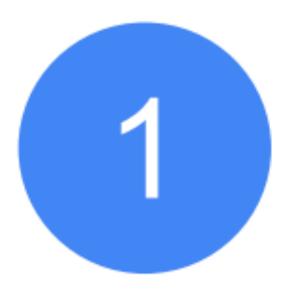
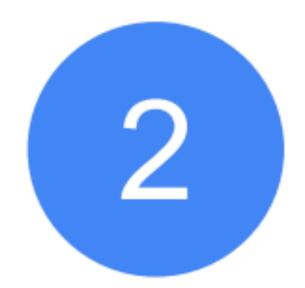
Predicting Motor Vehicle Crashes Severity

Amanda Curtis, Arthur Diep-Nguyen, Olti Myrtaj, Brandon Owens, Fabio Ricci

Erdős Institute Spring 2025



There are more than 5 million motor vehicles crashes (MVCs) every year in the US



Predicting the frequency and severity of MVCs as a function of built-environment features in the surrounding area can help

Overview



Built Environment

Features:

- Roads
- Crosswalks
- Transit Stops
- Walkways
- Stop Signs and Traffic Lights

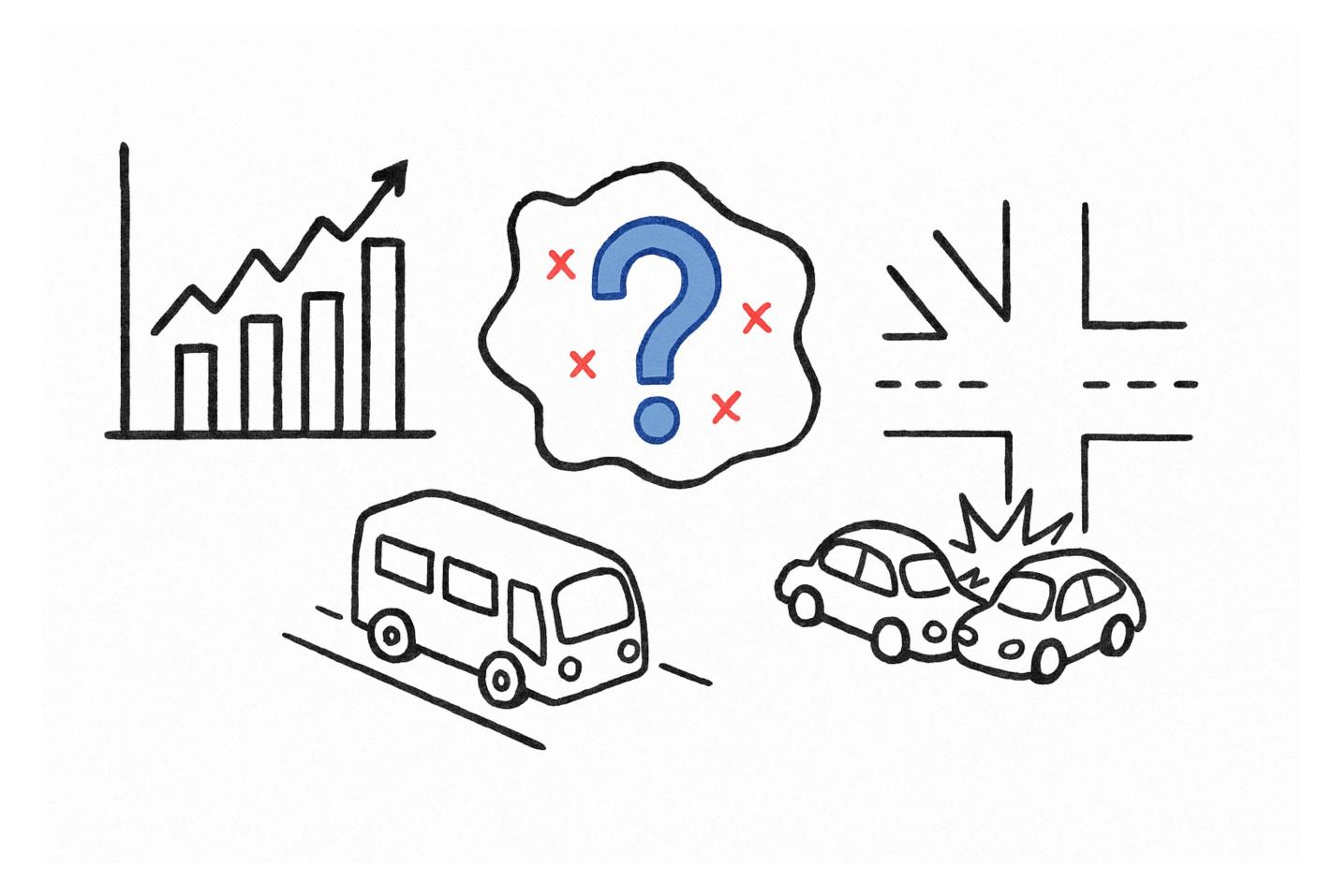


https://publichealth.harriscountytx.gov/Divisions-Offices/Divisions/Environmental-Public-Health/Built-Environment-BE-Program/Built-Environment-101



Research Question

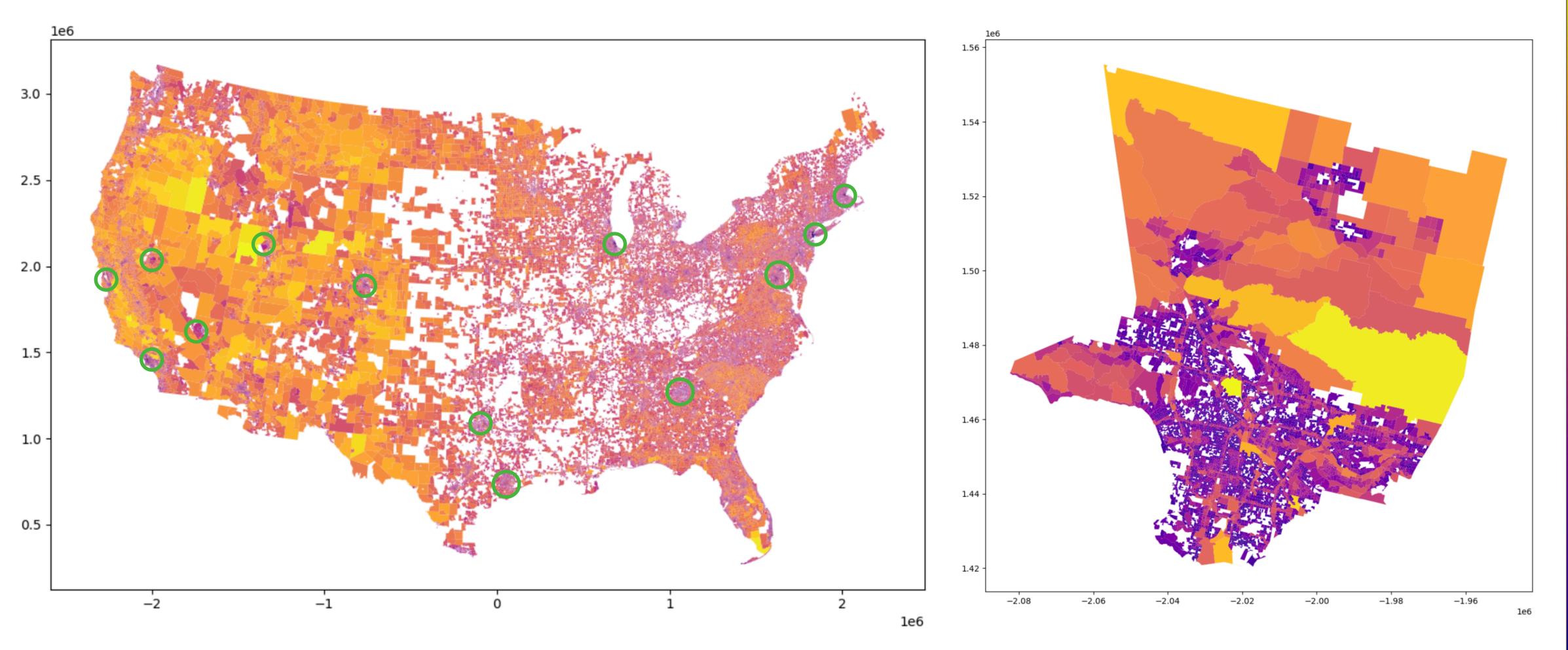
How can we use features of the built environment in a given area to predict the "crash density"?



Defining "Crash Density"



Engineered Target Variable = (Crashes * Severity)/(Population Density)



Crash density is lowest in major metro areas

Visualizing Crash Density

In LA County, crash density is greater along freeways and smaller in towns



Data Augmentation

EPA Data Set: Data based on Census Block Group

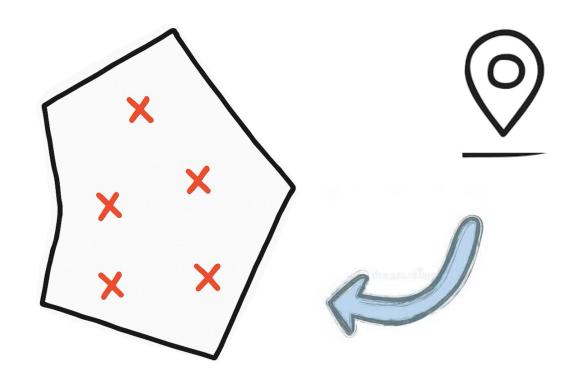
https://www.epa.gov/smartgrowth/smart-location-mapping

- > 200k census block groups
- Housing density, diversity of land use, walkability, neighborhood design, destination accessibility, transit service, etc.

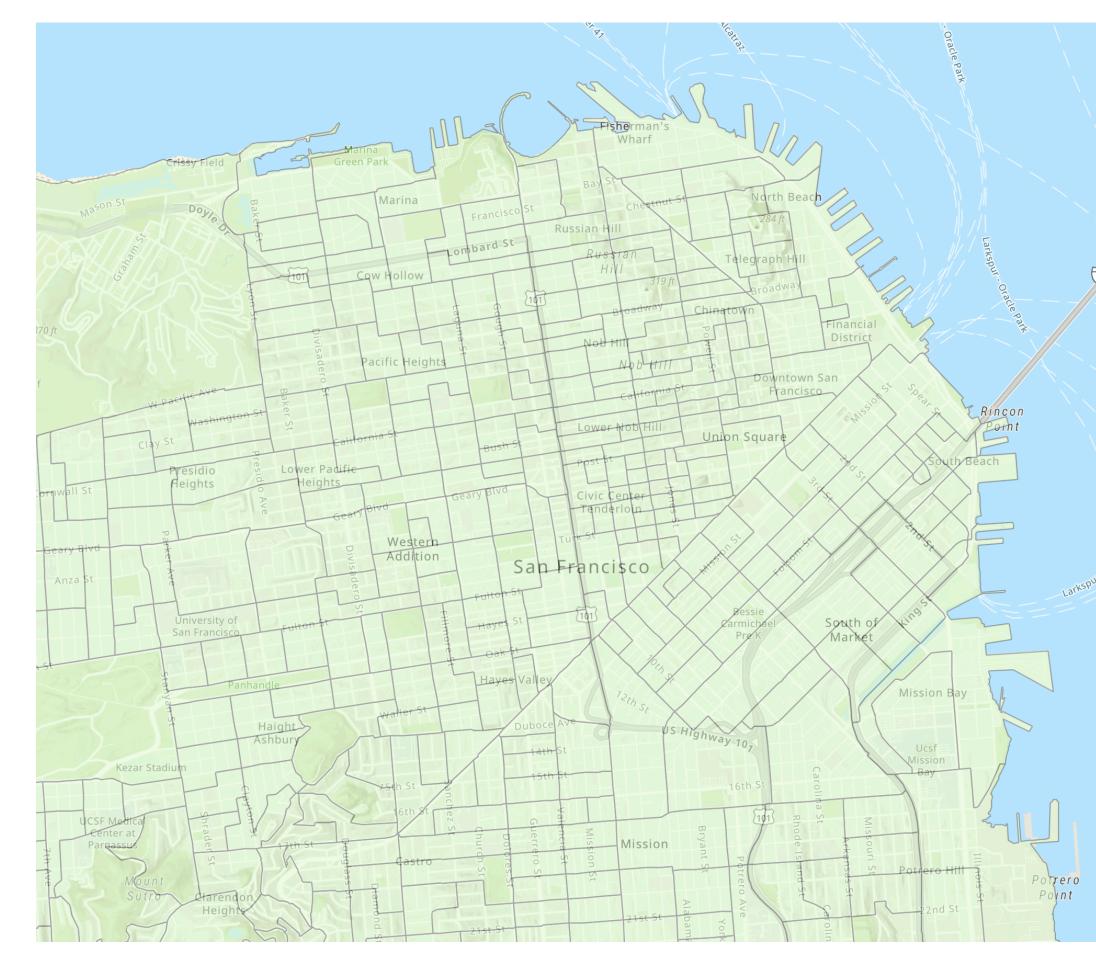
Kaggle Data Set: Single Crashes

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

- 7.7 million crashes from 2016-2023
- Location and severity of crashes



San Francisco, California - Census Block Groups



https://www.arcgis.com/apps/mapviewer/index.html?layers=2f5e592494d243b0aa5c253e75e792a4

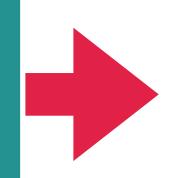


Data Processing

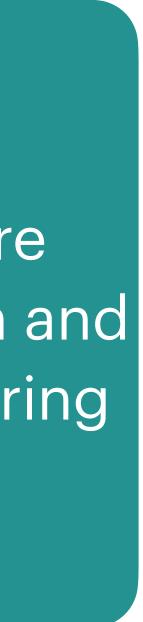
Import and clean data

Convert to same CRS (Coordinates Reference System) Aggregate crashes into census block groups using GeoPandas

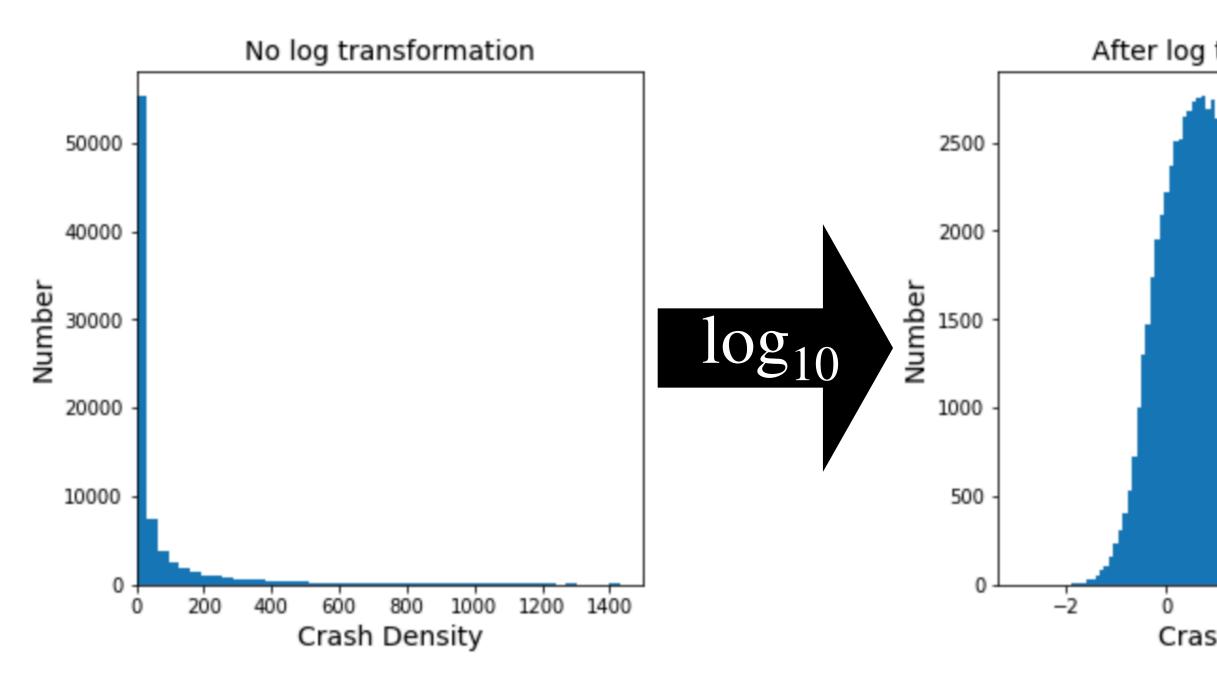
Merge Kaggle and EPA data



Feature selection and engineering

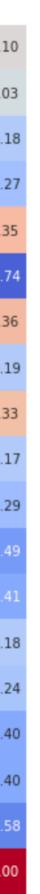


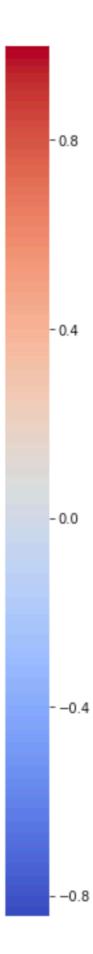
Feature Selection



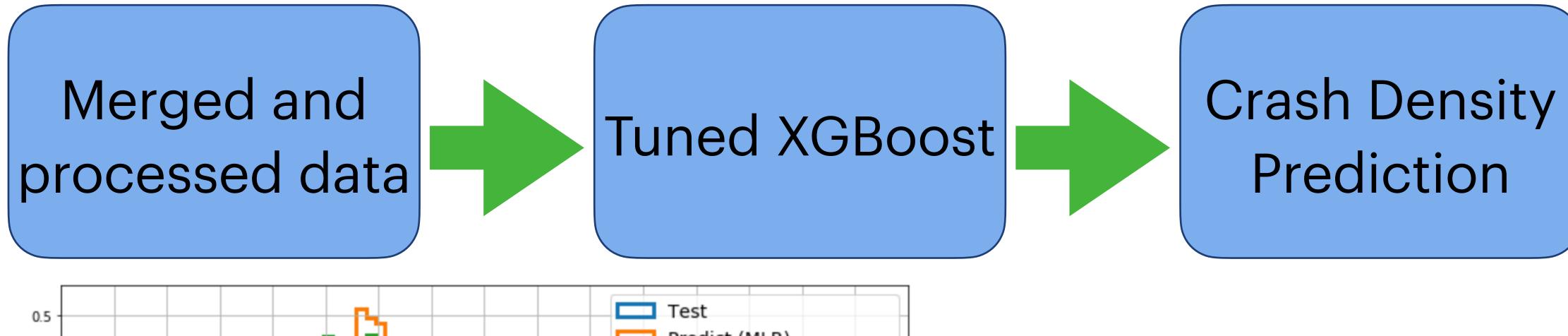
We apply a log_{10} scaling to highly skewed features to increase interpretability

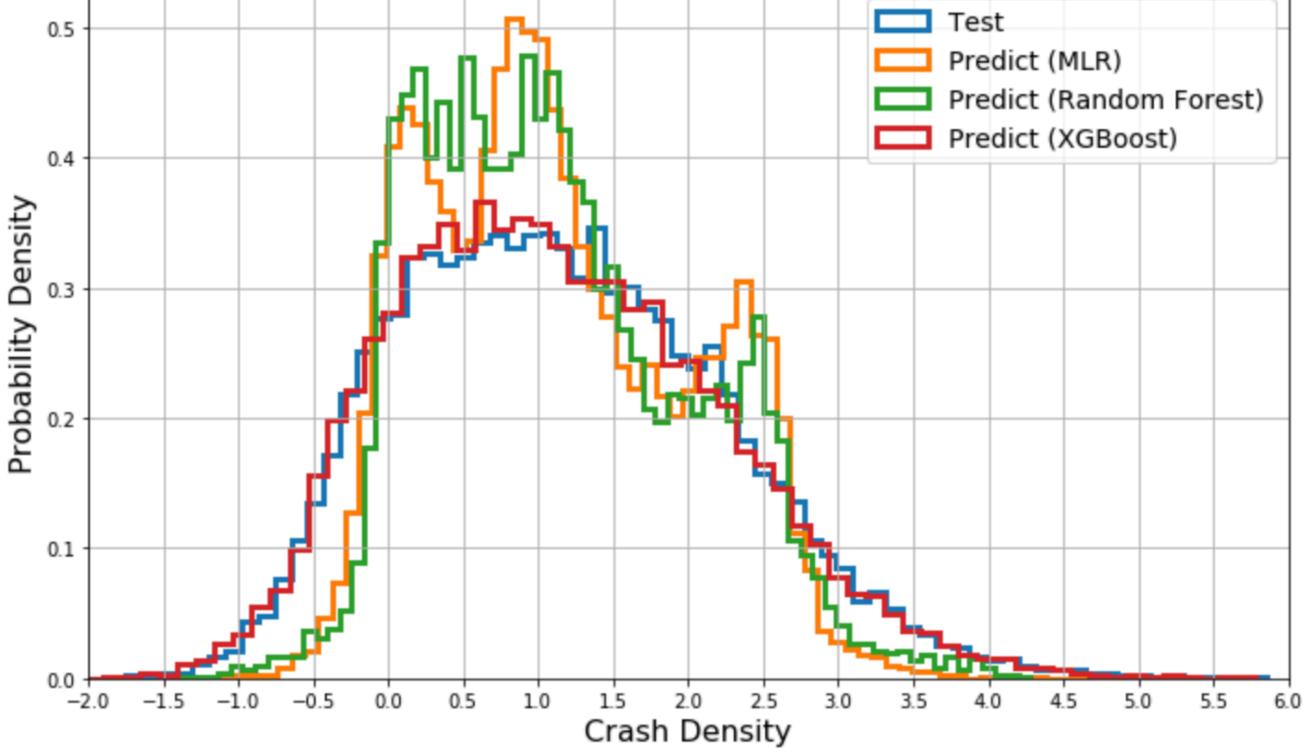
	1.00	-0.03	0.05	-0.03	-0.01	-0.15	-0.05	-0.05	-0.06	0.02	-0.05	-0.14	-0.05	0.02	0.00	-0.16	-0.03	-0.13	0.10
	-0.03	1.00	0.03	0.04	0.01	-0.07	0.10	0.01	0.11	0.05	0.01	-0.03	-0.16	-0.01	-0.04	-0.10	-0.09	-0.08	0.03
transformation	0.05	0.03	1.00	0.27	-0.59	0.22	-0.05	0.12	-0.02	0.10	0.16	0.13	0.19	0.18	0.21	0.19	0.22	0.28	-0.1
	-0.03	0.04	0.27	1.00	-0.84	0.38	-0.03	0.14	-0.01	0.13	0.19	0.26	0.22	0.11	0.15	0.17	0.34	0.38	-0.2
	-0.01	0.01	-0.59	-0.84	1.00	-0.45	0.06	-0.20	0.03	-0.16	-0.26	-0.28	-0.30	-0.28	-0.32	-0.32	-0.41	-0.47	0.3
	-0.15	-0.07	0.22	0.38	-0.45	1.00	-0.09	0.32	-0.06	0.29	0.44	0.64	0.52	0.21	0.28	0.53	0.59	0.80	-0.7
	-0.05	0.10	-0.05	-0.03	0.06	-0.09	1.00	-0.03	0.83	-0.07	-0.14	-0.14	-0.23	-0.04	-0.07	-0.06	-0.04	-0.15	0.3
	-0.05	0.01	0.12	0.14	-0.20	0.32	-0.03	1.00	-0.00	0.69	0.63	0.09	-0.00	0.14	0.16	0.23	0.24	0.29	-0.1
	-0.06	0.11	-0.02	-0.01	0.03	-0.06	0.83	-0.00	1.00	-0.04	-0.09	-0.11	-0.19	-0.02	-0.04	-0.04	-0.03	-0.10	0.33
	0.02	0.05	0.10	0.13	-0.16	0.29	-0.07	0.69	-0.04	1.00	0.51	0.12	-0.02	0.08	0.09	0.14	0.20	0.27	-0.1
	-0.05	0.01	0.16	0.19	-0.26	0.44	-0.14	0.63	-0.09	0.51	1.00	0.21	0.19	0.16	0.18	0.26	0.31	0.44	-0.2
	-0.14	-0.03	0.13	0.26	-0.28	0.64	-0.14	0.09	-0.11	0.12	0.21	1.00	0.47	0.15	0.20	0.33	0.32	0.60	-0.4
	-0.05	-0.16	0.19	0.22	-0.30	0.52	-0.23	-0.00	-0.19	-0.02	0.19	0.47	1.00	0.16	0.21	0.28	0.33	0.51	-0.4
	0.02	-0.01	0.18	0.11	-0.28	0.21	-0.04	0.14	-0.02	0.08	0.16	0.15	0.16	1.00	0.80	0.34	0.15	0.26	-0.1
2 4 6	0.00	-0.04	0.21	0.15	-0.32	0.28	-0.07	0.16	-0.04	0.09	0.18	0.20	0.21	0.80	1.00	0.43	0.19	0.33	-0.2
sh Density	-0.16	-0.10	0.19	0.17	-0.32	0.53	-0.06	0.23	-0.04	0.14	0.26	0.33	0.28	0.34	0.43	1.00	0.36	0.52	-0.4
	-0.03	-0.09	0.22	0.34	-0.41	0.59	-0.04	0.24	-0.03	0.20	0.31	0.32	0.33	0.15	0.19	0.36	1.00	0.53	-0.4
	-0.13	-0.08	0.28	0.38	-0.47	0.80	-0.15	0.29	-0.10	0.27	0.44	0.60	0.51	0.26	0.33	0.52	0.53	1.00	-0.5
	0.10	0.03	-0.18	-0.27	0.35	-0.74	0.36	-0.19	0.33	-0.17	-0.29	-0.49	-0.41	-0.18	-0.24	-0.40	-0.40	-0.58	1.0
kewed					·														





Modeling Process

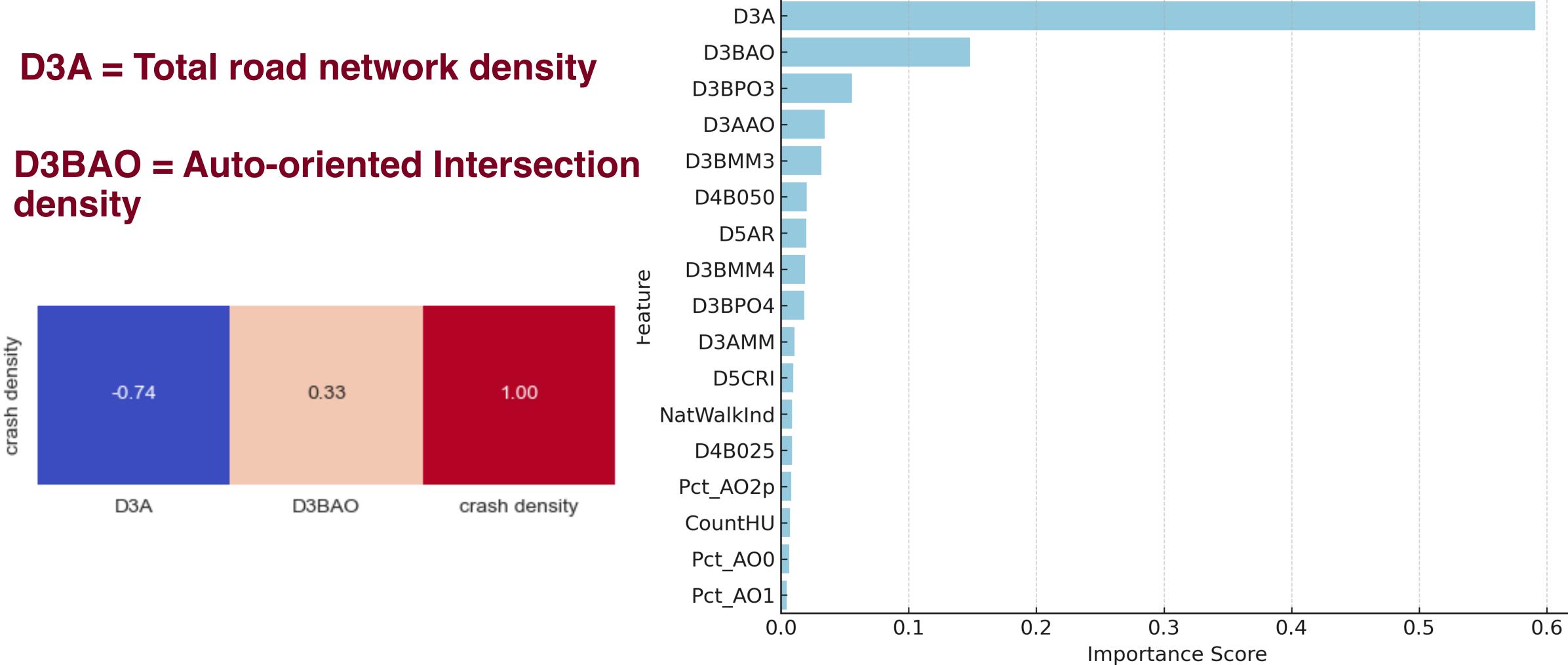




Model	RMSE
XGBoost	0.552
Random Forest	0.576
MLR	>0.6
Lasso	>0.6
Ridge	>0.6



Highlights of important features



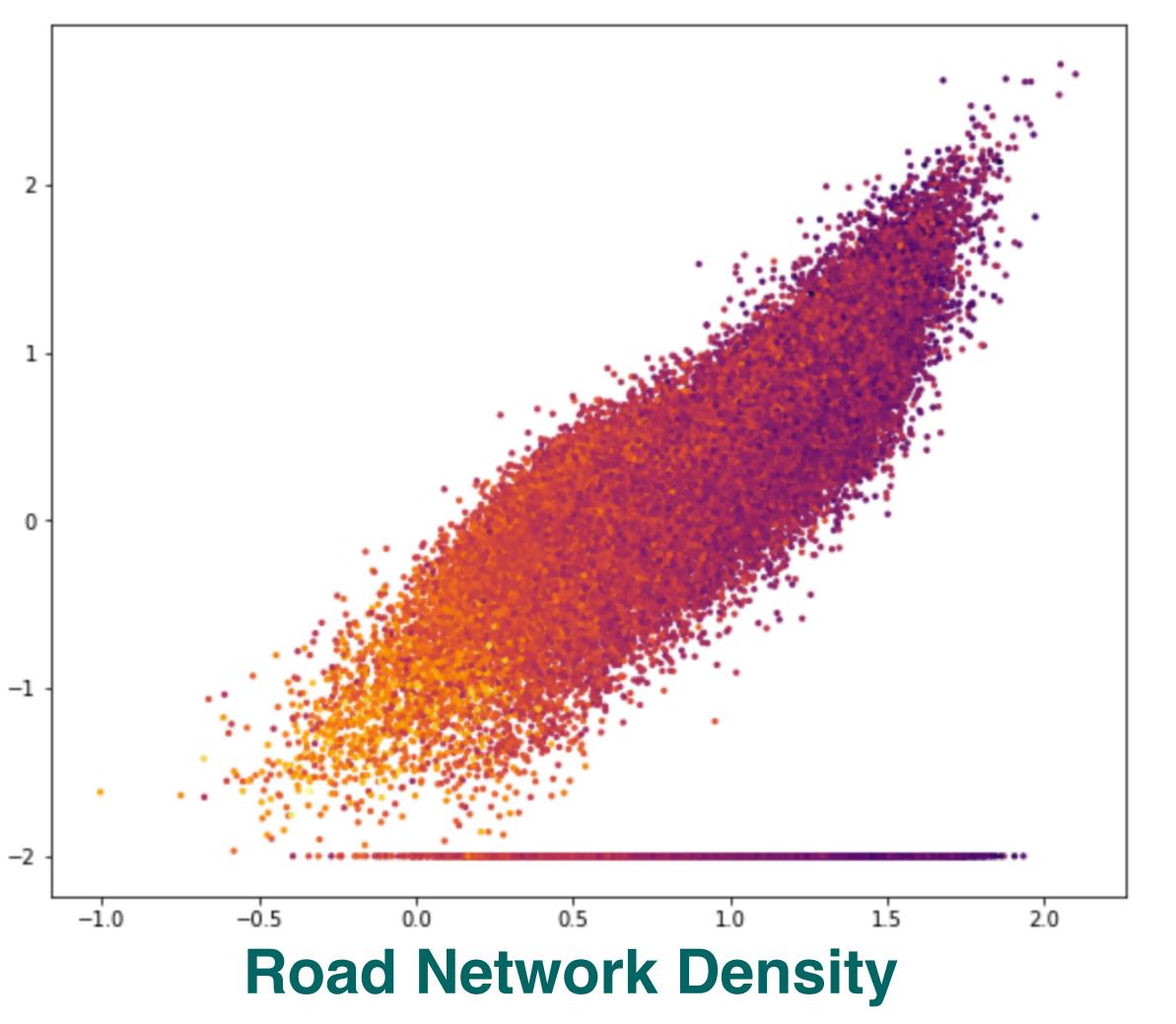
XGBoost Feature Importance

Conclusions

Densest road networks have lowest crash density

0 Oriel

5



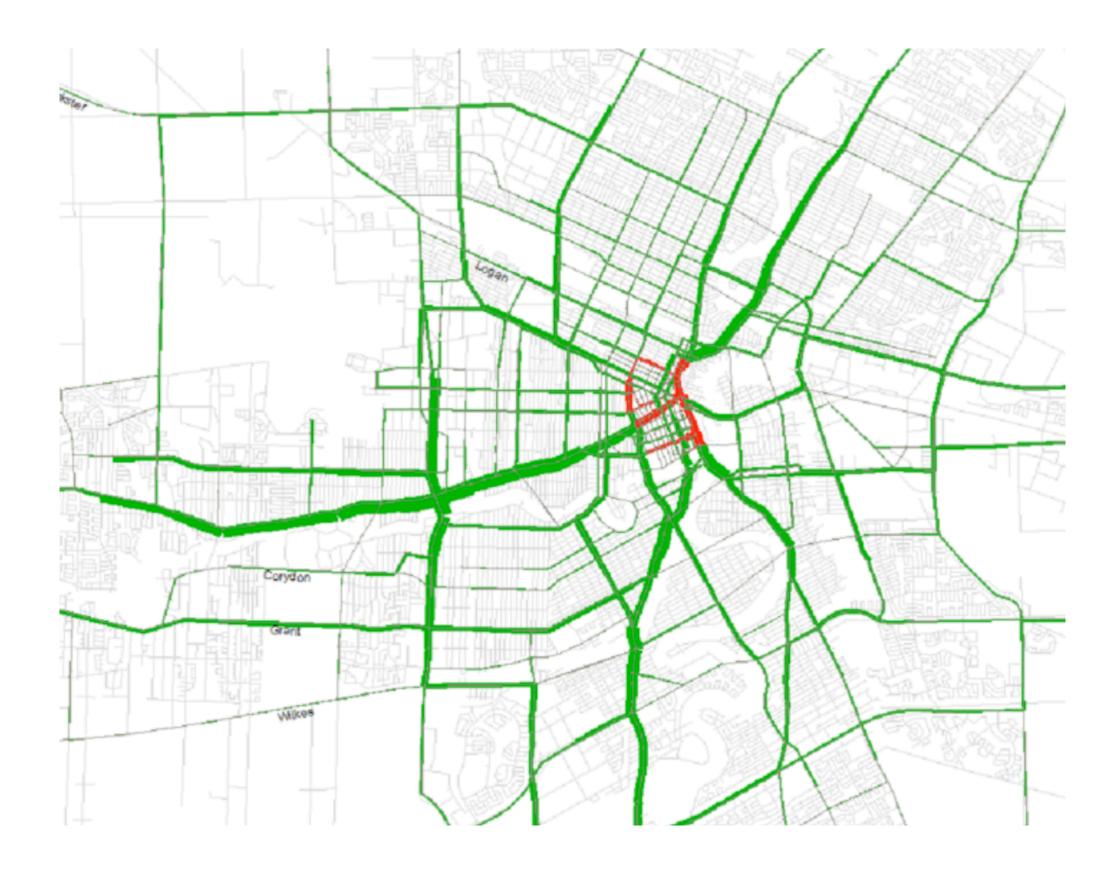


Future Work

Implement inference tests to study the impact built environment features have on Motor Vehicle Crashes

Design city plans that minimize number and severity of crashes using generative AI trained on **CBG** data







Thank You

ő The Erdős Institute

Thanks to Steven Gubkin, Alec Clott and Shravan Patankar