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
     - Count plot for cabin, traveler type
     - Find correlation with heatmap

  2) **Model Explainability**
     - ELI5

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

  4) **Cross-validationion**
     - 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 explainability**
     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

Please paste the GitHub Repo link.

GitHub Link: - https://github.com/Mahima2208/Airline_Passenger_Referal_Prediction_Classification

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 Count plot 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, Gradient Descent Machine**.

Conclusion:

- In the EDA part we observed that
  - It is apparent that people gave a high recommendation to the economic class in the 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 the economy class the highest negative ratings because they receive less infrastructure or service. Likewise, the business class has received the highest rating due to the quality service offered there, while the economy class has received the lowest rating due to its price or low attendance.
  - 'British airways' has the maximum number of trips and this can be attributed to its ultra-low-cost fare compared to other airlines.
  - Clearly, 'No' responses are more than 'Yes' responses in recommended, which means airlines have to focus on some aspects to make their fliers happy.

- According to our business needs, we will give first priority to recall and then to accuracy from a metrics point of view because we need to find how many people will recommend it.
- We can see that our models have performed very well all of the models have given recall greater than 90% which means our models are performing very well.
- Logistic Regression has the highest recall value It gave a recall of 95.12% followed by SVM which gave 94.91%.
- Support Vector Machine has the highest accuracy of the models but others also performed very well SVM gave 95.40% accuracy.

- Even after using Grid Search CV our models are giving similar accuracy.
- Naive Bayes Classifier and Random forest has the lowest recall of 93.95%.
- 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 for random forest 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, and ground_service has high shap feature value.

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