Prediction of Accident Severity

1. Introduction

1.1. Background

In traffic situations, passengers are prone to accidents on the roads. This can be due to different factors such as the weather conditions, the road conditions, the light conditions amongst other factors. These attributes are being stored by the traffic system in Seattle. It is highly recommended to be able to predict the severity of an accident based on the factors available to prepare for the casualty before the accident occurs.

1.2. Problem

The dataset provided for the Seattle city contains a total of 38 attributes (relating to the accidents that occur on the road) and the labelled data which describes the fatality of an incident. Given this dataset, the aim of this project is to select the necessary attributes that will be used to build a model that will help to predict the severity of an accident.

1.3. Interest

Residents of Seattle will find this helpful in predicting how severe an accident will be if they get into one based on the factors available. It will also be useful for traffic attendants and paramedic to prepare for accidents likely to happen. This will help reduce causality.

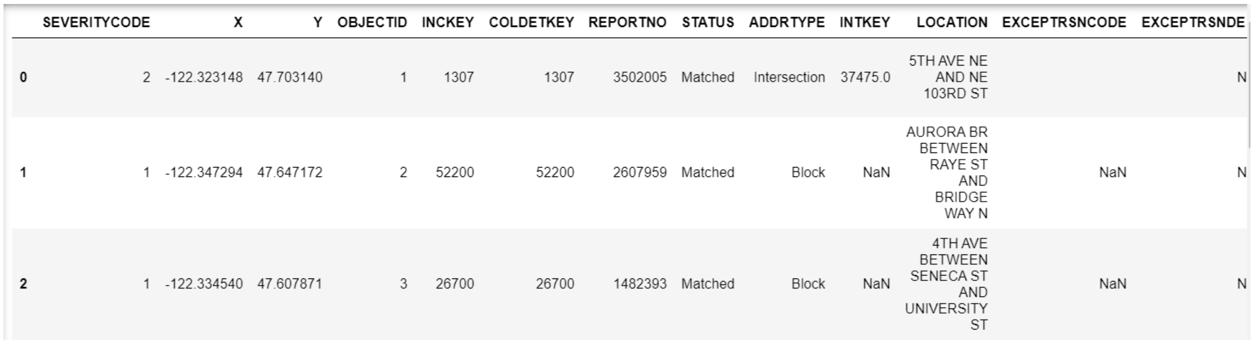
2. Data acquisition and cleaning

2.1. Data Sources

The dataset for Seattle road accidents was provided containing a total of 194673 observations and 37 attributes with many of them being categorical attributes.

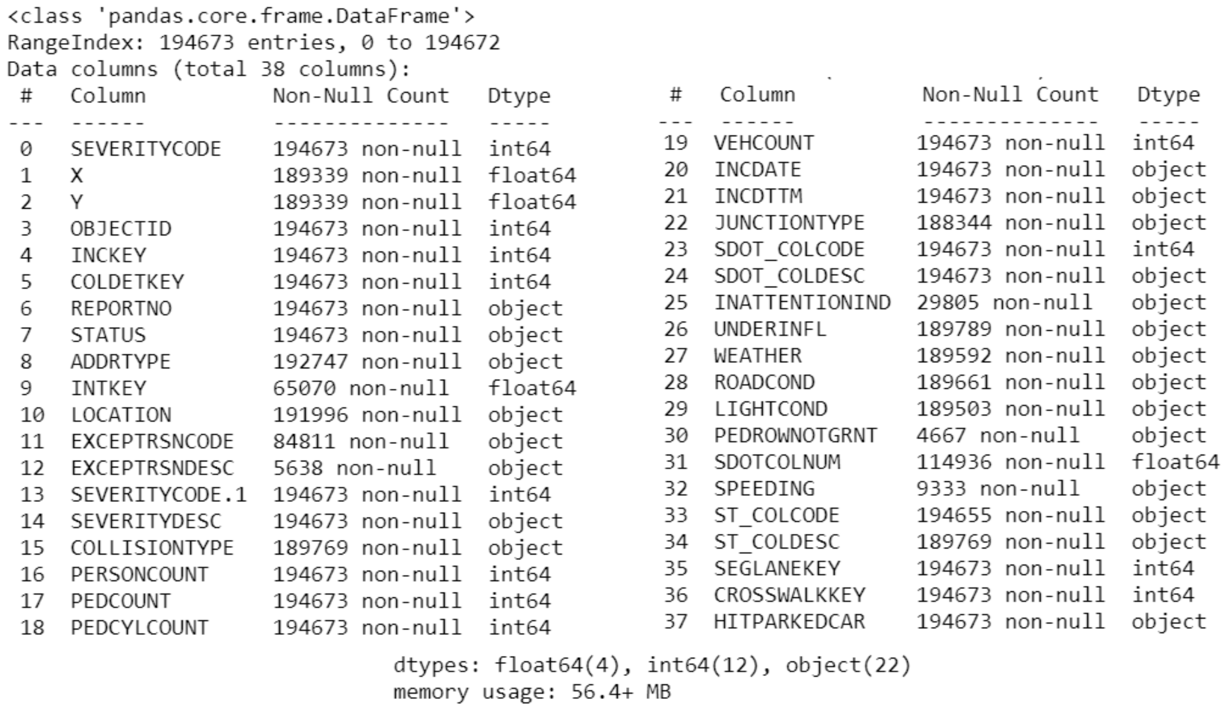
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Prediction of Accident Severity



	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSNCODE	EXCEPTRSNDE
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	5TH AVE NE AND NE 103RD ST		N
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	NaN	N
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	4TH AVE BETWEEN SENECA ST AND UNIVERSITY ST	NaN	N

Figure 1 Data frame from data read from given source.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 38 columns):
 #   Column          Non-Null Count   Dtype        #   Column          Non-Null Count   Dtype
---  ------          --------------   -----       ---  ------          --------------   -----
 0   SEVERITYCODE    194673 non-null  int64        19  VEHCOUNT        194673 non-null  int64
 1   X               189339 non-null  float64      20  INCDATE         194673 non-null  object
 2   Y               189339 non-null  float64      21  INCDTTM         194673 non-null  object
 3   OBJECTID        194673 non-null  int64        22  JUNCTIONTYPE    188344 non-null  object
 4   INCKEY          194673 non-null  int64        23  SDOT_COLCODE    194673 non-null  int64
 5   COLDETKEY       194673 non-null  int64        24  SDOT_COLDESC    194673 non-null  object
 6   REPORTNO        194673 non-null  object       25  INATTENTIONIND  29805 non-null   object
 7   STATUS          194673 non-null  object       26  UNDERINFL       189789 non-null  object
 8   ADDRTYPE        192747 non-null  object       27  WEATHER         189592 non-null  object
 9   INTKEY          65070 non-null   float64      28  ROADCOND        189661 non-null  object
 10  LOCATION        191996 non-null  object       29  LIGHTCOND       189503 non-null  object
 11  EXCEPTRSNCODE   84811 non-null   object       30  PEDROWNOTGRNT   4667 non-null    object
 12  EXCEPTRSNDESC   5638 non-null    object       31  SDOTCOLNUM      114936 non-null  float64
 13  SEVERITYCODE.1  194673 non-null  int64        32  SPEEDING        9333 non-null    object
 14  SEVERITYDESC    194673 non-null  object       33  ST_COLCODE      194655 non-null  object
 15  COLLISIONTYPE   189769 non-null  object       34  ST_COLDESC      189769 non-null  object
 16  PERSONCOUNT     194673 non-null  int64        35  SEGLANEKEY      194673 non-null  int64
 17  PEDCOUNT        194673 non-null  int64        36  CROSSWALKKEY    194673 non-null  int64
 18  PEDCYLCOUNT     194673 non-null  int64        37  HITPARKEDCAR    194673 non-null  object
                              dtypes: float64(4), int64(12), object(22)
                              memory usage: 56.4+ MB
```

Figure 2 Data frame information about available columns

2.2. Data Cleaning and Preprocessing

Out of the numerous attributes, only selected attributes were used because they relate to the severity code based on their description in the metadata. The attributes are: 'ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'WEATHER','ROADCOND','LIGHTCOND', 'PERSONCOUNT','PEDCOUNT', 'VEHCOUNT', 'HITPARKEDCAR'. The other attributes were dropped either because they do not relate to the target variable or because they have a lot of missing values.

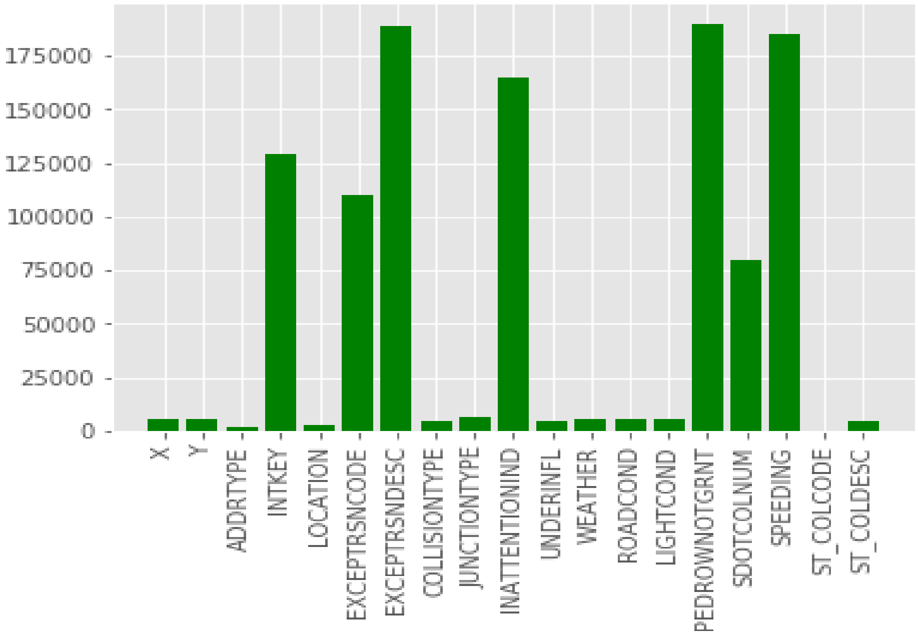


Figure 3 To analyze ratio of NaN across all selected columns

The following attributes are categorical values and needed to be changed to numerical values using the categorization of data.

- ADDRTYPE
- COLLISIONTYPE
- JUNCTIONTYPE
- UNDERINFL
- WEATHER
- ROADCOND
- LIGHTCOND
- SPEEDING
- HITPARKEDCAR

Applied required conditioning on columns like removal of rows with NaN values, type conversion etc.