Car Accident Severity Capstone Projet (Week 1)

Table of contents

Introduction: Description of the Problem

In this project a model to predict the **probability** of having a **severe car accident** taking into account the conditions of the road will be build up. In order to do so, the city of **Seattle** will be used as an **example**. A database containing information about car accidents in Seattle will be used to feed the model.

The **stakeholders** of this project will be **drivers** using their vehicles in Seattle and its surroundings. Of course, this project may also interest app developers which may use this example an extrapolate it to other cities and/or states.

The main idea behind this project is quite simple: "***forcasting car accidents is the best way to avoid them***".

Data

The data that will be used in this project comes from the ***Traffic Management Division of Seattle***. A raw database in a form of a CSV file will be downloaded from: "https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv". This **database contains** precious information about all registered **car accidents in Seattle from 2004 to Present**.

Understanding the Data

The raw database comes as a **CSV file** containing **194673** entries (car accidents in Seattle) and **38** attributes with several redundant data (columns) that will not be used to build the model. The target of our model will be the "**severity**" of the possible car accident.

The problem has been already simplified; the raw database has only **two different values to describe the severity** (see **Figure 1**):

1) **Property Damage Only Collision** 136485 accidents representing *ca*. 70% of the data.

2) **Injury Collision** 58188 accidents representing *ca*. 30% of the data.

SEVERITYDESC	
Property Damage Only Collision	136485
Injury Collision	58188

SEVERITYCODE	
1	136485
2	58188

Figure 1: Type of car accident severity describe in the Seattle car accident database.

The dataset includes the following fields:

SEVERITYCODE	int64	PERSONCOUNT	int64		
X	float64	PEDCOUNT	int64		
Y	float64	PEDCYLCOUNT	int64		
OBJECTID	int64	VEHCOUNT	int64		
INCKEY	int64	INCDATE	object		
COLDETKEY	int64	INCDTTM	object	ST_COLCODE	object
REPORTNO	object	JUNCTIONTYPE	object	ST_COLDESC	object
STATUS	object	SDOT_COLCODE	int64	SEGLANEKEY	int64
ADDRTYPE	object	SDOT_COLDESC	object	CROSSWALKKEY	int64
INTKEY	float64	INATTENTIONIND	object	HITPARKEDCAR	object
LOCATION	object	UNDERINFL	object		
EXCEPTRSNCODE	object	WEATHER	object		
EXCEPTRSNDESC	object	ROADCOND	object		
SEVERITYCODE.1	int64	LIGHTCOND	object		
SEVERITYDESC	object	PEDROWNOTGRNT	object		
COLLISIONTYPE	object	SDOTCOLNUM	float64		
		SPEEDING	object		

Figure 2: Fields of the Seattle car accident database.

Only few of the database attributes (columns) are relevant for the prediction of the probability of having an accident and the severity of it, namely:

- the condition of the road during the accident (**ROADCOND** attribute)

- the weather conditions during the time of the accident (**WEATHER** attribute)
- the light conditions during the accident (**LIGHTCOND** attribute)
- category of junction at which the accident took place (**ADDRTYPE** attribute)
- longitude (**X** attribute)
- latitude (**Y** latitude)

As a starting point, the attributes above will be used as main independent variables to build up our predicting model. Then, other parameters such as the **time of the accident** (**INCDTTM**) will be included in the model in order to explore possible improvements.

During the dataset preparation and cleaning, in order to remove possible biases in our model some of the entries containing positive (true) parameters such as **inattention** (**INATTENTIONIND** attribute), **speeding** (**SPEEDING** attribute) or **drug influence** (**UNDERINFL** attribute) will be removed. The effect of removing or filling data for the entries missing important attributes such as ROADCOND (with 5012 empty entries ca. 3% of the dataset) will be analyzed when the model will be built.

Evolution of the Car Accidents in Time

One of the first things that we need to verify is the evolution of accidents in time. In principle, over the years the quality of the streets, roads and even the cars themself may have improved, thus, driving might have become safer and less accident may have happended in the last years. We can verify this hypothesis looking at the trend of car accidents in the last few years with a horizontal bar chart.

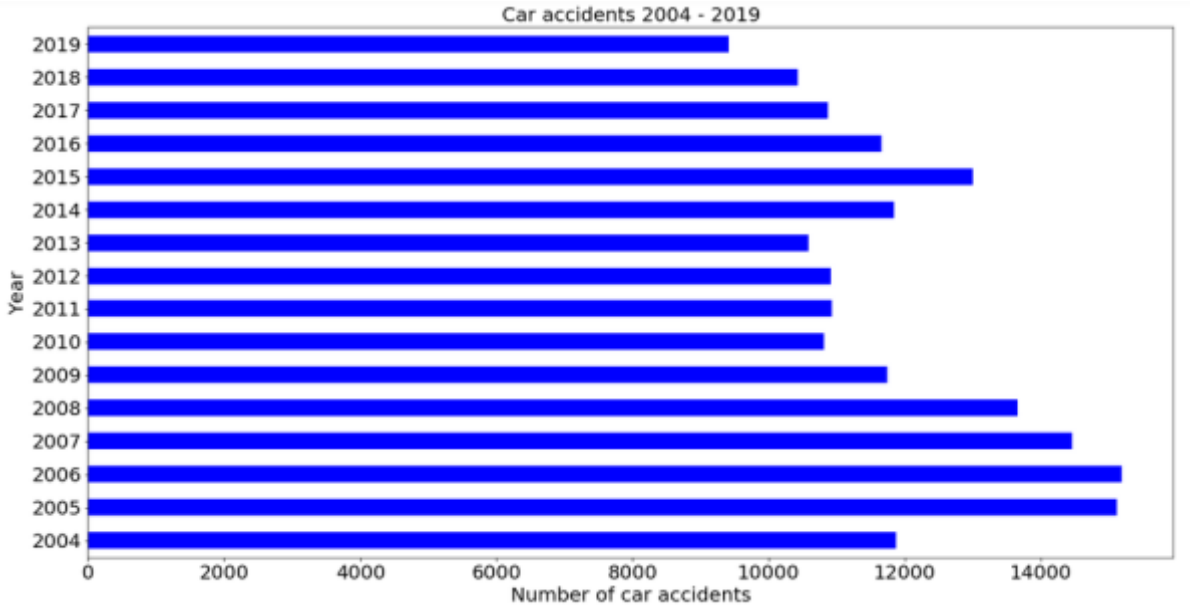


Figure 3: Evolution of car accidents in Seattle from 2004 until 2019.

As it can be observed in the horizontal bar chart above, the car accidents have decreased in the last decade. However, this decreasing trend might not be significant enough to discard the entries from the 2000's decade; thus, all the years will be taken into account in order to construct the predicting severity model.

Density of Car Accidents by Districts

Another important trend that can be explored is the density of car accidents in each district. In order to do so, the first thing that needs to be done is a transformation of latitude and longitude coordinates into zipcodes. To transform the latitude and longitude coordinates the GeoPy client (https://github.com/geopy/geopy) will be used.

At this point only the X and Y attributes will be used. However, some precautions need to be made, we need to verify if there are some instances with blank data and in this case we need to delete them in order to make the transforming function ("get_zipcode") work properly. Indeed, after verification we find that **5334** entries have no latitude and longitude coordinates. Thus we need to remove them from the dataset that will be used to construct the model.

Only a small part of the dataset (**the first three hundred entries**) **will be used** in order **to analyze the density of car accidents** in each districts (represented by the zipcodes), since the get_zipcode function needs to make calls to external APIs for each pair of coordinates. Thus, in order to transform the whole dataset into zipcodes the required time would be several days.

After examining the results of the "get_zipcodes" function, only 4 entries had to be removed since the zipcodes found: 98106-1499, 98133-6124 and 98109-5210 had more than 5 digits. The 296 remaining instances of car accidents in Seattle are represented in **Figure 4**:

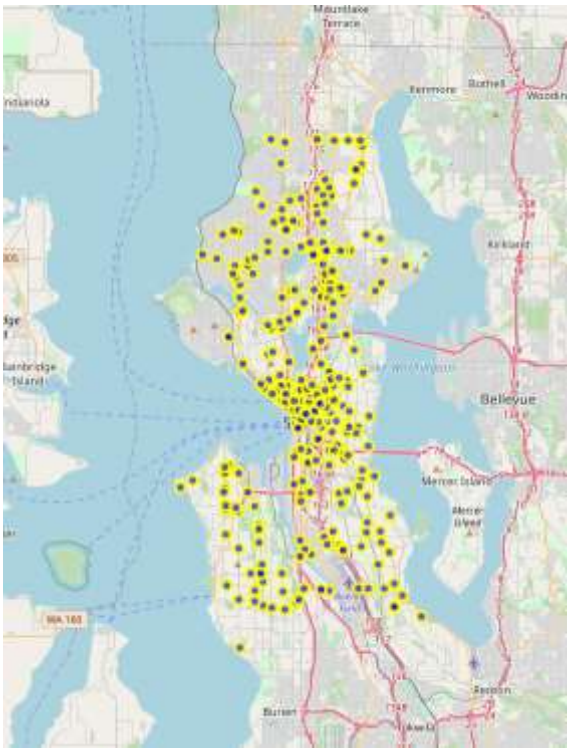
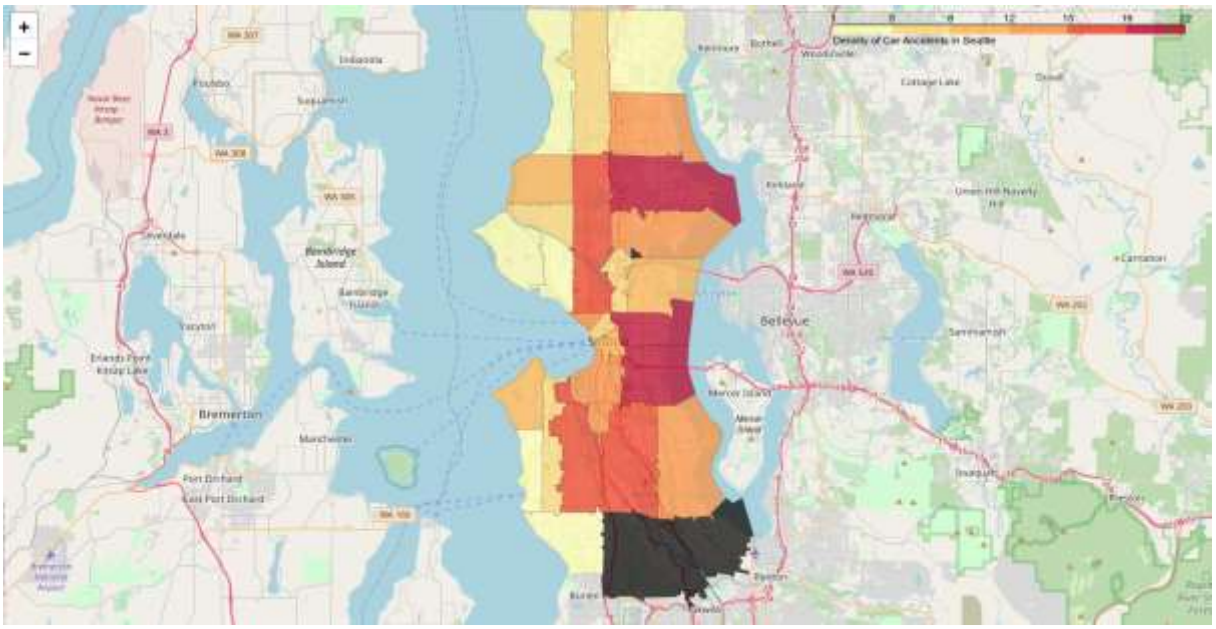


Figure 4: Sample of 296 car accidents in Seattle (from 2004 to 2020).

In order to represent the 296 car accidents in the map, the Folium Python library has been used.

When representing the density of car accidents by “districts” (**Figure 5**) it can be seen that most of the accidents have happened in the east-center part of the city and one of the north-east districts.



Nevertheless, comparing both images (**Figures 4** and **5**), one can clearly see that the zipcodes might not be the best way of representing the car accident **densities** in the maps, thus, a clustering method will be investigated.

Cleaning Data

Now the data needs to be cleaned before been used to build up our model. All the entries containing “Unknown”, “Other” and/or empty will be erased from the dataset that will be used to build the model. Only the attributes that are interesting for the study (**ROADCOND**, **WEATHER**, **LIGHTCOND**, **ADDRTYPE**) will be checked and treated.