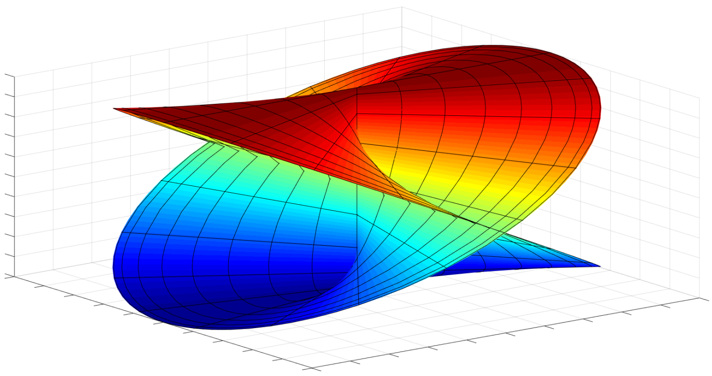
UNDERSTANDING OVER-INDEBTEDNESS IN PORTUGAL: DESCRIPTIVE AND PREDICTIVE MODELS

BY

LEONARDO VANNESCHI
DIEGO COSTA PINTO

NOVA Information Management School (NOVA IMS)
Universidade Nova de Lisboa
Campus de Campolide, 1070-312 Lisboa, Portugal



Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa NOVA Information Management School (NOVA IMS)

Title:
Understanding Over-Indebtedness in Portugal: Descriptive and Predictive Models

Authors:
Leonardo Vanneschi and Diego Costa Pinto

Edition:
Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa NOVA Information Management School (NOVA IMS)

ISBN:
978-972-8093-20-4

Digital Version:
http://hdl.handle.net/10362/114093

This work is supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., in the context of the project AICE (DSAIPA/DS/0113/2019)

How to cite:
Vanneschi, L. Pinto, D. C. (2021). Understanding over-indebtedness in Portugal: descriptive and predictive models. Lisboa: Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa. NOVA Information Management School (NOVA IMS). ISBN: 978-972-8093-20-4. Link: http://hdl.handle.net/10362/114093

Acknowledgements:
The authors gratelly thank all the researchers and collaborators involved in this project: Prof. Mário Boto Ferreira (FP-UL), Prof. Mauro Castelli (NOVA IMS), Prof. Márcia Herter (UE), Jerônimo Soro, Filipa de Almeida, Fernando Peres, Victor Costa, Dave Montali, Carina Castagna, Saleh Shuqair, and Yasmina El Fassi.

Contents

Chapter 1
Introduction

Over-indebtedness is a recurring problem in Portugal. After facing different economic cycles - between financial crises and prosperity periods - Portuguese consumers have been striving to keep their household finances stable and avoid being over-indebted. In the last decades, credit consumption expanded in Portugal, developing different debtors' classes and social strata. Indeed, many families continue to face financial difficulties, and a representative part of the Portuguese population still cannot pay their debts. This caused an increase in the family effort rate to about 70% in the recent years (DECO, 2018). Also, recently, about 4.4 million Portuguese consumers were indebted (Bank of Portugal, 2014). Thus, this project aims to gain insights on over-indebtedness, from different perspectives that range from the social to the economic point of view (Marques Frade, 2000). In particular, this project examines over-indebtedness from a psychological and from a data science perspective.

Previous research on over-indebtedness has concentrated on individuals' socio-economic, personal, and situational circumstances (e.g., Berthoud Kempson, 1992; van Staveren, 2002). Research indicates that vulnerability for over-indebtedness is mainly determined by socio-economic factors (Angel, Einbock, Heitzmann, 2009) and financially relevant life events such as job loss (for a review see Kamleitner Kirchler, 2007). However, most of this research did not explore the contribution that artificial intelligence can give in characterizing and predicting consumers' over-indebtedness.

1.1 Understanding Over-indebtedness: from Psychology to Machine Learning

In this project, we suggest that the systemic impact of financial crisis in Portugal not only promotes over-indebtedness but it crafts a specific profile of over-indebted consumers which may be distinguished from other profiles, namely the one that puts the emphasis on lack of self-regulation and careless management of one's budget as causal factors in a culture of intense consumerism.

As a result, the same psychological factors such as attitudes towards debt (e.g., Chien Devaney, 2001; Livingstone Lunt, 1992), time preferences (e.g., Groenland Nyhus, 1994), tendency to decide based on improper heuristics (e.g., Slovic, 2011), financial literacy (e.g., Luzardi Tufano, 2008), among others, may play markedly different moderator roles, precede or result from debt problems, depending on the over-indebtedness profile of the consumer.

Given this scenario, this project proposes the use of Machine Learning (ML) for developing descriptive and predictive models, to understand the influencing factors of over-indebtedness on Portuguese consumers. Descriptive models will be obtained using Unsupervised ML algorithms like Self Organizing Maps and Agglomerative Hierarchical Clustering and will be used for establishing consumer clusters and guidelines for over-indebtedness regulation and consumer financial empowerment.

Predictive models will be obtained using Supervised ML algorithms such as Decision Trees, Support Vector Machines, various versions of supervised Artificial Neural Networks and Genetic Programming (GP). A focus will be given to a recent and extremely promising version of GP, called Geometric Semantic GP (GSGP), recently developed with the crucial contribution of the PI and other team members. GSGP has returned an impressive number of applicative successes (Vanneschi, 2017), often outperforming the state of the art methods in many different areas. This success is due to the property of GSGP of inducing a unimodal error surface for any supervised learning problem. This property holds independently from the complexity and dimension of the data, which makes GSGP an appropriate candidate for tackling the problems related to this project, which are characterized by vast and extremely complex data. A rigorous comparison between the different used algorithms is crucial and will be carried on by means of Automated ML (Feurer, 2015), a set of techniques for automatizing the ML process, enabling the evaluation of thousands of models with multiple combinations of parametrization, and different types of feature selection methods.

1.2 Profiling Over-indebtedness in Portugal

Based on the biographical data available, we used the reported causes of over-indebtedness as proxies to identify and define different profiles based on subjective internal causes (related to poor money management, excessive expenditure), on subjective external causes potentially derived from the economic austerity imposed in the country or on other external causes not directly linked to the economic crisis. We expect to find differences between these profiles in participants behavior towards credit, as suggested by their informed financial situation, such as number, types and debt status of credits the participants may have and their estimates of monthly income and expenditure.

In sum, this project aims to: (1) characterize and describe over-indebtedness of Portuguese consumers using unsupervised ML; (2) create reliable supervised ML models to help to predict the factors that influence over-indebtedness. These models should help investigate and verify the influence of psychological factors such as attitudes towards debt (e.g., Chien Devaney, 2001; Livingstone Lunt, 1992), time preferences (e.g., Groenland Nyhus, 1994), tendency to decide based on improper heuristics (e.g., Slovic, 2011), financial literacy (e.g., Luzardi Tufano, 2008), among other factors. To test for the existence of distinguishable over-indebtedness cluster profiles that results from long-term exposure to an environment of severe austerity, we analyzed a database of 1,654 households (over-indebted consumers) who contacted the debt advisory services of the Portuguese Association for the Consumer Defense (DECO Portugal) and also getting insights from the Direction of Consumer (DGC) for counsel and deep understanding of the roots of over-indebtedness on how to deal with their financial debts.

Chapter 2
Credit Use and Over-Indebtedness

Over the last decades, Portuguese consumers have been gradually more open to the idea of using credit as a way of obtaining a liquidity that their paychecks would not otherwise allow (Watkins, 2000). With a more extensive and regular use of credit, the idea of being indebted has become less dreaded and progressively normalized as an inherent state shared by most consumers in order to obtain necessary goods and services (Merskin, 1998). Such tendency has been spreading worldwide with most Western societies reporting consecutive increases in household debt levels, as well as cases of over-indebtedness (e.g., Betti, Dourmashkin, Rossi Yin, 2007; Brown, Garino, Taylor, Price, 2005; Kida, 2009; Pattarin Cosma, 2012).

2.1 Factors Driving Over-indebtedness

Over-indebtedness may be described as the persistent incapability to meet all payment obligations and life expenses when they are due. Becoming over-indebted has a pronounced impact in households, leading to social stigma, deterioration of health and well-being, relationship difficulties, financial exclusion, and reduced labor market activity (e.g., Alleweldt et al., 2013).

Although low income leaves households at higher risks of becoming over-indebted (Aizcorbe, Kennickel, Moore, 2003; Sulivan Fisher, 1988), over-indebtedness is not necessarily a situation exclusive of households with low income (Betti, Dourmashkin, Rossi, Yin, 2007; Canner Luckett, 1991). Indeed, over-indebtedness has been shown to have multiple causes. Some individualistic factors are known to make consumers more vulnerable to debt repayment difficulties. Financial illiteracy or lack of knowledge about financial products and concepts (Lusardi Tufano, 2015), biased thinking (Slowik, 2012), materialism and impulsive behavior (Garðarsdóttir Dittmar, 2012; Watson, 2003) are among the most prominent of these factors. Both financial illiteracy and lack of self-control have been found to be positively associated with over-indebtedness (Gathergood, 2012).

Also measures of materialism (i.e., the importance given to material goods and assets as means for achieving major life goals) have been associated to a tendency to accumulate debt (Garðarsdóttir Dittmar, 2012; Watson, 2003). Furthermore, the effects of low financial literacy in the accumulation and repayment of debts can be aggravated by consumers biased decisions stemming from the use of heuristics (Thaller Sustein, 2008). One example is the present-bias preference or the asymmetrical perception of the value of present gains against future losses, which leads to increased use of credit and disregard for the accumulation of interest (Slowik, 2012). Indeed, individuals displaying present-bias preferences (i.e., desire for immediate consumption) show a higher tendency to have more credit-card debt (Meier Sprenger, 2010; Strömbäck, Lind, Skagerlund, Västfjäll, Tinghög, 2017).

External factors, such as socio-economic characteristics, adverse situational circumstances and significant life events, also work as important determinants of over-indebtedness. For instance, younger consumers and more numerous households (especially with more children) are associated to debt repayment difficulty (Canner Luckett; Godwin, 1999). Such is also the case of 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 experience of adverse life events in the last 12 months are associated to households with debt repayment strain in comparison to a control group (Tokunaga, 1993). Abrupt changes in socio-economic conditions can thus launch (mostly middle-class) households into financial strains and increased risk of indebtedness as it happened during the European sovereign debts crisis that followed the 2008-2009 World economic recession. The Portuguese debt crisis is a case in point.

2.2 Consequences of Over-indebtedness

To be indebted at certain stages of the life cycle may allow households to reduce liquidity constraints, it increases consumption and improves households' economic welfare (e.g., Hall, 1978). However, over-indebtedness becomes most problematic when it exceeds household resources leading to the inability to meet all payment obligations and cover living expenses over long time periods. Indeed, the burden of over-indebtedness has been shown to have a negative impact on both psychological and physical health.

Individuals with unmet loan payments show more suicidal ideation and are at a higher risk of depression than those without such financial difficulties (e.g., Gathergood, 2012, Hintikka, Viinamäki, Tanskanen, Kontula, Koskela, 1998; Turunen Hiilamo, 2014). Unpaid financial obligations have been also associated with poorer subjective health, deterioration of health-related behaviour and physical illness (Bridges Disney, 2010; Chmelar, 2013; Clayton, Linares-Zegarra, Wilson, 2015; Guiso Sodini, 2013; Lenton Mosley, 2008; Sweet et al., 2013; Turunen Hiilamo, 2014). Confirming this pattern, a recent longitudinal study of Finnish adults found an association between over-indebtedness and an increased incidence of various chronic diseases (Blomgren, Maunula, Hilamo, 2016). Warth et al. (2019) found a negative relationship between over-indebtedness and sleep quality.

Notably, it can play a major role in a variety of health problems, from hypertension (Buxton, Marcelli, Short, 2010; Gangwisch et al., 2006; Meng, Zheng, Hui, 2013), to diabetes (Buxton et al., 2010, Grandner, Chakravorty, Perlis, Oliver, Gurubhagavatula, 2014; Morselli, Guyon, Spiegel, 2012; Zizi et al. , 2012) and mortality (Cappuccio, D'Elia, Strazzullo, Miller, 2010; Gallicchio Kalesan, 2009; Grandner, Hale, Moore, Patel, 2010; Grandner Patel, 2010).

Taken together, such detrimental consequences are worrisome given the increasing number of over-indebted households across Europe and around the world (e.g., Betti, Dourmashkin, Rossi, Yin, 2007; Harvey, 2010; Kempson, 2015; Pivetti, 2008), and highlights the importance of further research to better understand the relationship between over-indebtedness and different indicators of well-being and quality of health.

2.3 Over-indebtedness: a Machine Learning Approach

Previous research has identified several individualistic and situational factors which have been found to be associated with increased risk of over-indebtedness. However, the extant literature also suggests that categorizing someone as "being over-indebted", may be putting under the same conceptual umbrella distinct profiles of indebtedness. These profiles may involve not only different risk factors but also specific combinations of these factors. Social measures and public policies targeted at helping over-indebted consumers are likely to be less effective if they adopt a "one size fits all" approach rather than considering the differential features of distinct debt profiles.

However, most of this research did not explore how artificial intelligence can classify and predict consumers' over-indebtedness, helping to assist in the Alternative Dispute Resolution (RAL) of consumer debt. From this viewpoint, the idea of the project of employing Machine Learning (ML) to characterize and predict over-indebtedness is entirely novel.

This project covers different scientific areas and addresses various challenges related to the past and current research activities of the team. In particular, team members demonstrated the capacity to analyze vast amounts of transactional data (Vanneschi et al., 2018), (Castelli et al., 2017a), and produced state-of-the-art results on many applicative domains using ML, in fields like forecasting of the vessels position at sea (Vanneschi et al, 2017a), predicting the relative position of computer tomography slices (Castelli et al., 2016), predicting the proteins tertiary structure (Castelli et al., 2015a), analysis of reviews of Amazon's products (Castelli et al., 2017b), forecasting of energy consumption (Castelli et al., 2015b), (Castelli et al., 2015c), electoral redistricting (Vanneschi et al., 2017b), predicting pharmacokinetic parameters (Vanneschi et al., 2014), predicting the unified Parkinson's disease rating scale assessment (Castelli et al., 2014), predicting high performance concrete strength (Castelli et al., 2013a), and classification of land cover/land use (Castelli et al, 2013b). These numerous applicative successes encourage us to use the previously studied ML methods for characterizing and predicting over-indebtedness. A first significant first step has already been done by the team members (Ferreira et al., 2019).

The importance of the objective is highlighted by the state of affairs triggered by the European debt crisis, which calls for new evidence based guidelines to help consumers gain control of their financial decisions and prevent the social contagion detrimental effects of scarcity and a generalized fear of over-indebtedness. We are certainly not the first to take such endeavor, but where past research has often looked at relatively stable differences at the level of financial literacy (e.g., Lusardi Mitchell, 2011) and the use of improper decisional heuristics (e.g., Thaler Sustein, 2008) as the main causes of poor financial decisions, the hallmark of our team's recent research is the emphasis on the flexibility and potential of human decision making.

Consumer's financial decisions may vary greatly depending on the characteristics of the financial market environment; the way the consumer represents and gives meaning to such environment, and how her goals and motivations stimulate reasoning in such environment. Next, we contr ast some new insights to approach over-indebtedness, based on our research with other important strategies already put forward by previous work. Thaler and Sustein (2008) argues for the disclosure of relevant information to individual consumers as the paradigmatic (but not the only) means of enabling better decisions.

Although this is undoubtedly an important point, we take a different approach: affording consumers with cumulative learning from multiple financial decisions via interactive computer based simulation of alternative world like financial scenarios. Previous research using a sequential judgment paradigm to manipulate the prevalence of heuristic versus rule based reasoning (Ferreira et al., 2006, Studies 3 4) already revealed the potential of this repeat play learning strategy. We found that decision contexts where heuristics provide suitable answers increased heuristic judgments whereas contexts promoting abstract thinking increased control and deliberate rule-based reasoning. A second common measure to fight over-indebtedness is financial education.

However, although several studies have confirmed the negative association between innumeracy and households' financial decision making, there is mixed evidence on the effectiveness of financial education programs (Lusardi, 2009), which suggests the need for a better understanding of this phenomenon. Moreover, common financial problems that are solved by mass providers in seconds often challenge even consumers with quite sophisticated numerical skills. For instance, when the evolution of interest rates and debt payments are non-linear, decision processes typically fail because consumers tend to assume linear trends when conceiving this kind of financial problems (Soll, Keeney, Larrick, 2014).

In sum, the lack of numerical skills may be just one factor leading to impoverished representations of financial problems that consumers take for granted. We based this view on previous research (Mata et al., 2014; Kruger et al., 2014). Using change detection methods borrowed from text comprehension literature, we were able to identify systematic distortions/omissions of numerical and inferential problems' representations.

To deal with this issue, we propose to a) identify consumers' specific misrepresentations of financial data using artificial intelligence methods; b) use these results to assist in the Alternative Dispute Resolution (RAL) of consumer debt. In sum, this project checks for the existence of different debt profiles: a) distinguishable features of over-indebted consumers based on the consumers perceived causes of their own debt situation; and b) the situational and/or dispositional factors that better differentiate these profiles, using machine learning. By doing so, we expect to find distinguishable debt profiles directly related to the socio-economic consequences of the Portuguese financial crisis or to low self-control or careless budget management.

Chapter 3
Machine Learning Algorithms for Descriptive and Predictive Modelling

Machine Learning (ML) [Mitchell, 1997, Shalev-Shwartz and Ben-David, 2014] is a field of study whose objective is to program computers to automatically learn to solve a problem, or accomplish a task. ML is useful when manually programming a computer to carry out a task is either impractical or infeasible. Typical cases are either problems that are so complex to be beyond human capabilities, like the ones characterized by vast amounts of data, or tasks that living beings perform routinely, yet our introspection on how we do it is not sufficiently elaborated to allow us to extract a well defined algorithm, like for instance driving, speech recognition, image understanding or client categorization. Other tasks where ML is useful are the ones where adaptativity to changes in the environment is a necessary requirement, like for instance time series forecasting, handwritten text decoding or spam detection. In its most accepted definition:

> "Machine Learning is the study of algorithms that automatically improve by means of experience" [Mitchell, 1997].

In this definition, *learning* is intended as *improving by means of experience*. Even though the term "learning" can have several meanings and interpretations, we believe that "improving by means of experience" is one of the most intuitive and close to our everyday experience. For instance, it includes the idea of "trial and error", that is very often implemented by many living beings when they are about to learn how to solve new tasks: learning often implies numerous consecutive attemps (or trials) of solving the task. If a trial gives a positive result, it will be rewarded by similar future trials; on the other hand, if a trial gives a negative result, it is customary to identify it with an erroneous behaviour, and thus not repeat it in the future attempts. Iterating the process, the trials should become more and more effective with time, until the task gets solved. A simple example consists in the method rats use to select food: when rats encounter food items with new look and smell, they will first eat a small amount of it. According to the flavour and the physiological effect of the food, the rats will later decide if eating more or not. If the food produces an ill effect, that food will be associated with illness, and not eaten again. If it tastes good and does not produce any negative effect on the health of the rat, it will probably be eaten again. Also human beings use trial and error several times to learn tasks. For instance, when a person is learning how to play tennis, she will probably try to hit the ball by performing particular movements of the arms, shoulders and legs. Those movements will be identified as effective of erroneous, according to the result of the shot, and this result will affect the next attempts to hit the ball. As a last example of how much trial and error is used by humans for learning new tasks, the students that have recently attended a course of introduction to programming should agree on how many wrong attempts, with subsequent mistake identifications and adaptions, were needed before becoming able to write correct computer programs.

This process is what has inspired the introduction of the field of ML. But what do we exactly want machines to learn? Even though it is impossible to give general definitions to cover such a vast field as ML, we believe that we can cover the large majority of the situations saying that one of the most frequent objectives of ML is the one of *learning a function*. In large part of the situations, ML is dealing with a problem that can be defined as follows. Given a set of data pairs:

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$$

the objective is to find (or approximate) a function (or relation) ϕ, such that:

$$\forall i = 1, 2, ..., n : \ \phi(\mathbf{x}_i) = y_i$$

In the most general definition, $\mathbf{x}_i$ and y_i can be *any* kind of object (numbers, vectors, matrices, expressions, images, movies, sentences, other objects from the real world, etc.), however the most typical situation is the one in which the $\mathbf{x}_i$ are m-dimensional vectors of objects of any type (including, but not necessaily, numbers), while the y_i are scalar values.

Before having a closer look at the problem of learning, it is useful to fix some terminology:

- D is called dataset;
- $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ are called input data, input vectors, instances or observations;
- $\{y_1, y_2, ..., y_n\}$ are called expected outputs, or target values;
- the sought for function ϕ, i.e. the function that perfectly matches all possible data in the input domain into the corresponding targets, is called target function;
- learning is a process that allows us to obtain a function f that approximates the target function ϕ;
- function f, i.e. the function obtained as a result of the learning process, is called data model, or simply model;
- finally, we will talk of supervised learning in case the target values $\{y_1, y_2, ..., y_n\}$ are known for each observation, and unsupervised learning otherwise.

Last but not least, we will say that model f has a good *generalization ability* if f behaves like the target function ϕ also for data that do not belong to D.

In this manuscript, we will use both supervised learning and unsupervised learning algorithms. These strategies will be used, respectively, to generate predictive and descriptive models for credit use and over-indebtness in Portugal. In both approaches, the objective will be partitioning the observations (users) into groups. As mentioned earlier, in supervised learning the target values $y_1, y_2, ..., y_n$ are known and the objective is finding a function that matches the input data into those values. On the other hand, in unsupervised learning no target value is known, and thus the objective of the partitioning is typically done on the basis of the mutual similarities between objects, in such a way that similar data are categorized in the same group, while different data are categorized in different groups. These two approaches differ between each other also in the way new data are treated: in case of supervised learning, the learned model f is applied to the new data, and the result is interpreted as a prediction (supervised learning is usually employed to generate *predictive models*). On the other hand, when the learning is unsupervised, the new datum is inserted in the group that contains the elements that are more similar to it (unsupervised learning is usually employed to generate *descriptive models*). These two approaches are so different between each other that different names exist to indicate them: when the learning is supervised, the problem of partitioning data into groups is called *classification*, while when it is unsupervised it is called *clustering*.

The classification algorithm chosen in this manuscript is Support Vector Machines, while the clustering algorithm is Self-Organizing maps. These algorithms are described in the continuation of this chapter.

3.1 Features

A feature is a characteristic of the objects that have to be classified or, more generally, for which a prediction is needed. So, datasets are usually a collection of values (or instances) of features. So, given a dataset of the form:

$$D = \left[\begin{array}{cccc|c} x_{11} & x_{12} & \dots & x_{1m} & y_1 \\ x_{21} & x_{22} & \dots & x_{2m} & y_2 \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} & y_n \end{array}\right]$$

We normally use the following termnilogy:

- For each $j = 1, 2, ..., m$, column $\{x_{1j}, x_{2j}, ..., x_{nj}\}$ represents a feature, and for each $i = 1, 2, ..., n$, element x_{ij} is a feature value, or instance.
- For each $i = 1, 2, ..., n$, line $\{x_{i1}, x_{i2}, ..., x_{im}\}$ is a dataset instance, observation or sample, and y_i is the corresponding target value.

In case of classification, features are appropriate or useful if they allow us to make a difference between one class (or more) and the others. This is why an appropriate choice of the features is often crucial in supervised ML. Let us consider, for instance, the following toy dataset, whose objective is to classifiy animals into roosters and dogs:

	# paws	# eyes	having a crest	body fat	blood pressure	target
animal 1	2	2	True	7%	97	Rooster
animal 2	4	2	False	18%	118	Dog
animal 3	4	2	False	22%	126	Dog
animal 4	2	2	True	10%	101	Rooster

This dataset contains four observations, each one representing a different animal. Each animal is represented by five features: *number of paws* and *number of eyes*, which are integer numbers, *having or not having the crest*, a Boolean value, *body fay rate*, which is a percentage, and *blood pressure*, which is a floating point number. Observing this dataset, we can immediately notice that:

- Number of paws and having/not having the crest are *good* features: they clearly allows us to tell dogs from roosters.
- number of eyes is a totally useless feature: its value is the same for both classes, and so the feature is constant in the whole dataset.
- Body fat rate and blood pressure may help to make the classification, but if we use these two features the classification may be harder than if we simply use one among number of paws and having/not having the crest.

Examples of models that allow us to make a perfect classification for each instance in the dataset are:

if (having crest) **then** Rooster **else** Dog

or:

if (number of paws == 2)
  then Rooster
  else if (number of paws == 4)
    then Dog

Both these models use a restricted number of features, compared to the total number of features that appear in th dataset. Removing several features from the dataset, possibly leaving only number of paws and/or having/not having a crest, may significantly help the work of a classifier. The presence of useless features, or of features which make the classification harder, in fact, increments the search space and makes the model's optimization harder.

Feature selection is the process of choosing the features that are useful to make the prediction, disregarding all the others. It is often a very hard and complex task, and in can, in principle, be based on previous knowledge of the problem, or on mathematical relationships between data. *Feature extraction* is the process of combining one or more existing features to create a smaller number of more insightful features. Contrarily to feature selection, in feature extraction features are typically not chosen or disregarded, but only combined. Feature selection and feature extraction can both be used, or only one of them can be used. Reducing the dimensionality of the feature space can be a crucial task to improve the generalization ability of a ML system, so choosing or creating the appropriate features is a fundamental step from which the performance of the whole system can depend. They are usually applied before beginning the learning process, and for this reason, they are usually integrated in a so called *data preprocessing* phase, a phase that usually contains also a step of data cleaning, aimed at removing mistakes, imperfection or noise from the data.

Modern datasets have hundreds to tens of thousands of variables or features. Feature selection and feature extraction have three main objectives:

- improving the prediction performance of the models,
- providing faster and more cost-effective predictors, and
- providing a better understanding of the underlying process that generated the data.

Besides this, there are many other potential benefits of feature selection/extraction: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, etc. Methods for feature selection can essentially be partitioned into:

- Filters;
- Wrappers;
- Embedded methods.

Wrappers utilize the learning machine of interest as a black box to score subsets of variable according to their predictive power. Filters select subsets of variables independently of the chosen predictor. Embedded methods perform variable selection in the process of training and are usually specific to given learning machines.

The most popular kinds of filters (although by far not the only ones known) are:

- Correlation based methods;
- Information Theory based methods;

Both these methods have the objective of ranking the features according to their "usefulnes" in helping prediction, so that only the k top-ranked ones can be used for generating the predictive model. The intuition is that if a feature is independent from the target, it is uninformative for predicting it. Of course, these methods introduce a new parameter k, that can have a crucial influence on the performance of the system, and that can only be set by means of experimental comparisons.

The idea of correlation-based feature selection is simple: calculate the correlation between all features and the target, and then rank the features according to this correlation value. One of the most known measures is the Pearson correlation coefficient. For a particular feature, given the vector of all the feature values $\mathbf{x} = x_1, x_2, ..., x_n$ and the vector of the target values $\mathbf{y} = y_1, y_2, ..., y_n$, the Pearson correlation between X and Y is:

$$Corr = \frac{cov(\mathbf{x}, \mathbf{y})}{\sqrt{var(\mathbf{x})\ var(\mathbf{y})}}$$

where *cov* is the covariance of two vectors and *var* is the variance of one vector, so:

$$Corr = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ is the average of the elements of vector $\mathbf{x}$. By definition, $Corr$ is a value in $[-1, 1]$. Usually, the measure that is used to perform the ranking is $Corr^2$, because a negative correlation can be useful (it is enough to consider the feature with a negative sign in the model). One possible drawback of correlation criteria is that they can only detect linear dependencies between features and target. A simple way of lifting this restriction is to make a non-linear fit of the target with single variables and rank according to the goodness of that fit.

Concerning information theory-based feature selection, the ranking of features is done using mutual information between features and the target:

$$Inf = \sum_{x_i}\sum_{y_i} P(X = x_i, Y = y_i) \cdot log\ \frac{P(X = x_i, Y = y_i)}{P(X = x_i) \cdot P(Y = y_i)}$$

This measure is appropriate in case the features are discrete variables. The case of continuous variables (and possibly continuous targets) is harder and one can consider discretizing the variables.

Besides correlation and information theory, another possible measures to rank the features is the χ^2 between features and targets, which also aims at quantifying the depencence between features and target.

One common criticism of variable ranking is that it may lead to the selection of a redundant subset. The same performance could possibly be achieved with a smaller subset of complementary variables. Still, one may wonder whether adding presumably redundant variables can result in a performance gain. Actually, it is an experimental fact that, in classification, better class separation may be obtained by adding variables that are presumably redundant. More precisely perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them; but very high variable correlation (or anti-correlation) does not mean absence of variable complementarity. Furthermore, experimental evidence tells us that a variable that is completely useless by itself can provide a significant performance improvement when taken with others, and two variables that are useless by themselves can be useful together. These two last observations lead the scientific community to the idea that filters can have important limitations, and they can be overcome by means, for instance, of wrappers or the use of embedded methods.

The ML process for predictive models can be represented as in Figure 3.1. As we

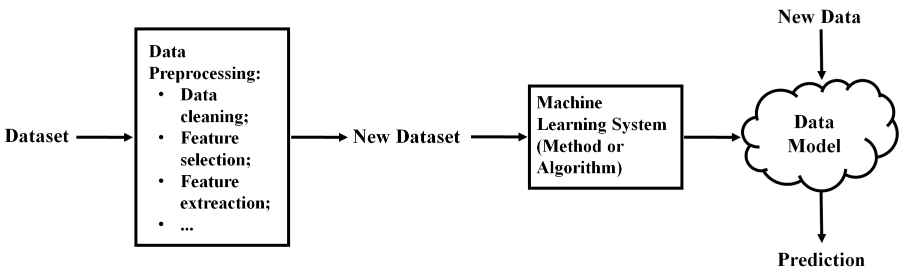


Fig. 3.1 The Machine Learning process for the generation of predictive models.

can see, the objective of data preprocessing is usually the one of generating a new dataset, that is generally smaller and possibly more informative, than the original one. This step is often crucial to facilitate the work to the ML system and often allows us to generated better models. When data preprocessing terminates, one is supposed to choose a ML algorithm to train the model.

3.2 Clustering by Unsupervised Neural Networks

Artificial Neural Network (ANNs) are computational methods that belong to the field of Machine Learning [Mitchell, 1997, Kelleher et al., 2015, Gabriel, 2016]. The aim of ANNs is to implement a very simplified model of the human brain. In this way, ANNs try to learn tasks (to solve problems) mimicking the behavior of the brain. The brain is composed by a large set of specialized cells, called *neurons*. Each single neuron is, in itself, a very simple entity, and the power of the brain is given by the fact that neurons are numerous and strongly interconnected between them, by means of connections called *synapsis*. The brain learns because neurons are able to communicate between each other. Biological neurons can receive stimuli and, as a consequence, emit (electric) signals, that can stimulate other neurons. When a biological neuron emits its signal, we say that it “fires”. In analogy with the human brain, ANNs are computational methods that use a large set of elementary computational units, called themselves (artificial) neurons. ANNs due their power to the numerous interconnections between neurons. Each neuron is able to only perform very simple tasks and ANNs are able to perform complex calculations because they are typically composed by many artificial neurons, strongly interconnected between each other and communicating with each other. In this section, we present ANNs for unsupervised learning, and specifically for clustering tasks. Several ANNs use unsupervised learning. In this document, we study:

- Competitive Learning Neural Networks (CLNNs);
- Kohonen Neural Networks (or Self Organizing Maps – SOMs).

Given that SOMs can be seen as an extension of CLNNs, our analysis begins with a presentation of CLNNs. For deepening both these typse of ANNs, the reader is referred to the book [Kohonen et al., 2001].

3.2.1 Competitive Learning Neural Networks

The architecture of CLNNs is shown in Figure 3.2, where the number of inputs n is

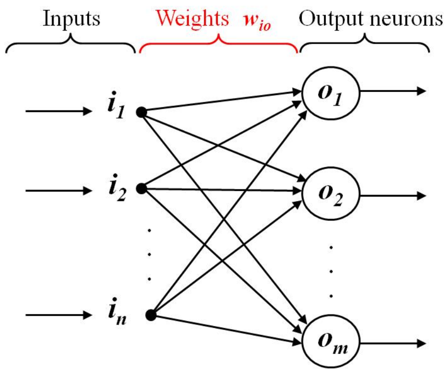


Fig. 3.2 The architecture of a Competitive Learning Neural Network (CLNN).

equal to the cardinality of the input vectors in the training set (i.e. number of input variables, or features), while the number of output neurons m is equal to the number of clusters in which we want to group data, that has to be known *a priori*. The functioning of a CLNN is very simple: whenever an input vector $\mathbf{x}$ is presented to the network, the network elects a *winner* among the output neurons. The network works in such a way that input vectors $\mathbf{x}$ belonging to the same cluster will produce the same output neuron as winner. The winner output neuron is decided in the following way: for each output neuron o, let $\mathbf{w}_o$ be the vector of the weights of the connections that enter in o. For a given input vector $\mathbf{x}$, the output neuron k that is defined as the winner is the output neuron for which the distance (typically the Euclidean distance) of the weight vector $\mathbf{w}_k$ to the input vector $\mathbf{x}$ is minimal. In other words, an output neuron k is chosen as the winner if, for all output neurons o:

$$\|\mathbf{w}_k - \mathbf{x}\| \leq \|\mathbf{w}_o - \mathbf{x}\|$$

As it is customary for any ANN, also for CLNNs the learning phase corresponds to a phase in which the weights of the synapsis are modified. In particular, once the winner k has been chosen, only the weights of the connections entering into k are modified, and they are modified in such a way that they get *even closer* to the input vector $\mathbf{x}$:

$$\mathbf{w}_k(t+1) = \mathbf{w}_k(t) + \eta(\mathbf{x}(t) - \mathbf{w}_k(t))$$

where η is a learning rate constant similar to the one used by the other types of Neural Networks that we have studied so far. In simple words, the neuron whose weights are the closer to the input vector is chosen as winner, and its weights are further shifted towards the input vector itself. CLNNs are one of the few ANNs for which a random initialization of the weights is unconvenient. Similarly to what we have already seen for other clustering algorithms, it is usually more convenient to select a subset of the input vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$, where M is the number of output neurons and, for each $i = 1, 2, ..., M$, we assign $\mathbf{w}_i$ to the vector $\mathbf{x}_i$. This method, however, may have problems in case a high number of input vectors belonging to the same cluster exist. In general, it is suitable to chose the M input vectors with the maximum diversity between each other for the initialization of the weights.

Figure 3.3 shows on a Cartesian plane (for simplicity we use a 2-dimension plane) the typical situation after the learning of a CLNN: at the end of the learning phase the weight vectors (represented by colored shapes in the figure) should approximate the centroids of the various clusters (the training set points are represented in non-colored shapes in the figure).

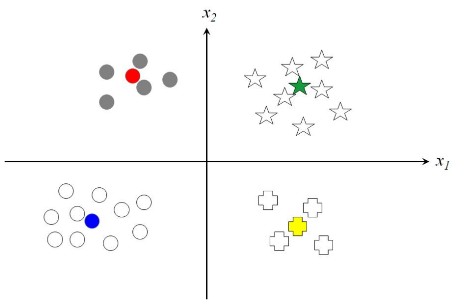


Fig. 3.3 A visualization of the typical situation after the learning of a CLNN: the weight vectors (represented as colored shapes here) approximate the centroids of the clusters.

3.2.2 *Self Organizing Maps*

Kohonen ANNs, or Self Organizing Maps (SOMs), can be seen as an extension of the CLNNs studied so far. The main difference between SOMs and CLNNs is that in SOMs the output neurons are organized in a *topological stucture*, that is often a 2-dimensional grid. In this way, it is possible to identify a neighborhood for the output neurons. The architecture of a SOM is shown in Figure 3.4.

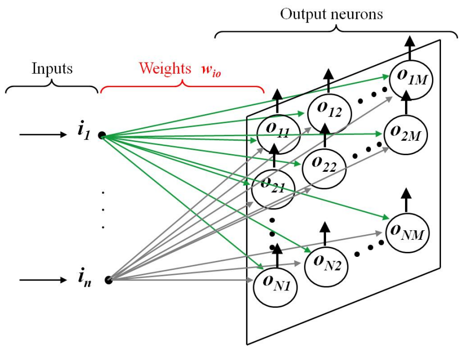


Fig. 3.4 The architecture of a Self Organizing Map (SOM).

A SOM works in such a way that input vectors that are "close to each other" (in the sense of the Euclidean distance) will chose as winners output neurons that are "close to each other" (in the output neurons topology). Given an input vector, a winner output neuron is decided exactly as for CLNNs: the winner is the output neuron for which the distance between the vector of the weights of the entering connections and the input vector is minimal. The difference between SOMs and CLNNs is in the way weights are updated: in the SOMs not only the weights of the connections entering in the winner neuron are updated, but also the weights of all the other connections, and the *strenght* of the modification is as large as the output neuron is near to the winner neuron in the output topological stucture. The formula used by SOMs for updating weights is:

$$\mathbf{w}_o(t+1) = \mathbf{w}_o(t) + \ \eta \ \ g(o,k) \ \ (\mathbf{x}(t) - \mathbf{w}_o(t))$$

where o is any output neuron, k is the winner neuron and $g(o,k)$ is a function of the distance of neuron k to neuron o in the topological output structure. Function g must have the following properties:

- If the distance between o and k is equal to zero (i.e. $o = k$), then the value of $g(o,k)$ must be equal to 1.
- If the distance between o and k is "large", then $g(o,k)$ must be "small" (almost equal to 0).
- g must be monotonically decreasing.

In other words. if a neuron o is "far" from the winner neuron k in the topological output structure, then its weights do not have to be updated, or have to be updated only in a minimal amount. Neuron weights have to be modified more and more strongly as the neurons are nearer and nearer to the winner neuron in the output topology. An example of a g function is, for instance, the following Gaussian function:

$$g(o,k) = e^{-d(o,k)}$$

where $d(o,k)$ is the (Euclidean) distance between neuron o and neuron k in the output topology.

Besides clustering, SOMs are often used for transformations ("mapping") of N-dimensional data into 2-dimensional data, preserving topological properties (useful for instance in data visualization).

3.3 Support Vector Machines

3.3.1 Binary Classification, Linearly Separable Problems

Let us define the binary classification problem as the task of separating a set of *positive* samples (instances labelled with $+$) from a set of *negative* samples (instances labelled with $-$), and let us assume that the problem is linearly separable. The situation is represented in Figure 3.5(a) for the elemenary case of bidimensional data. Observing the figure, it is clear that several separating straight lines exist. Assuming that you were allowed to choose one to perform the classification, which one would you choose? It is intuitive that it is not such a good idea to have a straight line that is too close to positive or negative examples. In fact, data might be affected to noise, so the position of points may be unprecise and thus a line that is too close to a point may not be accurate. Futhermore, unseen points are generally different from the training ones, and a straight line that is too close to training points may fail to correctly classify unseen ones. On the other hand, a straight line like the dashed one pictured in Figure 3.5(a), that is neither too close to the positive examples nor to the negative ones, looks like a more appropriate choice. Informally, we could say that the best line may be the one that has the property of standing in the middle of the "widest street" that separates the negative examples from the positive examples. The distance from the decision surface to the closest data point determines what we

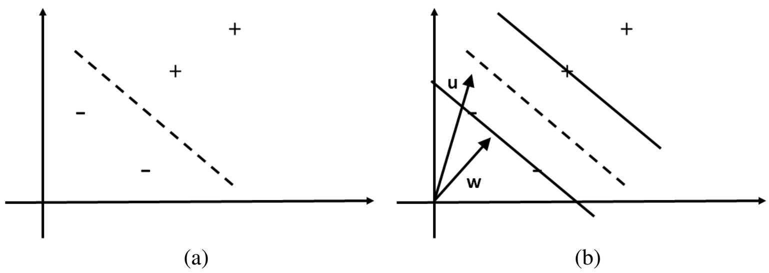


Fig. 3.5 A simple linearly separable, binary classification problem in two dimensions.

call the *margin*, or *gutter*, of the classifier. Thus, the best line is the one that maximizes the distance between the line itself and the margins. Margins are shown with solid lines in Figure 3.5(b). The approach of SVMs is informally called the "widest street" approach, because the three parallel straight lines in Figure 3.5(b) vaguely reminds a street. This method of construction necessarily means that the decision function for a SVM is fully specified by a (usually small) subset of the data, which defines the position of the separator. These points are referred to as the *support vectors* (the points intercepting the margins in Figure 3.5(b)). Other data points play no part in determining the decision surface that is chosen.

Now imagine that we have a vector (like **w** in Figure 3.5(b)) of any length, constrained to be perpendicular to the median line of our street. Imagine that we also have some unknown and a vector (like **u** in the figure) pointing to it. We want to understand whether that unknown is on the right side of the street or on the left side of the street. What we have to do to answer the question is to project that vector **u** down to **w**, that is perpendicular to the street. Remembering that the projection of a vector along side another can be expressed in function of the "dot product" of the two vectors, we could now measure whether $\mathbf{w} \cdot \mathbf{u}$ is greater or equal than some given constant, or, without lost of generality, for a given constant b:

$$\mathbf{w} \cdot \mathbf{u} + b \geq 0$$

If this expression is true, then our unknown point is a positive example, otherwise it is a negative example. So, this is the shape of our decision rule. The problem is that we do not know what constant b and what vector **w** to use. All we know is that **w** has to be perpendicular to the median line of the street, but so far it can be of any length. So, we do not have enough constraints to fix a particular b or a particular **w** yet. What we need is some constraint in such a way that we are actually able to calculate a **w** and a b. One assumption we could do is that if we take that vector **w** and we multiply it for a training positive example $\mathbf{x}_+$, then this product has to be larger or equal than 1:

$$\mathbf{w} \cdot \mathbf{x}_+ + b \leq 1 \tag{3.1}$$

In other words, an unknown point can be anywhere in the street, but if you are a positive sample, then we are going to assume that the decision function returns a value that is equal or greater than 1. Likewise, for a negative example $\mathbf{x}_-$, we will assume that our decision function will have to return a value that is equal to or less than -1:

$$\mathbf{w} \cdot \mathbf{x}_- + b \leq -1 \tag{3.2}$$

In other words, we are imposing a separation of -1 to +1 for all of the training samples. So, now we could just try to solve the system composed by Equations (3.1) and (3.1), in order to find a **w** and a b that guarantees this kind of separation for all the training samples.

In order to make the system more manageable, we now introduce a variable y_i such that $y_i = +1$ for positive samples and $y_i = -1$ for negative samples. Multiplying both previous equations by y_i brings the two equations to be the same, in fact we obtain:

$$y_i(\mathbf{w} \cdot \mathbf{x}_+ + b) \geq 1$$
$$yi(\mathbf{w} \cdot \mathbf{x}_- + b) \geq 1$$

So, now we can say that, for any sample training sample $\mathbf{x_i}$, being it positive or negative, we have:

$$y_i(\mathbf{w} \cdot \mathbf{x_i} + b) - 1 \geq 0 \tag{3.3}$$

Now, we can imagine that we make one more assumption, i.e. for every $\mathbf{x_i}$ in the gutter:

$$y_i(\mathbf{w} \cdot \mathbf{x_i} + b) - 1 = 0 \tag{3.4}$$

In other words, the result of this equation will be exactly equal to zero for the points circled in red in Figure 3.6(a), and greater than zero for the points circled in blue in the same figure. The points on the gutter, the ones circled in red, are what we call the

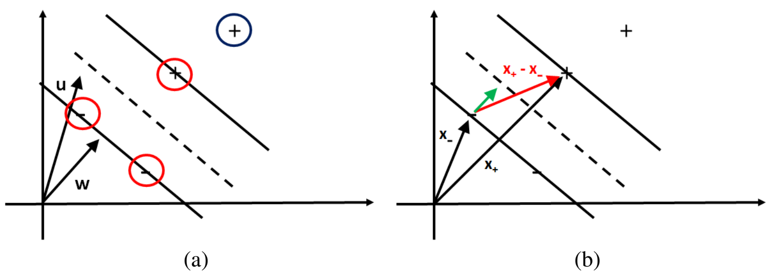


Fig. 3.6 Support vectors and separations, for the same problem as in Figure 3.5.

support vectors, and are the only training points we use to build our classifier. Our objective is to maximize the distance between the two gutters of the street, so, let us try to express that distance. For doing this, let us consider a positive example $\mathbf{x}_+$ and a negative example $\mathbf{x}_-$, and let us consider the difference between vectors $\mathbf{x}_+$ and $\mathbf{x}_-$. Vectors $\mathbf{x}_+$ and $\mathbf{x}_-$ are shown in black, while their difference is shown in red in Figure 3.6(b). If we only had a unit vector normal to the median line of the street, like the one shown in green in Figure 3.6(b), then we could consider the "dot-product" (i.e. the scalar product) between that unit vector and the difference between $\mathbf{x}_+$ and $\mathbf{x}_-$, and that would be the length of the street. But, one of our first assumptions was that $\mathbf{w}$ is normal to the median line of the street. So, we could say that the width of the street is equal to:

$$width = (\mathbf{x}_+ - \mathbf{x}_-) \cdot \frac{\mathbf{w}}{||\mathbf{w}||} \tag{3.5}$$

In fact, $\frac{\mathbf{w}}{||\mathbf{w}||}$ is a unit vector and, given that $\mathbf{w}$ is normal to the median line of the street, also $\frac{\mathbf{w}}{||\mathbf{w}||}$ is normal to the median line of the street.

Now, from Equation (3.4), that constraints the samples in the line of the gutter, we have that, for positive samples, $y_i = 1$, and thus:

$$\mathbf{w} \cdot \mathbf{x}_+ = 1 - b$$

So, from the previous observation, and doing an analogous reasoning for a negative example that stands in the gutter, we can develop Equation (3.5), obtaining:

$$width = (\mathbf{x}_+ - \mathbf{x}_-) \cdot \frac{\mathbf{w}}{||\mathbf{w}||} = \frac{\mathbf{x}_+ \cdot \mathbf{w} - \mathbf{x}_- \cdot \mathbf{w}}{||\mathbf{w}||} = \frac{1 - b + 1 + b}{||\mathbf{w}||}$$

This allows us to rewrite the Equation (3.5) as:

$$width = \frac{2}{||\mathbf{w}||} \tag{3.6}$$

In other words, the width of the street is maximized if $\frac{||\mathbf{w}||}{2}$ is minimized, or, in other words, if $||\mathbf{w}||$ is minimized. The reader is invited to notice the analogy between this minimization and regularization methods. Maximizing the width of the street, we are implicitly applying a regularization method to the coefficients of the decision rule. This is a further argument in favour of the selection of the maximum separating line, to foster generalization.

Joining Equation (3.4) with Equation (3.6), we are now able to express the final standard formulation of an SVM as a minimization problem:

Find $\mathbf{w}$ and b, such that:

- $\frac{1}{2}\,\mathbf{w}^T \cdot \mathbf{w}$ is minimized,
- under the constraint: $y_i(\mathbf{w}^T \cdot \mathbf{x_i} + b) - 1 = 0$.

So, the problem is now reduced to the optimization of a quadratic function, subject to linear constraints. Quadratic optimization problems are a standard, well-known class of mathematical optimization problems, and many algorithms exist for solving them. We could in principle build a SVM using standard quadratic programming, but much recent research was devoted to studying the particular structure of the kind of quadratic problem that emerges from a SVM. As a result, there are more complex, but much faster and more scalable, methods for building SVMs. The details of the mathematical derivation of such methods is beyond the scope of the book. However, it is interesting to understand the shape of the solution of such an optimization problem. The solution involves constructing a dual problem, where a Lagrange multiplier α_i is associated with each constraint $y_i(\mathbf{w}^T \cdot \mathbf{x_i} + b) \leq 1$ in the primal problem:

Find $\alpha_1, \alpha_2, ..., \alpha_N$, such that:

- $\sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x_i}^T \cdot \mathbf{x_j}$ is maximized,
- under the constraint: $\sum_i \alpha_i y_i = 0$,
- and: $\alpha_i \geq 0$ for all $i = 1, 2, ..., N$.

The solution is of the form:

- $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x_i}$
- $b = y_k - \mathbf{w}^T \cdot \mathbf{x_k}$, for any $\mathbf{x_k}$ such that $\alpha_k \neq 0$

In the solution, most of the α_i are equal to zero. Each non-zero α_i indicates that the corresponding $\mathbf{x_i}$ is a support vector. The classification function is then:

$$f(\mathbf{x}) = \text{ sign } (\sum_i \alpha_i y_i \mathbf{x_i}^T \cdot \mathbf{x} + b) \tag{3.7}$$

As we can see, the most expansive calculations are the dot products between the unseen point $\mathbf{x}$ and the support vectors $\mathbf{x_i}$. This makes SVM particularly efficient in terms of computational speed, compared to several other Machine Learning systems. In the continuation, we see a numerica example for a very simple case study.

3.3.2 Non-Linearly Separable Problems: the Kernel Trick

So far, we have studied SVMs for linearly separable problems, possibly with a few exceptions or some noise. In order to understand the idea that is at the base of the use of SVMs for non-linearly separable problems, let us consider an example like the one represented in Figure 3.7(a). Even though the points can be represented along just one dimension, the reader can agree that no linear function can separate the positive examples from the negative examples. One way to solve this problem is to map the data on to a higher dimensional space, and then to use a linear classifier in the higher dimensional space. For example, Figure 3.7(b) shows that a linear separator can classify the data if we use a quadratic function to map the data into two dimensions (and several other types of transformations could be imagined, besides the quadratic one). The general idea is to map the original feature space to some higher-dimensional feature space, where the training set is linearly separable. Of course, this needs to be done in a way that preserves relevant dimensions of relatedness between data points, so that the resultant classifier should still have generalization ability. SVMs provide an efficient way of doing this mapping to a higher dimensional space, which is referred to as *the kernel trick*. As we have studied so far, the SVM linear classifier relies on a dot product between data point vectors. Let $K(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i}^T \mathbf{x_j}$. Then the classifier we have seen so far is:

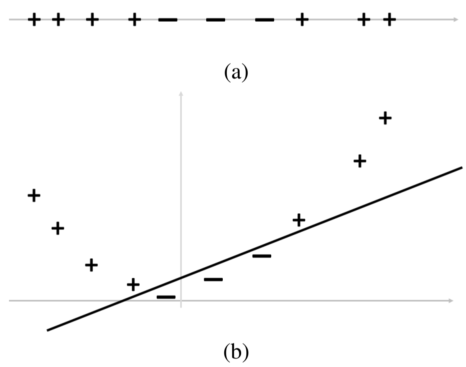


Fig. 3.7 (a): representation of the points for a non-linearly separable problem; (b): projections of the points on a higher-dimensional feature space, by means of a quadratic function. The mapped problem is now linearly separable.

$$f(\mathbf{x}) = \text{ sign } (\sum_i \alpha_i \, y_i \, K(\mathbf{x_i}, \mathbf{x}) \, + \, b) \tag{3.8}$$

Now, assume we map every data point into a higher dimensional space via some transformation $\Phi : \mathbf{x}' \rightarrow \phi(\mathbf{x}')$. Then the dot product becomes $\phi(\mathbf{x_i})^T \phi(\mathbf{x_j})$. If it turned out that this dot product could be computed simply and efficiently in terms of the original data points, then we would not have to actually apply the mapping Φ. Rather, we could simply compute the quantity $K(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i})^T \phi(\mathbf{x_j})$, and then use the function's value in Equation (3.8). A kernel function K is such a function, that corresponds to a dot product in some expanded feature space.

Even though several types of kernels can be defined, the vast majority of work with kernels uses one of two families of functions of two vectors, which we define below, and which define valid kernels. The two commonly used families of kernels are polynomial kernels and radial basis functions. Polynomial kernels are of the form:

$$K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}^T \mathbf{z})^d$$

The case of $d = 1$ is a linear kernel, which corresponds to SVMs, as they were studied in the previous sections (with the only, minor, difference that constant 1 changes the threshold). The case of $d = 2$ gives a quadratic kernel, and is very commonly used. The most common form of radial basis function is a Gaussian distribution, calculated as:

$$K(\mathbf{x}, \mathbf{z}) = e^{-(\mathbf{x}-\mathbf{z})^2/(2\sigma^2)}$$

Beyond these two families, there has been interesting work developing other kernels, some of which gave promising experimental results. Interestingly, other Machine

Learning algorithms, like for instance Genetic Programming, can also be used to automatically generate appropriate kernels for SVMs [Sullivan and Luke, 2007].

Chapter 4
Employed Data and Obtained Results

4.1 Data

In this manuscript, we present experiments obtained using data from consumers that have over-indebtness problems and exploit public assistance provided by the Portuguese Association for the Consumer Defens (DECO Portugal). The dataset contains information from 1,654 Portuguese consumers, who have taken advantage of the assistance of the debt advisory services in Portugal in 2016 and 2017, because they were experiencing over-indebtness issues and had, generally speaking, a significant risk of poverty (they were 802 in 2016 and 852 in 2017). The usual requests of these individuals when they enter in contact with advisory services concern advices on on how to organize their family budget, how to manage their debts with the credit holders (e.g., bank, insurance companies, stores), or which credits should they pay first. The experts from the advisory services can, in some cases, even suggest if there are some goods that these citizens should sell, and in this case which ones. Those goods usually range from simple consumption goods (e.g., mobile phone, computer) to important long-term goods, such as cars or houses. The dataset we have employed containes a vast set of features that offer a rather complete information regarding the consumers' financial situation, including, but not limited to family socio-demographics, total income, total expenses, employment information, as well as all credit details. In the work presented in the continuation of this chapter, we considered variables such as socio-demographic characterization (marital status, level of education completed, number of people in the household), an opinion of the consumers concerning the causes for their own over-indebtedness, and information about their economic state, such as for instance the total income and expenses or information about their credits and debts (amount of monthly use of credit cards, housing credit or car credit, if applicable, personal credit and other types of credit or debts, etc.). Each consumer is represented by one observation of the dataset.

4.2 Experimental Results

4.2.1 Descriptive Modelling

The descriptive model that we generated used the Kohonen R Package (R Studio), with the following parameter configuration: $rlen = 3,000$ iterations; $\alpha = 0.05$; topo = hexagonal; and grid size = 100 cells (10×10). After 3,000 iterations, the mean distance between the observations of each node was reduced to 0.015. As final outcome, SOM training extracted three well separated clusters, with clearly distinguishable characteristics: low-income households (n = 490, 31.27%), low credit control households (n = 586, 37.40%) and crisis-affected households (n = 491, 31.33%).

Concerning Cluster 1 (Low-income households), 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).

If we analyze Cluster 2 (Low credit control households), we can observe that 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).

Finally, we observe that Cluster 3 (Crisis-affected households) 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.

4.2.2 Predictive Modelling

In our experimental study aimed at generating preductive models for over-indebtment, several different classifiers were compared between each other, after that their respective parameters were optimized independently. Those classifiers were: (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. To perform this comparison, a cross-validation was implemented 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 were 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. 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 4.1 shows a comparison of their performance, using Accuracy Score and Log Loss. The algorithm that generated the best model after this exhaustive search was a version of the Support Vector Machine algorithm, the Nu-SVC. Table 4.1 presents detailed Machine Learning algorithms comparative performance.

4.2.3 Discussion

Over-indebtness can be understood and predicted, in order to have more effective personalized interventions on the population, earlier. The descriptive and predictive models provided by ML are ideal for this kind of study, and AI is everyday more a candidate to become a new state of the art technologies that public and private organizations can use to support citizens. The availability of vast amounts of data

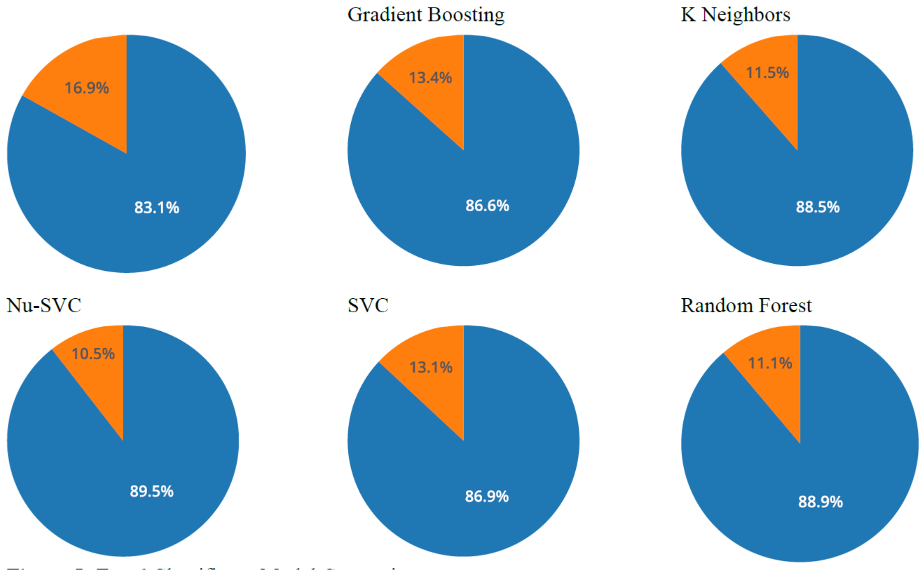


Fig. 4.1 Experimental comparison between the following supervised learning algorithms: SVC, Nu-SVC, Extra Trees, Random Forest, Gradient Boosting, and K Nearest Neighbors.

Table 4.1 Results returned by the best performing supervised learning algorithms.

	Mean accuracy	**Std accuracy**	**Mean f1_macro**	**Std f1_macro**	**Mean log loss**	**Std log loss**	**Mean precision**	**Std precision**	**Mean recall**	**Std recall**
Gaussian NB	0.646509	0.024092	0.608481	0.025551	2.975186	0.2064	0.717132	0.050065	0.671085	0.024783
K Neighbors	0.856223	0.018888	0.897674	0.019018	1.162856	0.701356	0.900295	0.018939	0.897983	0.01836
LDA	0.774867	0.027593	0.775904	0.030079	0.567435	0.062709	0.779909	0.029101	0.781957	0.025002
Log Reg	0.817216	0.010291	0.820368	0.01172	0.443736	0.026994	0.823359	0.01501	0.821844	0.0085
Decision Tree	0.725458	0.012966	0.7207	0.01981	0.60555	0.075023	0.750819	0.012217	0.736014	0.013784
Random Forest	0.865861	0.027188	0.8685	0.027157	0.418591	0.067184	0.868907	0.026631	0.869309	0.026868
Extra Trees	0.837118	0.025094	0.838138	0.026499	0.511755	0.02616	0.840766	0.024465	0.842403	0.022295
Gradient Boost.	0.836369	0.012675	0.83954	0.012855	0.493676	0.022844	0.842056	0.012014	0.83894	0.013125
Nu SVC	0.892248	0.014584	0.894278	0.014343	0.305083	0.056059	0.89562	0.015901	0.895422	0.01386
SVC	0.840369	0.026918	0.84071	0.025842	0.346097	0.047502	0.849353	0.027162	0.83909	0.025384

should pave the way for the development of a large amount of research, aimed at using AI to study and predict over-indebtness. The developed technologies and obtained results seem to point towards a future in which data science and ML will provide better information, so that it will be possible to make more informed decisions, while understanding the possible outcomes and costs. The value of ML stands in its ability to process vast amounts of data, beyond the scope of human capability, and then reliably convert them into clear and applicable insights.

Thanks to studies like the one presented in this document, it is nowadays clear that Ai is destined to play an important role in society in the future. In the form of ML, it is our most promising tool for the development of timely and effective interventions to counteract poverty and over-indebtment. The greatest challenge to AI, in partcular in this type of application, is not whether the technologies will be capable enough to be useful, but rather ensuring their adoption in daily practice. For these technologies to be widely adopted in public institutions, in fact, for instance AI systems and models must be approved by regulators, integrated within existing platforms, standardised to a sufficient degree that similar methods work in a similar fashion, taught to the main users, paid for by public or private organisations and updated over time, consistently with the research advancements in the field. These challenges will eventually be overcome, but they will take much longer than it will take for the technologies themselves to mature.

Chapter 5
Conclusions

5.1 Summary of Findings

Based on a categorization of causes of over-indebtedness as dispositional or situational, we identified three different over-indebtedness profiles. By doing so, this project uncovers the factors that might help explaining over-indebtedness among consumers who face severe economic austerity in Portugal. Our findings indicate that situational factors such as the crisis and other external causes explain the majority of over-indebted profiles, while dispositional causes related to consumerism and lack of self control are prevalent only in few of the cases.

The analysis of the three over-indebtedness profiles revealed some important disparities in terms of social-demographic distributions. One example is unemployment rate and family size, factors that affect credit consumption among Portuguese consumers.

Regarding credit consumption, some consumers tend to show more credit card consumption than others, both in number of credit cards and in total in debt. Given that there were no differences in monthly installments of credit card among profiles, this shows that some Portuguese consumers have a hard time and may take longer to eliminate debt. The same pattern is observed in household income and household expenses. This may contribute to explaining the significant difference found between profiles in the money available after regular expenses (income minus expenses).

Besides, a key difference between the clusters is the amount of credit usage, in some cases the average amount of credit is almost 200 % higher than the other profiles. This can indicate that an important part of the Portuguese population cannot control their credit consumption, and consequently fall into over-indebtedness.

Regarding the socio-demographic information, marital status for over-indebted consumers differs from what is referred in Canner and Luckett (1991), where households of divorced and separated consumers were more likely to report debt payment difficulties. In the data gathered here divorced/separated consumers were relatively few compared to married and even single consumers. However, the results pattern obtained here is comparable to the over-indebted profile gathered by a CES report (Frade, Lopes, Jesus, Ferreira, 2008).

An overall comparison of the three over-indebtedness profiles, indicates that some clusters raise from situational factors. The current profiling allowed us to not only distinguish between consumers in dimensions other than the reason for their over-indebted situation but also to make a rough estimation of social propensity of these profiles by looking at the frequency of households classified in each profile. Some of the cases were affected by the crisis (48.0%) and other external causes (40.3%), accounting for the majority of over-indebted cases in our data, while the self-control and psychological causes only captured 11.7% of the cases. This marked difference in frequencies reveals that most over-indebted individuals from our sample present external and largely uncontrollable factors as the causes for their over-indebtedness.

5.2 Societal Implications

This is the first work to attempt a characterization of over-indebted individuals using an approach from psychology to machine learning. Three different consumer profiles and a differential characterization of each, emerged from our data. This suggests that over-indebted consumers are a heterogeneous group both in demographic and financial aspects as well as perceived causes for over-indebtedness. Such heterogeneity has important implications for the society.

First, it adds by characterizing over-indebtedness into several sub-groups. Such profiling of over-indebted consumers suggests that causes for over-indebtedness are indeed multi-categorical, with some affecting certain demographic groups more than others. As a result, the same psychological factors such as attitudes towards debt (e.g., Chien Devaney, 2001; Livingstone Lunt, 1992), time preferences (e.g., Groenland Nyhus, 1994), tendency to decide based on dysfunctional heuristics (e.g., Slowic, 2012), financial literacy (e.g., Lusardi Tufano, 2015), may play markedly different moderator roles depending on the over-indebtedness profile of the consumer.

Second, our results may be of importance for public policy at creating interventions to reduce over-indebtedness. Since different profiles are associated to different causes, tailored interventions to tackle each profile's specific causes might not only increase the adherence to these interventions among over-indebted households but also their efficacy.

Indeed, some consumer profiles fit well with the notions of impulsiveness and lack of self-regulation as causes for over-indebtedness. As such, interventions tailored to curb such behaviors (e.g., focusing on developing self-regulation skills) might be helpful. For the consumers under the crisis or external causes profiles, on the other hand, interventions more focused on enhancing consumers financial literacy might be more useful given their general lower education levels. Such consumers seem more likely to suffer from scarcity feelings, given their poorer and less favourable background. Thus, feelings of scarcity might kick in as well with consequences for self-regulation and implications for interventions.

Finally, given the self-report nature of our findings, it is possible some consumers perceive the causes for their over-indebtedness as more external than they really are. However, the high discrepancy in numbers suggests most over-indebted situations are indeed likely driven by external causes rather than “laisser-faire” consumption patterns.

The preliminary nature of the current study begs for further research on the identification and description of what seems to be qualitatively different profiles of over-indebtedness using machine learning. Similarly, whereas it seems likely that lack of education contributed to the crisis and external causes situation, it is up to test whether these groups would derive a differential benefit from interventions tailored at increasing financial literacy. If that proves to be the case, then our work and future research on over-indebtedness profiling may contribute to the development of best public practices and social interventions to empower consumers in general and over-indebted consumers in particular.

We hope this project contributes to the understanding of this complex phenomenon of over-indebtedness and its causal prevalence in Portugal. The high number of consumers falling under over-indebtedness paints a picture which we hope contributes to call attention to the vulnerable consumers who might be affected in future financial crisis, so government can take measures accordingly.

References

[Aizcorbe et al., 2003] Aizcorbe, A. M., Kennickell, A. B., and Moore, K. B. (2003). Recent changes in us family finances: Evidence from the 1998 and 2001 survey of consumer finances. *Fed. Res. Bull.*, 89:1.

[Alleweldt et al., 2013] Alleweldt, F., Kara, S., Graham, R., Kempson, E., Collard, S., Stamp, S., and 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. *DG Health and Consumers of the European Commission. Available online at: www. civic-consulting. de*.

[Angel et al., 2009] Angel, S., Einböck, M., and Heitzmann, K. (2009). Politik gegen und ausmaß der überschuldung in den ländern der europäischen union.

[Berthoud and Kempson, 1992] Berthoud, R. and Kempson, E. (1992). *Credit and debt: the PSI report*. Number 728. Policy Studies Institute.

[Betti et al., 2007] Betti, G., Dourmashkin, N., Rossi, M., and Yin, Y. P. (2007). Consumer over-indebtedness in the eu: measurement and characteristics. *Journal of Economic Studies*.

[Blomgren et al., 2016] Blomgren, J., Maunula, N., and Hiilamo, H. (2016). Over-indebtedness and chronic disease: a linked register-based study of finnish men and women during 1995–2010. *International journal of public health*, 61(5):535–544.

[Brown et al., 2005] Brown, S., Garino, G., Taylor, K., and Price, S. W. (2005). Debt and financial expectations: An individual-and household-level analysis. *Economic Inquiry*, 43(1):100–120.

[Buxton and Marcelli, 2010] Buxton, O. M. and Marcelli, E. (2010). Short and long sleep are positively associated with obesity, diabetes, hypertension, and cardiovascular disease among adults in the united states. *Social science & medicine*, 71(5):1027–1036.

[Cappuccio et al., 2010] Cappuccio, F. P., D'Elia, L., Strazzullo, P., and Miller, M. A. (2010). Sleep duration and all-cause mortality: a systematic review and meta-analysis of prospective studies. *Sleep*, 33(5):585–592.

[Chien and Devaney, 2001] Chien, Y.-W. and Devaney, S. A. (2001). The effects of credit attitude and socioeconomic factors on credit card and installment debt. *Journal of Consumer Affairs*, 35(1):162–179.

[Chmelar, 2013] Chmelar, A. (2013). Household debt and the european crisis.

[Clayton et al., 2015] Clayton, M., Liñares-Zegarra, J., and Wilson, J. O. (2015). Does debt affect health? cross country evidence on the debt-health nexus. *Social science & medicine*, 130:51–58.

[Feurer et al., 2019] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., and Hutter, F. (2019). Auto-sklearn: efficient and robust automated machine learning. In *Automated Machine Learning*, pages 113–134. Springer, Cham.

[Gabriel, 2016] Gabriel, J. (2016). *Artificial Intelligence: Artificial Intelligence for Humans*. CreateSpace Independent Publishing Platform, USA, 1st edition.

[Gallicchio and Kalesan, 2009] Gallicchio, L. and Kalesan, B. (2009). Sleep duration and mortality: a systematic review and meta-analysis. *Journal of sleep research*, 18(2):148–158.

[Gangwisch et al., 2006] Gangwisch, J. E., Heymsfield, S. B., Boden-Albala, B., Buijs, R. M., Kreier, F., Pickering, T. G., Rundle, A. G., Zammit, G. K., and Malaspina, D. (2006). Short sleep duration as a risk factor for hypertension: analyses of the first national health and nutrition examination survey. *hypertension*, 47(5):833–839.

[Garðarsdóttir and Dittmar, 2012] Garðarsdóttir, R. B. and Dittmar, H. (2012). The relationship of materialism to debt and financial well-being: The case of iceland's perceived prosperity. *Journal of Economic Psychology*, 33(3):471–481.

[Gathergood, 2012] Gathergood, J. (2012). Self-control, financial literacy and consumer over-indebtedness. *Journal of economic psychology*, 33(3):590–602.

[Grandner et al., 2014] Grandner, M. A., Chakravorty, S., Perlis, M. L., Oliver, L., and Gurubhagavatula, I. (2014). Habitual sleep duration associated with self-reported and objectively determined cardiometabolic risk factors. *Sleep medicine*, 15(1):42–50.

[Grandner et al., 2010] Grandner, M. A., Hale, L., Moore, M., and Patel, N. P. (2010). Mortality associated with short sleep duration: the evidence, the possible mechanisms, and the future. *Sleep medicine reviews*, 14(3):191–203.

[Grandner and Patel, 2009] Grandner, M. A. and Patel, N. P. (2009). From sleep duration to mortality: implications of meta-analysis and future directions. *Journal of sleep research*, 18(2):145.

[Groenland and Nyhus, 1994] Groenland, E. and Nyhus, E. (1994). Determinants of time preference, saving, and economic behavior: A comparison between norway and the netherlands. In *IAREP/SABE conference. Erasmus University Rotterdam, July*, volume 10.

[Guiso and Sodini, 2013] Guiso, L. and Sodini, P. (2013). Household finance: An emerging field. In *Handbook of the Economics of Finance*, volume 2, pages 1397–1532. Elsevier.

[Hall, 1978] Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. *Journal of political economy*, 86(6):971–987.

[Harvey, 2011] Harvey, D. (2011). The enigma of capital and the crisis this time. In *Business as Usual*, pages 89–112. New York University Press.

[Hintikka et al., 1998] Hintikka, J., Viinamäki, H., Tanskanen, A., Kontula, O., and Koskela, K. (1998). Suicidal ideation and parasuicide in the finnish general population. *Acta Psychiatrica Scandinavica*, 98(1):23–27.

[Kamleitner and Kirchler, 2007] Kamleitner, B. and Kirchler, E. (2007). Consumer credit use: A process model and literature review. *European Review of Applied Psychology*, 57(4):267–283.

[Kelleher et al., 2015] Kelleher, J. D., Namee, B. M., and D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. The MIT Press.

[Kempson, 2015] Kempson, E. (2015). Over-indebtedness and its causes across european countries. *Consumer debt and social exclusion in Europe*, pages 137–153.

[Kida et al., 2009] Kida, M. et al. (2009). Financial vulnerability of mortgage-indebted households in new zealand–evidence from the household economic survey. *Reserve Bank of New Zealand Bulletin*, 72(1):5–12.

[Kohonen et al., 2001] Kohonen, T., Schroeder, M. R., and Huang, T. S. (2001). *Self-Organizing Maps*. Springer-Verlag, Berlin, Heidelberg, 3rd edition.

[Livingstone and Lunt, 1992] Livingstone, S. M. and Lunt, P. K. (1992). Predicting personal debt and debt repayment: Psychological, social and economic determinants. *Journal of economic psychology*, 13(1):111–134.

[Luckett and Canner, 1991] Luckett, C. A. and Canner, G. B. (1991). Payment of household debts. *Federal Reserve Bulletin*, (Apr):218–229.

[Lusardi, 2008] Lusardi, A. (2008). Financial literacy: an essential tool for informed consumer choice? Technical report, National Bureau of Economic Research.

[Lusardi and Mitchell, 2011] Lusardi, A. and Mitchell, O. S. (2011). Financial literacy around the world: an overview. *National Bureau of Economic Research Working Paper Series*, (w17107).

[Lusardi and Tufano, 2009] Lusardi, A. and Tufano, P. (2009). "debt literacy, financial experiences, and overindebtedness." working paper no. w14808, national bureau of economic research.

[Meier and Sprenger, 2010] Meier, S. and Sprenger, C. (2010). Present-biased preferences and credit card borrowing. *American Economic Journal: Applied Economics*, 2(1):193–210.

[Meng et al., 2013] Meng, L., Zheng, Y., and Hui, R. (2013). The relationship of sleep duration and insomnia to risk of hypertension incidence: a meta-analysis of prospective cohort studies. *Hypertension Research*, 36(11):985–995.

[Merskin, 1998] Merskin, D. (1998). The show for those who owe: normalization of credit on lifetime's debt. *Journal of Communication Inquiry*, 22(1):10–26.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.

[Morselli et al., 2012] Morselli, L. L., Guyon, A., and Spiegel, K. (2012). Sleep and metabolic function. *Pflügers Archiv-European Journal of Physiology*, 463(1):139–160.

[Pattarin and Cosma, 2012] Pattarin, F. and Cosma, S. (2012). Psychological determinants of consumer credit: the role of attitudes. *Review of Behavioural Finance*.

[Shalev-Shwartz and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA.

[Slowik, 2011] Slowik, J. (2011). Credit card act ii: Expanding credit card reform by targeting behavioral biases. *UCLA L. Rev.*, 59:1292.

[Strömbäck et al., 2017] Strömbäck, C., Lind, T., Skagerlund, K., Västfjäll, D., and Tinghög, G. (2017). Does self-control predict financial behavior and financial well-being? *Journal of Behavioral and Experimental Finance*, 14:30–38.

[Sullivan and Fisher, 1988] Sullivan, A. C. and Fisher, R. M. (1988). Consumer credit delinquency risk: Characteristics of consumers who fall behind. *Journal of Retail Banking*, 10(3):53–64.

[Sullivan and Luke, 2007] Sullivan, K. M. and Luke, S. (2007). Evolving kernels for support vector machine classification. In *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, GECCO '07, page 1702–1707, New York, NY, USA. Association for Computing Machinery.

[Sunstein and Reisch, 2017] Sunstein, C. R. and Reisch, L. A. (2017). *The economics of nudge*. Routledge.

[Tokunaga, 1993] Tokunaga, H. (1993). The use and abuse of consumer credit: Application of psychological theory and research. *Journal of economic psychology*, 14(2):285–316.

[Turunen and Hiilamo, 2014] Turunen, E. and Hiilamo, H. (2014). Health effects of indebtedness: a systematic review. *BMC public health*, 14(1):1–8.

[Watkins, 2000] Watkins, J. P. (2000). Corporate power and the evolution of consumer credit. *Journal of Economic Issues*, 34(4):909–932.

[Watson, 2003] Watson, J. J. (2003). The relationship of materialism to spending tendencies, saving, and debt. *Journal of economic psychology*, 24(6):723–739.