Analyzing Collision Events in Seattle and Predicting Severe Occurrences

Sayan Das
1/3/20

Why Do We Care About Collision Severity?

Problem

- **6,452,000 motor vehicle crashes** and **37,133 crash-related deaths** in 2017 alone.
- Significant costs in terms of **life, money and property** incurred
- Number of vehicles in the US **keeps growing every year**
- Need to **adopt greater safety measures** on the road

Solution

- Insurance companies play a significant role in **minimizing** and **handling automobile collisions.** Provide services by **charging insurance** premiums
- Need to more **accurately determine insurance rates** through prediction
- Use machine learning to **predict severe collisions** and **explore important factors**

Audience

- Big insurers in the US such as ***Statefarm*****,** ***Esurance*** **and** ***Allstate*** have stated **location as one of their top criteria** for determining rates.

- Insurance companies could use this model to **determine more accurate rates** for their customers.

- Use this information to **alert their clients regarding red flags** when buying expensive cars in more accident prone locations which would likely increase their insurance premiums.

Other Applications

- Alerts for **rideshare** companies

- *Google* and *Apple* who provide **mapping apps**

- **Traffic control departments** could **collect this data firsthand** and make it available to the aforementioned companies.

Data Description

- Data acquired from the City of Seattle Open Data Portal, consisting of **212,760 instances** of vehicle collisions in Seattle with **40 features** with timeframe ranging from **2004-2018**.

- Data for each variable was extracted using the *ArcGIS REST API* in *.csv* format, available on the Seattle GeoData page.

- *Neighborhood* feature extracted using *reverse_geocoder* library. *Tomtom API* used to extract *Speed* variable. *HERE API* used to acquire *Road Length* and *Road Congestion* variables.

Data Cleaning

Process

Selecting Initial Variables

- *Removed redundant variables*
- *Removed variables with many unique categories*
- *Removed variables with many missing values*

Renaming Columns

Handling Missing Values & Formatting

- *Latitude/Longitude*
- *Address Type*
- *Weather*
- *Road/Light Condition*
- *DUI*
- *Junction Type, Collision Type*
- *Severity Description*

Transforming Date Variables

Dealing With Outliers

Adding New Variables

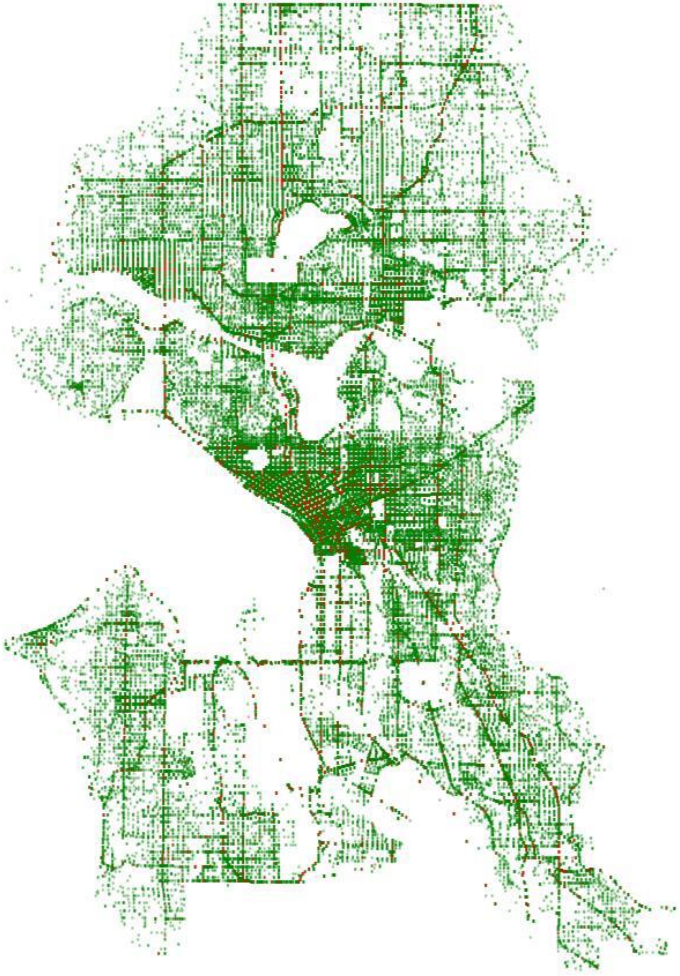
- *Neighborhood*
- *Speed*
- *Road Length*
- *Road Congestion*

Geographical Factors

Key Findings

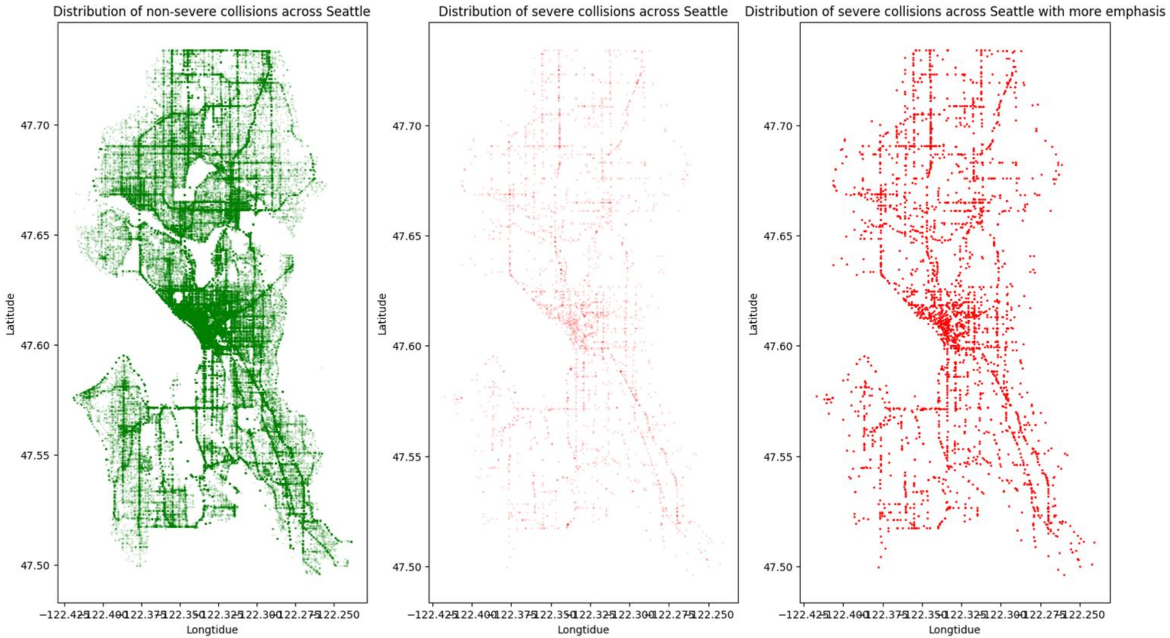
- Most collisions are concentrated in the Seattle neighborhood
- The center of Seattle have the highest density of collisions
- Locations along roads and highways are common collision prone areas

- Collision locations visualized using scatterplot of coordinates
- Points on plot essentially maps out the entire Seattle area
- Severe (red) incidents sparsely distributed compared to non-severe (green) incidents



Key Areas of Dense Severe Collisions

- *Center of the city*: highest density of crashes
- *Aurora Ave North*: road going north
- *Rainier Ave South*: road going south-east
- *15th Ave Northwest*: road going north-west
- *Lake City Way Northeast*: road going north-east
- *24th Ave East*: vertical road towards the east

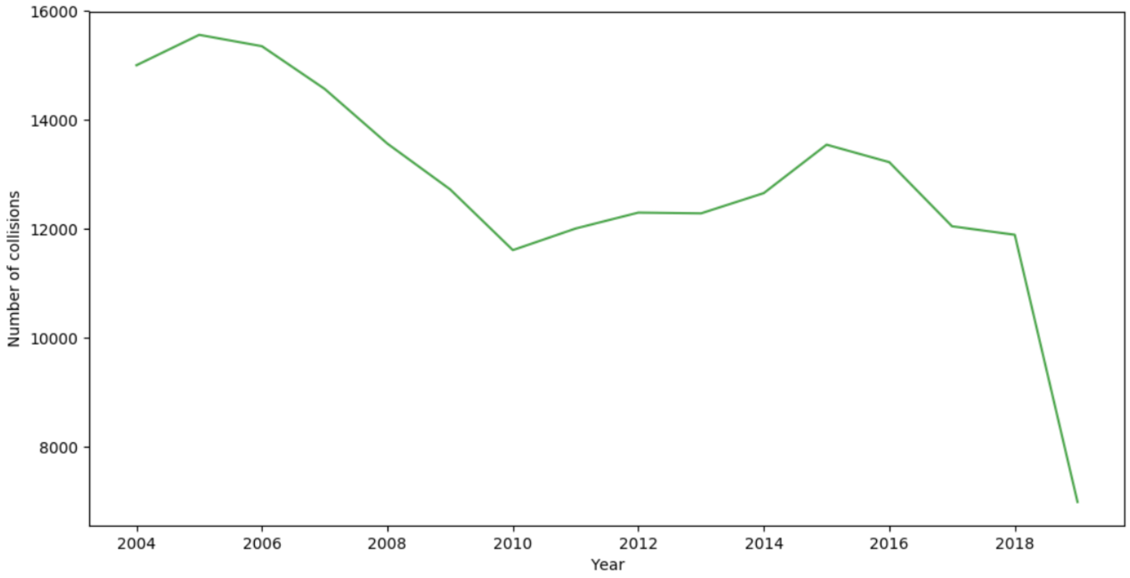


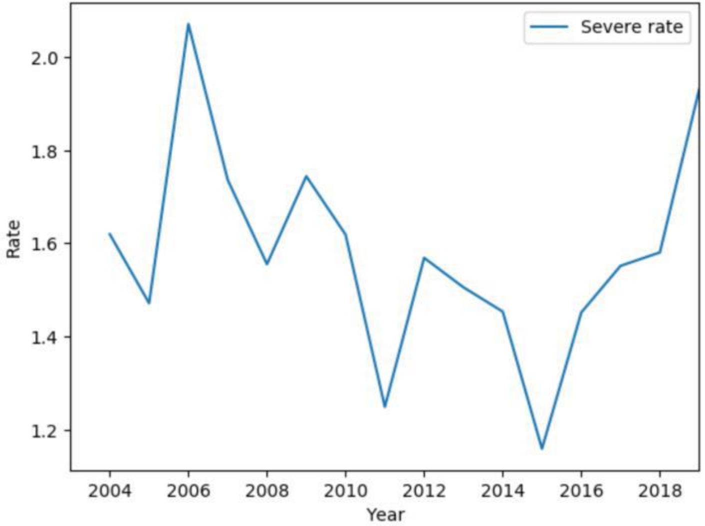
Time Series Analysis

Collisions Over The Years

- Average collisions over the years shows a **downward trend from 2006 to 2011** followed by an **upward pattern till 2016**.

- From **2017 to 2019, there is a sharp downward trend**.





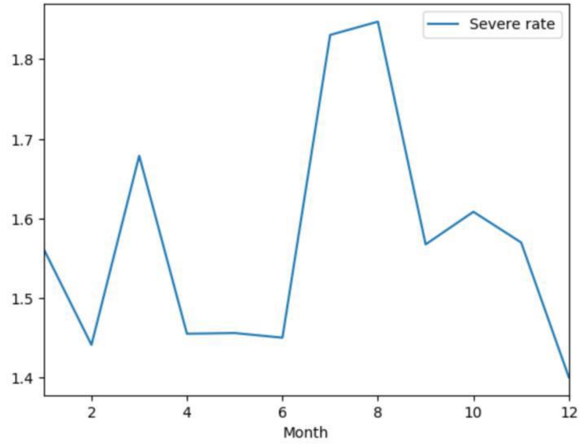
Avg Severity Rate Over The Years

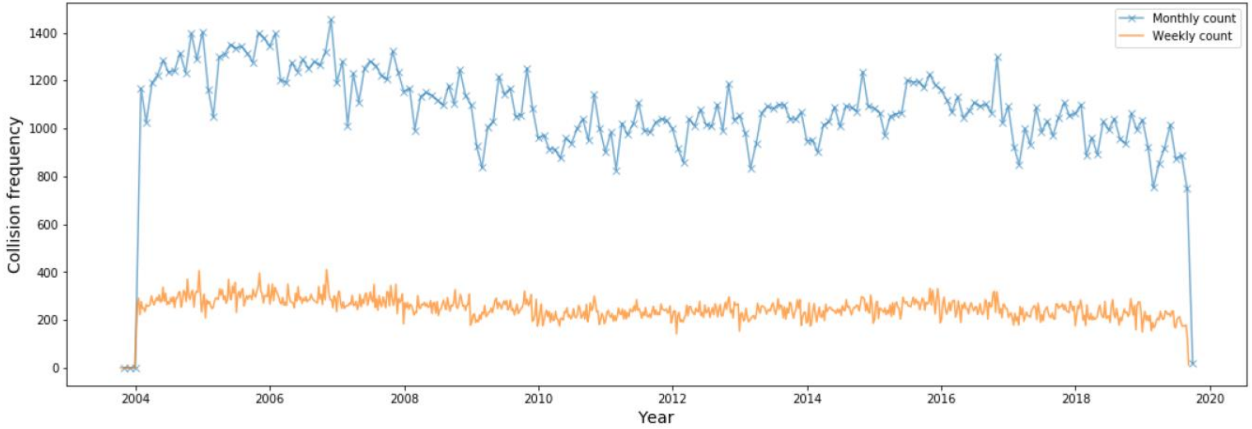
- Sharp increase from 2005-2006 followed by slump till 2008

- Small increase till 2009, then sharp downward trend till 2011

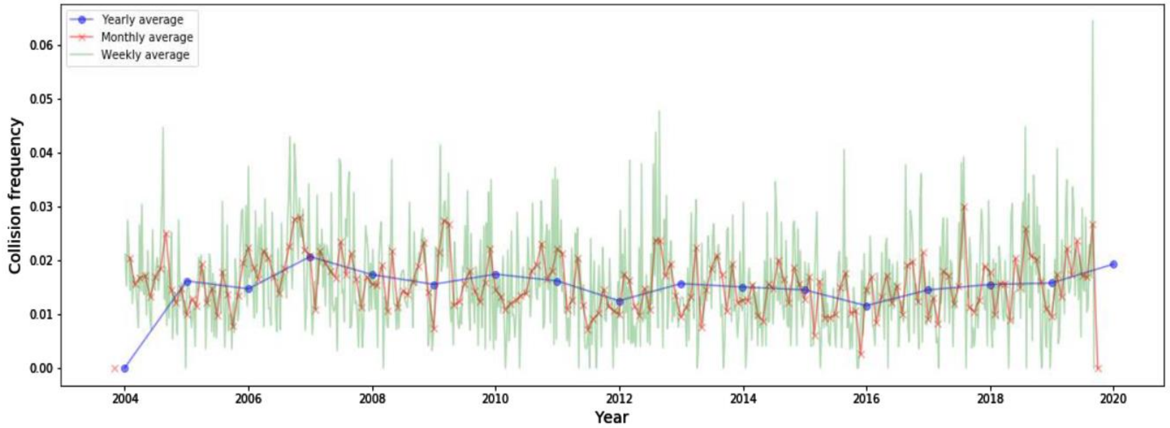
- Sharp Increase till 2012 followed by downward pattern till 2015

- From 2015, sharp upward trend





Average severity rate (top left) is **distinctly higher during July and August**. **April to June** as well as winter months of **Feb and Dec** see the **lowest rates**. Average monthly and weekly plots (top right) shows **indistinct peaks**.



Avg Yearly/Weekly/Monthly Collision Rates Across Time

Points **higher in value represent more severe collisions**. All points have **values closer to 0 due to class imbalance**. Due to the much larger proportion of non-severe cases, an **averaged point is likely to be skewed towards 0**.

Key Findings

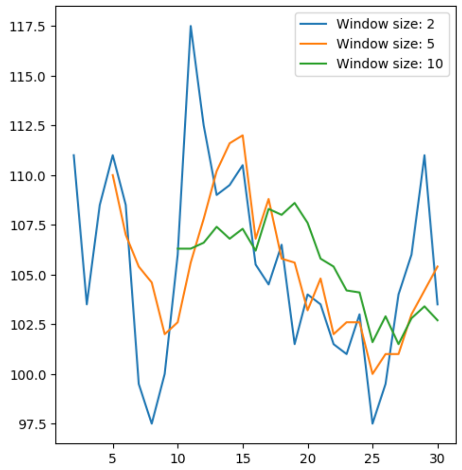
Chi-squared test for independence between day of month and severity suggests a **weak association**. Day of the week vs severity also suggests a **weak relationship.**

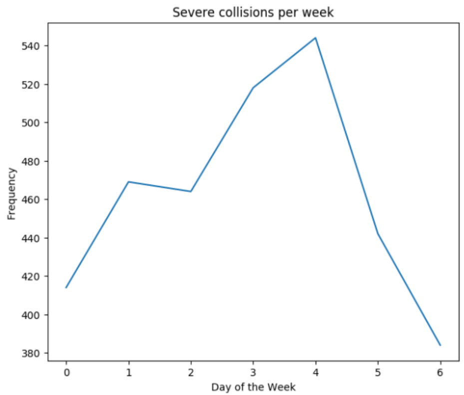
Curve for non-severe (top left) moving avg has a **smoother downward trend**. The severe trend sees **fluctuations during the early days** but has a **downward trend from day 15** onwards.

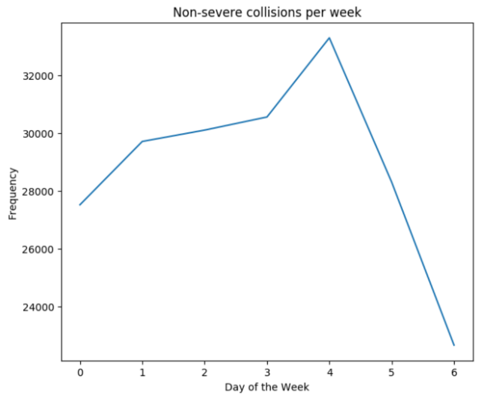
In the bottom plots, **Friday (4) sees the highest avg collisions** while **Sunday (6) sees the lowest**. Trends for severe and non-severe cases are identical.

Day of month vs Severity
Chi-squared: 14.2364
DF: 30
P-value: 0.99

Day of the week vs Severity
Chi-squared: 5.7106
DF: 6
P-value: 0.456377

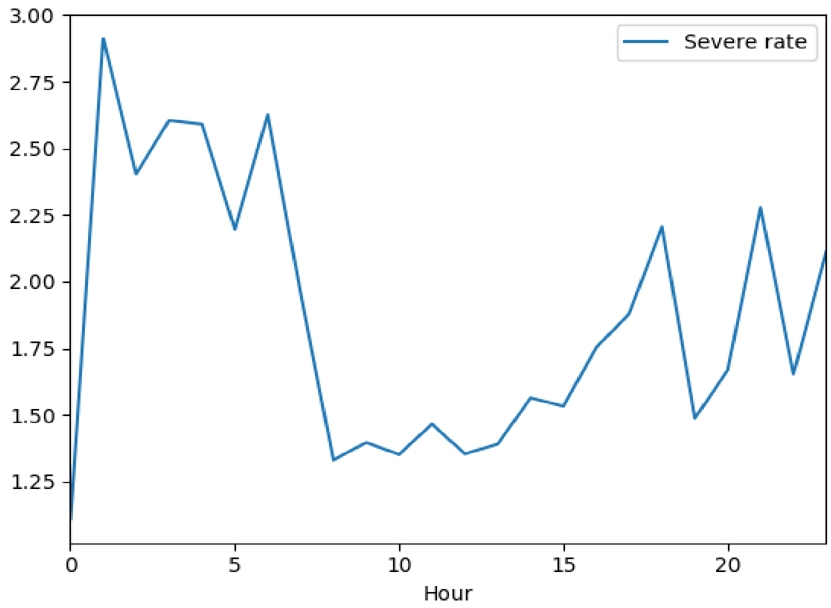






Severity Rate Across Hour of the Day

- Hourly average severity collision rates shown with '0' indicating 12am and '23' indicating 11pm.
- Hours between **1am and 6am** see a **higher rate of severe collisions**. The case is similar for 4pm-6pm, 9pm and 11pm.
- Surprisingly the **the lowest rate is seen at 12am**.
- Hours **between 8am and 3pm see the lowest severe collision rates** which is somewhat expected since light conditions tend to be favorable during this period.
- Chi-squared test between hour and severity shows a **significant association**.



Hour vs Severity
Chi-squared: 249.5802
DF: 23
P-value: <0.00001

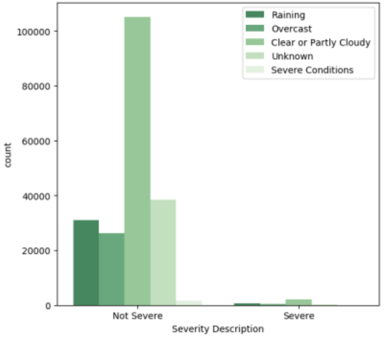
Multivariate Analysis

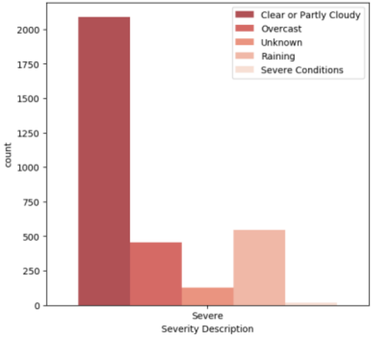
Key Findings

Chi-squared test for independence between weather-related variables and severity suggests a **strong association**.

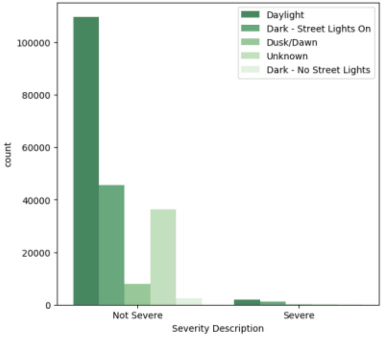
However, bar graphs **do not suggest** an obvious relationship.

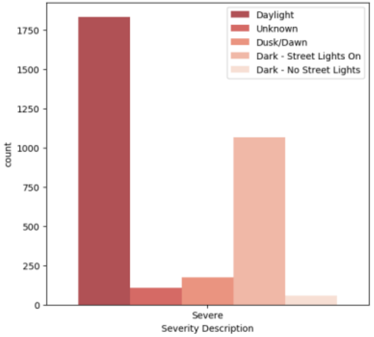
Weather vs Severity
Chi-squared value: 492.9949
Degrees of freedom: 4
P-value: <0.00001



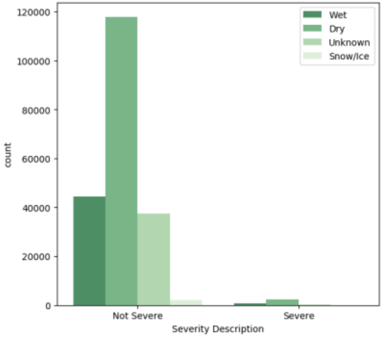


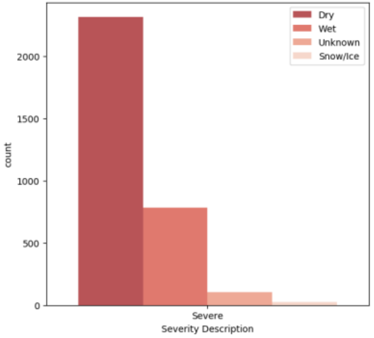
Light Condition vs Severity
Chi-squared value: 566.0597
Degrees of freedom: 4
P-value: <0.00001





Road Condition vs Severity
Chi-squared value:511.8966
Degrees of freedom: 3
P-value: <0.00001





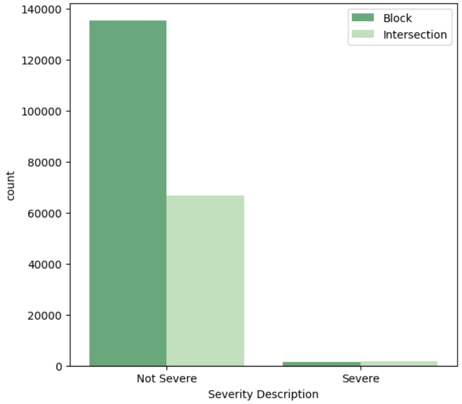
Key Findings

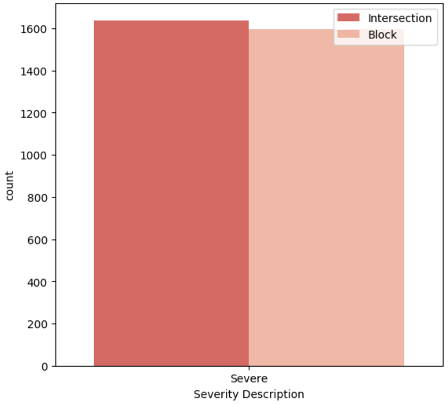
Chi-squared test for independence for both Address Type and Junction Type against severity suggests a **strong association**.

The bar graphs for Address Type suggests that for the non-severe class, the **'block' category is twice more frequent** suggesting some association with severity.

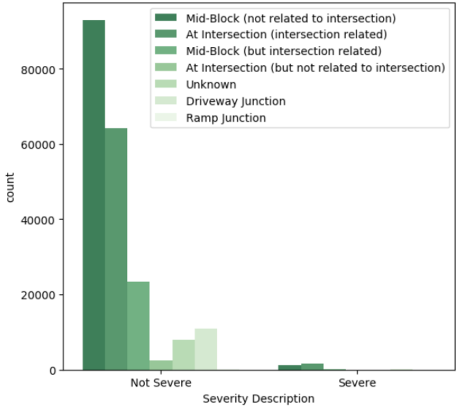
Similarly, for Junction Type, **'mid-block' is more frequent for non-severe** instances whereas **'intersection' is more frequent for severe** cases.

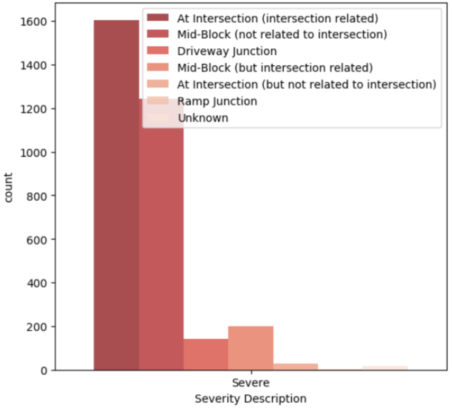
Address Type vs Severity
Chi-squared: 443.8558
DF: 1
P-value: <0.00001





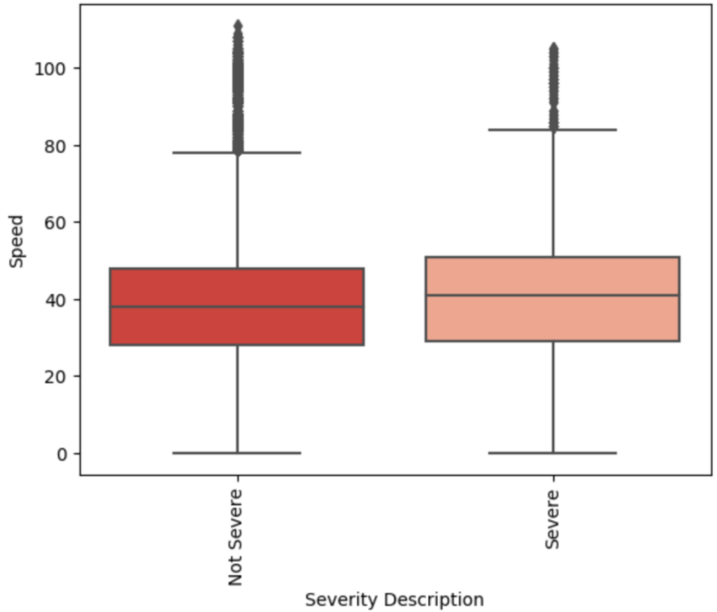
Junction Type vs Severity
Chi-squared: 547.7693
DF: 6
P-value: <0.00001

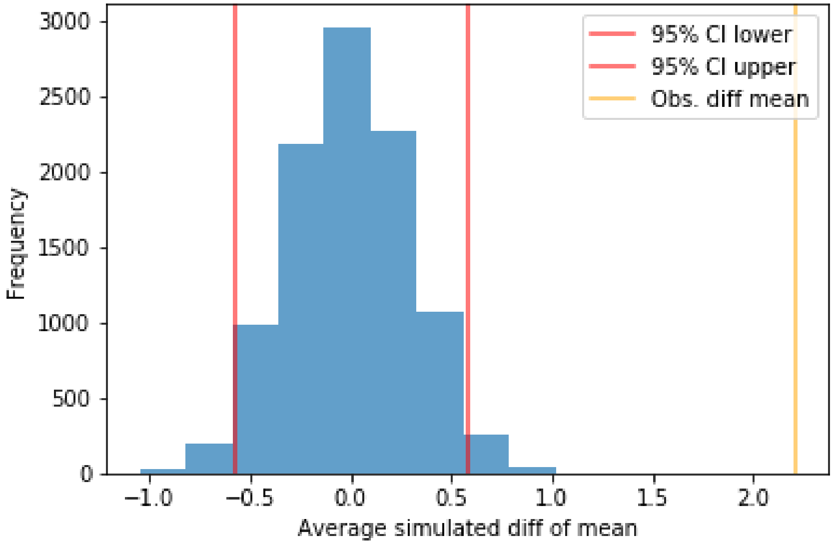




Key Findings (Speed vs Severity)

- Boxplots suggest that median traffic flow **speed is slightly higher for severe cases**.
- Hypothesis testing via simulations using permutation replicates indicates that the **probability of observing the actual difference in mean speed between severe and non-severe cases is significant**, whereby validating the boxplot findings.





Stats

Obs diff mean: 2.212

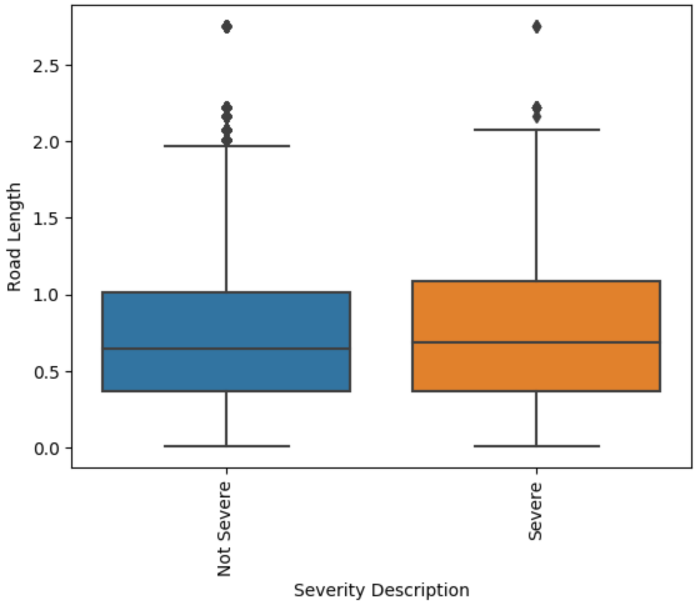
95% CI: -0.599, 0.603

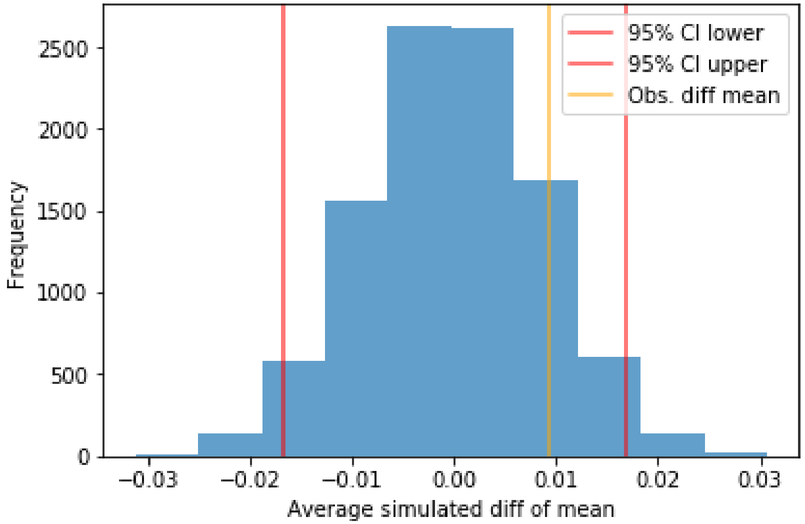
Significance level: 5% (0.05)

P-value: 2.055e-13

Key Findings (Road Length vs Severity)

- Boxplots suggest that median **road length is slightly higher for severe cases**.
- Hypothesis testing indicates that the **probability of observing the actual difference in mean road length between severe and non-severe cases is not significant** (p-value > 0.05)





Stats

Obs diff mean: 0.0093

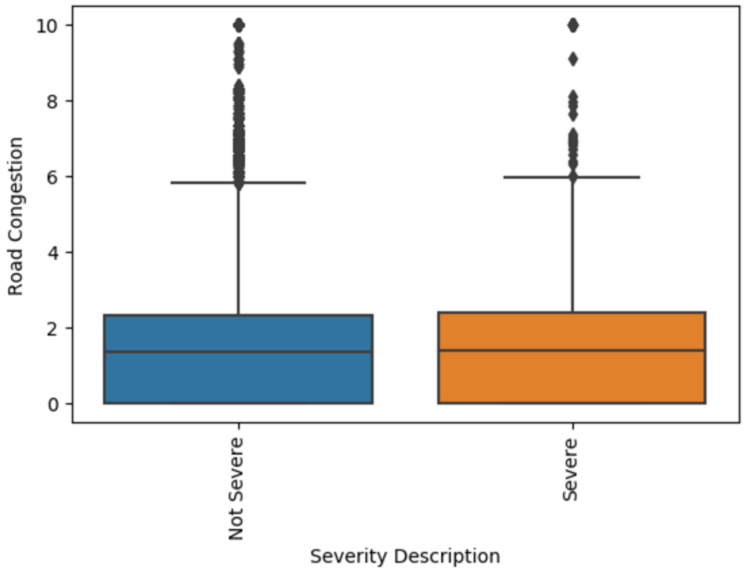
95% CI: -0.166,0.0168

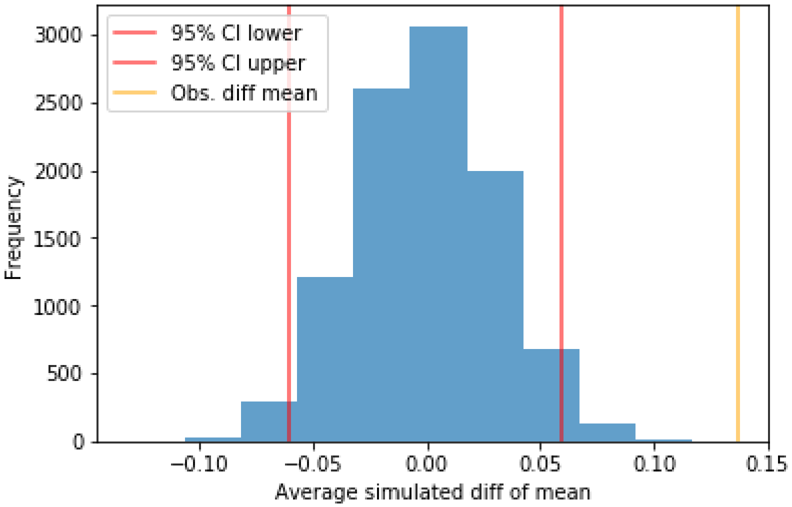
Significance level: 5% (0.05)

P-value: 0.138

Key Findings (Road Congestion vs Severity)

- Boxplots suggest that median **road congestion is slightly higher for severe cases**.
- Hypothesis testing indicates that the **probability of observing the actual difference in mean road congestion between severe and non-severe cases is significant**.





Stats

Obs diff mean: 2.212

95% CI: -0.0601, 0.0597

Significance level: 5% (0.05)

P-value: 8.788e-6

Modeling and Evaluation

Process

Feature Selection

Pre-processing
- *Label encoding*
- *Feature scaling*

Train test split

Model Selection
- *LightGBM*
- *Logistic Regression*
- *Random Forest*

Hyperparameter Tuning
- *Grid Search*
- *Cross Validation*

Model Evaluation
- *ROC AUC, Accuracy, Misclassification*
- *Confusion Matrix*

Tuning ROC AUC Threshold

Final Model

Final Model Details

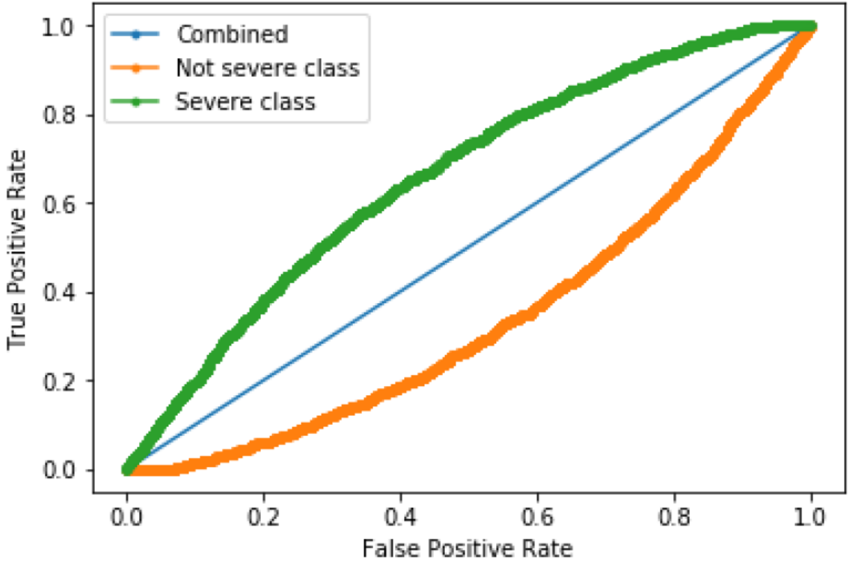
- **Algorithm**: LightGBM
- **Key Parameters**:
  - *Max depth*: 5
  - *Number of leaves*: 10
  - *Learning rate*: 0.1
  - *Min data in each leaf*: 20
- **Threshold**: 0.02

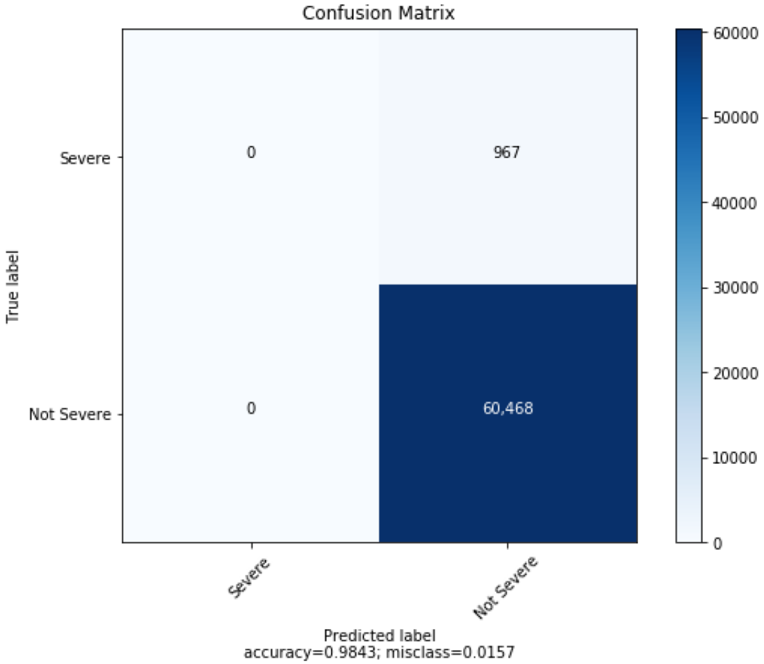
Model Evaluation

- **Best ROC AUC score**: 0.66
- **Confusion matrix**:
  - *True positives*: 493
  - *False negatives*: 474
  - *False positives*: 17697
  - *True negatives*: 42771
- **Accuracy**: 0.7042
- **Misclassification error**: 0.2957

Key Findings (ROC Curve)

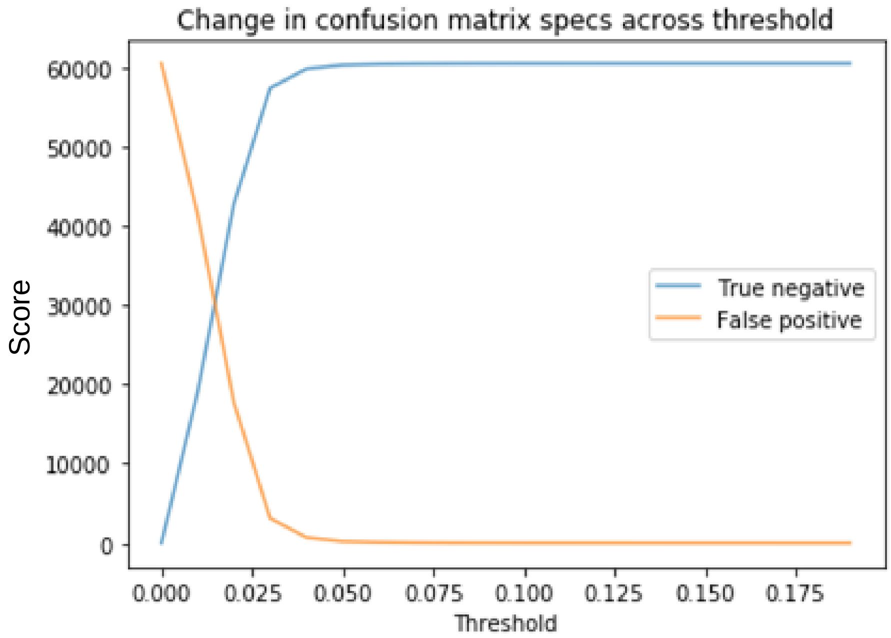
- Final model yielded a severe class **ROC AUC score of 0.66**
- Curves for non-severe and combined classes also shown

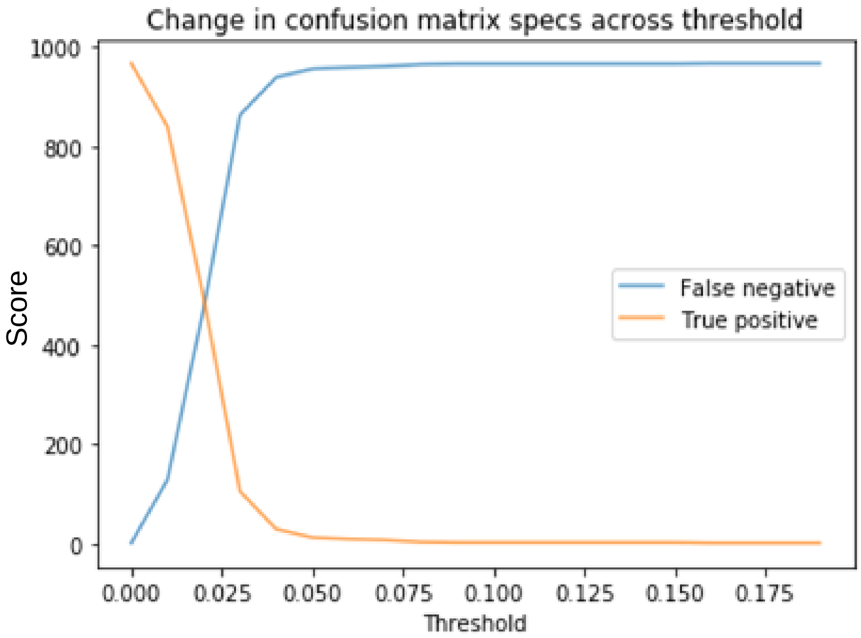




Key Findings (Confusion Matrix)

- None of the severe cases were being predicted due to class imbalance.
- Hence, the thresholds were tuned to increase the number of TP results and decrease the number of FN results (shown on slide 26).





Key Findings (Threshold Sweep)

- Threshold values are swept from 0 to 1 with a step size of 0.01 to visualize changes in the confusion matrix values.
- True positives increase with increasing false positives and decreasing false negatives
- True negatives decrease with increasing false positives
- Significant changes are only observed at a threshold lower than 0.05

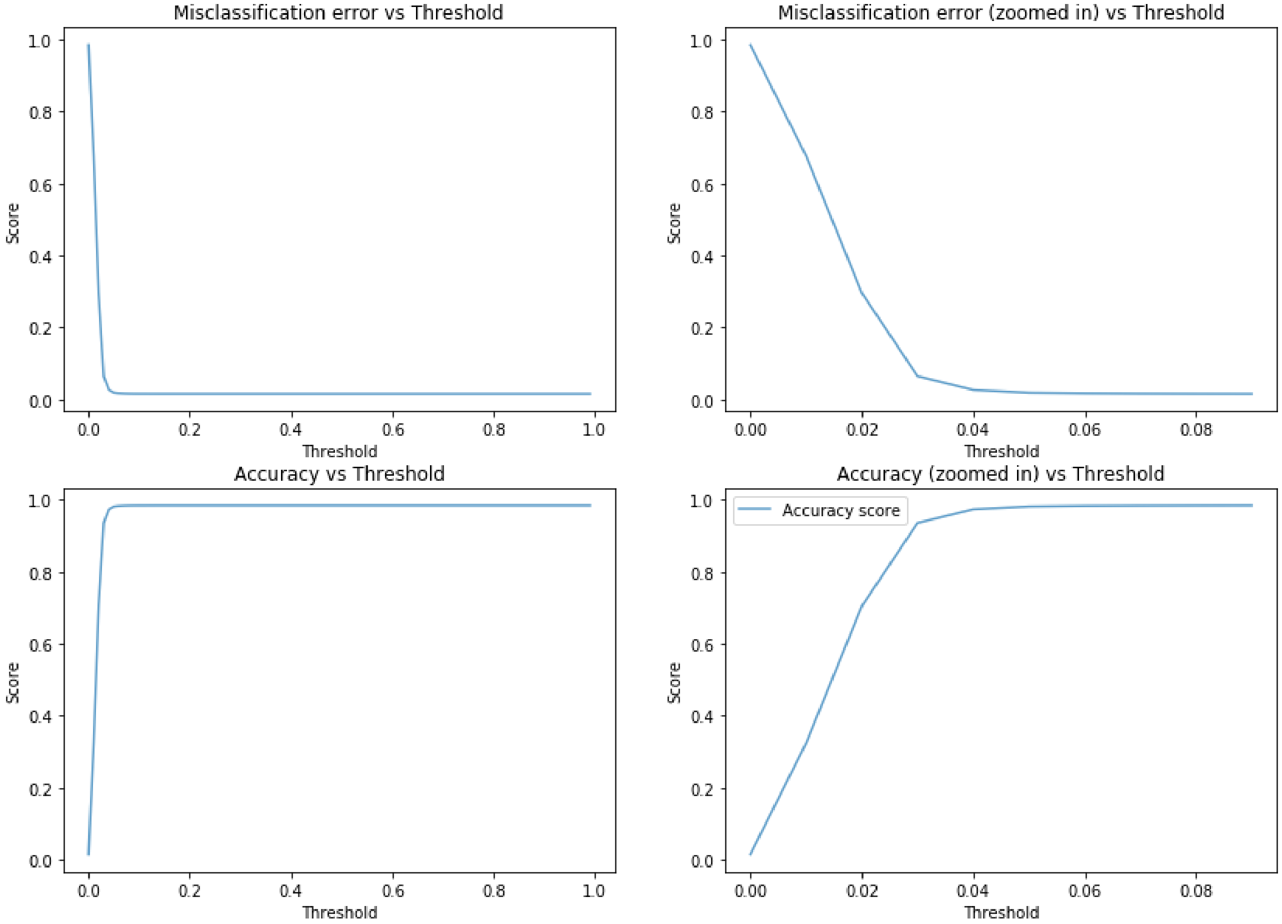
Key Findings (Accuracy and Misclassification)

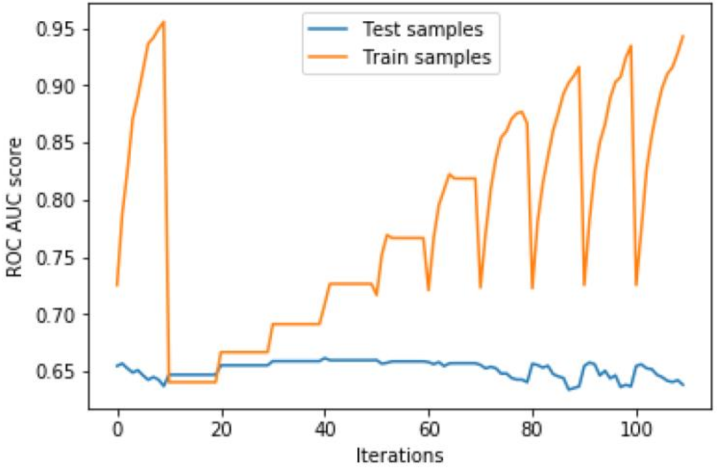
As threshold is decreased, accuracy decreases while AUC remains the same as we are just moving along the AUC curve.

As threshold decreases, events go from TN to FP whereby **decreasing the accuracy and increasing the misclassification error**

Similarly, values go from FN to TP for decreasing threshold.

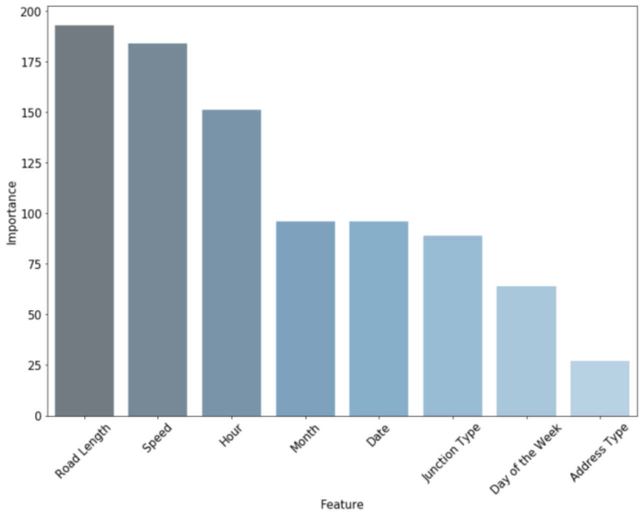
In our application where detecting severe collisions is the most important criteria, this is a **fair tradeoff to consider**.





Handling Overfitting

- Top left graph shows ROC AUC scores for train and test samples across *number of leaves* and *max depth* hyperparameters.
- **Points between 10-20 offer the best model in terms of overfitting**.
- Each test sample point above 20 increases train performance by a higher rate, the **ideal points being between 10-40.**



Important Features

- **Road Length** and **Speed** are the **best predictors** of severe collisions followed closely by **Hour**
- Month, Date and Junction Type have similar effects on the model
- **Day of the Week** and **Address Type** are the **worst predictors**

Conclusion

- The center of the city as well as routes along major roads going north, south and east have a relatively higher density of severe incidents

- Class imbalance made prediction of the severe (minority) class difficult. Oversampling methods such as SMOTE were not effective in dealing with the imbalance problem since not many features explained the variance in severity classes well.

- Location-based features such as ***Weather***, ***Road/Light Condition*** and ***Neighborhood*** provide little information. The more important variables turned out to be ones related to traffic flow, road dimensions, date and time.

- An optimized LightGBM model provided a ROC AUC score of 0.66 and zero true positive values due to imbalance. This was tackled by adopting a threshold value of 0.02 which yielded 493 TP, 474 FN, 17697 FP and 42771 TN. Accuracy was 0.7042 with a corresponding misclassification error of 0.2957.

- Since it's more critical to predict severe collisions correctly, a lower threshold was selected in order to generate more true positives at the expense of a higher false positive rate.