IBM Capstone Project – Modelling Accident Severity Level from a Car Collision

Business Understanding

The impact of road traffic accidents goes beyond property damage. In most cases, human lives are at stake. In more serious cases, fatalities can occur. While road accidents can be reduced through better urban designs and enforcement of the law, more can be done to understand how various factors can influence the severity of the accidents.

Besides the common factors such as weather conditions, road conditions and visibility, other factors such as the type of intersection, the location, the time of day, etc. could also have an impact on both the probability and severity of a car collision if it does happen.

With the increase usage in embedded technology in modern cars, it is possible for cars to have greater awareness of the environment such as current location, weather and visibility condition, route to the destination, etc.

Using these environment and location inputs, it would be beneficial to develop a model to predict the accident severity given the current situation. This would be very helpful to drivers as such a model can give the drivers warnings and encourage them to drive safety in such situations. Hence reducing the chance of an accident.

Data Understanding

The data provides a list of accident traffic records from the Seattle Police Department. Based on the period from 2004 to 2020, 194,673 accidents have been recorded.

Each accident is recorded as a row with 37 independent attributes that provide various details on the accident. The dependent variable SeverityCode is used to indicate the different levels of severity caused by the accident.

There are generally 4 levels of severity based on the meta description of the accidents:

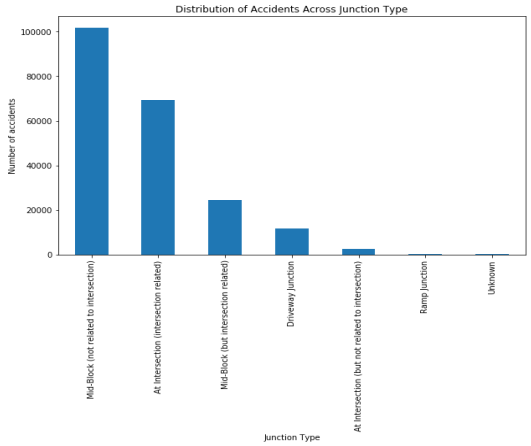
- 3 – Fatality
- 2b – Serious injury
- 2—injury
- 1—property damage

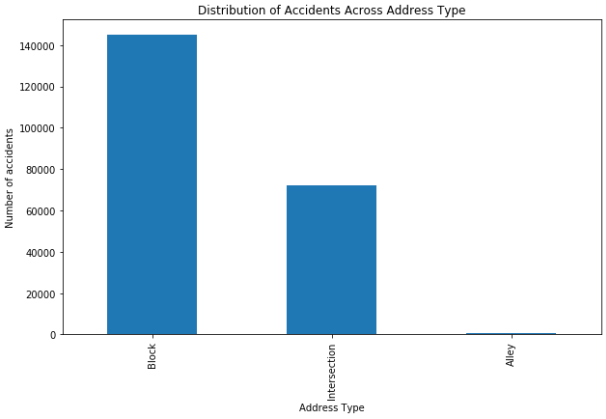
The attributes provide more insights into the details of the accidents and can be broadly categorized into the following area:

1. Location data: The location where the accident took place and characteristic of location such as the junction type.
2. External environment data: Data such as the weather condition, road condition and visibility are available.
3. Impact data: The number of injured persons and other vehicles are also recorded.
4. Driver condition data: In some cases where the accident is caused by driver inattention or influence under alcohol is also recorded.

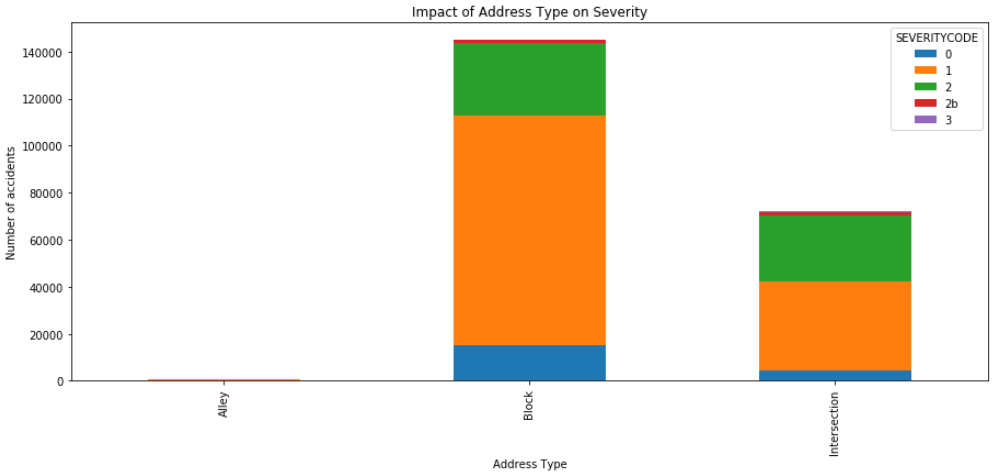
Based on the data category, we can visualize how it could impact the level of severity.

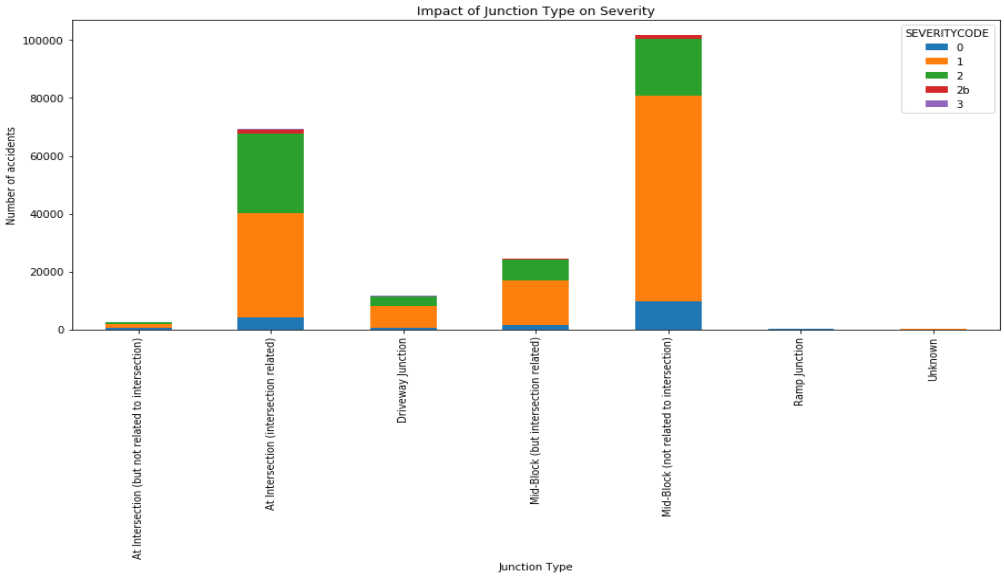
Analysis on Address and Junction Type





We can see that certain address and junction types do have more occurrence of traffic accidents. This could provide insight that road design plays a part in reducing accidents.

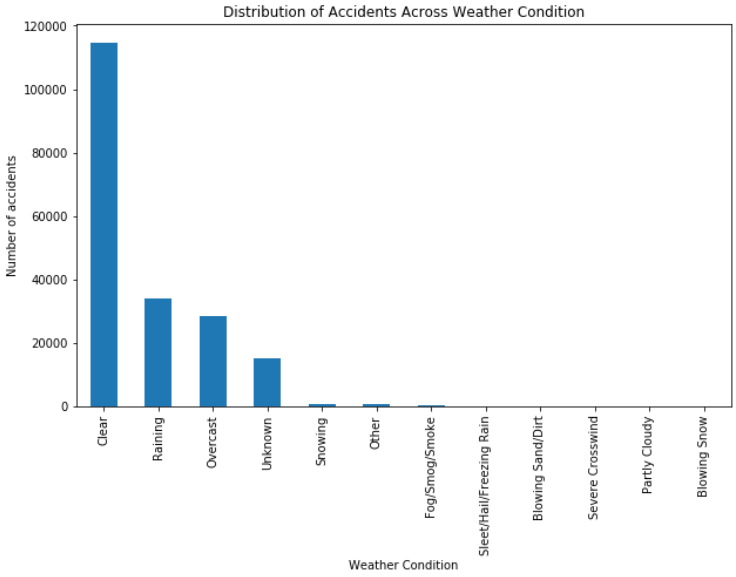




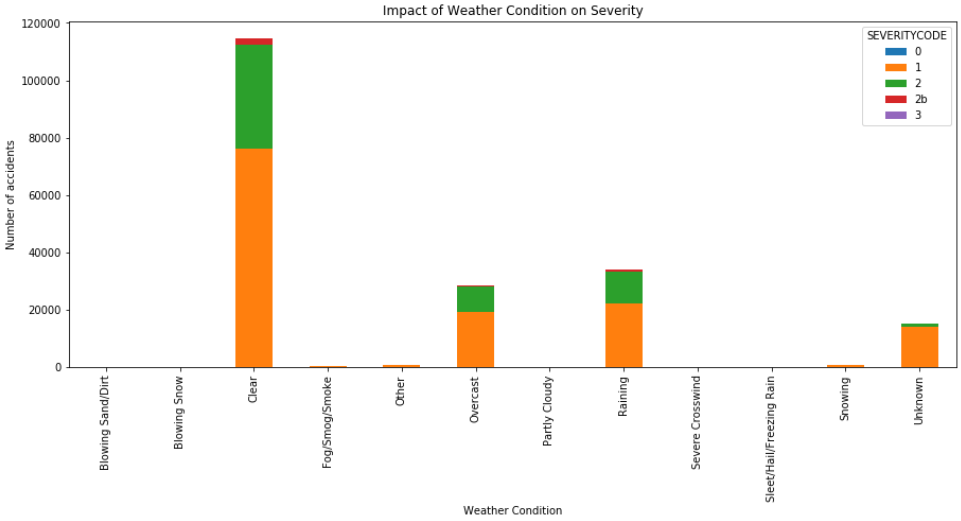
By grouping them into severity level, we can also observe that intersections have a greater proportion of accidents with higher severity level.

Analysis on Environmental Impact

Weather Condition

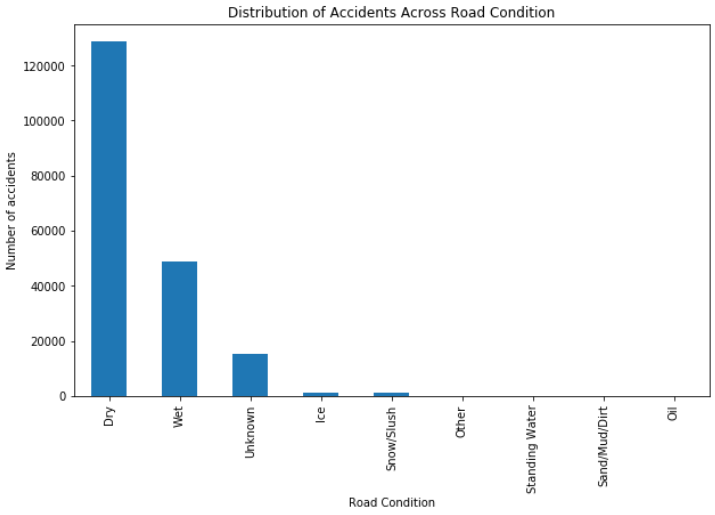


Looking at the distribution of the accidents by weather type, a majority of the accidents happens on clear day. This is might seem to be counterintuitive. This could also be due to the location where majority of the time the weather is clear, resulting in a heavier representation of the accidents.

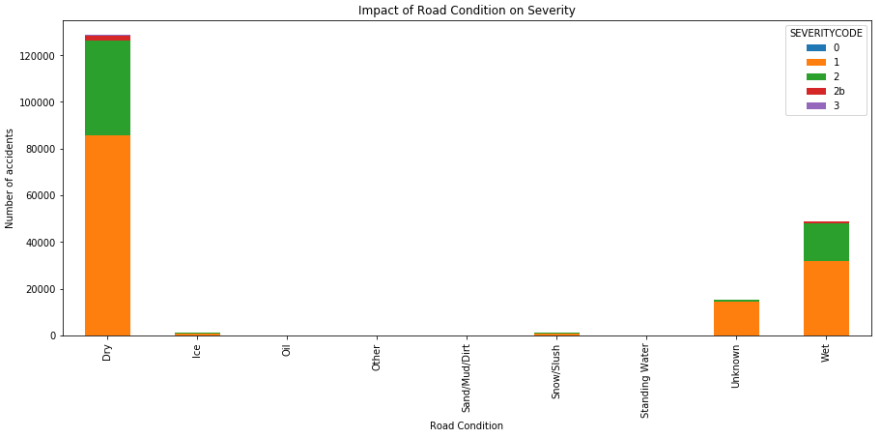


On closer inspection of the data, we also note that clear weather has a higher proportion of severity level 2 accidents compared to rainy days. This can be explained that on a rainy days, drivers would have exhibited more care, resulting in lesser accidents.

Road Condition

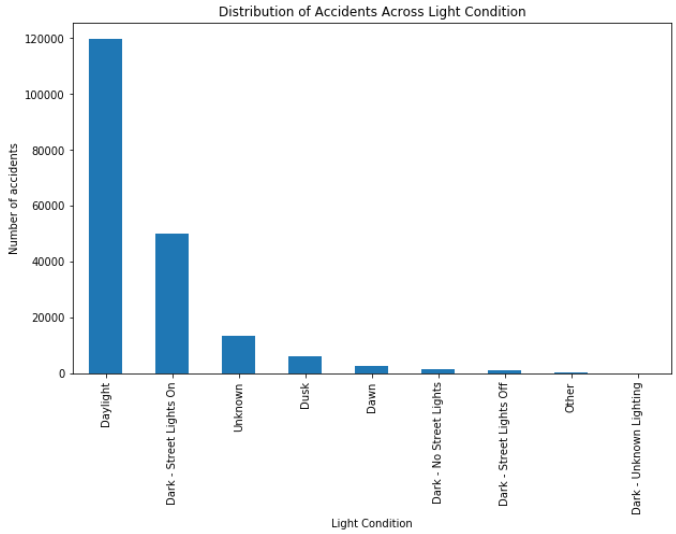


We can also observe a distribution of the number of traffic accidents across road conditions.

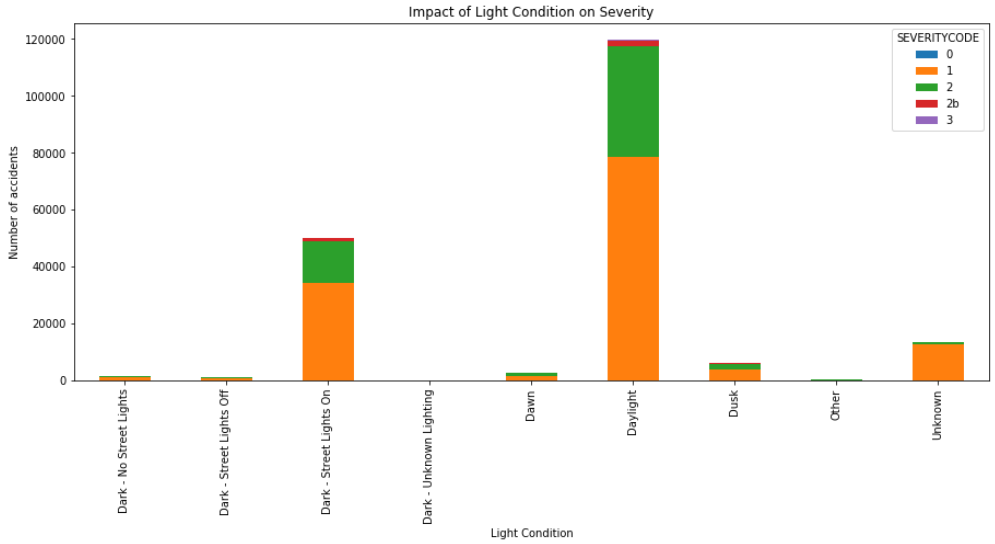


Looking at the proportion of accidents by severity level, road conditions might not play an important factor in terms of contribution to the severity level of the accident.

Light Condition



Similarly, we observe that majority of the accidents happened in daylight. This could again be due to biasness in the dataset resulting from human behaviour such as more traffic in the day due to working hours.

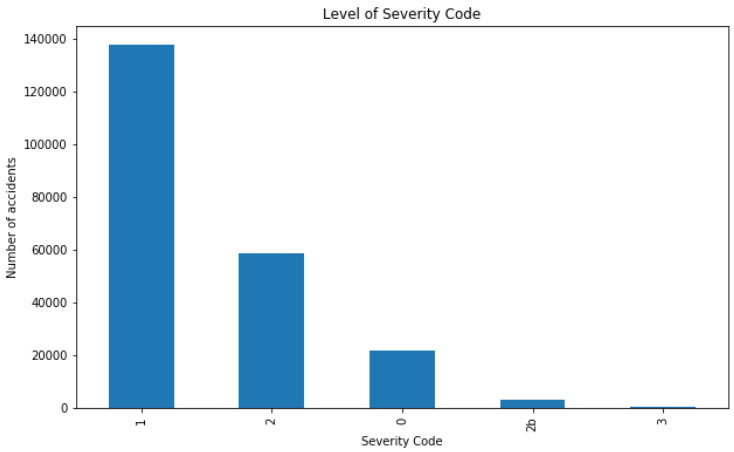


A breakdown of light condition on severity level shows that poor lighting may not necessarily result in more severe accident. This could indicate that drivers generally let their guard down in good driving conditions and exercise cautious when the driving conditions are not ideal.

Analysis on Target Variable

We also note in the dataset that not all information is always available. In some cases, certain key feature values are blank and not available. Further data processing is required to clean up the data before further analysis.

In addition, the dataset is not balance with most of the data (136,485) representing accident of severity 1 and only a small set of data (58,188) representing severity 2. No other severity level is found in the dataset. This could result in a model that is bias due to the lack of proper balance of the various scenarios in the dataset. Sampling of the data might be required to transform the data into a more balanced dataset.



Data Preparation

Feature Set Selection

The dataset in its original form has many different attributes. However not all attributes are of relevant to the proposed use case that we are looking to solve. In the context of the use case, we are looking at a model to help warn the driver of potential accident. In the earlier section, we have categorized the data into 4 categories and we will look at extracting the feature set from these 4 categories:

1. Location data: The location data is focused mainly in the Seattle area. As I would like to keep the model more generic, I have decided not to use the exact location as an input. However, feature like address type and junction type might give good information on how road design has an impact on the accident severity.
2. External environment data: The external environment features are used as part of the feature list as well as they played a part in overall environment constituting to the collision.
3. Impact data: As the impact data is a result of the collision, I would not use this as part of the dataset.
4. Driver condition data: While the driver condition data has interesting information on the state of the driver, it would not be feasible to get these data in a real-life scenario (e.g. if the driver is paying attention). However, it is possible to check if the car is speeding and hence this in one of the data used in the feature.

In summary, the following data points are used as the feature set:

- WEATHER
- ROADCOND
- LIGHTCOND
- SPEEDING

- ADDRTYPE
- JUNCTIONTYPE

Target Variable

The target variable used in this case is the Severity Code which indicates the severity of the collision. In the context of the use case, we generally want to split the category into 2 mainly groups:

- Collisions that result in no injury (i.e. Severity Code 1)
- Collisions that result in personal injury (i.e. Severity Code 2, 2b, 3)

Data Cleaning

Given the selected feature and target variable, we observed that the data is not clean and suitable for modelling. Hence the following steps have been performed to process the data.

1. Removing of empty rows
   There are various rows where the feature and target data are not available. Hence these rows are removed to avoid poor modelling due to missing inputs.

2. Removing Others values
   For the values in the feature datasets, there are cases where the data is not empty but the value is indicated as unknown or others. As it is not clear what it is meant as others, these rows are also removed to keep the data clean.

3. Categorizing the Severity Code
   To simplify the severity code, we want to group them into only 2 main categories. For severity code 1, it would indicate that the collision did not result in injury. For severity 2 and 2a, it would indicate that injuries have occurred. We have removed rows with severity 3 as the number of data available is very low which will result in wrong modelling.

4. Down sampling of data
   We also observe that majority of the dataset has an outcome of Severity Code level 1 which resulted in an unbalance dataset. To resolve this issue, a down sampling of the majority dataset is performed to create a balance dataset.

5. One-Hot encoding
   As majority of the feature data is categorical data, a one-hot encoding is applied where a new binary variable is added for each of the categorical values. This will prevent the model from assuming a natural ordering between the categorical values which can result in poor performance.

Methodology

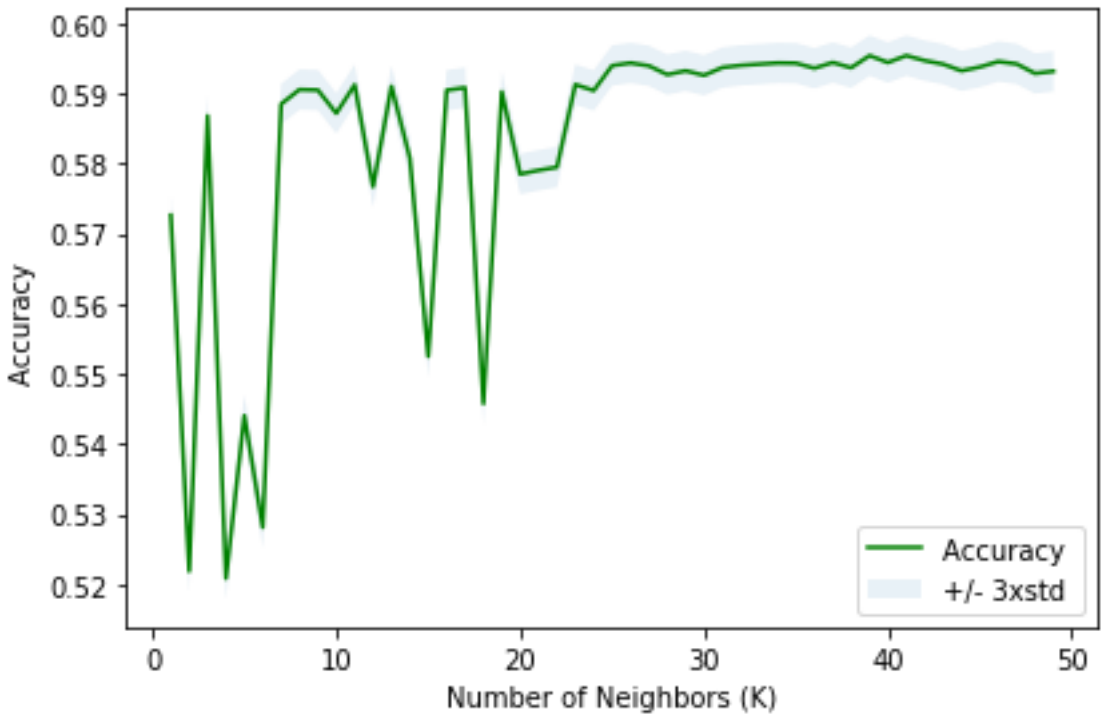
After processing the dataset, we are now ready to build and train the various models. For this use case, we want to have a model that can gives a clear category of the Severity Code and at the same time it would also be useful to have a different model which can give the Severity Code and the probability of that Severity Code. Hence, we can consider employing the following 3 models:

1. K-Nearest Neighbours (KNN)
2. Decision Tree
3. Logistic Regression

The start the training, we have also split the dataset into training and testing data set with a proportion of 75% of the data is used for training.

K-Nearest Neighbours (KNN)

Using the K-Nearest Neighbours model, we can see that the accuracy hovers around slightly above 59% as K increases. Based on the results, K = 26 was used which resulted in an accuracy of 0.59437.



Decision Tree

We also apply the training and testing dataset into a decision tree model. By varying the depth of the tree, we can observe at a depth of 6 gives the best accuracy of 0.59731. However, it is also noted that the tree depth did not significantly change the level of accuracy.

```
DecisionTrees's Accuracy:  1 0.5928121400989093
DecisionTrees's Accuracy:  2 0.5928121400989093
DecisionTrees's Accuracy:  3 0.5958945870875957
DecisionTrees's Accuracy:  4 0.5958945870875957
DecisionTrees's Accuracy:  5 0.5968769053587155
DecisionTrees's Accuracy:  6 0.5973172549285278
DecisionTrees's Accuracy:  7 0.5968430323148838
DecisionTrees's Accuracy:  8 0.5967752862272204
DecisionTrees's Accuracy:  9 0.5961994444820812
```

Logistic Regression

In the case of KNN and Decision Tree, we are only able to get a categorical result on the severity code. With logistic regression, we will also be able to get the probability of the severity code.

Using the same training and testing dataset and applying logistic regression, our model achieved an accuracy of 0.59498.

Results & Evaluation

Below is the result of the various models developed based on the dataset.

Algorithm	Jaccard	F1-score	Log Loss
KNN (K=26)	0.594370	0.593672	NA
Decision Tree (Depth 6)	0.597114	0.596785	NA
Log Regression	0.594980	0.592254	0.670874

Based on the results of the various model, we are generally able to achieve an accuracy ~0.59 for the various models by exploring the various parameter used (e.g. K in KNN and the tree depth). This shows that the features chosen in the dataset does have an impact to the collision severity. However, it should also note that the model accuracy is not high and might accurately predict the outcome correctly. Given that Log Regression also provide the probability of the predicted value and its accuracy is similar to the rest of the model, it is a good candidate for our use case.

Conclusion

With the advancement in self-driving cars and increase use of technology in cars, there is a potential use of incorporating machine learning and new data inputs to assess the environment risk which could result in a car collision. Using data like location, weather condition, lighting condition, etc. could help to predict these risks and warn the driver accordingly.

From the development of the various model, the accuracy achieved from the dataset is not as high as expected. This could also due to possible biasness in the data. For example, if most of the days are clear in the year, it is likely that most of the collision would happen in those clear days' vs the non-clear days which are less frequent. Similarly, certain road conditions (e.g. sand and mud) might not be common. Such data biasness has not been addressed with the current use of the dataset.

However, through the development of the various models, we can see that machine learning techniques can be used to explore historical data and get better insights to predict traffic accidents. The model explored can be extended to incorporate more features so that we can improve the accuracy. By doing so, there is potential to extend the use in other scenarios such as drive route planning, improving road design, etc. This will enable better design and help to reduce the rate of traffic accidents and their severity.