Report - Tuesday Group 4

Yuchen Xu, Mario Ma, Yudi Wang, Yiteng Tu

Introduction/Motivation: In this project, we utilized flight data from the U.S. Department of Transportation and weather data from the National Centers for Environmental Information. Data was collected during the holiday season (November through January of the following year). Using this merged dataset, we developed models to accurately predict flight cancellations and delays. Based on our model,

1.To **avoid cancelled flights**, passengers should: choose the right airport, select airlines with lower delay rates, travel on less busy days and avoid peak holiday periods.

2.To **arrive early or on time** to their destination, passengers should: choose the right airport, select airlines with lower delay rates, travel during non-peak times and opt for better visibility conditions.

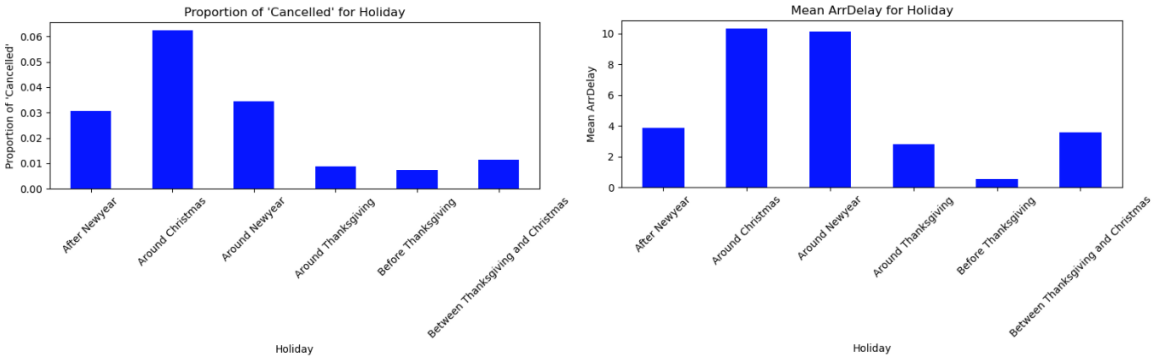
3. Using the LightGBM model, we achieved an MAE of 20.46 minutes. Key features for predicting delays are Origin, Destination, Airline, Holiday, along with weather factors like visibility, wind speed, and temperature.

Background Information about Data and Exploratory Data Analysis:

We merged flight data with U.S. Department of Transportation weather data, using matched weather station IDs and airports. For airports missing station IDs, we identified up to three nearby stations within 50 km to impute missing data. Missing weather data for specific airports (e.g., FCA, PGV, PIH) was set to NA. All times were standardized to Central Time (CT), and weather data before at most an hour of each flight's departure at both origin and destination airports was imputed to ensure accuracy.

1.For variable selection, we removed data unavailable pre-departure, such as actual departure times and weather strings (e.g., "HourlySkyConditions") and inaccessible variables like "HourlyPressureChange."

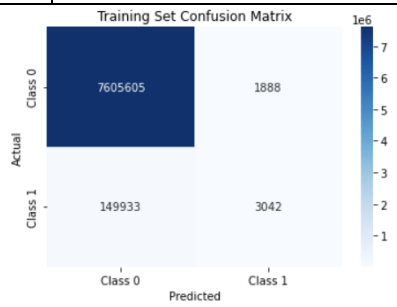
2.For data processing, date-related variables were converted to categories, with "DayofMonth" replaced by a new category. "TimeBlk" categorized by hour and "WindSpeed" was treated as categorical in 45-degree intervals. In the weather data, 'T' in "HourlyPrecipitation" was imputed as 0.05, and 's' values in "HourlyWindSpeed" were removed.



3. From the two charts above regarding cancellation rate and average ArrDelay, we can observe a strong correlation between our two targets, cancellation and ArrDelay, with the date. The cancellation rate and ArrDelay both peak around Christmas and New Year. To capture these relationships in the model, we created additional categories by grouping "DayofMonth" into six categories based on three key holidays.

4.Missing values were handled by imputing 0 for "HourlyPrecipitation" and "HourlyWindGustSpeed," using cluster means for other numeric variables, and assigning an "NA" category for categorical variables.
After processing, the final data contained 9,700,585 rows and 34 features with no missing values.
Part 1: Simple tips/takeaways to **avoid cancelled flights during the holiday season And Part 3: Prediction Model for Cancelled Flights**
We one-hot encoded categorical variables and split the dataset into an 8:2 training-to-test ratio. Due to the highly imbalanced nature of the data, we tested two approaches: one using SMOTE to resample the training set, and another without resampling. We evaluated logistic regression, LightGBM and Random Forest.

Model	Test Accuracy	ROC-AUC of testing
regression without SMOTE	0.98	0.81
regression with SMOTE	0.71	0.77
LightGBM without SMOTE	0.97	0.88
Random Forest without SMOTE	0.76	0.79



After analyzing each model's accuracy, ROC-AUC, and confusion matrix, we chose regression without SMOTE, as SMOTE led to overfitting, misclassifying non-canceled flights as canceled—an unacceptable error given its potential impact on passengers.
From our model, here we can give some tips:
1.**Choose the Right Airport**: For flights from ORD to JFK, choose MDW over ORD to reduce delays by about 7.93 minutes.
2.**Select Airlines with Lower Delay Rates**: Avoid high-delay airlines like JetBlue and Allegiant Air. Select Delta, Republic, or Endeavor Air to minimize delays.
3.**Travel on Less Busy Days**: Choose to travel on Wed, Tue, or Sun to reduce delays. Avoid Mondays.
4.**Avoid Peak Holiday Periods**: Travel outside peak holiday periods. Traveling "Around New Year" can increase delays by 10%.
Part 2: Simple tips/takeaways to arrive early or on time to their destination during the holiday season And Part 3: Prediction Model for Arrival Times
Here, we treat ArrDelay as a continuous variable. Instead of using a binary 0-1 variable to indicate whether an arrival is early, we use its positive or negative value to represent arriving early or delayed, with 0 indicating on-time arrival.

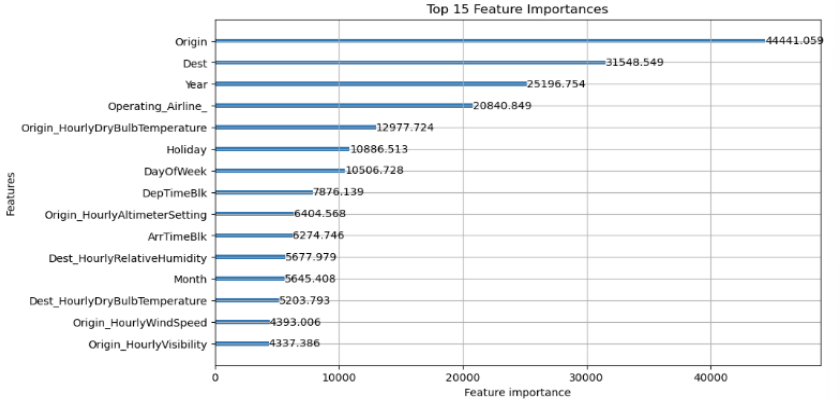
Models	Logistic Regression	Random Forest	LightGBM	Netural Network
TEST MAE(min)	23.09	23.04	20.46	23.09

We explored LightGBM, Random Forest, Neural Network and Logistic Regression models for predicting flight delays, where the response variable is ArrDelay, and the features include all available weather and flight data including flight-related data (e.g., origin, destination, airline), weather data (e.g. temperature, humidity, wind speed), and temporal information (e.g. day of the

week, month). We also transformed Day of Month into a Holiday category, which was added to the model.

We split the dataset into training and test to optimize performance and avoid overfitting. Our final model is LightGBM, and we evaluated its performance using metrics such as R^2, RMSE, and MAE. Ultimately, we selected LightGBM based on its performance with the MAE metric, achieving an MAE of 20.46 minutes on the testing set, as LightGBM proved particularly well-suited for handling large and complex datasets efficiently. Additionally, LightGBM offers interpretability, can handle NaN values directly, and has a shorter training time compared to other models. To further enhance model performance, we applied a log transformation to normalize the skewed delay data, Huber Loss to reduce the impact of outliers, and Randomized Search CV for efficient hyper-parameter tuning.

The feature importances (as shown in the figure below) highlight the most influential variables in predicting delays, with Origin, Dest, Year, and Operating Airline being among the top contributor, which aligns with some plots in our data visualization. Additionally, we used SHAP values to interpret the model further, providing insights into how each feature impacts delay predictions at an individual level.



Based on our model, here we can give some tips:

1.Select airlines with lower delay rates. For example, choosing Delta over JetBlue from ORD toJFK can decrease delay time by around 20 minutes.

2.Travel during non-peak times and avoid holidays. Shifting travel from "Between Thanksgiving and Christmas" to "Around New Year" can increase delays by 15%.

3.Consider weather condition like visibility. A drop in visibility from 20 km to 5 km is associated with a 14.48% increase in delays, according to our model predictions.

4.Plan Departure Times Carefully: Flights after 6 PM and on Sundays or Mondays are more likely to delays.

Conclusion/Discussion: In conclusion, our LightGBM model using flight and weather data shows that choosing the right airport and airline, avoiding peak holiday travel, and considering visibility can effectively reduce delays. These insights can help passengers improve travel punctuality during the holiday season.

However, there are some limitations to our model. While LightGBM handles large datasets well, it struggles with extreme outliers, which affects prediction accuracy. Additionally, our R^2 results were relatively low, indicating that the model does not fully capture the variance in delay times. This may be partly due to issues with data cleaning, as some data inconsistencies or missing values could still be influencing the model's performance. Further refinement of the data and exploration of additional features may improve its accuracy.

Contributions：

	Yuchen Xu	Mario Ma	Yudi Wang	Yiteng Tu
Code	model construction and diagnostic	Data download, model construction	Data cleaning, model construction and diagnostic	Data download, cleaning and visualization
Summary	Model part	Review	Model part	Introduction, data cleaning sections
Shiny App	Review	Frame building, realization and Beautification	Review	Review
Presentation	Part2 and Part3	Summary and shiny	Part1	Data cleaning

References:

[1] Seongeun Kim & Eunil Park(2024). Prediction of flight departure delays caused by weather conditions adopting data-driven approaches
[2] Sun Choi, Young Jin Kim, Simon Briceno, Dimitri Mavris(2016) Prediction of weather-induced airline delays based on machine learning algorithms
[3] Blog from Visual Crossing Weather
[4] YuYanyinga, HaiMoa, LiHaifeng(2019) A Classification Prediction Analysis of Flight Cancellation