Capstone Project - 3
Airline Passenger Referral Prediction

By- Shri Prakash Yadav

Points of Discussion

- Problem Statement
- Data Summary
- Bivariate Analysis
- Univariate Analysis
- Covariance
- Model Creation
- Feature Importance
- Conclusion
- Challenges

Problem Statement

- Airline is a fast and premium mode of travel.
- For a decent portion of population nowadays
- airways is the first choice. For airline giants it is
- important to know whether a passenger likes
- their services or not. It would be very
- informative for them if they can analyse and
- predict whether any passenger is going to
- recommend their flight or not. Analysing
- reviews can also help them with the services
- or features that passenger likes the most.

Data Summary

- The dataset airline review consists of 65947 rows and 17 columns which are listed below:

- **seat_comfort** : seat comfort rating (0-5)
- **cabin_service** : cabin service rating (0-5)
- **entertainment** : entertainment rating (0-5)
- **ground_service** : cabin service rating (0-5)
- **value_for_money** : value for money rating (0-5)
- **overall** : overall rating given by reviewer (0-10)
- **food_bev** : rating given to food and beverage (0-5)

Date Summary

Dates

- **review_date** : date on which review was given
- **flown _date** : date of flight

Others

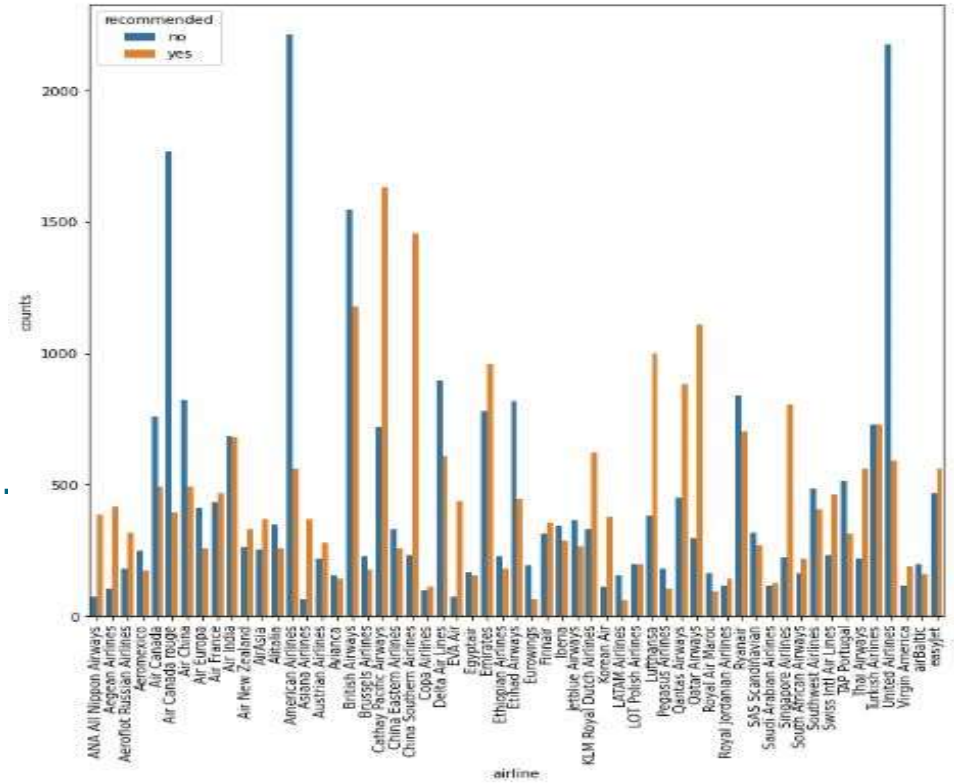
- **airline** : name of the airline
- **aircraft** : name of the aircraft
- **author** : name of the author who has given the review
- **routs** : route of flight
- **review** : review given by author
- **recommended** : whether a traveller going to recommend the flight or not

Data Summary

- **cabin_type** : type of cabin
- **traveller_type** : type of traveller

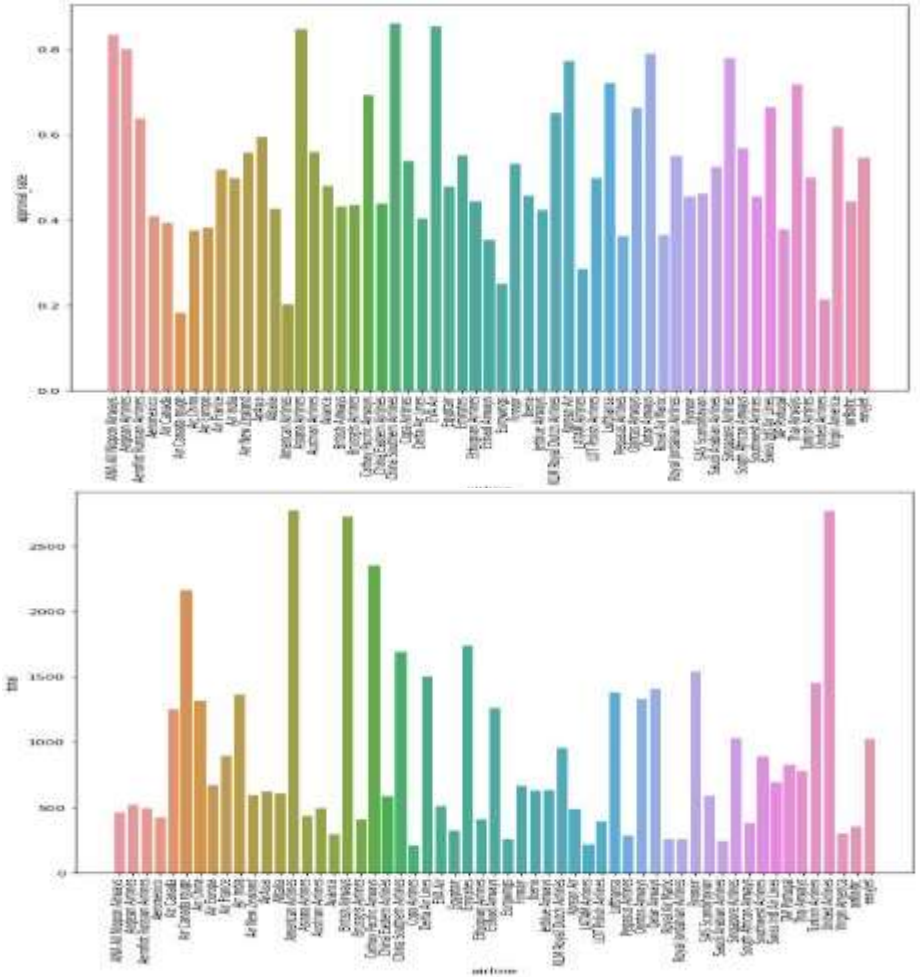
Bivariate Analysis

- American airlines and United airlines have highest number of no recommendations.
- Cathay Pacific and China Southern airlines have highest number of yes recommendations.



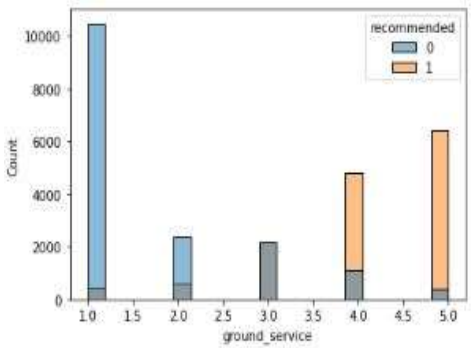
Bivariate Analysis

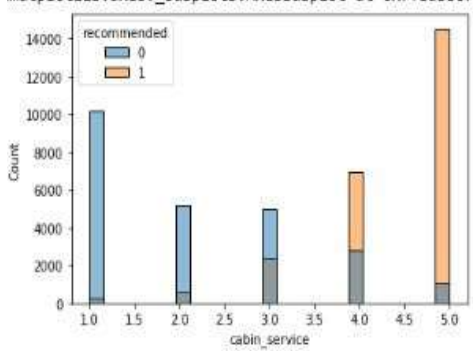
- China Southern, Asiana and Eva Air airlines have highest yes recommendation rate.
- Air Canada rouge and American airlines have the lowest yes recommendation rate.
- American airlines have received highest number of reviews.
- Copa airlines have received minimum number of reviews.

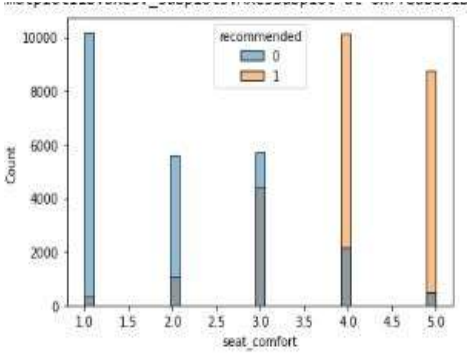


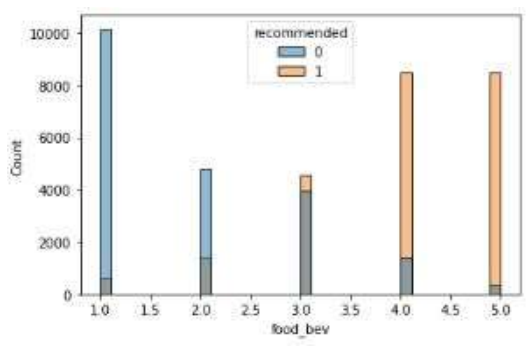
Bivariate Analysis

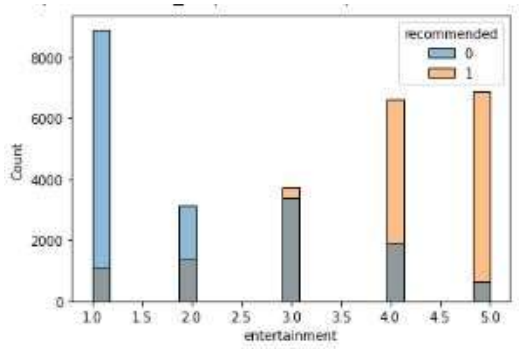
- For ground_service, cabin_ service, seat_comfort, food_ bev, entertainment, value_for_ money if rating is less than or equal to 2 the recommendation is mostly no and for rating values greater than or equal to 4 the recommendation is mostly yes.
- For rating equal to 2 the recommendation could be both yes or no.

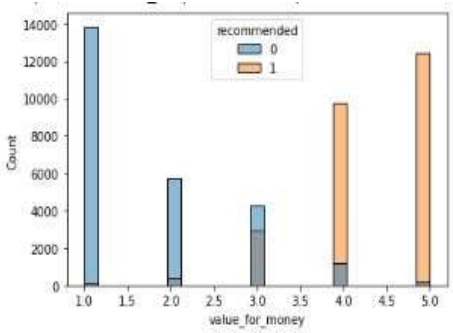






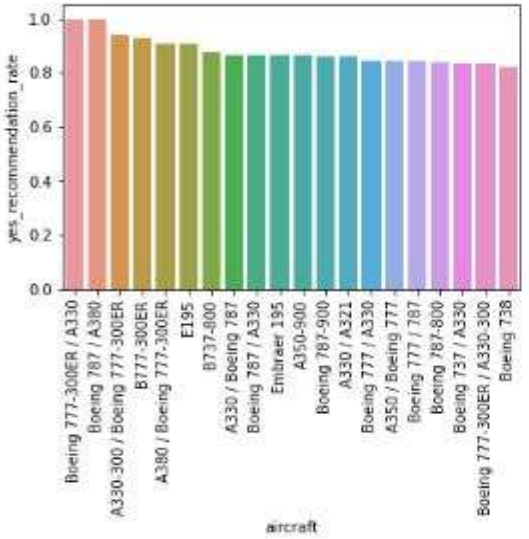


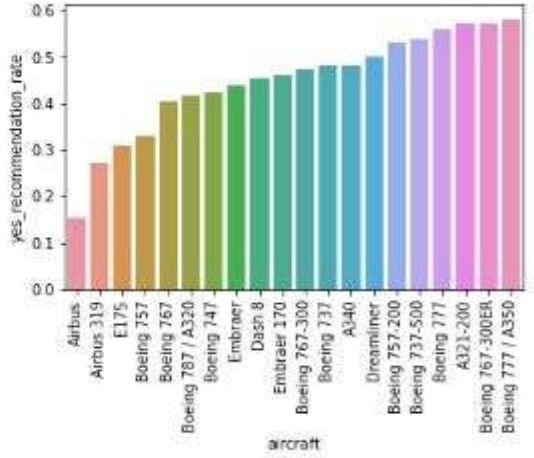


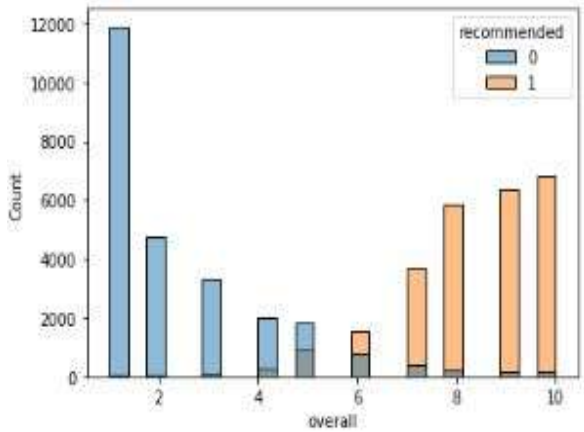


Bivariate Analysis

- Boeing 777-300ER/A330 and Boeing 787/A380 has the highest yes recommendation rate.
- Airbus has the lowest yes recommendation rate.
- For overall values greater than 6 recommendation is mostly yes.
- For overall values less than 6 the recommendation is mostly no.

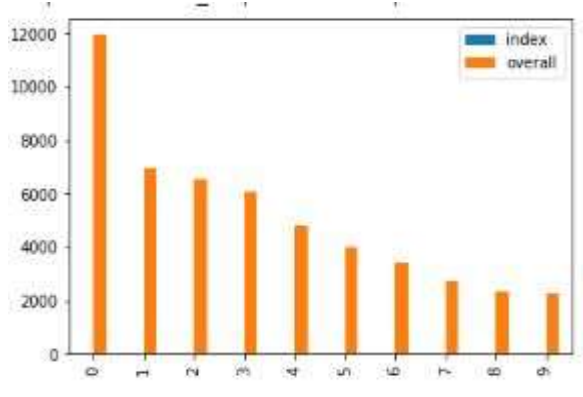


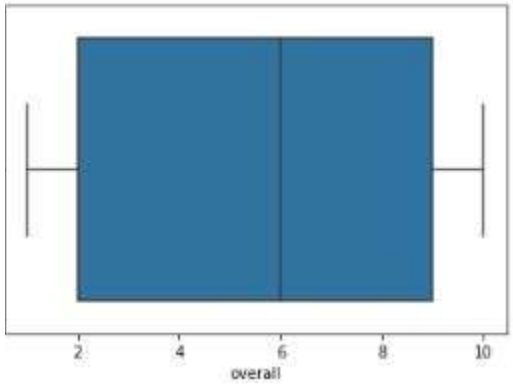


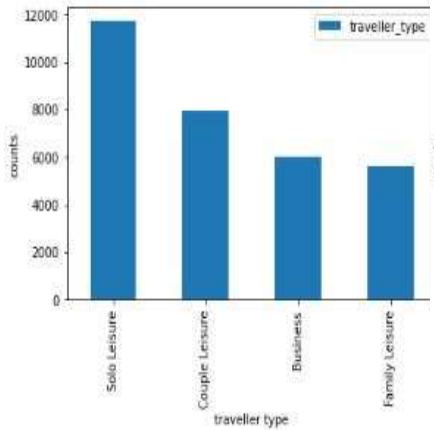


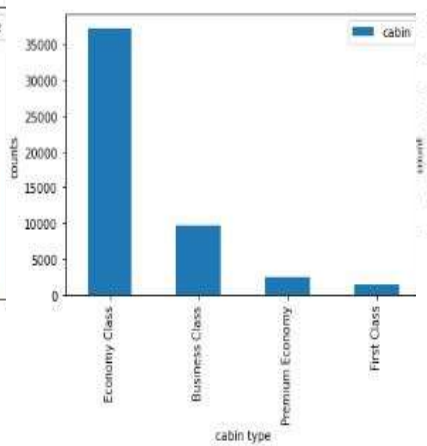
Univariate Analysis

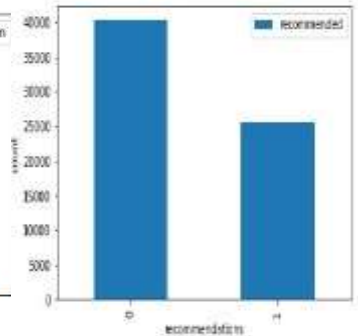
- The mean of overall is 6.
- For most of the reviews overall recommendation is below 6.
- Solo leisure type of travellers are highest.
- Travellers mostly uses economy class.
- The dataset is imbalanced.





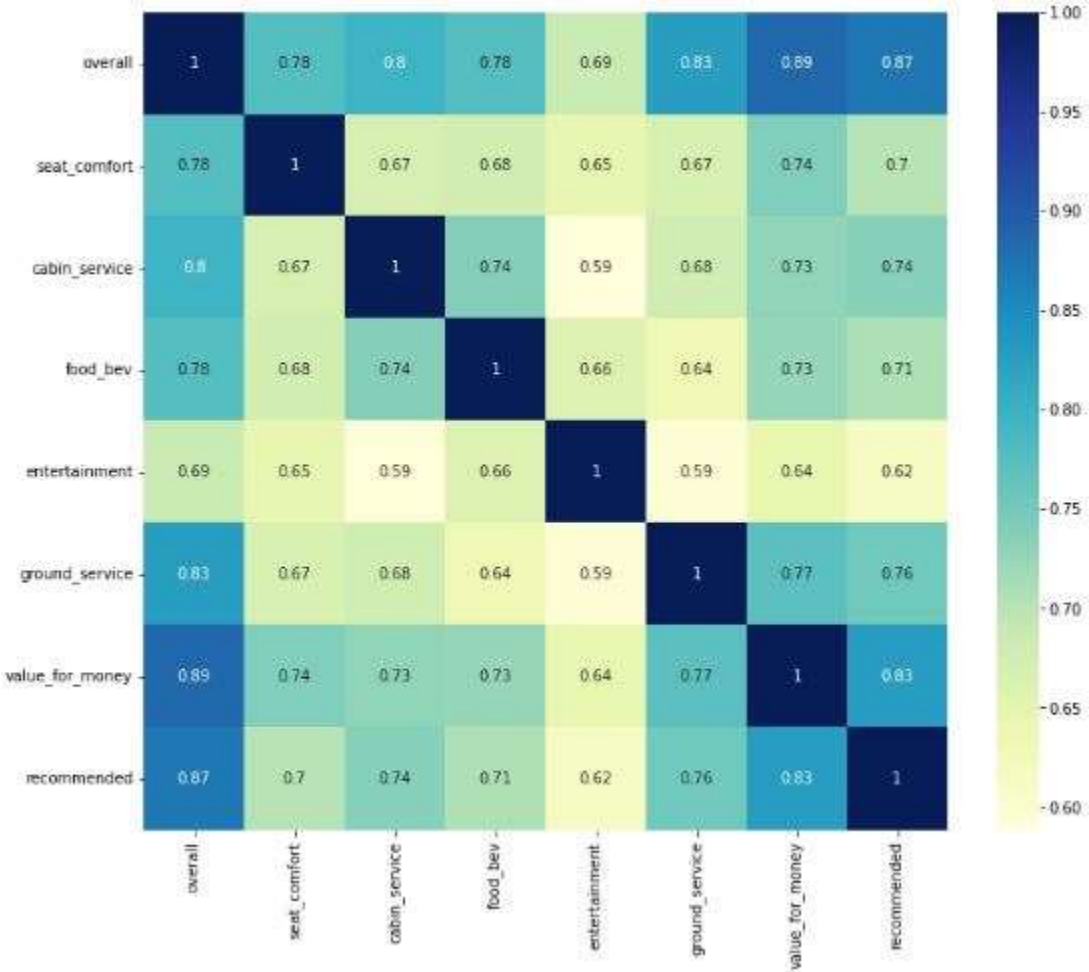






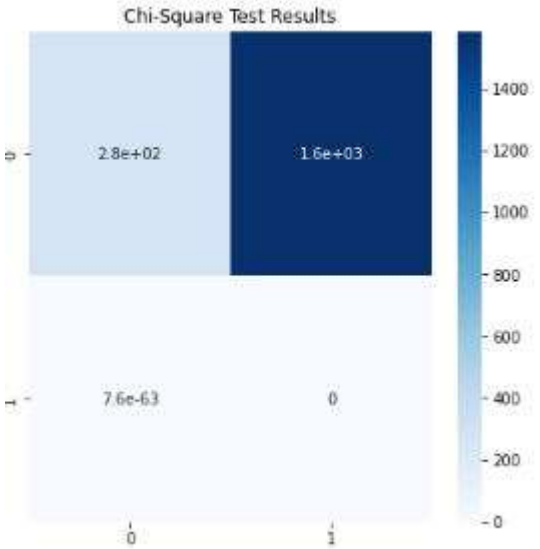
Covariance

- All the numerical features are strongly correlated with target variable.
- Recommended highest correlated with value_for_ money and overall.



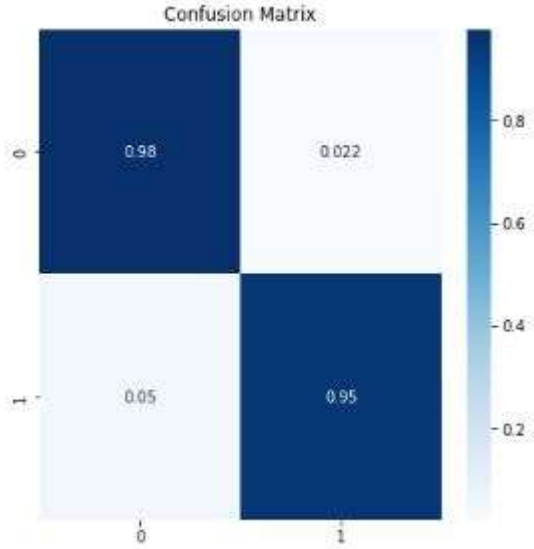
Covariance

- For both cabin_type and traveller type p_value is less than 0.05 means both are correlated with our target variable.



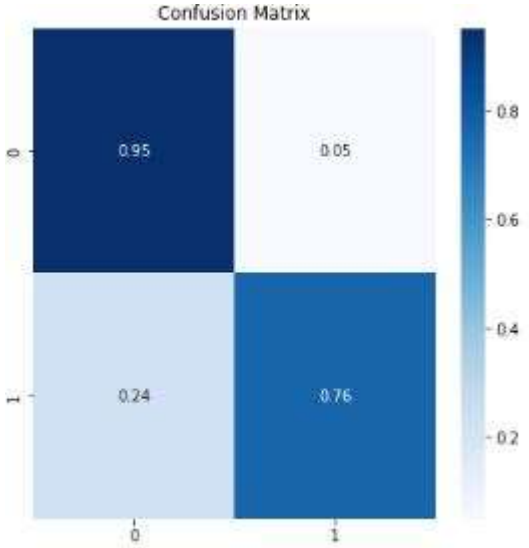
KNN Classifier

- For a distance base model KNN was performing very good.
- The best n_neighbors calculated using gridsearch cv was 85.
- Accuracy score for KNN was 0.96.
- F1 score for KNN was 0.95.



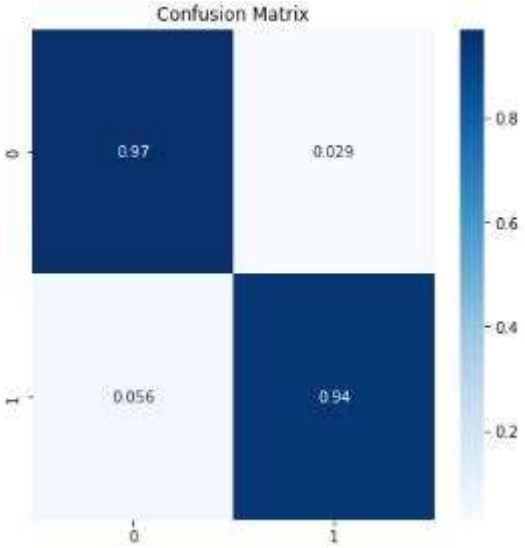
Naive Bayes

- Naive bayes was not a good choice as the dimension of dataset was high.
- The accuracy score for NB was 0.86.
- F1 score for NB was 0.83.



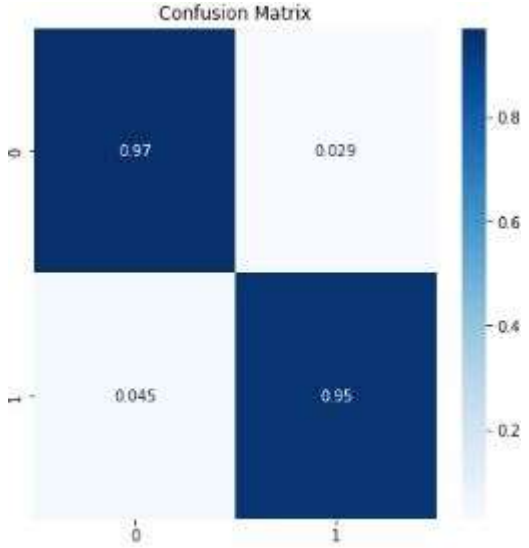
Logistic Regression

- Logistic Regression model was working like KNN model.
- Accuracy score for LR was 0.96.
- F1 score for LR model was 0.94.



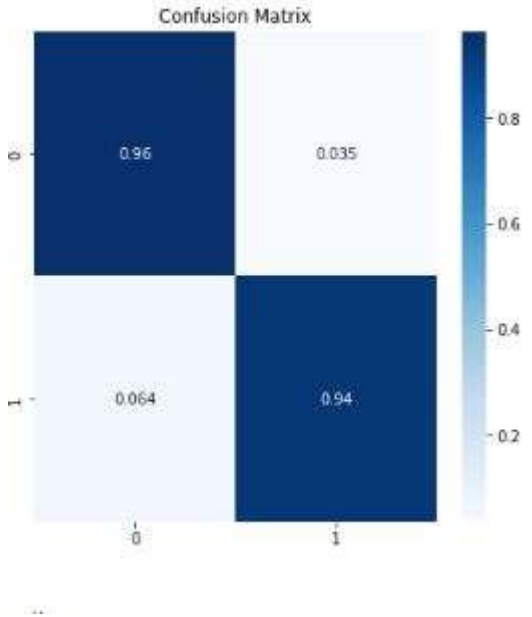
Support Vector Machine Classifier

- Support Vector Machine was working better than KNN and LR model.
- Accuracy score for SVM model was 0.96.
- F1 score for SVM was 0.95.



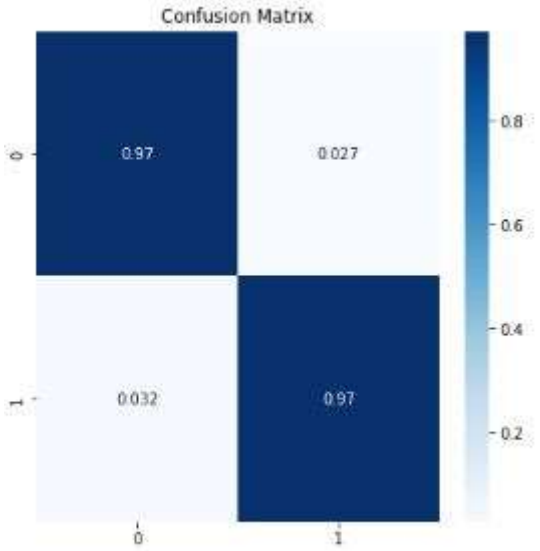
Decision Tree

- As the dataset was high dimensional still Decision Tree was performing good.
- Accuracy score for DT was 0.95.
- F1 score for DT was 0.94.



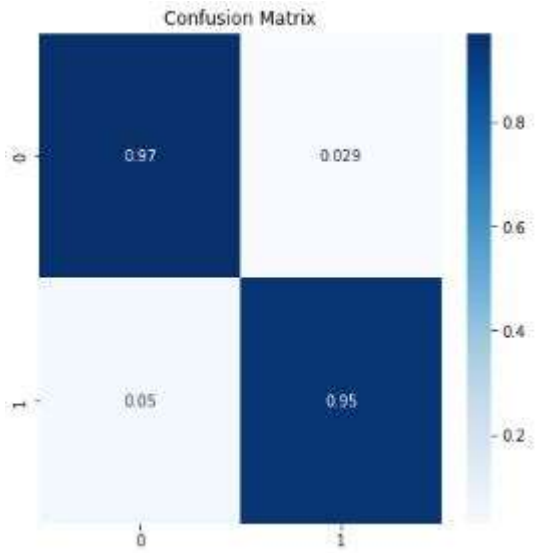
Random Forest

- The best model than I could train was Random Forest.
- Accuracy score for RF model was 0.968.
- F1 score for RF was 0.96.



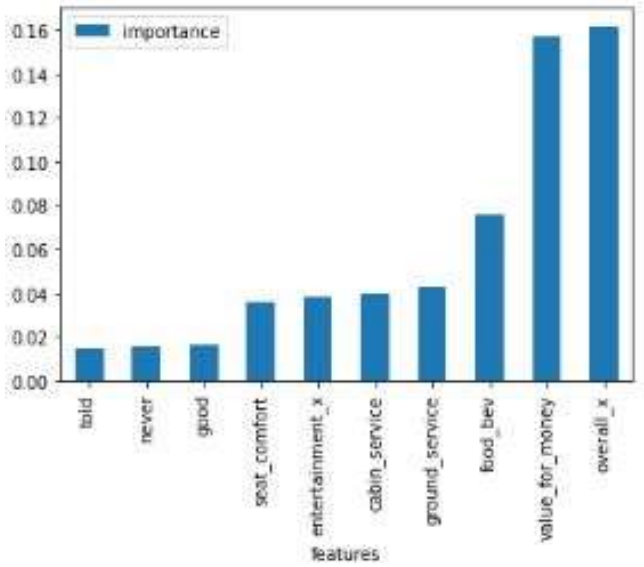
Gradient Boosting Classifier

- Gradient Boosting was also performing very good on our dataset.
- Accuracy score for GB classifier was 0.96.
- F1 score for GB classifier was 0.95.



Feature Importance

- The most important features which determines the recommendations are
- Overall, value_for_money, food_bev, ground_service, cabin_service, entertainment, seat_comfort, good, never and told.
- Good, never and told features come from bag of word vectorization of reviews.



Conclusion

That’s it! I performed EDA and modelling on airline review dataset. I was able to find various trend and dependencies from dataset. The best model that I could train was Random Forest whose accuracy score was 0.97 and F1 score was 0.96. The most important pre existing features were overall, value_for_money and food_bev rating.

Challenges

The dataset had reviews as one of its features. Converting this textual feature to numerical feature was one of the challenges. The other challenge that I faced in this project was that the dataset had a lot of null values. Handling the null values was another challenge.

THANK YOU