Final Report

Examining Factors that affect Airline Passenger's Satisfaction by using machine learning algorithm

Ervin Gubin Moung [1,*], Tuen Yong Hao [1] , Lim Hong Yao [1] , Lim Min Xuan [1] , Peter Chin Meng Meng [1]

[1]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu 88400, Sabah, Malaysia
*Correspondence: ervin@ums.edu.my

Abstract: Airline passenger's satisfaction is one of the factors that affect the reputation of an airline reputation. Numerous studies have applied Machine Learning algorithms to determine the factors that affect the satisfaction of passengers and most of their results indicated that the airline provided services as the most influential factor that contribute to the airline passengers' satisfaction. Thus, this paper proposes an investigation applying Machine Learning techniques which are Random Forest classifier, Logistic Regression, and Gradient Boosting Machine classifier to determine the nature of the existing factors associated with the airline passenger's satisfaction based on the dataset provided by the American Customer Satisfaction Index (ACSI) Travel Report 2015. From the model evaluation stage, the results obtained from three classification models will be compared by using the performance metric of each model in order to examine the stability and accuracy of the three classifiers. The result from this study also has shown that certain kinds of services are mainly contributed to the decision-making of classifiers on determining the satisfaction level of airline passengers.

Keywords: Airline, Passenger's satisfaction, Reputation, Services, Machine Learning, Random Forest Classifier, Logistic Regression, Gradient Boosting Classifier, Classification, Performance Metrics

1. Introduction

Growth in service began to accelerate in the 1960s and accelerated again after the double-dip recession in early 1980.[1] Based on the commerce ministry data, service sectors accounts for 60% of India's GDP and 70% of Karnataka's GDP [2]. Thus, from this statement, service sectors play an important role in the global economy in helping every country in this world generate profit. In order to improve the efficiency of service sectors that is highly sensitive to the world's economy, there are several field and area that have to be studied, investigated, and analyse. one of the fields is customer satisfaction. Customers' satisfaction determines the loyalty of customers towards a certain type of service or brand among tons of them. Thus, airlines work desperately to find a way to boost customers satisfaction in order to have sustainable customer loyalty.

Passengers are more likely to judge an airline company and their flight experience based on their level of satisfaction throughout the whole flight journey. The reduction in in-flight service can negatively affect an airline's customer satisfaction rating. Thus improving the quality of the in-flight service should be emphasized as it is one of the factors that make one airline company successful.

In this paper, the researchers will start off with state-of-art that state the method of collecting information for this research. The researcher will show the motive and methodology using classification and regression algorithms to determine what are the main factors that affect airline passenger's satisfaction.

2. State-of-the-Art

2.1. Application of Sentiment Analysis with Machine Learning Approaches

Sentiment analysis is a significant method for extracting emotions from the textual information, such as web articles, product evaluations, movie reviews, Twitter data, and so on. Twitter data typically offer information about a person's viewpoint on a variety of topics. Air passengers usually share their travel experiences on Twitter. Air travel is becoming increasingly popular on Twitter as a result of the current trend. This input can be significant if processed using machine learning techniques since it can reveal insights on the passenger's degree of comfort during the journey studies from Rane and Kumar, 2018 [3] used decision trees, random forests, Gaussian Nave Bayes, SVM, K-nearest neighbours, logistic regression, and AdaBoost to compare six US-based airline firms. They trained the classifiers on 80% of the data and used the remaining data for testing. They divided the tweets into three sentiment categories. They stated that logistic regression, AdaBoost, random forest, and SVM performed well on the model with more than 80% accuracy, but that improvements can be made by including more tweets in the analysis.

2.2. Application of data-mining and analytics method on classifying the airline passenger satisfaction and loyalty

Previous studies from Wong, J.-Y., & Chung, P.-H., 2008 and Javed Parvez & Sahayadhas, 2020 [4,5] on retaining and monitoring the airline passenger satisfaction and loyalty found to apply the data analytics and data mining approaches which involve the use of machine learning algorithm to obtain the useful insights and knowledge from the existing passenger database. According to Larose, 2015 and Sedkaoui, 2018 [6,7], data mining is a process of discovering meaningful patterns and trends in a large scale of data whereas data analytics is the act of examining, cleaning, manipulating, and modelling data which comprise the goal of data mining to find relevant information, suggesting conclusions, and assist in decision-making. In other words, both data mining and data analytics are crucial to the researchers who intend to obtain new knowledge from their data models by using the appropriate machine learning algorithm, mathematical and statistical techniques. Wong, J.-Y., & Chung, P.-H., 2008 have applied the Decision tree and logistics regression algorithm as their data mining techniques and adopted the result generated from the Decision tree model that had a slightly higher accuracy (82.6%) than the logistic regression model (82.5%) to identify the characteristics of the loyal passengers of one of the Taiwanese airline. Their study demonstrated that both the algorithms performed a high and consistent accuracy when differentiating the loyalty and unloyalty airline passengers after applying the cross-validation method. Besides, Javed Parvez & Sahayadhas, 2020 compared the accuracy of three of the algorithms that are Random Forest algorithm, Gradient Boosting Machine, and Decision tree utilizing the data analytics techniques to determine the most suitable algorithm for classifying the satisfactory level of airline passengers based on the overall ratings and determining the loyalty of airline passenger by recommending the airline based on their satisfaction on various categories. Their results demonstrated that the Random Forest algorithm (42%, 92.7%) and Gradient Boosting Machine (46%, 94.6%) were capable of efficiently classifying passenger satisfaction and loyalty based on the enormous scale of the data.

3. Motivation

Satisfaction is a consumer's subjective assessment of the extent to which desires and demands arising from the acquisition or consumption of a provided product or service are satisfied, and reusability is the likelihood that the customer will continue to use the product or service after purchasing it (Insil Park, 2009 and Czepiel, J. A. 1997) [8,9]. Passengers who need to travel on an airline and want to be satisfied with the travel experience including in-flight services, check-in services, in-flight entertainment, cleaning, etc. Customer satisfaction is a key element for modern businesses as it can significantly contribute to a continuing effort of service quality improvement. In order to meet customer expectations and achieve a higher level of quality, airlines need to

develop a specific passenger satisfaction measurement mechanism. Below are the research problems and research objectives related to it.

Research Problems:

1. What are the most important attributes/factors in predicting airline passenger's satisfaction?

2. What is the relationship between airline passenger's satisfaction and service quality?

Research Objectives:

1. To investigate the prediction predicting airline passenger's satisfaction for using machine learning algorithms.

2. To evaluate the machine learning model for airline passenger's satisfaction.

4. Methodology

4.1. Overview of the machine learning workflow

In order to determine how accurate it is the prediction of passenger satisfaction, there few methods that have been implemented to ensure the accuracy and efficiency of the model. First is data preparation. Data preparation is one of the most important steps before data modelling. This is because unwanted data and noisy data are removed from the data sets in this stage. At the same time, missing values are identified during data preparation. Next is model selection. In this investigation, the random forest model, logistic regression model, and Gradient boosting model are selected to classify the satisfaction level of airline passengers. Lastly is model evaluation. In this stage, the accuracy of each model is evaluated.

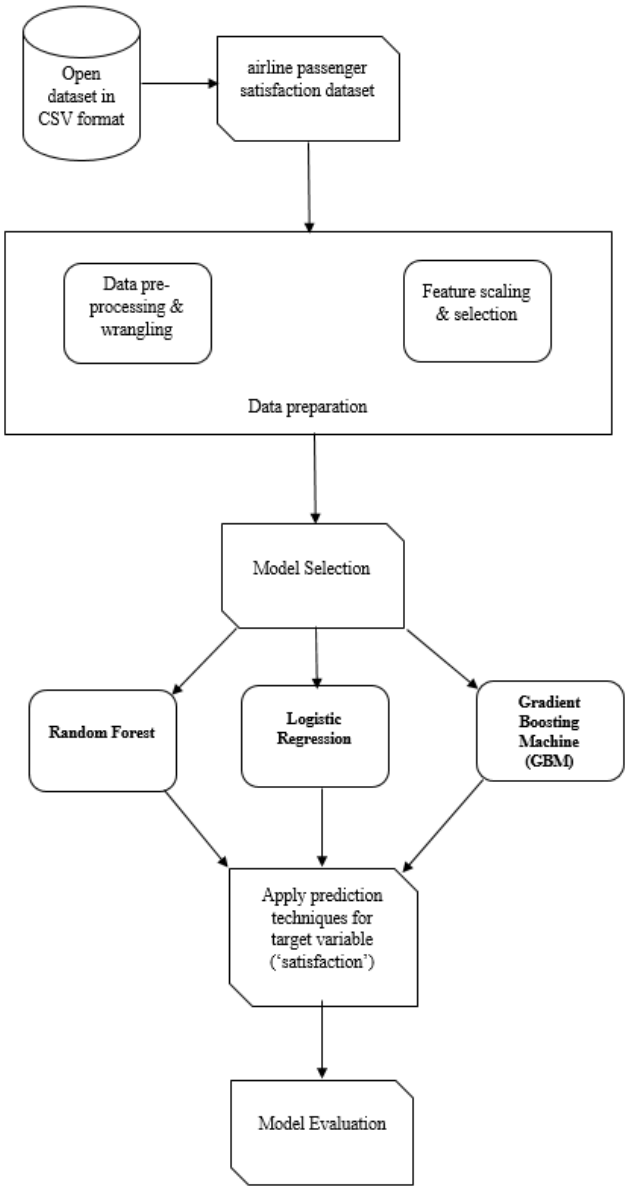


Figure 1. The overall view of machine learning workflow for this project

4.2. Random Forest

Random forest is a machine learning algorithm that combines the output of multiple decision trees to get the prediction result. decision trees are the building blocks of the random forest models. Generally, the decision tree used the Classification and Regression Tree (CART). The decisions which were represented by leaf nodes were made based on the given condition which is represented by nodes. In this prediction model, all the factor that has a high potential to affect the prediction are inserted as the nodes of the decision tree.

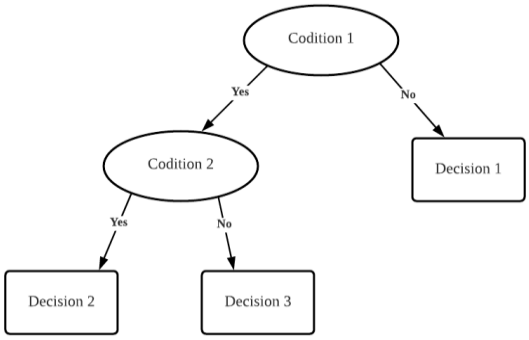


Figure 2. the general view of decision making based on the conditions

The data that has been ensembled in the decision tree is then become a training set for the random forest model by using a bootstrap sample. The model studied the patterns of the training set by referring to its condition and final result. The more data is used for the model training, the higher the accuracy of the prediction can get.

4.3. Logistic Regression

Logistic regression is a regression model that models the probabilities for classification problems with two possible outcomes. Logistic regression models use logistic functions to squeeze the output of linear equations between 0 and 1. The formula of the logistic function is

$$p = \frac{1}{1 + \square^{-\square}}$$

In general, logistic regression is a generalized linear model using the same basic formula of linear regression which focuses on regressing the probability of a categorical outcome. The formula of linear regression is defined as:

$$y = \square_0 + \square_1 \square + \ldots\ldots \square \quad \square \square \quad \square$$

where y is the predicted value of the dependent variable for any given value of the independent variable(x); $\square_0$ is the intercept, the predicted value of y when the x is 0; $\square_1$ is the regression coefficient which is how much y is expected to change as x increases; x is the independent variable that is expected to be influencing y; By substitute the equation of linear regression into the logistic function, the logistic regression formula is defined as:

$$\ln\left(\frac{\square}{1-\square}\right) = \square_0 + \square_1 \square + \ldots\ldots \square \quad \square \square \quad \square$$

where *p* is the probability of belonging to one class, $\frac{\square}{1-\square}$ is the odds ratio, and $\square_1 \ldots \square \quad \square$ are regression coefficients that are to be estimated based on the data. The commonly used method to estimate these coefficients is the maximum likelihood. In general, the goal of logistic regression is to use the best method to correctly predict the outcome category of a single case model. To achieve this goal, a model was created that includes all predictors useful for predicting response variables. An appropriate outcome type variable is required for data checking. In the calculation of logistic regression, the probability of success is greater than the probability of failure. The analysis results are in the form of an odds ratio. From the analysis, the p-value, which is the probability of finding the given test statistic if the null hypothesis of no relationship were true, is highly regarded to indicate whether the model fits the data well. A model whose p-value is closer to 0 will fit the data well. The model's performance can be evaluated by using Mean Absolute Error (MAE) which is used to measures how far predicted values are away from observed values. Mean Absolute Error (MAE) shows us how big the expected error from the forecast on average. Finally, the results and observations are taken.

4.4. Gradient Boosting Machine(GBM)

Gradient Boosting Machine is a kind of advanced machine learning technique to form a strong learner model by comprising numerous base-learners models such as decision trees [10,11]. The purpose of comprising the base-learners models is to obtain a more accurate estimation result of the dependent variable that holds the minimum loss function where the formulas of the loss function (1) and the Gradient Boosting algorithm (2) are shown as below:

$$\square_{\square}(\square) = \square_{\square}\left[\frac{\square(\square,\square(\square))}{\square\square(\square)}\middle|\square\right]_{\square(\square)=\hat{\square}^{\square-1}(\square)} \quad (1)$$

$$(\rho_t, \theta_t) = \arg\min_{\rho,\,\theta} \sum_{i=1}^{N} \left[-g_t(x_i) + \rho h(x_i, \theta)\right]^2 \quad (2)$$

Generally, the main principle of the Gradient Boosting Machine model is a simultaneous mathematical step to combine the estimation result such as the negative gradient of loss function of the previous conducted weak-learners model until a certain amount of the iteration has been completed. The model's performance also could be evaluated by inspecting the calculated loss function besides using the confusion matrix and Mean Absolute Error (MAE). Thus, the outcome of the final estimation result is believed to be reliable as it is improved by combining the weaknesses of all of the conducted base-learners. Gradient Boosting Machine also provides high flexibility that permits to decide the type of base learners model by invoking a different kind of algorithm such as linear regression instead of a decision tree. Hence, it supports performing either regression or classification tasks.

4.5. Data pre-processing stage

The raw dataset which consists of the airline passenger's satisfaction information is transformed as a data frame under a Python programming environment. The column 'Unnamed: 0' has been removed due to its irrelevant content. The dependent variable's column 'satisfaction' and one of the independent variable's columns 'customer_type' is encoded into integer type variable from object type by using the ordinal encoding technique whereas the other object type independent variables' column which are 'Gender', 'customer_class' and 'type_of_travel' encoded into integer type by using the dummy encoding technique. In order to provide a high performance of the classification model, 393 of missing values identified in this dataset will be handled by invoking the Multivariate imputation by chained equations (MICE) imputation algorithm for the GBM model whereas the Logistics Regression and Random Forest models will drop the data contain missing values in order to handle the missing values problem. After transforming the raw dataset into a data frame that is suitable for further construction of the classification model, the details about the transformed dataset is illustrated by using the describe() and mode() function under the Python programming language. Furthermore, A correlation plot is constructed to investigate the relationship between the satisfaction level of airline passenger and all of the independent variables that are related to the type of service.

4.6. Performance metrics

The performance metric used in this research is the Confusion Matrix. Confusion matrix is used to describe the performance of classification model and it consists of values that represent the true positive (TP), true negative (TN), false negative (FN), false positive (FP). True positive and true negative are the values that the model correctly predicted. All the values represented in confusion matrix produced by each of the models are contributed to the calculation of accuracy, precision, and recall of the model. The further explanation on the details of accuracy, precision, and recall will be presented in section 5.1.

4.7. Dataset description

The dataset can be accessed via the Kaggle website at the following link: https://www.kaggle.com/binaryjoker/airline-passenger-satisfaction. The dataset is about the survey of passenger satisfaction in US airlines which is taken from the American Customer Satisfaction Index (ACSI) Travel Report 2015 at the following link:

https://www.theacsi.org/industries/travel/airline. In this study, the dataset named 'air_passenger_satisfaction.csv' is chosen for modelling by using classification and regression machine learning algorithms. The information of the dataset consists of 129880 rows of data with 24 columns of variables. The dependent variable or response variable in this dataset is 'satisfaction' which corresponds to passenger satisfaction and the predictor variables or independent variables are 'Unnamed: 0' which corresponds to customer identification collected, 'Gender' corresponds to either male or female, 'customer_type' corresponds to customer loyalty, 'age' corresponds to numeric age in terms of years, 'type_of_travel' corresponds to either personal or business travel, 'customer_class' corresponds to customer classes in the airline, 'flight_distance' corresponds to numeric flight distance of travelled in miles, 'inflight_wifi_service' corresponds to numeric rating on WiFi service in the flight, 'departure_arrival_time_convenient' corresponds to numeric rating on the convenience of departure and arrival time, 'ease_of_online_booking' corresponds to numeric rating on the difficulty of online booking, 'gate_location' corresponds to numeric rating on the location of gates, 'food_and_drink' corresponds to numeric rating on the quality food and drink, 'online_boarding' corresponds to numeric rating on the online boarding service, 'seat_comfort' corresponds to numeric rating on whether the seat is comfortable or not, 'inflight_entertainment' corresponds to numeric rating on the entertainment facilities inside the cabin, 'onboard_service' corresponds to numeric rating on the boarding service provided by flight attendants, 'leg_room_service' corresponds to numeric rating on the standard seat pitch, 'baggage_handling' corresponds to numeric rating on the service of transporting passenger luggage, 'checkin_service' corresponds to numeric rating on the service of check-in at the service counter, 'inflight_service' corresponds to numeric rating on the services provided during the flight, 'cleanliness' corresponds to numeric rating on the cleanliness of the cabin, 'departure_delay_in_minutes' corresponds to numeric rating on the delay of departure time calculated in minutes, and 'arrival_delay_in_minutes' corresponds to numeric rating on the delay of arrival time calculated in minutes.

4.8. Data modelling

There is a total of three models that were trained by using the dataset known as airline_passenger_satisfaction.csv for training purposes after the data pre-processing stage. The chosen dataset was manually divided into training and testing sets with a ratio of 80.0% and 20.0% respectively for the use of three model's training stage. The partition of the chosen dataset is summarised in **Table 1.** As illustrated in **Table 1**, the training dataset consists of 1039904 rows of transformed data which are about the airline passenger information, and the testing dataset consists of 25976 rows of transformed data. Then the random forest classifier is imported in order to call the random forest function. Next, the training sets are inserted into the random forest model for the machine to learn. At the stage of building a logistic regression-based model, the features are normalized accordingly before the data partition. Furthermore, the model is built by calling the LogisticRegression function and fitting the data into the appropriate scale. In GBM model construction, the model's hyper-parameters were fine-tuned for preventing the outcome of an overfitting result by using one of the unnormalised training datasets. Then the GBM model was fit into the normalised training dataset for training purposes. According to a dataset threshold study, the k-fold cross-validation technique is recommended to be applied during the model's training phase with a large scale of the dataset as this technique could average the prediction error by recursively training the model. Hence, in order to achieve a more reliable prediction result, the 5-fold validation technique was applied to the GBM model's hyperparameter tuning process and the training stage.

Table 1. Dataset partition

Dataset	Total	Percentage
Training	103904	80.0%
Testing	25976	20.0%
Total	129880	100%

5. Results and Discussion

In order to explore the nature of attributes that influence the airline passengers' satisfaction which is categorised as two categorical variables in terms of neutral or dissatisfied and satisfied. The correlation plot shown in **Figure 3** is used to mainly examine the relationship between the dependent variable ('satisfaction') and all of the independent variables that are relevant to the airport's provided services. As a general analysis based on **Figure 3**, there is a total of 3 features which are 'departure_arrrival_time_convienient', 'customer_class_Eco', 'customer_class_Eco Plus' and type_of_travel_Personal Travel' were found to negatively influence the satisfaction level of airline passengers where there is a total of 14 features which are 'inflight_wifi_service', 'ease_of_online_booking', 'food_and_drink', 'online_boarding', 'seat_comfort', 'inflight_entertainment', 'onboard_service', 'leg_room_service', 'baggage_handling', 'checkin_service', 'inflight_service', 'cleanliness', 'customer_class_Bussiness' and 'type_of_travel_Business Travel' were found to positively influence the satisfaction level of airline passengers.

By further comparison between the correlation coefficient of each of the independent variables that are related to the airport's provided services, the services labelled with 'customer_class_Eco' and type_of_travel_Personal Travel' with the correlation coefficient smaller than or equal to -0.45 were found to be the most negatively associated with the satisfaction level of airline passengers whereas the 'online_boarding', 'inflight_entertainment', 'customer_class_Bussiness' and 'type_of_travel_Business Travel' with the correlation coefficient greater than or equal to 0.45 were found to be the most positively associated with the satisfaction level of airline passengers.

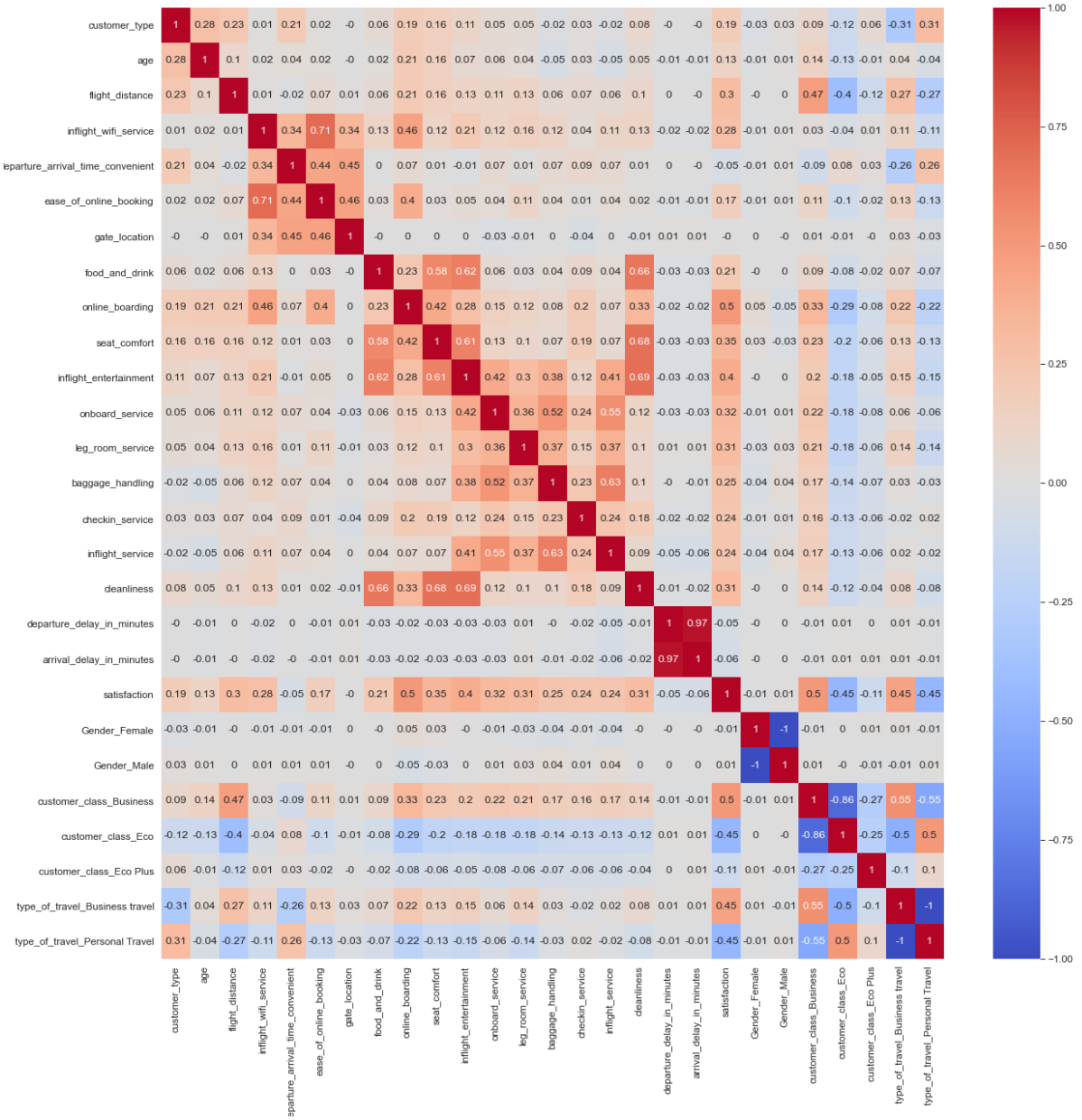


Figure 3. Correlation plot between the dependent variable and independent variables.

5.1 Models' Performance on Airline Passengers' Satisfaction

For each of the supervised models built, the training and testing steps were conducted 100 times. Due to the objective of this research is mostly related to the classification task, thus, the mean measures of the models' classification evaluation such as accuracy, precision, and recall will be computed for further comparison of the models' performance by combining the result from each of the testing phases conducted. **Table 2** illustrates three models' performance on classifying the satisfaction level of airline passengers based on one of the testing models that generated the greatest evaluation's performance from 100 testing models.

Accuracy is one of the measures of a classification model's performance that determine the correctness of a model on classifying the category of the target variable based on its characteristics. The formula for computing the accuracy of a classifier is shown in **Formula 1**. The data from **Table 2** clearly indicates that the Random Forest algorithm-based model (96.14%) and Gradient Boosting Machine model (96.12%) could perform the classification of the satisfaction level more accurately than the logistic regression model (87.45%).

Formula 1:

$$\square\square\square\square\square\square\square\square = \frac{\square\square + \square\square}{\square\square + \square\square + \square\square + \square\square}$$

Precision and Recall are one of the classifier's evaluation metrics that measure the correctness of the classifier on classifying the target attribute of the target variable such as the correctness on classifying the satisfied airline passenger from all of the satisfied airline passenger classified by the model. **Formula 2 and Formula 3** illustrate the calculation of a model's precision and recall respectively. As from **Table 2**, the evaluation result of the three models' performance states that the GBM model has the highest precision score (94.77%) computed which in other words the GBM model could precisely classify the satisfied airline passengers with the least number of airline passengers that were wrongly classified as the satisfying category (FP). Apart from the comparison between the three models' precision scores, the Random Forest classification model was found that own the highest recall score (97.17%) among three of the classifiers. To be more particular, the Random Forest classifier could precisely classify the satisfied airline passengers with the least number of airline passengers that were wrongly classified as the non-satisfying category(FN).

Formula 2:

$$\square\square\square\square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square}$$

Formula 3:

$$\square\square\square\square\square\square = \frac{\square\square}{\square\square+\square\square}$$

Area Under the Receiver Operating Characteristic Curve (AUC) is also an evaluation metric for a classification model's performance that determines the ability of the classifier to differentiate the classes of the target variable. By referring to **Table 2**, three of the classifiers were having a relatively high AUC value that is above 0.90 which indicate that all of these models were strong classifiers that could correctly differentiate the satisfaction level of the airline passengers.

Table 2. Confusion matrix of three models on the testing dataset

Model	TP	TN	FP	FN	AUC	Accuracy (%)	Precision (%)	Recall (%)
Random Forest	10210	14577	1109	80	0.956	96.14	93.88	97.17
Logistic Regression	9517	13255	1368	1836	0.930	87.45	83.57	87.07
GBM	10556	14411	427	582	0.990	96.12	94.77	96.11

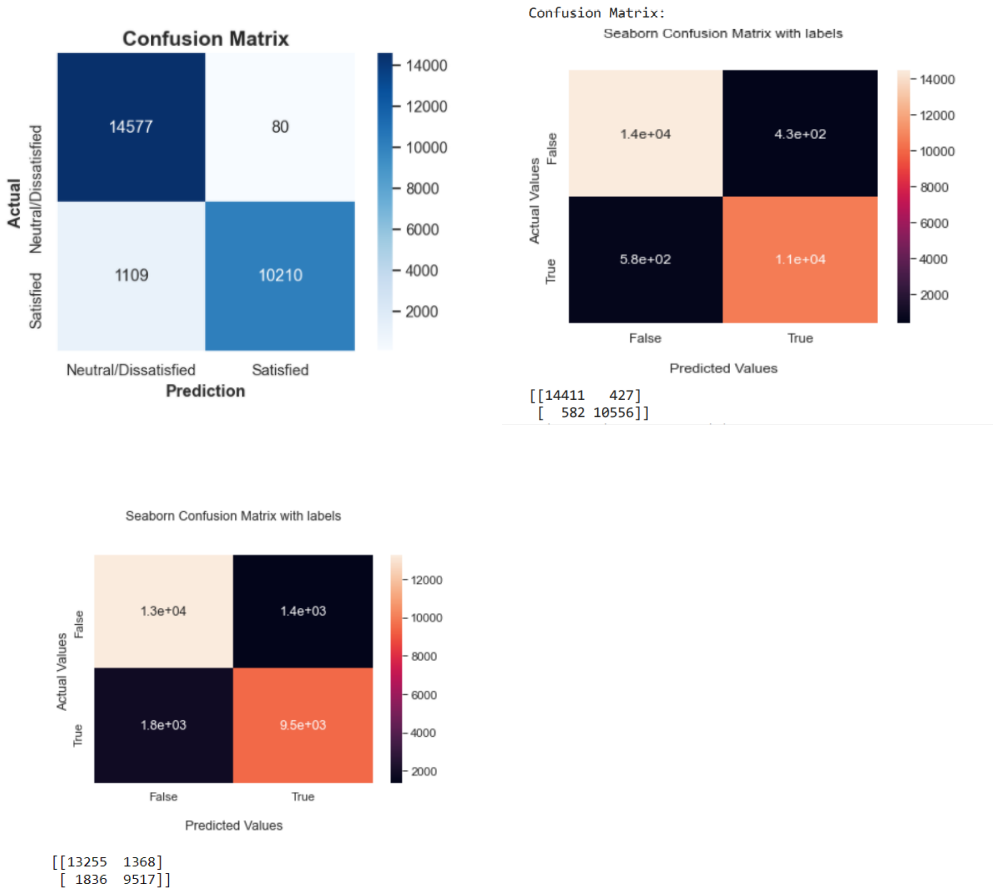


Figure 4: confusion matrix produced by the Random Forest, Logistics Regression, and GBM model

5.2 Models' Performance Report with Confidence Intervals

Confidence interval (CI) is an estimated range of values likely to involve an unknown population parameter. It is commonly expressed as a percentage (%) of the population mean lying between an upper and lower interval. 95% confidence interval is chosen for calculating the confidence interval of the accuracy, precision, and recall on three supervised models.

Table 3 presents the calculated mean confidence intervals for the models' performance on 100 testing datasets respectively. As a general view of all models' performance from **Table 3**, the logistic regression has the lowest performance but the longest range of the confidence interval by comparing with the Random Forest model and GBM model on classifying the satisfaction level of airline passengers. In other words, the Random Forest and GBM model's performance could be considered more stable than the Logistic regression model's performance as the predicted label of both models will have not much deviation for each of the experiments conducted.

Table 3. Confidence intervals for the models' performance on the testing dataset.

Model	Accuracy±□□□□□(%)	Precision±□□□□□(%)	Recall±□□□□□(%)
Random Forest	96.13±0.01	93.88±0.01	97.17±0.01
Logistic Regression	87.45±0.04	83.57±0.06	87.07±0.06
GBM	95.58 ±0.008	94.04±0.007	95.73±0.01

5.3 Feature Importance of three supervised models

Feature importance gives the notion which contributes to the decision of classification models on classifying the label. The feature importance values from Logistics Regression are tabulated in **Table 4**, the feature importance graph and the list of feature importance coefficients of the Random Forest model visualised using **Figure 5** and **Figure 6** respectively for Random Forest whereas

Figure 7 is the feature importance graph and feature importance coefficient list of the Gradient Boosting Machine.

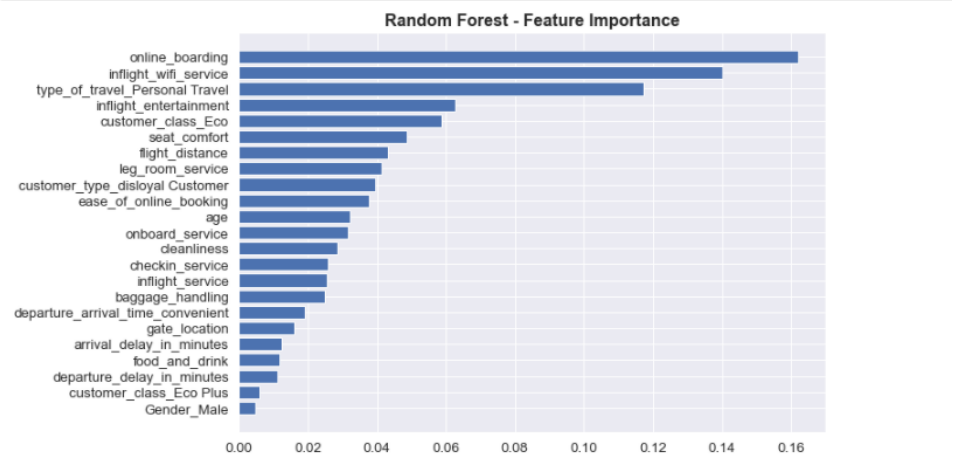


Figure 5. Feature importance graph of Random Forest model

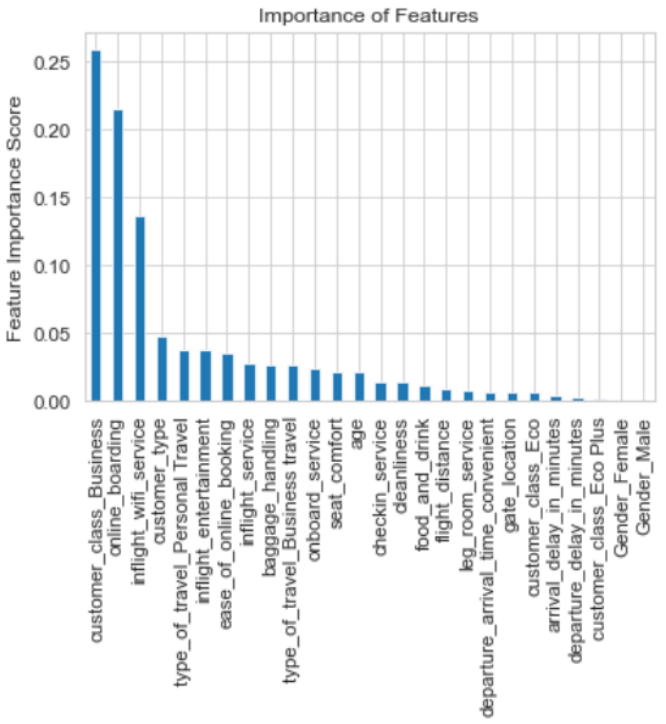
	features	importance
18	Gender_Male	0.004595
22	customer_class_Eco Plus	0.006041
16	departure_delay_in_minutes	0.011110
6	food_and_drink	0.011764
17	arrival_delay_in_minutes	0.012466
5	gate_location	0.015924
3	departure_arrival_time_convenient	0.019020
12	baggage_handling	0.024709
14	inflight_service	0.025522
13	checkin_service	0.025784
15	cleanliness	0.028436
10	onboard_service	0.031496
0	age	0.032031
4	ease_of_online_booking	0.037776
19	customer_type_disloyal Customer	0.039635
11	leg_room_service	0.041277
1	flight_distance	0.043021
8	seat_comfort	0.048637
21	customer_class_Eco	0.058566
9	inflight_entertainment	0.062772
20	type_of_travel_Personal Travel	0.117382
2	inflight_wifi_service	0.140031
7	online_boarding	0.162006

Figure 6. Importance value for each of the features of the Random Forest model

Table 4. Sort coefficients for features on Logistic Regression model

Features	Coefficient (Logistic Regression)
online_boarding	3.01
customer_type	2.03
inflight_wifi_service	1.97
checkin_service	1.62
onboard_service	1.51
type_of_travel_Business travel	1.36

leg_room_service	1.27
cleanliness	1.11
inflight_service	0.58
baggage_handling	0.54
customer_class_Business	0.50
seat_comfort	0.35
inflight_entertainment	0.30
gate_location	0.12
Gender_Male	0.02
Gender_Female	-0.04
flight_distance	-0.08
food_and_drink	-0.13
customer_class_Eco	-0.20
customer_class_Eco Plus	-0.30
age	-0.61
departure_arrival_time_convenie	-0.64
ease_of_online_booking	-0.73



	features	importance
20	Gender_Male	0.00045
19	Gender_Female	0.00057
23	customer_class_Eco Plus	0.00200
17	departure_delay_in_minutes	0.00313
18	arrival_delay_in_minutes	0.00355
22	customer_class_Eco	0.00581
6	gate_location	0.00594
4	departure_arrival_time_convenient	0.00620
12	leg_room_service	0.00824
2	flight_distance	0.00896
7	food_and_drink	0.01176
16	cleanliness	0.01343
14	checkin_service	0.01372
1	age	0.02103
9	seat_comfort	0.02187
11	onboard_service	0.02396
24	type_of_travel_Business travel	0.02625
13	baggage_handling	0.02637
15	inflight_service	0.02789
5	ease_of_online_booking	0.03503
10	inflight_entertainment	0.03753
25	type_of_travel_Personal Travel	0.03760
0	customer_type	0.04782
3	inflight_wifi_service	0.13607
8	online_boarding	0.21548
21	customer_class_Business	0.25932

Figure 7. Feature importance graph and the feature importances' coefficient of Gradient Boosting Machine model

Figure 5 and **Figure 6** illustrate the feature importance graph and the importance coefficient of each feature that is used by the Random Forest classifier for distinguishing between the satisfaction level of airline passengers respectively. **Figure 5** demonstrates that there are three most important features that crucially affect the model on classifying the satisfaction level of airline passengers which are 'Online boarding' (0.162006), ' inflight_wifi_service' (0.140031), and 'type_of_travel_Personal Travel' (0.117382).

Table 4 indicates the sorted importance coefficient in ascending order of each of the features applied in the classification task by using logistic regression. The result from **Table 4** has shown that the features named 'online_boarding' (3.01), 'customer_type' (2.03), and 'inflight_wifi_service' (1.97) as the most influential feature on the logistic regression model to differentiate the satisfaction level of airline passengers.

Figure 7 depicts the feature importance graph and the following list of the feature importance's coefficient. As the result of **Figure 7**, the most influential features on the GBM classifier are named 'customer_class_business' (0.25032), 'online boarding' (0.21546), and 'inflight_wifi_service' (0.13607).

As a conclusion for this section, the features 'online_boarding_service' and 'inflight_wifi_service' were found as the most influential features that affect the decision making made by three of the classifiers when distinguishing between the satisfaction level of airline passengers. Thus, as a business-like suggestion for this airport's management department, it is recommended to maintain the quality of the wifi service provided in a flight and the online boarding service as both of them are highly associated with the satisfaction level of the airline passengers. In addition, the quality of other provided airport's services and the treatment for each of the customer flight classes are also recommended to be emphasized in order to positively influence the airline passengers' satisfaction.

6. Conclusion

This paper studied the quality of airline service and airline passengers' satisfaction by incorporating the supervised Machine Learning algorithm. The problem statement and the research objective conducted in this study are major to investigate the relationship between the satisfaction level of airline passengers and the airport's provided service and the evaluation of the Machine Learning algorithm's performance on classifying the satisfaction level of airline passengers. During the experiment, three different models are developed by using the Machine Learning algorithm which are the Random Forest algorithm, Logistics Regression, and Gradient Boosting Machine (GBM). During the classification models' construction, the computational cost to build the GBM models was found that relatively higher than the computational cost of building the Logistics Regression model and Random Forest model. Three of the classification models were fitted with the normalised and transformed training dataset which is prepared from the dataset's pre-processing stage. Based on the evaluation result of three classification models, the GBM model and Random Forest model have the greatest performance on distinguishing between the satisfaction level of airline passengers meanwhile the stability of the models' performance could be promised at 95% of confidence level. Other than that, we also have successfully determined the most common variables that affected the most on three of the classifiers to distinguish between the satisfaction level of airline passengers. Additionally, we have also identified the kinds of services that are either positively or negatively associated with the satisfaction level of airline passengers based on the correlation plot constructed.

7. Future Works

As the results of this study recommend that airlines company should focus on customers' needs for certain services especially the online boarding service and wifi service provided in a flight. Thus, the authors were interested to conduct a deep investigation of the relationship between those of the recommended service to be emphasized and the airline passengers consuming behavioural. Other than that, the authors are also interested to gain a deep insight into the satisfaction degree of both Malaysian and non-Malaysians to use the services provided by the Malaysia airline industry such as KLIA airport and so on by conducting an analysis of the satisfaction level of airline passengers towards the services provided by the local airline industry. Additionally, a correlation analysis of airline customer satisfaction using a deep neural network (DNN) and support vector machine (SVM) models was intended to compare customer satisfaction with existing airline services in future research.

Acknowledgements: The authors are pleased to thank the individuals who contributed to this study.

References

[1] G. C. Saha and Theingi, "Service quality, satisfaction, and behavioural intentions: A study of low-cost airline carriers in Thailand," Manag. Serv. Qual., vol. 19, no. 3, pp. 350–372, 2009, doi: 10.1108/09604520910955348.

[2] Halpern, N. (2018a). Airport business strategy. In N. Halpern, & A. Graham (Eds.), *The Routledge companion to air transport management* (pp. 154–170). Routledge: Abingdon. Halpern, N. (2018b).

[3] Rane A, Kumar A. Sentiment classification system of Twitter data for US airline service analysis. In: Proc: 42nd IEEE

computer software and applications conference, COMPSAC 2018. Tokyo, Japan. p. 769–73.

[4] WONG, J.-Y., & CHUNG, P.-H. (2008). Retaining Passenger Loyalty through Data Mining: A Case Study of Taiwanese Airlines. Transportation Journal, 47(1), 17–29. http://www.jstor.org/stable/20713696

[5] Javed Parvez, S., & Sahayadhas, A. (2020). Data Analytics for Monitoring the Satisfactory Parameters of Airline Passengers using Machine Learning Algorithms in Python. International Journal Of Innovative Technology And Exploring Engineering, 9(3), 1231-1235. doi: 10.35940/ijitee.c8677.019320

[6] Larose, D. (2015). Discovering Knowledge in Data: An Introduction to Data Mining (2nd ed., p. 22). Hoboken, New Jersey: John Wiley & Sons.

[7] Sedkaoui, S. (2018). Data Analytics and Big Data [Ebook] (p. 76). Hoboken, New Jersey: John Wiley & Sons.

[8] Insil Park, "Influence of Customer Satisfaction and Reuse Intention on Service Quality of Airline Outsourcing: Focusing on National Airlines", Tourism Management Research, Vol. 13, No. 39, pp.27-60, 2009.

[9] Czepiel, J. A., Rosenberg, L. J., Akerele, "Perspectives on consumer satisfaction", AMA Conference Proceedings, pp.119-123, 1997.

[10] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7. https://doi.org/10.3389/fnbot.2013.00021

[11] Omary, Z., & Mtenzi, F. (2010). Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning. International Journal for Infonomics (IJI), 3(3), 314-325.