

Flight Delay Prediction

Machine Learning Analysis Report

January - March 2016

Data Analysis and Predictive Modeling

Author: Gowtham Sai Bhuvanam

Executive Summary

This report presents a comprehensive machine learning analysis for predicting flight departure delays of 15 minutes or more. The project integrates flight operational data with hourly weather conditions across 10 major U.S. airports for the period January-March 2016.

Key findings include that departure delays are driven primarily by time-of-day effects with strong interactions between scheduled departure hour and day of week. Weather variables contribute non-linearly, particularly wind speed, precipitation, and visibility. The final XGBoost model achieved a ROC-AUC of 0.735 and recall of 0.543, demonstrating meaningful predictive capability for operational planning.

Metric	Value
Total Flights Analyzed	105,783
Delay Rate	19.97%
Airports Covered	10
Time Period	Jan-Mar 2016
Final Model ROC-AUC	0.735
Final Model Recall	0.543

1. Data Preprocessing

1.1 Flight Data Processing

The flight data preprocessing pipeline focused on extracting features that are known before departure while filtering out data quality issues and leakage risks.

Data Loading Strategy:

Three monthly CSV files (January, February, March 2016) were loaded with selective column reading to reduce memory footprint. Only 15 relevant features were retained from the original 100+ columns available in the Bureau of Transportation Statistics dataset.

Feature Selection Criteria:

- Temporal Features: Year, Month, DayOfMonth, DayOfWeek, FlightDate, CRSDepTime, DepTimeBlk
- Operational Features: UniqueCarrier, Origin, Dest, Distance, DistanceGroup
- Outcome Indicators: Cancelled, Diverted, DepDel15 (target variable)

Airport Filtering:

Flights were filtered to include only those with origin airports matching available weather data coverage: ATL, EWR, JFK, LAS, LAX, MCO, MIA, ORD, SEA, and SFO. This ensured complete weather-flight data alignment.

Handling Missing Values:

Missing values in the target variable (DepDel15) were found to be structurally linked to cancelled flights. These records were removed as they represent flights that never departed, making delay prediction irrelevant. This resulted in 105,783 valid flight records for analysis.

1.2 Weather Data Processing

Weather data was provided as nested JSON files containing daily summaries with embedded hourly observations. The preprocessing pipeline flattened this hierarchical structure into a tidy hourly format suitable for merging with flight data.

JSON Structure Exploration:

Each airport had three JSON files (one per month) containing a 'weather' array with 28-31 daily objects. Each daily object contained a nested 'hourly' array with 24 observations. The first step involved understanding this nested structure through exploratory loading.

Feature Extraction Process:

- Temperature: tempC, FeelsLikeC
- Wind: windspeedKmph, windgustKmph, winddirDegree
- Atmospheric: humidity, pressure, visibility
- Precipitation: precipMM
- Cloud Cover: cloudcover
- Weather Code: weatherCode (categorical indicator)

Data Type Conversions:

All numeric weather features were explicitly cast from string to numeric types. Missing or invalid values were coerced to NaN and subsequently validated. No missing values were found in the final hourly weather dataset, indicating complete coverage from the data provider.

1.3 Merging Flight and Weather Data

The integration of flight and weather data required careful key construction to match hourly weather observations with scheduled departure times.

Join Key Construction:

The merge was performed on three dimensions: airport (Origin = airport), date (FlightDate = date), and hour (dep_hour = hour). The departure hour was derived from CRSDepTime by integer division ($\text{CRSDepTime} // 100$) to convert HHMM format to 0-23 hour buckets.

Merge Validation:

A left join was used to preserve all flight records. Post-merge validation confirmed 100% weather coverage with zero missing values in weather features. This perfect match rate validated both the temporal alignment strategy and the completeness of the weather dataset.

Final Dataset Structure:

The merged dataset contained 105,783 flights with 26 features (16 flight features + 10 weather features). Redundant columns (duplicate airport and date fields) were removed, and the final dataset was saved as flight_weather_data.csv.

2. Exploratory Data Analysis

2.1 Target Variable Distribution

The target variable DepDel15 exhibited class imbalance with approximately 80% on-time flights and 20% delayed flights. This imbalance is realistic and reflects actual operational patterns. Rather than resampling, the analysis focused on using appropriate evaluation metrics (ROC-AUC, Precision-Recall) that account for class imbalance.

2.2 Temporal Patterns

Time-of-day emerged as the strongest individual predictor of delays through several analyses:

- **Delay Propagation Effect:** Delay probability increased steadily from morning (5-7 AM: ~10%) to evening (8-10 PM: ~30%), indicating cumulative delay accumulation throughout the day.
- **Day of Week Variation:** Fridays showed consistently higher delay rates across most hours, while Saturdays had slightly lower rates. Sunday evenings exhibited elevated delays due to return travel patterns.
- **Hour × Day Interaction:** The effect of departure hour was not uniform across days of the week, with Friday evenings showing particularly high delay risk. This non-additive relationship motivated the use of interaction-aware models.

2.3 Weather Effects

Weather features exhibited non-linear relationships with delays, explaining their weak linear correlations but real predictive value:

- **Wind Impact:** Wind speed and wind gusts showed positive correlation with delays ($r = 0.10$ and 0.08 respectively). The effect appeared threshold-based, with delays increasing sharply above certain wind speeds.
- **Precipitation:** Precipitation (precipMM) showed moderate positive correlation ($r = 0.09$) with delays. The relationship was non-linear, with even light precipitation increasing delay risk.
- **Visibility and Pressure:** Lower visibility and atmospheric pressure were associated with higher delays, though correlations were weak ($|r| < 0.1$). These effects were likely conditional on other weather factors.
- **Temperature:** Temperature showed minimal linear correlation ($r \approx 0.01$) but may interact with other weather conditions (e.g., freezing precipitation, heat-related runway issues).

2.4 Carrier and Airport Patterns

Delay rates varied significantly across carriers and airports, indicating operational heterogeneity:

- **Carrier Variation:** Delay rates ranged from ~5% to ~24% across carriers. Some carriers (OO, NK) showed high variability across hours, while others (AS, AA) maintained more consistent performance.
- **Airport Effects:** Major hub airports showed higher delay rates, likely due to congestion and complex connection networks. Origin airport proved to be an important categorical feature.

- **Route Distance:** Distance showed a weak negative correlation with delays ($r = -0.02$), suggesting longer routes may have more schedule buffer. However, the effect was minimal.

2.5 Feature Correlation Analysis

Correlation analysis identified multicollinearity among derived features and validated weather data quality:

- **Redundant Pairs Identified:** CRSDepTime and dep_hour ($r \approx 1.0$), Distance and DistanceGroup ($r \approx 0.95$), tempC and FeelsLikeC ($r \approx 0.9$), windspeedKmph and windgustKmph ($r \approx 0.85$). One feature from each pair was selected for modeling.
- **Weather Feature Relationships:** Expected atmospheric relationships were confirmed (e.g., humidity-cloudcover positive, visibility-humidity negative), validating data quality.
- **Target Correlations:** No feature showed strong linear correlation with DepDel15 (max $|r| = 0.14$ for dep_hour), confirming that delays are driven by interactions rather than simple linear effects.

2.6 Key Insights Summary

1. Departure hour is the most important single predictor, with a clear delay propagation pattern throughout the day.
2. Strong interaction effects exist between hour and day of week, requiring non-additive models.
3. Weather effects are real but non-linear and threshold-based, justifying their inclusion despite weak correlations.
4. Carrier and airport heterogeneity requires categorical encoding of these features.
5. The absence of strong linear correlations motivates tree-based modeling approaches that can capture complex interactions.

3. Modeling Approach

3.1 Problem Formulation

The prediction task was formulated as binary classification: predict whether a flight will experience a departure delay of 15 minutes or more ($\text{DepDel15} = 1$) based on pre-departure information only.

Evaluation Metrics:

Given the class imbalance and operational context, the following metrics were prioritized:

- **ROC-AUC:** Measures overall discrimination ability across all classification thresholds. Primary metric for model comparison.
- **PR-AUC (Precision-Recall AUC):** More informative than ROC-AUC for imbalanced datasets, focuses on performance on the minority class.
- **Recall:** Critical for operations planning - catching delayed flights is more important than avoiding false alarms.
- **Precision:** Important to avoid excessive false alarms that could erode trust in predictions.
- **F1-Score:** Harmonic mean balancing precision and recall.

3.2 Data Split Strategy

A stratified random split (80% train, 20% test) was used for model development. While temporal splits would be ideal for production deployment, the three-month window was too short for effective chronological splitting. Stratification ensured balanced class distribution in both sets.

3.3 Feature Engineering

Based on EDA findings and domain knowledge, features were curated to avoid leakage while maximizing predictive signal:

Features Removed (Leakage Risk):

- Cancelled, Diverted: Known only after departure

Features Removed (Redundancy):

- CRSDepTime: Redundant with dep_hour
- Distance: Redundant with DistanceGroup
- FeelsLikeC: Derived from tempC
- Year: Constant (zero variance)

Features Removed (Model-Specific - Logistic Regression only):

- DepTimeBlk: Categorical encoding of dep_hour, poorly suited for linear models
- winddirDegree: Circular variable requiring special treatment
- weatherCode: Integer-encoded categorical, misleading for linear models

Final Feature Set:

- Numeric (13): dep_hour, Month, DayOfWeek, DayOfMonth, DistanceGroup, tempC, windspeedKmph, windgustKmph, humidity, pressure, visibility, precipMM, cloudcover
- Categorical (4): UniqueCarrier, Origin, Dest, weatherCode

3.4 Model Selection and Results

Three modeling approaches were evaluated, progressing from simple baselines to more sophisticated ensemble methods:

Model Performance Comparison:

Model	ROC-AUC	PR-AUC	Precision	Recall	F1-Score
Logistic Regression	0.679	0.321	0.423	0.221	0.291
Random Forest (Baseline)	0.729	0.434	0.437	0.467	0.451
Random Forest (Tuned)	0.730	0.436	0.443	0.461	0.452
Gradient Boosting (Tuned)	0.730	0.440	0.461	0.404	0.431
XGBoost (Final)	0.735	0.448	0.403	0.543	0.463

3.4.1 Logistic Regression (Baseline):

Logistic Regression served as the interpretable baseline. Performance was limited (ROC-AUC = 0.679) due to the inability to capture interaction effects and non-linear weather relationships. Low recall (0.221) indicated that the linear decision boundary missed many delayed flights. However, it established that predictive signal exists in the feature set.

3.4.2 Random Forest:

Random Forest showed substantial improvement (ROC-AUC = 0.729), validating the hypothesis that delays are driven by feature interactions. The ensemble of decision trees naturally captured hour × day interactions and non-linear weather thresholds. Tuning (max_depth=15, min_samples_leaf=100, n_estimators=200) provided marginal gains while preventing overfitting.

3.4.3 Gradient Boosting:

Gradient Boosting achieved similar ROC-AUC (0.730) but with different precision-recall trade-offs. It showed higher precision (0.461) but lower recall (0.404) compared to Random Forest, making it suitable for applications where false alarm costs are high.

3.4.4 XGBoost (Final Model):

XGBoost was selected as the final model due to its superior recall (0.543) and overall balance of metrics (ROC-AUC = 0.735, PR-AUC = 0.448). The higher recall is operationally valuable - catching 54% of delays with 40% precision provides actionable early warning for operations teams. The model was configured with: learning_rate=0.1, max_depth=4, min_child_weight=5, n_estimators=300, subsample=0.8.

3.5 Model Interpretation

Feature importance analysis (using SHAP values for tree-based models) revealed the relative contribution of each feature to predictions:

Top Predictive Features:

- 1. **dep_hour:** Most important feature, capturing the delay propagation effect throughout the day
- 2. **Origin airport:** Captures airport-specific operational complexity and congestion

3. **DayOfWeek:** Reflects weekly travel patterns and their interaction with time of day
4. **UniqueCarrier:** Accounts for carrier-specific operational practices
5. **Weather features (windgustKmph, precipMM):** Contribute conditional delay risk

4. Limitations and Future Work

4.1 Current Limitations

4.1.1 Temporal Coverage:

The three-month dataset (January-March 2016) limits generalization in several ways:

- Seasonal variation is not captured (missing summer weather, holiday travel patterns)
- Long-term trends and year-over-year changes cannot be assessed
- Model performance on future data is uncertain due to operational changes since 2016

4.1.2 Geographic Coverage:

Analysis was restricted to 10 airports with available weather data, excluding:

- Smaller regional airports with different operational characteristics
- International destinations where weather impacts may differ
- Hub-spoke network effects for connections beyond these airports

4.1.3 Feature Limitations:

Several potentially valuable features were unavailable or excluded:

- Upstream delay propagation (delays from previous flight legs on same aircraft)
- Airline scheduling practices (planned turn times, crew scheduling)
- Airport-specific factors (runway configuration, air traffic control delays)
- Severe weather events (thunderstorms, snowstorms require more than hourly aggregates)

4.1.4 Model Evaluation Constraints:

- Random rather than temporal split may overestimate generalization to future dates
- No external validation on completely held-out time periods
- Class imbalance limits precision, which may be problematic for some operational uses

4.2 Recommended Improvements

4.2.1 Data Enhancements:

- **Extend temporal coverage:** Include full year or multiple years to capture seasonal patterns
- **Add upstream delay features:** Track aircraft rotation to account for delay propagation
- **Incorporate severe weather indicators:** Add METAR reports, weather alerts, and radar data
- **Include air traffic flow data:** Ground delay programs, arrival rate restrictions
- **Add airport congestion metrics:** Scheduled operations per hour, runway utilization

4.2.2 Modeling Enhancements:

- **Implement proper temporal validation:** Use sliding window or blocked time splits
- **Address class imbalance:** Explore SMOTE, class weights, or threshold optimization
- **Develop ensemble methods:** Stack multiple models to leverage diverse prediction strategies
- **Add calibration:** Ensure predicted probabilities reflect true delay likelihood
- **Implement online learning:** Update model as new data arrives to adapt to changing patterns

4.2.3 Operational Integration:

- **Build confidence intervals:** Provide uncertainty estimates for predictions
- **Develop decision support tools:** Translate predictions into actionable recommendations
- **Create early warning system:** Alert operations teams when delay risk exceeds thresholds
- **Implement A/B testing:** Validate model impact on operational efficiency

4.3 Additional Data Sources

High-Value Data to Incorporate:

1. Aircraft-Level Information:

- Tail number tracking to identify specific aircraft history
- Aircraft age and maintenance records
- Previous flight delays on same aircraft (delay propagation)

2. Airport-Specific Factors:

- Real-time runway configuration and capacity
- Ground delay programs and air traffic management initiatives
- Airport construction or special events affecting operations

3. Enhanced Weather Data:

- Terminal Aerodrome Forecasts (TAF) for forward-looking predictions
- METAR observations at higher frequency (current hourly is too coarse)
- Weather radar and satellite imagery for severe weather detection
- Wind shear, turbulence, and ceiling/visibility trends

4. Network-Level Features:

- System-wide delay statistics (network congestion indicators)
- Airline schedule density at departure and arrival airports
- Connecting flight pressures and passenger flow patterns

5. Temporal Features:

- Holiday and special event indicators
- School vacation periods affecting travel demand
- Historical delay patterns for specific route/time/day combinations

5. Conclusions

This analysis demonstrates that flight departure delays can be predicted with meaningful accuracy using pre-departure information combining operational schedules and weather conditions.

Key Achievements:

- Successfully integrated flight and weather data across 10 major airports for 105,783 flights
- Identified time-of-day and day-of-week interactions as the primary delay drivers
- Validated that weather contributes non-linearly to delay risk
- Achieved ROC-AUC of 0.735 and recall of 0.543 with XGBoost model
- Established that tree-based models substantially outperform linear approaches for this task

Practical Implications:

The final model correctly identifies 54% of delayed flights with 40% precision. While precision could be higher, this recall level provides valuable early warning for operations planning, allowing airlines to:

- Proactively notify passengers of potential delays
- Adjust staffing and gate assignments
- Optimize aircraft rotations to minimize delay propagation
- Improve customer service by setting realistic expectations

Methodological Insights:

The analysis reinforces that delay prediction is fundamentally an interaction-driven problem. Linear models fail because delay risk is not additive - the effect of weather depends on time of day, the impact of hour depends on day of week, and carrier performance varies by airport. Tree-based models that naturally capture these interactions prove essential.

Path Forward:

The model provides a strong foundation but requires enhancement for production deployment. Priority improvements include temporal validation, upstream delay features, and calibration. With additional data sources and operational integration, this approach could deliver significant value for airline operations management and customer experience improvement.

Appendix: Technical Details

Software and Libraries:

- Python 3.13.5
- Data Processing: pandas, numpy
- Modeling: scikit-learn, xgboost
- Visualization: matplotlib, seaborn, missingno
- Model Interpretation: shap

Computational Environment:

- Data processing performed on standard laptop hardware
- No GPU acceleration required
- Training time: ~10-15 minutes for XGBoost with grid search

Reproducibility:

All analysis code organized in Jupyter notebooks:

1. 00_data_directory.ipynb - Data organization
2. 01_flight_data_preprocessing.ipynb - Flight data cleaning

3. 02_weather_data_preprocessing.ipynb - Weather data processing
4. 03_merge_flight_weather_data.ipynb - Data integration
5. 04_exploratory_data_analysis.ipynb - EDA and visualization
6. 05_baseline_modeling_data.ipynb - Baseline model development
7. 06_final_modeling_data.ipynb - Advanced modeling and tuning