Predicting and Analyzing US Air Traffic Delays using Passenger-centric Data-sources

Philippe Monmousseau, Aude Marzuoli, Eric Feron and Daniel Delahaye

Thirteenth USA/Europe Air Traffic Management Research and Development Seminar

Thursday, June 20, 2019
1 Introduction

2 Datasets and feature creation

3 Prediction results

4 Features analysis and simplification

5 Conclusion & Discussion
Introduction

2 Datasets and feature creation

3 Prediction results

4 Features analysis and simplification

5 Conclusion & Discussion
Current situation

First observation:

Flight delays are a major issue both in the United States and in Europe

In 2017:

- In Europe\(^1\):
  - 44.4% of flights departed with a delay greater than 5 minutes
  - 38.5% arrived with a delay greater than 5 minutes
- In the US\(^2\):
  - 27.0% of departing flights
  - 27.8% of arriving flights

\(^1\)https://www.eurocontrol.int/sites/default/files/publication/files/coda-digest-annual-2017.pdf
\(^2\)http://www.rita.dot.gov/bts/about
Current situation

Fact

Many studies have shown that flight delays do not represent total delays of door-to-door journey of passengers.
Current situation

Fact

Many studies have shown that flight delays do not represent total delays of door-to-door journey of passengers.

New prism is necessary

Passengers are the raison d’être of the ATS and should be at the center of the evaluation of its performance.

New impulse: NextGen in the USA and SESAR in Europe.
Passenger-centric data?

So far, information about passengers is dispersed:

- Airlines only have access to information about their passengers
- Airports have access to:
  - customs and security records
  - queue lengths
  - shuttle traffic, parking occupancy...
- Third-parties can collect traces from WiFi or Bluetooth beacons

No one has an overall view of the situation, only partial prism. Currently Bureau of Transportation Statistics (BTS) provides most used metrics for the US Air Transportation System.
Passenger-centric data?

Other sources of passenger-generated data:

- Mobile phone data:
  - Nation-wide analysis possible
  - Per mode analysis possible
  But proprietary data not easily available
- Social media (Twitter):
  - Useful during natural disasters
  - Used for airlines sentiment analysis
  But not fully leveraged yet
Content

1. Introduction

2. Datasets and feature creation

3. Prediction results

4. Features analysis and simplification

5. Conclusion & Discussion
Description of the dataset

Datasets range

Training set: January 1st, 2017 - December 31st, 2017
Testing set: January 1st, 2018 - February 28th, 2018

Flight-centric data: BTS

- Publicly available
- Usually 2-3 months later
- For each flight, on-time (delays) and status (cancellation) information
Description of the dataset

Datasets range

Training set: January 1st, 2017 - December 31st, 2017
Testing set: January 1st, 2018 - February 28th, 2018

Passenger-centric data: Twitter

- Publicly available
- Real-time availability
- Filtered using airlines and airports handles
Datasets and feature creation

Description of the dataset

Datasets range

Training set: January 1st, 2017 - December 31st, 2017
Testing set: January 1st, 2018 - February 28th, 2018

Table: Handles used for creating the Twitter dataset

<table>
<thead>
<tr>
<th>Category</th>
<th>Twitter handles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airlines</td>
<td>@united, @Delta, @AmericanAir, @SouthwestAir, @SpiritAirlines, @VirginAmerica, @JetBlue</td>
</tr>
<tr>
<td>Airports</td>
<td>@JFKairport, @ATLairport, @flyLAXairport, @fly2ohare, @DFWAirport, @DENAirport, @CLTAirport, @LASairport, @PHXSkyHarbor, @iah, @MiamiAirportMIA, @EWRairport, @MCOAirport, @Official_MCO, @SeaTacAirport, @mspairport, @DTWeetin, @BostonLogan, @PHLAirport, @LGAairport, @FLLFlyer, @BWI_Airport, @Dulles_Airport, @MidwayAirport, @Reagan_Airport, @slcairport, @SanDiegoAirport, @flyTPA, @flypdx, @flystl, @flySFO, @flynashville, @HobbyAirport, @AUSTinAirport, @KClAirport</td>
</tr>
</tbody>
</table>
Prediction goals

The aim is to estimate per hour the following BTS quantities:

- **Number of:**
  - Delayed departing flights: NumDepDelay
  - Delayed arriving flights: NumArrDelay
  - Cancelled flights: NumCancelled

- **Percentage of:**
  - Delayed departing flights: PercDepDelay
  - Delayed arriving flights: PercArrDelay
  - Cancelled flights: PercCancelled

- **Total amount of:**
  - Delays at departure (in minutes): MinDepDelay
  - Delays at arrival (in minutes): MinArrDelay
First exploration

(a) BTS data: volume evolution

(b) Twitter data: volume of airlines-related tweets

Observations:
- Clear 24-hour pattern for both BTS and Twitter
- Some occasional activity spikes for the Twitter data
- Non directly correlated with spikes in delays or cancellations
First exploration

Observations:
- Clear 24-hour pattern for both BTS and Twitter
- Some occasional activity spikes for the Twitter data
- Non directly correlated with spikes in delays or cancellations

(a) BTS data: volume evolution

(b) Twitter data: volume of airports-related tweets
Will it be easy?

Figure: Hourly average of some BTS values over the year 2017. The hourly standard deviation is represented by the vertical bars.

<table>
<thead>
<tr>
<th>BTS label</th>
<th>Average $\mu$</th>
<th>Average $\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total flights</td>
<td>865.45</td>
<td>71.32</td>
</tr>
<tr>
<td>Delayed departing flights</td>
<td>316.23</td>
<td>89.97</td>
</tr>
<tr>
<td>Delayed arriving flights</td>
<td>305.48</td>
<td>93.96</td>
</tr>
<tr>
<td>Cancelled flights</td>
<td>12.37</td>
<td>25.81</td>
</tr>
<tr>
<td>Total delay at departure (min)</td>
<td>11,547.09</td>
<td>6,282.40</td>
</tr>
<tr>
<td>Total delay at arrival (min)</td>
<td>11,494.82</td>
<td>6,659.81</td>
</tr>
<tr>
<td>% delayed departing flights</td>
<td>0.364</td>
<td>0.098</td>
</tr>
<tr>
<td>% delayed arriving flights</td>
<td>0.351</td>
<td>0.098</td>
</tr>
<tr>
<td>% cancelled flights</td>
<td>0.014</td>
<td>0.030</td>
</tr>
</tbody>
</table>

Table: Average over the period 8am-8pm of the 2017 hourly mean and standard deviation of the BTS values.
Feature creation

Using information contained in the text of the tweets:

- Using keyword filtering:

<table>
<thead>
<tr>
<th>Filter</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancellation</td>
<td>cancellation, cancel, cancelled, postponed</td>
</tr>
<tr>
<td>Delay</td>
<td>delay, delayed, wait, waiting, late, postponed, hours</td>
</tr>
</tbody>
</table>

- Using LDA topic modeling:
  - 100 topics created using the tweets gathered from 2017
  - Calculate topic distribution of each tweet
  - Aggregate distributions per hour (mean, standard deviation)
Feature creation

Using information contained in the text of the tweets:

- Using keyword filtering:

<table>
<thead>
<tr>
<th>Filter</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancellation</td>
<td>cancellation, cancel, cancelled, postponed</td>
</tr>
<tr>
<td>Delay</td>
<td>delay, delayed, wait, waiting, late, postponed, hours</td>
</tr>
</tbody>
</table>

- Using LDA topic modeling:
  - 100 topics created using the tweets gathered from 2017
  - Calculate topic distribution of each tweet
  - Aggregate distributions per hour (mean, standard deviation)
Features summary

The following 8,327 features are considered:

- Hourly volume of tweets for each search handle (7 airlines and 34 airports giving 41 features): Num_tweets_handle
- Hourly volume of delay-related tweets for each search handle (41 features): Num_tweets_kwd_handle
- Hourly volume of cancelled-related tweets for each search handle (41 features): Num_tweets_kwc_handle
- Hourly average of topic distribution for each search handle (41x100 features): Mean_topic_handle
- Hourly standard deviation of topic distribution for each search handle (41x100 features): Std_topic_handle
- Month of the year, Day of the month, Day of the week and Hour in the day (4 features)
Sanity check

**Table: Most correlated feature per BTS label**

<table>
<thead>
<tr>
<th>BTS</th>
<th>Best Feature</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinDepDelay</td>
<td>Num_tweets_kwd_@AmericanAir</td>
<td>0.685</td>
</tr>
<tr>
<td>MinArrDelay</td>
<td>Num_tweets_kwd_@AmericanAir</td>
<td>0.668</td>
</tr>
<tr>
<td><strong>NumDepDelay</strong></td>
<td>Num_tweets_@SouthwestAir</td>
<td><strong>0.703</strong></td>
</tr>
<tr>
<td>NumArrDelay</td>
<td>Num_tweets_@SouthwestAir</td>
<td>0.684</td>
</tr>
<tr>
<td>PercDepDelay</td>
<td>Num_tweets_kwd_@AmericanAir</td>
<td>0.573</td>
</tr>
<tr>
<td>PercArrDelay</td>
<td>Num_tweets_kwd_@AmericanAir</td>
<td>0.585</td>
</tr>
<tr>
<td>NumCancelled</td>
<td>Num_tweets_kwc_@Delta</td>
<td>0.330</td>
</tr>
<tr>
<td>PercCancelled</td>
<td>Num_tweets_kwc_@Delta</td>
<td>0.291</td>
</tr>
</tbody>
</table>

Simple but efficient:

- BTS delays most correlated to delay keyword features
- BTS cancellation most correlated to cancellation keyword features

Cancellations are however poorly correlated to the Twitter features
Content

1. Introduction

2. Datasets and feature creation

3. Prediction results

4. Features analysis and simplification

5. Conclusion & Discussion
Methodology

- 3 regressors for each BTS label:
  - Decision Tree Regressor (DTR)
  - Random Forest Regressor (RFR)
  - Gradient Boosting Regressor (GBR)

- 2 performance measures:
  - $R^2$ score
  - Mean Absolute Error

- One benchmark:
  - Facebook’s timeseries forecasting tool (FB Prophet)

- From $h = 0$ to $h = 6$ hours ahead
Methodology

- 3 regressors for each BTS label:
  - Decision Tree Regressor (DTR)
  - Random Forest Regressor (RFR)
  - Gradient Boosting Regressor (GBR)

- 2 performance measures:
  - $R^2$ score
    \[
    R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}
    \]
  - Mean Absolute Error
    \[
    MAE = \frac{1}{n} \sum_i |f_i - y_i|
    \]

- One benchmark:
  - Facebook’s timeseries forecasting tool (FB Prophet)

- From $h = 0$ to $h = 6$ hours ahead
Methodology

- 3 regressors for each BTS label:
  - Decision Tree Regressor (DTR)
  - Random Forest Regressor (RFR)
  - Gradient Boosting Regressor (GBR)

- 2 performance measures:
  - \( R^2 \) score
  - Mean Absolute Error

- One benchmark:
  - Facebook’s timeseries forecasting tool (FB Prophet)

  From \( h = 0 \) to \( h = 6 \) hours ahead
Methodology

- 3 regressors for each BTS label:
  - Decision Tree Regressor (DTR)
  - Random Forest Regressor (RFR)
  - Gradient Boosting Regressor (GBR)
- 2 performance measures:
  - $R^2$ score
  - Mean Absolute Error
- One benchmark:
  - Facebook’s timeseries forecasting tool (FB Prophet)
- From $h = 0$ to $h = 6$ hours ahead
Performance results

Table: Immediate prediction $R^2$ score comparison

<table>
<thead>
<tr>
<th>BTS label</th>
<th>FB Prophet</th>
<th>DTR</th>
<th>RFR</th>
<th>GBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinDepDelay</td>
<td>3.80e-01</td>
<td>5.19e-01</td>
<td>6.57e-01</td>
<td>6.96e-01</td>
</tr>
<tr>
<td>MinArrDelay</td>
<td>3.24e-01</td>
<td>3.85e-01</td>
<td>5.58e-01</td>
<td>6.13e-01</td>
</tr>
<tr>
<td>NumDepDelay</td>
<td>6.39e-01</td>
<td>7.72e-01</td>
<td>8.58e-01</td>
<td>8.76e-01</td>
</tr>
<tr>
<td>NumArrDelay</td>
<td>6.47e-01</td>
<td>6.90e-01</td>
<td>7.70e-01</td>
<td>7.99e-01</td>
</tr>
<tr>
<td>PercDepDelay</td>
<td>-2.40e-02</td>
<td>2.92e-01</td>
<td>5.22e-01</td>
<td>5.84e-01</td>
</tr>
<tr>
<td>PercArrDelay</td>
<td>-1.90e-01</td>
<td>-1.49e-01</td>
<td>2.65e-01</td>
<td>2.94e-01</td>
</tr>
<tr>
<td>NumCancelled</td>
<td>7.23e-03</td>
<td>-3.80e-01</td>
<td>2.80e-01</td>
<td>3.08e-01</td>
</tr>
<tr>
<td>PercCancelled</td>
<td>-9.62e-02</td>
<td>-2.70e-01</td>
<td>1.86e-01</td>
<td>1.25e-01</td>
</tr>
</tbody>
</table>

Table: Immediate prediction MAE comparison

<table>
<thead>
<tr>
<th>BTS label</th>
<th>FB Prophet</th>
<th>RF regressor</th>
<th>GB regressor</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinDepDelay</td>
<td>4.09e+03 (min)</td>
<td>2.61e+03 (min)</td>
<td>2.40e+03 (min)</td>
</tr>
<tr>
<td>MinArrDelay</td>
<td>4.12e+03 (min)</td>
<td>2.80e+03 (min)</td>
<td>2.63e+03 (min)</td>
</tr>
<tr>
<td>NumDepDelay</td>
<td>72.1 (flights)</td>
<td>39.7 (flights)</td>
<td>37.1 (flights)</td>
</tr>
<tr>
<td>NumArrDelay</td>
<td>60.9 (flights)</td>
<td>44.2 (flights)</td>
<td>42.1 (flights)</td>
</tr>
<tr>
<td>PercDepDelay</td>
<td>1.15e-01</td>
<td>7.27e-02</td>
<td>6.75e-02</td>
</tr>
<tr>
<td>PercArrDelay</td>
<td>1.05e-01</td>
<td>7.84e-02</td>
<td>7.71e-02</td>
</tr>
<tr>
<td>NumCancelled</td>
<td>11.4 (flights)</td>
<td>10.7 (flights)</td>
<td>9.75 (flights)</td>
</tr>
<tr>
<td>PercCancelled</td>
<td>1.85e-02</td>
<td>1.82e-02</td>
<td>2.15e-02</td>
</tr>
</tbody>
</table>
Performance results

Table: Average over the period 8am-8pm of the 2017 hourly mean and standard deviation of the BTS values

<table>
<thead>
<tr>
<th>BTS label</th>
<th>Average μ</th>
<th>Average σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total flights</td>
<td>865.45</td>
<td>71.32</td>
</tr>
<tr>
<td><strong>Delayed departing flights</strong></td>
<td><strong>316.23</strong></td>
<td><strong>89.97</strong></td>
</tr>
<tr>
<td>Delayed arriving flights</td>
<td>305.48</td>
<td>93.96</td>
</tr>
<tr>
<td>Cancelled flights</td>
<td>12.37</td>
<td>25.81</td>
</tr>
<tr>
<td>Total delay at departure (min)</td>
<td>11,547.09</td>
<td>6,282.40</td>
</tr>
<tr>
<td>Total delay at arrival (min)</td>
<td>11,494.82</td>
<td>6,659.81</td>
</tr>
<tr>
<td>% delayed departing flights</td>
<td>0.364</td>
<td>0.098</td>
</tr>
<tr>
<td>% delayed arriving flights</td>
<td>0.351</td>
<td>0.098</td>
</tr>
<tr>
<td>% cancelled flights</td>
<td>0.014</td>
<td>0.030</td>
</tr>
</tbody>
</table>

Table: Immediate prediction MAE comparison

<table>
<thead>
<tr>
<th>BTS label</th>
<th>FB Prophet</th>
<th>RF regressor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MinDepDelay</strong></td>
<td>4.09e+03 (min)</td>
<td>2.61e+03 (min)</td>
</tr>
<tr>
<td><strong>MinArrDelay</strong></td>
<td>4.12e+03 (min)</td>
<td>2.80e+03 (min)</td>
</tr>
<tr>
<td><strong>NumDepDelay</strong></td>
<td><strong>72.1 (flights)</strong></td>
<td><strong>39.7 (flights)</strong></td>
</tr>
<tr>
<td>NumArrDelay</td>
<td>60.9 (flights)</td>
<td>44.2 (flights)</td>
</tr>
<tr>
<td>PercDepDelay</td>
<td>1.15e-01</td>
<td>7.27e-02</td>
</tr>
<tr>
<td>PercArrDelay</td>
<td>1.05e-01</td>
<td>7.84e-02</td>
</tr>
<tr>
<td>NumCancelled</td>
<td>11.4 (flights)</td>
<td>10.7 (flights)</td>
</tr>
<tr>
<td>PercCancelled</td>
<td>1.85e-02</td>
<td>1.82e-02</td>
</tr>
</tbody>
</table>
Content

1. Introduction
2. Datasets and feature creation
3. Prediction results
4. Features analysis and simplification
5. Conclusion & Discussion
Features analysis and simplification

Feature importance

Top ten features:

1. Hour
2. Num_tweets_@SouthwestAir
3. Num_tweets_@Delta
4. Num_tweets_kwd_@Delta
5. Num_tweets_kwd_@SouthwestAir
6. Num_tweets_kwd_@AmericanAir
7. Num_tweets_@AmericanAir
8. Num_tweets_kwd_@JetBlue
9. Num_tweets_kwd_@united
10. Month

(a) Top ten features from the Random Forest regressor used for predicting the number of delayed flights at $h = 0$
(a) Top ten features from the Random Forest regressor used for predicting the number of delayed flights at $h = 0$

(b) $R^2$ score comparison of the performance of the different regressors using the full feature set and using the sole feature 'Hour'.
Feature importance

Extract features gathering 99% of the total feature importance for the Random Forest Regressor

Table: Feature categorization of the reduced feature set explaining 99% of the number of delayed flights prediction performance

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw number of tweets</td>
<td>26</td>
</tr>
<tr>
<td>Number of tweets with delay keywords</td>
<td>5</td>
</tr>
<tr>
<td>Number of tweets with cancellation keywords</td>
<td>4</td>
</tr>
<tr>
<td>Average distribution of a topic</td>
<td>453</td>
</tr>
<tr>
<td>Standard deviation of the distribution of a topic</td>
<td>180</td>
</tr>
</tbody>
</table>
Feature importance vs. actual ranking

Table: Airlines and airports categorization of the reduced feature set

<table>
<thead>
<tr>
<th>Rank</th>
<th>Airlines</th>
<th>Freq.</th>
<th>Airports</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Delta</td>
<td>137</td>
<td>ATL</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>AmericanAir</td>
<td>127</td>
<td>DFW</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>United</td>
<td>116</td>
<td>LAX</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>SouthwestAir</td>
<td>95</td>
<td>PHL</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>JetBlue</td>
<td>55</td>
<td>SEA</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>SpiritAirlines</td>
<td>18</td>
<td>JFK</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>VirginAmerica</td>
<td>15</td>
<td>DEN, CLT</td>
<td>5</td>
</tr>
</tbody>
</table>

Table: Delay ranking on the year 2017

<table>
<thead>
<tr>
<th>Rank</th>
<th>Airlines</th>
<th># delays</th>
<th>Airports</th>
<th># delays</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SouthwestAir</td>
<td>615,095</td>
<td>ATL</td>
<td>129,196</td>
</tr>
<tr>
<td>2</td>
<td>AmericanAir</td>
<td>282,508</td>
<td>LAX</td>
<td>90,729</td>
</tr>
<tr>
<td>3</td>
<td>Delta</td>
<td>280,975</td>
<td>ORD</td>
<td>88,127</td>
</tr>
<tr>
<td>4</td>
<td>JetBlue</td>
<td>238,230</td>
<td>DEN</td>
<td>81,401</td>
</tr>
<tr>
<td>5</td>
<td>United</td>
<td>184,120</td>
<td>SFO</td>
<td>68,184</td>
</tr>
<tr>
<td>6</td>
<td>SpiritAirlines</td>
<td>47,412</td>
<td>DFW</td>
<td>64,661</td>
</tr>
<tr>
<td>7</td>
<td>VirginAmerica</td>
<td>28,938</td>
<td>PHX</td>
<td>55,171</td>
</tr>
</tbody>
</table>
Performance improvements

Observations:
- Performance two-by-two comparable
- Faster training times
- Slightly better performance most of the times

Figure: $R^2$ score comparison
Content

1. Introduction

2. Datasets and feature creation

3. Prediction results

4. Features analysis and simplification

5. Conclusion & Discussion
Conclusion

Summary of this work:

- Investigation into Twitter as a reliable passenger-centric data-source
- Method to transform Twitter stream into useful feature set
- Real-time passenger-centric data contains extra information than history of flight-centric data for prediction
- Best results obtained for predicting the number of delayed departing flights
Conclusion

Summary of this work:

- Investigation into Twitter as a reliable passenger-centric data-source
- Method to transform Twitter stream into useful feature set
- Real-time passenger-centric data contains extra information than history of flight-centric data for prediction
- Best results obtained for predicting the number of delayed departing flights
Conclusion

Summary of this work:

- Investigation into Twitter as a reliable passenger-centric data-source
- Method to transform Twitter stream into useful feature set
- Real-time passenger-centric data contains extra information than history of flight-centric data for prediction
- Best results obtained for predicting the number of delayed departing flights
Conclusion

Summary of this work:

- Investigation into Twitter as a reliable passenger-centric data-source
- Method to transform Twitter stream into useful feature set
- Real-time passenger-centric data contains extra information than history of flight-centric data for prediction
- Best results obtained for predicting the number of delayed departing flights
Discussion

Possible next steps:

- Refine model for a useful prediction per airport and/or per airline
- Expand model to another region (EU) with available flight-centric data for validation
- Integrate other readily available data-sources (both flight and passenger-centric)
Discussion

Possible next steps:

- Refine model for a useful prediction per airport and/or per airline
- Expand model to another region (EU) with available flight-centric data for validation
- Integrate other readily available data-sources (both flight and passenger-centric)

Thank you for your attention