Predicting the severity of vehicle collisions based on environmental factors – a case study

This case study is the Applied Data Science Capstone Project in the IBM Data Science Professional Program. This study does not only serve as a demonstration of the skills that I have acquired, hopefully it can also demonstrate the power and versatility of common data science methods. The audience of this article are not necessarily just scientist and researchers but the general public and especially those with an interest in the applications of data science or those concerned about road safety.

All the code as well as the used data set can be found on GitHub
https://github.com/dcadosch/Coursera_Capstone

1. Introduction

At least since we use wheeled motor vehicles for transportation, traffic collisions are an unfortunate but inherent aspect of it. The severity of accidents varies for a number of reasons. Roads may not be designed in a safe way or crucial parts of vehicles may fail while driving due to insufficient maintenance or due to engineering faults of the vehicle manufacturer. The driver may be distracted or violate safety rules and laws by speeding or driving under the influence of alcohol or drugs. There are also environmental factors that may influence the severity of a collision. A law-abiding driver who takes appropriate care of her or his vehicle can only influence a few of these factors that govern the outcome of a collision. However, it is difficult to judge these factors without a thorough analysis of the data of past collision events.

In this case study I am creating a machine learning model that estimates the severity of collision events based on a number of environmental factors. The severity of collisions is divided into multiple classes, which means that we need to employ a classification algorithm. In order to be able to choose an optimal classification algorithm, we will deploy multiple methods and compare their performance by measuring several evaluation metrics.

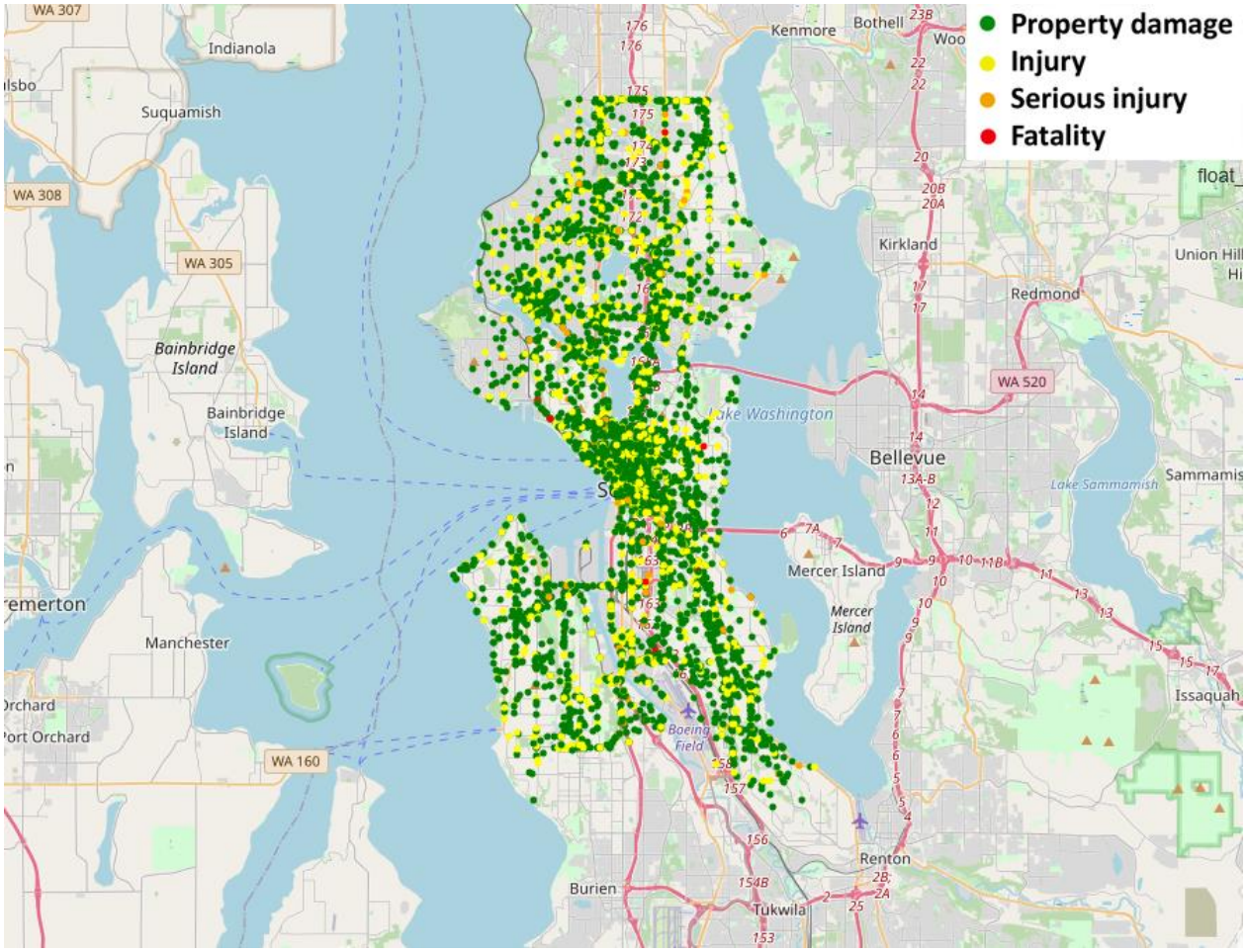
This case study will focus on traffic accidents in an urban environment in the United States of America. Hence, it might not be applicable to situations that differ considerably from such a scenario. The general approach of this study might nonetheless serve as a guide to create a similar model with data from another source.

2. Data understanding

The data to train and evaluate the model was obtained from the 'City of Seattle Open Data Portal' (https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d). The data set contains all reported vehicle collisions in Seattle from January 1, 2004 to October 9, 2020. The table contains 40 columns and 221,525 incidents. Not all of the columns contain relevant or usable information for the task at hand. Furthermore, not all incidents have a complete set of information. These incidents may need to be removed from the data set.

The most relevant feature is the severity of an accident. In the original data set the severity is divided into four classes: property damage, injury, serious injury, and fatality. The model will be trained to predict the severity class based on other available information. Independent features that a driver may observe before a collision occurs are for example the road and light conditions, the weather and the kind of road (s)he is in. All these features are divided into about ten classes each. Other features such as the month of the year or the time of the day may be extracted from the timestamp in the data set. Collision data that can only be obtained after the collision occurred such as the type of collision or whether any person involved was driving under influence will not be used in the training of the classification algorithms.

To get an initial overview of the nature of the data that will be used it might be helpful to have a geographic visualization. The following map shows all recorded collision incidents in 2020 up to October 9 for which the coordinates were available. The severity of incidents is divided into four categories and indicated by their color. From this map we can see the boundaries of the geographic area from which the data stems. We can further see that the incidents are not homogeneously distributed. There are certain areas where incidents are clustered. That clustering might be due to higher amounts of traffic, particularly dangerous roads or simply due to stochastic effects. The severity of accidents shows no obvious pattern that would suggest that certain areas cause more serious accidents. Since the geographic distribution of accidents and their corresponding severity does not seem to provide any additional information, we will not use the location data any further in our study.

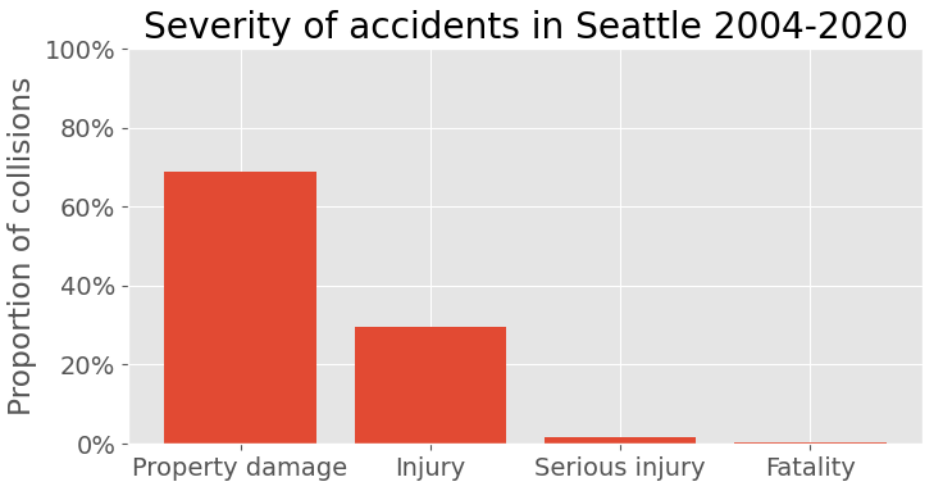


3. Methodology

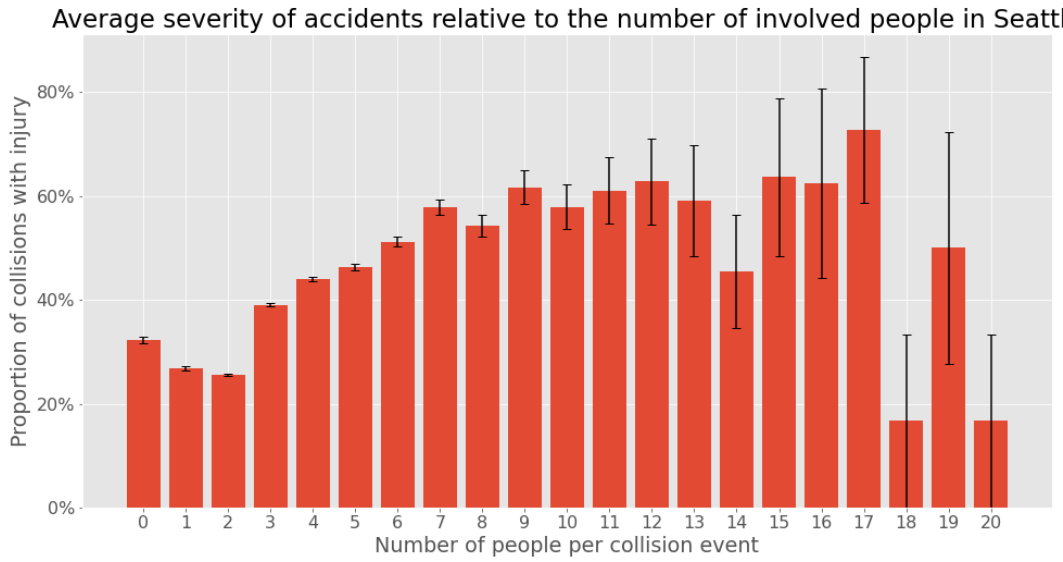
3.1 Data cleaning and examination

Before we analyze the data, it is appropriate to take a more in-depth look at the data. Some aspects and features might need to be cleaned or processed to be able to get a better understanding and also to increase the quality of the predictions that the model is supposed to make. First, we look at severity code labels for every incident. According to the metadata of the data set the labels are '1', '2', '2b', and '3', which correspond to 'property damage', *'Injury'*, *'Serious injury'*, and *'Fatality'*, respectively. However, the data set also contains 21,616 incidents without any label or with the label *'0'* which does not correspond to any severity class. These incidents cannot be used for the training or evaluation of the model and have to be removed from the data set.

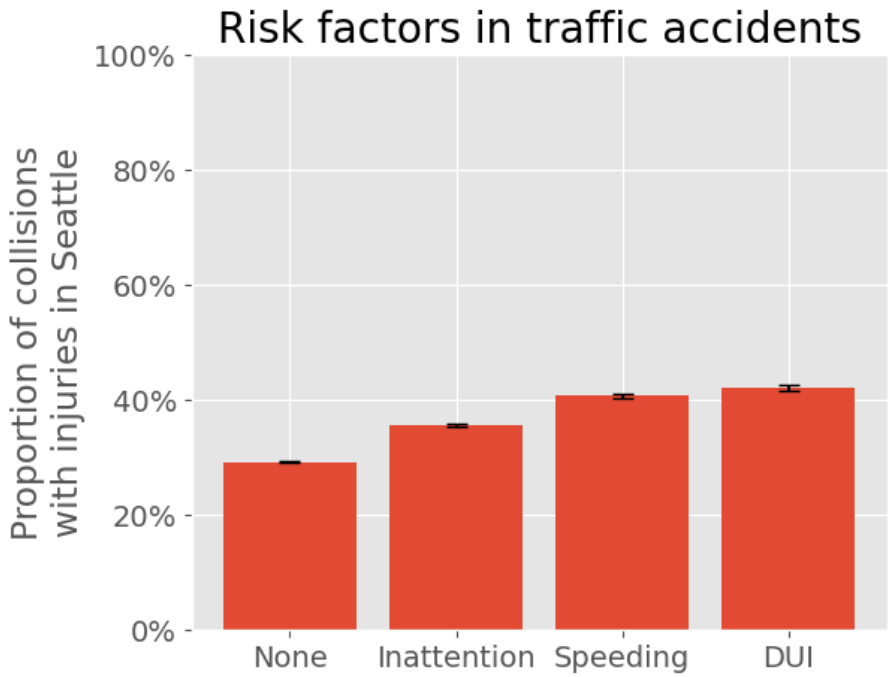
Next, we look at the distribution of the remaining incidents according to their severity. Unsurprisingly we see that that the majority of all collisions (68.9%) only cause property damage. 29.4% of all collisions resulted in (light) injury but only a very small percentage involved serious injury or fatality (1.6% and 0.2%). Of course, it is very fortunate that only a small proportion of accidents result in a serious outcome. On the other hand, this also means that the distribution among classes is very imbalanced. Furthermore, the different classes are not necessarily very distinct from each other but rather subdivisions on a gradient of collision severity. Thus, it is unlikely that we can train a reasonably accurate machine learning model for all four classes. This means that we should simplify the categorization. We do this by grouping incidents in only two classes - collisions that only caused property damage and collisions that have caused physical harm. This will still result in an imbalanced data set with a ratio of about 3 to 7 but this is something that we can handle and every class still has a reasonably high number of incidents to inform the model.



The data set also contains information about the number of people (pedestrians, cyclists, people in vehicles) who were involved in an accident. The number of people who are involved in a collision event correlates with the severity of that collision because the probability of at least one person being (seriously) injured or dead increases with the number of people. Thus, we will not use these features directly. However, the information whether at least one pedestrian, cyclist or vehicle driver was involved in a traffic accident could give us some information about the risk that those three groups face in an accident. So we will create three new binary features that indicate whether at least one pedestrian, cyclist, or vehicle driver was involved in an accident.



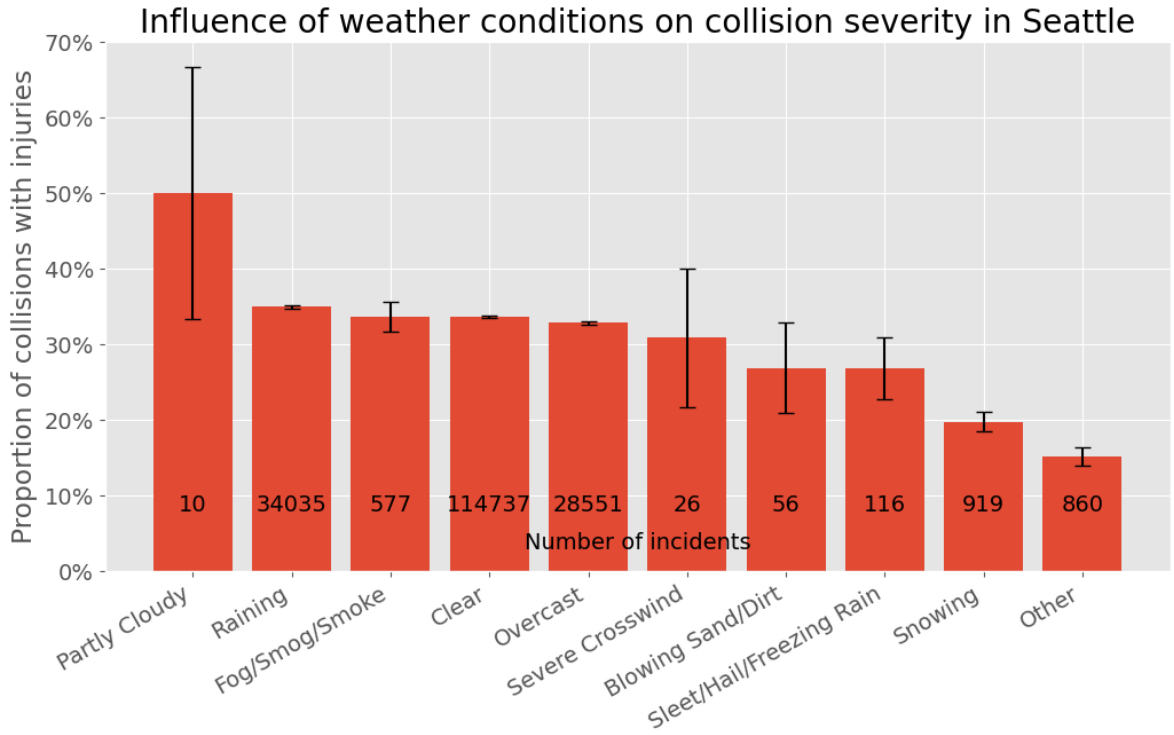
The data set contains information about whether the collision happened due to inattention, whether a driver was under the influence of drugs or alcohol (DUI), or whether speeding was a factor. This information is not suitable to be used to inform the model since this can usually only be determined retrospectively. Nonetheless it could be interesting to perform a quick analysis to assess how much these factors influence the severity of a collision.



The bar chart reveals that inattention, speeding and driving under influence (DUI) substantially increase the risk of a collision that involves injuries. Inattention increases the risk by 22.2%, speeding by 39.4%, and DUI by 44.6%.

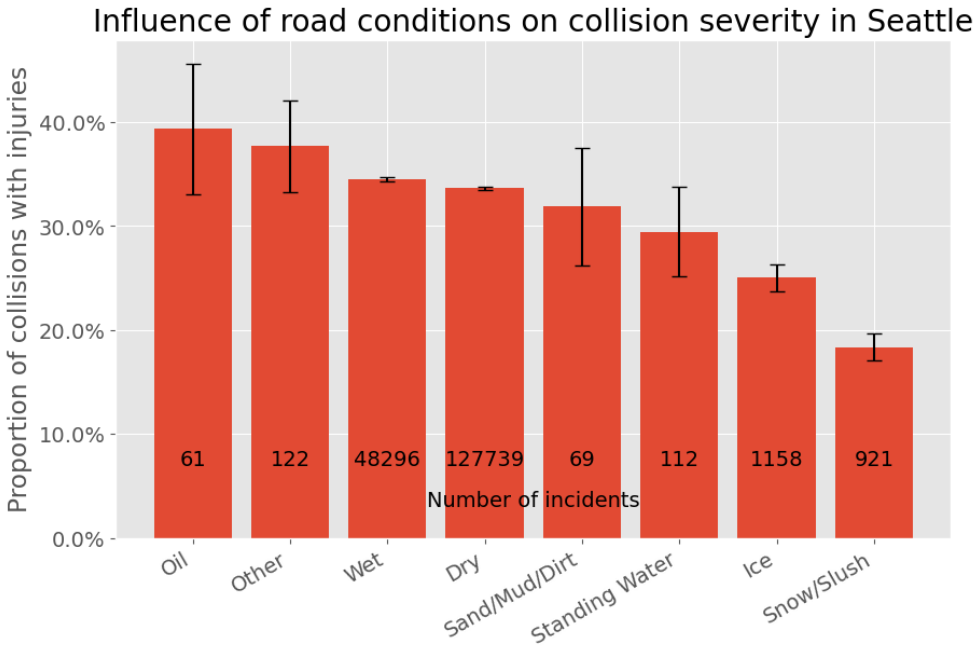
Among the features that are definitely expected to be informative for the model are the weather, the road conditions, and the light conditions. All three features categorize the conditions in a few classes. For the evaluation of these features it would be helpful to visualize their influence on the severity of collision events.

The *'Weather'* feature has 12 different classes. The class *'Blowing Snow'* class has only one incident with this designation and will be dismissed due to its insignificance. Furthermore, the class *'Unknown'* has 15,131 incidents. Since this class has no informative value the corresponding incidents will also be removed from the data set.



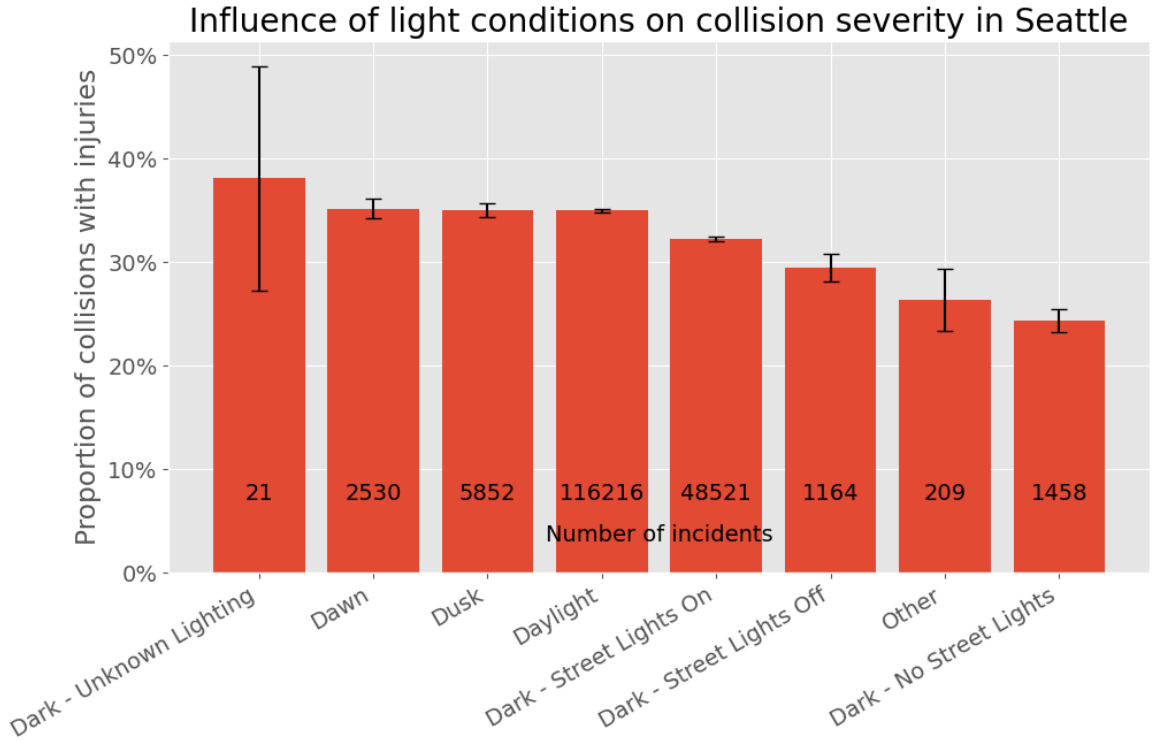
The differences between the 'weather' classes are in general not very large. Only 'Snowing' and 'Other' display a substantially lower risk for collisions with injuries. It is likely that traffic is slower and people might be more attentive when it is snowing, which could explain the lower risk. One also needs to consider that the overwhelming majority of accidents occur when it is either *'Clear'*, *'Overcast'* or *'Raining'*. Other weather conditions are very rarely reported.

Road conditions are subdivided into nine classes. One class is again *'Unknown'* and will be removed. That reduces the data set by another 1,502 incidents.



Similarly to the weather conditions we can observe that most road conditions do not have a large influence on the probability of a severe collision. In accordance with the previous observation we can see that *'Ice'* and *'Snow/Slush'* are clearly displaying a lower risk for an accident with injury. However, we also need to note that most accidents are reported to have occurred while the road was either dry or wet.

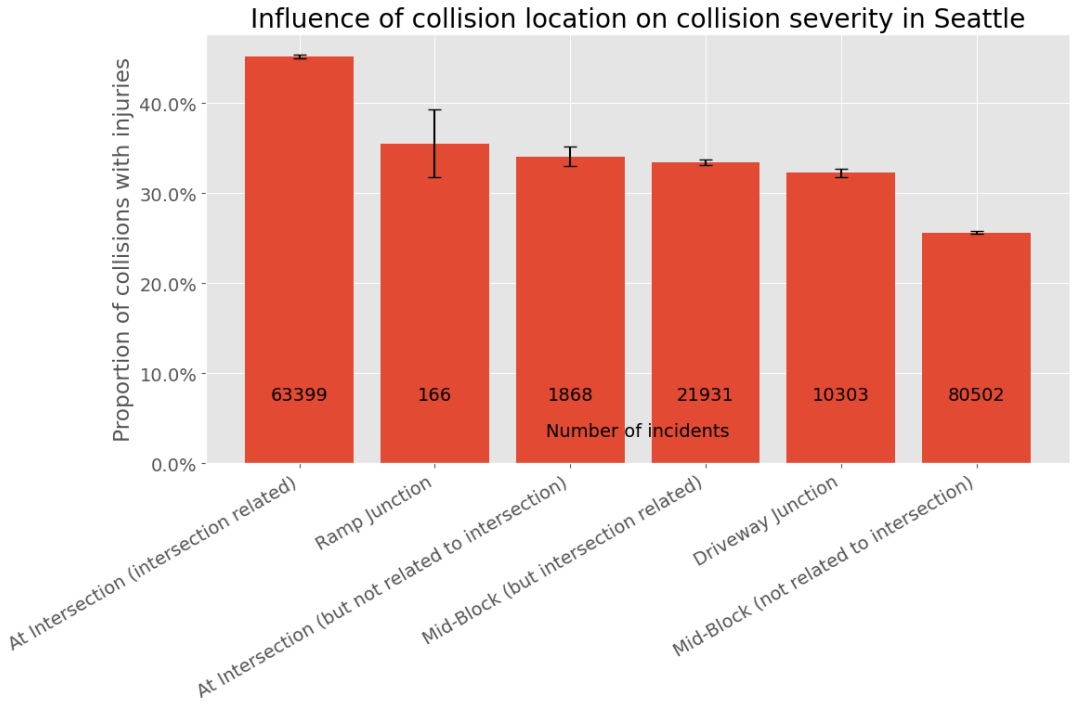
The last environmental feature is the light condition. By removing the 'Unknown' class we lose another 2,374 incidents from our data set.



Interestingly, adverse light conditions seem to negatively correlate with the severity of collision events. This further supports the hypothesis that difficult conditions increase the attention of drivers and promote more cautious driving. Similar to the other environmental features we need to keep in mind

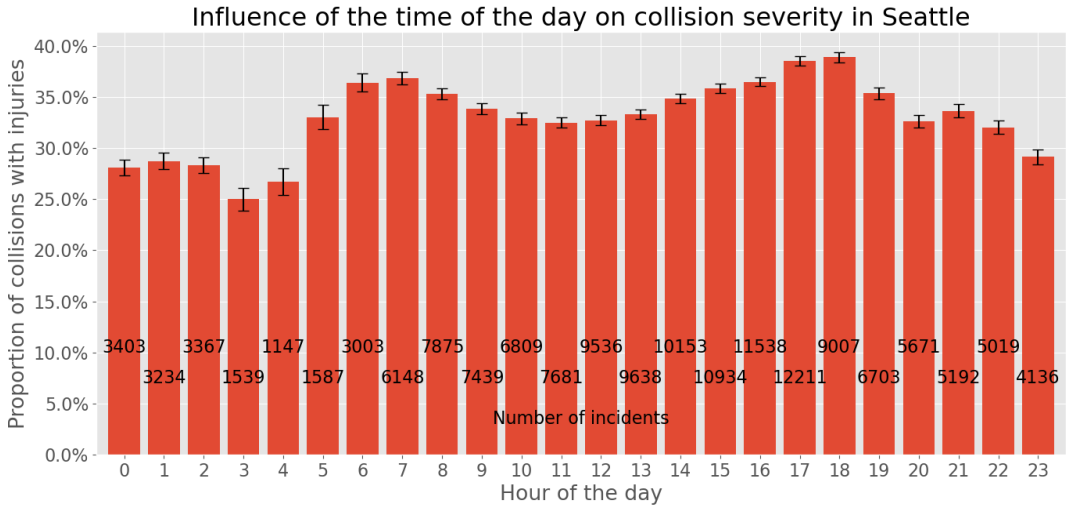
that the distribution of collision events is not evenly distributed between classes at all. Almost all accidents are reported to occur at daylight or in the dark with the street lights on.

Even though we decided not to use the exact location of incidents due to the apparent lack of informative value there is some categorical information that we can use. One feature in the data set provides information about the junction type where the collision occurred. This could inform a driver and our classification algorithm about locations where the risk of a severe accident is higher or lower. We dismiss again the incidents that are labeled *'Unknown'* which discards 8 rows.



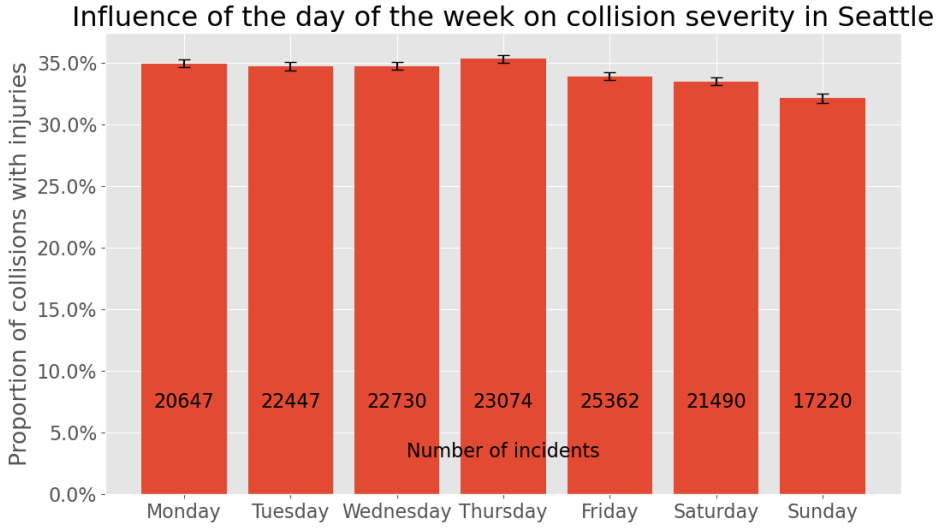
This location feature might have a higher informative value than the previous environmental features. We can clearly observe that collision that occur at intersections are substantially more severe than at any other location. Accidents that occur mid-block tend to be less severe than the rest. The other locations are somewhere in between and vary only mildly. A further contrast to the previous features is the more equal distribution of incidents between the classes. Most incidents are actually assigned to the two most extreme classes in terms of their influence on risk.

Besides environmental and location factors, we also have information about the time of collision events. From the timestamp of a reported incident we can extract categorical features that might improve the performance of our prediction algorithm. Unfortunately, there are many incidents for which the exact time is not available. When we discard these incidents, we lose 27,923 rows. The first aspect we investigate is to what degree the hour of the day may influence the risk of a severe collision.



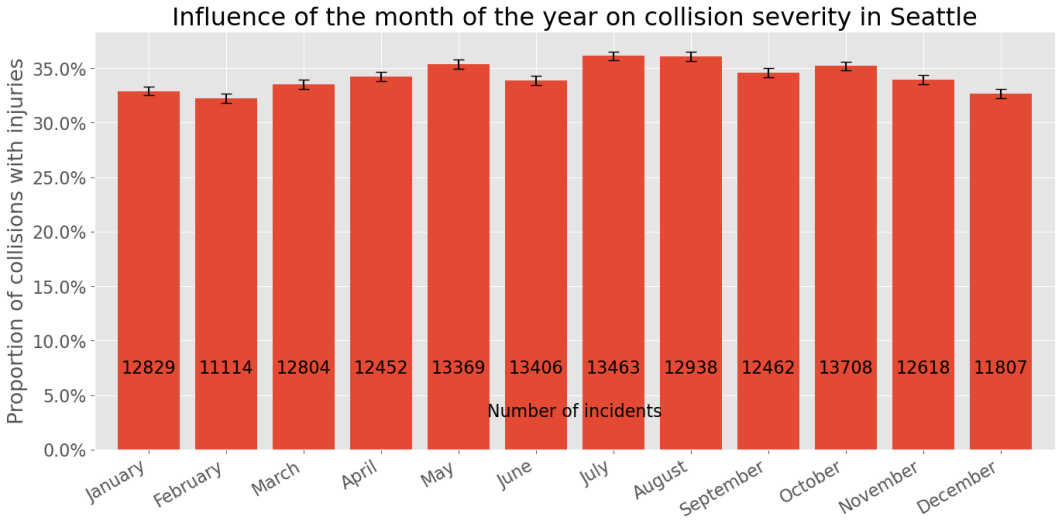
We can recognize a pattern throughout a day. Typically, the severity of accidents is the lowest during the night, which is in agreement with previous observations regarding the light conditions. Interestingly, the risk for a severe collision event seems to be the highest during the rush hours in the morning and in the evening with a trough in between.

Besides the time of the day, it could be beneficial to check whether the day of the week has an influence on the severity of accidents.



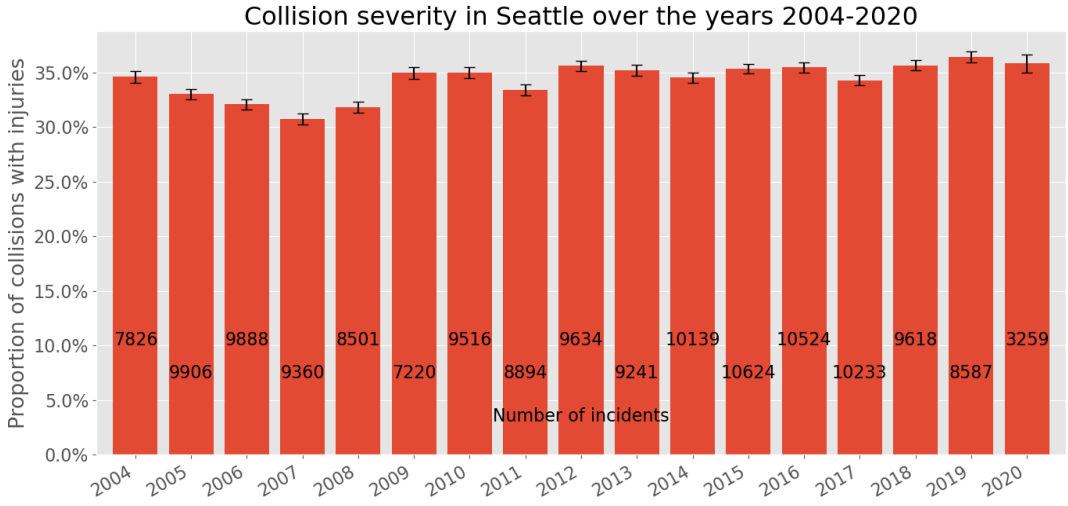
The day of the week shows less distinct differences. Only Friday, Saturday and Sunday display a somewhat reduced risk for the severity of accidents.

After the time of the day and the day of the week, it is consistent to further consider the potential influence of a month.



In general, the months in the middle of the year, especially July and August, appear to be more risky. The winter months from December to February display a lower risk for accidents involving injuries. This is consistent with the previous observation that snowy weather is associated with a smaller proportion of severe collisions.

Last but not least, we will check whether there is a trend throughout the years from 2004 to 2020 in terms of road safety.



There is not much difference between years in regard to the severity of collisions. The years 2005 to 2008 display a significantly lower risk for collisions resulting in injuries. Overall, one could maybe recognize a very light trend towards a higher proportion of collisions with injuries. Noteworthy is also the very low number of traffic accidents in 2020. The data set only contains events up to October 9 of 2020 but it is already clear that 2020 will have only about half the number of accidents of a typical year. This is most probably due to the ongoing COVID-19 pandemic and the resulting economic recession and reduced individual mobility. It is at least one positive aspect of an otherwise dire situation.

3.2 Data description

	SEVERITYCODE	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	PEDES	CYCL	VEHIC	YEAR	MONTH	HOUR	DAYOFWEEK
0	1	Driveway Junction	Clear	Dry	Daylight	0	1	1	2013	March	17	Thursday
1	1	Mid-Block (not related to intersection)	Overcast	Dry	Dark - Street Lights On	1	0	1	2006	January	17	Sunday
2	0	At Intersection (intersection related)	Overcast	Wet	Daylight	0	0	1	2019	September	15	Monday
6	0	At Intersection (intersection related)	Clear	Dry	Daylight	0	0	1	2020	July	10	Friday
8	0	Mid-Block (not related to intersection)	Clear	Dry	Daylight	0	0	1	2006	April	16	Tuesday

After removing all incidents without a complete set of information, our data set was reduced from originally 221,525 to 152,970 entries. This number should still be sufficiently large to provide an adequate training for our classification algorithms. When we remove all columns that does not contain information that is useful or that only retrospectively describes collision events, we end up with a data set that has 12 columns – one target feature and 11 explanatory features. The eleven features that should predict the severity of a collision event are: the junction type, the weather, the road conditions, the light conditions, the involvement of at least one pedestrian, the involvement of at least one cyclist, the involvement of at least one vehicle, the year of the accident, the month, the day of the week, and the hour of the day. For comparison reasons we will also keep a data set that does not contain any of the additional features that we extracted. That means that the only explanatory features are the junction type, the weather, the road conditions, and the light conditions. This should allow us to estimate how much the additional features improved that performance of the classification algorithms.

3.3 Data preparation

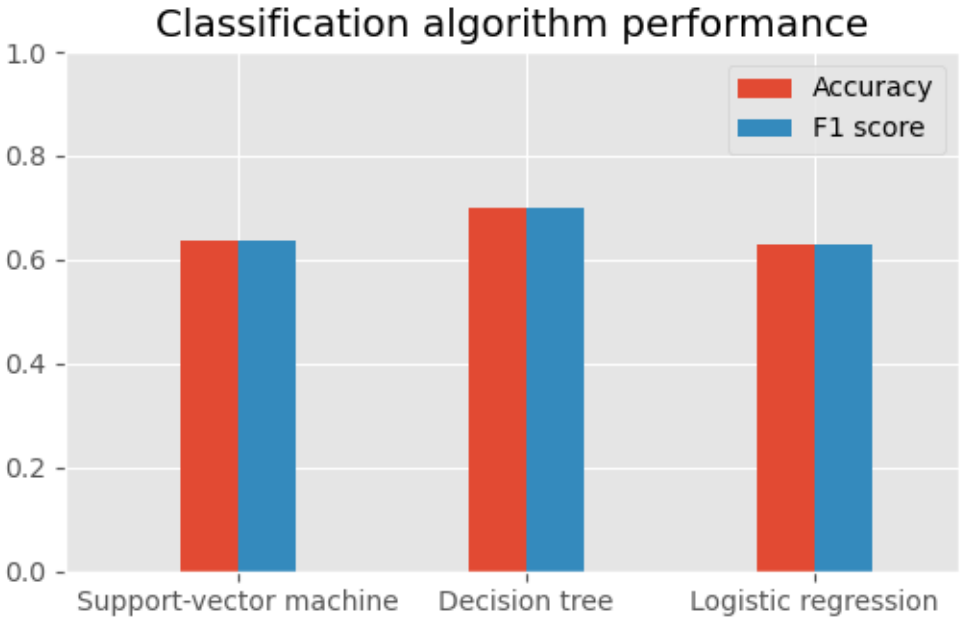
Before we can apply the classification algorithms, we need to prepare the data set accordingly. We will assume that all explanatory features, also the one relating to time, are categorical features. That means that we will need to perform a one-hot encoding on all of them. One-hot encoding means that we transform a single feature with multiple classes into many features that are only binary. This means that our data set subsequently contains 95 explanatory features instead of the former eleven. The data set without extracted features that we kept for comparison reasons contains 33 features after the one-hot encoding.

Before we apply the classification algorithms, we need to address the imbalance in our data set. The data set contains 100,538 incidents with only property damage and 52,432 incidents that resulted in injury. A naive algorithm could always predict a collision without injury and would achieve about a 66% accuracy. Of course, such a naive algorithm is not very useful. In order not to give a naive approach an unfair advantage we need to balance the data set. There are a couple of common methods available to balance an imbalanced data set. We will at this point not discuss all possible options. In this case we decided to apply over-sampling to our data set. This means that we randomly draw from samples from the minority class (collisions with injury) and add them to the data set until the two classes are equally represented in the data set. After the over-sampling procedure our data set contains 201,076 incidents. The same procedure will also be performed on the comparison data set.

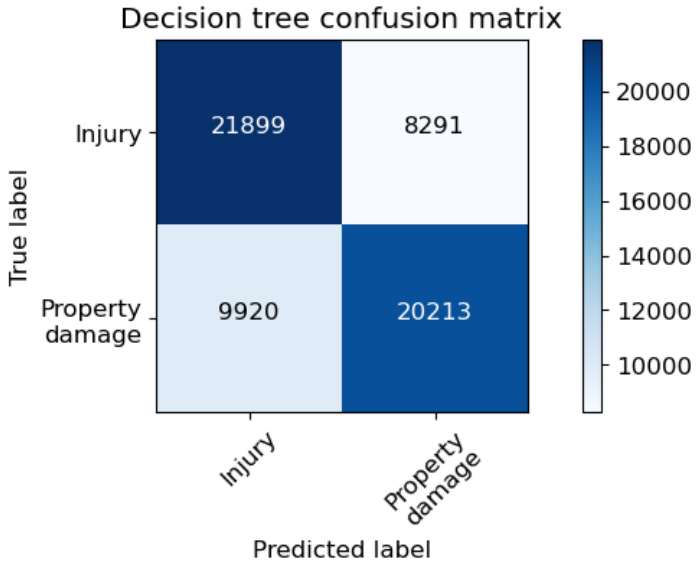
There are several well-established classification algorithms. Because all our features are binary, we chose to train a support-vector machine, a decision tree, and a logistic regression algorithm. Unfortunately, it is not the scope of this case study to discuss the inner workings, and advantages and disadvantages of these algorithms. All three algorithms are supervised machine learning algorithms and require us to provide them with a training set and subsequently test their performance on a test set. For this reason,

we will randomly split the data set at a 7:3 ratio. This means the training set contains 70% and the test set 30% of all incidents. All algorithms will receive the same training set and will be evaluated on the same test set. To test and compare the performance of the algorithms we will assess them by calculating their accuracy and their F1 score.

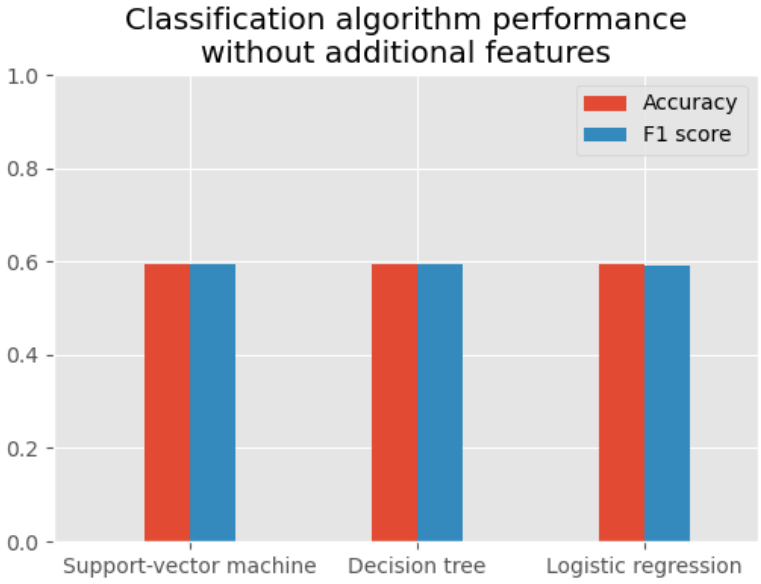
4. Results

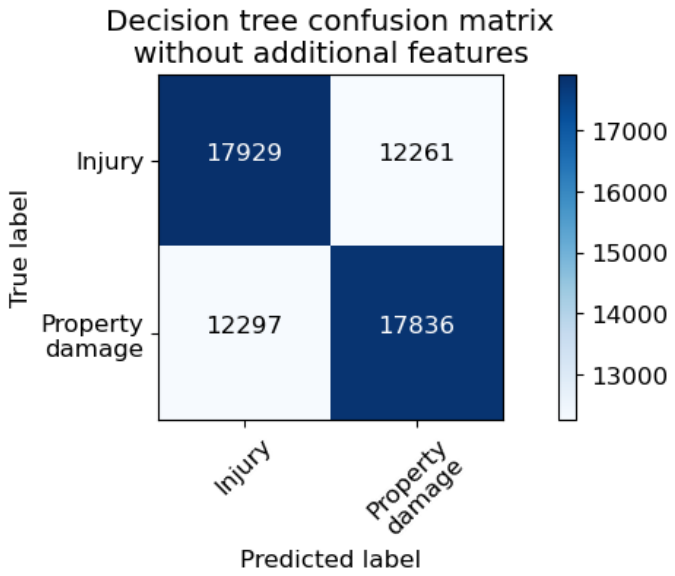


All three classification algorithms achieve an accuracy of over 62% and an F1 score of more than 0.62. The decision tree model achieves the best performance with an accuracy of 70% and an F1 score of 0.70. Looking at the confusion matrix of the output of the decision tree model we recognize that it is fairly balanced with a slight tendency of predicting more often *'Injury'* when it is actually *'Property damage'* than the other wrong assignment. Correspondingly, the model is better at correctly identifying incidents that involved injury than those that involved only property damage.



The performance metrics of the classification algorithms in the data set that did not contain the extracted features were substantially worse. However, all algorithms performed almost equally well. All achieved an accuracy of 59% and an F1 score of 0.59. The confusion matrix for the decision tree model is very balanced and shows approximately the same proportion of false positives and false negatives.





5. Discussion

When we assess the performance scores of the models it is important to consider the worst possible scores as the baseline. In a perfectly balanced data set we could employ a naive algorithm that only predicts collisions with injuries or without. Such an algorithm would be correct in 50% of all cases in a balanced data set. That means that such an algorithm would also have a 50% accuracy. Hence, 50% accuracy should be regarded as our baseline. The best performing algorithm was the decision tree model that achieved 70% accuracy and an F1 score of 0.70. This is substantially better than the hypothetical naive algorithm that we proposed before. One might be thinking that such an accuracy is still not particularly good. However, we also need to consider that the information at our disposal to make these predictions is rather crude and is based on only comparably few features. Furthermore, the apparent

influence of many features on the severity of collision seemed rather weak. With these considerations in mind it is actually quite remarkable that our model is able to reasonably well predict situations in which the outcome of a collision event is more severe.

The performance of the model could be improved if more information about the collision events would be available. At the moment, the model has to treat every incident equally even though in some incidents more than one person might have been injured. Such collision events should correspondingly receive a higher weight when the model is trained. Furthermore, the quality of the available data was not without doubt. Occasionally there were incidents that had a combination of information that seemed contradictory. A better understanding of the data and maybe also a check for consistency might improve the data quality and thus the model performance.

6. Conclusion

The goal of this case study was to investigate whether we can use information about environmental factors to predict the severity of traffic collision events. We wanted to train a model with publicly available data from the City of Seattle to inform drivers and other traffic participants about conditions that favor a more or less severe outcome of a potential collision event.

We showed that by analyzing the data set and extracting further information we can train a machine learning model that has a substantially higher performance and accuracy than an approach that would be based purely on guessing based on the proportion of severe accidents. Such a model could hypothetically be used to warn drivers about dangerous conditions and situations. If used on a wide scale in a similar environment to a city like Seattle, it could potentially reduce the number of severe collision events and protect the health and lives of a large number of people.