

Capstone Project

Airline Passenger Referral Prediction



Content:

- Problem statement
- Inferences from visualization of features
- Feature engineering
- Building Classifier Models
- Comparing Different Models Performance
- Conclusion
- Improvement
- Challenges



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.

Data Summary:









Data set name-- data_airline_reviews.xlsx

Shape-- (131895,17)

Columns--

['airline', 'author', 'review_date', 'customer_review', 'aircraft', 'traveller_type', 'cabin', 'route', 'date_flown', 'overall_score', 'seat_comfort', 'cabin_service', 'food_bev', 'entertainment', 'ground_service', 'value_for_money', 'recommended']

Understanding Data:

Aircraft	Boeing 787
Type Of Traveller	Business
Seat Type	Business Class
Route	Newark to Doha via Toronto
Date Flown	April 2021
Seat Comfort	
Cabin Staff Service	
Food & Beverages	
Inflight Entertainment	
Ground Service	
Wifi & Connectivity	
Value For Money	
Recommended	

Person feedback left by confirmed customers of most of the world's major airlines make up the dataset.

Data Cleaning:

- NAN values in alternate rows
- Dropped 70711 duplicate rows(53.61% of total rows)
- Left with rows 61183

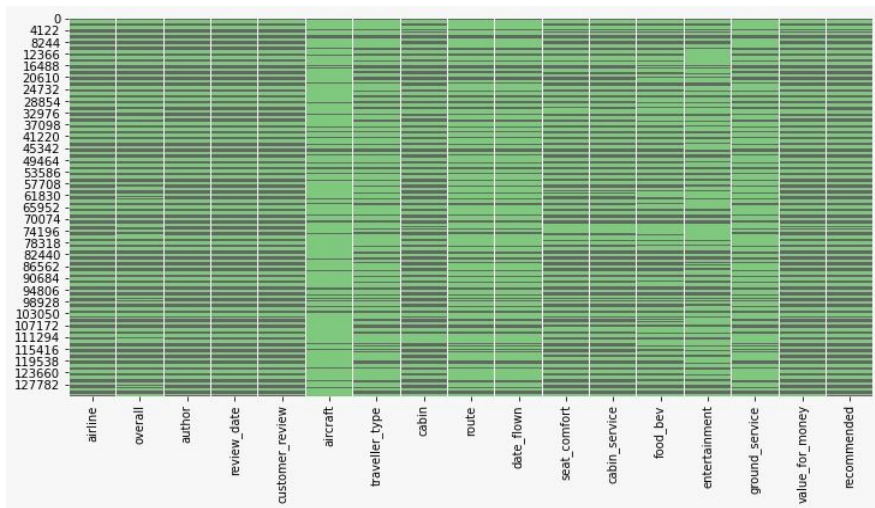


Fig: Before removing the alternative rows with all NaN values.

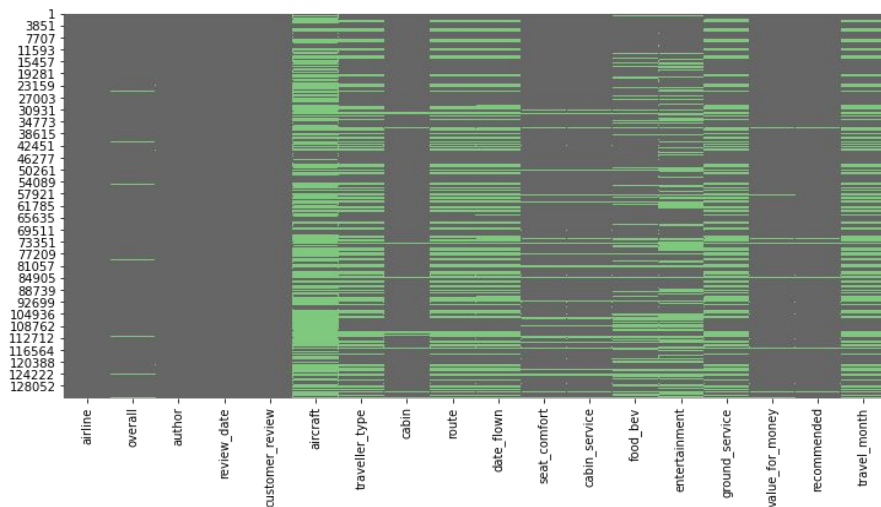
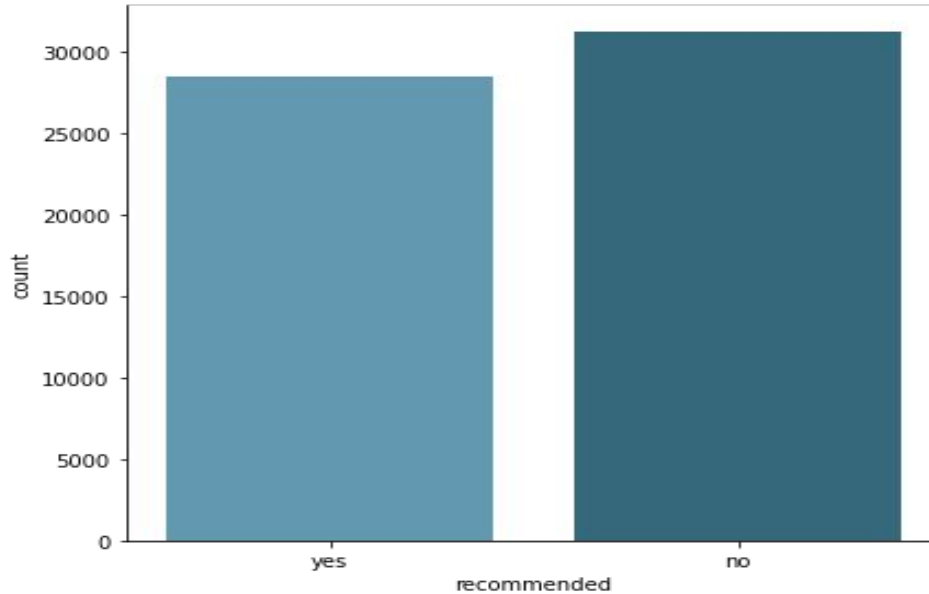


Fig: After removing the alternative rows with all NaN values.

Checking Imbalance in Target Column:

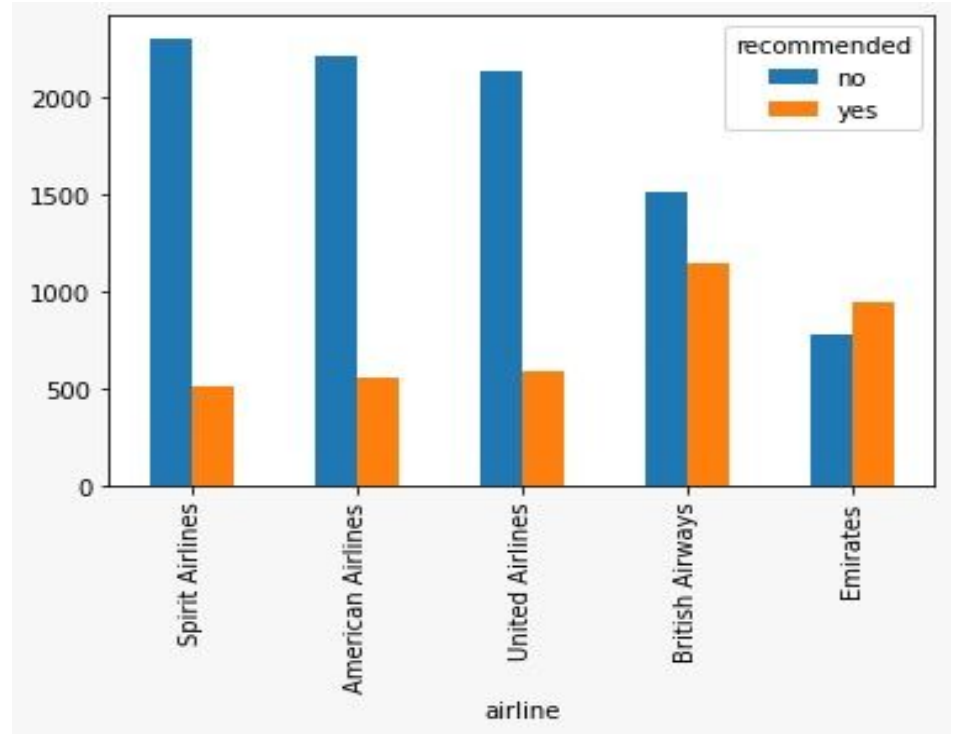
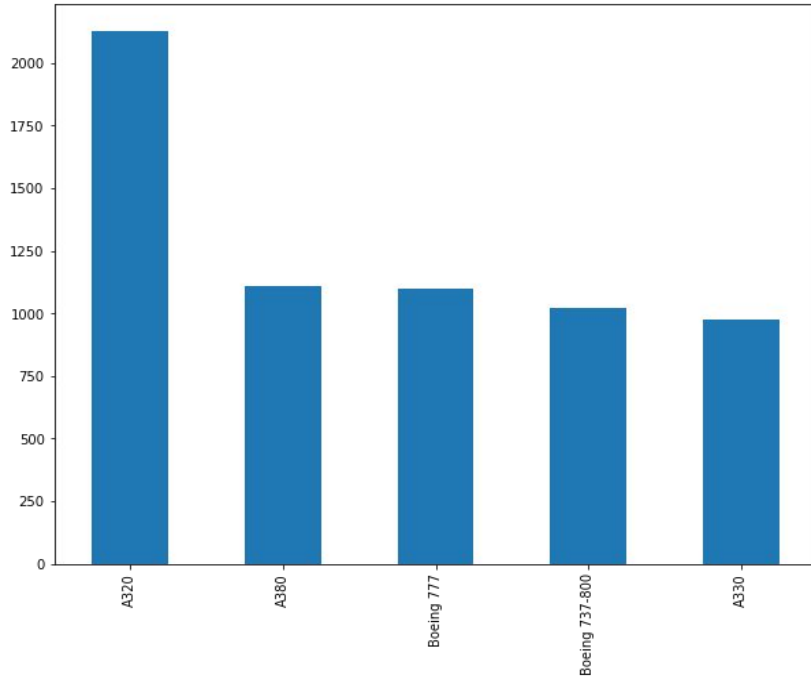


Dataset given is not under the influence of imbalances

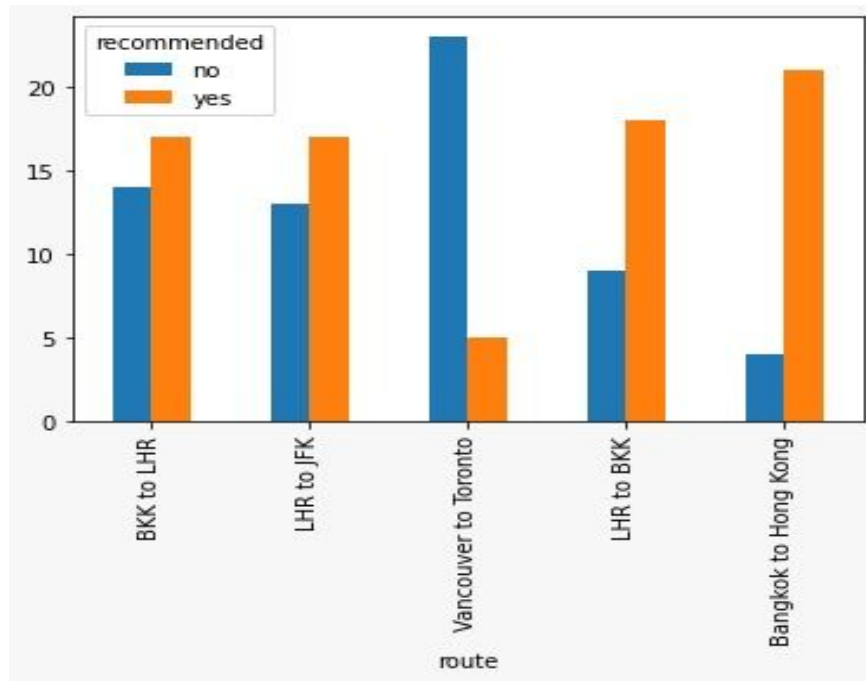
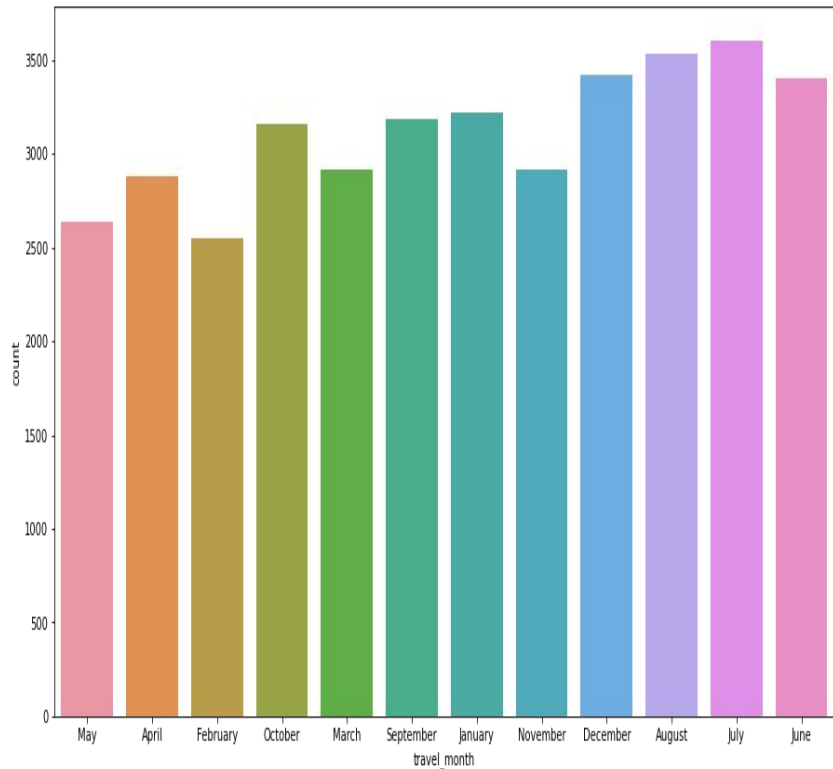
Inferences from Visualization of Features:

- Top 5 aircraft types and airlines
- Which season of a year people prefer to travel and what are the top 5 routes?
- Do people travel in group and which cabin class, they prefer?
- How Overall rating relates to passenger Recommendation?
- Top words in passenger Reviews
- Correlation among different rating types

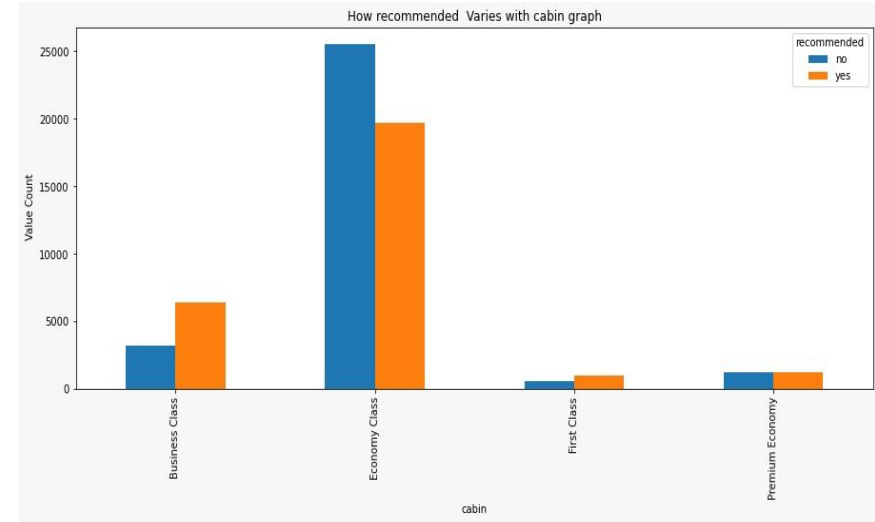
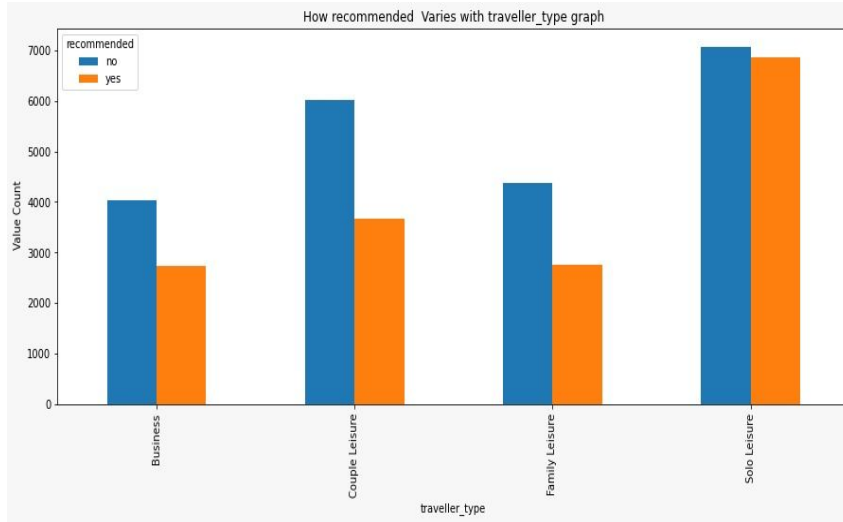
Top 5 aircraft types and airlines:



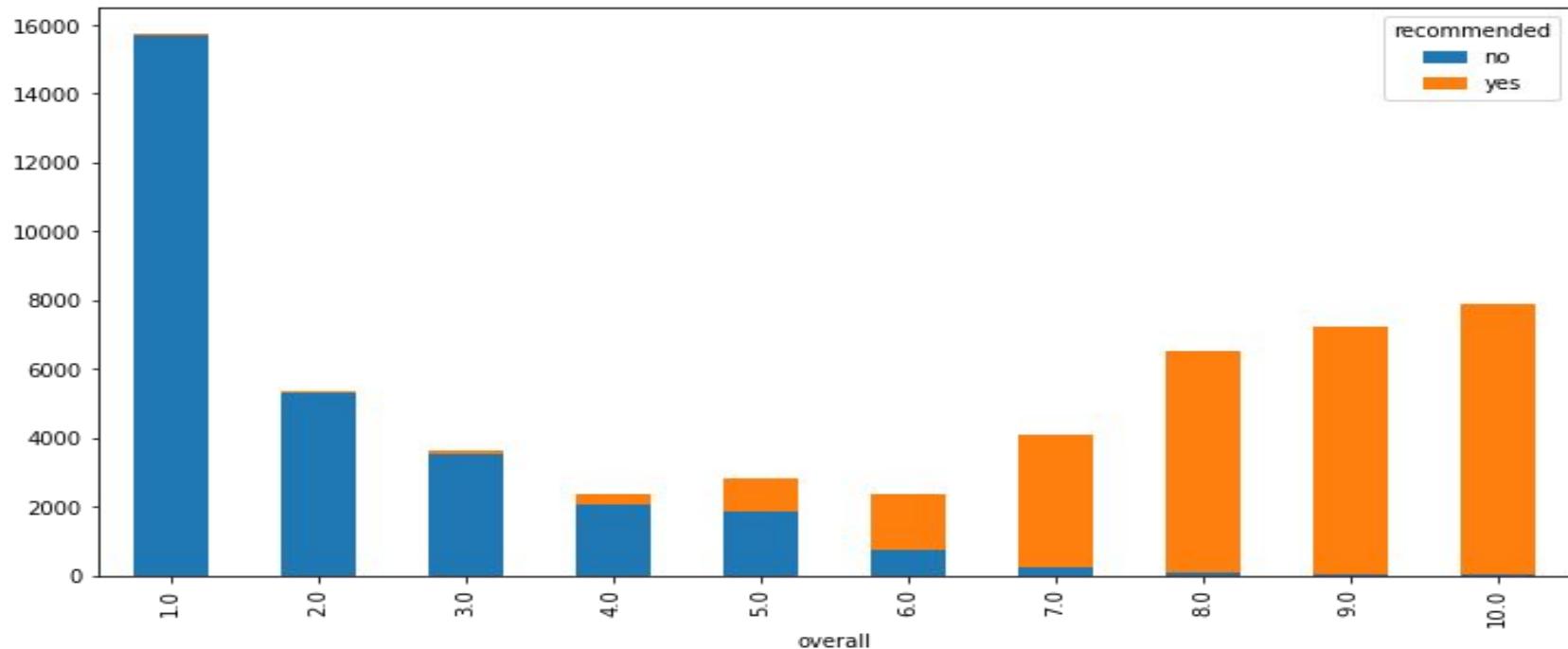
Which season of a year people prefer to travel and what are the top 5 routes?



Do people travel in group and which cabin class, they prefer?

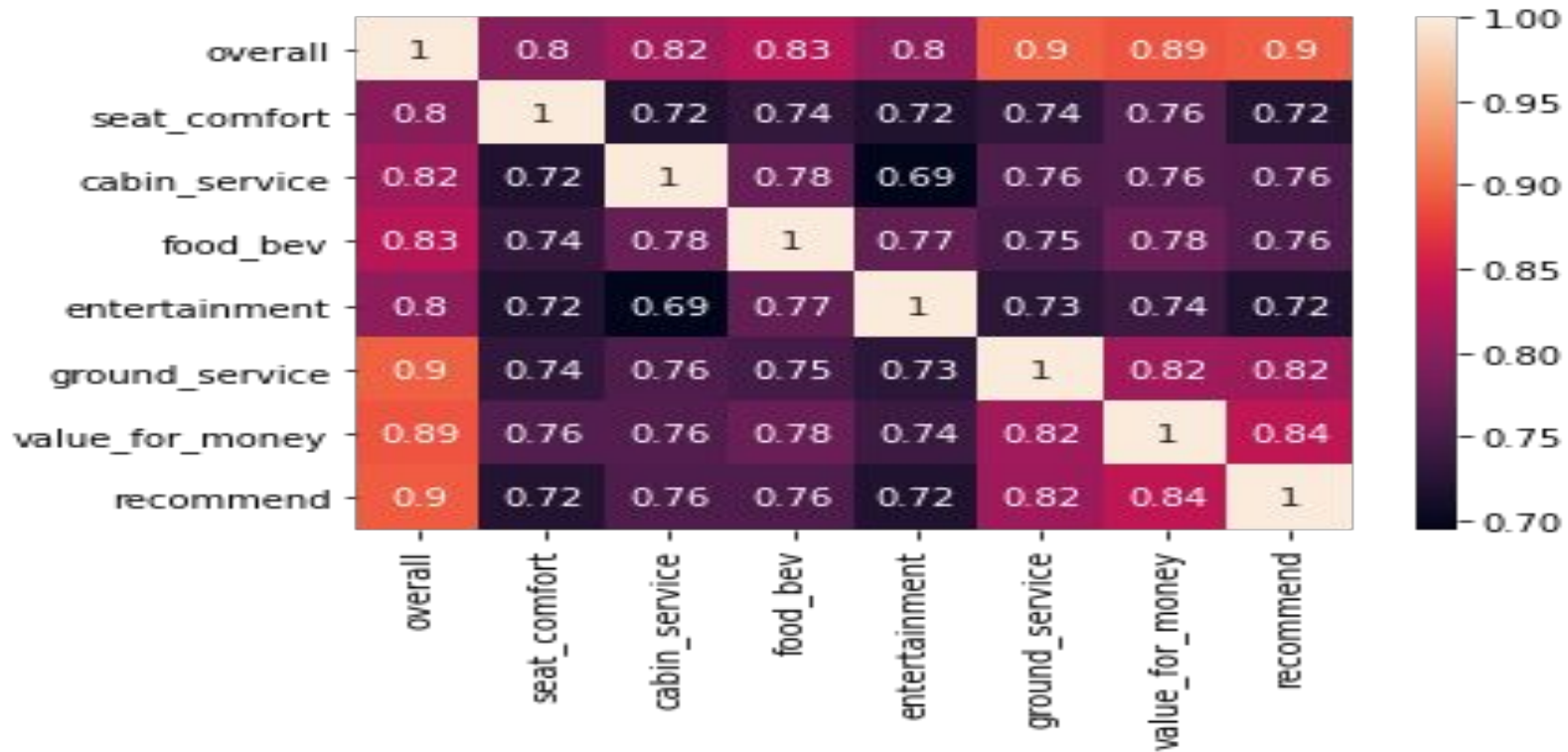


How Overall rating relates to passenger Recommendation?





Correlation among different rating type features and target feature:



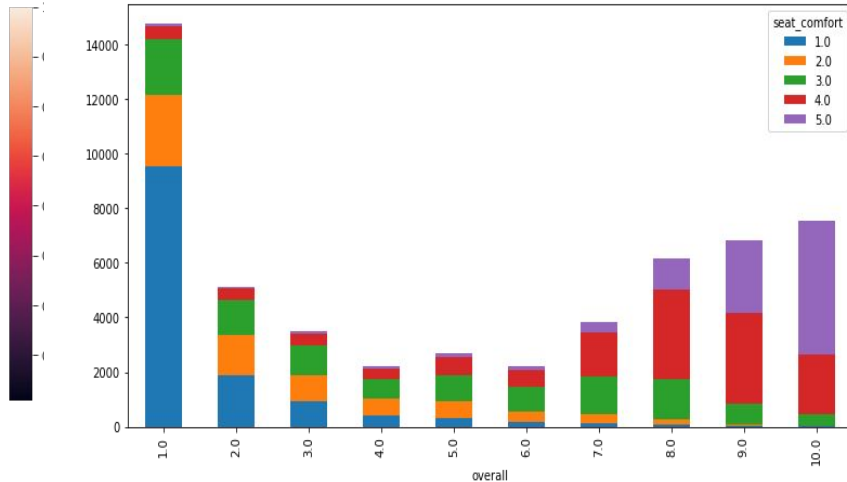
Imputation of missing values and Feature Engineering:

- Imputation of missing values in target variable.
- Imputation of missing values in Independent variable.
- Handling categorical variables and Date.
- Handled anomaly in Target variable.
- Handling text column:
 - Bag of word(BoW)
 - Tf-IDF
 - Sentiment VADER

Imputation of missing values in Target variable

- 1451 missing values in the target variable.
- Imputation using simple model with help of if and else statement on overall feature.
- Imputation using Review column by building a Naive Bayes model with Review column and Target columns.
- Ultimate model is chosen for the one with Overall feature as it gives higher accuracy.

Imputation of missing values in Independent variable:



- Imputation of missing values in sub-rating columns using overall column value.
- Imputation of missing values in overall columns using average sub-rating column value.

Handling categorical variables and Date:

- Categorical features and date includes ['airline', 'author', 'review_date', 'aircraft', 'traveller_type', 'cabin', 'route', 'date_flown', 'travel_month']
- Drop categorical variables with large number of unique values.
- Performed One hot encoding for categorical variables.
- Dropped features with more than 60% null values.
- Date was split into months and day columns.

Handling text column:

- **BoW:** Create a count vector of each of the words in the customer reviews.
- **TF-IDF:** Create weighted vector using term-frequency and Inverse of the document frequencies to represent each of the reviews.
- **Review Sentiment (VADER):** VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative. VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

Review polarity vs recommend:

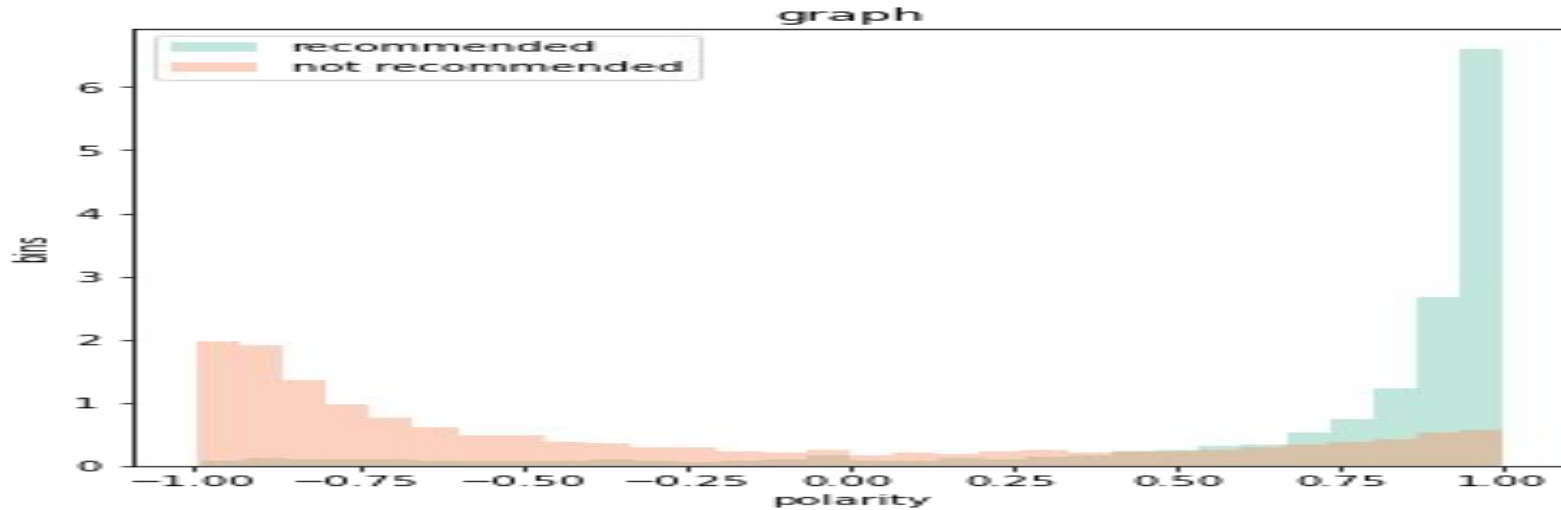
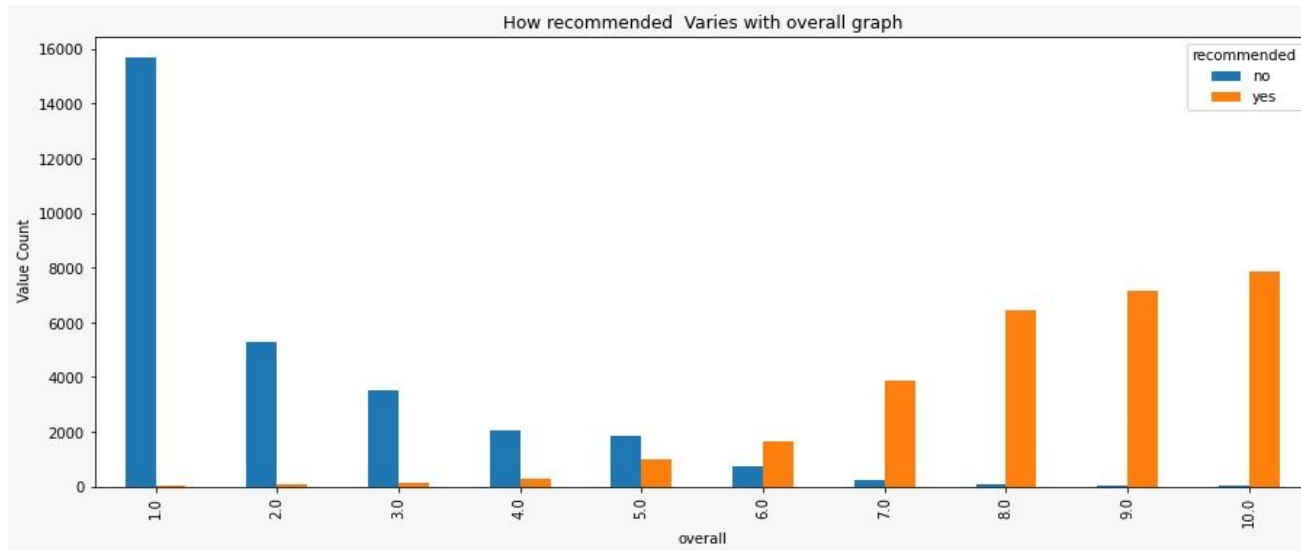


Fig: Histogram polarity

- The histogram shows the correlation of the review polarity and the recommends for the dataset.

Handling Anomaly in Target variable:



overall	recommended	values
1.0	no	16942
1.0	yes	53
2.0	no	5771
2.0	yes	73
3.0	no	3500
3.0	yes	200
4.0	no	2000
4.0	yes	500
5.0	no	1800
5.0	yes	1000
6.0	no	800
6.0	yes	1600
7.0	no	200
7.0	yes	3800
8.0	no	100
8.0	yes	6500
9.0	no	50
9.0	yes	7200
10.0	no	20
10.0	yes	7800

Fig: Before anomaly treatment

- Datapoint having overall rating 1.0 and 2.0 and still getting recommended as yes are considered Anomaly.
- Similarly datapoint having overall rating 9.0 and 10.0 and still getting recommended as No are considered Anomaly.

Train Test Split:

- Dataset size after cleaning and featurization (61183,97)
- Used 80% of data in training and 20% on test.
- Train data set has shape-(48946, 97) and Test dataset has shape-(12237, 97)

Building Classifier Models:

Models used-

- Logistic Regression
- LinearSVC
- MultinomialNB
- DecisionTreeClassifier
- Random Forest
- GradientBoostingClassifier

Different Classifier Models' Performance:

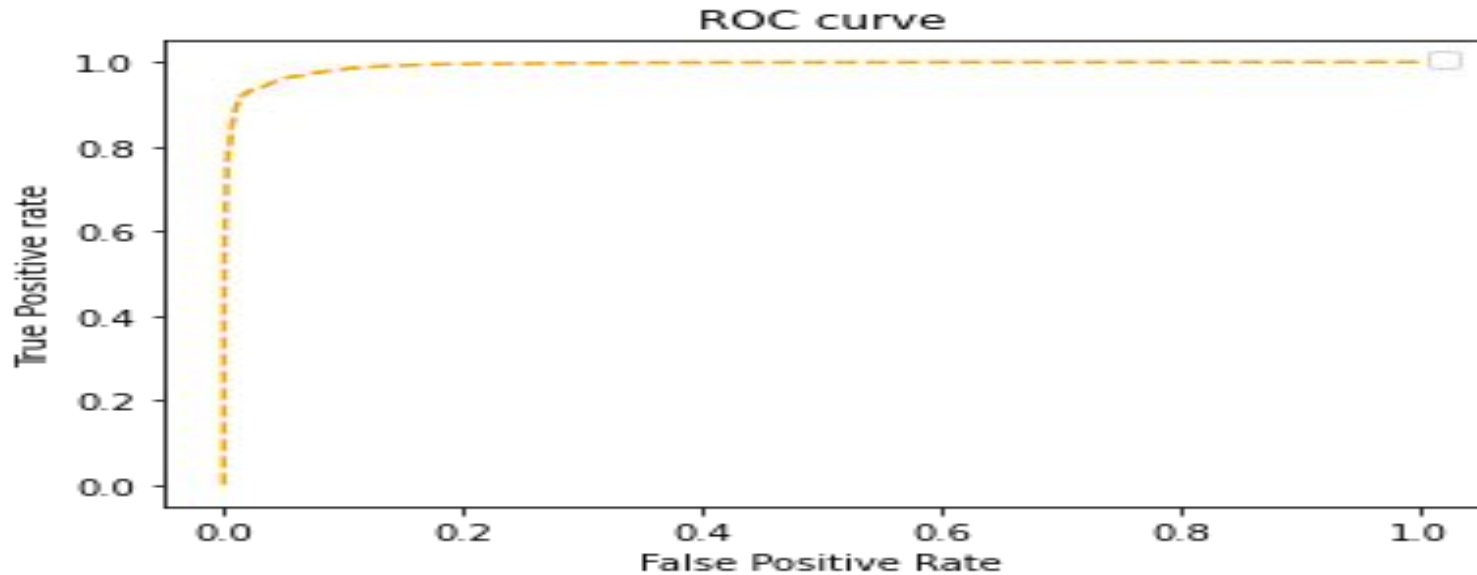
	Model_Name	Train_Accuracy	Test_Accuracy	Precision_Train	Precision_Test	Recall_Train	Recall_test	ROC_AUC_Train	ROC_AUC_Test	AUC	Model_training_time
0	LogisticRegression	96.18	96.19	95.95	96.45	96.04	95.59	96.17	96.17	0.961664	1.651352
1	LinearSVC	96.18	96.18	95.97	96.46	96.03	95.57	96.18	96.16	0.961591	0.378130
2	MultinomialNB	87.68	87.70	82.68	82.96	93.83	93.66	87.95	87.92	0.879215	0.021195
3	DecisionTreeClassifier	100.00	94.68	100.00	94.61	100.00	94.30	100.00	94.66	0.946642	0.361679
4	RandomForestClassifier	99.07	96.11	99.05	96.59	99.00	95.27	99.07	96.08	0.960788	5.452318
5	GradientBoostingClassifier	96.19	96.10	96.04	96.41	95.96	95.43	96.18	96.07	0.960723	17.829539

- Logistic Regression is the best performing model in term of accuracy as well as precision and recall value on test set

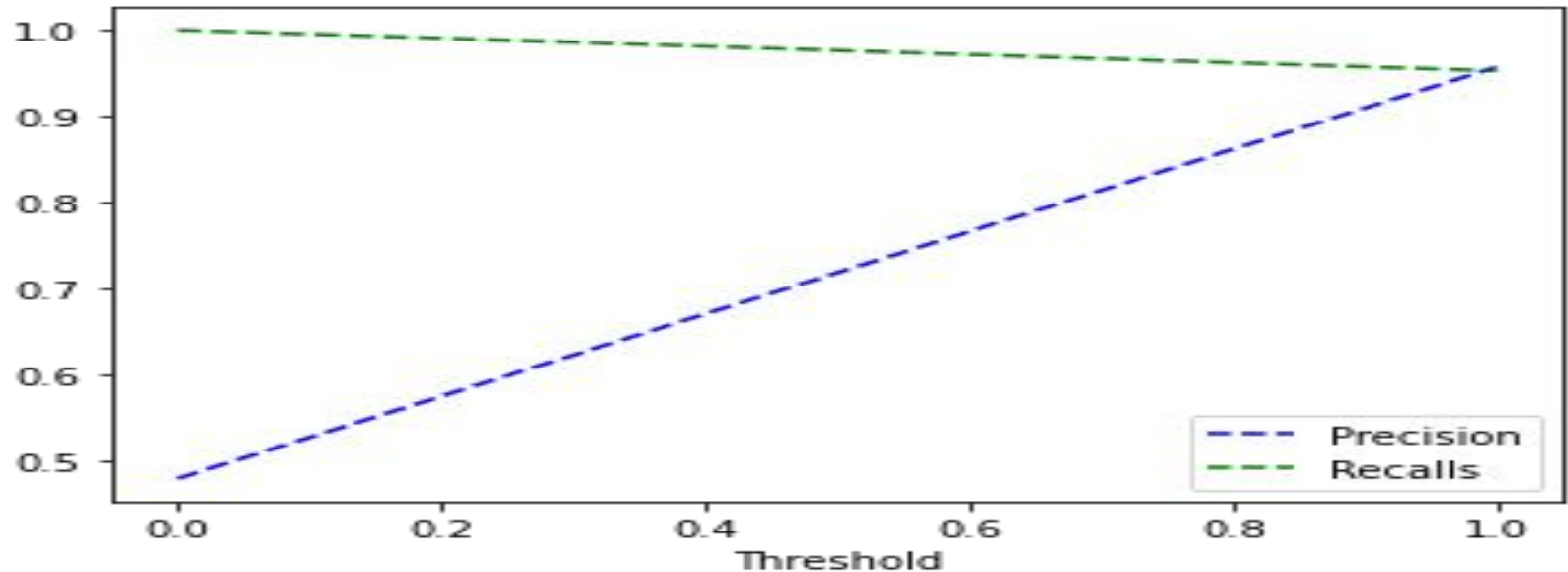
Cross validation techniques used

- Used K-Fold and RepeatedKFold techniques in LR model
 - Implemented K=10 fold for KFold and K=4 for RepeatedKFold.
 - Each fold gave an accuracy of above 95% above.

Model performance visualization using ROC-AUC curve

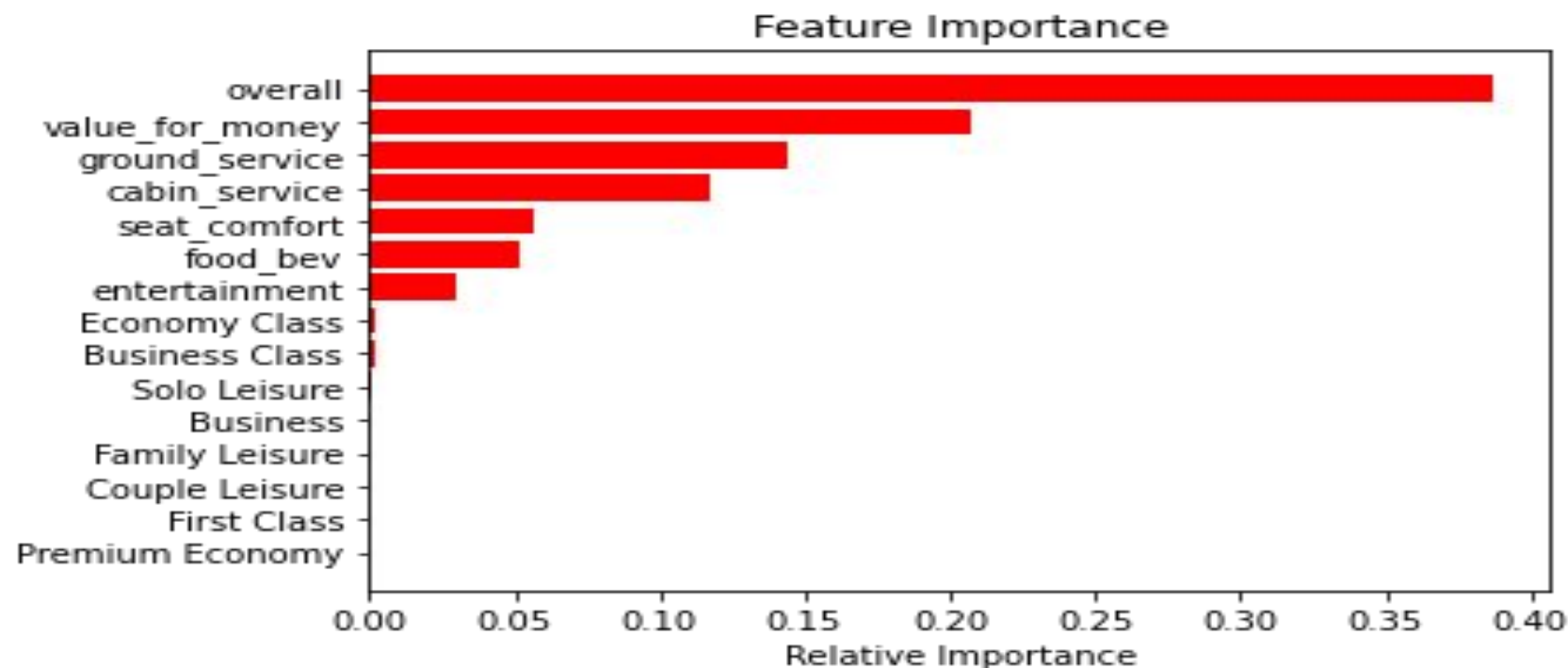


Precision and recall for different threshold values



- Low value of threshold, higher the recall.

Feature Importance



Conclusion

- We have built classifier models using 6 different types of classifiers and all these are able to give accuracy of more than 95%.
- The most important features are Overall rating and Value for money that contribute to a model's prediction.
- The classifier model developed will enable airlines ability to identify impactful passengers who can help in bringing more revenue.

Improvement

- Extra features such as flight delay time, pilot experience can be added to improve more accurate prediction.
- Increasing the data size.
- Working on removing more anomaly from data.

Challenges

- Imputing target variable missing values.
- Anomaly in the Dataset.
- Limited data.
- Chance of overfitting.
- Most of the engineered features was not much improving the model accuracy.

Q & A