IBM CAPSTONE PROJECT

RAMESH KUMAR SHARMA

Data

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

The data consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

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

Furthermore, because of the existence of null values in some records, the data needs to be preprocessed before any further processing.

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, among others.

To get a good understanding of the dataset, I have checked different values in the features. The results show, the target feature is imbalance, so we use a simple statistical technique to balance it.

```
In [26]:  1  pre_df["SEVERITYCODE"].value_counts()
```

```
Out[26]: 1    136485
         2     58188
         Name: SEVERITYCODE, dtype: int64
```

As you can see, the number of rows in class 1 is almost three times bigger than the number of rows in class 2. It is possible to solve the issue by downsampling the class 1.

In [27]:
```
from sklearn.utils import resample

```

In [28]:
```
pre_df_maj = pre_df[pre_df.SEVERITYCODE==1]
pre_df_min = pre_df[pre_df.SEVERITYCODE==2]

pre_df_maj_dsample = resample(pre_df_maj,
                              replace=False,
                              n_samples=58188,
                              random_state=123)

balanced_df = pd.concat([pre_df_maj_dsample, pre_df_min])

balanced_df.SEVERITYCODE.value_counts()
```

Out[28]:
```
2    58188
1    58188
Name: SEVERITYCODE, dtype: int64
```

Thank You