**SFU Business Analytics Hackathon 2017**

**Case: Customer Retention at Grocery Consolidators**

Grocery Consolidators started in Charlottesville and Richmond, Virginia in 2010, and has expanded successfully since then. The company takes orders from customers for products from a number of different small speciality retailers. Customers place orders online for products before midnight. Orders are downloaded after midnight and sent to the respective retailers. The retailers bag the orders and the next morning, Grocery Consolidators sends a truck to pick up the orders and bring them to the warehouse, where they are sorted into packages for each customer. Trucks are then sent to central locations where customers can pick up their combined orders. This greatly reduces the “last mile” delivery cost. For an extra fee, deliveries can be made to the customer’s doorstep, but that is not often used.

Since this is not a subscription service (like wireless or cable companies), lost customers cannot be identified by cancelled subscriptions. In that case, a decision to consider a customer as lost must be made on the basis of how long it has been since the last order. A fixed interval is problematic, as there is a lot of variation in how frequently customers order. The company has decided to calculate the standard deviation of each customer’s order interval, and if they have not ordered for more than two standard deviations of the order interval, they are considered lost, and the “lost” variable set to “Y”. If not, they are considered retained, and the “lost” variable is set to “N”.

**Your task is to determine if lost customers can be predicted using the data available in the order and email contact database.**

**MODEL SCORING:**

The data has been pre-sampled to have an Estimation, a Validation, and a Holdout sample using the “Sample” variable, with a 50/25/25 split. This is done to facilitate your own modeling on a subset of the data set (the Estimation set) and testing how well it does with a known target on another subset (the Validation set). However, you may ignore this distinction and use the records with the known target however you like. Holdout records are the test set for the competition.

Any records that are in the holdout sample have the “lost” variable missing (NA), e.g.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **custid** | **Pred1** | **Pred2** | **Sample** | **lost** |
|  | 1 | A | Estimation | N |
|  | 2 | A | Validation | Y |
|  | 3 | B | Validation | N |
|  | 3 | C | Holdout | NA |
|  | 5 | D | Estimation | N |
|  | 1 | A | Holdout | NA |
|  | 2 | B | Validation | Y |
|  | 3 | B | Validation | N |
|  | 6 | C | Estimation | N |
|  | 5 | C | Holdout | NA |
|  | 2 | A | Estimation | Y |
|  | 1 | D | Holdout | NA |

Having all samples in the same file allows for simultaneous data cleaning and modification of all data.

You will need to add a new variable to the data which must be labeled “**Score**” (case-sensitive), and which must contain your pr*edicted probability*, from your best model, of a customer being lost.

Use your best model to score the entire data set. The scores on the Holdout sample only will be compared with the true lost value available to the judges. The models will be judged on the cumulative captured lift chart, using only the percent captured at the 40% sample proportion level.

Save the scored dataset as a .csv file. For example, in R:

write.csv(Dataset, “C:/yourdirectories/…/yourteamname.csv”)

Your scored dataset can be submitted at any time to http://badm.ca/scoring. The file must contain at least two variables, one named “custid” and one named “Score”. Your CSV file must have the column names in the first row.

**FINAL SUBMISSION:**

**Final scored dataset is due at 2:15 p.m uploaded at http://badm.ca/scoring**  
When you upload your file, the system will give you a confirmation message which says “Your score has been submitted”. If you receive an error message please check your file and ensure it has the correct column headings and csv format.

**DATA DICTIONARY** for dataset CUSTOMERETENTION.csv  
Variable Description  
Custid Computer generated customer identification  
lost Y if customer is assumed lost; N if assumed retained; NA in the Holdout Sample  
created Date when the contact was first created in the database-when customer joined  
firstorder Date when customer placed first order  
lastorder Date when customer placed last order  
esent number of emails sent (see note below)  
eopenrate number emails opened divided by number of emails sent  
eclickrate number of email links clicked divided by number of emails sent  
avgorder average order size for the customer  
ordfreq number of orders divided by lastorder – firstorder in days ; 0 if only one order  
paperless Y if customer subscfribed for paperless communicaton   
refill Y if customer subscribed for automatic refill  
doorstep Y if customer subscribed for doorstep delivery  
vflag a database bookkeeping variable  
favday Customer’s favorite delivery day  
city city where customer resides  
Sample Estimation, Validation, Holdout … 50/25/25

**Notes:**

--There is no oversampling of this data

--Missing Variable indicator: “NA”

--30781 records  
--**Do not use “esent” as a predictor**. Be prepared to explain why that would be a bad idea.   
Also be aware that, because the target variable is developed from the database, it may be possible to compute new predictors from the dataset that do well on the holdout sample, but are not useful for future prediction.

**DATE MANIPULATION HELP:**

The dataset contains dates that may be useful. Here are some quick guidelines on date manipulation relevant to these date variables, in R. First convert the character variables to date variables. For example for the character variable “created” in the dataset “Dataset”

Dataset$created.date <- as.Date(Dataset$created, format = "%d/%m/%Y")

--Note the Y is upper case

Internally (you won’t see it) this creates an integer, with 0 on Jan 1 1970. So now you can add and subtract the date variables. For example, assuming you have converted “lastorder” and “create” to dates, the difference in days is:

Dataset$dif.days <- Dataset$lastorder.date – Dataset$created.date

This will return a “difftime” variable, in days.

To create a new variable that Is the maximum (or minimum – use “min”) of a date variable:

Dataset$maxcrt.date <- max(Dataset$create.date, na.rm = TRUE)

Models don’t like date and difftime variables, so if you want to use, for example, differences in dates in a model, convert the difference in days to a numeric variable:

Dataset$dif.num <- as.numeric(Dataset$dif.days)

You may have to Refresh the dataset to view the new variables.

**MANAGERIAL PRESENTATION and POSSIBLE ADDITIONAL ANALYSIS:**

Starting at 2:15 p.m (after you have submitted your scored data). you will begin preparing a five-minute powerpoint presentation that answers the questions “so what?” and “what next?” - you will have approximately 45min to work on your presentations. All **presentations will need to be emailed to** [**info@badm.ca**](mailto:info@badm.ca) **by 3pm at the latest**. \*Include your team number as the email subject line.

Management will be interested in the value of lost customers. A high probability of losing a customer that spends little would not justify much investment in retention. Similarly, retention efforts in a low-loss-probability customer would be wasted. One could estimate expected revenue losses for your customers for a fixed time frame, and provide a summary of these losses and describe how to use them.

AVOID TECHNICAL JARGON.

**At 3:15pm be ready to present at your tables**. Judges will come to your table and you will have 5minutes maximum to present from your desk using your laptop. Since you might be visited by 2 judges, separately, you need to stay at your table between 3:15 and 3:45 so be ready at all times. Teams will be judged on communication 30% (clarity, logic, persuasiveness), business understanding 40% (tactical and strategic considerations), and technical competence 30%. Then you will need to please rearrange the desks to how they were at the start of the day and return to the main presentation room by 3:55 for the finalist announcement and final round. At 4pm the 6 finalist teams will be announced – each of the finalist teams will then take a turn to present to the entire audience and the entire judge panel for 5minutes, plus 5minutes of questions. While one team is presenting, the other finalists who haven’t presented yet will have to wait outside the room. Although your presentation should avoid technical jargon, the judge questions may include some technical questions so be ready. Then at 5pm, winners will be announced.