Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1). **Ajinkya Shingote** E-mail: shingoteajinkya65@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/Ajinkya6597/Airline-Passenger-Refral-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)

The given data includes airline reviews from 2006 to 2019 for popular airlines around the world with multiple choice and free text questions. Data is scrapped in Spring 2019. The main objective is to predict whether passengers will refer the airline to their friends. We 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. Our main aim is to use the previous passenger data and predict the referrals by that passenger. The data have various influencing factors to the recommended part of the data set such as airline company and their aircraft carrier, overall rating out of 10, customer name, review date, review content, cabin, route, and some other relative factors. We get mixed data types which also included some of the outliers and almost half null values. Most of the rating categories are numerical while review which has a great influence over recommendation is categorical data. So, we have done exploratory data analysis to check out the major factors affecting the predicted recommendations by the passenger. We have done feature engineering followed by EDA to check out the important features and drop out the less important ones. After feature engineering, we have done natural language processing (NLP) review reviews and content then one-hot encoding for the categorical data. After preparing the final data set which is ready to feed to the ML model, we split it to train and test data and applied the Logistic Regression model, Decision Tree, Random Forest, Random Forest with grid search cv, Support Vector Machines (SVM), K-Nearest Neighbor (KNN) model, XG Boost, and compared different model by its evaluation matrix. After that, we did some Model explainability analysis using shap and finally drew the conclusion.