

AI 539 Final Project: Gotham City Cabs

Quick Overview

Each team can have a maximum number of 2 members, and would need to perform both projects.

This is a regression problem, where a model needs to be trained to predict the travel time of a cab from one point in the city to the other. You would see that the times in the data are around year 2034. However, the user still can extract the day of the week from the times given in the data.

Problem Setup



It is around the year 2034 in the city of Gotham, and the last time Batman got into a fight with Joker, the Batmobile (Batman's fancy car) was seriously damaged. Apparently, it takes his butler, Alfred, a while to fix the car and during that time Batman needs to use a cab to do his world saving business.

Alfred needs your help to come up with a good prediction of the taxi trip duration between multiple points of the Gotham city. If he can make such predictions, then that significantly helps with Batman's missions.

Lucius (Batman’s tech support staff) has been able to pull out a rich dataset of the recorded taxi durations between various parts of the city and is sharing that with you for your modeling purposes.

The input features of the aforementioned data file are:

- **pickup_datetime**: a variable containing a date and a time specifying the date and the time the taxi picked up a passenger. For instance, you may observe a `pickup_datetime` of “6/14/2034 3:00:00 AM”, which indicates the time the taxi picked up the passenger. Note that you may also obtain the day of the week, or the season information from this dataset. For instance, if we look up the 2034 calendar (search it on Google), you would see that “6/14/2034” is a Wednesday.
- **NumberOfPassengers**: The number of passengers loaded to the cab. Note that all passengers are loaded together and are dropped off together. So in case you see 5 people loaded, it means they all got to the cab together and all got off at the destination.
- **pickup_x**: This is a variable that represents the x coordinate of the location the taxi picked up the passenger.
- **pickup_y**: This is a variable that represents the y coordinate of the location the taxi picked up the passenger.
- **dropoff_x**: This is a variable that represents the x coordinate of the location the taxi dropped off the passenger.
- **dropoff_y**: This is a variable that represents the y coordinate of the location the taxi dropped off the passenger.

The response variable is:

- **duration**: which is the duration of the trip in seconds.

Modeling Instructions

You would need to use the file `Train.csv` to fit your models. Your model can take `pickup_datetime`, `NumberOfPassengers`, `pickup_x`, `pickup_y`, `dropoff_x` and `dropoff_y` as inputs and should be able to predict the quantity `duration`.

Please make sure to communicate with the instructor and Ed about any potential questions related to the data.