Summary

Group 7 Ruijing Chen / Tianyu Yao / Kangxin Zheng

Project Background and Dataset

The holiday season from November to January is one of the busiest for airlines in the US. Identifying flight delay and cancellation patterns can help reduce disruptions. This project aims to predict holiday flight cancellations and estimate delays using flight data and weather conditions at origin and destination airports, identifying key factors that influence cancellations and delays.

Flight data includes time period, airline, origin, destination, performance metrics, cancellations, and delay reasons. Weather data, sourced from the NWS and NCEI, includes temperature, precipitation, visibility, wind direction, and wind speed. During data integration, we matched hourly weather conditions to flights based on the nearest weather stations to airports and adjusted times to Central Time. The final dataset contains about 9 million flights for modeling.

Data preprocess

1.Handle 'HourlyPresentWeatherType' and impute NAs with 'NORM' indicating normal weather conditions.

2.Split 'HourlySkyConditions' into numerical variable 'Sky Height' and categorical variable 'Sky Conditions'

3.Applying sin and cos function to 'HourlyWindDirection' as wind direction is cyclical.

4.Extract hour and minute of time data.

5.Impute the missing values based on the same Origin, Month, and Day Of Month.

6.Drop the rows that cannot be imputed by the method above.

Exploratory Data Analysis

The EDA reveals that most flights are on time or early, with relatively few experiencing large delays or cancellations. The main causes of delay are Late Aircraft Delay and Carrier Delay, which together account for the majority of delay time. Certain airlines, like Allegiant, JetBlue, and Frontier, show a higher proportion of large delays, while Trans States, Empire, and Peninsula Airlines have higher cancellation rates. Additionally, cancellation rates peak on Mondays and during early morning hours (0-6 AM), while midweek and daytime flights tend to have lower cancellation rates.

Cancellation prediction Model and analysis

1. Modeling Procedure

We used a logistic regression model for this analysis due to its simplicity and interpretability.
Outcome (Dependent Variable): Flight cancellation (1 = cancelled, 0 = not cancelled).
Predictors (Independent Variables):
Date-related: Year, Month, Day of the Month, Day of the Week.
Flight-related: CRS Departure Time (CST), CRS Elapsed Time, Operating Airline.
Airport-related: Origin, Destination (Dest).

Weather-related: Altimeter Setting, Dry Bulb Temperature, Precipitation, Present Weather Type, Humidity, Sky Conditions, Visibility, Wind Direction.
To address class imbalance, we undersampled the data so that the number of positive (canceled) samples was equal to the number of negative (non-canceled) samples. This helped mitigate issues arising from imbalanced datasets. We then split the data into training (70%), validation (15%), and test (15%) sets for model evaluation.

2. Preprocessing and Dimensionality Reduction

Correlation Check and VIF Calculation: Before modeling, we checked for multicollinearity between numeric variables by calculating the Variance Inflation Factor (VIF).
Principal Component Analysis (PCA): We applied PCA to the variables Altimeter Setting, Humidity, Visibility, and Sky Conditions to reduce multicollinearity and improve model stability.
Categorical Variables: We used one-hot encoding to convert categorical variables.
Sparse Feature Removal: We excluded features where the proportion of non-zero values was less than 5%, preventing the risk of a singular matrix during computation.

3. Model Results

The model achieved a Pseudo-R^2 of 0.6, meaning it explains 60% of the variation in flight cancellations. The accuracy on the validation set was 88%, indicating good predictive performance. By calculating the odds ratio (exponentiating the coefficients), we assessed the relative importance of the predictors. The analysis showed that the most influential factors for flight cancellations are: Weather Condition, Visibility, Operating Airline.

4. Strengths and Weaknesses

Significance of Variables: The p-values for each individual predictor, as well as for the overall model, were below 0.005, indicating that all the variables included are statistically significant.
Residuals Independence Assumption: During the model diagnostics, we found a slight violation of the assumption of independence of residuals. Although this issue was minimal, it suggests that there may be some correlation between residuals that could affect the model's validity. Further checks or model adjustments might be necessary.

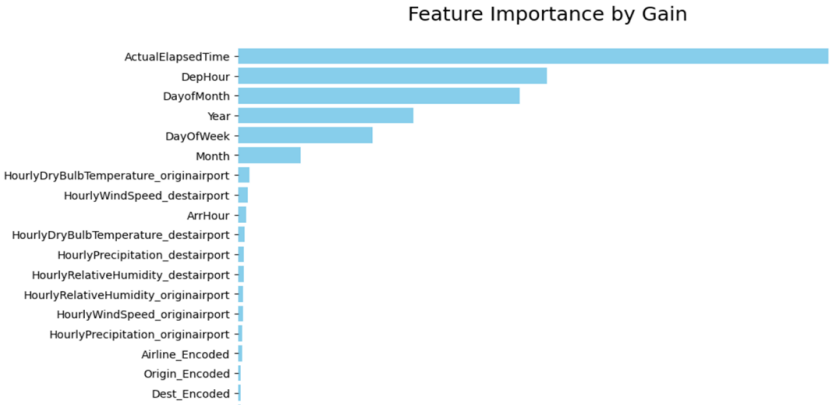
5. Takeaways for Avoiding Flight Cancellations

Monitor Weather Conditions including weather type, visibility, and wind direction at both your departure and destination airports. Check the Operating Airline.

Delay prediction Model and Analysis

The reasons we choose LightGBM as our model is that it can process large datasets faster. Additionally, LightGBM has proved better performance than other models on predicting flight delay by recent studies. LightGBM uses an ensemble of decision trees to prevent overfitting. Applying leaf-wise tree growth makes it faster and more efficient, especially with big datasets. The target variable in this model is the arrival delay minutes and predictors including flight information and weather data as the previous model. We trained the model on a training dataset comprising 70% of the data and 15 % of the validation dataset for hyperparameter tuning and evaluated it on a separate test dataset with the remaining 15% of the data.
Besides the techniques applied in the previous model for preprocessing, target encoding is a technique used to encode categorical variables by replacing each category with a numerical value based on the target variable, with smoothing to handle rare categories more effectively.
The final performance of the model is evaluated by RMSE equals 30.18, suggesting that, on average, the model's predictions differ from the actual delays by about 30.18 minutes.

Feature importance, are calculated within LightGBM by gain. Gain represents the improvement in the model's objective function (e.g., mean squared error for regression) brought about by each split on a particular feature across all the trees in the model. For each feature, LightGBM calculates the sum of the gains contributed by every split involving that feature across all trees. Features with higher cumulative gains are considered more important because they contribute more to reducing the error or improving the accuracy of the model.
The most important features are actual elapsed time ,followed by temporal patterns including departure hour, day of week, and flight date, and weather patterns including temperature, wind speed, precipitation and humidity.



We use Partial Dependence Plot to take a closer look for single variables. For example, flights with duration longer than 2 hours will decrease delay minutes. For departure hours, arrival delay minutes are low and stable in the early morning hours, increase from early morning toward afternoon and peak around 18:00.For day of week, delay minutes are lower during the midweek, possibly due to lower flight traffic.
So, from the analysis above, the takeaways to arrive on time/early are: People should be careful taking those with flight duration less than 2 hours. Also, people should choose flights that depart in the early mornings and in the midweek. Besides that, wind speed is an important factor in that higher wind speed will lead to longer delay minutes, same for departing from/arriving at colder cities.
Strength about this predicting model is that LightGBM has proven outstanding performances than other models. Besides the preprocessing done in the predicting canceled-or-not model ,we apply target encoding with smoothing on categorical variables on the training dataset, which captures the direct relationship between each category and the target variable. It also provides a smoothed estimate based on a combination of the category mean and the global mean, allowing the model to handle rare categories more effectively. Weakness is that there is still space for RMSE improvement indicates worse performance on handling extreme outliers. Also, LightGBM is highly dependent on tuning hyperparameters, which needs more effort(methods like GridSearch).

Contribution:
Ruijing Chen: writing the time difference conversion code, making shiny app and responsible for the introduction and Exploratory Data Analysis part in the summary & ppt & presentation.
Tianyu Yao: merge flight and weather data, build delay prediction model, provide part of the data processing form for shiny app, responsible for Delay prediction Model and Analysis in summary & ppt & presentation.
Kangxin Zheng: data cleaning, Cancellation prediction Model and analysis, providing some data processing forms to shiny app, responsible for Cancellation prediction Model and Analysis in summary & ppt & presentation.

References:
Kiliç, Kerim, and Jose M. Sallan. 2023. "Study of Delay Prediction in the US Airport Network" *Aerospace* 10, no. 4: 342.
J. Tao, H. Man and L. Yanling, "Flight delay prediction based on LightGBM," 2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Changsha, China, 2021, pp. 1248-1251.
https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html
Stefanovič, Pavel, Štrimaitis, Rokas, Kurasova, Olga, Prediction of Flight Time Deviation for Lithuanian Airports Using Supervised Machine Learning Model, *Computational Intelligence and Neuroscience*, 2020, 8878681, 10 pages, 2020. https://doi.org/10.1155/2020/8878681