Analysis of Road Accidents in Catalonia

Iana Pokrovskaia

October 13, 2020

1. Introduction

According to ACEA (Europe Automobile Manufacturers Association), there are 312.7 million motor vehicles in circulation on the EU's roads, which is more than one to every two Europeans. For a long time, cars have been an important part of our daily life, and it is not going to change anytime soon. Many types of public transport use the same roads as cars. As a result, the number of people driving the roads every day is enormous. This leads to the high importance of road safety since road injuries are one of the most common causes of accidental death. However, improving road conditions also requires a lot of resources, which makes it necessary to take a systematic approach to determine which roads are more dangerous and need immediate attention.

An effective way to define the actual riskiness of the specific road is to analyze data about road accidents. This project aims to find features that affect road safety negatively, weigh their importance, conduct a statistical analysis of the current state of the roads, as well as suggest options for improvements.

It is obviously interesting for car drivers and citizens in general. Besides, this analysis can be useful for people responsible for future road improvements. The traffic police also may be interested in this analysis, since it identifies the conditions that require additional attention.

2. Data acquisition and cleaning

2.1 Data sources

For this analysis, I have used Catalonia Road Traffic Injuries & Deaths dataset from Kaggle (link), which consists of data from 2010 to 2020. It provides a lot of data about car collision circumstances, including the types of the road intersection, weather/light conditions, allowed speed limit and whether it was exceeded, date and time of the accident, and more.

2.2 Data cleansing

Firstly, I translated the dataset from Catalan to English for better understanding. After basic analysis, I captured a lot of missing values in some additional attributes like allowed speed, regulation of priority, type of section and so on. I decided to replace these values with “Not specified”, since there is a lot of ‘Not specified’ values in the dataset and it holds the same meaning. Furthermore, I changed the datatype of allowed speed column from float to string, because it presented not a continuous value, but a list of options in numeric format. Also, in the ‘ALLOWED_SPEED’ column I found the value 999, which could not be correct, so it was replaced with ‘Not specified’.

In all other respects, the dataset was perfect for further processing.

2.3 Feature selection

This dataset has a huge number of features, so an important task is to sort essential features from irrelevant ones.

To highlight important features, I analyzed the number of unique values and found out that for columns ‘ROAD’, ‘CITY’, ‘TIME’, ‘KM’, their number is 679, 854, 1298 and 2150 respectively. Since the total number of entries in this dataset is 16774, there is no benefit in using these features data during analysis. These columns have been deleted. There were also a large number of unique values in the ‘DATE’ column. The exact date of the accident is not suitable for further analysis, but information about the month may be useful. Therefore, based on the ‘DATE’ column, the ‘MONTH’ column was formed, and the ‘DATE’ column was deleted. The ‘YEAR’ column was deleted too.

I decided to remove the features containing data that can be obtained only after an accident. Although the data from the ‘N_DEATHS’ and ‘N_MAJ_INJURIES’ columns will be used for the subsequent analysis to determine the hazard level of the road, the rest of such features have no practical use for this project. This list includes columns ‘N_VICTIMS’, ‘N_VEHICLES’, ‘N_PEDESTRIANS’ and so on. ‘FOG_INFLUENCE’, ‘TRAFFIC_INFLUENCE’ and similar columns were also removed, and the same for ‘ACC_TYPE’, ‘ACC_CLASSIFICATION’ and ‘SPEED_LIMIT_DISPLAY’. The ‘WEEKDAY’ column contains information from the ‘WORKING DAY’ column, so ‘WORKING DAY’ can be deleted too.

To determine the strength of correlation between values, it is necessary to convert them to a numerical format. A temporary dataset was created using LabelEncoder from sklearn library. The obtained values of the Pearson coefficient are shown in Figure 1.

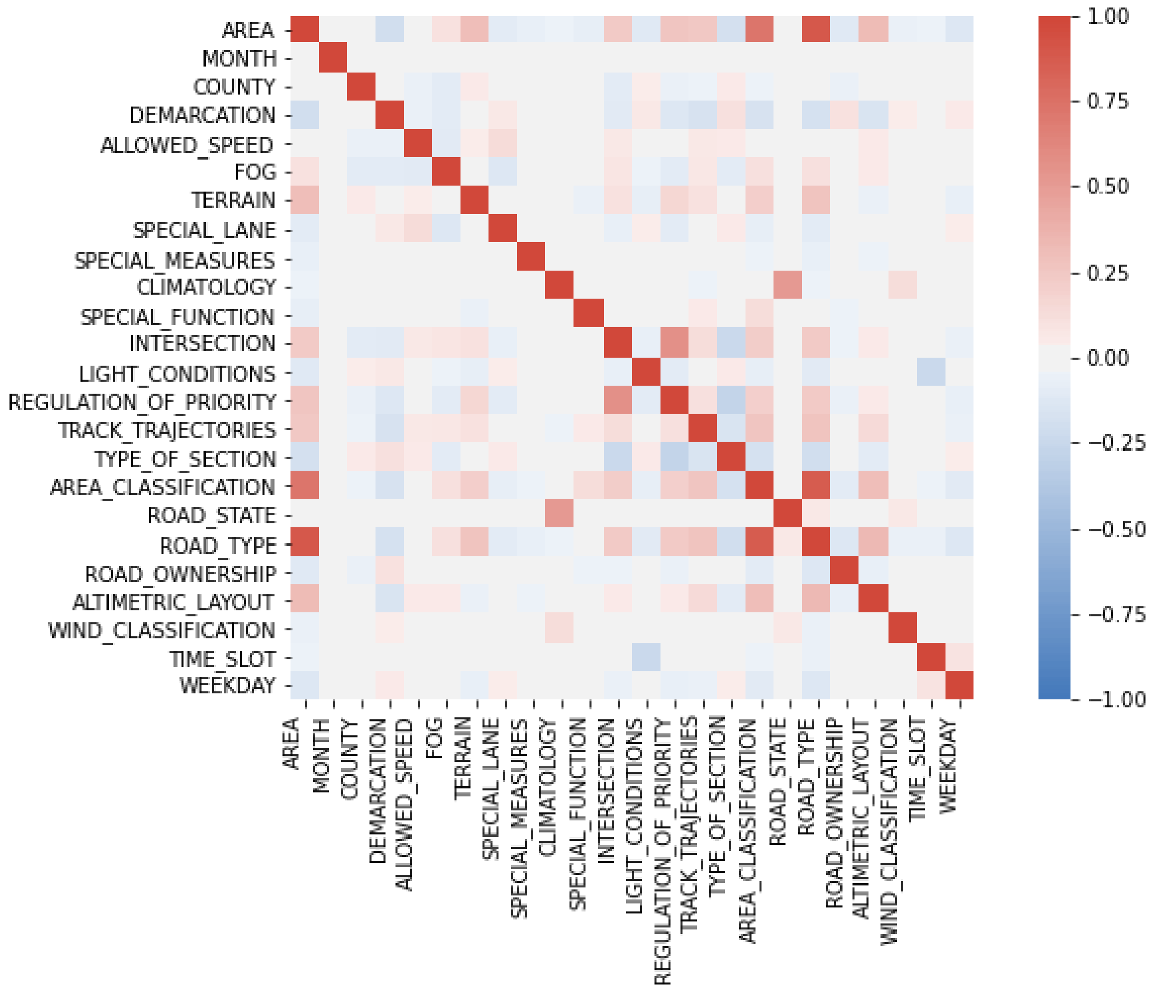

Figure 1. Pearson coefficient heatmap

As can be seen on the heatmap, there is practically no correlation between the values (the red diagonal line corresponds to identical features). The found relationships with the absolute value of the Pearson coefficient exceeding 0.7 are given in Table 1. P-values in all the cases are exactly 0.0.

Table 1. Features with the high value of Pearson coefficient

First feature	Second feature	Pearson coefficient
‘AREA’	‘AREA_CLASSIFICATION’	0.723
‘AREA’	‘ROAD_TYPE’	0.884
‘AREA_CLASSIFICATION’	‘ROAD_TYPE’	0.873

Thus, all three features are related. Values of ‘AREA’ and ‘AREA_CLASSIFICATION’ columns are close: ‘Urban area’ and ‘Road’ for the first and ‘Urban area’, ‘Road’ and ’Crossing’ for the second. Thus, the ‘AREA’ column can be dropped. The ‘ROAD_TYPE’ column can be dropped, too, since most of the values divide into two groups: ‘Urban road’ and ‘Conventional road’, which is directly related to the area type.

After all, 22 features were selected.

3. Methodology: Exploratory Data Analysis

3.1 Target variable

The target variable for this analysis is the road hazard ratio. To calculate it, the variables 'N_DEATHS' and 'N_MAJ_INJURIES' were used. First, the mortality and injuries ratio for the number of accidents is determined for each unique value of features. This determined the contribution of each road feature to the overall hazard ratio, which was calculated as a sum of deaths coefficient and major injuries coefficient divided by 2.

3.2 Features contribution

After further normalization, the maximum possible and average contribution of each feature relative to the total coefficient was determined, which is shown in Figure 2.

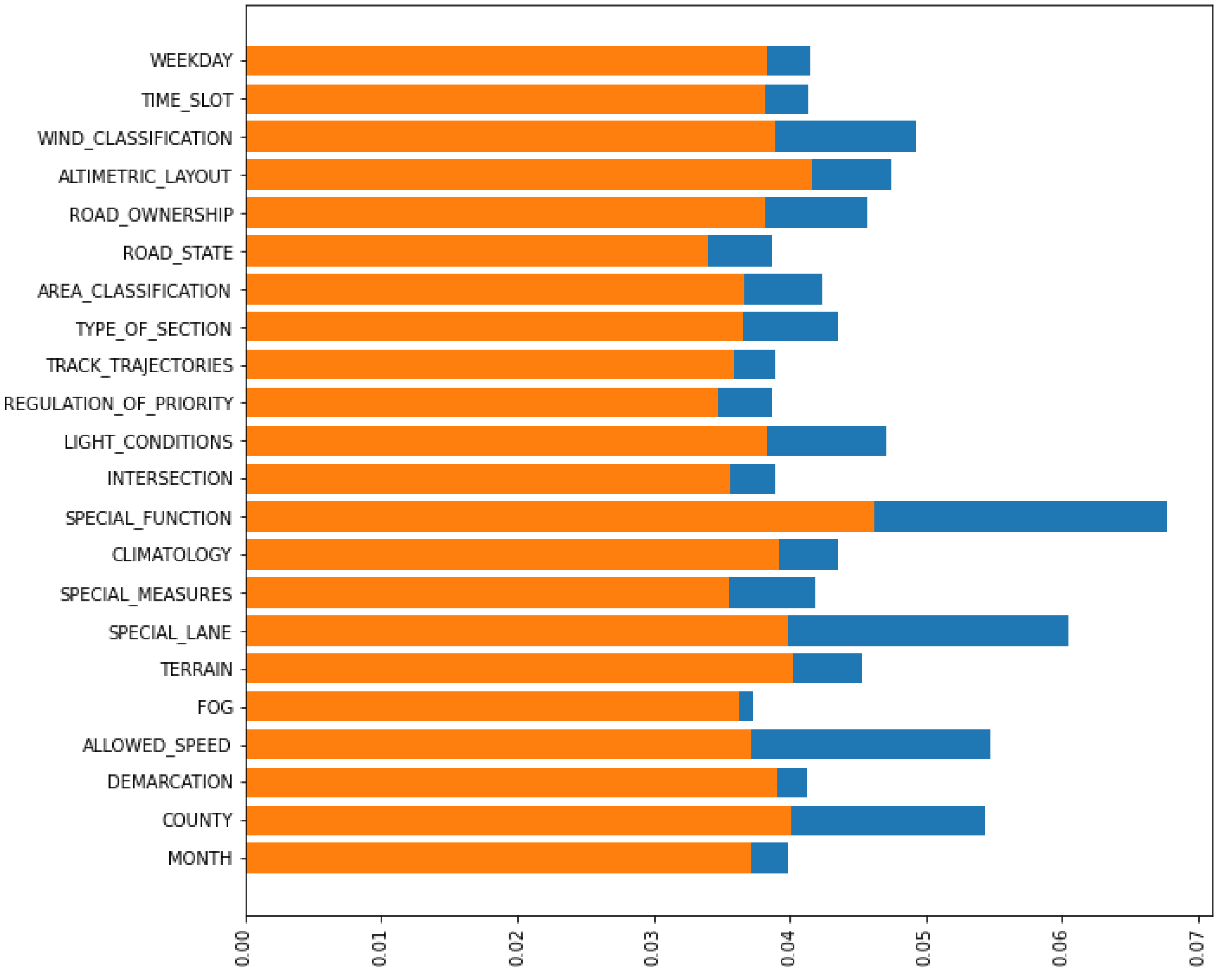

Figure 2. Maximal (blue) and average (orange) contribution of features

As can be seen in the figure, the average contributions of all features are approximately equal.

3.3 Groups of features

For further processing, the features were divided into three groups: road conditions, weather conditions, circumstances. The resulting lists are given un Table 2.

Table 2. Groups of features

Road conditions	Weather conditions	Circumstances
COUNTY	FOG	MONTH
DEMARCATION	CLIMATOLOGY	SPECIAL_MEASURES
ALLOWED_SPEED	LIGHT_CONDITIONS	INTERSECTION
TERRAIN	ROAD_STATE	REGULATION_OF_PRIORITY
SPECIAL_FUNCTION	WIND_CLASSIFICATION	TYPE_OF_SECTION
TRACK_TRAJECTORIES		ALTIMETRIC_LAYOUT
AREA_CLASSIFICATION		TIME_SLOT
ROAD_OWNERSHIP		WEEKDAY
		SPECIAL_LANE

Variables from the Road conditions column are the main ones for determining the hazard ratio of roads, the Weather conditions column allows to determine the hazard ratio in relation to certain weather conditions. The column Circumstances is dependent on time period and specific location, and it is most volatile. The contribution of the components from each column to the overall hazard ratio is shown on Figure 3.

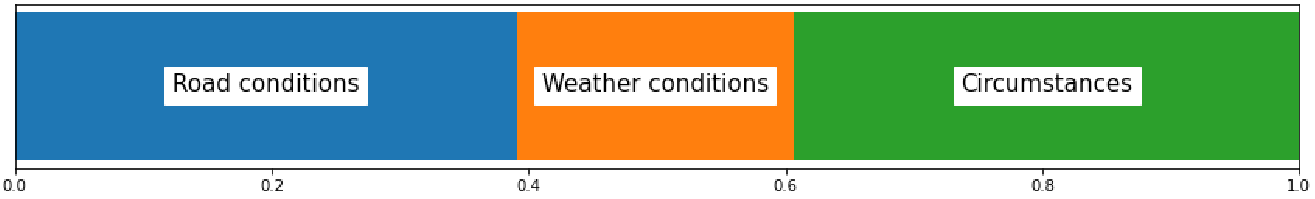

Figure 3. Contribution of the groups of features

Similar to the general hazard ratio, for these groups, the contributions of individual components can also be calculated. The results are shown on Figure 4.

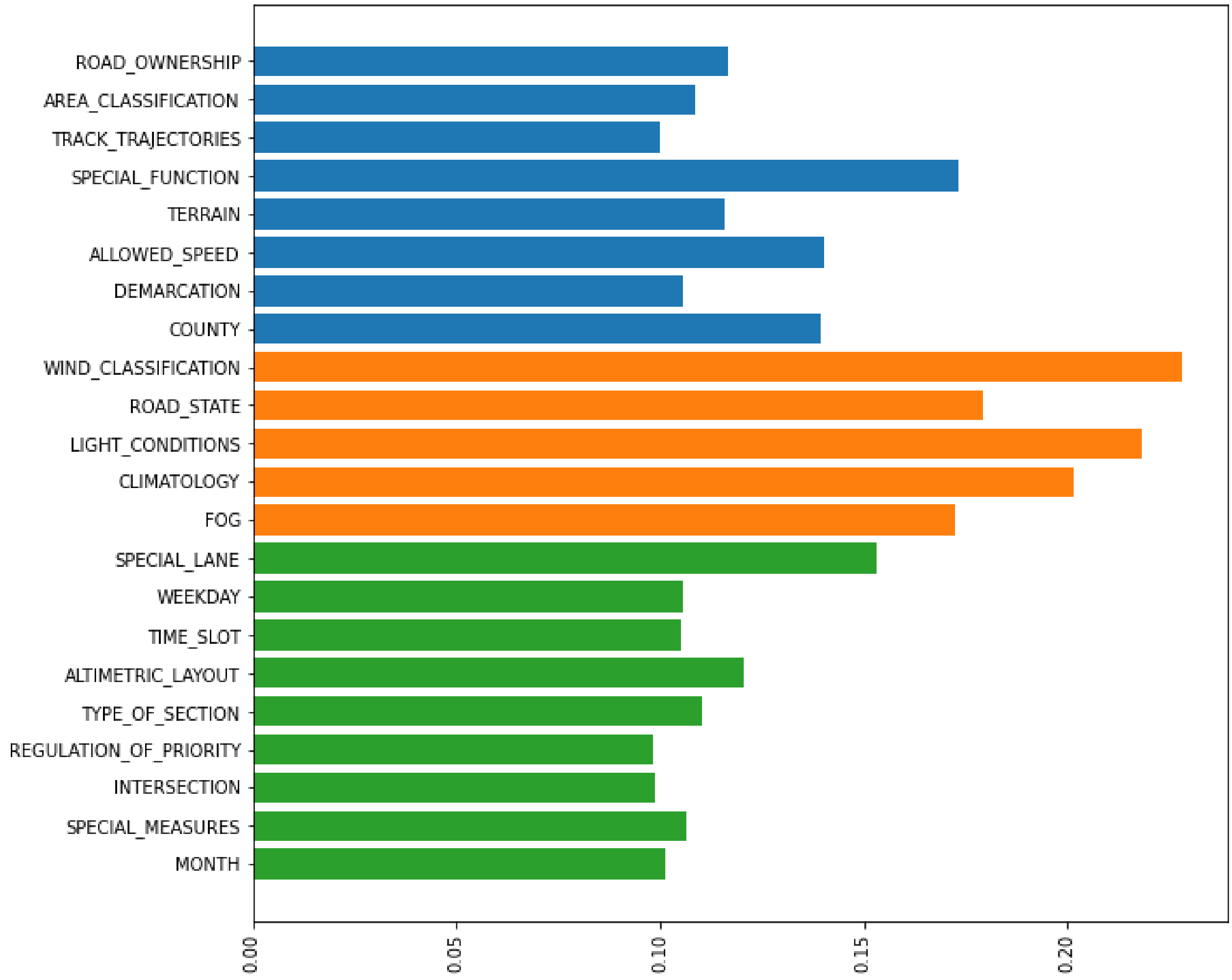

Figure 4. Contribution of features in the groups

3.4 Calculation of target variable

The final step is to actually calculate the hazard ratio for every entry of the original dataset. For this, additional columns 'HAZARD_RATIO_all', 'HAZARD_RATIO_road', 'HAZARD_RATIO_weather', 'HAZARD_RATIO_circumst' are formed. The statistics of the obtained hazard ratio values for various groups of features are shown in Table 3.

Table 3. Obtained hazard ratio values

Calculation	All features	Road conditions	Weather conditions	Circumstances
mean	0.818	0.762	0.862	0.850
std	0.026	0.044	0.018	0.026
min	0.746	0.690	0.710	0.772
25%	0.798	0.724	0.856	0.832
50%	0.812	0.750	0.856	0.850
75%	0.841	0.800	0.856	0.871
max	0.908	0.894	0.987	0.946

On average, the hazard ratio is between 0.7 and 0.9. Values less than or equal to 25% will be classified as low, 25-75% as medium, and the rest as high. This project will use the value obtained for road conditions. Figure 5 shows a histogram of the distribution of these hazard ratio values.

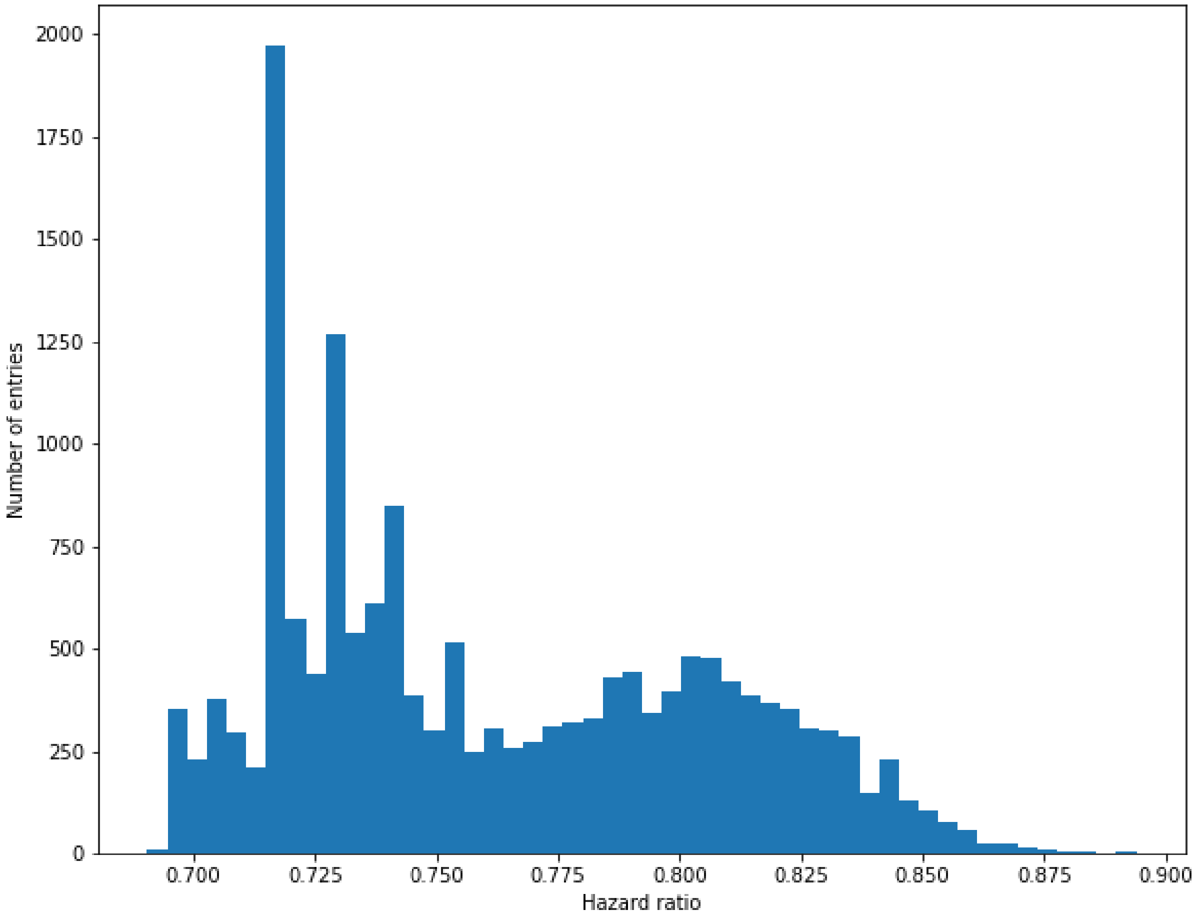

Figure 5. Hazard ratio values distribution (bins = 50)

4. Methodology: Machine Learning models

4.1 Preparing the dataset

To determine the most effective model, dataset is divided into 2 parts, training and testing on which will be carried out separately, and then interchangeably. This division also raises the important problem of the order of entries in the dataset, since all entries with same year of occurrence are going one after another. This can negatively affect the accuracy of the model, so the entries are shuffled. Both parts are encoded using label encoder.

4.2 Regression models

Since the target variable is continuous, a number of regression models were used in the project:

- Linear regression;
- Ridge regression;
- Lasso regression;
- Random forest regression;
- Gradient booster regression.

The R2-score was used as an evaluation metric. The results for various models and parts of the dataset are shown in Table 4.

Table 4. R2-score for different models

Regression model	R2-score for train-test split	R2-score for the remaining entries
Linear regression	0.49	0.51
Ridge regression	0.69	0.67
Lasso regression	0.66	0.67
Random forest regression	0.92	0.92
Gradient booster regression	0.98	0.98

5. Methodology: Hazard ratio estimation

5.1 Preparing the dataset

To obtain information about the roads, the same database was used, since it already contained information on the necessary features. It was sorted by individual roads and only those with one unique set of characteristics were taken. This is due to the fact that long roads consist of different parts with different properties, which is difficult to determine from road accident data. After all, 221 roads were chosen for calculation.

5.2 Hazard ratio calculation

To calculate the hazard ratio, the two most effective models based on the analysis results in (4.2) were used: random forest regression and gradient booster regression. The results of both options for the top-5 safest and top 5-most dangerous roads are shown in Table 5.

The histogram of hazard ratio values distribution is shown on Figure 6.

Table 5. Hazard ratio for different roads

Road	Random forest hazard ratio	Gradient booster hazard ratio
B-250a	0.719	0.707
BV-5005	0.721	0.724
BV-5033	0.726	0.711
C-149b	0.726	0.723
BV-2001	0.727	0.707
GI-514	0.829	0.830
T-712	0.829	0.840
TV-3322	0.829	0.850
T-233	0.829	0.845
N-1411	0.837	0.844

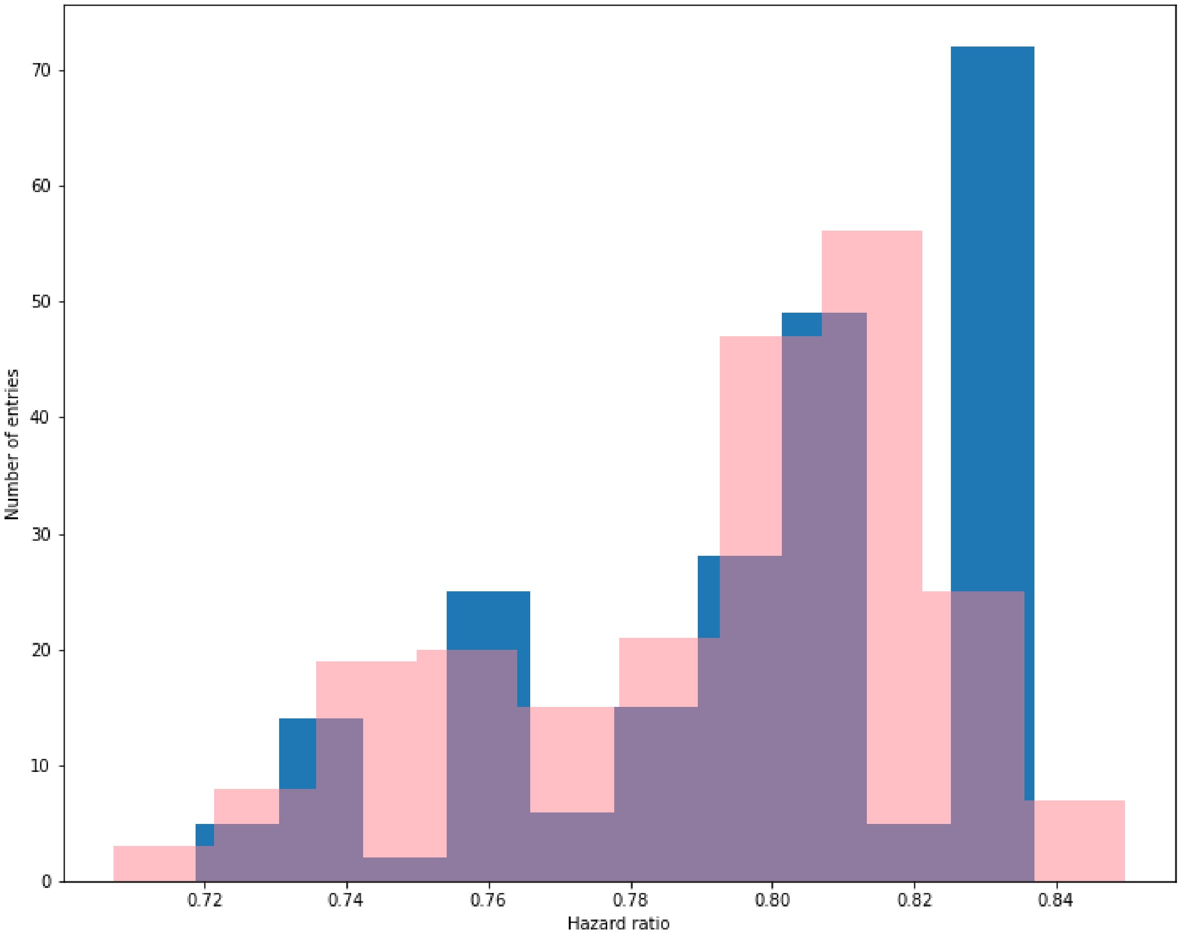

Figure 5. Hazard ratio values distribution

(bins = 10, blue – random forest regression, pink – gradient booster regression)

6. Results and Discussion

During the work, hazard ratios were determined for various components such as road conditions, weather conditions and circumstances, based on the dataset of Road accidents in Catalonia. In the presence of a large number of features, their average contribution to the hazard ratio is approximately equal, which confirms the idea that there is no universal way to improve road safety.

At the stage of machine learning, regression models were built to determine the hazard ratio based on road conditions. On average, the hazard ratio was between 0.7 and 0.9. Values less than or equal to 25% were classified as low, 25-75% as medium, and the rest as high.

The most effective were random forest regression and gradient booster regression. They were used to calculate the hazard ratio of specific roads. Based on data on 221 roads, a histogram of hazard ratio values distribution was obtained. Most of the roads are classified as high hazard category.

However, when calculating the hazard ratio of entries, most of the dataset was in the low hazard ratio zone. This is due to the presence of a large number of parts of the same road that have completely different conditions and, subsequently, hazard ratios. It may also be due to the fact that during the calculation of the hazard factor for the entire dataset, all features were used, and, therefore, the weather conditions too. Catalonia has good weather, so the contribution of positive weather conditions could artificially underestimate the hazard ratio.

7. Conclusion

The aim of this project was to build a model to determine the hazard ratio of the roads based on certain features. The results showed the need to collect similar data for all roads, and not just for actual accidents, as this way the data becomes negatively biased.

In the course of further use and development of this model, in the case of the availability of suitable datasets, it is possible to generate information on the hazard ratio of all roads in a certain region. Based on the hazard ratio data, it is possible to determine which roads should be paid special attention to over what period of time, which will prevent accidents.