Mentor- Dr. Ranju Mohan

Car Accident Severity Prediction Report

SIP-2022,IIT Jodhpur

Submitted by:- Toshini Agrawal

Table of contents

Introduction

Some cities around the world share one common thing: they are dense. They have a high amount of people in a lesser space compared to others. This set includes Tokyo, Mumbai, Barcelona and Seattle. Here, they have high issues of traffic-related problems. One such issue is "Traffic Collisions" and it has many negative outcomes such as:

- It can result in loss of human life or a life-changing injury.
- It can result in financial loss/property damage to the people and the city.
- It can cause a loss in productivity due to being stuck in traffic jams that would last for hours.
- It develops unsafe road scenarios for other drivers.

What could cause an accident?

There are several types of attributes that can be a factor in causing an accident such as Temperature. It can be a range where high temperatures can decrease a driver's capability to perform physical and intellectual tasks. Similarly, time of the day is an important factor such as before and after office hours when the traffic is more hence increasing the probability of accidents. Also, weather such as snow or rainfall can decrease the visibility of the driver, contributing to the collision.

Furthermore, the quality of the road such as road width, road alignment and the nature of the road can affect the rate of collision as drivers on smooth roads tend to have a higher speed limit than on bumpy roads increasing the chances of collision.

Data Description

The data sheet named "Data collision" consists of 195k rows and 38 columns and although the data is pre-refined ,still it has some values where cleaning might me record for analysis.Data is of a mixed nature consisting of int,float,date categorical variable so it will require normalisation.

Data Cleaning and Feature selection

There were a lot of missing values in the dataset ,so they had been replaced by 'Unknown' or 'N' depending on the features.

After examining the data, the following features have been taken into account 'LIGHTCOND', 'SPEEDING', 'INATTENTIONIND', 'JUNCTIONTYPE','WEATHER','COLLISIONTYPE', 'ROADCOND', and 'HITPARKEDCAR'.

The type of collision can be figured out from the the probability of an injury whether its property damage (94.5%) or cyclist,pedestrian injury (over 85%) injury.

The two factors weather and road conditions tend to show similar effects as high probability of property damage when there is worse weather and worse road conditions but also sometimes more cautious driving also can be seen .

More injury collision can be seen at intersection and speeding,junction types and inattention tend to have higher likelihood involving injury than baseline -37.5%.

Methodology

The main dependent attribute is ‘SEVERITY CODE’ . Here number 1 is assigned to property damage and ‘2’ is assigned to injury collision. Here after normalising the features through the class preprocessing.StandardScaler by removing the mean and scaling the unit variance.

After that the dataset is split into training and testing data with a test data comprising of 20 % of the dataset. So, now after balancing severity code and input feature , the data has been analysed over 4 machine learning models which are Linear Regression, Decision tree, K-Nearest neighbours and Random Forest.

Why these 4 models ?

Linear Regression- The regression helps makes the estimation faster and more easily over other datasets.

Decision Tree-As the dataset is quite large , it is difficult to look out for each path and their respective conclusions so ,decision tree makes the process easier as it helps in tracing each path and creates a comprehensive analysis and identifies nodes which requires further processing.

K-NN - It helps in giving the most accurate predictions without much hassle as it asks just two things - 1) No of neighbours and 2) the distance metrics for the distance between each points.

Random Forest - It is one of the best methods to work with high dimensional data and their subsets and as this data has many feature attributes it really helps identify their path and reach to a conclusion.

Analysis

Decision Tree

Here the max_depth is taken as 5 so as there is no underfitting or overfitting of the data ,less max_depth will under fit the data and although as the increasing the data would make it more expressive but increasing too much will allow outlines hence creating a noisy sample. The class DecisionTreeClassifier with the criterion as entropy and method metrics.accuracy_score is used to measure the accuracy of the model which came out to be 0.738 and f1 -score as 0.83 . Now as f1-score is the harmonic mean of precision and recall it is perfectably able to tell how the model performs on the unbalanced data.Hence , so far decision tree performs well on unbalanced data. The decision tree made for the given data is in the Figure 1

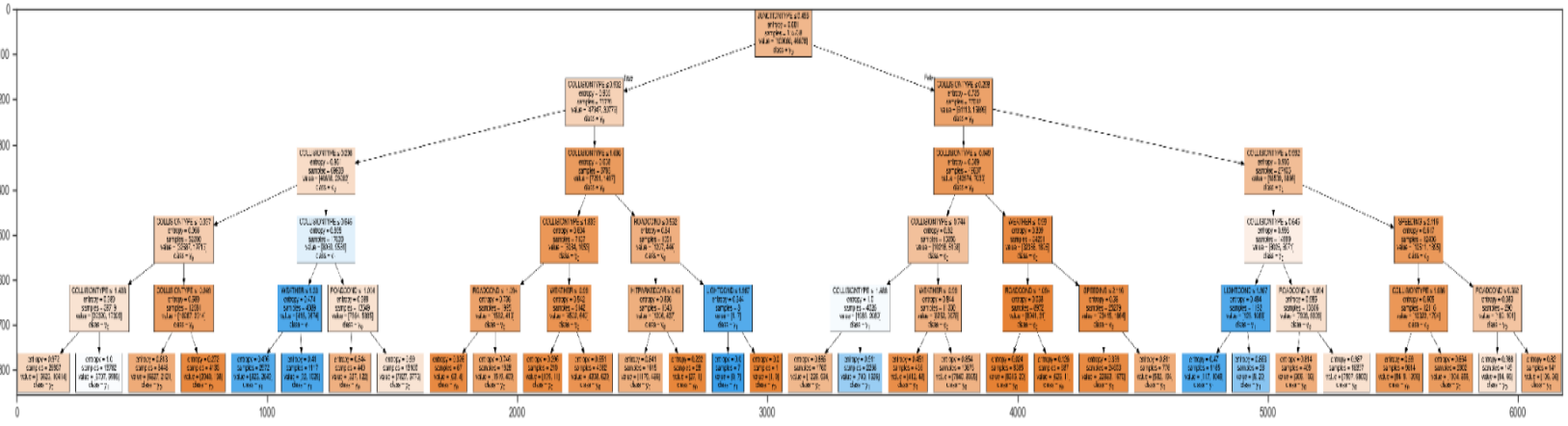

Figure 1.Decision Tree

K-Nearest Neighbours

The class KNeighborsClassifier is used and the value of k is chosen to be 18.

Why K=18?

As, smaller values of k(<10) on large datasets tend to create unstable decision boundaries. Also, when the k value is initalised it is hard to just assign a value specially on a huge dataset. So, in the prediction and training section ,there is no way to know beforehand which value will result in the best output.So, the best way to predict a value is to plot a graph between the error rate on the range of K-values in order to narrow down the choices.

By plotting the graph from the range 1-30 for K-values and determining the k values which has low error rate.

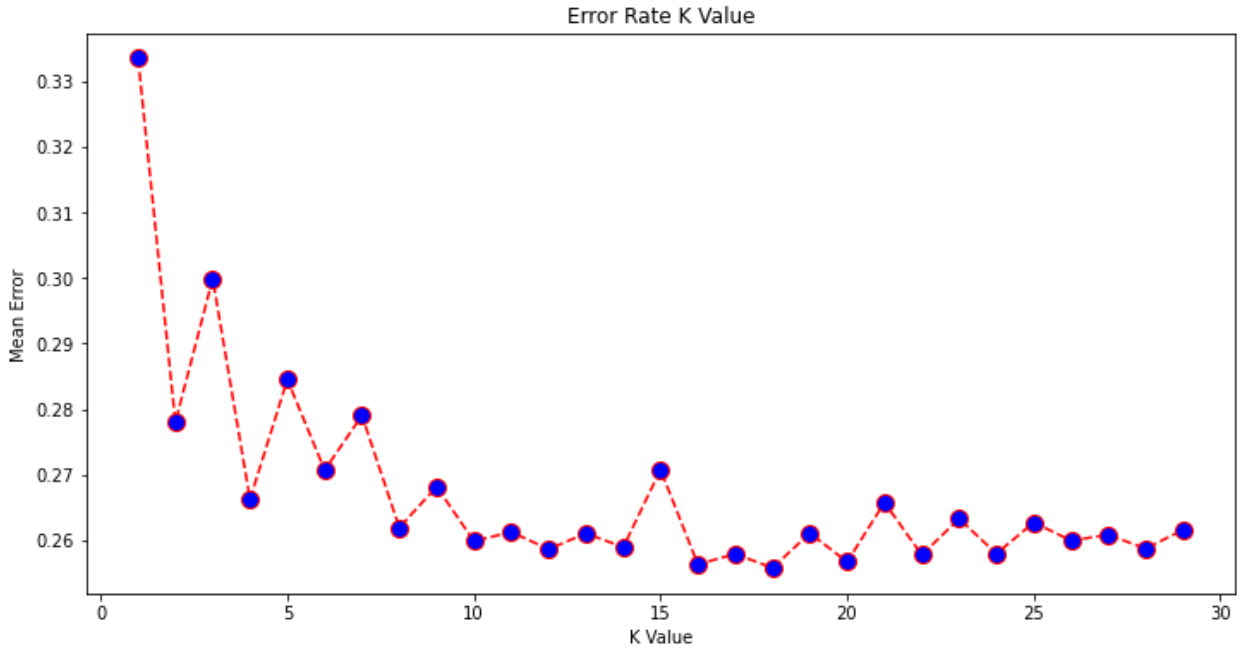

Figure 2 Error Rate VS K-value Graph

Now, here the lowest error rate is to be seen at K=18 .Now , here the accuracy came out the highest at K=18 which was 0.744 which is higher than Decision tree and the f1-score came out be to be 0.84 hence ,both KNN is better at dealing with the unbalanced data and provides a higher accuracy ,hence the best model till now in analyzing the severity data.

Linear Regression

The class LogisticRegression is imported for Linear Regression because to use the ‘liblinear’ function which uses coordinate descent algorithm which is highly useful for multivariate functions.It is recommended for high dimensional data so that is why the class LogisticRegression is used ,which is not be confused with LinearRegression class. Here, the accuracy came out to be 0.69 which is the lowest of all models,due to the fact that it is a regression model which is most simplistic yet a little slow compared to other models .Here, the f1-score is lowest which is 0.82 hence proving that it is least effective in handling the unbalanced data .

Random Forest

Here, the class RandomForestClassifier is used and stored in variable ‘rfc’ . Now the random forest is a collection of decision tree where it prevents in overfitting the data by generating multiple trees.As it searches for the best feature from the random subset of features rather than data as a whole ,it performs more accurately and is more robust . Here ,through the classification report , the f1-score came out to be 0.84 which is similar to KNN but it takes over in the accuracy which is 0.75 .Hence generating the best accuracy in all the given models.

Result and Discussion

First of all taking into account the 8 features on which the severity code is predicted ,it can be seen that

Feature 1- JUNCTIONTYPE

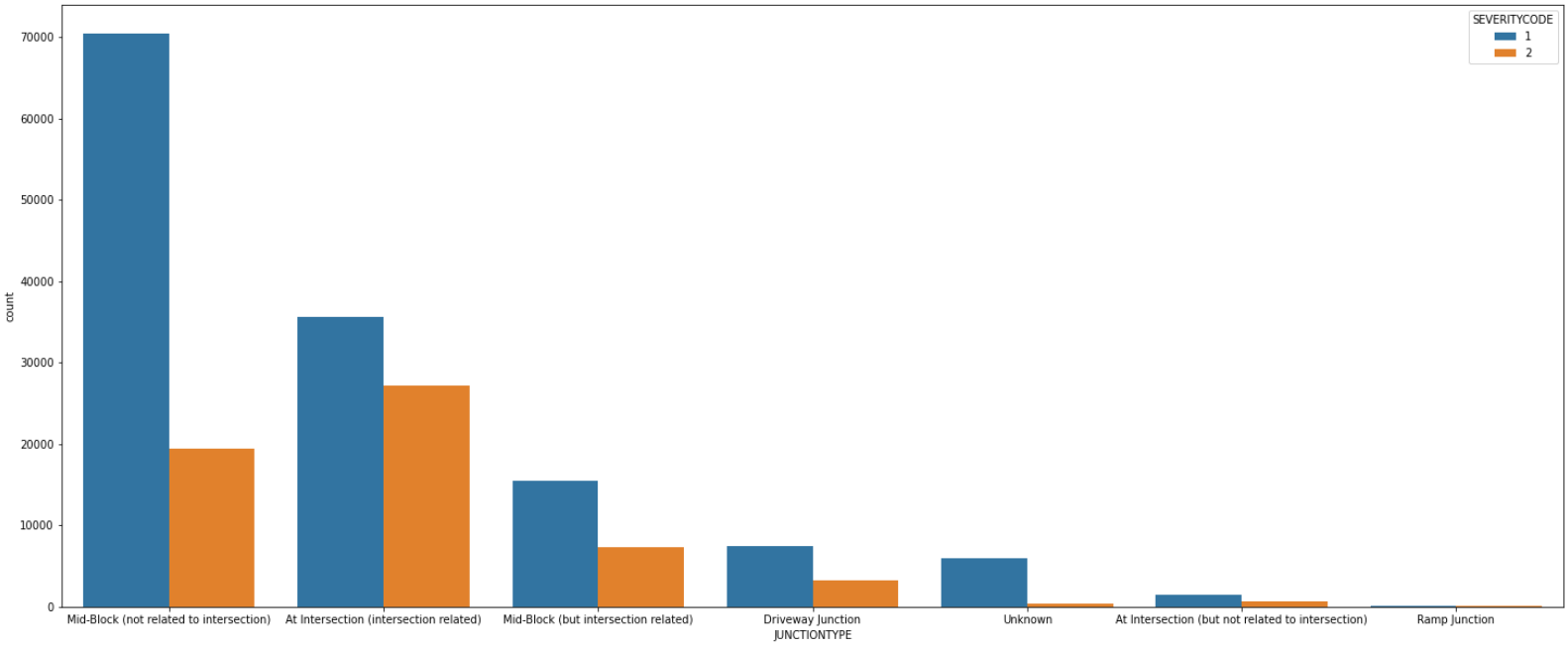

Figure 3-JunctionType Graph

The most number of property damages takes place at mid-block and most injury happens at intersection as seen in the figure 3.

Feature 2 -LIGHTCOND

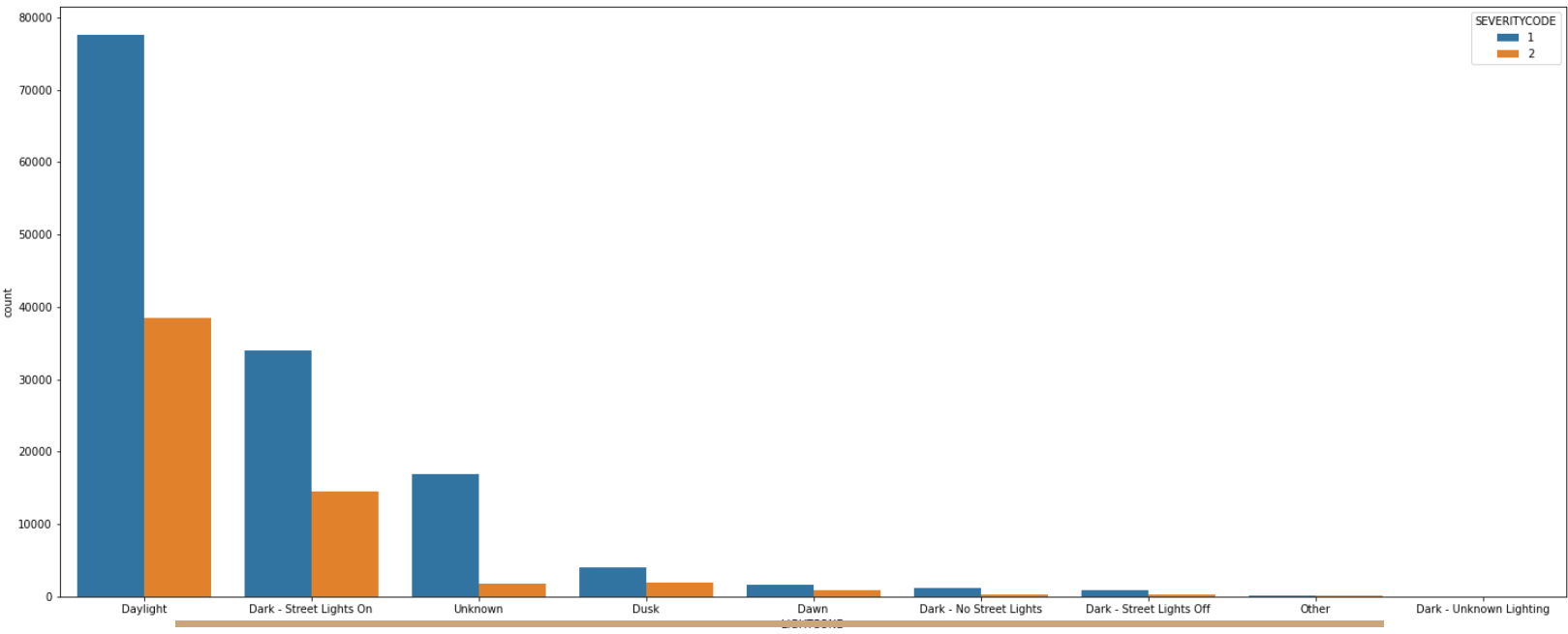

Figure 4-LightCondition Graph

The majority of accidents occur at DayLight which results in highest amount of property damage and injuries.

Feature 3-COLLISIONTYPE

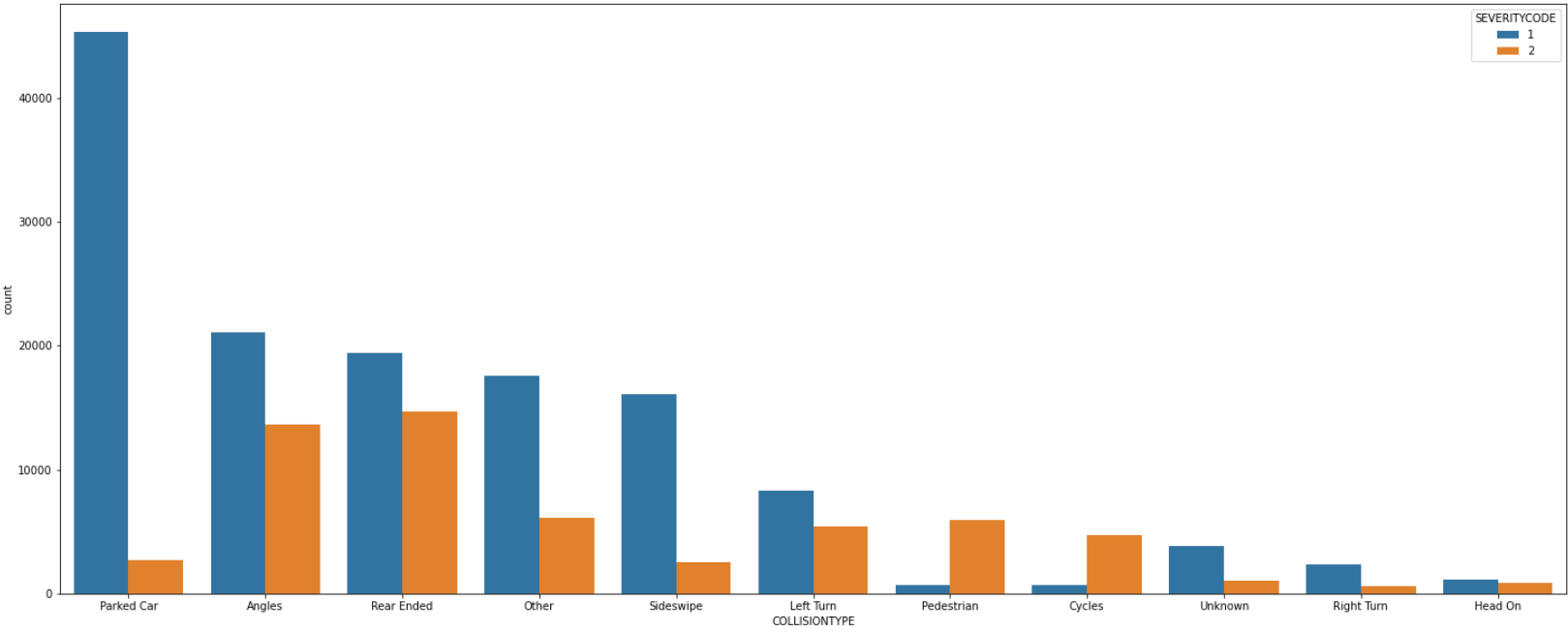

Figure 5- Collision Type Graph

The majority of property damage happened at parked car and the most injuries happended at rear-end.

Feature 4- WEATHER

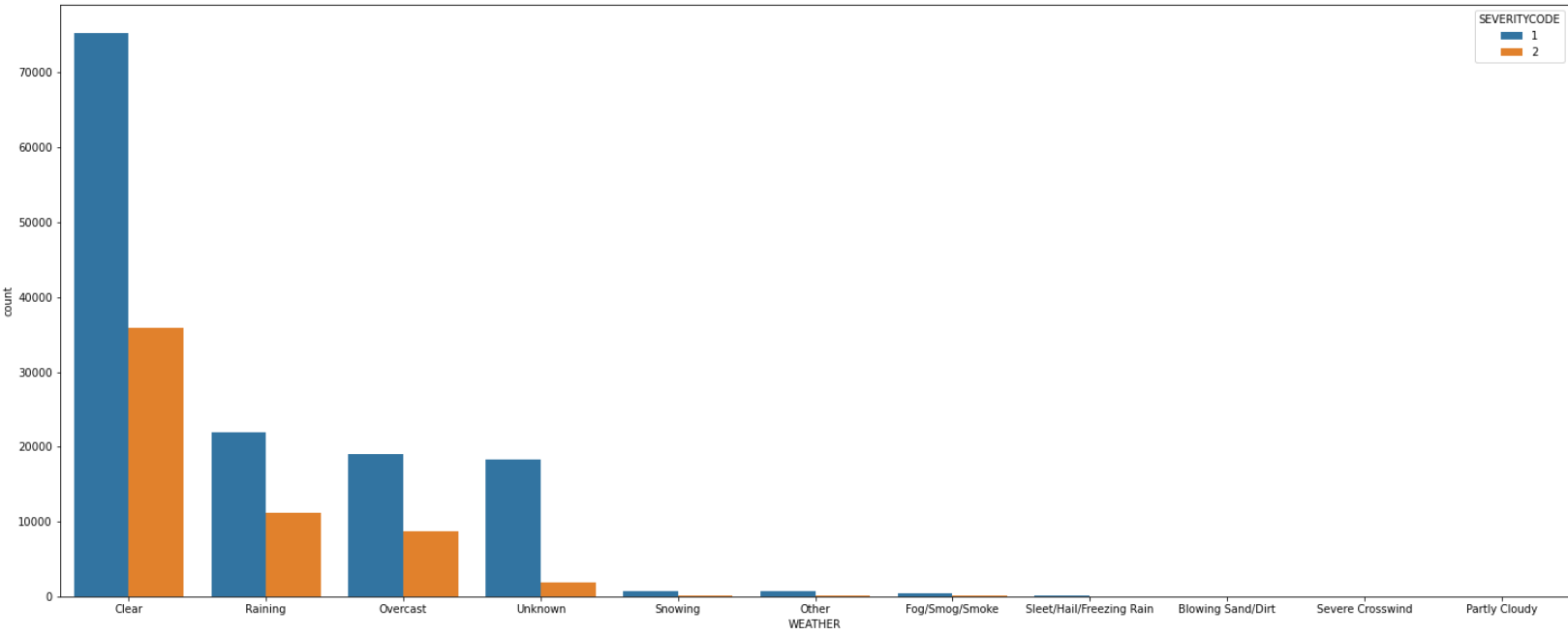

Figure 6 Weather Graph

Surprisingly , the most amount of property damage and injuries happened when the weather was clear , maybe because more alertness and more attentive mode od drivers during unclear conditions such as raining,overcast and other. Although, raining still has high amount of collisions property damage more than injuries.

Feature 5- SPEEDING

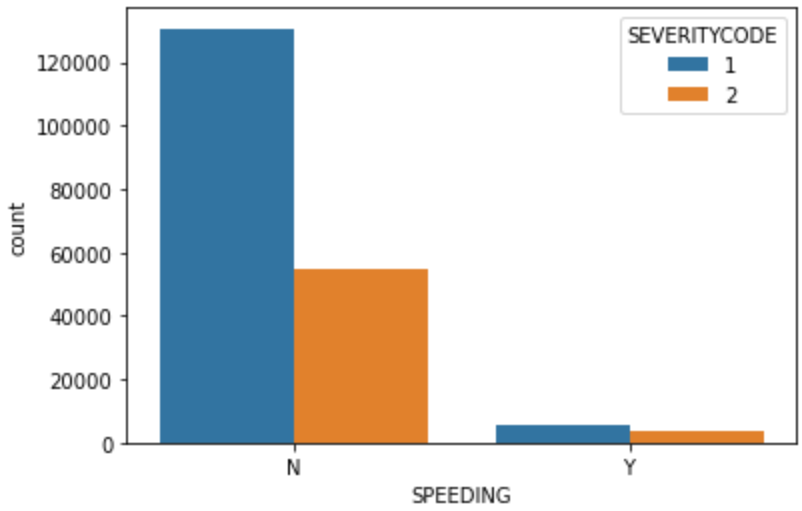

Figure 7-Speeding Graph

Speed has higher probability in injuries compared to baseline .

Feature 6- ROADCONDITIONS

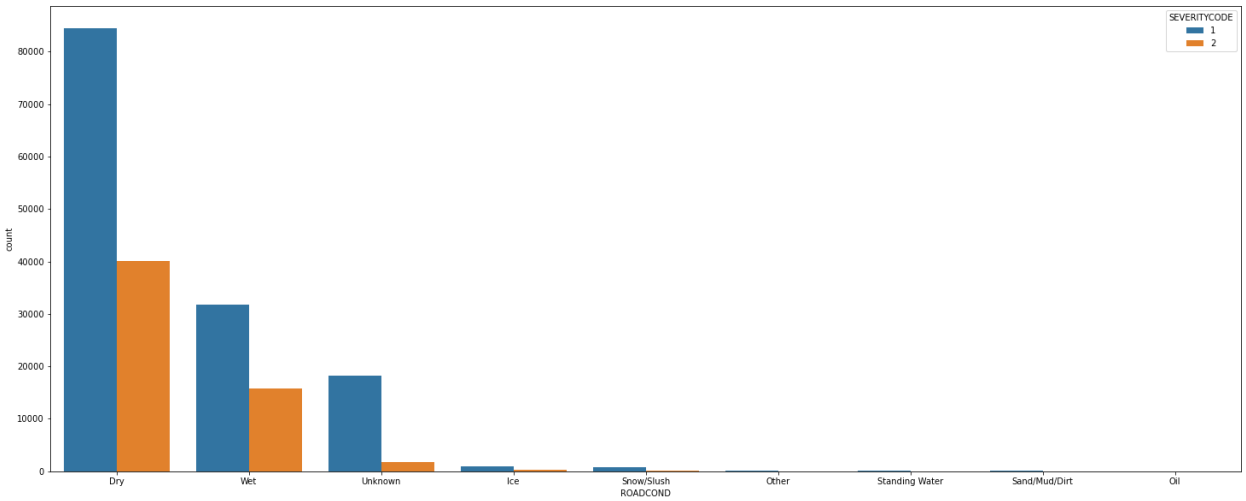

Figure 8-Road Conditions Graph

Road conditions appear to have similar results to weather. Worse road conditions may result in different driving habits which may have higher probability of property damage, but with more careful driving, may minimize the injury risks.

Feature 7-INATTENTION

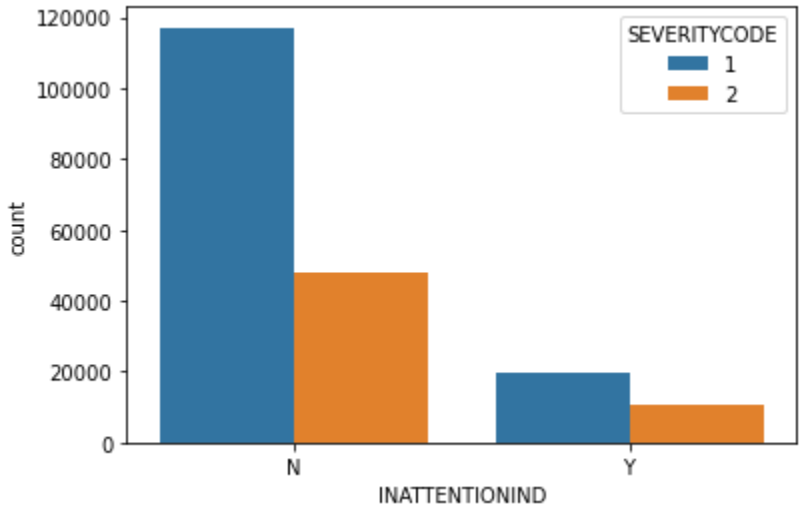

Figure 9-Inattention Graph

Inattention has higher likelihood of injuries than baseline.

Feature 8-HITPARKEDCAR

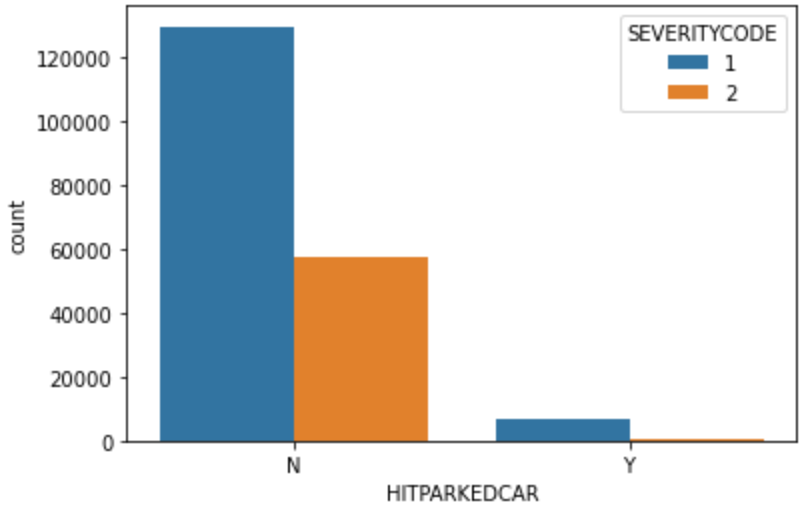

Figure 10- HitParkedCar Graph

Hit Parked car has higher property damage 93.8% of the time.

Future Advancements

The data does not include some important factors such as age,gender and model of the vehicle which can help in analysing the data more deeply. Also, based on the predictions a model can be created which can help travellers get real time information about the severity code at each junction in order to prevent them from involving in accidents.

Conclusion

This project was taken in order to analyze the severity of damage caused to property and to people during high traffic intensity. As, many values were dropped the data was less to give any further results. At par the Random Forest performed the best classifying the data for severity code to be predicted on the testing dataset.

References

1)Xavier Basagaña,Juan Pablo Escalera-Antezana, Payam Dadvand,Òscar Llatje,Jose Barrera-Gómez,Jordi Cunillera, Mercedes Medina-Ramón and Katherine Pérez(Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain)”High Ambient Temperatures and Risk of Motor Vehicle Crashes in Catalonia, Spain (2000–2011): A Time-Series Analysis”.Published online 2015 Dec.

2)Byeongjoon Noh,Hansaem Park,Hwasoo Yeo(Applied Science Research Institute, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseung-gu, Daejeon, Republic of Korea)”Analyzing vehicle–pedestrian interactions: Combining data cube structure and predictive collision risk estimation model”Published online 2021 December 17

3)Yuping Hu,Ye Li , Helai Huang , Jaeyoung Lee , Chen Yuan , Guoqing Zou “A high-resolution trajectory data driven method for real-time evaluation of traffic safety”Published online 2022 Feb

4)YounshikChung(Yeungnam University, Gyeungsan 38541, South Korea)”An application of in-vehicle recording technologies to analyze injury severity in crashes between taxis and two-wheelers” Published online 2021 Decmeber 24

5)DanniZhang,FeiChen,JiayunZhu,ChenzhuWang,JianchuanCheng,YunlongZhang,WuBo,PingZhang(School of Transportation, Southeast University, No.2 Dongnandaxue, Nanjing, Jiangsu 211189, PR China)”Research on drivers' hazard perception in plateau environment based on visual characteristics”Published online 2021 December 24.

6)QiangZeng,QianfangWang,XiaofeiWang(School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510641, PR China)”An empirical analysis of factors contributing to roadway infrastructure damage from expressway accidents: A Bayesian random parameters Tobit approach”Published online 2022 May 26

7)HaoliangChang,LishuaiLi,JianxiangHuang,QingpengZhang,Kwai-SangChin(Department of Advanced Design and Systems Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China,Faculty of Aerospace Engineering, Delft University of Technology, Postbus 5, 2600 AA Delft, Netherlands,School of Data Science, City University of Hong Kong, Kowloon, Hong Kong SAR, China,Department of Urban Planning and Design, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China)”Tracking traffic congestion and accidents using social media data: A case study of Shanghai” Published online 2022 February 26