Airline Passenger Referral Prediction

Sanchit Misra, Mohit Jain,
Tushar Hande.
Data science trainees,
AlmaBetter- Bangalore.

Abstract:

Nowadays consumers are increasingly relying on online resources to aid in purchasing decisions. Online resources also help companies to learn from customer reviews, how different types of customers have different priorities, and how customer choice affects their reviews. Now the power of peer opinion has taken to the skies too, with the rise of websites that allow passengers to post and read comments on airlines, airports, and even individual seats on particular planes. According to research, 91% of 18-34-year-olds trust online reviews as much as personal recommendations, and 93% of consumers say that online reviews influenced their purchase decisions. In this project, we have performed predictive analysis for qualitative reviews on the data provided. We have also performed an explanatory analysis of the different classes of airline services.

Keywords: Text mining, Natural language processing, Recommendation, Accuracy.

Problem Statement:

Data includes airline reviews from 2006 to 2019 for popular airlines around the world with multiple choice and free text questions. Data is scraped in Spring 2019. The main objective is to predict whether passengers will refer the airline to their friends.

Data description:

The dataset contains airline names, author information, overall rating, and recommended column. The dataset contained 131895 attributes and 17 features. Recommended is the target variable and the other 16 are independent variables.

More about the variables:

- airline: Name of the airline.
- overall: Overall point is given to the trip between 1 to 10.
- author: Author of the trip
- reviewdate: Date of the Review customer review: Review of the customers in free text format

- aircraft: Type of the aircraft
- travellertype: Type of traveler (e.g. business, leisure)
- cabin: Cabin at the flight date flown: Flight date
- seatcomfort: Rated between 1-5
- cabin service: Rated between 1-5
- foodbev: Rated between 1-5 entertainment: Rated between 1-5
- groundservice: Rated between 1-5
- valueformoney: Rated between 1-5
- recommended: Binary, target variable.

Steps involved:

1. Exploratory data analysis (EDA):

Exploratory Data Analysis (EDA) is an approach to analyzing the data using visual techniques. It is used to discover trends, and patterns, or to check assumptions with the help of statistical summary and graphical representations.

Some insights from the data using EDA:

- The range for the overall rating is from 1 to 10.
- The month feature's highest peak is 7
- Spirit Airlines with a maximum frequency in the dataset holds the top spot for most journeys taken followed by American and United airlines.
- Airbus A320 Aircraft with a maximum frequency in the dataset holds the top spot for most journeys taken followed by Boeing 777 and Airbus A380 aircraft.
- Bangkok to Hong Kong journey with a maximum frequency in the dataset holds the tops position followed by Bangkok to London and London to New York.
- There is a slight variation between recommended and not recommended in the recommended column.

2. Data Preprocessing

2.1 Handling duplicates:

When an entry appears more than once, it receives a disproportionate weight during training. Thus, models that succeed on frequent entries will look like they perform well, while in reality, this is not the case. Additionally, duplicate entries can ruin the split between train, validation and test set in cases where identical entries are not all in the

same set. This can lead to biased performance estimates that will lead to disappointing models in production.
We will remove duplicates to prevent overfitting.

2.2 Handling Missing values: Many machine learning algorithms fail if the dataset contains missing values. We may end up building a biased machine learning model which will lead to incorrect results if the missing values are not handled properly. Missing data can lead to a lack of precision in the statistical analysis.
Two primary ways to handle missing values.

1. Deleting the missing values
2. Imputing the missing values

2.3 Encoding categorical variables: The machine learning algorithms can't understand categorical data. It requires all independent and dependent variables i.e input and output features to be numeric.
Some of the ways to transform the categorical data.

1) Label encoding: The label encoder assigned a number to each category.
2) One Hot encoding: Dummy or one hot encoder convert each category to a binary column that takes the value 0 or 1.

We applied both ways to our data.

3. Text Mining

3.1 POS Tagging- Part-of-speech (POS) tagging is a Natural Language Processing the process that refers to categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context.

3.1 Tokenization − The given document is treated as a string and recognized as a single word in the document i.e. the given document string is split into one unit or token.

3.2 Removal of Stop word − In this process the removal of constant words such as a, an, but, and, of, the, etc. By removing these words, we remove the low-level information from our text in order to give more focus to the important information

3.3 Lemmatization: Lemmatization entails reducing a word to its canonical or dictionary form. The root word is called a 'lemma'. The method entails assembling the inflected parts of a word in a way that can be recognized as a single element. lemmatization allows end users to query any version of a base word and get relevant results.

3.4 Term Frequency Inverse Document Frequency (TF-IDF): Text from data needs to be transformed into a vector. The process of transforming text into a vector is referred to as text vectorization. Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector. It combines 2 concepts, Term Frequency (TF) and Document Frequency (DF).

Tf matrix calculated as follows:

$$tf(t) = \frac{No.of\ times\ term\ 't'\ occurs\ in\ s\ document}{Frequency\ of\ most\ common\ term\ in\ the\ document}$$

IDF matrix Calculated as follows:

$$idf_i = \log(\frac{n}{df_i})$$

TF-IDF is the multiplication of the term frequency matrix with its IDF.

$$w_{i,j} = tf_{i,j} \times idf_i$$

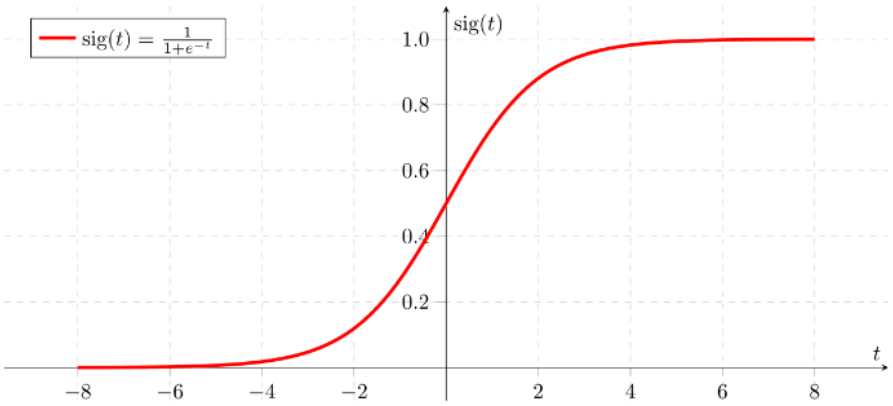
4. ML Modeling:

4.1 Logistic Regression:

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for a given set of features (or inputs), X.

The hypothesis function for logistics regression is:

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$



4.2 Naive Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

GaussianNB: It should be used for features in decimal form. GNB assumes features to follow a normal distribution.

MultinomialNB: It should be used for the features with discrete values like word count 1,2,3…

4.3 Passive Aggressive Classifier:

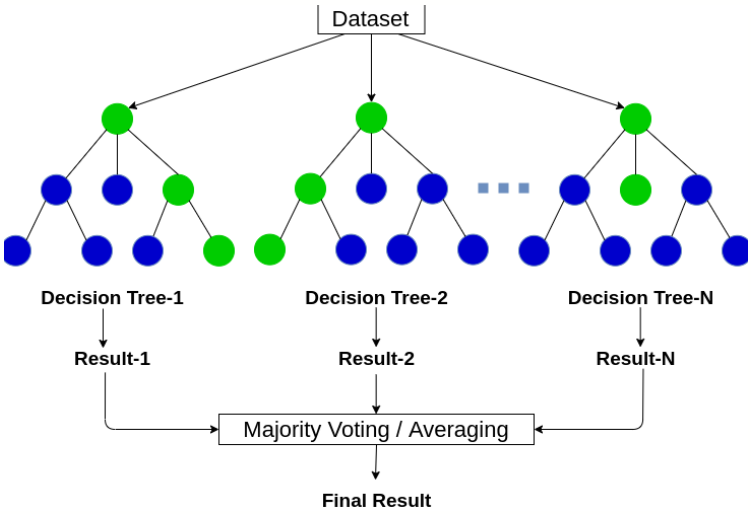
It is one of the few 'online-learning algorithms '. In online machine learning algorithms, the input data comes in sequential order and the machine learning model is updated step-by-step, as opposed to batch learning.

How its work?

- **Passive:** If the prediction is correct, keep the model and do not make any changes. i.e., the data in the example is not enough to cause any changes in the model.
- **Aggressive:** If the prediction is incorrect, make changes to the model. i.e., some change to the model may correct it.

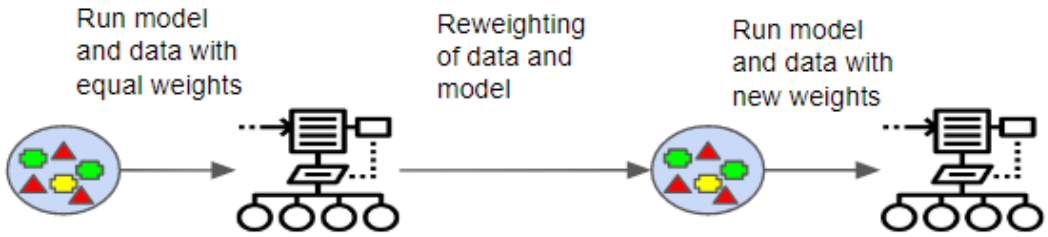
4.4 RandomForest Classifier:

The random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by the majority voting classifier. The output doesn't depend on one decision tree but on multiple decision trees.



4.5 GradientBoosting Classifier:

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting



5. Evaluation Metrics:

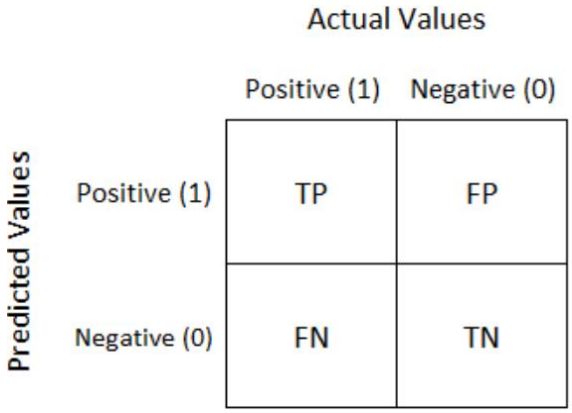
5.1 Accuracy:

Accuracy simply measures how often the classifier correctly predicts. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

5.2 Confusion Matrix:

A confusion matrix is defined as the table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.



5.2.1 Precision: Precision explains how many of the correctly predicted cases actually turned out to be positive.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

5.2.2 Recall: Recall explains how many of the actual positive cases we were able to predict correctly with our model.

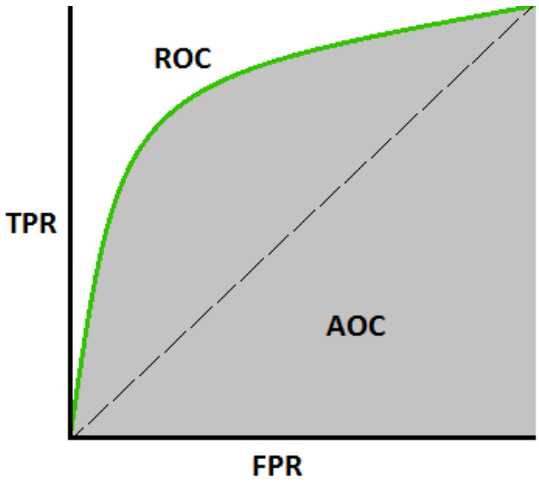
$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

5.2.3 F1 Score: It gives a combined idea about Precision and Recall metrics. It is maximum when Precision is equal to Recall. It is the harmonic mean of precision and recall.

$$F1 = 2.\frac{Precision \times Recall}{Precision + Recall}$$

5.3 AUC ROC:

The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR (True Positive Rate) against the FPR (False Positive Rate) at various threshold values and separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes.



6. Conclusion:

- The highest peak of the month feature is 7. According to legend, July is the month with the most travel. December is the second-most popular month.
- Most trips are taken by Spirit Airlines, which has the highest frequency in the dataset.
- The most trips taken were made by Airbus A320 aircraft, which had the highest frequency in the dataset, followed by Boeing 777 and Airbus A380 aircraft.
- The top spot pertains to the Bangkok to Hong Kong trip that occurred most frequently in the dataset, followed by Bangkok to London and London to New York trips.
- In the column for traveller type, it is noticeable to us that Solo Leisure travellers represent the majority of the population. In the cabin column, the majority of passengers prefer the Economy class.
- Following the use of bivariate analysis, we discovered that all travellers highly favors the economy class. Some Couple Leisure and Business class travellers choose to fly in business class. Among all traveller types, first class is the least popular.
- Due to the linear and balanced dataset, logistic regression outperformed the other algorithms well. The gradient boosting approach came in second.
- For this sort of dataset and its specified problem statement, accuracy and f1 score are the optimal evaluation matrix that is taken into consideration.