Data Science Capstone: Car Accident Severity Prediction

1. Introduction

1.1 Background

Road traffic injuries are one of the major public health problems in the world. According with the World Health Organization, approximately 1.35 million people die each year as a result of road traffic crashes.

1.2 The problem

The project's objective is to predict the severity and probability of a car accident analyzing data from Seattle city in USA, doing attribute selection, feature engineering and applying machine learning algorithms to choose one model with the best performance after evaluation.

2. Data

The data is downloaded from Seattle GeoData web site (https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0). This is the information about the dataset.

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 221738 entries, 0 to 221737
Data columns (total 40 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   X                 214260 non-null  float64
 1   Y                 214260 non-null  float64
 2   OBJECTID          221738 non-null  int64
 3   INCKEY            221738 non-null  int64
 4   COLDETKEY         221738 non-null  int64
 5   REPORTNO          221738 non-null  object
 6   STATUS            221738 non-null  object
 7   ADDRTYPE          218024 non-null  object
 8   INTKEY            72027 non-null   float64
 9   LOCATION          217145 non-null  object
 10  EXCEPTRSNCODE     101335 non-null  object
 11  EXCEPTRSNDESC     11785 non-null   object
 12  SEVERITYCODE      221737 non-null  object
 13  SEVERITYDESC      221738 non-null  object
 14  COLLISIONTYPE     195287 non-null  object
 15  PERSONCOUNT       221738 non-null  int64
 16  PEDCOUNT          221738 non-null  int64
 17  PEDCYLCOUNT       221738 non-null  int64
 18  VEHCOUNT          221738 non-null  int64
 19  INJURIES          221738 non-null  int64
 20  SERIOUSINJURIES   221738 non-null  int64
 21  FATALITIES        221738 non-null  int64
 22  INCDATE           221738 non-null  object
 23  INCDTTM           221738 non-null  object
 24  JUNCTIONTYPE     209759 non-null  object 
 25  SDOT_COLCODE     221737 non-null  float64
 26  SDOT_COLDESC     221737 non-null  object 
 27  INATTENTIONIND   30188 non-null   object 
 28  UNDERINFL        195307 non-null  object 
 29  WEATHER          195097 non-null  object 
 30  ROADCOND         195178 non-null  object 
 31  LIGHTCOND        195008 non-null  object 
 32  PEDROWNOTGRNT    5195 non-null    object 
 33  SDOTCOLNUM       127205 non-null  float64
 34  SPEEDING         9936 non-null    object 
 35  ST_COLCODE       212325 non-null  object 
 36  ST_COLDESC       195287 non-null  object 
 37  SEGLANEKEY       221738 non-null  int64  
 38  CROSSWALKKEY     221738 non-null  int64  
 39  HITPARKEDCAR     221738 non-null  object 
dtypes: float64(5), int64(12), object(23)
memory usage: 67.7+ MB
```

The dataset contains 40 columns and 221738 rows. It is composed by 17 numeric variables and 23 string variables. The target column is SEVERITYCODE, which also has a description column (SEVERITYDESC). This leave 38 possible predictors for our purpose.

0	Unknown
1	Property Damage Only Collision
2	Injury Collision
2b	Serious Injury Collision
3	Fatality Collision

The values in target variable make the dataset unbalanced, it will be considered when splitting for training and testing the model.

```
df['SEVERITYCODE'].value_counts()
1     137776
2      58842
0      21656
2b      3111
3        352
Name: SEVERITYCODE, dtype: int64
```