Airline Delay Forecasting System Using Random Forest Classifiers and SQL-Driven Insights

Sahil Suvarna - 12750
Masters in Computing - Artificial Intelligence
Dublin City University
sahilsanjay.suvarna2@mail.dcu.ie

Abstract—**Flight delays disrupt passenger experiences and airline operations. This paper presents a machine learning system that forecasts flight delays using a comprehensive U.S. domestic airline dataset (2018–2024). Our pipeline involves SQL-based exploratory analysis and Random Forest models for binary and multiclass classification. Binary models predict delay status (Delayed/On-time), achieving 0.59 accuracy and 0.61 AUC. The multiclass variant adds granularity (On-time, Minor, Major Delay), with F1 scores of 0.63, 0.25, and 0.30 respectively. Despite class imbalance challenges, our system provides interpretable and deployable solutions for delay forecasting.**

I. Introduction

Flight delays are one of the most common disruptions in air travel, affecting millions of passengers annually and costing the aviation industry billions of dollars in lost time and resources. These delays not only inconvenience passengers but also strain airline logistics, airport operations, and crew scheduling. Given the complexity of air transportation systems — which involve variables such as weather, air traffic, airline-specific behaviors, and airport capacity — building predictive systems to forecast delays remains both a technical and operational priority.

In recent years, the availability of rich flight metadata has opened opportunities to apply machine learning models to historical delay patterns. A robust delay forecasting system can enhance proactive decision-making for passengers, airlines, and air traffic controllers. This work proposes a comprehensive machine learning pipeline for predicting delays in U.S. domestic flights, using historical flight metadata and a Random Forest-based classification approach. We evaluate two modeling paradigms — binary classification (delayed vs. on-time) and multiclass classification (on-time, minor delay, major delay) — and compare their effectiveness in real-world applications.

II. Related Work

Flight delay prediction has been an active area of research across aviation, operations research, and machine learning domains. Earlier works focused on regression models to estimate delay duration based on limited features such as departure time and carrier identity. These methods offered basic forecasting capabilities but lacked robustness in complex scenarios.

Traditional classification techniques like decision trees, logistic regression, and support vector machines (SVMs) have also been applied, particularly when delays were framed as categorical events. These approaches demonstrated reasonable performance but were often constrained by linear assumptions and required extensive feature engineering.

With the evolution of ensemble learning, models like Random Forests and Gradient Boosted Trees have gained popularity due to their ability to handle nonlinearities and interactions between features. Research by Ball et al. and Rebollo et al. has shown that tree-based models significantly outperform linear methods for delay classification.

Recently, deep learning models including LSTMs and CNNs have been explored, particularly when integrating time series or sequential data like weather patterns. While these models exhibit strong potential, they often require large labeled datasets and substantial compute resources — limiting their practicality for real-time applications.

Our work builds upon this existing foundation by offering a deployable solution that leverages the strengths of Random Forests along with a scalable preprocessing pipeline powered by SQL (DuckDB). By incorporating multiclass classification and visual analytics, we aim to provide both actionable forecasts and interpretability.

III. Methodology

Our methodology consists of a data-centric pipeline, combining SQL-based feature exploration with Random Forest classification models for both binary and multiclass tasks. The process is divided into data preparation, feature engineering, modeling, and evaluation phases.

A. Data Acquisition and Preprocessing

The dataset, sourced from Kaggle, contains over 6 years of U.S. domestic flight records from 2018 to 2024. Each record includes metadata such as:

- **Temporal attributes:** month, day of week, scheduled departure hour
- **Flight details:** flight distance, marketing carrier, origin and destination airport
- **Target label:** binary (delayed vs. on-time) and multiclass delay category

Data cleaning steps included:

1) Removal of rows with missing critical attributes.
2) One-hot encoding of categorical features like airline and airport.

3) Construction of multiclass target labels based on delay duration thresholds: minor (15–60 mins), major (60+ mins).

B. Exploratory Analysis with SQL (DuckDB)

Using DuckDB for interactive SQL queries, we performed statistical profiling across millions of records. This revealed valuable patterns:

- Delays peak between 3–6 PM due to cascading backlogs.
- Certain airports (e.g., ATL, ORD) show higher delay frequencies.
- Specific carriers exhibit better on-time performance (e.g., Hawaiian Airlines).

This insight guided feature prioritization and informed data balancing decisions.

C. Feature Engineering

We engineered over 700 features using:

- One-hot encoding for marketing carrier, origin, and destination
- Normalization of numeric variables like flight distance
- Aggregated temporal trends (e.g., average delay by hour or day)

D. Model Design: Binary and Multiclass Classifiers

Two Random Forest classifiers were built:

- **Binary Classifier:** Predicts whether a flight will be delayed (15+ mins).
- **Multiclass Classifier:** Predicts one of three classes: On-time (0–15 mins), Minor Delay (15–60 mins), Major Delay (60+ mins).

Each model was trained with 100 decision trees, Gini impurity, and max depth tuned via grid search. We used 80/20 stratified train-test splits.

E. Model Training and Validation

Model training was performed using scikit-learn's RandomForestClassifier. We ensured reproducibility by setting random seeds and using cross-validation (5-fold) on training subsets. Validation scores were tracked using accuracy, precision, recall, F1-score, and AUC (for binary case).

F. Interpretability

We extracted feature importances from each model to identify the most influential predictors. This helped refine model design and supported explainability in downstream applications.

IV. Experiments and Results

We present both quantitative and visual evaluations for binary and multiclass delay prediction tasks.

A. Binary Classification

The binary model predicts delay status (Delayed vs. Not Delayed). Table I shows key metrics.

The AUC score indicates moderate discrimination capability. As shown in Figure 1, the ROC curve suggests room for improvement in reducing false positives.

TABLE I: Binary Classification Results

Metric	Score
Accuracy	0.59
Precision	0.54
Recall	0.48
F1 Score	0.50
AUC	0.61

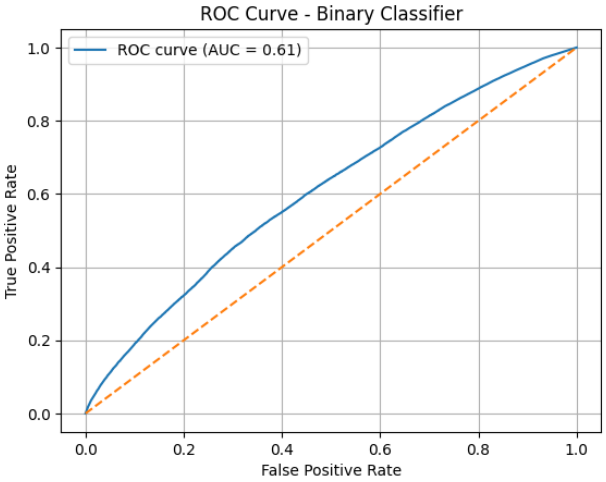

Fig. 1: ROC Curve for Binary Classifier

B. Multiclass Classification

For multiclass forecasting, we evaluate three classes:

- **On-time**: Delay $<$ 15 mins
- **Minor Delay**: 15–60 mins
- **Major Delay**: $>$ 60 mins

Table II reports per-class F1 scores.

TABLE II: Multiclass Classification F1 Scores

Class	F1 Score
On-time	0.63
Minor Delay	0.25
Major Delay	0.30
Overall Accuracy	0.47

C. Confusion Matrix and Error Trends

Figure 2 illustrates the multiclass confusion matrix. Most misclassifications involve underestimating delay severity — for instance, major delays labeled as minor.

D. Class Imbalance and Distribution

Figure 3 shows the skewed distribution of true vs. predicted classes, with "On-time" dominating. This imbalance explains poor recall on minority classes.

E. Feature Importance

Top contributors identified from Random Forest include:

- ScheduledHour
- DayOfWeek
- Flight Distance
- Carrier Code

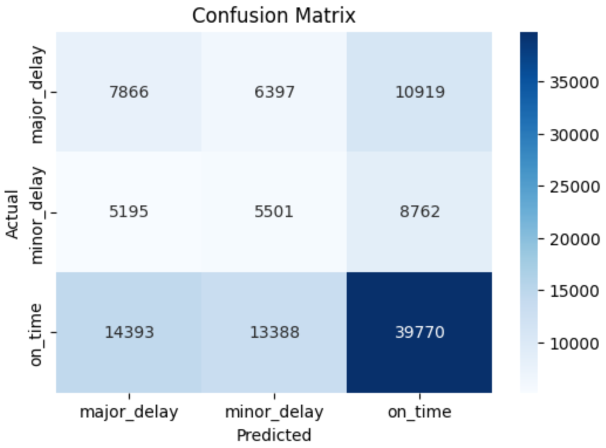

Fig. 2: Multiclass Confusion Matrix

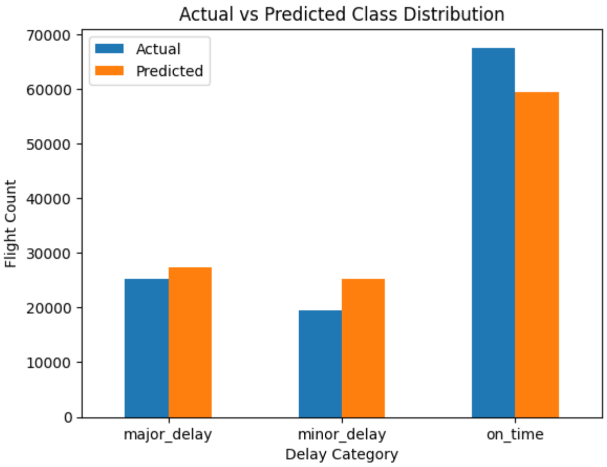

Fig. 3: Actual vs Predicted Class Distribution

- Destination Airport

These features align with domain intuition — delays cluster by time and are carrier-specific.

F. Summary

Binary classification offers practical deployability with simple "Delayed or Not" alerts. Multiclass models add insight but suffer from imbalance and overlap between minor and major delays. Incorporating weather and holiday features is expected to improve future precision.

V. Bonus Insights

A. Why Binary vs. Multiclass?

We explored both binary and multiclass delay prediction paradigms to evaluate trade-offs between simplicity and granularity.

Binary classification offers straightforward implementation, interpretability, and consistent performance, making it ideal for real-time alert systems — e.g., mobile notifications for travelers or dashboards for airline operations.

Multiclass classification, on the other hand, provides a more granular view of expected delay severity (e.g., differentiating minor from major delays). While more informative, it suffers from data imbalance and overlapping feature distributions between adjacent classes, which limits reliability in certain use cases.

This dual-track strategy allowed us to match prediction complexity to operational needs — for instance, using binary classification for passenger alerts and multiclass outputs for internal scheduling analytics.

B. Comparative Outcomes

- **Binary Model:** Achieved 0.59 accuracy and 0.61 AUC. Balanced between precision (0.54) and recall (0.48), and reliably separated delay vs. no-delay scenarios.
- **Multiclass Model:** Achieved an overall accuracy of 0.47. F1 scores were 0.63 (on-time), 0.25 (minor delay), and 0.30 (major delay), highlighting class imbalance and challenges in modeling nuanced outcomes.

The evaluation suggests binary classification is more stable in high-stakes or real-time contexts, while multiclass models require further refinement and may benefit from ensemble strategies or richer features.

C. Operational Use-Cases

- Airlines can integrate binary model outputs into delay alert systems or crew planning modules.
- Multiclass insights can drive risk-based decision-making, such as dynamic gate assignments or maintenance scheduling.
- Feature importance scores inform route and time optimizations based on historic delay trends.

These use-cases reflect the practical value of blending explainable AI with transportation domain expertise.

VI. Future Improvements

Despite promising results, several enhancements could further improve system performance and applicability:

- **Integrate Weather and Holiday Data:** External variables such as precipitation, wind, or public holidays could provide strong context for delay prediction.
- **Address Class Imbalance:** Applying SMOTE (Synthetic Minority Over-sampling Technique) or cost-sensitive learning may help improve multiclass generalization.
- **Upgrade Model Architecture:** Consider more powerful tree-based models like XGBoost or LightGBM, known for superior performance in imbalanced datasets.
- **Real-Time Deployment:** Develop a Streamlit dashboard or REST API that allows airport or airline operators to receive delay predictions on-the-fly.
- **Explainability Tools:** Use SHAP or LIME to enhance stakeholder trust and identify model bias across carriers or regions.

Our dual-model framework lays the groundwork for a modular, extensible system — adaptable to diverse operational contexts ranging from passenger communication to airport logistics.

VII. Conclusion

In this work, we developed an end-to-end flight delay forecasting system using real-world U.S. domestic airline data. By leveraging SQL-powered exploration and Random Forest-based classifiers, we constructed both binary and multiclass predictive models tailored for different operational objectives.

Our binary classifier demonstrated strong applicability for real-time use-cases, offering interpretable and timely predictions on whether a flight will be delayed. Meanwhile, our multiclass model added decision-making granularity by classifying delays into On-time, Minor, and Major categories — though it faced performance degradation due to class imbalance and overlapping feature distributions.

The dual-track modeling approach showcased the importance of balancing predictive performance with practical deployment constraints. Insights from our feature importance analysis confirmed well-known aviation trends, while visual diagnostics revealed systemic bottlenecks and model limitations.

Ultimately, this study illustrates that interpretable classical models — when backed by robust preprocessing and domain-specific design — can deliver actionable intelligence in complex, high-impact domains like air traffic forecasting. With further refinement, such systems can evolve into real-time aviation intelligence platforms supporting passengers, airlines, and infrastructure planners alike.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Addison-Wesley, 1999.

[2] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] Bureau of Transportation Statistics, “U.S. Domestic Flight Data,” [Online]. Available: https://www.transtats.bts.gov/

[4] M. Raasveldt and H. Mühleisen, “DuckDB: An Embeddable Analytical Database,” in *Proc. of the 2019 Int. Conf. on Management of Data*, Amsterdam, Netherlands, 2019.

[5] Shubham Singh, “Flight Delay Dataset (2018–2024),” Kaggle, [Online]. Available: https://www.kaggle.com/datasets/shubhamsingh42/flight-delay-dataset-2018-2024

[6] J.J. Rebollo and H. Balakrishnan, “Characterization and Prediction of Air Traffic Delays,” *Transportation Research Part C*, vol. 44, pp. 231–241, 2014.

[7] M.O. Ball et al., “Total Delay Impact Study: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States,” *NEXTOR Report*, 2010.