Summer 2024

Predicting Road Accident Injury Severity for Drivers in Automobile Crashes in United States Using Machine Learning Models and AI

Emil Agbemade
University of Central Florida, emil.agbemade@ucf.edu

Benedict Kongyir
Oklahoma State University, bkongyi@okstate.edu

Predicting Road Accident Injury Severity for Drivers in Automobile Crashes in United States Using Machine Learning Models and AI

By:
Benedict Kongyir, Oklahoma State University
(bkongyi@okstate.edu)
Emil Agbemade, University of Central Florida
(emil.agbemade@ucf.edu)

July 8, 2024

Abstract

This study analyzes data from the National Highway Traffic Safety Administration's 2021 Crash Report Sampling System to identify key factors contributing to the severity of injuries in car accidents. By utilizing various machine learning algorithms and cross-validation techniques, we assessed metrics such as accuracy, sensitivity, precision, specificity, and the area under the curve (AUC) to evaluate the effectiveness of predictive models. All data preprocessing and model building was done using KNIME Analytical software [9]. Our findings reveal significant correlations between certain variables such as airbag injection, weather conditions, intoxication, vehicle state, driver distractions, and injury severity. These insights underscore the importance of stringent safety measures, including proper restraint system usage and advanced driver-assistance technologies, in reducing the risk of severe injuries in car accidents. Recommendations for policy enhancements and preventive measures are discussed to improve overall vehicle safety.

Introduction

Car accidents remain a significant public health concern, leading to numerous injuries and fatalities each year. Understanding the factors that contribute to the severity of injuries in these incidents is crucial for developing effective safety regulations and policies. This report presents the results of a data analytics effort aimed at identifying the primary elements that most significantly influence injury severity in car accidents.

The dataset used in this study includes variables such as date, time of day, weather condi-

tions, collision details, intoxication levels, vehicle state, and driver distractions. By analyzing these variables through several machine learning algorithms, we aimed to uncover trends, correlations, and insights that could inform the enhancement of car safety regulations and policies.

Data for this study was sourced from the National Highway Traffic Safety Administration's Crash Report Sampling System for the year 2021, comprising detailed information from police reports on each collision. Our analysis highlights several key factors that significantly increase the severity of injuries sustained in car accidents, demonstrating a strong predictive correlation with injury severity.

The importance of these factors underscores the critical need for proper safety measures and precautions during driving. Effective utilization of restraint systems, adequate handling of airbags, and measures to prevent ejection and fire involvement are essential for minimizing the severity of injuries in car accidents. Our findings suggest that improving vehicle safety precautions is a vital step. This could involve enhancing the design and implementation of safety belts, promoting public education on the proper use of airbags, and utilizing advanced materials and technologies to reduce the risk of fire in collisions.

Additionally, the importance of these variables points to the necessity of focused preventive measures. Legislators are advised to impose more stringent speed limits, vigorously enforce anti-drunk driving statutes, and support sophisticated driver-assistance technologies to warn motorists of risky pre-crash actions. These actions target the situational and behavioral factors that lead to catastrophic accident outcomes.

In the following sections of this report, we will delve deeper into how these findings impact business applications, the meaning and processing of the data, the models employed, and the insights derived from them. We will also discuss the application of the aforementioned recommendations in greater detail.

Data Preparation and Data Understanding

The data for this project was sourced from the National Highway Traffic Safety Administration's (NHTSA) Crash Report Sampling System (CRSS) and covers police-reported crashes from 2016 to 2021. The datasets used include:

- Accident dataset: 46 variables such as Manner of Collision, time of the crash, Atmospheric Conditions, and work zone, with 54,200 observations.
- Vehicle dataset: 88 variables including Vehicle Identification number, model year, Travel speed, and area of impact, with 95,785 observations.
- Person dataset: 59 variables such as Age, sex, injury severity, seating position, airbag, and Police Reported Alcohol Involvement, with 133,734 observations.
- Distract dataset: Variables include Distraction and weight.

The combined dataset consisted of 204 variables and 128,393 observations. All variables were originally recorded as numeric, including the target variable, injury severity. Data

preprocessing involved cleaning the datasets and converting certain numeric variables into nominal categories to facilitate analysis. This preprocessing step ensured that the data was suitable for subsequent machine learning modeling and analysis.

Table 1: Showing Final Selected Variables for the Models Building

No.	Variable	Description	Data Type
1	BODY_TYPE	NCSA Body Type	Nominal
2	Region_Str	Region of country	Nominal
3	DAY_WEEK_Str	Day of week	Nominal
4	MONTH_binned	Month of year	Nominal
5	URBANICITY_Str	Urban or rural area	Nominal
6	HOUR_MISS	Hour at which crash occurred	Numeric
7	MAN_COLL_MISS	Manner of collision	Nominal
8	RELJCT1_MISS	Interchange area	Binary
9	CRASH_LOC	Crash location	Nominal
10	WORKZONE?	Work Zone	Binary
11	LIGHTING	Light condition	Nominal
12	WEATHER_CONDITIONS	Weather condition	Nominal
13	MOD_YEAR_MISS	Vehicle model year	Numeric
14	VEH_AGE	Vehicle age	Numeric
15	Overspeeding?	Measure of over speeding	Numeric
16	HIT_RUN_NEW	Hit & run occurred	Binary
17	JACKNIFE	Vehicle experienced jackknife	Binary
18	Under/Override	Under/override with another vehicle	Nominal
19	Rollover	Vehicle rolled over	Binary
20	DAMAGE_EXTENT	Severity of car damage	Nominal
21	FIRE	Fire present	Binary
22	Drive_Alcohol	Alcohol involved	Binary
23	SURFACE_COND	Road surface condition	Nominal
24	PreCrashScenario	Scenario before crash	Nominal
25	PresCrashStatbility	Stability of car before crash	Nominal
26	ACC_TYPE_NEW	Type of crash	Nominal
27	AGE_MISS	Age of the driver	Numeric
28	Airbag_Deployment	Airbag deployed	Binary
29	Ejection_Degree	Degree of ejection	Nominal
30	RESTRAINT_USE	Restraint equipment used by occupant	Nominal
31	AlcTes_Result	Alcohol test result	Numeric
32	DRUG_INVOLVEMENT	Drugs involved	Binary
33	Hospital_Transport	Hospital transport used	Nominal
34	Distraction_Cause	Distraction type	Nominal
35	INJ_SEV_Binned (DV)	Severity of injury	Nominal

Methods

For all our models in our study, we have implemented k-fold cross-validation with a value of $k = 10$. The k-fold cross-validation technique involves dividing the entire dataset into k equally sized subsets or folds. The model is then trained on $k - 1$ of these folds, while the remaining fold is used as a test set to evaluate performance. This process is repeated k times, with each fold serving as the test set once, allowing the model to be tested across all available data. The results from each fold are then averaged to provide a comprehensive measure of the model's performance. This method helps assess a model's effectiveness and minimize bias and variance, ensuring it generalizes well to new, unseen data [10].

Variables Importance

Variable importance is an essential component in machine learning as it is often used to identify the most contributing factors to the model. It is also sometimes used for feature selection. It is used to reduce the number of predictors in the model by eliminating fewer contributing factors.

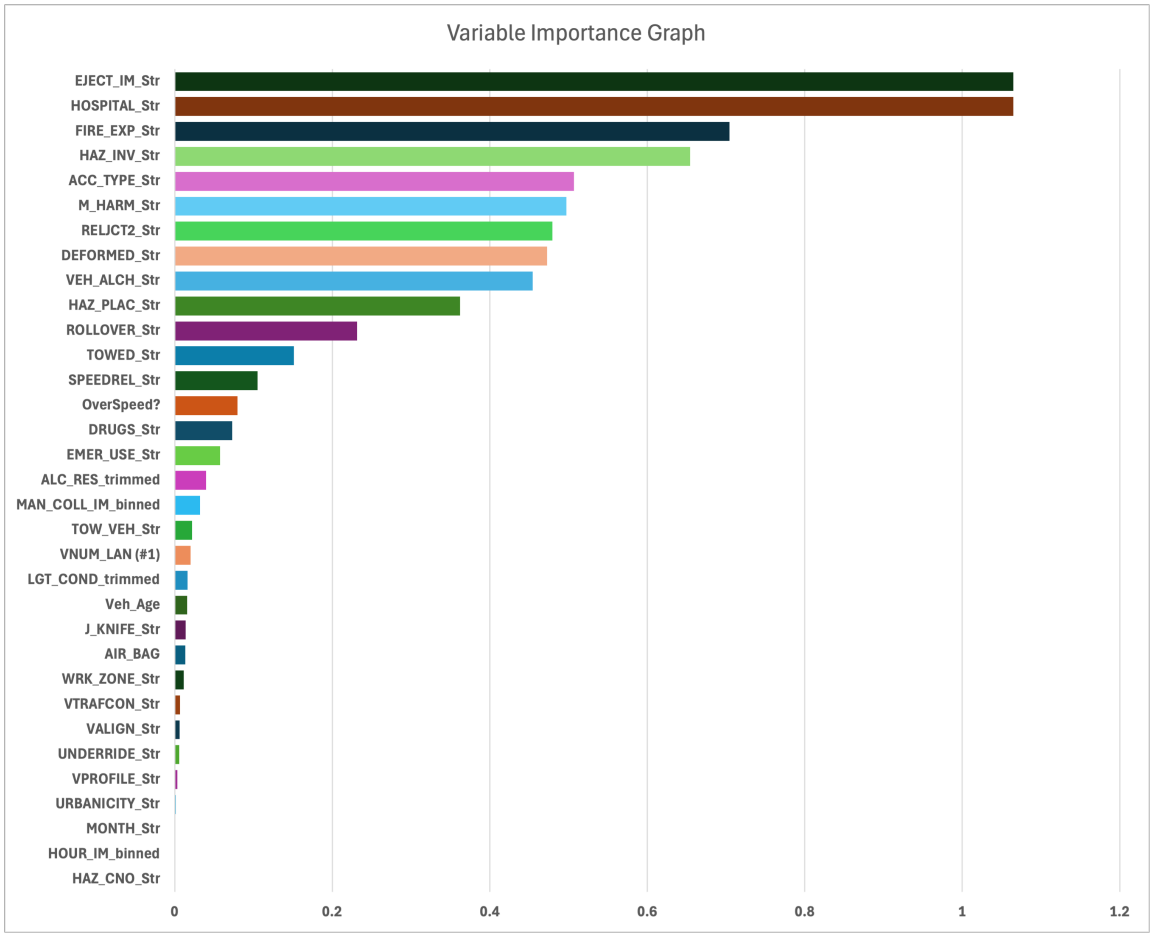

Figure 1: Variable importance of selected predictors

Model Building

Logistic Regression Model

Logistic regression is a binary probabilistic classification model that uses supervised learning to specify an event's outcome as belonging to one class or the other. It is a special regression model with a binary outcome variable. Logistic regression gains its popularity in predictive modeling through its ability to transform the results of a linear regression model into a number between 0 and 1 using the link function [8, 13].

Decision Tree Model

A decision tree is a classification tool that classifies data into a finite number of classes based on the values of the input variables. Decision tree models owe their popularity to easy interpretability. They are easy to explain to non-experts and often perform better with more categorical predictors. Decision tree models can sometimes be used for dimension reduction, that is, they can help identify the most important predictors and discard non-significant predictors [14, 3]. About 64% of all accident cases requiring transportation of victims to hospital are high-injury cases, and only about 19% of accidents that do not require transportation of victims to hospital are high-injury cases.

Random Forest

A random forest is an ensemble of several decision trees. Thus, it can produce a high prediction accuracy rate most of the time. However, it is often criticized for lack of interpretability. It has a wide range of applications in finance, healthcare, security, insurance, etc. [2, 11].

Neural Network Model

Artificial Neural Network (ANN) models were developed to mimic human thought processes. It is one of the most popular predictive models in recent years due to its ability to perform very well in most situations. A principal component in ANN is the perceptron. The perceptron is a mathematical structure that picks predictors, forms a linear combination of those predictors, and uses the activity function to obtain an initial result. The activation function then transforms the output of the activity function into a binary output that can be interpreted by the computer [1, 7].

Gradient Boosting Model

Gradient boosting trees, like random forests, is an ensemble of several decision trees. The main difference between the random forest and the gradient-boosting tree model lies in their training process. For random forests, the trees are independently trained on separate subsets of predictors and observations, whereas with gradient-boosting trees, the trees are trained sequentially using the residuals of the previous tree to improve the present model. Based on

this, gradient-boosting trees have the potential to perform better than random forests, but not always [6, 4].

Testing and Evaluation

The following performance metrics were employed to evaluate and select the best model for implementation: overall model accuracy rate, sensitivity, precision, and the Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) Curve.

Sensitivity (True Positive Rate or Recall)

Sensitivity measures the proportion of actual positive cases that are correctly identified by the model. High sensitivity means fewer actual positives are missed (low false negatives). It is calculated as:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Specificity (True Negative Rate)

Specificity measures the proportion of actual negative cases that are correctly identified by the model. High specificity means fewer actual negatives are misclassified as positives (low false positives). It is calculated as:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Precision (Positive Predictive Value)

Precision measures the proportion of positive predictions that are actually correct. High precision means the model's positive predictions are reliable. It is calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

The ROC & the AUC Curve

The Receiver Operating Characteristic (ROC) curve is a graph that plots the true positive prediction rate against the false positive prediction rate. The Area Under Curve (AUC) measures how well the model can distinguish between classes [12]. It is one of the most popular metrics often used to evaluate and compare the performance of classification models[5].

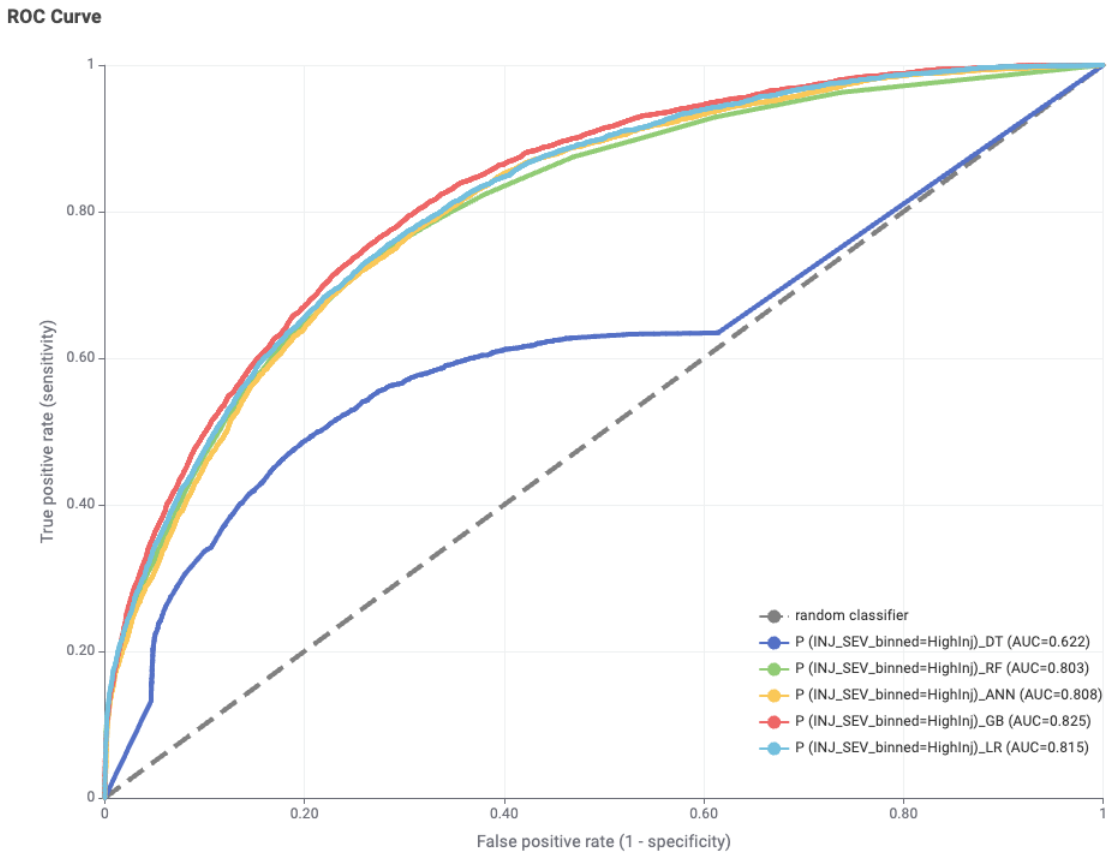

Figure 2: Showing Receiver Operating Characteristic(ROC) and the Area Under Curve (AUC)

Accuracy, Sensitivity, and Precision

Model	**Accuracy(%)**	**Sensitivity(%)**	**Precision(%)**	**AUC(%)**
Decision Tree (DT)	77.7	86.7	85.7	62.2
Random Forest	83.6	98.4	84.1	80.3
Gradient Boosting Trees	84.1	96.9	85.4	82.5
Artificial Neural Network	83.1	95.8	85.2	80.8
Logistic Regression	83.8	96.0	85.1	81.5

Table 2: Comparing Models Using Accuracy Rates, Precision, Sensitivity, Specificity and AUC

Conclusion and Recommendations

Model Deployment

Our analysis identified significant risk factors in auto accidents that could be effectively addressed through preventive measures and technological advancements. These include enforcing stricter driving evaluation and education legislation to combat issues like seatbelt non-usage, drug presence, improper airbag deployment, and accidents in work zones. Implementing regular driver reexaminations could help maintain driving skills and road etiquette, especially for older drivers. Furthermore, leveraging advancements in automotive technology, such as smart cars with automated safety features, holds promise in mitigating human error and enhancing overall driver safety.

Recommendations

- **Legislative Reforms:**Advocate for legislative changes to mandate periodic driver reevaluations to ensure ongoing competency and adherence to safety protocols.
- **Technological Integration:** Encourage the integration of advanced safety technologies in vehicles, such as automated braking systems and intelligent driver assistance features, to prevent accidents and reduce injury severity.
- **Educational Initiatives:** Promote educational campaigns that raise awareness about safe driving practices, leveraging insights from our study to emphasize behaviors that minimize injury risks.
- **Industry Innovation:** Collaborate with automotive manufacturers to design and implement safer vehicle systems based on our findings, prioritizing features that address critical injury factors identified in our analysis.
- **Emergency Response Enhancement:** Support emergency response organizations in adopting data-driven strategies to improve resource allocation and emergency medical care for accident victims.

In summary, our study underscores the critical role of data-driven analysis in understanding and mitigating the severity of injuries in auto accidents. From the model assessment metrics above, the best model is the Gradient Boosting Trees with an accuracy rate of 84.1%, sensitivity of 96.9%, precision of 85.4%, and AUC of 82.5%. By deploying the best model among the five models, we identified key variables influencing injury severity, providing actionable insights for policymakers, technologists, educators, and emergency responders to collectively enhance road safety. Embracing these recommendations can significantly contribute to reducing the human and economic toll of auto accidents, fostering a safer driving environment for all.

References

[1] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[2] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.
[3] Leo Breiman et al. *Classification and Regression Trees*. Wadsworth International Group, 1984.
[4] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 785–794.
[5] S. Dakurah et al. "A Model for Pricing Insurance Using Options". In: *Journal of Research in Business, Economics and Management* 10.3 (2018), pp. 1971–1988. URL: https://scitecresearch.com/journals/index.php/jrbem/article/view/1440.
[6] Jerome H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232.
[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
[8] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, 2000.
[9] KNIME AG. *KNIME Analytics Platform*. 2024. URL: https://www.knime.com/.
[10] Benedict Kongyir and Emil Agbemade. "Bootstrap Regression for Investigating Macroeconomics Factors Affecting USA Home Prices". In: *Data Science and Data Mining* 20 (2024). URL: https://stars.library.ucf.edu/data-science-mining/20.
[11] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22.
[12] Sarang Narkhede. *Understanding AUC - ROC Curve*. Medium, June 2018. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.
[13] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M. Ingersoll. "An Introduction to Logistic Regression Analysis and Reporting". In: *The Journal of Educational Research* 96.1 (2002), pp. 3–14.
[14] J. Ross Quinlan. "Induction of Decision Trees". In: *Machine Learning* 1.1 (1986), pp. 81–106.