Problem:

Predicting the occurrence of injury or death as the result of a road traffic collision in NYC. A model could provide live predictions to guide emergency response services.

Evaluation Metric: AUC

We used AUC as our evaluation metric for hyper-parameter and model selection. A false negative is always much more costly than a false positive. However, the desired sensitivity vs. sensitivity may change as the supply & demand of ambulances varies. AUC is base rate invariant and can capture this tradeoff.

Feature engineering:

Models:

Baseline model: Average injury rate for lat/long bins
Linear models: Logistic regression, SVM, Naive Bayes

Nonlinear models: Random Forest, XGBoost, Neural Networks

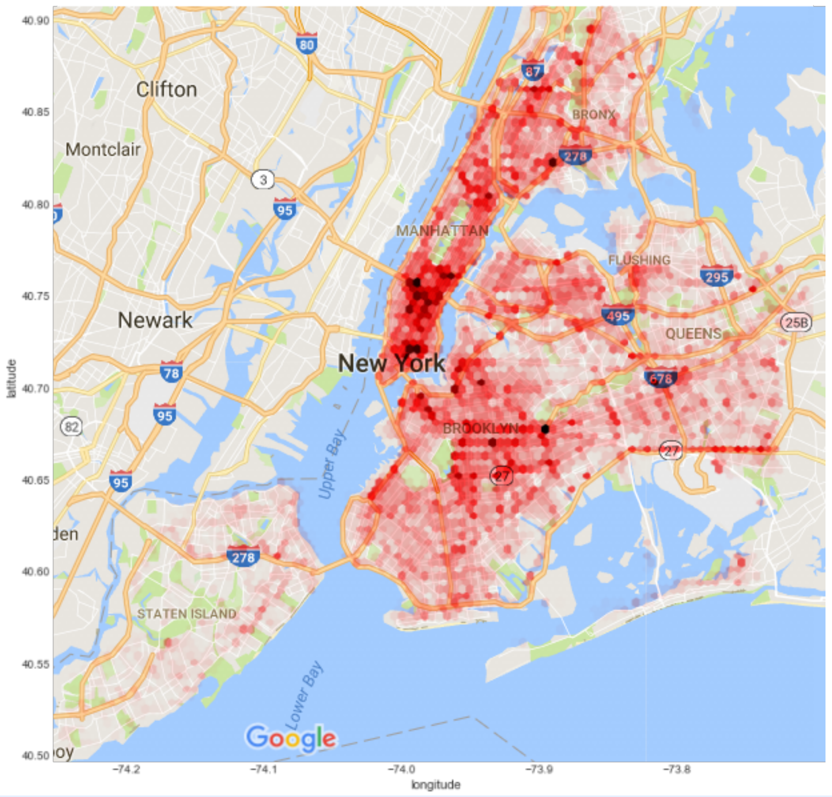
Insights:

- Number and type of vehicles was the most influential feature across the board.
- XGBoost found time & location features very influential, Random Forest preferred details about vehicle types.
- Temperature and humidity features also informative. Suggestion is that they are a proxy for number of people on the street. Distance to nearest subway behaves similarly.
- Other weather features (Snow, Rain, Fog) were not influential, possibly because drivers adjust their driving style in these conditions.
- Bicycle only collisions exhibit different influential features - driving style seems to play a larger part with speeding, stop-light violations and drag-racing reports being influential.

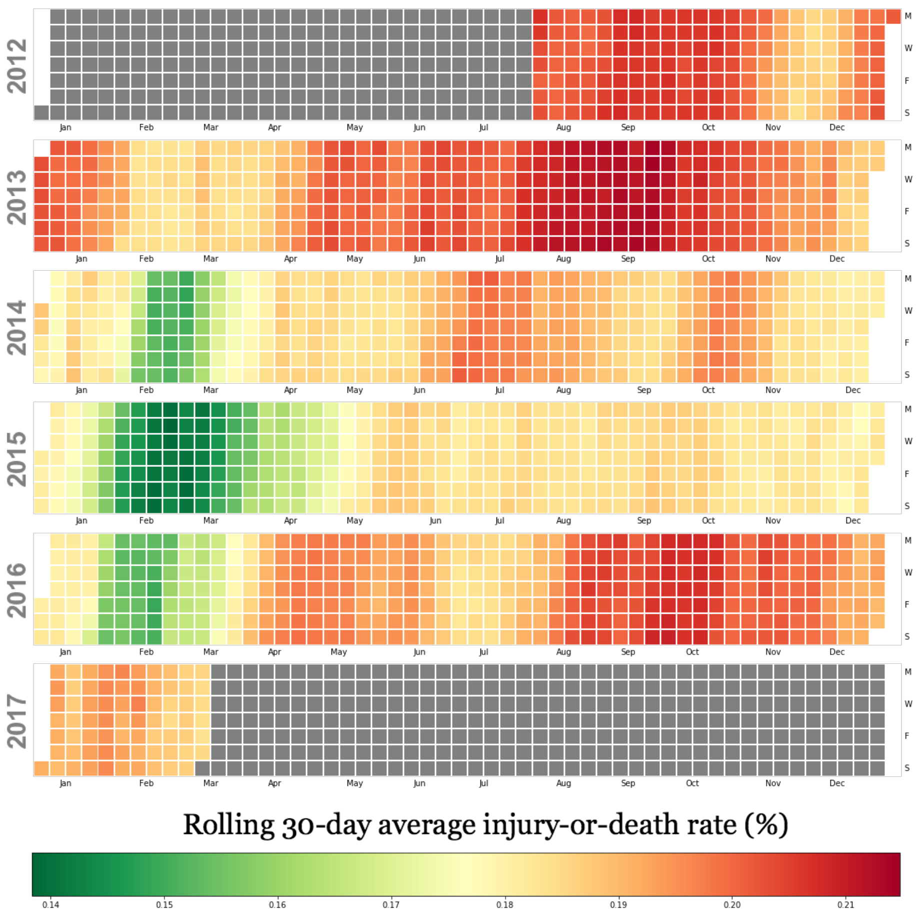
Data:

Data on 998,266 collisions in NYC from 07/01/2012 to 03/11/2017 from NYC OpenData. We joined with external datasets for additional features.

Geospatial trends: Injury/death occurrences



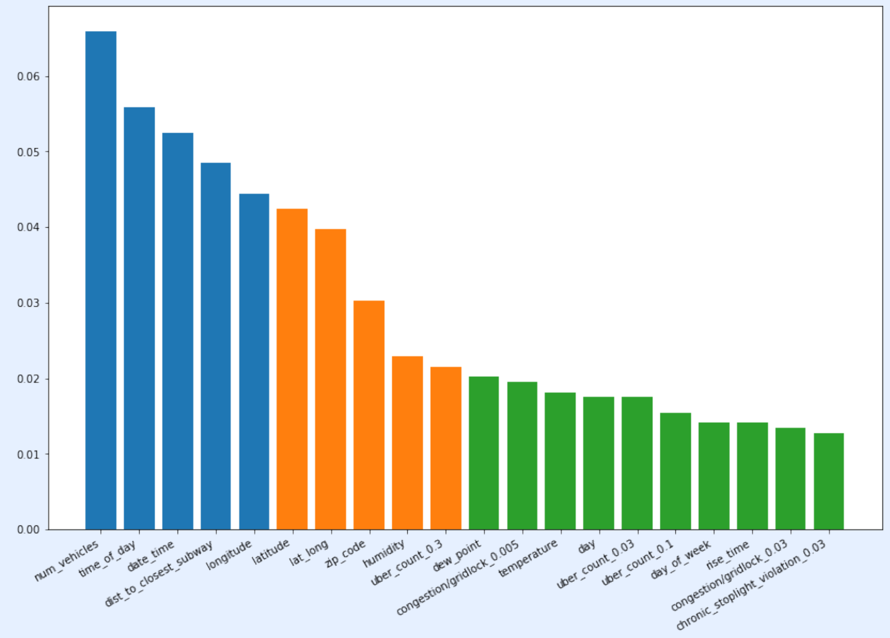
Temporal trends: Injury/death rates



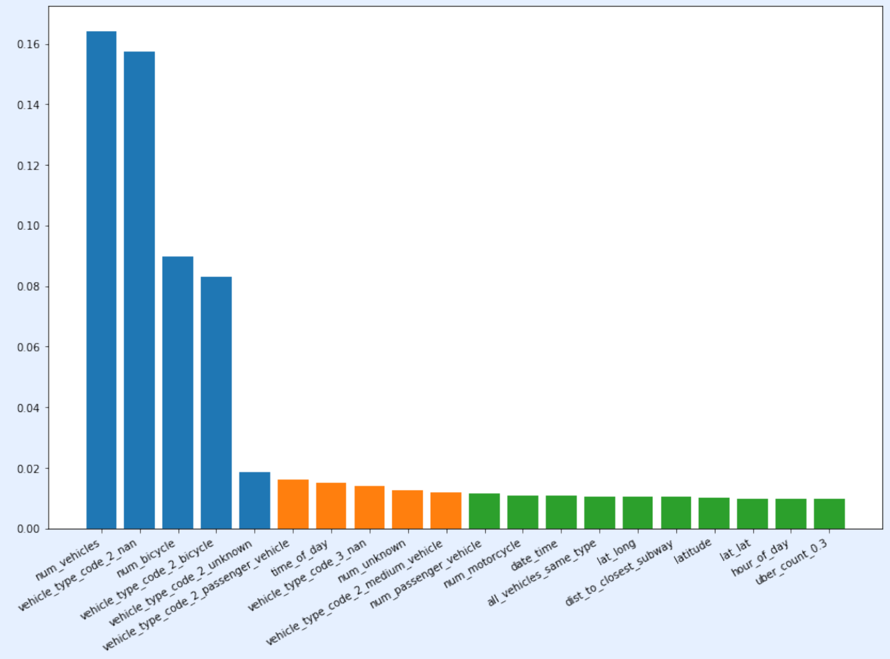
Rolling 30-day average injury-or-death rate (%)

Feature importances:

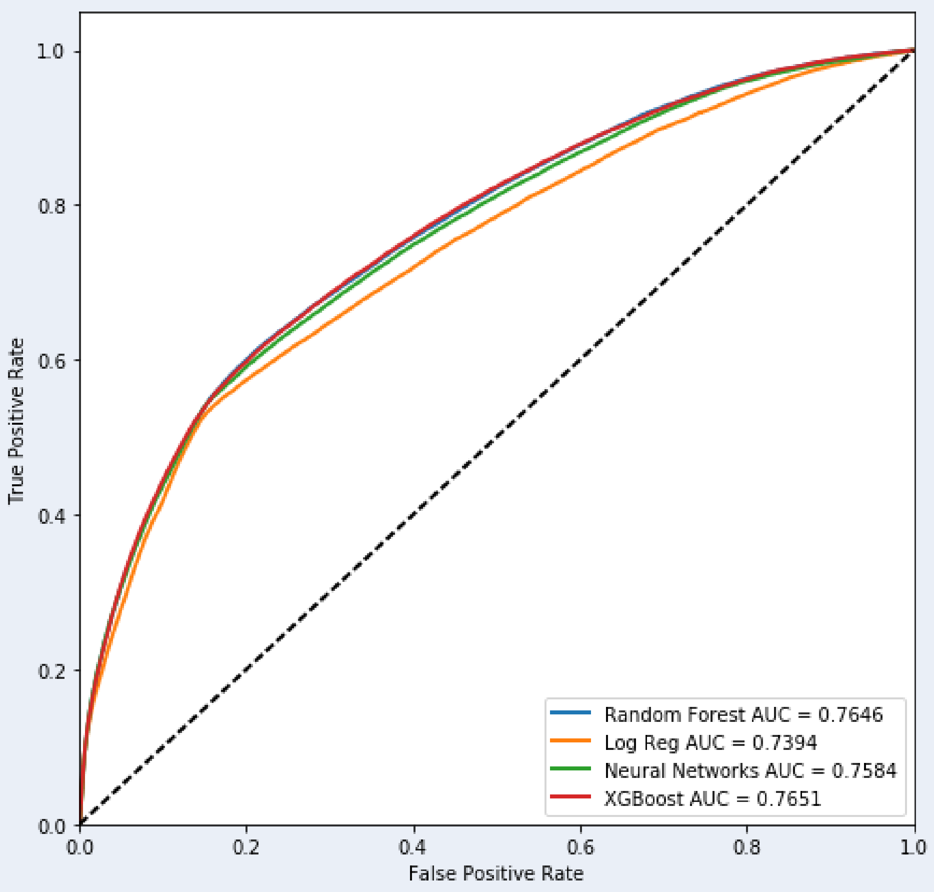
XGBoost



Random Forest



ROC Curve



Final model: XGBoost

estimators: 1000
max depth: 4
learning rate: 0.01
regularization: 100