**IBM Data Science Professional Certificate Capstone Project**

**Predicting the Severity of Car Accident**

BY: VARUN PATEL

19 September 2020

**Introduction**

Nowadays road accident are very common. These incident causes property as well as life damage, which is not good. According to the survey approx. 62% accident occur due to rash driving. There are many other major factors like weather condition, light condition etc. as technology is advancing day by day, making it possible to learn the pattern and causes of accident based on previously collected data. Doing so will help us to find the condition which result in most accidents. And spread awareness about the impact of different weather conditions. Making it available to public will help then to change their route and avoid any such situation, we can harness the power of data science and modern technologies. To reduce the numbers of accident. We can combine IoT, data science and deep learning to develop a tool that can automatically sense the weather condition and collect different required data. And warn the driver if there are any chances of accident, and what can be the severity of accident. This could also help us to avoid long traffic caused by accident, by telling us that there may be a chance that there is an accident. So we could change our route. This project cover the prediction of accident severity with the help of different machine learning algorithms.

**Data**

The dataset I am using is provided by Applied Data Science Capstone course by IBM on Coursera. This dataset contain the accident (collisions) record of Seattle. This dataset contains 37 columns and 194673 rows. Table 1 shows all attributes with their datatype and description. I have cleaned the dataset so that it can be used to train different machine learning algorithms.

I have used 10 columns for training the model:

1. SEVERITYCODE
2. ADDRTYPE
3. COLLISIONTYPE
4. VEHCOUNT
5. JUNCTIONTYPE
6. UNDERINFL
7. WEATHER
8. ROADCOND
9. LIGHTCOND
10. HITPARKEDCAR
11. TOTAL\_PERSON

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data type, length** | **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 | Description of the general location of the collision |
| EXCEPTRSNCODE | Text |  |
| EXCEPTRSNDESC | Text |  |
| SEVERITYCODE | Text | 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 | 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. |
| PEDCYCOUNT | 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 | The date and time of the incident. |
| JUNCTIONTYPE | Text | Category of junction at which collision took place |
| SDOT\_COLCODE | Text | A code given to the collision by SDOT. |
| SDOT\_COLDESC | Text | A description of the collision corresponding to the collision code. |
| INATTENTIONIND | Text | Whether or not collision was due to inattention. (Y/N) |
| UNDERINFL | text | Whether or not a driver involved was under the influence of drugs or alcohol. |
| WEATHER | Text | A description of the weather conditions during the time of the collision. |
| ROADCOND | Text | The condition of the road during the collision. |
| LIGHTCOND | Text | The light conditions during the collision. |
| PEDROWNOTGRNT | Text | Whether or not the pedestrian right of way was not granted. (Y/N) |
| SDOTCOLNUM | Text | A number given to the collision by SDOT. |
| SPEEDING | Text | Whether or not speeding was a factor in the collision. (Y/N) |
| ST\_COLCODE | Text | A code provided by the state that describes the collision |
| ST\_COLDESC | Text | A description that corresponds to the state’s coding designation. |
| SEGLANEKEY | Long | A key for the lane segment in which the collision occurred. |
| CROSSWALKEYKEY | Long | A key for the crosswalk at which the collision occurred. |
| HITPARKEDCAR | Text | Whether or not the collision involved hitting a parked car. (Y/N) |

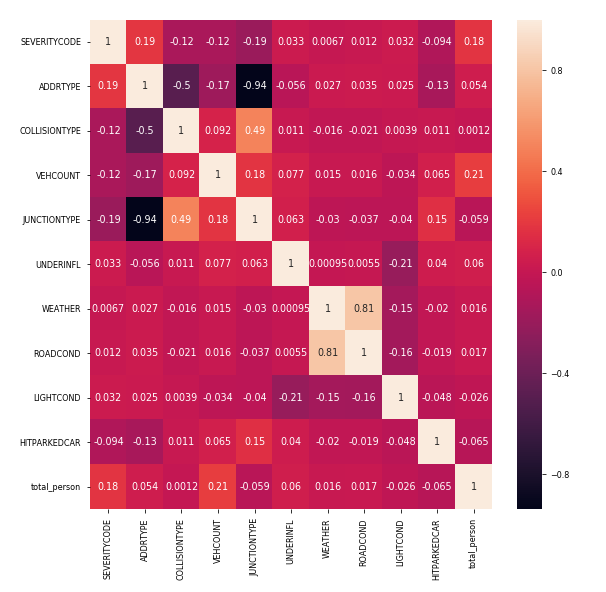
**METHODOLOGY**

The data is visualized for correlation. Negatively correlated features are selected to be dropped. Feature importance is plotted to visualize and only features with high importance are taken into consideration for predicting accident severity. The multi class label is converted to binary class by merging “Serious” and “Fatal” to Serious class.

Feature Selection: The dataset has 34 attributes describing the incident of an accident. There are mixed types of data such as continuous and categorical. Manually dropped few columns due to its inconsistency in values such as Accident ID, and Location ID. For selecting the best features, below functions are used from sklearn library.

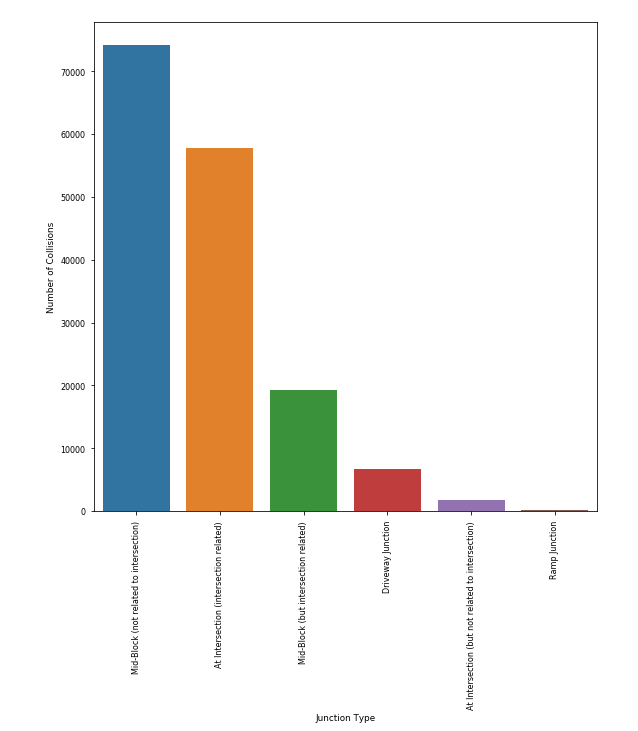
**Co-Relation Heatmap**

A heatmap is a graphical representation of data in which data values are represented as colors. That is, it uses color in order to communicate a value to the reader. This is a great tool to assist the audience towards the areas that matter the most when you have a large volume of data.

****

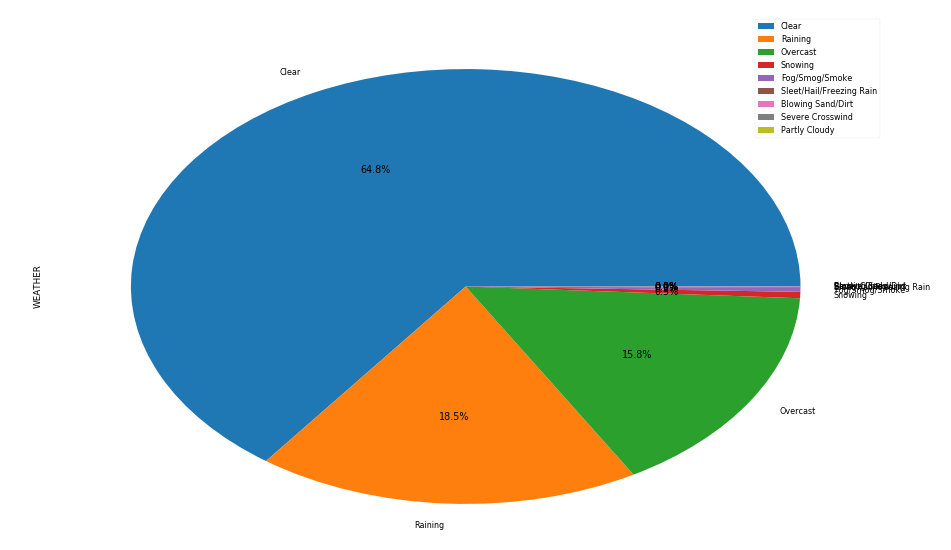
**GRAPHS:**

**1.**

****

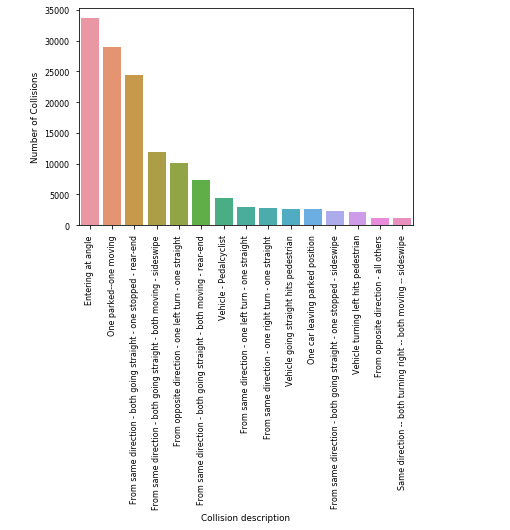
It can be seen that the number of collisions that take place at the intersection of roads or intersection related collisions are more than those not related to intersections. This information can be used to make rules specifically for the intersections. same goes with making rules for mid block.

2.



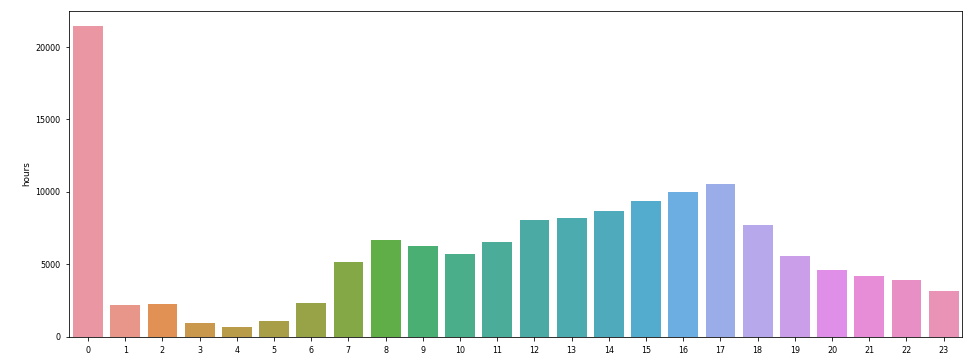
we can see here approx 65% accident occur in clear weather and 18.5% in raining weather.

**3.**

****

here we can conclude that most accident occour when a car take turn a steep angle. this create a blindspot as driver can see ahead because of obstacle. resulting in accident.

4.



here we can clearly see major accident occured during midnight, other than this high accident rate is at 5pm

**RESULT**

We used 3 supervised machine learning algorithms and we were successfully able to achieve an accuracy of around 73 %. Table 2 represents accuracy score and f1\_score of all 3 algorithm

|  |  |  |
| --- | --- | --- |
|  | **ACCURACY SCORE** | **F1-SCORE** |
| **LOGISTIC REGRESSION** | 0.70 | 0.66 |
| **K-NEAREST NEIGHBOUR** | 0.71 | 0.68 |
| **DECISION TREE** | 0.72 | 0.68 |

Table: 2

**DISCUSSION**

Hence we can see that major factor that lead to collision are collision type, weather condition, road condition, light condition. We can use this model with little or no change if we want to use it for some other city, because ever place may have its own category to classify the severity of accident. Features like weather, road, and light can be used to determine that will collision occurs or not. For determining severity of accident features like people count, vehicle count, speed, etc. can be used

**CONCLUSION**

Accidents are matter of great concern both people and government should take it seriously, place specific rules should also be made. People should be made aware of rules, combined work of people, government and science can reduce accident in upcoming years.