Applied Data Science Capstone Project –

Car accident severity (Week 1)

1. Introduction | Business Problem

In order to reduce the frequency of car collisions in a community, an algorithm is recommended to predict the severity of an accident given the key parameters such as current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

2. Data Understanding

The data was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present.

Our predictor is 'SEVERITYCODE'. It is used measure the severity of an accident from 0 to 4 within the dataset. Our attributes are used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

Severity codes are as follows:

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

3. Data Preprocessing

The dataset in the original form is not ready for data analysis. In order to prepare the data, first, we need to drop the non-relevant columns. In addition, most of the features are of object data types that need to be converted into numerical data types.

After analyzing the data set, I have decided to focus on only four features, severity, weather conditions, road conditions, and light conditions.

4. Balancing the Dataset

Our target variable SEVERITYCODE is only 42% balanced. SEVERITYCODE in class 1 is nearly three times the size of class 2.

We can fix this by down sampling the majority class.

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

5. 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 good to use with logistic regression.