# LOOK BEFORE YOU LEAP

## LEVERAGING PREDICTIVE MODELS TO IMPROVE AUTOMOTIVE SAFETY AND TRAVEL TIME

**Authors**: Hannah Do, Kamil Sachryn

[CSCI 35300/79502] Fall 2020

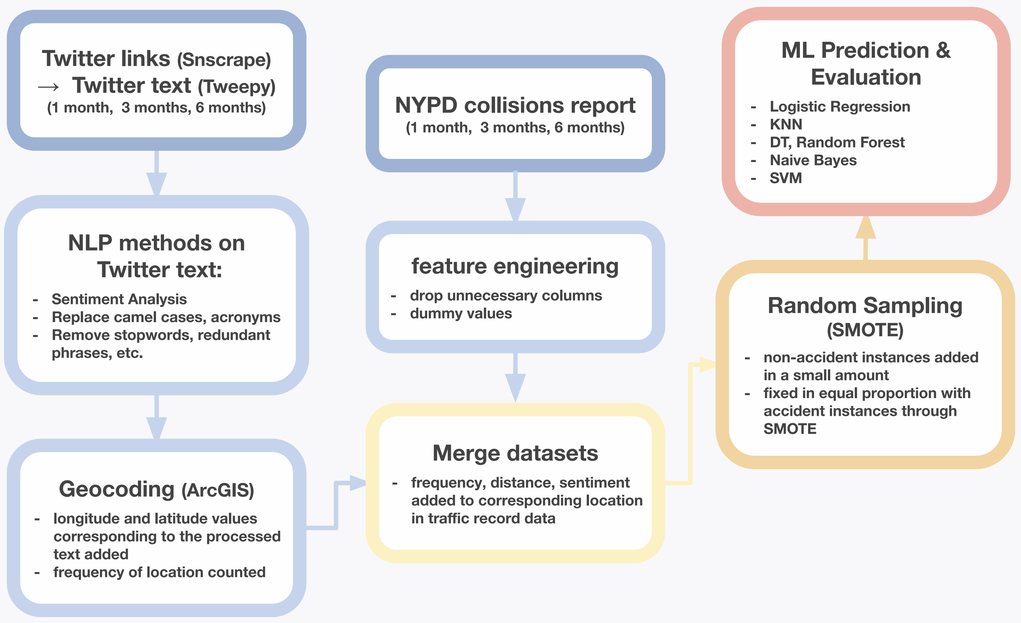
## OBJECTIVE

Recent growing amount of vehicles and traffic has been slowing down intensive road users, which costs over 10% of the logistic drivers' working hours. Human-based reports have high labor cost with significant delays, and current automatic incident detection algorithms have difficulties in reaching high accuracy due to unexpected circumstances and expensive costs. Our objective is to find an optimal traffic predicting algorithm using real-time and comparatively inexpensive Twitter data to avoid congestion and optimize time cost for many people in daily lives.

## PREVIOUS WORKS

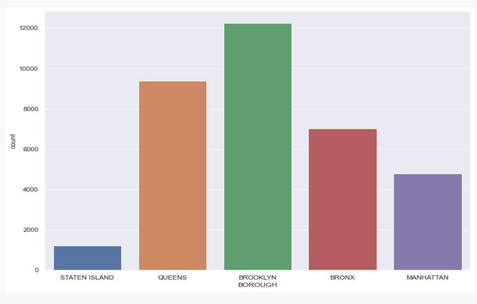
Antoine Hébert et al. (2019) predicted risk of accidents with high spatiotemporal resolution over a large area and time, allowing identification of significant risk factors. The method was to dynamically combine three large datasets, through feature engineering for imbalanced classes, and implementing BRF in Spark. Meanwhile Gu et al. (2016) added real-time Twitter data into the dataset, measuring discrepancy between different types of Twitter accounts and retrieving possible risk locations through NLP methods. There were many informative state-of-the-art models on accident risk prediction, however, we have focused on the methods mentioned above - language processing of twitter data and incorporation to traffic accident record through feature engineering and random sampling of prediction.

## METHOD



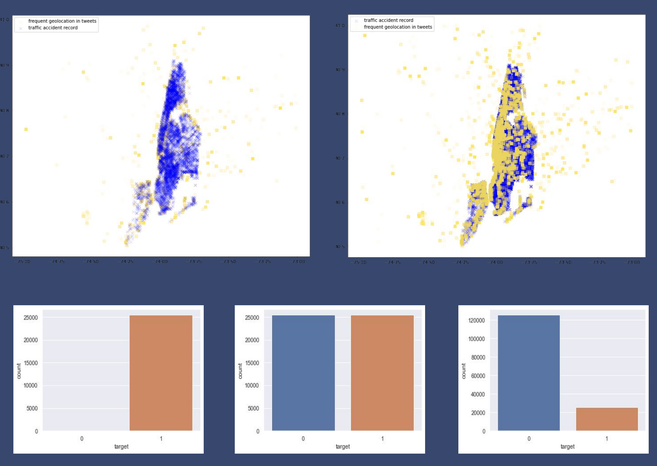


*Scatterplot of traffic accident record (6 months)*




*Distribution of data in different boroughs (6 months)*

## RESULTS & EVALUATION

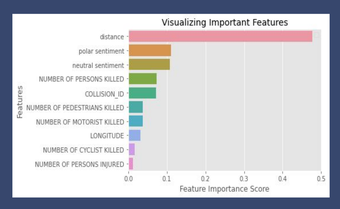



*Addition of negative instances through SMOTE and random generation*

***Scatterplot of accident record and twitter geolocations***

The locations from the geocoded tweets (yellow) and accident record (blue) were compared on a scatter plot. The twitter geocoded locations in yellow returned a wider distribution as the 511 tweets covered new york state, instead of new york city, and showed higher density near the intersection between the city and Queens, or upper area of Brooklyn. The data distribution in Staten Island returned a similar result, however the traffic record shows higher concentration near Bronx and Brooklyn.

| | 3 months | | | 6 months | | |
|---|---|---|---|---|---|---|
| | no twitter 1:1 | smote 1:1 | negative boost 4:1 | no twitter 1:1 | smote 1:1 | negative boost 4:1 |
| *precision* | 0.502156 | 0.933666 | 0.751797 | 0.676999 | 0.947914 | 0.787041 |
| *recall* | 1.000000 | 0.986144 | 0.852816 | 0.653652 | 0.979288 | 0.885943 |
| *f1-score* | 0.668580 | 0.959188 | 0.799127 | 0.665121 | 0.963346 | 0.833569 |



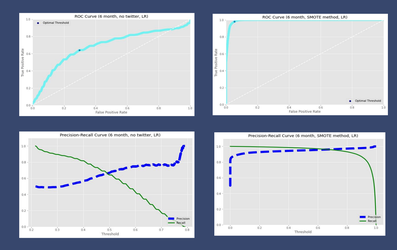

*Feature importance scores retrieved through Random Forest classifier*

***Accuracy of ML prediction on different datasets***

**negative boost : negative instances augmented with random generation*

The best performance resulted in Logistic Regression model with SMOTE method on 6 months data. The figures on the right represent precision-recall curve to verify the balanced trade-offs, and it shows the difference before and after the addition of twitter features.

Feature importance score was also used to confirm the important features as the additional twitter features, verifying that the new features have increased accuracy of the overall performance.




*ROC curve and Precision-Recall curve*
*Accident record vs. SMOTE ( 3 months)*

## BARRIERS

There were several bottlenecks that slowed down the project :

1. Twitter API had a rate limit in retrieving texts - used 3 different Twitter Developer Accounts to work around
2. Missing data - the initial dataset we chose 'The US Accident Dataset' had a huge amount Queens data missing - we switched to NYPD Open Data although it had less feature sets.
3. Geocoding each twitter text required a significant amount of time even with small data - made it difficult to achieve the 6-months data collection

## MOVING FORWARD

To continue research on this project, we hope to have an unlimited access to Twitter data, along with further text processing of the tweets. Clear understanding and sentiment-analysis of the tweets would be required for smooth geocoding. And adding additional datasets such as NYC bus stop locations, weather condition or road infrastructure would be another alternative to increase accuracy of the algorithm, along with fine tuning of the ML models.

## ACKNOWLEDGEMENT

This project was supervised by Professor Anita Raja, instructor of Machine Learning (CSCI 79502) at Hunter College.
The project title 'Look Before You Leap' - provided by Isaac Lapides.