A Cost-Sensitive Approach to Strategic Flight Delay Prediction[*]

Ruimeng Liu

Lehigh University, Bethlehem, PA 18015, USA
rul325@lehigh.edu

Abstract. Flight delays impose significant economic losses and operational challenges on the global aviation industry. Existing prediction research primarily focuses on real-time forecasts within short time windows (T-0 to T-2 hours) before departure and often falls into the "accuracy trap" when dealing with highly imbalanced data. Although models may achieve high accuracy, their low recall rates fail to capture actual delays, lacking practical business value. This paper proposes a strategic flight delay prediction framework targeting 12 to 24 hours before departure (T-12/T-24). To address the lack of real-time data in long-term predictions, we design a strategic feature system based on "Scheduled Traffic Congestion" and "Historical Network Inertia". Experiments on 28 million flight records across the United States from 2018 to 2022 demonstrate that the strategic model exhibits stronger robustness in cross-year predictions compared to real-time models (ROC AUC 0.70). Although the model's accuracy (61%) is lower than the zero-rule baseline (81%) while pursuing a high recall rate (68%), further cost-sensitive analysis reveals its true value. By effectively warning of high-cost missed detections (Type II Error), the model reduces comprehensive operational costs by approximately 42% compared to the baseline. This study confirms that the "better false alarm than missed alarm" prediction strategy yields significant practical benefits in aviation strategic planning.

Keywords: Flight delay prediction · Cost-sensitive learning · Imbalanced classification · Strategic planning · Machine learning

1 Introduction

Flight delays remain a persistent and costly challenge in the global aviation system, generating substantial economic losses and operational disruptions for airlines, airports, and passengers. According to industry reports, even small deviations from scheduled operations can propagate rapidly through the air traffic network, leading to cascading delays, crew misalignment, and passenger dissatisfaction. As a result, accurate flight delay prediction has long been regarded as a critical decision-support tool for airline operations and air traffic management.

[*] Supported by research funding.

Existing research in flight delay prediction has predominantly focused on *tactical-level* forecasting, typically within a short time window of 0–2 hours before departure. Within this horizon, models can leverage rich real-time information, such as upstream flight delays, aircraft turnaround status, and near-term weather conditions. These approaches often report high predictive accuracy, frequently exceeding 80%. However, such results are misleading in practice due to two fundamental issues.

First, aviation delay data exhibits severe class imbalance: delayed flights usually account for less than 20% of all observations. Under this setting, models optimized for accuracy tend to converge to the *zero-rule* solution—predicting all flights as on-time—thereby achieving deceptively high accuracy while failing to detect actual delays. This phenomenon, commonly referred to as the *accuracy trap*, renders many reported models operationally ineffective, as they provide little actionable warning for high-risk flights.

Second, short-horizon predictions, while accurate, offer limited operational value for strategic decision-making. Predictions made only minutes or hours before departure leave insufficient time for airlines to adjust capacity, reposition aircraft, reschedule crews, or proactively inform passengers. From a planning perspective, *strategic-level* predictions made 12–24 hours in advance (T-12/T-24) are far more valuable. However, extending the prediction horizon introduces a new challenge: real-time operational features become unavailable, and models relying on such information experience a sharp degradation in performance.

To address these limitations, this paper proposes a strategic flight delay prediction framework designed explicitly for long-horizon forecasting under strict temporal constraints. Rather than relying on real-time delay propagation features, we construct a leakage-free feature system based on *scheduled traffic congestion* and *historical network inertia*, capturing structural risks inherent in the air traffic network. Furthermore, recognizing that the operational cost of missed delay warnings far exceeds that of false alarms, we formulate the prediction task as a cost-sensitive learning problem.

Using 28 million U.S. domestic flight records from 2018 to 2022, evaluated under a rigorous time-series split, we demonstrate that while strategic models inevitably sacrifice accuracy, they achieve substantially higher recall and stronger robustness across years. More importantly, cost-based evaluation shows that prioritizing recall leads to significant reductions in overall operational cost. The results confirm that, for strategic flight planning, a “better false alarm than missed alarm” paradigm provides superior practical value compared to traditional accuracy-driven approaches.

2 Related Work

In recent years,Data-driven flight delay prediction has become one of the hot topics in machine learning forecasting.

2.1 Tactical Prediction

Early research are more focused on predict 0-4 hours in advance, for example, Chakrabarty [1] implemented approximately 85% accuracy by using random forest model and gradient boosting algorithm on history data from American Airlines. In fact, in Chakrabarty's experiments, the method of data partitioning is random split strategy, but the distribution of samples in dataset are distributed along timeline,result in a probability of using "feature data" to predict past. It's not feasible in actual deployment. Besides, Chakrabarty's testing is performed on validation set after oversampling with SMOTE. Issues that may lead to data breaches. Moreover,the original data set is extremely imbalanced – 20% samples status are delay.Therefore, these models often fall into the "accuracy trap," where they tend to predict all flights as "on time" (Zero Rule) in pursuit of high accuracy. While the statistical metrics appear favorable, their recall rate for actual delays is extremely low, rendering them ineffective at providing airlines with meaningful risk warnings.

2.2 Strategic Prediction and Network Horizons

To support airline capacity scheduling, the forecasting window must be advanced to 12–24 hours prior to departure (T-12/T-24). Pioneering research by Rebollo and Balakrishnan [3] indicates that as the forecasting time window extends, the predictive capability based on real-time conditions—such as preceding flight delays—exponentially diminishes. They proposed leveraging aviation network graph features to capture delay propagation effects. Our research builds upon this approach but employs stricter leak-free feature engineering, substituting real-time propagation features with "scheduled congestion" and "historical network inertia" to achieve more robust predictions within the T-24 window.

2.3 Cost-Sensitive Learning in Aviation Context

Although cost-sensitive learning has become a standard paradigm in high-risk domains such as financial fraud detection and medical diagnosis, it remains underutilized in flight delay prediction research. Most existing literature implicitly assumes symmetric error costs (i.e., $Cost_{FP} = Cost_{FN}$), which contradicts the actual economic logic of aviation operations. From a machine learning theoretical perspective, Elkan [5] notes that in scenarios with extreme class imbalance and asymmetric error costs, traditional classifiers targeting maximum accuracy often produce misleading decision boundaries, leading to systematic neglect of minority class samples. Specifically in the aviation domain, Ball et al. [6], in their study for the Federal Aviation Administration (FAA), quantified the comprehensive impact of flight delays, pointing out that the costs of passenger stranding, itinerary disruptions, and crew overtime caused by irregular operations significantly exceed the costs of preventive measures (such as standby arrangements or capacity adjustments). Based on this theoretical and empirical evidence, this study argues that in strategic-level prediction, the cost of false negatives (Type

II Error) far outweighs that of false positives (Type I Error). Therefore, we introduce an asymmetric cost matrix evaluation framework aimed at filling the gap in economic utility assessment in existing research, demonstrating that sacrificing some accuracy for high recall is the optimal strategy from a commercial perspective.

3 Methodology

3.1 Data Source and Preprocessing

Experimental dataset is extracted from U.S. BTS(Bureau of Transportation Statistics)https://www.bts.gov/airline-data-downloads , which is an on-time performance database. I retrieved 28 million flights from 2018 to 2022, 5 years in total.To assess the primary impact of meteorological factors on flight delays, we integrated historical hourly weather data from the Meteostat API. Weather records—including temperature, wind speed, precipitation, and visibility—were synchronized with flight records based on the nearest airport weather station data and scheduled departure times.

Before feature engineering, we performed rigorous data cleaning. Flights marked as "cancelled" or "diverted" were excluded, as their delay duration could not be determined. The missing values in numerical features were filled using the median strategy to maintain the robustness of the distribution, while the missing values in categorical features were labeled "UNK". The prediction task was defined as a binary classification problem. The target variable `DepDel15` is defined according to the U.S. Federal Aviation Administration (FAA) standard: if the actual departure time of a flight is delayed by more than 15 minutes compared to the scheduled departure time, it is marked as delayed ($y = 1$); otherwise, it is considered on-time ($y = 0$). However, this dataset exhibits significant class imbalance, with positive class (delayed) samples accounting for only about 19% of the total samples.

A common methodological flaw in prior literature is the use of random k-fold cross-validation, which inadvertently introduces temporal data leakage by incorporating future information into the training set (e.g., utilizing December flight data to train models for predicting January flights). To ensure a realistic evaluation of forecasting capabilities and mitigate look-ahead bias, we implemented a rigorous temporal partitioning strategy:

- **Training set (2018–2020):** Employed for model development and training.
- **Validation set (2021):** Utilized for hyperparameter optimization and early stopping.
- **Test set (2022):** A completely held-out independent period, reserved exclusively for final performance assessment to simulate real-world operational conditions.

A significant challenge in strategic prediction (T-12/T-24) is the unavailability of real-time operational data. Unlike tactical models that utilize upstream delays (e.g., *"departure delay of the incoming aircraft"*), strategic models must rely solely on *a priori* information available 24 hours before departure. To address this, we engineered a set of leakage-free features capturing structural network risks.

3.2 Scheduled Traffic Pressure (Network Congestion).

Airport congestion is a primary driver of delay propagation. Since actual departure counts are unknown at the T-24 horizon, we constructed a proxy for capacity saturation using the flight schedule.We aggregated flight records based on the Computerized Reservations Systems (CRS)departure time (rounded to the nearest hour). For a given flight f scheduled at date d and hour h, we calculated:

- **Origin Scheduled Traffic:** The total number of flights scheduled to depart from the origin airport during time window $[h, h+1]$.
- **Destination Scheduled Traffic:** The total number of flights scheduled to arrive at the destination airport, capturing downstream network constraints.

These features represent the *planned* load on the airport infrastructure, which remains constant regardless of daily operational disruptions.

Historical Network Inertia (Structural Risk). We hypothesize that flight delays exhibit "inertia"—a structurally congested airport or a struggling airline tends to perform poorly over consecutive days. To capture this without data leakage, we implemented a Rolling Window Statistic with a strict temporal cut-off.For each entity E (where $E \in \{Origin, Airline, Route\}$), we calculated the historical delay rate R over a window of W days (where $W \in \{7, 30\}$):

$$R_{E,t}^{(W)} = \frac{\sum_{i=t-W}^{t-1} \mathbb{I}(y_{E,i} = 1)}{\sum_{i=t-W}^{t-1} N_{E,i}} \tag{1}$$

where t is the current flight date, and $\mathbb{I}(\cdot)$ is the indicator function for a delay. Crucially, the window is defined as $[t - W, t - 1]$ (closed on the left, open on the right). This ensures that statistics for a flight on day t are derived strictly from data up to day $t - 1$, completely eliminating look-ahead bias.

Meteorological and Temporal Context. While local weather forecasts are ideal, we utilized actual METAR data as a proxy for high-accuracy forecasts (a standard practice in baseline comparisons [3]). We incorporated hourly variables including Temperature, Visibility, Wind Speed, and binary flags for extreme weather (e.g., Snow, Heavy Rain). Additionally, temporal features such as *Month*, *Day of Week*, and *Season* were encoded to capture cyclical demand patterns.

3.3 Cost-Sensitive Learning Framework

We frame the prediction problem as a cost minimization task. Based on business logic, we define the cost matrix as follows:

$$Total\_Cost = N_{FN} \times 5 + N_{FP} \times 1 \tag{2}$$

This implies that the penalty for a false negative (missed delay) is five times that of a false positive (false alarm). To implement this in LightGBM, we utilize the `scale_pos_weight` parameter to enforce cost-sensitive learning.

Listing 1.1. Cost-sensitive configuration in LightGBM

```
model = LGBMClassifier(
    n_estimators=500,
    # Force the model to focus on the minority class,
    # sacrificing accuracy for improved recall
    scale_pos_weight=5.16,
    reg_lambda=10.0
)
```

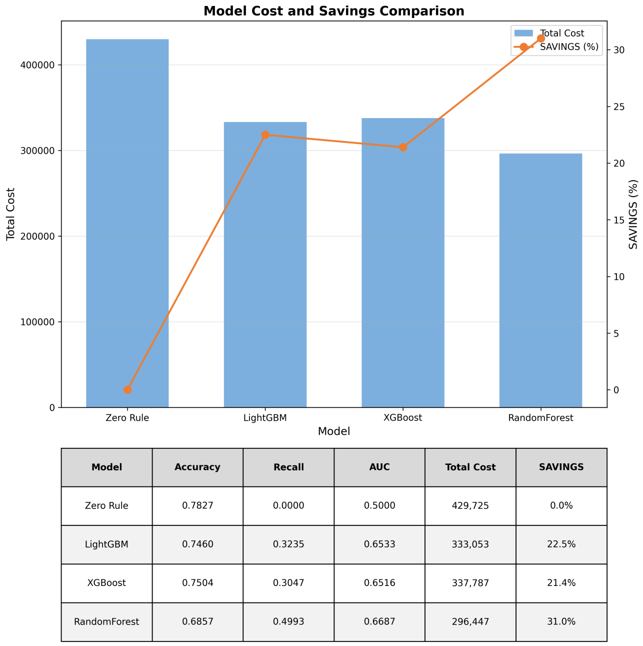
4 Experimental Results

4.1 Evaluation and Baseline setup

Like we mentioned before, Accuracy is only one aspect of evaluation; it sometimes fails to demonstrate a model's true capabilities. Moreover, we know the dataset used in this experiment suffers from extreme imbalance. Only 19% of samples represent delayed flights, meaning even if the model predicts all flights as on-time, it would achieve 80% or higher accuracy (Zero Rule baseline). Therefore, model evaluation in this paper requires a more comprehensive and meaningful approach. Therefore, this paper will first focus on the ROC AUC score to determine whether the model can genuinely learn useful classification capabilities from the extracted features. Next, recall will be examined to assess the proportion of delayed flights correctly identified by the model. Finally, its value will be judged based on total cost.

T-0 Real-time Model The T-0 model serves as the baseline model in this paper. It determines the immediate status of a flight—whether delayed or on time—by analyzing real-time data from the two hours preceding takeoff within the dataset.In Choi et al.[2] experiments, using multiple classifiers to make comparison. We compared results with classifiers: Zero Rule, XGBoost, LightGBM and RandomForest.

In Figure 1, we observe an interesting phenomenon: while we initially predicted an accuracy rate of approximately 81% for the Zero Rule, the final accuracy only reached around 79%. Although the difference is not significant, this may indicate potential missing values in the original dataset.



Model	Accuracy	Recall	AUC	Total Cost	SAVINGS
Zero Rule	0.7827	0.0000	0.5000	429,725	0.0%
LightGBM	0.7460	0.3235	0.6533	333,053	22.5%
XGBoost	0.7504	0.3047	0.6516	337,787	21.4%
RandomForest	0.6857	0.4993	0.6687	296,447	31.0%

Fig. 1. T-0 Horizon Results

Results from other classifiers show that accuracy rates hover around 75% for all except Random Forest, yet none surpass the Zero Rule's performance. While this might be considered a failed experiment in traditional machine learning, we achieved significant gains in Recall—particularly Random Forest's Recall nearing 50%, nearly doubling Choi's result[2].

Returning to the Zero Rule, its Total Cost reached 429,275 due to incorrectly predicting all delayed flights. In contrast, other classifiers in this experiment achieved savings ranging from 22% to 30%. This stems from our aggressive class weighting during training, where the model was designed to capture more delayed flights. The 66.87% AUC precisely demonstrates that within this baseline, the model exhibits a degree of risk resilience and the ability to effectively learn dataset features.

4.2 T-12 and T-24 Horizon Comparison

Training the T-12 and T-24 models is essential. T-0 represents real-time data, while most existing research is based on results from 1-hour, 2-hour, and 4-hour intervals[1]. Choi's predictions were made 15 minutes in advance[2]. However, these results lack practical utility because predictions made too close to the event may not allow airlines or passengers sufficient time to react or mitigate losses. Thus, predictions made 12 and 24 hours in advance address this critical gap.

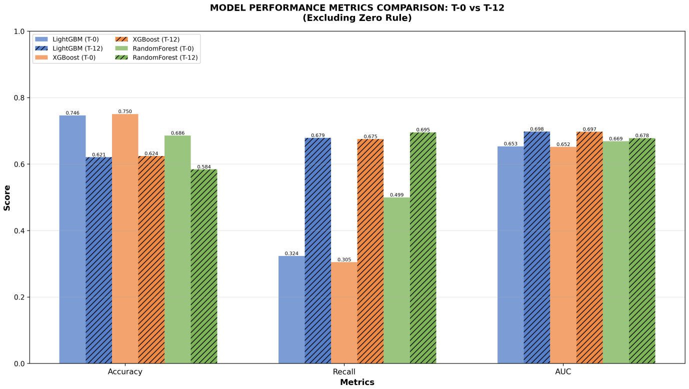


Fig. 2. Rcall, Accuracy and ROC AUC ofT-12 Horizon Model

Yet another issue arises: since predictions extend beyond half a day, certain fields in the original dataset become redundant, such as ‘rate_past_2h‘. This stems from prior research by Y Wang et al.[7], which indicated that over 40% of all flight delays stem from delay propagation. Delay propagation occurs when a preceding flight's delay likely causes subsequent flights on the same route to also be delayed, creating a chain reaction. This phenomenon plays a significant role in predicting flight delays.

Given the current inability to utilize such fields for model training, we must identify substitutes. Therefore, we decided to focus on OD Pair Historical Statistics, Scheduling Density, and Turnaround Pressure. After joining these “historical features,” multiple parameter adjustments yielded performance results for the T-12 and T-24 models.

Results of Two Models Respectively In the performance comparison between the T-12 model and the T-0 model shown in Figure 2, we observe that the T-0 model underperforms the T-12 model in all metrics except accuracy. Upon closer examination, the Random Forest model demonstrates superior performance across all metrics except accuracy, where it trails other classifiers. Macroscopically, the T-12 model's accuracy fluctuates between 58% and 62%, representing the optimal result achieved after parameter tuning. This significant accuracy reduction compared to T-0 serves as a warning signal. Our analysis suggests that predicting flight delays 12 hours in advance requires incorporating more features for training, substantially increasing the model's learning complexity. However, it is exciting to note that while sacrificing less than 15% of accuracy, not only did the AUC improve—approaching nearly 70%, indicating the model better captures delay patterns—but recall nearly doubled from the baseline. A recall rate nearing 70% signifies the model can now detect delays for the vast majority of flights, representing an outstanding achievement. Therefore,

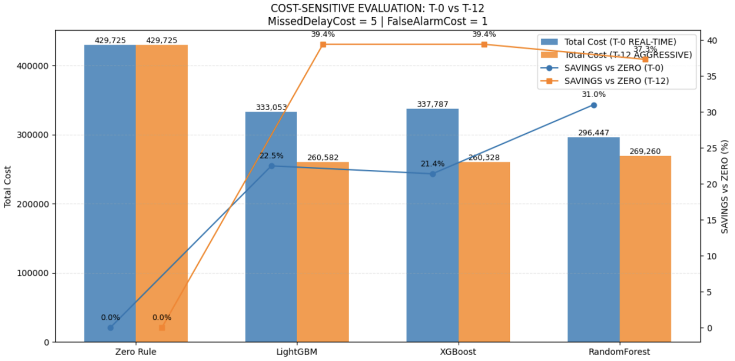


Fig. 3. Total Cost Compare with T-0 Model

after replacing some features related to delay propagation, the added historical features effectively helped the model learn how to capture delayed flights.

Figure 3 further demonstrates that T-12 not only outperforms the baseline model in terms of performance but also achieves significantly greater loss reduction. Calculations show that T-12 achieves a maximum loss avoidance of 39.4% on XGBoost, representing an 18% improvement over T-0. This implies that when relying on this model's predictions, airlines and passengers facing flight delays could mitigate substantial losses by devising effective countermeasures within 12 hours. Comparing T-12 and T-0 results reveals that the model achieves greater expected loss reduction by sacrificing some accuracy—signifying a significant enhancement in overall model performance and strategy optimization.

Compare to T-24 Horizon Building upon the T-12 model, the experiment was extended to the T-24 model. Like the T-12 model, the T-24 model also belongs to the strategic phase. At this stage, the model primarily learns and makes predictions through schedule planning and historical statistical features. However, the extended time interval means it becomes more challenging to search for preceding flights when predicting the target flight. Under consistent training logic, this may result in slightly lower prediction accuracy compared to the T-12 model. To capture more delayed flights, a more aggressive recall approach is adopted here. Let us observe the results in Figure 4.

In terms of results, the current model has nearly achieved performance on par with the T-12 model. Regarding the AUC metric, the T-12 and T-24 models show virtually no difference across various classifiers, with XGBoost and LightGBM still reaching nearly 70%. This indicates that at this experimental stage, as the prediction lead time increases, the model can still capture key features to determine whether a delay will occur.

Simultaneously, the steadily improving recall rates further substantiate this point. The Random Forest classifier achieved a recall rate of 71.8% for delayed

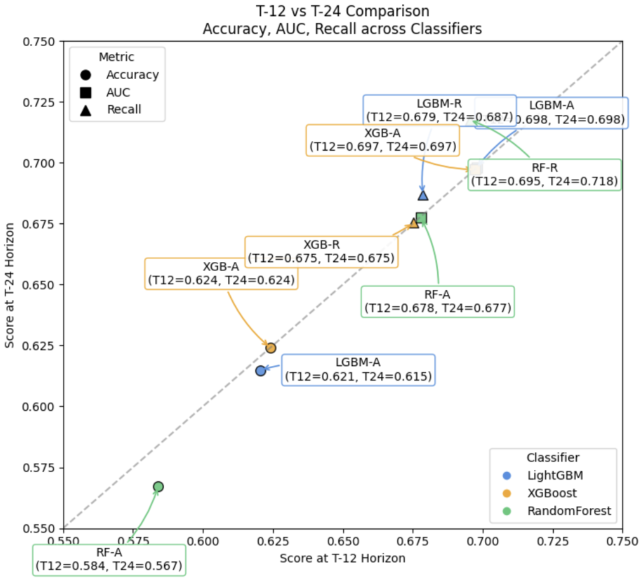


Fig. 4. Comparison Between T-12 Model and T-24 Model

samples; LightGBM also saw a slight increase to 68.69%, while XGBoost maintained its previous level.

Extending the prediction lead time correspondingly reduces accuracy slightly. LightGBM and Random Forest sacrificed only about 1%-2% accuracy, while XGBoost's accuracy remained virtually unchanged. This demonstrates XGBoost's remarkable stability in this experiment and on this dataset.

In terms of total cost and savings, the three classifiers remained nearly identical, with LightGBM achieving a marginal 0.1 percentage point improvement to 39.5%. From a loss recovery perspective, all three classifiers performed consistently well. Compared to the Zero Rule, they captured 68% of delays, reducing total costs to 250k. This resulted in approximately 42% cost savings, representing the most significant business value demonstration in this study.

4.3 Feature Importance Analysis

Figure 6 shows the 20 most important features identified from experiments on the Random Forest model, which performed best in the comprehensive experiments. Table 1 summarizes the top ten most important features. From these two sets of results, this paper distills three core logics revealing delays:

Temporal Dynamics Dominate From the results, we can see that Sched Dep Hour (Rank 1) and Part Of Day Early Morning (Rank 2) occupy the top positions, accounting for over 24% of the total weight. This actually reflects the

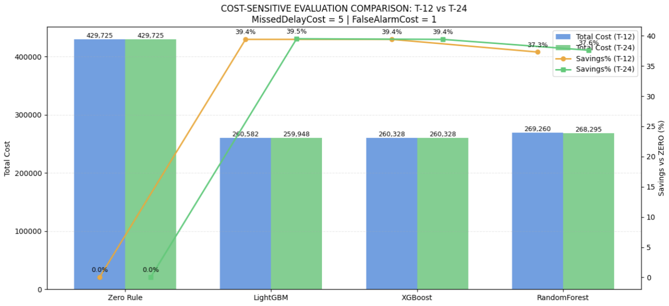


Fig. 5. Total Cost Comparison

Table 1. Top 10 Influential Features for Flight Delay Prediction (Random Forest)

Rank	Feature Name	Importance Score
1	Sched Dep Hour	0.1351
2	Part Of Day Early Morning	0.1080
3	Airline Delay Rate 7D	0.1079
4	Airline Delay Rate 30D	0.0967
5	Origin Delay Rate 7D	0.0946
6	Origin Delay Rate 30D	0.0729
7	Route Delay Rate 30D	0.0609
8	Part Of Day Evening	0.0488
9	Temp Avg	0.0328
10	Year	0.0309

aviation industry's well-known “ripple effect.” We can easily observe that early morning flights almost never experience delays. This is because the aircraft are already parked at their gates, free from the disruption of delay propagation from preceding flights.

Since the aforementioned delay propagation is the primary cause of most flight delays, evening flights are the most prone to delays. Any minor delay in preceding flights accumulates throughout the day, culminating in evening departures. The model discussed in this paper effectively captures this phenomenon. It recognizes that the departure time is a more critical factor than the specific flight details.

Recent history is more important than long-term history. Additionally, Table 1 shows that the 7-day airline delay rate (0.1079) is higher than the 30-day airline delay rate (0.0967). The same pattern is observed for origin airports. There is a reason why the 7-day delay average is higher. The 7-day average reflects the current operational state. If an airline experiences widespread delays this week due to strikes, system failures, or other reasons, the 7-day average captures this immediately, whereas the 30-day average is diluted. The model

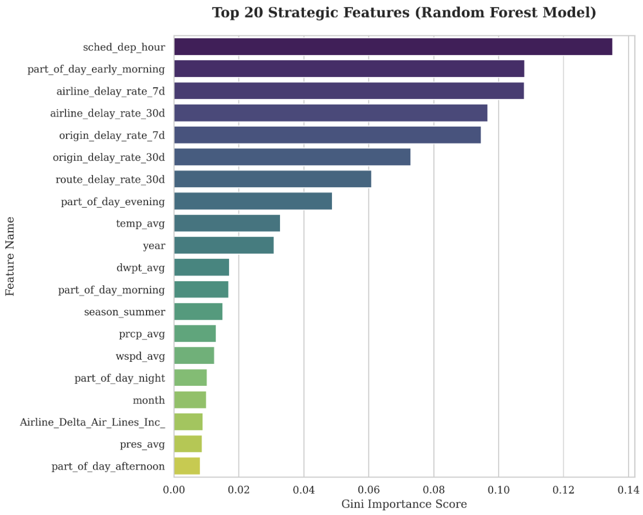


Fig. 6. Top 20 Important Features(Random Forest Model)

Table 2. Flight Delay Cause Analysis

Cause	**Avg Minutes**	**Severity**
Weather	71.44	High
Late Aircraft	55.33	High
Carrier	48.51	Medium
Security	31.96	Medium
NAS	30.85	Medium

effectively captures this phenomenon, which this paper refers to as short-term operational inertia.

T-24 and T-12 Phase: Systemic Factors Override Weather Conditions In common perception, we often assume that adverse weather is the primary factor determining flight delays. Therefore, this article aims to provide statistical analysis on this point. Table 2 shows the average delay duration for each cause of flight delays, while Table 3 presents the total delay duration and proportion of delays caused by each category overall. Looking solely at Table 2, weather-related delays do indeed have the longest average duration. However, when combining Tables 2 and 3, we observe that weather's negative impact is secondary—both in terms of total delay duration and proportion of delay causes.

Another reason is that at the T-24 (24 hours prior) stage, specific weather forecasts are often imprecise. Moreover, common adverse weather like snow or thunderstorms is typically already factored into the schedule. Thus, only extreme weather becomes significant. However, in the feature ranking, time-driven

Table 3. Total Delay Minutes by Cause

Cause	Total Minutes (M)	Share (%)
Late Aircraft	126.93	41.54
Carrier	113.49	37.14
NAS	45.50	14.89
Weather	19.02	6.23
Security	0.62	0.20

Note: Values in millions of minutes (except percentage).

systemic congestion and historically determined airline efficiency emerge as the primary sources of delays. This explains the phenomenon raised at the beginning of this section.This also demonstrates that the model presented in this paper is indeed a strategic model, rather than one based on meteorological forecasts.

5 Conclusion

In summary, our proposed cost-sensitive flight delay prediction framework successfully addresses the challenges posed by extreme class imbalance and long forecasting horizons. By shifting the focus from pursuing surface accuracy to enhancing recall, we avoid the "accuracy trap" and demonstrate robust predictive capabilities even when forecasting 12–24 hours in advance. The proposed model achieves approximately 61% accuracy, which is lower than the baseline level of 81% when consistently predicting normal flights. However, its delay capture capability is significantly enhanced, with a recall rate approaching 70%. This trade-off prioritizes "better to be wrong than to miss," and it proves valuable: the model effectively alerts for the vast majority of high-risk delayed flights, reducing overall operational costs by approximately 42% compared to the baseline.

Experimental results confirmed several key insights regarding flight delays. First, flight delays exhibit distinct temporal dynamics. Early morning flights rarely experience delays, which gradually accumulate throughout the day, peaking during evening flights. By leveraging features such as scheduled departure times, the model captured this cascading effect, indicating that a flight's scheduled departure time critically influences its delay risk. Second, recent historical performance proved more significant than long-term averages. The delay rate over the preceding 7 days exerted greater influence than the 30-day average, reflecting short-term "inertia" in operational efficiency. That is, if an airline or airport experienced high delays the previous week, the likelihood of delays the following day increased substantially—a trend successfully encoded by our feature engineering. Third, within the 12–24 hour forecast window, systemic network factors outweigh immediate weather conditions. While extreme weather may cause severe delays for individual flights, our analysis shows that nationwide, delays stemming from structural factors—such as inherent network congestion and inefficient airline scheduling—account for a larger proportion of total delay dura-

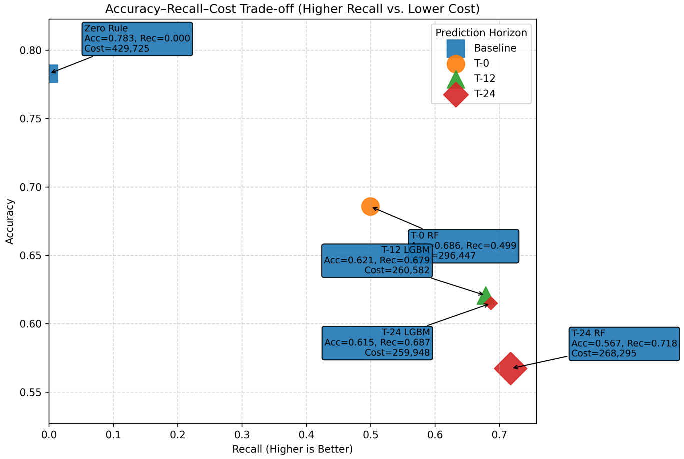


Fig. 7. Accuracy–Recall–Cost Trade-off Across Prediction Horizons

tion. Accordingly, the most significant predictors in our strategic model—such as planned traffic volume and historical delay rates at departure/arrival airports—exceed the predictive power of real-time weather variables, underscoring that structural congestion risks dominate as delay drivers during advance planning phases.

Overall, this study demonstrates that cost-sensitive prediction methods prioritizing high recall yield substantial benefits for aviation strategic planning. By providing advance warnings for the majority of potential delays, airlines and airports can proactively allocate resources or adjust schedules, mitigating the cascading effects and economic losses caused by delays. Our research validates that sacrificing some accuracy for higher recall represents an optimal strategy in flight delay prediction. The predictive philosophy of “better to over-report than under-report” not only enhances delay capture rates but also translates into significant cost savings, offering fresh insights into how machine learning models can generate tangible value in flight operations management.

References

1. Chakrabarty, N.: A data mining approach to flight arrival delay prediction for American airlines. In: 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON), pp. 102–107. IEEE (2019)
2. Choi, S., Kim, Y.J., Briceno, S., Mavris, D.: Prediction of weather-induced airline delays based on machine learning algorithms. In: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), pp. 1–6. IEEE (2016)
3. Rebollo, J.J., Balakrishnan, H.: Characterization and prediction of air traffic delays. Transportation Research Part C: Emerging Technologies **44**, 231–241 (2014)

4. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
5. Elkan, C.: The foundations of cost-sensitive learning. In: International Joint Conference on Artificial Intelligence, vol. 17, no. 1, pp. 973–978. Lawrence Erlbaum Associates Ltd (2001)
6. Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A.R., Peterson, E., Sherry, L., Trani, A., Zou, B.: Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States. University of California, Berkeley. Institute of Transportation Studies (2010)
7. Yanjun Wang, Yakun Cao, Chenping Zhu, Fan Wu, Minghua Hu, Vu Duong, Michael Watkins, Baruch Barzel, and H. Eugene Stanley, "Universal patterns in passenger flight departure delays," *Scientific Reports*, vol. 10, no. 1, p. 6890, 2020.