Aviation Flight Cancellation Prediction

Author: Hari Prasath P

**Summary:**

This project demonstrates the process of predicting flight cancellations using a Random Forest Classifier. While the dataset in this project is synthetic, most common attributes have been included as available in real-world aviation datasets.

The notebook is organized as follows:

**1. Introduction :**

The aim of the project is to predict if a flight is going to be canceled based on input variables, including delays in airline ratings and due to weather or technical issues.

**2. Explanation of Dataset:**

The following dataset contains 1000 samples, which I have collected random on internet and some feature variables are: departure delay, arrival delay, airline rating, impact of the weather, technical issues, and if the flight was cancelled or not.

**3. Exploratory Data Analysis (EDA):**

This section covers the exploratory visualization of the dataset. Key distributions-for example, delay distributions-are plotted to understand the underlying data. Cancellation rates based on ratings of the airlines, and concerning weather and technical issues, are visualized as these may or may not relate to the target variable.

**4. Feature Engineering :**

Then, in the hope of increasing the model's predictive power, a new feature would be created: Total Delay. This is the sum of the departure and arrival delays. Following this stage-after this engineering step-the correlation between the features should be revisited to assess the impact this has caused.

**5. Model Building:**

The idea here would be to create a Random Forest Classifier that's going to predict whether flights get cancelled or not. First, check the performance of the trained model by running a classification report, along with a confusion matrix showing the results of the model tested against actual values.

**6. Feature Importance :**

A Random Forest model was plotted for feature importance to help in understanding features that would most markedly predict cancellations.

**7. Cross-Validation :**

Cross-validation was performed so as to check if this model generalizes well on different subsets of data.

**8. Hyperparameter Tuning :**

Hyperparameter tuning was done through GridSearchCV to get the best parameters for the Random Forest model so it can give optimum performance.

**9. ROC Curve :**

Plot the ROC curve, which shows the relationship existing between sensitivity and specificity. The area under the curve gives a measure of the model performance.

**10. Conclusion :**

The project was able to show how machine learning can be applied to predict flight cancellations. The notebook does quite an exhaustive analysis, from exploratory data analysis to model evaluation and optimization. Further work may include testing more complex models or even using other data to increase the predictive accuracy.

This project highlights my skills in data science, from data exploration and feature engineering to machine learning and model evaluation. Secondly, the insights obtained from this analysis can be utilized in a real-world scenario in the field of aviation to attain operational efficiencies.

Thank You !