Transportation & Urban Planning:
Relations among NYC Traffic Accidents, Crash Conditions, and Weather

Introduction

In a modern era, it is undeniable that a city's transportation system is an essential element of human civilization, improving human living standards and contributing to the economic growth of a city. What comes with an ever-growing reliance on urban transportation is an increase in road accidents. As such, road accidents have been actively studied because of the growing demand for safety conditions and prevention strategies in the hopes of reducing the impact of car accidents on road-users and vehicles. This study aims to develop a model that predicts harmful accidents in NYC, the most congested cities in the United States. Harmful accidents are defined as accidents resulting in at least one injury, one death, or both, under the circumstances of certain time, location, weather and collision conditions. Specifically, the research question is: **What time, location, weather, and collision conditions predict harmful accidents that result in casualty in NYC?** The implication of this project is two-fold: 1) detecting patterns of circumstances that lead to harmful accidents and 2) given the limited resources of the NYPD and NYC medical system, especially in those days of terrible traffic, predictions of harmful accidents allow urgent and efficient police patrol and ambulance dispatch for most needed road-users.

Current literature has identified multiple factors that are related to road accidents. These factors include road conditions, driver conditions such as drunken driving, driver behavior, age, gender, etc., weather conditions like precipitation, rain, snow, fog, etcs., and road accidents rates in different locations [1-4]. Research in this field has adopted a range of statistical analysis, like Bayesian model, Poisson model, and regression analysis [5-7]. For those conducted specifically in the US context, one research [8] investigates the connection between precipitation and the number of road accidents from 1975 to 2000 and show a negative relationship between precipitation and severe crashes in the US, namely that the risk of a road accident is increased with increase in the time since the last precipitation. Another study [9] that focuses on freeways in Southern California suggests a strong relation of driving speed, wet roads, and high traffic volume to traffic collisions.

Despite the rigorous attention being paid to this field of study, to our knowledge, no prior research has been conducted in the context of New York City, one of the busiest cities in the world. For anyone who has experienced the bustle and hustle of NYC, one thing they couldn't forget is its traffic. According to INRIX transportation analysis, New York City had the worst traffic congestion in the US in 2020. This "traffic hell" inevitably results in a considerable amount of car accidents. The NYPD records demonstrate that even in 2020, the year of a raging pandemic, there were still an average of 303 car accidents on the city's streets every single day. As such, our project aims to expand the existing literature on road accidents by systematically investigating reported car accidents in NYC from March, 2019, to March, 2021. Additionally, by including not only weather conditions and driver conditions but accidents conditions (e.g., number of cars involved, type of crashed vehicles, contributors other than the driver), our study has a potential to build a model targeting NYC traffic conditions. The last contribution of this study is to adopt analysis techniques (i.e., logistic regression and decision tree) that differ from prior research so we can focus on the prediction of harmful accidents and assign timely resources to involved individuals and cars when accidents occur in the future.

Methods

Data collection and description

To answer our research question, we draw data from NYC Open Data (the NYC Police Department records on traffic accident) and use a substantial dataset of motor vehicle collisions[1] from March, 2019, to March, 2021. In this dataset, each record represents an individual collision, with detailed information on time and location of the accident, vehicles and victims involved, and contributing factors. This study merges the NYC Traffic Accidents data with another dataset—Speed Limits of NYC Street[2]—to gain more information about the traffic circumstances at the accidental moments. Another primary data source is the Weather dataset[3] that contains historically hourly weather data on specific date and time. National Centers for Environmental Information (NOAA) provides local climatological data[4] of approximately 1,000 locations within the US . This study selects weather data from 5 weather stations (JFK International Airport, Linden Airport, Laguardia Airport, Newark Liberty International Airport, and NY City Central Park) in or close to NYC. The raw weather data is cleaned on an hourly basis including different weather conditions. Cleaned weather data is merged to the NYC Traffic Accidents data with speed limits information by location (latitude and longitude) and time (e.g., an accident occurred at 1:15pm is merged to weather data at 1pm). When merging the two datasets, we dropped those accidents that do not have location information. For further analysis, we dropped the data points that do not have location information (latitude and longitude).

Specifically, the variables used in our statistical modeling methodology include number of cars involved in the collision, borough and time of accidents, speed limit, contributing factors of accidents, car types, temperature, humidity, precipitation, visibility, daylight (from the time of sunrise to sunset, varied by day), rain, and snow. To measure harmful accidents, we combine the injury and deaths information from the NYC Traffic Accidents as *casualty*. An accident that results in at least one injury, death, or both count as a harmful case (assigning 1 as "True" for the outcome variable *casualty*), whereas accidents without any injury are deemed as harmless (assigning 0 as "False" for the outcome variable *casualty*). The accident condition variables definitions are shown in Table 1.

Table 1
Accident condition predictors in the model

Number of cars involved in the collision (numerical variable)	Number of cars is calculated by summing up the given information on different types of cars from the raw data (e.g., type information for two cars is converted to 2 cars involved in the collision)
Borough (categorical variable)	There are five boroughs in NYC: Bronx, Brooklyn, Manhattan, Queens, Staten Island
Time (categorical variable)	Time is included in the model 8 categories with an interval of 3 hours
Speed limit (numerical variable)	The speed limit of a given NYC street
Contributing factors (categorical variable)	Over 50 different contributing factors in the raw dataset are categorized into 6 categories: Affected: accidents occurred due to external factors

[1] https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
[2] https://data.cityofnewyork.us/Transportation/VZV_Speed-Limits/7n5j-865y
[3] https://www.ncei.noaa.gov/maps/lcd/
[4] https://www.ncei.noaa.gov/maps/lcd/

	other than the driver or the crashed vehicle (e.g., pedestrians, reaction to uninvolved cars) Dangerous: accidents occurred due to dangerous driving that breaks traffic law Improper: accidents occurred due to improper driving but doesn't break any laws Equipment: accidents occurred due to dysfunction of the crashed car Inability: accidents occurred due to dysfunction of the driver of the crashed car Unspecified: unspecified reasons and NaN
Car type (categorical variable)	Over 600 of car types are categorized into 5 types: Small: small cars such as three-door, sedan, and taxi Medium: medium cars such as van, bus, school bus Large: large cars such as truck, trailer, flat Single: motor, bike, skate, pedestrian Unspecified: unspecified car types or unintelligible car types such as E450 and PC
**In addition to the above accident condition predictors, the statistical model involved 7 Weather data variables listed in the data description, with four of them are numerical (temperature, humidity, visibility, and precipitation) and three of them are categorical (daylight, rain, and snow)* **Note that numerical variables are standardized for further analysis*	

Methodology

Four major different statistical models are compared in order to determine the accuracy of predicting harmful cases in NYC from 2019-2021. Model accuracy is assessed by four metrics: accuracy score, f1 score, prediction-recall, and FNR (false negative rate). Note that we decided to use FNR as our metric for final model selection because we believe in the context of car accidents and given the real-world application of this model is to dispatch police and medical resources efficiently to accidents that most needed, it is essential to have low FNR—that is, lower chance of predicting an accident as harmless when in reality that this certain accident is harmful. Our statistical models are shown below:

1) Logistic regression models with no penalty:
   1.1) Simple logistic regression model using all predictors;
   1.2) Baseline logistic regression model using selected predictors that are identified from examining the marginal relationship between predictors and casualty;
   1.3) Baseline logistic regression model using selected predictors with interaction;
2) Lasso-like logistic regression models:
   2.1) Logistic regression models with L1 penalty
   2.2) Logistic regression models with L1 penalty using selected predictors with interaction;
3) Decision tree model
4) Random forest model

Additionally, for the logistic regression models (with or without penalty), we performed further analysis of the improving model with predictions above the selected threshold in order to decrease FNR. Lastly, after running the random forest model, we identified 12 top predictors and re-ran them as selected predictors in another logistic regression model for comparison purposes. Model performances are discussed in the Results section.

Results

We first present the descriptive statistics of variables in the model followed by casualty prediction and model selection results. The accuracy of our selective algorithms is discussed, together with the output of the classification mode.

Preliminary and descriptive analyses

Fig. 1 presents the distribution of recorded accidents from March 2019 to March 2021 across different categories for each categorical variable. Within this two-year range, the cleaned dataset involves a total of 280,787 car accidents. Brooklyn had the largest share of NYC's traffic accidents (88502, 31.52%), followed by Queens (29.85%, 838320), Manhattan (17.83%, 50061), Bronx (48560, 17.29%), and Staten Island (3.51%, 9844). As for the time of accidents, the results show that afternoons from 3pm to 6pm (20.24%, 56828), 12pm to 3pm (18.22%, 51169), and evening from 6pm to 9pm (15.16%, 42555) are the top three time-ranges when accidents were mostly to occur. Morning times from 9am to 12am (14.88%, 41792) 6am to 9am (10.7%, 30034), and late evening from 9pm to midnight (9.92%, 27850) were the following common occurrences. As for contributing factors of accidents, dysfunction of the driver (30.08%, 84470), improper driving (24.52%, 71001), and dangerous driving (12.64%, 68858), are the topic three contributors other than those accidents where contributing factors are unspecified. Occasionally, accidents occurred due to external factors other than the driver such as slippery pavement or animals' actions (6.76%, 18970) and it is extremely rare that an accident occurred due to dysfunction of the vehicle (0.72%, 2010). Over a half of the accidents involved a small-size car (52.28%, 146791) and one-third of the accidents involved road-users other than a motor vehicle (31.08%, 87279), such as pedestrians and bikes. Large vehicles like trucks (10.79%, 302929) and medium-size vehicles like vans (5.03%, 14128) had relatively small share of the accidents. The three categorical weather data reveals that accidents occurred more frequently during daylight time (64.69%, 181093) than night time (35.51%, 99694), during non-rainy day (90.86%, 255121) than snow day (9.14%, 25666), and during non-snow day (98.89%, 277677) than snow day (1.11%, 3110). Note that the imbalance results of the categorical weather variables are due to the natural imbalance distribution of daylight time vs. nightlight time, rainy day vs. non-rainy day, and snow day vs. non-snow day across time in the first place.

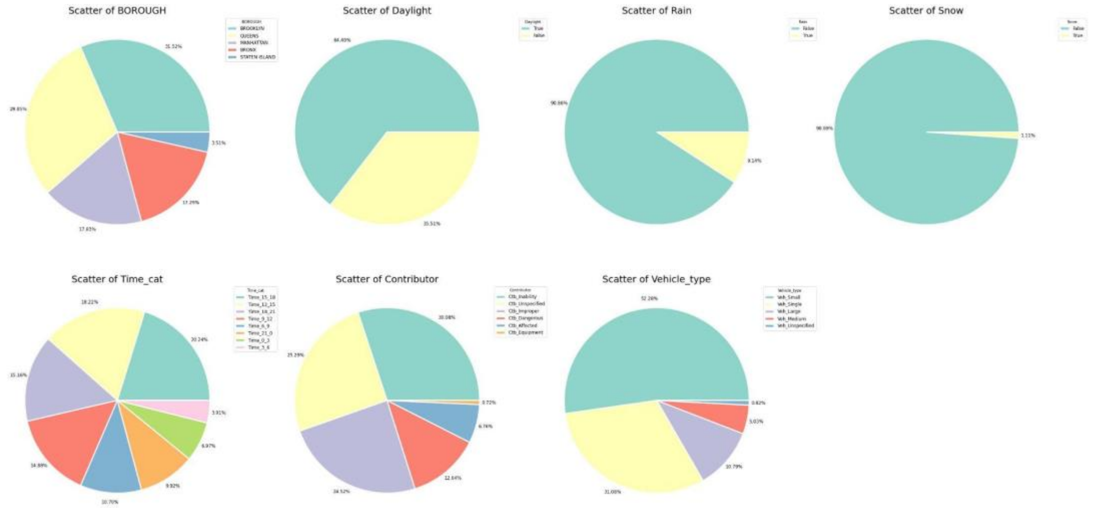


Fig. 1 Distribution of accidents across categorical variables

Descriptive statistics of numerical variables involved in the present study are shown in Fig. 2. As for the three weather variables, we learned that the distribution of *temperature* is bimodal with two peaks around 40 degrees Fahrenheit and 80 degrees Fahrenheit. The distribution of *humidity* is a little bizarre in that many accidents occur at the extreme end of humidity (when relative humidity falls into the 90% to 100% range) but when humidity is lower than 80%, the distribution is fairly normal. *Precipitation* and *visibility* don't seem to be strong predictors of harmful accidents, as the distribution is either extremely right skewed (*precipitation*) or left skewed (*visibility*). From the predictor *collision_count*, number of cars involved in an accident, we learned that the majority of the car accidents (67.1%, 188508) involved two cars, followed by a single car being involved (24.6%, 69189), and only occasionally, an accident involved more than 3 cars (7.4%, 20805). An examination of the distribution of *speed_limit*, the speed limit of the street where an accident occurred, reveals that 85% of the accidents happened at a location that has a speed limit of 25mph and only 1% (2799) of the accidents occurred at a location that has the highest speed limit in our record (45mph).

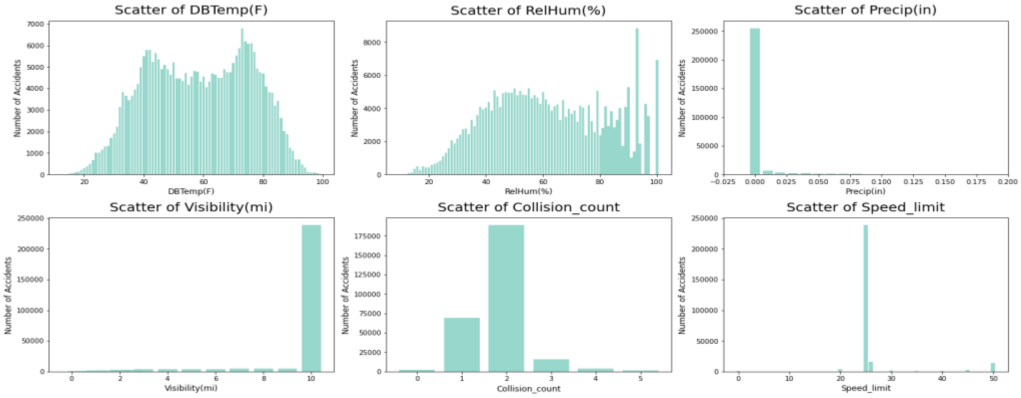


Fig. 2 Distribution of accidents across numerical variables

Fig. 3 shows the distribution of traffic accidents in NYC by severity, with the left plot demonstrating harmless accidents without any injury and the right plot demonstrating harmful accidents with at least one injury, one death, or both. A comparison of density of these two scatterplots reveals that the majority of car accidents did not lead to harmful consequence as measured by injury or death (75%, 210736), but still 25% (70051) of the accidents resulted in injury, which costed significant lost on not only individual, but economical and societal levels.

To gain more insights of the relationship between the model's predictors and outcome variables (non-injury vs. casualty), Fig. 4 demonstrates the marginal association of each of our 13 predictors with whether or not an accident is harmful. Specifically, the marginal association

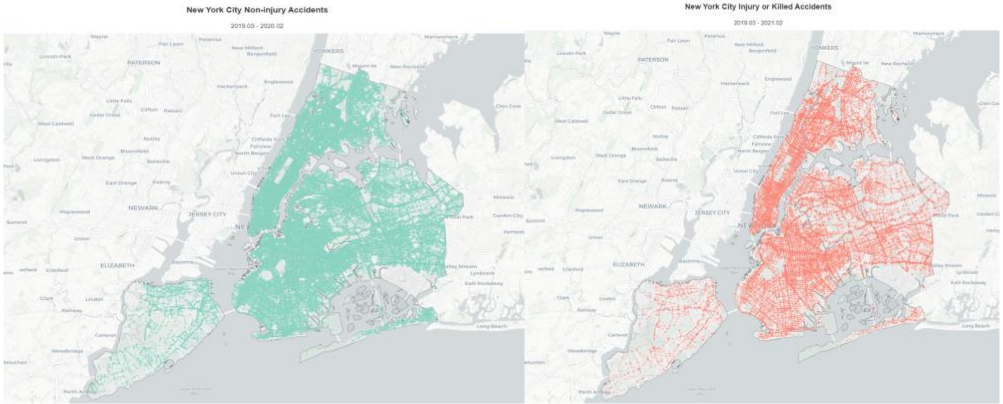


Fig. 3 NYC maps of non-injured accidents vs. injured or fatal accidents

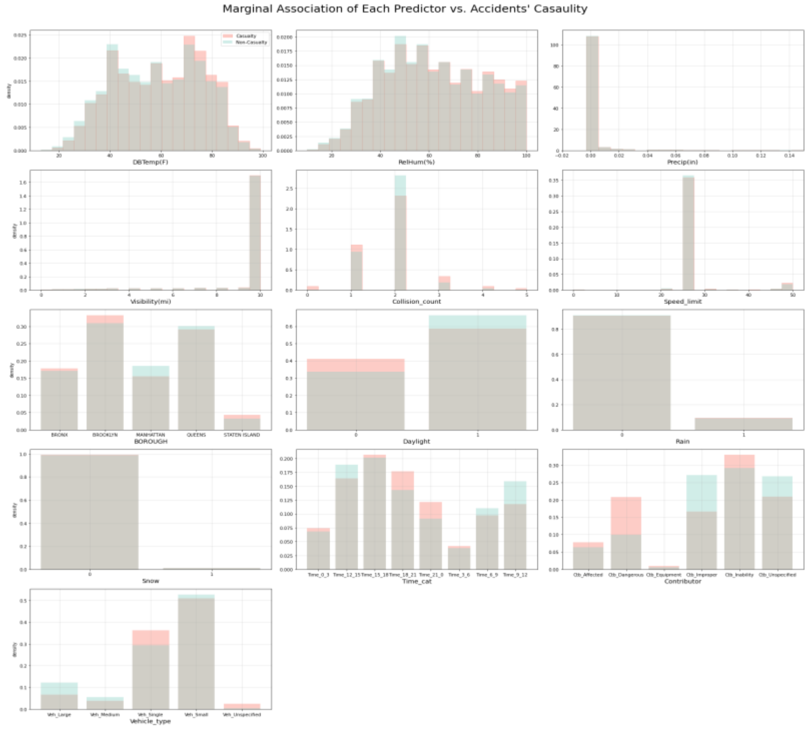


Fig. 4 Marginal association of each predictor vs. whether or not an accident is harmful

provides us with information about the associations between various values of the variables and the outcome variables (harmless vs. harmful accidents). By plotting distribution densities of "casualty" overlaid with "non-casualty" for each predictor, we can begin to see which predictors have the large shift in density among harmful vs. harmless outcomes. For instance, the notable shifts for *collison_count* observations, *borough* observations, *time_category* observations, *contributor* observations, and *vehicle_type* observations, signal how well those accident conditions variables might work for predicting outcomes. Moreover, when temperature is below 60 degrees of Fahrenheit, the density of harmless accidents is higher than that of harmful accidents, but the pattern reverses for temperature above 60 degrees of Fahrenheit. A major shift in density is also detected in *daylight* observations, with night times having more density of harmful events than day times.

Prediction model performance

We started with a simple logistic regression model with all predictors (**M1**), but the FNR is about 0.4654. Therefore, we made our first attempt to improve the model by selecting predictors that are deemed as important (**M2**) by examining the marginal association (Fig. 4) between predictors and outcome. We selected those variables that have a large shift in density such as *Daylight*, *Collision_count*, etc. The output of this model[5] suggests that: 1) there is a negative relationship between daylight hours and casualty; 2) there is a positive relationship between number of cars involved in an accident and casualty; 3) comparing to the referent borough, Bronx, Brooklyn and Staten Island have relatively higher risk of harmful accidents, whereas Manhattan and Queens are less likely to have harmful accidents; 4) comparing to the referent time category 3am-6am, the time intervals 12pm-3pm, 3pm-6pm, 6pm-9pm, and 9pm-12pm have higher risks of harmful accidents, whereas the time intervals 0am-3am and 9am-12am are less risky; 5) comparing to unspecified contributors of car accidents, all other contributors except improper driving are likely to result in harmful accidents, with dangerous driving having the highest relative risk; and 6) comparing to unspecified car types, all four size categories have lower risk

[5] Please refer to the notebook for detailed coefficients under the 2.1 model labeled as baseline model

but an examination of the relative negative magnitude indicates that single vehicle types are more likely than the other categories to be involved in harmful accidents. This baseline selected-predictor model yields an FNR of 0.3913, declined significantly from the simple logistic regression model. Our next attempt to improve the baseline selected-predictor model is to set up a threshold that differs from the default threshold value of 0.5. After identifying the best threshold at a value of 0.76 and manually calculating the predicted results based on the new threshold, the FNR of the model declines to 0.2176.

Furthermore, given that the *dangerous driving of the contributor* observations and the *daylight* observers appear to have the most impact on the outcome variable, we performed a logistic regression model (**M3**) with the interaction term *dangerous*daylight*, yielding an FNR of 0.3843 that is slightly better than M2. The coefficient of this interaction term is 0.0261, suggesting that compared to the relationship between *daylight* and *casualty* when the contributor of accidents is unspecified, such relationship increases when the driver was in a dangerous driving condition. In other words, even during the daytime, dangerous driving still resulted in higher risk of harmful accidents. Changing the threshold to an ideal level of 0.7576 yields an FNR of 0.2176.

The second major statistical model that we built for predicting harmful accidents is fine-tuned Lasso-like logistic regression. A baseline Lasso model (**M4**) with all predictors yields a FNR at 0.3984, which performed worse than M3. With the selected threshold of 0.73, the model achieves a FNR of 0.208, outperforms all other models. The third major statistical model we attempted to build is a basic decision tree model (**M5**). Given the binary nature of our outcome variable, we believed it is worth trying another approach of classification. For building a basic decision tree model, we first identified a depth value of 7 as the best threshold, and set the max depth accordingly in our decision tree model. This first attempt of the decision tree model yielded a FNR of 0.3618, which is slightly better than M4 and better than M3. The fourth major statistical model we built is a random forest model (**M6**). This attempt yields a FNR of 0.1952, so far the best performing model with default threshold. The random forest model also generates 15 top predictors that appear to be impactful when predicting our outcome variable *casualty*. Therefore, we performed another Lasso-like logistic regression model with interaction terms across all of these predictors (**M7**), yet this model does not outperform the random forest model, yielding a FNR of 0.3822.

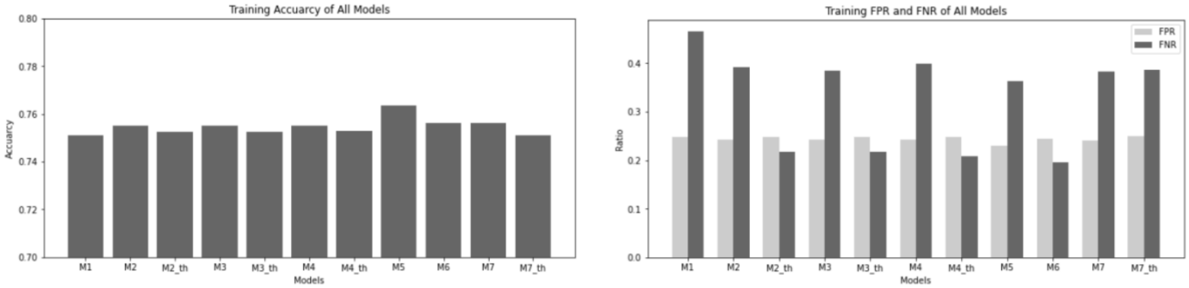


Fig. 5 Accuracy scores and FNR for all models with default threshold and adjusted threshold

Fig. 5 demonstrates model accuracy as measured by accuracy score and FNR across the attempted models (model labels ended up with *_th* showing output after adjusting the default threshold to the best threshold respectively). As we can see, the random forest model (M6) has the lowest FNR rate but its accuracy score is not the highest. *Fig. 6* shows the precision-recall curves for each model with default threshold and adjusted threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. An examination of the precision-recall curve plot also reveals the better performance of the random forest model (M6) as the area under the curve is relatively high. This suggests that the classifier is returning relatively accurate results (high precision), as well as returning a majority of all positive results (high recall) in a relative sense.

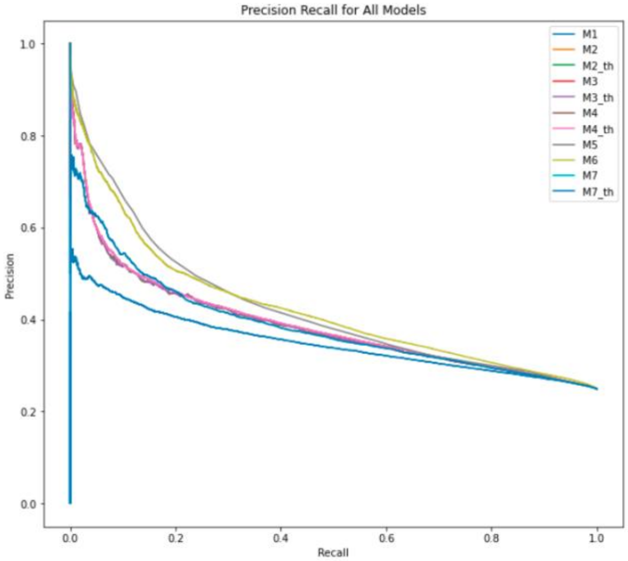


Fig. 6 Precision-recall curves for all models with default threshold and adjusted threshold

Conclusion and Limitations

By comparing the models in terms of FNR, we identified the random forest model (M6) with all predictors as our final model. M6 yields the lowest FNR even after we adjusted the threshold for the other models. However, M6 does not yield the highest accuracy score. This raises the question of lowering FNR at the expense of model accuracy. This is a trades-off that data scientists need to deal with constantly. In our case, a low FNR would suggest that the model has a low chance of predicting a real harmful accident to be negative but only accomplishing it with a relatively lower accuracy score. This could lead to a scenario of false positives, namely that predicting an accident is harmful but in reality, is harmless. A consequence of this in our context would be mistakenly sending police officers, ambulances, and medical workers to an accident site where no one got injured. On the other hand, our selected best model resulted in the lowest FNR among all attempted models. This is to avoid the scenario when there is a real harmful accident, but no one is assigned to help the victims. Balancing the cost (a potential of wasting resources) and benefit (not saving resources when there really are people involved in an accident who need help) of a low FNR but a relatively low accuracy score, we decided to choose the model yielding the lowest FNR but not the highest accuracy score.

To better understand the model predictions, *Fig. 7* displays the relative risk of harmful accidents by zip code of NYC (predictions generated from the test set) . As we can see from the map, our model predicts that harmful accidents are mostly likely to occur in Staten Island and least likely to occur in Manhattan. This could be the case because although Manhattan tends to have heavy traffic flow, the speed limit is low. As we learned from the output of our model, *speed limit* is positively related to *casualty*. This means that it is likely in Staten Island the speed limit is high as the traffic flow is not that heavy as compared to that in Manhattan or Queens. It is also worth noting that some specific areas in Brooklyn and the Bronx are at high risk of harmful accidents as well.

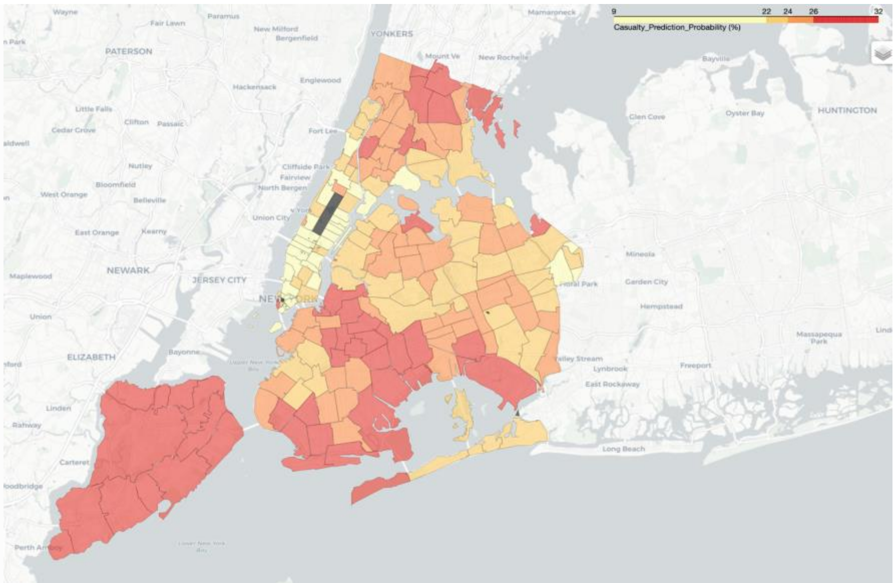
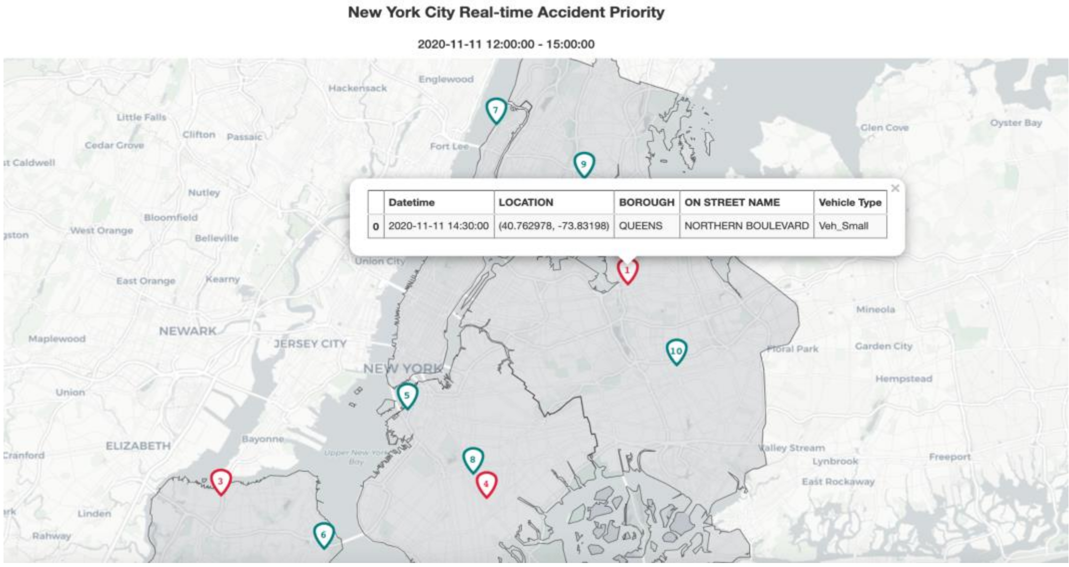


Fig. 7 Relative risk of harmful accidents by zip code in NYC

As discussed in the introduction, the ultimate goal of our model is for a scenario in which police offices and medical centers are receiving phone calls regarding car accidents occurring on the streets of NYC and need to dispatch resources to people who need them. To realize this goal, *Fig. 8* demonstrates a potential interface for police officers and medical workers. By gathering the information of all our predictors such as what types of cars are involved in the accident, the contributing factors, the speed limit of the street, and so on[6], this interface could automatically rank the reported accidents by the probability of a certain accident being harmful. Higher probability would be labeled in red and lower risk would be labeled in green. As such, relevant departments can dispatch resources efficiently, especially when the traffic is terrible and/or resources are limited, to the accident sites that need most help.


New York City Real-time Accident Priority

2020-11-11 12:00:00 - 15:00:00

	Datetime	LOCATION	BOROUGH	ON STREET NAME	Vehicle Type
0	2020-11-11 14:30:00	(40.762978, -73.83198)	QUEENS	NORTHERN BOULEVARD	Veh_Small

Fig. 8 A demo of NYC real-time car accidents ranked by the probability of being harmful

[6] Note that the figure only showing vehicle type as one example but in reality, this interface should include all of the model's predictors and the date/time are set for demonstration purposes

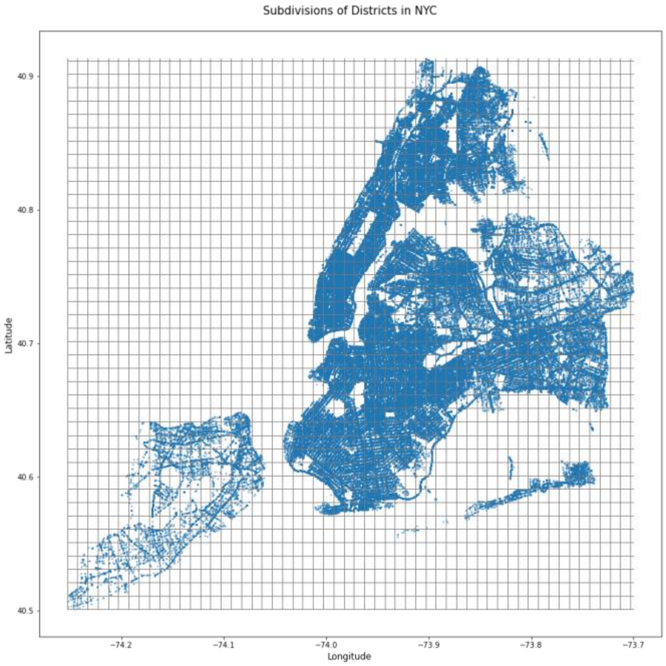


Fig. 9 A demo of NYC divisions of accidents using longitude and latitude information

Despite the rigorous effort being put into this project, there are a few limitations worth discussing. First, applications of this work can extend beyond dispatching police and medical resources efficiently. With more time, we can add other features into the model so predicting results can be used for other decision-making purposes like. For example, knowing that harmful accidents are likely to occur in certain regions of NYC, what prevention strategies (e.g., pedestrian pavements, speed limit, roadblocks, etc.) could be adopted. Also, with more information on in-state vs. out-state driver licenses or vehicle plates and their relationships with harmful accidents together with predictors included in our model, relevant departments can restrict road use during certain time of a day or weather conditions for certain cars given their license condition or original states of their vehicle plates. Second, with more time, we would measure relative risks by city zones, generated from dividing the cities by longitude and latitude information, as opposed to using zip codes. City regions divided by Zip code, as shown above, are not as fine-grained as zone areas defined by grids of the city area as shown in *Fig. 7*, and can be too large to pinpoint a specific area that is most likely to have an accident and that needs additional prevention strategies. Finally, we admit that the attempted models are by no means exhaustive and further actions are needed to improve model performance. For instance, although we have tried two decision tree models (basic and random forest), other approaches like AdaBoost, Bagging, and Boosting are worth trying as well if we'd have more time training the models. Additionally, our attempts to answer the research question based completely on classification, but it could also be answered by regression analysis. For instance, by aggregating harmful cases for each subdivided region and wrangling predictors accordingly, our research question can not only be answered by classification analysis (i.e., whether or not there will be harmful cases) but by regression analysis (i.e., how many more accidents are likely to occur in certain crash and weather conditions).

References

1. Ansari S, Akhdar F, Mandoorah M, Moutaery K. Causes and effects of road traffic accidents in Saudi Arabia. Public health. 2000; 114(1):37–39. https://doi.org/10.1038/sj.ph.1900610 PMID: 10787024
2. Odero W, Khayesi M, Heda P. Road traffic injuries in Kenya: magnitude, causes and status of interven- tion. Injury control and safety promotion. 2003; 10(1-2):53–61. https://doi.org/10.1076/icsp.10.1.53. 14103 PMID: 12772486
3. Vanlaar W, Yannis G. Perception of road accident causes. Accident Analysis & Prevention. 2006; 38(1):155–161. https://doi.org/10.1016/j.aap.2005.08.007
4. Shaw L, Sichel HS. Accident proneness: Research in the occurrence, causation, and prevention of road accidents. Elsevier; 2013.
5. Guo Y, Li Z, Wu Y, Xu C. Evaluating factors affecting electric bike users' registration of license plate in China using Bayesian approach. Transportation research part F: traffic psychology and behaviour. 2018; 59:212–221. https://doi.org/10.1016/j.trf.2018.09.008
6. Guo Y, Osama A, Sayed T. A cross-comparison of different techniques for modeling macro-level cyclist crashes. Accident Analysis & Prevention. 2018; 113:38–46. https://doi.org/10.1016/j.aap.2018.01.015
7. Delen D, Tomak L, Topuz K, Eryarsoy E. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. Journal of Transport & Health. 2017; 4:118– 131. https://doi.org/10.1016/j.jth.2017.01.009
8. Eisenberg D. The mixed effects of precipitation on traffic crashes. Accident analysis & prevention. 2004; 36(4):637–647. https://doi.org/10.1016/S0001-4575(03)00085-X
9. Golob TF, Recker WW. Relationships among urban freeway accidents, traffic flow, weather, and light- ing conditions. Journal of transportation engineering. 2003; 129(4):342–353. https://doi.org/10.1061/ (ASCE)0733-947X(2003)129:4(342)