### Capstone Report: Predict Accident Severity

### Lorena Wang

### 1. Introduction

Road traffic injuries and deaths has been a global severe problem. According to the statistic report from World Health Organization (WHO):

* Every year 1.35 million people are killed as a result of a road traffic crash.
* Between 20 and 50 million people suffer non-fatal injuries, which most can lead to disabilities in their life.
* Road traffic injuries are estimated to be the eighth leading cause of death globally for all age groups and the leading cause of death for children and young people aged 5–29 years old. More people now die in road traffic crashes than from HIV/AIDS.

This severe problem needs our attention, since human lives are irreplaceable. Thanks to the development of data science, we can get the insight of the traffic collision data and predict the severity of car accidents based on the complex various factors, such as weather, road condition, light condition, speeding etc. Machine Learning is an ideal method as this is a scientific approach for modelling and predicting the parameter of interest demanding only a low budget. We can also identify which factors have more impact on this problem and people can take actions to them.

This project is aimed to use Machine Learning to build models to predict the severity of the traffic accidents according to the factors in the car collision data in the city of Seattle, USA. This predictive model can be applied to multiple practical conditions to save life, such as safe route planning, emergency signal light control, vehicle allocation, signals placement or even AI car design.

This study of car accidents in Seattle can help other cities in similar conditions, and the methodology can be used widely among other countries. The stakeholders of the present problem involve state and local government agencies, non-governmental organizations, regional authorities, AI technology companies, and even individuals.

### 2. Data acquisition and Cleaning

### 2.1 Data source

The car collision data is obtained from Seattle Govt’s website, which records the severity of the accident and the current factors, including road condition, light, speed, weather, etc. The detailed information can be checked in Data Attributes.

Data Source:

<http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data>

Data Catalogue:

<https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf>

Data Attributes:

| **Attribute** | **Data type, length** | **Description** |
| --- | --- | --- |
| OBJECTID | ObjectID | ESRI unique identifier |
| 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, 255 | Description of the general location of the collision |
| EXCEPTRSNCODE | Text, 10 | Not specified |
| EXCEPTRSNDESC | Text, 300 | Not specified |
| 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 |
| SEVERITYDESC | Text | A detailed description of the severity of the collision |
| COLLISIONTYPE | Text, 300 | 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. |
| PEDCYLCOUNT | 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, 30 | The date and time of the incident. |
| JUNCTIONTYPE | Text, 300 | Category of junction at which collision took place |
| SDOT\_COLCODE | Text, 10 | A code given to the collision by SDOT. |
| SDOT\_COLDESC | Text, 300 | A description of the collision corresponding to the collision code. |
| INATTENTIONIND | Text, 1 | Whether or not collision was due to inattention. (Y/N) |
| UNDERINFL | Text, 10 | Whether or not a driver involved was under the influence of drugs or alcohol. |
| 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. |
| PEDROWNOTGRNT | Text, 1 | Whether or not the pedestrian right of way was not granted. (Y/N) |
| SDOTCOLNUM | Text, 10 | A number given to the collision by SDOT. |
| SPEEDING | Text, 1 | Whether or not speeding was a factor in the collision. (Y/N) |
| ST\_COLCODE | Text, 10 | A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary. |
| ST\_COLDES | Text, 300 | A description that corresponds to the state’s coding designation. |
| SEGLANEKEY | Long | A key for the lane segment in which the collision occurred. |
| CROSSWALKKEY | Long | A key for the crosswalk at which the collision occurred. |
| HITPARKEDCAR | Text, 1 | Whether or not the collision involved hitting a parked car. (Y/N) |

### 2.2 Data Cleaning

### Data cleaning is one of the most important steps to make sure an optimal result in the data insight analysis and the final predicting models. To explore the date and time feature in traffic collisions, first the INCDTTIME is transformed to date type and the Month, Weekday, Hour information is generated. Then the key variables are dropped, since they cannot be used to predict traffic collision severity. Some duplicate variables are also dropped. For now, the data set is ready for the exploratory data analysis, which we can use data statistics and visualization methods to identify the relationship in each factor and the severity code of the traffic collision.

However, we still need to preprocess the data set for building the predictive models. The serious types of the traffic collision account for only a small percentage, which means the data set is highly skewed. This will result in the prediction bias problem since the machine learning algorithms will get more train on the less severity types of traffic collision and fails to identity the minority classes. To solve this problem, we can under-sample the majority class and over-sample the minority class to balance the data. Here each severity level types of collision are sampled to 10000.

### In the predicting model section, this project adopted two kinds of targets to build the predictive models. Multi-Class predictive model use the four targets in the original dataset. Two-Class predictive model combines the 1,2-severity code collision to the first target, which are less serious traffic collisions resulting in prop damage and injury. The second target combines the 2b,3-severity code collision, which are much more serious types including serious injury and fatality.