IEOR E4650 Business Analytics

Group Project Report

Team Alabama

Lanqing Li	ll3444
Weishan Lu	wl2778
Zimin Lu	zl2956
Yiran Shen	ys3381
Lu Liu	ll3454

Introduction

Cars are the most common and popular transportation in recent decades. Car accidents are one of the accident types that cause many injuries and deaths every year. Besides life safety, car accidents will also cause severe traffic jams which negatively affect people's daily life. And the impacts on traffic depend on the severity of car accidents.

In this project, we regard the traffic department as our client and assume that we can quickly obtain data about road conditions and car accident information when an accident occurs. According to our effective prediction models based on historical data, when a car accident happens, the traffic department can predict the severity and impacts on traffic, then react immediately to prepare for the traffic control.

We used a countrywide car accident dataset collected from 2016 to 2021. Starting from Exploratory Data Analysis, we cleaned data and visualized the relationship between variables. Based on that, we built prediction models including logistic regression, LDA, Decision Tree and Random Forest. In the end, we discussed some extensions of our work.

1 Data Preprocessing

The original data set contains 2,845,341 rows and 47 columns of car accident records that happened in 49 states in the US, which was collected from February 2016 to December 2021.

Firstly, we removed some helpless columns, dropped all missing values, and added some new columns based on the original data. For example, we added the "Duration(min)" column, which is obtained from the difference between "start_time" and "end_time". After checking the correlation of all features *[Figure 1.1]*, we found that "Temperature" and "Wind_Chill" have a strong relationship, so we dropped the "Wind_Chill" column. Also, the definition of "Traffic_Calming" contains "Bump", so we removed "Bump". Secondly, we observed the outliers of weather attributes and made reasonable removals, based on common sense and outlier selection criteria. For example, the temperature is restricted to below 130(F), the humidity is restricted to between 0% and 100%, the pressure is restricted to the mean plus or minus two standard deviations, the visibility is restricted to within 0(mi) to the mean plus three standard deviations, the wind speed is restricted to between 0(mph) and 60(mph), and the precipitation is restricted to within 2.5(in) per hour. Finally, the dataset obtained after data preprocessing has 2,081,305 rows and 39 columns, includes:

Table I: Attributes in the Dataset after Preprocessing

Total Attributes (39)	**Attribute Names**
Traffic Attributes (10)	id, severity, start_time, end_time, start_Lat, start_Lng, end_Lat, end_Lng, distance(mi), duration(min)
Address Attributes (6)	street, side (left/right), city,state, zip-code, country
Weather Attributes (8)	temperature(F), humidity(%), pressure(in), visibility(mi), wind_direction, wind_speed(mph), precipitation(in), condition (e.g., rain, snow, etc.)
POI Attributes (11)	Amenity, Crossing, Give-Way, Junction, No-exit, Railway, Roundabout, Station, Stop, Traffic Calming, Traffic Signal
Time Attributes (4)	Sunrise/Sunset, Hour, Day, Month, Year

2 Data Visualization

We first took a look at the "Severity" and found that 93.43% of the car accidents in our data have severity level at 2 *[Figure 2.1]*. The data set is extremely imbalanced, so we will use some techniques to resample it later in the modeling part. Then we did exploratory data analysis by type of attributes.

Address attributes:

We made a map showing the number of car accidents by each state *[Figure 2.2]*. We found that this data set doesn't have statistics in states Alaska and Hawaii. California is the state with the most car accidents and Florida has the second-most in our data set. By city, almost 100,000 car accidents happened in Miami, which is about double of the number of accidents that happened in the second and third most cities Los Angeles and Orlando *[Figure 2.3]*. And we also found that the number of accidents that happened on the right side of the street is about four times that of accidents that happened on the left side *[Figure 2.4]*.

Time attributes:

We used the start time of accidents to generate four columns including "Hour", "Day", "Month", and "Year". We found that car accidents occurred more on weekdays than weekends *[Figure 2.5]* and most of the accidents happened during 6-8am and 13-18pm, which are the time slots people usually commute to work *[Figure 2.6]*. For accidents with severity level at 4, the percentage of accidents occurred at midnight is much higher than that of other severity levels.

Weather attributes:

We used violin plots to visualize the distribution of each weather attribute at each severity level and found that accidents with severity level at 2, 3, and 4 usually occurred when humidity is higher, compared to when accidents with severity level 1 occurred *[Figure 2.7]*. And the shape of the violin plot for severity level 4 is widest around 80-100% humidity, indicating that the higher the humidity is, the more severe accidents might occur.

POI attributes:

According to pie plots of each POI attribute, the presences of crossing, junction, and traffic signals are comparatively high when car accidents occurred. The plot shows that around 1/3 of accidents with severity level 1 occurred when there is a crossing nearby, and more than 40% of accidents with severity level 1 occurred when there is a traffic signal nearby in our data set *[Figure 2.8]*. And 11% of severe accidents (level 4) occurred near a junction or a traffic signal *[Figure 2.9]*.

3 Classification Modeling

With road situations and accident information as input features, we wish to train classification models to predict the severity of a traffic accident, helping our client to prepare for it. Our dataset shows the majority of severity 2 (93%), so we tried to deal with the imbalance problem in modeling. We picked several classical classification models and divided them into two parts — parametric and non-parametric modeling. For all models, we used numerical weather features, POIs and State as dependent variables, and set 70% of observations to be the training set and 30% of observations to be the testing set. Considering that predicting the accidents with severity 3 or 4 correctly is more important, we used weighted recall as the criterion for model selection.

3.1 Parametric modeling: Logistic Regression and LDA

We choose linear models like Logistic Regression and LDA to classify severity. Since the dataset is very imbalanced, in addition to the original training set, we also tried SMOTE to oversample the minority classes (1,3,4), RandomUnderSampler to undersample the majority class (2) and finally combine the two sampling methods. In the test set, we have 93.4% severity 2, 2.8% for severity 3, 2.8% for severity 4, only 1% for severity 1.

Using the original training set, we applied grid search for both models, shown in the classification report.*[Table 3.1][Table 3.2]* We can see that the models only show accuracy of 93.489% and 92.07%, with really poor recalls of minority classes. The multi-class logistic model performs just a little bit better than predicting only severity 2. So we can say that the two models cannot tell the difference between level 2 and other severity levels.

Next, we used SMOTE to oversample severity level 1, 3, 4 into 40000, 130000 and 130000, lowering the percentage of class 2 into 82%. Under this condition, SMOTE doesn't seem to improve accuracy, but a bit of improvement in recall of class 3 & 4. *[Table 3.3][Table 3.4]* The reason for this accuracy may be that SMOTE generates noise when generating the balanced data. By reducing the majority class into 500000 observations, RandomUnderSampler presents us with similar results as SMOTE. Combining undersampling and oversampling gives out even worse accuracy, but the best in recall of class 3 & 4!*[Table II]* If we choose the linear parametric models, maybe we have to trade off between accuracy and recall of serious accident types.

Looking at the confusion matrix *[Table 3.5]* derived from the logistic regression model using the *original training set*, we can see that class 1 usually can be classified as 2 and class 3 & 4 usually can be misclassified as each other (except for class 2). For further improvements, we try to combine class 1 & 2 and 3 & 4, changing into the binary class problem. With 94.4% of the less severe accidents, our models can only achieve 95% accuracy in the logistic model, with highest recall of 13% in LDA.*[Table 3.6][Table 3.7]* Still, these two models do not perform well telling serious accidents from less severe accidents. However, the weighted precision, recall and f1-score improve a little in binary classification. In future improvements, we can try training other models in binary classification along with balancing techniques to achieve better performances.

The performances of these parametric models show that our dataset is not suitable using linear, parametric modeling to classify severity. Possible reasons may be that the features in predicting severity have non-linear relationships, so linear models cannot do very well; also these parametric models cannot reveal the complex relationships among features in reality with easy, closed forms; although we try to deal with imbalanced problem, the minority classes still have little information because of our sparse features. Next part we try non-parametric models.

Table II: Results of Parametric Classification Models

Model	Total Accuracy	Weighted Recall	Class 3 Recall	Class 4 Recall
Logistic (original)	93.5%	93%	8%	1%
LDA (original)	92.1%	92%	19%	3%
Logistic-binary (original)	95%	95%	5%	
LDA-binary (original)	94%	94%	13%	
Logistic (combined balancing)	90%	90%	30%	10%
LDA (combined balancing)	85%	85%	27%	24%

3.2 Non-parametric modeling:

Since parametric models are not very suitable for this problem, we turned to non-parametric models and chose Decision Tree and Random Forest for classification.

First, we used grid search to find the best parameters through 5-fold cross validation. For Decision Tree, the best parameters are min_samples_split=2 and min_samples_leaf=1, both of which are the same as the default values, while the weighted recall always increases as the parameter max_depth increases. *[Figure 4.1]* So we kept these parameters at their default values and fitted the model, whose weighted recall is 91.31%. For Random Forest, since it's an embedded algorithm based on Decision Tree, we continued to use the parameters for Decision Tree and tried to search for the best value of the parameter n_estimators. We chose between 10, 30, 50, 80 and finally decided on 30, because it had the highest weighted recall, and a relatively small number of estimators can avoid the risk of overfitting. *[Figure 4.2]* The weighted recall of Random Forest is 94.06%, higher than that of Decision Tree.

What's more, Random Forest outperformed Decision Tree in all of the metrics, *[Table 4.8]* so we focused on Random Forest in the following study. Figure shows the importance of the top 20 features in the Random Forest model. Pressure, humidity and temperature are the most important features.

Imbalanced Dataset

Due to the imbalanced dataset, the recalls of classes with severity 3, 4 are not satisfying. To solve the imbalance problem, we tried resampling methods, including undersampling and SMOTE. When we implement both methods, the recall of Class 3, 4 can improve from 0.26, 0.20 to 0.45, 0.34 respectively. *[Table III]*

- Undersampling: We used all the data with severity 1, 3, 4 and randomly picked 300,000 samples with severity 2.
- Oversampling: We used the SMOTE method and increased the number of samples with severity 1, 3, 4 to 60000, 200000 and 200000 respectively.
- Undersampling + SMOTE: Based on the dataset used for undersampling, which has 300,000 samples of Class 2, we also used SMOTE to increase the number of samples with severity 1, 3, 4 to 40000, 150000 and 150000 respectively.

PCA

Because we used POIs and State as inputs, there were too many features, so we used PCA to reduce dimensionality. We applied PCA on the dummy variables. We selected the first 10 components, combined them with weather features to make up a new dataset, and fitted the model again. The scores were very close to the original model, so PCA is not effective enough to improve the performance of our model. *[Table III]*

Table III: Results of Random Forest Models Using Different Methods

Models	Weighted Precision	Weighted Recall	Weighted F1	Accuracy	Recall		
					1	3	4
Original Model	92.64%	94.06%	92.88%	94.06%	0.33	0.26	0.20
Undersampling	91.80%	91.17%	91.47%	91.17%	0.43	0.40	0.30
SMOTE	92.14%	93.48%	92.62%	93.48%	0.35	0.30	0.22
Undersampling + SMOTE	91.64%	88.65%	89.98%	88.65%	0.45	0.45	0.34
PCA	92.59%	94.01%	92.88%	94.01%	0.35	0.27	0.21

Change to Binary Dataset

As what we did for parametric models, we combined data with severity 1, 2 and 3, 4, denoted them as 0, 1 respectively and used this binary dataset instead. We also implemented resampling methods to weaken the influence of imbalance. When we implemented undersampling on Class 0 and oversampling on Class 1, the recall improved to 50.94% compared with 27.04% for the original model. *[Table IV]*

Table IV: Results of Binary Random Forest Models

Models	Number of samples		Precision	Recall	F1	Accuracy
	Severity 0	Severity 1				
Original Model	1,375,749	81,164	61.52%	27.04%	37.56%	95.00%
Undersampling	300,000	81,164	32.29%	45.91%	37.92%	91.68%
SMOTE	1,375,749	300,000	50.54%	30.48%	38.02%	94.47%
Undersampling+SMOTE	300,000	200,000	26.76%	50.94%	35.09%	89.52%

4 Regression Model for Duration(min) Prediction

In order to provide the predictive information for the police department to organize and clear traffic after accidents, we build several regression models to predict Duration(min) – time between the recorded start time and end time for an accident.

We have about 30 features that we considered important in determining the accidents' influence on traffic. Since our numerical features are mostly related to weather and cities' traffic conditions usually vary greatly in their sensitivity to the weather condition, which could lead to bias with mixed data in model fitting. In this case, we decided to use data from *New York* (4734) to fit the model and predict the accident's influence on traffic in this report. We would like to test the effectiveness of these features so that we were able to involve them into the prediction scheme.

We split data with 80% training and 20% testing. We did the splitting only once before running our two models for each algorithm so that we can ensure we use the same training dataset and testing dataset for each model.

Outliers

The response variable is the *Duration(min)* in our dataset. We first took a look into the distribution. In New York, the mean of duration is around 238 mins (3.96h), and the median is about 127 mins (2.11h). From the distribution we find there must be a set of extreme values, then we want to get rid of outliers in the dataset. We made graphs *[Figure 4.1]* for duration. After comparing different outlier selection criteria, we choose 95 percentile - 833 mins (13.8h) as the outliers lower bound, which seems quite reasonable in real traffic situations.

Nonlinear Transformation – take log of Duration(min) as Y

Linear regression assumes a linear relationship between all our input variables and the output variables. By fitting into a linear regression, we find the relationship between Duration(min) and features selected is not linear through the residual plot *[Figure 4.2]*. Therefore, we try to take log of Y and fit again, although the prediction result is not good, we find the variance of residuals are close to constant now, which indicates log transformation of Duration(min) is a feasible way to process data for better regression model fitting.

Random Forest Model for Duration(min) Prediction

Afterwards, we predicted the log Duration based on these models and evaluate them by metrics. We tried several regression models, including generalized linear regression, SVR, KNN, and finally found Random Forest is the most accurate one *[Table V]*.

Table V: Comparison of Regression Models

Test Accuracy	Linear Regression	Random Forest	KNN	SVR
R Squared	0.04	0.48	0.14	0.09
MSE	0.96	0.50	0.86	0.91

According to the Random Forest Model, we visualize the score of feature importance *[Figure 4.3]*. An insightful finding is that the location is very important in prediction of accident durations, and the weather condition like visibility and precipitation is not so influential on the response variable. Also, the importance score is relatively small for all the features. It seems reasonable because the correlation between those features and response variable Duration is not high overall.

5 Conclusion and Future Improvements

In this project, we trained classification models to predict the severity of accidents, and regression models to predict the duration of accidents. For both classification and regression, the Random Forest algorithm has the best performance. The location and time (day of week and hour) is very influential to the prediction result. Besides, weather features, like Pressure and Humidity, are the most important in our models. However, we also find lots of problems in our study, and the possible improvements in the future are as follows.

In the regression part, the model prediction accuracy is not so desirable. It is mainly because of the lack of correlation between features and the response variable, which means the features selected cannot explain the time of accidents influence on traffic very well. Therefore, we may need more data related to location (like road condition), as well as drivers' information (like driving years, ages), which might help with the model development and predictive accuracy.

We found that sometimes recorders may not have a clear standard for severity grading. So we highly recommend the Department of Transportation to develop a methodology for recording data and classifying severity, as well as provide training workshops for the recorders to make the dataset more effective.

We applied the PCA method to reduce dimensionality, but it doesn’t work very well. Maybe we should try to explore more methods, such as LDA and SVD, to reduce dimensionality.

Appendix

Figure 1.1 Correlation among Each Feature

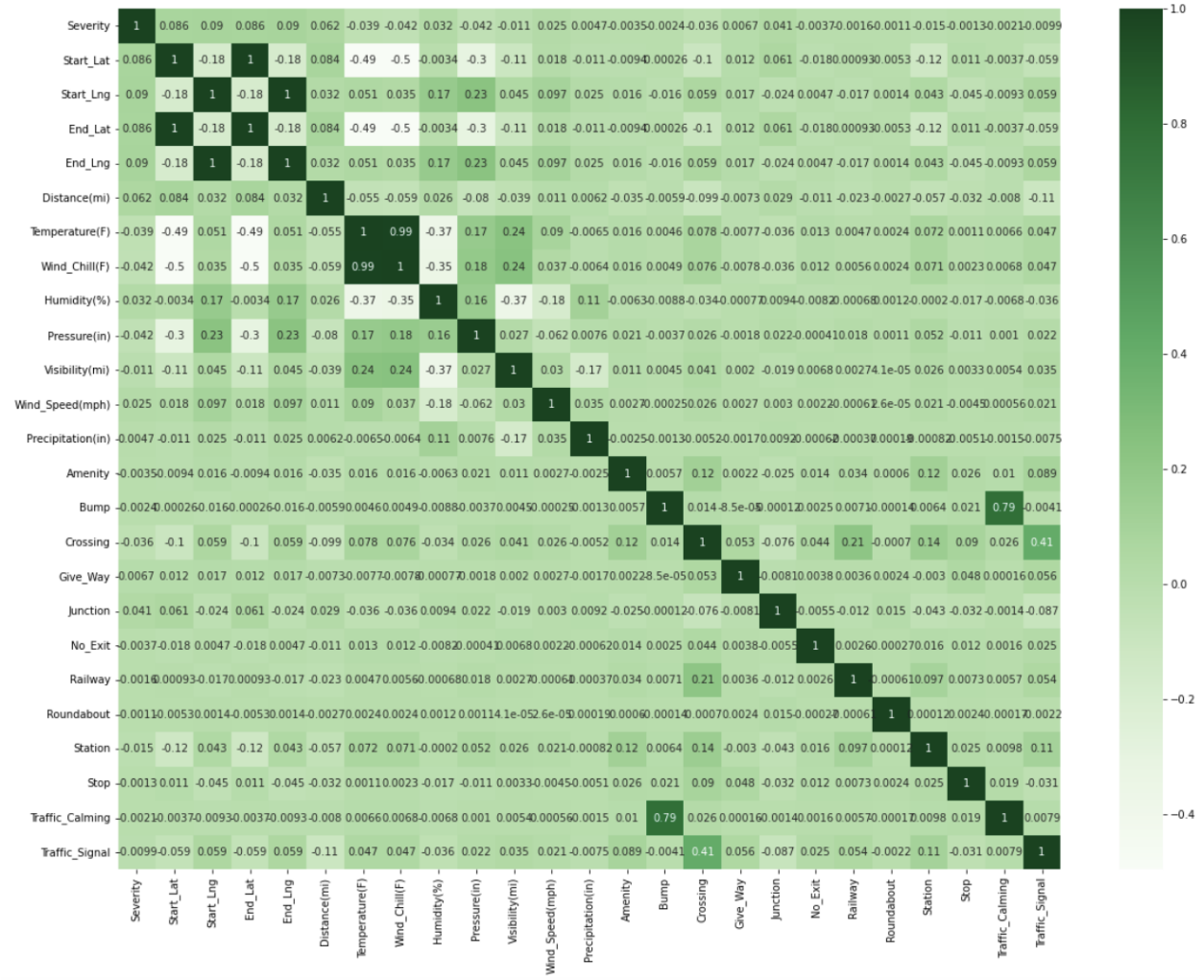

Figure 2.1 Pie Plot of Column “Severity”

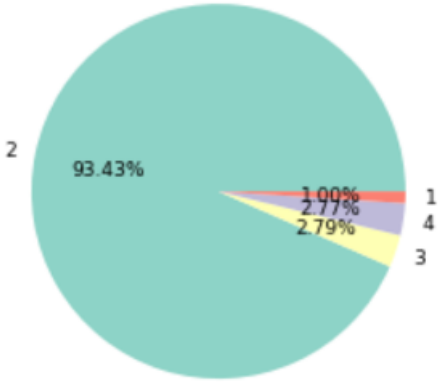

Figure 2.2 Number of US Accidents by States

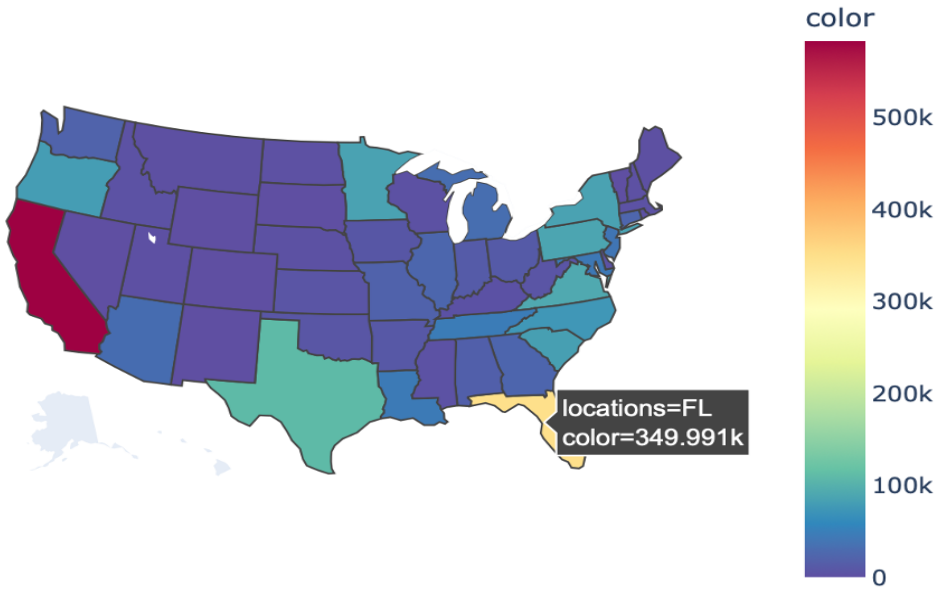

Figure 2.3 Number of Car Accidents in 15 cities

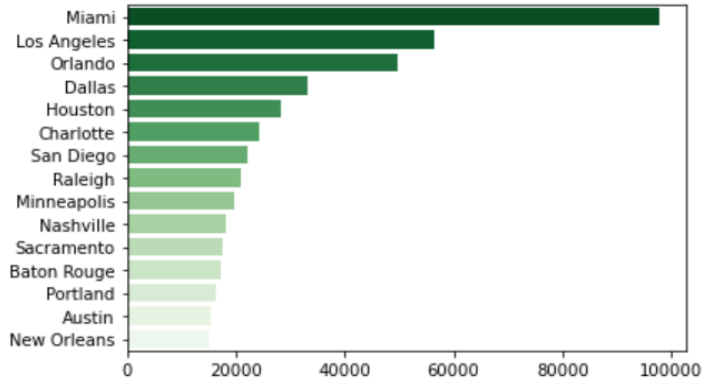

Figure 2.4 Pie Plot of Column “Side”

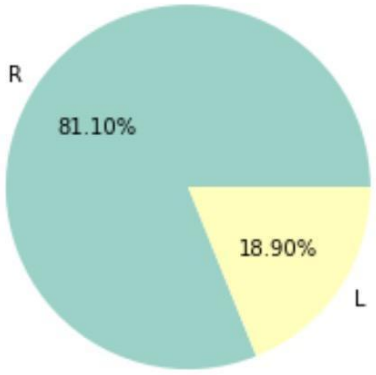

Figure 2.5 Number of Accidents by Day of the Week

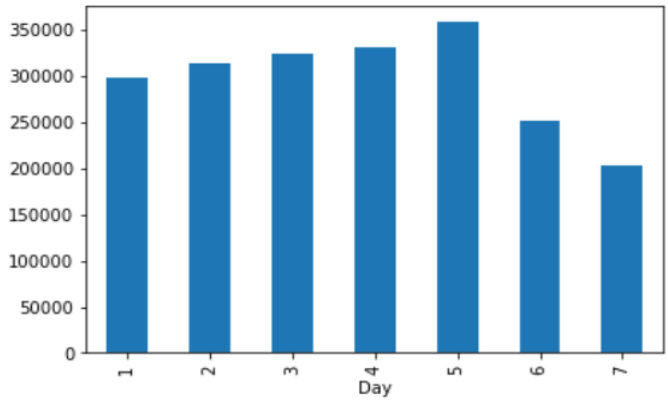

Figure 2.6 Number of Accidents by Hour for Each Level of Severity

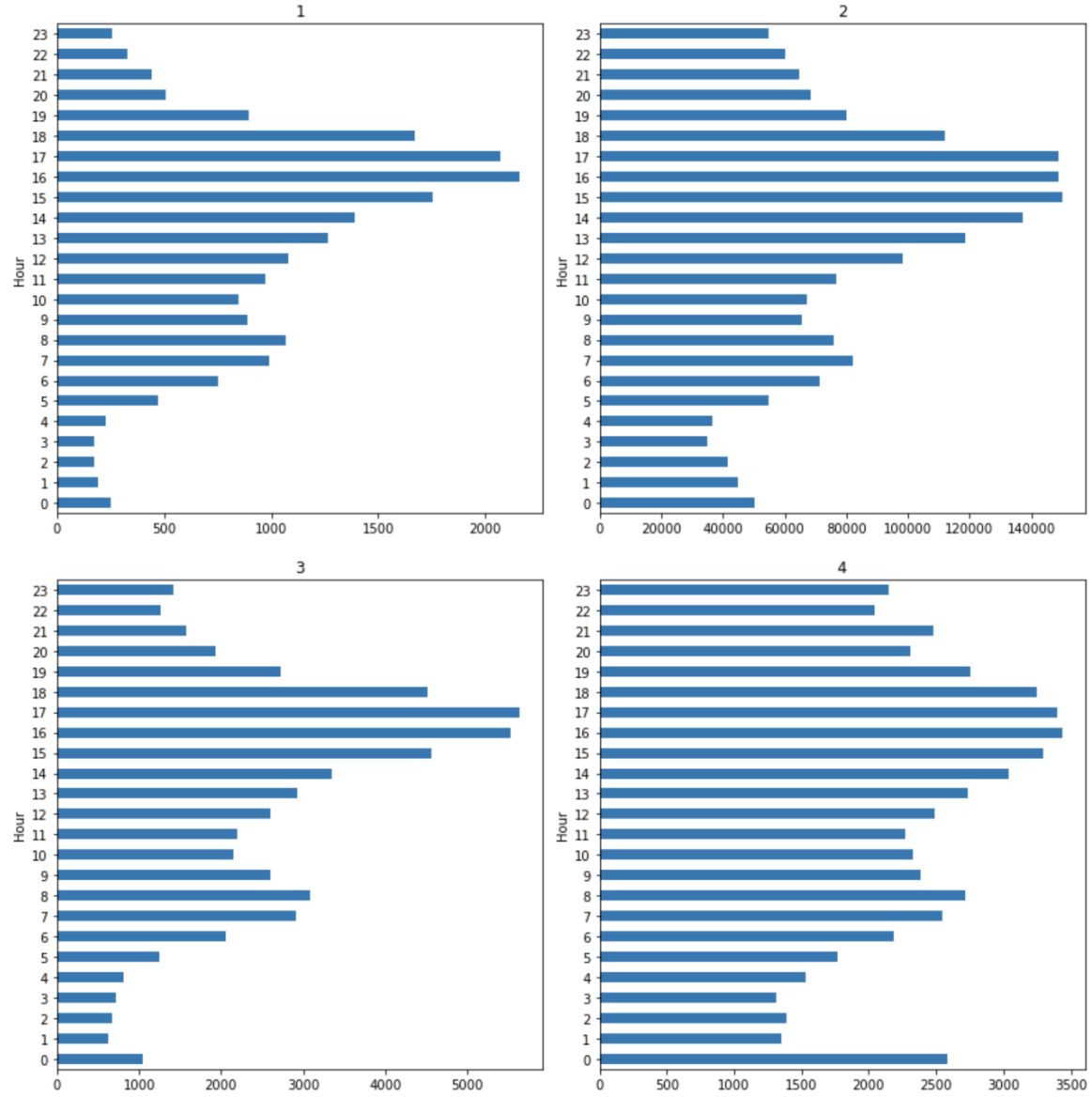

Figure 2.7 Violin Plot for “Humidity(%)” at Each Severity Level

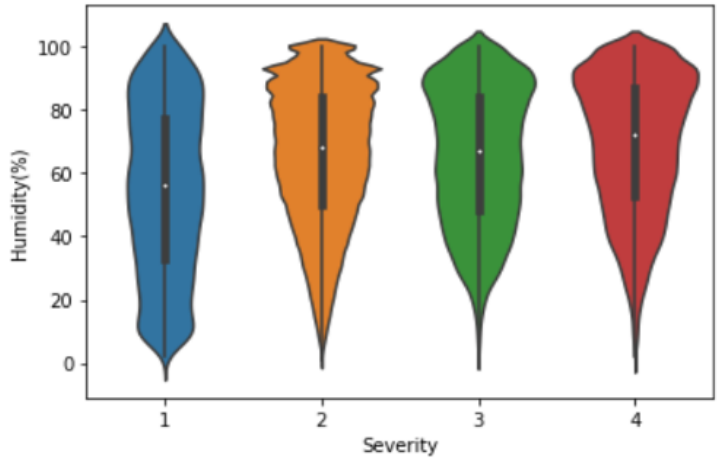

Figure 2.8 Percentage of Presences of Crossing and Traffic Signals at “Severity”=1

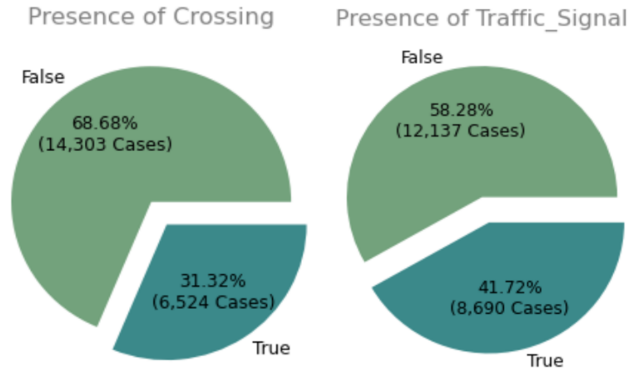

Figure 2.8 Percentage of Presences of Junction and Traffic Signals at “Severity”=4

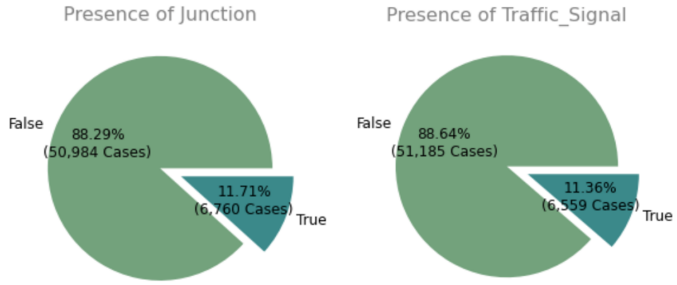

Table 3.1 Original training set: classification report of multi-class logistic regression

	precision	recall	f1-score	support
1	0.64	0.03	0.05	6211
2	0.94	1.00	0.97	583322
3	0.50	0.08	0.13	17455
4	0.57	0.01	0.01	17404
accuracy			0.93	624392
macro avg	0.66	0.28	0.29	624392
weighted avg	0.91	0.93	0.91	624392

Table 3.2 Original training set: classification report of multi-class LDA

	precision	recall	f1-score	support
1	0.13	0.19	0.15	6211
2	0.94	0.98	0.96	583322
3	0.40	0.19	0.26	17455
4	0.23	0.03	0.05	17404
accuracy			0.92	624392
macro avg	0.43	0.35	0.36	624392
weighted avg	0.90	0.92	0.91	624392

Table 3.3 SMOTE training set: classification report of multi-class logistic regression

	precision	recall	f1-score	support
1	0.55	0.11	0.19	6211
2	0.94	0.99	0.97	583322
3	0.40	0.18	0.25	17455
4	0.45	0.01	0.02	17404
accuracy			0.93	624392
macro avg	0.59	0.32	0.36	624392
weighted avg	0.91	0.93	0.91	624392

Table 3.4 SMOTE training set: classification report of multi-class LDA

	precision	recall	f1-score	support
1	0.12	0.21	0.15	6211
2	0.94	0.96	0.95	583322
3	0.37	0.20	0.26	17455
4	0.12	0.07	0.09	17404
accuracy			0.91	624392
macro avg	0.39	0.36	0.36	624392
weighted avg	0.90	0.91	0.90	624392

Table 3.5 Confusion matrix of original multi-class logistic regression

Predicted	1	2	3	4	All
Actual					
1	166	6004	40	1	6211
2	82	582092	1098	50	583322
3	1	16054	1354	46	17455
4	9	17074	194	127	17404
All	258	621224	2686	224	624392

Table 3.6 Original training set: classification report of binary class logistic regression

	precision	recall	f1-score	support
0	0.95	1.00	0.97	589533
1	0.61	0.05	0.10	34859
accuracy			0.95	624392
macro avg	0.78	0.52	0.53	624392
weighted avg	0.93	0.95	0.92	624392

Table 3.7 Original training set: classification report of binary class LDA

	precision	recall	f1-score	support
0	0.95	0.99	0.97	589533
1	0.46	0.13	0.20	34859
accuracy			0.94	624392
macro avg	0.70	0.56	0.59	624392
weighted avg	0.92	0.94	0.93	624392

Figure 4.1 Weighted Recalls of different max_depth for Decision Tree

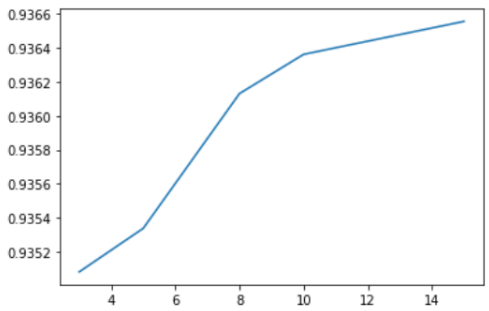

Figure 4.2 Weighted Recalls of different n_estimators for Random Forest

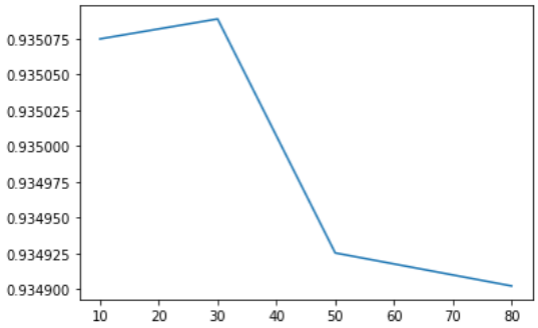

Table 4.8 Comparison on Metrics of Decision Tree and Random Forest

Model	Best Parameters	Weighted Precision	Weighted Recall	Weighted F1	Accuracy
Decision Tree	min_samples_split=2, min_samples_leaf=1 max_depth=None	91.27%	91.31%	91.29%	91.31%
Random Forest	n_estimators=30, min_samples_split=2 min_samples_leaf=1, max_depth=None	92.64%	94.06%	92.88%	94.06%

Figure 4.3 Feature Importance for Random Forest Model

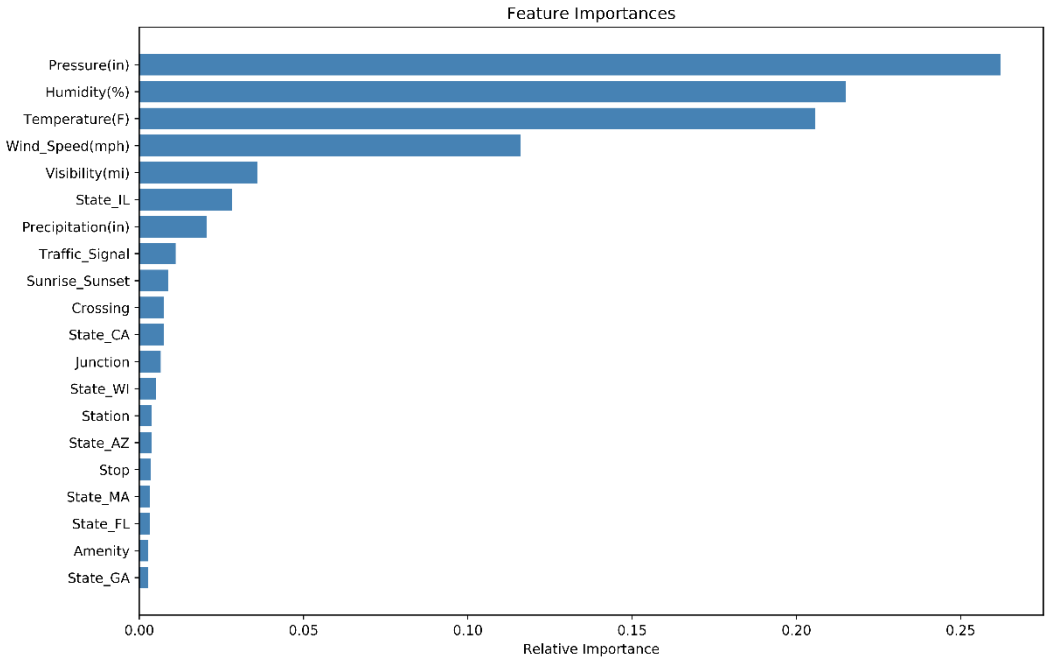

Figure 4.4 Explained Variance of Components

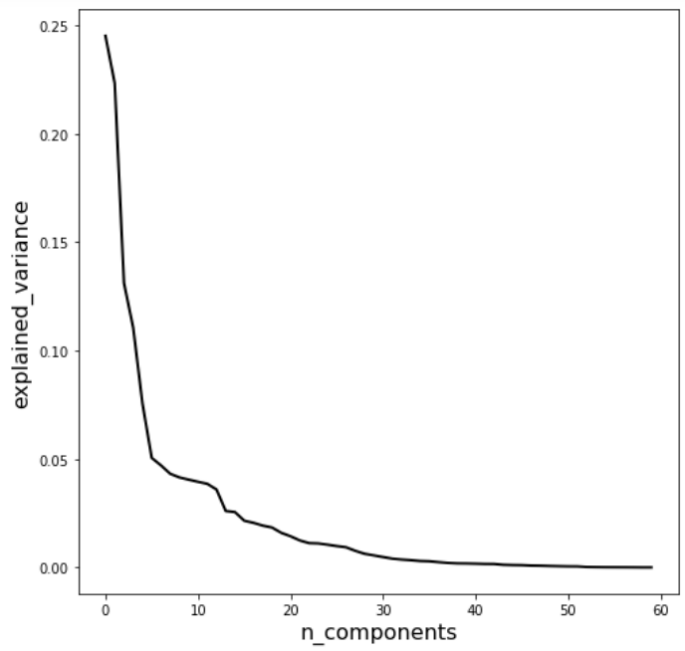

Figure 4.1 Distribution of Duration(min)

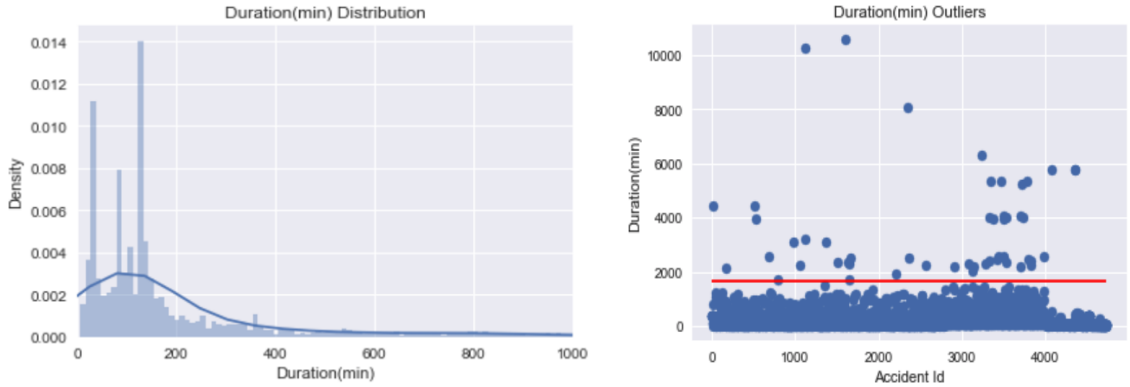

Figure 4.2 Residuals of Linear Regression

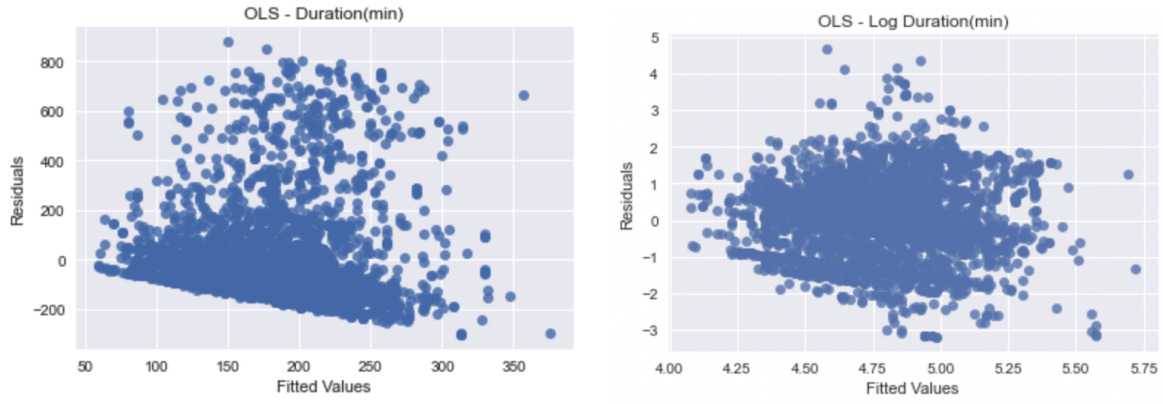

Figure 4.3 Feature Importance of Random Forest Regression Model