COURSERA FINAL CAPSTONE

PREDICTING THE CAR ACCIDENT SEVERITY

September 4, 2020

Kerman Sanjuan

Index

Chapter 1

Introduction

1.1 Background

Everyone knows the danger of the roads. Just in USA there more than 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 hotspot 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, the lack of variety of the data 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 is continuedly 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 dataset provides. There are 45 different data attributes, which are distributed in this way

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

Aktore lista eta pelikula lista izango dira erabiliko ditugun datu egitura nagusiak. Izan ere, hauek gabe ia ezinezkoa izango litzateke pelikula eta aktore guztiak batera kudeatzea.

2.3.1 Pelikula lista

Pelikula bakoitza aldi bakarreann soilik agertzen denez fitxategian, ez dugu pelikula hori jadanik agertu den ala ez konprobatu behar. Hori dela eta, pelikulez osatutako Arraylist bat erabiltzea aukeratu dugu.

2.3.2 Aktore lista

Pelikulak ez bezala, aktore bakoitza askotan ager daiteke fitxagian. Aktore berri bat gehitu nahi dugun bakoitzean aktore lista osoa konprobatu beharko bagenu, errepikapenak saihesteko, gure algoritmoa kostu konputazional handia edukiko luke. Horregatik HashMap erabiltzea erabaki dugu.

Dena den, bi klase hauek pelikula eta aktore guztiak dituzten listak dira. Hauetaz gain beste bi Arraylist erabiliko ditugu:

-Pelikula bakoitzak bere aktoreen lista edukiko du.

-Aktore bakoitzak bere pelikulen lista edukiko du.

2.4 FEATURE SELECTION

2.5 DATA WRANGLING

Chapter 3

Exploratory Data Analysis (EDA)

Chapter 4

Model Development

Chapter 5

Conclusion

Chapter 6

Future Directions

Chapter 7

References