Seattle Traffic Accident Severity Prediction

Narcís Gironès Sancho

16 September 2020

I. Introduction

1. Background

Do you live in a safe city? Can your children walk safe and free? Is Seattle a city for tourists? In the past 5 years, Seattle has received many complaints about unsafe neighborhoods, most of them related to car accidents. In addition, the big car rental companies complain that last quarter, tourists no longer rent cars since Seattle is one of the cities with the most car collisions in the United States. With all of this in mind, the top priority for Seattle's municipal government is to ensure that it can reduce the large number of accidents and return to being the safe city it was before.

2. Business Problem

The objective of this capstone project is to analyze and study the collision dataset for the city of Seattle, Washington and find patterns and determinate key factors such as weather, visibility, and road conditions to create the best traffic accident severity prediction. It will use various analytical techniques and machine learning classification algorithms such as k-nearest-neighbors, Support Vector Machine, etc.

3. Target Audience

We can provide the results of this analysis to government departments, car rental companies, insurance companies and emergency services, and make recommendations to reduce accidents.

II. Data

1. Source

The data that will be used to conduct the study is based on the collection of traffic accidents from 2004 to present in the city of Seattle, Washington. Provided by the International Business Machines (IBM) here.

2. Metadata

The dataset has 194,673 instances (rows) with 38 features (columns) of traffic accidents in Seattle. Each collision has the attribute of a severity code that it can be:

3—fatality
2b—serious injury
2—injury
1—property damage
0—unknown

More about the metadata can be found here.

III. Methodology

1. Data Analysis

How can we know which are the best features to most accurately determine the severity code prediction? To do this, we must analyze which feature has a correlation with the collision severity code using this method:

```
dataframe.groupby([feature])['SEVERITYCODE'].value_count()
```

In addition, a new column with the day of the week of the collision has been created to find out if there are more accidents on the weekend than during the work week.
Once we have what features we are going to use, we will proceed to the next step.

2. Data Encoding

Before we start to process and model the data, we must know what type of data we have. That is, if the data are of type numeric or object.
Many of the Machine Learning models work with numerical data (int or float), and that is why we must convert our data set to a numeric type, using this method with an example:

```
df['ADDRTYPE'].replace(to_replace=['Alley','Block','Intersection'],
                                        value=[0,1,2],inplace=True)
```

3. Feature Set and Normalization

Now we are ready to create our feature set to work with our Machine Learning models. Our feature set consists of 14 columns, including features such as weather condition, road condition, and light condition.

```
X=features[['ADDRTYPE','JUNCTIONTYPE','COLLISIONTYPE','VEHCOUNT','PEDCYLCOUNT',
'PERSONCOUNT','PEDCOUNT','SDOT_COLCODE','ROADCOND','LIGHTCOND','WEATHER','
PEDROWNOTGRNT','SPEEDING','HITPARKEDCAR']]
```

Finally, we need to remove the NaN data and normalize our dataset to avoid any bias caused by the different scales of each characteristic.

```
    features.dropna(inplace=True)
    preprocessing.StandardScaler().fit(X).transform(X)
```

Then, we can proceed to train our models.

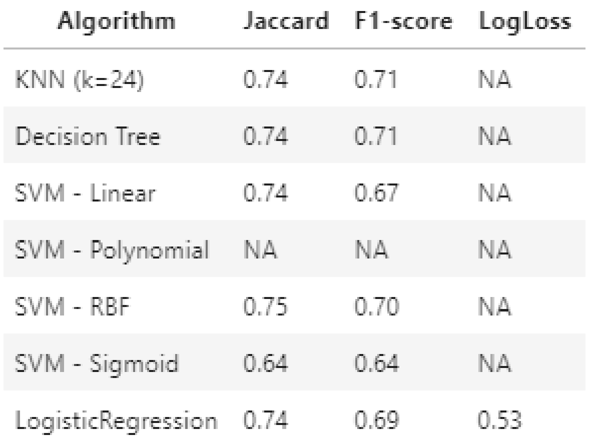
4. Machine Learning Model

All possible Machine Learning classification models have been studied since it is about classifying which severity code the accident has, such as k-Nearest Neighbors, Support Vector Machine, Decision Tree and Logistic Regression.
To create and test the models, our feature set must first be divided into a training set (80%) to train the models and test set (20%) to test the models, using this method:

```
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2,
                                                     random_state=4)
```

IV. Result

After training and testing all the models, the Decision Tree achieved the highest accuracy with 07158 and Jaccard index with 0.7455, however the differences in accuracy between models were small.

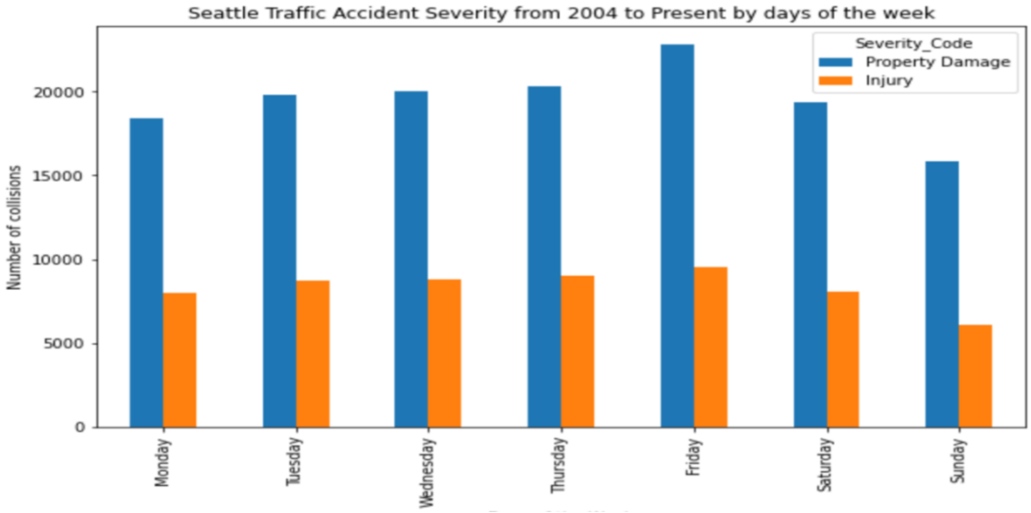


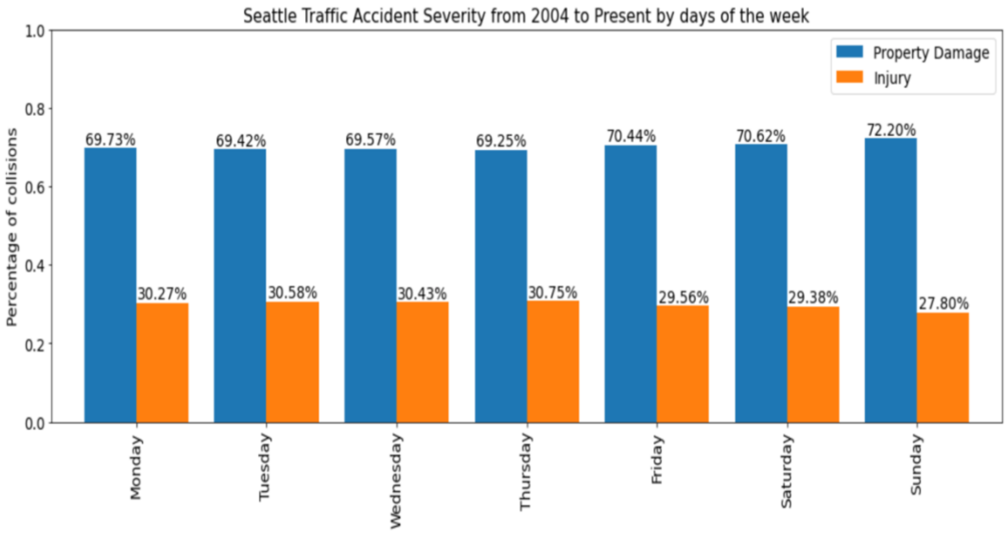
Algorithm	Jaccard	F1-score	LogLoss
KNN (k=24)	0.74	0.71	NA
Decision Tree	0.74	0.71	NA
SVM - Linear	0.74	0.67	NA
SVM - Polynomial	NA	NA	NA
SVM - RBF	0.75	0.70	NA
SVM - Sigmoid	0.64	0.64	NA
LogisticRegression	0.74	0.69	0.53

Computing time was the important characteristic in deciding which model is the best, because k-Nearest Neighbors and Decision Tree had the same result with accuracy and Jaccard index. Therefore, I recommend the Decision Tree model for implementing the Traffic Accident Severity.

V. Discussion

After finishing this project, I have realized how important methodology is for Data Science, especially the first three phases (from problem to approach, from requirements to collection, and from understanding to preparation). It is important to invest time in these three initial phases for building your model to have a successful outcome.
When I was preparing the data, I wondered if there had been more accidents on the weekend than during the work week (M-F). Surprisingly there is no clear difference between number or severity of weekend and weekday traffic accidents

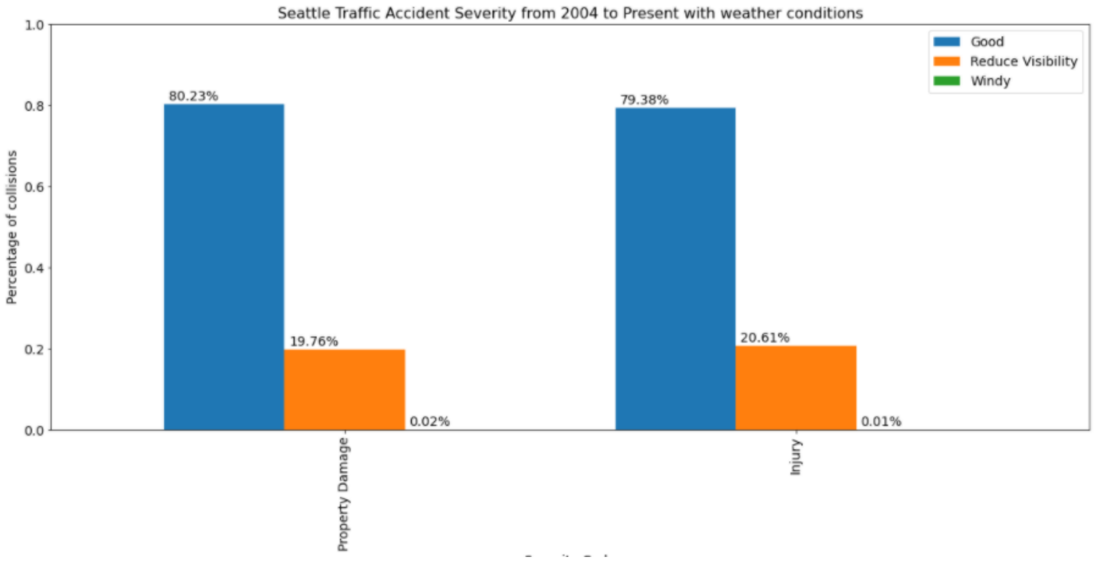




One theme on my mind was, how I can find pattern to prevent accidents using this data set? I analysed the features of the weather, the light conditions and the road conditions in case I found an anomaly that causes more accidents than normal. But the results have not been as expected. There have been more accidents in good circumstances than in adverse conditions*.*

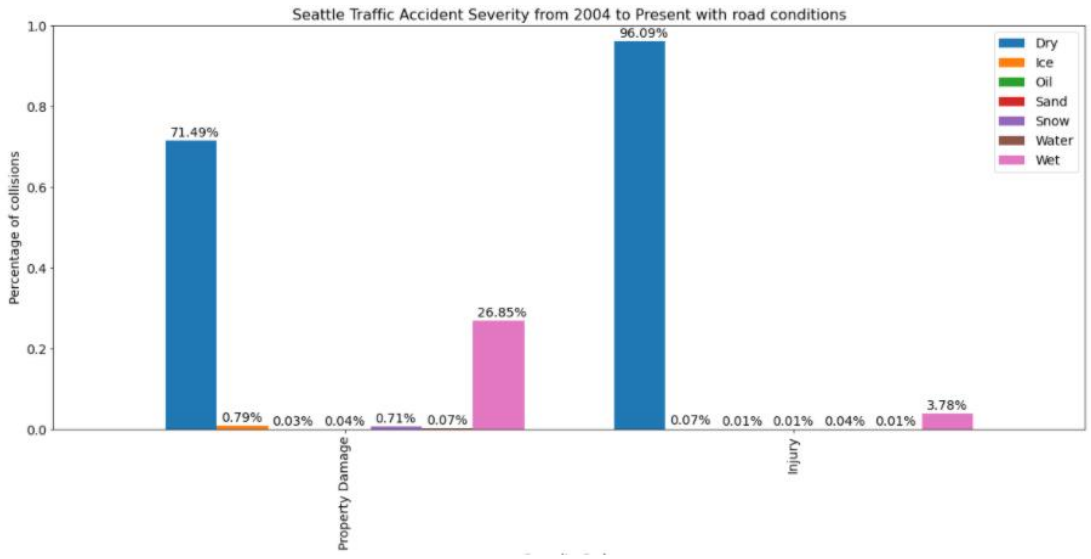
- Weather condition:

The weather conditions are classified:
Good: "Clear", "Overcast", "Partly Cloudy"
Reduce Visibility: "Raining", "Snowing", "Fog/Smog/Smoke", "Sleet/Hail/Freezing Rain", "Blowing Sand/Dirt"
Windy: "Severe Crosswind"



As shown in the weather condition graph, the set of accidents with the "Reduce Visibility" type could be studied to obtain the necessary information to incorporate speed or visibility signs on the road.

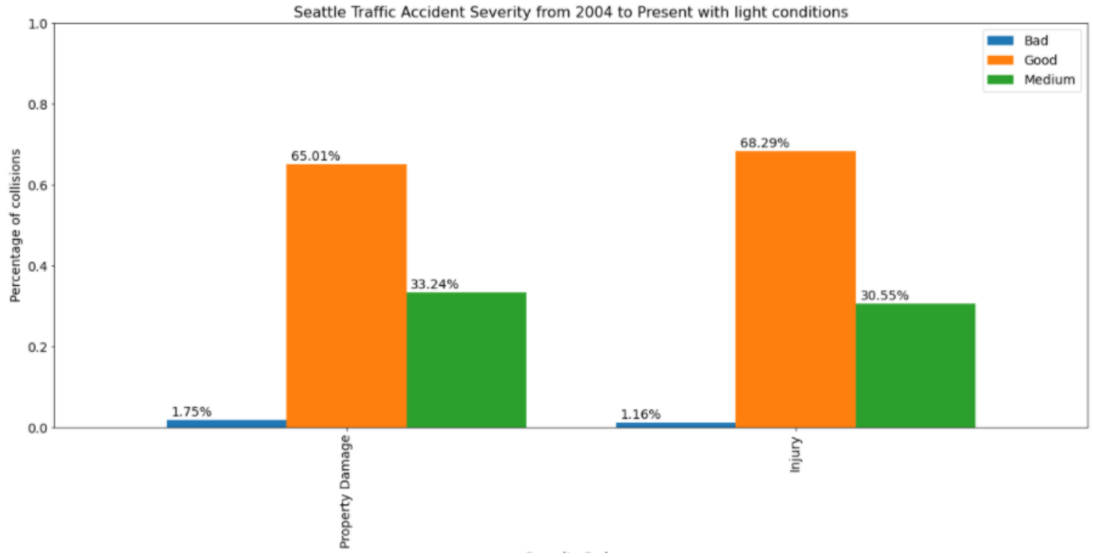
- Road condition:



Based on the graph, one could analyze the accidents suffered on wet roads in case of finding a pattern and being able to prevent accidents by incorporating traffic signs.

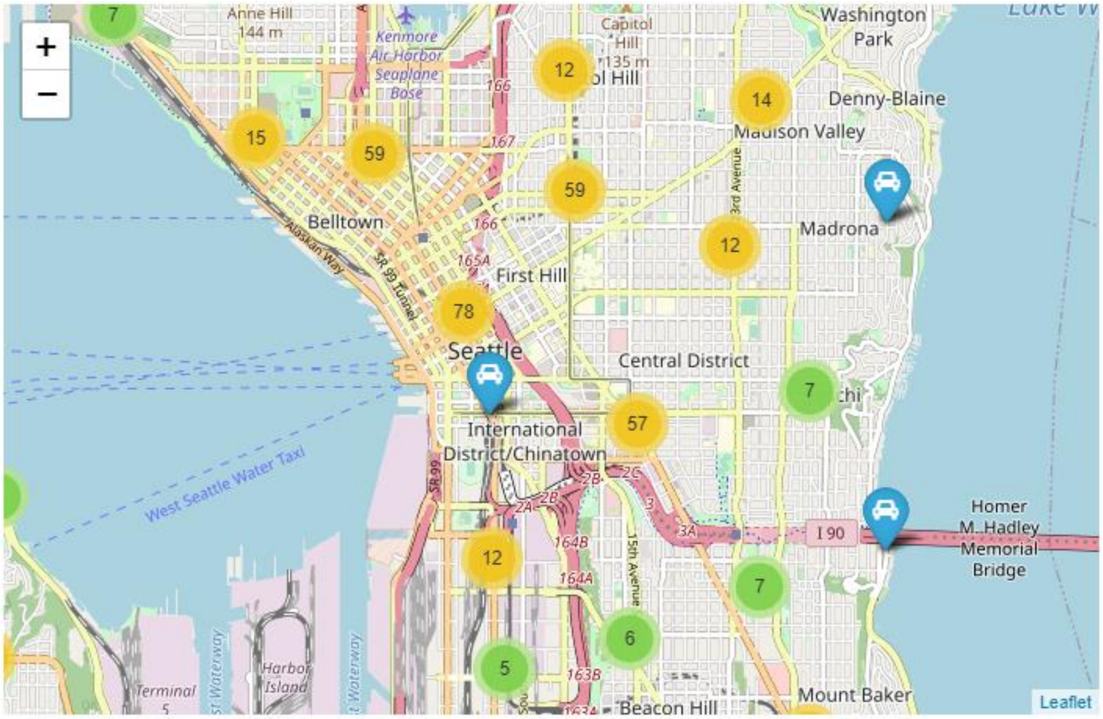
- Light condition:

The light conditions are classified:
Bad: "Dark - No Street Lights", "Dark - Street Lights Off", "Dark - Unknown Lighting"
Good: "Daylight"
Medium: "Dark - Street Lights On", "Dusk", "Dawn"



As a future point, the accidents produced by the "Medium" type of light condition could be studied in depth in case more road lights could be renewed or added.

Using the package ‘folium’, it is possible to visualize the accidents that occurred in Seattle on a map and obtain the possible high risk points.



Due to the high compilation time, the study was only possible with the first 1000 accidents. Although some areas on the map have a higher number of collisions, no specific sections of road were identifiable as a high-risk points.

In the future, using a high SPEC computer, this type of analysis could be a valuable source of information for city planning and locating areas for infrastructure improvements. This type of data has the capacity to prevent accidents and help Seattle become a safer city to drive in.

Based on these observations and analysis, more caution and speed limits should be enforced during reduce visibility and wet conditions since they are the second lead cause of car accidents. Adding new signs and lights will be very helpful in cautioning drivers at dusk, dawn and at night.

VI. Conclusion

This study used machine learning algorithms to analyse traffic accident data from 2004 to the present for Seattle, Washington in an effort to identify predictors of traffic accident severity. Based on the results of this analysis, wet conditions contribute to about one-quarter of traffic accidents resulting in property damage, while 96% of accidents resulting in injury occurred in dry conditions. This finding highlights the importance of infrastructure maintenance and speed enforcement rather than weather conditions as the biggest contributors to road accidents in Seattle.

Another important finding from this analysis is that the majority of accidents occurred under 'Good' or 'Medium' lighting conditions, with just about 1% of accidents occurring under 'Bad' lighting conditions. Reduced visibility conditions were present in equal proportions of accidents resulting in property damage and injury. Further analysis of the most high-risk areas is needed to assess street lighting and signage to improve traffic safety and prevent future accidents.