IBM Data science Capstone project

Predicting Seattle car accident severity

Renuka Devi Ulaganathan

October 3, 2020

Introduction:

With the prevailing traffic conditions around the globe, there is a need for an alert system that would predict the possiblity of an accident based on given conditions. Using machine learning techniques, a model can be built and trained from the past collision data available. The model thus trained shall be used to predict the occurence of a collision and its severity and alert the parties involved.

1. Business understanding:

The Seattle government, in an effort to reduce the number of car collisions, wants to reduce the car accidents in Seattle. To implement this, a model to be developed to predict the possiblity of car accidents given factors like accident spot, driving speed,weather, road, light conditions etc. This model will help in warning the local Seattle government, traffic police and the drivers on the targeted roads that will help in reducing the requency of accidents.

The target audience of this project is Seattle government, traffic police department,city traffic surveillance team, car drivers and the local residents in the neighbourhood.

2. Data understanding:

2.1 Source of the data:</u>

The collision data used for this project is taken from Seattle Department of Transport (SDOT) provided in capstone project introduction. The link for the data is here (https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). The metadata about the dataset can be found here (https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf).

The data is be downloaded from the portal and saved into a python dataframe for futher analytics.

```
--2020-10-11 06:24:14--  https://s3.us.cloud-object-storage.appdomain.cloud/cf
-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
Resolving s3.us.cloud-object-storage.appdomain.cloud (s3.us.cloud-object-stora
ge.appdomain.cloud)... 67.228.254.196
Connecting to s3.us.cloud-object-storage.appdomain.cloud (s3.us.cloud-object-s
torage.appdomain.cloud)|67.228.254.196|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 73917638 (70M) [text/csv]
Saving to: ‘collision_data.csv’

100%[=====================================>] 73,917,638  42.9MB/s   in 1.6s

2020-10-11 06:24:16 (42.9 MB/s) - ‘collision_data.csv’ saved [73917638/7391763
8]

Dataset downloaded successfully

/opt/conda/envs/Python36/lib/python3.6/site-packages/IPython/core/interactives
hell.py:3020: DtypeWarning: Columns (33) have mixed types. Specify dtype optio
n on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STAT
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matc
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matc
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matc
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matc
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matc

5 rows × 38 columns

Taking a look at the collision dataset, we have **194,673 rows** and **38 columns.**

```
The number of columns is :  38
The number of (rows,columns) is :  (194673, 38)
The size of the dataset is : 7397574
The data types of different columns are below :


SEVERITYCODE        int64
X                 float64
Y                 float64
OBJECTID            int64
INCKEY              int64
COLDETKEY           int64
REPORTNO           object
STATUS             object
ADDRTYPE           object
INTKEY            float64
LOCATION           object
EXCEPTRSNCODE      object
EXCEPTRSNDESC      object
SEVERITYCODE.1      int64
SEVERITYDESC       object
COLLISIONTYPE      object
PERSONCOUNT         int64
PEDCOUNT            int64
PEDCYLCOUNT         int64
VEHCOUNT            int64
INCDATE            object
INCDTTM            object
JUNCTIONTYPE       object
SDOT_COLCODE        int64
SDOT_COLDESC       object
INATTENTIONIND     object
UNDERINFL          object
WEATHER            object
ROADCOND           object
LIGHTCOND          object
PEDROWNOTGRNT      object
SDOTCOLNUM        float64
SPEEDING           object
ST_COLCODE         object
ST_COLDESC         object
SEGLANEKEY          int64
CROSSWALKKEY        int64
HITPARKEDCAR       object
dtype: object

Taking a back up of the dataframe for any reference down the line, as we will
now proceed with data cleaning
```

The target or dependent variable is **'SEVERITYCODE'** which determines the severity of the accident. Below are the possible values for this variable. However, the dataset provided in the capstone project has only **'1'** and **'2' severity codes**. so, our maachine learning technique will apply on the provided data source.

Severity codes:

0: Little to no Probability (Clear Conditions)

1: Very Low Probability — Chance or Property Damage

2: Low Probability — Chance of Injury

3: Mild Probability — Chance of Serious Injury

4: High Probability — Chance of Fatality

Out of the 38 columns present in the dataset, it is necessary to determine the predictor or independent variables that would influence the severity of the accident. Observing the data, the variables **'ADDRTYPE','WEATHER','ROADCOND','LIGHTCOND','UNDERINFL', 'COLLISIONTYPE','JUNCTIONTYPE' and 'SPEEDING'** seem to more likely to impact the accident severity. Let us plot these variables against severity code and observe relationship between them.

Identify the unique values for each of these variables:

```
The frequency of address type is:

Block           126926
Intersection     65070
Alley              751
Name: ADDRTYPE, dtype: int64

The frequency of weather conditions is:

Clear                       111135
Raining                      33145
Overcast                     27714
Unknown                      15091
Other                         5913
Snowing                        907
Fog/Smog/Smoke                 569
Sleet/Hail/Freezing Rain       113
Blowing Sand/Dirt               56
Severe Crosswind                25
Partly Cloudy                    5
Name: WEATHER, dtype: int64

The frequency of road conditions is:


Dry               124510
Wet                47474
Unknown            15078
Other               5144
Ice                 1209
Snow/Slush          1004
Standing Water       115
Sand/Mud/Dirt         75
Oil                   64
Name: ROADCOND, dtype: int64
```

```
The frequency of light conditions is:


Daylight                    116137
Dark - Street Lights On      48507
Unknown                      13473
Dusk                          5902
Other                         5405
Dawn                          2502
Dark - No Street Lights       1537
Dark - Street Lights Off      1199
Dark - Unknown Lighting         11
Name: LIGHTCOND, dtype: int64

The distinct 'under influence' values and its respective counts are:

N    105158
0     80394
Y      5126
1      3995
Name: UNDERINFL, dtype: int64

The distinct speeding values and its respective counts are:

N    185340
Y      9333
Name: SPEEDING, dtype: int64

The frequency of collision types is:

Parked Car    47987
Angles        34674
Rear Ended    34090
Other         28607
Sideswipe     18609
Left Turn     13703
Pedestrian     6608
Cycles         5415
Right Turn     2956
Head On        2024
Name: COLLISIONTYPE, dtype: int64

The frequency of junction types is:

Mid-Block (not related to intersection)              89800
At Intersection (intersection related)               62810
Mid-Block (but intersection related)                 22790
Driveway Junction                                    10671
Other                                                 6329
At Intersection (but not related to intersection)     2098
Ramp Junction                                          166
Unknown                                                  9
Name: JUNCTIONTYPE, dtype: int64
```

2.2 Data Cleaning:

```
The total number of null/NaN values for each variable is as follows:

 SEVERITYCODE            0
X                     5334
Y                     5334
OBJECTID                 0
INCKEY                   0
COLDETKEY                0
REPORTNO                 0
STATUS                   0
ADDRTYPE              1926
INTKEY              129603
LOCATION              2677
EXCEPTRSNCODE       109862
EXCEPTRSNDESC       189035
SEVERITYCODE.1           0
SEVERITYDESC             0
COLLISIONTYPE         4904
PERSONCOUNT              0
PEDCOUNT                 0
PEDCYLCOUNT              0
VEHCOUNT                 0
INCDATE                  0
INCDTTM                  0
JUNCTIONTYPE          6329
SDOT_COLCODE             0
SDOT_COLDESC             0
INATTENTIONIND      164868
UNDERINFL             4884
WEATHER               5081
ROADCOND              5012
LIGHTCOND             5170
PEDROWNOTGRNT       190006
SDOTCOLNUM           79737
SPEEDING            185340
ST_COLCODE              18
ST_COLDESC            4904
SEGLANEKEY               0
CROSSWALKKEY             0
HITPARKEDCAR             0
dtype: int64

Let us replace the null or NaN values of SPEEDING to N, null, NaN and 0 values
of UNDERINFL to N and all other null values of predictor variables to 'other'
```

Now, Let us check if there are still any null values for these independent variables.

```
SEVERITYCODE            0
X                    5334
Y                    5334
OBJECTID                0
INCKEY                  0
COLDETKEY               0
REPORTNO                0
STATUS                  0
ADDRTYPE                0
INTKEY             129603
LOCATION             2677
EXCEPTRSNCODE      109862
EXCEPTRSNDESC      189035
SEVERITYCODE.1          0
SEVERITYDESC            0
COLLISIONTYPE           0
PERSONCOUNT             0
PEDCOUNT                0
PEDCYLCOUNT             0
VEHCOUNT                0
INCDATE                 0
INCDTTM                 0
JUNCTIONTYPE            0
SDOT_COLCODE            0
SDOT_COLDESC            0
INATTENTIONIND     164868
UNDERINFL               0
WEATHER                 0
ROADCOND                0
LIGHTCOND               0
PEDROWNOTGRNT      190006
SDOTCOLNUM          79737
SPEEDING                0
ST_COLCODE             18
ST_COLDESC           4904
SEGLANEKEY              0
CROSSWALKKEY            0
HITPARKEDCAR            0
dtype: int64
```

Good! We have got rid off Nan and NULL values for the independent variables.

2.3 Understanding of Target variable:

```
The severity codes by count is as follows:
 1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5363025b70>
```

2.3.1 Downsampling and Balancing the dataset:

As we see, the severity code '1'(property damage) is more than doubles the time of severity code '2'(Injury). Hence the dataset is highly unbalanced. It will be difficult to predict or apply machine learning algorithms on unbalanced dataset. Hence, I shall downsample the dataset to balance it.

Let us take these variables to understand what kind of impact it has on SEVERITYCODE.

```
The dataset is now balanced and can be used for machine learning. Severity cod
e frequency is below.

2    58188
1    58188
Name: SEVERITYCODE, dtype: int64
```

2.3.2 Feature selection:

As part of feature selection, let us drop the rest of the fields that are not part of independent variable list.

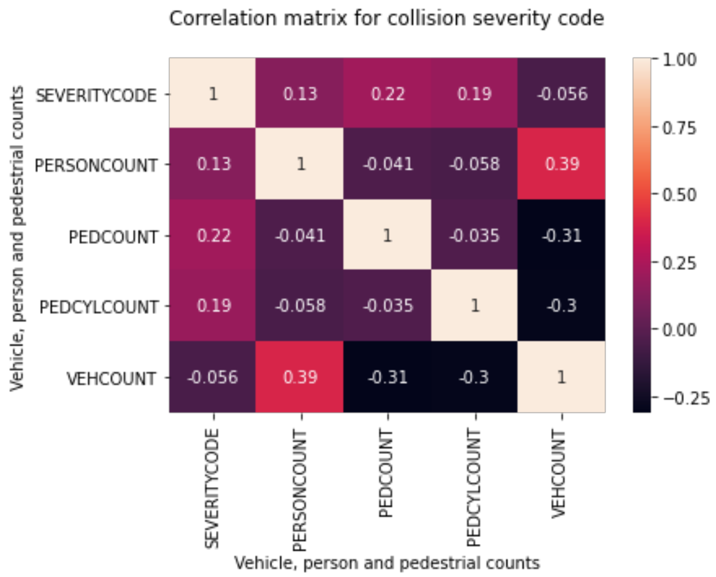
```
The variables part of dataframe are listed below:

SEVERITYCODE      int64
ADDRTYPE         object
COLLISIONTYPE    object
PERSONCOUNT       int64
PEDCOUNT          int64
PEDCYLCOUNT       int64
VEHCOUNT          int64
JUNCTIONTYPE     object
UNDERINFL        object
WEATHER          object
ROADCOND         object
LIGHTCOND        object
SPEEDING         object
dtype: object
```

As a next step, want to understand if the variables **PERSONCOUNT,PEDCOUNT,PEDCYLCOUNT and VEHCOUNT** has impact on the severity code before doing **dimension reduction** . Let us draw the **correlation matrix** with these numerical data to analyze further.

2.3.3 Correlation Matrix of numeric variables:

```
Text(32.99999999999999, 0.5, 'Vehicle, person and pedestrial counts')
```



From the correlation matrix above, we can see that the correlation coefficient is less than 0.5 for all the variables mapped and hence none of them have stronger relation between them. Comapring the variables against SEVERITYCODE, the variables **PERSONCOUNT,PEDCOUNT,PEDCYLCOUNT and VEHCOUNT** have **weaker relationship** with **severity code.** Hence these can be dropped from the dataframe.

```
Dropping PERSONCOUNT,PEDCOUNT,PEDCYLCOUNT,VEHCOUNT from the Data Frame
```

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	JUNCTIONTYPE	UNDERINFL	WEATHER
25055	1	Intersection	Angles	At Intersection (intersection related)	0	Raining
65280	1	Intersection	Angles	At Intersection (intersection related)	0	Clear
86292	1	Intersection	Angles	At Intersection (intersection related)	N	Unknown
155111	1	Block	Sideswipe	Mid-Block (not related to intersection)	N	Clear
64598	1	Block	Head On	Mid-Block (not related to intersection)	0	Clear

```
Data Frame after drpping the pereson and vehicle counts:

SEVERITYCODE       int64
ADDRTYPE          object
COLLISIONTYPE     object
JUNCTIONTYPE      object
UNDERINFL         object
WEATHER           object
ROADCOND          object
LIGHTCOND         object
SPEEDING          object
dtype: object
```

3. Exploratory data analysis:

The variable **UNDERINFL** currently has values **'0','N','Y' and '1'**. To keep it unform for machine learning, let us map 'N' to '0'and Y to '1'.

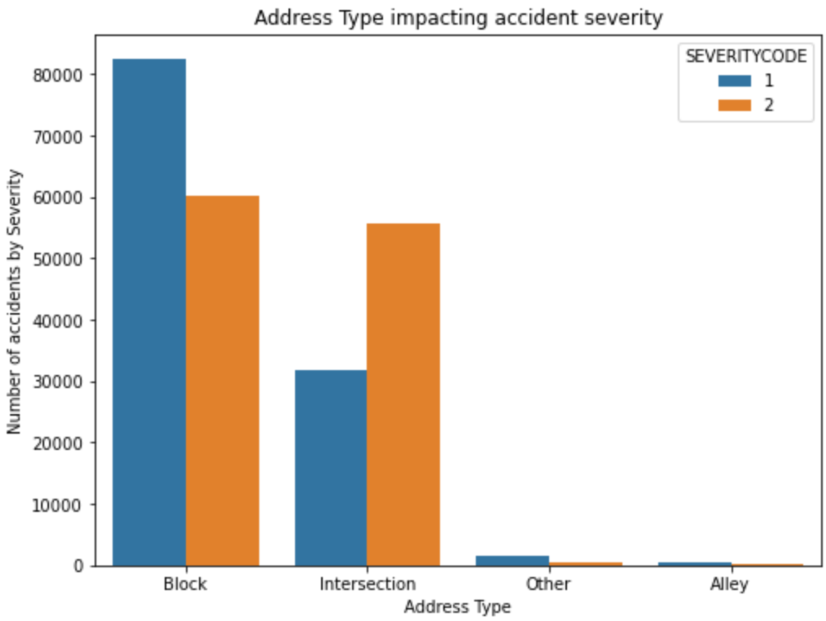
```
The values variable UNDERINFL is mapped to values 0 and 1
```

3.1 Relationship between ADDRTYPE and SEVERITYCODE:

	ADDRTYPE	SEVERITYCODE	ADDRCOUNT
2	Block	1	82450
3	Block	2	60192
5	Intersection	2	55638
4	Intersection	1	31862
6	Other	1	1462
0	Alley	1	602
7	Other	2	382
1	Alley	2	164

3.1.1 ADDRCOUNT vs SEVERITYCODE - Data Visualization

```
Text(0, 0.5, 'Number of accidents by Severity')
```



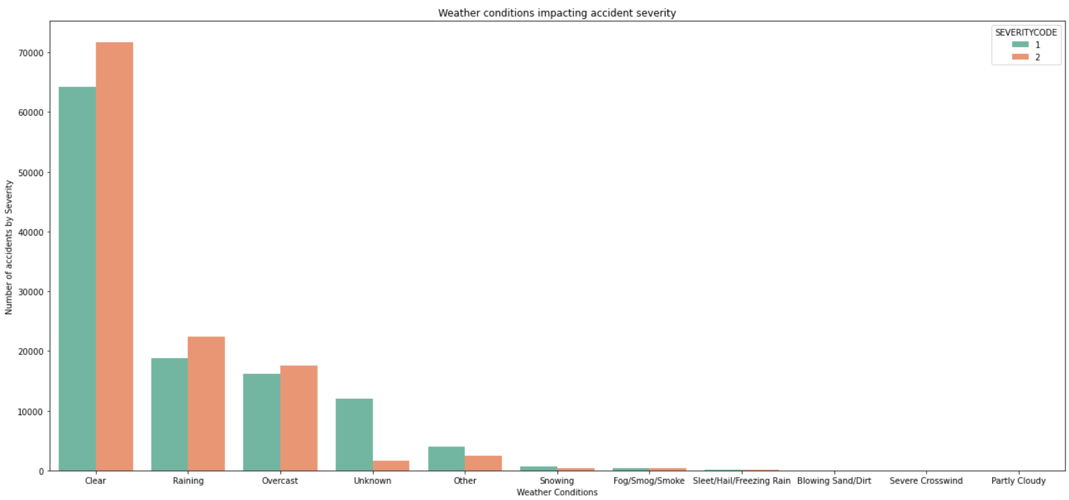
From the above, we can see that **ADDRCOUNT** has strong relationship with SEVERITYCODE influencing the accident severity. Hence, it should be considered as part of the independent variables to predict SEVERITYCODE.

3.2 Relationship between WEATHER and SEVERITYCODE:

	WEATHER	SEVERITYCODE	WEATHCOUNT
3	Clear	2	71680
2	Clear	1	64212
13	Raining	2	22352
12	Raining	1	18816
9	Overcast	2	17490
8	Overcast	1	16178
20	Unknown	1	12070
6	Other	1	4044
7	Other	2	2400
21	Unknown	2	1632
18	Snowing	1	606
5	Fog/Smog/Smoke	2	374
19	Snowing	2	342
4	Fog/Smog/Smoke	1	336
16	Sleet/Hail/Freezing Rain	1	70
17	Sleet/Hail/Freezing Rain	2	56
0	Blowing Sand/Dirt	1	30
1	Blowing Sand/Dirt	2	30
15	Severe Crosswind	2	14
14	Severe Crosswind	1	12
11	Partly Cloudy	2	6
10	Partly Cloudy	1	2

3.2.1 WEATHER vs SEVERITYCODE - Data Visualization

```
Text(0, 0.5, 'Number of accidents by Severity')
```



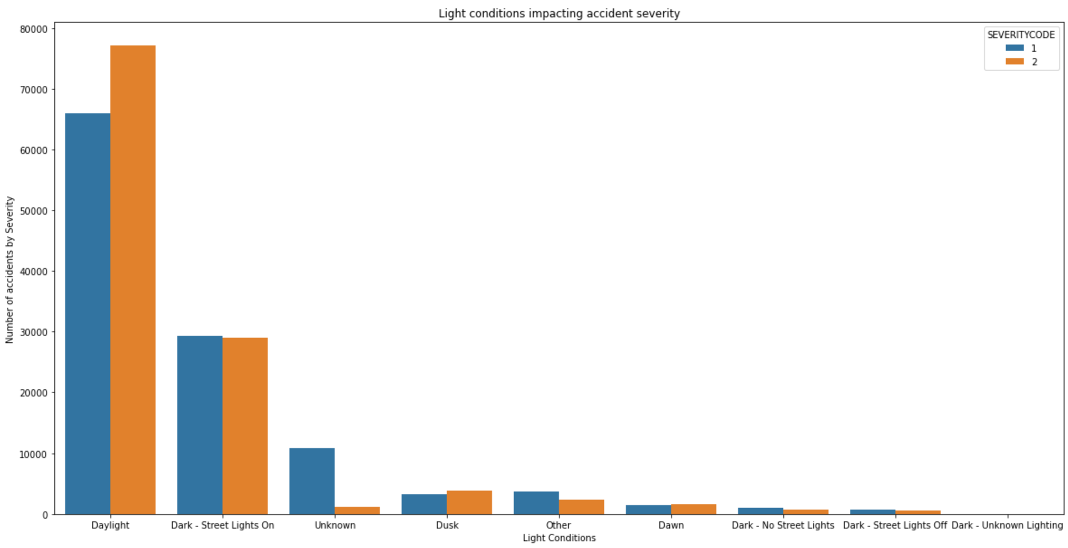
From the above, we can see that **WEATHER** has strong relationship with SEVERITYCODE influencing the accident severity. Hence, it should be considered as part of the independent variables to predict accident severity.

3.3 Relationship between LIGHTCOND and SEVERITYCODE:

	LIGHTCOND	SEVERITYCODE	LIGHTCOUNT
11	Daylight	2	77088
10	Daylight	1	65918
4	Dark - Street Lights On	1	29316
5	Dark - Street Lights On	2	28950
16	Unknown	1	10914
13	Dusk	2	3888
14	Other	1	3666
12	Dusk	1	3296
15	Other	2	2284
9	Dawn	2	1648
8	Dawn	1	1422
17	Unknown	2	1210
0	Dark - No Street Lights	1	1064
2	Dark - Street Lights Off	1	774
1	Dark - No Street Lights	2	668
3	Dark - Street Lights Off	2	632
7	Dark - Unknown Lighting	2	8
6	Dark - Unknown Lighting	1	6

3.3.1 LIGHTCOND vs SEVERITYCODE - Data Visualization

```
Text(0.5, 1.0, 'Light conditions impacting accident severity')
```



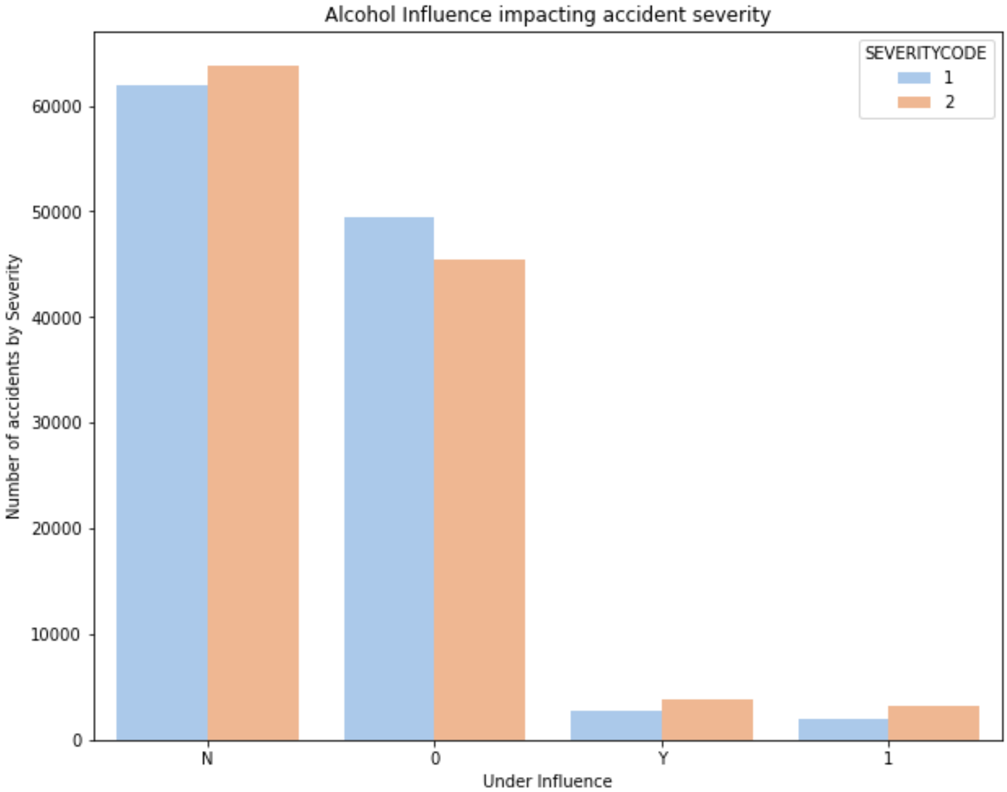
From the above, we can see that **LIGHTCOND** has strong relationship with SEVERITYCODE influencing the accident severity. The accident has resulted in injury mainly during daylight and dark when street lights are on. Hence, it should be considered as part of the independent variables to predict accident severity.

3.4 Relationship between UNDERINFL and SEVERITYCODE:

	UNDERINFL	SEVERITYCODE	INFLCOUNT
5	N	2	63850
4	N	1	62000
0	0	1	49524
1	0	2	45402
7	Y	2	3878
3	1	2	3246
6	Y	1	2802
2	1	1	2050

3.4.1 UNDRINFL vs SEVERITYCODE - Data Visualization

```
Text(0.5, 1.0, 'Alcohol Influence impacting accident severity')
```



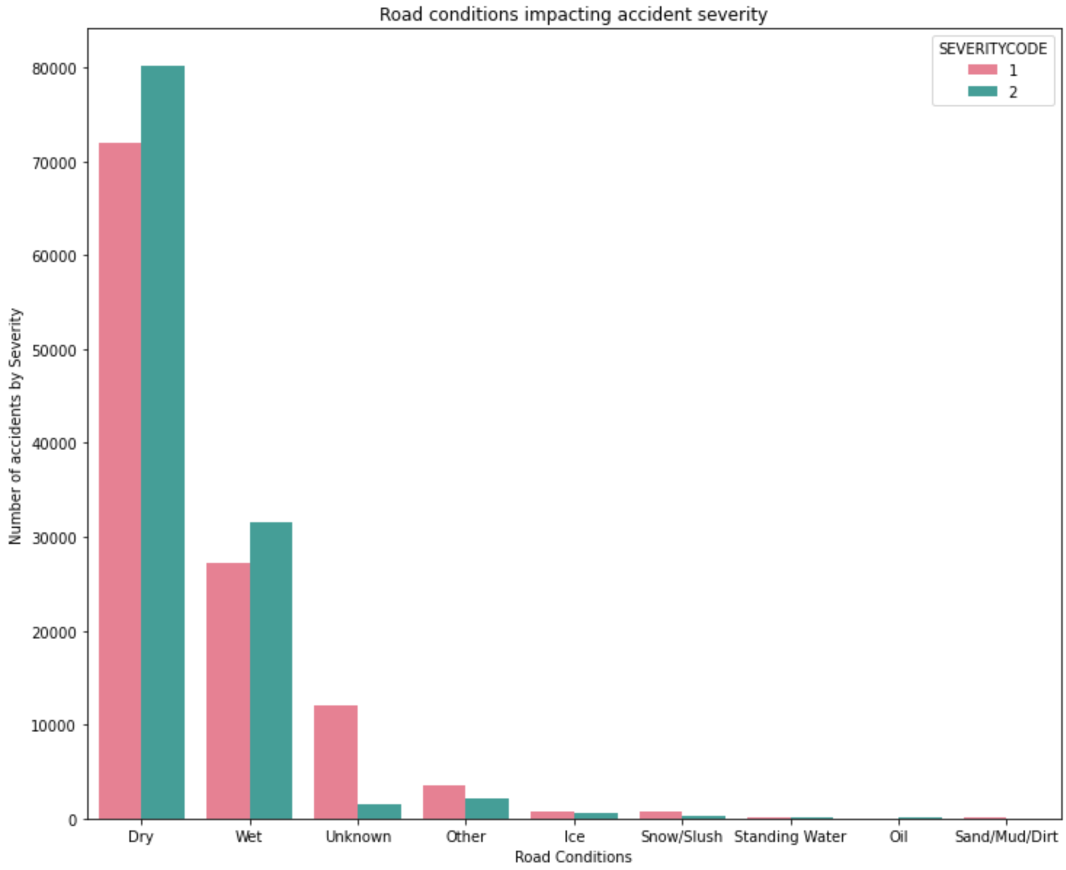
From the above, we can see that **UNDERINFL** has weaker relationship with SEVERITYCODE in influencing the accident severity. We can see that the accident has resulted in property damage or injury when the driver is not under alcohol influence. Hence, it should **not be considered** as part of the independent variables to predict accident severity.

3.5 Relationship between ROADCOND and SEVERITYCODE:

	ROADCOND	SEVERITYCODE	ROADCOUNT
1	Dry	2	80128
0	Dry	1	71872
17	Wet	2	31510
16	Wet	1	27238
14	Unknown	1	12114
6	Other	1	3488
7	Other	2	2206
15	Unknown	2	1498
2	Ice	1	810
10	Snow/Slush	1	722
3	Ice	2	546
11	Snow/Slush	2	334
13	Standing Water	2	60
12	Standing Water	1	54
5	Oil	2	48
8	Sand/Mud/Dirt	1	48
9	Sand/Mud/Dirt	2	46
4	Oil	1	30

3.5.1 ROADCOND vs SEVERITYCODE - Data Visualization

```
Text(0.5, 1.0, 'Road conditions impacting accident severity')
```



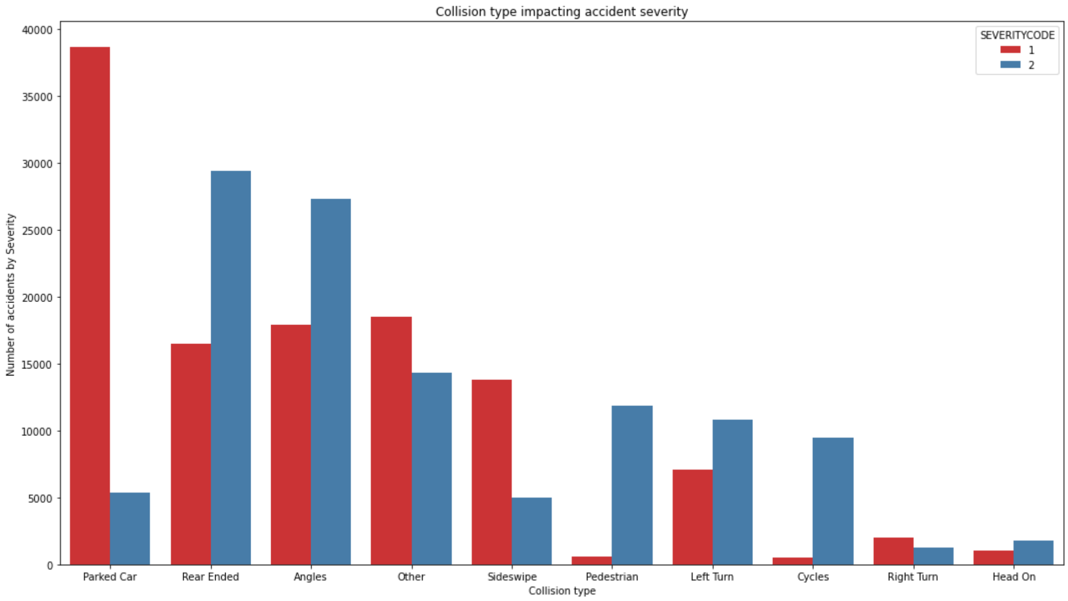
From the above, we can see that **ROADCOND** has strong relationship with SEVERITYCODE influencing the accident severity. The accident has resulted in injury mainly when the road was dry and wet. Hence, it should be considered as part of the independent variables to predict accident severity.

3.6 Relationship between COLLISIONTYPE and SEVERITYCODE:

	COLLISIONTYPE	SEVERITYCODE	COLLNCOUNT
10	Parked Car	1	38614
15	Rear Ended	2	29342
1	Angles	2	27248
8	Other	1	18450
0	Angles	1	17868
14	Rear Ended	1	16466
9	Other	2	14306
18	Sideswipe	1	13794
13	Pedestrian	2	11872
7	Left Turn	2	10822
3	Cycles	2	9488
6	Left Turn	1	7052
11	Parked Car	2	5324
19	Sideswipe	2	5012
16	Right Turn	1	1998
5	Head On	2	1744
17	Right Turn	2	1218
4	Head On	1	1018
12	Pedestrian	1	586
2	Cycles	1	530

3.6.1 COLLISIONTYPE vs SEVERITYCODE - Data Visualization

```
Text(0.5, 1.0, 'Collision type impacting accident severity')
```



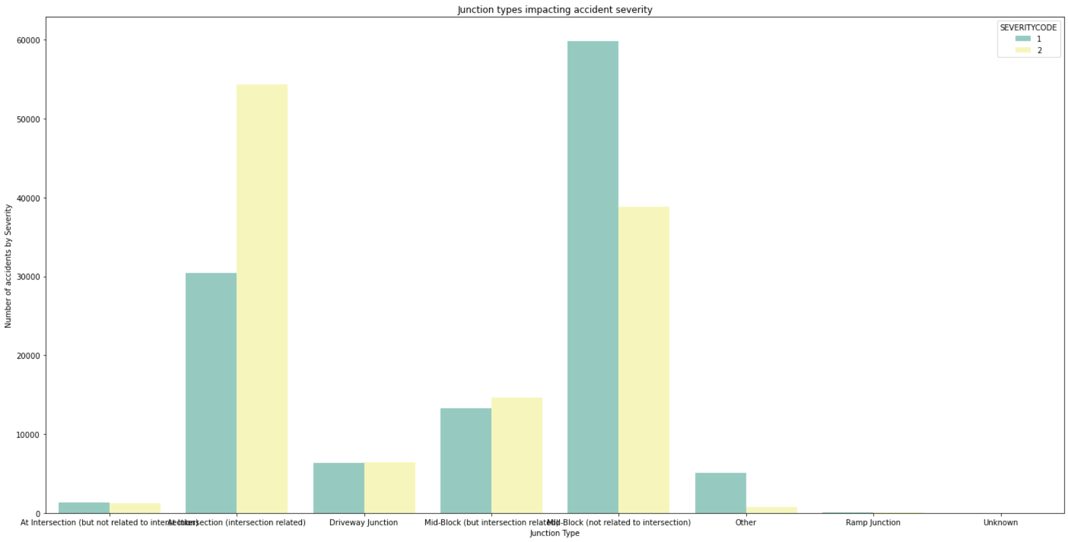
From the barplot, we can see that **COLLISIONTYPE** has strong relationship with SEVERITYCODE influencing the accident severity. Hence, it should be considered as part of the independent variables to predict accident severity.

3.7 Relationship between JUNCTIONTYPE and SEVERITYCODE:

	JUNCTIONTYPE	SEVERITYCODE	JUNCNCOUNT
0	At Intersection (but not related to intersection)	1	1290
1	At Intersection (but not related to intersection)	2	1246
2	At Intersection (intersection related)	1	30446
3	At Intersection (intersection related)	2	54348
4	Driveway Junction	1	6336
5	Driveway Junction	2	6468
6	Mid-Block (but intersection related)	1	13258
7	Mid-Block (but intersection related)	2	14594
8	Mid-Block (not related to intersection)	1	59890
9	Mid-Block (not related to intersection)	2	38808
10	Other	1	5074
11	Other	2	800
12	Ramp Junction	1	80
13	Ramp Junction	2	108
14	Unknown	1	2
15	Unknown	2	4

3.7.1 JUNCTIONTYPE vs SEVERITYCODE - Data Visualization

```
Text(0.5, 1.0, 'Junction types impacting accident severity')
```



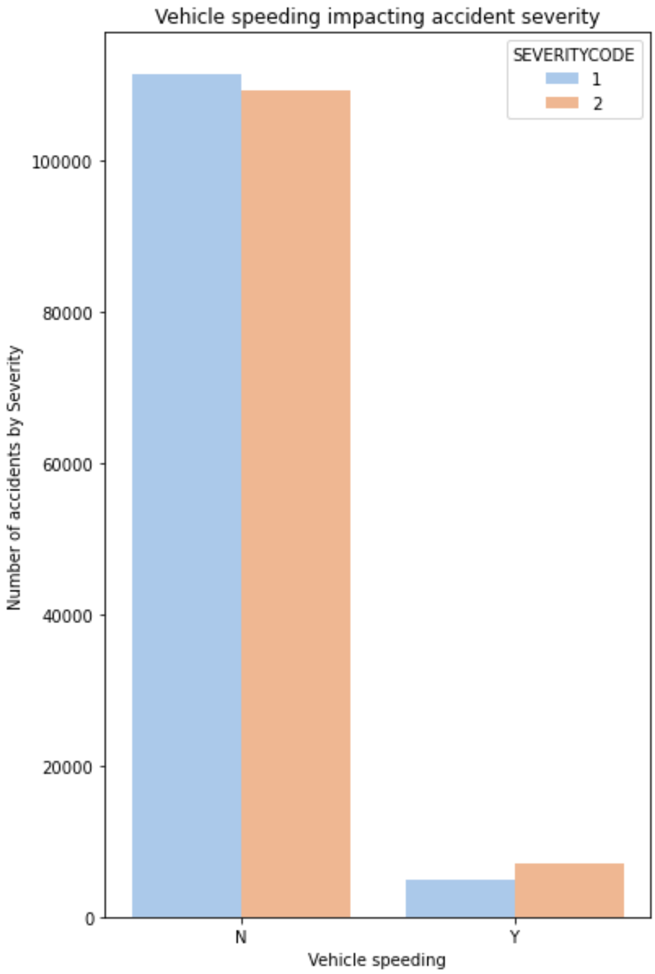
From the barplot, we can see that **JUNCTIONTYPE** has strong relationship with SEVERITYCODE influencing the accident severity. The accidents are more at midblock and intersection points. Hence, it should be considered as part of the independent variables to predict accident severity.

3.8 Relationship between SPEEDING and SEVERITYCODE:

	SPEEDING	SEVERITYCODE	SPEEDCOUNT
0	N	1	111478
1	N	2	109314
3	Y	2	7062
2	Y	1	4898

3.8.1 SPEEEDING vs SEVERITYCODE - Data Visualization

```
Text(0.5, 1.0, 'Vehicle speeding impacting accident severity')
```



From the barplot above, we can see that **SPEEEDING** has less impact on SEVERITYCODE influencing the accident severity. The accident has been reported more from vehicles which were not speeding. Hence, it should **not be considered** as part of the independent variables to predict accident severity.

At the end of Data visualization, we undersgtnad that **UNDERINFL and SPEEDING** cannot be relied upon to predict SEVERITYCODE. Let us remove them from dataframe.

```
UNDERINFL and SPEEDING has been dropped from the Data Frame.
```

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	JUNCTIONTYPE	WEATHER	ROADCONI
25055	1	Intersection	Angles	At Intersection (intersection related)	Raining	Wet
65280	1	Intersection	Angles	At Intersection (intersection related)	Clear	Dry
86292	1	Intersection	Angles	At Intersection (intersection related)	Unknown	Unknown
155111	1	Block	Sideswipe	Mid-Block (not related to intersection)	Clear	Dry
64598	1	Block	Head On	Mid-Block (not related to intersection)	Clear	Dry

```
The variables used for Data modeling and Evaluation are below:

SEVERITYCODE       int64
ADDRTYPE          object
COLLISIONTYPE     object
JUNCTIONTYPE      object
UNDERINFL          int64
WEATHER           object
ROADCOND          object
LIGHTCOND         object
SPEEDING          object
dtype: object
```

4. Data Preprocessing:

The above indpendent variables are now categorical variables. To apply machine learning algorithms, we have to convert the categorical values to a dummy numeric values.

Let us do **Label encoding** to assign a unique numeric value to each catagory of variables.

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	JUNCTIONTYPE	WEATHER	ROADCOND	LIGI
1	1	2	0	1	6	8	2
2	1	2	0	1	1	0	5
3	1	2	0	1	10	7	8
4	1	1	9	4	1	0	5
5	1	1	2	4	1	0	5

Now, we have the target variable balanced and the input feature standardized. Now ,the data is ready to be fed to build data models