Collision Data Analysis, Seattle City

- Vishal Dongare (September 13, 2020)

1. Introduction

Seattle city is well known for its challenging weather and driving conditions. It represents majority of the mega cities in the USA in terms of many factors that contribute to the severity factor of a vehicle collision. In this project, the collision data from the city of Seattle recorded between years 2004-2020 was analyzed and I have built several machine learning models to predict the severity of the collision based on various factors. I have then compared the models based on various evaluation parameters.

This project is particularly useful for the Seattle police and traffic department, interested in knowing how severe the collision is going to be based on several factors such as road conditions, number of vehicles involved, weather conditions and others. They would also want to understand the impact of different factors on the outcome of such collisions. The whole idea behind this project is to build a robust machine learning model that can predict the severity of a collision and provide insights into how to prevent such collisions so that the drivers can be alerted in time.

I have used various techniques of data science in this project, such as data cleaning, feature engineering, data exploratory analysis, building predictive models, and model evaluation. I have also extracted insights from the results to help stakeholders prevent or limit the severity of such collisions and to help them improve a response time in case of emergencies.

2. Data

2.1 Data source

The data for this project was collected from the shared data included in IBM's Applied Data Science Capstone Project (link). This dataset includes many useful features to build a robust predictive machine learning model. The metadata can be found here.

Data at a glance :

- The data has been recorded by SPD and traffic records.
- The data contains all types of collisions that occurred between 2004-2020 in the city of Seattle.
- It has a total of around 195000 records and 38 columns
- The data table consists of various columns such as location, date and time, severity, collision type, weather etc.

2.2 Data cleaning

After a preliminary analysis, I removed a few unnecessary columns from initial 38 columns.

Here is the list the removed columns:

'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'INCDATE', 'SDOT_COLDESC', 'SDOTCOLNUM', 'ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY'.

Some of the columns in the above list were unique identifiers for each incident, for example 'OBJECTID', 'INCKEY', 'COLDETKEY', which may not be useful in building a predictive model. Columns that contained a description text were also discarded, because the corresponding 'code' column for each such column was also present in the table. For example, 'ST_COLDESC' column containing collision code description provided by the state was removed and 'ST_COLCODE' column that contained the corresponding code was kept for analysis.

Columns with large data values missing were also removed as they would not have provided useful insights about the severity of collision. Finally, two redundant columns 'SEVERITYCODE.1' and 'INCDATE' were also dropped from the dataset and remaining 20 feature columns and 1 target column ('SEVERITYCODE ') were selected for further analysis.

After initial data cleaning, I also checked for missing values in columns and found out that those values can be filled with some other categorical value that is already present in the given column. Therefore, I replaced missing values in 'INATTENTIONIND', 'PEDROWNOTGRNT', and 'SPEEDING' columns with corresponding negative entry as these are yes/no categorical columns. And, for columns 'ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'WEATHER', 'ROADCOND', and 'LIGHTCOND', the categorical values 'unknown' and 'other' were already present. Therefore, missing values in these columns were replaced by either 'unknown' or 'other'.

I have also discarded location column to make the project computationally inexpensive. I have included the location analysis in future considerations of this project.

Lastly, the data was further prepared for exploratory analysis. The following section shows the impact of various features on the collision severity. I have also analyzed the relationship between the number of collisions and the date and time data to gain useful insights.

3. Data exploration and feature engineering:

The target variable for this study is the severity of the collision. There are two target categories: severity codes 1 and 2. Severity 1 includes collisions that resulted in only property damages and severity 2 contains collisions that resulted in property damages as well as injuries to the people involved.

3.1 Categorical features:

3.1.1 Time and date:

It is important to understand the impact of time of the day, month, and year on the number of collisions and severity. This analysis might present some insights into how these factors affect the severity of the collisions.

Figure 1 shows the time of the day impact. It can be observed that the number of collisions increase in the afternoon peaking at around 5 pm and the severity follows. Figure 2 shows a normalized distribution of severity 1 and 2 and it clear that the collisions between 12 to 5 pm are more severe.

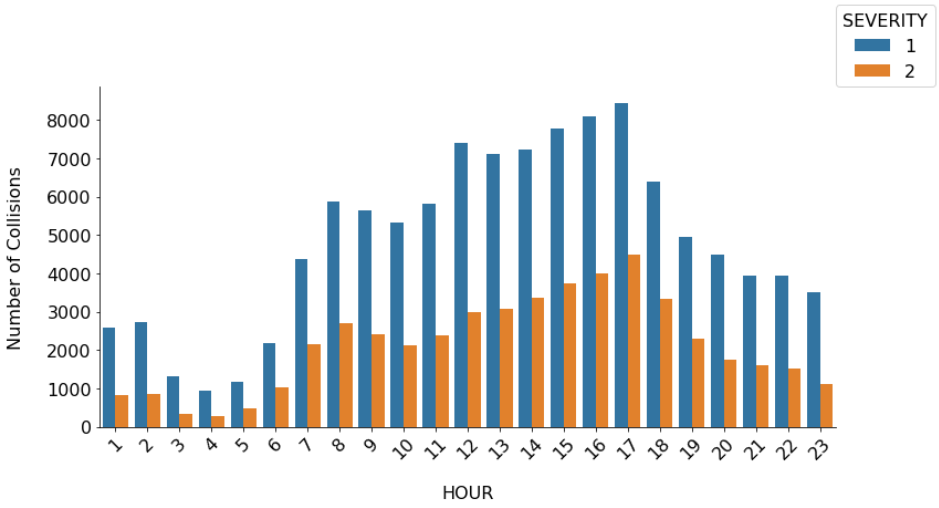

Figure 1. Hour versus number of collisions categorized by severity

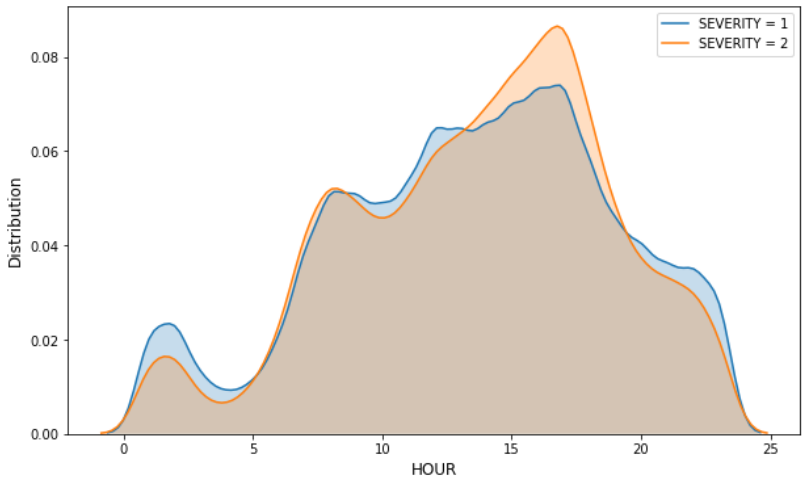

Figure 2. Normalized distribution of hour categorized by severity

On the other hand, it is evident that the number of collisions is almost uniformly distributed when plotted against month (figure 3).

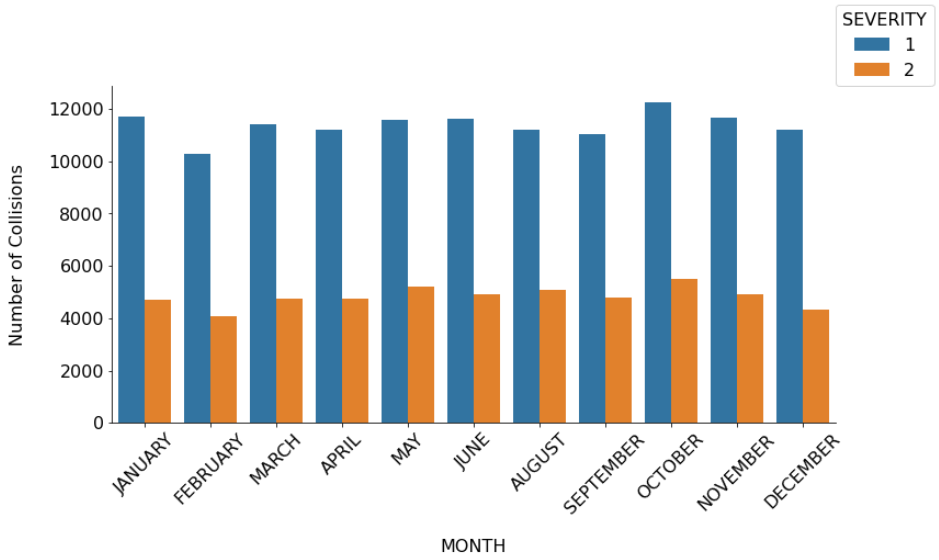

Figure 3. Month versus number of collisions categorized by severity

I have also plotted number of collisions categorized by the severity against the year they occurred (figure 4). Over the years, the number of collisions seem to have decreased except a slight uptick in the years 2014 and 2015. This decrease might be attributed to the technological advancements and public awareness.

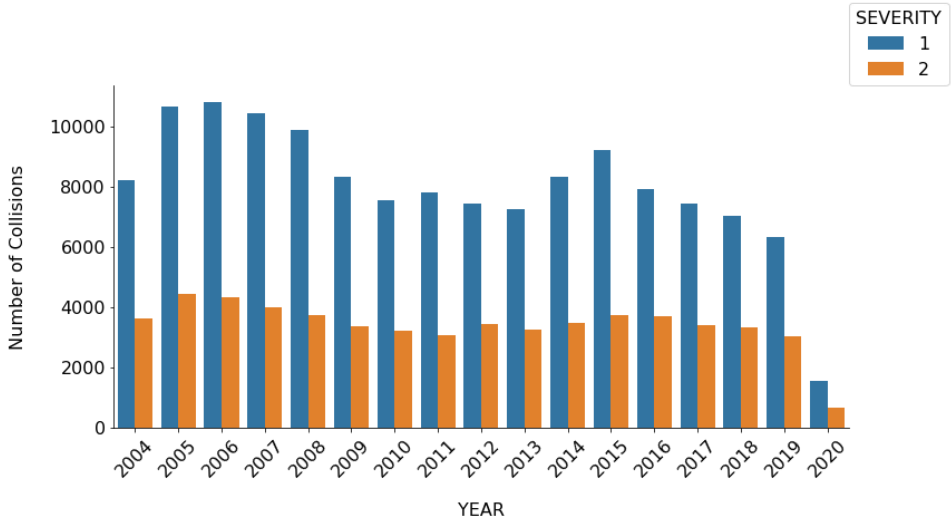

Figure 4. Number of collisions over the years categorized by severity

From this date time analysis, it can be concluded that more sever collisions occur in the afternoon and month has no impact on collisions. As a result of this analysis, I have removed date and time columns for the predictive model building because the month of the year does not impact number of collisions and severity, and around 32000 datapoints were missing from the time column. However, this analysis produced some useful insights that may help reduce the severity of the collisions.

3.1.2 Various driving conditions

In this section, I have explored how various driving conditions, such as light availability, road conditions, and weather, affect collision severity.

Figure 5 shows the number of collisions categorized by severity verses various weather conditions plot. Intuitively, it is expected that the extreme conditions such as raining, and snowing, would result in more collisions. However, from the graph, it is evident that highest number of collisions occurred in clear weather.

Likewise, when plotted against the number of collisions and severity, features light condition (Figure 6) and road condition (figure 7) produced similar results. It can be observed that the maximum number of collisions occurred in clear daylight and with dry road conditions.

The reason behind this behavior is that usually people are already aware of these extreme driving conditions. As a result, they drive their vehicles carefully or avoid driving completely resulting in a smaller number of collisions. However, we can not undermine the impact worse driving conditions.

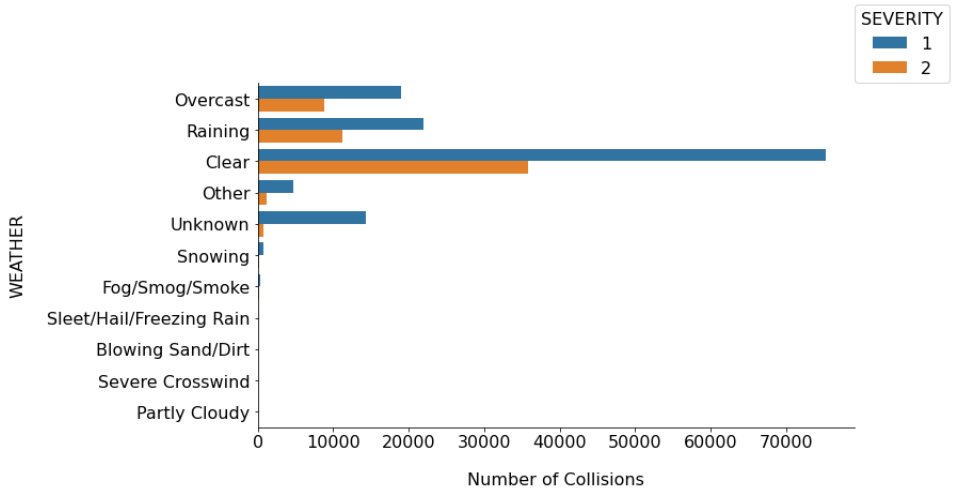

Figure 5. Number of collisions categorized by severity based on various weather conditions

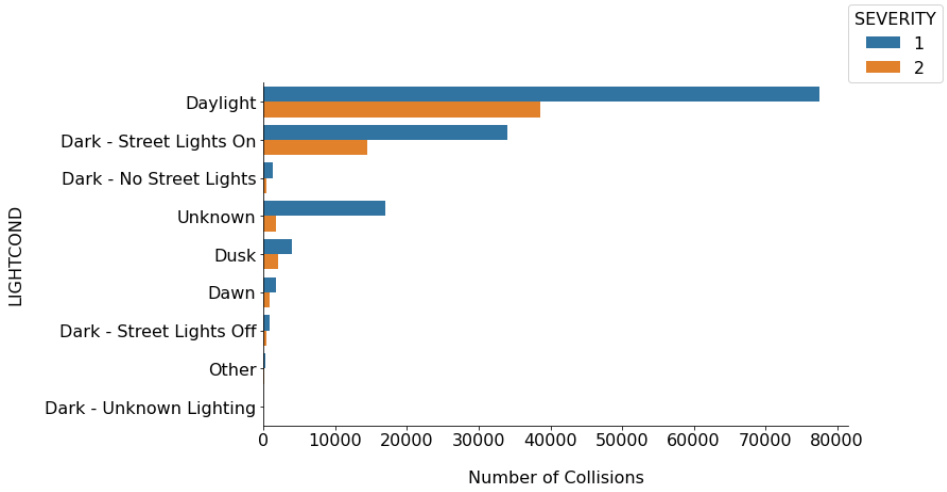

Figure 6. Number of collisions categorized by severity based on various light conditions

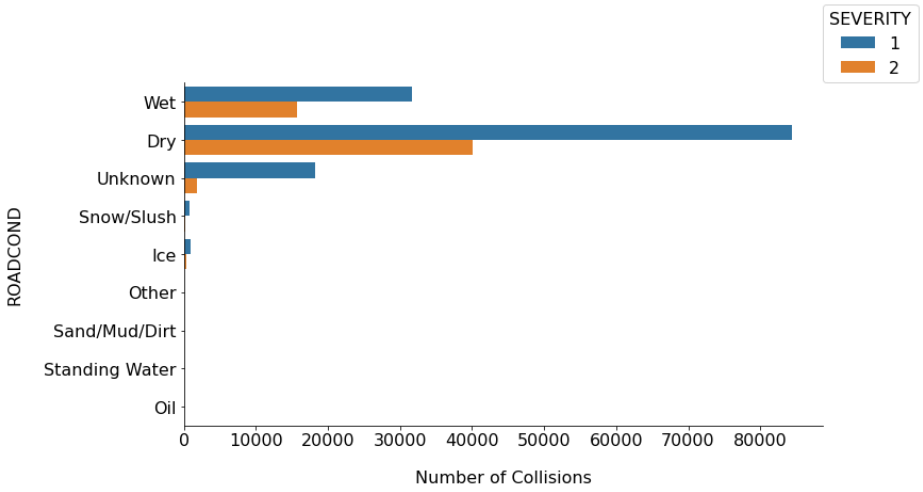

Figure 7. Number of collisions categorized by severity based on various road conditions

3.1.3 Collison type and junction type:

Figure 8 shows various collision types categorized by severity.

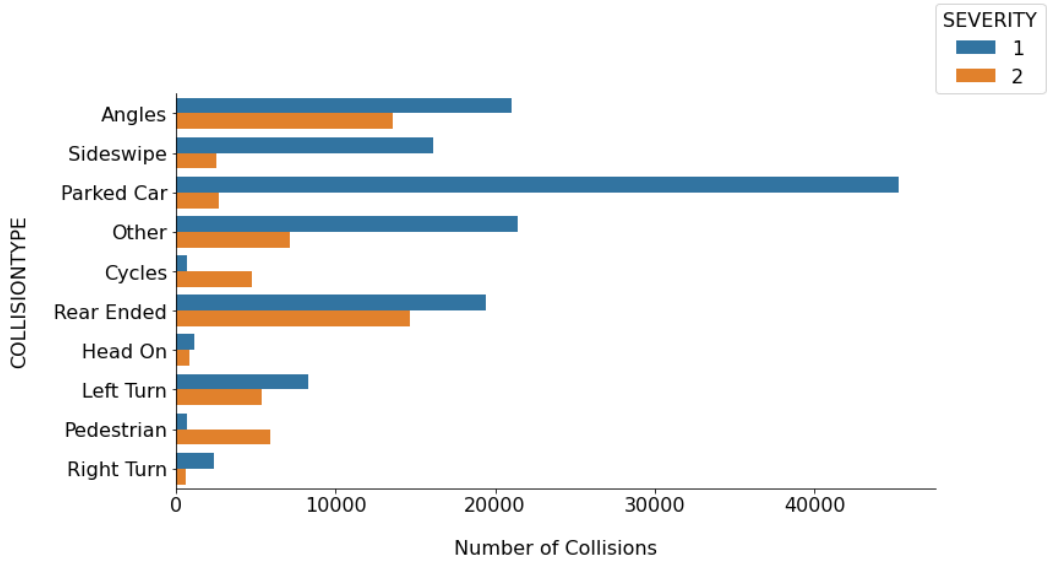

Figure 8. Various collision types categorized by severity

Maximum severity type 1 collisions occur when at least one of the vehicles involved is parked. And this type of collision results in minimal injuries. Whenever bicycles or pedestrians are involved, collision severity 2 occurs more frequently. The ratio of severity type 2 to type 1 is remarkably high when the collision types are ‘angled’ or ‘rear ended’.

From a plot of junction type versus collision (figure 9), it can be observed that the maximum number of collisions occur at the middle of a block. However, the collisions occurring at the mid-block produce fewer severity type 2 collisions when compared to the collisions occurring at the intersections.

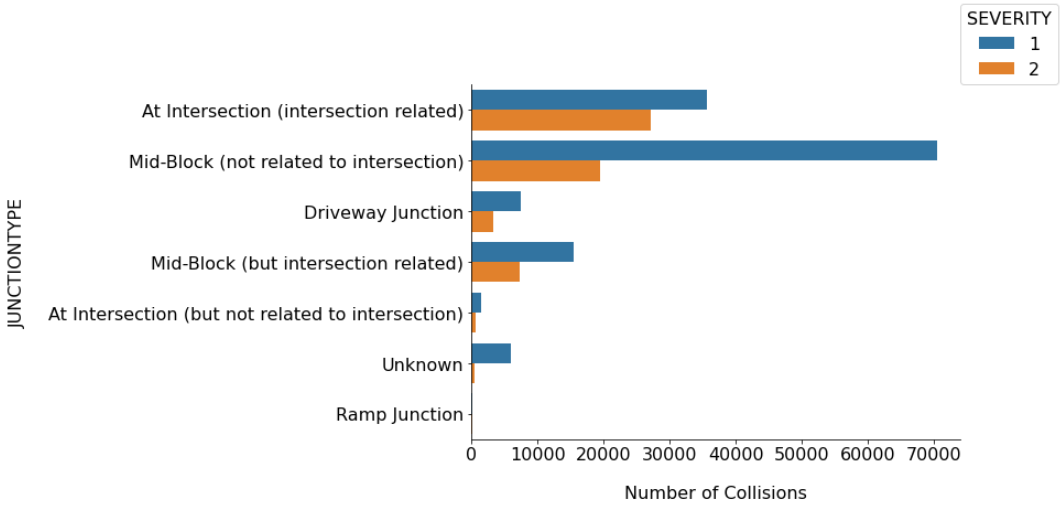

Figure 9. Number of collisions categorized by severity based on various collision types

3.2 Numeric features

Figure 10 shows boxplots of four features: number of persons, number of pedestrians, number of bicycles and number of vehicles involved in collisions. There are no extreme outliers found in any of the four columns. Therefore, I have included these features in the predictive model building process.

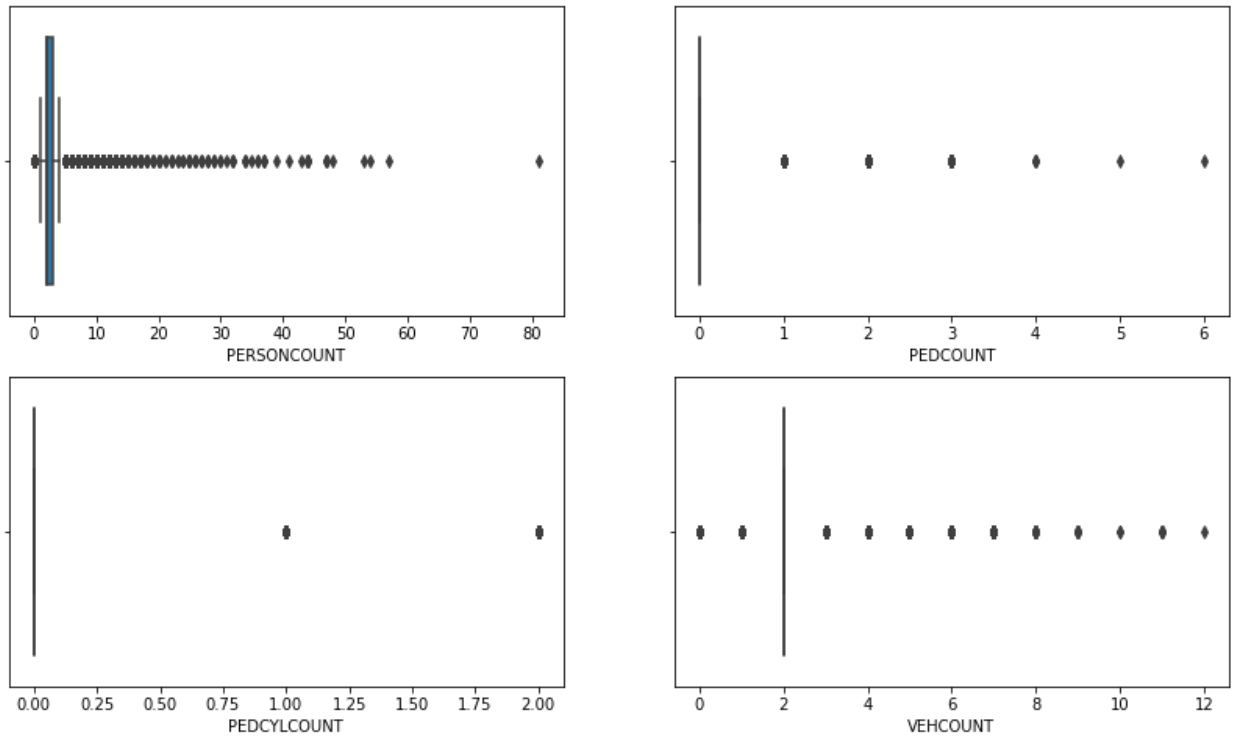

Figure 10. Boxplots of various numerical feature to identify extreme outliers

4. Model building

Our dataset contains several categorical and numerical features. And our target is also a categorical feature. It has two categories: severity 1 and severity 2. This makes it a classification problem, where based on various features, we will try to predict the severity type of a collision. Severity 1 represents only property damages and severity 2 represents property damages as well as injuries to the people involved. I selected following four candidates for the classification model for this project. I used accuracy, precision, recall and f1 score as the evaluation parameters in this analysis.

- Logistic regression
- Random forest
- Naïve Bayes
- Gradient boost

5. Results and Discussion

Table 1 shows the performance of all four models. I have also plotted confusion matrices for all four models (figure 11). It is evident that models performed poorly in classifying severity 2, with all models displaying a considerably low recall and f1 score. Best scores in each category are highlighted.

Table 1. Performance parameters of four classification models

	Logistic regression	Random forest	Naïve Bayes	Gradient boost
Accuracy	**0.76**	0.75	0.73	**0.76**
Precision Severity 1	0.77	0.78	**0.80**	0.77
Recall Severity 1	**0.95**	0.91	0.83	**0.95**
F1 score Severity 1	**0.85**	0.84	0.81	**0.85**
Precision Severity 2	0.72	0.63	0.55	**0.73**
Recall Severity 2	0.32	0.38	**0.49**	0.32
F1 score Severity 2	0.44	0.47	**0.52**	0.44

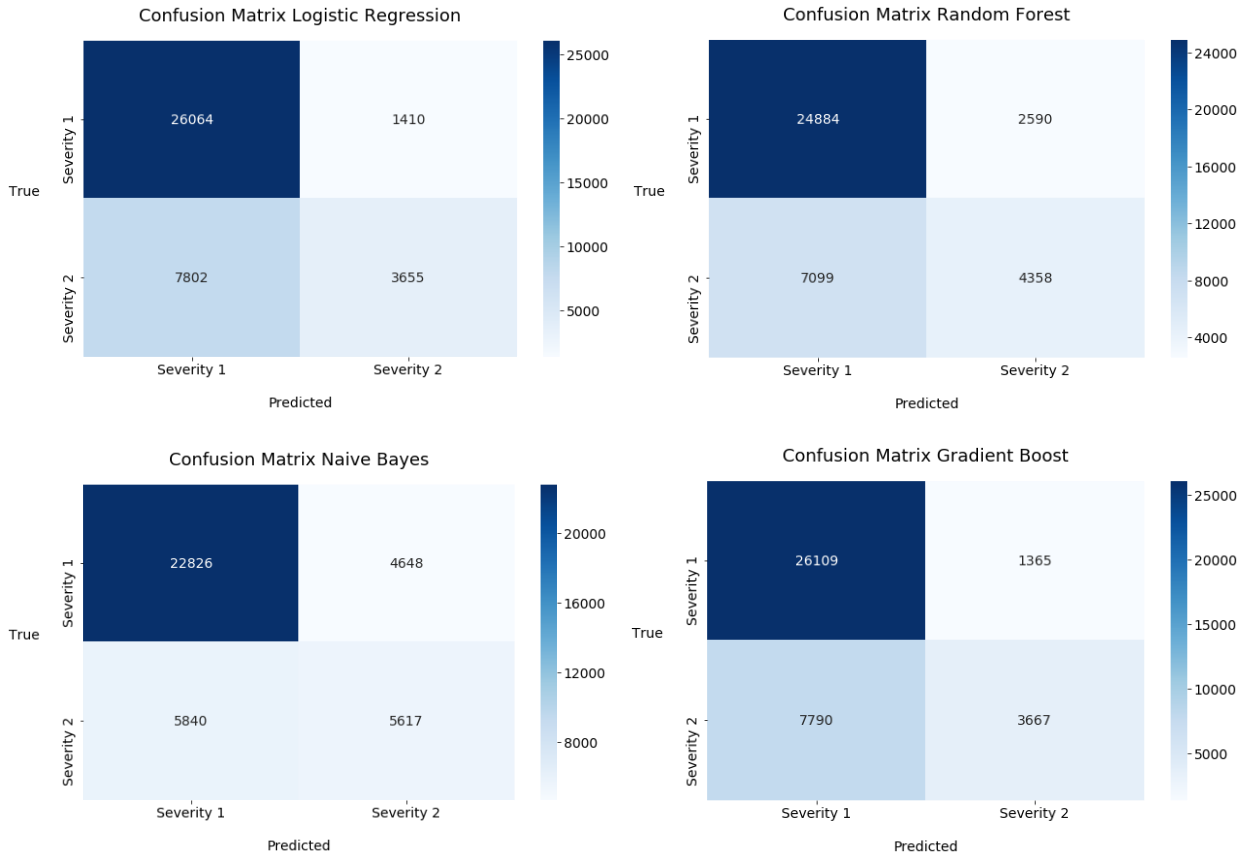

Figure 11. Confusion metrices for all four models

Upon further investigation, I observed that the number of collision entries in the training data for severity 1 is almost 2.34 times more than that of severity 2 entries. This unbalance could be the reason why all four models were biased towards severity 1 classification. To resolve this issue, I chose 'under sampling' technique. In this technique, number of entries from the dominant class are dropped until all classes have equal samples. Table 2 lists all model performance parameters after under-sampling. I have also plotted corresponding confusion metrices in figure 12.

Table 2. Performance parameters of four classification models after under-sampling

	Logistic regression	Random forest	Naïve Bayes	Gradient boost
Accuracy	**0.72**	0.71	0.60	**0.72**
Precision Severity 1	0.76	0.74	0.56	**0.77**
Recall Severity 1	0.65	0.65	**0.98**	0.63
F1 score Severity 1	0.70	0.69	**0.71**	0.69
Precision Severity 2	0.69	0.68	**0.92**	0.68
Recall Severity 2	0.79	0.76	0.21	**0.81**
F1 score Severity 2	0.73	0.72	0.34	**0.74**

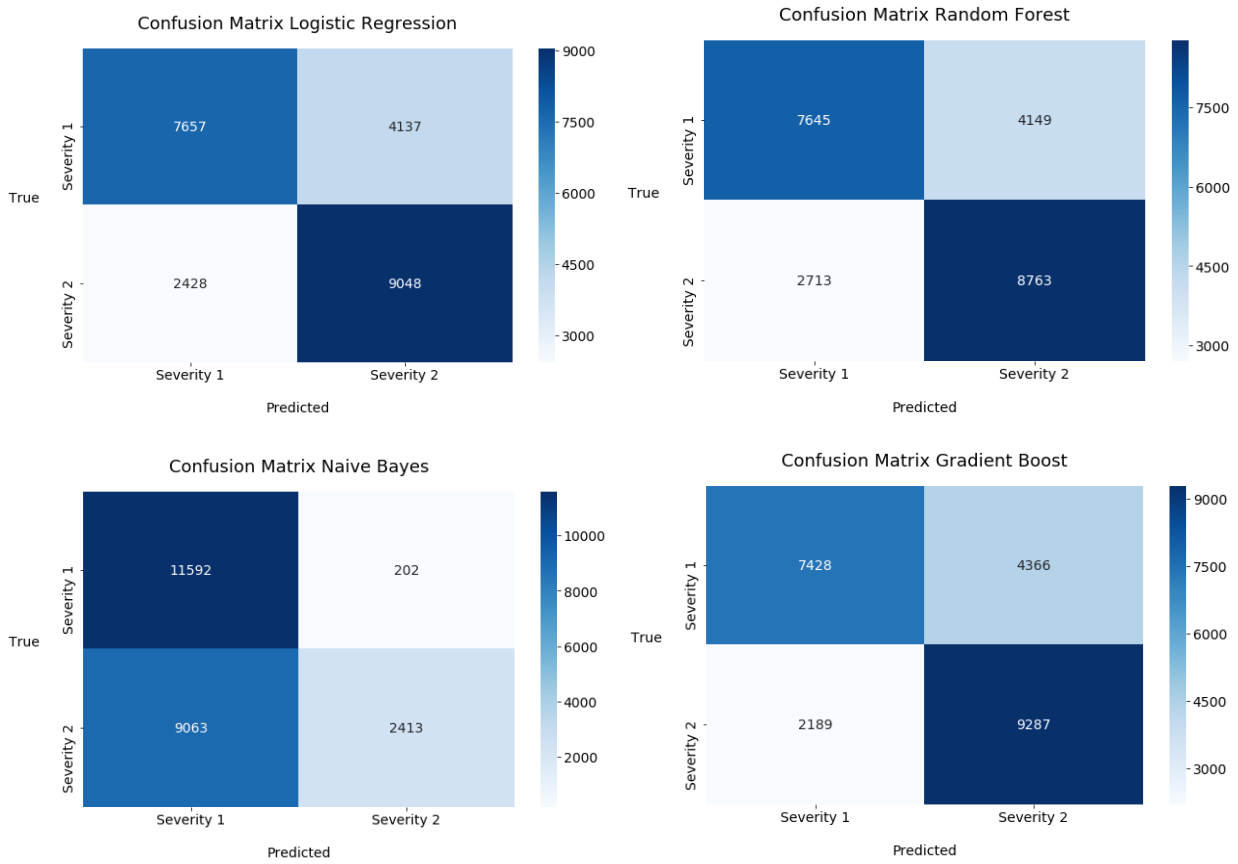

Figure 12. Confusion metrices for all four models after under-sampling

After applying under-sampling, the models performed better in making unbiased classification except the Naïve Bayes model. Although, there was no significant increase in overall accuracy scores for the three models: Logistic Regression, Random Forest and Gradient Boost, their performance in classifying severity 2 increased significantly.

It can be observed that, the overall performance of the gradient boost algorithm was better than other three models and it managed to predict slightly more accurate and unbiased results. In this type of classification problem, it is preferable that the model produces more false positive results than true negative results in classifying severity 2. In this regard, Gradient boost model performed better producing 4366 false positive predictions and only 2189 true negative results when compared to the other models.

6. Conclusions

6.1 Insights

Following insights were obtained from the data exploratory analysis.

1. Most collisions occur in the afternoon with more severe collisions at around 5 pm. Office rush might be contributing to this increase. Conversely, number of collisions are not affected by what month of the year they occur.
2. The number of collisions has been decreasing steadily year over year despite of an exponential growth in total number of vehicles on the roads. This means that public awareness and technological advancements has a positive impact in reducing the number of collisions and severity.
3. Although, intuitively worse weather, road and daylight conditions should produce more collisions, that is not the case. Maximum collisions occur on clear days and on dry roads.
4. Collisions involving a parked vehicle produce more severity 1 collisions than any other type of collision. However, ‘angled’ and ‘rear-ended’ type collisions result in more severity 2 collisions than any other type.
5. Collisions involving pedestrians and bicycles almost always result in injuries to the people involved .

6.2 Predictive model

In this project, I built four predictive models to classify collisions severity based on several features, such as weather condition, road condition, number of persons and vehicles etc. All four models produced biased results towards severity 1 when I used all the training dataset. The reason behind this bias was unbalanced data. To resolve this issue, the under-sampling technique was used and the three models out of four produced unbiased results. Based on the overall performance, Gradient Boost predictive model proved to be the best candidate producing 72% accuracy.

7. Future considerations

1. Location data can be used to gain more insights from the Collison data, so that the plan to reduce collision severity can be customized based on specific location
2. Date and time data can be included into predictive model building.
3. To tackle unbalanced data issue, techniques other than under-sampling, such as weighed sampling and over-sampling, should be considered for further analysis.
4. Hyper-parameter tuning can be used to further improve the performance of predictive models used in this project.