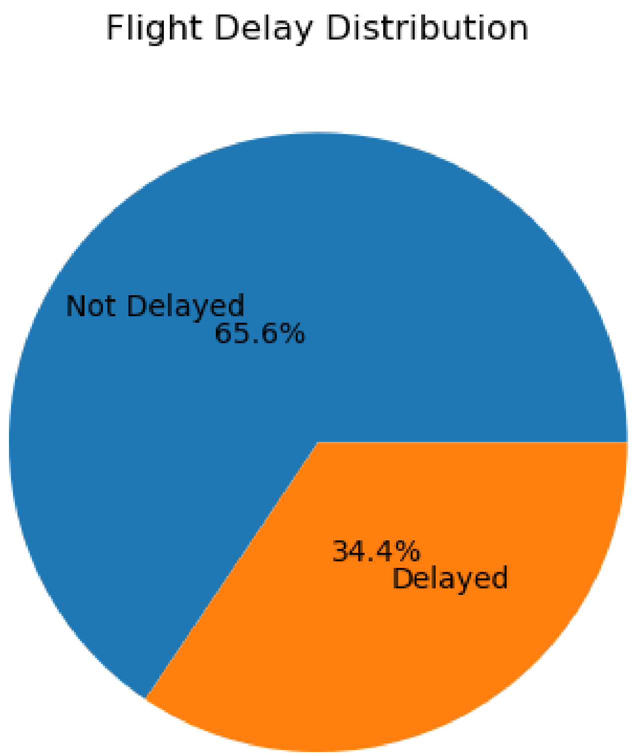
PREDICTING FLIGHT DELAYS

A DATA-DRIVEN APPROACH TO ENHANCING AIRLINE EFFICIENCY

AGENDA

1. The Problem
2. Data Analysis & Modeling
3. Recommendations
4. Going Forward

ACKNOWLEDGING FLIGHT DELAYS



THE IMPACT OF FLIGHT DELAYS ON AIRLINES

Operational Inefficiencies

Flight delays disrupt scheduling and complicate resource management.
These interruptions can lead to being over or understaffed in certain areas.

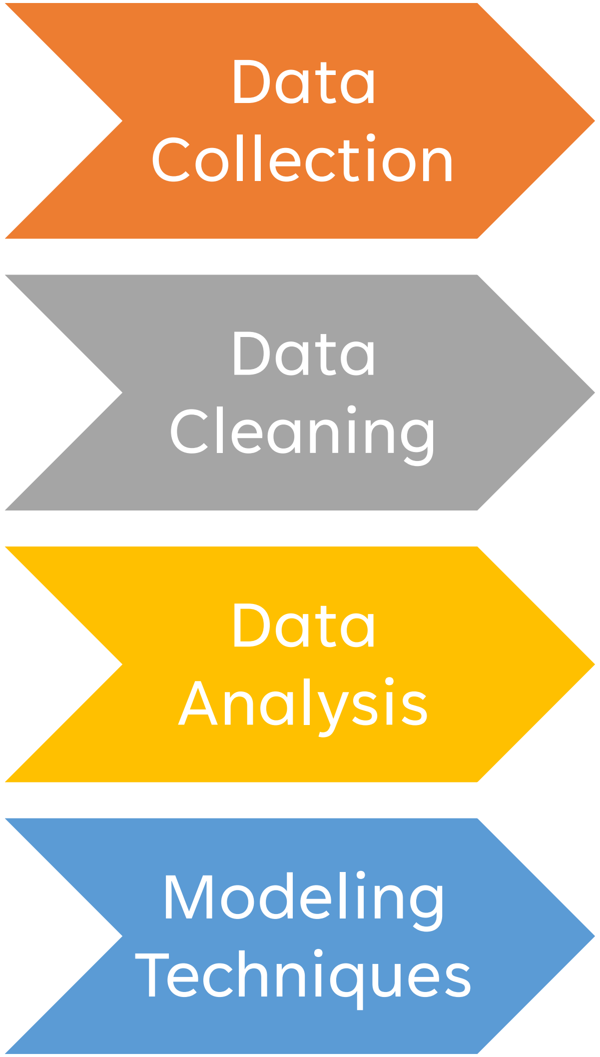
Financial Costs

Delays result in direct costs, such as passenger compensation and refunds.
Additional expenses arise from increased fuel consumption and the need for flight rescheduling.

Reputational Risks

Frequent delays can damage an airline's reputation and erode customer trust.
Negative experiences may lead to decreased customer loyalty and a loss of market share.

DATA ANALYSIS AND MODELING



DATA COLLECTION

Data Collection → 2018 USA Domestic Flights → Airline and Airport Names

Data Cleaning

Data Analysis

Modeling Techniques

- Flight Information
- Airline Names
- Airport Names

→ Combined Data Frame

DATA CLEANING

Data Collection

Data Cleaning

Irrelevant Data

Missing & Incorrect Data

Data Analysis

Modeling Techniques

	Date	Airline	Flight_number	Origin	Dest	Planned_depart_time	Actual_depart_time	Dep_delay
2124	2018-01-01	9E	3789	DFW	JFK	09:00:00	09:00:00	NaN
2132	2018-01-01	9E	3838	DTW	CWA	19:55:00	19:55:00	NaN
2179	2018-01-01	9E	3945	DTW	IAH	20:07:00	20:07:00	NaN
2188	2018-01-01	9E	3967	EWR	MSP	16:30:00	16:30:00	NaN
2248	2018-01-01	9E	4047	ATL	FAR	21:20:00	21:20:00	NaN

Data Collection

Data Cleaning

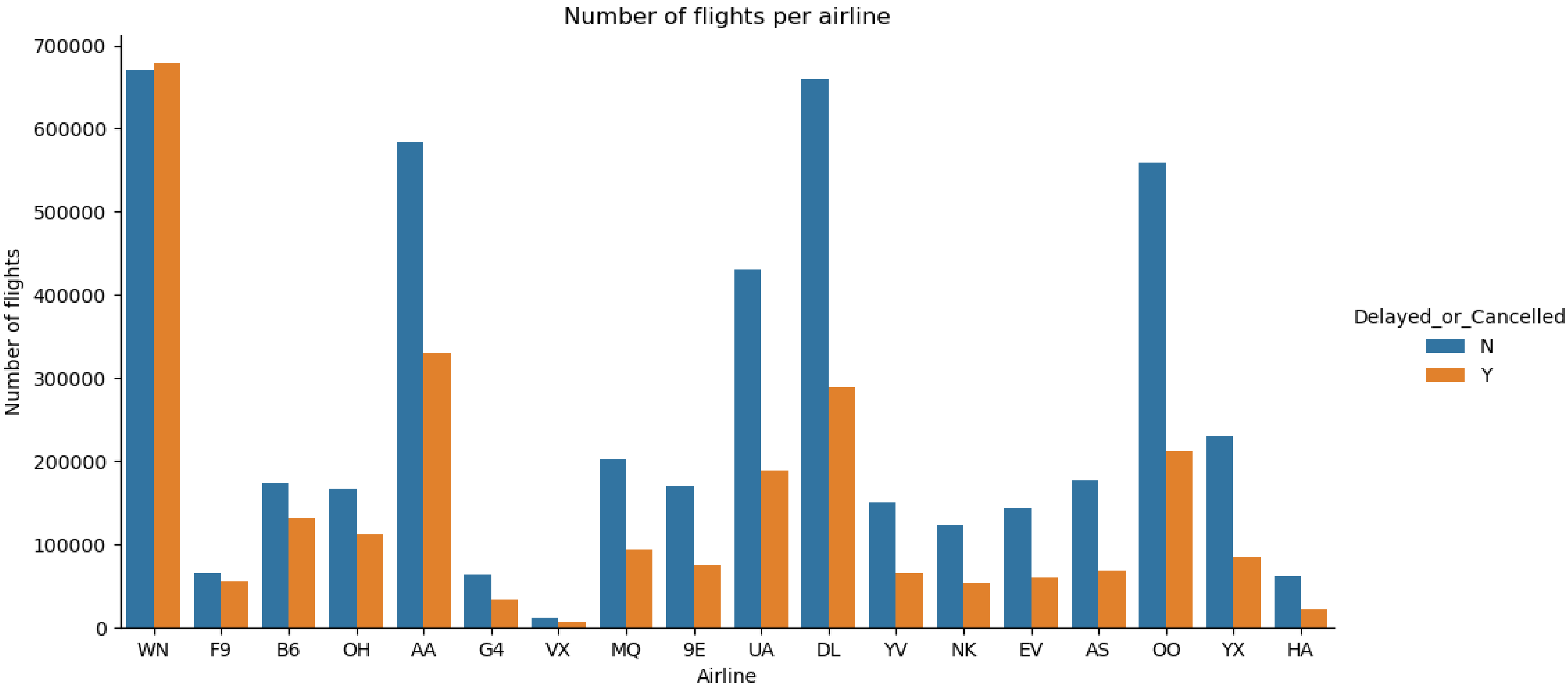
Data Analysis

Airline & Airports

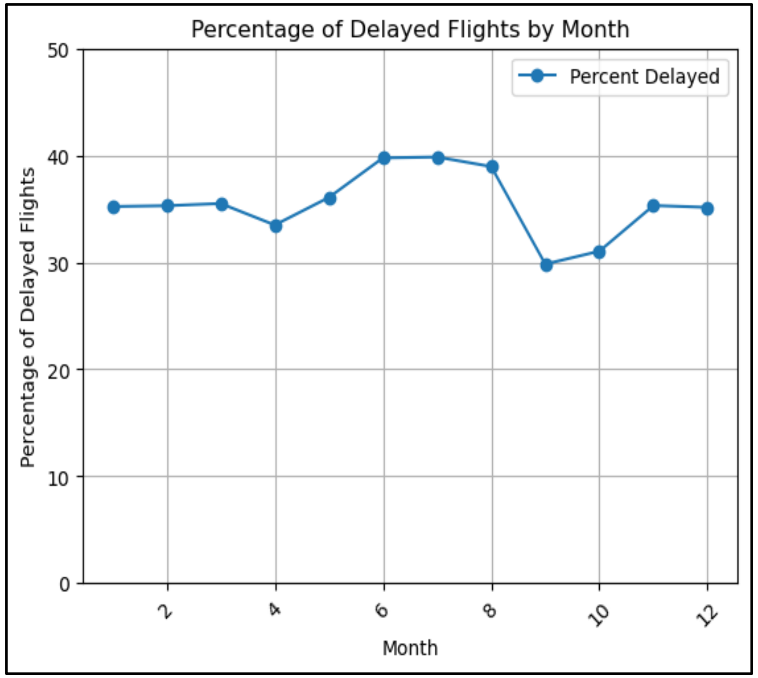
Day, Month, & Time

Modeling Techniques

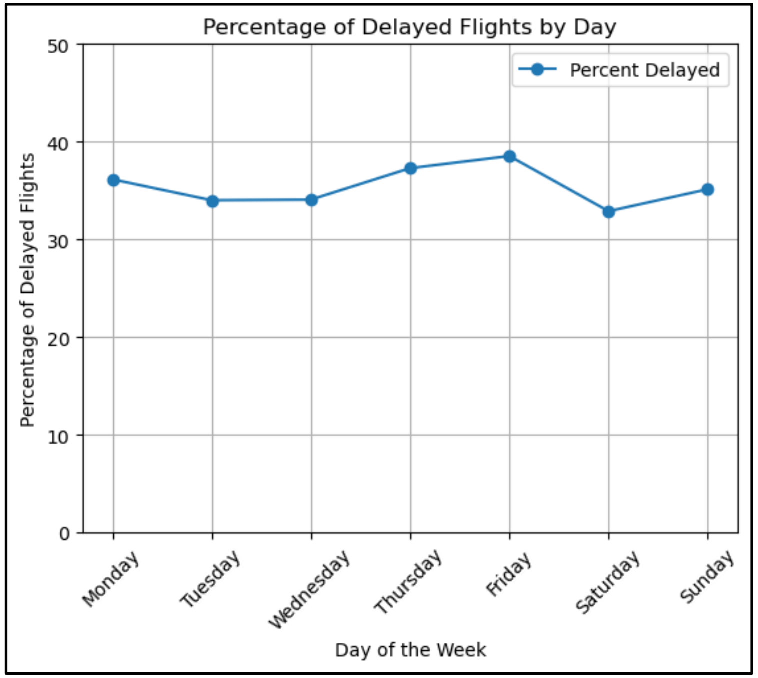
DATA ANALYSIS



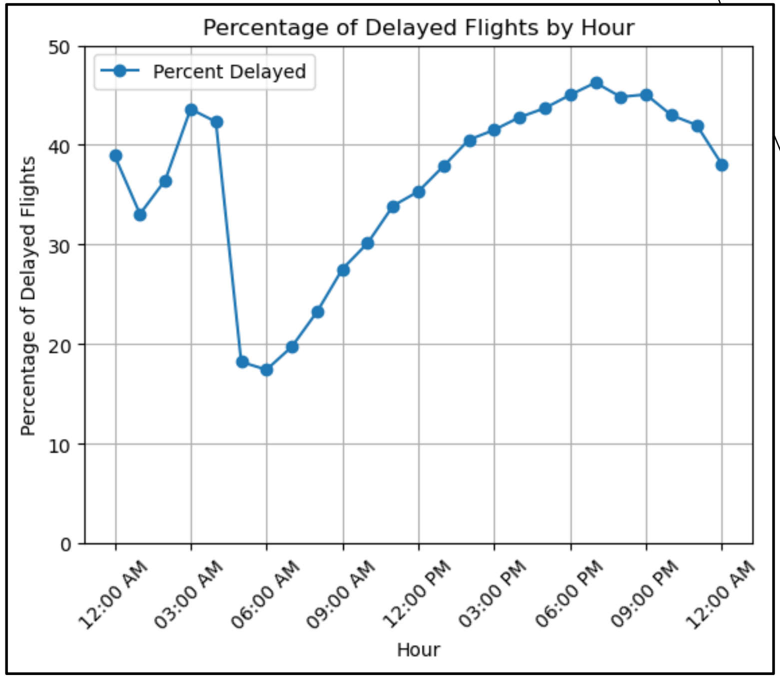
DATA ANALYSIS



Summer Months have the most delays



No noticeable difference between day of the week



More delays as the day progresses

MODELING

Data Collection

Data Cleaning

Data Analysis

Modeling Techniques

KNN & Random Forest

LightGBM & XGBoost

K Nearest Neighbors	Random Forest	LightGBM & XGBoost	LightGBM Categorical
• Simple Baseline Model • Works well if there are strong similarities between flights	• Good at capturing nonlinear relationships • Good at capturing interactions between different factors	• More powerful than the tree based random forest model • Performed faster and better than the previous two	• Allows for categorical variables, eliminating the need for one-hot encoding • Added cyclical encoding to address day/year rollover

MODELING

Data Collection

Data Cleaning

Data Analysis

Modeling Techniques

KNN & Random Forest

LightGBM & XGBoost

MAE: 12.8 Minutes

	Actual	Predicted	difference
5	17.0	18.173099	-1.173099
15	5.0	-0.827005	5.827005
16	39.0	20.761955	18.238045
17	12.0	15.407304	-3.407304
23	23.0	12.741737	10.258263
25	1.0	1.299814	-0.299814
35	41.0	25.097353	15.902647
41	2.0	4.340073	-2.340073
42	12.0	-0.705812	12.705812
44	46.0	28.299812	17.700188

RECOMMENDATIONS

Advance Delay Notifications

Inform passengers early, considering ±15 minutes error margin.

Optimize resource allocation

Schedule personnel based on predicted delay lengths.

Flight Schedule Adjustments

Analyze historical flight data to identify and adjust consistently delayed flights

FUTURE IMPROVEMENTS FOR FLIGHT DELAY PREDICTIONS

Collect More Data

More data and features can help identify relationships and interaction between different factors, improving model performance. This can include:

- Reason for the delay
- Weather Conditions
- Data from multiple years

Track Flight Patterns

Tracking the same flights throughout the year to identify if there are specific flights that are always delayed.

Tracking the same plane throughout the day to determine if there is a cascade effect that contributes to delays.

CONCLUSION

Predictive analytics empower airlines to take control of delays.

A data-driven approach can improve efficiency, customer satisfaction, and therefore profitability.

With data, the sky is the limit.