Predicting the Severity of Pedestrian Collisions in the City of Toronto

2023-01-09

Key Points:

i.
ii.

1. Background

Road safety is important for any society. It can enhance economic growth by ensuring the safe movements of resources, including human capital, across an economy. For this reason, the City of Toronto's Vision Zero Road Safety Plan is critical for the sustenance of its economy. Therefore, the goal of this project is to support this plan by investigating the factors that influence the probability that a pedestrian collision will result in serious injury or fatality.

2. Modelling Methodology

The outcome variable in this project is a binary variable, which equals 1 if the pedestrian collision resulted in major injury or fatality and 0 otherwise. Therefore, I used a classification algorithm (Random Forest) to model the probability that the outcome variable is equal to 1. The analysis was conducted with the R statistical programming language. To ensure the complete reproducibility of the results of this study, I employed the "Renv" package to document all the packages, including their metadata, I used in the project. Also, I used GitHub and Git to version control the project and to promote collaboration. Some of the major activities I conducted in this step are:

1. I extracted pedestrian collision data from the involved_persons data.

2. I dropped variables with missing or empty information from the provided datasets. I did not use imputation to fill in the missing information because it may introduce bias in the datasets.

3. checked for duplicate rows in the two datasets. The involved_persons data had duplicate rows; for those rows, I maintained 1 and dropped the other duplicates.

4. created the outcome binary variables (serious_fatal) from the injured_persons data; it equals 1 if the pedestrian collision (involved_injury_class column) is fatal or major and 0 otherwise. Also, I created 10 binary variables from the 2 datasets (variable definitions are in Table 1 in the appendix).

5. serious injury or fatal pedestrian collisions account for about 11% of the information in the outcome variable. Estimating the model with such data could result in unreliable inference. I used the Synthetic Minority Oversampling Technique (SMOTE) to solve the imbalance problem. After the

SMOTE analysis, the 11% statistic increased to about 31%.

6. conducted hyper-parameter tuning on mtry and min_n Random Forest algorithm (RF) parameters; mtry is the number of variables randomly selected as candidates in each split, min_n is the minimum number of data points in each node required for it to be split further. This was to ensure that I get a best-fit RF.

3. Findings

created the outcome binary variables (serious_fatal) from the injured_persons data; it equals 1 if the pedestrian collision (involved_injury_class column) is fatal or major and 0 otherwise. Also, I created 10 binary variables from the 2 datasets (variable definitions are in Table 1 in the appendix). serious injury or fatal pedestrian collisions account for about 11% of the information in the outcome variable. Estimating the model with such data could result in unreliable inference. I used the Synthetic Minority Oversampling Technique (SMOTE) to solve the imbalance problem. After the SMOTE analysis, the 11% statistic increased to about 31%. conducted hyper-parameter tuning on mtry and min_n Random Forest algorithm (RF) parameters; mtry is the number of variables randomly selected as candidates in each split, min_n is the minimum number of data points in each node required for it to be split further. This was to ensure that I get a best-fit R

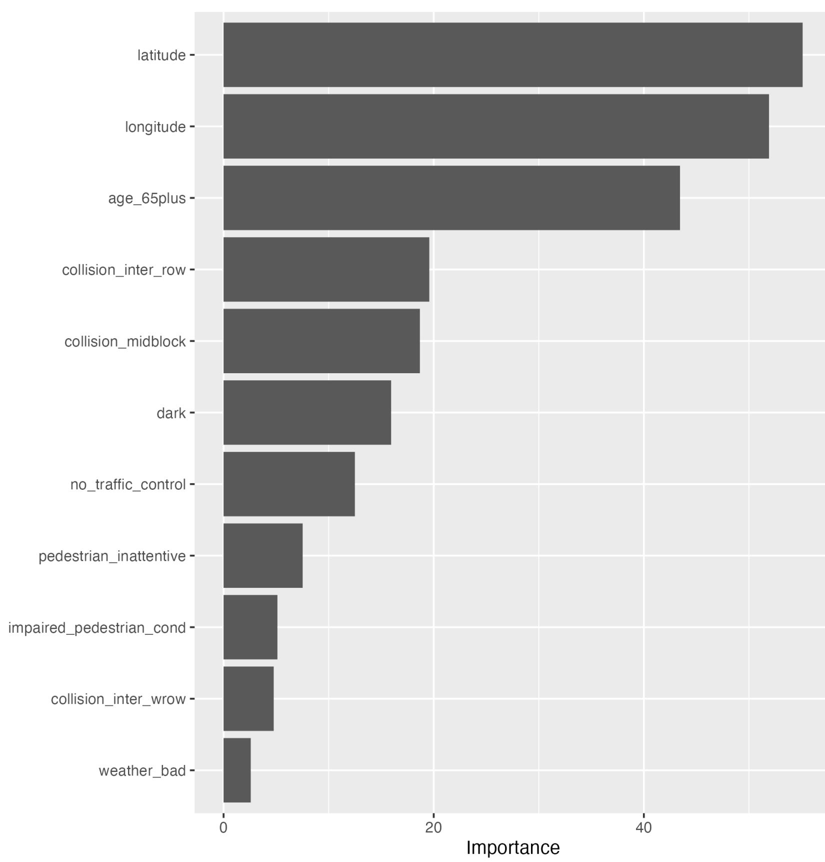

Figure 1: Variable Importance Plot

4. Model Accuracy

reated the outcome binary variables (serious_fatal) from the injured_persons data; it equals 1 if the pedestrian collision (involved_injury_class column) is fatal or major and 0 otherwise. Also, I created 10 binary variables from the 2 datasets (variable definitions are in Table 1 in the appendix). serious injury or fatal pedestrian collisions account for about 11% of the information in the outcome variable. Estimating the model with such data could result in unreliable inference. I used the Synthetic Minority Oversampling Technique (SMOTE) to solve the imbalance problem. After the SMOTE analysis, the 11% statistic increased to about 31%. conducted hyper-parameter tuning on mtry and min_n Random Forest algorithm (RF) parameters

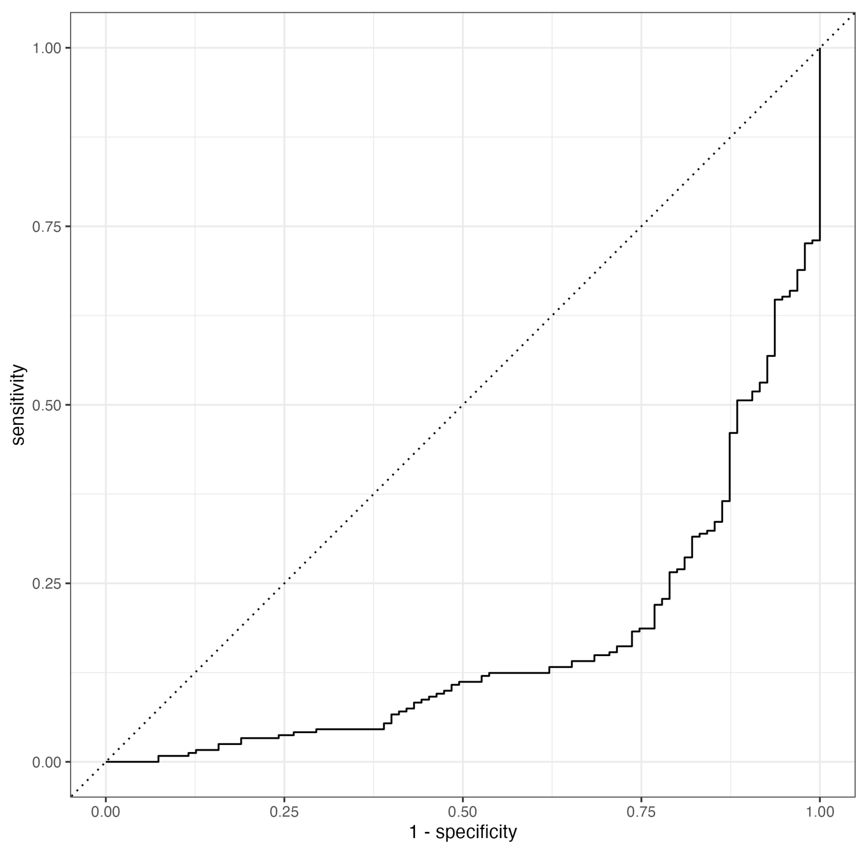

Figure 2: Receiver Operating Characteristic Curve

5. Future Work

reated the outcome binary variables (serious_fatal) from the injured_persons data; it equals 1 if the pedestrian collision (involved_injury_class column) is fatal or major and 0 otherwise. Also, I created 10 binary variables from the 2 datasets (variable definitions are in Table 1 in the appendix). serious injury or fatal pedestrian collisions account for about 11% of the information in the outcome variable. Estimating the model with such data could result in unreliable inference. I used the Synthetic Minority Oversampling Technique (SMOTE) to solve the imbalance problem. After the SMOTE analysis, the 11% statistic increased to about 31%. conducted hyper-parameter tuning on mtry and min_n Random Forest algorithm (RF) parameters

6. Data Quality Issue

reated the outcome binary variables (serious_fatal) from the injured_persons data; it equals 1 if the pedestrian collision (involved_injury_class column) is fatal or major and 0 otherwise. Also, I created 10 binary variables from the 2 datasets (variable definitions are in Table 1 in the appendix). serious injury or fatal pedestrian collisions account for about 11% of the information in the outcome variable. Estimating the model with such data could result in unreliable inference. I used the Synthetic Minority Oversampling Technique (SMOTE) to solve the imbalance problem. After the SMOTE analysis, the 11% statistic increased to about 31%. conducted hyper-parameter tuning on mtry and min_n Random Forest algorithm (RF) parameters