****

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

**Seoul Metropolitan Area**

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

**Table of Contents**

[I.Introduction](#_kuk22zlu43ph) 2

[II. Data](#_k2xxmu80a2eb) 2

[III. Model Implementation](#_91rtc0d7d1y6) 7

[IV. Model Comparison](#_9id73hfxi8v8) 16

[V. Conclusion and Recommendations](#_iy1o0b7rzquj) 17

# 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 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.

## **Descriptive Stats**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Descriptive Statistics for continuous variables | | | | | | | | |  | |
|  | **Age** | **Trip Duration** | **Flying Companion** | **# of Trips Last Year** | **Airfare** | **Access Time** | **Income** | **Age** | |
| **Min** | 17.00 | 0.00 | 0.00 | 0.00 | 5.00 | 4.00 | 1.00 | Min | |
| **25%** | 29.00 | 4.00 | 1.00 | 1.00 | 33.00 | 30.00 | 1.00 | 25% | |
| **Median** | 37.00 | 5.00 | 2.00 | 2.00 | 40.00 | 40.00 | 2.00 | Median | |
| **Mean** | 39.72 | 25.87 | 2.46 | 3.322 | 44.65 | 46.64 | 2.709 | Mean | |
| **75%** | 50.00 | 8.00 | 3.00 | 3.00 | 52.00 | 60.00 | 4.00 | 75% | |
| **Max** | 80.00 | 730.00 | 14.00 | 122.00 | 120.00 | 180.00 | 7.00 | Max | |

**Categorical variables - Percentage Table**

|  |  |
| --- | --- |
| Airport | |
| ICN | 48.80% |
| GMP | 51.20% |

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

|  |  |
| --- | --- |
| 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% |

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

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

|  |  |
| --- | --- |
| Trip Purpose | |
| Leisure | 67.18% |
| Business | 21.44% |
| Study | 5.47% |
| Other | 5.91% |

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

|  |  |
| --- | --- |
| 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% |
| 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% |

|  |  |
| --- | --- |
| Providence Residence | |
| Seoul | 39.39% |
| Incheon | 4.60% |
| Kyungki-do | 26.26% |
| Chungcheong-do | 2.80%` |
| Kyungsang-do | 5.47% |
| Jeonra-do | 1.75% |
| Kangwon-do | 1.00% |
| Other | 18.60% |

|  |  |
| --- | --- |
| Airline | |
| KE (Korean Airline) | 30.85% |
| OZ (Asiana) | 22.54% |
| Korean LCC | 17.72% |
| Foreign Airlines | 28.88% |

|  |  |
| --- | --- |
| Nationality | |
| Korean | 77.46% |
| Chinese | 6.78% |
| Japanese | 8.10% |
| South East Asian | 5.60% |
| Other | 3.06% |

|  |  |
| --- | --- |
| Number of Transport (Number of transportation modes used to arrive to the airport) | |
| China | 28.23% |
| Japan | 32.17% |
| Southeast Asia | 35.67% |
| Other | 3.94% |

|  |  |
| --- | --- |
| Mode of Transportation | |
| 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% |

**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 were 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 travelers 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.

**Quantitative Variables**

**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.

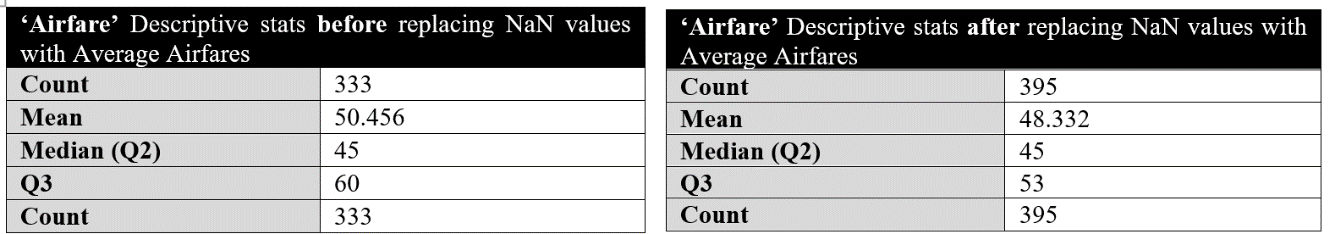
**19. Access Time:** On average, the examinees have 46.64 minute to access the airport with median access time of 40. The shortest access time is 4 and the longest access time is 180 minutes. 75% of the interviewees access to the airport 60 minutes or earlier, and 25% of total have 30 or less minutes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

**Data Processing - 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 includes 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.

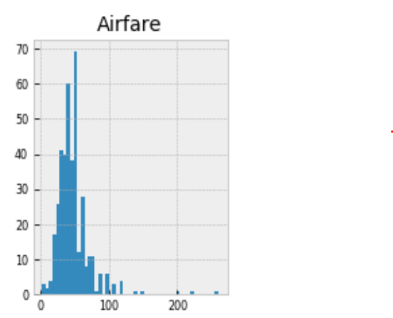
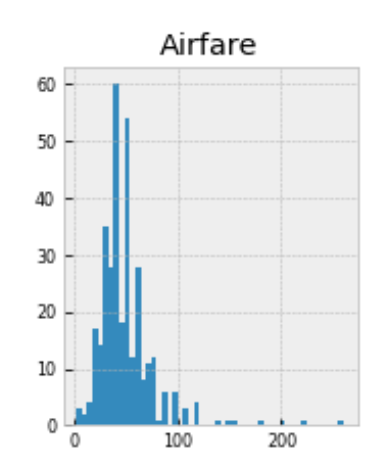
Before replacing the average values, we have dropped NA values from ‘Airline’, ‘Departure’, and ‘Destination’ variables, which led to 149 missing values in the ‘Airfare’ variable out of 476 fields. After observing our reference data, we have discovered that not all the average airline costs were provided (ex. Foreign Carriers). Although not all 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 must 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.

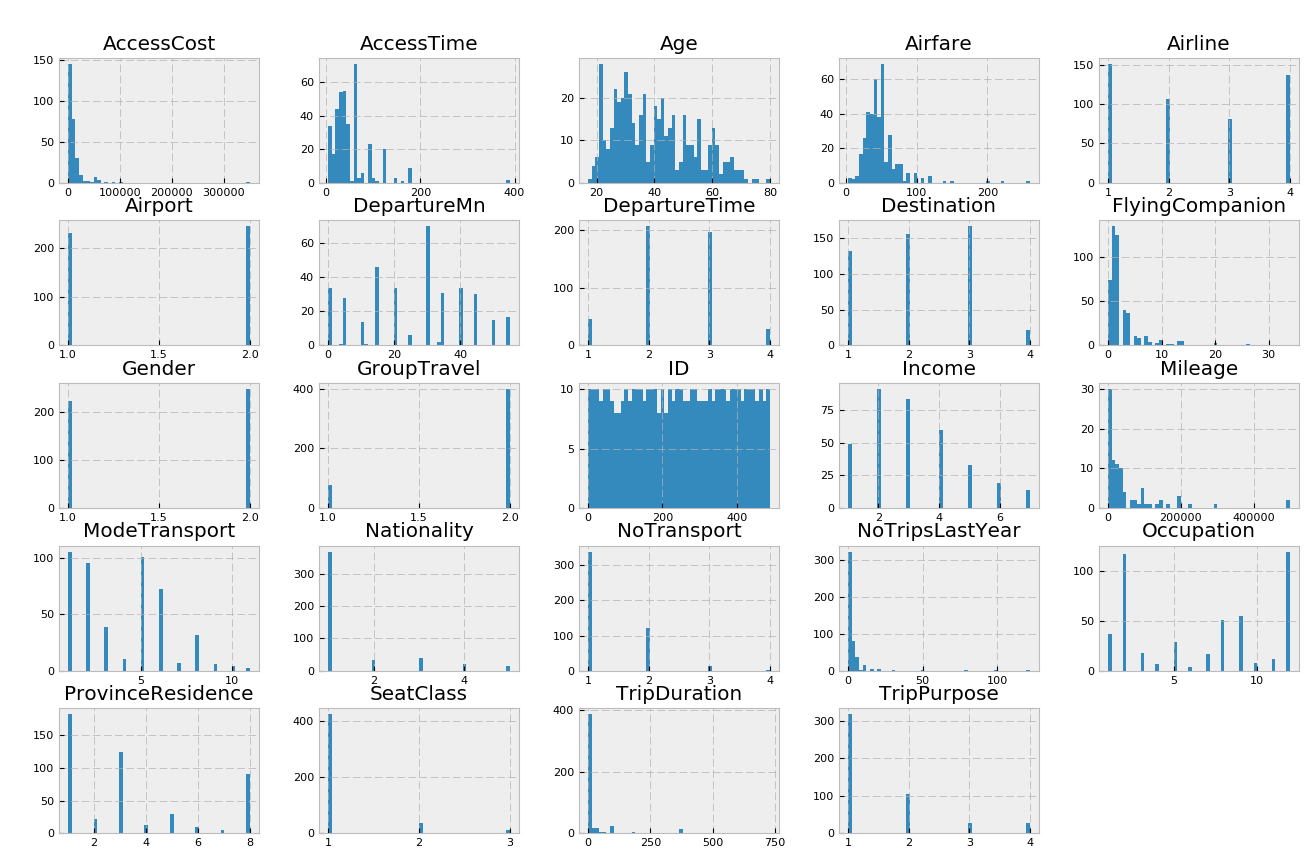


1 Airfare =10,000 Korean Won, KRW

**Original data – Airfare After replacing Average Airfare:**



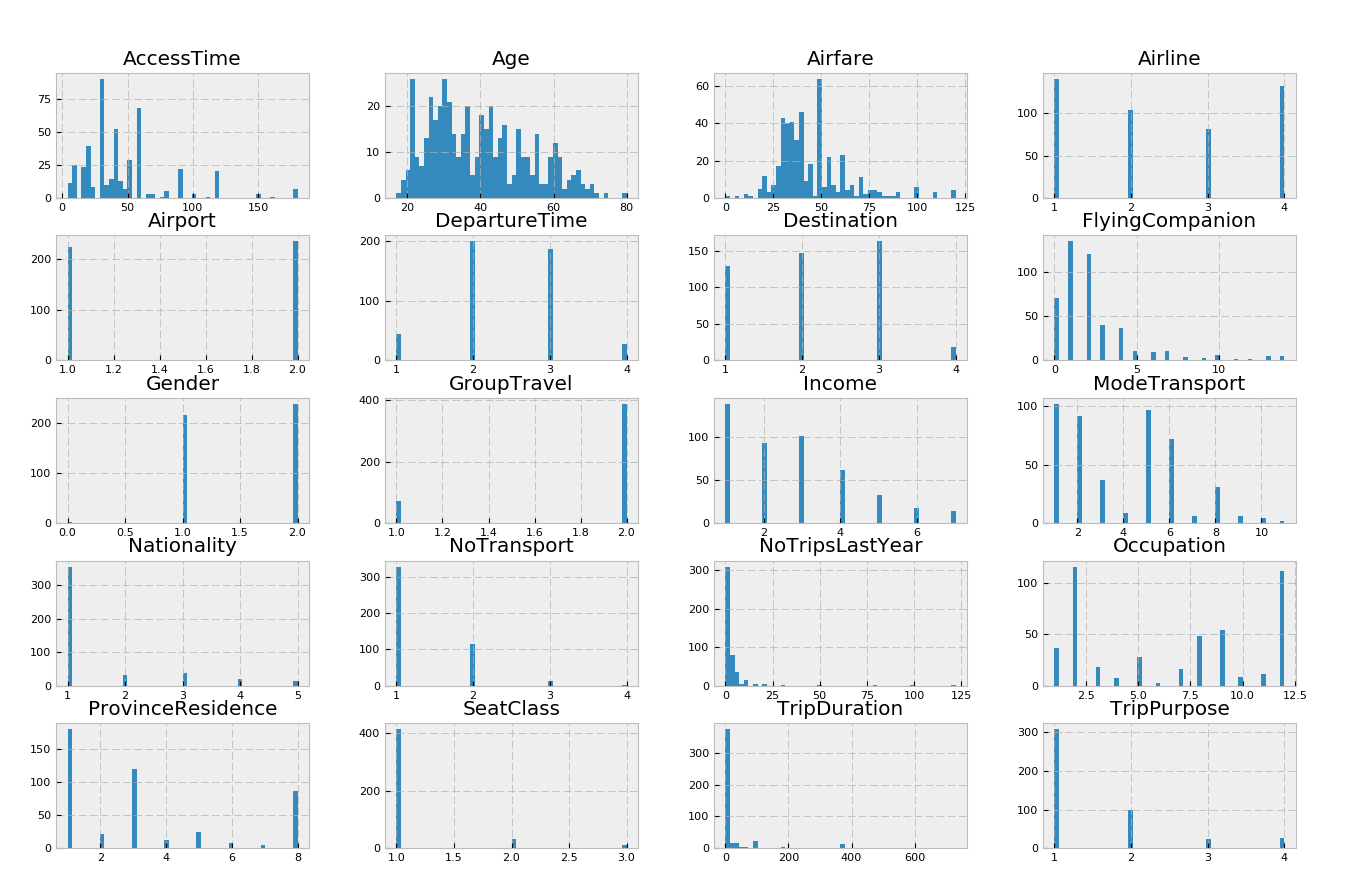
After examining histograms of our data to check for skewing of the curve, we have found that both graphs have similar distributions. Also, the descriptive stats yield similar results before and after; mean values were very close at 50.456 and 48.332, and the median airfare stayed the same at 45. Therefore, 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 our histograms to further evaluate outliers and NA values in the remaining variables. Based on our histogram evaluations, 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.

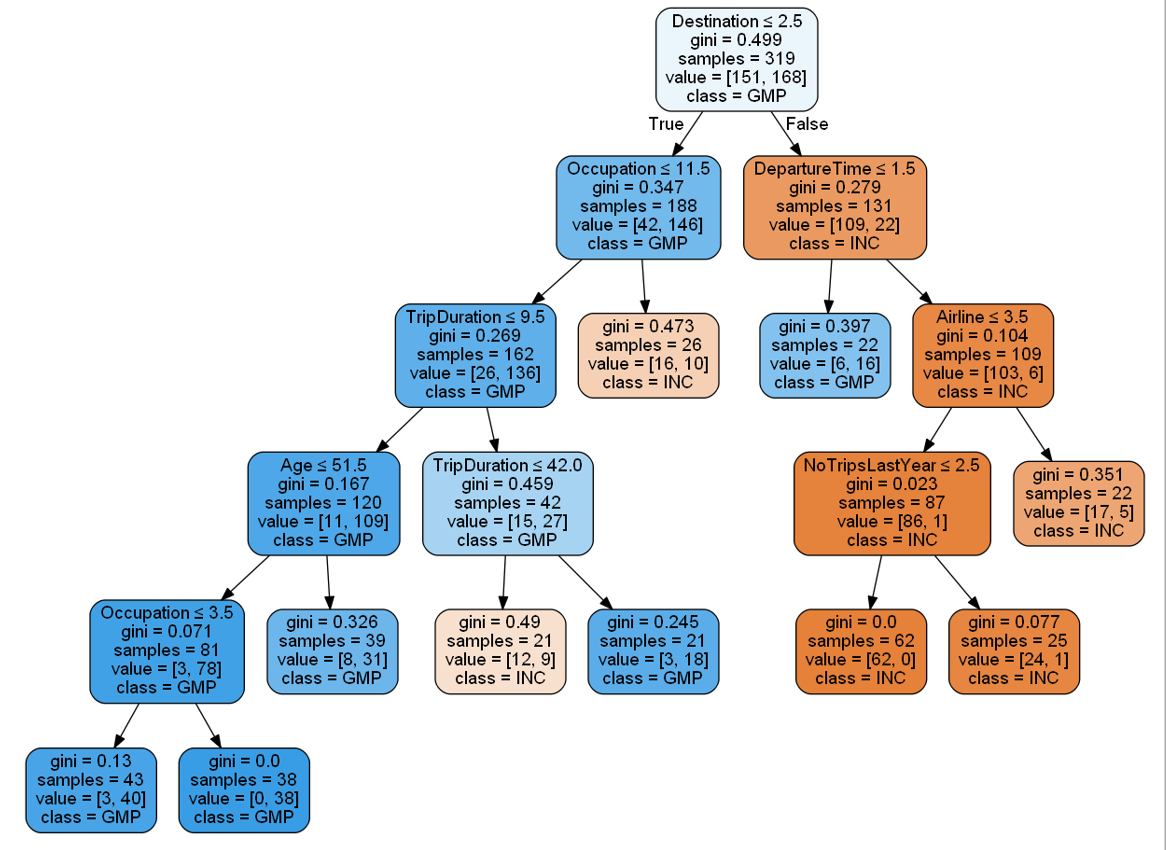
# Since most of the missing income is under occupation 12, the rest categories are not significant and occupation in the 5,8,9 categories have a decent size from our total population as well. We processed to group by occupation and get the average income for each of the groups to fill in our NA for Income. For the remaining Airfare NA, we take the average by grouping by Airport, Airline, and Destination. To fill NA for AccessTime, we have grouped Airport and ProvinceResidence together.

# The histograms below show the results after we have removed the outliers and filled in NaN values for all our selected features. The trimmed data contains 457 (93%) observations after cleaning our raw data.

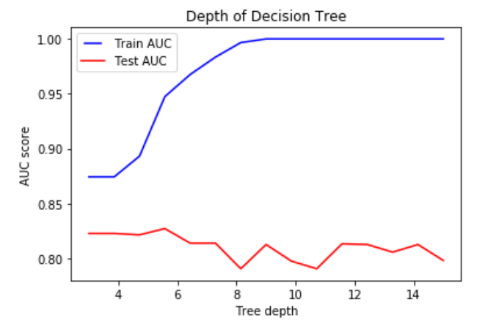
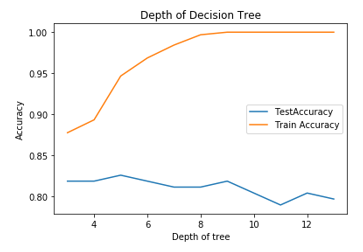


# III. Model Implementation

# Decision Tree Airport Choice Model



**Decision Tree Results for the Airport Choice Model *(max\_depth*: 5*, min\_samples\_split* = 20)**

**AUC Comparison (Train vs Test) - Airport model Accuracy Rate (Train vs Test) - Airport model**

For the initial evaluation of our decision tree model, we have compared both 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\_split* 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 *min\_samples\_split* = 20, as this had much higher accuracy, precision and recall rates compared to *min\_samples\_split* = 10.

We have started 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 most 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.

|  |  |
| --- | --- |
| Confusion Matrix – Airport choice model | |
| Accuracy Rate | 81.20% |
| Recall Rate | 86.40% |
| Precision Rate | 77.00% |
| Test Error Rate | 18.80% |

## 

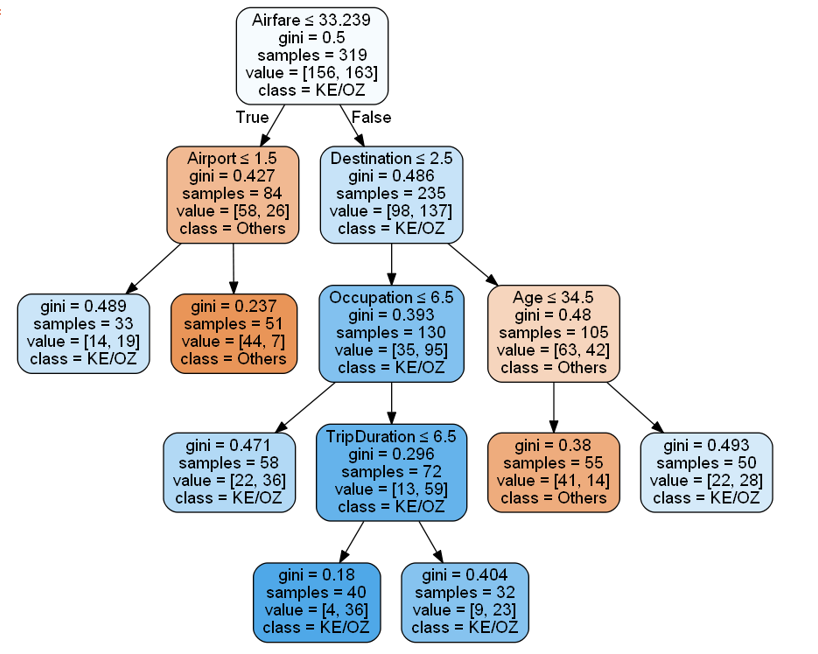
**Confusion Matrix for Decision Tree Results - Airport Choice Model**



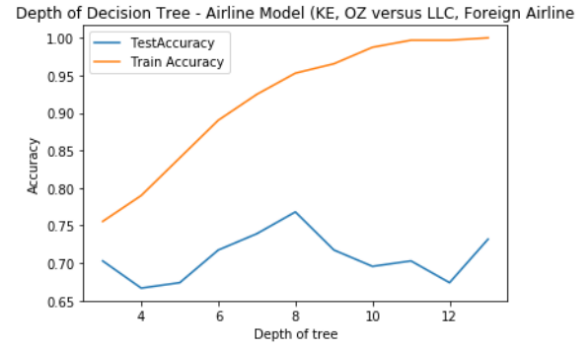
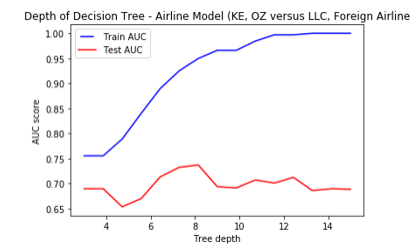




## **Decision Tree Model – Airline Model 1 (KE, OZ versus Others)**



**Decision Tree Results for the Airline Choice Model 2 (*max\_depth* = 8, *min\_samples\_split* = 30)**



**AUC Comparison (Train vs Test) - Airline Model 1 Accuracy Rate (Train vs Test) - Airline Model 1**

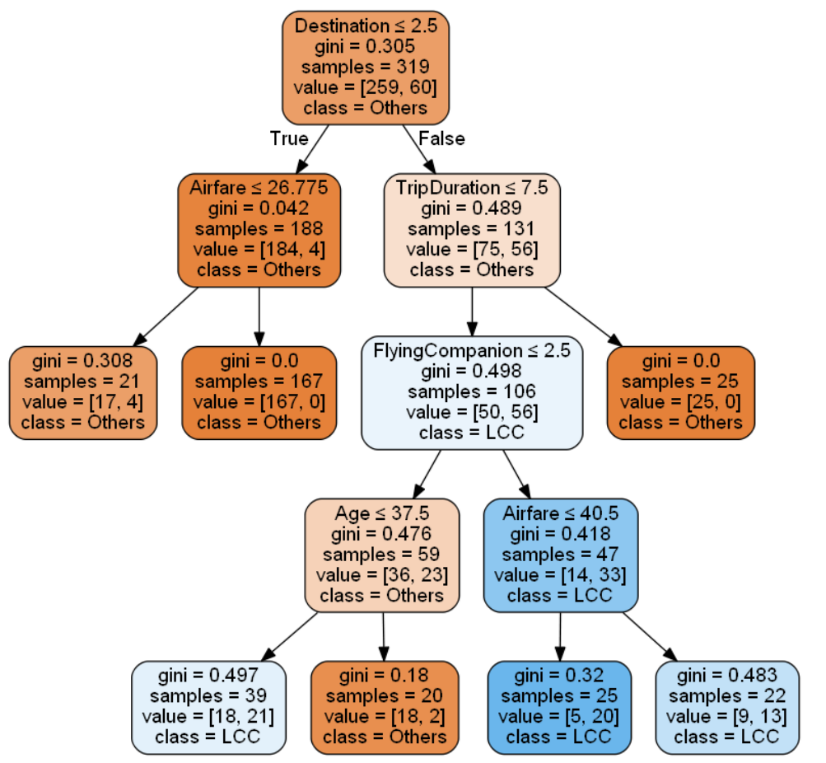
For the Arline Model 2, we have grouped the major Korean airlines together, Korean Airline (KE) and Asiana Airline (OZ), versus the rest of airlines. As we go down the nodes to the left side of our tree, we can see that the airfare and airport choice (INC) is an important decision factor for choosing the Korean and Asiana Airlines. Notable variables for the major Korean airlines are destination, occupation and the trip duration. As the major Korean airlines have the main hub at the Incheon Airport, we can see why the nodes have been broken down this way.

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 minimum sample sizes as 30.

|  |  |
| --- | --- |
| Confusion Matrix – Airline choice Model (Major Korean Airlines vs Other Airlines) | |
| Accuracy Rate | 67.40 % |
| Recall Rate | 75.30% |
| Precision Rate | 86.40% |
| Test Error Rate | 32.60% |

**Confusion Matrix for Decision Tree Results - Airline Choice Model 1 (KE & OZ Airlines vs Others)**

## **Decision Tree Model - Airline Model 2 (Korean LCC vs Other Airlines)**



**Decision Tree Results for the Airline Choice Model 1 (*max\_depth* = 6, *min\_samples\_split* = 20)**

## 

**AUC Comparison (Train vs Test) – Arline model 2 Accuracy Rate (Train vs Test) – Airline Model 2**

For the Arline Model 1, we have grouped Kroean LCC Airline versus other airlines. Looking at the result, we can notice that *Destination, TripDuration, FlyingCompanion and Airfare* are the important deciding factors for choosing the Korean LLC Airline as we move down to the right side of our depths towards our nodes with LCC classes.

For the 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* = 6. Therefore, we have set our *max\_depth* to 6, and kept the minimum sample sizes as 20, our confusion matrix showed high increase in the recall rate (tripled) and increased accuracy rate with decreased test error rate. Increasing the depth of our tree will result in overfitting in our data.

|  |  |
| --- | --- |
| Confusion Matrix – Airline choice Model (Major Korean Airlines vs Other Airlines) | |
| Accuracy Rate | 83.30 % |
| Recall Rate | 76.20% |
| Precision Rate | 47.10% |
| Test Error Rate | 16.70% |

**Confusion Matrix for Decision Tree Results – Airline Choice Model 2 (Korean LCC vs Others)**

* **Logistic Regression**

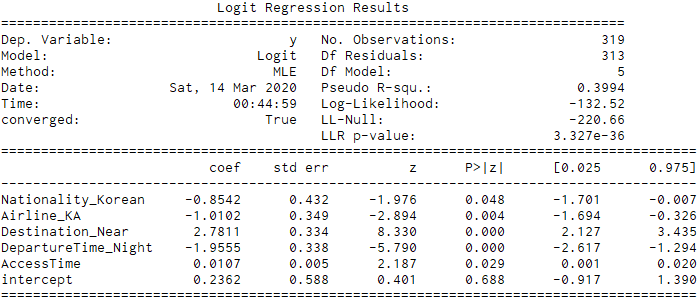
Our response variable is set to be binary for the airport choice, and the expected result is more appropriate to be a curved line instead of a straight regression line. Thus, we chose to apply logistic models for our study. We then used AIC and BIC scores to compare between models. These scores help to choose the best subsets in our 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 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 also means that a model is more likely to be the true model (*Dziak*, *2019*). We have also considered the adjusted R2 where it pays a price for the inclusion of unnecessary variables in the model.

1. **Logistic Regression - Airport choice model:**

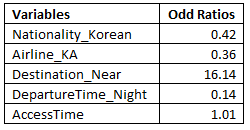
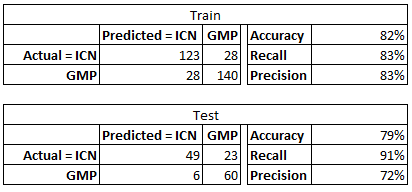
For describing the customer’s airport choice behavior between GMP (class 1) or ICN (class 0), we built two logistic models using *Nationality, Destination, Airline, DepartureTime, and AccessTime* (logit model 1) and *NoTransportation* (logit model 2) as independent variables. We split our 457 observations into train and test data. We picked a set of 319 random observations (70% of our trimmed data)for the train data and used the rest for testing.

Evaluating the air traveler’s airport choice decision in our data involves multiple factors, and the scale of these variables vary by the different groups in our data. However, given our data, some common factors which are affecting the decision most are as followed.

First of all, expats usually favor 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. Also, airlines in our data provide flights for different routes/destinations and consequently, Incheon (ICN) and Gimpo (GMP) do not share the same airline list. Therefore, airline and destination would be very significant factors for choosing an airport. While travelling, departure time plays an important factor for the passengers in order to maintain their schedules and that may affect 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.

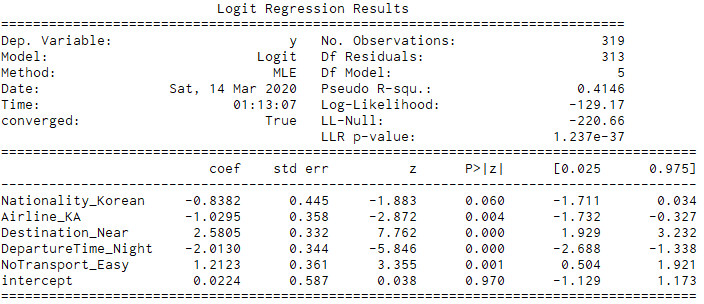
**Airport Logit Model 1 with AccesssTime (Access Time to Airport):**

**AIC: 277 BIC: 299.6**

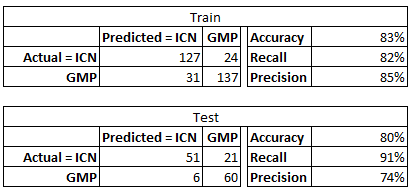


The Destination (Japan or China vs. others) and DepartureTime (redeye or regular) for both Korean airlines and Foreign airline, were significant factors that affect customer choice behavior at 1% significance level. Nationality as Korean or non-Korean, and AccesssTime are significant at 5% level.

Holding all others constant, the odds 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.

**Airport Logit Model 2 with NoTransportation (Number of Transportation):**

**AIC: 270.3 BIC: 292.9**

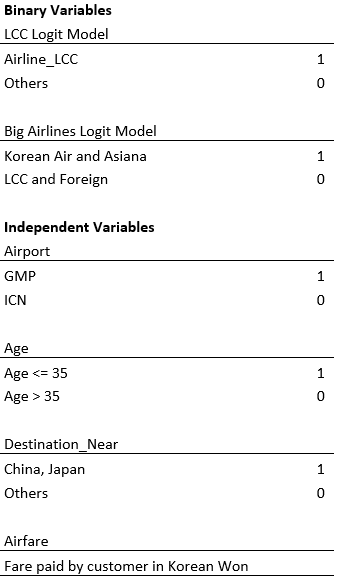
The Destination\_Near (Japan or China vs others), and DepartureTime\_Night (redeye flight or regular) are the significant factors that affect customer choice behavior at the 1% significance level. Airline\_KA is significant at 5% level and Nationality\_Korean is significant at 10% level.

The odds 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 odds ratio of NoTransport\_Easy is 3.36 meaning that the number of transports being one has the odds of choosing GMP is 336% higher than the number of transports being more than one.

The more transportations an air traveler must use to get to the airport the more likely they will have to spend longer time on the road to get there. For the logit model 2 above, we can see by substituting AccesssTime with NoTransport\_Easy, all other variables stayed significant in the model while reducing both AIC and BIC scores and increase R-square from 39% to 41%. In addition, NoTransport\_Easy is a lot more significant compared to AccesssTime and has bigger magnitude. We can say that model 2 is closer to the true underlying reasons and most likely be the true model.

1. **Logistic Regression - Airline choice model:**

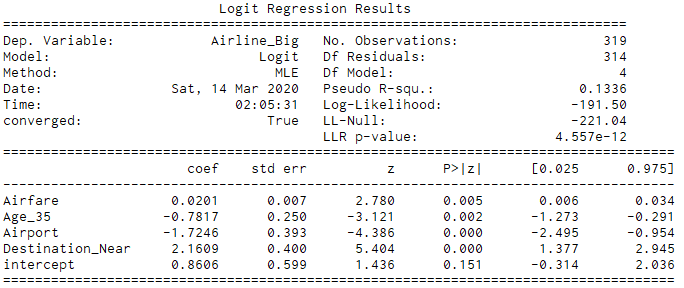
All the airlines are not shaped by the same set of facilities. These facilities may matter to some passengers and may not. However logically, there are some common factors that probably come into play every now and then while choosing an airline. Among which, age is a crucial factor i.e. a large portion of travelers 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.



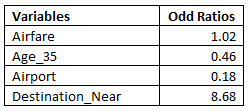
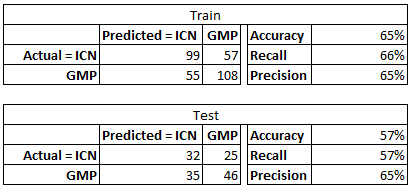
We developed two models: Big Korean Airlines vs. Others and LCC vs. Others.

After rounds of testing, comparing AIC scores between models and referring to our Decision Tree, we found that Airfare and Destination play very important roles on customer’s choice whether they want to use low-cost airline or not. While for the decision between using well known airline or not were significantly described by the Airport selection, Destination, along with the age range and Airfare.

The LCC model is 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).

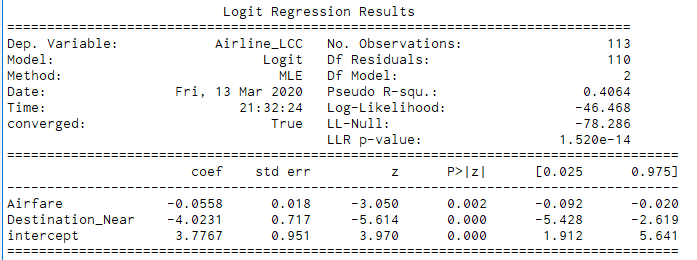
**Major Korean Airlines (KE, OZ) vs Other Airlines Logit Model:**

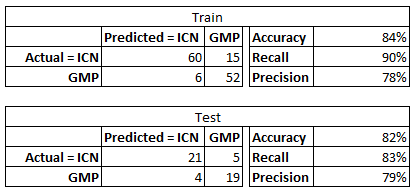
**AIC: 393 BIC: 411.8**

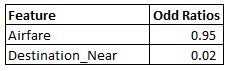


The Destination (Japan or China vs other), and Airport (GMP or ICN) are the significant factors that affect customer choice behavior at 1% significance level. Airfare and Age whether below 35 or not affect customer’s choice at 5% significance level.

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 10,000 won increase 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.

**Korean LCC Airline vs Other Airlines - 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.

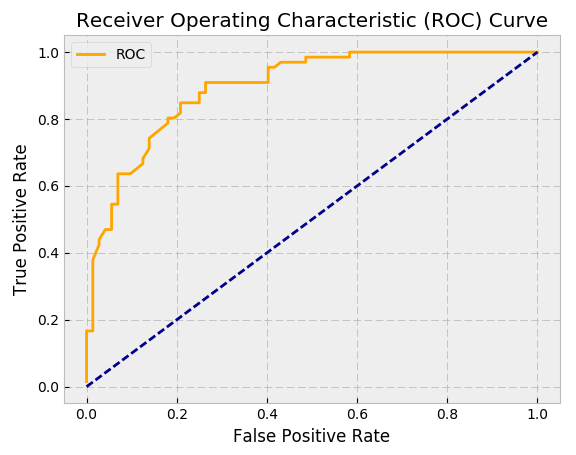
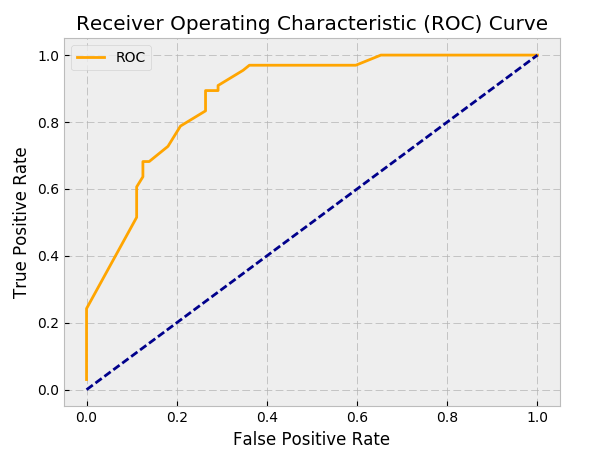
Given LCC is a low-cost airline which offers affordable travel tickets to customers, for every Korean Won increases in Airfare, the odds of the customer choosing LCC airline deceases by 5%. The estimated odds of choosing GMP are 98% lower among customers whose Destination is China or Japan holding all other constant. Although we only have a small sample size to run this model, we saw a significant improved in the model performance. We are not considering AIC or BIC score since this model and the Major airline model didn’t use the same set of data, we can see that R-square number is a lot better (40% compared to 13%) as well as the confusion matrix rates for LCC model.

# IV. Model Comparison – Airport Choice Model

# 

* + - 1. **Logistic Regression - Airport Choice**

The area under the ROC curve shows the performance of an algorithm for the data classification compared to a random classification.

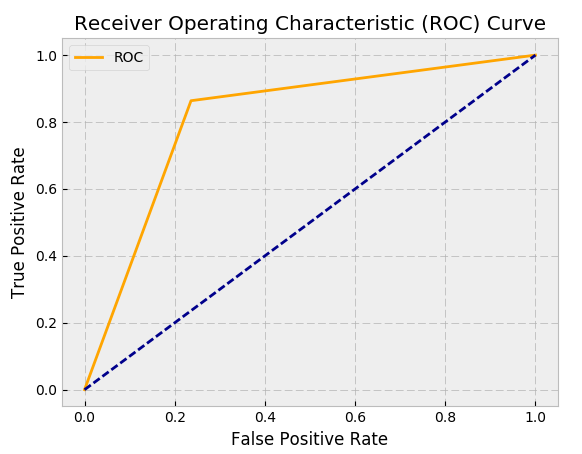


ROC Curve for Airport Logit Model with NoTransport

ROC Curve for Airport Logit Model with AccesssTime

Examining the ROC curves for both airport logit models, the model including the number of transports provides a better visibility of the data than the model including access time.

2. **Decision Tree – Airport Choice**

 Comparing the ROC curve and AUC scores for airport logit model with the number of transports with the decision tree airport model, the logit model performs slightly better as its AUC score is about 0.05 higher than the airport decision tree model

ROC Curve for Airport Decision Tree Model

**Confusion Matrix Comparison – Airport Choice Model**

|  |  |
| --- | --- |
| Confusion Matrix – Airport choice model: **Logistic Regression Model 1** | |
| Accuracy Rate | 79% |
| Recall Rate | 91% |
| Precision Rate | 72% |

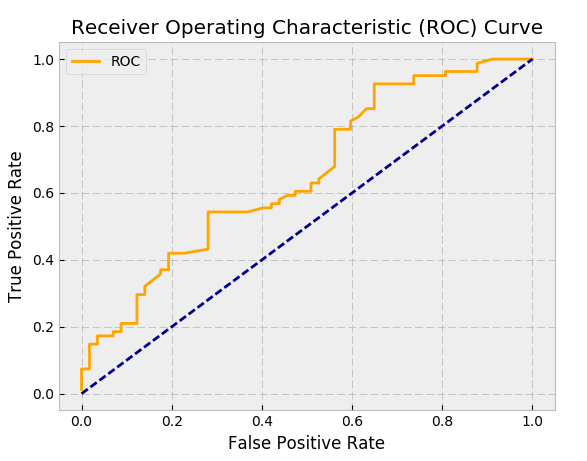
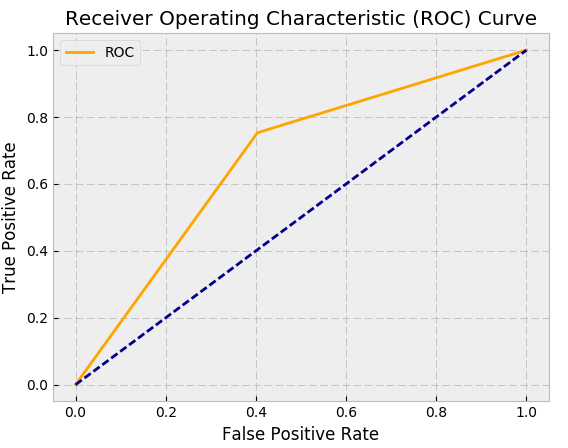
|  |  |
| --- | --- |
| Confusion Matrix – Airport choice model: **Logistic Regression Model 2** | |
| Accuracy Rate | 80% |
| Recall Rate | 91% |
| Precision Rate | 74% |

|  |  |
| --- | --- |
| Confusion Matrix – Airport choice model: **Decision Tree Model** | |
| Accuracy Rate | 81% |
| Recall Rate | 86% |
| Precision Rate | 77% |

# IV-I. Model Comparison – Airline Choice Model

**Airline Choice:**

**1. Logistic Regression – Airline Model 1: Major Korean Airlines (Korean Air and Asiana Airlines vs others)**



ROC Curve for Major Decision Tree

ROC Curve for Major Airline Logit Model

ROC Curve for Big Airlines Decision Tree Model

ROC Curve for Big Airlines Logit Model

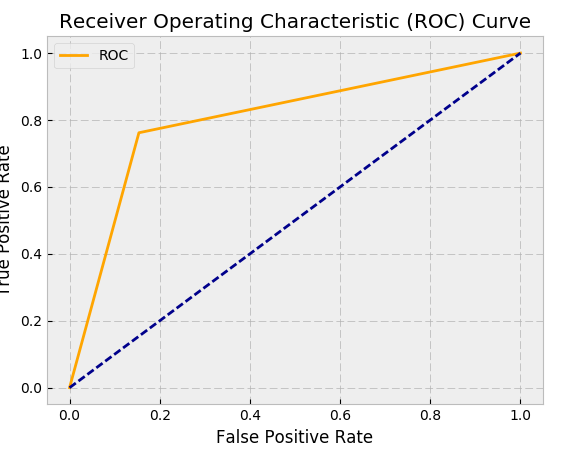
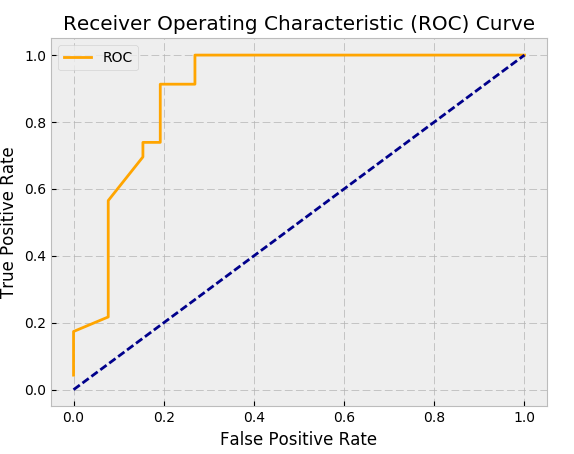
Between big airlines (Korean Air and Asiana Airlines vs others) logit and decision tree model comparison, logit model’s AUC score is approximately 0.02 higher than the decision tree model. Hence, logit model appears to be a better descriptive model than decision tree.

**Confusion Matrix Comparison – Airline Choice Model 2**

|  |  |
| --- | --- |
| Confusion Matrix – Arline choice model 1: **Logistic Regression Model** | |
| Accuracy Rate | 57% |
| Recall Rate | 57% |
| Precision Rate | 65% |

|  |  |
| --- | --- |
| Confusion Matrix – Arline choice model 1: **Decision Tree Model** | |
| Accuracy Rate | 67% |
| Recall Rate | 75% |
| Precision Rate | 86% |

2**. Logistic Regression – Airline Model 1: Korean LCC vs Other Airlines Choice:**



ROC Curve for LCC Logit Model

ROC Curve for LCC Decision Tree

**Confusion Matrix Comparison – Airline Choice Model 1**

|  |  |
| --- | --- |
| Confusion Matrix – Airline choice model 2: **Logistic Regression Model** | |
| Accuracy Rate | 82% |
| Recall Rate | 83% |
| Precision Rate | 79% |

|  |  |
| --- | --- |
| Confusion Matrix – Airline choice model 2:  **Decision Tree Model** | |
| Accuracy Rate | 81% |
| Recall Rate | 86% |
| Precision Rate | 77% |

For the Korean LCC vs. other airlines, logit model again performs slightly better than the decision tree. Logit model’s AUC score is about 0.09 higher than the decision tree model. Moreover, looking at the confusion matrix, the performance is not very significantly different in logit and decision tree model.























|  |  |  |  |
| --- | --- | --- | --- |
| **AUC** | **Airport** | **Major Korean Airlines** | **LCC Airline** |
| Logit | 0.88 | 0.65 | 0.90 |
| Decision Tree | 0.81 | 0.63 |  |

**Model Comparison Conclusion**

Overall, the logistic regression and decision tree models performed similarly in the airport model and the airline model 1, looking at the ROC, AUC and confusion matrix. For this data set, both the logistic regression and the decision tree models had comparable results, but we did notice a slightly better performance in the airport choice model and the airline model 2.

When we were evaluating the first airline model with the major Korean airlines versus other airlines, we have noticed that the confusion matrix results weren’t performing as well as we have hoped. Therefore, we have grouped the airlines differently with the 2nd airline model, comparing the Korean LLC versus other airlines. As a result, we have noticed increased accuracy, recall and precision rates.

# V. Conclusion and Recommendations

The objective of our study is to determine how air travelers in the Seoul Metropolitan area choose one airport and airline over another for traveling, and what factors influence their decision. Incheon International Airport (ICN) is the largest airport in South Korea, and one of the busiest airports in the world located between Yeongjong and Yongyu Islands (west of Incheon’s city center). Before Incheon was built in 2001, Gimpo (GMP) was the main international airport for Seoul located by the western end of the city. Our survey data has customers of the two biggest airlines in South Korea: Korean Air, and Asiana. In addition, the data has Korean LCC (low-cost airline) as well as a set of customers using foreign airlines.

Given that GMP is the smaller and older airport and mainly has flights to China or Japan, Destination plays a very important role in customer’s decision on choosing between GMP or ICN in our model. The preliminary results of this paper show airline, customer’s nationality, number of transportations needed to get to the airport and whether it is a redeye flight or not are also significant in customer’s airport choice. Considering GMP was built long time ago and is closer to the city center, if the customer doesn’t want to do complicated transport to the airport, they will more likely to go with GMP and not ICN since it is a lot further out. On the other hand, travelers are not allowed to stay overnight at the GMP airport, thus we expect if the person has late evening or nighttime flight, they will most likely to choose ICN since ICN is bigger and open 24/7. If a customer is not Korean and not familiar with Seoul area airports, they might not be aware of the GMP’s condition and will be slightly more likely to choose it while if the customer is Korean, they would know the policy as well as history of GMP and would prefer to use ICN.

For the airlines, Korean Air (KE) and Asiana Airlines (OZ) are the two biggest airlines in Korea. These are the only carriers that fly from just about everywhere to Seoul direct. Both airlines provide good services, with Korean Air being slightly better with higher cost. For LCC and other foreign airlines, airfare and destination are the most significant deciding factors for customers from our model evaluations. Additionally, customer’s age and airport choices were also significant factors on airline choice as well.

Considering the significance of destination to China or Japan, number of transportations to airport, and different departure times play vital roles in customer’s airport choice. Therefore, we suggest that the airports in the Seoul Metropolitan area to consider following recommendations to increase their airport traffic. For the Incheon Airport (INC), we recommend to improve mode of transportation to the airport such as more frequent airport shuttles from the city. Also, INC can work with the airlines to offer more flights to Japan and China area with lower rate. As for the Gimpo Airport (GMP), we suggest to extend their operating time, since the airport doesn’t allow air travelers to stay after 11 pm. The GMP airport is limiting all customers from booking air tickets during the time they are closed. With the strict operating hours, GMP has limited airlines offered at this airport. If GMP can work with the city officials to expand the airport and work with airlines to offer more flights out to different destinations, GMP would have a very huge potential of attracting more customers.

The airlines are necessitated to take measure to survive in the competitive industry. Based on the analysis, the core factors that the airlines need to focus are airfare and destination. Considering our analysis, we have noticed that on average, younger air travelers tend to consider the lower airfare. Airlines should offer competitive range of airfare so that the sales increase. Parallelly, analyzing the most frequent destinations is crucial for airline industry. In this case, China and Japan are the routes that airlines need to put more focus to increase number of passengers. Smaller airlines (i.e. Korean LCC) need to focus on price primarily to compete with the larger airlines as providing more options on destination might be cost sensitive.

For further research, we want to expand our survey to have more observations to further validate the effects of the factors we determined that might have strong impact on airport and airline choice for the travelers in Seoul area. Some airlines have exclusive direct flights to certain area/airports, we would like to include destination airport variable as well as if the traveler has connecting flight or not. These two variables would help us provide a more concrete descriptive information on customer’s choice of airport and airline in Seoul area.

**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).