Predicting Car Accident Severity

IBM Capstone Project

Yae Won Kim

30.October.2020

Table of Contents

1. Introduction

1.1. Problem

The road accident not only interferes with traffic but also leads to personal injury. It is important to prevent road accidents because serious accidents as well as minor accidents can pose a great threat to human life. With the development of technology, it is possible to predict the fatality of an accident by considering various factors such as weather and road conditions using machine learning techniques, so that the advancement of technology can contribute to preventing/predicting severe accidents.

1.2. Interest

Not only the government but also the general public can make their lives more prosperous by being provided with information related to traffic safety. In bad circumstances, the government can control traffic or warn drivers. Drivers can reschedule their trip or be aware of poor traffic conditions and pay more attention to driving.

2. Dataset

2.1. Data sources

The dataset to be used for this project is open data released by the Seattle Government and contains all types of collisions data from 2004 to the present with 194,673 rows and 37 columns. All attributes are not considered, and only useful attributes will be used. This dataset contains data such as the severity of the collision based on the fatality and disabling injury counts, as well as how many people were injured in the accident. Besides, weather, road and light conditions are also recorded. There are different types of data such as numerical and categorical values here.

2.2. Data cleaning

It is checked whether there is any duplicate record, and it is confirmed that there is not a duplicate record. For improving accuracy and effective prediction, all attributes are not considered, and other columns are removed, leaving only 8 attributes that appear to be relevant. Selected attributes are

- ADDRTYPE - Collision address type (Alley, Block, Intersection)
- JUNCTIONTYPE - Category of junction at which collision took place
- WEATHER - A description of the weather conditions during the time of the collision.
- ROADCOND - The condition of the road during the collision.
- LIGHTCOND - The light conditions during the collision.
- SPEEDING - Whether or not speeding was a factor in the collision. (Y/N)

With those attributes, we will predict the severity of the collision.

- SEVERITYCODE - A code that corresponds to the severity of the collision

Severity code has four values:

- 0 means unknown
- 1 means damage
- 2 means injury
- 3 means fatality

After dropping irrelevant columns, it is found that all columns except SEVERITYCODE column have a lot of null data. The six columns are categorical value, so we will impute with the most frequent level(value) of that variable for each attribute(column).

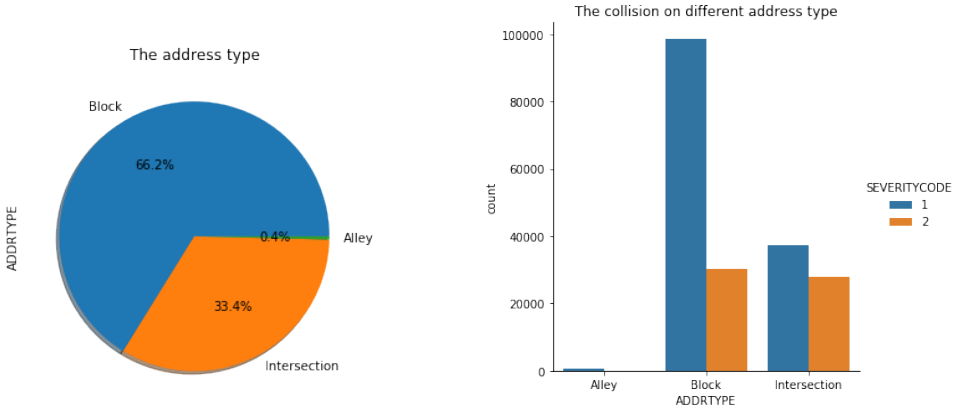
3. Methodology

3.1. Exploratory Data Analysis (EDA)

To analyse the dataset and find insights, EDA is conducted to see what the data can address with visualisations before the modelling.

3.1.1 ADDRTYPE

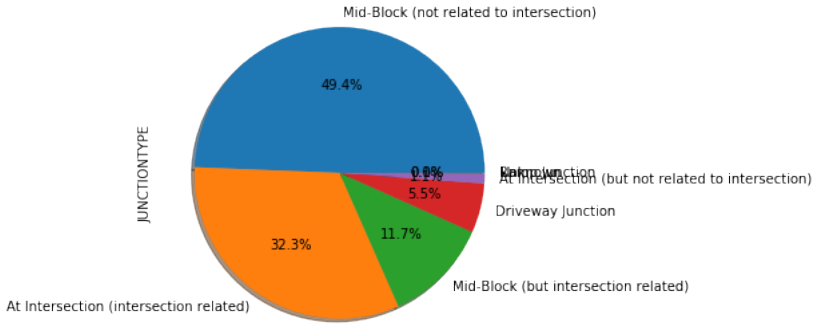
Firstly, for the address type (ADDRTYPE), as figure 1 shows, the collision occurred at block type of address the most and at the intersection the second-most. The collision rarely happened at the alley. For both severity codes, the collision occurred at the block more than at the intersection.



<Figure 1. The collision on different address type>

3.1.2. JUNCTIONTYPE

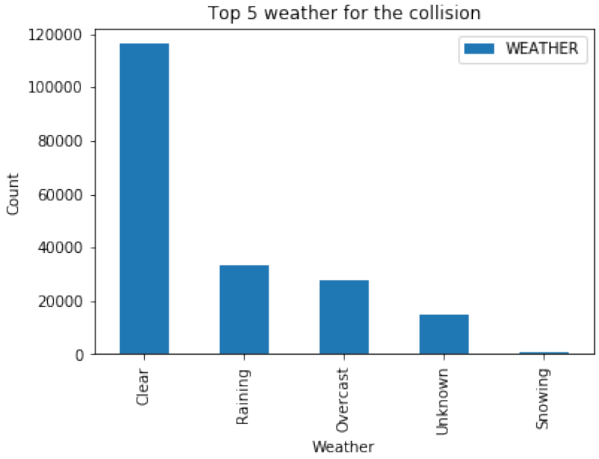
In terms of junction type, it is identified that the collisions occurred at mid-block (not related to intersection) the most frequently and at intersection (intersection related) the second-most frequently.



<Figure 2. The frequency of collision depending on junction type>

3.1.3. WEATHER

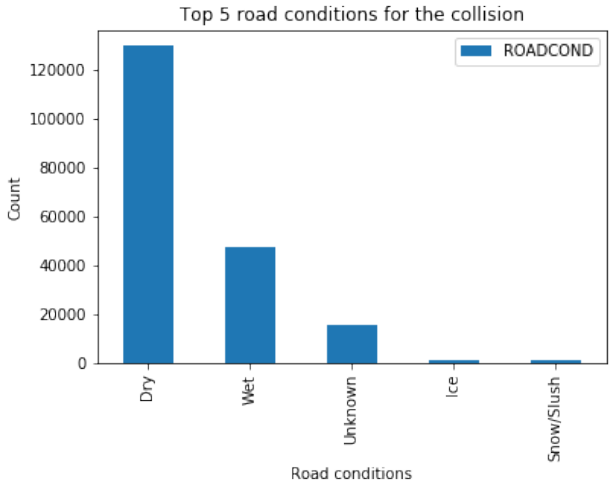
Interestingly, when collision occurred, most of the weather was clear, and unexpectedly, raining and overcast were overwhelmingly low.



<Figure 3. Top 5 weather for the collision>

3.1.4. ROADCOND

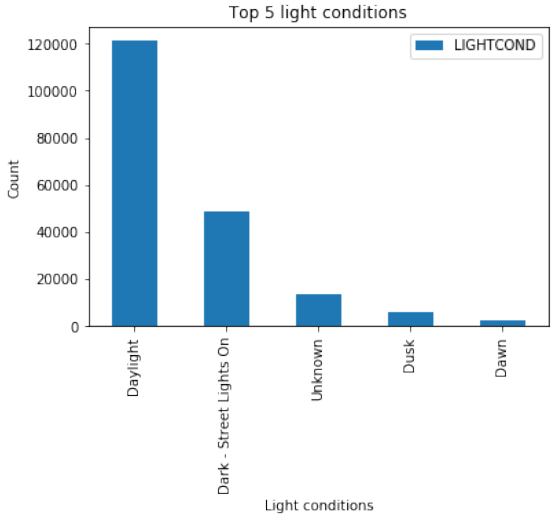
Top 5 road conditions were Dry, Wet, Unknown, Ice, Snow/Slush, and in most cases it can be seen that the road condition was dry.



<Figure 4. Top 5 road conditions for the collision>

3.1.5. LIGHTCOND

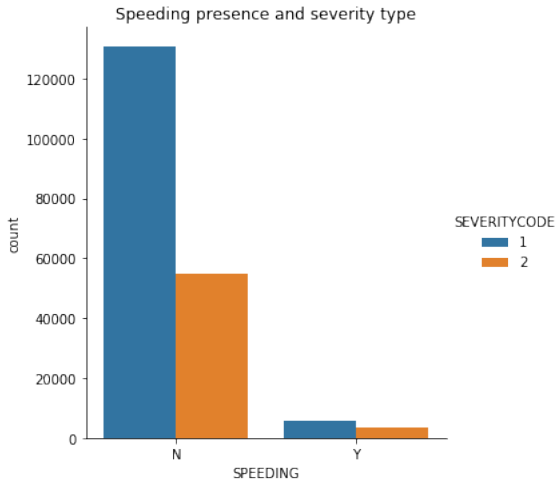
For light condition, dark and dawn showed a low number of cases, and daylight was the most common cause. It can be assumed that most collisions occurred during the day.



<Figure 5. Top 5 light conditions for the collision>

3.1.6. SPEEDING

When there was a collision, a large percentage of drivers did not speed, and even when considering the two different severity types, the proportion of drivers of each severity type not speeding was much higher.



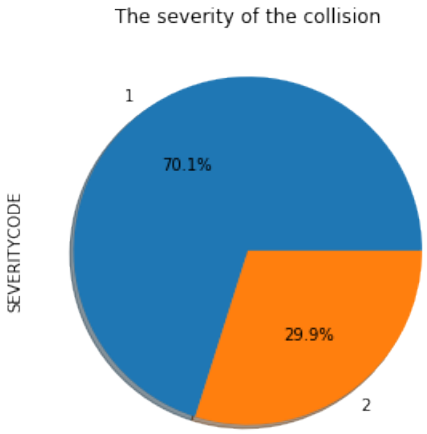
<Figure 6. Speeding presence and severity type>

3.1.7 SEVERITYCODE

The distribution of the severity of collision shows there are only two types of collision appeared in the dataset in terms of the severity. The severity of 1 (prop damage) with 70.1% occurred more than the severity of 2 (injury) with 29.9%. As the size of samples is not equal, this can cause a biased ML model, so the dataset should be balanced before processing.

	SEVERITYCODE
1	136485
2	58188

<Figure 7. The count of each severity code>



<Figure 8. The distribution of the severity of the collision>

To handle imbalanced data, it would be nice if it was possible to collect more dataset. However, if it cannot be collected, there are two ways such as undersampling and oversampling instead.

Oversampling may result in overfitting problem, so random undersampling method is used here for training data after splitting data into training and testing data in further process.

3.2. Balancing data in training data

Before starting machine learning modelling, separate the Y label which is SEVERITYCODE and we want to predict here from the dataset. Then, split the dataset into training and testing data at an 80:20 ratio and conduct undersampling on the training data set only to prevent creating a biased model. By setting the sampling strategy as 'majority', it will undersample the majority class which is the class with the largest number of examples. In this case, SEVERITYCODE 1 becomes the majority class having 135,485 examples (cases) in the total dataset and SEVERITYCODE 2 becomes the minority class. It will be undersampled so that both classes would have the same size of samples which is the sample size of the minority class. A balanced class distribution with training dataset will be created by selecting majority class instances at random to be removed.

3.3. Encoding categorical variables

As selected attributes are categorical variables, they cannot be used directly in predictive modelling. In this case, those categorical variables should be encoded to numbers before fitting and evaluating a model. For this task, an ordinal encoder is used to encode categorical variables to numbers.

3.4. Modelling

There are two types of models in predictive modelling which are classification and regression model. To build prediction models in this project which predict the severity of the collision, since all dependent variables are categorical variables, classification algorithms are considered for modelling. Three classification algorithms are selected such as logistic regression, decision tree and support vector machine and used to predict the severity of the collision.

4. Results & Discussion

To evaluate the performance of three models, accuracy, precision and recall are calculated to measure the performance of three models. Accuracy is calculated as the proportion of correctly predicted labels out of the total number of labels. It can be said that the higher the accuracy, the better. In terms of accuracy, all three models give the accuracy of over 60%. Among them, logical regression showed the highest performance. SVM also showed similar performance. However, the accuracy of 60% cannot be said to be very good accuracy.

	LR	DT	SVM
Accuracy	0.627584	0.603981	0.622627
Precision	0.406728	0.398441	0.403972
Recall	0.566154	0.665972	0.581444

<Figure 9. The performance of logistic regression, decision tree and SVM>

To improve accuracy, different resampling methods can be considered such as bootstrapping and cross-validation. As ordinal encoder is used for encoding categorical variables, a one-hot encoding can be used for categorical data if it is assumed that there is no relationship between categories. Furthermore, since this model got six attributes, ensemble learning methods which combine several learning algorithms could have been conducted to improve robustness over a single estimator and get better-predicted performance.

5. Conclusion

To sum up, modelling was performed with only 7 columns of the existing 37 attributes. For unbalanced datasets, undersampling was used to balance the dataset to create a biased model. Since categorical variables were used for prediction, classification algorithms were used such as logistic regression, decision tree and support vector machine (SVM). Based on the performance with accuracy for each model, it is concluded that logistic regression and support vector machine is the best model for predicting the severity of the collisions among the three models used. However, the accuracy of 62% is not high enough for predictive modelling. Thus, it is recommended that different resampling methods and can be used to improve the accuracy and robustness over a single estimator. Furthermore, this project simply focused on prediction, but it would have been more helpful to find insights into which variables influence collisions more than just prediction.