IBM Data Science Professional Certificate

Applied Data Science Capstone Project

Predicting Severity of Car Accidents in Seattle

By: Rama Chaitanya Samanchi

September 2020

Introduction

Driving a car is important today as it gives people power personal control and autonomy. So, the project deals with an aspect that is interesting from a Data Analyst's perspective while we look at the safety associated with driving a car. It would be great if we can model collisions associated with cars which can aid in understanding the factors associated with a collision. But what is the Business Interest involved in such a study? What are the factors which are important for the study? The following two sections will answer these questions to kick-start our project (with safety!).

Business Problem

Problem Statement

The objective of the project is to develop a model that can predict the severity of a car collision based on several important factors that influence a collision. The study requires a classification model that can predict the impact of an accident severity in a metro city called Seattle, Washington.

Background

For each accident taken place, the local traffic department of Seattle records each incident with a unique key and multiple factors related to the collision like the location, the number of people involved in, weather conditions, road conditions, light conditions, speeding, etc. These details are some of the important details which contribute to the severity of an accident. Certainly, these are the factors to investigate and analyze the data to build a model that can predict the severity of the collision to start with. The study will also provide the factors by importance in predicting the severity.

Target Stakeholders

The model is very helpful for the traffic government to take mitigative actions to reduce car collisions based on the factors which have a large influence in predicting the severity of the collision. For example, the traffic department can take preventive actions if observed that road condition is playing an important role in the classifying the severity of an incident. They can not only repair the road on which current accidents have taken place but also take preventive action by identifying such roads in similar condition and fix them to reduce the possibility of incidents in the future. Similarly, Smart Street Lighting Systems can be deployed to improve the visibility on the streets if Lighting condition was found to have a significant role in the severity of the collision. Therefore, the administration is the first and foremost target stakeholders of the project.

Also, the emergency service provider and healthcare systems will have an effective tool in predicting the severity of the collision to provide necessary aid to potentially save the lives and improve the life-sensitivity. They are second target stakeholders to get benefitted by the model as an accurate model is handy in predicting the severity in taking proactive action.

Data Understanding

Dataset Description

The dataset is recorded by the traffic department of Seattle and available on the government site which is provided through Coursera. We were given an option either to choose the one provided or select a new one. I am going ahead with the one provided by Coursera for the simplicity and effective building of the model. The dataset contains 194673 rows as the dataset captured the collision for multiple years i.e. from 2004 to 2020 (May). The dataset has 37 columns which include description columns as well which will be dropped while building the model.

Columns

Column Name	Data Type	Description
OBJECTID	ObjectID	ESRI unique identifier
SHAPE	Geometry	ESRI geometry field
INCKEY	Long	A unique key for the incident
COLDETKEY	Long	Secondary key for the incident
ADDRTYPE	Text, 12	Collision address type: • Alley • Block • Intersection
INTKEY	Double	Key that corresponds to the intersection associated with a collision
LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • 3—fatality • 2b—serious injury • 2—injury • 1—prop damage • 0—unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text, 300	Collision type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state
INJURIES	Double	The number of total injuries in the collision. This is entered by the state
SERIOUSINJURIES	Double	The number of serious injuries in the collision. This is entered by the state
FATALITIES	Double	The number of fatalities in the collision. This is entered by the state
INCDATE	Date	The date of the incident. INCDTTM Text, 30 The date and time of the incident.
JUNCTIONTYPE	Text, 300	Category of the junction at which collision took place
SDOT_COLCODE	Text, 10	A code is given to the collision by SDOT.
SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.

PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text, 10	A number is given to the collision by SDOT.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text, 10	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	Text, 300	A description that corresponds to the state's coding designation.
SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)

Key pattern observations

1. These are the columns and their number of null values.

```
In [9]: Data_Collisions.isnull().sum()
```

```
Out[9]: SEVERITYCODE           0
        X                   5334
        Y                   5334
        OBJECTID               0
        INCKEY                 0
        COLDETKEY              0
        REPORTNO               0
        STATUS                 0
        ADDRTYPE            1926
        INTKEY            129603
        LOCATION            2677
        EXCEPTRSNCODE     109862
        EXCEPTRSNDESC     189035
        SEVERITYCODE.1         0
        SEVERITYDESC           0
        COLLISIONTYPE       4904
        PERSONCOUNT            0
        PEDCOUNT               0
        PEDCYLCOUNT            0
        VEHCOUNT               0
        INCDATE                0
        INCDTTM                0
        JUNCTIONTYPE        6329
        SDOT_COLCODE           0
        SDOT_COLDESC           0
        INATTENTIONIND    164868
        UNDERINFL           4884
        WEATHER             5081
        ROADCOND            5012
        LIGHTCOND           5170
        PEDROWNOTGRNT     190006
        SDOTCOLNUM         79737
        SPEEDING          185340
        ST_COLCODE            18
        ST_COLDESC          4904
        SEGLANEKEY             0
        CROSSWALKKEY           0
        HITPARKEDCAR           0
        dtype: int64
```

2. Dropped some columns which are descriptions and have too many null values.

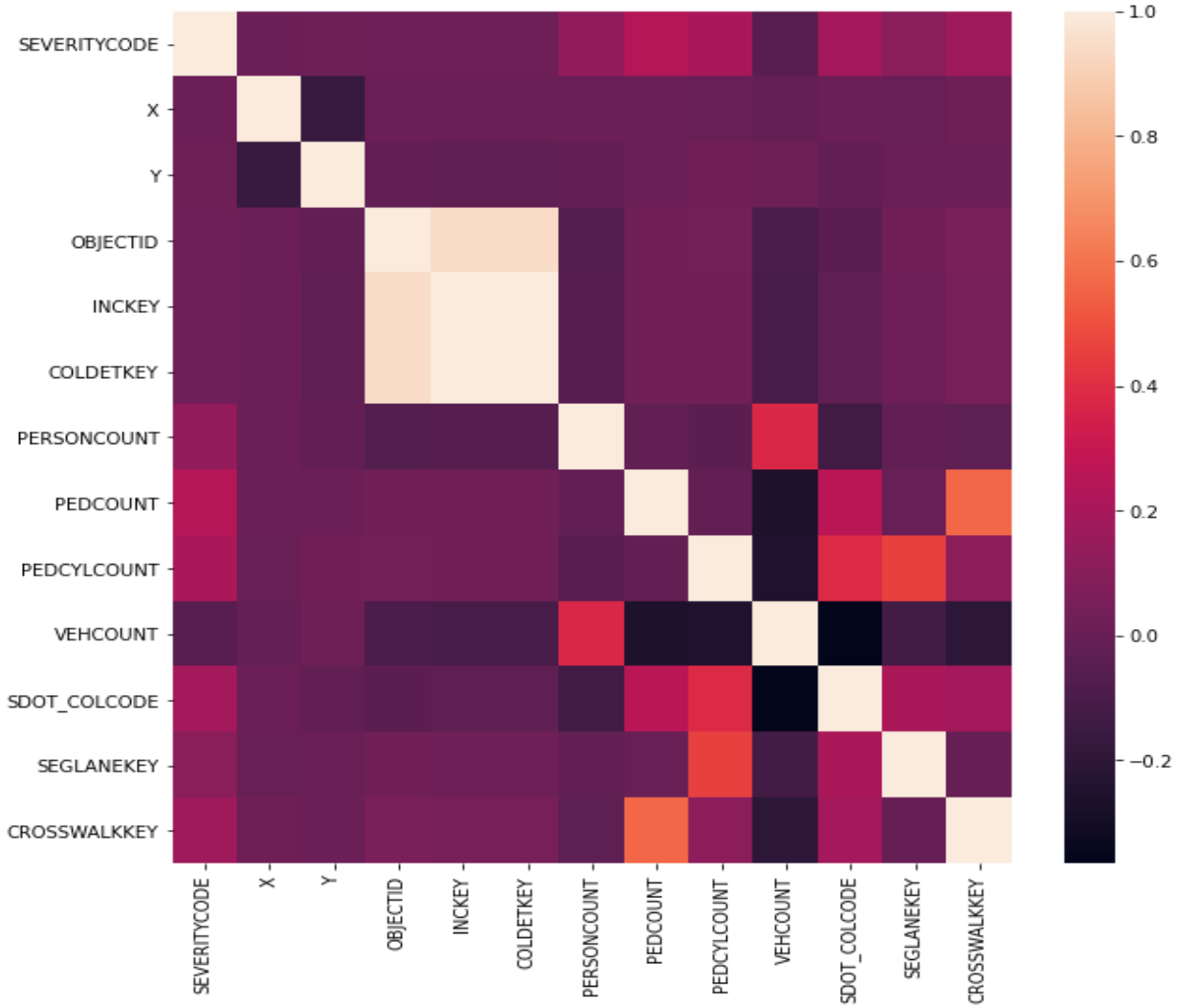
```
In [11]: New_Collisions = Data_Collisions.drop(['INTKEY', 'EXCEPTRSNCODE', 'SEVERITYCODE.1',
                                                'EXCEPTRSNDESC', 'SEVERITYDESC', 'SDOT_COLDESC',
                                                'INATTENTIONIND', 'PEDROWNOTGRNT','SDOTCOLNUM',
                                                'SPEEDING', 'ST_COLDESC'], axis=1)
```

3. The Target variable Severity Code is imbalanced with two classifications 1 & 2.

```
In [12]: New_Collisions['SEVERITYCODE'].value_counts()

Out[12]: 1    136485
         2     58188
         Name: SEVERITYCODE, dtype: int64
```

4. An early correlation heat suggests that there is not much correlation with some columns missing due to text datatype.



All these issues will be taken care of in the next section Data Preparation.