A blue circle and triangle shaped letters

Description automatically generated with medium confidence

**Prioritizing Pedestrian Safety: A Comprehensive Risk Assessment of Montreal's Signalized Intersections using Advanced Statistical Techniques**

**Work presented to**

Dr. Aurélie Labbe

**As part of the course**

Advanced Statistical Learning

MATH60611A

**By**

Siddhant Arora (11302420)

Philippe Béliveau (11320035)

Tuesday, March 12th, 2024

# Table of Contents[[1]](#footnote-0)

[Introduction](#_pgs3cl5t3po) …………………………………………………………………………………………………………………………………………. 1

[Data Exploration](#_jmm33kktoe4k) …………………………………………………………………………………………………………………………………. 1  
  
[Data Processing](#_ms3ogwx8lwpn) ……………………………………………………………………………………………………………………………….… 2  
  
[Modeling](#_2160s865zfqg) ………………………………………………………………………….…………………………………………………………………… 3

[Key Characteristics of Hazardous Intersections](#_4tj1de6vv7ba) ……………………………………………………………………………  
  
[Conclusion](#_yh4y16mz0f7h) …………………………………………………………………..……………………………………..………………………………. 7

# 

# 

# 

# 

# 

# 

# Introduction

The objective of this project is to assess the risk levels of intersections in Montreal. To achieve this, we are leveraging a comprehensive database comprising over fifty features related to the characteristics of around 1800 intersections. These features include specific location data, traffic volume, incidence of accidents, and signage information. Some features may overlap in the information they provide, necessitating a refinement process to streamline our analysis and recommendations.

Our primary objective is to analyze the frequency of accidents occurring at intersections over the past decade. However, we recognize that relying solely on accident counts may not provide a comprehensive understanding of an intersection's safety level. For instance, intersections with inadequate signage may pose significant hazards despite a relatively low accident count. Therefore, we aim to develop a predictive model capable of estimating the true number of accidents that could have occurred over 10 years based on intersection characteristics.

If our model demonstrates strong predictive performance, it would suggest that the variables used to estimate accidents serve as reliable indicators of intersection hazard levels. Consequently, our report will present a comprehensive analysis of the factors contributing to intersection safety, to provide actionable insights for enhancing road safety in Montreal.

# Data Exploration

## Target Variable

Let's begin by examining the target variable “acc” (number of accidents). The distribution of this variable is notably right-skewed (Figure 1), with approximately 35% of its observations clustered near zero. In literature, such distributional properties are often characterized by a zero-inflated Poisson distribution. Given that our variable represents accident counts, it naturally aligns with count data principles, falling within the framework of Poisson-type distributions, which will guide our analysis. Figure 1: Target variable’s distribution

However, it's important to note that the distributional assumptions of Poisson are not fully met in our dataset. Specifically, there is no linear relationship between the features and the logarithm of the expected accident count. Additionally, the mean of the variable is not equal to its variance. The mean of the target variable is found to be equal to 2.21 while its variance is at 10.10. Thus, our analysis considers testing with other count distributions that can handle overdispersion such as Negative Binomial.

## Multicollinearity

The data shows an important amount of collinearity between the variables, 40 pairs of variables show correlation coefficients over 70%. Some of the variables contain the same information, while some are even simply a transformed version of the original one. Our analysis takes this into account and aims at eliminating some of this colinearity with the help of different regularization techniques.

# Data Processing

We start by removing redundant variables such as X, X.1, street\_1, street\_2, rue\_1, rue\_2, x, y, date\_ and the intersection number, as these variables are too granular and specific to the intersections or are empty or duplications. Furthermore, the dataset does not contain any important amount of missing values. We do not remove highly correlated variables, nor apply any transformation on the features, we aim at letting the regularization technique and the PCA do their work and eliminate pairs of correlated variables that could affect the estimate and performance of the model. However, we remove half\_phase and keep new\_half\_r as they are strongly correlated (0.98) and as new\_half\_r has more information compared to half\_phase (with 12 intersections being corrected). Lastly, pi and ped\_100 are perfectly collinear (1), so we keep pi instead of ped\_100 because of interpretability as average annual daily flow is more interpretable than dividing it by 100. And similarly, fi and traffic\_10000 are perfectly collinear (0.99), we keep fi and drop traffic\_10000 because of the same reason. For the remaining highly correlated variables we let the regularization techniques do their job.

## Neighbourhood Variable

Additionally, some neighbourhoods tend to have a higher number of accidents per intersection, such as Villeray\_Saint\_Michel\_Parc\_extentision with a count of 3 accidents on average per intersection, against a mean of 1.49 for all intersections. We aim to incorporate this variable into our accident prediction model. However, before proceeding, it's crucial to assess if we possess enough observations across different neighbourhood levels. Upon examination, we find that several neighbourhoods contain a limited number of observations. Consequently, we opt to group levels with fewer than 50 observations. As a result of this grouping, we create a new level termed "Other neighbourhood," which encompasses 182 observations.

# Modelling

## Generalized Linear Model

Given the characteristics of our target variable, we opted for a Generalized Linear Model (GLM) framework with different families of distributions related to count data. These included Poisson, negative binomial, and zero-inflated Poisson. We explored multiple distribution families because none of the assumptions associated with each distribution were fully met. Our goal was to compare the performance of these GLM models using different count distributions and select the most suitable one.

Furthermore, we applied Lasso-type regularization techniques to address the high correlation and perform variable selection. We prefer Lasso over Ridge in this case as we arguably have a very high amount of correlated variables, and instead of averaging the effect of two correlated variables, we aimed to shrink one of the coefficients of the correlated variables towards zero, thereby facilitating better interpretation and providing clearer insights for the recommendation aspect of the project. Ridge, as we know, would shrink the coefficients but won’t perform variable selection.

As we can see from the table 1 below, the Poisson model performs slightly better than the other distribution with the same amount of variables. Very little amount of regularization is applied, and the model is still able to reduce the number of features from 63 to 43.

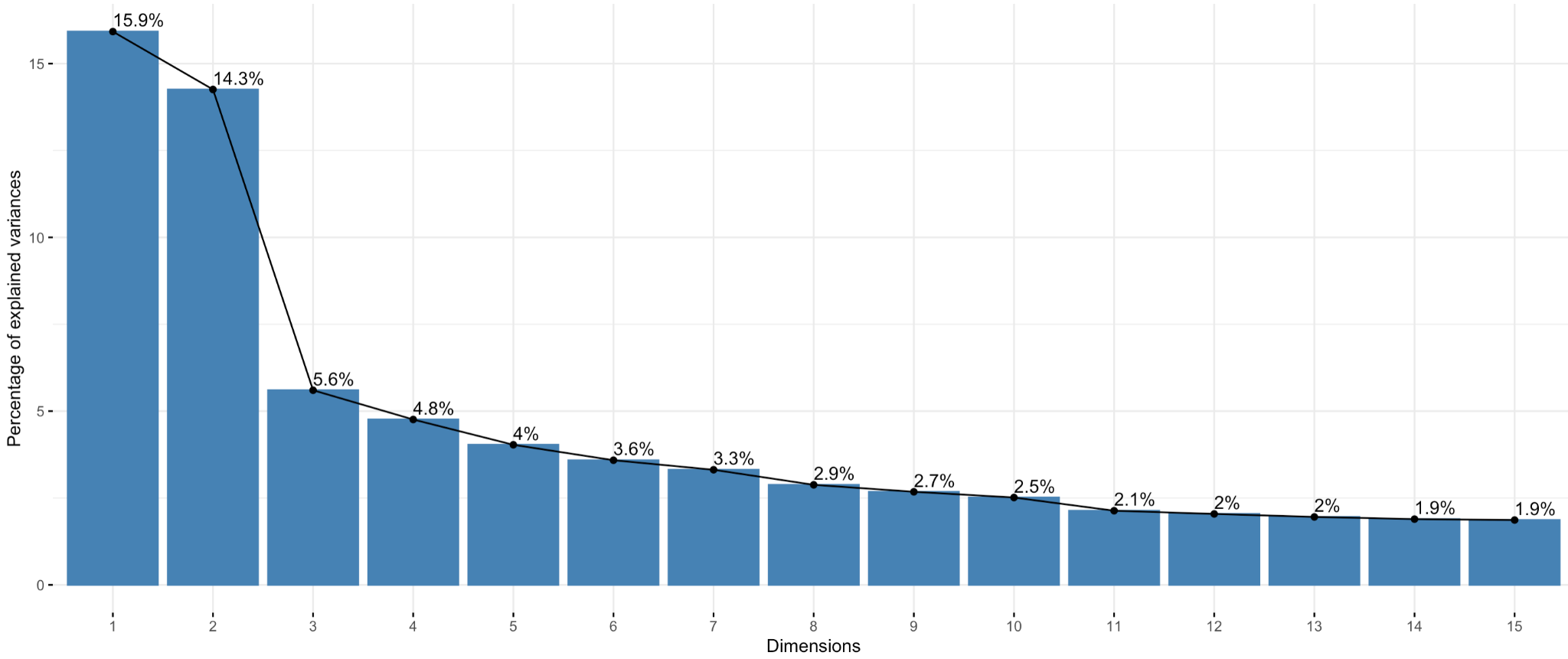
| **GLM model family** | **Lambda** | **MAE Test** | **MSE Test** | **Number of Variable Selected** |
| --- | --- | --- | --- | --- |
| **Poisson** | **0.018** | **1.58** | **5.98** | **43** |
| Negative binomial | 0.005 | 1.61 | 6.94 | 50 |
| Zero-inflated Poisson | No regularization applied[[2]](#footnote-1) | 1.60 | 6.53 | 43[[3]](#footnote-2) |

Table 1: Results of GLM models

We can observe that the method applying the least regularization on the dataset is performing the worst, which shows the relevancy of applying regularization technique on our dataset. Furthermore, the most important variables selected by the GLM model assuming a Poisson distribution are all\_pedest (-0.20), ln\_pi (0.53), ln\_fi (0.38), (0.26), median (-0.22), all\_red\_an (-0.64), Ahuntsic-Cartierville (-0.35), Saint-Laurent (-0.63), Ville Marie (0.15) and St-Leonard (0.25) neighbourhood. These preliminary results show some important intersection characteristics that could potentially lead to higher numbers of accidents, such as specific neighbourhood, the average annual daily flow for pedestrians and vehicles passing at the intersection (pi/fi) and the presence of any measure to protect pedestrians, such as green-straight arrow (green\_stra) or the presence of traffic management at the intersection (median).

## Dimensionality Reduction

Another analysis we aimed to do was to assess the change in performance of a random forest and boosting algorithm when using Principal Components as variables instead of utilizing the original variables. Also, in light of reducing the amount of collinearity of our variables, we aim to use PCA's power to produce orthogonal components, which are uncorrelated variables. We also believe that by using the principal components compromising the most variability instead of the original variable, we will be limiting the number of irrelevant features picked at random at every iteration to build the split with. We believe that this could help build more uncorrelated trees as PCA are all uncorrelated to each other in their composition and allow the random forest to make its split more efficient as it will have access to a smaller set of features that contains more information than if they were the original one (that is to say, for example, 20 PCs would probably explain more variation than 20 random original variables).. Additionally, we aim to evaluate the possibility of using regularized PCA to interpret the factors contributing to the intersection's danger level.

  
Figure 3: Elbow/Scree plot for PCA analysis

Starting with transforming our variables into PCs, we find that the cumulative proportion of variance explained by the first 15 components, before regularization, amounts to 70%. As we will regularize our PCs and by definition have to reduce the amount of variation explained by them, we prefer to include more PCs than not. Thus, we are going to include the first 35 components regardless of the elbow rule (Figure 3), and even considering this could have an impact on the ease of explanation of the model’s output.

Then, using a penalty of 32 for each of the 35 PCs, we are still able to account for 72% of the variation of the data and get sparse enough PCs so that the 3 most important variables in the components account for it in the majority. The table 2 below reveals the most important variables composing each PC.

| **Components** | **The Three Most Important Variables** | **Coefficients** |
| --- | --- | --- |
| 1 | avg\_crossw/tot\_road\_w/tot\_crossw | -0.47/-0.4/-0.39 |
| 2 | pi/cti/cri | 0.92/0.24/0.23 |
| 3 | ln\_cli/ln\_cti/ln\_fli | -0.63/-0.45/-0.43 |
| 4 | fti/ln\_fti/ln\_fi | 0.64/0.47/0.36 |
| 5 | fri/fli/fi | -0.65/-0.62/-0.33 |
| 6 | any\_ped\_pr/green\_stra/new\_half\_r | 0.71/0.57/0.43 |
| 7 | east\_veh/south\_veh/west\_veh | 0.61/-0.54/0.52 |
| 8 | lt\_restric/lt\_prot\_re/ln\_cli | -0.77/-0.63/-0.09 |
| 9 | curb\_exten/tot\_road\_w/tot\_crossw | -0.97/-0.2/0.12 |
| 10 | grouped\_boroughVille\_Marie/ln\_distdt/distdt | 0.61/-0.55/-0.42 |
| 11 | grouped\_boroughOther/grouped\_boroughAhuntsic\_Cartierville/distdt | -0.97/0.24/-0.1 |

Table 2: Most important variables composing each Principle Component

We observe that the first six principal components (PCs), which account for the majority of the variability in the data, are primarily explained by specific intersection characteristics. These include the width of the crosswalk and of the road along each approach of the intersection (avg\_crossw/tot\_crossw), the average annual daily flow of pedestrians (pi) and vehicles (fi/fri/fli), and the number of potential interactions between a pedestrian and vehicle turning either left or right over each 15 min intervals and the presence of a semi-protected pedestrian phase (green\_strat/new\_half\_r). Interpreting these PCs is straightforward and intuitive, as each PC comprises a combination of variables representing similar information, although in a transformed manner.

However, our analysis doesn't end here. To further understand the significance of these characteristics, we aim to investigate how both the random forest and boosting models utilize these principal components to make predictions.

## Tree-Based Model

### Random Forest

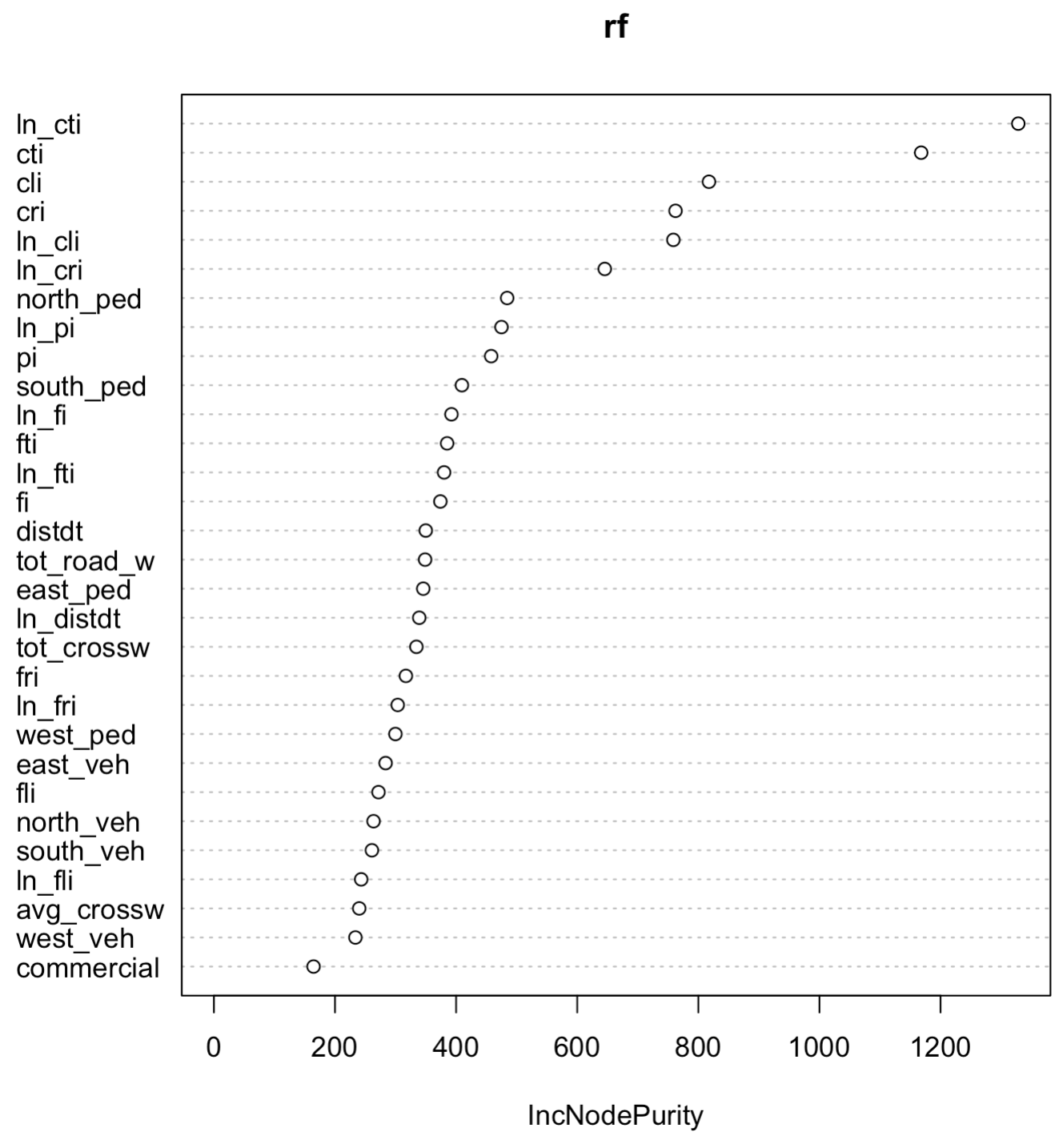
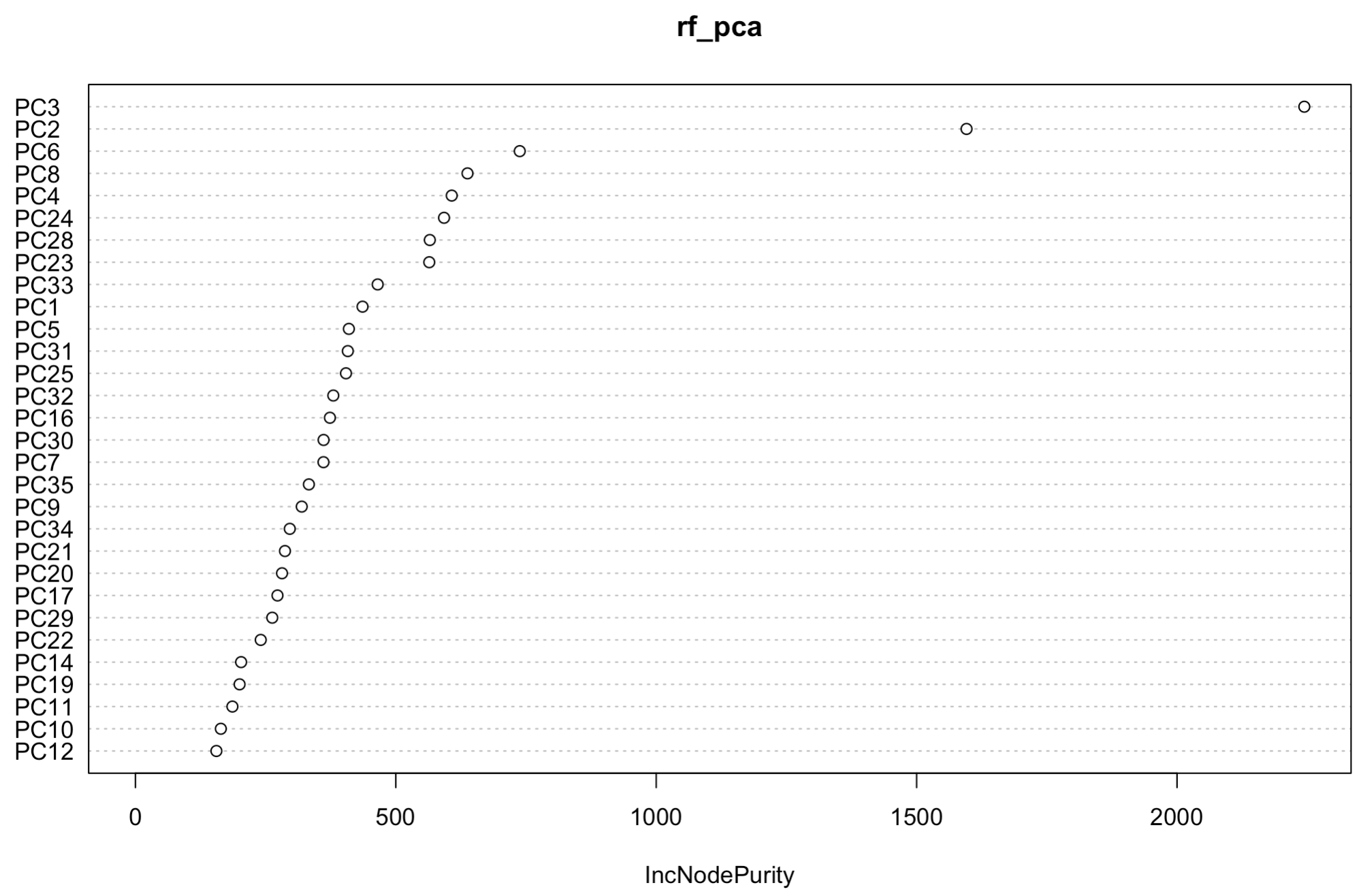
We trained two random forest models, where both models differ in their choice of predictors:   
(1) Random Forest with Regularized PCs and (2) Random Forest with Original Variables.

Using the OOB, we tune the optimal number of tree builds and the best number of randomly chosen predictors at every iteration.

| **Approach** | **Parameter** | **MAE Test set** | **MSE Test set** |
| --- | --- | --- | --- |
| **Random forest with PCA** | **ntree=700, mtry=11** | **1.55** | **5.38** |
| Random forest with Original variables | ntree = 500, mtry = 11 | 1.62 | 6.20 |

Table 3: Results of Random Forest models

We observe similar performance measures across the two approaches, but the random forest with PCA yields better results (Table 3). This shows signs of interest in using PCs as variables instead of the original one.

  
Figure 4: Variable importance by Random Forest models

In terms of variables used by the model, it's notable that the PCA approach heavily relies on sparse principal components 3, 2, and 6 compared to others (Figure 4). Interpreting these principal components involves examining the most important variables composing them, and this is made possible considering their significant regularization. In our analysis, PC3 pertains to potential pedestrian-vehicle interactions over 15-minute intervals, while PC2 relates to the average annual daily flow for pedestrians at intersections, and PC6 refers to measures to protect pedestrians, such as green straight arrows.

In contrast, the original variables encompass similar aspects, such as potential interactions between vehicles and pedestrians, daily flows of vehicles and pedestrians, and road width. However, the PCA approach effectively condenses this information into fewer components that contain the set of variables related to each other, facilitating a more streamlined interpretation.

### Boosting Algorithm - Tree vs Regression as a Learner

To enhance our analysis and gain deeper insights into the usefulness of using principal components as variables, we run two boosting algorithms, one with trees as learners and the other with a regression model. Assuming a Poisson distribution, table 4 shows the results.

| **Model** | **Family** | **Hyperparameters** | **MAE Test set** | **MSE Test set** | **Number of Variables selected** |
| --- | --- | --- | --- | --- | --- |
| Boosting - Treeswith PC’s | Poisson | n.trees=600, interaction.depth = 8, shrinkage =0.05, n.minobsinnode = 30 | 1.98 | 10.46 | 25 |
| Boosting - Trees with Original variables | Poisson | n.trees = 500, interaction.depth = 6, shrinkage = 0.01, n.minobsinnode = 30 | 1.99 | 10.57 | 46 |
| **Boosting - Regression  with PC’s** | **Poisson** | **147 iterations** | **1.91** | **10.45** | **26** |
| Boosting - Regression with Original variables | Poisson | 771 iterations | 1.97 | 10.6 | 46 |

Table 4: Results of Boosting models

Using cross-validation, we find the best set of parameters for the different models. The best model appears to be the boosting with simple linear regression as learners with the PCs.

Among the variables utilized by the best model, the top three in terms of relative importance are as follows: (1) The number of pedestrian-vehicle prohibited interactions (cti), with a relative importance of 28.16. (2) The number of pedestrian-vehicle left-turning potential interactions (cli), with a relative importance of 12.84. (3) The number of pedestrian-vehicle right-turning potential interactions (cri), with a relative importance of 7.61. (4) Notable features used are the average daily flow of vehicles and the distance of the intersections from downtown.

These models perform worse than any of the other methods, thus one could argue that the use of this set of variables doesn’t describe well enough the level of danger of an intersection.

# Key Characteristics of Hazardous Intersections

The variables consistently utilized in predictions across our best-performing method include metrics related to pedestrian-vehicle interactions during turning maneuvers (CLI, CRI), presence of pedestrian signalization such as green straight arrows (green\_stra) and all-red signal phasing (all\_red\_an), pedestrian traffic flow (pi), and specific neighbourhood indicators such as Ahuntsic-Cartierville, Ville Marie, and Saint-Laurent. We then proceeded to compare the statistics of the 25 most dangerous and safest intersections predicted by the random forest with principal components (PCs), our top-performing model, to discern differences between the safest and riskiest intersections. Notable variations in mean statistics of key variables flagged by our model were observed. For instance, variables capturing pedestrian and vehicular traffic volumes at intersections, coupled with distances from downtown, exhibited significant disparities. Additionally, features like tot\_crossw (sum of crosswalk widths along each approach) and tot\_road\_w (sum of road widths outside crosswalks along each approach) also showcased noticeable discrepancies. Riskier intersections tended to exhibit wider expanses, hinting at a potential correlation between larger road dimensions and heightened accident likelihood.

Additionally, we note that the GLM model with Poisson regression incorporates the variable `all\_red\_an`, which indicates the presence of an all-red pedestrian phasing. And rightly, when looking at a scatter plot of this variable, we observe an important difference in the frequency of accidents when the intersection lacks an all-red pedestrian phasing. We thus recommend that the city of Montreal treat such intersections lacking this feature. Another noteworthy characteristic contributing to intersection hazard is the presence of a semi-protected pedestrian phase, allowing pedestrians on all sides of an intersection to begin crossing before motor vehicles receive a green signal. Also, interestingly, our random forest model identifies the variable representing the presence of a green straight arrow as important. Through examination of a scatter plot, it becomes apparent that intersections without a green straight arrow are more susceptible to higher accident counts.

However, we are uncertain about the nature of this relationship. Is it possible that the presence of a green straight arrow confuses both drivers and pedestrians, leading to accidents due to a perceived right of way? Alternatively, could it be that green straight arrows are more commonly installed in inherently hazardous intersections, where there is a higher volume of pedestrian and vehicular traffic? We recommend that the city of Montreal investigates the underlying nature of this relationship to better understand its implications for intersection safety.

# Conclusion

In conclusion, our analysis demonstrates that the random forest with principal component analysis (PCA) emerges as the top-performing model. These results underscore the significance of employing principal components (PCs) instead of the original variables. Contrary to our initial assumptions, the sparse PCA method proves to be effective in interpreting and documenting the process by which our model generates predictions.

Using our best-performing model, we conducted predictions on the entire dataset and ordered the intersections from most dangerous to least dangerous. To allow us to rank based on the predicted number of accidents, we did not round the predicted values to get unique rankings, even though a continuous prediction doesn’t make sense for count data. Our goal is not to provide precise predictions of the number of accidents at each intersection. Instead, we aim to advise the city of Montreal on which intersections should be prioritized for attention and intervention.

Our findings underscore the complex interplay between traffic flow, road design, signalization, and urban settings in contributing to intersection hazard levels. These insights equip city planners with a data-driven hierarchy of intersections requiring targeted safety enhancements, thereby paving the way for informed infrastructural developments to bolster pedestrian safety. To enhance pedestrian safety at signalized intersections, the City of Montreal should reevaluate and potentially redesign intersection signalization, particularly concerning the use of green straight arrows and the implementation of exclusive pedestrian phases. Infrastructure improvements could include narrowing road widths and shortening crosswalk distances, especially in high-traffic neighbourhoods. The adoption of all-red signal phasing could act as a buffer to reduce accidents. It is also essential to continue data analysis to monitor the effectiveness of these interventions and to engage in public awareness campaigns that emphasize shared responsibility among road users.

1. Cover page and Table of Contents are not counted as a part of 8-page limit [↑](#footnote-ref-0)
2. The function “zeroinfl” of the package “pscl” does not allow us to use regularization directly as it does not have a separate parameter for it to perform regularization. [↑](#footnote-ref-1)
3. We give the variables selected by the poisson model to the zero inflated poisson model as the distribution remains the same i.e. poisson as otherwise with the original number of variables the model gives an error probably because of highly correlated variables and since “zeroinfl” does not perform regularization. [↑](#footnote-ref-2)