“CAPSTONE PROJECT REPORT”

TOPIC: CAR ACCIDENT SEVERITY

AUTHOR: MURAD POPATTIA

Introduction (Business Problem):

In this project we will try to find potential location for a car accident given the weather and road conditions. Specifically, this report will be targeted to stakeholders interested in increasing public road safety in Seattle, Washington, United States.

Unfortunate incidents often occur on the road. The unpredictability of these accidents make them so dangerous. It is usually the impatient nature of the human which leads to an accident. But wouldn't it be a blessing if the number of accidents that occur on a daily basis be reduced? That is the problem which will be addressed in this project. As we know, accidents can happen anytime and anywhere. However, there are many factors which might influence the severity of the accident. For instance, these include the weather conditions, time of the day, the speed of the car and the area the car is being driven. These factors greatly contribute whether the accident will be severe or not. Having the information about the above-mentioned factors can be used to predict if the accident happens can be severe or not. The indication that there is a possibility of a severe accident if it happens might warn the car drivers to drive more carefully and hence prevent accidents.

This would greatly reduce the loss of life and also damage to the property. This would also help routing software to give a warning which would alert the drivers and their insurance companies which would help them in saving cost.

Data:

The dataset used was provided by Coursera and it contains all collisions provided by SPD and recorded by Traffic Records, including all types of collision since 2004 to present.

It provides many columns with details on each type of colision. We excluded most of them and kept the columns that represents information data that would be available to some kind of traffic app, for example, in real time. The selected columns were:

- **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
- **INJURIES:** The number of total injuries 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:** Date of Accident
- **INCDTTM:** The Date and Time of Accident
- **JUNCTIONTYPE:** Category of junction at which collision took place
- **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.
- **SPEEDING:** Whether or not speeding was a factor in the collision. **(Y/N)**
- **HITPARKEDCAR:** Whether or not the collision involved hitting a parked car. **(Y/N)**

An overview of the data columns can be seen from the screenshot below:

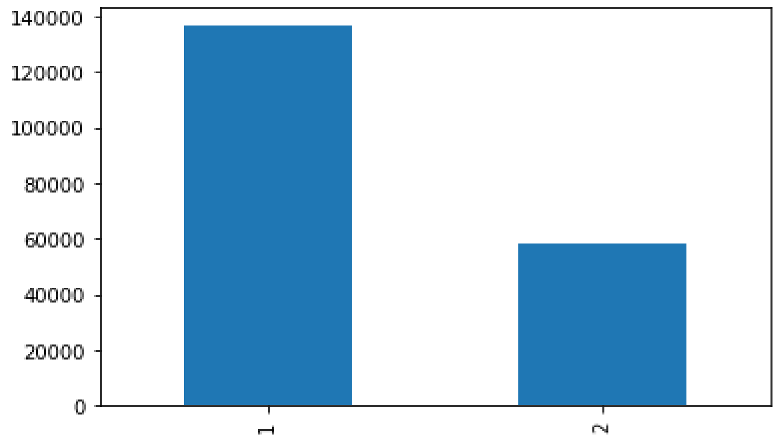
```
In [27]: df.columns

Out[27]: Index(['SEVERITYCODE', 'X', 'Y', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS',
                'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC',
                'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT',
                'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE', 'INCDTTM',
                'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC', 'INATTENTIONIND',
                'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT',
                'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY',
                'CROSSWALKKEY', 'HITPARKEDCAR'],
               dtype='object')
```

The data labels only contain two classes of severity, hence this can be treated as a binary classification problem, for which we can use a logistic regression classifier in order to determine the severity of the accident.

```
In [26]: # Checking dataset balanced or not
         %matplotlib inline
         df['SEVERITYCODE'].value_counts().plot(kind='bar')

Out[26]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0f66af5518>
```



The data is also imbalanced hence, we need to take appropriate measures in order to scale the data to get correct predictions.

Methodology:

(i): Data Pre-processing:

The data labels only contain two classes of severity, hence this can be treated as a binary classification problem, for which we can use a logistic regression classifier in order to determine the severity of the accident. Moreover, Since the data consists of many unnecessary columns and missing values, the data was cleaned using different data cleaning methods. Since the prediction was mainly based on Person Count, Vehicle Count, Attention ID, under influence verification, Weather, Road conditions, Lighting conditions; the other trivial columns were removed.

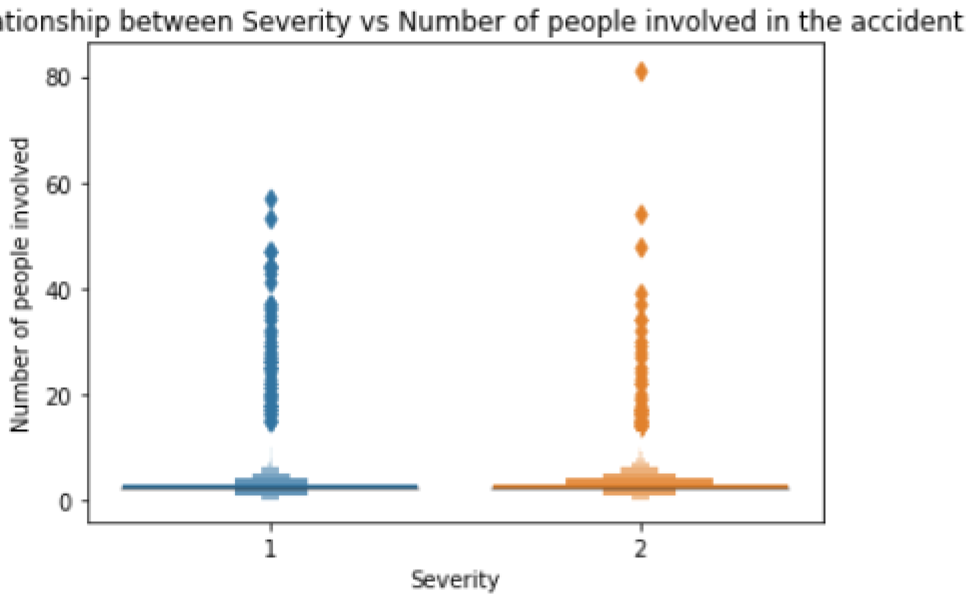
The columns in the data consisted of null values which were hindering in increasing the accuracy of the prediction model. Therefore, the missing values in the columns were replaced with the values of highest frequency. This prevented reduction in the training data set and provided higher efficiency.

Using the **One-Hot-Encoding** method, columns containing different categorical labels were split into more columns. Columns which contained the values initially were given a value 1 and the rest were given a value of 0.

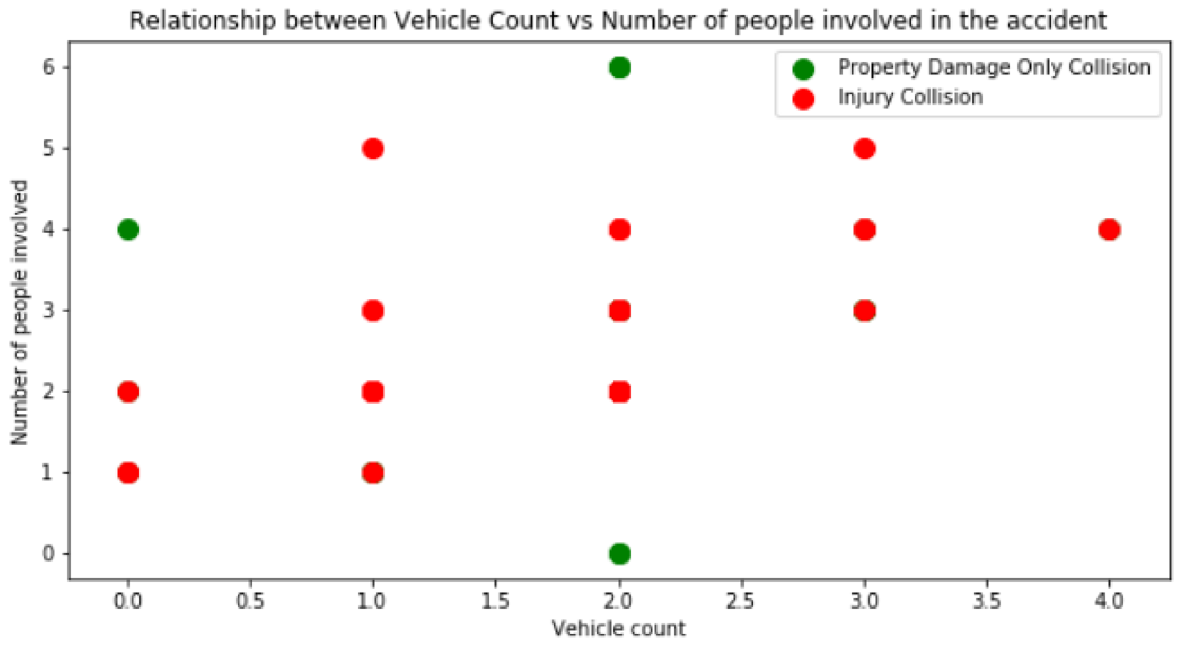
Many columns contained different values involving a number and a character. The values were standardized by replacing the characters ‘Y’ and ‘N’ with 1 and 0 respectively. This would ensure that the classification model would have integer values as input.

(ii): Exploratory Data Analysis:

The data in the columns was used for exploratory analysis. Firstly, a boxplot of severity vs the number of people involved in the accident was plotted. It was observed that majority of the accidents involved around 1 to 2 people. Through this we could conclude that if there are higher people traveling, the possibility of severe collision is less since more people mean more traffic which leads to less possible speeding. The box plot below shows this analysis:



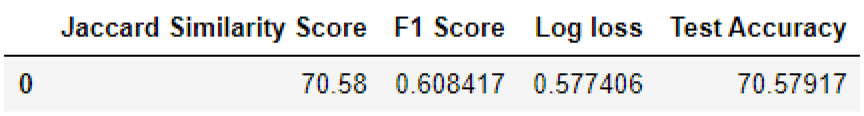
Another important observation was found out by using the scatter plot by plotting the vehicle count and the number of people involved in the accident. Seeing this plot, we could conclude that conclude that when the number of vehicles involved in the accident were high, there was more injury collision than the collisions which only damaged the property.
The scatter plot below shows this analysis:



As mentioned above, there are only two classes hence, we can treat this is as a binary classification problem and use a logistic regression classifier for the task. GridSearchCV has been implied for parameter tuning of the model.

Results:

A linear regression model was successfully created to predict the severity of the road accidents by evaluating various factors that occurred in the training set data. The best parameters were found for the model and the model accuracy was measured. The model accuracy was found to be 70.5%. Below are some performance metrics used for model evaluation.



	Jaccard Similarity Score	F1 Score	Log loss	Test Accuracy
0	70.58	0.608417	0.577406	70.57917

Conclusion:

Further additions to the project will involve improving the efficiency of the model by providing a more diverse dataset which would contain multiple degrees of severities. Moreover, we can also use a better algorithm with more hyperparameter tuning. The dataset as we see is imbalanced hence, we can use down-sampling as well in order to avoid biasness in results. However, as a benchmark, this has achieved respectable accuracy.