**Milestone 3**

**White Paper**

**Predicting Flight Delays Using Machine Learning**

Bobga-Herman Gwanvoma

Bellevue University

DSC 680 – Applied Data Science

Professor Amirfarrokh Iranitalab

April 6, 2025

**Project Overview**

This project focuses on developing a machine learning model capable of predicting flight delays using historical flight data, weather conditions, and airline operations. By accurately forecasting delays, both passengers and airlines can minimize disruptions, optimize scheduling, and improve the overall customer experience. The model analyzes various contributing factors to delays, aiming to provide reliable forecasts to enhance the efficiency of air travel operations.  
  
**Business Problem**

Flight delays are a significant challenge for the aviation industry, leading to financial losses for airlines, inconvenience for passengers, and inefficiencies in airport operations. Predicting delays before they occur allows airlines to adjust their schedules, inform passengers in advance, and make more informed decisions. By analyzing various factors contributing to delays, this project aims to create a predictive model capable of providing reliable forecasts of delays, helping mitigate their impact on all stakeholders involved.  
  
**Research Questions**

This study seeks to answer the following research questions:

1. What are the key factors influencing flight delays?
2. How accurately can machine learning models predict flight delays using historical and real-time data?
3. Which machine learning algorithms perform best in predicting flight delays?

**Data Explanation**

* The datasets used in this project are open-source and sourced from:
* The U.S. Department of Transportation (BTS) flight delay data.
* Kaggle flight delay datasets.
* Real-time weather and airline operational data.

These datasets contain detailed information on flight schedules, airline performance, departure and arrival times, weather conditions, and various causes of delays, such as weather, air traffic, and carrier operations. The data is in CSV format and requires preprocessing, including handling missing values, feature selection, and normalization before it can be used for machine learning model training.

**Methods**

The data underwent several data preprocessing steps:

Handling Missing Data: Missing values were imputed with the mean for numerical columns to prevent model bias, Feature Engineering: We created a new feature high\_severity\_delay, which categorizes flights with delays greater than 30 minutes, and Normalization: Numerical features such as carrier\_delay, weather\_delay, and arr\_delay were normalized to ensure they were on the same scale for machine learning models.

**Machine Learning Models:**

Random Forest: This ensemble learning method was used to improve classification accuracy by constructing multiple decision trees. It was trained on the features relevant for predicting flight delays, such as carrier\_delay, weather\_delay, nas\_delay, and arr\_delay.

XGBoost: A gradient boosting model that captures complex patterns in the data. It was also trained on the same features to predict the likelihood of flight delays.

Both models were evaluated using metrics such as accuracy, confusion matrix, and ROC curve to assess their performance.

**Evaluation Metrics:**

Accuracy: To measure the overall correctness of the models, Confusion Matrix: To evaluate the true positives, true negatives, false positives, and false negatives, and ROC Curve: To assess the performance in distinguishing between delayed and non-delayed flights.

**Analysis**

The models were trained and evaluated using a test set, and both Random Forest and XGBoost achieved exceptional performance, with perfect accuracy (100%) on the test data. However, it’s important to note that the class imbalance (with most instances of low-severity delays) may have influenced the results.

Model Performance: Both models demonstrated perfect classification performance in terms of sensitivity, specificity, and balanced accuracy, indicating their ability to predict flight delays correctly in this specific dataset.

Challenges: The class imbalance in the dataset, where most delays are minor, made the models more likely to predict the majority class correctly, potentially leading to overfitting. The weather data plays a critical role, but missing weather-related features were handled using imputation techniques.

**Visualizations  
Figure 1: Random Forest ROC Curve**The Random Forest ROC curve illustrates its perfect classification ability, with a high Area Under the Curve (AUC) of 1, confirming that it can distinguish between delayed and non-delayed flights effectively.  
A screenshot of a computer code

AI-generated content may be incorrect.  
A graph of a forest model

AI-generated content may be incorrect.  
**Figure 2: XGBoost ROC Curve**

Similarly, the XGBoost ROC curve also shows perfect classification performance with an AUC of 1.   
A screen shot of a computer code

AI-generated content may be incorrect.

A graph of a curve

AI-generated content may be incorrect.

**Figure 3: Confusion Matrix for XGBoost Model**The confusion matrix for the XGBoost model shows perfect predictions for both delayed and non-delayed classes.  
A black and white text

AI-generated content may be incorrect.  
A screenshot of a computer

AI-generated content may be incorrect.

**Conclusion**

Both Random Forest and XGBoost models achieved 100% accuracy on the test data. The ROC curves and confusion matrices confirmed the effectiveness of these models in predicting flight delays. However, due to the imbalanced dataset, additional steps, such as resampling or synthetic data generation (SMOTE), are recommended to improve generalization.  
**Ethical Considerations**

* Data Privacy: All data used is anonymized, ensuring no personally identifiable information is included.
* Bias & Fairness: Measures were taken to ensure that the model doesn’t introduce bias against certain airlines, airports, or routes.
* Impact of False Predictions: Misclassifying delays can have operational consequences. It’s important to validate the model rigorously before deployment.

**Recommendations**

To enhance the model's accuracy and mitigate overfitting, I recommend the following:

* Cross-Validation: Implement k-fold cross-validation to prevent overfitting and ensure the model’s robustness.
* SMOTE: Use Synthetic Minority Over-sampling Technique to balance the dataset and improve model performance for both classes.
* Hyperparameter Tuning: Further optimize the XGBoost and Random Forest models through hyperparameter tuning.

**Future Applications**

This model can be deployed in real-time flight operations to predict delays and notify passengers in advance, thus improving customer satisfaction. Airlines could use this system to optimize flight schedules, reduce cancellations, and minimize operational disruptions.

**Appendix**

* Data Preprocessing Steps: Details on data cleaning and feature engineering.
* Model Training Code: Code snippets for training Random Forest and XGBoost models.
* Hyperparameter Settings: Settings used for model training and optimization.

**References**

U.S. Department of Transportation flight delay data. (n.d.). Retrieved from https://www.transtats.bts.gov/

Kaggle flight delay datasets. (n.d.). Retrieved from <https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023>