STAT642 Group 10

Andrew Armstrong, Abhay Bhat, Alex Graber, Eugene Olkhov, Elizabeth Perkins

Project Proposal – Will I Be Late?

## Introduction and Background:

Approximately 119,000 daily riders take advantage SEPTA’s Regional Rail system, which provides service to the Philadelphia metropolitan area.[[1]](#footnote-1) The SEPTA (Southeastern Pennsylvania Transportation Authority) Regional Rail system consists of commuter rail service on 13 branches to more than 150 active stations in Philadelphia, Pennsylvania, and its suburbs and satellite cities.

SEPTA reports On-Time Performance (OTP) to measure service reliability. OTP identifies the number of trains for all rail lines that arrive at their scheduled destination at the scheduled time. However, by industry standard, a train may arrive up to 5 minutes and 59 seconds after its scheduled time and still be considered on-time. SEPTA has set an On-Time Performance target such that 91% of its trains arrive on time.[[2]](#footnote-2) Thus, even with 100% “on time” performance, trains may still arrive late, forcing commuters to deal with uncertainty and lost time – especially if they rely on back-to-back connections.

## Research Question / Objective:

In this study, we seek to identify factors that are predictive of delay, including individual stations, line interactions, day-of-week, holidays, and weather. Ultimately, we will create a model(s) for each line that will predict – given certain conditions – the likelihood that a given train will run late. This information will be of use to commuters, who can use the information to potentially alter their schedule, and to SEPTA, which can use the information to prepare for occurrences which increase the probability of delay.

## Methodology:

While the SEPTA OTP data is easily accessible due through Kaggle.com, the weather information is more challenging to access. For hour-by-hour weather information, we will have to write a custom function to call the DarkSky API per day and per locale and aggregate the weather information. Once our dataset has been collected, we will have to review, clean, and verify it.

Once cleaned, we can begin the modeling process. Models used for data mining and prediction may include linear and logistic regressions, Naïve Bayes, random forest, and/or SVM. In any of these cases, we will iteratively improve our models to use k-fold cross validation to ensure optimal model hyperparameter selection and to vet that our models do not overfit our dataset.

## Hypotheses:

We anticipate that weather events (precipitation, high winds), delays on other lines (network effects), and time of day (commute vs. non-commute hours), and time of year (slippery rail season) will all be strongly predictive variables in the likelihood at train is delayed. We believe that these factors may have different influence on lateness – commuter trains are more likely to be 1-5 minutes delayed due to high volume of passenger transfer and higher volume of trains on the tracks, while weather events, such as storms, may be at fault for the more severe delays.

## Challenges:

The major challenges we anticipate are fivefold: (1) Collecting weather information to the granularity we want requires custom programming (and potentially a cost outlay, as the API provides only limited free access). (2) Cleaning the data may require dropping a considerable number of rows due to missing and/or outlier data, and potentially resampling the data so we don’t have rare cases (only 30% of the trains run perfectly on time). (3) Though the overall data file is quite large, once we split by line we may run into problems where we have limited training data. (4) We have many potential independent variables to mine, but many will be highly correlated (especially within the weather data). Optimizing or justifying feature selection may be time-intensive and model-dependent. (5) We have many models available from which to choose, and subsequently many ways to tune each model. Attempting to find the “best” model may take a back seat to optimize a “good enough” model which is also explicable.

## Data Description:

We will use the On-Time Performance dataset provided by SEPTA to Kaggle.com. This “otp.csv” from the Kaggle dataset contains On-Time Performance information as from 23 March, 2016 to 6 November, 2016 [[3]](#footnote-3):

* train\_id
* direction ('N' or 'S' direction is demarcated as either Northbound or Southbound)[^1](https://www.kaggle.com/forums/f/1300/septa-regional-rail/t/21259/why-are-all-trains-originating-in-thorndale-marked-as-southbound)
* origin (See map below - you'll see 'Warminster', 'Glenside',...'Airport Terminal..')
* next\_station (Think of this as the station stop, at timeStamp)
* date
* status ('On Time', '5 min', ... This is a status on train lateness. 999 is a suspended train)
  + Note: only 32% of trains run perfectly “on-time”, with another 30% running 1-3 minutes late and the remaining ~1/3 of the trains running >3 minutes late
* timeStamp

We will also collect weather data using the DarkSky API[[4]](#footnote-4) for the Philadelphia area for all start- and end-points of the Regional Rail lines we model. The data will be aggregated into a single file, allowing us to link it to the OTP data, and includes the following information:

* time
* summary
* icon description
* precipIntensity
* precipProbability
* temperature
* apparentTemperature
* dewPoint
* humidity
* pressure
* windSpeed
* windBearing
* visibility
* cloudCover
* precipType

1. SEPTA. “SEPTA Revenue & Ridership Report – March 2018.” http://septa.org/strategic-plan/reports/revenue-ride.pdf. [↑](#footnote-ref-1)
2. SEPTA. “SEPTA - Regional Rail.“ www.kaggle.com/septa/on-time-performance. [↑](#footnote-ref-2)
3. SEPTA. “SEPTA - Regional Rail.“ www.kaggle.com/septa/on-time-performance. [↑](#footnote-ref-3)
4. Dark Sky. “Dark Sky API Overview.” https://darksky.net/dev/docs. [↑](#footnote-ref-4)