COURSERA FINAL CAPSTONE

PREDICTING CAR ACCIDENT SEVERITY

September 5, 2020

Kerman Sanjuan Malaxechevarria

Index

Chapter 1

Introduction

1.1 Background

Everyone knows the danger of the roads. Just in USA there are 6 million car accidents per year, involving more than one million deaths. Car accidents are considered one of the biggest causes of mortality in the world.

Moreover, the accidents have more consequences. Due to the fact that road traffic accidents lead to a large number of fatal incapacitating injuries, the consequences of these accidents are fundamentally reflected in the social sphere. This concerns job losses and the related financial hardships, loss of amenity and a fatal impact on the functioning of the whole family. We should not forget that the psychological impact of the consequences of road traffic accidents do not only to affect the direct participants, but also their families. The costs associated with road traffic accidents are shouldered by the whole society.

Accident prediction is important for optimizing public transportation, enabling safer routes, and cost-effectively improving the transportation infrastructure, all in order to make the roads safer. Given its significance, accident analysis and prediction has been a topic of research in the past few decades.

After taking over all these facts, it would be interesting to be able to predict the dangerousness of a road/place in case of accident, simply by knowing the characteristics of it, right?

1.2 Problem

With the obtained data, project aims to predict the severity of an accident, depending on the features of it. For example, the numbers of cars which are involved, speed, road condition, weather... etc

1.3 INTEREST

The main audience of this report could be any organization or government, obviously the ability to predict the consequence of an accident is something that generates a lot of interest. The applications of this predictions could be used for example, for real-time accident prediction, studying accident hot-spot locations, casualty analysis and extracting cause and effect rules to predict accidents, or studying the impact of precipitation or other environmental stimuli on accident occurrence.

Chapter 2

Data acquisition and cleaning

2.1 Acquisition of data

All these data is obtained from Cornell University, specifically from Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath from "A Countrywide Traffic Accident Dataset.", and "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights."

2.2 About the data

First, I decided to use the data which Coursera give it to me, but the lack of variety of the data when talking about the severity of the accident force myself to choose another data source. The obtained data contains information about traffic accidents, covering the 49 states of United States of America. The data has been continuously obtained since 2016, so we're working with actual data. Currently, there 3.5 million rows (we will use just 175k), each row contains information about a single accident. With all these data, we should be able to create a prediction model and discover some interesting information about this kind of events.

For explaining the data, I will use the resource that this data set provides. There are 45 different data features, which are distributed as follows

Type	Description
Amenity	Refers to particular places such as restaurant, library, college, bar, etc.
Bump	Refers to speed bump or hump to reduce the speed.
Crossing	Refers to any crossing across roads for pedestrians, cyclists, etc.
Give-way	A sign on road which shows priority of passing.
Junction	Refers to any highway ramp, exit, or entrance.
No-exit	Indicates there is no possibility to travel further by any transport mode along a formal path or route.
Railway	Indicates the presence of railways.
Roundabout	Refers to a circular road junction.
Station	Refers to public transportation station (bus, metro, etc.).
Stop	Refers to stop sign.
Traffic Calming	Refers to any means for slowing down traffic speed.
Traffic Signal	Refers to traffic signal on intersections.
Turning Loop	Indicates a widened area of a highway with a non-traversable island for turning around.

As we see, there's a POI table, where all Points Of Interest are defined, these are the main descriptions:

Total Attributes	45
Traffic Attributes (10)	id, source, TMC [23], severity, start_time, end_time, start_point, end_point, distance, and description
Address Attributes (8)	number, street, side (left/right), city, county, state, zip-code, country
Weather Attributes (10)	time, temperature, wind_chill, humidity, pressure, visibility, wind_direction, wind_speed, precipitation, and condition (e.g., rain, snow, etc.)
POI Attributes (13)	All cases in Table 1
Period-of-Day (4)	Sunrise/Sunset, Civil Twilight, Nautical Twilight, and Astronomical Twilight
Total Accidents	2,243,939
# MapQuest Accidents	1,702,565 (75.9%)
# Bing Accidents	516,762 (23%)
# Reported by Both	24,612 (1.1%)
Top States	California (485K), Texas (238K), Florida (177K), North Carolina (109K), New York (106K)

Most of these data is categorical, so it will be drop-out the table for a correct model developing. By the way, there some no relevant features which will be dropped out too.

2.3 DATA CLEANING

Now we have talked about the data, is time to start working with it, first of all, we have to clean the data, but, what it means? We will see it

2.3.1 Remove unnecessary information

So, as we see before, there's data that we don't need, so we will get rid of it. Which are examples of irrelevant data?

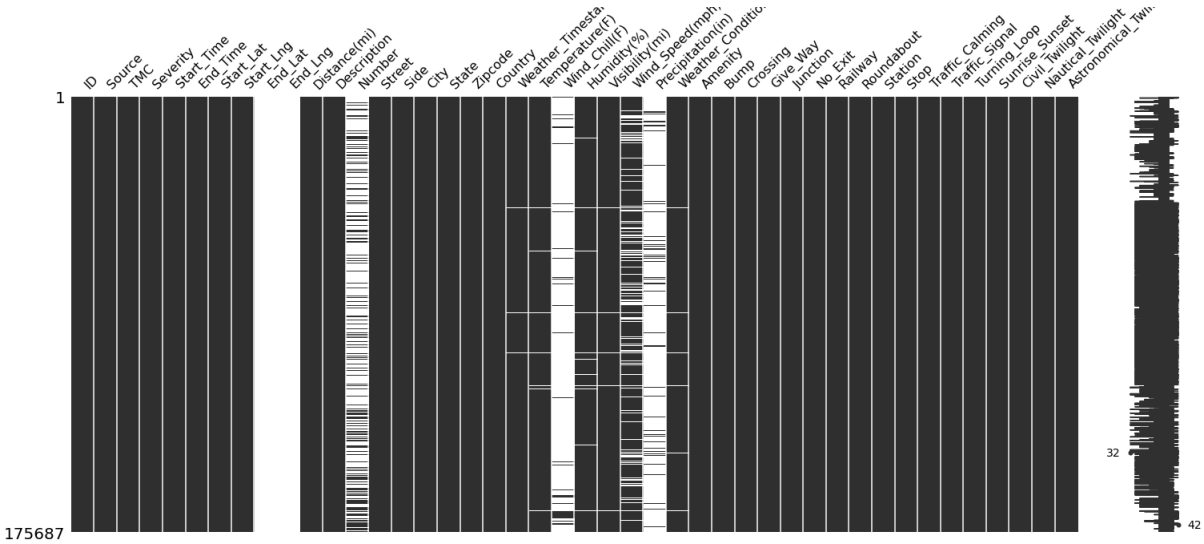
- Pressure:
- All address attributes: number, street, side (left/right), city, county, state, zip-code, country
- Wind Direction: No relevant
- Timezone: No relevant
- Airport Code: No relevant
- Nautical Twilight: Not relevant
- Astronomical Twilight: Not relevant

The Twilight we will use will be the Civil one, which is the period when enough natural light remains that artificial light is not needed. As you see, all of this information has no importance when talking about car accidents, so we drooped it. There's more data, but know we will drop the data with a bunch of missing values.

2.3.2 Delete the lack of data

In this type of data-set, is normal to have some missing values, but in some cases, can be a problem. In the cases where the missing values are above the 15% of the samples, we will drop it. So, lets see how much missing values we have on a visual way.

Figure 2.1: Missing values before cleaning.



As we see, there some features with more than 15% , so we will drop it, these are the features we will drop:

- Start Time
- End Time
- Latitude
- Longitude
- Description
- Weather Timestamp

- Wind chill
- Precipitation
- End latitude
- End longittude

Now, We've gone from 47 to 25 features, but there are still features to drop.

2.3.3 Delete some categorical values

A categorical variable (sometimes called a nominal variable) is one that has two or more categories, but there is no intrinsic ordering to the categories. For example, gender is a categorical variable having two categories (male and female) and there is no intrinsic ordering to the categories. This features are hard to compute, so we're going to delete the ones with more than 5 different categories.

The remaining categorical features are the following ones: ID, Source,Weather Condition, Sunrise Sunset, and Civil Twilight. In the case of ID, will be used as Index, so we can't remove it, but it will not be relevant. By the way, the rest of values can be changed to discrete values, for example. Except the Source one, which can't be a discrete value, so we will delete it. Weather have 50 different types, but is important when we talk about car accidents, so using Hot-Enconding will solve the problem.

Figure 2.2: Hot encoding example.

SAFETY-LEVEL (TEXT)	SAFETY-LEVEL (None)	SAFETY-LEVEL (Low)	SAFETY-LEVEL (Medium)	SAFETY-LEVEL (High)	SAFETY-LEVEL (Very High)
None	1	0	0	0	0
Low	0	1	0	0	0
Medium	0	0	1	0	0
High	0	0	0	1	0
Very-High	0	0	0	0	1

2.4 DATA WRANGLING

After all this cleaning, we ended up having just 24 features out of 47, which is a pretty god selection. Now it's time to replace all the missing values and transforming categorical values to discrete ones.

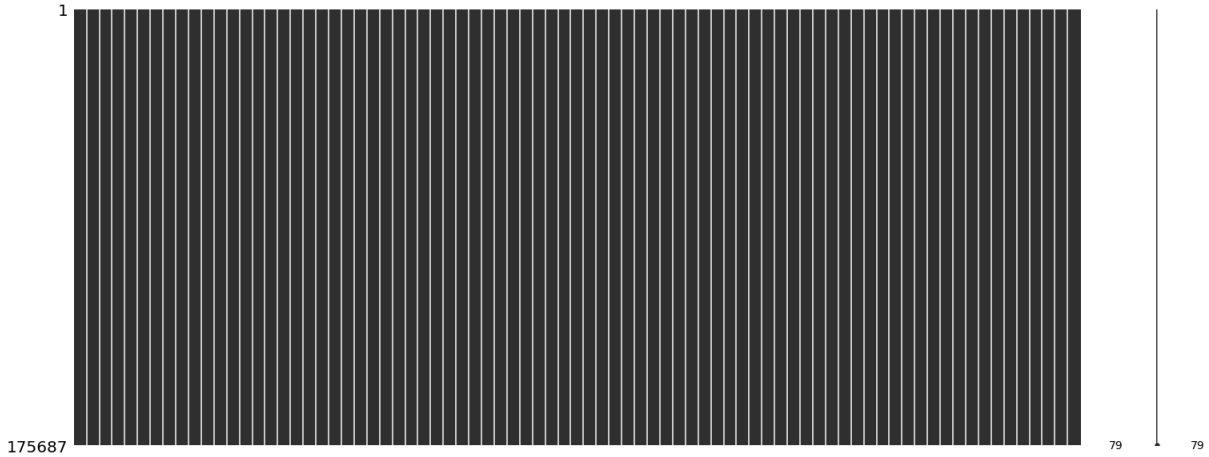
2.4.1 Replace missing values

First, we will replace the missing values with the mean of the feature, this is the best way to avoid having a negative impact on the data.

2.4.2 Transformation of categorical values

As we said, we will use hot-encoding technique, this will make an increase the number of features of the data set, but it's necessary to ensure the quality of our model (the weather is an important feature).

Figure 2.3: Missing values after cleaning and filling empty values.



2.5 FEATURE SELECTION

Now, we will see the relation between this features and severity.

Chapter 3

Exploratory Data Analysis (EDA)

Chapter 4

Model Development

Chapter 5

Conclusion

Chapter 6

Future Directions

Chapter 7

References