Predicting car crashes based on weather

Many thanks to our mentor, Julian Rosen

Guiding questions and data source

- What effect does weather have on car crashes?
- More precisely, given a city, interval of day (e.g., an hour, a day), and the weather, can one predict how many car crashes there will be?
- There is a huge dataset of car crashes in the United States maintained by Moosavi et al. It is available at https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents.
- At the time of the project it has 7.7 million records.
- We decided to focus on the 10 largest cities in the US.

Features

- Our features were the city, the time of day as an hour block, and the weather.
- The cities and weather were encoded as one-hot variables.
- For the time of day we used a circular encoding: $\beta_1 \cos\left(2\pi \cdot \frac{hour}{24}\right) + \beta_2 \sin\left(2\pi \cdot \frac{hour}{24}\right)$.
- The circular encoding is meant to keep things like 11pm and 1am close together.
- Some of our models also included interaction terms coming from the categorical variables.

Methods

- We tried a variety of regression methods including minimizing the residual square error, ridge regression, and elastic net regression.
- For each of those three we had a model that included and did not include the interaction terms.
- Our baseline model was simply the average over all (city, hour) pairs. That is, not including the weather at all. For example, there were about 4.03 crashes in Los Angeles between 12pm and 1pm.
- We compared the models using the mean squared error for 5-fold cross validation.

Results and further direction

- None of our six models beat the baseline model though the closest was the ridge regression with interaction terms.
- The baseline model can be viewed as a special case of a regression on a decision tree, so one could try to build a decision tree or random forest model and hope to get better performance that way.
- A problem is that some (city, hour) pairs have no accidents in the dataset. We would have to find the weather for all of those to fill in the data with zero accidents for that block. Remedying this could provide better performance as well.