

# Applied Data Science Capstone Final Assignment

## Car Accident Severity

Data Scientist: José Victor Tobias Romero

Date: September 2020

## Table of Contents

<b>Car Accident Severity (Business Understanding)</b>	<b>3</b>
<b>Data Understanding</b>	<b>4</b>
Data Summary:	5
Missing data	5
Data Visualization	10
<b>Data Transformation (For details, checkout the Notebook)</b>	<b>11</b>

## Car Accident Severity (Business Understanding)

Car accidents are very common in every part of the world. Car accidents can cause injuries, disabilities, property damage and can result in death, so, trying to minimize or get the needed medical assistant in the proper way is very important.

There are a lot of factors that can contribute to the severity of the accident like the speeding, the road, if there are multiple cars, alcohol, drugs, street racing, lack of maintenance, unfamiliar with the road, lack of visibility, distractions, traffic safety culture, etc.

The National Safety Council of USA says that in 2019, an estimated 38,800 people lost their lives because of car crashes. In 2018, an estimated 39,404.

In my country, Guatemala, the government estimates more than 10,000 car accidents in a year and more than 1,700 people lost their lives.

In 2013, more than 54 million people worldwide sustained injuries from car accidents and an estimated of 1.4 million lost their lives. The effects of Car accidents can be physical and psychological, so car accidents are a serious problem and we have to minimize its effects.

Because of the car accidents is a very common type of event, it is necessary to try to estimate the severity of the result of an accident to know how to take actions to reduce the severity and to give the proper medical care in the proper time and to control the traffic to reduce the accumulation of vehiculos (like long lines of cars or the car barely moving).

If we can predict the possibility of a car accident, we can warn us and take actions to avoid them. The possibilities can be predicted with factors like the weather, road, etc.

In this document, we are going to analyse a data set with data of car accidents with a variety of labels and observations and the severity, that represents the fatality of an accident, of the car accidents of every case.

# Data Understanding

For this project, we are going to use the Dataset “Example Dataset” and we can find it in the next link: [Dataset](#)

We can find the metadata for the dataset in the next link: [Metadata](#)

The label that we want to predict is the label “severity” that describes the fatality of a car accident. In total there are a total of 37 attributes of features but not all of them are useful, we are going to analyze them to find the proper features to utilize.

The data will be used to train a model and the model will give predictions of the severity of the car accidents. For that purpose, we are going to normalize the features, apply features engineering and we are going to split the data in groups, one for train and one for testing the trained model. The proportion of each group will be 80% for training and 20% for testing. Because we are going to predict a categorizing variable, we are going to train various categorizing models and we are going to choose the one with best accuracy.

Just inspecting the data we can see that there are some features that do not add value to the model like the description, ObjectID, IntKey, increment of identifiers of the incident, etc. The ignore features with the initial analysis are the following:

- X
- Y
- OBJECTID
- INCKEY
- COLDETKEY
- REPORTNO
- STATUS
- INTKEY
- EXCEPTRSNDESC
- SEVERITYDESC
- INCDATE (Because, there is a timestamp of the incidente with INCDTTM)
- SDOT\_COLDESC
- SDOTCOLNUM
- ST\_COLDESC

After removing the features that do not add value to the model. The goal of the project is to predict the Car Accident Severity using the remaining features. Now we can analyse the remaining features:

## Data Summary:

	SEVERITYCODE	SEVERITYCODE.1	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	SDOT_COLCODE	SEGLANEKEY	CROSSWALKKEY
count	194673.000000	194673.000000	194673.000000	194673.000000	194673.000000	194673.000000	194673.000000	194673.000000	1.946730e+05
mean	1.298901	1.298901	2.444427	0.037139	0.028391	1.920780	13.867768	269.401114	9.782452e+03
std	0.457778	0.457778	1.345929	0.198150	0.167413	0.631047	6.868755	3315.776055	7.226926e+04
min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00
25%	1.000000	1.000000	2.000000	0.000000	0.000000	2.000000	11.000000	0.000000	0.000000e+00
50%	1.000000	1.000000	2.000000	0.000000	0.000000	2.000000	13.000000	0.000000	0.000000e+00
75%	2.000000	2.000000	3.000000	0.000000	0.000000	2.000000	14.000000	0.000000	0.000000e+00
max	2.000000	2.000000	81.000000	6.000000	2.000000	12.000000	69.000000	525241.000000	5.239700e+06

## Missing data

The missing data is as follows (bold columns have missing data):

SEVERITYCODE

False 194673

Name: SEVERITYCODE, dtype: int64

ADDRTYPE

False 192747

True 1926

Name: ADDRTYPE, dtype: int64

LOCATION

False 191996

True 2677

Name: LOCATION, dtype: int64

EXCEPTRSNCODE

True 109862 ← have a lot of missing data

False 84811

Name: EXCEPTRSNCODE, dtype: int64

SEVERITYCODE.1

False 194673

Name: SEVERITYCODE.1, dtype: int64

COLLISIONTYPE

False 189769

True 4904

Name: COLLISIONTYPE, dtype: int64

**PERSONCOUNT**

False 194673

Name: PERSONCOUNT, dtype: int64

**PEDCOUNT**

False 194673

Name: PEDCOUNT, dtype: int64

**PEDCYLCOUNT**

False 194673

Name: PEDCYLCOUNT, dtype: int64

**VEHCOUNT**

False 194673

Name: VEHCOUNT, dtype: int64

**INCDTTM**

False 194673

Name: INCDTTM, dtype: int64

**JUNCTIONTYPE****False 188344****True 6329****Name: JUNCTIONTYPE, dtype: int64****SDOT\_COLCODE**

False 194673

Name: SDOT\_COLCODE, dtype: int64

**INATTENTIONIND****True 164868 ← have a lot of missing data****False 29805****Name: INATTENTIONIND, dtype: int64****UNDERINFL****False 189789****True 4884****Name: UNDERINFL, dtype: int64****WEATHER****False 189592****True 5081****Name: WEATHER, dtype: int64**

**ROADCOND****False 189661****True 5012****Name: ROADCOND, dtype: int64****LIGHTCOND****False 189503****True 5170****Name: LIGHTCOND, dtype: int64****PEDROWNOTGRNT****True 190006 ← have a lot of missing data****False 4667****Name: PEDROWNOTGRNT, dtype: int64****SPEEDING****True 185340 ← have a lot of missing data****False 9333****Name: SPEEDING, dtype: int64****ST\_COLCODE****False 194655****True 18****Name: ST\_COLCODE, dtype: int64****SEGLANEKEY****False 194673****Name: SEGLANEKEY, dtype: int64****CROSSWALKKEY****False 194673****Name: CROSSWALKKEY, dtype: int64****HITPARKEDCAR****False 194673****Name: HITPARKEDCAR, dtype: int64**

Because **EXCEPTRSNCODE**, **INATTENTIONIND**, **PEDROWNOTGRNT** and **SPEEDING** has a lot of missing data and there's no description in the metadata file, we are going to drop the column.

Detailed information of the values of the features with missing data are as follow:

Block 126926  
 Intersection 65070  
 Alley 751  
 Name: ADDRTYPE, dtype: int64

BATTERY ST TUNNEL NB BETWEEN ALASKAN WY VI NB AND AURORA AVE N  
 276  
 BATTERY ST TUNNEL SB BETWEEN AURORA AVE N AND ALASKAN WY VI SB  
 271

...  
 41ST AVE SW BETWEEN SW 102ND ST AND SW 104TH ST 1 1  
 8TH AVE NE AND NE 90TH ST 1  
 40TH AVE S AND S WEBSTER ST 1  
 Name: LOCATION, Length: 24102, dtype: int64

Parked Car 47987  
 Angles 34674  
 Rear Ended 34090  
 Other 23703  
 Sideswipe 18609  
 Left Turn 13703  
 Pedestrian 6608  
 Cycles 5415  
 Right Turn 2956  
 Head On 2024  
 Name: COLLISIONTYPE, dtype: int64

Mid-Block (not related to intersection) 89800  
 At Intersection (intersection related) 62810  
 Mid-Block (but intersection related) 22790  
 Driveway Junction 10671  
 At Intersection (but not related to intersection) 2098  
 Ramp Junction 166  
**Unknown 9**  
 Name: JUNCTIONTYPE, dtype: int64

N 100274  
 O 80394  
 Y 5126  
 1 3995  
 Name: UNDERINFL, dtype: int64



Clear	111135
Raining	33145
Overcast	27714
Unknown	15091
Snowing	907
Other	832
Fog/Smog/Smoke	569
Sleet/Hail/Freezing Rain	113
Blowing Sand/Dirt	56
Severe Crosswind	25
Partly Cloudy	5

Name: WEATHER, dtype: int64

Dry	124510
Wet	47474
<b>Unknown</b>	<b>15078</b>
Ice	1209
Snow/Slush	1004
Other	132
Standing Water	115
Sand/Mud/Dirt	75
Oil	64

Name: ROADCOND, dtype: int64

Daylight	116137
Dark - Street Lights On	48507
Unknown	13473
Dusk	5902
Dawn	2502
Dark - No Street Lights	1537
Dark - Street Lights Off	1199
Other	235
Dark - Unknown Lighting	11

Name: LIGHTCOND, dtype: int64

32	27612
10	23427
14	16883
32	16809
10	11247
50	9089
14	8888
11	8636
28	6925
13	5363

```

4886
50  4465
...
4    20
7    18
54   1
87   1
60   1
Name: ST_COLCODE, Length: 115, dtype: int64

```

I have put in bold the data that we have to deal with for the next step.

Now, we need to check the current data types:

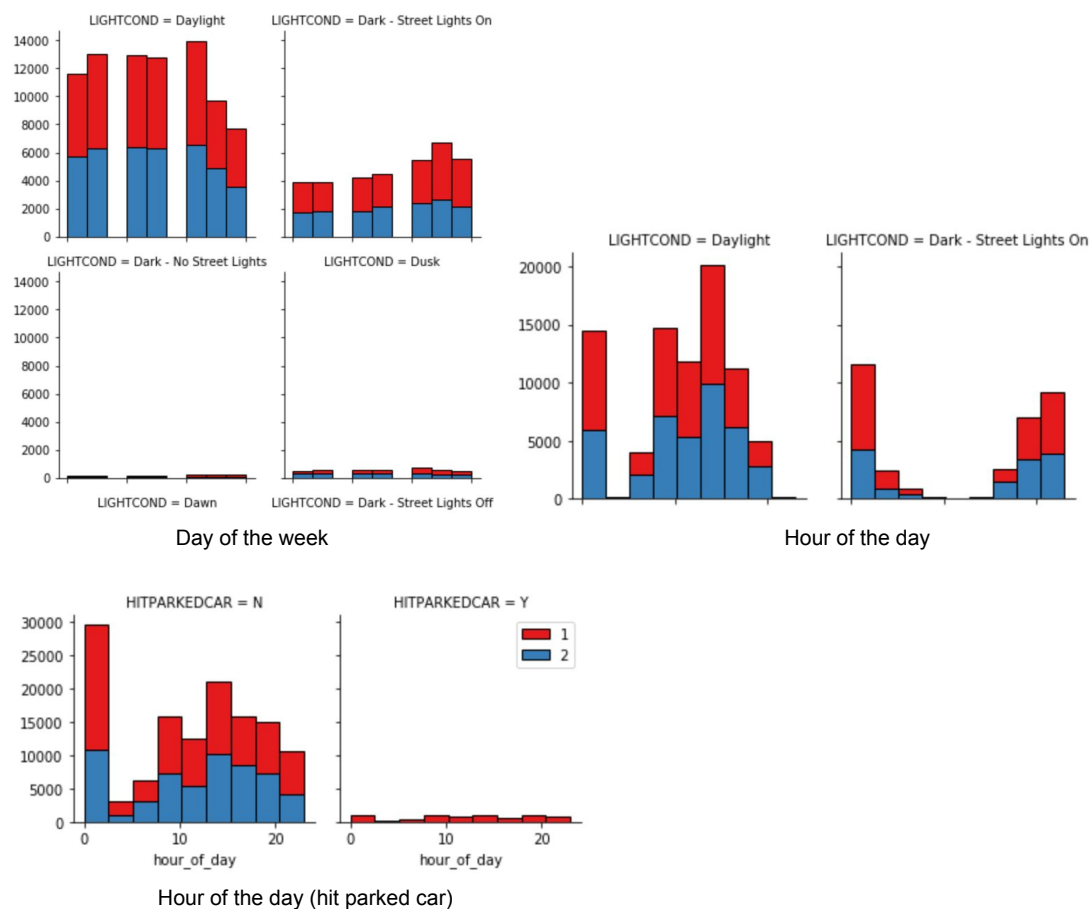
```

SEVERITYCODE    int64
ADDRTYPE        object
SEVERITYCODE.1  int64
COLLISIONTYPE   object
PERSONCOUNT    int64
PEDCOUNT       int64
PEDCYLCOUNT     int64
VEHCOUNT        int64
INCDTTM         object
JUNCTIONTYPE    object
SDOT_COLCODE    int64
UNDERINFL       object
WEATHER         object
ROADCOND        object
LIGHTCOND       object
ST_COLCODE      object
SEGLANEKEY      int64
CROSSWALKKEY    int64
HITPARKEDCAR    object

```

So, we need to change INCDTTM to datetime and visualize the data to find some util patrons. First we will inspect if there's some importance if the accident occurs in the weekend and then we are going to examine the incident in some hours of the day:

## Data Visualization



## Data Transformation (For details, checkout the Notebook)

First, we are going to deal with the unknown values changing them to NaN. In this case we need to change the features: JUNCTIONTYPE, ROADCOND, ST\_COLCODE and ST\_COLCODE has a blank space as value.

After analyzing the data, we can see that the location is not relevant (there are too many values and there is no direct influence over the severity) and for that case, we are going to drop it.

In most of the cases we have less than 5% of the rows with missing values (In statistical language, if the number of the cases is less than 5% of the sample, then the researcher can drop them), so we are going to drop the values that match that rule and deal with the other missing values. In this case we drop the values for COLLISIONTYPE.

We also need to drop the duplicated column SEVERITYCODE.1.

We need to change categorical values to numerical values, in this case for: UNDERINFL, CROSSWALKKEY and HITPARKEDCAR. Then we need to correct the type of the columns. All of the steps to clean, fill missing values and transformation can be found in the Notebook, they're not included here to try keep this document simple for the non technical people and reduce the complexity of the document.

Now, we need to transform the categorical values to binary variables with hot encoding techniques and in this case we are going to use the get\_dummies function of Pandas to encode the following features: **ADDRTYPE**, **COLLISIONTYPE**, **JUNCTIONTYPE**, **WEATHER**, **ROADCOND** and **LIGHTCOND**.

After all the data transformations, we have the final shape of (189769, 58) and the only thing missing is to normalize the data to reduce bias.

Check all the required procedures in the Notebook that is in the Github repository.