EVALUATING THE RISK OF TRAFFIC CRASH IN CHICAGO

Li Du
DATS6501 Data Science Capstone Project
Dec 4, 2018

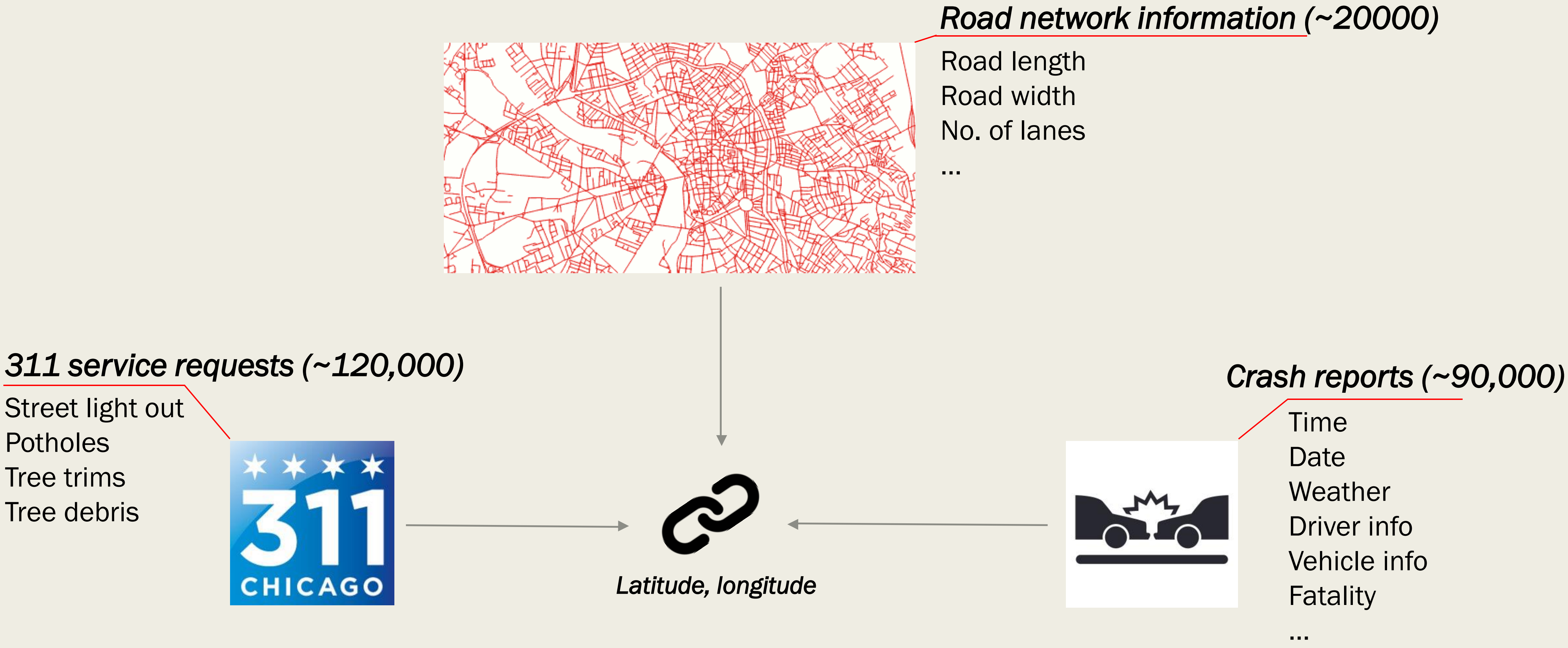
Motivation & Objectives

- Traffic crashes has been one of the major causes to transportation related deaths
  - *In 2016, over **37,000** lives lost due to traffic accidents*
- Traffic crashes result in significant amount of economic costs*:
  - ***$1.5 million****/fatal injury;* ***$80,700****/non-fatal disabling injury;* ***$9300****/property damage collisions*

- Can we predict the traffic crashes? ------- *Most probably not, unfortunately…*
  - *How about to some extent, for example crash risk, so that the stakeholders can better allocate resources or take preventive measures?*

* https://www.sawayalaw.com/blog/breaking-cost-accident/

Datasets



Road network information (~20000)

Road length
Road width
No. of lanes
…

311 service requests (~120,000)

Street light out
Potholes
Tree trims
Tree debris

Latitude, longitude

Crash reports (~90,000)

Time
Date
Weather
Driver info
Vehicle info
Fatality
…

Methodology

How can we represent the risk of crash associated with each road segment?

Probability generated from a classifier may serve this purpose

- Positive samples:

Latitude	Longitude	Road width	Street lights out	Hour of day	…

- Negative samples:
  - *Negative samples are of the same structure and format*
  - *Negative samples are "synthetic" by identifying the road segments without crash events and expand them across time (24h x 7days)*
  - *There are way more negative samples than positive samples, so the two classes needs to balanced. This can be done by randomly sampling the negative samples with sample size equal to the positive samples*

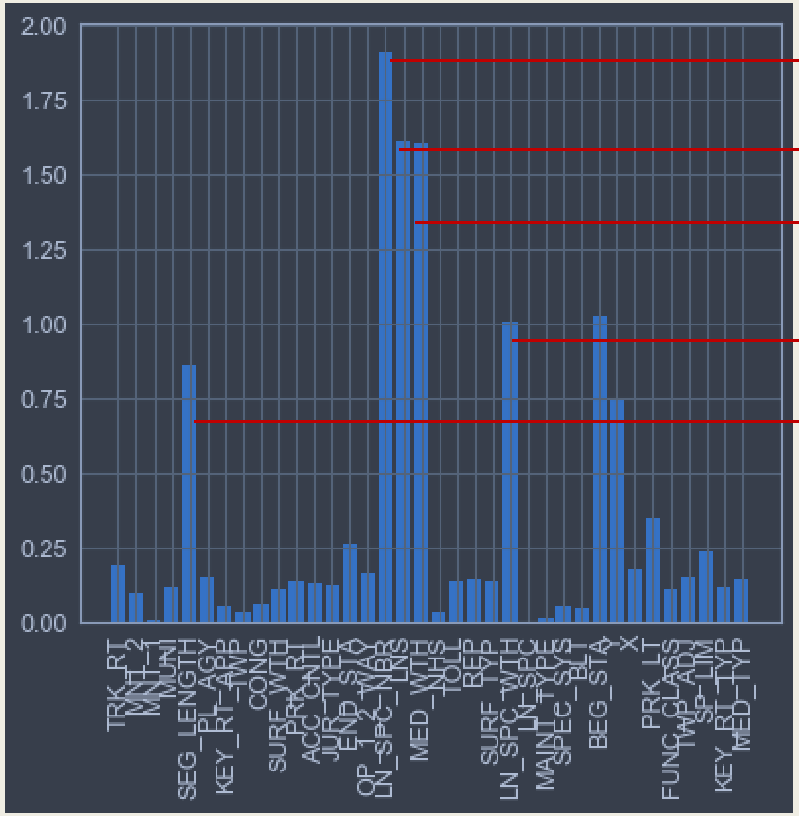
Methodology (cont’d)

- Modeling of traffic volume represented as annual averaged daily traffic (AADT)
  - *Predictors: geographical reference, road network information*
  - *Outcome: predicted AADT values associated with road segment where such measurements are not available*
- Modeling of crash risk
  - *Predictors: time, day, road network information, AADT*
  - *Outcome: predicted probability of crash using a classifier*
- Visualization

Modeling: AADT

- Data quality check and preprocessing
- Feature selection:
  - *Drop features that did not pass the quality checks (e.g. to many missing records, text etc.)*
  - *Mutual information*

Mutual information



Number of special lanes

Total number of lanes

Median width

Width of special lanes

Road segment length

The mutual information between two random variables measures the dependency. The function as implemented by ***sklearn.feature_selection.mutual_info_regression*** relies on nonparametric methods based on entropy estimation from k-nearest neighbors distances*.

https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html

Modeling: AADT

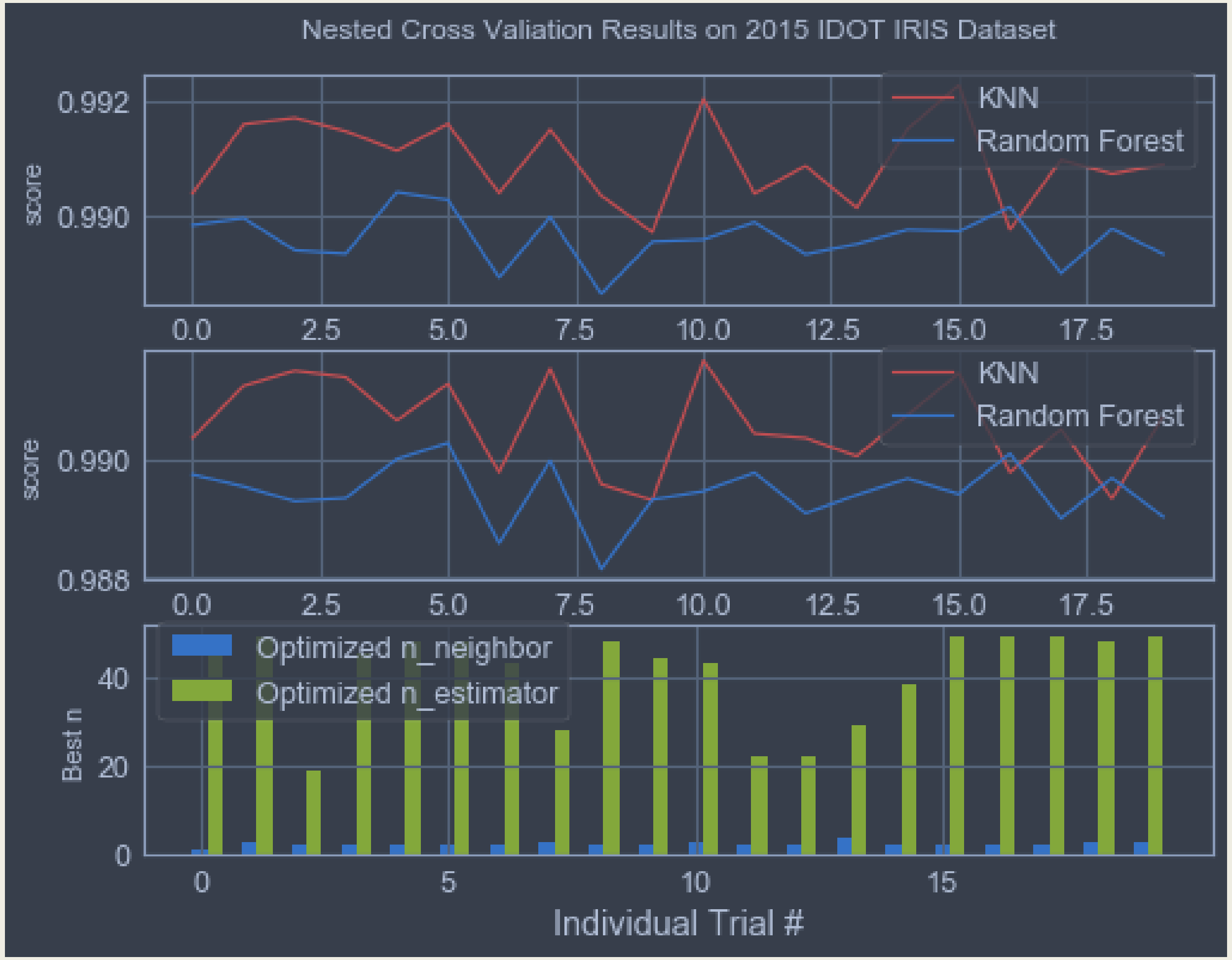
- Data quality check and preprocessing
- Feature selection
  - *Drop features that did not pass the quality checks (e.g. to many missing records, text etc.)*
  - *Mutual information*
  - *Test the performance of a few candidate models with different variable combinations (4 predictors retained in the final model)*
- Model selection and hyperparameter tuning
  - *Nested cross validation*

Modeling: AADT

Performances of four candidate models with 10 different seeds for the split of train and test data

	KNN	Random forest	Gradient boosting	Ada boosting
R^2	0.985 ± 0.003	0.990 ± 0.002	0.981 ± 0.003	0.924 ± 0.013
Mean square log error	0.132 ± 0.003	0.120 ± 0.004	0.242 ± 0.009	1.21 ± 0.094

Modeling: AADT



Modeling: AADT

- Data quality check and preprocessing
- Feature selection
    - *Drop features that did not pass the quality checks (e.g. to many missing records, text etc.)*
    - *Mutual information*
    - *Test the performance of a few candidate models with different variable combinations (4 predictors retained in the final model)*
- Model selection and hyperparameter tuning
    - *Nested cross validation*
    - <u>*KNN with n = 2*</u>

Modeling: crash

- Random forest model was selected due to its interpretability
- Feature selection was done based on the ones selected from the AADT model and evaluating the new variables such as 311 report data and crash day/time
- Training set was created by combining positive samples and negative sample and balancing the two classes
- The performance was evaluated by the classification accuracy (95%)

Final results and visualization

- The final results are visualized by Tableau
  - *Interesting insights*
  - *Dashboard application*

Caveats and future work

- Random forest has been reported to generate probabilities that are biased. Other methods such logistic regression with lasso may be potentially helpful.
- Think about other metrics or modification of current metric to better represent the idea of crash risks
- Weather condition reported by the crash reports may result in bias estimation, and should be reconsidered and maybe removed