Final Project Report
Manhattan Traffic Collision Map

Chongyang Ren MSAUD 23'
Junling Zhuang MSCDP 23'
Yaoze Yu MSAUD 23

Problem Definition

Traffic collisions are a major concern in urban areas, and can lead to severe injuries or fatalities. Manhattan is known for its high-density road network and heavy traffic, making it important to understand the factors that contribute to traffic collisions in this area. In order to reduce the frequency and severity of collisions, it is necessary to identify the intersections that are most prone to accidents and determine the factors that contribute to their danger.

Research Question

What factors contribute to the frequency and severity of traffic collisions at intersections in Manhattan, and how can this information be used to predict the risk of collisions at different locations?

Literature review

1. Luis F. Miranda-Morenoa, Patrick Morencyb, Ahmed M. El-Geneidyc. (2011). The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections
2. Ayesha Shafique, Guo Cao, Zia Khan, Muhammad Asad, Muhammad Aslam. (2022). Deep Learning-Based Change Detection in Remote Sensing Images: A Review
3. Xiaojiang Li, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, Weixing Zhang. (2015). Assessing street-level urban greenery using Google Street Viewand a modified green view index

Data Sources

1. Motor Vehicle Collisions - Crashes (https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95- From NYC Open Data)
2. Land Cover Raster Data (2017) – 6in Resolution (https://data.cityofnewyork.us/Environment/Land-Cover-Raster-Data-2017-6in-Resolution/he6d-2qns- From NYC Open Data)

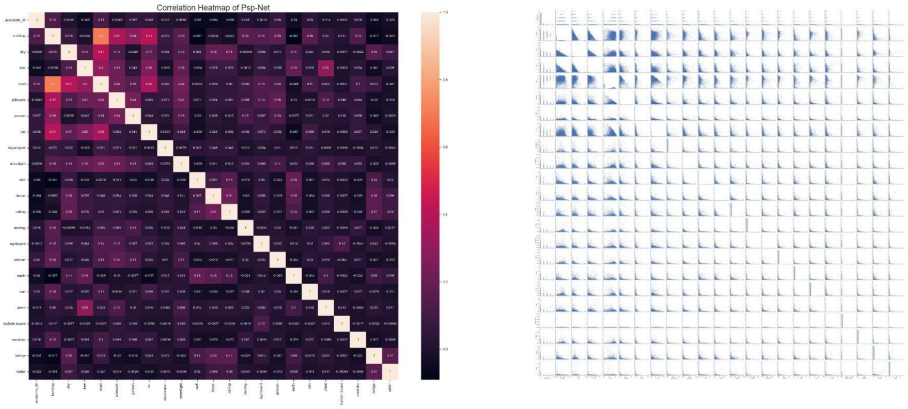
Brief Description

The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC.

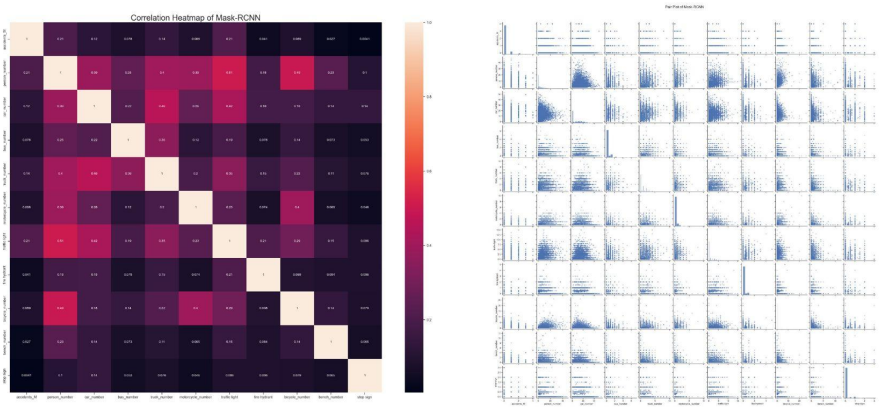
This dataset contains records of crash date, crash time and crash location and further detailed description. We extract **number of persons injured** and **number of persons killed** as the most important variable.

Exploratory Analysis

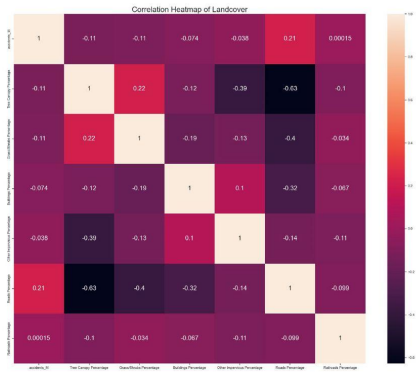
Exploratory Analysis - Psp Net

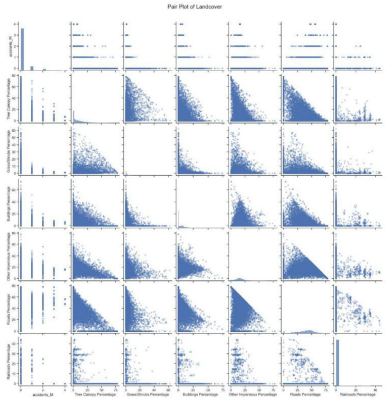


Exploratory Analysis - Mask RCNN

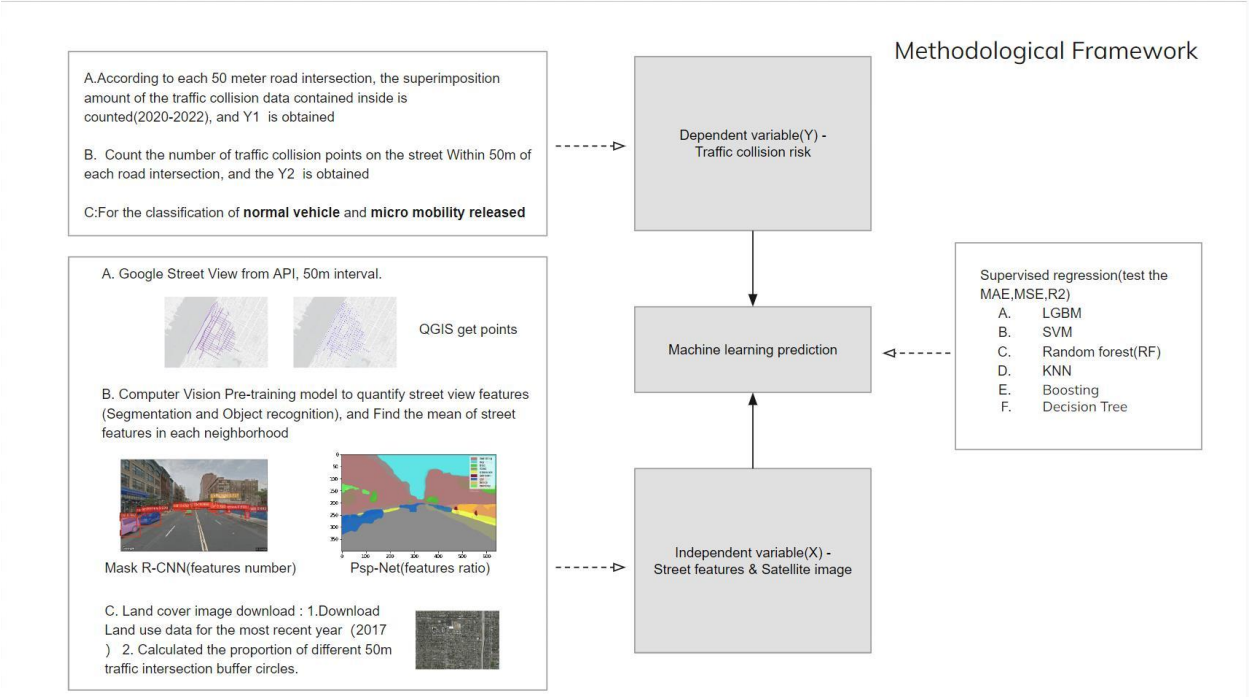


Exploratory Analysis – Land Cover





Methodology Framework



1. Defining the Traffic Collision Risk (Dependent Variable Y)

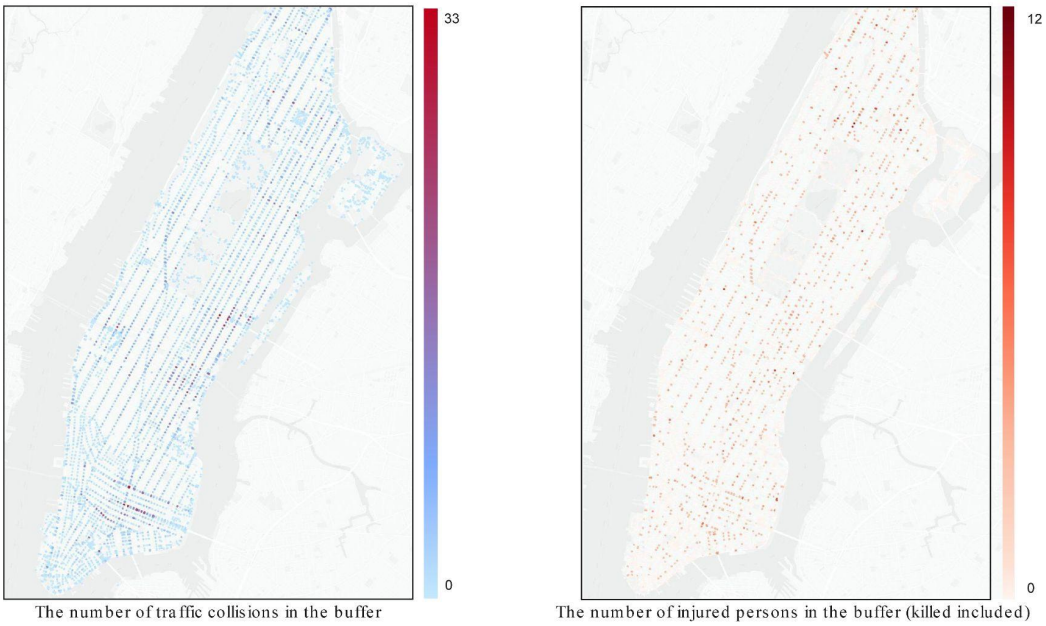
4.3. BE: land use, demographics, transit and road network

BE variables in the vicinity of each intersection are generated using Geographic Information Systems (GIS) data obtained from various sources. To account for the impact of buffer dimension, different buffer sizes were tested, including 50, 150, 400 and 600 m. Because the area under study (central neighbourhoods in Montreal) is more dense, has a rich mix of land uses, and has high transit accessibility, this study uses smaller buffers than those published by Pulugurtha and Repake (2008) and Schneider et al. (2009). The 50-m buffer was used to find how an intersection's immediate surroundings affected pedestrian activity. A 150-m buffer examines the effects of characteristics within close proximity to the intersection. The 400-m and 600-m buffers served as proxies for how characteristics at a walking distance or neighbourhood scale affect the level of pedestrian activity at a particular intersection. A list of the variables with a short definition is provided in Table 2. To extract demographic data, census data at the census tract level were intersected with each buffer generated around the intersections. To classify intersections according to the number of approaches (three-legged versus four-legged intersections), a dummy variable was generated.

a.

b. According to the paper"The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections": Section 4.3 we referred to, **the 50-m buffer was used to find how an intersection's immediate surroundings affected pedestrian activity, which fit our demand of researching how people get affected by traffic collisions happening around the intersections.**
c. According to each 50 meter road intersection, the superimposition amount of the traffic collision data contained inside is counted(2020-2022), and the Y of the neighborhood scale is obtained.
d. Count the number of traffic collision points on the street Within 50 meters of each road intersection.
e. For the classification of normal vehicles and micro mobility released.

Buffer Zone Analytical Visualization

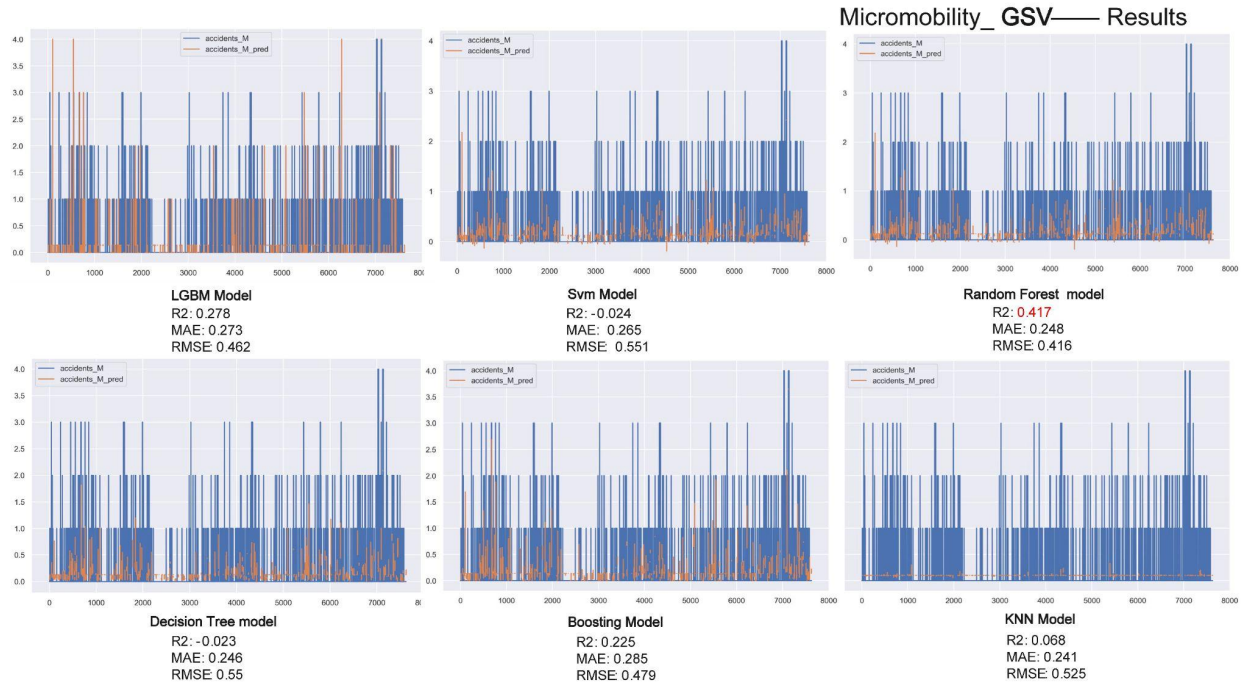


The number of traffic collisions in the buffer

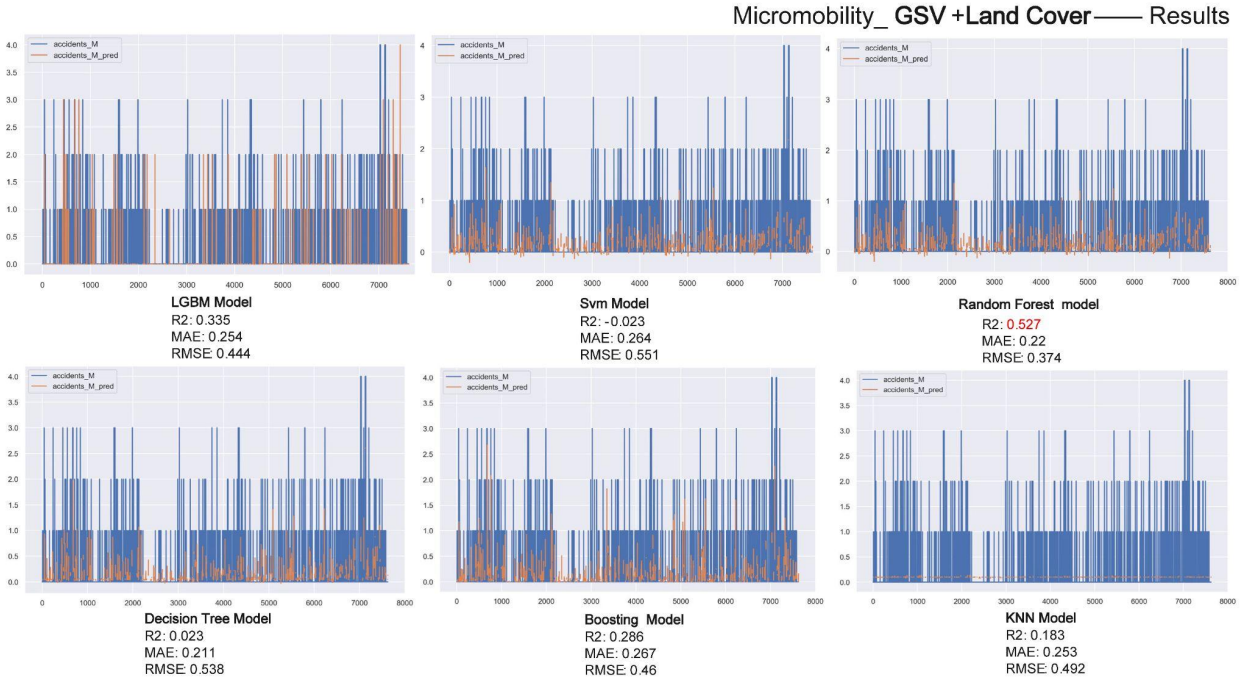
The number of injured persons in the buffer (killed included)

2. **Measuring the street views + Satellite Image (Independent Variable X)**
    a. GSV collection
        i. Download street pictures through google Api, 50m interval, 800x600 pixels.
        ii. Use Mask R-CNN to get the type and number of features.
        iii. Use Psp-Net to get the ratio of features.
        iv. Calculate the mean of street features in the same buffer circle
    b. Land cover image download
        i. Download 2017 NYC land cover image.
        ii. Calculated the proportion of different 50m traffic intersection buffer circles.
3. **Machine learning prediction**
    a. Use the training data of Manhattan 2020-2022 to predict the risk of collisions at different locations?
    b. Supervised regression(test the MAE,MSE,R2) to select a model of best performance.
        i. LGBM
        ii. SVM
        iii. Random forest(RF)
        iv. KNN
        v. Boosting
        vi. Decision Tree

Results

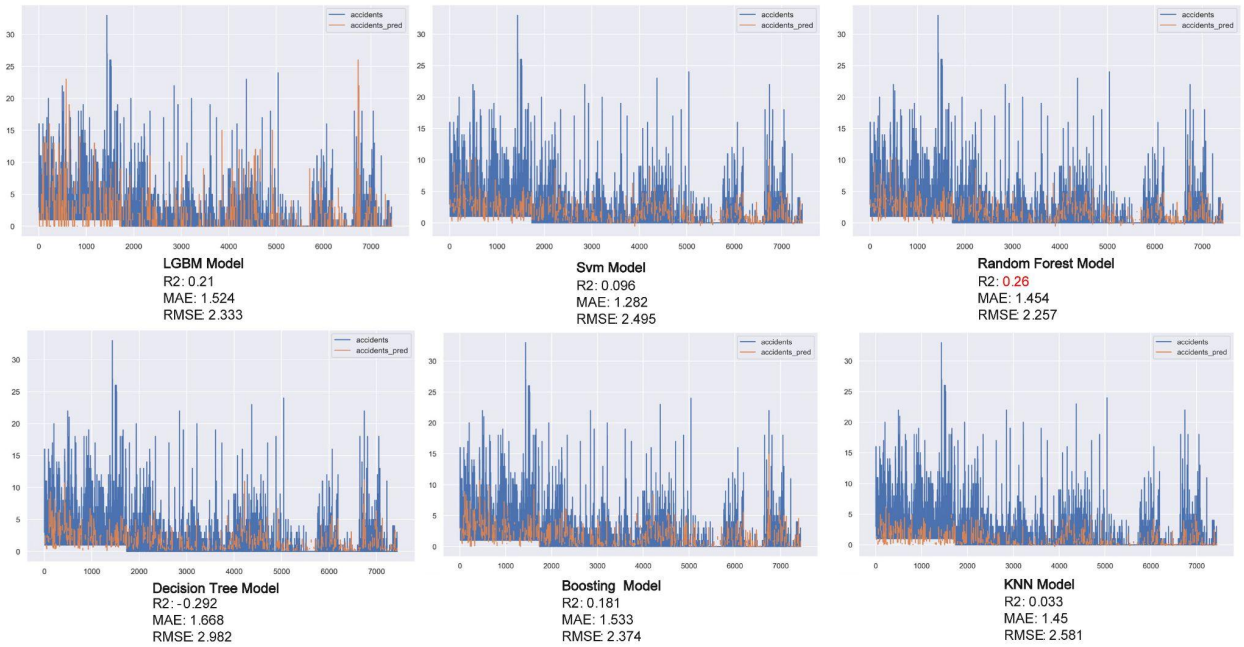


1. Firstly, for Micromobility, we conducted a regression analysis between the number of accidents events within the traffic intersection buffer, considering only GSV data, and the street-level built environment features segmented by Mask R-CNN and PSPNet. The LightGBM, Gradient Boosting, and Random Forest models all achieved an R2 score greater than 0.2. The MAE and RMSE values were centered around 0.2 and 0.4, respectively.



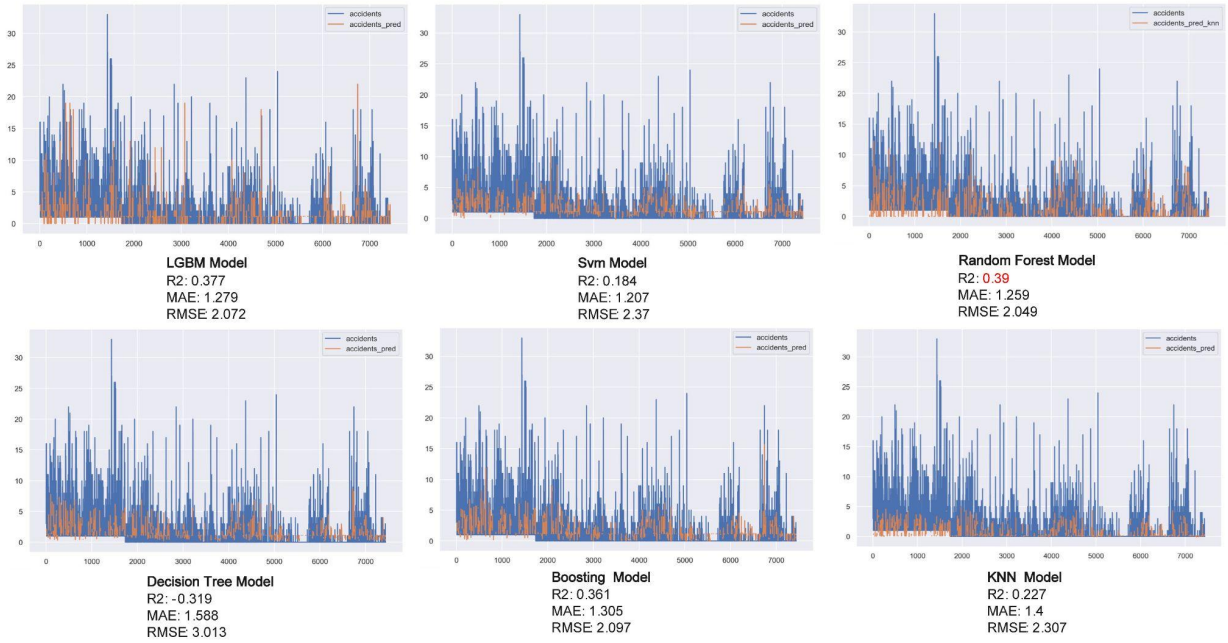
2. However, when we incorporated Land Cover into our analysis, the overall R2 scores experienced a significant improvement, particularly for the Random Forest model, which increased from 0.41 to 0.52. The RMSE values for each model also decreased by approximately 0.5 to 0.1 when compared to the cases considering only GSV data.

Normal_ **GSV**—— Results

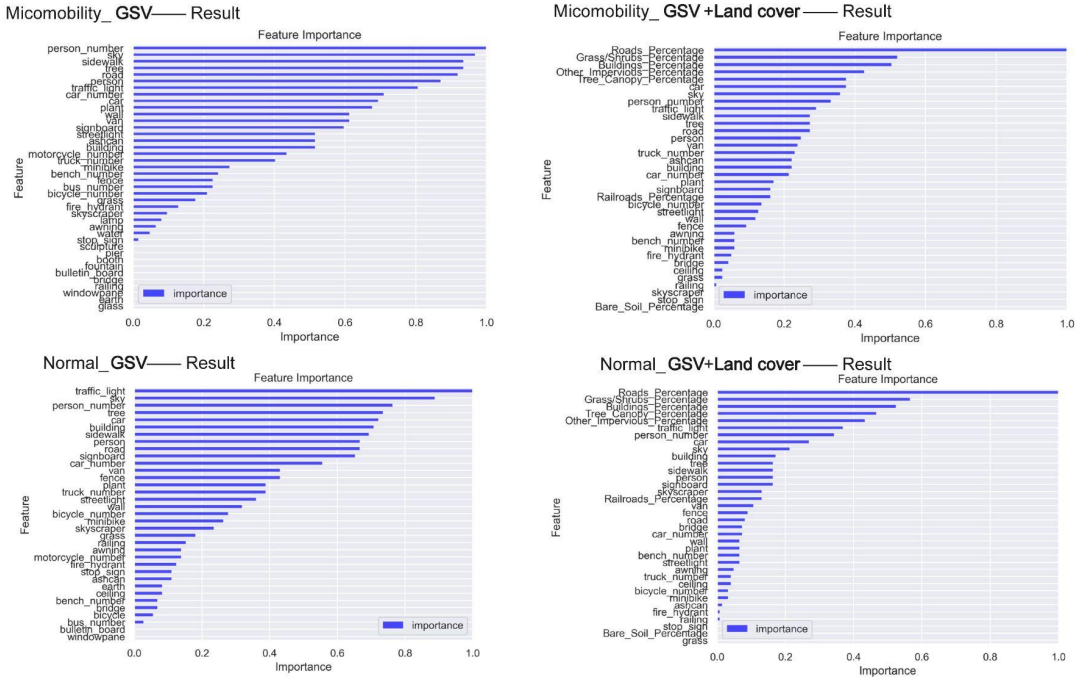


3. We employed the same approach for normal vehicles as well. When considering only GSV data, the Decision Tree Model, LightGBM Model, and Random Forest Model all exhibited relatively high R2 scores, though the Decision Tree Model displayed a negative correlation overall. At the same time, the MAE and RMSE values, in comparison to those for micro mobility, increased significantly, reaching 1.5 and 2.3, respectively.

Normal_ **GSV+Land Cover** —— Results



4. Upon incorporating Land Cover into our analysis, the overall performance of the models improved considerably. Among the six models, the Random Forest model demonstrated the best performance.



5. We then ranked the features' importance during the regression of the Random Forest model using GINI importance. We found that, without considering Land Cover, parameters such as traffic light, tree, car, buildings, sky, and person had a relatively high impact. However, after including Land Cover, the road percentage

became the most influential factor.

Conclusions & Implications & Limitations

- Random Forest (RF) model takes the slowest time but has the highest R^2 and the best performance among all models.
- According to the paper “The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections”: Table 3 we referred to, A correlation matrix for each buffer size is generated, including 50, 150, 400 and 600m. Therefore, The diameter size of the buffer zone should also be adjusted to 150, 400, 600 meters for comparative analysis.

Table 3
Correlations between BE and pedestrian activity, traffic and collision frequency.

	50 m buffer			150 m buffer			400 m buffer			600 m buffer		
Variables	*A*	ln(*P*)	ADT	*A*	ln(*P*)	ADT	*A*	ln(*P*)	ADT	*A*	ln(*P*)	ADT
Pedestrian collisions	1.00			1.00			1.00			1.00		
Log of pedestrian volume	0.36	1.00		0.36	1.00		0.36	1.00		0.36	1.00	
Pedestrian volume	0.37	0.84		0.37	0.84		0.37	0.84		0.37	0.84	
AADT	0.40	0.11	1.00	0.40	0.11	1.00	0.40	0.11	1.00	0.40	0.11	1.00
Commercial	0.32	0.49	0.06	0.24	0.42	0.05	0.11	0.36	0.02	0.07	0.35	−0.02
Residential	−0.07	−0.14	−0.20	0.05	0.01	−0.16	0.10	0.16	−0.13	0.10	0.16	−0.10
Open space	−0.03	−0.23	0.52	−0.07	−0.27	0.43	−0.10	−0.34	0.28	−0.05	−0.27	0.24
Parks/recreational	−0.09	−0.13	−0.06	−0.10	−0.15	−0.03	−0.13	−0.16	−0.06	−0.09	−0.12	−0.07
Industrial	−0.04	−0.08	0.07	−0.02	−0.05	0.09	−0.01	−0.06	0.12	−0.02	−0.03	0.11
Number of schools	0.00	0.00	−0.03	−0.08	0.12	−0.01	0.09	0.31	−0.02	0.09	0.27	−0.03
Number of jobs	0.16	0.26	0.06	0.17	0.33	0.17	0.10	0.36	0.08	0.02	0.20	0.09
Population	0.10	0.38	−0.17	0.12	0.42	−0.17	0.15	0.47	−0.15	0.16	0.47	−0.12
Employees	0.06	0.39	−0.18	0.08	0.42	−0.18	0.11	0.46	−0.15	0.11	0.45	−0.13
Seniors	0.10	0.19	−0.13	0.12	0.20	−0.13	0.18	0.23	−0.07	0.19	0.20	−0.05
Children	0.09	0.24	−0.13	0.12	0.28	−0.11	0.20	0.34	−0.06	0.22	0.35	−0.03
Presence of metro station	0.07	0.17	−0.01	0.20	0.32	0.14	0.08	0.32	0.05	0.04	0.24	0.04
Number of bus stops	0.42	0.23	0.56	0.31	0.37	0.36	0.08	0.34	0.00	0.02	0.32	−0.07
Km of bus routes	0.32	0.29	0.49	0.26	0.30	0.38	0.00	0.18	0.05	−0.08	0.08	0.07
Road length (km)	0.06	−0.08	0.48	0.08	0.14	0.18	0.07	0.27	−0.01	0.07	0.31	−0.02
Class 1 – Primary Highway	−0.02	−0.15	0.31	−0.02	−0.16	0.31	−0.08	−0.20	0.08	−0.10	−0.21	−0.02
Class 2 – secondary highway	0.10	0.07	0.12	0.05	0.10	0.06	0.01	0.17	−0.04	0.01	0.19	0.00
Class 3 – arterial road	0.22	0.01	0.52	0.17	−0.01	0.49	0.01	−0.08	0.21	0.01	−0.02	0.09
Class 4 – local road	−0.26	−0.03	−0.49	−0.10	0.15	−0.43	0.10	0.38	−0.20	0.11	0.39	−0.14
Number of street segments	−0.04	−0.07	0.05	−0.01	0.02	−0.04	0.00	0.03	0.06	−0.02	0.06	0.08
Number of intersections	−0.02	−0.12	0.33	−0.02	−0.01	0.12	−0.04	0.04	0.03	−0.02	0.12	0.03
Portion of major roads	0.29	0.03	0.60	0.17	−0.06	0.57	−0.06	−0.22	0.25	−0.10	−0.22	0.17
Average street length	0.11	0.00	0.43	−0.10	−0.13	−0.23	−0.07	−0.24	0.06	−0.07	−0.18	0.11
Intersection type	0.16	0.30	−0.05	0.16	0.30	−0.05	0.16	0.30	−0.05	0.16	0.30	−0.05

- After debugging the model with relatively optimal performance again, it is hoped that in the future, the risk of traffic collisions at road intersections in other low data areas can be predicted.