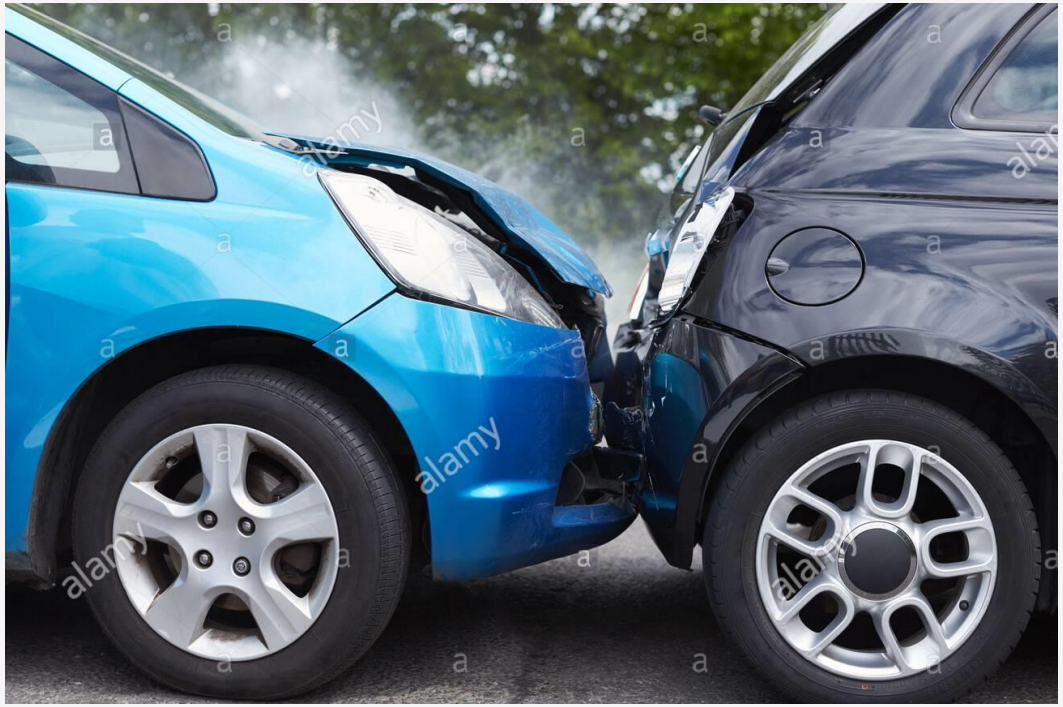
IBM Data Science Capstone Project

Predicting traffic accident severity based on Seattle historical records

Background

- 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.

By using historical traffic accident data, can we build a model to predict the severity of a traffic accident?



Motivation

- By building a model that can be used to predict severity of traffic accidents, we can use it for various scenarios:

1. Better road design and city planning
2. Increased driver level of caution in situations where there is high risk of an accident
3. Improved road routing by recommending routes that are of lower risk

Data Exploration

- We will develop a model based on the data was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present.
- Based on the data, we can generally categorized the data into the following categories

Location data

- The location where the accident took place and characteristic of location such as the junction type.

External environment data

- Data such as the weather condition, road condition and visibility are available.

Impact data

- : The number of injured persons and other vehicles are also recorded.

Driver condition data

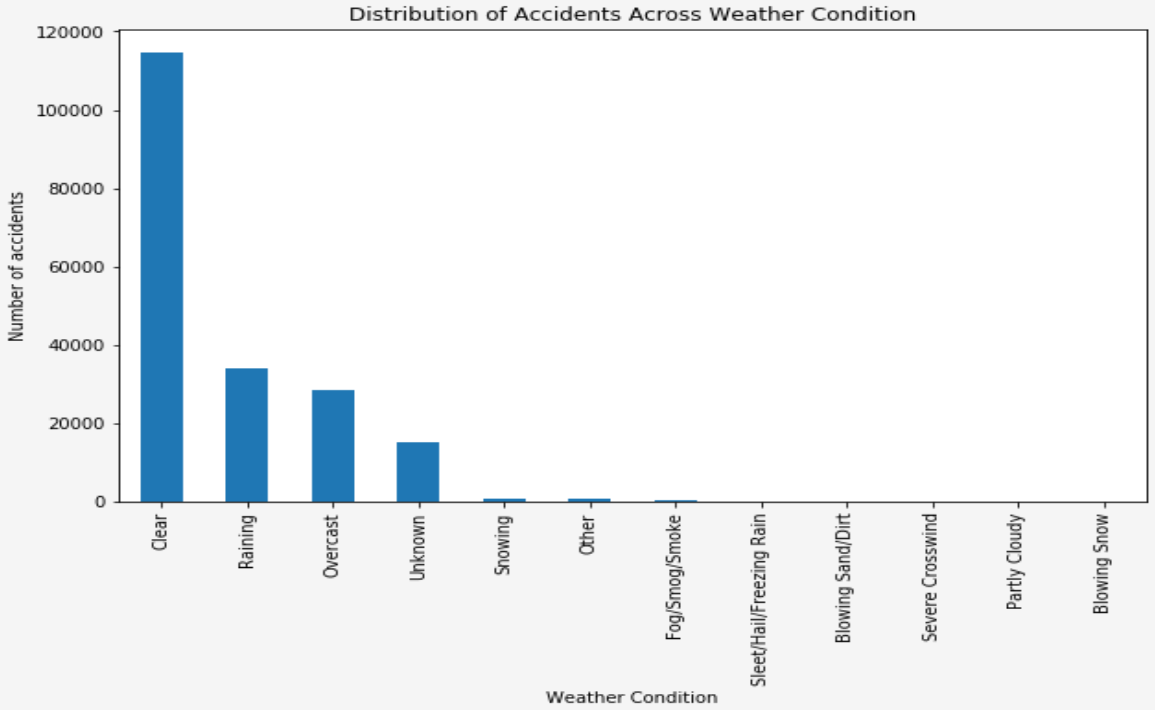
- In some cases where the accident is caused by driver inattention or influence under alcohol is also recorded.

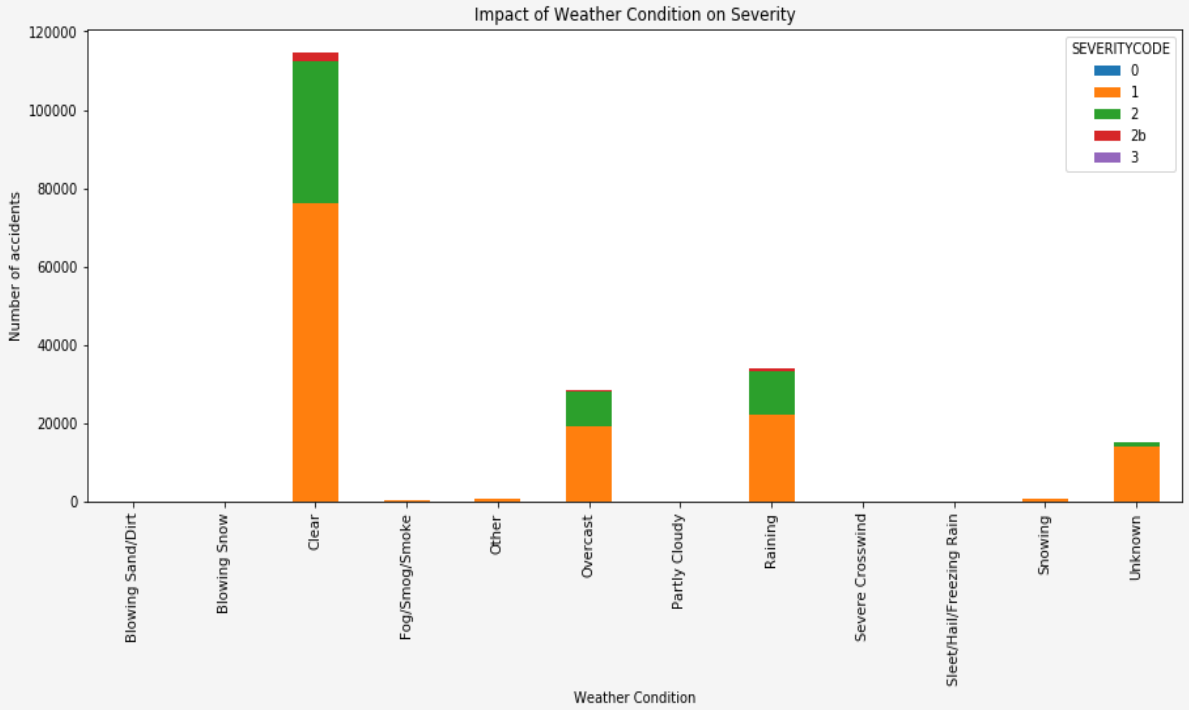
Severity Code

1: Property damage
2: Injury
2b: Serious injury
3: Fatality

Data Visualization – Weather Condition

- By visualizing the data, we can observe how the distribution of the number of traffic accidents across various factors and its impact on severity level.



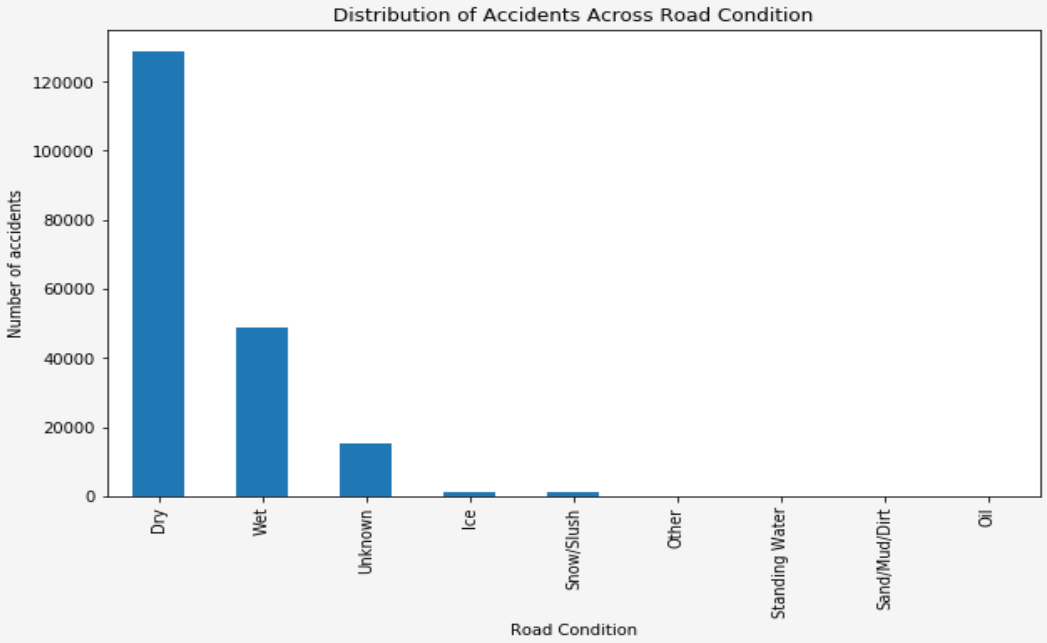


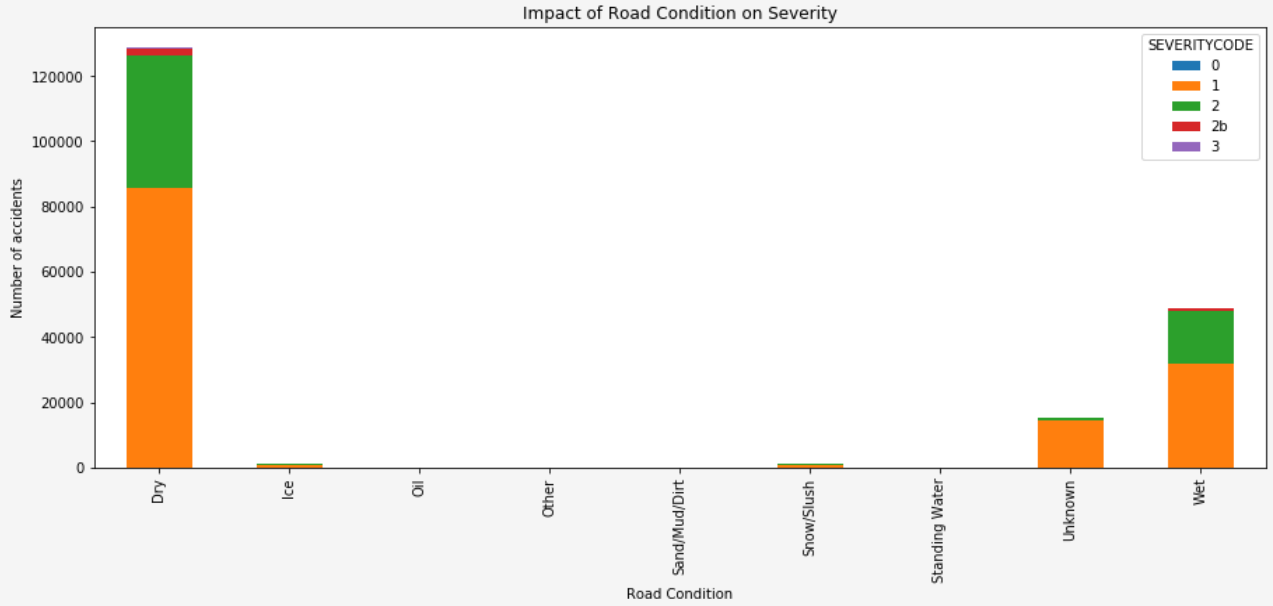
Insights

We would have expected on rainy days, there would be higher proportion of accidents with higher severity. However the data shows that this is not the case. This could due to greater caution exercised by the drivers in such situations.

Data Visualization – Road Condition

- By visualizing the data, we can observe how the distribution of the number of traffic accidents across various factors and its impact on severity level.



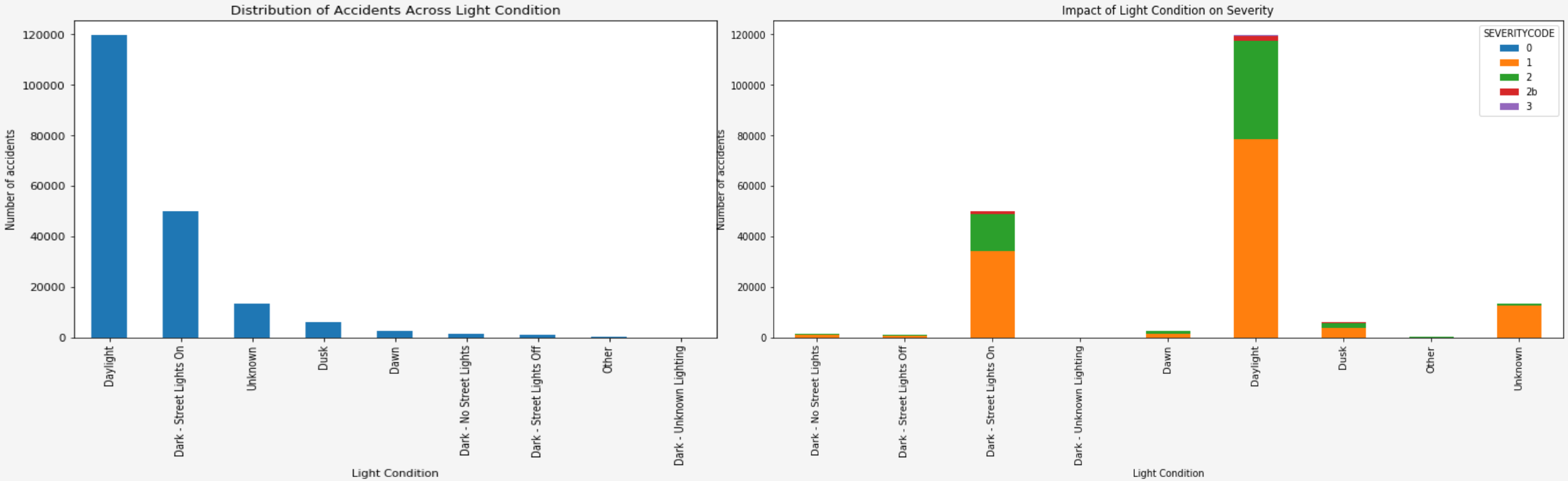


Insights

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.

Data Visualization – Light Condition

- By visualizing the data, we can observe how the distribution of the number of traffic accidents across various factors and its impact on severity level.

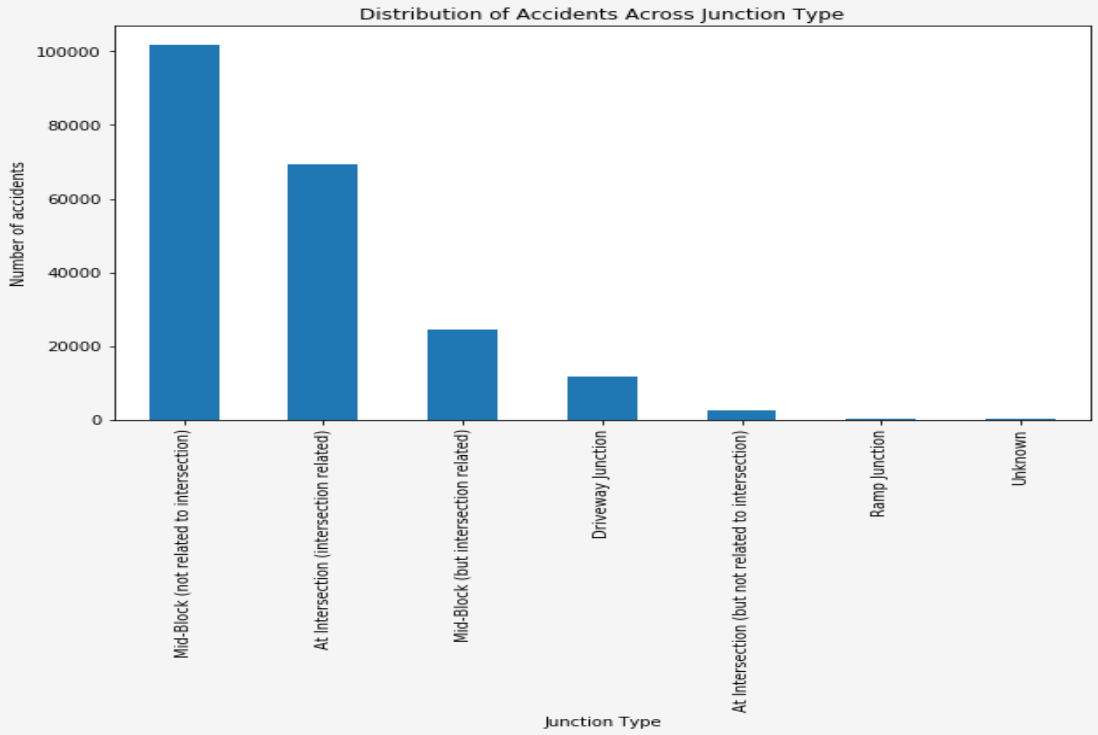


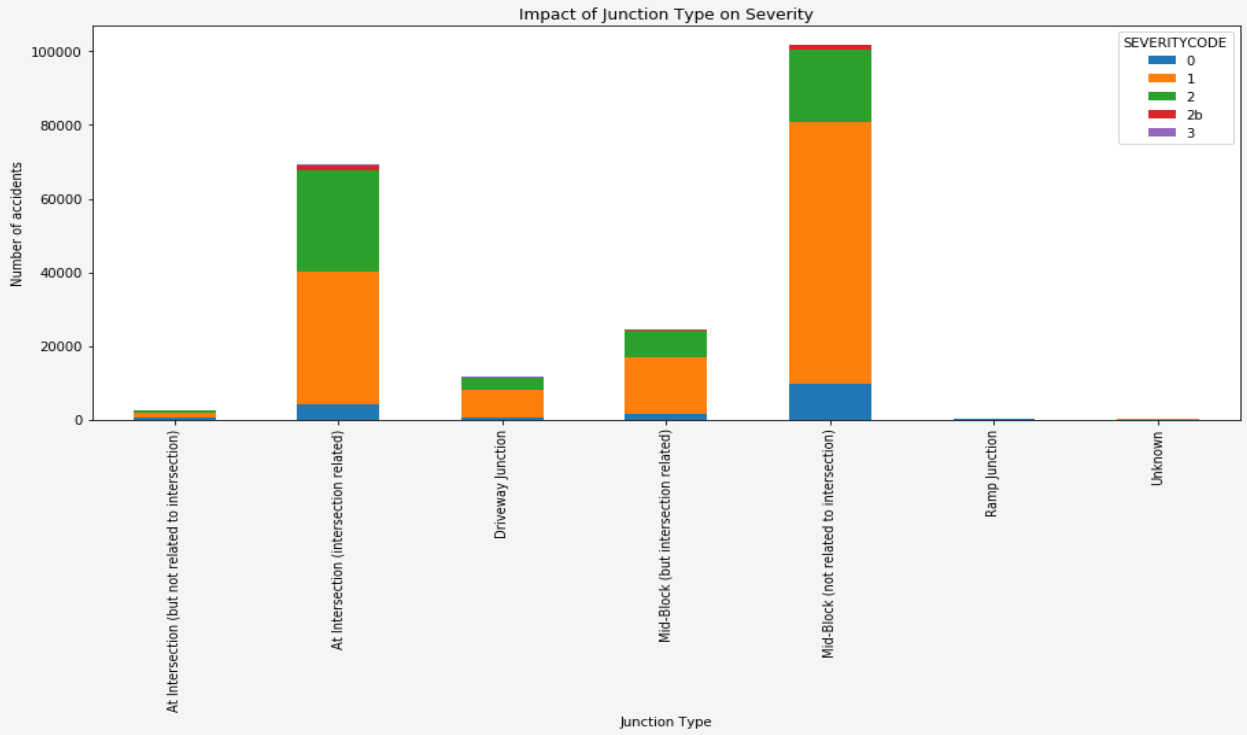
Insights

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.

Data Visualization – Junction Type

- By visualizing the data, we can observe how the distribution of the number of traffic accidents across various factors and its impact on severity level.



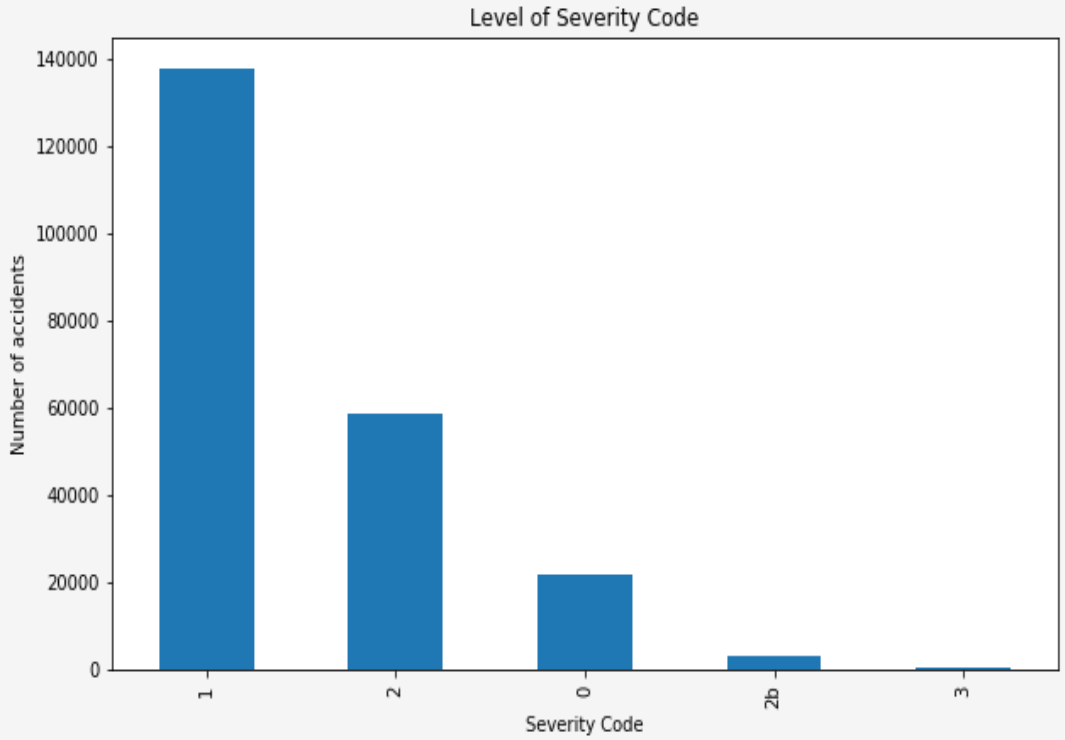


Insights

We can also observe that intersections have a greater proportion of accidents with higher severity level.

Data Preparation – Feature Selection

- Based on the dataset, the following datapoints are used as the feature set to predict the severity level.
    - WEATHER
    - ROADCOND
    - LIGHTCOND
    - SPEEDING
    - ADDRTYPE
    - JUNCTIONTYPE



Data Preparation - Preprocessing

- Before applying the data to the various machine learning models, the data has to be pre-processed to ensure that the data is clean and suitable for training.

Removing of empty rows

- These rows are removed to avoid poor modelling due to missing inputs.

Removing Others values

- For the values in the feature datasets, there are values indicated as unknown or others.
- These rows are also removed to keep the data clean.

Categorizing the Severity Code

- To simplify the severity code, we want to group them into only 2 main categories.
- 0 for accidents that result only in property damage.
- 1 for accidents that resulted in injury.

Down sampling of data

- Majority of the dataset has an outcome of Severity Code level 1 which resulted in an unbalance dataset.
- Down sampling of the majority dataset is performed to create a balance dataset.

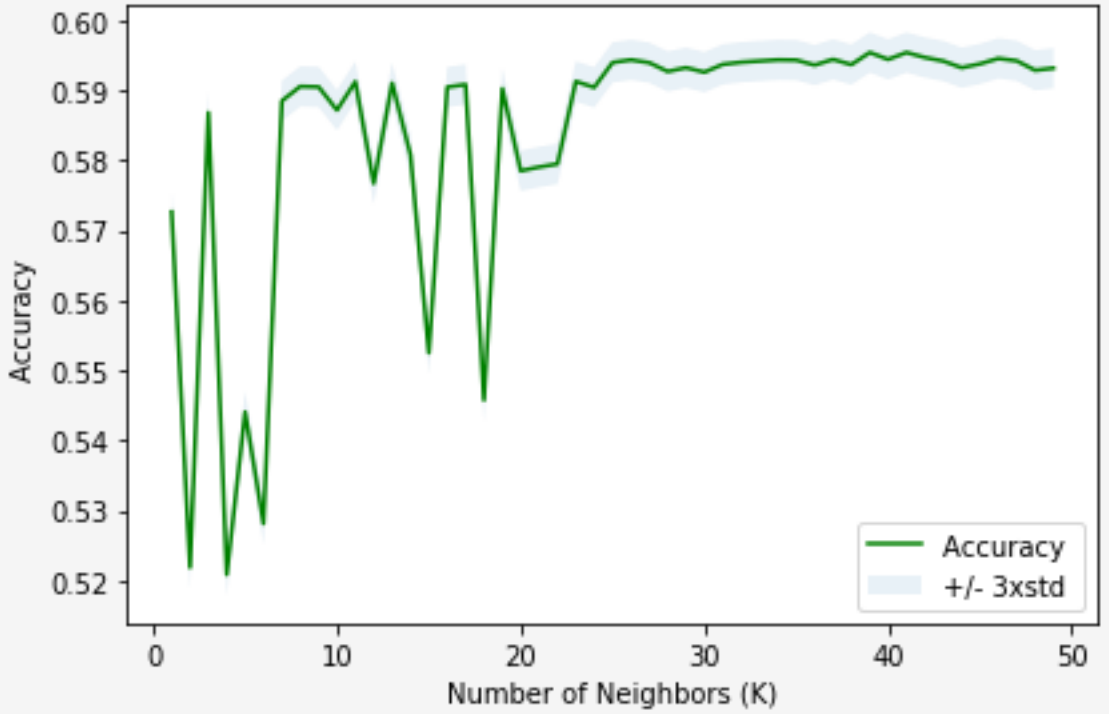
One-Hot encoding

- Most of the feature data is categorical data.
- One-hot encoding is applied where a new binary variable is added for each of the categorical values.

Methodology

- The dataset is split into a training dataset (75%) and testing dataset (25%).
- The following models are developed:
    - K Nearest Neighbours (KNN)
    - Decision Tree
    - Logistic Regression

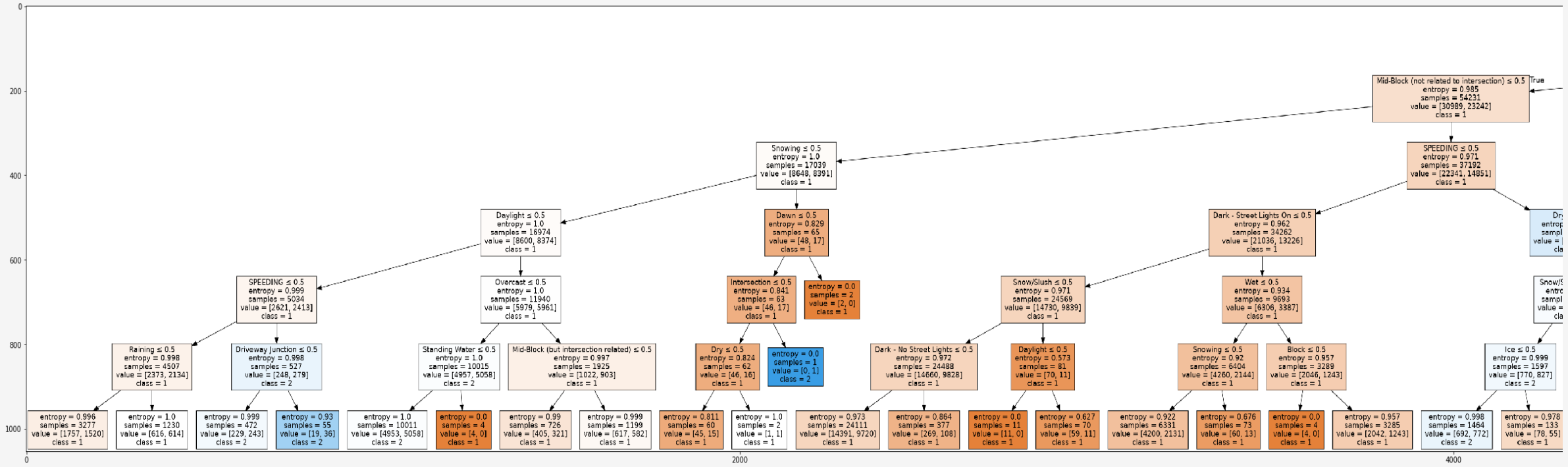
Methodology – K Nearest Neighbors



Model Evaluation

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.

Methodology – Decision Tree



*Partial decision tree for illustration purpose only.

Model Evaluation

By varying the depth of the tree, we can observe at a depth of 6 gives the best accuracy of 0.59731.

Methodology – Logistic Regression

```
: from sklearn.linear_model import LogisticRegression

LR = LogisticRegression(C=0.01, solver='liblinear').fit(X_train,y_train)
yhat_log = LR.predict(X_test)
yhat_log_prob = LR.predict_proba(X_test)

print("Log Regression Accuracy: ", metrics.accuracy_score(y_test, yhat_log))
print("Log Regression Log Loss: ", metrics.log_loss(y_test, yhat_log_prob))
```

```
Log Regression Accuracy:  0.5949800149041393
Log Regression Log Loss:  0.6708742641502969
```

Model Evaluation

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 a summary of how the various model performed 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

Summary

- 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.
- 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.