Predicting the Car Accident Severity in Seattle

Sakif Ahmed

September 21, 2020

Data sources and Data Preprocessing

Data Sources

The example dataset comes from the web page of Coursera Course Applied Data Science Capstone. The .csv file contains 19,4673 pieces of collision records of Seattle. The data for this capstone project is offered by the SDOT Traffic Management Division and recorded by Traffic Records Group. It covers the annual collisions data from 2004 to the present. The time-frequency of this dataset is weekly and it shows the traffic collision records in Seattle. The example datasets contain 194,673 pieces of records starting from 2004. The attributes in the datasets cover the weather condition, road condition, collision type and fatality.

The other relevant data comes from the database of Seattle government. The main data source is from the Seattle Department of Transportation. The basic background information is offered by this department.

Data Cleaning

The original dataset is not suitable to work with. There are 37 attributes with a lot of missing values. To clean the dataset, the rows with missing values were dropped. After cleaning the dataset from missing values, the dataset now contains 110,586 pieces of collision records.

Feature Selection

After data cleaning, there were 110,586 samples and 37 features in the data. Upon examining the meaning of each feature, I have decided to focus on only six features: SEVERITYCODE, JUNCTIONTYPE, PERSONCOUNT, WEATHER, ROADCOND, LIGHTCOND. These features were selected due to their suitability for the upcoming building of the machine learning models.

In addition, I have selected SEVERITYCODE as the target variable as it is best suited for car accident severity prediction.

Feature Engineering and Balancing of Data

The selected features have two types of variables: numerical and categorical. To best suit the machine learning model, I have converted the categorical features to numerical values. This action was done by using the df.replace() method.

On the other hand, the targe feature SEVERITYCODE has a problem too. The data set of this feature is imbalanced (Figure.1(a)). This imbalance may lead the machine learning model to be biased. This problem can be fixed by down sampling the majority class by using the resample method (Figure.1(b)).

Table 1. Description of the selected features

Feature	Data Type, Length	Description
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
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
PERSONCOUNT	Double	The total number of people involved in the collision
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

	SEVERITYCODE
1	136485
2	58188

(a)

	SEVERITYCODE
2	58188
1	58188

(b)

Figure. 1 The values of feature ser SEVERITYCODE: (a) imbalanced (b) after down sampling

Exploratory Data Analysis

To explore the relationship among the feature sets, I have conducted exploratory data analysis ploting the feature variable against the target variable. To achive effective visualization of the features, bar plot is used.

Relationship between Junction type and Severity

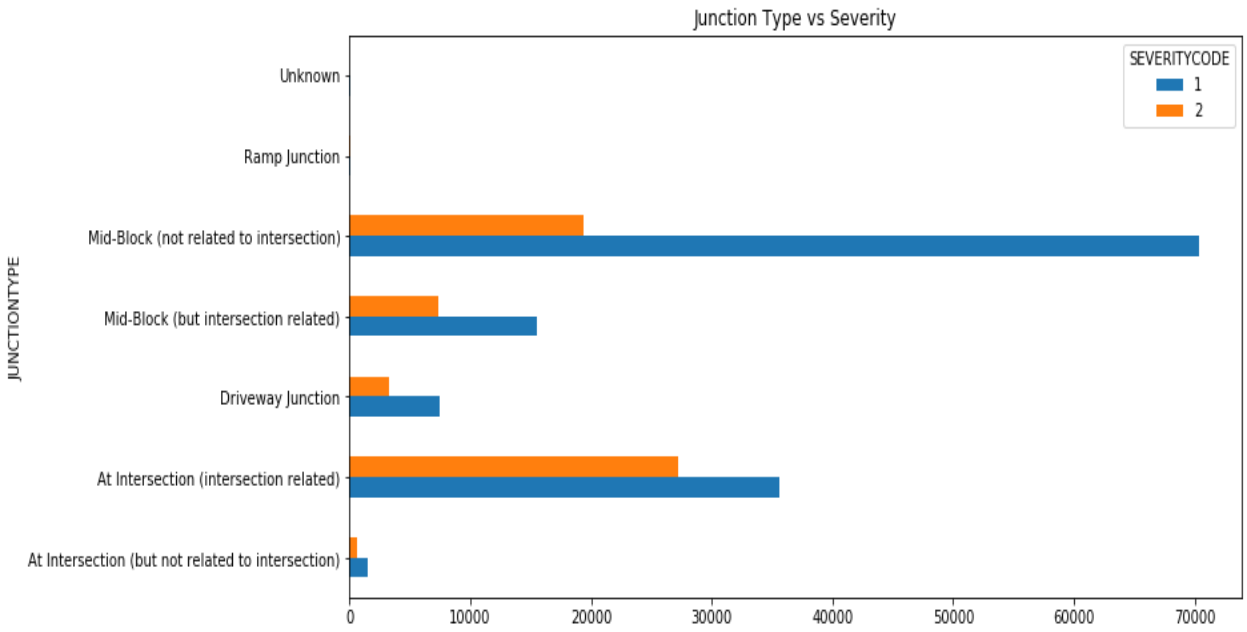

Figure.2 Visual representation of the relationship between junction type and severity

The feature junction type is a categorical variable which refers to the category of junction at which collision took place. It is obvious that there are some junctions more prone to car collision than other, most notably Mid-Block (not related to intersection) and At Intersection (intersection related) (Figure.2). Furthermore, we can also find from the visual representation of the relationship that Mid-Block (not related to intersection) junction has very high property damage. On the other hand, although At Intersection (intersection related) has lower property damage, this junction has the highest number of accident injuries.

Relationship between Number of Persons and Severity

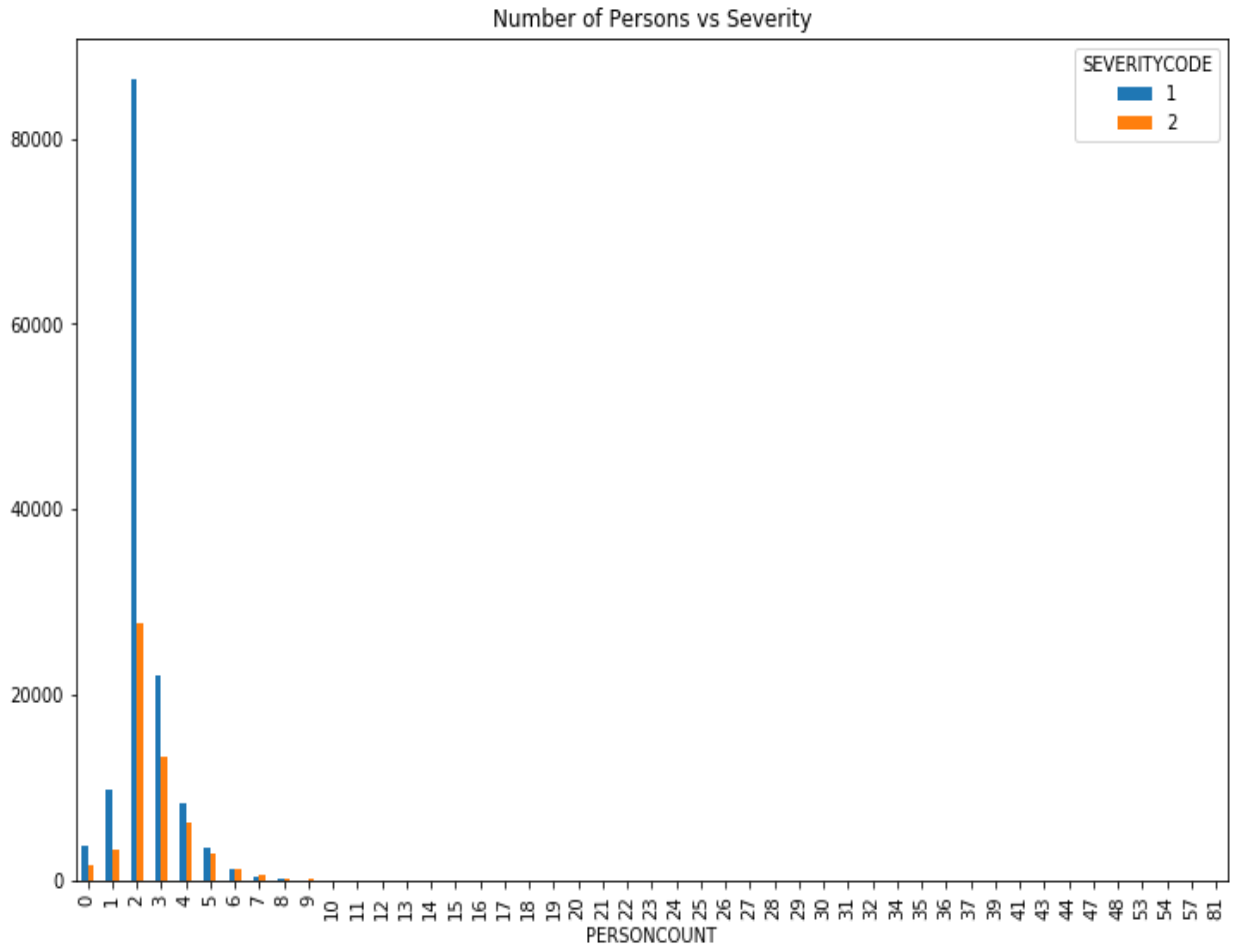

Figure.3 Visual representation of the relationship between number of persons and severity

The feature which denotes the total number of people involved in the collision in the dataset is known as PERSONCOUNT. From the visual representation of the relationship between number of persons and severity (Figure.3), it can be stated that car accidents involving 2 persons is quite severe, both on property damage and injuries. This visualization can also be an indicator to avoid driving alone or with a second person in the city as the maximum amount of car accident involves 2 persons. This indicator can be enhanced if there was an attribute whether the driver alone or with someone.

Relationship between Weather and Severity

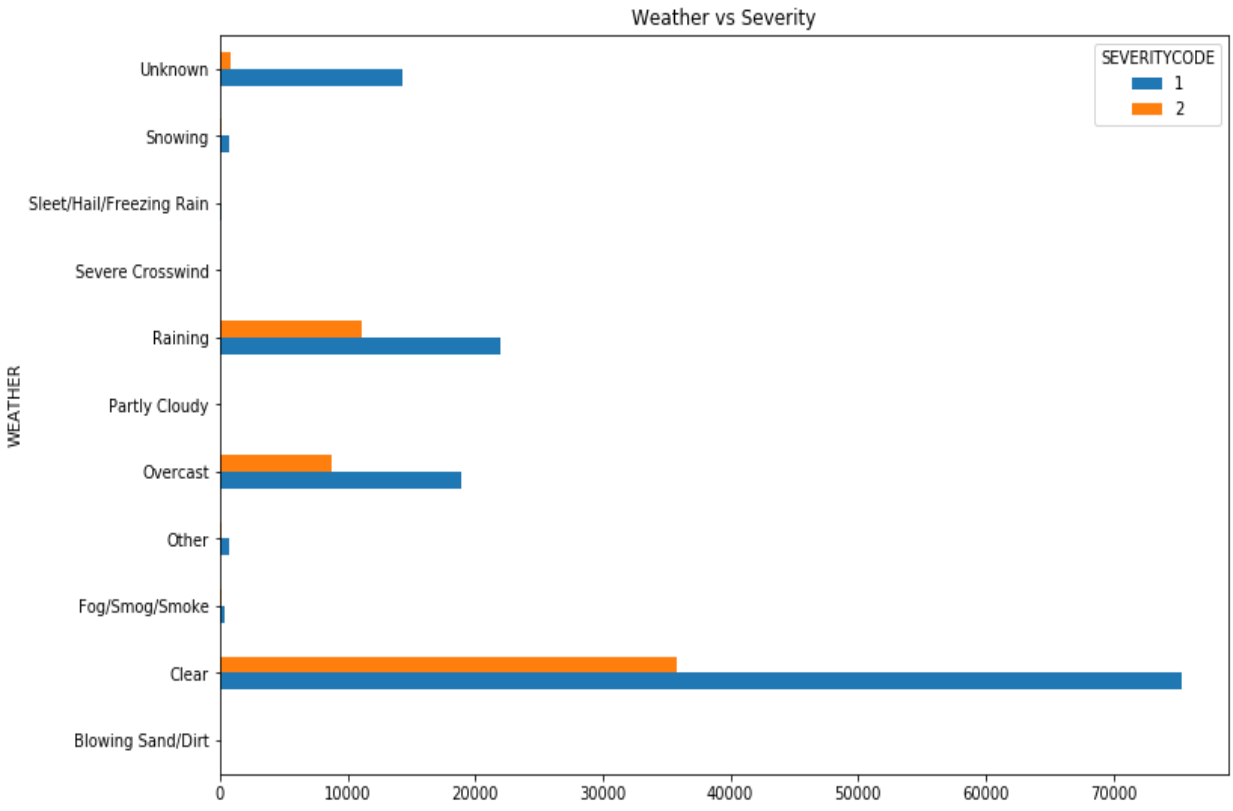

Figure.4 Visual representation of the relationship between Weather and severity

Weather is one of the most powerful attributes in this dataset to compare with car accident severity. It is categorical variable that provides a description of the weather conditions during the time of the collision. The visualization of this relationship of weather and severity provided rather an odd occurrence (Figure.4). From the visualization, it is quite clear that rather than in bad weather conditions, most car accidents occur in clear weather conditions. Both, the property damage and the incident of injury is highest in the clear weather conditions. This may be due to the low traffic movement during the bad weather conditions.

Relationship between Light Conditions and Severity

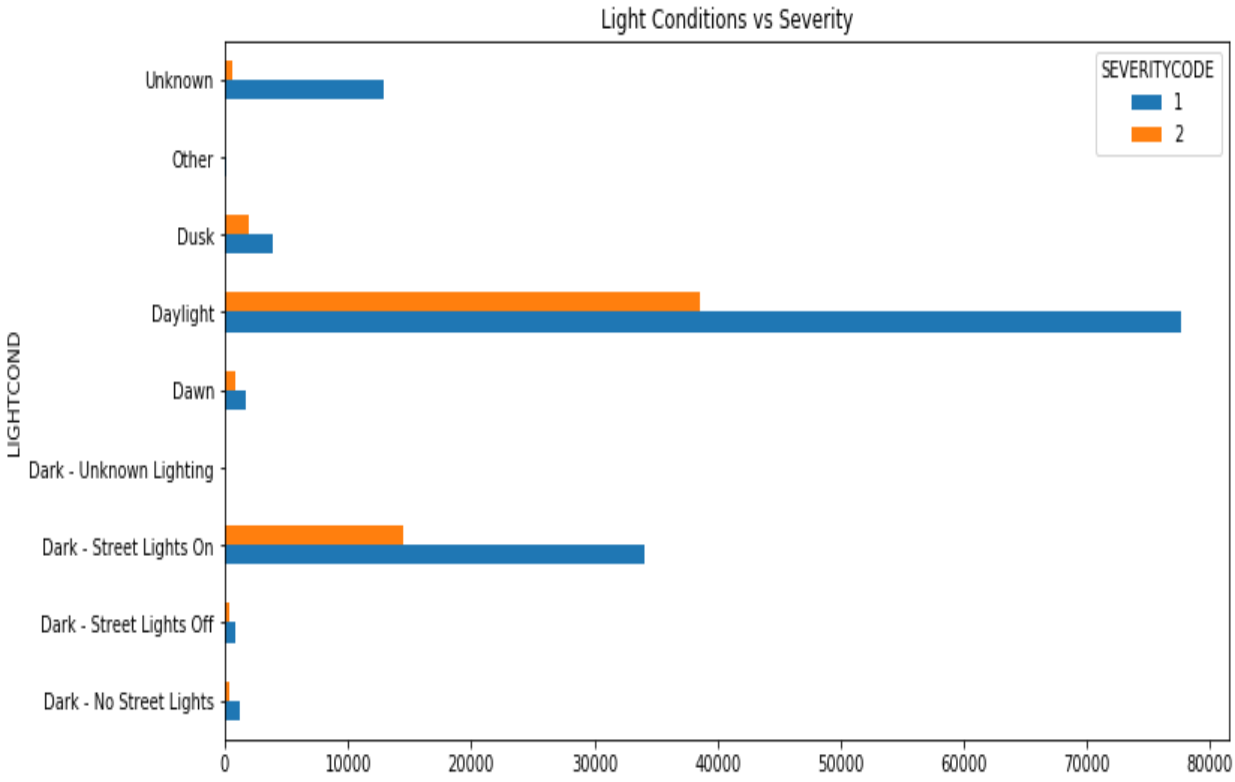

Figure.5 Visual representation of the relationship between Light conditions and severity

Light condition is a categorical feature that describes light conditions during the collision. This feature is quite important to compare with car accident severity. From Figure.5 the visual representation shows that most car accidents occur during the daylight. In addition, it can be also stated that property damage is quite high in daylight car accidents rather than in darker light conditions.

Relationship between Road Conditions and Severity

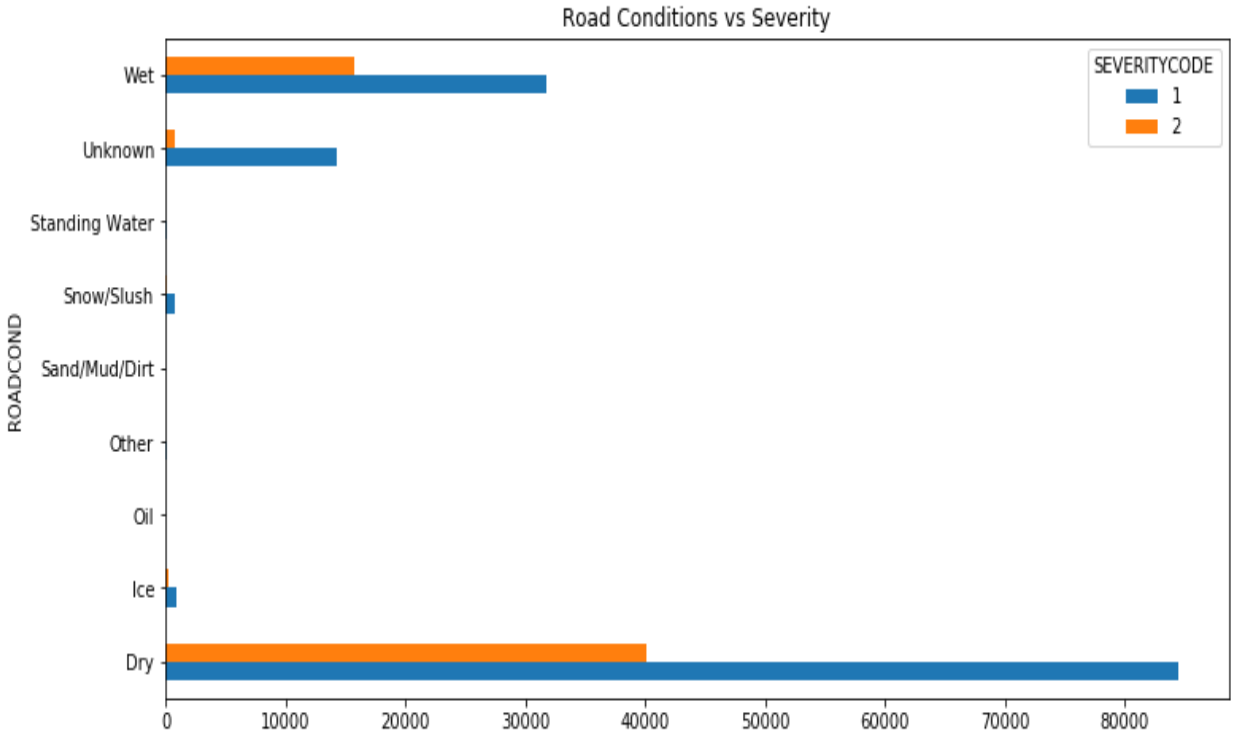

Figure.6 Visual representation of the relationship between Road conditions and severity

Road condition is another attribute that can be effectively plot with the target feature severity. This variable describes the condition of the road during the collision. From the visualization in Figure 6, it can be stated that car accidents happen in mainly two road conditions: wet and dry. In both cases the property damage was high. It can also be denoted that, most car accidents occurred in dry road conditions.