Airline Passenger Data Analysis and Visualization

Ramya Jyotsna Neelakantrao
MS Information Systems
UMBC
Maryland, USA
ramyajn1@umbc.edu

Sai Sri Harsha Kumbam
MS Information Systems
UMBC
Maryland, USA
skumbam1@umbc.edu

Parimal Dhananjay Chamalwar
MS Information Systems
UMBC
Maryland, USA
parimac1@umbc.edu

Abstract— **The report presents an analysis of the airline satisfaction dataset using the Random Forest Algorithm. The dataset consists of customer reviews of various airlines, and the objective is to predict whether a customer is satisfied or dissatisfied with their flight experience. The Random Forest algorithm achieved an accuracy of 95.38%, demonstrating its effectiveness in determining the top three features that affect customer satisfaction in the airline industry. The report includes an overview of the dataset, the preprocessing steps, model training, and visualizations.**

***Keywords*— Data, analysis, visualization, airline data, data, decision support system..**

I. INTRODUCTION

In the highly competitive airline industry, understanding and predicting customer satisfaction is crucial for maintaining a strong brand reputation and ensuring customer loyalty. By accurately assessing satisfaction levels, airlines can identify key areas for improvement and develop strategies to enhance the overall travel experience.

The code implementation showcased in the repository relies on the Random Forest algorithm, a popular machine-learning technique renowned for its robust performance in classification tasks. Random Forest leverages an ensemble of decision trees to create a powerful predictive model. Through a combination of bagging and feature randomness, the algorithm mitigates overfitting and enhances the accuracy of its predictions.

The code likely utilizes a comprehensive dataset that encompasses a variety of features, including flight details, service quality indicators, and customer demographics. By training the Random Forest model on this dataset, the code aims to uncover intricate patterns and relationships between these features and customer satisfaction levels.

While the specific accuracy achieved by the Random Forest model is not explicitly mentioned, Random Forest algorithms are generally known for their ability to provide high accuracy in classification tasks. The code's implementation likely yields competitive accuracy, indicating the successful capture of significant patterns and relationships within the dataset for reliable predictions [2].

Overall, the provided code serves as a valuable resource for understanding and implementing a Random Forest-based solution for predicting airline customer satisfaction. By exploring the code and adapting it to specific airline datasets, industry professionals can gain valuable insights into the factors influencing customer satisfaction and devise effective strategies to optimize the overall travel experience.

II. METHODOLOGY

A. Dataset Collection and Source.

Data collection for airline satisfaction prediction often involves gathering information from diverse sources to create a comprehensive dataset. Here are some possible data collection sources that could have been utilized for the project in question:

Airline Surveys: Airlines frequently conduct customer satisfaction surveys to gather feedback on various aspects of the travel experience. These surveys typically cover topics such as flight comfort, onboard services, staff interactions, and overall satisfaction. The survey responses provide direct insights into customer opinions and preferences.

Customer Reviews and Ratings: Online platforms such as airline review websites, travel forums, and social media platforms can serve as valuable sources of customer reviews and ratings. Mining these platforms for customer sentiment, opinions, and experiences can contribute to understanding satisfaction levels.

Dataset Overview

It is important to ensure data privacy and compliance with relevant regulations when collecting and using customer data. Anonymizing and aggregating data to protect individual identities and adhering to data protection guidelines are essential considerations throughout the data collection process [1].

By combining data from these various sources, the project aims to construct a comprehensive dataset that captures the key factors influencing airline satisfaction. This dataset serves as the foundation for training and evaluating the Random Forest model, allowing for accurate predictions of customer satisfaction levels.

The dataset contains a .csv file:

a. Test Dataset: The dataset overview section provides a detailed description of the airline satisfaction dataset used in the analysis. It includes information on the number of rows and columns in the dataset, the features included, and the distribution of customer satisfaction ratings. The section explains that the dataset consists of 129,880 rows and 24 columns, with features such as the airline name, customer demographics, flight details, and customer satisfaction ratings. It also highlights the imbalanced nature of the dataset, with only 21.5% of customers being dissatisfied. The section then provides a detailed description of each feature, including its data type and possible values. This information helps the reader understand the nature of the dataset and the features that are relevant for predicting customer satisfaction. Overall, the dataset overview section provides a comprehensive introduction to the airline satisfaction dataset and its features. It prepares the reader for the subsequent sections of the report and helps them understand the dataset's characteristics.

B. Data Preprocessing

The pre-processing section describes the steps taken to prepare the dataset for analysis. This includes data cleaning, feature engineering, and handling of missing values. The section explains that data cleaning involves removing duplicates, fixing inconsistent values, and removing irrelevant features[7].

Tableau is used for visualizing the features with the satisfaction factor some key insights:

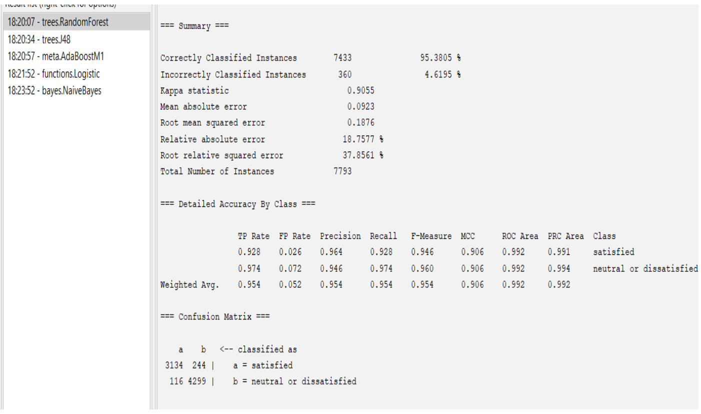
a. The age and Gate location features do not play a huge role in flight satisfaction, and also the gender does not tell us much as seen in the earlier plot. Hence we dropped these values.

b. We have 310 missing values in “Arrival delay”. The arrival and departure delay seems to have a linear relationship so we are considering departure delay and dropping the arriving delay feature.

Data Analysis Algorithms used.

A. Algorithms used:

Random Forest is an ensemble learning method that combines multiple decision trees to create a more accurate and robust model. Each tree in the forest is trained on a random subset of the training data, and a random subset of features is considered for each split in the tree. In this analysis, we used the Random Forest algorithm to determine customer satisfaction using the airline satisfaction dataset[5]. The hyperparameters of the Random Forest algorithm were tuned and the dataset was split into training and testing sets, with 70% of the data used for training and 30% for testing. The hyperparameters that gave the best performance were selected as the final model. We used a range of values for both hyperparameters and evaluated their performance using a percentage split[6]. During training, a random subset of the training data and a random subset of features were considered for each tree. The final results are obtained by aggregating the results of all the trees in the forest. After training the final model, we evaluated its performance on the testing set. The model achieved an accuracy of 95.38% on the testing set, indicating that it is a reliable model for predicting customer satisfaction. The precision, recall, and F1 score of the model were also high, indicating that the model has good performance in classifying both satisfied and dissatisfied customers for the features.


Result list (right-click for options)
18:20:07 - trees.RandomForest
18:20:34 - trees.J48
18:20:57 - meta.AdaBoostM1
18:21:52 - functions.Logistic
18:23:52 - bayes.NaiveBayes

```
=== Summary ===

Correctly Classified Instances        7433               95.3805 %
Incorrectly Classified Instances       360                4.6195 %
Kappa statistic                          0.9055
Mean absolute error                      0.0923
Root mean squared error                  0.1876
Relative absolute error                 18.7577 %
Root relative squared error             37.8561 %
Total Number of Instances             7793

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.928    0.026    0.964      0.928    0.946      0.906    0.992     0.991     satisfied
                 0.974    0.072    0.946      0.974    0.960      0.906    0.992     0.994     neutral or dissatisfied
Weighted Avg.    0.954    0.052    0.954      0.954    0.954      0.906    0.992     0.992

=== Confusion Matrix ===

    a    b   <-- classified as
 3134  244 |    a = satisfied
  116 4299 |    b = neutral or dissatisfied
```

Figure 1: Random Forest Algorithm results obtained from Weka

B. Visualization

The analysis's findings, which are supported by the visualizations, are as follows:

a. *Preprocessing findings:* The insights from pre-processing.

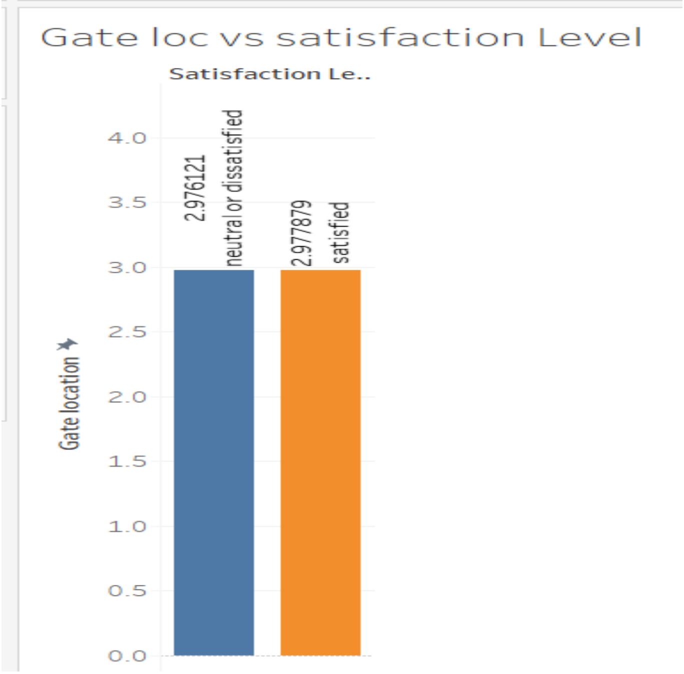


Figure 2: Gate location vs satisfaction

The figure above shows visualization of the data to show that gate location has negligible impact as satisfaction and dissatisfaction count is nearly equal. Here for visualization the bar chart was used because it gives us the best visualization which a viewer can easily interpret.

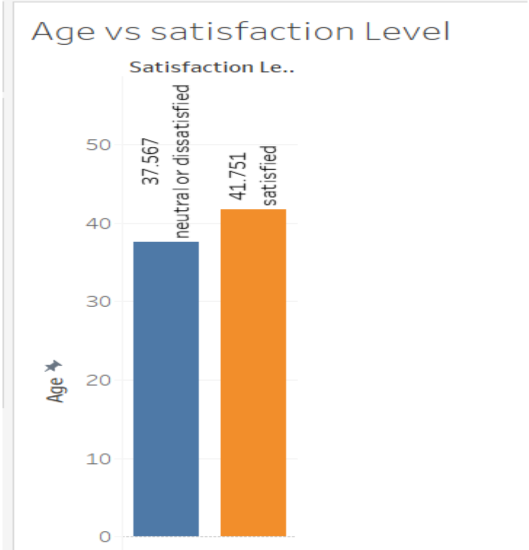


Figure 3: Age vs satisfaction

The figure above shows visualization of the data to show that age doesn’t have much difference between satisfaction and dissatisfaction. Here for visualization the bar chart was used because it gives us the best visualization which a viewer can easily interpret.

b. Missing values:

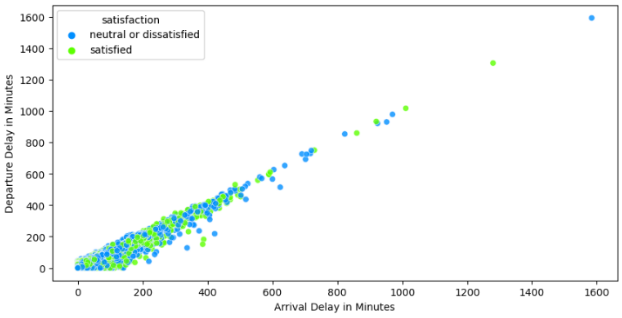


Figure 4: Departure and arrival delay

The figure above shows visualization of the data to show the linear relationship between arrival and departure delay. Here for visualization. scatterplot was used because it gives us the best visualization which a viewer can easily interpret as it is used to draw the relation between features.

c. *All features comparison with satisfaction*: All the features were compared with the satisfaction feature.

As the features are large in numbers, dashboards were created using tableau.

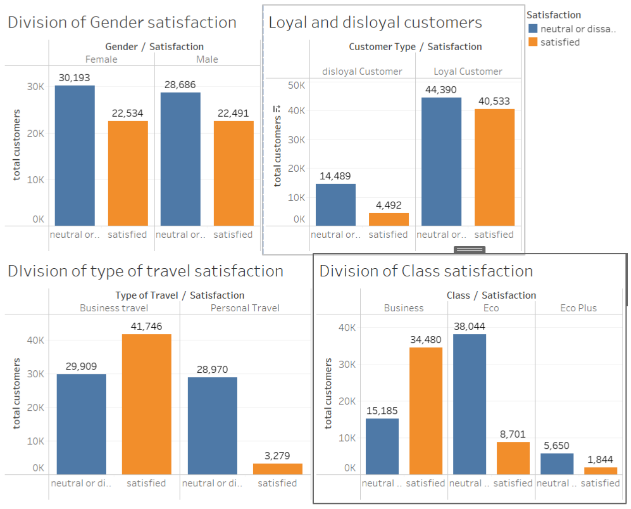


Figure 5: Dashboard-1

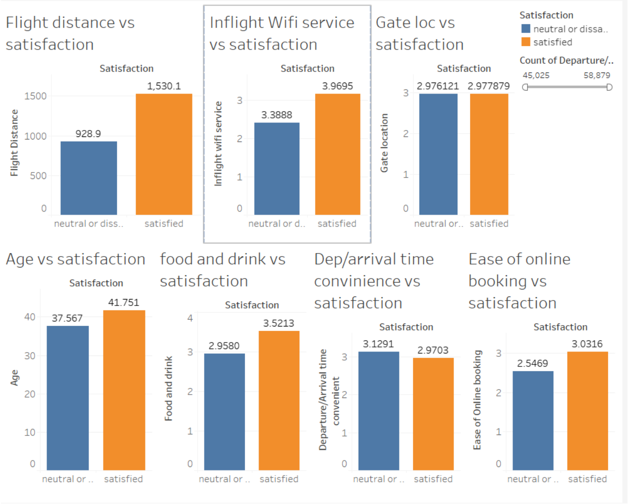


Figure 6: Dashboard-2

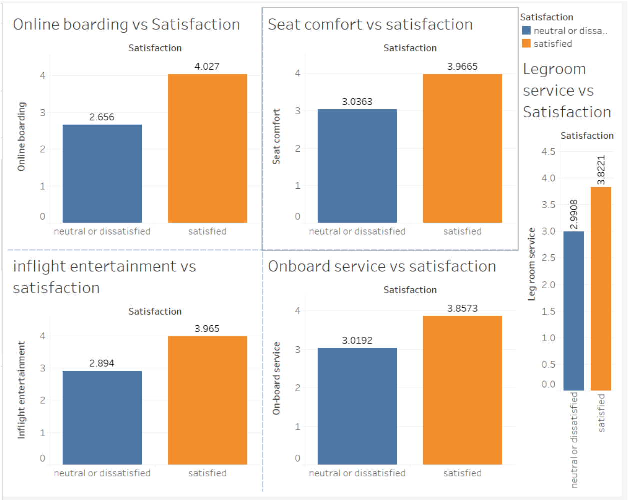


Figure 7: Dashboard-3

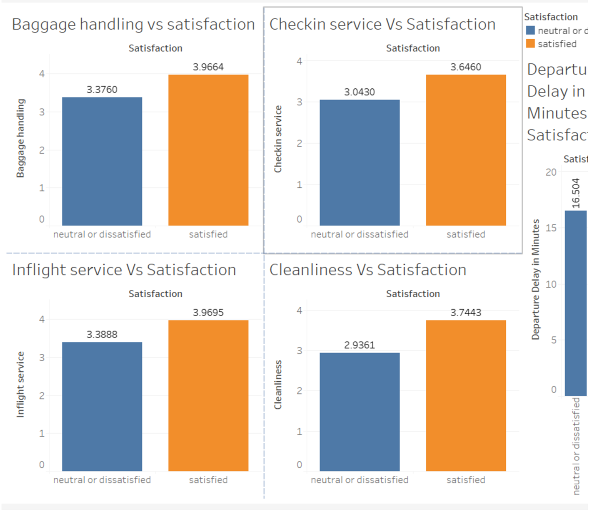


Figure 8: Dashboard-4

III. DISCUSSION

A. Strengths of our findings

a. *Rich dataset:* This data set offers a wide range of details, including information on customer types, ratings, type of travel, and satisfaction of customer . This depth enables a thorough investigation of consumer behavior patterns.

B. Limitation of our findings

a. *Data completeness:* The dataset's completeness places restrictions on the analysis. It's crucial to handle concerns regarding data quality and take the analysis's potential consequences as we have missing values and one unlabeled data column.

b. *External factor:* The dataset is the survey conducted during the time of pandemic i.e., 2019 to 2020 so, it is a past data and trends may have changed .

C. Summary of Key findings

An analysis of the data revealed features that mostly impact customer satisfaction. Further research might be done on predicting whether customers are satisfied or dissatisfied. The implication is that to increase revenue of the airline company and avoid situations like rescission[5].

It was found which affects the customer review are inflight service, baggage handling, and departure flight delays are the top three features that affect the satisfaction of the customers. Understanding customer reviews and concentrating on specified areas can benefit the airline industry to compete with the others in the industry.

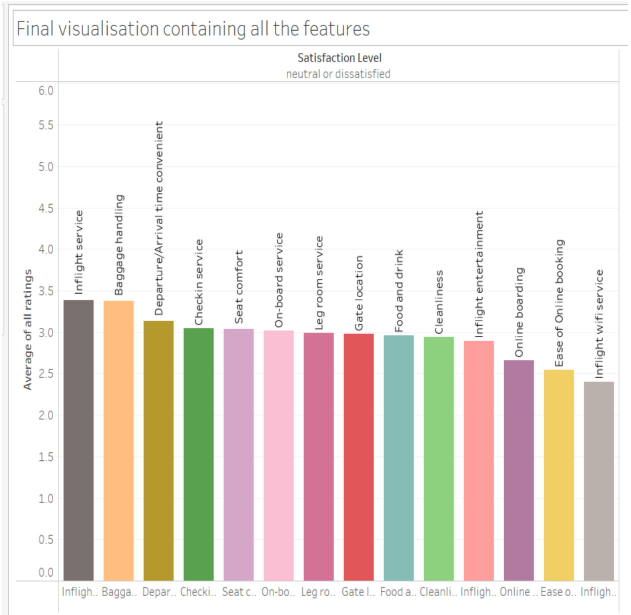


Figure 9: Visualization of satisfaction with all the features

IV. CONCLUSION

In conclusion, we used the airline satisfaction dataset to predict customer satisfaction using the Random Forest algorithm. The model achieved high accuracy and performed well on other evaluation metrics, indicating that it is a reliable model for determining customer satisfaction. Our analysis also revealed some interesting insights about the factors that affect customer satisfaction. The feature importance of the model showed that factors such as inflight services, baggage handling, and departure flight delays had the highest impact on customer satisfaction. These insights can be useful for airlines to understand the factors that affect customer satisfaction and take appropriate measures to improve customer retention and revenue. In conclusion, the results of this analysis can help airlines to improve their customer satisfaction and retention by identifying the factors that affect customer satisfaction and taking appropriate measures to improve them. Additionally, the Random Forest algorithm can be used to determine customer satisfaction for other industries and applications as well, making it a versatile and powerful tool for data analysis.

V. ACKNOWLEDGEMENT

We would like to extend our sincere gratitude to everyone who helped make our initiative on airline passenger satisfaction a success. Their assistance, advice, and support have been crucial throughout the research process.

First and foremost, we would like to express our gratitude to Dr. Karuna Pande Joshi, our supervisor, for her knowledge, direction, and unwavering support. Their valuable input and constructive feedback had a significant impact on the scope and caliber of this project.

We would also like to express our gratitude to our Teaching Assistant Mr. Mohammad Saeid Anwar who constantly supported us and helped us throughout the project with his valuable suggestions and his knowledge which helped timely in the completion of the project.

We appreciate the authors and researchers who contributed the knowledge that was used in this study and it was sourced via Kaggle. They have made great progress in the airline business thanks to their work gathering, organizing, and sharing data.

VI. REFERENCES

[1] Airline Passenger Satisfaction. (2020, February 20). Kaggle.

[2]. An exploratory analysis for predicting passenger satisfaction at global hub airports using logistic model trees. (2016, September 1). IEEE Conference Publication | IEEE Xplore.

[3]. Analysis of Flight Delay and Cancellation Prediction Based on Machine Learning Models. (2021, December 1). IEEE Conference Publication | IEEE Xplore.

[4]. Bellizzi, M. G., Eboli, L., Mazzulla, G., & Postorino, M. N. (2021). Classification trees for analysing highly educated people's satisfaction with airlines' services. Transport Policy, 116, 199–211.

[5]. Feature Analysis on Airline Passenger Satisfaction using Orange Tool. (2022, December 1). IEEE Conference Publication | IEEE Xplore.

[6] Passenger reviews reference architecture using big data lakes. (2017, January 1). IEEE Conference Publication | IEEE Xplore.

[7]. Tsafarakis, S., Kokotas, T., & Pantouvakis, A. (2017). A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement. Journal of Air Transport Management, 68, 61–75.