**Airline passenger referral prediction**

**Ashfaque sayyed**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract:**

Air business as we know has been largely affected due to Covid-19 and most of the airline now is sitting on the verge of Bankruptcy because of this situation. Due to industrialization and modernization is a huge competition between airlines to retain their customers by providing best service.

A customer review dataset consisting of around 17 features is given by almabetter to predict whether a customer will refer airline to his friend or not with the help of machine learning classification models.

1. **Problem Statement**

Data contains airline reviews from 2006 to 2019 from popular airlines around the world with multiple choices and free text questions. Data is scraped in spring 2019. The main objective is whether the passengers will refer the airline to their friends

**2. Introduction**

The Airline passenger Referral system has become the most important criteria globally for the airline industry in order to address the surge which has been created after global pandemic so as to remain in the global market competition. Airline referral system generally works on customer reviews which is basically sentiment given by the customer depending upon various factor like seat comfort, their trip distance, route they have travelled, timing, the airline frequency, ground service etc. on the basis of which sentiment reviews are analysed and machine learning model on classification is prepared which helps airline industries to focus on the factor resolving which it can actually help them in business growth better than the competitors.

**3.Data descriptions**

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.

Traveller type: Type of traveler (e.g. business, leisure)

cabin: Cabin at the flight.

date flown: Flight date

seat comfort: Rated between 1-5

cabin service: Rated between 1-5.

foodbev: Rated between 1-5

entertainment: Rated between 1-5

ground service: Rated between 1-5 value for money: Rated between 1-5.

recommended: Binary, target variable.

**4. Steps involved:**

**Data collection:**

Data collection is the process of collecting, measuring and analyzing different types of information using a set of standard validated techniques. The main objective of data collection is to gather information-rich and reliable data, and analyses them to make critical business decisions. Once the data is collected, it goes through a rigorous process of data cleaning and data processing to make this data truly useful for businesses. It refers to the process of finding and loading data into our system. Pandas library is used to loading our data in our system in python. Using pandas we can manipulate data easily.

**Data Cleaning:**

The next task was data cleaning which was easy with this dataset. Data cleaning refers to the process of removing unwanted variables and values from your dataset and getting rid of any irregularities in it. Such anomalies can disproportionately skew the data and hence adversely affect the results. Some steps that can be done to clean data are:

• Handling missing values: There are always some missing values in dataset. If we don’t remove or handle those missing values then that can cause a trouble in our analysis. Removing or replacing those missing values with something meaningful is very important so that our data will have no missing values.

• Removing duplicates: Drop the duplicates rows.

• Formatting data to proper dtype.

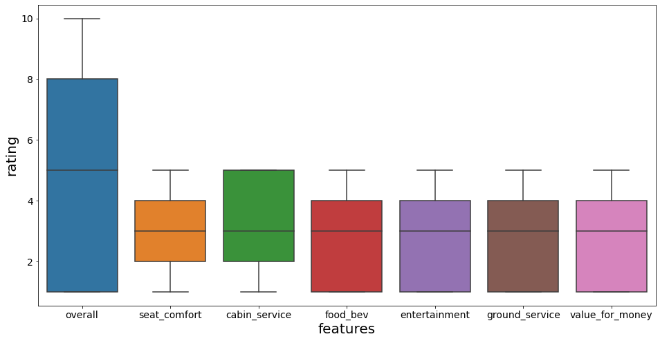
• Adding or removing columns required for analysis.

**Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) plays a vital role in the analysis of the data variables which are important from the aspect of feature engineering. It will help us to distribute and relate between dependent and independent variables. We have gone through an analysis of every independent as well as the dependent variable to check which independent factor affects the dependent factor.

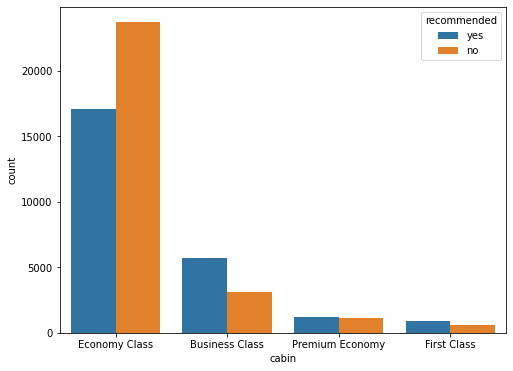
**4.1. Outlier detection**

Outlier detection is the process of detecting and subsequently excluding outliers from a given set of data. An outlier may be defined as a piece of data or observation that deviates drastically from the given norm or average of the data set. An outlier may be caused simply by chance, but it may also indicate measurement error or that the given data set has a heavy-tailed distribution.



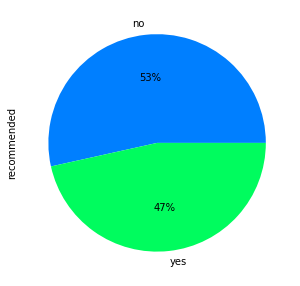
The outliers are not present in our data.

**4.2. Which type of cabin has maximum recommendations?**

****

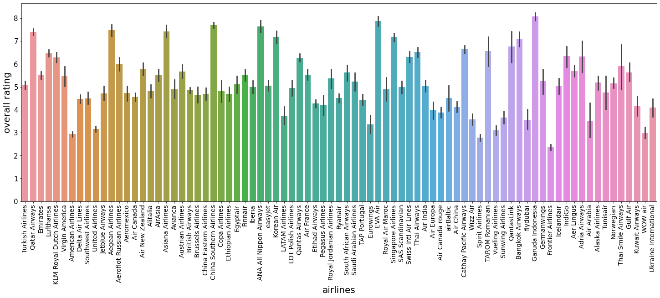
For the Economy class, Number of 'NO' recommendations are more than 'YES' recommendations. For business class and first class, the Number of 'YES' recommendations are more than 'NO' recommendations. For the Premium account number of 'YES' recommendations and 'NO' recommendations are approximately equal.

**4.3. What is the total recommendation percentage for all airlines?**

****

The overall recommendation percentage for all airlines is 47% which is less than recommended 'NO' by 6%. The data s not imbalanced.

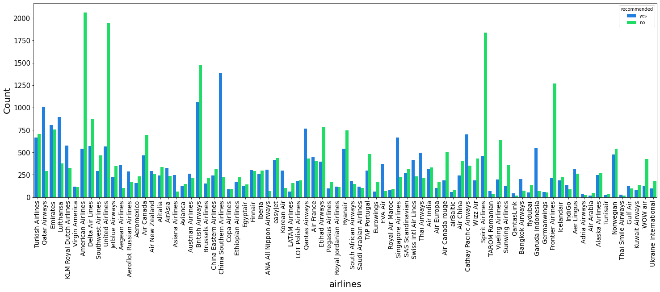
**4.4. What is the maximum overall rating got by different airlines?**

****

1.The maximum overall ratings are received by Qatar airlines, Aegean airlines, Asiana Airlines, China Southern Airlines etc ( rating is around 7.5-8).

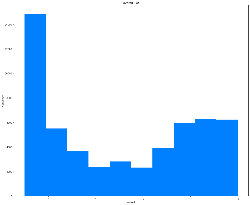
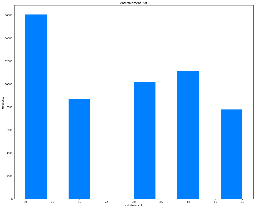
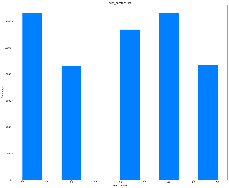
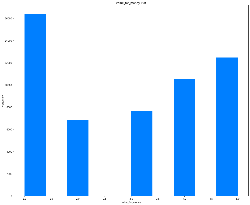
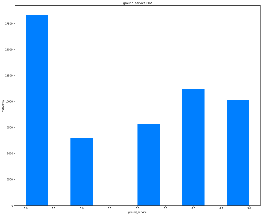
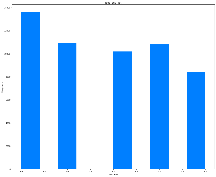
2.The minimum overall rating is around 2.5 received by frontier airlines.

**4.5. Which airlines got the maximum and minmum recommendations?**

****

American airlines, united airlines, spirit and frontier airlines received maximum 'NO recommendations. China southern airlines, Qatar, and British airways received maximum 'YES' recommendations. Thai smile, Tunisair, Air Arabia, and Adria airways received minimum 'Yes' recommendations.

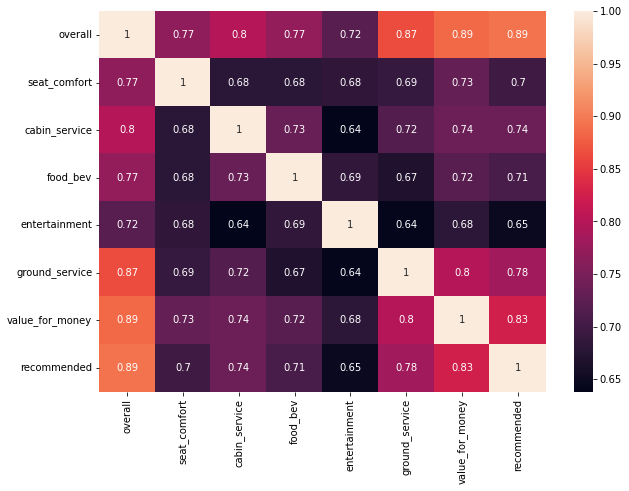
**4.6. Checking the frequency of values**

**   **

1. Cabin service got the maximum rating of 5.
2. Overall rating got by the airlines is poor equal to 1.
3. Maximum customers rate food\_bev as poor equal to 1.
4. Most of the customers have rated airlines as 1 indicating expensive(value for money).

**4.7. Heatmap**

Let’s check the heatmap plotted concerning independent variables.

****

'Overall','food bev','cabin\_service', 'value\_for\_money' etc are positively correlated with recommendation. 'Overall' is most correlated with the recommendation.

**Model Training:**

Model training is the process of fitting a data into machine learning model from which model learns the patterns in data to predict the dependent variable. Model do it so by assigning a weight to each variable. After our model is trained, we test our model on test data to check how our model is performing. The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. In this project we have used 80% data for training purpose and 20% data for test set. The train-test procedure is appropriate when there is a sufficiently large dataset available.

**Fitting different models**

For modeling we tried various regression algorithms like:

1. **Logistic Regression**
2. **Decision tree**
3. **Random forest**
4. **K-Nearest neighbour**
5. **Random forest with GridSearch CV**

**Tuning the hyperparameters for better accuracy**

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and avoiding overfitting. also called hyperparameter optimization, is the process of finding the configuration of hyperparameters that results in the best performance. The process is typically computationally expensive and manual.

We used Grid Search CV for hyperparameter tuning.

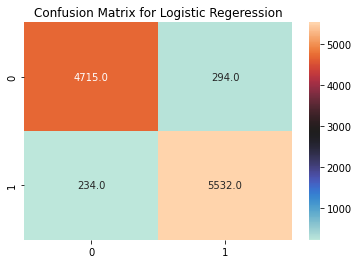
**Grid Search CV:**

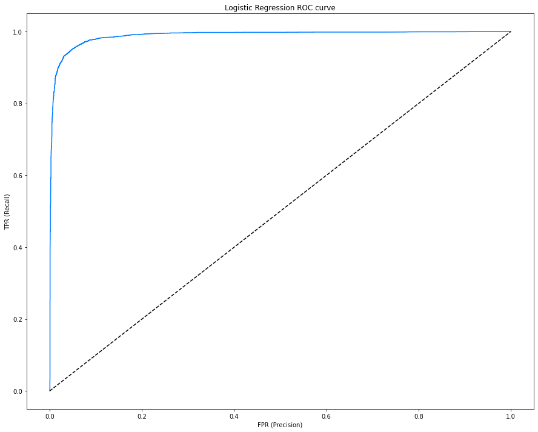
Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

**5. Algorithms:**

1. **Logistic Regression:**

Logistic regression is a classification technique that predicts the likelihood of a single-valued result (i.e. a dichotomy). A logistic regression yields a logistic curve with values only ranging from 0 to 1. The likelihood that each input belongs to a specific category is modelled using logistic regression. Logistic regression is a fantastic tool to have in your toolbox for classification purposes. For classification situations, where the output value we want to predict only takes on a small number of discrete values, logistic regression is an important technique to know. The logistic function offers a number of appealing characteristics. The probability is represented by the y-value, which is always confined between 0 and 1, which is exactly what we wanted for probabilities. A 0.5 probability is obtained for an x value of 0. A higher likelihood is also associated with a higher positive x value, while a lower probability is associated with a greater negative x value. In logistic regression to learn the coefficients of features in order to maximize the probability of correctly classifying the classes. For this maximum likelihood, concept is used.

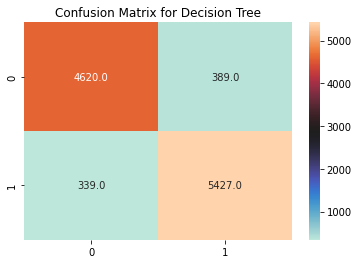


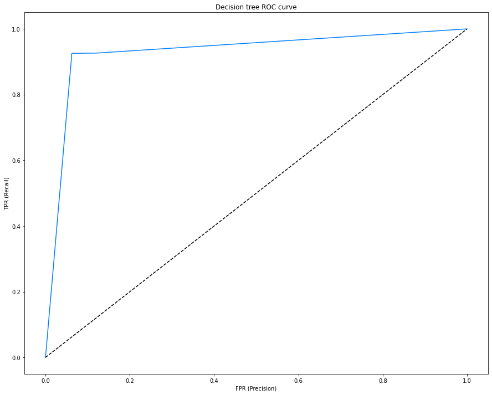


.

1. **Decision tree:**

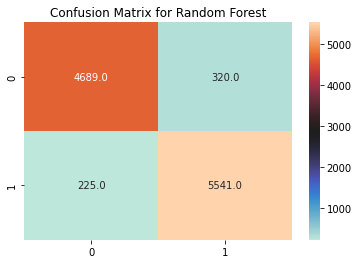
A decision tree is a supervised learning technique used to solve categorization problems. Both categorical and continuous input and output variables are supported. The decision to make strategic splits has a significant impact on a tree's accuracy. The decision criteria for classification and regression trees are different. To decide whether to break a node into two or more sub-nodes, decision trees employ a variety of techniques. The homogeneity of the generated sub-nodes improves with the generation of sub-nodes. To put it another way, the purity of the node improves as the target variable grows. The decision tree separates the nodes into sub-nodes based on all available variables and then chooses the split that produces the most homogenous sub-nodes.

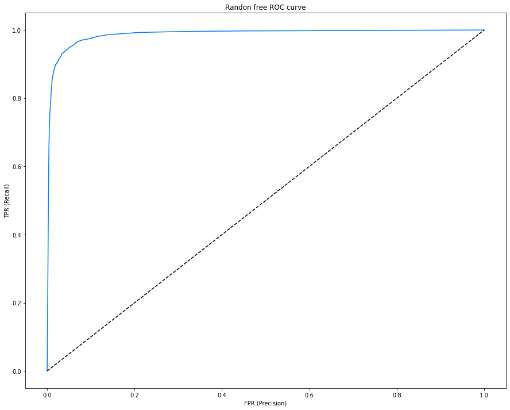




1. **Random Forest:**

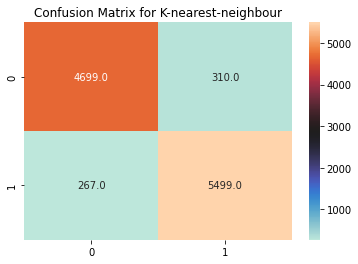
We create several trees in the Random Forest model rather than a single tree in the CART model. From the subsets of the original dataset, we create trees. These subsets can contain a small number of columns and rows. Each tree assigns a categorization to a new object based on attributes, and we say that the tree "votes" for that class. The classification with the highest votes is chosen by the forest.

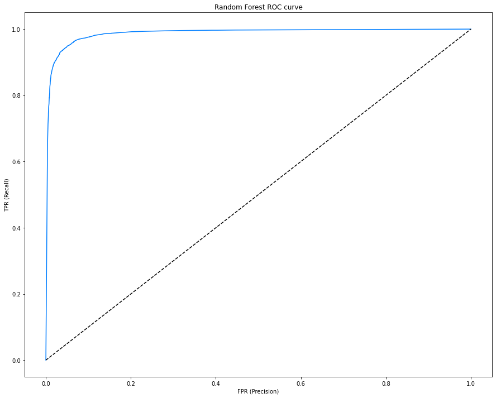




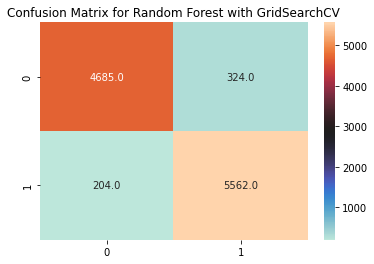
1. **K\_nearest Neighbour Model**

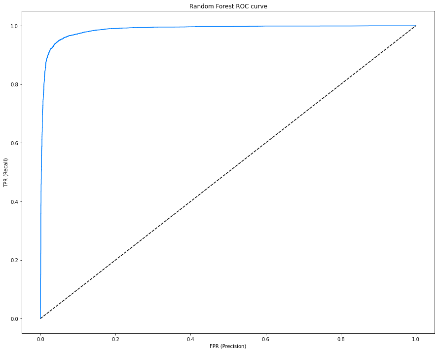
K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbours are classified.



****

1. **Random forest Model with GridSearchCV**





The accuracy for K\_nearest Neighbour model is 94.94% and by applying GridSearchCV on we got the accuracy with 95.1%

**6. Conclusion:**

1.  'Overall','food bev','cabin\_service', 'value\_for\_money' etc are positively correlated with recommendation. these parameters should be improved to provide better service and hence it will improve recommendation chances for airlines.

2.entertainment has 0.65 of correlation which is less than others.

3. 'Overall' is most correlated with the recommendation. 4. Logistic regression, decision tree, random forest,and  KNearest neighbor gave good results in terms of accuracy. The highest accuracy obtained is 95.1% with logistic regression.

5.Random forest with Gridsearch CV also gave good accuracy approximately equal to logistic regression (95.08%).

6. American airlines, united airlines,spirit and frontier airlines received maximum 'NO' recommendations.

7. Chins southern airlines, Qatar and British airways received maximum 'YES' recommendations. Thai smile, Tunisair, Air arabia, adria airways received minimum 'Yes' recommendations.

8.For Economy class, Number of 'NO' recommendations are more than 'YES' recommendations.

9.For business class and first class, Number of 'YES' recommendations are more than 'NO' recommendations.

10.For the Premium account number of 'YES' recommendations and 'NO' recommendations are approximately equal.

**References-**

1. stackoverflow.com
2. towardsdatascience.com
3. GeeksforGeeks
4. Analytics Vidhya