Identifying the Best Statistical Machine Learning Model for Predicting Road Fatalities in the United States

Varun Subramaniam, PUBH 6886, Final Project

GW | Public Health

ABSTRACT

- This investigation sought to identify the best statistical models for predicting road fatalities in the United States.
- I first compared the 10-Fold Cross-Validated Root-Mean-Squared Errors and R^2 values for 7 linear models predicting the number of fatalities per crash in a 2010-2014 dataset. I then compared the 10-Fold CV accuracy of predictions of fatal injury status from 3 non-linear models using several predictors from a very similar 2018 road fatalities dataset. The non-linear models presented better fits with fatality than did the linear models.
- The optimal model for predicting fatal injuries is a random forest with 5 predictors per split. This model finds blood alcohol content, passenger status, age, and no airbag deployment to be the top four predictors of fatal injury following a crash.

INTRODUCTION AND PROBLEM STATEMENT

- In this investigation, I applied statistical modeling to two existing datasets on road fatalities (described in detail in the Data Description section).
- According to the National Highway Traffic Safety Administration (NHTSA), road traffic accidents are a leading cause of death in the United States, with an estimated 42,915 fatalities in 2021 alone. In addition to the human toll, these accidents also have significant economic costs, with total financial losses due to fatal motor vehicle crashes in the United States estimated to exceed $340 billion annually.[1]
- Studies have identified several factors that contribute to road traffic fatalities in the United States, including impaired driving, adverse weather conditions, and speeding among many others.[2] However, existing research has mainly focused on identifying these factors, rather than using statistics to determine those that *most* exacerbate fatalities.
- I primarily sought to address a pressing public health question: **what are the top predictors of road fatalities from motor vehicle crashes within the United States?**

DATA DESCRIPTION

- The NHTSA collects road accident data from various sources: primarily, reports and form-submissions from first-responders. Each year, these data are compiled into a single dataset, known as the Fatality Analysis Reporting System, or FARS for short.[3]
- The first dataset I used was the 2014 FARS (henceforth, FARS14), available publicly online from the "Downloadable Data" page of the Data Analysis and Social Inquiry Lab (DASIL), housed at Grinnell College. FARS14 reports a range of information on every reported crash across the U.S. over a four-year timeframe.[3] Of the 58 variables in FARS14, I chose "Number of Fatalities" as my numeric response variable, and 25 others as candidate predictors. Information on each chosen variable in FARS14 is provided here.
- The second dataset I used was the updated 2018 FARS (FARS18), available directly from the NHTSA as "person.csv" under "FARS2018NationalCSV.zip." FARS18 reports a range of information about the persons involved in a crash.[4] I chose "Severity of Injury" (which includes fatal injury as a category) as my categorical response variable and 7 others as candidate predictors. Information on each chosen variables in FARS18 is provided here.
- Given the vast size of both datasets, I applied comprehensive cleaning protocols to each to maximize completeness. I removed any row with unknown, unreported, and missing data across *all* selected variables. After cleaning, FARS14 contained 37,702 observations across all 26 variables and FARS18 contained 29,942 observations across all 8 variables.

METHODS

Part I: Testing Predictive Accuracy of 7 Linear Models Using FARS14

- I first tested the accuracy of 7 different linear models in predicting the continuous number of fatalities from road accidents in FARS14 based on different predictor combinations. For each linear model, I calculated both 10-Fold Cross Validation (CV) Root-Mean-Squared Error (RMSE) and R-Squared values using the integrated *results* output from the *train* function in the *caret* R Package. I used these measures to assess the predictive performance of each linear model.
- The first model I generated was a **simple linear regression via least-squares estimation.** I also generated **Forward-** and **Backward-Stepwise** Models, as well as **Principal Component** and **Partial Least-Squares Regressions** (PCR and PLS respectively). For each of these four models, the number of predictors minimizing 10-Fold CV RMSE was taken as the optimal tuning parameter. I further generated **Ridge** and **LASSO** models, choosing optimum lambda values once again as those that minimized 10-Fold CV RMSE.
- 10-Fold CV R^2 values indicated that all 7 linear models performed poorly in predicting number of fatalities (see Results section).

Part II: Testing Predictive Accuracy of 3 Non-Linear, Tree-Based Models Using FARS18

- Seeing the overwhelming failure of several linear models in predicting number of fatalities, I turned to the FARS18 dataset, which offered more granularity in measuring human impacts of road accidents. FARS18 still offered fatality data but included these observations under the ordinal umbrella category of injuries: an especially promising feature for tree-based models.
- Once again using the integrated output of the *train()* function, I computed the 10-Fold CV Accuracy values of 3 non-linear models in predicting discrete injury status based on my 7 selected predictors in FARS18. First, I created training and testing datasets, composed of 67% and 33% of all observations in FARS18, respectively. I then used these variables to grow a **basic classification tree**, with *minsplit* set to 2000 and selected the optimal classification tree as that containing the alpha value maximizing 10-Fold CV Accuracy within 1-Standard Error. Next, I generated a **boosted classification tree** using the same training and testing datasets, increasing *n.trees* from 5 to 50 in increments of 5 and *shrinkage* from 0.20 to 0.45 in increments of 0.05. I set *interaction.depth* at 1:6 and *n.minobsinnode* at 1,000. The optimal boosted classification tree included the four parameters that maximized 10-Fold CV Accuracy. Lastly, I generated a **random forest** model with *ntree = 500* and *mtry = 1:7.* The number of predictors considered at each split that maximized 10-Fold CV Accuracy were used in the final random forest model.
- The non-linear model with the highest 10-Fold CV Accuracy was used to create a variable importance plot. Mean Decrease in Gini (MDG) Index, stored within the output of the *train()* function, was plotted for each candidate predictor, with the highest MDG values corresponding to the top predictors of fatal injury status in FARS18

RESULTS

Fig 1. Plot of Lambda vs 10-Fold CV RMSE for choosing tuning parameter for LASSO Regression (the optimal linear model) to predict number of fatalities in FARS14.

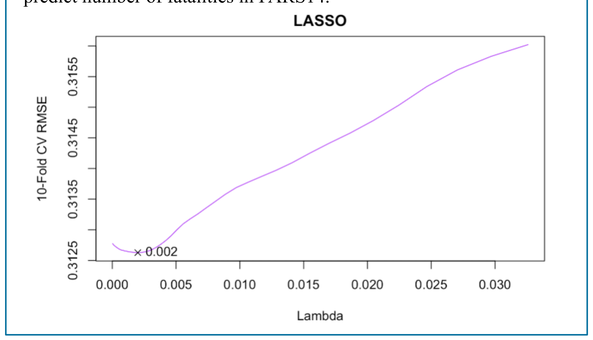


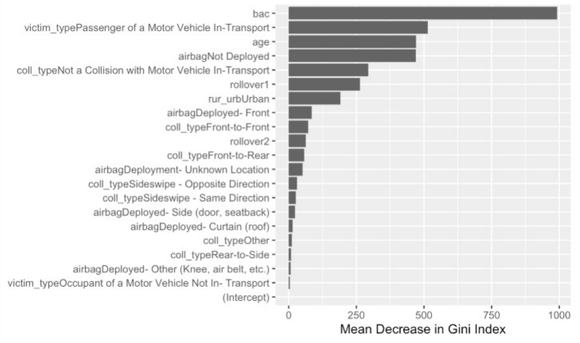
Fig 2. 10-Fold CV RMSE and R^2 values for each linear model predicting number of fatalities based on various predictors in FARS14.

Linear Model	Tuning Parameters	10-Fold CV RMSE	10-Fold CV R-Squared
Linear Regression	NA	0.31280	0.03048
Forward-Stepwise	14 Predictors	0.31356	0.02294
Backward-Stepwise	10 Predictors	0.31349	0.02282
PCR	20 Predictors	0.31271	0.03087
PLS	25 Predictors	0.31279	0.03072
Ridge	Lambda = 0.030	0.31267	0.03010
LASSO	Lambda = 0.002	0.31263	0.03077

Fig 3. 10-Fold CV Accuracy values for each non-linear model predicting injury status based on 7 predictors in FARS18.

Non-Linear Model	Tuning Parameters	10-Fold CV Accuracy
Basic Classification Tree	Alpha = 0.01	0.66675
Boosted Classification Tree	Shrinkage = 0.45 Interaction Depth = 5 Min. Obs. In Node = 1000 N Trees = 50	0.69675
Random Forest	5 Predictors per Split	0.70581

Fig 4. Variable importance plot showing top predictors (by MDG) of fatal injury in optimal random forest model for FARS18.



SYNTHESIS AND DISCUSSION

- The FARS14 dataset, though useful, did not lend itself to statistical predictions of road fatalities, as the vast majority (97%) of crashes corresponded to 1 fatality. In other words, varying combinations of predictors did not seem to affect the response variable (number of fatalities in each crash) to a significant degree. This is corroborated by **Fig 2.**, which shows extremely low 10-Fold CV R^2 values across all seven linear models (even for LASSO, which was found to be the best linear model).
- All three of the non-linear models proved far more accurate in predicting fatal injury status, compared to the fit of any linear model in predicting raw number of fatalities. Further works could also assess the accuracy of Naïve Bayes and KNN Regression models.
- As shown in **Fig 3.**, **Random Forest with 5 predictors per split maximized 10-Fold CV Accuracy at 0.706**. This specific model best predicted fatal injuries based on the 7 selected predictors in FARS18.
- **Fig 3.** shows that the accuracy was very similar for the boosted classification tree and random forest models. The difference is likely due to my chosen parameters. However, both models outperform the basic classification tree by over 3%.
- Vadhwani and Thakor also used the 2018 FARS release to model various predictions of fatal injury. They found an optimized XGBoost model to be best; however, they also noted the high accuracy for random forest models (among others).[5] My study only assessed 3 non-linear models and could be improved by incorporating several other candidate models for predicting fatal injury status.
- **Fig 4.** shows that blood alcohol content, passenger in-transport status, age, and non-deployment status for airbags are the Top 4 predictors of fatal injuries from accidents. While BAC, age, and airbag-failure are consistent with well-established trends in road fatalities, the predictive role of passenger status on fatal injury appears to be a novel finding.[2] More research is required to determine why passengers might have higher odds of encountering fatal accidents than drivers or other vehicular passengers. Nevertheless, it appears that the random forest model overall could be applied to real-world situations, as it correctly identifies several known exacerbators of road fatality as top predictors.
- This analysis shows that the FARS18 dataset is far better for predictive modeling, thanks to its more granular measurements of injuries. This dataset still allows for assessments of fatality without having to predict the exact *number* of deaths (which is almost always 1 in FARS14).

REFERENCES

1: National Center for Statistics and Analysis (2022) Early Estimates of Motor Vehicle Traffic Fatalities And Fatality Rate by Sub-Categories in 2021 NHTSA National Center for Statistics and Analysis, Washington, D.C.

2: Rolison,J.J. *et al.* (2018) What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. *Accid. Anal. Prev.*, **115**, 11–24.

3: National Center for Statistics and Analysis (2016) 2014 FARS/NASS GES Coding and Validation Manual National Highway Traffic Safety Administration, Washington, D.C.

4: National Center for Statistics and Analysis (2020) 2018 FARS Coding/Validation Manual. National Highway Traffic Safety Administration, Washington, D.C.

5: Vadhwani,D. and Thakor,D. (2022) Predictive analysis of injury severity of person across angle crashes using machine learning models. *Int. J. Crashworthiness*, **0**, 1–14.

Watch my YouTube Video HERE!

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC