Classical Machine Learning For Airline Passenger Satisfaction : Evaluative Study

Ness Ichhaporia
12th Grade, Delhi Public School Surat Surat, India

DOI:10.37648/ijrst.v14i04.005

[1]*Received: 23 September 2024; Accepted: 01 November 2024; Published: 08 November 2024*

ABSTRACT

In the highly competitive aviation industry Customer satisfaction is key to building brand loyalty andreputation. The airline therefore gives importance to every touchpoint. From booking to baggage collection to exceed passenger expectations and stand out in the market. We have used 4 best-known classical machine learning models: Random Forest, LightGBM, Catboost, XGBoost and compared them in order to find the best model. To further investigate we used SHAP for qualitative analysis. In our research we found out that the most important feature contributing to customer satisfaction is type of travel.

Keywords—*customer satisfaction; classical machine learning; SHAP*

INTRODUCTION

The airline industry is one of the most important sectors in the travel industry. It facilitates tourism, trade, connectivity, generates economic growth, provides jobs, improves living standards, etc. Airline industry runs on many factors [5]. Oneof the key factors for a reputable airline industry is its customer satisfaction.

Customer satisfaction is a broad concept, as its meaning varies for everyone. For some, it may depend on on-board service, while for others, it may relate to the ease of onlinebooking. In our research with the help of machine learning we have incorporated a lot of features like age, gender, gatelocation, food, etc. to study and examine customer satisfaction.

Machine Learning (ML) is a branch of Artificial intelligenceand computer science that focuses on the using data and algorithms to enable AI to imitate the way humans learn, gradually improving its accuracy.

An ML model functions as a trained program designed todetect patterns within data and generate predictions. Essentially, these models are mathematical functions thatprocess data inputs to deliver specific outputs. To identifythe most effective model, we've conducted a direct comparison of several ML models, evaluating their performance based on AUC (Area Under the Curve) and accuracy.

The area under the ROC curve (AUC) represents the probability that the model, if given a randomly chosen positive and negative example, will rank the positive higherthan the negative.

Finally, we have used the SHAP values to further explain thecontribution of each feature.

RELATED WORK

A. A Logistic Regression Model of Customer Satisfaction of Airline

This paper used logistic regression to develop customer satisfaction model for Precision Air [3]. Five dimensions or variables have been considered: on time performance; aircraft safety; schedule integrity; on board services; and customer service.

While in our research we have incorporated a total of 23 features which are leg room service, cleanliness, departure delay in minutes, etc. Wehave also used 4 ML models for our research insteadof only one which are Random Forest,

[1] How to cite the article: Ichhaporia N.; November 2024; Classical Machine Learning For Airline Passenger Satisfaction : Evaluative Study; *International Journal of Research in Science and Technology*, Vol 14, Issue 3, 47-53, DOI: http://doi.org/10.37648/ijrst.v14i04.005

LightGBM, Catboost, XGBoost.

B. Determinants of customer satisfaction with airlineservices: An analysis of customer feedback big data

In this study, Structural equation modeling method (SEM) is used in the proposed research model revealing that customers' affective values have notable effects on their satisfaction with airline service. Structural equation modeling is a multivariate, hypothesis-driven technique that is based on a structural model representing a hypothesis about the causal relations among several variables [4].

In our study, we have used in combination several ML Learning models to find the most efficient feature correlated to airline passenger satisfaction.

IMPLEMENTATION

A) Dataset

This dataset examines airline passenger satisfaction surveysto find key factors influencing passenger satisfaction and dissatisfaction [1]. By analyzing various aspects of the travel experience as detailed in Table 1. The goal is to identify which factorsare most strongly associated with satisfaction levels. and to predict overall traveler satisfaction based on these insights.

Table 1

Features	Meaning
Gender	Gender of the passengers
Customer Type	The customer types
Age	The actual age of the passengers
Type of Travel	Purpose of the flight of the passengers
Class	Travel class in the plane of the passengers
Flight distance	The flight distance of this journey
Inflight Wi-Fi service	Satisfaction level of the inflight Wi-Fi service
Departure/Arrival time convenient	Satisfaction level of Departure/Arrival time convenient
Ease of Online booking	Satisfaction level of online booking
Gate location	Satisfaction level of Gate location
Food and drink	Satisfaction level of Food and drink
Online boarding	Satisfaction level of online boarding
Seat comfort	Satisfaction level of Seat comfort
Inflight entertainment	Satisfaction level of inflight entertainment
On-board service	Satisfaction level of On-board service
Leg room service	Satisfaction level of Leg room service
Baggage handling	Satisfaction level of baggage handling
Check-in service	Satisfaction level of Check-in service
Inflight service	Satisfaction level of inflight service
Cleanliness	Satisfaction level of Cleanliness
Departure Delay in Minutes	Minutes delayed when departure
Arrival Delay in Minutes	Minutes delayed when Arrival
Satisfaction	Airline satisfaction level

B) Correlation Matrix

Correlation matrix helps us to measure the relationship between every feature of the dataset. In this dataset the correlation of every feature is explained by the word satisfaction. The range of the matrix is from –1 to 1, where 1 means the perfect positive correlation, 0 as no correlation and –1 as perfect negative correlation. The correlation matrix is then passed through a heatmap function. Lighter color represents positive correlation whereas darker color represents negative correlations. Neutral correlations show mid tone color

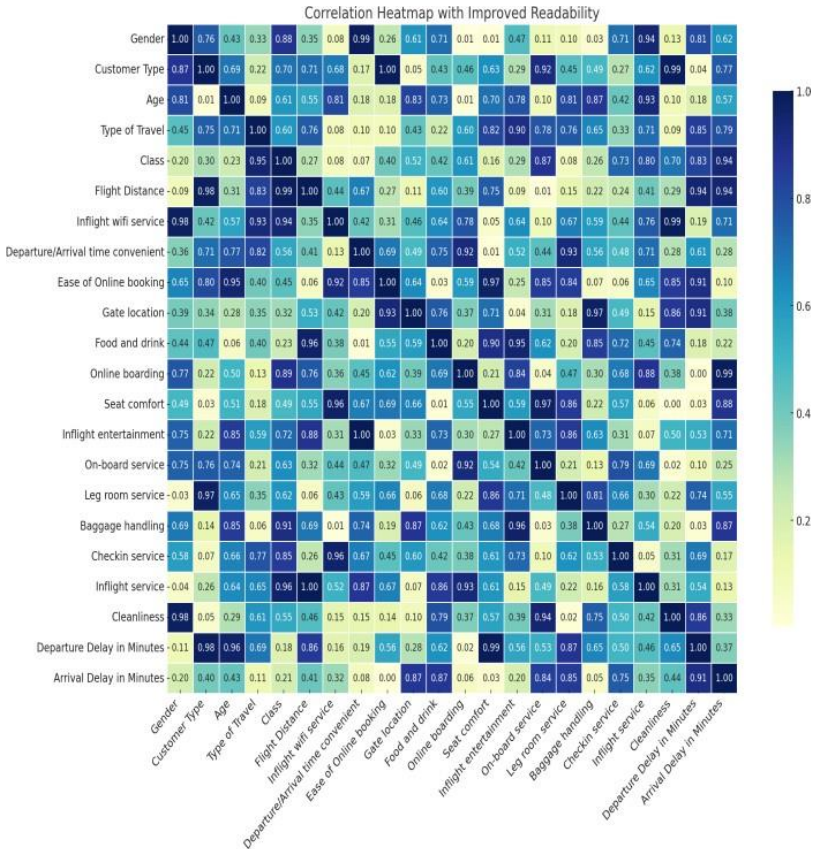

Fig 1. Correlation Matrix

According to this matrix we can observe that arrival delay in minutes is highly correlated to departure delay in minutes.

C) Data Refining and Standardizing

In this study, a data preprocessing pipeline was facilitated to transform categorical variables into numerical representations to facilitate model training [6]. The datasetcontains a total of 129880 entries and 23 columns. There were 25976 NaN, NaN represents missing values. The missing values were filled in with 'Arrival Delay in Minutes" with the median.

Several categorical features in the dataset, such as gender, customer type, travel class, and satisfaction were transformedinto numerical values. Every feature was transformed into custom numerical values. Gender "Female" was mapped to 1, "Male" to 0, and missing value were mapped to –1. Hence the range of standardizing ranged from –1 to 1.

We have used Standard Scaler to normalize the dataset. Standard scaler normalizes a dataset by transforming the features thus they have a mean value of 0 and a standard deviation of 1 [7]. It ensures that every feature is contributing equally to the model and prevents dominance of one feature in the model.

D) Parameters for Random Forest

Parameters	Value
max_depth	25
min_samples_leaf	1
min_samples_split	2
n_estimators	1200
random_state	42

Table 2. Default parameters for Random Forest

Max depth will limit the depth of the trees and will help prevent overfitting. Min samples leaf defines the minimum number of samples required to be in a leaf node. Min samples split allows the tress to split if there are at least 2 samples. N_estimators defines the number of decision treesin the random forest. Random state is a seed used by the random number generator. It ensures that the model's resultis reproducible.

E) Parameters for LightGBM

Parameters	Value
colsample_bytree	0.85
max_depth	15
min_split_gain	0.1
n_estimator	200
num_leaves	50
reg_alpha	1.2
reg_lambda	1.2
subsample	0.95
subsample_freq	20

Table 3. Default parameters for LightGBM

colsample_bytree specifies the fraction of features to be randomly selected for each tree. Max depth is the maximum depth of each decision tree in the model. min_split_gain is the minimum gain required to make a split in a tree. n_estimator is the number of boosting trees in the model. num_leaves is the maximum number of terminal nodes in each tree. reg_alpha adds a penalty to the absolute values ofleaf weights in the objective function. reg_lambda helps control overfitting by shrinking the coefficients. Subsample is the fraction of training data that is to be randomly sampledfor each tree. subsample_freq specifies how often to perform subsampling.

QUANTITATIVE ANALYSIS

We have compared the value of precision, recall, f1-score,accuracy and AUC of all the different models. The table beneath displays precision, recall, f1-score, accuracy and AUC. Accuracy provides us with the ratio of the sum of true predictions to that of total samples and AUC gives us the ration of TPR and FPR.

Model	Precision	recall	F1-score	Accuracy	AUC
Random forest	0.95556	0.97969	0.96747	0.96304	0.96072
LightBGM	0.95747	0.98099	0.96909	0.96489	0.96265
Catboost	0.95856	0.97777	0.96807	0.96381	0.96187
XGBoost	0.93695	0.95951	0.94810	0.94106	0.93849

Table 4

(IJRST) 2024, Vol. No. 14, Issue No. 4, Oct-Dec **e-ISSN: 2249-0604, p-ISSN: 2454-180X**

Table 4 shows that with default parameters LightBGM generated the highest AUC being 0.96265. Among all the models the best performing model is LightBGM as the AUCof LightBGM is highest of all the models.

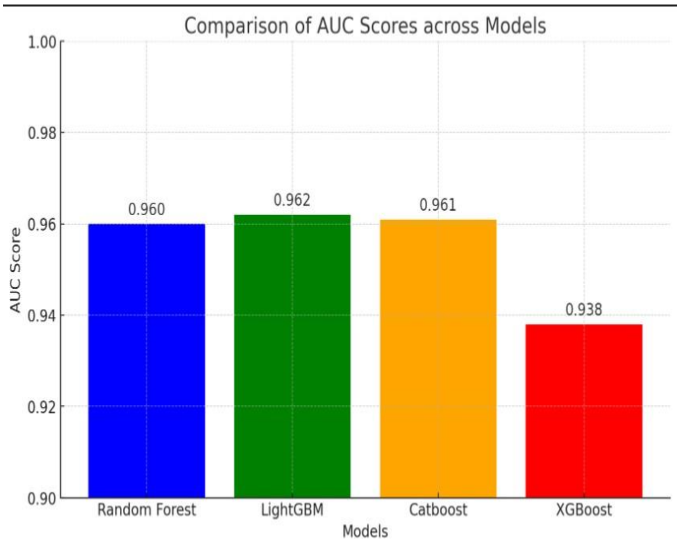

Fig 2. Bar Graph of all models comparing AUC Scores

We can further observe that accuracy of LightBGM is the highest as compared to the other models. It is shown in fig 3 below.

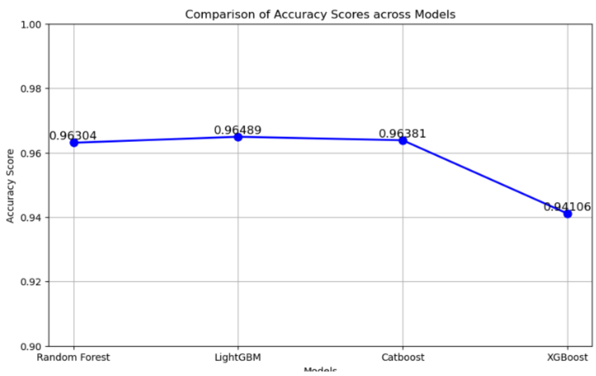

Fig 3. Line Graph of accuracy of all models

QUALITATIVE RESULTS

In machine learning, model explainability is a crucial attribute that helps us understand how different features contribute to predicting the outcome.

SHAP (Shapley Additive explanations) values are a way toexplain the output of any machine learning model. It uses agame theoretic approach that measures each player's contribution to the outcome [2]. In machine learning, each feature is assigned an importance value representing its contribution to the model's output. SHAP values show us how each feature affects each final prediction, the significance of each feature compared to others, and the model's reliance on the interaction between features.

A bar graph shows us the contribution of each feature inpredicting the factor that affects the most in customer satisfaction. We have displayed SHAP values of the topmodel for further explanation which is LightBGM.

A. SHAP Graph for LightBGM

This is the SHAP graph of LightBGM, our best performingmodel by applying the correlation matrix.

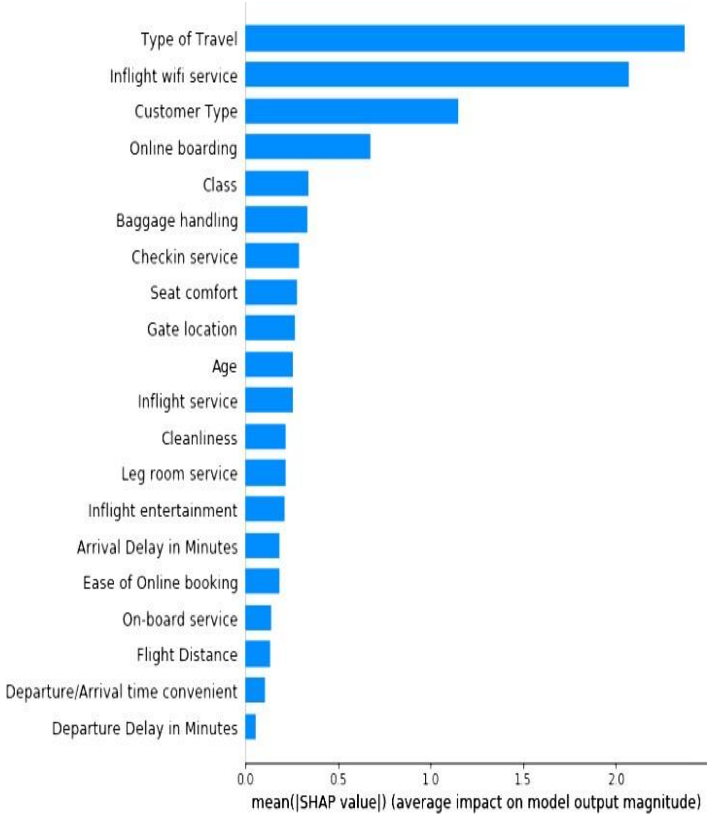

Fig 4. Bar Graph of LightBGM

From Fig 4 we can interpret that type of travel is our highestcontributing feature to customer satisfaction as it should be, because if a person buys business class, then the person would get better service, more space, more comfortable seat,better food than the one in economy class. The second highest feature is inflight Wi-Fi service as many people like to access internet in flight for various activities like entertainment, work, etc.

ACKNOWLEDGMENT

This research project was conducted independently in pursuit of higher education. It is not associated with anyschool and is not part of a school curriculum.

CONCLUSION

In this paper, we used 4 ML models: Random Forest, LightBGM, Catboost, XGBoost [8]. We have compared all the models, and the result is that LightBGM is the best performing model. Across every experiment, LightBGM hasshown consistent results of a well performing model. We have used SHAP to further explain the contributions of eachfeature to predict satisfaction level of passengers. In our research we have analyzed that the top 2 features affecting customer satisfaction are type of travel and inflight Wi-Fi service.

REFERENCES

1. https://www.kaggle.com/datasets/teejmahal20/airline-passenger- satisfaction

2. Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D. and Groh, G., 2022, October. SHAP-based explanation methods: a review for NLP interpretability. In *Proceedings of the 29th international conference on computational linguistics* (pp. 4593-4603).

3. Josephat, Peter, and Abbas Ismail. "A logistic regression model of customer satisfaction of airline." *International Journal of Human Resource Studies* 2.4 (2012): 197.

4. 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.

5. Apampa, Olatunji. "Evaluation of classification and ensemble algorithms for bank customer marketing response prediction." Journalof International Technology and Information Management 25, no. 4 (2016): 6..

6. A. Ghatnekar and A. D. Shanbhag, "Explainable, Multi-Region Price Prediction," 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 2021, pp. 1-7, doi: 10.1109/ICECETS2533.2021.9698641.

7. Ali, Peshawa Jamal Muhammad, Rezhna Hassan Faraj, Erbil Koya,Peshawa J. Muhammad Ali, and Rezhna H. Faraj. "Data normalization and standardization: a technical report." Mach Learn Tech Rep 1, no. 1 (2014): 1-6.

8. Patel, Krish, and Aakash Shanbhag. "Exploring ML for Predictive Maintenance Using Imbalance Correction techniques and SHAP." In 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), pp. 1-10. IEEE, 2022.

.