Accident Severity Prediction

Introduction/ Business Problem

The Washington State Department of Transportation Crash Data Portal provides crash information for accidents that occurred state-wide. According to the 2019 data, there were 45,524 accidents on all roads. Of those:

- 235 were fatal crashes
- 973 were suspected of serious injury accidents
- 2,798 were suspected of minor injury accidents
- 9,412 were possible injury crashes
- 32,106 were no apparent injury collisions

Our motivation is to use the weather, location and road condition data provided in the dataset, made available by the Seattle Department of Transportation Traffic Management Division, to arrive at a correlation to predict the severity of road accidents. This tool/data can then be made available to the public and the Seattle traffic authorities to possibly prevent/reduce severe or fatal accidents in the future by taking precautionary measures.

Data Understanding

We chose the unbalanced dataset provided by the Seattle Department of Transportation Traffic Management Division with 194673 rows (accidents) and 37 columns (features) where each accident is given a severity code. It covers accidents from January 2004 to May 2020. Some of the features in this dataset include and are not limited to Severity code, Location/Address of accident, Weather condition at the incident site, Driver state (whether under influence or not), collision type. Hence we think its a good generalized dataset which will help us in creating an accurate predictive model.

The unbalance with respect to the severity code in the dataset is as follows.

SEVERITY CODE	Count
1	136485
2	58188

Data Pre-processing

An unbalanced dataset is used, provided by the Seattle Department of Transportation Traffic Management Division with 194673 rows (accidents) and 37 columns (features) where each accident is given a severity code. The steps taken in pre-processing the dataset are as follows.

1. **Removal of irrelevant columns or features**

   Columns containing descriptions and identification numbers that would not help in the classification are dropped from the dataset to reduce the complexity and dimensionality of the dataset. 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'INTKEY', 'EXCEPTRSNCODE' and more belong to this category. Certain other categorical features were removed as they had a large number of distinct values, example: ‘LOCATION’.

After performing this step, the dimensionality dropped from 37 to 18.

```
Data columns (total 37 columns):
```

V

```
Data columns (total 15 columns):
```

2. **Identification and handling missing values**

To identify columns and rows with missing values is the next step. Empty boxes, 'Unknown' and 'Other' were values considered as missing values. These were replaced with NA to make the dataset uniform.

```
df.replace(r'^\s*$', np.nan, regex=True)
df.replace("Unknown", np.nan, inplace = True)
df.replace("Other", np.nan, inplace = True)
```

Columns ("INATTENTIONIND","PEDROWNOTGRNT","SPEEDING") which had more than 20% of its values missing were noted down and were dropped. For columns ("X","Y","COLLISIONTYPE","JUNCTIONTYPE"….) which had less than 20% of its values missing, the respective rows were removed since most of the columns in this dataset are categorical type, goal was to not impute the non-numerical columns; hence it did not make sense to replace the values.

Once the above two strategies were performed, the dataset reduced from having 194673 rows and 15 columns to having 143747 rows and 15 columns.

```
Int64Index: 143747 entries, 0 to 194672
Data columns (total 15 columns):
```

3. **Balancing the dataset**

With the above two pre-processing steps complete, a dataset (143747 rows) with 94821 rows for severity code 1 and 48926 rows for severity code 2 is obtained. Training an algorithm on an unbalanced dataset w.r.t the target category will result in a biased model. The model will have learnt more about one the category that has more data. In order to prevent this, a new balanced dataset (97852 rows) is created by randomly sampling out 48926 rows with severity code 2 and then concatenating it with 48926 rows with severity code 1. The dataset is then shuffled to randomize the rows.

```
Int64Index: 97852 entries, 139054 to 72625
Data columns (total 15 columns):
```

4. **Encoding of data**

The dataset is split into two datasets, X and Y, where Y contains the target feature (SEVERITYCODE) and X contains all the independent features/variables.

Machine Learning models are trained only on numerical data; hence all categorical features in the dataset have to be encoded so that the algorithms can be trained on those features. The 'get_dummies' method from pandas library is used to convert/encode each and every categorical feature. After application, number of features in dataset X increased from 14 to 50.

```
Int64Index: 97852 entries, 139054 to 72625
Data columns (total 50 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   X                                                       97852 non-null  float64
 1   Y                                                       97852 non-null  float64
 2   PERSONCOUNT                                             97852 non-null  int64
 3   PEDCOUNT                                                97852 non-null  int64
 4   PEDCYLCOUNT                                             97852 non-null  int64
 5   VEHCOUNT                                                97852 non-null  int64
 6   ADDRTYPE_Block                                          97852 non-null  uint8
 7   ADDRTYPE_Intersection                                   97852 non-null  uint8
 8   COLLISIONTYPE_Angles                                    97852 non-null  uint8
 9   COLLISIONTYPE_Cycles                                    97852 non-null  uint8
 10  COLLISIONTYPE_Head On                                   97852 non-null  uint8
 11  COLLISIONTYPE_Left Turn                                 97852 non-null  uint8
 12  COLLISIONTYPE_Parked Car                                97852 non-null  uint8
 13  COLLISIONTYPE_Pedestrian                                97852 non-null  uint8
```

5. **Splitting into training and testing datasets**

   The datasets X and Y are split into X_train, Y_train, X_test,and Y_test. The first two will be used for training purposes and the last two will be used for testing purposes. The split ratio is 0.8, 80% of data is used for training and 20% of is used for testing.

```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.2,random_state=0)
```

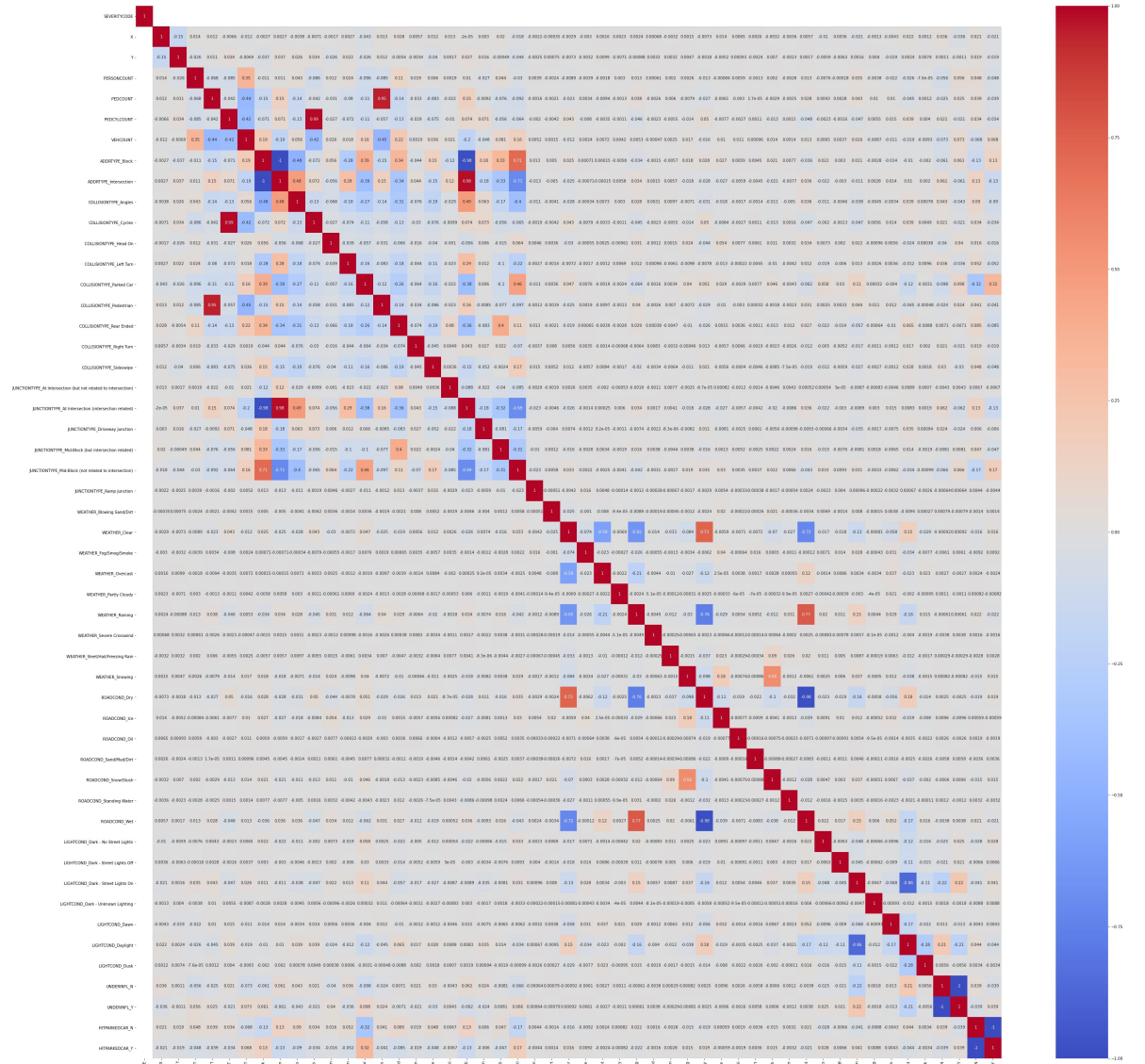
6. **Normalizing/ Feature scaling of data**

   Feature scaling of data is done to normalize the data in a dataset to a specific range. It also helps improve the performance of the ML algorithms. Standard Scaler metric is used to scale/normalize all the numerical data for both, the X_train and X_test datasets. This completes the pre-processing stage, we can move on to training our models.

Understanding Correlation in Dataset

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. Finding the correlation among the features of the dataset helps understand the data better. For example, in the below figure (correlation plot using matplotlib), it can be observed that some features have a strong positive/negative correlation while most of them have weak/ no correlation.

Examples, There is a strong positive correlation between 'PEDCYLCOUNT' and 'COLLISIONTYPE_Cycles'. This means that if the collision involves cycles, at-least one cyclist is involved in the accident. There is a strong negative correlation between 'ROADCOND_Wet' and 'ROADCOND_Dry', meaning that if the road is wet it cannot be dry. This is how we can get a deeper understanding of the data using correlation plots.


SEVERITYCODE
X
Y
PERSONCOUNT
PEDCOUNT
PEDCYLCOUNT
VEHCOUNT
ADDRTYPE_Block
ADDRTYPE_Intersection
COLLISIONTYPE_Angles
COLLISIONTYPE_Cycles
COLLISIONTYPE_Head On
COLLISIONTYPE_Left Turn
COLLISIONTYPE_Parked Car
COLLISIONTYPE_Pedestrian
COLLISIONTYPE_Rear Ended
COLLISIONTYPE_Right Turn
COLLISIONTYPE_Sideswipe
JUNCTIONTYPE_At Intersection (but not related to intersection)
JUNCTIONTYPE_At Intersection (intersection related)
JUNCTIONTYPE_Driveway Junction
JUNCTIONTYPE_Mid-Block (but not related to intersection)
JUNCTIONTYPE_Mid-Block (not related to intersection)
JUNCTIONTYPE_Ramp Junction
WEATHER_Blowing Sand/Dirt
WEATHER_Clear
WEATHER_Fog/Smog/Smoke
WEATHER_Overcast
WEATHER_Partly Cloudy
WEATHER_Raining
WEATHER_Severe Crosswind
WEATHER_Sleet/Hail/Freezing Rain
WEATHER_Snowing
ROADCOND_Dry
ROADCOND_Ice
ROADCOND_Oil
ROADCOND_Sand/Mud/Dirt
ROADCOND_Snow/Slush
ROADCOND_Standing Water
ROADCOND_Wet
LIGHTCOND_Dark - No Street Lights
LIGHTCOND_Dark - Street Lights Off
LIGHTCOND_Dark - Street Lights On
LIGHTCOND_Dark - Unknown Lighting
LIGHTCOND_Dawn
LIGHTCOND_Daylight
LIGHTCOND_Dusk
LIGHTCOND_Other
UNDERINFL_N
UNDERINFL_Y
HITPARKEDCAR_N
HITPARKEDCAR_Y
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00