Classification of Car Accident Severity in Virginia

Yisha He
University of Virginia
Department of Statistics

Data Overview

⋆ Data covers car accidents in Virginia from February 2016 to December 2019.

⋆ Sourced from published literature [1] [2].

⋆ 79,957 observations with 22 variables.

⋆ 74,799 observations after data cleaning.

Name	Description	Name	Description
Severity	Car Accident Severity. (2 levels: Severe/Not severe)	TMC	Traffic Message Channel (TMC) code that describes the accident. 14 levels
Location	GPS coordinate of the accident start point. (Latitude, Longitude)	Distance	Length of the road extent affected by the accident (in miles).
Time	Recorded Time of the accident.	Side	Relative side of the street in address field. (R/L)
Weekday	Date of week that accident happens. (7 levels: Mon, Tue, Wed, Thu, Fri, Sat, Sun)	Signal	Presence of traffic signal in a nearby location. (1/0)
Duration	Duration of the accident (in minutes).	Junction	Presence of junction in a nearby location. (1/0)
Weather Condition	54 levels: Clear, Cloudy, Fog, Rain, Heavy Rain, T-Storm, Snow, Light Snow, Overcast ...	Bump	Presence of speed bump in a nearby location. (1/0)
Temperature	Temperature. (in Fahrenheit)	Railway	Presence of railway in a nearby location. (1/0)
Humidity	Humidity. (in percentage).	Roundabout	Presence of roundabout in a nearby location. (1/0)
Pressure	Air pressure. (in inches).	Station	Presence of station in a nearby location. (1/0)
Wind Speed	Wind speed. (in miles per hour).	Stop	Presence of stop sign in a nearby location. (1/0)
Civil Twilight	Period of day based on civil twilight. (day/night)	Turning Loop	Presence of turning loop in a nearby location. (1/0)

Motivation

Motivation:

⋆ Traffic accidents are concerns for public safety.

⋆ Classifying severe accidents may help regulators and traffic departments allocate resources to release traffic pressure.

⋆ Help reduce future traffic accidents.

Questions:

⋆ Can the characteristics of a car accident predict its severity level?

⋆ If so, what are the main factors that cause a severe accident?

Data Manipulation

⋆ Duration: highly skewed to the right, so logarithms are taken.

⋆ Weather condition: similar levels are merged, ending in 10 levels.

⋆ Weekday: Monday to Friday become "weekday"; Saturday and Sunday become "weekends".

⋆ Time: Accident time is divided into four variables: Year, Month, Day, Hour.

⋆ Remove variables with near zero variances: Junction, Bump, Railway, Roundabout, Station, Stop, Turning Loop.

Logistic Classification

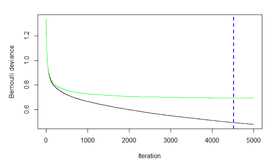
⋆ Binomial logistic regression selected by AIC criterion:

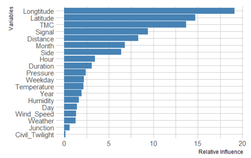
$$Severity \sim TMC + Location + Distance + Side + Temperature + Humidity + Pressure + WindSpeed + WeatherCondition + Junction + Signal + CivilTwilight + Year + Month + Hour + Weekday + log(Duration)$$

Accuracy %	False Positive %	False Negative %	Sensitivity %	Specificity
72.3%	13.4%	14.3%	73.2%	71.2%

⋆ Binomial logistic regression with Lasso penalty selects the same model as using AIC criterion.

⋆ Boosting the logistic model. Parameter ntrees is selected to be 4500 using 3-fold cross validation. Classification performance improves after boosting the logistic model.





Accuracy %	False Positive %	False Negative %	Sensitivity %	Specificity
85.6%	6.8%	7.7%	85.7%	85.4%

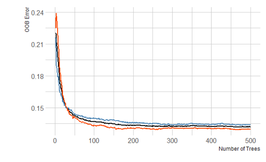
Tree Based Classification

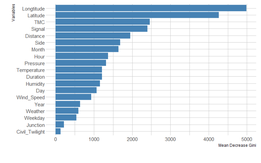
⋆ Decision tree model has low specificity rate.

Accuracy %	False Positive %	False Negative %	Sensitivity %	Specificity
72.9%	16.2%	10.9%	79.6%	65.2%

⋆ Random forest model has the highest prediction accuracy among all models applied.

⋆ According to mean decrease in Gini index, location, TMC, traffic signal, road side and road distance are the main variables that classify accident severity.





Accuracy %	False Positive %	False Negative %	Sensitivity %	Specificity
86.7%	5.9%	7.4%	86.2%	87.3%

Exploratory Analysis

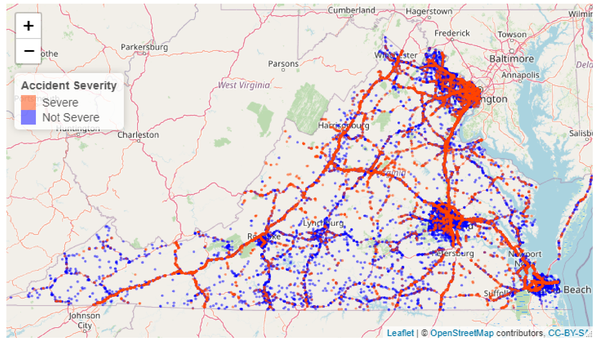


Figure 1: Car Accidents Heat Map in Virginia. Washington D.C., Richmond and Virginia Beach have the highest density of traffic accidents in Virginia.

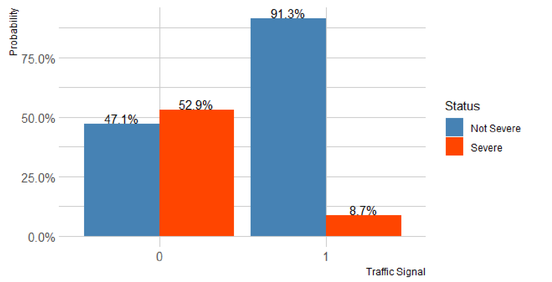


Figure 2: Distribution of Accident Severity by Traffic Signal. Severe accidents are more likely to happen far from traffic signals.

Figure 3: Counts of Accident by Month. Most accidents occur after August and October has the highest accident rate in a year.

Figure 4: Density of Accident Severity by Hour of Day. Compared to minor accidents, severe accidents occur less in the morning but more in the afternoon.

Discussion

Summary

⋆ Random Forest model and boosted logistic model perform well in classifying accident severity.

⋆ Variables about location, road condition and time are important in classifying accident severity, while weather variables are less important.

Limitations

⋆ Model may not work in real world. There may have more levels of accident severity.

⋆ Other algorithms may perform better: SVM, Adaboost, Neural Network ...

⋆ Correlations are discovered but causal relationships remain unclear.

References

[1] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

[2] 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, ACM, 2019.