Bobga-Herman Gwanvoma

Bellevue University

DSC 680 – Applied Data Science

Professor Amirfarrokh Iranitalab

Q&A proposed in Milestone 2

**Predicting Flight Delays Using Machine Learning**

1. How do you handle missing data in your models?

We handled missing data by using imputation strategies. For numerical columns with missing values, we imputed the missing values with the mean of the respective columns. This approach was chosen to minimize bias, but alternative strategies like k-nearest neighbors imputation or multivariate imputation could be explored for better accuracy in future iterations. Additionally, we made sure that the missing data did not disproportionately affect the target variable.

2. What measures did you take to deal with class imbalance in the dataset?

To address the class imbalance (where most of the flights had no significant delay), we used undersampling of the majority class (non-delayed flights) during model training. This ensured that both classes (delayed and non-delayed) were represented more equally, preventing the model from being biased toward predicting the majority class. In future work, SMOTE (Synthetic Minority Over-sampling Technique) could be used to generate synthetic samples for the minority class.

3. Why did you choose Random Forest and XGBoost for this project?

We chose Random Forest and XGBoost because both are powerful ensemble methods that excel in handling complex, non-linear relationships between features and the target variable. Random Forest works well with large datasets and can capture intricate patterns by constructing multiple decision trees, while XGBoost (a gradient boosting model) is known for its predictive power and efficiency in handling high-dimensional data, making it ideal for this classification problem.

4. What do the performance metrics (accuracy, sensitivity, specificity) indicate about your model?

* Accuracy: Both models achieved 100% accuracy on the test set, meaning the models predicted all instances correctly. However, this might be influenced by the class imbalance in the dataset.
* Sensitivity (Recall): This metric reflects how well the model identifies actual delayed flights (true positives).
* Specificity: This metric shows how well the model identifies non-delayed flights (true negatives). The perfect scores for both sensitivity and specificity indicate that the models were able to correctly classify both delayed and non-delayed flights without errors.

5. How can the model be deployed in a real-world setting to predict delays?

The model can be deployed in real-time systems for airlines. By integrating with real-time flight data (e.g., current weather conditions, air traffic), the model can predict delays before departure, allowing airlines to adjust schedules and inform passengers. For deployment, the model could be wrapped in an API that receives input features, performs the prediction, and returns delay forecasts.

6. How do you ensure that your model doesn’t introduce any bias against specific airlines or airports?

We carefully preprocessed the data and ensured that no features in the model favored any particular airline or airport. The undersampling of the majority class also helped prevent bias toward non-delayed flights. Additionally, we ensured that the data was balanced and reflected a variety of airlines, airports, and flight types. In future, further fairness measures, like bias detection algorithms, can be implemented to ensure equitable predictions.

7. How do you plan to test the model on new, unseen data to ensure it generalizes well?

To ensure the model generalizes well, we plan to:

* Use cross-validation during training to evaluate performance on multiple subsets of data and reduce overfitting.
* Test the model on data from different periods (e.g., seasonal data or data from other airports/airlines) to assess its ability to handle new, unseen data.
* Continuously retrain the model with new flight data and update it for real-time prediction accuracy.

8. Can this model be extended to predict delays in international flights?

Yes, this model can be extended to predict delays for international flights as well. However, additional factors such as customs processing times, international weather conditions, and longer flight durations need to be incorporated. The model may need retraining with data specific to international flights to ensure that it generalizes to these different contexts.

9. How does weather data impact the prediction of delays in your model?

Weather conditions are one of the most significant factors influencing flight delays. The model uses weather-related features like departure and arrival weather conditions to predict delays. We found that severe weather conditions (e.g., storms or snow) were strong predictors of delays. Further improvements could involve integrating real-time weather data to enhance the model's performance.

10. What steps would you take if the model starts to misclassify delays in the future?

If the model starts misclassifying delays, we would:

* Monitor the model’s performance over time and look for signs of drift (where the data or relationships between features and target change).
* Retrain the model with new, relevant data (such as data on new delays or updated features).
* Experiment with advanced resampling techniques or hybrid models (e.g., combining XGBoost and Random Forest for better predictive accuracy).
* Use error analysis to understand where the model is making mistakes and adjust accordingly, for example, by fine-tuning model hyperparameters or incorporating new features.