Flight Delay Prediction Using Ensemble Models

- ## Objective

  Develop an ensemble-based machine learning model to predict flight delays using historical flight data and contextual features such as holidays and weather conditions.

- ## Dataset and Setup

  - ✓ **Flight Data:** Sourced from BTS for January 2025
  - ✓ **Weather Data:** Synthetic data matching FL_DATE and ORIGIN
  - ✓ **Tools:** Python, pandas, scikit-learn, imbalanced-learn, holidays

- ## Preprocessing

  - ✓ Firstly collect the data from BTS dataset .
  - ✓ Check the first ten inputs , info and description.
  - ✓ Remove duplicates and cancelled flight and diverted flights .
  - ✓ Handle missing values from data using fillna method .
  - ✓ Create and select features like duplicates flights and handle missing values. reapplying to ensure data consistency. Create delayed features , create day of week features.
  - ✓ Scale numerical features to fit the data.
  - ✓ Handle class imbalance .
  - ✓ Train test split the data .
  - ✓ Models to train Random forest Classfier, Gradient boosting Classifier, Train Stacking Classifier .
  - ✓ Result of model performances.
  - ✓ Load required libraries and model to predict the delay of the flight.
  - ✓ Predict and display output using user input .
  - ✓ Model evaluation of the summary .
  - ✓ Comparison summary of all model evaluation of three models.
  - ✓ we add the additional options like holiday and wheather disruption, so here it is . So first , we can create the data, like wheather data .We can combine unique data and origin combinations, Simulate wheather data and save it . Check the first ten of from Data.

- **Model Building**

  Here we use the models like

- ✓ Random Forest Classifier.
- ✓ Stacking Classifier – Here the random forest + gradient Boosting classifier = logistic Regression
- ✓ Gradient Boosting Classifier.

- Used random_state = 42

- In this we got the highest performance is Stacking Classifier.

- **Evaluation Metrics**

Model	Accuracy	Precision	Recall	F1-Score	RMSE
Stacking Classifier	**0.976985**	**0.976245**	**0.977744**	**0.976994**	**0.151707**
Random Forest	0.974300	0.971756	0.976976	0.974359	0.160312
Gradient Boosting	0.944381	0.946759	0.941673	0.944209	1.235838

- **Insights**

  - ✓ Departure delay (DEP_DELAY) is the most predictive feature
  - ✓ Distance and carrier information also contribute significantly
  - ✓ Weather features such as wind speed and precipitation had a noticeable impact on delayed flights
  - ✓ Flights on holidays were slightly more prone to delays, validating inclusion of is_holiday.

- **Challenges**

  - ✓ BTS dataset required extensive cleaning, especially with date formats and categorical fields.
  - ✓ Simulated weather data had to be carefully generated to match the flight schedule.
  - ✓ Class imbalance initially caused poor recall; resolved using SMOTE.
  - ✓ Feature engineering required experimentation to avoid data leakage.

- **Conclusion**
  - ✓ Ensemble models outperform individual classifiers.
  - ✓ Stacking classifier provides best F1-score.
  - ✓ External features like weather and holidays enhance model generalization.
- **References**

1. **Dataset used -** https://transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?pn=1
2. **Chatgpt**
3. **Some other option-** https://github.com/HwaiTengTeoh/Flight-Delays-Prediction-Using-Machine-Learning-Approach/tree/main/data