**IBM CAPSTONE**

**1. Introduction**

**1.1. Background**

Traffic accidents are a significant source of deaths, injuries, property damage, and a major concern for public health and traffic safety. They involve a great deal of human exposure and economic interest. Only in the US, 4.4 million people are injured every year in car accidents, and over 38.000 people die on the road. This implies a total cost for US citizens of $871bn, of which $380bn belongs to medical expenses.[[1]](#footnote-1) Avoiding car accidents might be almost impossible, but understanding and reducing severity risk can have benefits in terms of health and less economic impact of governments, health institutions and insurance companies.

**1.2. Problem**

Data can contribute to understanding the drivers of severity of car accidents and help identify patterns in order to develop preventive measures. Property damage vs physical injuries have different impacts in diverse fields, such as legal treatment, driver health and economic compensation.

Bear in mind however, fatalities are out of the scope of this study.

**1.3. Interest**

Several stakeholders can be interested in such an analysis. Firstly, administrations would be able do city planning and distribute resources more efficiently depending on the areas of highest accident severity probability.

Secondly, data insights would help car manufacturers develop new technologies and components to minimize risk factors and improve the safety of the vehicle.

Moreover, insurance companies would benefit from a model that differentiates between personal and property damage. Having a better understanding of each type will help on risk assessment, policy amount and client targeting.

Finally, drivers themselves would be able to take precautions by understanding possible risk factors and plan their travels in a way to maximize their safety.

**2. Data Description and Cleaning**

The data I used is the CSV file provided by IBM Capstone in Coursera. This database, called “Collisions” and published by SDOT Traffic Management Division, contains a total 37 attributes related to severity of accidents (our prediction target), incident location, road state and number of agents involved, among others. Data ranges from 2004 to 2019 and includes 194,673 samples.

**2.1. Feature Selection**

However, many of these attributes were redundant or couldn’t be included in the model as they were consequences of the accident, rather than a possible driver:

|  |  |  |
| --- | --- | --- |
| Reason for dropping | Redundant index referring to another Feature | Consequence of the incident, cannot be input to predict severity |
| Dropped Features | COLDETKEY  INTKEY  LOCATION  SDOT\_COLCODE  SDOTCOLNUM  ST\_COLCODE  SEGLANEKEY  CROSSWALKKEY  EXCEPTRSNCODE  EXCEPTRSNDESC | INJURIES  FATALITIES |
| Kept Features | OBJECTID  ADDRTYPE  SEVERITY  SEVERITYDESC  COLLISIONTYPE  PERSONCOUNT  PEDCOUNT  PEDCYLCOUNT  VEHCOUNT  INCDATE | UNDERINFL  WEATHER  ROADCOND  LIGHTCOND  SPEEDING  HITPARKEDCAR  INCDTTM  JUNCTIONTYPE  SDOT\_COLDESC  INATTENTIONIND |

For more info on the metadata, please see Annex 1.

Note: in this case severity was restricted to “property damage” and “physical injury”.

**2.2. Data Cleaning**

- NaN: Deleted all NaN rows, since dataset was large, and it was considered preferable to keep quality data.

- Binarization: Transformed Yes/No cells into binary numbers, as well as severity.

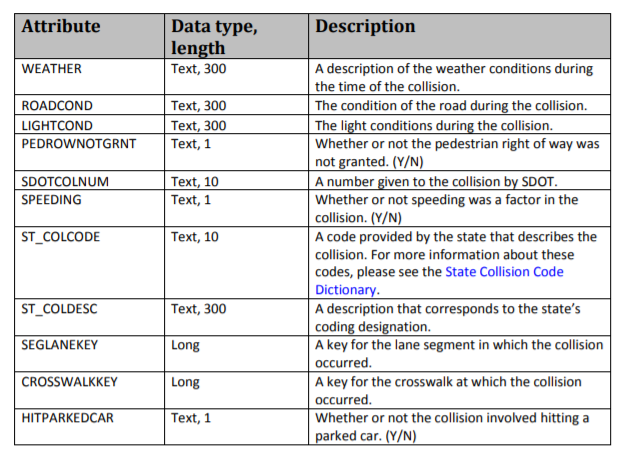
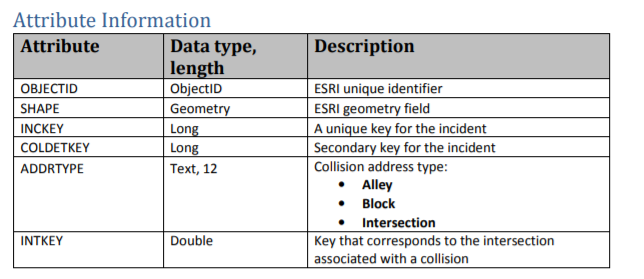
- Pearson correlation: after previous Feature selection, no other features had a large enough index to prove relevant redundancies.

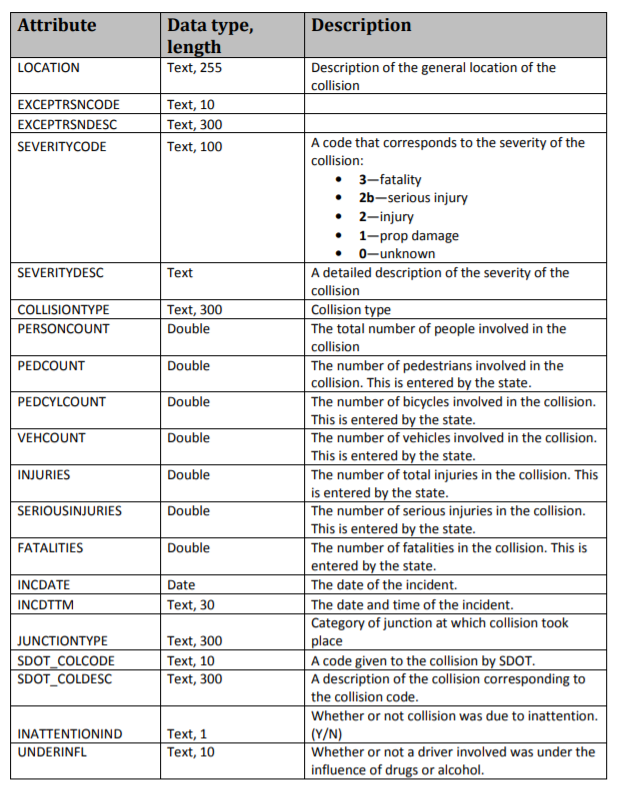
**2.3. How will the data solve the problem?**

With this revised dataset, I will study which are the most suitable features to predict the target variable: SEVERITY. Independent variables include information about the road’s state, location of the incident, amount of vehicles/people involved and other factors as well as driver factors such as speeding or presence of alcohol.

Through the incoming study, I will create a model that will explain which of the previous 20 variables are the most determinant factors regarding accident severity. By determining these factors, it will be possible for stakeholders to develop preventive measures, ensuring fast action from health institutions and allocating resources in a more efficient manner.

**ANNEX 1**





1. Association for Safe International Road Travel <https://www.asirt.org/safe-travel/road-safety-facts/#:~:text=Annual%20United%20States%20Road%20Crash,enough%20to%20require%20medical%20attention.> [↑](#footnote-ref-1)