

BUAN 5310 Statistical Learning

An Exploration of Airport and Airline Purchasing Decisions in Korea

Annie Hellebust, Qiuye Liu, Yifan Xiang, Ying Xue

2019 SQ

Seattle University

## 1. INTRODUCTION

In this project, we aimed to understand individuals' criteria when choosing between Incheon and Gimpo airports near Seoul as well as their criteria when choosing among airlines (Korean Air, low-cost Korean carriers, Asiana Airlines, and other foreign carriers).

For this project, we tested several models both to measure airport choice and airline choice. These models included logit, decision tree, support vector machine, and neural network models. To determine the best model, we focused most on accuracy, as we were interested in maximizing the number of correctly identified test cases and both the airline and airport variables were not highly skewed.

In this paper, we will focus our discussion on the decision tree and logit models because these models offer the most insight into understanding the variables most individuals take into account when they choose an airport. Additionally, the decision tree model consistently outperformed all other models in terms of accuracy for predicting both airport and airline choice.

## 2. DATA CLEANING

Our data set was somewhat messy, with missing values in nearly all variables. We elected to delete all observations with a missing value for either airline or airport, as these were the key variables our model sought to explain. Additionally, there were very few observations with missing values for airline or airport.

Further, we elected to remove variables with >30% missing values from our models, with the exception of airfare. For airfare, we filled in the missing values with the median airfare value given an observation values for airport, airline, and flight destination.

For other numeric variables with missing values, we also filled in missing values with averages. For numeric variables with outliers, we reassigned the outlying observations to a capped value.

For categorical variables, we treated missing values as their own category. For heavily skewed categorical variables, we created super-categories from multiple groups using domain knowledge.

## 3. MODEL AND METHODOLOGY

We split each dataset using hold-out method with 70% of the data as training data and 30% as testing data. Confusion matrix, training accuracy and test accuracy were generated to evaluate the models.

To model airport choice and airline choice, we created four different types of supervised models to understand the variables individuals take into consideration when choosing between a flight at GMP vs ICN airport: logit, decision tree, neural network, and support vector machine. Although we attempted many variations on each model (with different tuning parameters), we included only our highest-performing models for each model type in our attached code. Additionally, this paper will only discuss the logit and decision tree model in depth, as these models offer the greatest amount of insight into the decision-making process because they are highly interpretable.

## 4. AIRPORT MODEL

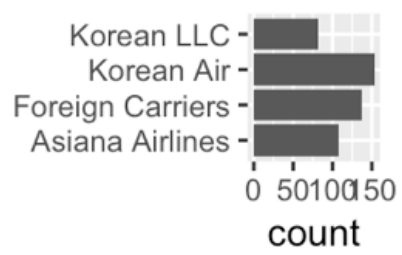
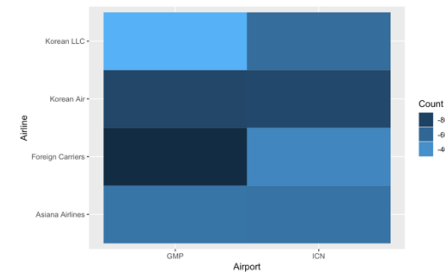
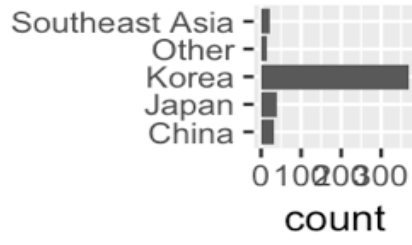
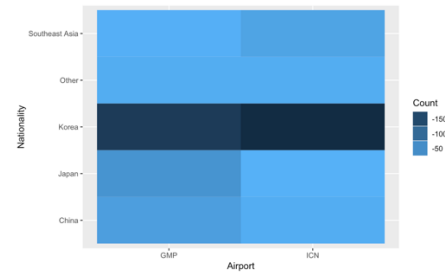
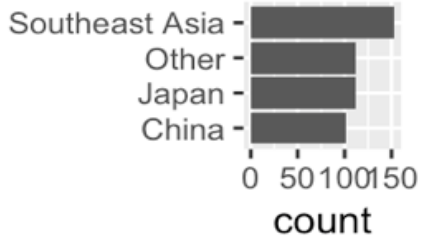
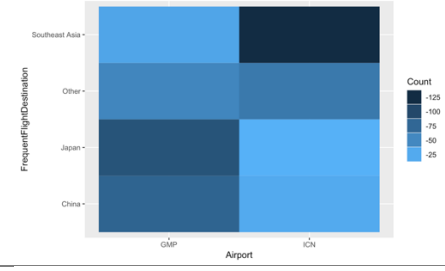
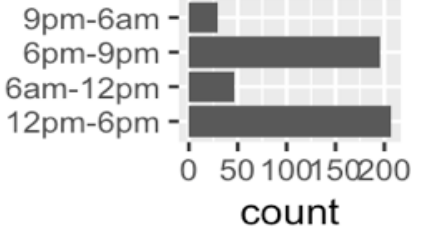
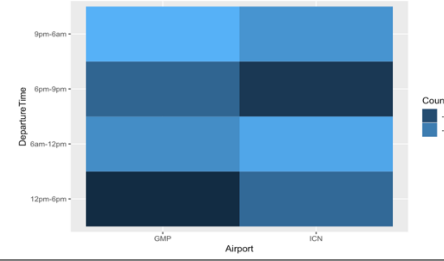
For this model, random assignment would have resulted in 50% accuracy; all models improved upon this by over 25 percentage points.

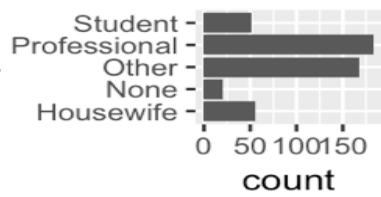
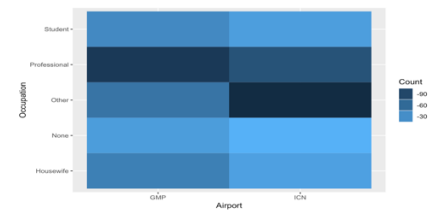
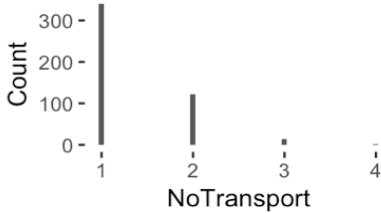
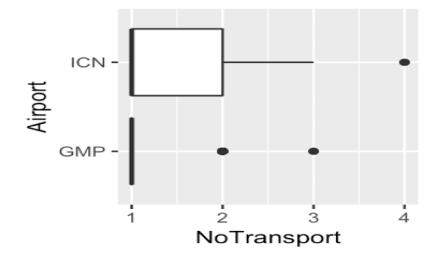
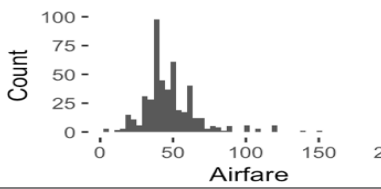
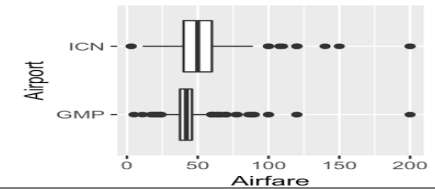
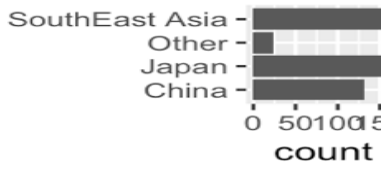
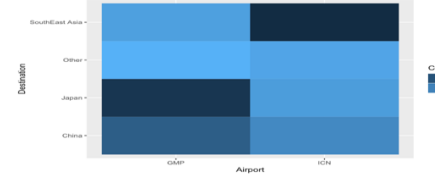
**Table: Airport Model Accuracy Metrics**

Model	Decision Tree	Logit	SVM (radial kernel, cost=1.8)	Neural Network (2 hidden layers: 350 nodes and 500 nodes)
Training accuracy	0.92	0.88	0.86	0.99
Test accuracy	0.83	0.83	0.83	0.85

## 4.1. Exploratory Data Analysis and Variable Selection

**Table: Variable Selection**

Variable	Variable description and reason it was included in our model	Graph of variable's distribution	Graph of variable vs. airport choice
<b>Airline</b>	The airline of the flight; chosen because it passed both chi squared and step-wise tests and because different airlines fly from different airports		
<b>Nationality</b>	Nationality of buyer; chosen because it passed both chi squared and stepwise tests and because individuals from non-Korean airlines may be more likely to fly from a global rather than regional airport		
<b>Frequent Destination</b>	<b>Flight</b> Location to which buyer frequently flies; chosen because it passed both chi squared and stepwise tests and because individuals who frequently fly longer distances may be more likely to fly from a global rather than a regional airport		
<b>Departure Time</b>	Time at which this particular flight departed (as a factor variable); chosen because it passed both chi squared and stepwise tests and because certain departure times are more favorable than others		

<b>Occupation</b>	Occupation of buyer; chosen because it passed both chi squared and stepwise tests and because job may influence purchasing preference; further, occupation is correlated with income/spending ability and use of business class vs. economy	 
<b>NoTransport</b>	How many forms of transit the buyer must ride to reach the chosen airport; chosen because it passed both anova and stepwise tests and individuals prefer airports that are simpler to get to; further, NoTransport is correlated with the providence in which an individual lives as well as the expense associated with traveling to the airport and thus can stand in for those variables as well	 
<b>Airfare</b>	Price the buyer paid for a given ticket; chosen because it passed the anova test and because most airline tickets are commodity goods purchased (for a given destination) on price alone	 
<b>Destination</b>	Location to which this particular flight flew; chosen because it passed both chi squared and stepwise tests and because certain destinations may be more common for a global rather than a regional airport	 
<b>Interaction term between Airfare and Destination</b>	Impact of airfare given a particular destination; chosen because airfare is most dependent upon destination. Including this variable enables a more reasonable price comparison, as it controls for destination.	

#### 4.2. Logit Model

For our logit model, we considered ICN to be the baseline airport and examined the impact of airline, nationality, frequent flight destination, destination, departure time, occupation, number of types of transit to reach the airport, and airfare on choice of airport. A discussion of our findings can be found in the table below.

**Table: Airport Logit Model Analysis by Variable**

Explanatory Variable	Coefficient	P-Value	Commentary
NoTransport	0.3031	0.214	In our model, we used NoTransport to represent the distance to one airport or another; we found that individuals tend to pick airports where they need to take fewer modes of transport to reach their flight. Controlling for the other variables, individuals who required more forms of transit to reach the airport were more likely to fly out of ICN. While informative with respect to understanding the impact of transportation, this result is not statistically significant.
Airfare	-0.5184	0.172	Excluding the impact of destination on airfare, we see overall that flights purchased out of GMP were more expensive than those purchased out of ICN. However, this is mitigated by the fact that for each given destination (except China, which is the baseline), we see that individuals will pay more to fly out of ICN.
Airfare-Destination interaction term: China	Baseline		
Airfare-Destination interaction term: Japan	1.2752	0.019	
Airfare-Destination interaction term: Other SE Asia	2.4720	0.009	
Airfare-Destination interaction term: Other	0.9805	0.329	
Airline: Asian Airlines	Baseline		We found that individuals who used Korean and Asian airlines were the most likely to use ICN, while those who used foreign carriers or Korean low-cost carriers were much more likely to use GMP. This makes sense, as GMP is an older, likely less expensive airport to fly from (and thus will have more low-cost carriers).
Airline: Foreign Carriers	-1.1298	0.063	
Airline: Korean Air	0.0746	0.891	
Airline: Korean LLC	-0.6846	0.450	
Nationality: Chinese	Baseline		Controlling for other variables in the

Nationality: Japanese	-2.4848	0.187	model, we found that Korean citizens were most likely to use GMP, while Japanese citizens and individuals from outside SE Asia were more likely to fly from ICN.
Nationality: Korean	1.5531	0.069	
Nationality: Other SE Asian	17.2712	0.998	
Nationality: Other	-0.0823	0.952	
Frequent destination: China	Baseline		Individuals who frequently fly to destinations in China or Japan were the most likely to select GMP airport; all other locations were more likely to fly to ICN. This is intuitive, as ICN is typically used for long-haul flights while GMP is a regional airport.
Frequent destination: Japan	0.0991	0.886	
Frequent destination: SE Asia	1.8727	0.002	
Frequent destination: Other	1.5458	0.006	
Destination: China	Baseline		Individuals flying to destinations in China were the most likely to select ICN airport; all other locations were more likely to use GMP. Given the addition of the Airfare:destination interaction term, this finding excludes the impact of airfare.
Destination: Japan	-3.0527	0.008	
Destination: Other	-0.3299	0.887	
Destination: SE Asia	-2.3972	0.204	
Departure Time: 12pm-6pm	Baseline		We find that individuals departing during the 6am-12pm time block are most likely to fly from GMP. Those leaving between 6pm and 6am are most likely to fly from ICN.
Departure Time: 6am-12pm	-1.9373	0.005	
Departure Time: 6pm-9pm	1.5864	0.001	
Departure Time: 9pm-6am	3.6174	0.003	
Occupation: Housewife	Baseline		Individuals without an occupation, whose occupation is housewife, or whose occupation is student, are most likely to fly from GMP. Those with a professional occupation or any other form of occupation are more likely to fly from ICN.
Occupation: None	-34.0862	0.997	
Occupation: Other	1.6576	0.013	
Occupation: Professional	1.1707	0.069	
Occupation: Student	-0.0604	0.941	
Intercept	-3.3891	0.000	

**Table: Airport Logit Model Confusion Matrix**

		Predicted	
Actual		GMP	ICN
	GMP	59	11
	ICN	13	61

**Table: Airport Logit Model Precision and Recall (by Airport)**

	GMP	ICN
Precision	0.82	0.85
Recall	0.84	0.82

We found that our logit model for airport was approximately equally good at predicting ICN vs. GMP, with an overall accuracy of 0.83.

#### 4.3. Decision Tree Model

In decision tree models, variables that appear at the top of the tree are typically more important in determining a given output than those that are present at the bottom of the tree. This is because decision trees typically use a greedy approach as the tree is built. Our decision tree model indicates that destination, airfare, and departure time were the most important determining factors when selecting an airport. A detailed visual description of our decision tree can be seen below, where we see that the flight destination of southeast Asia and a departure time between 6am and 12pm heavily impact airport choice.

**Table: Airport Decision Tree Model Confusion Matrix**

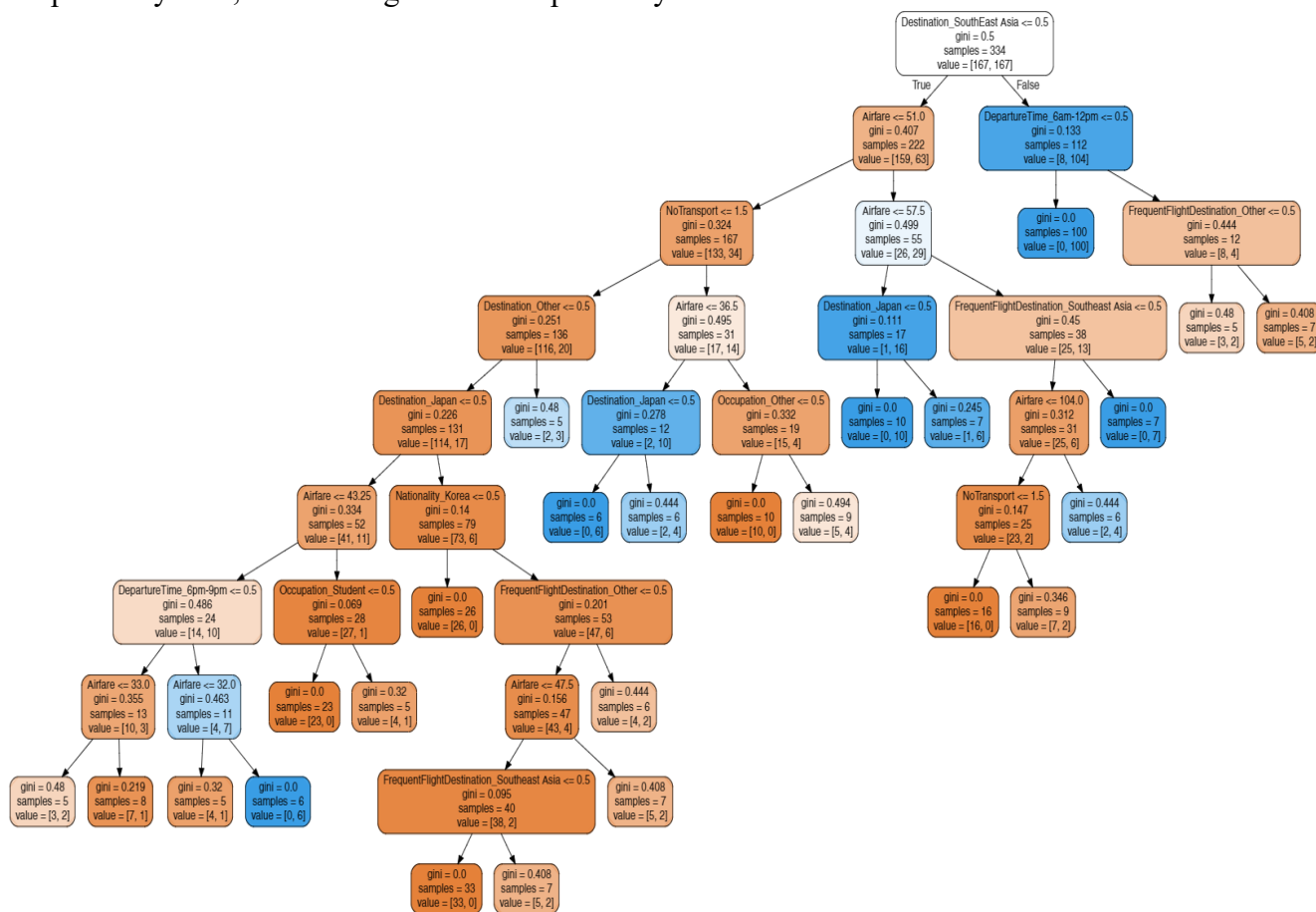
		Predicted	
Actual		GMP	ICN
	GMP	65	5
	ICN	20	54

**Table: Airport Decision Tree Model Precision and Recall (by Airport)**

	GMP	ICN
Precision	0.76	0.92
Recall	0.93	0.73

We found that our decision tree model for airport most often miscategorized ICN as GMP, though errors in the other direction were rare (only 20% of total errors). For this model, our overall accuracy was 0.83.

Below, we have a figure representing our tree model. In this figure, each color represents one airline with most people chosen at a certain node. The darker the color, the higher the purity. Blue nodes are primarily ICN, while orange nodes are primarily GMP.



## 5. AIRLINE MODEL

For this model, random assignment would have resulted in 25% accuracy; all models improved upon this by over 30 percentage points.

**Table: Airline Model Accuracy Metrics**

	Decision Tree	Logit	SVM (RBF performed best)	Neural Network
Training Accuracy	0.72	0.58	0.58	0.95
Test Accuracy	0.55	0.50	0.55	0.56

We also regrouped the levels of airlines into two: Korean airlines (includes Korean Air, Korean LLC, and Asian Airlines) vs. foreign carriers. For this model, random assignment would have resulted in 50% accuracy. The same variables selection were applied and stepwise produce the lower AIC of 553.19. Our highest test accuracy of 0.7153 when using Linear SVM.

**Table: Airline Model Accuracy Metrics: Domestic vs. Foreign Carriers**



	Decision Tree	Logit Regression	SVM (Linear performed best)	Neural Networks
Training Accuracy	0.84	0.75	0.77	0.99
Test Accuracy	0.68	0.70	0.72	0.71

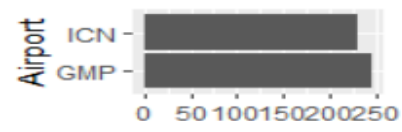
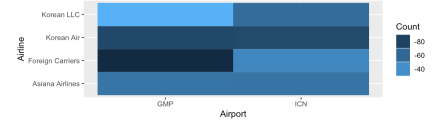
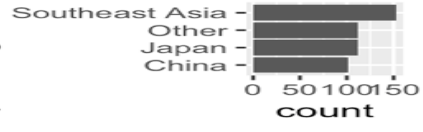
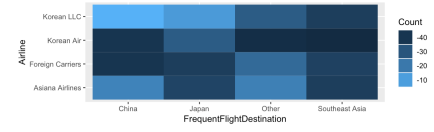

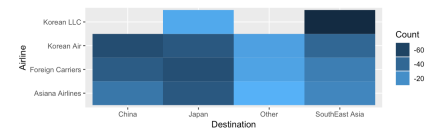
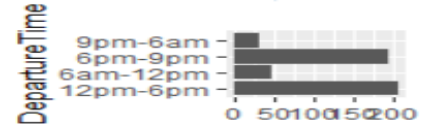
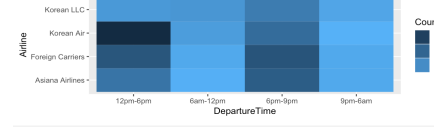
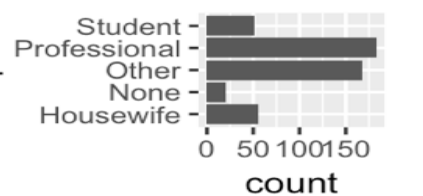
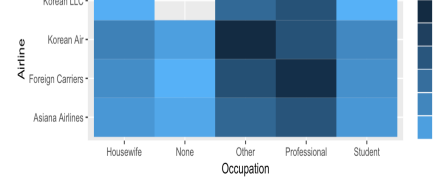
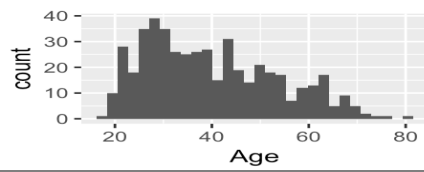
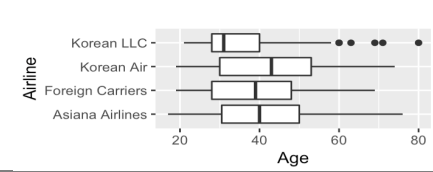
### 5.1. Exploratory Data Analysis and Initial Variable Selection

ANOVA, Chi-squared tests and stepwise regression were used to identify an initial set of independent variables. We used the methods with lowest AIC to decide which subset of variables will be used in the models. Because the difference between AIC are small, after testing all 3 sets of variables across all the models, finally select our variables for our final logit and decision tree models based on the measurements such as difference between training and testing accuracy and importance scores.

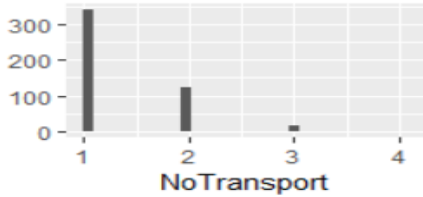
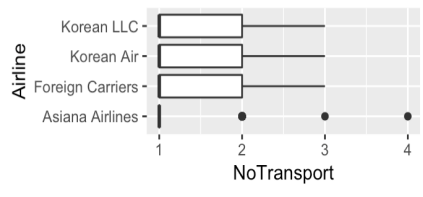
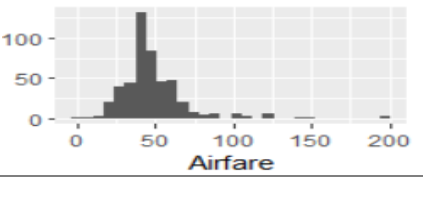
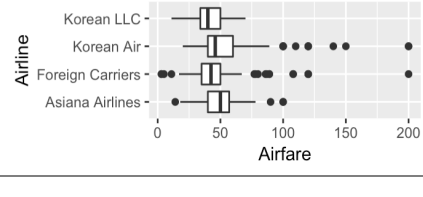
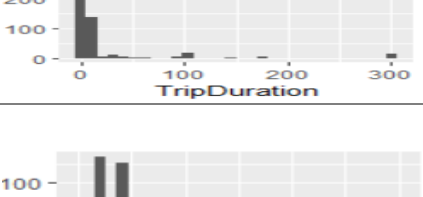

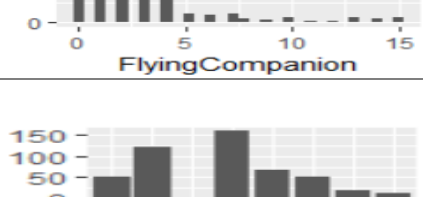
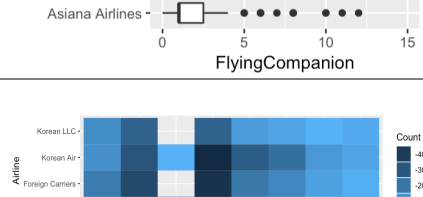
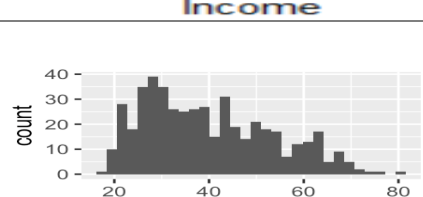
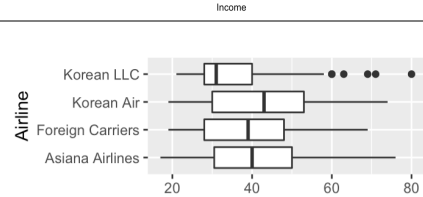


#### *Initial Variables selected by different methods*

Selection Method	AIC	Initial Selected Variable
ANOVA / Chi-squared tests	1105.15	Airport, Nationality, TripPurpose, ProvinceResidence, GroupTravel, FrequentFlightDestination, Destination, DepartureTime, Occupation, Age, TripDuration, FlyingCompanion, NoTripsLastYear, Airfare
Forward stepwise	1068.14	Airport, Destination, Airfare, DepartureTime, FlyingCompanion, NoTransport, FrequentFlightDestination, TripDuration, NoTripsLastYear
Backward stepwise	1070.10	Airport, TripDuration, FlyingCompanion, FrequentFlightDestination, Destination, DepartureTime, SeatClass, Airfare, NoTransport, Occupation, Income

**Table: Airline Model Variable Selection**

Variable	Variable description and reason it was included in our model	Graph of variable's distribution	Graph of variable vs. airline choice
Airport	The airport where the airline is based or flying from; chosen because it passed both chi squared and step-wise tests and because different airlines fly from different airports		
Frequent Flight Destination	Location to which buyer frequently flies; chosen because it passed both chi squared and stepwise tests and because individuals who frequently fly to a destination prefer a certain airline		
Destination	Location to which this particular flight flew; chosen because it passed both chi squared and stepwise tests and because airlines only fly to certain destinations		
Departure Time	Time at which this particular flight departed (as a factor variable); chosen because it passed both chi squared and stepwise tests and because certain departure times are more favorable than others		
Occupation	Occupation of buyer; chosen because it passed both chi squared and stepwise tests and because job may influence purchasing preference; further, occupation is correlated with income/spending ability and use of business class vs. economy		
NoTripsLastYear	Numbers of trips last year; chosen because it passed both Anova test and stepwise test		

# An exploration of airport and airline purchasing decision in Korea

NoTransport	How many forms of transit the buyer must ride to reach the chosen airport; chosen because it passed stepwise(forward) test and individuals prefer airports that are simpler to get to and the airlines that fly from that airport		
Airfare	Price the buyer paid for a given ticket; chosen because it passed the anova test and because most airline tickets are commodity goods purchased (for a given destination) on price alone		
TripDuration	Numbers of day for this trip; chosen because it passed both chi squared and stepwise tests and because trip duration is related to the trip purpose which may affect the airline chosen		
FlyingCompanion	How many people are flying with you; chosen because it passed both stepwise test and anova test and it is correlated with GroupTravel. Travel individually or with group may affect the airline choice		
Income	Income of an individual(categorical); chosen because it passed stepwise test and because income affects one's purchasing power		
Age	Age of an individual(numeric); chosen because it passed Anova test and age is relative to purchasing power and certain airline may have market strategy target at certain age group and		

## 5.2. Logit Model

Multinomial logistic regression is frequently used to analyze econometric discrete choice problems based on utility maximization theory. The utility maximization rule states that an individual will select the alternative from a set of available alternatives that maximizes his or her utility, and the utility could be constructed from a set of weights that are linearly combined with the explanatory variables (attributes) of a given observation using a dot product:  $\text{score}(\mathbf{X}_i, k) = \beta_k \cdot \mathbf{X}_i$ , where  $\mathbf{X}_i$  is the vector of explanatory variables describing observation  $i$ ,  $\beta_k$  is regression coefficients corresponding to Choice  $k$ , and  $\text{score}(\mathbf{X}_i, k)$  is the U(Utility) associated with assigning observation  $i$  to Choice  $k$ . The predicted choice is the one with greater utility of all other alternatives in the individual's choice set. In this project, If UKorean LLC,  $c > U_{\text{Asiana Airlines}, c}$ , the consumer would choose Korean LLC. Thus, in this model, it's crucial to figure out which are important attributes or characteristics of individuals that affect the consumers' choice of airlines and to calculate the difference in utility between pairs of airline alternatives, particularly difference and the sign of the coefficient is positive or negative.

MNLogit Regression Results						
Dep. Variable:	Airline	No. Observations:	334			
Model:	MNLogit	Df Residuals:	271			
Method:	MLE	Df Model:	60			
Date:	Sun, 09 Jun 2019	Pseudo R-squ.:	0.3224			
Time:	17:33:29	Log-Likelihood:	-306.65			
converged:	False	LL-Null:	-452.52			
		LLR p-value:	3.553e-32			
=====						
Airline=Foreign Carriers	coef	std err	z	P> z	[0.025	0.975]
Intercept	-2.4835	1.193	-2.082	0.037	-4.821	-0.146
Airport[T.ICN]	-1.6438	0.541	-3.040	0.002	-2.703	-0.584
Destination[T.Japan]	1.1299	1.197	0.944	0.345	-1.217	3.476
Destination[T.Other]	1.7742	1.821	0.974	0.330	-1.795	5.343
Destination[T.SouthEast Asia]	7.7788	2.108	3.691	0.000	3.648	11.910
DepartureTime[T.6am-12pm]	-0.2383	0.868	-0.275	0.784	-1.939	1.462
DepartureTime[T.6pm-9pm]	-0.0480	0.385	-0.125	0.901	-0.803	0.707
DepartureTime[T.9pm-6am]	-0.1378	0.827	-0.167	0.868	-1.759	1.484
Occupation[T.None]	-0.8068	0.922	-0.875	0.382	-2.615	1.001
Occupation[T.Other]	0.4679	0.647	0.723	0.469	-0.800	1.735
Occupation[T.Professional]	0.4437	0.604	0.735	0.463	-0.740	1.627
Occupation[T.Student]	0.4848	0.827	0.586	0.558	-1.137	2.106
Age	0.1469	0.229	0.641	0.521	-0.302	0.596
FlyingCompanion	-0.6481	0.567	-1.143	0.253	-1.759	0.463
FlyingCompanion:Airfare	0.0017	0.004	0.466	0.641	-0.006	0.009
Airfare	0.5068	0.554	0.915	0.360	-0.578	1.592
Airfare:Destination[T.Japan]	-0.9285	0.666	-1.394	0.163	-2.234	0.377
Airfare:Destination[T.Other]	-0.3503	0.691	-0.507	0.612	-1.705	1.004
Airfare:Destination[T.SouthEast Asia]	-3.5873	1.006	-3.564	0.000	-5.560	-1.615
TripDuration	-0.2006	0.150	-1.337	0.181	-0.495	0.093
Income	-0.0475	0.127	-0.374	0.709	-0.296	0.201
-----						

\* Asiana Airlines is the base outcome

### Foreign Carriers vs Asiana Airlines

**Destination\_SouthEast Asia** (China is base category) – with largest positive magnitude of 7.78 with statically significant p-value, it shows that compared with other destination, the consumer heading to Southeast Asia are more likely to choose Foreign carriers over Asiana Airline compared to other destination.

**Airfare \* Destination\_SoutheastAsia** with negative coefficient, excluding the impact of destination on airfare, we see overall consumers are more likely to choose Asiana if they have to pay higher airfare for

An exploration of airport and airline purchasing decision in Korea

southeast Asia compared to destination of China.

**Airport\_ICN** (Airport\_GMP is base category) –with negative coefficient of -1.6438, customers that departure from Airport\_GMP rather than Airport\_ICN are more likely to prefer Asiana Airlines over Foreign Carriers , given all other predictor variables in the model are held constant.

**Age** –with positive coefficient holding all other variables in the model constant. In other words, the older the customer, the more likely for the consumers to choose the Foreign Carriers.

**FlyingCompanion** – with negative coefficient, holding all other variables constant in the model, the more companions the customers have, the more likely they would choose for Asiana Airlines over Foreign Carriers.

**Airfare** – with positive coefficient, consumers are more likely to choose for Foreign Carriers relative to Asiana Airlines, holding all other variables in the model constant.

**Occupation\_None** (Occupation\_housewife is base category) - with coefficient of -0.8068, the lowest score within the group, it shows that the customers without occupation are most likely to Asiana Airlines relative to Foreign Carriers, given all other predictor variables in the model are held constant.

**DepartureTime\_6am-12pm** (DepartureTime\_12pm-6pm is base category) - with coefficient of -0.2383, the lowest score the group. it shows that the customers plan to departure during this period are most likely to prefer Asiana Airlines relative to Foreign Carriers given all other predictor variables in the model are held constant.

**Income** - negative coefficient shows consumers with higher incomes are more likely to choose Asiana Airlines over foreign airlines.

	Airline=Korean Air	coef	std err	z	P> z	[0.025	0.975]
Intercept		-0.4500	1.087	-0.414	0.679	-2.581	1.681
Airport[T.ICN]		0.1306	0.498	0.262	0.793	-0.845	1.106
Destination[T.Japan]		0.3165	1.159	0.273	0.785	-1.955	2.588
Destination[T.Other]		-3.7920	2.145	-1.768	0.077	-7.996	0.412
Destination[T.SouthEast Asia]		0.8209	1.832	0.448	0.654	-2.770	4.411
DepartureTime[T.6am-12pm]		0.8868	0.746	1.189	0.234	-0.575	2.349
DepartureTime[T.6pm-9pm]		-1.3815	0.397	-3.480	0.001	-2.160	-0.604
DepartureTime[T.9pm-6am]		-2.4258	0.932	-2.603	0.009	-4.253	-0.599
Occupation[T.None]		0.4552	0.853	0.534	0.593	-1.216	2.127
Occupation[T.Other]		0.7741	0.598	1.294	0.196	-0.399	1.947
Occupation[T.Professional]		0.4909	0.571	0.859	0.390	-0.629	1.611
Occupation[T.Student]		1.3920	0.822	1.694	0.090	-0.219	3.003
Age		0.3928	0.227	1.733	0.083	-0.051	0.837
FlyingCompanion		0.3751	0.531	0.707	0.480	-0.665	1.415
FlyingCompanion:Airfare		-0.0013	0.003	-0.371	0.711	-0.008	0.006
Airfare		0.2861	0.574	0.498	0.618	-0.840	1.412
Airfare:Destination[T.Japan]		-0.2269	0.629	-0.361	0.718	-1.459	1.006
Airfare:Destination[T.Other]		1.2971	0.745	1.740	0.082	-0.164	2.758
Airfare:Destination[T.SouthEast Asia]		-0.2737	0.829	-0.330	0.741	-1.899	1.351
TripDuration		-0.2689	0.167	-1.614	0.107	-0.595	0.058
Income		0.2410	0.125	1.928	0.054	-0.004	0.486

### ***Korean Air vs Asiana Airlines***

**Destination\_Other** (China is base category) – with the largest negative magnitude of coefficient of -3.79, compared with other destination, it shows that the consumer heading to other destinations are more likely to choose Asiana Airlines over Korean Air compared to Japan, China and SouthEast Asia.

**DepartureTime\_9pm-6am** (DepartureTime\_12pm-6pm is base category) - with comparatively significant coefficient of -2.42, the lowest score among the group, it shows that the customers plan to departure during this period are most likely to choose Asiana Airlines relative to Korean Air.

**Airfare \* Destination\_Other** with positive coefficient, excluding the impact of destination on airfare, we see overall consumers are more likely to choose Korean LLC over Asiana Airlines if they have to pay more for the destination of other destination.

**Age** – positive coefficient shows that the older the more likelihood for consumers to Korean LLC over Asiana.

**Airfare** – with positive coefficient, controlling for other variables in the model, the higher the airfare, the more likely the individuals would choose Korean Air relative to Asiana Airlines

**FlyingCompanion** – with positive coefficient, controlling for other variables in the model, individuals with FlyingCompanion were more likely to choose Korean Air relative to Asiana Airlines

**Airport\_ICN** (Airport\_GMP is base category) – With positive sign of coefficient, consumers are more likely to choose r Korean Air over Asiana Airlines while departure from Airport\_ICN.

**Occupation\_Student** (Occupation\_housewife is base category) - with highest positive coefficient among the group compared with housewife, it shows that students are more likely to choose Korean Air over Asiana compared to housewife.

**Income** - positive coefficient shows consumers with higher incomes are more likely to choose Korean Air over Asiana Airlines.

	Airline=Korean LLC	coef	std err	z	P> z	[0.025	0.975]
Intercept		-30.9402	5.88e+04	-0.001	1.000	-1.15e+05	1.15e+05
Airport[T.ICN]		-2.5320	1.284	-1.972	0.049	-5.049	-0.015
Destination[T.Japan]		28.5085	8.42e+04	0.000	1.000	-1.65e+05	1.65e+05
Destination[T.Other]		-1.6633	1.41e+05	-1.18e-05	1.000	-2.77e+05	2.77e+05
Destination[T.SouthEast Asia]		26.4334	8.42e+04	0.000	1.000	-1.65e+05	1.65e+05
DepartureTime[T.6am-12pm]		0.2524	1.274	0.198	0.843	-2.245	2.750
DepartureTime[T.6pm-9pm]		-0.5192	0.668	-0.777	0.437	-1.829	0.791
DepartureTime[T.9pm-6am]		0.5562	1.071	0.519	0.604	-1.544	2.656
Occupation[T.None]		-21.0522	1374.370	-0.015	0.988	-2714.768	2672.664
Occupation[T.Other]		1.7671	1.270	1.391	0.164	-0.723	4.257
Occupation[T.Professional]		1.6370	1.240	1.320	0.187	-0.794	4.068
Occupation[T.Student]		0.4922	1.706	0.288	0.773	-2.852	3.837
Age		-0.8304	0.419	-1.980	0.048	-1.652	-0.009
FlyingCompanion		-1.3966	1.449	-0.964	0.335	-4.236	1.443
FlyingCompanion:Airfare		0.0106	0.009	1.152	0.250	-0.007	0.029
Airfare		-4.9147	7.52e+04	-6.54e-05	1.000	-1.47e+05	1.47e+05
Airfare:Destination[T.Japan]		-6.8692	8e+04	-8.59e-05	1.000	-1.57e+05	1.57e+05
Airfare:Destination[T.Other]		1.3980	1.65e+05	8.45e-06	1.000	-3.24e+05	3.24e+05
Airfare:Destination[T.SouthEast Asia]		1.2360	7.84e+04	1.58e-05	1.000	-1.54e+05	1.54e+05
TripDuration		-0.6560	0.868	-0.755	0.450	-2.358	1.046
Income		-0.1500	0.220	-0.682	0.495	-0.581	0.281

***Korean LLC vs Asiana Airlines***

**Airport\_ICN** (Airport\_GMP is base category) – With negative sign of coefficient, controlling for other variables in the model, consumers are more likely to choose Asiana Airlines over Korean LLC while departure from Airport\_ICN.

**Destination** (China is base category) – with largest positive magnitude of 28.50 and 26.43 compared with Destination\_China, it shows that the consumer heading to Japan and SouthEast Asia are most likely to choose Korean LLC over Asiana Airlines than compared to other destination.

**Airfare** – with negative coefficient, controlling for other variables in the model, the higher the airfare, the more likely the individuals would choose Asiana Airlines relative to Korean LLC

**Airfare \* Destination\_Japan** excluding the impact of destination on airfare, we see overall consumers are less likely to choose Korean LLC over Asiana if they have to pay more for destination of Japan.

**Age** – with negative coefficient, it shows that the younger the more likelihood for consumers to choose Asiana Airlines over Korean LLC.

**FlyingCompanion** – with negative coefficient, controlling for other variables in the model, individuals with more FlyingCompanion were more likely to choose o Asiana Airlines over Korean LLC.

**TripDuration-** With negative sign of coefficient, controlling for other variables in the model, consumers are more likely to choose Asiana Airlines over Korean LLC for longer trip duration, while holding all other variables in the model constant.

**DepartureTime\_6pm-9pm** (DepartureTime\_12pm-6pm is base category) - with lowest coefficient of - 0.5129 within the category, controlling for other variables in the model, customers departure during this period are most likely to prefer Asiana Airlines relative to Korean Air compared other departure time.

**Occupation\_None** (Occupation\_housewife is base category) - with largest negative magnitude compared with the housewife, controlling for other variables in the model, individuals without occupation are more likely to choose Asiana Airlines over Korean LLC compared to other occupations.

**Income** - with negative coefficient , controlling for other variables in the model, consumers with lower income level are more likely to choose Asiana Airlines over Korean Air.

***Table: Airline Multinomial Logistic Model Confusion Matrix***

		Predicted			
Actual		Asiana Airlines	Foreign Carriers	Korean Air	Korean LLC
	Asiana Airline	5	11	15	2
	Foeign Carriers	4	20	13	8
	Korean Air	6	6	27	2
	Korean LLC	1	0	4	20

According to the size of magnificent and correspondent p-value of coefficients across three comparison models, we found out that the destination, airfare, airport, age, departure time and occupation affects more for the consumers' choice, we also found that our logit model for airline was much better at predicting Korean Air and Korean LLC with accuracy of 65.85% and 80% respectively.

### 5.3. Decision Tree Model

Our decision tree model yielded 0.5486 on variables selected by forward stepwise. This result indicates that when consumers choose airlines, they consider these variables as the most important factors. According to Table 3 and Table 4, among all the airlines, Decision Tree gives the best prediction results on Korean LLC.

**Table: Airline Decision Tree Model Confusion Matrix**

		Predicted			
Actual		Asiana Airlines	Foreign Carriers	Korean Air	Korean LLC
	Asiana Airline	18	7	8	0
	Foreign Carriers	7	21	15	2
	Korean Air	13	3	23	2
	Korean LLC	1	4	3	17

**Table: Airline Decision Tree Model Precision and Recall**

	Asiana Airlines	Foreign Carriers	Korean Air	Korean LLC
Precision	0.46	0.60	0.47	0.81
Recall	0.55	0.54	0.56	0.68

In order to analyze consumers' airline choice behavior, we traced along paths from the root to leaf nodes to see how each classification had been made. As shown in our tree model figures, each color represents one airline with most people chosen at a certain node. The darker the color, the higher the purity. In this case, orange represents Asiana Airlines, green represents Foreign Carriers, blue is Korean Air and purple is Korean LLC. Passengers who choose Korean LLC are more likely to go to Southeast Asia and the tickets are relatively cheaper. For passengers who choose Foreign Carriers, Southeast Asia is not their destination but they fly to China or Japan frequently. Their flights' departure time are usually between 6pm to 9pm and their airfares are relatively expensive. One third of passengers in this training dataset choose to fly with Korean Air. Majority of them are not going to Southeast Asia. Passengers are more likely to be those whose departure time is not between 6pm to 9pm. The trip duration is less than a week and the airfare is more expensive than 24.5. However, the remaining passengers who go to Southeast Asia spend more on airfare.

To find out the best subset of variables that produce higher test accuracy, we used the Decision Tree classifier implementation in scikit library to compute the importance score for each feature. The higher the feature importance, the more important the feature is. The importance of a feature is computed



An exploration of airport and airline purchasing decision in Korea

as the total reduction of the criterion brought by that feature which is the Gini importance. The table below shows that features selected from forward stepwise and the importance scores reported by Decision Tree.

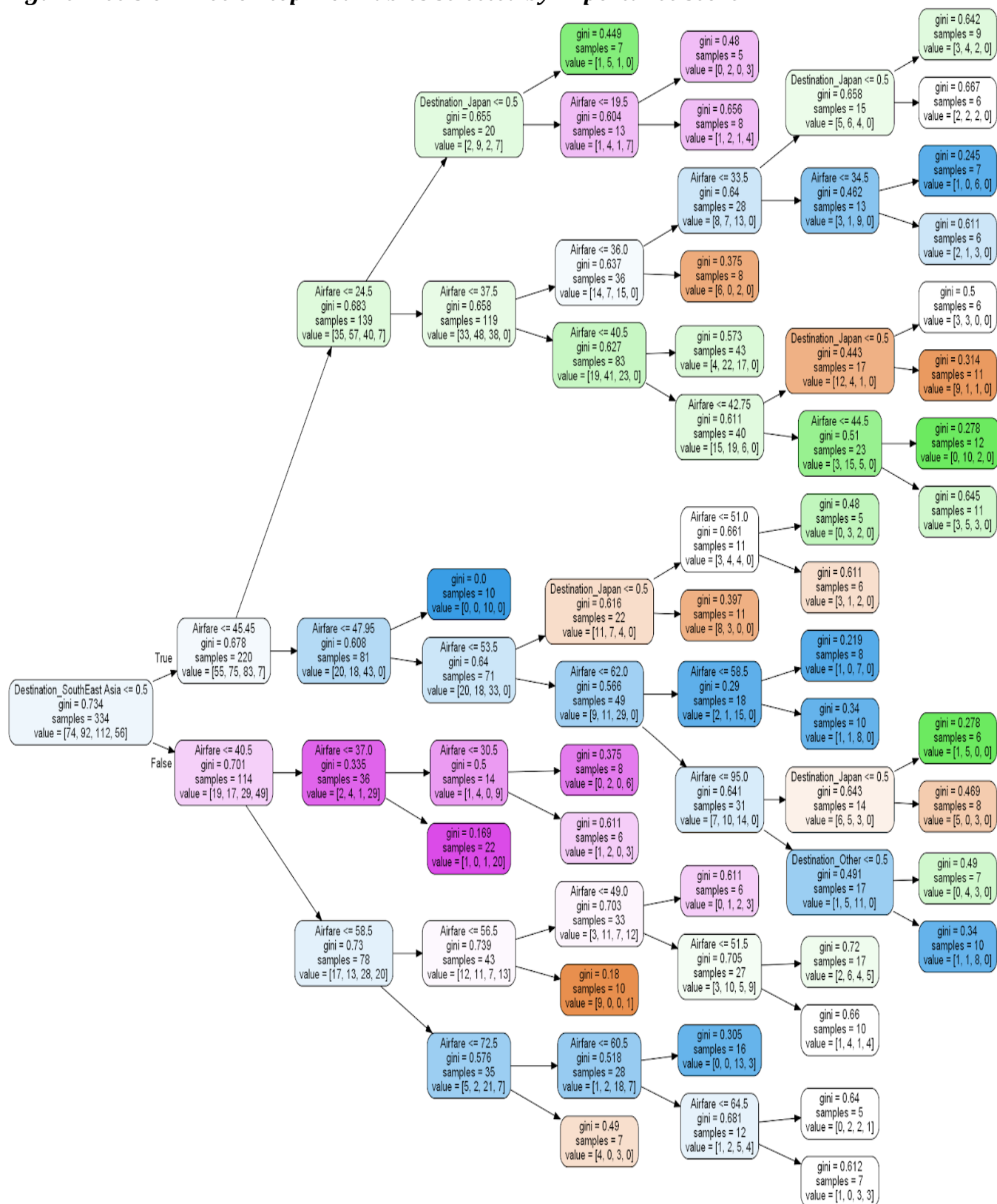
We then used the top-2 and top-3 features to retrain the Decision Tree classifier. We observed that the test accuracy for the Decision Tree models increased to 0.6042 when using only the top-2 features: Airfare and Destination.

***Table: Features with importance scores reported by Decision Tree***

Features	Importance Score
<b>Airfare</b>	<b>0.3941</b>
FlyingCompanion	0.0606
NoTransport	0
<b>TripDuration</b>	<b>0.1029</b>
NoTripsLastYear	0.0683
Destination_Japan	0.0496
Destination_Other	0
<b>Destination_SouthEast Asia</b>	<b>0.1268</b>
DepartureTime_6am-12pm	0
DepartureTime_6pm-9pm	0.0689
DepartureTime_9pm-6am	0
Airport_ICN	0.0227
FrequentFlightDestination_Japan	0.0589
FrequentFightDestination_Other	0.0360
FrequentFlightDestination_Southeast Asia	0.0113



**Figure: Decision Tree on top-2 variables selected by importance score**



## 6. CONCLUSION

In our analysis, we found that destination, airfare, and departure time most heavily impact airport choice. We found that airfare and destination most heavily impact airline choice.

We found that individuals flying within Southeast Asia beyond China and Japan were much more likely to choose ICN. This is intuitive, given that ICN has more long-haul flights than does the regional hub, GMP. Additionally, of individuals flying outside of China and Japan in SE Asia, those with a departure time in the morning were much more likely to fly ICN than GMP. For those who flew to Japan and China, airfare was a large determining factor in their airport selection.

For airline choice, we found that individuals who are price sensitive and flying to Southeast Asia are more likely to choose Korean LLC. Among those who are going to Southeast Asia, individuals who would like to have a departure time between 6pm to 9pm are more likely to take Foreign Carriers. Others fly with Korean Air more often.

In order to improve market share, carriers could consider expanding their flight destinations to new countries/cities in Southeast Asia. However, this may not be economical given competition with low-cost airlines, as airfare is also a key factor in airline selection. Of note, only a few people going to Japan or China choose a Korean LLC; this is likely due to lack of route for these low-cost carriers. If low-cost carriers were to add routes to these locations, they may be able to grow revenues, especially considering China is a fast growing air travel market. What's more customers who departure from 9pm-6am are less likely to choose Korean LLC, company may arrange routes for other departure time to increase seat occupancy rate and the profit.