Technical Document on Airline Passenger Referral Prediction Data

Introduction:

This document provides a comprehensive overview of the dataset and the process involved in predicting airline passenger referrals based on reviews and ratings. The dataset contains information on various aspects of airline services and customer reviews, aiming to determine whether passengers are likely to recommend the airline to others.

Dataset Description:

Size: Initially, the dataset comprised 131,895 rows and 17 columns.

Columns:

airline: Name of the airline.

overall: Overall rating given to the trip (scale: 1 to 10).

author: Author of the trip.

review date: Date of the review.

customer review: Free text review by customers.

aircraft: Type of aircraft.

traveller type: Type of traveller (e.g., business, leisure).

cabin: Cabin class during the flight.

date flown: Flight date.

seat comfort, cabin service, FoodBev, entertainment, ground service: Ratings (scale: 1-5) for various flight aspects.

Data Preparation:

The dataset underwent several preprocessing steps to ensure data quality and suitability for analysis.

1.Handling Missing Values:

- Rows with entirely missing values were dropped.
- Columns irrelevant for analysis were removed.

2. Duplicate Removal:

- Duplicate rows were eliminated to ensure data integrity.

Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into the dataset and identify patterns, anomalies, and potential relationships between variables. This involved:

- Summary statistics for numeric columns.
- Visualization techniques (e.g., histograms, box plots, correlation matrices) to explore data distributions and correlations.

Feature Engineering

New features were engineered based on the existing data to enhance model performance and capture additional information:

- Binary Rating Column:
    - Created a binary column indicating positive or negative reviews based on the overall rating.
- One-Hot Encoding:
    - Applied one-hot encoding to cabin type, Airlines, and traveller type to create categorical features.

Model Building:

Various classification models were implemented to predict passenger referrals based on the provided features. The following models were used:

1. Logistic Regression
2. Decision Trees
3. Random Forest
4. Support Vector Machines (SVM)
5. Naive Bayes
6. K-Nearest Neighbours (KNN)
7. Neural Networks
8. Gradient Boosting models
9. XGBoost

Model Evaluation:

- **Accuracy**: Models achieved an average accuracy of approximately 95%.
- **Scope for Improvement**: Identified opportunities to refine anomaly detection methods and enhance the replacement of the recommendation column.

Conclusion:

The project successfully addressed the objective of predicting passenger referrals based on airline reviews and ratings. By employing various models and feature engineering techniques, a high accuracy rate was achieved. However, there are areas for improvement, particularly in refining anomaly detection and further enhancing model performance.

The technical documentation provides a detailed account of the dataset, data preparation steps, feature engineering, model building, and evaluation metrics, allowing for transparency and replicability of the analysis.