Car accident severity

Introduction:

Road accidents are the most unwanted thing to happen to a person and family, still it happens very frequently. Nearly 1.3 million people lose life due to car accidents every year all over the globe. Drivers are aware of traffic rules and conditions still accidents do happen and result in loss of life and property. Data related to such incidents are available in public domain which can be used to analyze the scenario, conditions and severity accident. Thus, enabling the governments and concerned authority to make roads safer and make people aware of situations which can lead to accidents.

Business Problem:

A car accident results in huge cost, according to US government data shows each road accident result in loss of $60k approx. this cost can be attributed to the damage to property, injuries treatment, loss of workday, emergency response, insurance claims. Thus, avoiding a car accident can save lot of money and life, resulting in positive impact on the economy.

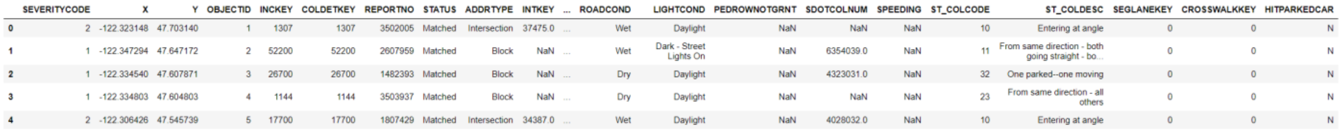
As car accidents and it severity is dependent on many factors, taking into consideration factors like weather condition, speed of the car, state of the road, traffic. Using these factors to predict the severity of the outcome of the accident can help the city administration take necessary measures and impose restriction to minimize loss, in terms of man, material and money.

Thus, a model able to predict any such outcome can be very desirable.

Data:

To design any such model which can predict the severity of the road accidents, we will be using Data-Collisions data set provided in the course of this module.

This data set provides the details like., data/time, junction type, road condition, weather, light condition, speeding, under influence, severity. Thus, this can help in predicting the severity in case we know the other factors.



	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING	ST_COLCODE	ST_COLDESC	SEGLANEKEY	CROSSWALKKEY	HITPARKEDCAR
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	NaN	NaN	NaN	10	Entering at angle	0	0	N
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	NaN	6354039.0	NaN	11	From same direction - both going straight - bo...	0	0	N
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	NaN	4323031.0	NaN	32	One parked--one moving	0	0	N
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	NaN	NaN	NaN	23	From same direction - all others	0	0	N
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	NaN	4028032.0	NaN	10	Entering at angle	0	0	N

Data Cleaning and features:

Data was downloaded from the Coursera course. The data had lot of missing values; some features also had only entries where answer was ‘N’.

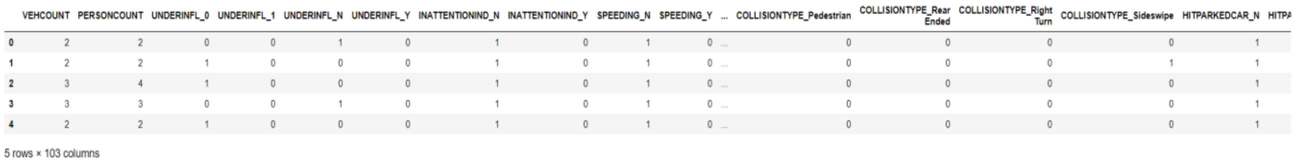
The data set had redundant columns like a column which gives a code to the severity of accident and another column explains the severity of the accident in words.

Missing data was filled looking into the handling of unknown situation in each feature. The data set with all the values need to be made understandable by the machine.

To do that we performed One-hot encoding on the various features of the dataset. The features were

'UNDERINFL','INATTENTIONIND','SPEEDING','LIGHTCOND','ROADCOND','WEATHER','HITPARKEDCAR','SDOT_COLDESC','JUNCTIONTYPE','COLLISIONTYPE','HITPARKEDCAR','ADDRTYPE'

This type of encoding increases the dimension of the dataset quite considerably. In our case the dimension of the data changed from “5 rows × 38 columns” to “5 rows × 103 columns”.



	VEHCOUNT	PERSONCOUNT	UNDERINFL_0	UNDERINFL_1	UNDERINFL_N	UNDERINFL_Y	INATTENTIONIND_N	INATTENTIONIND_Y	SPEEDING_N	SPEEDING_Y	...	COLLISIONTYPE_Pedestrian	COLLISIONTYPE_Rear Ended	COLLISIONTYPE_Right Turn	COLLISIONTYPE_Sideswipe	HITPARKEDCAR_N	HITPA
0	2	2	0	0	1	0	1	0	1	0	...	0	0	0	0	1	
1	2	2	1	0	0	0	1	0	1	0	...	0	0	0	1	1	
2	3	4	1	0	0	0	1	0	1	0	...	0	0	0	0	1	
3	3	3	0	0	1	0	1	0	1	0	...	0	0	0	0	1	
4	2	2	1	0	0	0	1	0	1	0	...	0	0	0	0	1	

5 rows × 103 columns

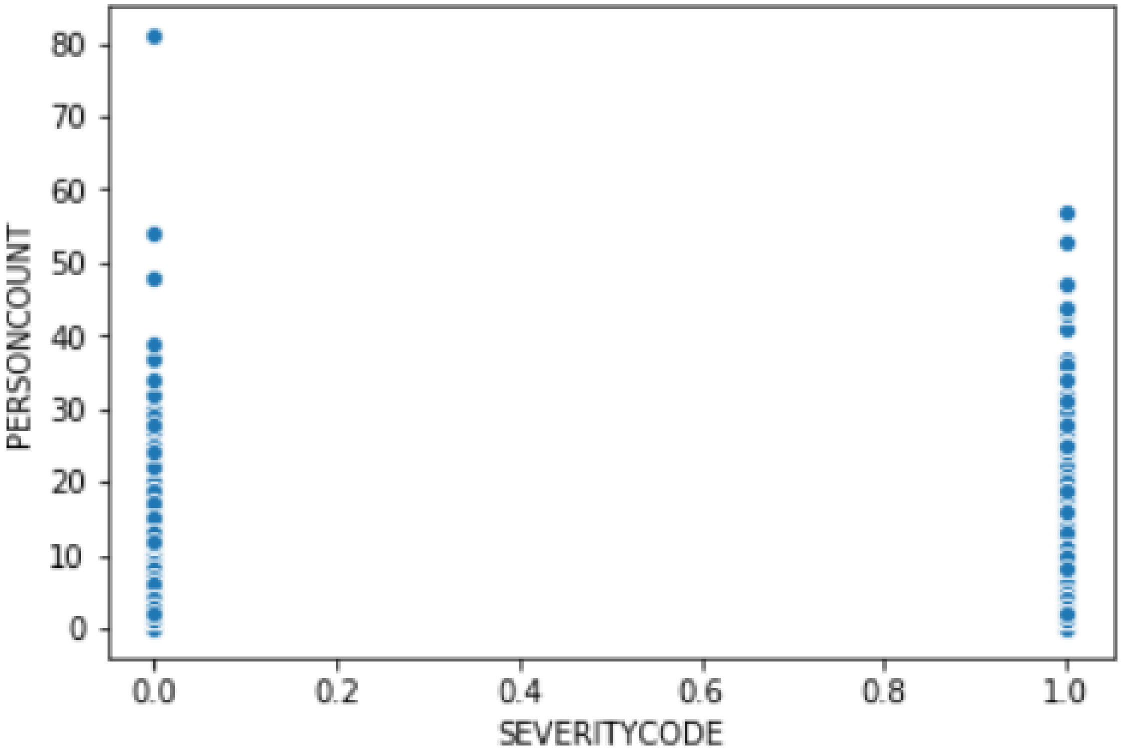
To identify the severity of the collision we are considering the severity code as the reference class. We treated the severity label 1 as ‘HIGH’ severity and label 0 as ‘LOW’ severity.

Data Analysis:

Usually a road accident many factors play the key role.

Road conditions like “wet” road increases breaking distance, which results in collisions in case the break needs to be applied suddenly.

Light conditions play its own role in the collisions occurring in the road. When vehicles are moving on the road the due to light conditions perceiving the speed of the other vehicles in traffic becomes difficult and hence resulting in collisions.



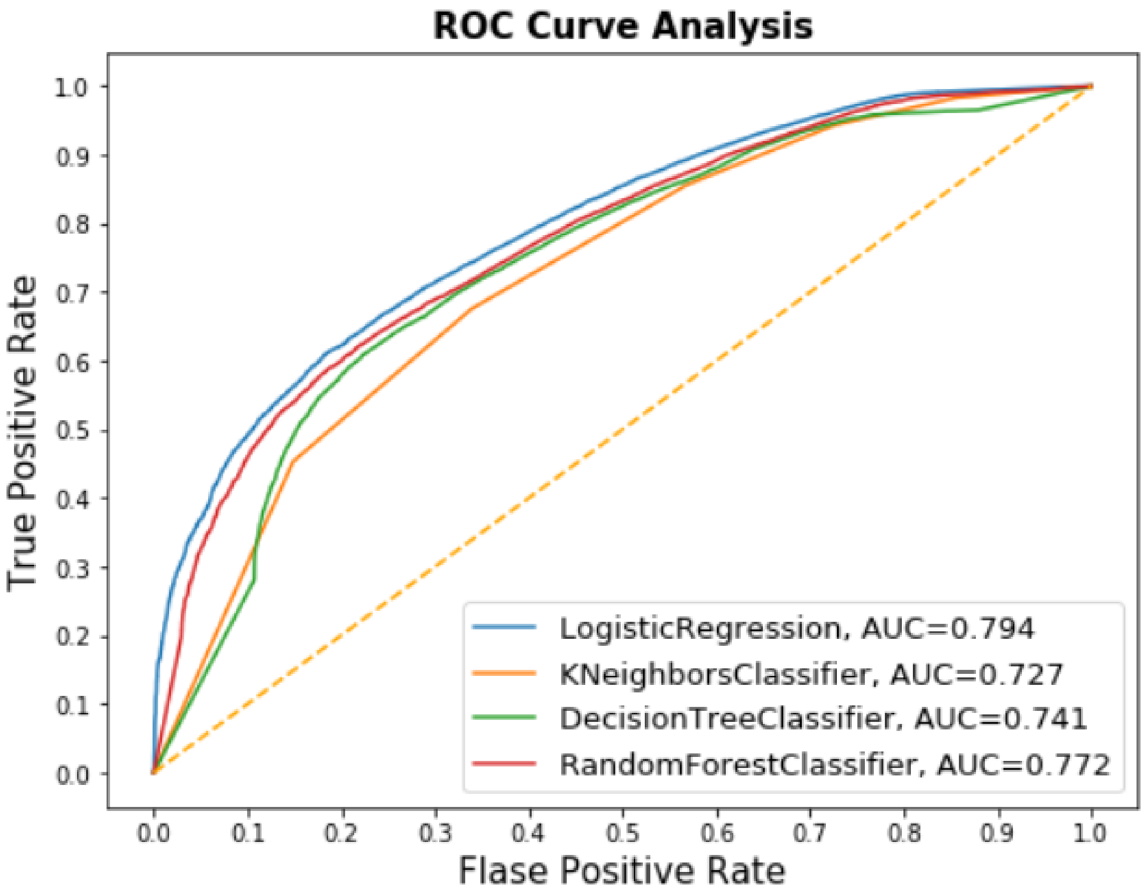
The number of persons involved in the collisions increase the chances of number of people suffering serious injuries or loss of lives. This is also clear from the plot. The increase in the number of people increase the probability of ‘HIGH’ severity of accident.

Model Training:

We have used four algorithms to train model.

1. Decision Model
2. Logistic Regression
3. KNeighbors
4. Random Forest

We trained the model and also calculated the accuracy. As accuracy does not give us a lot insight into the quality of the trained model. We thus also plotted ROC curve. We compared the ROC curve of all the trained model.



Looking at this it is clear that no model is a clear winner in terms of performance. Still the KNeighbors outperform other model a slight margin.