Modeling Passengers’ Airport and Airline Choice Behavior

# 1. Introduction

This report presents a comprehensive analysis of passengers’ airport and airline choice behaviors in the Seoul Metropolitan Area, focusing on two major airports: Gimpo (GMP) and Incheon (ICN). The objective of this project is to investigate the factors influencing passengers' choices between these airports and their preferred airlines. This investigation applies various modeling approaches to understand the complex decision-making processes of passengers. The findings aim to provide substantive policy implications, enhancing the efficiency and effectiveness of service offerings by airports and airlines.

# 2. Data

## 2.1 Summary of EDA Results

The dataset consists of survey responses from 488 individuals, encompassing 27 variables that include both socio-demographic (e.g., age, gender, occupation, income) and alternative-specific attributes (e.g., flight information, travel time, mode of transport). Key variables of interest are 'Airport choice' (binary: ICN or GMP) and 'Airline choice' (categorical: Korean Air, Asiana Air, Korean LCC, and Foreign Carriers).

Respondent ID: A unique identifier for each survey participant.

Airport Choice: The selected airport for the respondent's travel, such as Incheon (ICN) or Gimpo (GMP) Airport.

Airline Choice: The airline chosen by the respondent, including options like Korean Air, Asiana Airlines, Korean LCC, and other foreign airlines.

Age: The respondent's age.

Gender: The gender of the respondent, categorized as male or female.

Nationality: The nationality of the respondent, with options like Korea, China, Japan, Southeast Asia, or other.

Purpose of Trip: The main reason for travel, such as leisure, business, study, or other.

Trip Duration: The total number of days spent on the trip.

Flying Companion: The number of people accompanying the respondent on the flight.

Province Residence: The respondent's current residential province within or outside Korea.

Group Travel: Indicates whether the respondent is traveling as part of a group.

NoTripsLastYear: The number of international trips taken by the respondent in the previous year.

Frequent Flight Destination: The most common destination area for the respondent's flights.

Flight Destinations: The specific destination region for the current trip.

Flight Number: The flight number associated with the respondent's trip.

Departure Hour: The hour of departure for the respondent's flight.

Departure Minute: The minute of departure for the respondent's flight.

Departure Time: A categorical representation of the departure time slot.

Seat Class: The class of seating chosen by the respondent, such as economy, business, or first class.

Airfare: The cost of the airfare in 10,000 KRW units.

NoTransport: The number of transportation modes used to arrive at the airport.

Mode of Transport: The primary mode of transportation used to reach the airport.

Access Cost: The total cost of transportation to the airport in KRW.

Access Time: The time taken to travel to the airport, in minutes.

Occupation: The occupational category of the respondent, such as entrepreneur, student, or retired.

Income Class: The income bracket of the respondent, ranging from 30 million won or less to over 200 million won.

Mileage Airline: The airline associated with the respondent's frequent flyer mileage.

Mileage: The amount of frequent flyer miles accrued by the respondent.

A screenshot of a computer

Description automatically generated

## 2.2 Summary of Correlation among variables:

Diving into the correlation matrix from the travel data set unveils various strengths of relationships between pairs of variables. A red square points to a positive correlation, and a darker blue square indicates a negative one. The strength of the relationship is shown by the color intensity.

A graph of a plane

Description automatically generated with medium confidence

Key observations emerge upon closer inspection. 'ModeTransport', 'AccessCost', 'AccessTime', and 'Mileage' show positive correlations with each other, logically reflecting that the type of transport influences the cost and time taken to reach the airport, as well as the distance traveled.

Airfare has a moderately positive correlation with 'SeatClass', suggesting a direct link between the class one flies in and the price paid. It's an expected relationship, grounded in the understanding that higher seat classes are typically priced higher.

The mix of blue across the matrix, especially in areas like 'ID', 'Airport', and 'Airline', shows a lack of strong linear relationships with other variables. This lack of correlation indicates independence, meaning these factors don't significantly sway others within the data set.

## 2.4 Data Preprocessing, Null Values Analysis and Treatment

The graph below presented a clear visualization of the completeness of each variable in the dataset. 'Mileage' stands out with the most significant number of missing entries, nearly reaching 400. This is followed by 'Mileage/Airline' and 'AccessCost', which also show considerable gaps but to a lesser extent. The drop in missing values is quite sharp after these top three, with 'Airfare' and 'FlightNo' at roughly half the missing count of 'Mileage/Airline'.

A graph with different colored bars

Description automatically generated

Toward the lower end of the graph, a substantial number of columns, such as 'Nationality' and 'TripDurationCat', display a complete set of data, suggesting that these fields were either well-maintained during data entry or have been filled in during an earlier preprocessing stage.

Columns with high null value count is dropped because of their non usability in the model. However, keeping the end goal of predicting the airline and airport choice, certain attributes are important to include for better results of model. Two such attributes are picked from the list, “Access Time” & “Departure Hr” for which missing values are filled using average or mode.

**Treatment: “Access Time”:**

We initially had 97 missing values in the dataset. After using the group mean values by 'ProvinceResidence', 'ModeTransport', and 'NoTransport' as proxies, the number of missing 'AccessTime' values was reduced to 15. This shows the effectiveness of the method in dealing with most of the missing data.

For interpretation, we're looking at the group counts, which indicate the sample size for each combination of 'ProvinceResidence', 'ModeTransport', and 'NoTransport'. For example, there are 37 instances where the 'ProvinceResidence' is 1, 'ModeTransport' is 1, and 'NoTransport' is 1. These counts are essential as they give us an idea of the sample sizes that contribute to the mean 'AccessTime' used for imputation.

Knowing the size of these groups helps to validate the imputation: larger group sizes generally mean a more reliable mean estimate. Smaller groups might lead to less reliable imputation and could be the reason why we still have 15 missing values after the imputation process; if all 'AccessTime' values are missing for a particular group, no mean can be calculated, and hence, no imputation is possible for that group

**Treatment: “Departure Hour”:**

While Treating the Departure Hour, Mode was used instead of Mean a treatment of missing values as the mode represents the most common departure hour for a given destination and time, which is more meaningful for a scheduling variable than an average.

Using the mode for 'DepartureHr' is indeed sensible when we're interested in the most frequent departure time rather than an average that might not correspond to any actual departure time as Hours of the day are often treated as categorical data in scheduling contexts because they represent distinct categories rather than quantities. This approach ensures that the filled values align with real-world scheduling and keeps the data in a format that can be directly used for planning or operational decision-making.

Initially, there were 34 missing values in the 'DepartureHr' column. After the imputation process, this number was reduced to 3, indicating the mode was successfully applied as a fill-in for most missing instances.

The group counts provide a snapshot of the dataset, indicating the number of records available for each 'Destination' and 'DepartureTime' pairing. For instance, there are 9 instances where the 'Destination' is 1.0 and 'DepartureTime' is 1, and 73 instances where 'DepartureTime' is 2. These counts are crucial for understanding the robustness of the mode value used for each group. Larger groups provide a more reliable mode value, while smaller groups might not offer a significant mode for imputation, which could explain why there are still 3 missing values post-imputation.

In essence, the imputation strategy here ensures that the most common 'DepartureHr' for each 'Destination' and 'DepartureTime' is used to fill gaps, preserving the underlying distribution of departure hours across different destinations and times.

As part of the **Data Preprocessing** following treatment is carried out:

**Dropping Missing Destinations:** We start by eliminating any records where the 'Destination' field is missing. This ensures that every entry in our dataset has a defined destination, which is crucial for any travel-related analysis.

**Filtering Destination Categories:** Next, we remove entries categorized under 'Destination' as “others” as it is not relevant to the objective of model.

**Removing Missing Airlines:** Finally, we discard records with missing 'Airline' data. Similar to 'Destination', the 'Airline' field is essential, and its absence would compromise the integrity of any subsequent analysis focused on airline preferences or performance.

# 3. Summary of Feature categorization:

This section outlines all the feature categorization performed regardless of which model a feature is used in.

For the categorization, the chi-square test of independence is a common statistical test used to determine if there is a significant association between two categorical variables

**Transport Mode:**

When examining the new transport mode groupings in relation to airport choice, the chi-square test of independence provides these results:

Chi-Square Statistic: 9.971

p-value: 0.01881

Degrees of Freedom: 3

The p-value is below the 0.05 threshold, indicating a statistically significant association between the mode of transport groups and the airport choice. This means that the way passengers travel to the airport (like by car, taxi, bus, or train) is related to which airport they choose, and this isn't just due to random chance.

We can see how many people chose each mode of transport for each airport. For example, the "Car/Taxi/Bus" group has more people choosing Gimpo (GMP) Airport over Inchoen (ICN) Airport, while the "Rail/KTX/Limousine" group has a noticeably higher number choosing airport ICN over GMP.

Based on the statistical tests and analyses we've done, here are the groupings for mode of transport that show significant associations with airport and airline choices:

Mode of Transport Groupings for Airport and Airport Choice:

Personal & Road Transport: Combines Car (1), Taxi (2), and Bus (3). This group is associated with airport 2 more than airport 1.

Public City Transit: Combines Express Bus (4) and Subway (5). The association with airports is less distinct for this group, but it still shows a preference.

High-Speed & Premium Transit: Combines Airport Limousine (6), Rail (7), and KTX (Express Rail) (8). This group prefers airport 1.

Specialty & Miscellaneous Travel: Combines Shuttle provided by travel agency (9), Air (10), and Other (11). The sample size is small, so the preference is not very clear, but there is a slight inclination towards airport 2.

The associations here suggest that the type of transport a passenger uses to get to the airport can influence their choice of airport and potentially the airline they fly with. The specific reasons behind these preferences could be numerous, including convenience, cost, travel time, or other factors related to the services offered by the airports and airlines.

This tells us that certain modes of transport might be more convenient or preferred for getting to one airport over another, possibly due to factors like distance, travel time, or accessibility.

A graph of different types of groups

Description automatically generated with medium confidence

**Occupation:**

For the groupings of occupations, the chi-square test of independence results are as follows:

Chi-Square Statistic: 13.577

p-value: 0.00354

Degrees of Freedom: 3

These results tell us that there is a statistically significant association between the occupation groups and the choice between Korean versus Foreign airlines. The p-value is well below the conventional threshold of 0.05, which means that the pattern we're seeing in the data—where different occupation groups might prefer different airlines—is likely not just happening by chance.

Groupings:

High-Level Professionals: Combines 'Entrepreneur, Senior management' (1), 'Business (Corporate worker)' (2), and 'Professionals (doctor, lawyer, professor)' (5).

Service & Labor: Combines 'Sales, Service' (3), 'Government, Military' (4), and 'Industrial, Manufacturing' (6).

Self-Employed & Others: Combines 'Self-employed' (7), 'Other' (12), and any other small groups that don't fit well into the above categories.

Non-Working Individuals: Combines 'Student' (8), 'Housewife' (9), 'Retired' (10), and 'None' (11).

# 3. Models and Results

This section outlines the methodologies and key findings from the application of different modeling techniques to describe passengers' choice behaviors. Each model's results are discussed in detail, highlighting the features and factors that most significantly impact passenger choices.

## 3.1 Airline Models

This section outlines all the models used for predicting Airline choice.

## 3.1.1 Airline: Logistic Regression Model

The logistic regression model was developed to predict passengers' choices between a Korean major airline and other airlines. The model included a variety of predictor variables such as Age, Trip frequency, Departure airport, Gender, Destination, Seat class, Trip purpose, Nationality, and Occupational categories.

A number and numbers on a white background

Description automatically generated

The model's performance metrics indicate an overall accuracy of 69.23%, which suggests that it correctly predicted the airline choice for approximately 69 out of 100 passengers. Precision, which measures the proportion of true positives among all positive predictions, is 63.27%, indicating that when the model predicts a passenger will choose a Korean major airline, it is correct around 63 times out of 100. The recall of 75.61% shows that the model successfully identified approximately 76 out of every 100 passengers who actually chose a Korean major airline.

A screenshot of a computer

Description automatically generated

Examining the coefficients from the logistic regression output, we observe that factors such as 'Incheon' (indicating the departure airport), 'Economy' (indicating the seat class), and 'Southeast Asiad' (indicating the destination) have positive associations with choosing a Korean major airline, while variables such as 'Study' and 'SoutheastAsian' (indicating nationality) show negative associations.

The p-values associated with the coefficients suggest which variables are statistically significant predictors of airline choice. For instance, 'HighLevelProfessionals' has a p-value of 0.015, which is below the standard significance level of 0.05, indicating a statistically significant negative association with choosing a Korean major airline.

In conclusion, the logistic regression model has provided insightful predictive power regarding passengers' airline choices.

The confusion matrix indicates that the model has some tendency to predict passengers as choosing Korean major airlines when they actually chose another airline (false positives), which is reflected in the lower precision compared to recall. This indicates that while the model is relatively good at identifying passengers who will choose Korean major airlines, it also mistakenly predicts Korean major airlines for passengers who do not choose them.

The overall performance suggests that the model has merit, but there is room for improvement, particularly in distinguishing between passengers who choose Korean major airlines and those who do not.

## 3.1.2 Airline: Decision Tree Model

The accuracy of the model is approximately 65.93%, which suggests that the model correctly predicts the airline choice for roughly 66 out of every 100 passengers. This is a moderate level of accuracy and indicates that there's room for improvement.

A number and numbers on a white background

Description automatically generated

The precision of the model, which is about 64.71%, indicates the proportion of passengers who were correctly predicted to choose a Korean major airline out of all the passengers predicted to do so. This means that when the model predicts a passenger will choose a Korean major airline, it is correct about 65 times out of 100.

The recall of the model is approximately 53.66%, which is the ability of the model to find all the relevant instances of Korean major airline choices. In other words, the model identifies roughly 54 out of every 100 passengers who actually chose a Korean major airline

To summarize, while the decision tree model has demonstrated a reasonable ability to predict airline choice, it's more conservative when it comes to predicting passengers choosing a Korean major airline (higher precision than recall). It tends to miss more actual Korean major airline choices (lower recall). This may indicate that the model is more selective in predicting a Korean major airline choice, potentially at the cost of missing some of these actual choices.

## 3.1.3 Airline: Support Vector Machine

A number and numbers on a white background

Description automatically generated

SVM model shows a decent recall, indicating a good capability to identify passengers who will choose a Korean major airline. However, the precision is moderately low, meaning that there is a higher proportion of false positives among the predicted Korean major airline choices. The accuracy of the model is reasonable but suggests there is room for improvement.

## 3.1.4 Airline: Neural Network:

A number and numbers on a white background

Description automatically generated

The neural network model has moderate performance in predicting airline choice, with roughly equal chances of correctly identifying or missing a Korean major airline choice. Precision and recall are both close to 50%, suggesting that the model's ability to distinguish between the choices is not much better than a random guess.

## 3.1.5 Airline: Model Comparison and Evaluation

Logistic Regression: Best balance with accuracy at 69.23%, precision at 63.27%, and recall at 75.61%.

Decision Tree: Lower accuracy at 65.93%, with higher precision at 64.71% and lower recall at 53.66%.

SVM: Moderate accuracy at 64.84%, with precision at 59.18% and the highest recall at 70.73%.

Neural Network: Lowest accuracy at 56.04%, precision at 51.35%, and recall at 46.34%.

Overall, the logistic regression model performs best in terms of both accuracy and precision, while the SVM model has the best recall. The neural network model lags behind the others in all three metrics.

# 3.1.6 Airline: Conclusions and Recommendations

Based on the performance metrics, the logistic regression model emerges as the most effective for predicting airline choice. Here are some conclusions and recommendations derived from its analysis:

**Conclusions:**

Balanced Performance: The logistic regression model offers a well-rounded performance, with the highest accuracy and precision among the tested models. It is reliable for both predicting Korean major airline choices correctly and limiting false positives.

Strong Recall: Although the SVM showed a slightly higher recall, indicating its strength in identifying most of the actual positive instances, its overall lower precision makes it less reliable for precise targeting.

Predictor Influence: Certain variables such as departure airport and seat class have shown a positive influence on the likelihood of choosing a Korean major airline, which can inform strategic emphasis in marketing campaigns.

**Recommendations:**

Targeted Marketing: Leverage insights from the logistic regression model to tailor marketing efforts that align with the positive predictors of choosing a Korean major airline.

Service Enhancement: Focus on enhancing aspects of service that correlate with Korean major airline preference, such as improving the experience at specific departure airports or in certain seat classes.

Further Research: Given the recall capability of the SVM, combining insights from both the logistic regression and SVM models could yield a more comprehensive understanding. Additional variables and model tuning could potentially improve the predictions further.

Data-Driven Strategy: Incorporate these analytical insights into business strategies to optimize service offerings and improve customer satisfaction, ultimately aiming to increase the share of passengers choosing Korean major airlines.

## 3.2 Airport Models:

This section outlines all the models used for predicting Airport choice.

## 3.2.1 Airport: Logistic Regression Model

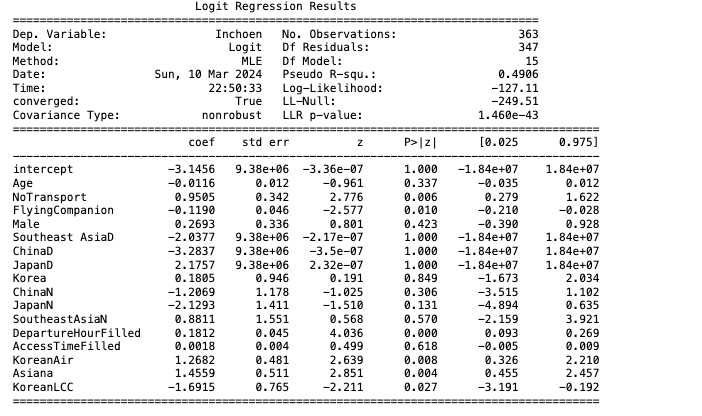
A number and numbers on a white background

Description automatically generated

Accuracy: The model correctly predicts the airport choice approximately 72.53% of the time, which is quite high.

Precision: Very high at 96.43%, indicating when the model predicts Incheon Airport, it is correct most of the time.

Recall: At approximately 52.94%, the model identifies a little over half of the actual cases of Incheon Airport choice. This could suggest a tendency to predict Gimpo Airport more often than is accurate.



In summary, the logistic regression model for airport choice appears to be a strong predictor, especially in terms of precision. The high precision with a moderate recall suggests the model is quite conservative, favoring accuracy when predicting Incheon Airport and seldom misclassifying Gimpo as Incheon. However, it also means that the model is more likely to overlook some passengers who do choose Incheon Airport.

## 3.2.2 Airport: Decision Tree Model

Accuracy: The model correctly predicts the airport choice 78.02% of the time, which indicates a high level of overall predictive power.

Precision: With a precision of 94.29%, the model is very reliable when it predicts that a passenger will choose a particular airport , Incheon. This means that it has a very low rate of false positives for this prediction.

A number and numbers on a white background

Description automatically generated

Recall: The recall rate is 64.71%, showing that the model is moderately successful at identifying all passengers who actually choose Incheon. There is some room for improvement in capturing all actual cases of Incheon being chosen.

In a nutshell, the decision tree model is quite accurate and especially precise in predicting passengers' airport choices. However, the recall indicates that it may miss a third of the passengers who do choose Incheon. This model could be particularly useful in scenarios where the cost of a false positive (incorrectly predicting Incheon) is high and less critical for missing out on some actual Incheon choosers.

## 3.2.3 Airport: Support Vector Machine

Accuracy: At approximately 75.82%, the model has a good ability to correctly identify the actual airport chosen by passengers.

Precision: It is highly precise, with about 93.94%. This indicates that when the model predicts a passenger will choose a certain airport, it is correct most of the time.

Recall: The model has a recall of around 60.78%, which means it correctly identifies a bit over half of the passengers who actually choose that specific airport

A number on a white background

Description automatically generated

In summary, the SVM model for predicting airport choices is quite accurate and precise, though the recall suggests there's some opportunity to capture more of the actual choices of the particular airport being predicted.

## 3.2.4 Airport: Neural Network

Accuracy: The model achieves an accuracy of 56.04%, indicating it correctly predicts the airport choice for slightly more than half of the passengers

Precision: At approximately 51.35%, the precision suggests that when the model predicts a specific airport choice, it is correct about half the time.

Recall: The recall of 46.34% indicates that the model identifies less than half of the actual choices for the specific airport.

A number on a white background

Description automatically generated

Overall, the neural network model shows a need for improvement across all metrics. It has the lowest accuracy, precision, and recall among the models discussed, suggesting its predictions are not as reliable.

## 3.2.5 Airport: Model Comparison and Evaluation

Logistic Regression: Accuracy: 72.53%, Precision: 96.43%, Recall: 52.94%.

Decision Tree: Accuracy: 78.02%, Precision: 94.29%, Recall: 64.71%.

SVM: Accuracy: 75.82%, Precision: 93.94%, Recall: 60.78%.

Neural Network: Accuracy: 56.04%, Precision: 51.35%, Recall: 46.34%.

The Decision Tree model outperforms the others in terms of accuracy and recall, while Logistic Regression and SVM show higher precision. The Neural Network model has the lowest scores in all three metrics.

## 3.2.6 Airport: Conclusions and Recommendations

**Conclusions:**

Robust Performance: The Decision Tree's solid performance indicates its strength in capturing the underlying decision-making patterns of passengers when choosing airports.

Influential Predictors: The significant predictors identified by the model could be leveraged for strategic enhancements in service delivery and targeted marketing efforts.

Model Insights: The Decision Tree model, due to its interpretability, provides valuable insights into the specific conditions that lead to one airport being chosen over another.

**Recommendations:**

Operational Efficiency: Utilize the Decision Tree model's findings to improve operational efficiency at airports by better aligning resources with passenger preferences and expected traffic flows.

Marketing Strategies: Implement targeted marketing campaigns that address the identified factors influencing airport choice to attract more passengers to preferred airports.

Service Enhancements: Tailor services and amenities at each airport in line with the characteristics of passengers who are more likely to choose that airport, as indicated by the model's predictors.

Continuous Improvement: Regularly update the model with new data to ensure that predictions remain accurate over time and reflect changing passenger behaviors.

The insights gleaned from the Decision Tree model provide a strategic advantage in understanding and responding to passenger preferences, which can ultimately enhance passenger satisfaction and loyalty.