SEATTLE UNIVERSITY

BUAN 5310 19SQ Statistical Learning

AIRPORT & AIRLINE CHOICE ANALYSIS

Ankita Pathak

Nancy Jain

Sanyukta Ghai

Sourabh Gupta

Table of Contents

1. Introduction

According to International Air Transport Association (IATA) there is more than 7% increase in Air Travel compared to last year. The demand for passenger air transportation is the driving force for business decisions in airline industry and it also drives travel choices in multi airport regions. In a multiple airport region, passengers must decide on both which airport and which airline to use. Competition between (origin) airports for passengers cannot be analyzed without considering the airlines' reactions to the airports. Hence, it is important to analyze what drives air travel among travelers and passengers and what factors they take into consideration while choosing airline and airport. In this paper, we conduct a detailed analysis of the choice of airport and airline for resident travelers in the Seoul Metropolitan Area.

We have created airport choice models between two airports in Seoul. One is Gimpo airport which is smaller and old but closer to the city and another is Incheon airport which is larger and new hub airport but away from the city. Furthermore, using discrete choice and data mining models, we analyzed six factors more in-depth. Similarly, for airline choice models, we have classified four airlines - Korean Air (KE), Asiana Air (OZ), Korean LCC, Foreign Carriers - into two types, International and Domestic for logistic and data mining models and kept it as it is for multinomial model. International airline includes Foreign Carriers and Domestic airline includes Korean Air (KE), Asiana Air (OZ) and Korean LCC.

We analyzed raw data collected from completed questionnaires consisting of 24 questions from 488 people and used the discrete choice data as input for a discrete and data mining models. The results of exploratory analysis on the data show that the passenger's gender is not a significant factor in explaining airlines' choices. Nationality and age, on the other hand, is one of the most important factors. This shows that, when trying to influence airlines in their airport choice, airports and policy makers also must consider the passenger's demographics.

The remainder of this paper is organized as follows. In the next section, we present a brief review of data used in the model and its exploratory analysis. The third section describes the airport and airline models and their key findings, methodology and model specifications. Finally, last section discusses model validation, and summarizes the findings of the research and provide some recommendations.

2. Data

- Summary of our EDA results
  - Missing values: There were many missing values in the survey data:
    - **DepartureMinute** - flight departure time had 120 missing values
    - **Airfare** – Price of the ticket had 155 missing values
    - **AccessCost** – Cost from residence to Airport had 197 missing values
    - **AccessTime** – Commute time from residence to Airport had 97 missing values

- **Income** – Earing of the survey participant had 132 missing values
- **Mileage** – Air miles from airlines for frequent travel customer had398 missing values

- Replacement for missing values: Various methods could be used to treat these missing values like replacing quantitative variables with mean or median value of the parameter, or replacing the values with the parameter's mode value, or replacing it with the data given in Airport_Airline_data (raw data) file.
  - **Age** – This variable had only one missing value. As age is normally distributed, we filled this missing record with the mean value of this variable.
  - **AirFare** – As already mention, AirFare had 155 missing values. We mapped those missing values with "AveragePrice" variable given in the "Price Info" sheet in "*Airport_Airline_Data*" file. After this we were left with only 6 records with missing values, which we removed from the data. "AveragePrice" consists of average airfare by Airport, Airline, Destination and seat class.
  - **AccessTime** – There were 97 missing values for this parameter. We filled all the missing values using "TravelTime,minutes" variable given in the "Airport Province Distance" sheet in "*Airport_Airline_Data*" file.
  - **Airlines** – Airlines had 10 missing values. Since, we did not have any information to fill this column, we removed those records.
  - **Gender** – Gender had 3 missing values. Instead of filling them with mode value of the category, we replaced them with "Not Mentioned".
  - **Destination** – Destination had 5 missing values, we removed those records since, we cannot get this information from anywhere else
  - **Seat Class** – SeatClass had 4 missing values. Since, we are not using this variable in any model we did not treat the Null values here.
  - **Access Cost** – AccessCost had 197 missing values. Since almost 40% of the data is missing, we cannot use mean or any other central tendency. Also, we have no other information on this parameter, hence, we did not consider this in any of our model.
  - **Mileage** – Mileage had 398 missing values. We don't have any other data which can be used to fill/treat these missing records. Also, we cannot use mean or other central tendency value because 398 are almost 80% of records. So, we are not using this parameter for our study.

- Removing Unnecessary parameters
  - **ID** – Unique Identification number for each record in survey data. This is not required for our study.
  - **Gender** – Both genders (Male and Female) has almost same statistics in terms of Airlines and Airport, so this parameter is not being used for our study.
  - **Trip Purpose** – Most of the surveyed people have either Leisure (66%) or Business (21%) as their trip purpose, so this is not being used for this study.
  - **Trip Duration** – No of days the trip lasted. This is not relevant for this study

- **Flying Companion** – This parameter describes "Who the participant is travelling with?". This parameter is not relevant for our study.
- **Group Travel** – This parameter describes "If the travel was in group or not". It doesn't give us any information regarding airline or airport choice. So, this parameter is removed from the final dataset.
- **No of trips last year** – This is not relevant for the study of airline and airport choice.
- **Frequent Flight Destination** – Participant's information regarding frequent destination can be used to give offers on selected route but we did not find it relevant in choosing airport and airline. Hence, we removed it from the final dataset.
- **Flight No** – This parameter could be useful to fill some of the missing values, but this is missing for most of the records.
- **Departure Hr, Mn** – This variable is correlated with DepartureTime, which we already took into consideration. Hence, we did not consider this parameter for our study.
- **Seat Class** – This could be a relevant parameter for airline selection but as per the sample most people are using economy as seat class. We did not consider this in our study.
- **No of transport, Mode of Transport** – This is the number and mode of transports used to commute from residence to airport. This is not useful for this study, but could be useful to study public transportation in city.
- **Access Cost** – This gives us the information about money spent on commute from residence to airport. This could be a good parameter in airport selection; however, it has many missing values, hence, we removed it from final dataset.
- **Access Time** – Commute time from residence to airport. This parameter is highly correlated with province and cost, so removed from the dataset as we are taking province as a parameter.
- **Occupation and Income** – These parameters are not relevant for this study. So, we removed them from the dataset.
- **Mileage Airline and Mileage** – These parameters could be helpful in deciding airline but 80% of records have missing values for this column. So, we removed it from the dataset.

- Descriptive Statistics of the final dataset

Skim summary statistics
n obs: 466
n variables: 8

Variable type: factor

variable	missing	complete	n	n_unique	top_counts	ordered
Airline	0	466	466	4	Kor: 148, For: 133, Asi: 105, Kor: 80	FALSE
Airport	0	466	466	2	GMP: 241, ICN: 225, NA: 0	FALSE
DepartureTime	0	466	466	4	12p: 203, 6pm: 191, 6am: 43, 9pm: 29	FALSE
Destination	0	466	466	4	Sou: 160, Jap: 154, Chi: 131, Oth: 21	FALSE
Nationality	0	466	466	5	Kor: 361, Jap: 39, Chi: 32, Sou: 20	FALSE
Province	0	466	466	8	Se: 181, Ky: 122, Je: 86, Ky: 27	FALSE

Variable type: numeric

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100	hist
Age	0	466	466	39.71	13.59	17	29	37	50	80	
Airfare	0	466	466	49.24	24.52	3	36.22	46.83	56	260	

3. Model and Results

3.1. Model 1 – Logistic Regression

Model description

Logistic regression is a classification algorithm that is used to predict the probability of a categorical dependent variable. In Airport and Airline models, the dependent variable is a binary variable that contains data coded as 1 meaning a yes or success Or a 0 meaning a no or failure.

Logistic Regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (our features), by estimating probabilities using its underlying logistic function.
These probabilities must then be transformed into binary values in order to actually make a prediction. This is the task of the logistic function, also called the sigmoid function.

Strengths:

- Outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid overfitting.
- Logistic models can be updated easily with new data using stochastic gradient descent.
- Logistic models are highly interpretable, they do not require input features to be scaled
- Logistic models do not require any tuning, it is easy to regularize, and they outputs well-calibrated predicted probabilities.

Weaknesses:

- Logistic regression tends to underperform when there are multiple or non-linear decision boundaries.
- They are not flexible enough to naturally capture more complex relationships.

Why use this model?

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variable.

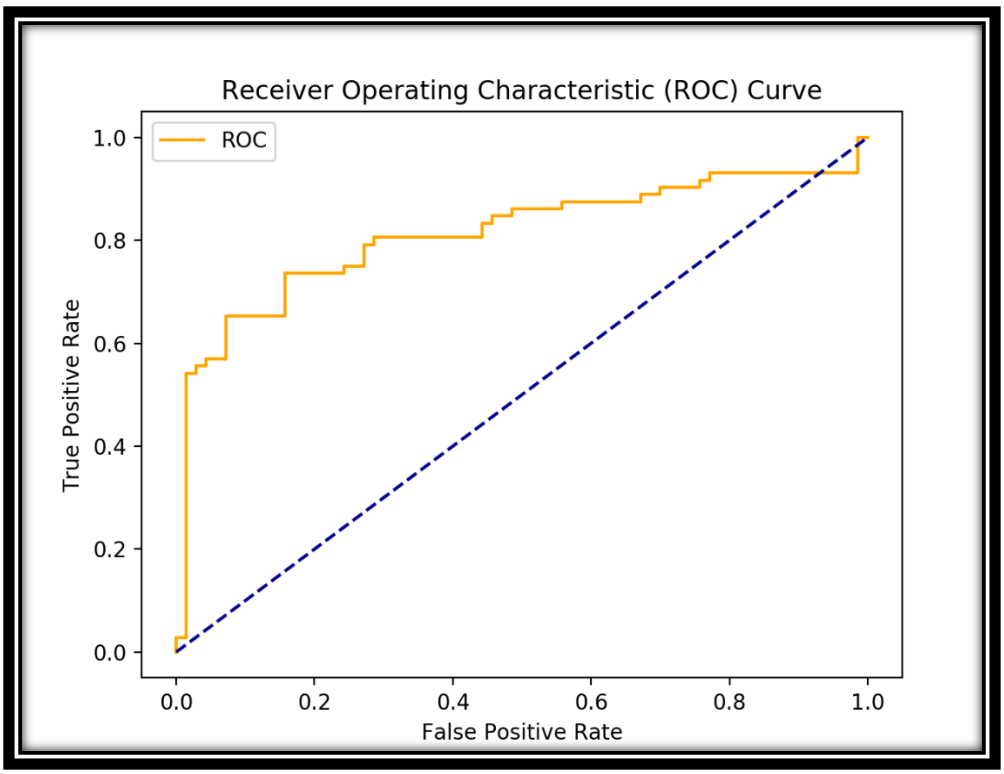
3.1.1. Airport Choice

Performance Measures	Threshold = 0.5	Threshold = 0.7	Threshold = 0.3
Accuracy	0.79	0.76	0.78
Precision	0.85	0.88	0.78
Recall	0.72	0.625	0.80
F1 Score	0.78	0.73	0.79

Key Findings

ROC Curve for Airport Choice:

A ROC (Receiver Operator Characteristic Curve) can help in deciding the best threshold value. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as we vary the threshold for assigning observations to a given class. ROC curve will always end at (1,1). The threshold at this point will be 0.



Below we will try to select the best threshold for the trade-off. According to the criticality of the business, we need to compare the cost of failing to detect positives vs cost of raising false alarms. On increasing the threshold from 0.5 to 0.7, the true positive rate and false positive rate has decreased. However, on decreasing the threshold value from 0.5 to 0.3, the true positive rate and false positive rate has increased. Hence, best threshold value for the model is 0.5 because the accuracy of 79% is maximum at 0.5 with minimum trade-offs.

Airport Choice: On checking the trade-offs between true positive rate and true negative rate, it is clear that the with Area Under the Curve (AUC) of 0.81, the Airport choice logistic regression classifier is returning accurate results (high precision) as well as returning a majority of all positive results (high recall).

Confusion Matrix with Threshold = 0.5

True positive rate (Recall, positive hit) = a / (a+b) = 0.72
False positive rate (Type 1 error, false alarm) = c /(c+d) = 0.12

	Predicted 0	Predicted 1
Actual 0	61(True Negative) d	9(False Positive) c (Type1 error)
Actual 1	20(False Negative) b (Type2 error)	52(True Positive) a

Confusion Matrix with Threshold = 0.7

True positive rate (Recall, positive hit) = a / (a+b) = 0.625
False positive rate (Type 1 error, false alarm) = c /(c+d) = 0.08

	Predicted 0	Predicted 1
Actual 0	64(True Negative) d	6(False Positive) c (Type1 error)
Actual 1	27(False Negative) b (Type2 error)	45(True Positive) a

Confusion Matrix with Threshold = 0.3

True positive rate (Recall, positive hit) = a / (a+b) = 0.80
False positive rate (Type 1 error, false alarm) = c /(c+d) = 0.22

	Predicted 0	Predicted 1
Actual 0	54(True Negative) d	16(False Positive) c (Type1 error)
Actual 1	14(False Negative) b (Type2 error)	58(True Positive) a

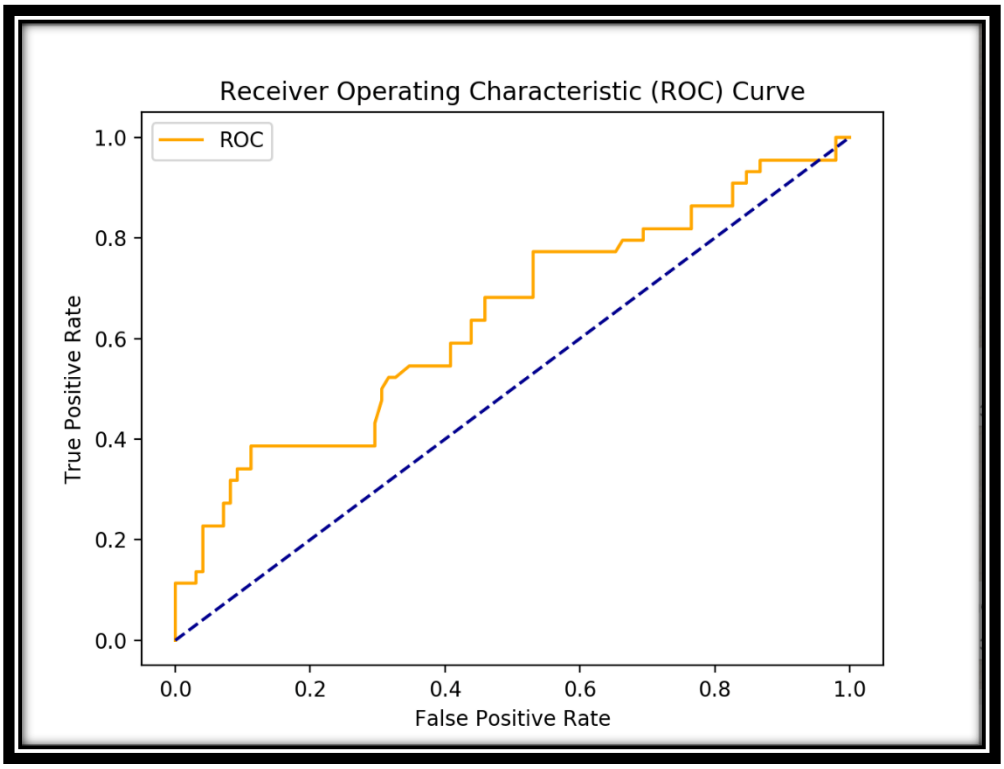
3.1.2. Airline Choice

Performance Measures	Threshold = 0.5	Threshold = 0.7	Threshold = 0.3
Accuracy	0.71	0.72	0.57
Precision	0.60	1.00	0.375
Recall	0.27	0.11	0.54
F1 Score	0.37	0.20	0.44

ROC Curve for Airline Choice:

Below we will try to select the best threshold for the trade-off. On increasing the threshold from 0.5 to 0.7, the true positive rate and false positive rate has decreased. However, on decreasing the threshold value from 0.5 to 0.3, the true positive rate and false positive rate has increased. Hence, best threshold value for the model is 0.7 because the accuracy has increased to 72% with minimum trade-offs.

Airline Choice: For Airline choice, logistic regression classifier, the AUC is low however, the model is giving better results in terms of accuracy and precision measures when the threshold value has been increased to 0.7 from 0.5.



Confusion Matrix with Threshold = 0.5

True positive rate (Recall, positive hit) = a / (a+b) = 0.27
False positive rate (Type 1 error, false alarm) = c /(c+d) = 0.08

	Predicted 0	Predicted 1
Actual 0	90(True Negative) d	8(False Positive) c (Type1 error)
Actual 1	32(False Negative) b (Type2 error)	12(True Positive) a

Confusion Matrix with Threshold = 0.7

True positive rate (Recall, positive hit) = a / (a+b) = 0.11
False positive rate (Type 1 error, false alarm) = c /(c+d) = 0

	Predicted 0	Predicted 1
Actual 0	98(True Negative) d	0(False Positive) c (Type1 error)
Actual 1	39(False Negative) b (Type2 error)	5(True Positive) a

Confusion Matrix with Threshold = 0.3

True positive rate (Recall, positive hit) = a / (a+b) = 0.54
False positive rate (Type 1 error, false alarm) = c /(c+d) = 0.40

	Predicted 0	Predicted 1
Actual 0	58(True Negative) d	40(False Positive) c (Type1 error)
Actual 1	20(False Negative) b (Type2 error)	24(True Positive) a

3.2. Model 2 – Decision Trees

Model description: A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision.

Strengths:

- Inexpensive to construct
- Extremely fast at classifying unknown records
- Easy to interpret for small-sized trees

- Accuracy is comparable to other classification techniques for many simple data sets

Weaknesses:

- Overfitting occurs when the algorithm captures noise in the dataset.
- The prediction model gets unstable with a very small variance in data.
- A highly complicated Decision tree tends to have a low bias which makes it difficult for the model to work with new data.

Why use this model?

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data). We have considered Pre-Pruning here. Pre-pruning prevents the generation of non-significant branches. It involves using a 'termination condition' to decide when it is desirable to terminate some of the branches prematurely as the tree is generated. When constructing the tree some significant measures can be used to assess the goodness of a split. If partitioning the tuples at a node would result the split that falls below a prespecified threshold, then further partitioning of the given subset is halted otherwise it is expanded. High threshold result in oversimplified trees, whereas low threshold result in very little simplification.

3.2.1. Airport Choice

Key Findings

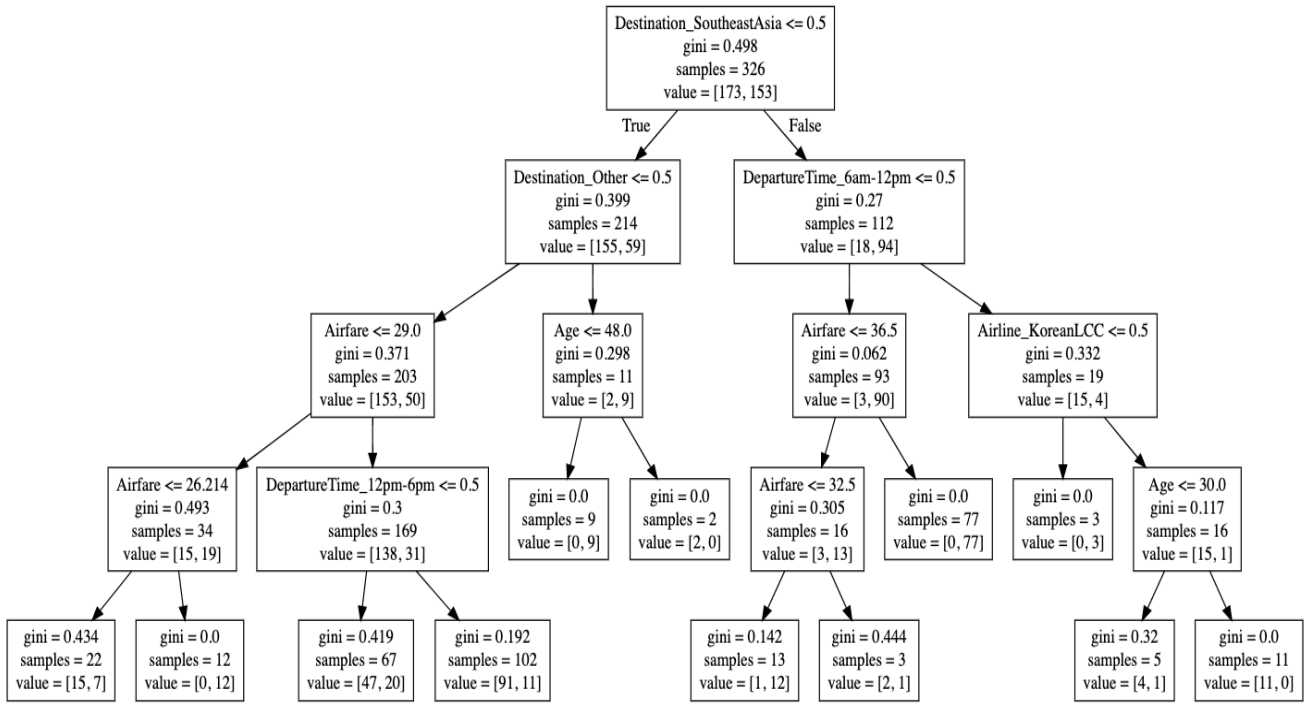
After pruning the decision tree model for airport accuracy and precision has increased considerably.

Pruned Decision Tree Model: Airport
Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	67(True Negative) d	1(False Positive) c (Type1 error)
Actual 1	19(False Negative) b (Type2 error)	53(True Positive) a

Accuracy: 0.8571
Precision: 0.9815
Recall: 0.7361

F1 score: 0.8413



3.2.2. Airline Choice

Key Findings

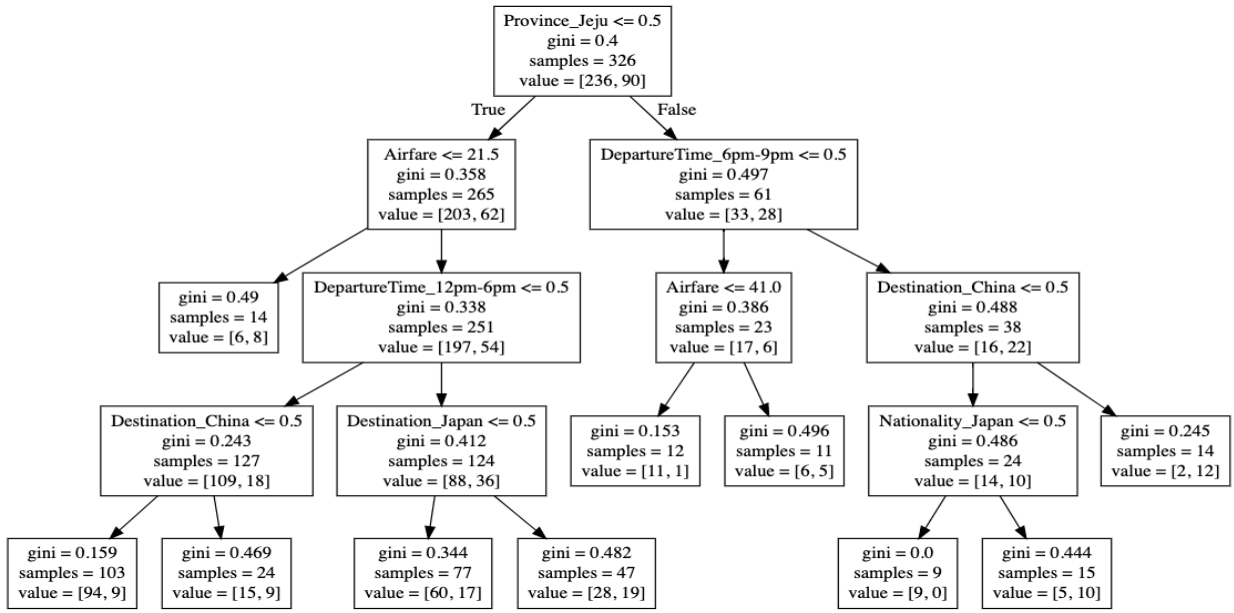
After pruning the decision tree model for airline accuracy and precision has increased considerably.

Pruned Decision Tree Model: Airline
Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	95(True Negative) d	2(False Positive) c (Type1 error)
Actual 1	27(False Negative) b (Type2 error)	16(True Positive) a

Accuracy: 0.7928

Precision: 0.8889
Recall: 0.3721
F1 score: 0.5246



3.3. Model 3 – SVM

Model description

Support Vector Machine is supervised learning model that can analyze the data for classification and regression problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts.
SVM model is a representation of the examples as in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

In addition to performing linear classification, SVM can perform non-linear classifications as well called kernel tricks, implicitly mapping their inputs to high dimensional feature spaces.

Strengths

- It works well with clear margin of separation.
- It is effective in high dimensional data.
- It is also effective in cases where number of dimensions are greater than number of sample data.

- It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- With non-linear SVM, we can capture more complex relationship between data points without having to perform much transformation on our own.

Weaknesses

- The training time is much longer as it is much more computationally intensive.
- It doesn't perform well, when we have large data set because the required training time is higher
- It also doesn't perform very well, when the data set has more noise i.e. target classes are overlapping
- SVM doesn't directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

Why used this model?

SVM model works best for classification where target variable has exactly two classes. Here in both cases, we have exactly two classes, so we are using SVM. But we are using this model just for the bench marking.

3.3.1. Airport Choice

Key Findings

	Kernel		
Performance Measures	Linear [[66 1] [18 57]]	Polynomial, 3 [[63 3] [20 54]]	Radial bias function [[60 6] [19 55]]
Accuracy	0.79	0.83	0.82
Precision	0.92	0.94	0.90
Recall	0.66	0.73	0.74
F1 Score	0.77	0.82	0.81

3.3.2. Airline Choice

Key Findings

	Kernel

Performance Measures	Linear [[84 13] [32 11]]	Polynomial, 3 [[90 7] [35 8]]	Radial bias function [[94 3] [35 8]]
Accuracy	0.67	0.7	0.73
Precision	0.45	0.533	0.73
Recall	0.25	0.186	0.18
F1 Score	0.32	0.275	0.30

3.4. Model 4 – Neural Network

Model description

Among the various machine learning approaches in the sub-field of data classification, neural-network methods have been found to be a useful alternative to the statistical techniques. An artificial neural network is a mathematical model, inspired by biological neural networks, are used for modeling complex relationships between inputs and outputs or to find patterns in data.

Neural networks are typically organized in layers. Layers are made up of a number of interconnected nodes which contain an activation function. Patterns are presented to the network via the input layer, which communicates to one or more hidden layers where the actual processing is done via a system of weighted connections. The hidden layers then link to an output layer. Most ANNs contain some form of 'learning rule' which modifies the weights of the connections according to the input patterns that it is presented with. In a sense, ANNs learn by example.

Strengths

- Information such as in traditional programming is stored on the entire network, not on a database. The disappearance of a few pieces of information in one place does not prevent the network from functioning.
- After ANN training, the data may produce output even with incomplete information. The loss of performance here depends on the importance of the missing information.
- Corruption of one or more cells of ANN does not prevent it from generating output. This feature makes the networks fault tolerant.
- The network's success is directly proportional to the selected instances, and if the event cannot be shown to the network in all its aspects, the network can produce false output

Weaknesses

- Artificial neural networks require processors with parallel processing power, in accordance with their structure. For this reason, the realization of the equipment is dependent.
- Unexplained behavior of the network is the most important problem of ANN. When ANN produces a probing solution, it does not give a clue as to why and how. This reduces trust in the network.
- There is no specific rule for determining the structure of artificial neural networks. Appropriate network structure is achieved through experience and trial and error.

Why use this model?

Neural networks helps to classify the model. It is a classification layer on top of the data we can store and manage. In this project, we have used Neural Networks because they help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on. Just like SVM, we are using this model result as benchmark.

3.4.1. Airport Choice

Performance Measures	Logistic 3 hidden layer of 10 neurons	Identity 3 hidden layer of 10 neurons	Tanh 3 hidden layer of 20 neurons
Accuracy	0.76	0.80	0.83
Precision	0.83	0.86	0.84
Recall	0.65	0.72	0.81
F1 Score	0.73	0.78	0.83

Key Findings

On increasing the number of neurons in the hidden layers with activation as tanh, the accuracy has increased considerably.

3.4.2. Airline Choice

Performance Measures	identity	tanh
Accuracy	0.71	0.74

Precision	0.56	0.64
Recall	0.29	0.40
F1 Score	0.38	0.50

Key Findings

We are using Neural Networks as a benchmark model hence using only tanh as a activation parameter.

3.5. Model comparison & evaluation

3.5.1. Airport Choice

	Models Comparison			
Performance Measures	Logistic Regression Threshold = 0.5	Decision Tree	SVM, Polynomial, 3	Neural Networks Activation = tanh
Accuracy	0.79	0.86	0.83	0.83
Precision	0.85	0.98	0.94	0.84
Recall	0.72	0.74	0.73	0.81
F1 Score	0.78	0.84	0.82	0.83

3.5.2. Airline Choice

	Models Comparison			
Performance Measures	Logistic Regression Threshold = 0.7	Decision Tree	SVM, Radial bias function	Neural Networks Activation = tanh
Accuracy	0.72	0.79	0.73	0.74
Precision	1.00	0.89	0.73	0.64
Recall	0.11	0.37	0.18	0.40
F1 Score	0.20	0.52	0.30	0.50

Airport Choice

Best model for Airport Choice: Decision Tree

Accuracy: 0.86
Precision: 0.98
Recall: 0.74
F1 score: 0.84

Airline Choice

Best model for Airline Choice: Decision Tree

Accuracy: 0.79
Precision: 0.89
Recall: 0.37
F1 score: 0.52

4. Conclusion

4.1. Our solutions

Main aim of our study was descriptive analysis and analyzing peoples' choice. Even though, we have used neural network and support vector machine, they only provide benchmark scores and cannot be used for description as they are not easily interpretable. Support vector machine takes long time to predict and fit for large data and neural networks are considered as black boxes. With logistic regression there are issues of poor feature inference. When the variables are highly correlated or highly nonlinear, the coefficients of logistic regression will not correctly identify the gain/loss from each individual feature.

Best model for our project for descriptive analysis is decision tree since decision tree is easier to interpret and it provides advantage of equivalent accuracy with interpretability. If this becomes a predictive project in future then we can use neural network and support vector machine, since, they have better predictive power, but they do not have good descriptive power.

4.2. Future Improvements

Future works based on (Loo, 2008) :

1. Airport characteristics like including airport access time, access cost, access mode, parking facilities, check-infacilities, lounge, restaurant and shopping facilities, transfer facilities, baggage, customs and immigration facilities, and airport tax or passenger charge

2. Flight characteristics, including flight frequency, in-flight travel time, number of stops, transfer arrangements, congestion or punctuality of flights, airlines serving the route and aircraft type.
3. Modes of transportation and access time, access cost can be good parameters to study public transport to Airport.

Limitations and Recommendations:

1. Based on results Data is biased, 77% of people in dataset are from Korea.
2. As per the data, mostly people prefer their flight departure time during the day. This is another bias in the data.
3. Also, ICN is a much larger airport with many other features. The data considers both airports equally. This might affect the results.

References

Loo, B. P. (2008). Passengers' airport choice within multi-airport regions (MARs):some insights from a stated preference surveyat Hong Kong International Airport. *Journal of Transport Geography 16*, 117–125.