Accident Severity Prediction

Rudrani Bhadra r2bhadra@uwaterloo.ca
Department of Mathematics
University of Waterloo
Waterloo, ON, Canada

Abstract

Rapid increase of traffic volume on urban roads over time has changed the traffic scenario globally. It has also increased the ratio of road accidents that can be severe and fatal in the worst case. To improve traffic safety and its management on urban roads, there is a need for prediction of severity level of accidents. In this work, various tree-based machine learning models have been used for accident prediction and compared for prediction of road accident severity.

Keywords: Classification, Data Analysis, Ensemble methods

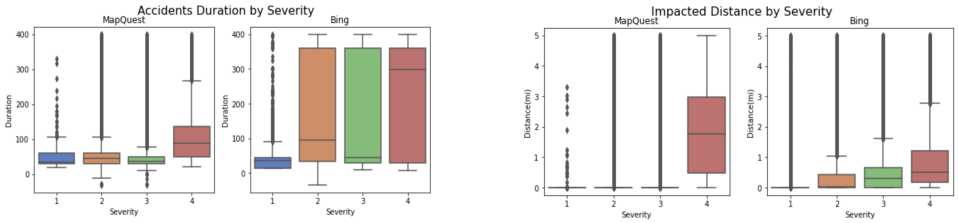
1. Introduction

Road traffic accidents are a major case of injuries, deaths, disabilities and loss. The economic and societal impact of traffic accidents cost people hundreds of billions of dollars every year. And a large part of losses is caused by a small number of serious accidents. Reducing traffic accidents, especially serious accidents, is nevertheless always an important challenge. Accidental severity is one of the popular research areas. The proactive approach, one of the two main approaches for dealing with traffic safety problems, focuses on preventing potential unsafe road conditions from occurring in the first place. For the effective implementation of this approach, accident prediction and severity prediction are critical. If we can identify the patterns of how these serious accidents happen and the key factors, we might be able to implement well-informed actions and better allocate financial and human resources.

In this project, different tree-based models have been used to predict the severity of accidents. The first objective of this project is to recognize key factors affecting the accident severity. The second one is to develop a model that can accurately predict accident severity. For a given accident, without any detailed information about itself, like driver attributes or vehicle type, the model is supposed to be able to predict the likelihood of this accident being a severe one.

2. Literature Review

Traffic accident data analysis has been considered by many researchers. In DW et al. (2011), they have used logistic regression and worked on the severity of traffic accidents in the United States. Their model's sensitivity and specificity were 40% and 98% respectively. B et al. (2016) used Support Vector Machine and multi-layer perceptron to analyze road accidents. Support Vector Machine outperformed with 94% accuracy. P et al. (2017) used



(a) (b)

Figure 1: Distribution of accidents for different sources

machine learning models like Decision Tree, Naive Bayes, and Support Vector Machine for classification. RE et al. (2019) used Naive Bayes, Adaboost, Random Forest, and Logistic Regression to find highways with high risk of accident for traffic agencies. Random forest outperformed with 76% accuracy. In T and S (2010), they analyzed important road-way related variables that can affect road accident severity and used Decision Tree, Naive Bayes, and K-Nearest Neighbours to make decision rules for road safety measures.

3. Methodology

3.1 Dataset

The dataset is called US Accidents (3.5 million records) which is obtained from Kaggle, Moosavi et al. (2019a), Moosavi et al. (2019b). This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. The dataset contains 12 traffic attributes, 9 address attributes, 11 weather attributes, 13 POI attributes, 4 attributes related to period of day. Since the dataset is huge, preprocessing has been done to clean the dataset and to keep only the essential features required for the model to classify an accident as severe or not.

3.2 Data Preprocessing

The following steps have been done:

- Since the data comes from two sources *MapQuest* and *Bing*, analysis was done to find out whether they report the severity level in the same way. From the box plots in Fig 1, the overall duration and impacted distance of accidents reported by *Bing* are much longer than those by *MapQuest*, and that the same severity level holds different meanings for *MapQuest* and *Bing*. Since *MapQuest* seems to have a clear and strict threshold for severity level 4, cases of which only account for a tiny part of the whole dataset compared to *Bing* which doesn't seem to have a clear-cut threshold, especially in duration, but the data is more balanced, I decided to select *MapQuest* as we care more about serious accidents and the sparse data of such accidents is the reality we have to confront.

- Columns such as 'Weather_Condition' and 'Street' has been cleaned and renamed for simplification. Both weather conditions and street type names can be good predictors of serious accidents.
- Some features such as 'TMC', 'Distance(mi)', 'End_Time', 'Duration', 'End_Lat', and 'End_Lng' can be collected only after the accident has already happened and hence cannot be predictors for serious accident prediction. Hence, they are dropped from the dataset, while others such as 'Precipitation(in)', 'Wind_Chill(F)' are dropped as 60% of the data is missing.
- Columns related to continuous weather features like 'Temperature(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)'and 'Wind_Speed(mph)' were first grouped by location ('Airport_Code') and time ('Start_Month') and then the missing values replaced by median value of each group. For categorical weather features, majority rather than median was used to replace missing values.
- 'Minute' can also be an important predictor. But directly using it would produce an overabundance of dummy variables. Therefore, the frequency of 'Minute' was utilized as labels, rather than 'Minute' itself. Similar to 'Minute', some location features like 'City', 'Zipcode', 'County' and 'Airport_Code' that have too many unique values have been labeled by their frequency.

3.3 Data Resampling

The accidents with severity level 4 are much more serious than accidents of other levels, and so I have decided to focus on level 4 accidents and regroup the levels of severity into level 4 versus other levels. Class 0 is for accident levels 1, 2 and 3, while class 1 is for accident level 4. Since the dataset is an imbalanced one, the combination of over- and under-sampling will be used since the dataset is large enough. Level 4 accidents was randomly oversampled to 100,000 and other levels was randomly undersampled to 100,000.

3.4 Data Analysis

In this section, I have analyzed the severity of accidents versus some of the features in the dataset.

3.4.1 Year

Fig 2 shows the distribution of accidents in the US throughout 2016 to 2020. It seems that the number of severe accidents have increased lately in the years 2019 and 2020.

3.4.2 Month

Fig 3 shows the distribution of accidents throughout the months. The count of other levels accidents is mostly consistent throughout, whereas the number of level 4 accidents rapidly increased from May to June and remained stable until September then increased slightly again from October.

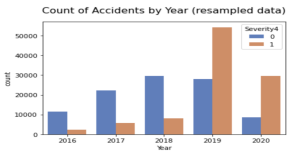


Figure 2: Count of Accidents by Year

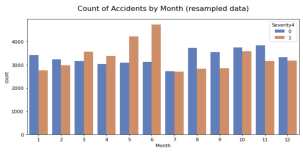


Figure 3: Count of Accidents by Hour

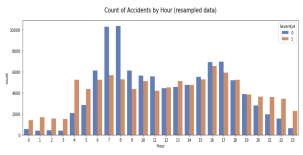


Figure 4: Count of Accidents by Hour

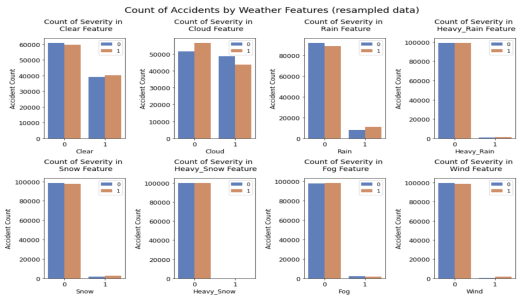


Figure 5: Count of Accidents by Weather

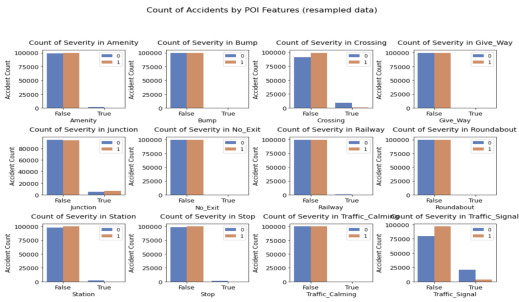


Figure 6: Count of Accidents by POI features

3.4.3 Hour

Fig 4 shows the distribution of accidents throughout the day. Most accidents happened during the daytime, especially AM peak and PM peak. When it comes to night, accidents were far less but more likely to be serious.

3.4.4 weather condition

As seen from Fig 5, accidents are little more likely to be serious during rain or snow while less likely on a cloudy day.

3.4.5 POI features

As seen from Fig 6, accidents near traffic signal and crossing are much less likely to be serious accidents while little more likely to be serious if they are near the junction. Other POI features are so unbalanced that it is hard to tell their relation with severity from plots.

3.4.6 Period-of-Day

From Fig 7 accidents were less during the night but were more likely to be serious.

3.4.7 Side

From Fig 8 shows that the right side of the line is much more dangerous than left side.

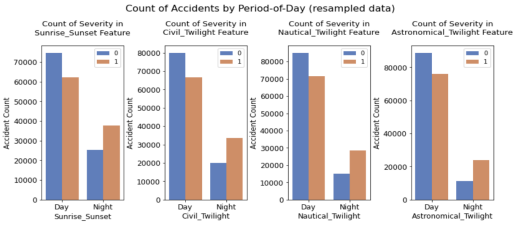


Figure 7: Count of Accidents by Period-of-Day

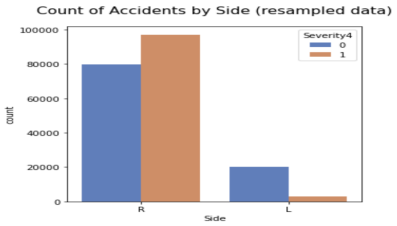


Figure 8: Count of Accidents by Side

3.4.8 Street

Fig 9 shows the correlation between key street words and accident severity level 4. Interstate Highway (denoted by 'I') turns out to be the most dangerous street. Other street features like road, drive, and avenue seem to be relatively safe.

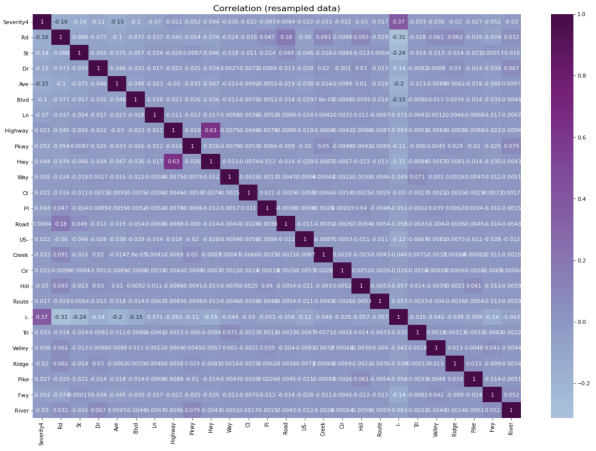


Figure 9: Correlation between key street words and severity

3.5 Model Selection

In order to predict if an accident to occur will be severe or not, the following classification models have been used: logistic regression, decision tree, random forest, gradient boosting, and adaboost.

Logistic regression has been chosen as the base model as it is a binary classification problem.

Decision tree is chosen as it is the simplest tree-based method and can be compared against other tree-based ensemble methods.

Random forest is selected as one of the classifiers as it builds trees in parallel to each other which are not correlated to each other and applies the general technique of bagging

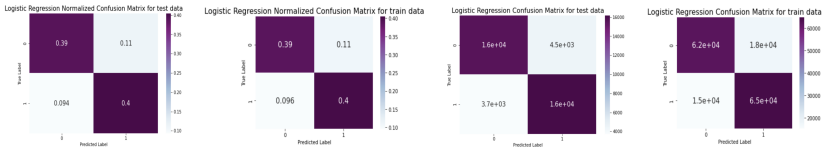


Figure 10: Confusion Matrices for Logistic Regression

to tree learners. This leads to better model performance as it decreases the variance of the model without increasing the bias.

Gradient booting and adaboost are selected due to their effectiveness. Like other boosting methods, gradient boosting and adaboost combine weak learners into a strong learner in an iterative manner.

3.6 Evaluation Metrics

For supervised classification of an unbalanced dataset, accuracy is not optimal. However, a balanced dataset is sampled from the original dataset and a relatively balanced one is achieved. Thus, accuracy score would be the main evaluation metric for the data. In addition, other metrics for some models such as F-1 score, precision and recall are provided as well. Given a confusion matrix, precision, recall, and F-1 score are calculated as follows:

$$Precision = \frac{True\ Positive}{Total\ Predicted\ positive}$$

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4. Results

For all five methods, the dataset is split into training and test datasets. 80% of the dataset is used for training and the remaining 20% is used for testing. All models were trained with the best parameters obtained from grid search cross validation.

4.1 Logistic Regression

The logistic regression model was trained with penalty='L2'. Even with the best parameter setting, logistic regression yielded poor results for both training data and test data. Its training accuracy was 79.2% while its test accuracy was 79.4%. Fig 10 shows its confusion matrices.

4.2 Decision Tree

The decision tree model was trained with min_samples_split=5. With the best parameter setting, decision tree yielded very good results for both training data and test data. Its

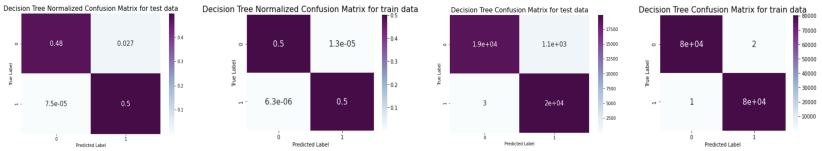


Figure 11: Confusion Matrices for Decision Tree

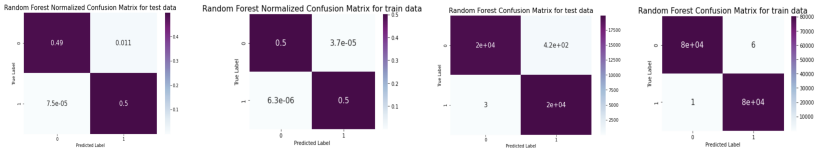


Figure 12: Confusion Matrices for Random Forest

training accuracy was 100% while its test accuracy was 97.4%. Fig 11 shows its confusion matrices. Fig 13 is the feature importance plot which shows the most useful features to predict severity. Among them, interstate highway('I-') is the most important feature. In addition to these features, spatio-temporal features and weather features like pressure, temperature, humidity, and wind speed are also very important.

4.3 Random Forest

The random forest model was trained with n_estimators=40 and max_depth=40. With the best parameter setting, random forest yielded very good results for both training data and test data. Its training accuracy was 100% while its test accuracy was 98.9%. Fig 12 shows its confusion matrices. Fig 14 shows its feature importance plot. The top 15 important features of random forest model are almost as same as decision tree model.

4.4 Gradient Boosting

The gradient boosting model was trained with n_estimators=50 and max_depth=4. With the best parameter setting, gradient boosting produced average results for both training

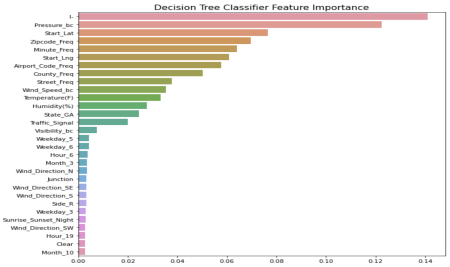


Figure 13: Line chart showing the important features in decision tree

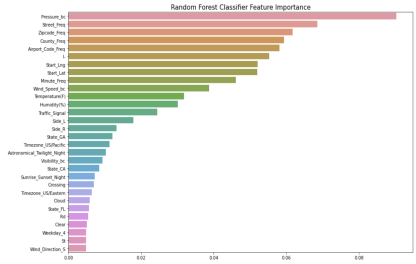


Figure 14: Line chart showing the important features in random forest

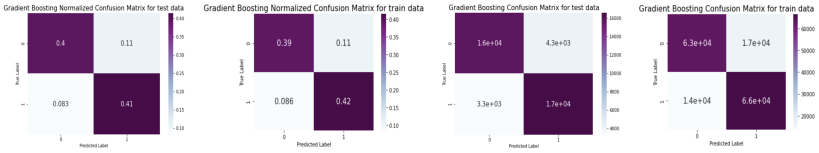


Figure 15: Confusion Matrices for Gradient Boosting

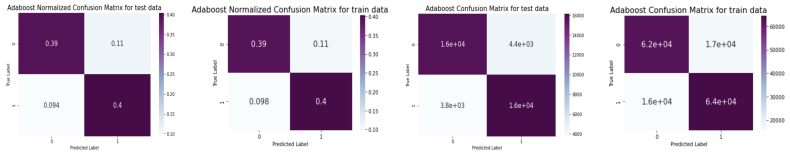


Figure 16: Confusion Matrices for Adaboost

data and test data. Its training accuracy was 80.9% while its test accuracy was 81%. Fig 15 shows its confusion matrices.

4.5 Adaboost

The adaboost model was trained with n_estimators=50. With the best parameter setting, adaboost produced poor results for both training data and test data. Its training accuracy was 79.3% while its test accuracy was 79.7%. Fig 16 shows its confusion matrices.

Table 1: Performance of five models

Methods	Precision		Recall		F1 Score		Accuracy	
	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.783	0.780	0.812	0.812	0.797	0.796	0.792	0.794
Decision Tree	0.999	0.947	0.999	0.999	0.999	0.973	1.000	0.973
Random Forest	0.999	0.979	0.999	0.999	0.999	0.989	1.000	0.989
Gradient Boosting	0.795	0.798	0.825	0.837	0.809	0.817	0.809	0.810
Adaboost	0.790	0.784	0.800	0.808	0.800	0.796	0.793	0.797

5. Conclusion

In this project, I have studied how to predict the likelihood that an accident can be severe using tree-based ensemble methods. As seen in Table 1, it can be concluded that random forest performs the best overall followed by decision tree. Both logistic regression and adaboost did not perform well in this classification task. As for future work, deep learning models can be applied on this problem, detailed relations between some key factors and accident severity can be further studied, and more ways can be studied as to how to deal with imbalanced datasets.

References

Sharma B, Katiyar VK, and Kumar K. Traffic accident prediction model using support vector machines with gaussian kernel. *Proceedings of fifth international conference on soft computing for problem solving*, pages 1–10, 2016.

Kononen DW, Flannagan CA, and Wang SC. Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. *Accident Analysis Prevention*, 43(1):112–122, 2011.

Moosavi, Sobhan, and Mohammad Hossein Samavatian. A countrywide traffic accident dataset. 2019a.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. *In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019b.

Tiwari P, Kumar S, and Kalitin D. Road-user specific analysis of traffic accident using data mining techniques. *International Conference on Computational Intelligence, Communications, and Business Analytics*, pages 398–410, 2017.

AlMamlook RE, Kwayu KM, Alkasisbeh MR, and Frefer AA. Comparison of machine learning algorithms for predicting traffic accident severity. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, page 272–276, 2019.

Beshah T and Hill S. Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in ethiopia. *AAAI Spring Symposium: Artificial Intelligence for Development*, 24:1173–1181, 2010.