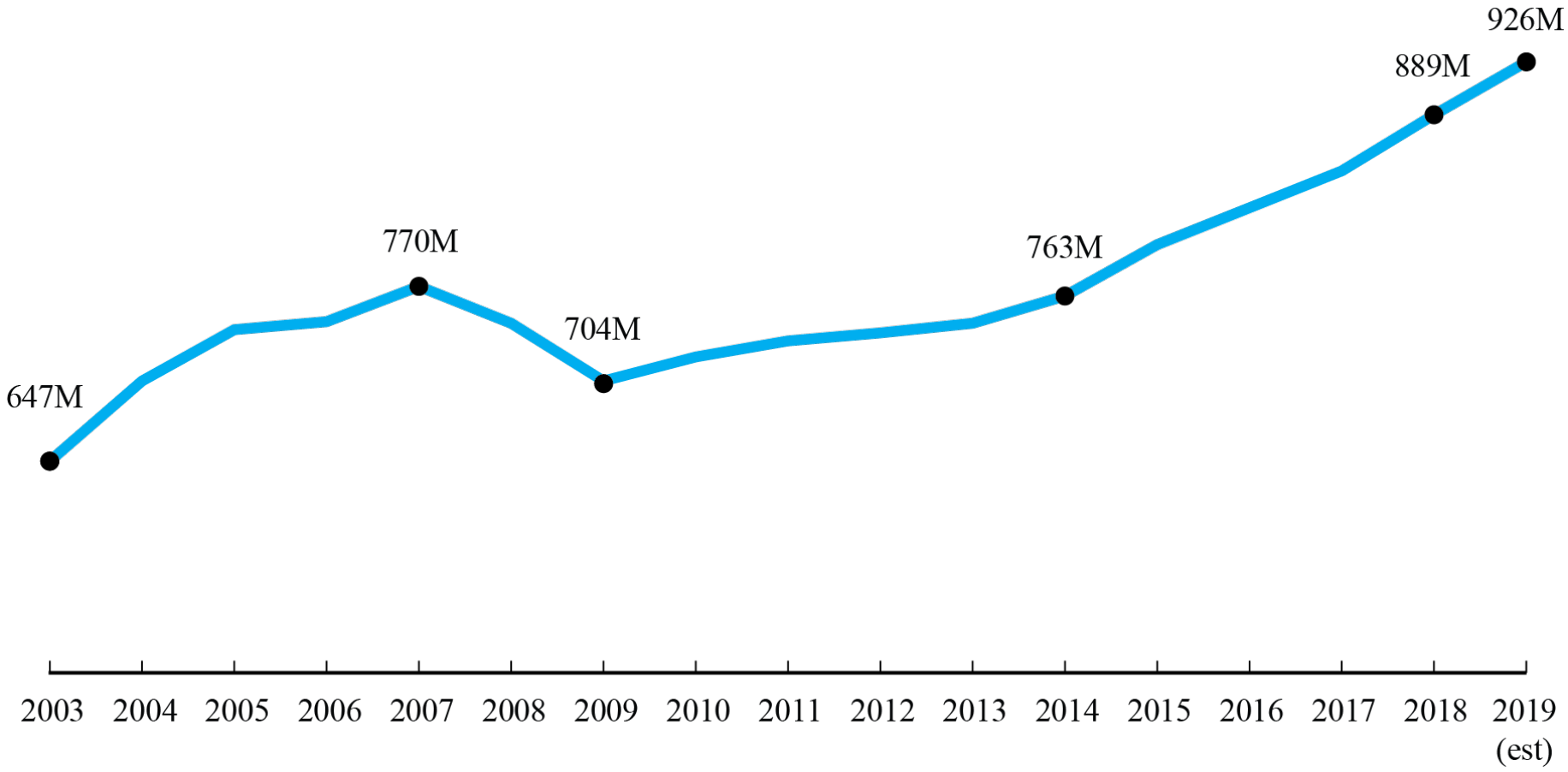
Optimizing flight choices for punctuality

Binh Hoang

Tremendous growth in domestic air travel



Source: Bureau of Transportation Statistics

Goals

- Build classification model to understand domestic flight delays
- Suggest tips for travelers looking to optimize their flight choices for punctuality

saving time = saving $

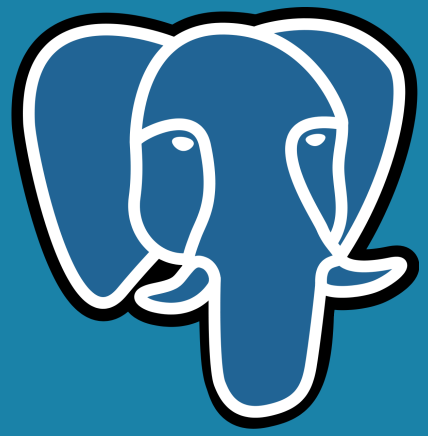
Methodology

Data: Kaggle 2015 U.S. flights (100k flights subset), Visual Crossing Weather

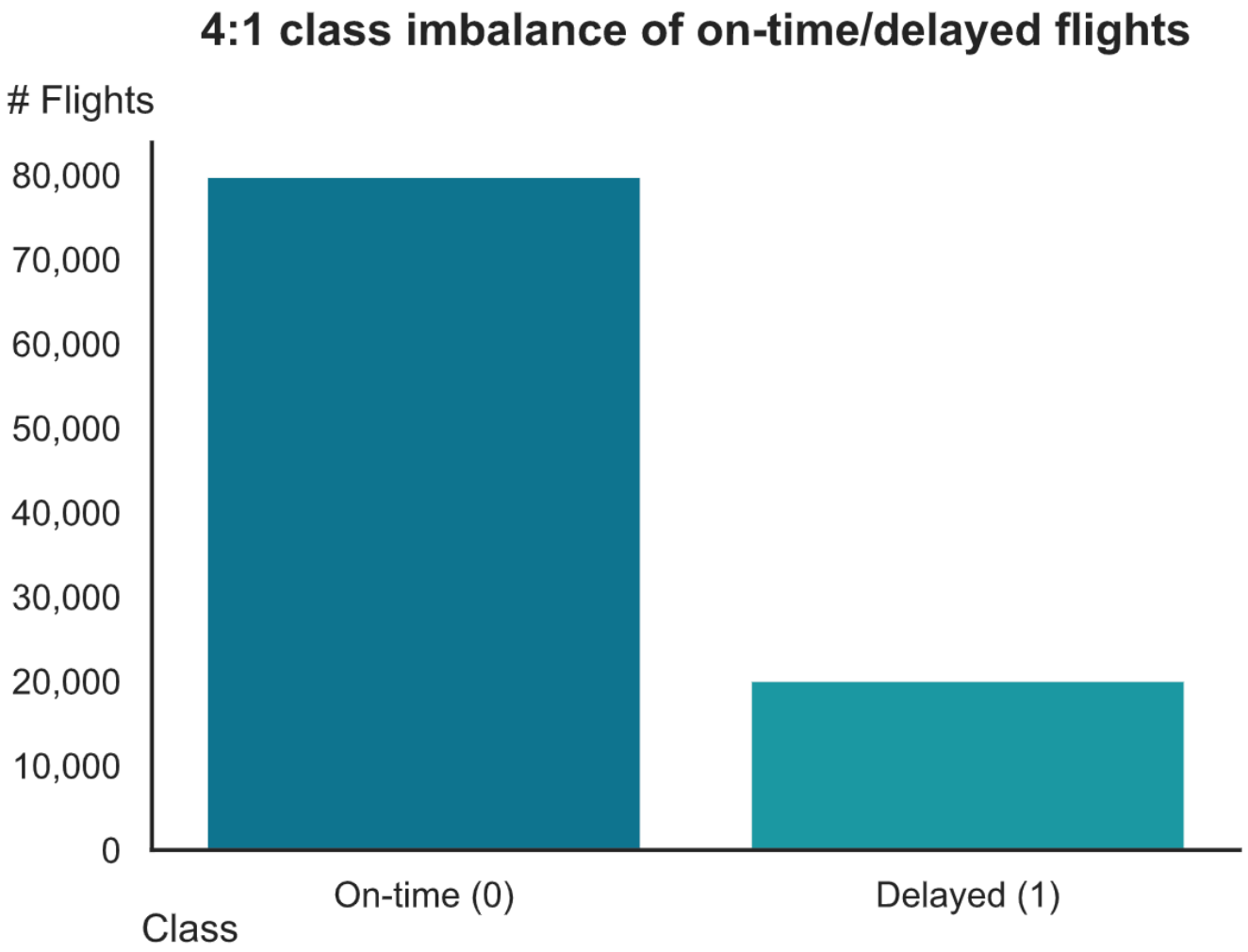
Tools: SQL, Tableau

Models: **logistic regression,** random forest, xgboost

Metric: F1 (precision/recall)



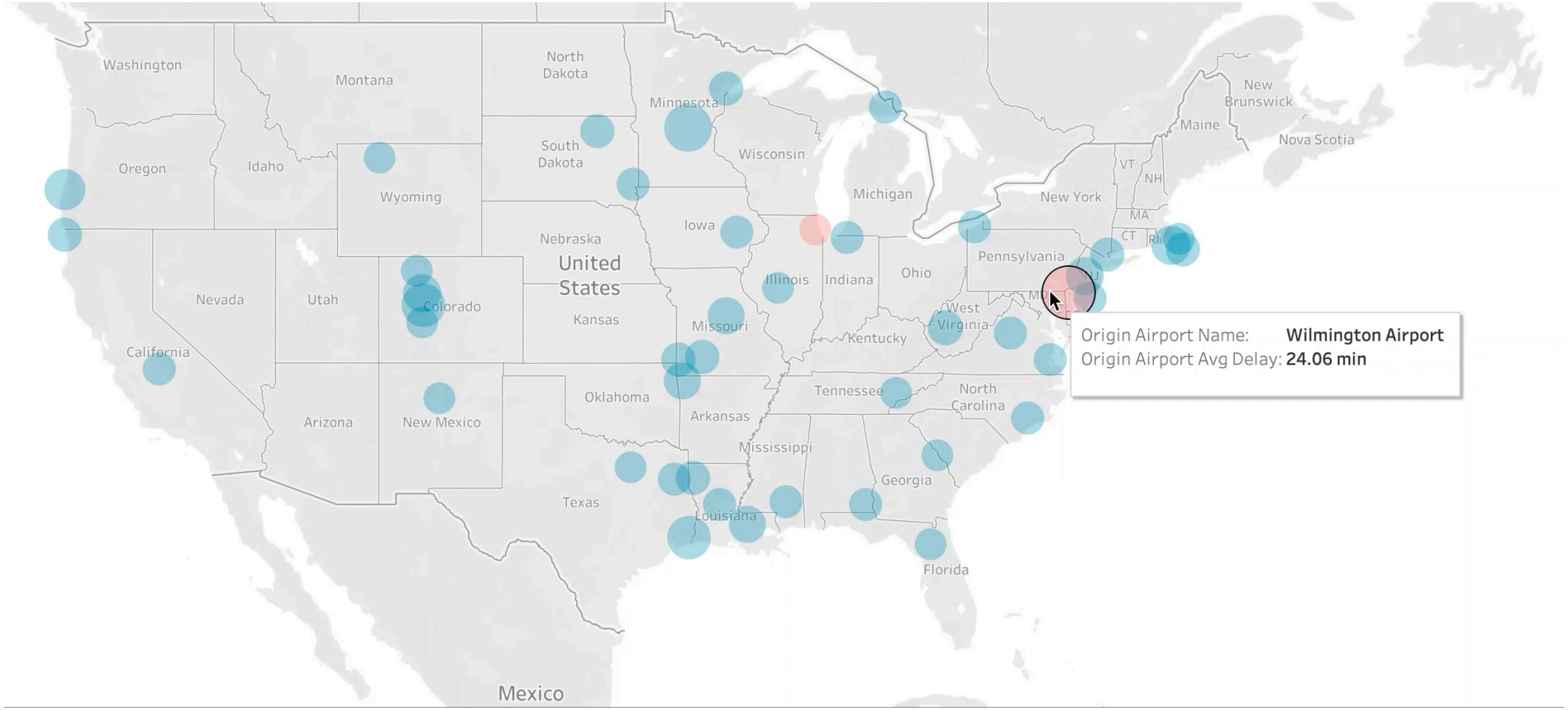
Class imbalance



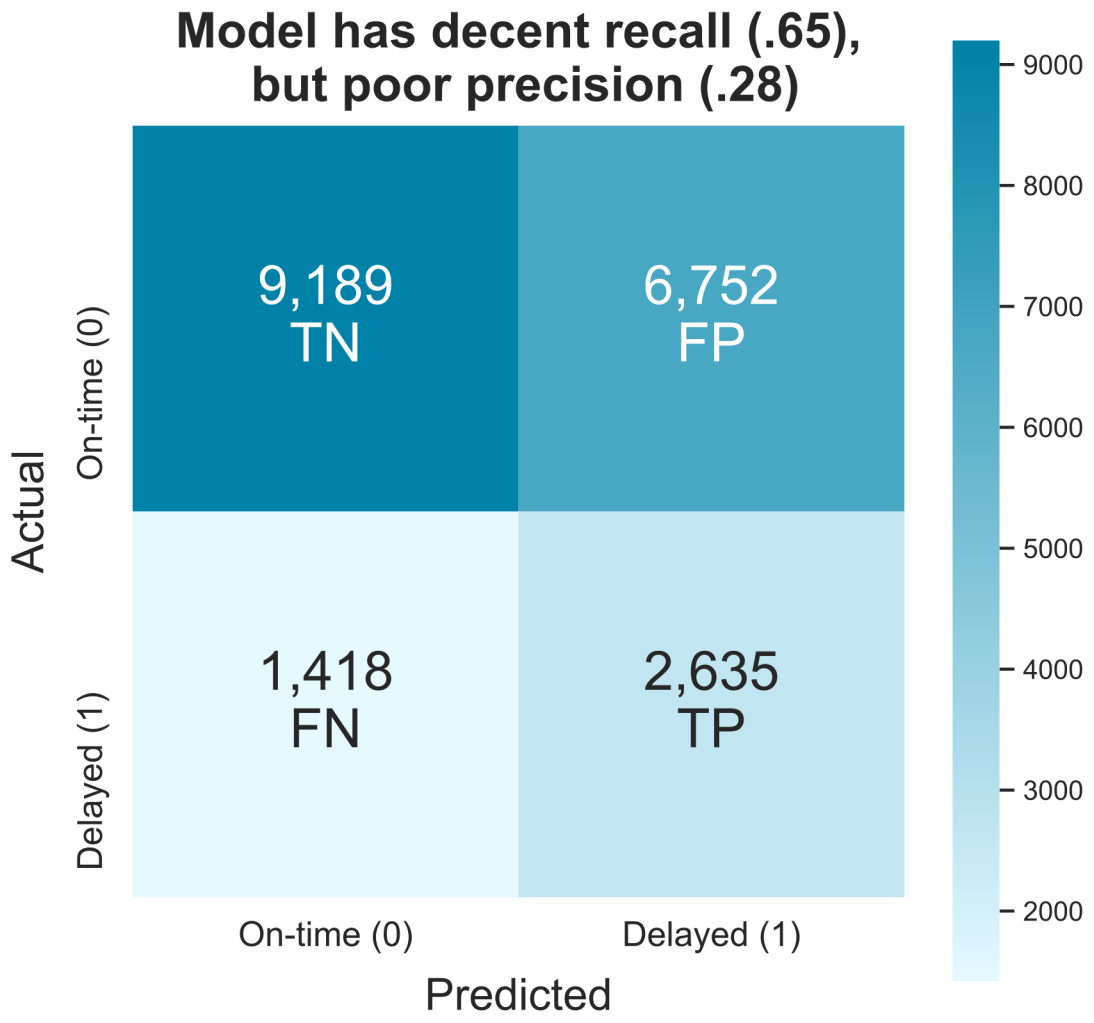
Strategies:

- Balanced class weights
- 47% threshold (to optimize F1)

Feature engineering example: origin airport avg arrival delay

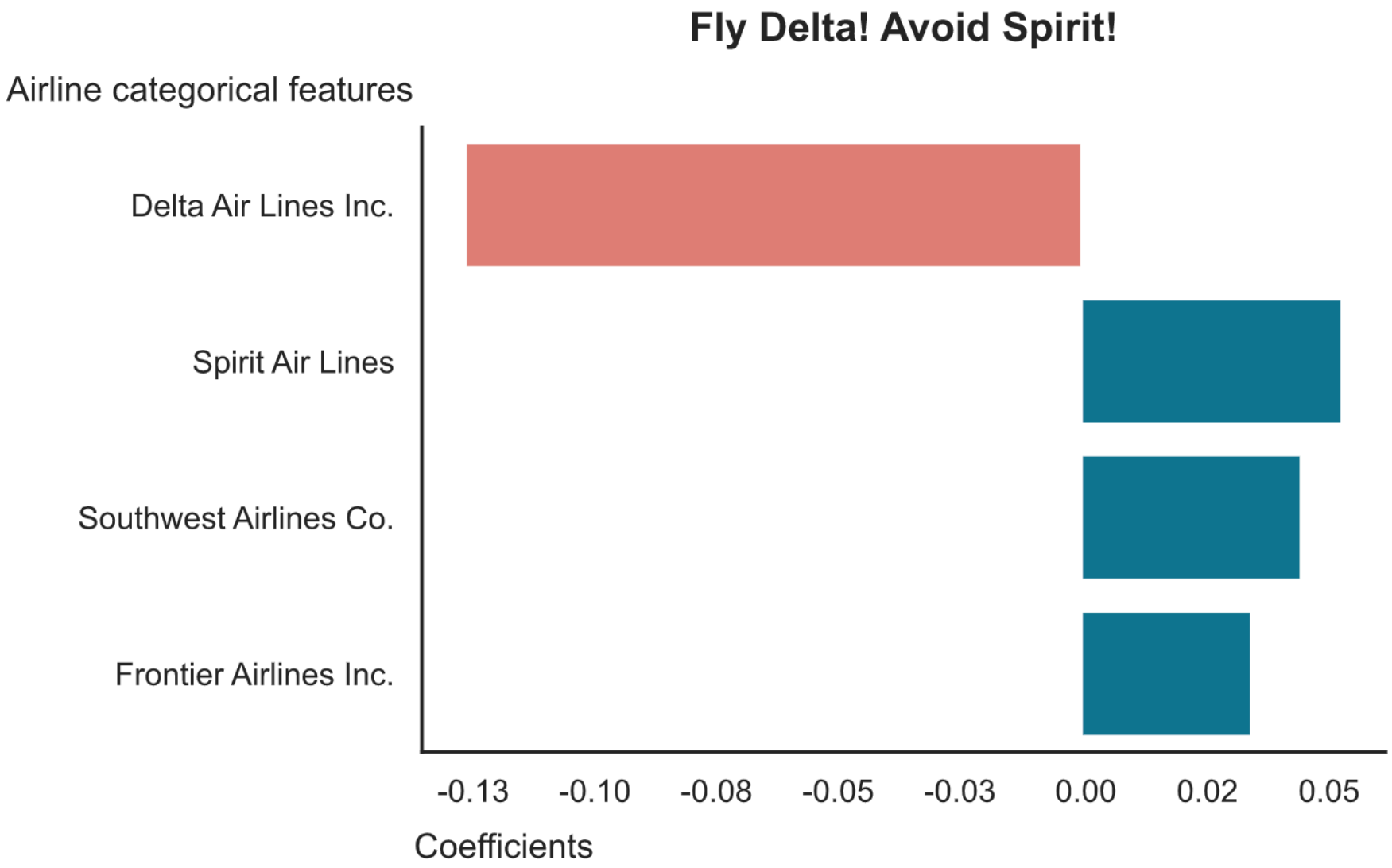


High bias/low variance

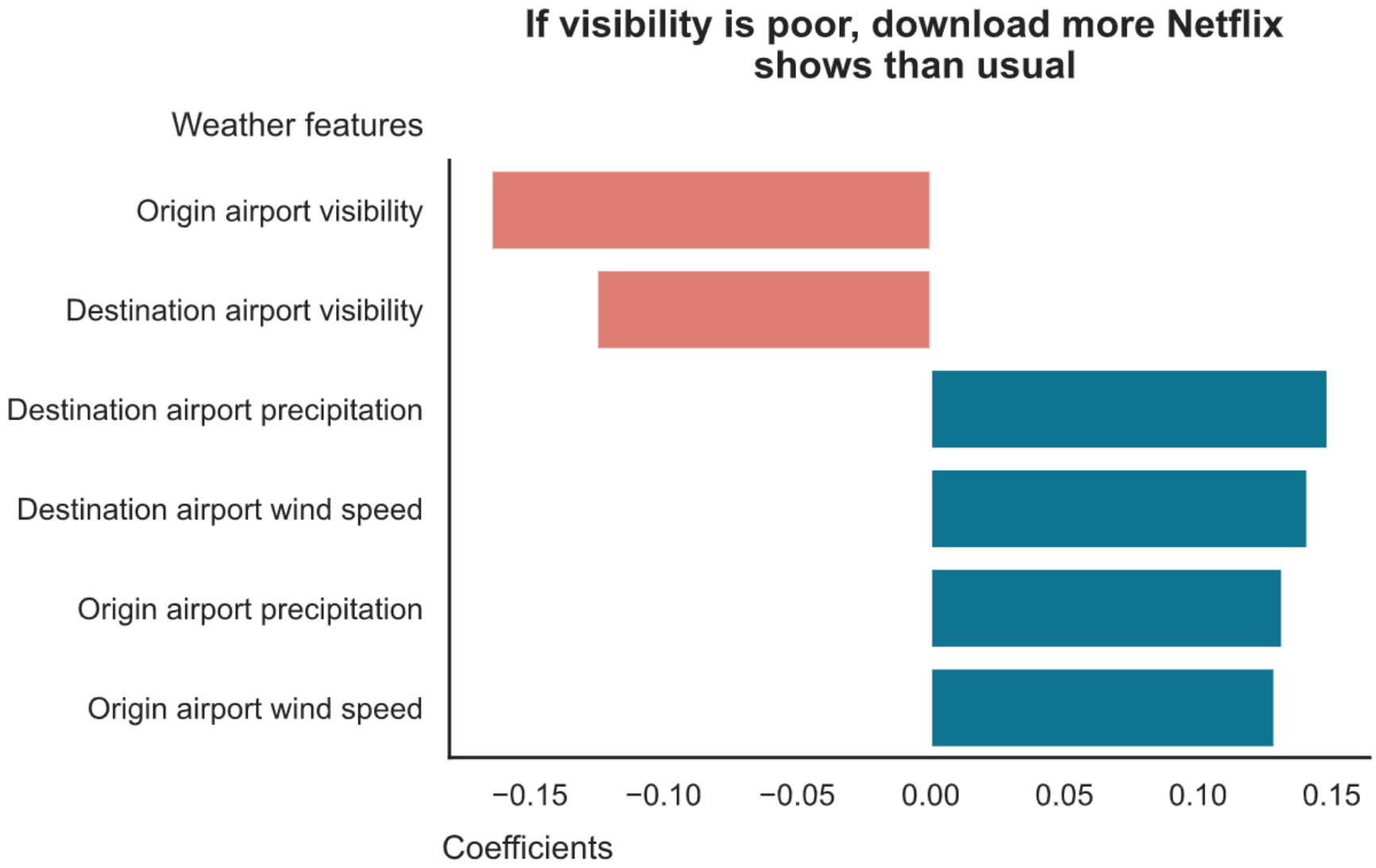


	Train	Test
F1	0.40	0.39
Precision	0.28	0.28
Recall	0.66	0.65

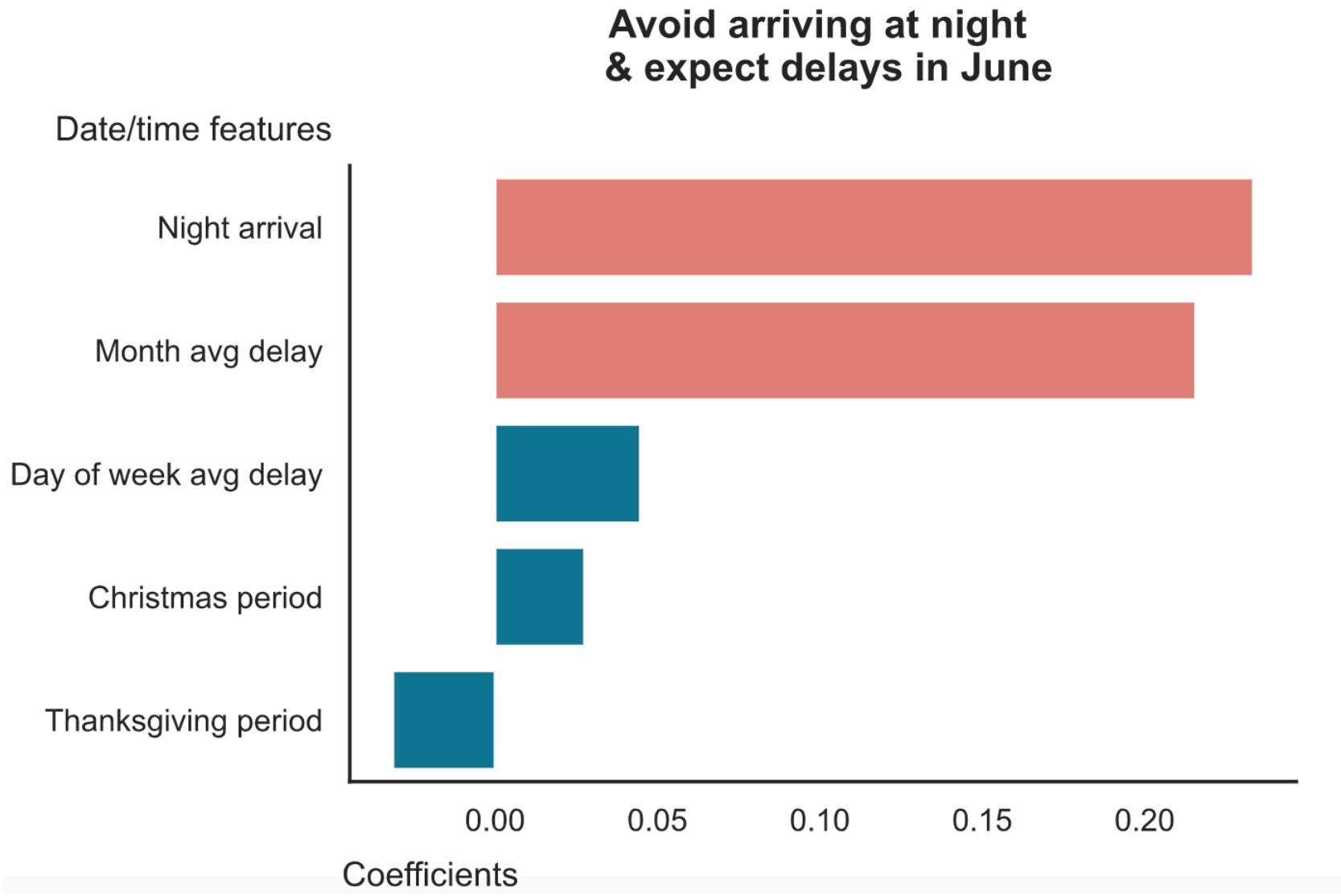
Tip #1


Fly Delta! Avoid Spirit!
Airline categorical features
Delta Air Lines Inc.
Spirit Air Lines
Southwest Airlines Co.
Frontier Airlines Inc.
-0.13
-0.10
-0.08
-0.05
-0.03
0.00
0.02
0.05
Coefficients

Tip #2


If visibility is poor, download more Netflix shows than usual
Weather features
Origin airport visibility
Destination airport visibility
Destination airport precipitation
Destination airport wind speed
Origin airport precipitation
Origin airport wind speed
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
Coefficients

Tips #3/4


Avoid arriving at night
& expect delays in June
Date/time features
Night arrival
Month avg delay
Day of week avg delay
Christmas period
Thanksgiving period
0.00
0.05
0.10
0.15
0.20
Coefficients

Mihashi made bad flight choices

✓Spirit Air Lines

✓June

✓Night arrival

(other features not listed here)

Delay probability:

73% – oh no!

Mihashi from Ookiku Furikabutte

Let’s do a flight makeover!

✓Spirit –> Delta

✓June –> May

✓Night arrival –> Day arrival

(other features not listed here)

Mihashi from Ookiku Furikabutte

Delay probability:

42% – much better!

Conclusion

Do:

- Fly Delta
- If visibility is low, download more Netflix shows than usual

Don’t:

- Fly Spirit
- Fly in June
- Fly at night
- Fly from Chicago O’Hare/Wilmington, DE

Future work

- Scale up using Google Cloud
- Improve F1 through additional feature engineering/modeling
- Create Flask for users to get flight delay probabilities from logistic model

Thank you! Fly safely and on time.

Appendix

Class definitions:

- In my code, is_not_on_time is the target where 1 is the positive class and 0 is the negative class.
- is_not_on_time = 1 if
    1. Flight is 15 minutes or more later than scheduled arrival time (matches Federal Aviation Administration's definition of flight delay); in code, where arrival_delay >= 15
    2. Flight is cancelled
    3. Flight is diverted
- Else is_not_on_time = 0

Appendix

Filling arrival_delay null values:

- arrival_delay is null where where flight was cancelled or diverted
- Because I featured engineered avg_arrival_delay at the airline/origin airport/destination airport aggregation level, I want to penalize those airlines/airports that cancel/diverted flights, so I filled arrival_delay null values with the 99th percentile of arrival_delay, which in my dataset was

Appendix

Additional resources to understand flight delays:

1. Understanding the Reporting of Causes of Flight Delays and Cancellations (https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations)
2. Why There Are So Many Flight Delays in the Summertime (https://lifehacker.com/why-there-are-so-many-flight-delays-in-the-summertime-1796465726)
3. Why You Should Never Fly at Night on These 30 Airlines (https://www.smartertravel.com/airlines-to-never-fly-night/)