Machine-Learning-Based No Show Prediction in Outpatient Visits

C. Elvira[1], A. Ochoa[2], J. C. Gonzálvez[2], F. Mochón[3]*

[1] Hospital Clínico San Carlos, Madrid (Spain)
[2] Zed Worldwide, Madrid (Spain)
[3] Universidad Nacional de Educación a Distancia (UNED) (Spain)

Received 15 July 2016 | Accepted 3 February 2017 | Published 27 March 2017

LA UNIVERSIDAD
EN INTERNET

Abstract

A recurring problem in healthcare is the high percentage of patients who miss their appointment, be it a consultation or a hospital test. The present study seeks patient's behavioural patterns that allow predicting the probability of no-shows. We explore the convenience of using Big Data Machine Learning models to accomplish this task. To begin with, a predictive model based only on variables associated with the target appointment is built. Then the model is improved by considering the patient's history of appointments. In both cases, the Gradient Boosting algorithm was the predictor of choice. Our numerical results are considered promising given the small amount of information available. However, there seems to be plenty of room to improve the model if we manage to collect additional data for both patients and appointments.

Keywords

Big Data, Healthcare, Machine Learning, Consultations, Gradient Boosting, Predicting No Shows.

DOI: 10.9781/ijimai.2017.03.004

I. Introduction

Healthcare demand has slightly different behaviours in the public and private sectors, in both quantitative and qualitative terms, regardless of the health system [1] [2] in the country. This can be explained mainly by variations in funding and differences in the portfolio of services offered [1]. Recognizing the intrinsic characteristics of healthcare demand becomes essential for stakeholders who hold any responsibility over it, be they service providers or policy planners.

Understanding this demand may provide economic and social benefits, for example through savings and reduction of waiting lists. The common denominator is always the added value that this knowledge brings. In any case, it is necessary to analyse demand in a scientific manner so that healthcare providers can react accordingly. Big Data techniques play a crucial role here, as it would be very difficult to do so without them. It is worthwhile noting that this analysis has been traditionally carried out using historical data. This is not the same in other economic fields, where techniques of prediction or behaviour anticipation of demand already have a long history and scientific foundation supporting them. For example, it is a given that electric energy is generated based on a minute by minute forecast of demand – otherwise supply cuts would be frequent. However, this approach is unusual within the healthcare domain, especially when circumscribed to the public sphere.

The second major pillar to consider in the relationship between demand and supply is the actual effectiveness of the provided healthcare; the more available supply, the better the response to its demand. Therefore, maximizing efficiency becomes paramount. Here too, anticipated knowledge of demand behaviour plays an important role. For example, in the case of outpatient consultations, where patients frequently miss their appointment, this lack of attendance has two direct effects. The first one obviously involves patients themselves, who postpone the chance to be treated for a medical condition. The second one affects healthcare procurement, as the time lost by one patient's non-attendance implies that another patient misses the opportunity to be seen by the doctor. This is the so-called opportunity cost. In private sector settings, you have to add another opportunity cost, for lost revenue during this idle time.

As an example of Big Data applications [3] of clinical data we find cases such as how to treat patients differently based on their characteristics ("treatment personalization") or in help systems of radio-diagnostic equipment that provide suggestions based on the differences of simple tones of grey (which are just points 1 or 0 in digital language) after the statistical analysis of millions of previous expositions. These are not impending developments, they are already here and making the most of them is an obligation that should not be delayed because at the end it is about the most valuable asset of human beings, health.

Going one step further in the analysis implies paying particular attention to outpatient healthcare, which makes the greatest impact in terms of number of patients being cared for in a public hospital, with magnitudes exceeding 30 ambulatory cases per admission in many cases. Therefore, we are dealing with an activity that affects a large number of people (patients), additionally absorbing significant hospital resources.

In health systems [2] with universal public coverage, the chronic mismatch between the demand for assistance and the supply of resources leads to waiting lists [4] with response times that are frequently unacceptably long, considering what would be the optimal time for citizens. On the other hand, general historical data in hospitals shows that there is a significant percentage of patients who do not attend their previously-committed outpatient appointment and that in

* Corresponding author.

E-mail address: fmochon@cee.uned.es

some cases this may amount to 10% or more of non-attendance. In terms of production or of responding to healthcare demand, wasting this percentage of available resources is an unacceptable luxury as long as there is a list [4] of other patients waiting to receive their assistance. Additionally, it implies an intrinsic waste of idle resources in the system [2].

Upon these considerations, if the percentage of patient non-attendance to their outpatient appointments could be reduced [5], it'd be possible to reduce waiting lists [4] and citizens could be better served, while use of health resources would be improved (via increased efficiency) [6].

In order to achieve this goal, it seems a good starting point could be to learn about the behaviour of patients who do not attend their appointments [7] and try to find out whether there is any pattern in their behaviour [8] [9] which then allows to carry out specific actions for each detected population strata. Until not so long ago, there were no technological tools available for the analysis of data related to predictive stratified studies on non-attendance, since databases are large (they can exceed one million annual appointments for a large hospital). The emergence of Big Data techniques [3] in recent years has made it possible to carry out these studies - a clear example of their usefulness in real life.

This article is structured in the following sections. Firstly, a description of the available information is presented. The next section discusses the operation of a predictive model which includes, as explanatory variables, the information related to medical appointments of different patients. Aiming to improve the results, the following section provides the model with the available data on the previous appointments that a patient has had. This information is used to construct a second predictive model. Next, the training of the model is carried out to try to improve prediction accuracy. The last section presents the work conclusions and discusses possible lines of research to try to improve the results.

II. Description of Available Information

This research lays out a study carried out in a university hospital [10] [11] in Madrid, the San Carlos Clinical Hospital. The hospital provides practically all clinical specialties and an outpatient activity. Consequently, it processes about eight hundred thousand outpatient consultations a year and, additionally, must perform a similar number of outpatient diagnoses (radiological, analytical, day hospital sessions, ambulatory surgical procedures, etc.).

We'd like to thank this hospital for its spirit of improvement and research, for facilitating data for the study while maintaining the absolute anonymity of all records used and the strict compliance with the legislation on personal data protection. A retrospective study of at least one year is therefore proposed with all available records from the field of consultations and examinations to identify whether there is any pattern that defines the behaviour of patients who do not attend their scheduled appointment. This way, strategies for action and improvement on specific groups could be defined, taking into account the already mentioned positive repercussions on efficiency, performance and benefits for the patient.

There are two data sets with information on medical appointments of different patients from January 2015 to September 2016. One of the data sets refers to the ancillary appointments that precede diagnosis and the other one to consultations.

Consultations are acts in which there is the intervention of the patient and medical staff, basically a doctor, with a diagnosis purpose or clinical follow-up. Ancillary processes are acts that are usually related to technological equipment for diagnosis, although a doctor interpretation may be necessary later on.

We can define appointment as the information regarding an attendance commitment for a date, time and place / assistance device. A consultation as the act of assistance with purpose of diagnosis or follow-up of a clinical process carried out by health personnel.

Both data sets contain the following information:

- Patient identifier (unique alphanumeric sequence that guarantees patient anonymity).
- Demographic: gender and age.
- Date and time when the appointment is requested and when it actually takes place.
- Region (province) and place of the appointment.
- Medical speciality and type of appointment (monographic or not).
- Type of appointment (first appointment, review, prevention).
- Whether the patient attends or not.

Analysing the data set we proceed to delete the province, since it remains constant, and to add the CONSULTATION variable to indicate when an appointment belongs to the first or second data set.

Both data sets are then pre-processed, eliminating incomplete records, solving inconsistencies and correcting errors (for example, date formats). After this set of operations and the merging of both data sets, a combined data set with 2,362,850 records (one record per appointment) is obtained. Each record contains the following 12 variables:

- 'PATIENT ID',
- 'GENDER'
- 'AGE',
- 'APPOINTMENT DATE',
- 'APPOINTMENT TIME ',
- 'DATE_REQUEST',
- 'MEDICAL_SPECIALITY',
- 'MONOGRAPHIC',
- 'APPOINTMENT_TYPE',
- 'CONSULTATION',
- 'CENTER' (different building where the patient will be attended)
- 'ACCOMPLISHED'.

III. Construction of a Predictive Model

Starting from this final data set, a predictive model [12] is constructed in order to predict the value of the ACCOMPLISHED variable, which reflects, based on the remaining variables, whether the patient attends the appointment or not.

Regarding the prediction we want to make, we may find three possible scenarios:

1. Patients who had previously requested an appointment and attend the doctor´s consultation.
2. Patients who had previously requested an appointment and do NOT attend the consultation.
3. Patients who had previously NOT requested the appointment and attend the doctor´s consultation.

Table 1 shows the figures and percentages of each of the cases previously mentioned.

TABLE I. Patients Classification According to Whether or not They Attend the Appointment and if They Do it Without Appointment

#	Class	# Consultations	%
1	Show	1.997.090	85%
2	No Show	237.029	10%
3	Show without appointment	128.731	5%
	Total	2.362.850	100%

Case 3 is excluded from our analysis since it would not make sense to try to predict the attendance of patients who have not requested an appointment, as we would not have information about them. Eliminating Case 3 records from the data set leaves 2,234,119 records, 90% of which correspond to patients who previously requested an appointment and attended it. The remaining 10% corresponds to patients who previously requested an appointment but failed to attend. That is, we are facing a classification problem in which classes are very unbalanced.

According to the “Show” / “no show” distribution of our data set, an algorithm stating that a patient always goes to the appointment would be making a mistake of only 10% which may seem acceptable. However, the overall accuracy is not reliable measure to assess the quality of the results with unbalanced datasets.

The first approximation that has been made on the data set was to consider only the available information about the target appointment - the one to be predicted. It is carried out by building a model based on Gradient Boosting Machine (GBM) [13], a classification algorithm which has shown very good results in different tasks, both in the use of discrete or continuous variables, as in the treatment of unbalanced data sets [14].

In our numerical work we used the H2O.ai implementation of the GBM model [15]. In this data set we obtain an average per class accuracy of approximately 60%.

Accuracy is calculated as shown in Figure 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

Fig. 1. Accuracy calculation.

From this initial exploration of the dataset, follows the relative variable ranking shown in Figure 2.

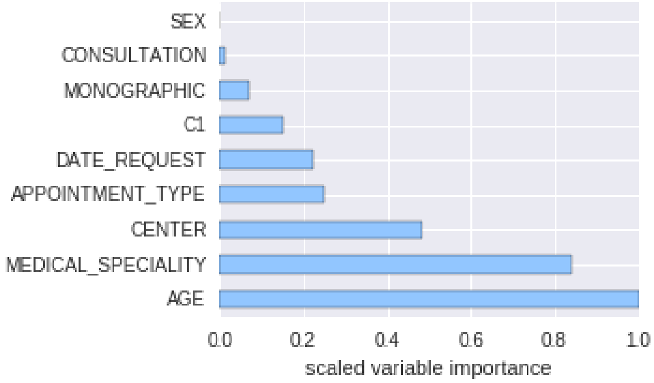


Fig. 2. Variables relative importance.

According to the results obtained it is observed that when it comes to making a classification decision, the most relevant variable (the variable with the most predictive power) is AGE, followed by the MEDICAL_SPECIALITY (0.83) and the health CENTER where the appointment takes place (0.48). Likewise, the variable SEX is not found to be a relevant variable, in other words, there are no differences between men and women regarding their attendance to previously arranged appointments. Note C1 is the date of the appointment.

IV. Inclusion of Previous Appointments in the Predictive Model

To further improve the results, we are providing the model with the data on the patient’s previous appointments. Put another way, we are checking whether the information regarding previous appointments kept or not by the patient can improve the functioning of the algorithm.

To this end, we will have to create new variables from the data that is provided, on the history of patients regarding their previous appointments.

In the first place, we will create the data set which associates each patient with the ordered history of their appointments. The new variables created are:

- FIRST_DATE: Date of first appointment.
- LAST_DATE: Date of last appointment.
- LENGTH: Number of appointments made.
- SERIES: Chain containing the following bundled information about each patient’s appointments:
    - SPECIALITY
    - MONOGRAPHIC
    - TYPE_OF_APPOINTMENT
    - CONSULTATION
    - MEDICAL_CENTER
    - DELAYS - number of days since the previous appointment
    - H_D - appointment time interval (the day is divided into 4-hour intervals).
    - D_W - day of the week
    - M – month
    - DAYS_Request - number of days since the appointment was requested
    - DAYS_First appointment - number of days since the first appointment
    - ACCOMPLISHED

It should be highlighted that for each appointment we will have a tuple like the previous one, storing under the SERIES variable, in a bundled form, all tuples, which constitute all the appointments made by a patient. This allows us to increase the number of variables that describe an appointment.

We can also add calculated variables that will allow us to add information about the patient, such as: the number of past appointments, the number of appointments attended, the number of days elapsed between appointments, the sum of delays between appointments, both in the history record and in the period of the k-last appointments considered by the model.

With these new defined variables we are able to use the information of each patient, considering their previous appointments, in order to create a new data set for our model.

When considering more than one appointment in the model, we will have to establish a mechanism to identify each appointment. For that purpose, we will use a number that we will add as a suffix to the name of the variable (NAME OF VARIABLE_i, being the suffix i-appointment).

For example, if we use information from two appointments in our data set, we will find the variable ACCOMPLISHED_0 that is the one we want to predict and the variable ACCOMPLISHED_1 that will take the values S (Yes, with Scheduled Appointment), N (no) or U (Yes, without appointment) depending on whether the patient attended his or her last appointment or not. Thus, if we decided to take into account only the last two appointments for the analysis, the data set would contain the following information:

- Information about the patient: PATIENT_ID| AGE | SEX
- Information about appointments made by a patient:

  FIRST_APPOINTMENT | LAST_APPOINTMENT | n_APPOINTMENTS | n_DAYS | Delay_sum

- Information about the immediately preceding APPOINTMENT to the one to be predicted (n-1, marked by the suffix "_1"):

  SPECIALITY_1 | MONOGRAPHIC_1 | TYPE_OF_APPOINTMENT_1 | CONSULTATION_1 | MEDICAL CENTER_1 | Delays_1 | H_d_1 | D_w_1 | M_1 | DAYS_Request_1 | DAYS_First_Appointment_1 | ACCOMPLISHED_1

- Information about the appointment you want to predict:

  SPECIALITY_0 | MONOGRAPHIC_0 | TYPE_OF_APPOINTMENT_0 | CONSULTATION_0 | CENTER_0 | Delays_0 | H_d_0 | D_w_0 | M_0 | DAYS_Request_0 | DAYS_First_Appointment_0 | ACCOMPLISHED_0

Note that the variable ACCOMPLISHED_0 is restricted to the values {show, no show} but only by appointment. However, for previous appointments, the variable ACCOMPLISHED_i is considered as attendance regardless of whether or not the patient had arranged a previous appointment, since it is now relevant whether or not the patient attended.

Initially a data set is constructed that takes into account the information of two appointments, the one to be predicted and the immediately previous one. This new data set contains 1,715,029 records. The number of records is now smaller; there will be as many records as n-1 appointments per patient, since each record corresponding to an appointment will have in the variable SERIES the information about the immediately previous appointment. This way the oldest appointment of a patient (in terms of time) will no longer appear in the data set as a record, as it will already be incorporated under the variable SERIES of the penultimate appointment in the same time.

In order to determine whether this enrichment of the data has any effect on the average accuracy (as was done with the initial data set), we proceed to apply this new data set to different predictive models. This time the result obtained for the accuracy is an average of 70% between the two classes (10 percentage points better than for the original data set).

Although the different models used have had similar results, a model of Gradient Boosting Machine (GBM) was chosen since it was the one which generated better results.

We now proceed to the exploration of this new data set using the GBM application. The importance of the variables in this new data set is shown in figure 3. The variable with a higher predictive power is now the one that indicates if the patient attended or not the previous appointment (ACCOMPLISHED_1). The variable that follows it in predictive importance is the SPECIALITY 0, above the AGE, which in the previous model turned out to be the variable with greater relative predictive importance.

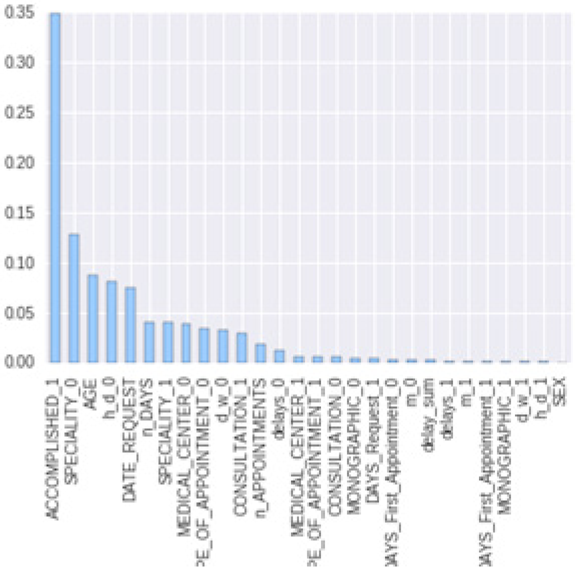


Fig. 3. Variables relative importance for the second dataset.

In order to determine if the number of a patient's previous appointments considered in the model have any relevance in the results, the same experiment is carried out with models of the i-last appointments for i = {3, 4 and 5). Despite testing other predictive algorithms, e.g. General Linear Model GLM [A5] and Deep Learning [17], the obtained results were similar and do not significantly improve those obtained with i = 2. This seems to suggest that in order to improve the results, additional information would be needed.

V. Model Training: Trying to Improve Accuracy

Since taking into account a higher number of appointments has not improved the results, in order to proceed with the investigation, the data set will be used with i=2. That is, two appointments, the current one and the immediately previous one.

Once the classification model has been defined, we will proceed to construct a predictive model, for which the training of the model will be necessary. The latter data set is divided into two parts:

1. One with the appointments available for the period 2015-01-01 -- 2016-05-31 to be used for training, validation and testing. For this training process the following procedure has been followed:
   a) Training of the model, 80% of total data set records.
   ii) 60% of records have been used for the training
   iii) 20% of records have been used for validation of the model.
   d) Test data: remaining 20% of the records.
2. Another one with the appointments from 2016-05-31 until 2016-09-27 that will be used as a test data set. The results presented in this article are obtained applying the trained model to this data set.

Once the model has been trained, it is applied to the test data, obtaining the probability of each appointment belonging to either the "show" or the "no show" class, which allows us to construct the ROC1 curve in Figure 4, with a value of 0.7404 for the Area under the Curve (AUC) [18].

Taking different values of the ROC curve, we can construct different confusion matrices. Ideally, we should have at hand a relative cost function which allowed us to select the value of the ROC curve that would then allow us to obtain the most appropriate confusion matrix

[1] Receiver Operating Characteristic curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied

for the problem that we want to solve. That is, we should be able to quantify the value that has a false positive for the business (thinking that the patient will attend the appointment when actually there is a non-attendance) or a false negative (thinking the patient will not attend when in fact there is an attendance), with the purpose of minimizing costs this way.

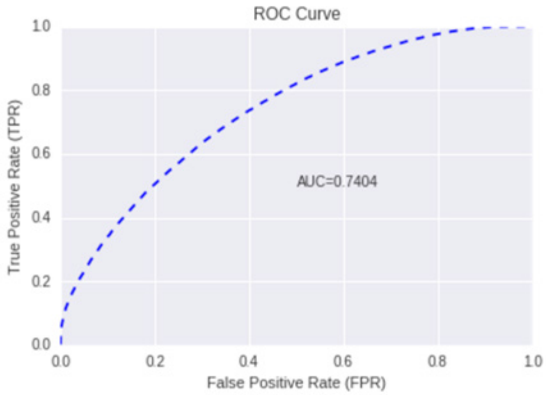


Fig. 4. ROC curve.

Since we do not have such a cost function, no business rules will be established for the study that allow us to perform that quantification, we simply intend to predict the attendance or non-attendance to the appointment. The value that makes the maximal average of the accuracy by class has been taken as a threshold value of the ROC curve. That threshold value is at 0.899529962584.

This value allows us to build the Table 2 confusion matrix[2].

TABLE II. Confusion Matrix for the Threshold Value= 0.899529962584

		PREDICTION				
		No show (0)	Show (1)	Total	Error	Rate
REAL VALUES	No Show (0)	19.955	12.315	32.270	38,16%	(12.315/32.270)
	Show (1)	81.613	209.096	290.709	28,07%	(81.613/291.394)
	Total	101.568	221.411	322.979	29,08%	(93.928/322.979)

The values of the rows correspond to the real values, while the values of the columns correspond to the prediction values of the model. Therefore, out of the 322,979 appointments included in the data set of test, patients did not attend 32,270 appointments, whereas patients did attend 291,394 appointments.

According to these data, the classification model is making:

- An error of 38.16% in predicting non-attendance. That is, of the 32,270 patients who truly did not attend their appointments, the model was right with 19,955 and failed with 12,315.
- An error of 28.01% in predicting attendances. That is, of the 291,394 patients who actually attended their appointments, the model was right with 209,781 and failed with 81,613.

As mentioned above, in order to evaluate the quality of the results obtained, we would need to establish a function that measures the relative cost of the decisions that are made based on the results obtained from the model. In this sense, it would be possible to improve the error of one class (worsening the error of the other), by modifying the value of the decision threshold. As the improvement in the error of one class implies a worsening of the other, it is necessary to find the value that optimizes the results.

In the case of medical appointments such as those used for making this article, the most obvious examples for the application of a prediction model as the one previously described would correspond to the following business situations:

1. A model that allows minimizing doctor idle time caused by patient non-attendance. In this case, we would be interested in minimizing the prediction error of attendances. Therefore we should establish what the cost for the business is when a doctor is not attending other patients because the patient of the current appointment has not attended.
2. A model that allows minimizing patient waiting times avoiding overbooking. In this case, we will be interested in minimizing the error in the prediction of non-attendances. We should then establish the cost for the business of having patients wait and therefore waste their time (or of doctors having to lengthen their day), because a doctor has more patients than he/she can actually attend.

Applying this model, with the data set available, to any of the two business cases described above is not very realistic as it does not show information such as the number of doctors who are attending at the same consultation of one speciality. Usually in hospitals and primary healthcare centres, consultations of the same speciality are cared for by more than one doctor, which makes the care flow and therefore, doctor idle time or patient waiting time, directly dependent on that variable. However that information is not available in the data set.

VI. A practical Application of the Model

In relation to medical appointments it is common practice to make use of notification systems based on the sending of SMS to the patient on the dates close to the appointment. These SMS remind the patient of the details of the appointment, in order to minimize forgetfulness and non-attendance, or to otherwise seek the patient's notification of non-attendance, which would allow rescheduling the appointment and assigning that time slot to another patient.

However, sending SMS is not free; it means a cost for the institution that provides the medical service. Using a prediction system such as the one described, despite the results not being spectacular, could reduce this cost without a worsening of patient attendance ratios.

Normally these SMS notification systems send a message to all patients who have a scheduled appointment. In the case at hand, since our file contains 323,664 patient appointments, the system would send the same number of messages.

Using this system, it would be possible to limit the sending of SMS to those patients that the model predicts will not attend the appointment, in the case that concerns us 101,568 SMS. This would mean a 66% reduction in the sending of messages. According to the data of the confusion matrix presented above, the model would recommend sending SMS to patients who are actually going to attend and would be leaving out of the sending 12,315 patients who have been classified as attending but who did not effectively attend, making therefore an error of approximately 4% on the total data set.

VII. Final Thoughts

In view of the results, it can be stated that the information collected in the data set does not seem sufficient, neither in terms of patient description, nor in terms of appointment characteristics, so as to

[2] A matrix of confusion is a tool that allows visualizing the performance of an algorithm that is used in supervised learning. Each column in the matrix represents the number of predictions for each class, while each row represents the instances in the real class.

construct a solid predictive model. The improvement of the results, that is to say, the improvement of the capacities of the classifier presented in this work, seems to depend on an improvement of the amount of information available, both for patients and appointments.

Patient information could be supplemented with more socio-demographic information. Likewise, with regard to appointments, it seems logical to think that supplementing information with data related to the procedures and processes to be performed on the patient can provide the classifier with relevant information to better predict categorization.

Finally, it also seems reasonable to think that the severity of a disease and its consequences can be a significant variable in a patient's decision to attend an appointment or not. While it is true that these are very subjective concepts and each individual interprets them in a different way, health is something that the average individual usually takes very seriously. Therefore, providing this information from the patient's medical history could improve the model.

References

[1] Lameire, Norbert, Preben Joffe, and Michael Wiedemann. "Healthcare sys-tems—an international review: an overview." Nephrology Dialysis Trans-plantation 14.suppl 6 (1999): 3-9.

[2] World Health Organization. The world health report 2000: health systems: improving performance. World Health Organization, 2000.

[3] Aldana J, Baldominos A, García JM, Gonzálvez JC, Mochón F, Navas I; (2.016). "Introducción al Big Data", Editorial García Maroto Editores SL.

[4] Romero E. Granja, et al. "Study of the derivations to an external consultation of Internal Medicine: can be managed the waiting list?." Anales de Medicina Interna-Madrid-Órgano Oficial de la Sociedad Española de Medicina Interna-. Vol. 21. No. 2. ARAN, 2004.

[5] Guerrero MA, Gorgemans S. "Absentismo de pacientes citados en las consultas de Atención Especializada del Consorcio Aragonés Sanitario de Alta Resolución: repercusión económica y demoras." XVI Encuentro de Economía Pública: 5 y 6 de febrero de 2009: Palacio de Congresos de Granada. 2009.

[6] Jabalera ML, Morales JM, Rivas F. "Factores determinantes y coste económico del absentismo de pacientes en consultas externas de la Agencia Sanitaria Costa del Sol." Anales del Sistema Sanitario de Navarra. Vol. 38. No. 2. Gobierno de Navarra. Departamento de Salud, 2015.

[7] Fonseca E, Vázquez P, Mata P, Pita S, Muiño ML. "Estudio de la inasistencia a las citaciones en consulta en un servicio de dermatología" Piel 2001; 16: 485-489.

[8] Salinas, EA, De la Cruz R, Bastías G. "Inasistencia de pacientes a consultas médicas de especialistas y su relación con indicadores ambientales y socioeconómicos regionales en el sistema de salud público de Chile." Medwave 14.09 (2014).

[9] Perez M, Rendon MM. Características asociadas con la inasistencia a la consulta de promoción y prevención en salud en una IPS de la Ciudad de Medellín 2016. Diss. 2016.

[10] Giunta D, et al. "Factors associated with nonattendance at clinical medi-cine scheduled outpatient appointments in a university general hospital." Patient Prefer Adherence 7 (2013): 1163-70.

[11] Pereira-Victorio CJ, et al. "Absentismo de pacientes a la consulta externa es-pecializada en un hospital de tercer nivel en España." Medicina General y de Familia 5.3 (2016): 83-90.

[12] Max K, Kjell J. "Applied Predictive Modeling". Springer 2013 Edition. ISBN 978-1-4614-6848-6. DOI 10.1007/978-1-4614-6849-3.

[13] Natekin A, Knoll A. "Gradient boosting machines, a tutorial". Frontiers in Neurorobotics, December 2013. https://doi.org/10.3389/fnbot.2013.00021.

[14] Teramoto R. "Balanced Gradient Boosting from Imbalanced Data for Clinical Outcome Prediction". Statistical Applications in Genetics and Molecular Biology. Volume 8, Issue 1, Pages 1–19, ISSN (Online) 1544-6115, DOI: https://doi.org/10.2202/1544-6115.1422, April 2009.

[15] Click C, et al. "Gradient Boosted Models with H2O". Published by H2O.ai, Inc. 2016.

[16] Nykodym T, et al. "Generalized Linear Modeling with H2O". Published by H2O.ai, Inc. 2016.

[17] Candel A, et al. "Deep Learning with H2O". Published by H2O.ai, Inc. 2016.

[18] Provost F, Fawcett T. "Data Science for Business". O'Reilly Media. July 2013.

Carlos M. Elvira

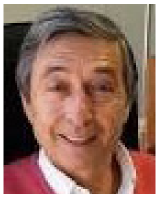
Carlos M. Elvira is M.D. since 1.998 from Universidad de Cantabria (Spain) and Ph.D. since 2.012 from Universidad Rey Juan Carlos (Spain). He is currently chief of the "Admission, Codding and Health Information Department" at Hospital Clínico San Carlos (Madrid-Spain). He is also a member of the Public Health and Medicine History Department at the Faculty of Medicine- Universidad Complutense of Madrid as associate professor.

Alberto Ochoa

Alberto Ochoa has a PhD. in Computer Science from the International Centre for Informatics and Electronics, Moscow 1992. He is one of the founders of Estimation of Distribution Algorithms (EDAs), a branch of evolutionary computation that combines statistical machine learning and evolutionary theory to build predictive models of objective functions. For over 20 years led research projects in evolutionary optimization, image analysis, complex networks, parallel and distributed computing, probabilistic graphical models and applications of information and copula theories to optimization and machine learning. Senior Data Scientist at Zed Worldwide during the last three years.

Juan Carlos Gonzalvez

Juan Carlos Gonzalvez has a Bachelor Degree in Chemistry from Universidad Autonoma de Madrid and he's got an MBA from IE Business School. He counts on more than eighteen years' experience in Telecom, Technology, Media and Internet sectors. He is currently Chief Innovation Officer at the Zed Worldwide Spanish multinational company, where he is currently working in the following innovation lines: Big Data, Advertising, Mobile Payments and Mobile Financial Services or Security and Privacy among others. He leads several R&D projects, in collaboration with different European universities, funded by various public and semi-public Spanish and European Institutions, under the multiple existing R&D program aids, like Horizon 2020.



Francisco Mochón

Francisco Mochón has a PhD in economics from the Autonomous University of Madrid and from Indiana University and is a Fulbright scholar. Currently he is full Professor of Economic Analysis at UNED, Madrid. He has been Advisor to the Ministry of Economy and Finance of Spain, Director General of Financial Policy of the Government of Andalusia, CEO of the research firm ESECA and Chief Financial Officer (CFO) of Telefónica of Spain. He has been the Chairman of the Social Board of the University of Malaga. Currently Prof. dr. Mochón is a member of the advisory committee of U-TAD and member of the advisory committee of the Futures Market of Olive Oil (MFAO). He has published numerous research articles and is the author of more than fifty books on economics, finance and business. Currently his research interests are the Economics of Happiness in the business environment and the Digital Economy. He has been director of the MOOC course "Felicidad y práctica empresarial".