Airport & Airline Modeling Project

Ty Hak, Yabo Duan, Rusteen Farivar, Grant Napolitano

3/10/2024

Introduction

This study aims to analyze air passenger flow at Gimpo International Airport and Incheon International Airport in the Seoul metropolitan area, Korea. Two airports were opened in different years and offer multiple terminals with a range of facilities including shops and transportation services. Leon (2012) highlights that passengers are dealing with an overwhelming number of options and attributes when deciding on airport selection. According to Hess and Polak (2006), flight frequency, access cost, and flight time also remain significant factors, emphasizing the importance of practical considerations in the decision-making process. By constructing the Logit and decision tree models, we would expect the influence on airport and airline choice intentions of air passengers from perspectives of passenger demographic characteristics, travel information, ground transportation, etc. Both airports play crucial roles in boosting Seoul with regional and global connectivity and economic development. This research can hold several potential implications in planning and developing transportation infrastructure such as expansion, and public transit connectivity as well as in promoting and assessing tourism or marketing strategies and the economic impact of airports.

Objective and Scope

Our main objective is to determine the reasons why an individual may choose one airport or airline over another. These reasons may be related to internal motivating factors, like the purpose of the trip. The reasons may also be more external logistic factors, such as the number of transportation methods required, the time required to access the airport, or even the time the flight takes place. The goal would be to provide explainable criteria that can inform any relevant stakeholders, while also identifying any possible executive decisions along the way. Our dependent variable of choice for the airport models will be if the airport Gimpo is selected (1 or 0). If the value is 0, that simply means the airport Incheon was selected instead. Our dependent variable of choice for the airline models will be if the airline KE is selected (1 or 0). If the value is 0, that means another airline was chosen. Our dataset was collected from a survey of 488 respondents. This survey covers a wide array of responses related to airport/airline choice, various demographic details, access to the airport, time of departure, airfare, and others.

Data

Data Cleaning and Organization

As mentioned prior, the data was collected by a survey of 488 respondents across a total of 27 attributes, which include airline and flight information, passenger demographic/socio-demographic characteristics, travel purpose, ground transportation, etc. 14 out of 27 attributes contain missing values, with Mileage and MileageAirline both sharing this issue the most. In addition, for DepartureHr and DepartureMn columns, we would use DepartureTime to accurately capture this relationship more effectively, which has no missing values while also being able to measure temporal effects on our dependent variables. Similarly, we could drop AccessCost as well, since we have NoTransport, ModeTransport as references for ground transportation.

Generally, we filled missing values by modes of columns or modes corresponding to reference columns since the rest columns are categorical or discrete data. For the Age and Gender columns, only one and three values were missing in these two columns. As one of the dependent variables, there are 10 missing values in the Airline column, so we tried FlightNo as a reference, but the result shows all rows missing in Airline are missing in FlightNo. Therefore, we tried to fill 6 of the missing values by the mode corresponding to Destination and TripDuration and fill the remaining by the mode of Airline. Similarly, we filled 4 missing values by the mode of SeatClass corresponding to Airfare. For AccessTime, we referred to modes corresponding to ModeTransport, NoTransport, and ProvinceResidence and filled the remaining by median of AccessTime. Furthermore, we dropped Income and Airfare since we have SeatClass as one critical variable.

Feature Selection

During the exploratory data analysis, important steps were taken to determine potentially impactful features for the model, and this included creating a series of tables that helped to capture the relationships between the various categories and our dependent variables. The first table plotted out how many people selected each specific category. The second table covered how many people selected each specific category, also choosing Gimpo. The third and most notable table finds the percentage of how many people who chose a specific category plus Gimpo, out of all of the people who chose that specific category. A part of the results (Figure 1.1) is shown below.

Section of Proportion Table - Category X Airport[Gimpo] (Figure 1.1)

Category	Nationality	TripPurpose	Destination	DepartureTime
1	45.62%	44.27%	66.17%	74.47%
2	78.79%	65.09%	83.87%	64.62%
3	97.62%	62.96%	14.29%	37.88%
4	9.52%	61.29%	20.83%	3.33%
5	50.00%			

For example, 97.62% of those who chose nationality category #3 (Japanese) chose Gimpo airport. TripPurpose is included to show what a less interesting set of percentages looks like, which is helped by including the color-based conditional formatting. Destination also holds particular interest, with its category #2 being Japan. But by understanding the similarity between nationality and destination, we can already assess a high chance of multicollinearity between the two variables just by assessing their proportions. Departure time is also shown to indicate some sort of relationship and will be a feature worth exploring further in the models as well. For the Category X Airline [KE] percentages shown below in (Figure 1.2), it was found that TripPurpose and GroupTravel showed signs of higher representation. GroupTravel may be expressed in its binary form, yes or no, but may also be further explored in its continuous form, given to us as FlyingCompanion (the number of companions traveled with). What's interesting to note is in the "occupation" category, 18.49% of those who identified as a corporate worker had chosen KE as their airline. While the executive value of this number cannot currently be determined, it may serve as a valuable supporting variable for any models going forward. Following this logic, certain selections in ProvinceResidence may also be valuable, but it is important to note that categories with a small number of responses may not be intuitive to include in a general predictive model.

Section of Proportion Table - Category X Airline[KE] (Figure 1.2)

Category	GroupTravel	Occupation	TripPurpose	ProvinceResidence
1	44.71%	40.54%	29.41%	32.43%
2	29.60%	18.49%	37.74%	9.09%
3		30.00%	29.63%	42.97%
4		50.00%	45.16%	7.14%
5		31.03%		45.16%
6		75.00%		50.00%
7		38.89%		16.67%
8		37.25%		20.88%
9		39.29%		
10		50.00%		
11		50.00%		
12		32.54%		

Descriptive Statistics Table (Figure 1.3)

	Mean	Standard Dev.	Minimum	Maximum	25%	50% (Median)	75%
ID	244.500	141.018	1	488	122.75	244.5	366.25
Airport	1.510	0.500	1	2	1	2	2
Airline	2.416	1.208	1	4	1	2	4
Age	39.945	13.666	17	80	29	37.5	50
Gender	1.531	0.500	1	2	1	2	2
Nationality	1.484	1.013	1	5	1	1	1
TripPurpose	1.518	0.862	1	4	1	1	2
TripDuration	27.441	74.988	0	730	4	5	8
FlyingCompanion	2.820	4.002	0	34	1	2	3
ProvinceResidence	3.391	2.599	1	8	1	3	5
GroupTravel	1.826	0.380	1	2	2	2	2
NoTripsLastYear	3.262	8.997	0	122	1	2	3
Destination	2.184	0.892	1	4	1	2	3
DepartureTime	2.432	0.750	1	4	2	2	3
SeatClass	1.121	0.395	1	3	1	1	1
Airfare	50.457	23.929	3	260	40	50.457	50.457
NoTransport	1.334	0.552	1	4	1	1	2
ModeTransport	3.801	2.380	1	11	2	3	6
AccessTime	50.822	40.553	4	390	30	40	60
Occupation	6.768	4.172	1	12	2	8	12
Income	2.822	1.434	1	7	2	2	4

Models and Results

Logistic Regression Models

Logit Regression Model 1 - Airport [Gimpo] (Figure 2.1)

```
==============================================================================
Dep. Variable:         Airport[Gimpo]   No. Observations:                  488
Model:                          Logit   Df Residuals:                      481
Method:                           MLE   Df Model:                            6
Date:                Wed, 06 Mar 2024   Pseudo R-squ.:                 0.4189
Time:                        23:12:40   Log-Likelihood:                -196.50
converged:                       True   LL-Null:                       -338.15
Covariance Type:            nonrobust   LLR p-value:                3.068e-58
================================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
FlyingCompanion                0.0626      0.035      1.779      0.075      -0.006       0.132
Destination[Japan]             2.9913      0.381      7.853      0.000       2.245       3.738
Destination[China]             1.8064      0.364      4.965      0.000       1.093       2.519
Destination[SEAsia]           -1.1575      0.434     -2.665      0.008      -2.009      -0.306
NoTransport                   -1.0929      0.232     -4.720      0.000      -1.547      -0.639
DepartureTime[6am-12pm]        2.7957      0.486      5.747      0.000       1.842       3.749
DepartureTime[9pm-12am]       -3.5352      1.120     -3.157      0.002      -5.730      -1.341
================================================================================================
```

In logistic regression, the estimated coefficients are measured in terms of log odds (logarithm of odds). The estimated coefficients represent the change in the log odds of the event for a one-unit change in the corresponding predictor variable, holding all other variables constant. To convert log odds to odds, use the inverse of the natural logarithm which is the exponential function e^(estimate). To find the probability, we divide the odds ratio by (1 + the odds ratio).

Odds and Probability Table - Airport [Gimpo] (Figure 2.2)

Coefficient	Log(Odds)	Odds Ratio	Probability	P>|z|
FlyingCompanion	0.0626	1.065	51.56%	0.075
Destination[Japan]	2.9913	19.912	95.22%	0.000
Destination[China]	1.8064	6.088	85.89%	0.000
Destination[SEAsia]	-1.1575	0.314	23.91%	0.000
NoTransport	-1.0929	0.335	25.11%	0.000
DepartureTime[6am-12pm]	2.7957	16.374	94.24%	0.000
DepartureTime[9pm-12am]	-3.5352	0.029	2.83%	0.002

With this table, we now can intuitively explain our coefficients.

- Flying Companion: Controlling for all other variables, the probability of choosing Gimpo Airport over Incheon increases by approximately 6.26% for each additional flying companion.
- Destination[Japan]: Controlling for all other variables, the probability of choosing Gimpo Airport over Incheon is approximately 95.22%.
- Destination[China]: Controlling for all other variables, the probability of choosing Gimpo Airport over Incheon is approximately 85.89%.
- Destination[SEAsia]: Controlling for all other variables, the probability of choosing Gimpo Airport over Incheon is approximately 23.91%.
- NoTransport: Controlling for all other variables, the probability of choosing Gimpo Airport over Incheon is approximately 25.11%.
- DepartureTime[6am-12pm]: Controlling for all other variables, the probability of choosing Gimpo Airport over Incheon is approximately 94.24%.
- DepartureTime[9pm-12am]: Controlling for all other variables, the probability of choosing Gimpo Airport over Incheon is approximately 2.83%.

Regarding statistical significance:

- Destination[Japan], Destination[China], Destination[SEAsia], NoTransport, and DepartureTime[6am-12pm] are all statistically significant at the .001 level.
- DepartureTime[9pm-12am] is statistically significant at the .01 level.
- FlyingCompanion is statistically significant at the .1 level.

Logit Regression Model 2 - Airline [KE] (Figure 2.3)

Dep. Variable:	Airline[KE]	No. Observations:	488
Model:	Logit	Df Residuals:	481
Method:	MLE	Df Model:	6
Date:	Thu, 07 Mar 2024	Pseudo R-squ.:	0.05279
Time:	13:51:46	Log-Likelihood:	-291.06
converged:	True	LL-Null:	-307.28
Covariance Type:	nonrobust	LLR p-value:	1.340e-05

	coef	std err	z	P>|z|	[0.025	0.975]
TripPurpose[Leisure]	-1.3320	0.226	-5.896	0.000	-1.775	-0.889
TripPurpose[Study]	-1.2158	0.444	-2.736	0.006	-2.087	-0.345
TripPurpose[Business]	-0.8524	0.246	-3.472	0.001	-1.334	-0.371
GroupTravel[Yes]	0.8552	0.263	3.248	0.001	0.339	1.371
Destination[SEAsia]	-0.5314	0.223	-2.380	0.017	-0.969	-0.094
Province[Seoul]	0.4745	0.235	2.021	0.043	0.014	0.935
Province[Kyungki-do]	0.9035	0.253	3.565	0.000	0.407	1.400

Here is the second logistic model that covers whether or not someone will choose the KE airline over another airline.

Odds and Probability Table 2 - Airline[KE] (Figure 2.4)

Coefficient	Log(Odds)	Odds Ratio	Probability	P>|z|
TripPurpose[Leisure]	-1.332	0.264	20.88%	0.000
TripPurpose[Study]	-1.2158	0.296	22.87%	0.006
TripPurpose[Business]	-0.8524	0.426	29.89%	0.001
GroupTravel[Yes]	0.8552	2.352	70.17%	0.001
Destination[SEAsia]	-0.5314	0.588	37.02%	0.017
Province[Seoul]	0.4745	1.607	61.64%	0.043
Province[Kyungki-do]	0.9035	2.468	71.17%	0.000

Now, we can repeat our process for interpretation like how we did in the first model (Figure 2.2).

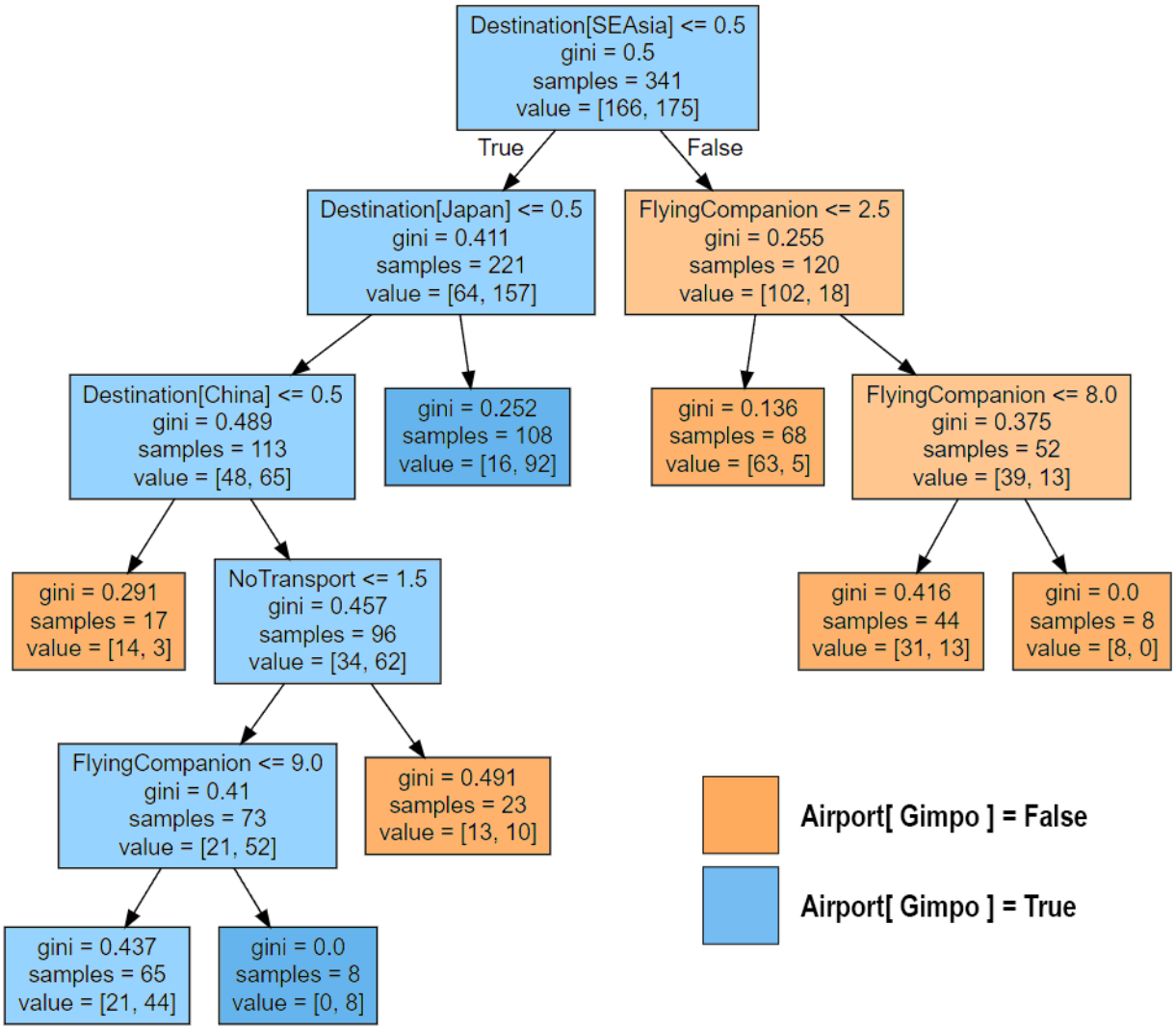
- TripPurpose[Leisure]: Controlling for all other variables, the probability of choosing KE airline over others is approximately 20.88% lower for leisure trips compared to other trip purposes.
- TripPurpose[Study]: Controlling for all other variables, the probability of choosing KE airline over others is approximately 22.8% lower for study-related trips compared to other trip purposes.
- TripPurpose[Business]: Controlling for all other variables, the probability of choosing KE airline rport over others is approximately 29.89% lower for business-related trips compared to other trip purposes.
- GroupTravel[Yes]: Controlling for all other variables, the probability of choosing KE airline over others is approximately 70.17% higher for passengers traveling in groups compared to those traveling alone.
- Destination[SEAsia]: Controlling for all other variables, the probability of choosing KE airline over others is approximately 37.02% lower if the destination is in Southeast Asia compared to other destinations.
- Province[Seoul]: Controlling for all other variables, the probability of choosing KE airline over others is approximately 61.64% higher if the passenger's province is Seoul compared to other provinces.
- Province[Kyungki-do]: Controlling for all other variables, the probability of choosing KE airline over others is approximately 71.17% higher if the passenger's province is Kyungki-do compared to other provinces.

Regarding statistical significance:

- Province[Kyungki-do], and TripPurpose[Leisure] are statistically significant at the .001 level.
- TripPurpose[Study], TripPurpose[Business], and GroupTravel[Yes] are statistically significant at the .01 level.
- Destination[SEAsia] and Province[Seoul] are statistically significant at the .05 level.

Decision Tree Models

Decision Tree 1 - Airport [Gimpo] (Figure 2.5)



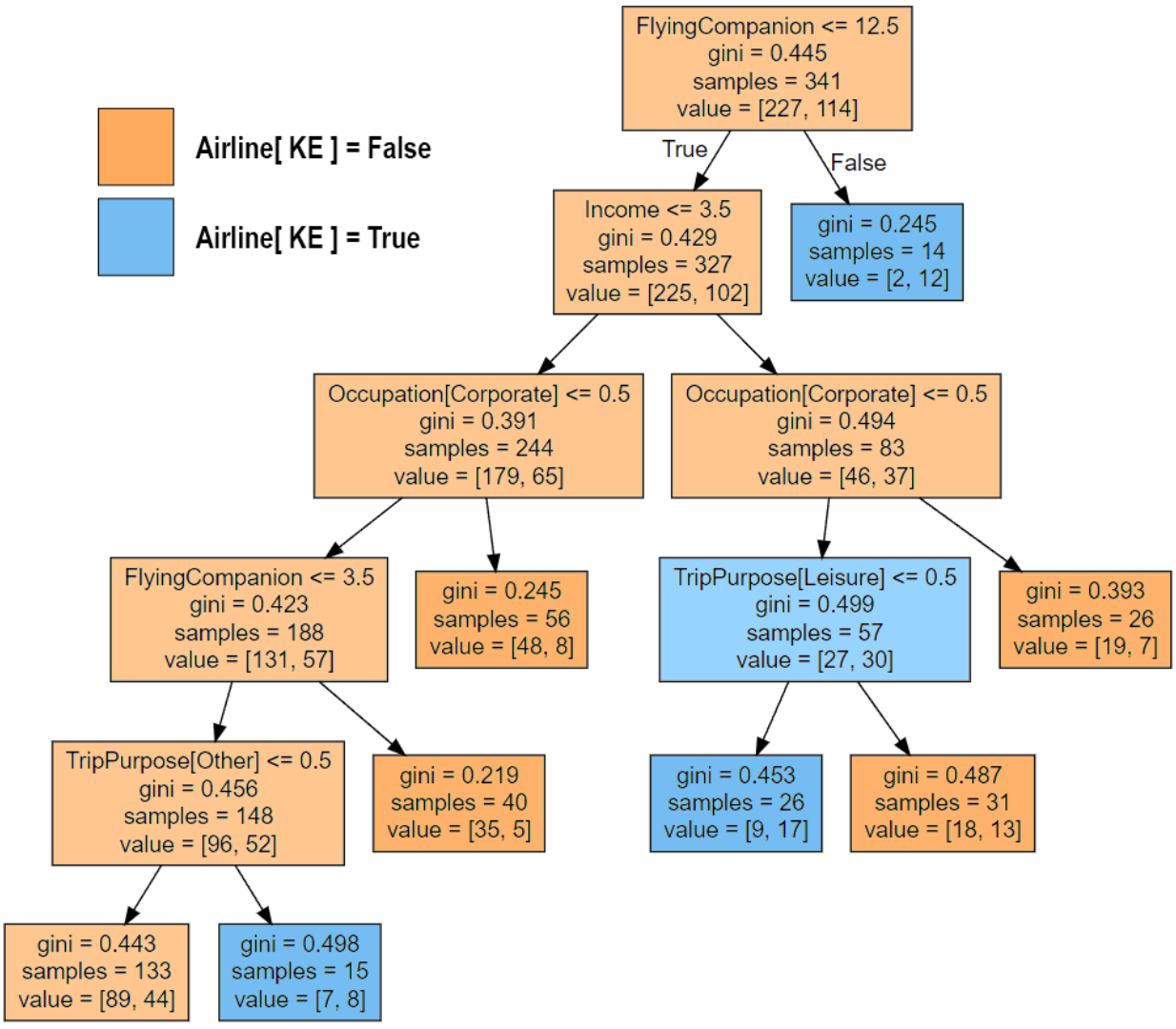
Compared to the logit regression model, the decision tree helps to visualize the most apparent relationships. In particular, if the destination is Southeast Asia, no classifications are predicting that the individual would choose Gimpo. If the destination is Japan, the predicted airport is Gimpo, and the gini of the terminal node is solid at .252. If the destination is China, the tree model predicts that Gimpo would not be the airport of choice. Lastly, the number of transportation methods required to reach Gimpo has a negative predicted effect, tying back to their established negative relationship in our logit model as well. The performance of the model, shown below, illustrates a satisfactory accuracy given the low depth of the model and overall provided sample size.

Decision Tree 1 - Performance Metrics (Figure 2.6)

Accuracy	78.9%
Precision	78.7%
Recall	79.7%
Specificity	78.1%
FP Rate	21.9%
FN Rate	20.3%

Following suit is the second decision tree for whether or not an individual chooses the airline KE.

Decision Tree 2 - Airline [KE] (Figure 2.7)



This decision tree was created to estimate whether an individual would choose the airline KE, and this airline in particular was chosen due to it providing the largest subset of observations. This is important when presented with the constraint of a small dataset, such as the one used throughout this project. A greater level of depth was provided for the input of the model to improve accuracy as much as possible, but the tree's explainability was properly retained. To start, the number of people flying with an individual had an impact on whether KE was chosen, and the airline being chosen if the number was 13 or higher. Notably, occupation also has an impact on an individual's choice. If they were a corporate worker, it is predicted that KE would not be chosen, and if they were not a corporate worker, KE would be chosen. This provided an interesting split to identify airline choices based on occupation. To further support this point, if the trip's purpose was for leisure, it was predicted that KE would be chosen as well. Overall, the performance metrics for this decision tree still leave more to be desired. This may come down to multiple factors, but given the constraints of the dataset and the even smaller subset based on each potential airline choice, it is understandable that the tree's performance would be some degree lower than the tree for airport choice (Figure 2.5).

Decision Tree 2 - Performance Metrics (Figure 2.8)

Accuracy	73.5%
Precision	61.9%
Recall	29.5%
Specificity	92.2%
FP Rate	7.8%
FN Rate	70.5%

Conclusions and Recommendations

From a policy perspective, we have three suggestions for the city of Seoul to implement based on our research. The first suggestion is from our primary logit model, specifically the NoTransport coefficient which implies that with an increase in transportation methods needed to reach an airport, a traveler is 25% less likely to choose Gimpo airport compared to Incheon. This is surprising since Gimpo Airport is closer to the city of Seoul. This research finding indicates that the transportation infrastructure for Gimpo Airport may need to be re-evaluated and expanded upon. We suggest creating a direct transit line from parts of Seoul to Gimpo airport such as subway or buses to minimize transfers, as well as clear information on the optimal path to Gimpo airport. In addition, we recommend improving the time/cost savings of transportation to Gimpo by expanding park-and-ride facilities and shuttles to/from hotels to create more direct routes to the airport for citizens and travelers. From a marketing perspective, the city should consider promoting the proximity advantage of Gimpo airport to local citizens and airlines could even offer discounted fairs for flights out of Gimpo to make its use more compelling. The key goal here is to make Gimpo's proximity advantage more enticing to reduce passenger transportation time from traveling to a further location, Incheon.

Our second suggestion also comes from our first logit model (Figure 2.1) which indicates the correlation between Gimpo Airport and flights to Japan and China. Passenger flights to Japan and China are 95% and 85% more likely to go through Gimpo airport than Incheon. Given the high demand, we suggest Gimpo work with airlines to add more flights from Gimpo to major Japanese and Chinese cities, and to leverage Gimpo's proximity for short regional routes. Given the large-scale demand for these locations, we also recommend promoting Gimpo airport as the regional hub for travel to North Asia from Seoul and Korea in general and ensuring gates and operational capacity to Japan and China are met.

Lastly, our research has uncovered that passengers with long-haul flights to Southeast Asia are 37% less likely to go through Gimpo Airport. We suspect this phenomenon is due to the latest facility upgrades and the sheer size of Incheon in comparison to Gimpo. Because of this, we suggest Gimpo transfer some existing Southeast Asia flights to Incheon gradually while increasing capacity and flights for closer North Asian flights. Eventually, Gimpo may phase out long-haul flights and provide smooth transfers to Inchoen for connecting Southeast Asian flights as Inchoen is better equipped. Additionally, Incheon should market itself as the premier Southeast Asia hub, improving amenities and increasing appeal for long-haul passengers with more robust lounges, food options, and customer service.

Final Words

To conclude in full, this research project has provided an invaluable perspective on the various factors that weigh into people's decision-making when it comes to airport/airline selection. With a larger dataset and more applications into less interpretable but more powerful models such as neural networks, the possibilities are endless, and it is something to look forward to in the future given more volume to work with. Moving forward, there's an exciting horizon of opportunities awaiting exploration, promising even deeper insights into consumer behavior and preferences in the aviation industry.

Works Cited

Leon, S. (2011). Airport Choice Modeling: Empirical Evidence from a Non-Hub Airport. *Journal of the Transportation Research Forum*, 5-16.

Polak, J. W., & Hess, S. (2006). Exploring the potential for cross-nesting structures in airport-choice analysis: A case-study of the Greater London area. *Transportation Research Part E* , 63-81.