IE7280

Project Report

Ming Luo

1. **Background:**

A retail company sells clothing on its website and via catalogs. They sent a catalog mailing to all customers in early fall of 2012 and recorded all the purchases. We aim to build a prediction model to find the potential customers who are likely to buy clothing due to the catalog mailing and how much they will buy, which is one kind of promotion in the industry, next year. In addition, there are 15 predictor variables and we would like to know which variables or interactions of variables are significant for prediction.

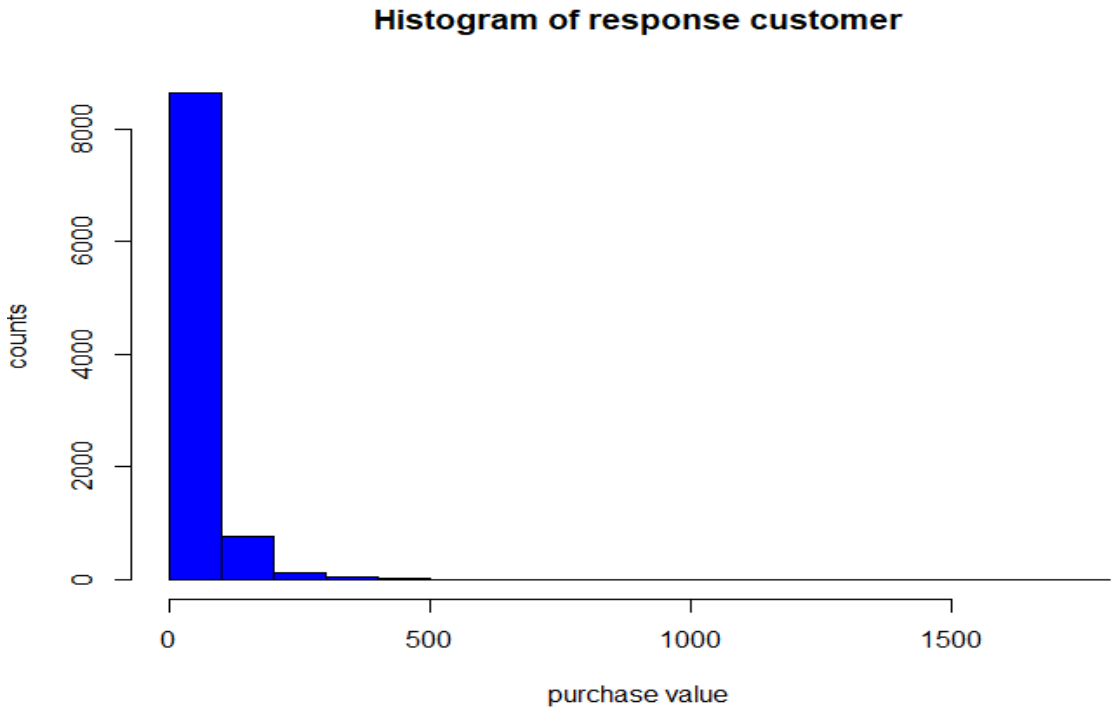
2. **Data Exploration And Preprocessing:**

Predictor variables		
Name	Description	Range
datead6	Date added to file	Date
datelp6	Date of last purchase	Date
lpuryear	Latest purchase year	Many missing value
slstyr	Sales ($) this year	[0, 1748]
slslyr	Sales ($) last year	[0, 2290]
sls2ago	Sales ($) 2 years ago	[0, 2942]
sls3ago	Sales ($) 3 years ago	[0, 2844]
slshist	LTD dollars	[0, 5091]
ordtyr	Number of orders this year	[0,8]
ordlyr	Number of orders last year	[0,11]
ord2ago	Number of orders 2 year ago	[0.9]
ord3ago	Number of orders 3 year ago	[0,10]

ordhist	the time since the customer purchased for the first time (LTD) order	[0.39]
falord	LTD fall orders	[0,106]
sprord	LTD spring orders	[0,21]

Since the ranges of predictors have a large gap, we can see that we should standardize the predictors.

Target variable	
Name	Description
targdol	Dollar purchase resulting from catalog mailing during fall 2012



There are a total of 101,532 customers. Among them, only 9571 customers (9.43%) purchased the clothing. And the distributions of the nonzero purchase amounts are highly right-skewed with some unusually large amounts. Therefore, we will do a log transformation to the y into ln(y+1). This log transformation has no impact on 0 purchases since they will be 0 as well after the transformation. Also, we did the same log transformation to the predictors.

In addition, the data are randomly divided into a training set with 50,418 observations and the remaining 51,114 into a test set.

Inconsistency:

- As we know that the variable "ordhist" should equal the sum of variables "falord" and "sprord". However, after we tested, there are 8792 customer's LTD orders are inconsistent with the sum of LTD spring orders and LTD fall orders. And based on the domain knowledge, we will consider the variable "ordhist" as an accurate one. In addition, "ordhist" contains the same information as "falord" and "sprord". Thus, we will remove the variables "falord" and "sprord" when fit to the model.

- In addition, the year of the latest purchase obtained from "lpuryear" variable and from "datelp6" variable is not consistent. However, the information in "lpuryear" is included in "datelp6", thus, we will remove "lpuryear" and use the information in "datelp6" as the accurate late purchase data.

- There are 205 cases that have the number of orders recorded but no sales amount recorded this year. There are 121 cases that have the number of orders recorded but no sales amount recorded last year. There are 124 cases that have the number of orders recorded but no sales amount recorded 2 years ago. There are 237 cases that have the number of orders recorded but no sales amount recorded 3 years ago. Since these cases are only a small portion of the overall dataset and we don't have other useful information to extrapolate the sales amount, thus, we will remove these cases before fitting into the model.

- In addition, sales for this year, last year, 2 years ago, and 3 years ago are correlated to LTD dollars since LTD dollars are the sum of sales amount for all the past years. Thus, we will use individual year sales as predictors instead of LTD dollars since it contains more detailed information.

Feature engineering:

- We will build a new variable “f_s” to combine the information in the variables “falord” and “sprord”. If people have orders only in fall, we will give the corresponding “f_s” value as 0.5. If people have orders only in spring, we will give the corresponding “f_s” value as 0.5. And if people buy in both fall and spring, we assign 1 to that customer.

- We will build a new variable “weight_avg_s” to indicate the consistency of past purchases and discount the order purchases more than the recent purchase. First, we will calculate the average sales (sales per order) for each of the past 4 years. Then we calculate the weighted average of the past 4 years by giving higher weight to the avg sales that were made recent. The weight for the current year is 4/1+2+3+4 = 0.4, last year is 3/1+2+3+4 = 0.3, 2 year ago is 0.2, and 3 years ago is 0.1.

- Since we can’t use date information directly, we build a new variable to calculate how long this customer has been with us by finding the difference between 2013 and time since the customer purchased for the first time. Thus, if 2012 is the year that the customer purchased for the first time, then we will assign 1 to it.

3. Modeling:

As suggested by the project instruction, since we have an imbalanced dataset we will use a two-step modeling approach involving logistic regression followed by multiple regression. The first step is to use logistic regression to model the response probabilities, p(y>0). Then, use multiple regression to model the conditional mean of the purchase amount given that the customer is a responder, E(y|y>0). Then we will estimate the the purchase amounts for the customers in the test set with the information we get from both steps: E(y) = E(y|y>0) * p(y>0). And the predicted value of the purchase amount for a customer in the test set is $e^{E(y)}$.

<u>logistic regression:</u>

We train the logistics regression model on the training set with oversampling technique. Because the percentage of respondents is very low in each data set, we will oversample respondents (duplicate each respondent in the training set 10 times) in the training set when fitting the logistic regression model to improve the accuracy of the estimates of the regression coefficients.
The actual proportion of respondents is q = 0.0925. After oversampling, the proportion of respondents rises to q’ = 0.5047, which is m = 5.4562 times of actual proportion. Thus, the bias-adjusted estimated response probability p is given by

$$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{p'}{1-p'}\right) - \ln\left[\frac{q'/(1-q')}{q/(1-q)}\right] = \ln\left(\frac{p'}{1-p'}\right) - \ln\left[\frac{m(1-q)}{1-mq}\right]$$

p’ the predicted probability of whether a customer will respond to the promotion on the test dataset.

Multiple regression:

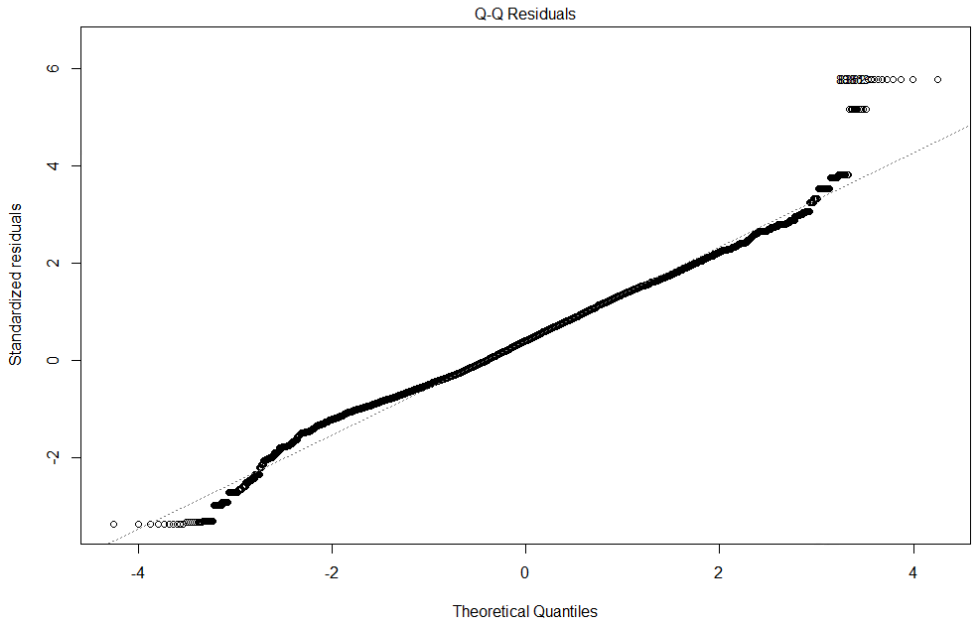
Since the interaction between variables could be significant in predicting the sales, we build our models with the individual variables and interactions between paired variables. And we apply the stepwise model selection technique to find the best models and remove unnecessary variables. The significant predictors of our final model are: ordhist, ordtyr, slstyr, slslyr, ordlyr, sls2ago, ord2ago, sls3ago, ord3ago, and paired interactions between them. The adjusted R-squared is 0.7565, which means our model is a good fit. And the p-value is very small, which is 2.2e-16, which also means our linear regression model as a whole is statistically significant in explaining the variability in the y.

Model fitting:

Then we apply the multiple regression model to our test set and multiply the result with corresponding probability that is predicted by the logistic regression model, then do the inverse log transformation to get the predicted purchase amount. In addition, there are some predictions that contradict our common sense, so we modified them with our domain knowledge and common sense. For example, if some predicted purchase amount is negative, we change them to 0. And few of them are extremely and unreasonably large, so we change them to the max purchase amount in the training set.

4. **Model Assessment**

Below is the normal QQ plot of residuals. We can see that our residuals approximately follow a normal distribution in most of the parts except for the upper end and lower end. This could be because of the modification we made above. Generally speaking, we can say that our model is a good fit.



In addition, the mean error is -3.658. It means that our model underestimated the sale by 3.658 on average. However, 3.658 is very small, so it indicates that our model's prediction ability is good.