Predictive Analysis of High-Damage Vehicle Crashes: investigating Key contributing factors

Ali Gorji Sedfidmazgi

Abstract

This research primarily investigates the factors most contributing to high-damage vehicle crashes, which are over 1,500 $. The focus is on a range of factors including traffic control devices, environmental conditions, lighting, road conditions, and crash types. The study's core lies in its use of SHAP (SHapley Additive exPlanations) values to delve into the contributions of each variable toward predicting high-damage crashes. The findings reveal that certain traffic control devices and device condition are strongly associated with higher damage. By identifying key contributer of high-damage crashes, the study provides valuable insights for enhancing road safety and minimizing the financial and societal impacts of vehicle crashes.

KEYWORDS: CRASH, DAMAGE, CONTRIBUTION, SHAP VALUES

Question:

In the realm of traffic safety, understanding the factors that most significantly indicate high-damage crashes, specifically those with damages exceeding $1,500, is crucial. Research indicates that a confluence of environmental, infrastructural, and situational elements plays a pivotal role. The study aims to identify key contributors to high-damage crashes, providing insights into potential areas for intervention.

Hypothesis 1: Traffic control devices and environmental conditions significantly impact the severity of crash damage.

Hypothesis 2: Certain features will have a stronger predictive power for high-damage crashes.

Method

Data and Preprocessing

I used crash data from the Chicago City Data Portal. Data belongs to the last 6 months of 2023.

The study utilized a dataset comprising various variables including : Posted Speed Limit, Traffic Control Device, Device Condition, Weather Condition, Lighting Condition, First Crash Type, Alignment, Roadway Surface Condition, Crash Hour

My target variable was damage that classified crash damage into three classes:

1. Over $1,500
2. Between $501 - $1,500
3. $500 Or Less

I did some preprocessing as follows:

1. Removing rows that had non-descriptive values (e.g. “UNKNOW”)
2. One hot encoding of categorical variables
3. Splitting data into training and testing with the portion of 80 to 20

Since in the process of one hot encoding of categorical variables, the number of variables increased dramatically, I had to do a feature selection to find the most important variables. Feature selection is a critical process in machine learning that involves identifying and selecting the most relevant variables for use in model construction. Its significance lies in enhancing model performance by reducing overfitting, improving accuracy, and reducing training time. In the context of predictive analytics for traffic safety, feature selection helps in isolating the most impactful factors from a vast dataset, which might include diverse attributes like traffic conditions, environmental factors, and crash characteristics. The chosen features not only contribute to a more efficient and interpretable model but also provide crucial insights into the key factors that influence the outcome, in this case, the damage of vehicle crashes.

The variables selected below were effective in predicting the damage of the crash:

Table 1. Selected features

Traffic Control Device	**Device Condition**
Bicycle Crossing Sign	No Controls
Delineators	Worn Reflective Material
Flashing Control Signal	Blowing Snow
Other Railroad Crossing	**Weather Condition**
Other Reg. Sign	Cloudy/Overcast
Police/Flagman	Fog/Smoke/Haze
Railroad Crossing Gate	Sleet/Hail
Rr Crossing Sign	**Lighting Condition**
Stop Sign/Flasher	Daylight
Traffic Signal	**First Crash Type**
Yield	Animal
Alignment	Overturned
Curve On Grade	Pedestrian
Curve On Hillcrest	Train
Straight On Grade	**Roadway Surface Cond**
	Ice

Model and Evaluation

In this study, I used XGBoost, short for eXtreme Gradient Boosting, which is a powerful machine-learning algorithm renowned for its efficiency and performance. It operates by constructing a sequence of

decision trees, each designed to correct the errors of its predecessor, leading to a highly accurate ensemble model.

However, the effectiveness of XGBoost is not just a product of its sophisticated design but also of its fine-tuning through hyperparameter optimization. Proper tuning of these parameters is essential because it helps in striking a balance between the model's ability to learn complex patterns and its tendency to overfit the training data. I used the F1 score as a classification metric.

Model Performance

The XGBoost model achieved an F1 score of `0.706` on the test set, indicating a good balance between precision and recall in the model's predictions.

Findings

Feature Importance

Feature importance in XGBoost refers to a metric that quantifies the contribution of each feature to the model's predictions.

Table 2. Feature Importance values

0.224552	**First Crash Type** (Pedestrian)
0.093702	**Traffic Control Device** (Stop Sign/Flasher)
0.090530	**Lighting Condition** (Daylight)
0.089548	**Alignment** (Straight on Grade)
0.084966	**Device Condition** (No Controls)
0.082532	**Traffic Control Device** (Traffic Signal)

SHAP

I used SHAP (SHapley Additive exPlanations) Value Analysis to determine the type and magnitude of contribution (negative or positive) of each feature on the prediction of high-damage crashes (OVER $1,500). In contrast to feature importance, SHAP values explain how each feature affects the model's output for a specific instance, allowing for a deeper understanding of model behavior. SHAP plot reveals what feature value makes what kind of contribution to the model, the color red is equal to high values (here 1) while blue is a sign of low values (here 1), and the position of dots concerning vertical line show magnitude and sign of contribution. The right side of the vertical line indicates positive SHAP values, while the left side indicates negative SHAP values and the distance of dots to the line shows magnitude. Each dot is a sample in test data.

As we can see in Figure 1 below, **Lighting Condition** (Daylight) which is shown by plenty of offset red dots shows a huge negative impact on OVER $1,500 damage crashes, the same thing is true for **Traffic Control Device** (Traffic Signal) but the instances with positive effect are more with lower impact (Blue dots). The minority of **Traffic Control Device** (Stop Sign/Flasher) points shows a huge positive impact on OVER $1,500 damages, while the majority of dots tell another story. **Device Condition** (No Controls) mostly has a low negative impact. Some **First Crash Type** (Pedestrian), **Weather Condition** (Cloudy/Overcast) show a big negative impact, unlike the rest of the points. The other variables, approximately, have a low positive contribution with some exceptions.

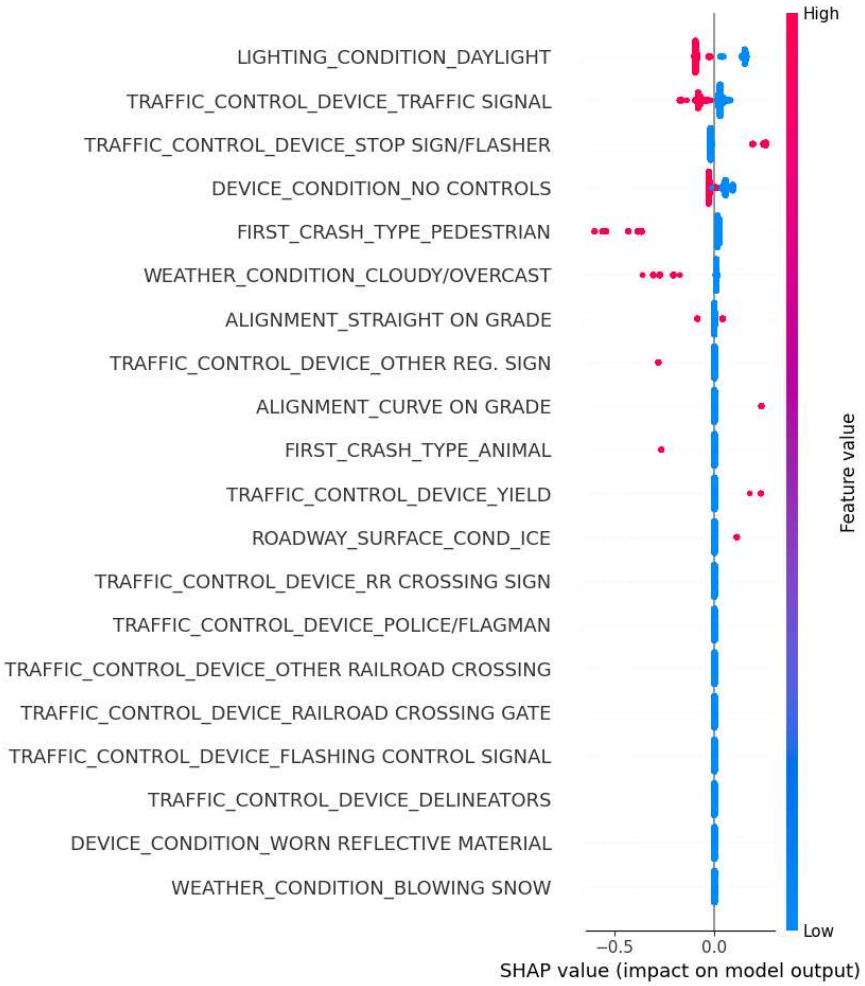

Figure 1. SHAP plot

REFERENCES

HTTPS://DATA.CITYOFCHICAGO.ORG/TRANSPORTATION/TRAFFIC-CRASHES-CRASHES/85CA-T3IF/ABOUT_DATA