Analysis on Car Accident Severity in Seattle based on Machine Learning Approach

Seohyun Lee

Abstract

Predicting the severity of the car accident is an important task to provide essential information not only to local drivers but also traffic authority and insurance designer. In this work, based on annual collision data in Seattle, severity of injury analysis model was built using machine learning approach. After applying various machine learning algorithm including K-nearest neighbors, decision tree model, and support vector machine (SVM) with radial basis function (RBF), the test dataset accuracy was evaluated. It is expected that the severity of car collision can be estimated with the prediction accuracy as high as approximately 80 percent, using various environment conditions, as this study revealed.

Introduction

Although New York City is known as one of the most congested cities in the United States, Seattle is tied with New York as the fourth worst traffic congestion in the country, according to TomTom Index Traffic Report in 2014 (TomTom.com 2014). Urban Mobility Report (Lasley 2019) released in 2019 by Texas A&M University's Transportation Institute also analyzed that Emerald City is suffered from the worst traffic congestion, which causes 3.1 billion dollars of annual cost. Due to such traffic congestion, a crash occurred every 4.5 minutes and a person died in a crash every 16 hours in Seattle, as reported in Annual Collision Data Summary Reports by Washington State Department of Transportation in 2015 (WSDOT 2015).

In order to reduce injuries caused by car accidents, it is required to understand the relationship between severity of the injury and the contributing factors to the car collision severity (Rifaat and Chin 2007), including surrounding conditions such as road condition, light condition, and weather, as well as driving situation such as speeding and junction types.

In this report, based on car collision data in Seattle, the severity of the car accident is analyzed with various attributes in each crash case to build a model predicts the severity based on the contributing factors affecting severity. This study aims to provide essential information particularly to local drivers and insurance designers, by clarify

© Seohyun Lee

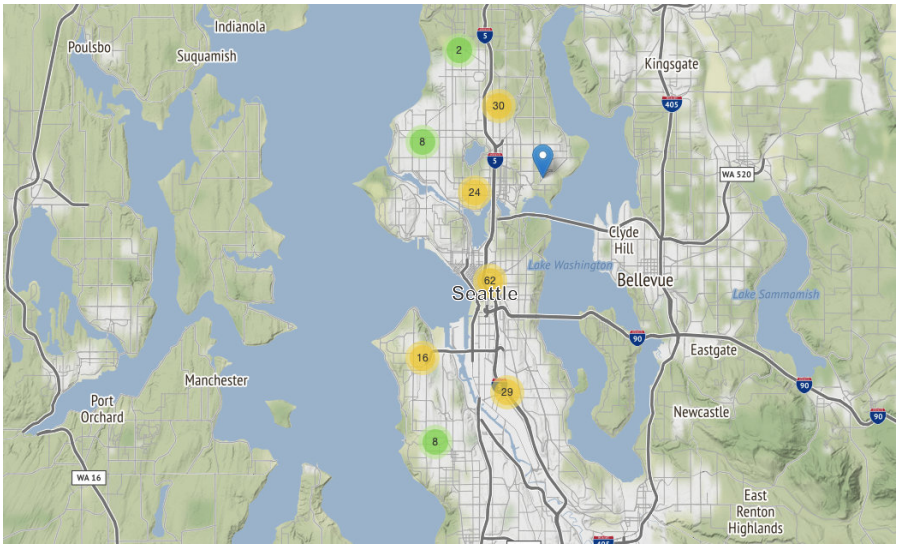

FIGURE 1. Car accident map in Seattle 2015. Markers are clustered by the number of incidents.

the relationship between various driving conditions and severity of injuries that can be caused by possible car accidents.

Data

The dataset exploited in this project was acquired from here. The raw data comprise 194673 of rows and 38 of columns, which can be interpreted as 194673 of cases and 38 of attributes in individual case. The attributes includes latitude and longitude information of the incident, address type, coolision type, identification number, road condition, light condition, speeding, etc. Additionally, the severity of each collision was labeled from **0** to **3** as SEVERITYCODE, where **0** stands for unknown, **1** indicates property damage, **2** represents injury (**2b** is serious injury), and **3** means fatality. All rows were labeled with the SEVERITYCODE, but either **1** or **2**.

Data Cleaning

Because the raw dataset includes duplicated columns and null elements, it is required to clean the dataset by removing some rows and columns that might hinder from building an accurate model.

First of all, the dataset was trimmed to contain only following attributes: SEVERITYCODE, ADDRTYPE, COLLISIONTYPE, PERSONCOUNT, VEHCOUNT, JUNCTIONTYPE, ROADCOND, LIGHTCOND, WEATHER, which respectively

stands for severity of the collision, address type, collision type, number of person involved in the accident, number of vehicles involved in the accident, type of junction, road condition, light condition, and weather condition. Except the column named SEVERITYCODE, the attributes can be considered as independent properties. The dependent property is SEVERITYCODE, which is the label of each collision record, which should be predicted after building a model.

Second, the rows containing null elements are removed, since null element hardly contribute to constructing an accurate model.

Finally, because some attributes includes 'Unknown' or 'Other' element, the rows containing such element were also removed.

First five rows of the trimmed dataset are as shown in Table 1.

SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	VEHCOUNT
2	Intersection	Angles	2	2
1	Block	Sideswipe	2	2
1	Block	Parked Car	4	3
2	Intersection	Angles	2	2
1	Intersection	Angles	2	2

JUNCTIONTYPE	ROADCOND	LIGHTCOND	WEATHER
At Intersection (intersection related)	Wet	Daylight	Overcast
Mid-Block (not related to intersection)	Wet	Dark - Street Lights On	Raining
Mid-Block (not related to intersection)	Dry	Daylight Car	Overcast
At Intersection (intersection related)	Wet	Daylight	Raining
At Intersection (intersection related)	Dry	Daylight	Clear

TABLE 1. First five rows of the dataset after data cleaning

Data Exploration

To better understand the dataset, distributions of each attribute were depicted using various types of graph. In order to count the number of properties of individual column, bar or pie graph can help figure out the property distribution. This step is an essential process to determine which model is the most appropriate for given dataset.

From the dataset shown in Table 1, except for the SEVERITYCODE which is the target attribute, there are two types of attributes: the columns with categorical values (ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, ROADCOND, LIGHTCOND, and WEATHER) and the columns with numbers (PERSONCOUNT and VEHCOUNT). Because the data type of former attributes is string while the later data type is integer, it is required to apply different visualization method to avoid error. For example,

since the ADDRTYPE has string type data as elements, bar graph or pie graph is appropriate method, rather than box-whisker plot. In contrast, PERSONCOUNT can be analyzed with box-whisker plot, where convenient statistical values including mean and interquartile range (IQR) are easily recognizable.

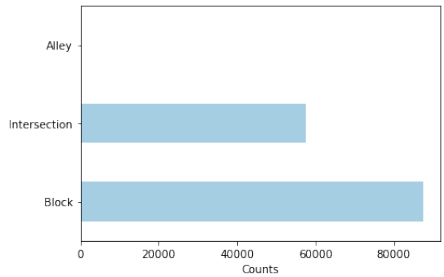

FIGURE 2. Property distribution of ADDRTYPE

First, from the bar graph of ADDRTYPE as shown in Fig. 2, it is clear that the attribute named ADDRTYPE has three different properties, Alley, Intersection, and Block, and most of the accidents took place at intersections and blocks, rather than alleys.

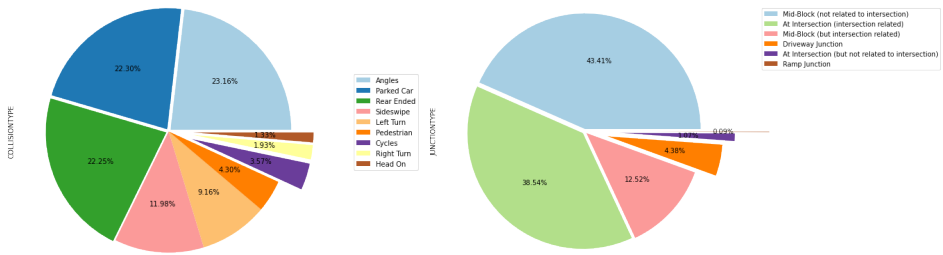

FIGURE 3. Property distribution of COLLISIONTYPE (left) and JUNCTIONTYPE (right).

Then, the distribution of categorical values in COLLISIONTYPE and JUNCTIONTYPE attributes can be visualized using pie chart as shown in Fig. 3, to figure out the number of sub-properties and the occupancy of them. In COLLISIONTYPE which represents the type of car collision, three major collision types are revealed as Angles, Parked Car, and Rear ended, followed by Sideswipe, Left turn, Pedestrian, Cycles, Right turn, and Head on. In JUNCTIONTYPE which indicates the type of junction where car collision took place, two major places are Mid-Block (not related to intersection) and Intersection (intersection related), followed by Mid-block (but

intersection related), Driveway juction, Intersection (but not related to intersection), and Ramp juction.

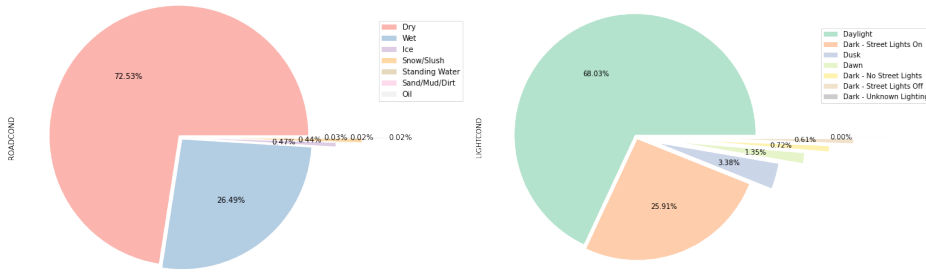

FIGURE 4. Property distribution of ROADCOND (left) and LIGHTCOND (right).

Likewise, ROADCOND, LIGHTCOND can be visualized with pie chart as well. As shown in Fig. 4, in ROADCOND which refers to the condition of road when the car collision happened, over 72% of the road was turned out to be dry while around 26.5% of car accidents took place on wet road condition. Ice, Snow/Slush, Standing Water, Sand/Mud/Dirt, and Oil were trivial cases. Additionally, LIGHTCOND which means the light condition at the incident showed that Daylight covers aobut 68% of total cases, followed by Dark - Street Lights On which occupies approximately 26%. Dusk, Dawn, Dark - No Street Lights, Dark - Street Lights Off, Dark - Unknown Lighting followed but with very small portions.

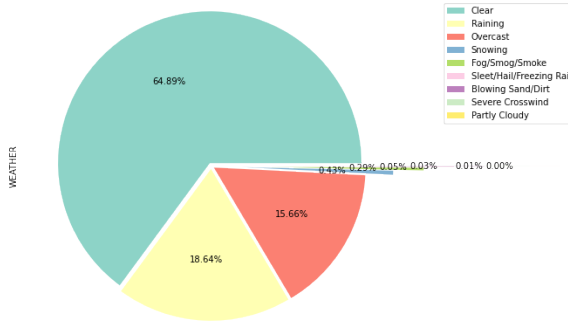

FIGURE 5. Property distribution of WEATHER

Furthermore, WEATHER attributes which indicates the weather condition on the day of the car collision showed three major properties as shown in Fig. 5, where

Clear occupies with almost 65% of total case, followed by Raining with 18.6% and Overcast with approximately 16%. Other weather conditions such as Snowing, Fog/Smog/Smoke, Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, Severe Crosswind, and Partly Cloudy followed but with the portions smaller than 1%.

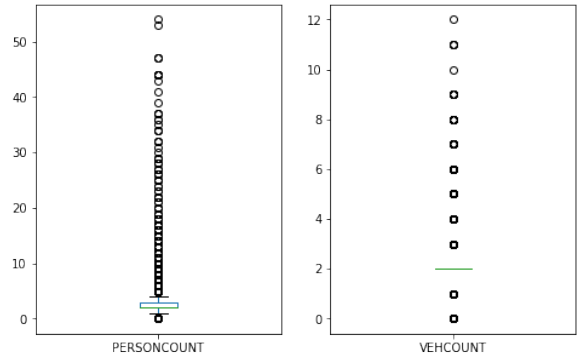

FIGURE 6. Property distribution of PERSONCOUNT (left) and VEHCOUNT (right)

Finally, the attributes the element of which consist of integers instead of string data type, PERSONCOUNT and VEHCOUNT which respectively refer to the number of people and number of vehicles involved in a single collision, was explored by using box-whisker plot.

As shown in Fig. 6, the number of people involved in an individual car accident is approximately 2.61 with standard deviation of 1.38, and some outliers are distributed to 54. In case of the number of vehicle, most of the values are densely aggregated near 2 with mean of 2.05 and standard deviation of 0.53 accompanying outliers up to 12, implying that the majority of collisions were recorded as the accidents between two cars.

Methodology

In order to build a model that can predict the severity of individual car collision case, the dataset was split into training dataset and test dataset with 7 : 3 ratio, to test the model accuracy after training.

The programming language used in overall building and training is Python 3.7, and the package exploited in training is scikit-learn version 0.20.3. Additionally, mac OS Catalina version 10.15 with graphic cards of Radeon Pro 560X 4 GB and Intel HD Graphics 630 1536 MB was used in training the model.

Analysis

K-Nearest Neighbors

One method that can be applied to build prediction model for collision dataset is K-nearest neighbors. To apply this method, it is required to calculate the best K value which maximize the accuracy score in prediction training. From the number 1 to 10, the average accuracy score and standard deviation of the accuracy was calculated to determine the appropriate K value for building a K-Nearest Neighbors based training model. Figure 7 shows the computed accuracy which is dependent of K.

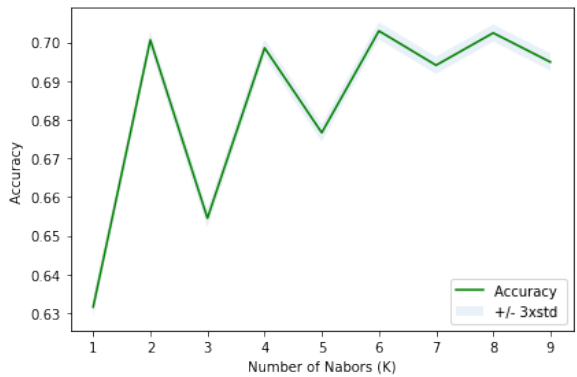

FIGURE 7. Accuracy change over the value K

From the K-accuracy graph, it is recognizable that K=6 produces the best accuracy among the value from 1 to 10. Therefore, the dataset was trained to find 6-nearset neighbors.

Decision Tree

Since the most of the attributes in dataset are categorical value the data type of which is string, decision tree model can be considered as a promising method to predict severity of the collision. In order to apply decision tree model to the trimmed dataset, the categorical type of attributes were encoded as integer label as a pre-processing.

Therefore, the attributes named ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, ROADCOND, LIGHTCOND, and WEATHER were transformed to have integer values to fit the original string data, while the columns named PERSONCOUNT and VEHCOUNT remained as is. Then, the transformed attributes can be prepared as an array of integers, which plays a role as a training dataset X. The SEVERITY column is now training dataset y, which is the target of training dataset X.

Using Decision Tree Classifier function with max depth of 4, the decision tree model can be depicted as shown in Fig. 8, when 30% of dataset was treated as a test dataset.

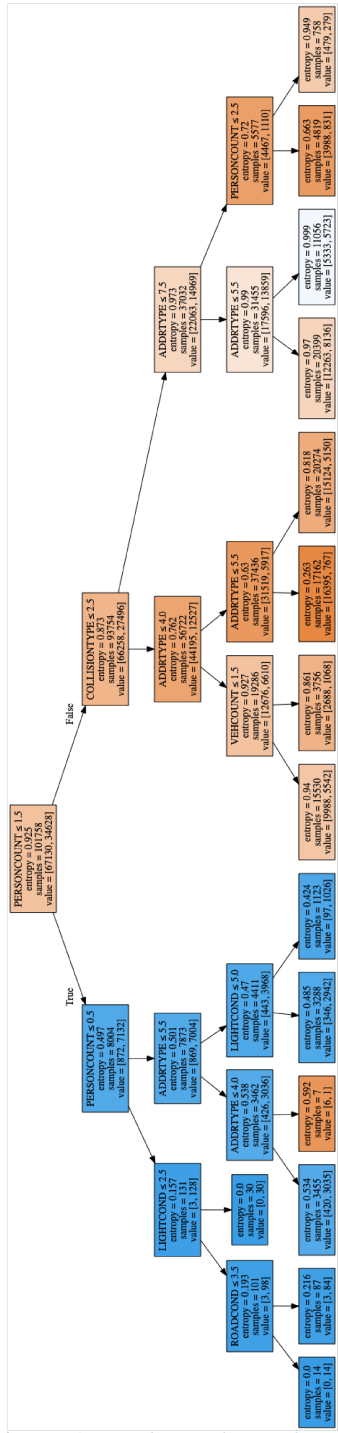

FIGURE 8. Decision tree model applied to collision dataset

Support Vector Machine

Another training method to build a severity prediction model is Support Vector Machine (SVM). Among various transformation scheme, Radial Basis Function (RBF) classifier is the most commonly used function, where support vectors automatically determines the centers and weights to minimize the test error (Scholkopf et al. 1997).

Therefore, using SVM with RBF, the dataset was trained and the confusion matrix was produced.

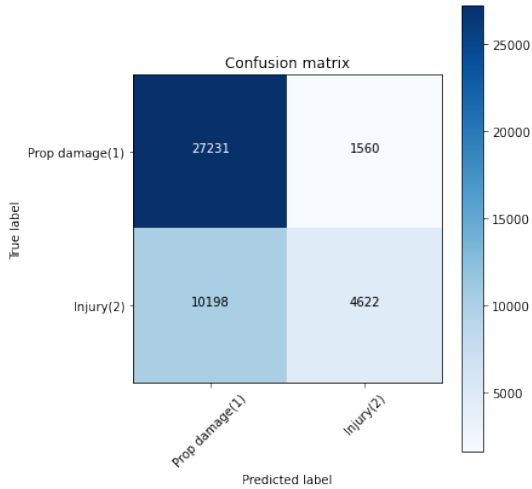

FIGURE 9. Confusion matrix of the training result conducted with SVM method

As shown in Fig. 9, for property damage labeled with 1, true positives are 27231 out of total 37429 cases predicted as 1, while 4622 out of total 6182 cases predicted as injury (2) were truly labeled.

Results and Discussion

In the previous section, the car collision dataset was trained to build a prediction model using three different methods: K-nearest neighbors, decision tree, and support vector machine. For each training method, the training set accuracy and test set accuracy can be evaluated. Additionally, the F_1 score which is a statistical measure of test set accuracy computed using precision and recall, defined as 1, was calculated.

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \tag{1}$$

where recall indicates the proportion of relevant items among the total selected group while precision refers to the proportion of selected group among the total relevant

items.

Table 2 and Fig. 10 show the training and test accuracy with F_1 score computed after training with each method.

Training method	Train accuracy	Test accuracy	F_1 score
K-nearest neighbors (K=6)	0.71	0.70	0.79
Decision tree	0.72	0.72	0.81
Support vector machine	0.72	0.72	0.82

TABLE 2. Training and test accuracy computed by three different training methods

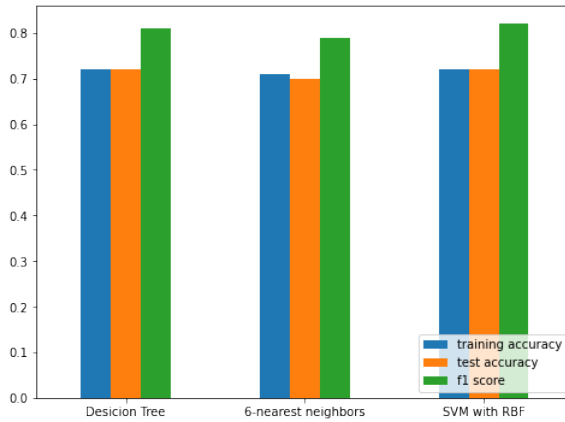

FIGURE 10. Training and test accuracy computed by three different training methods

In terms of calculated accuracy, support vector machine scheme turned out to be the best training method, for predicting the severity of the car collision, based on the dataset acquired from Seattle.

Conclusion

In this project, the severity of car accidents was predicted based on machine learning algorithm using three different training methods: K-nearest neighbors, decision tree, and support vector machine. The dataset was provided by Seattle car collision data, which are labeled by either property damage (1) or injury (2). After training, support vector machine with radial basis function produced the best F_1 score, 0.82, among the applied training methods. Therefore, it can be concluded that support vector machine is the proper method in severity prediction.

References

Lasley, Phil. 2019. 2019 URBAN MOBILITY REPORT.

Rifaat, Shakil Mohammad, and Hoong Chor Chin. 2007. Accident severity analysis using ordered probit model. *Journal of advanced transportation* 41 (1): 91–114.

Scholkopf, Bernhard, Kah-Kay Sung, Christopher JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio and Vladimir Vapnik. 1997. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing* 45 (11): 2758–2765.

TomTom.com. 2014. *TomTom Index Traffic Report.* Available at <https://www.tomtom.com/en_gb/traffic-index/>.

WSDOT. 2015. *Annual Collision Data Summary Reports.* Available at <https://www.wsdot.wa.gov/mapsdata/crash/pdf/2015_Annual_Collision_Summary.pdf>.