Predict Probability of Sales on Inbound Call Center Traffic

## Project description

This project aims at predicting the sale on a call based on caller's demographic, socio economic data, location and the source of traffic (paid, SEO or affiliate). A sale comprises sale of any combination of cable, internet or phone service to a residential caller in any part of US depending on the availability of service at the caller's address. Call center sales agents have access to order entry tool for almost all the major cable, internet and phone service providers in the country which enables them to sell any provider the caller is looking for.

## Background on the problem:

Ordering cable, internet and phone services is a complex process that requires customization and selection of many features on caller’s part and also because the price a customer pays on a monthly basis, varies with the choice of channels selection for a cable product, speed of internet package and in case of phone whether in-state or out-of-state calling is selected or not. People usually need assistance in order to order services and hence call in.

Before calling, callers generally search online to get an idea of what they should order and who to reach out to in order to get services. The search is mostly carried out on some of the popular search engines such as google, bing, yahoo etc. This particular client markets its services on these search engines to generate demand and to drive calls to the sales center.

Since sales center is cost intensive, running the business profitably depends largely on achieving operational efficiencies in the call center. One of the areas requiring optimization is maximizing the revenue generated on each call by predicting likelihood of conversion of a call and prioritizing that call over the other waiting calls in the queue.

**Introduction of the client:**

The client for this project is a provider of outsourced online customer acquisition solutions principally to large, consumer-facing organizations and to clients offering products or services to business users. A business seeking to attract customers online can do so by procuring advertising from search engines and other providers, directing consumer attention to a website or a sales center, and then selling interested consumers its products or services. Client provides outsourced solutions to businesses for some or all components of this set of online customer acquisition services. As an outsourcer, the client will procure paid search advertising, develop and manage branded web sales portals, and establish and maintain sales centers for its clients.

Client provides several benefits to corporations that seek to grow their online customer acquisition business process. First, it maintains and uses proprietary databases and algorithms that allow for cost-effective procurement of online advertising that drives customer action. Second, it has expertise in optimizing websites and sales centers for optimal levels of sales conversion. Third, it offers its solution to its clients on a fee-per-sale basis, by which clients pay it a pre-negotiated commission for each customer it procures.

Historically, the client has focused on providing customer acquisition solutions to US corporations operating in the telecommunications sector; nine of the ten largest US cable telecommunications companies (as measured by subscribers) acquire customers through the client. The client later commenced operations in Canada and Mexico, and is pursuing further growth opportunities in Latin America, Europe, and Asia Pacific. It has also recently started providing its customer acquisition services to the electric utility industry and renewable energy sector, and is examining additional opportunities in the insurance, retail banking, and consumer technology sectors.

## Call center operations:

Since the client markets a large number of telco providers, the call center is set up to have multiple skills. Each skill has call center agents who are trained to be expert in selling a particular brand of product and services. There are over 500+ agents selling 20+ brands of telco providers. Calls are driven to the center on free toll free numbers (TFNs) which are mapped to various skills. Once a call is answered by a sales agent, she helps the caller with order placement or otherwise if a caller has any query. Once the conversation is over, the call center agent disposes the call by selecting a disposition code. On average, the call center answers around 500K calls per month.

## Data Set:

Data on every call that hits the call center switch is recorded from the time it lands at the switch to the final outcome on the call. In order to predict the sales outcome on a call, calls data for May 2016 has been used (file name is CallsDataMay2016.csv) which has over 440K rows and 12 variables. Of the 440K calls offered, hit the switch, roughly 400K calls were answered by the agents. A summary of variables is obtained by running the summary() command on the read data.

A brief description of the variables in the data set is as follows:

* Call.Date : Captures the day of month when the call arrived. The dataset has calls answered for all 31 days in May.
* Call Start Time: The start time of the call
* Call End Time: The end of the call
* Outcome: This captures whether a call got successfully connected to an agent or not. For this analysis, all connected calls have been taken. This variable does not have any preditive capability and will be dropped.
* AgentID: Every agent in the center is assigned a unique ID captured by AgentID field.
* TFN: The toll free name dialled by the customer to reach us to order services. A TFN has 10 digits.
* BTN: Phone number of the customer
* CallSkill: Agents are divided into various skill in the call center. A skill comprises a group of agents who are expert in selling a particular group of services. Some agents are multiskilled. CallSkill is the skill of the agent who answered the call.
* OrigianlSkill: In case all agents on a skill are busy, the call is answered by the next available agent on another skill. The agent who was originally supposed to answer the call is captured by OriginalSkill variable.
* CableCompany: THere are 24 skills in the center.
* Affiliate: In addition to generating demand by marketing on search engines, some third party parnters also send traffic to the call center. They use a mix of marketing channels to drive traffic. The name of an affiliate is captured by this variable. The call center's own sources of traffic are grouped under "Digital Globe Services" to differentiate from the rest.
* Disposition: At the end of a call the call center agent classifies a call as sales/non-sales calls captured by the disposition variable. To better understand what happened on a call, disposition field helps capture the caller intent. A sale could be a sale of one service (A single play sale), two services (a double play sale), three services (a triple play sale), or four services (a quad play). The revenue generated on a call depends on the number of services sold.

## Data Cleaning and Data Wrangling

As we are interested in the calls that successfully connect to an agent, we filter out the calls that have an outcome of connect using the following command:

*calls.may <- calls.may[calls.may$Outcome=="CONNECT",]*

Running summaries of the variables, using *summary* command, reveals that the BTN variable has some alpha-numeric characters and some words. Since the prediction of call outcome as sale or no- sale depends on BTN, any row containing words or incomplete BTN should be removed. For example, some records have BTN as “Anonymous”, these records are using the following R command:

*calls.may <- calls.may[which(calls.may$BTN != "Anonymous"),]*

Similarly, the record containing useless data are removed. The number of characters in the BTN field should be 10 characters. Any BTN which more than 10 characters, either because of a leading ‘+’ or a leading 1, are truncated to 10 characters using the following R command:

*calls.may$BTN <- ifelse(nchar(calls.may$BTN) > 10,*

*substr(calls.may$BTN,*

*nchar(calls.may$BTN) - 9,*

*nchar(calls.may$BTN)), calls.may$BTN)*

Calls from TFN numbers should also be removed as they do not help with the prediction of the call. These are also removed.

*calls.may <- calls.may[-which(grepl("^855|^0|^1",calls.may$BTN)),]*

A new column “Area Code” is created from the BTN field to get the first 3 digits of the caller’s area code which helps with the location of the caller.

*calls.may <- calls.may[-which(grepl("^855|^0|^1",calls.may$BTN)),]*

After cleaning the BTN field, the next step is to clean the name of cable companies in the Cable company field. This is achieved using some of the commands mentioned below:

*# Converting all cable company names to lower case to eliminate case issues*

*calls.may$CableCompany <- str\_to\_lower(calls.may$CableCompany)*

*# Trim white spaces*

*calls.may$CableCompany <- trimws(calls.may$CableCompany, which = "both")*

And there are many more commands to fix some other issues detailed in the complete code file.

The disposition column is converted to a sale/no-sale column to create the predicted variable which will be used later in the analysis and for creating a predictive model.

*calls.may$Sale <- ifelse(grepl("Play -", calls.may$Disposition, ignore.case = T),*

*"Sale","No Sale")*

**Approach used to solve the problem:**

I have used joint probabilities of individual factors to predict the probability of sale(conversion) based on a call given by the formula:

Where Pr(C|x) is the joint probability of individual factors’ probability and n is the number of statistically significant factors used to calculate the joint probability. Statistically significant factors are determined using anova.

The generalized conditional probability formula states that:

Where

* 1. Pr(C|x) is the probability of conversion given one of the factor is present
  2. Pr(x|C) is the probability of a certain factor given conversion such as
     1. Ratio of Conversion | TFN
     2. Ratio of Conversion | Area code etc.
  3. Pr(C) is the probability of over conversion calculated as

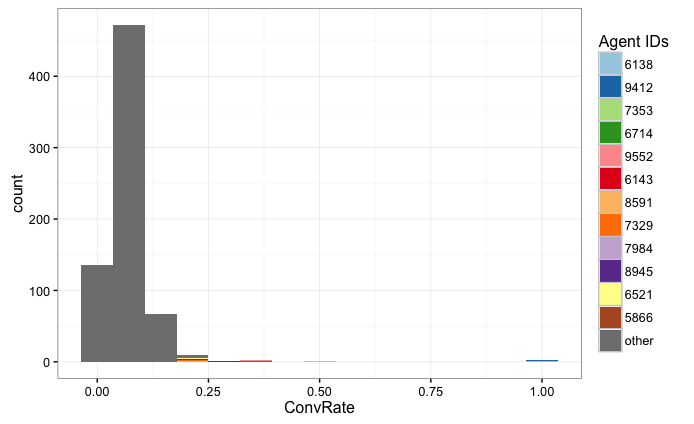
Pr(C) comes out to be 6.89%.

* 1. Pr(x) is probability of each level for each factor, calculated as follows:

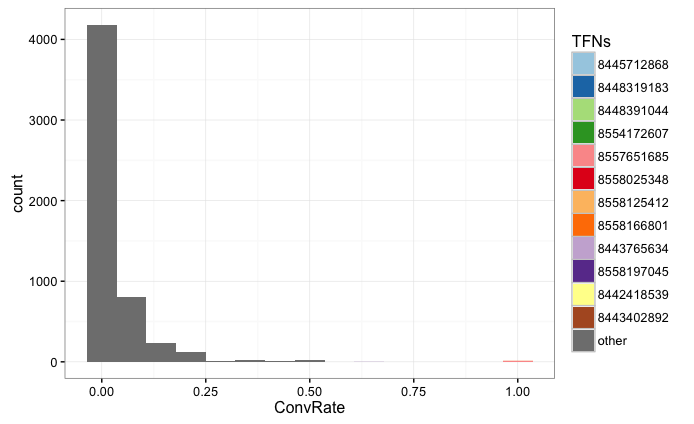
## Exploratory Data Analysis:

After the overall sales conversion ratio (Pr(C)) is calculated, a graphical analysis is performed to determine the relationship between individual factors and sales conversion.

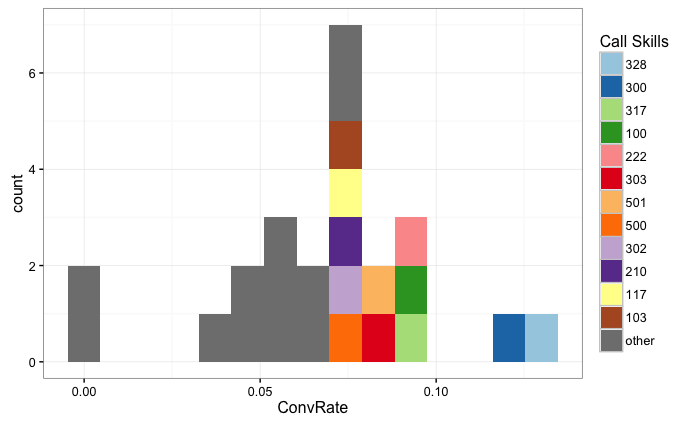
*Plot of Agent IDs Sales Conversion Ratios*



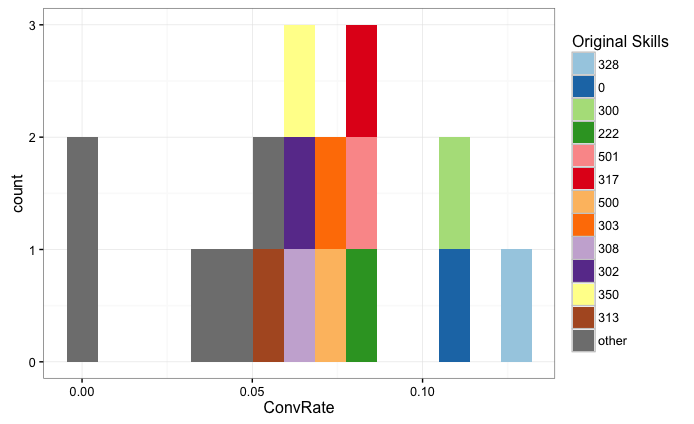
*Plot of TFN Sales Conversion Ratios*



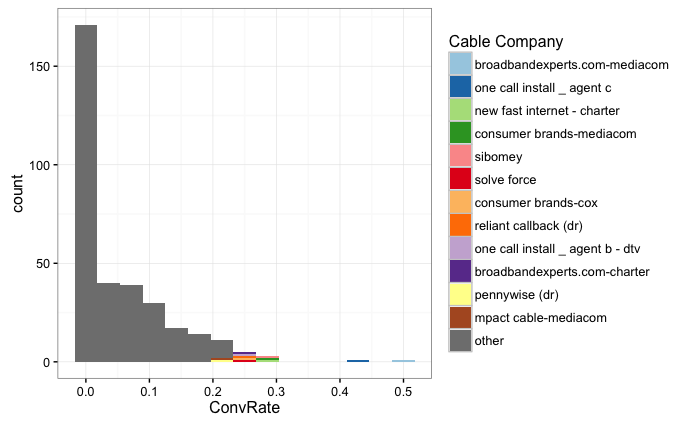
*Plot of Call Skill Sales Conversion Ratios*



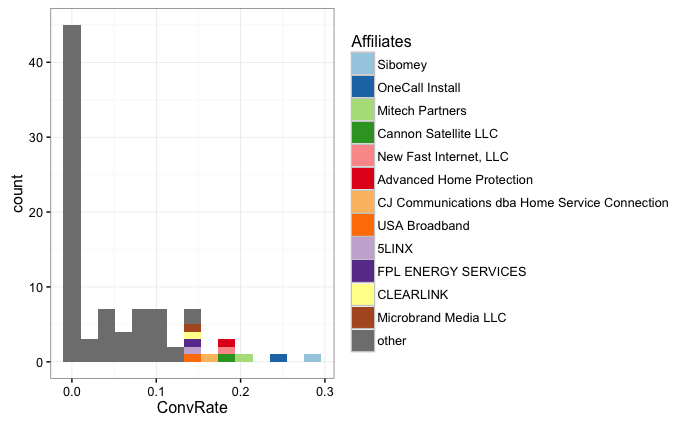
*Plot of Original Skill Sales Conversion Ratios*



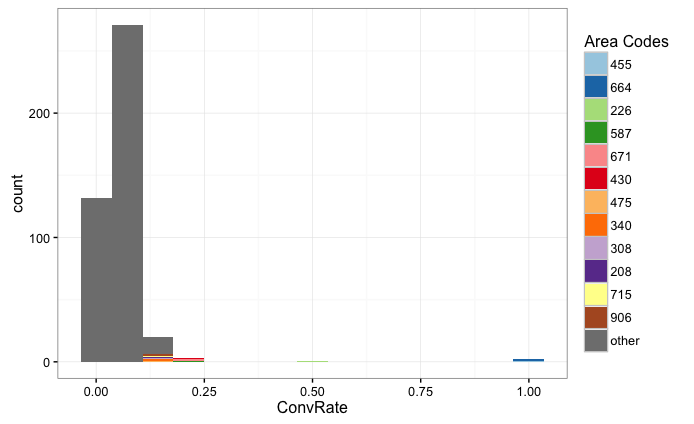
*Plot of Cable Company Conversion Ratios*



*Plot of Affiliate Conversion Ratios*



*Plot of Area Codes Sales Conversion Ratios*



**Calculate Factor Probabilities:**

We use anova test to determine which factors are significant in the data set. Since there are over 400K rows in the data, we sample the data set for 1000 rows to run the anova on by using the following command:

*n<-sample(nrow(calls.may),1000)*

*calls.maysample <- calls.may[n,]*

We start by running anova on all factors and then eliminate the factors one-by-one based on p-value. After running tests, we find that Area Code and Affiliate are statistically significant. The results of anova are as follows:



We then calculate conditional probabilities for these factors.

Lastly, we write a function to calculate the joint probabilities as a sum of probabilities calculated above.

*prob.of.sale<-function(AreaCode, Affiliate){*

*prob <- PrConv.given.AreaCodeProbOverall[PrConv.given.AreaCode$AreaCode==AreaCode]\*100+PrConv.given.Affiliate$ProbOverall[PrConv.given.Affiliate$Affiliate==Affiliate]\*100*

*paste(round(prob/2,2),"%")*

*}*

**Results and Recommendations:**

The project aimed at identifying the factors that are significant in determining the probability of sales conversion on a call. We have seen that Area Code and Affiliate are the two factors that can help predict the sales conversion. This knowledge can be used in one or all of the following ways to improve the overall sales conversion and efficiencies in the call center that would eventually result in higher revenues and margins for the business

* **For Increasing Sales**:

Based on the results, a call ranking mechanism can be developed that ranks the incoming calls in the queue, based on their probability of conversions, so the higher probability of sale call gets answered first, improving customer experience and results in more conversions. It is generally conceived that longer wait time lead to poor customer experience that impact sales

* **For Decreasing Labor Costs:**

Currently the sales center is staffed to answer every call generated from the core marketing of the client (call volume that client generates through its core marketing efforts and is roughly 70% of the total call volume) resulting in lower conversion from this channel and higher direct labor costs. Interactive voice response (IVR) can be implemented on the core volume to weed out low propensity calls and higher conversion volume could be routed directly to the call center agents. This would result in lower direct labor costs which would help in improving the overall margins of the business.

* **For Decreasing Ad Costs:**

As mentioned above, the core marketing efforts drive 70% of the total call volume which requires a ton of ad dollars spend. A reverse lookup on area codes can help determine the less or non-profitable areas where significant savings in ad dollars can be achieved resulting in direct improvement in contribution margins.

**Further Research:**

The results of this study can be used to do more analysis to include time of day, day of week and seasonality factors to further fine tune the probability calculations. This can also be extended for call volume forecasting for the call center. Most of the core marketing is done on google and bing. We can combine google and bing keywords, being bid on to generate volume, in the analysis to create a feedback loop based and adjust bids based on the profitability of a campaign.