Predicting Car Collision Severity

Roopika Palukurthi

October 5, 2020

1. Introduction

1.1 Background

There are 6.734 million crashes that occurred in 2018. Over 2.7 million people have been injured resulting in over 36,000 fatalities. Avoiding such accidents would not just save lives but also would mitigate the cost of damages to people and property. One way to avoid such collisions is to be able to intimate the driver of possibility of a crash and its severity so that the driver can take appropriate actions to avoid a collision. We will look at how we can use the data on external and internal factors that can predict an accident and its severity.

1.2 Problem

Data that might contribute to determining the probability of a collision and it's severity might include external factors like the weather, road conditions and light conditions and some internal factors like the location and it's address and junction type of the drivers current destination or the whether the driver is speeding.

1.3 Target Audience

There are multiple stakeholders for such a project. The most important being the drivers themselves who would have the advantage of not getting injured or even face death in a situation of collision. The transport department or the traffic control department whose job will become less burdensome if the drivers themselves can avoid accidents. Finally, the auto-insurance companies that would reduce their costs by avoiding payment for the damaged vehicles by avoiding the collisions.

2. Data acquisition and cleaning

2.1 Data sources

We will use the data provided by Seattle Police Department on all collisions and their traffic records. The data is on all collision types taken place in Seattle since 2004 to present date. The data contains information on each collision along with external conditions as well as information on violations that may have caused the collision. The metadata is provided at this location to understand the data. The data set however lacks the data on when an accident did not occur, hence

we can only train our model on the severity of an accident assuming it will occur. The data also contains only two severity types 1 for property damage and 2 for injury. But the metadata dictionary contains other values which are not represented in the data.

2.2 Data cleaning

Following issues were encountered while observing the data:

- Missing values – There were three types of columns with missing values
    - Columns with a large set of missing values – such columns were dropped as it is not feasible to replace missing values in columns with over 50% missing values and doing so would have skewed the data
    - Columns with a small set (<3%) of missing values – since most of these columns were categorical, missing values were replaced by most frequent value of the column. If there were no significant most frequent values, rows with missing values were dropped
    - Columns with value either 'Y' or blank – assuming the values is only filled if it is 'Y', missing values were replaced with 'N'
- There were four rows which were duplicative based on the unique ID which were dropped
- Duplicate columns like SeverityCode.1 are dropped as they do not contribute to the modelling
- Location specific data such as location description and X,Y coordinates are also dropped as they would be a hindrance in making the model generalized.
- Another attribute i.e. the incident date time has also been separated into month, day of week and hour of day to analyze how it affects the probability of Severity 1 or Severity 2 defect

Another significant issue was the imbalance between records with Severity 1 and Severity 2. The records consisted of 130,549 Severity 1 rows and 57,786 (<30%) of Severity 2 rows. Severity 2 rows have been up-sampled to have 130,549 rows in order to balance the data.

2.3 Feature selection

After data cleaning, there were a total 261,098 samples and following 31 features in the data. Of these features many were dropped that provided information that indicated post-collision situation, hence cannot be used to predict collision before it takes place. The features include COLLISIONTYPE, HITPARKEDCAR, INJURIES, FATALITIES among many others. These are more causal to predicting the severity of the collision than correlated. In additional to the causality, many of these features cannot be measured prior to the collision.

After removing the causal attributes, we are left with following 9 features. The description are reference in the metadata file.

Table 1. Feature Selection

Features	Description
WEATHER	The description of the weather conditions during the time of collision
ROADCOND	The condition of the road during the time of collision
LIGHTCOND	The lights conditions during the time of collision
SPEEDING	Whether the driver was speeding at the time of collision
JUNCTIONTYPE	The type of junction where the collision took place
HOUR	The hour within the day at which time the collision took place
DAYOFWEEK	The day of week when the collision took place
MONTH	The month in which the collision took place
ADDRTYPE	The type of address where the collision took place

3. Exploratory Data Analysis

3.1 Calculation of target variable

The target variable we are trying to calculate in SEVERITYCODE which can take up to following 4 different values as per the metadata file –

- **3 -** fatality
- **2b -** serious injury
- **2 -** injury
- **1 -** prop damage
- **0** - unknown

The data we have only represent two values, **1** (prop damage) and **2** (injury). This is a limitation. For the sake of analysis we will assume Severity 2 encompasses 2, 2b and 3 codes else they will be considered Severity 1.

We can analyze every feature variable against SEVERITYCODE to analyze the affect on the same. Since all the features are categorical variables, we are restricted in using many of the visualizations.

3.2 Relationship between severity and address type

While plotting the normalized frequency of Severity 1 and 2 collisions by the Address type – it is clear that the probability of a severity 2 collision is higher at intersections and lower otherwise as indicate in Figure 1 below.

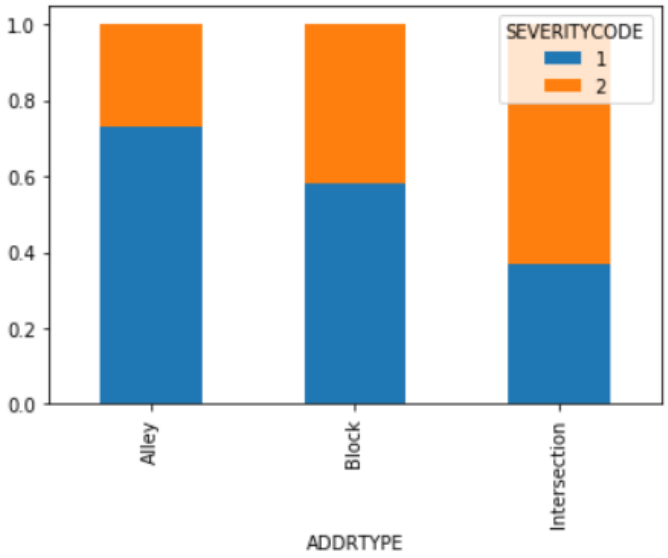

Figure 1. The share of Severity 1 and 2 collisions across different address types

3.3 Relationship between severity and junction type, road condition and light condition

If you observe each variable independent, there is no significant relationship between severity and attributes like junction type, road condition and light condition as seen in Fig 2 below

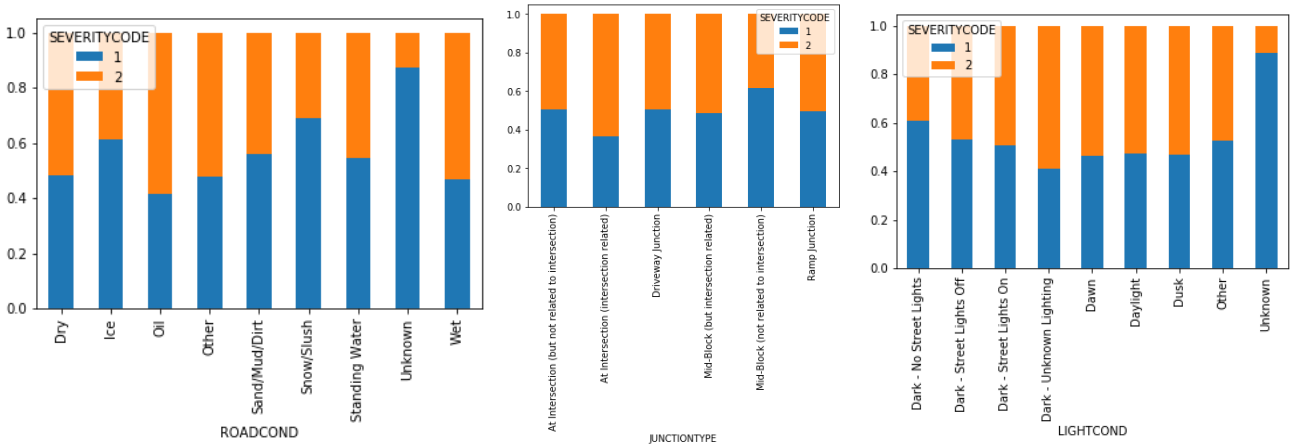

Figure 2. The share of Severity 1 and 2 collisions across different road conditions (left), junction types(center) and light conditions (right)

But when you combine them, we observe that there are certain combinations where one severity has much higher share than the others. Some of the Severity 2 dominated combinations and Severity 1 combinations are indicated by the appropriately colored arrows

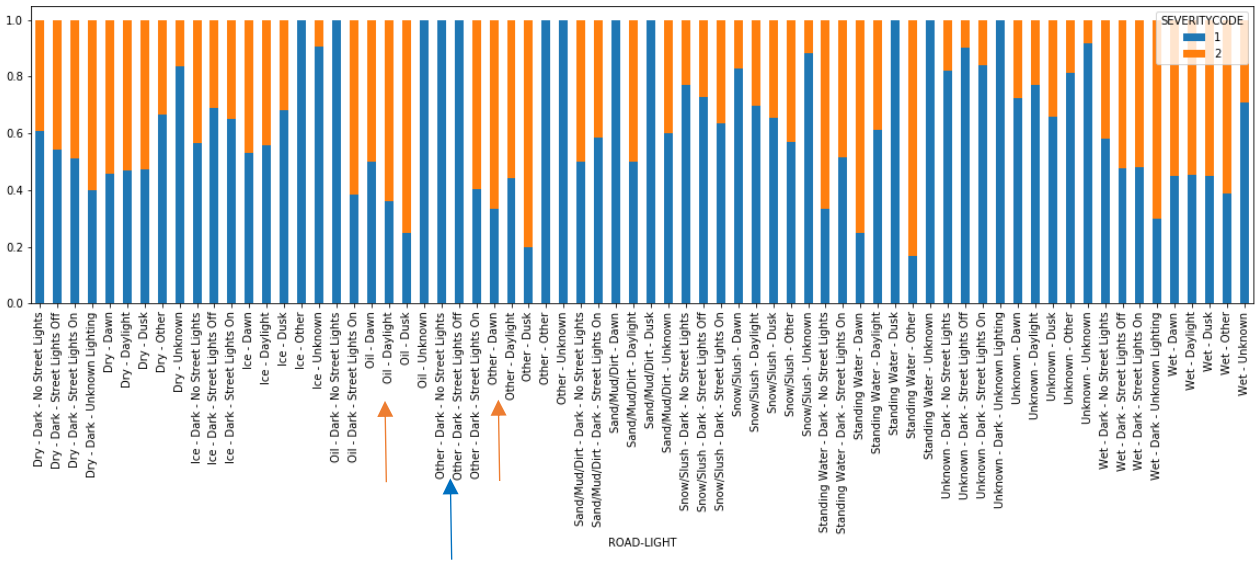

Figure 3. The share of Severity 1 and 2 collisions across different combinations of road conditions and light conditions

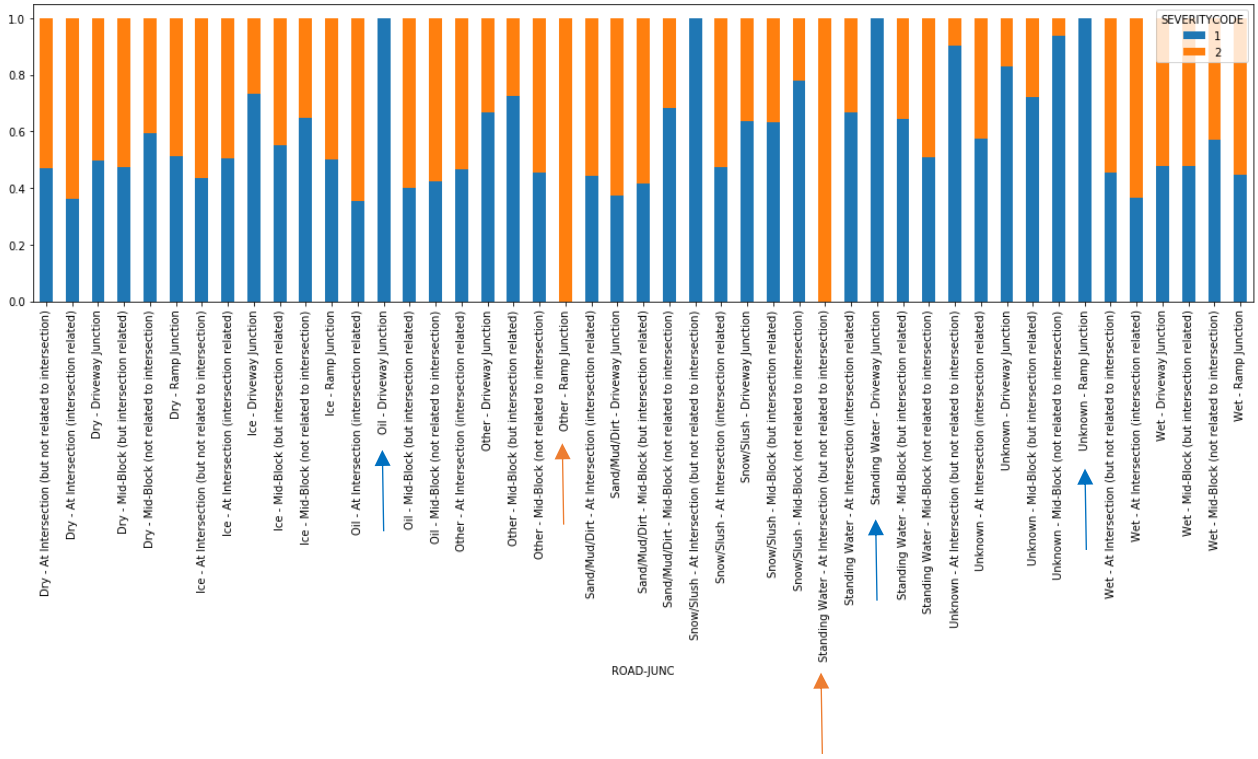

Figure 4. The share of Severity 1 and 2 collisions across different combinations of road conditions and junction types

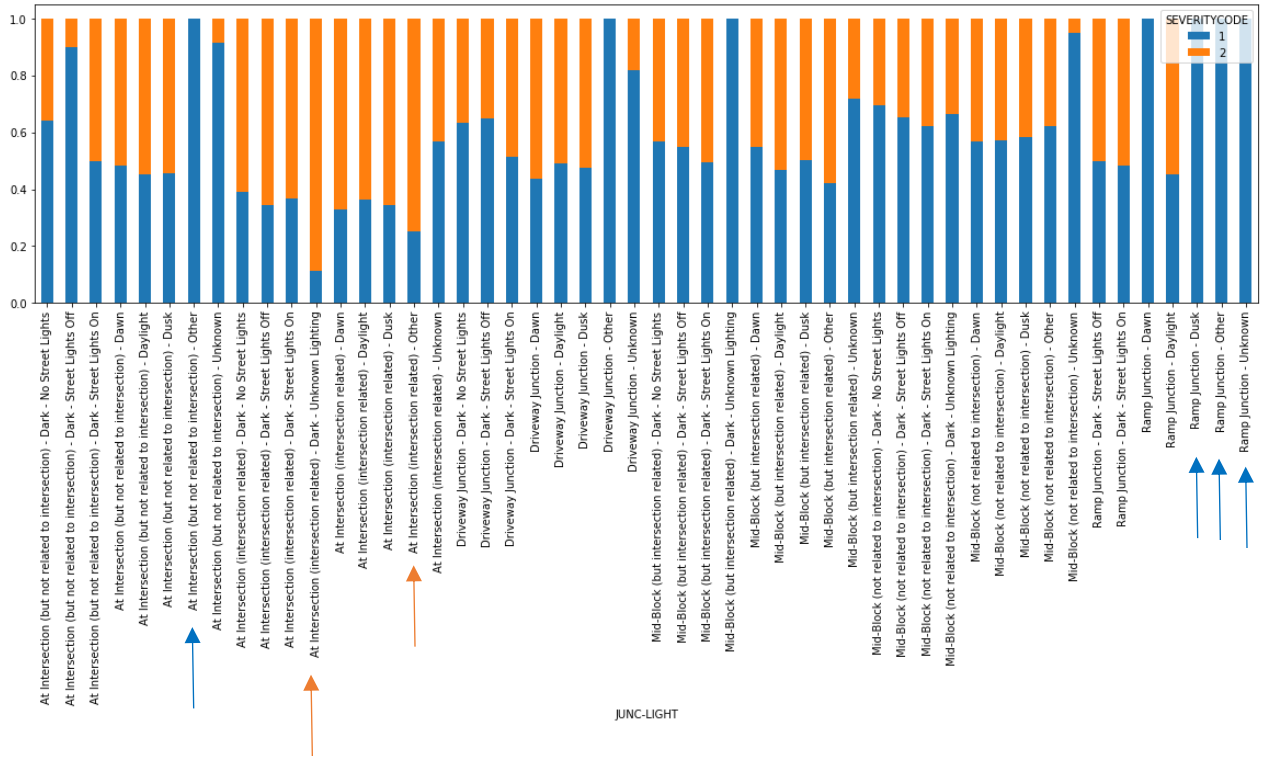

Figure 5. The share of Severity 1 and 2 collisions across different combinations of junction types and light conditions

3.4 Relationship between severity and speeding

We hypothesized that probability of a Severity 2 collision will be higher if the driver was speeding. We tested the same with the data we have. We can observe below, the share of Severity 2 collision is slightly higher when the driver is speeding.

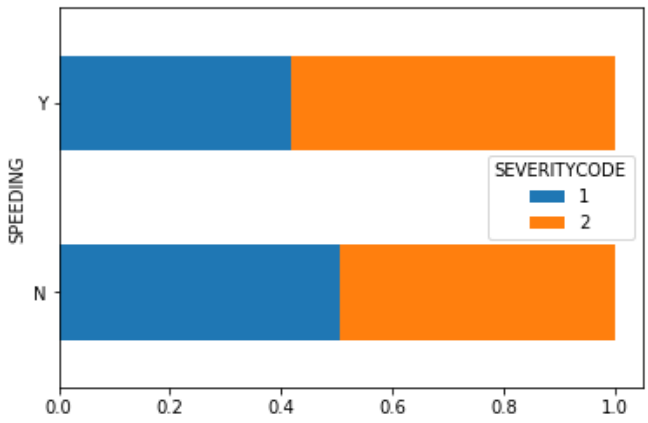

Figure 6. The share of Severity 1 and 2 collisions based on the driver speeding

3.5 Relationship between severity and time of collision

When we look at the histogram plots collision by month of the year and day of the month, it is clear that Severity 1 collisions are concentrated in the winter months i.e. December, January and February and relatively more on Sundays compared to other days.

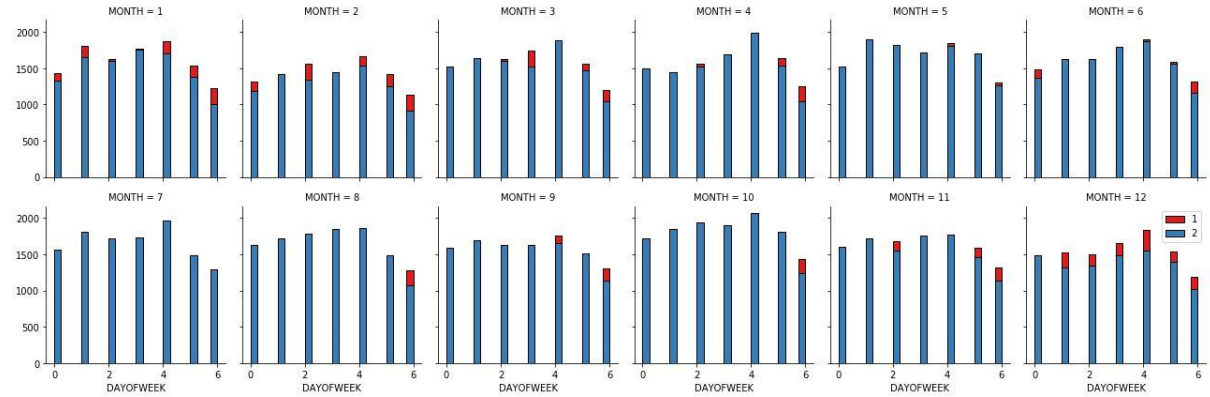

Figure 7. Histogram plot of severity by month of year and day of time.

When we observe the histogram plot of collisions by day of the week and the hour of the day, you see more Severity 1 collisions are concentrated at midnight and hours before and after it i.e. the night time.

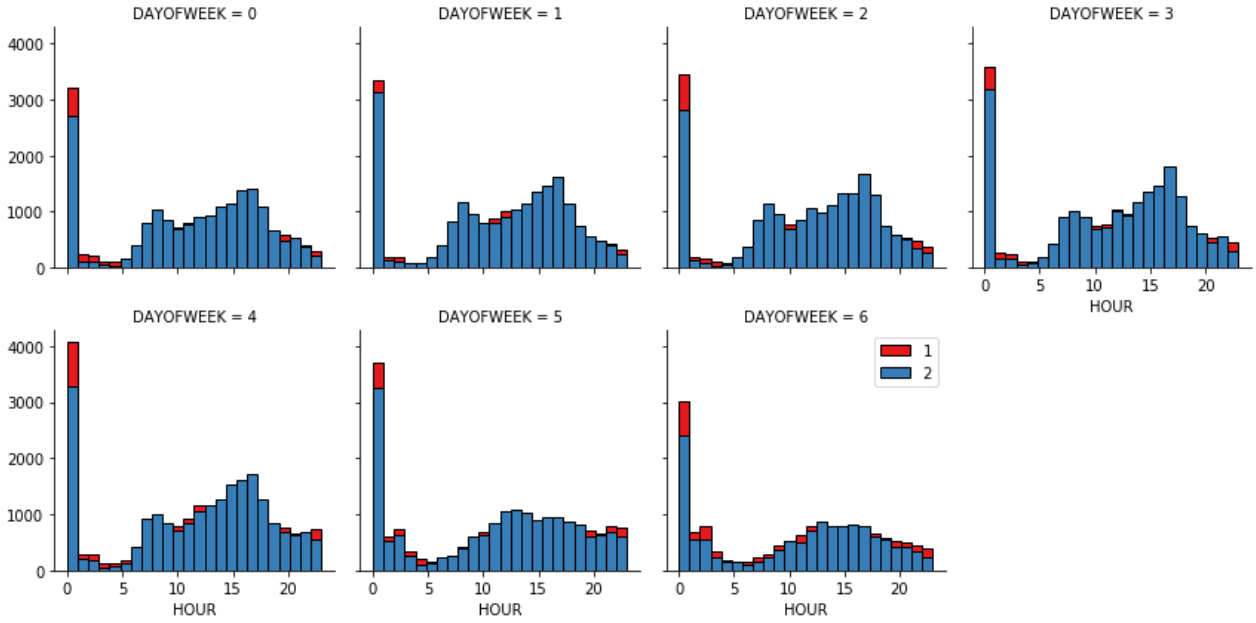

Figure 8. Histogram plot of day of time and hour of day

3.6 Final feature set

Since most of the features are categorical variables, we have created dummy variables to represent each value of the attribute. For month of the year, we have split them into a single dummy i.e. Winter

4. Predictive Modeling

The aim here is to classify the incident as Severity 1 or Severity 2 collision. We can use classification modelling to focus on the prediction of severity of the collision. Had the severity been on a continuous scale, we could have considered regression models or unsupervised models which would rather than concentrating on probability of the severity of collision would have predicted the extent (severity) to which the collision is problematic.

Ideally, we should be also having negative data i.e. no collision occurring but with the limited data we have we will classify between Severity 1 and Severity 2 collisions.

Under the classification models, we can look at K-Nearest Neighbor classification, Decision Tree Classification, Random Forest Classification and Support Vector Machine model. Since we have a large data set and a large feature set, SVM model might be too slow to process the data for modelling.

Had we had imbalanced data, Decision Tree Classification or Random Forrest Classifier would have been best. However, since we have up-sampled the minority class, we can include K-Nearest Neighbor algorithm for analysis.

4.1 Regression models

4.1.1 Performances of different models

We applied above mentioned three classification models on the data. For performance evaluation, we have used Accuracy Score, F1 score and Jaccard Similarity score on all three models. In addition to this, we have also created a confusion matrix to test the precision and recall.

Of all the models, Random Forrest Classifier is the best performing model but only by a marginal difference as compared to the other two models. Another observation on the performance in terms of time taken – KNN model was significantly slower than other models taking more than 15 minutes as compared to the response of Decision Tree and Random Forest.

For the sake of this evaluation, we will assume Severity 1 collision as positives and Severity 2 collisions as negatives.

Of the three models, KNN model's performance was comparatively different from the other two. KNN model was better than other models at predicting true positives while inferior in predicting true negatives.

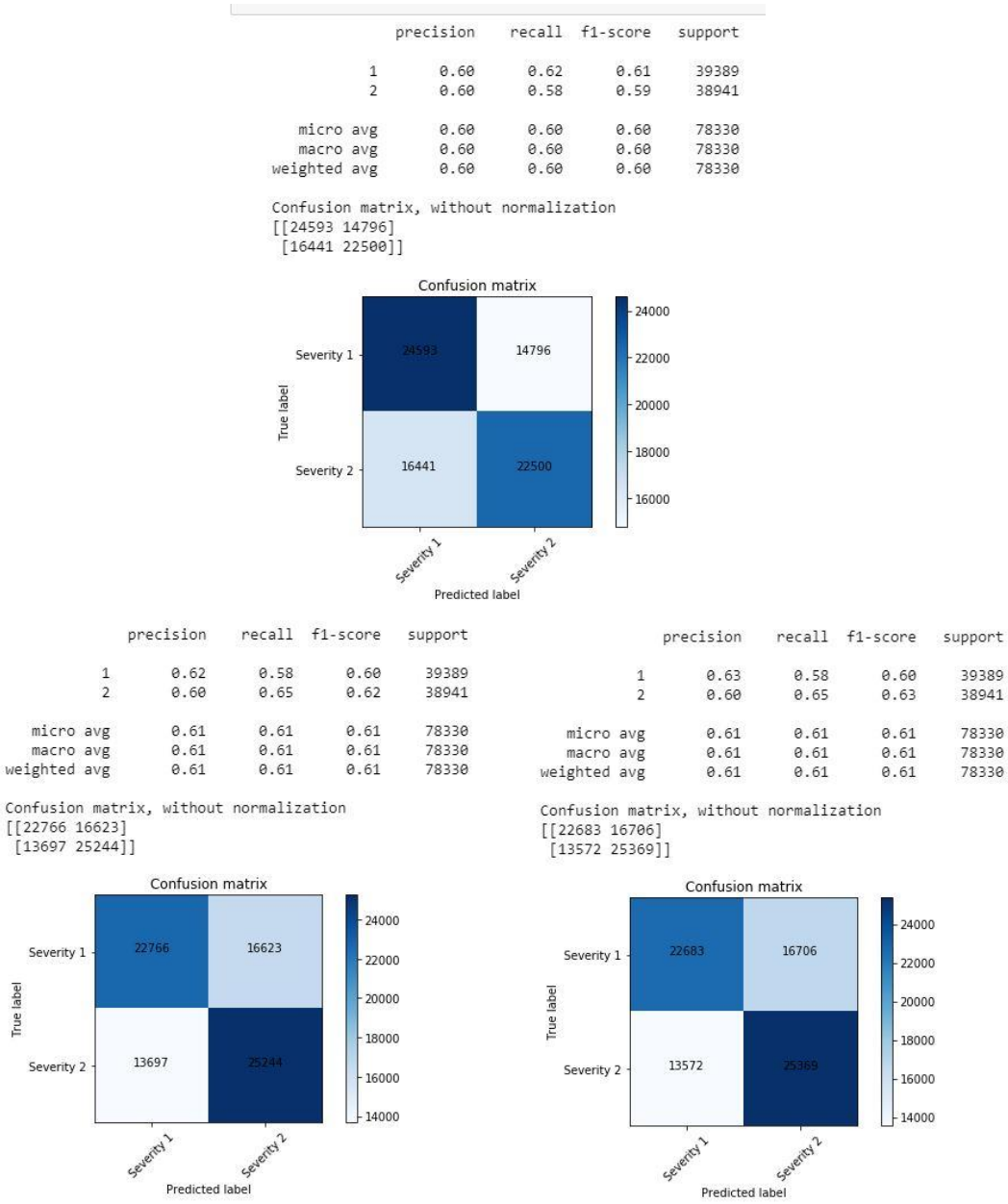

Figure 9. Confusion matrix for KNN model (top), Decision Tree Classification model (bottom left) and Random Forest Classification model (bottom right)

Table 1. Performance of classification models. Best performance labeled in red.

	KNN	Decision Tree	Random Forest
Accuracy Score	0.6012	0.6129	0.6134
F1 score	0.6009	0.6124	0.6129
Jaccard Similarity	0.6012	0.6129	0.6134
No. of True Positives	24593	22766	22683
No. of False Positives	14796	16623	16706
No. of False Negatives	16441	13697	13572
No. of True Negatives	22500	25244	25369

5. Conclusions

In this study, we analyzed different models in predicting severity of a collision. We had data for two severity values to play with i.e. 1 – prop damage and 2 – injuries. We were able to successfully predict over 60% of the classifications using different models. To determine the best performing model, we prioritized looking at the true negatives i.e. severity 2 detections as that would be more costly is predicted falsely leading to injuries. Keeping that in mind, we conclude that Random Forest Classifier is the best model for collision severity prediction.

6. Future directions

As mentioned before, the data is skewed towards only two severity types, having more data on the other values would improve prediction. So as a next step, we should look at if there is any data available that would give us more granular split.

In addition to this, the aim of this project was to predict the probability of a collision in addition to severity of the collision. We have only achieved the latter part assuming a collision is going to

occur. In such cases, having a continuous variable for severity may help us understand lest severe cases which would not lead to a collision or even have more robust data on where we move away from classification and into unsupervised models where we can classify certain situations into no collisions. A solution might be to use this data on future conditions to identify cases where there was no collision when the drive was not warned.

Lastly, we can improve on the accuracy score if we had more features like total traffic on the road, total pedestrians on the road, visibility, and other such information. Having some continuous features will also help make the models better.