Statistical Models
12.12.24
Devi Mahajan

Final Project Part 2

Business Problem

A travel insurance company is interested in understanding the factors that contribute to a flight being delayed in order to evaluate claims and alter their policies for flight reimbursements.

Logistic Regression Model

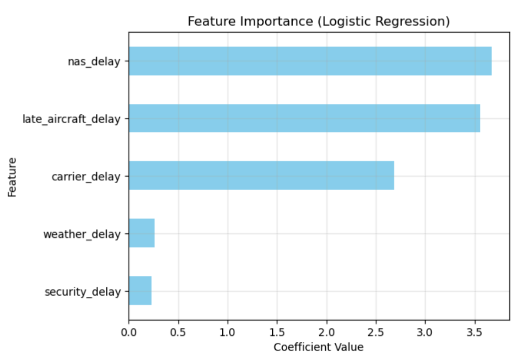
I have built a logistic regression model using a cleaned [Kaggle dataset]() of flight delay data from US airports, categorized by carriers. The dataset contains 150128 rows. A logistic regression model was chosen because my goal was to categorize flights based on them being delayed or not delayed. A logistic regression model is best suited for situations where there is a binary outcome – in this case, 'delayed' or 'not delayed'. The model formula is as follows:

`logit(p) = 2.2751 + 2.6860 * carrier_delay + 0.2634 * weather_delay + 3.6682 * nas_delay + 0.2293 * security_delay + 3.5569 * late_aircraft_delay`

The tuned model accuracy is 0.92. Here, the response variable is whether or not a flight has been delayed by more than fifteen minutes (as indicated by the variable `arr_del15`, which gives the number of flights that were delayed by more than fifteen minutes, and for our purposes has been adapted to a binary variable). The model has 6 parameters, and the 5 features included are as follows:

- `nas_delay`: delay attributed to the National Airspace System
- `late_aircraft_delay`: delay attributed to late aircraft arrival
- `carrier_delay`: delay attributed to the carrier
- `weather_delay`: delay attributed to the weather
- `security_delay`: delay attributed to security

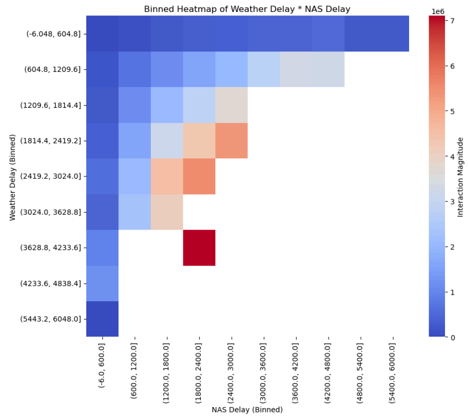
Here, `weather_delay` and `security_delay` are the weakest features and have the least influence on whether or not a flight is delayed, while `nas_delay` and `late_aircraft_delay` are the strongest features.



Key Considerations

- None of the features are highly correlated with one another, suggesting that each delay is attributed to only one cause

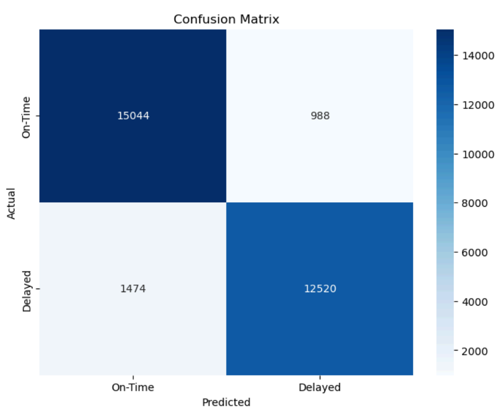
- Interaction terms such as `nas_delay * weather_delay` and `carrier_delay * late_aircraft_delay` were explored during model development but were found to have minimal contribution



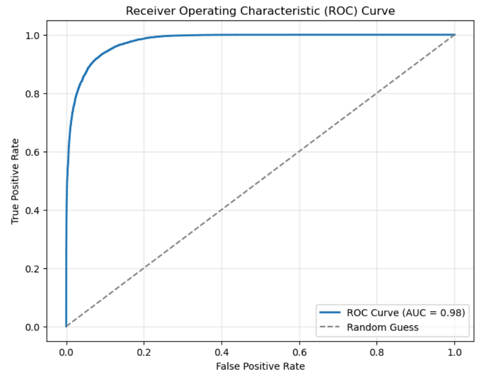
- Exploring different interaction terms (e.g. `late_aircraft_delay * nas_delay`) could improve the model slightly but show no logical overlap and might contribute to overfitting
- Variables were scaled to address differences in magnitudes across predictors
- We omit `airport` (as a one-hot encoded categorical variable) to avoid multicollinearity as well as to improve explainability of the final model

Performance Evaluation

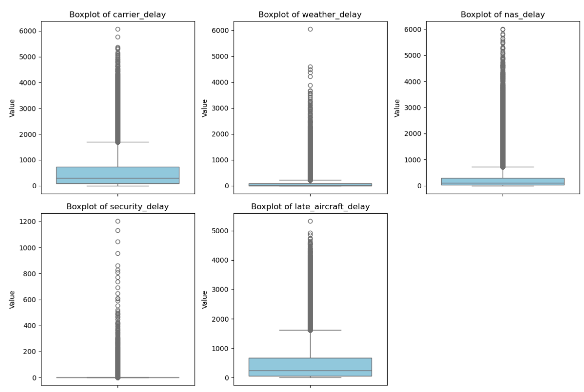
- Below is a confusion matrix generated to visualize the performance of the model



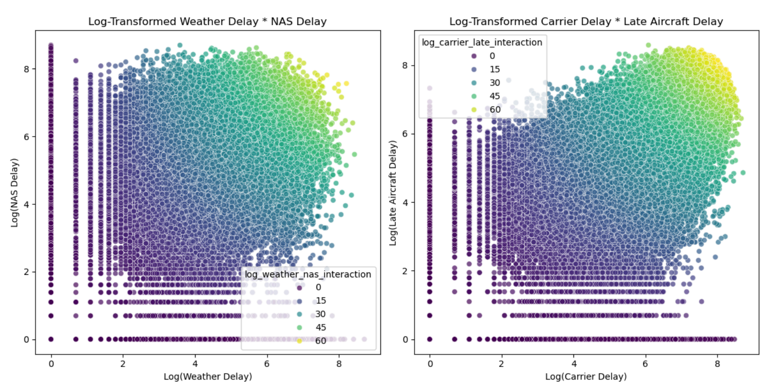
- We have the following metrics:
    - Accuracy: **92%**
    - Precision (delayed flights): **93%**
    - Recall (delayed flights): **89%**
    - F1-score (delayed flights): **91%**
- Since the model has high accuracy and precision, we can say that it performs well overall and rarely predicts delays when a flight is actually on time
- However, 89% recall is discouraging, as it shows that the model sometimes fails to predict a delayed flight and instead classifies it as on time – ideally we would target a higher recall and fewer false negatives overall



- The shortcomings of the model relate to the presence of large delays that create significant outliers in each feature, as seen in the boxplots below



- The model is highly sensitive to outliers, which may be solved by applying log transformations



Other Insights

- We assumed a linear relationship between the predictors and the log-odds of the response, and though we attempted to use interaction terms to mitigate this, the terms did not improve the model and were instead discarded
- To capture non-linear relationships we could utilise supervised learning approaches such as Random Forest models
- The current features may miss seasonal effects such as extreme winter weather, as well as geography-specific effects such as hurricanes and tornadoes; this could be improved or investigated by one-hot encoding `airport` and using `month` as a feature

Appendix 1

Model reports:

```
Intercept: 2.2750592844247635

Coefficients:
               Feature  Coefficient
2            nas_delay     3.668182
4  late_aircraft_delay     3.556894
0        carrier_delay     2.686042
1        weather_delay     0.263401
3       security_delay     0.229323


Accuracy: 0.9180043961899687

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.94      0.92     16032
           1       0.93      0.89      0.91     13994

    accuracy                           0.92     30026
   macro avg       0.92      0.92      0.92     30026
weighted avg       0.92      0.92      0.92     30026


Best Hyperparameters: {'C': 100, 'penalty': 'l2', 'solver': 'liblinear'}
Tuned Model Accuracy: 0.92

Classification Report for Tuned Model:
              precision    recall  f1-score   support

           0       0.91      0.94      0.92     16032
           1       0.93      0.89      0.91     13994

    accuracy                           0.92     30026
   macro avg       0.92      0.92      0.92     30026
weighted avg       0.92      0.92      0.92     30026


Feature Importance:
                Feature  Coefficient
2            nas_delay     3.668182
4  late_aircraft_delay     3.556894
0        carrier_delay     2.686042
1        weather_delay     0.263401
3       security_delay     0.229323
```

Please find reproducible Python code attached separately