Predicting Car Accident Severity

Martin Ng

02 October 2020

Introduction

Car accidents, or in general traffic accidents are a serious problem of the modern society. The World Health Organisation estimates that very year, road accidents result in more than 1.3 million deaths, 20 to 50 million of non-fatal injuries and costs economies 3% of their annual gross domestic product through lost resources, productivity and collateral damage. It is thus important to determine the factors leading to accidents, in order to develop strategies to eliminate or mitigate them to reduce the occurrences of traffic accidents.

Traffic accidents lead to a variety of consequences, ranging from altercations, minor property damages to the more severe loss of human lives. Having studied the factors causing traffic accidents, a subsequent, important step is to then determine what affects the level of severity of accidents, and if we can effectively predict the severity of the accidents.

Problem

Aside from understanding the factors causing accidents, it is also imperative to understand what causes severe accidents so that we can predict them and developed targeted, prioritised strategies to reduce high severity accident occurrences first, as an efficient use of limited resources.

Interest

With such insights, country agencies can efficiently allocate resources to reduce high severity accidents by eliminating mitigatable factors (e.g. improving lighting conditions at specific junctions) and ameliorate the consequences of un-mitigatable accidents (e.g. deploy more medical/evacuation personnel at regions where and/or periods during which high severity accidents are likely to occur to increase survivability of those involved).

Data Acquisition & Understanding

Data Source

In order to answer the question on which factors affect the severity of accidents, the data should include information/attributes on the weather conditions, location, number and types of parties involved, other event factors and preferably the labelled data attribute of accident severity.

For this report, the data source kindly provided by the course instructors here was used. Metadata of the dataset can be found here.

Data Description

The provided dataset contains 194,673 entries (rows), with 38 different features (columns). Each entry contains information regarding an accident incident, generally including information on:

- **Severity of the accident**
  - This includes severity class/code, severity description
- **Location of the accident**
  - This includes (X, y) coordinates, address, location type, junction type
- **Date-Time of the accident**

    - This includes the date and the time
- **Environment conditions**
    - This includes the weather, road surface conditions, lighting conditions
- **Parties involved**
    - This includes the number of pedestrians, vehicles, cyclists involved
- **Event information**
    - This includes information on the type of collision, the description of the collision and if the vehicle was speeding

Feature Selection

An initial looking into the feature and dataset was carried out to identify the few key relevant features for subsequent processing and analysis.

There are several redundant features found in the dataset. For example, *SEVERITYDESC* describes the type of severity, either as "injury collision" or "property damage only" and this is similar to the information *SEVERITYCODE* presents. It is necessary to remove *SEVERITYDESC* or we run the risk of causing target leakage. *SEVERITYCODE* will be kept as a feature.

There are other features such as *ST_COLCODE, ST_COLDESC, COLLISIONTYPE, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM* which contain information about the collision event, i.e. how the accidents occur, whether it involves rear-ending, accident at an angle etc. While these are useful, they are information that are generated after the accident has occurred by SDOT and are not readily and reliably available before and during the collision, they are thus removed.

The *SPEEDING* feature presents information on whether speeding (i.e. speed above a stipulated speed limit) was a factor of the collision or not. This feature is surprisingly not very useful as it does not indicate how much above what speed limit (so that a change to the speed limit may be prescribed), nor how this classification was derived. It may have been subjectively derived by SDOT (e.g. if speeding happened, but SDOT did not judge it to be a factor). The *INATTENTIONIND* feature has similar attributes and thus would also not be included.

Other redundant features were discarded and the table below shows a list of features that will be kept for the subsequent Exploratory Data Analysis (EDA).

TABLE 1. INITIAL FEATURES SELECTED

Feature Categories	Features to keep	Features to drop
Severity of incident	*SEVERITYCODE*	*SEVERITYDESC*
Junction & location	*ADDRTYPE, JUNCTIONTYPE, CROSSWALKKEY, SEGLANEKEY*	*INTKEY, X, Y, LOCATION*
Date Time	*INCDATE, INCDTTM*	-
Parties Involved	*PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, WEATHER, ROADCOND, LIGHTCOND, HITPARKEDCAR*	-

Event Information	-	*ST_COLCODE, ST_COLDESC, COLLISIONTYPE, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, SPEEDING, INATTENTIONIND*
Others	*STATUS*	*OBJECTID, INCKEY, COLDETKEY, REPORTNO, UNDERINFL, PEDROWNOTGRNT, EXCEPTRSNCODE, EXCEPTRSNDESC*

Data Cleaning

Firstly, the selected data *INCDTTM* and *INCDATE* column values were converted to the appropriate datetime format, for better manipulation. This generated a number of missing values, in particular from the *INCDTTM* where there are a lot of missing time values. We decided to keep these entries during the EDA, but eventually we remove these missing value entries when we eventually decided that the time feature was of importance to our analysis, and that the SEVERITYCODE distribution within the entries with missing time values were similar to the overall dataset true distribution.

To resolve the other missing values coming from *ROADCOND, LIGHTCOND, WEATHER*, we used the filtered out those that had 'Unmatched' values in the *STATUS* column as these had a high proportion of missing *ROADCOND, LIGHTCOND, WEATHER* values. For the remaining entries with null values, they constitute a small proportion <1% of the total dataset and were also removed.

For the remaining missing values in *ADDRTYPE, JUNCTIONTYPE*, we found that a proportion of them had missing values in both *ADDRTYPE, JUNCTIONTYPE* and these were dropped as they will not provide much value. We would that a large proportion of the remaining missing values in the *JUNCTIONTYPE* column contain the *ADDRTYPE* 'Block' value. As it is unknown what the led to the missing values, we decided to relabelled all remaining missing values in these two columns as 'Unknown'.

Data Processing

As we've kept a couple of categorical features, we utilised both Label Encoding during the EDA to visualise data correlation, and One-Hot Encoding on the final feature set so that they can be input into the model. To enable ease of manipulation, the feature names were all also converted to lowercase.

We created a *day*, *month* and *hour* feature from the *incdate* and *incdttm* features for EDA and model buildings. Other features such as *severitycode*, *hitparkedcar* were all relabelled to appropriate values of 0 and 1 according.

Exploratory Data Analysis

Target Variable: Severity

The main target variable in this case is accident severity, represented by the feature column *severitycode*. It was found that the dataset is imbalanced, with the proportion of the minority class corresponding to *severitycode* of 1 at about 30%.

While this may be the naturally occurring distribution of incident severity, down sampling of the majority class or up sampling of the minority class can be considered as strategies to improve the performance of the model. However, as a start, we will build the model with the original dataset with the true distribution.

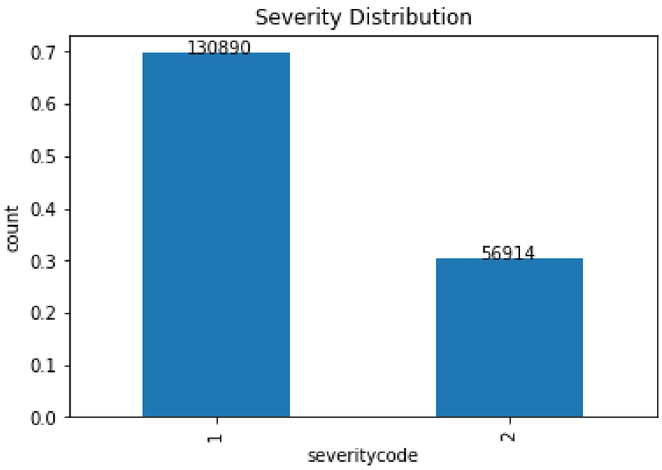

FIGURE 1. DISTRIBUTION OF SEVERITYCODE

Seasonality

A common understanding of accident occurrence is that it varies with timing, day and seasons. As such we looked into the overall seasonality trend of incidents. What we observed In Figure 2 is that there is a general decrease in total incidents per year from 2004 to 2020, which is a positive trend, and may be indicative of the efforts by various parties to promote safe driving environments.

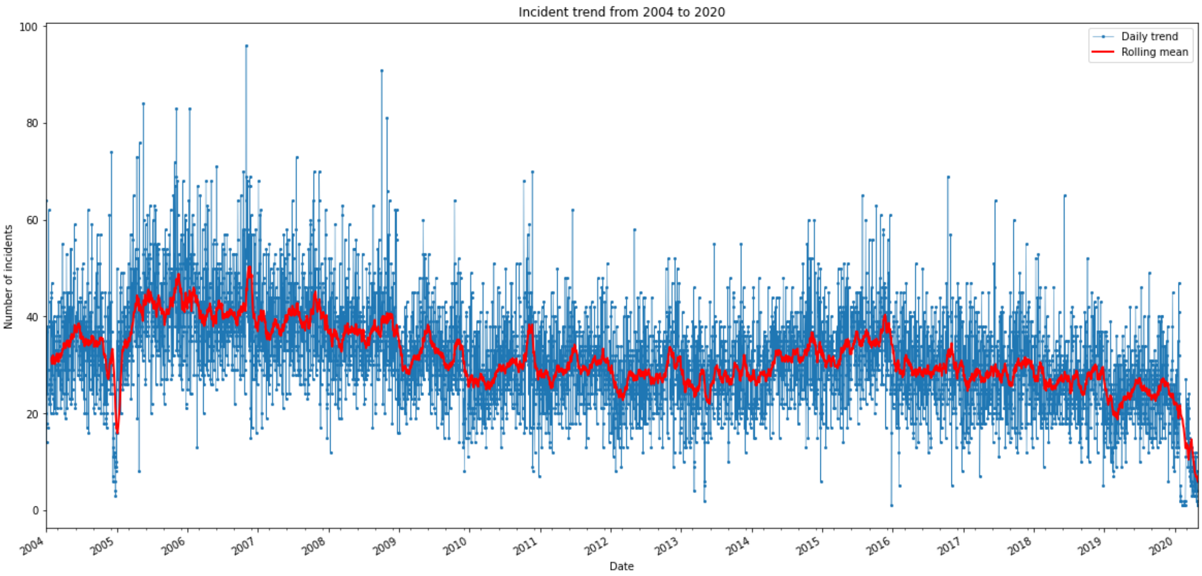

FIGURE 2. CUMULATIVE DAILY INCIDENTS OVER 2004-2020 WITH A ROLLING MEAN TAKEN OVER A PERIOD OF 30 DAYS.

We also found in Figure 3 that while there are 2 slight peaks of incidents occurring during the months of about the about May-June period and the other around October-November. We find that the proportion of incidents however seem stable across the months, with incidents of *severitycode* 1 hovering approximately 28-30% of the total incidents.

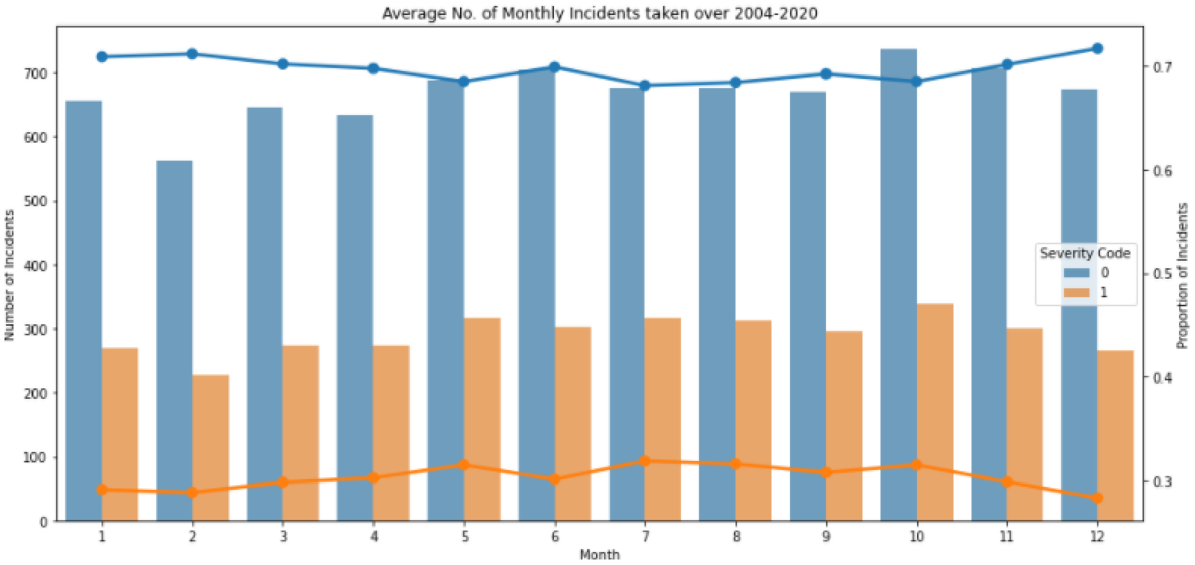

FIGURE 3. AVERAGE MONTHLY INCIDENTS TAKEN OVER 2004-2020 ACCORDING TO THE SEVERITY. DISTRIBUTION OF THE INCIDENT SEVERITY ARE SHOWN BY THE LINE PLOTS.

Looking at the weekday incident pattern in Figure 4, we see that on average, there are more incident occurrences on Fridays (the end of the work week), and a considerably lower occurrence on Sundays. Figure 4 also shows that there is a decreasing proportion of severe incidents as we head into the weekends, but the change is small (31% - 28%).

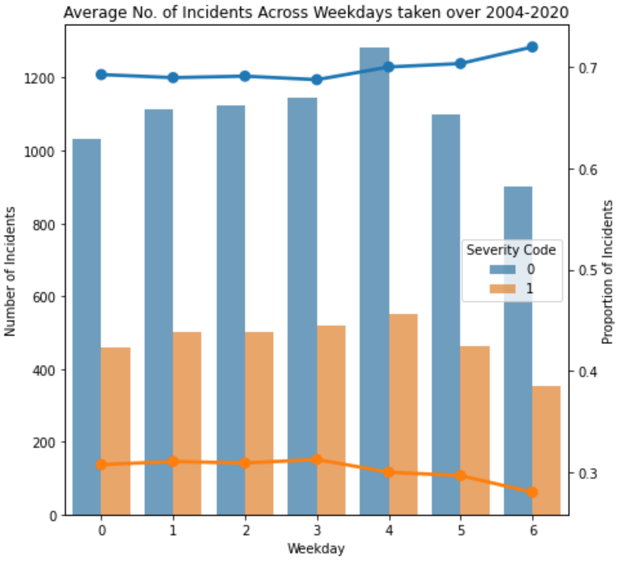

FIGURE 4. AVERAGE INCIDENTS PER WEEKDAY TAKEN OVER 2004-2020 ACCORDING TO THE SEVERITY. DISTRIBUTION OF THE INCIDENT SEVERITY ARE SHOWN BY THE LINE PLOTS.

In Figure 5, we see that there are 2 distinct peak periods, one about 8 am when people are going to work and the other at about 5 pm when people are returning from work. The proportion of incidents happening seems to be strongly affected by the hour, peaking at 35% at 5pm from a low of 21% at 3am.

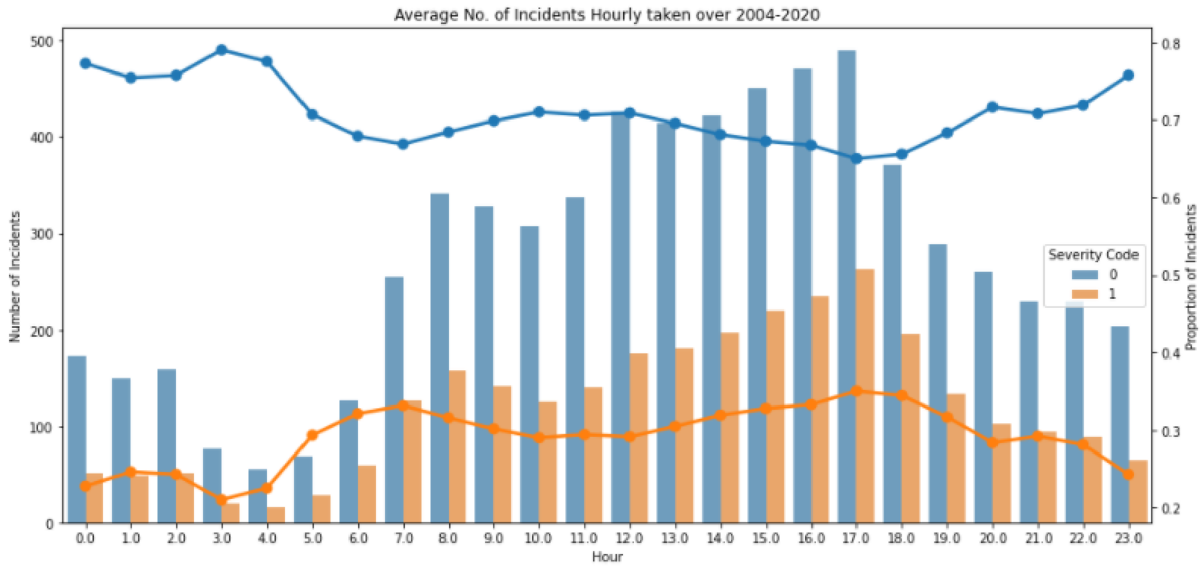

FIGURE 5. AVERAGE HOURLY INCIDENTS TAKEN OVER 2004-2020 ACCORDING TO THE SEVERITY. DISTRIBUTION OF THE INCIDENT SEVERITY ARE SHOWN BY THE LINE PLOTS.

Overall, it seems that the hour feature may be more predictive of the severity of the incident due.

Location Conditions

We see that there are mild correlations between *severitycode* and the *junctiontype* and *addrtype* features (Figure 7). Overall, we see a huge proportion of incidents occurring at the blocks, in particular, mid-block, unrelated to the intersection (Figure 6). However, we also see that there is a strong correlation between *junctiontype* and *addrtype* in Figure 7. We thus dropped one feature (*addrtype* in this case) to reduce multicollinearity and skewed results, given that the information provided by *addrtype* is also captured by the *junctiontype*.

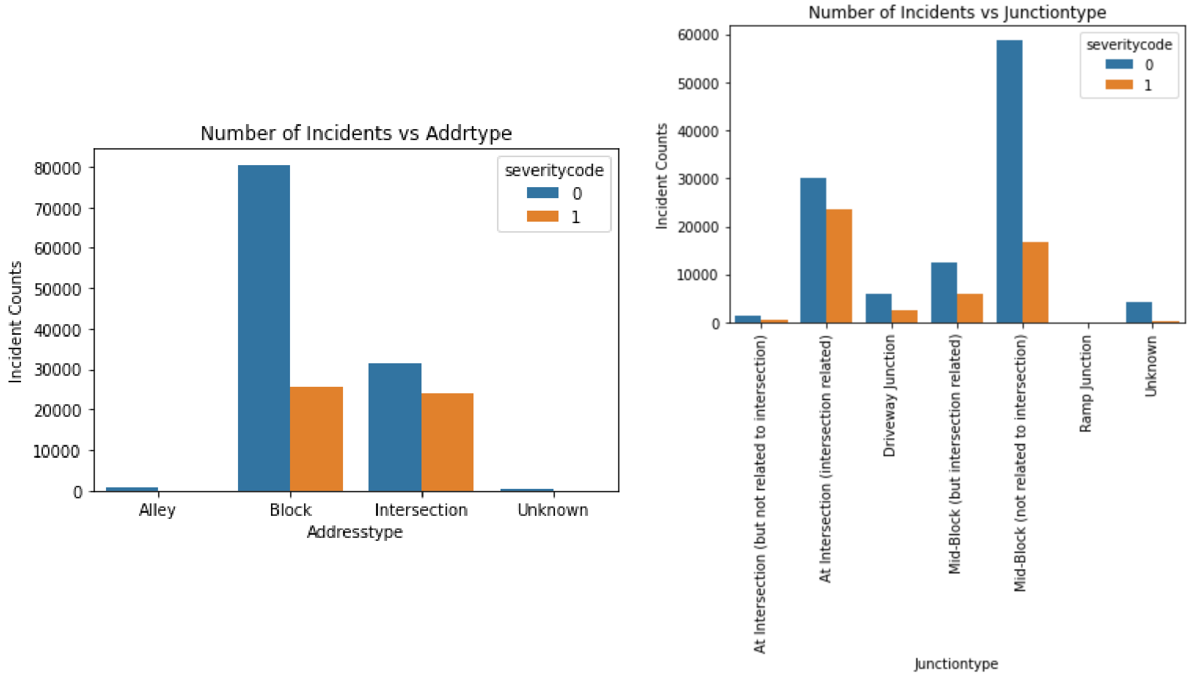

FIGURE 6. NUMBER OF INCIDENTS ACCORDING TO ADDRTYPE AND JUNCTIONTYPE.

For the remaining *seglanekey, crosswalkkey, hitparkedcar* features we see that there are mild correlations to *severitycode*. This is to be expected, given that incidents which involve crosswalk and is likely to involve pedestrians, which is more likely to result in injury related incidents. *Seglanekey* showed some mild positive relation to *severitycode*, this is likely due to the fact that *seglanekey* indicates if the incident happens at the cycling lanes or not.

We will then keep the other features *junctiontype, seglanekey, crosswalkkey, hitparkedcar*.

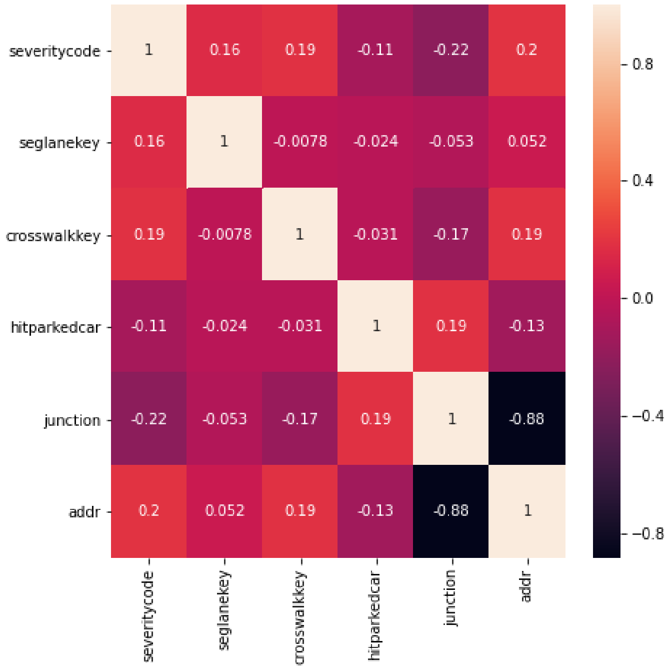

FIGURE 7. CORRELATION HEATMAP OF LOCATION FEATURES

Environment Conditions

We observe that there is some correlation between the severity code and the weather, *lightcond*, *roadcond*, with the *weather* having the greater correlation to *severitycode*. What we also observe is the strong correlation between the weather and the *roadcond*, which is to be expected, as ice and snow on the road can only be present when the weather freezing rain/snowing.

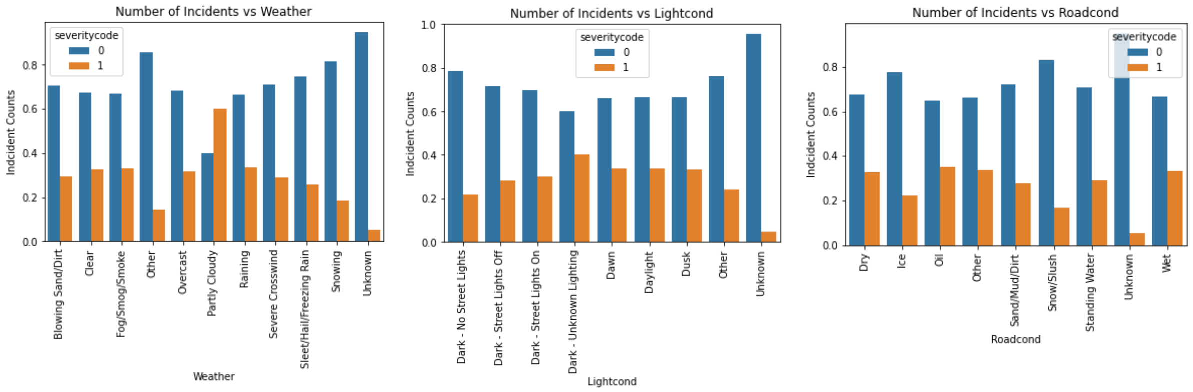

FIGURE 8. NUMBER OF INCIDENTS ACCORDING TO THE VARIOUS ENVIRONMENTAL FEATURES

Ideally, we would remove either *weather* or *roadcond*. However, we see that there are certain *roadcond* values such as Oil and Unknown which are not represented by a corresponding *weather* value. As such, we will keep these features.

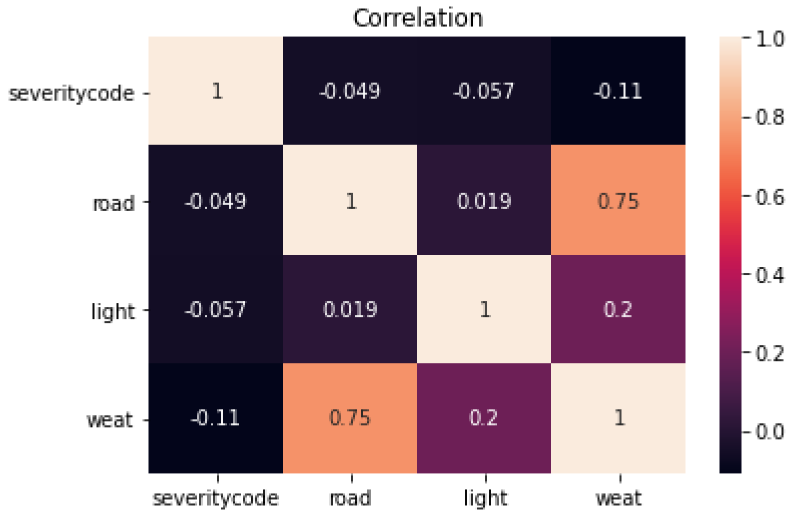

FIGURE 9. CORRELATION HEATMAP OF ENVIRONMENT FEATURES

Others

We explored the remaining features *personcount, pedcount, pedcylcount, vehcount*. Expectedly they show correlation with *severitycode*, as generally, with more people or vehicle involved in the accident, the more likely it is that there will be a serious injury. We also observe expectedly, the strong correlation between these 4 factors

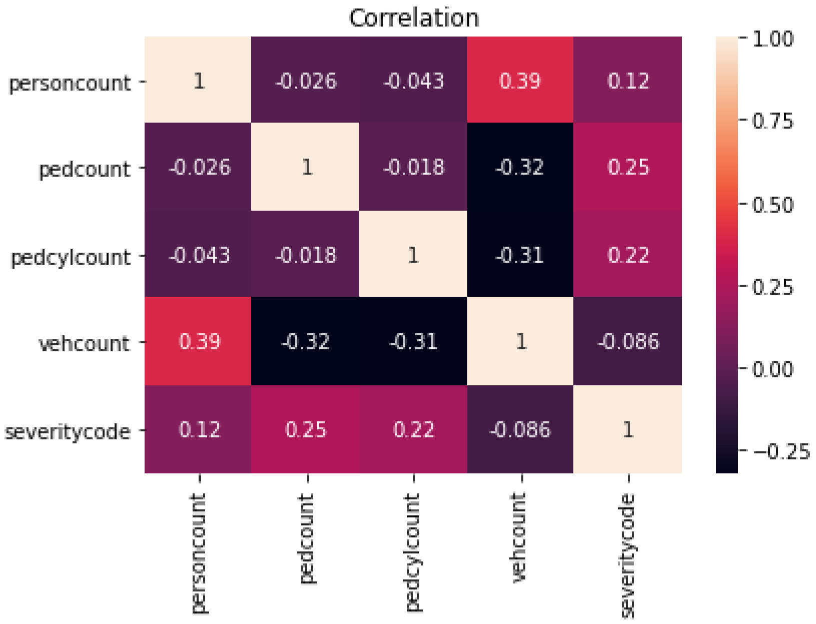

FIGURE 10. CORRELATION HEATMAP OF OTHER FEATURES

We see that *severitycode* is less correlated to the *vehcount* but most correlated to the *pedcount*. This is understandable as the *severitycode* is based on the presence of personnel injury, which has a higher change of occurring if a pedestrian or pedal cyclist (with minimal protection as compared to those in vehicles) is involved.

Model Building

Given that this is a binary classification problem, we explored the use of Logistic Regression (LR) and the ensemble Random Forest Classification (RF) models. Gradient Boosting Classifiers (GBC) which were found to normally provide superior results for binary and multi-class classification tasks was also used.

The dataset was first split 80/20 into the training and test sets. The training set was then normalised via scikit-learn's StandardScaler function to reduce the effect of the different numeric scales of the values in the initial dataset. This will reduce the model's tendency to give weight to large value features. Once we have fitted and transform the training dataset, we use the same StandardScaler model to transform the test dataset.

Following that, we build, train and tested the 3 different models, each using the same modus operandi:

1. Use default model settings
2. Run modelling process on different subsets of the training dataset using cross validation
3. Evaluate trained model with test dataset

We did not perform hyperparameter tuning in this instance.

Results

From Table 2, we find that in general RF model seems to work better than both the LR and GBC models. However, they all suffer from poor recall scores, which in the prediction of accident severity, is the more important evaluation metrics than precision as a higher recall will minimise the situation of False Negatives, i.e. predicting an incident to have lower severity when in reality, it is severe.

TABLE 2. MODEL SCORES BASED ON ORIGINAL DATASET

Model	Jaccard	F1-Score	Precision	Recall
Logistic Regression	0.22	0.70	0.76	0.23
Random Forest Classifier	0.29	0.70	0.69	0.39
Gradient Boosting Classifier	0.23	0.70	0.76	0.25

It is likely that the poor performance is due to the imbalanced dataset. As such we will balance the training dataset, before re-training the models.

Result of Resampled Dataset

We decide to perform undersampling of the majority cases in the training dataset while leaving the test dataset untouched so that it remains representative of the actual distribution observed in reality. This will provide a better more accurate evaluation of the model.

We undersampled the training dataset so that all the minority class within the training dataset is preserved while the majority cases are reduced to a 1:1 ratio with the minority cases. After undersampling, we ended up with 79,418 entries as compared to 129,881 from before.

From Table 3, we see that with a balanced dataset, we are able to improve all the Recall scores across all models. However, we see that the F1 and Precision scores have decreased slightly. The Jaccard score on the other hand improved, but is still rather dismal.

TABLE 3. MODEL SCORES BASED ON RESAMPLED DATASET

Model	Jaccard	F1-Score	Precision	Recall
Logistic Regression	0.37	0.68	0.71	0.65
Random Forest Classifier	0.34	0.65	0.69	0.63
Gradient Boosting Classifier	0.39	0.67	0.72	0.70

Feature Importances

When looking into the feature importance (Figure 11) of the GBC model, we find that there are only a few main features which contributes largely to the model's predictions, namely, *personcount, pedcount, pedcylcount, vehcount* and *juntiontype*. It is also interesting to see that unknown *roadcond, lightcond* and *junctiontype* conditions contribute largely to the prediction.

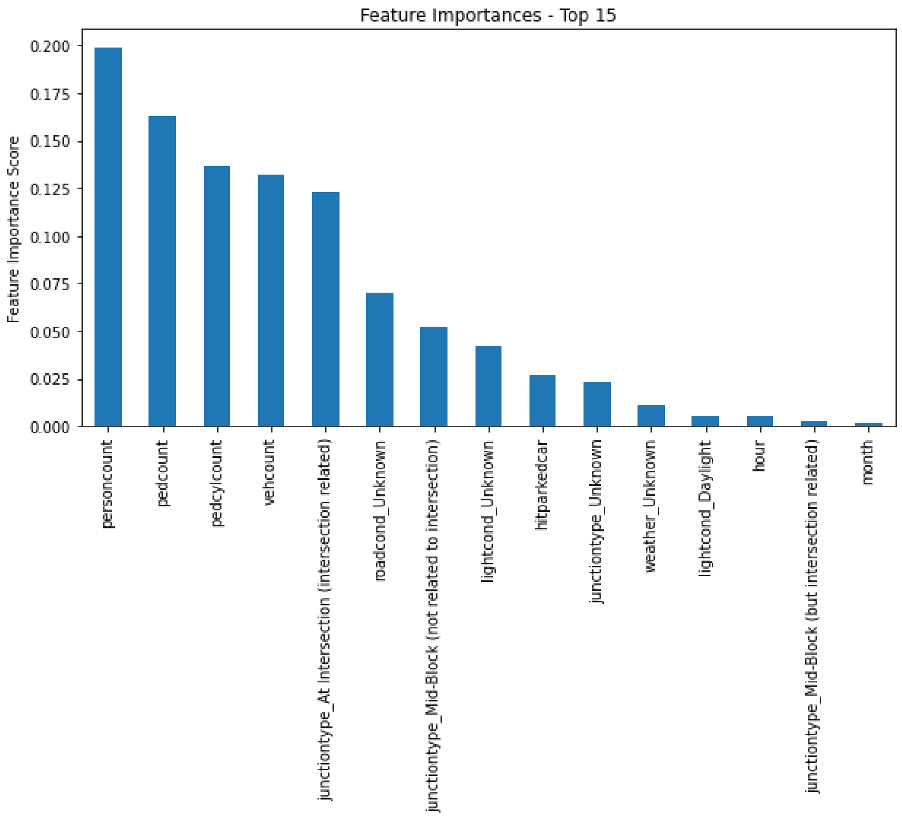

FIGURE 11. TOP 15 FEATURE IMPORTANCES FROM THE GBC MODEL.

Conclusion

We see an improvement in the Jaccard and Recall scores of all models after rebalancing the training dataset. With a balanced dataset, the GBC model has improved and performed better in all aspects than the RF model, with the LR model close in predictive performance.

A closer look into the classification reports above show that most model suffer from poor precision for the severe case, i.e. when *severitycode* is 1, i.e. there is a tendency for the model to make False Positives.

Overall, we see that GBC returns the best performance compared to LR and RF, boasting a recall of 0.70 vs 0.64 and 0.63 from LR and RF respectively. The poor Jaccard score for all models however meant that the models tend to fail in predicting a large portion of the dataset correctly, likely due to poor precision.

Future Directions

To further improve prediction of the models, hyperparameter tuning of the models can be carried out. However, to further greatly improve the scores, it is highly likely that we may have to look deeper into feature or data engineering to obtain better predictive features.

It is also no surprise that the number people, vehicle involved in the incident could be a predictor of the *severitycode*. This insight provides little value and as such, we may consider removing them from future analysis.

A further look into entries with unknown values in the *roadcond, lightcond, junctiontype* should also be carried out to understand why these values are unknown, given the relative importance they are in prediction.