# Capstone Project Submission

**Team Member's Name, Email and Contribution:**


**Team Member's Name** **Email**
Ajinkya Shingote shingoteajinkya65@gmail.com


**Contribution:**
**Ajinkya Shingote:**

- Exploratory Data Analysis
- Train/Test Split
- Building Machine Learning Algorithm
- Observations
- Summarization
- Conclusions
- Technical Document
- Power Point Presentation

**GitHub Repo link.**

Github Link: https://github.com/Ajinkya6597/Air-Line-Refral-Prediction

**Project Name:** Airline Passenger Refral Prediction.

**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.

**Approach:**

- I have tested the data and done some Exploratory data analysis to build machine learning models for the prediction of the dependent factor which is the recommendation of airlines by the passenger to his\her friend.
- The given information in its crude structure was not straightforwardly utilized as a contribution to the model. A few components designing was completed where barely any elements were changed, few were dropped, and few were added.I have Engineered new features based on the existing features which are date of travel, review text, overall rating etc. I have done imputation of missing values in the target variable, I also did imputation of missing values in the independent variable. We handled categorical variables and date columns. I used NLP for handling the review text feature.
- 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.

## Conclusion:

- We can see that people have given both 1 or 0 which we will consider from now on as positive and negative recomendation so to interpret it effectively to the solo leisure. This may because of the poor infrastructure or the service recieved by the people and positive recommedation may be because of low price for solo. But this is approximate analysis based on the data provided.
- Also we can see that people gives the high positive recommendation to economic class in cabin. From this we can conclude that people love to travel in economic class as of low price also in same way we can see people give highest negative recomendation to economy class maybe because less infrastrure or service provided to them. Also we can see people have given higest positive recommedation to Business class it may be because of the quality of service provided to them in Business class and simmilarly negative recoomendation because of high price of business class or less travelling percentage.
- From month vs no. of recommendation. We can see that people tents to travel most in the month of July considering the total of positive and negative recommendation combined.
- 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 highest negative recommendation has been given to overall rating of 1.0 which is really a matter of concern.
- In seat comfort people has given highest positive recommended 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 highest negative recommendation as compare to its positive recommendation. Here we come to a conclusion it must be removed as early as possible.
- In cabin service rating people has given highest recommendation to rating to cabin service rating 5 as compare to its counterpart. From this we can conclude that cabin service is doing pretty good.
- In food and beverage rating people have given 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 entertainment also we can see most people has given highest negative recommendation to entertaiment rating 1 which shows that airline has to improve their entertainment system as well.
- In ground service also we can see most people has given highest negative recommendation to ground service rating 1 which shows that airline has to improve their ground service.
- In value for money also we can see most people has given highest negative recommendation to value for money rating 1 which shows that airline has to make

their flight service more cost effective.

- In model Selection we can see that Random Forest and XGBoost Model is having the same high Model Accuracy with a score 0.957082 but we can also see that recall, precision, f1-score and roc_auc_score of XGBoost model combined is giving higher score than Random Forest from which we have chosen XGBoost Model for further prediction.
- In 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 model.
- In Shap summary scatter plot we can see in scatter plot high overall,value for money,numeric_review,cabin service,ground_service positive features and low airline_British_airways is increasing positive prediction and it is common for all models. Also we can see that overall,value for money,numeric_review,cabin service,ground_service has high shap feature value.