Coursera - Applied Data Science Capstone
Ana Cristina Bittar de Oliveira

Capstone Project – Week 1

1. A description of the problem and a discussion of the background. (**15 marks**)

How to prevent or, at least, reduce the severity of road accidents is a very important question. A better understanding of the probable causes of collisions could be a good approach in order to suggest actions to be taken or improvements to be made in traffic engineering systems willing to prevent them.

The main potential consequences of a car collision are:

- Injuries
- Traffic jams
- Related costs (government, insurance companies, individuals, *et al*)

Seattle Traffic Management Division releases weekly updates about all collisions recorded by Traffic Records since 2004. The main goal is to analyze that data to identify patterns and make predictions about risk and severity of the accidents based on some common attributes. With this information an alert system could be developed to provide guidance on what to do in short, medium, or large term.

Some examples of actions to be taken could be:

- deviate traffic in pre-defined times or weather conditions (preventive)
- improve traffic signs
- reduce speed in specific roads in determined weather conditions
- reduce speed in specific roads (permanently)
- make structural changes on traffic
- educative campaigns and advertisement
- recommendations to improve the data collection
- other

The present work involves data analysis related to collisions in the city of Seattle. It would be out of scope to generalize this model for sites other than those included on the data to be analyzed, although intuition indicates that, for similar traffic systems, the results could be extended. The expected result is a model capable to predict an accident severity in pre-determined conditions.

2. A description of the data and how it will be used to solve the problem. (**15 marks)**

This project will use the dataset shared in the Capstone. The CSV file was obtained directly from Seattle Open Data Portal.

https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data

The dataset provides information about all types of collisions occurred in the city of Seattle since 2004 and is updated weekly. It was also provided metadata for better understanding of the dataset.

Attribute Information:
https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

The data will be used to make predictions about probability and severity of car accidents based mainly on:

- Weather conditions
- Road conditions
- Light conditions

Further analysis (maybe using additional datasets) could be made to check if it's possible to correlate some driver conditions (as speed, influence of drugs or alcohol, inattention) with other attributes like time of the day, day of the week, events like concerts or games (if the data is available) and define whether or not those attributes could be used for predictions. This is out of scope of the present report.

Some attributes are redundant, or the number of observations is irrelevant, and will be excluded.

Info about the original data file:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 221389 entries, 0 to 221388
Data columns (total 40 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   X                213918 non-null  float64
 1   Y                213918 non-null  float64
 2   OBJECTID         221389 non-null  int64
 3   INCKEY           221389 non-null  int64
 4   COLDETKEY        221389 non-null  int64
 5   REPORTNO         221389 non-null  object
 6   STATUS           221389 non-null  object
 7   ADDRTYPE         217677 non-null  object
 8   INTKEY           71884 non-null   float64
 9   LOCATION         216801 non-null  object
 10  EXCEPTRSNCODE    100986 non-null  object
 11  EXCEPTRSNDESC    11779 non-null   object
 12  SEVERITYCODE     221388 non-null  object
 13  SEVERITYDESC     221389 non-null  object
 14  COLLISIONTYPE    195159 non-null  object
 15  PERSONCOUNT      221389 non-null  int64
 16  PEDCOUNT         221389 non-null  int64
 17  PEDCYLCOUNT      221389 non-null  int64
 18  VEHCOUNT         221389 non-null  int64
 19  INJURIES         221389 non-null  int64
 20  SERIOUSINJURIES  221389 non-null  int64
 21  FATALITIES       221389 non-null  int64
 22  INCDATE          221389 non-null  object
 23  INCDTTM          221389 non-null  datetime64[ns]
 24  JUNCTIONTYPE     209417 non-null  object
 25  SDOT_COLCODE     221388 non-null  float64
 26  SDOT_COLDESC     221388 non-null  object
 27  INATTENTIONIND   30188 non-null   object
 28  UNDERINFL        195179 non-null  object
 29  WEATHER          194969 non-null  object
 30  ROADCOND         195050 non-null  object
 31  LIGHTCOND        194880 non-null  object
```

```
 32  PEDROWNOTGRNT    5192 non-null    object
 33  SDOTCOLNUM       127205 non-null  float64
 34  SPEEDING         9928 non-null    object
 35  ST_COLCODE       211976 non-null  object
 36  ST_COLDESC       195159 non-null  object
 37  SEGLANEKEY       221389 non-null  int64
 38  CROSSWALKKEY     221389 non-null  int64
 39  HITPARKEDCAR     221389 non-null  object
dtypes: datetime64[ns](1), float64(5), int64(12), object(22)
memory usage: 67.6+ MB
```

The table below shows statistics about some numerical variables related to severity. In a supervised model those attributes could be used to define different weights for severity scale.

	PERSONCOUNT	*PEDCOUNT*	*PEDCYLCOUNT*	*VEHCOUNT*	*INJURIES*	*SERIOUSINJURIES*	*FATALITIES*
Count	221,389	221,389	221,389	221,389	221,389	221,389	221,389
Mean	2.227161	0.038136	0.027350	1.731057	0.373962	0.015209	0.001685
Std	1.470190	0.201815	0.164508	0.829259	0.732158	0.158072	0.044701
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	2.0	0.0	0.0	2.0	0.0	0.0	0.0
50%	2.0	0.0	0.0	2.0	0.0	0.0	0.0
75%	3.0	0.0	0.0	2.0	1.0	0.0	0.0
Max	93	6	2	15	78	41	5

Rows with missing values for the main attributes were removed.

	SEVERITYCODE	*WEATHER*	*ROADCOND*	*LIGHTCOND*
Count	194699	194699	194699	194699
Unique	5	11	9	9
Top	1	Clear	Dry	Daylight
Freq.	133575	114556	128304	119384

All the observations with SEVERITYCODE = “0”, described as “Unknown” were removed from dataset.

Some aspects that should be observed are:

- There’s very detailed geographic information, but no demographic about the conductors
- The data went back 15 years, but traffic volume increases rapidly, but the first model won’t differ the significance of most recent occurrences

First attempt will be an unsupervised model.