Customer Analytics and Customer Insights
WS 2021/22
Assignment 1

Solution 1a:

Table 1: Summary of descriptive statistics

```
     Choice          Email_25        Email_Taxi         Gmail        
 Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.000   Median :0.000   Median :0.0000   Median :0.0000  
 Mean   :0.148   Mean   :0.333   Mean   :0.3333   Mean   :0.3093  
 3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
    Add_Eug            Age           Tickets        RoundTrip      
 Min.   :0.0000   Min.   :18.00   Min.   :1.00   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:30.00   1st Qu.:1.00   1st Qu.:1.0000  
 Median :0.0000   Median :38.00   Median :1.00   Median :1.0000  
 Mean   :0.3797   Mean   :38.04   Mean   :1.99   Mean   :0.8697  
 3rd Qu.:1.0000   3rd Qu.:46.00   3rd Qu.:3.00   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :71.00   Max.   :5.00   Max.   :1.0000  

     yahoo             Edu           AlaskaFF        Add_Ore       
 Min.   :0.0000   Min.   :0.0000   Min.   :1.000   Min.   :0.000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.000  
 Median :0.0000   Median :0.0000   Median :2.000   Median :0.000  
 Mean   :0.2237   Mean   :0.1053   Mean   :1.741   Mean   :0.251  
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:1.000  
 Max.   :1.0000   Max.   :1.0000   Max.   :3.000   Max.   :1.000  
```

Table 2: Correlation Plot of Nominal Variables

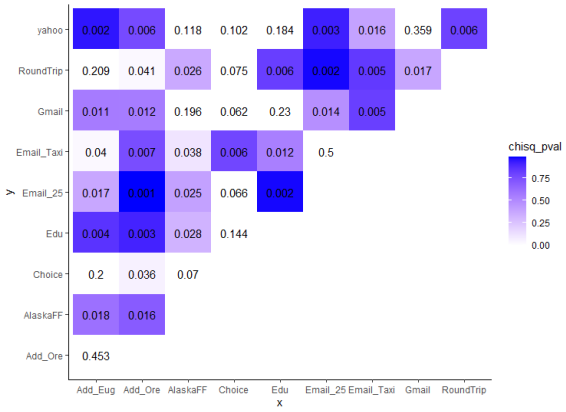
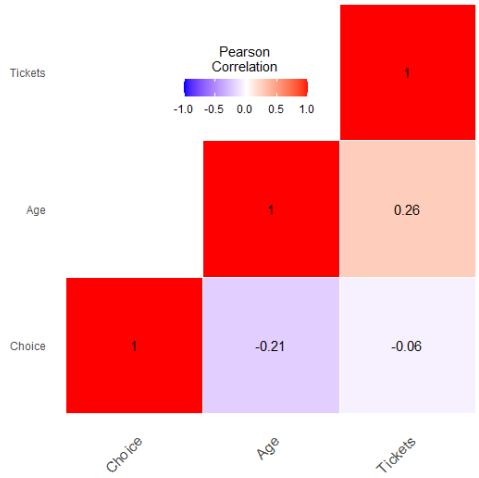


Table 3: Correlation of Metric Variables



As we can see from Table 1, the variable age is a metric(continuous) variable that has a mean of 38.04, with minimum and maximum ages of 18 and 71. The variable Tickets which is also a continuous variable, we see the number of tickets booked by the passenger in the itinerary range from 1 to 5, with a median of 1, showing that at least 50% of passengers have booked just one ticket, with a median of 1.99. The variables RoundTrip, Add_Ore, Add_Eug, yahoo, Edu, AlaskaFF , Gmail, Email_25 and Email_Taxi are dummy variables. RoundTrip with values 0 and 1, where 0 shows no roundtrip and 1 shows a roundtrip. AlaskaFF, has values 1,2 and 3, and a mean of 1.741, and the univariate graph of the variable shows that most of the fliers fall in category 2, which is the frequent flyer category, and very few in categories 1 and 3. Add_Eug has a mean of 0.380 and Add_Ore has a mean of 0.251. This shows that 63% of the passengers are from the state, either from Eugene or Oregon or Springfield. Email_25, Email_Taxi and Gmail have almost the same means, which shows the equal number of passengers belonging to each email category.

Solution 1b: To determine the relationship between the coefficients one uses correlation to understand how the features are associated with each other. To measure the association of metric variables, i.e., Age and Tickets in our data we use Pearson's coefficient correlation method. Age and Tickets have a low and significant

Table 4: Specifications of Different Models

Estimate	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	-0.738472	-0.63156047	-9.94617E-17	0.294022149 ***	0.304416966 ***
Email_25	0.595335***	0.487724***	8.949309e-02***	0.067430302***	0.056403655***
Email_Taxi	0.214881		3.111693e-02	0.023439783	
Gmail	0.289448*	0.33079*	4.632813e-02*	0.035591625*	0.037414471*
yahoo	-0.173487		-1.010196e-02	-0.008608585	
Edu	0.514541**	0.547872**	8.554096e-02***	0.098948965***	0.100433732***
AlaskaFF	0.285926*	0.280966*	4.098990e-02*	0.028059339*	0.027396065*
Add_Ore	-0.341035**	-0.344982**	-6.430302e-02***	-0.052662835***	-0.052861041***
Add_Eug	-1.567497***	-1.572962***	-2.184289e-01***	-0.159825828***	-0.160321749***
Age	-0.05003***	-0.052614***	-1.501601e-01***	-0.005156855***	-0.005377971***
Tickets	-0.035545		-1.449108e-02	-0.00437474	
RoundTrip	0.473584*	0.462169*	3.720424e-02 *	0.0392409*	0.038717902*
LogLik	-1093.553	-1095.44	-4095.998	-990.43	-992.0939
AIC	2211.107	2208.88	8217.995	2006.86	2004.188

correlation of -0.21 and -0.06 with Choice. There is also a low and significant association of 0.26 between Age and Tickets. However, Cramer's V is used for the nominal variables as can be seen in Table 2. We can see that there is a low to moderate correlation between variables. Only one significant but a low correlation of 0.006 can be seen of the outcome variable Choice with Email_taxi. Correlation between Choice with Add_eug, Add_Ore, Gmail, Edu and yahoo is 0.2, 0.036, 0.062, 0.144 and 0.102 respectively, which is statistically insignificant. Low but significant correlation can be seen among the independent variables, like yahoo and Add_Eug(0.002), RoundTrip and Email_25(0.002), Email_25 and Add_Ore(0.001),Email_25 and Edu(0.002), and so on. The significant level is measured by p-value<0.05.

Solution 2:

In table 4, we have 5 models. Model 1 is a logistic regression model consisting of all variables in the dataset, Model 2 is a logistic regression model with only significant variables, model 3, is a standardized linear regression model containing all the feature variables in the dataset, Model 4 is a linear regression model on unscaled dataset and model 5 is a linear regression model on an unscaled dataset containing only significant variables.

Model 1 is a good model specification since the outcome variable, Choice is a binary variable, values 0 and 1, and so using a Logit model gives the flexibility to understand what is the probability of the outcome given the drivers of the outcome. We employ a logistic regression model. Although, The LRT and AIC of the linear regression models 4 and 5 are somewhat better than Logistic regression, but Logistic regression is a better choice as it helps in predicting the probability of an outcome based on the drivers of the outcome whereas linear regression will predict the change in outcome given change in the independent variable. It is therefore good to use logistic regression to predict the probability of booking an Airbnb.

The model is as follows:

$$P(Choice = 1) = \frac{exp\,(\beta_0+\beta_1\cdot Email\_25+\beta_2\cdot Email\_25+\beta_3\cdot Gmail+\beta_4\cdot yahoo+\beta_5\cdot Edu+\beta_6\cdot AlaskaFF+\beta_7\cdot Add_{Ore}+\beta_8\cdot Add_{Eug}+\beta_9\cdot Age+\beta_{10}\cdot Tickets+\beta_8\cdot RoundTrip)}{1+exp\,(\beta_0+\beta_1\cdot Email\_25+\beta_2\cdot Email\_25+\beta_3\cdot Gmail+\beta_4\cdot yahoo+\beta_5\cdot Edu+\beta_6\cdot AlaskaFF+\beta_7\cdot Add_{Ore}+\beta_8\cdot Add_{Eug}+\beta_9\cdot Age+\beta_{10}\cdot Tickets+\beta_8\cdot RoundTrip)}$$

Equation 1

In R:

```
Airbnb_log_reg <-
glm(Choice~Email_25+Email_Taxi+Gmail
+yahoo+Edu+AlaskaFF+Add_Ore+Add_Eu
g+Age+Tickets +RoundTrip, family =
"binomial")
```

The reason for choosing this model is since there is no high correlation between independent variables, the interaction between features do not return any coefficients or significant coefficients. Also, removing the insignificant features (with p-value > 0.05) from the model does not have a significant impact on the model fit as can be seen in Table 4. Therefore, the proposed model is given by equation-1.

Solution 3a:

The customer characteristics that act as significant drivers to book an Airbnb are the Age, Address, their frequent flyer status, whether they have a roundtrip or not and the email account of the customers.

As we can see in table 5, Age is a significant variable with a negative β coefficient of -0.05, passengers are 0.951 times less likely to book an Airbnb when the age is higher. Possible reasons why younger travellers are more likely to book Airbnb, as older people may have established loyalty with some hotels like Marriot which might not be so established in younger people. Also, younger people may be very adept in technology as compared to older people.

We can also see the significance of variables Add_Eug and Add_Ore. Both have a negative coefficient, which shows that people who live in Eugene or Oregan are less likely to book houses via Airbnb. If People are residents of Eugene and Oregon then the probability to book an Airbnb goes down by 20.9% and 71.1%.

We also see that there is a positive effect of AlaskaFF on the outcome variable, meaning a 1 unit increase in AlaskaFF, increases the Choice variable by 0.285926 units or the chances to book Airbnb increase by 33% if the passenger is a frequent flyer member.

The variables Edu, Gmail and RoundTrip all have a significant positive effect on the outcome variable. The passengers with an Edu email account are 67.3% more likely to book an Airbnb property, as compared to the 33.6% likelihood of Gmail account holders. And the roundtrip increases the likelihood of booking with Airbnb by 60.6%.

Solution 3b: As can be seen from the model, Table 5 and Table 6, that relative to the welcome email promotion, the email campaign offering a $25 discount has a positive effect on booking. The probability of booking Airbnb increases by 81% when a discount of $25 is offered. Although, there is no significant effect of Email_taxi on the outcome variable.

The relationship between email marketing and the outcome variable is causal because with a unit increase in the marketing

variable, Email_25 there is a significant positive increase of 0.595 units in the Choice variable. It shows that the choice variable is dependent on the Email_25 variable.

Solution 3c: The email domain that is working best for email marketing is Edu, followed by Gmail account. The probability to book increases by 66% in the case of Edu and 33.6% in the case of Gmail account. The reason is that Edu accounts mostly belong to people at an educational institution like teachers, students and staff. This could be related to awareness, propensity towards technological innovation and also income.

The people who are less innovative or not up-to-date with the latest technologies and innovations might still be using yahoo accounts and not switched to a new email provider, and so for those people, it is less likely to use a booking platform. On the contrary, passengers using Edu accounts might be students who are more prone to keep pace with the technology trends and it is easier for them to use the booking platform.

The domain will help in understanding the characteristics of the passengers, about how modern or conservative they are, about their awareness of the innovations taking place, about their economic status as well. The account of Edu mostly will belong to students, as compared to the staff of the educational institutions. Since, most of the people with Edu accounts will be students, i.e. not earning, they are most likely to respond to affordable accommodations and discounts. This type of information is crucial in targeting and positioning the brand and campaigns.

Solution 4a:

Segmenting the customers based on Acquisition probability:

Based on the acquisition model, Airbnb found out that the statistically significant

Table 5: Summary of Logic Regression Model 1

```
Call:
glm(formula = Choice ~ Email_25 + Email_Taxi + Gmail + yahoo +
    Edu + AlaskaFF + Add_Ore + Add_Eug + Age + Tickets + RoundTrip,
    family = "binomial", data = airbnb_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3649  -0.6001  -0.4107  -0.2385   2.9227

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.738472   0.398124  -1.855  0.06361 .
Email_25     0.595335   0.135804   4.384 1.17e-05 ***
Email_Taxi   0.214881   0.140331   1.531  0.12571
Gmail        0.289448   0.137410   2.106  0.03516 *
yahoo       -0.173487   0.178739  -0.971  0.33174
Edu          0.514541   0.184538   2.788  0.00530 **
AlaskaFF     0.285926   0.111699   2.560  0.01047 *
Add_Ore     -0.341035   0.126186  -2.703  0.00688 **
Add_Eug     -1.567497   0.152549 -10.275  < 2e-16 ***
Age         -0.050030   0.007042  -7.105 1.21e-12 ***
Tickets     -0.035545   0.052655  -0.675  0.49964
RoundTrip    0.473584   0.206279   2.296  0.02168 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 6: Odds ratios of significant variables

Email_25	Gmail	Edu	AlaskaFF	Add_Ore	Add_Eug	Age	RoundTrip
1.814	1.336	1.673	1.331	0.711	0.209	0.951	1.606

the marketing campaign is the email campaign with a $25discount on bookings on Airbnb, Email_25. Knowing the key driver for increasing bookings on its platform, Airbnb then had to identify who these customers were so that it can position its marketing offering campaign of $25 discount. From the acquisition model, we were able to see that the probability of booking was higher for those who were on the frequent flyer list (category 2), were not residents of the area, Eugene and Oregon and had an Edu or Gmail account as opposed to yahoo email addresses, with a focus on young customers, college-going students, with Edu accounts. Airbnb can make a deal with the airlines, and use the airline's data to send the marketing offering of $25 discount to the passengers who are the frequent flyer members of the airline and have an Edu account, potential students or staff at an educational institution and to the non-residents. The campaign should be focused on non-residents.

The acquisition probability of this customer segment is given by using equation 1 as follows, considering the value of age to be the mean of the variable (38.04) and mean of Tickets variable, 1.99:

P(Choice =1) =

$$P(Choice = 1) = \frac{\exp\left(\begin{array}{c}-0.738472 + 0.595335 \cdot 1 + 0.21488065 \cdot 0 \\ +0.28944 \cdot 2 + -0.17348655 \cdot 0 + 0.514541 \cdot 1 + 0.285926 \cdot 1 \\ -0.341035 \cdot 0 - 1.567497 \cdot 0 - 0.050030 \cdot 38.04 - 0.03554505 \cdot 1.99 \\ +0.473584 \cdot 1\end{array}\right)}{1 + \exp\left(\begin{array}{c}-0.738472 + 0.595335 \cdot 1 + 0.21488065 \cdot 0 \\ +0.28944 \cdot 2 + -0.17348655 \cdot 0 + 0.514541 \cdot 1 + 0.285926 \cdot 1 \\ -0.341035 \cdot 0 - 1.567497 \cdot 0 - 0.050030 \cdot 38.04 - 0.03554505 \cdot 1.99 \\ +0.473584 \cdot 1\end{array}\right)} = 0.433$$

Solution 4b:

1-Acquisition cost promotional campaign of $25, along with $3000 per 1000 passengers paid to the airline for the email marketing offer of $25 discount. Therefore, the total cost incurred by Airbnb on the Email_25 campaign is $3000+$25*1000 = $28000. Number of acquired customers = 0.433*1000 = 433. Cost per acquired customer = ($28000)/433 = $ 64.665

2- Email_Taxi promotional email and the same customer segment as in case 1, we get an acquisition probability of 0.343 and cost per acquired customer = $28000/ 0.360*1000 = $ 81.633

3-Welcome Email promotion and the same customer segment as in cases 1 and 2, we get acquisition probability of 0.297, Cost = $3000 and Cost per acquired customer = $3000/1000= $10.101

4- In case of Email_25 and taking a customer segment with similar characteristics as in the previous cases except that we consider residents of Eugene now, we get acquisition probability = 0.138 and Cost per acquired customer = $ 202.899