**IBM Data Science Capstone:**

Car Accident Severity Report

* **Background:**

Imagine you have to travel for a Business meeting and you take your car off and start driving. The weather is not good leading to poor visibility conditions. On the way, you come across a huge traffic - cars lined behind one another and there is hardly any chance for the traffic to move swiftly.

You come to know that a car has met with a terrible accident being the cause for the traffic jam and the passengers are severely injured.

We come across such situations or read about it almost every day thinking of a solution where we can avoid such accidents or receive an alert of the severity forehand.

* **Introduction | Business Understanding:**

In an effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful predicting the severity.

* **Data Understanding | Data Section:**

Our predictor or target(dependent) variable will be 'SEVERITYCODE' as per the dataset since it is the measure of the severity of an accident ranged from 0 to 4. The attributes highly contributing to the severity of an accident in our dataset are 'WEATHER', 'ROADCOND', 'LIGHTCOND' (weather, road conditions and light conditions) represented as independent variables.

The severity code representation is as follows:

0 - unknown

1 - prop damage or property damage

2 - injury

2b - serious injury

3 - fatality

* **Feature Engineering:**

By looking at our dataset we come to know that our data is not fit for analysis. We will need to remove some columns which will not serve any purpose for our model and also we need to balance and normalize our dataset.

Also, some columns are of type object which are need to be converted to numeric type for the ease in our model building.

* **Methodology:**

Our data is now ready to be fed into machine learning models.

We will use the following models:

**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.

* **Result and Evaluation:**

For KNN:

The model is more accurate when k=

For Decision Tree:

The model is more accurate when depth =

For Logistic regression:

The model is more accurate for

* **Discussion:**

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 downsampling the majority class with sklearn's resample tool. We downsampled 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).