Car Accidents in Seattle

By Evaristo Acosta

1. INTRODUCTION

Deaths in -or by- automobiles are a pressing public health concern. It is the second leading cause of accidental death in the country.

In 2018, Seattle alone had over 180 traffic accidents that resulted in serious injury or death. Since 2015, Seattle has made it its goal to drop the number of accidents to 0 by 2030, using Sweden's Vision Zero plan. While the number has fallen from over 326 in 2006, the city has a long way to go before it can hit that lofty goal.

Even if the people survyve, they are still suffering serious injuries in hundreds of car accidents each year. Those who survive traffic accidents can face hefty medical bills, thousands of dollars in lost wages and property damage, pain and suffering, and lost quality oflife.

A collision could cause permanent disability, stripping the victim of the life he or she might have had. For compensation for all of these serious losses, car accident victims can turn to the Seattle civil court system.

In other words, car accidents are a serious problem for both citizens and the state for all the expenses and inconveniences that it entails. Looking for the causes and how to reduce them then becomes a relevant problem.

2. PROBLEM

The goal is to use data that can help determine which characteristics increase the severity of accidents, such as road conditions, weather conditions, the exact time and location of the accident.

The Seattle Police Department (SPD) has recorded all car accidents from 2004 to the present. Based on these historical data (194,673 records) we can come to understand the factors that influence an increase in injuries. With the intention of avoiding accidents and better planning our next trip to Seattle.

Reducing the severity of accidents can be beneficial to the whole of society, both for the Seattle government, which seeks to protect citizens and reduce the costs that accidents cause, and for car drivers, who can avoid risky situations .

3. DATA

In this project i will be using the Data-Collisions dataset, provided in this course for be analyzed.

The dataset is a table that contains accidents and identifies their severity, location, and a long list of attributes. I will use some of them to try to know where and under what circumstances they are most likely to occur.

Through this project i will be analyzing the effects of some independents variables on the severity and frequency of the car accidents.

Using the tools acquired throughout these courses, I will analyze the dataset looking for to identify those things that citizens and the state can do to avoid them.

First i will preprocess the datset and decide which attributes are unnecessary and which i will use in this project. I will choose what to do with the empty cells and once this is done I will proceed to work with the data.

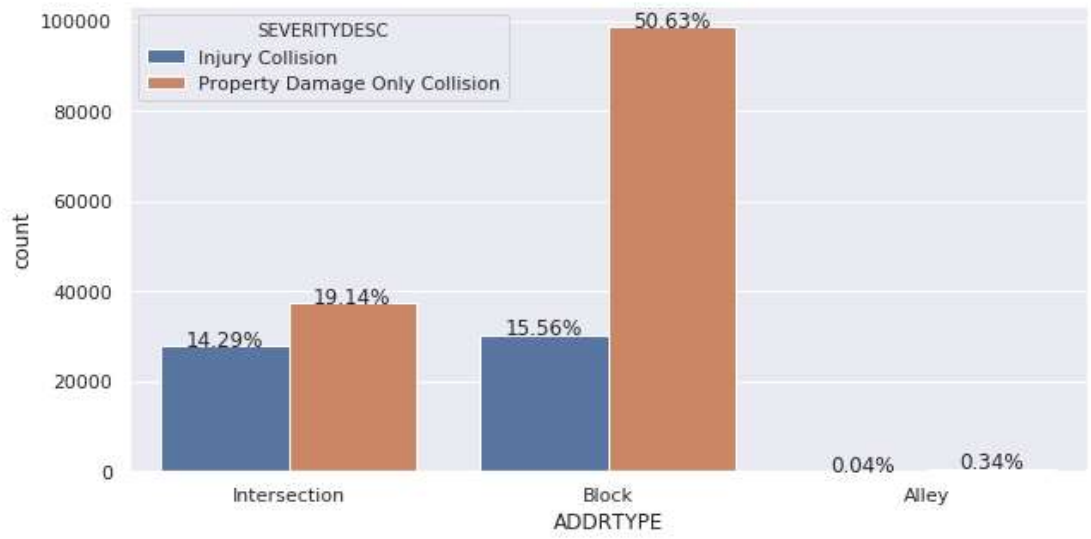
4. DATA ANALYSIS

ANALYZING USING VISUALIZATION:

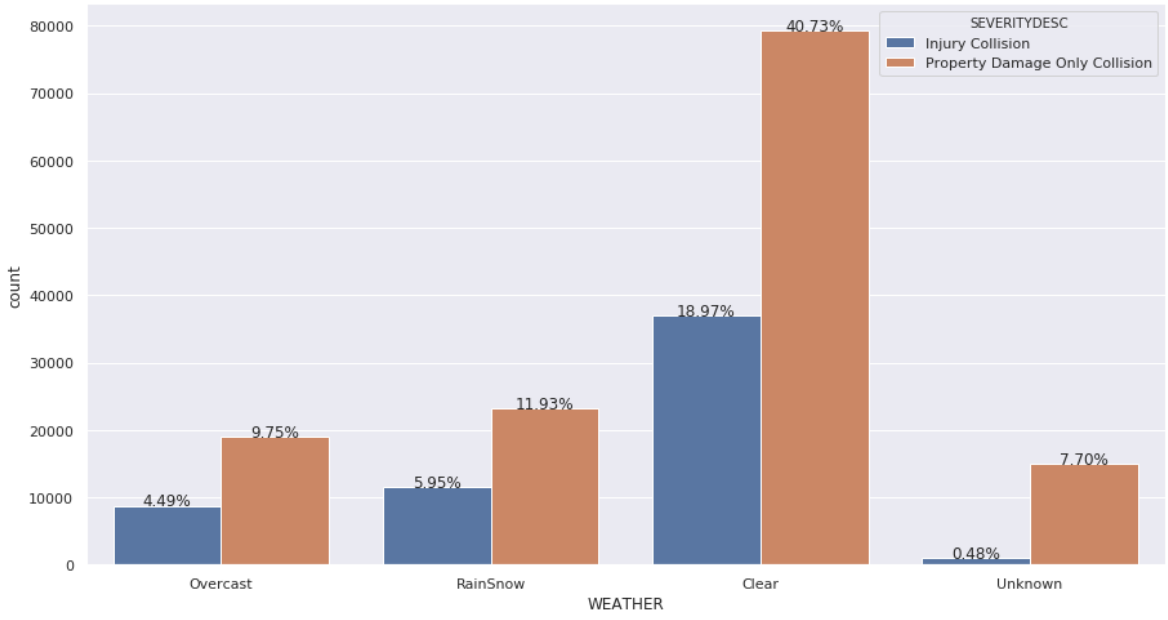
As we want to know what factors will have impact on injury collision compared to property damage, I plot the count and percentage of each type of each feature/attribute. The main goal is to find whether the relative ratio of each feature type differs.

INDIVIDUAL FEATURE PATTERNS

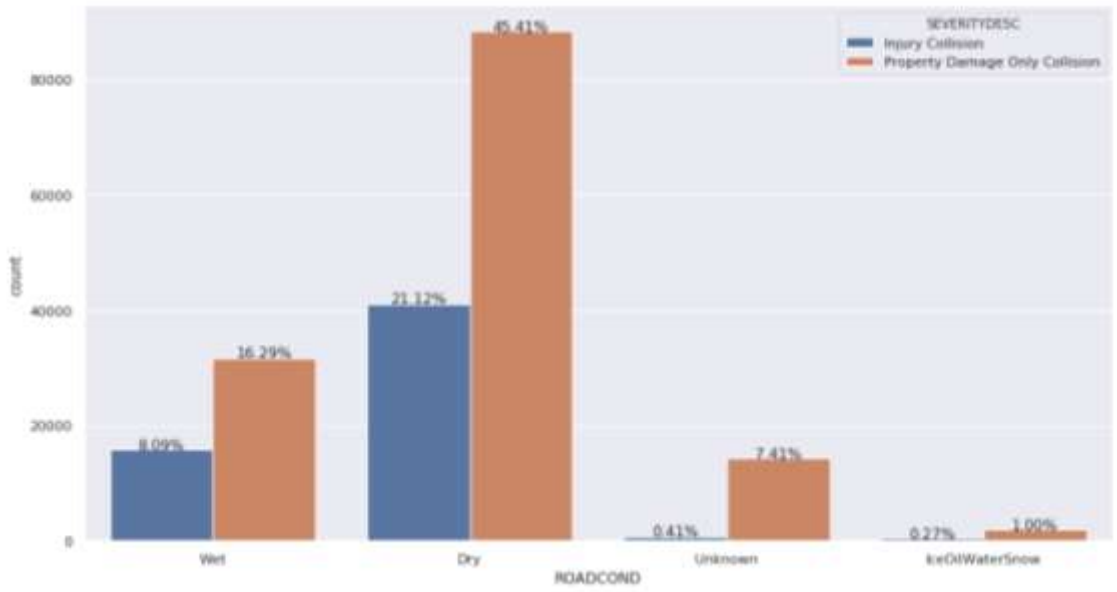
Common sense tells us that the place of impact affects the severity of the collision. In the following graph we can see that "At the intersection" is an important factor, where accidents are more likely to involve injuries, only slightly less than the chances of property damage. However, most accidents happen on the block, this is explained by the type of collision. Most of these collisions occur while parking, that also explains the high percentage of property damage.



Another factor that is usually considered relevant is the weather at the time of the collision, however the expected impact was not found. The weather does not seem to change the relationship in the severity of accidents between injuries and property damage. As can be seen in the graph below, the relationship remains constant.

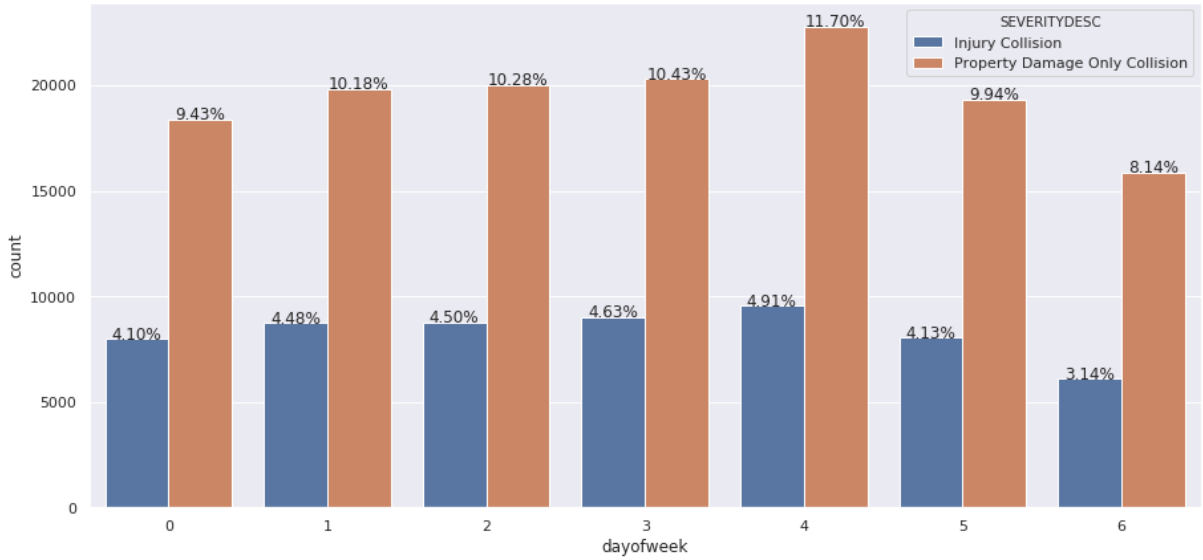


Regarding the condition of the road, the condition of the same does not seem to significantly modify the percentage of accidents with injuries. However, there appears to be a significant increase in accidents in general when the road is wet.

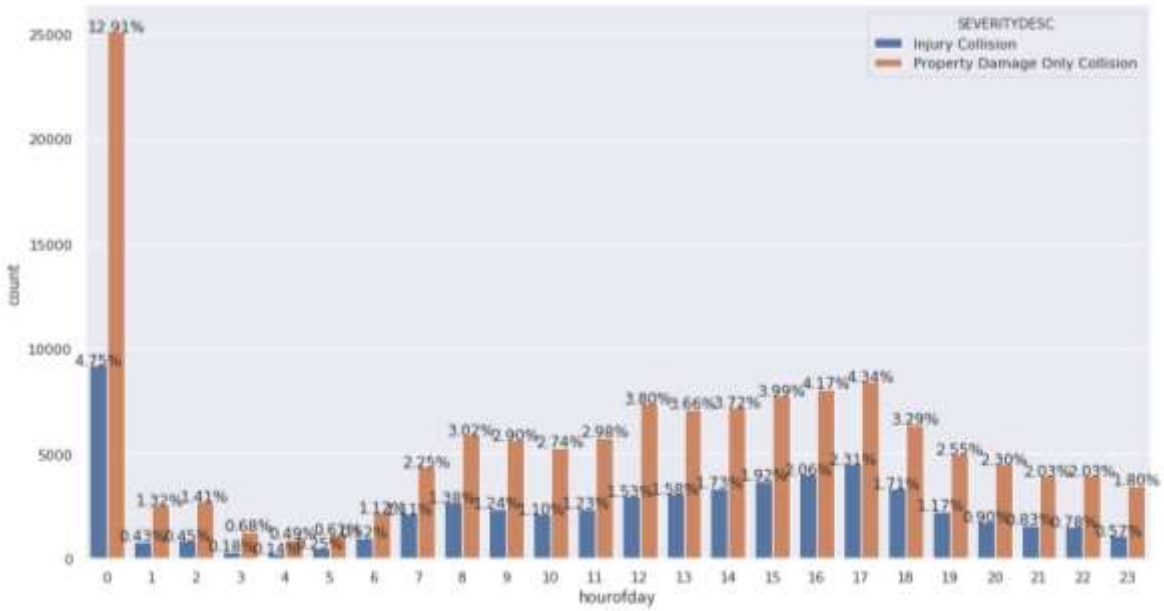


I then proceeded to analyze the impact of the time factor on collisions.

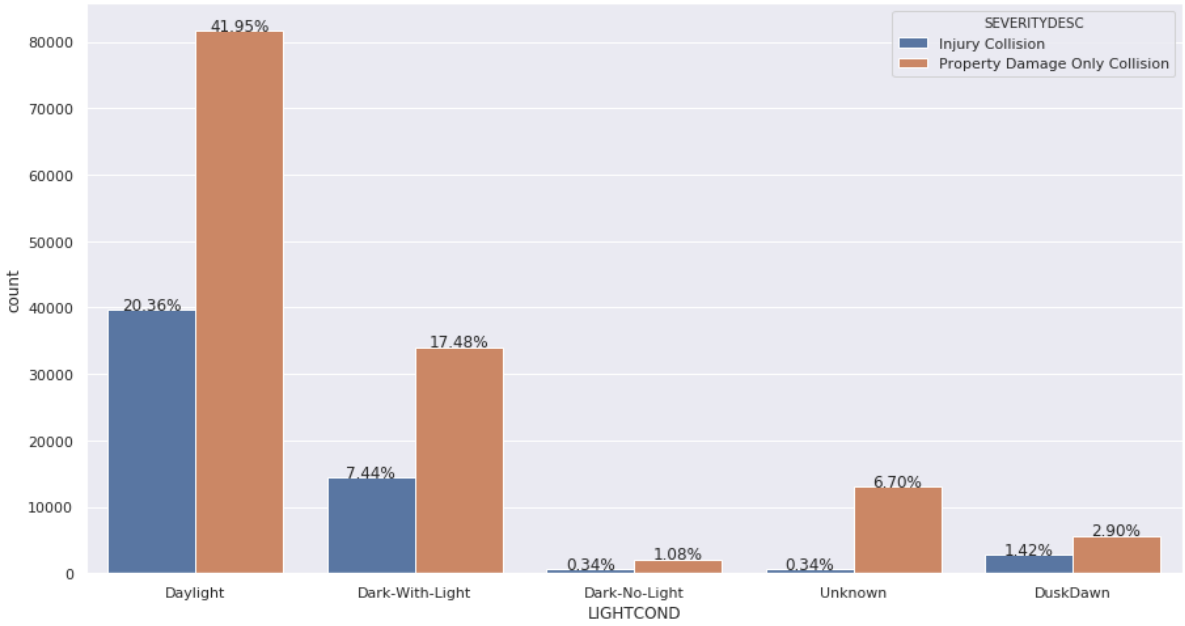
Analyzing the impact of the days of the week, we can see a small increase towards the end of the week, especially on Friday. But there is no pronounced increase on weekends, with Sundays being one of the calmest days.



As for the hours of the day, there is a pronounced reduction in accidents during the early morning hours and an increase in working hours. A peak is observed at midnight, but this shock is ignored, considering that it is probably due to an error in the data collection, which in the absence of data is completed by default with "00:00". The graph is shown below.



The graphic of the lighting condition is correlated with this idea.



5. METHODOLOGY

First i clean the data and realize a feature selection. As we want to analyze what factors will probably cause a car collision and the severity of the accident, we would drop those unrelated features and useless information, only keep 15 features/attributes of the original dataset.

MACHINE LEARNING MODEL SELECTION

The machine learning models used are Logistic Regression, Decision Tree Analysis, and K-Nearest Neighbor.

Logistic regression is a statistical model that, in its basic form, uses a logistic function to model a binary dependent variable.

Decision tree analysis breaks a data set into smaller subsets while at the same time incrementally developing an associated decision tree. The end result is a tree with decision nodes and leaf nodes.

K Nearest Neighbors is a simple algorithm that stores all available cases and classifies new cases based on the distance to their neighbors.

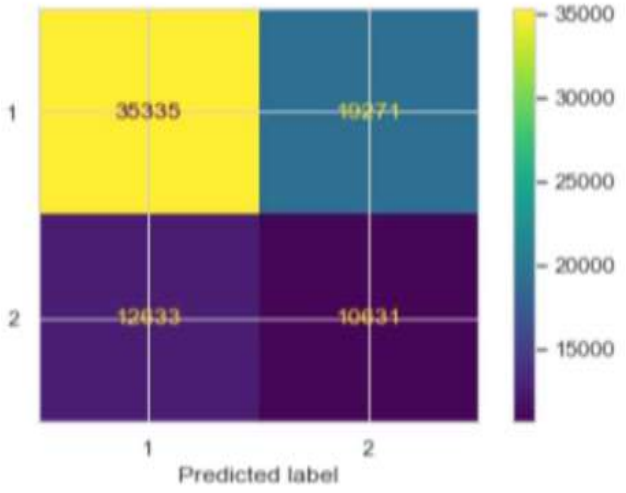
Do not use the Support Vector Machine (SVM) model because this model becomes inaccurate when working with large volumes of data, our database has more than 190,000 rows of data.

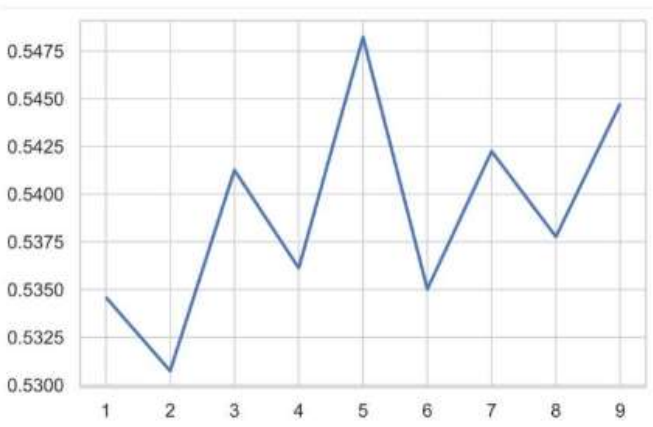
6. RESULT

The priority of the model is to predict "Injury Collision" type over the prediction of the "property damage only collision" type. The better indicator for this is the recall ratio of the Injury Collisions, so we are not so interested in the performance of the model in general but in the performance predicting Injury collisions.

Let's see then the performance of the different models.

K- Nearest Neighbors



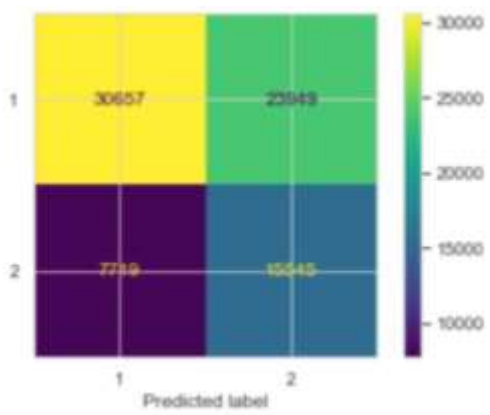


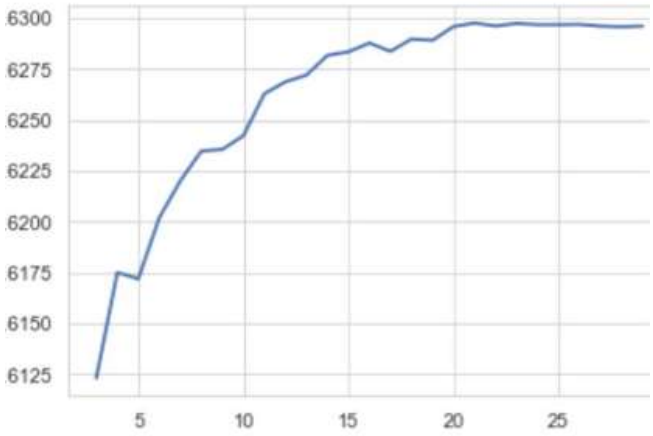
As we can see in the second graph, the best K for the model is 5.

Also, of this analysis can provide that the K-NN model show an accuracy of 59% and a recall rate of 46%.

	precision	recall	f1-score
Property collision	0.74	0.65	0.69
Injury Collision	0.36	**0.46**	0.40
accuracy			0.59

Decision Tree Analysis

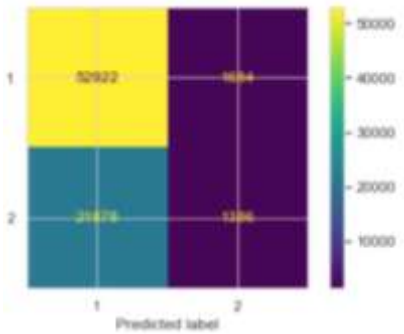




The Decision Tree Model show an accuracy of 59% with a recall rate of 67%.

	precision	recall	f1-score
Property collision	0.80	0.56	0.66
Injury Collision	0.39	**0.67**	0.50
accuracy			0.59

Logistic Regression



In the last analysis, we run a logistic regression model. The model has an accuracy of 70%, but the recall rate was only 6% has we can see in the next table.

	precision	recall	f1-score
Property collision	0.71	0.97	0.82
Injury Collision	0.45	0.06	0.11
accuracy			0.70

7. DISCUSSION

If we compare the three models, we can see that the most useful to predict "injury collisions" is the decision tree model. However, the model is still feasible for improvement, having a larger database could help to adjust the model, thus improving the predictions.

There is a long way to go in the quest to reach zero accidents. Having better and larger databases is essential to achieve the objectives.

Alogorithm	Average F-1 Score	Type	Precision	Recall
Decision Tree	0.61	Property collision	0.80	0.56
		Injury Collision	0.39	**0.67**
k-Nearest Neighbor	0.60	Property collision	0.74	0.65
		Injury Collision	0.36	**0.46**
Logistic Regression	0.61	Property collision	0.71	0.97
		Injury Collision	0.45	**0.06**

8. CONCLUSION

As we see before, the Decision Tree model was the better predictor of the Injury Collisions, but the model has a large error margin. And the improvement requires large and more exhaustive datasets. More information is required to have a better opportunity of identifying what increases the probability of an accident.

After having seen and analyzed the data, we can conclude that some simple actions could help reduce accidents and their severity.

Greater caution when driving during peak hours and on Fridays, special care at street crossings, and caution when driving when the road is wet are some of the actions that drivers can implement.

On the other hand, the government can improve the drainage of the streets so that they dry faster after the rains, signal the crossings of the streets and place speed bumps in the areas of greater traffic, as well as promoting the use of public transport , reducing the circulation of vehicles during peak hours.