Design and Development of a Discovery Service for Drivers Within the Connected Car Context Using Predictive Machine Learning Methods

Javier Goikoetxea and Alex Rayón

Abstract Currently, cars are equipped with a large number of electronic sensors that fulfil several functions. These vary from receiving and issuing a signal, to allow the automation through the permanent exchange of data and information. However, in this last field the decentralization of suppliers and manufacturers, and the lack of fully connected car solutions, have limited the creation of new solutions for both the drivers at a microeconomic level and for the general safety at a macroeconomic level. The way in which companies have developed their services to tackle these challenges have been through business rules. Historically, companies mixed up the geolocation of the vehicle with the proposals of the businesses with their own interests. This was a product-oriented approach, rather than a driver-oriented approach we propose. Additionally, we propose the usage of machine learning techniques that could scientifically show which activations are better to improve the value proposals for the drivers. Considering this context, we present a discovery platform for the drivers that could permit the recommendation of a service or a product when needed with the final focus of saving money. We also identify which variables are the most important ones in the maintenance and usage of the car. Considering a wide variety of variables, we show which ones explain better the behaviour of the drivers and show them ways to save money accordingly.

1 Introduction

In the last years, the automotive domain is changing. Among the different challenges the motor industry is facing is that vehicles are evolving from efficient engines into software machines. We are witnessing the beginning of a new era in the automotive industry where the concept of technology connected into a car arrives to completely

J. Goikoetxea · A. Rayón (✉)
University of Deusto, Bilbao, Spain
e-mail: alex.rayon@deusto.es

J. Goikoetxea
e-mail: javier.goikoetxea@opendeusto.es

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8_9

transform the vehicle. According to a study by PriceWaterhouseCoopers (2018), 40% of the mileage will be made by autonomous vehicles in the European Union by 2030. In that same year, the pool of vehicles is also expected to fall from 280 to 200 million. This represents a 28.5% reduction in the passenger car pool. This means a change of paradigm. Since there will be less demand for vehicles, they will have to last longer. It is also heading towards a world where the connected car is on the way to becoming a key car feature. An opportunity is therefore arising for the aftermarket world as fewer existing vehicles will have more activity and greater longevity is expected in those cars. Furthermore, according to the same report, vehicle manufacturers are gradually becoming software development companies to ensure that the needs of the autonomous car are met. While all this arrives, the world of the connected car is carving a niche in the evolution of this more autonomous society in terms of mobility.

This chapter aims to focus on this reality that the automotive industry is already experiencing today and that is beginning to change the management and treatment of the expenses that a person makes around the car. Twenty years ago, a vehicle had to be comfortable, safe, fast and fit with the fashion of the moment. Nowadays a car is something else. The vehicle has become a computer on wheels. This circumstance can be seen every time car manufacturers make a new commercial launch. In 1980, Spain had 10 million vehicles and just over 2,000 km of highways. At that time, the vehicles did not have any driving assistance system. The vehicles had to be efficient and beautiful. Today there are more than 30 million vehicles in Spain and more than 16,550 km of highways. The challenge lies not so much in waiting for manufacturers to develop new car models equipped with sufficient technology, but rather in embracing the opportunity to transform a conventional vehicle into a connected car through the installation of an electronic device into it. In Spain, there are over 30 million vehicles registered, out of which 25 million are simple cars. Of this figure, only slightly less than 5 million cars are less than 4 years old. So, we can assume that there is a potential market of almost 20 million cars aged over 4 years. The introduction of technology in this segment will help manage the car and its expenses more efficiently. As an average, it is calculated that in the EU a passenger car costs around €2,000 per annum (Grupo Next, 2019). Any variation below this would suggest significant savings for the drivers.

Within this context, we present NEXT Group (NEXT in advance) as a Spanish company, located at Madrid (www.gruponext.es). This company is focused on mobility data treatment and cost-efficiency driving models. NEXT has created a real-time communication platform with georeferenced and enriched data, to bring proposals and tailor-made solutions, depending on the end-user ontology. This company integrates with external systems (e.g. managers campaigns) and/or partners for data monetization by customized commercial campaigns. The most relevant use case of this platform is the Connected Car solution. The connected car solution generates and collects information from the car, using an *On Board Diagnostic* (OBD) connection. The focus of getting that data is to process and transform it into worth, based on big data analysis. All in all, the mission of the company is to generate and analyse relevant and monetizable information on the mobility ecosystem in order to provide ad hoc services to the end-user and B2B clients with both digital and

physical experiences. The connected car platform solves the vehicle needs, when needed and only what is needed. It is relevant to note that nowadays that vehicles generate data that can be used to discover when a vehicle will need different services. Unfortunately, there is not a standard designed system to manage and collect car data. Aftermarket car services are numerous and can encompass actions such as refuelling, performing maintenance, parking the car or simply fixing a breakdown that has just occurred.

The way of turning a simple car into a smart car is via connecting a device into the car using three main components (see Figs. 1 and 2):

A. The Platform is a high-performance computer system, equipped with redundancy characteristics in all its critical elements, to allow the provision of the service without temporary interruptions. The platform also has interfaces with the information systems of the partners that provide real-time services to drivers.
B. The on board OBD device, connected to the car port of each of the vehicles, reads data from the various sensors that the car has, to monitor the status of its various components, to send them to the platform for processing and exposure (ordered and managed). Generically, we can call these data the telemetry of the

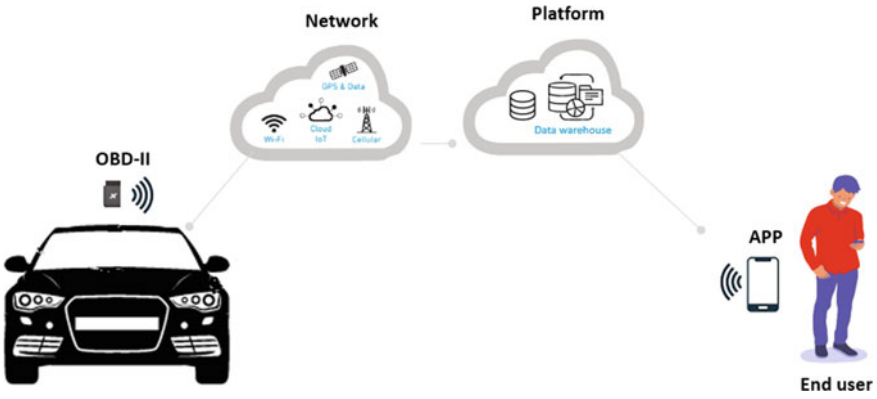

Fig. 1 High-level diagram of how the overall system

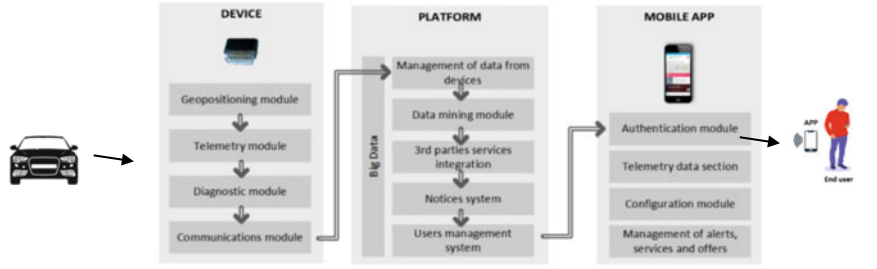

Fig. 2 Technical IT artefact solution high-level diagram

vehicle (Amouzegar & Patel, 2013). In addition, the device also incorporates a component to determine the geographical position in which the vehicle is using GPS technology. The most common technology of this type of device supports is formed by a GPS module, a Global System for Mobile communications (GSM) communications system, an accelerometer, a gyroscope, a small microprocessor, the diagnostic process and a backup battery (Duncan et al., 2015). The data will be collected from the OBD device that is part of our IT artefact, which will allow us to know what is happening in the car and its mobility in real-time and approach the user to propose a maintenance solution, when it is needed. DAS is a self-installing device which means that we will not have to use professional installers to board a device in a vehicle. This will simplify the testing. All this data is sent to a central computer to be analysed and processed normally thanks to the SIM allocated inside the artefact. The device contains a SIM card which, conveniently activated in the operator's 3G/4G network, facilitates the periodic sending of car info. The device is complemented by the creation of a communications protocol. This communications protocol will allow us to send to a server the device info.

C. A mobile application allows the driver to manage the system/service himself, from its activation to its deactivation, including its configuration, the subscription of new value-added services and some data visualization. It also offers information about the saved money in real-time, for car management purposes. This mobile application must be available for both Android terminals (versions equal to or greater than 4.4.x—KitKat) and iOS (versions equal to or greater than 9.0). The granularity of data collection as well as the frequency of sending them to the server can be handled manually by the operator of the artefact. According to Fogg (2009), the App, as end-user endpoint system, is a critical factor for the success of the system.

These three components consider technical aspects of efficient design. Like any element of automotive use, they must go through rigorous evaluation and testing methods to meet the required specifications (Resetar, 2016).

Currently, cars are equipped with a large number of electronic sensors that fulfil several functions from receiving and issuing a signal to automation through the permanent exchange of data and information. But the company platform we present is a full connected car solution. Coppola and Morisio (2016) define the connected car as '*a vehicle capable of accessing to the Internet, of communicating with smart devices as well as other cars and road infrastructures, and of collecting real-time data from multiple sources*'. This allows, for instance, that the car can communicate with the surrounding infrastructure (gas stations, workshops, spare parts suppliers, tolls…). The connected car will therefore affect the road network, the way of interacting with it, road safety and also the relationship between the vehicle and people. The vehicle, therefore, ceases to be a system that only transports us, to become a tool that can help us to satisfy our needs. It will cover them in real-time with the most accurate products and services available at that time. This technology must be able to control some parameters of the car and interact with the vehicle and its driver. Embedding

technology in the vehicle is the catalyst for developing on-demand service models. From this point, it is very easy to control and determine how much a vehicle drives, how it is driven, what inspections need or what does it pay for petrol simply by parking the vehicle next to the gas station pump hose.

2 Problem Statement

With the previous section in mind, the idea is how to design a better line of services discovering when a service/product is needed to suggest a real-time solution opportunity to the driver with the final focus of saving money. In this sense, the problem can be studied from two sides: (i) variables affecting car consumption and (ii) the role of technology in reducing costs.

2.1 Variables Affecting Car Consumption

It is to be said that there are very few people who have dared to explore this terrain. Since the automobile has the second-largest share on the consumer´s durable goods pie (the first one is the house), the disbursements around it are also very relevant (Ferber, 1967).

Fuel consumption is one of the most relevant costs in-car use. Parry (2005) indicates in his article that he has estimated a reduction in fuel consumption of 9.1% for having embedded technology into the vehicle. It works as an element of persuasion and reminder, transforming drivers into more informed people and therefore more sensitive to spending.

It can be talked about car expenses but certainly not about the decisions taken beyond those expenses. Kim et al. (2018) indicate in their article that outgoings may vary depending on the type of driving, the driving circumstances, the type of car, the speed with which it is driven and other factors related to the car use. And this is true. All these factors affect the consumption of the vehicle but our research line highlights the consumer changes behaviour more than the effective car consumption.

However, there are other studies (Shanhan, 2019) that show that 44.33% of buyers of electric or hybrid vehicles are also buyers of home solar panels and 12.67% indicate that they have not yet bought the solar panels but they are going to buy soon. If both percentages are aggregated, it can be deducted that 57% of the people who drive a sustainable vehicle (*in environmental terms*) are also eco-driven in other activities of their ordinary life. This is a relevant factor since technology can help to condition the attitude of people and people who adopt an attitude as a pattern of behaviour, evolve according to that pattern (Fogg, 2009).

2.2 The Role of Technology in Reducing Costs

There is some literature that has focused on measuring some parameters of the car, just from a car activity control point of view (Amouzegar & Patel, 2013). These authors are focusing on car maintenance control through radiofrequency systems installed on roads. The ability to check a car remotely through technology (e.g.: change of oil, change of tires or change of brake pads) is part of what we are looking for in the literature review. According to car manufacturers, managing a vehicle on time has always an impact on an indirect cost reduction, since a possible deterioration is expected if this vehicle does not attend the revision on time. In this way, Lin et al. (2009) mentioned in their research that because of the remote on-line diagnostic system connected to a car, the time of fleet management and repairs can decrease significantly.

There is a current opinion that is looking at the impact of selling mobility instead of cars. According to empirical analysis from Firnkorn and Müller (2011), private vehicles were reduced as a result of a consumer reaction. This theory confirms that a vehicle is watched as an expense generator and there is, according to the same study, a part of users who want to pay only for their mobility.

Other studies have underlined the link between the use of technology and energy (Oppong-Tawiaha et al., 2020) and between the management of energy efficiency through the use of technology and gamification (Sousa et al., 2019). Until now, no writing evidence has been found in the scope of car use and savings management.

Apart from this, different authors point out that it is necessary to deepen in the analysis of the type of maintenance service that can be offered to a driver in real-time thanks to the adoption of new technologies (He et al., 2014). This call for research is focused on the management of the automatic calls for car revision (Lin et al., 2009), the collection of vehicle data for new services (Reininger et al., 2015) and the new business models for the automotive world (Jittrapirom et al., 2017).

In a nutshell, this chapter comes to fill in the gap of the inexistence in modern mobility literature of identifying which variables are most important about car expenditures and share the results in an empiric analysis of those variables.

3 Description of Previous Solutions

The way in which companies have developed their services to tackle these challenges has been through business rules. They have developed a system whereby the back end of the platform is able to collect different variables of mobility. These variables affect vehicle data as well as mobility data. The former refers to parameters of the car's engine (reading of levels, adequacy of the engine development, variables inherent to the control of the vehicle's electronic sensors, etc.) and the latter refers to the collection of mobility data (location, speed, vehicle movements in general, etc.). Before incorporating the description and analysis of the variables, all the data mining

referring to its management and assimilation with real-time solutions is carried out based on business rules. These business rules have been defined by a set of criteria that mixed the geolocation of the vehicle with the definition of the business or the response that it was intended to launch towards the user.

Just to figure out how these business rules work, let's describe a real use case. Considering a situation when a specific vehicle passed a certain geographic point, regardless of its nature and its affinity for the product in question, the (business) rule was launched indiscriminately. In this sense, if a vehicle passed a certain kilometre point on a road, a small piece of information was sent to the vehicle, via the app, related to the advantages of refuelling at a particular gas station that was only a few kilometres from it was geo-localized. Thus, the rest of the variables necessary to determine the chance to send that precise opportunity to that vehicle were not considered. The commercial stimulus was sent only to meet the variable of the business rule: the occupant is passing by a certain geographic point.

This model for determining variables only by the description of a simple rule assumed that stimuli were being sent to customers just when a part of the business model was fulfilled, which was the drive-through a finished point. Thus, the need for creating more sophisticated algorithms to take into account other equally relevant variables to enrich the model was identified. This enriched model through an algorithm went from considering only the geographic variable to taking into account all possible variables. Elements such as the ontological needs of the drivers, the weather, the traffic density, the previous times in which they have stopped, the duration of those stops, the variable of the identification of the assignment of the suitability of the vehicle were considered. Another very relevant factor that could be analysed is the interest of that customer for the brand based on the affinity criterion of the social networks and other characteristics that went from identifying a simple model to a very sophisticated model where many other factors were taken into account.

In addition, we wanted to determine the ontological level of each user to use the factor of interests in intentions in the reformulation of the business rules. In this way, we wanted to go from using fixed variables to new variables that would help us to identify user purchasing trends. For this, we needed to incorporate a new methodology to develop the clustering of users.

On top of this, the company would like to know users' ontology. In order to do so, the company would need to understand what the behaviour of a user in sinister terms was. The company tried not to use the typical variables (speed depending on the type of road, number of kilometres travelled, smoothness or aggressiveness at the wheel …) but wanted to develop a new model that would allow us to determine new correlations to understand the behaviour of a user and their possible accident rate. Apart from this, the second main focus of this new approach to the end-user is to understand their interest, needs and intentions. All of those concepts are below the concept of ontology. For this reason, the company would like to develop a new system for treating the data. Up today, ontologies have become common on the internet world, to rank and categorize products and services for sale reasons (Noy & McGuinness, 2001). What the company is looking for is to analyse the users in order

to prioritize their needs suggesting products and services in real-time, if needed and when those products are needed.

4 Solution Proposal and Results

The main objective of this work is to generate a model to predict, given a user whose mobility data we have, their ontological profile and, within this, their insurance risk profile (cars). We define the risk profile (SINCO score) of a user not as a value from 1 to 6 (which would force us to use regression methods), but as a class (class 1, class 2, … up to class 6 to which each user belongs). The risk profile will be trained against what is called the SINCO score, which is nothing more than the historical accident analysis shared by car insurance companies in Spain. It is a database that contains the accident rate and that allows car insurance companies to master the price of each of their car policies. For this reason, we call a classifier the model which can predict the class (SINCO score) based solely on the mobility data of a user.

In order to create and configure the infrastructure needed to capture data, its pre-processing, generate the customer´s ontology and finally the creation of subsequent models, we would need the following elements (see Fig. 3):

(1) Data capture and pre-processing: From a T3 request dedicated solely to the execution and monitoring of Python scripts, it is asynchronously, and multi-thread/multi-task initiated. We pre-process NEXT RAW data together with those of the client. We then add an incident and event notification system through AWS, SNS and Cloudwatch. All of these are preprocessed with AWS lambdas.
(2) User ontology (Data processing): The resulting data (list of users together with the categories of the POIs) coming from the pre-processing of the lambdas are then submitted to the Geo-Profiling Cluster to obtain the ontology for each user.
(3) Models: using Keras to control and execute the training in supervised and non-supervised models in Tensorflow, a cluster of 10 machines with dedicated GPUs has been built and configured.

In mathematical terms, a classifier occurs when we have input data (we call them ‘X’ and in our case they are the vectors with mobility data or independent variables), some output variables (we call it ‘Y’ and in our case they are the classes from 1 to 6 of SINCO, the dependent variable) and we use an algorithm to learn the ‘mapping function’ between the input data X and the output variables Y. This function is defined as $Y = f(X)$ and we call them classifiers or supervised predictive models. It is a classifier (or supervised model) because the process of an algorithm learning from a set of historical data (the SINCO results for each user along with their mobility data in the past) resembles the human learning process: the algorithm performs predictions iteratively over the historical data and is corrected until optimal performance is achieved.

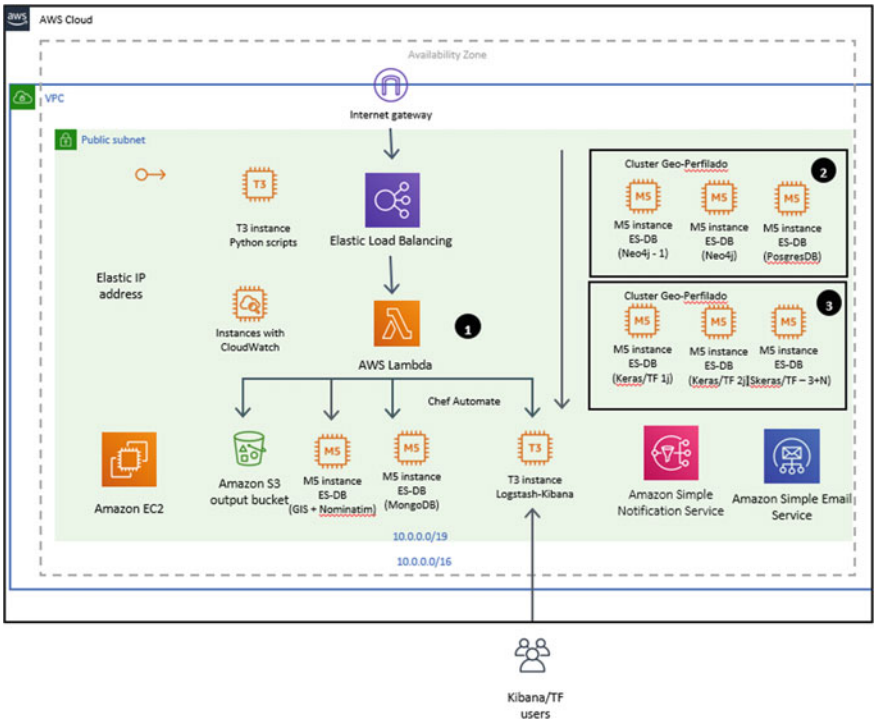

Fig. 3 The architecture of the proposed solution

The simplest algorithm to train a classifier consists of a linear function that separates each of the classes for all dimensions. The problem with simple linear classifiers is that, in real cases with many training data, many dimensions, and many classes, many possible solutions are presented with the same error. To solve this problem and at the same time achieve a function that better classifies classes that cannot be 'separated' in a linear way (as is the case with most of the real problems and in our case), we use the Support Vector Machines (SVMs) algorithm. SVM seeks the separation of the classes by maximizing the margin between the data closest to each other of each class (the 'support vectors') and applying on these a non-linear function that affects the entire class to which the' support vector belongs (see Fig. 4).

The mobility data of each user collected by the Next Group Company (www.grupponext.es) are data that we obtain from the GPS incorporated into the OBD-II device connected to the user's vehicle as explained in the previous points. These data are made up of the user's ID, longitude, latitude, vehicle speed, and the data collection time (timestamp) that the OBD-II device sends to the Grupo NEXT servers every 45 s. This data (see Fig. 5) is called raw data, since it is the primary data obtained from the capture device (the OBD-II device in the user's vehicle). In general, as in this case, it is data that individually, without further processing, does not contain

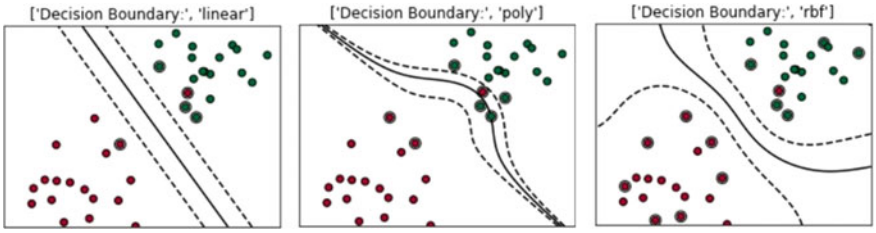

Fig. 4 SVM solution: **a** linear SVM; **b** Kernel SVM; **c** Radial SVM

I0003143	36,64100	−4,49500	0	True	18/06/2014 20:00:56	18/06/2014 18:00:45	Apagado
I0003066	41,63400	−4,73700	37	True	18/06/2014 20:01:30	18/06/2014 18:01:29	Viaje
I0003102	38,94400	1,29900	0	False	18/06/2014 20:02:13	18/06/2014 18:01:40	Encendido
I0003066	41,63200	−4,74000	0	True	18/06/2014 20:02:29	18/06/2014 18:02:29	Viaje
I0003068	41,51900	2,42000	0	False	18/06/2014 20:03:18	18/06/2014 18:02:46	Encendido
I0003102	38,94400	1,29900	62	False	18/06/2014 20:02:52	18/06/2014 18:02:51	Viaje
I0003102	38,94400	1,29900	62	False	18/06/2014 20:03:21	18/06/2014 18:02:51	Viaje
I0003073	42,49800	−1,67300	0	False	18/06/2014 20:03:41	18/06/2014 18:03:01	Encendido
I0003066	41,62800	−4,74400	29	True	18/06/2014 20:03:29	18/06/2014 18:03:29	Viaje
I0003102	38,95800	1,30800	97	True	18/06/2014 20:04:04	18/06/2014 18:03:51	Viaje
I0003068	41,52000	2,42100	16	True	18/06/2014 20:03:58	18/06/2014 18:03:57	Viaje

Fig. 5 Raw data available within the architecture of NEXT

much information. Therefore, its processing is necessary to obtain data with 'useful' information.

Raw data should not be used directly to train models (whether supervised or unsupervised classifiers). If we did so, what we would get is a classifier that would only be able to distinguish based on a given latitude and longitude or the closest point nearby to that corresponding to each class. For example, we would consider similar data the information coming from a user who works in the Jarama circuit (Madrid, Spain) or in a restaurant near it.

To obtain data that we can use to train a classifier, an inference process is necessary. In our case, starting from raw data representing geographic coordinates, we need to 'find out' which place, name it a school, company, store, shopping centre, highway, etc. (also called POI or Points of Interest) is at those coordinates. Once the POI of each coordinate is obtained and knowing the timestamp of each user (and the time that remains in each POI as well as the speed of the vehicle), we can infer the places visited by each user, how often and, based on this, draw subsequent inferences: where do you sleep, where do you work, if you take your children to school, the age of your children, etc. This process is called reverse geocoding. Due to the nature of the raw data (which is based on vehicle mobility data and not on the user), we encountered a problem when making the inference of which points of interest visited by the user: we knew where the vehicle was, but not the user.

To solve this problem and obtain the highest degree of confidence when assigning a POI to a user, we worked on various hypotheses based on the different POIs near the vehicle's parking place. For example, if a parked vehicle is 10 metres from 5 different stores, we assigned a low probability (0.01) to the potential visit of the user to each POI, since we cannot infer a visit to one store compared to another. In another

example, we introduced the hypothesis that the recurring parking of a vehicle during working hours indicates the user's workplace, or the recurring parking of the vehicle during night hours the usual residence.

To obtain the data corresponding to each POI, it is necessary to use georeferenced databases, which provide us with the POI category (egg School, business, hospital, …) and within each category, the corresponding subcategories (egg School 'primary', 'electronics' trade, 'children's' hospital, …). This classification by categories and subcategories is called a taxonomy and each individual category is a taxon. In our case, we use a mix of OSM and Google Maps taxonomies:

Even though obtaining the relationship of POIs that a user has visited and the relationship of the user with each POI (e.g. 'Works in', 'sleeps in', 'has children of school age', etc.), it is then necessary to define the inferences that we are going to set for all users and how we are going to store and exploit them digitally. Or what is the same, what type of database we will use and how we will carry out the queries or to obtain the data that it is of interest for us from each user.

The set of inferences and data that we have predefined for each user is called the user ontology. This ontology or user profile (profile in its most complete and comprehensive state—not to be confused with a static profile that does not contain user relationships with each data), we store it digitally in graph databases. More specifically, it is a NEO4J graph database. To exploit this data (queries) we use the SPARQL language, which allows us to make queries using the relationships between each user with their data. For example, we can obtain all users who have school-age children and work less than 5 km from their school.

The process of selecting the data (dimensions or variables) and relationships in a user's ontology, we call it 'dimensionality reduction'. If N is the number of dimensions or variables that we have obtained for the user, in our case it will be all the different categories of POIs, in addition to the set of all possible relationships (we call it R) between them and the user (the relationships of the ontology). The real number of possible dimensions will be the Cartesian product of $N \times R$. In our case we have an N of 10,200 (number of taxa) and an R of 1,100 (total of possible relationships defined in the ontology), then N (the overall dimensions) will be 11,220,000. If M is the number of users with data and Y results (understood results as a SINCO class of the history and Y the 'true' observations of the function $Y = f(x)$ that we saw previously), it leads us to the maximum number of Users with useful observations to train a supervised model will be, in our case, 40,500 (the number of common users).

The problem we encounter is that (mathematical proof aside), when N is much greater than M ($N >> M$), any supervised classifier that is trained by any algorithm will result in unstable models with low performance (the so-called the 'curse of dimensionality'). So, we need M to be much greater than N ($M >> N$). Even if we eliminate the Rs (that is, we assume that in the ontology all relations are of the type '*the user is interested in a certain POI category*'), we would still have an N of 10,200 and M of 40,500. The only option is to reduce N to those dimensions that contain more discriminatory information between the different classes. Out of the dimensionality reduction techniques, the most developed one and with the best results for the type of data we have is the Principal Component Analysis (or PCA

according to its acronym). This process is applied to all variables and with this we obtain an N″ with sets of 30–100 dimensions, which are the ones we proceeded to train the supervised models using SVMs.

The PCA technique consists of extracting the normal values out of all the observations and all the dimensions, in such a way that when sorting them from highest to lowest, we obtain for each dimension the 'information provided by each dimension in the discrimination of each class'. Now we only need to select which of these will be the ones we will use to train our supervised classifier. There are two techniques. One is to select the dimensions manually by an expert when the domain gives rise to it (not in our case, since it is difficult to know which dimensions of an ontology are important for SINCO beforehand). The second is automating it by generating so many supervised models using sets of them (30, 40, 50… sets of variables) and having them compete with each other to obtain the optimal number of dimensions.

This process (together with training using SVMs) is the process with the highest computational cost, since the number of models that compete with each other will ideally be formed by the Cartesian product of the size of the set of selected variables (the set T of sets 30, 31, 32, ... up to N″/2) and the number of possible dimensions (10,200). In order to avoid a computationally unmanageable number of classifiers, we have decided to reduce the number of sets to 150 experiments (and it is still 3 days of computational time for each PCA process plus dimension selection)

Once we obtain a trained classifier, we must be able to test it with users who have not been used in the training process. To do this, the first thing to focus on is the creation of a subset with the N data, we call it N″', which consists of 33% of the users with SINCO data. This is our test data set for each model generated. This causes N to be reduced (from 40,500 to 30,000) and N″ from 10,500.

The results of the classifier on the test data can then be clustered as follows:

- True Positives (TP): when the real class of the data point was 1 (True) and the predicted class is also 1 (True).
- True Negatives (TN): when the actual class of the data point was 0 (False) and the predicted class is also 0 (False).
- False Positives (FP): when the actual class of the data point was 0 (False) and the predicted one is 1 (True).
- False Negatives (FN): When the actual class of the data point was 1 (True) and the predicted value is 0 (False).

If we group this data in a matrix, we obtain for each classifier a matrix that we call a 'confusion matrix'.

Starting from the confusion matrix, we can then obtain the following metrics:

- Precision (what proportion of positive identifications have really been correct?) = TP/(TP + FP).
- Recall (what proportion of real positives has been correctly identified?) = TP/(TP + FN).
- Accuracy (which of the predictions did our classifier correctly identify?) = (TP + FN)/(TP + TN + FP + FN).

Establishment of the processes for the evaluation of supervised and predictive models (aimed to replicating the SINCO risk score and developing the complete user ontology):

(1) Extraction of 'blind' customers for independent evaluation—1.000 blind customers have been drawn (with whom no training nor internal evaluation in any model have been carried out) out of the 40.500 common customers (from the client and as from NEXT). Additionally, two other groups were created: one comprising 30,600 customers to train the models and another one of 6,000 customers to internally evaluate the models

(2) As a specific request from the client, some other variables on the type of customer driving together with the variable of temporality have been included (…). These variables have been incorporated into the models to be trained again, forcing their crunching regardless of their weight in the discrimination of the data, in order to evaluate them against the rest of the variables in depth. Time dimensions have also been incorporated into current variables

- Total driving time—the sum of a customer's total driving minutes on any type of road. It is defined as the time with the vehicle at a speed >0 and normalized by the total time as a customer.
- Total high speed time—It is defined as the total time with the vehicle at a speed >100 km/h of a customer on any type of road. It is then normalized for total time at speed >0.
- (×3) Total time on Highway, Secondary and Regional/Other roads—total time with the vehicle at a speed >0 km/h for each type of road. Normalized by the total driving time.
- Total time of violations—total time with the vehicle at a speed 20% higher than that allowed on any type of road.
- Total number of infractions—number of trips, out of the total made, with at least one infraction committed for speeding.

(3) Analysis of context, events and meta-data—No customer's context has been incorporated at any time nor the meta-data corresponding to unique events or situations. For example, special events such as concerts or demonstrations, which give us information about the customer's profile, have not been linked to geolocations.

Of all the variables used, we are left with those with the highest incidence on the data, ranked from the highest to the lowest number of events. We highlight with a red dot those that have a disproportionate weight over the rest and on which we will include a third level of taxonomy (e.g. (Amenity, Pub) (Amenity, Pub, 00:00 to 06:00 LV, NOT Work_Place)) (see Fig. 6).

	office	estate_agent
	amenity	hospital
	leisure	golf_course
	shop	daityourself
	railway	platform
	toursim	artwork
	historic	monument
	shop	sports
	amenity	theatre
	military	checkpoint
x	amenity	pub
	tourism	attraction
	amenity	public_building
	amenity	parking
	craft	window_construction
	office	company
	amenity	social_facility
	building	appartments
x	leisure	marina
	leisure	playground
	office	yes
	amenity	bank
	amenity	fuel
	shop	mall
	shop	tyres
x	amenity	brothel
	leisure	stadium
	natural	beach
	shop	furniture
	tourism	hotel
	shop	bakery
	highway	service
	amenity	cafe
	amenity	bar
	highway	trunk
	highway	primary
	shop	hairdresser
	building	yes
	amenity	bus_station
	amenity	pharmacy
	highway	secondary
	place	house
	tourism	museum
	highway	pedestrian
	amenity	restaurant
	leisure	sports_centre
x	amenity	place_of_worship
x	highway	tertiary
	shop	supermarket
	highway	unclassified
	amenity	school
	highway	residential
	shop	tobacco
	shop	hardware
x	tourism	caravan_site
x	shop	garden_centre
	amenity	post_office
	amenity	police
	amenity	marketplace
	highway	proposed
	amenity	library
	historic	memorial
	highway	construction
	amenity	bicycle_parking
	shop	clothes
	natural	spring
	amenity	language_school
	building	commercial
x	amenity	embassy
	building	garage
	place	locality
	railway	station
	highway	footway
	amenity	university
	historic	ruins
	highway	track
	building	terrace
	amenity	townhall
	leisure	garden
	building	industrial
	club	culture
	office	government
	aeroway	aerodrome
	highway	services
	highway	cycleway
	leisure	pitch
	shop	kiosk
	shop	optician
	place	square
	amenity	clinic
	amenity	college
	man_made	surveillance
	building	office
	highway	living_street
	shop	department_store
	amenity	fast_food
	amenity	motorcycle_parking
	amenity	community_centre
	highway	motorway
	shop	car
	leisure	swimming_pool
	building	residential
	shop	lottery
	office	ngo
	shop	greengrocer
	office	telecommunication
	building	train_station
	shop	gift
	building	school
	shop	jewelry
	building	roof
	shop	car_repair
	amenity	training
	amenity	vehicle_inspection
	amenity	vending_machine
	amenity	waste_transfer_station
	building	church
	shop	convenience
	highway	primary_link
	building	dormitory
	shop	motorcycle
	tourism	information
x	military	barracks
x	military	bunker
	natural	heath
	natural	water
x	amenity	kindergarten
	amenity	charging_station
	shop	variety_store
	shop	wine
	office	foundation
	amenity	exhibition_centre
	shop	pet
	shop	music
	shop	trade
	office	educational_institution
	railway	subway_entrance
x	shop	bicycle
	building	house
	building	garages
	office	it
	historic	castle
	building	hospital
	office	insurance
	shop	car_parts
	leisure	common
	tourism	guest_house
	amenity	bicycle_rental
	tourism	camp_site
	shop	yes
	shop	electronics

Fig. 6 Variables that have a disproportionate weight over the rest

5 Results Evaluation and Conclusions

We have presented the variables that, out of over 1.500 variables analysed (without including the variables with time dimension), best represent the behaviour of the driver. Evaluation and influence of these variables in the models has also been presented, being the best the one that presents an accuracy of 92.6% and a recall of 74.1%. The 150 variables with the highest incidence on data have been ranked

from upper to lower level of events. We have highlighted with a red dot those that have a disproportionate weight over the rest and on which we will include a 3rd level of taxonomy.

A total of 900 computing hours have been conducted to date (in 53 uninterrupted days) for data processing and generation of supervised models. We have incorporated the time dimension variable into modelling, but with some limitations. For example, when a customer is driving on a county road (day or night) or the hours and days of the week when a customer visits a Pub or a Restaurant (weekend vs. midweek). Perhaps more granular time variables could be incorporated.

Apart from that, from a business point of view, we could also include variables to analyse users from an economic perspective: higher margin contribution, greater capacity for cross-selling and up-selling and last but not least, less churn. That could provide NEXT a further step in improving its' business value proposition.

References

Amouzegar, F., & Patel, A. (2013). Vehicle maintenance notification system using RFID technology. *International Journal of Computer Theory and Engineering,* 312–316.

Coppola, R., & Morisio, M. (2016). Connected car: Technologies, issues, future trends. *ACM Computer Surveys, 49*(3), 1–36.

Duncan, M., Charness, N., Chapin, T., Horner, M., Stevens, L., & Richard, Y. A. (2015). *Enhanced mobility for aging populations using automated vehicles.* BDV30 977-11

Ferber, R. (1967). Determinants of investment behavior. In *NBER books, national bureau of economic research*, Inc, number ferb67–1.

Firnkorn, J., & Müller, M. (2011). What will be the environmental effects of new free-floating car-sharing systems? The case of car2go in Ulm. *Ecological Economics, 70*(8), 1519–1528.

Fogg, B. J. (2009). A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology (Persuasive '09).* Association for *Computing Machinery.* New York, NY, USA, Article 40, 1–7.

He, W., Yan, G., & Xu, L. D. (2014). Developing vehicular data cloud services in the IoT environment. *IEEE Transactions on Industrial Informatics, 10*(2), 1587–1595.

Jittrapirom, P., Caiati, V., Feneri, A. M., Ebrahimigharehbaghi, S., Alonso González, M. J., & Narayan, J. (2017). *Mobility as a service: A critical review of definitions, assessments of schemes, and key challenges. Smart Cities – Infrastructure and Information, 2*(2)

Kim, J., Hwangbo, H., & Kim, S. (2018). An empirical study on real-time data analytics for connected cars: Sensor-based applications for smart cars. *International Journal of Distributed Sensor Networks, 14*(1), 1550147718755290.

Lin, J., Chen, S., Shih, Y., & Chen, S. (2009). *For vehicles by integrating the technology of OBD, GPS, and 3G* (Vol. 56). Word Academic of Science.

Noy, N. & Mcguinness, D. (2001). Ontology development 101: A guide to creating your first ontology. *Knowledge Systems Laboratory, 32.*

Oppong-Tawiah, D., Bassellier, G., & Pinsonneault, A. (2020). Tracing the next-generation platform firm: A typology of digital platforms as new organizing forms. *ECIS 2020 Research-in-Progress Papers,* 9.

Parry, I. W. H. (2005). Is pay-as-you-drive insurance a better way to reduce gasoline than gasoline taxes? *American Economic Review, 95*(2), 288–293.

PricewaterhouseCoopers. (2018). *En 2030, Europa tendrá 80 millones de coches menos como consecuencia del transporte compartido y de la digitalización.* Available at: https://www.pwc.es/es/sala-prensa/notas-prensa/2018/2030-europa-ochenta-millones-menos-coches.html

Reininger, M., Miller, S., Zhuang, Y., & Cappos, J. (2015). A first look at vehicle data collection via smartphone sensors. In *2015 IEEE Sensors Applications Symposium (SAS)* (pp. 1–6).

Resetar, M. (2016). *Innovative approach to vehicle diagnostics.* Available at: https://docplayer.net/33754531-Innovative-approach-to-vehicle-diagnostics.html.

Shahan, Z. (2019). EV Ownership + Rooftop Solar Ownership—New Report & Charts. *CleanTechnica.* Available at: https://cleantechnica.com/2019/12/25/ev-ownership-rooftop-solar-ownership-new-report-charts/.

Sousa, S., et al. (2019). Gamification as a way to involve young adults in energy efficiency and sufficiency—A case study. *ECEEE Summer Study Proceedings.*