Predicting Car Accident Severity in Seattle City

Daniel Orozco Venegas

October 13, 2020

Table of contents

1. Introduction

1.1 Background

The Seattle City population in 2019 was 747, 300 people and had grown at a 22.78% rate since 2010, at the same time, 81% of Seattle households owned at least one vehicle in 2019. Furthermore, in 2019 Seattle City reported 9412 car accidents. Therefore, it is of great interest to reduce car accidents, and it can be achieved by making citizens aware of how likely they are to be implicated on a severe car accident. However, predicting car accidents and their severity is not a task humans can accomplish when thousands of cars, pedestrians and pedalcyclists are involved in the prediction.

1.2 Problem

Data that might help to determine car accident severity might include weather condition, speeding, light condition, road condition, and how crowded a place is. This project aims to predict whether and how severe a car accident will be based on these data.

1.3 Interest

The Seattle City government is interested in reducing car accidents because it negatively impacts quality of life, and has high costs, both human and economical. Other major cities governments may also be interested in the project as they could implement it in their cities. The National Health Service may as well be interested because reducing car accidents will improve public health, and reduce health assistance costs.

2. Data

2.1 Data Source

The Seattle City car accident statistics, including weather condition, speeding, light condition, road condition, and number of involved people, can be found in one Kaggle dataset here, and it has 194, 673 events. This dataset, however, only have data since 2004 until 2020, but there is no data available before 2004. Also, only ~5% of the observations have values for the speeding feature, which is considered as one of the most important features to be included on a car accident severity prediction.

2.2 Data cleaning

There is no other dataset about Seattle City car accidents specifying speeding for enough observations to be combined or replace the dataset from here. Therefore, as the studied dataset only has speeding values for ~5% of the events, the speeding feature was discarded for further consideration.

Some of the columns that had different values to express the same were normalized. Also, any row containing null values was removed from the dataset, this is because the number of rows with null values was small and the resulting dataset still preserved a large number of samples. No normalization was required as the dataset had its numerical values already normalized. One outlier with 81 people involved in the accident was removed too.

2.3 Feature selection

After data cleaning, there were 182, 895 samples and 37 features in the data. By examining the meaning of each feature, it was determined that there was redundancy in the features and that some of them were irrelevant for the car accident severity prediction, some columns were simply used like keys.

Now, other features might have seem like redundant, for example the number of people involved in the accident, number of pedestrians involved, and the number of pedalcyclists, but there is a big difference in the probabilities of suffering serious injuries if people involved in the accident were inside of a car or not. Therefore, the different counts of different kind of people involved in car accidents were preserved.

After discarding redundant and irrelevant features or columns, 12 features were preserved, this is the list of selected features: number of people involved in the accident, number of pedestrians, number of pedalcyclists, number of vehicles, if a parked car was hit, address type, collision type, junction type, under influence of drugs, weather, road condition, and light condition.

3. Exploratory Data Analysis

3.1 The target variable

Car accident severity was a feature in the dataset, therefore, it was not necessary to calculate it. The severity code feature is the target variable and the dataset was ready to be tested with the model. The only change was that the severity code had a text datatype, then the severity codes had to be changed to a number format.

Table 1. Severity code formatting from text to integer numbers.

Code meaning	Original string datatype	New integer datatype
Unknown	0	0
Prop damage	1	1
Injury	2	2
Serious injury	2b	3
Fatality	3	4

3.2 Relationship between car accident severity and road condition

Road condition is widely accepted as one of the most important aspects that affect driving and car accidents, and that wide acceptance was contradicted by our dataset. Actually, ~67% of the car accidents happened under dry roads, and only ~25% happened under wet roads, the rest of road conditions only appear in ~8% of the car accidents.

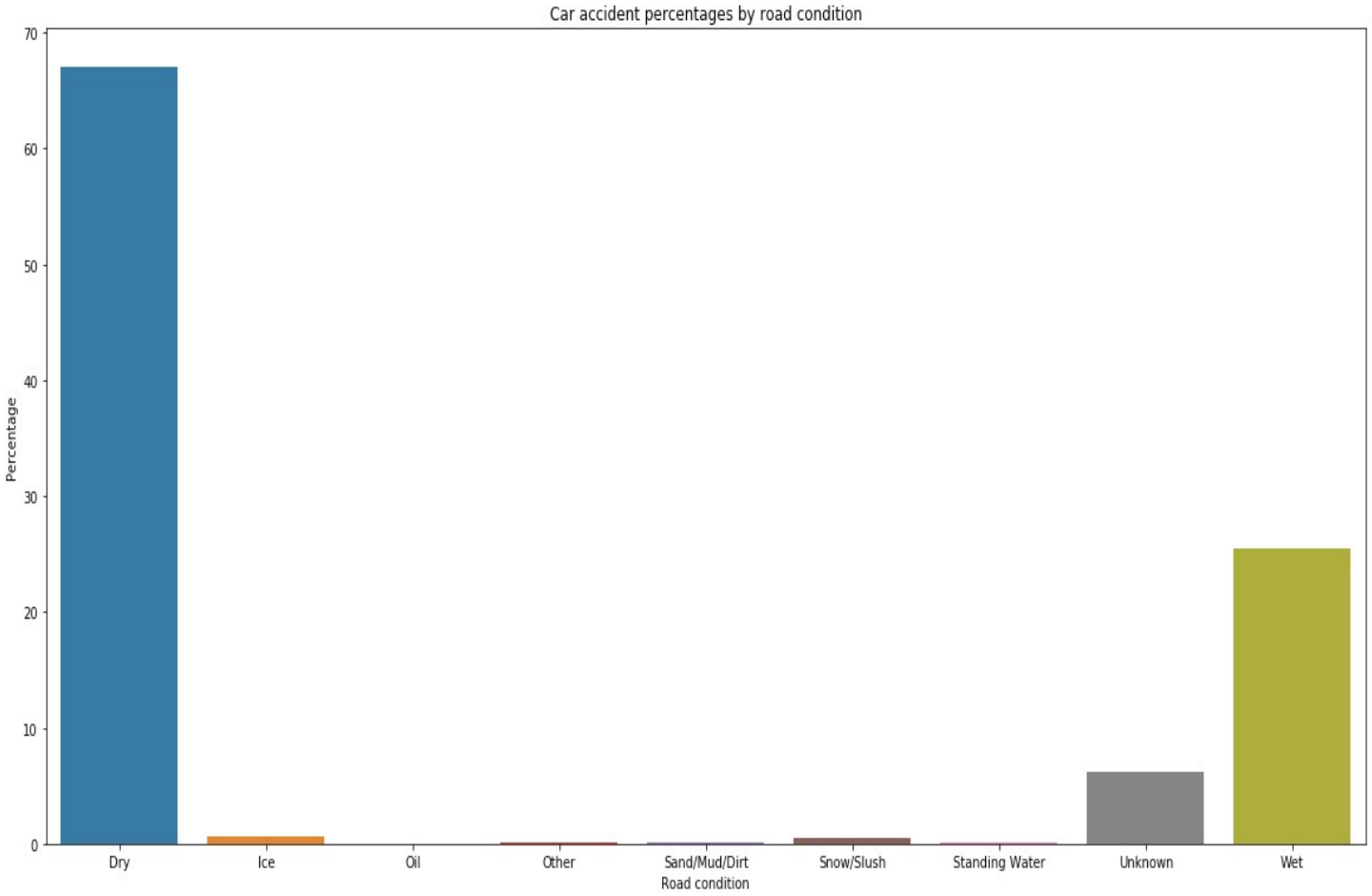

Figure 1. Car accidents by road condition

Now, considering the car accident severity, ~44% of the accidents occurred under dry roads but there were only property damages, ~21% of the accidents were under dry roads with injured people, ~17% of car accidents happened in wet roads and ended up only with property damages, and ~8% of car accidents were under wet roads with injured people. Therefore, it could be hypothesized that car accidents are more likely to occur under dry roads because dry roads stimulate speeding. However, wet roads still are considerably correlated with more severe car accidents as ~45% of car accidents under wet roads ended up with injured people. The percentages of car accidents by road conditions and severity are shown in Table 2.

Table 2. Car accident percentages by road conditions and severity.

ROADCOND	SEVERITYCODE	Car accidents	Percentage
Dry	1	83442	44.31%
Dry	2	40310	21.4%
Dry	3	2219	1.18%
Dry	4	266	0.14%
Ice	1	912	0.48%
Ice	2	269	0.14%
Ice	3	18	0.01%
Ice	4	1	0.0%
Oil	1	36	0.02%
Oil	2	24	0.01%
Other	1	82	0.04%
Other	2	42	0.02%
Other	3	3	0.0%
Sand/Mud/Dirt	1	47	0.02%
Sand/Mud/Dirt	2	22	0.01%
Snow/Slush	1	815	0.43%
Snow/Slush	2	164	0.09%
Snow/Slush	3	8	0.0%
Standing Water	1	80	0.04%
Standing Water	2	29	0.02%
Standing Water	3	2	0.0%
Unknown	1	10834	5.75%
Unknown	2	713	0.38%
Unknown	3	29	0.02%
Unknown	4	1	0.0%
Wet	0	1	0.0%
Wet	1	31355	16.65%
Wet	2	15780	8.38%
Wet	3	753	0.4%
Wet	4	69	0.04%

Also, most of the car accidents, independently of the road condition, ended up with property damages or injured people.

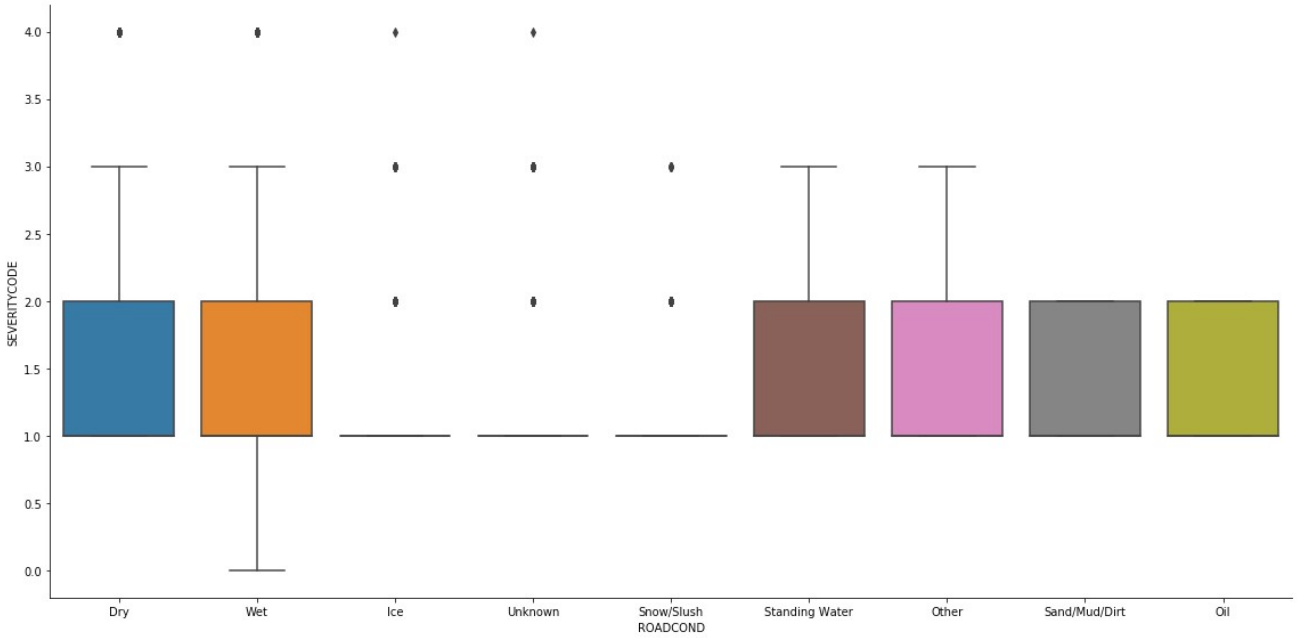

Figure 2. Car accident severity by road condition

3.3 Relationship between car accident severity and light condition

Light condition is another factor widely accepted as one of the main aspects that influence car accidents, but only ~28% of car accidents happened under dark light conditions, and ~62% were under daylight, also, ~3% were under dusk conditions.

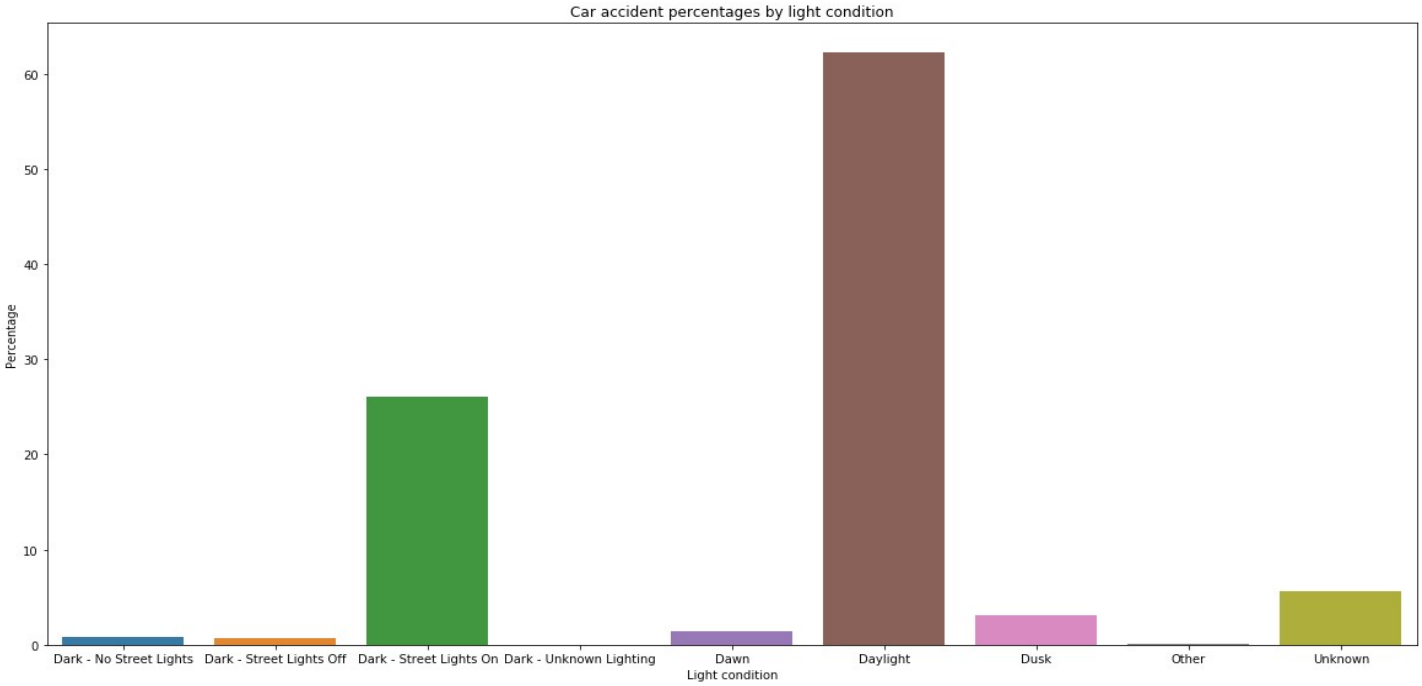

Figure 3. Number of car accidents by light condition

Now, if car accident severity is considered, along with light condition, we can see that ~8% of the accidents happened under dark light conditions ended up with injured people or a higher severity. The percentages are shown in Table 3.

Also, most of the car accidents, independently of the light condition, ended up with property damages or injured people.

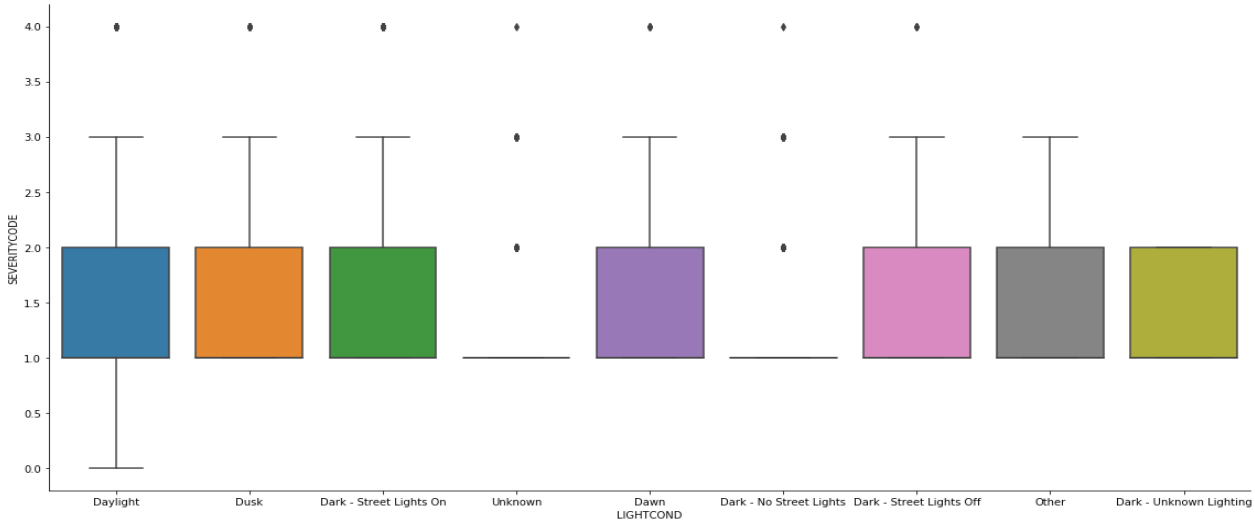

Figure 4. Severity code by light condition

3.4 Relationship between car accident severity and weather

When it comes to weather, ~60% of car accidents occurred under clear weather, but ~20% happened with raining weather, and ~17% were under overcast weather.

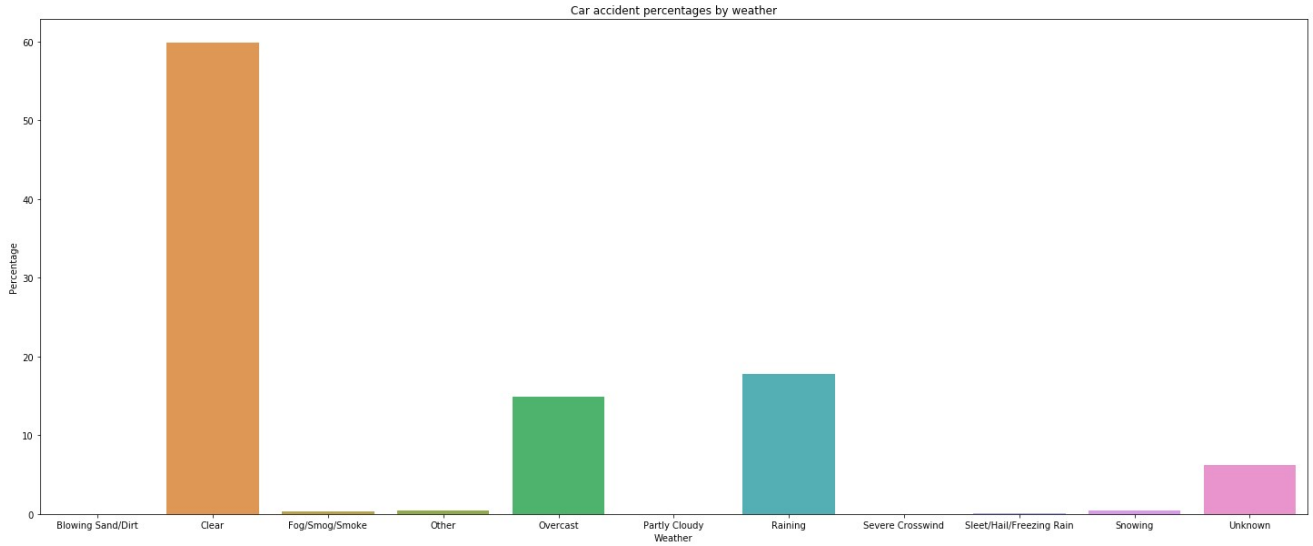

Figure 5. Car accidents by weather conditions

An interesting statistic is that, even though car accidents under overcast weather only represent ~17% of all the car accidents under study, ~5% of all the car accidents ending up with injured people happened under overcast weather. The same is seen under raining weather, ~6% of all the car accidents were under raining weather and ended up with injured people. Table 4 shows the percentages of car accidents by weather and severity.

Table 3. Car accidents by light condition and severity

LIGHTCOND	SEVERITYCODE	Car accidents	Percentage
Dark - No Street Lights	1	1144	0.61%
Dark - No Street Lights	2	334	0.18%
Dark - No Street Lights	3	25	0.01%
Dark - No Street Lights	4	1	0.0%
Dark - Street Lights Off	1	851	0.45%
Dark - Street Lights Off	2	314	0.17%
Dark - Street Lights Off	3	28	0.01%
Dark - Street Lights Off	4	4	0.0%
Dark - Street Lights On	1	33501	17.79%
Dark - Street Lights On	2	14502	7.7%
Dark - Street Lights On	3	1007	0.53%
Dark - Street Lights On	4	145	0.08%
Dark - Unknown Lighting	1	16	0.01%
Dark - Unknown Lighting	2	8	0.0%
Dawn	1	1662	0.88%
Dawn	2	832	0.44%
Dawn	3	60	0.03%
Dawn	4	5	0.0%
Daylight	0	1	0.0%
Daylight	1	76487	40.61%
Daylight	2	38784	20.59%
Daylight	3	1774	0.94%
Daylight	4	167	0.09%
Dusk	1	3883	2.06%
Dusk	2	1945	1.03%
Dusk	3	110	0.06%
Dusk	4	14	0.01%
Other	1	160	0.08%
Other	2	54	0.03%
Other	3	4	0.0%
Unknown	1	9899	5.26%
Unknown	2	580	0.31%
Unknown	3	24	0.01%
Unknown	4	1	0.0%

Independently of the weather condition, most of car accidents ended up with property damage or injured people.

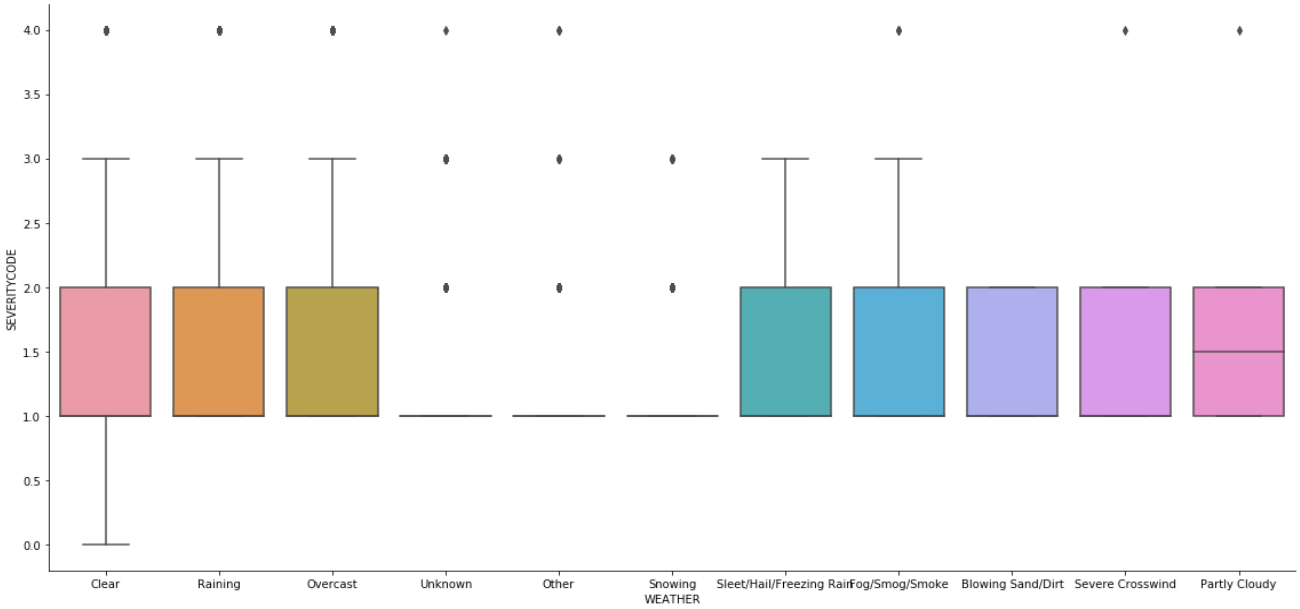

Figure 6. Car accidents by weather and severity

Table 4. Car accidents by weather and severity

WEATHER	SEVERITYCODE	Car accidents	Percentage
Blowing Sand/Dirt	1	36	0.02%
Blowing Sand/Dirt	2	13	0.01%
Clear	0	1	0.0%
Clear	1	74364	39.49%
Clear	2	36087	19.16%
Clear	3	2005	1.06%
Clear	4	225	0.12%
Fog/Smog/Smoke	1	371	0.2%
Fog/Smog/Smoke	2	187	0.1%
Fog/Smog/Smoke	3	3	0.0%
Fog/Smog/Smoke	4	3	0.0%
Other	1	645	0.34%
Other	2	118	0.06%
Other	3	7	0.0%
Other	4	3	0.0%
Overcast	1	18740	9.95%
Overcast	2	8778	4.66%
Overcast	3	442	0.23%
Overcast	4	53	0.03%
Partly Cloudy	1	5	0.0%
Partly Cloudy	2	4	0.0%
Partly Cloudy	4	1	0.0%
Raining	1	21757	11.55%
Raining	2	11193	5.94%
Raining	3	525	0.28%
Raining	4	50	0.03%
Severe Crosswind	1	18	0.01%
Severe Crosswind	2	7	0.0%
Severe Crosswind	4	1	0.0%
Sleet/Hail/Freezing Rain	1	85	0.05%
Sleet/Hail/Freezing Rain	2	28	0.01%
Sleet/Hail/Freezing Rain	3	2	0.0%
Snowing	1	716	0.38%
Snowing	2	167	0.09%
Snowing	3	10	0.01%
Unknown	1	10866	5.77%
Unknown	2	771	0.41%
Unknown	3	38	0.02%
Unknown	4	1	0.0%

4. Predictive modeling

The classification models are the best choice to predict car accident severity. Therefore, a classification was implemented in order to produce the predictions based on the previously discussed features.

4.1 Classification model

4.1.1 Applying standard algorithms and their problems

I applied support vector machines, but support vector machines are not efficient for large datasets, and the results are only slightly better than with random forests. Hence, support vector machines were not considered for this project. Also, the target variable, the severity, is unbalanced and that requires types of models that have a better performance with unbalanced labels.

Another problem is that random forests need all the features to be numbers, and most of our selected features were strings.

4.1.2 Solution to the problems

In order to handle the unbalanced target variable, an efficient model for unbalanced labels was required, and random forests are among the best classifiers when your dataset has unbalanced labels, therefore, random forests were chosen to implement our prediction.

As for the feature's datatype, all of the feature values that were not already numbers were mapped to numeric values to allow the random forest to built the model and produce the predictions.

4.1.3 Performance of the random forest model

The performance was evaluated with two metrics, the accuracy score and the confusion matrix metrics. Confusion matrix was added because that is the most common and efficient way to evaluate unbalanced labels. The accuracy score metric for the final test set is shown in Table 5.

Table 5. Performance of the random forest model evaluated with accuracy score

	Accuracy score
Random forest classifier	73.07%

The confusion matrix for the final test set is shown in Figure 7.

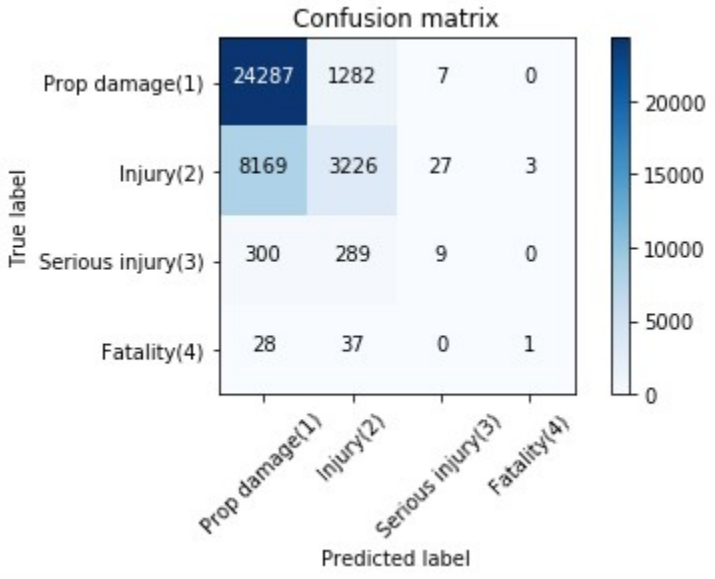

Figure 7. Confusion matrix for the test set

Based on the confusion matrix, the model predicted 95.38% of the “property damage” severity car accidents correctly, 28.03% of car accidents were correctly predicted with an “injury” severity. However, the model failed to correctly predict 98.57% of “serious injury” car accidents, and 98.53% of car accidents that ended up with fatalities. Actually, many of the car accidents that had an injury severity were misclassified with a “property damage” severity, which may well be because the model is biased towards the “property damage” severity.

5. Conclusion

In this study, I analyzed the relationship between car accident severity and the external conditions in Seattle City roads. The features that were identified as the most important features affecting car accident’s severity are road condition, light condition, weather condition, number of pedestrians involved, and number of pedalcyclists involved. I built a classification model to predict car accident’s severity going from property damages, to fatalities. This model can be very useful in helping governments to reduce car accidents and their severity as it predicts how likely a severe car accident is under specific external conditions and advice drivers to be more careful when those external conditions are present, as well as increasing traffic police controls.

6. Future directions

Even though, the model has a ~70% accuracy, it failed to correctly classify most of the car accidents with serious injuries or fatalities, and it misclassified many of the car accidents with an injury severity as property damages severity.

Another point for improvement is that the dataset did not have information about speeding for most of the car accidents, and speeding is widely accepted as one of the most important factors in car accidents severity.