WA-Assignment#3- HotelClickStream

Course – Web & Social Analytics (ITSM 6209)

Team Members – Rakesh Singh

Professor - Dr. Wenjing Duan

Date – 20-June-2021

**Question 1)**

a) Please provide a summary table showing the top 10 *domain names (DOMAIN\_NAME)* that generated the most volume of transactions the report should look like the following Table

**Ans**:

The data below was prepared from excel by applying a pivot, the top 10 domain names are highlighted below in the table.

|  |  |  |
| --- | --- | --- |
| Rank | Domain Names | # of Transactions |
| 1 | marriott.com | 524 |
| 2 | hilton.com | 438 |
| 3 | hotels.com | 400 |
| 4 | expedia.com | 380 |
| 5 | priceline.com | 309 |
| 6 | choicehotels.com | 297 |
| 7 | jetblue.com | 229 |
| 8 | hotwire.com | 217 |
| 9 | bestwestern.com | 120 |
| 10 | orbitz.com | 109 |

A quick graph has been plotted to visualize how these top 10 domains are performing in terms of transactions.

**Fig. 1. Visual representation of top 10 domains by transaction**

**Summary of Observations:**

Marriott is at 1st rank having a gap of almost 20% from the 2nd rank holder Hilton. Priceline on the other hand is a distant fifth as there is a considerable difference between expedia and priceline while expedia is a close fourth and similarly, hotwire and bestwestern at 8 and number 9 position respectively has good distance of 80%. Finally, to conclude Marriott made the maximum volume of transactions and secured the first rank with 524 transactions. Out of top 10 domain names the last in the ranking is orbitz at number 10 with 109 transactions only.

b) Please provide a summary table showing the top 10 *reference domain names (REF\_DOMAIN\_NAME)* that generated the most volume of transactions the report should look like the following Table

**Ans**:

|  |  |  |
| --- | --- | --- |
| **Rank** | Reference Domain Names | # of Transactions |
| 1 | google.com | 620 |
| 2 | yahoo.com | 222 |
| 3 | bing.com | 129 |
| 4 | aol.com | 53 |
| 5 | comfortinn.com | 48 |
| 6 | jetblue.com | 43 |
| 7 | qualityinn.com | 29 |
| 8 | comfortsuites.com | 22 |
| 9 | kayak.com | 20 |
| 10 | mywebsearch.com | 17 |

The data above was prepared from excel by applying a pivot, the top 10 domain names are highlighted above in the table.

**Fig. 2. Visual representation of top 10 references domains by transaction**

**Summary of Observations**:

Google secures Rank 1 in reference domain name with the most transaction of 620 wherein yahoo stands at Rank 2 with 222 transactions. If we look at it between Rank 1 and Rank 2 there is a whooping difference of 398 transactions this also means that Google leads transaction with a significant difference. Bing is at Rank 3 while Aol is a distant fourth. Minor observation for rest of the domain that they remain close to each other with not significant difference with respect to transactions received.

**Question 2**: Please use the Binary Outcome (Logistic/Logit) regression technique to answer the question on “*what are the factors that influence people’s decision on whether to book directly on a hotel website or from other third party website?”* Please use *DIRECT\_D* as your Dependent Variable (DV); and *REF\_D, LOG\_PRICE, TRANS\_FREQ, DURATION, HOUSEHOLD\_SIZE, CHILDREN\_D*, and *CONNECTIONSPEED\_D* as your Independent Variables (IV). Please report and *interpret* your regression results, which should include the interpretation of the regression coefficients.

**Ans 2**:

The analysis below for Binary Outcome (Logistic/Logit) regression technique was performed on RStudio using Rcommander and the results was published on RStudio console the snapshot below has been taken from the console with the name of local directory where it’s saving these outputs.

Text, table

Description automatically generated with medium confidence

**Fig. 3. Binomial (Logit/Logistic) regression summary**

**Analysis & Observations -**

Looking at the data above using the regression coefficients it is evident that Reference website (REF\_D), Transaction frequency (TRANS\_FREQ), Log of Product Total Price (LOG\_PRICE), the Population of Children (CHILDREN\_D) and Connection Speed(CONNECTIONSPEED\_D) have a Positive Correlation with Transaction Occurring on Direct Website(DIRECT\_D), while Duration of Usage(DURATION) and Size of Household (HOUSEHOLD\_SIZE) have a Negative Correlation with it.

Let’s look at the correlations, the positive correlation narrates that the variables are directly proportional, and the negative correlation states the variables are inversely proportional. While from all these independent variables, only REF\_D,TRANS\_FREQ,DURATION and CHILDREN\_D hold Statistical Significance as its P-Values < 0.05. The output beings statistically significant means the output is most likely not due to mere chance.

Now considering the Logit Coefficients which implicate the change in log odds of the dependent variable upon one unit increase in the independent variable. From the above analysis the following can be concluded:

* For every 1 unit increase in REF\_D, the log odds of adopting Direct Website Transaction will increase by 0.735 with odds ratio of 2.086. This means that another user will have 108.6% higher odds of adopting transactions via direct website
* Similarly, with every 1 unit increase in TRANS\_FREQ, the log odds of adopting Direct Website Transaction will increase by 0.115 with odds ratio of 1.122. This means that another user will have 12.2% higher odds of adopting transactions via direct website
* Since DURATION has a negative correlation with DIRECT\_D, for every 1 unit increase in DURATION, the log odds of adopting Direct Website Transaction will decrease by 0.019 with odds ratio of 0.9814. This means that the more time spent by customers on the hotel website, the less chances the transaction will be completed from the site by lowering the odds by 1.86%
* Finally, with a increase in CHILDREN\_D by 1 unit the log odds of direct transaction increases by 0.256 with odds ratio of 1.2915, which means there is a 29.15% higher chance of users directly transacting from the website

**Question 3. a) Please use the Count Data (Poisson) regression model to answer the question on “*what are the factors that influence people’s booking frequencies?”* Please use *TRANS\_FREQ* as your DV; and *REF\_D, LOG\_PRICE, PAGES\_VIEWED, HOUSEHOLD\_SIZE, CHILDREN\_D*, and *CONNECTIONSPEED\_D* as your IVs. Please report and *interpret* your regression results, which should include the interpretation of the regression coefficients.**

**b) Please repeat the analysis in question a) using the Negative Binomial Regression model. Please report and *interpret* your regression results and coefficients.**

**c) Please summarize your observations by comparing the results from a) and b).**

Ans: a) Overview: In [statistics](https://en.wikipedia.org/wiki/Statistics), Poisson regression is a [generalized linear model](https://en.wikipedia.org/wiki/Generalized_linear_model) form of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) used to model [count data](https://en.wikipedia.org/wiki/Count_data) and [contingency tables](https://en.wikipedia.org/wiki/Contingency_table). Poisson regression assumes the response variable Y has a [Poisson distribution](https://en.wikipedia.org/wiki/Poisson_distribution), and assumes the [logarithm](https://en.wikipedia.org/wiki/Logarithm) of its [expected value](https://en.wikipedia.org/wiki/Expected_value) can be modeled by a linear combination of unknown [parameters](https://en.wikipedia.org/wiki/Parameter). A Poisson regression model is sometimes known as a [log-linear model](https://en.wikipedia.org/wiki/Log-linear_model), especially when used to model contingency tables.

Negative binomial regression is a popular generalization of Poisson regression because it loosens the highly restrictive assumption that the variance is equal to the mean made by the Poisson model.

Poisson regression models are [generalized linear models](https://en.wikipedia.org/wiki/Generalized_linear_model) with the logarithm as the (canonical) [link function](https://en.wikipedia.org/wiki/Link_function), and the [Poisson distribution](https://en.wikipedia.org/wiki/Poisson_distribution) function as the assumed probability distribution of the response.

Poisson regression with multiple predictor variables:

log(λ\_i) = β\_0 + β\_1\***TRANS\_FREQ** + β\_2\***DURATION** + β\_3\* **HOUSEHOLD\_SIZE** + β\_4\***CHILDREN\_D** + β\_5\***CONNECTIONSPEED\_D** + β\_6\***REF\_D** + β\_7\***LOG\_PRICE**

where the observed values TRANS\_FREQ ~ Poisson with λ=λ\_i for a given x\_i

**Inferences**

We use Rcmdr to compile the results just as in the second question above. For Poisson regression, REF\_D, LOG\_PRICE, PAGES\_VIEWED, HOUSEHOLD\_SIZE, CHILDREN\_D, and CONNECTIONSPEED\_D were used as the independent variables and TRANS\_FREQ was used as the dependent variable.

The summary output is shown in fig. 8 and we derive the following interpretations from it:

1. Pr(>|Z|) has the same meaning as in fig. 7. We see that REF\_D, PAGES\_VIEWED, CHILDREN\_D and CONNECTIONSPEED\_D are the most statistically significant variables because of their extremely low Pr(>|Z|) values (which are extremely close to 0). They are followed by variables LOG\_PRICE and HOUSEHOLD\_SIZE with larger values farther from zero. In our Poisson regression model, **all** the independent variables are statistically significant.
2. A unit increase in the in the value of REF\_D causes the log count to **decrease** by 0.228. By taking the exponent of the coefficient, e^0.228 = 1.26, we can conclude that a unit increase in REF\_D would **decrease** transaction frequency by a factor of 1.26.
3. A unit increase in the in the value of LOG\_PRICE causes the log count to **increase** by 0.026. By taking the exponent of the coefficient, e^0.026 = 1.03, we can conclude that a unit increase in LOG\_PRICE would **increase** transaction frequency by a factor of 1.03.
4. A unit increase in the in the value of PAGES\_VIEWED does not affect the count. By taking the exponent of the coefficient, e^0.002 = 1.00, we can conclude that PAGES\_VIEWED does not affect the transaction frequency.
5. A unit increase in the in the value of HOUSEHOLD\_SIZE does not affect the count. By taking the exponent of the coefficient, e^0.012 = 1.00 we can conclude that HOUSEHOLD does not affect the transaction frequency.
6. A unit increase in the in the value of CHILDREN\_D causes the log count to **decrease** by 0.23. By taking the exponent of the coefficient, e^0.23 = 1.26, we can conclude that a unit increase in CHILDREN\_D would **decrease** transaction frequency by a factor of 1.26
7. A unit increase in the in the value of CHILDREN\_D causes the log count to **increase** by 0.9. By taking the exponent of the coefficient, e^0.9 = 2.46, we can conclude that a unit increase in CHILDREN\_D would **decrease** the transaction frequency by a factor of 2.46.

Table

Description automatically generated

**Fig. 4 Poisson regression summary for 3a.**

Ans: b) Overview: Negative binomial regression is a popular generalization of Poisson regression because it loosens the highly restrictive assumption that the variance is equal to the mean made by the Poisson model. The traditional negative binomial regression model, commonly known as NB2, is based on the Poisson-gamma mixture distribution. This model is popular because it models the Poisson heterogeneity with a gamma distribution. The equation is same as that of Poisson regression but with allowing the mean and variance to be different.

**Inferences**

We use the same set of independent and dependent variables as in 3a. The summary output for negative binomial regression is shown in fig. 9 and we derive the following interpretations from it:

1. We see that REF\_D and CHILDREN\_D are the most statistically significant variables. Each of them has a p-value close to zero. CONNECTIONSPEED\_D is somewhat statistically significant, with its p-value close to 0.05. We now discuss how the coefficients affect the dependent variable **for the variables that are statistically significant**.
2. A unit increase in the value of REF\_D would cause a **decrease** in the difference of logs of transaction frequency by 0.29 units.
3. A unit increase in the value of CHILDREN\_D would cause a **decrease** in the difference of logs of transaction frequency by 0.23 units.
4. A unit increase in the value of CONNECTIONSPEED\_D would cause an increase in the difference of logs of transaction frequency by 0.23 units.

Table

Description automatically generated

**Fig. 5. Negative Binomial regression summary for 3b**

Ans: c)

As discussed above, both regression models depict similar positive and negative correlation, however, Poisson regression model shows statistically significant correlation with 6 coefficients. Also, negative binomial regression depicts statistically significant correlation with 3 coefficients on this data set.

To compare which regression model is better we need to validate the mean and variance of the TRANS\_FREQ from Q1. Mean = 2.98 and Variance = 16.9. Poisson regression is a restrictive model (Mean~Vairance) i.e. Mean is similar or equal to variance, hence, it is giving unreasonable results due to over-dispersion in count data. As we know, poisson regression is a reactive model ((Mean<Variance) which appears to be working better on this dataset.