**Predicting the Severity of Accidents in Seattle**

**Shwetha Acharya**

**Oct 22, 2020**

1. **Introduction**
   1. **Background**

**In the recent past we have seen an increase in the number of road accidents in Seattle. While some of these accidents just cause property damage some are serious, resulting in heavy injuries and fatalities to drivers and passengers and pedestrians. With this trend, driving on highways and roads poses a risk to the lives of people. There may be different reasons for these accidents such as poor road conditions, light and weather conditions and so on. The idea is to use data science techniques to understand if we can predict the gravity of the accidents on account of collisions.**

* 1. **Problem**

**Data that might contribute to determining accident severity might include road conditions, light and weather conditions, car speeding, drunken driving. This project aims at predicting accident severity based on historic data.**

* 1. **Interest**

**This analysis and prediction will help drivers and passengers of various vehicles such as trucks, cars as well as pedestrians taking these road routes as it will serve as a safety warning to them. In addition, it would help traffic cops monitoring these areas – making them more vigilant and taking steps to avoid collisions and accidents, thus saving/securing lives of people.**

1. **Data Acquisition**
   1. **Data Sources**

**Data for this study is taken from Seattle Department of Transportation - Traffic Management Division, Traffic Records Group\*\*. It includes information on all types of collisions from 2004 till date (2020). The key features in the dataset encompass accident severity in terms of numeric codes, collision description, address type (whether intersection, alley or block), X and Y co-ordinates, number of vehicles and pedestrians involved, light, road and weather conditions, speeding, whether driving under the influence**.

* 1. **Data Cleaning**

**Our target variable is the accident " severity" in terms of human fatality, traffic delay, property damage and hence SEVERITYCODE is the target variable – y. There are 37 features - 'X', 'Y', 'OBJECTID', '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'**

**On observing, we see that following columns can be dropped for the below reasons**:

* **EXCEPTRSNCODE, EXCEPTRSNDESC – Metadata doesn’t explain much about these columns. Also, columns have many null values and do not have any impact on the prediction.**
* **SEVERITYCODE.1 is a duplicate column**
* **OBJECTID and REPORTNO are just unique identifiers or keys to identify every row and collision report and no impact on accident predictions.**
* **SEVERITYDESC is just the accident description - Numerical values are captured in SEVERITYCODE.**
* **INCDATE - It is a subset of INCDTTM**
* **INCKEY and COLDETKEY – These are unique keys for the incident to identify collision**
* **On inspecting SDOT\_COLCODE, SDOT\_COLDESC, ST\_COLCODE, ST\_COLDESC**

1. **ST\_COLCODE and ST\_COLDESC -refer to code provided by the state that describes the collision**
2. **SDOT\_COLCODE and SDOT\_COLDESC – refer to code given to the collision by SDOT.**

* **One of the above can be dropped as it is a code describing the same collision but given by 2 authorities - SDOT/State. Given that, ST\_COLCODE has some null values when compared to SDOT\_COLCODE, ST\_COLCODE and ST\_COLDESC can be dropped. Also, SDOT\_COLDESC - description of SDOT\_COLCODE, can be dropped**
* **LOCATION – Location/address is given with X and Y co-ordinates/latitude-longitude and hence location can be dropped**
* **Also, more than 95% of data is missing for ‘SPEEDING’ and ‘PEDESTRIAN-RIGHT-OF-WAY-NOT-GRANTED’ and hence they can be dropped**
* **INTKEY can be dropped too, as it is a key to represent address type - “Intersection” of a given location.**

**X and Y represent latitudes and longitudes of the accident sites. DBSCAN clustering algorithm is used to cluster them into zones.**

**Categorical variables such as Address Type, Junction Type, Weather, Road and Light condition are converted to numeric values. INCDTTM is a date column and fine grained to classify into weekend and non-weekend days.**

**After cleaning up this data, the above set of refined attributes is put through feature selection to pick only the relevant features.**

* 1. **Feature Selection**

|  |
| --- |
| HITPARKEDCAR |
| SDOT\_COLCODE |
| CROSSWALKKEY |
| INCDTTM |
| VEHCOUNT |
| PEDCYLCOUNT |
| PERSONCOUNT |
| PEDCOUNT |
| SEGLANEKEY |
| UNDERINFL |
| COLLISIONTYPE |
| ROADCOND |
| WEATHER |
| LIGHTCOND |
| (X,Y) - ZONES |
| JUNCTIONTYPE |

**\*\*Data used for this analysis is taken from coursera courtesy SDOT Traffic Dept**