US Car Accidents Severity Analysis

Wanyue Xiao, Yingxue Gao and Yiyuan Cheng

School of Information Studies, Syracuse University, Syracuse, NY 13244 USA

xwanyue@syr.edu; ygao65@syr.edu; ycheng26@syr.edu

Abstract

Motivated by the demand to develop a traffic prediction strategy since car accidents are inevitable in reality. Patterns involved in several crashes could be detected or even predicted if an appropriate machine learning model were developed. Based on the dataset collected from different traffic control agencies, this paper trains several prevailing machine learning models to predict the severity level of car accidents. Experiments results show that tree-based algorithms outperformed the rest classifiers when it comes to binary categorical classification problems.

1 Introduction

Since the car was invented in 1886, it has gradually become a widely used means of transportation in the United States. However, statistics from Association for Safe International Road Travel show that more than 38,000 people lost their lives and more than 4.4 million are injured in the United States each year due to car accidents. In order to have a better understanding about the causes of car accidents, the team analyzed the US Accidents dataset using data mining methods and created different models to predict the severity of accidents.

Data mining is the process of finding correlations or patterns from large amounts of data. The data mining techniques can be separated into two groups, descriptive approaches and predictive approaches. The descriptive category includes Association and Clustering, and the predictive category contains Classification and Regression (Figure 1). In this report, we used Association Rule Mining, Logistic Regression and several Classification approaches, such as Decision Tree, Support Vector Machine (SVM), K Nearest Neighbor (KNN) etc. to analyze the dataset. Among these different models, Gradient Boosting Machine (GBM) has the best performance. The Shiny App for creating and tuning different machine learning models is also generated (See Appendix).

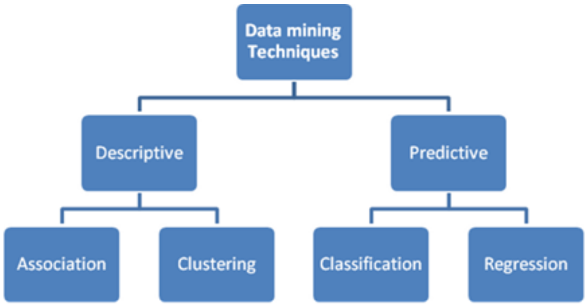


Figure 1 Data mining techniques [1] .

2 Project Design

2.1 Experiment Design

The purpose of the project is to analyze the important features and how they affect the severity of the car accident, which consists of two steps. There are five components for the project analysis process: data preprocessing, exploratory data analysis, data sampling, model generation, and performance evaluation. First, the team fixed the data quality issues such as missing values, outliers, and duplicates in the extracted data. Some variables are transformed in order to provide more intuitive and precise information. The team also tailored three kinds of data with different data manipulation methods such as categorization, numerization, and standardization applied, for specific machine learning models. Second, exploratory data analysis was conducted to discover the distributions and patterns among variables, which helps identify potential key features. Then, an equal size of low and high severity accident records is sampled out respectively to create a series of train and test data for model construction.

Three types of machine learning algorithms used in this analysis, which are 1) Association Rule Mining, 2) classification models including Decision Tree, Random Forest, K Nearest Neighbor, Support Vector Machine, Gradient Boosting Machine, and 3) Logistic Regression. Finally, there is a comparison of the performance across different machine learning algorithms. The team can obtain an accurate and comprehensive conclusion of the important features and their effects on accident severity.

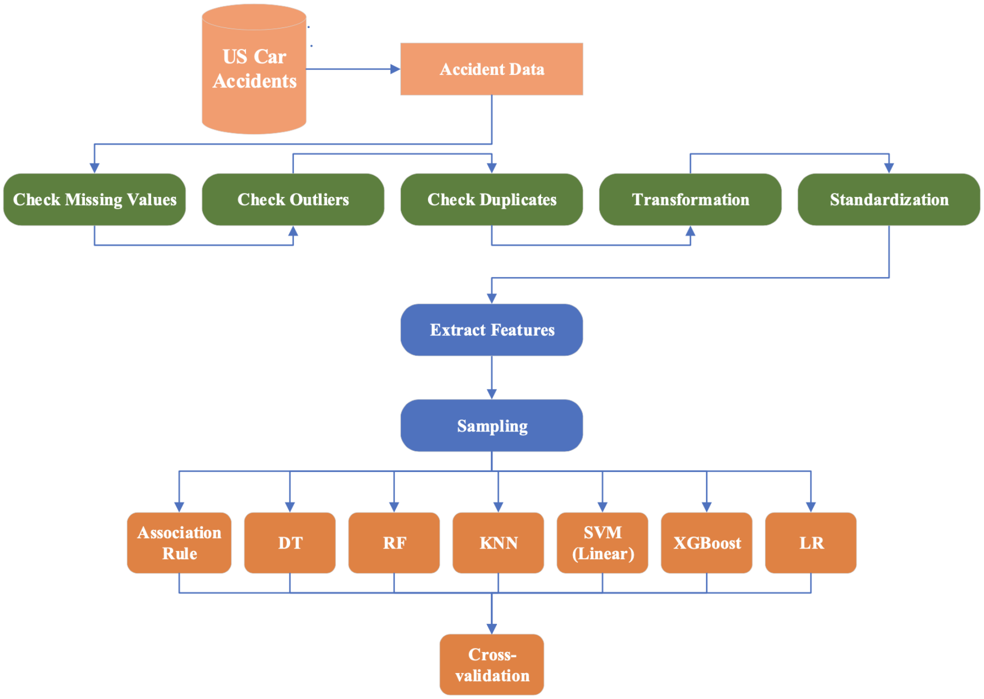


Figure 2 Experiment Design Flow Chart.

2.2 Dataset

The dataset is about countrywide car accidents that happened in 49 states in the United States from February 2016 to December 2019, containing 3.0 million records in total. The team randomly sampled and downloaded 50 thousand records and 49 variables for this analysis. The target variable of this research is accident severity, which is labeled as Low (severity 1 and 2) and High (severity 3 and 4). The rest are explanatory variables mainly describing the location, time, weather, traffic, and report of the accidents.

2.3 Machine Learning Method

Association Rule Mining Aims at identifying the important correlations and frequent patterns among objects in the data repositories [2, 3, 4], association rules are extracted mainly based on two measures, minimal support ($support(X) = count(X)/N$) and minimal confidence ($confidence(X => Y) = support(X, Y)/support(X)$). In this analysis, the Apriori algorithm is used. It includes two steps: candidate generation process to extend items to the frequent subsets, and then test these candidates against the data [2]. Apriori is efficient during the candidate generation process, however, this process is time, space and memory consumption since it needs to scan the dataset multiple times [4].

Decision Tree It is a kind of classification technique which uses a tree-like model to continuously split data according to a certain parameter, with each node corresponding to certain records in the training dataset [1]. As a binary logical tree, the root node represents the target variable and the decision nodes represent the important predictor variables it found, which well illustrates the classification process.

K Nearest Neighbor This algorithm considers observations that are close to each other as the same class and k nearest neighbor rule will be generated [8, 9]. It is a classification method which classifies an unclassified sample based on its nearest neighbors identified by distance metric [9]. Euclidean distances are usually calculated to measure the similarities for most KNN classifiers [9].

Support Vector Machine Classification is accomplished by identifying which side of the final hyperplane a data point falls under this algorithm. The support vector machine mainly considers data points (support vectors) near the hyperplane. So, it has an advantage in dealing with noisy data. For this case, linear function is used since this is a linear separable question.

Random Forest Ensemble methods are techniques that use multiple algorithms to produce one optimal predictive model. Bagging and boosting are two well-known ensemble methods. Random forest is proposed based on bagging method. It generates the model by combining several trees which are built separately together [5]. Firstly, several bootstrap samples are generated from the training dataset. Then each bootstrap sample is used to build a tree. Finally, these trees are combined together for predicting the testing data [6].

Gradient Boosting Machine Similar to Random Forest, Gradient Boosting is an ensemble method which generates a final result based on a collection of individual models. The predictability of those individual methods might be weak and even tend to overfitting yet combining those models in an ensemble will lead to an overall much improved result. In this project, XGBoost, which is also called Extreme Gradient Boosting, will be used. It is a special gradient boosting method which uses more accurate approximations to find the best tree model by computing second-order gradients and advanced regularization (such as L1 and L2).

Logistic Regression It is one of statistical regression models which could be used to predict binary classes by utilizing sigmoid function. A linear predictor (logit) represents the logarithmic odds, a variable describing the ratio of success to ratio of failure, of independent variables. The coefficients computed enable one to understand the effect of an independent variable towards the dependent variable.

2.4 Performance Evaluation

As for the association rule mining, the team narrows the most interesting rules by improving the value of minimal support and minimal confidence. Also, one can limit the number of items in the left-hand side and specify the items in the right-hand side. An important measurement of rules is lift ($lift(X => Y) = confidence(X => Y)/support(Y)$). The higher the lift, the better the rule.

The performance of other machine learning models is evaluated by three-times repeats of 10-fold Cross-Validation implemented on the train data, and a confusion matrix on an independent test data. All results are reported using the following three performance metrics: sensitivity, specificity, accuracy, precision, and recall. In addition, the team also computes the Receiver Operating Characteristics (ROC) curve and Area Under Curve (AUC), which is a preferred metric for evaluating the performance of a binary classifier.

3 Data Preprocessing

Variables removal Before continuing data preprocessing, columns that contain the same value (such as the Country) or contain unique values (such as ID) should be removed.

Missing Values and Empty Values The first step of data preprocessing is checking the columns with missing values and empty values. Normally, for columns that do not have many missing values, one can drop rows with missing values directly. In this scenario, the dataset should be processed separately. For columns whose missing rate reaches to 60%, it is unnecessary to keep them. For the rest, one can replace those null values by using the KNN Imputation algorithm which is useful for matching a point with its closest k neighbors in a multi-dimensional space.

Outliers Management Examining the boxplot distribution in figure 3, one can find that distance, pressure, temperature, visibility, and wind_speed have outliers. In this project, one defines that values are below 0.25% quantile or are above 99.75% are outliers. For those outliers one can use average to replace them if the distribution of the original column is normal distribution. Otherwise one can use mode to replace them. For example, checking the outliers that are above 99.75% quantile in the Distance column, the result is 11.54 mile which is absurd in reality. Then the average of the distance column will be used to replace those outliers. For the wind_speed variable, according to the research, wind speed could range from 0 to 73 which indicated Hurricane. The maximum is 47 which is reckon as “Fresh Gale”. Hence, those outliers will be kept.

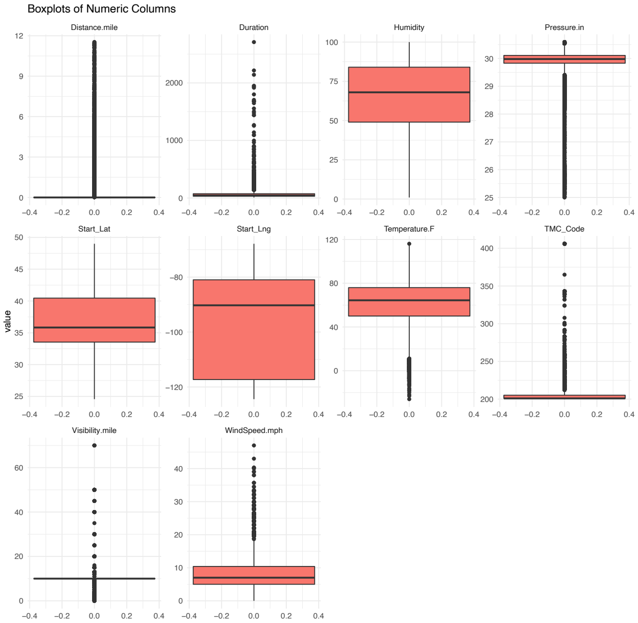


Figure 3 Boxplot of Numeric Variables

Feature Transformation Given that the dataset contains time related variables, such as “Accident Start Time” and “Accident End Time”, one can make an assumption that the duration between start time and end time might be a potential indicator.

Compress Category of Weather Condition The original categorical variable, weather condition, contains 70 levels, with a majority of them having fewer than 100 observations. Besides, there are many trivial descriptions that are hardly distinguishable in future analysis. For example, over 10 terms such as “Drizzle”, “Light Drizzle”, “Light Rain”, “Light Rain Shower”, and “Rain / Windy” are used to describe a normal rain. The unnecessary categories are summarized into the most common and distinct weather conditions. Wind-related information is removed because it is already covered by another variable, wind speed. Snow and ice, fog and dust are combined as they result in similar traffic conditions -- snow and ice denote a low temperature and slippery road, and the latter denotes a low visibility. Therefore, those 70 levels are condensed to 9 categories: clear, fair, heavy rain, rain, snow or ice, overcast, cloudy, fog or dust, and thunderstorm.

Categorization To conduct Association Rules Mining, all the numeric variables need to be converted to categorical variables. TMC code is directly categorized since each code represents one kind of accident description. Distance, temperature, humidity, and duration are divided according to their quantile distribution. As for the pressure, lower than 29.2 inches can be considered as a low pressure and often related to a storm weather whereas high pressure that is greater than 30.2 inches always denotes a sunny day. Wind speed is partitioned into “0-2”, “3-4”, and “>5” based on the Beaufort Force level. Since 10-mile-visibility is a common definition of good visibility, this variable is separated at 5 and 10 miles. Start and end hour are converted to parts of the day.

One-Hot Encoding and Standardization Data fed into the decision tree, random forest, support vector machine, gradient boosting machine, and logistic regression needs to be numeric. So, previous categorical variables are transformed using one-hot encoding. Those variables should also be standardized to have each weighted equally (except for the decision tree model).

Data Sampling The records of low severity accidents are twice as many as the high-severity counterparts in the dataset, which would give rise to a biased data analysis. Thus, 10000 low severity and 10000 high severity records are sampled out as the train data, another 5000 records also in a balanced 1:1 ratio serve as the test data. To sum it up, 25000 rows and 21 variables are finally selected for model generation and validation.

4 Exploratory Data Analysis

Figure 4 demonstrates the distribution of different severity in various weather conditions. Given that the number of low is twice as much as the number of high, one might need to divide the number of “Low Severity” by two and make judgement. Weathers like “Heavy Rain”, “Snow or Ice” has a higher High Severity count compared with the Low. The rest seen to have a similar value.

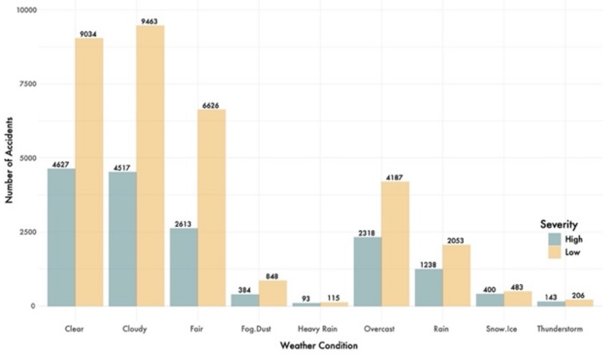


Figure 4 Stacked Bar Plot by Weather Conditions

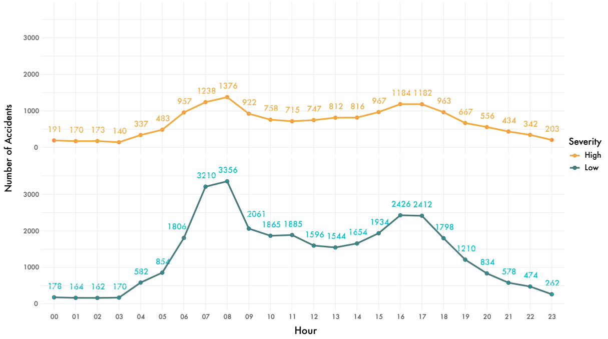


Figure 5 Time Series Plot by Hour

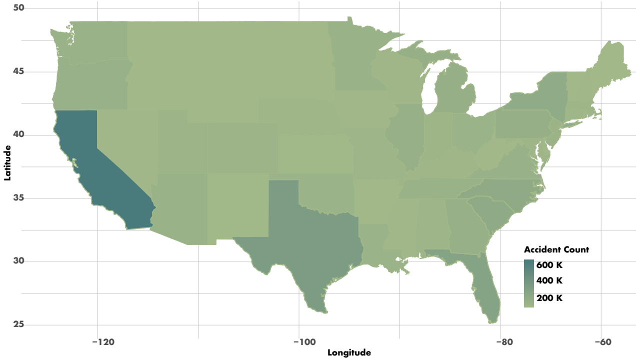


Figure 6 Map Plot by Counts of Car Accident

Figure 5 is a time series plot which records accumulated accident count by different hour in a day. The two sub plots follow a same pattern but with a different intensity. Both the number of high and low tends to increase during 3am to 9am and 4pm and 5pm. This could be explained by the regular or the routine hour since people will go to work and back to home. It is interesting to find that the number of accidents is higher in daylight. Since intuitively, people tend to have a severe car accident in night.

Figure 6 is an American map plot. The color deepens as the number of accidents counts increase. It pretty obvious that California has the highest accumulated number of accidents up to 2019. The state which has the second highest and third highest counts are Taxes and Florida. Those observations could be explained by the vast amount population in the states. Besides, states near the east seaboard seems to have a higher value compared with the inland states.

5 Results

5.1 Association Rule Mining

Table 1 shows the top 5 rules related to low severity accidents. One can summarize that the absence of junction and traffic signal around seems to cause low severity accidents. Those accidents tended to be reckoned as "Normal Accident" (which represents TMC code 201) and to be reported by channel MapQuest. The finding that those accidents most likely happened in daytime is consistent with the time series plot in the visualization section.

Table 1 Top 5 Association Rules related with Low Severity Accident

LHS	RHS	Supp-ort	Confi dence	Lift	Count
TMC_Code = 201, Juction = False, Traffic_Signal = True	Severity = Low	0.1323	0.9304	1.3903	5204
TMC_Code = 201, Traffic_Signal = True	Severity = Low	0.1327	0.9290	1.3882	5221
Soruce = MapQuest, Juction = False, Traffic_Signal = True	Severity = Low	0.1359	0.9137	1.3653	5347
Soruce = MapQuest, Traffic_Signal = True	Severity = Low	0.1363	0.9115	1.3620	5364
Juction = False, Traffic_Signal = True, Nautical_Twilight = Day, Astronomical_Twilight = Day	Severity = Low	0.1322	0.8992	1.3437	5202

Correspondingly, table 2 shows the top 5 rules related to high severity accidents. Compared with table 1, it is interesting to find that the high-severity accident tended to happen when crossing and stations were missing in the nearby location and the pressure index was between 19.5 to 30.2. Traffic signals, still, are important to the criteria of high severity accidents. Besides, those accidents also happened in daytime and were reported by channel MapQuest.

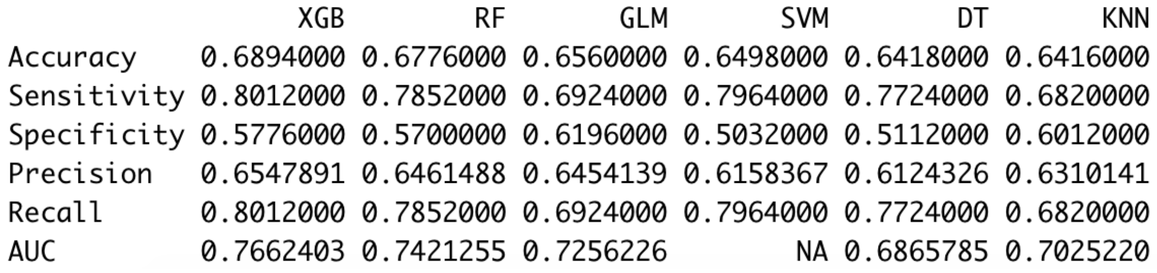
Table 2 Top 5 Association Rules related with High Severity Accident

LHS	RHS	Supp-ort	Confi dence	Lift	Count
Source = MapQuest, Pressure.in = 29.5-30.2, Crossing = False, Station = False, Traffic_Signal = False	Severity = High	0.1830	0.4251	1.2853	7200
Source = MapQuest, Pressure.in = 29.5-30.2, Crossing = False, Traffic_Signal = False	Severity = High	0.1846	0.4235	1.2803	7262
Source = MapQuest, Pressure.in = 29.5-30.2, Crossing = False, Station = False, Traffic_Signal = False, Astronomical_Twilight = Day	Severity = High	0.1571	0.4180	1.2637	6168
Source = MapQuest, Pressure.in = 29.5-30.2, Station = False, Traffic_Signal = False,	Severity = High	0.1837	0.4178	1.2631	7229
Source = MapQuest, Crossing = False, Station = False, Traffic_Signal = False,	Severity = High	0.2393	0.4171	1.2611	9415

5.2 Performance Evaluation

Based on the results obtained from data preprocessing, we first analyze the predictive power of six classifiers, i.e. DT, RF, KNN, SVM, XGB, and GLM. The models are tested on the training dataset by using the 10-fold cross validation and experimental results are shown in Table 3.

Table 3 Performance Comparison of six classifiers with 10-fold Cross-validation



```
                  XGB        RF       GLM       SVM        DT       KNN
Accuracy    0.6894000 0.6776000 0.6560000 0.6498000 0.6418000 0.6416000
Sensitivity 0.8012000 0.7852000 0.6924000 0.7964000 0.7724000 0.6820000
Specificity 0.5776000 0.5700000 0.6196000 0.5032000 0.5112000 0.6012000
Precision   0.6547891 0.6461488 0.6454139 0.6158367 0.6124326 0.6310141
Recall      0.8012000 0.7852000 0.6924000 0.7964000 0.7724000 0.6820000
AUC         0.7662403 0.7421255 0.7256226        NA 0.6865785 0.7025220
```

Table 3 indicates that the optimized XGBoost provides the highest performance in terms of Accuracy, Sensitivity, Precision, and Recall compared to the other methods for the prediction of severity level. RF is the classifier which has the second highest accuracy rate and second highest precision value. It is worthwhile to mention that even though XGB does not achieve the highest accuracy, it performs well in identifying potential high-severity accidents with the highest sensitivity value of 0.8. It is also evident that the DT model performs worst in this task. In addition, the algorithms of GLM and KNN show the acceptable performance with the AUC value larger than 0.7. To assure the distinct and high quality of the target figure, ROC curves corresponding with GLR, DT, RF, KNN, and XGB are shown in Figure 7, which illustrates the consistent findings with Table 3.

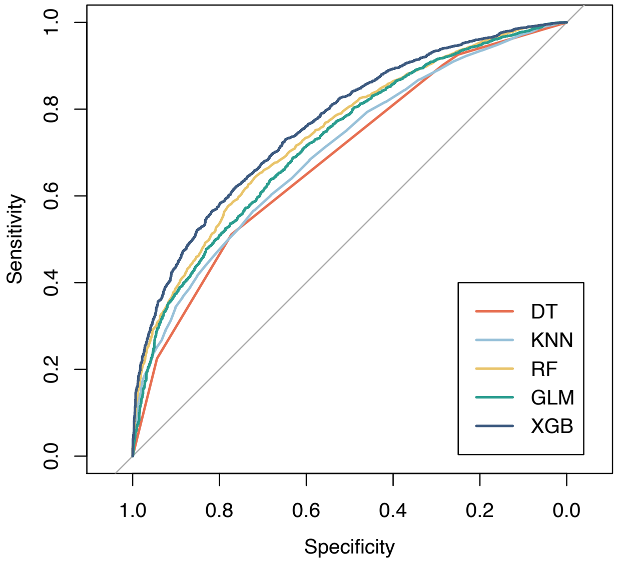


Figure 7 ROC Comparison of Five Classifiers with 10-fold Cross-validation

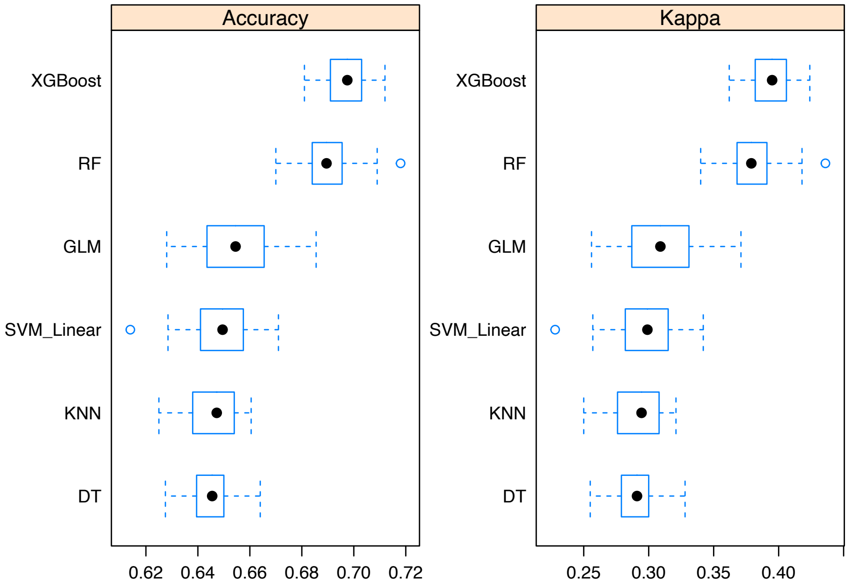


Figure 8 Accuracy Comparison of Six Classifiers with 10-fold Cross-validation

6 Conclusion

In this research, the team applied seven machine learning algorithms to extract the most significant features affecting the severity of car accidents and conducted unbiased and low-variance validation. It is found that the Gradient Boosting Machine provides the most accurate classification, which reaches 68.9%. According to the comparison results, road condition, weather, and TMC code are three most critical features for classifying accident severity. No traffic signal, no crossing, and no station in the nearby location tends to cause a severe car accident. A lower pressure between 29.5 and 30.2 inches, which denotes a stormy weather, is more likely to be associated with high severe accidents. TMC code is another important indicator of severity, with 201 related to a normal accident whereas a severe accident tends to have other TMC codes. Therefore, it can be concluded that the severity of traffic accident can be predicted by some objective condition like road and weather condition.

Data Availability

The datasets and source codes for this study are freely available to the public at kaggle website: https://www.kaggle.com/sobhanmoosavi/us-accidents.

Shiny Application

https://yiyuan-cheng.shinyapps.io/IST707_Final/.

References

[1] Alsagheer, R. H., Alharan, A. F., and Al-Haboobi, A. S. (2017). Popular decision tree algorithms of data mining techniques: A review. *International Journal of Computer Science and Mobile Computing*, *6*(6), 133-142.

[2] Kumbhare, T. A., and Chobe, S. V. (2014). An overview of association rule mining algorithms. International Journal of Computer Science and Information Technologies, 5(1), 927-930.

[3] Maragatham, G., and Lakshmi, M. (2012). A recent review on association rule mining. *Indian Journal of Computer Science and Engineering (IJCSE)*, *2*(6), 831-836.

[4] Kotsiantis, S., and Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71-82.

[5] Denil, M., Matheson, D., and De Freitas, N. (2014, January). Narrowing the gap: Random forests in theory and in practice. In *International conference on machine learning* (pp. 665-673).

[6] Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

[7] Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4), 325-327.

[8] Weinberger, K. Q., Blitzer, J., and Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems* (pp. 1473-1480).

Appendix 1 Team Contribution

	Contributors	**Proportion**
1. Project proposal	Yiyuan Cheng	0.5
	Yingxue Gao	0.5
2. Data preparation	Yiyuan Cheng	0.25
	Yingxue Gao	0.25
	Wanyue Xiao	0.5
3. Data visualization	Wanyue Xiao	1
4. Models	Yiyuan Cheng	0.3
	Yingxue Gao	0.4
	Wanyue Xiao	0.3
5. Poster production	Yingxue Gao	1
6. Poster presentation video clip	Wanyue Xiao	1
7. Shiny App	Yiyuan Cheng	1
8. Final report	Yiyuan Cheng	0.5
	Yingxue Gao	0.2
	Wanyue Xiao	0.3