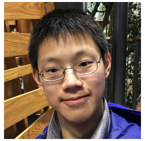
Reducing Flight Delays at Southwest Airlines:

A Data-Driven Strategy

Meet the Team: Group 4

Ayushi Goel



Licheng Zhong

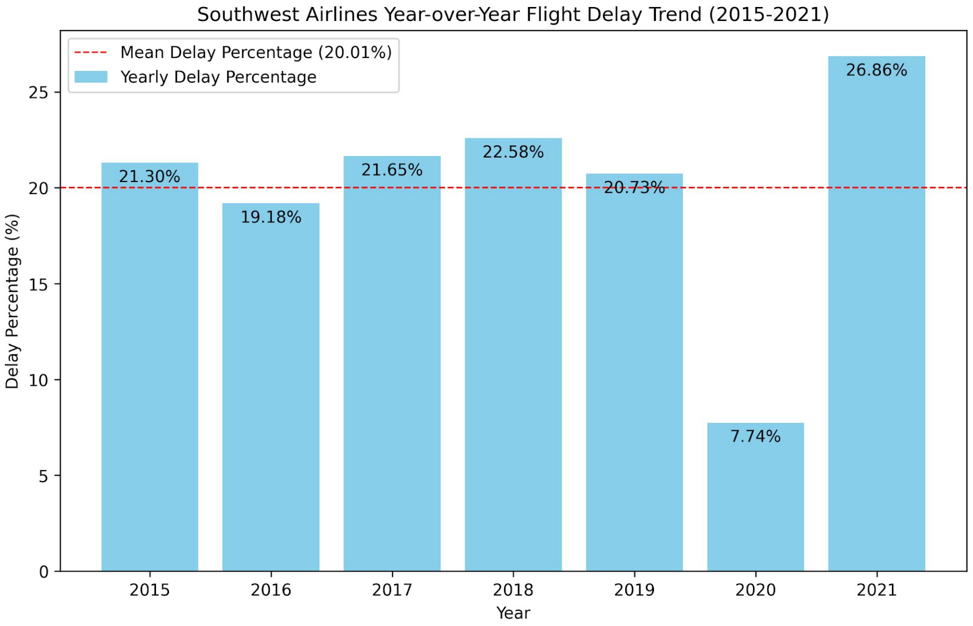
Louis Wu

Sammy Cayo

Target Audience *- Senior Leadership at Southwest Airlines & Co.*

Problem Statement

- From 2014-2021, Southwest Operated 14.94 million flights
- 3.09 million were delayed
  - 20.01% delay rate
  - 7.74% in 2020 attributed to global pandemic
- Financial Impact of at least $2.85 Billion
- Operational breakdown
  - Increased fuel consumption (23.5%)[1]
  - Labor cost (55.0%)[1]
  - Maintenance costs (8.4%)[2]
  - Cancelled flights compensation payout



Berkeley

Source - See Appendix

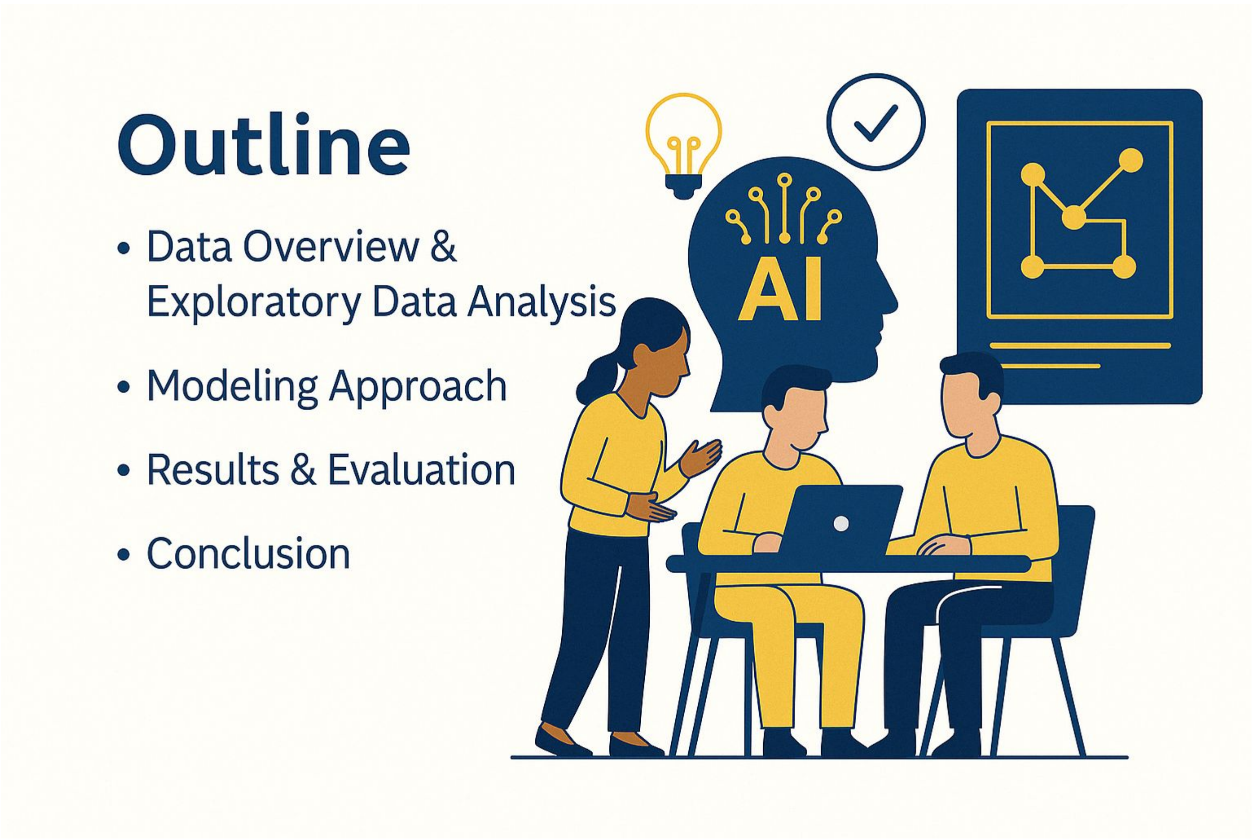
Objective

- Machine learning-based prediction to decrease delays
- Classification-based models
  - Logistic regression
  - Random Forester
  - XGBoost
  - Multi-layer Perceptron (Neural Network)
- Goal: 5% delay reduction
  - $142 Million operational cost savings
  - 154,696 fewer delays
  - ↑ Customer satisfaction
  - ↑ Shareholders value

Outcome

- Actual Delay vs Model Prediction
  - 551,740 total delays in 2019
  - 236,912 true delay prediction (True Positive)
    - ~42% of all delays
  - 36,612 not delayed (False Positive)
- Cost savings for 2019
  - ~ $218 Million opportunity to reduce cost (see Appendix)
  - Not all delays can be avoided
  - 5-10% cost reduction target
    - ~$10.9 - $21.8 million for 2019

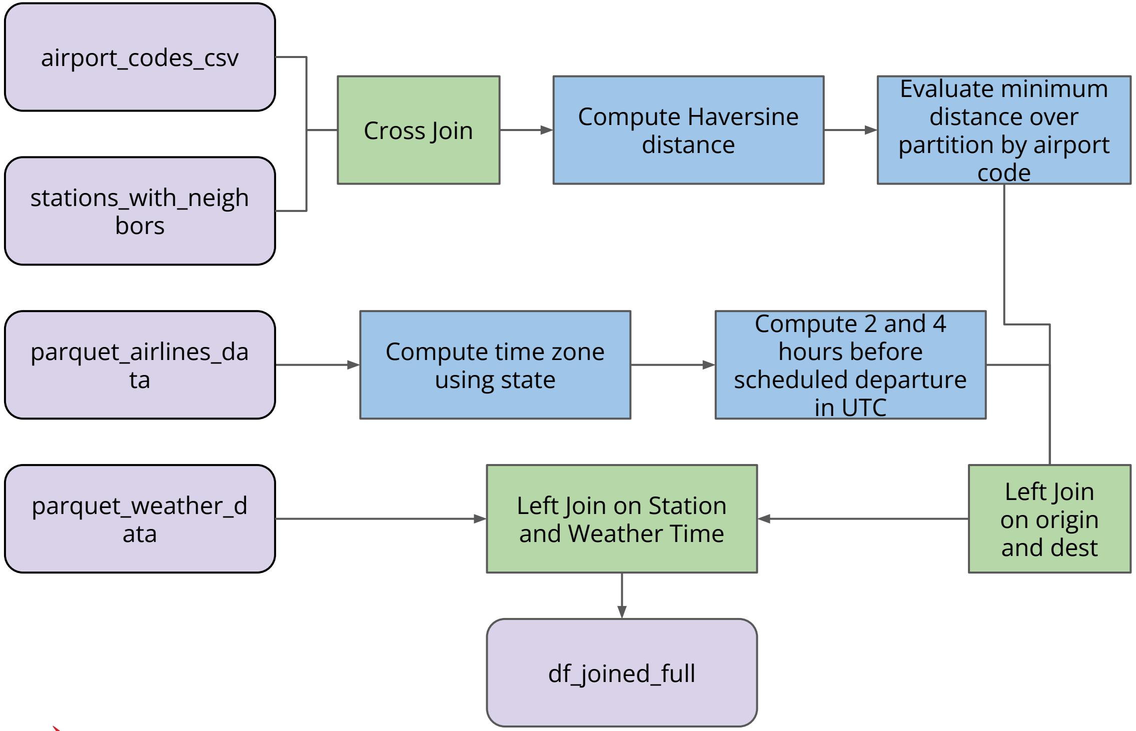
Image Source - *https://career-advising.ndsu.edu/blog/2023/01/13/how-to-set-better-goals/*



Berkeley

Image Source *- https://openai.com/index/dall-e-3/*

Join Pipeline (2015-2021)



Total Time: 19 minutes
Total File Size: 6.6 GB

- Deduplicated airlines data from 2015-2019
- Performed cross join first, which minimizes number of records created in intermediate CTEs
- Performed final range join directly in spark DF, outside of SQL
- Bucketed weather and flight times by hour to minimize number of comparisons required

Aircraft Tail Number Features

- **Prev Cancelled**: whether the previous flight leg for a given tail number was cancelled
- **Prev Origin**: the origin of the previous flight leg a given tail number
- **Minutes between Flights**: difference between arrival datetime UTC of the previous leg and the scheduled departure time UTC of the current leg
- **Prev Arr Delay**: difference in minutes between scheduled departure time and actual departure time
- **Prev Arr Delay New**: non-negative difference in minutes between scheduled departure time and actual departure time
- **Prev Arr Delay 15**: Whether the delay of the previous flight exceeded 15 minutes
- **Triplet**: a string denoting the origin of the previous leg, the current origin, and the destination of a given tail number

Incorporating Recent Data (2022-2024)

Airlines Data (7.3 GB)

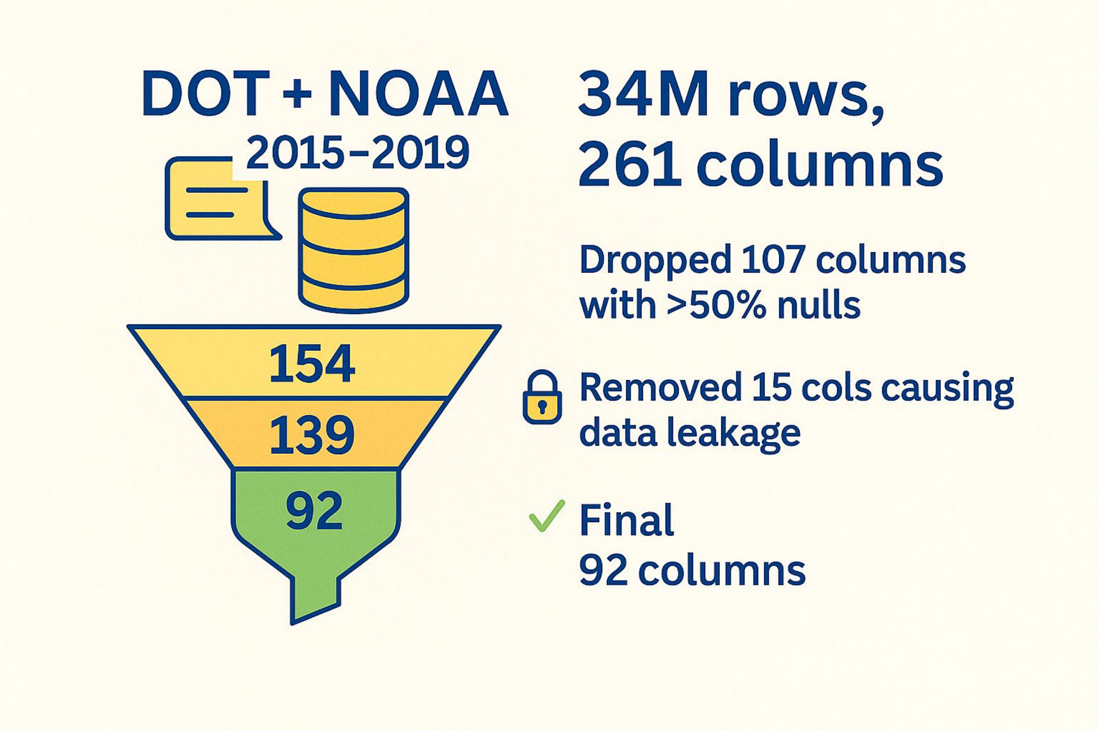
- Downloaded by month as .csv files
- Uploaded to DBFS through browser
- Checkpointed as parquet file
- 21 million additional rows

Weather Data (11 GB)

- Downloaded by year as .tar.gz files
- Uploaded to DBFS through CLI
- Copied from DBFS to local driver
- Extracted on local driver
- Copied extracted files to DBFS
- Checkpointed as parquet file

Berkeley

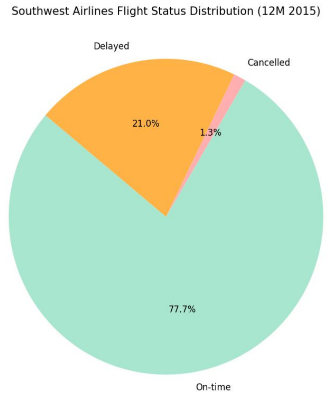
Data Description

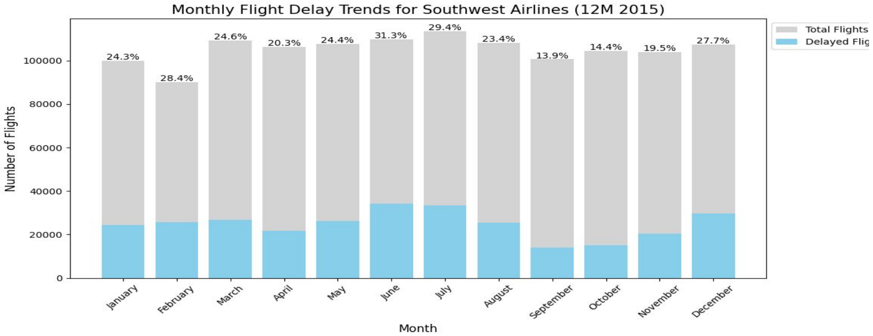


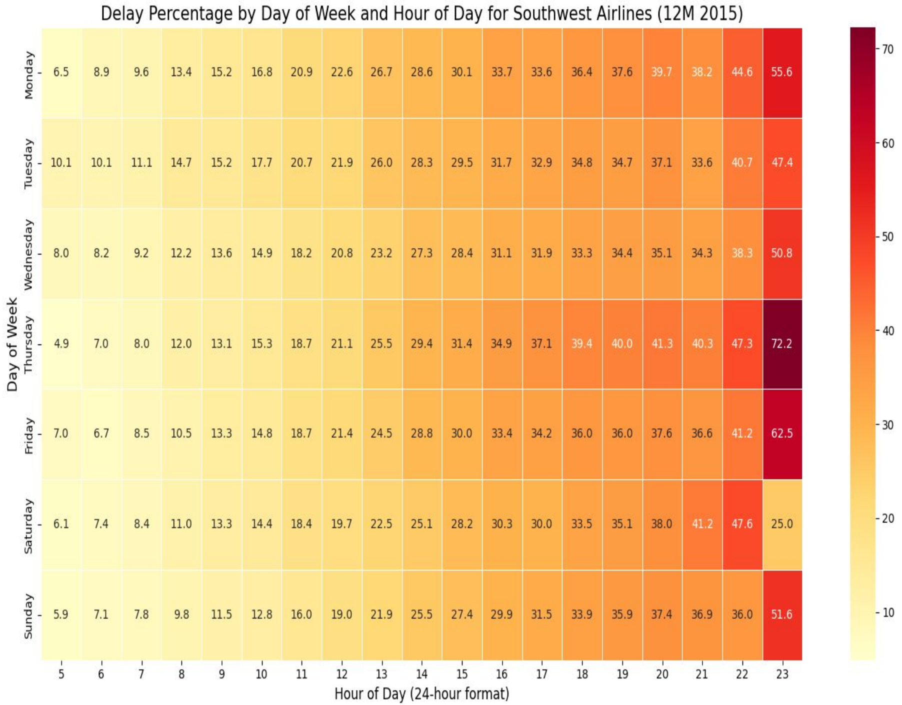
Berkeley

Image Source *- https://openai.com/index/dall-e-3/*

Timing is Everything

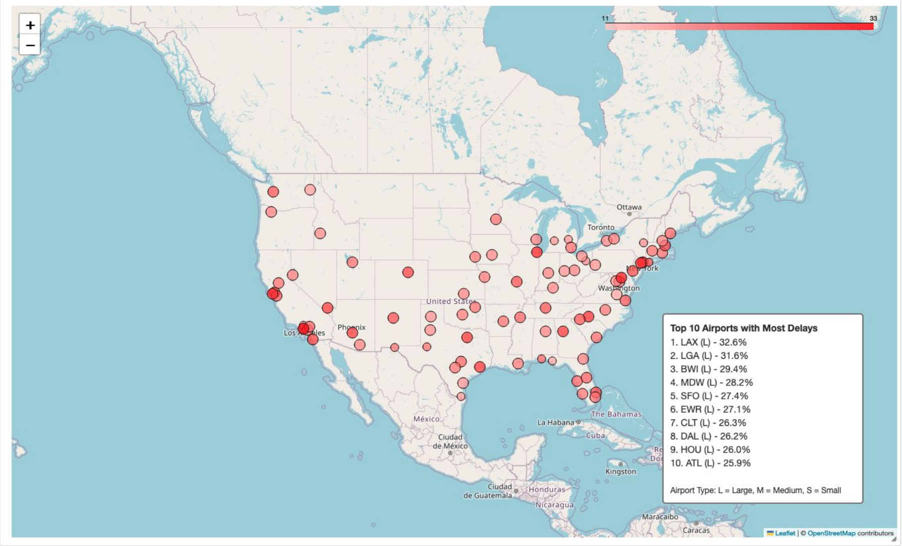




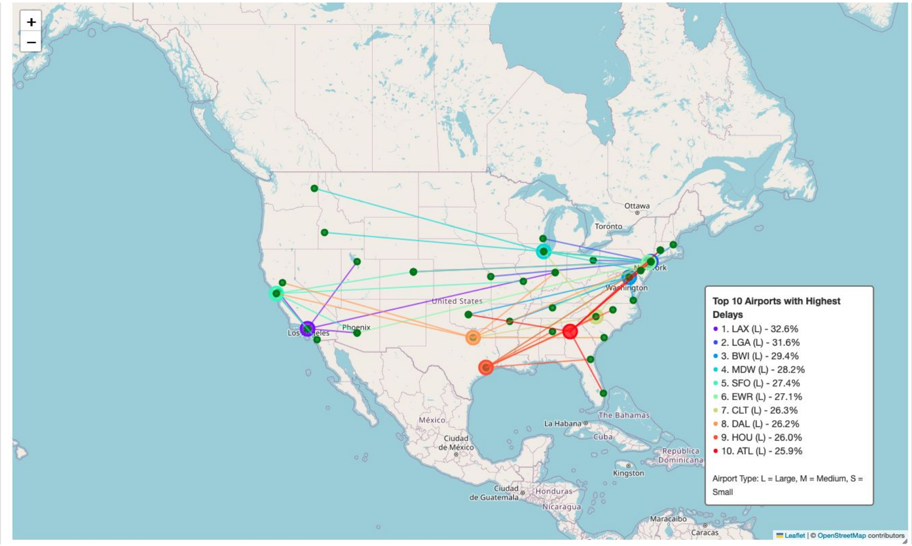


Berkeley

Where Things Are Taking Off... Late



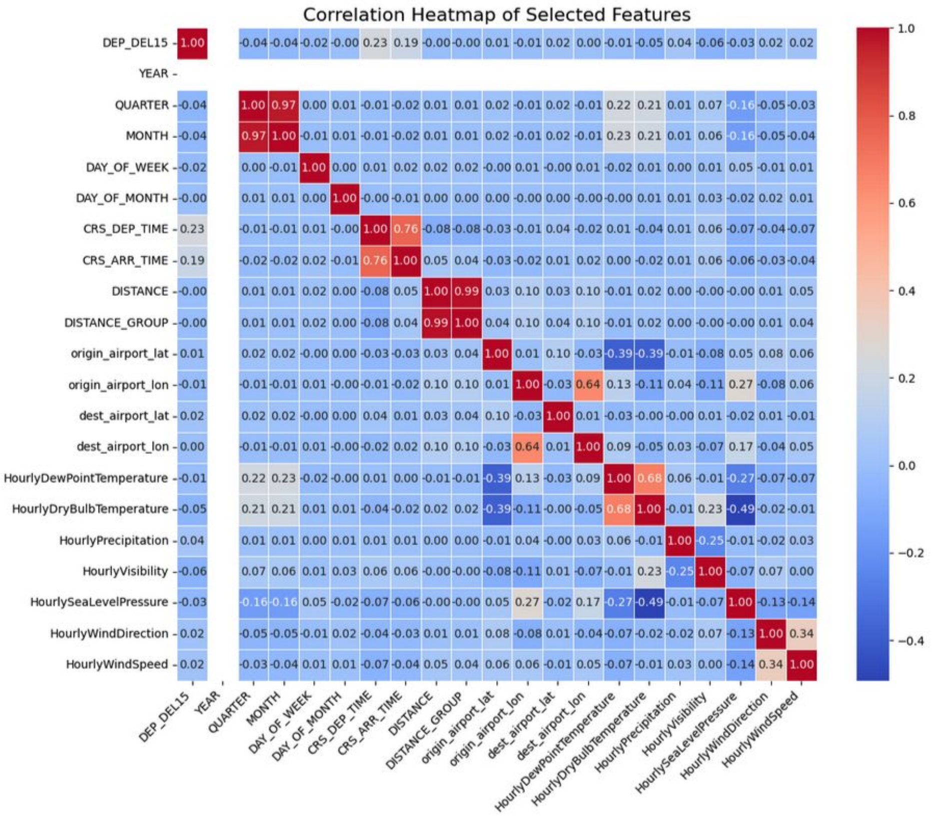
(12M 2015)



(12M 2015)

Berkeley

What (Barely) Correlates with Delays - Pearson correlation



Comprehensive Feature Engineering for Flight Delay Prediction

- Airport Profile
- Time-Based Profile
- Weather-Based Profile
- Southwest Airlines Profile

Berkeley

Image Source *- https://openai.com/index/dall-e-3/*

Airport Profile Features

Feature	Description	Null Handling	Temporal Integrity
Origin Airport Daily Operations	Total number of flights departing from each origin airport on a given day	None needed (always populated)	Current day only
Origin Airport 30-Day Rolling Volume	Sum of flights from the origin airport over the past 30 days	0 for first 30 days (no history available)	Growing window until 30 days history
Origin Airport 1-Year Delay Rate	Annual delay percentage at origin airport	Global fallback (15%) for first year (2015) rows	Expanding window using all prior data
Route	The origin and destination of the flight	Not needed	Concat the origin and destination in a single window
Route Traffic Volume	Number of flights between specific origin-destination pairs over the past year	0 for new routes or first year (2015) rows	Expanding window using all prior data
Southwest Market Share	Percentage of flights operated by Southwest at each origin airport over the past year	0 when no data available for Southwest flights	Rolling 365-day window
Southwest Origin 30-Day Delay Rate	Recent Southwest delay performance at origin airport (past 30 days)	Global fallback (15%) for missing data in the previous 30 days	Growing window until 30 days history
Southwest Route Historical Performance	Southwest's historical delay rate on specific routes over the past year	Global fallback (15%) for missing route data or first year rows	Expanding window using all prior data
Southwest Relative Performance Index	How Southwest compares to other airlines at the same airport (delay rate ratio)	Default value of 1.0 when no data available or division by zero occurs	Ratio with epsilon smoothing to prevent division by zero

Time-Based Profile Features

Feature	Description	Null Handling	Temporal Integrity
time_bucket	15-minute departure intervals	Derived from CRS_DEP_TIME (always populated)	Current flight only
dep_hour	Hour of day for scheduled departure	None needed	Current flight only
time_of_day_category	Morning/Midday/Evening/Night	Categorical fallback to "night"	Current flight only
is_weekend	Weekend flight indicator	None needed	Current flight only
holiday_season	Peak travel period indicator	None needed	Current flight only
prior_day_delay_rate	Previous day's delay rate at origin airport	3-level fallback: prior day → airport avg → 15% global fallback	Strict date ordering
same_day_prior_delay_percentage	Percentage of flights delayed earlier in the day at the same airport	Additive smoothing (prevents 0/0) and nulls default to 0% delay rate	Same-day ordering
time_based_congestion_ratio	Current vs historical congestion ratio for the same time bucket (hour + 15-min interval) on the same day of the week at the same airport	3-level fallback: historical average → airport avg → default capacity (10 flights)	365-day lookback excluding current day

Weather Profile Features

Feature	Description	Calculation Method	Null Handling
extreme_precipitation	Flag for heavy precipitation	95th percentile of historical precipitation data	0 if missing
extreme_wind	Flag for high wind conditions	95th percentile of historical wind speed data	0 if missing
extreme_temperature	Flag for extreme temperatures	5th/95th percentiles of historical temperature data	0 if missing
low_visibility	Flag for poor visibility	5th percentile of historical visibility data	0 if missing
extreme_weather_score	Weighted weather risk score	Weighted sum of extreme conditions based on their historical delay impact	Scaled to [-1,1]
heat_index	Perceived temperature	NOAA heat index formula for $T \geq 80°F$ and $RH \geq 40\%$	Raw temp otherwise
rapid_weather_change	Significant weather shifts	Z-score > 3 in temp/wind over 24h window	0 if missing data
temp_anomaly_z	Temperature deviation	Z-score vs. airport-month historical average	0 if no history
precip_anomaly_z	Precipitation deviation	Z-score vs. airport-month historical average	0 if no history

Berkeley

Southwest Airlines Profile Features

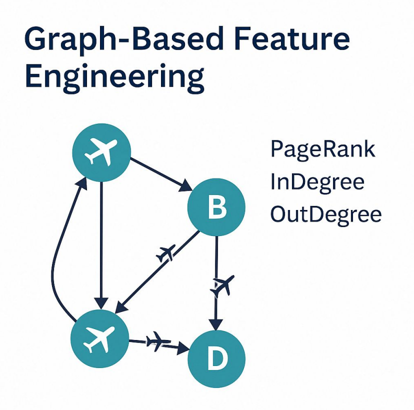
Feature	Description	Calculation Method	Null Handling
sw_time_of_day_delay_rate	Southwest's delay rate by origin and time bucket	Expanding window average with origin/global fallbacks	Uses origin average → global median
sw_day_of_week_delay_rate	Bayesian-smoothed delay rate by route and weekday	(Delays + 3*global_p30)/(Flights + 3)	Built-in smoothing prevents nulls
sw_aircraft_delay_rate	Aircraft performance metric	Hierarchical: aircraft → route → global median	Always populated
sw_origin_hub	Dynamic hub identification	Top 15th percentile of Southwest flight volume	0/1 encoding
sw_schedule_buffer_ratio	Schedule padding ratio	Current vs 1-year historical average	Defaults to 1.0
sw_origin_time_perf	Hybrid airport/time performance	Time bucket → time category → global fallback	Hierarchical coalesce
sw_route_importance	Normalized route significance	(Flight count + distance) normalized	Always 0-2 range

Berkeley

Feature Engineering Graph Based

Graph Feature Category	Description	Calculation Method	Lag method
`PageRank`	Measure of influence of high-traffic airports based on flight connection	Distinct airport ID as Vertices and flight routes as Edges (src: origin airport ids, dst: destination airport ids)	Year
`InDegree`	Measure of high-traffic airport arrival patterns	Count of incoming connections from an airport	Quarter
`OutDegree`	Measure of high-traffic airport departure patterns	Count of outgoing connections from an airport	Quarter

GraphFrames



PageRank
- Airport linked to other airports are ranked higher

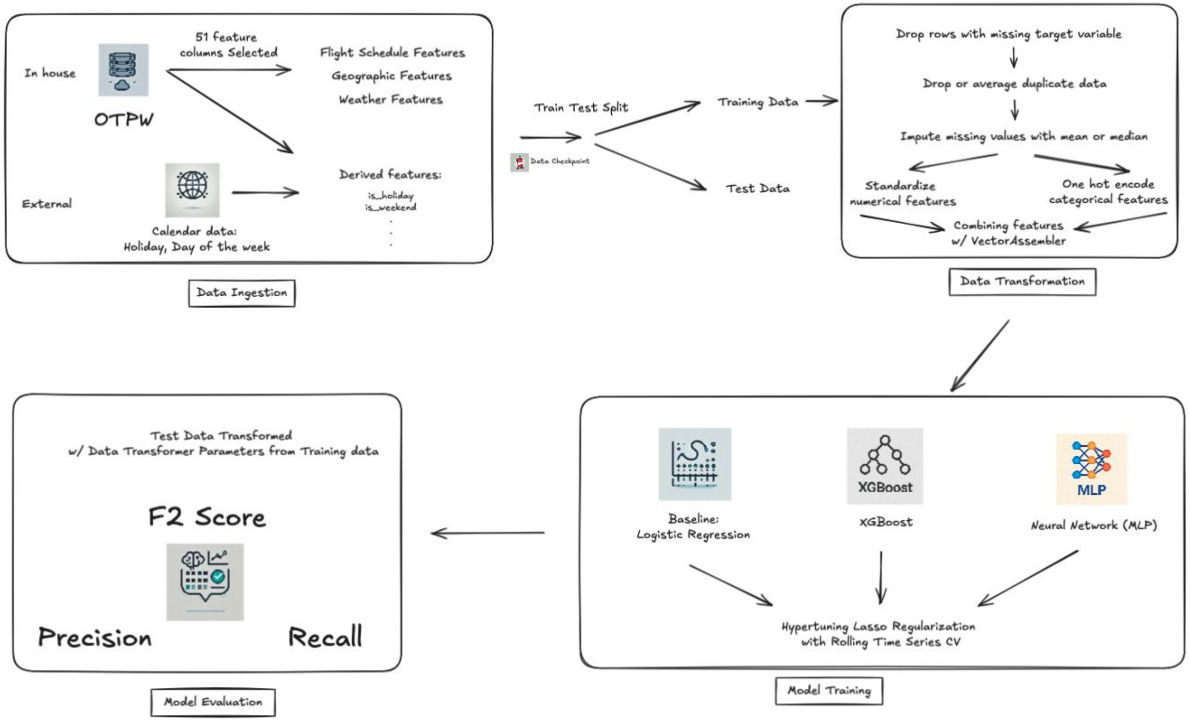
InDegree
- Popular destination have higher values

OutDegree
- Major Hubs have higher values

Berkeley

Image Source *- https://openai.com/index/dall-e-3/*

ML Pipeline



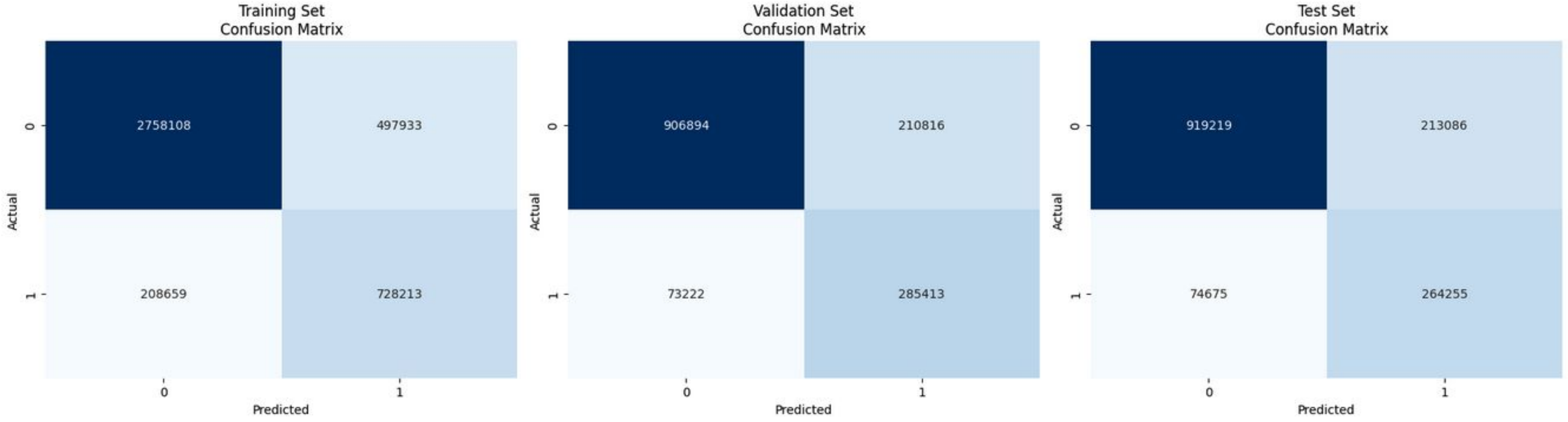
Berkeley

Baseline Results - Logistic Regression 2015-2019

No Regularization; Runtime = 2 minutes, GPU - 5 Workers

Metrics Comparison:

dataset	precision	recall	f2
Training	0.594	0.777	0.732
Validation	0.575	0.796	0.739
Test	0.554	0.780	0.721



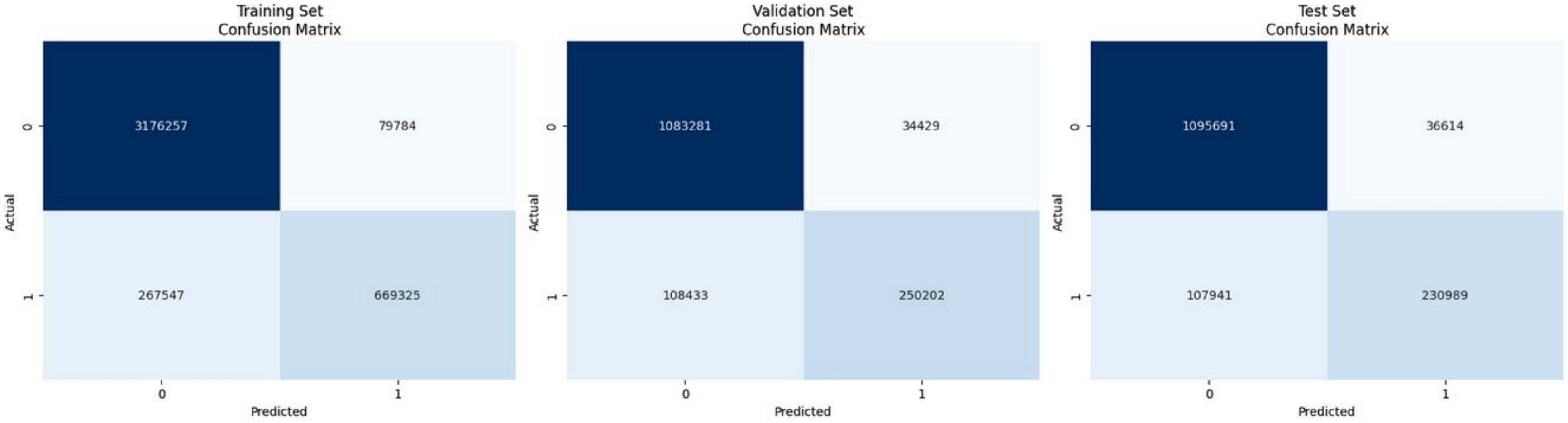
Berkeley

XGBoost Results - 2015-2019

(n_estimators=100, max_depth=6, learning_rate=0.3); Runtime = 3 minutes, GPU - 5 workers

Metrics Comparison:

dataset	precision	recall	f2
Training	0.893	0.714	0.744
Validation	0.879	0.698	0.728
Test	0.863	0.682	0.711

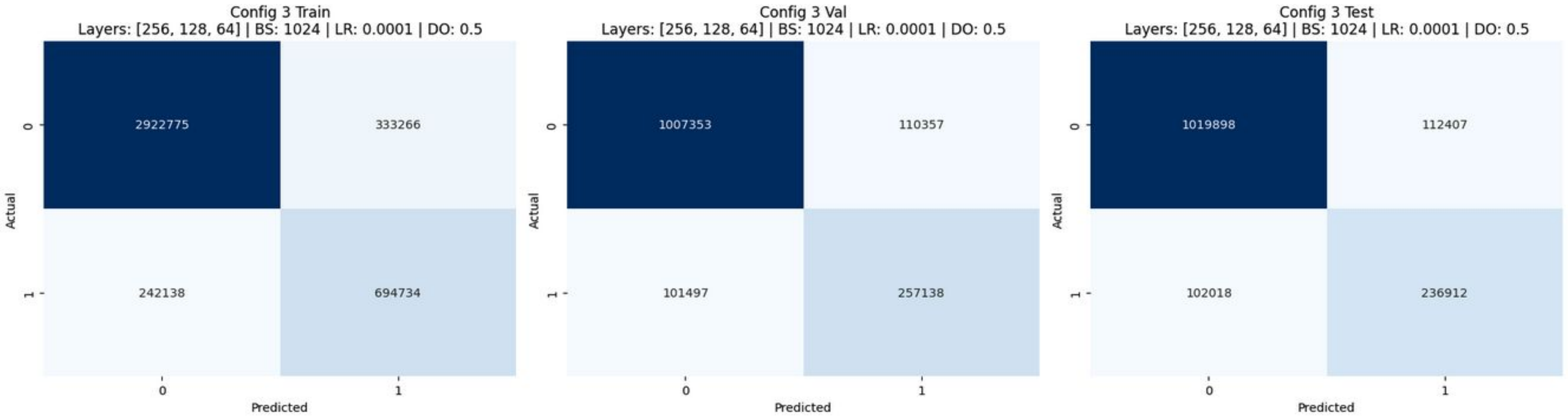


Neural Network Results - 2015-2019

Runtime = 27 minutes, GPU - 5 workers

Final Comparison:

Layers	Batch	LR	DO	L2	TrnF2	TrnRcl	ValF2	ValRcl	TstF2	TstRcl
[256, 128, 64]	1024	0.0001	0.5000	0.0010	0.7274	0.7415	0.7135	0.7170	0.6947	0.6990
[128]	256	0.0010	0.4000	0.0010	0.6617	0.6426	0.6686	0.6609	0.6641	0.6705
[128]	256	0.0005	0.2000	0.0010	0.6889	0.6780	0.6661	0.6457	0.6506	0.6346
[128, 64]	512	0.0003	0.3000	0.0010	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000



Conclusion

- Best Model: XGBoost
  - Comparable F2 (0.73) with Logistic Regression (0.74)
    - Higher than NN’s 0.66
  - All metrics are around 0.7 or above
    - Logistic Regression has a bad precision: 0.57
- Number of features: 49
- Hyper parameters:
  - n_estimators=100
  - max_depth=6
  - learning_rate=0.3

Top 10 features from XGBoost by Gain

1. prev_cancelled
2. sw_market_share (by origin airport)
3. minutes_between_flights
4. origin_type (large vs medium vs small airport)
5. day_of_week
6. dest_type (large vs medium vs small airport)
7. Prior_day_delay_rate (by origin airport)
8. time_of_day (morning, midday, evening, night)
9. Prior_delays_today (by origin airport)
10. Sw_origin_time_perf (by origin airport and 15 min time bucket)

Berkeley

Questions?

Appendix: Clarification of Financial Calculations

The financial estimates provided in this project are **approximations** and should be interpreted with caution. The total economic impact of flight delays and potential savings from reducing delays are derived from **publicly available data, not Southwest Airlines' internal financial records**.

Source of Economic Impact Estimate:

The cost per delayed flight is taken from the article **"Flight Delays in Numbers – Not Only Painful For Passengers"**, which states that **each delayed flight costs approximately $920**.

Calculation Details:

- **Total Delayed Flights (2015–2021): 3.09 million**
- **Cost Per Delayed Flight: $920**
- **Total Economic Impact of Delays:**
  - **3,090,000 × 920 = $2.84 billion**
- **Potential Savings from a 5% Reduction in Delays:**
  - **5% of Total Delayed Flights:**
    - **3,090,000 × 0.05 = 154,500 flights**
  - **Potential Savings:**
    - **154,500 × 920 = $142.1 million**

Berkeley

Appendix: Clarification of Financial Calculations

- 2019 total delay rate: 551, 740
- True prediction delay: 236,912
- Delay opportunity
  - 236,912 * 920 ~ $218 Million
- 5% cost reduction
  - 236, 912 * 5% = 11,845.6 * 920 ~ $10.9 Million
- 10% cost reduction
  - 236, 912 * 10% = 23,691 * 920 ~ $22 Million

Appendix: Operational Costs Breakdown

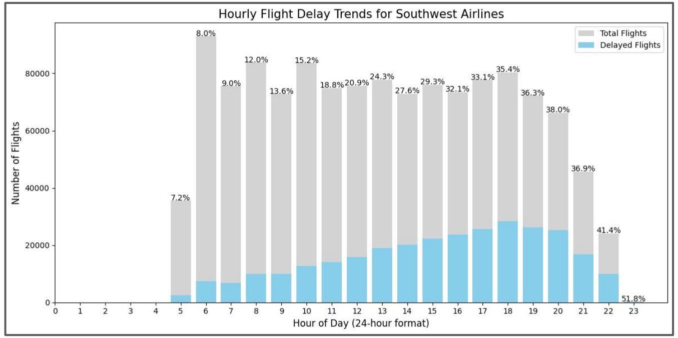
Sources

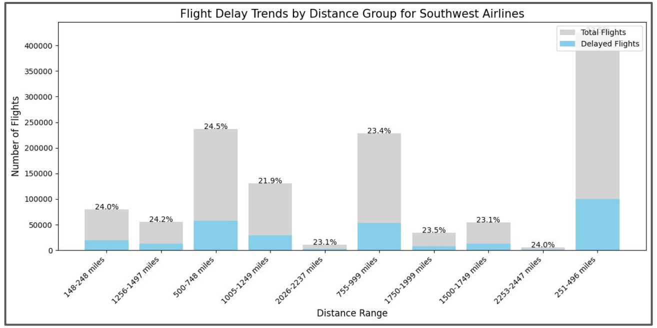
1: https://www.sec.gov/ix?doc=/Archives/edgar/data/0000092380/000009238022000007/luv-20211231.htm

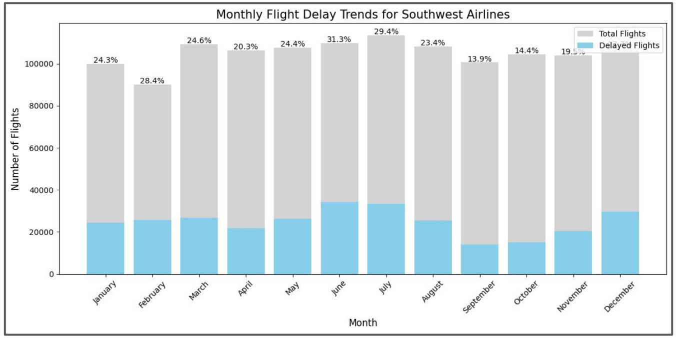
2: https://www.iata.org/en/publications/newsletters/iata-knowledge-hub/unveiling-the-biggest-airline-costs/

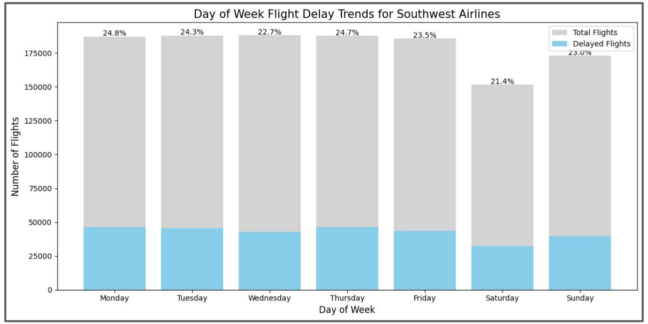
Berkeley

Appendix:









Berkeley