Capstone Project Submission

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1). **Saugata Deb**
E-mail: saugatad56@gmail.com

- Framework establishment
- Line Plot
- Data Manipulation
- Data Preprocessing
- Feature Engineering
- NLP Implementation
- Model Selection
- Model Deployment
- SHapley Additive exPlanations
- Shap Summary
- PPT presentation

2). **Ankit Patil**
E-mail: ankit.patil67@gmail.com

- Data Cleaning
- Data Analysis
- Error Handling
- Data Manipulation
- NLP Implementation
- One hot encoding
- Feature Engineering
- Normalization
- Model Selection
- Hyperparameteric Tuning

3). **Chukkapalli Naga Sai**
E-mail: nagasaichowdary1111@gmail.com

- Debugging Error
- Data Sorting
- Technical Documentation
- ppt Presentation
- Approach Towards Plan
- Seaborn,matplotlib
- Heatmap
- Evaluation Matrix

4). **Shreyash Movale**
E-mail: shreyash9m@gmail.com

- Data Sorting
- Matplotlib
- ppt Presentation
- Data Visualization
- Approach toward Plan

- Line Plot,Barplot,Histogram
- Heatmap
- Evaluation Matrix
- Data Preparation

Please paste the GitHub Repo link.

Github Link:-
https://github.com/SaugataDeb/SaugataDeb-Airline_Passenger_Referral_Prediction

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

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.

Conclusion:
So here we come at the end of our project Airline Passenger Referral Prediction. Let's take a short recap on what we have done. In exploratory analysis we first find the duplicate we found 70711 we then drop those duplicates after dropping those duplicates we did find the info we found 61684 entries. Then we dropped those which have all null values. Then we found the percentage of passengers in different cabins using a pie plot. We found the economic class highest. Also we found Spirit Airways is the most frequently used airways. We found flight A320 to be the most frequent aircraft. Also people prefer to travel solo. July is the month where people travel most.

In feature description we did natural language processing to convert the customer_reviews sentiment based on polarity to numeric reviews. We did one hot encoding on categorical features.

In model selection we first did a train and test split in 4:1 or 80:20 split. We created functions to store evaluation metrics values.Then we did model deployment. XGBoost model had shown highest model accuracy along with highest recall,precision,fi_score and roc_auc_score. We select XGBoost for classification of our prediction.

In model explainabilibility we used 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.

References:
- GeekforGeeks
- Kaggle
- Analytics Vidya