What is a Crash Prediction Model? Look into this.

* Especially for Toronto
* Has there been a crash prediction model developed for Toronto? Ans. Doesn’t look like it.

Do I have to combine different data sets into one dataset before conducting analysis? Ans. Yes

Datasets

* Toronto KSI
* Toronto All Collisions
* Toronto Road Network
* Road Features (Speed limit etc.)
* Toronto Traffic Speeds
* Toronto Traffic Volumes
* Toronto Neighborhood Demographics
* Toronto Traffic Control (Data is there but have to merge it).
* Toronto Bike Lanes
* Road Infrastructure Data (Sidewalks, Number of Lanes etc.)
* Weather Data

Models

* Gradient Boosting/AdaBoost
* XGBoost
* Random Forest
* Linear Regression
* Neural Networks
* K Nearest Neighbors Classifier
* Deep Learning Neural Network

Error Statistics

Regressions

* Mean Squared Error: [sklearn.metrics.mean\_squared\_error — scikit-learn 1.3.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html#sklearn.metrics.mean_squared_error)
* Mean Absolute Percentage Error: [sklearn.metrics.mean\_absolute\_percentage\_error — scikit-learn 1.3.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_percentage_error.html#sklearn.metrics.mean_absolute_percentage_error)
* DMBA RegressionSummary
* Explained Variance Score: [sklearn.metrics.explained\_variance\_score — scikit-learn 1.3.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.explained_variance_score.html#sklearn.metrics.explained_variance_score)

Classification

* DMBA ClassificationSummary

AUC/ROC/Lift Charts

Questions to Model

* What is the probability that a collision will occur on this intersection at a given day?
* How many collisions are predicted to occur at this intersection monthly/annually?
* If a collision occurs at this intersection, what is the probability that it will be fatal?
* If a collision occurs, what is the predicted number of fatalities/injuries likely to happen?

You could also model collision “rates” like Ella. Essentially no. of collisions by neighborhood/hex cell divided by traffic volume. The ML model would predict the collision rate, or the fatality rate.

\*\*I will have to define the problem as a classification or a regression problem.

* David said classification might be easier to work with.

\*\*You might have to change your objective to “if a collision occurs at this intersection, what is the probability that it will be fatal?”

* This will enable you to avoid having to create negative examples.
* Could also make it a classification problem: major/fatal or minor. Or 3-stage classification – minor, major, fatal.

Decide on which metrics to compare the performance on the models on.

* The research paper says accuracy might not be a good choice and recall might be better.

The project could also be just predicting bike collisions or pedestrian collisions, does not have to be all collisions.

Include some cluster analysis and PCA in there.

Two important decisions to make:

1. What am I modeling? How will I make it into a risk map? Which dataset is my primary dataset?
2. What kind of spatial join will I use. Geopandas or Hex bin. Geopandas spatial join.

Figure out what I am modeling and decide which dataset will be the main dataset for merging.

Figure out join type to use (spatial join or hex cell)

* You can try both joins (spatial and hex) to see which one gives you more rows.
* Spatial join, if done right, should possible be more accurate than joining by hex cells.

You could potentially join the road segment map + intersection map to get every single road segment and intersection in Toronto.

* Or I could just do my project on intersections. Might be easier because I have more intersections data such as avg. travel speeds.

CRS Choices: 3857 (Distance in meters) or 4326 (Distance in degrees).

Generally, point geometry does not seem to cause an issue if you leave it’s CRS alone and is joined with sjoin\_nearest.

**Project Steps**

1. Finalize project objectives.
2. Gather datasets.
3. Merge all datasets together.
4. Conduct exploratory analysis for modeling. (make sure to include traffic volumes).
5. Data pre-processing.
6. Feature Engineering.
7. Modeling – Supervised and Unsupervised
8. Performance review and hyper-parameter tuning.
9. Pick best model/Model Selection
10. Interpretation and conclusion.
11. Create risk map of roads/intersections.

Write a medium blog post about this project.

**Processed**

1. Sensitive Zones
2. Collision Data (KSI + All Collisions)
3. Intersections and Travel Speeds
4. Traffic Volumes
5. Traffic Control
6. Cycling Network
7. Road Features
8. Posted Speed Limits
9. Automated Speed Enforcement
10. Intersections

**Datasets Left**

1. Weather (Only if doing collisions)
2. Locations of interest
3. Zoning
4. Residential Areas

**Ready for Merge**

1. Speed Enforcement
2. Traffic Cameras
3. All Collisions
4. KSI
5. Cycling Network
6. Intersections Travel Speeds
7. Posted Speed Limits
8. Road Features: Traffic Calming
9. Road Features: Lanes
10. Roads and Intersections (Primary Dataset)
11. Sensitive Zones
12. Traffic Control

**Merged**

1. Cycling Network
2. Lanes
3. Pedestrian Crossover
4. Traffic Calming
5. Speed Enforcement
6. Traffic Cameras
7. Speed Limits
8. Sensitive Zones
9. Traffic Volumes (May have to change merge procedure)
10. Collisions
11. KSI

**Still to do**

1. Collisions (As no. of collisions)
2. KSI (As no. of KSI collisions)

\*\*\*\***After all the merging is complete, double check all common columns between dataframes to see if spatial joints had issues.**

* **Currently, it seems that there is a mismatch between road names from road network data frame (from Toronto centreline) and road names from other data frames (such as speed limits)**
* **Double check if this issue still persists if the base road network is the Ontario Road Network Dataframe**

**\*\*I might have to estimate traffic volumes by building a model.**

* **After imputing traffic volumes with the model, run two different models with and without traffic volumes and check accuracy.**

**\*\*Or I can take only roads that I have traffic volumes for.**

**\*\*Or I take traffic volumes in a 1 km radius which leads to very little null values.**

**\*\*Or since you have traffic volumes for mostly arterial roads, you could build two models – one for local roads, one for arterial roads.**

**Feature Engineering**

1. Collisions/Total\_Volume
2. KSi/Total\_Volume
3. Percentage of Collisions that are KSI

**Skew Treatments**

* Cap and Floor
* Box Cox Transformation
* Log Transformation

**Decisions to Make**

1. Whether to use log transformation or cap & floor to reduce skews. Or both.

Q1: Does using LeaveOneOut CV make the results better?

Ans. Leave one out is too computationally expensive. Only try on best model.

Q2: Does Cap & Floor instead of Log Transformation make results better?

Q3: Consider removing total\_count, as it is correlated with the other traffic volume variables.

Q4: Map the very high volume road segments that you clipped to model them separately.

Q5: Check MAPE scores between model 4 and 5 again.

Notes:

* Collision Rates (With Volume) is the best model so far
* Both Collision Rates models are better than Collision Count Models
* For some reason, the models have a hard time differentiating between 5-20 classes and 80-100 classes. They classify a lot of 5-20s as 80-100, and a lot of 80-100s as 5-20. Otherwise all other classes are accurately classified.
  + Try changing the classification thresholds (e.g. 80 up only)