Anales del Instituto de Actuarios Españoles, 4ª época, 25, 2019/119-145

SCORING AND PREDICTION OF EARLY RETIREMENT USING MACHINE LEARNING TECHNIQUES: APPLICATION TO PRIVATE PENSION PLANS

CALIFICACIÓN Y PREDICCIÓN DEL RETIRO TEMPRANO USANDO TÉCNICAS DE MACHINE LEARNING: APLICACIÓN A LOS PLANES PRIVADOS DE PENSIONES

Jose de Jesus Rocha Salazar[1] · María del Carmen Boado-Penas[1*]

[1]*Institute for Financial and Actuarial Mathematics, University of Liverpool, Liverpool, Liverpool L69 7ZL, United Kingdom.*

Fecha de recepción: 11 de octubre de 2019
Fecha de aceptación: 9 de noviembre de 2019

Abstract

Artificial intelligence techniques have become very popular in public and private organizations since they allow a more accurate decision-making process. Private insurance companies have ventured into this field by implementing algorithms that allow a better understanding of available data. The knowledge of retirement decisions allows the insurance companies to detect early retirement at a given time so that they have the adequate budgetary provision in place. In this paper, machine learning algorithms and data from private pension plans are used to predict whether a person retires before or after 65 years old in function of both individual characteristics and macroeconomic factors.

José de Jesús is grateful for the financial support from the National Council on Science and Technology (CONACYT) CVU-413231. María del Carmen Boado-Penas is grateful for the financial assistance received from the Spanish Ministry of the Economy and Competitiveness [project ECO2015-65826-P].

*Corresponding author (carmen.boado@liverpool.ac.uk)

DOI 10.26360/2019_6

Keywords: Insurance company, machine learning, early retirement, supervised learning.

Resumen

Las técnicas de inteligencia artificial se han vuelto muy populares en las organizaciones públicas y privadas debido a que permiten un proceso de toma de decisiones más preciso. Las compañías de seguros privadas se han aventurado en este campo mediante la implementación de algoritmos que permiten una mejor comprensión de los datos disponibles. El conocimiento de las decisiones de jubilación permite a las compañías de seguros detectar el retiro temprano en un momento dado para tener una provisión presupuestaria adecuada. En este documento, los algoritmos de aprendizaje automático y datos de planes de pensiones privados se utilizan para predecir si una persona se jubila antes o después de los 65 años en función de características individuales y factores macroeconómicos.

Palabras clave: Compañía de seguros, aprendizaje automático, retiro temprano, aprendizaje supervisado.

1. Introduction

The retirement began to be institutionalized and become a reality in the life of workers in the 19th century (Costa, 1998). Later it began to expand and gain strength especially in developed and industrialized countries in the 20th century (Atchley, 1996). Two reasons explain its emergence. The first one is that workers agreed with their employers the retirement age because at some point in their lives they were going to face weakening of their physical abilities to perform the job (Atchley, 1993). Second, old workers faced difficulties with finding work because employers preferred younger workers, limiting their obtaining of income and making necessary the development of a system that provided them financial sustenance in this stage of life. The retirement is colloquially defined as a change of state in the

working life of individuals, a definitive abandonment of the labour force (Zickar, 2012). Definitive abandonment is key in this definition because many individuals change jobs several times throughout their lives passing through a short period of unemployment. This short period of time away from the labour force does not necessarily mean a permanent retirement (Denton and Spencer, 2009). Retirement is also defined as the event where workers leave the labour force due to a decrease in psychological commitment to work and an increase in behavioural withdrawal (Wang and Shi, 2014). The individual's new financial support would be the pension, which is different to the salary from an economic activity (Szinovac, 2003).

The study of retirement decisions has strengthened during the last years due to the negative impact that early retirement has on the sustainability of public pension and profits of private ones. It is commonly known that traditional public pension systems such as pay as you go have a strong dependence on the demographic structure of countries. This makes that factor such as the increase in life expectancy and a decrease in birth rate cause loss of solvency (Turner, 2009; Heinrich and Weil, 2012). The early retirement encourages this effect turning workers into solely consumers and paying the pensions for longer. In the private pension schemes where an authorized financial company accepts the resources from a natural person in order to invest and ensure a pension in the old age, the demographic structure has not the same impact. These ones are affected mostly by variables such as proportion of contributions, fluctuations of interest rates, number of pensioners and early retirement. The early retirement in private pension plans involves fewer contributions to the system, less investment, and therefore lower net earnings.

Several reasons can lead individuals to an early retirement. Some studies explain that macroeconomic factors such as unemployment rate and stock market affect the decision of people to advance their retirement or delay it (Bosworth and Burtless, 2010; Coile and Levine, 2011). Studies such as that of Breinegaard, Jensen, and Bonde (2017) associate early retirement with organizational and management factors. Stress at work has also been a factor associated with early retirement (Wang, Zhan, Liu, and Shultz, 2008; Lin, 2001). Other

articles show that personality traits can predict the timing and routes of people's retirement (Blekesaune and Skirbekk, 2012; Feldman and Beehr, 2011). The policies and conditions for the claim of pensions can encourage early retirement since people tend to claim their pension as soon as they meet the conditions of age and time worked as established by Muldoon and Kopcke (2008). Also, the absence of recessions and the presence of a stable economic environment, stimulate an early retirement since workers have no personal financial reasons to stay in the workforce for longer (Hurd and Rohwedder, 2010).

In the context of a private company that offers personal pension plans, the prediction of early retirement is of high interest because it allows foreseeing how much and when the budgetary provision will be needed to make the pension payments almost immediately. Unfortunately, and despite the studies to understand the early retirement, a methodology, and computing system has not been developed to track early retirement of individuals in real-time and act on it, by implementing it, in the internal systems of a company.

The purpose of this research is to develop an early retirement score system using machine learning techniques to predict whether an individual with certain characteristics and macroeconomic conditions is more likely to retire before or after 65 years old[1]. This analysis uses data offered by a private insurance company in Mexico that provides pension plans with certain regulation. This regulation allows workers to claim pension benefits before age 65 under certain conditions. We acknowledge that the regulatory framework has an impact on the individuals' decisions. However, we will not discuss this regulation since the objective of this paper is purely computational, i.e. recognition of patterns based on observed claims and individuals' characteristics.

After this introduction, the paper is structured as follows. Section 2 gives summarised explanations about the use of machine learning as predictive tools. In section 3 the data used in the analysis is presented.

[1] 65 years old is the threshold commonly used to define early retirement (Feldman, 1994).

Section 4 presents the methodology used in the study. Section 5 shows the results. Section 6 provides conclusions.

2. Machine Learning as Predictive Tool and Advantages

Constantly companies take important decisions based on predictions made by their internal systems. Traditionally, the techniques used to make predictions were based on statistical models such as regressions. Recently, machine learning algorithms have spread and dominated the internal processes of companies due to the ease of their computational implementation and production of reliable and predictive models (Talwar and Kumar, 2013). For example, platforms such as Amazon and Netflix use these models to predict the preferences of the customers based on their previous choices (Bell, Koren, and Volinsky, 2008). Models for prediction of weather changes also use these algorithms and update continuously online as the environmental conditions change (Alavi, Gandomi, and Larry, 2016). Within the financial sector, machine learning algorithms such as decision tree and neural networks have been used to detect credit card fraud. These models help companies to minimize the loss from this financial crime (Bolton and Hand, 2001; Delamaire, Abdou, and Pointon, 2009; Juszczak, Adams, Hand, Whitrow, and Weston, 2008; Pozzolo, Caelen, Le Borgne, and Waterschoot, 2014). Studies in the insurance sector have used artificial neural networks to evaluate the financial capability of insurance companies and predict insolvency (Olaniyi, Ajibola, Ibiwoye, and Sogunro, 2012). Other studies have used feedforward neural networks with the back-propagation algorithm to build decision models for five insurances including life, annuity, health, accident, and investment-oriented insurances (Lin, Huang, and Lin, 2008).

Machine learning has some advantages over traditional statistical and econometrical models. First, in data processing machine learning algorithms are more efficient in the manipulation of big data sets. Second, machine learning allows more effective ways of model

complex relationships than simple linear regressions (Varian, 2014). Third, conventional statistical models as those ones used in econometrics focus more on finding causality relationships between the variables by estimating partial correlations (ceteris paribus) and under certain statistical assumptions.

Machine learning focuses on the prediction and classification using data, very computational not necessarily in a statistical way. Machine learning is not very focused on finding causality between the variables (Zheng et al., 2017). Finally, another attractive advantage of these algorithms is learning. Previously companies processed information with data available at a certain time, and calibrated the models in the systems for that period of time. Subsequently, there was no updating in the parameters of the model despite the fact that new data was constantly being generated. The innovative idea of machine learning is that methodologies are designed for computers to learn constantly as data is generated over time. This makes predictions of the future more accurately.

A new trend is arising in which it is recommended to use both traditional statistical models and machine learning algorithms in an integral way to strengthen the result on prediction both in short and long term. There is evidence that machine learning may be better at making short-term predictions and statistical techniques as econometrics turn out to be better in the long term. This is because machine learning methodologies deal with the heterogeneity of data and therefore are better at capturing short-term predictions. Econometrics, on the other hand, is better with long-run trends, i.e. linear or regular patterns (Liu and Xie, 2018).

In the case of this study, only machine learning algorithms will be used to make predictions since the objective of the model is purely of computational interest to create a score to predict early retirement.

3. Data

For the purpose of this study, information on retirement events was obtained from an important private insurance company in Mexico which offers personal pension plans. These pension plans can be acquired by natural people on a voluntary basis and under certain agreements. The available information consists of individual characteristics of each event of retirement from 2005 to 2017, totalling 1,500 claims in the period. These claims coincide with the ones reported to the Mexican Institute of Social Security (IMSS)[2]. This indicates that the sample is composed of definitive retirements.

In order to consider the macroeconomic environment, variables such as the return of government bonds, stock market, and the unemployment rate were obtained from the database of the Bank of Mexico.

The representation of the variables[3] in the algorithms is shown in the table 1.

Table 1
Variables Used in the Algorithms

Variable	Representation
Gender	This variable takes value 1 if the person is male and 0 female.
Disease	This variable is 1 if the person has a condition related to the 10 most common mortal diseases in Mexico and 0 otherwise.
Level of Education	1 means the person has elementary or secondary education, 2 high school, 3 graduate studies and 4 postgraduate studies.
Marital Status	1 if the person is married, 0 otherwise.
Salary	The last salary obtained at the moment of retirement.
Employment Status	This variable takes value 1 if the person is an employer and 0 if the person is an employee.
Dependants	The number of people who depend economically from the

[2] IMSS is the institution in Mexico which manages the public pensions governed by the law of 1973.
[3] It would have been interesting to include variables related to tax incentives but the company did not provide this information.

	pensioner.
Unemploy-ment Rate	Unemployment rate at the moment of retirement.
Credit Score	The score fluctuates between 400 and 850, 850 is the best rating.
Return of Government Bonds	Return of the government bonds at the moment of retirement.
Stock Market	Stock Market at the moment of retirement.

Source: Own elaboration. Characteristics observed in the sample of claims in private retirement plans.

The construction of the dependent variable considers as early retirement that one which occurs before 65 years old, according to Feldman (1994) and Wang and Alterman (2017). Thus, the dependent variable will be a binary one that takes 1 if the person retired before 65 years old and 0 otherwise.

3.1 Individual characteristics and expected trends

Gender

The total sample has 909 men and 591 women. According to statistics from OECD studies, in 2016 the man was the main financial support in OECD countries having an employment rate of 78% in Mexico and 74% in the rest of the countries against of 44% and 58% in women, respectively[4]. Commonly, the women are who stays at home doing the housework. Under this context, it is expected that being woman increases the probability of early retirement since in most of the cases she has not a strong financial responsibility to work for longer.

Disease

The health has been a common predictor of retirement (Clark and Spengler, 1980). A deteriorated health condition is an important cause of early retirement according to Leinonen, Laaksonen, Chandola, and

[4] See socioeconomic studies of the OECD, México, January 2017.

Martikainen (2016). Also, studies such as Burtless and Quinn (2000) explain that workers with poor health status and who perform very demanding jobs tend to be the first to retire. The health conditions of clients are monitored constantly by insurance companies due to the appearance of physical disabilities which are some of the risks covered by the personal pension plans. The database for this analysis provides information on whether the worker has health conditions related to the most common mortal diseases in Mexico. According to INEGI statistics[5], these diseases are heart illness, diabetes mellitus, cancer, liver disease, cerebral-vascular illnesses, lung diseases, pneumonia, kidney failure, congenital malformations, and bronchitis.

Salary

The pattern observed in the last years of the twentieth century suggests that individuals prefer to remain in the workforce for longer to get increased income (Atchley, 1989). Brown (1996) shows that high social class is related to late retirement. Individuals of lower social status generally feel less satisfaction, autonomy, and appreciation at work which make them retire early. From this evidence is expected that individuals with lower salaries are more likely to retire early. The average salary in the sample is around 1,300.00 USD a month. The salary in the sample varies from 750 to 2,000 USD, which is considered the salary of the middle class and upper-middle-class in Mexico.

Dependants

Espenshade, Kamenske, and Turchi (1983) show that larger families tend to have more consumption to short term and the per capita income declines. Also, studies such as Gensler (1997) indicate that large families are related to low wages and education. Based on this, it is expected that individuals with major number dependants work for

[5] For more details see the statistics of INEGI.
http://www.inegi.org.mx/est/contenidos/proyectos/registros/vitales/mortalidad/tabulados/ConsultaMortalidad.asp.

longer to compensate and improve the low welfare in the family. In the sample of the study, 34.71% of individuals have 4 dependants while only 0.73% has 1 dependant. At first glance, it seems to be a large number of dependents compared to European countries or northern America. The sample considers people who have retired in the last 12 years, people who were born in the fifties and who began to procreate in the sixties and seventies. In Mexico, according to statistics from the National Population Council (CONAPO), the average number of children per woman in the seventies was 6.1. Taking this into account, it is seen that the number of dependants in the sample is below the average for that generation.

Level of education

Stenberg and Westerlund (2013) using longitudinal population register data in Sweden from 1982 to 2010 find that higher education increases the labor survival rates of individuals aged 61-66 by around 5%. Other researches such as Peracchi and Welch (1994) suggest that skilled workers that are often related to higher education are less likely to leave full-time employment. Form this, it is expected that higher education decreases the probability of early retirement.

In the sample, over 23% of individuals have graduate studies, over 25% elementary education or high school and over 26% have postgraduate studies. This means that most of the workers in the sample have postgraduate studies. Also, workers with elementary education or high school have a high frequency. According to statistics from the National Employment and Occupation Survey (ENOE), the generations of the sixties and seventies used to finish only elementary education and high school. This explains the high frequency of individuals with only basic education. Over the years and as Mexico grew economically and became involved in a globalized and competitive world, individuals chose to acquire higher educational degrees. On the other hand, the high frequency of individuals with postgraduate studies in the sample may be due to the profile that is commonly associated with a private insurance company. Individuals with higher educational levels tend to have better jobs,

higher income and therefore have the possibility of choosing a private personal plan.

Credit score

This variable was included as a measure of an individual's financial health regarding personal debt. The greater the score the better the financial health of the person. Usually, a person with a high personal debt and bad financial management will remain working for longer to discharge his debt. This variable is commonly used in the banking sector to predict the credit default of consumers and companies. Individuals and companies with low scores are rejected or given smaller credits. The sample is composed by workers who have a score between 600 and 800[6]. In general, it is a good credit score.

Marital status

The marital status is a variable that can also influence early retirement (Figueira, Haddad, Gvozd, and Pissinati, 2017). For example, if the spouse receives a high income that allows the couple to have a good quality of life, the other one might retire early. The possibility of early retirement can increase even more if the person is a woman as explained before. The sample is composed by 846 individuals who are married.

Employment status

The social environment at work is often the second family and home of the employee. This is why conditions at work and how comfortable employee feels influence retirement decisions. If the worker does not have a good relationship with colleagues or the environment is simply unfriendly and stressful, he will retire early (Elovainio et al., 2005; Carr et al., 2016). This situation disappears when the individual is an

[6] For more details see the interpretation of the score, https://www.burodecredito.com.mx/

employer and can decide the environment he wants to work in. The sample for this analysis has 602 individuals who are employers.

3.2 Economic environment and expected trends

Government bonds

This variable was introduced as an indicator of a fixed rate and was obtained from statistics of the Bank of Mexico. In the sample, the return of government bonds faced by individuals varies in a range of between 6% and 8%.

Unemployment rate and stock Market

These variables were obtained from statistics of Bank of Mexico and were included as macroeconomic indicators that can influence retirement decisions. Bosworth and Burtless (2010) explain that when the unemployment rate decreases, people tend to retire early because there are no jobs in the market. Coile and Levine (2011) show that fluctuations in the stock market have an impact on the retirement decisions of workers with a high level of education. Particularly, decreases in stock market cause a delay in retirement because people decide to stay in the labour market longer to rebuild their lost wealth.

The trends explained above, imply in their majority relationships of causality and were mentioned with the objective of supporting the inclusion of the variables in the machine learning algorithms. Since machine learning does not focus on causality some of these trends might change on the recognition of patterns.

4. Methodology and Model

4.1 Learning process

The vital part of machine learning algorithms is "learning". It is assumed that there exists a function "F" applied to a set of data and the objective of the apprentice system is to guess what type of function it is. A hypothesis is created about the function to be learned, let us say "L". It is understood that "$L(x)$" is the function to be estimated and implemented by the system that has as input the vector $x = (x_1, x_2, \ldots, x_k)$ with "k" components. The creation of "L" is based on a set of data called training set with "n" input vectors. In several cases, there is also a test set to evaluate the fit and accuracy of the "L" function with respect to the "F" function. The learning process in these algorithms can be done in two ways: supervised learning and unsupervised learning. The learning that is a matter for us in this paper is supervised learning.

Formally, a supervised learning consists of a set of "n" ordered pairs $(x_1, y_1) \ldots (x_n, y_n)$, where x_i is a vector of characteristics and y_i is the label of the vector. In the case of this study, the vector x_i represents the characteristics or attributes of retired people and y_i is the classification of the pensioner as "retired before 65 years old" or "retired at 65 or after". The test data will be a set of "m" vectors without classification, $(x_{n+1}, \ldots, x_{n+m})$. The goal is to label the test set as "retired before 65 years old" or "retired at 65 or after" based on the learning from the training set.

4.2 Supervised learning algorithms

This paper is focused on three models of supervised learning to see the behaviour of data in each one and select that one with the best performance to create the scoring of early retirement prediction. These three models are the most used for binary classification problems.

Random Forest

Random forest is a classifier that consists of a collection of tree-structured classifiers $\{h(x, k), k = 1, \ldots\}$ where the $\{k\}$ are independent identically distributed random vectors and each tree casts

a unit vote for the most popular class at input x (Breiman, 2001). It is a substantial modification of bagging that builds a large collection of de-correlated trees and then averages them. In many ways, the performance of random forests is very similar to boosting, but they are simpler to train and tune.

The random forest technique can be used in two ways, to estimate a regression or classification. The regression is used when the output variable is continuous; the classification function is used when the output variable is categorical. For example, if we want to determine or predict the systolic pressure of a person based on height, weight and age, a regression would be used. If we want to determine if a person will retire before or after 65 years old (as is our case, yes/no) depending on his gender, health status, monthly salary, etc., the classification function would be used.

This technique has some advantages over other classification models. First, it avoids the problem of over-fitting. Second, if the variable is continuous the same algorithm and trees can be used. Finally, the random forest methodology identifies the most important and relevant variables for the prediction.

Logistic regression

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a *logit transformation* of the probability of the presence of the characteristic of interest.
This model starts with the following.

Let y_i be the binary variable and x_i the vector of explanatory variables. We define:

$$P_i = Prob(y_i = 1) = Prob\left[u_i > -\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j}\right)\right]$$
$$= 1 - F\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j}\right)\right] \quad (1)$$
$$= F\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j}\right)$$

Where F, is the comulative distribution function of u_i and P_i is the probability of retiring before 65 years old. $y_i = 1$ if individuals retire before 65 years old and 0 otherwise.

The model can be written in terms of odds:

$$\left(\frac{P_i}{1 - P_i}\right) = exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j}\right) \quad (2)$$

Or in terms of probability of early retirement occurring as:

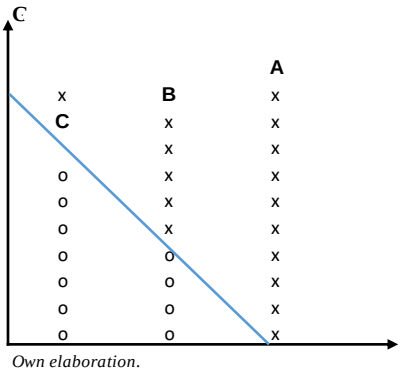
$$P_i = \frac{exp(x_i{'}\beta)}{1 + exp(x_i{'}\beta)} \quad (3)$$

Where P_i is the probability of early retirement, x_i the vector of explanatory variables and β the parameters to be estimated.

One of the advantages of the logistic regression is that it does not require all the assumptions of the linear regressions to generate good estimators. It is also possible to work with non-linear relationships between the dependent variable and the explanatory variables since a logarithmic transformation of the linear regression is used.

Support Vector Machine

The intuition behind Support Vector Machine is the following: consider the following graph, in which "x´s" represent individuals who retired early, "o's" denoting individuals who retired late and the decision boundary which is a line given by the equation $\theta x^T = 0$ (this line is also called the separating hyperplane). Three points have also been labelled A, B and C.



Own elaboration.

Graph 1.

SVM technique represents the sample points in space, separating the classes into 2 spaces as wide as possible by a hyper-plane defined as the vector between the 2 points, of the 2 classes, closest to which is called support vector. When the new samples are put in correspondence with said model, depending on the spaces to which they belong, they can be classified in one or other class (Cortes and Vapnik, 1995). For example, point A is very far from the hyper-plane, so it can be classified as early retirement with a high level of confidence. On the other hand, point C is closer to the borderline, being labelled as early retirement but not with the same level of confidence as point A.

Mathematically the problem is the following.

$$min_{\gamma,\omega,b}\frac{1}{2}|\omega^2| \quad (4)$$

$$s.t.\ \ y^i(\omega^T x^i + b) \geq 1, i = 1,2,\ldots,m$$

Where γ is the distance to the decision boundary, (ω, b) is the orthogonal vector of the hyperplane and y^i the label of the training data. The objective is to find the optimal hyperplane that allows a correct classification.

5. Results

The data was divided into two parts: 70% for the training set and 30% for the test set. The test set was used to evaluate the fitness of the models. The split is made randomly and each model considers same training and test sets.

The performance of the model on the prediction can be measured by the confusion matrix. This matrix shows the proportion of true positives (proportion of positive cases that were correctly classified TP), false positives (proportion of positive cases that were incorrectly classified FP), true negatives (proportion of negatives cases that were correctly classified TN) and false negatives (proportion of negative cases that were incorrectly classified FN). The table 2 shows the confusion matrix of the three models.

Table 2
Confusion Matrix

Random Forest			
		Predicted	
		Early	Late
Real	Early	177	41
	Late	62	170
Logistic Regression			
		Predicted	
		Early	Late
Real	Early	173	35
	Late	66	176

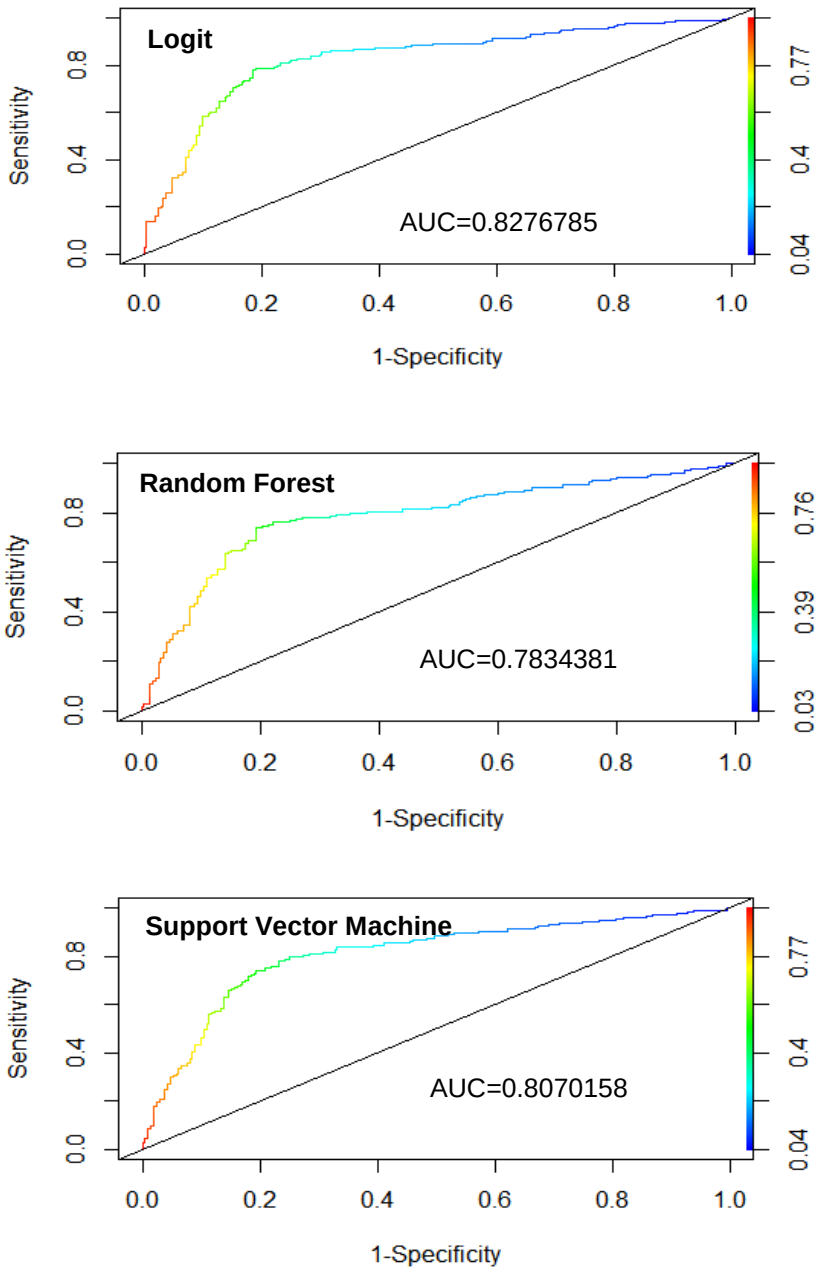
Support Vector Machine			
		Predicted	
		Early	Late
Real	Early	174	40
	Late	65	171

Own elaboration.

The confusion matrixes were constructed considering a threshold of 0.5. This means that if the predicted probability of the event in the test set was greater than 0.5, it classified as "Early Retirement", otherwise "Late Retirement". The numbers along the main diagonal represent the correct decisions made in the test set. From the table 2, it can be obtained that the decisions made correctly were 347, 349 and 345 for random forest, logistic regression and SVM, respectively. Although there is not much difference between these values, it is observed that logistic regression has the best performance to predict true events of early and late retirement.

The proportion of true positives is also known as *sensitivity*, which is the probability of predicting an early retirement on the test set when the person actually retired early. The proportion of true negatives is also known as *specificity* which means the probability of predicting a late retirement on the test set when actually the person retired late.

The measures of sensitivity and specificity are used to construct the ROC Curve which is the graph of 1-specificity versus the sensitivity for each possible threshold value or cut-off point on the scale of results of the test under study (Fawcett, 2006). Then, we have probabilities in both axes of the ROC curve, it will be contained in the square$[0,1]X[0,1]$. The graph 2 shows the ROC curve of the three models of this analysis.



Graph 2. ROC Curves**.** Source: Own elaboration. AUC denotes the accuracy of the models.

It is observed that the three models have similar shape and curvature on the ROC graphs. The area under the ROC Curve is known as accuracy (AUC) and represents the probability of discriminating correctly between early and late retirement.

Despite the small differences, the logistic regression keeps as the model with the best performance based in accuracy. Another advantage of the logistic regression is the ease with which the score is created. Usually, the creation of scores with vector machine and random forest requires additional steps in the internal algorithms of the systems, while with logistic regression the score is generated directly and naturally.

Table 3 shows the score for 2 clients in the test set based on the logit model.

Table 3
Score Based on Logit Model

Variables	**Individual 1**	**Individual 2**
Gender	Male	Female
Disease	No Disease	Disease
Level of Education	Postgraduate Studies	Elementary or Secondary Education
Marital Status	Married	Single
Salary	26,000	25,000
Employment Status	Employer	Employee
Dependants	4	5
Unemployment rate	2.93	4.6
Credit Score	624	694
Return of Government Bonds	8.89	6.05
Stock Market	19,520.66	44,100.16
Score of Early Retirement (Probability of early retirement)	**0.0404**	**0.9495**

Source: Own elaboration.

Certain expected patterns can be observed from the attributes of these two individuals and the scores obtained. Individual 1 has a higher level of education than individual 2, which indicates that workers with a lower level of education tend to retire early. Also, Individual 1 earns a little bit more than individual 2, which supports the idea that workers with higher incomes tend to work for longer. Individual 1 is male and 2 female, which implies that men remain in the labour force for a longer time than women. Individual 2 is subject to a higher unemployment rate than individual 1, which confirms the idea that a high unemployment rate pushes worker to early retirement. Individual 2 is subject to a higher stock market confirming the trend that a high stock market is related to early retirement. Regarding the returns on bonds, the individual 2 who faced a lower return, obtained a higher score(probability) of retirement before 65 years. The individual 1 has a lower credit score than individual 2 and less score(probability) of early retirement, which coincide with the expected trend of this variable. As for health, Individual 2 has a condition related to some of the mortal diseases, which make her very likely to retire early.

6. Conclusions

This study showed how to create a score system to predict early retirement using personal attributes and macroeconomic variables. The algorithms used for the analysis were support vector machine, logistic regression and random forest. Logistic regression was the model with the best performance and the most suitable for scoring.

The scoring system can be used in any insurance company that has the necessary data to make predictions. The threshold used in this analysis is 65 years old because this is the normal retirement age and the regulation allows to retire before and after this age. In other cases, with a more restrictive or different regulation, a higher (or different) threshold can be fixed. This study can also be done beyond the concept of early retirement and analyse whether a worker retires before or after a certain age with purely predictive purposes.

The advantage of machine learning techniques is that new variables can always be incorporated into the model in order to improve the accuracy of the results. This work also gives an example of how data and machine learning methodologies might be used, if the availability of data was bigger. For example, in the case of a movement from a public pension system to a private one of individual accounts, the availability of data would be BIG and machine learning algorithms would be suitable to process that amount of data.

The creation of score systems in pension schemes is innovative and offers insurance and government companies an alternative to track and control early retirement. Predictive data modelling is gaining strength in different fields of the industry and pension sector must implement artificial intelligence in its processes to improve its results.

Future research aims to analyse how the predictions would change if we apply the models in the public sector considering the country specific regulations or in countries with a different macroeconomic environment and different behaviour patterns. Also, the inclusion of variables related to tax incentives would strength the prediction if available.

References

Alavi, A. H., A. H. Gandomi, and Larry, D. J. (2016). The progress of Machine Learning in Geosciences: Preface. *Geoscience Frontiers*, 7(1), 21-31.

Atchley, R. C. (1989). A continuity theory of normal aging. *The Gerontologist*, 29(2), 183-190.

Atchley, R. C. (1993). *Critical perspectives on retirement. Voices and visions of ageing: Toward a critical gerontology*. New York: Springer, 3-19.

Atchley, R. C. (1996). Retirement. *Encyclopedia of Gerontology 2*, San Diego, CA: Academic Press, 437-449.

Bell, R., Koren, Y., and Volinsky, C. (2008). *The BellKor 2008 Solution to the Netflix Prize.*

Blekesaune, M., and Skirbekk, V. (2012). Can personality predict retirement behaviour? A longitudinal analysis combining survey and register data from Norway. *European Journal of Ageing,* 9(3), 199-206.

Bolton, R. J., and Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit Scoring and Credit Control,* 7, 235–255.

Bosworth, B. P., and Burtless, G. (2010). Recessions, wealth destruction, and the timing of retirement. *Working Paper* 2010-22. Chestnut Hill, MA: The Center for Retirement Research at Boston College.

Breiman, L. (2001). Random forests. *Machine Learning,* 45(1), 5-32.

Breinegaard, N., Jensen, J. H., and Bonde, J. P. (2017). Organizational change, psychosocial work environment, and non-disability early retirement: A prospective study among senior public employees. *Scandinavian Journal of Work, Environment and Health,* 43(3), 234-240.

Brown, M. T. (1996). Annual review, 1990–1996: Social class, work, and retirement behavior. *Journal of Vocational Behaviour,* 49(2), 159-189.

Burtless, G. and Quinn, J. F. (2002). Is working longer the answer for an aging workforce? *Issue Brief* No. 11. Chestnut Hill, MA: The Center for Retirement Research at Boston College.

Carr, E., Johnson, G. H., Head, J., Shelton, N., Stafford, M., Stansfeld, S., and Zaninotto, P. (2016). Working conditions as predictors of retirement intentions and exit from paid employment: A 10-year follow-up of the English Longitudinal Study of Ageing. *European Journal of Ageing.* 13, 39-48.

Clark, R. L., and Spengler, J. J. (1980). *The economics of individual and population aging*. Cambridge University Press.

Coile, C., and Levine, P. B. (2011). The market crash and mass layoffs: How the current economic crisis may affect retirement. *The B.E. Journal of Economic Analysis and Policy*, 11(1), 22.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

Costa, D. L. (1998). The evolution of retirement: An American economic history, 1880-1990. *The National Bureau of Economic Research*, 6-31.

Delamaire, L., Abdou, H., and Pointon, J. (2009). Credit card fraud and detection techniques: A review. *Banks and Bank Systems*, 4(2), 57–68.

Denton, F. T., and Spencer, B. G. (2009). What is retirement? A review and assessment of alternative concepts and measures. *Canadian Journal on Aging/La Revue Canadienne du Vieillissement*, 28, 63–76.

Elovainio, M., Van Den Bos, K., Linna, A., Kivimäki, M., Ala-Mursula, L., Pentti, J., and Vahtera, J. (2005). Combined effects of uncertainty and organizational justice on employee health: Testing the uncertainty management model of fairness judgments among finnish public sector employees. *Social Science and Medicine*, 61, 2501-2512.

Espenshade, T. J., Kamenske, G., and Turchi, B. A. (1983). Family size and economic welfare. *Family Planning Perspectives*, 15(6), 289-294.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.

Feldman, D. C. (1994). The decision to retire early: A review and conceptualization. *The Academy of Management Review*, 19(2), 285–311.

Feldman, D. C., and Beehr, T. A. (2011). A three-phase model of retirement decision making. *American Psychologist*, 66(3), 193-203.

Figueira, D. A. M., Haddad, M. C. L., Gvozd, R., and Pissinati, P. S. C. (2017). Retirement decision-making influenced by family and work relationships. *Revista Brasileira de Geriatria y Gerontologia*, 20(2), 206-2013.

Gensler, H. (1997). Welfare and the family size decision of low-income, two-parent families. *Applied Economics Letters*, 4(10), 607-610.

Heinrich, H., and Weil, D. N. (2012). On the dynamics of the age structure, dependency, and consumption. *Journal of Population Economics*, 25(3), 1019-1043.

Hurd, M. D., and Rohwedder, S. (2010). Effects of the financial crisis and great recession on American households. *The National Bureau of Economic Research Working Paper* 16407.

Juszczak, P., Adams, N. M., Hand, D. J., Whitrow, C., and Weston, D. J. (2008). The peg and bespoke classifiers for fraud detection. *Computational Statistics and Data Analysis*, 52(9), 4521–4532.

Leinonen, T., Laaksonen, M., Chandola, T., and Martikainen, P. (2016). Health as a predictor of early retirement before and after introduction of a flexible statutory pension age in Finland. *Social Science and Medicine*, 158, 149-157.

Lin, T. C. (2001). Letter to the editor: Impact of job stress on early retirement intention. *International Journal of Stress Management*, 8(3), 243-247.

Lin, Y. J., Huang, C. S., and Lin, C.C.C. (2008). Determination of insurance policy using neural networks and simplified models with factor analysis technique, WSEAS *transactions on Information Science and Applications*, 10(5), 1415-1425.

Liu, Y., and Xie, T. (2018). Machine learning versus Econometrics: prediction of box office. *Applied Economics Letters*, 26(2), 124-130.

Muldoon, D., and Kopcke, R. W. (2008). Are people claiming social security benefits later?. *CRR Issue Brief* Number 8-7. Chestnut Hill, MA: The Center for Retirement Research at Boston College.

OECD (2017). Socioeconomic studies of the OECD, México.

Olaniyi, O., Ajibola, E., Ibiwoye, A., and Sogunro, A.B. (2012). Artificial neural network model for predicting insolvency in insurance industry. *International Journal* of *Management and Business Research,* 2(1), 59-68.

Peracchi, F., and Welch, F. (1994). Trends in labor force transitions of older men and women. *Journal of Labor Economics,* 12(2), 210-242.

Pozzolo, A. D., Caelen, O., Le Borgne, Y.A., and Waterschoot, S. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications,* 41(10), 4915-4928.

Stenberg, A., and Westerlund, O. (2013). Education and retirement: does University education at mid-age extend working life? *IZA Journal of European Labor Studies,* 2, 16.

Szinovacz, M. (2003). Contexts and pathways: retirement as institution, process and experience. In Adams, G. A., and Beehr, T. A. (Eds.) *Retirement: Reasons, Processes, and Results.* New York: Springer, 6-52

Talwar, A., and Kumar, Y. (2013). Machine Learning: An artificial intelligence methodology. *International Journal of Engineering and Computer Science,* 2(12), 3400-3404.

Turner, A. (2009). Population priorities: the challenge of continued rapid population growth. *Philosophical Transactions of the Royal Society,* 364, 2977–2984.

Varian, H. R. (2014). Big Data: New tricks for Econometrics. *Journal of Economics Perspectives,* 28(2), 3-28.

Wang M., Zhan, Y., Liu, S., and Shultz, K.S. (2008). Antecedents of bridge employment: a longitudinal investigation. *Journal of Applied Psychology*, 93(4), 818–830.

Wang, M., and Shi, J. (2014). Psychological Research on Retirement. *Annual Review of Psychology,* 65, 209–233.

Wang, M., and Alterman, V. (2017). *Retirement*. Oxford Research Encyclopedia of Psychology. New York, NY: Oxford University Press.

Zheng, E., Tan, Y., Goes, P., Chellappa, R., Wu, D. J., Shaw, M., Sheng, O., and Gupta, A. (2017). When Econometrics meets Machine Learning. *Data and Information Management*, 1(2), 75-83.

Zickar, M. J. (2012). The evolving history of retirement within the United States. In Wang, M. (Ed.). *The Oxford Handbook of Retirement*. New York: Oxford University Press, 10–21.