Airport Choice Modeling Report

By

Siqi Zhang

1. INTRODUCTION

Air travel is the fastest method of transport around and less accident prone than other methods of travel. Air travelers can reach most domestic destinations in hours, and international travel seldom takes more than 24 hours. Passengers choose a particular origin airport and particular airlines for a variety of reasons. This paper reports on an investigation of air traveler behavior in choosing between two departure airports and a particular airline. Data from the Airport Choice Survey of 488 air passengers were used to study the characteristics of airport and airline choice for both local residents and non-local residents.Geographically, a country or region is first divided into mutually exclusive and exhaustive air traffic zones. Air travellers, however, can and do choose between different airports and among different airlines not only based on vicinity but also a series of other attributes. Other factors can be categorized into three types. The first type is related to socio-demographic information of passengers, such as passenger's nationality, occupation, income,etc. The second type is about flight characteristics, including travel time, mode of transport, seat class of airplane, etc. The third type is closely associated with airport characteristics, including airport access time, access cost, distances from passenger's home to the airport, etc.

This paper generates two models to conduct this research. One is Logit Model, which is based on the theory of maximizing utility. It is used widely to examine and describe the relationship between a binary response variable. The other model we generated is the decision tree model. It uses a decision tree to go from observations about an item to do descriptions and conclusions about the target value. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making, which can be intuitive and clear.

The purpose of this study is to find a powerful model and determine which makes it possible to know the variables through survey data that influence passengers' choice of the different airports and various routes. Moreover, this paper also aims to analyze travel choices made by air transportation users in multi airport regions because it is a crucial component when planning passenger redistribution policies for airline companies and for the airline industry. It may allow airport planners and airport managers to identify important passenger choice behaviors without incurring the added time and expense of administering a formal passenger survey.

2. DATA

2.1 Missing Value Processing

The table below is the counts of missing values we have for each variable.

Variable	Number of Missing Value	Variable	Number of Missing Value
Airline	10	SeatClass	4
Age	1	Airfare	155
Gender	3	AccessCost	197
Destination	5	AccessTime	97
FlightNo	142	Income	132
DepartureHr	33	MileageAirline	237
DepartureMn	120	Mileage	398

Age has one missing value only, so we filled it with median. Gender also has few missing values, so we filled them with the mode (i.e.: Female). Three of the seat class missing values have airfares, and all three airfares are lower than or close to the average price of economy class, so we filled the missing value of seat class with economy. According to the bar graph below (Figure 1), we can find that most of people in Nationality 1 (i.e.: Korea) tend to choose Airline 1 (Korean Air), most people of the other four nationalities choose Airline 4 foreign airlines (since the distribution of Nationality 4 South Asia is relatively balanced, we also counted them in Airline 4). From the bar graph below (see Figure 2), we can find that the distribution of destinations is different across different airlines, so we will use the mode value of each airline to fill the missing value of destination. For income and access time, we will add one more level called “Unknown” (coded to 0) for missing values. Intuitively income is a big factor and we want to keep the variable so we try to add a third category for it.

Figure 1: Airline across Nationality

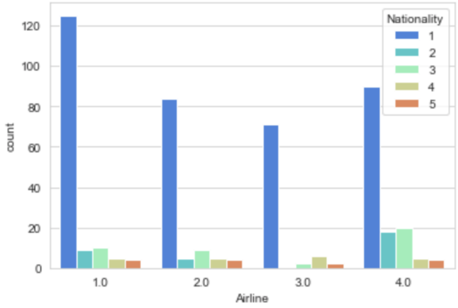
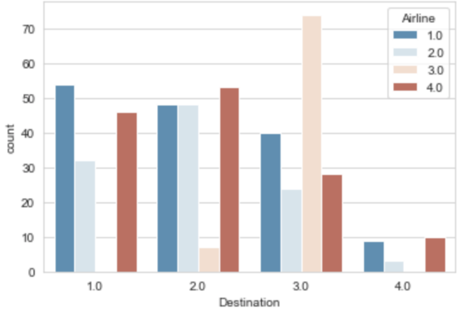


Figure 2: Destination across Airline



Since flight no. is not highly correlated with our analysis, so we will ignore the missing values in flight number. Departure hour and minutes have the departure time variable instead, so we will ignore these two variables and use departure time directly which has no missing values. Airfare has 155 missing values, which may bring bias to our model if imputed from other values, so we will not use it in our logistic model but we will use this variable in our decision tree model. This is the same for access cost. Since mileage airline and mileage are not highly correlated with our analysis and have too many missing values, we will ignore these two variables.

2.2 Re-category Processing

Age: We found the density of passengers above 60 years old from this sample is smaller than that of others who are below 60 years old. Therefore, we recategorized age into two groups, one is less than 60 years old, another is greater than or equal to 60.

Nationality: Based on the *EDA, Airport_Nationality and EDA, Airline_Nationality*, the number of Korean passengers share dramatic differences than passengers from China, Japan and other countries. So we regrouped nationality into two categories: Korean and non-Korean traveler.

Trip Purpose: There are four levels in the purpose of trip variable, and based on the following. Travelers who travel for leisure are much more than travelers who travel for other purposes. Therefore, we recategorized them into two categories: leisure travel and other.

Province Residence: Based on the purchaser may or may not reside in the Incheon Airport (ICN) and Gimpo Airport (GMP) catchment area (within or around a 120-minute drive), we regrouped Seoul, Incheon, Kyungki-do, and Chungcheong-do together. And Kyungsang-do, Jeonra-do, Kangwon-do and other areas are grouped together.

Trip Duration: We recategorized the duration of trip into one day trip and multi-day trip.

Number of Trips Last year: Most of the frequency of trips are lower than or equal to 3, so we regroup them into less than or equal to 3 and greater than 3.

Occupation: We recategorized occupation into “Entrepreneur, Senior management”, because people who are in charge of entrepreneur and senior management also have different trip purposes, travel frequencies and trip duration, etc.

Access Time: We recategorized access time into three categories. If the access time is less than or equal to 60 minutes, it will be coded to 1. If the access time is greater than 60 minutes, it will be coded to 2.If the access time is unknown, it will be coded to 0.
Airfare: We re-categorized airfare into three categories. One is airfare under 50; one is airfare between 50 and 100; another one is over 100; the last group is unknown.
Income: For income class, we found the passengers whose income less than 100 million have dramatic differences on choosing different airports and airlines. Therefore, we recategorized into people whose income less than 100 million as one group , 100 to 200 million as one group, greater than 200 million as the other group, and unknown income as the last group.
Airline: There are two ways to re-category airline. We divided airlines into Korean and Non-Korean groups. This method will be used in airline models. In addition, we divided airlines into full service carriers which include Korean Air, Asiana Airlines, and Foreign airlines. Non-full service carrier is Korean LCC.

Here are the summary of variables after re-category:

Binary Variables

Airport
Gimpo (GMP) Airport = 1
Inchoen (ICN) Airport = 0

Airline (Re-category Method 1)
Korean Air(KE), Asiana Airlines (OZ), Korean LCC = 1
Foreign Airlines = 0

Airline (Re-category Method 2)
Korean Air(KE), Asiana Airlines (OZ), Foreign Airlines = 1
Korean LCC= 0

Independent Variables

Age
Age less than 60 = 1, Age greater than or equal to 60 = 0

Gender
Female = 1, Male = 0

Nationality
Korean = 1, Non-Korean = 0

Trip Purpose
Leisure = 1, Other = 0

Province Residence
Seoul, Incheon, Kyungki-do, Chungcheong-do = 1
Kyungsang-do, Jeonra-do, Kangwon-do, other =0

Group Travel
Yes = 1, No = 0

Number of Trips Last Year
Greater than or equal to 3 = 1, Less than 3 = 0

Frequent Flight Destination
Southeast Asia, China, Japan = 0, North/South America, European =1, Other, None = 2

Flight Destinations
China = 0, Japan = 1, Southeast Asia = 2, Other = 3

Departure Time
6am - 12pm = 0, 12pm - 6pm = 1, 6pm - 9pm = 2, 9pm - 6am = 3

Seat Class
Economy = 0, Business = 1, First Class = 2

Airfare
Less than 50 = 0, Between 50 and 100 = 1, Greater than 100 = 2, Unknown = 3
Access Time
Less than 60 minutes = 0, Greater than 60 minutes = 1, Unknown = 2

Number of Transportation
One Transportation mode = 0, other = 1

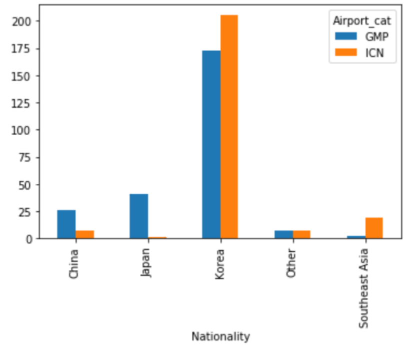
Occupation
Entrepreneur, Senior management = 0, other = 1

Income Class
Between 30 to 100 million = 0, Between 100 to 200 million = 1,
Greater than 200 million = 2, Unknown = 3

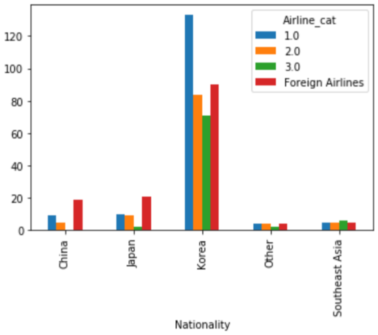
2.3 EDA

Here are exploratory data analysis of the various which need to be re-categorised. *EDA, variable* plots the variable before re-categorized, *EDA,variable(R)* plots the variable after re-categorized.

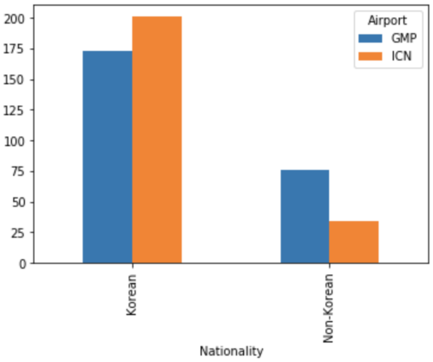
Nationality:



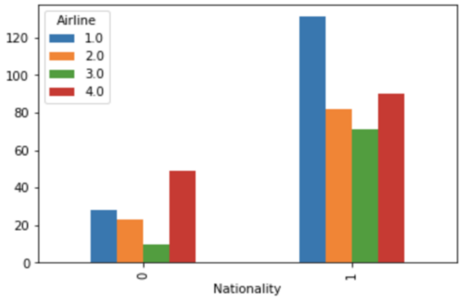
EDA, Airport_Nationality



EDA, Airline_Nationality

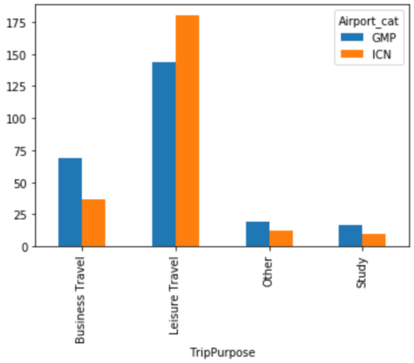


EDA, Airport_Nationality (R)

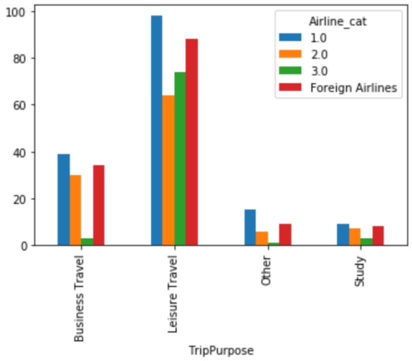


EDA, Airline_Nationality (R)

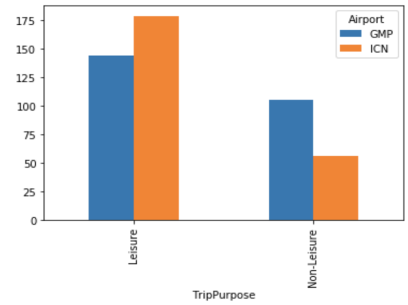
Trip Purpose:



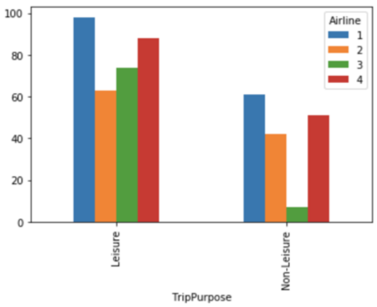
EDA, Airport_ TripPurpose



EDA,Airline_TripPurpose

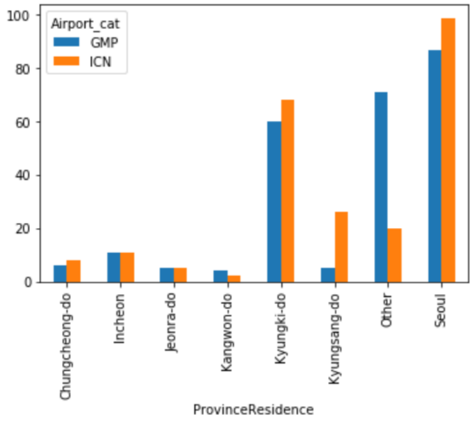


EDA, Airport_ TripPurpose (R)

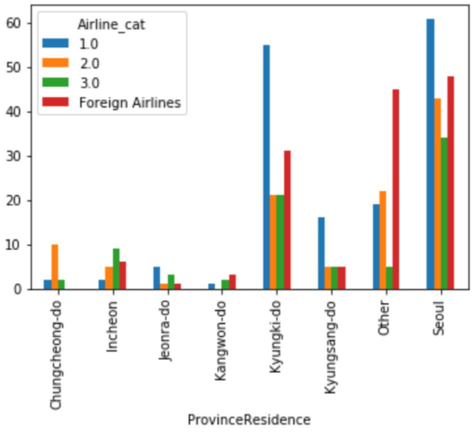


EDA,Airline_TripPurpose (R)

Province Residence:



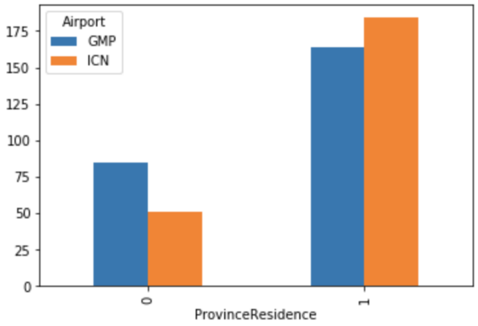
EDA, Airport_ ProvinceResidence



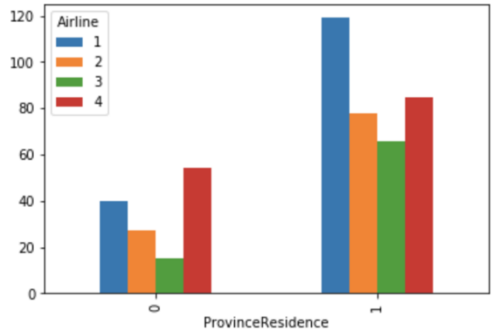
EDA, Airline_ProvinceResidence

1: Close to both airport, and travel time within 120 mins

0: Far from both airport, and travel time over 120 mins

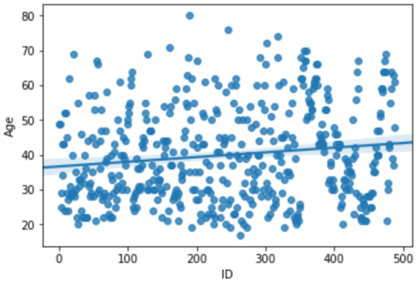


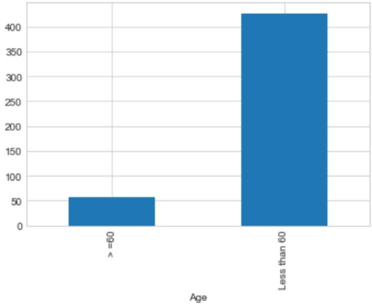
EDA, Airport_ ProvinceResidence (R)



EDA, Airline_ProvinceResidence (R)

Age:

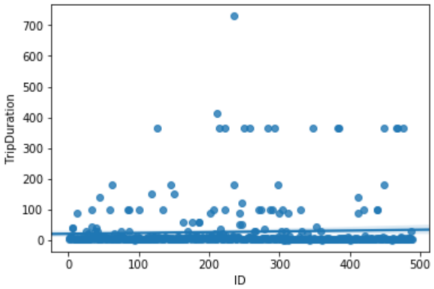


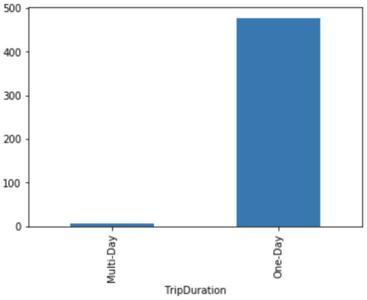


EDA, Age

EDA, Age (R)

Trip Duration:

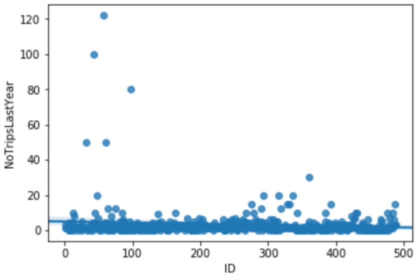


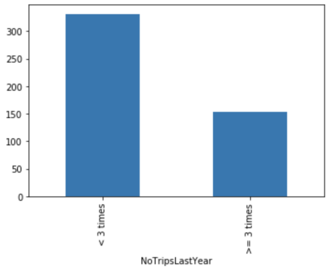


EDA, TripDuration

EDA, TripDuration (R)

Number of Trips Last Year:

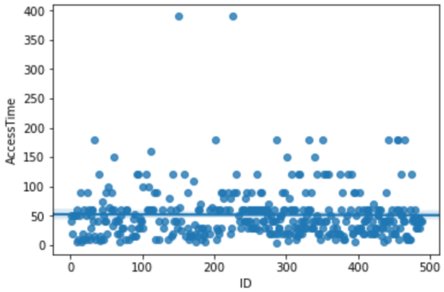


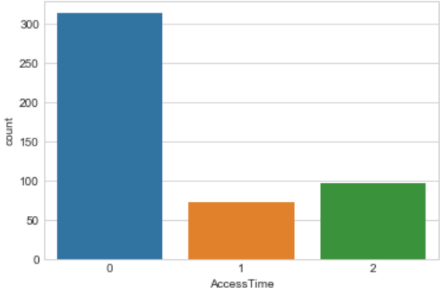


EDA, Number of Trips Last Year

EDA, Number of Trips Last Year(R)

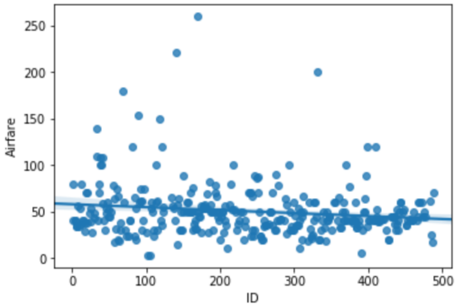
Access time:

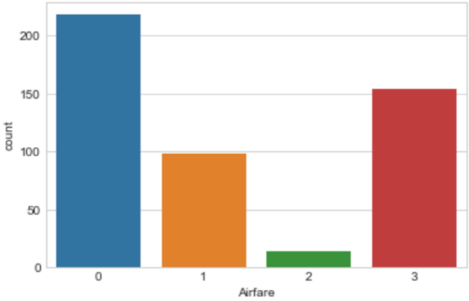




EDA, Access time *EDA, Access time（R）*

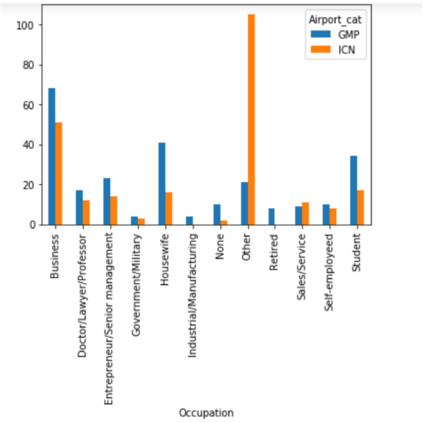
Airfare:

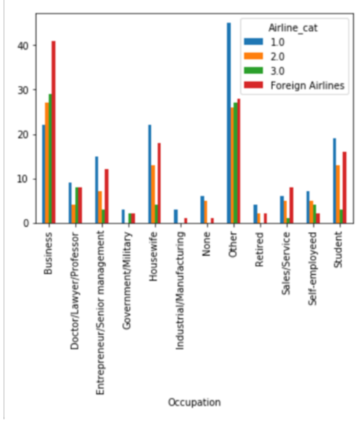




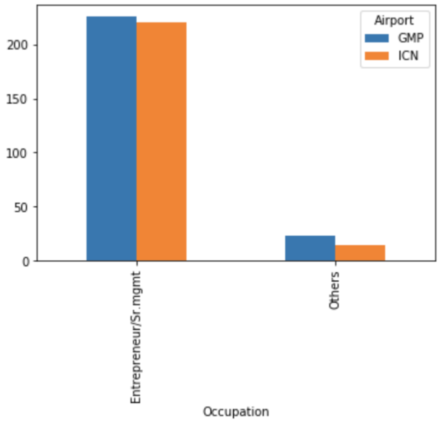
EDA, Airfare *EDA, Airfare(R)*

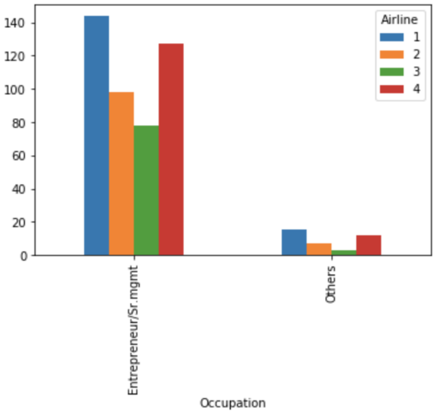
Occupation:





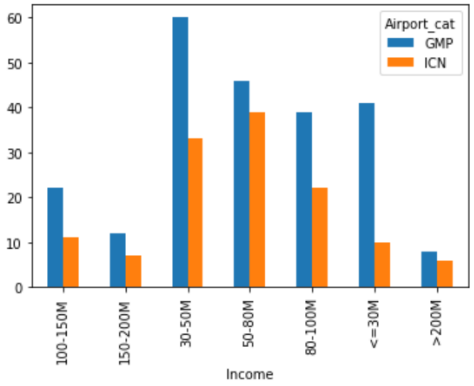
EDA, Airport_ Occupation *EDA, Airline_ Occupation*



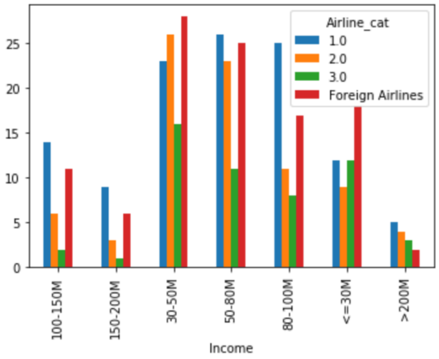


EDA, Airport_ Occupation (R)　　　*EDA, Airline_ Occupation (R)*

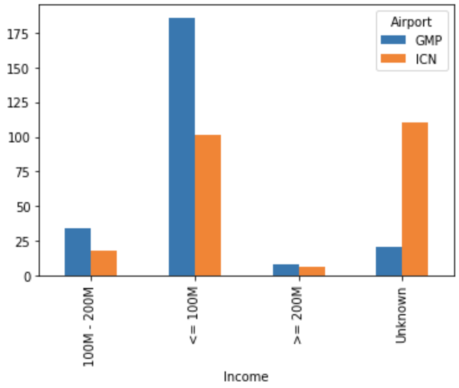
Income Class:



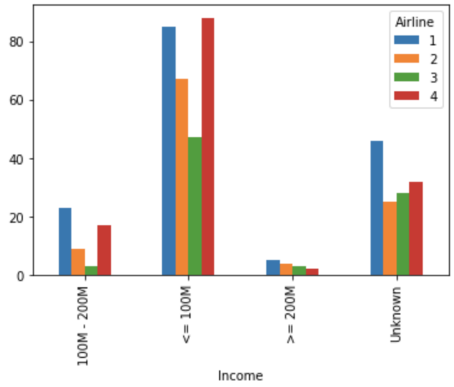
EDA, Airport_ Income



EDA, Airline_ Income



EDA, Airport_ Income (R)



EDA, Airline_ Income (R)

3. Models and Results

3.1 Logit Model

Logistic regression model is one of the most popular and one of the most widely used machine learning models people use in classification problems. In this project, we are using logistic regression to understand the behaviors in people's choice of airport or airline in Korea. In classification problems, logistic regression uses a discrete binary dependent variable (0 or 1) and takes in various independent variables. We would pick several independent variables from our survey data and fit a regression on our dependent variable to gain insights from it.

3.1.1 Model Description

Airport Model 1:

When building model 1, we considered the effect of different occupations on airport selection. Occupation may affect the trip purpose, frequent flight destination, airfare, income, and departure time.

$$\mathit{Airport} = \beta_0 + \beta_1\mathit{Airline} + \beta_2\mathit{Age} + \beta_3\mathit{Gender} + \beta_4\mathit{TripPurpose} + \beta_5\mathit{FrequentFlightDestionation} + \beta_6\mathit{Nationality} + \beta_7\mathit{Destination} + \beta_8\mathit{Airfare} + \beta_9\mathit{NoTransport} + \beta_{10}\mathit{Occupation} + \beta_{11}\mathit{Income} + \beta_{12}\mathit{DepartureTime}$$

Airport Model 2:

This model we mainly considered the effect of convenience to airport, including age, province residence, nationality, destination, income, access time, and departure time.

$$\mathit{Airport} = \beta_0 + \beta_1\mathit{Airline} + \beta_2\mathit{Age} + \beta_3\mathit{ProvinceResidence} + \beta_4\mathit{Nationality} + \beta_5\mathit{Destionation} + \beta_6\mathit{Income} + \beta_7\mathit{AccessTime} + \beta_8\mathit{DepartureTime}$$

Airport Model 3:

We also want to confirm the effect of airfare, so we added airfare in this model.

$$\mathit{Airport} = \beta_0 + \beta_1\mathit{Airline} + \beta_2\mathit{Age} + \beta_3\mathit{ProvinceResidence} + \beta_4\mathit{Nationality} + \beta_5\mathit{Destionation} + \beta_6\mathit{Income} + \beta_7\mathit{AccessTime} + \beta_8\mathit{DepartureTime} + \beta_9\mathit{Airfare}$$

Airline Model 1:

$$\mathit{Airline} = \beta 0 + \beta 1\mathit{NoTripsLastYear} + \beta 2\mathit{Income} + \beta 3\mathit{Airfare} + \beta 4\mathit{Destination} + \beta 5\mathit{GroupTravel} + \beta 6\mathit{ProvinceResidence} + \beta 7\mathit{SeatClass} + \beta 8\mathit{AccessTime} + \beta 9\mathit{TripPurpose}$$

3.1.2 Feature selection

Airport as the dependent variable:

For the feature selection, we used the forward selection method, which means we added the variables one by one in the model to check the pseudo R-squared. Since we have two different re-category methods for Airline (one is Korea and non-Korea airline, the other is full service airline and not full service airline), we check the pseudo R-squared for both re-category methods. We found that the pseudo R-squared of “Korea and non-Korea” re-category method is higher than “full service and not full service” airline. So, for the following Airport as a dependent variable models, we will use the “Korea and non-Korea” re-category method.

For the first model, since we consider the occupation effect on airport choice, we first added airline and occupation. Different occupations may have different trip purposes, so we added trip purpose then. Age and gender may also lead to a different behavior for airline choice. Since some occupations may have high frequency to one specific destination. Some people with high income may not be very sensitive to airfare, so we add airfare and income in our model. And different departure time or trip purposes may also have an effect on airport choice based on different occupations. Other variables do not have significantly effect on airport choice in this group, so we didn't include them in model 1.

For the second model, we then added the Age, ProvinceResidence, Nationality one by one for the first three models. Trip purpose has no significant effect on pseudo R-squared, and it is not statistically significant in our model (p-value is much higher than 0.05), so we dropped that variable. Age and ProvinceResidence may jointly have an effect on airport choice, so we put them together. Then we kept adding Destination and Income which also have an effect on Airport choice. Access time and departure time will also affect the choices of airport, and it can increase the pseudo R-squared significantly.

For the third model, we added one more Airfare variable based on model 2. And we found that airfare also has an effect on airport choice.

Airline as dependent variable:
For feature selection, we once again used the forward selection method. As we mentioned, we have two different re-category methods for Airline (Korea vs non-Korea airline and full service airline vs full service airline), we compare two methods and we found that "full service vs not full service" airline has a much higher pseudo R-squared. So, we use full service vs not full service to re-category the airline and use it as a dependent variable.

By regrouping the airline by full service vs not full service, we assume our model should start with variables that have a close relationship with the cost of the trip. So, we start with income, airfare and seat class as they are the ones that are cost oriented. In the process of implementing the airline model, we notice even though pseudo R-squared keeps going up but variables tend to have a low p-value, in another word, variables tend to be less statistical significant. It could be a collinearity problem to some extent which can be a limitation that we would examine later in this paper.. We tried to add on variables and continue with forward method to add on variables and improve R-squared.

3.1.3 Model implementation and Evaluation

Airport Model:
By running cross validation on our airport model , model 1 gives a K-Fold Validation Error 0.19 and the accuracy 0.83. Model 2 gives a K-Fold Validation Error 0.15 and the accuracy 0.86. Model 3 gives a K-fold Validation Error 0.15 and the accuracy 0.87. We think those are considerably good results.

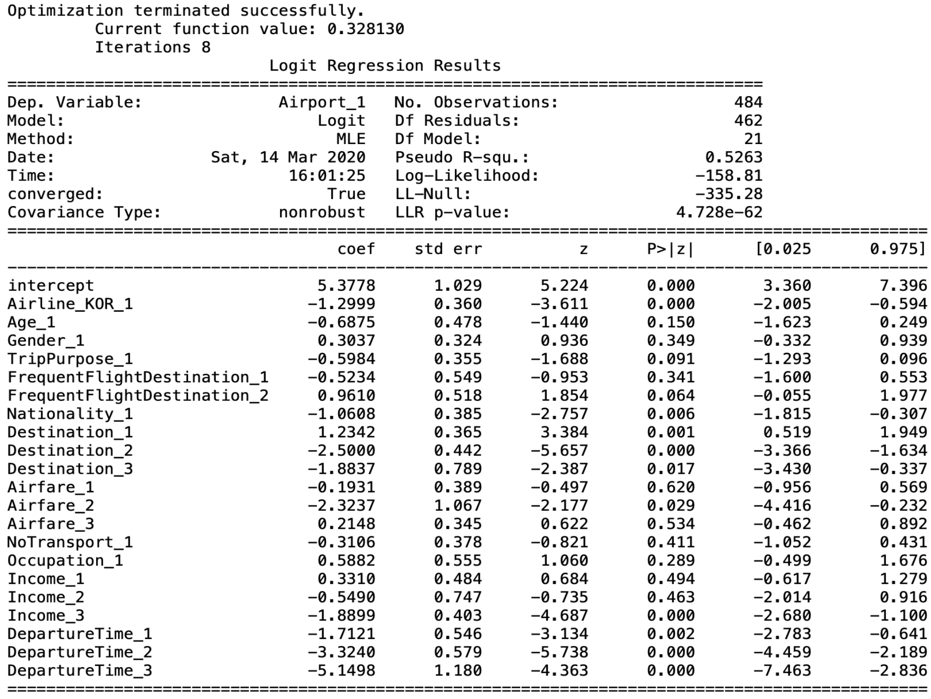
Airline Model:
By running cross validation on our airline model , the model gives a K-Fold Validation Error 0.15 and the accuracy 0.86. We think those are considerably good results; However, variables still have fairly low statistical significance.

3.1.4 Interpretation

Airport Model 1:

$$\mathit{Airport} = \beta_0 + \beta_1\mathit{Airline} + \beta_2\mathit{Age} + \beta_3\mathit{Gender} + \beta_4\mathit{TripPurpose} + \beta_5\mathit{FrequentFlightDestionation} + \beta_6\mathit{Nationality} + \beta_7\mathit{Destination} + \beta_8\mathit{Airfare} + \beta_9\mathit{NoTransport} + \beta_{10}\mathit{Occupation} + \beta_{11}\mathit{Income} + \beta_{12}\mathit{DepartureTime}$$

Output:



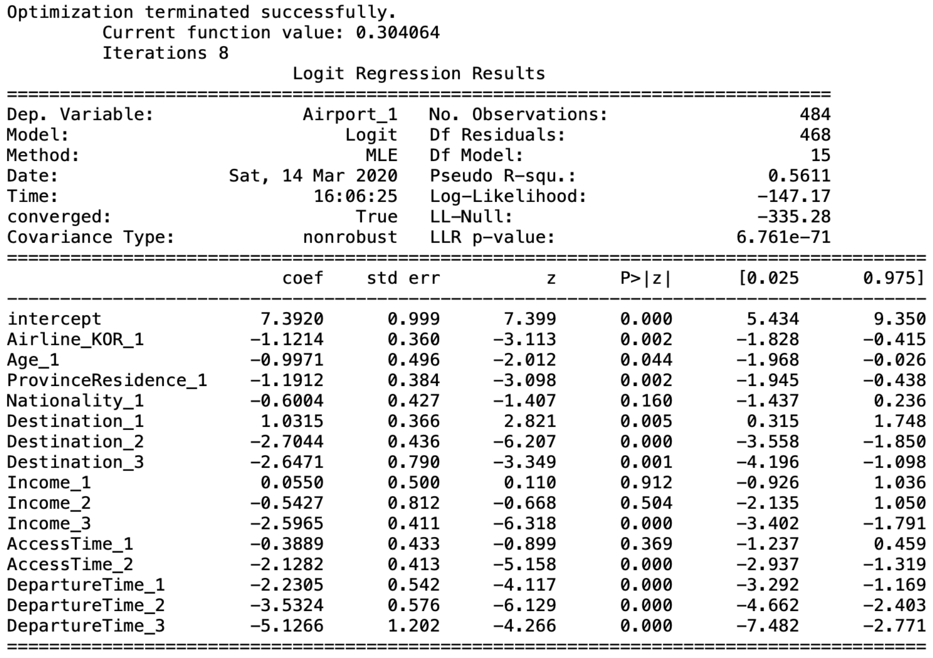
```
Optimization terminated successfully.
         Current function value: 0.328130
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:              Airport_1   No. Observations:                  484
Model:                          Logit   Df Residuals:                      462
Method:                           MLE   Df Model:                           21
Date:                Sat, 14 Mar 2020   Pseudo R-squ.:                  0.5263
Time:                        16:01:25   Log-Likelihood:                -158.81
converged:                       True   LL-Null:                       -335.28
Covariance Type:            nonrobust   LLR p-value:                 4.728e-62
========================================================================================
                                  coef   std err         z     P>|z|    [0.025    0.975]
----------------------------------------------------------------------------------------
intercept                       5.3778     1.029     5.224     0.000     3.360     7.396
Airline_KOR_1                  -1.2999     0.360    -3.611     0.000    -2.005    -0.594
Age_1                          -0.6875     0.478    -1.440     0.150    -1.623     0.249
Gender_1                        0.3037     0.324     0.936     0.349    -0.332     0.939
TripPurpose_1                  -0.5984     0.355    -1.688     0.091    -1.293     0.096
FrequentFlightDestination_1    -0.5234     0.549    -0.953     0.341    -1.600     0.553
FrequentFlightDestination_2     0.9610     0.518     1.854     0.064    -0.055     1.977
Nationality_1                  -1.0608     0.385    -2.757     0.006    -1.815    -0.307
Destination_1                   1.2342     0.365     3.384     0.001     0.519     1.949
Destination_2                  -2.5000     0.442    -5.657     0.000    -3.366    -1.634
Destination_3                  -1.8837     0.789    -2.387     0.017    -3.430    -0.337
Airfare_1                      -0.1931     0.389    -0.497     0.620    -0.956     0.569
Airfare_2                      -2.3237     1.067    -2.177     0.029    -4.416    -0.232
Airfare_3                       0.2148     0.345     0.622     0.534    -0.462     0.892
NoTransport_1                  -0.3106     0.378    -0.821     0.411    -1.052     0.431
Occupation_1                    0.5882     0.555     1.060     0.289    -0.499     1.676
Income_1                        0.3310     0.484     0.684     0.494    -0.617     1.279
Income_2                       -0.5490     0.747    -0.735     0.463    -2.014     0.916
Income_3                       -1.8899     0.403    -4.687     0.000    -2.680    -1.100
DepartureTime_1                -1.7121     0.546    -3.134     0.002    -2.783    -0.641
DepartureTime_2                -3.3240     0.579    -5.738     0.000    -4.459    -2.189
DepartureTime_3                -5.1498     1.180    -4.363     0.000    -7.463    -2.836
========================================================================================
```

From the output of logistic regression model 1, we can find that the log likelihood of the model is -158.81, and pseudo R-squared is 0.5263 accordingly. Most of the variables are statistically significant on a 5% confidence interval. The odds of choosing GMP airport for people taking Korean airline is 72.74% lower than non-Korean airline. The odds of choosing GMP airport for people lower than 60 years old is 49.71% lower than people greater than 60 years old. The odds of choosing GMP airport for female people is 35.49% higher than male. The odds of choosing GMP airport for leisure purpose trip is 45.03% higher than other trip purpose. The odds of choosing GMP airport for frequent to European or American countries is 40.75% lower than frequent to Asia countries, and other group 161% higher than frequent to Asia countries. Compared to non-Korean travelers, Korean travelers have 65.38% less likely to choose GMP airport. Travelers traveling to South Asian are 91.79% less likely to go GMP airport compared to travelers going China. Travelers buying air ticket expense (i.e.:airfare) greater 1,000,000 Korean Won are 90.21% less likely to go GMP airport. The odds of choosing GMP airport for a number of transportation greater than one are 26.70% less likely to choose GMP airport compared to number of transportation one. Compared to Entrepreneur and Senior management, other occupations are 80.07% more likely to choose GMP airport. Income level between 100 to 200 million is 39.24% more likely to choose GMP airport. Income level greater than 200 million is 42.25% less likely to choose GMP airport. Compared to departure time from 6 am to 12 pm, people departed from 12 pm to 6 pm are 81.95% less likely to choose GMP airport; people departed from 6 pm to 9 pm are 96.40% less likely to choose GMP airport; people departed from 9 pm to 6 am are 99.41% less likely to choose GMP airport.

Airport Model 2:

Airport = β_0 + β_1Airline + β_2Age + β_3ProvinceResidence + β_4Nationality + β_5Destionation + β_6Income + β_7AccessTime + β_8DepartureTime

Output:



```
Optimization terminated successfully.
         Current function value: 0.304064
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:              Airport_1   No. Observations:                  484
Model:                          Logit   Df Residuals:                      468
Method:                           MLE   Df Model:                           15
Date:                Sat, 14 Mar 2020   Pseudo R-squ.:                  0.5611
Time:                        16:06:25   Log-Likelihood:                -147.17
converged:                       True   LL-Null:                       -335.28
Covariance Type:            nonrobust   LLR p-value:                 6.761e-71
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
intercept               7.3920      0.999      7.399      0.000       5.434       9.350
Airline_KOR_1          -1.1214      0.360     -3.113      0.002      -1.828      -0.415
Age_1                  -0.9971      0.496     -2.012      0.044      -1.968      -0.026
ProvinceResidence_1    -1.1912      0.384     -3.098      0.002      -1.945      -0.438
Nationality_1          -0.6004      0.427     -1.407      0.160      -1.437       0.236
Destination_1           1.0315      0.366      2.821      0.005       0.315       1.748
Destination_2          -2.7044      0.436     -6.207      0.000      -3.558      -1.850
Destination_3          -2.6471      0.790     -3.349      0.001      -4.196      -1.098
Income_1                0.0550      0.500      0.110      0.912      -0.926       1.036
Income_2               -0.5427      0.812     -0.668      0.504      -2.135       1.050
Income_3               -2.5965      0.411     -6.318      0.000      -3.402      -1.791
AccessTime_1           -0.3889      0.433     -0.899      0.369      -1.237       0.459
AccessTime_2           -2.1282      0.413     -5.158      0.000      -2.937      -1.319
DepartureTime_1        -2.2305      0.542     -4.117      0.000      -3.292      -1.169
DepartureTime_2        -3.5324      0.576     -6.129      0.000      -4.662      -2.403
DepartureTime_3        -5.1266      1.202     -4.266      0.000      -7.482      -2.771
=======================================================================================
```

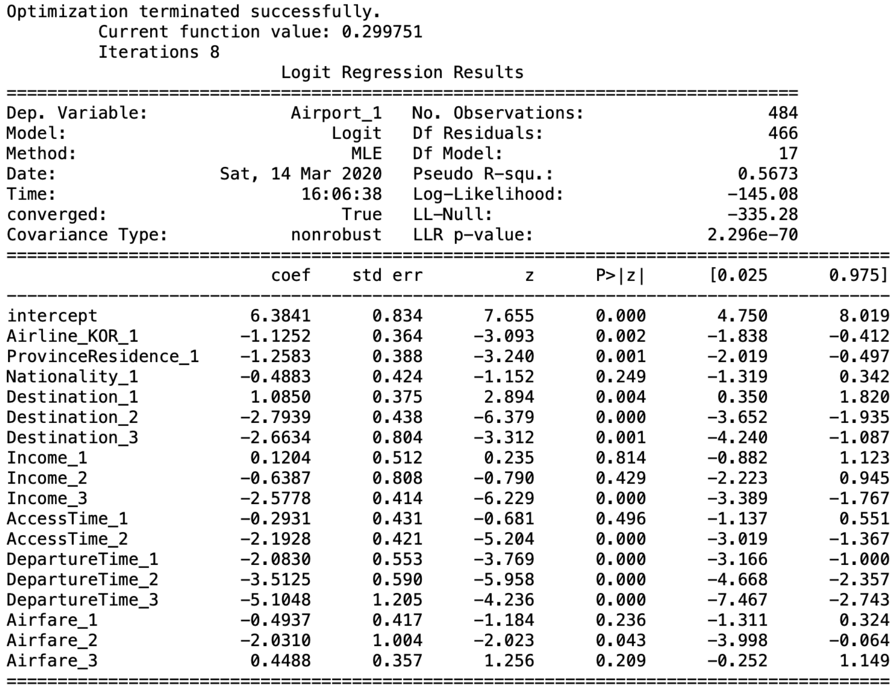
From the output of logistic regression model 2, we can find that the log likelihood of the model is -147.17, and pseudo R-squared is 0.5611 accordingly. Most of the variables are statistically significant on a 5% confidence interval. The odds of choosing GMP airport for people taking Korean airline is 67.41% lower than non-Korean airline. The odds of choosing GMP airport for people lower than 60 years old is 63.10% lower than people greater than 60 years old. The odds of choosing GMP airport for people living in Seoul, Incheon, Kyungki-do, and Chungcheong-do is 69.61% lower than people living in Kyungsang-do, Jeonra-do, Kangwon-do and other. The odds of choosing GMP airport for Korean is 45.14% lower than non-Korean. The odds of choosing GMP airport for Japan destination is 180.53% higher than destination China, and the odds ratio for Southeast Asia destination is 93.31% lower than destination China, and the odds ratio for other destinations is 92.91% lower than destination China. Income level greater than 200 million is 92.45% less likely to choose GMP airport. Compared to departure time from 6 am to 12 pm, people departed from 12 pm to 6 pm are 89.25% less likely to choose GMP airport; people departed from 6 pm to 9 pm are

97.08% less likely to choose GMP airport; people departed from 9 pm to 6 am are 99.41% less likely to choose GMP airport.

Airport Model 3:

$$Airport = \beta_0 + \beta_1 Airline + \beta_2 Age + \beta_3 ProcinceResidence + \beta_4 Nationality + \beta_5 Destionation + \beta_6 Income + \beta_7 AccessTime + \beta_8 DepartureTime + \beta_9 Airfare$$

Output:

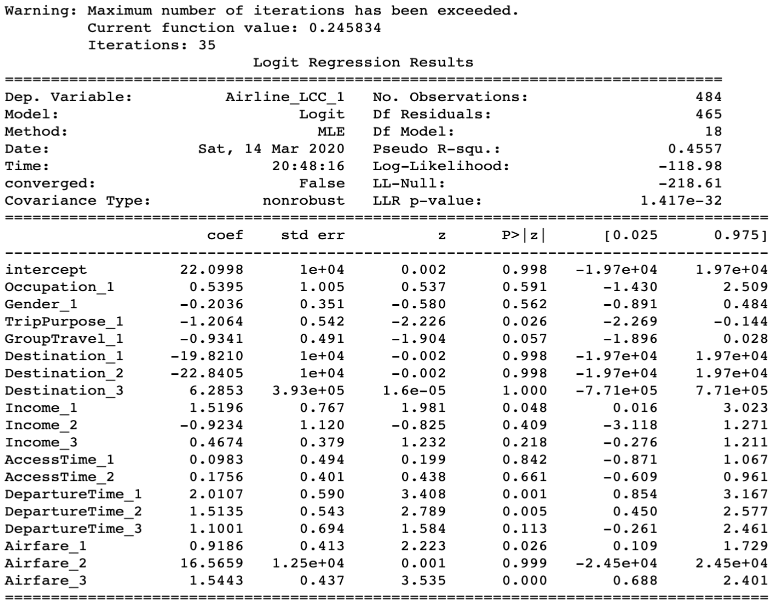


```
Optimization terminated successfully.
         Current function value: 0.299751
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:              Airport_1   No. Observations:                  484
Model:                          Logit   Df Residuals:                      466
Method:                           MLE   Df Model:                           17
Date:                Sat, 14 Mar 2020   Pseudo R-squ.:                  0.5673
Time:                        16:06:38   Log-Likelihood:                -145.08
converged:                       True   LL-Null:                       -335.28
Covariance Type:            nonrobust   LLR p-value:                 2.296e-70
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
intercept               6.3841      0.834      7.655      0.000       4.750       8.019
Airline_KOR_1          -1.1252      0.364     -3.093      0.002      -1.838      -0.412
ProvinceResidence_1    -1.2583      0.388     -3.240      0.001      -2.019      -0.497
Nationality_1          -0.4883      0.424     -1.152      0.249      -1.319       0.342
Destination_1           1.0850      0.375      2.894      0.004       0.350       1.820
Destination_2          -2.7939      0.438     -6.379      0.000      -3.652      -1.935
Destination_3          -2.6634      0.804     -3.312      0.001      -4.240      -1.087
Income_1                0.1204      0.512      0.235      0.814      -0.882       1.123
Income_2               -0.6387      0.808     -0.790      0.429      -2.223       0.945
Income_3               -2.5778      0.414     -6.229      0.000      -3.389      -1.767
AccessTime_1           -0.2931      0.431     -0.681      0.496      -1.137       0.551
AccessTime_2           -2.1928      0.421     -5.204      0.000      -3.019      -1.367
DepartureTime_1        -2.0830      0.553     -3.769      0.000      -3.166      -1.000
DepartureTime_2        -3.5125      0.590     -5.958      0.000      -4.668      -2.357
DepartureTime_3        -5.1048      1.205     -4.236      0.000      -7.467      -2.743
Airfare_1              -0.4937      0.417     -1.184      0.236      -1.311       0.324
Airfare_2              -2.0310      1.004     -2.023      0.043      -3.998      -0.064
Airfare_3               0.4488      0.357      1.256      0.209      -0.252       1.149
=======================================================================================
```

We added airfare in this model, and we can find that the log-likelihood is -145.08, and pseudo R-squared is 0.5673. The odds ratio of choosing GMP airport for airfare between 50 and 100 is 38.96% less than airfare less than 50. The odds ratio of choosing GMP airport for airfare greater than 100 is 86.88% less than airfare less than 50. The odds ratio for unknown groups is 56.64% higher than airfare less than 50 groups.

From our airport model, we conclude that non-Korean travelers and taking non-Korean airlines travelers are more likely to choose GMP airport. Older people are more likely to choose GMP airport. Female, leisure travel purpose, Japan destination traveler or other occupation people (i.e.: occupation besides entrepreneur, senior management) are more likely to choose GMP airport. On the other hand, the number of transportation to airports greater than one is more likely to choose INC airport. People with pretty high income (i.e.: greater than 200 million) are more likely to choose INC airport. Departure time from 12 pm to 6 am, or people living in Seoul, Incheon, Kyungki-do, and Chungcheong-do are more likely to choose INC airport.

Airline Model 1:



```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.245834
         Iterations: 35
                           Logit Regression Results
==============================================================================
Dep. Variable:          Airline_LCC_1   No. Observations:                  484
Model:                          Logit   Df Residuals:                      465
Method:                           MLE   Df Model:                           18
Date:                Sat, 14 Mar 2020   Pseudo R-squ.:                  0.4557
Time:                        20:48:16   Log-Likelihood:                -118.98
converged:                      False   LL-Null:                       -218.61
Covariance Type:            nonrobust   LLR p-value:                 1.417e-32
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
intercept          22.0998      1e+04      0.002      0.998   -1.97e+04    1.97e+04
Occupation_1        0.5395      1.005      0.537      0.591      -1.430       2.509
Gender_1           -0.2036      0.351     -0.580      0.562      -0.891       0.484
TripPurpose_1      -1.2064      0.542     -2.226      0.026      -2.269      -0.144
GroupTravel_1      -0.9341      0.491     -1.904      0.057      -1.896       0.028
Destination_1     -19.8210      1e+04     -0.002      0.998   -1.97e+04    1.97e+04
Destination_2     -22.8405      1e+04     -0.002      0.998   -1.97e+04    1.97e+04
Destination_3       6.2853   3.93e+05    1.6e-05      1.000   -7.71e+05    7.71e+05
Income_1            1.5196      0.767      1.981      0.048       0.016       3.023
Income_2           -0.9234      1.120     -0.825      0.409      -3.118       1.271
Income_3            0.4674      0.379      1.232      0.218      -0.276       1.211
AccessTime_1        0.0983      0.494      0.199      0.842      -0.871       1.067
AccessTime_2        0.1756      0.401      0.438      0.661      -0.609       0.961
DepartureTime_1     2.0107      0.590      3.408      0.001       0.854       3.167
DepartureTime_2     1.5135      0.543      2.789      0.005       0.450       2.577
DepartureTime_3     1.1001      0.694      1.584      0.113      -0.261       2.461
Airfare_1           0.9186      0.413      2.223      0.026       0.109       1.729
Airfare_2          16.5659   1.25e+04      0.001      0.999   -2.45e+04    2.45e+04
Airfare_3           1.5443      0.437      3.535      0.000       0.688       2.401
===================================================================================
```

For Airline, we can find that the log-likelihood is -118.98, and pseudo R-squared is 0.4557. The odds ratio of choosing a non-full service airline for trip for leisure is 29.93% more likely in comparison to a trip for another category. The odds ratio of choosing a non-full service airline for group travel is 39.29% less likely in comparison to a non group travel..

From our airline model, we conclude that customers with high income over 200M are more likely to choose a full service airline in comparison to low income customers and it is aligned with intuition. People are more likely to choose a full service airline when they fly to Japan and southeast Asia in comparison to China as a destination. Overall, the airline model provides less insight in comparison to our airport model even though the accuracy rate is still significantly high.

3.2 Decision Tree

Decision trees allow us to analyze fully the possible consequences of a decision. It provides a process to quantify the values of outcomes which can be illustrated by using different impurity measures. The model is built based on the values of the attributes from training and validation dataset. In current 484 observations, 70% of data are used for training dataset, 30% of total data are used for validation. The best model is selected by the model comparison operator using the validation dataset during the decision process. Finally, the quality of the performed classification can be studied on the test dataset. Then compare these models, we can interpret the passenger behavior base on Airport and Airline choice from the passenger survey.

3.2.1 Airport Choice

(1) Feature selection

In this session, Airport is set as the dependent variable which includes two attributes as a purpose of finding passenger choice of Incheon (ICN) or Gimpo (GMP) airport. For the independent variables, one model excludes all the insufficient and irrelevant variables which contain too many missing values as paper mentioned previously, another one also excludes all the insufficient variables and variable Assess Time and Income. The variable selection as below table 1 shown. In the following session, the process of developing the model and the corresponding explanation will be presented.

Table 1. Variables selection for each model

Airport Choice Model 1	Airport Choice Model 2
Airline	Airline
Gender	Gender
Nationality	Nationality
TripPurpose	TripPurpose
TripDuration	TripDuration
FlyingCompanion	FlyingCompanion
ProvinceResidence	ProvinceResidence
GroupTravel	GroupTravel
NoTripsLastYear	NoTripsLastYear
Destination	Destination
DepartureTime	DepartureTime
SeatClass	SeatClass
Airfare	Airfare
NoTransport	NoTransport
ModeTransport	ModeTransport
Occupation	Occupation
AccessTime	
Income	

(2) Model implementation and Evaluation

Two models are built in this session to explain the target variable Airport. Both models are using GINI measures of node impurity. Figure 3 displays the results of the model 1. We found that Destination and Departure Time clearly classify the group to describe the passenger choice. However, Access Time and Income didn't explain well in this model. The variables AccessTime show if the value of the attribute is larger than 1.5 will be classified as False. The attribute “2” in Access Time is an unknown value category. Similarly, the variables Income show if the value of the attribute is larger than 2.5 will be classified as False. The attribute “3” in Income is also an unknown value category. Hence, these unknown values will be split to be an individual node in the decision model, which is hard to interpret in a real-life situation since we couldn't guess or know the reason why the surveyee didn't respond to their access time of transportation and income in the survey.

As a result, we came up with a new model that exclude these two independent variables as figure 4 shown. In this model, we could observe a more reasonable classification based on the value of the attribute.

Figure 3: Model 1- Decision Tree Model with all Variables

Accuracy: 86%

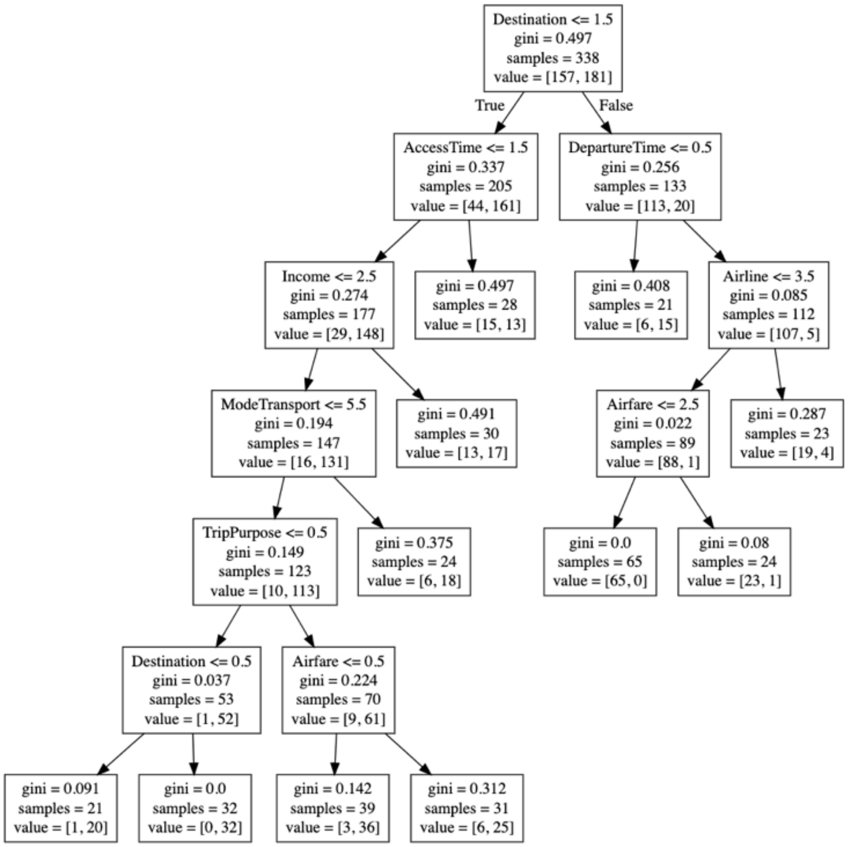
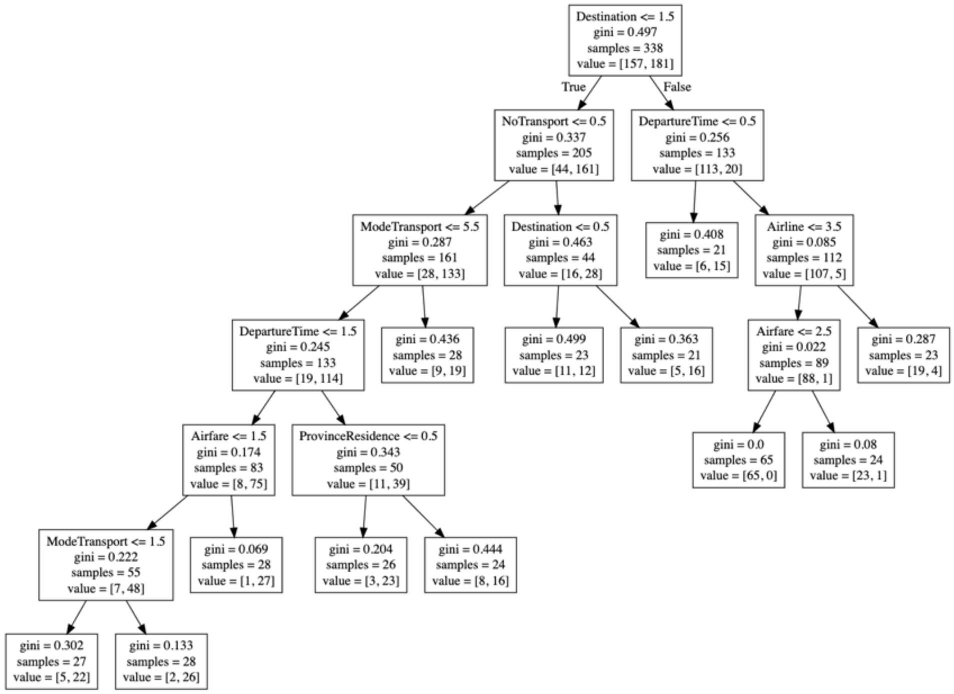


Figure 4: Model 2 - Decision Tree Model excludes Income and Access Time

Accuracy: 80%



(3) Interpretation

Both models start the root node of the Destination, then move down the tree. More notably, both show the three important components that determine airport choice, which are Destination, Departure Time, and Airline. They explain the passengers whose flight destination is South Asia or others and departure time is 6 am to 12 pm are more likely to choose Gimpo Airport. On the other hand, in Model 2, the result shows the passenger whose flight destination is South Asia or others, departure time is after 12 pm, and take the foreign airlines are more likely to choose Incheon Airport. Moreover, the passengers whose flight destination is China, and have only one number of transportation modes used to arrive at the airport are more likely to choose Gimpo Airport.

Overall, the destination of the fight and flight departure time plays a crucial role in choosing the airport. The possible explanation is probably because most of the flights depart from Gimpo airport are to Japan or China. If the flight is to South Asia and departs in the early morning, people from Seoul would need to choose a near one with less traveling time.

3.2.2 Airline Choice

(1) Feature selection

In the airline decision, we used the re-category dependent variable. Two different re-category groups are used in two different airline choice models. One is Korean Airlines such as Korean Air (KE), Asiana

Airlines (OZ) and Korean LCC versus the Foreign Airlines, another one is Korean LCC (Low-cost Carrier) versus full-service airline which will be noted as Non-LCC Airline in the following part.

For the independent variable, we are using the same variables from Airport model 2 and include the Airport in the airline model. Airfare will be excluded in independent variables in the second tree model of LCC airline due to this variable might have a joint effect on the dependent variable of the Airline. Because the airfare is decided directly by which kind of airline. Choosing low-cost carriers has decided that the airfare will be lower than other kinds of the airline. Therefore, it’s not legitimate to include Airfare in the model of choosing LCC or Non-LCC airline. The variable selection as below table 2 shown. In the following session, the process of developing the model and the corresponding explanation will be presented.

Table 2. Variables selection for airline models

Airline Choice Model 1	Airline Choice Model 2
Airport	Airport
Gender	Gender
Nationality	Nationality
TripPurpose	TripPurpose
TripDuration	TripDuration
FlyingCompanion	FlyingCompanion
ProvinceResidence	ProvinceResidence
GroupTravel	GroupTravel
NoTripsLastYear	NoTripsLastYear
Destination	Destination
DepartureTime	DepartureTime
SeatClass	SeatClass
NoTransport	NoTransport
ModeTransport	ModeTransport
Occupation	Occupation
AccessTime	AccessTime
Income	Income

(2) Model Implementation and Evaluation

Two models are built in this session to explain the target variable Airline. Both models are using GINI measures of node impurity. Figure 5 and Figure 6 display the results of model 1 and model 2 accordingly. Two models address a slightly different result. They start from different root node, and explain the different important variables on airline choice. Model 1 classifies the data start from Airport, Group Travel, Departure Time and Province Residence. Model 2 classifies the data start from Destination, Flying Companion, Departure Time, Airfare, and Mode Transportation.

Figure 5: Model 1 - Decision Tree Model of Korean Airline and Non-Korean Airline Choice
Accuracy: 71%

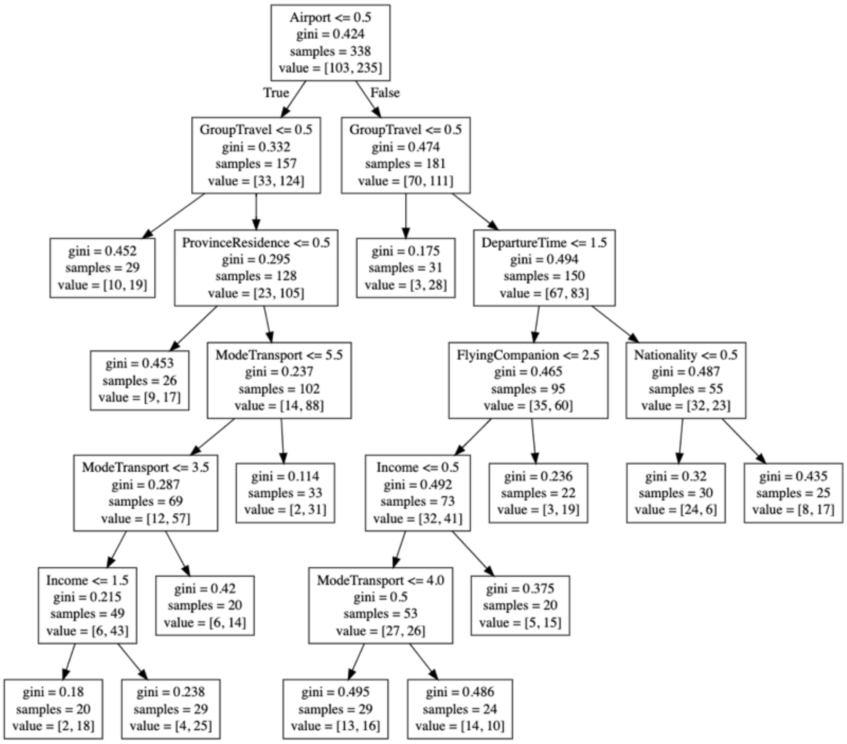
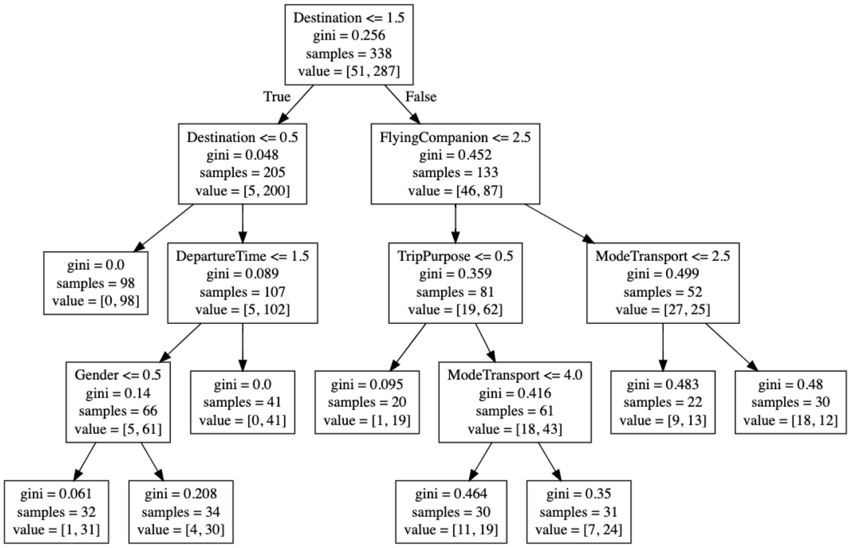


Figure 6: Model 2 - Decision Tree Model of LLC (Low-cost carrier) and Non-LLC Airline Choice

Accuracy: 82%



(3) Interpretation

Two models indicate the different factors that might impact the passenger behavior on airline choice. When the purpose is for finding passenger choice between Korean and non-Korean airlines. The outcome performs whether passenger travel with a group is an important component to affect the selection. In model 1, the result shows no matter which airport the passenger chose, if they travel with a group, they are more likely to take Korean Airlines. Also, when passengers choose Incheon Airport, travel with a group, and live near Kyungsang-do, Kyungsang-do or Kangwon-do area also more likely to choose Korean Airlines. The result is somehow intuitive. Most travel agencies will provide promotions or cheaper price of flight tickets for passengers since they collaborate with airline companies. In this dataset, nearly 80 % of surveyee are Korean. Most passengers travel with a group departing from Korean to choose Korean travel agencies, leading to the high possibility that they will take Korean airlines.

When the purpose is finding passenger choice between the low-cost carriers and full-service airlines. The model 2 performs the different factors that might impact the passenger behavior on airline choice. The outcome shows when the travel destination is China, passengers are more likely to choose full-service airlines instead of low-cost carriers. However, when the travel destination is Japan, and the departure time is after 6 pm, and before 6 am, passengers more likely to choose low-cost carriers instead of full-service airlines. Moreover, when the travel destination is China or Japan, travel with less than 3 people, and the purpose of travel is leisure are more likely full-service airlines.

3.3 Model Comparison

Airport Model: For "Airport" as the dependent variable model, we would prefer to choose logistic regression model, which has a higher accuracy. Also, the odds ratio differences between different categories are clear. Tree Model's interpretation on variables are also clear and straightforward in comparison to the logistics model, but the decision tree is prone to overfitting. In particular, one of the problems with our study is the limited data size and a decision tree model's overfitting can be exacerbated.

Airline Model: For "Airline" as the dependent variable model, we would prefer to choose decision tree model, even if the logit model has a higher accuracy. In logit model, many variables are not significant enough (i.e.: p-value is pretty large). So, the decision tree is more straightforward and suitable to explain the relationship between airline choice and other variables.

4. Conclusion and Recommendations

4.1 Comments and Conclusion

Conclusion: Based on our analysis, we conclude that foreign travelers tend to choose GMP airport. Female, leisure travel purpose, Japan destination and South Asia traveler or other occupation people (i.e.: occupation besides entrepreneur, senior management) are more likely to choose GMP airport. However, number of transportation modes used to arrive at the airport is more likely to choose INC airport. INC airport is more convenient for these travelers. People with pretty high income (i.e.: greater than 200 million) are more likely to choose INC airport. Departure time from 12 pm to 6 am, or people living in Seoul, Incheon, Kyungki-do, and Chungcheong-do are more likely to choose INC airport.

Limitation: data size in our model is fairly small. Logistics regression has too many categorical variables. There might be Collinearity problems in our variable selection process in the logistics model. Pseudo R-squared is biased high as we mostly focus on pseudo R-squared.

Comments: In the future, we would like to validate our results on a larger data set. If we have time series data, we can explore the seasonal change of people's airport and airline choice. Also, we can explore people's behavior change on choosing the airport and airline with time goes by.

4.2 Recommendations:

For GMP airport, we think they can develop more domestic airlines, which may help them expand service for more Korean travelers. Building subways between main cities and GMP airport, which can help more people reach to GMP airport faster and more conveniently. Also, GMP should explore more airlines based on business needs, which can help them serve more business travel and high income travelers. For INC airport, we think they can develop more foreign airlines to serve foreign travelers.