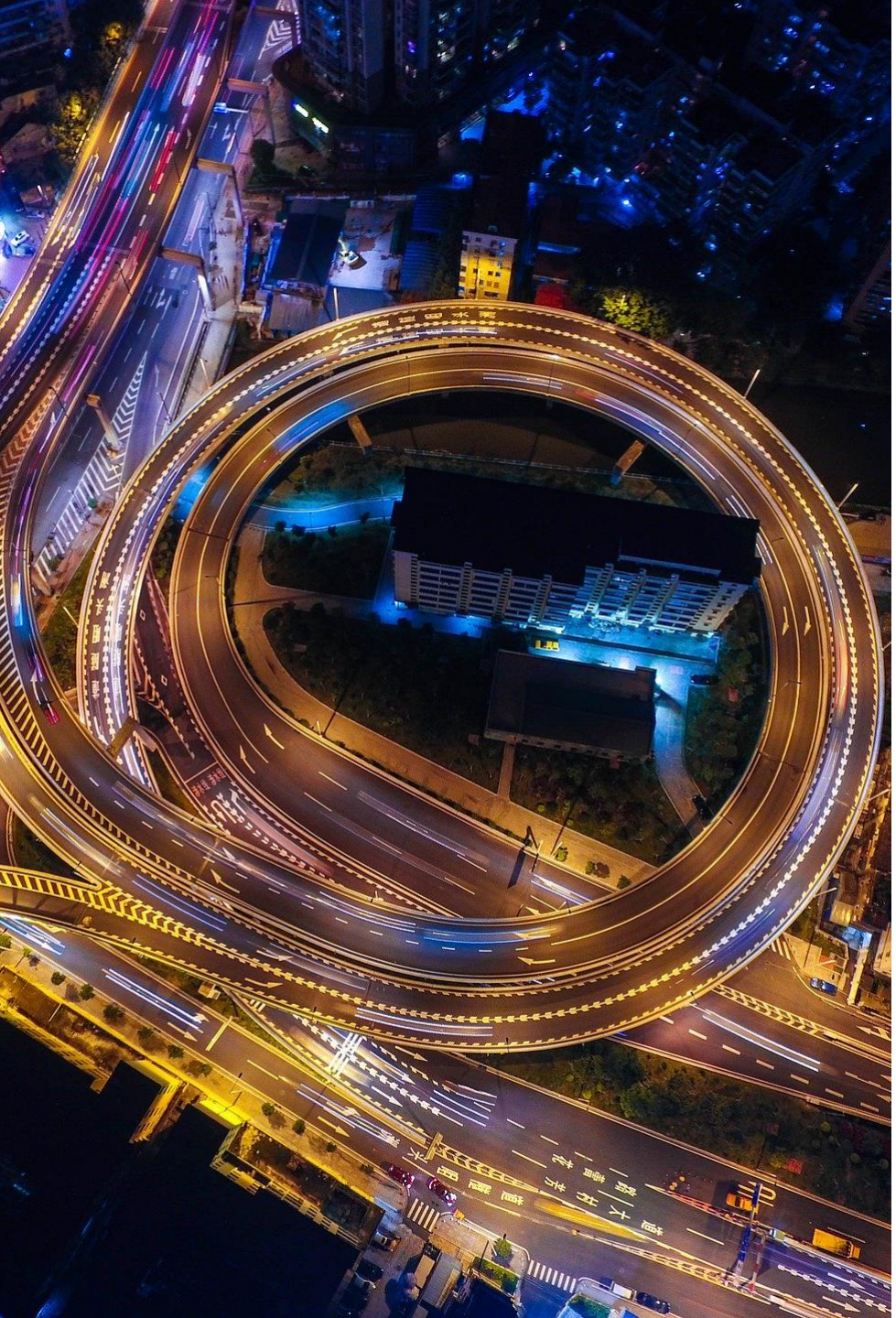


Predicting Traffic Accident Severity

IBM APPLIED DATA SCIENCE CAPSTONE

MARTIN NG

HTTPS://GITHUB.COM/GALACTICPENGWIN/IBM-DATA-SCIENCE-PROFESSIONAL-CERTIFICATE

Problem

Traffic accidents lead to a variety of consequences, ranging from altercations, minor property damages to the more severe loss of human lives.

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 important to investigate 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.

Data

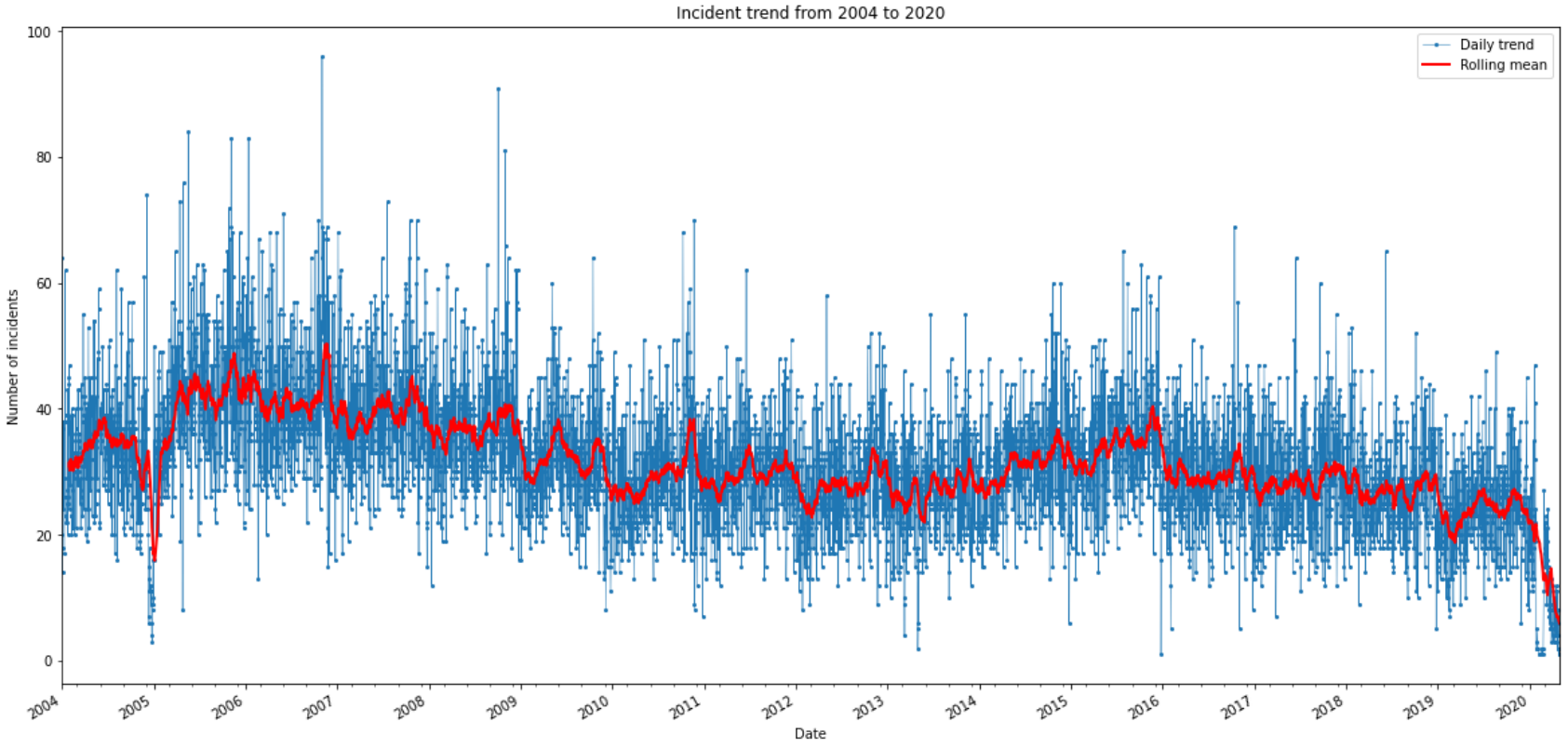
The raw dataset used consists of 194,673 accidents recorded from the period of Jan 2004 to May 2020 occurring in Seattle, US. This dataset was kindly provided by the IBM Coursera course instructors and can be found here. Metadata for the dataset can be found here. The dataset was compiled from various sources such as from Seattle Department of Transport (SDOT).

Irrelevant features, highly correlated features were dropped. A portion of entries with missing values in certain features were dropped, while others were replaced.

Remaining categorical features were encoded via one-hot encoding.

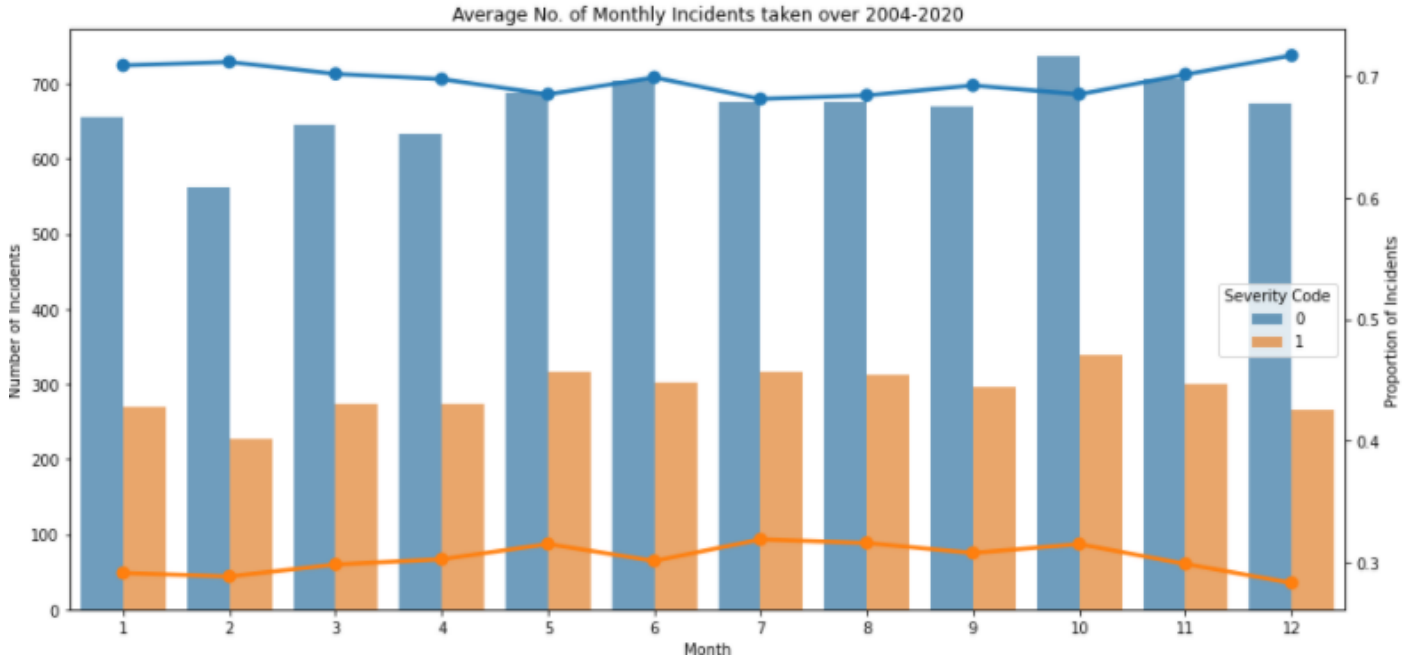
Overall the cleaned data contained 46 columns.

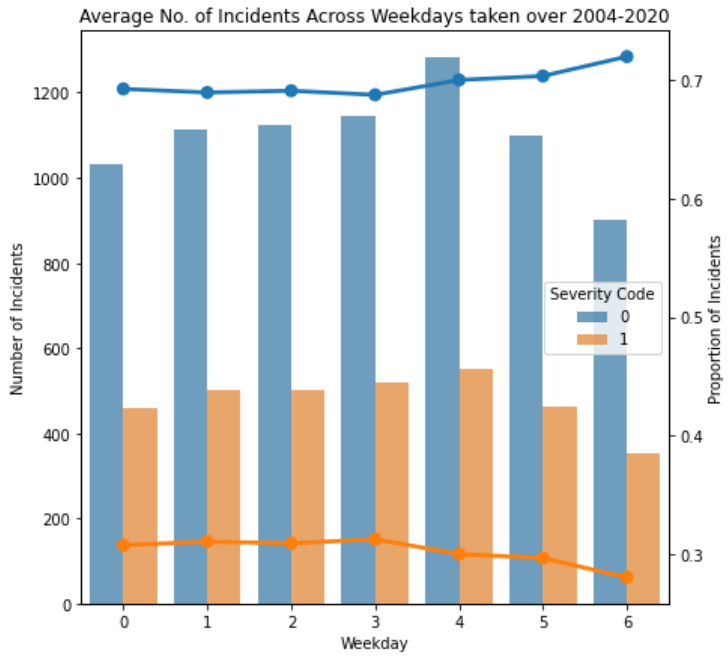
Seasonality



There is a general decrease in total incidents per year from 2004 to 2020, and may be indicative of the efforts by various parties to promote safe driving environments.

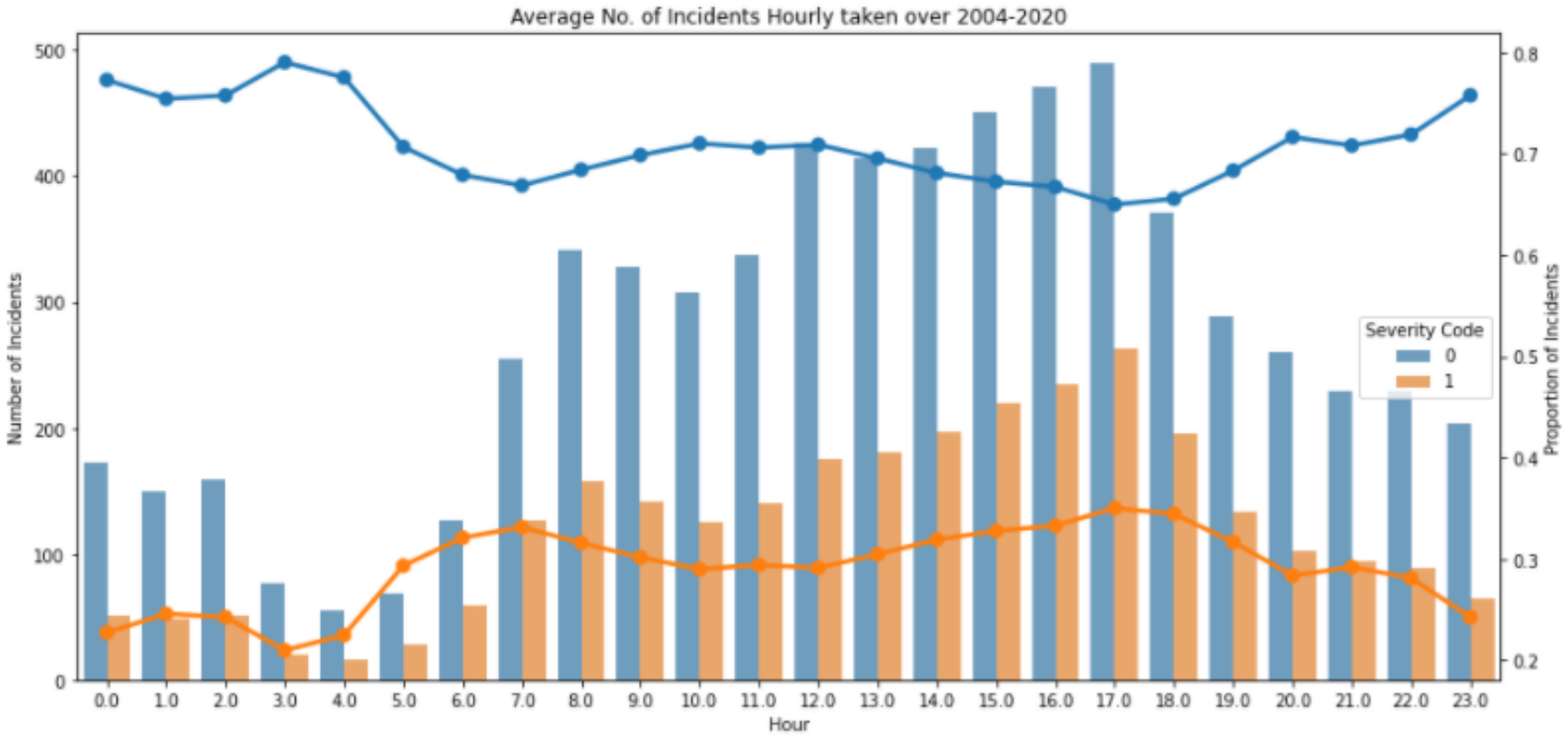
Monthly & Day of Week Trend





- 2 mild peaks occurring during May-June & Oct-Nov periods.
- Peak on Friday, general decrease in incident occurrences towards the weekend.
- Overall, weak trend observed in the average monthly and day of week incident rate. Proportion of severe incidents remained rather stable, between 28-31%.

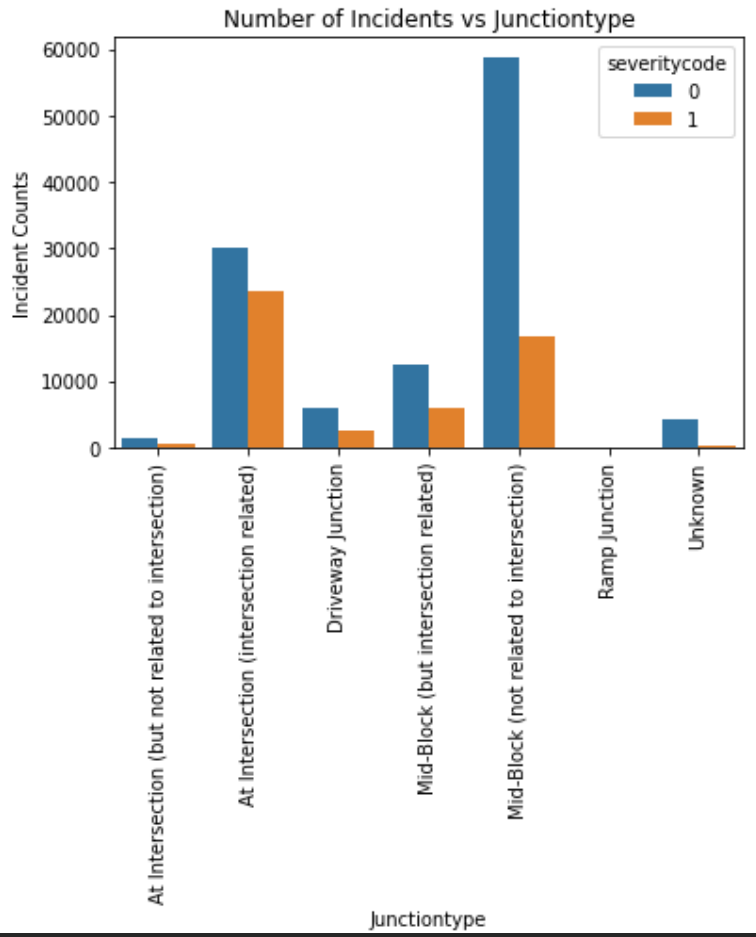
Hourly Trend

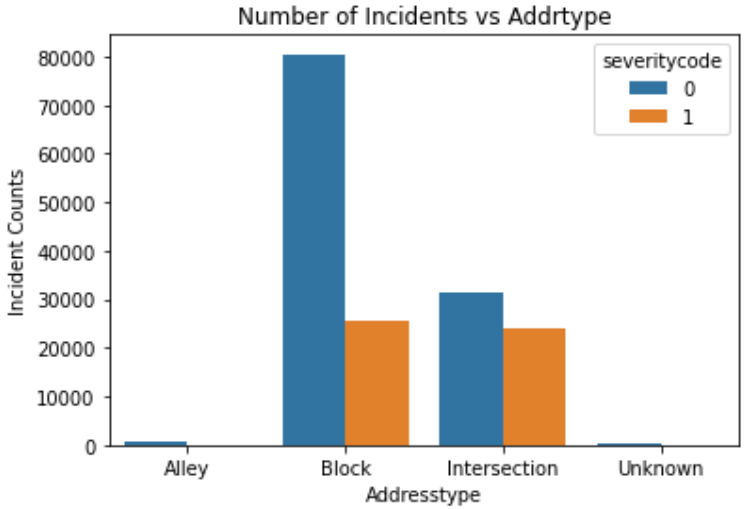


2 distinct peak periods, one about 8 am and 5 pm when people are going and returning from work.

Proportion of severe incidents seems to be strongly affected by the hour, peaking at 35% at 5pm from a low of 21% at 3am.

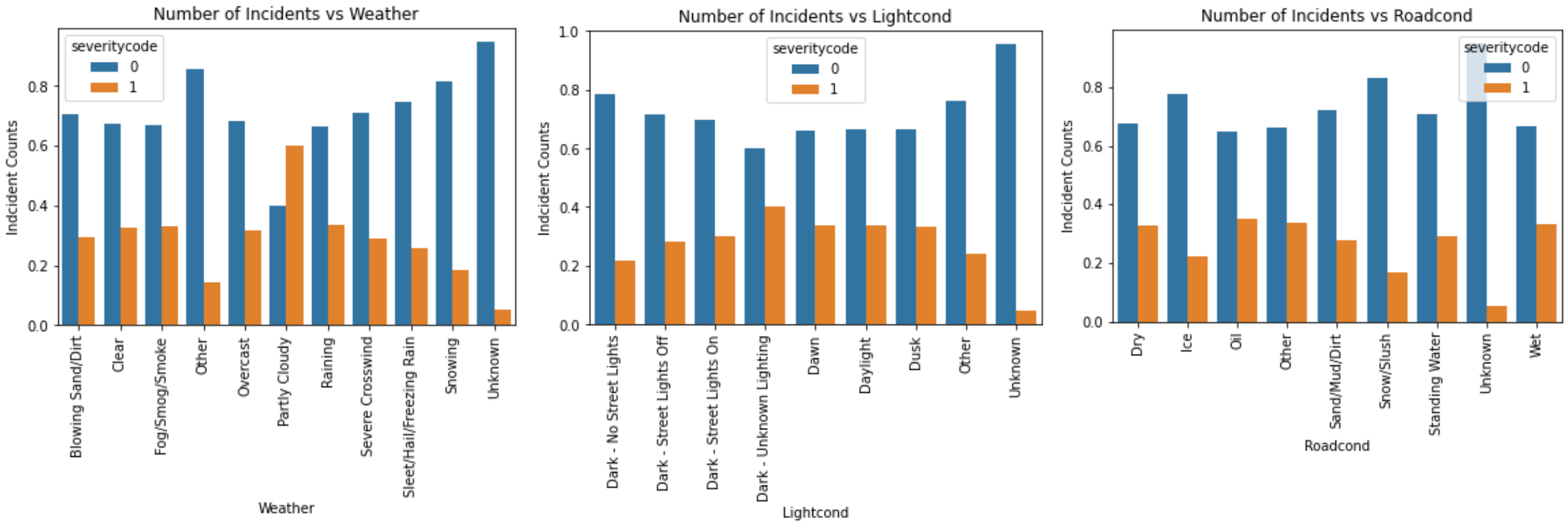
Location





- Proportion of severe incidents seems to be strongly affected by the ***junctiontype*** and the ***addrtype*** features.
- We observe that incidents are more likely to occur at intersections and blocks, however, the proportion of severe incidents occurring at intersections is higher.

Environment



- Here we see huge variances in the proportion of severe incidents, across all environmental features. However, we also see that unknown conditions tend to lead to a lower proportion of sever incidents. We may wish to investigate this occurrence.

Classification Model

Logistic Regression (LR)

Random Forest Classifier (RF)

Gradient Boosting Classifier (GBC)

1. Used default model settings
2. Ran modelling process on different subsets of the training dataset using cross validation
3. Evaluated trained model with test dataset

Initial Results

In general RF model seemed 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. They also have poor Jaccard score.

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

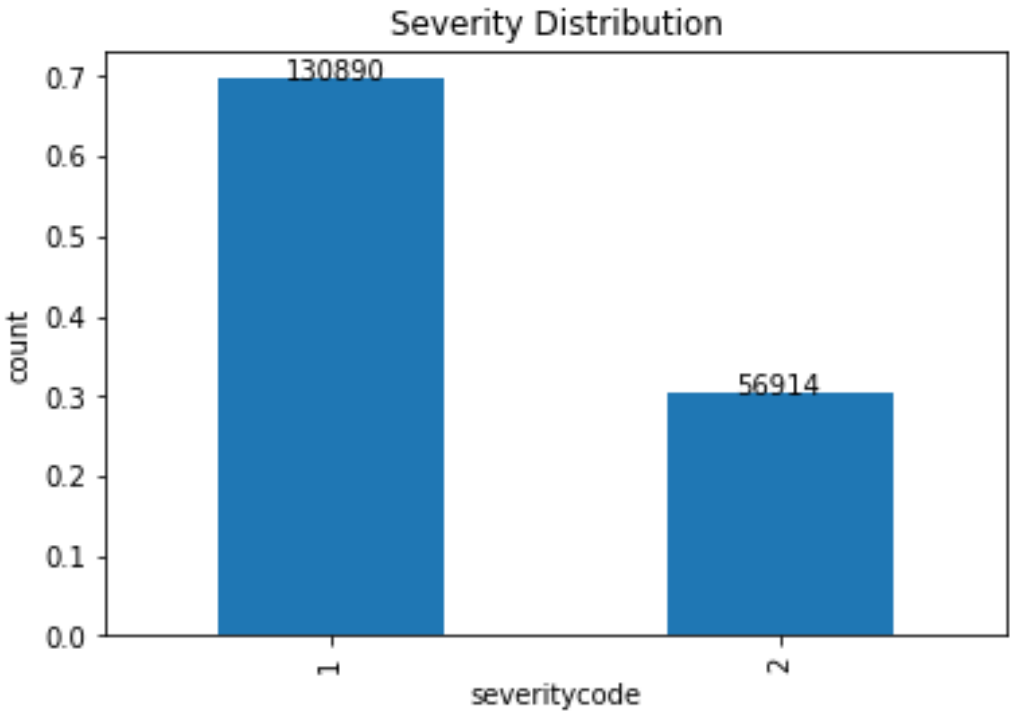
Imbalanced Dataset

Poor performance is likely due to the imbalanced dataset.

In the original dataset, ~30% of the entries correspond to the severe case.

While mildly imbalanced, it may lead to the training model spending most of its time on the low severity cases.

To balance the dataset, we use downsampling, to train on a subset of the low severity cases while maintaining a 1:1 ratio of severe to non-severe cases.



Final Results

With a balanced dataset, Recall scores across all models were improved drastically (from 0.23-0.39 to 0.63-0.70), while F1 and Precision scores have decreased slightly. The Jaccard score also improved, but is still dismal.

GBC model has outperformed both the LR and RF models, with a Recall of 0.70

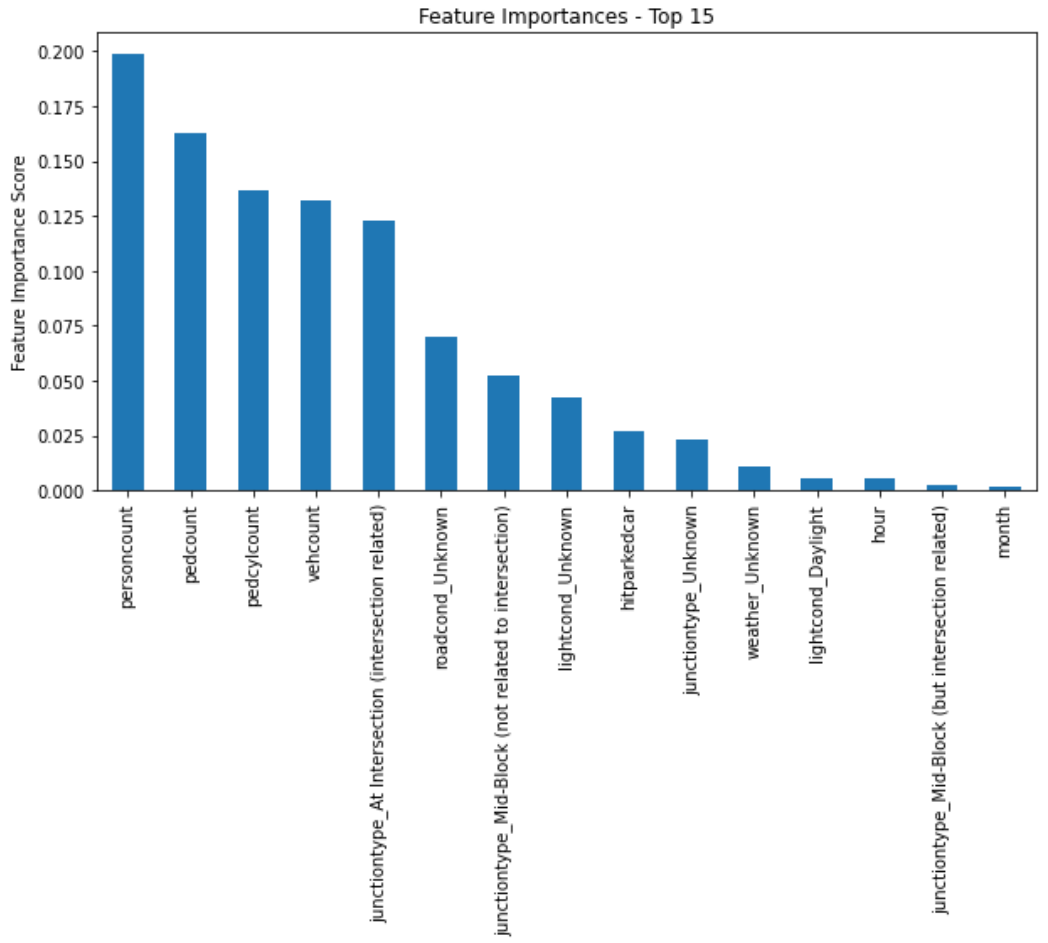
Model	Jaccard	F1-Score	Precision	Recall
Logistic Regression	0.37	0.68	0.71	0.65
Random Forest Classifier	0.34	0.65	0.69	0.63
Gradient Boosting Classifier	0.39	0.67	0.72	0.70

Feature Importances

When looking into the feature importance 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. This should be further investigated.



Conclusion & Future Directions

Built useful models to predict the severity of traffic incidents occurring in Seattle, US.

Accuracy of the models has room for improvement – poor Jaccard, and middling F1-score.

To improve prediction of the models in the future:

- Hyperparameter tuning of the models can be carried out
- However, to further greatly improve the scores, feature or data engineering must be carried out to obtain better predictive features.
- Look deeper into entries with unknown values in the ***roadcond, lightcond, junctiontype*** given the relative importance they are for the prediction.