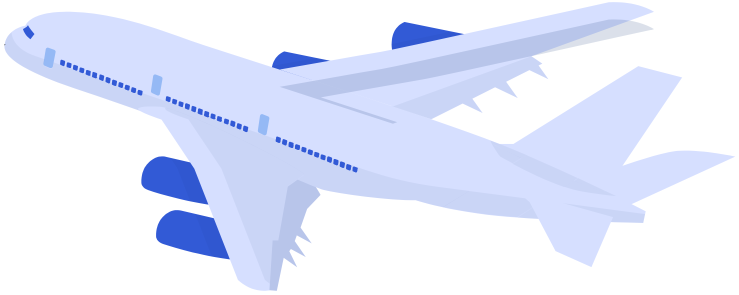
Late Flight Predictions



Linus Jen

Table of contents

1

Overview

Predicting Late Flights

- Given *flight_history* and *flight_test*, predict late flights
- Temporal aspect (date + time), airline and flight info provided
- Must create two meta-variables to append to *flight_test*
- *flight_history* has certain variables missing

Business Impact

Scheduling
Reduce staffing issues and ad-hoc plane switching

Customer Satisfaction
Greater trust leads to customer loyalty and brand loyalty

Cost Reduction
Reduce added costs such as flight credit, refunds, and labor costs

EDA

2

Modeling

Train w/ flight_history

Perform EDA, create meta-variables, and use CV to find "best" performing model

Test w/ flight_test

Treat as holdout/burn set, only use with best model

Data Overview

flights_history

Delays
Dept and arrival delay info show high correlation

Clean data
Generally, most columns are populated

Dist & airtime
Greater distance & airtime could have greater variability

Target
lateflight not provided!

Creating lateflight

1. ***lateflight* if *arr_time* > *sched_arr_time* (64,712 flights, ~39%)**
2. *lateflight* if *arr_delay* > 0 (67,370 flights, ~40.5%)

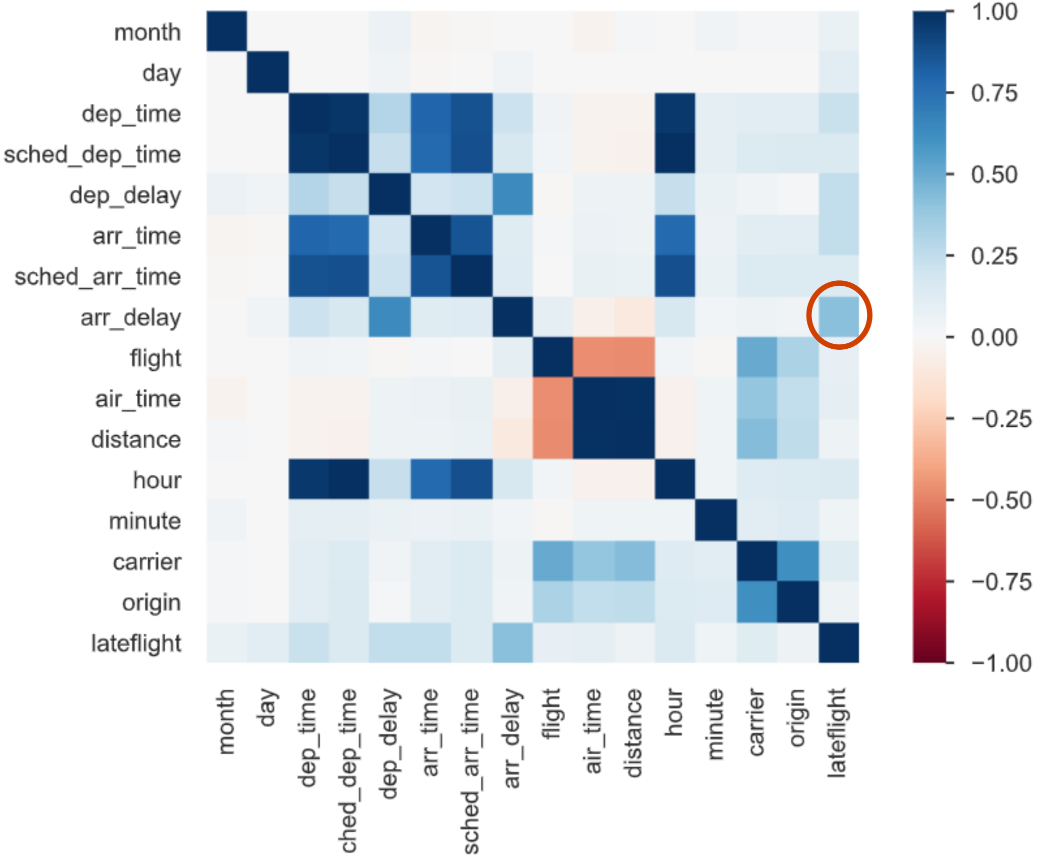
- Emphasize flights that are very late

Flight Delay Breakdown (min)

<5 min	5-15 min	15-30	30-60	60-120%
17%	23.6%	19.3%	18.2%	21.9%

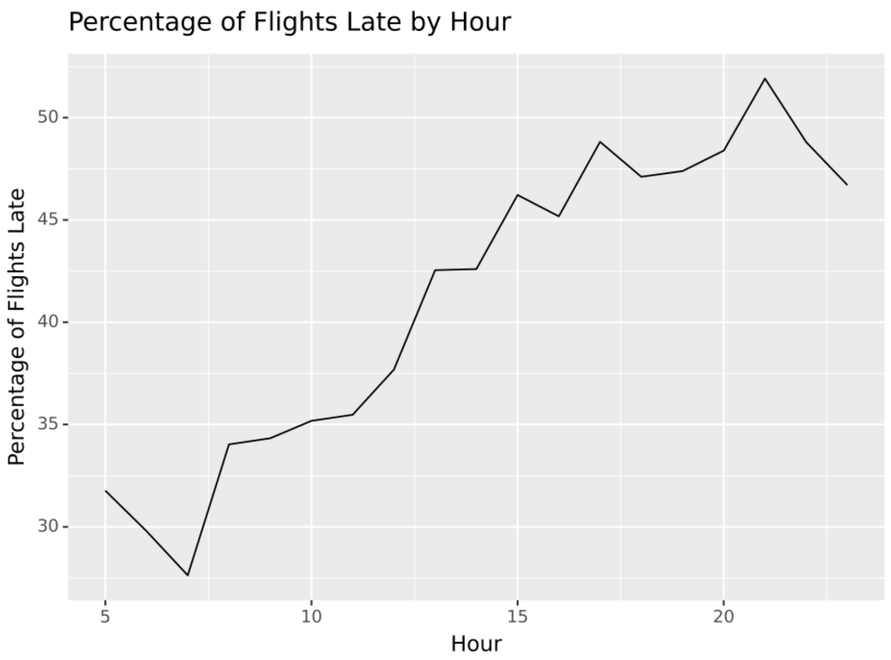
~60%!

Correlation Matrix



Dep & Arr Times

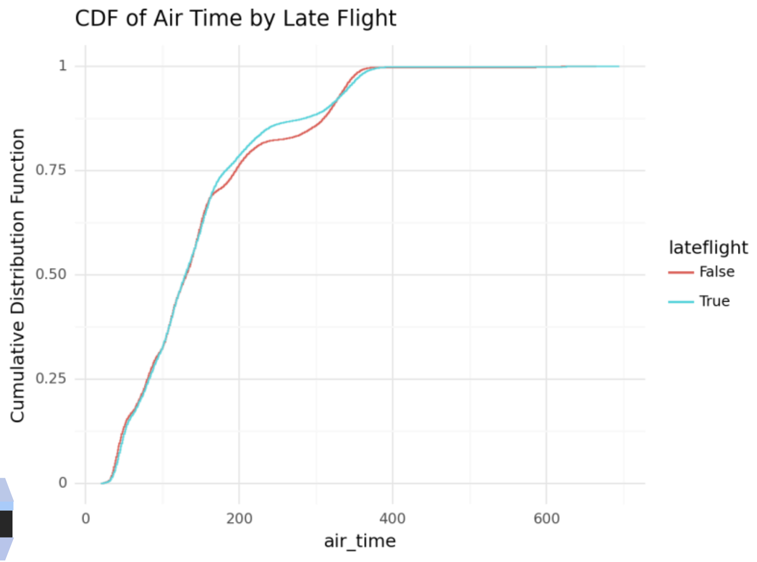
Departure Flights



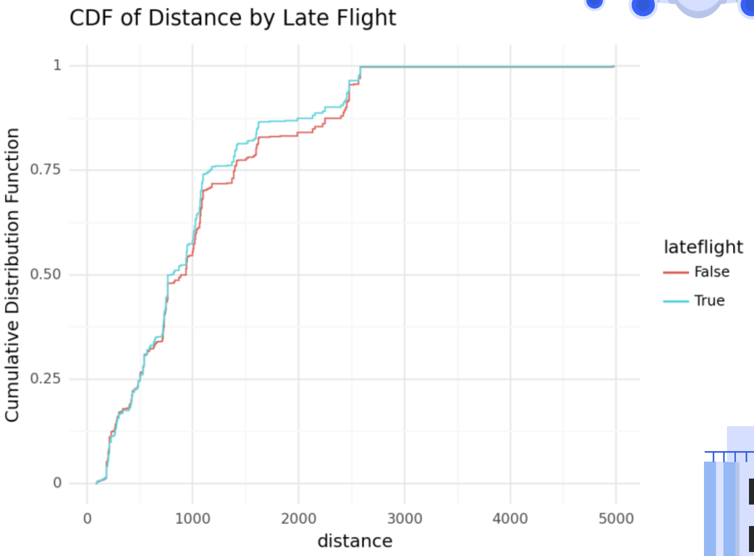
Notable Callouts

- Carriers F9 (Frontier) and FL (AirTran Airways) > 50%, while HA (Hawaiian Airways) < 25%
- EWR (Newark) > 43%, while JFK and LGA (LaGuardia) < 39%
- High variation amongst destination airports, including OKC and TUL (Tulsa) > 60%
- Thursday & Friday had the most late flights (>44%)

Air Time



Distance



Not much difference for late flights in distance and air time

Meta-Variables

Goal: Avoid OHE & lots of categories!

Flight Info

Capture flight lateness tendencies without categorization

- % flights late
- Median delay (min)

Destination

Capture destination lateness

- % flights late

Carrier

Capture carrier reliability

- % flights late

3

Modeling & Results

Methodology

- Using *flight_history*, create and connect meta-variables
- Get dummy variables
- Split data (80/20)
- SMOTE + undersampling for class imbalance
  - Because of SMOTE, won’t report fit precision, recall, F1, AUC
- Fit model (XGBoost) to improve AUC
- Predict on test set to get metrics

Variables

	sched_dep_time	sched_arr_time	weekday	carrier_perc_late	median_arr_delay	flight_num_perc_late	dest_perc_late
0	515	819	1	38.481	-7.0	35.897	42.728
1	529	830	1	38.481	-5.0	34.848	42.728
2	540	850	1	33.205	-11.0	29.457	36.314
3	545	1022	1	44.520	-4.0	35.484	46.122
4	600	837	1	34.094	-9.0	23.711	45.093

Train Results

Precision	Recall	F1
0.51	0.6	0.55

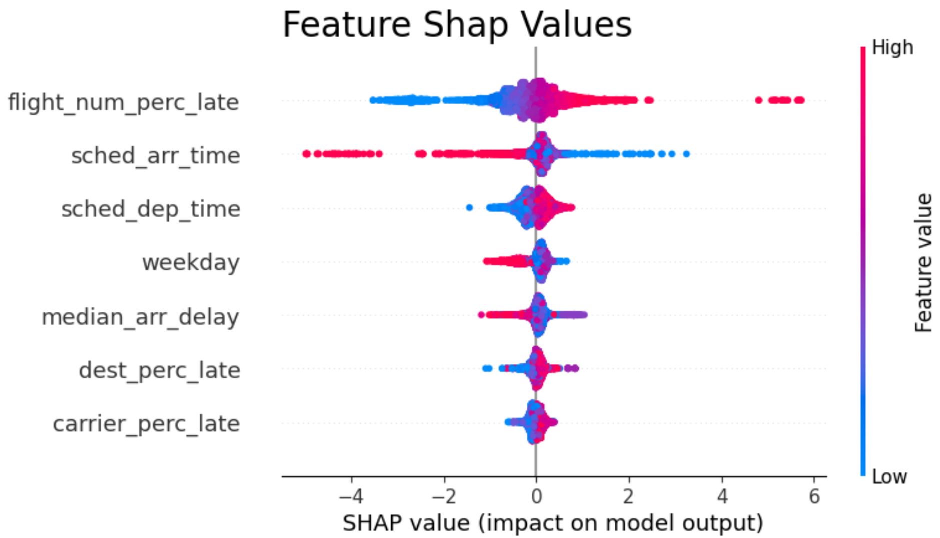
Precision	Recall	F1
0.54	0.57	0.55

	precision	recall	f1-score	support
0.0	0.63	0.60	0.61	3302
1.0	0.54	0.57	0.55	2698
accuracy			0.58	6000
macro avg	0.58	0.58	0.58	6000
weighted avg	0.59	0.58	0.59	6000

Feature Importance

Shap values give insight to black box models by showing prediction tendencies

- Percent of times past flights are late strong ties to late
- Early arrival times, late departures are more likely to be late
- Median delay for past flights contradicts expected findings



Next Steps

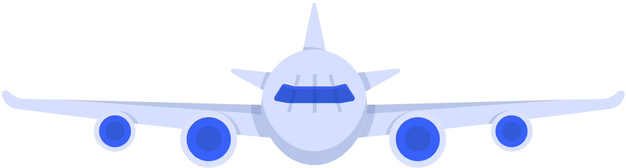
4

Limitations

Group categories with high variance in *lateflight* together

Only XGBoost was tested in the modeling process (without hyperparameter tuning)

Low metrics undeserving of any pilot plan



Thanks!

[GitHub]