**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. Ashfaque Sayyed ([Ashfaquesayyed014@gmail.com](mailto:Ashfaquesayyed014@gmail.com))   **Contribution:**  • Outlining project plan.  • Data wrangling and maintaining data integrity.  • EDA.  • Feature Engineering.  • Training and Testing Model.  • Hyperparameter Tuning. |
| **Please paste the GitHub Repo link.** |
| GitHub Link:- https://github.com/ashfaquesayyed/Airline-passenger-referal-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 airline passenger referral prediction dataset consist of various categorical and numerical columns. The dataset consist of total of 17 columns such as airline, overall, author, review\_ date, customer\_review, aircraft, traveller\_type, recommended etc.  I followed step-by-step processes for the project like data collection, data cleaning, EDA, Visualization, Model Training and Testing, Hyperparameter Tuning, and Evaluation.  In EDA I started by checking the head of the dataset, it contains various categorical and numerical columns. I dropped some categorical features which are irrelevant with respect to target variable “recommended”. Further I check unique values and duplicate values in dataset. Later I count the duplicate values and drop them. I also worked on NAN values, I count and dropped them. Later I remove all NAN values from target variable.  I used KNN imputer for the imputation of missing values. KNNimputer is a sci-kit learn class that is used to fill out or predict missing values.  In this approach, I specify a distance from the missing values which is also known as the K parameter. The missing value will be predicted in reference to the mean of the neighbors. After imputing values I reindex the data frame to make recommended target variable at the last of the dataset. Later I do some visualizations such as boxplots to find outliers, count plot, barplot, and correlation matrices for inferences.  Further, I split the dataset into two parts a training dataset (80%) and a testing dataset (20%). A function is also created to save the performance matrices such as Accuracy, precision, recall, f-1 score, roc- auc score. The model gets trained from the training data, after training we use regression models such as logistic regression, Decision tree, random forest, K- nearest neighbors and cross-validation technique one by one and evaluated the results.  The test results of all the models are evaluated and compared. We checked performance metrics such as Accuracy, precision, recall, f-1 score, roc- auc score.  All the models performed well gave accuracy above 93% and logistic regression gave the best result with an accuracy of 95.1%. |