Prediction of the Car Accident Severity for the City of Seattle

Coursera Capstone Project

Silviya Ninova

September 29, 2020

Introduction

In its ”Global status report on road safety 2018”, the World Health Organization (WHO) states that road traffic injuries took the life of about 1.35 million people.[1] It is the main cause of death for young people in the age group 5-29 years and the 8th leading cause of death globally.[1] The emotional cost is high - the families of the victims must cope with the sudden loss their loved ones, while survivors often have to suffer the devastating impact lifelong injuries have on their physical and mental health, along with a possibly lowered financial status. At an economical level, this translates as up to 3% of the annual gross domestic product going for costs related to road accidents.[1] It is thus of vital importance on global and local scale to adopt a strategy for road-accident prevention, so as to save lives in the generations to come.

High-income countries are reported to have an overall decrease in the fatal road accidents, as compared to low- and middle-income countries.[1] The United States of America is the exception of the rule and shows the opposite trend. The fatality rate per 100,000 population did drop by 4.4 between the years 2005-2014. As of 2014, however, the numbers have been slowly crawling back up again with a peak in 2016.[2] Identifying thus the major factors leading to an increased accident severity is of great value towards road traffic injury prevention. Potential victims could be saved from a hospital stay or even long-term damage, which would also influence positively the economical costs associated with these.

In the present report, I am going to investigate the road accidents on a local scale, namely for the city of Seattle with the aim to identify features on which the accident severity depends and to what extent these can be used to predict the risk of a collision with injured people. Such knowledge is enormously useful to the local regulation authorities, who in line with the strategy for road traffic injury prevention, can take the necessary precautions and alert the local population for high-level risks of collision.

Data

The data set used for this report is readily available online on the Seattle GeoData[9]. It contains information on all collisions provided by the Seattle Police Department and recorded by Traffic Records Groups from 2014 until 2020. The original data contains 194673 records and 37 features. Not all features in the original dataset, however, provide relevant information for this project, so I started with a selection of pertinent columns.

The target in this model is the `severity` of a accident. In the present dataset it can adopt two values: 1 for *property damage only collisions* and 2 for *injury collisions*. There are 136485 cases of the first class and 58188 of the second, meaning that the dataset is imbalanced. This is a common problem for binary classification problems, where the goal is to predict the ”rare” event, whereas most algorithms aim only at minimizing the overall error rate.[5]

More details on how this problem is addressed in the present report can be found in Section Methodology/Machine Learning Model.

Features proven to be relevant in the prediction of accidents include the time/date of the accident, the weather conditions and the road (type, condition, etc.).[3, 10, 11] This informations is already present in the used database in several column values.

Several time related features were taken into consideration. The time of the accident (`hour`) is a relevant factor, since collisions prevail at certain hours of the day, *i.e.* rush hour. In addition, traffic is more intense during workdays, so `day-of-week` was included. Next, the month of the year is considered, so as to account for the changing seasons and datetime. Finally, the yearly evolution of the road accidents (`year`) is investigated within dataset range 2004 to 2020.

The original `weather` feature reports 10 different meteorological conditions. In an attempt to simplify the information, the columns was described by 6 classes - sunny, rain (including sleet/hail/freezing rain), overcast, snow, fog/smog/smoke and other, which comprises the cases of blowing sand/dirt and severe crosswind. A class labelled as ”unknown” was present. Since it was not noted, one could speculate that the weather was clear with no particular phenomena. Given however this uncertainty, records with this value will be removed from the dataset along with any NaN values.

Information on the road where an accident took place is stored in a variety of features. The road conditions are stored under`road-condition` and were grouped into the most frequently observed 4 categories - dry, wet (including standing water), Ice/Snow/Slush and Other (including the presence of oil, sand/mud/dirt). Such discrimination is important when considering the stopping way during an emergency braking for instance. Next comes the road illumination, which distinguishes between daylight, twilight (both dusk and dawn), dark with lights on, dark with no lights (or off) and other. An overall increased risk of accidents can be associated with night drives where the visibility is decreased and the light pollution in towns increased.

Data concerning the type of junction where a collision happened is stored both under the feature address-type and junction-type. The former contains only three types (block, alley, intersection), while the latter has much more detailed information on the precise location. Given that the `junction-type` has much more missing values, the work will continue only with the `address-type`. To give clarity its name was changed to `road-segment`.

In a next step, I look into how the number of participants in a collision correlates with its severity. The features include overall number of `persons pedestrians`, `cyclists` and `vehicles` involved in the collisions, as well as whether the right of way of pedestrians was not granted (`pedestrian-row`). In this regard,

In this regard, it is interesting to see what type of accidents prevail for one or the other class. This is already provided by the `collision-type` column, which has 10 categories. The SDOT-code and state-code offer much more detailed information on the exact type of collision divided into 36 (SDOT) or 62 (state) categories. Grouping the data would be more helpful for its interpretation, so only the `collision-type` column was kept for the modelling. The SDOTCOLNUM does not provide any useful information, so it is dropped as well.

Reckless driving can also be a factor for accident prediction. The extent to the collision is a result of high speed (`speeding`), inattention (`inattention`), under the influence of drugs or alcohol (`under-influence`) was also investigated.

The last feature regards information on the exact address of the accident, stored under `latitude` and `longitude`. This is useful for a map representation of the data, so as to understand whether some neighbourhoods or roads tend to be the scene of severe collisions more often than others. The precise address in `LOCATION` does not bring any additional information, so it is dropped from further analysis.

Finally, all records with ”not enough information” for the EXCEPTRSNCODE feature (5638) were dropped from the data set, due to missing important information. All categorical values in all features were transformed into numerical ones. Any other changes to the data will be discussed in the following section.

Methodology

This section summarizes the data analysis performed on the dataset. In the first subsection, the exploratory analysis is presented, whereas the second subsection looks into the machine learning algorithms used in the present report.

Exploratory Data Analysis

While getting familiar with the data at hand, I try to understand to what extent the specific features can help with the discrimination between collisions with property damage or injury.

Time period

The time of the day is an indicator for the traffic flow and thus probability of an accident. Indeed, the majority of the collisions appear to happen at rush hour, reaching its maxima at about 8 and 17, when the frustration and traffic levels are high (see Figure 1 (left)). This is in line with the findings of other works.[10, 12] The general trends are maintained for the two target classes, with injury collisions proportionately more numerous at peak hours and less so over night. Given the similarity of the two curves, I divided the time of day into four periods, where one collision severity class is more predominant over the other (see Figure 1 (left)). It is important to stress out that the comparison is made within each class, since the dataset is imbalanced.

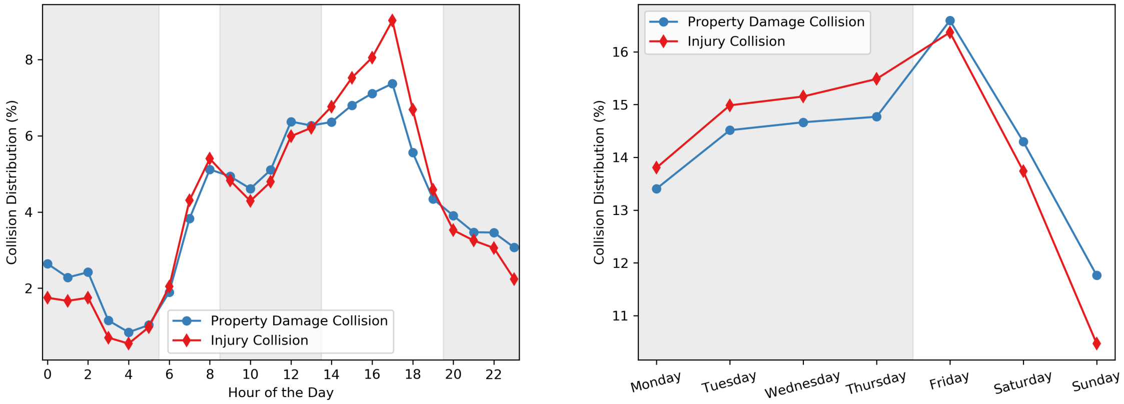


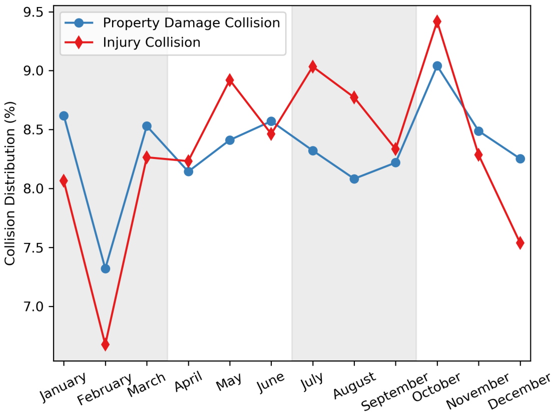
Figure 1 Proportion of collisions with property damage or injuries at each hour of the day (left) and day of the week (right). The chosen time windows are represented in alternating grey and white.

As far as the day of the week is concerned, the collisions of both classes happen on mostly on Friday and the least - on Sunday (see Figure 1). Monday and Saturday share similar levels, which are overall lower than those of Tuesday through Thursday. It is noticeable that when the normalized values of the two classes are considered, there are more accidents with injury in the first half of the week, whereas towards the weekend the percentage of property-damage collisions is higher. This feature adopts two values - for grouped days Monday-Thursday and Friday-Sunday.

Another time related factor is the month of the year. It is directly related to an extended daylight or dark periods together with challenging weather conditions. It is interesting to observe that an increased number of injury collisions happen between April and October (see

Figure 2) with a lower percentage in the winter months in comparison to the other severity class. This could be explained with the presence of more traffic participants with increased vulnerability, such as pedestrian and cyclists, during the hot months. In order to increase the distinction between the two target classes, the feature `month` was thus replaced with the feature `cold-months`, which distinguishes two classes - 0 for an accident in April-October and 1 for December-March.

Figure 2 Proportion of collisions with property damage or injuries in each month. The chosen time windows are represented in alternating grey and white.



Finally, we look at the evolution of collisions with the years 2004 to 2020. 2020 is however still ongoing and covers data only until April. This accounts for the artificial drop in number of accidents (see Figure 3 (left)). In addition, the last months coincide with the COVID-19 emergency and lockdown, so fewer cars than normal were in the traffic and thus fewer accidents were reported. Since the aim is to predict the severity of collisions under normal circumstances, no accidents from 2020 were considered in this report.

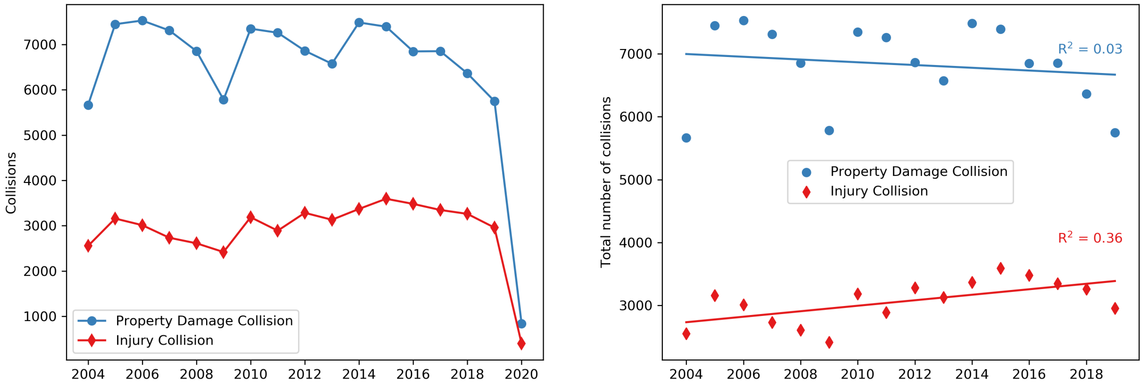


Figure 3 Total number of collisions for the period 2004-2020 (left); linear regression on the total annual numbers for each severity class (right).

The number of accidents over the remaining years has been varying (see Figure 3 (right)). As it can be seen in the graph above, several drops have been observed (2004, 2009, 2013, 2019) with the number of accidents going back up again in the following year, suggesting a cyclic pattern. Fitting a line over the collisions yield very poor results - the coefficient of determination (R^2) is close to 0 for property-damage accidents and at 0.36 for collisions with injuries. This is an indication that the number of accidents do not evolve linearly over the years. Using a polynomial fit, however, may cause overfitting of the curves without previous knowledge about the factors leading to this. It is difficult thus to draw any obvious conclusion.

Weather and road conditions

Meteorological conditions are expect to increase the probability of an accident. Heavy rain and fog reduces the visibility, standing water and high speed can lead to aquaplaning. Predicting the accident severity on the weather conditions is, however, not so straightforward. As it is clear from Figure 4 (left), both property-damage and injury collisions are most probable to occur when it is sunny, overcast or raining. In comparison, all other conditions, i.e. snow/fog, are much less pronounced. The difference in proportion between the two classes is however below 1%.

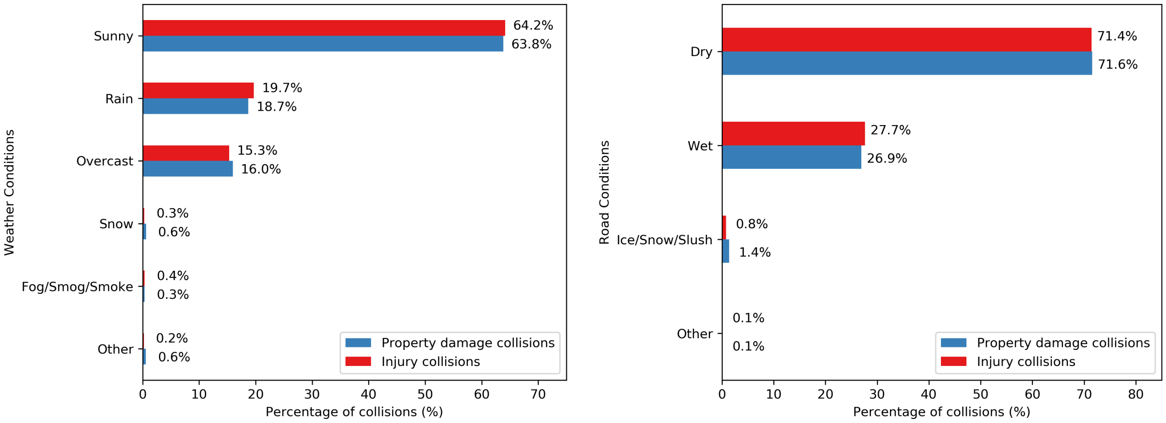


Figure 4 Proportion of collisions with property damage or injuries under different weather (left) and road (right) conditions.

The weather conditions affect also the road conditions. Wet roads, for instance, can cause a longer stopping distance, while the presence of oil or ice can easily lead to skidding. Similarly to the weather conditions, we do not observe any significant difference in terms of collision severity. Both classes have most accidents at dry or wet roads and to a much lower extent on ice (see Figure 4 (right)).

Street illumination

Similar situation to the weather and road conditions is also observed for the street illumination. Almost two thirds of the accidents occur during daytime and less than one third when it is dark and the street lights are on (see Figure 5). Next come collisions at twilight, which is not surprising, given that the majority of accidents actually happen late afternoon, early evening, when people come back from work. Interestingly, there are slightly more injury collisions during the day and twilight, as opposed to when it is dark. There is however no significant differences between the two severity classes in terms of street illumination.

Road segment

The situation changes when the severity classes are compared against the type of junction where the accident took place (see Figure 6). While almost two thirds of the collisions with property damage happen along a block, the collisions with injury are almost evenly split between block and intersection. The injury collisions are thus much more probable to take place at an intersection, which makes this a promising feature for the prediction analysis.

Figure 5 Proportion of collisions with property damage or injuries under different street illumination.

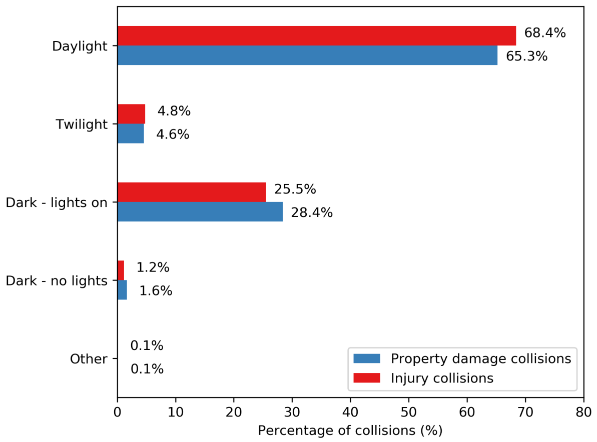
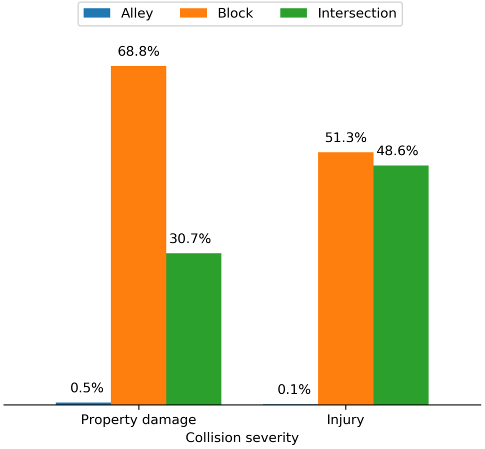


Figure 6 Number of collisions (in %) with respect to different junction types for the two collision severity classes.



Participants in the collision

The number and type of participants in a collision is another feature to investigate. Property-damage accidents tend to mostly involve up to two people or vehicles. Injury collisions are predominant for one-vehicle accidents and multiple-participants. When cyclists and pedestrians are involved, a collision with a higher severity level is more pronounced in percentage. These features indeed provide a certain level of discrimination between the two target classes, so they will be certainly useful during the supervised learning.

Most accidents happen with up to 10 people. The highest number of people involved in a collision with property damage is 57, whereas for injury - it is 81. The maximum number of pedestrians involved in an accident of class 1 is 3, whereas for class 2 - it is 6. The vast majority of damage collisions do not involve a pedestrian, whereas 10.5% of the injury collisions involve a pedestrian. Given that most of the collisions involve either 1 or 0 pedestrians, the feature is turned into a classification of whether pedestrians are involved or not.

The situation with the cyclists is similar to that with pedestrians. The majority of recorded accidents happen with no cyclists. This feature is therefore also turned into a binary one, indicating the participation of a cyclist or not.

Having a high number of vehicles in a collision is rare. Most of the accidents happen with two vehicles. Such crashes are mostly with property damages and to a lesser extent with injuries. When only one vehicle is involved, it is to be expected that most accidents will be with injuries. This can be ascribed to accidents involving collisions with pedestrians or cyclists.

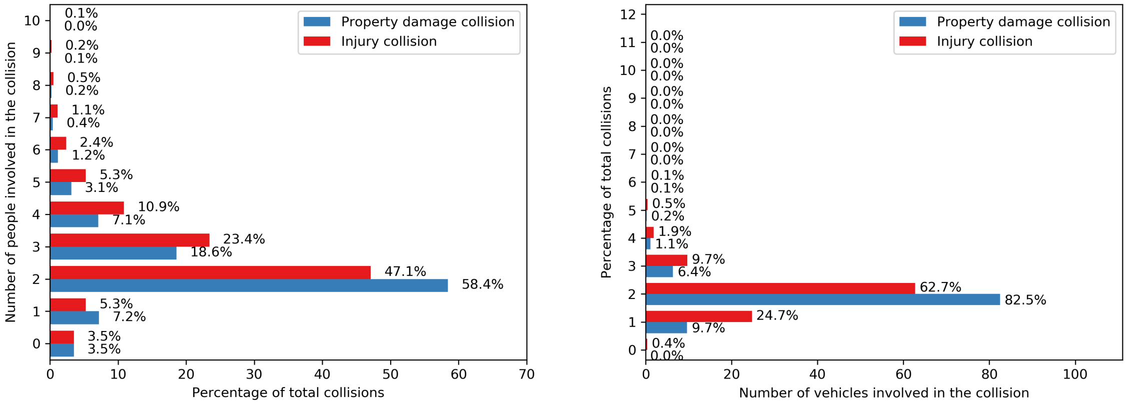


Figure 7 Number of people (left) and vehicles (right) participating in the collisions for the two severity classes.

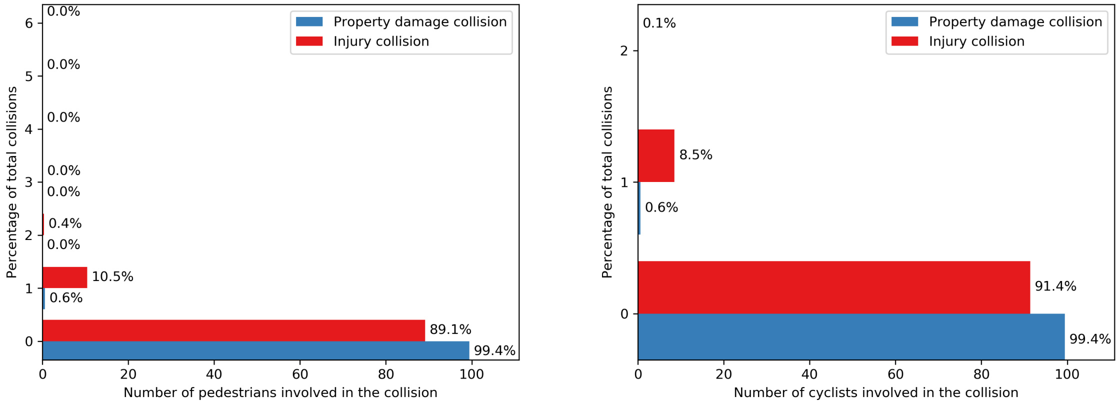


Figure 8 Number of pedestrians (left) and cyclists (right) participating in the collisions for the two severity classes.

Accident code

The accident codes contains information about the type of collision reported. While the code describe how the crash happened, this feature gives an insight into what what types of accidents should be prevented, so as to reduce the severity of accidents. Crashes with pedestrians or cyclists, for instance, are much more likely to involve injuries (see Figure 9 (left)). In addition, about one-fourth of all injury collisions are rear-ended crashes, which often includes spinal injuries, and one-fourth happen at an angle.

By comparison, the most prevalent collision type with property damage is actually going against a parked car. This can partially explain the reason for the proportionately high level of accidents on the block (see Figure 6). Next come angles, rear-ended and sideswipe collisions. Head-on or right-turn accidents appear to be rare overall, whereas turning left more often leads to crashes.

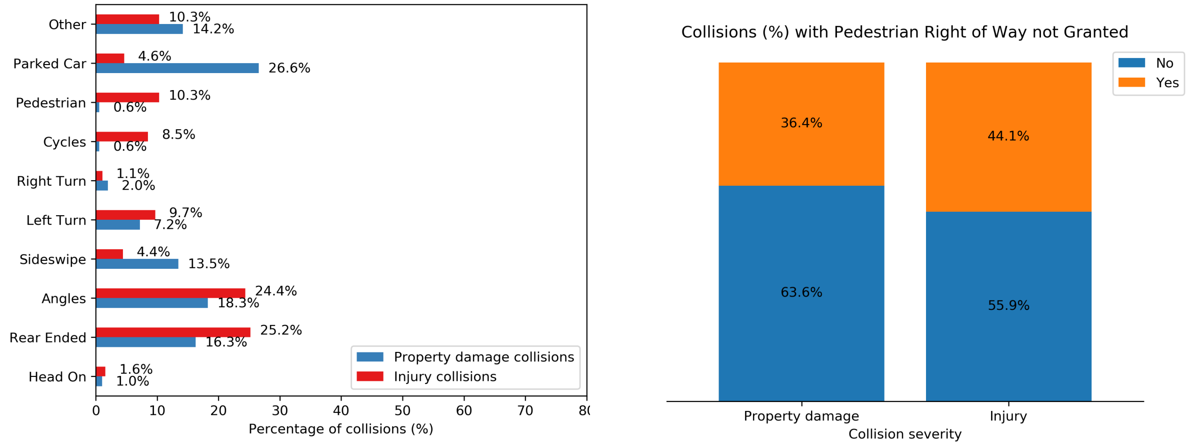


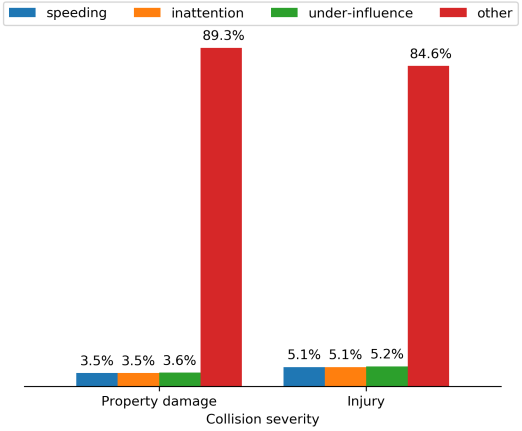
Figure 9 Proportion of collisions with property damage or injuries for the different collision types (left); Proportion of collisions with pedestrians, where the right of way was not granted (right).

Moreover, from the previous graph we already know that only 10.3% (injury collisions) and 0.6% (property collisions) of all accidents involve pedestrians. It is found that less than a half of the accidents resulting in injury, happen because the pedestrian right of way was not granted, the number being one third for the property damage collision.

Reckless driving

Other factors of relative importance are the driver's inattention, their state under the influence of drugs or alcohol, and driving at higher speed than permitted. The results are summarized in Figure 10. Fewer accidents happen as a result of such reckless driving and the overall numbers remain almost for these three features. In comparison, however, the proportion is slightly higher for collisions with injury (2%).

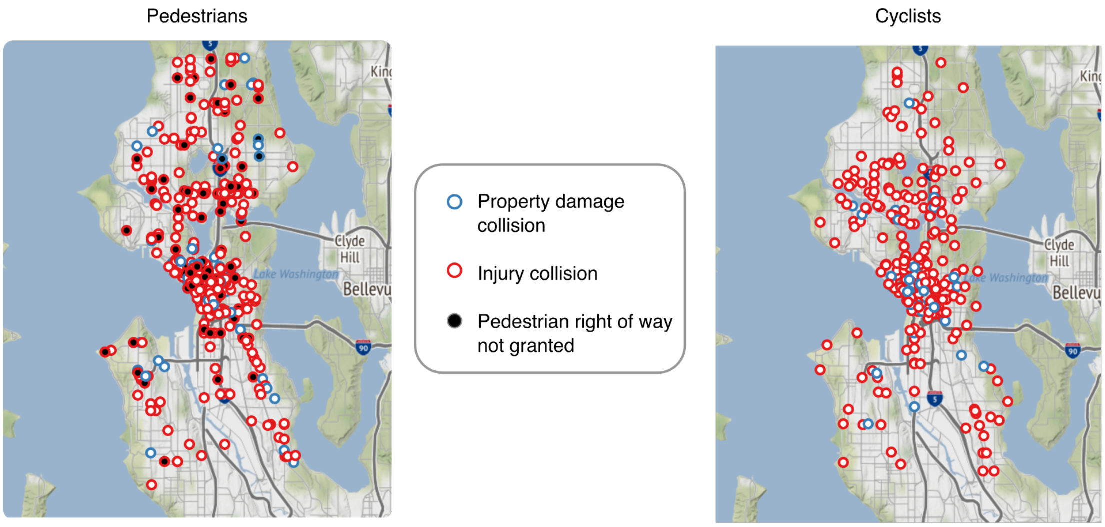
Figure 10 Number of collisions (in %) with respect to speeding, inattention and driving under the influence of alcohol or drugs for the two collision severity classes.



Collision location

The last feature considered for this supervised learning model is the collision location, represented by its latitude and longitude. Figure 11 contains only the records from 2019 where pedestrians or cyclists were involved in the accident. It is noticeable that most of the accidents happen in the city center and less in the northern/southern neighbourhoods. Interestingly, the northern part contains much more collisions with injuries for cyclists and pedestrians, a high proportion of which are as a result of their ignored right of way. It is difficult to comment on the reason for the accidents distributions without looking at cycling-path availability. This, however, goes beyond the scope of the present work.

Figure 11 Collisions with pedestrians (left) and cyclists (right) for the year 2019 by their location in the city of Seattle.



Machine Learning Model

The problem at hand is a classification one - namely to determine if a collision with property damage or injury will happen given certain conditions. A popular choice for similar studies on road accident prediction[3, 7, 10, 12] is the tree classification, which will be applied also for the current report.

There are several variations of the tree classification machine-learning algorithms, such as the Random Forest[4], Balanced Random Forest[5] and Gradient Boosting[6, 8]. The last two are especially valuable when working with imbalanced data sets, such as the present case. Indeed, there are twice as many cases of class one (property damage collisions) than class two (injury collisions). Workarounds to this problem are random sampling of the majority class, so as to obtain a balanced data set, or a cost-based approach, where additional weight is added to the instances of the minority class.

In the present report, the supervised learning is carried out using three methods - the Decision Tree (DT), the Random Forest (RF) and the Extreme Gradient Boosting (XGBoost). DT is based on the construction of one decision tree for the prediction of the target. While easy to understand, this prediction model is prone to overfitting. The RF corrects this tendency by introducing several decision trees (ensemble) and make decisions based on the average predictions. The third method, XGBoost, also uses an ensemble of decision trees. These are added step-wise and aim at minimizing the loss function.

The two classes were balanced within the cost-based approach for all three methods. 70% of the data was used for the training and 30% for the testing. Given the high number of hyperparameter coming with each method, their values were optimized using a randomized search (RandomizedSearchCV) as implemented in the `sklearn` package.

Results

Model Performances

The performance evaluation of the three models is done with the Receiver operating characteristic (ROC) and the Precision-Recall (PR) curves (see Figure 12). These metrics are often used for studies of imbalanced data sets, since the focus is on the classification efficiency for the positive class (rare event), which in our case is the injury collisions.

All three methods give a very similar performance, as can be seen from their ROC and PR curves (see Figure 12). Their efficiency is rather low, where the DT performs the worst and XGBoost marginally the best (see Table 1).

Figure 12 ROC (left) and PR (right) curves of the three methods.

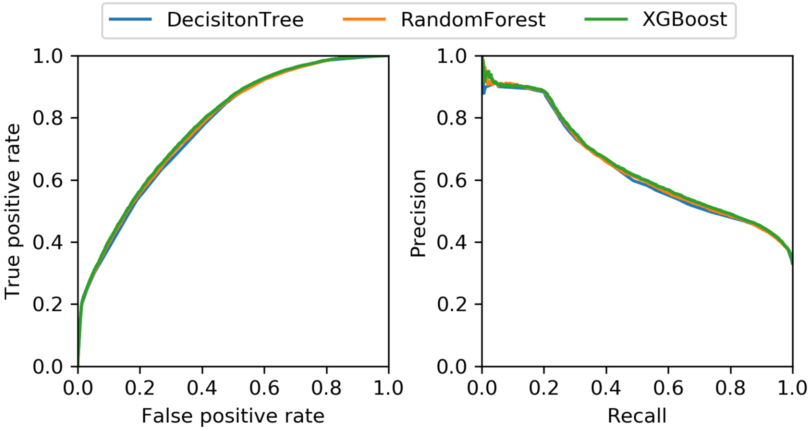


Table 1 Area under the ROC results for the three methods used.

	DT	RF	XGBoost
Area under ROC	0.68	0.69	0.70

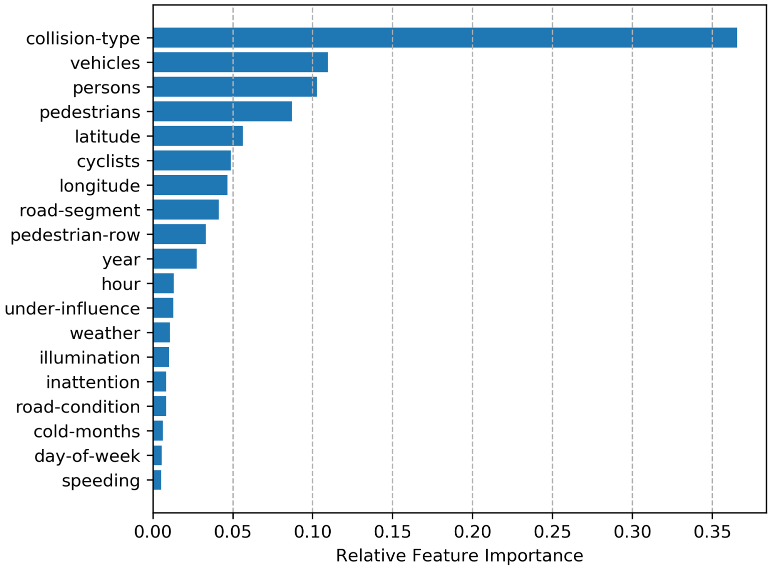
Feature Importance

Originally, 19 features were considered for the model. Their importance for the classification was estimated using their impurity decrease within the Random Forest method (see Figure 13). The most highly ranked feature is the collision type, followed by features regarding the participants in the accident, such as number of vehicles, people and pedestrians. This is not surprising, considering that certain accidents are much more likely to inflict injuries and these involve pedestrians for instance.

In addition, in terms of geographical location, it is found that the latitude is somewhat more important than the longitude. This is ascribed to the elongated form of the city where most of the accidents happen in its central part and to a lesser extent north or south of it. It is thus to be assumed that the importance of these two features is rather problem specific and possibly not transferable to other similar models.

Finally, below 3% importance can be found all features regarding reckless driving and time period. Surprisingly and with low importance are also the weather, road and light conditions. These features were already reported as important for the accident prediction[10] and injury severity prediction[3, 7]. They seem, however, to play a less crucial role as far as the present study is concerned.

Figure 13 Relative feature importance as calculated with the Random Forest Classifier for all features considered.



Discussion

The prediction model reports a maximal score of about 0.70 (ROC-AUC) (see Table 1). While this result is good, it does need to be improved. Indeed, the false positive rate is rather high for high recall, so many injury accidents are missed when testing and this is not an optimal solution. On a positive note, higher recall values also have a rather high precision. High true positive rate is in fact beneficial, being able to capture also collisions with an increased risk of injuries.

The most important feature is determined to be the collision type. While this category offers only information in hindsight, it provides also guidelines to what accidents are to be prevented, so as to decrease their severity.

The most frequent accidents involve rear-end or car crashes at an angle. In this view, it is recommended to supervise the observance of safety distance while driving on the most affected roads, so as to avoid rear-ended accidents and spine injuries. Road signs indicating the right of way should be checked for visibility in order to decrease crashes at angles.

Furthermore, it is suggested to work towards more safer roads for pedestrians and cyclists, a collision with which almost always leads to an injury. The state, illumination and visibility of the zebra crossings and cycling paths are to be controlled and if necessary improved.

Conclusions

In conclusion, the present report investigates the predictability of injury collisions as opposed to property-damage ones for the town of Seattle. It is found that the severity of the accident strongly depends on the collisions type. Among the most prone to inflict injury are rear-ended or angled car crashes and accidents with pedestrians and cyclists. The extreme gradient boosting method yielded the highest score of 0.7 for the area of the ROC curve.

Future work includes improving the predictability of the model by considering additional features. As far as the two-motor vehicle crashes are concerned, it is interesting to see the extent to which they depend on the traffic congestion, measured as velocity needed to leave a certain critical road segment. As far as the accidents with cyclists are concerned, it can be useful to explore the correlation between injury collisions, availability of cycling paths and their visibility to the other participants in the traffic.

Bibliography

[1] Global status report on road safety 2018. Technical report, World Health Organization, 2018.

[2] National highway traffic safety administration. Technical report, United States Department of Transportation, 2018. [Online; accessed 15-September-2020].

[3] Joaquín Abellán, Griselda López, and Juan de Oña. Analysis of traffic accident severity using decision rules via decision trees. *Expert Systems with Applications*, 40(15):6047 – 6054, 2013.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] C Chen and L Breiman. Using random forest to learn imbalanced data. University of California, Berkeley, 2004.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[7] Miao M. Chong, A. Abraham, and M. Paprzycki. Traffic accident analysis using machine learning paradigms. *Informatica (Slovenia)*, 29:89–98, 2005.

[8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics*, 28(2):337–407, 04 2000.

[9] Seattle GeoData. Collisions. `https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions?geometry=-123.320%2C47.452%2C-121.342%2C47.776`, 2018. [Online; accessed 16-September-2020].

[10] Antoine Hébert, Timothée Guédon, Tristan Glatard, and Brigitte Jaumard. High-resolution road vehicle collision prediction for the city of montreal. In *2019 IEEE International Conference on Big Data*, pages 1804–1813, 2019.

[11] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '19, page 33–42, New York, NY, USA, 2019. Association for Computing Machinery.

[12] Daniel Wilson. Using machine learning to predict car accident risk. Technical report, 2018.