Exploratory Data Analysis and Prediction of Passenger Satisfaction with Airline services

Maham Salah-Ud-Din
Faculty of Mathematics and Data Science
Emirates Aviation University
Dubai, United Arab Emirates
10055860@eau.ac.ae

Blessy Trencia Lincy.S.S
Faculty of Mathematics and Data Science
Emirates Aviation University
Dubai, United Arab Emirates
blessytrencialincy.simonshepherd@emirates.com

Hannah Al Ali
Faculty of Mathematics and Data Science
Emirates Aviation University
Dubai, The United Arab Emirates
hannah.alali@emirates.com

***Abstract*— The airline industry has bloomed greatly after the COVID-19. The customer satisfaction about the airlines is a key factor in the success of the airline industry. This study focuses on several aspects affecting the services of the airline and a comparative study is made by applying various algorithms like Gaussian Naïve Bayes, Gradient Boosting Classifier, Linear Support Vector Machine, Logistic Regression, Multilayer Perceptron Classifier and Random Forest Classifier. The result of our analysis to predict the customer satisfaction as satisfied or unsatisfied shows that the Gradient Boosting Algorithm achieved an accuracy of 96% and it is identified that the key feature of the analysis is the boarding process and Wi-Fi services.**

Keywords—* *Aviation, Exploratory Data Analysis, Classification, Machine learning, Prediction

I. Introduction

The aviation industry is constantly evolving and revolutionizing. It is turning the way it used to operate by bringing tech-driven transformations. The data that is being generated each day by the aviation industry is massive. Many important insights are hidden in this enormous data but handling a huge volume of data and extracting meaningful information from it is difficult. So, big data analytics comes to the rescue to get meaningful insights from the ever-growing aviation data. Those meaningful insights help in knowing the customer in a better way. As customers are the driving force of any industry, it's important to know if they are satisfied with the service or not. Their satisfaction will make them the airline's loyal customers, increasing the number of customers will increase the overall revenue. Using big data analytics and machine learning, airlines could know what factors are affecting the satisfaction level of the customer and how improving those factors could make their experience better throughout their travel journey. The aviation industry is becoming more competitive because air travel demand is increasing each day. To stand out in the market, airlines need to transform their way of operation to tech-driven by integrating big data analytics, AI, and machine learning in their operation. Aiding Big Data analytics and machine learning in the aviation industry will help airlines to understand their passengers' requirements in a better way and the Decision Support System can assist in making fruitful decisions for the businesses that will earn them more customers and benefit them.

II. Literature Review

Eunil Park conducted a study[1] in which he analyzed customer feedback data to identify the factors influencing passenger satisfaction with airline services. The sentiment analysis helps to identify the satisfaction level of customers. He noted that customers' emotions, social words, and monetary values are some factors that affect customer satisfaction. He assumed that the priorities of low-cost carrier and full-service carrier passengers would be different. Lestari, Y.D et al. discussed [2] that a binary logical statistical technique is beneficial to conduct statistical analysis to know the elements impacting satisfaction. Those elements are ticket prizes, passenger seat, cabin crew service, refreshments, and meals are the most significant factors that customer looks for in their entire journey. The statistical analysis also indicates that inflight entertainment and ground services are not primarily focused and prioritized by the customers to make them satisfied.

The aviation industry faced a significant decline during the coronavirus (COVID-19) pandemic. It impacted the aviation sector immensely. Making customers happy and satisfied during their travel journey can help airlines attain more customers and that's how Airlines can overcome those challenges. Random Forest Recursive feature elimination (RF-RFE)-Logistic feature selection model [3] is beneficial to know the factors that are impacting the satisfaction of the customers. The passenger's priorities changed during and after the COVID-19 for air travel. They often go for the airline that provides facilities like refunds of the ticket and has a clean cabin that gives satisfaction to the customer that they are in a clean environment. COVID-19 impacted every person mentally more than physically, and air travel was one of the main reasons[4] that spread it all around the world. That is why passengers prioritize cleanliness and disinfection rather than all other factors.

The technology has evolved tremendously. Customer feedback could be taken from social media instead of taking surveys. The most honest reviews are given on social media as it provides freedom to people to speak about anything in any manner. Online reviews can assist businesses in understanding the quality of services they are offering. They would help businesses to know about their performance and where they stand. They can aid businesses in identifying what factors are important for customers and how changing their services based on their feedback can satisfy them and benefit the business. Semantic Network Analysis was conducted [5] to know the meaningful insights from the reviews of passengers and analyze the relationship between frequent keywords. The study found out that seat comfort, ground services, airline name, ticket price, and staff service can help make the customers happy and satisfied by conducting Linear Regression analysis.

Most of the customers specifically tweet to express their experience about their air travel with a specific airline. So, analyzing the feedback from the tweets using machine learning helps to know the emotions behind the whole customer journey. The factors that influence the customer [6] could be selected by n-gram and Glove dictionary. Then the different kinds of emotions can be identified using an artificial neural network and support vector machine. Association analysis contributes to identify the reasons behind those emotions. However, customer satisfaction varies with countries. All region's people are quite different therefore their choices and priorities would differ. Aymeric Punel [7] noted that in-flight services are prioritized by the Asian people rather than the ticket price and it is vice versa for North American and South American people. However, seat comfort is a common factor that is important for passengers in all regions. This study helped businesses to know that their services should be according to the regions and understanding the people in those regions is significant to make them satisfied with the airline services. The airline industry is very competitive and demands top-notch quality service to lead other competitors. Knowing the features that will satisfy the passengers will help the airline to lead among all other airlines [8].

Bhargav conducted a study [20] using the Novel Hybrid Random Forest algorithm and KNN classifier algorithm to analyze the passenger satisfaction and compared their performance. Ouf in their study [21] used deep neural networks with the adaptive moment estimation Adam optimization algorithm to the airline passenger satisfaction dataset to increase the classification performance and made a comparison with random forests, artificial neural networks (ANNs), and support vector machine techniques. Gorzałczany in their study [22] used a modern fuzzy-genetic business-intelligence solution depicted both by high accuracy and high interpretability to the airline passenger satisfaction decision support system. In this section, the existing techniques and analysis upon the airline passenger satisfaction data is discussed.

III. Data

The dataset used in this study is "Passenger Airline Satisfaction" which is an open-source dataset taken from Kaggle[19]. The dataset consists of 103905 airline passengers' feedback and 24 attributes which includes both numerical and categorical attributes. Some of the categorical attributes in the dataset are 'Customer Type', 'Gender', 'Type of Travel', 'Class', and 'Satisfaction'. The target variable in our work is the attribute 'Satisfaction' which can be classified as satisfied or dissatisfied/neutral. This dataset considers the dissatisfied/neutral scenarios as same differentiating the satisfied scenario. Some important numerical features from the dataset are seat comfort, inflight entertainment, flight distance, Wi-Fi service, legroom service, check-in service, etc..

IV. Processing Framework

The overview of the processing framework is shown in Figure 1. The first step in the processing framework involves the preprocessing and cleaning of the dataset to ensure the data quality. The next step is to perform exploratory data analysis (EDA) to unveil correlations and patterns among attributes.

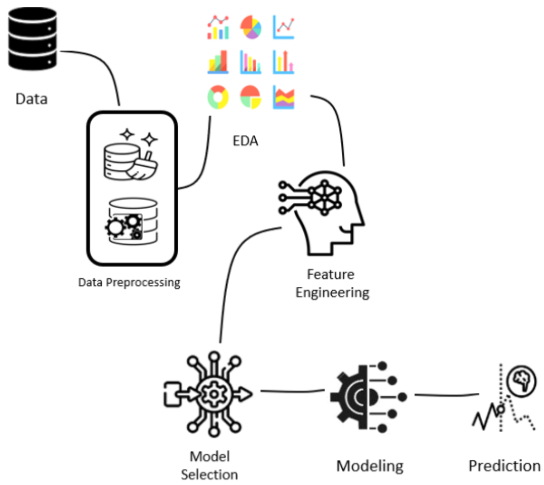

Fig. 1. *Overview of the Processing Framework*

The next step is to employ feature engineering techniques to transform important features into a suitable format. Finally, various machine learning models are selected and evaluated to determine the best fit for the data. We also analyzed and identified the important features that contribute significantly to the predictive performance of the chosen model.

V. Data Preprocessing

Data preprocessing is the first step in machine learning and data analysis. It is a data mining technique that transforms the raw data into a form that makes us understand it. Data preprocessing is the necessary step [9] that helps to increase the reliability of data, ease the mining process, and reduce the time the model needs to be trained. It involves removal of the incomplete, noisy, and inconsistent data. So, it improves the quality of the data that is selected for the analysis. Data integration is performed first if the data is collected from multiple sources. Data Cleaning is a vital step in data preprocessing because clean data is incredibly important for effective analysis. It involves removing the incomplete and incorrect data, removing the duplicate data, and handles the outliers in that data. If the data is not cleaned, the output of the machine learning model would be inaccurate, that is why it has significant importance. It helps to identify patterns and connect related information.

VI. Exploratory Data Analysis

The Exploratory Data Analysis [10] is a vital step that should be performed before modelling. The techniques of EDA reveal the true nature of the data. It helps to identify the relationships between each attribute. The better you understand the data, their relationships, and patterns, the better will be the analysis. It reveals the true nature of data and provides valuable information. The first thing that needs to be done in EDA is to evaluate whether the data is relevant to the problem and whether the data collected is enough or whether more data should be collected. EDA checks the values that lie far away from a standard set of values which are known as outliers. EDA uncovers the behaviour of variables and their correlation. The correlation plays vital role in locating crucial variables on which other variables depend.

A. *Target Attribute*

The pie chart represented in the Figure 1 below shows the distribution of the target class as ‘satisfied’ or ‘unsatisfied’.

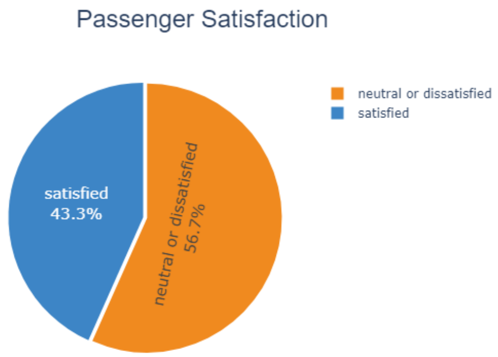

Fig. 2. *Passenger Satisfaction*

The chart shows that the dataset is slightly imbalanced as 43.3% of passengers are satisfied whereas 56.7% of the passengers are not satisfied by the air travel experience.

B. *Visualizing Categorical Attributes*

Gender Distribution: The gender distribution is represented using pie plot in Figure 3 which shows that 49.3% of male and 50.7% of female records are present in the dataset.

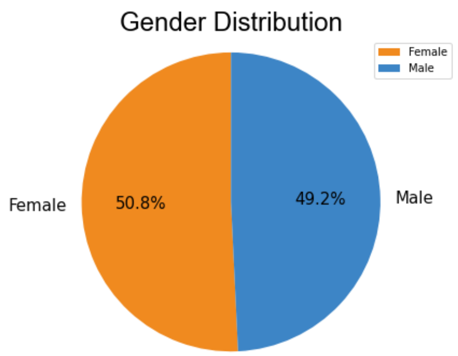

Fig. 3. *Gender Distribution*

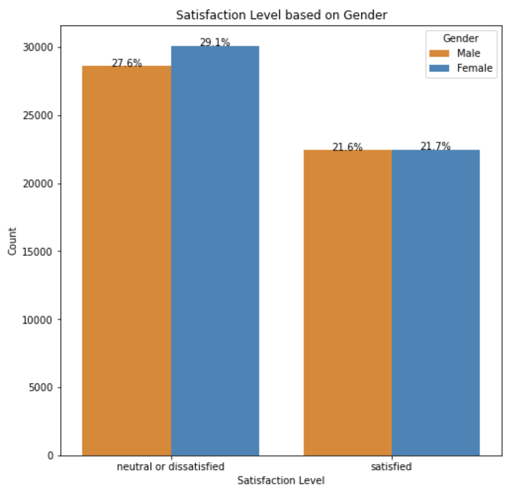

Fig. 4. *Satisfaction level based on gender*

The bar chart in Figure 4 visualizes how many males and females are satisfied and how many are dissatisfied with the airline service. It shows that more females are dissatisfied by the air travel experience but still, there is a slight difference only. So, satisfaction level based on gender is quite balanced, and satisfaction of passengers does not seem to depend on gender.

Customer Type Distribution: The pie chart in Figure 5 shows that 81.7% of passengers are loyal customers and only 18.3% of the passengers are those who rarely travel with the particular airline. The Figure 6 shows that out of 18% of disloyal customers, only 4% of them are satisfied with the airline service. Out of 81% of loyal customers, only 39% of them are satisfied.

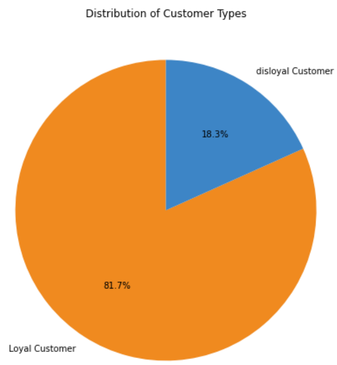

Fig. 5. *Customer Type Distribution*

So, in both customer types, a greater number of people are dissatisfied. But loyal customers are more satisfied on comparison with the disloyal customers.

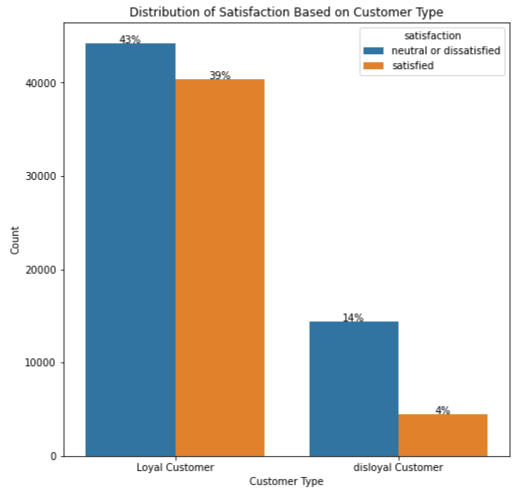

Fig. 6. *Distribution of Satisfaction based on Customer Type Distribtution*

Class Distribution: The bar chart in Figure 7 shows that most satisfied passengers are from business class and most dissatisfied people are from economy class. The class difference majorly affects the satisfaction level of the passengers.

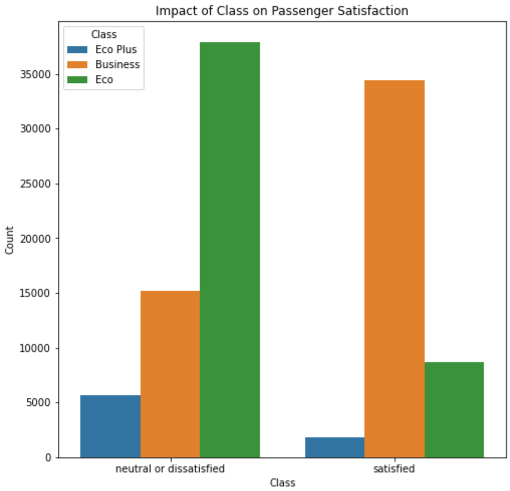

Fig. 7. *Class Impact on Passenger Satisfaction*

Business travellers opt for business class; they enjoy amenities and hence become more satisfied. But economy class travellers are mostly personal travellers who opt for a budget-friendly class and the business class is about the comfort.

Type of Travel Distribution: The pie chart in Figure 8 shows that 69% of people travel for business purposes whereas only 31% of people travel for personal reasons.

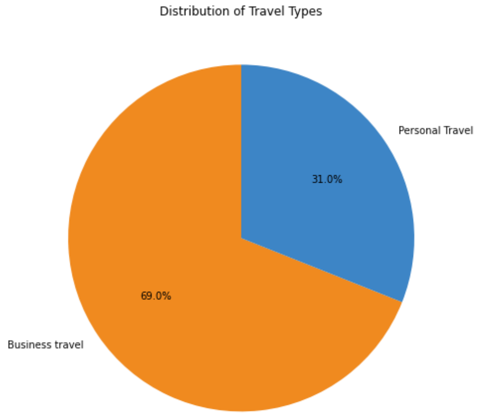

Fig. 8. *Distribution of Travel Types*

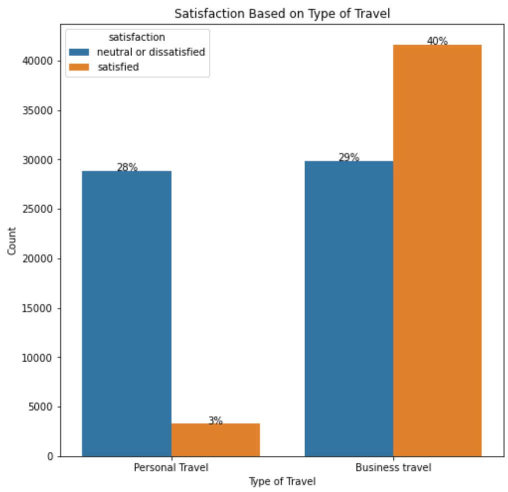

Fig. 9. *Distribution of Satisfaction based on Type of Travel*

The bar chart in Figure 9 shows that out of 69% of people traveling for business purposes, 40% of them are satisfied whereas only 3% of people out of 31% are satisfied that are traveling for personal purposes. It depicts that business travellers are more satisfied than personal travellers.

Age Distribution: The bar chart in Figure 10 shows that most people who are satisfied with the overall airline experience are in between the age group of 40 – 60 and the least people satisfied are in between the age group of 0-18 years. The youth and elder people have least satisfaction rate. So, the age varies with the satisfaction level of the passengers.

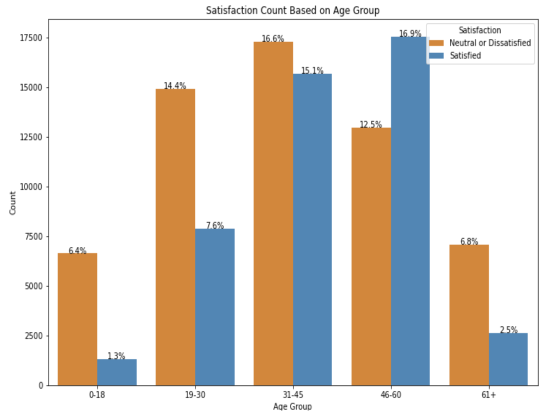

Fig. 10. *Satisfaction count based on Age group*

Other Factors: The horizontal bar chart in Figure 11 shows that baggage handling and inflight services contribute the most to satisfy the passengers. The class in the Figure 11 represents the business class, economy class or economy plus class. Even though the satisfaction level is different for each class, baggage handling and inflight services are the most crucial factors for each class. After these two, on-board services and extra legroom seat have great significance for business class passengers. Inflight entertainment and cleanliness are important for economy-plus class passengers. On-board and check-in service is important for economy class passengers.

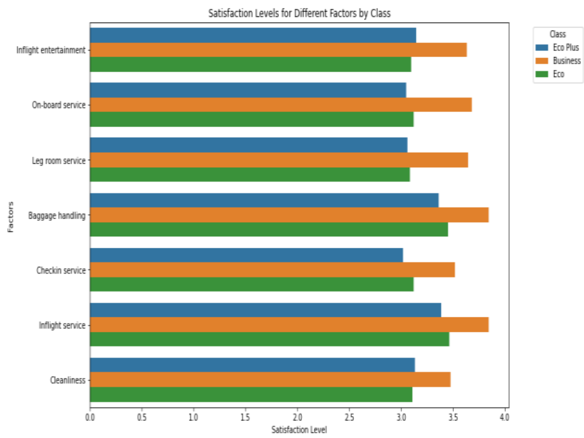

Fig. 11. *Satisfaction Levels for different factors*

C. *Feature Engineering*

Feature engineering is the concept of using the domain knowledge to chose and transform the data for creating the model for our evaluation.

1) Encoding

The encoding technique is applied to categorical variables. The categorical variables can't be directly given to the machine learning algorithm. One-hot encoding is a type of nominal encoding whereas Label Encoding is a type of ordinal encoding.

Label Encoding: Label Encoder is simple for transforming the target class that has two categories only. As the 'satisfaction' attributes have only two classes, 'satisfied' and 'neutral or dissatisfied', label encoding was the best of all encoding techniques. It has a clear mapping of 0 and 1 for each category. In this case, 'neutral or dissatisfied' is represented as 0, and 'satisfied' is represented as 1.

One-hot Encoding: One-hot encoding[11] transforms categorical variables into zeros and ones. The length of these vectors is equal to the number of classes or categories that our model is expected to classify. One-hot encoding is suitable to transform Gender, Customer Type, Type of Travel and Class because they have nominal categories, and the order doesn't matter for each one of them. One-hot encoding is represented by zeros for absence and one's for presence which helps machine learning algorithm to understand that they don't have any ordinal relationship among them.

2) Train-Test Split

The sklearn library is used to split the data into training set and test set and to apply the machine learning models. The dataset is split in 70:30 ratio where 70% of the data will be used for training the model and 30% of the data will be used for testing the model to evaluate its performance.

D. Model Selection

We selected six models which includes Random Forest Classifier, Gradient Boosting Classifier, Linear SVC, Gaussian Naïve Bayes, MLP Classifier, and Logistic Regression to identify which one best fits to our dataset.

a) Random Forest:

A machine learning model that uses an ensemble of decision trees to make its predictions is a Random Forest Model Classifier [12]. It takes a random sample of the data and then builds an ongoing series of decision trees on the subsets. So, a whole bunch of decision trees are created together which gives a larger group or model. The more decision trees used with different criteria, the better the random forest model will perform because they essentially increase the precision accuracy, and the ones that are not relevant and uncertain are ignored.

b) Gradient Boosting

Gradient Boosting is a supervised learning algorithm. It tries to find the best function that minimizes the expected loss between our labels and the predictions given by the function [13]. It is an ensemble method that performs classification by combining the outputs from individual decision trees. It is suitable for our data as it prevents overfitting and handles complex relationships in an efficient way. It helps to know which feature has contributed the most in prediction.

c) Linear SVC

A Linear SVC is a supervised learning algorithm that can divide data into two categories. Each object that needs to be classified is represented as a point in an n-dimensional space and the coordinates of these points are called features. It performs the classification test by drawing a hyperplane that is a line in 2D or 3D [14]. Linear SVC is selected as it is a straightforward choice for binary classification problems. It efficiently deals with noisy data.

d) Gaussian Naïve Bayes

Gaussian Naïve Bayes is a probabilistic classification algorithm. It is based on Bayes theorem which considers the features of the data to be conditionally independent. It means that the presence or absence of one feature does not affect the presence or absence of another one. The probability of each class is calculated [15], and it selects the highest probability as the predicted class as shown in (1).

$$P(A|B) = [P(B|A)P(A)]/P(B) \tag{1}$$

where, P(A|B) = Posterior Probability (Probability of event after event B is true) and P(A) = Prior Probability (Probability of event happening).

e) MLP Classifier:

MLP Classifier is selected because it is a supervised learning algorithm that can be trained on the 'satisfaction' target class [16]. It is a neural network that has many hidden layers. The number of inputs will be based on a number of features with an additional node of bias. All nodes of the input are connected to the inner layer known as the activation layer or hidden layer. These number of hidden layers depends on how deep the network is and each of these activations would be either identity [f(x) = x], logistic[f(x) = 1/(1+exp(-x))], tanh [f(x) = tanh(x)] or relu [f(x) = max (0, x)]. The activations are connected to the output layer.

f) Logistic Regression

Logistic Regression is similar to linear regression but specifically for classification problem [17]. Logistic refers to the log odds probability that is modelled. The term odds is the ratio of the probability that an event occurs to the probability that it doesn't occur. It is suitable for our dataset as we need to predict the probability of customer's satisfaction or dissatisfaction.

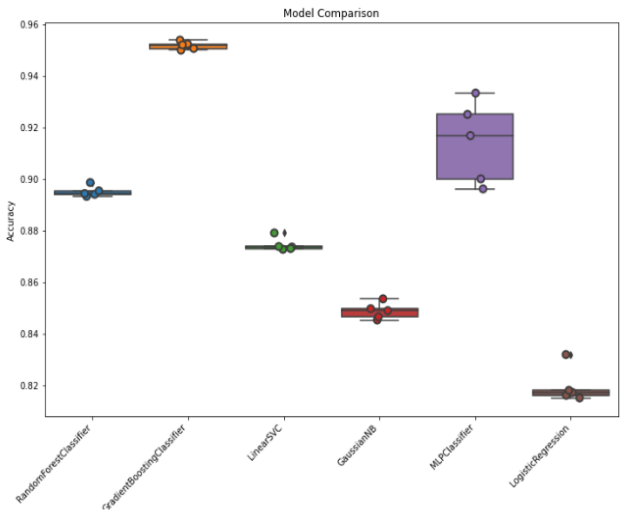

Fig. 12. *Model Comparison*

All models are evaluated and compared, and their accuracy is stored as data frame. Then the model is visualized using a boxplot that compares each model accuracy as shown in Figure 12. Then, the mean accuracy of each model is calculated which shows that Gradient Boosting is the best for our dataset by providing highest accuracy of 95%.

TABLE I. MODEL ACCURACY

Model Name	**Accuracy**
GaussianNB	0.848756
GradientBoostingClassifier	0.951677
LinearSVC	0.874385
LogisticRegression	0.819623
MLPClassifier	0.914213
RandomForestClassiifer	0.895071

E. Modeling

As the Gradient Boosting performed the best among all the models, it is selected for modeling, The Gradient Boosting is about taking a model that is a weak predictive model by itself. Combining that model with other models of the same type to produce an accurate model is a gradient-boosting model. In Gradient Boosting, boosting refers to combining weak learners sequentially to achieve a strong learner. The weak learners are the decision trees in gradient boosting. Each tree attempts to minimize the errors of the previous tree. Adding many trees in sequence on which each decision tree focuses on the errors from the previous one, makes boosting a highly efficient and accurate model. In every iteration, a new tree is added, and it fits on a modified version of the initial dataset. Since the trees are added sequentially, the boosting algorithm learns slowly. As it learns slowly, it performs better. The final model aggregates the result from each step and a strong learner is achieved [13, 18]. The algorithm is based on the following formula (2):

$$F(x)=F0(x)+\eta h1(x)+\eta h2(x)+\ldots+\eta hm(x) \qquad (2)$$

Where, $F(x)$ is the final model's prediction, $F0(x)$ is the initial model, η is the learning rate, controlling the contribution of each weak learner, $hm(x)$ represents the weak learner at the m-th iteration.

F. Feature Importance

The features that contributed the most to predict the model are visualized using the bar plot shown in Figure 13. The plot shows that online booking and inflight Wi-Fi service contributed the most to the model's predictions. The higher score implies that those features have a greater impact on the model that is being used.

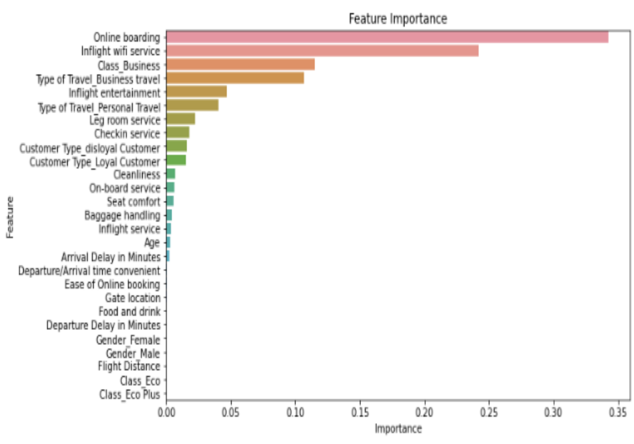

Fig. 13. *Feature importance*

G. Results

The confusion matrix is used to evaluate the performance of the Gradient Boosting Algorithm. It shows that 13713 times the model correctly classifies the 'satisfied' instance and 18496 times the model correctly identifies the 'dissatisfied' instance.

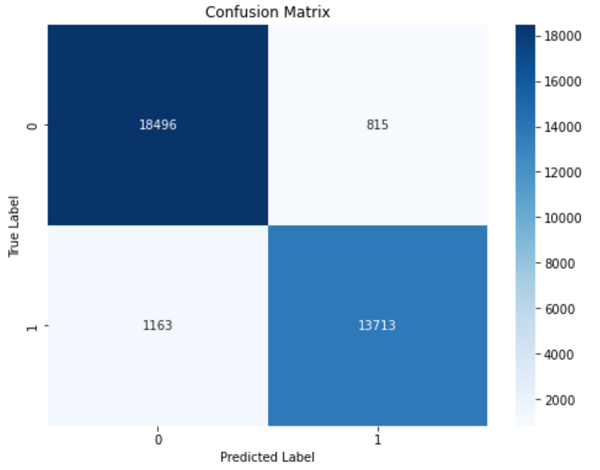

Fig. 14. *Confusion Matrix*

TABLE II. CLASSIFICATION REPORT

	Precision	Recall	F1-score	Support
Neutral and Dissatisfied(0)	0.94	0.96	0.95	19311
Satisfied(1)	0.94	0.92	0.93	14876
Accuracy			0.94	34187
Macro avg	0.94	0.94	0.94	34187
Weighted avg	0.94	0.94	0.94	34187

The classification report provides the performance metrics which shows that the model is 94% accurate means it correctly classified whether the passenger is satisfied or not 94% of the time. The ratio of true positives to the total prediction is also 94%(Precision). The model correctly identifies 96% of actual 'satisfied' instances and 94% correctly identifies 'neutral and dissatisfied' instances (Recall).

CONCLUSION

By comparing various models, the Gradient Boosting Algorithm achieved an accuracy of 96% based on precision, recall, and F1-score metrics. By performing Exploratory Data Analysis, we observed that the class difference majorly affects the satisfaction level of passengers. The most satisfied passengers are traveling from business class for business purposes. As they enjoy amenities from the business class, they tend to be more satisfied, but the economy class passengers don't enjoy those luxuries which makes them dissatisfied. Through EDA, we also got to know that inflight service and baggage handling are the most important factors to make any class passengers satisfied i.e., either business, economy plus, or economy. The feature importance after modeling shows that online boarding i.e. online check in and inflight Wi-Fi service contribute the most to the satisfaction of the passenger.

ACKNOWLEDGMENT

We thank the Emirates Aviation University for supporting us in performing this analysis and study.

REFERENCES

[1] Park, E., Jang, Y., Kim, J., Jeong, N.J., Bae, K. and Del Pobil, A.P., 2019. Determinants of customer satisfaction with airline services: An analysis of customer feedback big data. Journal of Retailing and Consumer Services, 51, pp.186-190.

[2] Lestari, Y.D. and Murjito, E.A., 2020. Factor determinants of customer satisfaction with airline services using big data approaches. Jurnal Pendidikan Ekonomi Dan Bisnis (JPEB), 8(1), pp.34-42.

[3] Jiang, X., Zhang, Y., Li, Y. and Zhang, B., 2022. Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model. Scientific Reports, 12(1), p.11174.

[4] Pereira, F., Costa, J.M., Ramos, R. and Raimundo, A., 2023. The impact of the COVID-19 pandemic on airlines' passenger satisfaction. Journal of Air Transport Management, p.102441.

[5] Ban, H.J. and Kim, H.S., 2019. Understanding customer experience and satisfaction through airline passengers' online review. Sustainability, 11(15), p.4066.

[6] Kumar, S. and Zymbler, M., 2019. A machine learning approach to analyze customer satisfaction from airline tweets. Journal of Big Data, 6(1), pp.1-16.

[7] Punel, A., Hassan, L.A.H. and Ermagun, A., 2019. Variations in airline passenger expectation of service quality across the globe. Tourism management, 75, pp.491-508.

[8] Tian, H., Presa-Reyes, M., Tao, Y., Wang, T., Pouyanfar, S., Miguel, A., Luis, S., Shyu, M.L., Chen, S.C. and Iyengar, S.S., 2021. Data analytics for air travel data: a survey and new perspectives. ACM Computing Surveys (CSUR), 54(8), pp.1-35.

[9] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 1-22.

[10] Wongsuphasawat, K., Liu, Y., & Heer, J. (2019). Goals, process, and challenges of exploratory data analysis: An interview study. arXiv preprint arXiv:1911.00568.

[11] Saxena, S. (2023) What are categorical data encoding methods: Binary encoding, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/ (Accessed: 22 January 2024).

[12] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[13] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

[14] DataTechNotes (2020) Classification example with Linear SVC in python, Classification Example with Linear SVC in Python. Available at: https://www.datatechnotes.com/2020/07/classification-example-with-linearsvm-in-python.html (Accessed: 22 January 2024).

[15] Naiem, S., Khedr, A. E., Marie, M., & Idrees, A. M. (2023). Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in the Detection of DDOS in Cloud Computing. IEEE Access.

[16] DataTechNotes (2020) Classification example with Linear SVC in python, Classification Example with Linear SVC in Python. Available at: https://www.datatechnotes.com/2020/07/classification-example-with-linearsvm-in-python.html (Accessed: 22 January 2024).

[17] Logistic regression in machine learning (2023) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/understanding-logistic-regression/ (Accessed: 22 January 2024).

[18] Wizards, D.S. (2023) Understanding the gradient boosting algorithm, Medium. Available at: https://medium.com/@datasciencewizards/understanding-the-gradient-boosting-algorithm-9fe698a352ad (Accessed: 22 January 2024).

[19] Klein, T. (2020) Airline passenger satisfaction, Kaggle. Available at: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?resource=download (Accessed: 31 January 2024).

[20] Bhargav, B., & Prabu, R. T. (2023, April). Airline Passenger Satisfaction Prediction Using Novel Hybrid Random Forest Model Comparison with K-Nearest Neighbour Model. In 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) (pp. 1-6). IEEE.

[21] Ouf, S. (2023). An Optimized Deep Learning Approach for Improving Airline Services. Comput. Mater. Contin, 75(1), 1213-1233.

[22] Gorzałczany, M. B., Rudziński, F., & Piekoszewski, J. (2021). Business Intelligence in airline passenger satisfaction study—A fuzzy-genetic approach with optimized interpretability-accuracy trade-off. Applied Sciences, 11(11), 5098.