Traffic Accident Severity Prediction in the City of New York

Milica Skipina
Faculty of Technical Sciences, University of Novi Sad
Novi Sad, Serbia
skipinamilica@gmail.com

Djordje Batic
Faculty of Technical Sciences, University of Novi Sad
Novi Sad, Serbia
djordjebatic@gmail.com

***Abstract*—Throughout the world, roads are shared by cars, buses, motorcycles and other motor vehicles, as well as pedestrians and other travelers. Efficient travel made possible by motor vehicles facilitates economic and social development and supports higher quality of life. Yet each year, vehicles are involved in crashes that are responsible for millions of deaths and injuries. In 2018 alone, New York City saw 228,047 traffic accidents, of which 287 were fatal. In this paper, we propose a solution that leverages open datasets of a city like New York City, in order to create reliable, explainable and efficient machine learning accident severity prediction models.**

We present a benchmark dataset, which was obtained by extracting accident, weather, road and traffic related information from various data sources as well as a preprocessing pipeline which accounts for severe class imbalance related to accident severity prediction problem.

Experimental results show that we can achieve a macro-average f1-score of at least 0.56 for the accident severity prediction task. In addition, we proposed an accident prediction solution which achieves a macro-average f1-score of 0.70. Finally, we used causal inference techniques like propensity score matching and inverse probability weighting to propose an action that would lead to 4% decrease in rate of serious crimes.

***Index Terms*—road accidents, machine learning, classification algorithms, ensemble learning, urban planning, causal inference**

I. Introduction

Road traffic crashes are a leading cause of death in the United States for people aged 1–54 [1]. It is estimated that fatal and nonfatal crash injuries will cost the world economy approximately $1.8 trillion dollars in the period between 2015 and 2030 [2]. Taking into account the human burden as well as the significant economic toll, it is important to consider possible ways that we can prevent or efficiently react to the road accidents. This can be done through education, safety provisions and better urban planning [3], but also through support of emergency ambulance services [4].

With the emergence of efficient data acquisition, storing and processing methods, as well as high-accuracy machine learning algorithms many researchers have tackled this problem, which will be thoroughly examined in the Section II.

In this study, we discuss the possibility of developing a traffic severity prediction models that can be implemented in real-time emergency ambulance services system inside NYC area. We present a benchmark dataset, which consists of data extracted from various Open Data sources and contains features related to accidents, weather, roads and traffic. Explainability and interpretability are important for engagement but also help avoid unintended consequences. We propose an inherently interpretable solution based on gradient boosting and decision tree based algorithms and discuss its performance and efficiency, as well as the key features that drive its decision. This research will offer real-time support for emergency ambulance services. By predicting the severity of the accident, we can assign more effective strategies for emergency response - thus improving the survival rate. Furthermore, it will promote a better understanding of the factors that influence the severity of accidents and help urban planners and city legislators develop effective prevention strategies.

A detailed review of existing solutions related to traffic accident prediction and their methodology, performance and limitations will be provided in Section II. Next, in Section III, we focus on data collection, processing and exploration techniques. Section IV is dedicated to the proposed system, while Section V deals with evaluation methods and results. Section VI is dedicated to discussion of the proposed solution and it's robustness. Finally, conclusions and the identified future work are presented in Section VII.

II. Literature Review

There has been significant effort in the machine learning community focused on predicting the road traffic accidents, spurred by the availability of Open Data datasets (ie. data gathered by governments, states, provinces and municipalities).

In [5], the need for collecting data from multiple sources is discussed. Authors used three public datasets provided by the city of Montreal and the government of Canada. First dataset, Montreal Vehicle Collisions, contained 134,489 records of the date, the hour and the location of the accident. Next, National Road Network dataset [6] was included. It contains the geometry of all roads in Canada, of which 44,111 road segments belonging to the island of Montreal were selected. Finally, Historical Climate Dataset [7] containing temperature, humidity measure, humidity percentage, wind direction, wind speed, visibility, atmospheric pressure, Hmdx index, and observation of atmospheric phenomena such as snow, fog, rain, etc. was used. After testing various models, authors report 85% detection of road vehicle collisions with a false positive rate of 13% using the Random Forest algorithm.

According to [8], novel Frequent Pattern tree (FP tree) based variable selection methods offer better accident risk prediction performance compared to random forest based methods regardless of the type of prediction models(i.e. k-nearest neighbor or Bayesian network). The method works by first identifying all the frequent patterns in the traffic accident dataset. Next, for each frequent pattern, a novel Relative Object Purity Ratio (ROPR) metric was introduced. The ROPR is used to calculate the importance score of each explanatory variable which in turn can be used for ranking and selecting the variables that contribute most to explaining the accident patterns. Authors achieved 61.11% accident prediction score with false alarm rate of 38.16% on the Virginia I-64 interstate highway.

In [9] authors were predicting traffic crash likelihood at different severity levels. They were using crash, weather and traffic data collected from loop detector stations. The researchers extracted traffic data in the time interval between 5 and 10 min prior to crash occurrence. To generate the dataset of non-crash cases, they randomly selected 20 five-minute intervals without crashes from the crash-free days. The time and loop station were randomly chosen for each non-crash case.

Authors in [10] considered a binary classification problem. They used several datasets including the motor vehicle crashes, detailed road network, and hourly weather data. They dealt with class imbalance and incorporate the spatial structure of the road network into the predictive model. Informative sampling approach was used for constructing negative samples. For each positive example they randomly changed the value of only one feature among Hour, Day and road ID and if it was not a positive example, they added it to the dataset. They evaluated four classification models, i.e., Support Vector Machine, Decision Tree, Random Forest, and Deep Neural Network (DNN). Random Forest and DNN generally perform better than other models.

III. Data Integration

We used 3 public datasets [11] - [13] provided by the city of New York, as well as information scraped from IBM Weather API [14]:

A. Motor Vehicle Collisions

NYC Open Data provides three different tables which beside data of crash events contain data on the people and vehicles involved in the crash. Due to many missing and invalid data in the Vehicles and Person datasets, we were using only crash events data. Dataset contains information from all police reported motor vehicle collisions in NYC from the July 1, 2012 and is updated daily. In order to be able to merge this data with [13] we used the crash records from September 13, 2014 to November 24, 2019 which is in total 1,164,268 crashes. Fig. 1 shows the map of New York City with recorded traffic accidents represented as red dots. Dataset contains basic information as date, time and location of crash as well as number of injured or killed persons which we had used as output for our models for predicting accident severity. As the total number of samples where someone has been killed was small, we labeled accident as non severe (label 0) if there was no injured or killed and severe (label 1) otherwise. As number of persons injured or killed represents target variable and location is the important feature for merging this dataset with others, we removed all the rows with the missing value for at least one of the these three features. After removing these rows, we have 1,026,365 rows left of which 224,426 rows had missing name of the street on which the collision occurred.

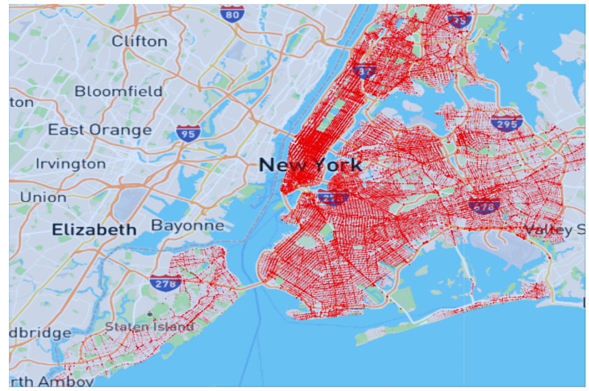

Fig. 1. Map of New York City with traffic accidents represented as red dots

B. IBM Weather API

To determine how much impact the weather has on a car accidents, we scraped weather data from [14]. Scraped data contains hourly information of temperature, measure of felt temperature, dew, humidity, wind, wind speed, wind gust, pressure, precipitation, condition, day/night and clouds type. Weather and crash data were merged by date and hour.

C. LION Single Line Street Base Map

Linear Integrated Ordered Network (LION) is a single line representation of New York City streets containing street related information such as street name, traffic direction, street width, speed limit in miles per hour, designation whether the segment is a part of New York City bicycle or truck route network, snow removal priority, number of travel lanes, number of parking lanes, total number of lanes in the roadway, indicator whether the segment is accessible for pedestrians or not, spatial coordinates of the beginning and the end of the segment.

We merged this dataset with [11] by matching each crash event with the closest road segment. To reduce processing time, we considered only the road segments of the street at which accident had happened. As mentioned before, dataset with crash event details contained rows with missing values for street name on which accident happened even though the location (longitude and latitude) of the accident was known. Prior to merging this dataset with dataset containing crash event details, we had to map location of these samples to the street name. For every row with the missing street name, we found the street line which was the closest to the location of the accident. If there was no street line with the distance less than 0.001 the crash is considered as outlier and removed.

D. Traffic Volume Counts

Data from this dataset collected by Department of Transportation (DOT) represents hourly information of traffic volume count for the specific street for each direction. There are data for only 422 unique dates for the period September 13, 2014 to November 24, 2019 which is in total 1899 days. Because of this, after merging the dataset with the first three datasets based on "segment_id" feature we were left with 22,872 traffic accident records. Thinking that the traffic volume has significant influence on the crash severity, we decided to train our machine learning models with two (distinct) merged datasets: one without, and the other with traffic count information. We then compared their performance as well as most relevant detection features which is further examined in Section VI.

IV. Methodology

In this section, we first describe the challenges related to training the models with our dataset, and provide a methodology for overcoming them. Then, we propose a method for generating a synthetic dataset (with negative values) to ensure reduction in bias during accident prediction task. Next, we evaluate well-known machine learning methods on both datasets (with and without traffic count information). Lastly, we focus on treatment effect estimation and propose an action that would lead to decrease in the rate of serious accidents.

A. Feature Engineering

Though date and time columns can provide valuable information for the model, they can be unusable in their standard format ('DD-MM-YYYY HH:mm') so we decided to extract date, moth, year and time information into separate columns. We created two new columns: month and day, while for the time column we kept only the hour of the accident, since information about the minute of the accident doesn't represent a sufficiently useful information. Additionally, we transformed day of the accident to the column which represents day of the week. Since our new features (month, hour and day of week) represent cyclical data in nature, and data points which are presented in the one dimensional plane as furthest apart (i.e. hour 0 and hour 23) are in fact the closest, we transformed them into two dimensions using a sine/cosine transformation:

$$x_{sin} = sin(\frac{2 * \pi * x}{max(x)}) \tag{1}$$

$$x_{cos} = cos(\frac{2 * \pi * x}{max(x)}) \tag{2}$$

Location of the accident in our dataset is a geospatial feature. It is represented by longitude and latitude which actually represent three dimensional space. We mapped these two features to x, y and z by applying the following transformations:

$$x = cos(lat) * cos(lon) \tag{3}$$

$$y = cos(lat) * sin(lon) \tag{4}$$

$$z = sin(lat) \tag{5}$$

For categorical features, we used one-hot encoding which increased the number of columns. Feature 'Condition' from weather dataset contained 59 unique values where many of them appeared only a few times in data. In order to reduce the number of features of final model, we have replaced every Condition value with one of the following:

- CLD for normal or cloudy conditions (e.g. 'Mostly Cloudy', 'Cloudy', 'Partly Cloudy', 'Mostly Cloudy / Windy',...),
- LVS for low visibility conditions (e.g. 'Fog', 'Haze', 'Thunder', 'Drizzle and Fog',...)
- RD for dangerous road conditions (e.g. 'Rain', 'Heavy Rain', 'Snow', 'Snow and Sleet',...).

For each combination of segment which belongs to two-way street and a specific day there are two different rows in [13] containing information about traffic volume count by hour for every direction. In order to integrate this dataset with [12], we merged these two rows by calculating the mean value for every column.

As numerical features differ from each other by range, we standardized features by using the Z-normalization method before applying k-nearest neighbors algorithm (KNN). The standard score of a sample x is calculated as:

$$z = \frac{x - \mu}{\sigma} \tag{6}$$

where μ is the mean of the training samples, and σ is the standard deviation of the training samples.

B. Training the Models with Imbalanced Data

Nature of our dataset introduces challenge of working with imbalanced data. Around 85% of all accidents belong to non-serious (no injury or death) class. Class unbalance is addressed with the use of the synthetic minority over sampling technique (SMOTE) [15]. In order to create a perfectly balanced dataset for training, we first used slight under-sampling technique on majority class followed by oversampling of the minority class.

C. Negative Examples Generation

As accident prediction is a binary classification problem and our dataset contains only positive samples, we had to generate negative examples for training the model. Total number of positive and negative samples for every day after adding negative samples was twice the average number of car crashes by day. For every non-cras3h example we selected one location from the already known locations from our crash dataset. We were randomly choosing one of the locations from the street where there was no collision for the whole day. For the hour of the non-crash example we have randomly chose one number from 0-23. After we have selected the date, time and location, features from other datasets were joined in the identical way that we already described. Final result was a balanced dataset with 1,104,190 positive and 1,035,866 negative samples.

D. Trained Models and Ensemble Learning

We decided to focus on training gradient boosting and decision tree based models. They were chosen because of their robustness and generalization capabilities. Since we are working with tabular data, and not images, text or time series data - deep learning based approaches don't offer significant improvement over decision tree based methods. Furthermore, decision trees offer lower computational complexity as well as intrinsic interpretability, something which deep-learning models, which are often characterized as "black-box" models lack. We trained XGBoost, CatBoost, LightGBM, Extra Trees and KNN models. In order to tune the hyper-parameters of our models, we used grid search 5-fold cross-validation [16] method. Lastly, we used ensemble learning technique on two best performing models (XGBoost and LightGBM) by averaging their class probabilities. The best model for predicting car accident severity is LightGBM with following parameters: estimators = 120, num_leaves = 45, learning_rate = 0.1 and num_leaves = 45 and 'max_depth': -1. CatBoost model trained for the accident prediction task has the following parameters: depth = 10 and learning_rate = 0.05.

E. Treatment effect estimation

Average treatment effect estimation from observational data is a central topic in causal inference, and has received a great deal of attention in recent years. Potential outcomes framework [17] allows us to define the average treatment effect as:

$$\tau := \mathbb{E}\{Y(1) - Y(0)\} \tag{7}$$

where Y(1) and Y(0) are the potential outcomes of treated and non-treated groups such that Y=Y(T). Identification of τ assumes identical distributions of covariates and potential outcomes. This is something that can't be guarantied given our dataset, however such an assumption can be achieved through applying the propensity matching and inverse probability weighting (IPW) techniques. Solving the counterfactual problem, relies on identifying pairs of treated (T = 1) and control (T = 0) units whose covariates are similar or identical to each other. *Rubin et al.* [18] proved that this can be achieved through propensity score matching, where propensity score is defined as:

$$e(X) = P(T = 1|X) \tag{8}$$

where X represents the observed covariates. However, identifying the pairs with similar propensity score introduces a distribution bias on covariates X. We overcome this by applying the inverse probability matching (IPW) technique [18]. After applying the IPW, average treatment effect estimator can be represented as:

$$r = \frac{1}{n_1} \sum_{i:t_i=1} \frac{y_i}{e(X_i)} - \frac{1}{n_0} \sum_{i:t_i=0} \frac{y_i}{1 - e(X_i)} \tag{9}$$

where n_i represents number of samples of each group. Intuitively, IPW allows us to create a pseudo-population where covariates are the same between the treatment and control group. Treatment of interest was a city-wide 25 mph limit. We created a propensity score estimator using the logistic regression model, while the inverse probability matching was performed with IBM's *Causallib* Python package [19].

V. Results

In this section, we present the results of our proposed approach for predicting car accidents and car accident severity separately and reflect on the factors that influence the model's decision by displaying the feature importance.

A. Accident Severity Prediction

All models were tested on dataset of 72,657 samples, of which 58,840 were non severe accidents. As dataset is highly unbalanced, we evaluated the models by using the accuracy and macro-average precision, recall and F1-score. The performance of our predictive models are summarized in Table I.

We used ensemble learning technique on the LightGBM and XGBoost models, as they had the best performance, however this approach didn't lead to an increase in the F1-score.

TABLE I
RESULTS OF MODELS FOR PREDICTING ACCIDENT SEVERITY

		Precision	***Recall***	***F1-score***
Extra Trees	0	0.84	0.65	0.74
	1	0.24	0.48	0.33
	accuracy			0.62
	macro avg	0.54	0.57	0.53
LightGBM	0	0.83	0.83	0.83
	1	0.28	0.29	0.29
	accuracy			0.73
	macro avg	0.56	0.56	0.56
XGBoost	0	0.83	0.82	0.83
	1	0.28	0.30	0.29
	accuracy			0.72
	macro avg	0.56	0.56	0.56
CatBoost	0	0.83	0.82	0.83
	1	0.27	0.29	0.28
	accuracy			0.72
	macro avg	0.55	0.56	0.55
KNN	0	0.82	0.79	0.80
	1	0.21	0.25	0.23
	accuracy			0.69
	macro avg	0.52	0.52	0.52
Ensemble[1]	0	0.83	0.82	0.83
	1	0.28	0.30	0.29
	accuracy			0.72
	macro avg	0.56	0.56	0.56

[1] XGBoost and LightGBM

B. Feature Importance

The feature importance obtained by LighGBM classifier after hyper-parameter tuning is presented in Fig. 2. We observe that the features with highest influence on the model's decision are features related to location (LL), temperature (FeelsLike), time of the day (HOUR), street width (StreetWidth_Min), month (MONTH) and day of the week (DOW).

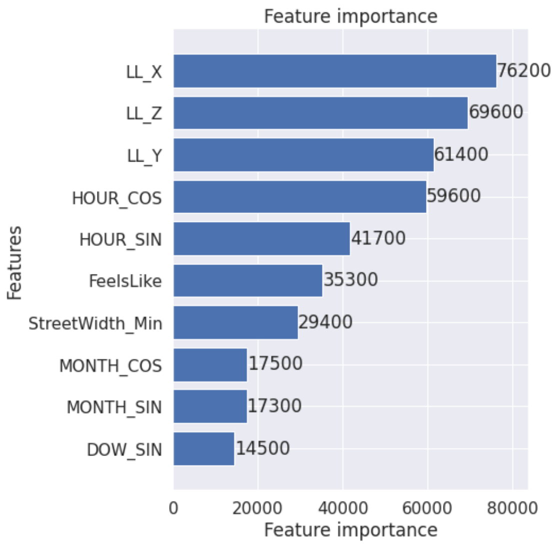

Fig. 2. Feature importance for the LightGBM model trained without traffic count data

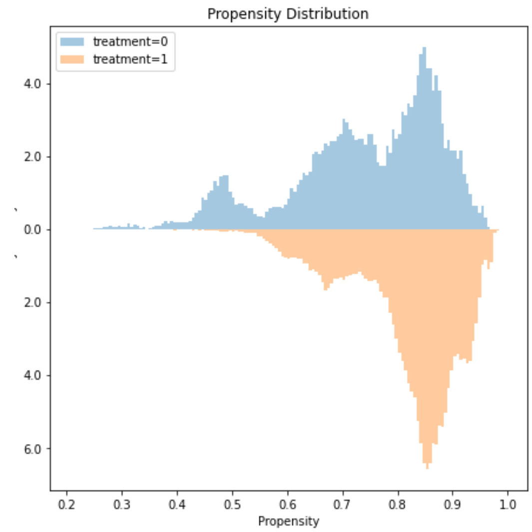

Fig. 3. Propensity distribution after applying the inverse probability weighting.

C. Accident Prediction

Models were trained on the same dataset which has been used for predicting accident severity with addition of generated negative samples. We used 10% of the data for the model evaluation which is in total 214,006. Table II presents the results obtained on the test set. We observe that macro average and accuracy scores are the same since the positive and negative samples are roughly equal in size. The best accuracy achieved is around 70%, while the features display similar importance scores as the ones shown in the Fig. 2.

TABLE II
RESULTS OF MODELS FOR PREDICTING ACCIDENTS

		Precision	*Recall*	*F1-score*
Extra Trees	0	0.64	0.60	0.62
	1	0.65	0.68	0.66
	macro avg	0.64	0.64	0.64
LightGBM	0	0.67	0.65	0.66
	1	0.68	0.71	0.69
	macro avg	0.68	0.68	0.68
XGBoost	0	0.69	0.67	0.68
	1	0.70	0.71	0.70
	macro avg	0.69	0.69	0.69
CatBoost	0	0.70	0.69	**0.69**
	1	0.71	0.72	**0.71**
	macro avg	0.70	0.70	**0.70**
KNN	0	0.46	0.57	0.51
	1	0.48	0.37	0.42
	macro avg	0.47	0.47	0.46
Ensemble[1]	0	0.69	0.67	0.68
	1	0.70	0.71	0.70
	macro avg	0.69	0.69	0.69

[1]CatBoost and XGBoost

D. Treatment effect estimation (T = 25 mph limit)

In order to estimate the propensity scores of the treatment, we trained a logistic regression model on the dataset without traffic count data. We decided to use this dataset because it contains significantly more observations. Initially, inverse probability didn't achieve an equal distribution between the two groups. The reason for this is that the dataset had many samples from the treatment group (25 mph limit or less) with covariates only belonging to that group (i.e. very few samples with similar covariates are found in the control group). After inspecting the dataset, we found that the vast majority of streets with 2 parking lanes or "Sector" snow priority are in fact 25 mph limit streets. After removing these observations and performing the same steps we achieved the desired distribution which can be seen in the Fig. 3. We can observe that the two groups belong to a roughly similar distribution. Thus, we can claim that the inverse probability weighting was successful. After applying the formula for average treatment effect (9), we estimate that a city-wide 25 mph limit would lead to 4% relative decrease in the rate of serious accidents. This would lead to 2,223 less severe accidents per year. However, even though this number is significant, repercussions of enforcing such rule should be carefully examined since it might lead to certain challenges associated with logistics and overall transportation system inside the city.

VI. Discussion

Our best model achieves macro-average f1-score of 0.56. In order to achieve better predictive performance, we decided to examine the effectiveness of introducing traffic count data. We believed that traffic density has a high influence on the outcome of the accident. Using the same model hyper parameters, we trained the models with traffic count data. Best performing model (Extra Trees) achieves macro-average f1-score of 0.69. Even though this is a significant increase, this solution is not as robust, since this dataset contains far less observations compared to the first dataset. In the Fig. 4, we can

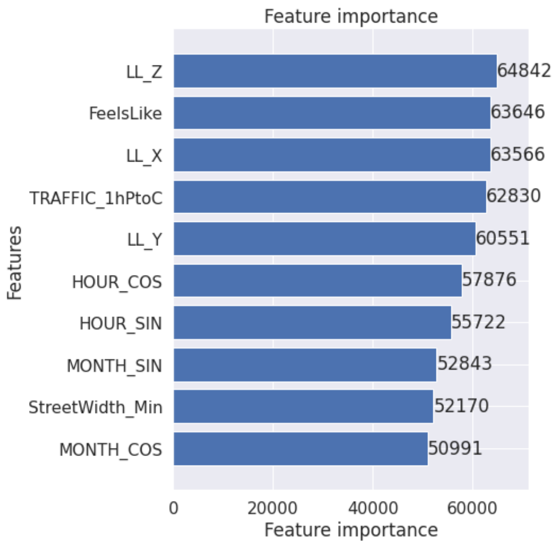

Fig. 4. Feature importance for the ExtraTree model trained with traffic count data

observe that the features have roughly the same contribution between the two approaches. We can also confirm our initial assumption that traffic count information has a significant influence on the severity of the traffic accidents.

As apparent from our approach, an open problem that remains for the research community to address is creation of publicly available datasets which contain sufficiently dense information related to the accidents.

VII. Conclusions and Future Work

A. Conclusions

In this paper, we presented a machine learning solution for predicting the severity of traffic accidents in New York City. After carefully collecting the relevant datasets and information related to traffic accidents, we trained various models. To address the class unbalance, we propose an approach which contains majority class undersampling and synthetic minority class oversampling techniques. We then evaluated various supervised learning algorithms where XGBoost and LightGBM were identified as the most suitable candidates in terms of accuracy, computational complexity and explainability. Next, we proposed an accident prediction method by generating negative samples. Finally, we presented an average treatment estimator and evaluated effectiveness of introducing a city-wide 25 mph speed limit.

We demonstrated that LightGBM model can achieve a 0.56 macro average F1 score for dataset without traffic count information. Furthermore, introducing traffic count information achieves better macro average F1 score performance (0.69), however this solution is not as robust, as it contains less observations than the first dataset. In the case of accident prediction, we achieved a f1-score of 0.70. Finally, examining the treatment of 25 mph speed limit after applying the inverse probability weighting showed an estimated 4% relative decrease in the rate of serious accidents.

B. Future Work

In future work, we aim to improve the model performance by adding additional features from more diverse data sources. Inclusion of information such as traffic sign proximity, population density, per-block population density and bike traffic data can introduce more dense information and improve the prediction performance. Additionally, including 3D content such as 3D models of buildings or geographic context such as road curvature can help us determine the drivers mean field of view, possible visual obstructions as well as dangerous road segments. Furthermore, collecting the city-wide traffic count information through New York City ”Vision Zero” [20] system can help affirm our assumption that traffic count information is important factor of accidents severity.

References

[1] Centers for Disease Control and Prevention (CDC), National Center for Injury Prevention and Control (NCIPC). Web-based Injury Statistics Query and Reporting System (WISQARS).

[2] Chen, Simiao & Kuhn, Michael & Prettner, Klaus & Bloom, David. (2019). The global macroeconomic burden of road injuries: estimates and projections for 166 countries. The Lancet Planetary Health. 3. e390-e398. 10.1016/S2542-5196(19)30170-6.

[3] Safe Urban Form: Revisiting the Relationship Between Community Design and Traffic Safety

[4] Dukiya, J., & Zagi, A.B. (2016). EMERGENCY AMBULANCE SERVICE SCHEME IN ROAD TRAFFIC ACCIDENT RESCUE OPERATION; THE CASE OF FCT ABUJA, NIGERIA.

[5] A. Hébert, T. Guédon, T. Glatard and B. Jaumard, ”High-Resolution Road Vehicle Collision Prediction for the City of Montreal,” 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 1804-1813.

[6] City of Montreal, “Montreal vehicle collisions.” [Online]. Available: http://donnees.ville.montreal.qc.ca/dataset/collisions-routieres

[7] Government of Canada, “National road network.” [Online]. Available: https://open.canada.ca/data/en/dataset/ 3d282116-e556-400c-9306-ca1a3cada77f

[8] “Historical climate dataset.” [Online]. Available: http://climate.weather.gc.ca/

[9] XU, Chengcheng, et al. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accident Analysis & Prevention, 2013, 57: 30-39.

[10] Yuan, Z., Zhou, X., Yang, T., Tamerius, J., & Mantilla, R. (2017, August). Predicting traffic accidents through heterogeneous urban data: A case study. In Proceedings of the 6th international workshop on urban computing (UrbComp 2017), Halifax, NS, Canada (Vol. 14, p. 10).

[11] https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

[12] https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-lion.pa

[13] https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts-2014-2019-/ertz-hr4r

[14] https://www.ibm.com/weather

[15] Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 2002, 16, 321–357

[16] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830

[17] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5):688, 1974.

[18] Rosenbaum P.R., Rubin D.B. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983a;70:41–55.

[19] Shimoni, Yishai et al. (2019). An Evaluation Toolkit to Guide Model Selection and Cohort Definition in Causal Inference.

[20] https://www1.nyc.gov/content/visionzero/pages/