STAT 6021 Final Project Report
Predicting Repeated Buyers

Date:12/7/2015

Team10
Adam Jiang
Sharath Pingula
Antoine Rigaut
Ruisi Xiong

Executive Summary

Consumer brands typically send coupons to their customers as part of an acquisition and conversion strategy. Their goal is to convert their customers into loyal purchasers. Being able to identify and predict with higher accuracy the repeat buyers that an offer will incentivize is beneficial for improving business targeting strategy and more importantly, boosting sales. Our goal is to use customer purchasing history prior to the offer date and the characteristics of the offer for constructing a logistic regression model that identifies the customers who will make a repeat purchase after receiving an offer.

Data sizing and feature engineering allowed us to profile the targeted customers based on their purchase history. For this project, we used three data sets: ***transaction history data***, ***offer details***, and ***offer history***. As in most transaction data sets, the ***transaction history*** file was massive; it contained 350 million rows of observations for a weight of 22 GB. Each observation corresponded to one product transaction. After data reduction by filtering transactions of only those customers that have purchased a product in the same category or company or brand as the coupon offered, we were left with 27 million rows of customer transaction history. The ***offer history*** data set contained details about the customers who were targeted by the coupon campaign, including our dependent variable - their purchasing response to those offers. One customer in four who received a coupon made a repeat trip after the offer.

We engineered over 50 features based on customer purchase history. Our main objective was to capture their overall purchase behavior and how actively they purchased products from the same category, company, department or brand as the offers.

We started with an Area Under Curve (AUC) baseline value of 0.58 on our initial logistic regression model. Since then, we have improved our model to an AUC value of 0.71 by adding new variables and employing various diagnostics such addressing outliers, applying transformations and reducing multicollinearity through variable selection. Our final model retained 23 features. We could understand the effect of those features on future chance of repeat purchase by studying the model coefficients.

We found that total retail expenditures for each customer, the amounts spent in the same category or company as the offer offered, the department of the offer, the type of retail chain of the offer were the most important features. In addition to predicting repeat customers, our model can be used in understanding what type of purchase behaviors would influence a customer for a repeat purchase. Our study could help retail chains or coupon distribution companies in targeting their customers more precisely while reducing their marketing expenditures and distribution costs.

A. Problem Definition

Our objective was to identify the buyers who would make a repeat purchase after receiving an incentive (such as a coupon) based on their past purchase history and the characteristics of the offer.

Consumer brands typically offer coupons to their customers as part of an acquisition and conversion strategy. Their goal is to convert their customers into loyal purchasers. It would be helpful for them to identify and predict the customers for whom the offer would trigger a repeat purchase, in order to improve their targeting strategy. Our task was to identify these customers who would make a repeat purchase after receiving an offer, particularly by analyzing the recency, frequency and monetary value of customer purchasing behavior.

In this project, we were given at least one year of customer purchasing history prior to the offer date as well as detailed information regarding the offers. The training set we used contained information about customers who received an offer (as part of an acquisition campaign which lasted from March to April 2013) and whether or not they made a repeat purchase following their offer date.

B. Description of the Data

For this project, we were provided with 3 data sets: transaction history, offers detail, and offer history.

- **Transaction history** contained information on customer transaction. Each customer had multiple transactions, with each row of transaction describing an individual purchase of one particular product. Example features included the customer's ID, product's brand, company, category, size, quantity purchased, and price.
- **Offers detail** contained information about the 37 offers offered to various customers. Example features included offer ID, product's category, company, product's brand, dollar value of the offer, and retail chain.
- **Offer history**. This data set described a customer's behaviour when previously offered a coupon. Some example features includes customer ID, offer ID, geographical region ID, offer date, number of times the customer made the purchase after offer, and finally, our response variable, repeater. 1 signifies that this customer has used the coupon and made a purchase when offered, 0 signifies that this customer did not use the offered coupon.

The following plot shows how effective each offer was in attracting repeat buyers:

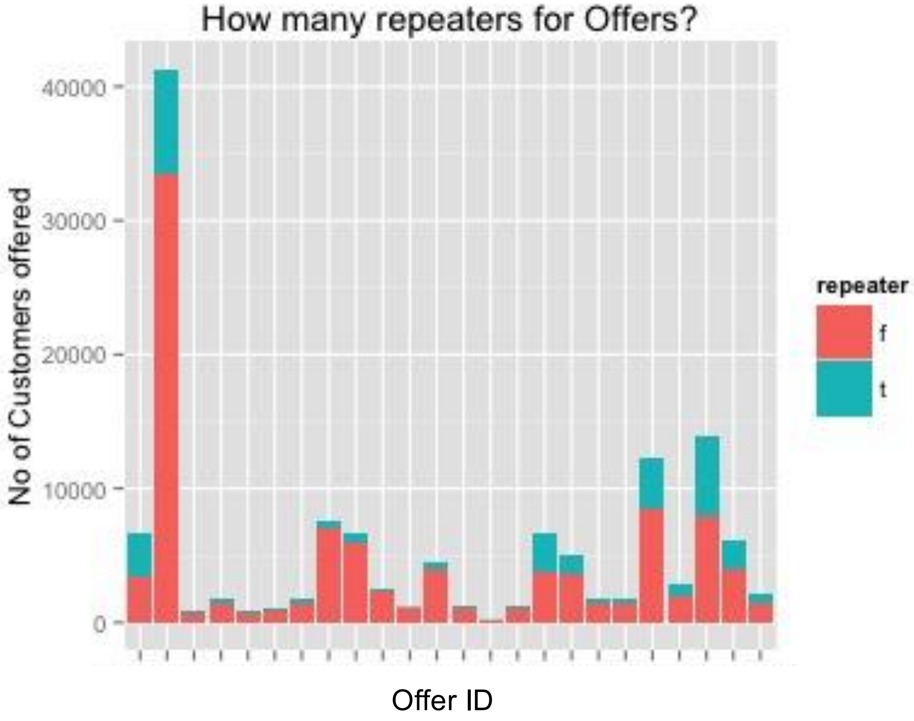


Figure 1. Proportion of repeaters for each type of offers (37 types)

The distribution of log of total spend values are displayed in the boxplot below, indicating that there are a large number of extreme values.

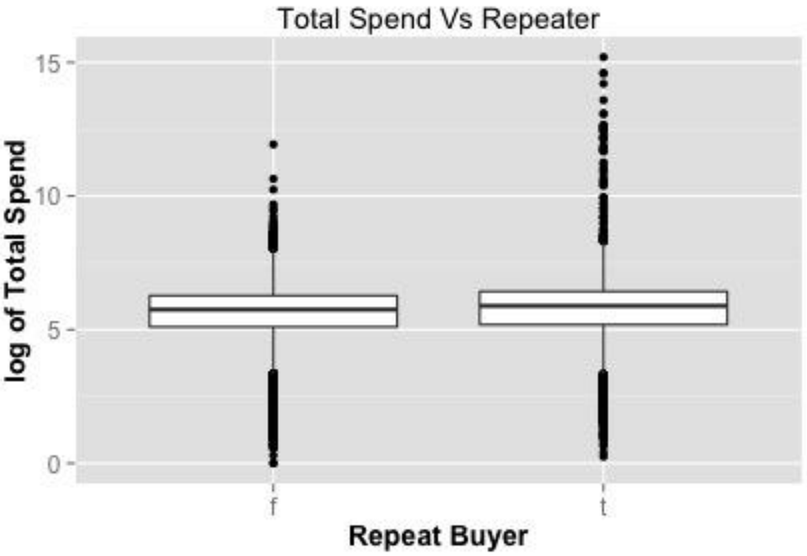


Figure 2. Log of total retail expenditures per customer in the year of transaction history for non-repeater (left) vs. repeaters (right)

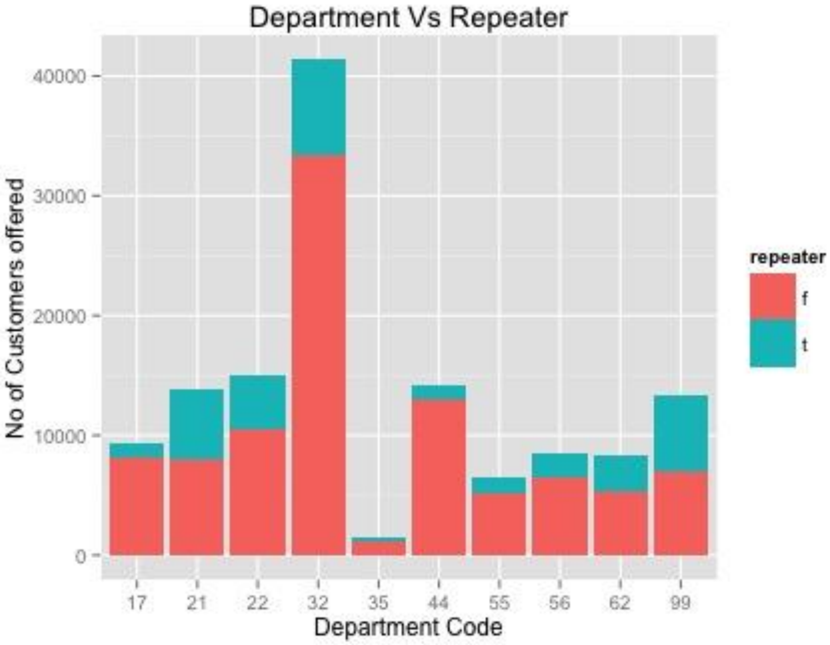


Figure 3. Repeat buyers across the different department of the offers. It appears that offers are successful in some department but not others.

C. Data Cleaning

We were interested in this project for two reasons:

The problem is a classic in data science. Often time in industry, we need to build a predictive model predicting which customer would most likely respond to a incentive, whether it be an coupon, email, snail mail, or website banner ad. These kinds of models can provide huge value to companies across any industry by allowing companies to more accurately target advertisements to customers who are more likely to purchase, thus increasing sales and lowering advertisement expenses. We believe that this project would allow us to get hands-on experience with this type of predictive modeling using real world data. Through it, we learned how to use feature engineering to process the data and consolidate customer transaction histories in our final data set.

The original data set was large. After decompressing, the transaction data was as large as 22GB and contained 349,655,789 rows of observations. This was enticing to us because we wanted to come up with multiple ways to work with these large data sets. We shortly realized that data of this size will simply not be workable with our devices, so we had to be creative and come up with ways to reduce the data size in a way that makes sense to our problem. There was a substantial number of transactions which were not related to any of the coupons that we were interested in. For example, there could hypothetically be rows of transaction data on Acqua Panna bottled water, however, none of the coupon offers would be of that category (bottled water), or that brand (Acqua Panna) or of that company (Nestle). We assumed that while these transactions do have value, their impact would be minimal and it would make sense to remove them for the purpose of data reduction. After removing these transactions, and only keeping transactions that related to either category or company of the coupons, we managed to reduce the data to a more manageable size of 1.6 GB, or about 27 million lines.

D. Feature Engineering

Feature engineering was one of the main components of this project. We were working with three separate data sets, with the transaction data set containing up to thousands of rows of data for just one customer. It was not possible to build a model by simply merging the data sets, or using any of the existing variables as predictors.

We generated 52 predictors from the transaction history dataset in order to capture patterns of recency, frequency and monetary value of purchasing behaviors. Our new features were based on how many times, in which quantity and which dollar value did a customer buy from the category, company or brand related to the coupon offered. Many of our features took recency into account, meaning that we had the same feature but for different periods, such as the 30 days, 60 days, 90 days, and 180 days before the coupon was offered. A more descriptive list of variables is shown in Appendix 1.

The monetary value of the coupon was also used as a feature. We made the hypothesis that the "department" of the offer was significantly related to the decision of the customer to make a repeat trip. A customer will react differently depending on whether the offer is about foodstuff or electronics. We hence used the department of the offer as a categorical predictor of about 10 levels.

We have clustered retail chains into high, medium and low based on number of coupons in offer from them and used it as one of the predictors. It proved to significant in predicting the probability of repeat purchase.

E. Dealing with Outliers and Data Quality

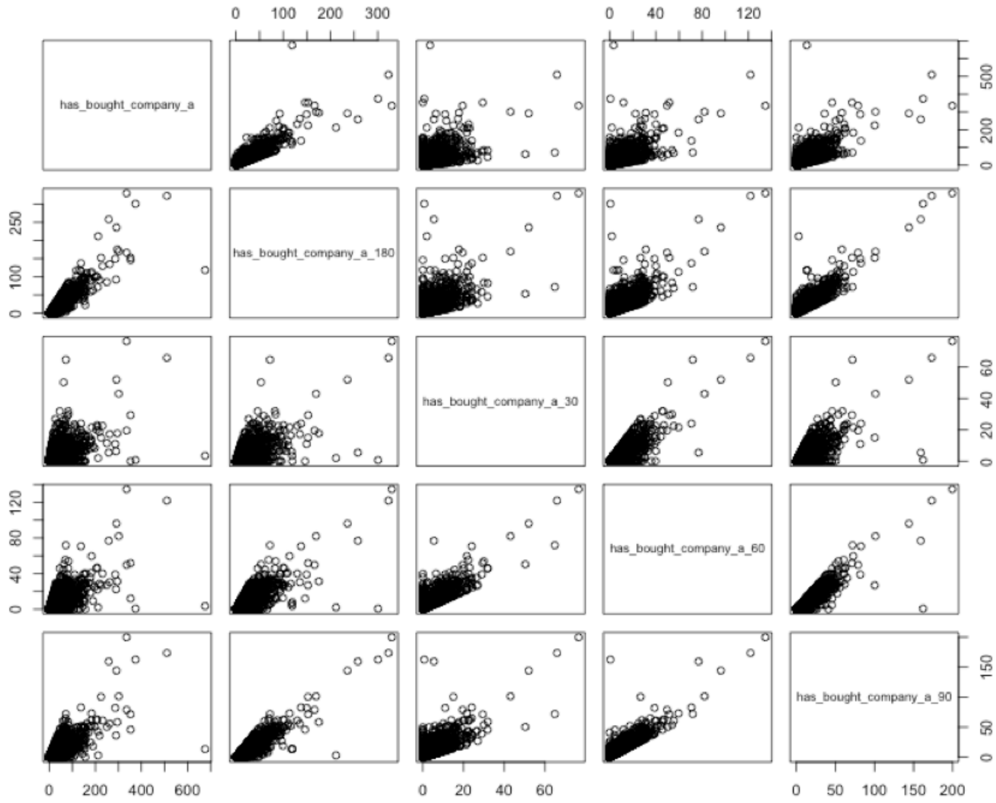
Once we had engineered our features of interest, we had to deal with some data quality issues that often occur in marketing and customer analytics.

One challenge was the phenomenon of customers returning their products to the store, resulting in negative values on some of our predictors. For instance, one predictor was the total expenditures of customers on the coupon product category in the last 180 days. Hypothetically, a customer could buy a product prior to that period and then return it in the last 180 days, resulting in a negative value. Another challenge we faced was that the customers in the records have demonstrated quite different purchasing behavior, in terms of the purchasing scale. This occurred in our data in two regards:

Firstly, there were outliers. Some customers spent as much as 4 million dollars on products in the entire one-year transaction history, while the median is about 500 dollars. Moreover, the distribution of our continuous predictors was usually highly skewed towards extreme values. This had the potential to create distortions in our predictor estimate, with obvious problems for inference.

A log transformation of our continuous predictor was our preferred solution, but this meant that we had to deal with negative values as a first step. After finding out that these negative values occurred in only 300 customers (out of 150,000 observations), we simply dropped them.

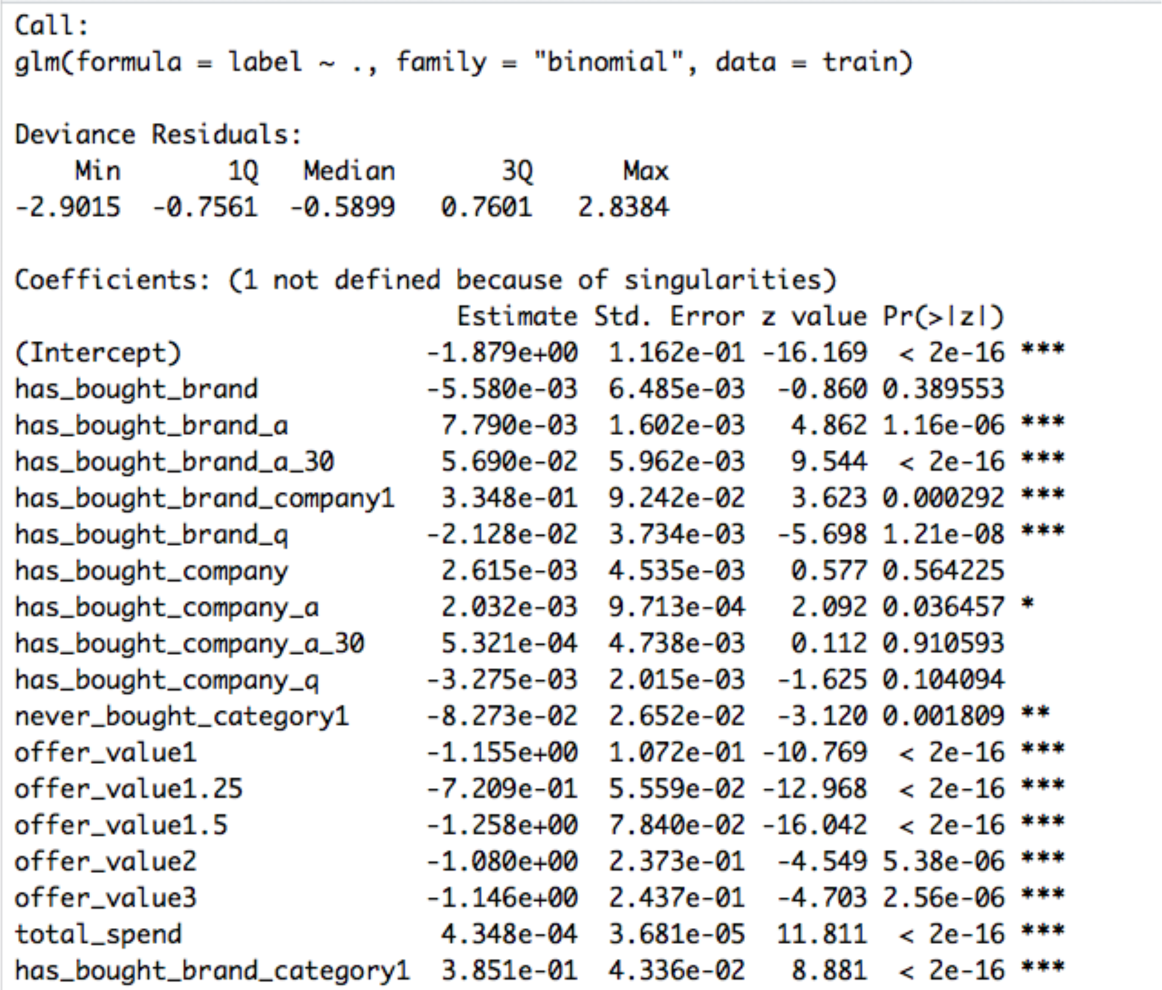
We also realized that some of the new features were collinear. The following pair plots show that in most of the cases 30, 60, 90, 180 days and total days, features are correlated. We have used *regsubsets* and heuristically chosen features to mitigate the effect of collinearity on the regression coefficients. In the final model, we have chosen the features that depict the ovarall purchase behavior and the features that capture the purchase behaviors in most recent 30 days.



F. Model

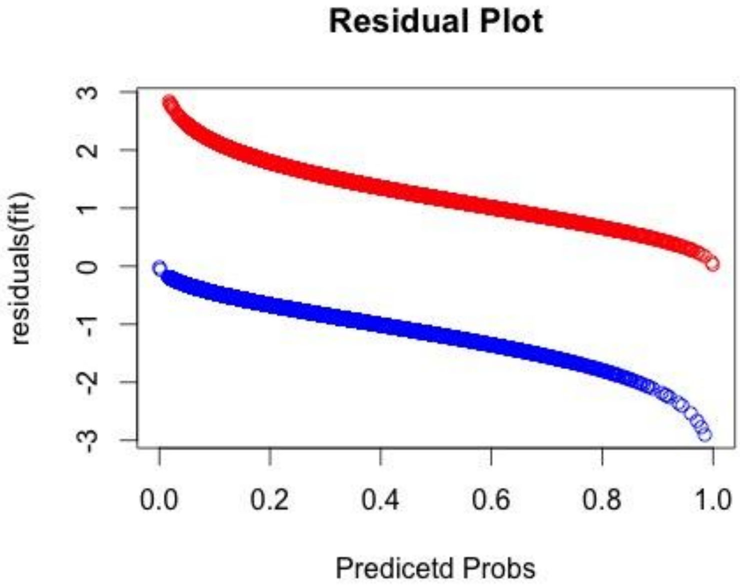
We modeled the “repeater” outcome through a logistic regression model.
Following is the snapshot of our model summary. We can interpret impact of various features on log of odds of repeat purchase from their corresponding coefficient values. For example total_spend value, one of the significant predictors has coefficient value of 4.348e-04. It can be interpreted as an increase of total_spend by 100 dollars could improve the chance of repeat purchase by 4.5% (exp(0.043)) when all other features are held constant.

Similarly other features can also be studied for their role in explaining probability of repeat purchases.

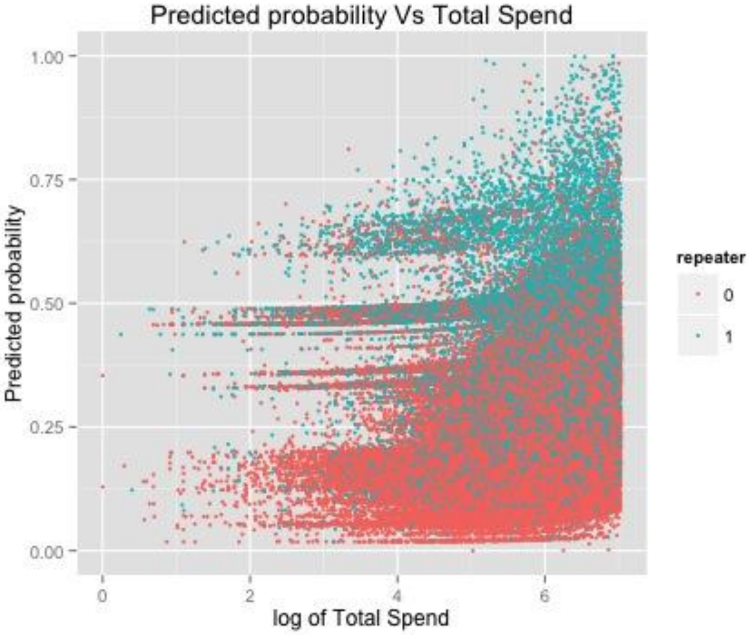


```
Call:
glm(formula = label ~ ., family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.9015  -0.7561  -0.5899   0.7601   2.8384  

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)               -1.879e+00  1.162e-01 -16.169  < 2e-16 ***
has_bought_brand          -5.580e-03  6.485e-03  -0.860 0.389553    
has_bought_brand_a         7.790e-03  1.602e-03   4.862 1.16e-06 ***
has_bought_brand_a_30      5.690e-02  5.962e-03   9.544  < 2e-16 ***
has_bought_brand_company1  3.348e-01  9.242e-02   3.623 0.000292 ***
has_bought_brand_q        -2.128e-02  3.734e-03  -5.698 1.21e-08 ***
has_bought_company         2.615e-03  4.535e-03   0.577 0.564225    
has_bought_company_a       2.032e-03  9.713e-04   2.092 0.036457 *  
has_bought_company_a_30    5.321e-04  4.738e-03   0.112 0.910593    
has_bought_company_q      -3.275e-03  2.015e-03  -1.625 0.104094    
never_bought_category1    -8.273e-02  2.652e-02  -3.120 0.001809 ** 
offer_value1              -1.155e+00  1.072e-01 -10.769  < 2e-16 ***
offer_value1.25           -7.209e-01  5.559e-02 -12.968  < 2e-16 ***
offer_value1.5            -1.258e+00  7.840e-02 -16.042  < 2e-16 ***
offer_value2              -1.080e+00  2.373e-01  -4.549 5.38e-06 ***
offer_value3              -1.146e+00  2.437e-01  -4.703 2.56e-06 ***
total_spend                4.348e-04  3.681e-05  11.811  < 2e-16 ***
has_bought_brand_category1 3.851e-01  4.336e-02   8.881  < 2e-16 ***
```

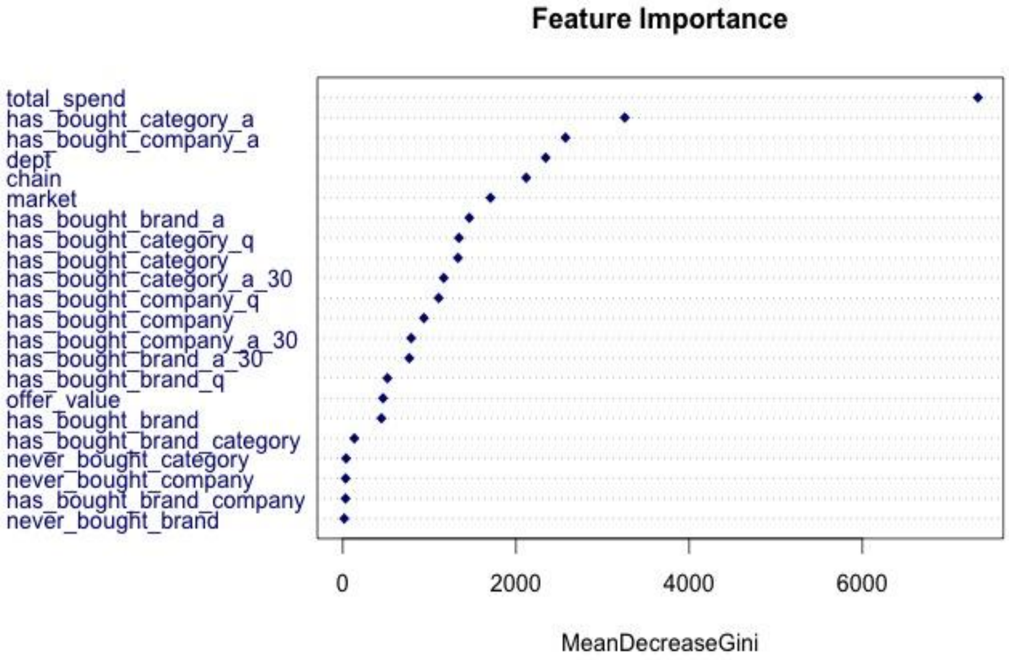
Initially our logistic regression model had convergence issues. It was mainly due to highly collinear features and extreme values. After employing various model diagnostics as explained above our regression achieved convergence. Following plot shows the residual against predicted probabilities. It was a lot of improvement from the initial models, which had issues of convergence.



The following plot shows predicted probabilities of our model against log of total_spent. Predicted probabilities are more aligned with the actual labels in the customers with either very large total_spend values or very small total_spend values.

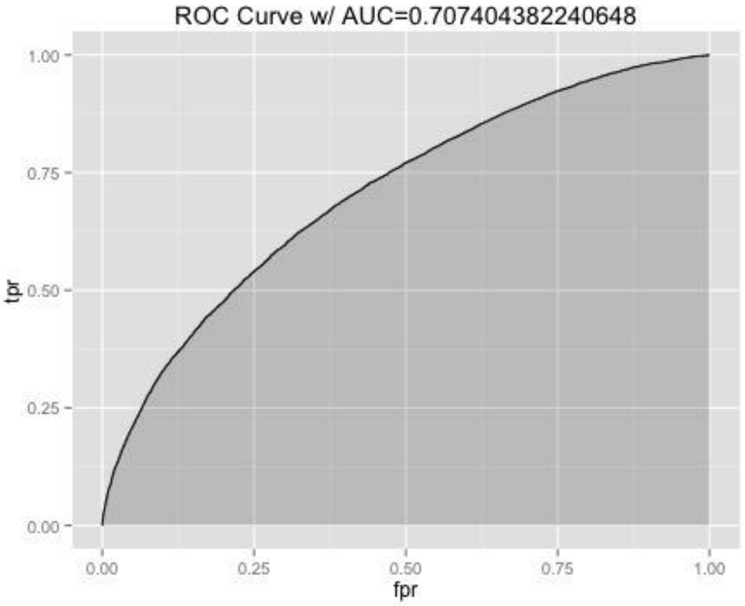


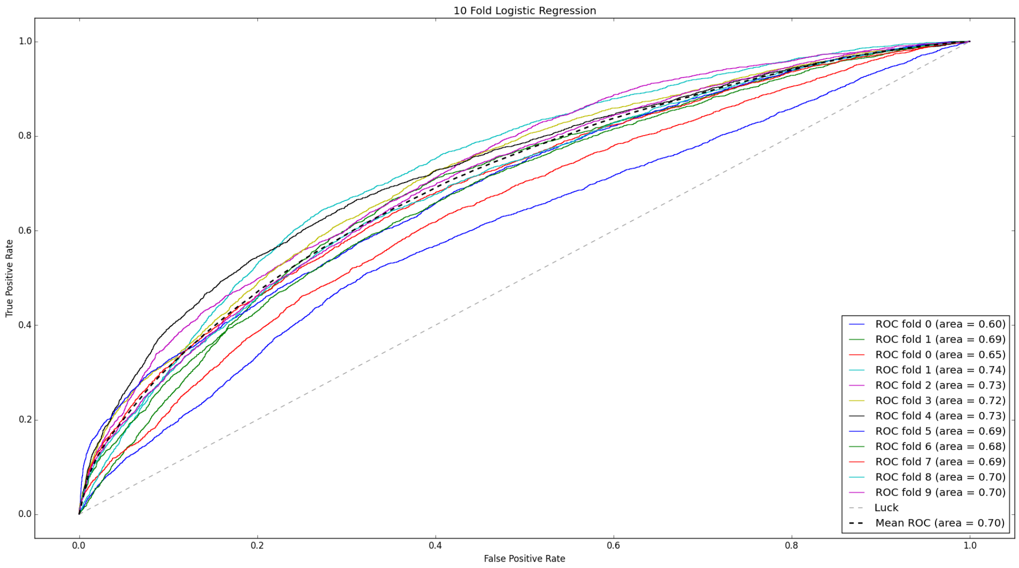
The following feature importance plot was generated through random forest model using ‘randomForest’ package in R. The features generated through feature engineering have played a very important role in our models.



G. Model Evaluation

The data is divided into train (70%) and test (30%) sets. We have trained our model on train set and validated on test for evaluating it effectiveness. We have achieved auc value of 0.707 on the test set in our final model. We have 22 features in the final model after applying transformations and addressing collinearity issues and extreme values etc.





We have evaluated the model with 10 fold cross validation to test over fitting scenarios.

H. Regularization

Given the large number of variables in our model, we resorted to lasso regularization to decrease our model’s variance. The penalty term was optimized using standard statistical packages. Lasso has the added benefit of performing variable selection.

I. Conclusion

We have found that attributes such as total value spent by a customer, amount spent in the category or company same as the coupon, department (Ex: food, electronics etc.) of the coupon, retail chain that is offering the coupon were strongly associated with the probability of repeat purchase. Coupon distribution companies if considered above factors could potentially increase repeat purchases, sales volume while reducing their marketing expenditure and distribution costs.

J. Appendices

1. The full list of predictors we have derived for feature engineering

Based on the predictor has_bought_brand, we prepared predictors called has_bought_brand_30, has_bought_brand_60...etc. Such predictors include:

- has_bought_brand: The number of times the customer has bought a product from the same brand as the coupon offered before receiving the coupon.

- has_bought_brand_a: The dollar amount the customer has spent from buying a product from the same brand as the coupon offered before receiving the coupon.
- has_bought_brand_q: The quantity of product the customer has purchased from the same brand as the coupon offered before receiving the coupon.
- has_bought_company: The number of times the customer has bought a product from the same company as the coupon offered before receiving the coupon.
- has_bought_company_a: The dollar amount the customer has spent from buying a product from the same company as the coupon offered before receiving the coupon.
- has_bought_company_q: The quantity of product the customer has purchased from the same company as the coupon offered before receiving the coupon.
- has_bought_category: The number of times the customer has bought a product from the same category as the coupon offered before receiving the coupon.
- has_bought_category_a: The dollar amount the customer has spent from buying a product from the same category as the coupon offered before receiving the coupon.
- has_bought_category_q: The quantity of product the customer has purchased from the same category as the coupon offered before receiving the coupon.

Predictors that do not have the recency factor include:

- never_bought_category: If the customer has never purchased any product that is in the same category as the offered coupon, label 1, otherwise 0.
- never_bought_brand: If the customer has never purchased any product that is in the same brand as the offered coupon, label 1, otherwise 0.
- never_bought_company: If the customer has never purchased any product that is in the same company as the offered coupon, label 1, otherwise 0.
- offer_quantiy: Number of coupon offer.
- offer_value: The discount amount of the coupon.
- total_spent: The total dollar amount of money this customer has spent in his/her transaction history.
- has_bought_brand_category: If the customer has purchased a product that is the exact same combination of brand and category as the coupon offered, label 1, 0 otherwise.
- has_bought_brand_company_category: If the customer has purchased a product that is the exact same combination of brand, company, and category as the coupon offered, label 1, 0 otherwise