Predicting Flight Delays Using Machine Learning: Classification and Regression Models on BTS Aviation Data

Ayah M. Batayneh
Department of Computer Information Systems
Data Science
Jordan University of Science and Technology
Irbid, Jordan
ambatayneh23@cit.just.edu.jo

Lama F. Emran
Department of Computer Information Systems
Data Science
Jordan University of Science and Technology
Irbid, Jordan
lfemran23@cit.just.edu.jo

Majd A. Aleyadeh
Department of Computer Information Systems
Data Science
Jordan University of Science and Technology
Irbid, Jordan
maaleyadeh23@cit.just.edu.jo

***Abstract*— Flight delays remain a persistent challenge in air transportation systems, affecting operational efficiency, costs, and passenger experience. This study examines the use of machine learning techniques to predict flight delay occurrence and arrival delay duration using one year of BTS aviation data. A representative subset of approximately one million flights was extracted from the original dataset. A comprehensive preprocessing pipeline was applied, including feature engineering, treatment of skewed distributions, and outlier handling. Two modeling tasks were explored independently. For classification, delayed flights were defined as those with arrival delays exceeding 15 minutes, while regression models were used to estimate delay duration. Multiple machine learning models were evaluated using appropriate performance metrics for each task. Results indicate that tree-based ensemble models outperform simpler baselines, demonstrating their effectiveness in capturing complex operational patterns. These findings highlight the potential of data-driven approaches to support proactive delay management in airline operations.**

Keywords*— *Flight Delay Prediction, Machine Learning, Supervised Learning, Classification, Regression, Arrival Delay

1. INTRODUCTION

Flight delays are a serious issue that airlines deal with, and they affect both economic performance and passenger experience. Small schedule issues often grow into more widespread delays, increasing fuel consumption, congestion, labor costs, and passenger dissatisfaction. Since air traffic growth continues to expand while the capacity of airports remains constrained, the effects of these delays become more severe, pointing to the **importance of better delay prediction and management**. The main research problem is the challenge of predicting delay behavior in a system influenced by many factors. Accurate delay prediction is key to improving scheduling, reducing unnecessary costs and increasing the overall reliability of air travel. In this study, predictions focus on the **post-departure** phase, where the objective is to estimate arrival delay using operational information available after the flight has departed.

In this project we use a large and completely raw dataset from the *Bureau of Transportation Statistics (BTS) TranStats system* [23] , which makes our work closely connected to real-world conditions. The full dataset has **7+ million U.S. flight records** from 2024 each described by **35 features**, from which we used a representative sample for analysis. These features include airport codes, delay minutes, scheduled and actual times, cancellations and distances, providing detailed information about each flight. Because the data comes directly from BTS, it is messy and contains natural inconsistencies, missing values and a wide range of numeric, categorical and time-based features. This shows the true complexity airlines face when dealing with daily operations. The dataset also grows continuously as new flights are reported, making it more reflective of the real and constantly changing conditions in airline operations.

Despite extensive research on flight delay prediction with machine learning, several **limitations** remain, as noted in the related work section. Many earlier studies use binary classification to predict if a flight is delayed or not, this approach **oversimplifies** the problem and cannot deliver the minute-level forecasts required for operational decisions such as crew planning or gate scheduling. Many studies also relied on **pre-cleaned or simplified datasets**, often from Kaggle or limited to individual airports, which overlook the full complexity and scale of authentic aviation operations. As a result, models produced in such studies frequently have **poor generalization** particularly when exposed to real-world noise, missing values, or the intricate relationships among different delay factors. Although some recent papers incorporate **deep learning** and **advanced feature engineering**, only a limited number integrate these techniques into a complete big-data pipeline capable of **processing raw, large-scale datasets**. This creates a noticeable gap between what research models can do and what airlines need. The **challenge** then is to turn the huge and messy BTS data into a working prediction system that can understand delay patterns and give accurate minute-level predictions.

The goal of this study is to create a regression-based machine learning framework capable of predicting flight delay duration in minutes using the arr_delay variable as the target, under a post-departure prediction setting. This approach aims to provide detailed, minute-level predictions that exceed the limitations of binary classification. In addition to predictive modeling, the study performs extensive exploratory data analysis to reveal operational patterns, pinpoint key contributors to delay and analyze temporal and geographic trends.

2. RELATED WORK

Flight delay prediction has attracted significant research attention due to its economic and operational impact. Literature has evolved from descriptive statistical analysis to complex machine learning (ML) and deep learning (DL) architecture. Existing studies can be categorized into five distinct approaches: foundational classification, advanced ensemble techniques, regression-based magnitude prediction, deep learning integrations, and holistic optimization systems.

A. Foundational Machine Learning and Classification:

Early research focused primarily on binary classification (delayed versus on-time). Nibareke and Laassiri [1] used Big Data analytics with Spark SQL to identify peak delay times and high-traffic routes, establishing a baseline for understanding delay patterns.

Building on this, several studies applied standard supervised algorithms. Kumar et al. [2] and other researchers compared Random Forest (RF), Support Vector Machines (SVMs), and K-Nearest Neighbors (KNN), consistently finding that SVM and RF outperform simpler models due to their ability to handle high-dimensional aviation data.

Comparative studies highlighted that while Linear Regression often fails to capture non-linear dependencies, SVM can achieve accuracy as high as 97% in binary classification tasks [3].

Furthermore, web-based implementations have demonstrated the practical deployment of these foundational models (RF, KNN, and Naïve Bayes) for real- time user interfaces [4].

B. Advanced Ensemble Methods and Data Handling:

To address the limitations of basic models, particularly 'class imbalance' (where on-time flights heavily outnumber delayed ones), recent studies have shifted toward ensemble methods and advanced sampling techniques. Kiliç and Sallan [5] demonstrated that Gradient Boosting Machines (GBMs) outperform standard models when applied to US Airport Networks, specifically when coupled with undersampling techniques.

Similarly, other studies utilizing XGBoost, CatBoost, and LightGBM [6] have shown that oversampling techniques like SMOTE are superior to undersampling, achieving up to 95% accuracy [7]. These studies emphasize that handling data imbalance is as critical as algorithm selection.

For instance, applying weighted evaluation metrics to Decision Trees has proven effective in neutralizing the dominant effect of on-time/non-delayed flights in imbalanced datasets [8].

Feature engineering has also proved critical; for instance, incorporating "Network Centrality" attributes into Random Forest models significantly improved robustness [9].

C. Regression and Minute-Based Prediction:

Moving beyond binary classification, a subset of the literature addresses the more challenging task of predicting exact delay magnitude (Regression).

Dhadake et al. [10] utilized Random Forest Regression to output precise delays in HH:MM format, emphasizing weather and schedule vectors. Comparative studies on regression techniques found that RIDGE regression outperforms LASSO for extended delay time series forecasting [11].

More recently, Janarthanan and Balasubramanian [12] identified Gradient Boosting Regressors (GBRs) as the most effective method for minimizing Root Mean Squared Error (RMSE) in the National Airspace System, outperforming linear baselines. Innovative feature engineering has also emerged; for example, The "Flight Delay Path Previous" (FDPP- ML) algorithm introduced path-based features -tracking previous flight delays-, reducing Mean Absolute Error (MAE) by approximately 39% compared to traditional models [13].

Aggregate level predictions using LightGBM have also shown promise in predicting airport-wide departure delays [14].

D. Deep Learning and Hybrid Architectures:

The current leading-edge involves Deep Learning and Hybrid architectures designed to capture temporal dependencies and complex network effects.

Gole et al. [15] proposed a Deep Learning model optimized with the Levenberg-Marquardt algorithm, achieving significantly higher precision than standard Recurrent Neural Networks (RNNs).

To address the overfitting issues common in pure deep learning models [16], hybrid approaches have been developed. Chavan et al. [17] combined LSTM (for temporal data) with XGBoost (for feature selection), achieving 98.40% accuracy.

Similarly, Dai [18] proposed a clustering- based hybrid to improve accuracy by 5.3% over conventional methods. Other novel hybrids include Deep Feed Forward Regression Networks (DFFRNs) [19] with spatio-temporal features and fusing Random Forest with Maximal Information Coefficients (RFR-MIC) [20] to better integrate multi-route flight information.

Comprehensive reviews such as those by Adamalapelli et al. [21] confirm that these hybrid and stacking-based models consistently outperform standalone techniques.

E. Integrated Optimization Systems:

The highest level of complexity connects prediction with operational mitigation.

A notable study proposed a "Ground Handling Process Optimization" [22] model that links Random Forest predictions with a Genetic Algorithm (NSGA- II). This system not only predicts delays but uses the results to dynamically optimize ground services, demonstrating a shift from passive prediction to active system recovery.

3. METHODOLOGY AND PREPROCESSING

This study makes use of a large-scale aviation dataset gathered from the Bureau of Transportation Statistics (BTS) TranStats database. While the full dataset contains over 7 million U.S. flight records for the year 2024, a representative sample of 1,000,025 flights was selected for analysis to maintain computational efficiency while preserving the distribution of key features. The dataset includes 34 attributes, such as flight dates, carrier identities, airport codes, scheduled and actual times, and multiple delay types (carrier, weather, NAS, etc.), providing a comprehensive view of real-world airline operations. The full dataset and sampling code can be accessed on GitHub for verification and reproducibility.

The data illustrates the operational complexity and variability of the aviation system by recording both scheduled and actual performance measures. Using **ARR_DELAY** as the target variable, the main analytical goal is to create a regression-based machine learning system that forecasts the exact length of flight delays in minutes. The investigation of flight performance patterns and delay trends over various time periods, routes, and carriers is further supported by an interactive visualization dashboard.

Data Quality Assessment

Table 1. Missing Data Summary.

Feature(s)	Description
CANCELLATION_CODE	98.62% missing in CANCELLATION_CODE
LATE_AIRCRAFT_DELAY, NAS_DELAY, WEATHER_DELAY, CARRIER_DELAY, SECURITY_DELAY	79.64% are missing in LATE_AIRCRAFT_DELAY, NAS_DELAY, WEATHER_DELAY, CARRIER_DELAY, and SECURITY_DELAY
AIR_TIME, ACTUAL_ELAPSED_TIME	1.61% are missing in AIR_TIME and ACTUAL_ELAPSED_TIME
ARR_TIME, WHEELS_ON, and TAXI_IN	1.39% are missing in ARR_TIME, WHEELS_ON, and TAXI_IN
WHEELS_OFF and TAXI_OUT	1.36% are missing in WHEELS_OFF and TAXI_OUT
DEP_DELAY	1.32% missing in DEP_DELAY
DEP_TIME	1.32% missing in DEP_TIME
ARR_DELAY	1.62% missing in ARR_DELAY

Table 2. Outlier Analysis Summary.

Feature(s)	Description
DEP_DELAY	Extreme right tail, values exceed 1.5×IQR (~12.7%)
TAXI_OUT, TAXI_IN	High positive skew, extreme long durations (~5–7%)
CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, LATE_AIRCRAFT_DELAY	Delay categories with long right tails (IQR > 1.5×IQR) (1–2%)
DISTANCE	Flights with abnormally long or short distances (~5.6%)
DIVERTED, CANCELLED	Binary columns with rare "1" values considered as outliers statistically (0.2%) and (1.4%)

The dataset was examined for data quality issues, including duplicates and inconsistent numeric formats. No duplicated rows were found. However, several time-related features showed mixed numeric semantics. Six columns (CRS_DEP_TIME, DEP_TIME, WHEELS_OFF, WHEELS_ON, CRS_ARR_TIME, ARR_TIME) represent clock times in **hhmm** format, which differs from other time features measured as durations in minutes.

In contrast, eleven columns (such as DEP_DELAY, TAXI_OUT, TAXI_IN, AIR_TIME, and CARRIER_DELAY) correctly represent duration values in minutes; while numerically valid, these are semantically distinct from the hhmm time columns.

Additionally, one non-numeric identifier (OP_CARRIER_FL_NUM) was stored as a float64 type. Although this flight number is not a quantitative variable, it was encoded numerically for memory efficiency and modeling consistency.

Data Preprocessing

To prepare the raw flight data for machine learning, we implemented a robust preprocessing pipeline designed to handle the specific challenges of aviation data, such as heavy skewness, missing values in delay columns, and physical data errors. The pipeline follows these logical steps:

1- Physical Constraint Filtering (Domain Knowledge)

Before statistical treatment, we applied domain-specific filters to remove erroneous entries that violate physical constraints.

Examples: Removed flights with TAXI_OUT > 120 minutes, AIR_TIME > 1260 minutes, or DISTANCE > 3000 miles. These values typically indicate data entry errors or extreme outliers that do not represent normal domestic operations.

2- Handling Missing Values

Deletion: The column CANCELLATION_CODE was dropped as it contained ~98.6% missing values, providing little utility for delay prediction.

Logic-Based Imputation: For specific delay breakdown columns (e.g., CARRIER_DELAY, WEATHER_DELAY), missing values were imputed with 0. In this domain, a missing value signifies that no delay occurred for that specific reason.

Median Imputation: For continuous operational metrics (e.g., AIR_TIME, TAXI_IN), missing values were imputed using the median. This strategy is robust against the right-skewed nature of flight time data compared to the mean.

3- Feature Transformation & Outlier Treatment

Clipping: Extreme outliers in delay columns (e.g., DEP_DELAY, ARR_DELAY) were clipped to a range of [-30, 1440] minutes (24 hours). This preserves the information of a "long delay" without allowing extreme anomalies (e.g., 2000+ minutes) to dominate the model's loss function.

Log Transformation: To address the heavy right-skew in delay variables, we applied a signed log transformation: $x' = sign(x) * log(1 + |x|)$. This compresses the scale of delays, making relationships more linear.

Power Transformation: For other skewed continuous features, we applied the Yeo-Johnson Power Transformer. This method automatically finds the optimal parameter to make the feature distribution as Gaussian (normal) as possible, which is essential for models sensitive to data distribution.

4- Feature Engineering

Time Conversion: Columns representing time in HHMM format (e.g., 1330 for 1:30 PM) were converted into continuous "minutes from midnight" to preserve the cyclical and ordinal nature of time.

Redundant Feature Removal: City and State columns were removed in favor of Airport codes (ORIGIN, DEST) to reduce dimensionality while retaining location precision.

Methodology

Classification

For the classification task, the objective was to predict flight delay categories after departure. The methodology followed a

structured approach covering target definition, model selection, feature evaluation, and model comparison:

1. Target Definition

The arrival delay variable (ARR_DELAY) was transformed into a categorical target by binning continuous values into two classes: **delayed** (arrival delay > 15 minutes) and **not delayed** (arrival delay ≤ 15 minutes). This framing allowed the problem to be addressed as a binary classification task suitable for standard machine learning techniques.

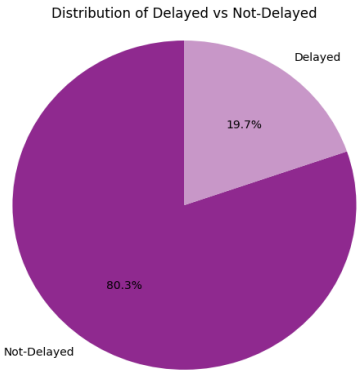

Fig.1. Distribution of Delayed vs Not Delayed Class

Fig. 1. illustrates the distribution of the binary arrival delay classes. Approximately 80.3% of flights are classified as not delayed, while only 19.7% fall into the delayed category. This indicates a clear class **imbalance**, with on-time flights significantly outnumbering delayed ones.

Due to this imbalance, accuracy alone is not a sufficient metric for evaluating classification performance, as a model could achieve high accuracy by favoring the majority class. Therefore, the **F1-score** which balances precision and recall, was used as the primary metric for model comparison and selection in the classification task.

2. Model Selection

A variety of machine learning classifiers were implemented to establish performance baselines and explore different algorithmic approaches:

- **Linear Models:** Logistic Regression as a baseline for interpretable linear classification.
- **Probabilistic Models:** Gaussian Naive Bayes to provide a probabilistic reference point.
- **Ensemble Models:** Random Forest, Gradient Boosting, AdaBoost, and LightGBM to capture complex relationships and improve predictive performance.
- **Deep Learning:** A multilayer perceptron (MLP) neural network for non-linear classification.

3. Feature Evaluation and Selection

Feature selection techniques were applied to identify the most informative predictors. Both mutual information and model-based feature importance from tree-based models (Random Forest) were used. This enabled evaluation of models trained on all features versus a smaller subset of top-performing features, providing insight into the trade-off between model complexity and predictive ability.

4. Model Comparison

Classifiers and feature sets were compared using a structured evaluation strategy:

- **Cross-Validation:** Stratified K-Fold cross-validation was employed to preserve the original class distribution within each fold, ensuring stable and reliable performance evaluation despite the class imbalance.
- **Performance Metrics:** Models were assessed using accuracy, F1-score, ROC-AUC, recall, and specificity to provide a comprehensive view of classification performance.
- **Feature Set Comparison:** Models trained on the complete feature set were compared against those trained on selected subsets to evaluate whether simpler models could maintain robust predictive performance.

Regression

The regression task aimed to predict the exact arrival delay in minutes after flight departure. The methodology followed a structured approach covering target definition, model selection, feature evaluation, and model comparison:

1. Target Definition

The continuous arrival delay variable (ARR_DELAY) was used as the target. Unlike the classification task, which grouped delays into categories, regression models were designed to estimate the precise number of minutes a flight was delayed, providing more detailed operational insights.

2. Model Selection

A diverse set of regression algorithms was implemented to explore different modeling approaches:

- **Linear Models:** Linear Regression, Ridge, Lasso, and ElasticNet for interpretable linear relationships.
- **Neighbor-Based Models:** K-Nearest Neighbors Regressor for similarity-driven predictions.
- **Ensemble Models:** Random Forest Regressor, Gradient Boosting Regressor, AdaBoost Regressor, XGBoost Regressor, and LightGBM Regressor to improve predictive performance through aggregation and boosting.
- **Deep Learning:** A sequential neural network with dense, dropout, and batch normalization layers to capture highly non-linear patterns.

3. Feature Evaluation and Selection

To identify the most informative predictors and reduce model complexity, the following feature selection approaches were applied:

- **Feature Importance Ranking:** Tree-based models were used to assess the influence of each feature on arrival delay.
- **Filter Methods:** Correlation-based selection and mutual information were used to identify relevant features.

4. Model Comparison

Models were evaluated using standard statistical metrics and validation techniques:

- **Error Metrics:** Mean Absolute Error (*MAE*), Mean Squared Error (*MSE*), and Root Mean Squared Error (*RMSE*) were computed, with both mean and standard deviation reported across cross-validation folds.
- **Explained Variance:** R^2 Score was used to measure how much of the variance in arrival delays was captured by each model.
- **Cross-Validation:** K-Fold cross-validation ensured stability and generalizability of the evaluation metrics across different data subsets.
- **Feature Set Comparison:** Models trained on all features were compared against those trained on selected subsets to examine the trade-off between model simplicity and predictive performance.

5.RESULTS AND DISCUSSIONS

Overview

This study aimed to predict flight delays using post-departure operational data, with two complementary approaches: **classification**, to predict whether a flight would be delayed or not, and **regression**, to estimate the exact arrival delay in minutes.

Models ranged from simple linear baselines to advanced ensemble and boosting methods, all evaluated using K-Fold cross-validation. Key predictors included departure delay (DEP_DELAY), carrier information, and flight scheduling features.

Classification Results

Flights were categorized into delayed (arrival delay > 15 min) and not delayed (≤ 15 min).

Table 3. Classification Base-Model Performance Summary.

Model	Accuracy	**F1-Score**	ROC-AUC
LightGBM	93.42%	81.77%	94.20%
Random Forest	93.36%	81.63%	93.84%
Gradient Boosting	93.36%	81.60%	93.84%
Neural Network	93.39%	81.53%	93.92%
AdaBoost	93.29%	81.47%	93.11%
Logistic Regression	92.39%	80.66%	93.31%
Naive Bayes	88.57%	74.57%	92.42%

Top Features (classification):

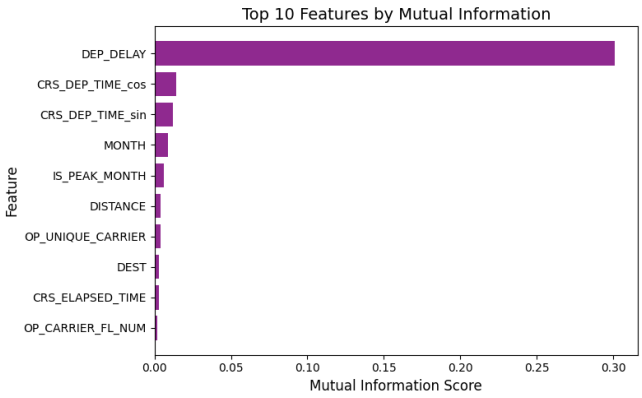

Fig.2. Top 10 Features Mutual Information (classification)

Mutual Information quantifies the dependency between each feature and the target variable. Higher MI scores indicate stronger relationships.

Fig.2. Mutual Information analysis shows that DEP_DELAY is the most informative feature by a large margin, indicating a strong dependency with arrival delay. Temporal features derived from scheduled departure time (sine and cosine transformations) also contribute meaningful information, while route- and calendar-based variables provide smaller but non-negligible signals.

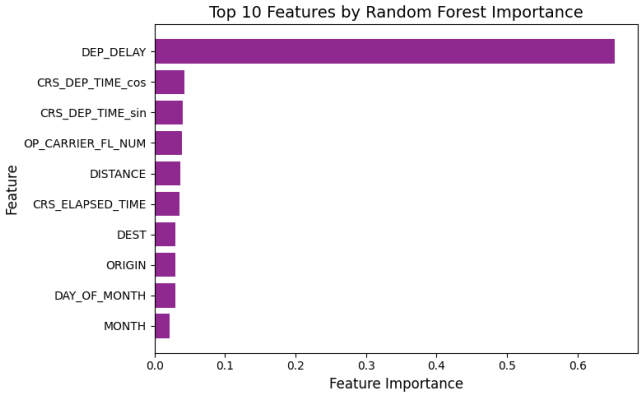

Fig.3. Top 10 Features Random Forest Importance (classification)

Random Forest considers feature interactions and non-linear relationships during training.

Fig.3. Random Forest feature importance further confirms the dominance of DEP_DELAY, while assigning higher relative importance to operational and route-related variables such as carrier flight number, distance, and elapsed time. Compared to Mutual Information, tree-based importance highlights features that contribute most to predictive performance within the model structure.

Impact of Feature Selection on Classification Performance

Feature selection using the top 10 features identified by Random Forest importance led to a slight reduction in classification performance across most models. Tree-based ensemble models, particularly LightGBM and Random Forest, achieved their best results when trained on the full feature set, indicating that these models benefit from additional features to capture complex patterns. Logistic Regression showed minimal

performance change after feature reduction, suggesting lower sensitivity to feature removal.

Overall, while feature selection improved model simplicity and interpretability, using all available features produced the highest accuracy and ROC-AUC for ensemble classifiers.

Classification Discussion

- **Best Model:** LightGBM performed best (ROC-AUC 94.20%), efficiently handling non-linear relationships and large datasets.
- **Feature Insights:** Flights that depart late are highly likely to arrive late; carrier and scheduled time provide additional predictive value.
- **Complexity vs Performance:** Baseline models (Logistic Regression) performed reasonably (~92% accuracy), but boosting models improved performance significantly.
- **Neural Network Comparison:** The MLP achieved similar accuracy but required longer training, making LightGBM more practical.

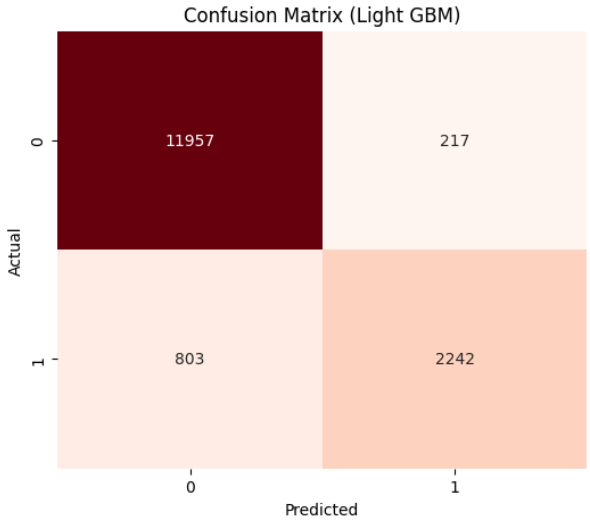

Fig. 4. Confusion Matrix (Light GBM)

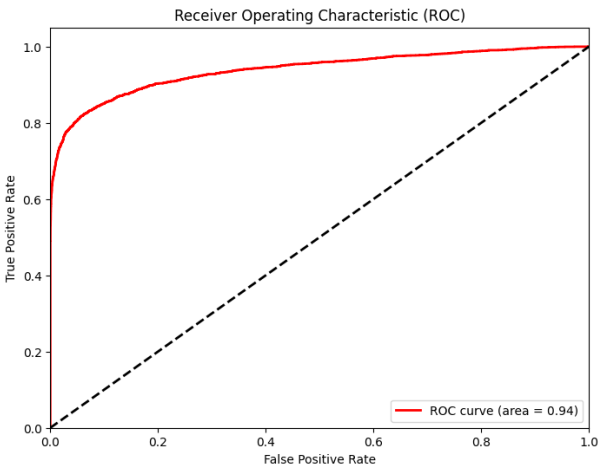

Fig. 5. ROC curve (Light GBM)

Fig. 4. highlights that LightGBM correctly identifies the majority of both delayed and on-time flights, with few misclassifications. Fig. 5. confirms the strong discriminatory ability of the model, justifying the use of F1-score as the main metric given class imbalance (delayed = 19.7%, not delayed = 80.3%). Stratified K-Fold cross-validation ensured that these results are stable and generalizable across the dataset.

Regression Results

The regression task predicted **ARR_DELAY** as a continuous variable.

Table 4. Regression Model Comparison (5-Fold CV ± std).

Model	R^2 Score	MAE (min)	RMSE (min)
Linear Regression	0.60 ± 0.01	0.50 ± 0.00	0.63 ± 0.00
Ridge	0.60 ± 0.01	0.50 ± 0.00	0.63 ± 0.00
Lasso	0.59 ± 0.01	0.50 ± 0.00	0.63 ± 0.00
ElasticNet	0.60 ± 0.01	0.50 ± 0.00	0.63 ± 0.00
K-Nearest Neighbors	0.66 ± 0.01	0.45 ± 0.01	0.58 ± 0.01
Random Forest	0.78 ± 0.00	0.33 ± 0.00	0.47 ± 0.00
Gradient Boosting	0.78 ± 0.00	0.33 ± 0.00	0.46 ± 0.00
AdaBoost	0.77 ± 0.01	0.36 ± 0.00	0.48 ± 0.00
XGBoost	0.75 ± 0.01	0.36 ± 0.00	0.49 ± 0.00
LightGBM	0.78 ± 0.00	0.33 ± 0.00	0.47 ± 0.00
Neural Network	0.77	0.33	0.46

Top Features (regression):

Embedded Methods (Top 5)

- **Random Forest:** DEP_DELAY, OP_CARRIER_FL_NUM, CRS_ELAPSED_TIME, ORIGIN, DISTANCE
- **LightGBM:** DEP_DELAY, CRS_ELAPSED_TIME, DISTANCE, ORIGIN, DEST
- **Gradient Boosting:** DEP_DELAY, CRS_ELAPSED_TIME, MONTH, OP_UNIQUE_CARRIER, ORIGIN

Filter Methods (Top 5)

- **Mutual Information:** DEP_DELAY, CRS_DEP_TIME_cos, CRS_DEP_TIME_sin, OP_UNIQUE_CARRIER, ORIGIN
- **Correlation:** DEP_DELAY, OP_UNIQUE_CARRIER, ORIGIN, DEST, IS_PEAK_MONTH

Impact of Feature Selection on Regression Performance

The impact of feature selection on regression performance was evaluated using multiple methods, including Random Forest importance, Mutual Information, and correlation-based filtering. Across all approaches, models trained on the full feature set consistently achieved slightly better performance, with lower MAE and RMSE and higher R^2 values. Feature selection generally led to modest increases in prediction error, indicating that removing lower-ranked features reduced the models' ability to capture subtle relationships in the data. Ensemble models such as LightGBM and Gradient Boosting remained relatively robust after feature reduction, though their best performance was still observed when all features were retained.

Overall, these results suggest that while feature selection can reduce model complexity, using the complete feature set yields more accurate arrival delay predictions.

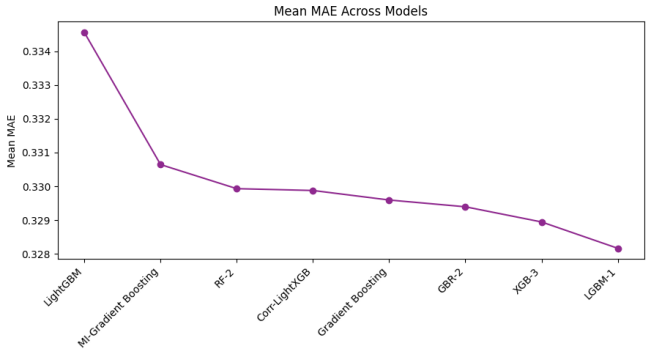

Fig.6. Mean MAE Across Models (regression)

Fig.6. shows the final regression leaderboard based on Mean Absolute Error (MAE). The top three models are LightGBM-1 (MAE ≈ 0.328), followed by XGBoost-3 (MAE ≈ 0.329), and Gradient Boosting Regressor-2 (MAE ≈ 0.329). These models achieved the lowest prediction errors, indicating the most accurate estimation of arrival delays. Models using feature selection methods, such as mutual information or correlation-based top features, performed slightly worse but remained competitive, suggesting that a carefully chosen subset of predictors can still provide high predictive performance.

Overall, ensemble boosting methods consistently outperform simpler baselines, demonstrating their strength in capturing complex patterns in the dataset.

Regression Discussion

- **Best Model:** LightGBM achieved the lowest MAE on the normalized dataset (~0.328) and the highest R^2 (~0.94), indicating strong predictive performance.
- **Feature Influence:** DEP_DELAY is the dominant predictor, with flight duration and route features supporting performance.
- **Challenges:** Extreme outliers remain difficult to predict, increasing RMSE slightly.
- **Neural Network Comparison:** Comparable performance but longer training makes GBMs more practical.

5. CONCLUSION

This study investigated post-departure flight delay prediction using both classification and regression approaches. By leveraging a large, raw dataset from the BTS TranStats system, we explored a wide range of machine learning models and feature selection strategies to identify the key drivers of arrival delays and assess predictive performance.

Key Findings:

- **Best Models:** LightGBM emerged as the top performer for both tasks. For *classification*, it achieved an F1-score of 81.8% (ROC-AUC 94.2%), effectively distinguishing delayed from on-time flights despite class imbalance. For *regression*, it produced the lowest mean absolute error (MAE) and the highest R^2 (~0.94), demonstrating accurate predictions of arrival delays and the model's ability to capture the variance in the data.
- **Dominant Features:** Across both tasks, departure delay (DEP_DELAY) was the primary predictor of arrival delay. Secondary contributors included carrier, scheduled departure time, and late aircraft delay. Feature selection showed that using only the top 5–10 features preserved nearly the same predictive power as the full feature set, allowing for simpler and computationally efficient models.
- **Practical Implications:** High-performing models like LightGBM can support airlines in post-departure operational decisions, such as resource allocation, gate management, and early warning systems. Simplified models can maintain accuracy while reducing data collection and computational overhead.

6. FUTURE WORK

While this study focused on post-departure prediction of arrival delays, several opportunities exist to extend and deepen the analysis. One potential direction is to explore **pre-departure delay prediction**, examining how operational factors before takeoff can forecast both departure and subsequent arrival delays. Integrating weather data and other external factors could allow a more comprehensive study of how environmental conditions influence delays. This would enable modeling of **departure delays and their effects on arrival delays** as a connected process, providing airlines with actionable insights for proactive planning and scheduling.

Future studies could also investigate feature interactions and real-time predictive pipelines, combining multiple data sources to improve the accuracy and operational relevance of delay forecasts.

7. REFRENCES

[1] Using Big Data-machine learning models for diabetes prediction and flight delays analytics

[2] Flight Delay Prediction Based On Aviation Big Data And Machine Learning

[3] Improving Passenger Experience: Predicting Airline Delays Through Machine Learning

[4] Improved Methods for Predicting Flight Delay using Machine Learning Techniques

[5] Study of Delay Prediction in the US Airport Network

[6] Predictive Modeling of Flight Delays at an Airport Using Machine Learning Methods

[7] Enhancing Flight Delay Predictions Using Advanced Machine Learning and Data Analytics

[8] Airline Flight Delay Prediction Using Machine Learning Models

[9] Flight Delay Prediction Using Random Forest with Enhanced Feature Engineering

[10] Flight Delay Prediction by Machine Learning

[11] Flight Delay Prediction Using Different Regression Algorithms in Machine Learning

[12] Airline Flight Delay Prediction Using Machine Learning Models

[13] A novel intelligent approach for flight delay prediction

[14] A Methodology for Predicting Aggregate Flight Departure Delays in Airports Based on Supervised Learning

[15] Flight delay prediction based on deep learning and Levenberg-marquart algorithm

[16] Flight delay prediction based on aviation big data and machine learning

[17] Machine Learning Model - based Prediction of Flight Delay

[18] A hybrid machine learning-based model for predicting flight delay through aviation big data

[19] Spatio-Temporal Feature Engineering and Selection-Based Flight Arrival Delay Prediction Using Deep Feedforward Regression Network

[20] A novel hybrid method for flight departure delay prediction using Random Forest Regression and Maximal Information Coefficient

[21] Towards Real-Time and Accurate Flight Delay Predictions: A Comprehensive Review of Techniques

[22] Ground Handling Process Optimization Model Linked to Flight Delay Prediction Results

[23] Bureau of Transportation Statistics (BTS) TranStats system