Mining Patterns from Large-Scale Flight Data

Feature Engineering and Classification for Delay Prediction at ATL

The Data Challenge

Big Data in Aviation

Modern aviation generates massive operational data volumes. Flight schedules, weather, sensors, air traffic control—hidden patterns require systematic mining.

Scale of the Problem

4+ million flight records annually with 109 attributes per flight. Complex interdependencies, missing data, and need for scalable workflows.

Project Objectives

Knowledge Discovery Goals

- Extract actionable insights from historical flight data
- Develop predictive models for operational decisions
- Build systematic preprocessing pipeline
- Validate on world's busiest airport

Success Criteria

- High prediction accuracy on test data
- Balanced performance across delay types
- Scalable to real-world deployment
- Interpretable and explainable results

Data Source & Scope

4M

Flight Records

US Bureau of Transportation Statistics 2022 data covering domestic flights nationwide

109

Features

Attributes covering operations, timing, airlines, and airports

184K

ATL Flights

Focus on world's busiest airport with high traffic volume and diverse airline mix

Initial Data Quality Assessment

Duplicate Records

12,847 duplicate records identified requiring removal to prevent training bias

Missing Values

23 features with 100% missing values and varying levels across other features

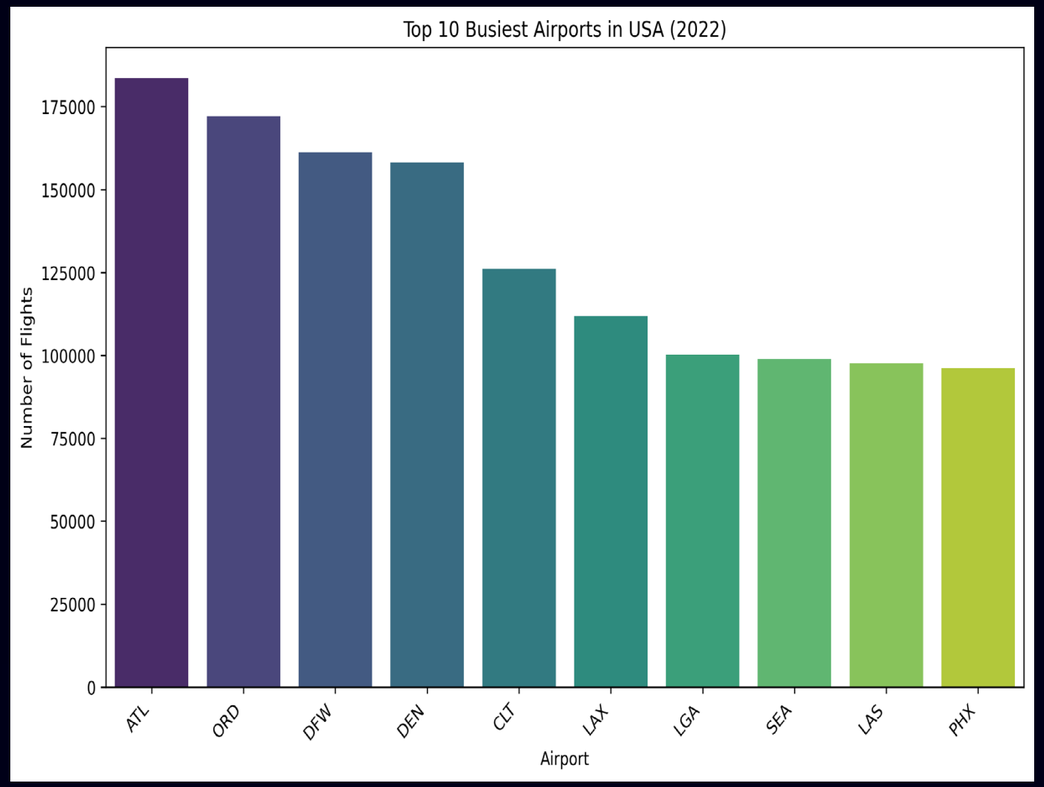
Format Issues

Inconsistent time formats (HHMM encoding) requiring standardization

Class Imbalance

80% on-time vs 20% delayed flights affecting model training

Data quality determines model quality—comprehensive cleaning pipeline required before applying algorithms.



Traffic Distribution Analysis

Top 10 airports handle 28% of all flights, with ATL leading at 183K flights. Significant concentration in major hubs with long tail of smaller airports.

Systematic Cleaning Steps

01

Remove Duplicates

Eliminated 12,847 duplicate records to prevent training bias and ensure data integrity

02

Drop Null Features

Removed 23 features with 100% missing values that provide no information

03

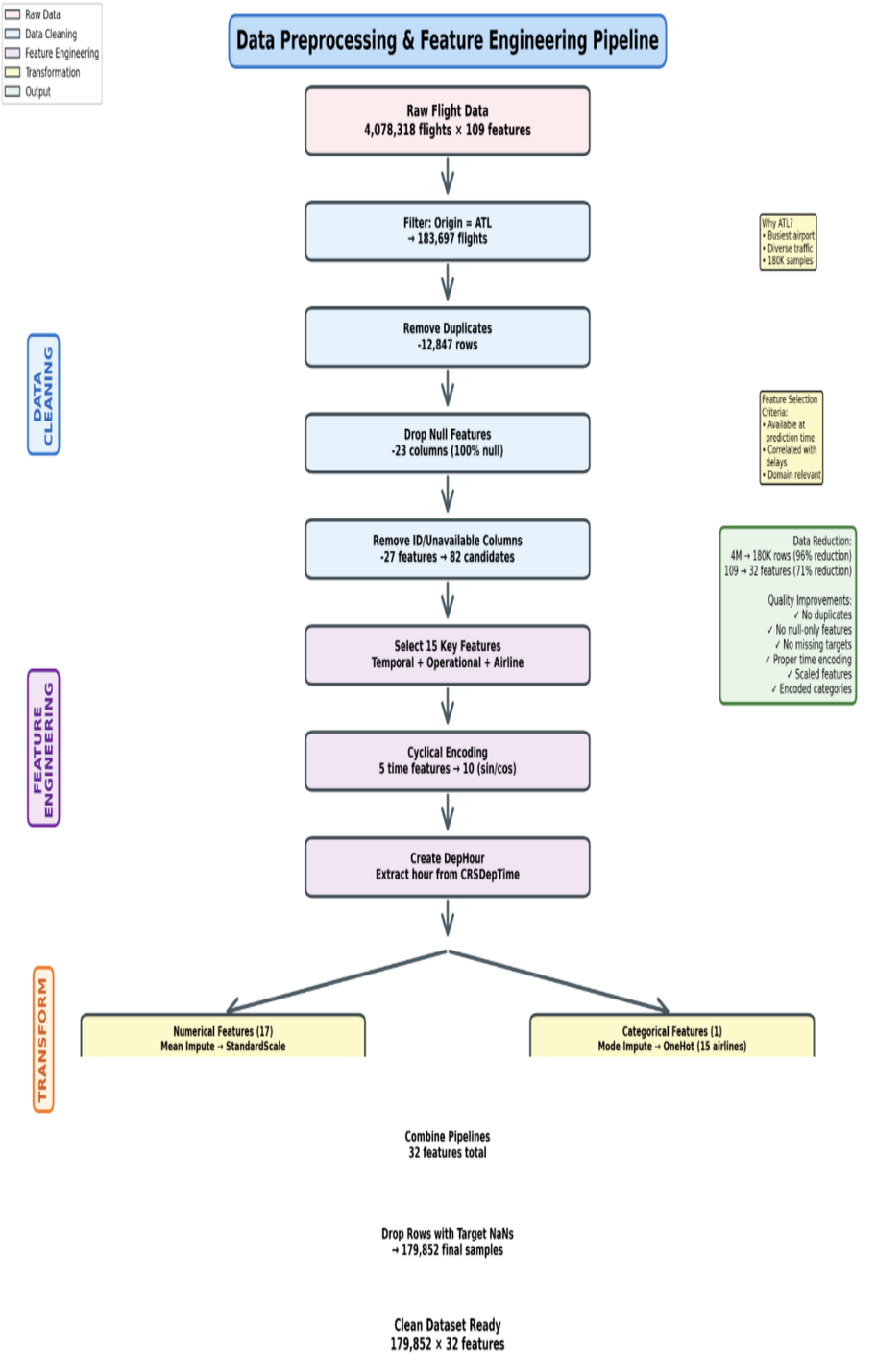
Eliminate ID Fields

Identified and removed 27 ID/unavailable features that cause overfitting

04

Feature Reduction

Reduced from 109 to 82 candidate features for further analysis



Feature Selection Process

Selection Criteria

- Available at prediction time
- Relevant to delays via domain knowledge
- Correlated with target variable
- Avoid unique identifiers

From 82 to 15 Features

Temporal: Month, DayofMonth, DepHour

Operational: TaxiIn, TaxiOut, AirTime, Distance

Scheduled/Actual Times: 5 time features

Delay Indicator: DepDel15

Categorical: Airline

Missing Value Handling

1

Missing Data Patterns

Different patterns across feature types requiring strategic approach

2

Imputation Strategy

Numerical: mean imputation. Categorical: most frequent. Target: row deletion

3

Two-Stage Approach

Select features first, then impute to preserve statistical properties

The Temporal Feature Challenge

Problem with Time Data

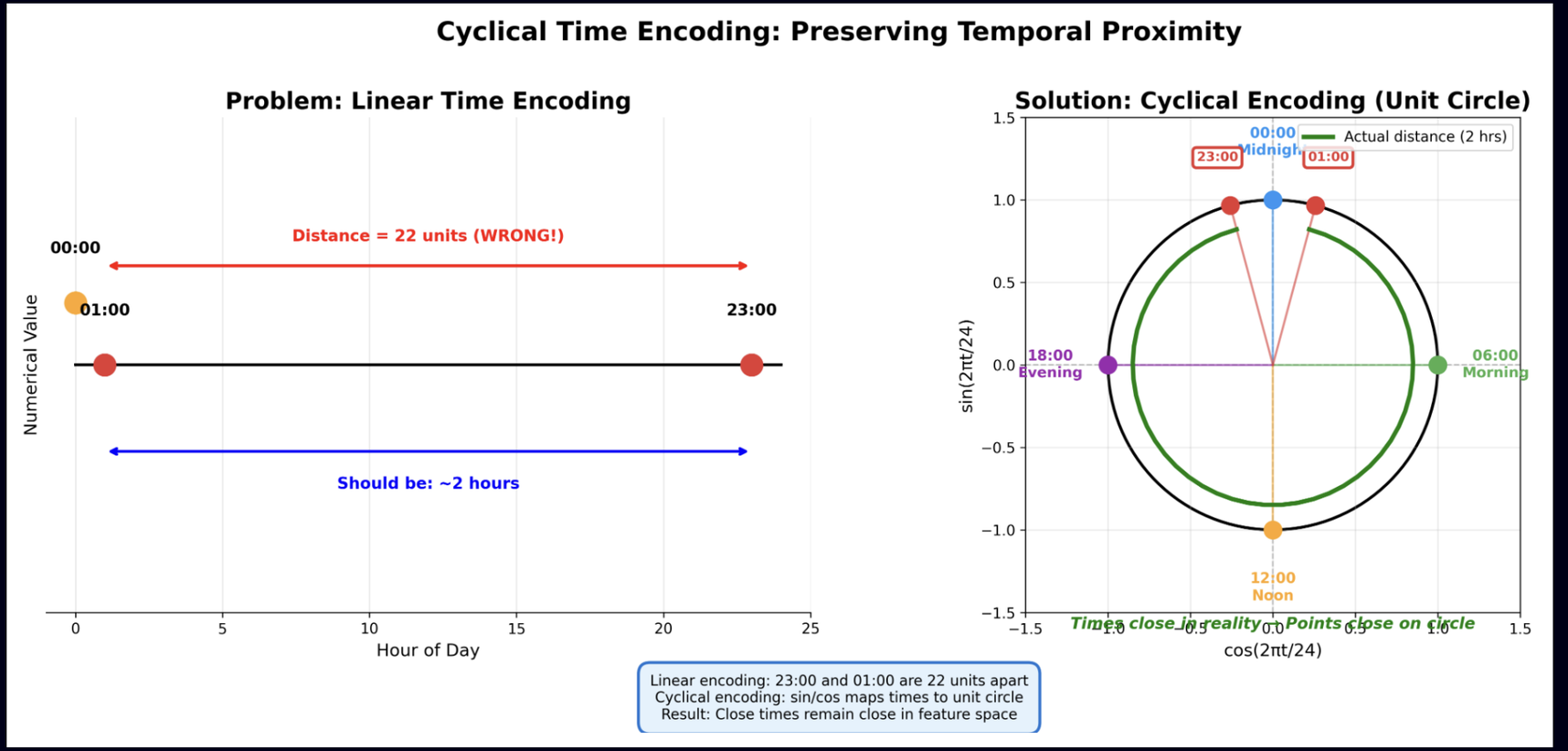
Time encoded as integers (HHMM format). 23:59 and 00:01 are numerically distant (2358 units) but temporally adjacent.

Linear distance doesn't reflect temporal proximity—models learn incorrect patterns.

Real-World Impact

- Cannot capture late night/early morning similarities
- Misses circadian patterns in operations
- Artificially breaks temporal continuity
- Reduces prediction accuracy

Cyclical Encoding Solution



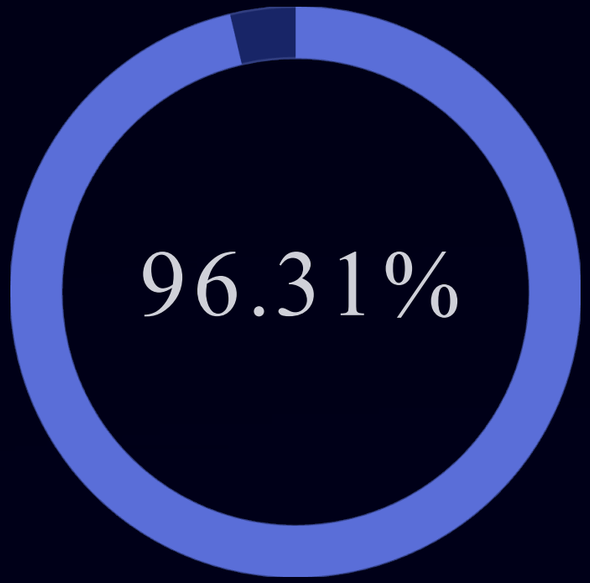
Geometric Interpretation

Maps 24-hour cycle onto unit circle. Midnight at (0,1), noon at (0,-1). Adjacent times map to nearby points.

Convert HHMM to minutes: m = (t ÷ 100) × 60 + (t mod 100)

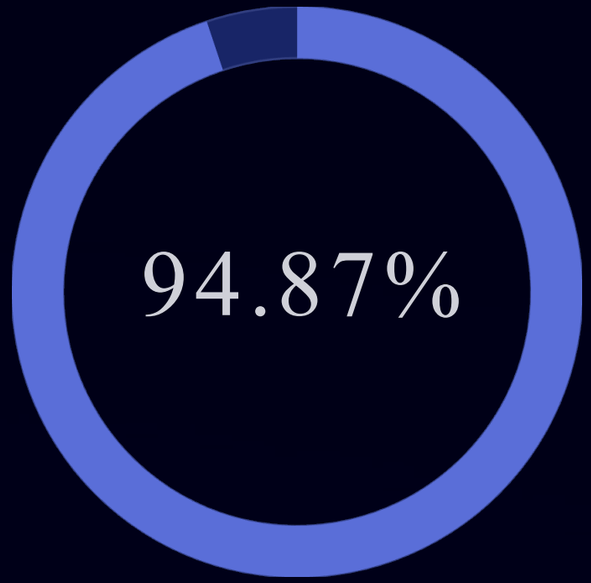
Apply sin(2πm/1440) and cos(2πm/1440) creating two features per time dimension

Impact Quantification



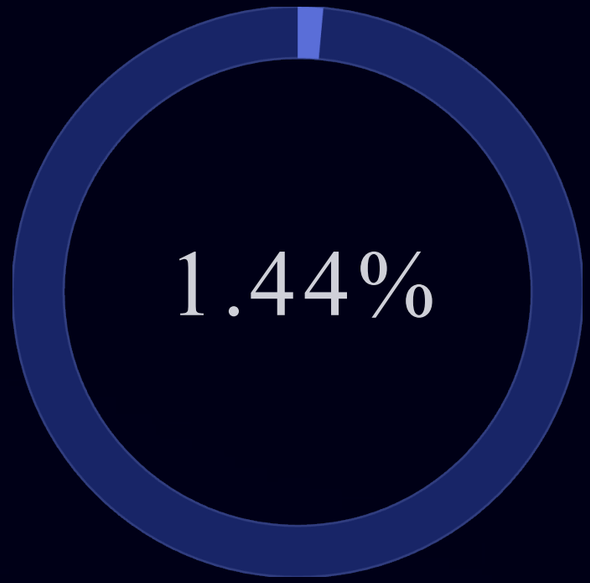
With Cyclical Encoding

Accuracy achieved using proper temporal representation



Without Cyclical Encoding

Baseline accuracy with raw HHMM format

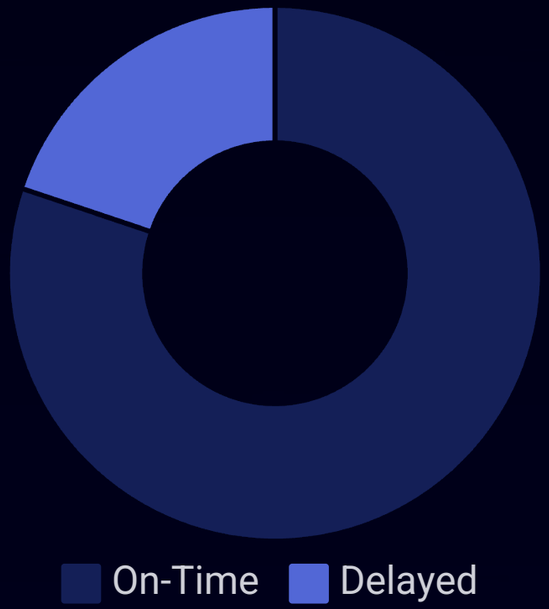


Improvement

~520 additional correct predictions on test set—largest single preprocessing contribution

Validates feature engineering effort and demonstrates importance of domain-appropriate encoding.

Class Imbalance Characteristics



Distribution Analysis

On-time flights: 144,410 (80.3%)

Delayed flights: 35,442 (19.7%)

Ratio: approximately 4:1

Implications

Classifiers may bias toward majority class. Standard accuracy can be misleading—need class-specific metrics.

Handling Strategies Considered

Oversampling

SMOTE, ADASYN to increase minority class representation

Undersampling

Reduce majority class to balance distribution

Cost-Sensitive Learning

Apply class weights to penalize errors differently

Natural Distribution

Our choice: Keep natural distribution, use robust algorithms and class-specific metrics

Results Validation

97%	92%	0.94	0.96
On-Time Precision	Delayed Precision	Macro F1	Weighted F1
98% recall for on-time flights	89% recall for delayed flights	Average across classes	Accounting for class size

Key Takeaway: Random Forest's ensemble structure naturally handles 4:1 imbalance. No explicit resampling needed—multiple trees capture minority patterns effectively.

Algorithm Selection Rationale

Random Forest

Ensemble method with non-linear capabilities, robust to noise and outliers

Logistic Regression

Linear model providing interpretable results with fast training

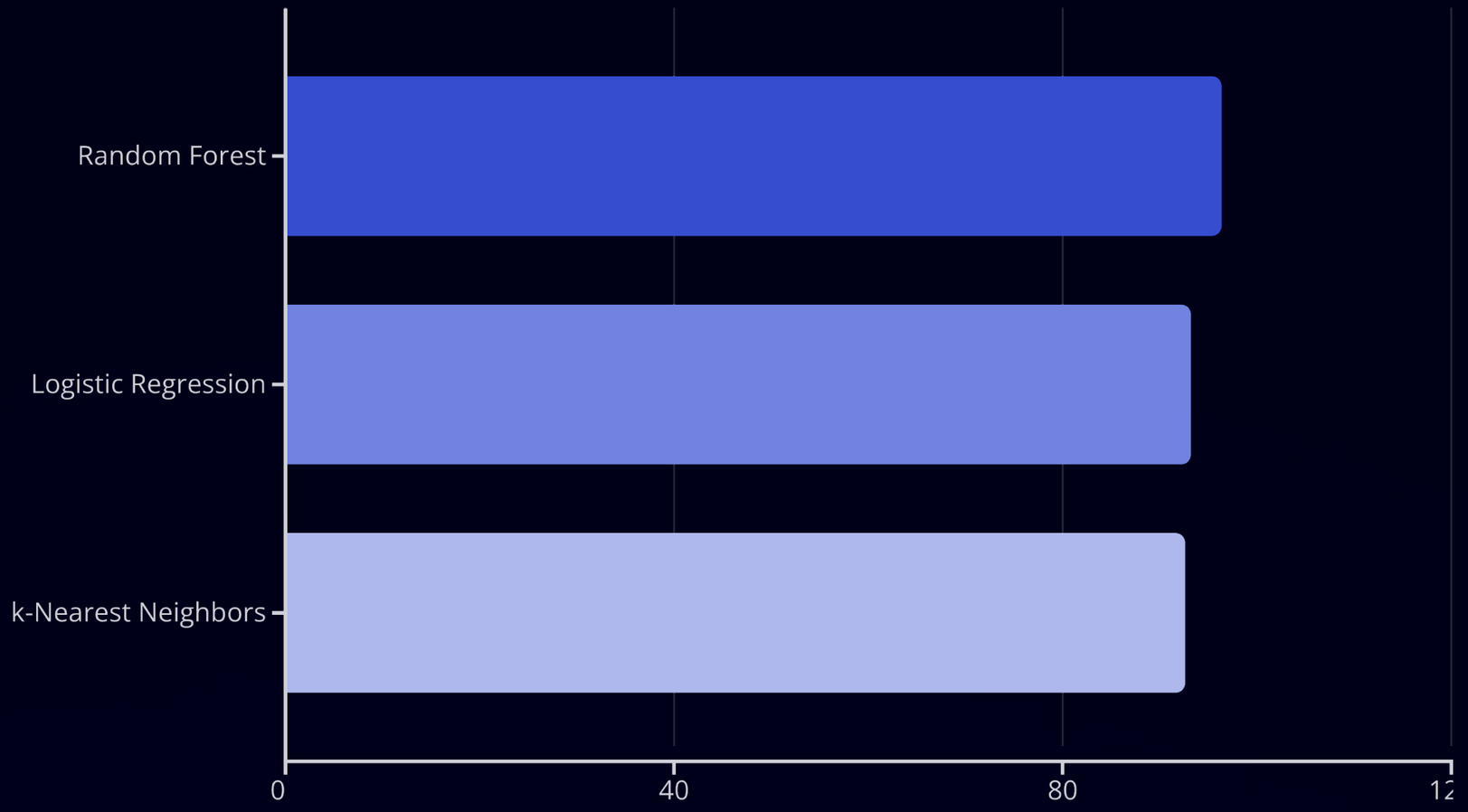
k-Nearest Neighbors

Non-parametric, instance-based approach with simple implementation

Train all on identical preprocessed data, evaluate on same test set with multiple metrics to identify best performer.

Performance Results

Clear winner: Random Forest with 3+ point advantage over alternatives



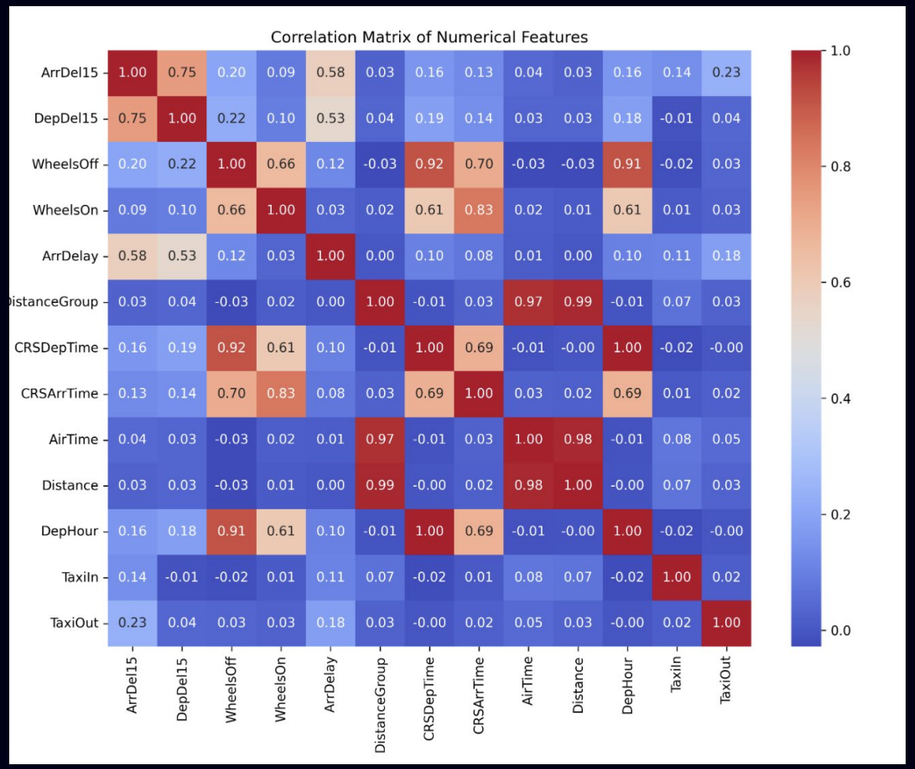
3% difference equals 1,100 correct predictions on 36K test set. Consistent across multiple metrics and robust to different random seeds.

Temporal Patterns Discovered

1. **Early Morning (5-7 AM)**
   <5 min average delay
2. **Afternoon (12-4 PM)**
   5-10 min average delay
3. **Evening (6-10 PM)**
   15-25 min average delay

Explanation: Cascading delay effect throughout daily operations. Same aircraft/crew operate multiple flights—morning delays propagate through schedule with network effects compounding over time.

Feature Correlation Patterns



Strong Correlations

DepDel15 ↔ ArrDel15: 0.93 (departure delays predict arrival)

Distance ↔ DistanceGroup: 0.99 (redundant features)

- Moderate correlations among time features

Weak Correlations

Distance ↔ Delays: weak correlation

- Distance alone not predictive
- Other factors dominate delay patterns
- Validates multi-feature approach

Confusion Matrix Deep Dive

True Negatives

28,236 correctly predicted on-time flights (strong majority class performance)

True Positives

6,371 correctly predicted delays (good minority detection)

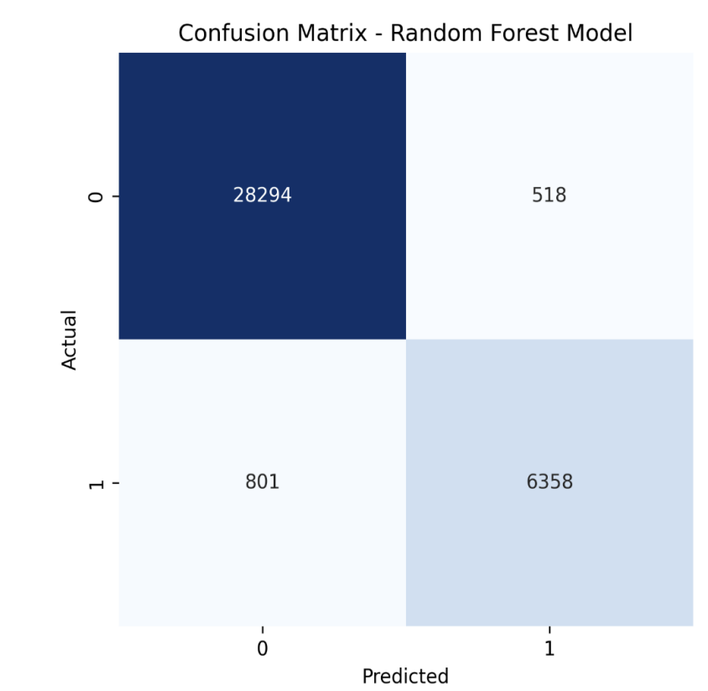
False Positives

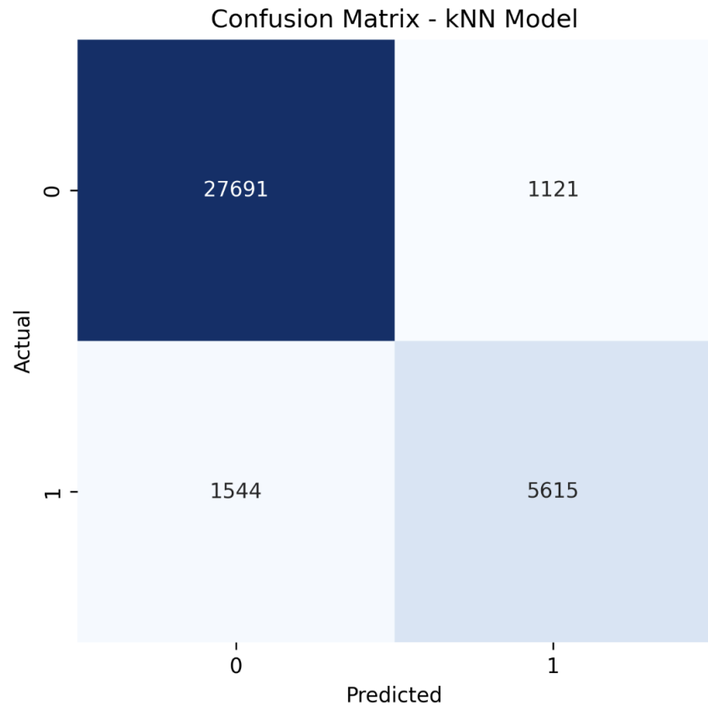
576 incorrect delay predictions (2% false alarm rate)

False Negatives

788 missed delays (11% miss rate)

Logistic Regression shows higher false negatives, kNN has more scattered errors. Random Forest achieves best balance overall.





Hyperparameter Search

01

RandomizedSearchCV

Efficient exploration of parameter space with 10 combinations tested

02

5-Fold Cross-Validation

Robust evaluation across different data splits

03

Parameter Space

Trees, depth, split criteria, leaf size explored systematically

04

Parallelized Execution

Completed in ~30 minutes with efficient computation

Optimization Results

Best Parameters Found

Optimal configuration improved baseline from 96.28% to 96.31% accuracy.

Small improvement validates good default parameters and confirms model robustness.

Insights from Tuning

- Default Random Forest parameters already strong
- Limited overfitting even with full depth
- Bootstrap sampling provides regularization
- Additional tuning could explore ensemble size

Computational Performance

180K	5min	15min	2sec
Dataset Size	Preprocessing Time	Training Time	Prediction Time
179,852 records × 32 features processed efficiently	Fast data cleaning and transformation	Random Forest model training duration	36K test samples—suitable for real-time deployment

Linear scaling with data size. Random Forest parallelizes across trees. Can handle monthly batches of millions of flights.

Deployment Architecture

Production Pipeline

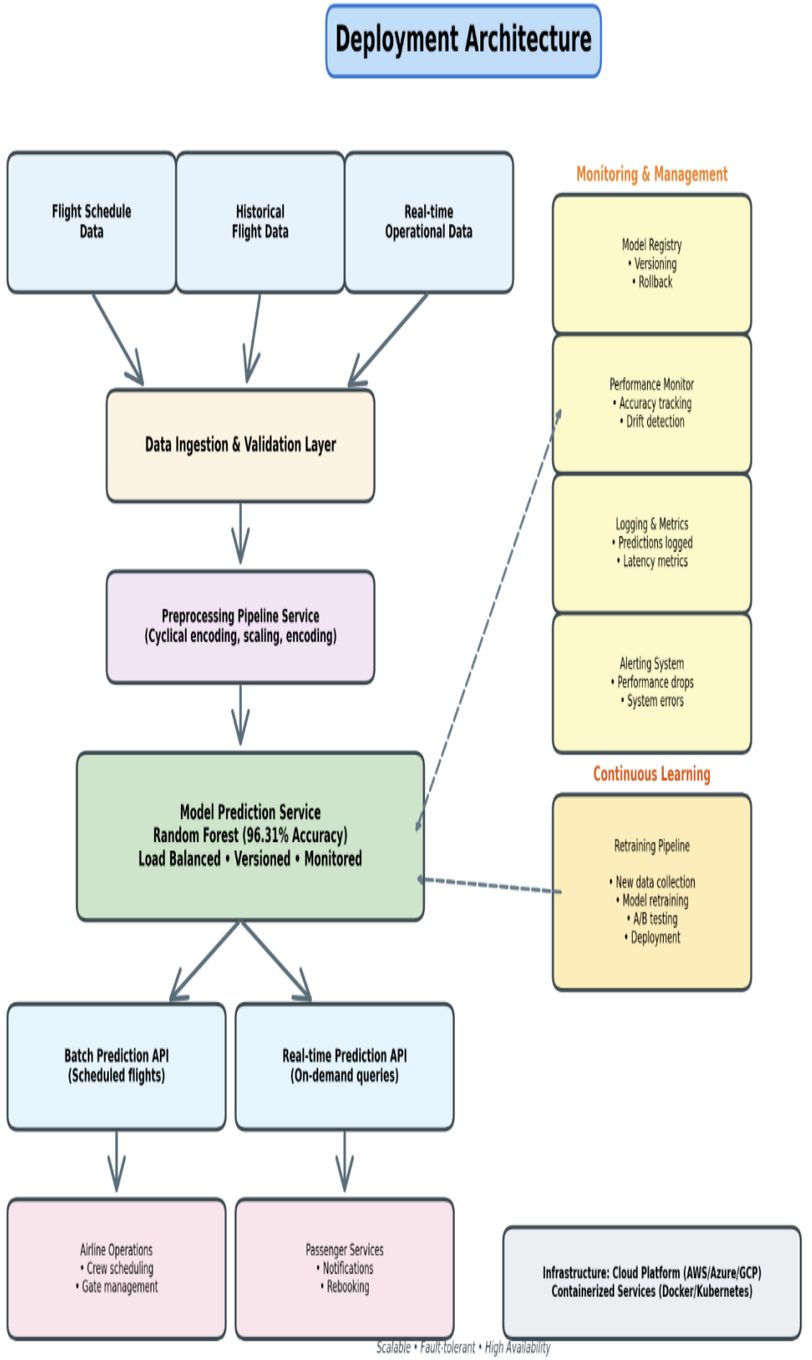
Batch prediction for scheduled flights with real-time API for on-demand queries

Model Management

Preprocessing pipeline serialization, model versioning, monitoring, and automated retraining

Integration Points

Airline operational systems, airport management platforms, passenger notifications, resource allocation



Current Limitations

Data Limitations

ATL-only training limits generalization. 2022 data only (single year). No external weather or air traffic control information. DepDel15 may not be available at prediction time.

Model Limitations

Binary classification loses magnitude information. 11% of delays still missed. Black-box nature limits interpretability. No explicit causality, only correlation.

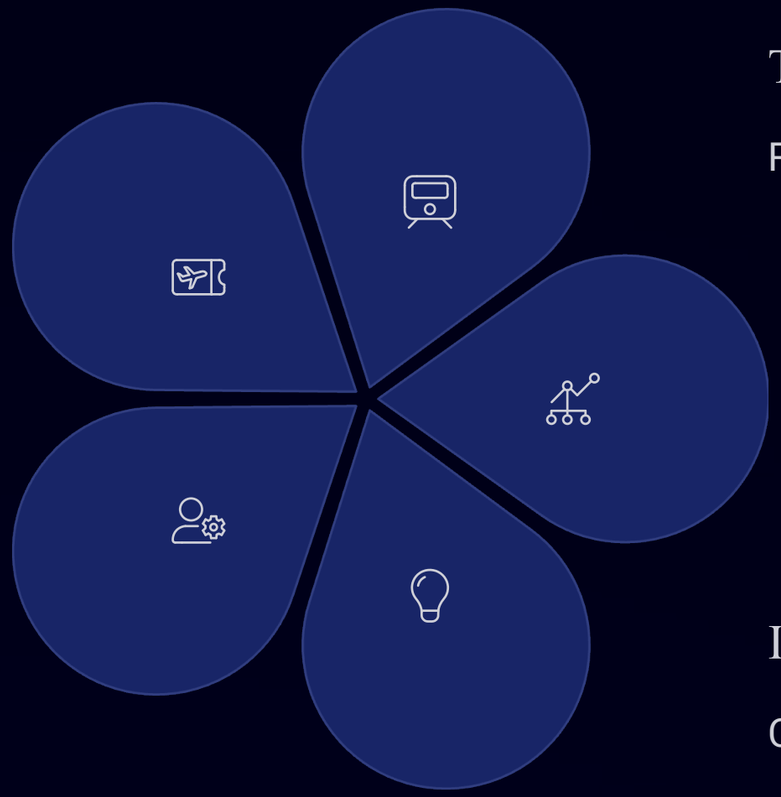
Broader Applications

Other Aviation Problems

Cancellations, diversions

Resource Optimization

Airport-wide allocation



Transportation Domains

Rail, bus systems

Operational Prediction

General framework

Industry Impact

Cost reduction, passenger experience

Thank You