mohamed osama

Capstone Project

IBM Coursera Data Science

PREDICTING THE SEVERITY OF ROAD VEHICLE ACCIDENTS

I Table of content

1 Business Understanding

1.1 Background

In recent years, the number of vehicles in operation and the number of drivers holding a valid driving license has increased continuously worldwide. One of the busiest countries in terms of road traffic are the United States of America with almost 280 million vehicles in operation. Alongside the increase of traffic, the number of road vehicle accidents has increased as well. Looking at the statistics for the United States, in 2018 there have been around 6 million car accidents, with a total of 12 million vehicles involved**. [1]** The described situation of the road vehicle accidents results in different consequences. Surely, the financial influence is immense. In a worldwide average, road vehicle accidents result in a cost of approx. 3% of the gross domestic product. Furthermore, physical integrity, not only of the drivers but also of pedestrians is in danger. So, to speak, almost 3 million people are injured every year in car accidents just in the US alone. In 2018, 36.560 people died in car accidents, which results in a fatality rate of 11.18 deaths per 100.000 capita.**[2]** In order to reduce the number of fatalities in car accidents, it is of highest important, that paramedics arrive as soon as possible and with the right equipment, the right number of ambulances and an appropriate rescue team size. To do so, it is necessary to evaluate the severity of the occurred accident as soon and as accurate as possible. A potential situation, in which the paramedics arrive at the accident scene and realize that further support is required might result in critical delays regarding the treatment and transportation of the accident victims.

1.2 Goal and objectives

To prevent the described situation and allow a quick evaluation of the accident severity, it is desired to develop an algorithm which can predict the severity of an accident by external environmental input factors, that can be observed right away from whoever is calling the emergency number from the accident scene. This algorithm would help to reduce the fatality rate from accidents by ensuring a quick and appropriate arrival and treatment from paramedics and would be of a huge help for **rescue services** all over the country. In turn, this would also lead to financial benefits regarding several stakeholders as cost factors such as hospital stays, life insurance payments etc. would be reduced. Therefore, the implementation of the algorithm would be also advantageous for any **(health) insurance firm**.

In order to achieve the defined goal, following objectives have been stated:

- Obtain a data set which includes the severity of road vehicle accidents, external environmental factors, and further information
- Develop a supervised machine learning model based on the obtained and cleaned dataset which requires as few features as possible to predict the severity of an accident
- Ensure that all features of the model can be visually/directly obtained by whoever is calling the emergency number

2 Data Understanding

2.1 Data source

The dataset used in this project was created by SDOT Traffic Management Division, Traffic Records Group, and contains all recorded collisions in the city of Seattle from 2004 to present. The dataset is updated weekly. The dataset used in this project (as of 28.08.2020) contains 194673 datapoints with a total of 38 columns. As every accident (data point) is labelled with a severity code which indicates the severity of the accident it is suitable for the previously described goal of this project. The dataset distinguishes between “property damage”, labelled as 1, and “injury” labelled a 2. Included in the 38 columns are several attributes which fulfil the defined condition of being “external environmental”. Some of these are light condition, road condition, date, number of vehicles involved etc. The dataset can be downloaded here and the metadata can be downloaded here.

2.2 Feature selection

Taking a closer look into the dataset, it can be seen that many of the columns contain interorganizational codes which are not relevant for the case of this study. Therefore, these columns are dropped (e.g. Incident key, report number etc.). Furthermore, there are redundant and irrelevant columns such as “Severitycode1” which contains the same information as the column “Severitycode”. Additionally, there are several attributes which are interesting but cannot be used for the described use case as they cant be observed right away and visually. An example for this is the column “Underinf” which describes whether or not a driver involved in the accident was under the influence of drugs. Another example is the column “speeding”, which describes whether or not one of the drivers was speeding up. Even though these attributes are interesting to understand the reason of the accident, they cant be observed before a more thorough investigation has taken place. Therefore, these columns are also dropped. As it is desired to implement the model nationwide, the location of the observed accidents is not taking into consideration, as a bias due to local circumstances should be prevented. The location column (displayed as “X” and “Y”) is also dropped. The features which have been chosen for training the model are:

- PERSONCOUNT → Number of person involved in the accident
- VEHCOUNT → Number of vehicles involved in the accident
- DATE → Date of the accident
- JUNCTIONTYPE → Intersection, Mid-Block etc.
- WEATHER → Weather condition during accident
- ROADCOND → Road condition during the accident
- LIGHTCOND → Light condition during the accident

All of these attributes can be observed easily and by anyone, which makes them suitable for use in the to be trained model.

Before continuing with cleaning and pre-processing the data, it has been checked whether the date has an influence on the severity of the accidents. To do so, first the date column has been changed to a “datetime” type. Afterwards, the interesting information has been extracted, which is the day of the week [df.dt.dayofweek]. Accordingly, a column was obtained containing numbers ranging between 0 and 5 which indicate the day of the week. After splitting the dataset by the severity code, two histograms have been plotted to observe whether the severity of accidents depend on the weekday (compare figure 1).

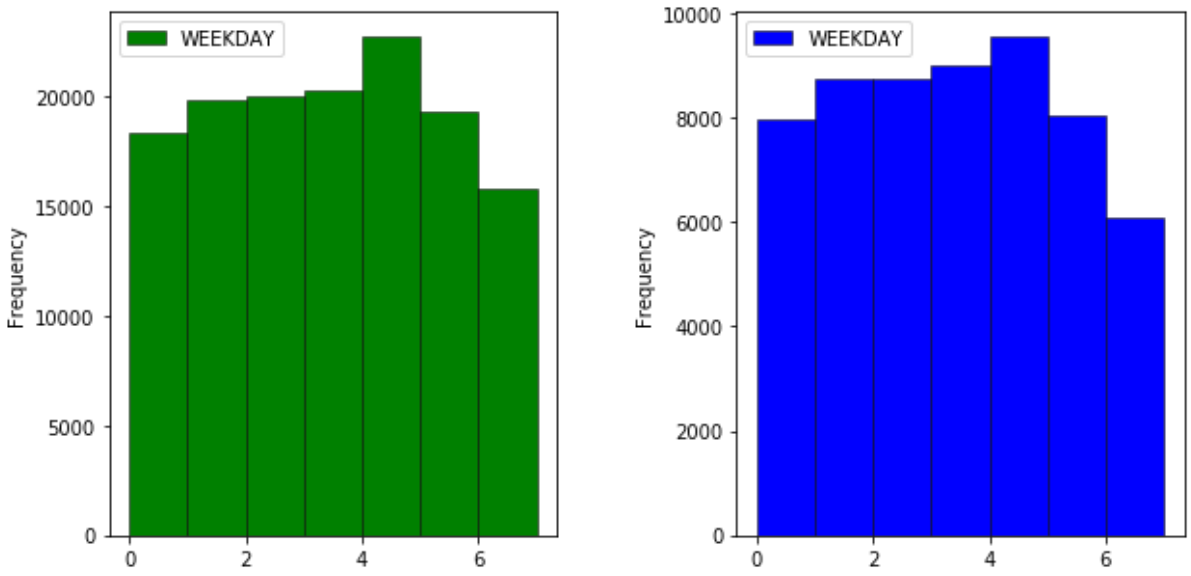

Figure 1: Frequency weekday for both severity codes

As can be seen in Figure 1, there is no relevant difference between the data set with severity=1 and severity=2. It appears that even though the day of the week seems to have an influence on the frequency of accidents, the day of the week doesn’t have a significant influence on the severity of the accident, therefore the date column has been dropped.

	SEVERITYCODE	PERSONCOUNT	VEHCOUNT	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND
0	2	2	2	At Intersection (intersection related)	Overcast	Bad_Conditions	Daylight
1	1	2	2	Mid-Block (not related to intersection)	precipitation	Bad_Conditions	Dark - Street Lights On
2	1	4	3	Mid-Block (not related to intersection)	Overcast	Good_Conditions	Daylight
3	1	3	3	Mid-Block (not related to intersection)	Clear	Good_Conditions	Daylight
4	2	2	2	At Intersection (intersection related)	precipitation	Bad_Conditions	Daylight

Figure 2: Data Frame after dropping irrelevant columns

As a result of the data understanding section, the shown data frame of Figure 2 (shows first five entries of the data frame) has been obtained. In the following, the chosen features will be further observed, the data will be cleaned, visualized, and the data set will be balanced in order to prevent a bias.