Using artificial intelligence to overcome over-indebtedness and fight poverty

Mário Boto Ferreira[a]; Diego Costa Pinto[b], Márcia Maurer Herter[c], Jerônimo Soro[d], Leonardo Vanneschie[e], Mauro Castelli[f], Fernando Peres[g]

[a]Universidade de Lisboa, Faculdade de Psicologia, Portugal

[b]NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Portugal

[c]Universidade Europeia, UNIDCOM and Europeia Research Centre for Business and Law, Portugal

[d]Universidade de Lisboa, Faculdade de Psicologia, Portugal

[e]NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Portugal

[f]NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Portugal

[g]NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Portugal

This is the accepted author manuscript of the following article published by Elsevier:

Boto Ferreira, M., Costa Pinto, D., Maurer Herter, M., Soro, J., Vanneschi, L., Castelli, M., & Peres, F. (2020). Using artificial intelligence to overcome over-indebtedness and fight poverty. [Advanced online publication on 19 October 2020]. Journal of Business Research, 1-15. https://doi.org/10.1016/j.jbusres.2020.10.035

Using Artificial Intelligence to Overcome Over-Indebtedness and Fight Poverty

ABSTRACT

This research examines how artificial intelligence may contribute to better understanding and to overcome over-indebtedness in contexts of high poverty risk. This research uses Automated Machine Learning (AutoML) in a field database of 1,654 over-indebted households to identify distinguishable clusters and to predict its risk factors. First, unsupervised machine learning using Self-Organizing Maps generated three over-indebtedness clusters: low-income (31.27%), low credit control (37.40%), and crisis-affected households (31.33%). Second, supervised machine learning with exhaustive grid search hyperparameters (32,730 predictive models) suggests that Nu-Support Vector Machine had the best accuracy in predicting families' over-indebtedness risk factors (89.5%). By proposing an AutoML approach on over-indebtedness, our research adds both theoretically and methodologically to current models of scarcity with important practical implications for business research and society. Our findings also contribute to novel ways to identify and characterize poverty risk in earlier stages, allowing customized interventions for different profiles of over-indebtedness.

Keywords: over-indebtedness, poverty risk, economic austerity, credit control, artificial intelligence, automated machine learning.

Using Artificial Intelligence to Overcome Over-Indebtedness and Fight Poverty

1. Introduction

According to the United Nations (UN), more than 780 million people (about 11% of the world's population) live below the international poverty line (United Nations, 2019a). While this percentage represents a significant decline in extreme poverty from past decades, the number of poor families globally remains unacceptably high (Njuguna & McSharry, 2017). It is thus not surprising that for the United Nations 2030 Agenda for Sustainable Development, ending "poverty in all its forms everywhere by 2030" is the first global challenge in terms of sustainable development goals (United Nations, 2019b).

Most of the research and public policy about poverty highlights emerging and least developed countries (e.g., Africa, Latin America; for a review see Njuguna & McSharry, 2017). Less attention has been given to localized deprivation of extreme poverty within developed countries (Shaefer & Edin, 2013) and to over-indebtedness as a major factor of poverty due to severe economic austerity in these countries. Such factors can considerably foster clusters of scarcity in more fragile developed countries, sharing features comparable to those of underdeveloped countries.

The current research focuses on over-indebtedness (i.e., the recurrent incapability to repaying credits when they are due) and its risk factors among Portuguese households in the aftermath of the European sovereign debt crisis. As a result of this crisis, Portuguese society was besieged by severe austerity due to so-called collective overspending (e.g., Panico & Purificato, 2013). According to the Organization for Economic Co-operation and Development (OECD), by 2014, Portugal was characterized by a poverty rate significantly higher than the European

average (Arnold & Rodrigues, 2015) with more than 2.6 million people living at risk of poverty (Statistics Portugal, 2017). Economy recovery has since then taken place but provisional data from 2019 still shows that 17.2% of the population (2.2 million people) was at risk of poverty (Statistics Portugal, 2019).

Different theoretical accounts of consumers' over-indebtedness vary on the emphasis they put on situational (e.g., European sovereign debt crisis) versus individualistic risk factors (e.g., careless overspending) (Angel, Einbock, & Heitzmann, 2009; Berthoud & Kempson, 1992; Kamleitner & Kirchler, 2007; van Staveren, 2002). Research aimed at testing such accounts have indeed linked many of these risk factors to over-indebtedness. However, most studies have provided evidence for the causal role of each of these factors "ceteris paribus" (i.e., assuming that all the remaining factors are held constant); however, actual cases of over-indebted households are likely to be multifactorial. Remarkably, how different risk factors combine in producing concrete situations of over-indebtedness is a highly important issue to the goal of ending poverty that has received less research attention.

In sum, our approach asserts that over-indebtedness is a multifaceted concept that does not speak with one voice. Rather, it embraces different kinds of over-indebted consumers, each one presenting a different profile and hence a different configuration of risk factors. The independent contribution to the over-indebtedness of each of these factors has already been established by previous business and psychology research as we review in more detail later. However, they are usually treated independently from each other rather than in interaction with each other.

Here, we address and evaluate the existence of different profiles of over-indebtedness by looking for distinguishable clusters or combinations of characteristics (based on household biographical and financial data) that would allow us to categorize over-indebted consumers in an exhaustive and mutually exclusive way. For this, we investigate a large field data set of over-indebted households who contacted the debt advisory services of the Portuguese Association for the Consumer Defense (DECO Portugal). Given the extension and complexity of the data set, we used artificial intelligence (AI) to look for distinguishable clusters or combinations of features that would allow us to classify and predict consumers' over-indebtedness.

Artificial intelligence is a term commonly used to describe the ability bestowed on digital computers, or computer-controlled systems, to accomplish tasks like intelligent beings (Nilsson, 2014). We resorted to a particular sub-area of AI, called Machine Learning (ML) (Marsland, 2015) to autonomously extract patterns from over-indebtedness data. Specifically, we employed Automated Machine Learning (AutoML; Feurer et al., 2015), which enabled us the evaluate thousands of models generated by state-of-the-art algorithms with multiple combinations of parametrization, and different types of feature selection methods. AutoML outperforms traditional ML approaches, allowing for a more substantial contribution of the present work to business research.

We begin by briefly reviewing prior business and psychological accounts of consumers' use of credit and discuss several risk factors of over-indebtedness. Next, we present a systematic review of the literature on artificial intelligence and machine learning to show the relatively scant use of these important methods in the realm of business research. The current machine learning approach specifically aimed at describing and predicting different profiles of over-indebtedness is then presented and empirically tested. Our ultimate goal is to contribute to the development of

predictive models that can help practitioners and public policy makers to make better interventions regarding economic decisions and contribute to reducing poverty risk at earlier stages.

2. Credit use and over-indebtedness

Economic theories of consumption such as the life cycle hypothesis purpose that people take on debt based on expected future income when they are young and then save during middle age to maintain consumption level later in life (Modigliani, 1966). In practice, many consumers seem to deviate from these theoretical predictions when it comes to borrowing and saving. Indeed, along the 20th century, consumers have been gradually more open to the idea of using credit as a way of obtaining liquidity that their pay-checks would not otherwise permit (Watkins, 2000) as a way to promote their financial well-being (Brüggen et al., 2017).

Together with more extensive use of credit came a shift in how consumers react to debt. The idea of being in debt has become progressively less dreaded and more normalized. Nowadays, it is often perceived as an inherent condition shared by many of us in the process of obtaining necessary goods and services, such as a place to live or getting a college degree (Celsi et al., 2017; Merskin, 1998). Following this tendency, most Western societies reported over the last few decades an increase in consumer credit use and household debt levels (e.g., Betti, Dourmashkin, Rossi, & Yin, 2007; Brown, Garino, Taylor, & Price, 2005; Kida, 2009; Pattarin & Cosma, 2012).

Debt turns out to be troublesome to individuals and governments alike when it reaches such high levels that households become over-indebted. Over-indebtedness may be defined as recurrent incapability to repaying credits when they are due, sometimes self-reported, sometimes

calculated from economic information from the household (e.g., a high debt-to-income ratio). As such over-indebtedness is related to the more overarching concept of scarcity, usually defined as a condition of having insufficient resources to cope with financial demands (Zhao & Tomm, 2018).

Being over-indebted has a profound impact in households, leading to increased risk of deprivation and poverty and consequent social stigma, deteriorated health and well-being, relationship difficulties and breakdown, financial exclusion, and reduced labor market activity (Alleweldt et al., 2013). Although some studies have found that households with lower incomes are more prone to have difficulties repaying their debts (Aizcorbe, Kennickel & Moore, 2003; Sullivan & Fisher, 1988); other studies found no evidence of such relation (Betti et al., 2007; Canner & Luckett, 1991), which suggests that while low income may increase households' risk of over-indebtedness, being over-indebted is not a situation exclusive to low socioeconomic status households.

2.1. Risk factors of over-indebtedness

Over-indebtedness has been related to several possible causes or risk factors. One is poor financial literacy due to a lack of appropriate formal education. Consumers' lack of knowledge concerning financial products and concepts make households vulnerable to debt repayment difficulties. However, although several studies have confirmed the association between innumeracy, financial illiteracy, and households' poor financial decision-making (Lusardi & Mitchell, 2011; Lusardi & Tufano, 2015), there is mixed evidence on the effectiveness of

financial education programs in avoiding decisions potentially leading to over-indebtedness (see Dellande et al., 2016; Lusardi, 2008).

Effects of low financial literacy in accumulation and repayment of debts are likely to be aggravated by consumers' proneness to rely on heuristics when making decisions, which make them the prey of several reasoning biases (e.g., Thaler & Sustein, 2008). For instance, the asymmetrical perception consumers display between present gains and future losses encourages the increased use of credit and disregard for the accumulation of interests (Slowik, 2012). In the same vein, individuals displaying present-bias preferences (i.e., desire for immediate consumption) show more tendency of having credit-card debt and higher debts in credit-cards (Meier & Sprenger, 2010; Strömbäck, Lind, Skagerlund, Västfjäll, & Tinghög, 2017).

These (and other) reasoning biases are often explored by financial institutions that tailor their communication and product advertisement to turn consumer's heuristic-based judgments into their favor, sometimes leading consumers to bad financial decisions regarding debt accumulation and repayment (Bar-Gill & Warren, 2008). However, reliance on heuristic-based judgments seems to be dependent not only on contextual factors but also on individual differences in rational behavior. Specifically, on the ability to second guess, analyze and override appealing but biased outputs, replacing them with more accurate decisions (Stanovich, 2009; Stanovich & West; 2008; Stanovich, West, & Toplak, 2011; Toplak, West, & Stanovich, 2017).

Interventions based on nudging and disclosure of relevant information to individual consumers have been successfully developed by behavioral economics (e.g., Hertwig & Grüne-Yanoff, 2017; Loibl, Jones, & Haisley, 2018; Thaler & Sustein, 2008) as means to promote better financial decisions. However, even carefully designed messages may have only a small

impact in counteracting the negative effects of reasoning biases on consumers' behavior (e.g., Bertrand & Morse, 2009).

Furthermore, consumers' self-control and the need to resist impulsive consumption (e.g., using credit cards), likely depends on the availability of limited and easily depletable cognitive resources (Baumeister, Vohs, & Tice, 2007; Inzlicht & Schmeichel, 2012). Limited self-control may work both as a cause and as a consequence of over-indebtedness. On one hand, there is substantial evidence showing individual differences in self-control (e.g., Eigsti et al., 2006; Mischel, 1958) suggesting that some people may be more vulnerable to a social environment that encourages impulsive consumption than others; on the other hand, previous research also shows that over-indebted households face a spiral of difficult decisions (that other households typically do not face) that result from small budgets requiring the meticulous calculation of expenses and juggling of sporadic incomes. This state of affairs progressively depletes their self-control capacity (e.g., Mani, Mullainathan, Shafir, & Zhao, 2013; Vohs & Heatherton, 2000; Zhao & Tomm, 2018). Indeed, the demands caused by over-indebtedness in particular and scarcity in general, tend to hijack the cognitive system depleting cognitive resources, such as attention, working memory, and executive control (Bertrand, Mullainathan, & Shafir, 2004; Mullainathan & Shafir, 2013). Regardless, the dispositional lack of self-control or its subsequent depletion by circumstances of severe austerity, impairs consumers' cognitive capacity shifting decision behavior away from reasoned options towards more intuitive and impulsive choices (Vohs & Faber, 2007).

Over-indebtedness has also been related to other more situational risk factors, such as adverse local circumstances or significant life events. Younger consumers and more numerous households (especially with more children) are associated with debt repayment difficulty

(Canner & Luckett, 1991; Godwin, 1999), as well as households with divorced/separated people (Canner & Luckett, 1991). Adverse life events are reported frequently as a reason for late payments (Canner & Luckett, 1991) and the presence of adverse life events in the last 12 months are associated with households with debt repayment strain in comparison to a control group (Tokunaga, 1993).

Abrupt changes in socio-economic conditions can launch (mostly middle-class) households into financial strains and increased risk of indebtedness. The European sovereign debt crisis that followed the 2008-2009 World economic recession is a case in point. After the bailout of the Portuguese debt in 2010, several austerity measures ensued. There was a steep increase in taxes for employees and businesses and substantial cuts in the monthly income of public workers and retirement pensions. Unemployment soared and social benefits were cut. Such measures put together led to a dramatic increase in the financial vulnerability of the Portuguese households (similar scenarios unfolded in Greece, Ireland, and Spain). By the end of 2014, in a population of about 10 million, 2.6 million were over-indebted (i.e., with a debt-to-income rate of more than 35%) and 700.000 had entered in default (Bank of Portugal, 2014; Statistics Portugal, 2017).

In the last few years, particularly since 2016, the Portuguese economy has started a slow recovery with all major credit rating agencies moving Portugal's debt from junk territory to "stable" or "positive" outlook by 2018. Interestingly, despite the decline in unemployment and the progressive removal of cuts in monthly income, the household effort rate of the Portuguese families increased from 70.8% in 2017 to 73% in the first semester of 2018 (DECO, 2018). This once more suggests that over-indebtedness is a complex and multifaceted phenomenon that needs to be better understood.

2.2. Profiling and predicting over-indebtedness: a machine learning approach

Most of the aforementioned research on risk factors underlying over-indebtedness has been done in a top-down manner. Several risk factors (e.g., financial illiteracy, prevalence in the use of improper heuristics, lack of self-control, markers of economic austerity) are related to over-indebted households in some cases, and interventions based on these factors (e.g., financial education programs, nudging) have shown to be sometimes (but not always) successful in counteracting over-indebtedness. This indicates that the identified factors play an important role but are not always sufficient conditions for over-indebtedness.

Actual cases of over-indebtedness are likely to result from different combinations of risk factors. In this sense, we hypothesize that the notion of over-indebtedness in itself may be a misnomer because it puts under the same conceptual umbrella distinct types or profiles of indebted households. However, the degree with which the different combinations of factors underlying over-indebtedness carve different profiles of over-indebted households is an empirical question that we begin to answer in the present work.

In this paper, we suggest a bottom-up approach capable of a) exploring possible different profiles of over-indebted households, and b) identifying the main features of the profiles (if and when they emerge from the data) to develop predictive models of classification of new cases (households) under risk of over-indebtedness. Such a bottom-up approach is methodologically challenging but achievable using artificial intelligence to develop descriptive and predictive models concerning the risk factors of over-indebtedness of Portuguese consumers.

Before presenting our methodological approach and to contextualize the research here reported, we first assess prior work using artificial intelligence in business research.

3. Artificial intelligence in business research

We conducted a systematic literature review searching several online scientific databases (e.g., EBSCO, Elsevier Science Direct, Emerald, JSTOR, SCIELO, Scopus, and Tailor & Francis) to identify empirical and conceptual studies examining artificial intelligence or machine learning specifically in business research.

We used the following keywords in the search process, along with business research: *machine learning*, *artificial intelligence*, *support vector machines*, *automated machine learning*, and *AutoML*. After obtaining the initial set of articles, we applied a snowballing procedure for examining the references in these articles and find additional studies. The search process was completed in May 2020 (see Table 1 for details). We found a total of 11 articles in business research directly linked to artificial intelligence and/or machine learning. 72.7% were empirical articles and 27.3% were conceptual papers. The large majority of the articles (81.8%) analyzed single algorithms (or a single algorithm family) and only 18.2% of the studies employed multiple algorithms. Finally, none of the business research studies found in this search used automated machine learning (AutoML). This shows that the employment of Automated Machine Learning (AutoML) to characterize and predict over-indebtedness is still rather new.

Insert Table 1 about here

Probably, the most similar approach to the one here presented was advanced by Montiel et al. (2017). These authors used feature selection and supervised learning techniques, such as Logistic Regression and Random Forests to generate predictive models of over-indebtedness. However, several points distinguish the present work from Montiel et al. (2017). First, Support Vector Machines, the supervised machine learning technique used here is substantially different from the ones used in Montiel et al. (2017). In our work, we employed state-of-the-art algorithms using Automated Machine Learning (AutoML; Feurer et al., 2015), in which Support Vector Machines were able to outperform several alternative algorithms used in business research such as Logistic Regression and Random Forests (on the data at our disposal in a preliminary experimental phase). Second, descriptive modeling is absent in Montiel et al. (2017), while it is a fundamental part of the present work. Third, Montiel et al. (2017) used data from a banking institution relative to French individuals and households. This is a noticeable difference since all indicators suggest that the risk of over-indebtedness and poverty in Portugal is more serious than in France. In addition, our data is originated at a consumer protection institution, not a bank, which may highlight the societal relevance of our field dataset.

Other less directly related but important research include the use of Gradient Boosted Models by Agarwal et al. (2018) to connect financial outcomes and phone-based social behavior to predict financial wellbeing in the US, and the use of various data mining algorithms by Alomari (2017) for default prediction of peer-to-peer loans and learning associations between various attributes of loan applications. It is also worth mentioning the work by Eletter et al. (2010), where Artificial Neural Networks were used for evaluating credit applications to support loan decisions.

4. Methods

In the present work, we use Unsupervised ML algorithms such as Self Organizing Maps and Agglomerative Hierarchical Clustering to obtain descriptive models aimed at finding different over-indebtedness consumer clusters (corresponding to distinct profiles). Furthermore, we resort to Supervised ML algorithms such as Support Vector Machines, using Automated Machine Learning (AutoML) (Feurer et al., 2015) to obtain predictive models aimed at successfully identifying and classifying (in different profiles) new cases (households) under risk of over-indebtedness.

Our objective is to contribute to preventing over-indebtedness and reduce poverty by proposing artificial intelligence tools that are capable of a more fine-grained characterization of actual cases of over-indebted households, and that build on this characterization to better assess the risk of future cases of over-indebtedness.

This section is structured as follows. In section 4.1, we describe the employed over-indebtedness real-world field data. Section 4.2 presents the employed ML approach. Section 4.3 shows the system used to generate our descriptive model, aimed at producing a clustering of over-indebtedness categories (Self-Organizing Maps). Finally, section 4.4 describes the algorithm used to generate our over-indebtedness predictive model (Support Vector Machines).

4.1. Data Collection

In this work, we used data gathered from consumers under assistance for over-indebtedness of the Portuguese Association for the Consumer Defense (DECO Portugal). We analyzed the data of the population of 1,654 consumers nationwide who contacted the debt

advisory services in Portugal during the years of 2016 and 2017. In particular, a total of 802 consumers contacted the debt advisory services in 2016 and 852 consumers in 2017. When consumers contact the debt advisory services, they are over-indebted and cannot pay their bills anymore, having a high risk of poverty. These consumers ask for help on how to organize their family budget, how to consolidate their debts among the credit holders (e.g., bank, insurance companies, stores), or which credits should they pay first. In extreme cases, the debt advisory services can suggest which goods they should give up, from simple consumption goods (e.g., mobile phone, computer) to important long-term goods, such as cars and their houses.

The dataset comprises a broad range of variables to understand the full picture of consumers' financial health: family socio-demographics, total income, total expenses, employment information, as well as all credit details. The features considered for the analyses were: socio-demographic characterization (marital status, level of education completed, number of people in the household), the perceived causes for over-indebtedness (from a predetermined pool of causes), and data concerning their economic situation, including the total income and expenses of the household as well as data concerning their credits and debts (amount of the monthly installments for credit cards, housing credit, car credit, personal credit and other types of credit or debts; total monthly installment concerning all credits). Each household is represented by one record (one observation) of the dataset with many features to describe their characteristics and behavior. Appendix A summarizes the main variables analyzed in this study.

4.2. Machine Learning (ML) Approach

In this paper, we present a study aimed at distinguishing and describing the cluster profiles of over-indebtedness of Portuguese citizens, based on their main characteristics and

behavior. This research also intends to create predictive models to classify the consumers into the appropriate over-indebtedness profiles. Based on these research objectives, this research approach combined unsupervised and supervised Machine Learning (ML) techniques, to jointly analyze descriptive and predictive models of over-indebtedness (Mitchie et al., 1994).

To achieve this task, we employed Automated Machine Learning (AutoML) (Feurer et al., 2015), a technique to automate the ML process, enabling the evaluation of thousands of models generated by many classifiers (ML algorithms) with multiple combinations of parametrization, and different types of feature selection methods. The methodological approach was divided into 4 phases: (1) data preparation, (2) data selection, (3) unsupervised ML, and (4) supervised ML.

In the data preparation phase, data preprocessing activities were executed to prepare and normalize the data of 2016 and 2017, and to generate new features to extract information hidden in the existing features (feature engineering). With the prepared data, the second phase (data selection) performed the feature selection to eliminate data redundancy. For every set of features, it was applied an analysis of extreme outliers to treat the noise (or *errors*) in the data — each set receiving its specific outlier treatment instead of performing the same action for the entire dataset. Consequently, this procedure of outlier treatment does not affect the entire dataset and there is no risk of losing important information for other features. The presence of extreme outliers was detected using univariate and multivariate analysis in all numeric variables. In univariate analysis, these variables had few observations with values two times higher of the upper limit (one of the criteria used to filter extreme outliers). The multivariate analysis of extreme outliers further supported most of the extreme outliers selected by univariate analysis. As an outcome, the extreme outliers removed represent 5.25% (87 observations). Therefore,

from a total of 1,654 observations, 1,567 were used to generate and test the models. To remove the potential bias associated with the different order of magnitude of the values of the input features, we performed a normalization process. In this way, all the numerical features range in the interval [0;1]. The normalization only used information calculated on the training set. Thus, the minimum and the maximum of each feature were calculated only on the training samples. One-hot-encoding was applied to the categorical features. The process for obtaining the one-hot encoding of a categorical variable first requires that the categorical values are mapped into integer values. Subsequently, each integer value is represented as a binary vector that contains all zero values, except the index of the integer which contains a one. This transformation is necessary, when there is no ordinal relationship between the categories, to remove any bias associated with the integer representation of the categories.

Concerning unsupervised ML, we have used clustering analysis to describe the data through a set of distinguishable clusters, grouping consumers with common characteristics and consumption behavior. For cluster profiling analysis, the approach of this study also designed and developed an algorithm to automate the cluster profiling generation, filtering the main statistical differences among the clusters and highlighting the specific characteristics of each cluster. After clustering analysis, the supervised ML phase evaluated several estimators (generated by AutoML design) and selected the best method ("winner" model) to classify the profile of consumers based on their characteristics and behavior. The final method was chosen by model performance metrics that were selected carefully following the nature of the dataset and the target feature (profile-clustering). The following sections detail the procedures for unsupervised and supervised ML in this study. For both unsupervised and supervised ML, we do not consider Deep Learning. While the use of Deep Learning was proposed in recent literature on

ML for social good (Al-Hashedi, Soon, & Goh, 2019; Khatua, Cambria, & Khatua, 2018; Sawhney et al., 2018), the problem considered in this paper is characterized by a limited amount of data. Therefore, considering that Deep Learning provides accurate models when a vast amount of training data is available, we do not rely on this ML technique.

4.3. Unsupervised ML: Self-Organizing Maps

To identify and describe the consumers' profiles groups of over-indebtedness, we employed Self-Organizing Maps (SOM). SOM is an unsupervised learning computational method, belonging to the field of artificial neural networks (Haykin, 1994). This technique was recently employed to address several complex tasks over different domains and produced competitive results. For instance, it was used in conjunction with Deep Learning and Principal Component Analysis to successfully solve a human sentiment classification problem (Ali et al., 2019); it was used in the field of electricity consumption and power systems (Ghadiri, & Mazlumi, 2020); and it was used for heartbeat analysis (Lee, Song, & Lee, 2020). Thus, the flexibility of SOMs makes it a suitable technique in the context of our study.

SOMs are commonly used as a clustering and visualization technique in exploratory data analysis. For instance, they can be used to group sets of data observations according to their mutual similarities (clustering). Given a particular distance metric, typically groups are formed to maximize the intra-cluster distances and minimize inter-cluster distances. A trained SOM, that can make such a categorization of data, is normally referred to as a descriptive model of the data. Another typical use of a SOM is to transform input data of arbitrary dimension into a lower (typically two) dimensional discrete map, and to perform this transformation adaptively, preserving the topological features of the original input space.

Among the different SOM variants, we considered in this work the Kohonen Network (Kohonen, 2013). This SOM has a feed-forward structure, where neurons are set along an *n*-dimensional grid: typical applications assume a 2-dimensions rectangular grid (e.g., 10×10). Each neuron is fully connected to all the source nodes in the input layer, and the connection weights are initialized with small random values, or with appropriate input values. This single-layer neural network represents a distribution of input data items using a finite set of models. These models are automatically associated with the nodes of the grid, so similar models become automatically associated with nodes that are adjacent in the grid, whereas less similar models are situated farther away from each other in the grid (Kohonen, 2013). In this way, the grid gradually becomes a 2-dimensional transformation of the input space, preserving the topology of the input data.

Training a SOM requires several iterative steps. For a generic input pattern (or data observation) **x**, the following steps must be executed (Resta, 2012): (1) evaluate the distance between **x** and the vector of weights of the synaptic connections entering in each neuron. For instance, the Euclidean distance between the input vector **x** and the weight vector can be considered; (2) select the neuron (node) with the smallest distance to **x** (i.e., "winner neuron" or Best Matching Unit – BMU); (3) correct the position (i.e., by modifying the weights) of each node according to the results of Step (2), to preserve the network topology. This iterative process continues until a stopping criterion is reached. Typically, the stopping criterion considers a weighted average over the Euclidean norms of the difference between the input vector and the corresponding best matching unit.

Once the training procedure is concluded, the result consists of a descriptive model that considers how the input space is structured and projects it into a lower-dimensional space, where

closer nodes represent neighboring input patterns. Thus, a SOM is particularly suitable for visualizing hidden patterns from the multi-dimensional input data. The Elbow method and Silhouette were used as cluster definition techniques to determine the optimal number of three clusters (see Appendix B for details).

The grid size defined for this study was 100 cells (dimension x = 10 and dimension y = 10) presenting good results, with a good distribution of observation in the nodes, and it did not generate any empty node (nodes without any observations). The *topo* is a parameter to define the way nodes are arranged in the grid (100 nodes, dimension x = 10, and dimension y = 10). The nodes of the grid can be arranged as rectangular or hexagonal, it defines the number of immediate neighbors, rectangular shapes have 4 immediate neighbors, and hexagonal shapes have 6 immediate neighbors. The *alpha* parameter defines the learning rate, defining the amount of change in each interaction. The default value is to decline linearly from 0.05 to 0.01 over each iteration. The *whatmap* defines which data layers were used, and the configuration used the final feature selection: categorical variables: (1) cause classification (crisis and other causes not related to the crisis); and numerical variables: (2) income per capita, (3) total expenses, (4) effort rate with credit card, (5) effort rate with housing credit, (6) effort rate with car credit, (7) effort rate with personal credit, and (8) effort rate other types of credit or debts.

The *dist.fcts* parameter is a vector of distance functions to be used to calculate the distances among nodes: Tanimotodistance (for categorical data/factors) (Lipkus, 1999) and Euclidean distance (for numeric features) (Gower & Legendre, 1986). The *keep.data* parameter defines the return of a trained map. If *keep.data* is true, it returns original data and mapping information; if false, it only returns the mapping information (trained map).

4.4. Supervised ML: Support Vector Machines

Support Vector Machines (Cortes & Vapnik, 1995) are supervised ML techniques that can be used for addressing classification and regression tasks. The objective of Support Vector Machines is to establish the equation of a hyperplane that divides the space, leaving all the points of the same class on the same side, and separating points belonging to different classes.

Among the possible hyperplanes, a Support Vector Machine selects by construction the one that maximizes the distance (margin) of the hyperplane from the closest data points of each class (support vectors). This hyperplane is usually called maximum separation hyperplane, and it is usually addressed as a predictive model. Once a Support Vector Machine is trained (i.e., the maximum separation hyperplane has been achieved), the prediction of new unlabeled information can be performed. New observations will be categorized as belonging to the same class as the points that stand on the same side of the maximum separation hyperplane. This results in a robust classifier that maximizes the probability of classifying a new data point in the correct class, thus ensuring an appropriate generalization ability. When the points are not linearly separable, Support Vector Machines transform (through a function called *kernel*) the original space of data, to map into a new higher dimensional space, where the data points are linearly separable. Then, the maximum separation hyperplane can be achieved in this new mapped space. Support Vector Machines can be used to address both classification and regression problems. In the first case, it is common to refer to them as Support Vector for Classification (SVC). Thus, in the continuation of the paper, we will refer to Support Vector Machines as SVC. For a full understanding of the properties of Support Vector Machines and the definition of kernel functions, the interested reader is referred to Schölkopf et al. (2002).

To define the design of the grid search approach, the characteristics of the dataset and the target feature (cluster profile) were considered. These characteristics also were considered to select the ML algorithms to be tested in the grid search, to classify the profiles of over-indebtedness. The exhaustive search of the grid search hyperparameters tuning can be done for several algorithm types, exploiting many approaches to develop the best model possible to classify the profiles of over-indebtedness. Consequently, a wide range of classifiers (machine learning algorithms) was used: (1) Nu-support Vector Machine, (2) Support Vector Machine, (3) Gradient Boosting, (4) Extra Trees, (5) Random Forest, (6) Decision Trees, (7) Gaussian Naive Bayes, (8) K Nearest Neighbors, (9) Linear Discriminant Analysis, and (10) Logistic Regression. Each classifier has a different set of parameters following the algorithm design (development) and objectives of optimization. Figure 1 shows the algorithms and the parameters tested by the Grid Search for each ML method.

Insert Figure 1 about here

This research uses a three-step approach designed to use Grid Search Hyperparameters Tuning (see Figure 2 for details). The hyperparameter tuning phase is the first step of this approach. This step consists of a set of experiments to test the range of possible values or existing options of a parameter to find good configurations. Therefore, 32,730 intermediate models were tested for these hyperparameters. As an outcome, we obtained a range of parameters for each ML algorithm and this generated 6,546 candidate models.

Insert Figure 2 about here

The second step (training performance) aims to select the most appropriate configuration of each algorithm (e.g. Nu-Support Vector Machine, Support Vector Machine, Gradient Boosting, etc.). So far, only the training set has been used to assess the different models (using 5-folds cross-validation). To evaluate the performance of the models, two main metrics were used: accuracy score and logistic loss (log loss).

- **Accuracy Score:** the accuracy score is obtained through the division of total correct predictions overall observations, assuming values between 0 and 1. The goal of each predictor is to maximize the accuracy score (approximately 1).
- **Logistic Loss:** the logistic loss (or log loss) provides a nuanced view of the models' performance, considering the uncertainty of a prediction and calculating how many times it varies from the actual label. If the predicted probability diverges from the actual label, the log loss increases, assuming values between 0 and infinite. The objective of the classifiers is to minimize the log loss value, consequently, a perfect model would have a log loss equal to 0.

The third step is the model selection and generalization ability. After the best model for each algorithm is found — with its most appropriate hyperparameter combination — the test set is used to assess these models and the best one is selected (winning model).

5. Findings

5.1. Self-Organizing Maps and Over-Indebtedness Clustering

An Automated ML algorithm was developed to create and select the best descriptive model and generate cluster profiling automatically. The automated profiling description highlighted the distinguishable characteristics of each cluster, showing the values of similar clusters and ranking the variables by statistical tests for numeric and categorical variables. Several descriptive models for Self-Organizing Maps were generated with different parameters and sets of features, comparing the cluster results performance, number of clusters, and profiling. The final selection was based on the analysis and capacity of cluster description (descriptive ability) in accordance with over-indebtedness analysis considerations. The final descriptive model used the Kohonen R Package (R Studio), using the method *supersom* (Supervised SOM) with the following parameter configuration: *rlen* = 3,000 iterations; *alpha* = 0.05; *topo* = hexagonal; and *grid size* = 100 cells (10 x 10). After 3,000 iterations, the mean distance between the observations of each node was reduced to 0.015. Figure 3 shows the progress of SOM training and the decrease of the mean distance to the closest unit distance over time (iterations).

Insert Figure 3 about here

However, it is important to note that some variables did not achieve statistical significance in the cluster profiling analysis. Indeed, the differences among groups are not statically significant for education level ($\chi^2_{(4\ ,\ 1455)} = 0.9608, p = 0.9157$) nor years of education ($F_{(2\ ,\ 1564)} = 0.5813, p = 0.5593, \eta_p^2 = 7e\text{-}04$). The total income is also not statistically significant to distinguish the groups ($F_{(2\ ,\ 1564)} = 0.9568, p = 0.3844, \eta_p^2 = 0.0012$), only income per capita is statistically significant ($F_{(2\ ,\ 1564)} = 162.6146, p < 0.001 \eta_p^2 = 0.1721$).

5.2. Over-Indebtedness Cluster Profiling and Description

As final outcome, SOM training extracted 3 clusters with distinguishable characteristics: low-income households (n = 490, 31.27%), low credit control households (n = 586, 37.40%), crisis-affected households (n = 491, 31.33%).

Cluster 1 – Low-income households: In this cluster, 100% of consumers have over-indebtedness problems due to causes not related to the crisis. Over-indebtedness stems in this group from low-income levels as the cluster includes medium-sized families (*M* = 2.65 people) with the lowest income per capita (401.94 euros per month, Z-score mean = -0.34). Furthermore, the consumers of this group have the lowest total credit monthly installment (453.65 euros per month, effort rate = 40%, Z-score mean = -0.46), the lowest credit card monthly effort rate (149.54 euros per month, effort rate = 12%), and the lowest housing credit monthly installment (M = 80.21 euros per month, effort rate = 6%) of the three clusters. This group presents the lowest level of unemployment (6.6%), which is 12.6% below the dataset mean, and is mostly employed in the private sector (51.3% of the consumers, 7% above the dataset mean). One of the main issues reported as a cause of the financial difficulty is the increase in family members (12.8% of the households).

Cluster 2 – Low credit control households: This cluster includes cases of over-indebtedness predominantly due to other causes (83.96% of the observations) and a few crisis-related cases (16.04% of the observations). Households have the highest income per capita and the smallest mean number of people in the household. Notably, there are several indications of low credit control when compared to other groups. Although these households have the highest income per capita (686.35 euros per month, Z-score mean = 0.54) and the lowest number of

people in the household (M=1.78, Z-score mean = -0.48), they present the highest credit effort rate (M=75%, Z-score mean=0.27) and personal credit rate (246.00 euros per month, effort rate = 28%). On the other hand, these consumers have the lowest car credit effort rate (19.88 euros per month, effort rate = 2%) and the lowest household expenses (570 euros per month).

Cluster 3 – Crisis-affected households: Cluster 3 presents cases of over-indebtedness that are mostly due to the crisis (83.7% of people) and a few pertaining to other causes not related to the crisis (16.3% of people). This cluster is characterized by low income per capita (413.15 euros per month, Z-score mean = -0.3) and includes the largest families (2.76 people in the household) and the highest household expenses (790.69 euros per month) of the three clusters. The main causes for over-indebtedness are unemployment (40.5%), which is, 21.3% higher than the dataset mean; salary cuts (12.2%), 6% higher than the dataset mean; and spouse's unemployment (8.4%), which is 4% higher than dataset mean. These consumers have the highest provision with housing (209.63 euros per month, effort rate = 20%, Z-score mean= 0.27) and with other credits or debts (79.54 euros per month, effort rate = 10%, Z-score mean 0.33). Table 2 provides the cluster profiling automated feature selection.

Insert Table 2about here

5.3. Support Vector Machine and Over-Indebtedness Prediction

The supervised ML phase used Automated ML supervised Grid Search Hyperparameters Tuning in Python to search the most appropriate experimental setting. In this study, an exhaustive search using Grid Search Hyperparameters Tuning generated 32,730 intermediate

models during the cross-validation stratified 5-fold, creating 6,546 candidate models in total. To define the design of the grid search approach, the characteristics of the dataset and the target feature were considered (cluster profile). These characteristics also were considered to select the ML algorithms to be tested in the grid search, to classify the profiles of over-indebtedness.

The exhaustive search of the Grid Search Hyperparameters Tuning can be done for several algorithm types, exploiting many approaches to develop the best model possible to classify the profiles of over-indebtedness. Consequently, a wide range of classifiers was used: (1) Nu-support Vector Machine, (2) Support Vector Machine, (3) Gradient Boosting, (4) Extra Trees, (5) Random Forest, (6) Decision Trees, (7) Gaussian Naive Bayes, (8) K Nearest Neighbors, (9) Linear Discriminant Analysis, and (10) Logistic Regression.

When searching for the best estimator, the objective is to split the data into two sets (training set vs. test set) and use them as input for the algorithm to learn the parameters that best describe the patterns underlying the data and afterward assess their performance. The hyper-parameters are the algorithm's parameters that are not directly learned within estimators and need to be defined prior to training.

For the present study, an exhaustive cross-validation search was implemented in order to reach the best possible hyperparameter combination for each algorithm. First, the data was split into training and test sets. Then, for cross-validation, the training set was further divided into 5 partitions for the study. These 5 folds of data are used in the same way as training (80%) and test set (20%), in the sense that 4 folds are combined as input to learning the data and one-fold is used to evaluate the quality of the resulting model. The objective is to compare the performance of each hyperparameter. Thus, the process is run several times, each with a different combination of hyperparameter values. Once the best arrangement is found, the algorithm is trained with the

elected hyperparameters on all the 5 folds — and the learning phase is repeated on the entire training set.

The performance of each winner classifier is presented in Figures 4a and 4b. During this analysis, 4 models were removed from the selection, because their performance was much lower than the one of the other algorithms. The removed methods were: (6) Decision Trees, (7) Gaussian Naive Bayes (9) Linear Discriminant Analysis, and (10) Logistic Regression. In particular, the poor performance of Linear Discriminant Analysis and Logistic regression seems to suggest that the problem under analysis is particularly complex. As a consequence, the aforementioned techniques cannot provide good-quality models because they cannot solve non-linear problems since their decision boundary is linear. Focusing on Gaussian Naïve Bayes, its poor performance is mainly due to the "naïve" assumption made by the algorithm: it assumes conditional independence between every pair of features given the value of the class variable. Finally, the poor performance of Decision Trees could be motivated by the fact that they are unstable, with a small change in the data leading to a large change in the structure of the optimal decision tree, making it difficult for the Decision Trees algorithm to learn and express these features.

Insert Figures 4a and 4b about here

After the performance analysis on the training set, the six algorithms that returned the best results in their categories were: SVC, Nu-SVC, Extra Trees, Random Forest, Gradient Boosting, and K Nearest Neighbors. For the models generated by these six algorithms, their

generalization ability was assessed by evaluating their performance on the test set (unseen data). Figure 5 shows a comparison of their performance, using Accuracy Score and Log Loss.

Insert Figure 5 about here

The algorithm that generated the best model after this exhaustive search was a version of the Support Vector Machine algorithm, the Nu-SVC. The Nu-SVC is similar to SVC but uses a regularization parameter to control the number of support vectors implementing a penalty on the misclassifications that are performed while separating the classes. As the Nu-SVC was the winning classifier, in the next section it will be described the functioning of this algorithm. The best-fitting model of Nu-SVC generated automatically had the following parameters (see Table 3 for details): Nu (0.08), Kernel (RBF), Gamma (Scale), Decision Function Shape (ovr), and Class Weight (balanced). Table 4 presents detailed Machine Learning algorithms comparative performance.

Insert Tables 3 and 4 about here

---

6. Discussion

The first global challenge in terms of sustainable development goals elected by the United Nations 2030 Agenda is ending poverty in all its forms and everywhere by 2030 (United Nations, 2019b). Over-indebtedness is a major factor of poverty. Research in business and

psychological sciences has related the risk of over-indebtedness and scarcity to several different factors. Financial illiteracy (Lusardi & Mitchell, 2011), the prevalence of intuitive judgments and decisions based on heuristics (Thaler & Sustein, 2008), impulsiveness, and lack of self-control (Mani et al., 2013) are dominant theoretical accounts typically used to explain how households become over-indebted.

Furthermore, socio-economical abrupt changes often prompted by financial crisis such as the 2008 Global Financial Crisis and the European sovereign debts crisis and the ongoing economic shock caused by the COVID-19 pandemic are known to increase consumers financial vulnerability and ability to pay their debts when due (e.g., Alleweldt et al., 2013; Canner, Luckett, Cook, & Middleton, 1991; Tokunaga, 1993).

However, none of the aforementioned factors determines by itself over-indebtedness. Actual cases of over-indebted households result from different combinations of factors and these combinations are likely to form meaningful clusters or profiles. In other words, over-indebtedness is not a unified concept but may come in many forms.

To empirically support the claim that over-indebtedness does not speak with one voice this research used a bottom-up approach based on Automated Machine Learning (AutoML) to develop descriptive and predictive models of distinguishable profiles of over-indebtedness (based on specific configurations of households features).

6.1. Theoretical and Methodological Contributions

Our findings suggest that the consumers' socio-economical features do not vary randomly but clustered together in three emerging profiles. Economic crises are often pointed out

in the public arena as being among the main situational causes of over-indebtedness. However, in the aftermath of the Portuguese financial and socio-economic crunch, our findings indicate that over-indebtedness is associated mainly with other situational causes for both low-income households and low credit control households, with only one profile of over-indebtedness (accounting for less than one-third of the cases) directly related to the economic crisis.

In light of this profile classification, it seems reasonable to conclude that although the social-economic crisis that besieged Portugal certainly increased the financial vulnerability of households, it can hardly be considered the immediate cause of all or even most cases of over-indebtedness. Other situational causes not directly related to the crisis characterize the majority of over-indebted families.

Furthermore, the emergence of the low income and low credit control profiles suggest that lack of self-regulation may be more a consequence of the emotional strain and cognitive overload that progressively deplete self-control capacity (e.g., Mani et al., 2013) in the first of these profiles; whereas dispositional low levels of self-control are more likely to be a cause (or important risk factor) of over-indebtedness (e.g., Eigsti et al., 2006) for the latter profile.

In the same vein, although heuristic-based judgment may contribute to decision biases across all profiles, in the case of the low credit control profile, failures to second guess intuitive (but biased) responses and replace them with more deliberate decisions are more likely to work as a predecessor of over-indebtedness due to individual differences in rational behavior (e.g., Stanovich, 2009). For low-income and crisis-affected families, the same failures are more likely to begin as a consequence of the depletion of cognitive resources associated with over-indebtedness and then contribute to accentuate a spiral of biased decisions. By empirically distinguishing various profiles, the bottom-up approach we adopted shows potential for

explaining in a coherent way how different psychological mechanisms may interact with situational risk factors to carve specific types of over-indebtedness.

Methodologically, our findings contribute to business research presenting an AutoML perspective for social good. AutoML automates the configuration and selection of a complex machine learning model of over-indebtedness and fosters the generation of robust models – that is, models that are resistant to data variations and able to provide a more accurate data generalization (compared to single ML algorithms testing), reducing possible errors and biases that may occur using a human-based design of specific machine learning model.

Although recent studies have explored analytical approaches in business research (e.g., Delen & Zolbanin, 2018), most previous research did not analyze these phenomena using AutoML in a more comprehensive way. In one of the first analytic attempts in business research, Bejou et al. (1996) explored Artificial Neural Networks (ANNs) on customer relationship management. Fish et al. (2004), and Hamid and Iqbal (2004) also explored ANNs to model brand market share and forecast volatility of S&P 500 futures prices, making important contributions to the field.

Only recently, business research approaches have become more sophisticated, using not only ANNs but also different ML algorithms such as random forests (Coussement & Bock, 2013) and SVMs (Moro et al., 2016). These approaches investigated ML algorithms in domains such as online reviews (Singh et al., 2017), online gambling (Coussement & Bock, 2013), social media performance (Moro et al., 2016), and academic performance (Fernandes et al., 2019). In this context, and to the best of our knowledge, the work here reported represents the first attempt of exploiting AutoML in business research (see Table 1 for details). By doing so, our AutoML approach was able to promptly return accurate results on an external validation dataset of new

over-indebted cases (unknown to the algorithm during the learning phase). Such results have important practical and social implications.

6.2. Practical and Social Implications

In terms of practical and social implications, it follows from the above that the development of measures to fight over-indebtedness would considerably gain if one could a) differentiate among different profiles in a sample of identified over-indebted households; and b) based on this classification, not only estimate the risk of future cases of over-indebtedness but also anticipate the more adequate measures to reduce poverty risk.

Given that over-indebtedness is not a unified concept as indicated by the surfacing of different profiles (involving different risk factors), interventions to counteract and prevent households from becoming over-indebted should be adapted to each profile. Different policy measures put forward in Portugal (and other European Countries) have contributed to the socio-economic recovery (e.g., promoting entrepreneurship and employment creation, reposition of income). Although this certainly reduced the financial strain over many Portuguese households, it is not surprising, in light of our results, that positive signs of economic recovery do not necessarily translate into an overall reduction of over-indebted households. Thus, even if the macro economical relief felt in the last few years has helped reduce over-indebtedness, there are at least two other profiles much less dependent on crisis-related factors that might continue to grow in number. Indeed, in 2019 the indebtedness of Portuguese families attained the maximum values of the last three years, with a steady number (of about 29.000 new cases a year) of households recurring to DECO over-indebtedness support office.

By using AutoML, our model drastically reduced the time and costs of designing and developing a financial solution for over-indebted households. After testing several thousands of different algorithms using AutoML, it was possible to predict the profile of over-indebted households with a high accuracy level. Results of Support Vector Machines indicated that Nu-Support Vector Clustering had the best accuracy in predicting causes for over-indebtedness (89.5%). Such high predictive power is revealing of the applied value of using machine learning approach in the current domain of households' financial debts. Debt advisory services such as the Portuguese Association for the Consumer Defense (DECO Portugal) would gain considerably if they could use this type of predictive models when attending over-indebted families. For that, there is a beta version of the software that will contribute to reducing the time needed to identify consumers' debt profiles from the current 1-2 weeks to a few seconds. This software uses the AutoML model here presented to classify new cases of over-indebtedness into one of the three profiles and is further able to provide feedback to non-overindebted consumers based on their similarity (feature overlap) with the different profiles of over-indebtedness. The App has a user-friendly interface making it suitable to be used directly by the consumers or as a tool for more in-depth expert financial analysis. In sum, the more precise and agile these models become the more helpful their software implementations can be in providing information to not only better counseling indebted consumers but also to anticipate the risk of future cases of over-indebtedness, reducing poverty risk.

Furthermore, the systematic and continuous use of classification and prediction models of machine learning could and probably should play an important role in informing new policies to empower households and fight the negative effects of scarcity and over-indebtedness. In this sense, a recent follow-up project developed in 2020 in partnership with the Portuguese

Government (Ministry of Economy – Consumer Protection) is planning tailored interventions to prevent over-indebtedness to better fit each profile. This project is estimated to be conducted until 2022, helping an estimated number of 2,000 consumers per year.

6.3. Limitations and Future Research

This work opens several possible research opportunities using AutoML in business research. Based on our findings using AutoML to predict over-indebtedness, we foresee that any organization or company could effectively use such solutions to address their practical business problems. For instance, AutoML can be used in healthcare, marketing, retail, transportation, and many other areas that are not covered in the current research. Future research should be conducted to examine AutoML applications to other business research areas.

One of the limitations of the current research concerns the lack of data in the profiles concerning several of the psychological and situational risk factors. Adding to the database questions or tasks that could provide us with measures of consumers' tendency to rely on improper heuristics, individual differences in self-control, innumeracy, attitudes towards credit, mental accounting, well-being, etc., would be crucial to be able to confirm the initial results here reported and to refine our analyses and conclusions. Future research could add more fine-grained information to allow improving artificial intelligence tools' ability to classify and describe the over-indebtedness profiles.

In addition, there were no differences in educational levels across the three profiles. Considering the educational level as a proxy of literacy in general and financial literacy in particular, this suggests that financial illiteracy did not play a distinguishable causal role in our analysis. Given the low level of financial literacy typically found in surveys conducted in

Portugal, we suspect that innumeracy and financial illiteracy may have contributed to all profiles of over-indebtedness. However, we cannot be sure since the available data did not include a direct measure of financial literacy.

Finally, the use of AutoML could be extended as a robust methodology to describe and analyze other forms of poverty. Indeed, poverty is likely to refer to a myriad of different forms of scarcity closely related to distinctive social-economic realities. The bottom-up approach followed in this paper appears suitable to study such diversity. Research and theory in business and psychological sciences concerning the explanations and causes of poverty may then be used to give meaning to the Artificial intelligence results and to complement them with other methods (e.g., behavioral experiments).

References

Agarwal, R., R., Lin, C.-C., Chen, K.-T., & Singh, V. K. (2018). Predicting financial trouble using call data—On social capital, phone logs, and financial trouble. *Plos One*, *13* (2), e0191863. doi.org/10.1371/journal.pone.0191863

Aizcorbe, A., Kennickell, A., & Moore, K. (2003). Recent Changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances. *Federal Reserve Bulletin*, 1-32.

Al-Hashedi, M., Soon, L. K., & Goh, H. N. (2019). Cyberbullying Detection Using Deep Learning and Word Embeddings: An Empirical Study. *In Proceedings of the 2nd International Conference on Computational Intelligence and Intelligent Systems.* ACM.

Ali, M. N. Y., Sarowar, M. G., Rahman, M. L., Chaki, J., Dey, N., & Tavares, J. M. R. (2019). Adam deep learning with SOM for human sentiment classification. *International Journal of Ambient Computing and Intelligence*, *10* (3), 92-116.

Alleweldt, F., Kara, S., Graham, R., Kempson, E., Collard, S., Stamp, S., & Nahtigal, N. (2013). The over-indebtedness of European households: Updated mapping of the situation, nature and causes, effects and initiatives for alleviating its impact – Part 1: Synthesis of Findings. http://www.civic-consulting.de Accessed 26 September 2016.

Alomari, Z. (2017). Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications. *New Zealand Journal of Computer-Human Interaction, 2*.

Angel, S., Einböck, M., & Heitzmann, K. (2009). Politikgegen und Ausmaß der Überschuldung in den Ländern der Europäischen Union. http://epub.wu.ac.at/278/ Accessed 15 April 2019.

Arnold, J., & Rodrigues, C. F. (2015). Reducing Inequality and Poverty in Portugal Economics. *Organization for Economic Co-operation and Development.* http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ECO/WKP(2015)76&docLanguage=En Accessed 15 April 2019.

Bank of Portugal (2014). Economic Bulletin. https://www.bportugal.pt/en/publications/banco-de-portugal/all/381 Accessed 17 November 2018.

Bar-Gill, O., & Warren, E. (2008). Making Credit Safer. *University of Pennsylvania Law Review, 157* (1), 1-101. doi:10.2307/40041411

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The Strength Model of Self-Control. *Current Directions in Psychological Science, 16* (6), 351-355. doi:10.1111/j.1467-8721.2007.00534.x

Berthoud, R., & Kempson, E. (1992). *Credit and Debt: The PSI Report*. London: Policy Studies Institute.

Bertrand M., & Morse, A. (2009). What Do High-Interest Borrowers Do with Their Tax Rebate? *American Economic Review: Papers & Proceedings, 99* (2), 418-423. doi:10.1257/aer.99.2.418

Bertrand, M., Mullainathan, S., & Shafir, E. (2004). A Behavioral-Economics View of Poverty. *American Economic Review, 94* (2), 419-423. doi:10.1257/0002828041302019

Betti, G., Dourmashkin, N., Rossi, M., & Yin, Y. (2007). Consumer over-indebtedness in the EU: measurement and characteristics. *Journal of Economic Studies, 34* (2), 136-156. doi:10.1108/01443580710745371

Brown, S., Garino, G., Taylor, K., & Price, S. W. (2005). Debt and financial expectations: an individual and household level analysis. *Economic Inquiry, 43* (1), 100-120. doi:10.1093/ei/cbi008

Brüggen, E. C., Hogreve, J., Holmlund, M., Kabadayi, S., & Löfgren, M. (2017). Financial well-being: A conceptualization and research agenda. *Journal of Business Research, 79*, 228-237.

Canner, G. B., & Luckett, C. A. (1991). Payment of household debts. *Federal Reserve Bulletin, 77* (4), 218-229.

Celsi, M. W., Nelson, R. P., Dellande, S., & Gilly, M. C. (2017). Temptation's itch: Mindlessness, acceptance, and mindfulness in a debt management program. *Journal of Business Research, 77*, 81-94.

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

Eigsti, I.-M., Zayas, V., Mischel, W., Shoda, Y., Ayduk, O., Dadlani, M. B., … Casey, B. J. (2006). Predicting Cognitive Control From Preschool to Late Adolescence and Young Adulthood. *Psychological Science, 17* (6), 478-484. doi:10.1111/j.1467-9280.2006.01732.x

Eletter, S. F., Yaseen, S. G., & Elrefae, G. A. (2010). Neuro-Based Artificial Intelligence Model for Loan Decisions. *American Journal of Economics and Business Administration, 2*, 27-34. doi: 10.3844/ajebasp.2010.27.34.

Feurer, M., Klein, A., Eggensperger, K, Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and Robust Automated Machine Learning. *Advances in Neural Information Processing Systems,28,* 2962-2970.

DECO (2018). Statistical Bulletin1st Semester 2018 - Over-indebtedness of families The "new" causes of over-indebtedness. *Gabinete de auxílio ao sobreendividado.* https://gasdeco.net/activeapp/wp-content/uploads/2018/11/GAS-DECO-data-1st-semester-2018.pdf Accessed 17 February 2019.

Dellande, S., Gilly, M. C., & Graham, J. L. (2016). Managing consumer debt: Culture, compliance, and completion. *Journal of Business Research, 69* (7), 2594-2602.

Ghadiri, S. M. E., & Mazlumi, K. (2020). Adaptive protection scheme for microgrids based on SOM clustering technique. *Applied Soft Computing, 88,* 1-21.

Godwin, D. D. (1999). Predictors of Households' Debt Repayment Difficulties. *Financial Counseling and Planning, 10,* 67-78.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification, 3* (1), 5-48. doi.org/10.1007/BF01896809

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. NY: Macmillan College Publishing.

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science*, *12* (6), 973-986. doi:10.1177/1745691617702496

Inzlicht, M., & Schmeichel, B. J. (2012). What Is Ego Depletion? Toward a Mechanistic Revision of the Resource Model of Self-Control. *Perspectives on Psychological Science*, *7* (5), 450-463. doi:10.1177/1745691612454134

Kamleitner, B., & Kirchler, E. (2007). Consumer credit use: A process model and literature review. *EuropeanReview of Applied Psychology/Revue Européenne de Psychologie Appliquée, 57* (4), 267-283. doi:10.1016/j.erap.2006.09.003

Khatua, A., Cambria, E., & Khatua, A. (2018). Sounds of silence breakers: exploring sexual violence on Twitter. *In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE.

Kida, M. (2009). Financial vulnerability of mortgage-indebted households in New Zealand - evidence from the Household Economic Survey. *Reserve Bank of New Zealand Bulletin*, *72*, 5-12.

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural networks, 37*, 52-65.

Lee, M., Song, T. G., & Lee, J. H. (2020). Heartbeat classification using local transform pattern feature and hybrid neural fuzzy-logic system based on self-organizing map. *Biomedical Signal Processing and Control*, *57*, 1-9.

Lipkus, A. H. (1999). A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry*, *26* (1-3), 263-265. doi.org/10.1023/A:1019154432472

Loibl, C., Jones, L., & Haisley, E. (2018). Testing strategies to increase saving in individual development account programs. *Journal of Economic Psychology*, *66*, 45-63. doi:10.1016/j.joep.2018.04.002

Lusardi, A. (2008). Financial Literacy: An Essential Tool for Informed Consumer Choice? *National Bureau of Economic Research*, 1-29. doi: 10.3386/w14084.

Lusardi, A., & Mitchell, O. S. (2011). Financial literacy around the world: an overview. *Journal of Pension Economics and Finance*, *10* (4), 497-508. doi:10.1017/S1474747211000448

Lusardi, A., & Tufano, P. (2015). Debt literacy, financial experiences, and over-indebtedness. *Journal of Pension Economics and Finance, 14* (4), 332-368. doi:10.1017/S1474747215000232.

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science, 341*, 976-980. doi:10.1126/science.1238041

Marsland, S. (2015). Machine learning: an algorithmic perspective. CRC press.

Meier, S., & Sprenger, C. (2010). Present-Biased Preferences and Credit Card Borrowing. *American Economic Journal: Applied Economics*, *2* (1), 193-210. doi:10.1257/app.2.1.193

Merskin, D. (1998). The show for those who owe: Normalization of credit on lifetime's debt. *Journal of Communication Inquiry, 22* (1), 10-26. doi:10.1177/0196859998022001003

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning. *Neural and Statistical Classification, 13*, 1-298.

Mischel, W. (1958). Preference for delayed reinforcement: An experimental study of a cultural observation. *The Journal of Abnormal and Social Psychology*, *56* (1), 57-61. doi:10.1037/h0041895.

Modigliani, F. (1966). The life cycle hypothesis of saving, the demand for wealth and the supply of capital. *Social Research*, *33* (2), 160-217).

Montiel, J., Bifet, A., & Abdessalem, T. (2017). Predicting over-indebtedness on batch and streaming data. Paper presentation ate the 2017 IEEE International Conference on Big Data, Boston, MA, 1504-1513.

Nepomuceno, M. V., & Laroche, M. (2015). The impact of materialism and anti-consumption lifestyles on personal debt and account balances. *Journal of Business Research, 68* (3), 654-664.

Nilsson, N. J. (2014). Principles of artificial intelligence. Morgan Kaufmann.

Njuguna, C., & McSharry, P. (2017). Constructing spatiotemporal poverty indices from big data. *Journal of Business Research, 70,* 318-327. doi.org/10.1016/j.jbusres.2016.08.005.

Panico, C., & Purificato, F. (2013). Policy Coordination, Conflicting National Interests and the European Debt Crisis. *Cambridge Journal of Economics, 37* (3), 585-608. doi: 10.1093/cje/bet009

Pattarin, F., & Cosma, S. (2012). Psychological determinants of consumer credit: the role of attitudes. *Review of Behavioral Finance, 4* (2), 113-129. doi:10.1108/19405971211284899

Resta, M. (2012). *Graph mining-based SOM: a tool to analyze economic stability. In Applications of Self-Organizing Maps*. *IntechOpen*. doi: 10.5772/51240. https://www.intechopen.com/books/applications-of-self-organizing-maps/graph-mining-based-som-a-tool-to-analyze-economic-stability.

Sawhney, R., Manchanda, P., Mathur, P., Shah, R., & Singh, R. (2018). Exploring and learning suicidal ideation connotations on social media with deep learning. *In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 167-175).*

Schölkopf, B., Smola, A. J., & Bach, F. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press.

Shaefer, H. L., & Edin, K. (2013). Extreme Poverty in the United States and the Response of Federal Means-Tested Transfer Programs. *Social Service Review, 87* (2), 250-268.

Slowik, J. (2012). Credit CARD Act II: Expanding Credit Card Reform by Targeting Behavioral Biases. *UCLA law review, 59*, 1292-1341.

Stanovich, K. E, & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*, 672-695. doi:10.1037/0022-3514.94.4.672

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. New Haven, CT, US: Yale University Press.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2011). Individual differences as essential components of heuristics and biases research. In K. Manktelow, D. Over, & S. Elqayam (Eds.), *The science of reason: A festschrift for Jonathan St. B. T. Evans* (pp. 335-396). New York: Psychology Press.

Statistics Portugal (2017). 2.6 million at-risk-of-poverty or social exclusion – 2016. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=281091354&DESTAQUESmodo=2&xlang=en. Accessed 14 November 2018.

Statistics Portugal (2019). 17.2% at-risk-of poverty - 2019. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=354099803&DESTAQUEStema=5414278&DESTAQUESmodo=2. Accessed June 14 2020.

Strömbäck, C., Lind, T., Skagerlund, K., Västfjäll, D., & Tinghög, G. (2017). Does self-control predict financial behavior and financial well-being? *Journal of Behavioral and Experimental Finance,14*, 30-38. doi:10.1016/j.jbef.2017.04.002

Sullivan, A. C., & Fisher, R. M. (1988). Consumer credit delinquency risk: Characteristics of consumers who fall behind. *Journal of Retail Banking, 10*, 53-64.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT, US: Yale University Press.

Tokunaga, H. (1993). The use and abuse of consumer credit: application of psychological theory and research. *Journal of Economic Psychology 14* (2), 285-316. doi:10.1016/0167-4870(93)90004-5.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample. *Journal of Behavioral Decision Making*, *30*, 541-554.

United Nations (2019a). Ending Poverty. https://www.un.org/en/sections/issues-depth/poverty/ Accessed 23 April 2019.

United Nations (2019b). The Sustainable Development Agenda. https://www.un.org/sustainabledevelopment/development-agenda/ Accessed 23 April 2019.

Van Staveren, I. (2002). *Global finance and gender*. Paper presented at the Gender Budgets, Financial Markets, Financing for Development conference, Berlin.

Vohs, K. D., & Faber, R. J. (2007). Spent resources: Self-regulatory resource availability affects impulse buying. *Journal of Consumer Research, 33* (4), 537-547. doi:10.1086/510228

Vohs, K. D., & Heatherton, T. F. (2000). Self-Regulatory Failure: A Resource-Depletion Approach. *Psychological Science*, *11* (3), 249-254. doi:10.1111/1467-9280.00250

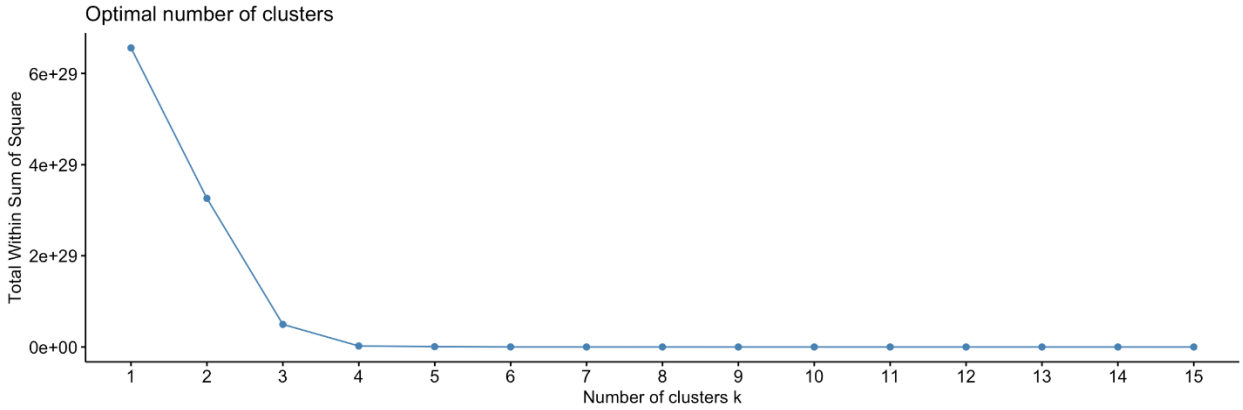
Watkins, J. P. (2000). Corporate power and the evolution of consumer credit. *Journal of Economic Issues*, *34* (4), 909-932. doi:10.1080/00213624.2000.11506321

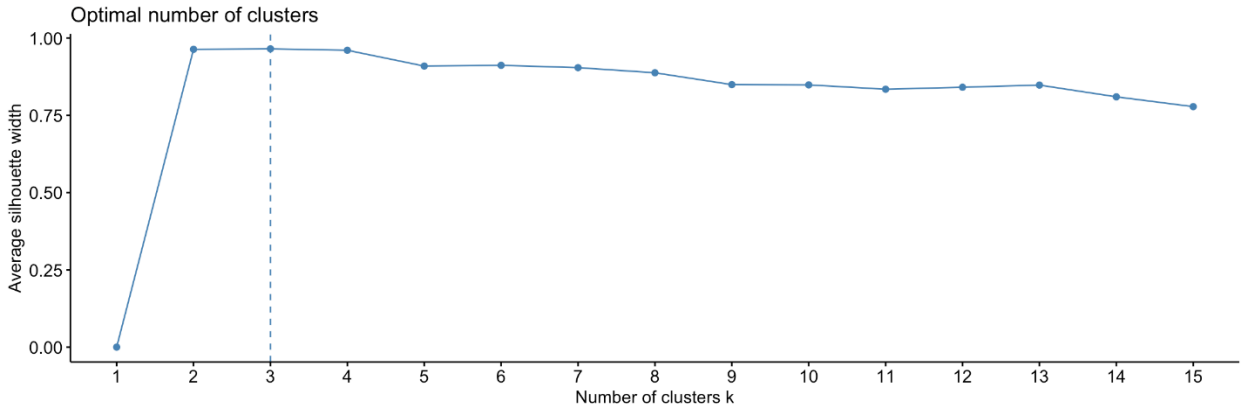
Zhao, J. & Tomm, B. (2018). Psychological responses to scarcity. *Oxford Research Encyclopedia of Psychology*. New York: Oxford University Press.

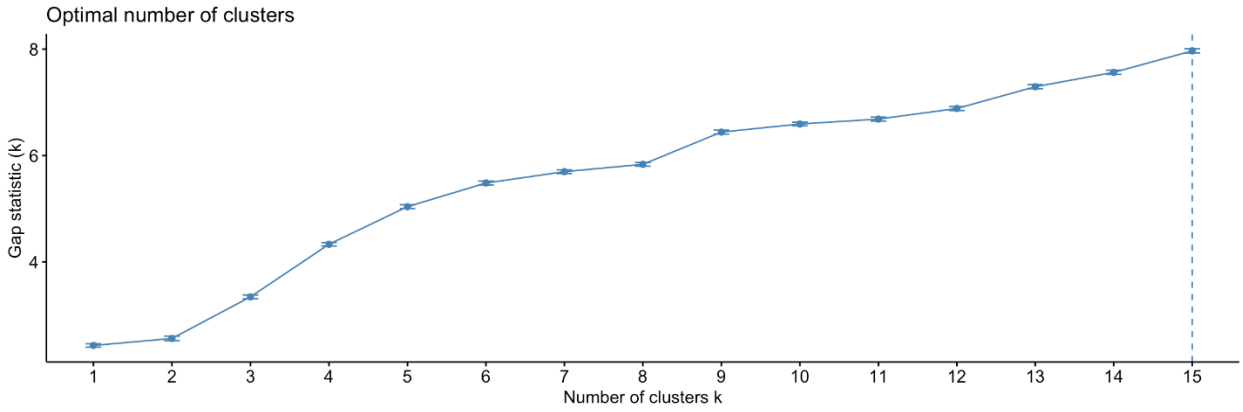
Appendix A: Dataset Variables

Feature	Data Type	Group	Note
Process Number	Categorical	N/A	Unique anonymous identifier of consumer process
Marital Status	Categorical	Social-demographic	
People in the household	Numeric	Social-demographic	
Level of Education	Categorical	Social-demographic	
Years of study	Numeric	Social-demographic	
Employment status	Categorical	Social-demographic	From a predetermined set of employment status
Causes of over-indebtedness	Categorical	Perceived Causes	From a predetermined set of causes
Cause classification	Categorical	Perceived Causes	Crisis and Other
Income Total	Numeric	Economic Situation	In Euros
Income per capita	Numeric	Economic Situation	In Euros
Income after Expenses (Net Income)	Numeric	Economic Situation	In Euros
Expenses of the household	Numeric	Economic Situation	In Euros
Expenses per capita	Numeric	Economic Situation	In Euros
Expenses – effort rate	Numeric	Economic Situation	% of income
All credits - monthly installment	Numeric	Economic Situation	In Euros
All credits - quantity	Numeric	Economic Situation	
All credits - effort rate	Numeric	Economic Situation	% of income
Credit Card - monthly installment	Numeric	Economic Situation	In Euros
Credit Card - quantity	Numeric	Economic Situation	
Credit Card - effort rate	Numeric	Economic Situation	% of income
Credit Card - participation	Numeric	Economic Situation	% of Credits Total
Housing Credit - monthly installment	Numeric	Economic Situation	In Euros
Housing Credit - quantity	Numeric	Economic Situation	
Housing Credit - effort rate	Numeric	Economic Situation	% of income
Housing Credit - participation	Numeric	Economic Situation	% of Credits Total
Car Credit - monthly installment	Numeric	Economic Situation	In Euros
Car Credit - quantity	Numeric	Economic Situation	
Car Credit - effort rate	Numeric	Economic Situation	% of income
Car Credit - participation	Numeric	Economic Situation	% of Credits Total
Personal Credit - monthly installment	Numeric	Economic Situation	
Personal Credit - quantity	Numeric	Economic Situation	
Personal Credit - effort rate	Numeric	Economic Situation	% of income
Personal Credit - participation	Numeric	Economic Situation	% of Credits Total
Other Credits - monthly installment	Numeric	Economic Situation	In Euros
Other Credits - quantity	Numeric	Economic Situation	
Other Credits - effort rate	Numeric	Economic Situation	% of income
Other Credits - participation	Numeric	Economic Situation	% of Credits Total
Highest credit type	Categorical	Economic Situation	

Appendix B: Unsupervised ML Analysis - Optimal Number of Clusters







FIGURES AND TABLES

Figure 1: Supervised Machine Learning algorithms that have been used in this work.

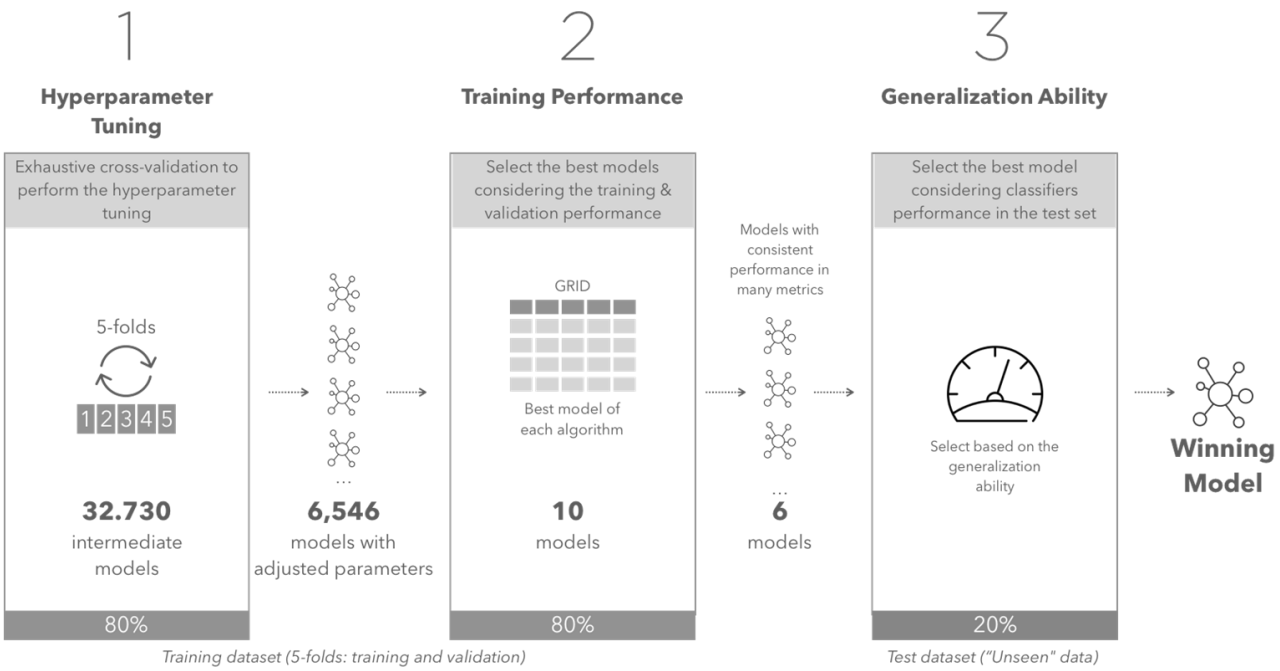


Figure 2: Grid Search Hyperparameters Tuning Process

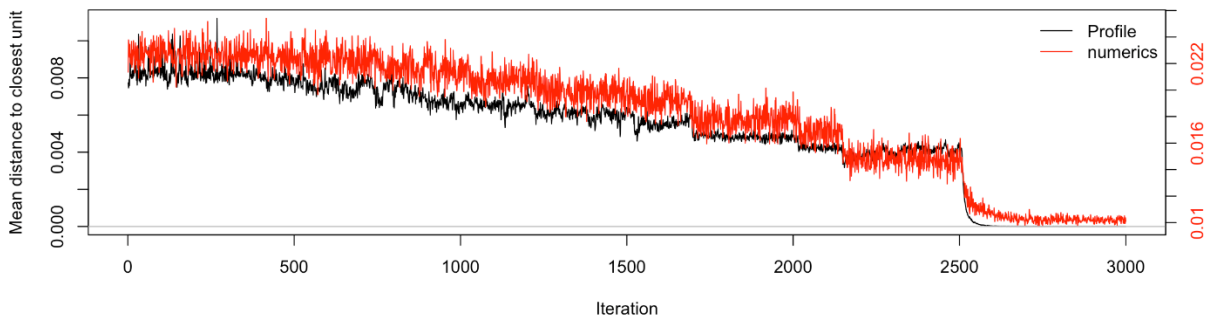


Figure 3: Self-Organizing Maps Training Progress

Accuracy Score

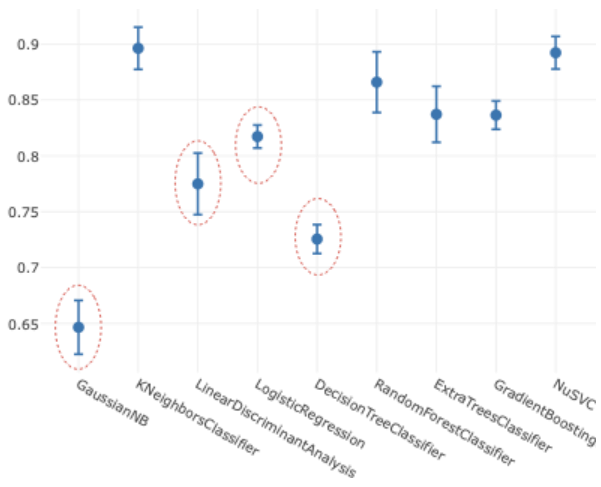


Figure 4a: Accuracy Score – Machine Algorithms

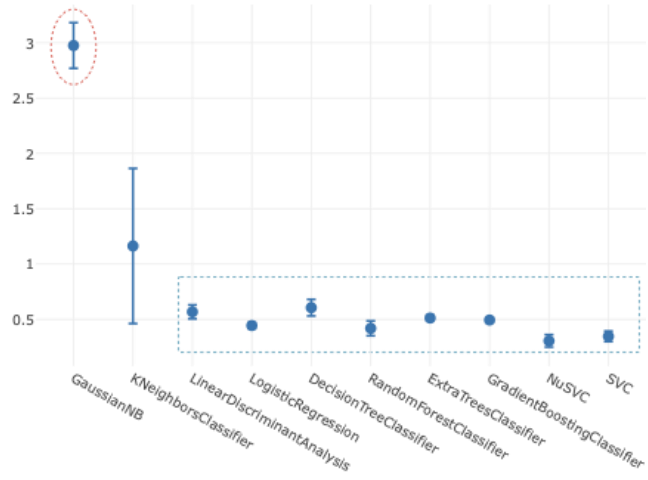


Figure 4b: Log Loss - ML Algorithms

Log Loss

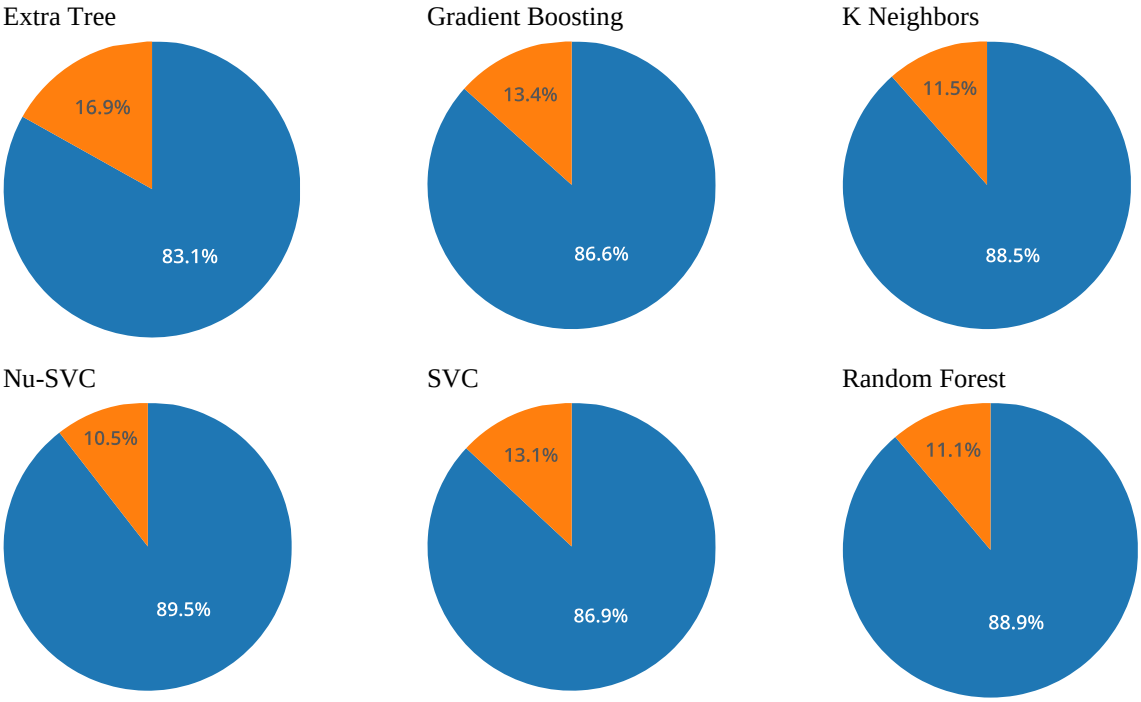


Figure 5: Top 6 Classifiers - Model Comparison.

Authors	Algorithms tested	Context	Empirical/Conceptual	AutoML
Bejou et al. (1996)	Artificial Neural Networks	Customer relationship management	Empirical	No
Coussement & Bock (2013)	Decision trees, generalized additive model, random forest, and GAMens	Customer churn prediction in the online gambling	Empirical	No
Delen and Zolbanin (2018)	None (review of descriptive, predictive, and prescriptive algorithms)	Analytics paradigm in business research	Conceptual	No
Fernandes et al. (2019)	Gradient Boosting Machine (GBM)	Predictive analysis of academic performance	Empirical	No
Fish et al. (2004)	Artificial Neural Networks	Model brand market share	Empirical	No
Hamid and Iqbal (2004)	Artificial Neural Networks	Forecasting volatility of S&P 500 futures prices	Empirical	No
Moro et al. (2016)	Support vector machines	Predicting social media performance metrics	Empirical	No
Orriols-Puig et al. (2013)	Fuzzy-CSar	Knowledge Discovery in Databases	Empirical	No
Singh et al. (2017)	Gradient boosting algorithm	Predicting the helpfulness of online consumer reviews	Empirical	No
Sivarajah et al. (2017)	None (review of analytical methods)	Big data analytics	Conceptual	No
Xu et al. (2016)	None (review of marketing analytics and big data analytics)	Big data analytics on new products	Conceptual	No

Table 1: Business Research Studies Using Machine Learning Algorithms

Numerical Variables	**F_statistics (ANOVA)**	**Cluster 1**	***p***	**Cluster 2**	***p***	**Cluster 3**	***p***
Income per capita	$F_{(2, 1564)} = 162.6146$, $p = 0.000$, $\eta p^2 = 0.1721$	€ 401.94	*NS*	€ 686.35	*	€ 413.15	*NS*
Household expenses	$F_{(2, 1564)} = 41.5088$, $p = 0.000$, $\eta p^2 = 0.0504$	€ 736.13	*NS*	€ 570.85	*	€ 790.69	*NS*
People in the household	$F_{(2, 1564)} = 124.4617$, $p = 0.000$, $\eta p^2 = 0.1373$	€ 2.65	*NS*	€ 1.78	*	€ 2.76	*NS*
All Credits - monthly installment	$F_{(2, 1564)} = 37.5157$, $p = 0.000$, $\eta p^2 = 0.0458$	€ 453.65	*	€ 732.30	*NS*	€ 683.35	*NS*
Credit Card - monthly installment	$F_{(2, 1564)} = 20.4283$, $p = 0.000$, $\eta p^2 = 0.0255$	€ 149.54	*NS*	€ 284.91	*	€ 193.74	*NS*
Car Credit - monthly installment	$F_{(2, 1564)} = 38.3585$, $p = 0.000$, $\eta p^2 = 0.0468$	€ 70.34	*NS*	€ 19.88	*	€ 193.74	*NS*
Housing Credit - monthly installment	$F_{(2, 1564)} = 47.3656$, $p = 0.000$, $\eta p^2 = 0.0571$	€ 80.21	*	€ 158.90	*	€ 209.63	*
Personal Credit - monthly installment	$F_{(2, 1564)} = 35.6587$, $p = 0.000$, $\eta p^2 = 0.0436$	€ 146.30	*NS*	€ 246.00	*	€ 138.10	*NS*
Other Credit - monthly installment	$F_{(2, 1564)} = 41.1646$, $p = 0.000$, $\eta p^2 = 0.05$	€ 7.25	*NS*	€ 13.61	*NS*	€ 79.54	*
All credits - effort rate	$F_{(2, 1564)} = 85.4148$, $p = 0.000$, $\eta p^2 = 0.0985$	40%	*	75%	*	68%	*
Credit Card - effort rate	$F_{(2, 1564)} = 36.0509$, $p = 0.000$, $\eta p^2 = 0.0441$	12%	*	29%	*	19%	*
Car Credit - effort rate	$F_{(2, 1564)} = 28.9662$, $p = 0.000$, $\eta p^2 = 0.0357$	8%	*NS*	2%	*	7%	*NS*
Housing Credit - effort rate	$F_{(2, 1564)} = 54.7266$, $p = 0.000$, $\eta p^2 = 0.0654$	6%	*	16%	*	20%	*
Personal Credit - effort rate	$F_{(2, 1564)} = 76.8049$, $p = 0.000$, $\eta p^2 = 0.0894$	12%	*NS*	28%	*	12%	*NS*
Other Credit - effort rate	$F_{(2, 1564)} = 42.1191$, $p = 0.000$, $\eta p^2 = 0.0511$	0.4%	*NS*	0.8%	*NS*	10%	*
Categorical Variables	**Chi-Square Statistics**	**Cluster 1**	***p***	**Cluster 2**	***p***	**Cluster 3**	***p***
Profile - Crisis	$\chi^2_{(2, 1567)} = 899.0587$, $p = 0.000$	0.0%	*	16.0%	*	83.7%	*
Profile - Other		100.0%		84.0%		16.3%	
Marital_Status - Married	$\chi^2_{(8, 1497)} = 104.2944$, $p = 0.000$	40.0%	*	21.5%	*	45.6%	*
Marital_Status - Single		23.6%		34.6%		20.0%	
Marital_Status - Other		36.4%		43.9%		34.4%	
Causes_Difficulties - Unemployment	$\chi^2_{(24, 1463)} = 541.697$, $p = 0.000$	6.6%	*	11.4%	*	40.5%	*
Causes_Difficulties - Family Growth		12.8%		2.4%		7.1%	
Causes_Difficulties - Other		80.6%		86.2%		52.4%	
Educational_Level - Basic	$\chi^2_{(4, 1455)} = 0.9608$, $p = 0.9157$	40.9%	NS	41.8%	NS	41.1%	NS
Educational_Level - Secondary		40.2%		38.2%		40.7%	
Educational_Level - College		19.0%		20.0%		18.2%	
Professional_Situation - Unemployed	$\chi^2_{(8, 1431)} = 119.5188$, $p = 0.000$	11.0%	*	11.8%	*	31.8%	*
Professional_Situation - Retired		10.8%		19.5%		6.4%	
Professional_Situation - Working		78.2%		68.7%		61.8%	
Highest_Credit - Credit Card	$\chi^2_{(8, 1567)} = 247.7595$, $p = 0.000$	26.1%	*	31.4%	*	27.3%	*
Highest_Credit - Housing		8.8%		25.8%		31.4%	
Highest_Credit - Personal Credit		28.6%		37.2%		13.2%	
Highest_Credit - Other		36.5%		5.6%		28.1%	

Table 2: Automated Feature Selection for Cluster Profiling (numerical and categorical features).

Parameter	Value	Description
Nu	*0.08*	It corresponds to the probability of acceptance to find observations outside the frontier. This value must be in the interval between 0 and <1. Consequently, 0.08 indicates 8% at the maximum of the training samples are acceptable to be incorrectly classified and at least 8% of the training set can act as support vectors.
Kernel	*RBF*	It specifies the kernel type to be used by the learning algorithm. In the winning model was used RBF – Radial Basis Function. Possible values*: 'linear', 'poly', 'rbf', 'sigmoid', and 'precomputed'*.
Gamma	*Scale*	This parameter defines the influence level which a single training observation can reach and affect the 'curve' (decision boundary) during the learning process. When gamma has a high value, the influence of one observation is higher to define the 'curve', if this value is "big" it can create "islands" and consequently can generate overfitting easily. Low values of gamma, the influence of one observation is lower to define the 'curve'. Possible values: • *'auto' – 1 / number of features* as gamma value • *'scale' – 1 / (number of features * X.var())* as gamma value Remark: It is a kernel coefficient only applicable for *'rbf', 'poly'* and *'sigmoid'*.
Decision function shape	*"ovr"*	The decision function used to separate one group from another using the observations of the dataset. Possible values: • *'ovr' – One Verse Rest* • *'ovo' – One Verse One*
Class Weight	*Balanced*	It is used to set the value of the *C* parameter for each class. If this value is not defined, all classes are assumed to have weight one. The '*balanced*' value uses the data of the target variable (y) to automatically adjust the weights of each class using the classes frequencies in the training set. Possible values: *'balanced'* and *dict*

Table 3: The winning model configuration (Nu-SVC).

	Mean accuracy	Std accuracy	Mean bal_acc	Std bal_acc	Mean f1_macro	Std f1_macro	Mean log loss	Std log loss	Mean precision macro	Std precision macro	Mean recall macro	Std recall macro
Gaussian NB	0.646509	0.024092	0.671085	0.024783	0.608481	0.025551	2.975186	0.206400	0.717132	0.050065	0.671085	0.024783
K Neighbors	0.856223	0.018888	0.857983	0.018360	0.897674	0.019018	1.162856	0.701356	0.900295	0.018939	0.897983	0.018360
LDA	0.774867	0.027593	0.781957	0.025002	0.775904	0.030079	0.567435	0.062709	0.779909	0.029101	0.781957	0.025002
Log Reg	0.817216	0.010291	0.821844	0.008500	0.820368	0.011720	0.443736	0.026994	0.823359	0.015010	0.821844	0.008500
Decision Tree	0.725458	0.012966	0.736014	0.013784	0.720700	0.019810	0.605550	0.075023	0.750819	0.012217	0.736014	0.013784
Random Forest	0.865861	0.027188	0.869309	0.026868	0.868500	0.027157	0.418591	0.067184	0.868907	0.026631	0.869309	0.026868
Extra Trees	0.837118	0.025094	0.842403	0.022295	0.838138	0.026499	0.511755	0.026160	0.840766	0.024465	0.842403	0.022295
Gradient Boost.	0.836369	0.012675	0.838940	0.013125	0.839540	0.012855	0.493676	0.022844	0.842056	0.012014	0.838940	0.013125
Nu SVC	0.892248	0.014584	0.895422	0.013860	0.894278	0.014343	0.305083	0.056059	0.895620	0.015901	0.895422	0.013860
SVC	0.840369	0.026918	0.839090	0.025384	0.840710	0.025842	0.346097	0.047502	0.849353	0.027162	0.839090	0.025384

Table 4: Supervised ML Analysis - Machine Learning Algorithms performance.