Corvinus University of Budapest

Classifying collision severity using machine learning on British road safety data

Authors: Ádám Burkus, Attila Sztreborny,

Bendegúz Birkmayer, Bojta Rácz,

Roland Tuboly

Social Data Science MSc

Instructor: Johannes Wachs

Applying and Interpreting Machine Learning

2025.

Table of content

Motivation and research questions

In 2024 road accidents claimed 446 lives in Hungary and an additional 4143 collisions resulted in serious injury out of 14 687 total road collisions (KSH, 2025). Although fatalities related to road collisions have been decreasing in the past few years, in our opinion these numbers are still high, and it is important to know which collisions will result in fatality. With the knowledge of what features and conditions enhance the probability of fatal collision, hospitals and police forces could prepare better for unexpected events and car manufacturers could also implement measures to increase the chance of survival in case of an accident.

With that being said, the main question of our research is how we can classify road collisions, which variables decide or explain the probability of whether a collision will result in death or not.

Data

We use road safety data about circumstances of personal injury road collisions in Great Britain from 2024 (GOV.UK, 2025). Originally, we downloaded three datasets, one containing details about the collision itself (*dft-road-casulty-statistcs-collision-2024.csv*), and the other two about the casualty (*dft-road-casulty-statistcs-casualty-2024.csv*) or the vehicle (*dft-road-casulty-statistcs-vehicle-2024.csv)* participating in the accident.

First, we filtered out the redundant, duplicated or simply unnecessary variables (e. g. police department or location which both have extremely many distinct values) from each dataset. Then we aggregated the datasets to collision level, because our goal is to classify collision severity. After the aggregation, we merged the three into one data frame, where we have 100 927 observations and 34 variables.

We modified the data and time of day variables into more interpretable numeric variables. We also modified the data types of our variables into proper format, because lots of categorical variables, where numbers represent categories, were originally in numeric format. After conversions we got 7 numeric and 27 categorical variables.

We created our dependent variable by constructing a binary variable from collision severity data, where 0 equals non-fatal (serious and slight collision severity combined) case and 1 equals fatal case. Out of all observations, around 1.5% of them are fatal and the rest is non-fatal, so our data is rather imbalanced.

Methodology

We imputed numeric variables which contained NAs using the mean of given variables, and we also imputed our categorical variables with a “missing” constant if they contained NA values.

First, we used Lasso Logistic Regression (λ = 0.1), so we had to standardize our variables as well. Apart from Lasso Logistic Regressions we created 9 different Random Forests each with 200 decision trees as well. We constructed our trees with different maximum depths (10, 20, 30) and different minimum leaf sizes (1, 5, 10).

In order to evaluate our models, we created a train-test split with 5-means cross validation (20%-80% ~ 80 741 – 20 186 obs.), where the proportion of fatal to non-fatal cases remained consistent. We calculated all three measures (mentioned in the following) on the test sets, but used Recall measure (TP / TP + FN) instead of Accuracy or AUC to select our best model. We chose this because Recall focuses on how well we detect true positives (fatal cases), and it works well on imbalanced datasets, such as ours.

Baseline models

We used some basic baseline models to compare them with our more complex models. Greater and positive deviation from them indicates that it was worth it to implement more sophisticated methods and that there is some real connection between the independent and dependent variables. As the dataset is highly imbalanced (~98.5% of the cases are non-fatal collisions) classifying every observation as non-fatal would result in a great, 98.5% accuracy. Despite the great accuracy, we wouldn’t get the answer to our main question: which are the deadly accidents, as the true positive rate would be 0.

Our other baseline model is random guessing, that randomly picks 98.5% of the data as non-fatal and the remaining 1.5% as fatal. Our goal is to find a model that can exceed these baseline models mostly in respect of true positive rate.

Results

Our first model was a Lasso Logistic Regression with λ = 0.1 parameter that was tuned by K-fold cross validation. With the best lambda value, the Lasso was using 86 variables (the categorical variables were split into several categories, therefore, it seems like we have more variables than in the beginning) to fit the model. It found that the most important factors that contribute to a fatal collision are the followings: age band of the casualty, the speed limit and the surface conditions of the road. We evaluated the Lasso Logistic Regression's fit on a separated test data, and the model excelled the baseline models. It predicted 78% of all fatal collisions. However, the accuracy was only 71%. It means that this model predicts a lot of false positive cases, in order to find most of the true positives. The AUC value is 0.83 that indicates a good performing model.

We also used Random Forest to predict which collisions are fatal and which aren't. The drawback of this model is the lack of interpretability of the variables' influence. However, the most relevant factors that decide the outcome of the accident were the following: the speed limit, whether it was an urban or rural area and junction control. This model had a significantly better accuracy than the Lasso, as here it was 76%. However, it found only 60% of the fatal collisions, which means the Random Forest had a stricter threshold for a collision to be classified as fatal. Great accuracy is easy to achieve in imbalanced datasets thus the significantly smaller true positive rate indicates that this model had a poorer performance compared to the Lasso Logistic Regression.

Table 1. – Accuracy, True Positive rate and AUC values for the 2 baseline models and the 2 more sophisticated models

Model	Accuracy	True Positive rate	AUC
Baseline – all non-fatal	0.9851	0	0.5
Baseline – random guessing	0.9704	0.0133	0.4991
Lasso Logistic Regression	0.7101	0.7767	0.8287
Random Forest	0.7620	0.603	0.8009

source: own compilation

Conclusion

As a conclusion we can say that the Lasso Logistic Regression performs better based on our results and data. While it predicted more false positive cases than the Random Forest, the Lasso caught the most fatal collisions, almost 4 out of 5. In this research question, Type II Errors are much more problematic than Type I errors as not recognizing a fatal collision can cost human life. Also, we found that the most relevant factors that influence the outcome of an accident weren't what we had expected in the beginning. Weather and light conditions, time of day, condition of the vehicle or the age and sex of the driver aren't among the 5 most relevant variables. Meanwhile, the speed limit, the surface condition of the road and whether it is urban or rural area seems relevant in this question.

Limitations

One limitation to our research could be that we examined only one year (2024) of road casualty data. Analyzing a longer time period, for example the last 10 years available, would make our models more robust and our classification results more accurate. Looking at an even longer period, on the other hand, would distort our results, because those factors that were important decades ago might not be as crucial now (vehicle safety, seat belts, road safety measures improved greatly over time).

Moreover, we only looked at British data and that is a good proxy for all Western European nations' collision patterns, but in Eastern Europe, Southern Europe or anywhere else in the world, conditions and driving traditions could be completely different. Southern European nations tend to be more carefree when driving, and Eastern European vehicle stock is way older and overused than their Western counterparts. That is why vehicle type could be more important in Eastern Europe and driver-related factors more relevant in the case of Southern Europeans.

References

KSH (2025): Road accidents resulting in personal injury on regional and county level – 2024-2025 quarterly https://www.ksh.hu/stadat_files/ege/hu/ege0077.html

GOV.UK (2025): Road safety data from Great Britain 2024 https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-accidents-safety-data