Predicting Traffic Incident Severity in Seattle, Washington — IBM Capstone Project

Kan Mo Bryan Lo

20 Oct 2020

1. Introduction/ Business Problem

Road traffic incidents causes fatalities every year. In 2019 alone, 36,120 lives were taken away due to all sorts of causes towards road incidents — driver misconduct, adverse weather, etc. Spendings on property recovery, medical costs, legal bills, and loss of earnings sums up towards billions of dollars on an annual basis. While it may seem that these incidents are beyond our expectations, identifying factors that contribute towards these incidents and their severity may be insightful for local authorities and drivers who wish to maintain a safe driving environment.

As open data are available for the public to examine details on previous incidents, specific causes can be found for incidents of large and small severity, where this project would attempt to model the road condition in order to predict whether it is safe and smooth for drivers to go onto a road trip in Seattle.

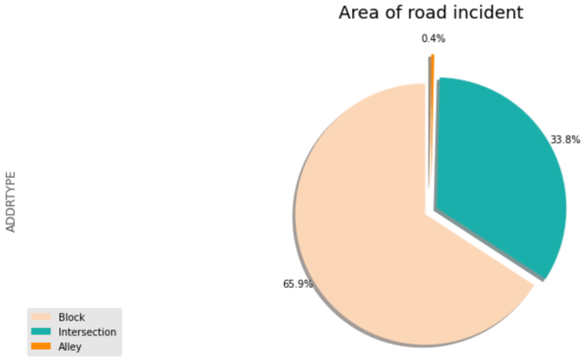
2. Data - describe the data that will be used to solve the problem and the source of the data

The dataset used for the project is extracted from official data from local authority that records information of every incident that happened in Seattle, Washington between year 2004 and 2020 (dataset available via here . Information regarding the location, place, time, and even conditions under which the incident occurred is presented. Given the project's objective being to predict the severity of incidents and their cause, the following features are chosen to be studied:

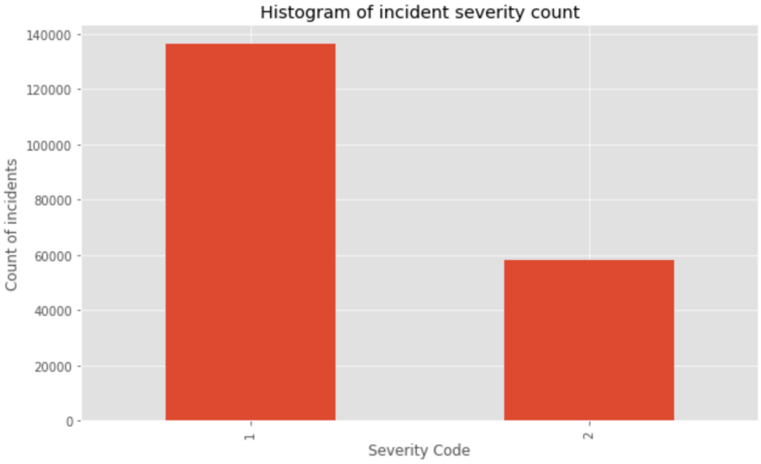
Dependent variable
SEVERITYCODE - A code that corresponds to the severity of the collision

Independent variables:
INATTENTIONIND - Whether or not the collision was due to inattention
UNDERINFL - Whether or not a driver involved was under the influence of drugs or alcohol.
SPEEDING - Whether or not speeding was a factor in the collision. (Y/N)
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.

With a total of 194,673 entries and 38 attributes, some 5,000 rows are found with null entries for independent variables such as ROADCOND and LIGHTCOND, where these rows are expected to be excluded for future modelling purposes.



Another discovery whilst looking into the day is that most (i.e. 66%) of the incidents occyr within blocks rather than intersections. However, we will not be looking into the matter as it is believed to be of minimal importance towards predicting the road condition.



With the severity code ranging from 0 to 3(i.e. 0, 1, 2, 2b, 3), incidents within the datasets only covers those with code 1 (property damage) and 2 (injury) only. Whilst the above bar chart depicts an unbalanced dataset, efforts will be made to balance out the count in order to facilitate a more accurate machine learning algorithm.The following algorithms will be used:

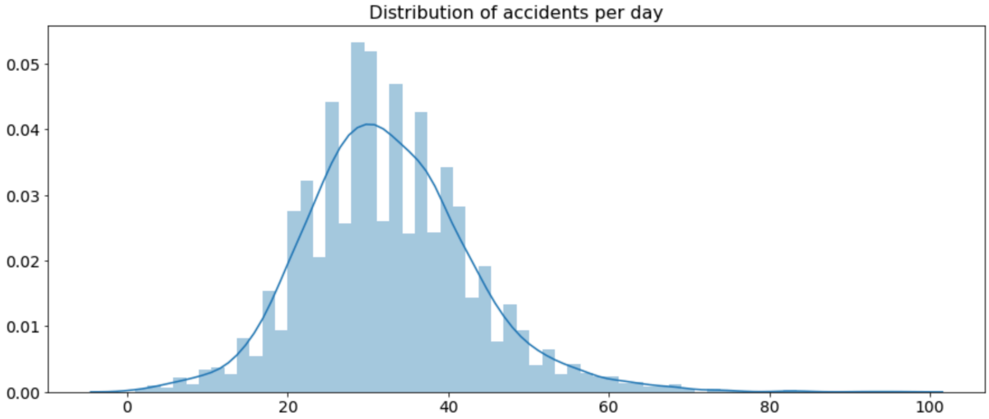
Logistic Regression - Using logistic functions to model binary output (dependent variable);
Decision Tree - Breaking down the prediction into smaller subsets and generating a tree-like logic flow to model the prediction;
k-Nearest Neighbour (kNN) - Grouping data points into categories/ groups based on similarity measures(or distance in between)

3. Methodology

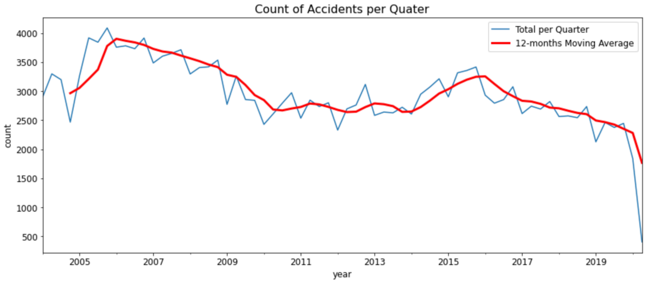
3.1. Exploratory Analysis

As a database, I used GitHub repository in this project. Looking into the frequency of incidents within timeframes, it is no surprise to see that the distribution of incidents count follows somewhat a normal distribution, which further supports the fact that the data itself fits real-life scenarios. And that the number of incidents each year is progressing towards a downslope trend.

This is further proved in the line plot above, where count of incidents per quarter decreases generally, with the 12-month moving average somewhat going up again between 2015 and 2016, and falling back down ever since. What is interesting here, is that the count of incidents actually follows a generic trend where the count first starts off not too high, and increases in Q2 and Q3 of the year, and falls to a yearly low in Q4.



Given the longer insolation duration during the summer and spring time, and lower in winter and autumn, it seems interesting that incidents tend to happen more often at brighter days of the year with clear skies and dry floor rather than darker ones with rough weather conditions and slippery roads.

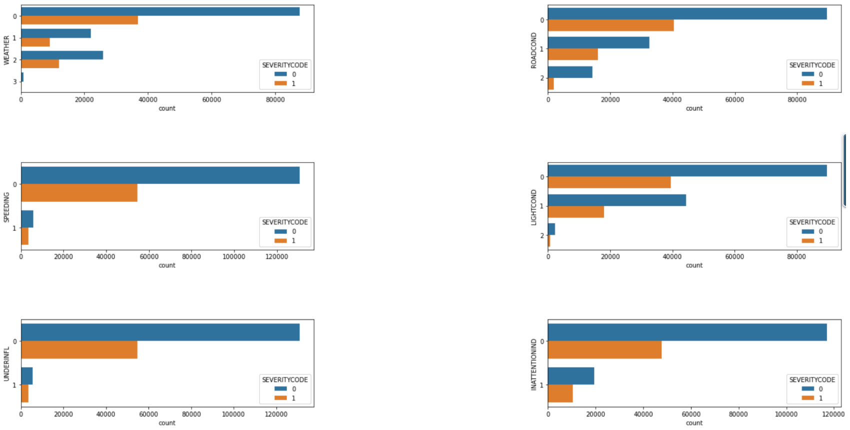


If we look into the geographical location of where these incidents occur, we can see the clusters of incidents (grouped by The State Collision Code Dictionary) in Seattle from the map below, where Downtown area as well as Northeastern area of Seattle are found with higher numbers of incidents throughout the 16-year time period. Local authorities may wish to look into enforcing straighter policies and upgrading infrastructures in order to reduce collisions around. For drivers, extra attention are needed when driving in respective neighbourhoods as well.

After checking that each incident is unique (by ensuring that each incident has its own unique ID "INCKEY"), a number of data cleaning gestures were done in order to facilitate a smooth modelling process:

- Replacing all attributes of the independent variables into integers (i.e. Y = 1; N = 0) such that these data points can be made better for model fitting
- Allocating the blank entries (i.e. ROADCOND) with respective score as per ratio of the filled cells (i.e. fill 0.4 of the blank cells with 0 where 0.4 equals the ratio of "0" in filled cells)
- Changing all other cells into int64 format (i.e. SPEEDING, INATTENTIONIND, UNDERINFL)

Counts of incidents according to severity and the independent variable can be visualised via the following plot:



Given that the count of cases with severity rating of 1 (property damage) is 2.35 times more than that of 2 (injury), SMOTE was deployed to balance out the dataset.

3.2. Model Testing and Results

With a sample test size of 0.3, and random state of 3, each of the aforementioned models were carried out, calculated the accuracy score, as well as have had a confusion matrix run such that a diagram can be plotted to display the performance of each model:

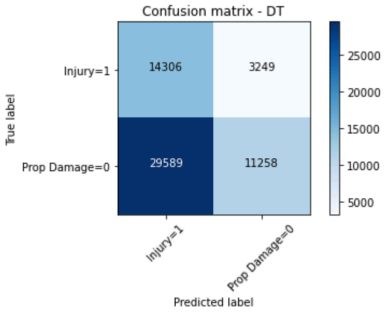
	Decision Tree	Logistic Regression	k-Nearest Neighbour
Accuracy Score	0.438	0.462	0.539

The Decision Tree Classifier was used to model the dataset, where the criterion of classification was ‘entropy’ and a maximum depth of 6 was chosen. A classification report and confusion matrix were generated alongside:

```
              precision    recall  f1-score   support

           0       0.78      0.28      0.41     40847
           1       0.33      0.81      0.47     17555

    accuracy                           0.44     58402
   macro avg       0.55      0.55      0.44     58402
weighted avg       0.64      0.44      0.42     58402
```

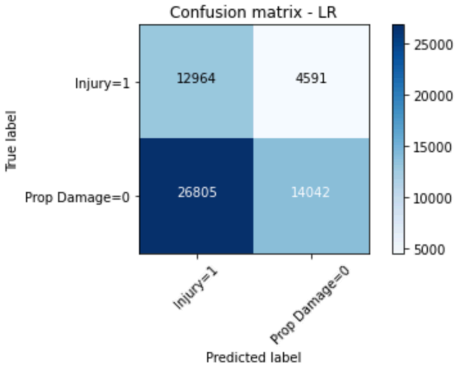


For Logistic Regression modelling, a C value of 0.01 and ‘liblinear’ solver are chosen to model the dataset. A slightly better performance is seen:

```
              precision    recall  f1-score   support

           0       0.75      0.34      0.47     40847
           1       0.33      0.74      0.45     17555

    accuracy                           0.46     58402
   macro avg       0.54      0.54      0.46     58402
weighted avg       0.63      0.46      0.47     58402
```

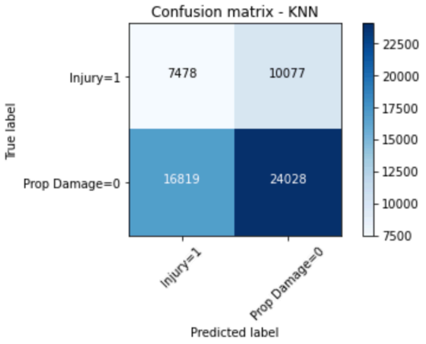


For k-Nearest Neighbour, a k value of 4 was chosen to begin the simulation, and then a *for* loop was deployed to model the prediction from *k = 9* all the way back to 1. It was found that the best performing value for *k* is 4:

```
              precision    recall  f1-score   support

           0       0.70      0.59      0.64     40847
           1       0.31      0.43      0.36     17555

    accuracy                           0.54     58402
   macro avg       0.51      0.51      0.50     58402
weighted avg       0.59      0.54      0.56     58402
```



4. Discussion - discuss observations, and recommendations based on the results

A summary of all the predicting performance of the 3 models are as follows:

model	f1 score	Accuracy Score
Decision Tree	0.42	0.438
Logistic Regression	0.47	0.462
k-Nearest Neighbour	0.56	0.539

Looking at the weighted average f1 score, which depicts the accuracy of the model, we can see that the KNN model performs the best (i.e. 0.56). Yet this metric is biased towards the precision and recall values as the higher *support* that the f1 score of either variable (0 or 1) gives, the more biased the weighted average would be bending towards. This result coheres with the comparison of the accuracy score where KNN performs the best as well, with a value of 0.539, representing around a probability of 50% of accurate prediction.

If we look closer at the precision and recall scores, which represents the proportion of relevant results and the accuracy of relevant results in successful prediction, we can see that Decision Tree performs best in predicting incidents with severity code 1 (property damage) (i.e. 0.78) , and both DT and LR perform best in incidents with code 2 incidents (injury) (i.e. 0.33). For recall, KNN is the best for code 1 incidents (i.e. 0.59) and DT is the best for code 2 incidents (i.e. 0.81).

Despite not high in evaluation metrics, but the confusion matrices provide insights towards the performance of the models for major stakeholders such as local drivers and authorities. Except for KNN, which predicts higher number of code 1 incidents when the true value is code 2, all other models predict either accurately, or predict a more severe incidents — from a safety standpoint, this is a good thing as it prevents drivers from suffering form congestions and potential incidents more effectively.

5. Conclusion

Much of the work here focuses on both internal and external stimuli that would potentially result in a traffic incident. From driver consciousness and influence, to weather and light conditions of the road — these attributes are perhaps some of the obvious ones that drivers would be able to relate with general road safety.

With the three algorithms generating mediocre results, better performances may be benefitted from the following:

- More balanced dataset
- Less blank cells where attributes such as SPPEDING and UNDERINFL are unknown
- Taking into account all influencing factors (i.e. date and time, location, SDOT code classification, etc.)

The same method of machine learning can be applied in other industries. For example, if machine learning could be applied to predicting underground incidents and general customer services, the ‘future’ may be more predictable as it may seem nowadays, and people will be better off preparing themselves for the many outcomes that the future may hold.

6. References

https://www.statista.com/topics/3708/road-accidents-in-the-us/