Promotional Sales Evaluation - A Predictive Model

Saurabh Annadate, Jonathan Lewyckyj, Molly Srour, Yiwei Zhang

MSiA 400: Predictive Analytics I

December 6th, 2018

EXECUTIVE SUMMARY

Our model attempted to predict whether a customer would respond to a promotion given by an online bookstore, in addition to potential purchase amount. The most significant predictors for this process as a whole are purchase frequency and the interaction of amount and frequency for the History category. For the classification process, a multitude of predictors have been identified as significant, including the variables recency, time on file, time-sensitive variables for purchase behavior over different time periods, and categorical variables, both aggregated and individual categories. A smaller subset of variables, including frequency, amount, amount per order, amount per book brought, and the History categorical variables have proved significant in predicting the order amount for the customers that have responded to the promotion. In terms of payoff, the overall model is able to predict with ~25% accuracy as compared to the actual response data from the test dataset.

I. Introduction

The overall goal of this project involved predicting whether a given customer will respond to a specific promotion, and how much they will spend if they do respond. Almost all of the given data relies on time, in some variety, since the best way to predict future purchases is to observe overall purchase history among customers. We focused on the importance of the "time factor" in numerous ways in our analysis, hypothesizing that partitioning data based on date ordered would prove extremely beneficial. Therefore, we concentrated on creating new variables dependent on the time frame of an order, hoping to find a "sweet spot"of time since the last order that would make a customer most likely to buy again. Our section on data cleaning and exploratory data analysis details our methods for creating these time-sensitive variables, along with their interactions, in the hopes of creating a successful predictive model.

Another goal of ours involved exhaustively searching the data for any relevant variables, knowing that we could rely on stepwise selection to reduce the number of predictors to the most significant subset. This included delving into the complex dataset of category variables, which separated orders by category of book ordered. With the assistance of random forest to determine the most significant of these categories, we wished to include these categories in our analysis. This category analysis is detailed at the beginning of our model fitting section.

After narrowing down which category variables are the most significant, we moved on to building the core of our analysis: our predictive models. Our model fitting section discusses our methods to create optimized logistic regression and multiple linear regression models, including

outlier removal, model diagnostics such as Cook's distance and VIF calculation, and stepwise selection. Finally, we discuss the accuracy of our models, both statistically and financially.

II. Data Cleaning and Exploratory Data Analysis

Our goal regarding data cleaning and exploratory data analysis involved identifying, combining and partitioning variables that we hypothesize to be potentially significant predictors. We first endeavoured to pinpoint significant information, if any, contained in the category variables. These variables, containing data regarding amount and frequency of orders separated by book category, proved cumbersome due to their sheer number. Rather than abandoning these variables, we looked into ways to reduce their dimensions by examining the relationships between categories. To begin, we looked at the correlation matrix (Appendix 1) between the category amount variables. No correlations were very high, with the highest correlations between certain category variables being slightly less than 0.50, but in the amount correlation matrix, we were able to identify some categories which seemed to be related, with a correlation greater than 0.3 considered as significant. It was observed that the following categories were intercorrelated: Fiction, Cartoon, Art, History, Travel Guides, Hobby, Contemporary History, and Nature. In our feature engineering steps, we created a variable Mgroup, adding together the amount values for these categories, and the variable Fgroup, adding together the frequency values for these categories. To further test these relationships, we ran a random forest model including all the category variables, which identified one group of category variables that are most related.

Observations regarding individual orders and order pricing, namely that sometimes a single book can cost upwards of $1000, caused us to question the significance of the "amount" variable representing the total amount spent by a single customer summed across all order data.

Therefore, we created a separate variable, qty, representing the total number of books bought by a single customer summed across all order data. Dividing the qty variable by time on file yielded a useful interaction that we later included in both the logistic and linear regression models. Additionally, we created a variable called amount upon quantity that created another metric comparing dollar amount and quantity of books by dividing amount by qty. Other interaction variables created included: dividing amount by total number of orders ("frequency" variable), dividing qty by frequency, and dividing frequency by time on file. Please refer to Appendix 2 for a summary on all engineered features.

For dealing with outliers, we created three criteria based on some of the interaction variables mentioned above. We removed observations where:

1. amount divided by the total number of orders (amtuponorders) exceeded $1000
2. the total number of books per total number of orders (qtyuponorders) exceeded 40
3. amount divided by total number of books ordered (amtuponqty) exceeded $600.

Overall these three criteria only removed 0.0355% (12 observations) of the total data, while improving the fit of our two models on the training data. While we cannot be certain that these data are incorrect, our goal was to create the most accurate model while not mispredicting large values in the test data, and these criteria best struck that balance. The scatterplots for the variables both before and after outlier removal are illustrated in Appendix 3. Later on, for our multiple linear regression model, we used Cook's Distance to identify a few more outliers.

As stated in the introduction, much of the data relies on time to chart consumer behavior, and we endeavoured to capture the effect of the time an order is placed on the response variable. While the variable "recency" charts the number of days since the most recent order a customer

has placed, the order date variable records the date of all orders a customer has placed. We utilized information about both the most recent order and the total orders, per customer, to build a number of novel features in the hopes of illuminating relationships between time and targeted amount. We separated customers by timeframe of order, partitioning customers into groups based on who had ordered in the last 1, 3, 6, and 12 months. Summary statistics for each group, by customer, were calculated, namely quantity of books ordered in this time frame, designated "qty", and total dollar amount, designated "price". Thus, we were able to separate the two variables discussed in the previous paragraph, amount and qty, into a few time-frame partitions. When considering the total number of orders per customer, we were able to calculate the percentage of the total amount purchased, per customer, that occured in these time frames. Therefore, we were able to observe sale trends for each customer over the past year.

When organizing customers based on order date in this manner, we decided to order time on file in a similar way. Partitioning customers based on time on file in the last 1, 3, 6, and 12 months gave us a separation between newer and older customers. During this analysis of time on file, we discovered an interesting observation about a small subset of the customers. In the training set, 87 individuals have a time on file value of zero; similarly, 271 individuals in the test set have a time on file value of zero. These customers are brand new to the online store database; therefore, they have no previous order data and so would be nearly impossible to create accurate predictions for. We were unable to remove these observations due to 52.9% of the new customers in the training set being responders. Since the overall percentage of responders in the dataset is 3.93%, removing the new customers would remove a significant portion of the responders. We therefore decided to manually impute predicted values for these observations.

The imputation process will be detailed in the next section, as it differed for the logistic and the linear regression models.

III. Model Fitting

(i). Random Forest Variable Selection on Book Categories

We ran two random forest models in order to identify potentially important book category variables with respect to classifying responders and predicting log-target amount. The input variables include frequency as well as amount variables for each category (60 variables in total). The output variable is a binary flag capturing whether a customer bought a book (logtargamt equals 0 or not) for the classification model, and logtargamt (for the subset of data that only contains responders) for the regression model. Library "caret" was used to tune both the classification and the regression random forest models.

The random forest analysis illuminated a few important category variables that accurately classified and predicted the customers' purchasing behavior. The CCR for the category-only classification model is 0.9625, slightly better than the null model (classifying everything to the category of logtargamt = 0, with a CCR of 0.9606), indicating that adding some categories might provide valuable information to our logistic and linear regression models. To identify the important category variables, we ran the variable importance function within the caret package and selected the top 6 categories (as reported above in section II) that were reported as highly important (with an importance score larger than 30) in both the classification and the regression model. We included these variables in our initial logistic and linear regression models.

We also created product interaction terms between the frequency and amount for each aforementioned important category in order to capture any interaction effect. These interaction

terms were added to the preliminary logistic and linear model fitting as well. A graph showing the final important categories selected by random forest is shown in the Appendix 4.

(ii). Classification Model: Binary Logistic Regression

To create a logistic regression model, the response variable logtargamt received the necessary transformation into a binary variable. This binary variable has a value of 1 if logtargamt > 0 and a value of 0 if logtargamt = 0. Our preliminary model included all the variables we created, in addition to the recency, frequency, amount, and time on file variables that were included in the original dataset. These included the time-related variables we had created - total quantity, dollar amount, binary flag if any orders have been placed in given time frame, and time on file in given time frame - for 1 month, 3 months, 6 months, and 12 months, in addition to our qty variable and the numerous interaction variables we created. We also included the Mgroup and Fgroup aggregated category variables along with the most significant category variables as found through our random forest model.

As stated in the previous section, the customers with a 0 value for time on file (i.e. brand new customers) did not have data related to previous orders, rendering it near impossible to predict their target amount. This necessitated a manual imputation method of predicted values. For this logistic regression, we decided that the response rate for the new customers in the training set (0.5287) adequately replaced the probability of logtargamt > 0 for these new customers. Since this manual imputation replaces the predictive model for new customers, we removed all new customers from both the training and the test set before performing the logistic regression. The new customers from the test dataset were placed back into the dataset later when validating our predictive model with the imputed value 0.5287 serving as the predicted

probability; more details regarding our steps with respect to these customers will be discussed in the next section.

Since the response rate for the training data was very low (3.93%), we utilized bootstrapping to oversample the responders (customers with logtargamt > 0) to ensure the validity of our model in predicting customers as responders. We assigned weights for each responder and non-responder in the training data and sampled with replacement to get our final training set. The percentage of responders in the final dataset is 20% and non-responders is 80%. The dimensions (number of observations and variables) of the final training data is the same as the original training data.

We ran backward stepwise regression with 42 variables minimizing AIC to arrive at an optimal model. The model summary has been presented in Appendix 5 Section (a). We checked the variance inflation factors for the final predictors (Appendix 5 Section (b)) and removed the variables that had VIFs significantly larger than 10 and low predicting power in the model. Note that we did not remove the variables Mgroup and Fgroup since they are highly significant at the 5% level in classifying responders, even though they had VIFs above 30. Please also note that we removed the variable amount even though it was significant in the original model. After removing variables with high VIF values, amount became non-significant at 5% level, suggesting multicollinearity with some of the removed variables, so we removed the amount variable as well. We refit the model with the remaining predictors after removing these variables and arrived at our final model. The final model is shown in Appendix 5 Section (c) and the final VIFs are shown in the Appendix 5 Section (d).

In our final logistic regression model, recency, frequency and tof are all highly significant (at 1% level) in classifying response. Variables capturing customer purchasing behavior, including whether a customer purchased in the recent 3, 6, 12 months, whether a customer just started in the recent 6, and 12 months are also all highly significant. Moreover, the interaction term between customers' purchasing amount and quantity and the interaction between quantity and time of file are also very important in classifying response. In terms of book category variables we added to the model, Mgroup and Fgroup proved to be statistically significant at the 1% level, in addition to the amount variable for History and Cartoon, the interaction between the amount and frequency variables for History, the interaction between the amount and frequency variables for Contemporary History and the interaction between the amount and frequency variables for Health.

The final number of variables in our logistic model is 23 out of the original 42. The residual deviance is 7360.4 on 8197 degrees of freedom and the final AIC is 7408.4. We also tried fitting the model without the bootstrap oversampling on responders, which generated a much lower AIC of around two thousand. However, since the model without bootstrap classified most customers into non-responders and has low predicting power on responders, we decided to use the model with bootstrap as our final model.

We graphed the Receiver Operating Characteristic curve for our final logistic model using the training as well as the test data, as is shown in Appendix 5 Section (e). The curve deviates from the 45 degree line by a significant amount, suggesting that our model has good predicting power. The area under the ROC curve is 0.70 for training data and 0.66 for the test set,

which shows that our model is decently accurate in classifying responders and non-responders, but also did not blindly classify everyone as non-responders.

(iii). Multiple Linear Regression Model

In the logistic regression model, we did not worry about the varying purchase amount; we merely wished to predict whether a customer would respond to the promotion and purchase anything at all. However, building the multiple linear regression model necessitated further exploration of the spread of the target amount variable. We first explored the subset of our training dataset where logtargamt was greater than zero and time on file is greater than zero (customers who responded to the survey and also have previous order data). This left us with 280 observations. We began with a number of scatterplot matrices which plotted the relationship between logtargamt and the variables provided in the book.csv dataset (Appendix 6 Section (a)). This gave us a sense of whether we should include higher order terms of our features in our model.

Before running our initial model, we made an executive decision regarding the new customers (those with time on file = 0). We decided to take a similar route to how we treated them in the logistic regression. Since a predictive model would be highly inaccurate for customers with no previous order data, we decided a manual imputation would be the best route for handling these customers. We removed the new customers from the training set and the test set before running the model. Later, we considered the average of the logtargamt of the new customers in the training set to be the predicted logtargamt of all customers in both the training set and the test set. The new customers were placed back into the dataset later when validating

our predictive model with the imputed value; more details regarding our steps with respect to these customers will be discussed in the next section.

Our initial linear regression model included all the possible features. While this kitchen-sink model by itself would not be expected to be illuminating, we followed this up by running backwards stepwise regression to eliminate variables that did not significantly improve the fit of our model. The stepwise regression (summarized in Appendix 6 Section (b)) showed the following variables as statistically significant at the 5% level: frequency, amount, qty6mo (total orders in the last 6 months), amtuponorders (amount divided by the total number of orders), qtyuponorders (total number of books per total number of orders), and the interaction between the amount and the frequency variable for the History category. A couple other variables were not significant at the 5% level but improved the adjusted R-squared enough to be included. This model had an R-squared of 0.3709 and an adjusted R-squared of 0.3402.

To evaluate the performance of our model, we ran the necessary checks for the main underlying assumptions of any good multiple regression model: normality, homoscedasticity, no outliers/influential observations, and no evidence of multicollinearity. The Q-Q plot of our model confirmed the assumption for normality, as the quantiles for the standardized residuals and theoretical quantiles largely followed a straight-line pattern. There were a few outliers observed in the Q-Q plot. The plot for residuals vs. fitted values confirmed the assumption for homoscedasticity, as no relationship was shown between the residuals and the fitted values. For outlier detection, the formula 4 / (n-(p+1)) yielded a threshold of 0.015. This threshold identified too many observations as outliers, so we decided on a threshold of 0.1 for Cook's Distance, which removed 4 outliers, giving us a training sample of 276 observations. To check for

evidence of multicollinearity among predictors, we looked at the Variance Inflation Factors for the variables in this model. The variables price3mo, amtupontof, and qtyupontof exceeded the VIF threshold of 15, as did the amount and frequency variables for Contemporary History. All model diagnostics are illustrated in Appendix 6 Section (c).

Next, we re-ran the model for multiple regression after removing outliers and not including the variables with a VIF exceeding 15. The final multiple linear regression model found the following variables as statistically significant at the 5% level: frequency, amount, amtuponorders (amount divided by the total number of orders), amtuponqty (amount divided by total number of books ordered), the amount variable for the History category (p value = 0.05089), and the interaction between the amount and frequency variables for the History category.

Our final multiple regression equation:

$$\begin{aligned} Logtargamt = {} & 3.193 - 0.04075 * frequency + 0.001263 * amount + 0.009667 * amtuponorders \\ & - 0.02372 * amtuponqty + 0.0019 * Mhistory19 \\ & - 7.702 * 10^{-5} * Mhistory19 : Fhistory19 \end{aligned}$$

This final model has an R-squared of 0.385, and an adjusted R-squared of 0.3713. The model diagnostics for the final model is illustrated in Appendix 6 Section (d).

One final check we performed to test our model involved exploring the spread of the test data. In order to optimize our model, we had previously removed outliers in the training set where amount upon order was larger than 1000, quantity upon order was larger than 40, and amount upon quantity was larger than 60. We observed similar outliers in the test set, with values well above these previously defined thresholds for the aforementioned predictors.

Wishing to preserve the integrity of the test data set, we decided to keep these observations but to impute a predicted value for them. We made this decision through observation of the wildy large predicted values created when predicting the target amount for these customers using our multiple regression model (on the order of 10^{40} or greater). Therefore, we chose as our imputed predicted value the maximum observed target amount for the training data. If we had left the massive predicted values in the model, our overall standard error would have been erroneously massive. Additionally, since we had removed the corresponding outliers in the training set, the model is unable to anticipate such extreme cases, so our imputation was absolutely necessary here, since we had endeavoured to include all observations in the test set. Furthermore, we justify our removal of the corresponding outliers in the training set by the increase in R^2 of our model following their removal. In total, only 2 such observations (0.007873% of the testing data) were found with the aforementioned outlier characteristics. Although this imputation was relatively minor with relation to the size of the test data set, its effect was appropriately large due to the extremeness of the outliers. The characteristics for these two outlier observations are illustrated in Appendix 7 Section (a).

IV. Model Validation

(i). Statistical Criterion

The predicted target amount for customers with previous order data (time on file > 0) in the test set was calculated directly from our predictive models: E(log target amount) = the product of the probability of logtargamt > 0 given by the logistic regression (after adjusting for oversampling) and the predicted logtargamt value given by the linear regression. For the customers in the test set with time on file = 0, we artificially created values that, to the best of

our ability, mimic the predictions given by the logistic and linear models if previous order data had been logged. As stated in the previous section, the probability of logtargamt > 0 for these customers was imputed as 0.5287 (the overall response rate in the training set). Similarly, the predicted logtargamt value that should have been found in the linear regression was imputed as the mean value of logtaramt among new users from the training dataset. E(log target amount) was calculated for these new customers in the test set by multiplying these two values together to get an artificially predicted target amount.

This product, the predicted log target amount, was then exponentiated and subtracted by 1 to obtain the predicted target amount. The sum squared errors of prediction (SSEP) for this final model is 2,863,585 for the test set. There are more than 20 thousand rows in the test set, causing the sum of squared error to be very large; therefore, we also calculated the root mean square error (RMSE) of our predicted value by taking the square root of the mean of the squared errors of prediction. The RMSE of our final model is $10.57, suggesting satisfactory model performance.

(ii). Financial Criterion

Using our predictive model, we identified the customers with the top 500 predicted target amounts. The sum of these predicted purchases totals $2261.83. Their total actual purchases totaled $6449.529, which represents the payoff of our predictive model. After noting the accuracy of our prediction compared to the payoff (35.07% of the actual payoff), we focused on the true financial criterion - the payoff percentage. We pinpointed the 500 customers who actually had the highest purchased amount during the promotional period. Their purchases totaled $27,035.11. Therefore, our payoff percentage is 6449.529/27035.11 = 23.85%. This

result indicates that even though our logistic and linear model both have high accuracy and R-square (correspondingly) in predicting the response variable, there is still space for improvement in terms of actual performance.

The payoff percentage of 23.85% gives one measure of our model's accuracy at predicting the purchase amount of the top 500 prospects; however, we decided to further optimize the financial success of our model by finding the number of prospects which maximizes the short term profit. With the assumption that the profit margin is 25% of the purchase amount, we maximized 0.25*(sales revenue from the top x predicted prospects) - 1*x to yield 852 as the optimum number of top prospects to target, giving a short term profit of $1319.74. A logical next step involves calculating the payoff percentage for the top 852 customers and compare it to the percentage for the top 500 customers. This percentage was slightly higher: 25.96% for the top 852 vs. 23.85% for the top 500. Since the difference is not much, we can say that per the results, majority of the maximum probable profit can be achieved by targeting the top 500 customers. Future exploration of financial accuracy could illuminate better ways to optimize fiscal payoff; however, this is not in the scope of our project.

V. Conclusions

In our final models, purchase frequency and the interaction of amount and frequency for the History category were significant in both classifying responders and predicting order amounts for responders. A multitude of predictors were significant in classifying responders or non-responders, including the original variables recency, frequency, and time on file, time-sensitive variables for purchase behavior over different time periods, and certain categorical variables, in addition to interactions among variables. For predicting order amounts for

responders, a smaller subset of variables, including frequency, amount, amount per order, amount per book brought, and the History categorical variables proved significant. The root mean squared error of the predicted target amount was $10.57, indicating statistically strong final models. The total purchases of our top 500 predicted amounts fell a bit short of the total orders to the actual top 500 customers, at a 23.85% payoff percentage. While this shows that there is room for improvement in our models, it is also a testament to regression to the mean and the difficulty in predicting outliers.

We felt that more information about this particular promotion could have helped our prediction accuracy. If we had known whether this promotion was general or specific to certain categories, our categorical variable analysis would have become more efficient. Also, promotion implementation was unclear: if the promotion was for a certain percent off one order, this would have incentivized different ordering behavior than a promotion that would apply to all orders. Additionally, information regarding promotional audience (e.g. whether it went out to all customers, a random sample of customers, or targeted groups of customers) would have also been beneficial. Another possible metric of success for a promotion could have been how many customers engaged with the promotion (e.g., clicked through an e-mail vs. deleted it), but didn't ultimately purchase. Further potentially beneficial metrics, such as frequency and amount of time spent on the book store's website, as well as what they typically do while browsing the website, could have also improved our model. Finally, customer service information may have also been beneficial. A customer who frequently engages customer service representatives could be more likely to purchase. However, negative recent experiences could decrease the probability of future purchases.

References

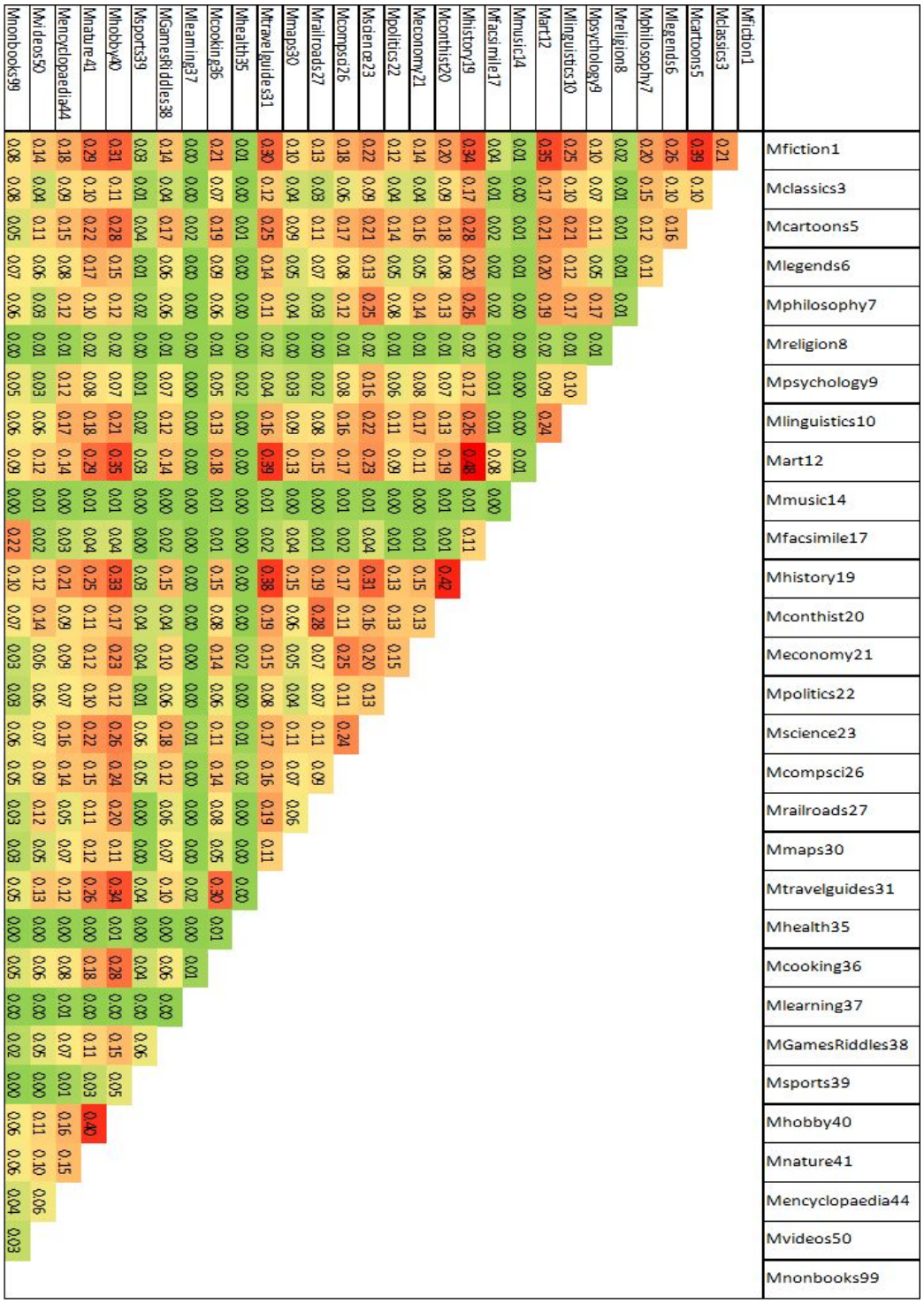
1. Tamhane, Ajit C. Predictive Analytics: Parametric Models for Regression and Classification Using R. Wiley. Draft.

Appendix

Appendix 1 : Correlation Chart
Appendix 2 : Engineered features
Appendix 3 : Outlier removal
Appendix 4 : Random Forest
Appendix 5 : Logistic Regression
Appendix 6 : Linear Regression
Appendix 7 : Testing on Test data

Appendix 1 : Correlation Chart

- Figure 1 : Amount Correlations



Appendix 2 : Engineered Features

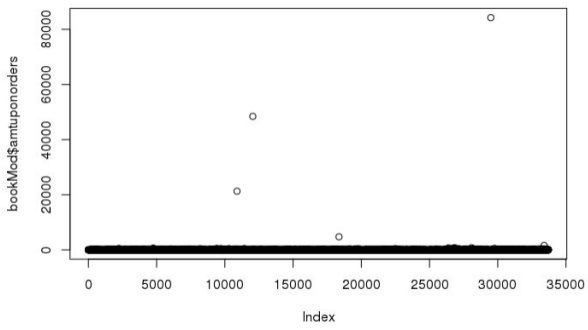
Following is a list of all features that were engineered for the analysis:

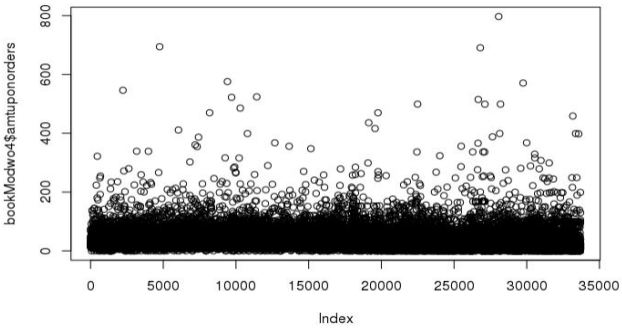
Feature	**Description**
Mgroup	sum of amounts for the buckets : fiction, cartoon, art, history, travel guides, hobby, contemporary history, nature
Fgroup	sum of frequencies for the buckets : fiction, cartoon, art, history, travel guides, hobby, contemporary history, nature
qty	Total quantity of books brought by the customer over the entire lifetime
qty1mo	Quantity of books brought in the most recent 1 month
price1mo	Total amount spent by the customer in the most recent 1 month
qty3mo	Quantity of books brought in the most recent 3 month
price3mo	Total amount spent by the customer in the most recent 3 month
qty6mo	Quantity of books brought in the most recent 6 month
price6mo	Total amount spent by the customer in the most recent 6 month
qty12mo	Quantity of books brought in the most recent 12 month
price12mo	Total amount spent by the customer in the most recent 12 month
onetothreeflagmo	Binary flag determined by whether the most recent purchase by the customer has been between the most recent one and three months
orderinrecent1mo	Binary flag determined by whether the most recent purchase by the customer has been within the most recent 1 month
orderinrecent3mo	Binary flag determined by whether the most recent purchase by the customer has been within the most recent 3 month

orderinrecent6mo	Binary flag determined by whether the most recent purchase by the customer has been within the most recent 6 month
orderinrecent12mo	Binary flag determined by whether the most recent purchase by the customer has been within the most recent 12 month
startedinrecent1mo	Binary flag determined by whether the customer started using the website in the past 1 month
startedinrecent3mo	Binary flag determined by whether the customer started using the website in the past 3 month
startedinrecent6mo	Binary flag determined by whether the customer started using the website in the past 6 month
startedinrecent12mo	Binary flag determined by whether the customer started using the website in the past 12 month
PerSales3mo	Percentage of total amount spent by a customer in the most recent 3 months
PerSales6mo	Percentage of total amount spent by a customer in the most recent 6 months
amtuponorders	Total amount spent divided by total number of orders
qtyuponorders	Total quantity of books ordered divided by total number of orders
amtuponqty	Total amount spent divided by total books ordered
ordersupontof	Total number of orders divided by the time on file
amtupontof	Total amount spent divided by the time on file
qtyupontof	Total quantity of books ordered divided by the time on file

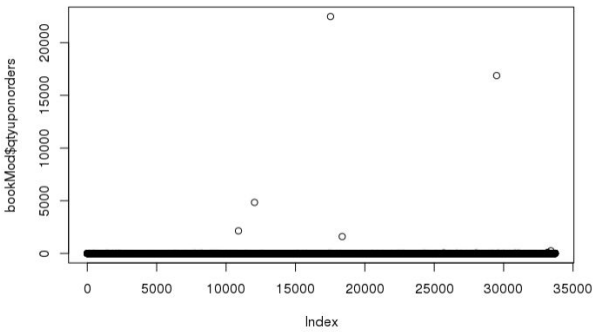
Appendix 3 : Outlier Removal

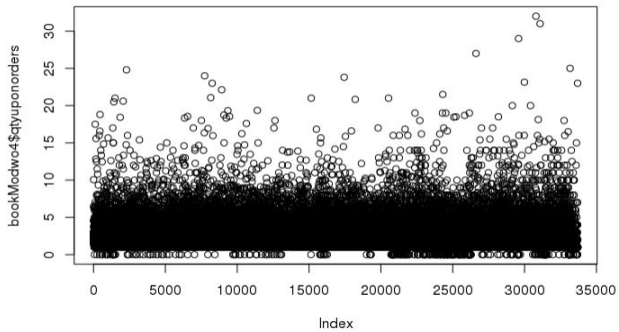
- amount divided by the total number of orders (before and after outlier removal)



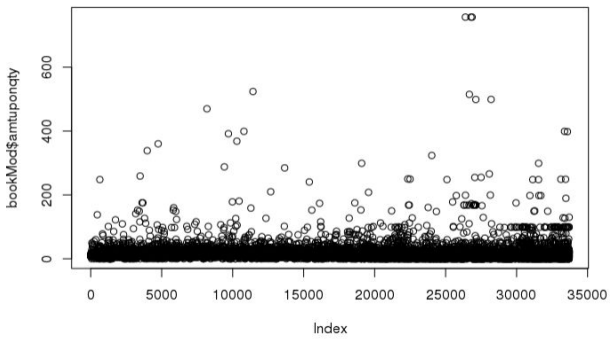


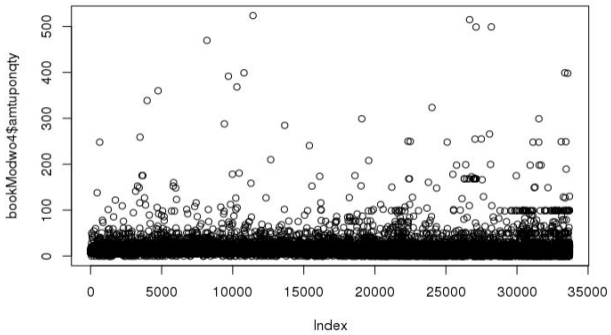
- the total number of books per total number of orders (before and after outlier removal)





- amount divided by total number of books ordered before and after outlier removal)





Appendix 4 : Random Forest

- Figure 1 : Significant factors from the Random Forest model

```
rf variable importance

  only 20 most important variables shown (out of 22)

                         Overall
frequency                100.000
amount                    99.995
tof                       91.133
recency                   49.963
Mhistory19                47.597
Mcartoons5                44.583
Mconthist20               34.795
Fhistory19:Mhistory19     34.716
Fhealth35:Mhealth35       34.257
Fconthist20               34.108
Mcartoons5:Fcartoons5     33.068
Fconthist20:Mconthist20   30.175
Mhealth35                 29.938
Mmusic14                  29.697
Fmusic14:Mmusic14         25.589
Fcartoons5                21.510
Fmusic14                  20.421
Fhealth35                 19.162
Fhistory19                13.765
Freligion8                 7.332
```

Appendix 5 : Logistic Regression

- Section (a) : Model Summary post Stepwise regression

```
Call:
glm(formula = as.factor(successFlag) ~ recency + frequency +
    amount + tof + Mgroup + Fgroup + price3mo + qty6mo + price6mo +
    qty1mo + price12mo + qty + onetothreeflagmo + orderinrecent3mo +
    orderinrecent6mo + orderinrecent12mo + startedinrecent6mo +
    startedinrecent12mo + PerSales6mo + amtuponorders + qtyuponorders +
    amtuponqty + ordersupontof + amtupontof + Mhistory19 + Mcartoons5 +
    Mconthist20 + Fconthist20 + Mhistory19:Fhistory19 + Fhealth35:Mhealth35 +
    Mconthist20:Fconthist20, family = binomial, data = bookModTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4584  -0.6578  -0.5115  -0.3496   2.6212

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.178e+00  1.455e-01  -8.098 5.58e-16 ***
recency                 -5.743e-04  1.323e-04  -4.341 1.42e-05 ***
frequency                4.750e-02  1.021e-02   4.652 3.29e-06 ***
amount                   1.491e-03  5.241e-04   2.846 0.004432 **
tof                     -3.489e-04  6.363e-05  -5.482 4.20e-08 ***
Mgroup                  -4.488e-03  1.100e-03  -4.079 4.52e-05 ***
Fgroup                   2.858e-02  1.180e-02   2.422 0.015418 *
price3mo                 3.420e-03  1.828e-03   1.871 0.061368 .
qty6mo                   3.320e-02  1.242e-02   2.673 0.007525 **
price6mo                -4.819e-03  1.810e-03  -2.662 0.007759 **
qty1mo                  -8.886e-02  3.275e-02  -2.713 0.006663 **
price12mo                4.186e-03  7.454e-04   5.616 1.95e-08 ***
qty                     -1.927e-02  6.040e-03  -3.190 0.001425 **
onetothreeflagmo        -2.400e-01  1.552e-01  -1.546 0.122094
orderinrecent3mo         5.410e-01  1.648e-01   3.284 0.001024 **
orderinrecent6mo         3.699e-01  1.262e-01   2.932 0.003371 **
orderinrecent12mo       -2.896e-01  1.219e-01  -2.376 0.017484 *
startedinrecent6mo       6.685e-01  1.907e-01   3.506 0.000454 ***
startedinrecent12mo     -8.083e-01  1.159e-01  -6.972 3.13e-12 ***
PerSales6mo             -3.891e-03  2.339e-03  -1.663 0.096224 .
amtuponorders           -1.076e-02  3.221e-03  -3.341 0.000836 ***
qtyuponorders            1.132e-01  3.360e-02   3.368 0.000756 ***
amtuponqty               1.104e-02  3.201e-03   3.449 0.000562 ***
ordersupontof            5.159e+00  2.196e+00   2.349 0.018808 *
amtupontof               1.618e-01  8.791e-02   1.840 0.065748 .
Mhistory19               3.476e-03  1.144e-03   3.038 0.002384 **
Mcartoons5              -6.345e-03  4.303e-03  -1.475 0.140312
Mconthist20             -1.607e-03  2.492e-03  -0.645 0.518854
Fconthist20              2.858e-02  2.879e-02   0.993 0.320854
Mhistory19:Fhistory19   -9.689e-05  3.514e-05  -2.757 0.005825 **
Fhealth35:Mhealth35     -1.342e-04  4.101e-05  -3.272 0.001069 **
Mconthist20:Fconthist20  1.887e-05  9.218e-06   2.047 0.040697 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8137.5  on 8220  degrees of freedom
Residual deviance: 7334.5  on 8189  degrees of freedom
AIC: 7398.5

Number of Fisher Scoring iterations: 7
```

- Section (b) : VIFs for the predictor variables in the Model arrived post Stepwise regression

```
                recency               frequency                  amount
               3.097481                6.664473               34.935608
                     tof                  Mgroup                  Fgroup
               3.137952               47.442827               36.562188
               price3mo                qty6mo                 price6mo
               6.221998                5.393104               12.628912
                 qty1mo                price12mo        onetothreeflagmo
               3.984717                5.225272                5.257890
                    qty       orderinrecent3mo       orderinrecent6mo
              36.710033                7.586520                4.820757
      orderinrecent12mo       PerSales6mo        startedinrecent6mo
               3.643355                3.204680                2.742570
     startedinrecent12mo       amtuponorders          qtyuponorders
               5.866612                7.284596                4.365509
          amtuponqty          ordersupontof              amtupontof
               3.967545                1.867296                3.190706
            Mhistory19              Mcartoons5              Mconthist20
               7.017989                1.283258               40.654232
          Fconthist20  Mhistory19Fhistory19  Mhealth35Fhealth35
              39.423506                2.507631                1.680020
Fconthist20Mconthist20
               3.381829
```

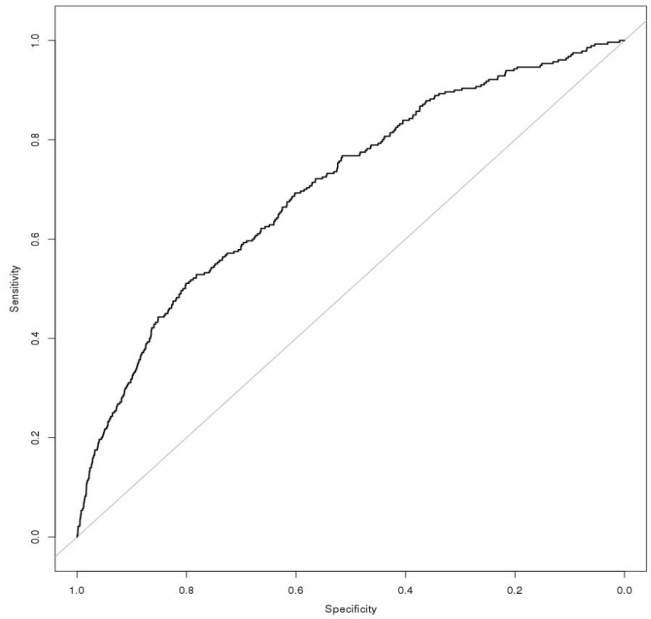
- Section (c) : Final Model Summary

```
Call:
glm(formula = as.factor(successFlag) ~ recency + frequency +
    tof + Mgroup + Fgroup + qty6mo + price6mo + qty1mo + price12mo +
    qty + orderinrecent3mo + orderinrecent6mo + orderinrecent12mo +
    startedinrecent6mo + startedinrecent12mo + PerSales6mo +
    amtuponqty + qtyupontof + Mhistory19 + Mcartoons5 + Mhistory19:Fhistory19 +
    Fhealth35:Mhealth35 + Mconthist20:Fconthist20, family = binomial,
    data = bookModTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3363  -0.6590  -0.5132  -0.3534   2.5945

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.007e+00  1.267e-01  -7.943 1.97e-15 ***
recency                 -6.203e-04  1.324e-04  -4.685 2.80e-06 ***
frequency                5.117e-02  8.432e-03   6.069 1.29e-09 ***
tof                     -3.702e-04  6.241e-05  -5.932 2.99e-09 ***
Mgroup                  -3.452e-03  7.894e-04  -4.373 1.22e-05 ***
Fgroup                   2.329e-02  8.676e-03   2.684 0.007273 **
qty6mo                   2.749e-02  1.085e-02   2.534 0.011262 *
price6mo                -2.788e-03  1.315e-03  -2.120 0.034019 *
qty1mo                  -5.439e-02  2.110e-02  -2.578 0.009941 **
price12mo                4.033e-03  6.794e-04   5.936 2.93e-09 ***
qty                     -6.185e-03  3.480e-03  -1.777 0.075491 .
orderinrecent3mo         4.612e-01  9.520e-02   4.844 1.27e-06 ***
orderinrecent6mo         3.272e-01  1.230e-01   2.660 0.007804 **
orderinrecent12mo       -3.030e-01  1.215e-01  -2.495 0.012610 *
startedinrecent6mo       7.162e-01  1.867e-01   3.836 0.000125 ***
startedinrecent12mo     -8.278e-01  1.157e-01  -7.154 8.42e-13 ***
PerSales6mo             -4.086e-03  2.313e-03  -1.767 0.077284 .
amtuponqty               2.384e-03  1.279e-03   1.863 0.062415 .
qtyupontof               2.833e+00  6.420e-01   4.412 1.02e-05 ***
Mhistory19               2.999e-03  1.044e-03   2.872 0.004075 **
Mcartoons5              -7.899e-03  4.267e-03  -1.851 0.064162 .
Mhistory19:Fhistory19   -7.786e-05  3.275e-05  -2.378 0.017429 *
Fhealth35:Mhealth35     -1.403e-04  3.976e-05  -3.529 0.000417 ***
Mconthist20:Fconthist20  2.154e-05  7.138e-06   3.018 0.002542 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8137.5  on 8220  degrees of freedom
Residual deviance: 7360.4  on 8197  degrees of freedom
AIC: 7408.4

Number of Fisher Scoring iterations: 7
```

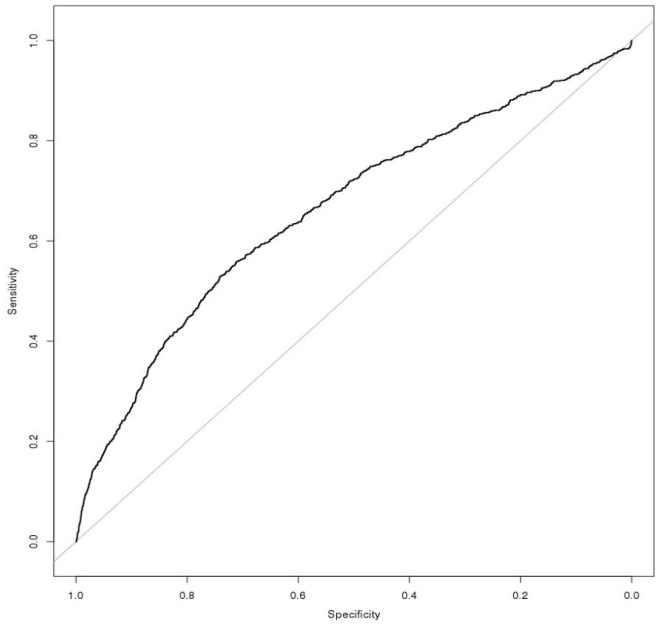
- Section (d) : VIFs for the predictor variables in the final Model

```
          recency         frequency            amount               tof            Mgroup
         3.086487          4.670598         28.556806          3.101346         39.978905
           Fgroup            qty6mo           price6mo            qty1mo          price12mo
        29.521778          5.065467          7.399457          1.786867          4.921295
              qty     orderinrecent3mo    orderinrecent6mo   orderinrecent12mo  startedinrecent6mo
        29.440039          2.482176          4.545415          3.624101          3.131594
startedinrecent12mo       PerSales6mo       amtuponqty        qtyupontof          Mhistory19
         2.733835          5.802671          1.347720          1.722391          6.305891
       Mcartoons5 Mhistory19Fhistory19 Mhealth35Fhealth35 Fconthist20Mconthist20
         1.263094          2.277666          1.611738          1.699281
```

- Section (e) : ROC curves for training and test data



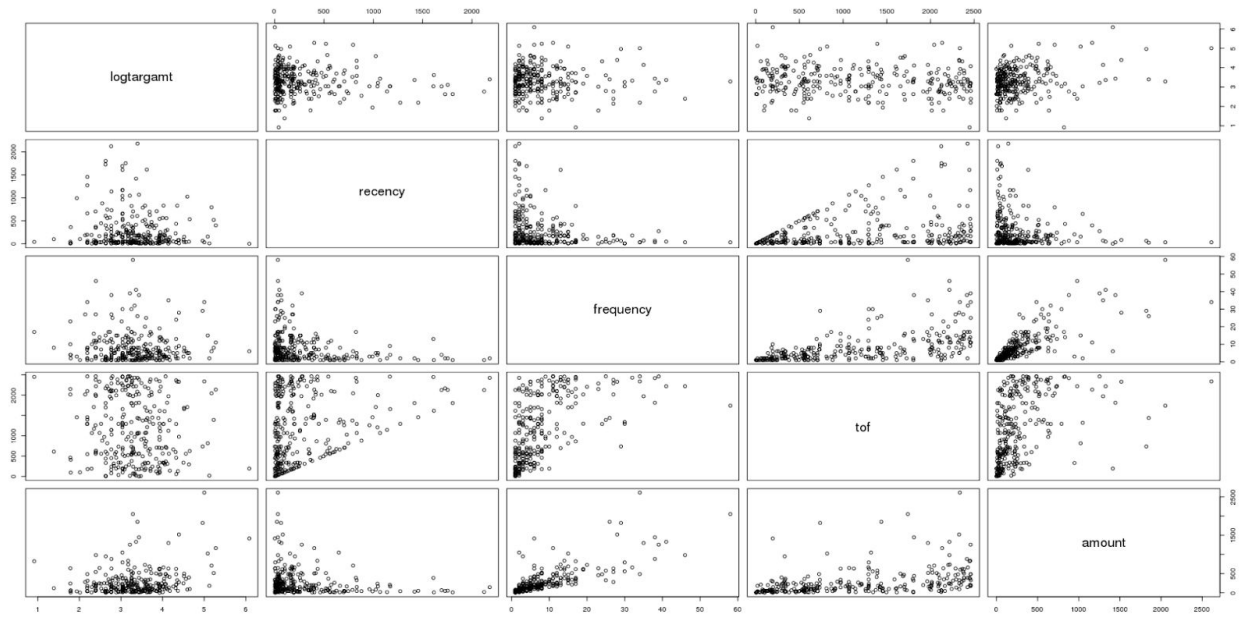
Training data



Test Data

Appendix 6 : Linear Regression

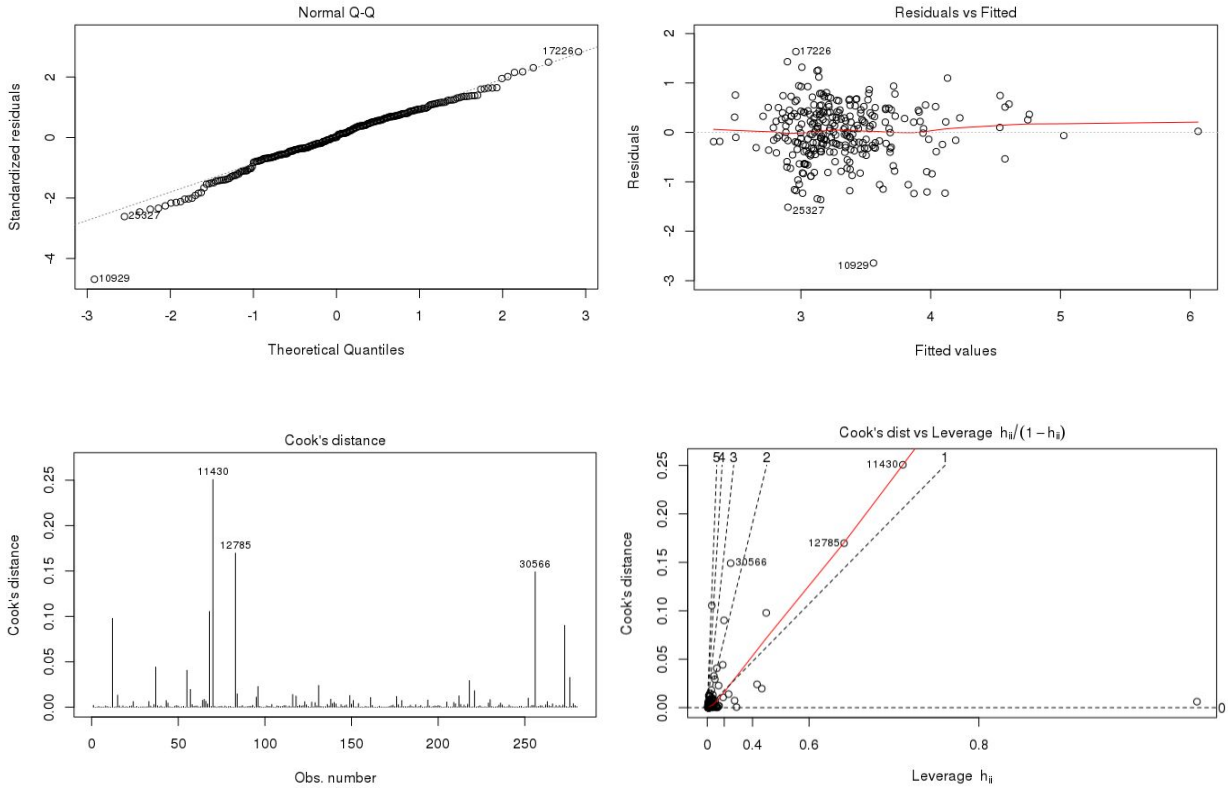
- Section (a) : Scatterplots of variables



- Section (b) : Initial Linear Regression Model Summary

```
Call:
lm(formula = logtargamt ~ recency + frequency + amount + qty3mo +
    price3mo + qty6mo + price6mo + orderinrecent12mo + amtuponorders +
    qtyuponorders + amtupontof + Mhistory19 + Mhistory19:Fhistory19,
    data = bookModTrain2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64296 -0.32262  0.00523  0.39614  1.63381

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.126e+00  1.582e-01  19.754  < 2e-16 ***
recency                -2.242e-04  1.484e-04  -1.511  0.13190
frequency              -4.424e-02  9.565e-03  -4.626 5.84e-06 ***
amount                  1.411e-03  3.319e-04   4.251 2.95e-05 ***
qty3mo                  4.102e-02  2.877e-02   1.426  0.15505
price3mo               -6.097e-03  3.403e-03  -1.792  0.07429 .
qty6mo                 -4.336e-02  1.988e-02  -2.181  0.03007 *
price6mo                5.110e-03  2.649e-03   1.929  0.05482 .
orderinrecent12mo      -1.887e-01  1.319e-01  -1.430  0.15385
amtuponorders          -3.593e-03  1.283e-03  -2.801  0.00547 **
qtyuponorders           1.347e-01  2.405e-02   5.601 5.31e-08 ***
amtupontof              1.084e-01  6.416e-02   1.690  0.09226 .
Mhistory19              1.819e-03  9.504e-04   1.914  0.05666 .
Mhistory19:Fhistory19 -8.936e-05  3.016e-05  -2.963  0.00333 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5827 on 266 degrees of freedom
Multiple R-squared:  0.3709,    Adjusted R-squared:  0.3402
F-statistic: 12.06 on 13 and 266 DF,  p-value: < 2.2e-16
```

- Section (c) : Initial Linear Regression Model Diagnostics



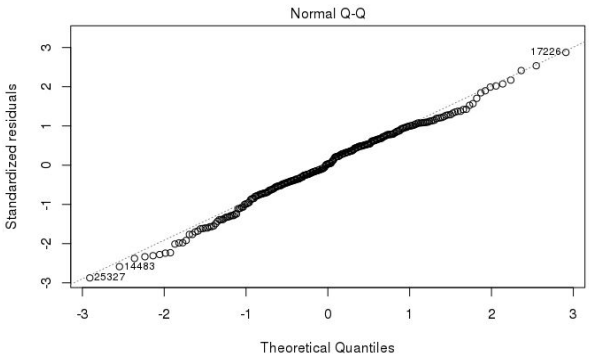
frequency	amount	qty3mo	price3mo	amtuponorders
6.883625	14.297105	9.265215	15.035247	4.071968
amtuponqty	ordersupontof	amtupontof	qtyupontof	Mhistory19
1.892545	4.618931	28.952873	27.762682	8.875483
Fconthist20	Mconthist20	Mhistory19	Fhistory19	
99.492622	105.175957	8.097797		

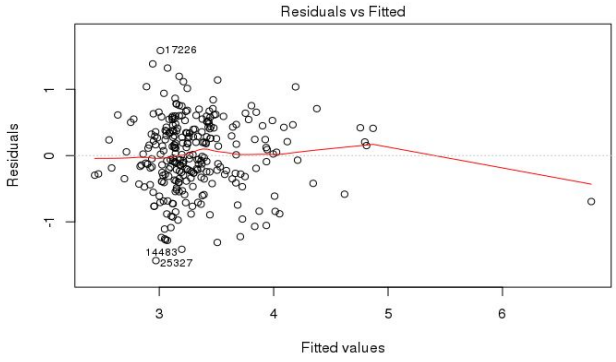
VIF for the predictors in the initial model

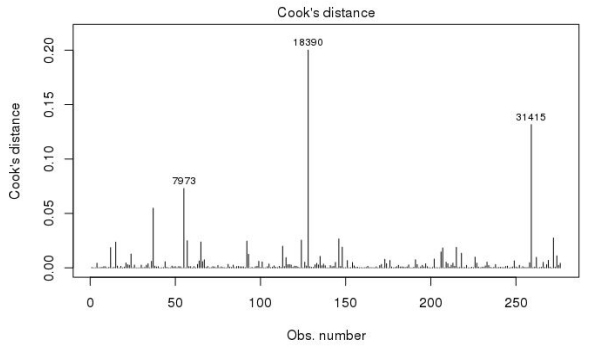
- Section (d) : Final Linear Regression Model Summary

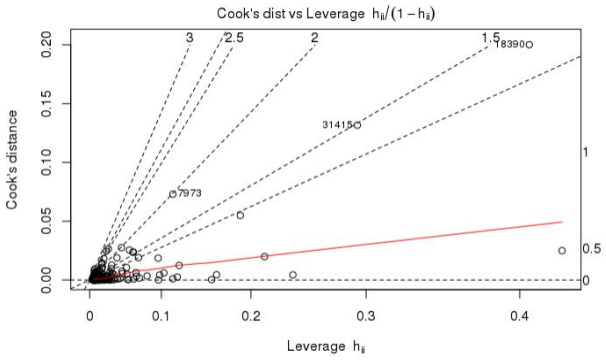
```
Call:
lm(formula = logtargamt ~ frequency + amount + amtuponorders +
    amtuponqty + Mhistory19 + Mhistory19:Fhistory19, data = bookModTrain3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.58455 -0.33093  0.01536  0.39635  1.58492

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.193e+00  1.041e-01  30.673  < 2e-16 ***
frequency             -4.075e-02  9.377e-03  -4.345 1.97e-05 ***
amount                 1.263e-03  3.214e-04   3.929 0.000108 ***
amtuponorders          9.677e-03  2.307e-03   4.194 3.73e-05 ***
amtuponqty            -2.372e-02  8.485e-03  -2.796 0.005549 **
Mhistory19             1.900e-03  9.687e-04   1.961 0.050899 .
Mhistory19:Fhistory19 -7.702e-05  2.989e-05  -2.576 0.010515 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5535 on 269 degrees of freedom
Multiple R-squared:  0.385,     Adjusted R-squared:  0.3713
F-statistic: 28.07 on 6 and 269 DF,  p-value: < 2.2e-16
```

- Section (d) : Final Linear Regression Model Diagnostics










```
frequency          amount     amtuponorders      Mhistory19 Mhistory19Fhistory19
 6.141723       11.269220          2.735749        7.079868            5.802651
amtuponqty
 1.331779
```

VIF for the predictors in the final model

Appendix 7 : Testing on Test data

- Section (a) : Characteristics for outliers in the test data

	id <int>	amtuponorders <dbl>	qtyuponorders <dbl>	amtuponqty <dbl>
12049	5900190	48444.73	4843.455	10.002102
29495	14158205	84247.71	16878.667	4.991372