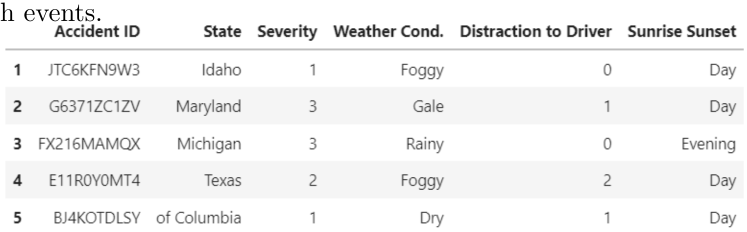
Big DataMining - Assignment 2

Alex Chantzaras

December 2021

1 Part 1 - Creating the Data set

The topic i chose was car accidents that took place in United States the previous year taking into consideration some significant causes that occured the time these accidents happened. After evaluating respective data sets i collected some of the most significant and explainable factors that we should distinguish in such events.



	Accident ID	State	Severity	Weather Cond.	Distraction to Driver	Sunrise Sunset
1	JTC6KFN9W3	Idaho	1	Foggy	0	Day
2	G6371ZC1ZV	Maryland	3	Gale	1	Day
3	FX216MAMQX	Michigan	3	Rainy	0	Evening
4	E11R0Y0MT4	Texas	2	Foggy	2	Day
5	BJ4KOTDLSY	of Columbia	1	Dry	1	Day

1.1 Creating values for every attribute of the Dataframe

In this part i had to create the possible values that each of my attribute would be assigned. Every accident has its own unique accident id which is created by a function named ”id generator”. Each case was connected randomly with the name of one of the fifty six States of America. Respectively, attributes ”Severity” , ”Weather Condition”, ”Sunrise Sunset” and ”Distraction to Driver” took their values from specific lists giving bias to the random values they were assigned. More specifically ”Distraction to Driver” takes higher values to more severe accidents especially when these accidents took place in States that is known to be more dangerous to drive in [1]. With the same way the attribute ”Sunrise Sunset” is connected with the attribute ”Severity” as it is proved that significant amount of accidents with fatal or serious injuries occur at night hours due to many circumstances [2]. The final step of part1 is to store the Data set in order to work with it in a second place.

2 Part 2 - Strategies for adding noise

In this section will be described the way of adding noise to every attribute of the initial data set.

2.1 Noise in ”States” attribute

The first noise that was added in the data set was the replacement of values for the States *Virginia*, *Indiana*,*New Hampshire* and *American Samoa* in column ”State” supposing that this problem was transported to the data set by the respective State agency files data were collected.

2.2 Noise in ”Severity” attribute

The noise added in attribute ”Severity” was connected to possible error in loading files from safe to drive in States[3].In this situation i assumed that files from such States consist of more accidents with lower severity values and specifically accidents with severity value ”1” which by error were introduced to all instances of accidents happened in these States.

2.3 Noise in ”Weather Cond.” attribute

Some other special conditions of other attributes values (”Sunrise Sunset” = Day and ”Weather Cond.” = Foggy) triggered some ”Nan” values to the column ”Weather Cond.”.

2.4 Noise in ”Distraction to Driver” attribute

Consequently, noise added to a subset of column ”Distraction to Driver” to instances that were assigned with the actual interpretation of the categorical value ”1” for Distraction due to data entry error.

2.5 Noise in ”Sunrise Sunset” attribute

In this column a range of instances failed to be loaded to our data set and as a result we face missing values issue.

2.6 Noise in ”Accident ID” attribute

Value ”Not Applicable” was assigned in column ”Accident ID” where state name refers to *New York*. Finally at the loading of our initial file to our program seems that by system fault, the first three characters of attribute ”Accident ID” were omitted.

3 Data Mining questions

- Which factors seem to raise the possibility of observing a severe accident?
- In which States should we apply politics to motivate people utilize public transportation in order to decrease accidents that are taking place during working hours and due to increased traffic jam?

- In which States should we force punitive laws for those who are driving under dangerous circumastances(talking on the phone, having consumed alcohol/drugs etc.) ?
- Should we beat the drum to citizens for avoiding to drive or be very careful when driving when some certain weather conditions occur? Which is this or these conditions?

4 References

1. https://tramontanalaw.com/10-most-dangerous-states-to-drive/
2. https://www.rospa.com/rospaweb/docs/advice-services/road-safety/drivers/driving-at-night.pdf
3. https://www.safewise.com/blog/safest-states-drivers/