Project Title: Predicting the Severity of Traffic Collisions in Los Angeles

Applied Data Science Capstone

Jitong Yang

September 5, 2020

1. Introduction

Motor vehicles are essential for most people to travel or work because of their convenience. Since the global economies, social welfare, and technology have been well developing in recent decades, there are increasing number of people start to buy a car for their daily uses. However, the risk that a traffic accident occurs also becomes increasingly higher along with more vehicles appear on the road. For example, if two or more cars collide each other, the road will become extremely crowded, especially when the weather condition is terrible, or the visibility is low. The collisions would result in unnecessary injuries to people involved in the accident and cause unexpected delays for other drivers to perform their travel plans. This is especially the case in metropolises.

Therefore, the government must consider about how to predict the severity of a potential traffic accident effectively based on a wide variety of factors. The primary goal is to place warnings that could help people to drive smoothly before the occurrence of different unusual circumstances. For instance, a torrential rain could result in a very slippery road surface, and thus drivers can hardly control the speed of their vehicles. Also, the probability that collisions occur in the evening is likely to be higher than that in the daytime as its visibility is relatively low. As a result, we can see that the degree of traffic accident severity could be different under different circumstances, and it is important to figure out which factor can best predict that severity.

In this project, I use five classification machine learning algorithms (i.e. K-Nearest Neighbors, Decision Tree, Logistic Regression, Support Vector Machine, Random Forest) to predict the severity of a traffic accident based on many potential attributes, such as weathers and road locations, in Los Angeles, which is one of the largest cities in the United States. Put another way, I investigate which features have the most significant impact on the severity using

exploratory data analysis and put them into model training. Moreover, I evaluate the quality of each machine learning model by calculating relevant accuracy scores using test data and find the best model. It can help the government to implement appropriate traffic protocols and significantly reduce the likelihood of traffic collisions under different scenarios.

In "Data" section, I introduce the dataset for my study and present which variables I would choose from it. Next, I present exploratory data analysis and use it to determine the choice of regression or classification and feature selection. The subsequent section is called "Results", which displays the empirical results of the analysis and assessments of model evaluation. Afterwards, we discuss the implications from the empirical results in "Discussion" section, and how they can recommend the government to make appropriate decisions.

2. Data

The description and use of data source are shown in this part. The dataset is retrieved from Kaggle.com, which contain around 3.5 million records of traffic accidents in the United States from 2016 to mid-2020. It contains a lot of available information for my studies, such as degree of severity, latitudes & longitudes of accident, states, cities, humidity, visibility, wind speed, weather conditions, presence of amenity or pedestrian crossing, the period of day (i.e. daytime or night), etc. I use Pandas functions in Python to slice the dataset for the traffic accidents in Los Angeles only since it is of my primary interest.

Moreover, I choose attribute "Severity" as the target (i.e. dependent variable), which describes the degree of traffic accident severity. It contains integers from 1 (least severe accident) to 4 (most severe accident). Also, I only select necessary attributes as features (i.e. explanatory variables such as weather conditions, presence of amenity or pedestrian crossing, the period of day, and so on) and create dummies if the variable does not contain numerical values

such as integer, float, or Boolean. I also create a new categorical variable “Weekday” because severity may vary across different weekdays. Then, I use my target and features to conduct exploratory data analysis to further narrow my feature list and train machine learning models.

3. Methodology

In this section, I discuss the exploratory data analysis since it could preliminarily identify the outliers and useful features in the dataset. Then, I specify appropriate machine learning models of interest for the studies, which are suggested by the exploratory data analysis.

3.1 Exploratory Data Analysis

Conducting exploratory data analysis is a requirement for further empirical studies because it enables us to discover the responsiveness of severity to other variables prior to the machine learning process. Also, it could help me to decide whether I should use regression or classification techniques for estimating my model. Specifically, I calculate descriptive statistics for numerical variables, visualize the dataset, and analyze the correlations among variables.

3.1.1 Descriptive Statistics

Table 1: Summary Statistics of all Numerical Variables

	Severity	Temperature (F)	Humidity (%)	Pressure (in)	Visibility (mi)	Wind_Speed (mph)
mean	2.372	66.691	61.244	29.900	9.115	4.868
std	0.502	9.003	20.440	0.158	1.964	3.205
min	1	37.9	3	28.83	0	0
1st quartile	2	60.1	50	29.81	10	3.5
median	2	66	64	29.91	10	4.868
3rd quartile	3	72	77	30	10	5.8
max	4	106	100	30.5	10	36.8
mode	2	64	78	29.91	10	4.868
skewness	0.737	0.412	-0.630	-1.042	-2.418	0.908
kurtosis	-0.932	0.156	-0.198	4.391	5.120	2.949
count	79169	79169	79169	79169	79169	79169

Table 1 shows the summary statistics of all numerical variables. We can see that the standard deviation of "Humidity" is the highest among those of all variables, and thus it is the most volatile. In contrast, "Pressure" has the smallest standard deviation, so it has the lowest volatility. By comparing mean, median, and mode, we could observe that they are close to each other for most variables. One exception is that the mode of "Humidity" is much larger than the mean and median, which indicates the sign of outlier for this feature.

It is also evident that the outlier present by comparing different quartiles. For most variables, the values change gradually from minimum to the third quartile, but they change significantly from third quartile to maximum. Furthermore, the measures of both skewness and kurtosis are obvious, which indicate the probability distributions of these features are deviate from a normal distribution. As a result, there are indeed outliers for all variables in this dataset.

Table 2: Average Severity under Different Weather Conditions

Weather Condition	Severity	Weather Condition	Severity	Weather Condition	Severity
Blowing Dust	2	Haze	2.358	Light Thunderstorms and Rain	2
Clear	2.456	Heavy Rain	2.239	Mist	2.5
Cloudy	2.263	Heavy T-Storm	2	Mostly Cloudy	2.363
Drizzle	2	Light Drizzle	2.25	Mostly Cloudy / Windy	2
Fair	2.201	Light Rain	2.315	Overcast	2.502
Fair / Windy	2.273	Light Rain / Windy	2.5	Partly Cloudy	2.314
Fog	2.412	Light Rain with Thunder	3	Partly Cloudy / Windy	2
Patches of Fog	2.5	Scattered Clouds	2.5	Thunder	2
Rain	2.327	Shallow Fog	2.333	Thunderstorm	2.5
Rain / Windy	2	Smoke	2.53		

The average severity under various weather conditions are shown in *Table 2*. It is clear that the means of traffic collision severity vary significantly across weathers, ranging from 2 to

3. In other words, the degree of traffic accident severity changes when the weather condition changes. Consequently, weather condition could be an effective predictor of severity.

Table 3: Average Severity under Daytime and Night

Period	Severity
Day	2.355
Night	2.403

I put the average severity under daytime and night in *Table 3*. We can see the difference of the means under daytime and night is small. As a result, the severity does not significantly depend on whether the period of day is day or night, although the severity under night is slightly higher. Therefore, variable "Sunrise_Sunset" may not be a good predictor of severity.

Table 4: Average Severity under Different Weekdays

Weekday	Severity
Sun	2.459
Mon	2.352
Tue	2.332
Wed	2.348
Thu	2.363
Fri	2.351
Sat	2.465

Table 4 demonstrates the average traffic accident severity under different weekdays. The means under all business days are similar in magnitudes, and they are lower than those under weekend. Therefore, we could find that the traffic accidents in weekend are slightly more serious than those in business days, but this difference is not evident. As a result, we might not consider "Weekday" to be of interest in model training.

3.1.2 Correlation Analysis

I also compute the sample correlation coefficients between severity and other features using Pearson method. This step helps me to determine the responsiveness of severity to other

variables and to narrow my feature list. If this responsiveness is extremely close to zero, we need to drop the corresponding feature from the list because it does not help me to predict the target effectively. *Table 5* below illustrates the results of correlation analysis, and I only focus on the co-movements between severity and other variables.

Table 5: Correlations among Variables

	Severity	**Temperature (F)**	**Humidity (%)**	**Pressure (in)**	**Visibility (mi)**	**Wind_Speed (mph)**
Severity	1*	-0.032	0.08	0.147	0.005	0.092
Temperature (F)	-0.032*	1	-0.477	-0.245	0.177	0.11
Humidity (%)	0.08*	-0.477	1	-0.028	-0.361	-0.036
Pressure (in)	0.147*	-0.245	-0.028	1	0.007	0.106
Visibility (mi)	0.005*	0.177	-0.361	0.007	1	0.085
Wind_Speed (mph)	0.092*	0.11	-0.036	0.106	0.085	1

Overall, the correlations between severity and all other explanatory variables are somewhat weak, especially that between severity and visibility is only 0.005. Therefore, visibility may not effectively explain the behavior of severity, and we could drop it from the feature list. In addition, we could also find that traffic severity is positively correlated with all features except temperature.

3.1.3 Data Visualization

Data visualization can better explain the characteristics of the dataset than tables because it straightforwardly presents the relationships between variables using appropriate graphs. In this part, I visualize average severity under different categorical variables using box plots and correlation analysis through scatter plots.

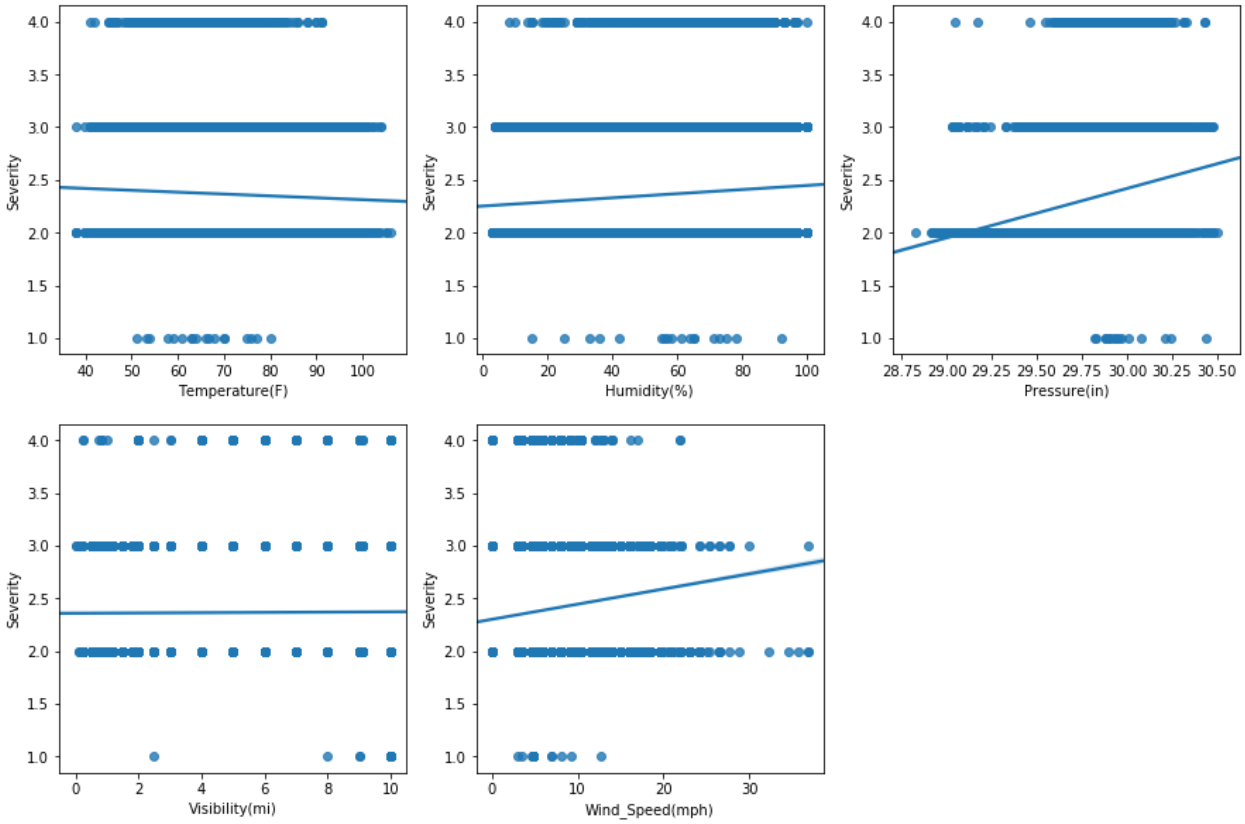

Figure 1: Scatter Plots of Correlations between Severity and Other Variables

Figure 1 depicts the correlations between severity and other features via scatter plots and linear regression lines. The charts are corresponding to the correlation analysis results I present in the last part. I observe that severity varies positively with pressure, humidity, and wind speed and varies negatively with temperature. In contrast, severity does not have an obvious co-movement with visibility.

Additionally, the observations are not appropriately fitted by the regression line, the reason is that severity is not continuous and have limited range. Consequently, regression techniques could not be useful for predicting severity, and we should use classification algorithms to account for the prediction. Next, we look at box plots, which display the average of severity under three categorical variables: weather condition, period of day, and weekdays.

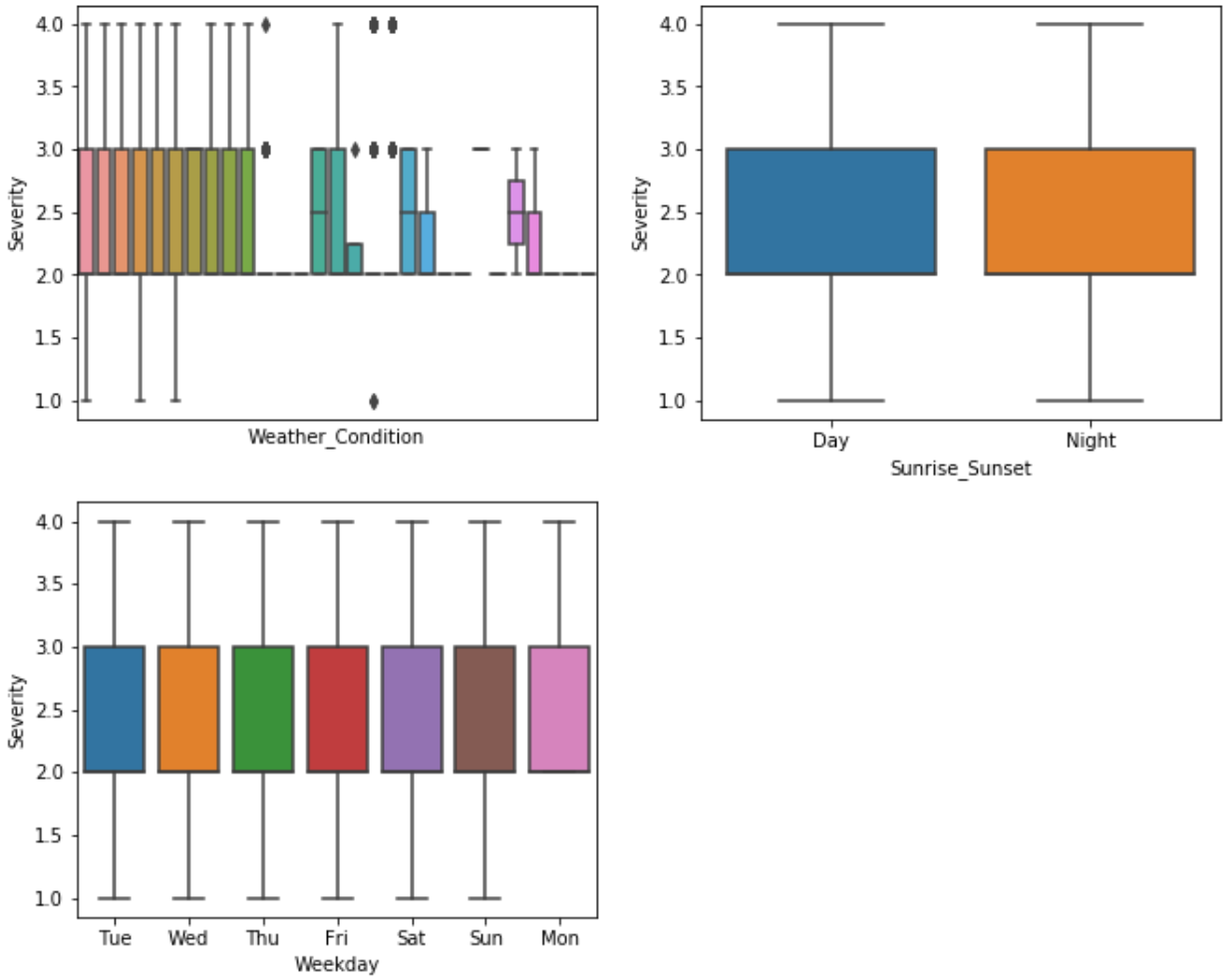

Figure 2: Relationship between Severity and Three Categorical Variables

Figure 2 visualizes the results from *Table 2* to *Table 4*, which depict how severity looks like under different categories. Since we could find the evidence that average severity differs across weathers, we could keep "Weather_Condition" in the feature list. In contrast, the period of day and weekdays are not useful to address the changes in severity because we find the box plots overlap significantly. As a result, we could remove "Sunrise_Sunset" and "Weekday" variable from selected features.

To sum up, I apply classification techniques for my studies, and the features I choose are temperature, humidity, pressure, wind speed, weather condition, and most dummy variables that

represent road locations. The exceptions are give_way, roundabout, and turning_loop in that they have same values for all observations and thus do not correlate with all other variables.

3.2 Machine Learning Algorithms

I apply five classification algorithms for my analysis, including K-Nearest Neighborhoods (KNN), Decision Tree, Logistic Regression, Support Vector Machine (SVM), and Random Forest. I split the dataset into training set (75% observations) and test set (25%). I use training set to train the model under each algorithm, and the test set is used to predict traffic accident severity and assess the quality and accuracy of each model. Additionally, for KNN model, I also find the optimal number of k using appropriate Python loops and functions.

For model evaluation, the goal is to choose the model with highest accuracy score because it can best predict severity. I compute Jaccard index and F1-score using predicted values and actual values for each model after the prediction of severity.

4. Results

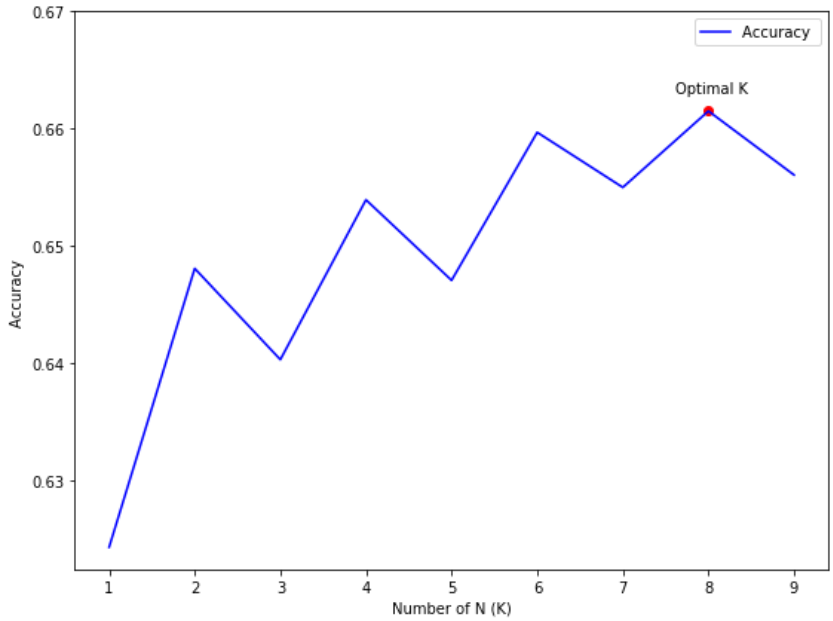

Figure 3: Optimal Number of Nearest Neighbors k

Before the training process of KNN, I find the optimal number of nearest neighbors k that could achieve the most accurate prediction. I calculate the accuracy score when k ranges from 1 to 9 and visualization of this result is shown in *Figure 3* above. When k equals to 8, the accuracy score is the highest at 0.6615.

After the training process of all models, I use test data to compute the predicted values of traffic accident severity. Then, I calculate Jaccard index and F1-score for each algorithm using fitted values and actual values. I put the results in *Table 6* below:

Table 6: Model Evaluation Scores

Algorithm	Jaccard index	F1-score
KNN	0.6615	0.6338
Decision Tree	0.6457	0.5227
Logistic Regression	0.658	0.6316
SVM	0.6621*	0.6348
Random Forest	0.6605	0.6525*

These five algorithms achieve a similar performance in predicting the severity because their accuracy index values are close to each other. We can see that SVM has the highest value of Jaccard index, and Random Forest has the highest value of F1-score. Therefore, these two models provide a relatively accurate prediction result among all algorithms.

In contrast, Decision Tree gives the least accurate results under both criteria. Especially, its F1-score is only 0.5227, which is significantly lower than all other values. As a result, Decision Tree is not useful to predict the severity of traffic collisions.

Additionally, I plot the distributions of predicted values and actual values for each model as it provides a more straightforward insight. The goal is to see whether predicted severity is close to actual one. The distribution graph is illustrated in *Figure 4* below.

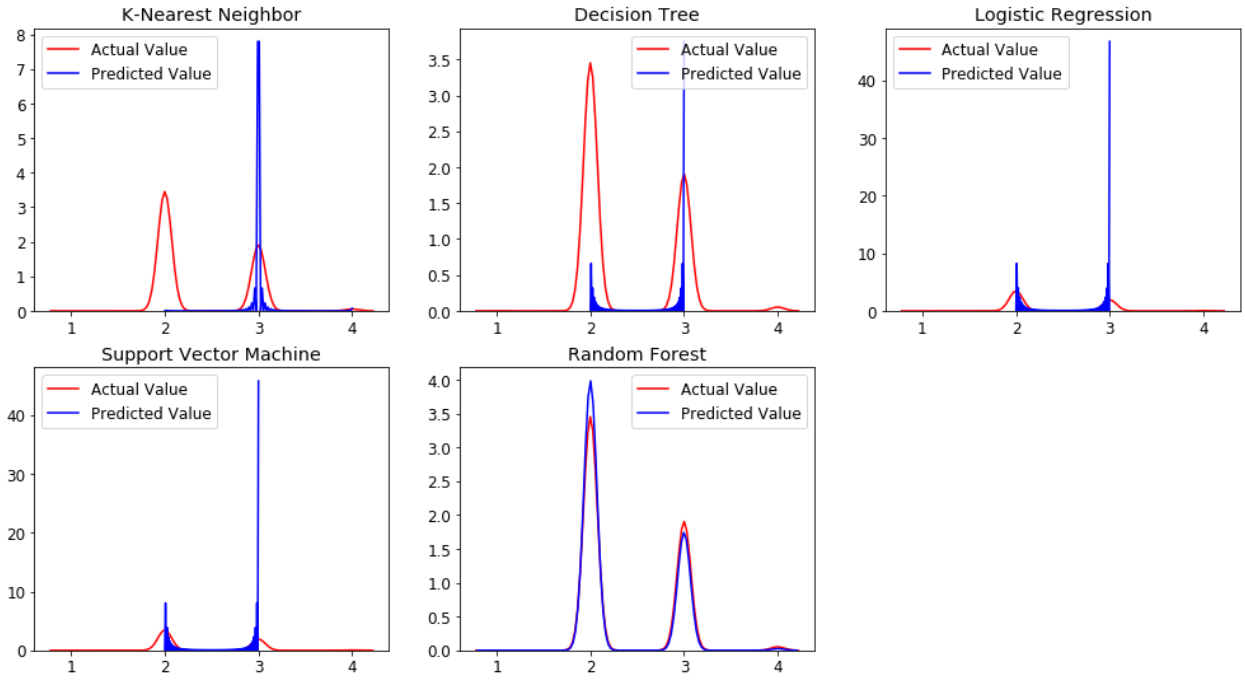

Figure 4: Distribution Plots of Predicted Value and Actual Value (Test Set) under Each Algorithm

It is obvious that KNN and Decision Tree do not effectively predict the severity because the distributions of predicted values are significantly different than those of actual values. Logistic Regression and Support Vector Machine perform better, but the distributions of predicted values are extremely leptokurtic, which indicate sign of large outliers in prediction. I also find that predicted values are quite similar to the true values if I use Random Forest. Therefore, the distribution plot suggests that Random Forest provides the most accurate severity prediction among all models.

5. Discussion

Therefore, weekdays and period of the day are not useful to forecast the severity of traffic accidents in Los Angeles, and we should not consider them while making predictions. In contrast, we could take weather conditions and road locations into account. In this scenario, severity is a categorical variable, and using regression models does not fit observations effectively. Instead, in this scenario, we use classification algorithms.

Different algorithms have different performance, so we should train various models and find the best one. Consequently, I train five classification models, calculate the accuracy score for each of them, and plot the distributions of fitted values and true values under each algorithm. Based on the empirical results, Random Forest is considered as the best model for predicting the traffic accident severity in Los Angeles.

6. Conclusion

The primary goal of this project is to find the appropriate determinants of traffic collisions severity in Los Angeles and use them to predict severity under different machine learning models. I first conduct an exploratory data analysis to examine whether the selected features could be a good predictor of severity using summary statistics, correlation analysis, and some visualization tools. The results indicate that weather conditions can predict severity effectively, while weekdays and period of the day do not.

I then split the dataset into training data and test data. I use training data to fit five classification models, including KNN, Decision Tree, Logistic Regression, SVM, and Random Forest. Then, I calculate the accuracy scores (i.e. Jaccard index, F1-score) and visualize the distribution of predicted severity and actual severity for each algorithm. I find that Random Forest best predicts the severity of traffic collisions in Los Angeles.

In my future studies, I would still discover about which factors could effectively determine the severity of traffic accidents in other cities, but the target could be economic losses arise from traffic accidents. In this case, I can use regression techniques for prediction since the target becomes a continuous variable, which can better show the estimated impact of explanatory variables on the dependent variable than classification. Also, I might investigate how COVID-19 pandemic impacts the traffics once the number of observations under COVID-19 periods becomes sufficient in the future. I would also further refine my machine learning models and achieve an accuracy score as higher as I can.