Airline Passenger Referral Prediction

Saugata Deb Shreyash Movale, Ankit Patil, Naga sai

Data science trainees,

Almabetter

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. Any bad decision may lead to severe outcomes where no stakeholder wants to invest without any future assurance. As an example of Singapore airlines who are among the hardest hit. As we know this situation is not permanent and it will be over but once this is over there will be a high surge as people will be back for holidays overseas. What can airlines do to tackle this situation? To answer this question, a machine learning model for classification is created from the airline_reviews dataset. This dataset has been provided to us by Almabetter in order to identify the important factors that lead to better customer satisfaction.

1. Problem Statement

Data is scraped in Spring 2019 from Skytrax website. Data includes airline reviews from 2006 to 2019 for popular airlines around the world with multiple choice and free text questions. The main objective is to predict whether passengers will refer the airline to their friends or not.

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
- **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

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.

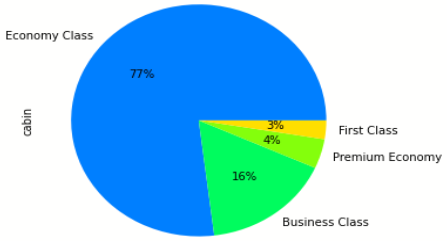
3.ExploratoryData Analysis

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.

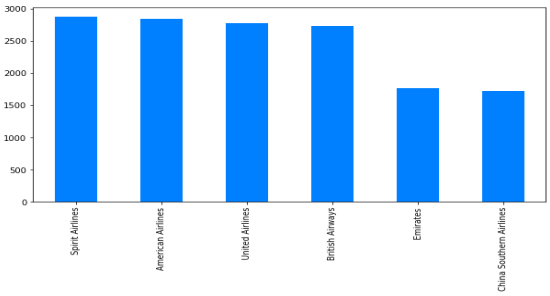
3.1 percentage of class of passenger in cabin

The pie plot on the down tells about the cabin in which most of the passengers travelled. It can be clearly derived from the plot that almost 77% of total passengers were "Economy Class" travelers and only 3% of total passengers were "First Class" travelers.



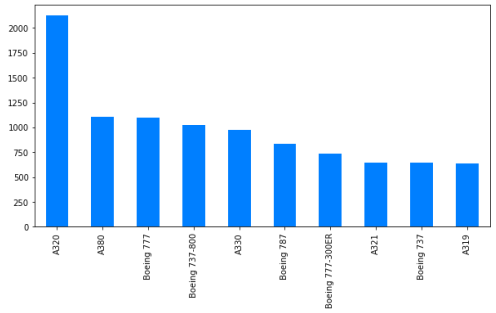
3.2 top 6 most frequently used airlines

These are the top 5 most frequently used airlines in the given data.



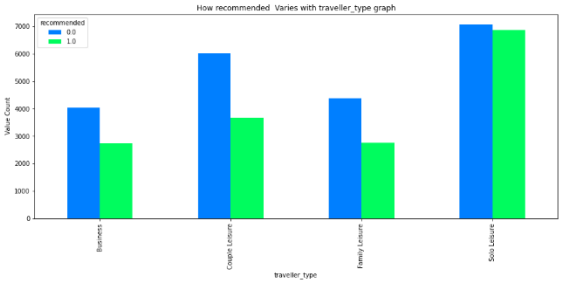
3.3 Top 10 aircrafts used

The below plot shows the top 10 aircrafts used in the given data. The A320 is the most used aircraft in the given data.



3.4 Types of customer categories and their opinion

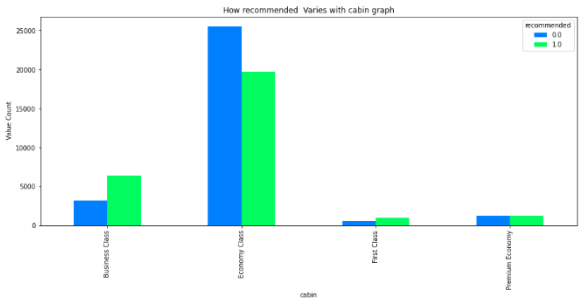
We can see that people have given both 1 or 0 which we will consider from now on as positive and negative recommendations so as to interpret it effectively to the solo leisure. This may be because of the poor infrastructure or the service received by the people and positive recommendation may be because of low price for solo. But this is an approximate analysis based on the data provided. In Traveller type we can see that both the recommendation trend as of yes or no increases from business to couple leisure and decreases to family then again increases high in solo leisure. Which indicates people prefer solo leisure higher than any of the other leisure's.

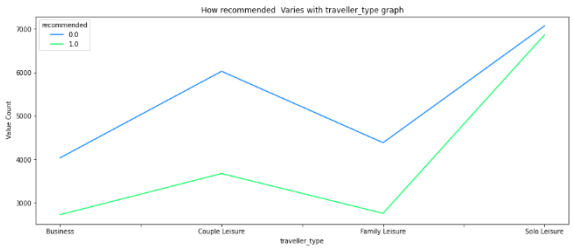


3.5 Cabin wise analysis

Also, we can see that people give highly positive recommendations to the economy class in the cabin. From this we can conclude that people love to travel in economy class as of low price. In the same way we can see people give the highest negative recommendation to economy class maybe because less infrastructure or service provided to them. Also we can see people have given

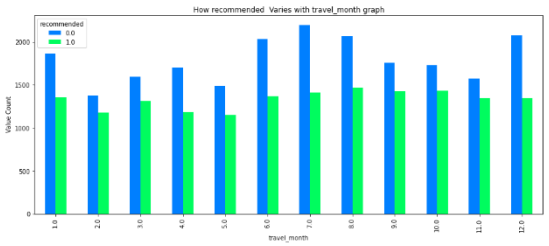
the highest positive recommendation to Business class. It may be because of the quality of service provided to them in Business class and similarly negative recommendation because of the high price of business class or less travelling percentage. In Cabin type we can see that both the recommendation trend of yes or no increases from business to Economy class and decreases to First class then again increases slightly in Premium class. Which indicates most people travel in economy class.

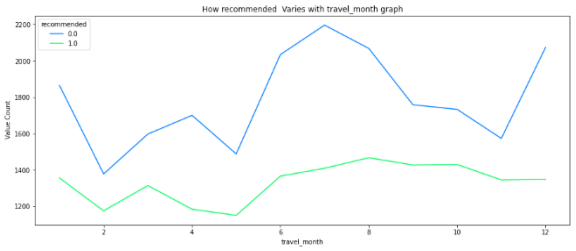




3.6 travel month analysis

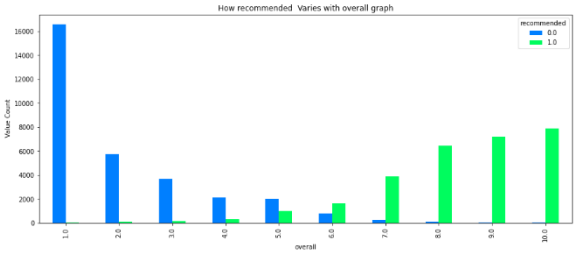
From month vs no. of recommendation. We can see that people tend to travel most in the month of July considering the total of positive and negative recommendations combined. In month we cannot see any preferable trend but here we can conclude people tend to travel highest during the month of July.

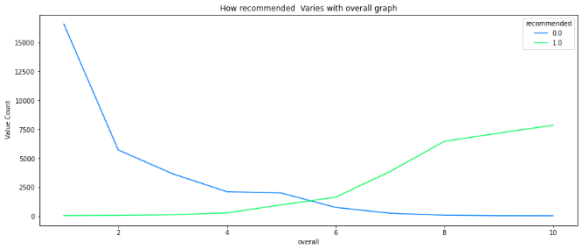




3.7 overall rating

From overall vs recommended graph we can see which is perfectly understandable that negative recommendation has been given to the overall rating of 1.0 and high positive recommendation has been given to the overall rating of 10. But it is very true that the highest negative recommendation has been given to an overall rating of 1.0 which is really a matter of concern. In overall rating we can experience very good insights which are also regular. We can see as the positive recommendation increases with the overall rating and also negative recommendation on the same decreases.

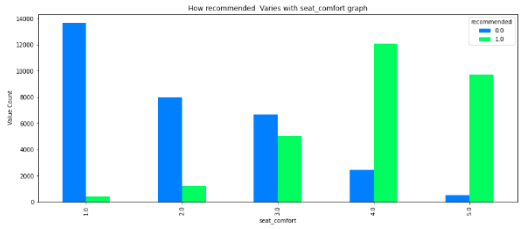


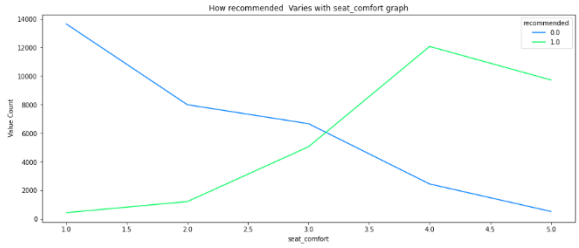


3.8 Seat comfort Analysis

In seat comfort people have given the highest positive recommendation to the seat of class 5 as compared to very low negative recommendation to the same. Also we can see seat of class 1 have been given the highest negative recommendation as compared to its

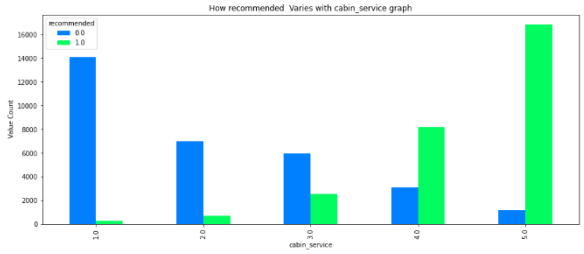
positive recommendation. Here we come to a conclusion it must be removed as early as possible. In seat comfort we can see as the positive recommendation increases with the overall rating and also negative recommendation on the same decreases also we can see an intersection in seat comfort rating 3.0 where we can see similar positive and negative recommendation.

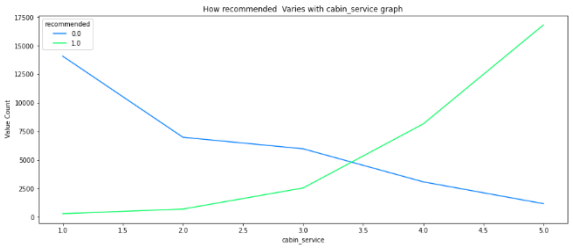




3.9 Cabin service Analysis

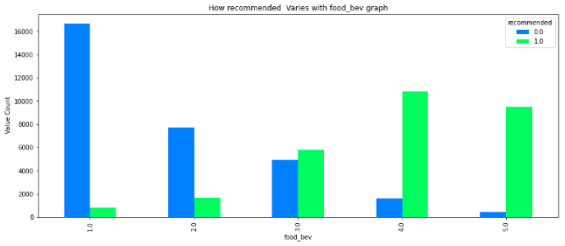
In cabin service rating people have given the highest recommendation to cabin service rating 5 as compared to its counterpart. From this we can conclude that cabin service is doing pretty good.In cabin service we can see the positive recommendation increases with the overall rating and also negative recommendation on the same decreases also we can see an intersection in cabin service rating 3.5 where we can see similar positive and negative recommendation.

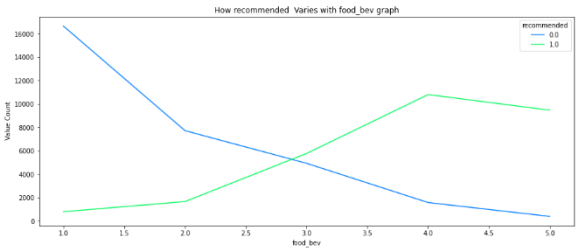




3.10 Food beverages

In food and beverage ratings people have given the highest negative recommendation to rating 1.0 from this we can conclude that airline service has to improve their food delivery and quality service. In food service we can see the positive recommendation increases with the overall rating and also negative recommendation on the same decreases also we can see an intersection in food service rating close to 3.0 where we can see similar positive and negative recommendation.

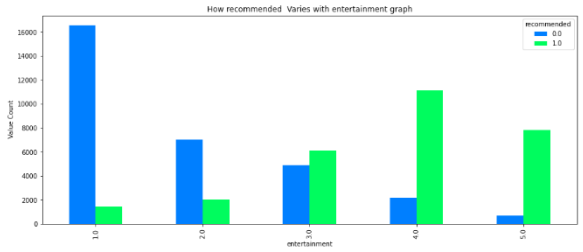


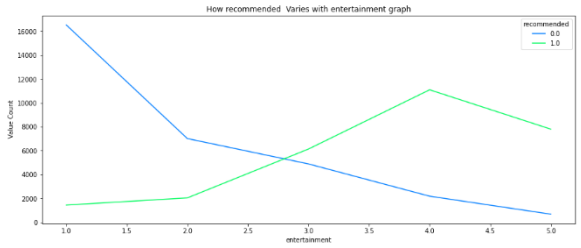


3.11 Entertainment Analysis

In entertainment also we can see most people have given the highest negative recommendation to entertainment rating 1 which shows that airlines has to improve their entertainment system as well. In Entertainment service too we can see the same as the positive recommendation increases with the overall rating and also negative

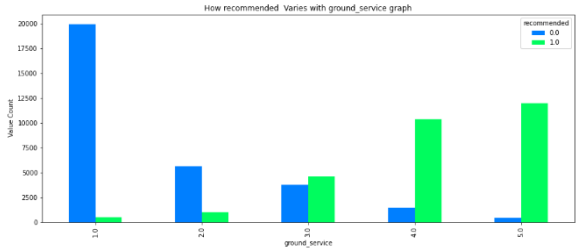
recommendation on the same decreases also we can an intersection in Entertainment service rating between 2.5 and 3.0 where we can see similar positive and negative recommendation.

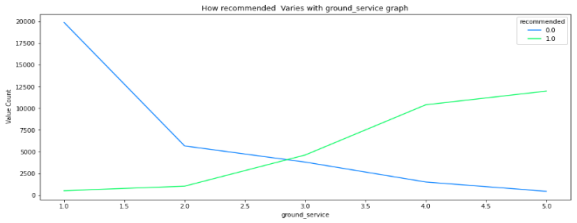




3.12 Ground service Analysis

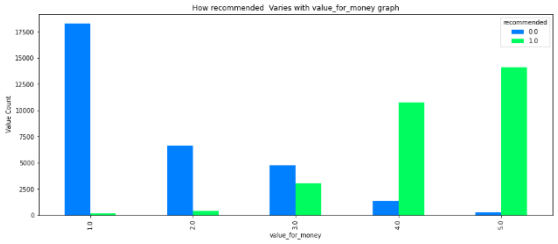
In Ground service also we can see the positive recommendation increases with the overall rating and also negative recommendation on the same decreases also, we can see an intersection in Ground service rating close 3.0 where we can see similar positive and negative recommendation. In ground service also we can see most people have given the highest negative recommendation to ground service rating 1 which shows that airlines must improve their ground service.

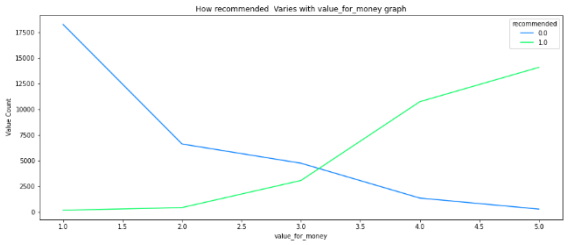




3.13 Value for money

In value for money also we can see most people have given the highest negative recommendation to value for money rating 1 which shows that airlines have to make their flight service more cost effective. Lastly in Value for money rating we can see the positive recommendation increases with the overall rating and also negative recommendation on the same decreases also we can see an intersection in Value for money rating greater than 3.0 where we can see similar positive and negative recommendation.





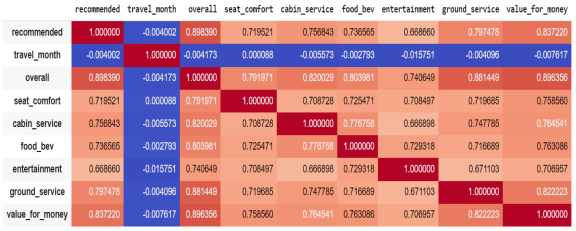
4. Correlation Analysis

The correlation analysis has been done to get a better understanding of dependent and independent variables' multicollinearity. Multicollinearity may not affect the accuracy of the model as much but we might lose reliability in determining the effects of

individual independent features on the dependent feature in your model and that can be a problem when we want to interpret your model.

4.1 Heatmap

Let's check the heatmap plotted concerning independent variables.



	recommended	travel_month	overall	seat_comfort	cabin_service	food_bev	entertainment	ground_service	value_for_money
recommended	1.000000	-0.004002	0.898390	0.719521	0.756843	0.736565	0.668660	0.797478	0.837220
travel_month	-0.004002	1.000000	-0.004173	0.000088	-0.005573	-0.002793	-0.015751	-0.004096	-0.007617
overall	0.898390	-0.004173	1.000000	0.791971	0.820029	0.803981	0.740649	0.881449	0.896356
seat_comfort	0.719521	0.000088	0.791971	1.000000	0.708728	0.725471	0.708497	0.719685	0.758560
cabin_service	0.756843	-0.005573	0.820029	0.708728	1.000000	0.776758	0.666898	0.747785	0.764541
food_bev	0.736565	-0.002793	0.803981	0.725471	0.776758	1.000000	0.729318	0.716689	0.763086
entertainment	0.668660	-0.015751	0.740649	0.708497	0.666898	0.729318	1.000000	0.671103	0.706957
ground_service	0.797478	-0.004096	0.881449	0.719685	0.747785	0.716689	0.671103	1.000000	0.822223
value_for_money	0.837220	-0.007617	0.896356	0.758560	0.764541	0.763086	0.706957	0.822223	1.000000

Overall and Recommended are highly correlated, Overall and Value for money are highly correlated

5 Feature description

- **airline**: Name of the airline in str format
- **overall**: Overall point is given to the trip between 1 to 10 in float format.
- **author**: Author of the trip in str format
- **review date**: Date of the Review customer review: Review of the customers in free text format in str need to be converted into DateTime Format
- **aircraft**: Type of the aircraft in str format
- **traveller type**: Type of traveler (e.g. business, leisure) consist of four class in str format
- **cabin**: Cabin at the flight date flown: Flight date in str format consist of 4 class.
- **seat comfort**: Rated between 1-5 in float format
- **cabin service**: Rated between 1-5 float format
- **foodbev**: Rated between 1-5 entertainment: Rated between 1-5 in float format
- **groundservice**: Rated between 1-5 in float format
- **value for money**: Rated between 1-5 in float format

5. Feature Engineering

The given information in its crude structure was not straightforwardly utilized as a contribution to the model. A few components designing were completed where barely any elements were changed, few were dropped, and few were added. The following is a rundown of the element designing completed with the gave informational index

We have Engineered new features based on the existing features which are date of travel, review text, overall rating etc.

We have done imputation of missing values in the target variable, we also did imputation of missing values in the independent variable. We handled categorical variables and date columns. We used NLP for handling the review text feature.

We also did one hot encoding on the categorical features like airline, cabin, traveller_type.

6. WORKING WITH DIFFERENT MODELS

6.1 Train/Test Split

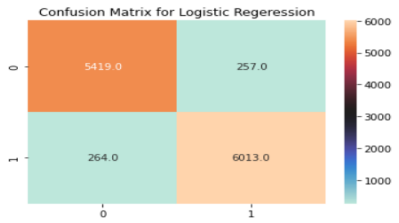
The train/test split was done as 80/20 % of data with a random state of 0. The final dataset was of shape (61183, 17) which was split to (48946, 17) as Train data and (12237,17) as Test data.

6.2 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.

	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.95	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.64%



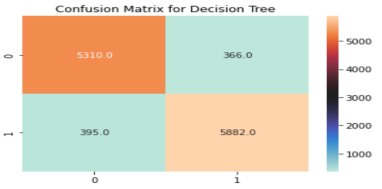
6.3 Decision Tree

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, then chooses the split that produces the most homogenous sub-nodes.

	precision	recall	f1-score	support
0.0	0.94	0.94	0.94	6277
1.0	0.93	0.94	0.93	5676
accuracy			0.94	11953
macro avg	0.94	0.94	0.94	11953
weighted avg	0.94	0.94	0.94	11953

Accuracy score % of the model is 93.63%

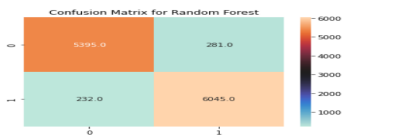


6.4 Random Forest

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.

	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.96	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.71%

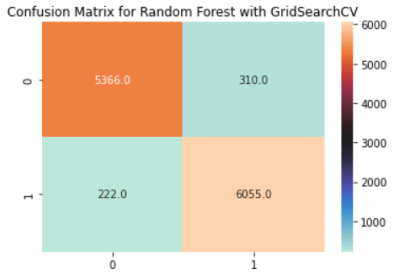


6.5 Random Forest with GridSearchCV

The best parameters for this grid search max_depth 12, min_sample_leaf 5(min_sample_split 100 and n_estimators as 80. S that the model accuracy we obtained is 95.5%.

	precision	recall	f1-score	support
0.0	0.95	0.96	0.96	6277
1.0	0.96	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.55%

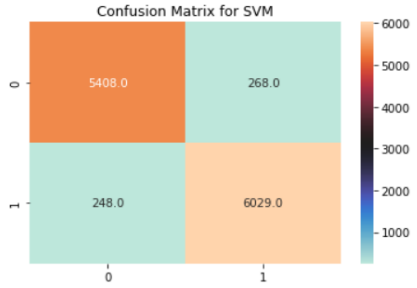


6.6 Support Vector Machine

SVM(Support Vector Machine) SVMs take a direct approach to binary classification by attempting to find a hyperplane in a feature space that "best" separates the two classes. In practise, however, finding a hyperplane that completely separates the classes using only the original features is challenging (if not impossible). SVMs get around this by expanding the idea of separating hyperplanes in two different ways. (1)Expand the feature space to the point where perfect separation of classes is (more) likely, and(2) apply the so-called kernel trick to extend the feature space.
Support Vector - the dividing line between two sets of points that maximises the margin between them A number of the training sites are nearly on the edge of the margin, as represented by the black circles in this diagram. The support vectors are the pivotal elements of this fit, and they are known as the key aspects of this fit.

	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.96	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.68%

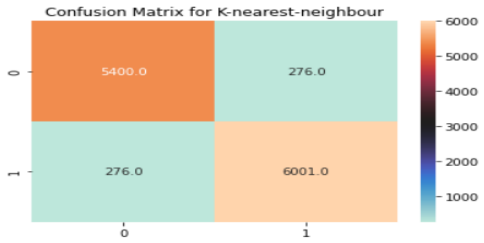


6.7 K_nearest Neighbour Model

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.

	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.95	0.95	0.95	5676
accuracy			0.95	11953
macro avg	0.95	0.95	0.95	11953
weighted avg	0.95	0.95	0.95	11953

Accuracy score % of the model is 95.38%

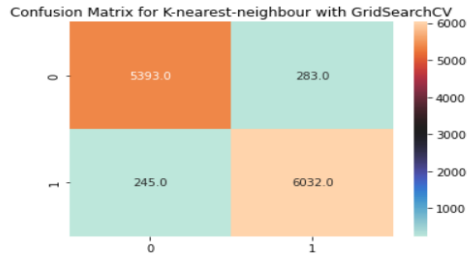


6.8 K_nearest Neighbour Model with GridSearchCV

The best n_neighbors value we get is 42. The accuracy for K_nearest Neighbour model is 95.38% and by applying GridSearchCV on we got the accuracy with 95.58%.

	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.96	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.58%

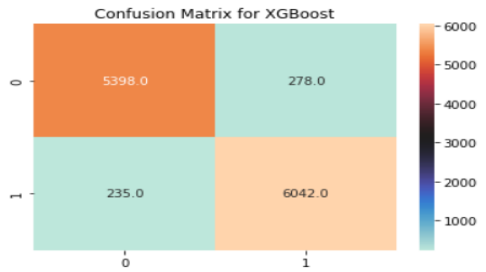


6.9 XGBoost Model

XGBoost is a distributed gradient boosting library that has been optimised for performance, flexibility, and portability. It uses the Gradient Boosting paradigm to implement machine learning algorithms. XGBoost is a parallel tree boosting (also known as GBDT, GBM) algorithm that solves a variety of data science problems quickly and accurately.

	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.96	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.71%



7. Conclusion

1. So here we come at the end of our project Airline Passenger Referral Prediction. Let's take a short recap on what we have done. In exploratory analysis we first find the duplicate we found 70711 we then drop those duplicates after dropping those duplicates we did find the info we found 61684 entries. Then we dropped those which have all null values. Then we found the percentage of passengers in different cabins using a pie plot. We found the economic class highest. Also we found Spirit Airways is the most frequently used airways. We found flight A320 to be the most frequent aircraft. Also people prefer to travel solo. July is the month where people travel most.
2. In feature description we did natural language processing to convert the customer_reviews sentiment based on polarity to numeric reviews. We did one hot encoding on categorical features.
3. In model selection we first did a train and test split in 4:1 or 80:20 split. We created functions to store evaluation metrics values.Then we did model deployment. XGBoost model had shown highest model accuracy along with highest recall,precision,fi_score and roc_auc_score. We select XGBoost for classification of our prediction.
4. In model explainabilibility we used Shap JS summary we can see positive features overall, value for money,numeric_review combined red color block pushes the prediction toward right over base value and causing positive model prediction and it is common for all models.

8. References:

- GeekforGeeks
- Kaggle
- Analytics Vidya