Severity of Car Accident

Yun Lin, Proud Ao, Yanzhang Chen, Tracy Zhang

Keyword

Random Forest, Neural Network, XGBoost, Text Mining, Logistic Regression, K-Nearest-Neighbor, Data Cleaning, Sentiment Analysis, and NA imputation.

1 Abstract

The goal of this Kaggle project is to predict the severity status of car accidents in U.S using statistical learning models based on the given data, and this report provides a clear and detailed description on how we build our classification model from start to the end, including introduction, exploratory data analysis, data cleaning, feature selection, model construction, analyzing results, discussing limitations and recommendations. The final model is based on random forest and uses 65 variables. It has accuracy score of 0.9435.

2 Introduction

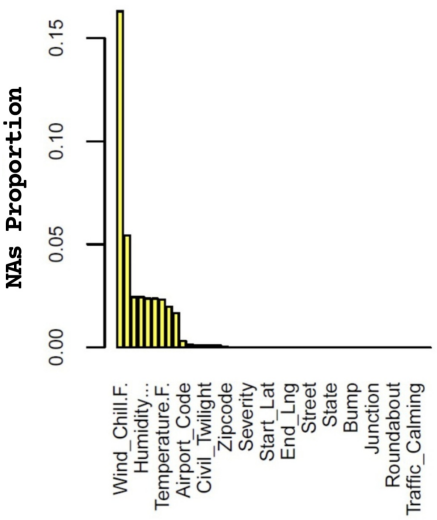
The most common cause of death of teenagers and younger adults is a vehicle accident. Over the past few years, fatalities in vehicle crashes have been declining, but there were still 39,404 fatalities in 2018. US-Accidents can be used for numerous applications such as real-time car accident prediction, studying car accidents hot spot locations, casualty analysis and extracting cause and effect rules to predict car accidents severity, and studying the impact of precipitation or other environmental stimuli on accident occurrence. The most recent release of the dataset can also be useful to study the impact of COVID-19 on traffic behavior and accidents. In order to reduce the severity of the harmful effects on human and property damage, it will be helpful to identify the main causes of traffic accidents. The severity of an accident is not random; it has patterns that can be recognized and avoided. Thus, finding an efficient model to successfully predict car accident severity is significant and urgent.

3 Data Analysis

3.1 Exploratory Data Analysis

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to Dec 2021, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. There are about 50,000 accident records in this dataset.

Our train data includes 35,000 observations and 43 predictors–11 numerical predictors and 32 categorical predictors. Test data includes 15,000 observations. Car severity is our response vector which includes 90% "MILD" accidents and 10% "SEVERE" accidents. Thus, we need a model that has prediction accuracy rate higher than 90%.
The graph below shown the proportion of NAs for each features:



3.2 Data cleaning and imputations

After a quick examination of NAs in data using **MICE** package, we found a total of 13211 missing values which accounts for 0.86% of total cells.
For Airport codes, we decided that it's not an important predictor because the airport codes only report the location of the weather station but we alraedy have the weather condition provided in other predictors.
For Zipcode, we used starting longitudes and latitudes with **KNNImputer** of $k = 1$ to impute the missing zipcodes with the first nearest neighbor.

Since not all zipcode were provided the Full 9-Digit ZIP, we decide to remove all the ZIP+4 Code for unifying the zipcode. For the rest of the predictors, we used the **MICE** package to deal with NA, and it has reduced the NA values to an insignificant amount. For models that can not compute NAs we used **IterativeImputer** package to impute the NA values.

3.3 Variable Generation and Text Mining

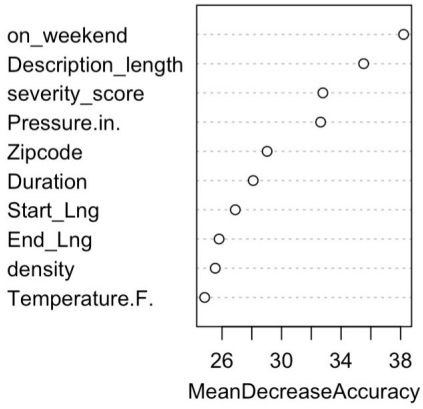
We can synthesize new information from existing predictors. For example, the variable Description contains summary report of the accident. Based on Description, we generated new variables such as description length, calculating from word count of the description.

We further explored description using **text mining**. Borrowing the idea of **sentiment analysis**, we first filtered out the important words and calculated frequencies for words from both the mild and severe accidents. We then assign weights to each word according their frequency and whether the words are in Mild or Severe accidents. Using our Mild Severe Dictionary, we were able to create a severity score for each description. Severity Score turned out to be important when we selected important predictors.
We also generated **Duration** and **onWeekend** variable and month and year factors from start and end timestamp. Furthermore, we generated **Speed** from Duration and distance.mi.

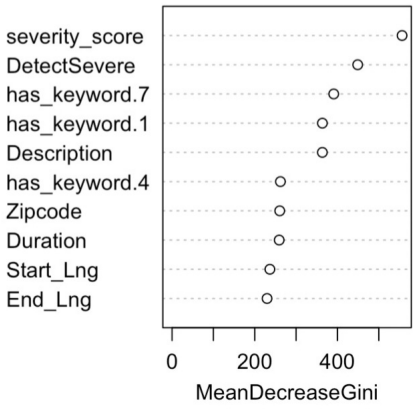
4 Model selection

4.1 Important Variables

We ran **random forest** to decide the important variables and the top 10 important variables by mean decrease in accuracy and **Gini index** after removal are shown below:



Based on Mean Decrease Accuracy, we can tell onWeekend is out weight the other variable, and based on Mean Decrease Gini score



4.2 Model Section and Comparison

We then ran various models on most important variables and our ranked accuracy rate are shown below:

Models	Score
Logistic Regression	0.93288
K-Nearest-Neighbors	0.9328
XGBoost	0.93484
Artificial Neural Networks (ANN)	0.92897
Random Forest	0.94373

For all the models, we trained through different combinations of hyperparameters within reasonable range using **cross validation** to find the optimal parameters for each models respectively. As the result we had, The best models we had on Kaggle public score was

random forest model. The **tuning parameters** of it for each tree was max tree depth 21 with 8 features.

5 Conclusion

Our final random forest model achieved an accuracy rate of 94.373 %. In the Kaggle contest, we achieved at **final rank 6** out of 40 teams. The most important predictors included reports of car accidents, which are listed under description variable. We found that when the report contains certain keywords, for example, "road closed," the car accident is severe most of the times. The location of accident is equally important. For example, though most car accidents took place in California, it had the lowest severity rate compared to other states. Besides, on weekend and at certain months, severe accidents tend to happen more frequently. Bad weather also affects severity a lot. For example, on rainy, stormy, and thunder days, cars tend to have worse crashes.

5.1 Future Improvement

We have much room of improvement for our data cleaning phase. First of all, we chose 2022 US census data to generate the population and population density variable based on Zipcode. However, observed car accidents only took place from 2016 to 2021. Thus it is better if we attached US census from the correct years. Secondly, we assumed all timezone to be UTC time, but there is a variable named timezone according to which we could have more precise datetime conversion. Thirdly, we had a categorical predictor called isNight that takes all time between 6pm to 6am to be night. However, to have more precise cutoffs of sunrise and sunset, we could have scraped online data, which will make isNight a much better predictor.
Secondly, importance of variable generated by Random Forest is largely affected by **multicollinearity** and also more biased for categorical variables with a large number of levels. We could have reduced multicollinearity in our data to generate better importance graphs. Reducing number of predictors, on the other hand, will also allow more interpretability on the model and reduce the risk of over fitting.
Thirdly, our model uses a max depth of 21 for each decision tree generated in the random forest. That means we could have possibly generated a tree with 2^{21} leaves at maximum. High complexity of model tends to overfit. given more time, we will be able to evaluate the degree to which we overfit.

References

Predicting Car Accidents Severity. Kaggle, www.kaggle.com/competitions/predicting-car-accidents-severity/overview.

US Census Data. Simple Maps, simplemaps.com/data/us-zips.