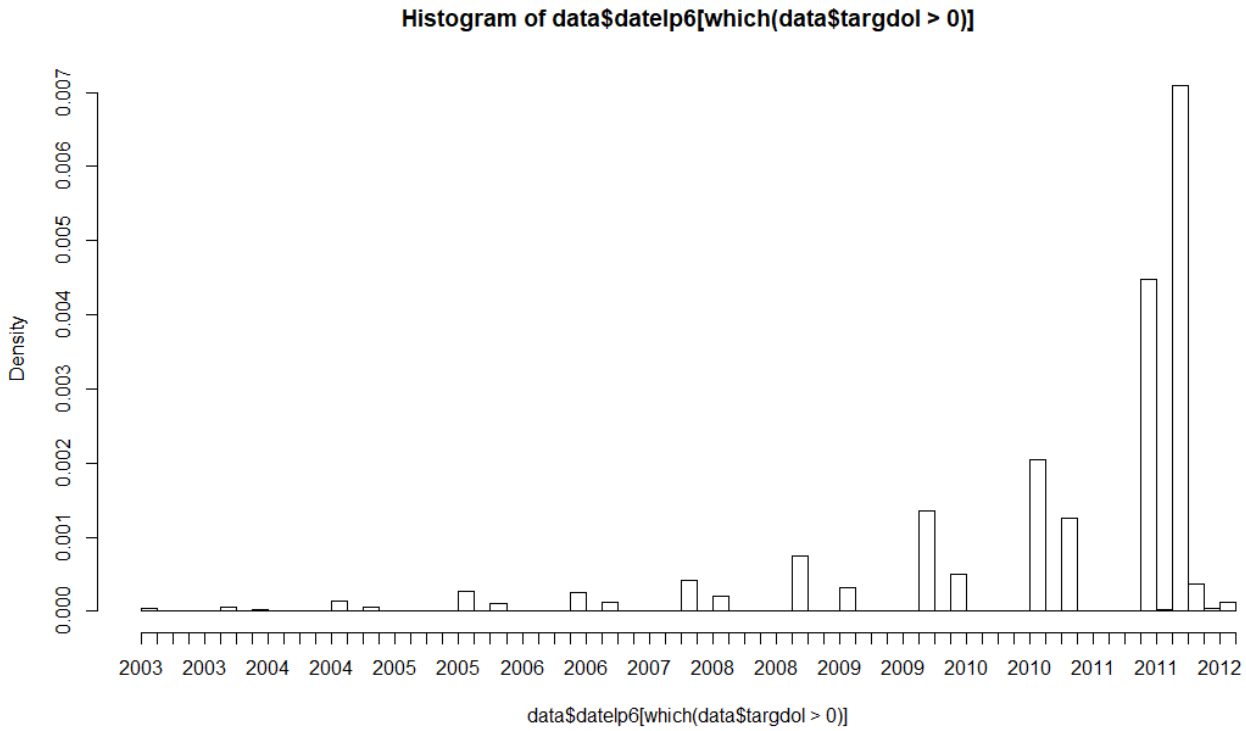

Figure 2d. Histogram of date of last purchase

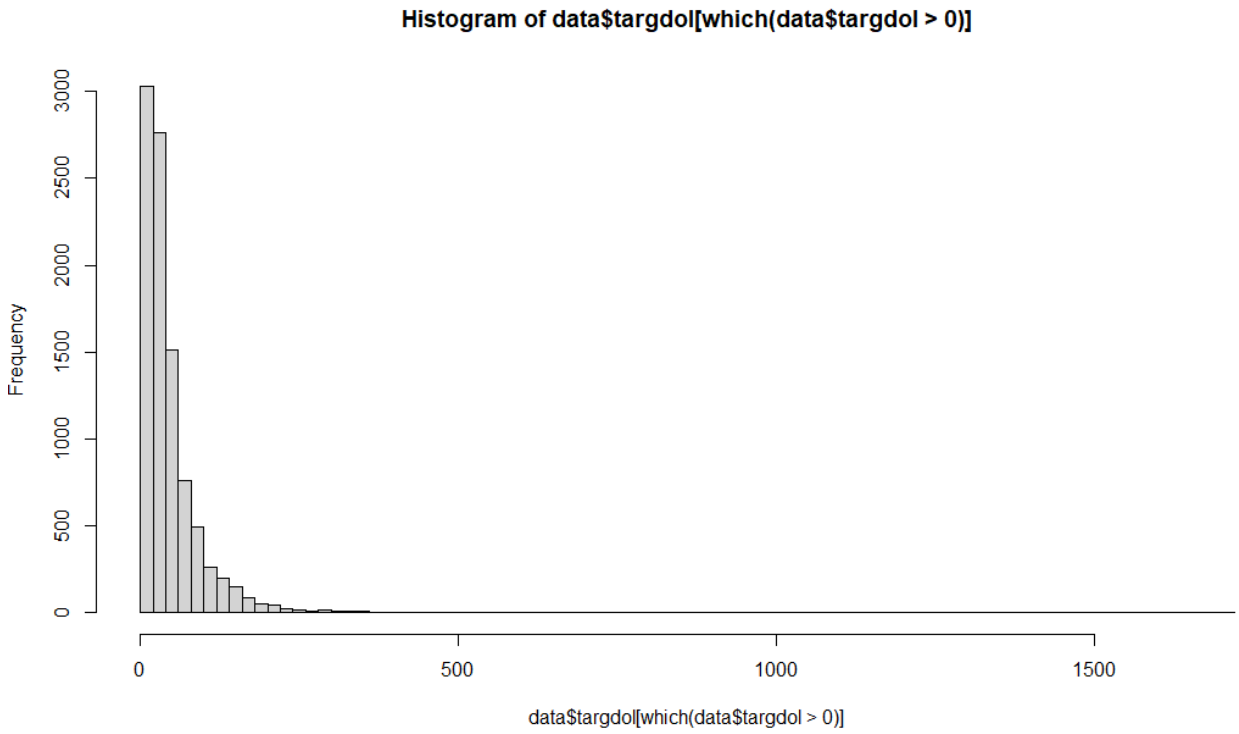

Figure 2e. Histogram of dollar purchase (tagdol) resulting from catalog mailing

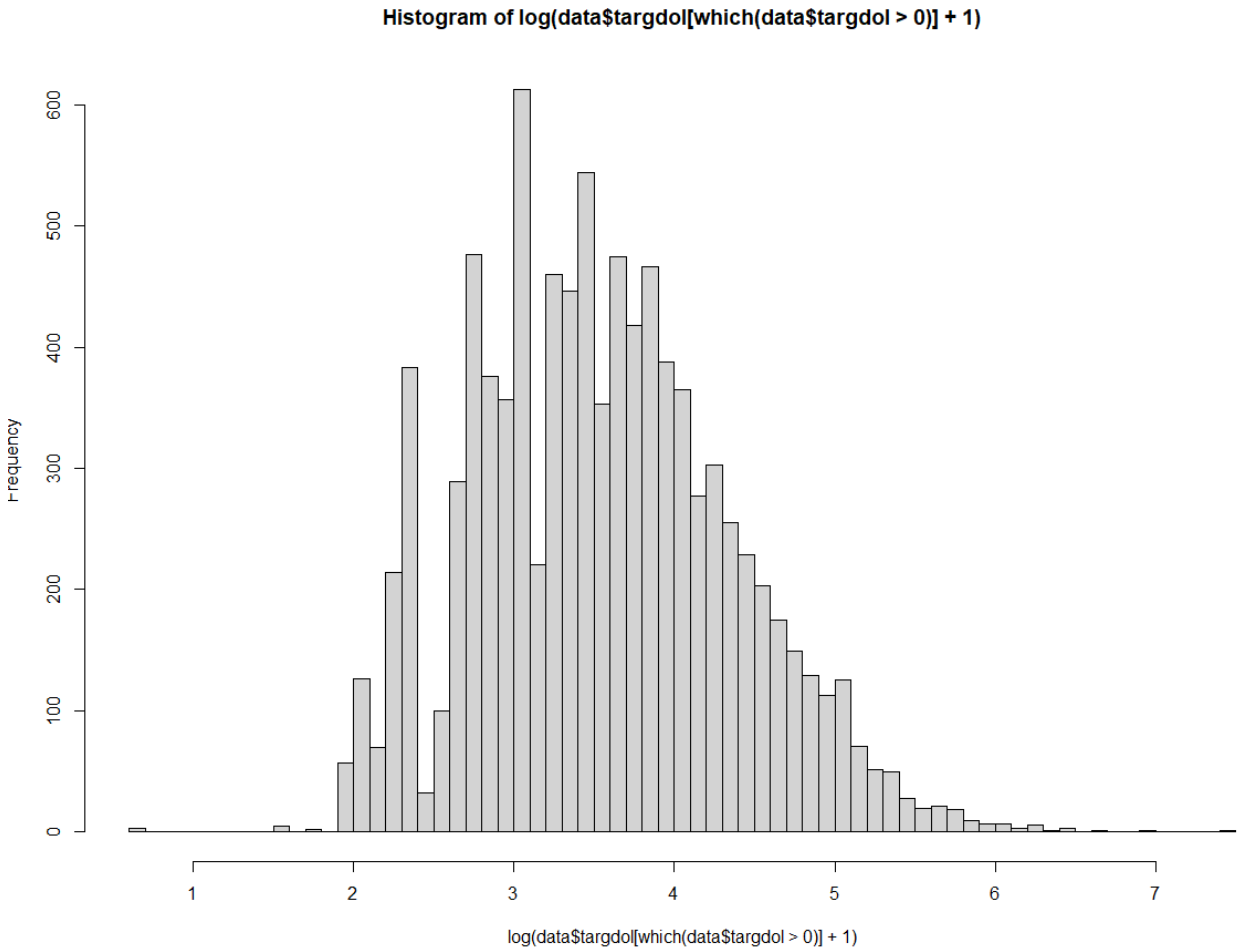

Figure 2f. Histogram of log-transformed targdol

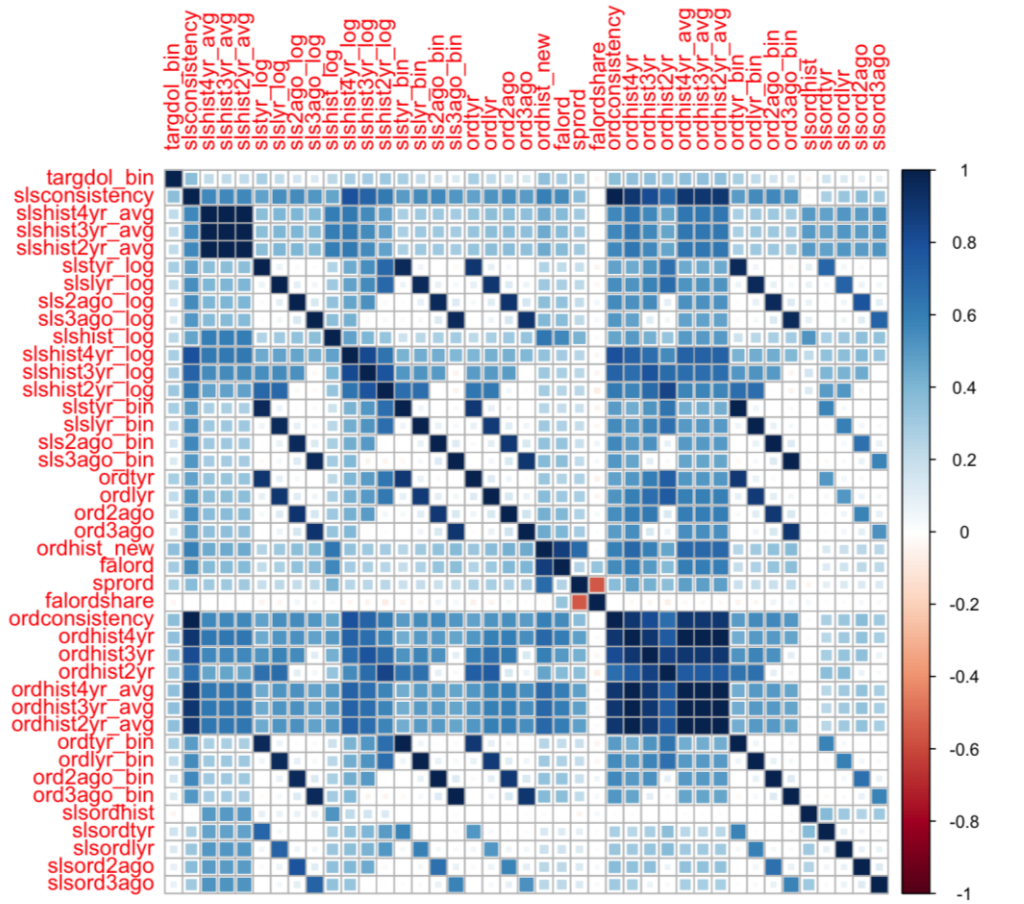

Figure 3.1a. Correlation plot among all predictor variables

```
Call:
glm(formula = respondent ~ ., family = "binomial", data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.2840  -0.7445  -0.5439   0.8183   4.0580

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.5926339  0.0776657 -33.382  < 2e-16 ***
slshist4yr_avg -0.0016452  0.0013846  -1.188 0.234769
slstyr_log     -0.0161494  0.0318436  -0.507 0.612051
slslyr_log     -0.0881539  0.0317440  -2.777 0.005486 **
sls2ago_log    -0.2120783  0.0323763  -6.550 5.74e-11 ***
sls3ago_log    -0.2075733  0.0304354  -6.820 9.10e-12 ***
slshist_log     0.0841415  0.0267514   3.145 0.001659 **
slshist4yr_log  0.1468400  0.0162130   9.057  < 2e-16 ***
slshist3yr_log  0.0368205  0.0193600   1.902 0.057186 .
slshist2yr_log  0.0443685  0.0193742   2.290 0.022016 *
ordtyr         -0.0492602  0.0475902  -1.035 0.300626
ordlyr         -0.1718973  0.0439982  -3.907 9.35e-05 ***
ord2ago        -0.0314185  0.0473788  -0.663 0.507245
ord3ago        -0.1826821  0.0512985  -3.561 0.000369 ***
ordhist_new     0.3824932  0.0167993  22.768  < 2e-16 ***
falord         -0.1665456  0.0187252  -8.894  < 2e-16 ***
falordshare     0.1663255  0.0418195   3.977 6.97e-05 ***
ordconsistency  0.6573846  0.0595877  11.032  < 2e-16 ***
slsordhist     -0.0086964  0.0006718 -12.945  < 2e-16 ***
slsordtyr       0.0039901  0.0008190   4.872 1.11e-06 ***
slsordlyr       0.0029091  0.0008034   3.621 0.000293 ***
slsord2ago      0.0036679  0.0009032   4.061 4.89e-05 ***
slsord3ago      0.0058249  0.0008436   6.904 5.04e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 78238  on 64336  degrees of freedom
Residual deviance: 65375  on 64314  degrees of freedom
AIC: 65421

Number of Fisher Scoring iterations: 5

[1] "--Vif--"
slshist4yr_avg     slstyr_log     slslyr_log    sls2ago_log    sls3ago_log    slshist_log
      9.768838      34.456012      33.534193      32.526184      26.151968       7.345738
slshist4yr_log slshist3yr_log slshist2yr_log         ordtyr         ordlyr        ord2ago
      7.760308      13.709138      15.192414       9.272303       8.659856       9.346313
       ord3ago    ordhist_new         falord    falordshare ordconsistency     slsordhist
      9.736113      12.430228       9.917346       2.725782      29.891484       5.540096
     slsordtyr      slsordlyr     slsord2ago     slsord3ago
      7.327113       6.477615       6.623802       6.649896
```

Figure 3.1b. Logistic regression on all variables (excluding those with perfect multicollinearity) and the corresponding VIFs shows high multicollinearity

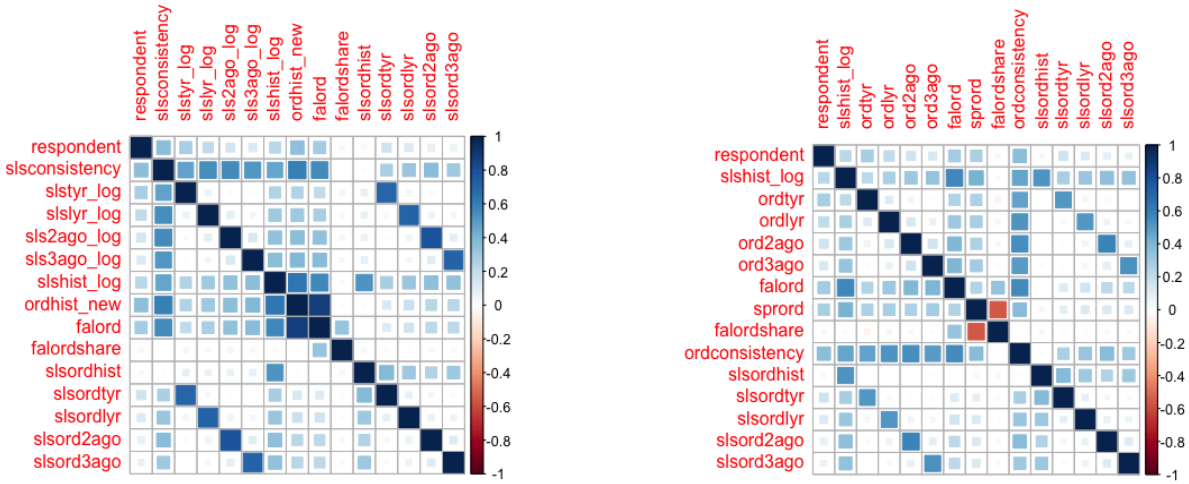

Figure 3.1c. Correlation plot among the two subsets of predictors, one involving primarily sales-related variables (left) and the other involving primarily order-related variables (right)

Model Description	**Stepwise Regression Model**		**Best Subset Regression Model focused on Sales-Related Data**		**Best Subset Regression Model focused on Order-Related Data**		**"Simple" Regression Model focused on Order-Related Data**		**"Simple" Regression Model focused on Sales-Related Data**	
Model Name	**lr_model1**		**lr_model2**		**lr_model3**		**lr_model4**		**lr_model5**	
AUC on Training Data	0.7699		0.7674		0.7679		0.7667		0.7675	
Coefficient	**Estimate**	**Signif.**	**Estimate**	**Signif.**	**Estimate**	**Signif.**	**Estimate**	**Signif.**	**Estimate**	**Signif.**
(Intercept)	-2.376	***	-2.394	***	-2.076	***	-2.241	***	-2.228	***
ordtyr	0.238	***	-	-	0.240	***	0.334	***	-	-
ordlyr	-0.053	*	-	-	-	-	-0.045	*	-	-
ord2ago	-0.213	***	-	-	-0.252	***	-0.293	***	-	-
ord3ago	-0.411	***	-	-	-0.383	***	-0.431	***	-	-
falord	0.171	***	-0.126	***	0.221	***	0.222	***	-0.129	***
sprord	0.353	***	-	-	0.350	***	0.359	***	-	-
ordhist_new	-	-	0.307	***	-	-	-	-	0.344	***
falordshare	0.154	***	-	-	-	-	-	-	-	-
ordconsistency	0.578	***	-	-	0.641	***	0.690	***	-	-
slsconsistency	-	-	0.623	***	-	-	-	-	0.651	***
slsordhist	-0.008	***	-0.004	***	-0.004	***	-	-	-	-
slsordtyr	0.006	***	-	-	0.004	***	-	-	-	-
slsordlyr	0.003	***	-	-	-	-	-	-	-	-
slsord3ago	0.002	***	-	-	-	-	-	-	-	-
slshist_log	0.104	***	0.103	***	-	-	-	-	-	-
slstyr_log	-	-	0.147	***	-	-	-	-	0.135	***
sls2ago_log	-	-	-0.091	***	-	-	-	-	-0.100	***
sls3ago_log	-	-	-0.126	***	-	-	-	-	-0.137	***
Significance Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1										

Table 3.1a. Summary of logistic regression candidate models

Model Description	**Full Model without Non-Significant Predictors (Lasso with λ = 0)**		**"Exhaustive Search" Model : Lowest BIC**		**Lasso with λ= 0.01**		**"Simple" Model**	
Model Name	*(no name - not used)*		**mr_model1**		*(no name - not used)*		**mr_model2**	
R^2_{Adj} on Training Data	0.1358		0.1308		0.1319		0.1268	
Coefficient	**Estimate**	**Signif.**	**Estimate**	**Signif.**	**Estimate**	**Signif.**	**Estimate**	**Signif.**
(Intercept)	3.26	***	3.26	***	3.28	***	3.27	***
slstyr_log	0.25	***	0.19	***	0.23	***	0.19	***
slslyr_log	0.18	***	0.17	***	0.16	***	0.17	***
sls2ago_log	0.20	***	0.20	***	0.19	***	0.17	***
sls3ago_log	0.17	***	0.20	***	0.19	***	0.17	***
slshist_log	0.06	***	0.07	***	0.06	***	0.07	***
slsconsistency	-0.50	***	-0.65	***	-0.59	***	-0.67	***
ord2ago	-0.10	**	-0.11	***	-0.11	***	-	-
ord3ago	-0.10	**	-0.11	**	-0.11	**	-	-
slshist3yr_log	0.04	**	-	-	-	-	-	-
slshist2yr_log	-0.07	***	-	-	-	-	-	-
slstyr_bin	-0.32	**	-	-	-0.24	**	-	-
sls2ago_bin	-0.24	*	-	-	-	-	-	-
ordlyr	-0.08	**	-	-	-	-	-	-
Significance Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1								

Table 3.2a. Summary of multiple linear regression candidate models

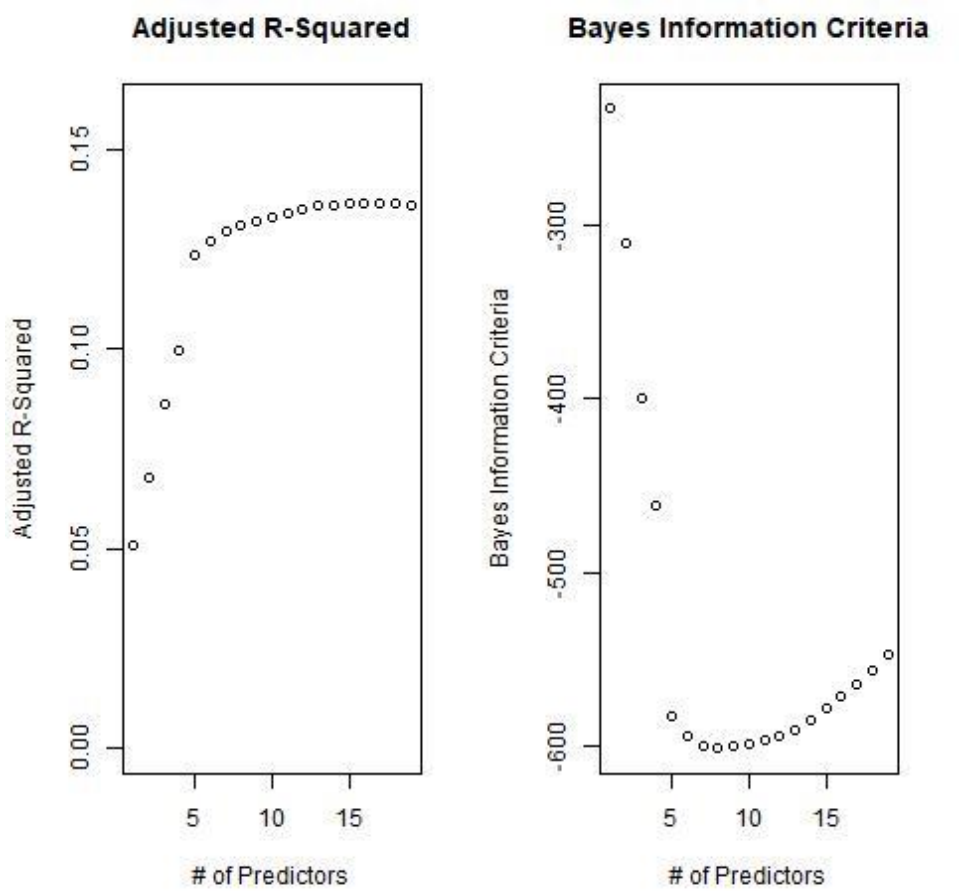

Figure 3.2a. R^2_{Adj} and Bayes Information Criteria For Different Model Sizes

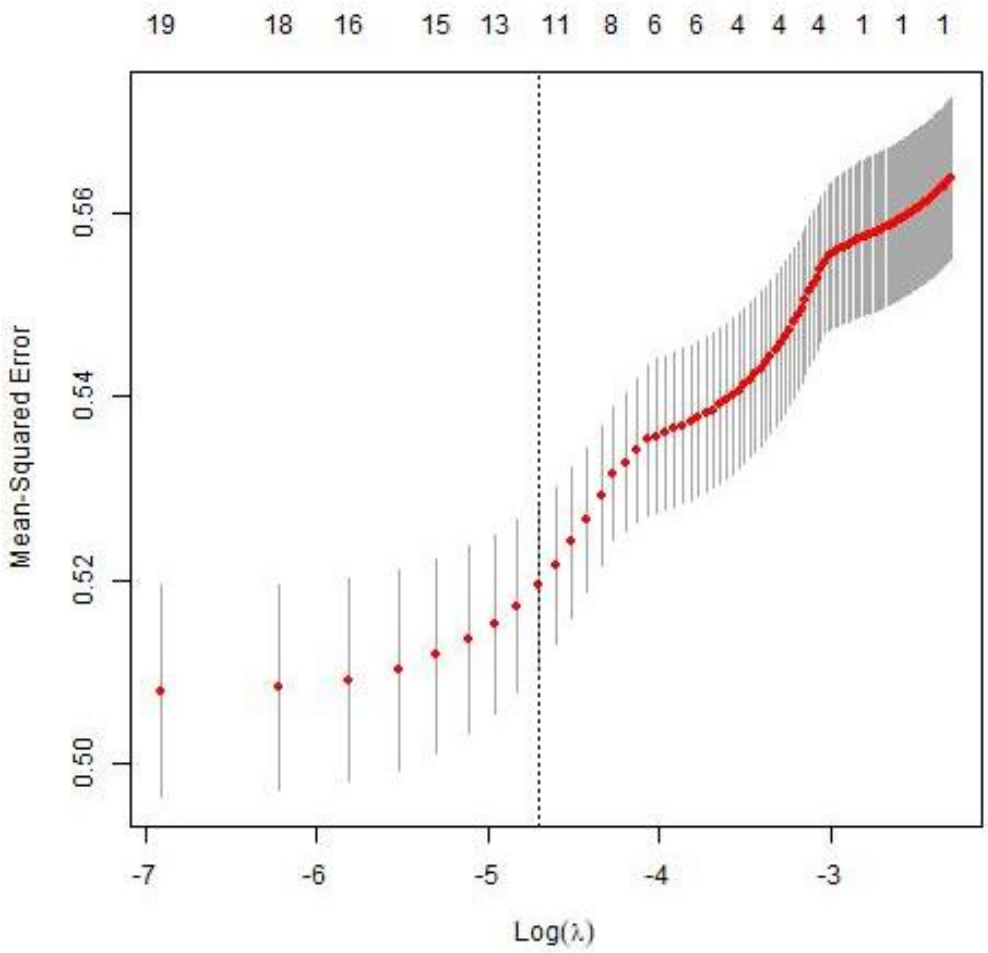

Figure 3.2b. Lasso Regression Results by Lambda Size

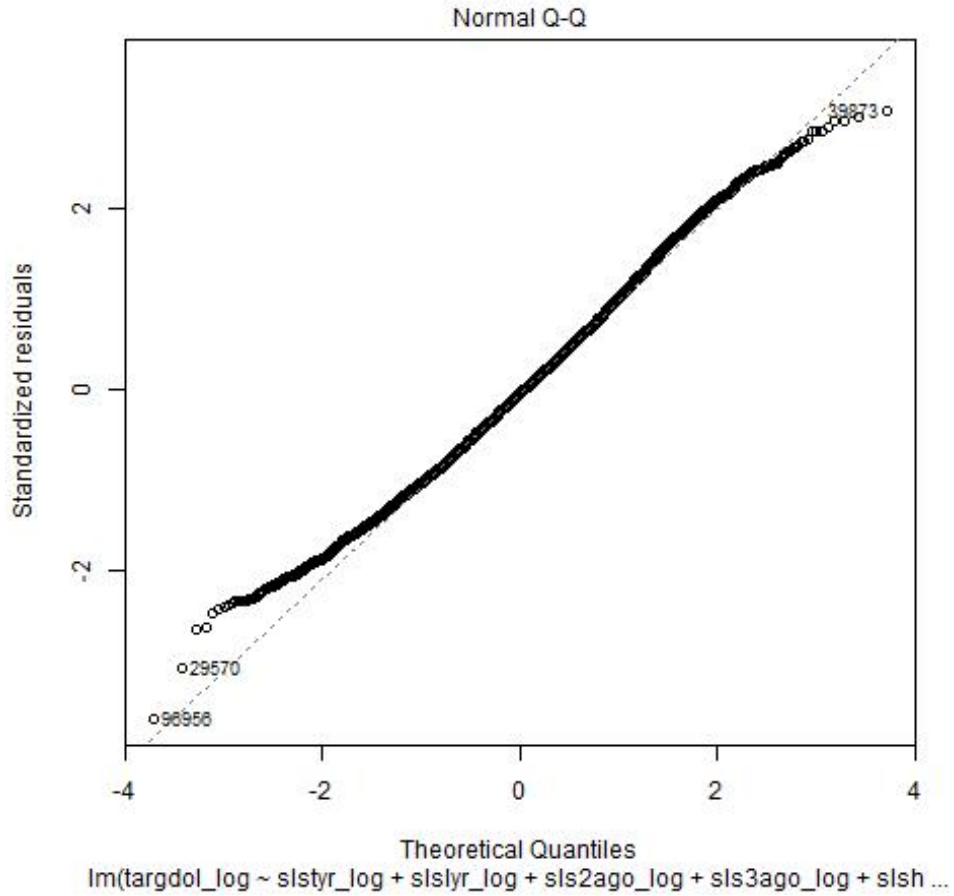

Figure 3.2c. Normal Q-Q Plot for "Simple" Multiple Regression Model

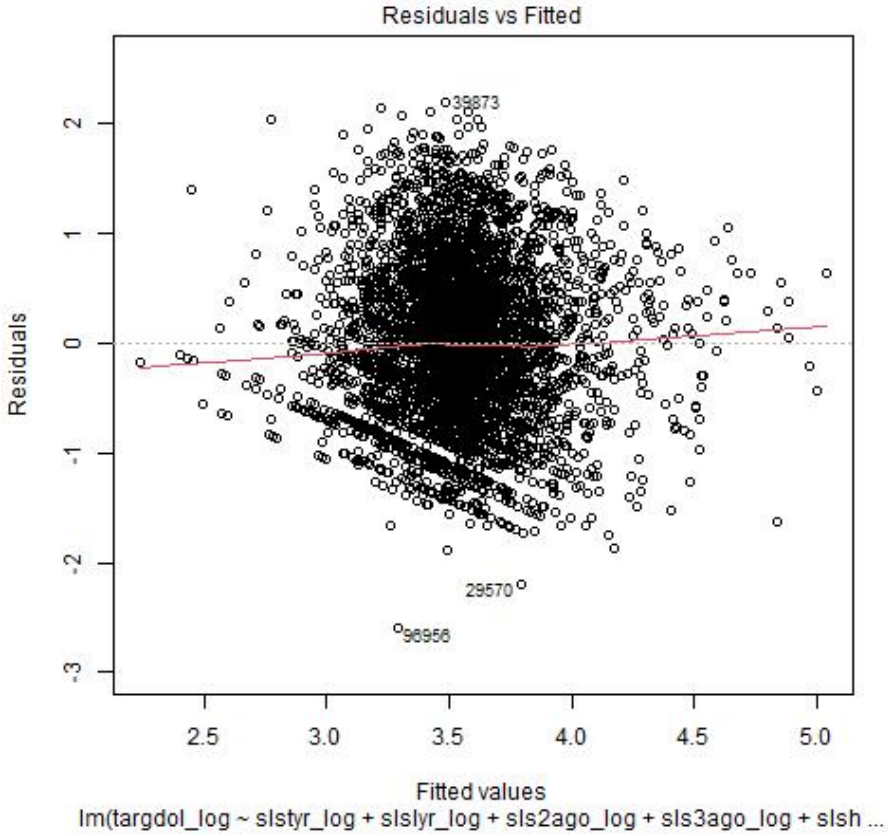

Figure 3.2d. Residuals vs Fitted Plot for "Simple" Multiple Regression Model

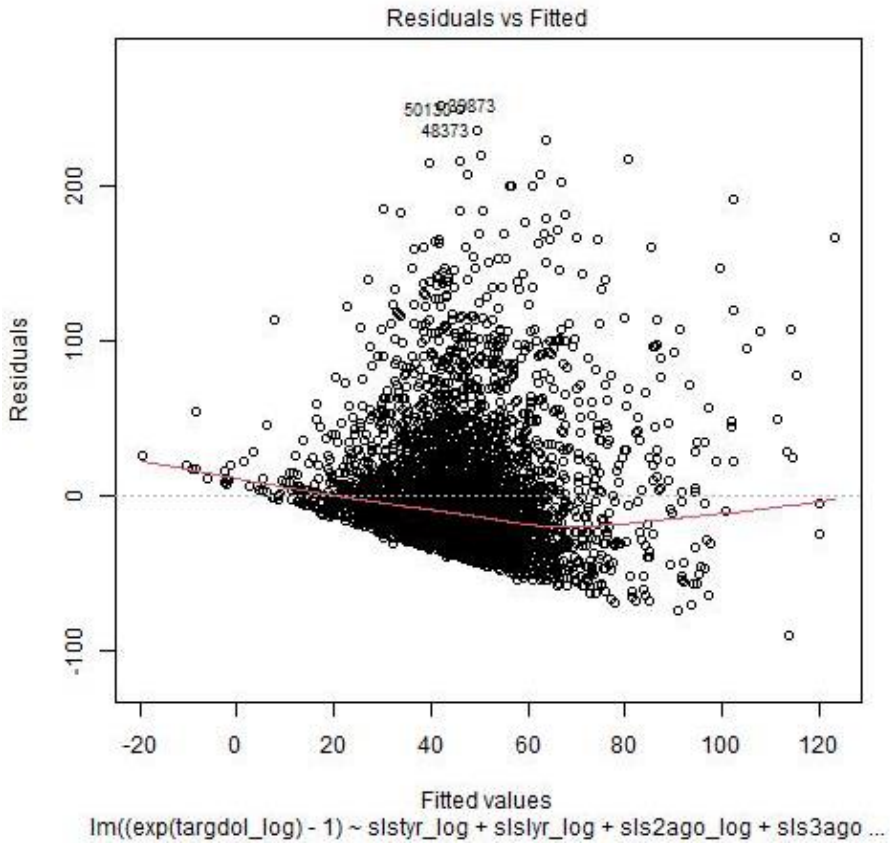

Figure 3.2e. Residuals vs Fitted Plot for "Simple" Multiple Regression Model <u>without</u> Log Transformation

Model Name	Max CCR	Optimal Cutoff Probability
lr_model1	0.9106	0.68
lr_model2	0.9109	0.66
lr_model3	0.9108	0.64
lr_model_simple1	0.9110	0.69
lr_model_simple4	0.9110	0.67

Figure 4a. Summary of logistic regression candidate models performance at hard classification on test dataset

Model Combination	MSE	R^2
mr_model1 & lr_model1	0.9774	0.1094
mr_model1 & lr_model2	0.9768	0.1099
mr_model1 & lr_model3	0.9765	0.1102
mr_model1 & lr_model4	0.9797	0.1072
mr_model1 & lr_model5	0.9799	0.1071
mr_model2 & lr_model1	0.9775	0.1093
mr_model2 & lr_model2	0.9779	0.1089
mr_model2 & lr_model3	0.9767	0.1099
mr_model2 & lr_model4	0.9798	0.1071
mr_model2 & lr_model5	0.9811	0.1060

Table 4b. Model validation output for criteria 1

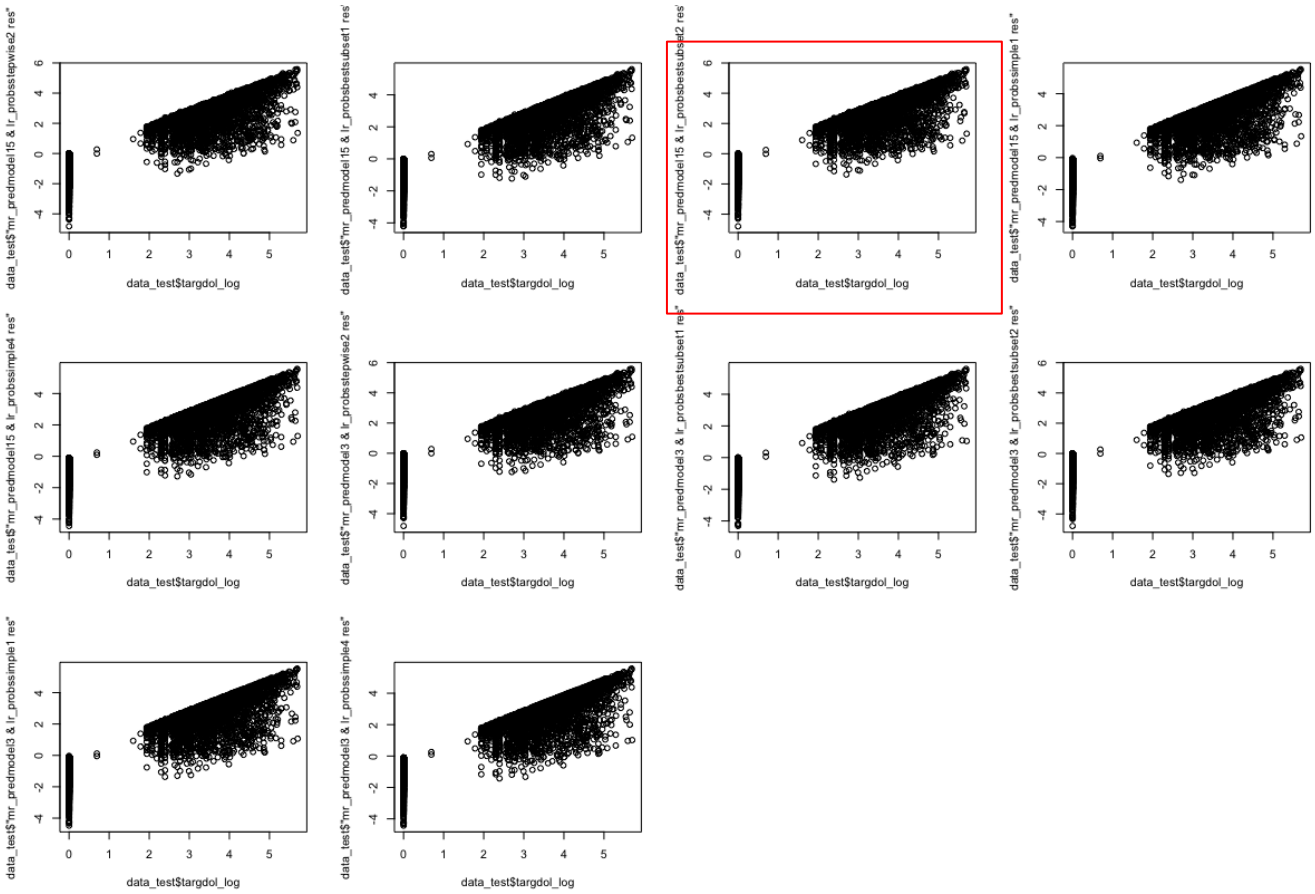

Figure 4a. Residual plots for all ten model combinations, with the "best" model combination according to criteria 1 outlined in red

Model Combination	% of Expected Purchase Amount from the Top 1,000 Customers Predicted
mr_model1 & lr_model1	81.9%
mr_model1 & lr_model2	70.5%
mr_model1 & lr_model3	80.2%
mr_model1 & lr_model4	80.3%
mr_model1 & lr_model5	73.2%
mr_model2 & lr_model1	83.1%
mr_model2 & lr_model2	73.3%
mr_model2 & lr_model3	81.5%
mr_model2 & lr_model4	80.3%
mr_model2 & lr_model5	73.2%

Table 4c. Model validation output for criteria 2

Model Combination	% of the Top 1,000 Customers Identified
mr_model1 & lr_model1	17.8%
mr_model1 & lr_model2	16.9%
mr_model1 & lr_model3	17.4%
mr_model1 & lr_model4	16.8%
mr_model1 & lr_model5	16.8%
mr_model2 & lr_model1	17.5%
mr_model2 & lr_model2	16.8%
mr_model2 & lr_model3	17.4%
mr_model2 & lr_model4	16.7%
mr_model2 & lr_model5	16.8%

Table 4d. Model validation output for criteria 3

	Multiple Regression Model		Logistic Regression Model Criteria		Combined Model Criteria
Model Name	**mr_model2**		**lr_model3**		**mr_model2 & lr_model3**
	R^2_{Adj}= 0.1268		**AUC** = 0.7679		**MSE** = 0.9767 R^2 = 0.1099 **% of Purchase Amount from Top 1,000 Customers Predicted** = 81.5% **% of the Top 1,000 Customers Identified** = 17.4% **Top 10 Customers (by ID) to Target in Future Promotions:** 1. 75384 2. 49704 3. 44474 4. 59256 5. 88930 6. 46216 7. 36848 8. 67332 9. 83882 10. 81807
Coefficient	**Estimate**	**Signif.**	**Estimate**	**Signif.**	
(Intercept)	3.27	***	-2.076	***	
ordtyr	-	-	0.240	***	
ord2ago	-	-	-0.252	***	
ord3ago	-	-	-0.383	***	
falord	-	-	0.221	***	
sprord	-	-	0.350	***	
ordconsistency	-	-	0.641	***	
slsconsistency	-0.67	***	-	-	
slsordhist	-	-	-0.004	***	
slsordtyr	-	-	0.004	***	
slshist_log	0.07	***	-	-	
slstyr_log	0.19	***	-	-	
slslyr_log	0.17	***	-	-	
sls2ago_log	0.17	***	-	-	
sls3ago_log	0.17	***	-	-	
Significance Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 4d. Summary of final chosen model