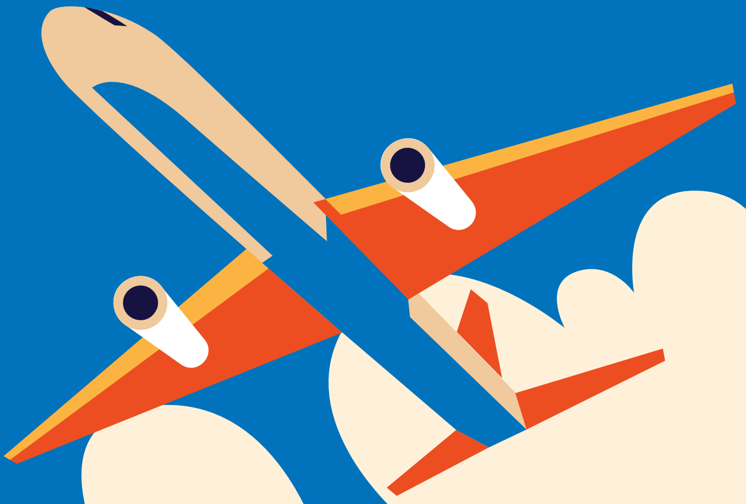
Estimating Flight Delays

DJ Ammirato, Gene Lam

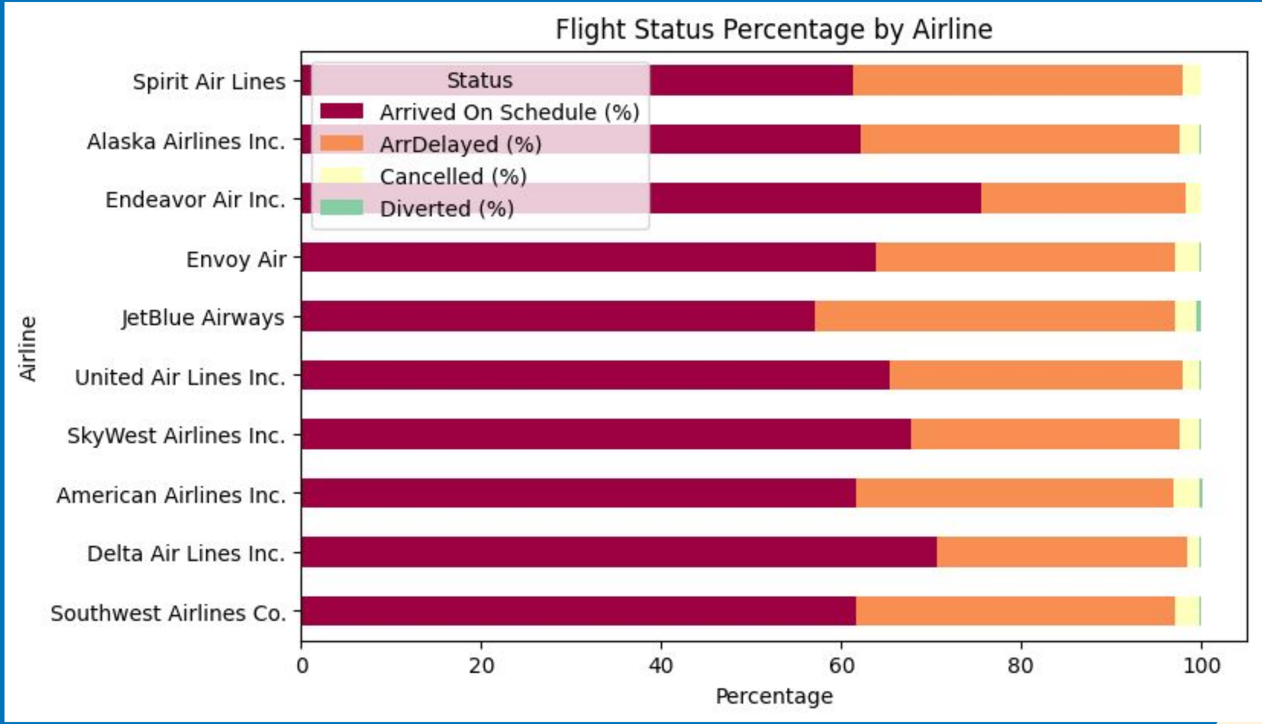


Objectives

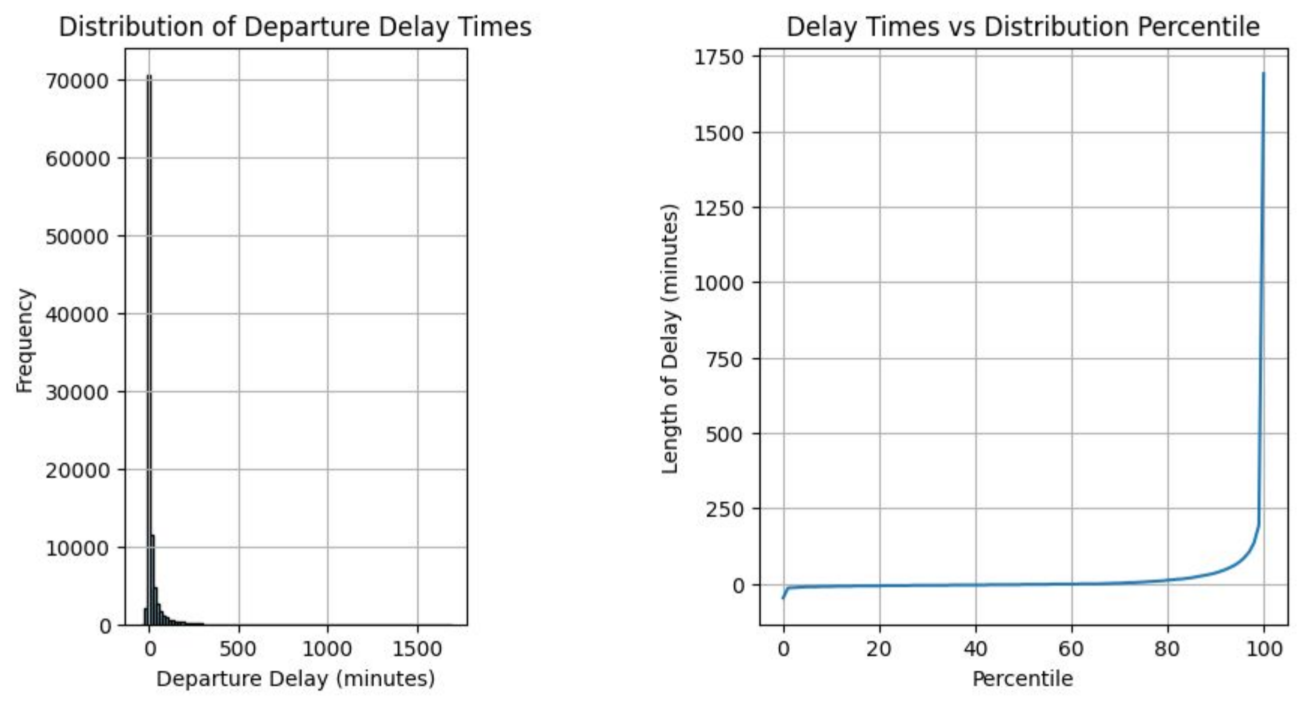
- Can we successfully predict whether or not future flights are going to be delayed? By how long?
- What features influences whether or not a flight is going to be delayed?
- Empower consumers to make more informed decisions when buying flight tickets
- Data comes from Kaggle and Bureau of Transportation Statistics
  - 2018 - 2023 US domestic flights (including US territories)
  - 39M rows, 61 features
  - We sample 100,000 rows and use 28 features

Exploratory Data Analysis

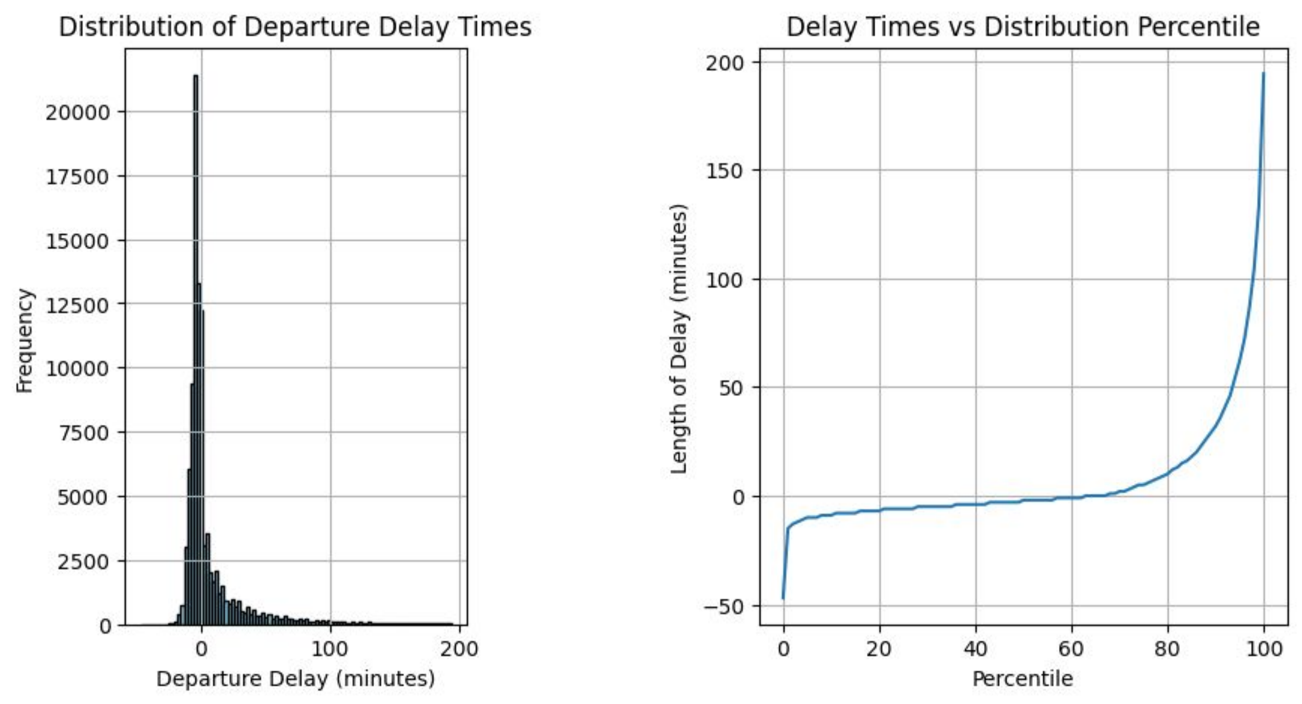
Top 10 Airlines in America



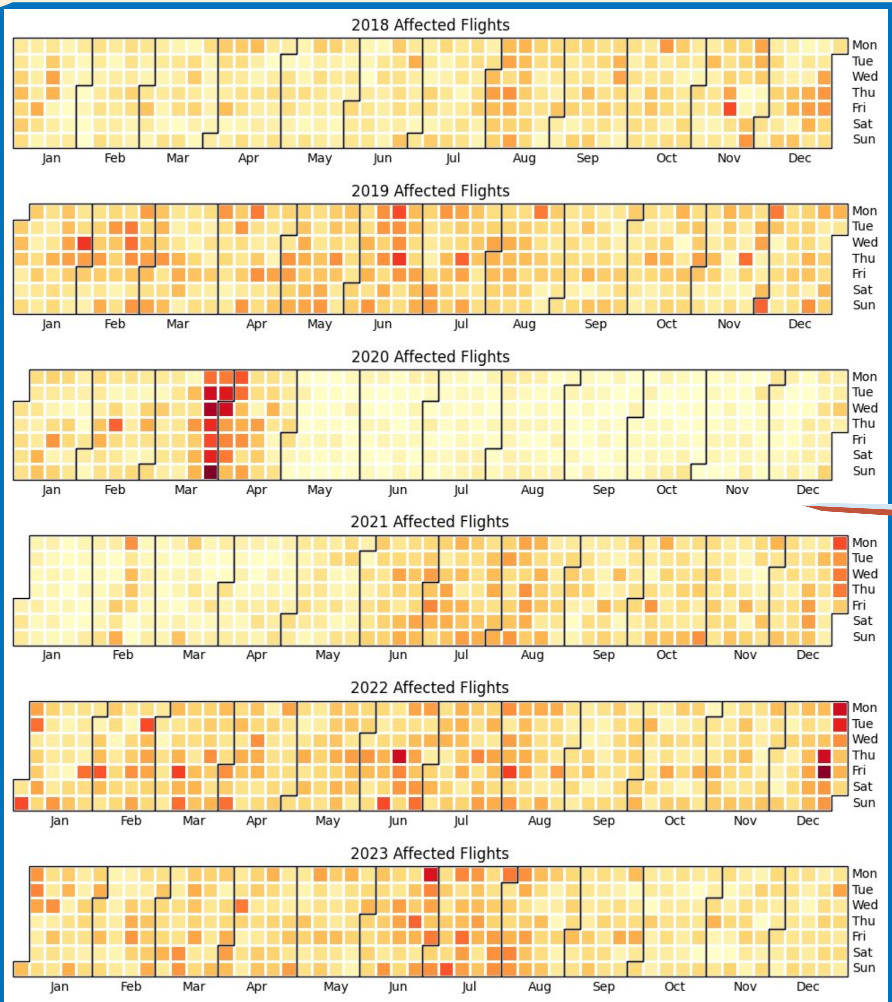
How long are we waiting?

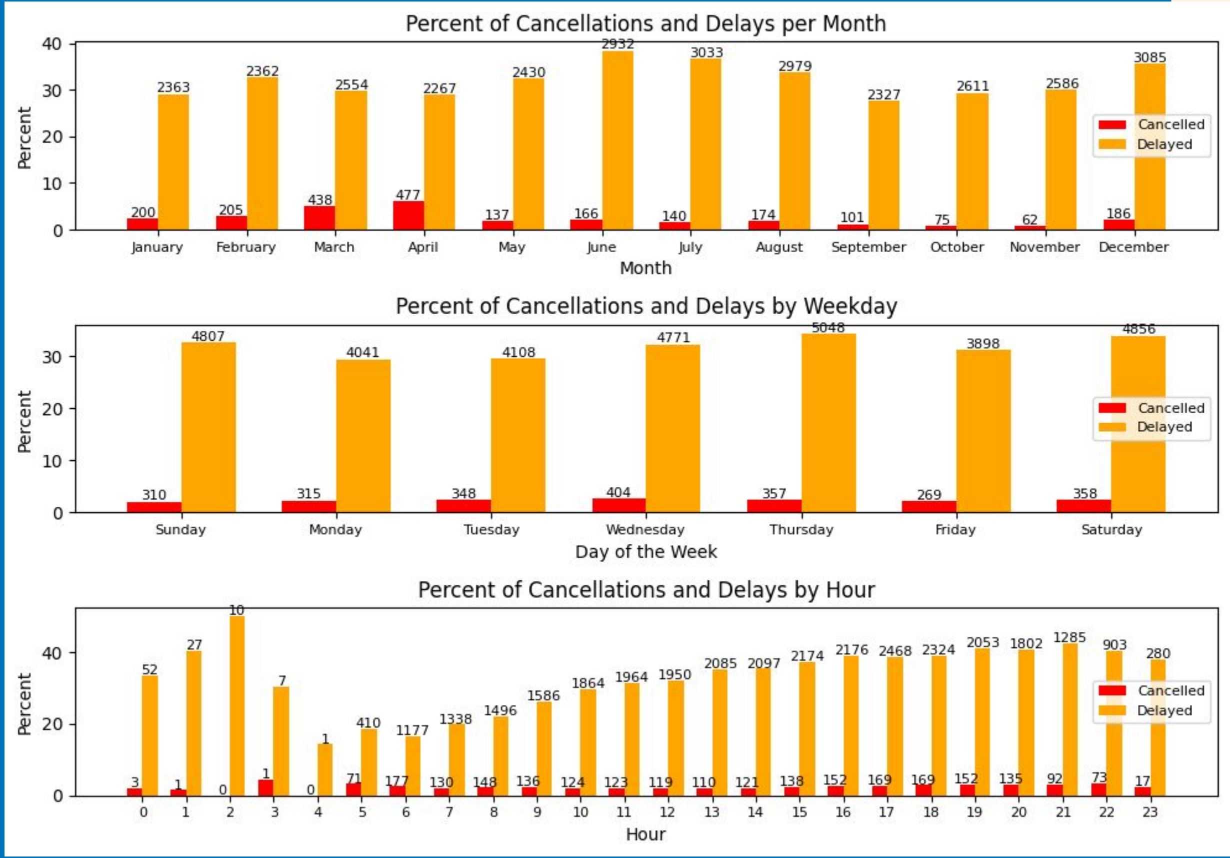


How long are we actually waiting?



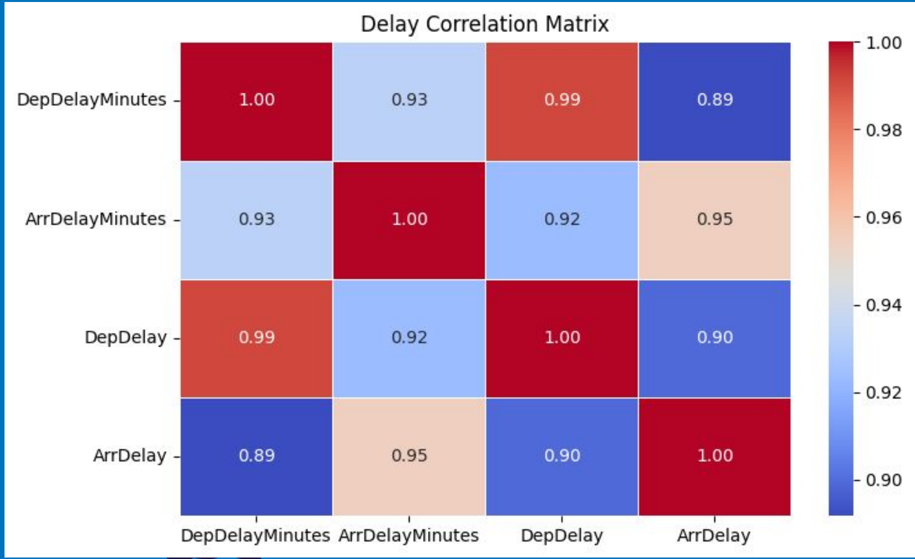
When are we waiting?

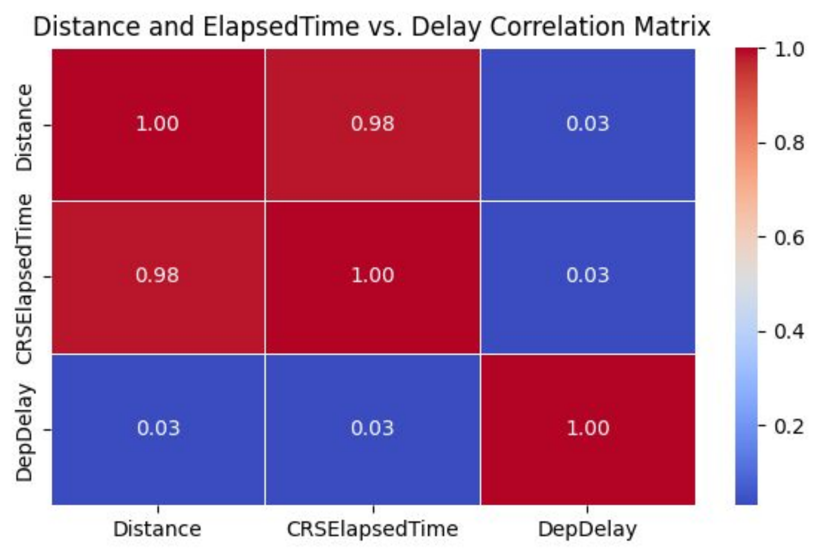




Percent of Cancellations and Delays per Month
Percent
Month
January
February
March
April
May
June
July
August
September
October
November
December
Cancelled
Delayed
2363
2362
2554
2267
2430
2932
3033
2979
2327
2611
2586
3085
200
205
438
477
137
166
140
174
101
75
62
186
Percent of Cancellations and Delays by Weekday
Day of the Week
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
4807
4041
4108
4771
5048
3898
4856
310
315
348
404
357
269
358
Percent of Cancellations and Delays by Hour
Hour
52
27
10
7
1
410
1177
1338
1496
1586
1864
1964
1950
2085
2097
2174
2176
2468
2324
2053
1802
1285
903
280
3
1
0
1
0
71
177
130
148
136
124
123
119
110
121
138
152
169
169
152
135
92
73
17

Correlation of Numeric Features





Modeling

Performance Metrics

Classification

- Recall: percentage of positive classifications correctly defined
  - Most important for value proposition
- Switched to F1 due to poor precision

Regression

- Mean-Squared Error: penalize larger errors to achieve high precision
- R2 score: conveys proportion of variance in delays that is conveyed by our features

Classification

Logistic Regression

59.59%

Training Accuracy

64.19%

Training Recall

24.91%

Training Precision

58.73%

Testing Accuracy

63.66%

Testing Recall

24.35%

Testing Precision

Random Forest

66.96%

Training Accuracy

67.95%

Training Recall

30.44%

Training Precision

63.80%

Testing Accuracy

57.53%

Testing Recall

26.12%

Testing Precision

Bayesian Optimization

- A method of automatically tuning the hyperparameters of a model
- Fits a series of models and compares their relative performance

Optimized Logistic Regression

59.56%

Training Accuracy

64.23%

Training Recall

24.90%

Training Precision

58.66%

Testing Accuracy

63.66%

Testing Recall

24.31%

Testing Precision

Random Forest

74.03%

Training Accuracy

76.52%

Training Recall

38.29%

Training Precision

65.93%

Testing Accuracy

53.48%

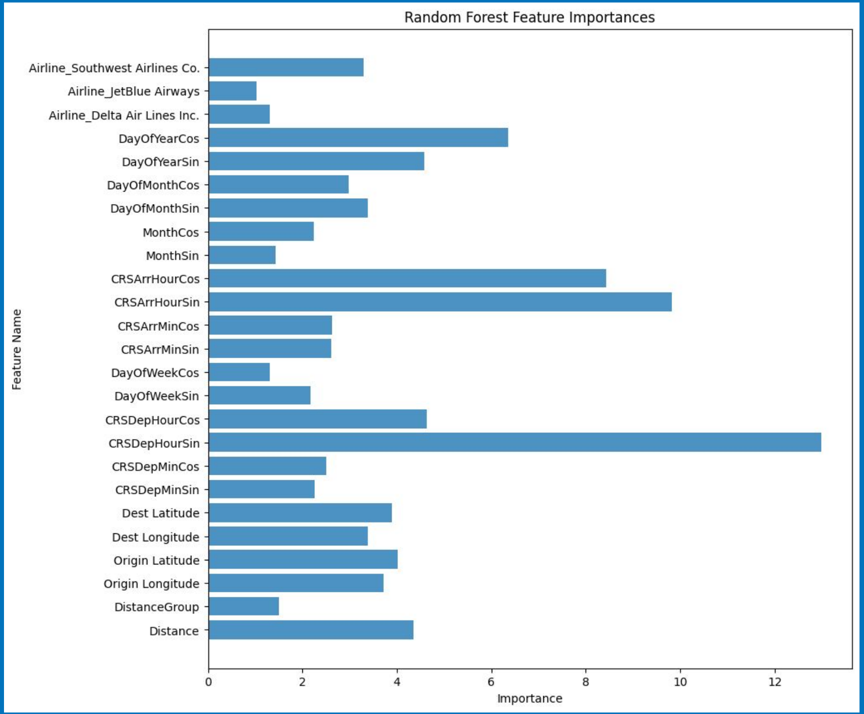
Testing Recall

26.81%

Testing Precision

Feature Importance

- Hour of scheduled flight departure and arrival
- Day of the year

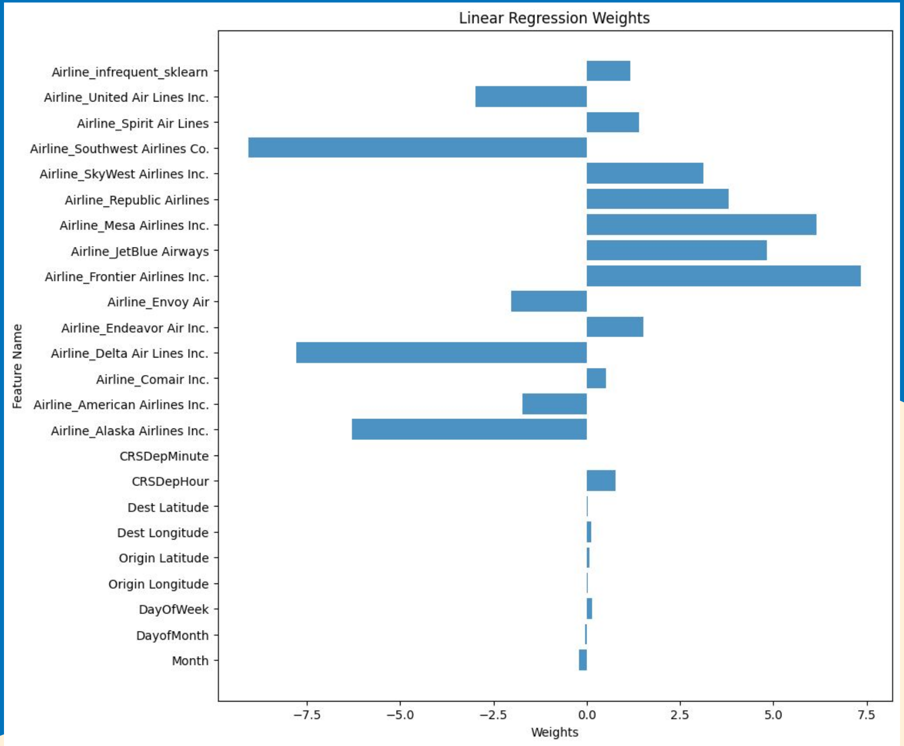


Regression

Linear Regression

Training
MSE: 1299.71
R2: 0.04

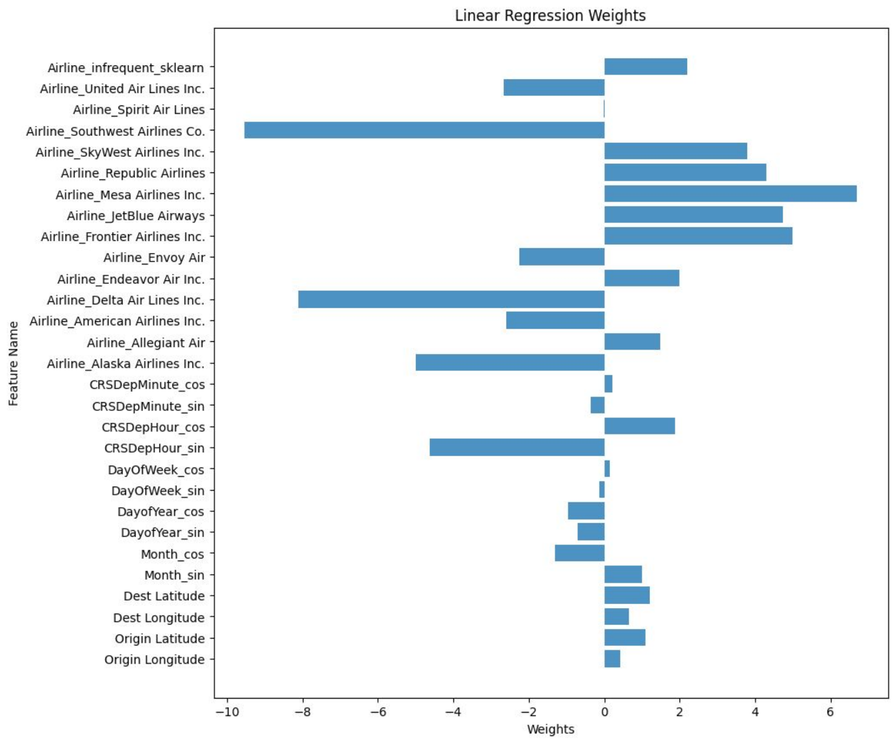
Test
MSE: 1299.71
R2: 0.04



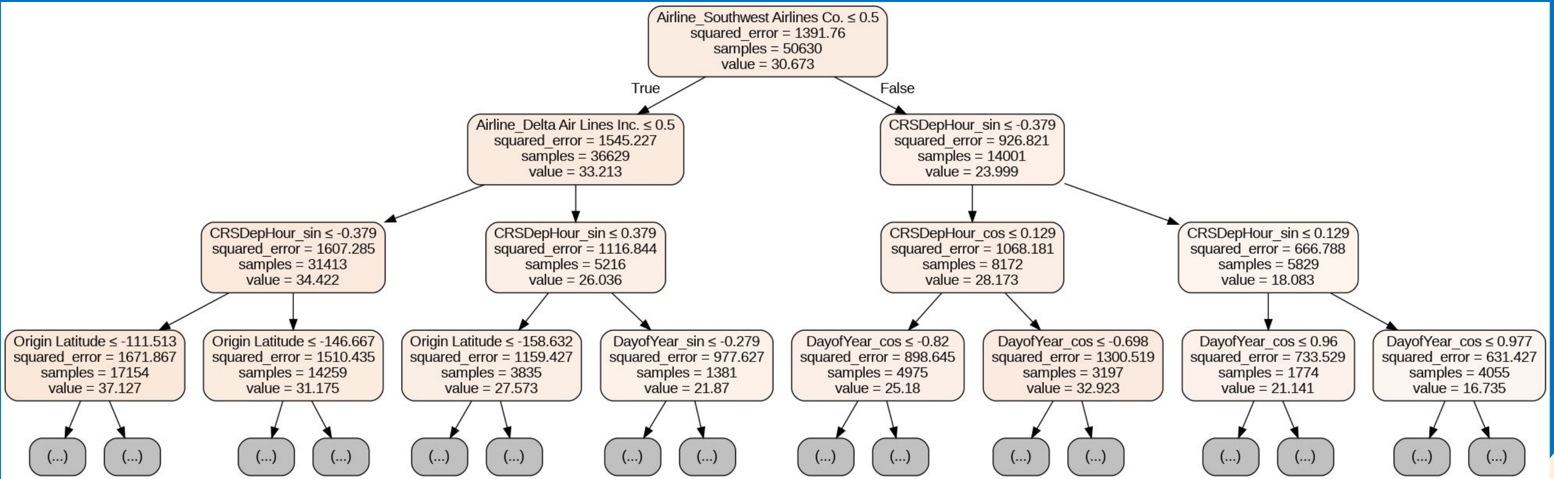
Optimized Linear Regression

Training
MSE: 1317.68
R2: 0.04

Test
MSE: 1317.68
R2: 0.04



Random Forest + Bayes Search



Training MSE: 1263.55
Training R2: 0.08
Test MSE: 1263.55
Test R2: 0.05

Implications and Insights

- Most important factors
  - Time of day
  - Day of the year
  - Airline
- These three factors are already major factors individuals consider when purchasing a plane ticket.
  - Information from our model can supplement what we already know.

Limitations and Future Work

- Could not utilize entire dataset
  - Required us to sample small fraction
- Limited number of features
  - Even after encoding

- Classifier for “Cancelled” flights
- Address overfitting in Random Forest
- Add more features:
  - Airline controversy
  - Make/model of the plane: Boeing vs Airbus
  - Ticket price
  - Number of seats sold

Reflections and Challenges

- Limited feature availability for our model
    - Even after encoding and feature engineering, our model effectively used airline, flight date, and airport location
- Should’ve spent more time on the EDA to identify more features
- Good practice of reviewing the entire course intimately
- Enjoyed data visualization

Thanks!