```
In [1]: %%HTML
        <style>
        .rendered_html table, .rendered_html th, .rendered_html tr, .rendered_html td {
            font-size: 90%;
        }
        </style>
```

Indicators of Crash Severity

UrbanLogiq Capstone

Sayanti Ghosh, Jes Simkin, Evan Yathon, Mike Yuan

Disclaimer ⚠

All images and data are either from open source Toronto data or fabricated data, and are **not** from the client city.

Agenda

- Business Objectives
- Data & Data Analysis
- Results & Data
  Product

About UrbanLogiq

"Building better communities with data"

- Primarily aggregate data for client cities
- Also provide analytic services for government bodies
- Extract useful insights from aggregated data for their clients

Vision Zero

All traffic fatalities and serious injuries are preventable

- A traffic safety initiative to eliminate deaths and severe injuries on roadways.
- One traffic death is one too many.

Source: https://visionzeronetwork.org/about/what-is-vision-zero/

Business Objective

Identify factors and their effect on collision severity

- Hand findings of the client city to UrbanLogiq
    - These findings can be used to influence public policy
    - Move towards Vision Zero
    - Want to see features that can be 'turned with a knob' and see what their effect is

Distilling the Business Objective

A matter of inference

- Methods with feature interpretability and importance are key
    - Regression methods
    - Tree based methods
- Prediction scores are not the objective

Data

Geospatial and Collisions data

- Collision dataset is a record of police reports spanning a ten year time frame
- Geospatial data
    - Buildings, Roads, Sidewalks, Zoning, and more!
- Open source data
    - Bus stops
    - Economic and demographic data

Observational Unit and Response

Our observational unit is any collision related to an intersection

Accident Date	Street Lighting	Intersection	Minor Injuries	Moderate Injuries	Major Injuries	Fatal Injuries	Response
2018-02-15	Not lit	Fake Street & Not Real Ave	1	1	0	0	0
2018-01-23	Lit	Spaghetti Crossing	1	0	0	1	1
2018-12-12	Lit	Meatball Junction	0	0	0	0	0
2018-05-16	Not lit	Fake Street & Not Real Ave	0	0	1	0	1

Observational Unit Significance

Any findings are conditional on a collision having occurred

- Interpret findings as influencing collision severity, **not** influencing a collision occurring
- This aligns with the client's goal: Vision Zero
  - People make mistakes, let's try to ensure they aren't severe
- An alternative would have been to examine a collision occurring

Endgame Goals

Inference and Importance

- Ideal output is a list of features ordered by importance, with regression coefficients
- How do we get there?

How to Deal with Categories?

Dummy Coding

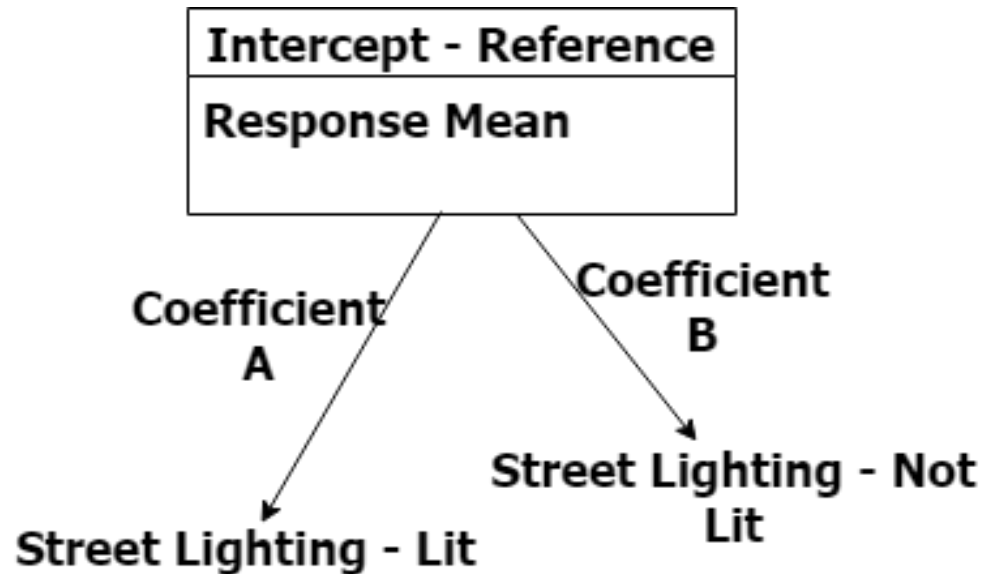| Intercept - Reference |
|---|
| Street Lighting - Not Lit |

Coefficient A

↓

Street Lighting - Lit

Intercept - Reference
Street Lighting - Lit

Coefficient B

↓

Street Lighting - Not Lit
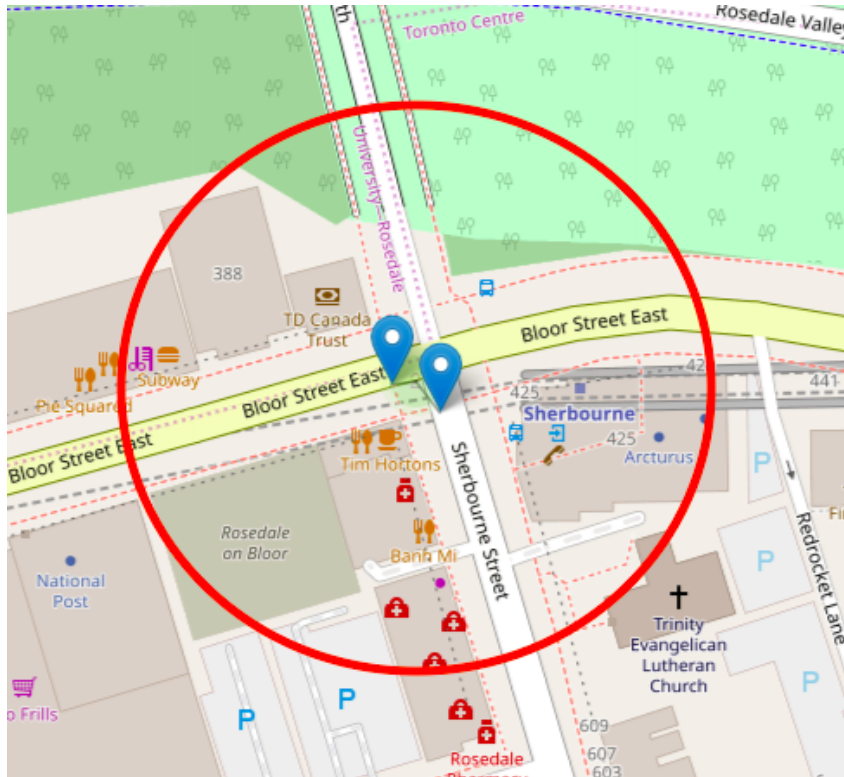
How to Deal with Categories?

Effect Coding[1]

1: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-effect-coding/

Feature Engineering

## Collision At Intersection

Engineered Features</div>

- Total Bus Stops
- Total Number of Buildings
- Mean Building Volume
- Total Residential Zoning Permits
- Total Commercial Zoning Permits
- Speed Limit Difference
- And much more!

Clustering by Neighbourhoods

Cities are big & diverse. Is our approach local enough?

- Not all intersections are created equal
- One set of indicators may not apply everywhere
    - eg. Downtown Core vs. Suburban Shopping Mall Complex
- Idea: Perhaps different areas of the city have different indicators
  💡
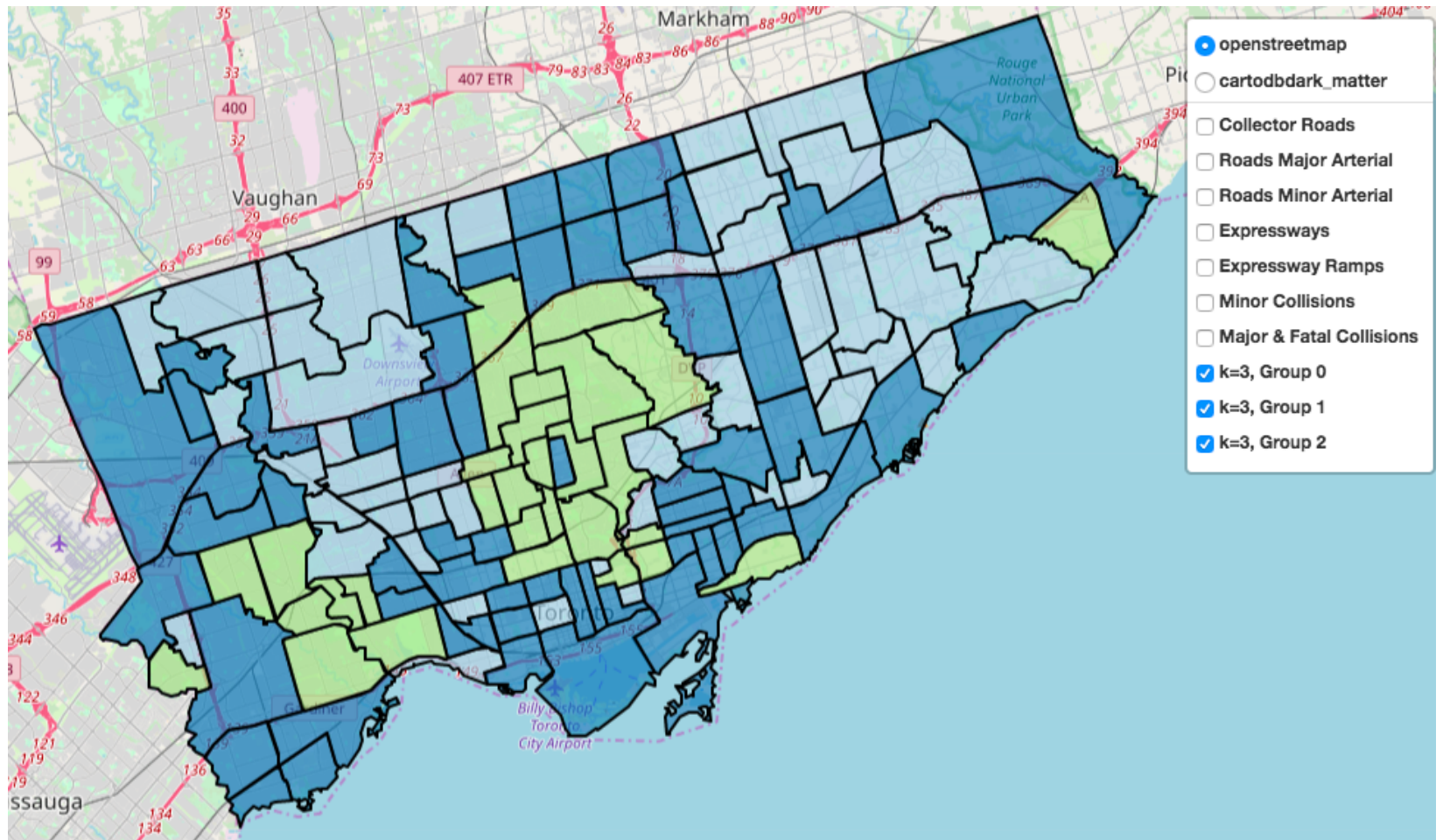
Feature Selection

More features, more problems

- L1 Regularization & XGBoost SHAP values
- UrbanLogic's domain expertise & intuition
- Interpretability

Source: giphy.com (https://gph.is/g/4MgdxbW (https://gph.is/g/4MgdxbW))

Models - Logistic Regression

Estimating odds effect of each feature

- Very useful to say that one unit increase of feature A leads to Z times increase in the odds of a more severe collision
- Created a bootstrap utility in order to give UrbanLogiq coefficient confidence intervals
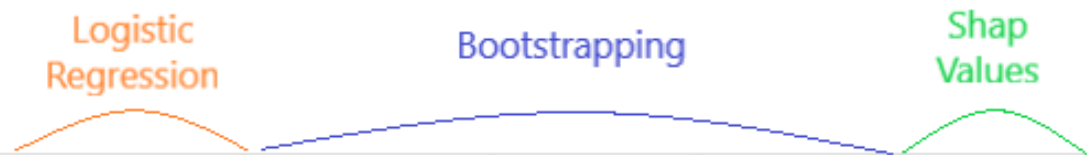
Models - XGBoost

Gettin' in SHAP

- XGBoost - To reduce variance and bias
- Using SHAP values to interpret complex XGBoost model prediction.
- SHAP values shows distribution of the impacts each feature has on the model output

SHAP Example
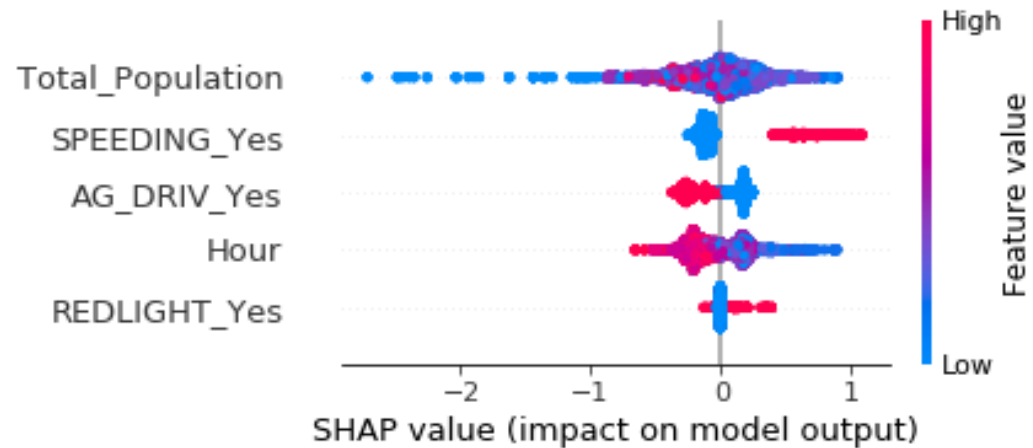
What does our Model Output Look Like?

Feature	Logistic Regression Coefficient	Bootstrapping Coefficient SD	Bootstrapping Lower 95% CI	Bootstrapping Upper 95% CI	Shap Values SHAP Value
Number of Bus Stops	0.25	0.08	0.10	0.40	0.25
Traffic Controls Not Functioning	-1.80	0.60	-2.95	-0.65	0.16
Collision occurred on a workday	-0.08	0.03	-0.12	-0.04	0.06
...	...	...	...	...	...

Predictors

Sample From Our Result

	All	Urban Core	Urban-Suburban	Suburban-Rural
Contributing Factors	No Traffic Controls	Nightlife Hours, Dark Street Light	Afternoon Rush Hour, High Speed Limit Difference	Number of Bus stops, Not Workday
Preventing Factors	Daylight	Daylight, Non-Rush Hour	CycleTrack, Non-Rush Hour	Workday, Not Functioning Traffic Control

Data Product

A scientific report and reproducible repo 🎉

- Format aligned with UrbanLogiq's best practices
  - Easy and ready to use!
- Repo includes:
  - Interactive maps of the data and features!
  - Well-documented functions and scripts!
  - Pickled models!

Interactive Map
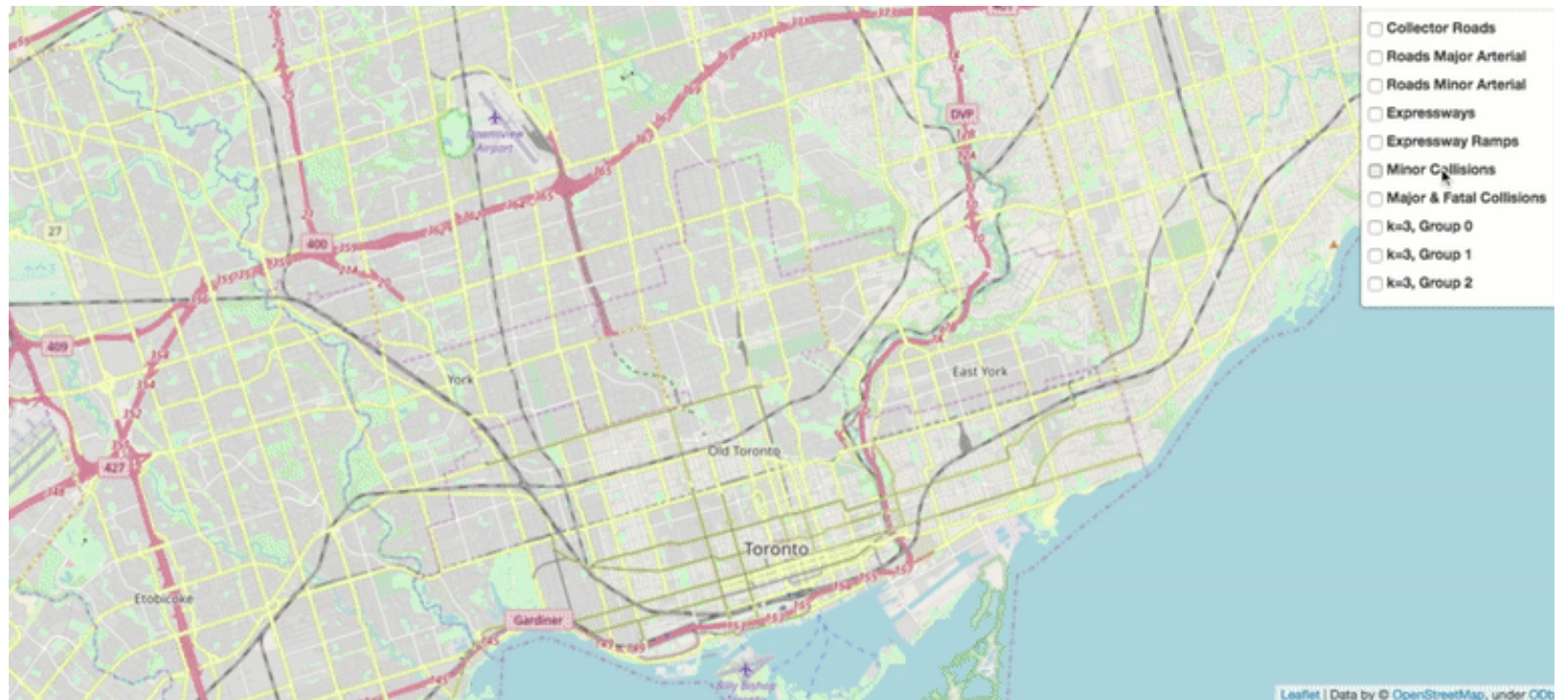
Conclusion

- Addressed business and technical objectives
- Defined an appropriate response variable
- Identified predictors for major and fatal collisions at the intersection level
- Confident in our data product
- Learned a lot

Thank You

- The UrbanLogiq team
- Capstone mentors

Questions?