COLEGIO UNIVERSITARIO DE ESTUDIOS FINANCIEROS

MÁSTER EN DATA SCIENCE PARA FINANZAS

Perfilado de conductores mediante técnicas de ML

Realizado por:

D. Jose López Galdón

D. Hugo César Octavio del Sueldo

Dirigido por:

D. Juan Manuel López Zafra

CUNEF (Colegio Universitario de Estudios Financieros)

MADRID, a 18 de junio de 2021

data

MDPI

Article

Driver profiling through Machine Learning techniques

Jose López Galdón [1] and Hugo César Octavio del Sueldo [2]

[1] jlopezgaldon@outlook.com
[2] octadelsueldo@gmail.com

Abstract: Traffic accidents represent a high cost for insurance companies as well as for society, in economic and social terms, because in all cases the costs include medical and rehabilitation expenses, legal and emergency services, property damage and production losses. Thanks to the use of telematics and data science we may be able to find patterns of behavior that explain the claims. During this research we will work with a database of more than 95,000 drivers that includes information collected over 6 years; for this, we have performed an important work of cleaning and engineering of variables, for finally clustering the drivers through a PAM, being the most representative variables the intensity of use of the vehicle and the driving experience. In addition, we have made a prediction based on whether or not they have suffered a crash using a decision tree, obtaining a 72.25% accuracy rate.

Keywords: insurance; data science; supervised and unsupervised algorithms; machine learning; clustering; decision tree.

1. Introduction

Insurance is an effective way of protecting individuals against the consequences of risks. It is based on transferring the risks to an insurer who is responsible for compensating all or part of the damage caused by the occurrence of an event. A fair price will accurately reflect the actual risk of the insured, otherwise the business may have problems in the client portfolio, since those clients who have less risks end up supporting those who are riskier, which causes a client´s churn with lower loss ratio and an entry of those with high loss ratio.... [1-4]

This directly harms insurance companies, policyholders and, consequently, society. Therefore, actuaries use risk-related information, i.e., they use those factors that model risk for insurance premium calculation, so that they can construct league tables based on expected losses.

In the case of automobiles, the traditional variables for determining the risk profile are personal characteristics, claims history and vehicle characteristics. However, premiums are often inaccurate in practice, as these factors do not have a direct causal relationship with actual driving risk.

Thanks to the development of networks, connectivity, IoT... UBI (Usage Based Insurance) products are increasingly popular within insurance companies. This insurance product model is based on schemes known as "pay-as-you-drive" (PAYD) and "pay-how-you-drive" (PHYD); the determination of the premium is based on variables determined from the actual data of the driver, such as driving time, distance traveled, type of roads traveled, speeds.... [5-7]

Thanks to the use of these technologies, the benefit is twofold, since the insured obtains a price that is more in line with his behavior, so that drivers with lower accident rates will not be penalized, and, on the other hand, insurance companies can effectively improve the accuracy of insurance prices. In addition, UBI products encourage drivers to

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Data* **2021**, *6*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

drive less or improve their driving habits, as they can benefit from the use of dynamic premiums.

This coupled with the use of data science makes it possible to mine the information collected by the devices, so that insurers can fine-tune their risk models. Insurers can model driver behavior and therefore able to predict future claims based on what has happened in the past.

According to studies on the subject, such as Ryan et al. (1998) [8-11], young drivers are those with the highest accident rates, which is why we will focus on this profile of drivers, specifically, drivers between 18 and 30 years of age.

As we have already mentioned, the main objective of our research is to segment young drivers based on their driving characteristics, driver information, such as age and vehicle qualities (power, weight, etc.). This segmentation (unsupervised algorithm) will help to classify drivers, using a decision tree (supervised algorithm), based on their accident rate, for which we will use the proxy of crashes (decelerations of more than 4G).

2. Data Description

The data used for the research come basically from a tier 1 Spanish insurance company; conveniently anonymized, we have a set of personal data of each driver, such as date of birth, sex or area of driving; we also have a set of personal data of the vehicles, such as make, model, weight or power; finally, we a last set of variables that mix both driver and vehicle characteristics such as the number of daily trips or journeys, the distance and time spent in the journey, maximum, minimum and average speeds, date and time, and, in case there's any, the value of the deceleration in g (depicted as G in the text); g stands for g-force, the measurement of the type of force per unit mass – typically acceleration – that causes a perception of weight. One unit of g-force (1 g) equals to the conventional value of gravitational acceleration on earth, g, of about 9.8 m/s

For the purposes of this research, we will understand as trip, journey, or move of a driver everything that occurs between each start and stop of the vehicle. Indeed, according to the technical characteristics of the electronic device for collecting information, it begins to transmit its position and to collect technical characteristics (in terms of speed, travel times, types of roads traveled, etc., which we will point out later) as soon as the ignition is turned on, ending when the engine is switched off. This is going to cause a certain level of "noise", from the moment that there are trips of zero distance, at null speed; obviously, one of the first tasks carried out consisted of the debugging of this and other types of errors that we will later report.

In total, we have more than 54 million trips, corresponding to 97,544 different drivers (rows) and 45 variables (columns) for a time window between April 11, 2007, and December 31, 2013.

In the case we are dealing with, our database was filtered with the possible existing errors, eliminating those records that met at least one of the following conditions:

- The average speed is greater than the maximum speed.
- The average speed is negative or higher than 200 km/h.
- The maximum speed is negative or higher than 250 km/h.
- The kilometers traveled are greater than 1,200 but less than 0.1.
- The duration of the trip is less than two minutes or more than 15 continuous hours.
- The parked time is negative.

We must consider that for each driver we know, for each trip, what was the maximum speed reached and what was the average; by obtaining the aggregate figures for

each driver, we are able to know their behavior patterns, as well as the maximum and minimum speed, average distance traveled per trip, average time used for this trip, record count, days since they have been in the product and others.

2.1. Variables

Below, we will detail some of the most important variables we have used for our research:

- Gender (categorical): Dichotomous variable indicating the gender of the user, distinguishing between male and female.
- Groups of autonomous regions (categorical): performing a prior clustering according to the driving features of the drivers (maximum speed, distance and time of the journey), we defined three main areas Central Zone, East and South, North Zone and Islands.
- Count (numerical): Total records for each driver.
- Days (numerical): Number of days registered between the first and the last journey of the driver.
- Weight to power ratio (numerical): The usual ratio in Spain, the inverse of the power-to-weight ratio.
- Age (numerical): Age of the driver at 12/31/13
- Experience (numerical): License seniority in years, at 12/31/13
- Brand (string): Vehicle brand
- Avg max speed (numerical): Average of the maximum speeds (km/h) of each driver
- Avg average speed (numerical): Average of the average speeds (km/h) of each driver
- Average distance (numerical): Average distance of each driver's journey in meters
- Average duration (numerical): Average duration of the trip in seconds.
- Number of crashes (numerical): Number of decelerations higher than 4G.

2.1.1. What is a crash?

One of the key variables in this research is crashes. We will use the term crash to indicate the existence of a significant deceleration, and so a proxy of the accident, of the analyzed vehicle. A crash of more than 4g is considered an "accident warning", according to the information provided by a company specialized (private email) in the analysis of accidents, which placed the average number of accidents without a tow truck at 4.13G. So, in the lack of the actual information of the existence of an accident, we have used the value of 4g as a conservative threshold of an accident.

2.2. Exploratory Data Analysis

Exploratory data analysis refers to the critical process of conducting initial investigations of data to discover patterns, detect anomalies, test hypotheses, and test assumptions with the aid of summary statistics and graphical representations. [12]

For this reason, we will now carry out an EDA to have a better understanding of the variables, as well as to lay the foundations of our analysis and help us for subsequent stages such as data processing. Given that we have 45 variables, we will focus on those that are most important.

2.2.1. Gender

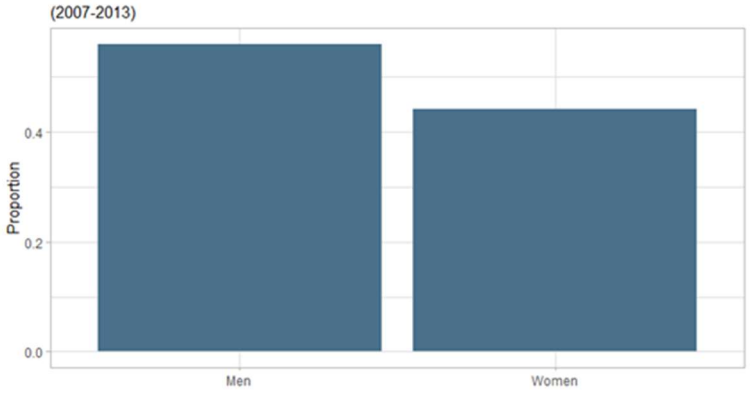


Figure 1. Gender distribution

Figure 1 represents the share of men and women in our database; most drivers are men. In table 1 we have the detail, so that men represent 55% and women almost 45%.

Table 1. Gender

Category	Count (n)	Share (%)
Men	54,531	55.94
Women	42,947	44.05

2.2.2. Group of autonomies

As mentioned above, there are 3 groups:

- 1: corresponds to the provinces in Cantabrian coast and Balearic and Canary Islands.
- 2: central Spain.
- 3: east and south of Spain.

As we can see, the group with the most drivers are 3 with almost 55% of the population, followed by 2 with more than 26% and 1 with barely 20%. In this case the variable has no nulls, so we do not have to treat them.

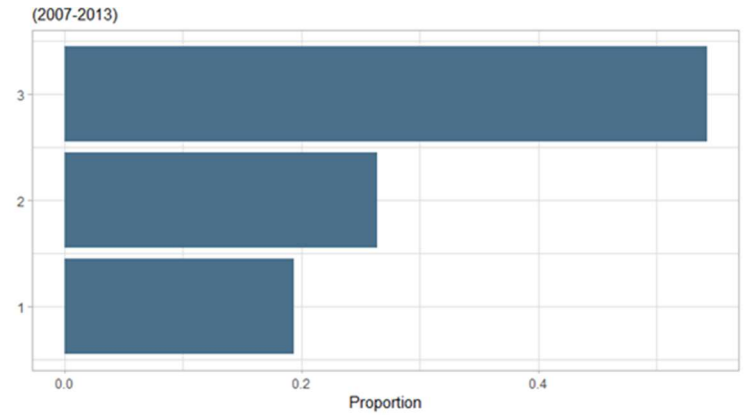


Figure 2. Share of drivers by autonomous regions' groups.

2.2.3. Count

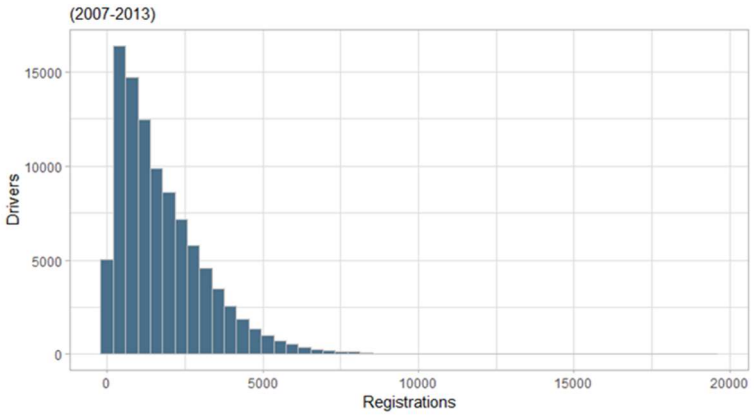


Figure 3. Count

Figure 3 shows the distribution of the records, i.e., the number of times a user has started the car. Sometimes it does not mean a trip since the driver can simply start the car and turn it off with no move. It presents a geometric-like distribution since the data are concentrated in the left part of the distribution. In this case, we see that many of our drivers are in the first bar; this will later be addressed.

2.2.4. Days

In this case, the distribution of the variable days seems to follow a bimodal distribution, since it has two nuclei where the data are concentrated. We found a first group of drivers with less than 500 days of seniority, that is, less than a year and a half, and the rest of the drivers, who have more than a year and a half of seniority.

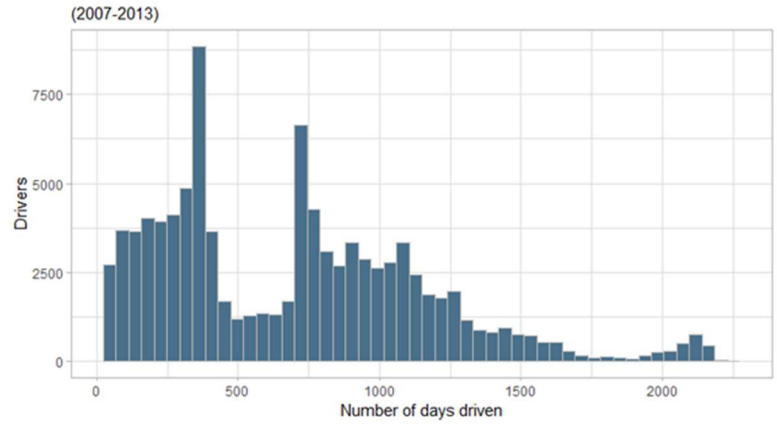


Figure 4. Days

2.2.5. Weight-to-power ratio

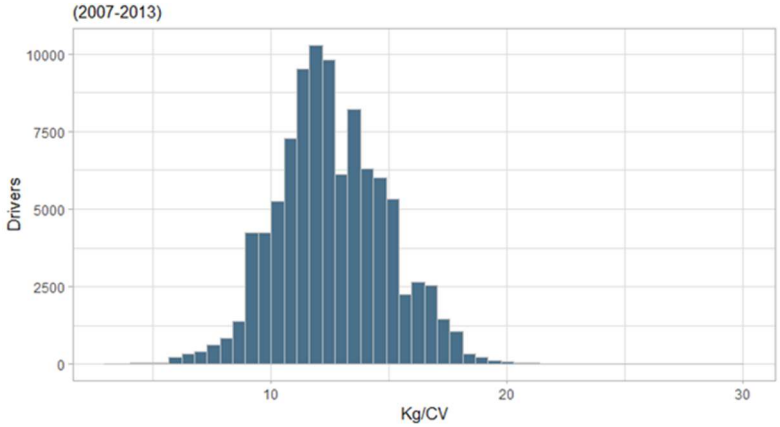


Figure 5. Weight-to-power ratio

In this case the less kg/hp, the more acceleration a vehicle will have. The minimum is 3.4 kg/hp, in this case it is a vehicle with a very good power-to-weight ratio, typical of a sports car or a motorcycle (it may be the vehicle with 555hp). On the other side we have a maximum of 30 kg/hp, this is a very slow vehicle, such as a 3.000 kg vehicle with 100hp (the average weight of our vehicles is 1.200 kg). The average is at 12 kg/hp with a low deviation, just 2 kg/hp.

2.2.6. Age

The variable Age refers to the age of the insured driver as of 12/31/13. As we can see in Figure 6, the data are concentrated in the minimum data, and there are some extreme values in the right tail. Clearly, the data are concentrated between 20 and 30 years old, being this a database of young drivers, this will be key since one of our main objectives will be the study of less experienced drivers; the filtering process to keep just the youngest drivers will be later addressed.

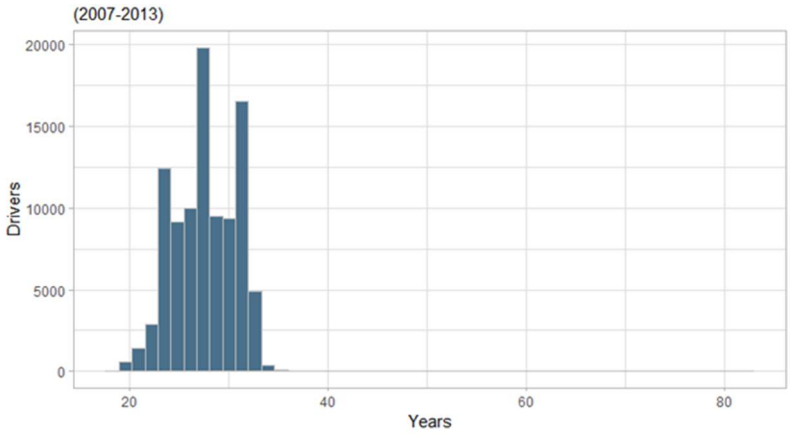


Figure 6. Age

2.2.7. Drivers' experience

The data are concentrated between license ages between 0 and 20 years, making sense since the drivers are between 18 and 40 years old.

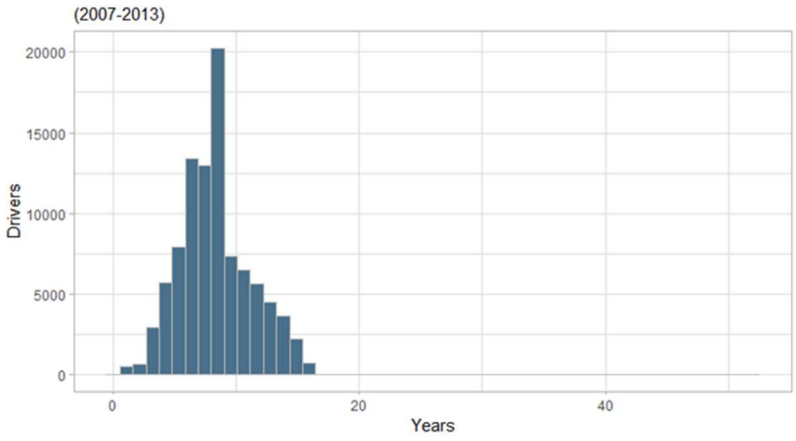


Figure 7. Drivers' experience

2.2.8. Brand

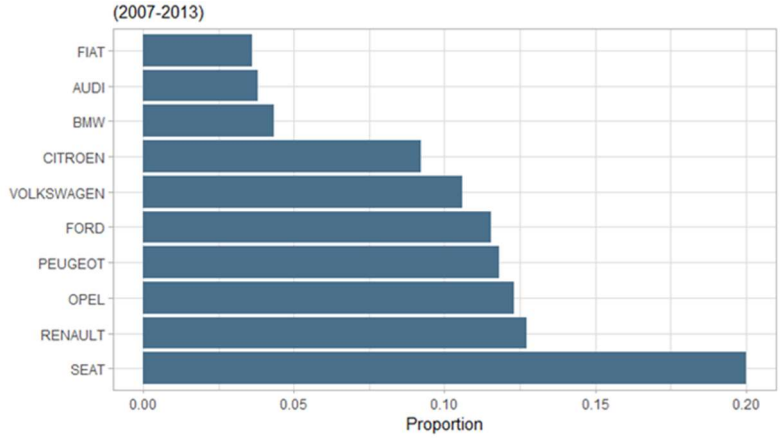


Figure 8. Brands

As Figure 8 depicts, most of the drivers own broadline brands such as SEAT, RENAULT, OPEL, PEUGEOT, FORD, VOLKSWAGEN... This makes sense, as these are young drivers who drive non-premium cars.

2.2.9. Average maximum speed

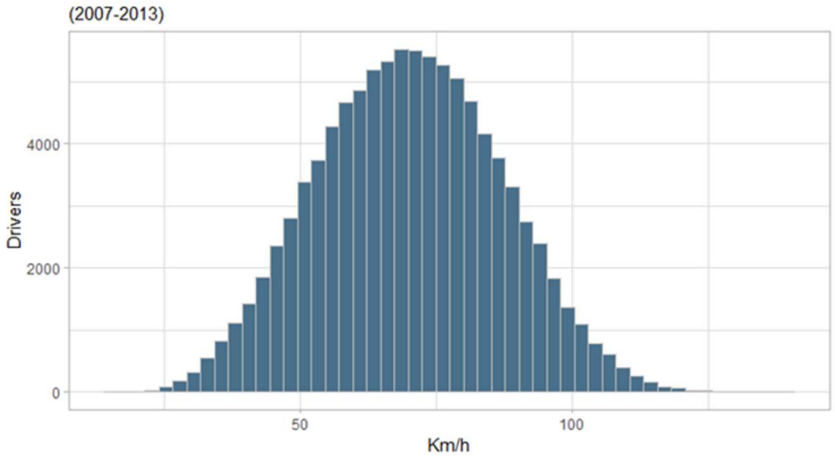


Figure 9. Average maximum speed

As we can suspect, the average maximum speed seems to fit a normal distribution. The minimum is found to be 15km/h, which is a very small figure. The maximum may also be an outlier, or the driver always drives at a maximum of 140km/h. As for the mean, we find it at 70 km/h with a standard deviation of 16km/h.

2.2.10. Average average speed

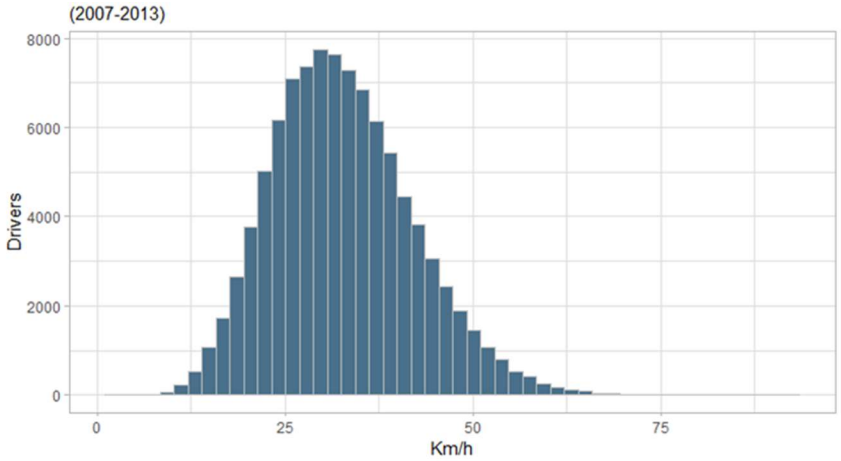


Figure 10. Average average speed

This variable stands for the average of the average speeds performed by each driver. Once again, even if slightly skewed to the right, it seems to fit a normal distribution. In this case, we find a very low minimum of 2 km/h, this may be due to a person who has barely moved the vehicle. The maximum is at 92km/h. The average is 32 km/h with a high standard deviation, close to 10km/h.

2.2.11. Average distance

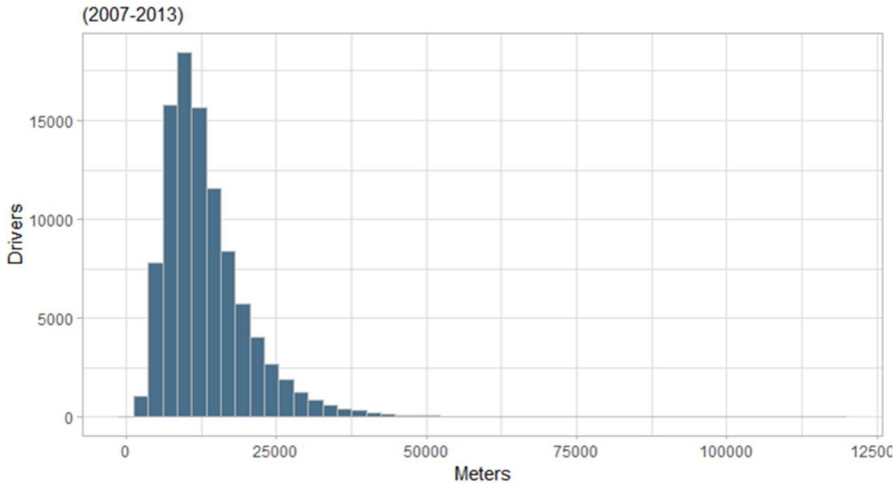


Figure 11. Average distance per trip

We have found that the minimum average distance is 400m, while the maximum is 119.110 (120km approximately). The mean is 13.3986 meters, around 13 km, with a quite high standard deviation, since it reaches up to more than 7km.

Once we have described the most important variables in the database, we enter now in the description of the methods in place for the present research.

3. Methods

3.1. Feature Engineering

3.1.1. Delete variables.

Once we have carried out an extensive exploratory analysis in our dataset, we already have a clear idea of the treatment to be carried out in each of the variables. Starting out with feature engineering, we proceed to eliminate a subset of twenty-one variables that do not provide any relevant information [13-14].

On the other hand, there is a series of variables already included in other variables; to avoid multicollinearity problems, we will eliminate them. It is the case of the tare weight and the power of the vehicle, already present in the *weight-to-power* ratio. Another similar case is related to speeds and distances; in this case we have decided to eliminate the maximum and minimum speeds of drivers. Likewise, we have eliminated the date of birth and the date of issuance of the driving license as we already use the age of the driver and the age of the license.

3.1.2. Nulls treatment

Regarding nulls, we have eliminated the null records of those variables that had a negligible number of missing values. Within these are the gender, the age of the license and the age that had 0.08%, 2.02% and 0.02% of missing values. This decision was made knowing that mostly no information was lost when deleting these rows.

Then, as for the rest of the variables, we choose for filling in with zero, knowing that most of the variables that did not have these data referred to either distances or meters traveled on roads of different capacities or speeds which makes all the sense to replace by zero in case of lack of information.

3.1.3. Transformations

To work in a comfortable way, we have transformed the variables of distances to kilometers (km) from meters and duration to minutes from seconds. Original variables have so been removed from our dataset.

3.1.4. New variables

As a first step, we will generate three types of roads based on the maximum driving speeds. These will be the so-called “high-capacity roads”, “urban_ways” and “rest_highways”. For the first case, we have added road type one and road type two where these routes refer to highways and dual carriageways. In the second case, we have added type two (national highways), three (regional highways), four (sub-regional highways), five (highly important local highways), six (secondary urban network) and seven (minor local highways). In the third case, we have left the sum of the type eight roads (Rest of roads not suitable for the circulation of vehicles) and type X in a separate group. Likewise, we will convert the distance traveled by type of road into kilometers and finally, we will remove the original variables.

Then, we have built a dichotomous variable, which indicates whether the driver has one of the most used brands or another. Among the most used brands in our dataset are Seat, Renault, Opel, Peugeot, Volkswagen. As mentioned, the rest of the brands were treated as "Rest of Brands".

We have also created a variable that indicates whether the driver has crashed or not. This variable will take one when there is at least one crash, or zero if there are no crashes in its records. This variable will be our target when making our classification model to predict those drivers who are going to crash or not. Continuing with similar variables, we have created a new variable that indicates the number of crashes discretized by "None", "One", "Two", “Three or more”.

In addition, we have created a new variable that represents the average total distance driven by the user. For this we will multiply the average distance by the total number of records of the driver.

Regarding the intensity of daily use of the vehicle, we have calculated it by dividing the number of records by the total number of days driven.

Likewise, we have calculated the number of days since the last start to know if it is a user who uses the car frequently or not. To do this, we will subtract the last date of the database (12-31-2013) from the date of the last startup.

In addition to this, we have calculated the annual distance traveled by the user; to do so, we will calculate the distance divided by days (total_distance_dia) and multiply by 365 so that we obtain the annual km of the driver.

Once the annual distance effectively traveled by each driver has been identified, we proceeded to make two corrections on it. The first, multiplying it by the years of experience of the driver elapsed since obtaining the driving license, applying a correction factor from the 4th year of license equal to the square root of the excess over three years, to define the variable kilometers of equivalent experience (km_equiv). We introduce here the concept of equivalent kilometers; the idea rests on the fact that an annual distance of 5,000 kilometers is not the same for a novice driver as it is for one with 10 years of experience. For this, it is necessary to make some correction of the experience. Figure 12 shows the effect of different correction factors for a driver who does 1,000 kilometers a year. If we do not consider the age, the 1,000 kilometers will always be the same (red line); On the other hand, we consider that a driver who has 4 years of experience and who has traveled 1,000 kilometers this year is as if in total he had accumulated 4 * 1000 = 4000 km (green and purple lines); this does not always hold because it seems quite clear that the mere accumulation of kilometers does not suppose, from a given moment, a proportional increase in experience; thus, from the third year on, we corrected the kilometers traveled in the year by a softer factor (purple line) than that of direct proportionality (green line).

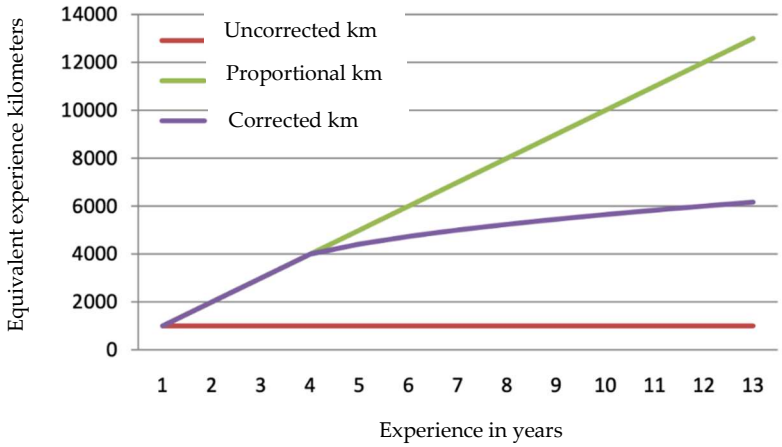


Figure 12. Experience in years vs Equivalent experience kilometers

Thus, for all intents and purposes, a driver with a 12-year-old license who has completed 1,000 annualized kilometers is as if he really had accumulated a total of 1,000 x (3 + √ (12-3)) = 6,000 km.

Once we have created the new variables and verified that we have not generated null values, we will filter by age following the objective of studying young drivers. It is here where we will only keep the users aged over 18 years and under 30 years and with an age of the license between 0 (less than 1 year of experience) and 13 years.

In addition, to avoid additional distortions, it was decided to analyze exclusively those drivers who presented a history in the file of more than thirty days and who, at least,

in that time, had made a minimum of two daily trips on average, that is, a minimum of 60 records in the 30 days of analysis. The reason for this is, again, the attempt to preserve the coherence and integrity of the data to be analyzed, since it was estimated that presence times of less than 30 days would not allow to derive precise behavior patterns; likewise, a driver who is on the database for more than 30 days but barely uses the car (a minimum of 60 trips in the observed time) may, or not, have a technical or commercial interest, but it is not possible to derive any behavior from her.

Thus, after this last decision, a total of 71,540 drivers were studied, that is, 73.33% of those originally present in the original registry base.

3.2. Behavioral patterns

One of the main objectives of this research is to find out if there are relationships between the incidents recorded by the driver (Crash) and certain variables related to road safety such as the intensity of vehicle use, gender and age of the driver or age of the license.

3.2.1. Daily usage intensity analysis

One of the variables that we have generated during our engineering has been the intensity of daily use; we addressed this issue through the share of the total number of trips that the driver has made to the number of days she has been in the program; for example, the intensity of use is different for a driver who uses the vehicle for 100 days to go to and from work than another one who, for only 50 days, also drives home for lunch every day. In both cases the total number of trips (assuming that in all the recorded days you would have gone to work) amounts to 200, however, the first driver has an intensity of use of 2 trips per day compared to 4 for the second. We will refer to this measure of intensity of use below.

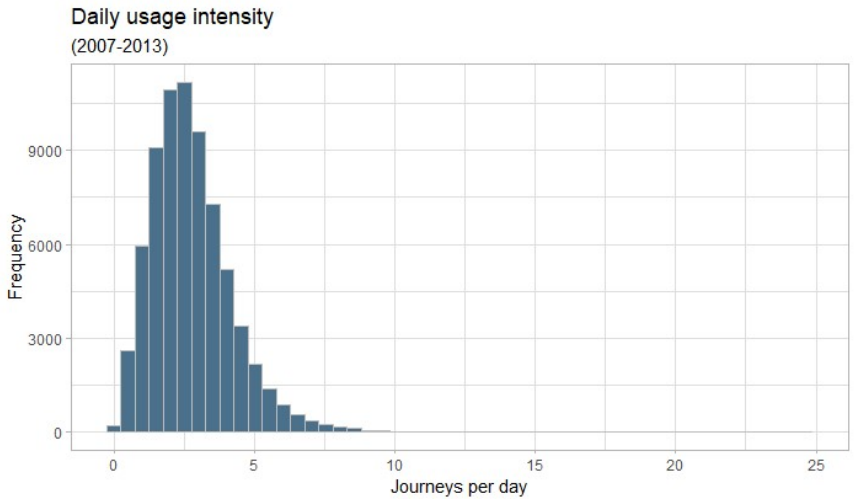


Figure 13. Daily usage intensity histogram

Looking at Table 2, the average number of daily trips is 2.78 with a standard deviation of 1.4628, showing a high variability. Drivers tend to use the car between 1 and 4 times a day.

Table 2. Daily intensity usage summary

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Daily usage intensity	0.0556	1.7642	2.5741	2.7830	3.5351	24.7104

The analysis of the data lead us to set the following hypothesis: the higher the intensity of vehicle use, the more exposed the driver should be to accidents, and therefore the higher the accident rate. To study this hypothesis, we will work with the quintiles of the trip, through an ascending sorting of the intensity of use variable and defining groups for each 20% of drivers.

3.2.2. Intensity of use & crash

As Figure 14 depicts, there is a clear relationship between the number of crashes and the intensity of use; as Table 3 shows, the increasing number of crashes with the increasing quintile suggests that a higher intensity relates to a higher number of crashes.

Table 3. Average crashes per quintile

Quintile	Average crashes
1st	0.4092116
2nd	0.5642997
3rd	0.6745876
4th	0.8403103
5th	1.0348081

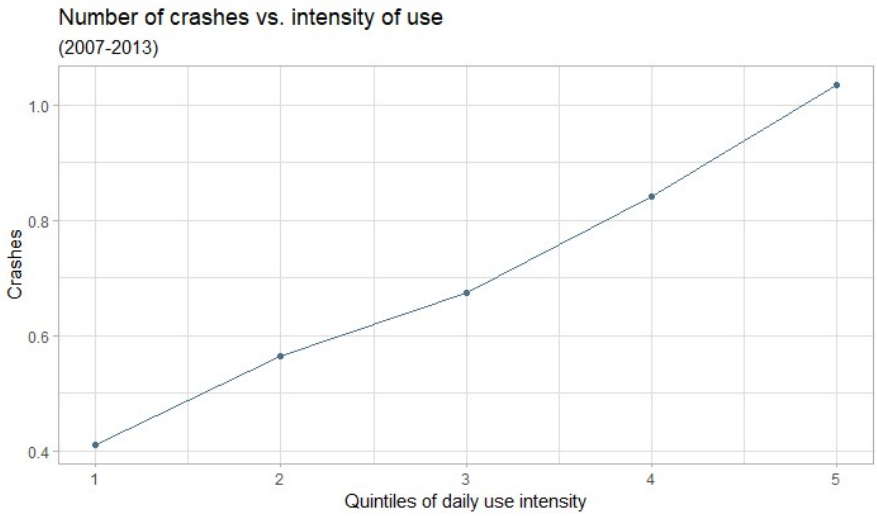


Figure 14. Number of crashes vs. intensity of use by quintile

3.2.3. Who are the drivers who make most use of the vehicle?

We enter now in the analysis of the drivers with the highest intensity of use.

As Figure 15 shows, men are the ones who use the vehicle the most, so, a priori, they will be the ones with the highest accident rate.

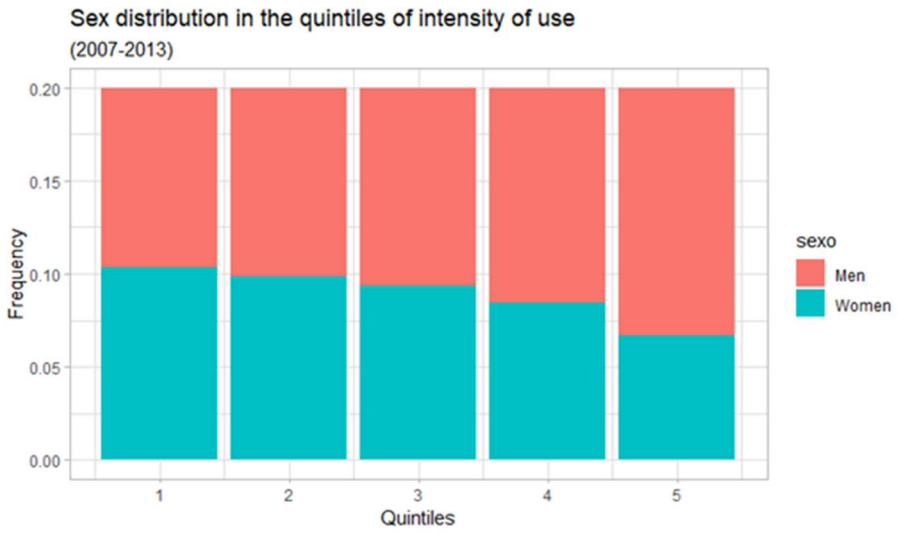


Figure 15. Gender distribution in the quintiles of intensity

To observe in greater detail whether there are differences in the crashes by gender and intensity of use, we have combined the previous plots in Figure 16.

As we can see, in any case, men have more crashes than women, the differences seem significant except for the third quintile. Therefore, we can affirm that the higher the intensity of use, the higher the crashes and, in the case of men, they are consistently more likely to have more crashes than women.

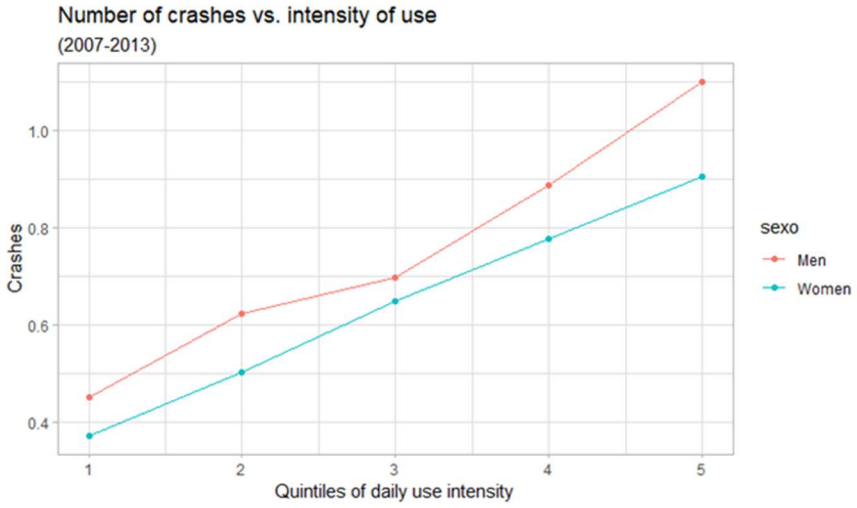


Figure 16. Number of crashes vs intensity of use by sex

We will now check the average experience (measured in years) against the quintiles of experience. We can in Figure 17 see that the 20% of drivers who use the vehicle the most are young drivers with less than 7 years of experience, compared to the others with more than 7 years. A priori, we could consider that these differences are insignificant or not very relevant, since the greatest difference between the quintiles is barely half a year. However, the scope of this research is young drivers, that is, those between 18 and 30 years of age.

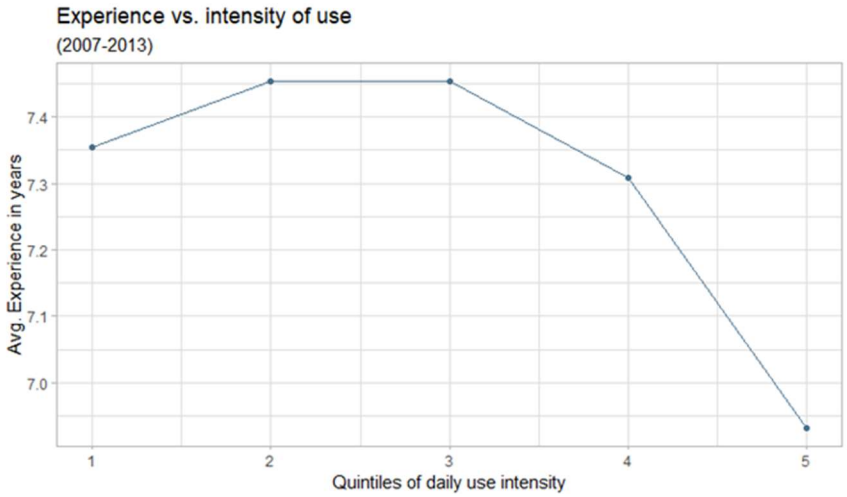


Figure 17. Experience vs intensity

It is true that in this case there seems to be differences only with the least experienced drivers (fifth quintile of intensity).

Table 4. Average experience in years by quintile of daily use intensity

Quintile	Average Experience in years	Upper limit	Lower limit
1st	7.354697	4.802208	9.907185
2nd	7.454012	4.965445	9.942578
3rd	7.454361	5.024163	9.884560
4th	7.307918	4.883101	9.732736
5th	6.930943	4.495996	9.365889

Finally, we decided to test whether there are differences with gender.

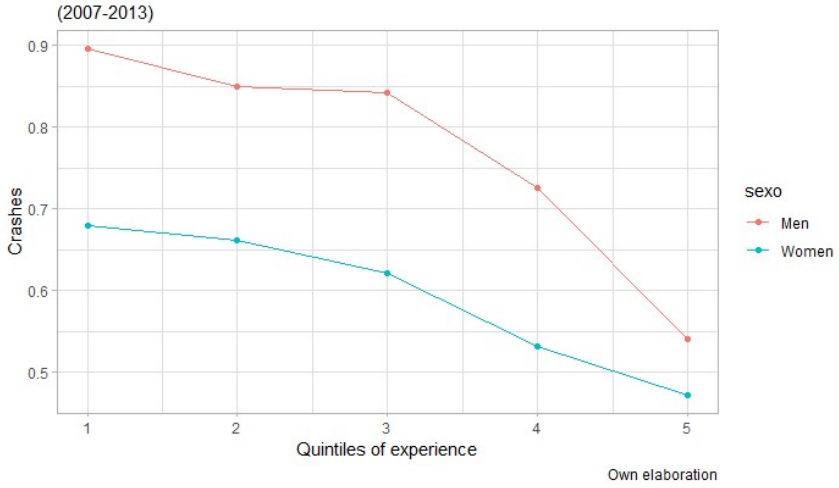


Figure 18. Crashes vs. experience quintiles by gender

In this case, we can observe how the less experienced drivers (first quintile of experience) are those with more crashes, while those with more experience (fifth quintile) have the fewest crashes. Moreover, in all cases women have fewer crashes (on average) than men in the same age groups, although it should be noted that this difference decreases in the last quintile.

We can conclude that there is a clear relationship between crashes, intensity of use, gender, and driver experience.

3.3. Clustering

We will now perform a cluster analysis with the following objectives:

- Segmenting drivers for explaining crashes according to groups.
- Defining the driver classification model through the segmentation.

3.3.1. Methodology

In our case, since we have not eliminated extreme values, the best option is working with a PAM cluster algorithm. By working with the median instead of the mean, we can avoid the problem of extreme values (the reason why algorithms such as K-Means are discarded) while we keep all the information they can provide.

The idea of K-Medoids clustering is to make the final centroids as actual data-points, resulting in a higher level of understanding [15-17].

Partitioning Around Medoids (PAM) is the one we have performed. It presents slight variations to the Lloyd's algorithm, basically in the updating step.

Steps to follow for PAM algorithm:

- Step 1 (Initialization): The initial k-centroids are randomly picked from the dataset of points.
- Step 2 (Assignment): For each point in the dataset, find the Euclidean distance between the point and all centroids. The minimum distance from the point to the centroid will be the assigning rule.
- Step 3 (Updating centroids): In the case of K-Means we were computing mean of all points present in the cluster. But the updating of the centroid by the PAM algorithm is different. For an m-point cluster, the algorithm swaps the previous centroid with all other (m-1) points inside the cluster and ends defining the point with the minimum loss as the new centroid. Minimum loss is computed by the following cost function:

$$M_1^*, M_2, \cdots, M_k = argmin \sum_{i=1}^{k} \sum_{x \epsilon S_i} ||x - M_i||^2$$

- Step 4 (Repeat): Repeat steps 2 and 3 unless convergence is achieved. Convergence refers to the condition where the previous value of centroids is equal to the updated value.

For computing reasons, we decided to use a sample of 30.000 records of the total 71,540, representing a 41.93% of the total population.

3.3.2. Optimal number of groups

Silhouette width (SW) is one of the most popular metrics when selecting the optimal number of clusters, by comparing the similarity of each point to its cluster and the one to the nearest neighboring cluster. This metric ranges from -1 to 1, where a higher value meaning a higher similarity to the pertaining cluster. Therefore, a higher value of the SW is desirable. In this case, we will compare the silhouette for a total of 2 to 5 clusters, since using more clusters could improve the silhouette but would generate higher costs when managing them [18-20]. As we can see in Figure 19, the optimum number is 2.

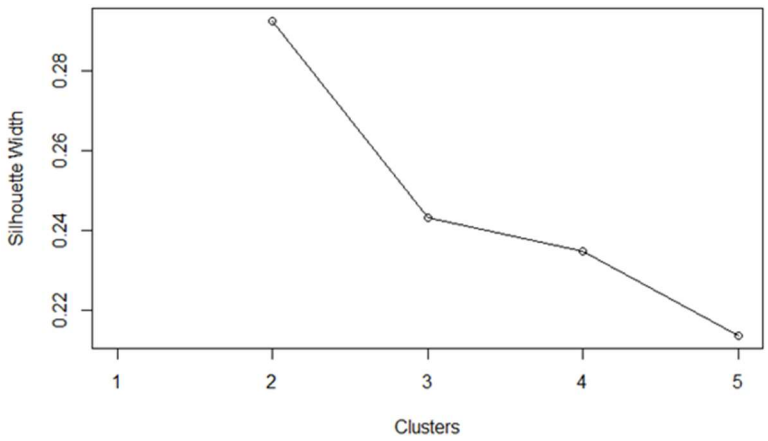
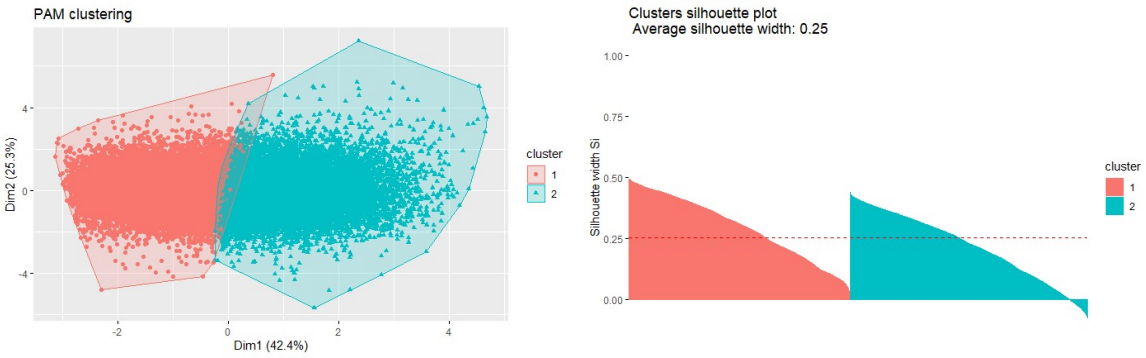


Figure 19. Silhouette width plot

3.3.3. Results



(**a**) (**b**)

Figure 20. Clustering plot: (**a**) Cluster representation in two dimensions; (**b**) Cluster's silhouette plot

Figure 20 depicts there is small overlap, an average SW of 0.25 and the first explain almost 70% of the total variance.

We can conclude that, from both the lack of overlapping, the total variance explained and the average 0.25 SW that the two clusters solution seems a good one.

3.3.3. Group analysis

We enter now in the analysis of the groups.

Figure 21 shows how driving experience is clearly linked to the group. Drivers with less than 4 years' experience belong to cluster 1, while those with more than 10 years' experience belong to cluster 2.

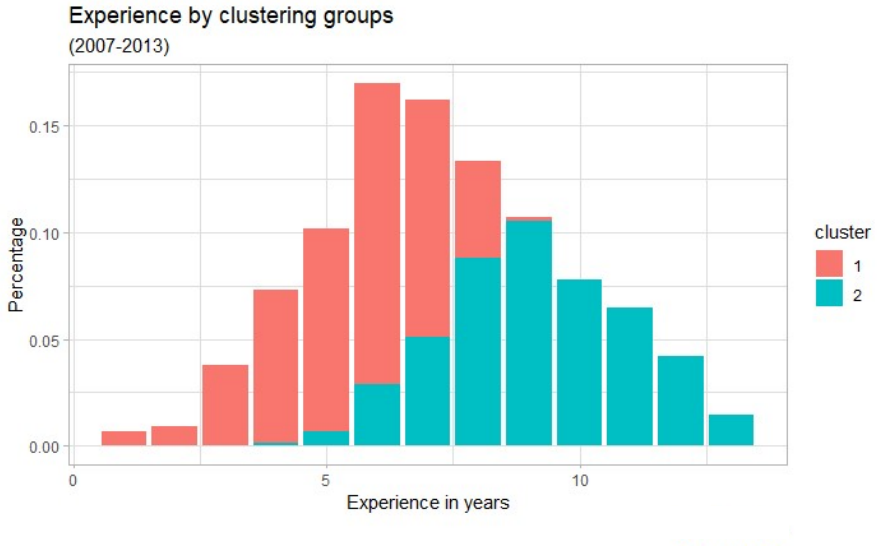


Figure 21. Experience by clustering groups

Therefore, we already see one of the big differences in the clusters, which is driving experience.

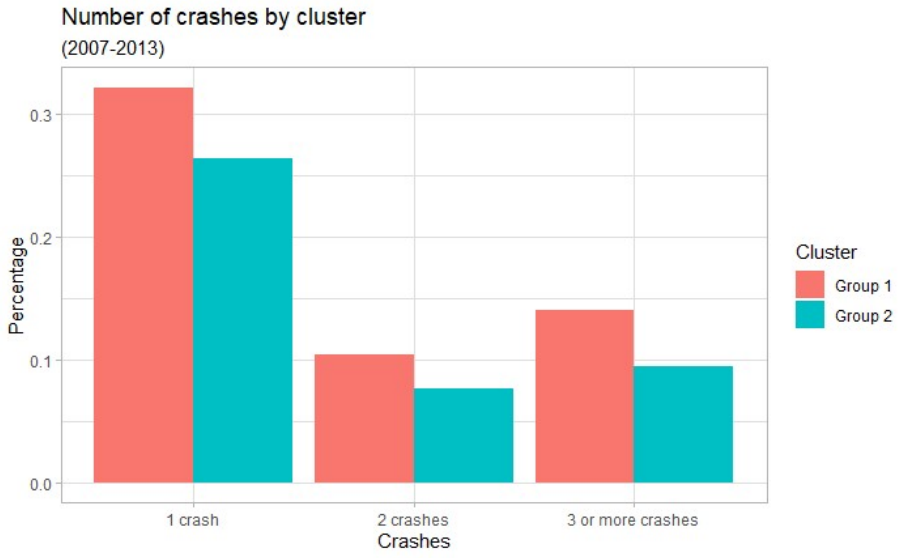


Figure 22. Number of crashes by cluster

Figure 22 shows the differences between the number of crashes by group. As we can see, group 1, the one with the less experienced drivers always retrieves a higher ratio of crashes whatever the number of them we use. The highest differences are found in the group of drivers with three or more, even if, after performing an ANOVA, the differences between the mean number of crashes per cluster are not significant.

3.3.4. First group

These are younger and less-skilled drivers; average and maximum speeds are slightly lower than those of the first group; after performing an ANOVA, we found that there were no significant differences between the speeds of the groups.; however, the intensity of use is higher than average, almost 3 trips per day. Finally, they have a higher-than-average accident rate (Table B2).

3.3.5. Second group

These are the more experienced and older drivers, with higher average and maximum speeds; a possible explanation comes from the fact they tend to drive in open roads with a higher intensity than in the urban cycle. The intensity of use (defined as times the driver uses the car per day, as previously stated) is slightly lower than in the other group. Finally, the accident rate (measured in terms of crashes) is lower in this group than in the second one (Table B1).

3.4. Crashes prediction

We have reached this final stage of the research after having carried out an extensive exploratory data analysis that allowed us to obtain insights into the behavior of drivers. This information was useful to put us in context about the variables that were important to cluster drivers based on their characteristics. This is where we will leverage these groups of drivers and their qualities to predict whether they will crash.

In context, we have made use of the Python programming language in its version 3.8.5 always using Jupyter Notebook as IDE for our research project. Likewise, we have used the pandas libraries in version 1.2.3 and numpy 1.19.2 for data manipulation, pycaret 2.2.0 and sklearn 0.23.2 for the machine learning models, matplotlib 3.3.4 for the graphic and visual part, and finally IPython 7.19.0 that will allow us to view our data file in a comfortable way.

3.4.1. Setting up the environment of PyCaret

Once we have loaded our file that contains the clusters inside, we proceed to initialize the context in which we will use PyCaret for modeling. Here we must define our target variable that will be "crash_flag", with binary values 1 or 0 whether the driver had crashes or not. Likewise, we have defined a series of variables to ignore that will not help us for the prediction of our model. We have left out a series of variables that almost fully explained the model to be used, so we decided to set them aside to create a model that should not rely on only one or a few input variables. Within this list, the most important are 'days_from_ultimo_start', 'dist_total_media_km', 'dist_total_day', 'annual_ distance', 'record_count' and 'days'.

Next, we have removed the outliers from our file. Outliers are identified by PCA linear dimensionality reduction using the singular value decomposition technique. The share of outliers is controlled by the outliers_threshold parameter at initialization. By default, 0.05 is used, which means that 0.025 of the values on each side of the tail of the distribution are removed from the training data.

Then, we have removed variables for their multicollinearity. Multicollinearity increases the variance of the coefficients, thus making them unstable and noisy for linear models. One way to deal with multicollinearity is to drop one of the two features that are highly correlated with each other. This can be achieved in PyCaret using remove_multicollinearity parameter within setup.

We have also applied feature selection based on their importance. Feature selection in the context of forecasting is the process used to select variables in the data set that contribute the most to predicting the target variable. Working with selected variables instead of all inputs reduces the risk of overfitting, improves accuracy, and reduces training time. In PyCaret, this can be achieved using the feature_selection parameter. It uses a combination of several supervised feature selection techniques to select the subset of features that are most important to modeling. The size of the subset can be controlled by the feature_selection_threshold parameter within the configuration. When the feature_selection parameter is set to True, a subset of variables is selected using a combination of several permutation importance techniques, including Random Forest, Adaboost, and Linear correlation

with the target variable. The size of the subset depends on the feature_selection_param. Generally, this is used to constraint feature space to improve modeling efficiency.

Finally, we have established a seed with the value 1322 that will allow us to replicate the results on future occasions.

Our base model will be the share of each class. So, we proceed to calculate the share of the classes based on the total count.

This share will be the benchmark we need to reach. The idea behind is that if the model classifies all the users as good drivers this will lead us to have an accuracy of 75.88% but without picking up any bad driver.

3.4.2. Choosing the number of leaves on the tree

Considering that within the objectives of the work is to make presentations to directors of insurance and car companies, we made the choice to choose the decision tree as a model to present to directors due to its ease of interpretation by people who do not have a background of technology. A classification tree is a set of conditions organized in a hierarchical structure, in such a way that the final decision to be made can be determined by following the conditions that are met from the root node to any of its leaves [21-22].

To begin with the analysis, we are going to verify the optimal growth that the tree should have. In this process we have used both the Gini index and the Entropy to choose the number of nodes. The Gini index is a measure of variability in the set of K classes of the node and the higher the purity of a node, the lower the Gini value. Entropy quantifies the disorder of a system. Here the disorder is node impurity. If a node is pure the entropy is zero being one in the case that the probabilities of the classes are the same, showing the maximum uncertainty. For running the algorithm and verifying which of the two methods we will stick with, we have crossed a decision tree model verifying the Accuracy for both the Gini index and the entropy and we verified that with three nodes we achieved the highest accuracy which can be achieved, as Figure 23 depicts [23].

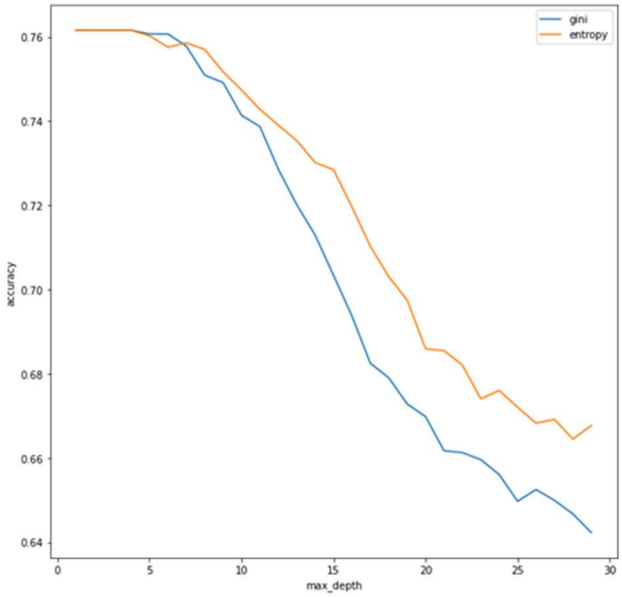


Figure 23. Accuracy vs depth

3.4.3. Training the model

Once we have verified this, we proceed to carry out our classification model defining the max_depth parameter in three and using the Gini Index criterion to select the most homogeneous nodes. After training our model, using K-folds Cross-Validation we

achieved an Accuracy of 0.7511 on average and an AUC of 0.5929, which does not meet our performance expectation. The results are presented below:

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7514	0.5772	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.7514	0.5860	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.7514	0.5928	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.7514	0.6174	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.7509	0.5789	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.7509	0.5912	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.7509	0.5878	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.7509	0.6133	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.7509	0.5699	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.7513	0.6147	0.0000	0.0000	0.0000	0.0000	0.0000
Mean	0.7511	0.5929	0.0000	0.0000	0.0000	0.0000	0.0000
SD	0.0002	0.0159	0.0000	0.0000	0.0000	0.0000	0.0000

Figure 24. Training metrics using 10 folds (cross-validation)

3.4.4. Tune Model

Once we trained our first model, we proceed to tune it, which consists of the process of optimizing the hyperparameters that the model configuration entails. We can increase the number of times we iterate our training data. Another important parameter is the "Learning Rate" which is usually a value that multiplies the gradient to bring it little by little closer to the global (or local) minimum to minimize the cost of the function. It is not the same to increase our values by 0.1 units than by 0.001, as this significantly affects the execution time of the model. The maximum allowed error of model might also be set. In our case, we have made use of PyCaret's tune_model function, optimizing the parameters to improve the AUC, the measure that did not convince us of the results of the first model. We present below the results of the tuned model.

Regarding the Precision-Recall curve, we can see that its Average Precision (AP) is equal to 0.34. This is a way to calculate the area under the PR or PR AUC curve. The Average Precision helps us to evaluate and compare the performance of models. The closer its value is to 1, the better our model will be.

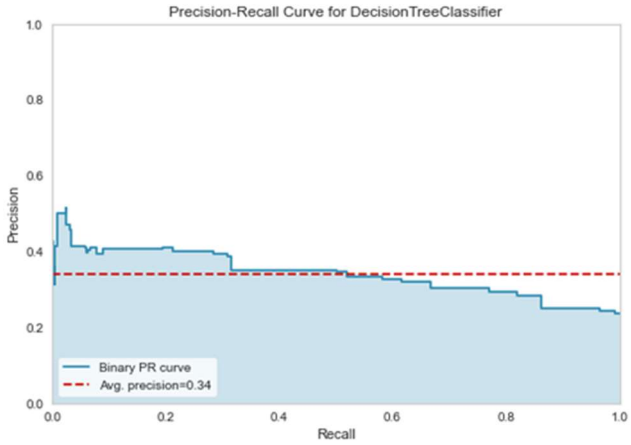


Figure 25. Precision-Recall Curve for DecisionTreeCalssifier

If we go to the importance of the selected variable, we can observe that the intensity in the daily car use represents more than 50% in the model classification decisions, as well

as driving on urban roads, the average speed and the minimum average duration are variables that influence the predictions made by our model.

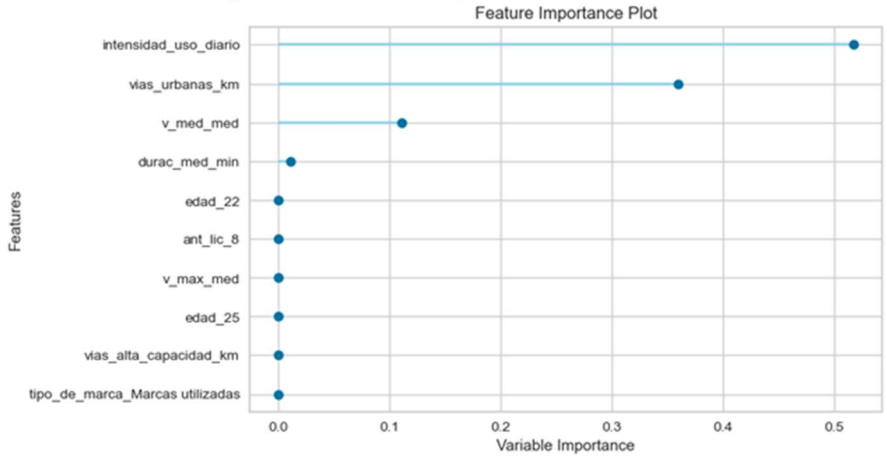


Figure 25. Feature importance plot

On the other hand, if we analyze the confusion matrix, we can see that it collects the zeros of our model very well (no crash), but not the ones (drivers that had crashes) since we get a recall of 0.07 in our dataset of test. This can give us the guideline that we should put more weight in the model to the ones than to the zeros to improve these values since, in the end, as businesspeople we are interested in having a better identification of the drivers that possibly cause us some costs than the ones that do not.

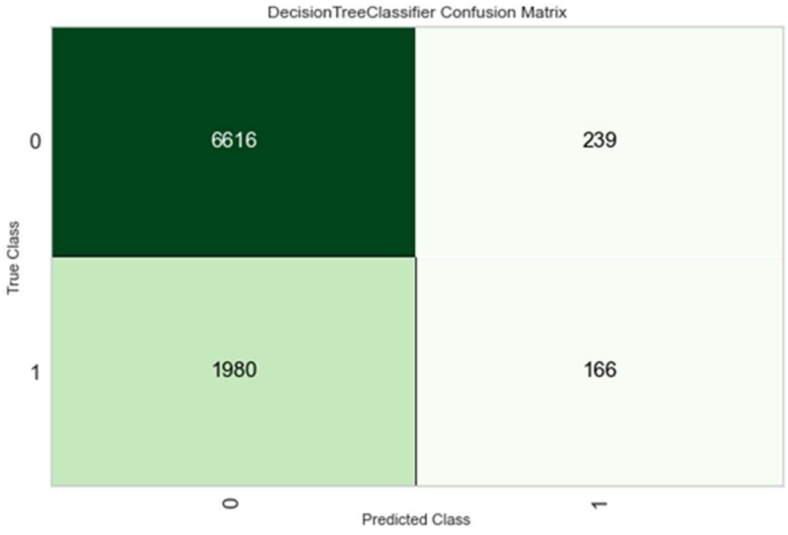


Figure 26. Confusion Matrix

Regarding the ROC AUC we see that we get an AUC of 0.65 for the zero and for the ones. In the ROC curves, we want the curve to be as close as possible to the upper left corner of the graph, so that increasing the sensitivity (the recall) does not cause our model to introduce more false positives.

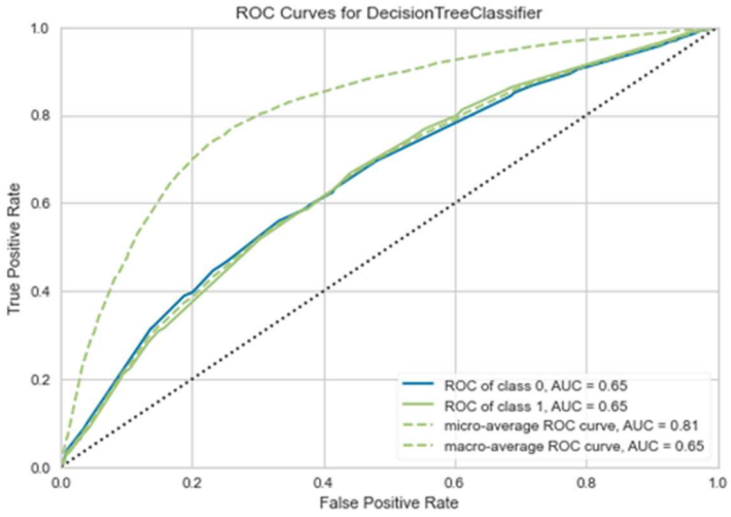


Figure 27. ROC Curves

As for our Classification Report, we can see a small summary of the results that we have been commenting on. As can be seen, the zeros in our file are well identified, which is shown in the established metrics of Precision (0.77), Recall (0.96) and F1 Score (0.85). Likewise, we can observe how the ones or people who will have a crash are not very well identified since the metrics reduce their values to 0.44, 0.07 and 0.13.

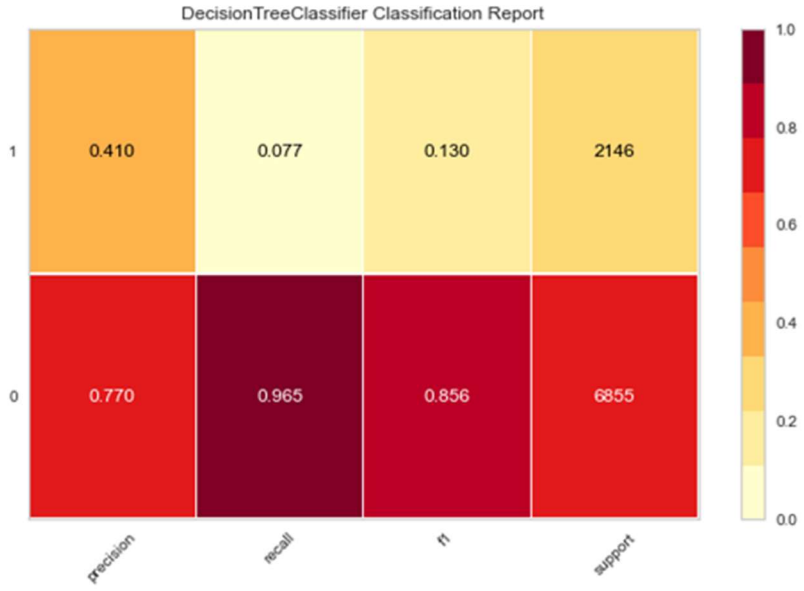


Figure 28. Classification Report

3.4.5. Interpreting the model

If we go in detail to observe the reason for the results for our first observation, we verify that this user has a probability of crash below the mean of 0.2489. Likewise, the 3,559 km driven on urban roads and the intensity of use of about 2,024 lead this person to decrease their probability of crash while the km driven on high-capacity roads increase their probability of having a crash [24-25].

Figure 29. SHAP plot

3.4.6. Alternative Model

At this point, we have considered another alternative as a way to measure the results since what interests us the most is to be able to predict even more of ones, which in view of the results, we were not achieving. So, what we proposed is to carry out a cost-based model. In our crash model the idea is to assign different costs to our TP, TN, FN and FP in order to find the model that saves us the greatest costs. This is where we will assign a profit of EUR 4000 for those true positives, that is, those that we predict as crash and are effectively crash since if we can identify them, we can charge a higher premium to these drivers despite the cost that it represents. On the other hand, we will assign a reward of EUR 2500 for those drivers who are identified as no crash and who are effectively no crash since we will give them rewards for their good driving, but at the same time we will save capital since they will not give us additional costs because they have good driving habits. The false negatives will cost EUR -3000 since they would be drivers that we have identified as negative when they were positive, that is, we are going to charge them less and they will also have crashes, so they are those drivers that we want to avoid. As for the false positives, we will put a cost of EUR 1000 since we will be charging them more than they should and also, they will not have incidents driving so it is likely that they can get other alternatives in the market, and they will go to the competition what we would lead to stop collecting those premiums.

Once we have assigned the costs, we can see in Figure 30 from Probability Threshold Optimization that the optimal probability of the threshold for the positive classifier is 0.355, which we will then pass to the predict_model function so then we can see the results of the prediction with this probability threshold.

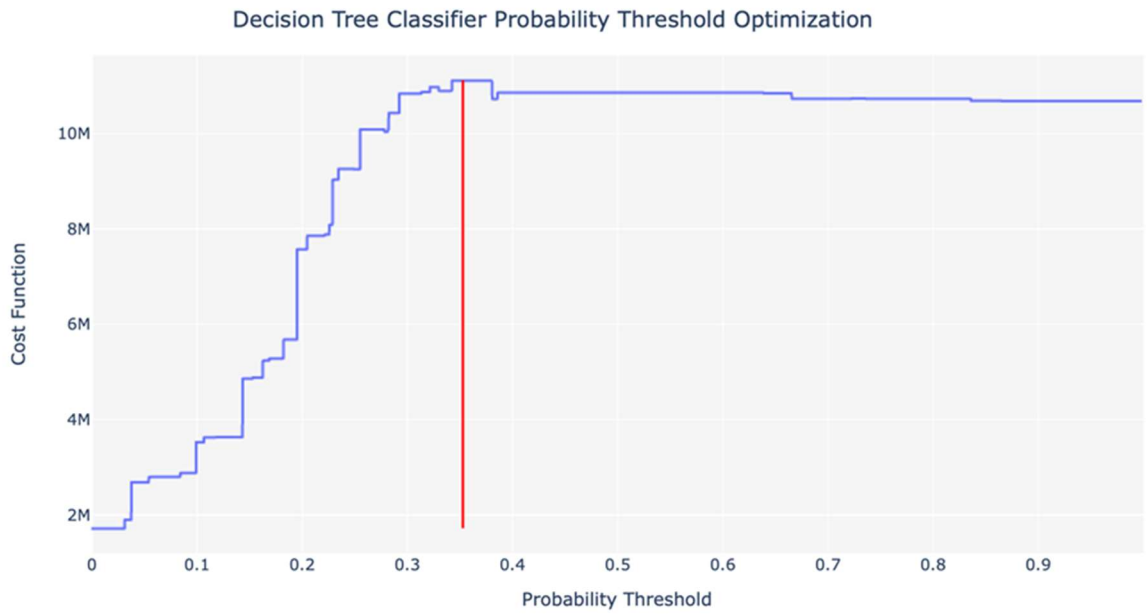


Figure 30. Probability threshold

Once we have made our predictions on the test sample, we can visualize how our indicators change.

Table 5. Decision Tree Classifier metrics

Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0.7225	0.6164	0.2195	0.3640	0.2738	0.1151	0.1207

Here we can check that we have decreased by about 4 percentage points in Accuracy and in Precision, but we have improved the Recall and the F1 Score since now we are capturing more ones in our model, and we go from a Recall of 0.02 to 0.219 while the F1 Score passes from 0.04 to 0.2738. We have achieved superior results than our first model by adding the costs to each of the classifiers according to the relative needs of the business.

4. Conclusions

After conducting extensive research on 97,000 drivers and 45 variables from more than 54 million trips between 2007 and 2013, which determine driving behavior, vehicle characteristics and driver attributes such as age, experience or gender, we can conclude that there are variables that explain the crash rate for drivers between 18 and 30 years old. Among these variables we highlight the driving experience or the intensity of use. The higher the intensity of use the higher the accident rate, this makes sense because the insured is more exposed to accidents. So, we may have a greater number of crashes. The experience is also a determinant in explaining the crashes and the less experience the greater the number of crashes. In both cases men are more likely to have accidents than women.

Thanks to the identification of these variables, we have carried out a segmentation of the drivers using a Partitioning Around Medoids (PAM). Using the elbow method, we determined that 2 is the one that maximizes the cluster silhouette. Using the Euclidean distance, we obtain an average silhouette of 0.25 and two groups of similar size:

- First cluster: These are younger and less-skilled drivers; average and maximum speeds are slightly lower than those of the first group; however, the intensity of use is higher than average, almost 3 trips per day. Finally, they have a high-er-than-average accident rate.
- Second group: These are the more experienced and older drivers, with higher average and maximum speeds; a possible explanation comes from the fact they tend to drive in open roads with a higher intensity than in the urban cycle. The intensity of use (defined as times the driver uses the car per day, as previously stated) is slightly lower than in the other group. Finally, the accident rate (measured in terms of crashes) is lower in this group than in the second one.

Using the information provided by the unsupervised algorithm, we perform a classificatory prediction of users who will or will not crash. For this, we use a decision tree, because it is one of the easiest models to interpret by non-technical people. After optimizing the algorithm, we have obtained the following metrics:

- Accuracy: 72.25
- AUC: 61.64
- Recall: 21.95
- Precision: 36.40
- F1 Score: 27.38

To achieve these results, we have assigned different weights to each of the possible classifiers in our algorithm, i.e., the False Negative will not have the same weight as the True Positive, since they will be those that we treat as low risk drivers, charging them a low premium when they will suffer a crash in the future. Beyond the results, one of the most important conclusions, which reinforces our initial hypotheses, is that the most significant variable for the tree is the intensity of use of the vehicle.

Author Contributions: Introduction, H.C.O.D.S. and J.L.G.; data description, J.L.G.; Feature Engineering, H.C.O.D.S.; Behavioral patterns, J.L.G.; Clustering, H.C.O.D.S. and J.L.G.; classification, H.C.O.D.S. and J.L.G.

Software:

- R version 4.0.5 (2021-03-31)
  - Dplyr version 1.0.5
  - Ggplot2 version 3.3.3
  - FactoMineR version 2.4
  - Factoextra version 1.0.7
  - Clustertend version 1.5
  - NBClust version 3.0
  - Cluster 2.1.1
- Python 3.8.5
  - Numpy version 1.2.3
  - Pandas version 1.19.2
  - Sklearn version 0.23.2
  - PyCaret version 2.2.0
  - Matplotlib version 3.3.4
  - IPython version 7.19.0

Hardware:

- Windows pc:
  - AMD Ryzen 5 3600 6-Core Processor
  - 32 GB RAM 3200 MHz
  - NVIDIA GeForce GTX 1660 SUPER
- MacBook Pro 13:
  - 2.3. GHz Dual-Core Intel Core i5
  - 8 GB 2133 MHz

Appendix A

Table A1. Variable summary

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Recording count	1	669	1,393	1,775	2,512	19,482
Days	1	316	707	697.4	1,008	2,240
Weight-to-power ratio	3.473	11	12.388	12.604	14.167	30
Age	18	25	28	27.78	30	82
Experience	0	6	8	8.329	10	52
Avg Max Speed	15.13	58	70	70.06	81.88	139.66
Avg Avg Speed	1.284	25.884	31.894	32.894	38.602	92.135
Avg Distance	412.7	8,443.2	11,699.3	13,396.6	16,466.8	119,110
Avg Duration	164.4	993.2	1,186	1,252.3	1,424.1	30,637.2

Appendix B

Table B1. Old drivers (second group)

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Intensity of use	0.08409	1.71065	2.45232	2.60764	3.30480	17.1380
Weight-to-power ratio	3.614	10.850	12.111	12.330	13.800	24.390
Age	25.0	27.0	29.0	28.4	30.0	30.0
Experience	4.00	8.00	9.00	9.165	11.00	13.00
Avg Max Speed	21.72	59.59	71.76	71.65	83.46	137.22
Avg Avg Speed	7.258	26.791	32.892	33.637	39.680	80.162
Avg Duration (minutes)	3.024	16.712	19.958	20.967	23.967	309.138
Avg Distance (km)	0.8713	8.7332	12.2202	13.9946	17.1977	113.242
Crashes	0.0000	0.0000	0.0000	0.6049	0.0000	91.0000

Table B2. Young drivers (first group)

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Intensity of use	0.06443	1.83882	2.70246	2.94125	3.75723	19.4444
Weight-to-power ratio	4.373	11.262	12.688	12.897	14.500	27.500
Age	18.00	23.00	25.00	24.68	26.00	30.00
Experience	1.00	5.00	6.00	5.582	7.00	9.00
Avg Max Speed	23.70	55.69	67.51	67.84	79.53	124.63
Avg Avg Speed	3.427	24.413	30.100	30.901	36.529	74.667
Avg Duration (minutes)	3.636	16.480	19.595	20.652	23.356	455.330
Avg Distance (km)	1.772	8.161	11.066	12.491	15.223	78.774
Crashes	0.0000	0.0000	0.0000	0.8177	1.0000	148.000

References

1. Zhang, H., Xu, L., Cheng, X., Chen, W., & Zhao, X. (2017). Big Data Research on Driving Behavior Model and Auto Insurance Pricing Factors Based on UBI. Lecture Notes in Electrical Engineering, 404–411. https://doi.org/10.1007/978-981-10-7521-6_49
2. Ma, Y. L., Zhu, X., Hu, X., & Chiu, Y. C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. Transportation Research Part A: Policy and Practice, 113, 243–258. https://doi.org/10.1016/j.tra.2018.04.013
3. Sarabia, J. M., Prieto, F., Jordá, V., & Sperlich, S. (2020). A Note on Combining Machine Learning with Statistical Modeling for Financial Data Analysis. Risks, 8(2), 32. https://doi.org/10.3390/risks8020032
4. Guillen, M., & Pesantez-Narvaez, J. MACHINE LEARNING AND PREDICTIVE MODELING FOR AUTOMOBILE INSURANCE PRICING MACHINE LEARNING Y MODELIZACIÓN PREDICTIVA PARA LA TARIFICACIÓN EN EL SEGURO DE AUTOMÓVILES.
5. Huang, Y., & Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data. Decision Support Systems, 127, 113156. https://doi.org/10.1016/j.dss.2019.113156
6. Bordoff, J., & Noel, P. (2010). Pay-as-you-drive auto insurance. Issues of the Day: 100 Commentaries on Climate, Energy, the Environment, Transportation, and Public Health Policy, 150.
7. Bolderdijk, J., Knockaert, J., Steg, E., & Verhoef, E. (2011). Effects of Pay-As-You-Drive vehicle insurance on young drivers' speed choice: Results of a Dutch field experiment. Accident Analysis & Prevention, 43(3), 1181–1186. https://doi.org/10.1016/j.aap.2010.12.032

8. Ryan, G., Legge, M., & Rosman, D. (1998). Age related changes in drivers' crash risk and crash type. Accident Analysis & Prevention, 30(3), 379–387. https://doi.org/10.1016/s0001-4575(97)00098-5
9. Lifestyle and accidents among young drivers. (1994, 1 junio). ScienceDirect. https://linkinghub.elsevier.com/retrieve/pii/0001457594900035
10. Machin, M. A., & Sankey, K. S. (2008). Relationships between young drivers' personality characteristics, risk perceptions, and driving behaviour. Accident Analysis & Prevention, 40(2), 541–547. https://doi.org/10.1016/j.aap.2007.08.010
11. Tränkle, U., Gelau, C., & Metker, T. (1990). Risk perception and age-specific accidents of young drivers. Accident Analysis & Prevention, 22(2), 119–125. https://doi.org/10.1016/0001-4575(90)90063-q
12. Patil, P. (2018, July 7). *What is Exploratory Data Analysis? - Towards Data Science*. Medium. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
13. Reid Turner, C., Fuggetta, A., Lavazza, L., & Wolf, A. L. (1999b). A conceptual basis for feature engineering. Journal of Systems and Software, 49(1), 3–15. https://doi.org/10.1016/s0164-1212(99)00062-x
14. Rençberoğlu, E. (2019, 3 abril). Fundamental Techniques of Feature Engineering for Machine Learning. Medium. https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114
15. Kalsoom, R., & Halim, Z. (2013). Clustering the driving features based on data streams. INMIC. Published. https://doi.org/10.1109/inmic.2013.6731330
16. Higgs, B., & Abbas, M. (2013, October). A two-step segmentation algorithm for behavioral clustering of naturalistic driving styles. In 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013) (pp. 857-862). IEEE.
17. Hanafy, M., & Ming, R. (2021). Machine learning approaches for auto insurance big data. Risks, 9(2), 42.
18. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. (1987, 1 november). ScienceDirect. https://www.sciencedirect.com/science/article/pii/0377042787901257
19. Model-based evaluation of clustering validation measures. (2007, 1 march). ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0031320306003104
20. Arora, P., Deepali, & Varshney, S. (2016, 1 January). Analysis of K-Means and K-Medoids Algorithm For Big Data. ScienceDirect, 78. https://www.sciencedirect.com/science/article/pii/S1877050916000971
21. Chakure, A. (2020, 6 November). Decision Tree Classification - The Startup. Medium. https://medium.com/swlh/decision-tree-classification-de64fc4d5aac
22. Moral-García, S., Castellano, J. G., Mantas, C. J., Montella, A., & Abellán, J. (2019). Decision tree ensemble method for analyzing traffic accidents of novice drivers in urban areas. Entropy, 21(4), 360.
23. Galarnyk, M. (2021, January). Understanding Decision Trees for Classification (Python). Medium. https://towardsdatascience.com/understanding-decision-trees-for-classification-python-9663d683c952
24. Rane, S. (2019, 9 November). SHAP: A reliable way to analyze model interpretability. Medium. https://towardsdatascience.com/shap-a-reliable-way-to-analyze-your-model-interpretability-874294d30af6
25. Dataman. (2021, 2 May). Explain Your Model with the SHAP Values - Towards Data Science. Medium. https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d