**Statistic Learning - Group Project**

**Group 11: Shuai Ma, Huy Le, Han Li, Wei Li**

Airport and Airline Choice Modeling: Understanding of Passengers’ Choice Behavior

# **1 Introduction**

Passenger’s airport choice in multi-airport regions (MARs) has been of great interest in the researches of local governments, airport authorities and airline companies. Like any other organization or company that sells products or services, airports, in the same way, have to compete with each other for customers. In this study, we use the revealed preference survey result from 488 passengers and try to find out the factors that impact the choice of airports. The two airports of interest are Gimpo Airport (GMP) and Incheon Airport (ICN). Presumably, distance between the airport and the city, conditions of the airport and other factors, such as age, income level, or occupation of passengers might impact the choice. When reviewing similar airport preference researches, we also found air fare, flight frequency, access time, ground transport expenses, etc. would mainly capture passengers’ choice of origin airport. The value of this research would be seen in several aspects. From the aspect of the airport authorities, they would like to know the factors, take action accordingly to compete with the other airports. From the local government perspective, understanding the factors can help balance the traffic of the two airports, improve the overall airports’ service, and provide better airport experience. For airline companies, this study also can provide useful insights that they can refer when arranging air routes between the two airports. So the first purpose of this paper is to explore the factors that are of critical significance affecting the choice of airports. The second task of this paper is focused on the choice of airlines. It’s similar to the research of the choice of airport. As we can see, there are fierce competition among airline companies. Understanding passengers’ choice behaviors can help win more customers. We use the same survey data from the 488 respondents and explore the significant factors in the choice of four kinds of airlines: Korean Air (KE), Asiana Air (OZ), Korean LCC, and Foreign Airlines.

In the modeling process, we first refer to the Logit Model (LM). Through LM, we select the variables that are statistically significant in the first step, then calculate the AIC and BIC of combinations of variables and at last we get a set of variables that can be put into LM and are statistically significant with low AIC and BIC. Then we run the Decision Tree Model (DTM). With the result of LM and DTM, we can properly interpret the factors that are affecting the passengers’ choice of airports/airlines. Furthermore, to compare the accuracy of LM and DTM with other models, we also applied K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Naive Bayes, and Support Vector Machine (SVM). The result of LM and the top layers of DTM provide good insights for other models’ variable selection. Though most of the other models perform very well, they are not interpretable. So for those models, we will only focus on the performance rather than interpretation.

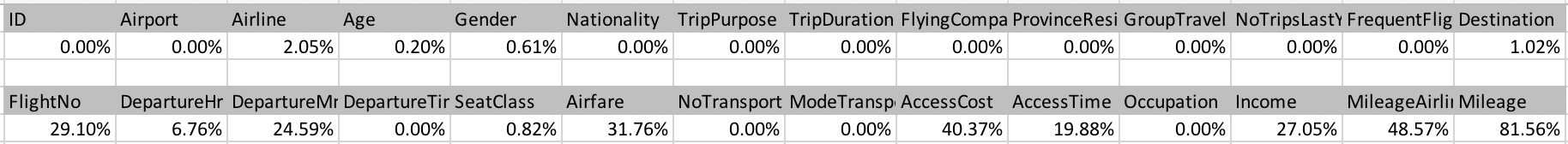
This paper will be organized into four main parts. First is the introduction of this research. The second part is exploratory data analysis (EDA), including the exploration of the raw data, imputation of the missing values, dealing with outliers, and the exploration of the cleaned data. The third part is the result of each model. We introduce each model here, explain the variable selection criteria for each model, present the results, and propose the insights. The last part is the conclusion part. We summarize the main findings of this study, propose recommendations according to the insights and point out the limitations.

# **2 Data**

# **2.1 EDA of Raw Data**

We used two files in this study. One is the survey result of passengers who chose one of the airports and one of the airlines. The raw data contains 488 records and 28 variables. We use this data set for the research on airport choice as well as the airline choice. So in this data set, one variable is the record of airport choice (GMP or ICN) and another is the airline choice (KE, OZ, Korean LCC, or Foreign Airlines). When modeling for the choice of airports, the airline choice records will be one of the independent variables. Likewise, when modeling for the choice of airlines, the airport choice data will be an independent variable. The data set also contains socio-demographic features, such as age, gender, occupation, income level, nationality, and province of residence.

It also contains some alternative-specific information for each respondent, such as trip purpose, group travel or not, frequent flight destination, destination for this trip, flight number, departure time, seat class, mode of transport, mileage airline companies and amount in mileage, trip duration, number of flying companions, number of trips last year, air fare, number of transportation modes that were used to get to the airport, access cost and access time. This data set will be the main input to models. To apply this data set into multiple models, we need to check the data integrity. The percentage of missing values of each variable is described as below.



From the table above, there are several columns that contains a large percentage of missing values, such as FlightNo (Flight Number), DepartureMn (Departure Minute), Airfare, AccessCost, Access Time, Income, MileageAirline, and Mileage. Some of them might be of little significance and thus can be ignored, while some of them might be able to imputed in a way. We will talk about the missing value imputation in 2.2.

The other data set we can make use of contains three parts, traffic information, price information, and airport province distance information. These can provide referencing information when we understand the survey data set and imputing the missing values in the survey data set.

## **2.2 Data Cleaning**

### **2.2.1 Missing Data**

The strategy of "best reasonable guess" will be applied in filling the missing values with the following principles:

1. Records with missing values of ***Airline, Age, Destination, Seat Class***and***Gender*** are removed, because the amount of missing values of these variables are very small and it is also difficult to set up a rule to fill out this information. Meanwhile, removing this data will not influence the result significantly. Therefore, we directly remove the records with those missing values.
2. For the variable of ***FrequentFlightDestination,*** we group all the missing value together as category 7 “unknown missing value”.
3. For ***DepartureHr, DepartureMn and DepartureTime,*** since they are highly correlated, we will only keep ***DepartureTime*** and remove the other two columns, while there is no missing value of ***DepartureTime.***
4. For the ***Airfare,*** we mainly refer to the information from table “Price\_Info” under the file “Airport\_Airline\_data”. We directly use the average price for the same value of “Airport, Flight.Destination, Airline”. However, there are still some missing values in the reference table. In that case, if the average price is blank, we group the sample data by “Airport, Flight.Destination, Airline” and get the mean of the “Airfare”. Then we use this mean value to fill out the missing value in the same table.
5. For the ***FlighNo***, since there are correlated variables like Airline in this model, we decided to drop it from our model. So there is no need to fill out its missing value.
6. For the ***AccessCost,*** we group the sample data by “Airport” and “ResidenceProvince”, then use the mean value to fill out the missing value.
7. For the ***AccessTime,*** we use the table “Airport Province Distance” under file “Airport\_Airline\_data” as reference information. We insert the missing value with “Travel Time” with the same airport and province.
8. ***Income*** is special in this data set, because there is 132 missing values out of 498 in total. Meanwhile, there is no reference table with additional information. If we directly group the data by “Age” and “Occupation”, the distribution of income may change a lot, which is not reasonable. In addition, we found that most of missing values are belong to Occupation = 12 - “Other”. Therefore, we decided to create a new category 8 of “Unknown Income” for ***Income.*** As a result, there will be 132 records with category 8 of ***Income.***
9. For the missing values in ***MileageAirline*** and ***Mileage***, we decide to leave it there, since they will not be used in our model. The reason for that is ***MileageAirline*** and ***Mileage*** are correlated with ***Airport*** and ***Destination.*** In order to get rid of many correlated independent variable in the model, we decide to drop them, so there is no need to fill out the missing values.

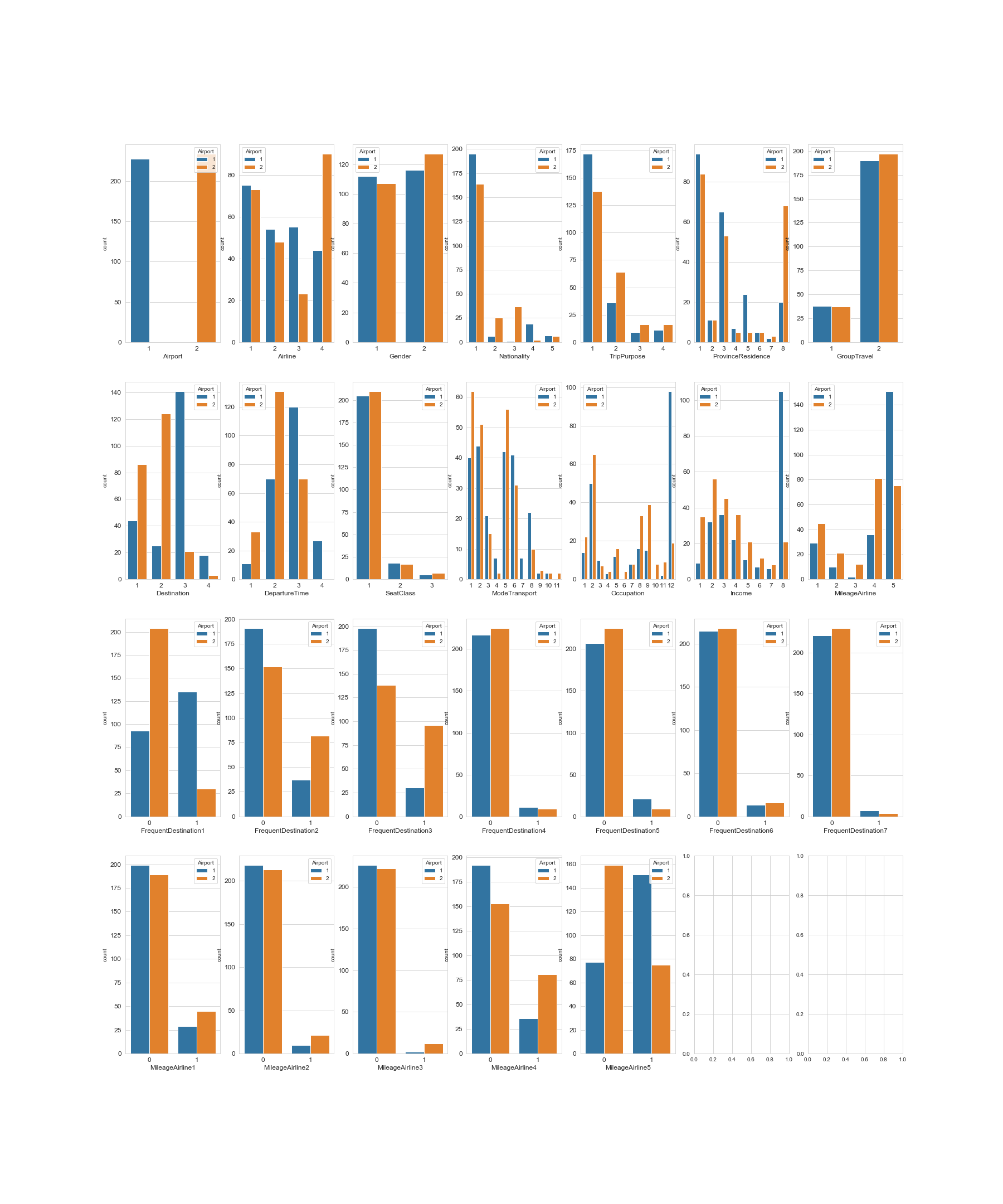
### **2.2.2 Outliers**

In this project, there is only 488 records. So we cannot remove too many extreme values. Based on EDA result, we decided to only remove one outlier in AccessCost, which is over 100,000.

## **2.3 EDA of Clean Data**

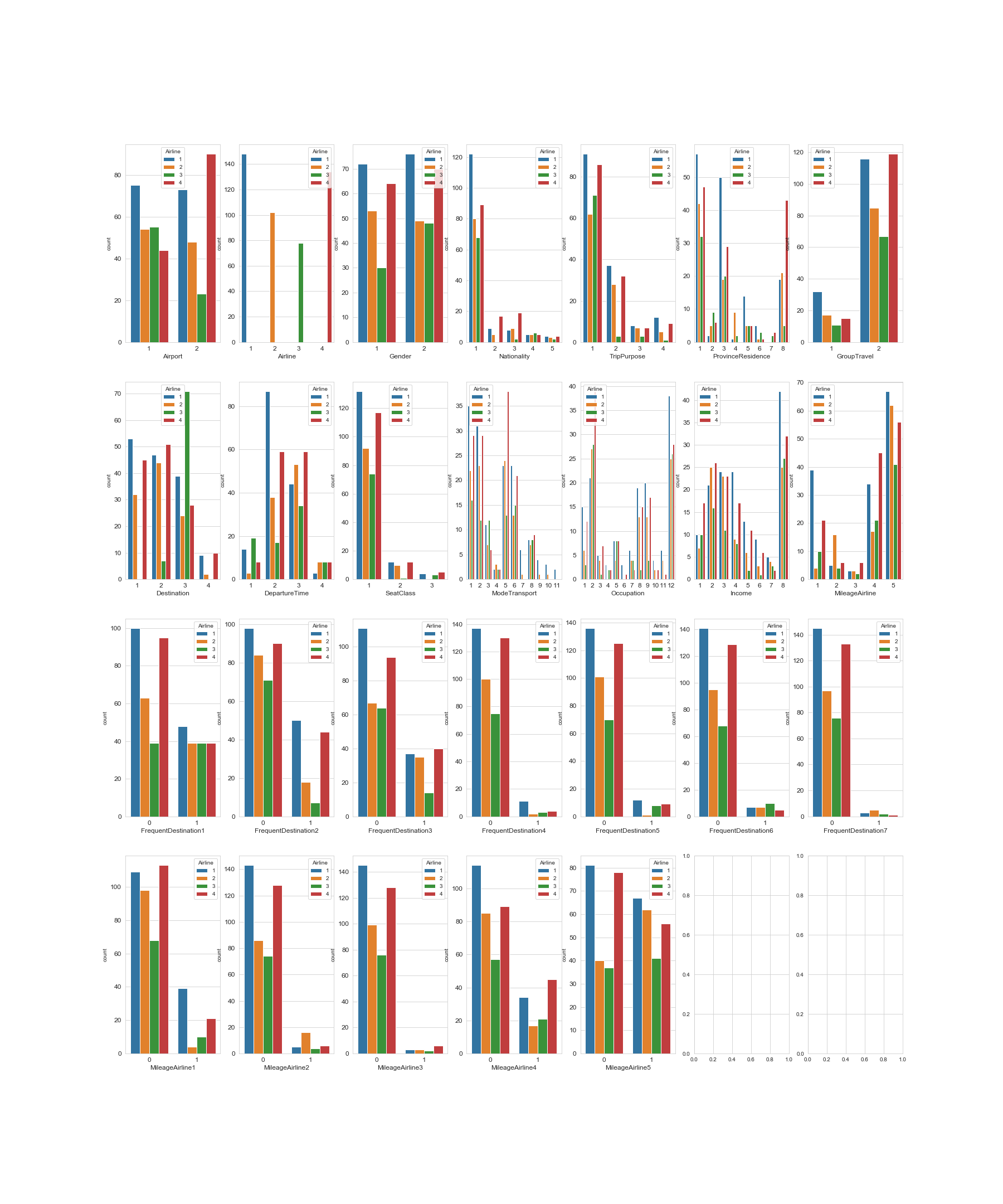
Since this is a classification task, after data cleaning, we are more interested in the features that have different distributions in different choice of dependent variables. These features have better classifying abilities that we can further put into models.

First, Figure 2.3.1 shows the distributions of all categorical variables in airport choice model. Blue bar is the count of choice of ICN, and orange is for GMP. We can get some insights from this graph. For example, we can regroup ‘Airline’ into Korean Airlines and Foreign Airlines; regroup ‘destination’ into Japan/China and Others; regroup ‘modeTransport’ into public transportation and individual transportation; ‘TripPurpose’ and ‘SeatClass’ are not good features; etc.



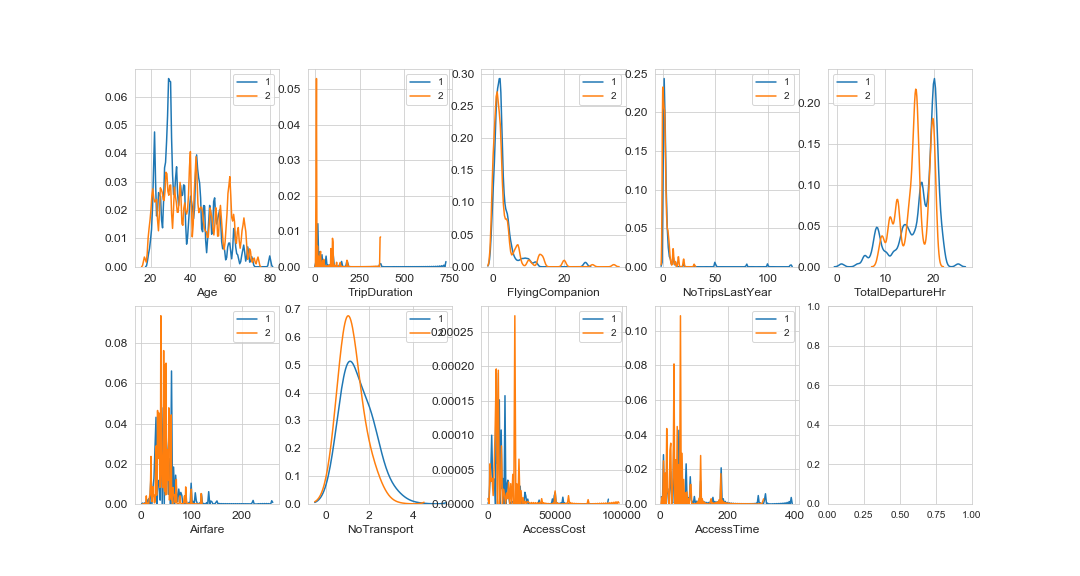
***Figure 2.3.1***

Figure 2.3.2 is the distributions of categorical variables in the Airlines Choice model. Blue bar is KE, orange OZ, green Korean LCC, and red Foreign Airlines. Here are some insights: “GroupTravel’, ‘Gender’, ‘FrequentDestination2/3’, ‘MileageAirline’ are not good classifiers; ‘Destination’ can be regrouped; etc.

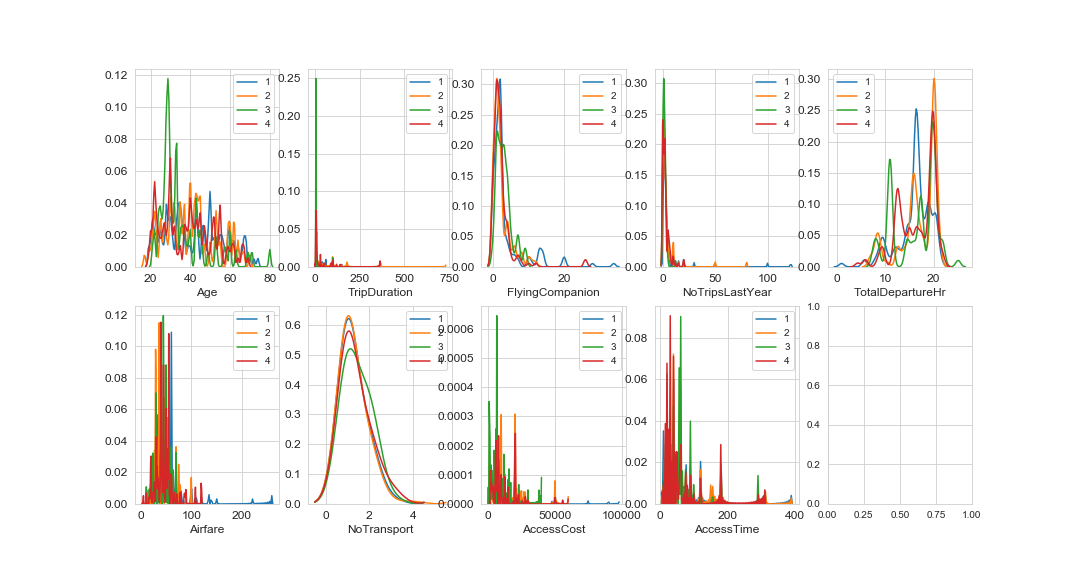


***Figure 2.3.2***

Then we look at the continuous variables’ distribution in the Airport Choice model, shown by Figure 2.3.3. Figure 2.3.4 is the continuous variables distribution in the Airline Choice model. Some of the distributions in the same variable are similar which indicates it’s not a good classifier, such as ‘NoTransport’. As for other continuous variables we need to put them into LM and see its statistics.



***Figure 2.3.3***



***Figure 2.3.4***

# 

# **3 Models and Result**

## **3.1 Models**

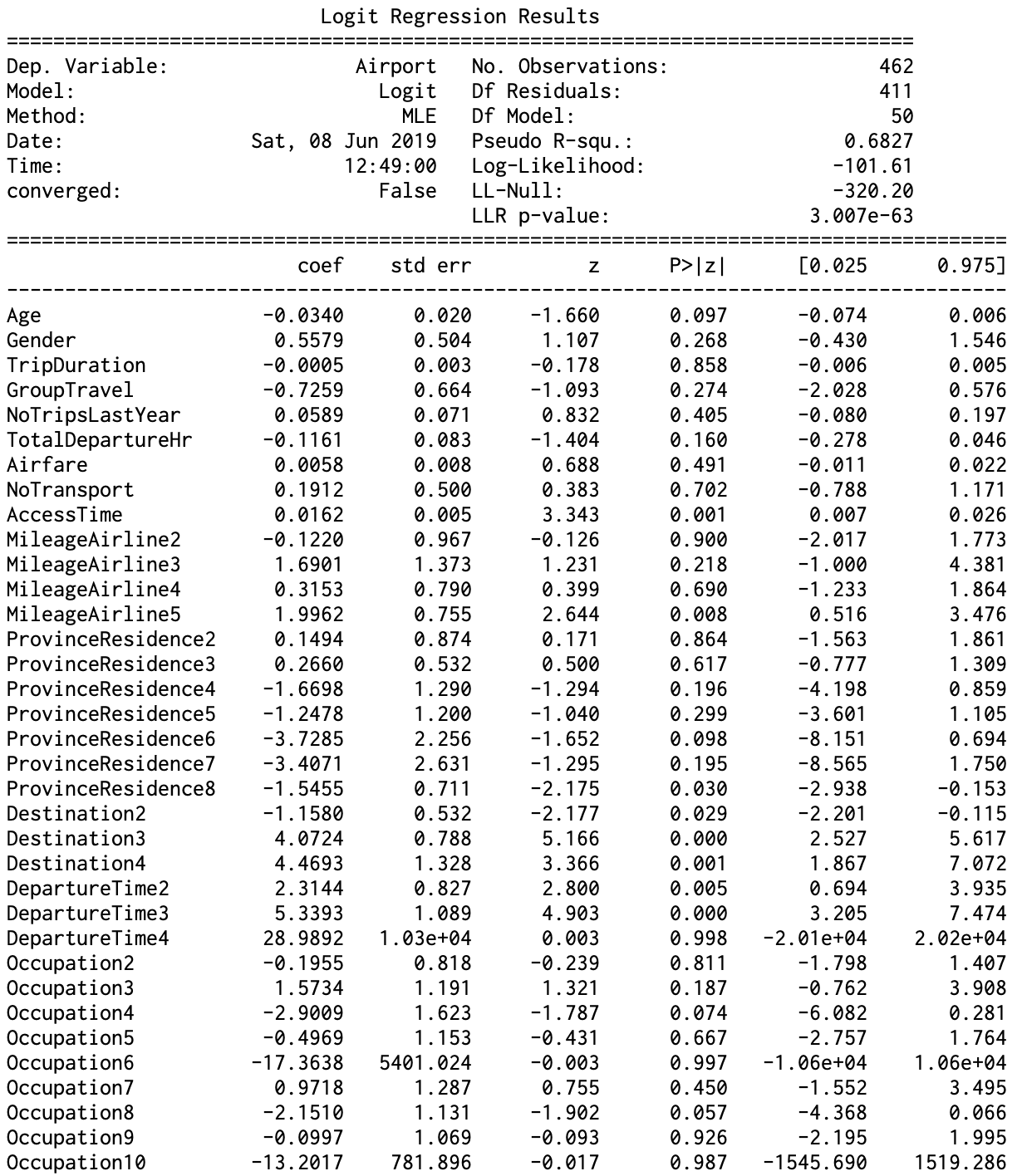
**3.1.1 Data**

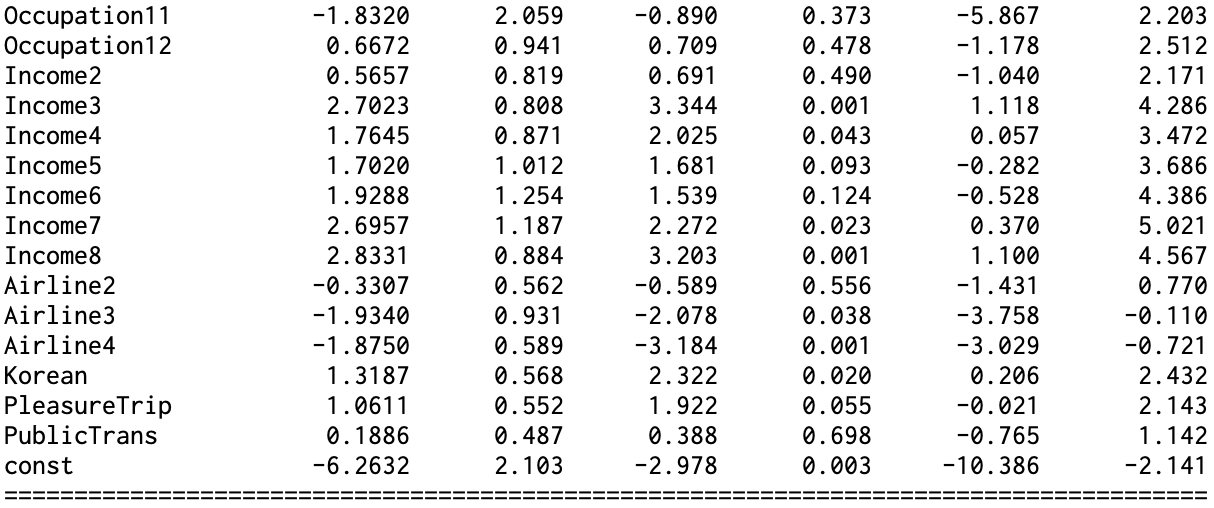
The data we use here contains 463 observations. For each model, we divide the data into 70% and 30% randomly as training data and testing data respectively. We train the model with training data. We predict the output of testing data with training model to get the testing accuracy, while we predict the output of training data with training model for training error.

**3.1.2 Logit Model**

In this project, there are two variables to explain, the choice of airport and airline. For airport, the result is binary, either 0 for GMP or 1 for ICN. While for airline, the result has 4 different airlines, 1 for Korean Airline, 2 for Asiana Airline, 3 for Korean LCC and 4 for Foreign Airline. In statistics, logistic model is widely used model with logistic function to model a binary dependent variable. So we build airport model with logit function. On the other hand, in statistics, multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems. Therefore, we establish airline model with multinomial logistic regression.

In order to better explain the choice of airport and airline, we need to select independent variables for each model. At first, we removed some correlated variables in the data. The next step is to put all the remaining variables as independent variables into logit and multinomial logit models respectively. The results are shown in the figure below.





***Figure 3.1.2.1 Airport Logit Model with all uncorrelated variables***

From ***Figure 3.1.2.1***, we would like to select the variables with high coefficient value and low p-value (<0.05) as our first selection result for airport model before calculating AIC and BIC. Therefore, the selected variables are ‘AccessTime’, ‘Destination’, 'DepartureTime’, ‘Income’, ‘Airline’, ‘Korean’, ‘PleasureTrip’ and ‘const’. Then the next step is to input the selected variables into logit model to calculate its AIC and BIC. We would like to select the lowest AIC and BIC result and, if possible, the highest R^2. The selected result is shown below in ***table 3.1.2.1***. We select all the variables after comparing a couple of models’ results. Therefore, we will use ***‘AccessTime’, ‘Destination’, 'DepartureTime’, ‘Income’, ‘Airline’, ‘Korean’, ‘PleasureTrip’*** to explain passengers’ choices of airport with logit model.

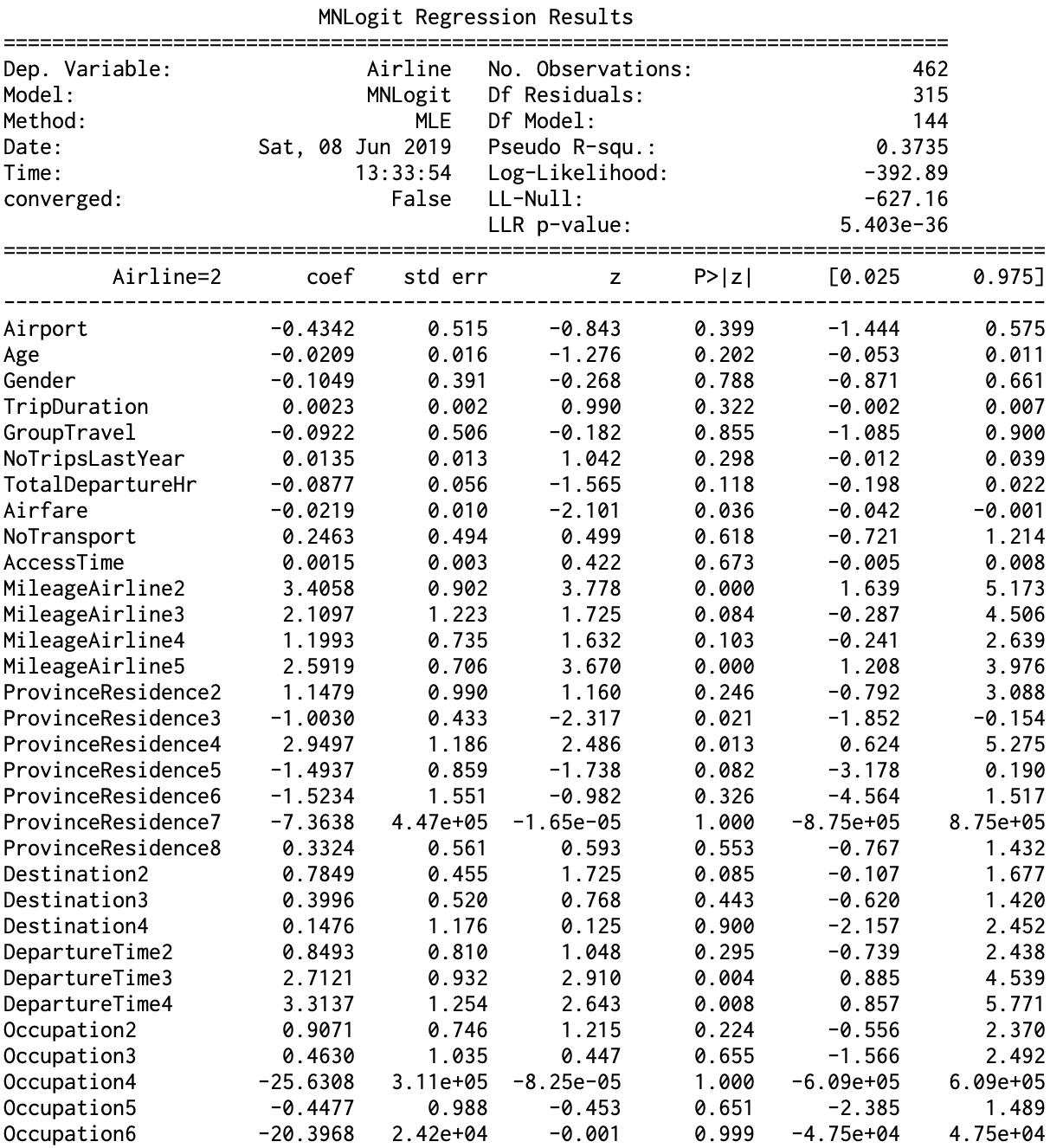
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AIC** | **BIC** | **Loglikelihood** | **McFaddens R2** | **Variables** |
| 354.45 | 433.03 | -158.23 | 0.5058 | ['AccessTime','Destination2','Destination3','Destination4','DepartureTime2','DepartureTime3','Income2','Income3','Income4','Income5','Income6','Income7','Income8','Airline3','Airline2','Airline4','Korean', 'PleasureTrip','const'] |

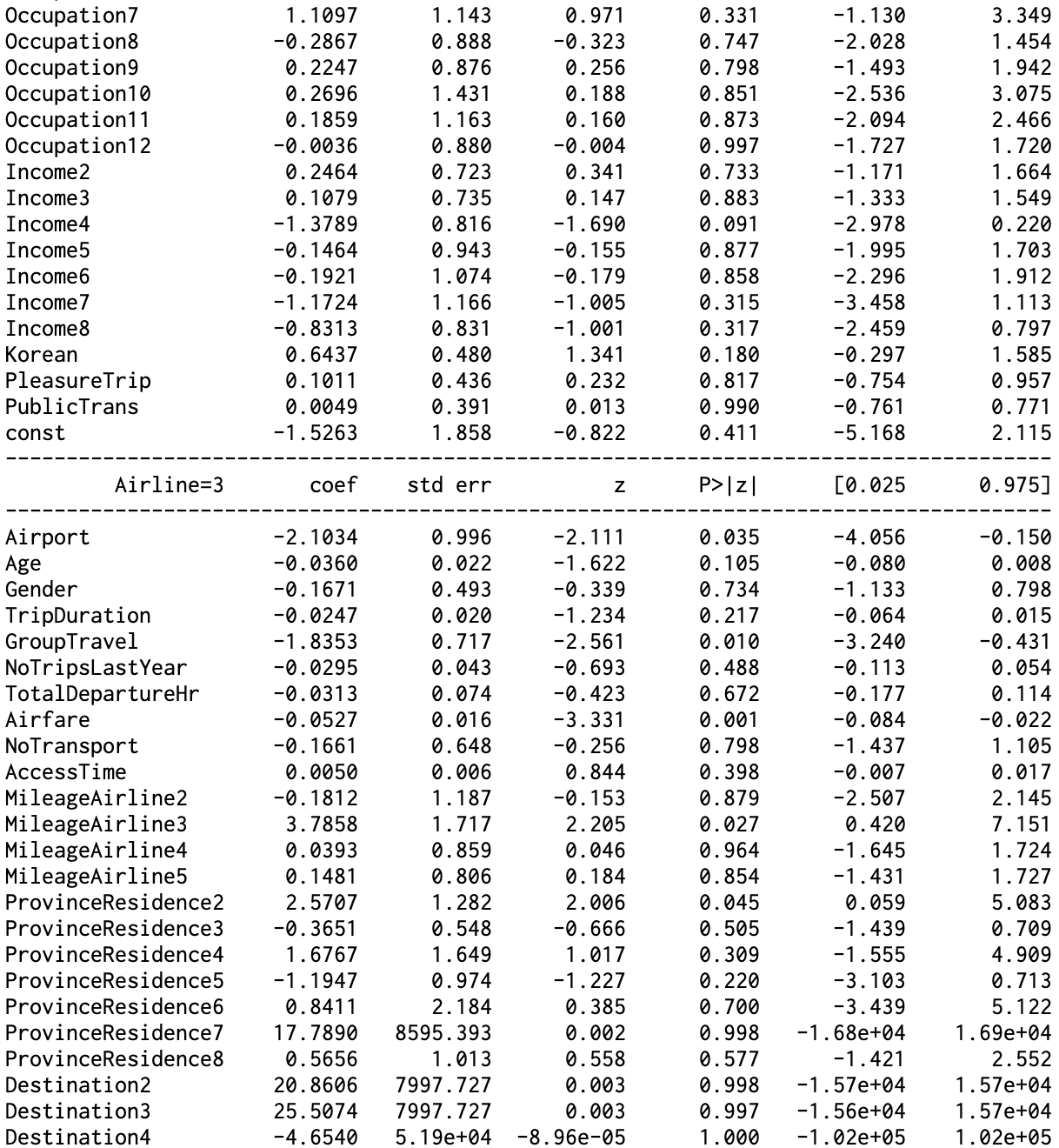
***Table 3.1.2.1 Final selected variables and the AIC, BIC results of airport model***

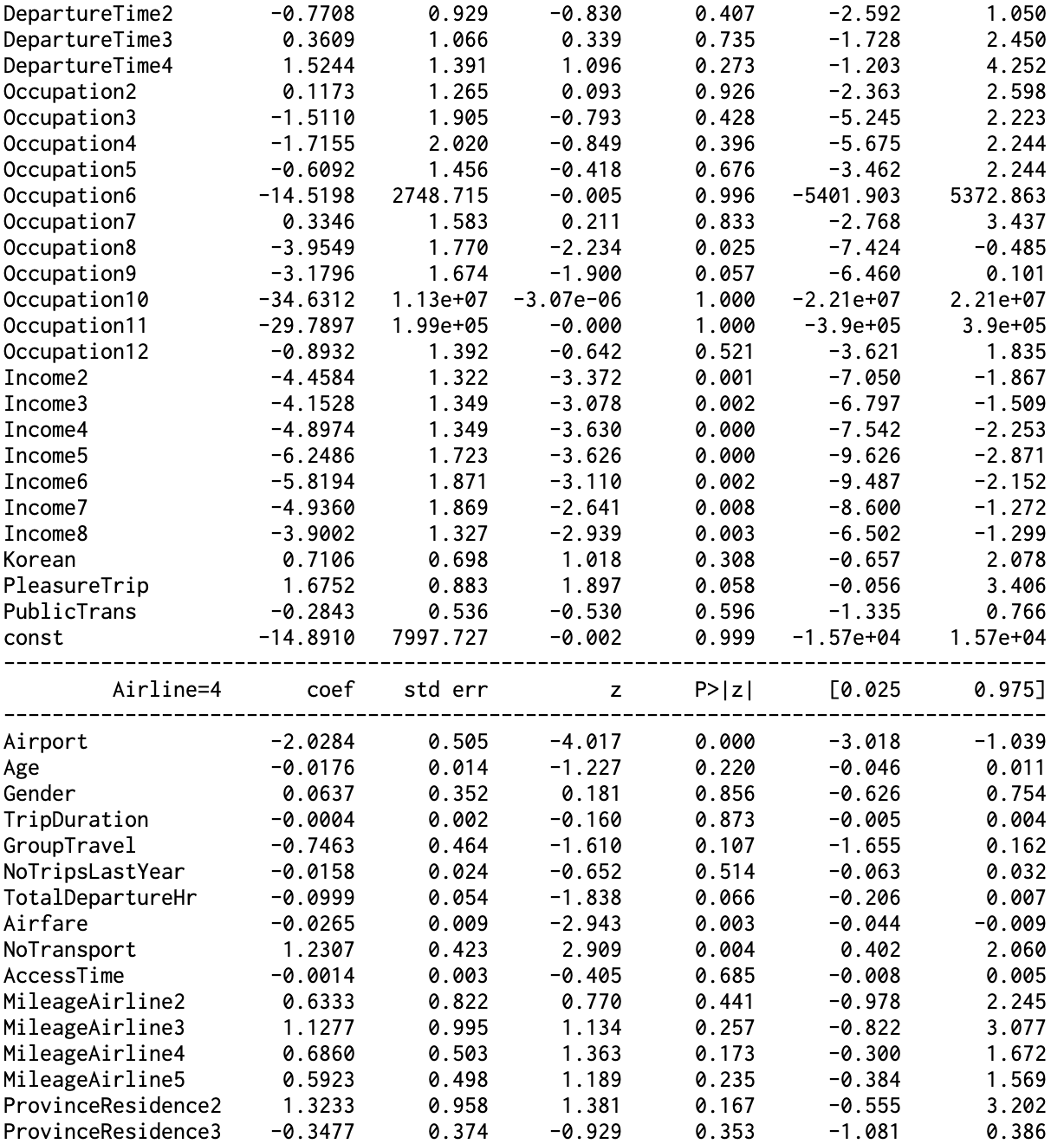
On the other hand, we use a similar method to select the variables for airline model. The result of multinomial model with all the variables is shown below in ***figure 3.1.2.2***. Similarly, as the criteria of high coefficient and low p-value(<0.05), we select ‘Airfare’, ‘GroupTravel’, ‘NoTransport’, ‘PleasureTrip’, ‘Destination’, ‘DepartureTime’, ‘Income’ for further filtering. After comparing the different models’ results, we find that the number of AIC, BIC between the model with the lowest AIC & BIC and the model with the highest R^2 is not significantly different. Therefore, we select the model with the highest R^2 result in this case. The result is shown in ***Table 3.1.2.2.*** We will use ***‘Airfare’, ‘GroupTravel’, ‘NoTransport’, ‘PleasureTrip’, ‘Destination’, ‘DepartureTime’*** and ***‘Income’*** as the the selected independent variables in airline multinomial logistic model.

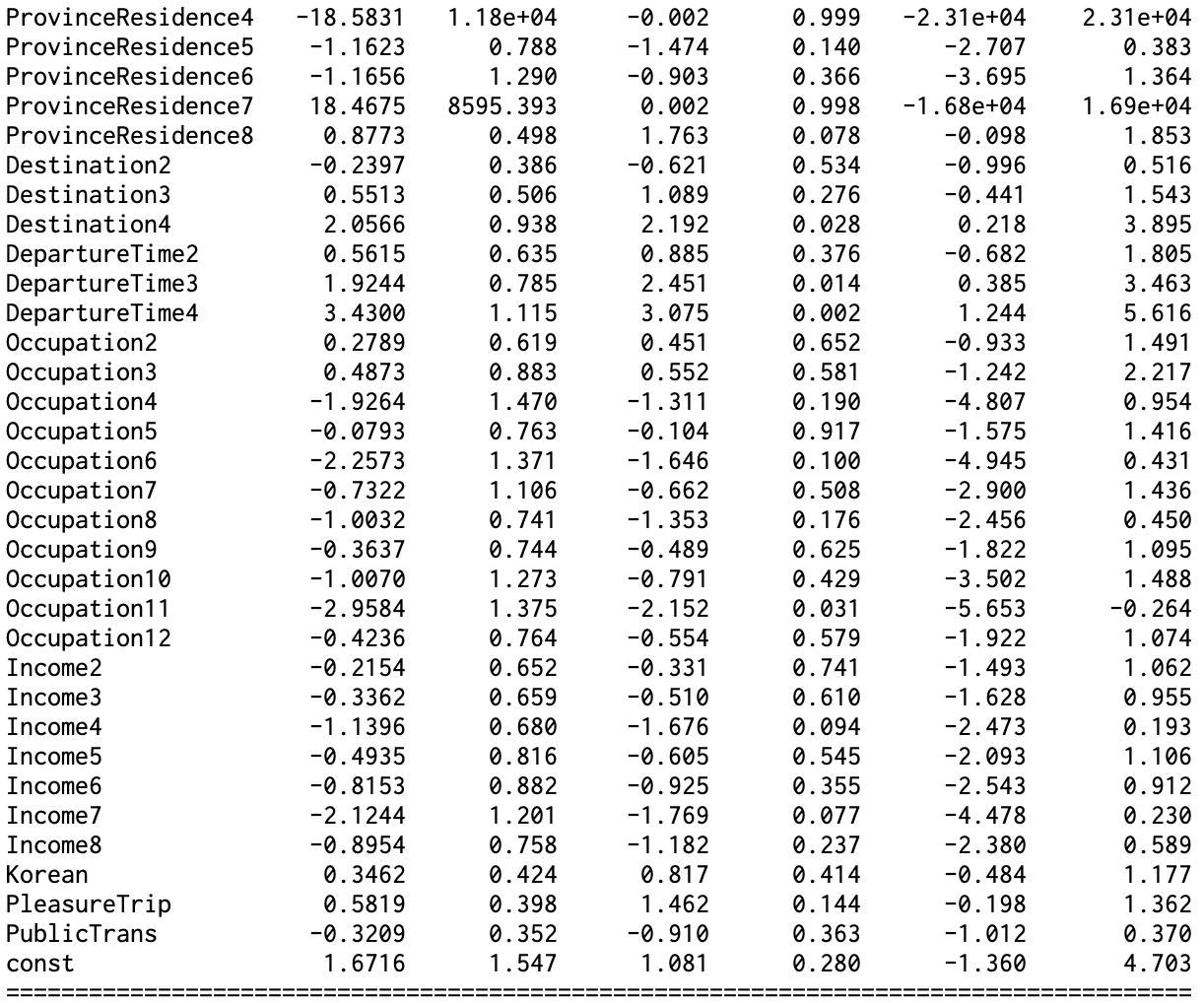
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AIC** | **BIC** | **Loglikelihood** | **McFaddens R2** | **Variables** |
| 1097.12 | 1320.44 | -494.56 | 0.2114 | ['Airfare','GroupTravel','NoTransport', 'PleasureTrip','Destination2','Destination3','Destination4''DepartureTime2','DepartureTime3','DepartureTime4','Income2','Income3','Income4','Income5','Income6','Income7','Income8'] |

***Table 3.1.2.2 Final selected variables and the AIC, BIC results of airport model***









***Figure 3.1.2.2 Airline MNLogit Model with all uncorrelated variables***

**3.1.2 Decision Tree**

The first method we used to build a classification model is the decision tree method. We decided to use decision tree because this is a popular machine learning tool which is easy to interpret the result, handle categorical and numeric variables. Decision tree does not require scaled data and is also able to capture linear and non-linear relationship. In this project, we train the classification decision tree including every single predictor in both model Airport and Airline choice. The first model is trained with default value of parameters (Full tree), then we pruned the tree by implementing the model with different values of max depth, min sample leaf, and min sample split and choosing the smaller tree with higher accuracy. The tree is split by choosing the best split set from a set of possible splits, in order to maximize the information gain at a tree node.

After some experimentation we came up with final models as described in the table below:

|  |  |  |
| --- | --- | --- |
| Tree model parameter | Airport choice | Airline choice |
| Max depth | 8 | 11 |
| Min sample leaf | 1 | 1 |
| Min sample split | 2 | 2 |

### **3.1.3 Nearest Neighbor (KNN)**

The nearest neighbor model is the instance based classifier, which is also known as the “rote learner.” The euclidean distance is used in calculation of distance between records. When determining the class of the new data, the KNN model takes the majority vote of the class labels among the k-nearest neighbors. The variables included in the KNN model is the same as the selected variables from Logit regression model. The numerical data are normalized to remove the impact of different scale.

### **3.1.4 Artificial Neural Network (ANN)**

Artificial Neural Network model is an assembly of interconnected nodes and weighted links. The output node sums up each of its input values according to the weights of its links. When deciding if the node should be activated, the model compares the output node against the defined threshold. The training process of ANN model is the process of learning the weight of the neurons. The variables included in the ANN model is the same as the selected variables from Logit regression model. The numerical variables are normalized to remove the impact from various scale.This study defined the three hidden layers with ten nodes in each layer (10,10,10). The activation method is set to default - “relu”, “identity”, and “logistic” in three models.

### **3.1.5 Naive Bayes**

The naive bayes classifier use the probabilistic framework to solve the classification problems. The variables included in the naive bayes model is the same as the selected variables from Logit regression model. The continuous variables in the model will be discretized or use probability density estimation to calculate probability.

### **3.1.6 Support Vector Machine (SVM)**

Support Vector Machine tries to find a decision boundary that can separate the data into targeted groups. We try three kernel functions: linear, rbf, and polynomial with the input independent variables that suggested by Logit Model. Continuous variables are also scaled before putting into the model.

## **3.2 Result**

### **3.2.1 K-fold Cross Validation**

In order to evaluate the performance of our model, we decided to choose K-fold method. With this method, all of observation in the dataset are eventually used for both training and validating. We chose k=5 because our dataset only has 463 observations. Otherwise, if the k is bigger, it could be overfitting, while if the k is small, it could be underfitting. K = 5 means that we divided the dataset into 5 equal parts. Each part has the same or similar number of observations. We use each part as the validating data separately, while we use unused parts as training data to build the training model. At last, we calculated the mean of the accuracy from 5 different training models’ results as the validation accuracy.

### **3.2.2 Result of All Models**

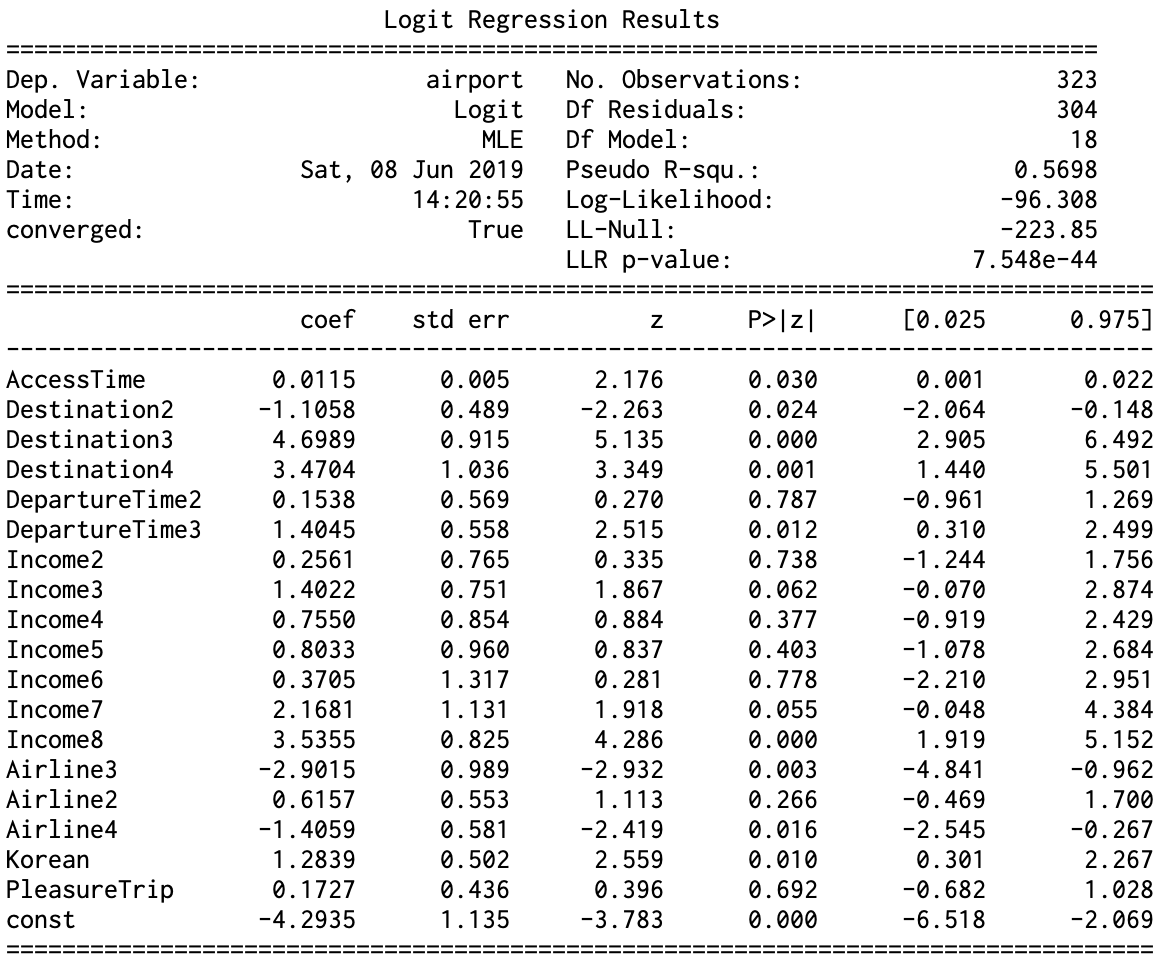
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Airport** | | | | |
| **Model** | **Specification** | **Accuracy (Training)** | **Accuracy (Testing)** | **K-fold Validation k = 5 Avg(Accuracy)** |
| Logit |  | 0.858 | 0.784 | 0.797 |
| Decision Tree | maxdepth=8, minsamleaf =1, minsamsplit = 2,others are default | 0.980 | 0.8129 | 0.758 |
| Decision tree | everything default | 1.000 | 0.805 | 0.747 |
| KNN | k=15 | 0.833 | 0.755 | 0.809 |
| ANN | 10,10,10 (default Activation: relu) | 0.950 | 0.777 | 0.898 |
| ANN | Activation: identity; 10,10,10 | 0.780 | 0.756 | 0.773 |
| ANN | Activation: logistic; 10,10,10 | 0.523 | 0.496 | 0.571 |
| Naive Bayes |  | 0.765 | 0.727 | 0.753 |
| SVM | linear | 0.842 | 0.734 | 0.819 |
| SVM | poly | 0.536 | 0.511 | 0.825 |
| SVM | rbf | 0.848 | 0.719 | 0.860 |
|  |  |  |  |  |
| **Airline** | | | | |
| **Model** | **Specification** | **Accuracy (Training)** | **Accuracy (Testing)** | **K-fold Validation k = 5 Avg(Accuracy)** |
| MN Logit |  | 0.536 | 0.489 | 0.435 |
| Decision Tree | maxdepth=11,minsamleaf =1, minsamsplit = 2others are default | 1.000 | 0.450 | 0.450 |
| Decision tree | everything default | 0.990 | 0.446 | 0.471 |
| KNN | k=15 | 0.507 | 0.403 | 0.476 |
| ANN | 10,10,10 (default Activation: relu) | 0.820 | 0.439 | 0.706 |
| ANN | Activation: identity; 10,10,10 | 0.557 | 0.374 | 0.455 |
| ANN | Activation: logistic; 10,10,10 | 0.331 | 0.295 | 0.320 |
| Naive Bayes |  | 0.272 | 0.245 | 0.483 |
| SVM | linear | 0.570 | 0.460 | 0.544 |
| SVM | poly | 0.356 | 0.273 | 0.689 |
| SVM | rbf | 0.570 | 0.460 | 0.642 |

## **3.3 Summary and Model Comparison**

### **3.3.1 Airport Model**

**Logit:**

Figure 3.3.1.1 shows the result of airport ***logit model***. When holding other variables constant, we could find that ***access time*** is statistically significant, and for every one minute change in access time, the log odds of choosing ICN airport (versus GMP airport) increases by 0.0115. ***Destination*** is also statistically significant, flying to Japan versus flying to China changes the log odds of choosing ICN by -1.1058, flying to Southeast Asia versus flying to China changes the log odds of choosing ICN by 4.6989, while flying to other region except Japan, China and Southeast Asia versus flying to China changes the log odds of choosing ICN by 3.4704. It indicates that the passengers flying to Japan are more likely to choose ICN airport, while the passengers flying to Southeast Asian are more likely to choose GMP airport. ***DepartureTime*** is statistically significant, too, while for the passengers flying between 6pm and 9pm versus between 6am and 12pm, it changes the log odds of choosing ICN by 1.4045, which means passengers who flies between 6pm and 9pm is more likely to choose ICN than between 6am and 12pm. ***Income*** is a special independent variable in this model, because only income 8 (Unknown Income) is statistically significant, so we won’t discuss the influence of income to the choice of airport here. ***Airline*** is also statistically significant. For example, Korean LCC versus Korean Air changes the log odds of choosing ICN by -2.9015, while foreign airline versus Korean Air changes the log odds of choosing ICN by -1.4059. The result indicates that comparing to Korean Air, the passengers with Korean LCC and foreign airlines are more likely to choose GMP airport than ICN airport. ***Korean*** is statistically significant, the result shows that Korean passenger looks more likely to choose ICN than GMP.



***Figure 3.3.1.1 Airport Logit Model***

The **decision tree** method depends mainly on using the information gain metric to determines the attributes which are most useful. In the Airport model, the attributes on top 3 levels of the tree are: Destination, Airfare, DepartureTime,TripDuration, MileageAirline5, and Airline. Destination is considered as the root node of the decision tree. With the optimized parameter, the airport model has generated 29 classification rules. The table below shows the most important rules.

|  |  |  |  |
| --- | --- | --- | --- |
| No | Rule | #Object | #Features |
| 1 | If the destination is Southeast Asia orOther, departure time is from 12 pm - 6 am, and the airline is domestic airlines the airport is I**ncheon**. | 89 | 3 |
| 2 | If the destination is Southeast Asia orOther, departure time is from 12 pm - 6 am, the airline is Foreign airlines,MileageAirline5 =1, the airport is **Incheon**. | 16 | 4 |
| 3 | If the destination is China or Japan, and the Air Fare is from 26.215 to 29, the airport will be **Incheon**. | 9 | 2 |
| 4 | If the destination is Southeast Asia orOther, departure time is from 6 am - 12 pm, MileageAirline5 =0, MileageAirline3 =0, the airport is **Gimpo**. | 12 | 4 |
| 5 | If the destination is Southeast Asia orOther, departure time is from 6 am - 12 pm, MileageAirline5 =0, MileageAirline3 =1, the airport is **Gimpo**. | 1 | 4 |

***Table 3.3.1.2: Airport tree model rules. \*Note: the tree diagram is included in the zip file***

The ***KNN model*** tries the parameters of k=1, k=3, k=5, k=10, k=15. The accuracy of each KNN model decreases when increasing the k values because the small k value leads to a model which is overly fit to the training data set. This result can be perceived from the model where k=1, that the accuracy of the training set equals to 100%. The accuracy gap between the training and testing data in KNN model is large, which is because the model contains too many categorical variables which can not precisely reflect the distance between categories in a meaningful way.

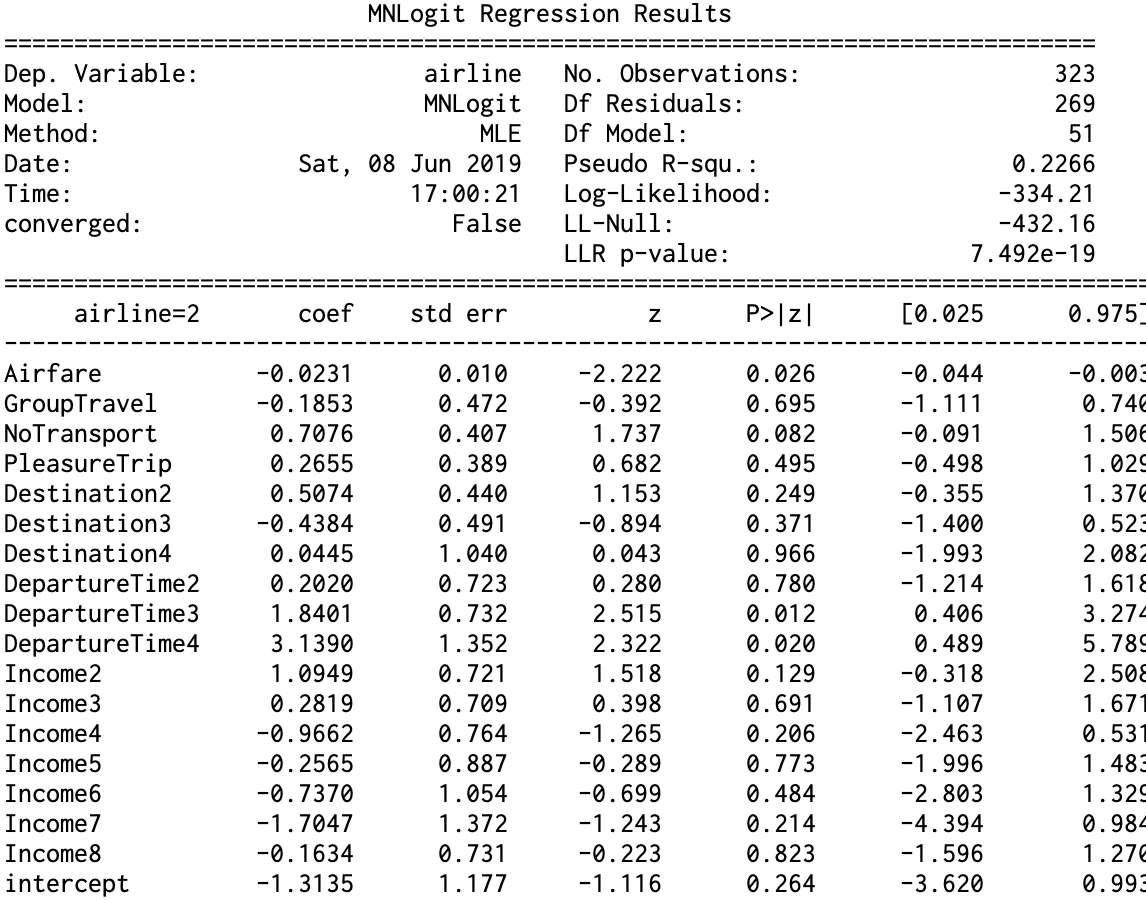
The **ANN model** with “activation = relu” performs the best when predicting the airport. The model with activation method of “identity” derive a satisfactory result. But the predicting power of the model with “activation = logistic” is very low.

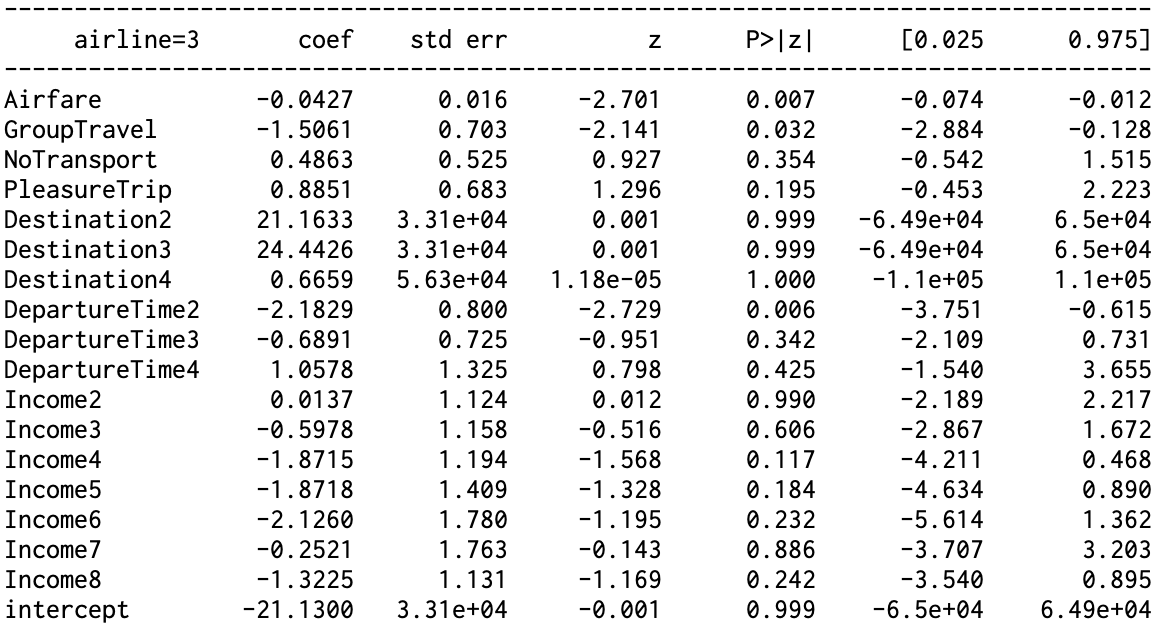
The ***Naive Bayes model***’s result is less accuracy than KNN model’s result. The accuracy of naive bayes is similar to ANN (10,10,10, activation = “identity”) model.

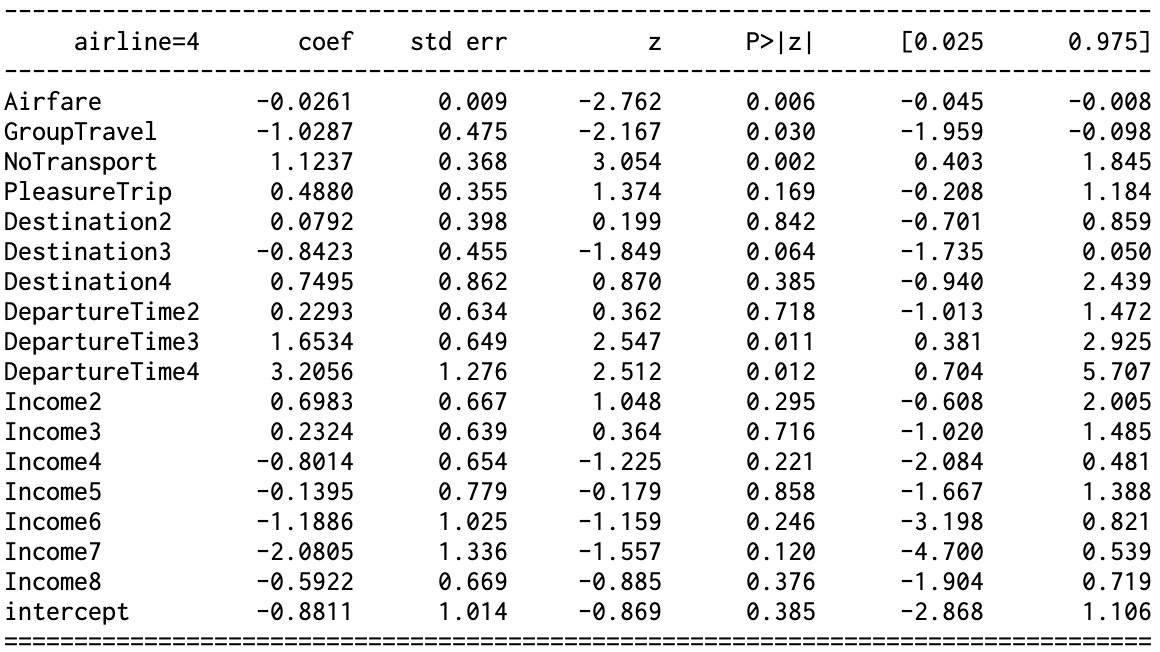
The ***Support Vector Machine***’s result is generally good, especially with “rbf” and “linear” kernel functions.

### **3.3.2 Airline Model**

**MNlogit:**







***Table 3.3.2.1 Airline Multinomial Logit Model***

Table 3.3.2.1 presents the result of airline multinomial logit model. Firstly, we will interpret the result between airline 1 and 2, which are Korean Air and Asiana Airline respectively. All the statistically significant variables are listed in ***Table 3.3.2.2***.

|  |  |  |
| --- | --- | --- |
| Variable | Statistical Explanation | How to affect passengers’ choices |
| Airfare | When holding other factors constant, one increasing unit of airfare will change the log odds of choosing Asiana Airline other than Korean Air by -0.0231. It is statistically significant at 95% confidence level. | Higher airfare is more likely to make passengers choose Korean Air other than Asiana Airline. |
| No.Transport | When holding other factors constant, one more transportation would change the log odds of choosing Asiana Airline other than Korean Air by 0.7076. It is statistically significant at 90% confidence level. | Passengers are more likely to choose Asiana Airline other than Korean Air when the flight contains more times of transportation to arrive at the airport. |
| Destination 3:  Southeast Asia | When holding other factors constant, flying to Southeast Asia other than flying to China would change the log odds of choosing Asiana Airline other than Korean Air by 1.8401. It is statistically significant at 95% confidence level. | Comparing to those who are flying to China, passengers who are flying to Southeast Asia are more likely to choose Asiana Airline other than Korean Air. |
| Destination 4:  Others but China, Japan or Southeast Asia | When holding other factors constant, flying to other regions excluding China, Japan or Southeast Asia other than flying to China would change the log odds of choosing Asiana Airline other than Korean Air by 3.1390. It is statistically significant at 95% confidence level. | Comparing to those who are flying to China, passengers who are flying to other regions excluding China, Japan or Southeast Asia are more likely to choose Asiana Airline other than Korean Air. |

***Table 3.3.2.2 Result of “Choosing Korean Air or Asiana Airline”***

Similarly, all the statistically significant variables of “choosing Korean Air or Korean LCC” are listed in ***Table 3.3.2.3***.

|  |  |  |
| --- | --- | --- |
| Variable | Statistical Explanation | How to affect passengers’ choices |
| Airfare | When holding other factors constant, one increasing unit of airfare will change the log odds of choosing Korean LCC other than Korean Air by -0.0427. It is statistically significant at 95% confidence level. | Higher airfare is more likely to make passengers choose Korean Air other than Korean LCC. |
| GroupTravel | When holding other factors constant, group travel other than single travel would change the log odds of choosing Korean LCC other than Korean Air by -1.5061. It is statistically significant at 95% confidence level. | Comparing to flying by themselves, passengers are more likely to choose Korean Air other than Korean LCC when they fly with someone together. |
| DepartureTime2:  12pm - 6pm | When holding other factors constant, flying between 12pm and 6pm would change the log odds of choosing Asiana Airline other than Korean Air by -2.1829. It is statistically significant at 95% confidence level. | Passengers are more likely to choose Korean Air other than Korean LCC if their flight is in the afternoon between 12pm and 6pm. |

***Table 3.3.2.2 Result of “Choosing Korean Air or Korean LCC”***

We could use the same way to explain the result of “choosing Korean Air or Foreign Airlines”. The result presents below in ***Table 3.3.2.3***.

|  |  |  |
| --- | --- | --- |
| Variable | Statistical Explanation | How to affect passengers’ choices |
| Airfare | When holding other factors constant, one increasing unit of airfare will change the log odds of choosing Foreign Airlines other than Korean Air by -0.0261. It is statistically significant at 95% confidence level. | Higher airfare is more likely to make passengers choose Korean Air other than Foreign Airlines. |
| GroupTravel | When holding other factors constant, group travel other than single travel would change the log odds of choosing Foreign Airlines other than Korean Air by -1.0287. It is statistically significant at 95% confidence level. | Comparing to flying by themselves, passengers are more likely to choose Korean Air other than Foreign Airlines when they fly with someone together. |
| No.Transport | When holding other factors constant, one more transportation would change the log odds of choosing Foreign Airline other than Korean Air by 1.1237. It is statistically significant at 90% confidence level. | Passengers are more likely to fly with Foreign Airline other than Korean Air when the flight contains more times of transportation to arrive at the airport. |
| Destination 3:  Southeast Asia | When holding other factors constant, flying to Southeast Asia other than flying to China would change the log odds of choosing Foreign Airline other than Korean Air by -0.8423. It is statistically significant at 90% confidence level. | Comparing to those who are flying to China, passengers who are flying to Southeast Asia are more likely to choose Foreign Airlines other than Korean Air. |
| DepartureTime3:6pm - 9pm | When holding other factors constant, flying between 6pm and 9pm other than morning from 6am to 12pm would change the log odds of choosing Asiana Airline other than Korean Air by 1.6534. It is statistically significant at 95% confidence level. | Passengers are more likely to choose Foreign Airline other than Korean Air if their flight is in the evening between 6pm and 9pm other than the flight in the morning from 6am to 12pm.. |
| DepartureTime4:9pm - 6am | When holding other factors constant, flying between 9pm and 6am other than morning from 6am to 12pm would change the log odds of choosing Foreign Airlines other than Korean Air by 3.2056. It is statistically significant at 95% confidence level. | Passengers are more likely to choose Foreign Airlines other than Korean Air if their flight is during the mid-night from 9pm to 6am other than the morning between 6am and 12pm. |

***Table 3.3.2.3 Result of “Choosing Korean Air or Foreign Airlines”***

In summary, the passengers’ choice of 4 kinds of airline could be concluded as:

1. Passengers with a high budget is prone to choose Korean Air over other airlines.
2. Passengers who travel in group is prone to choose Korean Air over Korean LCC or Foreign Airlines.
3. Passengers who travel to Southeast Asia prefer Asiana Airline and Foreign Airlines over Korean Air.
4. Passengers who travel to the airport with convenience, less transfer between transportation methods, prefer the Korean Air over Asiana Airline or Foreign Airlines.
5. Passenger who travels during the night time prefer Foreign Airlines. And passengers who travels in the afternoon prefer the Korean Air and Korean LCC. Another possible reason could be only Foreign Airlines provide the flight in the night.

In the Airline choice model, the decision tree classifier method has a lower accuracy performance, since Airline model has more than 2 choices. However, based on the information gain metric, the attributes was ranked similarly to the airport choice with Destination as root following by Airfare, DepartureTime, FrequentFlightDestination6, ProvinceResidence, TripDuration, and FlyingCompanion. The Airline choice tree model is larger and deeper than Airport tree. It generated 91 rules. That is much more than other trees. However, this model performs better accuracy and robust. Some rules are presented in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| No | Rule | #Object | #Features |
| 1 | If destination is China or Japan, departure time is from 6am to 6pm, the Frequent flight destination is not North/South America and Income level <=4, the Airline will be **Asiana Airlines**. | 4 | 4 |
| 2 | If destination is China or Japan, departure time is from 6am to 6pm, the Frequent flight destination is not North/South America and Income level > 4, the Airline will be **Foreign Airlines**. | 3 | 4 |
| 3 | If destination is Southeast Asia or other, Airfare > 58.5, TripDuration > 7.5, AccessCost <= 8700.0, The Airline will be **Foreign Airlines**. | 5 | 4 |
| 4 | If destination is China or Japan, DepartureTime is from 6am to 6pm, FrequentFlightDestination6 =0, Airfare <=45.555, Airfare <= 48.195, the Airline will be **Korean Air(KE)** | 14 | 4 |
| 5 | If destination is China or Japan, departure time is from 6pm to 6am, ProvinceResidence <= 4.5, Income <= 4.5, FrequentFlightDestination6 = 0, MileageAirline2 = 0, the Airline will be **Foreign Airlines**. | 24 | 6 |

***Table 3.3.2.4: Airline tree model rules. \*Note: the tree diagram is included in the zip file***

The ***KNN models*** apply to the airline classification is similar to the model for airport. The parameter of k is also set to 1, 3, 5, 10, and 15. However, the results show that with the increase of k value, the accuracy of training set increase, but the accuracy of testing set varies around 40% level.

The ***ANN models*** used to predict the airline get similar results. The ANN model with “activation = logistic” produces a bad result.

The ***Naive Bayes model*** produces the worst result when predicting multiple choices. The result here is the worst among all models.

The ***Support Vector Machine***’s result is generally good, especially with “rbf” and “linear” kernel functions.

# **4 Conclusions and Recommendations**

This project attempts to use data mining methods to analyze and classify passengers’ choice behavior in airport and airline. We ended up using decision tree model which gave us the highest accuracy and robusted result. The table(our table of accuracy) shows that Decision tree was the highest performing method, because its had the lowest training error and consistent accuracy in both testing and k-fold cross validation. By choosing specific parameter to prune the tree, it also achieved the best balance between accuracy, simplify, and robustness.

With the result from decision tree model, managers can extract the knowledge, such knowledge can be used to get an insight of air travellers’ decision patterns in order to utilize the airport facilities, improve their policies, enhance their strategies, and improve the quality of management system. For instance, according to rule number 1 of airport model, manager can arrange those trips to Southeast Asia which will depart from 6am to 12pm to **Gimpo** airport, and which will depart from 12pm to 6am to **Incheon** airport. In the Airline choice, there is a pattern that people to travel to Southeast Asia or Others, departure from 6am to 6pm, live in Seoul, Incheon, Kyungki-do, or Chungcheong-do and Income level is lower than 5 tend to choose **Foreign Airline**. Therefore, to attract these people using domestic airline, manager can adjust the airfare to match with passenger income.

In closing, it should be noted that our study has some limitations. First, different ways to regroup dummy variables can be explored. This can help classify task more precisely. Another possible trial is to collect a larger data set so that the training accuracy and testing accuracy would be closer. Furthermore, we also need to point out that the way we impute missing values might have problems. Through deeper understanding the survey background and a larger data set, a better way to impute the missing values can be better for modeling.