DTM 2025 - Machine Learning

Flight Delayed Prediction

Machine Learning project

OPEN IN COLAB

Presented by: Alessandro De Faveri

Project Objective

The goal of this study is to predict a flight's Delay Risk, a classification target that flags whether arrival delay will be ≥ 15 minutes. The model is designed for post-departure decision-making, leveraging the actual departure delay (DEP_DELAY) together with schedule and route information.

To achieve this, the analysis integrates:

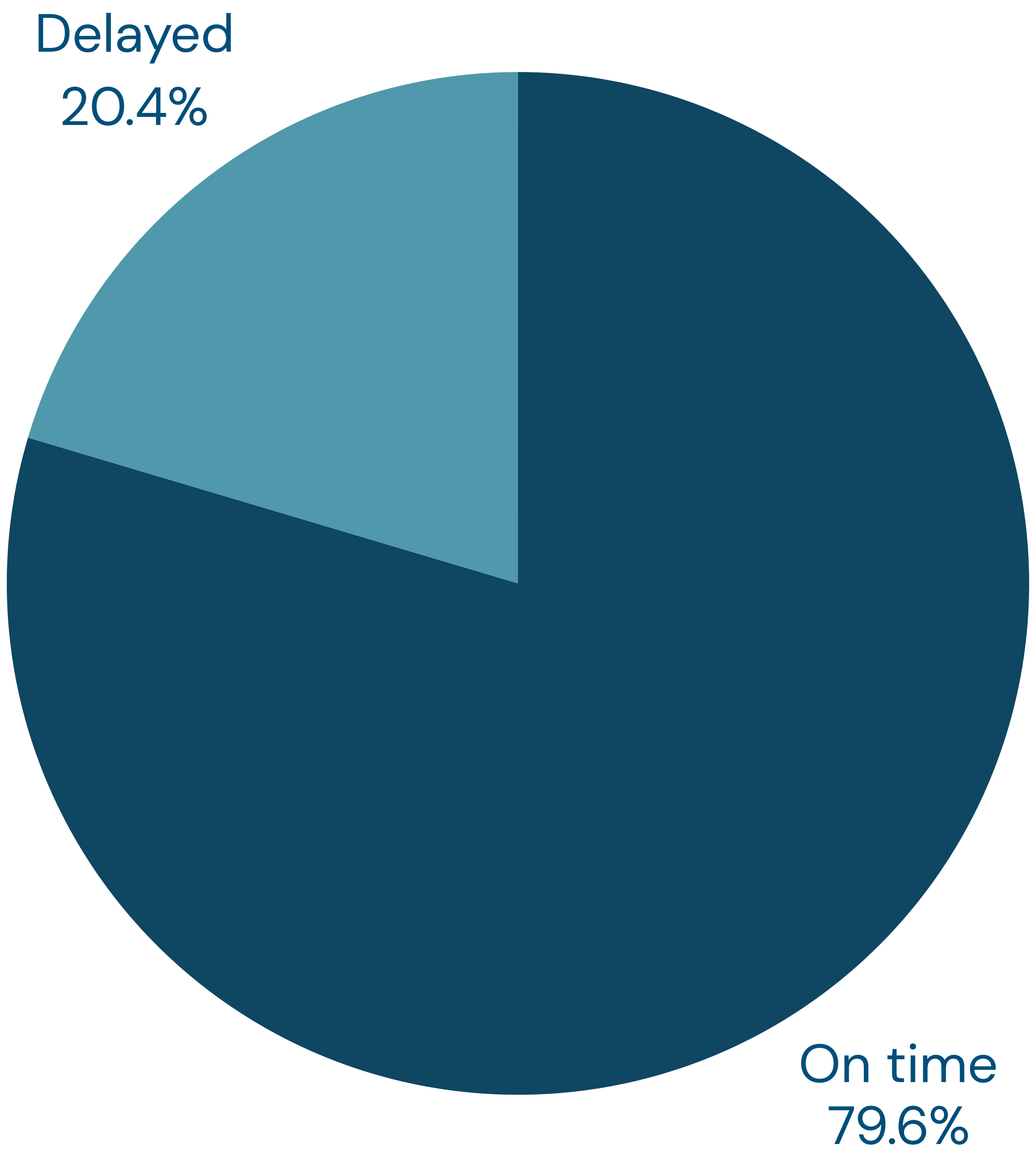
- Schedule & temporal signals such as departure hour, day of week, and seasonality.
- Route & operator context, including airline, origin/destination, and historical delay rates.
- Flight profile features such as distance, estimated duration, and short/long-haul flags
- Post-departure status via DEP_DELAY, which captures realized pushback lateness.

Data Set

Dataset: 687,860 flights from 2022 in USA
Share delayed: 20.4% (≈1 in 5 flights)
Missing ARR data: 2.9%
Missing DEP data: 2.6%

Basic delay statistics:

- Average arrival delay: 6.9 min
- Average departure delay: 12.5 min
- Average distance: 817 miles



Data Exploration

In this part, I identified patterns and correlations in flight delay data.

There is a strong correlation between departure and arrival delays, as well as seasonal patterns and airline-specific performance differences.

The most important are:

- **Figure 1**: The delay-rate by month shows a clear seasonal pattern: a gradual rise through spring, elevated levels in summer, a sharp trough in early autumn, and a renewed increase in December
- **Figure 2**: Reveals a strong monotone relationship: small departure delays frequently propagate, while substantial departure delays almost always lead to late arrival.

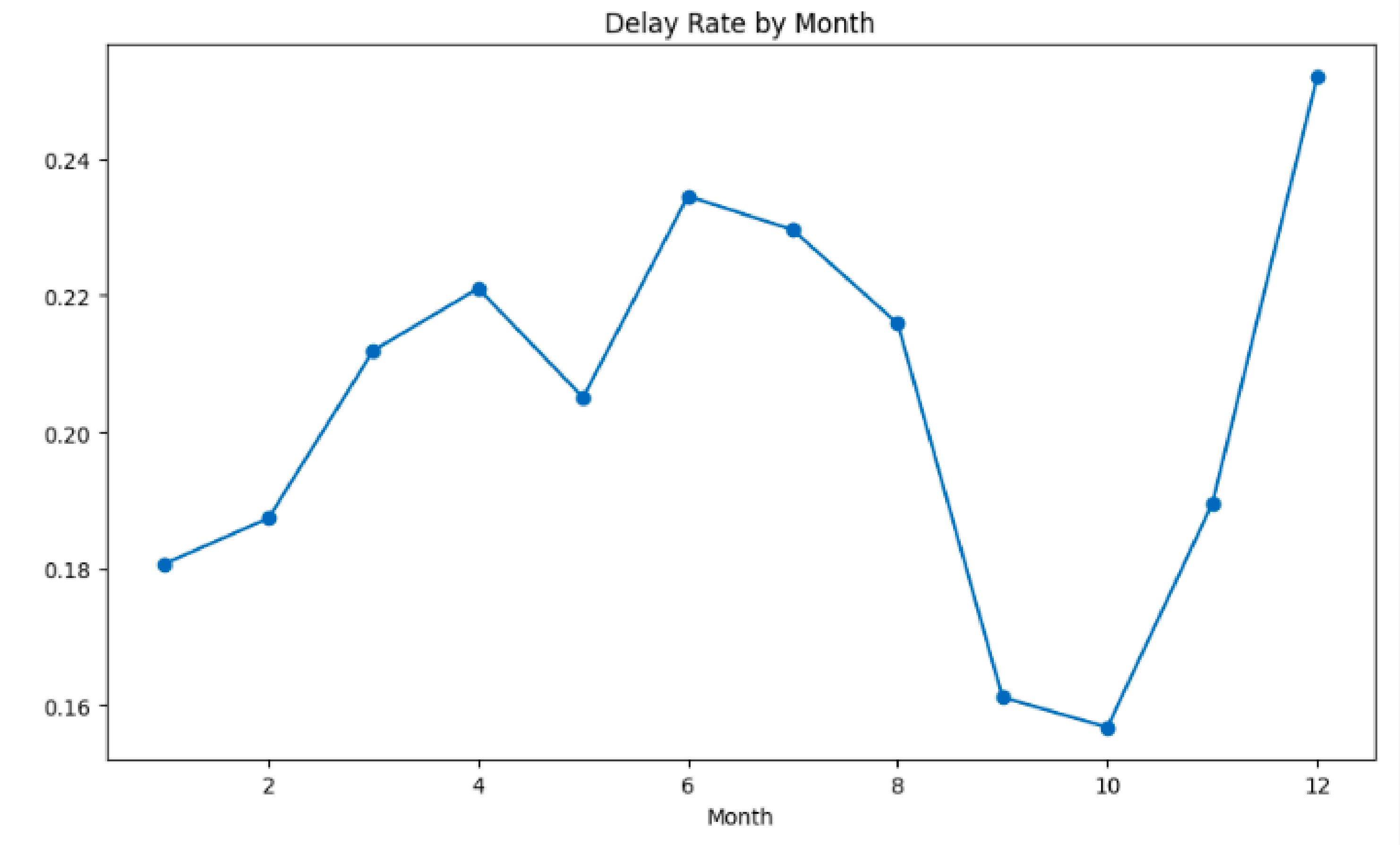


Figure 1

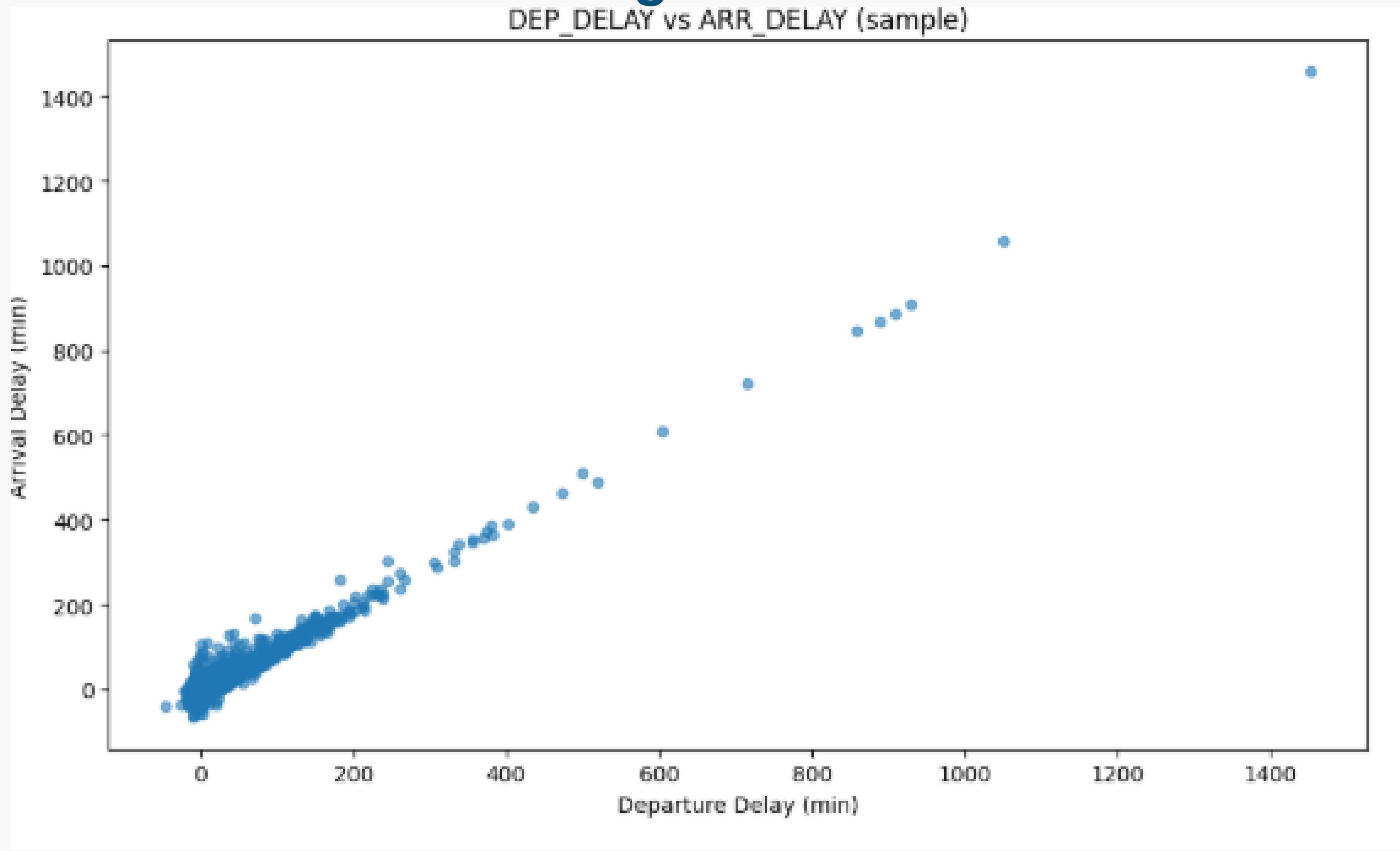
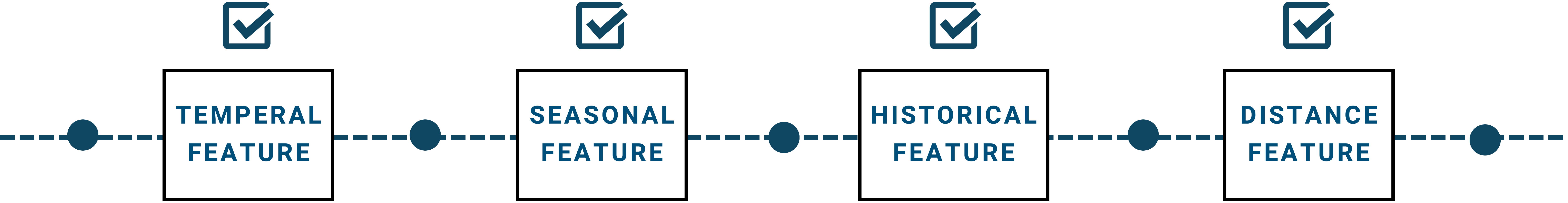


Figure 2

Feature Engineering



TEMPERAL FEATURE

- This allows the model to “understand” that a flight at 6:30 p.m. is more likely to be delayed than one at 6:00 a.m.

SEASONAL FEATURE

- The feature shows that flight patterns vary by season, with delays peaking in summer and winter, while spring and autumn tend to have more stable on-time performance

HISTORICAL FEATURE

- The feature calculates on only the training set. Example: if American Airlines historically has 25% delays, the model “remembers” this when it sees a new American flight.

DISTANCE FEATURE

- This feature analyzes the data based on the duration and the distance of the flight

Feature Engineering

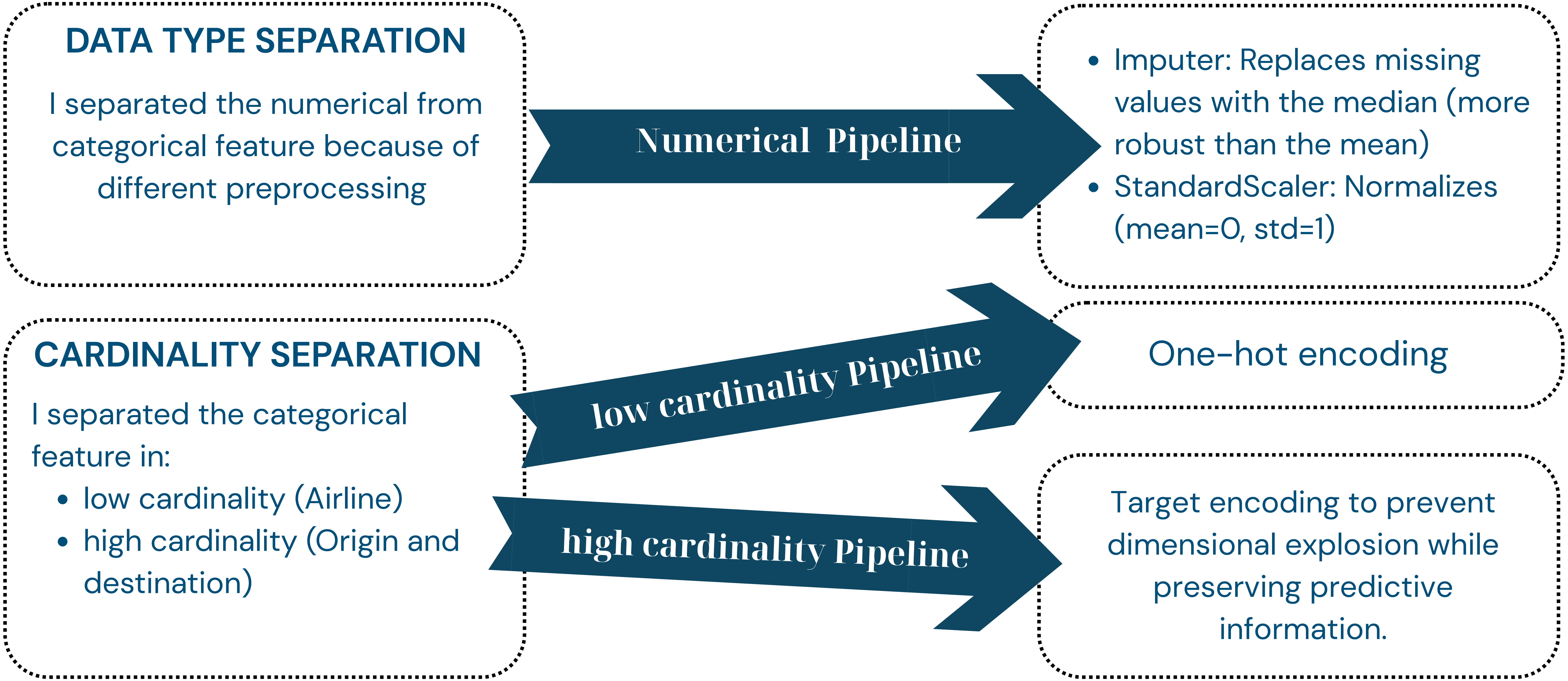
One crucial thing of my project is the split of the dataset.

WHY?

Problem: The model would see future data to predict the past! It's like knowing the results of the game while you're watching it.

Instead of random splits that create impossible future-past scenarios, we use temporal splitting where the model trains on historical data (Jan-Aug) to predict future flights (Sep-Dec), exactly like it would work in production."

Data Preparation



DATA TYPE SEPARATION

I separated the numerical from categorical feature because of different preprocessing

Numerical Pipeline

- Imputer: Replaces missing values with the median (more robust than the mean)
- StandardScaler: Normalizes (mean=0, std=1)

CARDINALITY SEPARATION

I separated the categorical feature in:

- low cardinality (Airline)
- high cardinality (Origin and destination)

low cardinality Pipeline

One-hot encoding

high cardinality Pipeline

Target encoding to prevent dimensional explosion while preserving predictive information.

Modeling Approach

PHASE 1: SETUP PIPELINE

- Automatic and consistent preprocessing
- No data leakage between training/validation/testing
- Complete reproducibility

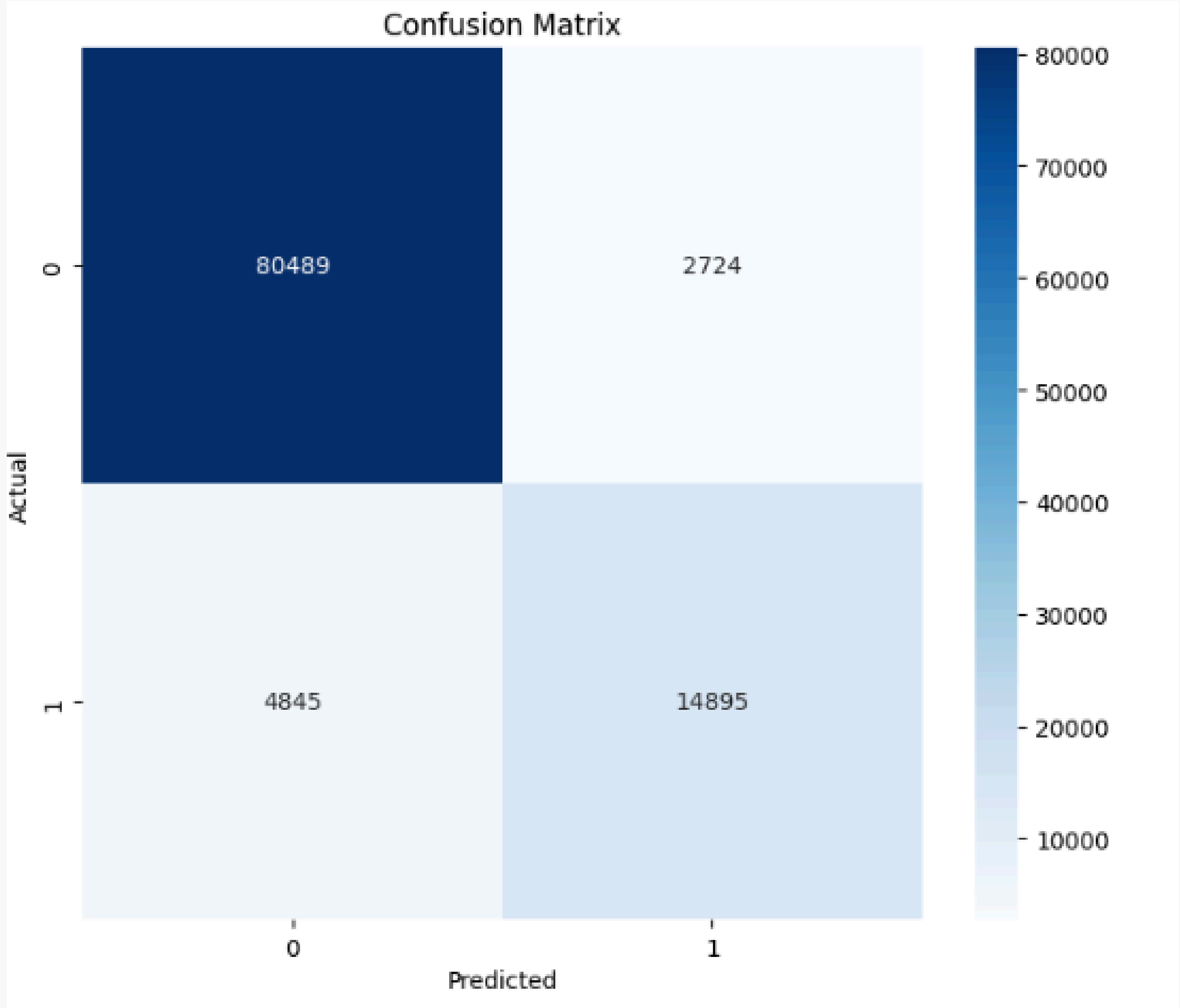
PHASE 2:HYPERPARAMETER OPTIMIZATION

1. Grid
    a. Conservative: Avoids overfitting with limited max_depth
    b. Balanced: Manages class imbalance
    c. Scalable: Range of manageable n_estimators
2. Why F-beta con β=1.5?
    a. β > 1: Favors recall (capturing more delays)
    b. Business logic: Better to predict false alarms than miss real delays
3. Temporal cross validation
4. Take a small sample before tuning
5. Randomized Search

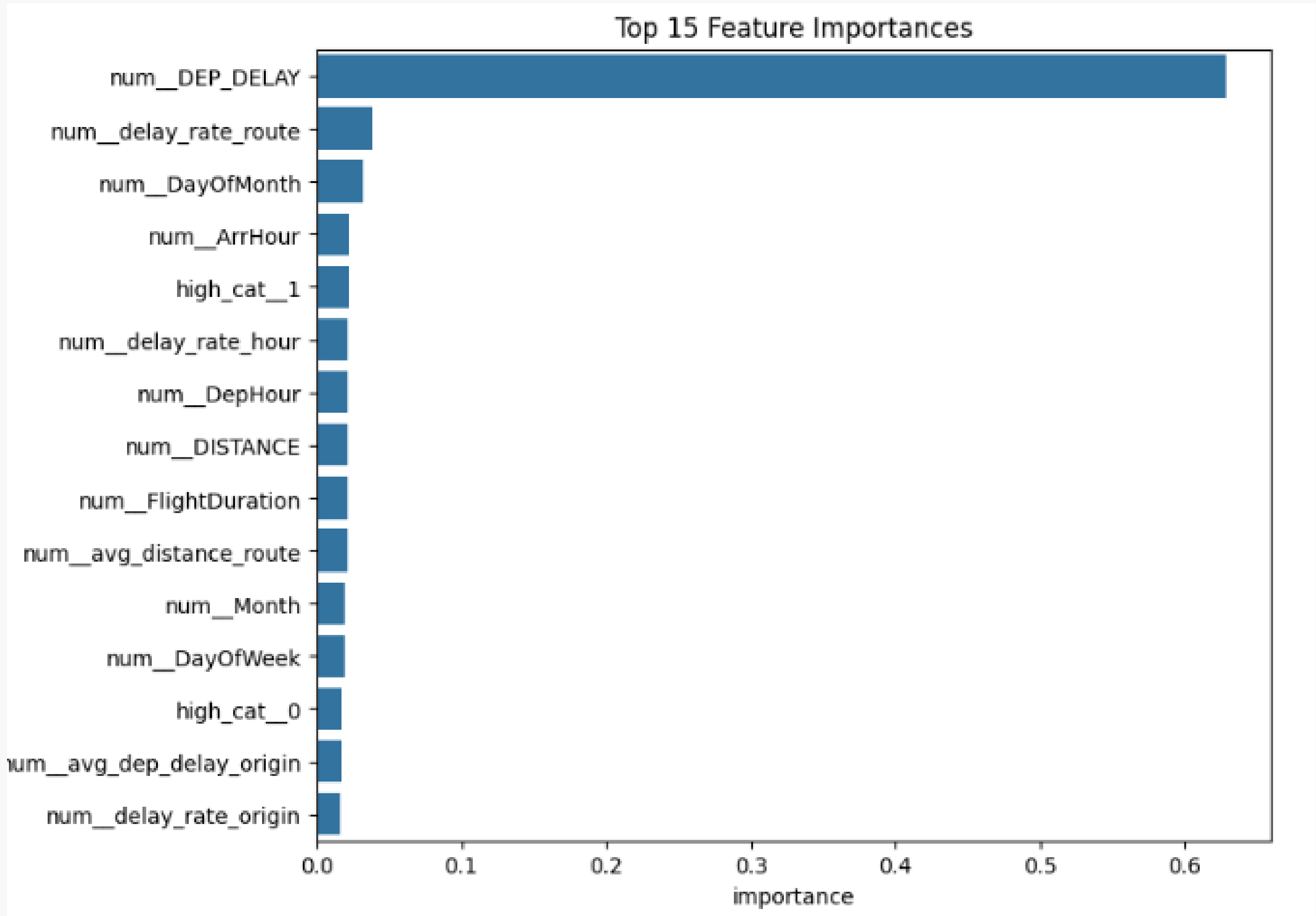
PHASE 3:FINAL TRAINING

- Training + validation

Evaluation on Test



Confusion matrix showing 92.6% accuracy with strong performance on both classes



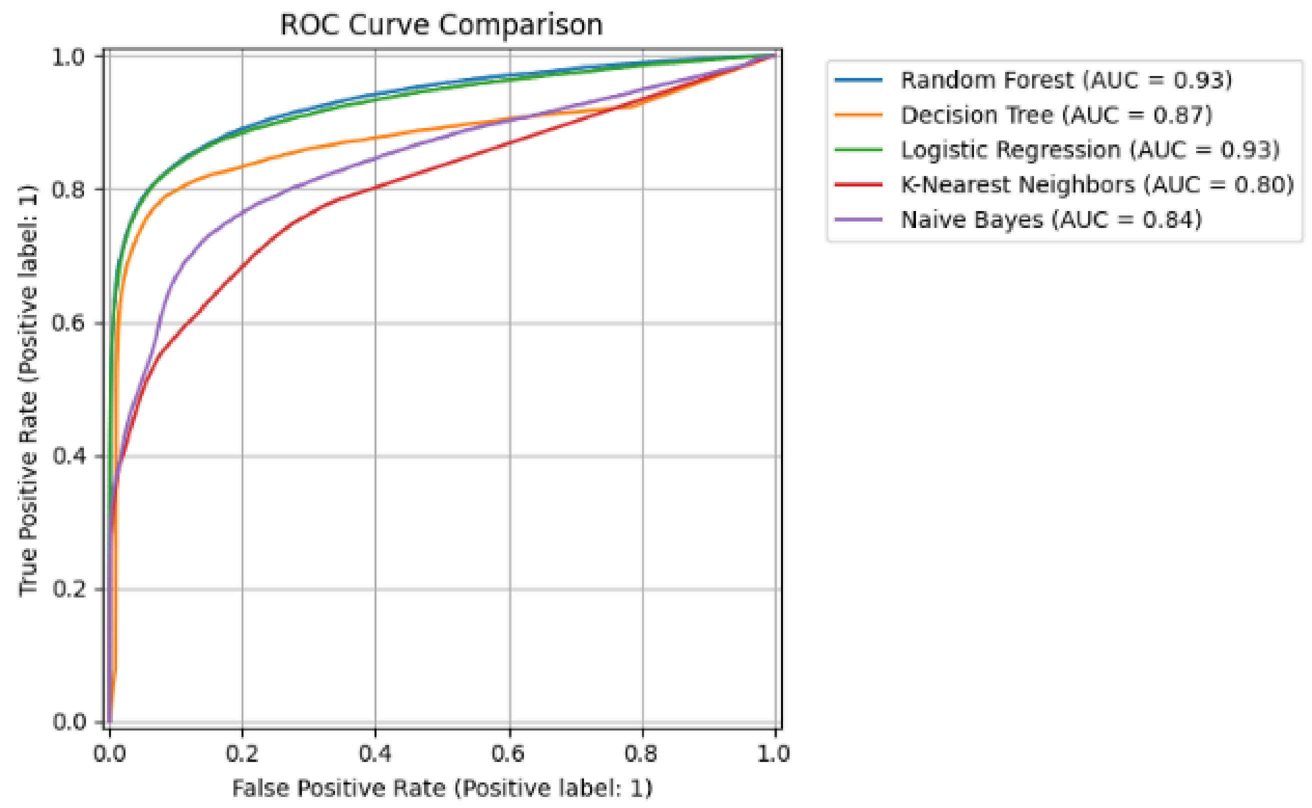
The feature importance analysis validates our domain knowledge, with departure delay emerging as the dominant predictor, alongside temporal and historical features.

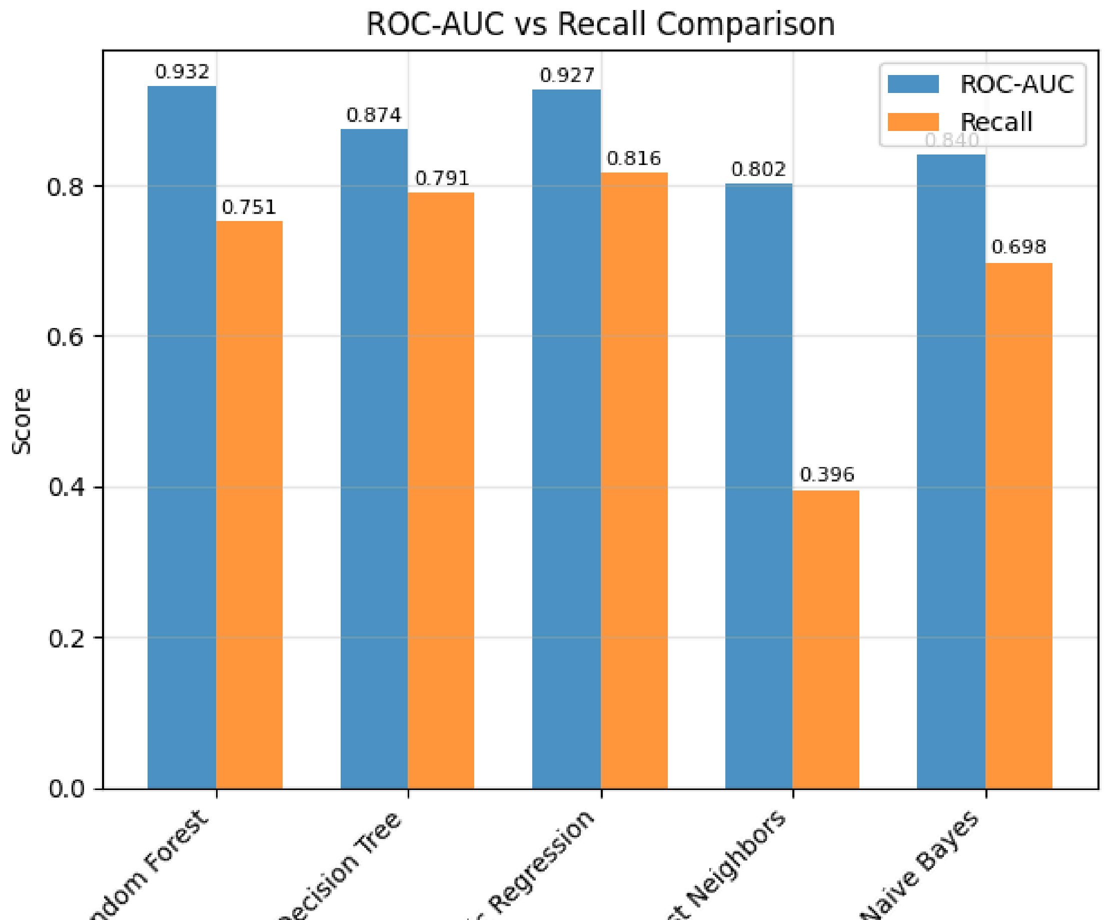
Model Comparison

I compared Random Forest with basic ML models and found that it performs the best in terms of ROC-AUC and F1 score.

Random Forest was not chosen at random, but as the optimal solution for the specific characteristics of the problem: natural handling of high-cardinality categorical features, robustness to outliers, business-friendly interpretability.

The overall benchmark confirms Random Forest as the optimal choice for general performance, while revealing strategic alternatives for specific scenarios.





Conclusion

Post-departure risk can be ranked reliably.

- Combining DEP_DELAY with schedule and route context gives a strong signal.

Proven performance on 2022 data.

- Random Forest: AUC 0.93, recall ~0.75 on the delay class → we catch most risky flights with controlled false alarms.

Explainable drivers.

- Largest lifts come from actual pushback lateness, time-of-day/season, and historical route/airline performance.

Operational impact.

- Probability outputs enable policy thresholds by hour/airport, earlier passenger comms, and smarter gate/crew replanning

Machine Learning Project - DTM 25

Thank you for listening

Flight delayed Prediction

Alessandro De Faveri