Machine Learning-Based Classification According to the Reason for Travel of Drivers Injured in Traffic Accidents in Barcelona (2023). Executive Summary

Joan Manel Ramírez Jávega*

Abstract

This project, which we will present here in a summarized and concise form, originated from a series of practical exercises in the *Mineria de dades* - Data Mining- course taken by the author between March and June 2024 as part of the Master in Data Science program at the Open University of Catalonia (UOC). Later, with the goal of furthering the previously obtained results, it was continued during this summer until the present outcome, always as a personal project, independent of any institution or formal assignment.

We would also like to express special thanks for the support, advice, critiques, guidance, and encouragement provided by our two brothers, Dr. Miguel C. Ramírez and Dr. Francisco Ramírez, as well as by our colleague and friend Marcel López over the past few months as we embarked on this project. We also emphasize that any errors of omission or judgment in this or other works can only be attributed to the fallibility of the author.

1 Introduction

In a large city like Barcelona, it is common for dozens of traffic accidents to occur every day. With approximately 1.66 million registered residents according to municipal data as of January 1, 2023, and up to 2.288 million people present at some point during March 22, 2024, the most recent working day recorded in data collected by *La ciutat al dia* from the Barcelona City Council's *Oficina Municipal de Dades* - Municipal Data Office-, this also represents hundreds of thousands of vehicles of all kinds—cars, vans, bicycles, etc.—used by everyone needing to travel.

Our working hypothesis is that the variables we will study show a statistically significant relationship with whether the injured drivers, including fatalities, involved in a traffic accident were traveling for a work-related reason—either commuting to or from work, as described by the *Diccionario de la Real Academia Española* or by insurance companies like Allianz—or during the course of work itself. We find this question relevant because traffic accidents have been the leading cause of death for people commuting to or from work and the second leading cause when considering all work-related incidents in Spain from 2019 until June 2024.[1]

To test this hypothesis, the selected dataset has been subjected to both descriptive analysis and a series of machine learning algorithms, implemented using the R programming language.[2] These analyses have allowed us to confirm that this dataset is indeed relevant for improving our understanding of this issue, and additionally, the challenge of data incompleteness has been addressed by imputing, with moderate reliability, a "Yes" or "No" label in response to the analytical question of whether the driver was traveling for a work-related reason.

*ramirezjm@protonmail.com

2 Data Description

In 2023, a total of 7,721 traffic accidents occurred that required the presence of Urban Guard officers, and in most cases, this was due to at least one injured driver. This data was later anonymized and subsequently published on the *Open Data BCN platform* by the Barcelona City Council. Since we have already conducted a detailed descriptive analysis of the dataset elsewhere,[3] here we will only present the most relevant conclusions in the context of the modeling task performed.

Given that the objective is related to the reason for which an injured driver was traveling, it is important to know the distribution of these drivers based on whether the reason is known (Figure 1.1 below) or unknown (Figure 1.2 below). In this regard, after excluding drivers with incomplete or incorrect data or those involved in accidents without injuries, the reason for travel was known for 2,823 drivers involved in 2,561 traffic accidents, while the reason was unknown for another 2,196 drivers involved in 2,029 traffic accidents,[4] with both groups generally exhibiting similar characteristics. Here, we will only present their territorial distribution since a much more detailed presentation has already been made in the three parts of this study published in the RPubs repository.

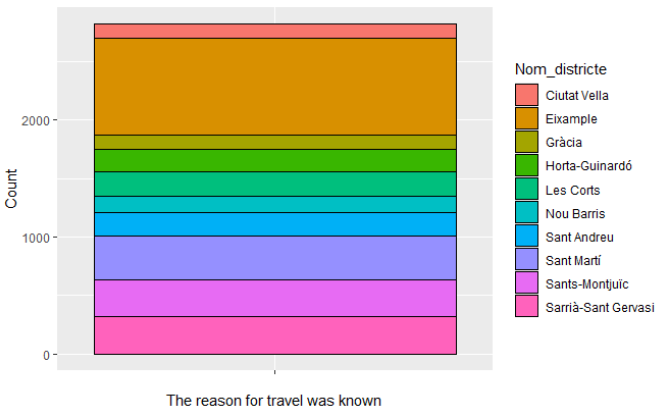

Figure 1: Stacked bar chart of the driver count whose reason for travel was previously known, based on the District where the accident occurred.

Both in the Figure 1 and in Figure 2, related to the territorial distribution by Districts of the city of Barcelona, they seem to correspond to their urban density—an emblematic case would be the Eixample—or the district's surface area. This is consistent with the fact that the districts of Eixample, Sant Martí, Sarrià-Sant Gervasi, and Sants-Montjuïc are also the four districts where the most accidents involving the Urban Guard were recorded. However, it is noteworthy that Les Corts, a district with a lower population density and smaller urbanized area, has almost the same number of injured drivers as Horta-Guinardó, which is more densely populated and more extensive. This could be explained by the fact that part of the Diagonal Avenue, one of the city's main thoroughfares, as

well as parts of the Ronda de Dalt and Ronda del Mig, run through Les Corts.

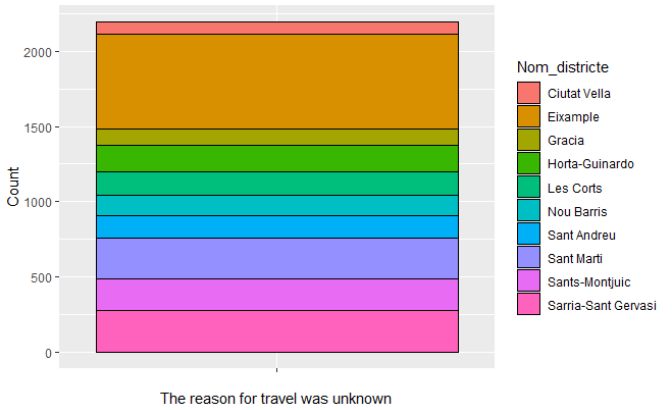

Figure 2: Stacked bar chart of the driver count whose reason for travel was previously unknown, based on the District where the accident occurred.

We will conclude this section by proposing the profile of the typical injured driver involved in an accident in Barcelona in 2023—in another context, we have already referred to the origin of the data presented.[3] In these accidents, typically, 1 person was slightly injured in the Eixample district, and it is most likely that at least one of the vehicles involved was a motorcycle if a driver was injured. This typical driver would be a man, aged between 29 and 37 years, with 5 or more years of driving license experience, requiring continuous medical attention for less than 24 hours immediately after the accident. Generally, the reason for their travel is unknown. We also add that this profile for 2023 coincides with the profile of people who died or were seriously injured, as collected by the Barcelona City Council, except that the predominant age group is between 45 and 54 years old.[5]

3 Data Description

To improve the predictive capacity of the classification model, it was deemed appropriate to enrich the selected dataset with external information, specifically the data from the Weekday Mobility Survey, which is conducted annually by the Metropolitan Transport Authority. In this case, the definition of travel schedules for occupational reasons was consulted, meaning those trips made to the workplace, educational center, or during the workday by the respondents in 2023 [6, p. 49].

The justification for this enrichment stems from the early observation, after conducting the descriptive analysis, of a bias in the data collected by the Urban Guard, as most injured drivers were also motorcycle and moped drivers.[3] This can be explained, firstly, by the fact that drivers of these types of vehicles are more likely to be injured in a traffic accident than drivers of cars, trucks, or vans; and secondly, by their widespread use in commuting within the city of Barcelona, far outnumbering bicycle or personal mobility vehicle trips.

Having identified this initial challenge, building a classification model that could be useful for determining whether a given driver was traveling for a work-related reason or not also required selecting the variables to be considered when constructing this model. This was necessary to reduce both dimensionality and noise in the

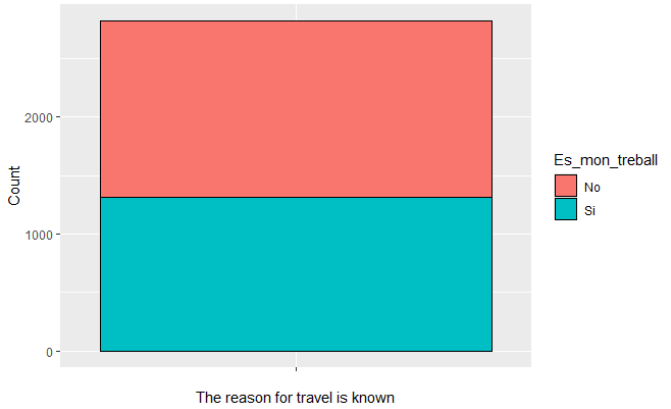

Figure 3: Stacked bar chart of the driver count based on whether their reason for travel was work-related and this reason was previously known.

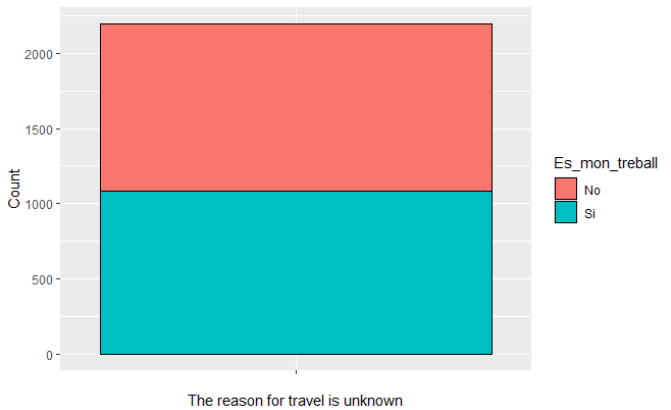

Figure 4: Stacked bar chart of the driver count based on whether their reason for travel was work-related and this reason was previously unknown.

training data, two traits that result in more reliable and easily explainable classification models.[7] In this specific study, we chose to train a Logistic Regression model using the data from the 2,823 drivers for whom the reason for their travel is known. The reason for choosing this methodology is that the method implemented in R language stands out for the explainability of its results, allowing us to determine which variables are statistically significant in explaining the model's resulting classification. Thus, out of the 29 variables considered, the Logistic Regression classification model identified only 11 as statistically significant.[4]

Once the relevant variables were selected, an eXtreme Gradient Boosting (XGBoost) classification model was built, primarily because it is the most useful in cases where class groups are unevenly distributed. Starting from the initial selection made by the Logistic Regression model, the model was refined, reducing the number of useful variables to just 9 in order to provide an optimal result in terms of accuracy (proportion of correct predictions), sensitivity (proportion of correct "Yes" predictions), and specificity (proportion of correct "No" predictions) (see Results Table, below). In the following Table 1, we detail the selected features.

It is also important to consider the selected variables as relevant, not only because the corresponding algorithm statistically evaluates them as optimal, but also because they align with what we

Variable Name	Data Type	Variable Description
"Edat"	Continuous Numerical	Complete years of age of the driver at the time of the accident.
"Tipus vehicle estandaritzat"	Categorical	Standardized into four categories for the "Tipus vehicle" variable: "Professional use vehicles", "Four-wheeled motor vehicles", "Two-wheeled motor vehicles", and "Vehicles without a driving license".
"Descripcio causa mediata"	Categorical	Details the immediate cause identified by the corresponding Urban Guard patrol that filed the accident report, referring to the type of maneuver or immediate circumstance that caused the accident.
"Numero lesionats greus"	Continuous Numerical	Number of people injured in the accident who required hospitalization for more than 24 hours.
"Nom mes"	Categorical	Month of the year when the accident occurred.
"Victimes CODIF"	Categorical	Specifies whether there were two or more victims in the accident.
"Vehicles CODIF"	Categorical	Specifies whether two or more vehicles were involved in the accident.
"Es laborable"	Categorical	Specifies whether the date of the accident was a working day in 2023.
"Es ocupacional"	Categorical	Specifies whether the accident occurred during occupational hours, from 5 AM to 4 PM, on a working day.

Table 1: Overview of the results.

know about the world of work. For this reason, it is consistent that the type of vehicle—especially professional use vehicles—or the fact that the accident occurred on a working day and during occupational hours are highly relevant for considering whether the travel was work-related. On the other hand, more intuitively, it is also coherent that the driver's age and the month in which the accident occurred are relevant, in addition to having clear analytical significance. In this regard, and thus for subsequent interpretation, it is of interest that both the number of serious injuries, the fact that the number of victims or vehicles involved in the accident exceeded two or not, as well as the immediate cause of the accident, also show statistically significant relationships with a work-related travel reason or not, although this should not necessarily be understood as the cause. It is also noteworthy that neither the variable related to biological sex nor whether any of the drivers involved had less than 5 years of experience were statistically significant when classifying the results using the Logistic Regression algorithm. The district where the accident took place was also excluded, suggesting that the administrative division of the municipality is not significant in classifying whether the driver was traveling for a work-related reason or not. In the ending Table 2, we detail the most relevant results.

It is observed that the classification model, when tested with data from drivers whose reason for travel was known, achieves an accuracy of 0.6731, while the proportion of correct "Yes" classifications is 0.6987, and the accuracy for classifying "No" is 0.6439 (see 4 in Table 2). The acceptance threshold for the "Yes" result, meaning the minimum required probability, is set at 70 per cent.

These results fall within the acceptable range for a classification model of this type [8, p. 218] and do not require a very high degree of certainty, as it is an imputation for the 2,196 drivers whose reason for travel was unknown. The relevance of carrying out this imputation lies in the fact that these drivers represented up to 43 per cent of the total injured drivers. In this way, we now have a complete set of valid data, in contrast to the previous situation where such a high proportion of "Unknown" labels rendered the reason for travel variable unreliable for analysis. Thus, we believe that this has contributed to enriching the study and knowledge of traffic accidents involving injuries where Guàrdia Urbana agents from Barcelona were involved.

Index	Description of the Result	Data Type	Results
1	Number of variables considered	Integer	29
2	Number of variables accepted	Integer	9
3	Acceptance threshold for the probability of positive classification	Percentile	70
4	Accuracy, sensitivity, and specificity	Decimal, decimal, decimal	0.6731, 0.6987, 0.6439
5	Number and proportion of drivers traveling for a work-related reason without using the classification model	Integer, percentile	1,316, 46.617
6	Number and proportion of drivers not traveling for a work-related reason without using the classification model	Integer, percentile	1,507, 53.383
7	Number and proportion of drivers traveling for a work-related reason using the classification model	Integer, percentile	1,083, 49.317
8	Number and proportion of drivers not traveling for a work-related reason using the classification model	Integer, percentile	1,113, 50.683

Table 2: Overview of the selected features.

The complete dataset, including the geocoordinates of the corresponding accident for each driver, can be downloaded at this link.

References

[1] Noriega D, Jara Y, Oliveres V. Dos muertes al día en el tajo, el drama que no cesa en España desde 2019. elDiario. 17-VIII-2024. Available from: https://www.eldiario.es/

economia/muertes-dia-tajo-drama-no-cesa-espana_1_11586949.html.

[2] R: A Language and Environment for Statistical Computing;. Available from: https://www.R-project.org.

[3] Ramírez-Jávega JM. Classificació atès al motiu de desplaçament dels conductors ferits en accidents de trànsit a Barcelona (2023) - Primera part: Selecció i preparació del joc de dades. RPubs. 13-VIII-2024. Available from: https://rpubs.com/ramirezjm/1210633.

[4] Ramírez-Jávega JM. Classificació atès al motiu de desplaçament dels conductors ferits en accidents de trànsit a Barcelona (2023) - Tercera part: Elecció del model. RPubs. 27-VIII-2024. Available from: https://rpubs.com/ramirezjm/1214120.

[5] Ajuntament de Barcelona. Dades bàsiques de mobilitat;. Available from: https://dades.ajuntament.barcelona.cat/dades-basiques-de-mobilitat/.

[6] Enquesta de mobilitat en dia feiner 2023. (EMEF 2023). Institut Metropili and Autoritat del Transport Metropolità; 2024. Available from: https://bcnroc.ajuntament.barcelona.cat/jspui/bitstream/11703/136649/2/EMEF%202023_Informe%20SIMMB%20ATM.pdf.

[7] Ramírez-Jávega JM. Classificació atès al motiu de desplaçament dels conductors ferits en accidents de trànsit a Barcelona (2023) - Segona part: Millores del conjunt de dades i modelització. RPubs. 13-VIII-2024. Available from: https://rpubs.com/ramirezjm/1210635.

[8] Bruce P, Bruce A, Gedeck" P. Estadística práctica para ciencia de datos con R y Python. Marcombo; 2022.