# Predicting Lateness of SEPTA Regional Rail

**Lukas Allard** ALLARDL@SEAS.UPENN.EDU
**Ajay Charan** AJCHARAN@SEAS.UPENN.EDU
**Emmanuel Genene** EGENENE@SEAS.UPENN.EDU

## Abstract

Timeliness is important for all public transportation platforms. SEPTA regional rail transports over 100,000 riders daily [2]. One late train means thousands of riders may subsequently be late to their destinations. This paper details an approach to approximating how late a SEPTA regional rail passenger would be to their desired station. Applying linear regression, and gaussian process regression models to SEPTA data for each station along specified regional lines, we were able to predict a passenger's lateness within 5 minutes.

## 1. Problem and Motivation

SEPTA uses a metric called "On-Time Performance" (OTP) to measure the timeliness of their Regional Rail Trains. OTP defines a train to be "On-Time" even if it is up to 5 minutes 59 seconds late [1]. However, OTP does not convey useful information to passengers. SEPTA passengers would benefit more if they knew approximately how late they may be to their destination. This would allow them to choose an alternative form of transportation if timeliness is an issue.

### 1.1. Related Work

SEPTA created a web-based tool called TrainView that allows viewers to track the status of active trains. The application shows the stations along the route, scheduled arrival times at each station, estimated arrival times, and actual arrival times all for a specific, currently active train [3]. In theory this application would provide passengers with all the important information they need to judge the timeliness of their trains; however, SEPTA's estimated arrival times are produced by a trivial approach that may not reflect accurate estimates. For example, if a train is shown on TrainView to have arrived at station X five minutes later than scheduled, the estimated arrival time for station Y is simply the scheduled time plus five minutes. This methodology fails to capture trends in the data and completely ignores any external factors that may cause a train to be late. Additionally, earliness is not reflected in estimated arrival times. For example, if a train arrives at station X five minutes early, the estimated arrival time for station Y is, by default, just the initial scheduled arrival time instead of the scheduled arrival time minus five minutes. Furthermore, a passenger using TrainView can only retrieve estimated time of arrivals for active trains. Thus, if someone is attempting to plan ahead (the train they would use for their upcoming trip isn't currently active), they would not be able to estimate how late they may be using TrainView.

In addition to TrainView, SEPTA has taken another approach to improving passengers' accessibility to train status information. After making their train data publicly available, some viewers have performed detailed analysis demonstrating important factors that influence train lateness. These online publications show that factors like weekday and specific terminals are related to how late a train will be; therefore, it is vital that important features such as these be integrated into the method of estimating how late a train will be [1].

## 2. Approximating Train Lateness

Unlike TrainView, we want to provide SEPTA passengers with accurate predictions of how late they may be to a desired station; thus, our approach incorporates machine learning algorithms that capture a number of features from the SEPTA data.

### 2.1. Data

The raw, public data provided by SEPTA regional rails included train ID numbers, origin station, destination station, next station, length of stop at the current station (seconds), date, latitude and longitude of the train, and status (number of minutes late). The available data ranged from March 2016 to May 2016 and included n>500,000

University of Pennsylvania, CIS 419/519 Course Project.

instances.

First, we rearranged the data by regional rail line. We included data for the following rail lines in our final algorithm: Newark/Norristown, Airport/Warminster, Elwyn/Suburban Station, Elwyn/West Trenton, Fox Chase/Chestnut Hill West, and Suburban Station/Cynwyd. These lines encompass over $50\%$ of all regional rail stops.

Preprocessing this data into usable features for our learning algorithms was the next step. We converted the date into "time of day", "day of the week", and "month". We used the origin station, destination station, and next station to derive the train's current station. Originally we generated a distance feature that described how far the train's current station is from the origin station in units of stations. We later converted this feature to actual physical distance using the GPS coordinates of the data; however, this did not change the results significantly so we did away with the change.

After this feature engineering, the final data we input into the learning algorithm was train ID number, current station (labeled 1...n), distance from current station to station of origin (in units of stations), length of stop at the current station (seconds), time of day (seconds past 00:00:00), day of the week (1-7), and month (3-5 since the data came from March-May) while status (how late the train is) remained as the label for each instance. All the features and a brief explanation of why they were used are shown in Table 1 and an example data instance is shown in Table 2.

*Table 1.* Feature Selection and Explanation

| Feature | Explanation |
|---|---|
| Train ID | Certain trains may have mechanical issues that make them slower than expected or consistently cause delays. Additionally, the same trains always take the same routes and some routes may be more prone to delays than others. |
| Distance From Origin | A train is more likely to be late to later stations along the track than earlier stations |
| Pause at Station | Long stops at the current station are likely to make a train late to following station(s). |
| Time of Day | Certain times of the day are busier than others. For example, trains are more likely to be behind schedule during rush hour. |
| Weekday | Certain days of the week experience higher train use and are more prone to delays. |
| Month | Train use increases in the winter months when it's too cold to ride bikes. This makes colder months more prone to delays. |

While handling the data, we discovered that trying to predict lateness based off of individual train ID's resulted in a time-series problem (i.e. the train's lateness at one station effected its lateness at all proceeding stations). To get around this dilemma, we chose to examine lateness based off of individual stations instead. Thus, we look at trends in the data to predict how many minutes late a passenger would likely be if they wanted to go to station X.

### 2.2. Modeling the Data

We chose to implement a linear regression classifier, a random forest classifier, and a gaussian process regression classifier to train models on the SEPTA data. Each of the models were fitted to the same 70% of the data for each station of each regional rail line we used. These models were then each tested on the remaining 30% of the data to compare how far off the predicted lateness was from the true lateness (in minutes). Ultimately, as expected, the random forest classifier proved to give the best lateness estimate for the majority of stations [4]. The gaussian process regression classifier was the next best and linear regression proved to be the worst. The average error (how many minutes off our models' predictions were from the true lateness) for each station is depicted for each classifier in Figure 1. The random forest model was able to predict average lateness of a train for a specified station within 2.5 minutes of the train's true lateness to that station. For most stations, the prediction was within 1.5 minutes of the train's true lateness. Because the random forest model produced the most accurate estimates of how late the train would be to the desired station, we applied it to the other lines and generated the plots in Figures 2-7.

### 2.3. Results and Impact

The overarching idea is that a potential SEPTA passenger taking a regional rail line will be able to input their destination station, their train ID number, and a timestamp that includes the time of day and date (taken from the computer when they submit their train information) and the learning algorithm will predict how late their train would likely be to that destination (in minutes). To demonstrate the usefulness of our project, we developed an interactive code wherein the user provides the information as outlined above and receives an instant prediction of how late they will be to their destination if they take that train. Figure 8 shows how a passenger would use the interface to predict lateness.

*Table 2.* Processed Data Features Example

| TRAIN ID | DISTANCE FROM ORIGIN | PAUSE AT STATION | TIME OF DAY | DAY OF WEEK | MONTH |
|---|---|---|---|---|---|
| 1 | 2 (STATIONS) | 33 (SECONDS) | 43,200 | 7 | 4 |

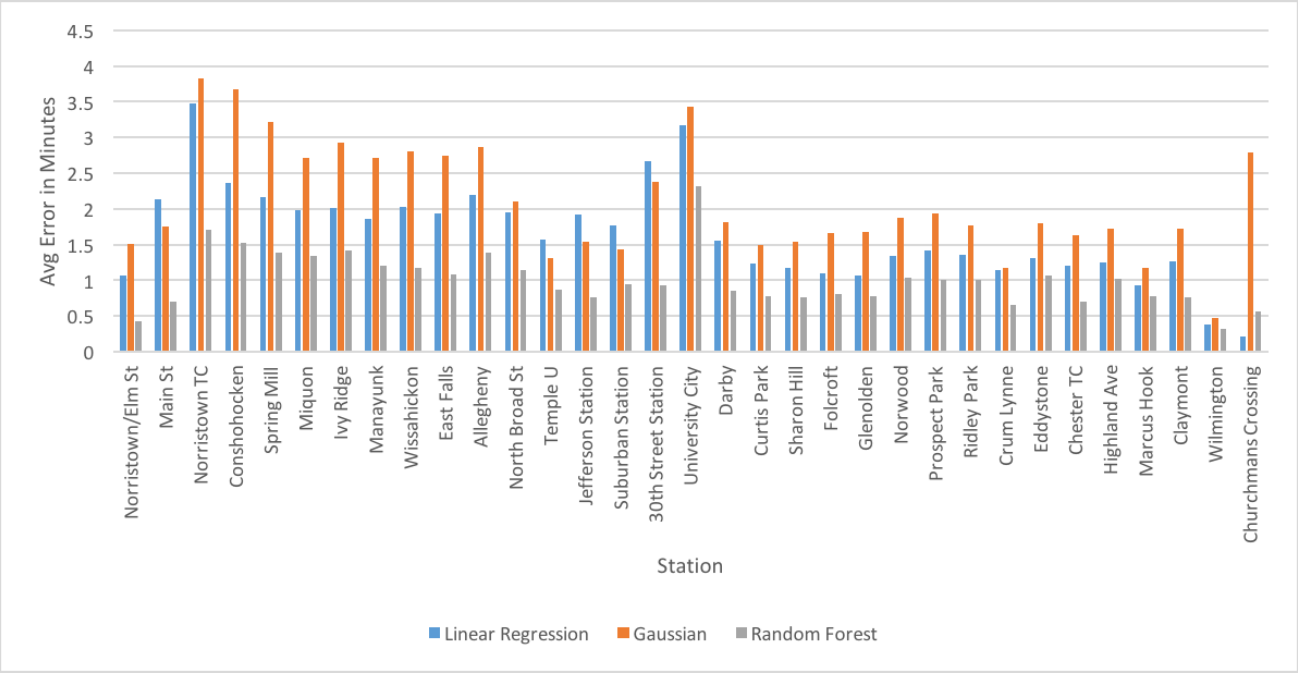


*Figure 1.* Average error (minutes) for each station predicted by Linear Regression Model, Gaussian Process Regression Model, and Random Forest Model

```
Please enter the source station: Marcus Hook
Which station do you want to know the status for?: Crum Lynne
Do you want to know the status at the current time? (y/n): y
The train is predicted to be 0:00:57 late (H:MM:SS).
Do you want to quit?(y/n): n
Please enter the source station: Marcus Hook
Which station do you want to know the status for?: Ridley Park
Do you want to know the status at the current time? (y/n): y
The train is predicted to be 0:01:00 late (H:MM:SS).
Do you want to quit?(y/n): n
Please enter the source station: Marcus Hook
Which station do you want to know the status for?: University City
Do you want to know the status at the current time? (y/n): y
The train is predicted to be 0:01:23 late (H:MM:SS).
Do you want to quit?(y/n):
```

*Figure 8.* Example of how a passenger would query how late they would be to their destination station

The estimated lateness our algorithm provides is a more reasonable prediction than SEPTA TrainView's current estimated delay because it considers the relevant features described in Table 1, can predict early arrivals, and can be found for any train not just active ones. If implemented into the SEPTA TrainView application to adjust their current estimated arrival times, the program would be a lot more realistic and useful to SEPTA regional rail passengers.

### 2.4. Future Work

Currently our project only uses a fraction of the total SEPTA regional rail data. Ideally our learning process would use all of the available data (discarding anomalies caused by the recent SEPTA strike). More training instances would likely help us produce more accurate predictions.

## Acknowledgments

Guidance provided by Dr. Eaton on how to avoid time-series complications.

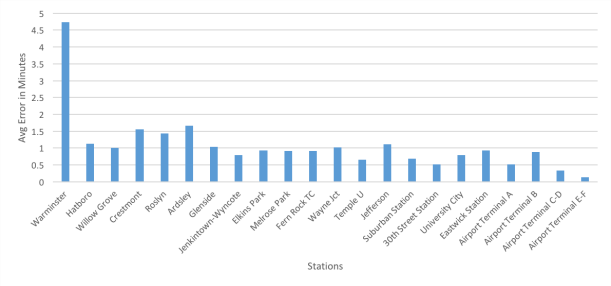


*Figure 2.* Airport/Warminster Line

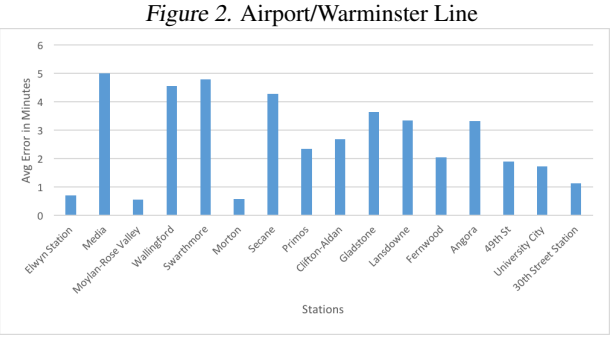


*Figure 3.* Elwyn/Suburban Station Line

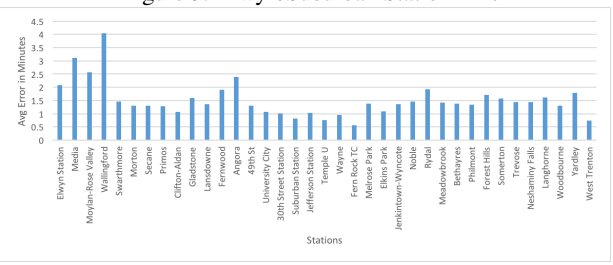


*Figure 4.* Elwyn/West Trenton Line

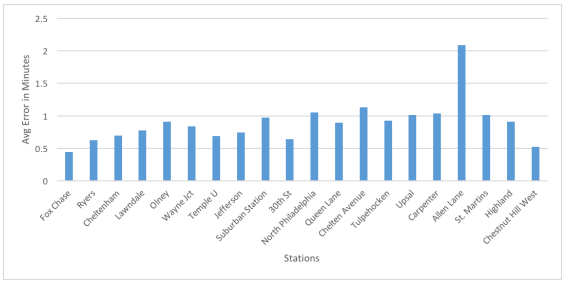


*Figure 5.* Fox Chase/Chestnut Hill West Line

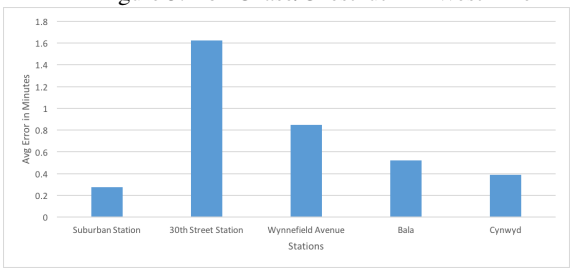


*Figure 6.* Suburban Station/Cynwyd Line

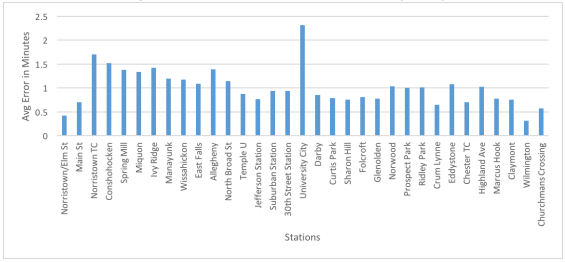


*Figure 7.* Newark/Norristown Line

## Resources

[1] SEPTA. (2016, October 02). Predict Arrival Times of Philadelphia's Regional Trains. Retrieved from https://www.kaggle.com/septa/on-time-performance/version/20.

[2] REVENUE & RIDERSHIP REPORT (2016): n. pag. Revenue & Ridership Management SEPTA. Web. 12 Nov. 2016. Retrieved from http://septa.org/strategic-plan/reports/revenue-ride.pdf.

[3] "SEPTA — TrainView — Real Time Regional Rail Status Information." TrainView — Real Time Regional Rail Status Information. SEPTA, n.d. Web. 11 Dec. 2016. Retrieved from http://www.septa.org/site/trainview.html.

[4] Van Der Spoel, Sjoerd, Chintan Amrit, and Jos Van Hillegersberg. "Predictive Analytics for Truck Arrival Time Estimation: A Field Study at a European Distribution Center." International Journal of Production Research (2016): 1-22. Web. 11 Dec. 2016.

# Predicting Lateness of SEPTA Regional Rail

## THE PROBLEM?

SEPTA estimates a train's arrival time via its TrainView application using a simple procedure that produces inaccurate predictions of lateness

(Estimated Arrival) = (Status)+(Scheduled Arrival)

### What's missing:

1. External influences
2. Predicting Early Arrivals
3. Estimating arrival times of future trips

## OUR SOLUTION

- Take into account external influences
- Test Gaussian Process Regression, Linear Regression, and Random Forest models on SEPTA data
- Allow passengers to predict lateness for any destination along regional rails at any time

| Train ID | Distance from Origin | Pause at Station | Time of Day | Day of Week | Month |
|---|---|---|---|---|---|
| 1 | 2 (stations) | 33 (seconds) | 43,200 | 7 | 4 |

# Results

- Random Forests produced the most accurate prediction of lateness for all rail lines
- Interactive code allows users to predict how late they will be to their desired station within 2.5 minutes in most cases

```
Please enter the source station: Marcus Hook
Which station do you want to know the status for?: Ridley Park
Do you want to know the status at the current time? (y/n): y
The train is predicted to be 0:01:00 late (H:MM:SS).
Do you want to quit?(y/n): n
Please enter the source station: Marcus Hook
Which station do you want to know the status for?: University City
Do you want to know the status at the current time? (y/n): y
The train is predicted to be 0:01:23 late (H:MM:SS).
Do you want to quit?(y/n):
```