Project 2 Proposal
Predicting Bus Delays

Catherine Magsino

Background

King County Metro operates the 10th largest fleet of buses in the United States, running over 1500 buses on a daily basis. With ridership falling above half a million each day, many of the riders depend on the bus system for their daily commutes. In everybody's busy schedules, it is helpful to understand when to expect that a bus will be delayed, in order to get to their destinations in time.

As a new Metis student, I've recently started riding the bus to Downtown Seattle from Renton. Inspired by multiple days in which my bus was ~20 minutes late, I've decided to work on a regression model that will help predict bus delays during my time at Metis. This will therefore help me plan my trips more effectively.

MVP

The MVP for this project will consist of data for the time period between September 23, 2018 to December 13, 2018, focusing on Route 143. This will reflect a time period similar to my time as a student at Metis, for the specific bus route that I take each day. Later iterations may include additional years and more bus routes.

Data

Data for this project will come from the following resources:

- King County Metro Data: https://metro.kingcounty.gov/GTFS/
- King County Metro Schedule: https://kingcounty.gov/depts/transportation/metro/schedules-maps/route/143.aspx
- Weather Underground: https://www.wunderground.com

The King County GTFS data comes in multiple zip files over specific periods of time. Each zip file consists of 12 .txt files that I will need to combine in order to retrieve the data I need. Since the GTFS data does not provide the number of minutes in which a bus was delayed, I will need to webscrape the schedule for the bus route. This will give me the time in which the bus should have arrived, and I'll be able to calculate the delayed time by subtracting the scheduled time from the actual arrival time. Additionally, I am interested to see if weather has an effect on bus delays. This data will be retrieved by webscraping daily/hourly weather data from Weather Underground.

Feature	Data Type	Additional Description
Route_id	string	
Trip_id	string	
Stop_id	string	

Date	datetime	
Day of the Week	datetime	Use datetime to get day of the week
Month	datetime	Use datetime to get month
Arrival Time	datetime	
Arrival Hour	datetime	Use datetime to get hour
Scheduled Time	datetime	Merged from King County Metro Schedule
Delay Time	integer	Number of minutes delayed
Exception Type	integer	Indicates service exceptions due to holidays
Number of remaining stops	integer	Calculated field based on the number of stops remaining at the location for the day
Temperature	integer	The hour’s temperature in farenheit
Precipitation	float	The hour’s number of inches of precipitation
Weather Condition	String	The hour’s weather condition: cloudy, mostly cloudy, light rain, rain, etc.

Known Unknowns

- The bus I take (143) is an express bus, so it does not run throughout the entire day
- Weather data is based on the hour and not the exact time of bus arrival