BUS in Time

Utilizing Machine Learning to Predict Bus Delays

Jungmin Ahn, Joseph Knoth, Steven Nguyen, Tommy Zhang

King County METRO

1 PROBLEM STATEMENT

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: Google Maps, OneBusAway

2 CURRENT SYSTEM

Linear model based on:

Speed + Distance to Next Stop

High error margin and utilizes only live bus attributes

3 ASSUMPTIONS AND CONSTRAINTS

- No events that cause significantly large delays
  - Weather, sports, etc.
- Buses have same driving pattern
- Constrained data sources
  - Traffic data covering our routes were difficult to find

4 OBJECTIVE

Design a machine learning model based on historical data from KCM and other sources that can improve the accuracy of bus arrival predictions

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

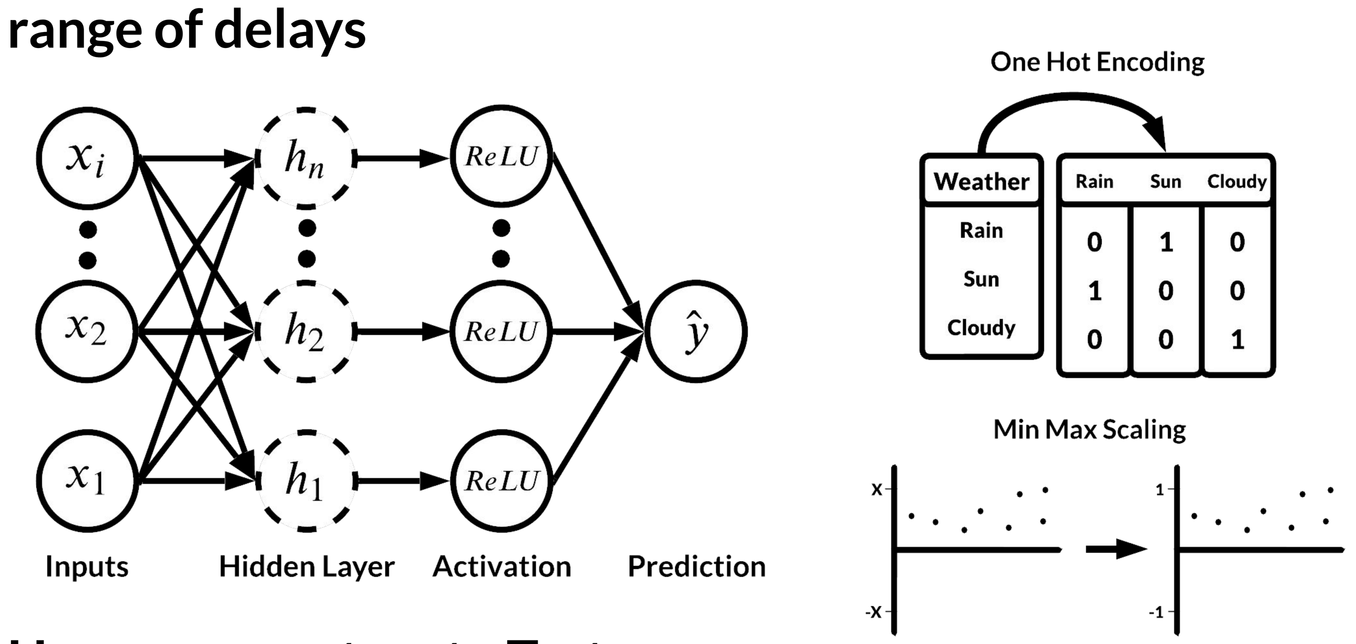
DATA ENGINEERING

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)
    - Applied to delay and speed

NEURAL NETWORK

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



Hyperparameters to Test:

Learning Rate 0.1 to 0.0001

$dim(h_n)$ Nodes in Hidden Layer 10 - 100 Nodes

RANDOM FOREST

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

Random Forest Model built using the Ranger package → Training and Testing on Dataset → Determine best model for prediction of delay

Delay = Scheduled Arrival Time - Actual Arrival Time

Hyperparameters to Test:

Number of Rows 5000 - 100000 rows

Number of Decision Trees 1000 - 20000 rows

NEURAL NETWORK RESULTS

2.5 min Mean absolute error for all routes using the regression model

71% accurate Predicting on a binned range of 6 minutes using the classification model

< 2 min Time required to train with 25 epochs (until convergence)

RANDOM FOREST RESULTS

2.4 min Mean absolute error for all routes using the random forest model

7.8 min Average time taken to train/predict on random forest model

4 routes Coverage of the random forest model with data provided to the model

Plots and Visuals

RapidRide A Line MSE | RapidRide E Line MSE | Importance Plot | Route and Traffic

5

6 IMPACT METRICS

42% of KCM customers say traveling by bus takes too long

25% say traveling by bus does not offer enough flexibility for their schedule

~50% ridership drop due to COVID, yet to be fully recovered

7 TAKEAWAYS

Changes Pre-COVID -> Now

Data sources often lack complete coverage of routes, making their reliability and granularity inadequate.

Better Ongoing Documentation

To ensure current and future production implementation success, documenting assumptions and reasoning is crucial due to evolving data sources.

8 NEXT STEPS

Model Iteration

Leverage the insights from the previous capstone (clustering) and this project's approach to enhance the development of an improved predictive model

Deploy to Production

Integrate the model architecture with KCM's live prediction system.

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