Car Accident Severity

Data science Project

1. INTRODUCTION

1.1 BACKGROUND

With increase in population as well as urbanization we can also see the increase in number of road travel which finally leads to accidents.As most of the people prefers the road ways as the mode of transportation their chances of getting into accidents also increases.Therefore developing the system that is capable of predicting the probability of getting into accidents as well as the type of severity based on the data provided will help many to escape and handle the situations properly.

1.2 BUSINESS PROBLEM:

1. Many people lose their lives while driving either by four wheelers or two wheelers just because they dont take precautions or don’t have information about the weather condition or the road condition or any external factors.
2. In some cases the hospitals are not always ready for sudden new patients, so using this predictions we can make the hospitals be prepared for such cases.
3. Another problem is traffic officers or any other security services can be alarmed to monitor the locations where more accidents are likely to occur.
4. Often people get confused when more number of options are avaliable to travle from source to destinations and in many cases they choose the one with short distance which may not be the safest way to travle.
5. Better if insrance is covered for the vehicle used to travel.

Hence this project will be predicting the severity of the accidents that are likely to happen which aims help the target audiance who are

- People likely to travel in strange weathers.
- Police,governments,traffic officers
- Hospitals
- Vehicle insurance companies

and to solve the above mentioned common problems.

2. DATA

The data source for this project : https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

2.1 FEATURES USED

The data in initial stage contains 37 features out of which we will be using only the effective 13 features.

The following factors are used to solve this problem:-

LOCATION : Description of the general location of the collision

SEVERITYCODE :1 – Prop Damage , 2 – Injury (Target variable)

COLLISIONTYPE :Collision type

PERSONCOUNT :Total number of people involved in the collision

PEDCOUNT :Total number of pedestrians involved in the collision

PEDCYLCOUNT :Total number of bicycles involved in the collision

VEHCOUNT :Total number of vehicles involved in the collision

WEATHER :Weather conditions

ROADCOND :Road Conditions

LIGHTCOND :Light Conditions

SPEEDING :Wheather speeding was cause for accident

JUNCTIONTYPE :Type of Junction where accident occured.

UNDERINFL :Either driver was under drug or alcohol influence.

2.2 Examle data(initial without cleaning):

	SEVERITYCODE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	WEATHER	ROADCOND	LIGHTCOND	JUNCTIONTYPE	SPEEDING	U
0	2	Angles	2	0	0	2	Overcast	Wet	Daylight	At Intersection (intersection related)	NaN	
1	1	Sideswipe	2	0	0	2	Raining	Wet	Dark - Street Lights On	Mid-Block (not related to intersection)	NaN	
2	1	Parked Car	4	0	0	3	Overcast	Dry	Daylight	Mid-Block (not related to intersection)	NaN	
3	1	Other	3	0	0	3	Clear	Dry	Daylight	Mid-Block (not related to intersection)	NaN	
4	2	Angles	2	0	0	2	Raining	Wet	Daylight	At Intersection (intersection related)	NaN	

2.3 DATA CLEANING

From the above pic of the example data which is the raw format with any modification, this type of dataset cannot be directly used for any operations in data science process. So it is very mandatory step to modify, clean, trabsform the dataset which suits for the model to analyse and prediction properly.

Features (the columns which are used for prediction) contains both the numerical as well as categorical data. So the first step is to segregate the features based on the type of data it holds.

For the numerical columns the missing values are replaced using the python library SimpleImputer which replaces the missing value(NaN) with the mean of that columns.

For the categorical columns the SimpleImputer used with strategy of most frequent value occured in a particular column

Among total of 37 columns in the initial raw dataset only 13 columns are used going further.

The above mentioned 13 columns are selected based on the good correlation with the target and the rest are removed because of the reasons like some of them conatined too many constant value or NaN, some with many categorical values which doesn’t make sense in converting into the format suitable for the machine learning models.

The categorical columns with reasonable unique values are converted using either

of the library in python i.e LableEncoder or OneHotEncoder

2.4 Explanation how data can help the prediction:

1. Starting with the main factors which are weeather condition,road condition,light condition all three can be considered we effective features for the predicting the severity.
2. Collision type and junction type describes which type of collision are more likely cause severity
3. Vehicle,bicycles,pedestrian count aslo affects prediction as more traffic can cause high probability for accidents.
4. Information about whether the driver was under drug or alcohol consumption while driving can warn the other drivers before their travel.
5. Location data can give alternate travel route if their prefered route is more likely to get into accidents.

3. METHODOLOGY

As this project deals with analysisng and predicting the severity and possibility of getting into accident , this can be clearly stated that this is a classification problem.

So according to the general principle of data science the necessary exploratory data analysis and the appropriate machine learning models which are suitable for classification problem shall be used in this project.

For analysing the data, different graphs were created to interpret how different features affect the severity code which is the target in this dataset.

For Visualization matplotlib and Seaborn, these two libraries of python are used.

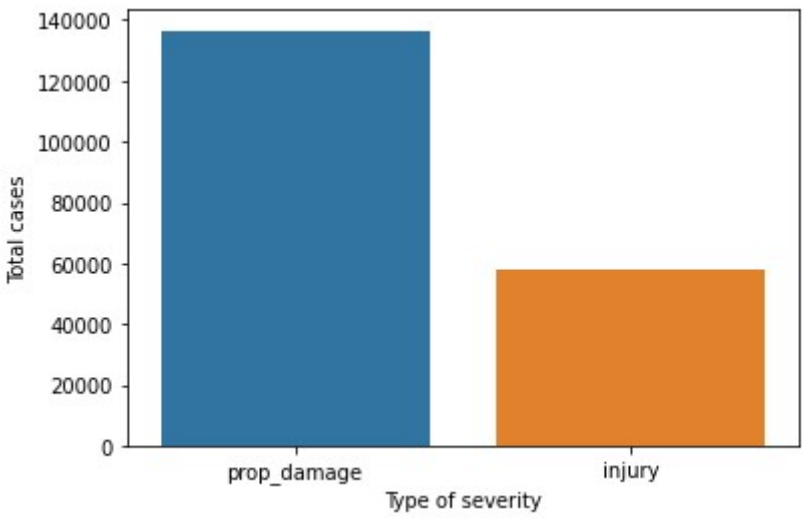
The target variable which is the “severitycode” indicates 1 for “property damage” (can be any damage which does not include any harm to human body like vehicle damage etc) and 2 for “injury”

3.1 Differene between the total number of cases for each severity

According to the graph which shows the total cases for each severity , it can be clearly stated the the property damage is very high as comparable to the accidents which involves the injury to the human body.

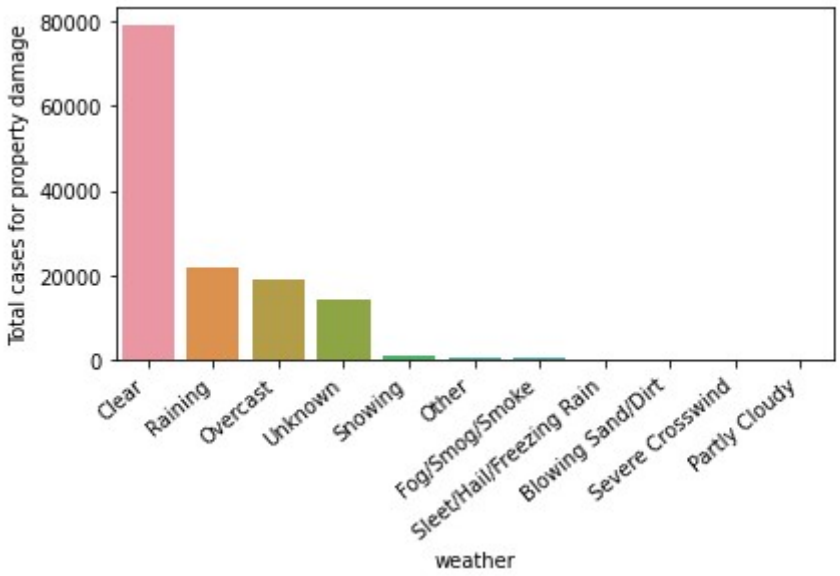
The total cases for property damage is 136485

and that for injury is 58188



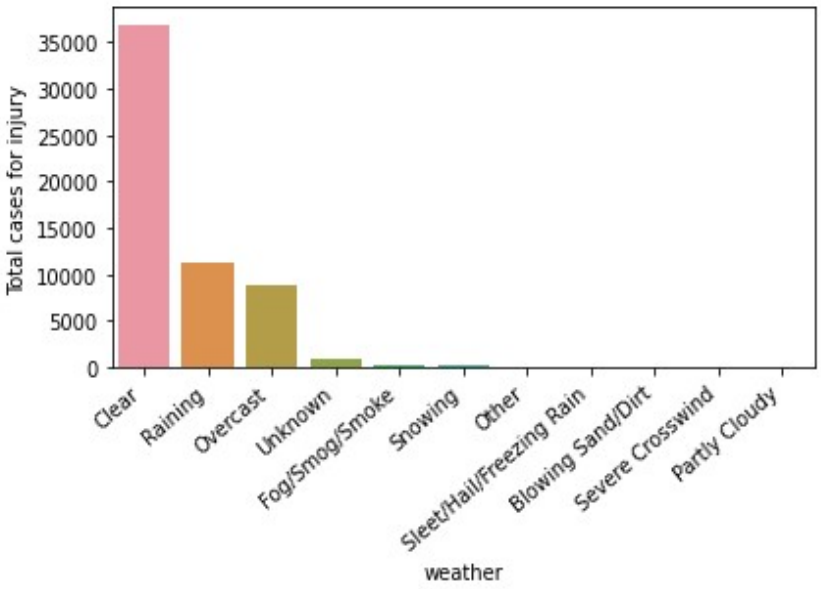
3.2 Affect of weather in accidents

Weather can be considered as one of the common reasons for sudden damage to the human properties. Various weathers can be directly related to acciednts as it makes the travel more dangerous, e.g during the rainy day the road get wet which reduces the friction between the vehical's wheel and road surface and finally leads to skiting of the vehicals. Same holds for the Snowy and more moistured day.



This bar graph shows the total number of cases for the property damage and strange thing noted here is that it's not the weather which causes most of the cases, as it can been seen that clear weather has most of the cases followed by rainy,overcast,snow

Same things holds for the accident cases with injury, with majority of accidents occuring during the clear weather.
Although as compared weather is not much influencial precautions needs to be taken for the rainy and snowing like weather conditions



From previous two graphs it can be stated the some other factors holds strong correlation with accidents as most of the accidents took place with clear weather.

3.3 Road Conditions

Another important factor is the road conditions while, as discussed the friction and the grip between the road and the vehicle plays major role during travel. Let’s see how this affects the number of cases.
To show the contributions Pie chart has been used

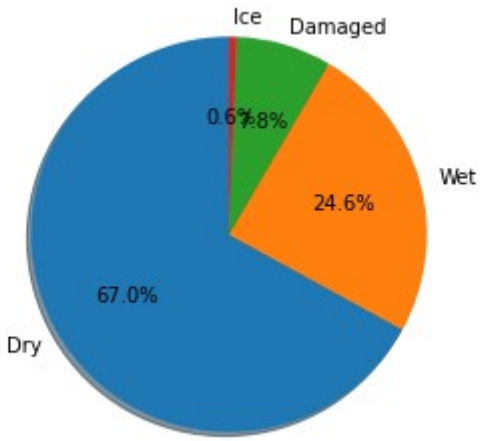
The percentage of contribution :

DRY : 67.0%

WET : 24.6%

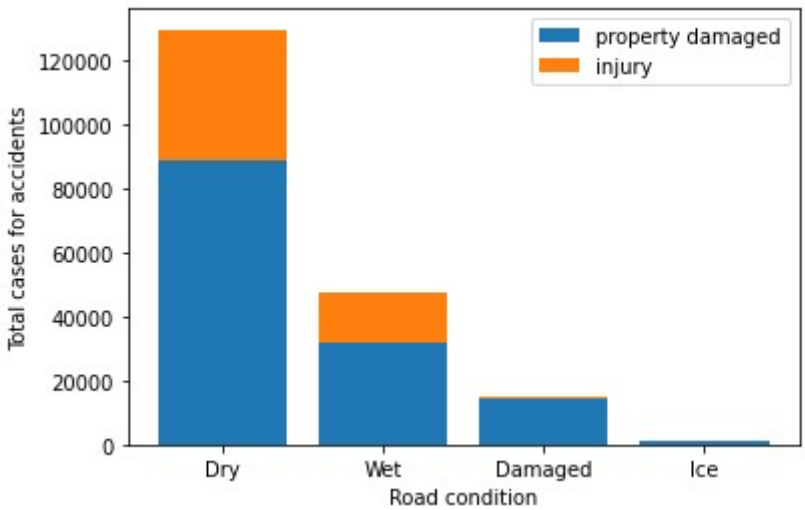
Damaged : 8.0%

ICE : 0.6%



Majority of the contribution comes with Dry and none damaged road followed by the wet and damaged roads.

Also from this chart we can guess that some other factor is affecting the accidents because the contribution of wet, damaged and ice roads as noticably small when compared to dry roads.
Here the damaged category refers to conditions like pit holes, paddles and roads with frequent ups and downs.

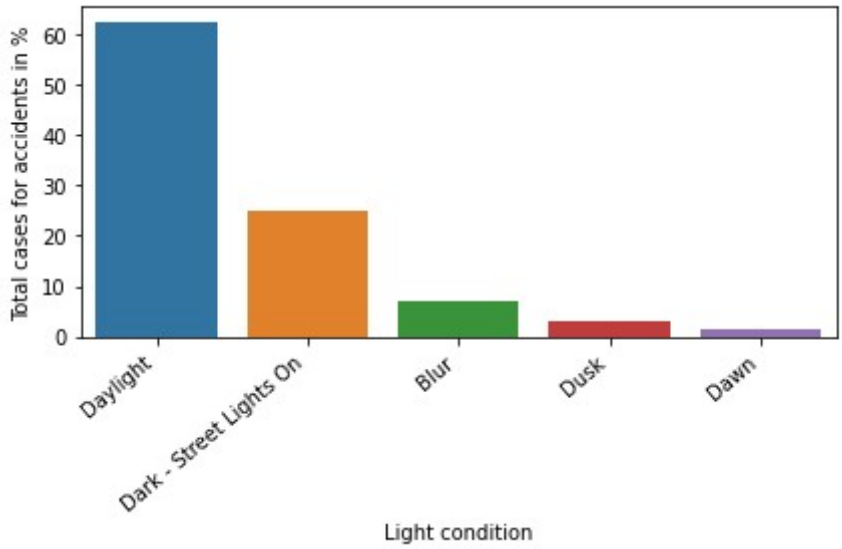


From the above stacked bar graph with combined data for severity it can be stated that road conditions has can be linked to high chances of property damage.

In the cases when Road is filled with ice the injury risk is almost near to neglegible.

3.4 Light Conditions while driving

Common thing is that most of the people are comfortable driving in day light. The visibility of surrounding things is important factor for the person who is driving.



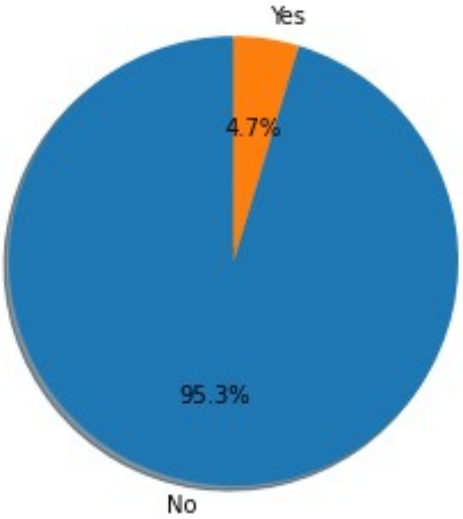
From the above graph most of the accidents took under daylight condition which is near to 60%, followed by the dark. Blur condition refers to the quality of the from window or due to water droplets in rainy weather.Dusk and Dawn are near to negligible.

3.5 Affect of any kind of addiction habbit of the driver

Health and consciousness condition of the driver also needs to be considered as this kind of variable is directly proportional to the probability of the accidenct.Here the cases like taking drugs or alcohol is considered.
From the data used, following pie chart shows the contribution of the driver under influence.

95.3% people with accident case found to be clear and didn’t took any kind of addiction.

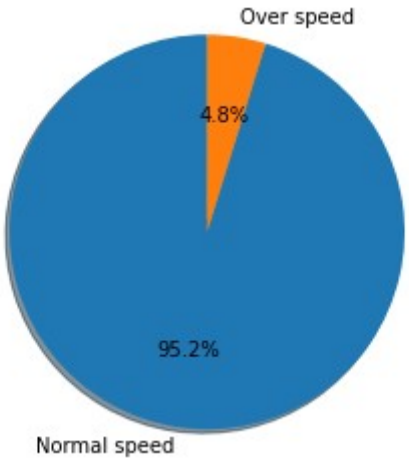
Where as 4.7% were under some kind of addiction. Although the percentage is less compared to another side but still this kind of correlation has direct effect because probability of getting into accidents is almost equal to 100%



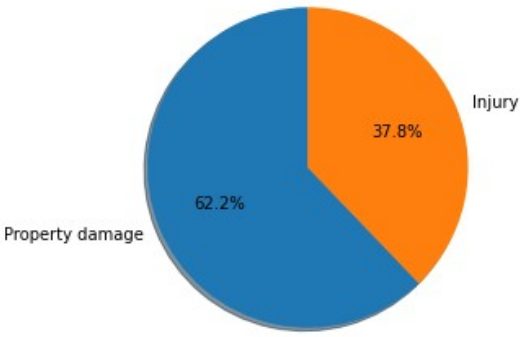
3.6 Affect of Speeding or crossing speed limit

To have the save journey, maintaining the speed limit is very crucial. Over speeding can lead to losing control over vehicle and also in violation in the traffic law.
Speed limit is set based on the curvature and the slope of the road.Hence over accelarating vehicle above the speed limit cause unmatch between other vehicles in the road which definitely leads to collision.

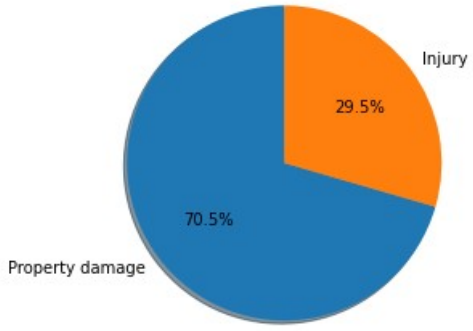
The sub Pie chart shows the % of type of severity.



For Normal Speed



For Over Speed

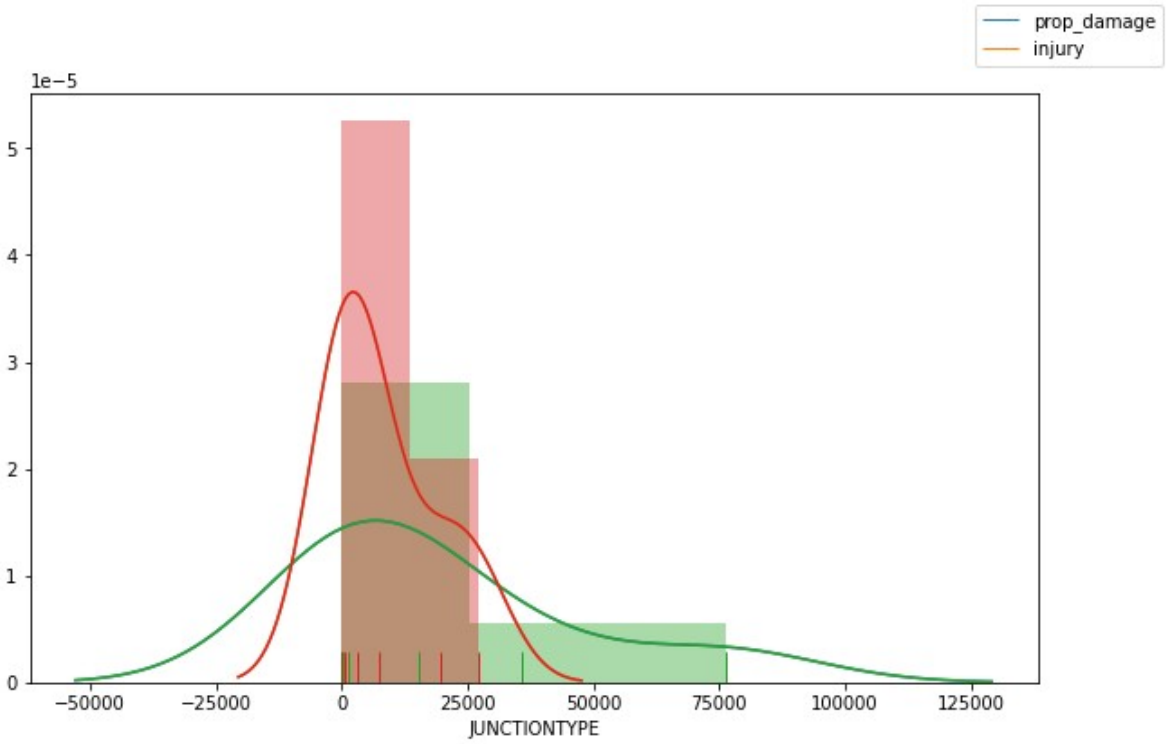


3.7 Affect of Junction type

Junction type is considered to be one of the most influencial factor for the accident cases. Junction type can be categorised into two main divisions i.e Intersections and Blocks.

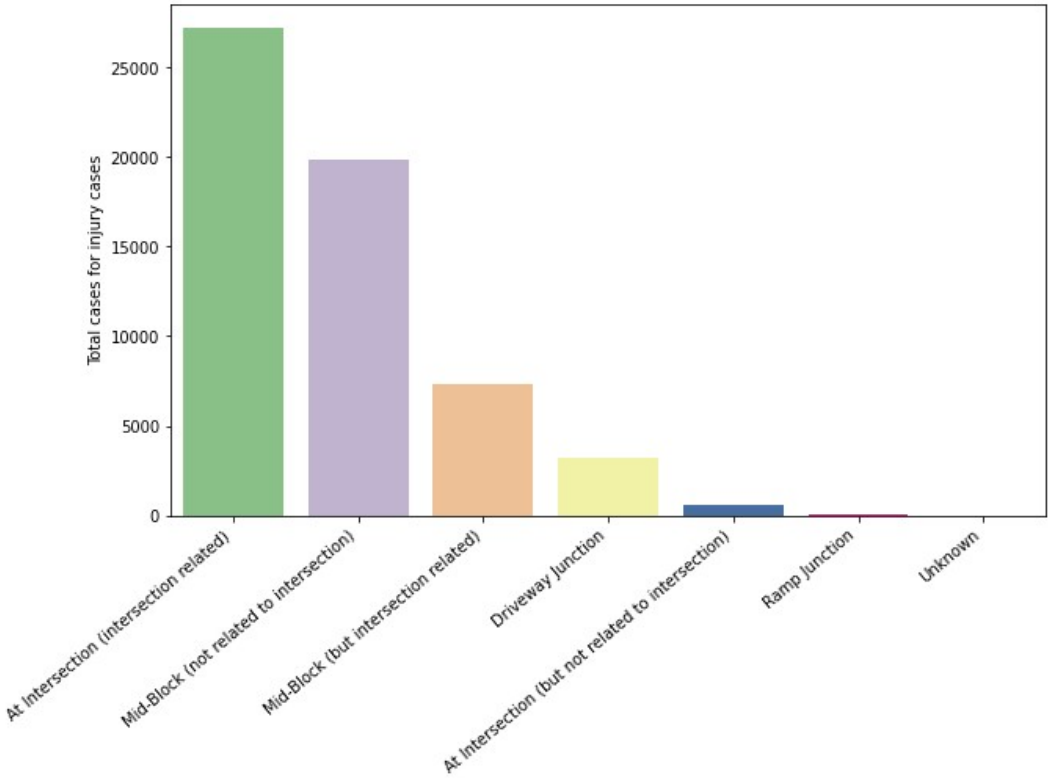
The distribution graph show how the number of cases varies for the property damage and injury type accidents.

Most of the accidents with mid block as junction types also includes the parked vehicle damage.



From the above distribution graph the width of the property damage is larger then injury, but for injury even the width is narrow for certain junction type the number of accidents are high.

Hence it will be better to investigate for which value of injury severity the number of cases are high using the bar graph.



The highest number of cases goes with the intersection followed by the mid block.
Driveway and Ramp junction doesn’t involve much into the accident cases.

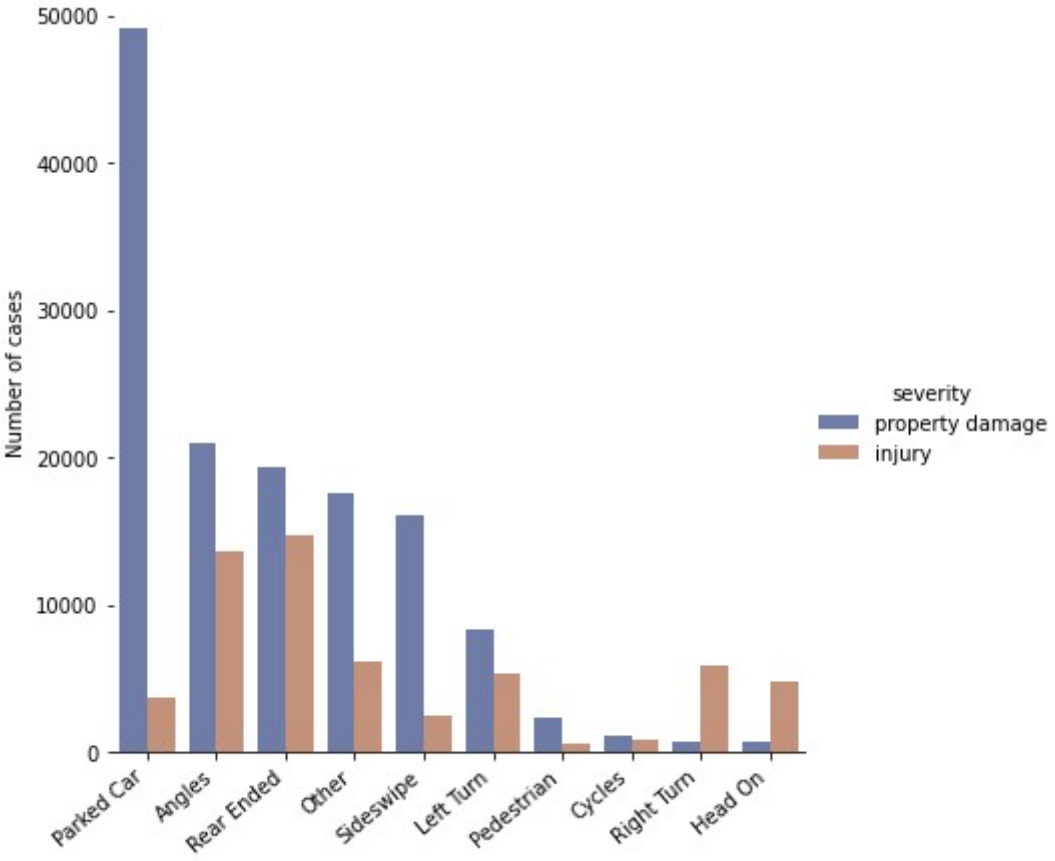
The below fig shows how accident because of intersection looks like.

3.8 Affect of collision type

Based on the type of collision between two objects, it can be determined how close is the case to accident severity.The suppied data contains 10 different types of collisions and the bar graph below show which of them has highest accidents cases.

The types of collisions involved are parked car which means that one of the objects hits the parked car in the mid block and mostly does not involve any injury cases.Right or left turn are the collisions occuring when a vehicle tries to take a turn either way. Pedestrians or cycles involved in collision mostly leads to property damage.

From the below grouped bar chart, property damage type of accidents involve parked car, angles, rear ended, sideswipe.
For injury type of accidents we have right turn, head on, rear ended, angles as the majority collision type.



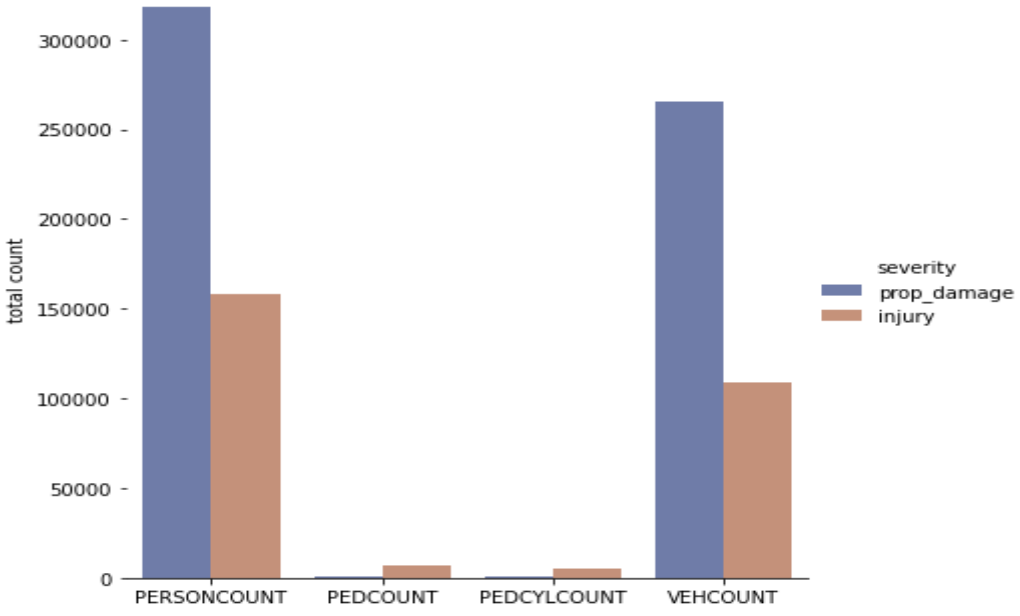
3.9 Affect of number of people,pedestrians,vehicles involved in accident

Number of people inside the vehicle also affect the accident cases, e.g if car has more than 5 people may cause reduce in the speed, also person driving may drive carefully.

Traffic condiction varies with number of vehicle, therefore it can also be considered as important factor.

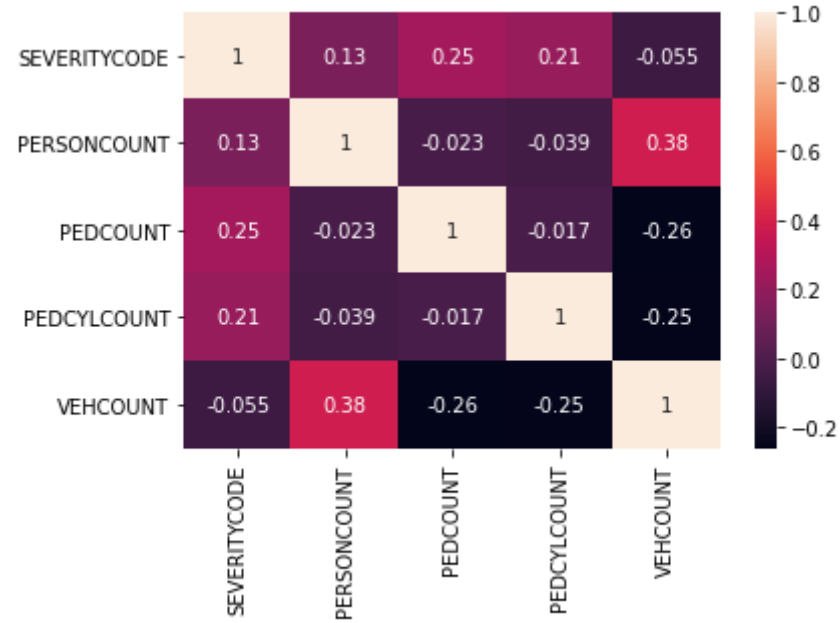
The grouped bar graph show the variance in number of cases along different types

Compared to all four types the person count and vehicle count has most number of cases. Other factors like pedestrians count and bicycle count can be ignored as it is nearly to negligible.

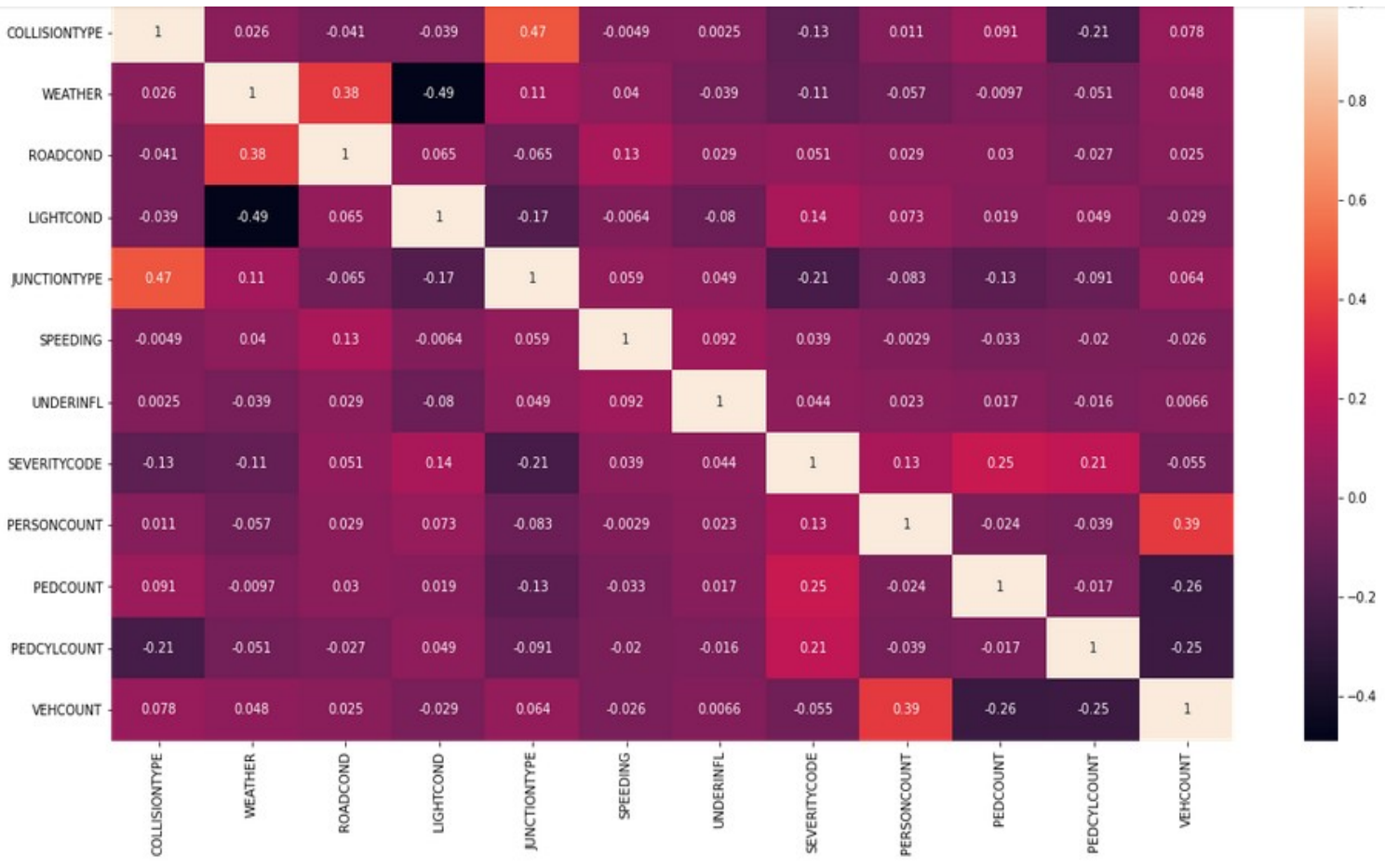


The following heat map show the correlation between the above four categories with the severity.

Extreme Positive show strong relation with injury type of accidents and Extreme negative shows strong relation with property damage type of accidents.



To conclude the Exploratory data analysis section the below heatmap shows the correlation with all the features included in the dataset.



4 PREDICTION MODLES

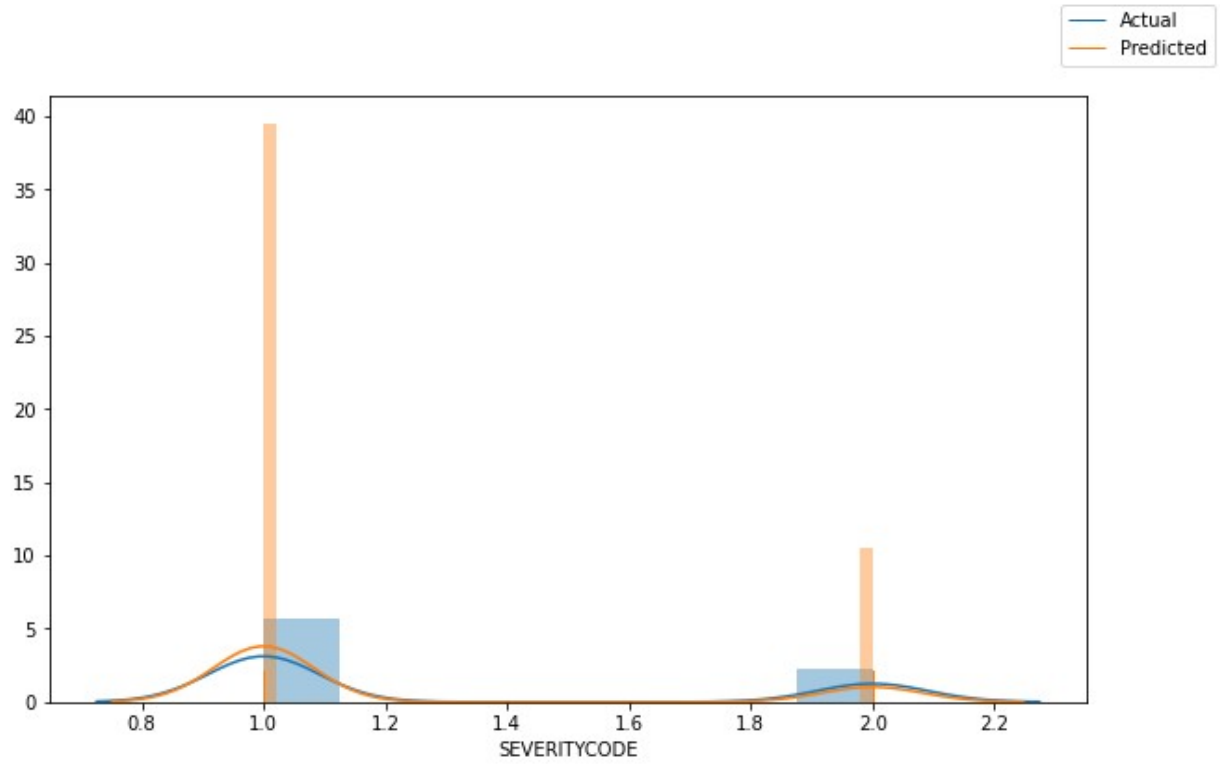
As the aim of this project is to analyse and predict the type of severity whcih is either property damage or injury, this problem strongly relates to classification problem.

Hence the supervised machine learning method will be applied for the dataset. Four most effective classification model which are 1
1) K means nearest neighbour
2) Random forest Classifier
3) SVM
4) Logistic regression
Along with XGB classifier are used to predict the accident severity.

As all of the above are classification models the appropriate metrics like Mean Absolute Error, Accuracy score, Jaccard score, f1 score, log loss score are condidered to deterimine which model performs best.

Data preprocessing steps like using Simpleimputer, label encoder or Onehot encoder are performed before training the model.

Below graph show the distribution between the actual and predicted value.



5 Result and Solution

The parameters for all the models are set to default.

The below table shows the metrics values for each model.

	MAE	Accuracy score	Jaccard score	F1 score	Log loss score
K Means nearest neighbour	0.29	0.71	0.67	0.70	9.9
Random forest classifier	0.26	0.73	0.70	0.71	9.04
SVM	0.24	0.75	0.73	0.70	8.34
Logistic regression	0.23	0.76	0.74	0.72	8.10
XGB classifier	0.24	0.75	0.72	0.73	8.5

Among all of the models **SVM** has the least mae and highest accuracy score.

From the predicted data and from the observations there is descent correlation between the all the selected features and target variable which is severitycode. The average accuracy for all the models is neary 70%, from the features if all the numerical feature’s value is reduced then the accident severity is close to 1, i.e the accodent is more likely to cause the property damage only.
Another main feature is the junction type because the model predicted exact category of severity according to the each match in the type of junction.

Similarly the collision type also shows strong correlation and can be considered as effective because the prediction outcome varies exactly with the match of type of collision.

Other features like weather, road, light condition doesn’t have much influence in the prediction outcome.

From the above observations the hospitals as well as traffic officers need to be prepared which comes under the route where junctions are more and the number of travellers are also high.
Apart from that when weather conditions are not as expected or the road condition are not proper both, the vehical owner and the insurance agency must be sure that their insurance package is coverd.

Conclusion

From this study it can be noted that how important are the features used in this project are important to relate the accident severity making the target auidance much prepared for unexpected situations.The model used for prediction has all the parameters as default and can be improved in terms of accuracy by performing tunnig the parameters.Based on the types of junctions and collision, the model can vary in prediction if more types are passed in the data.Hence providing the model using which the target auidance can plan for their benefits and precautions.