IBM Data Science

CAR ACCIDENT SEVERITY

CAPSTONE PROJECT

ARUN CHAUHAN

19^{th} *September* 2020

Introduction

Seattle is a seaport city on the West Coast of the United States. **Seattle** residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it's **no** surprise that **Seattle** sees car **accidents every day.** Our main objectives of this project are to analyze the accident data, Predict the chances and severity of an accident through various data science techniques that would eventually help the residents plan their travel more carefully.

Data

To achieve our goals we will be using Data Collision data (Data-collisions.csv)
All the data is recorded by Traffic control and SPD. This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment.

Out[5]:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight
...	...	...	...	...	...	...	...	...	...	...	...	...	...
194668	2	-122.290826	47.565408	219543	309534	310814	E871089	Matched	Block	NaN	...	Dry	Daylight
194669	1	-122.344526	47.690924	219544	309085	310365	E876731	Matched	Block	NaN	...	Wet	Daylight
194670	2	-122.306689	47.683047	219545	311280	312640	3809984	Matched	Intersection	24760.0	...	Dry	Daylight
194671	2	-122.355317	47.678734	219546	309514	310794	3810083	Matched	Intersection	24349.0	...	Dry	Dusk
194672	1	-122.289360	47.611017	219547	308220	309500	E868008	Matched	Block	NaN	...	Wet	Daylight

194673 rows × 38 columns

The dataset includes all the data from 2004 to present.
The dataset includes 37 attributes and a separate state collision code dictionary. Other datasets will be obtained from Open Government Data portal and open source research groups.
The unbalanced datasets will be inspected first for proper use. The datasets will allow us to train our ML models and predict the severity of accidents and chances of the same.

```
In [6]: df.describe()
```

Out[6]:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	INTKEY	SEVERITYCODE.1	PERSONCOUNT	PEI
count	194673.000000	189339.000000	189339.000000	194673.000000	194673.000000	194673.000000	65070.000000	194673.000000	194673.000000	19467
mean	1.298901	-122.330518	47.619543	108479.364930	141091.456350	141298.811381	37558.450576	1.298901	2.444427	
std	0.457778	0.029976	0.056157	62649.722558	86634.402737	86986.542110	51745.990273	0.457778	1.345929	
min	1.000000	-122.419091	47.495573	1.000000	1001.000000	1001.000000	23807.000000	1.000000	0.000000	
25%	1.000000	-122.348673	47.575956	54267.000000	70383.000000	70383.000000	28667.000000	1.000000	2.000000	
50%	1.000000	-122.330224	47.615369	106912.000000	123363.000000	123363.000000	29973.000000	1.000000	2.000000	
75%	2.000000	-122.311937	47.663664	162272.000000	203319.000000	203459.000000	33973.000000	2.000000	3.000000	
max	2.000000	-122.238949	47.734142	219547.000000	331454.000000	332954.000000	757580.000000	2.000000	81.000000	

Methodology

Once the data has been collected and analysed. I Started with inspecting and cleaning through the following ways

Removing of unrelated columns and Empty columns -

- "X", "Y", "EXCEPTRSNCODE" and "EXCEPTRSNDESC"

```
In [4]: df = df.drop('X',axis = 1)
        df
```

```
In [5]: df = df.drop(['Y','EXCEPTRSNCODE','EXCEPTRSNDESC'], axis =1)
        df
```

194669	1	219544	309085	310365	E876731	Matched	Block	NaN	AURORA AVE N BETWEEN N 85TH ST AND N 86TH ST	1	...	Wet
194670	2	219545	311280	312640	3809984	Matched	Intersection	24760.0	20TH AVE NE AND NE 75TH ST	2	...	Dry
194671	2	219546	309514	310794	3810083	Matched	Intersection	24349.0	GREENWOOD AVE N AND N 68TH ST	2	...	Dry
194672	1	219547	308220	309500	E868008	Matched	Block	NaN	34TH AVE BETWEEN E MARION ST AND E SPRING ST	1	...	Wet

194673 rows × 34 columns

```
In [6]: df.shape
Out[6]: (194673, 34)
```

Deleting Duplicate Column

Resultant Dataframe :

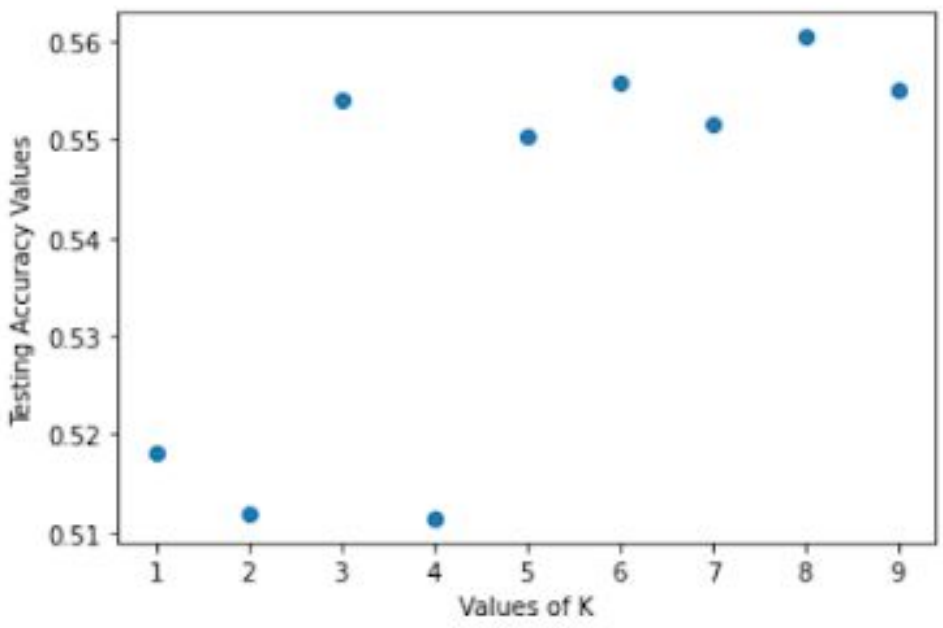
Out[7]:

	SEVERITYCODE	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	SEVERITYDESC	...	ROADCOND	LI
0	2	1	1307	1307	3502005	Matched	Intersection	37475.0	5TH AVE NE AND NE 103RD ST	Injury Collision	...	Wet	
1	1	2	52200	52200	2607959	Matched	Block	NaN	AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	Property Damage Only Collision	...	Wet	
2	1	3	26700	26700	1482393	Matched	Block	NaN	4TH AVE BETWEEN SENECA ST AND UNIVERSITY ST	Property Damage Only Collision	...	Dry	
									2ND AVE				

```
In [8]: new_df.shape
Out[8]: (194673, 33)
```

For my model to be unbiased and give accurate results I balanced the dataset and co-related the Severity index with conditions provided such as Weather, Lighting, Road using K-means approach.

Results



Conclusion

With Factors like Road condition, Lighting and Weather the chances and severity of road accidents increases more than any other factor.