**Analysis of Passengers’ Airport and Airline Choice Behavior in**

**Seoul Metropolitan Area**

Hoàng Anh Nguyễn (Yumi), Kimin Lee, Sahira Zaheen, Yunjung Ham (Yunee)

[**Abstract**](#_db1df6w77yt0) **2**

[**I. Introduction**](#_kuk22zlu43ph) **2**

[**II. Data**](#_k2xxmu80a2eb) **2**

[II-1. Descriptive Stats](#_1q4kwn1brc3) 3

[**III. Model Implementation**](#_91rtc0d7d1y6) **9**

[**III-1. Decision Tree - Airport Choice Model**](#_pkmd0posy6gt) **9**

[III-2: Decision Tree Model - Airline Model 1 (KE, OZ versus others )](#_6z3ol9dggog9) 10

[IIIb-Decision Tree Model - Airline Model 2 (Korean LLC vs Other Airlines)](#_qwcoomo0qxm) 11

[**IV. Model Comparison**](#_9id73hfxi8v8) **16**

[**V. Conclusion and Recommendations**](#_iy1o0b7rzquj) **16**

# Abstract

This paper aims to analyze the passengers' choice of two international airports in the Seoul Metropolitan Area, Gimpo Airport and Incheon Airport, which are the two airports that serve as main gateways to Seoul, Korea.

Our objective is to develop models to describe passengers’ airline and airport choice behavior in the Seoul Metropolitan Area. To address the former, we performed Logit(Logistic regression) from discrete choice models and Decision Trees from data mining models. To further validate our results, we performed K-nearest neighbors (KNN) and Neural Networks to evaluate our results from Decision Tree models. Our analysis shows that \_\_\_\_\_\_We also found that\_\_\_\_

After the model evaluation on passengers’ behavior based on the modeling results, we came up with policy implications based on the quantitative analysis. The implications include... \_\_\_\_

# I. Introduction

Over the last decade, the air transport industry faced a continuous increase in the number of airline passengers and air travel and experienced noticeable growth. The International Air Transport Association (IATA) forecasts that the current trends in air transport “suggest passenger numbers could double to 8.2 billion in 2037” (IATA). With this increasing trend, the number of departure flights is also expected to increase over the next decade. Considering this trend, it’s important for the airports and airlines to consider a systematic approach to analysis to forecast the air traveler’s demand and decisions. It is also critical to understand and identify the traveler’s needs and different factors that affect its decision to develop strategic marketing and tailor airline industry’s offerings.

There are many different dimensions and areas that impact air traveler’s decisions. The earlier studies found non-price factors that impact the airport choice such as access time and access cost to the departure airport. An international air traveler survey in Japan showed that two of the most important factors are access time and flight frequencies for choosing the departure airport (Furuichi, M., Koppelman, 1994). Other important non-price variables include “airport delay, flight frequency, the availability of particular airport–airline combinations, and early arrival times were found” that strongly affected choice probabilities (Leon, 2011). Additionally, another finding includes that business travelers are less sensitive to airfare increase than leisure travelers, noting that airfare is not a primary factor in choosing the departure airport. By looking at these different factors, the main target of this research is to develop a modeling approach that explicitly accounts for the varying choice behavior of different groups of passengers.

# II. Data

The survey data is from 488 respondents from air travelers in Seoul Metropolitan Area, who departed either from the Gimpo Airport or Incheon Airport. The data includes 27 variables including the Airport Choice (GMP, ICN), Airline Choices(Korean Air (KE), Asiana Air (OZ), Korean LCC, Foreign Carriers), socio-demographic variables such as age, gender, occupation, income and other variables such as flight information, travel time, and the mode of transport. We can notice many missing fields due to the nature of survey data, privacy issues etc.

To begin our exploratory data analysis, we have removed ‘ID’ and ‘FlightNo’ as these were unique identifiers that will not provide useful information or values for our models. Additionally, we removed any variables with a high percentage of missing values. The ‘Mileage’ had the highest NaN values with roughly 82% of the total 488 observations. We also removed MileageArline and AccessCost that had more than 40% of missing data. We have also removed ‘FrequentFlightDestination’ as we decided to use ‘Destination’ variable to further evaluate our model, and ‘DepartureHr’ and ‘DepartureMn’ have been removed and to use the ‘DepartureTime’ variable.

## **II-1. Descriptive Stats**

**Quantitative variables - Summary**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | **TripDuration** | **FlyingCompanion** | **NoTripLastYear** | **DepartureMn** | **Airfare** | **AccessTime** | **AccessCost** | **Mileage** | **Income** |
| **Min** | 17.00 | 0.00 | 0.00 | 0.000 | 0.00 | 3.00 | 4.00 | 0 | 1 | 1.000 |
| **25%** | 29.00 | 4.00 | 1.00 | 1.000 | 15.00 | 35.00 | 25.00 | 2000 | 8258 | 2.000 |
| **50%** | 38.00 | 5.00 | 2.00 | 2.000 | 30.00 | 45.00 | 40.00 | 6000 | 27998 | 3.000 |
| **Mean** | 39.97 | 27.44 | 2.82 | 3.262 | 25.98 | 50.46 | 51.83 | 11220 | 56384 | 3.126 |
| **75%** | 50.00 | 8.00 | 3.00 | 3.000 | 40.00 | 60.00 | 60.00 | 12000 | 61808 | 4.000 |
| **Max** | 80.00 | 730.00 | 34.00 | 122.000 | 55.00 | 260.00 | 390.00 | 350000 | 500000 | 7.000 |
| **Na’s** | 1 |  |  |  | 120 | 155 | 97 | 197 | 398 | 132 |

**Descriptive Statistics for categorical variables**

**Categorical variables - Percentage Table**

**Airport**

|  |  |
| --- | --- |
| **ICN** | 48.98% |
| **GMP** | 51.02% |

**Airline**

|  |  |
| --- | --- |
| **Korean Air (KE)** | 32.01% |
| **Asiana Airlines (OZ)** | 22.38% |
| **Korean LCC** | 16.95% |
| **Foreign Airlines** | 28.66% |

**Group Travel**

|  |  |
| --- | --- |
| **Yes** | **17.42%** |
| **No** | **82.58**% |

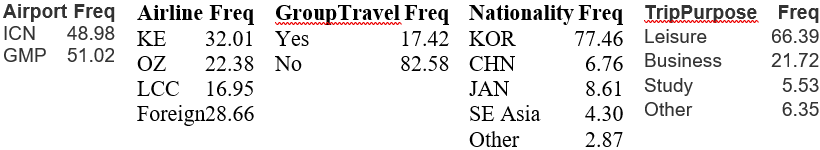
**Nationality**

|  |  |
| --- | --- |
| **Korean (KOR)** | 77.46% |
| **China (CHN)** | 6.76% |
| **Japan (JAN)** | 8.61% |
| **Southeast Asia (SE Asia)** | 4.30% |
| **Other** | 2.87 |

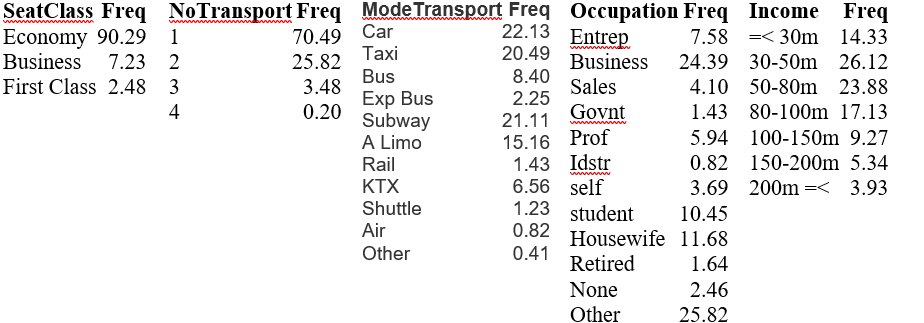
**Trip Purpose**

|  |  |
| --- | --- |
|  | **%** |
| **Leisure** | 66.39 |
| **Business** | 21.72 |
| **Study** | 5.53 |
| **Other** | 6.35 |

**ProvinceResidence**

****

****

****

**Descriptive Statistics for Trimmed Data**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | **Trip Duration** | **Flying Companion** | **No. of Trips Last Year** | **Airfare** | **Access Time** | **Income** |
| **Min** | **17.00** | **0.00** | **0.00** | **0.00** | **5.00** | **4.00** | **1.00** |
| **25%** | **29.00** | **4.00** | **1.00** | **1.00** | **33.00** | **30.00** | **1.00** |
| **Median** | **37.00** | **5.00** | **2.00** | **2.00** | **40.00** | **40.00** | **2.00** |
| **Mean** | **39.72** | **25.87** | **2.46** | **3.322** | **44.65** | **46.64** | **2.709** |
| **75%** | **50.00** | **8.00** | **3.00** | **3.00** | **52.00** | **60.00** | **4.00** |
| **Max** | **80.00** | **730.00** | **14.00** | **122.00** | **120.00** | **180.00** | **7.00** |

**Quantitative variables - Summary**

**Categorical variables - Percentage Table**

**Airport**

|  |  |
| --- | --- |
|  | **%** |
| **ICN** | **48.8** |
| **GMP** | **51.2** |

**Airline**

|  |  |
| --- | --- |
|  | **%** |
| **KE** | **30.85** |
| **OZ** | **22.54** |
| **LCC** | **17.72** |
| **Foreign** | **28.88** |

**Trip Purpose**

|  |  |
| --- | --- |
|  | **%** |
| **Leisure** | **67.18** |
| **Business** | **21.44** |
| **Study** | **5.47** |
| **Other** | **5.91** |

**Nationality**

|  |  |
| --- | --- |
|  | **%** |
| **KOR** | **77.46** |
| **CHN** | **6.78** |
| **JAN** | **8.10** |
| **Southeast Asia** | **5.60** |
| **Other** | **3.06** |

**Province Residence**

|  |  |
| --- | --- |
|  | **%** |
| **Seoul** | **39.39** |
| **ICN** | **4.60** |
| **Kyungki - do** | **26.26** |
| **Chungcheong-do** | **2.84** |
| **Kyungsang-do** | **5.47** |
| **Jeonra-do** | **1.75** |
| **Kangwon-do** | **1.00** |
| **Other** | **18.60** |

**Group Travel**

|  |  |
| --- | --- |
|  | **%** |
| **Yes** | **15.54** |
| **No** | **84.46** |

**Destination**

|  |  |
| --- | --- |
|  | **%** |
| **China** | **28.23** |
| **Japan** | **32.17** |
| **Southeast Asia** | **35.67** |
| **Other** | **3.94** |

**Departure Time**

|  |  |
| --- | --- |
|  | **%** |
| **6am - 12pm** | **9.63** |
| **12pm - 6pm** | **43.54** |
| **6pm - 9pm** | **40.70** |
| **9pm - 6am** | **6.13** |

**Seat Class**

|  |  |
| --- | --- |
|  | **%** |
| **Economy** | **90.59** |
| **Business** | **7.00** |
| **First Class** | **2.41** |

**Number of Transport**

|  |  |
| --- | --- |
|  | **%** |
| **1** | **71.55** |
| **2** | **25.16** |
| **3** | **3.06** |
| **4** | **0.22** |

**Mode of Transport**

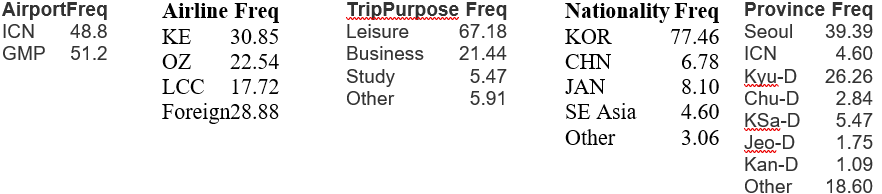
|  |  |
| --- | --- |
|  | **%** |
| **Car** | **22.10** |
| **Taxi** | **20.13** |
| **Bus** | **8.10** |
| **Express Bus** | **1.97** |
| **Subway** | **21.23** |
| **Airport Limousine** | **15.75** |
| **Rail** | **1.31** |
| **KTX (Express Rail)** | **6.78** |
| **Shuttle provided by travel agency** | **1.31** |
| **Air** | **0.88** |
| **Other** | **0.44** |

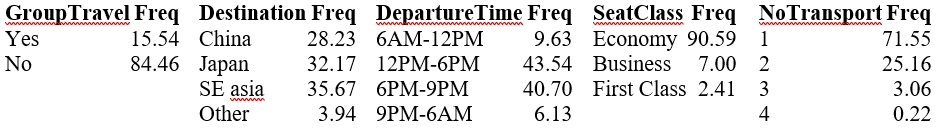
**Occupation**

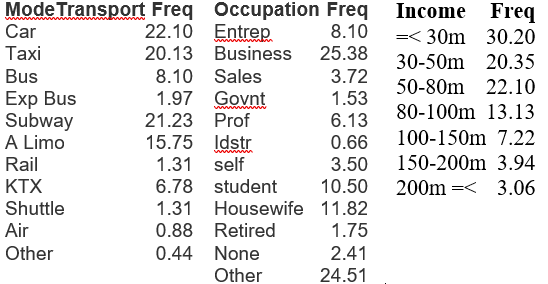
|  |  |
| --- | --- |
|  | **%** |
| **Entrepreneur, Senior management** | **8.10** |
| **Business (Corporate worker)** | **25.38** |
| **Sales, Service** | **3.72** |
| **Government, Military** | **1.53** |
| **Professionals (doctor, lawyer, professor)** | **6.13** |
| **Industrial, manufacturing** | **0.66** |
| **Self-employed** | **3.50** |
| **Student** | **10.50** |
| **Housewife** | **11.82** |
| **Retired** | **1.75** |
| **None** | **2.41** |
| **Other** | **24.51** |

**Income**

|  |  |
| --- | --- |
|  | **%** |
| **30 Million Won or less** | **30.20** |
| **30~50 Million Won** | **20.35** |
| **50~80 Million Won** | **20.10** |
| **80 ~100 Million Won** | **13.13** |
| **100 ~150 Million Won** | **7.22** |
| **150 ~200 Million Won** | **3.94** |
| **200 Million Won or more** | **3.06** |

****

****

****

**1. Airport**: Our processed data has 48.8% Incheon and 51.2% Gimpo airport. The raw had almost the same combination of two airports.

**2. Airline**: The trimmed data contains 30.85% Korean Air, 22.54% Asiana, 17.72% LCC, and 28.88% foreign airlines. While there was a slight loss in percentage for all three airline categories, Korean Air increased by approximately 1.5% in the data processing.

**3. Nationality**: The majority of the dataset is Korean 77.46% followed by Japanese 8.1%, Chinese 6.78%, Southeast Asian 4.60% and others 3.06% in our trimmed data. There was only tiny changes in percentages for all Nationality values by data trim.

**4. Trip Purpose:** The processed data has 67.18% leisure, 21.44% business, 5.47% study, and 5.91% other purpose of trip. The data is dominated by travellers for leisure purpose.

**5. Province Residence:** Large portion of airport users in our trimmed data reside in Seoul (39.39%) and Kyungki-do (26.26%) make up 65.65% of the total data, followed by Others (18.60%) and the rest ( 15.75% total from all provinces except Seoul, Kyungki-do and Others).

**6. Group Travel:** Only 15.54% of airport users in our trimmed dataset are using a package tour while the majority 84.46% users are travelling individual

**7. Destination:** 35.67% of the respondents have a destination to Southeast Asia, 32.17% to Japan, 28.23% to China, and 3.94% to others.

**8. Departure Time:** Most respondents travel 12pm - 6pm (43.54%) and 6pm -9pm (40.70%). Only a small portion of the respondents travel 6am -12 pm (9.63%) and 9pm-6am (6.13%).

**9. Seat Class:** Economy class (90.59%) carry weight of the trimmed data. Business class and first class take only 7% and 2.41% respectively.

**10. Number of Transport:**  Our data shows that the majority of respondents had 1(67.18%) to 2 (21.44%) transports to the airport.

**11. Mode of Transport:** Four transport modes including car (22.10%), taxi(20.13%), subway(21.23%), and airport limousine (15.75%) are the most popular. The other seven transport types take only 19.48% of our data.

**12. Occupation:** The top four occupations of respondents are business 25.38%, other 24.51%, housewife 11.82%, and student 10.50%. The rest occupations sum up to 27.79% of the trimmed data.

**13. Income:** Only 14.22% of the trimmed data is equal or more than 100 million won. The largest group is the one having income equal or less than 30 million won at 30.20%, followed by the group with income between 50 - 80 million won at 22.10% and between 30-50 million won at 20.35%. The average income is 2.709 which falls somewhere between 40 -50 million won, and the minimum income is below 30 million won with maximum income is over 200 million one while the distribution is right skewed.

**Quant**

**14. Age:** The mean age is 39.72 years old with minimum age 17 and maximum age 80 years old. The median age is 37, 25% of respondents are younger than 29 years old and 25% respondents are older than 50 years old in our trimmed data.

**15. Trip Duration:** In our data, the average trip duration is 25.87 days while 25% has four or less days of trip and 75% has five or less trip days. The maximum trip day is 780 which greatly increased the average trip days (can be an outlier?) with the minimum trip duration of 0 while the median is five days.

**16. Flying Companion:** The minimum number of flying companions is 0 with the maximum companion number of 14. The data indicates that 25% of total interviewee has one or no flying companions and 75% of the total has three or less than three companions in their flights. The average flying companion number is 2.46 with a median of 2.

**17. Number of Trips Last Year:** The distribution is greatly right skewed by the maximum of 122 with the minimum of 0. The average number of trips in the last year is 3.222 with median of 2. The first and third quartile are 1 and 3 respectively.

**18. Airfare:** The minimum airfare is 5 and the maximum is 120 with the first quartile of 33 and the third quartile of 60. The average airfare in our process data is 46.64 with the median of 40.

**19. Access Time:** On average, the examinees have 46.64 minute to access the airport with median access time of 40. The shortest access time is 4 and the longest access time is 180 minutes. 75% of the interviewees access to the airport 60 minutes or earlier, and 25% of total have 30 or less minutes to get to the airport.

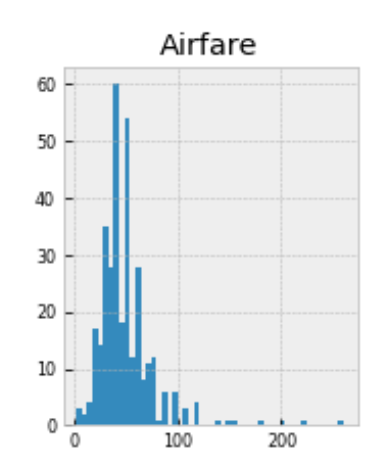
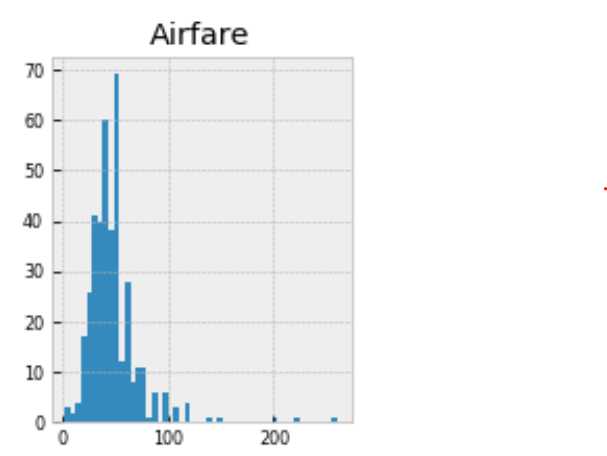
**Replacing Average Airfare**

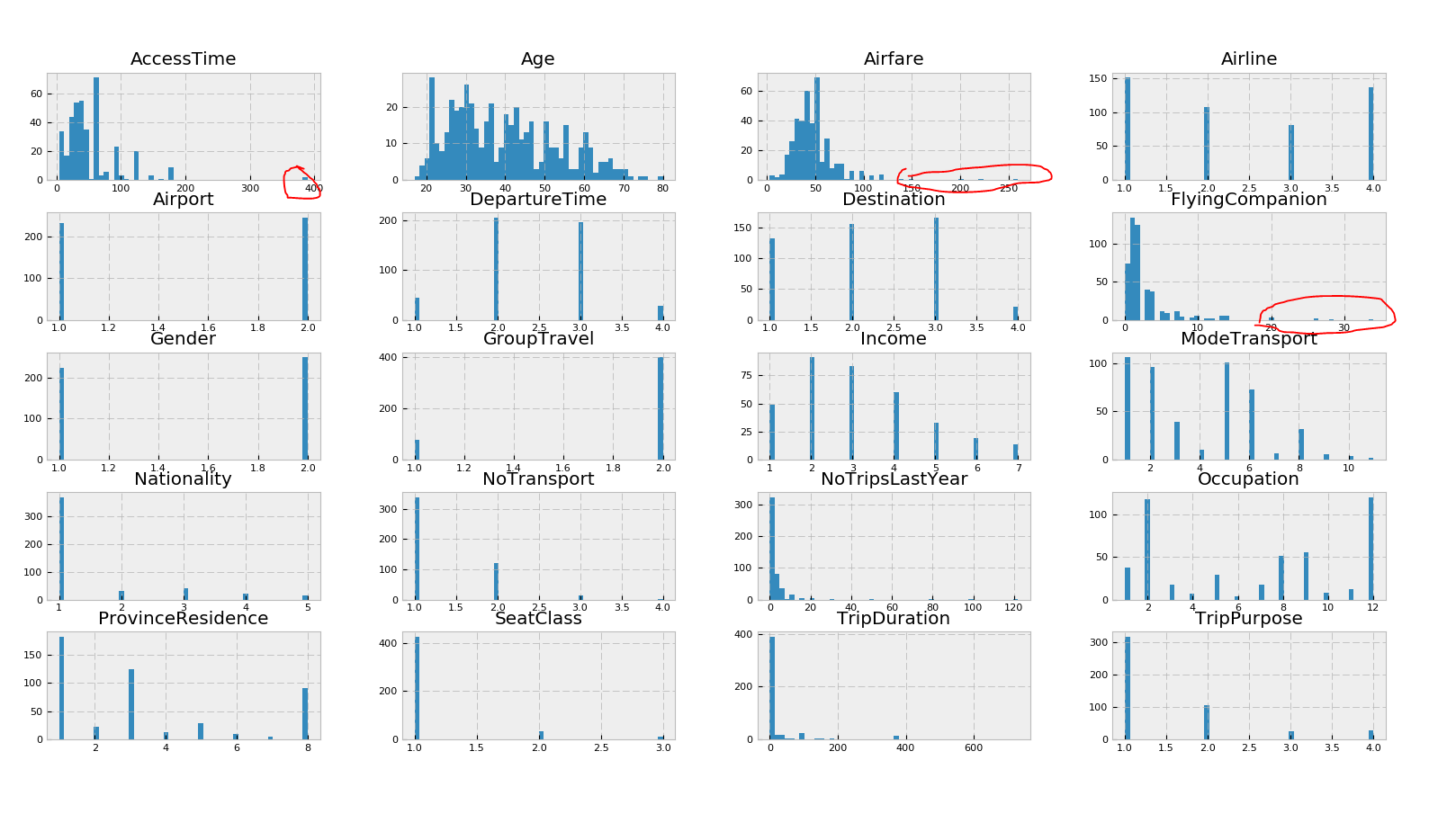
There were variables we wanted to further evaluate to see if it should be included in our sample data. The first variable is ‘Airfare’, where the raw data included 155 missing fields out of the total 488 fields. The missing fields were about 32%of the total fields, and we decided to reconsider this variable by replacing the missing fields with average airfare costs from our reference data.

Before replacing the average values, we have dropped NA values from ‘Airline’, ‘Departure’, and ‘Destination’ variables, which led to 149 missing values in the ‘Airfare’ variable out of 476 fields. After observing our reference data, we have discovered that not all of the average airline costs were provided (ex. Foreign Carriers). Although not all of the airline’s average airfares were provided, we were still able to reduce the missing fields by 50%, from 155 to 81. We now have 17% of missing fields in the Airfare variable, which we have reduced by half. However, we have to be cautious when we are adding average values into the missing fields as this could skew our data.

Filling in missing values should be done cautiously, and in order to make sure our data is not skewed, we have carefully observed the histogram of the Airfare variable, comparing before and after average values were replaced. We have also compared the descriptive statistics of the Airfare variable, such as mean, standard deviation and the quartiles. An ideal result we expected was to obtain distributions that are similar in shape.

|  |  |  |  |
| --- | --- | --- | --- |
| **‘Airfare’** Descriptive stats **before** replacing NaN values with Average Airfares | | **‘Airfare’** Descriptive stats **after** replacing NaN values with Average Airfares | |
| **Count** | 333 | **Count** | 395 |
| **Mean** | 50.456 | **Mean** | 48.332 |
| **Median (Q2)** | 45 | **Median (Q2)** | 45 |
| **Q3** | 60 | **Q3** | 53 |

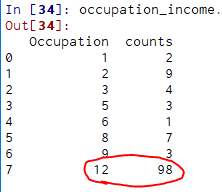
**Original data - ‘AirFare’** **After replacing Average Airfare: **

After examining a histogram of the data to check for skewing of the curve, we have found that both graphs have similar distributions. Also, the descriptive stats yield similar results before and after; mean values were very close at 50.456 and 48.332, and the median airfare stayed the same at 45. Therefore, we have decided to keep the Airfare variable with average values filled in, although not all airline’s average fares were available. As a result, we have reduced the missing values by 47% for the airfare variable (155 values to 81). 

To continue, we have observed the rest of histograms to further evaluate outliers and NA values in the remaining variables. Based on our histograms, we have found outliers in the Airfare variable over 1,250,000 Won, FlyingCompanion with over 20 people, and AccessTime with more than 350 minutes and removed these outliers.

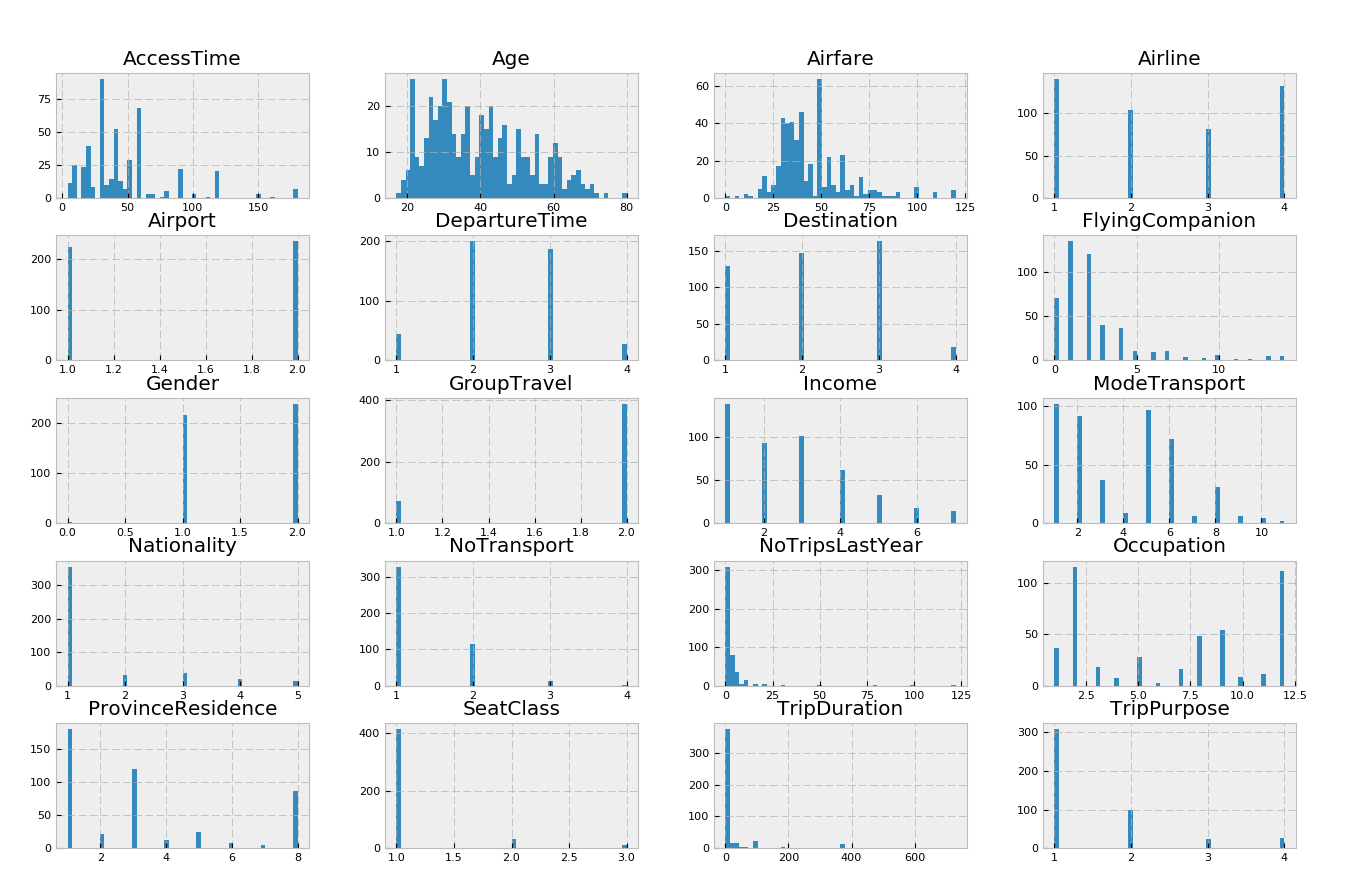
\

**Processing of filling in NA for Income, Airfare, and AccessTime:**



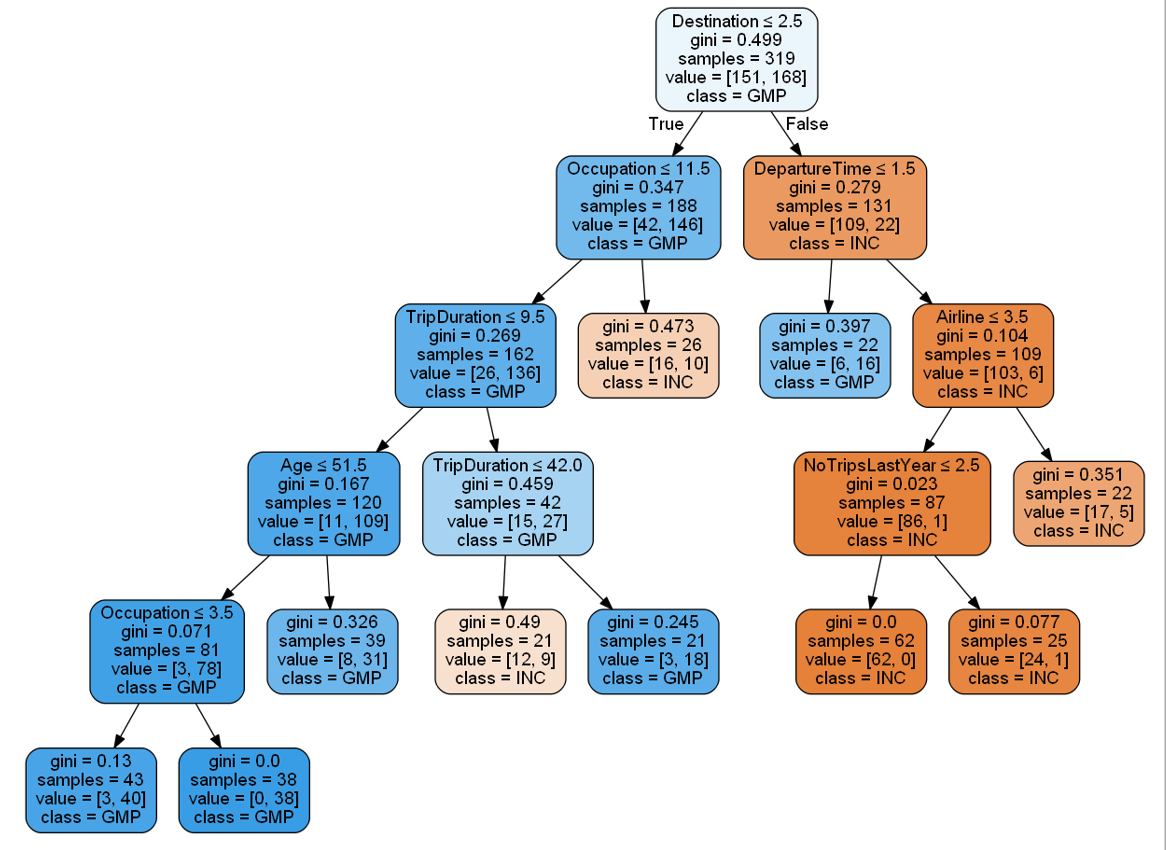
Since the majority of the missing income is under occupation 12, the rest are not significant and also 5,8,9 occupation groups have a decent size of the population as well. We processed to group by occupation and get the average income for each of the groups to fill in our NA for Income.

For the remaining Airfare NA, we take the average by grouping by Airport, Airline, and Destination.   
To fill NA for AccessTime, we group by Airport and ProvinceResidence.

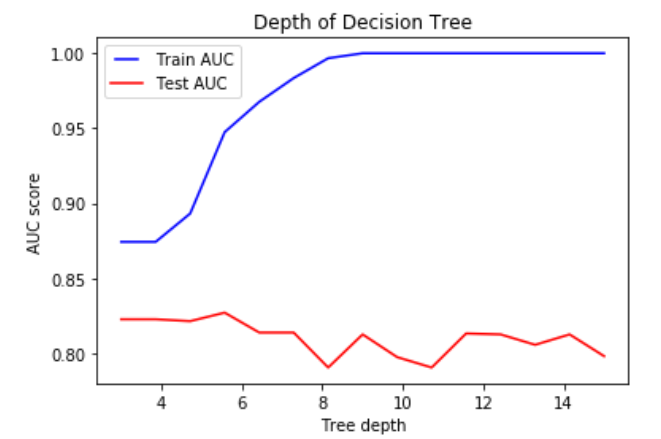
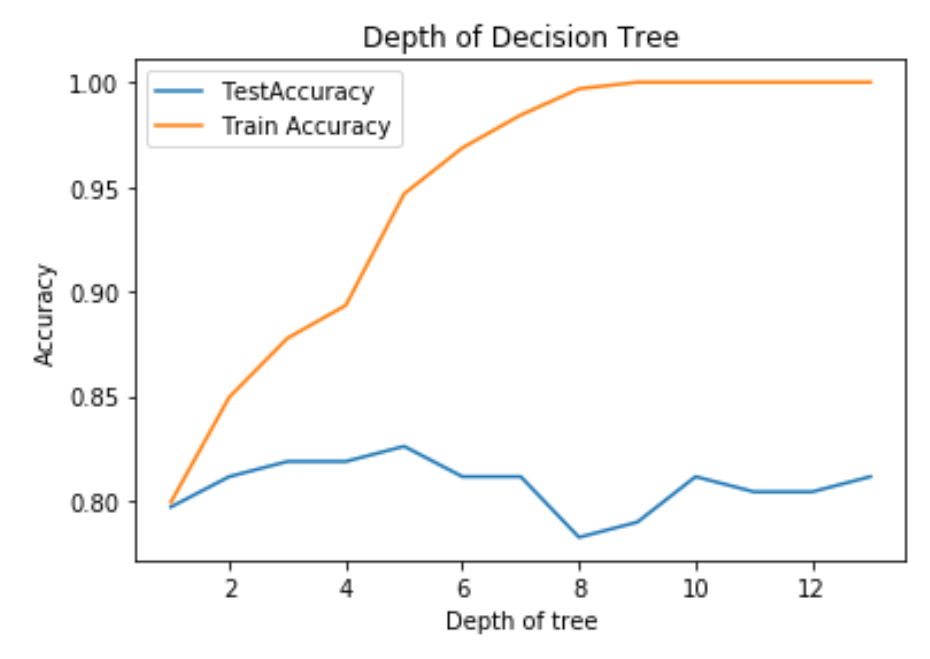
Above are the histograms after removed outliers and fill in NAs for all features.We saw that Airfare still has one 0 observation. Thus, we removed it from our data. Final data has 457 (93%) observations after cleaning. 

# III. Model Implementation

# **III-1**: **Decision Tree** - Airport Choice Model



III-1 Figure 1: Decision Tree Results for the Airport Choice Model

III-1 Figure 2: AUC Comparison (Train vs Test) - Airport model III-1 Figure 3: Accuracy Rate (Train vs Test) - Airport model

For the initial evaluation of our decision tree model, we have compared both of the AUC(Area Under Curve) and accuracy rates for our test and train data for determining the max depth of our tree model. These are useful metrics for binary classification models and can check for overfitting and underfitting of our model. Our initial model with default decision tree parameters resulted in a total of 9 depths as our baseline. For our testing, we have decided to split our data into 70% training and 30% for testing using our cleaned data. Looking at the result of our AUC and accuracy comparison graphs, we can notice that our training data’s accuracy increases while that’s not the case for the testing data, and we can notice overfitting in the data as we increase the tree depth (refer to tables III 2 and 3 below). We can also notice how the accuracy increases until you reach the max depth of 5 for our testing data, then it starts to decrease, also noticing overfitting in our data. Therefore, we have chosen the max depth of 5 for our decision tree model. Max depth above 5 didn’t have much impact on how our tree was expanding and the decisions stayed similar. For minimum samples in splitting the nodes, we have tested min\_samples\_lef for 10, 20 and 30 and compared the model results. To capture more generalized behavior of the air travelers, we have decided to prune the tree using a higher sample size of 30, and decided to use this for all other decision tree models.

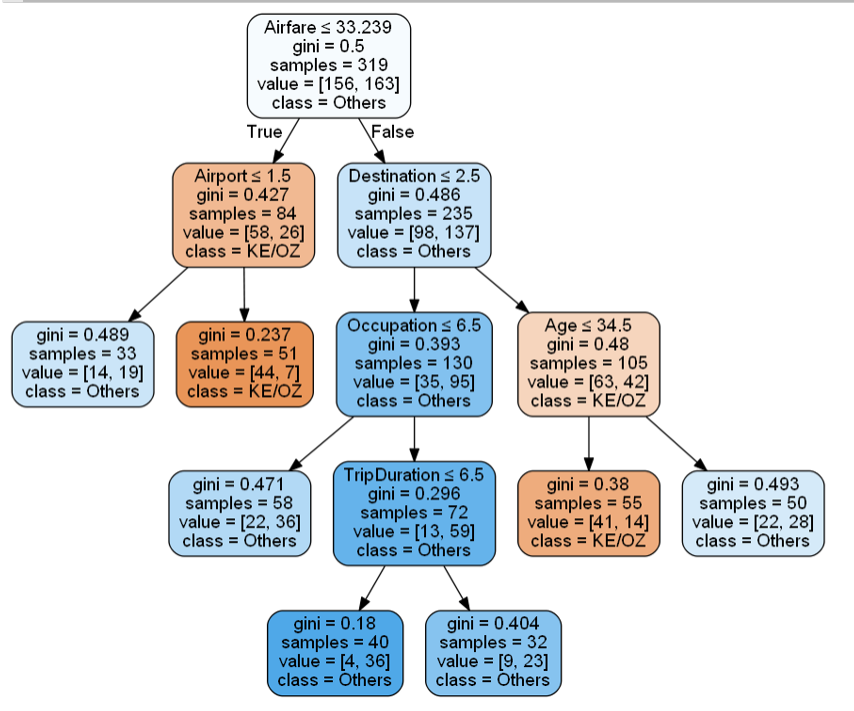
We can start evaluating our decision tree model at the root node, which is the best predictor of our model that asks whether the destination is less than 2.5. Given our data, anything less than 2.5 includes 1. China 2.Japan and it’s classified as the Gimpo Airport. Out of the 319 samples, 151 samples fall into Gimpo Airport, and 168 samples fall into Incheon Airport, which is a comparable amount of split for both airports. Looking at the results of our decision tree model, we can see that the majority of the left side of the blue nodes are classified as Gimpo Airport and right side of the orange nodes as Incheon Airport. Therefore, the air travelers heading to Japan or China are most likely to choose Gimpo Airport, and air travelers going to 3. Southeast Asia and 4.Other destinations will likely choose Incheon airport. Therefore, from this set of population, destination plays an important role in consumer’s airport choice.

Furthermore, Gimpo Airport currently supports “58 daily flights...to foreign destinations; Tokyo (Haneda), Shanghai, Beijing, Nagoya, Osaka and Taipei,” which supports the decision of our root node with Gimpo Airport’s main destinations to China and Japan. Also, with Incheon airport being the main hubs for major Korean Airlines and a bigger airport with more flights offered to other destinations, it supports the split made in the destination node. For destinations to China and Japan, occupation plays a role in choosing the incheon airport as we can see in the 2nd and 3rd depth of our tree. On the other hand, for destinations to Southeast Asia and other areas, departure time is a deciding factor for choosing Gimpo Airport.

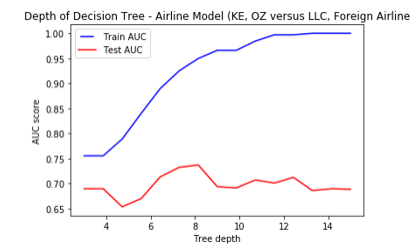
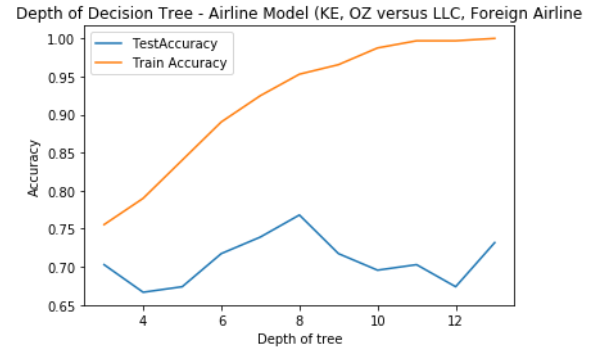
III-1 Figure 4: Confusion Matrix for Decision Tree Results - Airport Choice Model

|  |  |
| --- | --- |
| Accuracy | 81.20% |
| Recall | 86.40% |
| Test Error Rate | 18.80% |

## **III-2: Decision Tree Model - Airline Model 1 (KE, OZ versus others )**



III-2 Figure 1: Decision Tree Results for the Airline Choice Model 1

** **

III-2 Figure 2: AUC Comparison (Train vs Test) - Airline Model 1 III-2 Figure 3: Accuracy Rate (Train vs Test) - Airline Model 1

For the Arline Model 1, we have grouped the major Korean airlines together, Korean Airline (KE) and Asiana Airline (OZ), versus the rest of airlines. Looking at the result, we can see that the lower airfare (lower than 33.239) and the Airport choice (INC) is the decision factor of choosing the major Korean airlines. We can also see that as we go down the node to the right (higher airfare), destination and age are the deciding factors for choosing the major Korean Airlines.

To further evaluate our deciding factors, we have compared the gini index for the two decisions and factors leading to each decision. We can see that the airfare and airport (left side of our tree) is more likely to be correct as it holds more weight than the airfare, destination and age variables. Therefore, lower airfare and airport choice (INC) have impact on the air traveler’s choice on the major Korean Airliners (KE, OZ). In fact, Asiana Airlines’s main hub is at the Seoul Incheon International, as well as the Korean Airlines, which reflects the decisions made in the nodes.

In addition, we have also looked at the AUC for test and training data to determine the optimal size of our max\_depth for pruning our tree model. When comparing the AUC and accuracy rate table, we can see a similar pattern where the test data increases along with the train data but starts to decrease at max\_depth = 8. Therefore, we have set our *max\_depth* to 8, and kept the same minimum sample sizes as our decision tree for the airport choice model.

III-2 Figure 4: Confusion Matrix for Decision Tree Results - Airline Choice Model 1 (KE & OZ Airlines vs Others)

|  |  |
| --- | --- |
| Accuracy | 67.40 % |
| Precision | 75.30% |
| Recall | 86.40% |
| Test Error Rate | 32.60% |

## **III-3: Decision Tree Model - Airline Model 2 (Korean LLC vs Other Airlines)**

## 

## III-3 Figure 1: Decision Tree Results for the Airline Choice Model 2

## 

III-3 Figure 2: AUC Comparison (Train vs Test) - Arline model 2 III-1 Figure 3: Accuracy Rate (Train vs Test) - Airline Model 2

For the Arline Model 2, we have grouped Kroean LLC Airline versus other airlines. Looking at the result, we can notice that *Destination* and *Airfare* are important deciding factors for choosing the Korean LLC Airline.

In comparison, looking at the right side of our tree model, we can see that the *Destination, FlyingCompanion* are the two most important deciding factors for choosing other airlines versus the Korean LLC.

For this second airline model, we also evaluated AUC and accuracy rates for test and training data to determine the optimal size of our max\_depth for pruning our tree model. As a result, we saw a similar pattern where the test data increases along with the train data but starts to decrease around at max\_depth = 7. Therefore, we have set our *max\_depth* to 7, and kept the same minimum sample sizes as our decision tree for the airport choice model. Increasing the depth of our tree will result in overfitting in our data.

III-3 Figure 4: Confusion Matrix for Decision Tree Results - Airline Choice Model 2 (Korean LLC vs Others)

|  |  |
| --- | --- |
| Accuracy | 81.20 % |
| Recall | 28.60% |
| Precision | 35.30% |
| Test Error Rate | 18.80% |

**IIIc:** Logistic Regression:

Our response variable is set to be binary, the expected result is more appropriate to be a curve instead of straight regression line. Thus, we chose to apply logistic models for our study. We then use AIC and BIC scores to compare between models. These scores help choose the best subsets regression. AIC estimates a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model (Dziak, 2019).

1. **Airport choice:**

For describing the customer’s airport choice behavior between GMP (class 1) or ICN (class 0), we built two logistic models using Nationality, Destination, Airline, DepartureTime, and AccessTime (logit model 1) and NoTransportation (logit model 2) as independent variables. This set of variables

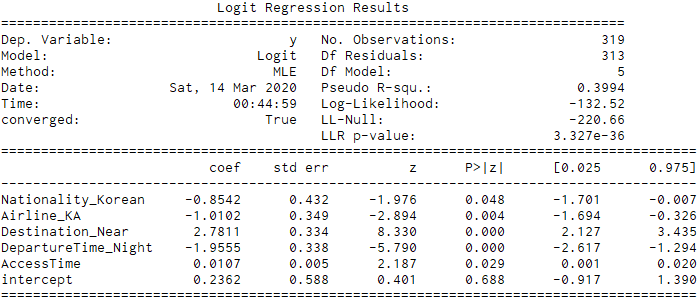


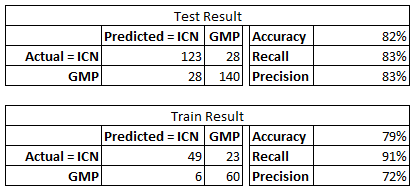
We splitted our 457 observations into train and test data. We picked a set of 319 random observations for the train data and used the rest for testing.

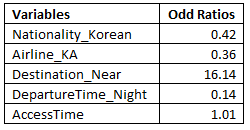
While choosing an airport multiple factors come into play. For different individuals the scale of these variables vary. However given our data, some common factors which are affecting the decision most are as followed.

Expats usually favour international airport over others as it harbors more flights for various combinations of oversea destinations. Hence, considering nationality was important to include in the model. Then airlines provide flights for different routes/destinations and consequently Incheon(ICN) and Gimpo(GMP) do not share the same airline list, which is why airline and destination would be very significant factors for choosing an airport. Usually while travelling, departure time plays an important factor as well for the passengers in order to maintain their schedules and that may affect on the respondents' airport selection as each airport may carry different flight schedules. Lastly, mobility to the airport is also a factor for passengers to choose an airport.

Airport Logit Model 1 with AccessTime:



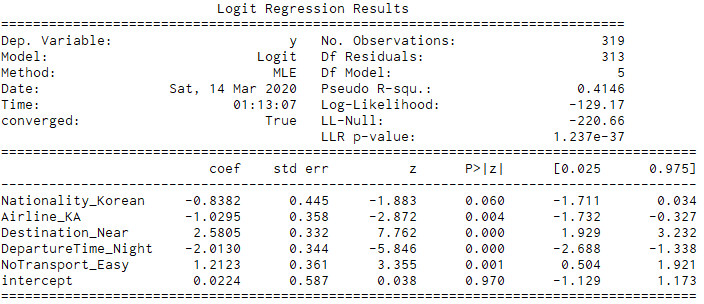
AIC: 277 BIC: 299.6



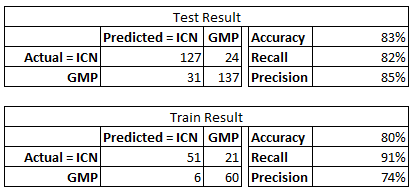
The Destination (Japan or China vs. others), whether Korean airlines or Foregn airline, and DepartureTime (redeye or regular) are the significant factors that affect customer choice behavior at 1% significance level. Nationality as Korean or non Korean, and AccessTime are significant at 5% level.

Holding all others constant, the odd ratio of Nationality\_Korean is 0.42 meaning that being Korean has the odds of choosing GMP airport is 0.42 as large as non Korean odds of choosing GMP airport. In other words, the estimated odds of choosing GMP airport are 58% lower for Korean Nationality. Given the knowledge that the majority of GMP flights are going to Japan or China, Destination plays a very important part in a customer’s airport choice. As we can see, estimated odds of choosing GMP airport are 1600% higher for customers going to China or Japan than who are going elsewhere. Although AccessTime is

Airport Logit Model 2 with NoTransportation:



AIC: 270.3 BIC: 292.9





The Destination\_Near (Japan or China vs others), and DepartureTime\_Night (redeye flight or regular) are the significant factors that affect customer choice behavior at 1% significance level. Airline\_KA is significant at 5% level and Nationality\_Korean is significant at 10% level.

The odd ratio of Nationality\_Korean is 0.43 meaning that being Korean has the odds of choosing GMP airport is 0.43 as large as non Korean odds of choosing GMP airport controlling for all other variables. In other words, the estimated odds of choosing GMP airport are 57% lower for Korean Nationality. Given the knowledge that the majority of GMP flights are going to Japan or China, Destination plays a very important part in a customer’s airport choice. As we can see, estimated odds of choosing GMP airport are more than 1300% higher for customers going to China or Japan than who are going elsewhere. The odd ratio of NoTransport\_Easy is 3.36 meaning that the number of transport being one has the odds of choosing GMP is 336% higher than the number of transport being more than one.

The more transportations an air traveler has to use to get to the airport the more likely they will have to spend more time on the road to get there. For the 2nd model above, we can see by substituting AcessTime with NoTransport\_Easy, all other variables stayed significant in the model while reducing both AIC and BIC scores. In addition, NoTransport\_Easy is a lot more significant compared to AccessTime and also has bigger magnitude. We can say that model 2 is closer to the true underlying reasons and most likely be the true model.

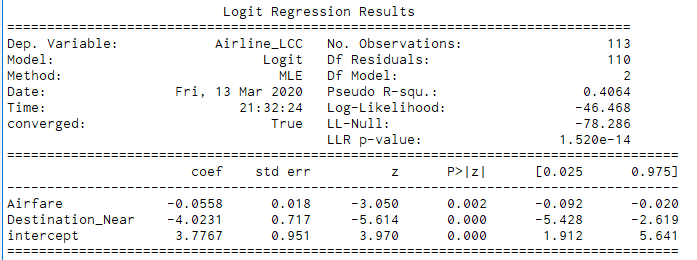
1. **Airline choice:**

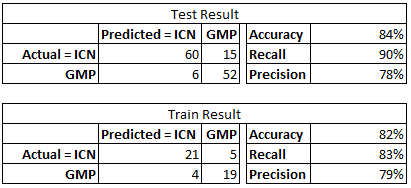
All the airlines are not shaped by the same set of facilities. These facilities may matter to some passengers and may not. However logically, there are some common factors that probably come into play every now and then while choosing an airline. Among which, age is a crucial factor i.e. a large portion of travellers over 35 years old tend to go for airlines with great services and facilities. They prefer comfort over anything whereas young people put relatively more focus on other factors. Varying on the capacity of airports, they operate flights and most of the time do not share the same list of airlines. Hence, airports also affect the decision of choosing an airline. Generally passengers are price sensitive and choose an airline which offers a flight within their budget, hence airfare is an important factor to include in the model. Over and above, destination is always the first filter for the passenger in order to choose an airline and as not all airlines always operate to the same destination, it proved to be significant for the model.

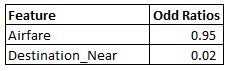
We developed two models: LCC vs. Others and Big Korean Airlines vs. Others.

For the first model where we want to determine what drives customers to choose the low cost airline instead of others. Since there is only a small number of observations for LCC, we took a sample size of 81 for customers who chose different airlines than LCC. We ran our model on the dataset of 162 observations with 81 customers who chose LCC (as class 1) and 81 is a random sample of customers who chose the other three airlines (as class 0).

LCC Logit Model:



AIC: 98.9 BIC: 107.11



# 

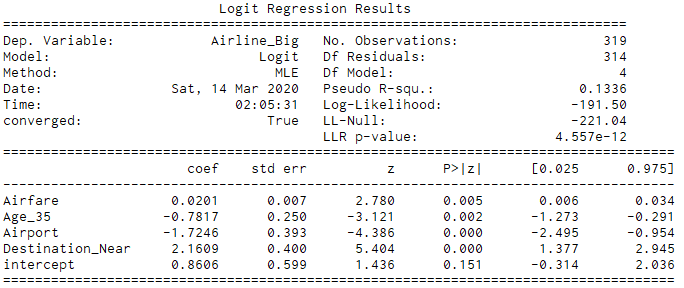
# 

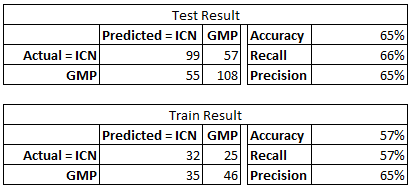
# 

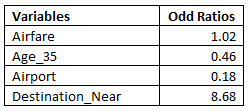
The Destination (Japan or China vs other) are the significant factors that affect customer choice behavior at 1% significance level. Airfare affects customer’s choice at 5% significance level.

Given LCC is a low cost airline which offers affordable travel tickets to customers, for every won increase in Airfare, the odds of the customer choosing LCC airline deceased by 5%. The estimated odds of choosing GMP are 98% lower among customers whose Destination is China or Japan holding all other constant.

Big Airlines Logit Model:



AIC: 393 BIC: 411.8



The Destination (Japan or China vs other), and Airport ( GMP or INC) are the significant factors that affect customer choice behavior at 1% significance level. Airfare and Age whether below 35 or not affect customer’s choice at 5% significance level.

The two largest Korean airlines are Korean Air and Asiana, with strong reputation, we expect to see slightly higher airfare rates for this group. Based on our model, for every won increases in Airfare the customer’s odds of choosing Korean Air or Asiana increases by 2%. In addition, older people tend to choose the better reputation airlines results in the estimated odds of choosing big name airlines are 54% lower among customers who are 35 years old or lower.

# IV. Model Comparison

# 

# V. Conclusion and Recommendations

**VI. References**

Dziak, John J., et al. "Sensitivity and specificity of information criteria." bioRxiv (2019): 449751.

Furuichi, M., Koppelman, F.S., 1994. An analysis of air travelers’ departure airport and destination choice behavior.

Transportation Research Part A 28, 187–195.

Skinner, Robert E. "Airport choice--an empirical study." *Journal of Transportation Engineering* 102.TE4 (1976).