PREDICTIVE ANALYTICS PROJECT

Predictive Modelling of Customer Response Behavior for an Online Bookseller

by

Sophie Du, Michael Fedell, Tova Simonson, Eileen Zhang

December 6, 2018

Professor Ajit C. Tamhane
MSiA 401: Predictive Analytics I

Table of Contents

Executive Summary

This paper discusses the strategy, methodology, and results of predicting customer behavior in response to a promotional offer from an online bookseller. The most important predictors for customer response likelihood relate to order date history, frequency and recency of purchases, total amount spent, how long they've been customers, and the amount spent on their largest order. When predicting amount of sales, the key predictors are frequency of orders, maximum price spent on an order, amount per order, total products purchased, and whether the customer purchased in the past month. A logistic regression model is first used to predict a customer's likelihood to respond to a promotion which is then used to scale the predicted spending amount as determined by a multiple regression model. The final models can be used to drive a marketing strategy which will optimize short-term profit at a total payoff of €11850.89, or 34.2% of the maximum payoff for the top 1516 responders.

Introduction

In order to evaluate customer behavior in response to promotional offers for a German online book seller, we developed a predictive model based on historical sales data from the seller. This data was broken up into a training and test set to develop and validate the models. Our approach to building the model began with an exploratory analysis to form a priori hypotheses and judge the quality of each predictor. These predictors were evaluated for use in the logistic regression model, which would estimate the probability of a customer responding to a promotional offer, as well as for use in the multiple regression model, which would predict how *much* responders spend on the promotion. After fitting our models to optimize the relevant statistical criteria, we examined the financial criterion to measure the potential business impact of our predictive models. Though our final models grew to leverage many predictors, we began with three main hypotheses to evaluate.

When approaching the logistic model, we hypothesized a relationship between order date history and likelihood to respond to a promotional offer. The `recency` variable, which is provided in the dataset, indicates number of days since last order. This allows us to draw conclusions from a linear relationship with a customer's most recent order. For example, an increase in amount of days since last purchase may correspond to a decrease in likelihood to respond. We hypothesized that this variable may not actually correlate with response linearly since a customer who purchased very recently (past few days) may be saturated and thus less likely to respond, while a customer who purchased in the time frame of one to three months ago may be ready to spend again and thus more likely to respond. Further, someone who has not purchased in the past year is probably less likely to respond.

In addition to a relationship with order date buckets, we believed customers who have returned a product in the past may be more or less likely to respond. A customer may return a product when they are not satisfied with it, which could make them less likely to purchase again. Conversely, the may be more likely to purchase in the future if they have a satisfactory return experience. In order to test this relationship, we included a binary variable which indicates if a customer has returned a product.

For both models we hypothesized interactions between amount, frequency, and time on file for customers. Moreover, amount and frequency are both likely to increase as time on file increases. Therefore, we decided both models should include interaction terms to standardize these variables for number of orders per time on file and amount per time on file. Overall, we expected that customers with a higher amount per time on file and more orders per time on file would be more likely to respond to the promotional offer because they order frequently and spend more money. In particular, the amount per time on file was of great interest since customers who spend a lot of money will benefit more from promotions than customers who will be saving a marginal amount if they respond.

The next section in this paper will discuss how we cleaned the data after exploratory data analysis. After, we will discuss the statistical methods used to fit and evaluate both our classification and linear models. We then validated those models on statistical and financial criteria, which will be discussed before our final conclusions.

Data Cleaning and Exploratory Data Analysis

To evaluate the usefulness of the included data, our group began with exploratory data analysis which would help identify predictors of interest to include in the binary logistic regression and multiple regression models. This search also included a basic examination of the included continuous predictors plotted against the response variable (logtargamt). A complete description of all available variables and the aforementioned plot can be found in the Appendix

(Table 5, Graph 4). We were given four data files (book, orders, training, testing) where training data and testing data were subsets of the book data (per customer basis). We noted that the test set contained approximately three times as many observations as the training set. We then examined the number of customers who responded to the promotion (logtargamt > 0) and found that the number of non-responders far outweigh the responders. Also, in the training data, there existed a column labeled `X` with only a single value for the first observation. Furthermore, the variables for `frequency` and `amount` in the book data represented the aggregation of the more granular `Fxx` and `Mxx` variables; we decided to only use the aggregate variables and thus disregarded all `Fxx` and `Mxx` variables.

We further explored each dataset by generating boxplots and histograms of some key variables to examine the distributions and found that most of the data was extremely long-tailed, indicating the existence of outliers. We also noticed that the date (orddate) information in the orders data used a non-standard format including both numbers and characters. Moreover, given that the data files were matched with common customer IDs, we considered merging the book data and orders data on the IDs and creating new variables which combined information of the columns with interactions. While customers could each have several orders, we grouped orders by the unique customer IDs, and calculated the total amounts and quantities of their orders.

We then performed data cleaning on inconsistencies and outliers in the data. First, we dropped the empty column "X" in the training data since all entries in the column are NAs. Then we re-formatted the variable order date (orddate) in the orders dataset from "D-Mon-Yr" (e.g. 2-Dec-09) into "YYYY-MM-DD" (e.g. 2009-12-02). To remove the outliers, we started with removing individual orders with the quantity of books ordered larger than 1000 (qty > 1000) in the orders dataset. As a result, we removed five observations in this step. Then, after merging training data, orders data, and book data by the unique customer ID, we double checked that among the customers with non-zero time on file (tof > 0), the total quantities of books ordered by each customer are all less than 1000 (sumQty < 1000). The customers with zero time on file (tof = 0) represent new customers who have not yet ordered books. Since no historical data exist for these customers, they were excluded from the training data and handled separately during model validation.

After fitting baseline models for the linear and logistic regressions, we used Cook's Distance to detect the influential observations using a confidence level of 10%, where$D_i > f_{p+1,n-(p+1),\alpha} = f_{7,8217,0.1}$. The confidence level 10% may seem very low, but here our goal is to identify influential observations, rather than estimate significance with a high level of confidence. In this step, we removed one observation (id = 5635683) from the linear model and no observations from the logistic model.

Model Fitting

Classification Model

After initial data cleaning and exploratory data analysis, we fitted a binary logistic model to estimate the probability of customers responding to promotion offers. We merged the book dataset with orders and trained dataset on customer ID to create a full training dataset. The initial variables included in the dataset were limited, and in order to better predict the probability of being responders, we added several of the engineered variables discussed previously. We hypothesized that the total quantity ordered by a customer as well as the average and maximum amount of money spent on any order would correlate strongly to whether an individual responded to the offer. To create these predictors, we grouped the orders by customer ID and calculated their total order quantities as `sumQty.` For each order, quantity was multiplied with associated price to get the net amount per order. Since each person's average daily spending and average daily number of orders placed were also thought to be potentially influential, the variables `amountPer` and `sumQtyPerTof` were created respectively to represent these two factors. Lastly, we added a predictor to indicate if a customer had ever returned an order.

We also thought that time-related variables would affect if customers respond to the offer since people who placed an order a long time ago might no longer shop on this website and thus were less likely to be responders. In order to group orders by their order date, we computed the time difference between each person's order dates and 08/01/2014, upper date of data and day of promotion. Each order was then categorized as `oneMonth`, `threeMonth`, `sixMonth`, `oneYear` and `overYear` (placed more than a year ago) using one-hot encoding and these time-related variables were treated as dummy variables.

One big issue with the original training dataset was the imbalance between responders and non-responders, as only 3.93% of customers were responders. Such a strong imbalance results in a model that heavily predicts no-response resulting in at least 96% accuracy. Therefore, resampling the training dataset was necessary to adjust the ratio of responders to non-responders. We replicated the observations of responders four times and then reunited these rows with those for non-responders to form the new training set, which contained roughly 13% responders and 87% non-responders.

Before fitting the model, we checked multicollinearity with our baseline model that included all fifteen variables, average and maximum amount of money each individual spent (avgNetOrder, maxNet), total quantity of book ordered (sumQty), recency, total past purchase (amount), number of orders (frequency), total time on file (tof), daily spending (amountPer), daily order number (sumQtyperTof), whether a return was made (returned) and five date-related variables. Among all fifteen predictors, sumQty was the only one with VIF > 10 (12.68), so it was removed from the model.

With the rebalanced training set, the model was then refitted with `sumQty` removed then altered with stepwise regression resulting in a model with AIC minimized. The stepwise regression selected recency, frequency, amount, `tof`, `amountPer`, `oneMonth`, `threeMonth`, `sixMonth`, `overYear`, `maxNet` and `sumQtyPerTof` to be in the model with AIC of 6328. However, the summary output for this model indicated that `amountPer` and `sixMonth` have p-value equal to 0.0651 and 0.0588 respectively, meaning that these two variables are not very significant when predicting the likelihood of being a responder. Therefore, `amountPer` and `sixMonth` were dropped from the model. After refitting the new model, all 9 variables turned out to be very significant with p-values smaller than 0.05. The AIC for this new model was 6330.7, which was only slightly higher than the one chose by stepwise regression. However, since the new model had fewer predictors, we decided to use the following model as our final logistic model:

$$\text{Respond} = \beta_0 + \beta_1\, \textit{recency} + \beta_2\, \textit{frequency} + \beta_3\, \textit{amount} + \beta_4\, \textit{tof} + \beta_5\, \textit{oneMonth} + \beta_6\, \textit{threeMonth} + \beta_7\, \textit{overYear} + \beta_8\, \textit{maxNet} + \beta_9\, \textit{sumQtyPerTof} + \varepsilon$$

Another criterion for assessing the logistic model goodness of fit is deviance. The difference between the null deviance and residual deviance of our final model is 468.6. This means that the inclusion of predictor variables decreased the model's deviance by 468.6 and the corresponding chi-square is extremely small (2.11e-116), indicating that our model is very significant. According to the final logistic model, recency, amount and total time on file are negatively correlated with response probability meanwhile `frequency`, `oneMonth`, `threeMonth`, `overYear`, `maxNet` and `sumQtyPerTof` have positive relationship with the likelihood of responding to promotion offers. These results align with our initial hypothesis especially regarding the effect of order data buckets.

To test the predictivity of the logistic model, we fitted the model on the merged test dataset. The fitted logistic model was trained on oversampled data and thus biased slightly towards successes (logtargamt > 0). Therefore, the resulted fitted probability with 4-fold oversampling should be corrected to obtain the true fitted probability of whether customers would respond to the promotion. We used $ln[p_1/(1-p_1)] = -ln(m) + ln[p_2/(1-p_2)]$ to compute the true response probability. Here, p1 denotes the true fitted probability with no oversampling, p2 denotes to the fitted probability after oversampling and m is the number of folds.

To evaluate the performance of the classification model on the testing set, we plotted the ROC curve, which illustrates the diagnostic ability of our classifier as its discrimination threshold is varied. The area under the ROC curve (AUC) of the model is 0.691, indicating there is 69.1% chance our final logistic model distinguished between a responder and a non-responder when fitting on the test dataset. Due to the highly imbalanced dataset, the ROC curve was chosen as the main evaluation metric. If we use correct classification rate (CCR), the CCR could still be very high even if the model failed to identify anyone as responders since there were nearly 97% of non-responders in the test dataset. Therefore, CCR would be pretty high (over 96%) regardless the actual predictivity of the classifier.

Multiple Regression Model

Using the same cleaned and merged training data for the logistic model, which included train, books, and orders, we fitted our linear model. We subset the data for just responders, those with a `logtargamt` greater than zero. As with the logistic model, we included five binary variables for order date buckets, a `returned` binary variable, interaction terms, net price, maximum price on a single order for a customer, and the total amount of products ordered by a customer. We decided to create and include these variables because of the limited variables provided and as a result of the a priori hypotheses we formed.

Before we fitted our final model, we created a baseline model that included the predictors we were interested in. Using this baseline model, we checked for outliers using Cook's Distance, as discussed in the last paragraph of the data cleaning section, resulting in the removal of one observation (id = 5635683). After preparing the data for responders, we proceeded with fitting the model using a stepwise regression in both forward and backward direction. By minimizing AIC, the stepwise function found the best model to be:

$$\ln \text{targamt} = \beta_0 + \beta_1 \, \textit{frequency} + \beta_2 \, \textit{maxPrice} + \beta_3 \, \textit{amountPer} + \beta_4 \, \textit{sumQty} + \beta_5 \, \textit{oneMonth} + \varepsilon$$

The minimized AIC for this model was -272.82. Four of the five variable are significant past the 95% level, with a p-value less than 0.05. One of the variables, oneMonth, which is a binary variable for customers who have purchased in the past month, is significant at the 90% level, has a p-value under 0.1. In order to evaluate our model, we used R-squared and Adjusted R-squared, which were 0.2995 and 0.2867 respectively. While these are seemingly low, data for customer behavior is more difficult to fit because of the high natural variation in consumer behavior and preferences, resulting in a lower goodness of fit measure. Also, many additional variables would be helpful and likely to fit the model better, but we did not have access to them. These will be discussed in the conclusion section. Overall, we believe our R-squared and adjusted R-squared values to indicate a good model fit. Our overall model has a very significant F-stat of 23.34 on five variables and 273 degrees of freedom with a p-value: < 2.2e-16.

Although there were no signs of assumption violations since all our predictors are significant, the overall model is significant, and we observed the expected signs for all of the coefficients, we double checked. For multicollinearity, we ran a VIF function on the regression model resulting in VIF values that were all below ten, with the highest values being 4.516 for sumQty and 4.210 for frequency. Since none of these values are over ten, there is no indication of multicollinearity. We then looked at the Q-Q Plot for normality, which did not indicate a normality violation as the points fell mostly on the line (graph 3). For heteroskedasticity, we saw a relatively uniform distribution once zooming into the clustered points (graph 4).

Model Validation

Although the models were developed individually on the relevant training data, their true utility would only be evident when brought together to form predictions on the test data. At a high level, the logistic regression would be used to predict the likelihood that a customer would respond to the offer, and the multiple regression would then predict how much those customers would spend if they responded. Both models were applied to the test set and then the *expected* logtargamt calculated by taking the product of likelihood of response and predicted logtargamt. These expected values could then be compared to the actual amounts present in the test data to evaluate accuracy. Additionally, the expected logtargamt was transformed back to expected amount in euros, representing the model's prediction for revenue generated by each customer's response to the promotion. A more detailed summary of the validation process follows.

The full test data set was generated by taking the provided ids and known logtargamt's then joining all the original and engineered features from the "book" dataset that were used in the model development stage. Additionally, outliers were removed based on extreme values of \`sumQty\` (total number of books ordered by customer), and \`amountPer\` (euros spent customer per time on file). Several customers in the test data had \`amountPer\` values more than 10 times greater than any value present in the training data and were thus removed. Lastly, all new customers (tof = 0) were set aside for the time being as the model would not be able to make meaningful predictions for customers without historical data.

The prepared test data was then processed by the logistic model, producing a response likelihood value between 0 and 1. Of these predictions, 75% were less than 5% likely to respond; although maximum likelihood extended to nearly 93%. Optimizing the classification threshold to maximize the F1-score resulted in a precision of 0.058 and a recall of 0.64, and F1-score of 0.107. Next, the multiple regression model was used to predict the amount spent by each customer. These predictions were densely clustered with the interquartile range captured between 3.184 and 3.332; although values extended to a maximum of 17.6. Before moving forward, expected values would need to be calculated for the new customers as well. Since these customers do not have any historic data, their response likelihood was recorded as 52.9%, the average response rate for new customers in the training data. Similarly, their predicted `logtargamt` was recorded as 1.753, the average `logtargamt` for new customers in the training data. Now that likelihoods and `logtargamt` predictions were available for all customers, an *expected* amount could be calculated for each by taking the product of both predictions. These scaled predictions were much more moderate with an IQR between 0.084 and 0.165. The `logtargamt`s would need to be transformed back to standard amounts in euros in order to be more informative. Both the expected and actual `logtargamt`s were transformed back to standard target amounts and then compared to evaluate the model's accuracy. The total SSEP was found to be 2,526,649 across 25,386 observations (RMSE = 99.53). However, nearly 40% of this SSEP is captured by the top fifteen observations by greatest error.

Once the model's statistical accuracy was evaluated, the model was ready to be examined in a more practical light. First, the model would be evaluated once more using a financial criterion. The top 500 customers by predicted response amount were selected for evaluation. The actual amount spent by these customers after the promotion was 6,672.74 euros, or approximately 24.7% of the predicted response value for the 500. Perhaps a more interesting question is to ask how the company may optimize their marketing strategy so that they maximize revenue and minimize cost. Assuming a profit margin of 25% on each purchase and a flat marketing cost of one euro per customer, the short-term profit function can easily be evaluated against the customers in the test set. Starting with the customers with highest predicted target amounts, the cumulative profit is maximized at 1,446.72 euros generated by marketing to the 1,516 top prospects. We can also examine the payoff and percent payoff for these optimal 1,516 prospects. The total payoff (actual amount spent) for the top 1,516 predictions is 11,850.89 euros which is about 34.2% of the revenue from the actual top 1,516 responders. For a response as sporadic as customer buying behavior, we believe that being able to account for 34% of the actual response will be a huge asset to marketing and strategy departments at this bookstore.

Conclusions

Several intuitions can be deduced from the preceding analysis and evaluation of relationships. We will look at the key predictors in each of the final models as well as the nature of their coefficients to better understand customer behavior. First of all, recency and frequency coefficients showed opposite signs in the logistic regression meaning that customers are less likely to respond if they have purchased very close to the promotion date, but more likely to purchase if they have made many purchases throughout their time as customers. Additionally, customers are most influenced towards responding if they have made an order in the past one to three months, validating our initial hypothesis. Lastly, we note that the total *quantity* of books purchased by a customer (normalized to their time on file) is a significant indicator of likelihood to respond, confirming an intuition that loyal and heavily invested customers are more likely to be interested in a promotion. Moving on to the multiple regression, we observe that customers will spend more money on the promotional offer if they have made large purchases in the past (indicated by maxPrice). The most obvious intuition is that of cumulative amount spent by a customer during their time on file. This predictor (amountPer) does indeed turn out to significantly predict how much they will spend on the promotion (positive relationship).

Although these variables do a decent job in predicting customer behavior, we believe that predictions could be improved by the capture and inclusion of several currently unavailable data. First and perhaps most obvious is the dates of any previous promotional offers. This would allow us to examine the change in customer behavior during promotional periods in the historical data, improving our understanding of the response likelihood. Additionally, we believe it may improve our model to include data around customer demographics such as age, income, family status, etc. Finally, we think that vast improvements could be made to our data with simply more volume in orders data. This could allow profiling of individual customers based on their interests and habits. Then a targeted promotion could be crafted to cater to the individual profile of those customers most likely to respond to any generic promotion.

Though the inclusion of several unavailable data may improve the specificity and accuracy of our model, its current state does a satisfactory job of explaining and predicting customer behavior in response to a promotional offer. We believe that the use and correct application of this model will significantly help the client predict and strategize around their customers' behavior, driving effective and efficient marketing strategies into the future.

References

Tamhane, Ajit C., and Edward C. Malthouse. *Predictive Analytics: Parametric Models for Regression and Classification Using R*. A JOHN WILEY & SONS, INC., PUBLICATION.

Appendix

Table 1: Logistic Regression Output[1]

	Coefficient
Recency	-0.000677 *** (4.7e-10)
Frequency	0.04942 *** (1.14e-10)
Amount	-0.0007059 *** (4.32e-5)
tof	-0.0003683 *** (2.62e-8)
oneMonth	0.2972 *** (0.00387)
threeMonth	0.6068 *** (9.96e-12)
overYear	0.5903 *** (4.36e-12)
maxNet	0.001873 ** (0.00195)
sumQtyPerTof	4.127 *** (1.21e-08)

[1] Standard errors are reported in parentheses. ***, **,*, . denote significance at the 0.1%, 1%, 5%, and 10% levels respectively.

Table 2: Logistic Regression VIF Values

Variable	Rencency	Frequency	Amount	tof	oneMonth	three Month	overYear	maxNet	sumQty PerTof
VIF	1.57	3.96	3.91	2.62	1.21	1.21	1.88	1.55	1.39

Graph 1: ROC Curve

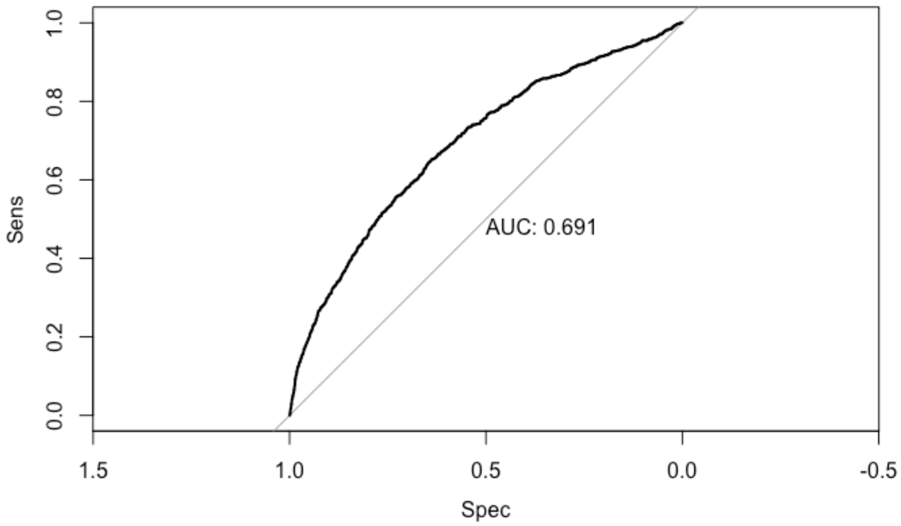


Table 3: Linear Regression Output[2]

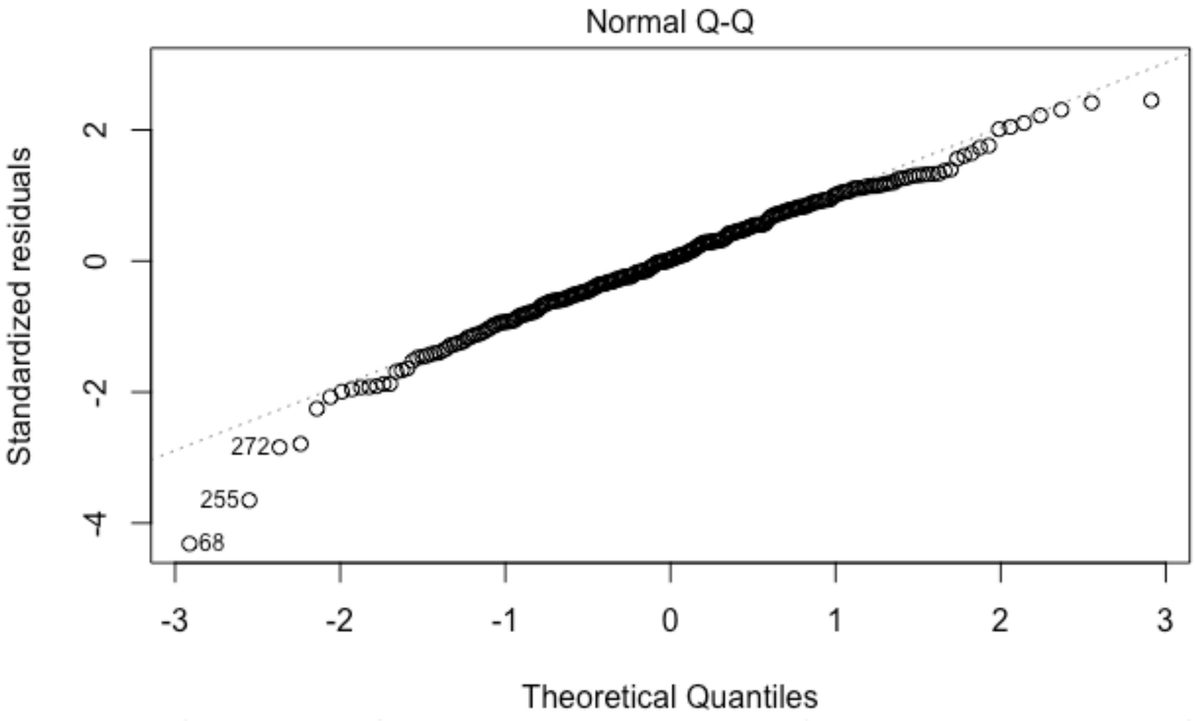
	Coefficient
Frequency	-0.058835 *** (0.058773)
maxPrice	0.002768 ** (0.001103)
AmountPer	0.170934 *** (0.052422)
sumQty	0.017819 * (0.002544)
oneMonth	-0.181183 . (0.101100)

Table 4: Linear Regression VIF Values

Variable	Frequency	maxPrice	amountPer	sumQty	oneMonth
VIF	4.210123	1.227934	1.348778	4.515691	1.103212

[2] Standard errors are reported in parentheses. ***, **,*, . denote significance at the 0.1%, 1%, 5%, and 10% levels respectively.

Graph 2: Normality - Linear Model



Graph 3: Heteroskedasticity - Linear Model

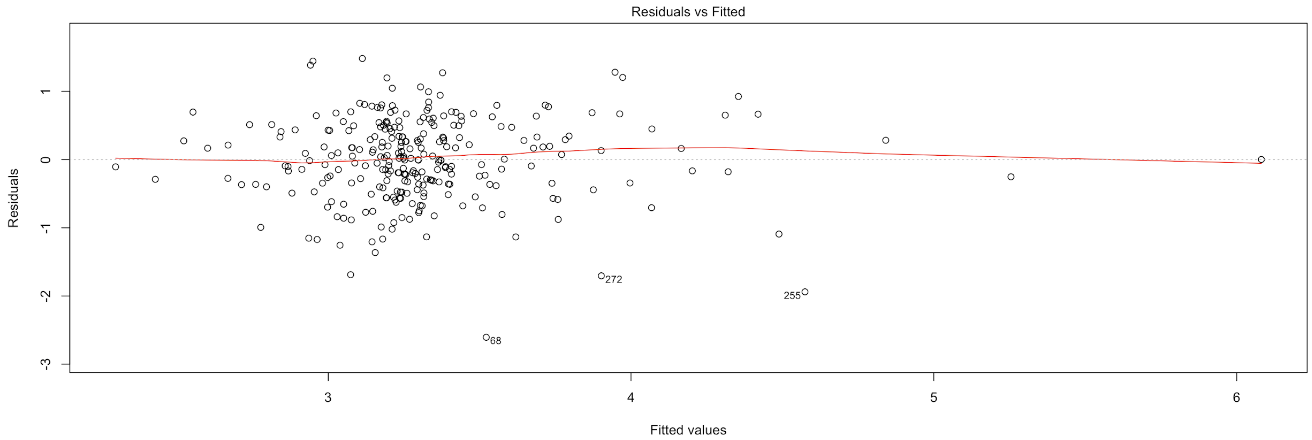
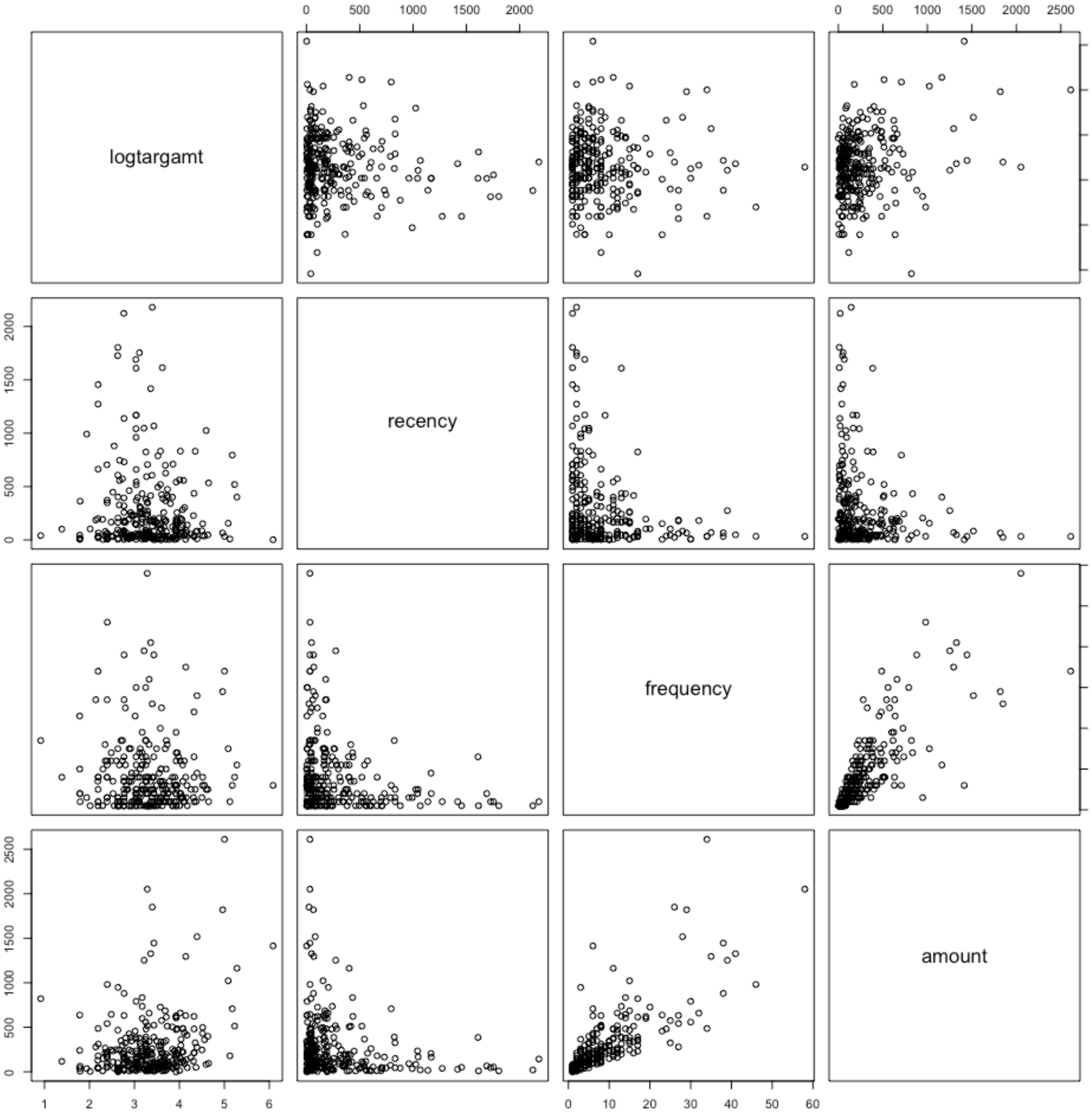


Table 5: Variable Descriptions

Variable	Description
id	Unique customer ID
logtargamt	Log amount purchased after promotion
respond	Binary variable (1 = responder, 0 = non-responder)
recency	Number of days since the last order
frequency	Number of orders
amount	Total purchase amount in euros
tof	Time a customer has been on file (days)
ordersPer	Number of orders per time on file: frequency/tof
amountPer	Amount spent per time on file: amount/tof
sumQty	Total products purchased
avgNetOrder	Average total price of each order
avgSumQty	Sum of quantity divided by order count
maxNet	Maximum order total - sum(qty*price) for orders
maxPrice	Maximum amount spent on a single order for a customer
returned	Binary variable (1 = returned product in pat, 0 = as never returned)
oneMonth	Binary variable (1 = has purchased in past 31 days, 0 = has not)
threeMonth	Binary variable(1 = purchased between 1 mo. and 3 mo. ago, 0 = has not)
sixMonth	Binary variable (1 = purchased between 3 mo and 6 mo ago, 0 = has not)
oneYear	Binary variable (1 = purchased between 6 mo & year ago, 0 = has not)
overYear	Binary variable (1 = purchased over a year ago, 0 = has not)

Graph 4: Base Variable - Response Correlations



Model Results

The logistic regression and mulitple linear regression models willbe combined together to come up with a predicted response for each customer in the test set.

This will be done by first predicting the customer's liklihood to respond to the promotion via the logistic regression model. After determining response liklihood, the mulitple linear regression will be applied to predict how *much* a customer will purchase after the promotion.

In summary: E(logtargamt) = P(logtargamt > 0) * E(logtargamt | logtargamt > 0)

Logistic Regression

The logisitc model was developed by fitting a binomial generalized linear model to the test data set using a collection of predictors both natural and engineered. The predictors used in the final model are as follows:

- avgNetOrder: average amount of money each customer ever spend
- sumQty: The total quantity each customer ordered
- recency: no. of days since the last order
- frequency: number of orders
- tof: each customer's time on file
- amountPer: total past purchase amount/time on file
- oneMonth: dummy variable- 1 if the customer placed order within one month prior to 08/01/2014
- threeMonth: dummy variable- 1 if the customer placed order within three month prior to 08/01/2014
- overYear: dummy variable- 1 if the customer placed order over one year prior to 08/01/2014
- sumQtyPerTof: frequnecy of placing orders

Multiple Regression

The mulitple linear regression was produced by several iterations of feature engineering and model variation. Ultimately, a best model was obtained via stepwise regression allowed to move in both forward and backward directions. The final predictors used are as follows:

- recency: no. of days since the last order
- frequency: number of orders
- amount:total past purchase amount in euros
- amountPer: total past purchase amount/time on file
- sumQty: The total quantity each customer ordered
- oneMonth: dummy variable- 1 if the customer placed order within one month prior to 08/01/2014

Synthesis

These two models will be brought together to form predictions on the expected amount purchased for customers in the test set.

We will narrow down the full test data to just the needed variables, while also filtering out several of the significant outliers (with more than 1000 orders)

```
predictions <- test.full %>%
  select(id, avgNetOrder, sumQty, recency, frequency, tof, amount, maxNet, maxPrice,
         amountPer, oneMonth, threeMonth, sixMonth, oneYear, overYear, sumQtyPerTof,
         logtargamt) %>% filter(tof > 0, sumQty < 1000, amountPer < 15)
```

The predictions given by the logistic regression are biased towards responders becuase the training data was oversampled to better balance responses. This bias will be corrected below

```r
p2 <- predict(log.best, newdata = predictions, type = 'response')
m <- (nrow(filter(train.rebalanced, respond == 1)) / nrow(train.rebalanced)) /
  (nrow(filter(train.full, respond == 1)) / nrow(train.full))
p1 <- exp(-log(m) + log(p2 / (1 - p2)))/(1 + exp(-log(m) + log(p2 / (1 - p2))))

predictions$probRespond <- p1
summary(predictions$probRespond)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## 0.002883 0.025395 0.034343 0.043129 0.049550 0.925039
```

The multiple regression can now be applied to the set to predict the amount purchased.

```r
predictions$predLogtargamt <- predict(ml.best, newdata = predictions)
summary(predictions$predLogtargamt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -0.7522  3.1812  3.2487  3.3153  3.3587 19.1409
```

Lastly, we need to include the new customers in the test set. These customers do not have any data on file and thus we will not be able to form sensible predictions for their response. Instead, we will keep use the average values for new customers: 52.9% likely to respond and spending an average logtargamt of 1.753316

```r
newCustomers <- test.full %>%
  select(id, avgNetOrder, sumQty, recency, frequency, tof, amount, maxNet, maxPrice,
         amountPer, oneMonth, threeMonth, sixMonth, oneYear, overYear, sumQtyPerTof,
         logtargamt) %>% filter(tof == 0)
new.avg.response.rate <- nrow(filter(train.full, tof == 0, logtargamt > 0)) /
  nrow(filter(train.full, tof == 0))
new.avg.logtargamt <- mean(filter(train.full, tof == 0)$logtargamt)
newCustomers$probRespond <- new.avg.response.rate
newCustomers$predLogtargamt <- new.avg.logtargamt
predictions <- rbind(predictions, newCustomers)
```

Now that the predictions for all customers are in place, they can be combined to yield an overall expected logtargamt based on liklihood to respond and predicted logtargamt.

```r
predictions$scaledLogtargamt <- predictions$probRespond * predictions$predLogtargamt
summary(predictions$scaledLogtargamt)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -0.67824  0.08412  0.11377  0.15183  0.16532  4.91407
```

The predicted logtargamts must now be transformed back to a normal, predicted amount in euros by performing the inverse of the original log transformation ($ln(amt + 1)$)

```r
predictions$predTargamt <- exp(predictions$scaledLogtargamt) - 1
predictions$targamt <- exp(predictions$logtargamt) - 1
summary(predictions$predTargamt)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
##  -0.49249   0.08776   0.12050   0.19752   0.17977 135.19315
```

```r
SSEP <- sum((predictions$predTargamt - predictions$targamt)^2); SSEP
```

```
## [1] 2534917
```

To help find significant error contributors, we can look at the Sq. Error term for each prediction

```
bad.apples <- predictions %>% mutate(sq.error = (predTargamt - targamt) ^ 2) %>% arrange(-sq.error) %>%
```

A separate measure of performance will evaluate the financial value in our model. This will be done by looking at the top 500 prospects as identified by predicted targamt (amount spent after promotion), and then examining how much these top prospects *actually* spent following the promotion.

```
top500 <- predictions %>%
  arrange(-predTargamt) %>%
  head(500)

payoff <- sum(top500$targamt); payoff
```

```
## [1] 6672.738
```

```
actual.top.500 <- predictions %>% arrange(-targamt) %>% head(500)
percentPayoff <- payoff / sum(actual.top.500$targamt); percentPayoff
```

```
## [1] 0.2468175
```

Now we can examine the optimal number of customers to target for the promotion by looking at the short term profit margins. If the profit margin is 25% of the `targamt` spent, and cost of marketing to a customer is 1 euro, then we will maximize total profit for marketing to the top `x` customers ordered by `predTargamt`.

We will do this by ordering the customers by descending predicted targamt and then calculating the cumulative profit where $profit = 0.25 * targamt - 1$

```
predictions$profit <- predictions$targamt * 0.25 - 1
opt.profit <- max(cumsum(arrange(predictions, -predTargamt)$profit)); opt.profit
```

```
## [1] 1446.723
```

This profit is obtained by marketing to the top x customers where x is calculated below

```
opt.customers <- which.max(cumsum(arrange(predictions, -predTargamt)$profit)); opt.customers
```

```
## [1] 1516
```

This new optimal marketing strategy will be used to calculate a new payoff and percent payoff

```
top.prospects <- predictions %>%
  arrange(-predTargamt) %>%
  head(opt.customers)

opt.payoff <- sum(top.prospects$targamt); opt.payoff
```

```
## [1] 11850.89
```

```
actual.opt.prospects <- predictions %>% arrange(-targamt) %>% head(opt.customers)
opt.percentPayoff <- opt.payoff / sum(actual.opt.prospects$targamt); opt.percentPayoff
```

```
## [1] 0.3424145
```

In summary:

```
list('SSEP' = SSEP,
     'Payoff' = payoff,
     'Percent Payoff' = percentPayoff,
     'Optimal Profit' = opt.profit,
     'Optimal Target Customers' = opt.customers,
     'Optimal Payoff' = opt.payoff,
     'Optimal Percent Payoff' = opt.percentPayoff)
```

```
## $SSEP
## [1] 2534917
##
## $Payoff
## [1] 6672.738
##
## $`Percent Payoff`
## [1] 0.2468175
##
## $`Optimal Profit`
## [1] 1446.723
##
## $`Optimal Target Customers`
## [1] 1516
##
## $`Optimal Payoff`
## [1] 11850.89
##
## $`Optimal Percent Payoff`
## [1] 0.3424145
```