Prediction of Car Accident Severity

Ishan Takkar

October 14, 2020

1. Introduction

1.1 Background

Road traffic crashes are one of the world's largest public health and injury prevention problems. The problem is all the more acute because the victims are overwhelmingly healthy before their crashes.

According to the World Health Organization (WHO), more than 1 million people are killed on the world's roads each year. A report published by the WHO in 2004 estimated that some 1.2 million people were killed and 50 million injured in traffic collisions on the roads around the world each year and was the leading cause of death among children 10–19 years of age.

1.2 Problem

The project's objective is to predict the probability and severity of a car accident before it happens in order to intimate the driver to avoid such incidents through the use of historical car accidents data and how aspects like weather, road conditions, traffic and many other factors affect the probability of an accident and how severe it can be.

1.3 Interest

The insights of this project can be helpful to the following parties:

- **Drivers**: They can get to know the probability and severity of the accident that can happen, they might consider driving more carefully or even choose a safer route.
- **Emergency Services/Road Safety/Traffic Department**: They can get to know the areas where the probability of an accident to happen is more. So, they can have more teams near those areas.

2. Data

2.1 Data Sources

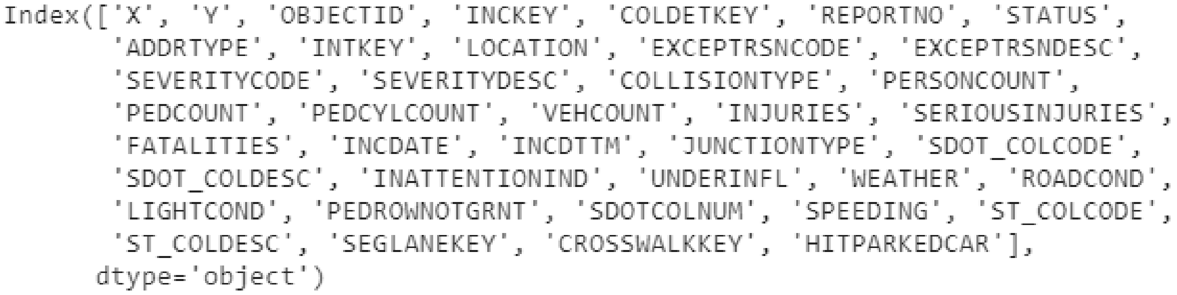
The data is taken from the Seattle Department of Transportation and recorded by Traffic Records Group. It covers the annual collisions data from 2004 to the present.
The dataset can be downloaded from here and the metadata from here.

2.2 Data Description

The dataset is huge as it consists of **221266 rows and 40 columns.**
The scope of the project is to predict the likelihood and severity of an accident. Therefore, we use **SEVERITYCODE** (i.e. the severity of the accident) as the dependent variable.

We can potentially use the other 39 columns to train the algorithm. There are many columns that are not relevant for this project or they contain non-standardized data and missing values hence the data needs to be cleaned up.

Columns:



```
Index(['X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS',
       'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC',
       'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT',
       'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'SERIOUSINJURIES',
       'FATALITIES', 'INCDATE', 'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE',
       'SDOT_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND',
       'LIGHTCOND', 'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE',
       'ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR'],
      dtype='object')
```

Fig. 2.1

2.3 Data Preparation

2.3.1 Unnecessary Columns

The dataset had many columns which don't any purpose for this project and dropping them would help in cleaning the data.

Unique Keys: Since, ID's are not predictors and there's no need for them.
('OBJECTID', 'INTKEY', 'COLDETKEY', 'INCKEY', 'SEGLANEKEY', 'CROSSWALKKEY')

Unknown Columns: There is no description of some columns in the metadata and hence, can't be used for analysis.
('EXCEPTRSNCODE', 'REPORTNO', 'STATUS', 'SDOTCOLNUM', 'ST_COLCODE', 'PEDROWNOTGRNT')

Description Columns: There are some columns in the dataset which are to describe other columns. There are of no use to use and will be dropped.
('EXCEPTRSNDESC', 'SEVERITYDESC', 'SDOT_COLDESC', 'ST_COLDESC')

After dropping the unnecessary columns we are left with **24 Columns**.

2.3.1 Dealing with Missing Data

Some columns in the dataset were left empty on purpose. These are binary variable in which the Y or 1 value is noted and rest is left empty because it's obvious that they either are N or 0.
('SPEEDING', 'INATTENTIONIND', 'UNDERINFL', 'HITPARKEDCAR')

These columns had such issues and the missing values were evaluated with 0 and the value Y and N were replaced with 0 and 1 respectively to easier analysis.

Once I chose the attributes and the target variable, I dropped the unnecessary columns and analysed more deeply the necessary ones.

2.3.2 Unknown and Other values

Many columns have Unknown and Other values which either needs to be replaced to removed. On further evaluation, these values were replaced with null values.

2.3.3 Date-Time Features

Extracted Months, Days and Hours from Date-Time variables for better analysis. ('INCDTTM', 'INCDATE').

Converted these to ('Hour,'Month','Weekday') columns and hence, dropped the ('INCDTTM', 'INCDATE') columns.

2.3.4 Null Values

The dataset had a lot of null values and cleaning them necessary before performing any data analysis. Since, all the values which were deemed to be not useful in this project were converted to null and now all these null values were dropped.

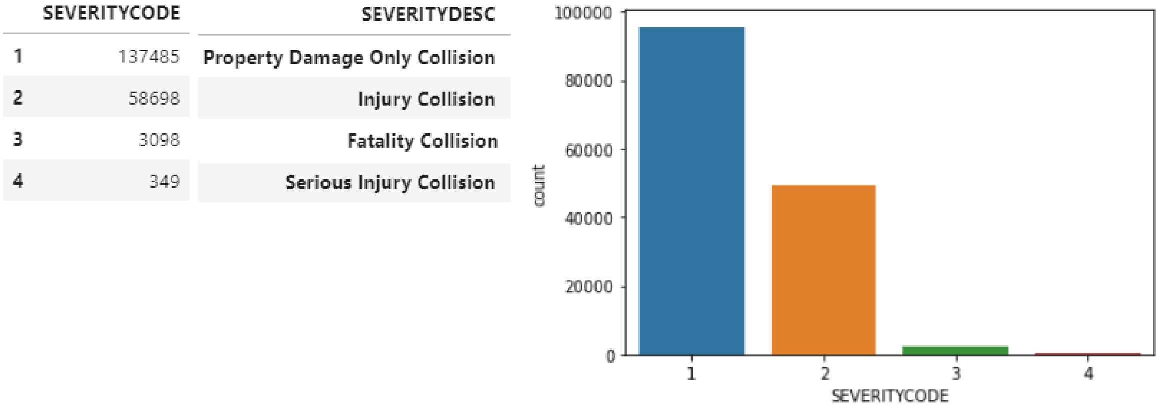
After, this step the dataset had **147743 rows and 25 columns.**

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Calculation of target variable

The target column of this project is ‘SEVERITYCODE’. Figure 3.1 shows the columns values mapped with its description and the count of each type of collision.



	SEVERITYCODE	SEVERITYDESC
1	137485	Property Damage Only Collision
2	58698	Injury Collision
3	3098	Fatality Collision
4	349	Serious Injury Collision

Fig 3.1 Severity Codes and Description with counts.

3.1.2 Relation between Environmental and Physical Conditions with Accidents.

1. **Relationship with Weather:**
   Weather can be a key factor for a collision to happen and how severe it could be. In Figure 3.2 it can be seen that the majority of collisions took place in **Clear** Weather.

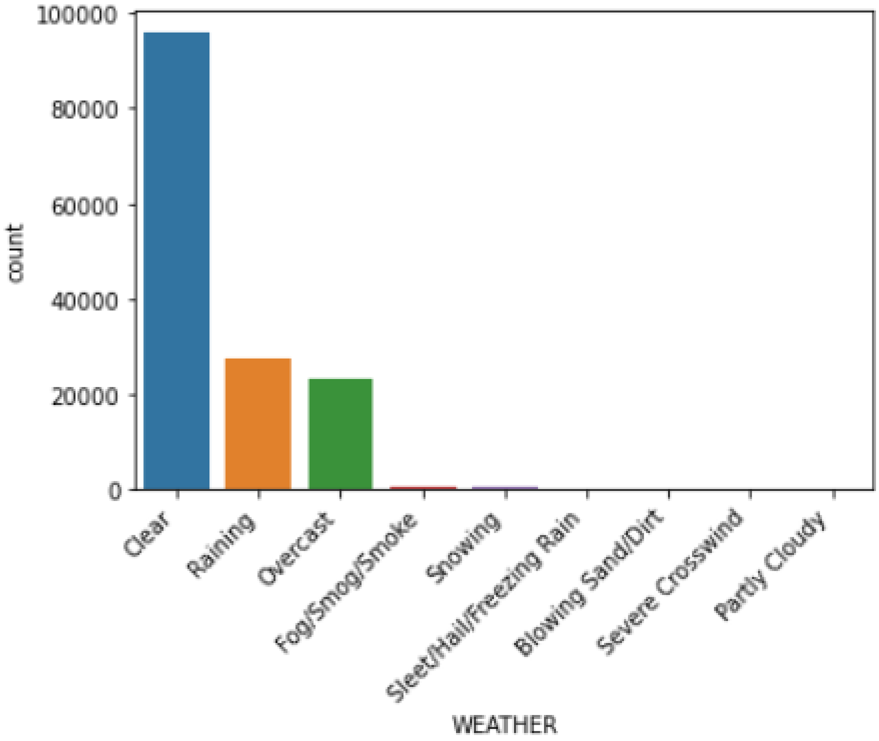


Fig 3.2 Relationship between Weather and Collisions.

2. **Relationship with Road Conditions:**
   Road Conditions can be a key factor for a collision to happen and how severe it could be. In Figure 3.3 it can be seen that the majority of collisions took place on **Dry** Roads.

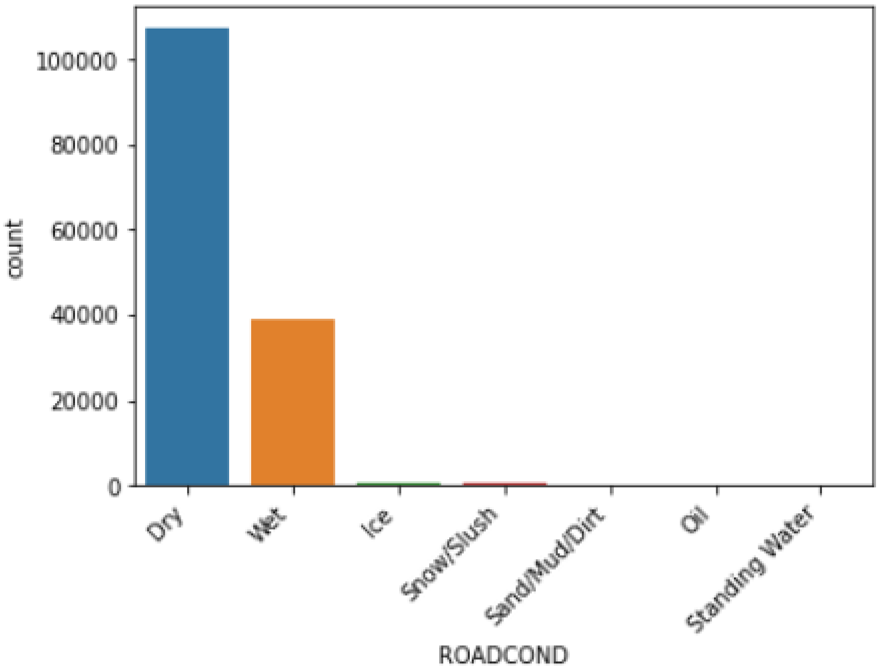


Fig 3.3 Relationship between Road Conditions and Collisions.

3. **Relationship with Light Conditions:**
   Light Conditions can be a key factor for a collision to happen and how severe it could be. In Figure 3.4 it can be seen that the majority of collisions took place in **Daylight.**

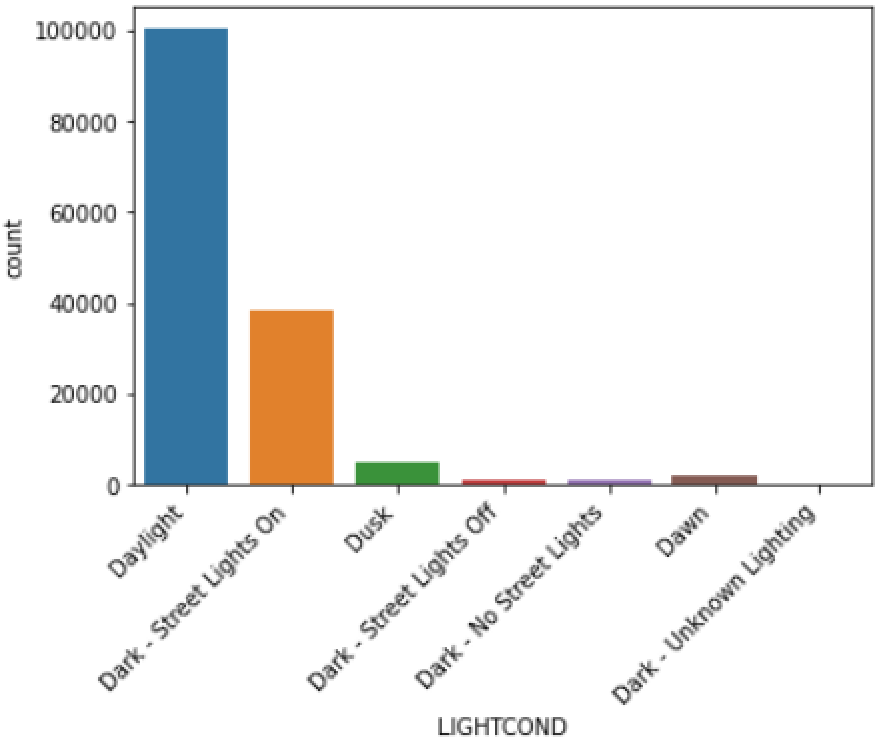


Fig 3.4 Relationship between Light Conditions and Collision.

3.1.3 Relation between Common Mistakes and Accidents.

Driving the car beyond the speed limits or not being attentive while driving or being under the influence of alcohol could be huge factor in causing a Collision. In this we analyze the factors such as ‘SPEEDING’, ‘UNDERINFL’, ‘INATTENTIONIND’ and their relationship between the collisions.

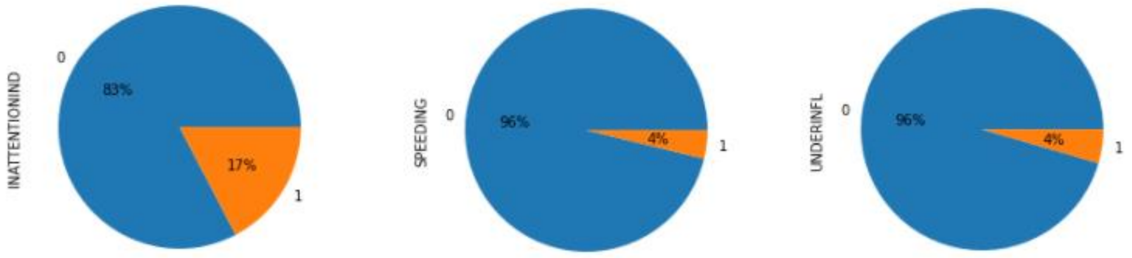


Fig 3.5 Relation between Driver Mistakes and Accidents.

In the Fig 3.5 it is seen that on average only a small percentage of accidents were caused because of these factors.

3.1.4 Relation between Hour, Month and Day with Accidents.

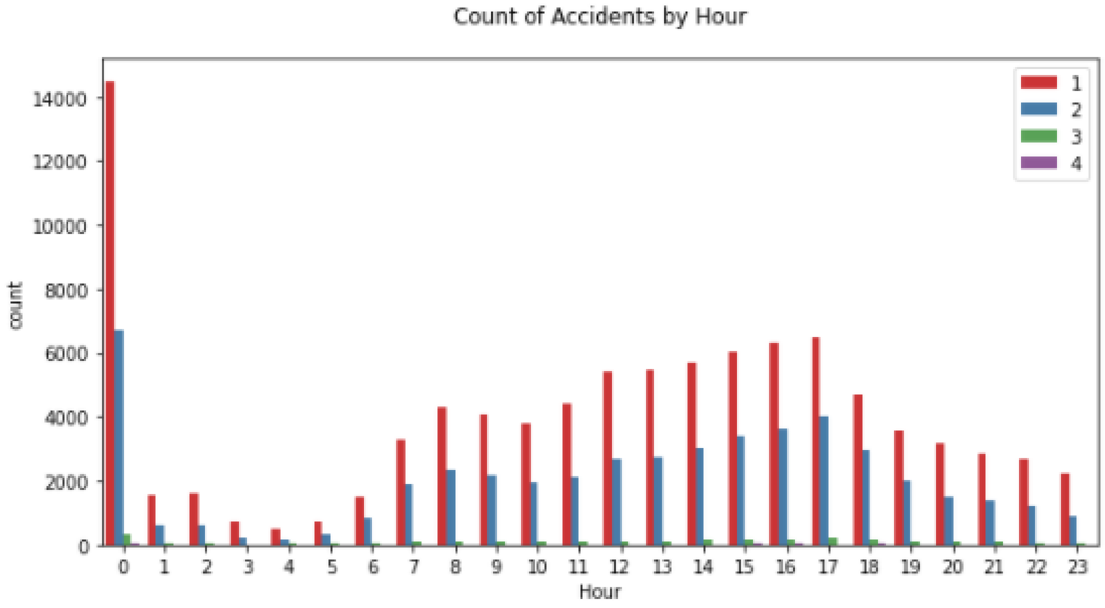


Fig 3.6 Hour vs. Accidents

In Fig 3.6 it can be noted that the majority of accidents took place at mid night.

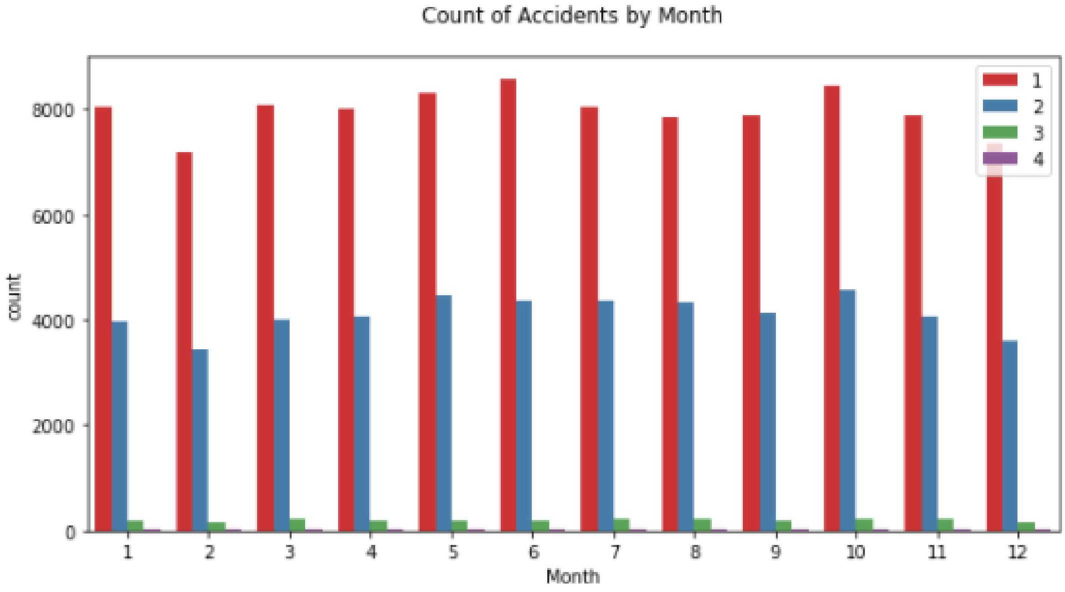


Fig 3.7 Month vs. Accidents

No trend can be noticed in Fig 3.7 as almost all month recorded same number of Accidents.

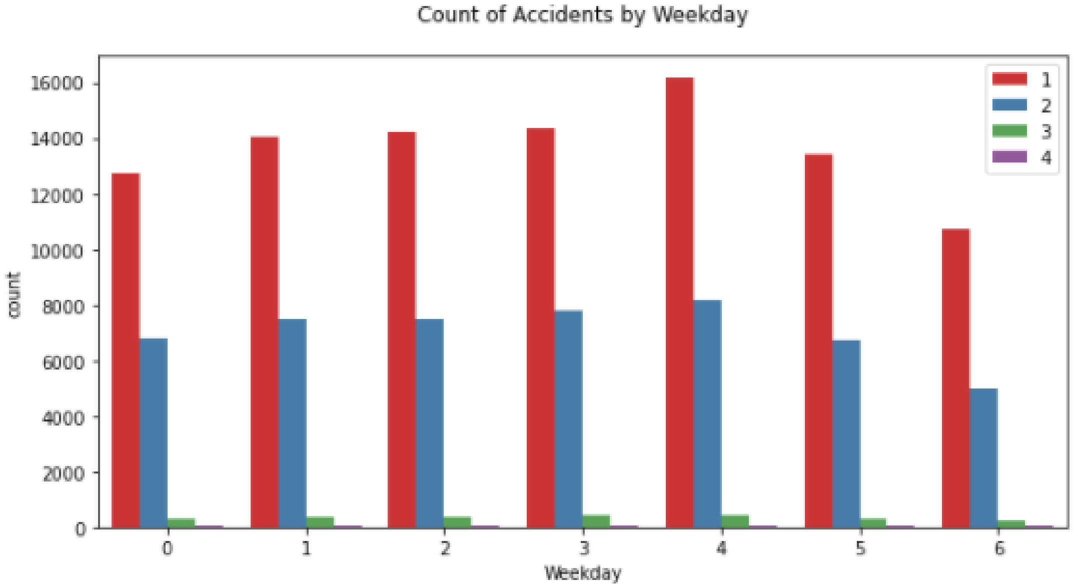


Fig 3.8 Weekday vs. Accidents

In Fig 3.8 it can be noted that slightly higher number of accidents took place on the 5th day of week i.e. Friday.

3.1.5 Relation between Junction Type and accidents

Junction Type can be a good indicator to identify what type of junction can cause a collision. In the fig 3.8 we try to identify the same.

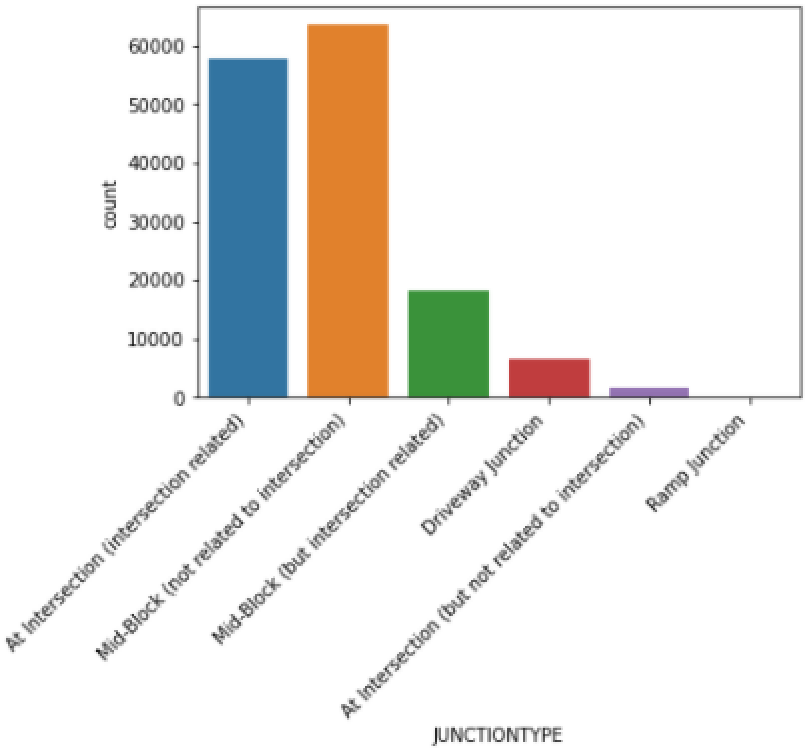
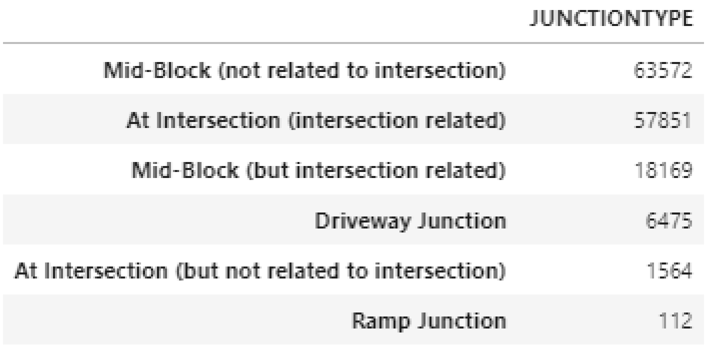


Fig 3.8 Junction Type vs. Accidents

It can be discovered that majority of accidents took place at **Mid-Block (not related to intersection)** and **At Intersection (intersection related).** Exact numbers are mentioned below.

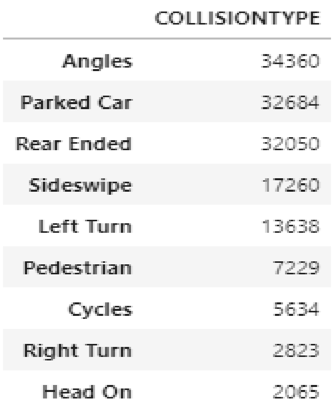


	JUNCTIONTYPE
Mid-Block (not related to intersection)	63572
At Intersection (intersection related)	57851
Mid-Block (but intersection related)	18169
Driveway Junction	6475
At Intersection (but not related to intersection)	1564
Ramp Junction	112

Fig 3.9 Number of Accidents at Junction Type

3.1.6 Relation between Collision Type and accidents

The type of Collision can help predict the severity of the accident that might have happened.



	COLLISIONTYPE
Angles	34360
Parked Car	32684
Rear Ended	32050
Sideswipe	17260
Left Turn	13638
Pedestrian	7229
Cycles	5634
Right Turn	2823
Head On	2065

Fig 3.10 Collision Type Count

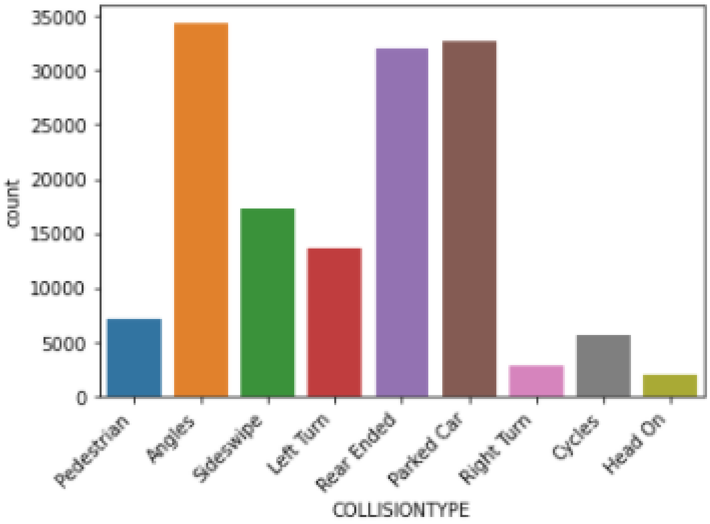


Fig 3.11 Collision Type vs. Accidents

In Fig 3.11 it is noted that Angles, Rear Ended and Parked Car contribute to the majority of the collisions that took place.

3.2 Feature Selection

After performing analysis on various features, we need to select some features that will help predict the severity of the collisions. A heat map can help us to find the correlation between the features.

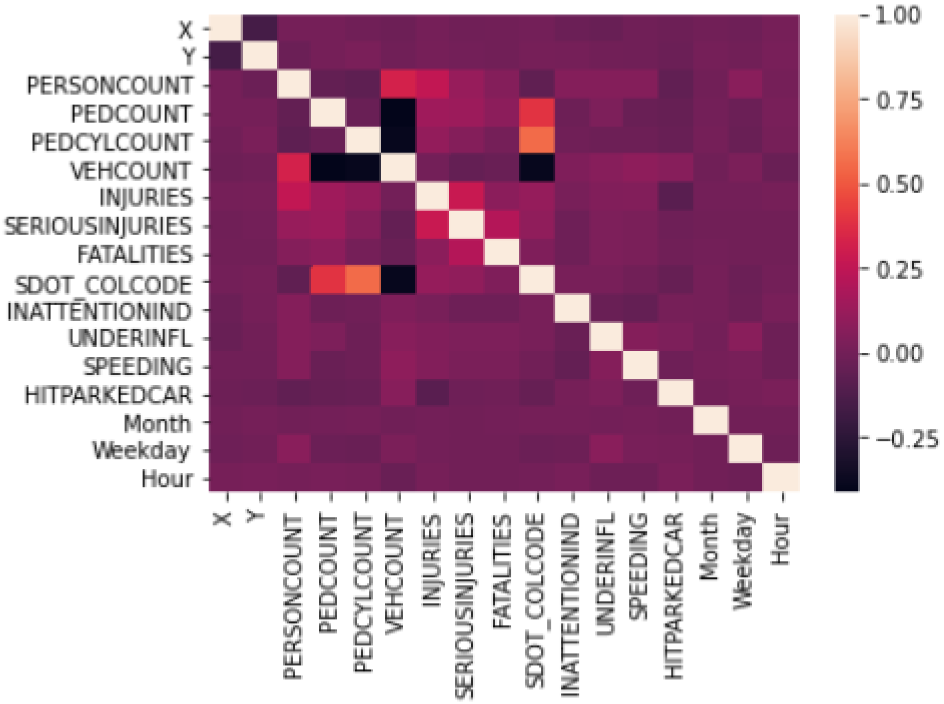


Fig 3.12 Heat Map

As it is very evident from the heat map that there is not any strong correlation between the features. However, we can try a few which have more than average correlation.

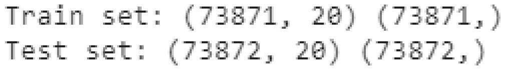
The features thus selected for modeling are as follows:

1. 'JUNCTIONTYPE'
   As in our analysis we found how type of junctions can have an impact on the severity of the collision.
2. 'PERSONCOUNT'
   This has a direct relation with the severity as if the number of person involved is high then there are good chances that the collision was very severe.
3. 'VEHCOUNT'
   As more vehicles are involved more are the chances of a big collision.
4. 'ADDRTYPE'
   Knowing where the collision took place can help identify more accurately.
5. 'PEDCOUNT'
   A direct relation is there that if more pedestrians are involved in the collision then it implies that the collision is severe.
6. 'COLLISIONTYPE'
   The type of collision can really help identify the severity especially if it involves a pedestrian.

After the selection of features, the categorical values were converted to Binary values for the modeling process using One Hot Encoding.

3.3. Predictive Modeling

Since, our target column is categorical we will be using a classification model to predict the severity of a collision.
The dataset is split into training and testing dataset. Following is the shape of the dataset after the split.



```
Train set: (73871, 20) (73871,)
Test set: (73872, 20) (73872,)
```

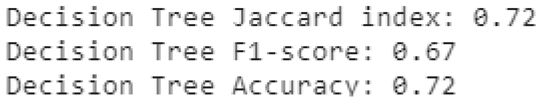
I implemented 4 models to identify which is the best suited for this project.

1. Decision Tree
2. K- Nearest Neighbor
3. Logistic Regression.
4. Support Vector Machine (SVM)

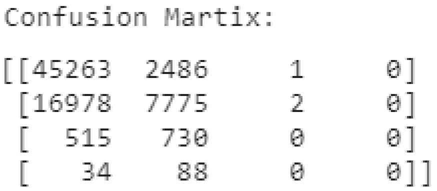
3.4 Model Evaluation

The models were evaluated using Jaccard Index, F1- Score and Accuracy Score and Log loss for Logistic Regression. Confusion matrix was created to look at the True and False, Positive and Negatives. A Classification Report was created to identify Precision and Recall. Below are the metrics which will help in identifying the best model.

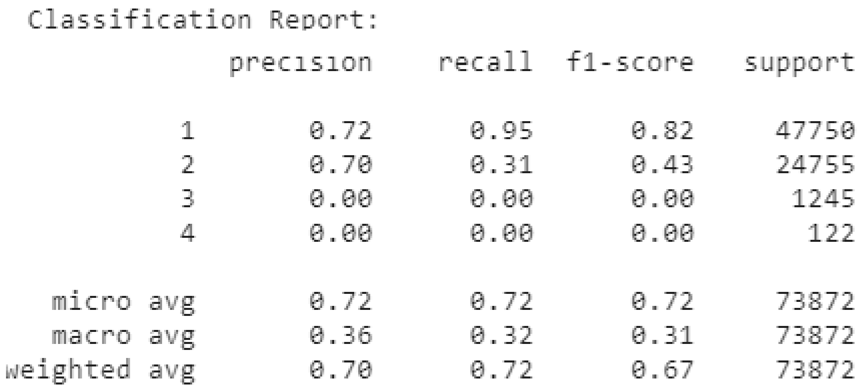
1. Decision Tress:



```
Decision Tree Jaccard index: 0.72
Decision Tree F1-score: 0.67
Decision Tree Accuracy: 0.72
```

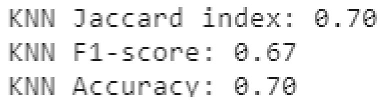


```
Confusion Martix:

[[45263  2486     1     0]
 [16978  7775     2     0]
 [  515   730     0     0]
 [   34    88     0     0]]
```

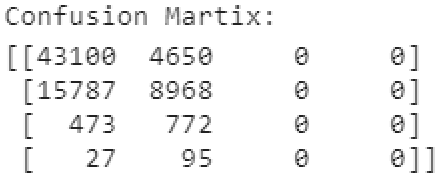


```
Classification Report:
              precision    recall  f1-score   support

           1       0.72      0.95      0.82     47750
           2       0.70      0.31      0.43     24755
           3       0.00      0.00      0.00      1245
           4       0.00      0.00      0.00       122

   micro avg       0.72      0.72      0.72     73872
   macro avg       0.36      0.32      0.31     73872
weighted avg       0.70      0.72      0.67     73872
```

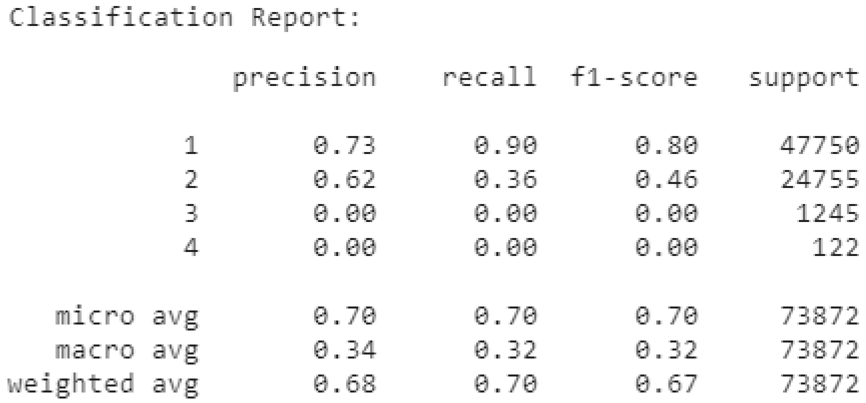
2. K- Nearest Neighbor



```
KNN Jaccard index: 0.70
KNN F1-score: 0.67
KNN Accuracy: 0.70
```

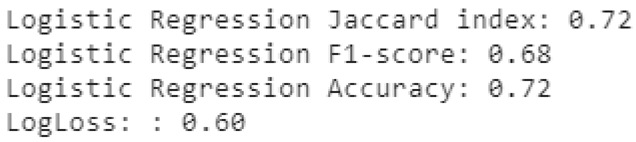


```
Confusion Martix:
[[43100  4650     0     0]
 [15787  8968     0     0]
 [  473   772     0     0]
 [   27    95     0     0]]
```

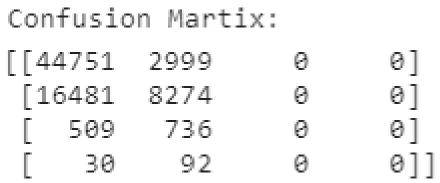


```
Classification Report:
              precision    recall  f1-score   support

           1       0.73      0.90      0.80     47750
           2       0.62      0.36      0.46     24755
           3       0.00      0.00      0.00      1245
           4       0.00      0.00      0.00       122

   micro avg       0.70      0.70      0.70     73872
   macro avg       0.34      0.32      0.32     73872
weighted avg       0.68      0.70      0.67     73872
```

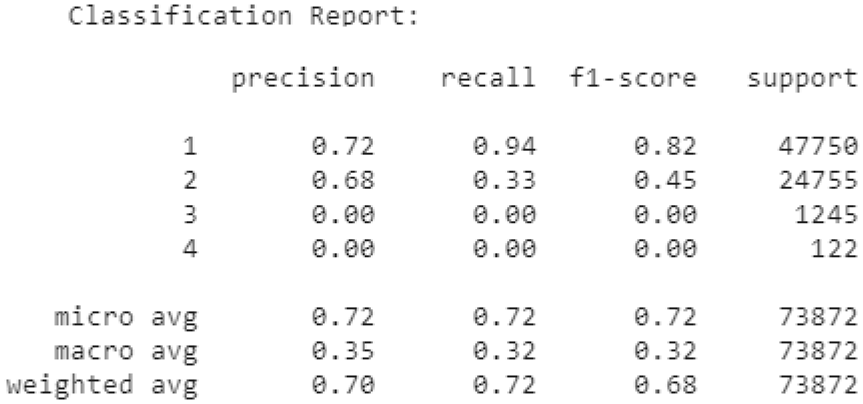
3. Logistic Regression



```
Logistic Regression Jaccard index: 0.72
Logistic Regression F1-score: 0.68
Logistic Regression Accuracy: 0.72
LogLoss: : 0.60
```

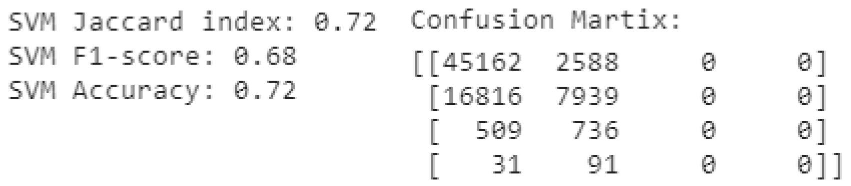


```
Confusion Martix:
[[44751  2999     0     0]
 [16481  8274     0     0]
 [  509   736     0     0]
 [   30    92     0     0]]
```

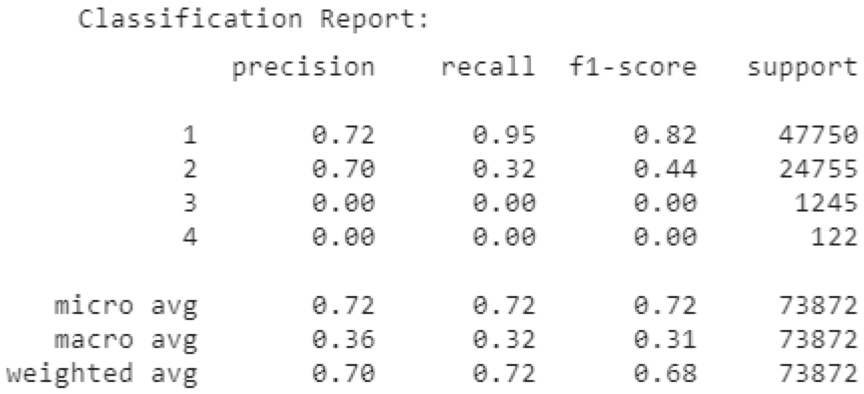


```
Classification Report:
              precision    recall  f1-score   support

           1       0.72      0.94      0.82     47750
           2       0.68      0.33      0.45     24755
           3       0.00      0.00      0.00      1245
           4       0.00      0.00      0.00       122

   micro avg       0.72      0.72      0.72     73872
   macro avg       0.35      0.32      0.32     73872
weighted avg       0.70      0.72      0.68     73872
```

4. Support Vector Machine (SVM)



```
SVM Jaccard index: 0.72  Confusion Martix:
SVM F1-score: 0.68       [[45162  2588     0     0]
SVM Accuracy: 0.72        [16816  7939     0     0]
                          [  509   736     0     0]
                          [   31    91     0     0]]
```



```
     Classification Report:
               precision    recall  f1-score   support

           1       0.72      0.95      0.82     47750
           2       0.70      0.32      0.44     24755
           3       0.00      0.00      0.00      1245
           4       0.00      0.00      0.00       122

   micro avg       0.72      0.72      0.72     73872
   macro avg       0.36      0.32      0.31     73872
weighted avg       0.70      0.72      0.68     73872
```

4. Results

On examining the scores, the confusion matrix and the classification report. The results of the models are somewhat close but the **SVM Model** is the best among all with an overall Accuracy of **72%.**

The model is a really good predictor of **Severity 1** collisions but the accuracy drastically drops for **Severity 2** and none of the models is able to predict for **Severity 3 and 4**.

Since, Severity 2 should be given more preference because it has a bigger impact in that case the **K-Nearest Neighbor Model** does a great job. Although, the overall accuracy of KNN Model is slightly lower that other models but shows the max number of True Positives i.e. 8968 for severity 2 collisions.

5. Discussion

In my analysis I found that features such as INJURIES are really overpowering and on including them the accuracy of models goes to about 96%. That’s why none of these features were included in the final modeling and evaluation.

Correlation does not imply causation:

It was found that majority of collision happened in **Clear weather, Dry Road, Daylight Light Condition** that does not mean that these conditions caused more accidents. In a year majority of days have

1. Clear skies
2. Dry Roads
3. More traffic in day time.

That's the reason why they result to majority of collisions, they do share a correlation but are not the cause of collisions and on including these features the accuracy of the model drops.

Recommendation:

The data is really skewed and majority of features have 1 or 2 values with really high proportion of the whole feature, down sampling of these features and inclusion of more features which could help in the prediction should be considered.

6. Conclusion

To conclude, Road crashes are a huge problem. Hopefully, this project will assist people to be more careful. Drive safe for yourself and for the safety of others.