Chris Pelkey
PREDICT 450-55
Solo 3

Introduction

This dataset contains various pieces of information from the XYZ Corporation. The initial dataset is a combination of customer data and ZIP code and Census data. This is inclusive of demographic data, like information on respondents ZIP code and what the make up of that ZIP code is, as well as individual items like previous mailings, responses to those mailings and money spent. This dataset contains 554 variables with 30,779 entries.

Exploratory Data Analysis

At the outset of the project, I knew that the data analysis portion of the project was going to be among the toughest portions of the project. With 554 variables, there is a lot of information to sort through to find out what might be relevant to the modeling of the mailing response.

Maybe to my own detriment, but I started out initially knocking out all variables of interest, in large part because a lot of this information is unknown by the survey data, so there are a lot of NA responses. This included whether people are excited about politics, home/garden, music, etc. These variables were eliminated from the mix.

Next up was the elimination of some of the Census data and some of the locational data. Because I already resolved to use the ZIP code variable in some way removing information like latitude and longitude, as well as tract number, block ID and other location data seemed repetitive.

Finally, large swaths of Census data were removed from the data set that I examined. While it might be important how much in terms of percentage the population in a certain area is Asian, I don't think it is particularly impactful on whether a certain person will respond to a mailing campaign. It may be useful to pull in an exact model, but we are trading predictability for accuracy of the fit at this time, so the fewer the variables in the model, the better.

After these initial cuts we can see in large part what is left is information on the respondents themselves.

Variables Created

Several variables were created to capture more accurately some of the trends that the data suggests. For example, variables were created that measure the life to date sales and transactions before 2009. Two more variables were created to measure

the total number of sales and total number of transactions per customer in the data set.

Variables Selected

After the initial exploratory data analysis, I was able to narrow down which variables I thought might be good to include in the modeling process. Much of this data comes from the initial customer data received from XYZ, but there is some information in the ZIP data pulled in and combined with the original data. Also, some variables were created from existing variables. In total, 17 variables were pulled into the data set for processing. Not each of these variables were used for all of the modes, these were just the variables scraped for NA, NaN and Inf.

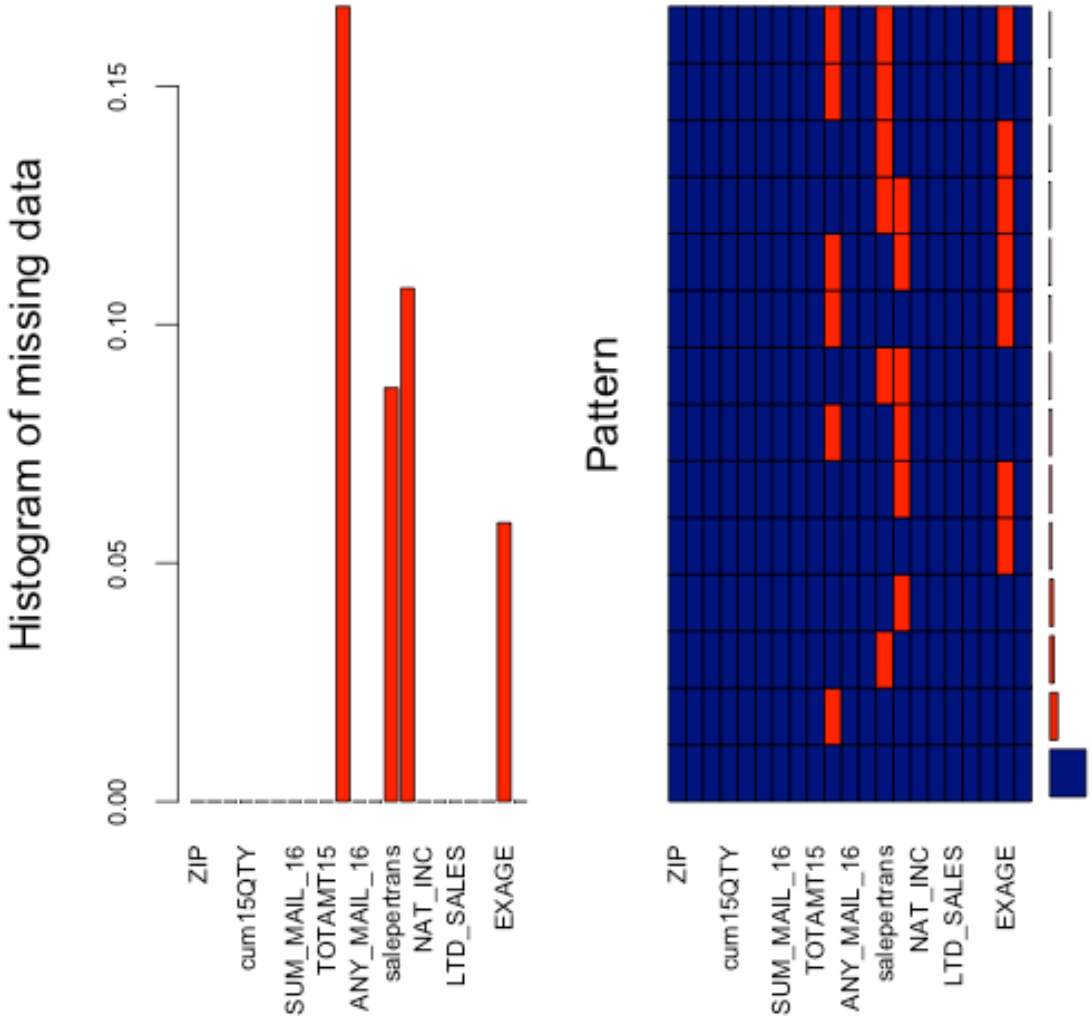
Field Name	Description
ANY_MAIL_16	Whether Mail Was Sent in Campaign 16
BUYER_STATUS	Status of the Buyer
cum15QTY	Cumulative Quantity Ordered
cum15TOTAMT	Cumulative Total Amount Spent
EXAGE	Age of the Buyer
LTD_SALES	Life to Date Sales
LTD_TRANSACTIONS	Life to Date Transactions
M_HH_LEVEL	Household Level Classification System
MED_INC	Median Household Income for ZIP
NAT_INC	National Income Percentile
PRE2009SALES	Life to Date Sales – 2009 Sales
PRE2009TRANSACTIONS	Life to Date Transactions – 2009 Transactions
salepercamp	Total Sales Divided by Campaign
salepertrans	Total Sales Divided by Number of Transactions pre-2009
SUM_MAIL_16	Total Mail Sent in Campaign 16
TOTAMT15	Total Amount Spent in Campaign 15
ZIP	5 digit zip code

Data Cleansing

In order to prepare for both the logistic and random forest classification models, all data has to be cleansed. This included transforming various variables into factors as opposed to numbers, and removing all NA, NaN and Inf values within the data set by dropping those instances.

Chris Pelkey
PREDICT 450-55
Solo 3

These variables actually led to having missing and infinite variable measures in the data. These can be seen in the graphic below. The variables also measure according to the table below the graph.



Missings in variables:

Variable	Count
salepercamp	2487
salepertrans	1294
HOMEOWNR	1606
EXAGE	872

Chris Pelkey
PREDICT 450-55
Solo 3

Modeling of the Data

Binomial Logistic Regression

The initial model that was built was a logistic regression model. This is the simplest model that we can play with. It involves determining a yes/no response from the variables that are input into the model.

This model was called with the following variables:

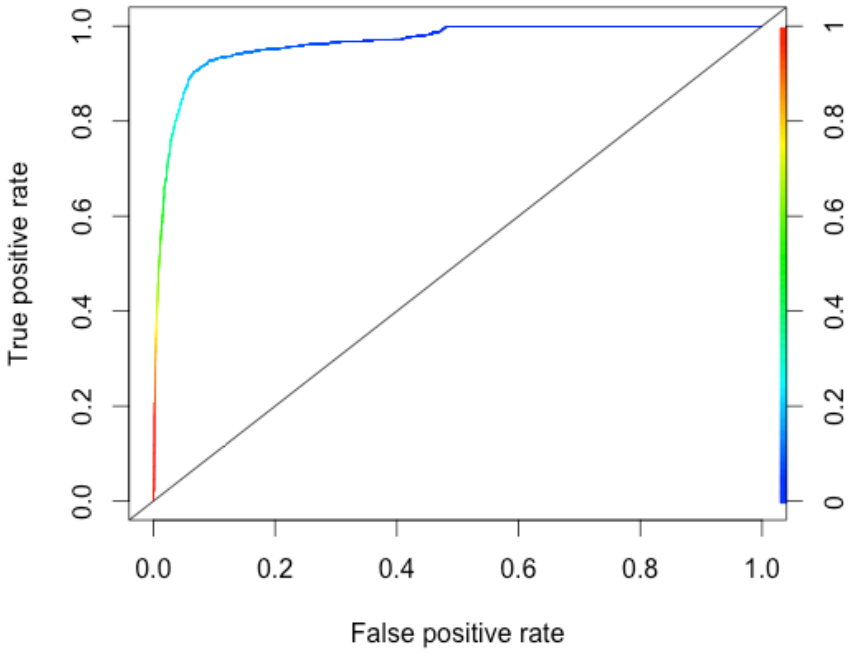
```
glm(formula = RESPONSE16 ~ ZIP + PRE2009SALES +
PRE2009TRANSACTIONS + cum15QTY + MED_INC + cum15TOTAMT +
SUM_MAIL_16 + TOTAL_MAIL_16 + TOTAMT15 + salepercamp +
ANY_MAIL_16 + salepertrans + NAT_INC + M_HH_LEVEL + LTD_SALES +
LTD_TRANSACTIONS + BUYER_STATUS + EXAGE, family = "binomial", data =
subdat3)
```

This model did fairly well considering the fact that it is a simple logistic model. Producing a confusion matrix on the run model with the actual response rate we see that we get a 93.5% accuracy rate.

	0	1
0	7309	108
1	438	622

This model produces some really interesting insights that we will examine more fully in the financial analysis section, but the 93.5% accuracy rate tests well and the ROC curve below looks good too. The area under the curve actually has a great statistic at 0.9631.

Chris Pelkey
PREDICT 450-55
Solo 3



First Random Forest

After a logistic model was run, the first attempt at a random forest model was prepared. This model is interesting and unique because while random forest models are typically great techniques to model data, they frequently over fit the data and they work as a black box. This means that these models tend to follow very closely the presented data, but that they do too well at that and these models aren't good for predicting other datasets. It is a black box, because you do not get access to what the models themselves look like, but rather can only predict future models by applying them with no knowledge of how the model is working.

The first random forest model I ran involved many of the same variables as the logistic model. It was called in the following way:

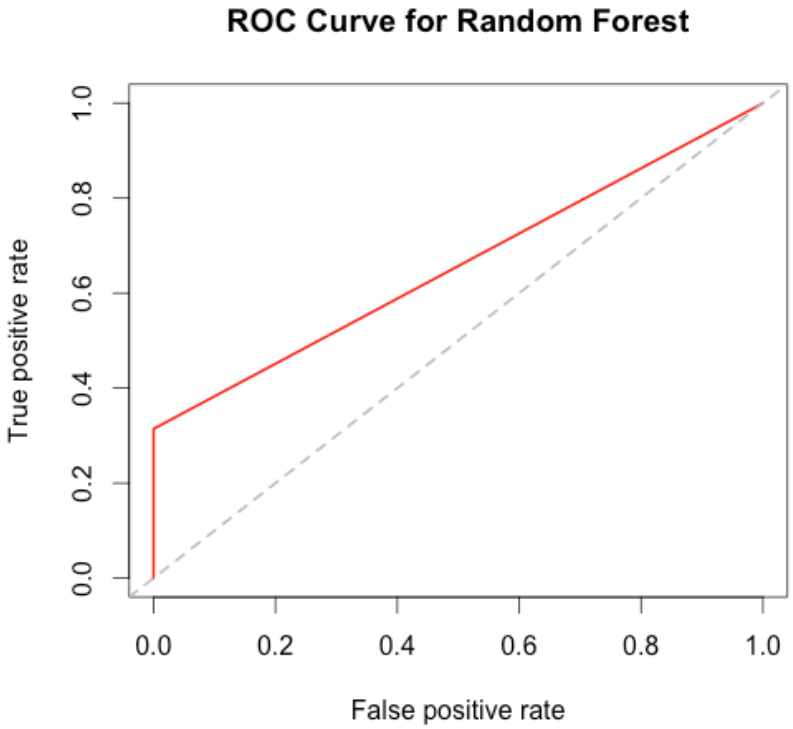
```
randomForest(FRESPONSE16 ~ ZIP + PRE2009SALES +
PRE2009TRANSACTIONS + cum15QTY + MED_INC + cum15TOTAMT +
SUM_MAIL_16 + TOTAL_MAIL_16 + TOTAMT15 + salepercamp +
ANY_MAIL_16 + salepertrans +  NAT_INC + M_HH_LEVEL + LTD_SALES +
HOMEOWNR + LTD_TRANSACTIONS + BUYER_STATUS + EXAGE,
data=subdat3, importance=TRUE, ntree=100)
```

This model did not do as well as it should have for being as complex as it is. Producing a confusion matrix on the run model with the actual response rate we see

that we get a 91.4% accuracy rate. This is a worse model because we lose out on more people that we could be making more money from.

	0	1
0	7417	0
1	727	333

The area under the curve has a lackluster statistic at 0.6571.



Second Random Forest

After a logistic model was run, the first attempt at a random forest model was prepared. This model is interesting and unique because while random forest models are typically great techniques to model data, they frequently over fit the data and they work as a black box. This means that these models tend to follow very closely the presented data, but that they do too well at that and these models aren't good for predicting other datasets. It is a black box, because you do not get access to what the models themselves look like, but rather can only predict future models by applying them with no knowledge of how the model is working.

Chris Pelkey
PREDICT 450-55
Solo 3

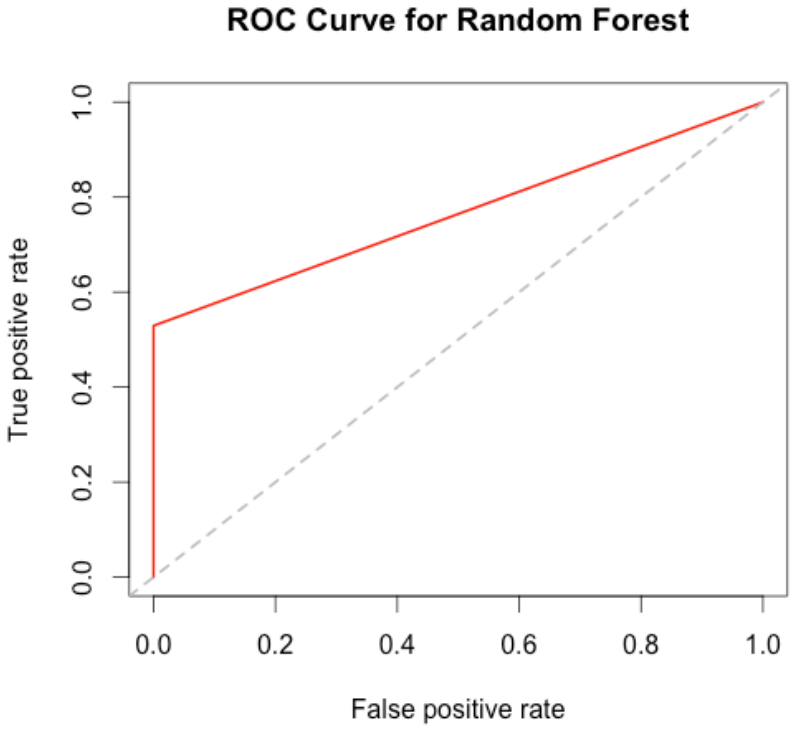
The second random forest model I ran involved many of the same variables as the first random forest model. Most specifically it removed the location data from the model and it did much better. It was called in the following way:

randomForest(FRESPONSE16 ~ M_HH_LEVEL + PRE2009SALES + PRE2009TRANSACTIONS + cum15QTY + MED_INC + cum15TOTAMT + TOTAL_MAIL_16 + TOTAMT15 + salepercamp + ANY_MAIL_16 + salepertrans + NAT_INC + LTD_SALES + LTD_TRANSACTIONS + BUYER_STATUS + EXAGE, data=subdat3, importance=TRUE, ntree=100)

This model did much better than the first random forest model did. Producing a confusion matrix on the run model with the actual response rate we see that we get a 94.1% accuracy rate, which is better than the logistic model.

	0	1
0	7417	0
1	499	561

This model produces some really interesting insights that we will examine more fully in the financial analysis section, but the 94.1% accuracy rate tests well and the ROC curve below looks good too. The area under the curve has a lackluster statistic at 0.7646.



Chris Pelkey
PREDICT 450-55
Solo 3

Financial Analysis

As can be seen in the financial analysis table below, the logistic regression really worked the best on this data. Looking at various cut off points, those people at above the 50% mark of whether they will respond appear to have the most maximization. As the cutoff raises the numbers begin to decline, until we hit a cut off of 85% likelihood of responding, and we see the profit jump back up again to $7,966. This matches exactly the best performing model of the random forest models run. The logistic model created with a 50% likelihood cut off produces the best profit maximization for this data of the models run and brings in an additional nearly $13,000.

	Logistic (.5)	Logistic (.6)	Logistic (.75)	Logistic (.85)	RF1	RF2
Direct mail unit cost	$1.00	$1.00	$1.00	$1.00	$1.00	$1.00
Without RFM Targeting						
Sample Size (All Customers Get Direct Mailing)	14,992	14,992	14,992	14,992	14,992	14,992
Customers Responding	1,060	1,060	1,060	1,060	1,060	1,060
Average Revenue per Customer	22.46	22.46	22.46	22.46	22.46	22.46
Direct mail cost per Customer						
Ave. Revenue Minus Mail Cost per Customer	22.46	22.46	22.46	22.46	22.46	22.46
Cost per mailer	14992	14992	14992	14992	14992	14992
Total Revenue Minus Mail Cost without Targeting	$8,816	$8,816	$8,816	$8,816	$8,816	$8,816
With RFM Targeting						
Targeted Customers						
Number of Customers Targeted	730	609	428	561	333	561
Average Revenue per Customer	30.86	31.90	33.09	30.98	28.82	30.98
Direct mail cost per Customer	1.00	1.00	1.00	1.00	1.00	1.00
Ave. Revenue Minus Mail Cost per Customer	29.86	30.90	32.09	29.98	27.82	29.98
Revenue Minus Mail Cost from Targeted Customers	$21,798	$18,818	$13,735	$16,819	$9,264	$16,819
Total Revenue Minus Mail Cost with Targeting	$21,798	$18,818	$13,735	$16,819	$9,264	$16,819
Profit Contribution/Lift of RFM Targeting	$12,982	$10,003	$4,919	$8,003	$448	$8,003
Per Customer Profit Contribution/Lift	$0.87	$0.67	$0.33	$0.53	$0.03	$0.53
Number of Customers in Database	14,922	14,922	14,922	14,922	14,922	14,922
Estimated Profit Contribution/Lift of Targeting	**$12,922**	$9,956	$4,896	$7,966	$446	$7,966

Chris Pelkey
PREDICT 450-55
Solo 3

Limitations

The results of this modeling of course have limitations to them. To begin, the modeling process uses a random forest model in part, which does not provide readily available access to what the model is, it acts like a black box. This means that outside of the models goodness of fit statistics, we are unable to see how the model itself was built, and we are therefore unable to verify whether this model makes logical sense. This is also based solely off of the 16th campaign of the mass mailing. While there is plenty of data available, we have no information on whether these trends are temporal and only because of the timing of the 16th campaign, or whether they will hold true for all future campaigns. The best recommendation to solve this would be to refresh the model on the 17th campaign as well simply to verify the model works correctly.

Next Steps

The logistic model is clearly the winner of the models created. This most likely makes sense because of the type of data present; many of these variables that were used had too many factor levels for the random forest to accurately predict them. Moving forward XYZ should prepare data in the same way that it captured data for campaign 16. This will allow them to use the predictive model gained from the logistic regression to score future customers to decide whether they are likely to respond to future campaigns.