### 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.

### 3. Exploratory Data Analysis

### The preprocessed data for exploratory contains 43776 traffic collisions in total. In the data cleaning process, the initial severity code has been transformed to 1, 2, 3 (initial 2b), 4 (initial 3). The new mapping is as follows:

|  |  |
| --- | --- |
| SEVERITYCODE | SEVERITYDESC |
| 1 | Property Damage |
| 2 | Injury |
| 3 | Serious Injury |
| 4 | Fatality |

Table 1. Transformed Severity Code

The 1-severity traffic collisions account for around 69% in the total samples. The 2- severity traffic collisions are around 40% of the 1-severity traffic collisions, make up 29% in the total samples. The 3-severity and 4-severity traffic accidents only occupy 1.6% and 0.2% of the total samples separately.

A screenshot of a cell phone

Description automatically generated

Figure 1. Count of different severity of traffic collisions

The Python folium library is used to see the geographic details of the Seattle traffic collisions. In this step, 100 traffic accidents are selected randomly and the Seattle map with those traffic accidents address is created as below.

A close up of a map

Description automatically generated

Figure 2. The Seattle traffic collisions map

I also tried to add labels to each spot, so that we can easily identify which type the specific traffic collision is on the map.

A close up of a map

Description automatically generated

Figure 3. The Seattle traffic collisions map with labels

Then the weather condition in these traffic accidents are analyzed. The distribution of weather conditions in all the traffic collision samples are showed in the pie chart below. The majority type of weather when traffic collisions happened is clear, account for 59.2% in the total samples, while the precipitation weather only takes up 18.4%, which raining takes hold of 18% in total. Comparing the traffic weather condition to the average Seattle climate can help us identity the relationship with weather and traffic collisions more clearly. On average, there are 152 sunny days per year in Seattle, occupying 41.6% in the total weather condition. Besides, Seattle has average 155 days of precipitation each year, including rain, snow, sleet and hail that falls to the ground, dominating 42.5% days in a year.

It’s really amazing to see that the probability of a car accident happens on clear days is 17.6% higher than the annual probability of clear days, while the probability of a car accident happens on precipitation weather is only 43% of the annual precipitation probability. Figure 5 shows the count of each type of traffic collisions in different weather condition. We can see that the clear weather dominates the most count in every severity type of collisions.

(Seattle Climate data source: <https://www.bestplaces.net/climate/city/washington/seattle#:~:text=The%20US%20average%20is%2028,average%2C%20155%20days%20per%20year.>

).

A close up of a logo

Description automatically generated

Figure 4. Weather condition of the traffic collisions

A close up of a logo

Description automatically generated

Figure 5. Count of different severity types of traffic collisions by weather.

Since we have transformed the date and time, we can see the statistics of four types traffic accidents by Month, Weekday and Hour. There is little gap in the distribution of traffic collisions by months. The least number of traffic accidents is in February, while the most is in April. In the Weekday statistics, the most traffic accidents happen in Friday. In the condition of hours statistics for traffic accidents, 0-1am shows an extremely high traffic collision rate, almost 5 times of the second most traffic collision hour: 17-18 pm. 3-4am, 4-5am, and 5-6am shows the similar lowest count of traffic collisions.

A close up of a fence

Description automatically generated

Figure 6. Count of traffic collisions by month

A close up of a sign

Description automatically generated

Figure 7. Count of traffic collisions by weekday

A screenshot of a cell phone

Description automatically generated

Figure 8. Count of traffic collisions by hour

We can also visualize the count of traffic accidents with different types of the address type. The Block type takes hold of 66.8% of the total traffic accidents, while Intersection and Alley counts for 32.2% and 0.53% separately. One thing worthy to be noticed is that even though the 1-severity traffic accidents are much higher likely to occur in block address, the number of 2 and 3-severity traffic accidents are similar in intersection and block address.

A screenshot of a cell phone

Description automatically generated

Figure 9. Count of traffic collisions by address type

Only 1% of the drivers involved was under the influence of drugs or alcohol in the total traffic collisions. But around 17.3% of the drivers involved in the serious traffic collisions (except for 1-severity traffic collisions) were under the influence, which means the traffic accident with under influenced drivers are more likely to be a serious type.

A screenshot of a social media post

Description automatically generated A picture containing screenshot

Description automatically generated

Figure 10. Count of traffic collisions by driver under influenced condition

There are 97.6% of the traffic accidents are in the condition with the pedestrian right of way granted in the total samples. Ungranted the pedestrian right of way shows an extremely low rate in 1-severity type of traffic accidents, which means this condition can decrease the probability of traffic accidents, especially 1-severity type.

A picture containing screenshot

Description automatically generated

Figure 11. Count of traffic collisions by PEDROWNOTGRNT

94.4% of the total traffic collision samples are not caused by the speeding factor. The rate of serious type traffic collision samples (expect 1-severity type) caused by speeding factor is higher than the 1-severity type collision.

A picture containing screenshot

Description automatically generated

Figure 12. Count of traffic collisions by if caused by speeding factor

### 4. Predictive Modeling

### In this project, classification machine learning algorithms are used to predict severity of the traffic collisions. Logistic Regression, K Neighbors Nearest (KNN), SVM and Decision Tree algorithms are used to classify the traffic collisions into the initial four severity types and the transformed two classes. In the following section, the accuracy of each algorithms is compared.

### 4.1 Multi-Class Predictive Modeling

### In Multi-Class predictive modeling, the decision tree classification has the highest test accuracy: 70.83%. KNN method has the second highest accuracy with 67.31%, when the predicted target is compared to 7 nearest neighbours.

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| LR | 56.23% |
| KNN | 67.31% |
| SVM | 65.59% |
| Decision Tree | 70.83% |

Table 2. Test accuracy for different algorithms

**A close up of a logo

Description automatically generated**

Figure 13. The accuracy of KNN algorithm with different number of neighbors (k)

Here shows the fi-score and confusion matrix of each algorithm:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Severity/ f1-score | LR | KNN | SVM | Decision Tree |
| 1 | 0.62 | 0.62 | 0.64 | 0.64 |
| 2 | 0.49 | 0.54 | 0.53 | 0.59 |
| 3 | 0.43 | 0.67 | 0.60 | 0.71 |
| 4 | 0.68 | 0.86 | 0.85 | 0.88 |

Table 3. f1-score for LR, KNN, SVM and Decision Tree in Multi-Class Predictive Models

**A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated**

Figure 14. Confusion matrix for LR, KNN, SVM and Decision Tree in Multi-Class Predictive Models

We can also see the feature importance in using the decision tree classifier. Besides the street collision code, the top 5 important features in Multi-Class Predictive Models are Under Influenced, Speeding, Light Condition, Junction Type and PedRoWNotGrnt.

A close up of a logo

Description automatically generated

Figure 15. Feature importance using Decision Tree in Multi-Class predictive model

### 4.1 Two-Class Predictive Modeling

### In Two-Class predictive modeling, the decision tree classification also has the highest test accuracy: 83.68%. KNN method has the second highest accuracy with 81.85%, when the predicted target is compared to 3 nearest neighbours.

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| LR | 75.88% |
| KNN | 81.15% |
| SVM | 78.56% |
| Decision Tree | 83.68% |

Table 2. Test accuracy for different algorithms

A close up of a logo

Description automatically generated

Figure 16. The accuracy of KNN algorithm with different number of neighbors (k)

Here shows the fi-score and confusion matrix of each algorithm:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Severity/ f1-score | LR | KNN | SVM | Decision Tree |
| 0 | 0.77 | 0.80 | 0.79 | 0.83 |
| 1 | 0.75 | 0.82 | 0.77 | 0.84 |

Table 4. f1-score for LR, KNN, SVM and Decision Tree in Two-Class Predictive Models

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

Figure 17. Confusion matrix for LR, KNN, SVM and Decision Tree in Two-Class Predictive Models

We can also see the feature importance in using the decision tree classifier. Besides the street collision code, the top 5 important features in Multi-Class Predictive Models are PedRoWNotGrnt, Speeding, Under Influenced, INATTENTION and Road Condition.

A close up of a logo

Description automatically generated

Figure 15. Feature importance using Decision Tree in Two-Class predictive model

**5. Conclusion**

This project analysed the relationship in main factors (weather, month, weekday, hour, address type, drivers’ condition, speeding and pedestrian right to way) and the severity of traffic collisions. It also visualized the traffic accidents on the Seattle map. The predictive models are built in LR, KNN, SVM and Decision Tree algorithms. Besides the original severity code degree, this project also makes a two-class level severity code. We can conclude that Decision Tree has the best predicting accuracy in Multiple-Class and Two-Class models, with 70.83% and 83.68% accuracy separately. The top factors in the decision tree algorithm are also visualized in this project.