Automated Fight Rescheduling using Gradient Boosted Tree Regression

Joshua E. Jodesty

Background & Hypothesis:

Flight delays are an ongoing problem and are expensive and inconvenient for passengers, airlines, and Air Traffic Control (ATC). Developing a system that reduces flight delays involves an understanding of the complex relationships between airlines, ATC, and airports, as well as ATC's application of regulations stipulated by the National Airspace System (NAS). Flight delays are difficult to reduce because delays, runway congestion, and ATC operations have a cyclic causality pattern. All causes of this pattern affect each other. This pattern implies that solutions for reducing delays are biased in favor of certain stakeholders in the aviation industry. Each stakeholder defends their own objectives. Although ATC plays a primary role in breaking this cycle of inconvenience, their primary objective is to prevent flight collisions and expedite air traffic flow. This objective is misaligned with airlines' primary objective to increase revenue curtailed by flight delays. Reducing flight delays can increase airline revenue with the opportunity to maximize flight frequency and/**or** customer satisfaction (Note: It is less likely for customer satisfaction to improve with increased flight frequency because it will also result in increased air-traffic congestion which will lengthen flight delays).

Despite these conflicts of interest, ATC and airlines share the incentive to reduce runway congestion. Therefore, reducing congestion is the most effective way to break the causality pattern and placate opposing stakeholders since reducing congestion is mutually beneficial to both ATC and airlines. The ATC reduces congestion by using flight plans submitted by the airlines to reschedule, reroute, and prolong take-off of flights when runways are congested. Although these operations are crucial to reduce runway congestion, these techniques can also contribute to the congestion it is intended to reduce due to the cyclical causality of flight delays and runway congestion.

Proposal:

I developed a scalable data pipeline using Apache Spark's Scala API that can train a cross-validated Gradient Boosted Tree Regressor to estimate runway congestion time. When deployed, ATC can automate flight rescheduling and rerouting by providing an application with a model that feeds flight plans and airline leased terminal times as dependent variables, appling a function that choses the flight plan with the lowest estimated runway congestion time, and replies to airlines with modified flight plans.

Technology Suite:

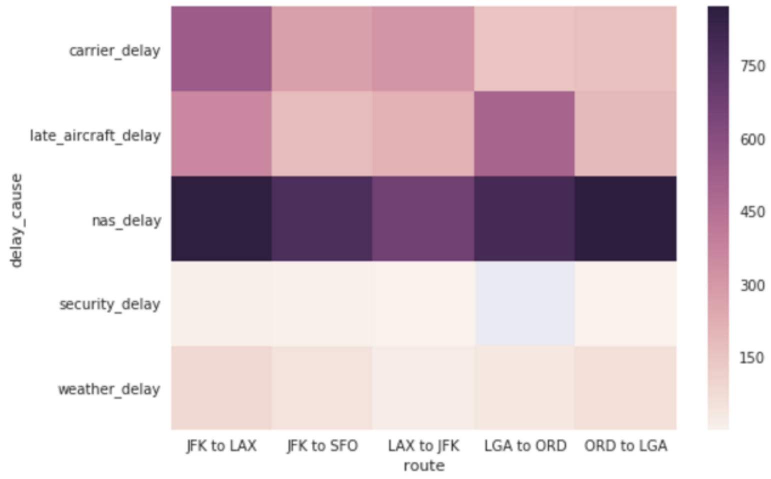
I used a Docker containerized distribution of Apache Spark to build the data pipeline. Although I'm spinning a single node cluster configuration and my dataset isn't large, I have the ability to improve model performance by scaling the pipeline to connect to the head node of spun cluster to process more data. I used Spark's Scala API for data pre-processing, feature engineering, and machine learning, and model evaluation. I decided conducted exploratory data analysis and model evaluation within this container using Python's Seaborn and SciKit-Learn.

Dataset:

I trained this model using "On-Time Performance" data provided by the United States Bureau of Transportation Statistics. This data contains domestic flight information by major airlines for non-stop domestic flights. In order to procure normally distributed data and avoid memory constraints of my personal machine, I filtered the dataset only include flights routed through airports in New York state and flights between November 2015 and October 2016 to encapsulate holidays that impact air and runway traffic.

Exploratory Data Analysis

Congestion Time Frequency:

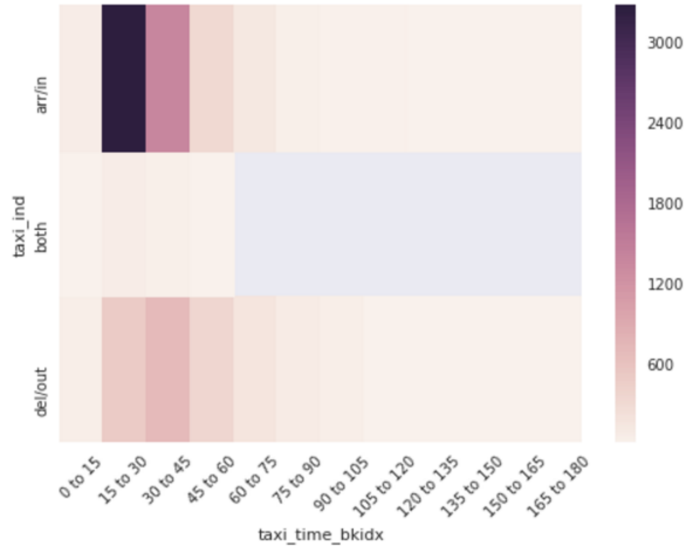


This is a heatmap of runway congestion time frequencies over 30 minutes between delay causes and flight routes. The dataset used for this visualization is filtered by flights with routes experiencing more than 600 occurrences of congestion time greater than 30 minutes.

(Note: Although there are routes for which recorded airports are outside N.Y., these routes indicate flights were routed ***through*** N.Y. and not to/from airports in N.Y.)

This heatmap indicates that most runway congestion of flights routed through N.Y. are associated with the National Airspace System [nas_delay] (airspace, navigation facilities and airports of the United States along with their associated information, services, rules, regulations, policies, procedures, personnel and equipment). This system is used for the ATC of commercial and military flights and is based on multiple classes of airspace for which certain rules apply. This information lead to my hypothesis that runway congestion, flight delays, and ATC/airport operations fit a cyclic causality pattern. Due to this discovery, I decided to engineer features for the model with this pattern in mind.

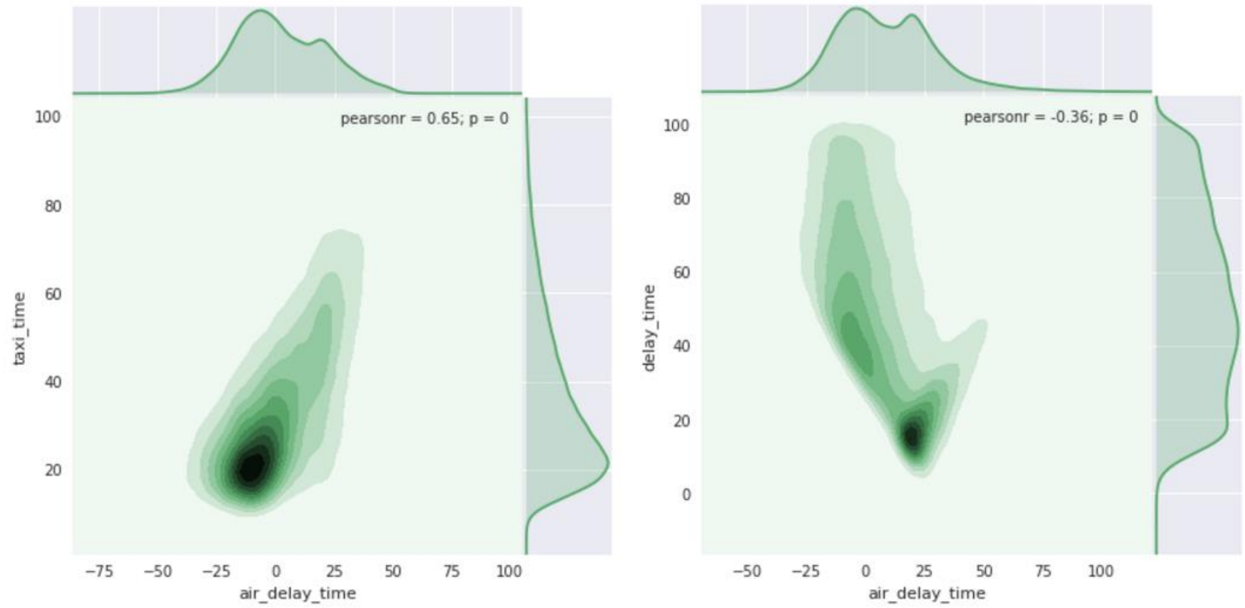
Congestion Time Frequency:



This is a heatmap of delay time frequencies over 30 minutes between arrival [arr/in], destination [del/out], equal [both] congestion types and congestion time windows. The dataset used for this visualization is filtered by flights with routes experiencing more than 600 occurrences of delay time greater than 30 minutes.

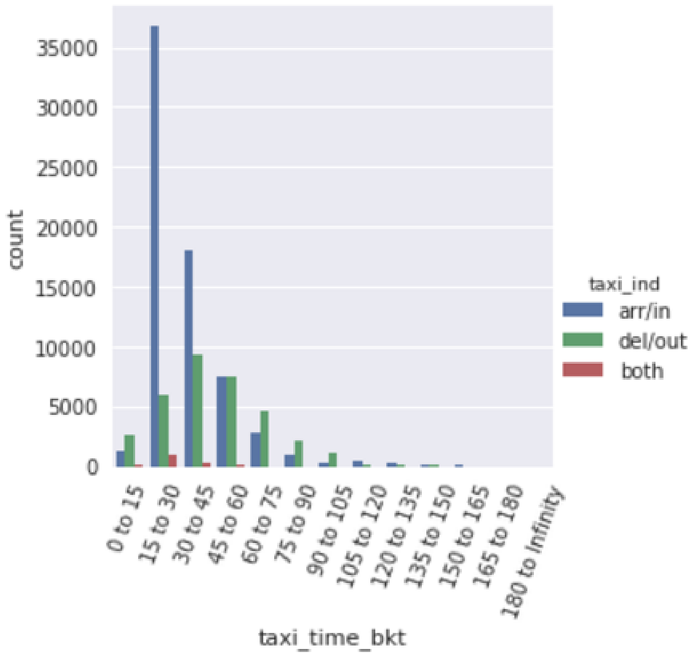
This heatmap indicates that most flight delays through N.Y. are experienced by in-bound / arrival flights. The following Kernel Density Estimation (KDE) visualizations provides insight into this

Kernel Density Estimation:



According to the KDE on the right, flight air-times are mostly less than what the Computer Reservation Schedule (CRS) reported them to be when the densities of congestion time (taxi_time) and the flight air time is delay are high below air time delays of zero minutes. This indicates that flights arriving early cause runway congestion and the Pearson Correlation Coefficient (personr) of 0.65 and density shapes supports the linear relationship between these two features which is why I engineered air_time_delay to be included in my feature set. According to the KDE on the left, flight delay times also follows this trend for some flights. Except in this scenario, delays with air-times less than what the CRS reported are not only instances of early arrivals, but may be due to ATC prolonging flight take-offs at origin airports due to runway congestion, or ATC prolonging airtime due to runway congestion at destination airports.

Congestion Time Frequency Histogram:



According to this histogram, flights with in-bound flight congestion time between 15 & 45 cause most runway congestion. Congestion times under 15 minutes are dominated by ATC prolonging flight take-offs.

Out-bound flight congestion time almost matches between 45 & 60 minutes and overtakes in-bound congestion times as it reduces after the 45-60 minute window.

Feature Engineering

y = label / depend variable: *runway congestion time in minutes*
x = features / independent variables:

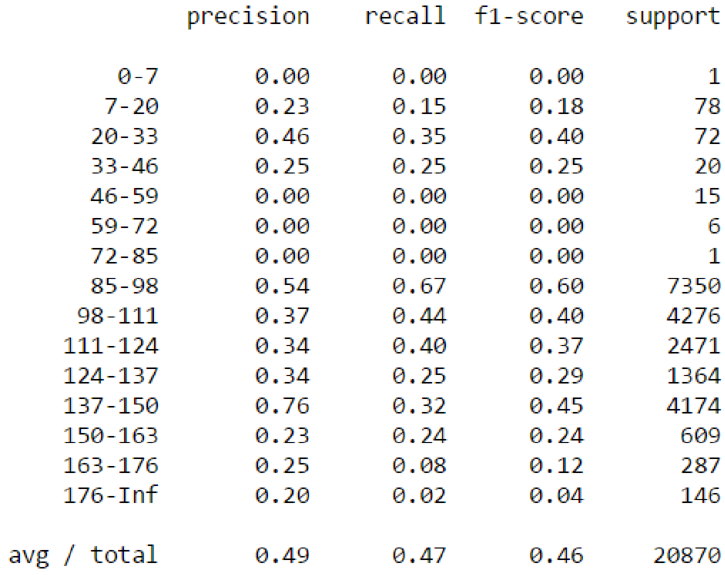
- Airlines, Flight routes
- Delay causes in minutes
- Air time delay: The difference between flights' actual and their scheduled air-times
- An indicator of whether runway congestion times and flight delays are due to origin or destination airports
- An indicator of arrival or departure delay significance (whether these delays are over 15 minutes)
- Buckets for arrival and departure delays, flight distances, and air time delays
- Times for which the wheels of the plain where activated and deactivated
- Primary delay cause
- **Topological representations of time:**
  I decided to additionally represent dates and time (month, day_of_month, hours, and minutes) as Cartesian Coordinates (x,y) for month of year, week of year, day of week, and minute of day to enable the model to learn time as a measure of Euclidean distance, and to enable the creation topological measures such coordinate distances of time. In other words, the machine computated delta of the first (hour 0 or 1) and last (hour 23 or 24) hours of the day is 23 hours, unlike the intuitive/human clock-like computated delta of 1 hour.

Model Performance

Error:

Root Mean Squared Error (**rmse**) - 12.060027132075415
Mean Squared Error (**mse**) - 145.44425442639516
Median Absolute Error (**mae**) - 8.552334236413296

Classification Report of Congestion Time Windows:

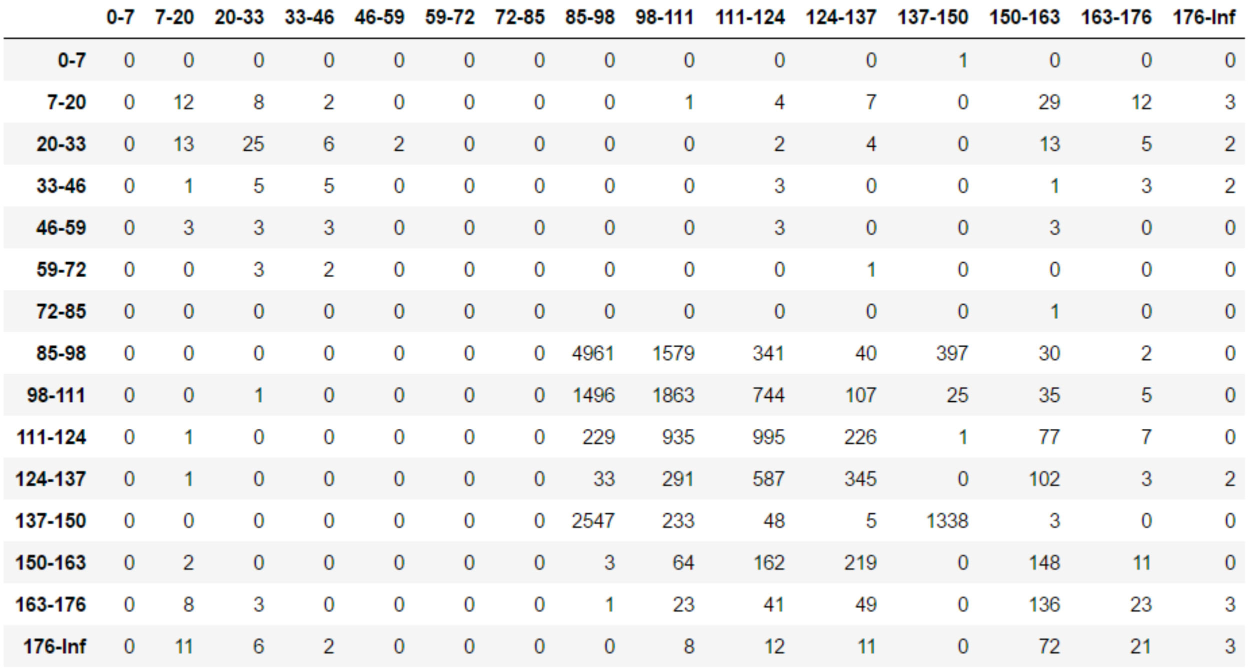


```
             precision    recall  f1-score   support

        0-7       0.00      0.00      0.00         1
       7-20       0.23      0.15      0.18        78
      20-33       0.46      0.35      0.40        72
      33-46       0.25      0.25      0.25        20
      46-59       0.00      0.00      0.00        15
      59-72       0.00      0.00      0.00         6
      72-85       0.00      0.00      0.00         1
      85-98       0.54      0.67      0.60      7350
     98-111       0.37      0.44      0.40      4276
    111-124       0.34      0.40      0.37      2471
    124-137       0.34      0.25      0.29      1364
    137-150       0.76      0.32      0.45      4174
    150-163       0.23      0.24      0.24       609
    163-176       0.25      0.08      0.12       287
    176-Inf       0.20      0.02      0.04       146

avg / total       0.49      0.47      0.46     20870
```

46% of the predicted runway congestion times are accurate. The model is the most accurate for congestion times between 85 and 98 minutes. There could be a primary delay cause for runway congestion times of this class.

50% of absolute deviations from median runway congestion time are below approx. 8.55 minutes, which suggests that the congestion time median might be between 85-150 minutes given the supports.

Confusion Matrix of Congestion Time Windows:



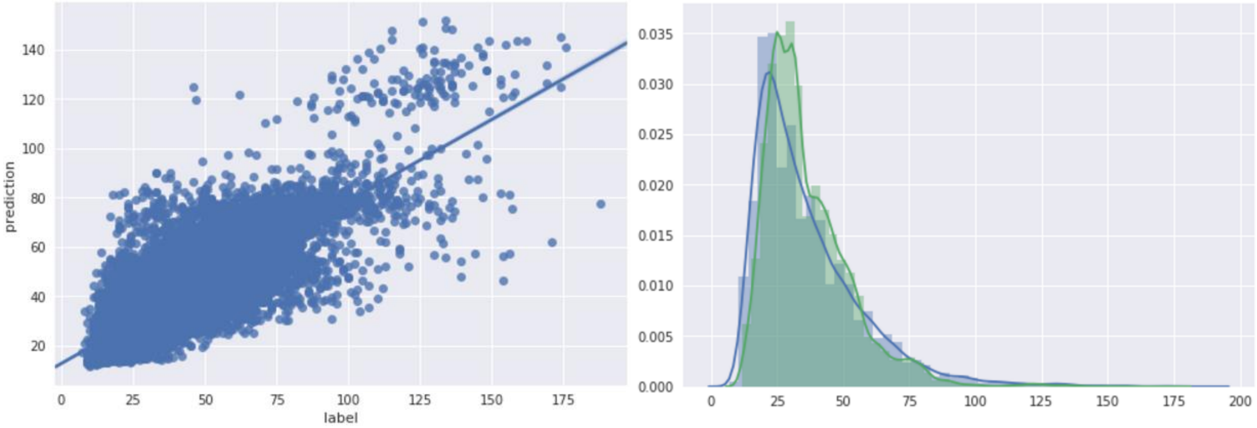
	0-7	7-20	20-33	33-46	46-59	59-72	72-85	85-98	98-111	111-124	124-137	137-150	150-163	163-176	176-Inf
0-7	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
7-20	0	12	8	2	0	0	0	0	1	4	7	0	29	12	3
20-33	0	13	25	6	2	0	0	0	0	2	4	0	13	5	2
33-46	0	1	5	5	0	0	0	0	0	3	0	0	1	3	2
46-59	0	3	3	3	0	0	0	0	0	3	0	0	3	0	0
59-72	0	0	3	2	0	0	0	0	0	0	1	0	0	0	0
72-85	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
85-98	0	0	0	0	0	0	0	4961	1579	341	40	397	30	2	0
98-111	0	0	1	0	0	0	0	1496	1863	744	107	25	35	5	0
111-124	0	1	0	0	0	0	0	229	935	995	226	1	77	7	0
124-137	0	1	0	0	0	0	0	33	291	587	345	0	102	3	2
137-150	0	0	0	0	0	0	0	2547	233	48	5	1338	3	0	0
150-163	0	2	0	0	0	0	0	3	64	162	219	0	148	11	0
163-176	0	8	3	0	0	0	0	1	23	41	49	0	136	23	3
176-Inf	0	11	6	2	0	0	0	0	8	12	11	0	72	21	3

Goodness of Fit:

Coef. of Determination (**r2**) - 0.6533716108894303

Left: The model explains approx. 65% the variability of the response data around its mean.

Right: Predicted (Green) & Actual (Blue) *runway congestion time* distributions



Obstacles:

Memory limitations of a containerized / single node Spark distribution made expensive model improvement techniques such as grid parameter tuning and cross validation too time consuming to perform iterations of exploratory data analysis and development effectively.

Codebase available upon request.