Team 8 Road Safety & Closing Roads II – Executive Summary

Assaf Bar-Natan, Jesse Frohlich, Jacob van Hook, Pedro Lemos

GitHub

Objectives:

- Use regression modeling to predict collision rates in New York City using public road data and other geographical features.
- Identify key road infrastructures that contribute to higher collision rates.

Data:

- Road features collected from NYC Open Data[1] including street width, traffic volume, trees, speed humps, speed limits, and bike lanes.

Data Processing:

- Use geographic data for objects such as trees or speed bumps to associate each to a road segment.
- Use geographic and time data of collisions in order to compute a rate of collisions for each road segment.
- Bifurcated road data when speed bumps were installed in order to accurately assess the difference before and after a speed bump was installed.

Models:

- **Baseline Model**: This model always predicts the mean collision rate across the entire city for every road segment, regardless of features. This is important to have because it gives us a basis of comparison by which to measure our other models. If a model cannot do better than just predicting the average every time, it is likely not a very good model.
- **Linear Regression**: This straightforward model uses linear regression and predicts a linear relationship between features, or combinations thereof, and the collision rate. This is also a good basis of comparison and can be used to assess the effectiveness of more complicated models.
- **Random Forest Regression and XGBoost**: We also want to test the effectiveness of a random forest regressor. This may prove useful as a way to determine relationships which we would not be able to see through standard linear regressions.
- **K-Nearest Neighbors Regression**: We want to look at the predictive power of a K-Nearest Neighbors model because roads with similar features are likely to have similar collision rates and can therefore serve as a predictor.

Results:

- All models outperformed the baseline model.
- The K-Nearest Neighbors model outperformed all other models, having a test MSE of 3.48×10^{-7}.

Implications:

- The most important feature in predicting the collision rate of a road segment is the average traffic volume, followed by the width of the street.

---

[1]Source: NYC Open Data (https://opendata.cityofnewyork.us/)