Research Report

Predicting Car Accident Severity Using Machine Learning Techniques

Created By: Hind Baageel

Sep 2020

1. Introduction

When you are travelling on a road, you are likely will be prone to road accidents if you are not careful. According to one interesting statics: traffic accidents countries more than 3% of their domestic products [1]. in this paper, we will study the effect of environment on the likelihood of getting a severe car accident. This paper starts by defining the characteristics of the chosen dataset. Then, this step will be followed by data cleaning step. Next, this study will develop an exploratory data analysis that will analyze collision dataset to derive valuable insights on features affects the most likelihood of getting a severe accident. Then, this paper will develop a predictive model that can predict the possibility of getting a severe car accident.

2. Data Description

In this paper, a dataset called Collisions—All Years, was chosen to help achieve the research goal. This dataset was gathered and organized by SDOT Traffic Management Division, Traffic Records Group. It includes a total of 37 features of the collision such as the collision date, number of involved persons, severity level and so on. Table 1 describes all the features of the current dataset.

Table 1: *List of Features in the dataset*

Feature Name	**Description**
OBJECTID	ESRI unique identifier
SHAPE	ESRI geometry field
INCKEY	A unique key for the incident
COLDETKEY	Secondary key for the incident
ADDRTYPE	Collision address type: Alley, Block, Intersection
INTKEY	Key that corresponds to the intersection associated with a collision
LOCATION	Description of the general location of the collision
EXCEPTRSNCODE	-
EXCEPTRSNDESC	-
SEVERITYCODE	A code that corresponds to the severity of the collision: 3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown
SEVERITYDESC	A detailed description of the severity of the collision
COLLISIONTYPE	Collision type
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state.

VEHCOUNT	The number of vehicles involved in the collision. This is entered by the state.
SERIOUSINJURIES	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	The number of fatalities in the collision. This is entered by the state.
INCDATE	The date of the incident.
INCDTTM	The date and time of the incident.
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	A code given to the collision by SDOT.
SDOT_COLDESC	A description of the collision corresponding to the collision code.
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	A number given to the collision by SDOT.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	A description that corresponds to the state's coding designation.
SEGLANEKEY	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)

First, this dataset was uploaded into a python environment for data exploration. The data type of each feature was checked. Some data types didn't match the required feature type. For example, INCDATE was saved as an object not a date. Thus, type casting was conducted to transform the datatypes into the required form.

Then, we analyzed the full description of the dataset as shown in Figure 1. It can be noticed that the maximum numbers of vehicles involved in a single collision are 11 vehicles and the maximum person which were involved in car collision count was 81.

In [99]:
```
df.describe()
```

Out[99]:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	INTKEY	SEVERITYCODE.1	PERSONCOUNT
count	194007.000000	188719.000000	188719.000000	194007.000000	194007.000000	194007.000000	64900.000000	194007.000000	194007.000000
mean	1.299865	-122.330518	47.619545	108640.304778	141273.862861	141481.481766	37560.973960	1.299865	2.444649
std	0.458200	0.029975	0.056160	62607.022410	86599.201441	86951.380156	51764.839238	0.458200	1.346964
min	1.000000	-122.419091	47.495573	1.000000	1001.000000	1001.000000	23807.000000	1.000000	0.000000
25%	1.000000	-122.348673	47.575956	54573.500000	70675.500000	70675.500000	28666.000000	1.000000	2.000000
50%	1.000000	-122.330219	47.615369	107164.000000	123611.000000	123611.000000	29973.000000	1.000000	2.000000
75%	2.000000	-122.311937	47.663667	162400.500000	203465.500000	203605.500000	33973.000000	2.000000	3.000000
max	2.000000	-122.238949	47.734142	219547.000000	331454.000000	332954.000000	757580.000000	2.000000	81.000000

Figure 1: *Dataset Description*

3. Methodology

Under this section, the used methods in the current work to develop a predictive machine capable of predicting a car accident are described.

3.1 Feature Engineering

After prime investigation of the dataset, it can be noticed that dataset requires some cleaning in order to produce efficient results. First, we started by analyzing the required features columns to see what is the needed columns to be kept and what needs to be removed when it doesn't provide sufficient important to the research goal.

The list of unnecessary features:

1- ST_COLDESC

2- SDOT_COLDESC: No need for a description, the SDOT code is enough

3- SEGLANEKEY

4- COLDETKEY: The primary, unique key is enough, no need for a secondary key.

5- EXCEPTRSNCODE: No sufficient information - mostly NaN

6- EXCEPTRSNDESC: No sufficient information - mostly NaN

7. SEVERITYCODE.1: Duplicated Column

8. SDOTCOLNUM: doesn't affect greatly the severity level

the rest of the features were included in the features data frame X.

3.2 Data Balancing

The dataset scanning showed the total number of each category according to severity. severe accidents reached a total of 13832 while mild collision reached a total of 33112. Thus, it can be concluded that mild collision is much higher than severe collision. That means that the data is unbalanced. This issue needs to be addressed to overcome biased classification decisions.

To balance the data, we used the ***IMBLEARN*** library to conduct under-sampling technique on the dataset. This technique work by taking the major group and reduce the number of samples until it reaches the other group**.** The final dataset count after under-sample:

Mild Collision	**58176**
Severe Collision	**58176**

3.3 Dealing with Missing Values

The dataset was further cleaned by examining the missing data. The two columns SPEEDING & PEDROWNOTGRNT contained NaN instead of No. For example, if the car was not speeding, the value of this column is NaN. Therefore, we replaced all the NaN value with 0 and all Y values with 1.

3.4 Converting Categorical Data into Numerical

some columns have values in string categories, we need to convert them into numerical values for later normalization step. For example, WEATHER was consisted of string category (clear, rainy, …etc). This will create an issue for the classification algorithm. Thus, it was converted into numerical values(clear: 1, rainy: 2, … etc)

4. Results & Findings

In this section, the results and findings of the conducted experiment is illustrated.

4.1 Data Visualization & Analysis

we will try to understand the relationships between different variables by using visualization tools.

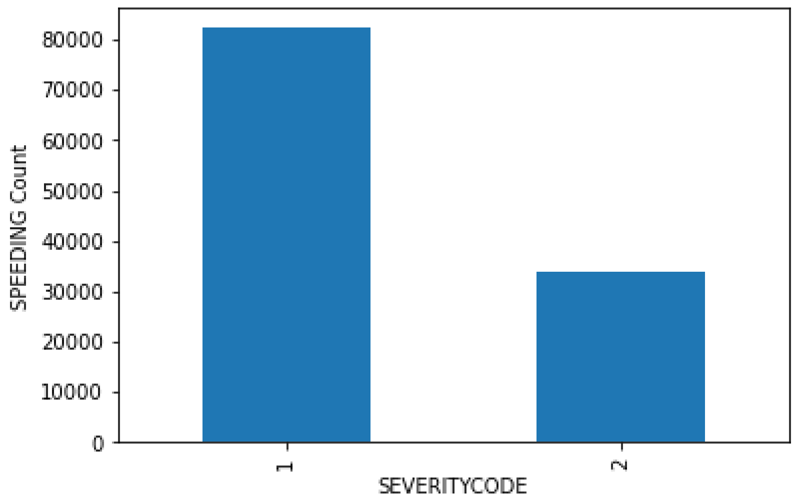

Figure 2: *Severitycode vs SPEEDING count*

from the Figure 2 , we can see that almost 30000 Accident involved speeding. Thus, this feature was included in the predictive model. Furthermore, we analyzed the weather data to see in what weather most collision happens. From Figure 3, we can see the majority of collisions happened in clear weather (2). followed by raining weather (7). Thus, we can see that rain has a huge impact on the likelihood of getting a collision.

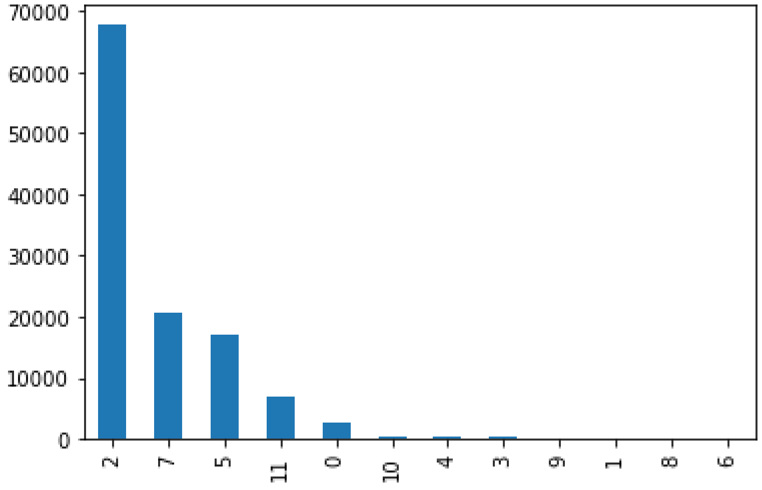

Figure 3: *Weather category counts*

Moreover, we plot the count of lightning condition, as shown in Figure 4, it is found that most collisions happened on daylight(1) followed by Dark Street Lights On (3). This implies that driving in dark light can increase the likelihood of getting a collision.

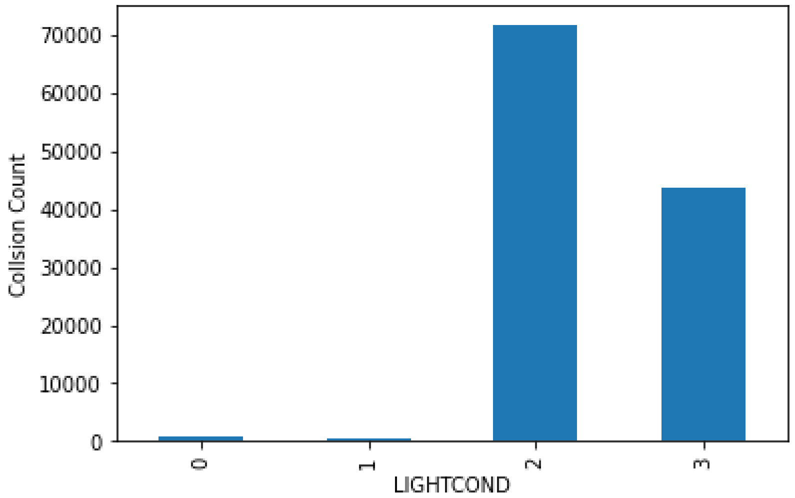

Figure 4: *Lightening Condition vs Collision Count*

Next, we started looking into accidents date to check what year contributed the most on the number of accidents.

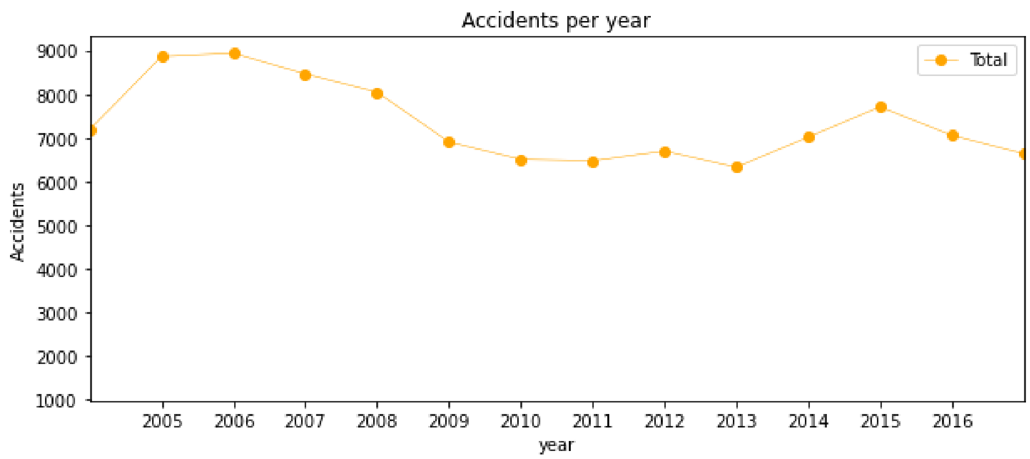

Figure 5: *Total Accidents per Years*

we can see that 2005 was the highest year in term of total accidents. The total number of accidents after 2005 started to fall until it raises back up on 2014 & 2015. The reason for this raise needs to be explored. Moreover, it can be noticed from Figure 6 that Friday is the day where most accident occurs.

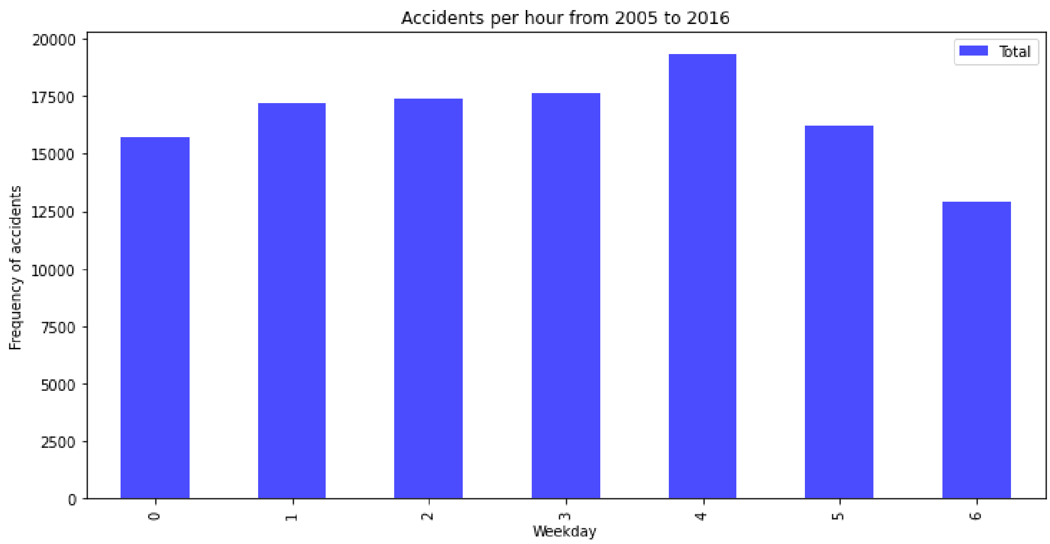

Figure 6: *Total Accidents per Weekday*

4.2 Developing the Predictive Model

In this step, we start building the predictive model using the training set and validation set. we started by splitting the dataset into training (80%) and validation set (20%). After splitting, the training set was left with 93081 samples while the test set contained 23271 samples.

Then, we chose the following supervisor model to be used as a predictive system:

- Naive Bayes

- K - Nearest - Neighbors
- Decision Trees
- Logistic Regression
- Random Forests

First, we normalized the values to avoid biases and increase system accuracy. Then, the predictive model was trained using the training set and tested on a different test set. The result of the predictive model is as follows:

Table 2: *Comparison Table*

Classification Algorithm	Accuracy
Naive Bayes	71.23%
K - Nearest - Neighbors	57.34%
Decision Trees	70.21%
Logistic Regression	71.11%
Random Forests	71.09%

From the above result, we can see that logistic regression has achieved the best classification results. Followed by the decision trees and random forest with an insignificant difference.

5. Conclusion

Road accidents are a serious danger to drivers and pedestrians. In this work, we implemented a predictive machine capable of predicting the likelihood of getting a severe collision based on some features such as the weather conditions, the road condition and many others. From analyzing the data, it was noticed that features like speed and consuming drugs and alcohol was one of the factors of having a collision. The final model has achieved an acceptable accuracy using logistic regression which is one of the supervised machine learning models. For future work, we encourage looking into a different dataset with a higher number of features to increase the classification accuracy.

Reference

[1] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] Pandas Dataframe: Plot Examples with Matplotlib and Pyplot, https://queirozf.com/entries/pandas-dataframe-plot-examples-with-matplotlib-pyplot