Assignment 1 – WA – Assignment#1-AirFrance

Course – Web & Social Analytics (ITSM 6209)

Team Members – Rakesh Singh

Professor - Dr. Wenjing Duan

Date – 28-May-2021

**Ans 1:**

**Descriptive Statistics & Observations:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Item | CTR | TCR | ROA | Net Revenue | Avg. Cost per Click | Avg. Revenue per Booking | Probability of Booking. |
| Count | 4510 | 4510 | 4510 | 4510 | 4510 | 368 | 4510 |
| Maximum | 200 | 900 | 3794.87 | 549524.05 | 10 | 5877.75 | 0.82 |
| Minimum | 0 | 0 | -1 | -8725.92 | 0 | 34 | 0 |
| Mean | 11.14 | 0.57 | 3.41 | 866.21 | 1.89 | 1024.26 | 0 |
| Std | 20.23 | 13.86 | 72.78 | 14246.3 | 1.32 | 704.32 | 0.015 |

The table above depicts the depth of the data, if you look at the count of Avg. Revenue per booking it differs from others as we have not considered **NA** in count. Standard Net revenue is appearing to be positive wherein the mean is sounds low, it is probably because of the user search engine adaptability pattern. The average revenue per booking is the average amount of money AirFrance earned with each booking. The average was $1024.26, which is a high amount of revenue for each individual booking. Though the number in minimum for Net Revenue is in negative, however overall, it is observed that AirFrance advertisement campaign was a success since most of the numbers are in positive.

**Scatter Plot with trend line:**

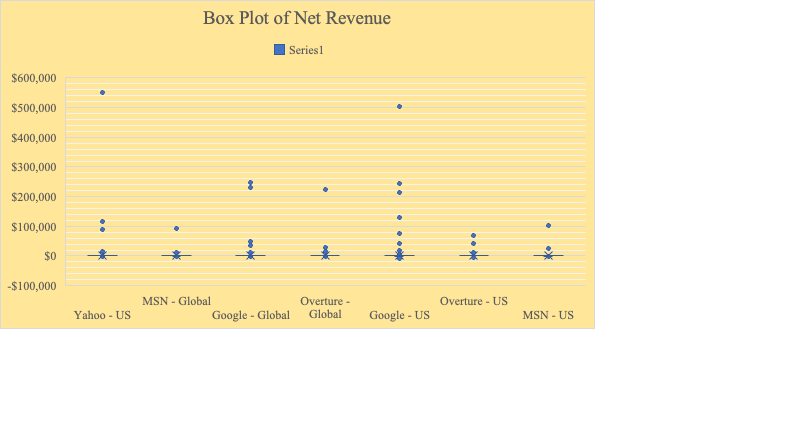
The scatter plot above establishes the relationship Net Revenue earned by publisher. The visualization clearly depicts that the revenue for some publisher went till 600K for a count of 4510 publisher. Though this doesn’t mean that all the publishers had a net revenue of approx. 600K. Looking at the trendline it is clearly visible that the number for publisher is slightly above 4500 (to be very precise 4510). There are publishers who would need to develop a strategy to ensure that the Net Revenue touches the number that can give them maximum profit and they can stay relevant to this competition.

Note – The above graph is produced with the raw data from Double Click sheet.

**Histogram for Net Revenue:**

It is visible in the histogram graph that Google – US made the maximum revenue for AirFrance and out-performed all other Publisher. Also, the contribution made in US is way higher than Global, which proves that the advertisement acceptance in USA is way more positive than the rest of the world. Looking at the Histogram it is also visible that MNS is at the bottom when it comes to the Net revenue even though they have Global existence as well.

**Box Plot for Net Revenue**



Unlike Histogram in the Box Plot above it is visible that Google – US and Yahoo – US are so close to each other (since we don’t have the numbers), when a AirFrance customer books the flight the probability of that booking is much higher for Google – US compared to rest. The differences are visible by outliers and the box plot eases the sight of that outlier. Maximum and minimum can also be clearly observed through box plot.

**Ans 2**:

**Pivot Table:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data** | | | | | | | | | | |
| **Publisher Name** | **Sum of Net Revenue** | **Sum of Click Charges** | **Sum of Clicks** | **Sum of Avg. Pos** | **Sum of Total Volume of Bookings** | **Sum of Avg. Cost/Click** | **Sum of Avg. Revenue/Booking** | **Sum of ROA.** | **Sum of Cost/Booking** | **Sum of Prob. Book** |
| Google - Global | 808,603.09 | 120,946.71 | 72,895.00 | 1.821 | 797 | 1.659 | 1,014.558 | 668.56% | 151.752 | 0.044% |
| Google - US | 1,391,841.20 | 353,640.60 | 192,109.00 | 1.918 | 1550 | 1.841 | 897.962 | 393.58% | 228.155 | 0.040% |
| MSN - Global | 133,363.89 | 12,160.36 | 11,217.00 | 1.778 | 129 | 1.084 | 1,033.829 | 1096.71% | 94.266 | 0.092% |
| MSN - US | 165,451.31 | 16,098.49 | 10,808.00 | 1.651 | 140 | 1.489 | 1,181.795 | 1027.74% | 114.989 | 0.082% |
| Overture - Global | 365,788.84 | 64,295.86 | 60,899.00 | 1.470 | 372 | 1.056 | 983.303 | 568.92% | 172.838 | 0.002% |
| Overture - US | 205,457.18 | 141,976.07 | 119,323.00 | 2.178 | 289 | 1.190 | 710.924 | 144.71% | 491.267 | 0.002% |
| Yahoo - US | 836,091.13 | 46,197.82 | 45,598.00 | 1.869 | 662 | 1.013 | 1,262.978 | 1809.81% | 69.785 | 0.071% |
| Grand Total | 3,906,596.63 | 755,315.92 | 512,849.00 | 12.686 | 3939 | 9.332 | 7,085.350 | 57.100246 | 1323.054 | 0.0033342 |

It is evident from the that Google – US is one of the major contributors and the Net Revenue for Google – US is highest compared to all Publishers, they also have the maximum volume of booking and clicks. Google – Global has secured 2nd position and has a better ROA rate when compared from Google – US. MSN – Global has generated minimum revenue and however the probability of booking is best amongst the competition. MSN – US performed better than MNS – Global since the volume of booking for MNS – US is more than Global. Also, MNS is the most underperformed publisher compared to all in the table above. The ROA for Yahoo – US is highest compared to all Publishers, Yahoo only possess competition threat to Google, however they have no existence Globally to support their competition against Google. Overture has a Global and US existence when we look at the table the Average Pos for Overture US is highest amongst all. Cost per booking for Overture – US is also the highest as noticed. If we look at the metrics like ROA Yahoo – US has a best rate of ROA, wherein Google – US even after generating maximum revenue their ROA is 2nd last. Another metric Probability of booking MSN is a clear winner. Average cost per click is considered best when it is lowest and Yahoo – US has performed amazingly well in this area.

Based on the above analysis we can conclude that a) Google is the most successful company among the four companies in this case. Google earns the highest revenue although it has the highest cost. b), the business in US is better than the business in the Global level.

**Ans 3**:

**Procedure followed**

* Initially the given data has 27 columns out of which two columns (AverageRevenueperBooking, BidStrategy) has high number of null values. So those two are dropped.
  + *27 -> 25*
* Before starting the conversion of categorical variables to numerical using dummy columns, some of unnecessary columns needs to be removed that have too many unique values. They are ‘PublisherID', 'PublisherName', 'Keyword', 'KeywordID', 'KeywordGroup', 'KeywordType', ‘Category’
  + *25 -> 18*
* Now the rest of the categorical features are converted to numerical and dummy features were created for each of those categorical feature.
  + *18 -> 45*
* Correlation is performed and features that are highly correlated with each other (> 0.8) are found. We only need to keep one variable and drop the other in each such pair to reduce redundancies.
  + Campaign\_Unassigned is highly correlated to MatchType\_Broad
  + ClickCharges is highly correlated to Clicks
  + Total Cost is highly correlated to Clicks
  + NetRevenue is highly correlated to Amount
  + ROA is highly correlated to Amount
  + *45 -> 40*
* Now the main step is performed where highly correlated features to our specific target variable i.e. Total Volume of Bookings are selected. The threshold used for this correlation is 0.5
  + *40 -> 2*
* The obtained relevant features varies with our y as following

*Chart, scatter chart

Description automatically generated*

**Why correlation?**

The filtering here is done using correlation matrix created using Pearson correlation. Correlation explains how one or more variables are related to each other. We will only select features that have correlation greater than 0.5 (taking absolute value) with our output variable. This is because a value closer to 1(after taking absolute value) implies stronger positive or negative correlation.

**Results**

1. The final set of independent variables are
   * Clicks
   * Amount
2. The coefficients obtained are
   1. Clicks - 0.03
   2. Amount - 0.87

* Intercept obtained is 0.00
* So essentially the regression equation becomes
  1. TotalVolumeofBookings = 0.87 \* Amount + 0.03 \* Clicks + 0

1. Explaining coefficients
   1. Clicks
      1. The coefficient is 0.03
      2. It means that a unit increase in "Clicks" results in an increase in average “TotalnumberofBookings” by 0.03 units, all other variables held constant.
      3. “Clicks” is positively correlated with our target variable
   2. Amount
      1. The coefficient is 0.87
      2. It means that a unit increase in "Amount" results in an increase in average “TotalnumberofBookings” by 0.87 units, all other variables held constant.
      3. “Amount” is highly positively correlated with our target variable
2. Summary

Text

Description automatically generated