CAR ACCIDENT SERVERITY PROBLEM

Introduction/Business Problem

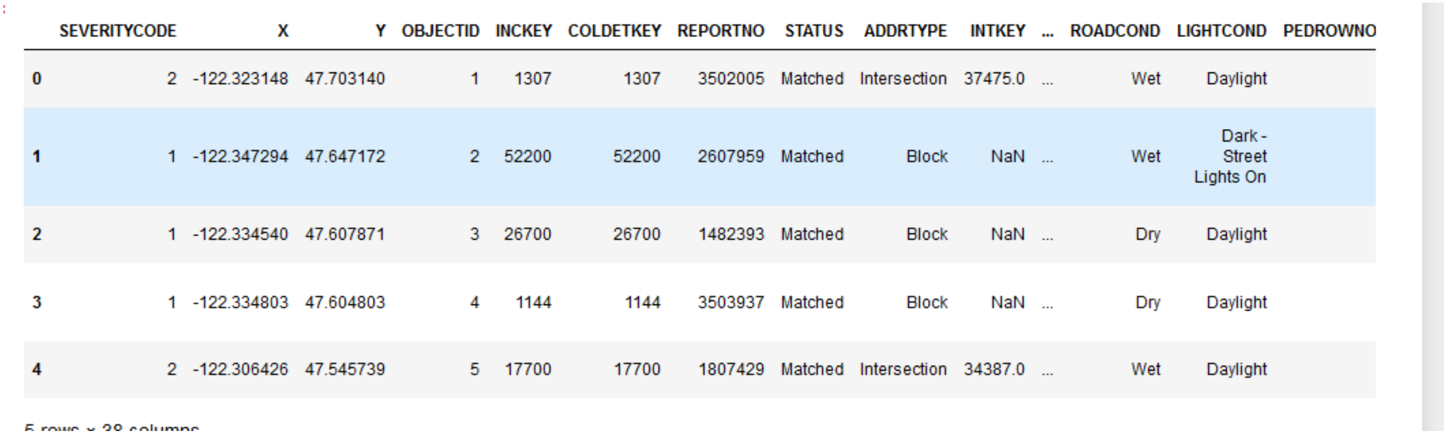
In Our Day today life, we may not experience things that happen suddenly always but things are still happening around which we don’t have control over. One of such incidents is Car accidents and there are serval reasons for that to occur. There are cases where lives are saved with immediate action and cannot be saved also. This project is done on regard to Prevent or reduce the such incidents. The project will work on different cases the accidents occur and provide a solution how it can be reduced after analyzing the reasons for the incidents. Below Project will give the Government/ People understanding on the accidents occur in certain locations and how to prevent the accident from happening.

Data section:

In section, we are going to understand the data set we are going pick as samples. The data set will consist of inputs such as location, severity or no of accidents occurred, weather condition, road condition, period of the day and other optional inputs like road signal, traffic condition and cause of accident. These inputs will be processed and used for the analysis purpose.

The severity will be numbered from 0-5 where 0 is least time accident occurred and 5 is the maximum value. The weather condition will be rainy or dry. Road condition will be normal, wet road or damaged. And finally the period of the day will be day, noon and night which also help us understand the cause of an accident.

Sample data set as below.



	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNO
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	

5 rows × 38 columns

Methodology Section

K-Nearest Neighbor (KNN)

KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

Decision Tree

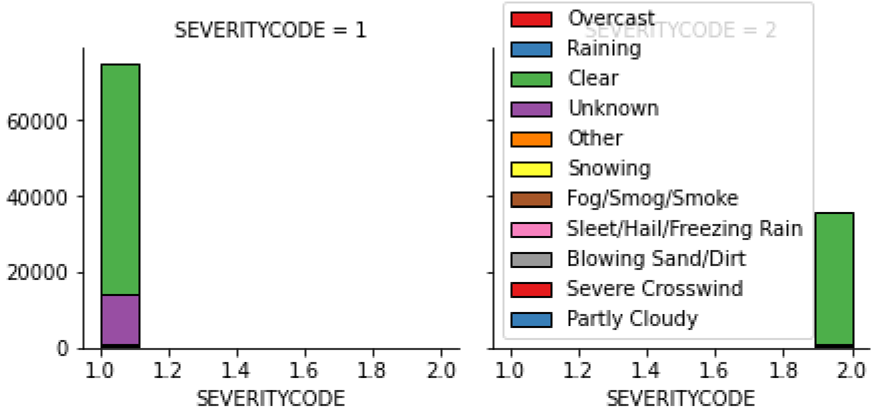
A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

Logistic Regression

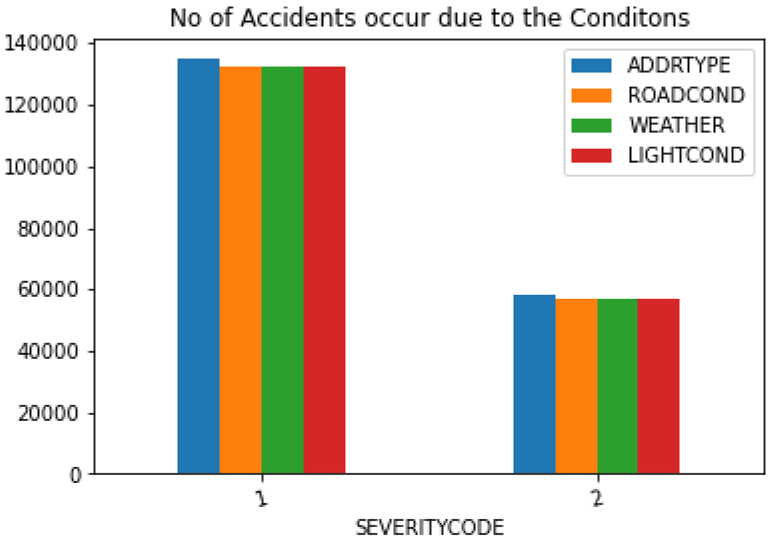
Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

Analysis Section

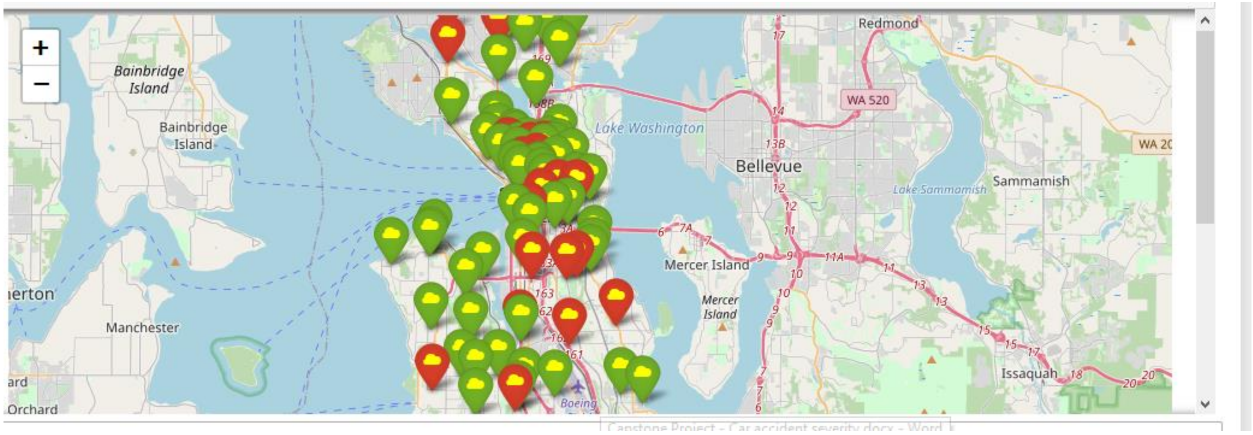
Seaborn chart to represent the data structure of Weather conditions for the Accidents.



With the Accidents predicting chart the no of accidents occurred In Severity1 is more than the Severity 2. The Accidents are occurring in equal sets for the Road Condition, Weather and Light Condition.



Map to project number of Accidents occurred in the Certain Areas



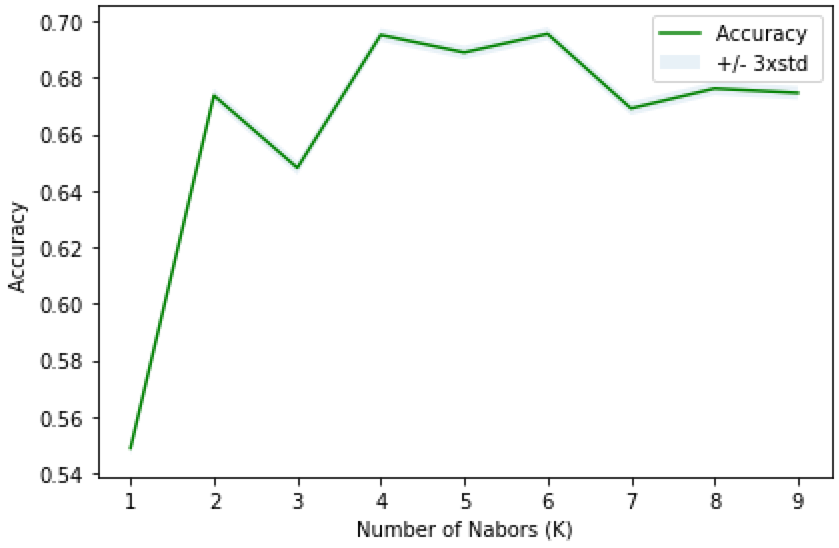
Train/Test Split

We will use 20% of our data for testing and 80% for training.

```
Train set: (155738, 4) (155738,)
Test set: (38935, 4) (38935,)
```

K-Nearest Neighbor (KNN)

As Per the KN Neighbor below the Highest level of Accuracy is achieved at Level 6.



```
Best accuracy: 0.6954924874791318 k= 6
```

Decision Tree

The Decision tree logics has been predicted with the Accuracy is Achieved at 6.9.

Decision Tree Accuracy: 0.6993082428683949

Logistic Regression

The Logistic Regression Analysis shows the Highest level of Accuracy is achieved at the level 6

```
Train set: (155738, 4) (155738,)
Test set: (38935, 4) (38935,)
0.5874048132576434
```

Results & Evaluation

Below Table shows the Predicted values and the Log Loss Value.

	KNN	Decision	LOG
F1-score	0.60	0.58	0.5805
Jaccard Score	0.64	0.70	0.6939
			LogLoss: : 0.59

Discussion

In the beginning of this notebook, we had categorical data that was of type 'object'. This is not a data type that we could have fed through an algorithm, so label encoding was used to created new classes that were of type int8; a numerical data type.

After solving that issue, we were presented with another - imbalanced data. As mentioned earlier, class 1 was nearly three times larger than class 2. The solution to this was down sampling the majority class with sklearn's resample tool. We down sampled to match the minority class exactly with 58188 values each.

Once we analyzed and cleaned the data, it was then fed through three ML models; K-Nearest Neighbor, Decision Tree and Logistic Regression. Although the first two are ideal for this project, logistic regression made most sense because of its binary nature.

Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and logloss for logistic regression. Choosing different k, max depth and hyparameter C values helped to improve our accuracy to be the best possible.

Conclusion

Based on historical data from weather conditions pointing to certain classes, we can conclude that particular weather conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).