IBM Applied Data Science Capstone Project

Car accident severity (Week 1)

Snehal Kolekar

1. Introduction

Road traffic accidents are an important public health concern globally. Approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads [1]. An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities. Therefore, it is necessary to study the factors causing them and finding the strategies to tackle the road accidents.

Traffic accidents not only cause millions of disabilities and deaths, and take a toll on the countries' finance, but are also considered a big problem in the way of improving public health. Numerous factors such as environmental, vehicle-related, host (driver, pedestrian), and their type of interaction affect characteristics of these accidents. In order to find strategies for minimizing traffic accidents, we need to understand the combination of numerous factors.

Our main aim is to find what are the main factors causing road accidents and can we predict the severity based on these factors? This study will be helpful for understanding the design of road network, finding the high risk zones, evaluating the severity in particular areas. So, the targeted group will be city planners and engineers who can design the road network with consideration of preventative mechanisms for Traffic Accidents and Road Safety.

2. Data and Methodology

2.1 Data

The example dataset is used here which is recorded by SDOT (Seattle Department of Transport). Data_Collisions.CSV file is downloaded & read into pandas read_csv function in Jupyter notebook [2]. There are 194,673 observations and 38 variables in this dataset as shown in Figure 1 and Figure 2.

3	1	-122.334803	47.604803	4	1144	1144
4	2	-122.306426	47.545739	5	17700	17700

5 rows × 38 columns

```
[34]: df.shape
```

```
Out[34]: (194673, 38)
```

Figure 1: Screenshot from Jupyter notebook showing shape of dataset

```
35]: df.head(15)
```

Out[35]:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDRO
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	
5	1	-122.387598	47.690575	6	320840	322340	E919477	Matched	Intersection	36974.0	...	Dry	Daylight	
6	1	-122.338485	47.618534	7	83300	83300	3282542	Matched	Intersection	29510.0	...	Wet	Daylight	
7	2	-122.320780	47.614076	9	330897	332397	EA30304	Matched	Intersection	29745.0	...	Dry	Daylight	

Figure 2: Screenshot showing output of df.head(15)

Since we would like to identify the factors that cause the accident and the level of severity, we will use SEVERITYCODE as our dependent variable Y, and try different combinations of independent variables X to get the result. Since the observations are quite large, we may need to filter out the missing value and delete the unrelated columns first. Then we can select the factor which may have more impact on road accidents, such as address type, weather, road condition, and light condition.

2.2 Data Preparation

Before building machine learning models, there is need of data wrangling/ data cleaning which is the process of converting data from the initial format to the format that may be better for analysis. There are some reasons why this original dataset is not suitable for further analysis, they are as below.

1. There are some columns in dataset which have unrelated data for the car accident severity analysis. Examples of such columns are OBJECTID and INCKEY.
2. The part of dataset is categorical such as WEATHER, LIGHTCOND, ROADCOND.
3. But machine learning models need numerical data not categorical data.
4. Several unknown/ missing values present in the dataset. Those missing values may hinder further data analysis.

In order to use dataset for machine learning models, first I will do some standard steps required for data analysis.

1. Identifying missing data with python's isnull() function and counting missing values in each column.
2. Dealing with missing values such as dropping rows which have crucial missing data and dropping columns if most of entries are unknown.
3. Discarding columns which are irrelevant to car accident severity analysis.
4. One hot coding to convert categorical data into numerical data
5. Normalizing data into similar range

2.3 Data Visualiztion

To visualize different factors related to car accidents severity, I will use python library- Matplotlib & generate the graphs by using matplotlib.pyplot as shown in Figure 3 & Figure 4.

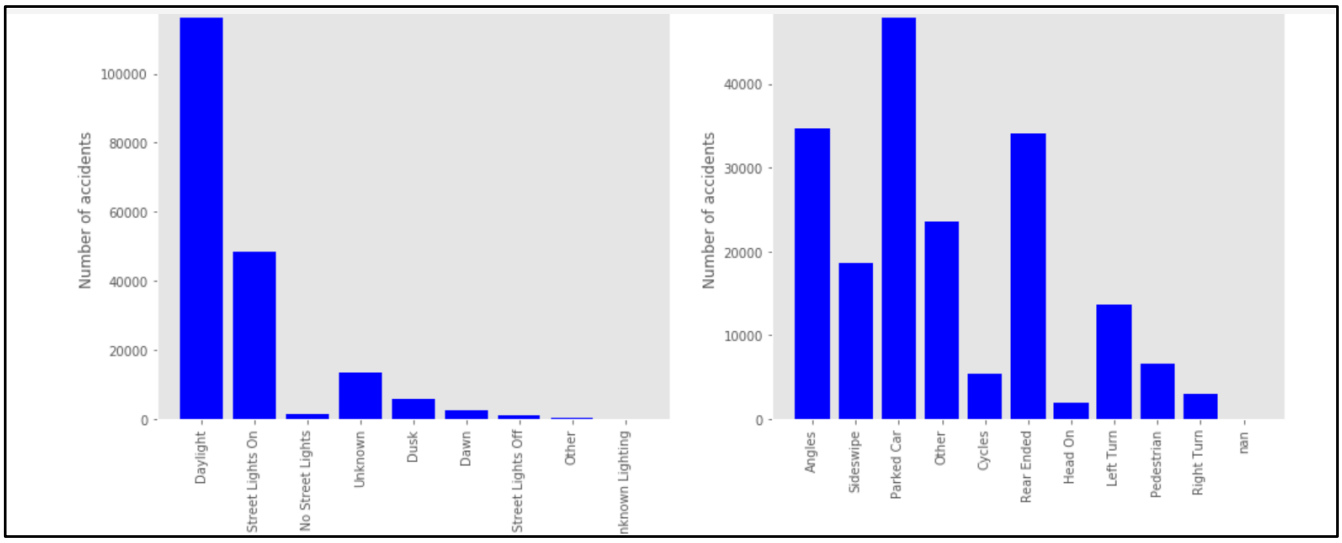

Figure 3: Screenshot of Jupytor notebook showing accidents related to local conditions such as light conditions and collision type

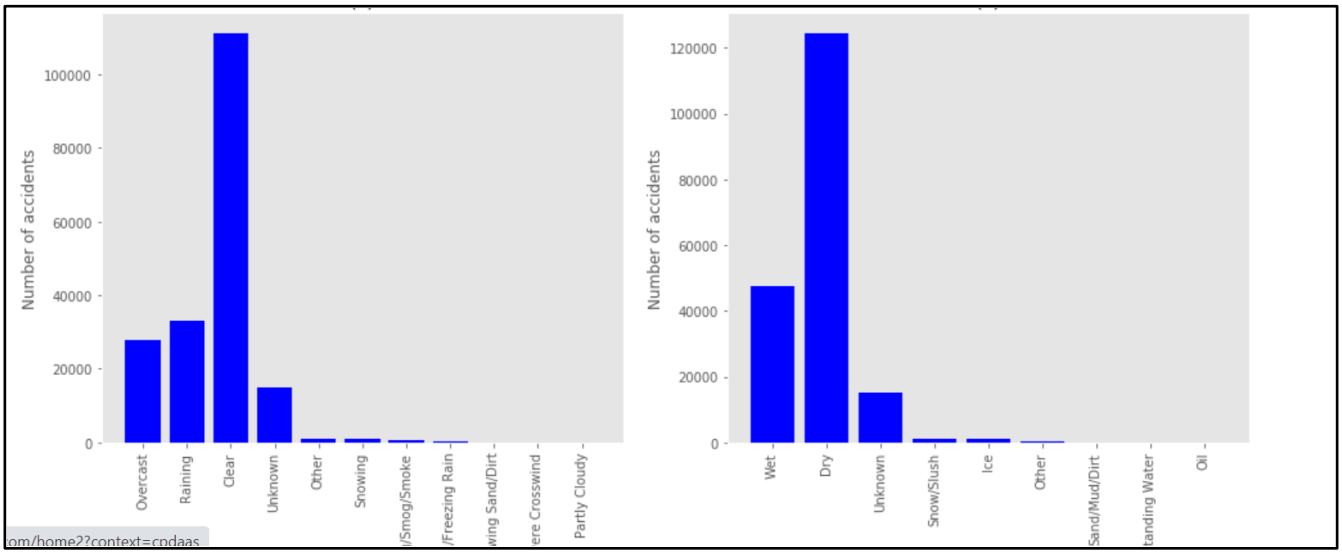

Figure 4: screenshot of Jupytor notebook showing accidents related to local weather & road condition

We can illustrate from Figure 3 as most of accidents are occurred during daylight and most of collisions included parked cars. From Figure 4, it is clear that most of the accidents are occurred during clear weather conditions and remaining are under overcast and raining conditions.

Methodology

After cleaning data, I will split data into training and testing dataset using TRAIN_TEST_SPLIT. Then I will build machine learning models for evaluation.

1. K-Nearest Neighbour (KNN) : This model will be used to predict severity of accidents in test dataset using train dataset whose conditions are nearly similar.

2. Decision Tree : In this model, the dataset is splitted and branched as a decision tree. Then, it is used to predict the severity in accidents in the test dataset.

3. SVM : The Support Vector Machine is also used to predict severity of accidents in test dataset.

After building these models, F1 score and Jaccard index are used to identify best model. That best model can be used for the further deployment & result will be helpful for targeted group.

References

1. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
2. https://www.coursera.org/learn/applied-data-science-capstone/supplement/Nh5uS/downloading-example-dataset