Analysing Vehicle Accident Severity

Elena Gerken
October 2020

1. Introduction

1.1 Background
Road safety is an important issue for modern societies. According to the CDC, road traffic crashes are the leading cause of death in the United States for people aged 1–54. Big cities need to respond to this by managing the growing volume of traffic to mitigate accidents and reduce fatalities and serious injuries on the road.

The city of Seattle has been recording vehicle accidents in order to look for patterns and innovative solutions to road safety issues. Examining the risk factors that contribute to more dangerous accidents can help to direct policies and police activities. The city will then be able to adjust and develop traffic-related policies and prevention operations accordingly, for example by providing better lighting or warning systems when weather conditions become dangerous.

1.2 Problem
Learning which conditions are risk factors that elicit more severe accidents and therefore cause more fatalities and serious injuries is crucial. This project will try to find models to predict when accidents are prone to becoming more dangerous. This should provide the necessary information to help make accidents less fatal.

2. Data

2.1 Data source
The city of Seattle published a dataset of vehicle collisions on its open data portal in 2018. The dataset shows 221,738 records of accidents going back to 2004 and consists of 40 columns, including id numbers from the state of Washington and the city of Seattle, report numbers, injury counts and markers such as if a parked car was hit and how many vehicles were involved.

The target variable we will be trying to predict with our models is the column ‘Severity’. This variable in the dataset has five possible outcomes:
- Property damage only collision
- Injury collision
- Serious injury collision
- Fatality Collision
- Unknown

Our goal is therefore to predict how dangerous a collision will be.

2.2 Feature selection
For the purpose of this analysis, we will be choosing a select few features out of the 40 columns that we assume to have the best prediction value:

- Weather condition (clear, raining, overcast, snowing, fog etc.)
- Road condition (dry, wet, standing water, ice etc.)
- Light condition (daylight, dark, streetlights on/off etc.)
- Whether or not the collision was due to inattention. (Y/N)
- Whether or not a driver involved was under the influence of drugs or alcohol.
- Whether of not a driver was speeding.
- The number of pedestrians involved in the collision.
- The number of cyclists involved in the collision.

We will be examining these features further in the exploratory data analysis section of this report to determine which factors have the biggest influence on the severity of an accident.

2.3 Data cleaning

After having selected the listed feature columns, I dropped all rows that were marked as ‘Unknown’ for the target variable ‘Severity'. The dataset we are now working with consists of 200,081 rows and 9 columns (8 features and 1 target variable).

The selected feature columns needed some cleaning, too. Most columns had a mix of 0, 1, Y and N values. The columns ‘PEDCYLCOUNT’ and ‘PEDCOUNT' listed *how many* cyclists or pedestrians were involved in an accident, for our purpose, the information *whether* a cyclist or a pedestrian was involved is sufficient, so I changed it to a binary value. This resulted in the columns ‘Speeding’, ‘Cyclist', ‘Inattention’, ‘Pedestrian’ and ‘Underinfluence' having binary values showing whether this condition was true (1) or not (0) for the particular accident.

The columns ‘Lightcondition', ‘Roadcondition' and ‘Weather’ contained much more information then the other feature columns. I summarise some information (like ‘Unknown’ and ‘Other’, or ‘Blowing Snow’ and ‘Snowing’) and renamed some values. This should help moving forward with selecting the values and getting useful information out of these features.

For now, we will leave these features as categorical values and change them into dummy values later. The predictive models will need the columns Lightcondition', ‘Roadcondition' and ‘Weather’ to be binary, but for the purpose of exploratory data analysis we will leave them intact.

This resulted in the following sample of the dataset:

Severity	Pedestrian	Cyclist	Inattention	Under influence	Speeding	Weather	Road condition	Light condition
Property Damage Only Collision	0	0	0	0	0	Clear	Dry	Daylight
Property Damage Only Collision	0	0	1	0	0	Rain	Wet	Dusk
Injury Collision	0	0	0	0	0	Clear	Dry	Dark_Street_Lights_On
Injury Collision	1	0	0	0	0	Rain	Wet	Dark_Street_Lights_On
Injury Collision	0	0	0	0	1	Clear	Ice	Dark_Street_Lights_On

2.4 Balancing the Dataset

The underlying dataset is imbalanced, meaning the target variable ‘Severity’ is heavily biased towards the outcome ‘Property Damage Only Collision’:

Severity	Total count	% of Dataset
Property Damage Only Collision	137,776	68.9 %
Injury Collision	58,842	29,4 %
Serious Injury Collision	3,111	1,6 %
Fatality Collision	352	0,18 %

This is obviously due to the nature of the data: Thankfully, fatal accidents are much more rare then property damage only collisions. However, for the purpose of a predictive model, we have to correct this imbalance. Otherwise, a model that would never predict a fatal collision would still be accurate 99,82% of the time. This number would give the impression of a good model, but it would fail in our attempt to predict fatal accidents.

To balance out the dataset and make a proper prediction model possible, I therefore down-sampled the results 'Property Damage Only Collision' and 'Injury Collision' and up-sampled the 'Fatality Collision' to 3,111 values respectively. This resulted in an even distribution of 25% for each of the outcomes. In total, the dataset now consist of 12,444 rows.

3. Exploratory Data Analysis

3.1 Preliminary remarks

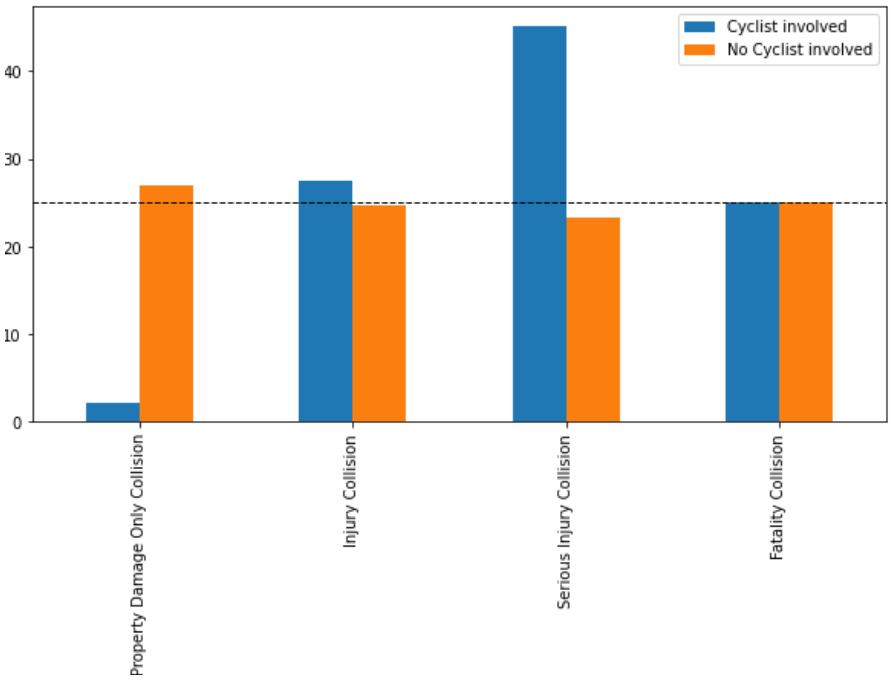
In the following paragraphs, we will be examining the relationship of our selected feature variables with the target variable 'Severity'.

Because we balanced the dataset and we normalised the feature data, if a feature has no impact on the outcome of the accident, each accident severity will appear with the same probability. As we have four different possible outcomes for the severity of an accident, a random distribution, or a feature with no impact, will result in about 25% for each result. Therefore, 25% is our threshold to determine if a feature has an impact on the severity of the accident or not.

The total amounts of accidents were checked prior to make sure that the charts had an appropriate amount of data to make a statement. If any of the results should be made based on less then 100 cases it will be mentioned in the according paragraph.

3.2 Cyclist

As our first variable, let's determine what kind of effect the involvement of cyclist has on an accident. We will therefore be comparing accidents in which no cyclist was involved to accidents in which at least one cyclist was involved.

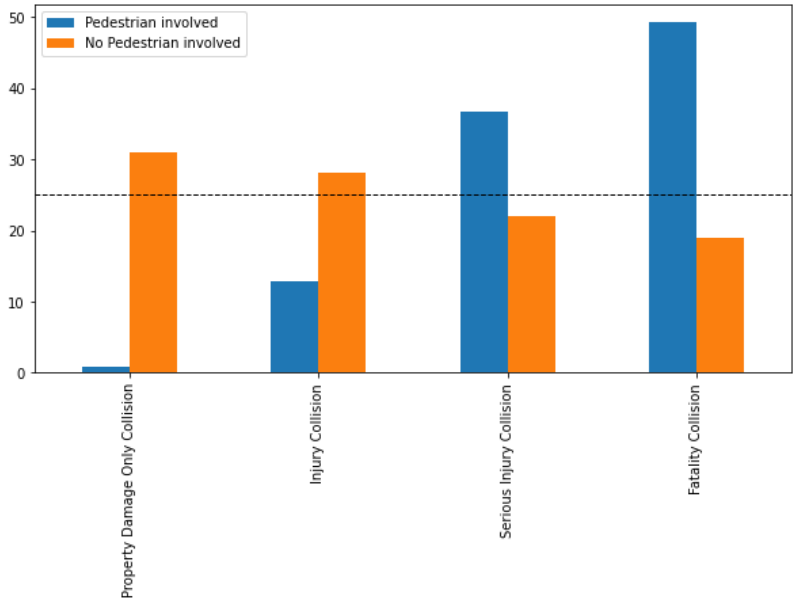


The above bar chart can be read like this: Looking at the orange bars, we can see that accidents that occurred without cyclists involved were about as likely to be property damage only as they were injury of fatal (about 25% respectively).

Accidents with cyclists however are not distributed evenly. When a cyclist is involved in a collision, the outcome is much more likely (> 45%) to end up with serious injuries. These accidents are on the other hand much less likely to turn out with only property damage (≃ 2%). Given that cyclists are much more vulnerable than passengers in a car, this result is not surprising. It shows however that this is a good variable to determine the severity of a collision.

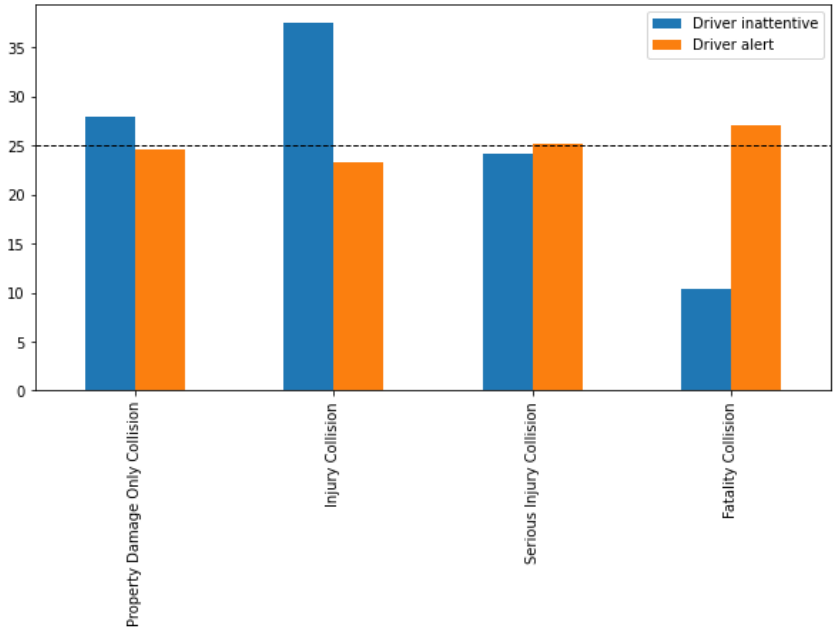
3.3 Pedestrians

This feature variable is quite interesting, because we can see that there is a positive effect on the severity of the accident when no pedestrians (orange bars) are involved. On the other hand, when pedestrians are involved in an accident, these are much more likely to end with serious injuries (> 36%) or fatalities (> 49%). This strong correlation indicates that the involvement of pedestrians has a big impact on the severity of the accidents.



3.4 Driver Inattention

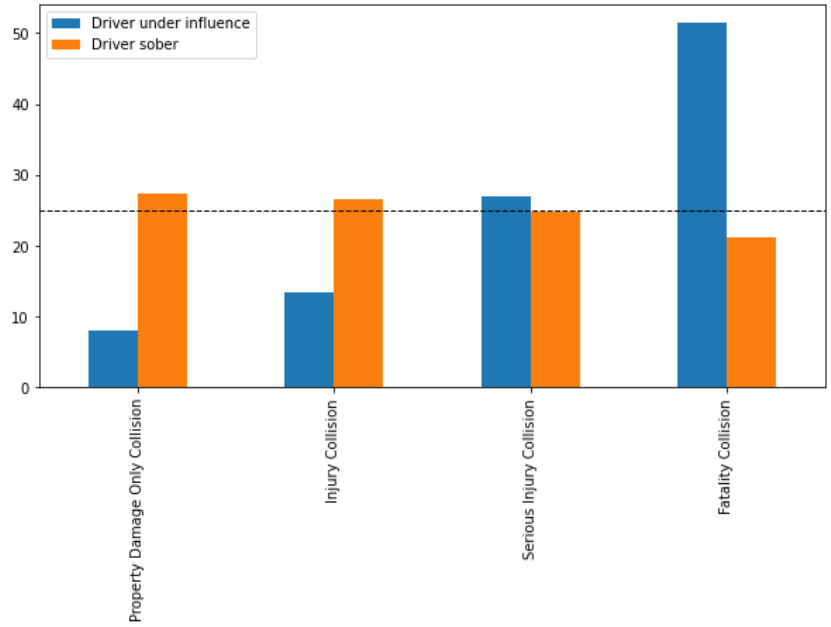
Next, we will look at the effect driver inattention has on the outcome of a vehicle collision. Again, looking at the orange bar chart, showing alert drivers, we can see that the severity of the accidents is distributed quite evenly at about 25%.



The blue charts however, showing accidents with inattentive drivers have a clear effect: These accidents tend to be injury collisions (> 37%), while they are less likely to cause fatalities (> 11%).

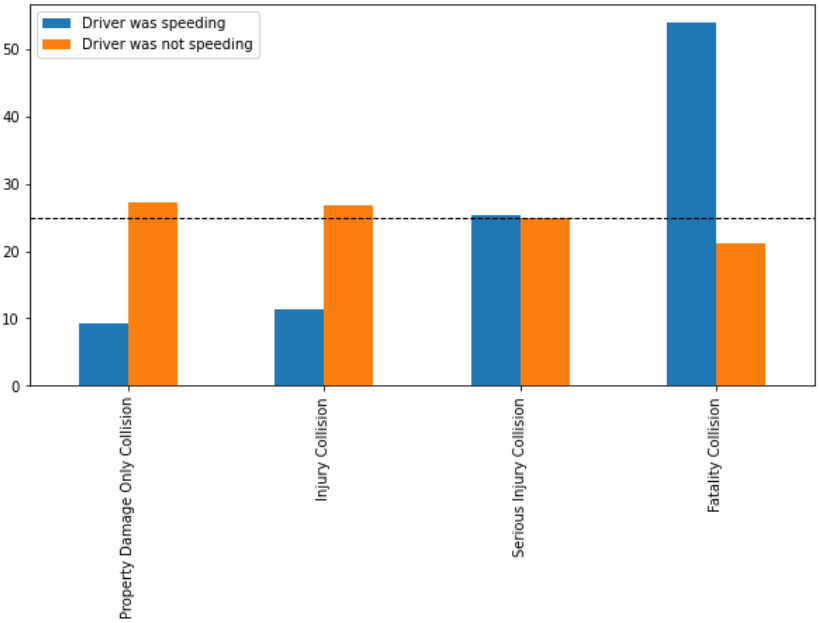
3.5 Drugs and Alcohol

People driving under the influence of drugs or alcohol are a major problem on the roads. The below chart clearly shows that accidents with drunk drivers are prone to end with fatalities (> 51%). Out of all the feature variables, this was one of the strongest correlations between a feature variable and the outcome of an accident being fatal.



3.6 Speeding

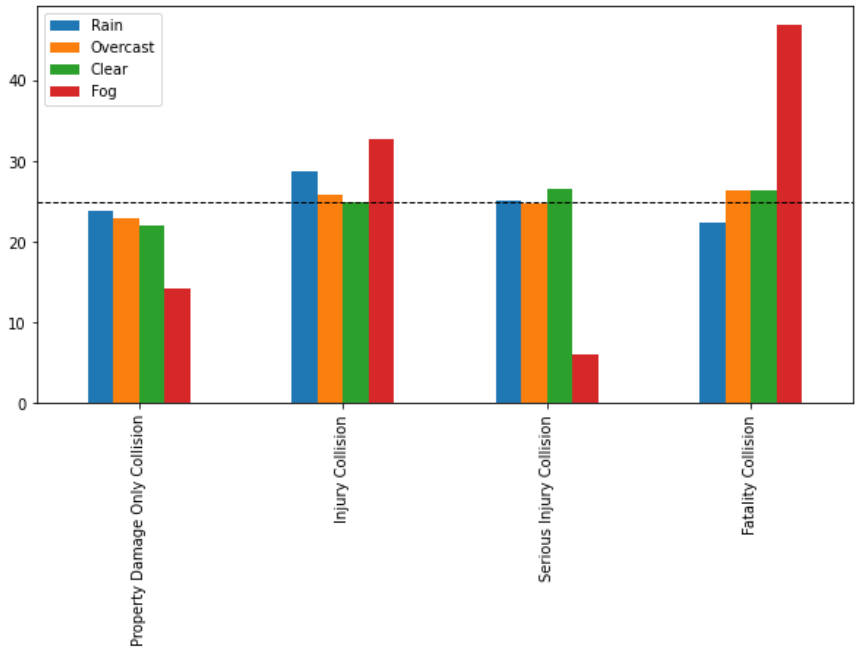
The effect of speeding on the outcome of an accident is equally strong as with drunk driving. Accidents where the driver was speeding are strongly correlated to fatalities (> 54%), while they are less likely to end with just property damage or light injuries.



3.7 Weather conditions

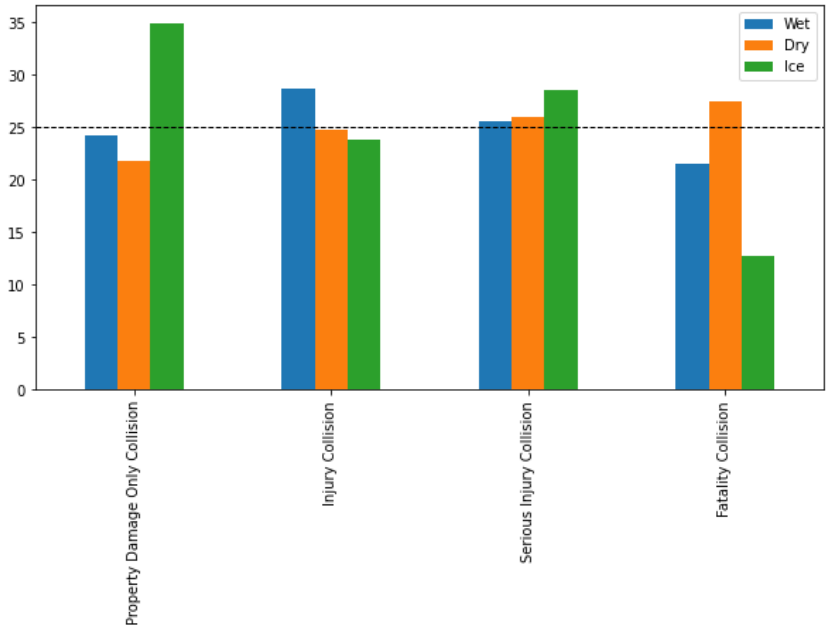
Next, lets have a look at the effect the weather conditions have on the outcome of vehicle accidents. Based on the 25% threshold we have been working with, we can see that rain, overcast and clear weather do not seem to have a big impact on the severity of accidents. While there seem to be slightly more injury collisions when it is raining, all accidents are still distributed quite evenly.

The feature ‘Fog’ seems promising, as it shows more fatalities compared to the other factors. However: In total, only 49 out of 12,444 accidents involved the feature Fog, making it less then 0,5% of the dataset. The correlation might be there, but it is not backed by a lot of data and should therefore be considered less safe then other correlations that were based on many more cases.



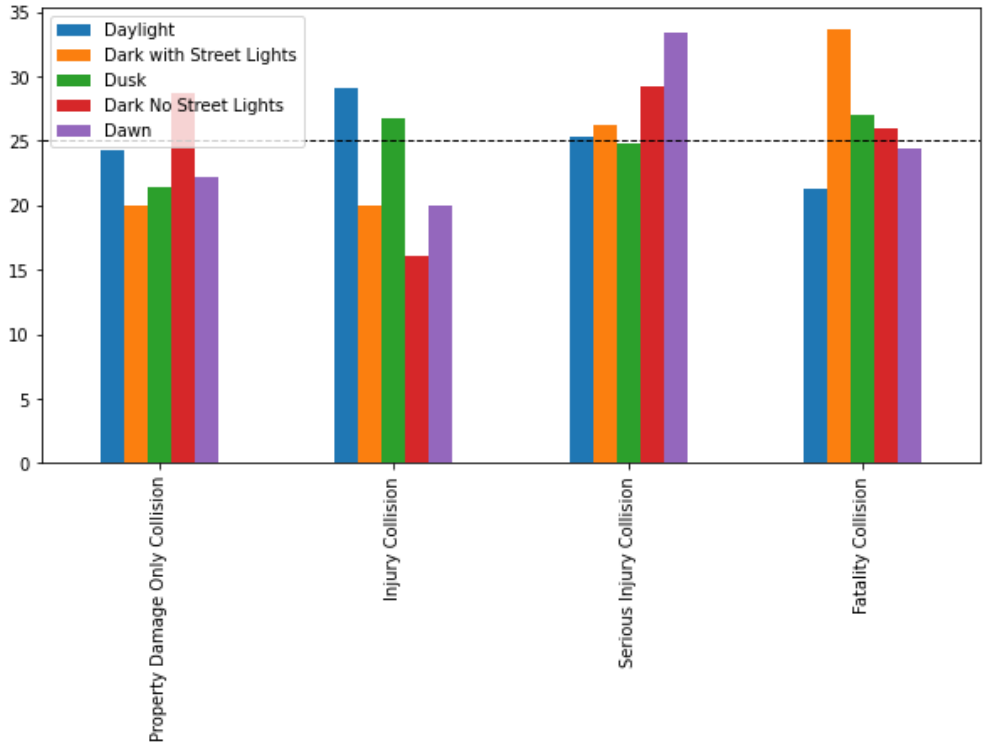
3.8 Road conditions

The effect of road conditions on the severity of accidents is not very strong. Wet and dry road conditions only fluctuate slightly around the 25% mark. Interestingly, Ice seems to have an almost positive effect on the severity of accidents, making them more likely to produce only property damage. However, this is again based on only 101 cases, so less then 1% of the total dataset.



3.9 Light conditions

Light conditions are hard to make out as a factor influencing the severity of collisions. More collisions seem to be fatal when it is dark with street lights on. There appears to be more serious injury collisions at dawn.



4. Predictive Models

4.1 Preliminary remarks

After the exploratory data analysis, I decided to include all the variables we examined into the training of the models. Even though some variables have a much stronger predictive value then others, all showed some kind of tendency to influence the prediction of the target variables and would therefore be valuable to make out models better.

I changed the columns ‘Weathercondition’, ‘Roadcondition’ and ‘Lightcondition’ into dummy variables. All of our feature variables now had a binary value of either 0 or 1 and did therefore not need any normalisation.

Before developing our prediction models, I split the data into train and test sets. 20% of the data was reserved to serve as a test set to evaluate the models performance.

4.2 Model Development

Due to the nature of our data, the fitting approach to predict the severity of vehicle collisions were classification models. That is because we only had four possible outcomes compared to a spectrum of outcomes that would suggest a regression model. I chose to develop K-nearest-neighbour, Support vector machine, decision tree and Logistic regression.

For K-nearest-neighbour I plotted models with different k-values to chose the model with the best accuracy.

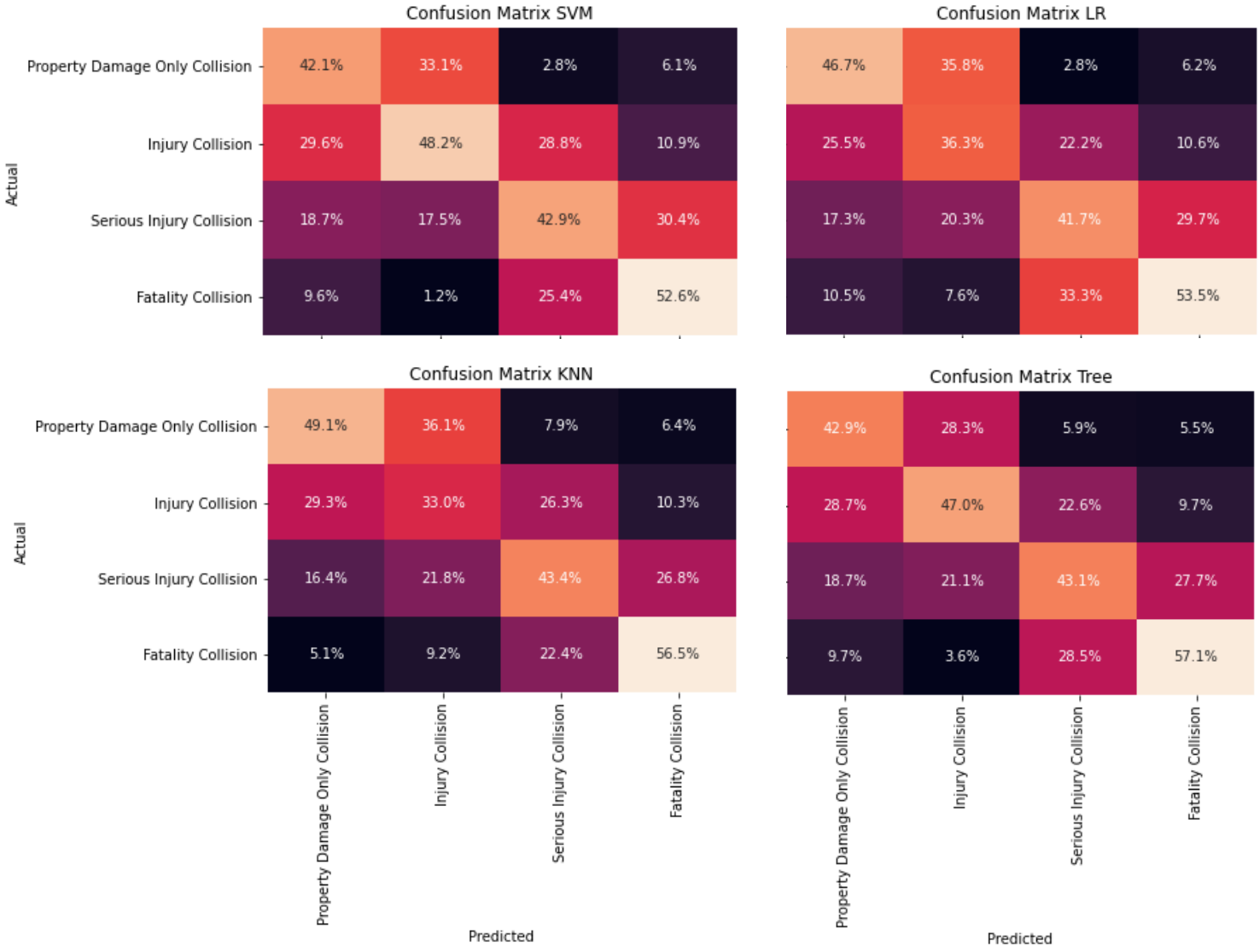
4.3 Model Evaluation

First, I calculated the F1 score for our four prediction models:

Model	F1 scores
K-nearest-neighbour	0.42
Decision Tree	0.44
Logistic Regression	0.43
Support Vector Machine	0.4

The best value for the F1 score being 1, while the worst outcome would be 0. An outcome of about 0.4 is not great, even though it is better then a random outcome, which would in our case be 0.25.

To get a better understanding of the models strength and weaknesses, I decided to generate a confusion matrix for each model. While the metrics showed slight differences in the numbers, they all clearly shared some similarities, too:



For all models, it seemed to be easiest to predict the variable ‘Fatality Collision’. The true positive rate for this outcome was between 52% and 57% depending on the model. The true positive rate for the outcome ‘Property Damage Only Collision’ was still good at about 42-49%. Similarly, the outcome ‘Serious Injury Collision’ was predicted correctly about 41-43%.

The Support Vector Machine and Decision Tree models did a better job at identifying ‘Injury Collisions’ correctly (47% and 48,2%), compared to 33% and 36.3% for the K-nearest-neighbour and Logistic regression model respectively.

The models all struggled with differentiating between accidents that were close in severity, for example confusing a ‘Property Damage Only Collision’ with ‘Injury Collision’ (28-36%) or predicting a fatal collision when in fact the accident only produced serious injuries (26-30%).

5. Results and Discussion

5.1 Results

In this study, I analysed the relationship between the severity of vehicle collisions with driver behaviour and external conditions. I found that drunk driving and speeding contributed most to deadly outcomes of accidents. Also, accidents tended to be much more dangerous when pedestrians were involved.

I developed four classification models to predict the severity of accidents. While they were far from perfect in terms of prediction value, this report could still show which behaviours and conditions contributed most to accidents ending with severe injuries or death.

Overall, the exploratory data analysis contributed most to the findings, as it showed the strong impact some features had on the outcome of the accidents.

5.2 Recommendations

To reduce fatal accidents on the road, police and policy makers should concentrate on drunk driving and speeding, as well as protecting pedestrians.

Additionally, I found that inattentive driving leads to less severe accidents. While these accidents obviously should still be avoided all together, they tend to be less fatal and should therefore be less penalised.

5.3 Future outlook

For future studies, it might be interesting to take a closer look at the location and time of accidents. Both information were available in the dataset but were not included in this report.

The location might give some insights into which intersections have a potential to be particularly dangerous. This could especially be interesting because we found that when pedestrians were involved in accidents these tended to be more severe. Finding intersections with many severe accidents could help decide where to install better safety measures like traffic lights.

The timing of accidents could also prove to be insightful. Our categories light and weather conditions had part of that information, but I could imagine that there are certain times of the day where more accidents happen (rush hour for example) or more severe accidents occur (drunk driving after-hours). This could help direct policing and attribute to making police operations more effective.