Predicting Severity of Seattle Collisions

Miguel Enrique Játiva Jiménez

September 24, 2020

Contents

List of Figures

1 Introduction

The **SDOT Traffic Management Division** stores all the traffic collisions that take place in Seattle. It would be useful to predict the severity of the collision given the context, this could be of interest to the Health Care System since it would allow it to assign the right amount of resources to each collision depending on the severity predicted.

2 Data acquisition and cleaning

2.1 Data sources

The data that will be used for this task is obtained from the Seattle Government more specifically from **SDOT Traffic Management Division**. This data set contains very useful information as the amount of vehicles involved in the collision, the number of people, the weather or road condition. The data can be obtained here.

2.2 Data cleaning

The data set has 194673 samples and 38 features. First of all, I checked the amount of missing values on each feature. Depending on the amount of missing values a decision was made, e.g., when the amount is close to the total number of samples, the feature is dropped, but, if the amount is small in comparison to the total number of samples, the missing values can be filled with the most frequent one in that column.

It is worth mentioning that some features like **Junctiontype** or **Weather** have a value named *Unknown*. These values have been taken into account as if they were missing values and the corresponding solution was applied.

2.3 Feature selection

After data cleaning there were 30 features left. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. For example, there was a feature named **Severitycode** and another named **Severitycode.1**, due to the similarity in the names I decided to check if both of them were equal. They happened to be equal so one of them was dropped. Another similar case is the one between **IncDate** and **IncDTTM**, the first one stores the date of the collision and the second one stores the date and time, therefore the first one was dropped.

There were also a group of features that seemed like they were Id features, i.e., **ObjectId**, **IncKey**, etc. To make sure that they were features that uniquely identify each sample, I checked the amount of unique values on them, therefore if the amount of unique values is the same as the total number of samples in the data set, it means that the feature is a uniquely identifier so it has to be dropped.

There were also some features containing descriptions that could be useful for visualization purposes but were dropped because they do not add value to the prediction. Another problem was found in the **UnderInf** feature, it had 4 unique values, i.e., *N, 0, Y, 1*. It is clear that the values of this feature should be *Yes* or *No*, represented as *N, Y* or *0, 1*, but not both ways, so every *0* was transformed to *N* and every *1* to *Y*.

3 Methodology

3.1 Exploratory Data Analysis

3.1.1 Evolution of the number of collisions trough years

The evolution of collisions trough years has been studied calculating the number of collisions each year and then plotting all together. As it is seen in Figure 1 since 2005 the number of collisions has been decreasing until 2013, when it suddenly increases till 2015, but, since 2015 the number has been decreasing. It is interesting to see that 2020 is the year when the number of collisions is the smallest, this may be due to the pandemic situation we are living, and we can take into account that there are still 3 months left to conclude 2020.

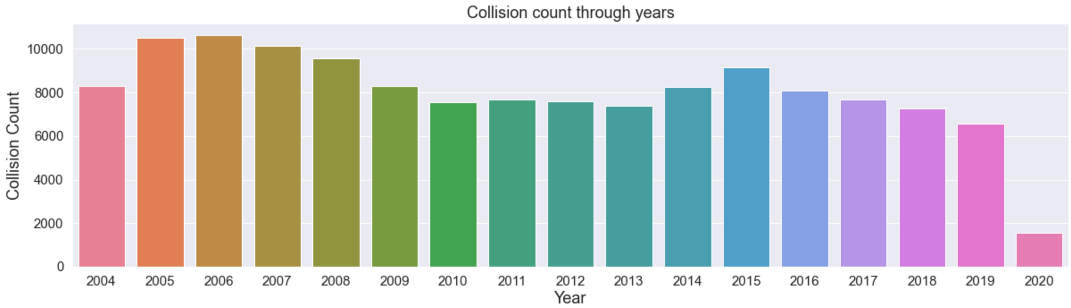

Figure 1: Collision count through years

3.1.2 Collisions given the time of the day

It was also studied how the number of collisions is spread over the day. It is observed in Figure 2 that the number of collisions increases from the **Early Morning** to the **Noon** *(12 PM - 16 PM)* when it hits its maximum value, but then it decreases until **Night** *(20 PM - 0 AM)*. It is seen a rough increase from **Night** to **Late Night** *(0 AM - 4 AM)* maybe due to the influence of alcohol/drugs.

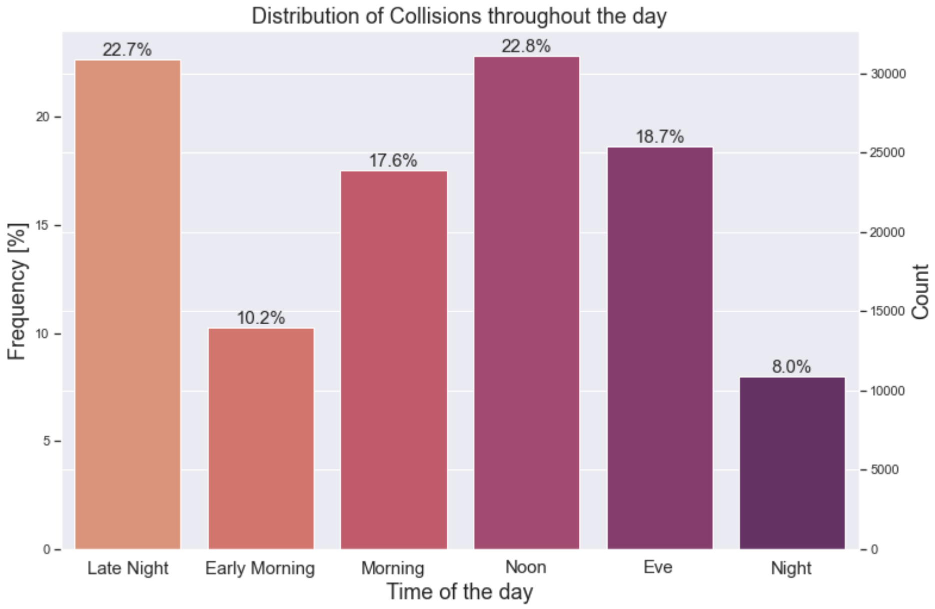

Figure 2: Distribution of collisions throughout the day

3.1.3 Collisions of people influenced by alcohol/drugs given the time of the day

The distribution of collisions when people are influenced by alcohol/drugs has been studied. As it is seen in Figure 3, the number of collisions increases from **Early Morning** until **Late Night** when it achieves its peak. It was expected that this would be the behavior of the data since alcohol/drugs are mostly consumed at **Night**/**Late Night**.

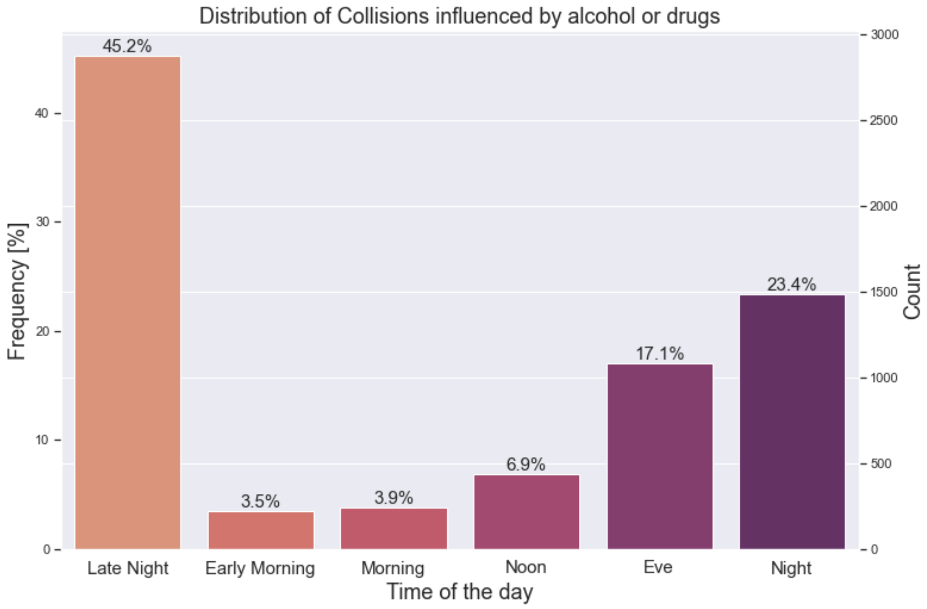

Figure 3: Distribution of collisions when influenced by alcohol/drugs

3.1.4 Collisions visualization using Seattle map

Using the **Folium** library, I created a Seattle map to visualize the collisions, using the **X** and **Y** features that represent the **Longitude** and **Latitude** respectively. Each collision location was represented as marker in the map. A marker cluster was used to group the markers depending on the zoom of the map. Looking at Figure 4 we can see that most of the collisions are located at the center of Seattle.

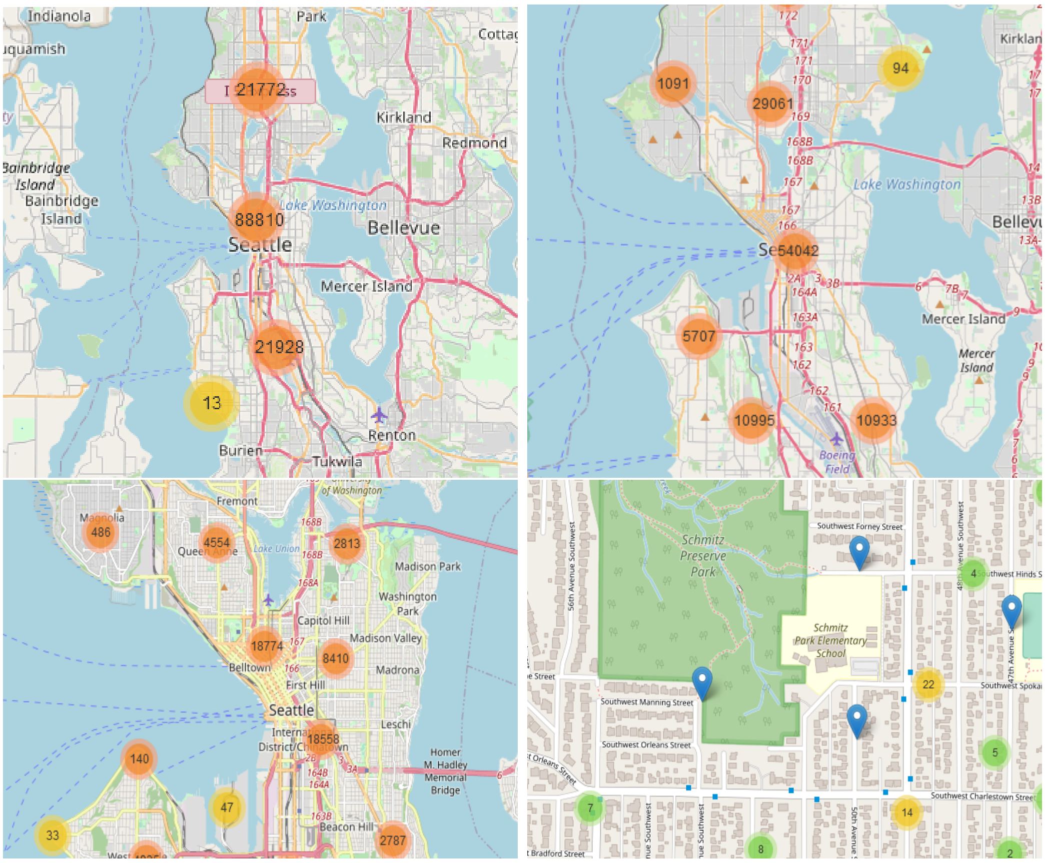

Figure 4: Marker cluster on Seattle map

3.2 Predictive Modeling

The problem faced is a classification one. The models need to predict the severity of the collision which is measured using 5 different classes. The data set used only contains samples with 2 of those 5 classes so a future improvement is to merge another data set more balanced with the actual one.

3.2.1 Decision Tree

It is observed that the tree model does a good job classifying the **Prop damage** samples, 40506 (TN, if we take the Prop damage class as negative) but its behavior is not good when classifying **Injury** samples, it classifies 14160 samples as **Prop damage** when they really are **Injury** (FN).

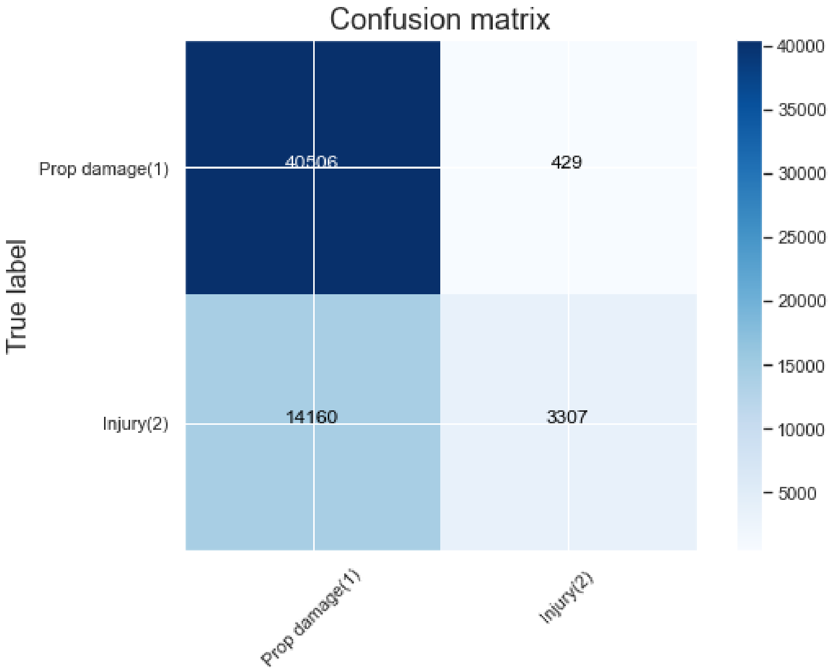

Figure 5: Confusion matrix Decision Tree

3.2.2 K Nearest Neighbors

The key to make a good K Nearest Neighbors model is to select a good value for K. For this task a loop was created, inside it, a model is created and its accuracy is calculated, this is done using numbers from 1 to 20 as K. These models were created using fit on the training data and calculating the score on the validation data. As it is seen in Figure 6 the accuracy increases along with the value of K. The K that provides the best accuracy is selected, in this case $K = 16$.

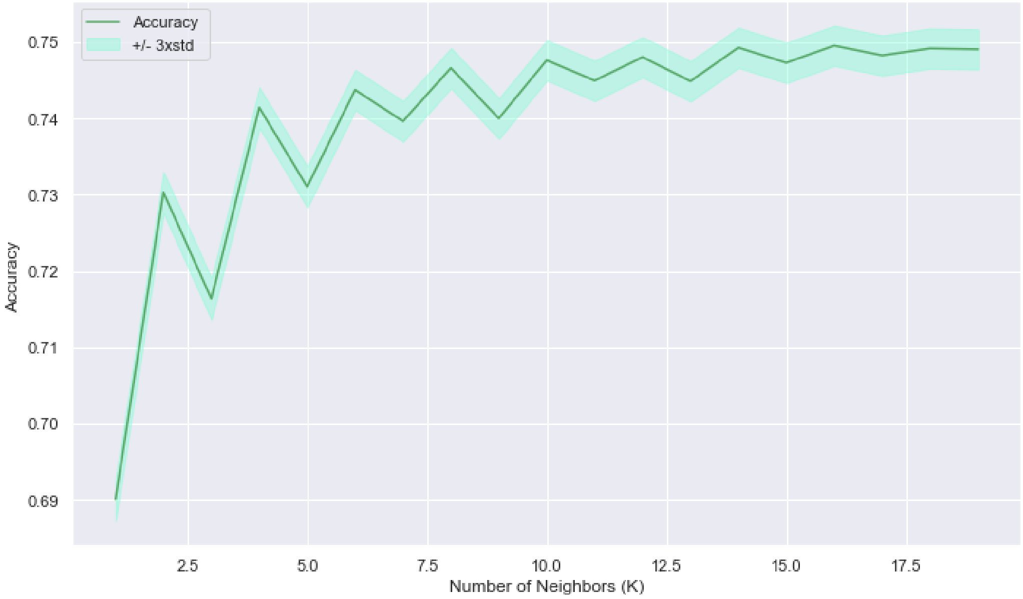

Figure 6: Accuracy for each K

It is seen that the behavior is similar to the one that the tree model offers (Figure 5), but we can see that the kNN model does a better job classifying the **Injury** class (Figure 7).

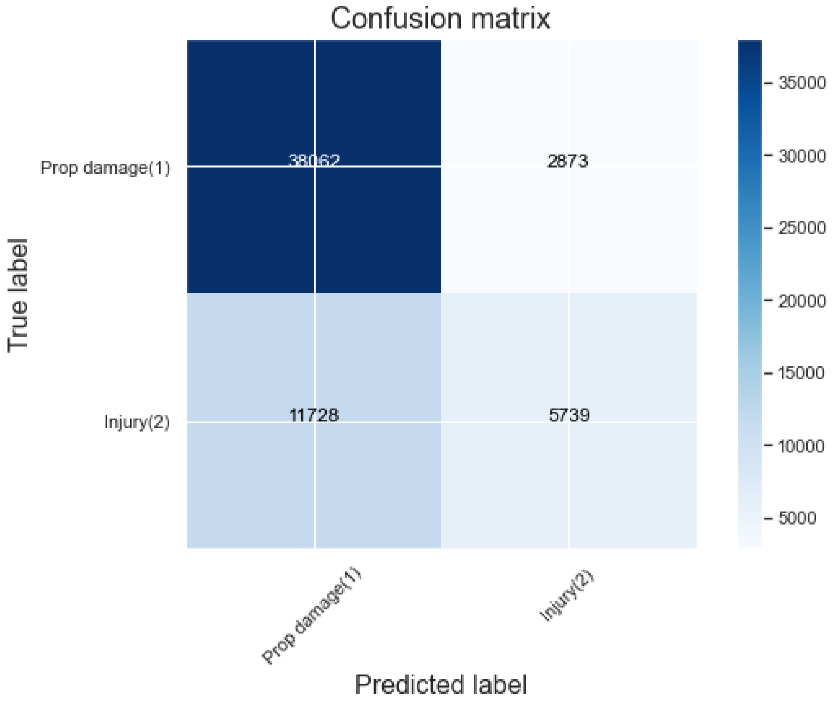

Figure 7: Confusion matrix kNN

3.2.3 Logistic Regression

We can see that the behavior is similar to the ones offered by the tree and kNN models (Figure 5 and Figure 7) but this model is the worst when classifying the **Injury** class.

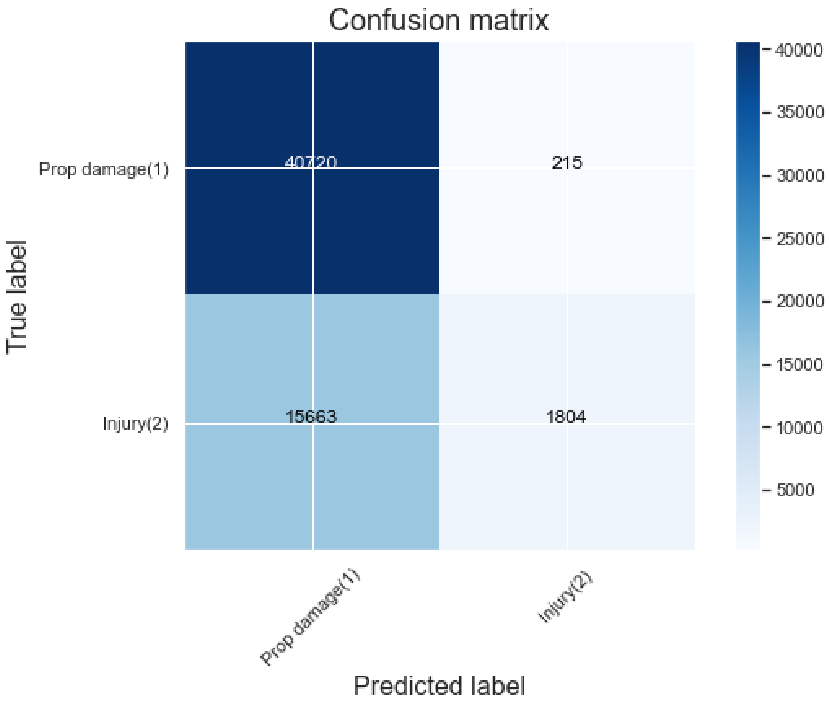

Figure 8: Confusion matrix Logistic Regression

4 Results

To obtain the following results the models mentioned above made predictions on the test data. The metrics used are *Accuracy, Jaccard, F1-score* and in the case of **Logistic Regression** *LogLoss* is used. The results are summarized in Table 1.

Table 1: Results table

	Accuracy	Jaccard	F1 - score	LogLoss
Decision Tree	0.750197	0.735203	0.687250	NA
K Nearest Neighbors	0.749991	0.722747	0.719748	NA
Logistic Regression	0.728126	0.719460	0.641937	0.672446

5 Discussion

Having a look at Table 1 we can say that the model that obtains the best accuracy is the **Decision Tree** model, but, if we take a look at the different confusion matrices (Figure 5, Figure 7 and Figure 8) we can say that the most consistent model is the **K Nearest Neighbors** model, because it does a good job predicting the *Prop damage* class and in comparison with the other models, it does a better job predicting the *Injury* class.

The models overall do not make a good job on predicting the *Injury* class, this may be due to the imbalanced data set we are working on. In our data set 70% of the samples belong to the *Prop damage* class and 30% belong to the *Injury* which is the class that the models struggle the most when predicting. This could be potentially solved with a more balanced data set.

6 Conclusion

In this study, I analyzed the evolution of the amount of collisions through years. I analyzed how the collisions distribute through the day and how being influenced by alcohol/drugs affects to collisions. I also created a map to analyze the locations where most of the collisions happen. I built classification models to predict the severity of a collision. These models can be very useful in helping the Health Care System decide the amount of resources that have to be assigned to that collision. This would make the Health Care System more efficient.