Predicting Car Crash Severity and Frequency in Maryland

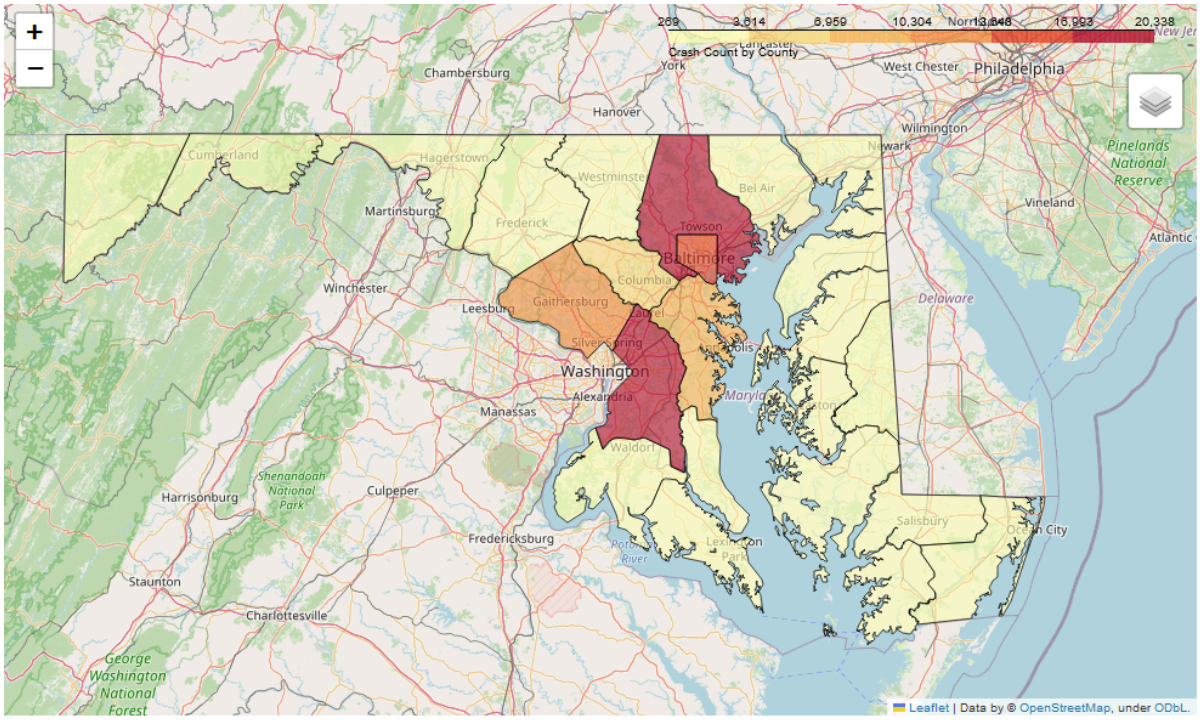
Dennis Piliptchak
and
Addis Yesserie

Abstract

In the state of Maryland alone, car crashes kill hundreds of people every year and injure thousands more. Traffic safety analysis is needed to determine the main causes of this dangerous subset of crashes. Our goal is to understand and predict the factors affecting crash occurrence, injury severity, collision type, and their underlying relationships in order to predict car crash severity in the future. We trained several different machine learning models to get the most important features for predicting car crash severity. The models we used were logistic regression, random forest, and XGBoost. XGBoost was the best-performing model with an overall accuracy of 87% and recall of 73%. It was also the best model at predicting fatal and injury crashes, which are the two more important classes compared to property damage crashes. In terms of the main causes, we found that distracted driving, incorrect use or lack of safety equipment, ignoring the right of way, and poor weather conditions were all factors contributing to a large percentage of fatal and injury crashes. We also found that high speed crashes had a small impact on the number of dangerous crashes, so future policy might be better directed toward stopping distracted driving or ensuring drivers, pedestrians, and cyclists use proper safety equipment.

Introduction

Traffic accidents pose a significant public safety concern, leading to numerous injuries and fatalities worldwide. According to WHO, every year the lives of approximately 1.19 million people are cut short by car crashes. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability. In the U.S., according to the National Highway Traffic Safety Administration (NHTSA), an estimated 20,190 and 19,515 people died in motor vehicle traffic crashes in 2022 and 2023 respectively in spite of the decline in traffic fatalities after the Covid-19 pandemic. Traffic deaths are continuing to slowly fall—but there is still a long way to go to achieve the expected safety standard set in the core mission of the U.S. Department of Transportation. In Maryland specifically, there are about 500 fatal crashes and several thousand injury crashes per year, primarily in places like Prince George's County and Baltimore County.



Maryland is dedicated to saving lives and preventing fatalities caused by motor vehicle crashes. The Maryland Highway Safety Office uses a data-driven approach to try and minimize the number of fatal crashes. Despite new advances in car safety technology like backup cameras, adaptive headlights, and advanced collision warning systems, motor vehicle crashes still happen at high rates.

Policymakers want to reduce the number and severity of all accidents, and to do that they need to know what policy to push. A data-driven approach like how the MD Highway Safety Office does is especially important because it can find the most important factors causing dangerous crashes. Without it, they would be primarily making educated guesses at what the real problems are. This could be misleading because there could be a factor that seems like it could be especially important, but in reality is not. The opposite can also be true, where a factor that was assumed to be only mildly important is actually the root cause of many crashes. Machine learning models like logistic regression, decision tree, and random forest are used in combination with traditional data analysis to better predict when and how car crashes will occur. Policymakers in the past have been more reactive in their response to growing numbers of crashes, but with this kind of predictive analytics, they can be more proactive and fight against dangerous factors ahead of time. Several studies have undergone this line of analysis in other locations like Chicago and Texas.

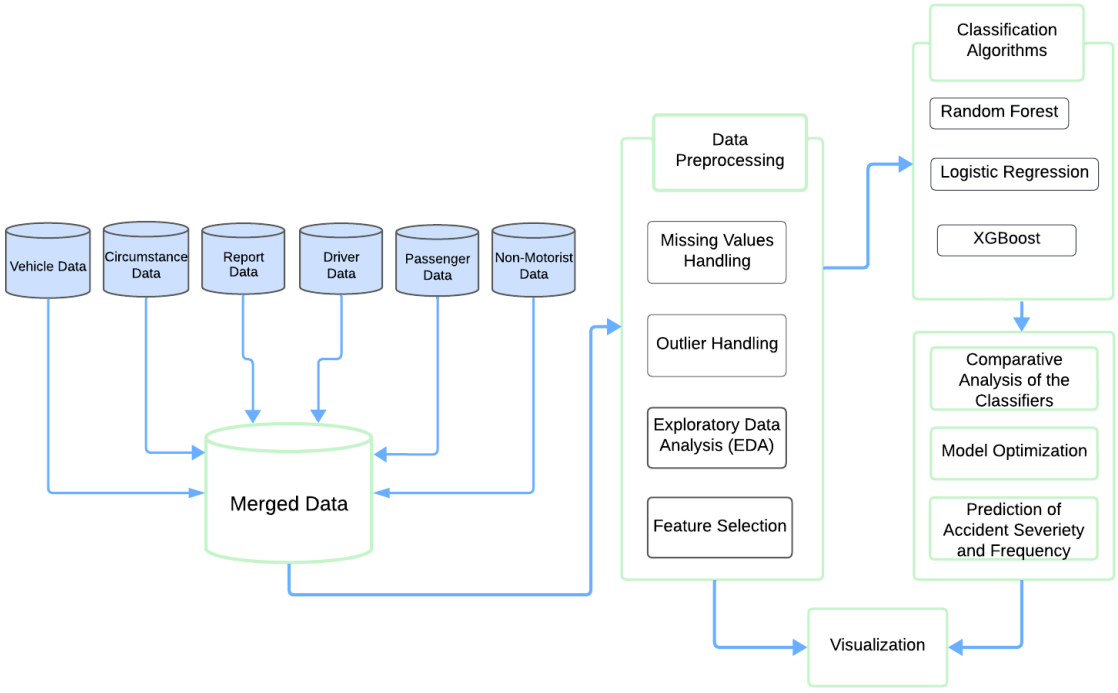
This project is designed to reduce the number of crashes across Maryland by training a machine learning model to find the most relevant risk factors for different kinds of vehicle crashes (fatal, injury, or property damage). Different risk factors include location, speed limit, driver status, road status, weather, etc. Crash severity prediction models enable different agencies to predict the severity of a newly reported crash with unknown severity or the severity

of crashes that may be expected to occur sometime in the future more broadly. The model was accompanied by maps visualizing roads and areas that are highly concentrated with crashes. Depending on which features the model believes to be the most important, state agencies could concentrate more resources on addressing those issues specifically.

Methodology

This study mixes both traditional analysis and machine learning models to classify crashes as fatal, injury, or property damage. As part of our results, we examine what the model believes to be the most relevant features, or the ones that are most highly weighted when classifying a crash. We take special care with specifically the injury and fatal crashes because they can be modeled more directly. Property damage crashes can happen under almost any circumstances, but injury and fatal crashes are more likely to have some kind of pattern associated with them.

Before creating the model, the first step is to merge the data together. The raw data comes in six categories: circumstance, driver, passenger, nonmotorist, vehicle, and report data. The records in these categories can be joined together by using a shared report number. Each crash has a unique report number. After joining the data, there will be one data file with all the combined data. We created two separate models for 2020 and the other years, with the reasoning explained later on.



As illustrated in the above chart, the next step after merging the data is data preprocessing. This includes data cleaning, where missing values and outliers are handled. Feature selection is

done to find the columns with the most null values and remove them. Lastly, EDA is done to gain some initial insight into the data and to answer some preliminary questions with both queries and visualizations.

After preprocessing is complete, the data is ready to be modeled. We used three different classification models: logistic regression, random forest, and XGBoost. We compared the three models to see which is the best one, performed model optimization with gridsearch, and used it to predict crash severity.

Data Collection

This project used data obtained from the Maryland Crash Database. Available data is compiled from police crash reports approved and submitted to the Maryland Department of State Police (MDSP) via the Automated Crash Reporting System (ACRS). Crash locations reflect the approximate locations of the incident based on longitudinal and latitudinal information provided by the officer through ACRS. The dataset provides general information about each incident tracked with a report number and includes report, driver, nonmotorist, passenger, vehicle and collision circumstances occurring on all county and local roadways, state, and interstate highways within Maryland. The dataset is updated daily, and the version used for this project contains data from January 2020 to December 2023. In a given year there are about 100,000 reported incidents.

It should be noted that the timespan of our data ranges from the start of 2020 to the end of 2023. This is very important because Maryland was still undergoing covid policies like isolation in 2020. Since the roadway environment was significantly different between 2020 and the years following, we decided to treat 2020 as a completely separate dataset. Training a model from 2020 data could result in model drifting, meaning that the circumstances of the 2020 data could be significantly different from today's circumstances, thus worsening the model.

The information from this dataset used for the model and visualizations include, among others, crash report type, crash date and time, road name and type, collision type, weather, surface condition, and driver substance abuse. Using Python libraries, the data will be read and cleaned to drop null values, unnecessary columns, and any incomplete records. We also added some visualizations to add to our exploratory analysis in more detail.

Data Preprocessing

Data preprocessing has several parts, and every part needs to be completed to the fullest to prepare the data to train a classification model with. Preprocessing involves data merging, data cleaning, feature extraction, and feature selection. Once this step is complete, the data should be structured enough to enter a machine learning pipeline to go into the three models, logistic regression, random forest, and XGBoost.

Data Merging

The raw data exists in six different categories. The six categories are circumstance, driver, passenger, nonmotorist, vehicle, and report data. Each category has columns related to it, so the driver data for example has data about the driver condition, whether he has been under the influence, and so on for the remaining categories. Each reported crash has a report number that can be used to join all the different categories together. Doing an outer join with all the categories ends up with the total number of crashes.

Data Cleaning

This dataset was not clean, and it actually had several mistakes in it. This is to be expected with the way that individual records are added. Officers use the Automated Crash Reporting System (ACRS), which in spite of its name is not automated; it is a customized user interface designed to make reporting crashes easier. Officers can still make mistakes when adding new records.

The process of data cleaning involves three major steps: transforming the shape of the data, removing or fixing incorrect records, and getting rid of null values. The shape of the data was fine the way it was, where one row corresponds to one report number. In other words, one row is one crash, and this is the proper shape needed to feed into our models.

As mentioned previously, several records have mistakes in them. For example, the coordinates of every crash should be inside Maryland borders, but several or not. This could be attributed to mistakes in the GPS or perhaps hit and run incidents that ended somewhere outside the state. There were other kinds of typos in several different columns, like car make and model, or time of day of the crash.

The combined dataset has 135 columns between the 6 categories outlined earlier. Most of these columns are categorical, so their values all need to be one-hot encoded. This will end up being a weakness of the model because of the heavy use of indirect evidence. When there is a high number of records that all have unique values in some columns, the models cannot use them as effectively as when the records are mostly similar. Since most crashes are unique in nature when considering all the possible features, it's important to try and categorize them in some ways to prepare for a meaningful analysis.

The number of columns was reduced from 135 down to about 40. Our process was to find the columns with the most null values. Columns with many null values are the features that are not present in the vast majority of crashes. For example, most passenger and nonmotorist data was removed because it was not collected in the first place for 99% of crashes.

Feature Selection

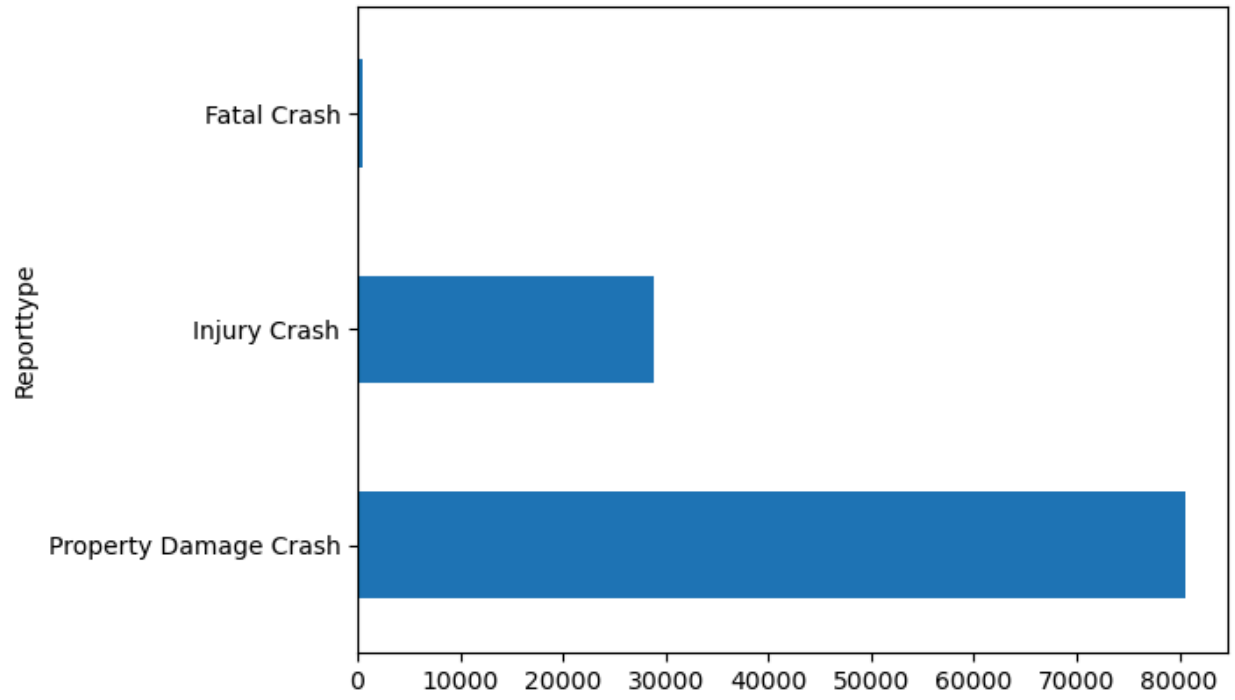
We started with a list of features that we thought could be interesting. We also reduced the number of features from 135 to 40 by examining the null values. More specifically, we opted to keep the columns with less than 50% null values. It should be noted that among the remaining columns, the one with the most null values had about 25%. This indicates that there were many

columns with a very high proportion of null values and many columns with a low proportion of null values, but not ones in between. For the remaining null values in each column, the “unknown” or “other” value is imputed when applicable.

Besides cutting features, we also did some feature engineering to produce some features that we thought could be useful at predicting crash severity. We changed the time of the crash to the hour of the crash. It is less useful for our model to time crashes by the minute because the minute has no bearing on the crash, but the hour does. We did something similar with driver dates of birth; we changed the feature from date of birth to generation. For example, it would be impossible to say that drivers born on August 6th, 1955 are more at risk, but it would be possible to say that baby boomers are more at risk than millennials in fatal crashes.

Exploratory Data Analysis (EDA)

EDA encompasses all of the previous steps taken, including data cleaning, feature selection, and feature extraction. In addition, we performed some exploratory queries to look at the structure of the data which could provide us leads to follow when performing a more detailed analysis. For example, we looked at the target class in some detail to see whether or not it was balanced. Naturally, fatal crashes are the smallest target class, with only a few hundred crashes per year. Property damage crashes are the largest target class and consist of the majority of crashes.



We started our EDA by looking at all the unique values about some of the columns we thought were the most important, like the type of circumstance, the type of collision, the condition of the

driver, time of the crash, day of the week of the crash, speed limit, driverless car involvement, and several others. Based on the count of each kind of result, we gauged how often that feature was involved in crashes more broadly as well as the distribution of values for that feature.

The data contained coordinates for every crash, and we wanted to use these to the maximum extent we could. The data also contained a road name and general location for each crash. We created an interactive heat map to find crash hotspots. These are usually at intersections, large roads like highways, or both. Policymakers can use this map to find particularly dangerous locations, even ones that are not expected to be.

Before moving on, we should discuss some of the hypotheses about car crash causes that the models and general analyses attempt to answer. These are several of the questions we wanted to answer or gain insight on, and there are many more not listed here:

1. Is drinking and driving still one of the most major causes of fatal crashes? Is it trending up or down?
2. Does having more cars on the road imply more crashes, particularly more injury and fatal crashes?
    a. Among people working mostly at home, most of them commute to work on Tuesday. Are there more crashes on Tuesday? Are there significantly more crashes during rush hour?
3. Are old people more at risk than young people? If so, how much more at risk? Are there more reckless young drivers involved in fatal crashes?

4. Is speeding as important a factor in injury and fatal crashes as we think it is? Could driver negligence be more important?
5. How safe are self-driving cars really?

Modeling

As described in the methodology, we used three different models to classify crashes as either fatal, injury, or property damage: logistic regression, random forest, and XGBoost. For each model, we used gridsearch to optimize its hyperparameters.

As part of the pipeline, any remaining null values in the remaining features would impute the average or most common value for both numeric and categorical columns. We did manually assign the null values of these columns to “other” or “unknown” when applicable already, but this pipeline serves as a redundancy to catch any unexpected values.

Accuracy was not a particularly valid metric for us due to the imbalance in the target class. Fatalities consist of less than 1% of total crashes, so creating a model that always chooses property damage crashes would have a high accuracy despite being useless. Instead, we viewed recall as the most important metric; if a model can accurately predict when injury and especially fatal crashes happen, despite how relatively few there are, then that would be very useful if used in the field. Some studies opt to add synthetic data to balance out the classes, but we chose to leave the data as is to keep the model training as realistic as possible.

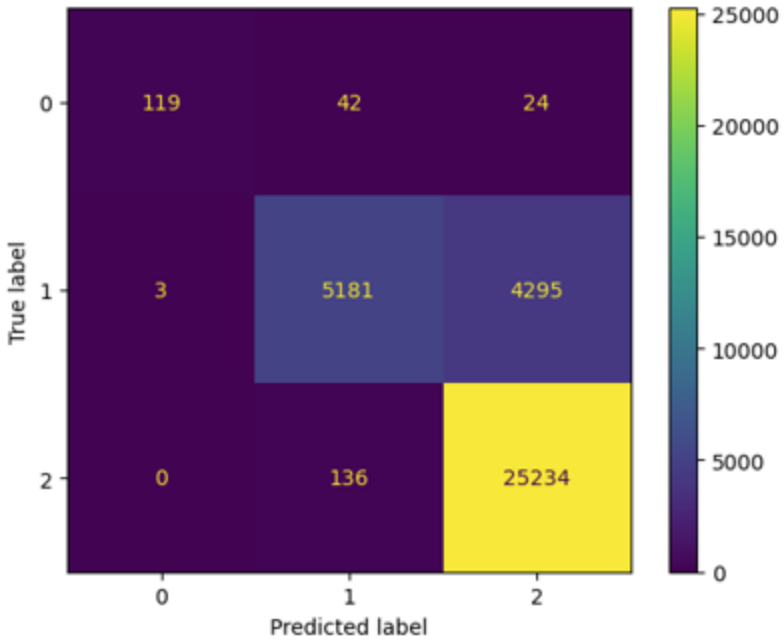
If the model is well-performing, then we can assume that its weights for all the different features are relatively reliable. The model can rank the features used to train it in order of importance using a kind of correlation coefficient where the higher the coefficient, the more important that feature is.

Results and Discussion

Between the three classification models, XGBoost performed the best; it had an accuracy of 87% and an average recall of 73%. Its recall with fatal crashes is the highest among all three models at 64%. This is the main reason why we consider it the best model. In terms of recall on fatal crashes, logistic regression scored a 30% and random forest scored a 0%.

About recall, the importance of false positives and false negatives should be emphasized. A false positive in this case means that the model predicts that a crash is either property damage or injury when it is actually a fatal crash. A false negative is when the model reports a fatal crash when in reality it is not. Both are bad in this case, since false positives could result in an underallocation of resources sent to a crash site while false negatives could result in an overallocation of resources. In terms of our models, they are quite good at minimizing false negatives, but have trouble preventing false positives.

	precision	recall	f1-score	support
0	0.98	0.64	0.78	185
1	0.97	0.55	0.70	9479
2	0.85	0.99	0.92	25370
accuracy			0.87	35034
macro avg	0.93	0.73	0.80	35034
weighted avg	0.89	0.87	0.86	35034



In terms of the most important features, the suspected driver injury was the most important. This seems obvious, but suspected injury could differ from reality. For example, a crazy crash could result in an unharmed driver, or it could result in a seemingly unharmed driver who is in shock, but in reality is in critical condition. The next most important feature is whether or not a pedestrian/cyclist was involved in the crash, which makes sense since they have little to no protection once they are hit. A lack or partial use of safety equipment was also a common cause of injury and fatal crashes, and over 40% of all fatal crashes did not have a deployed airbag. 52% of injury crashes did have a successfully deployed airbag. The large difference between the percentages of fatal and injury crashes in regard to airbag deployment indicates that airbags are keeping drivers more safe; collisions that would have been fatal may have been reduced to injury crashes due to airbags.

To answer some of the questions outlined earlier, some of our hypotheses were supported and others were rejected. We confirmed that drunk driving is still a major problem; over 31% of fatal crashes were alcohol or drug related. Head on collisions were the most dangerous type of crash. We can potentially draw a connection between head on collisions and drunk drivers since they have a tendency to veer into oncoming traffic.

We did not find any evidence of there being significantly more dangerous crashes on Tuesday. In fact, Friday is much more dangerous. There was also no significant difference in the frequency of dangerous crashes between young and old people; baby boomers/Generation X and Generation Z both hover at about 20% in both fatal and injury crashes. About self-driving cars, we found that only 0.5% of fatalities and 1% of injury crashes involve self-driving cars. This may not seem that impressive when considering how few self-driving cars there actually are on the road, but this is actually not the case. 25% of cars made since 2020 have level 2 autonomy, so self-driving cars do seem quite safe.

One key observation we found was that speeding did not seem to be an important factor for either injury or fatal crashes. When examining injury and fatal crashes, 4.32% and 1.70% of them respectively occurred with drivers either going above the speed limit or too fast for the conditions. Moreover, none of the classification models put the speed limit among the most significant features. It wasn't a useless feature, but it was not an especially important one either.

Based on these findings, there are three major points. The first is that drunk and distracted driving as a whole is still a major problem and should be addressed more severely. They are the most at risk of causing head on collisions with both opposing vehicles and pedestrians. Secondly, improper use of safety equipment puts the driver at higher risk of death; keeping working airbags and wearing the seatbelt correctly will make a big difference.

Lastly, dangerous high speed crashes are not as prevalent as they used to be. It's possible that recent advances in car safety specifically designed to protect drivers in high speed crashes have been successfully keeping drivers more safe in the past few years. For this reason, policymakers should potentially consider reallocating some resources to target these different issues. For example, some police that are currently stationed at speed traps might be better utilized at a sobriety checkpoint.

Future Development

Earlier we showed a heat map that displays crash hot spots on a local level. Policymakers in these local areas can use it to find spots to allocate more resources in, but this is still a manual process.

A potential future development would be to integrate the heat map into the machine learning models. This would be done by implementing a location-based clustering algorithm and comparing crash features of one cluster to other clusters. Doing so would let us see the relationship between a particular road or intersection and its unique effect on some crash features that are not there for most other clusters.

A good way to improve model performance would be to implement data segmentation. We briefly mentioned that most crashes are unique when looking at every feature together. This introduces a high amount of noise that could reduce the effectiveness of the model. PCA could eliminate some of the noise but at the cost of damaging some of the interpretability of the

model. The solution is to create several smaller-scope models designed to consider only a couple crash characteristics at a time where individual records can overlap and produce more clearly defined patterns in the data.

Conclusion

We used both traditional and AI-based analysis to predict crash severity and frequency in Maryland. The XGBoost model was the most well-performing with an average accuracy of 87% and an average recall of 73%, correctly predicting fatal crashes 64% of the time. Based on the models, we saw that pedestrians are the most at-risk group. Drunk or distracted driving and the incorrect use of driver safety equipment were two premier causes of fatal and injury crashes, whereas speeding was not as significant a feature as expected.

References

Al Mamlook, R. E., Abdulhameed, T. Z., Hasan, R., Al-Shaikhli, H. I., Mohammed, I., & Tabatabai, S. (2020, July). Utilizing machine learning models to predict the car crash injury severity among elderly drivers. In *2020 IEEE international conference on electro information technology (EIT)* (pp. 105-111). IEEE.

AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019, April). Comparison of machine learning algorithms for predicting traffic accident severity. In *2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT)* (pp. 272-276). IEEE.

Alkan, G., Farrow, R., Liu, H., Moore, C., Ng, H. K. T., Stokes, L., ... & Zhong, Y. (2021). Predictive modeling of maximum injury severity and potential economic cost in a car accident based on the General Estimates System data. *Computational Statistics*, *36*, 1561-1575.

Assi, K., Rahman, S. M., Mansoor, U., & Ratrout, N. (2020). Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol. *International journal of environmental research and public health*, *17*(15), 5497.

Bhowmik, T., Yasmin, S., & Eluru, N. (2021). A new econometric approach for modeling several count variables: a case study of crash frequency analysis by crash type and severity. *Transportation research part B: methodological*, *153*, 172-203.

Fiorentini, N., & Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures*, *5*(7), 61.

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27-36.

J. Zhang, Z. Li, Z. Pu and C. Xu, "Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods," in *IEEE Access*, vol. 6, pp. 60079-60087, 2018, doi: 10.1109/ACCESS.2018.2874979.

Rahman, M. S., Abdel-Aty, M., Hasan, S., & Cai, Q. (2019). Applying machine learning approaches to analyze the vulnerable road-users' crashes at statewide traffic analysis zones. *Journal of safety research*, *70*, 275-288.

“Road traffic injuries.” https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed 13 December 2023).

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of safety research*, *80*, 254-269.

Wang, X., & Kim, S. H. (2019). Prediction and factor identification for crash severity: comparison of discrete choice and tree-based models. *Transportation research record*, *2673*(9), 640-653.