Capstone Project Submission

Team Member's Name, Email and Contribution:

Contributor role :

1) **Data Wrangling -**
   i. Analyze the Data set

2) **Data cleaning-**
   i. Delete unnecessary data.
   ii. Null value treatment/Duplicate values treatment

- **Name: Sameer Satpute**
  Email: sameersatpute7@gmail.com
  Contribution:

  1) **Data Visualization-**
     - Pie chart for cabin
     - Countplot for cabin ,traveller type
     - Find correlation with heatmap

  2) **Model Explanability**
     - ELI5

  3) **Regression analysis-**
     - Decision tree
     - Random forest
     - KNN
     - SVM

  4) **Cross validation**
     - Optimize hyper tuning parameter for Random forest and KNN

- **Name: Mahima Phalkey**
  Email: mahimaphalkey@gmail.com
  Contribution:

  1) **Data Visualization-**
     - Histogram for recommend
     - Plot for feature with respect to recommended
     - Countplot for airlines

  2) **Model explanability**
     i. SHAP

  3) **Classification analysis-**
     - Logistic regression
     - XGBoost
     - Gradient Boosting
     - Naïve bayes

  4) **Cross validation**
     i. Optimize hyper tuning parameter for Gradient boosting and XGboost
     ii.

Please paste the GitHub Repo link.

GitHub Link: - https://github.com/sameersat96/Airline-passenger-referral-predicton.git
G-Drive link -

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

The airline industry encompasses a wide range of businesses, called airlines, which offer air transport services for paying customers or business partners. These air transport services are provided for both human travellers and cargo, and are most commonly offered via jets, although some airlines also use helicopters. The airline industry refers to companies that offer air transport services to paying customers, whereas the aviation industry includes all aviation-related businesses.

My First Step was to import the dataset using Pandas then data wrangling and know the features in the dataset. There are lot null value in the dataset I also use KNN imputer fill null values.

The next step is EDA in that I use different feature to know insights from dataset. I plot Countplot with respect to different features. Now after this correlation has been checked with help of heatmap there I can see highly correlated features.

Prepare dependent and independent variables for the train test split method. I apply **Logistic regression,XGboost,Random forest,KNN,SVM,Decision tree,Naïve bayes**.

Conclusion:

- In EDA part we observed that

    1. people gave a high recommendation to the economic class in cabin.This tells us that people like to travel in economy class due to the low price, but we can also see that they give economy class the highest negative ratings because they receive less infrastructure or service.
    2. From the countplot that 'Solo Leisure' has highest ratings among all whereas 'Family Leisure' has the least ratings.
    3. No' responses are more as compared to 'Yes' responses in recommended that means airlines have to focus on some aspects to make there fliers happy.
    4. Entertainment and ground service we can see most people has given highest negative recommendation to entertainment rating 1 which shows that airline has to improve their entertainment system as well.

- Logistic Regression has the highest recall value It gave the recall of 95.12% followed by SVM which gave 94.91%.

- Support Vector Machine has the highest accuracy from the models but others are also performed very well SVM gave 95.40% accuracy.

- From Eli5 we can see overall and value for money contributed more to give the positive recommendation and ground service and family leisure contributed to give negative recommendation for XGBoost .

- From Eli5 we can see overall and value for money contributed more to give the positive recommendation and Gradient Boosting model.

- According to our business need we will give first priority to recall and then to accuracy from metrics point of view because we need to find how many people will recommend.