Mulpuru lokesh **Augus2020**

Business understanding

Background

Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.

Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic product. More than half of all road traffic deaths are among vulnerable road users: pedestrians, cyclists, and motorcyclists. Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years.

Source: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries#:~:text=Approximately%201.35%20million%20people%20die,of%20their%20gross%20domestic%20product.

Problem

The dataset from SDOT available to us, tells us the following details:

Details of accident

1. Severity of collisions: Tells us about the extent of damage - property damage, injury or fatality
2. Collision type - We understand whether the collision was head on, whether pedestrians or cyclists were involved and similar data
3. We also also analyse time of accident date and time: This reveals whether more accidents occur on weekdays or weekends and whether accidents occur more at night

Affected

1. No. of persons involved - Reveals whether only one person hit a non mobilized object or more people were involved
2. No. of cyclists involved in accidents were cyclists were involved
3. No. of pedestrians affected by different accidents
4. No. of vehicles involved in accidents
5. If an accident involved pedestrians, whether they were granted their way
6. No. of accidents where parked cars were hit

Location factors

1. Type of address: Whether more accidents occur in alleys, blocks or intersections
2. In which junction types more accidents occur

Human factors

1. He/she was attentive/ unattentive
2. If the person was under influence
3. If the person was speeding

Environmental factors

1. If rainy days cause more accidents or sunny days
2. Whether dry or wet roads cause more accidents
3. If lighting conditions are a factor in accidents

The aim is to understand the causes of road accidents by analysing the parameters outlined above, namely:

1. Location factors
2. Human factors
3. Environmental factors

Eventually we build a machine learning model to classify road accidents and predict injury collisions.

Client

Road traffic crashes cost most countries 3% of their gross domestic product. Governments would be interested to understand the reasons behind road accidents. The aim of this project is to equip them with data driven insights to enable decision making to reduce the number of accidents.

Data understanding

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 38 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   SEVERITYCODE    194673 non-null  int64  
 1   X               189339 non-null  float64
 2   Y               189339 non-null  float64
 3   OBJECTID        194673 non-null  int64  
 4   INCKEY          194673 non-null  int64  
 5   COLDETKEY       194673 non-null  int64  
 6   REPORTNO        194673 non-null  object 
 7   STATUS          194673 non-null  object 
 8   ADDRTYPE        192747 non-null  object 
 9   INTKEY          65070 non-null   float64
 10  LOCATION        191996 non-null  object 
 11  EXCEPTRSNCODE   84811 non-null   object 
 12  EXCEPTRSNDESC   5638 non-null    object 
 13  SEVERITYCODE.1  194673 non-null  int64  
 14  SEVERITYDESC    194673 non-null  object 
 15  COLLISIONTYPE   189769 non-null  object 
 16  PERSONCOUNT     194673 non-null  int64  
 17  PEDCOUNT        194673 non-null  int64  
 18  PEDCYLCOUNT     194673 non-null  int64  
 19  VEHCOUNT        194673 non-null  int64  
 20  INCDATE         194673 non-null  object 
 21  INCDTTM         194673 non-null  object 
 22  JUNCTIONTYPE    188344 non-null  object 
 23  SDOT_COLCODE    194673 non-null  int64  
 24  SDOT_COLDESC    194673 non-null  object 
 25  INATTENTIONIND  29805 non-null   object 
 26  UNDERINFL       189789 non-null  object 
 27  WEATHER         189592 non-null  object 
 28  ROADCOND        189661 non-null  object 
 29  LIGHTCOND       189503 non-null  object 
 30  PEDROWNOTGRNT   4667 non-null    object 
 31  SDOTCOLNUM      114936 non-null  float64
 32  SPEEDING        9333 non-null    object 
 33  ST_COLCODE      194655 non-null  object 
 34  ST_COLDESC      189769 non-null  object 
 35  SEGLANEKEY      194673 non-null  int64  
 36  CROSSWALKKEY    194673 non-null  int64  
 37  HITPARKEDCAR    194673 non-null  object 
dtypes: float64(4), int64(12), object(22)
memory usage: 56.4+ MB
```

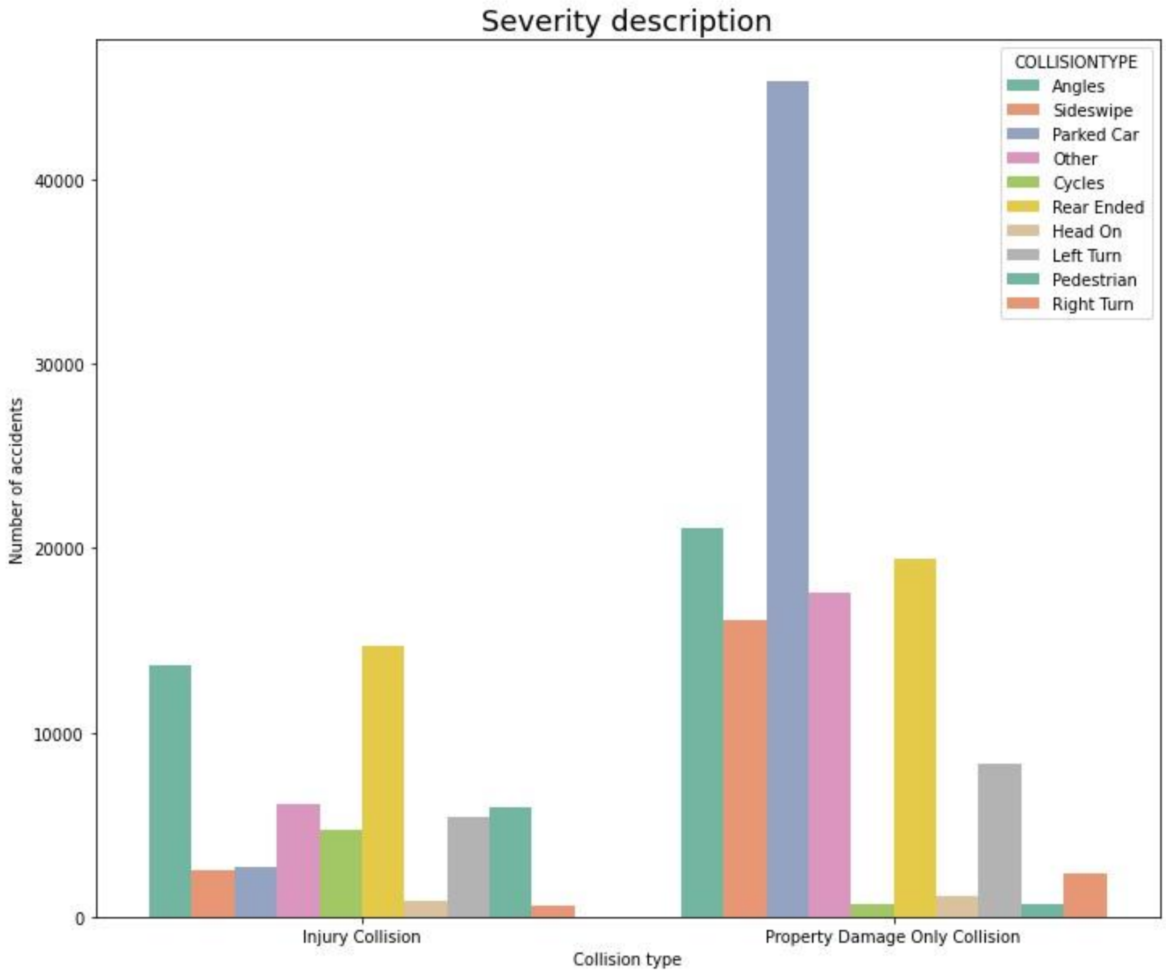
There are a lot of missing values in the dataset

	Attribute	Count of missing values
30	PEDROWNOTGRNT	190006
12	EXCEPTRSNDESC	189035
32	SPEEDING	185340
25	INATTENTIONIND	164868
9	INTKEY	129603
11	EXCEPTRSNCODE	109862
31	SDOTCOLNUM	79737
22	JUNCTIONTYPE	6329
2	Y	5334
1	X	5334
29	LIGHTCOND	5170
27	WEATHER	5081
28	ROADCOND	5012
15	COLLISIONTYPE	4904
34	ST_COLDESC	4904
26	UNDERINFL	4884
10	LOCATION	2677
8	ADDRTYPE	1926
33	ST_COLCODE	18

Data preparation

We clean the data and normalize it to achieve a feature rich dataset. We fill in the missing data and drop unnecessary columns.

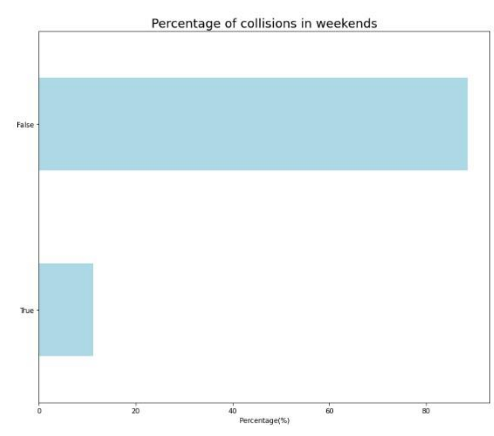
Exploratory data analysis



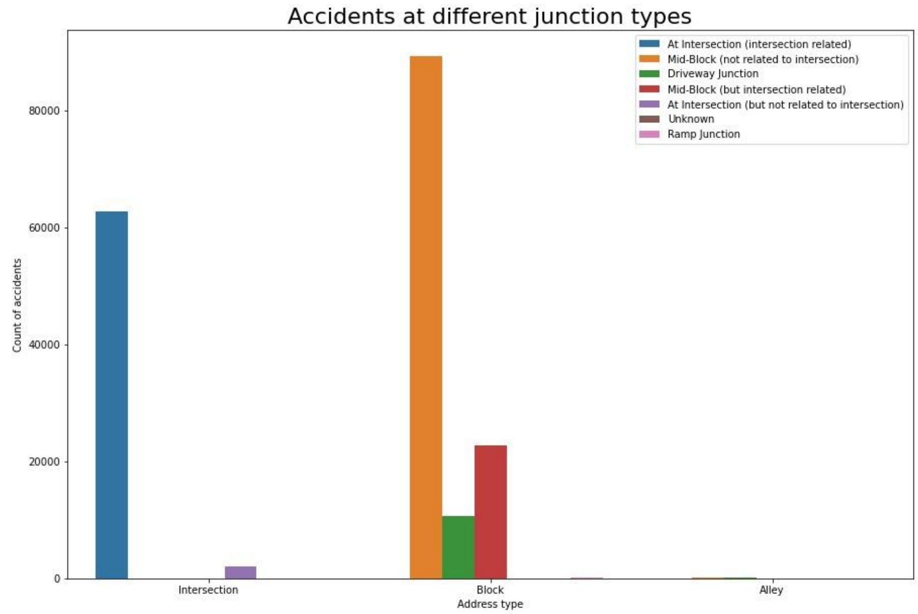
Observations:

1. More accidents occur on weekdays
2. In injury collision, major accidents occur due to vehicles hitting another vehicle's rear end or hitting pedestrians
3. In property damage collisions, mostly parked cars are hit

As is evident from the graph given alongside, more accidents occur on weekdays on weekends.



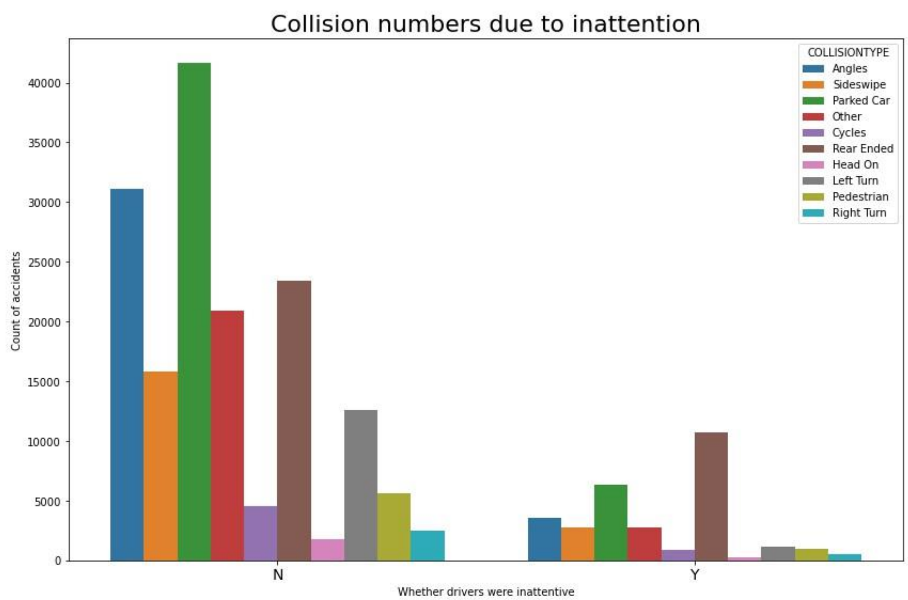
Analyzing location factors



Observations:

1. Some accidents unrelated to intersections occur at intersections.
2. In blocks, maximum accidents occur at midblock and away from intersections.
3. Very few accidents occur at alleys.

Analysing human factors



When drivers are inattentive, in maximum accidents they drive into rear end of moving vehicles. Whe n inattentive, they generally hit parked cars.

Interesting findings on collision outcomes:

	SEVERITYCODE	Intersection	Alley	Block
SEVERITYCODE	1.000	0.199	-0.026	-0.185
Intersection	0.199	1.000	-0.044	-0.970
Alley	-0.026	-0.044	1.000	-0.085
Block	-0.185	-0.970	-0.085	1.000

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
SEVERITYCODE	1.000	0.131	0.246	0.214	-0.055
PERSONCOUNT	0.131	1.000	-0.023	-0.039	0.381
PEDCOUNT	0.246	-0.023	1.000	-0.017	-0.261
PEDCYLCOUNT	0.214	-0.039	-0.017	1.000	-0.254
VEHCOUNT	-0.055	0.381	-0.261	-0.254	1.000

Maximum injury collisions occur at intersections involving pedestrians and cyclists.

Modelling

Analysis of attribute's characteristics which lead to collision

The problem is a binary classification problem on SEVERITYCODE:

```
SEVERITYCODE of collision:
       0 - Property collision
       1 - Injury collision
```

Feature selection and splitting into train and test sets

Correlation of selected features:

Attribute	Correlation with severity of collision
PEDCOUNT	0.246
PEDCYLCOUNT	0.214
PEDROWNOTGRNT	0.206
Intersection	0.199
Cycles	0.213
Pedestrian	0.245
At Intersection (intersection related)	0.202
PERSONCOUNT	0.131
SDOT_COLCODE	0.189

Different machine learning models were tried on this classification problem:

Results

Algorithm	Accuracy score	F1 score	Precision score	Recall score
Logistic Regression	74.99	36.48	76.67	23.94
Decision Tree	75.24	37.41	77.43	24.67
Random Forest	75.21	37.31	77.27	24.59
SVM	74.84	30.62	88.68	18.51
XG Boost	75.26	35.56	81.35	22.75

Discussion

During the data exploration process, I came across some interesting observations:

```
1. Maximum accidents occur during weekdays at intersections
2. Weather conditions do not play a significant role in accidents
3. Road and lighting conditions have a weak correlation with accidents
4. Being under influence doesn't cause noticeably more accidents than being i
nattentive
5. Between blocks, maximum accidents occur at mid-blocks
6. In collision accidents, maximum damage is done to parked cars
```

In this project I have identified the relation between accidents and several human, environmental and location attributes. Maximum accidents occur at intersections related to pedestrians or cyclists. I analysed different machine learning models to classify accidents as injury or collision accidents. The "XGBoost" model offered maximum accuracy. It correctly predicted 81.35% as injury collisions. This data could be used by governments to establish separate signals for allowing pedestrians and cyclists to cross at intersections. It is also aimed at us, whether we are pedestrians, cyclists or vehicle owners to be more careful at intersections to prevent an accident.

Conclusion

I could achieve an accuracy of ~75% using the XGB Classifier. There are a lot of variances which have not been accounted for. However, using this project we could really narrow down to the location(intersections) where maximum accidents occur and the most affected. We also understood that there is very less importance of human and weather factors in causing an accident. The prediction could be improved by capturing real time data during accidents.