PREDICTING THE SEVERITY OF AN ACCIDENT

1. BACKGROUND:

Accidents occur daily. A traffic collision is also known as a car accident. Whenever a vehicle has collided with some object like a person, animal, road cones, building etc. an accident might occur. So many accidents are fatal. There are so many factors that are responsible for the accident to take place. It might be possible that the person who is driving the car was not paying attention, the person might be drunk while driving, the person was driving at a higher speed and was not able to get it under control when an object came near it or the traffic light didn't work at a certain point for few seconds and 2 cars coming from different sides collided or improper roads, etc. People need to be aware and in a conscious state while driving the car, so as to avoid such accidents and the traffic systems also need to be monitored so as to ensure that such accidents do not happen.

There are different approaches used to monitor the traffic safety problems: reactive approach and proactive approach. In the reactive approach, the required improvements or changes are made to the variable that might act as a medium in an accident. Example, an unsafe work environment where the materials/instruments were not kept properly or in a safe place. Making sure that the instruments are kept in a proper way so that they do not harm anyone. Proactive approach, on the other hand, can include a collision prevention approach. Example, if a road is not in proper condition some kind of accident might take place. To avoid that, the access to that road itself will be blocked.

We mainly focus on the proactive approach, where certain calculations or estimations are already made to avoid any future accidents. The important factors in an accident are identified, it will be helpful to identify other similar potential accidents. And in this report, the word accident and collision are used interchangeably.

2. PROBLEM STATEMENT: (mention who would be interested.

How do accidents occur or what can be the possible reasons for an accident to occur? If an accident occurs, how will a person be able to identify the severity of the accident? Also, identify the factors responsible to understand how severe the accident actually is.

3. DESCRIPTION OF DATA:

The data set which will be used is the Data-Collisions provided through the Applied Data Science Capstone by Coursera. The data consists of the information about the accidents. The data has 194,673 rows and 38 columns. The columns are referred as the features of the dataset.

The features are:

- OBJECTID: it specifies the ESRI unique identifier.
- SHAPE: specifies the ESRI geometry field.
- INCKEY: specifies a unique key for incident occurred.
- COLDETKEY: specifies a secondary key for incident
- ADDRTYPE: specifies the address of the type of collision: Alley, Block, Intersection
- INTKEY: specified the key corresponding to the intersection which is associated with a collision.
- LOCATION: specifies where the collision/accident occurred.
- SEVERITYCODE: specifies a particular code which corresponds to the severity of the collision:

  3 – Fatality
  2b – Serious injury
  2 – Injury
  1 – Prop Damage
  0 - Unknown

- SEVERITYDESC: specifies a detailed description of the severity of the collision.
- COLLISIONTYPE: type of collision
- PERSONCOUNT: specifies the total number of people involved in the collision
- PEDCOUNT: specifies the number of pedestrians involved in the collision.
- PEDCYLCOUNT: specifies the number of bicycles involved in the collision. This is provided by the state itself.
- VEHCOUNT: specifies the number of vehicles involved in a collision.
- INJURIES: specifies the total number of injuries in the collision which is provided by the state.
- SERIOUSINJURIES: specifies the number of serious injuries in the collision as provided by the state.
- FATALITIES: specifies the number of fatalities in the collision.
- INCDATE: specifies the date of the incident.
- INCDTTM: specifies the date and time of incident.
- JUNCTIONTYPE: specifies the category of junction at which the collision took place.
- SDOT_COLCODE: specifies the code given to collision by SDOT.
- SDOT_COLDESC: specifies the description of the collision corresponding to the collision code.
- INATTENTIONIND: specifies whether the collision was due to inattention. (Y/N)
- UNDERINFL: specifies whether the driver involved in the collision was under the influence of drugs, alcohol or not.
- WEATHER: specifies the weather condition at the time of collision.
- ROADCOND: specifies the condition of the road during the collision.
- LIGHTCOND: specifies the conditions of light during the collision.
- PEDROWNOTGRNT: specifies whether or not the pedestrian right of way was not granted. (Y/N)

- SDOTCOLNUM: specifies the number given to collision by SDOT.
- SPEEDING: specifies whether or not speeding was a factor in the collision. (Y/N)
- ST_COLCODE: provides a code by the state that describes the collision.
- ST_COLDESC: gives a description that corresponds to the state's coding designation.
- SEGLANEKEY: specifies the key for the lane segment in which the collision occurred.
- CROSSWALKKEY: specifies the crosswalk at which the collision occurred.
- HITPARKEDCAR: specifies whether or not the collision involved hitting a parked car. (Y/N)

State Collision Code Dictionary

Code	Description
0	Vehicle Going Straight Hits Pedestrian
1	Vehicle Turning Right Hits Pedestrian
2	Vehicle Turning Left Hits Pedestrian
3	Vehicle Backing Hits Pedestrian
4	Vehicle Hits Pedestrian - All Other Actions
5	Vehicle Hits Pedestrian - Actions Not Stated
10	Entering At Angle
11	From Same Direction -Both Going Straight-Both Moving- Sideswipe
12	From Same Direction -Both Going Straight-One Stopped- Sideswipe
13	From Same Direction - Both Going Straight - Both Moving - Rear End
14	From Same Direction - Both Going Straight - One Stopped - Rear End
15	From Same Direction - One Left Turn - One Straight
16	From Same Direction - One Right Turn - One Straight
19	One Car Entering Parked Position
20	One Car Leaving Parked Position
21	One Car Entering Driveway Access
22	One Car Leaving Driveway Access
23	From Same Direction - All Others
24	From Opposite Direction - Both Moving - Head On
25	From Opposite Direction - One Stopped - Head On
26	From Opposite Direction - Both Going Straight - sideswipe
27	From Opposite Direction - Both Going Straight - One Stopped - sideswipe
28	From Opposite Direction - One Left Turn - One Straight
29	From Opposite Direction - One Left Turn - One Right Turn
30	From Opposite Direction - All Others
31	Not Stated
32	One Parked - One Moving
40	Train Struck Moving Vehicle
41	Train Struck Stopped or Stalled Vehicle
42	Vehicle Struck Moving Train
43	Vehicle Struck Stopped Train
44	Unicycle
46	Tricycle
47	Domestic Animal (horse, cow, sheep, etc)
48	Domestic Animal Other (Cat, Dog etc)
49	Non Domestic Animal (deer, bear, elk, etc)
50	Struck Fixed Object
51	Struck Other Object
52	Vehicle Overturned
53	Person Fell, Jumped, or was Pushed From Vehicle
54	Fire Started In Vehicle
55	Accidently Overcame By Carbon Monoxide Poison

56	Breakage Of Any Part Of the Vehicle Resulting In Injury or in Further Property Damage
57	All Other Non-Collisions
60	Vehicle Hits State Road or Construction Machinery
61	Vehicle Struck By State Road or Construction Machinery
62	Vehicle Hits County Road or Construction Machinery
63	Vehicle Struck By County Road or Construction Machinery
64	Vehicle Hits City Road or Construction Machinery
65	Vehicle Struck By City Road or Construction Machinery
66	Vehicle Hits Other Road or Construction Machinery
67	Vehicle Struck by Other Road or Construction Machinery
71	Same Direction - Both Turning Right - Both Moving - Sideswipe
72	Same Direction - Both Turning Right - One Stopped - Sideswipe
73	Same Direction - Both Turning Right - Both Moving - Rear End
74	Same Direction - Both Turning Right - One Stopped - Rear End
81	Same Direction - Both Turning Left - Both Moving - Sideswipe
82	Same Direction - Both Turning Left - One Stopped - Sideswipe
83	Same Direction - Both Turning Left - Both Moving - Rear End
84	Same Direction - Both Turning Left - One Stopped - Rear End

4. METHODOLOGY

4.1 SUPERVISED LEARNING:

It is the type of machine learning in which the models knows what type of data is to be predicted. There are certain input variables X and output variables Y and then we use an algorithm to learn the mapping function from the input to output.

It was chosen because, our solution aims to predict the severity of the accident. For which there was a dataset present which consists of many features like total number of persons, what type of accident/collision was it, how many people were injured, the type of junction at which the accident took place. And also a feature for which the values are to be predict, is the severity code. It indicates how severe the accident is.

The supervised learning can be further grouped into classification and regression.

4.1.1 CLASSIFICATION:

Classification a type of supervised learning was chosen because it will help in predicting a particular category. It is mostly used when our outcome is not a continuous value, but some particular value from certain set of values.

Our solution aims to predict the severity code as mentioned before. These include: 3 – Fatality, 2b – Serious injury, 2 – Injury, 1 – Prop Damage, 0 – Unknown.

There are different algorithms in classification which can be used to do so.

- Logistic Regression: In this, the probabilities that describe the possible outcomes of a single trial are done using a logistic function.

- Naïve Bayes: It is based on Bayes theorem with the assumption of independence between every pair of features.
- Stochastic Gradient Descent: Mostly used where there is large datasets. It takes one example, looks at the slope and tries to move downwards.
- K-Nearest Neighbours: it stores all the available cases with their respective classes or categories and classifies the new cases base on similarity measure like distance functions, like Euclidean distance, Manhattan distance, etc.
- Decision Tree: it creates a classification by building a decision tree in which each node specifies a test on the feature and each branch that goes further from that node corresponds to one of the many possible values for that particular feature.
- Random Forest: it is just like decision tree. But adds more randomness to it while creating the trees. Instead of searching for the important feature when splitting a node, it will search for the best feature among a random subset of features.
- Support Vector Machine: it is mostly used for two group classification. After being provided with labelled training ata for each class or category, it can categorize new text.

4.1.2 ALGORITHM USED:

Different algorithms were tried on the data. This can be seen in the following:

	Name	CROSS VALIDATION SCORE	Accuracy Score	F1-SCORE	RECALL SCORE	ROC AUC SCORE
0	K NEAREST NEIGHBOURS	0.735220	0.738389	0.829792	0.922786	0.624352
1	STOCHASTIC GRADIENT DESCENT	0.641212	0.721563	0.819504	0.914670	0.602140
2	DECISION TREE CLASSIFIER	0.739288	0.736948	0.829404	0.925320	0.620453
3	LOGISTIC REGRESSION	0.720810	0.724357	0.822691	0.925352	0.600054
4	RANDOM FOREST CLASSIFIER	0.740945	0.738544	0.829626	0.921150	0.625615
5	NAIVE BAYES	0.637965	0.634419	0.710374	0.648767	0.625546

Decision tree classifier was chosen. And its hyperparameters were tuned. And after hyperparameter tuning, the scores were as follows:

	Name	CROSS VALIDATION SCORE	Accuracy Score	F1-SCORE	RECALL SCORE	ROC AUC SCORE
0	DECISION TREE CLASSIFIER	0.748289	0.745949	0.838651	0.95541	0.616411

4.2 EXPLORATORY DATA ANALYSIS

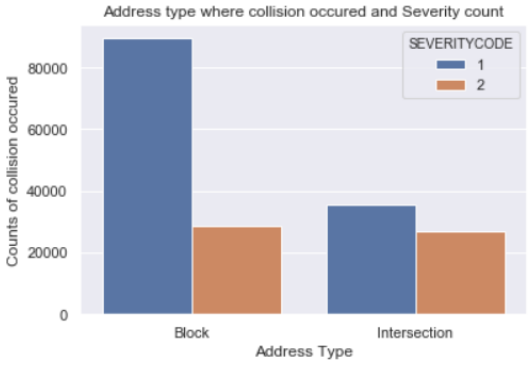
It is an approach where the data sets are analysed to summarize their characteristics or to identify patterns. It can also include visually representing the data.

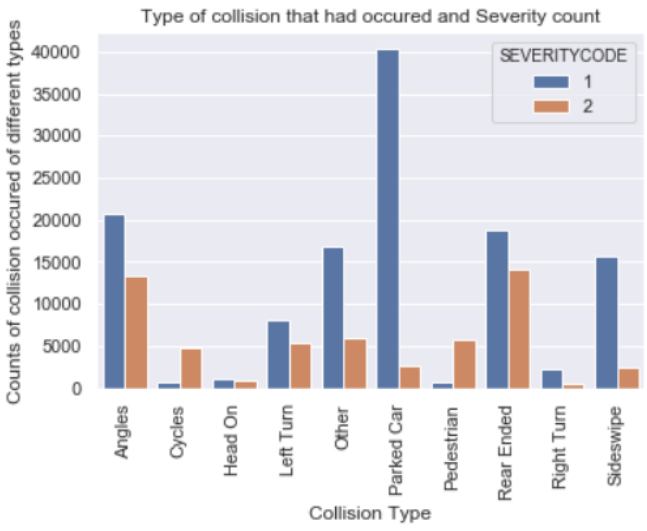
Describing our data:

```
updated_data.describe()
```

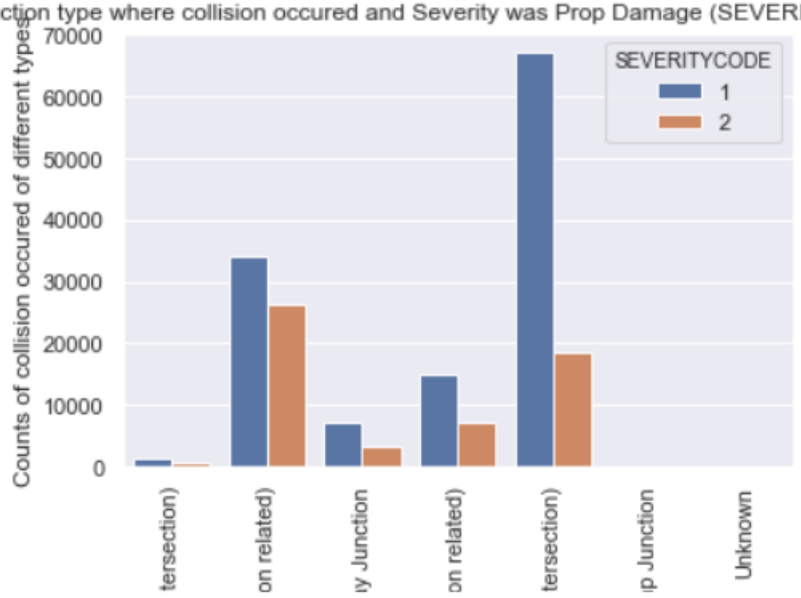
	PERSONCOUNT	VEHCOUNT	SEVERITYCODE
count	182660.000000	182660.000000	182660.000000
mean	2.476902	1.972358	1.309843
std	1.371036	0.563108	0.462430
min	0.000000	0.000000	1.000000
25%	2.000000	2.000000	1.000000
50%	2.000000	2.000000	1.000000
75%	3.000000	2.000000	2.000000
max	81.000000	12.000000	2.000000

4.2.1 VISUZALING THE DATA

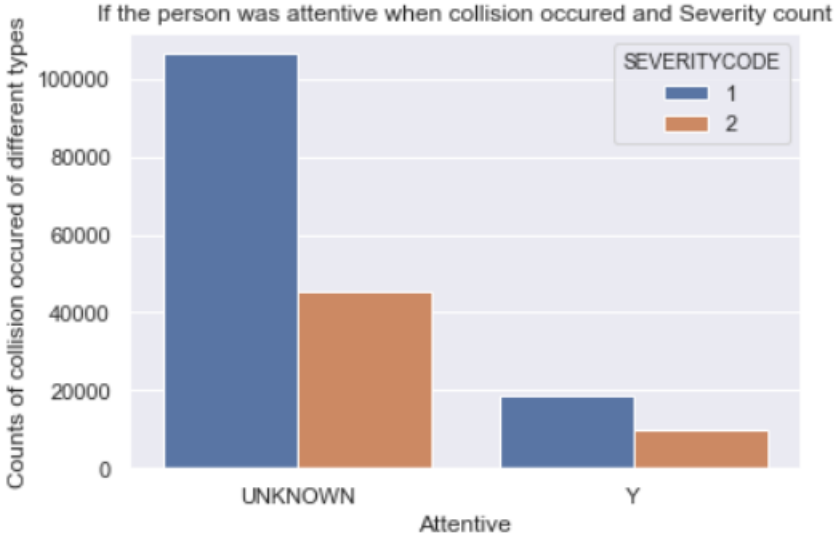




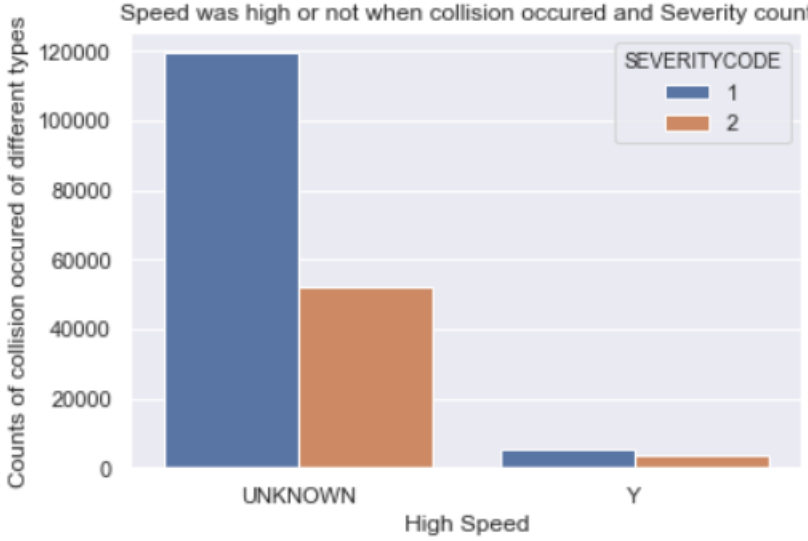
Junction type where collision occured and Severity was Prop Damage (SEVERITYCODE = 1)

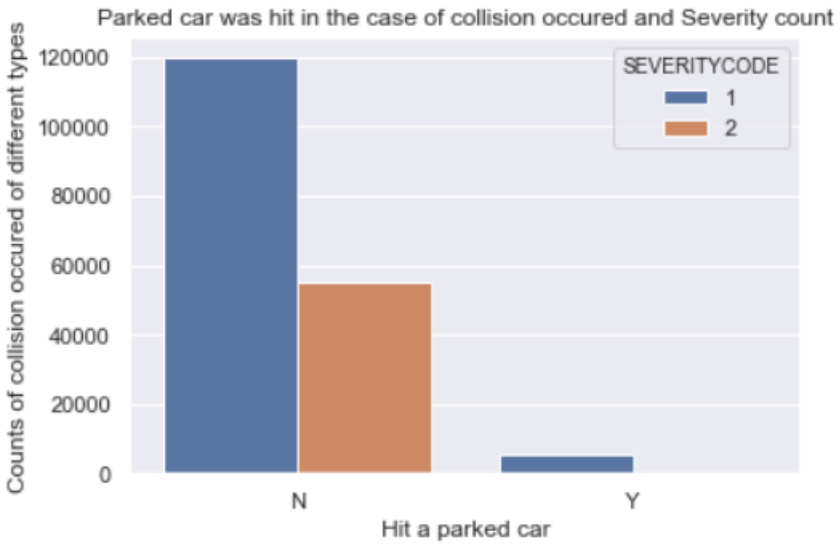


If the person was attentive when collision occured and Severity count



Speed was high or not when collision occured and Severity count





5. RESULTS

As mentioned above, the following results were obtained.

	Name	CROSS VALIDATION SCORE	Accuracy Score	F1-SCORE	RECALL SCORE	ROC AUC SCORE
0	DECISION TREE CLASSIFIER	0.748289	0.745949	0.838651	0.95541	0.616411

6. DISCUSSION

There were so much of data that was missing. Thus, few of the features had to be dropped or few of the rows with missing values were dropped. And because of this, only 2 classes present in SEVERITYCODE feature were present in the whole data of 194k rows, which was class 1 – prop damage and class 2 – injury.

7. CONCLUSION

More such data can be taken from different sources. They can be cleaned and combined together and can be used to make better predictions and ot be made sure that the data gathered is correct.