Is the Plane Ready Yet? Unraveling Flight Delays with Data Science

Team 6-2

April 19, 2024

Berkeley
UNIVERSITY OF CALIFORNIA


0805 0815 MILANO
DELAYED
0845 0855 PARIS
DELAYED
0905 0915 NEW YORK
DELAYED
0910 0920 FRANKFURT
DELAYED
0925 0955 LONDON

Team 6-2

Chi So

Derek Dewald

Michael Botros

Vinh Bui

AGENDA

- Define Metrics
- Summary Exploratory Data Analysis (EDA)
- Feature Engineering and Top Features
- Overview of Modeling Pipelines Explored
- Results and Discussion
- Conclusions

Understanding the F-beta Score

The general formula for non-negative real β is:

$$F_\beta = \frac{(1 + \beta^2) \cdot (\text{precision} \cdot \text{recall})}{(\beta^2 \cdot \text{precision} + \text{recall})}$$

- The F-beta score is a metric used in machine learning to evaluate the accuracy of a classification model.
- It considers both precision (the accuracy of positive predictions) and recall (the ability to find all the positive instances) in its calculation.
- The "beta" in the F-beta score determines the weight of precision and recall in the harmonic mean.
- We decide to use β = 0.5 because we favor precision

Adjusted weighted average for Cross Validation

Compensate for more recent time

$$Weighed\ Avg = \frac{\sum_{i=1}^{n} i * f\_beta_i}{\sum_{i=1}^{n} i}$$

Exploratory Data Analysis: Process

- **Step 1:**
  - Problem Statement: Can we predict 2 hours in advance whether a scheduled departure will be delayed by more than 15 minutes?
- **Step 2:**
  - Understanding the Data
    - Quality, Completeness, Consistency.
- **Step 3:**
  - Data Insights
- **Step 4:**
  - Feature Engineering

Exploratory Data Analysis: Data Sets

Table Name	Total Rows	Total Columns	Source	Table Description
Flights	31,269,523	109	DOT	Flight data as captured and reported by Department of Transportation, including pertinent information related to flight travel, departure, arrival, delays, take-off, landing, etc. Data includes US Domestic Commercial Air Traffic between 2015 - 2019.
Weather	402,403,814	124	NOAA	Hourly weather data as captured by weather stations through US. Information includes point of time and average reading related to temperature, precipitation, humidity.
Airport Codes	20,089	12	Data Hub	Descriptive Airport information, including Name, Latitude and Longitude. Primarily utilized for mapping of Flight and Weather Data. Data supplemented with 18 missing values, which included PR, GU and VI.

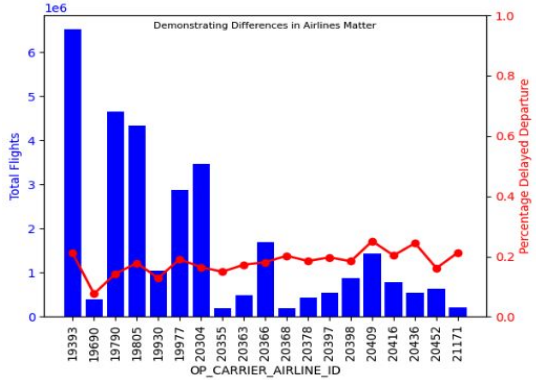
Links

Department of Transportation Direct Link; https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr
NOAA: National Oceanic and Atmospheric Administration Direct Link: https://www.ncei.noaa.gov/data/local-climatological-data/archive/
Airport Code: https://datahub.io/core/airport-codes

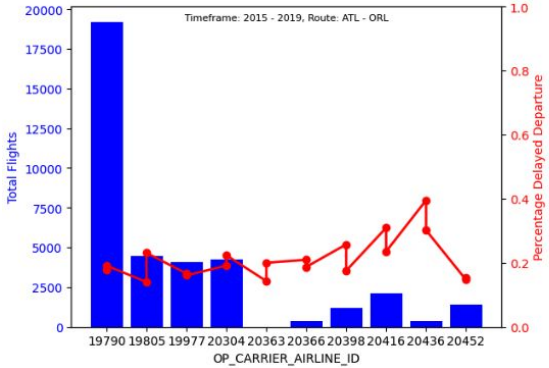
Data Insights

- **Airline Matters**
  - Not all Carriers Operate Equally
- **Airport Matters**
  - Not all Airports are the same
- **Distinctions based on Route**
  - Not all routes are created equal

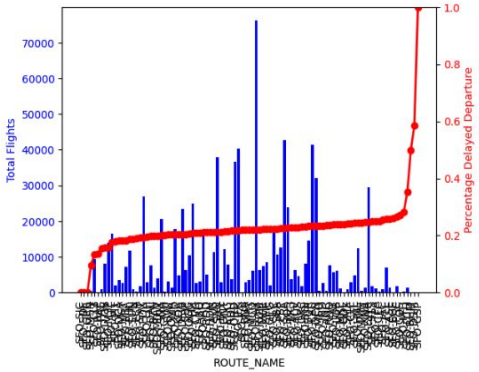
Graph 1. Performance 2015 - 2019 All Airlines by ID



Graph 2. Performance by Carrier on Individual Route

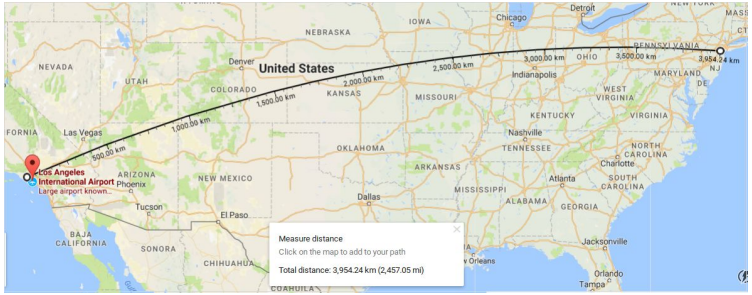


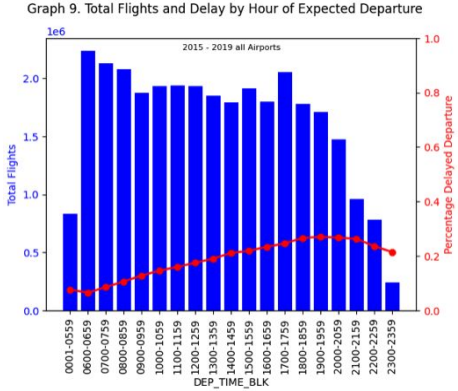
Graph 5. Total Flights and Perc Departure Delay from SFO 2015 - 2019

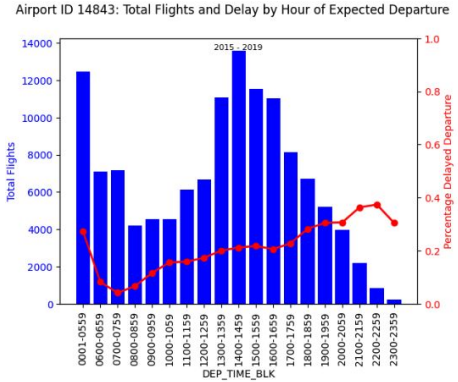


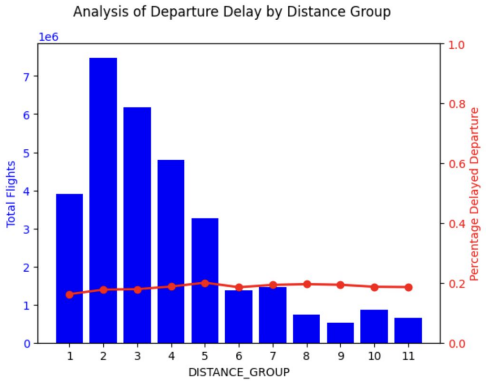
Data Insights

- **Time of Day**
  - Does time matter, or is time indicative of something else?
- **Distance**
  - Can Airlines really make it up in the air?







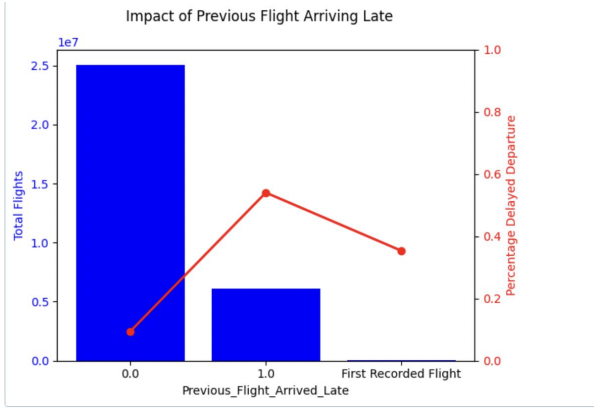


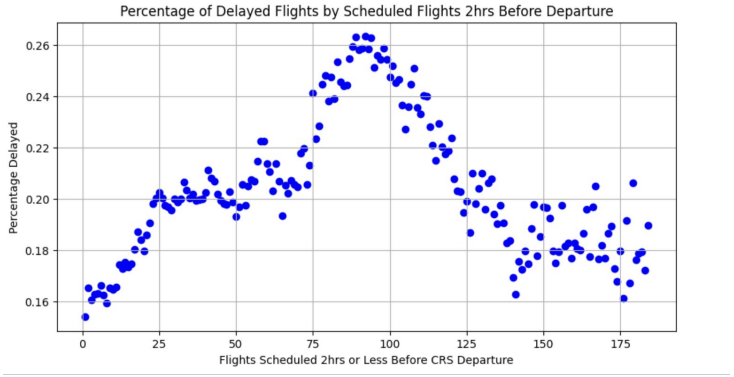
Feature Engineering

- **Raw weather did not improve F-Beta Score significantly**
  - How does weather impact
    - In the Air? - Captured in whether flight Arrived Late.
    - At the Airport? - Captured in impact trailing activity and delays
- **Commitment to deliver a Intuitive, Interpretable, and Effective Prediction Model**

Feature Engineering

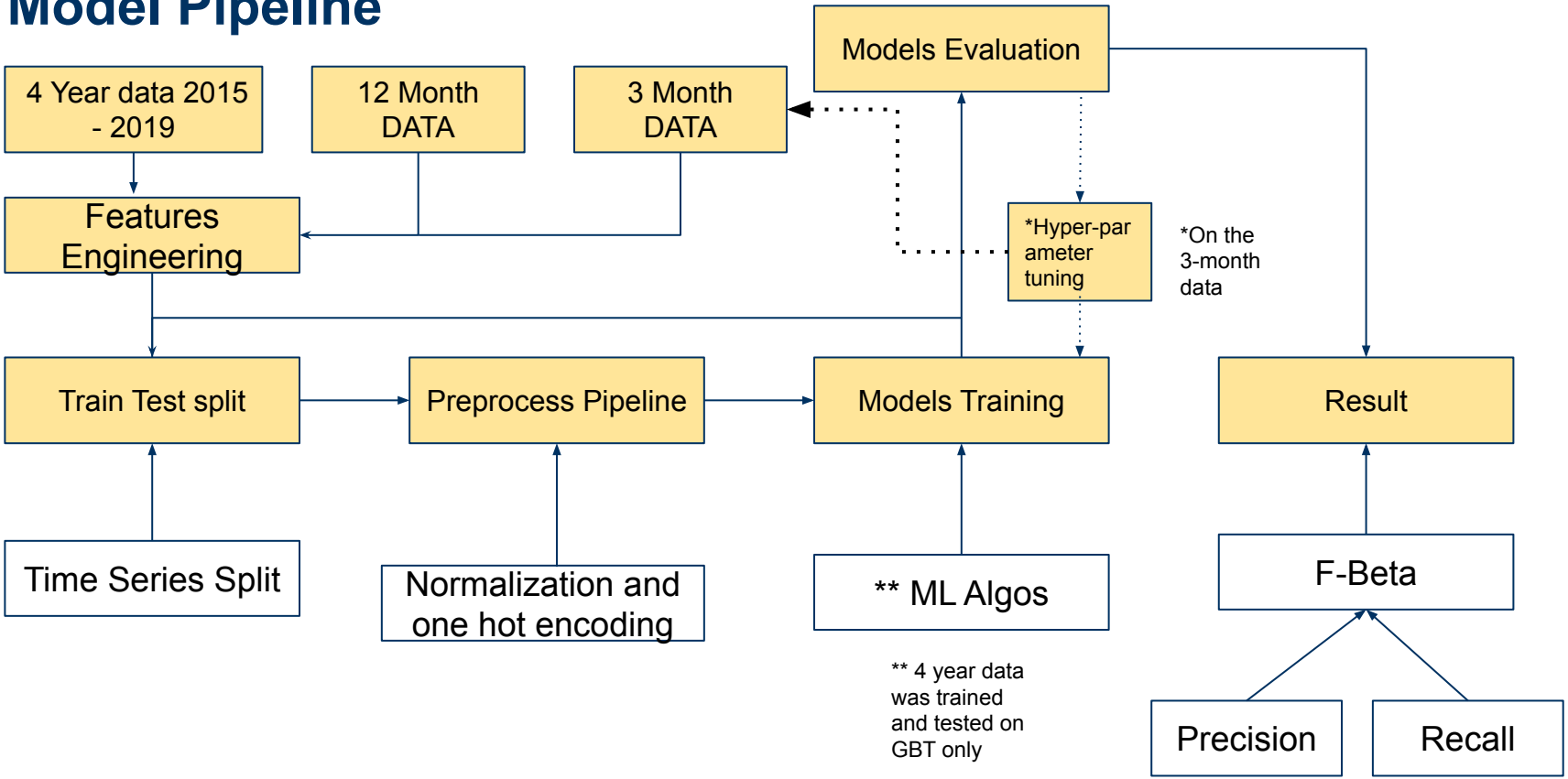
- **Modelled to capture underlying cause, not effect**
  - Whether previous flight arrived late and expected turnaround time
    - Airlines leverage and stretch operational efficiency models, this identifies areas of strain and where least amount of flexibility exists
  - Number of flights scheduled within next 2 hours
    - Captures activity at airport, simplifies impact of seasonality, surge demand
  - Number of flights and delays occurring 0-2 hours prior to prediction
    - Captures in real time impact of events such as impacts of weather, runway bottlenecks, system outages, holidays
- **Relationship between Routes**
  - Page Rank





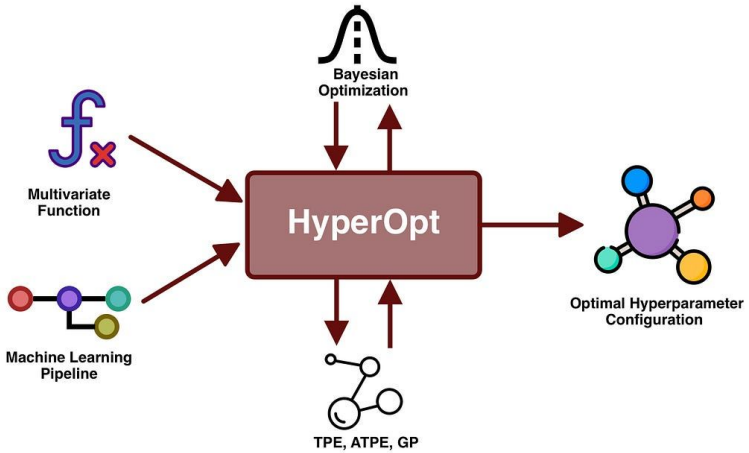
Overview of Modeling Pipelines

Model Pipeline



HyperOpt - More efficient ways to optimize

- Maximize F Beta-score
- Using Time Series Cross validation to calculate the f beta score
- Perform hyperparameters tuning on 12-month dataset
- Implement early stopping to reduce wasting resources



Results and Discussion

Final Results

On 12 - months dataset

Model	Train CV	Test
Logistic Regression	77.22%	80.3%
Random Forest	67.61%	71.38%
Gradient Boosted Trees	88.7%	90.0%
Multilayer Perceptron `[639,Sigmoid,7,Sigmoid, 2, Softmax]`	TBD	87.67%

Going up in scale

Trained GBT on 2015 → 2018, test on 2019

~ 90%

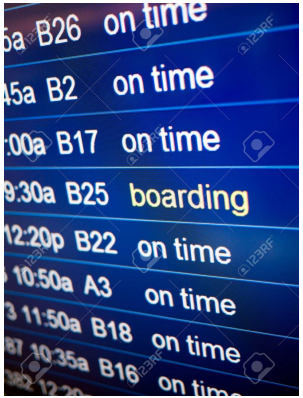
Conclusion

Problem: Can we predict when a departing plane will be delayed by more than 15 minutes, 2 hours in advance of scheduled departure?

Requirement: Develop a Intuitive, Interpretable and Effective Prediction Model

Solution:

- Gradient Boosted Trees
- 90%, 10% Improvement relative to baseline



Impact:

- Increase our ability to hold airlines accountable, and situationally identify opportunities to audit performance, and levy fines.

Next Steps

Preliminary Results

Use 14 features, including previous 2 and 12 weather features:

Model	Train CV	Test
Logistic Regression	76.76%	80.65%
Random Forest	67.79%	72.78%
GBT	77.74%	81.35%
Multilayer Perceptron	75.49%	79.2%

Project Abstract

- **Our goal:** Predict flight delays within a 2-hour window to enhance airline operations.
- **Dataset:** Utilize flight schedules, airline data, weather conditions, and historical delays.
- **Data Source:** flight information - https://transtats.bts.gov/
  Weather Data - https://www.ncei.noaa.gov/
- **Methods:** Apply machine learning algorithms (logistic regression, random forests, XGBoost, neural networks) via PySpark.
- **Techniques**: Conduct exploratory data analysis, feature selection, engineering, and parameter tuning.
- **Baseline Model:** Logistic regression achieves 80.6% F-beta score on selected features.
- **Evaluation:** Rigorously assess models based on F-Beta
- **Implications**: Provide airlines insights for preemptive action and empower passengers with travel planning.

Understanding the Data

	Observation	Description	Unique Observations	Null Value Observations	Delays Occuring with Null Value Observations
0	ORIGIN_AIRPORT_ID	Origin Airport, Airport ID. An identification ...	371	0	0
0	ORIGIN_AIRPORT_SEQ_ID	Origin Airport, Airport Sequence ID. An identi...	688	0	0
0	ORIGIN_CITY_MARKET_ID	Origin Airport, City Market ID. City Market ID...	344	0	0
0	ORIGIN	Origin Airport	371	0	0
0	ORIGIN_CITY_NAME	Origin Airport, City Name	362	0	0

Focus on Data Quality, Consistency and Cleaning

FL_DATE	TAIL_NUM	OP_CARR	ORIGIN_A	ORIGIN_A	ORIGIN_C	ORIGIN	ORIGIN_CITY_NAME	ORIGIN_S	ORIGIN_V	DEST_AIR	DEST_AIR	DEST_CIT	DEST	DEST_CIT	DEST_STA	DEST_STA	DEST_STA	DEST_WA	CRS_DEP	DEP_TIME	DEP_DI
2019-01-01	N914EV	4321	10135	1013505	30135	ABE	Allentown/Bethlehem/I	PA	23	11433	1143302	31295	DTW	Detroit, MI	MI	26	Michigan	43	535	529	
2019-01-01	N914EV	3618	11433	1143302	31295	DTW	Detroit, MI	MI	43	12884	1288403	32884	LAN	Lansing, MI	MI	26	Michigan	43	835	830	
2019-01-01	N914EV	3618	12884	1288403	32884	LAN	Lansing, MI	MI	43	11433	1143302	31295	DTW	Detroit, MI	MI	26	Michigan	43	1000	953	
2019-01-01	N914EV	4735	11042	1104205	30647	CLE	Cleveland, OH	OH	44	11433	1143302	31295	DTW	Detroit, MI	MI	26	Michigan	43	1014	1245	
2019-01-01	N914EV	7434	11433	1143302	31295	DTW	Detroit, MI	MI	43	12397	1239702	32397	ITH	Ithaca/Cortl	NY	36	New York	22	1540	1536	