

Predicting U.S. Flight Arrival Delay Minutes Using Historical Flight Status Data

Hejiong Zhao

*Department of Statistics, University of Washington, Seattle, USA
Kevinapply7@outlook.com*

Abstract. Flight delays can cause extra operating costs for airlines and negative experience for passengers. Accurate delay predictions can mitigate the effect. This study predicts U.S. domestic flight arrival delay minutes by using the Kaggle Flight Status Prediction dataset, which contains historical flight status records from 2018 to 2022. After removing cancelled and diverted flights, the study selects a sample of 300,000 completed flights, with 60,000 observations from each year. The target variable is ArrDelayMinutes with a range from 0 to 360 minutes. The study used temporal validation by using 2018-2021 flights for training and 2022 flights for testing. The experiment focuses on a departure-aware real-time prediction setting that includes actual departure delay information available before arrival. The study trained and compared ridge regression, random forest regression, and histogram gradient boosting regression. The departure-aware setting reduces the mean absolute error to approximately 5 minutes, and all trained models explain more than 91.8% of the test-set variance. The results show that actual departure delay is the most significant information for minute-level arrival delay prediction. Then model complexity is less important once this operational information is available.

Keywords: U.S. domestic flight delay prediction, Departure-aware prediction, Machine learning, Regression analysis, Temporal validation

1. Introduction

Air transportation is important in many fields. However, flight delays are a long-term problem in commercial aviation. Delays increase costs while reducing airport efficiency and passenger satisfaction. Also, delays can cause a chain reaction across the aviation network because of connections in flights. Therefore, accurately predicted delays can help airlines adjust operational plans and provide passengers with more reliable travel information.

Flight delays are generated by multiple interacting factors. The Bureau of Transportation Statistics identifies carrier operations, weather, National Aviation System conditions, security events, and late-arriving aircraft as major delay categories [1]. These factors can interact across different stages of a flight. For example, severe weather may reduce airport capacity, while the late arrival will delay the next departure. Zamkova et al. [2] found that delay duration is associated with operational characteristics such as flight type, aircraft capacity, aircraft utilization, and previous

flight delays. Their results show that earlier operational conditions are important for understanding subsequent arrival performance.

Data science has become an important approach to flight delay research. Carvalho et al. [3] conducted a systematic review of data science methods in this field and organized previous studies according to their research scope and computational methods. The review shows that flight delay research has gradually developed from conventional statistical analysis toward machine learning, large-scale data processing, and network-based prediction. Nevertheless, previous studies define prediction targets differently, including airport-level delay, individual-flight delay, binary delay status, delay categories, and continuous delay duration.

Several studies have revealed the usefulness of machine learning for flight-delay prediction. Rebollo and Balakrishnan [4] developed random-forest classification and regression models using temporal and network delay states. Their research showed that current delay conditions across airports and origin-destination connections have valuable information for forecasting future delays. Belcastro et al. [5] combined scheduled flight information with weather observations to predict individual arrival delays. The prediction was mainly formulated as classification above predefined delay thresholds. More recent studies have applied machine learning and deep learning methods. Gui et al. [6] developed a random forest based model using aviation big data and demonstrated that ensemble learning could provide strong binary delay-classification performance while reducing overfitting. Cai et al. [7] discussed flight delay prediction from a perspective of multiple airports and used time-evolving graphs to represent airport operations. Their convolutional model captured temporal patterns and changing spatial relationships among airports. It showed that delay prediction should consider interactions across airports. Overall, these studies indicate that nonlinear models, large datasets, and spatial temporal information can improve the delay prediction.

However, several research gaps still remain. First, many literatures treat flight delay prediction as a binary classification. They determine whether the delay exceeds a specific threshold (e.g., 15 minutes) [3, 5, 6]. It cannot distinguish between a short delay and a disruption of more than one hour. So, predicting delay duration in minutes can provide more detailed information for airports and passengers. Second, some models focus on specific airports, airlines, or specific periods. It restricts the generalizability. Third, some models use information that is only available after the flight has ended. Using actual arrival information can cause confidential data leakage and make unrealistic estimates of model performance.

To cover these gaps, this study investigates using historical flight-status information to accurately predict U.S. domestic arrival delay minutes. The Kaggle Flight Status Prediction dataset [8] is used to construct a 300,000 cleaned observations sample. The sample contains flight records from 2018 to 2022. The prediction is based on a departure-aware setting. This setting contains actual departure information available before arrival. The study evaluated ridge regression, random forest regression, and histogram gradient boosting regression models. These models used a temporal split, with data from 2018 to 2021 used for training and data from 2022 used for testing. Model performance is evaluated by mean absolute error (MAE), root mean square error (RMSE), and R-squared. The following sections describe the dataset and methodology, present the experiment and results, analyze the visual evidence, and discuss the limitations and conclusions.

2. Dataset and methods

2.1. Data source and sample selection

This study uses the Kaggle Flight Status Prediction dataset, from the U.S. Department of Transportation on-time performance data [8]. The downloaded archive contains annual combined flight files from 2018 to 2022. Each observation represents a flight and includes information about dates, airlines, origin and destination airports, scheduled departure and arrival times, distance, cancellation status, diversion status, and delay-related fields.

Cancelled and diverted flights were removed because these observations do not represent arrival-delay outcomes. Observations with missing ArrDelayMinutes values were also excluded. To reduce the influence of extreme outliers, the study retained arrival delays between 0 and 360 minutes. Flights delayed by more than six hours were excluded rather than winsorized, so their values were not replaced with the upper threshold. The experiment used a random seed of 42 to randomly select 60,000 flights from each year. The final sample contained 300,000 observations across five years. This sampling can prevent years with higher flight volumes from dominating model estimation. However, the sample does not represent an exact reflection of the annual volume distribution in the complete database.

2.2. Target variable and temporal split

The target variable is ArrDelayMinutes. It represents the absolute difference between the scheduled arrival time and the actual arrival time. Early or on-time arrivals were assigned a value of zero. Because the target variable is a continuous variable measured in minutes, the prediction task was formulated as a regression problem.

The experiment used temporal validation. Flights from 2018 to 2021 formed the training set, containing 240,000 observations. Flights from 2022 formed the test set, containing 60,000 observations. This temporal split is more practical because the deployed prediction model would use historical flights to predict future flights.

2.3. Prediction setting and feature design

The experimental setting is departure-aware arrival-delay prediction. The setting uses variables available before or at departure and updates the expected arrival delay after actual departure information becomes available. The retained features include month, quarter, day of month, day of week, scheduled departure and arrival times, scheduled elapsed time, distance, distance group, airline, operating carrier, origin and destination airports, origin and destination states, scheduled time blocks, route, weekend status, scheduled departure and arrival hours, actual departure time, departure-delay minutes, departure-delay status, the 15-minute departure-delay indicator, departure-delay group, and actual departure hour.

Because many final features are categorical identifiers, the feature-correlation heatmap in the results section focuses on key numeric features. It includes scheduled timing variables, distance variables, weekend status, departure-delay fields, actual departure hour, and the target variable. Categorical variables such as airline, airport, state, time block, and route were handled through model-specific encodings.

2.4. Leakage prevention and preprocessing

To avoid data leakage, actual arrival time, actual elapsed time, air time, wheels-on time, taxi-in time, arrival-delay group, and delay-cause variables were excluded from model features. Carrier delay, weather delay, National Aviation System delay, security delay, and late-aircraft delay are useful explanatory fields after a delay has occurred, but they are not valid predictors for a real-time model because they are typically known only after the flight outcome is observed.

2.5. Experimental environment and configuration

The analysis was implemented in Python. Scikit-learn was used for preprocessing, model training, and evaluation [9]. Pandas was used for data preparation and aggregation [10], while NumPy supported numerical operations. Matplotlib and Seaborn were used to generate the descriptive and diagnostic visualizations. Annual Parquet files were read directly from the compressed Kaggle archive with PyArrow.

Three regression algorithms were compared. Ridge regression used L2 regularization with an alpha value of 2.0. Random forest regression used 50 trees, a maximum depth of 18, a minimum of 10 observations per terminal leaf, and a 60% observation sample for each tree. Histogram gradient boosting used 180 iterations, a learning rate of 0.06, a maximum of 31 leaf nodes, and L2 regularization of 0.05. A random seed of 42 was maintained throughout the sampling and modeling.

Mean absolute error is the primary evaluation metric because it measures the average prediction error directly in minutes:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

Root mean squared error reveals greater penalties for large errors:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

R-squared measured the proportion of test-set variation explained by the predictions. Model selection was based primarily on MAE, with RMSE and R-squared used as supporting diagnostics. Predictions below zero were reset to zero because ArrDelayMinutes cannot be negative.

3. Results and analysis

3.1. Descriptive and temporal patterns

Table 1. Descriptive summary of the modeling sample

Statistic	Value
Sample size	300,000 flights
Years included	2018, 2019, 2020, 2021, 2022
Target variable	ArrDelayMinutes, filtered to 0-360 minutes
Mean arrival delay	11.33 minutes
Median arrival delay	0 minutes
Flights delayed more than 0 minutes	32.44%

Table 1. (continued)

Flights delayed at least 15 minutes	17.38%
75th percentile delay	6 minutes
90th percentile delay	33 minutes
95th percentile delay	66 minutes

In Table 1, the descriptive statistics show that arrival delay minutes are strongly right skewed. The mean arrival delay is 11.33 minutes, but the median is 0 minutes. Approximately 32.44% of flights have a positive arrival delay, and 17.38% are delayed by at least 15 minutes. The 75th percentile is 6 minutes, the 90th percentile is 33 minutes, and the 95th percentile is 66 minutes. These values show that most flights have little or no delay, while a smaller group of flights accounts for much longer disruptions.

3.2. Distribution and temporal patterns

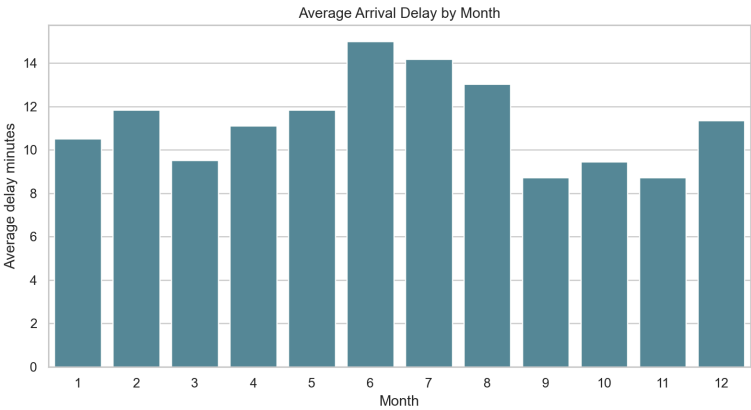


Figure 1. Average arrival delay by month (picture credit: original)

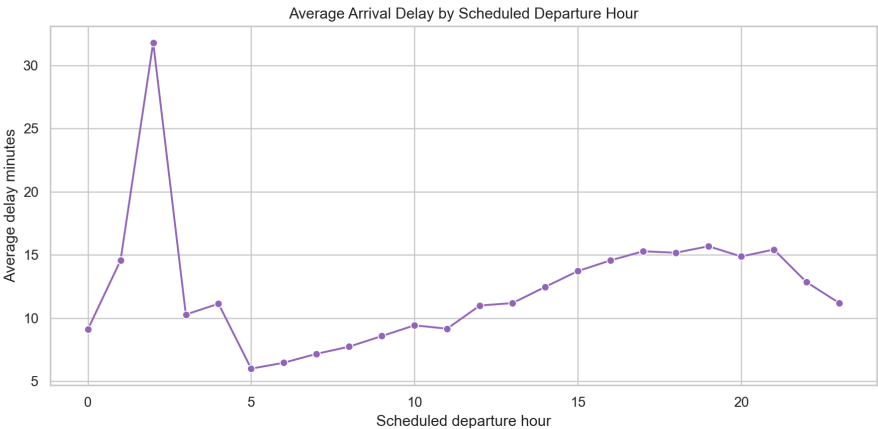


Figure 2. Average arrival delay by scheduled departure hour (picture credit: original)

Figure 1 shows that the average arrival delay varied across months. June had the highest average delay at approximately 15.00 minutes, followed by July at 14.18 minutes and August at 13.03 minutes. September and November had the lowest averages at approximately 8.71 minutes. This

pattern is consistent with more variation during the summer vacation travel, although the aggregated data cannot identify whether weather, airport congestion or other reasons caused the differences.

In Figure 2, early morning flights around 05:00 average about 6.00 minutes of delay, but flights from 17:00 to 21:00 average about 15 to 16 minutes. This pattern is consistent with the accumulation of operational disruptions over the day.

3.3. Feature correlation analysis



Figure 3. Correlation heatmap for key numeric departure-aware features and arrival delay (picture credit: original)

Figure 3 shows Pearson correlations among key numeric variables in the departure-aware features. The strongest relationship with arrival delay is the departure-delay group and departure-delay minutes. This result is consistent with the operational interpretation that the delay already accumulated by departure is often carried into arrival. It also shows that several timing features are naturally correlated with each other, such as scheduled departure hour and scheduled arrival hour. Distance and scheduled elapsed time are strongly related because longer routes require longer planned flight times. These patterns support the use of regularized linear models and tree-based models, both of which can handle correlated predictors when evaluated on a test dataset.

3.4. Model performance and diagnostics

Table 2. Departure-aware model performance on the 2022 test sample

Model	MAE	RMSE	R-squared
Ridge Regression	5.019	9.890	0.921
Random Forest Regression	5.056	9.877	0.921
Histogram Gradient Boosting	5.090	10.001	0.919

Table 2 shows that all retained features in departure-aware models explain more than 91.8% of the variation in the 2022 test sample, exceeding the 0.65 adequacy threshold. Ridge regression achieves the lowest MAE at 5.019 minutes and an RMSE of 9.890 minutes. Random forest regression produces a very similar MAE of 5.056 minutes and the lowest RMSE at 9.877 minutes. Histogram gradient boosting regression also performs strongly, with an MAE of 5.090 minutes and an R-squared value of 0.919. The three trained models are close in performance. This suggests that once actual departure delay is included, model complexity is not significant. A transparent regularized linear model can perform nearly as well as nonlinear ensemble models because actual departure delay summarizes much of the delay already realized before arrival.

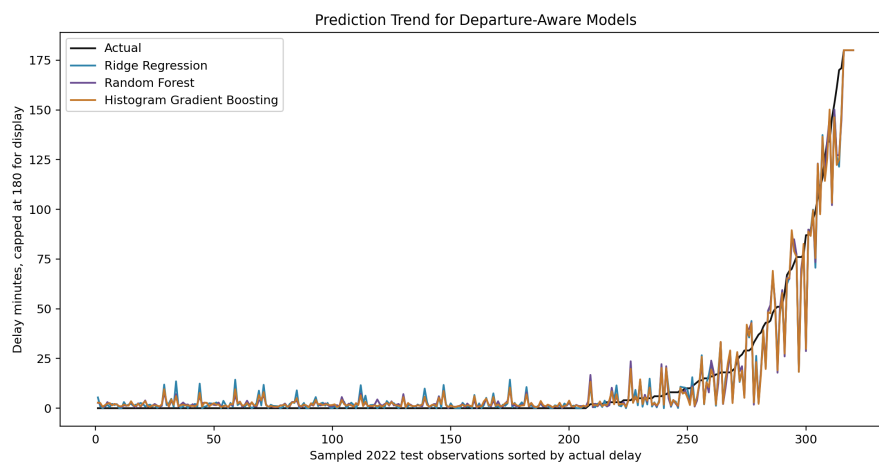


Figure 4. Prediction trend for departure-aware models on a sampled 2022 test subset (picture credit: original)

Figure 4 shows that the three models generally follow the increasing pattern of actual delay in the sampled test observations. Figure 5 further shows that most points follow the 45-degree reference line, although dispersion widens for severe delays. It shows that the models capture the general relationship between departure and arrival delays, but are less accurate for rare, severely delayed flights. The two plots provide diagnostic evidence that the models can produce stable and meaningful predictions.

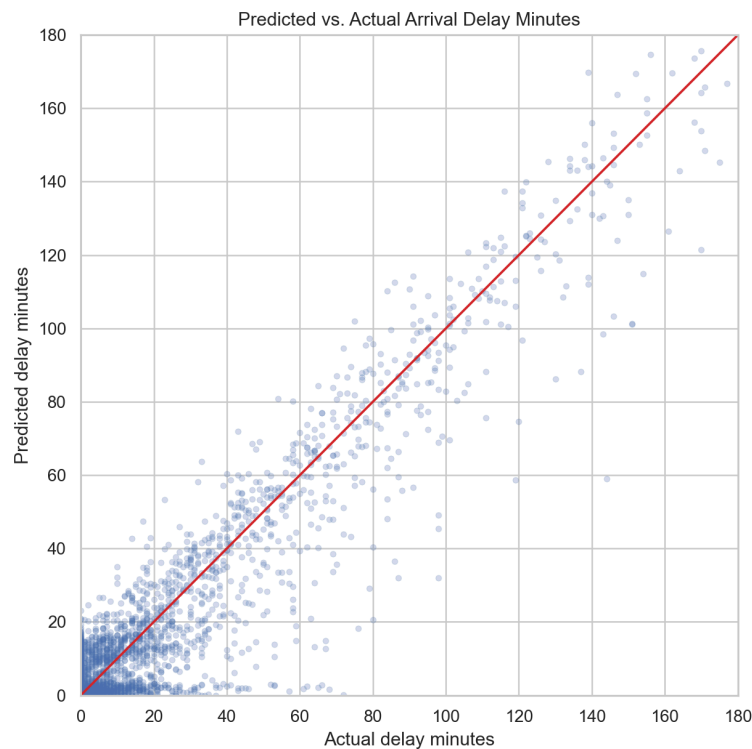


Figure 5. Predicted versus actual arrival delay minutes for the best departure-aware model (picture credit: original)

4. Discussion and implications

The results demonstrate that the departure-aware model is significant for minute-level delay prediction. Actual departure delay does not make data leakage because it is known before the flight arrives. It already covers accumulated information about disruptions before takeoff. This explains the reason that the retained models achieve R-squared values above 0.918 on the 2022 test set.

Ridge regression, random forest regression, and histogram gradient boosting all had strong performance. It suggests that model complexity is not the main determinant of predictive accuracy, once departure-aware information is included. Ridge regression is especially suitable for deployment because it combines high accuracy and easier interpretation.

The findings also support a staged prediction. Before departure, schedule information may be more suitable for delay classification. After actual departure information becomes available, regression models can update expected arrival delay minutes much more accurately. Such a two-stage framework would make prediction features with the information available at each operational moment.

Meanwhile, this study has several limitations. The experiment uses an equal annual sample rather than the full dataset. The target variable is filtered to a maximum of 360 minutes, which reduces the effect of extreme outliers but also removes some severe disruption cases. The model excludes external weather, air-traffic-control restrictions, airport capacity, gate availability, crew scheduling, or aircraft-rotation variables. In addition, the 2018-2022 period includes the COVID-19 pandemic, which may affect traffic patterns and delay mechanisms.

Future research should incorporate weather observations, airport capacity measures, traffic-management restrictions, and aircraft connection networks. Temporal validation could test whether model performance keeps stable in changing aviation conditions. Future studies can evaluate model

performance across airports, airlines, even seasons to examine whether prediction errors are unevenly distributed

5. Conclusion

This study investigated whether historical U.S. flight status data can be used to predict arrival delay duration in minutes. Using the Kaggle Flight Status Prediction dataset, the experiment selects a balanced sample of 300,000 completed flights from 2018 to 2022. The sample removes cancelled and diverted flights and filters arrival delays to 0-360 minutes. The study evaluates models by a temporal split in which 2018-2021 was used for training and 2022 was used for testing. The study evaluated model performance by using temporal validation, with data from 2018 to 2021 used for training and data from 2022 used for testing.

The experiment focuses on departure-aware prediction. The retained trained models got high R-squared values and correspond to a realistic real-time prediction setting. Ridge regression achieves the lowest MAE at 5.019 minutes, while random forest regression and histogram gradient boosting regression produce similar results. All three trained models explain more than 91.8% of the test-set variance.

Actual departure delay is the strongest available predictor of arrival delay minutes, once post-arrival leakage variables are excluded. This finding indicates that accurate arrival-delay prediction should be updated after actual departure information is observed. The study also shows that transparent regression models can be competitive with more complex ensemble models when the feature set is aligned with the intended prediction time. Future research can improve the framework by adding weather, airport-capacity, air-traffic-control, and aircraft-rotation information, and by evaluating prediction intervals and subgroup performance across operational contexts.

References

- [1] Bureau of Transportation Statistics. (n.d.). Understanding the reporting of causes of flight delays and cancellations. <https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations>
- [2] Zamkova, M., Rojik, S., Prokop, M., & Stolin, R. (2022). Factors affecting the international flight delays and their impact on airline operation and management and passenger compensations fees in air transport industry. *Sustainability*, 14(22).
- [3] Carvalho, L., Sternberg, A., Goncalves, L. M., Cruz, A. B., Soares, J. A., Brandao, D., Carvalho, D., & Ogasawara, E. (2021). On the relevance of data science for flight delay research: A systematic review. *Transport Reviews*, 41(4), 499–528.
- [4] Rebollo, J. J., & Balakrishnan, H. (2014). Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, 44, 231–241.
- [5] Belcastro, L., Marozzo, F., Talia, D., & Trunfio, P. (2016). Using scalable data mining for predicting flight delays. *ACM Transactions on Intelligent Systems and Technology*, 8(1), Article 5, 1–20.
- [6] Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, 69(1), 140–150.
- [7] Cai, K., Li, Y., Fang, Y.-P., & Zhu, Y. (2022). A deep learning approach for flight delay prediction through time-evolving graphs. *IEEE Transactions on Intelligent Transportation Systems*, 23(8), 11397–11407.
- [8] Kaggle. (n.d.). Flight status prediction [Data set]. <https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022>
- [9] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [10] McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 56–61).