# Predicting the risk of suffering chronic social exclusion with machine learning

Emilio Serrano*, Pedro del Pozo-Jiménez , Mari Carmen Suárez-Figueroa, Jacinto González-Pachón, Javier Bajo, Asunción Gómez-Pérez
emilioserra@fi.upm.es

Ontology Engineering Group, Universidad Politécnica de Madrid, Spain

**Abstract.** The fight against social exclusion is at the heart of the Europe 2020 strategy: 120 million people are at risk of suffering this condition in the EU. Risk prediction models are widely used in insurance companies and health services. However, the use of these models to allow an early detection of social exclusion by social workers is not a common practice. This paper describes a data analysis of over 16K cases with over 60 predictors from the Spanish region of Castilla y León. The use of machine learning paradigms such as logistic regression and random forest makes possible a high precision in predicting chronic social exclusion. The paper is complemented with a responsive web that allows social workers to calculate the risk of a social exclusion case to become chronic through a smartphone.

**Keywords:** Social exclusion, Social services, Data analysis, Machine learning, Data mining.

## 1 Introduction

*Social exclusion* is a complex and multi-dimensional process involving the lack of resources, rights, goods and services, and the inability to participate in the normal relationships and activities, available to most people in a society, whether in economic, social, cultural or political scopes [8]. Social exclusion affects not only the quality of life of individuals, but also the equity and cohesion of society as a whole. The economic crisis is undermining the sustainability of social protection systems in the EU [1]: 24% of all the EU population (over 120 million people) are at risk of poverty or social exclusion [1]. The fight against poverty and social exclusion is at the heart of the Europe 2020 strategy for smart, sustainable and inclusive growth.

In chronic medical diseases, there is strong evidence supporting that early detection results in less severe outcomes. This paper intends to provide social workers with methods and tools to bring this early detection, which is so beneficial in the medical field, to the challenging problem of chronic social exclusion. To this purpose, the paper contributes with (1) an analysis of the social services

* ORCID ID: 0000-0001-7587-0703

data of Castilla y León (CyL), which is the largest region in Spain and counts with around two and a half million inhabitants. This analysis allows getting insights into why social exclusion can become chronic. Furthermore, a (2) machine learning model capable of quantifying the risk of chronic social exclusion is build. Finally, a (3) a responsive web is deployed to allow queries by social workers through a number of devices such as smartphones, tablets, or laptops. A RESTful web service is also provided to integrate the predictive capabilities into other software applications.

The paper outline is as follows. After revising some of the most relevant related works in section 2, the process used to analyze the data is explained in section 3. Section 4 reports the outcomes of the experiments conducted. Section 5 explains, analyzes, and compares the results. Section 6 introduces the web service implemented. Finally, section 7 concludes and offers future works.

## 2 Related works

Risk prediction models are widely used in insurance companies to allow customers to estimate their policies cost. Manulife Philippines [2] offers a number of online tools to calculate the likelihood of disability, critical illness, or death before the age of 65; based on age, gender, and smoking status. Health is another application field where risk estimations are undertaken for preventive purposes. More specifically, the risk of heart disease can be estimated at different websites such as at the Mayo clinic web [3].

There are a number of data analysis works in social exclusion that are detailed enough to extrapolate some of their methods to the research presented here. Ramos and Valera [12] use the *logistic regression* (LR) model to study social exclusion in 384 cases labeled by social workers through a heuristic procedure. Lafuente-Lechuga and Faura-Martínez [7] undertake an analysis of 31 predictors based on segmentation methods and LR. Haron [6] studies the social exclusion in Israel and proposes the *linear regression* as a better alternative to the LR. Suh et al. [13] analyze over 35K cases of 34 European countries using LR. Although these works are significant contributions to the social exclusion problem; they do not provide social workers with an online tool or an implemented machine learning model to cope with social exclusion as presented here. Moreover, the use of linear classifiers exclusively such as LR may hinder models from achieving a better predictive power.

## 3 Methodology

The methodology employed for this data analysis research is the widely used *Knowledge Discovery in Databases* (KDD) process described by Fayyad et al. [5]. This includes the following steps: Selection, Preprocessing, Transformation, Data Mining, and Evaluation and/or Interpretation. Although the KDD is an iterative and incremental process, some of the decision made in the different steps are presented unlooped here for the shake of clarity.

### 3.1 Selection

Eleven databases (DBs) with social services information were available to select relevant data. More specifically, the DBs were implemented with the Oracle object-relational database management system.

After several meetings with the social workers experts, 63 relevant variables from those DBs were selected to further study and preprocess. The predictors were identified by their use in different applications by the social workers. Nonetheless, locating these variables in the DBs to select them was specially challenging: a mapping from the variables to the DB (schema, table, and column) was not available. For example, the SAUSS application[1] has a schema with over 800 tables under the hood, plus a large number of tables shared among other applications.

Another important decision made in this step, after several iterations in the methodology, was the class definition to represent a chronic social exclusion sate. Several prediction services were outlined but the main class to study was defined as "having received social aid during 60 months or more", not necessarily continuously.

### 3.2 Preprocessing

The preprocess phase included among others: (1) the data integration where multiple data sources from the selection are combined; (2) the data cleaning removing noise and inconsistent data such as negative income or dates of birth in the year 1900; (3) managing missing values; and, (4) dealing with the temporal dimension.

Regarding the missing values, a number of variables whose values are missing in over 90% instances were not considered. Moreover, a clear positive correlation between missing data and non-chronic social exclusion was observed in the exploratory data analysis. This supports the idea that these missing values are not random but indicate that the social worker has decided not to log a particular measurement. Therefore, a special value of "NR" (not registered) has been included. As Witten et al. [14] explain, people analyzing medical databases have noticed that cases may be diagnosed simply from the missing values indicating tests that a doctor has decided not to make. Imputing values in these cases would result in an information loss.

### 3.3 Transformation

In this phase, data are transformed into forms appropriate for mining. This includes, among others: (1) standardization of numeric variables; (2) transforming internal numerical codes into interpretative nominal values; (3) aggregation for the multi-instance learning (where each labeled case in the data comprises several different instances); and, (5) dealing with the imbalanced classification problem.

[1] `https://sauss.jcyl.es/sauss-sso/`

The result of this process is a dataset with 63 predictors (some of the most important ones are described in section 5) and 16535 instances: 4205 of the positive class and 12330 of the negative class. This situation is known as imbalanced classification: a high accuracy is achieved by just predicting always the negative class. Some approaches to cope with this situation include: (1) penalized models; (2) undersampling the over represented class (negative);(3) oversampling the underrepresented class (positive); and, (4) generating synthetic samples. Section 4 shows several experiments in these lines.

### 3.4 Data mining

In this phase, machine learning paradigms are applied to create a hypothesis that explains the observations. The *logistic regression* (LR) is widely use to predict the risk of social exclusion as explained in related works, section 2. Furthermore, it is an intuitive solution when a class prediction is wanted with a degree of confidence. Experiments were also conducted, although not shown in this paper, using *decision trees* (which typically tolerate imbalanced data) and *rule-based classifiers* (whose hypotheses are highly interpretable for social workers). Meta-classifers such as *Boosting* and *Random Forests* (RF) were also considered given their higher predictive power. Besides, these are good of the box solutions that improve the maintenance when rebuilding new machine learning models in the light of new cases.

### 3.5 Evaluation

For the evaluation of the models, the cross validation is typically considered when the performance allows it. Using 10 folds involves rebuilding the machine learning model for the data 11 times. Nonetheless, when oversampling methods are applied, the cross validation method leads to overoptimistic results since the testing fold considers instances that are also present in the training folds. Thus, the classic partition between training and testing data was undertaken ensuring:

- the splitting (80/20% is considered) preserves the overall class distribution of the data (25% of positive cases);
- and, the oversampling is performed after this splitting both in the training and the testing data.

## 4 Results

Table 1 details the accuracy, precision, and recall for LR and RF. For the logistic regression, the experiments use a multinomial logistic regression model with a ridge estimator implemented in Weka [14]. This model requires adjusting the ridge parameter for underfitting or overfitting situations. For the random forest, the randomForest package of R [9] is employed with its defaults parameters, which include the use of 500 decision trees.

The table shows 9 different experiments including classification: (1) with the imbalanced data; oversampling the positive class with random sampling with replacement (2) before and (3) after splitting the testing data; oversampling the positive class with SMOTE [4] (4) before and (5) after separating the testing data; oversampling the positive class with ROSE [10] (6) before and (7) after splitting the testing data; and, undersampling the nagative class (8) by random sampling with replacement and by (9) the *K-medioids* [11] segmentation method.

| | Logistic Regression | | | Random Forest | | |
|---|---|---|---|---|---|---|
| **Experiment** | **Accuracy** | **Precision** | **Recall** | **Accuracy** | **Precision** | **Recall** |
| Imbalanced | 81.3% | 69.3% | 47.4% | 80.6% | 71.3% | 40.3% |
| Oversampling 1 | 60.5% | 78.5% | 29% | 91.5% | 93.1% | 89.7% |
| **Oversampling 2** | 55.5% | 78% | 15.4% | 67.8% | **88.6%** | 40.9% |
| **SMOTE 1** | 54.9% | 82.1% | 12.5% | 89.2% | **96.1%** | 81.8% |
| SMOTE 2 | 67.9% | 72.6% | 57.5% | 82.4% | 88.3% | 74.7% |
| ROSE 1 | 58.8% | 56.2% | 73.4% | 88% | 90.7% | 84.6% |
| ROSE 2 | 61.4% | 59% | 74% | 73.4% | 81.7% | 57.4% |
| Undersampling 1 | 60.5% | 62.7% | 62.7% | 75.2% | 77.6% | 70.9% |
| Undersampling 2 | 59.9% | 59% | 65.3% | 74.6% | 78.3% | 68.1% |

**Table 1.** Experiments results.

## 5 Discussion

The results show that for the imbalanced dataset the accuracy is reasonable but precision and recall are very low. This situation is known as *accuracy paradox*: the accuracy is only reflecting the underlying class distribution. Besides, the *precision*, considering the chronic social exclusion as positive class, is the most interesting metric because it allows social workers to find hazardous cases and to focus limited resources on them. Precision is a measure of quality as recall is a measure of quantity. High precision means: when the prediction says positive, the social worker can be very confident that it will be a chronic case.

Another clear result is that, as expected, the oversampling methods offer better results when the validation instances are extracted after oversampling the positive class. In "oversampling 1" this happens because several instances that are exactly the same are considered both for training and testing. When more advanced methods are employed such as SMOTE and ROSE, instead of duplicating cases, new instances are created by generalizating the points where the minority class is valid. Therefore, splitting a validation set after using SMOTE and ROSE is possible although it leads to optimistic results. Considering this, the two machine learning models selected are "oversampling 2" as a *conservative prediction*, and "SMOTE 1" as an *optimistic prediction*.

As explained above, the precision is the most interesting metric for the models purpose. In this vein, the selected conservative and optimistic models achieve

a high precision: 88.6% and 96.1%, respectively. These results are especially remarkable when compared to the imbalanced alternatives.

## 6 Implementation

A web service for predicting the risk of suffering chronic social exclusion based on the machine learning models explained has been deployed and is being considered for integration into the social services of CyL. Due to usability reasons, instead of considering all the 63 predictors, only the ten most important variables are asked by the web service:

1. Age: calculated from day of birth to current date or date of death.
2. Level of studies: an ordinal indicator from illiterate to "higher education or vocational training".
3. Classification code: preliminary evaluation label given by the social workers whose values may be temporary, structural, undecided, or (as in most cases) unknown.
4. Annual income in euros: the training dataset contains a great deal of missing data for this variable, but it becomes highly relevant after imputing an average value.
5. Economic activity code: classified by the Spanish Ministry of Employment and Social Security.
6. Civil status.
7. Year of registration in local government.
8. Number of years as job seeker.
9. Professional qualification code: another indicator of education level ordering professional qualifications subject to recognition and accreditation. The code is given by the Spanish Ministry of Education, Culture and Sport.
10. National or foreigner: this binary variable (ternary with the "non registered" value) was obtained merging a number of nationality codes, most of them too unusual to offer a generic hypothesis about chronic social exclusion.

Limiting the attributes to 10, the accuracy, precision and recall of the model are 86.3%, 89.9% and 81.9%, respectively.

Figure 1 shows, on the left, the presentation website and, on the right, the form to query the service about a case. The *Bootstrap* front-end web framework has been employed for designing the website and web application. This ensures the web responsiveness and allows social workers to access the service from a number of devices such as: computers, tablets, and smartphones.

The prediction returns a risk percentage for chronic social exclusion and it is considered a positive case when the risk is over 50%. The same queries may be conducted via RESTful service by introducing the query string parameters in the URL.


**Fig. 1.** A web service to detect risk of chronic social exclusion.

## 7 Conclusion and future works

This paper introduces a service to predict the risk of suffering chronic social exclusion with machine learning. With a precision around 90% in the most conservative predictions, it offers a quick rule of thumb that can detect citizens who are in danger of been excluded of the society beyond a temporary situation. The application offers a RESTful web service for software applications and a responsive web interface to be consulted by social workers from their smartphones. An early detection is possible thanks to this service and hence, as in medical diseases, the recovery process can be accelerated.

This service is based on an intelligent model that is fed with data from a whole Spanish region: eleven databases from the social services of Castilla y León. The service is being considered for integration into the social services of CyL. The classical Knowledge Discovery in Databases (KDD) process has been used and instantiated to the particularities of the data and application field. The results of the analysis reveal the age as the most relevant factor for chronic social exclusion. Besides, five of the top ten predictors are work or education related.

The future works in this research include but are not limited to: the use of deep learning techniques for feature extraction, the consideration of unlabeled cases to train supervised neural networks, the inclusion of more potentially relevant predictors in the studied observations, and the generation of new prediction models.

## Acknowledgments

This publication would not have been possible without the inputs and collaboration of the Social Services of Castilla y León. This research work is supported by the "Junta de Castilla y León" under the public contract: "Servicios de elaboración de modelos matemáticos para realizar segmentación poblacional" (A2016/000271); by the EU Programme for Employment and Social Innovation (EaSI) under the project PACT ("people-oriented case management for social inclusion proactive model"); and, by the Spanish Ministry of Economy, Industry and Competitiveness under the R&D project Datos 4.0: Retos y soluciones (TIN2016-78011-C4-4-R, AEI/FEDER, UE).

## References

1. European Commission's DG for Employment, Social Affairs & Inclusion. `http://ec.europa.eu/social/main.jsp?catId=751`. Accessed: February of 2017.
2. Manulife Philippines. Calculate your risk, your partner's risk or both. `ttp://www.insureright.ca/what-is-your-risk`. Accessed: February of 2017.
3. Mayo Clinic. Heart Disease Risk Calculator. `http://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-disease-risk/itt-20084942`. Accessed: February of 2017.
4. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002.
5. U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Advances in knowledge discovery and data mining. chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.
6. N. Haron. *On Social Exclusion and Income Poverty in Israel: Findings from the European Social Survey*, pages 247–269. Springer US, Boston, MA, 2013.
7. M. Lafuente-Lechuga and U. Faura-Martínez. An álisis de los individuos vulnerables a la exclusión social en españa en 2009. *Anales de ASEPUMA*, (21), 2013.
8. R. Levitas, C. Pantazis, E. Fahmy, D. Gordon, E. Lloyd, and D. Patsios. *The multidimensional analysis of social exclusion.* London: Social Exclusion Task Force, Cabinet Office, 2007.
9. A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.
10. N. Lunardon, G. Menardi, and N. Torelli. Rose: A package for binary imbalanced learning. *R Journal, The*, 6(1):79–89, 2014.
11. H.-S. Park and C.-H. Jun. A simple and fast algorithm for k-medoids clustering. *Expert Syst. Appl.*, 36(2):3336–3341, Mar. 2009.
12. J. Ramos and A. Varela. Beyond the margins: Analyzing social exclusion with a homeless client dataset. *Social Work & Society*, 14(2), 2016.
13. E. Suh, P. TiffanyVizard, and T. AsgharBurchardt. Quality of life in europe: Social inequalities. *3rd European Quality of Life Survey*, 2013.
14. I. H. Witten, E. Frank, and M. A. Hall. *Data mining: practical machine learning tools and techniques.* Morgan Kaufmann, Burlington, MA, 2011.