PROJECT ABSTRACT

Flight Delay Prediction using Machine Learning

A dual gradient-boosted model predicting flight delays and delay minutes on US DOT BTS data, with a predictor demo and departure board. Delivered built-to-order with report, PPT and viva kit.

---

01. Title

Flight Delay Prediction using Machine Learning

02. Introduction / Background

Flight delays cost airlines and passengers billions yearly and follow learnable patterns — monsoon months, evening banks, specific airports and carriers. The US DOT Bureau of Transportation Statistics publishes on-time performance data at massive scale, a rare student project with genuine big-data character. This project treats the problem as the dual task it is — classify delay vs on-time and regress delay minutes — then decomposes predictions into the official BTS cause categories, with an interactive predictor and departure-board demo.

03. Problem Statement

Student builds usually shuffle the data and fit one classifier, leaking the future into training and ignoring the regression half of the problem. The project addresses this with temporal splits (train on earlier months, test on later), dual modelling with proper metrics (AUC, MAE), calibration analysis, and a cause-breakdown layer using the BTS taxonomy.

04. Objectives

- Engineer schedule, route, carrier and cyclical time features from BTS records.
- Train a gradient-boosted classifier (delay ≥ 15 min) and regressor (delay minutes).
- Evaluate with temporal splits, AUC/MAE and calibration curves.
- Attribute predicted delay across the five BTS cause categories.
- Build predictor + departure-board web demos.
- Deliver the complete student kit: notebook, models, demo, report, PPT and viva Q&A.;

05. Proposed Methodology

The pipeline has three stages. (1) Data: 2.1M-record BTS sample; cancelled/diverted flights handled per documented rules; cyclical encoding for departure time. (2) Modelling: temporal train/test split; XGBoost/LightGBM classifier and regressor tuned with cross-validation. (3) Evaluation and demo: AUC, confusion matrix, MAE and calibration on the later-months test set; models serialized and wired to the demo with the cause-attribution layer.

06. System Architecture / Working

The notebook exports both trained models and feature code. The single-file demo's predictor view takes airline, route, month, day and departure time and returns delay probability, expected minutes and the BTS cause breakdown. Departure-board mode scores upcoming flights ranked by risk. Analytics views show the monthly seasonality curve, national cause shares and model metrics.

07. Hardware / Software Requirements

- Python 3, scikit-learn, XGBoost/LightGBM
- pandas, NumPy, Matplotlib
- Jupyter Notebook
- US DOT BTS on-time performance data
- HTML5, CSS, JavaScript (demo)
- 8 GB+ RAM recommended for the 2.1M sample

08. Expected Outcomes

- Dual models with design-target AUC 0.81 and MAE 9.7 minutes.
- An interactive predictor and departure-board demo.
- Seasonality and calibration analyses for the report.
- A complete report, PPT and viva Q&A; on applied ML with big data.

09. Applications

- Teaching regression + classification dual modelling.
- Demonstrating temporal validation and calibration.
- Academic study of seasonality in transport data.

10. Future Scope

- Real-time weather-feed features.
- Delay propagation across connecting flights.
- Indian domestic data retraining.
- Airline-operations dashboard extension.