# Capstone Project Submission

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

**Team Member's Name, Email and Contribution:**

**Team Member's Name: Sanchit Misra**
**Email: sanchit.misra.a13@gmail.com**
**Contribution:**

1. **Imported dataset and visualize first glimpse of data**
2. **Data Wrangling:**
   - **Airline Passenger Referral Prediction shape**
   - **Null Values, Duplicates, Unique Values, Date Time feature**
   - **Encoded the features**
3. **Descriptive analysis of features in dataset**
4. **Exploratory Data Analysis:**
   - **Performed Univariate Analysis**
   - **Performed Bivariate Analysis**
   - **Performed Multivariate Analysis**
   - **Performed Outlier detection**
5. **Identified Multicollinearity**
6. **Data Preparation for natural language processing (NLP)**
7. **Text Cleaning:**
   - **Tokenization**
   - **Conversion to lowercase**
   - **Removing stop words**
   - **Lemmatization**
8. **Model Building:**
   - **Implemented TF-IDF Vectorizer**
   - **Divided data into dependent and independent variables**
   - **Train-Test split**
   - **Implemented various models**
   - **Evaluation metric for model performance**
   - **Plotted Confusion Matrix for each model**
   - **Hyper tuned model with GridsearchCV**
9. **Model performance on random text:**
   - **Implemented pipeline library**
   - **Predicted referral column based on random text given using GB model**
10. **Assessed feature importance using SHAP (Shapley Additive explanation)**
11. **A Colaboratory notebook, PowerPoint presentation and project summary were all created with team participation.**

**Team member's name: Tushar Hande**
**Email: handetushar3@gmail.com**
**Contribution:**

1. **Loaded dataset and imported dependencies for EDA and preprocessing steps.**
2. **Checked head, tail and the size of the dataset. Inspected all variables.**
3. **Checked null values counts and duplicates.**
4. **Statistical analysis of dependent and independent variables.**
5. **Analysis of the independent and dependent variables using EDA**
6. **Null value treatment and outlier handling using created functions.**

7. **Imported dependencies for NLP and downloaded required libraries.**
8. **Completed basic text preprocessing and modeling part.**
9. **Cheeked model performance and evaluation score with different metrices.**
10. **Model tuning using GridSearchCV and hyperparameter selection.**
11. **Conclusion and explainability of the models.**
12. **Contributed to develop technical document and ppt.**

**Team Member's Name: Mohit Jain**
**Email: jainmohit02.mj@gmail.com**
**Contribution:**

1. **Imported all the libraries for data exploration, Sorting, Cleaning and Visualization and Modelling.**
2. **Imported and mounted data set require for analysis from google drive.**
3. **Exploring data set such as number of columns and row with feature name and what is data type of each feature using python libraries like Info(), shape, describe(),Head(),Tail().**
4. **Checked Null Values, Duplicates, Unique Values, Date Time feature.**
5. **Exploratory Data Analysis: - Performed Uni variant analysis and bi variant analysis to check dependency of each independent variable with dependent variable and checked multicollinearity by heat MAP and dropped feature that is mostly affected.**
6. **Examine the outlier's detection in dataset.**
7. **Divided data into dependent and independent variables.**
8. **Prepared data for natural language processing (NLP).**
9. **Implemented text cleaning and tokenization process.**
10. **Performed Train-Test split on clean data.**
11. **Implemented various models and Checked Evaluation metric for model performance to identify best model as per the business requirement.**
12. **Checked model performance on randomly created data.**
13. **Evaluated and identified important feature using SHAP.**
14. **With help of all group member prepared presentation.**
15. **Helped group member to prepare technical documentations.**
16. **Made Conclusion on analysis and model evaluation metric.**

**Please paste the GitHub Repo link.**

GitHub Link:- https://github.com/jainmohit02/-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)**

Due to fierce competition in the airline industry, the airline company needs to focus on the passenger's experience and satisfaction. Customer feedback, in particular, is critical since it is an outcome measurement for business performance.
To achieve the purpose of this study we built the predictive model using customer review that is predicting that whether passengers will refer the airline to their friends. We also identified some important factors that influences the customer or passenger satisfaction.

Initially by doing EDA part we got some insights about the data. Overall rating is much effects on the satisfaction on the passenger. Cathay Pacific Airways of Hong Kong is the most likes airlines on the basis of overall ratings and 'Emirates Airlines' Airbus A380 is most popular aircraft.

We also did the outliers detection process but there were not outliers into the data. We built the model using Natural language processing so numerical features was not the priority.

For null values treatment we focused on the null values of the target column. We replaced it using overall ratings using threshold 5. If the overall ratings are less than 5, we replaced the recommended by no otherwise yes.

To build model using Natural Language Processing we used only two columns 'customer_review' and 'recommended'. Where 'customer_review' is text column and recommended is binary categorical target variable.

Initially we imported important libraries and toolkit which required to natural language processing and text preprocessing. We only downloaded things those are required.

For test preprocessing first we did Part of Speech (POS) tagging that categorize the word in a text in correspondence with a particular part of speech. In text cleaning we removed excluded from a-z and A-Z. Then we lowercase the text and split them through the space. After that we removed stop words from the text so we can only keep more informative text.

lemmatization allows end users to query any version of a base word and get relevant results, so we did lemmatization process. For vectorization process we used TF-IDF (term frequency-inverse document frequency). To get better result we used max_features hyperparameter.

For modeling part, we used Logistic regression, Naïve Bayes, PassiveAggressiveClassifier, RandomForestClassifier and GradientBoostingClassifier and we calculated accuracy score for each model.

For a better performance of the model, we tuned our model using GridSearchCV and to identify the importance features from the data we used Shap method.

For evaluation metrics we used Accuracy score, confusion matrix, precision, recall and f1 score. We also check and draw accuracy of some models using AUC ROC curve.

At the end we also built the predictive model and checked the output using some random reviews. We also checked and plotted the feature importance using Shap (Shapley Additive explanations).

Finally, we reached to the following conclusion:

- The highest peak of the month feature is 7. According to legend, July is the month with the most travel. December is the second-most popular month.
- The top spot pertains to the Bangkok to Hong Kong trip that occurred most frequently in the dataset, followed by Bangkok to London and London to New York trips.
- In the column for traveler type, it is noticeable to us that Solo Leisure travelers represent the majority of the population. In the cabin column, the majority of passengers prefer the Economy class.
- Following the use of bivariate analysis, we discovered that all travelers highly favor the economy class. Some Couple Leisure and Business class travelers choose to fly in business class. Among all traveler types, first class is the least popular.
- Due to the linear and balanced dataset, logistic regression outperformed the other algorithms well. The gradient boosting approach came in second.
- For this sort of dataset and its specified problem statement, accuracy and f1 score are the optimal evaluation matrix that is taken into consideration.