Predicting the Severity of Road Traffic Accidents

Dmitry Vasilyev

September, 2020

1. Introduction

1.1. Problem

Traffic accidents are a significant source of deaths, injuries, property damage, and a major concern for public health and traffic safety. Accidents are also a major cause of traffic congestion and delay. Effective management of accident is crucial to mitigating accident impacts and improving traffic safety and transportation system efficiency. Accurate predictions of severity can provide crucial information for emergency responders to evaluate the severity level of accidents, locating and eliminating accident blackspots, estimate the potential impacts, and implement efficient accident management procedures.

1.2. Interest

Model can be applied to predictions of accident severity which is an essential step in accident management process. By recognizing those key influences, this report can provide suggestive results for government and local authorities to take effective measures to reduce accident impacts and improve traffic safety.

2. Data acquisition and cleaning

2.1. Data source

Dataset includes detailed information about collisions provided by Seattle Police Department and recorded by Traffic Records. It includes geo location, address type, weather and light conditions, time of the day and many others. Our dependent variable is severity code, from metadata file codes are divided into five categories, however dataset contains only two – property damage(1) and injury(2). This is cumulative dataset with time frame from 2004 till present.

2.2. Data cleaning

Multiple issues were found while working with the dataset.

First of all, several categorical features such as Address Type 'ADDRTYPE' and Collision Type 'COLLISIONTYPE' had number of missing values, around 2.5%, that I decided to replace with mode values.

Second, categorical features Junction Type 'JUNCTIONTYPE', Weather 'WEATHER', Road 'ROADCOND' and Light Conditions 'LIGHTCOND' already had ‘Uknown’ values, so all missing values were replaced as ‘Unknown’ .

Third, Inattention 'INATTENTIONIND', Speeding 'SPEEDING', Hit Parked Car 'HITPARKEDCAR' and Pedestrian Right Of Way Was Not Granted ‘PEDROWNOTGRNT’ features were converted from categorical to numeric values: ‘NaN’ or ‘N’ into 0, and ‘Y’ into 1. Column Under Influence ‘UNDERINFL’ had mixed type of values (NaN, ‘N’, ‘Y’, 0, 1), that were all converted into numeric.

Fourth, feature that reflects the data and time ‘INCDTTM’ of the incident was converted into datatime64 format to extract data easier.

After fixing missing values I checked for outliers in the dataset. First feature I looked at was Number of Persons involved in the Collision: there were around 237 (<0.1%) observations with more than 10 people, so for simplicity any value higher than 10 will be equal to 10. Same way I applied filter for number of vehicles involved in the incident, but limited it to 5. For incidents that involved pedestrians (‘PEDCOUNT’) and bicycles ‘PEDCYLCOUNT’ I used feature binarization: so either any number was in collision or not.

2.3. Feature selection

After data cleaning there were 194673 samples and 38 features. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. For example, a Collision code provided by SDOT and State Collision code contain very similar information, moreover feature Collision Type and Address type already had sufficient information related to the incident classification. So I decided to keep last two features for the model.

Some features were not relevant, contained only unique values related to each incident.

Table 1. Simple feature selection after data cleaning

Kept Features	Dropped Features	Reason
Numeric features: 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT'	‘X’, ‘Y’, ‘OBJECTID’, ‘INCKEY’, ‘REPORTNO’, ‘INTKEY’, ‘SDOT_COLCODE’, ‘SDOTCOLNUM’, ‘ST_COLCODE’, ‘EXCEPTRSNCODE’, ‘CROSSWALKKEY’, ‘SEGLANEKEY’	Irrelevant or unique data that has no correlation with target variable.
Binary features: 'INATTENTIONIND', 'UNDERINFL', 'PEDROWNOTGRNT', 'SPEEDING', 'HITPARKEDCAR'	SEVERITYCODE.1	Duplicate data
Categorical features: 'ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'WEATHER', ROADCOND', 'LIGHTCOND', ‘LOCATION’	‘STATUS’, ‘EXCEPTRSNDESC’, ‘SEVERITYDESC’, ‘INCDATE’, ‘SDOT_COLDESC’	Duplicated data from other columns, or data is irrelevant.

One hot encoding technique was applied to all categorical features in order to have each features to be represented in a binary format.

3. Exploratory Data Analysis

3.1. Calculation of target variable

Severity code for each incident has multiple values according to the metadata file, however dataset contains only two values: 1 – any property damage, 2 – incidents that included injuries. Dataset contains 136485 observations with target value 1, and 58188 - with value 2. So we conclude that our dataset is imbalanced and we need to apply certain techniques to balance it before training the model. Our primary goal is to establish whether incident involved injuries (2) or not (1), so that local authorities can implement efficient accident management procedures..

3.2. Number of collisions in each location.

I was trying to find black spots using 'Location' feature, as certain addresses had more than one collision incident reported, so another column with number of incident at the location was added for all samples based on number of incidents at the location: 'NUMINCLOCATION'.
Distribution of collisions per location is shown below:

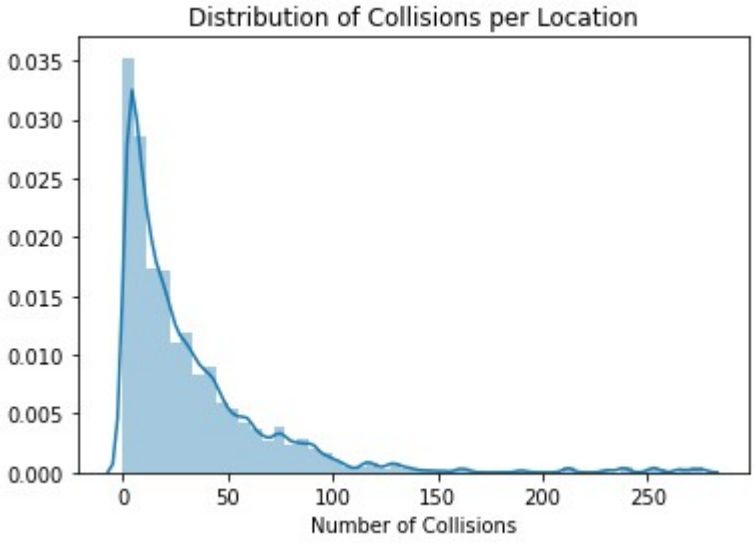

Figure 1. Distribution of Collisions per Location

3.3. Distribution of collisions during the week.

After creating another column 'DAYOFWEEK' containing Day of Week when incident happen, I was trying to find days with most collisions:

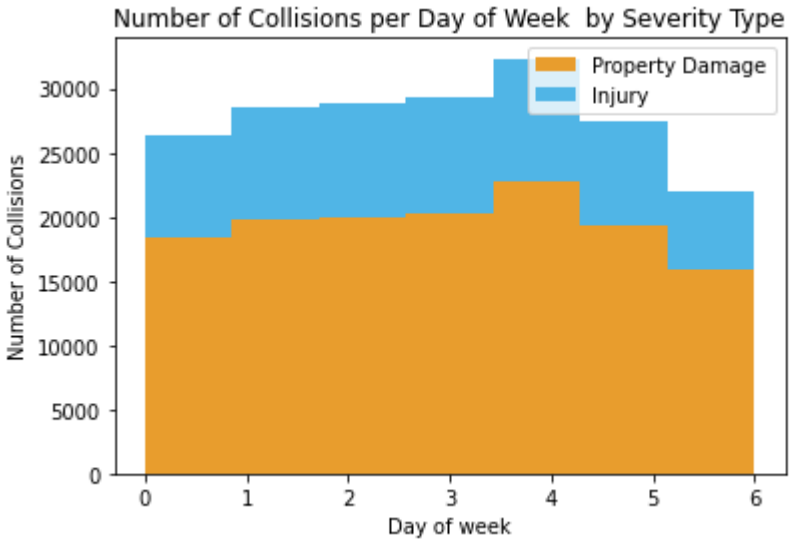

Figure 2. Distribution of Collisions During the Week

I decided to add another binary feature 'RISKDAY' for Fridays, by applying lambda function with x==4.

3.4. Distribution of collisions during the day.

After creating another column 'INCHOUR' containing hour of the day when incident happen, I was trying to find hours with most collisions:

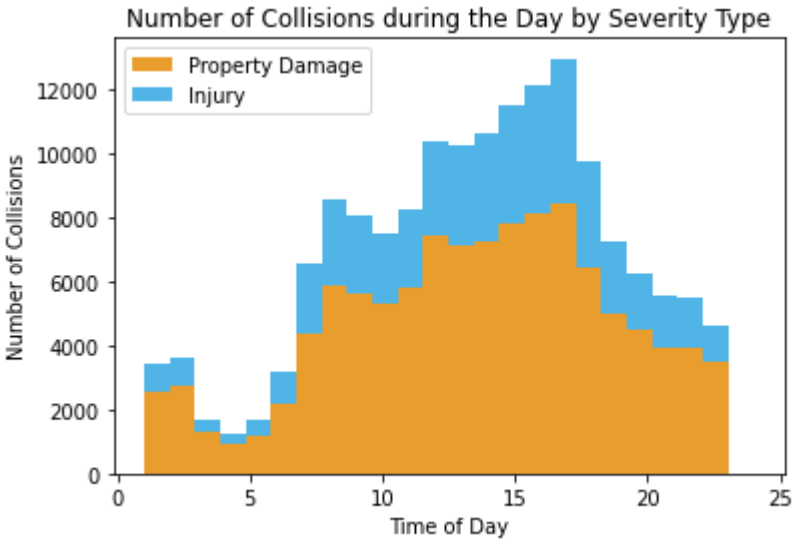

Figure 3. Distribution of Collisions During the Day

We see a spike of collisions during the day, so let's apply feature binarization for hours between 12:00 and 18:00. At the same time spike at midnight will not be taken into consideration as it also contains null values.

3.5. Number of persons involved in the collision by severity type

Majority of incidents that happened had 2 persons involved into the incident. Let’s see the histogram showing number of persons by severity:

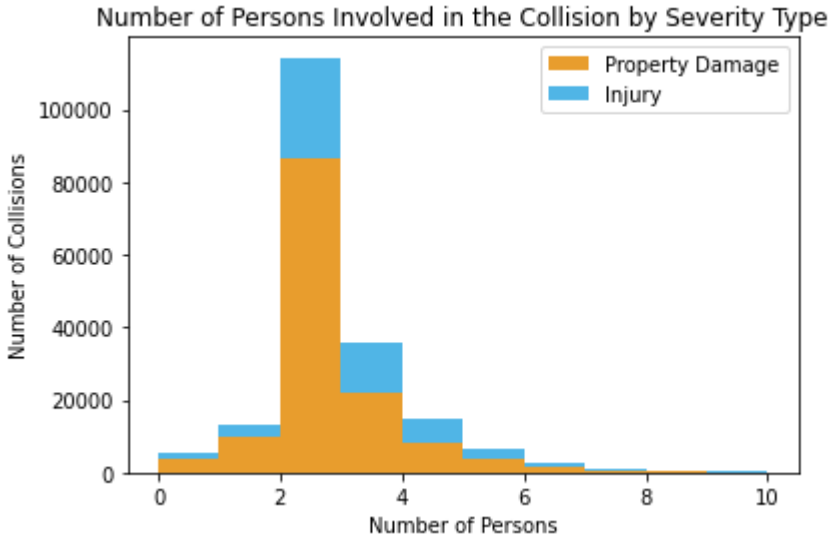

Figure 4. Number of Person Involved in the Collision by Severity Type

Let's assume that any value above 10 for 'Person Count' will be equal to 10 in order to remove outliers.

3.6. Number of pedestrians involved in the collision by severity type

As we see from the graph below feature binarization can be applied to that feature: so either pedestrians were involved in the collision or not.

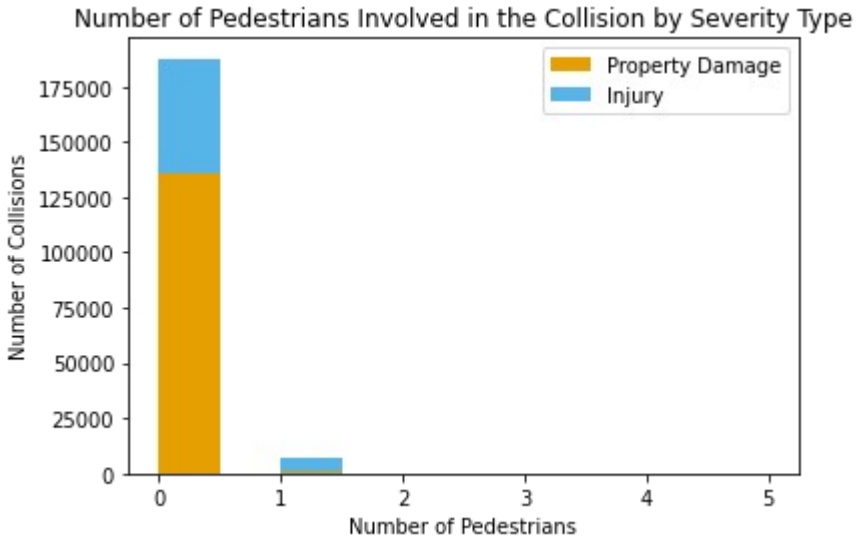

Figure 5. Number of Pedestrians Involved in the Collision by Severity Type

3.7. Number of bicycles involved in the collision by severity type

Same way we can apply feature binarization for Number of Bicycles involved in the collision column.

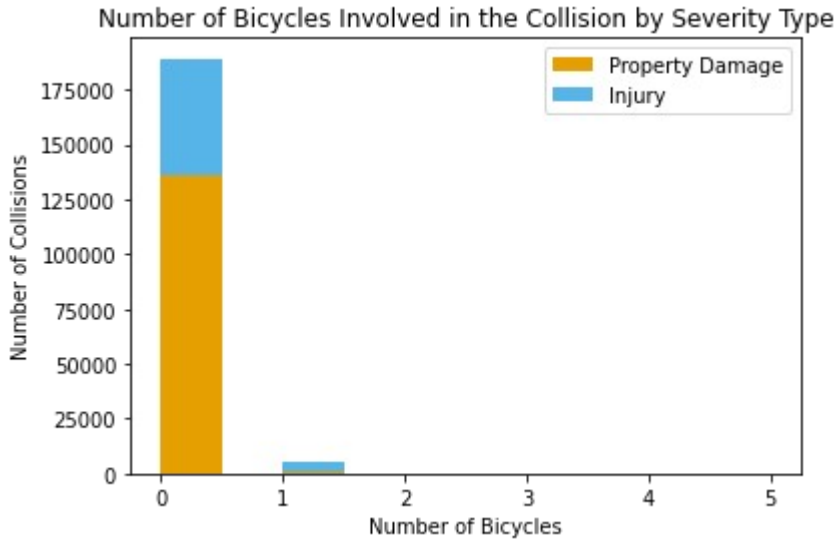

Figure 6. Number of Bicycles Involved in the Collision by Severity Type

3.8. Number of vehicles involved in the collision by severity type

Let’s see outliers in number of vehicles in the collision by severity type.

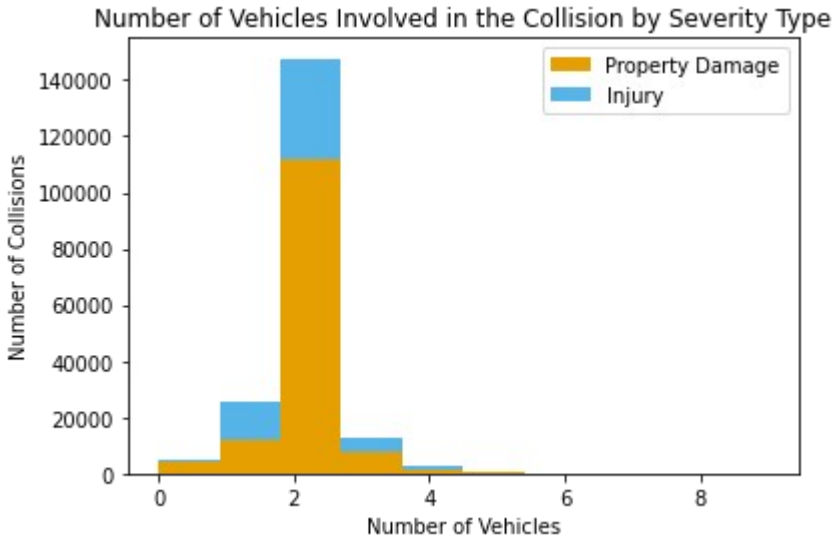

Figure 7. Number of Vehicles Involved in the Collision by Severity Type

We will assume that any value above 5 for 'Person Count' will be equal to 5 in order to remove outliers.