**FINAL REPORT**

College Of Information, University of North Texas

Haihun, Chen

March 13th, 2022

Project name**: Air Passenger Satisfaction**

**Abstract**

In the past 20 years, the air transportation industry has been growing rapidly. This growth of the industry provides more opportunities as well as challenges. While the opportunities come because of increasing demand, equivalent airlines pose threat. Apart from maximizing the prices, have you ever thought what airlines do to overcome these troubles? **Air Passenger Satisfaction**. Unhappy passengers mean fewer customers and less earnings or income. Therefore, it is important that passengers should have a great experience every time they travel. The satisfaction survey from consists of train and test datasets and we have passenger reviews. The dataset contains the combination of categorical and continuous variable which has been used for this study. The models have been built using decision trees, logistic regression, and Random Forest regression concepts. This study seeks to not only explain the paramount factors which impact the passenger satisfaction in the Airline industry but also change in those factors across different age groups. We used machine learning concepts and jupyter notebook software to predict and clean the data that is data processing and data manipulation respectively.

We have implemented the Airline passenger Satisfaction Predictor which utilizes customer flight data to determine drivers that influence the satisfaction level of airline clientele with the aim of increasing overall customer satisfaction. The provided flight data is obtained from Kaggle which contains the data sourced from a survey that included each customer’s rating of their experience along with various other variables belonging to the customer’s reviews, behavior, history, and specific flight information.

**Introduction**

Air transportation in all over the world has been growing fast because of rise in low-cost carriers. Now a days People are more likely to travel with air transport because flight tickets are cheaper than it used to be. Airport as a destination and transfer to passengers is playing an important role in air transportation. Airport industry is facing rapid changes as well. Most of the Passengers are used to fail in choosing the right airport destination, as mostly routes are determined by airlines. But however, now a days most of the passengers could enjoy choosing their airport destination freely due to easy understand of the flights and having many options to book the flights in a cheaper way. Airports need to response rapidly to the changes of air passenger to meet their demands. Seyanon (2012) stated that air passengers have various options among airports and airport marketers that has been increasing changes among themselves to meet the need of customers satisfaction better than in the competition.

In that view, passengers’ satisfaction is increasingly important that may affect in reputation and long-term profits of airports. For example, a business will get successful and profitable when passenger is satisfied, therefore, passengers must be satisfied. Otherwise, Business cannot run without customer. Eboli and Mazzulla (2009) explains there are two basic concepts of customer satisfaction surveys. One is Expectation is what customers expect from the service and the other perception is what customers receive from the service.

**Methodology**

The methodology used for this project is divided into 4 stages shown in the diagram below.

**Logistic Regression**

**Concatenating train and test dataset**

Train Data Set

Test Data Set

**Decision Tree**

**classifier**

**Random Forest Classifier**

**Figure 1: Workflow of the project**

**3.1. Data preparation/collection:**

The dataset for this project is obtained from Kaggle which contains the train and test dataset from a survey conducted by airlines on the satisfaction level of passengers/customers based on various factors. The dataset consists of 25 columns such as Age, Gender, Travel class, Arrival and Departure delays and features that influences customer satisfaction level such as On-board service, Cleanliness, Seat comfort, Baggage handling etc.  
The dataset consists of a column or feature named ***‘*Satisfaction*’*** which describes the overall satisfaction level of the customer. It has two values they are neutral or dissatisfied and satisfied. This **Satisfaction** feature is considered as the label feature since it conveys the overall experience of the customer based on the ratings given for other features. The dataset consists of 103904 and 25976 records in train and test respectively.

**3.2 Data Preprocessing / EDA:**

Before going to Data cleaning first we have loaded the train and test data set into the jupyter notebook software and then combined both train and test datasets into one data frame using pandas and concatenation concept concat() function.

From here we have implemented the data cleaning process for the combined dataset checked all the missing values and duplicate values in the data. After checking for the null values, we have seen that Arrival delay in minutes columns has null values, so we are dropping the row and then we are dropping duplicates columns which are unnamed columns in the data. We have encoded the dependent variable as 0: unsatisfied and 1: satisfied. We used Correlation concept in this

before and after changing the categorical values in the data.

To visualize the data, we need to convert all the categorical values to the int format so that we can plot all the visualizations and graphs which we need. Replacing 0 and 1 in the place of the categorical values and ready for analysis and predict with the feature.

Next we are dividing the train and test using X\_train, X\_test and Y\_train, Y\_test to predict the data using different classification and regression models.

**3.3 Model Selection:**

Three models have been used in this project to check for the maximum efficiency and accuracy and to evaluate the how much accuracy is getting on each model for the satisfaction field. we are doing analysis on these three algorithms to find out which model is best to predict this data.

* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier

**Logistic Regression:**

We used Logistic regression models to analyze passenger satisfaction and qualitative measures, by expressing the dependent variable as the passenger’s overall satisfaction, by taking the logistic predictors as the attributes related to the service quality and computing the customer reviews of every attribute on the passenger’s overall satisfaction.  The data should be independent of each other, and it is a supervised machine learning algorithm. It can only take discrete values for a set of features(columns) in the data and the number of categories used in the data. Using test and train datasets we have calculated the accuracy for this model, so we got 67% accuracy rate after doing the prediction for the satisfaction field.

Graphical user interface, text, application

Description automatically generated

**Decision Tree Classifier:**

Decision tree classifier is used based on the small chunks of the data which means splitting the data into separate partitions according to the attribute selection features. Target variable should be categorical that means satisfaction feature should be in categorical values. Then we tried to predict the model by importing the packages for this model. Using test and train datasets we have concluded that the algorithm has scored accuracy of 95% for this model for the satisfaction field.

Graphical user interface, application

Description automatically generated

**Random Forest Classifier:**

Suppose dataset has m features and the random forest will select the random features k where k>m from subset of training data with the corresponding labels. We need basic parameters to predict this model like minimum split, split criteria etc. By predicting this model, we got the accuracy around 96% for the satisfaction field.

Graphical user interface, text, application

Description automatically generated

**Data Visualizations**

In today’s world everyday huge amount of data is being generating day by day but sometimes if we want to analyze the data for certain trends, patterns, and graphs it may become difficult to understand if the data is in raw format. So, this data visualization gives us good and better understand and observe the data We have done some visualizations using our combined dataset by comparing the two columns.

In our project we have made some visualizations to understand the data in better way and can compare the graphs between two columns.

**Created boxplot to find out how many loyal and non-loyal customers are there using customer type:**

From the below picture we can see loyal customers are more than non-loyal customers using customer type field which means it gives positive response from the customers. As we are getting to know some information from this visualization in better way.

**Chart

Description automatically generated with medium confidence**

**Created box plot to find out how many are satisfied and unsatisfied with the journey:**

From the below picture we can identify that most of the people are unsatisfied with the journey we can say may be this is because of the behavior of the airhostess or communication between them or not having the proper facilities like seat comfort, food and drink, cleanliness etc. From this we can understand that we need to improve our facilities to get positive response from customer.

**Graphical user interface

Description automatically generated with medium confidence**

**Created boxplot to find out how many satisfied and unsatisfied based on the Age:**

From the below diagram we can see that 20 - 40 age most of the people are not satisfied with the journey as it shows in the graph with red color which is highest and from 40-60 age people are more satisfied with the journey shows in green color. But people who are less than or equal to 20 and greater than 60 are least satisfied.

**Chart, bar chart

Description automatically generated**

**Conclusion:**

We have predicted our data by using the customer reviews and satisfaction field to check whether the people are satisfied or unsatisfied using their feedback. By using three models’ logistic regression, decision tree classifier and random forest classifier. We have concluded that from these three models **Random Forest classifier** model gives the high accuracy with 96% so random forest fits the best model of our project for prediction.

**References:**

1. <https://www.upgrad.com/blog/random-forest-classifier/>
2. A. S. Rao, B. V. Vardhan and H. Shaik, "Role of Exploratory Data Analysis in Data Science," 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021, pp. 1457-1461, doi: 10.1109/ICCES51350.2021.9488986.
3. Biswas, S., Khare, N., Agrawal, P. et al. Machine learning concepts for correlated Big Data privacy. J Big Data 8, 157 (2021). <https://doi.org/10.1186/s40537-021-00530-x>
4. Haldun Akoglu, User's guide to correlation coefficients,Turkish Journal of Emergency Medicine,Volume 18, Issue 3,2018,Pages 91-93,ISSN 2452 2473, [https://doi.org/10.1016/j.tjem.2018.08.001.(https://www.sciencedirect.com/science/article/pii/S2452247318302164)](https://doi.org/10.1016/j.tjem.2018.08.001.(https:/www.sciencedirect.com/science/article/pii/S2452247318302164))
5. Rokach, L., Maimon, O. Z., Rokach, L., &amp; Maimon, O. 2008, Data mining with decision trees: Theory and applications (New Jersey: World Scientific)
6. Dey, N. 2019, Social Network Analytics: Computational Research Methods and techniques (London: Academic Press)