**Passengers’ airport choice within multi-airport regions (MARs)**

**And multiple Airlines**

Isha Apandkar, Pravalli Buddida, Sukanya Mishra, Swati Srivastava, Vaishali Agrawal

**1. Introduction**

Within the Seoul Metropolitan Area, two prominent airports, Gimpo and Incheon, cater to the diverse travel needs of individuals. Gimpo Airport, situated closer to the city centre, contrasts with Incheon Airport, a larger and newer hub positioned farther away. Understanding the factors influencing passengers' choice of airport and airline within this context is essential for optimizing operational strategies and enhancing service offerings.

This project endeavours to investigate the underlying drivers of passengers' choice behaviour regarding airport and airline selection within the Seoul Metropolitan Area. By analysing survey data collected from 488 respondents, comprising a comprehensive set of variables ranging from socio-demographic characteristics to alternative-specific attributes, the project aims to develop robust models that delineate the decision-making processes of air travellers.

The primary objective of this study is twofold: first, to elucidate the determinants shaping passengers' airport and airline preferences, and second, to employ various modelling techniques, including discrete choice models and data mining approaches, to unravel the complexities of air travel behaviour. Through meticulous model implementation and evaluation, this research endeavours to provide actionable insights into passenger decision-making dynamics, thereby informing strategic policy formulations and operational enhancements within the air travel industry. In this introduction, we delineate

the scope and significance of the project, outlining its overarching objectives and methodological approaches. Subsequent sections will delve into the empirical analysis, model development, and policy implications derived from the study's findings.

A group of airplanes on a runway

Description automatically generated

A map of a city

Description automatically generated

**2. Exploratory Data Analysis (Data)**

**2.1 Summary information of variables**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | **TripDuration** | **FlyingCompanion** | **DepartureHr** | **Airfare** | **AccessTime** |
| **count** | 487 | 488 | 488 | 454 | 333 | 391 |
| **mean** | 39.97 | 27.44 | 2.82 | 15.98 | 50.46 | 51.83 |
| **std** | 13.67 | 74.99 | 4.00 | 4.013 | 28.98 | 43.49 |
| **min** | 17 | 0 | 0 | 1 | 3 | 4 |
| **0.25** | 29 | 4 | 1 | 13 | 35 | 25 |
| **0.5** | 38 | 5 | 2 | 16 | 45 | 40 |
| **0.75** | 50 | 8 | 3 | 19 | 60 | 60 |
| **max** | 80 | 730 | 34 | 25 | 260 | 390 |

**2.2 Choice probability of some variables to help categorize for Airport model. (Values in %)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Airline | | | | Destination | | | Gender | | Departure Time | | | |
| Airport | Korean | Asiana | Korean LCC | Foreign | China | Japan | Southeast Asia | Male | Female | Morning | Afternoon | Evening | Night |
| Incheon | 46.5 | 50.0 | 70.4 | 29.1 | 33.3 | 15.8 | 86.7 | 48.4 | 45.2 | 23.3 | 32.3 | 61.0 | 96.0 |
| Gimpo | 53.5 | 50.0 | 29.6 | 70.9 | 66.7 | 84.2 | 13.3 | 51.6 | 54.8 | 76.7 | 67.7 | 39.0 | 4.0 |

**2.3 Choice probability of some variables to help categorize for Airline model. (Values in %)**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Airport | | Destination | | | Gender | | Departure Time | | | |
| Airline | Incheon | Gimpo | China | Japan | Southeast Asia | Male | Female | Morning | Afternoon | Evening | Night |
| Korean | 58.2 | 41.6 | 40.9 | 36.1 | 68.7 | 46.5 | 51.9 | 79.1 | 51.7 | 40.1 | 48.0 |
| Foreign | 41.8 | 58.4 | 59.1 | 63.9 | 31.3 | 53.5 | 48.1 | 20.9 | 48.3 | 59.9 | 52.0 |

The **summary statistics** in table 2.1s provides a comprehensive overview of several key variables within the dataset. Age observations total 487, with a mean age of approximately 40 years and a standard deviation of 13.67. Trip duration spans from 0 to 730, with an average duration of 27.44 days and a notable standard deviation of 74.99. Flying companion counts vary widely, with an average of 2.82 and a maximum count of 34. Departure hours range from 1 to 25, with a mean of 15.98 and a standard deviation of 4.013. Airfare values range from 3 to 260, with a mean of 50.46. Access times range from 4 to 390, with a mean of 51.83. These statistics offer valuable insights into the distribution and characteristics of each variable, aiding in the understanding of the dataset's underlying patterns and trends.

The **choice probabilities** in table 2.2 for various variables provide essential insights to aid in categorizing the Airport model effectively. In terms of airlines, Incheon Airport appears to attract a lower percentage of Korean carriers (46.5%) compared to Gimpo Airport (53.5%), indicating a slight preference for Gimpo among Korean airlines. The scenario is opposite for Korean LCC. Conversely, Gimpo Airport shows a higher percentage of foreign airlines (70.9%) compared to Incheon (29.1%), suggesting a stronger preference for Gimpo among

foreign carriers.

A graph of different colored bars

Description automatically generated

Regarding destination, both airports serve a diverse range of locations, with Gimpo Airport exhibiting higher percentages for China (66.7%), Japan (84.2%), and Southeast Asia (13.3%) compared to Incheon. Gender distribution shows a relatively balanced proportion of male and female passengers at both airports, with slight variations in percentages. Analysis of departure times reveals interesting patterns, with Incheon Airport experiencing higher activity during the evening (61.0%) and night (96.0%), while Gimpo Airport sees more flights in the morning (76.7%).

The choice probabilities in table 2.3 for various variables provide valuable insights to aid in categorizing the Airline model effectively. When considering the airport, Korean airlines demonstrate a higher preference for Incheon (58.2%) over Gimpo (41.6%), whereas foreign airlines show a higher preference for Gimpo (58.4%) compared to Incheon (41.8%).

A graph with green and blue squares

Description automatically generated

Korean airlines have a relatively balanced distribution across different regions, with slightly higher percentages for Southeast Asia (68.7%). In contrast, foreign airlines exhibit higher percentages for China (59.1%) and Japan (63.9%) compared to Southeast Asia (31.3%). Gender distribution shows that both Korean and foreign airlines serve a similar proportion of male and female passengers, with minor variations in percentages. Analysis of departure times reveals interesting patterns, with Korean airlines experiencing higher activity during the morning (79.1%) and foreign airlines showing a preference for evening flights (59.9%). These choice probabilities offer valuable insights into passenger preferences and behaviours, facilitating effective categorization within the Airline model.

**2.4 Removal of Variables**

A screenshot of a computer

Description automatically generated

Mileage Airline (217), Departure Minute (110), Access Cost (187) were not considered for modelling due to large number of missing values. Since ID and Flight Number will not add any value to the model, these variables were removed.

**2.5 Correlation Matrix**

As you can see below, we do not have any variables which are highly correlated.

A screenshot of a computer screen

Description automatically generated

**2.6 ANOVA Test**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **F-Statistic (Airline)** | **P-Value (Airline)** | **F-Statistic (Airport)** | **P-Value (Airport)** |
| **Age** | 1.0913 | 0.2972 | 6.8242 | 0.0093 |
| **Gender** | 0.5308 | 0.4670 | 0.4497 | 0.5028 |
| **Nationality** | 0.2512 | 0.6167 | 3.5395 | 0.0606 |
| **Trip Purpose** | 0.0165 | 0.8979 | 11.7041 | 0.0007 |
| **Trip Duration** | 3.4147 | 0.0658 | 0.9996 | 0.3179 |
| **Flying Companion** | 6.8263 | 0.0095 | 1.2320 | 0.2676 |
| **Province Residence** | 0.3053 | 0.5811 | 15.6895 | <0.0001 |
| **Group Travel** | 1.3115 | 0.2532 | 0.0122 | 0.9122 |
| **Number ofTrips Last Year** | 0.1978 | 0.6569 | 1.8800 | 0.1710 |
| **Frequent Flight Destination** | 0.0212 | 0.8844 | 29.1606 | <0.0001 |
| **Destination** | 0.0414 | 0.8390 | 121.4023 | <0.0001 |
| **Departure Hr** | 7.3071 | 0.0073 | 6.7098 | 0.0099 |
| **Departure Time** | 26.2542 | 0.0000 | 70.3310 | <0.0001 |
| **Seat Class** | 0.7919 | 0.3744 | 0.1291 | 0.7196 |
| **Airfare** | 1.6009 | 0.2070 | 3.5485 | 0.0602 |
| **Number of Transport** | 0.0044 | 0.9473 | 40.2543 | <0.0001 |
| **Mode of Transport** | 0.9071 | 0.3418 | 6.6028 | 0.0105 |
| **Access Time** | 0.5715 | 0.4504 | 0.9161 | 0.3390 |
| **Occupation** | 0.4351 | 0.5101 | 21.7961 | <0.0001 |
| **Income** | 4.0701 | 0.0447 | 0.6280 | 0.4285 |

The ANOVA tests conducted provide valuable insights into the relationships between various variables and the factors of Airline and Airport.

**3. Models and Results**

We developed and evaluated several machine learning models to address passengers' choice behaviour regarding airport and airline selection within the Seoul Metropolitan Area. Our goal was to explore the performance and characteristics of various algorithms in solving this task efficiently and accurately. The models we focused on include Logistic Regression, Decision Tree, Neural Network (NN), and Support Vector Machine (SVM). Each of these models offers unique strengths and capabilities, making them suitable for different types of problems and data. In this study, we will compare and evaluate the performance of these machine learning models on our dataset, considering factors such as accuracy, interpretability, computational efficiency, and robustness to noise and outliers. By understanding the strengths and weaknesses of each model, we aim to identify the most suitable approach for our specific problem domain and task requirements. These models are:

**3.1 Logistic Model (Airport)**

For categorization, we calculated the choice probability of each variable to determine the best category. Once the model was constructed, we focused on optimizing the significance of the variables. Categories that did not contribute significantly to the model, such as Nationality and Mode of Transport, were removed. Additionally, we categorized some continuous variables like Trip Duration, Departure Hr, Number of Trips Last year, Access Time to enhance their significance and reduce the p-value. Below is the final model with accuracy of 80.4%.

A screenshot of a computer

Description automatically generated

The training accuracy of 93.1% suggests that the model correctly predicts the choice of Incheon Airport for approximately 93.1% of the training data instances. Similarly, the testing accuracy of 80.4% indicates that the model correctly predicts the choice of Incheon Airport for approximately 80.4% of the testing data instances. The precision of 95.3% in the training set means that among all instances predicted as choosing Incheon Airport, 95.3% chose it. In the testing set, the precision is slightly lower at 90.2%. The recall of 88.88% in the training set implies that the model correctly identifies 88.88% of all instances where Incheon Airport was chosen. In the testing set, the recall is lower at 72.5%. Low AIC BIC for this model indicated that the model is a good fit.

A screenshot of a data

Description automatically generated

This log odds of selecting Incheon Airport decreases by 3.86 when all other variables are zero or not applicable. Passengers are more likely to choose Incheon Airport for Korean airlines flights (Korean and Korean LCC). The log odds of choosing Incheon airport increases by 2.3 if a passenger is flying by Korean airlines as compared to passenger flying by foreign/other airlines keeping other variables constant. For each additional flying companion, the log odds of selecting Incheon Airport decreases by 0.1337. Passengers are less likely to choose Incheon Airport when the access time is 40 minutes or less, as opposed to when it exceeds 40 minutes.

If the trip duration is less than or equal to 7 days, the log odds of selecting Incheon Airport over Gimpo Airport decreases by 1.28 as compared to passenger with trip duration greater than 7 days. The log odds of choosing Incheon airport increases by 1.4 if a passenger has departure hour of <=15 as compared to passenger having departure hour of >15 keeping other variables constant. If traveling in a group, the log odds of selecting Incheon Airport increases by 0.735 on average as compared to travelling alone. The log odds of selecting Incheon airport increases by 1.39 if a passenger is travelling for leisure as compared to a passenger travelling for business, study, or other purpose.

The log odds of selecting Incheon airport decreases by 1.28 if passenger resides in Seoul or other as compared to passengers residing in Incheon, Kyungki, Chungcheong or Kangwon. The log odds of selecting Incheon are positive for passengers travelling in afternoon, evening or night -3.2,7.4 and 10.3 respectively as compared to passengers travelling in the morning. If airfare increases by 1 unit then log odds of choosing Incheon airport increases by 0.017 on average. The log odds of selecting Incheon airport decreases by 1.64 if passenger has frequent flight destination of southeast Asia as compared to passenger having frequent flight destination of Europe or none.

Passenger with frequent flight destination of China, other, Japan, North and South America have less chance of selecting Incheon airport with reference to passenger with frequent flight destination of Europe or none. The log odds of selecting Incheon increases by 3.39 if passenger is traveling to Southeast Asia as compared to passenger with destination of Japan. The log odds of selecting Incheon increases by 1.9 if passenger is traveling to China as compared to passenger with destination of Japan. The log odds of selecting Incheon decreases by 2.6 when passenger has occupation of entrepreneur, senior management, student, housewife, retired, none as compared to passenger having occupation of sales, service, other.

The log odds of selecting Incheon decreases by 2.34 when passenger has occupation of Business (Corporate worker), Government, Military, Professionals (doctor, lawyer, professor), industrial, manufacturing, self-employed compared to passenger having occupation of sales, service, other. The log odds of selecting Incheon increases by 2.27 when passenger has income class of 50~80 Million Won as compared to passenger having income class of 30 Million Won or less. The log odds of selecting Incheon decreases by 2.35 when passenger has income class of 30~50 Million Won or 80 ~100 Million Won or 100 ~150 Million Won or 150 ~200 Million Won or 200 Million Won or more as compared to passenger having income class of 30 Million Won or less.

Most of the variables are highly significant for the choice of airport. Although airfare and group travel are insignificant as they have high p value.

**3.2 Logistic Model (Airlines)**

A screenshot of a computer

Description automatically generated

The test accuracy is approximately 67.39%, indicating that about 67.39% of the instances are correctly classified by the model. The precision for Korean Airline (class 1) is approximately 62.96%, indicating that about 62.96% of the instances predicted as Korean Airline belong to Korean Airline.  The recall for Korean Airline (class 1) is approximately 77.27%, indicating that about 77.27% of the actual Korean Airline instances are correctly identified by the model.

The accuracy and precision for the logistic regression model of Airlines is not very high. Also, the Pseudo R-squared value which gives the goodness of fit of the model is very less with 25%.  Training Accuracy (74.5%) is relatively good compared to Testing Accuracy (67%). This could be due to overfitting, discrepancies between training and testing data distributions, data preprocessing issues, or imbalances in the dataset.

Additionally, the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) provide measures of the relative quality of statistical models for a given set of data. Lower AIC and BIC values generally indicate better model fit. In this case, the AIC is approximately 412.15 and the BIC is approximately 478.41. Hence, alternative models like Decision Tree can be considered for better description of this model.

A screenshot of a computer

Description automatically generated

Categories that did not contribute significantly to the model, such as Age, Gender, Nationality, TripPurpose, ProvinceResidence, GroupTravel, NoTripsLastYear, NoTransport, ModeTransport are removed from the model.  Additionally, continuos variables such as TripDuration, AccessTime, DepartureHr are also categorised to enhance their significance and reduce the p-value.

Intercept(2.9259) represents the baseline log odds of selecting Korean Airlines over Other Airlines when all other variables are zero or not applicable.  For each passenger choosing the Incheon Airport, the log odds of the passenger choosing (Korean Airline) increases by 1.3205 on average compared to a passenger choosing Gimpo Airport, holding all other variables constant.  For each unit increase in the Airfare, the log odds of the passenger choosing Korean Airline changes by 0.0114 on average, holding all other variables constant.  For each unit increase in the Trip Duration (up to 5 days), the log odds of the passenger choosing Korean Airline changes by 0.6570 on average compared to a trip duration of greater than 5days, holding all other variables constant.  For each minute increase in the Access time (up to 40 minutes), the log odds of the passenger choosing Korean Airline changes by -0.7619 on average compared to an access time of greater than 40 minutes, holding all other variables constant. For each additional Flying Companion, the log odds of the passenger choosing Korean Airline changes by 0.0779 on average, holding all other variables constant. For each hour increase in the Departure hour (<= 15 hours), the log odds of the passenger choosing Korean Airline changes by -1.0207 on average compared to departure hours greater than 15 hours, holding all other variables constant.  For each instance of a Night departure, the log odds of the passenger choosing Korean Airline changes by -3.8564 on average compared to morning departures, holding all other variables constant.  For each instance of an Afternoon or Evening departure, the log odds of the passenger choosing Korean Airline changes by -1.7627 and -3.6268, respectively, on average compared to morning departures, holding all other variables constant.

For each instance of the respective frequent flight destination Southeast Asia, Japan, the log odds of the passenger choosing Korean Airline changes by -1.3043, and -0.7298 respectively, on average compared to other destinations, holding all other variables constant.  For each instance of the respective destination China and Japan, the log odds of the passenger choosing Korean Airline changes by -1.9193, and -1.6544, respectively, on average compared to destination of SE Asia, holding all other variables constant. For each instance of choosing Economy class, the log odds of the passenger choosing Korean Airline changes by 1.2005 on average compared to Business and First Class, holding all other variables constant.  For each instance of having a business occupation, the log odds of the passenger choosing Korean Airline changes by -0.7694 on average compared to other occupations, holding all other variables constant.  For each instance of belonging to the income group of 30-80 million Won, the log odds of the passenger choosing Korean Airline changes by -0.8422 on average, holding all other variables constant.

**Positive and Negative relations with dependent variable:**

While AccessTime, DepartureHr, DepartureTime, FrequentFlightDestination, Destination, Occupation(Business) and Income(30-80 Million Won) show a negative relation with Korean Airlines, Airport, Airfare, TripDuration (up to 5 days), FlyingCompanion, SeatClass (Economy) show a positive relation with Korean Airlines.

**Interpretation of p-values:**

In the above model, Night, Evening and Destination(China) are highly significant at 0.1% level of significance.  Airport(Incheon), AccessTime<=40, DepartureHr1, Afternoon, FrequentDestination(SEAsia), Destination(Japan), IncomeGroup(30-80) are significant at 1% level of significance.  TripDuration<=5, FlyingCompanion, SeatClass(Economy), Occupation(Business) are all significant at 5% level of significance.

**3.3 Decision Tree Model (Airport)**

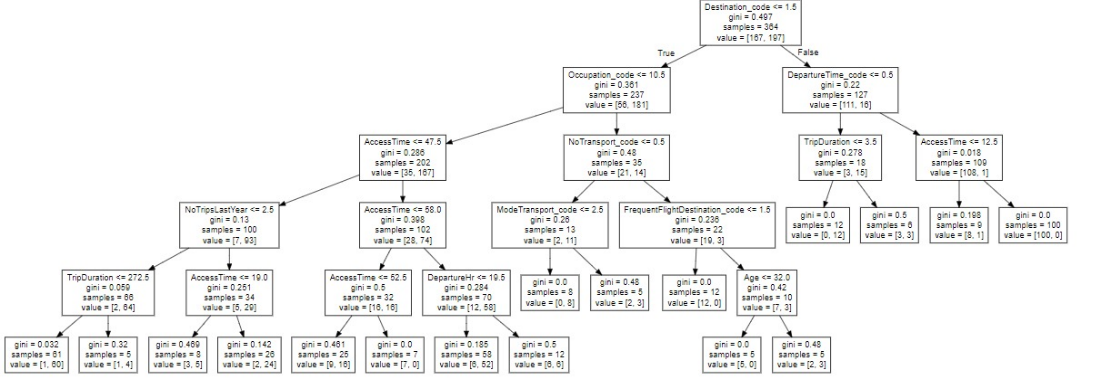
Decision tree builds a tree-like structure, partitioning the feature space to maximize homogeneity in airport and airline selection. Each partition is chosen to maximize the homogeneity of the target variable within the subset. The decision nodes represent features, while the leaf nodes represent the predicted class. The tree uses various features (variables) to make decisions at each node, where a decision leads to either another decision (internal node) or a final classification (leaf node).

The dataset used for training and testing the model includes various demographic and trip-related features such as gender, nationality, trip purpose, departure time, airfare, etc. Data is split into training and testing sets (80% training, 20% testing).

A screenshot of a computer

Description automatically generated

These results indicate that the model performs well, with high accuracy and balanced precision and recall, both during training and testing.  Training Accuracy (90%) is relatively good compared to Testing Accuracy (85%). This could be due to overfitting, discrepancies between training and testing data distributions, data preprocessing issues, or imbalances in the dataset. Overall, the model effectively predicts airport choices based on the given features.



From the Tree structure, it is evident that first decision is based on  Destination. This indicates that the primary distinction in classifying the airports is made based on the destination.  The occupation of an individual is the next significant variable with threshold <= 10.5. This implies that certain professions prefer a specific airport, perhaps due to the proximity of the airport to their workplace or the types of flights (domestic or international) they typically take.

The features or attributes that appear closer to the root of the tree such as Destination(<=1.5), Occupation(<=10.5), DepartureTime(<=0.5), NoTransport(<=0.5), TripDuration(<=3.5), AccessTime(<=12.5) are considered more significant. This is because they are used to make the first splits, which have a significant impact on the resulting branches of the tree. The features like Age(<=32), TripDuration(<=272.5), AccessTime(<=19), DepartureHr(<=19.5) can be considered less significant.

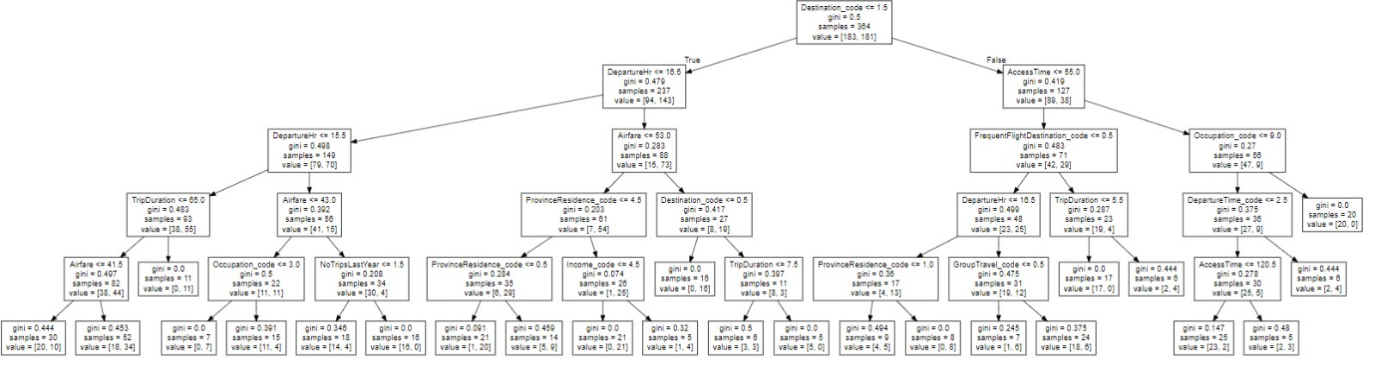
Each node represents a decision point where the tree splits based on the value of a variable, leading to branches that either lead to another decision node or to a leaf node representing a classification.  Leaf Nodes represent the final decision made by the tree, with a class label (Incheon or Gimpo). The  Gini value indicates the purity of the node, with 0 being completely pure.

**3.4 Decision Tree Model (Airlines)**

A screenshot of a computer

Description automatically generated

Overall, the model performs decently on both the training and testing sets, with slightly higher performance on the training set compared to the testing set. The precision and recall values are relatively balanced, indicating a fair ability to correctly predict both positive and negative cases.



From the Tree structure, it is evident that first decision is based on  Destination. This indicates that the primary distinction in classifying the airlines is made based on the destination (Korea).  The DepartureHr of an individual is the next significant variable with threshold <= 16.5. This implies that passengers with those departure hours prefer a specific airline.

The features or attributes that appear closer to the root of the tree such as Destination(<=1.5), DepartureHr(16.5), AccessTime(<=55) , Airfare(<=53), FrequentFlighDestination(<=0.5), Occupation(<=9) are considered more significant. This is because they are used to make the first splits, which have a significant impact on the resulting branches of the tree. The features like NoTRipsLastYear(<=1.5), ProvinceResidence(<=0.5), Income(<=4.5), GroupTravel(<=0.5) can be considered less significant.

**3.5 Neural Network Model  (Airport)**

Neural networks are adept at capturing complex patterns and relationships within data, which is crucial in understanding the diverse and multifaceted decision-making processes of passengers when it comes to airport selection. Traditional statistical models may struggle to capture non-linear relationships between predictors and outcomes. Neural networks can handle non-linearities effectively, making them suitable for modelling complex decision behaviours. Neural networks can automatically learn interactions between features, allowing for more nuanced insights into how different factors collectively influence airport choice.

Here, the dataset includes features such as age, flying companion, access time, gender, airline, nationality, trip purpose, etc., which are used to predict the airport choice.

Data is split into training and testing sets (80% training, 20% testing).

The model's performance is evaluated on both the training and testing datasets.

Training Confusion Matrix:

predicted 0 1

actual

0 139 29

1 14 182

Train Accuracy: 0.8818681318681318

Train precision: 0.8625592417061612

Train Recall: 0.9285714285714286

Testing Confusion Matrix:

predicted 0 1

actual

0 36 9

1 4 43

Test Accuracy: 0.8586956521739131

Test precision: 0.8269230769230769

Test Recall: 0.9148936170212766

**Key Findings/Results:**

Training Accuracy: 88.19% of the training data instances were correctly classified by the model, while Testing Accuracy: 85.87% of the testing data instances were correctly classified.

Precision: Indicates that out of all instances predicted as positive, 82.69% were correctly predicted for the testing set. Recall: Out of all actual positive instances, 91.49% were correctly classified for the testing set.

Confusion Matrix: In the training set, the model has correctly classified 139 instances as 0 (negative class) and 182 instances as 1 (positive class) out of actual instances. In the testing set, the model has correctly classified 36 instances as 0 and 43 instances as 1 out of actual instances.

**Model Interpretation:**

The model shows relatively good performance in both training and testing sets, with high accuracy, precision, and recall scores.

The confusion matrix indicates that the model has a slightly higher number of false negatives (9) compared to false positives (4) in the testing set.

This suggests that the model is better at identifying positive instances (1) than negative instances (0), but there is room for improvement in reducing false negatives.

**3.6 Neural Network Model (Airlines)**

Training Confusion Matrix:

predicted 0 1

actual

0 148 35

1 41 140

Train Accuracy: 0.7912087912087912

Train precision: 0.8

Train Recall: 0.7734806629834254

Testing Confusion Matrix:

predicted 0 1

actual

0 31 11

1 19 31

Test Accuracy: 0.6739130434782609

Test precision: 0.7380952380952381

Test Recall: 0.62

**Key Findings/Results:**

Training Accuracy: 79.12% of the training data instances were correctly classified by the model, while Testing Accuracy: 67.39% of the testing data instances were correctly classified.

Precision: Indicates that out of all instances predicted as positive, 73.81% were correctly predicted for the testing set. Recall: Out of all actual positive instances, 62% were correctly classified for the testing set.

Confusion Matrix: In the training set, the model has correctly classified 148 instances as 0 and 140 instances as 1 out of actual instances. In the testing set, the model has correctly classified 31 instances as 0 and 31 instances as 1 out of actual instances.

**Model Interpretation:**

The model shows a significant drop in performance from the training set to the testing set, indicating potential overfitting or generalization issues.

The confusion matrix reveals a higher number of false positives (11) compared to false negatives (19) in the testing set, suggesting that the model tends to misclassify negative instances more often.

There's a notable gap between training and testing accuracies, indicating that the model might not generalize well to unseen data.

Further optimization or regularization techniques may be necessary to improve the model's performance and generalization capabilities.

Overall, the neural network model provides insights into the factors influencing passengers' airline choices within the Seoul Metropolitan Area, contributing to the project's objectives of understanding passenger decision-making dynamics and informing strategic policy formulations within the air travel industry.

**3.7 SVM  (Airport)**

The Support Vector Machine (SVM) model excels in classifying complex datasets, offering invaluable insights for predicting passenger preferences for airports and airlines based on a myriad of variables, including age, flying companion, access time, gender, and trip purpose. A notable feature of our SVM models for both airport and airline preference prediction are the utilization of a polynomial kernel with a degree of 3. This specific configuration, kernel=poly, degree=3, is chosen for its enhanced performance in capturing the intricate, non-linear relationships among the variables, leading to better accuracy in our predictions.

The polynomial kernel allows the SVM to not only consider the linear interactions between variables but also to model the more complex, higher-order relationships that are pivotal in understanding passenger behavior. This capability is particularly useful in our context, where the decision-making process of passengers regarding airport and airline selection is influenced by a combination of numerous factors. By adeptly navigating through the high-dimensional space created by these factors, the SVM model with a polynomial kernel becomes a powerful tool, offering precise and reliable classifications that can significantly benefit the aviation sector in optimizing their services and marketing efforts to align with passenger preferences.

Training Confusion Matrix:

predicted 0 1

actual

0 139 28

1 0 197

Train Accuracy: 0.9230769230769231

Train precision: 0.9377777777777778

Train Recall: 0.9161676646706587

Testing Confusion Matrix:

predicted 0 1

actual

0 33 13

1 2 44

Test Accuracy: 0.8369565217391305

Test precision: 0.8573934837092732

Test Recall: 0.8369565217391304

Key Findings/Results:

Training Performance: The SVM model achieved a high training accuracy of 92.31%. The precision for class 1 (airport present) is 93.78%, indicating that when the model predicts an airport is present, it is correct 93.78% of the time. The recall for class 1 is 91.62%, suggesting that the model correctly identifies 91.62% of the actual airports.

Testing Performance: The model's accuracy on the testing dataset is 83.70%, which is slightly lower than the training accuracy but still relatively high. The precision for class 1 is 85.74%, indicating that the model correctly identifies 85.74% of the airports in the testing dataset. The recall for class 1 is 83.70%, suggesting that the model correctly identifies 83.70% of the actual airports in the testing dataset.

Confusion Matrix Analysis: The confusion matrix shows that the model performs well in predicting both classes, with relatively low false positives and false negatives. There are 28 false positives (predicted as airport but not actual airports) and 2 false negatives (actual airports predicted as not present) in the testing dataset.

Overall Interpretation: The SVM model for airport classification demonstrates strong performance both in training and testing, with high accuracy, precision, and recall. It effectively distinguishes between airports and non-airport locations.

**3.8 SVM (Airlines)**

Training Confusion Matrix:

predicted 0 1

actual

0 152 31

1 13 168

Train Accuracy: 0.8791208791208791

Train precision: 0.8827166133698797

Train Recall: 0.8793889442381426

Testing Confusion Matrix:

predicted 0 1

actual

0 26 16

1 13 37

Test Accuracy: 0.6847826086956522

Test precision: 0.6823899371069182

Test Recall: 0.6795238095238095

Key Findings/Results:

Training Performance: The SVM model achieved a training accuracy of 87.91%. The precision for class 1 (airline present) is 88.27%, indicating that when the model predicts an airline is present, it is correct 88.27% of the time. The recall for class 1 is 87.94%, suggesting that the model correctly identifies 87.94% of the actual airlines.

Testing Performance: The model's accuracy on the testing dataset is 68.48%. The precision for class 1 is 68.24%, indicating that the model correctly identifies 68.24% of the airlines in the testing dataset. The recall for class 1 is 67.95%, suggesting that the model correctly identifies 67.95% of the actual airlines in the testing dataset.

Confusion Matrix Analysis: The confusion matrix shows that the model has higher false positives and false negatives compared to the airport model.

There are 16 false positives (predicted as airline but not actual airlines) and 13 false negatives (actual airlines predicted as not present) in the testing dataset.

Overall Interpretation: The SVM model for airline classification demonstrates reasonable performance but is not as strong as the airport model.

While the model achieves decent precision and recall, it shows lower accuracy on the testing dataset compared to the airport model. The presence of higher false positives and false negatives indicates some limitations in the model's ability to accurately classify airlines.

In summary, the SVM model for airport classification performs significantly better than the model for airline classification, demonstrating higher accuracy, precision, and recall in both training and testing datasets. This suggests that the features used for airport classification might be more discriminative and easier to learn for the model compared to those used for airline classification.

**4. Conclusion**

In this study, we embarked on an analysis to understand and predict passenger choices regarding airports and airlines within the Seoul Metropolitan Area, aiming to optimize operational strategies and enhance service offerings in the aviation sector. By analysing survey data from 488 respondents and employing a variety of modelling techniques including Logistic Regression, Decision Trees, Neural Networks, and Support Vector Machines, we sought to capture the intricate decision-making processes of air travellers.

**Logistic Regression vs. Decision Tree for Airport and Airline Model**

Logistic Regression provides a probabilistic approach, offering clear interpretations of how each variable influences the likelihood of choosing a particular airport or airline. In the case of the airport model, logistic regression shows high accuracy, particularly in identifying factors that influence the choice of Incheon over Gimpo. Variables like airline type, flying companions, trip duration, and others are significant predictors. For the airline model, logistic regression identifies factors such as airport choice, airfare, and trip duration, but with a noted lower accuracy and precision compared to the airport model.

Decision Tree offers a more intuitive, visual representation of decision-making processes, which can be particularly useful for identifying critical decision points and the hierarchy of variable importance. Both the airport and airline models demonstrated the decision tree's effectiveness in partitioning the dataset based on key variables, leading to high accuracy in predictions. Decision trees are more flexible in handling non-linear relationships and can easily capture complex interactions between variables without needing extensive data preprocessing.

**Comparison and Suitability**

* **Interpretability**: Logistic regression provides coefficients that directly measure the impact of each predictor on the outcome, making it highly interpretable for policy formulation and strategic decisions. Decision trees offer a visual interpretation, showing the decision path clearly but might lack the direct quantitative impact measure provided by logistic regression.
* **Model Performance and Complexity**: Decision trees tend to outperform logistic regression when relationships between variables are non-linear or complex. They can also handle variable interactions more intuitively. However, logistic regression can perform better when relationships are approximately linear or when the primary interest lies in understanding the effect size of predictors.
* **Overfitting and Generalizability:** Decision trees are more prone to overfitting, especially with very complex trees. Logistic regression is generally more robust to overfitting with fewer variables but might underperform if key interactions or non-linear relationships are missed.

**Final Reflections:**

While logistic regression offers a more balanced approach, it exhibits lower accuracy and recall compared to the Decision Tree model. Although logistic regression may yield higher precision, in the context of the airline sector, precision might not hold utmost importance. Moreover, Decision Tree modeling has demonstrated comparable results to those of the SVM model. Consequently, the **Decision Tree model** emerges as a better choice for representing both airline and airport models in our project.

As we conclude, it becomes evident that for predicting passenger choices regarding airports and airlines, the Decision Tree model outperforms Logistic Regression owing to its adaptability, adept handling of non-linear relationships, and the clear, interpretable paths it provides. This makes Decision Trees especially conducive for strategic and operational planning in the aviation industry, where comprehending the hierarchy of influencing factors is key.