Passengers' airport and airline choices in Seoul Metropolitan Area: insights from survey and quantitative methods

Project from BUAN 5310 Machine Learning

Nick Helm, Grace Jung, Kelly Tsang

Introduction

In the Seoul Metropolitan area, two international airports, Incheon Airport and Gimpo Airport, serve as key transportation hubs. This report examines passengers' choices regarding airports and airlines, along with the influential factors identified from survey data gathered at these locations. Subsequently, we will explore policy implications derived from the quantitative analysis.

Exploratory Data Analysis

Descriptive Statistics and Data Visualization

Descriptive Statistics on Continuous Variables

N =488	Mean	Std Dev	Min	Max	N	#of Missing Value
Age	39.96	13.67	17	80	487	1
Trip Duration	27.44	74.99	0	730	488	0
Flying Companion	2.82	4	0	34	488	0
No trips Last Year	3.26	8.99	0	122	488	0
Departure Hr	15.98	4.01	1	25	454	0
Departure Mn	25.98	15.75	0	55	368	120
Airfare	50.46	28.98	3	260	333	155
No Transport	1.33	0.55	1	4	488	0
Access Cost	11,220.08	24,083.03	0	350,000	291	197
Access Time	51.83	43.49	4	390	391	97
Mileage	56,383.7	89,411.82	1	500,000	90	398

The descriptive statistics for continuous variables in the study of customer behavior concerning airport and airline choice offer a detailed look into the dataset's quantitative aspects. The sample size is 488.

Trip-related features

The average trip duration is about 27 days, and respondents typically fly with nearly 3 companions, suggesting a propensity for group travel. Departure hours mostly peak around 3 pm to 4 pm. The mean airfare paid by respondents is around KRW500,000, yet the access cost to the airport shows a significant variance, with an average cost of KRW11,220, and a notably high standard deviation, signaling a broad spectrum of expenses. Access time to the airport averages 52 minutes, with a considerable standard deviation, indicating that the convenience of reaching the airport could be a substantial factor in airport selection.

Passenger-related features

Respondents’ ages range from 17 to 80 years, with a mean age of approximately 40. Within the last year, respondents have taken an average of 3 trips. The standard deviation is substantial for mileage points, suggesting varying degrees of customer loyalty and distances traveled.

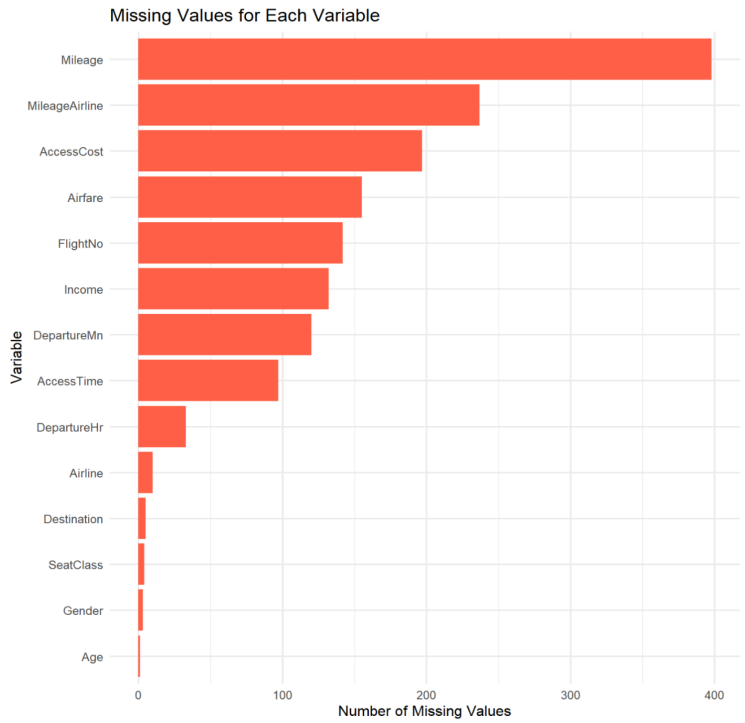
Descriptive Statistics on Nominal Variables

N=488	Mode	N	# of Missing Values
ID	1	488	0
Airport	2	488	0
Airline	1	478	10
Gender	2	478	10
Nationality	1	485	3
Trip Purpose	1	488	0
Province Residence	1	488	0
Group Travel	2	488	0
Frequent Flight Destination	1	488	0
Destination	3	483	5
Flight No	KE2815	346	142
Departure Time	2	488	0
Seat Class	1	484	4
Mode Transport	1	488	0
Occupation	12	488	0
Income	2	488	0
Mileage Airline	4	251	237

In terms of nominal variables, the data depict clear patterns in customer preferences. With 488 observations, modes indicate favored options in airports and airlines, with most respondents selecting a particular class, mode of transport, and occupation among other categorical factors.

Missing data is manageable, with the lowest being 10 missing values for airline choice, and the highest being 237 in mileage of frequent flyer program. The frequency of certain flights and the chosen departure times show consistency, further highlighting prevalent trends in customer choices. Notably, income levels show the mode at the 2nd category (30 to 50 million KRW), and the mileage airline points for airline loyalty programs exhibit a mode of 4, meaning no frequent flyer program membership, out of a subsample of 251.

These figures suggest there is a level of consistency in customer choices, albeit with some data missing, primarily in the mileage airline points category. Overall, the nominal data provide a snapshot of the passenger preferences that prevail within the dataset, which could be pivotal for understanding the decision-making process behind airport and airline selection.



To enhance our analytical process, we employed a visual representation of missing values within our dataset, facilitating a more intuitive understanding of variable completeness. This approach aids in determining the inclusion or exclusion of variables in our analysis, based on the prevalence of missing data.

Variables characterized by substantial missing values warrant meticulous examination to devise appropriate strategies for addressing these gaps. Such a methodical evaluation is crucial for ensuring data integrity and optimizing the reliability of our subsequent analyses. Missing values will be addressed later in the report.

	Age	TripDuration	Airfare	AccessCost	AccessTime
Age	1.000	-0.030	0.040	0.047	0.052
TripDuration	-0.030	1.000	0.047	0.112	-0.046
Airfare	0.040	0.047	1.000	0.085	0.049
AccessCost	0.047	0.112	0.085	1.000	0.253
AccessTime	0.052	-0.046	0.049	0.253	1.000

The correlation matrix serves as a pivotal tool for discerning statistical significance among variables, thereby facilitating a more informed selection of predictors for model evaluation. Variables exhibiting higher correlation coefficients merit closer scrutiny and are considered prime candidates for inclusion in subsequent modeling efforts.

Age is positively correlated with airfare, access cost, and access time to the airport, potentially reflecting increasing purchasing power as people age. Conversely, age exhibits a negative correlation with trip duration, implying shorter trips as individuals grow older. However, the correlation between age and other variables is relatively weak, all below 0.1 or -0.1.

Trip duration demonstrates a weak positive correlation with airfare and access cost to the airport, indicating a higher travel budget for longer trips. Conversely, trip duration exhibits a weak negative correlation with age and access time to the airport.

Airfare shows a positive correlation with access cost and access time to the airport, suggesting travelers are willing to invest more time and expense in transportation when purchasing costlier flight tickets. Nonetheless, the correlation is weak. Moreover, access cost displays a positive relationship with age, trip duration, airfare, and access time, with the strongest correlation observed with access time (0.253).

These findings offer insights into the modest linear relationships among continuous variables within the dataset, guiding subsequent data exploration and model development efforts.

Data Processing

Outliners

In our dataset, certain variables exhibit a wide range of values. For example, *airfare* spans from 30,000 to 2,600,000 KRW. We do not classify airfare outliers, as the costs are influenced by factors such as airline, seating class, purchase timing, and method. Special cases, such as staff-discounted airfare, require careful consideration. Likewise, *access time* to the airport varies from 4 to 390 minutes. Again, we refrain from classifying outliers in access time. Unique situations, like passengers staying in airport hotels, should be taken into account.

Missing Values

Several variables in our dataset have missing values, requiring some data manipulation.
For the four missing values in *airline*, we first check the flight number for clues about the operating airlines. If the flight number is also missing, we simply discard the observation.
Regarding the four missing values in *seat class*, we infer the possible seat class by examining airline, airfare, and destination. If airfare is missing, we assign the most frequent seat class from the dataset.

For the 165 missing values in *airfare*, we estimate the data by considering airline, destination and seat class, calculating the average airfare for the observations. As for the 98 missing values in *access time*, we determine the data by examining airport and province of residence, and calculating the average access time for the observations.

For the missing values in age, gender, nationality, and destination, we substitute with the mode of the corresponding variables. However, concerning the frequent flyer program, mileage, access cost to the airport, and income, each of these variables has over 100 missing values and is not suitable for analysis.

Recategorization

For nominal variables, we need to regroup them into binary variables based on choice probability and the number of observations. It is important to note that the regrouping for airport and airline models will differ. Initially, the regrouped variables will be put into analysis, and further regrouping may be done based on the performance of the models.

Modeling

Methodology & Model Description

We aim to develop models that describe passengers' airport and airline choice behavior respectively. To achieve a comprehensive understanding, we plan to employ logistic regression from discrete choice methods and decision trees from machine learning methods for analysis. Additionally, we will utilize other models such as neural networks and nearest neighbors for model evaluation. All models used in this analysis are supervised learning models.

Logistic regression is a statistical method primarily employed for binary classification tasks, focusing on predicting the probability of an instance belonging to one of two possible classes. It models the relationship between independent variables and the dependent variable using the logistic function. In the airport choice model, logistic regression predicts the likelihood of selecting either Incheon Airport or Gimpo Airport. Similarly, in the airline choice model, it predicts the probability of choosing either Korean full-service carriers (Korean Air and Asiana Airlines) or others (Korean low-cost carriers and foreign airlines).

A decision tree, a versatile supervised machine learning model for both classification and regression tasks, recursively partitions the data into subsets based on feature values, aiming for homogeneous subsets concerning the target variable. Each node represents a decision based

on a feature's value, while branches depict outcomes, and leaf nodes offer predicted outcomes or target values. In our analysis, variables were not regrouped in the decision tree.

A neural network comprises interconnected neurons organized into layers, each receiving input signals, conducting computations, and yielding output signals passed to subsequent layers. Meanwhile, Support Vector Machine (SVM), another supervised machine learning approach, targets classification and regression tasks. SVMs endeavor to determine the optimal hyperplane, effectively segregating data into distinct classes by maximizing the margin — the gap between the hyperplane and the nearest data points from each class, known as support vectors.

These two methods are employed to assess and compare the performance of the first two models, aiding in selecting the most fitting model tailored to the airport and airline choices based on the performance metrics.

Variable Selection

Several variables are dropped from the analysis. As previously mentioned, the *frequent flyer program*, *mileage*, *access cost to the airport*, and *income* are excluded due to the large number of missing values. Additionally, *departure hour* and *departure minute* are not included since equivalent information is available from the nominal variable departure time. *Flight number* and *respondent ID* are dropped as they are irrelevant to making airport and airline decisions. Lastly, only observations related to destinations in China, Japan, and Southeast Asia are included, as these are the common destinations served by Gimpo Airport and Incheon Airport.

We adopt a forward selection approach that adds or deletes one independent variable at a time based on statistics metrics.

Modelling Results - Logistic Regression

Logisitc regression is a power statistical technique because of its availability on probabilistic interpretation, ability to assess feature importance, and efficiency on computation.

Airport Choice Model

Our final model on airport choice is as follows:

Dependent variable: Airport - Incheon Airport				
	Coefficient	Standard error	z-value	p-value
Intercept	-2.1109	0.363	-3.566	0.000
Airline - Korean Air & Asiana Airlines	1.0556	0.525	2.904	0.004
Airline - Korean low-cost carriers	1.8460	0.337	3.513	0.000
Trip Purpose - Leisure	0.8079	0.044	2.398	0.017
Flying Companion	-0.1264	0.365	-2.852	0.004
Destination - Japan	-2.3362	0.380	-6.409	0.000
Province Residence - Seoul	0.8145	0.663	2.143	0.032
Province Residence - Incheon	1.2041	0.404	1.816	0.069
Province Residence - Kyungki-do	0.8017	0.573	1.987	0.047
Departure Time - 6am to 12nn	-2.6010	0.383	-4.543	0.000
Mode of Transport - Car	-0.8983	0.299	-2.344	0.019
No of Transport	1.0153	0.592	3.398	0.001
Log-Likelihood: -145.83				
Number of observations is 451				

The odds of a passenger using Incheon Airport are about 0.121 times the odds of them using Gimpo Airport, suggesting a lower likelihood of passengers using Incheon Airport when all other factors are excluded from consideration.

Several factors contribute to decreasing the likelihood of passengers choosing Incheon Airport. These include traveling with a larger group, heading to destinations in Japan, departing between 6 am and 12 noon, and using cars to reach the airport.

Conversely, passengers tend to prefer Incheon Airport when flying with Korean Air, Asiana Airlines, or Korean low-cost carriers. This preference is also observed among passengers traveling for leisure or residing in Seoul, Incheon, or Gyeonggi-do. Additionally, the likelihood of passengers selecting Incheon Airport increases with each additional number of transportation required to reach the airport.

Airline Choice Model

In our analysis of airline preferences, we simplify the nominal dependent variable *airline* into binary categories. We distinguish between Korean full-service carriers (Korean Air and Asiana Airlines) and other choices, which include Korean low-cost carriers and foreign airlines. Here is our final model on airline choice:

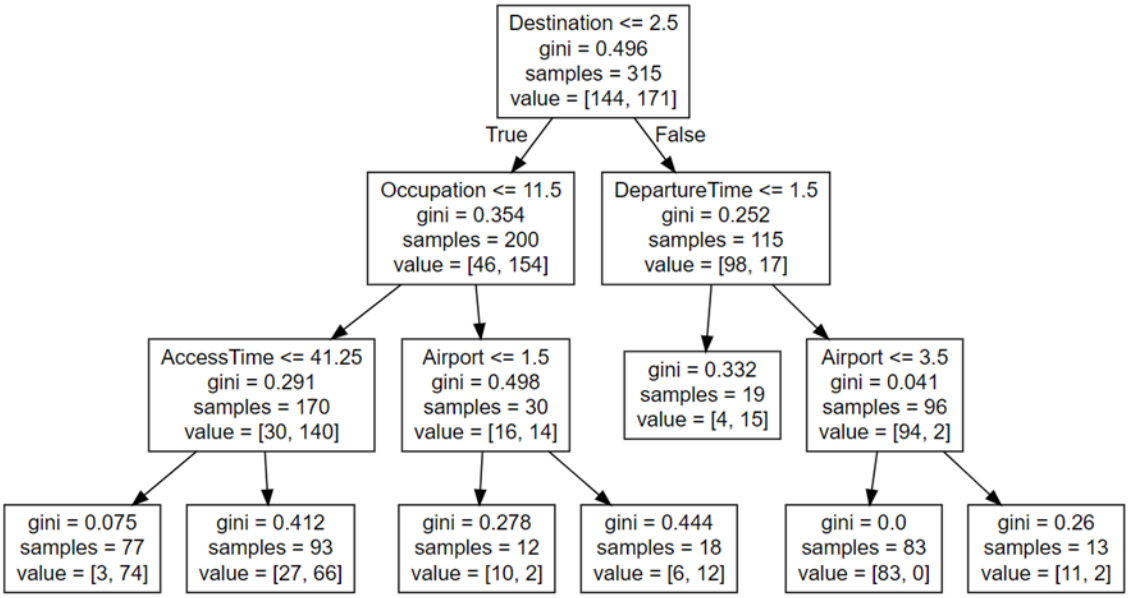
Dependent variable: Airline - Korean full-service carriers				
	Coefficient	Standard error	z-value	p-value
Intercept	-4.5040	0.795	-5.633	0.000
Airport - Incheon Airport	1.7319	0.433	4.000	0.000
Destination - China	2.3005	0.456	5.043	0.000
Destination - Japan	2.4200	0.499	4.848	0.000
Trip Duration	0.0041	0.002	1.680	0.093
Group Travel - Yes	1.0017	0.417	2.404	0.016
Depature Time - 12nn to 6pm	0.6183	0.285	2.167	0.030
Airfare	0.0280	0.009	3.014	0.003
Age	0.0186	0.010	1.787	0.074
Nationality - China or Japan	-0.6740	0.408	-1.650	0.099
Log-likelihood: -176.15				
Number of observations is 451				

In the airline choice model, most of the predictors are trip-related. The likelihood of passengers opting for Korean full-service carriers is relatively lower, approximately 0.01 times, compared to choosing Korean low-cost carriers or foreign airlines when other factors are not considered. One factor also contributing to this tendency is the nationality of passengers, particularly those from China or Japan.

On the other hand, passengers show a preference for Korean full-service carriers when departing from Incheon Airport, traveling to destinations in China or Japan, or participating in group travel. This preference is also noticeable among passengers with longer travel durations, departing between 12 noon and 6 pm, or paying higher airfares. Additionally, the likelihood of passengers selecting Korean full-service carriers tends to increase with age.

Modelling Results - Decision Tree

Airport Choice Model



The visualized decision tree illustrates the model's decision-making process, starting with the 'Destination' feature and subsequently making decisions based on additional features such as 'Occupation', 'DepartureTime', 'AccessTime', and 'Airport'. At each node, the decision tree provides essential information including the Gini impurity, the number of samples that reached that node, and the distribution of the target class among those samples, denoted as 'value'. The leaves of the tree, where the final classifications are made to predict outcomes, are particularly noteworthy; a Gini impurity of 0 at a leaf indicates a perfect separation of classes at that node, which is ideal for accurate predictions. Given that the maximum depth of the tree is 3, the tree is relatively shallow. This shallowness is beneficial as it aids in generalizing to new data by simplifying the decision rules, making the model potentially more robust and easier to interpret.

A DecisionTreeClassifier is configured with parameters that define a pruned tree structure to combat overfitting and improve its ability to generalize to new data. The depth of the tree is restricted to three levels through the max_depth parameter set to 3. This limitation prevents the

tree from growing too complex and fitting to noise in the training data. Furthermore, each leaf node is required to contain at least 10 samples, as dictated by the min_samples_leaf parameter.

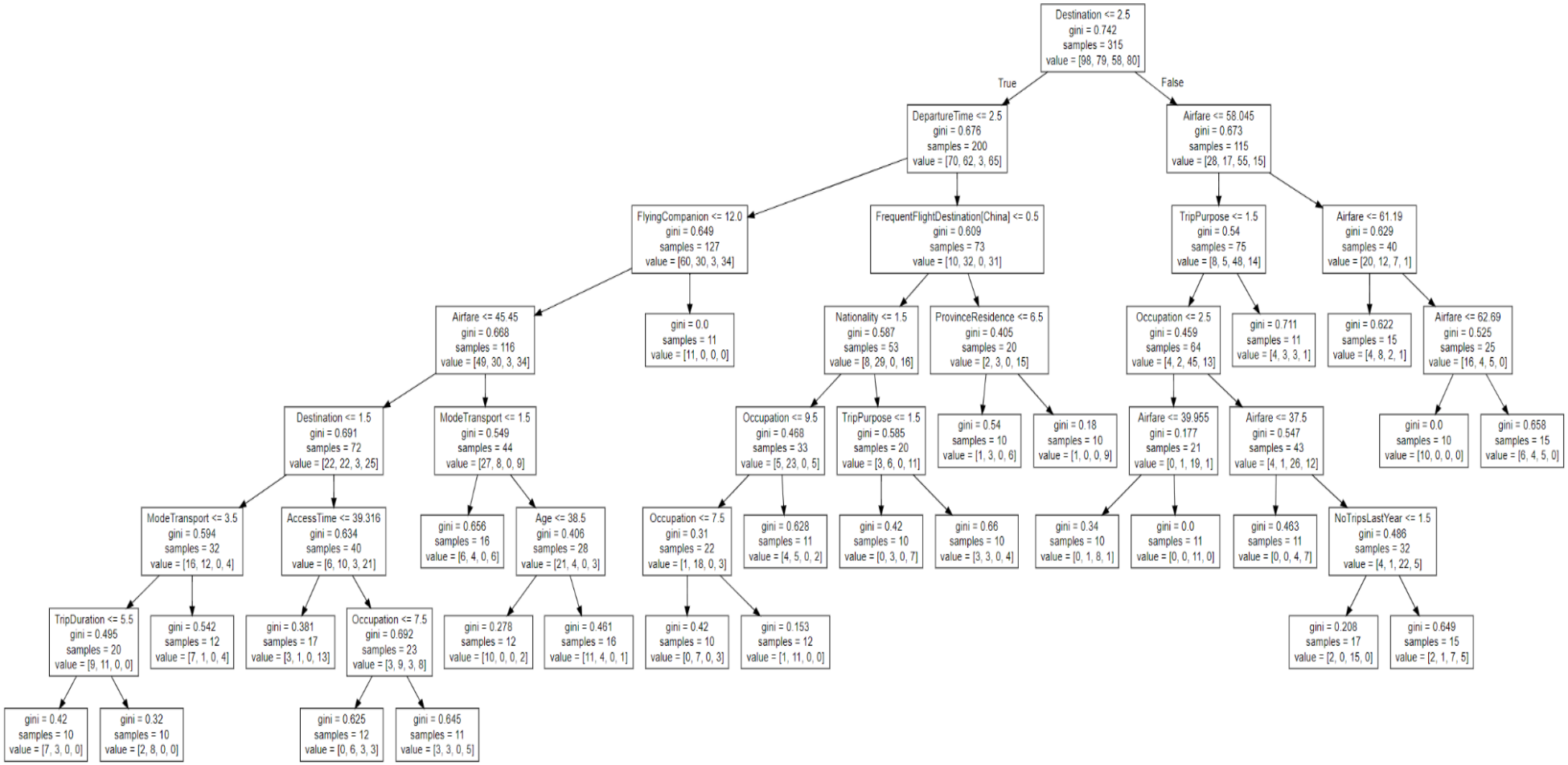
This condition ensures a minimum amount of data to make reliable decisions at the leaves of the tree. In addition, for any split to be considered in the tree, a minimum of 10 samples must be present, which is specified by the min_samples_split parameter. This rule helps in avoiding splits that are based on small, potentially unrepresentative data subsets. The random_state parameter is set to 100, guaranteeing that the results of the model are consistent across different runs. All other parameters are left at their default settings provided by scikit-learn, which are designed to work well for a wide range of problems.

The reported accuracy of the model is 0.8015, indicating that it correctly predicts the target variable 80.15% of the time on the test data. The precision of the model is notably high at 0.913, signifying that when the model predicts a positive class, it is correct 91.3% of the time. However, the recall is somewhat lower at 0.6462, which points out that the model is missing a significant number of actual positive cases. Furthermore, the confusion matrix offers a more detailed breakdown of the predictions, categorizing them into true positives, true negatives, false positives, and false negatives, thereby providing a comprehensive analysis of the model's performance.

The model demonstrates commendable performance on the test set, although the relatively low recall indicates a cautious approach towards predicting the positive class, suggesting the model may be erring on the side of specificity over sensitivity. The decision to utilize a shallow tree depth reflects a strategic effort to mitigate the risk of overfitting, ensuring that the model does not become too tailored to the training data at the expense of its ability to generalize to new data. The employment of a fixed random_state parameter is key to maintaining consistency in the model's performance across different runs, highlighting that alterations to this value could potentially affect the model's splits and, consequently, its overall performance.

The balancing act between precision and recall underlines a deliberate trade-off, which varies in importance depending on the application context. In certain scenarios, achieving higher precision and thereby reducing false positives may be prioritized over attaining a higher recall, which focuses on identifying all positive cases.

Airline Choice Model



The model, with a depth of 10, is designed to capture complex patterns, yet this depth raises concerns about potential overfitting, especially if the decision tree becomes overly intricate for the dataset at hand. Evaluation metrics and the confusion matrix reveal that the model's performance is far from perfect, with an accuracy of around 54.41%, indicating a substantial need for improvement. The similarity in precision and recall metrics suggests that the model maintains a balance in its predictions, neither being overly conservative nor excessively liberal in identifying the positive class. The complexity of the decision tree is evident from its visualization, which shows numerous splits and a variety of features being used for decisions.

This complexity hints at possible overfitting, particularly when considering that the training data might contain noise or patterns that are not representative of the broader dataset. Certain splits in the tree lead to nodes with very low Gini scores, indicating instances where the model is highly confident in its classifications. Nonetheless, the model employs a strategy to somewhat counter overfitting by setting a minimum sample requirement at each leaf node (min_samples_leaf=10), which helps ensure that the model does not tailor its predictions too closely to the training data.

Model Validation & Model Evaluation

Throughout our models, we utilize the holdout method to split the dataset into two segments: the training set and the test set. This method involves random partitioning the data, with 70% allocated to the training set and 30% to the test set in our analysis.

We extend our evaluation by applying Support Vector Machines (SVM) and neural networks. This approach enables us to assess model interpretability and generalization performance, prioritizing informative features while avoiding overfitting. In the SVM, we employ the linear kernel, which shares conceptual similarities with logistic regression, as both methods yield linear decision boundaries. For the neural network, we utilize a model comprising three neurons.

Neural Network Modeling

Airport Choice Model

The confusion matrix for the neural network reveals that it correctly predicted a total of 108 instances, summing up both true positives and true negatives, while making 28 incorrect predictions, combining both false positives and false negatives. This results in an accuracy of 79.41%, meaning the model successfully predicts the outcome correctly in approximately 79.41% of cases. The precision of the model stands at 82.46%, indicating that when it predicts the positive class, it is correct around 82.46% of the time. Furthermore, the recall of the model is

72.31%, showing that it correctly identifies about 72.31% of all actual positive cases. To fully assess the effectiveness of this performance, comparing these results to typical benchmarks or the performance of previous models in the specific task context is essential.

In summary, this simple neural network model shows a relatively high level of accuracy and precision with a slightly lower recall, suggesting it is more conservative in predicting positive classes and may miss some true positives. The performance metrics suggest that the model is relatively robust, but there's a tradeoff between precision and recall that could be further optimized, perhaps by adjusting the threshold for classification, adding more neurons or layers, or tuning other hyperparameters.

Airline Choice Model

The accuracy of the neural network is approximately 66.91%, indicating that just over two-thirds of the model's predictions on the test set are correct. The precision, at about 65.82%, shows that when the model predicts the positive class, it does so correctly around 65.82% of the time. With a recall of 74.29%, the model demonstrates a relatively better capability at identifying all relevant instances within the actual class, highlighting its effectiveness in capturing the positive cases. However, the confusion matrix reveals the presence of a significant number of false predictions, including 27 false positives and 18 false negatives, underscoring the potential for the model to improve its predictive accuracy and reduce errors.

Support Vector Machine

In the investigation of customer behavior regarding airport and airline selection, the application of Support Vector Machines (SVMs) with a linear kernel function has been a focal point for predictive modeling. The linear kernel's simplicity often makes it the first choice in SVM modeling due to its direct approach in finding a separating hyperplane in the feature space.

Airport Choice Model

The linear kernel delivered a notable performance with an accuracy of 81.62%, indicating a high level of correct predictions out of all cases. The precision of the model stood at 84.48%, signifying that when it predicted an airport choice, it was correct most of the time. The recall was slightly lower at 75.38%, reflecting the proportion of actual positive instances that were correctly identified by the model. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC) was calculated to be 0.8135, demonstrating the model's strong capability to discriminate between the classes of interest.

Airline Choice Model

Similarly, in the airline SVM model, the linear kernel also showed considerable results. The model achieved an accuracy of 63.24%, which, although lower than the airport model, still presents a significant predictive power. The recall rate was high at 70%, indicating the model's sensitivity in identifying airlines that customers are likely to choose. The precision was 62.82%, showing that there were more false positives compared to the airport model. The AUC for this model was 0.6303, suggesting moderate discrimination ability.

Model Comparison

To compare the results of logistic regression, decision tree, neural network, and Support Vector Machine (SVM), we employ accuracy, precision, and recall as performance metrics. Additionally, for our logistic regression model, we utilize the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and R-squared to compare different models and select the most appropriate one. AIC measures the relative quality of the statistical model, while BIC, sharing a similar criterion, takes into account model complexity. The coefficient of determination (R-squared) provides insight into the proportion of data variation explained by regression models.
Here is a summary of the performance of the methods:

	Airport Choice Model			
	Logistic Regression	Special Vector Machine	Neural Network	Decision Tree
Accuracy	0.7941	0.8162	0.8015	0.8015
Precision	0.8364	0.8488	0.8519	0.9130
Recall	0.7077	0.7538	0.7077	0.6462
AIC	315.66	*n.a.*	*n.a.*	*n.a.*
BIC	365.00	*n.a.*	*n.a.*	*n.a.*
R-squared	0.3285	*n.a.*	*n.a.*	*n.a.*
Area under ROC	*n.a.*	0.8135	*n.a.*	*n.a.*

	Airline Choice Model			
	Logistic Regression	Special Vector Machine	Neural Network	Decision Tree
Accuracy	0.6387	0.6324	0.6838	0.5441
Precision	0.6364	0.6282	0.6753	0.5493
Recall	0.7000	0.7000	0.7429	0.5441
AIC	372.30	*n.a.*	*n.a.*	*n.a.*
BIC	413.41	*n.a.*	*n.a.*	*n.a.*
R-squared	0.1842	*n.a.*	*n.a.*	*n.a.*
Area under ROC	*n.a.*	0.6303	*n.a.*	*n.a.*

Descriptive models (like logistic regression and decision trees) are more transparent and easier to interpret, making them suitable for applications where understanding the decision-making process is important. Non-descriptive models(like neural networks and SVM) are often more complex and less interpretable, but they can capture complex patterns and interactions that might be difficult for simpler models to grasp. They are chosen for their performance in applications where accuracy is more critical than interpretability.

Accuracy measures the overall correctness of model predictions, while precision gauges the proportion of true positives among all predicted positives, and recall measures the proportion of true positives that were correctly identified.In an airport choice model, all modeling results indicate high accuracy, with an even higher precision, but a lower recall. High accuracy indicates good overall correctness in classifying a large portion of predictions accurately. Precision and recall offer more nuanced insights; lower precision suggests potential false positives, indicating that the model may miss identifying some positive cases. To enhance the capture of all positive cases, collecting additional information through surveys is advisable. The decision tree model showed the most promising performance among the models in the airport choice model. This is likely due to its flexibility in handling different data types. Specifically, it can manage mixed data types, including both continuous and nominal variables in our case, without requiring extensive preprocessing. This characteristic helps reduce the risk of errors associated with data manipulation.

In an airline choice model, the task becomes more challenging. The model's accuracy is relatively moderate. For logistic regression, Support Vector Machine, and neural network models, a higher recall than precision suggests better performance at capturing positive instances compared to negative instances. For decision tree models, all three performance metrics indicate room for improvement. This performance gap may be attributed to limited information for analysis. Including more trip-related features, such as flight schedule and availability, flight service quality, and historical data on delays or early arrivals, could enhance model performance.

Insight and Policy Implication

According to air traffic statistics, in 2023, Incheon Airport hosted nearly 340,000 flights and over 56 million passengers, while Gimpo Airport hosted over 134,000 flights and approximately 23 million passengers. Compared to the fully international Incheon Airport, Gimpo Airport primarily serves domestic flights and offers limited international flights to China, Japan, and Taiwan. Gimpo Airport is primarily accessible by railway lines, while Incheon Airport offers access via railway, ferries, Korean Train eXpress (KTX), and both airports are linked by the Incheon International Airport Expressway.

From the airport perspective, Gimpo Airport does not pose as a competitor to Incheon Airport but rather serves as a supplementary facility to alleviate and bolster support for Incheon Airport. Our model results indicate that passengers consider mostly trip-related factors when selecting airports, with a higher inclination towards using Incheon Airport for leisure travel, residing near Seoul, Incheon, or Kyngki-do, and requiring extensive transportation options.

Gimpo Airport enjoys a strategic location advantage in proximity to the Seoul business area. Consequently, it would be prudent for the airport to target business travelers, thus optimizing their transportation time. Additionally, targeting early morning flights could prove advantageous. Lastly, the transportation department could incentivize airlines to utilize Gimpo Airport as an interchange point for transferring international passengers arriving at Incheon Airport to domestic flights at Gimpo Airport, with the Airport Railroad (AREX) All Stop Train offering a swift 30-minute transfer time.

Airport and airline choices often go hand in hand. Common trip-related factors include destination and departure time. Specifically, when selecting airlines, passengers consider airfare, trip duration, and whether they are traveling alone or in a group. To enhance the competitiveness of Korean full-service carriers in the local market, airlines should carefully manage airfare elasticity, expand destination offerings, and optimize daytime departures. Other

noteworthy factors, not covered in our analysis, include flight service quality, frequency of delays, and frequent flyer programs, as supported by articles on airline choices.

Conclusion

This report offers a brief discussion on passengers' airport and airline preferences, drawing from survey responses collected at Incheon Airport and Gimpo Airport. Through logistic regression and decision tree analysis, we find that passengers predominantly prioritize trip-related features when choosing both airports and airlines, which isn't surprising. These findings can serve as valuable insights for policymakers aiming to strategically position and enhance the competitiveness of Incheon Airport, Gimpo Airport, and Korean airlines. As 2024 marks the first fully recovered year from the impacts of COVID, there is ample opportunity for the flight businesses to thrive if the right strategies are implemented.