Utilizing Machine Learning to Predict Bus Delays
Jungmin Ahn, Joseph Knoth, Steven Nguyen, Tommy Zhang

King County Metro’s (KCM) current linear-based bus arrival prediction system is often inaccurate, negatively affecting the public trust of apps that use the system. Examples of apps that can benefit from a better prediction system: King County Metro's (KCM) arrival predictions and other sources that can improve the accuracy of bus arrival predictions.

Determine best range of delays
- Not outlier if (Q1 - 1.5IQR) ≤ x ≤ (Q3 + 1.5IQR)
- Applied to delay and speed

Hyperparameters to Test:
- Number of Rows: 5000 - 100000 rows
- Number of Decision Trees: 1000 - 20000 rows

Random Forest Model
Training and Testing on Dataset
Determine best model for prediction of delay

Delay = Scheduled Arrival Time • Actual Arrival Time

Hyperparameters to Test:
- Learning Rate: 0.1 to 0.0001
- dim(hn): 10 - 100 Nodes

Data Collection
Collected the following historical data from 2021-2022 for use in the models:
- Stop-Level KCM Data
- Seattle Weather Data (OpenWeatherMap)
- Traffic Data (WSDOT, Tracflow and SDOT)

Important columns include predicted/actual bus arrival times, hourly temperature, and traffic count

Neural Network
Created a regression model which predicts bus delay and a classification model which predicts on a binned range of delays

Neural Network Results
Mean absolute error for all routes using the regression model
- 2.5 min
- 71% accurate
- < 2 min

Random Forest
Created a random forest regression model in R which preprocesses the dataset, trains, and predicts the delay

Random Forest Results
Mean absolute error for all routes using the random forest model
- 2.4 min
- 7.8 min
- Coverage of the random forest model with data provided to the model
- 4 routes

Important columns include predicted/actual bus arrival times, hourly temperature, and traffic count

SQL
- Filtered out invalid times and joined datasets based on hour
- Delay calculation from actual and scheduled arrival time

Outlier Detection
- Neural Network
  - Not outlier if -900sec ≤ x ≤ 900sec (for classification compatibility and outlier removal)
- Random Forest
  - Not outlier if (Q1 - 1.5IQR) ≤ x ≤ (Q3 + 1.5IQR)

Special thanks to our project sponsors Al-Ahad Ekram, Valerie Harris and Sidney Quach, and our project advisors Patty Buchanan and Michelle Song.