Prediction Modelling of Severity of Road Accident.

Applied Data Science Capstone Project offered by IBM.

Himanshu

September,2020

1. Introduction

1.1 Background

Road Accidents are one of the most prevalent issues of Modern Society which poses a heavy toll of losses to mankind. WHO describes the road traffic system as the most complex and dangerous system with which people have to deal every day. As per Global status report on traffic safety, 2018 there are 1.35 million traffic deaths in the world occurring annually. Also, the report suggests that Road traffic accidents rank 1st for the cause of deaths of people in the 5-19 years of age and ranks 8th for people of all ages. Hence, Accident prediction becomes the need of an hour for the optimization of public transportation, enabling safer routes, and cost-effectively improving transportation infrastructure; all leading to safer roads. Due to it's high significance, the topic has attracted attention of many researchers since the last decade. Analyzing the impact of various events such as traffic events(e.g., congestion, construction and road hazards), weather(e.g., temperature, visibility and wind speed), points-of-interest(e.g., traffic signal, stop sign and junction) using the availability of Big data in modern day with utilisation of various Advanced Statistical Methods have ease the thought process and lead the researchers to form a Predictive Model more efficiently.

1.2 Scope

The Goal of the project is to create a Predictive Model for Severity of Traffic Accidents. The target or response variable is "Severity". The Case Study involves the study of the City of Washington, Seattle. The datasets and all other required information are obtained from the Seattle GeoData website:
https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data?selectedAttribute=ADDRTYPE
The project is based on Cross Industry Process for Data Mining(CRISP) Methodology. As it is a case specific project, CRISP will provide a clear intent for the same.

The following six steps will be followed throughout the project:

A. **Business Understanding**: Intention of Project is to create a Predictive Model to predict the “Severity” of the accident based on the dataset provided. Hence it is clear that a Supervised Machine Learning Model will be advisable to use. The clarity of “labeled data” will further help in creating a training and testing data.
B. **Data Understanding**: The Data as mentioned above is obtained from the Seattle website. This step will involve understanding the various aspects of Data obtained and a deep analysis for the same.
C. **Data Preparation**: Based on the results from the previous step, the raw data will be transformed into a usable subset. Once the dataset is chosen, it must be checked for questionable, missing or ambiguous cases.
D. **Modelling:** In this step, data will be expressed through whatever appropriate models provide meaningful insights and any new knowledge if present. Different patterns and structure of the data will be captured using the various Modelling techniques.
E. **Evaluation**: The Model selected will be tested. A pre-selected test set will be used on the trained model. The effectiveness of the model will be analysed. Based on the analysis results efficacy of the model will be determined.
F. **Deployment**: Deployment for the model presented will be done on Github as part of the Course Project Guidelines.

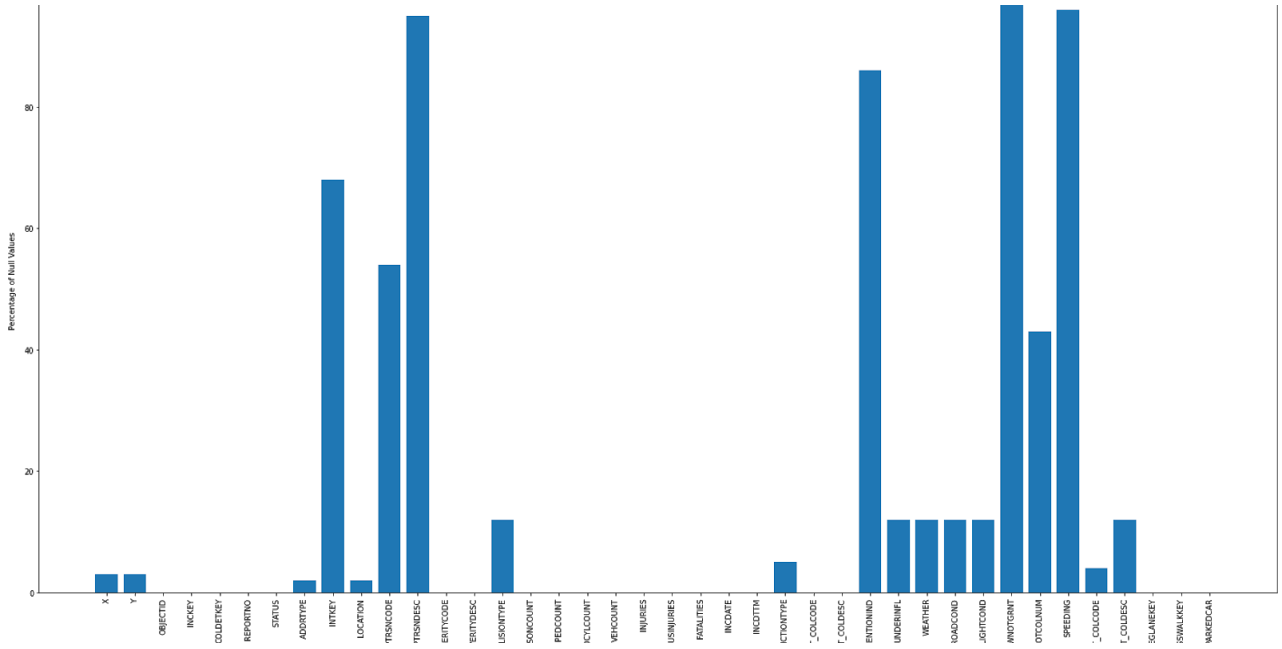
2. Data Understanding

A. **The dataset named “Collisions.csv”** obtained from the above mentioned website is imported in the form of Pandas Dataframe.
B. **Shape of the original Dataset : 221389 x 40** This denotes that we have 39 attributes as column “SEVERITYCODE” represents target or response variable. The dataset is comprehensive, however not all attributes are useful and hence we need to decide which attributes to keep and which to exclude from our final dataset. The discussion for the same will be done in the next section.
C. **Attributes and their data types:** Below table shows labels and their data types. It depicts the presence of both numerical and categorical variables. Hence, it should be dealt with in the next section.

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

```
{dtype('int64'): Index(['OBJECTID', 'INCKEY', 'COLDETKEY', 'PERSONCOUNT', 'PEDCOUNT',
        'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'SERIOUSINJURIES', 'FATALITIES',
        'SEGLANEKEY', 'CROSSWALKKEY'],
       dtype='object'),
 dtype('float64'): Index(['X', 'Y', 'INTKEY', 'SDOT_COLCODE', 'SDOTCOLNUM'], dtype='object'),
 dtype('O'): Index(['REPORTNO', 'STATUS', 'ADDRTYPE', 'LOCATION', 'EXCEPTRSNCODE',
        'EXCEPTRSNDESC', 'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE',
        'INCDATE', 'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLDESC', 'INATTENTIONIND',
        'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT',
        'SPEEDING', 'ST_COLCODE', 'ST_COLDESC', 'HITPARKEDCAR'],
       dtype='object')}
```

D. **Checking the number of NaNs** present in the Dataset in all columns. Below image shows the same.



{**'X': 3.0, 'Y': 3.0**, 'OBJECTID': 0.0, 'INCKEY': 0.0, 'COLDETKEY': 0.0, 'REPORTNO': 0.0, 'STATUS': 0.0, **'ADDRTYPE': 2.0**, '**INTKEY': 68.0**, '**LOCATION': 2.0**, '**EXCEPTRSNCODE': 54.0**, **'EXCEPTRSNDESC': 95.0,** 'SEVERITYCODE': 0.0, 'SEVERITYDESC': 0.0, **'COLLISIONTYPE': 12.0**, 'PERSONCOUNT': 0.0, 'PEDCOUNT': 0.0, 'PEDCYLCOUNT': 0.0, 'VEHCOUNT': 0.0, 'INJURIES': 0.0, 'SERIOUSINJURIES': 0.0, 'FATALITIES': 0.0, 'INCDATE': 0.0, 'INCDTTM': 0.0, '**JUNCTIONTYPE': 5.0**, 'SDOT_COLCODE': 0.0, 'SDOT_COLDESC': 0.0, **'INATTENTIONIND': 86.0**, **'UNDERINFL': 12.0**, **'WEATHER': 12.0**, **'ROADCOND': 12.0**, '**LIGHTCOND': 12.0**, **'PEDROWNOTGRNT': 98.0**, **'SDOTCOLNUM': 43.0**, **'SPEEDING': 96.0**, '**ST_COLCODE': 4.0**, **'ST_COLDESC': 12.0**, 'SEGLANEKEY': 0.0, 'CROSSWALKKEY': 0.0, 'HITPARKEDCAR': 0.0}

1. Analyzing “SEVERITYCODE” column:

```
1     137596
2      58747
0      21594
2b      3102
3        349
Name: SEVERITYCODE, dtype: int64
```

```
Property Damage Only Collision    137596
Injury Collision                   58747
Unknown                            21595
Serious Injury Collision            3102
Fatality Collision                   349
Name: SEVERITYDESC, dtype: int64
```

- As per above observation, 21594 rows have severity code = 0, and as per metadata, 0 refers to unknown severity hence it cannot be used to train/test models. Also, there are 21595 “Unknown” values in “SEVERITYDESC”. So we need to drop these rows.

2. Analyzing Redundant Columns:

The table below shows the labels for which no information is provided in the metadata or is not required for the final dataset. Hence, below columns can be dropped.

ATTRIBUTES	REASON FOR REDUNDANCY
OBJECTID	It is just a database key provided by ESRI which is not required in our case.
COLDETKEY	It is a Secondary key like “OBJECTID”
REPORTNO	There is no information on this in Metadatas and observing the database suggests it's just a number for record purposes which is not required to train the model.
STATUS	There is no information about this in Metadata and the column includes- “MATCHED” or “UNMATCHED”, so we can safely drop this attribute.
EXCEPTRSNCODE	No information provided in the Metadata and the below percentage of NaNs present shows 54 and 97% of blank cells for these two respectively, hence they can be dropped.
EXCEPTRSNDESC	
INCDATE	It is exactly same as “INCDTTM” which has better format of data available.
SDOTCOLNUM	It is similar to “INCKEY” and it has 43% data missing, so we can drop it.

3. **Analyzing Missing Information**:

As per the bar chart present above which depicts the percentage of NaNs present in each attribute, we can draw inference for the preliminary missing information with the help of Metadata and other data understanding. Also, we won’t discuss the attributes which we have decided to drop in the previous step. Below table presents a general understanding for the same however it must be noted that this is just a preliminary step. A more deeper action can be taken only after the process of Exploratory Data Analysis is accomplished which will be dealt in later section.

ATTRIBUTES	UNDERSTANDING AND ACTION
“X”,”Y” - Coordinates	As we don’t have information on the co-ordinates, we can remove the null rows. The relationship between INTKEY and Co-ordinates were observed to fill the missing value based on INTKEY but there was no any information.
“ADDRTYP”	This will be analysed later.
“INTKEY”	As per the understanding, “INTKEY” attribute is not required for the modelling. Also, 68% of the data is missing and hence it can be safely dropped.
“LOCATION”	This will be analysed later.
“COLLISIONTYPE”	This will be analysed later.
“JUNCTIONTYPE”	This will be analysed later.
“UNDERINFL”	This will be analysed later.
“WEATHER”	12% of the data is missing for these three attributes. However, as per understanding, these are important variables for training the model. On observing the data, which is provided below, for most of these null attributes, the “SEVERITYCODE” = 0, which refers to “UNKNOWN”, hence we can safely remove these data.
“ROADCOND”	
“LIGHTCOND”	
“'PEDROWNOTGRNT”	Around 98%, 96% and 86% of data is missing for these three attributes. However as per Metadata, these attributes accept either “Y” or “N” resembling “YES” or “NO” and the dataset has only “Y” present, so it can be safely assumed, that the missing cells represents “N”
“SPEEDING”	
“INATTENTIONID”	
“ST_COLCODE”	4% data is missing. Required analysis will be done later.
“ST_COLDESC”	12% of the data is missing for this case. However, we have the attribute “ST_COLDESC” which informs us about the collision. Hence, this is not required.

4. Handling Categorical Variables:

In order to train a model using techniques of machine learning, categorical variables are not used. Hence, it must be converted to numeric values. Below is the table for the same:

ATTRIBUTES	ACTION
“SPEEDING”	These attributes have values - “Y” for “YES”, “N” for “NO” and Null. The same can be replaced with 1 or 0 depending on “Y” or “N” and Null.
“INATTENTIONIND”	
“UNDERINFL”	
“PEDROWNOTGRNT”	
“HITPARKEDCAR”	
“LOCATION”	Location is not required as the accident coordinates are provided and is sufficient to visualize. Also, it wont play any role in modelling the data.
“SEVERITYCODE”	This attribute has values - 0,1,2,2b,3 for different Severities. “2b” can be converted to 3 and “3” can be converted to “4”.
“COLLISION TYPE”	The categories can be converted to numeric, appointing digits to the categories.
“JUNCTIONTYPE”	
“WEATHER”	The attributes have a bunch of values and “ONE-HOT ENCODING” can be used to convert categorical data to numerical data.
“ROADCOND”	
“LIGHTCOND”	
“ADDRTYPE”	

5. HANDLING TIMESTAMP:

The column INCDTTM has a format of date/time and can be converted to actual timestamp in separate columns.

Conclusion on Preliminary Data Understanding:

Following Data Cleaning Steps need to be implemented before diving into the Exploratory Data Analysis.

1. Remove rows for which Target Variable has “UNKNOWN” values.
2. Remove all redundant columns discussed above.
3. Handling Missing Information.
4. Handling Categorical Variables-
    a. Replacing categories with numeric values.
    b. One-Hot Encoding.
5. Handling Timestamp issue.
6. Final Data Cleaning based on preliminary results.