NOVA IMS
Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Overcoming over–indebtedness with AI

A case study on the application of AutoML to research

Victor Cardoso Reis Costa

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

OVERCOMING OVER–INDEBTEDNESS WITH AI

A case study on the application of AutoML to research

by

Victor Cardoso Reis Costa

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

Advisor: Prof. Dr. Mauro Castelli

November 2020

ACKNOWLEDGEMENTS

I would like to thank…

My incomparable supervisor, Prof. Dr. Mauro Castelli, who was always relentlessly patient, available and insightful throughout the development of this thesis and many other endeavors.

My good friend, Fernando Peres, with whom I worked side-by-side for over a year, striving for solutions to the problems identified in this research.

My parents, who throughout my entire life have always put my needs above theirs. I hope one day I can make you as proud to be my parents as I am of being your son.

And last, but definitely not least, my beautiful Giovanna Luna Lucas. Without you this would not have been possible. Thank you for being much more than I could ever have hoped to find in a life partner.

ABSTRACT

This research examines how artificial intelligence may contribute to better understanding and overcoming over-indebtedness in contexts of high poverty risk. This study uses a field database of 1,654 over-indebted households to identify distinguishable clusters and to predict its risk factors. First, unsupervised machine learning generated three over-indebtedness clusters: low-income (31.27%), low credit control (37.40%), and crisis-affected households (31.33%). These served as basis for a better understanding on the complex issue that is over-indebtedness. Second, a predictive model was developed to serve as a tool for policymakers and advisory services by streamlining the classification of over-indebtedness profiles. On building such model, an AutoML approach was leveraged achieving performant results (92.1% accuracy score). Furthermore, within the AutoML framework, two techniques were employed, leading to a deeper discussion on the benefits and inner workings of such strategy. Ultimately, this research looks to contribute on three fronts: theoretical, by unfolding previously unexplored characteristics on the concept of over-indebtedness; methodological, by proposing AutoML as a powerful research tool accessible to investigators on many backgrounds; and social, by building real-world applications that aim at mitigating over-indebtedness and, consequently, poverty risk.

KEYWORDS

over-indebtedness; poverty risk; credit control; artificial intelligence, machine learning; automated machine learning; automl

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ACP *Association for Consumers' Protection*. Fictitious name referring to the real Portuguese institution that helped the present research by providing data and relevant expert input.

AutoML Automated Machine Learning

AI Artificial Intelligence

ML Machine Learning

EU European Union

SOM Self-Organizing Maps

SVM Support Vector Machine

SVC Support Vector Classifier

Nu-SVC Nu-Support Vector Classifier

1 INTRODUCTION

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 paychecks would not otherwise permit (Watkins, 2000). In general, credit has become a way to promote 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 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).

However, as expected, debt is troublesome to individuals and governments alike when it reaches such high levels that households become over-indebted – i.e. incapable of repaying credits when they are due. Apart from the more straightforward financial consequences – such as increased risk of deprivation and poverty – over-indebtedness even carries additional indirect hardships upon a household. Social stigma, financial exclusion, deteriorated well-being and health, and even family breakdown are a few examples (Alleweldt et al., 2013).

Also, collective burdens stem from over-indebtedness, reflected, for instance, as a reduction in labor activity. Those employed while dealing with extreme debt might present a considerable drop in productivity due to stress and failure to concentrate at work. At the same time, for the unemployed, feelings of failure and lack of self-confidence that comes from social stigma, undermines their ability to access new employment (Alleweldt et al., 2013).

Broadening the discussion to encompass the concept of poverty, over-indebtedness has shown to be a major factor to foster localized scarcity within more fragile developed countries (Shaefer & Edin, 2013). Notably, severe economic austerity in such countries can play a big role to generalized hardships regarding consumers' debt. For instance, as a result of the European sovereign debt crisis, the Portuguese society was besieged by 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).

1.1 THEORETICAL MOTIVATION FOR THE RESEARCH

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 has 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). In practice, actual cases of over-indebted households are likely to be multifaceted.

The hypothesis of multiple causes surrounding over-indebtedness marked the initial motivation for the research that culminated in a paper (a joint work, to be published at the Journal of Business Research) and the present thesis. Nuances differ between the two documents and shall be specified later on after more context is provided. Nonetheless, for both outputs – the paper and the thesis – the object of study was the concept of over-indebtedness and its risk factors among Portuguese households from the aftermath of the 2008-2009 European financial crisis leading up to recent times.

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

1.2 PROJECT'S FRAMEWORK, SCOPE AND GOALS

To provide a new view, the existence of different profiles of over-indebtedness was evaluated by looking for distinguishable combinations of characteristics (based on households biographical and financial information) that would allow to categorize over-indebted consumers in an exhaustive and mutually exclusive way. For this goal, a large field data set of over-indebted cases was investigated. The cases were acquired from households who contacted the debt advisory services of a well-known Portuguese association focused on providing guidance for consumers during financial hardship. To preserve the institution's anonymity, it shall be referred to as *Association for Consumer Protection* (ACP) from now on.

Furthermore, given the extension and complexity of the data set, artificial intelligence (AI) was used to look for such distinguishable characteristics that would allow to describe and identify consumers' over-indebtedness. Namely, a clustering method was first applied to investigate the existence of patterns within the data, hinting on the definition of different profiles. Afterwards, predictive models were evaluated to assess the feasibility of tools functioning as classifiers of such profiles. The two employed AI techniques are further detailed throughout the study, both in terms of their theoretical backgrounds and also in relation to the methodology followed during implementation.

Going back to the aforementioned research outputs, while both followed the same initial development stages, they differ on emphasis and project scope:

- The *paper* (referred as such throughout this document for simplification) focuses primarily on the theoretical contributions to business achieved by AI investigation. Also, important to note that it was the product of a joint effort between researchers from Nova IMS and from the University of Lisbon's Psychology School (FPUL).

- This dissertation, on the other hand, focuses on the process of building AI tools from what was learned through the business research – i.e. training predictive models and conceiving AI embedded software. Apart from being a more technical-led research output, the distinction also translates the author's main scope of participation in the project as a whole.

Formally, the ultimate goal of this study 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 reduce poverty risk at earlier stages.

The study begins by formally defining over-indebtedness and discusses several risk factors of over-indebtedness. As a follow up, it provides a theoretical review on prior business and psychological accounts of consumers' credit use. It then evolves to describe over-indebtedness profiles through the lens of multi-dimensional analysis, facilitated by AI. From the developed clusters, it discusses the opportunities for practical applications. Specifically, it presents software conceived for both the ACP's internal processes and the Portuguese society as a whole. These applications are powered by AI, so the study details the methodology and results for building the desired predictive model.

Extending on the methodology for generating predictive models, the research makes its case on the benefits of using Automated Machine Learning (AutoML) for business investigation. It discusses how AutoML can be a scalable and robust tool to enhance research sectors not yet leveraging the capabilities of AI to its full potential.

As an added statement, throughout the process of building the aimed AI system, when the models' performance appeared unsatisfactory, the author dealt with the challenge by improving the guiding principles for automated learning and not manually interfering with the "learners" (i.e. the models), reinforcing the potential outcomes that an automated process may yield. These accounts will be further discussed during the Methodology chapter.

2 CONTEXT

Chapter 2 unfolds several topics to offer all the necessary context leading up to the development of the predictive model. From the socio-economic problem the model aims to mitigate, passing through the relevant theoretical background (both on artificial intelligence terminology and its applications to business research) and ending with the research's practical contributions, the chapter resembles the investigators' thought process that initially instigated the model's development. The predictive model itself is detailed in the following chapter (i.e. Chapter 3, Methodology & Initial Results).

2.1 OVER-INDEBTEDNESS: DEFINITION AND REVIEW

Section 2.1 opens by formally characterizing what constitutes over-indebtedness within the scope of this study. With that goal in mind, it is important to first recognize that, as stated by Fondeville et al. in a 2010 research for the European Commission, "*there is no standard definition of over-indebtedness used in the EU and, accordingly, no set of standardized, and harmonized, statistics on it*". The statement leads an investigation on the topic to decide upon the specific set of criteria used to identify over-indebted scenarios. For the present study, the following characterization put forward by the European Commission is taken as its standard definition:

> *"The unit of measurement should be the household because the income of individuals can be pooled – and indeed, is usually assumed to be pooled – between household members.*
>
> *Indicators need to cover all financial commitments of households – borrowing for housing purposes, consumer credit, paying utility bills, meeting rent and mortgage payments and so on – and not be confined to just one aspect.*
>
> *Over-indebtedness implies an inability to meet recurring expenses and, therefore, it should be seen as an ongoing rather than a temporary, or one-off, state of affairs.*
>
> *It is not possible to resolve the problem simply by borrowing more.*
>
> *For a household to meet its commitments requires it to reduce its expenditure substantially (or find ways of increasing its income)."*
>
> — *Fondeville et al., 2010*

To conclude, Fondeville et al. (2010) reduces to one sentence the understanding of such cases:

> "*An over-indebted household is, accordingly, defined as one whose existing and foreseeable resources are insufficient to meet its financial commitments without lowering its living standards, which has both social and policy implications if this means reducing them below what is regarded as the minimum acceptable in the country concerned.*"
>
> — *Fondeville et al., 2010*

Highlighting from the proposed definition above: "[…] whose existing and foreseeable resources are insufficient to meet its financial commitments". In that respect, one can postulate that 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). A reinforcing notion on the seriousness of a household's situation when dealing with the issue.

The following topic (2.1.1) considers numerous theoretical accounts on risk factors of over-indebtedness. It serves as background to formally discuss the research's hypothesis of a multi-faceted phenomenon.

2.1.1 Literature on 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. Consumer's lack of knowledge concerning financial products and concepts, makes 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 Lusardi, 2008; Dellande et al., 2016).

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 prey on 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 by 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., Loibl, Jones, & Haisley, 2018; Hertwig & Grüne-Yanoff, 2017; 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 presence 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 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 debt-to-income 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.

2.1.2 Considerations on existing literature

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) have been shown to be 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. Thus, this study hypothesizes 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 actually carve different profiles of over-indebted households is an empirical question that begins to be answered in the present work.

This study suggests 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). Such a bottom-up approach is methodologically challenging but achievable using artificial intelligence to develop descriptive models concerning the risk factors of over-indebtedness of Portuguese consumers.

2.2 ARTIFICIAL INTELLIGENCE: TERMINOLOGY AND PRIOR RESEARCH IN BUSINESS

Before advancing onto the described approach and in order to contextualize the research here reported, first, the relevant terminology relating to artificial intelligence is defined. It is then followed by a systematic literature review on prior work using AI in business investigation. This comes as an assessment on the study's relevance towards the perception of AI as a socio-economic research method, given its relatively scant use for such purpose.

2.2.1 Artificial intelligence introductory concepts

The current section contains a beginner-friendly definition of artificial intelligence terms regularly discussed throughout the dissertation. These can be considered fundamental concepts within the arena. Hence, readers with related background knowledge are most probably familiar with the topics.

Artificial intelligence (AI) 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). This study resorted to a particular sub-area of AI, called **Machine Learning** (ML) (Marsland, 2015) to autonomously extract patterns from over-indebtedness data.

"Extracting patterns" can be viewed from a descriptive perspective, such as finding a pattern that describes a dataset. Connected to the concept of **Unsupervised Learning**, in this scenario, one looks to build a descriptive model that communicates the data's intrinsic

patterns. For instance, it can be used to group sets of data observations according to their mutual similarities (**clustering**). Given a particular distance metric, typically, groups are formed in order to maximize the intra-cluster distances and minimize the inter-cluster distances.

On a predictive scenario, however, "extracting patterns" may translate into learning the data's intrinsic patterns that lead to a determined outcome. This second objective relates to **Supervised Learning**. Its general goal is to build predictive models that are able to estimate parameters from data and later use this "knowledge" on a new observation for either: (a) predict the most appropriate category it relates to (i.e. building a **classification** model of categorical target variables) or (b) predict a numeric value – i.e. building a **regression** model for numeric target variables.

Within the different forms of building AI, this study employed **Automated Machine Learning** (AutoML; Feurer et al., 2015), which enabled evaluating thousands of models generated by state-of-the-art algorithms with multiple combinations of parametrization. AutoML offers a clear benefit over traditional and more manual ML approaches, which shall be discussed during the Methodology chapter.

2.2.2 Artificial Intelligence in business research

A systematic literature review was conducted, 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. Again, it is important to re-emphasize that this comprehensive review was a product of the conjoint work of investigators from Nova IMS and FPUL and so deserve due credibility.

The following keywords were used in the search process, along with *business research*: *machine learning*, *artificial intelligence*, *support vector machines*, *automated machine learning*, and *AutoML*. The relevance of the term *support vector machines* will become clearer for the reader in chapter 3, when methodology and results are discussed. *AutoML* and au*tomated machine learning,* where the latter is simply the extended version of the former, will also be further discussed in the methodology.

After obtaining the initial set of articles, a snowballing procedure was applied for examining the references within these articles in the effort of finding additional studies. The search process was completed in May 2020 (see Table 1 for details). A total of 11 articles were found 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 a comprehensive comparison between several Machine Learning algorithms, similar to the idea of *AutoML*. This goes to show that the method employed during this research to characterize and predict over-indebtedness is still rather new.

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.

Probably, the most similar approach to the one here presented was the 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, while the present work did include both algorithms in its process, it also considered several other options. In fact, Logistic Regression and Random forests were outperformed by alternative algorithms. Also, a descriptive modeling is absent in Montiel et al. (2017), while it is a fundamental part of this study. A final diverging aspect is that 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, the data used here was originated at a consumer protection institution, not a bank, which may highlight the societal relevance of this work's field dataset.

Other less directly related but important research include Agarwal et al. (2018)' use of Gradient Boosted Models to connect financial outcomes and phone-based social behavior to predict financial wellbeing in the US; and Alomari (2017)'s use of various data mining algorithms for default prediction of peer-to-peer loans and for 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.

2.3 CLUSTERS: A MULTI-FACETED PERSPECTIVE MATERIALIZED

Now that proper background on the risks of extreme debt and regarding artificial intelligence was provided, the study continues to explore the existence of different over-indebted household profiles through the lens of multi-dimensional analysis. As previously alluded, an unsupervised machine learning method was applied to surface inherent clusters from ACP's data set. The specific algorithm used is named Self-Organizing Maps (SOM) and as a final outcome – i.e. after a comprehensive process of model evaluation and selection – it extracted 3 clusters with distinguishable characteristics: low income households, low credit control households, and crisis-affected households.

In order to maintain this section as a conceptual discussion, specific technical details about the algorithm's implementation and training were reserved for the Methodology chapter. Below are the socio-economic descriptions of each cluster; in other words, the 3 profiles of consumers facing over-indebtedness.

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 with the lowest income per capita. Furthermore, the consumers of this group have the lowest total credit monthly installment, the lowest credit card monthly effort rate, and the lowest housing credit monthly installment of the three clusters. This group presents the lowest level of unemployment, which is 12.6% below the dataset mean, and is mostly employed in the private sector. One of the main issues reported as a cause of the financial difficulty is the increase in family members.

Cluster 2 – Low credit control households:
This cluster includes cases of over- indebtedness predominantly due to other causes and a few crisis- related cases. 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 and the lowest number of people in the household, they present the highest credit effort rate and personal credit rate. On the other hand, these consumers have the lowest car credit effort rate and the lowest household expenses.

Cluster 3 – Crisis-affected households:

Cluster 3 presents cases of over-indebtedness that are mostly due to the crisis and a few pertaining to other causes not related to the crisis. This cluster is characterized by low income per capita and include the largest families and the highest household expenses of the three clusters. The main causes for over-indebtedness are unemployment, which is, 21.3% higher than dataset mean; salary cuts, 6% higher than dataset mean; and spouse's unemployment, which is 4% higher than dataset mean. These consumers have the highest provision with housing and with other credits or debts.

2.4 APPLICATIONS: RESEARCH'S PRACTICAL CONTRIBUTIONS

This section presents the practical applications stemming from the descriptive investigation produced by the unsupervised machine learning method (SOM) and conclusions observed regarding consumers' profiles. It elaborates on how the interest in assisting debt advisory services (such as ACP) and the Portuguese public as a whole led to conceiving two *softwares* and the classification model that powers both.

For additional context, consumers that approach the ACP 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 should they give up, from simple consumption goods (e.g., mobile phone, computer) to important long-term goods, such as cars and their houses. Also, they might serve as mediators between consumers and creditors to arrive at a satisfactory solution for both parties.

2.4.1 1st Application: assisting decision-making on over-indebtedness cases

Once surfaced the intrinsic consumer profiles from ACP's data, an opportunity presented itself to further supplement the association's mission on assisting Portuguese citizens. As one would expect, each household's scenario instigates a different strategy that ACP follows to help alleviate their financial hardships. Within this process of defining the most appropriate set of actions to take, some key information can steer the approach towards drastically different paths.

On a conceptual level, the degree of over-indebtedness and, perhaps more importantly, the causes that culminated into financial hardship, are key information that ACP's analysts assess to reach an initial conclusion regarding which general strategy should be applied.

For instance, a scenario where the main issue leading to the household's difficulties originates solely from what could be considered "irresponsible spending", ACP might not take the case at all. In another example, if the consumer is unable to afford a reasonable minimum of debt repayment, the association will probably organize the appropriate documentation and guidance for the household to declare insolvency but will not engage in actual negotiation with creditors.

Arriving at such conclusions (and numerous others) traditionally meant analyzing each individual metric provided by consumers, coupled with heuristic thresholds defined by the association's experts. Figure X illustrates a fragment of the reasoning one senior analyst would follow throughout the decision process. As an example, if the household's monthly effort rate is below 35%, ACP assumes as not being a pressing issue and, ultimately, is inclined to dismiss the case. In other recurrent scenarios, if effort rate is in indeed higher than an arbitrary threshold, then a set of sequential qualitative, univariate considerations are made regarding the household's information to decide upon the strategy to employ.

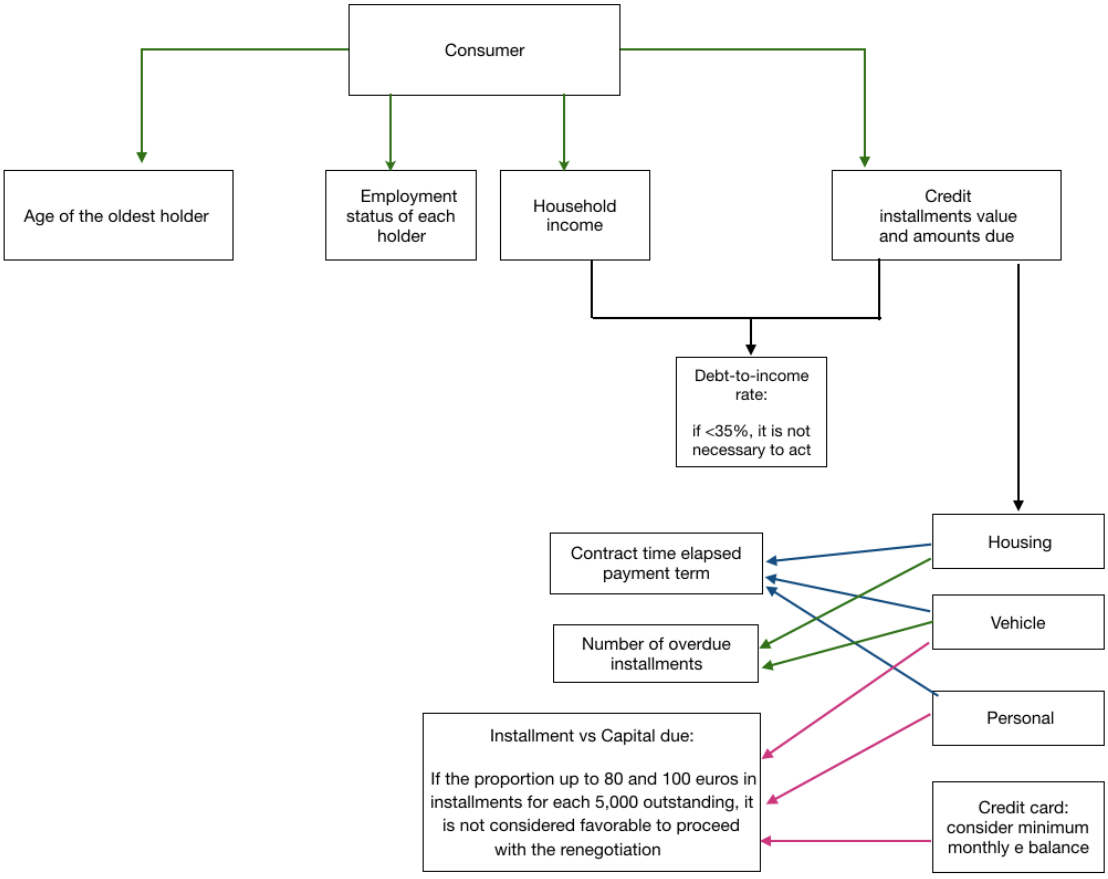

Figure 1: Fragment of the mental model followed by a senior ACP analyst throughout the decision process.

Although a tried and tested process, ACP understood that AI could be leveraged to improve some of its aspects. First and foremost, it is time-consuming. A senior analyst must carefully

investigate each variable and compare it against the set of heuristic rules that guides decision-making. Secondly, this process is essentially univariate and ignores the added information a multi-dimensional assessment provides. Lastly, it is not standardized. The weight given to each metric lies within the analyst's qualitative perception. This might reflect on significantly different end results depending on the individual.

After reviewing the clusters profiles, became clear that one endeavor responded to all 3 gaps listed above: a software with the embedded capability of instantly labeling each new household in one of the scenarios, while also presenting a systemic view of the case. Such application would speed-up the process of identifying root issues, through an inherently multi-dimensional method, while setting the same rationale to be shared by all internal stakeholders. Thus, responding on all 3 fronts.

Visually, the application's participation within decision-making is illustrated below (figure 3) by the conceptual diagram of ACP's internal process when dealing with a new case. As a high-level description, after collecting the household's financial information – either through a web portal or in-person at one of ACP's offices – an analyst first makes the rigorous assessment on taking or not the case. Afterwards, if accepted, he or she evaluates the most appropriate strategy.

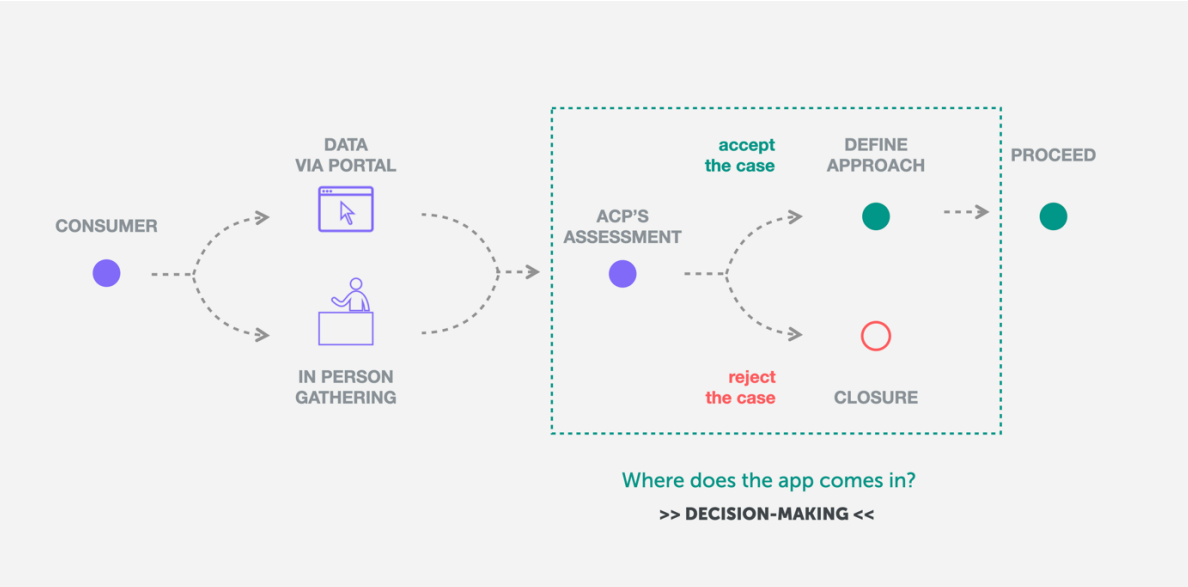

Figure 2: Conceptual diagram of ACP's internal process for new cases.

Identifying the consumer's profile speeds-up these first main decisions on the general strategy to take, because it specifically alludes to the relevant key information mentioned before. Here is where a predictive model became a valuable addition to the research. If satisfactory results were yielded from the model's training, it would efficiently translate the

software's expected capability of quickly labeling new cases. Details on the model's implementation are specified at chapter 3.

The software

The software was designed to reflect the "funneling behavior" reported by ACP: start by deciding on taking the case or not, and from then on, continuously narrow the scope of possible strategies. For that, each household's case is displayed independently within the user's interface and is based on a "slides" rationale – i.e. each following slide (or group of slides) presents more fine-grained information. Figure 4 presents a broad view of the entire interface illustrating a household's example case.

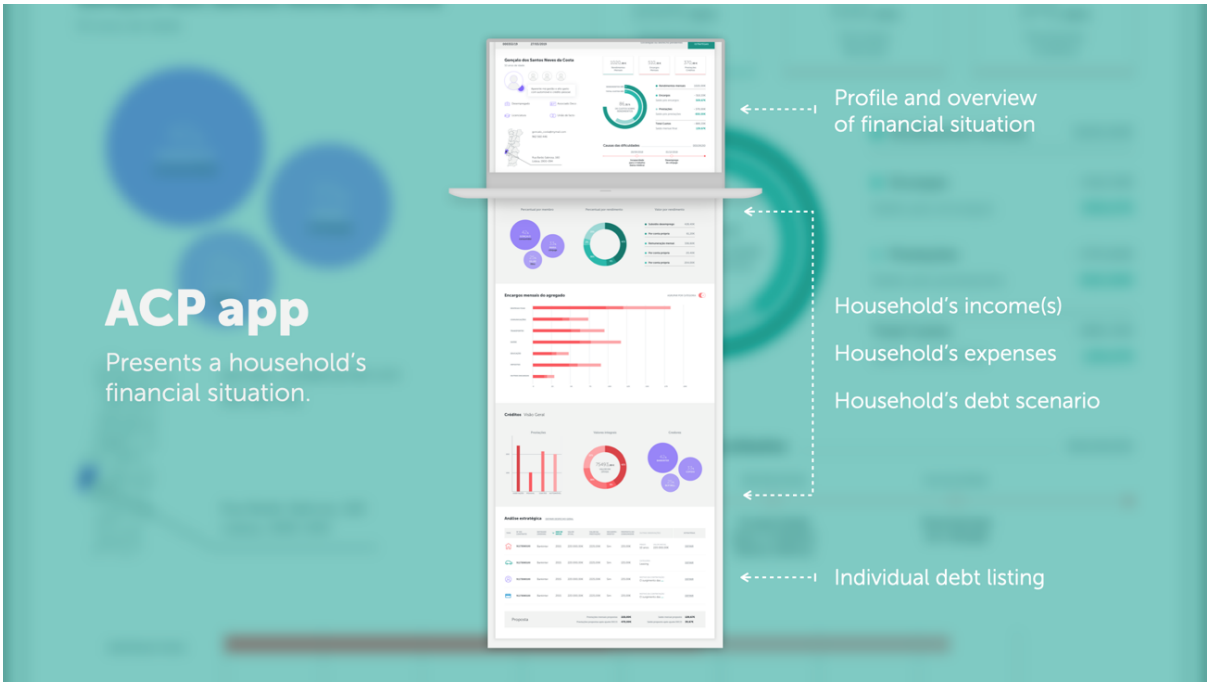

Figure 3: ACP's internal app. Specifically, interface of a household's case.

The "first slide" was designed to act as an overview and essentially presents the crucial information for the analyst to answer questions at a broader level – e.g. "should we take the case or not?". Namely it presents a summary of the household's financial situation and its over-indebtedness profile, classified by the predictive model. Figure 5 illustrates the section.

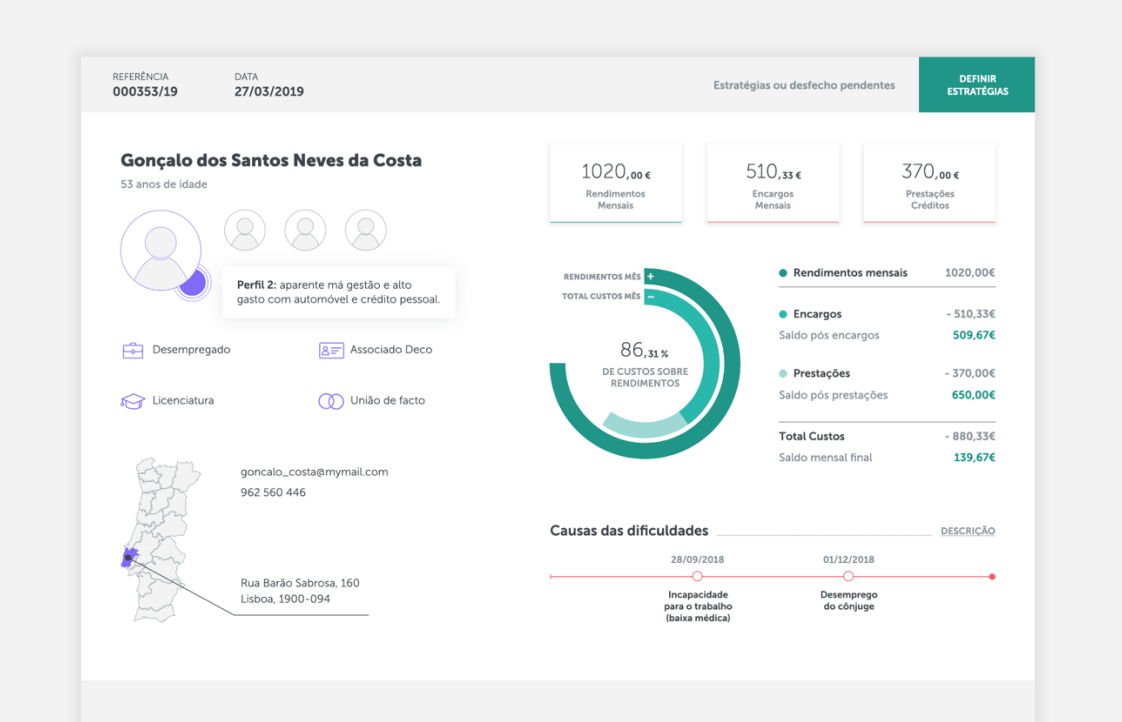

Figure 4: First section ("slide") of the interface designed for ACP.

The second, third and fourth "slides" dive deeper into each pillar of the household's financial information. Respectively, those are: sources of income, monthly expenses and credits overview. Figure 6 presents all 3 sections.

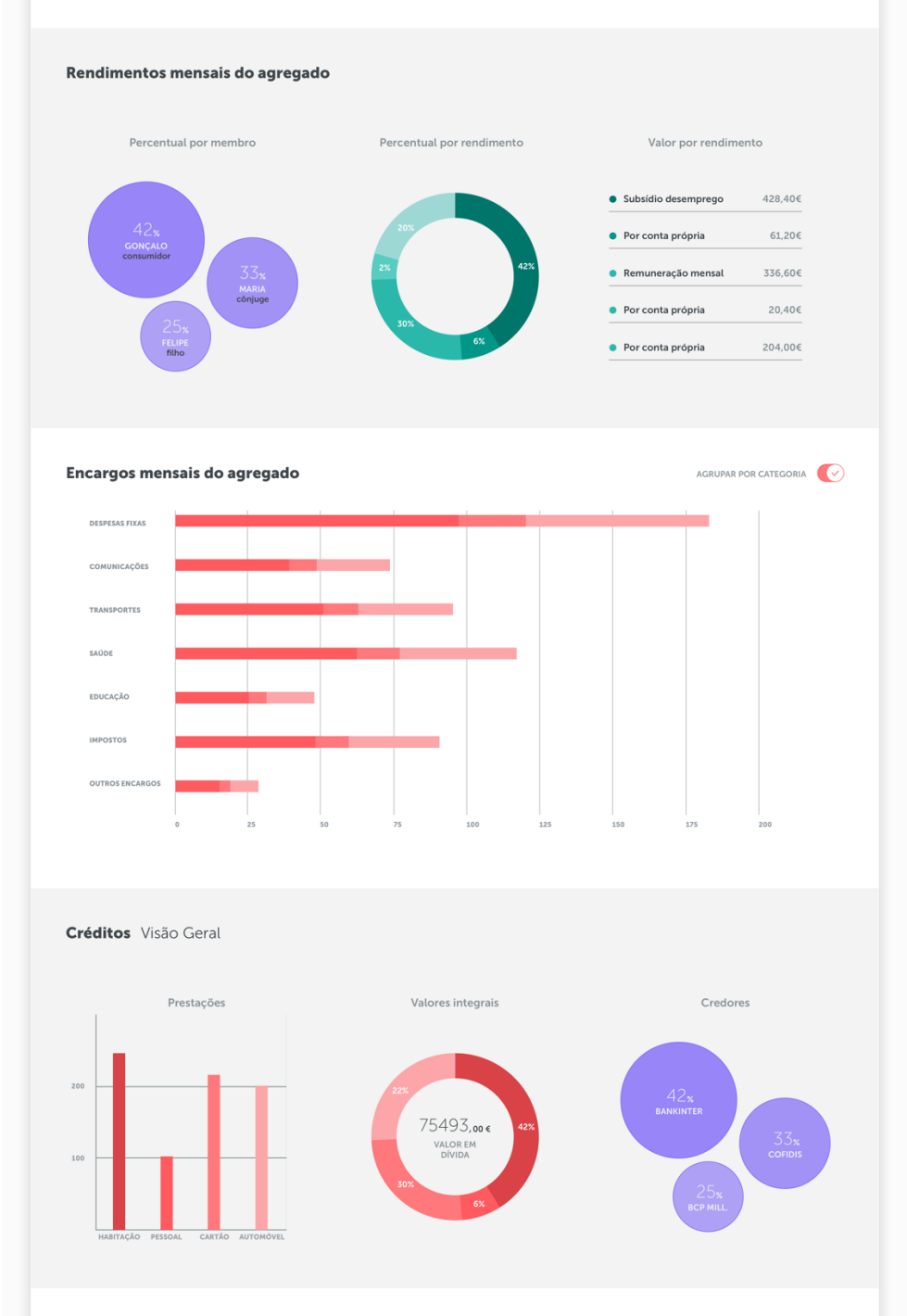

Figure 5: Second, third and fourth sections ("slides") of the interface designed for ACP.

The fifth and final "slide" presents the household's debts broken down into each credit contract. The analyst is able to assess the individual debt's information, such as monthly charge, creditor, type (e.g. credit card, housing and vehicle), among others. Figure 7 presents the described listing. Also, figure 8 shows insights surfaced by the software on how to deal with the household's debts at a granular level – i.e. per individual credit.

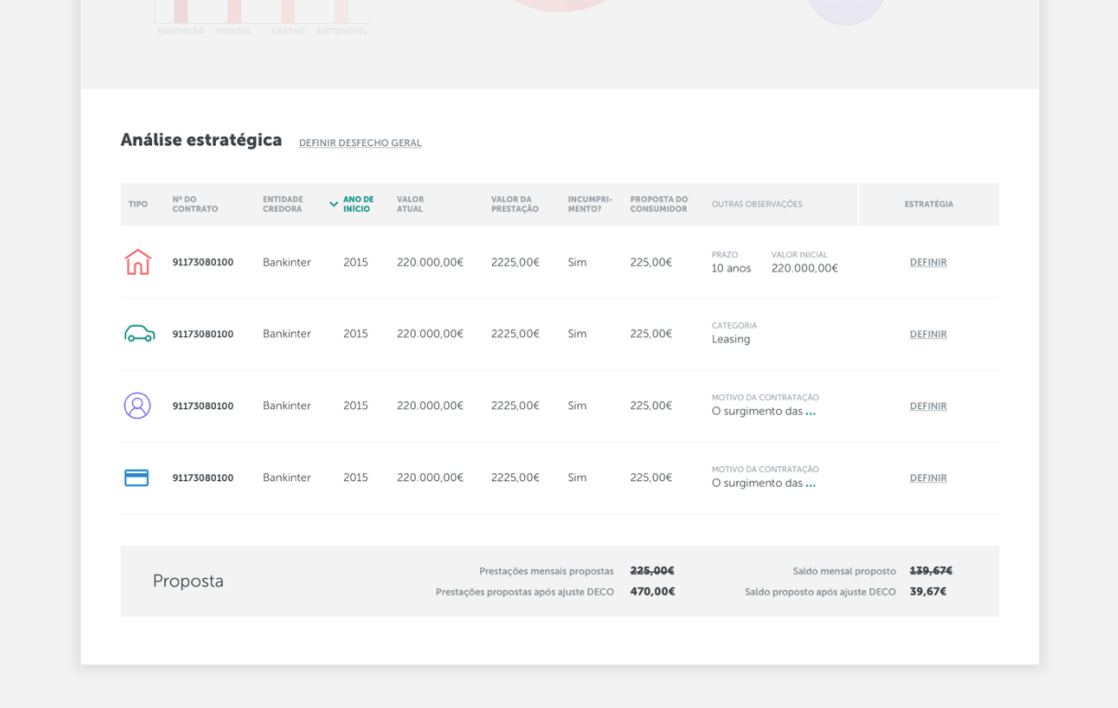

Figure 6: Fifth section ("slide") of the interface designed for ACP.

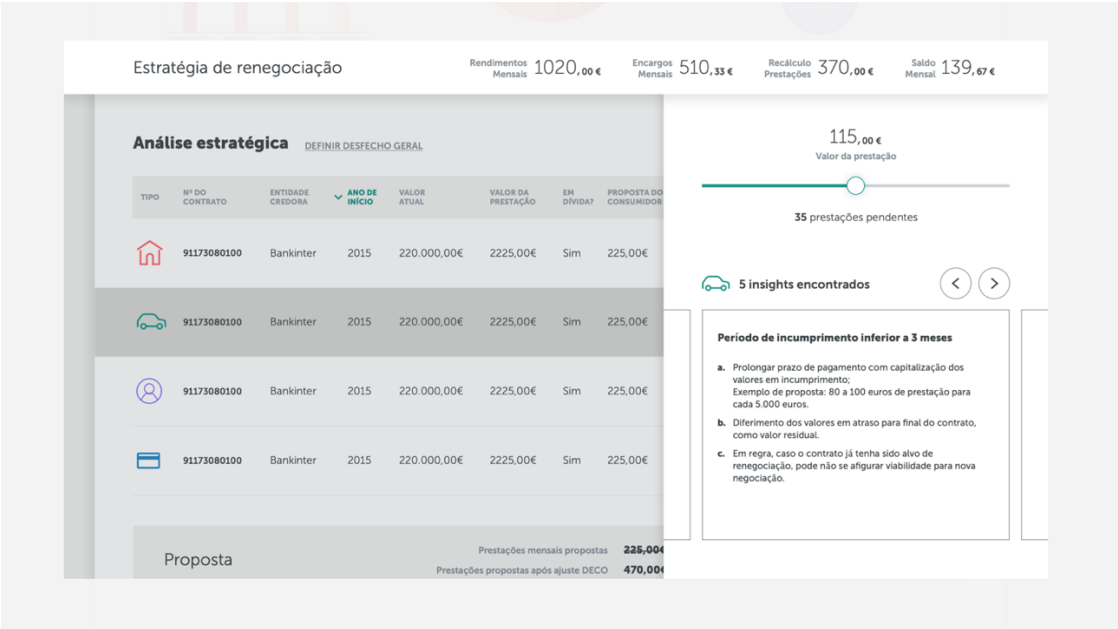

Figure 7: Additional layer on the fifth section ("slide") of the interface designed for ACP. Illustrates insights uncovered by the software.

2.4.2 *2nd Application: reaching the general public*

Once defined how the AI embedded application will support ACP's internal processes, a secondary goal was established for the project's practical contributions. One in which a broader scope of the population could be reached, while still feeding back into ACP's ecosystem.

With the general Portuguese public in mind, an additional application was conceived that served two purposes: offering relevant financial information to consumers and providing an initial, uncommitting access to potentially needed help.

Following upon prior research that established the ineffectiveness of simply presenting financial knowledge, this application strived to be compelling to users and, consequently, be successful in communicating its message. For such, instead of presenting static information, the application responds contextually to consumers' inputs and, therefore, is able to offer personalized feedback on the household's situation. The concept is to approximate a user-friendly, personalized consultant. At scale. Figure 8 illustrates the solution's conversational aspect.

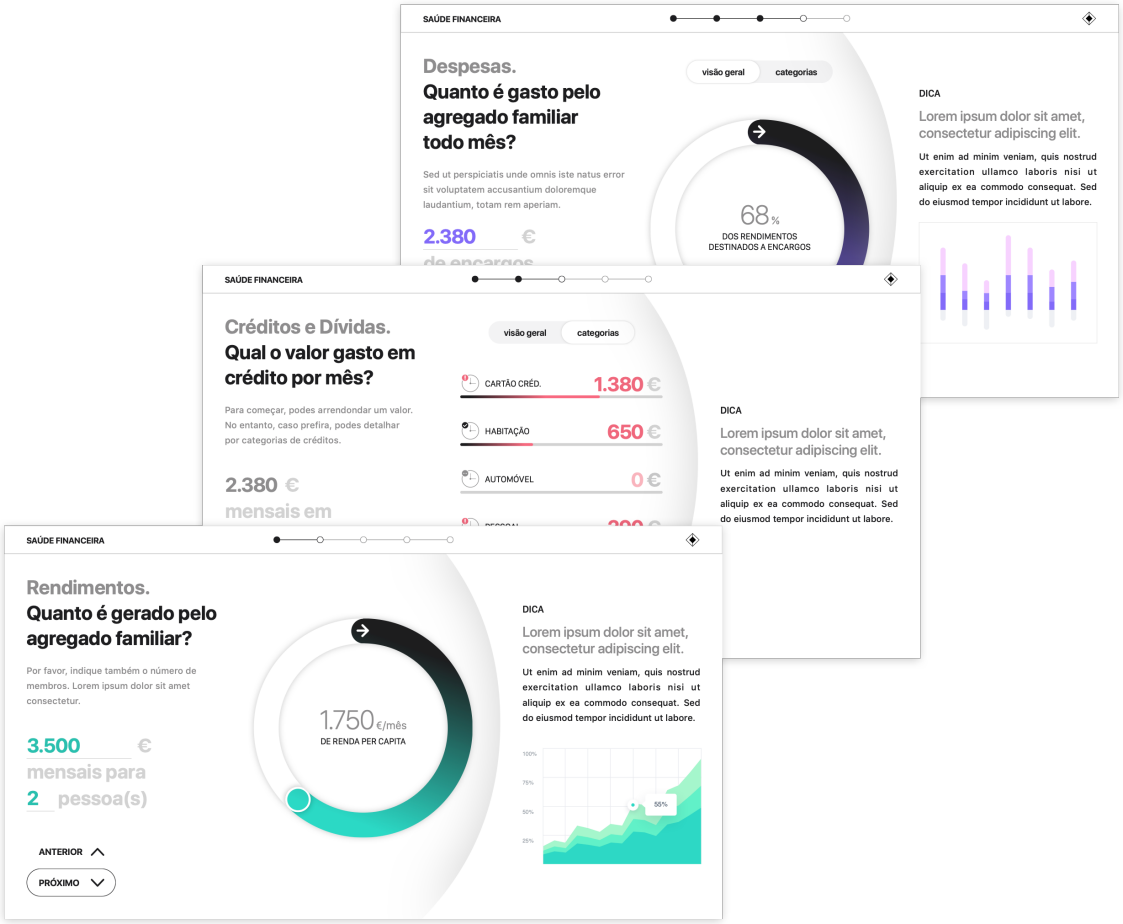

Figure 8: User interface of the application for the Portuguese general public.

At the same time, in an eventual scenario where the user-inputted data indicates a troublesome financial situation, the software customizes its output to include an easy connection to debt advisory services (ACP being one example), with the option of sharing the already provided household's statistics. Users are then encouraged to get in touch.

Going back to ACP's high-level diagram, while the first software assisted on the internal decision-making routine, the second software complements the initial data acquisition phases, helping the association scale its efforts. Ultimately, the application looks to reach a greater number of consumers in financial stress not only for its digital medium but, also, for eliminating the reported embarrassment of being candid regarding one's situation. It alleviates "social stigma". Figure 10 illustrates both applications contribution to ACP's processes.

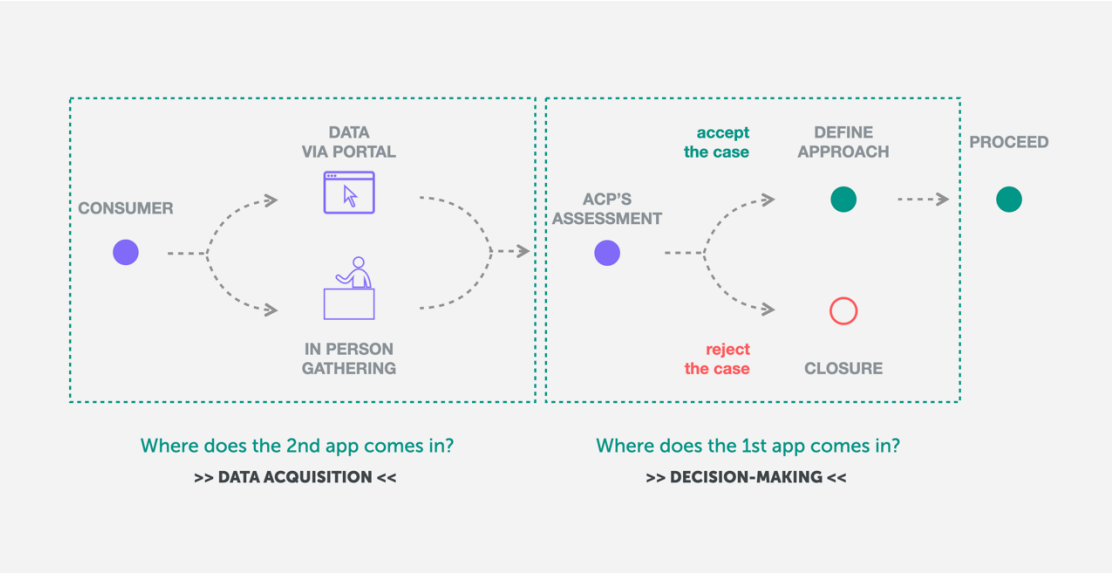

Figure 9: Revised diagram of ACP's internal process to include 2nd application's contribution.

3 METHODOLOGY & INITIAL RESULTS

Chapter 3 outlines the entire process followed to obtain the machine learning models; from data collection to model training. Each phase, illustrated by Figure 10, is methodologically discussed in detail below. It starts reviewing the dataset (section 3.1), how it was collected (by the ACP) and preprocessed. Next, section 3.2 describes the unsupervised modeling process that generated the target labels for predictive modeling, which is then presented, as well, in section 3.3. Considering the predictive model as the final outcome, its results are separately presented in chapter 4. All intermediate results leading to it (e.g. the clustering results and initial iterations of predictive modeling) are reported contextually throughout the current chapter.

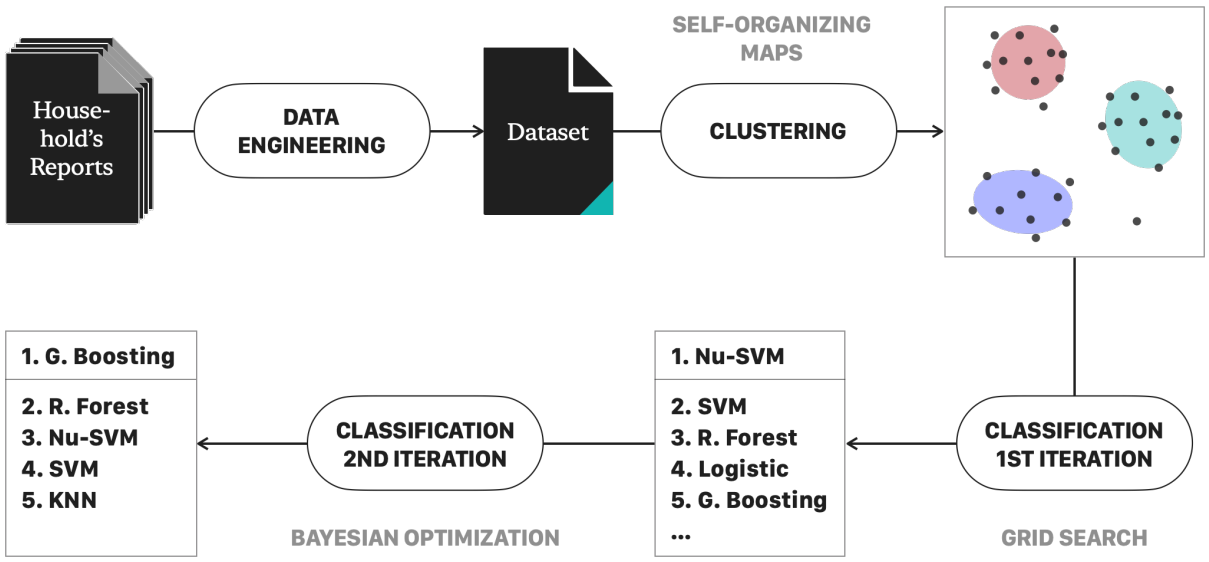

Figure 10: Project's phases.

3.1 THE DATASET

In this work, the data analyzed was gathered from consumers under assistance for over-indebtedness by the ACP. The data consisted 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.

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 analysis 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.

3.1.1 Data engineering

In the data engineering phase, data preprocessing techniques were executed to treat missing values, normalize the data and generate new features by extracting relevant information from the existing features (feature engineering). With the prepared data, the second phase (data selection) performed Pearson and Spearman correlation analysis on numeric variables and Cramér's V and Information Value on categorical attributes to arrive at a feature selection that eliminates redundancy while maintaining relevant information. 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 all the dataset minimizing the 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 than 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, a normalization process was performed. 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.

In the end, after treating missing values, removing outliers, encoding categorical features, normalizing numerical features and testing different feature selection combinations, the final elected and preprocessed set was composed by the variables listed below (note that the term *effort rate* refers to the household's "debt-to-income ratio"):

- cause classification (categoric) – crisis and other causes not related to crisis
- income per capita (numeric)
- total expenses (numeric)
- effort rate with credit card (numeric)
- effort rate with housing credit (numeric)
- effort rate with car credit (numeric)
- effort rate with personal credit (numeric)
- effort rate with other types of credit or debts (numeric).

This was the feature set used during unsupervised and supervised modeling. Nonetheless, while analyzing profiles generated by the descriptive model, all variables were considered.

3.2 UNSUPERVISED ML MODELING

As previously mentioned, Self-Organizing Maps (SOM) algorithm was applied to identify and describe the consumers' profiles of over-indebtedness. Among the different SOM variants, the Kohonen Network (Kohonen, 2013) was the version of choice. 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 a number of iterative steps. For a generic input pattern (or data observation) $\boldsymbol{x}$, the following steps must be executed (Resta, 2012):

(1) evaluate the distance between $\boldsymbol{x}$ and the vector of weights of the synaptic connections entering in each neuron. For instance, the Euclidean distance between the input vector $\boldsymbol{x}$ 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), in order 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, which 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.

3.2.1 Training settings and results

The descriptive model was trained using the Kohonen R Package (R Studio). Specifically, the method supersom (Supervised SOM). The ***grid size*** defined for this study was 100 cells (dimension x = 10 and dimension y = 10), which presented good results – i.e. generated a considerably homogenous distribution regarding observations count per nodes and did not return any empty nodes (nodes without observations). Figure 11 exposes such results.

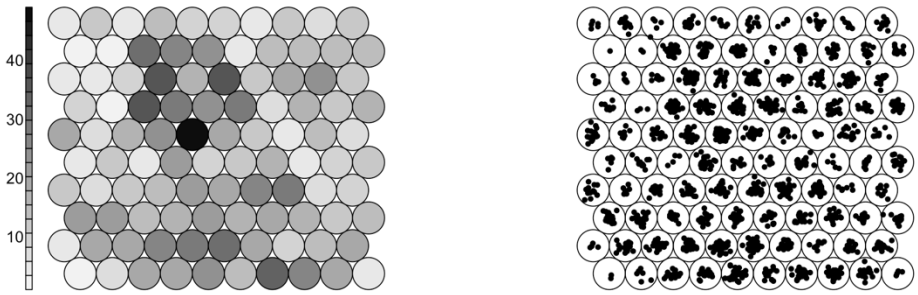

Figure 11: SOM's results: observations' count per nodes. The map on the left translates the counts through color intensity, while the one on the right represents each observation as a dot.

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 dist.fcts parameter is a vector of distance functions to be used to calculate the distances

among nodes: Tanimoto distance (for categorical data/factors) (Lipkus, 1999) and Euclidean distance (for numeric features) (Gower & Legendre, 1986).

In sum, the final descriptive model used 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 12 shows the progress of SOM training and the decrease of the mean distance to the closest unit distance over time (iterations).

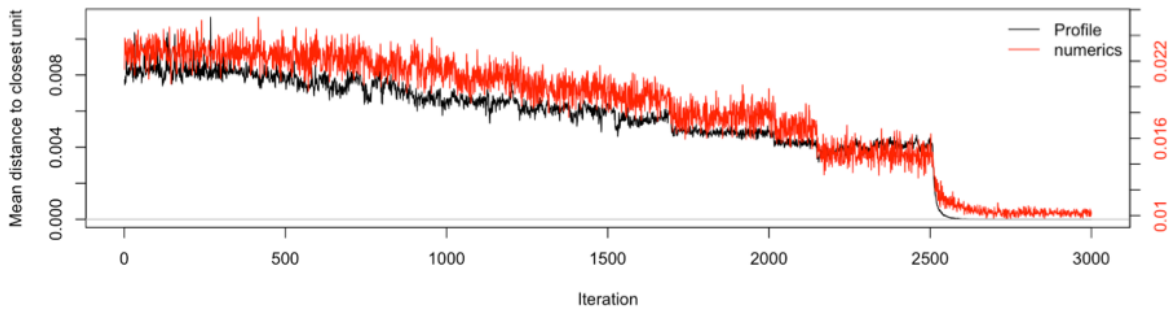

Figure 12: Self-Organizing Maps training iterations vs Mean distance to closest unit.

3.2.2 Clustering statistical results

Profiling description highlighted the distinguishable characteristics of each cluster, showing the values of similar clusters and ranking the variables in accordance with 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 then based on the analysis and capacity of cluster description (descriptive ability) in accordance with over-indebtedness analysis considerations and, also, complemented with traditional assessments on the optimal number of clusters, such as Elbow and Silhouette methods. Plots for Elbow and Silhouette are shown in Figure X and variables' significance per cluster are presented further on, in Figure X.

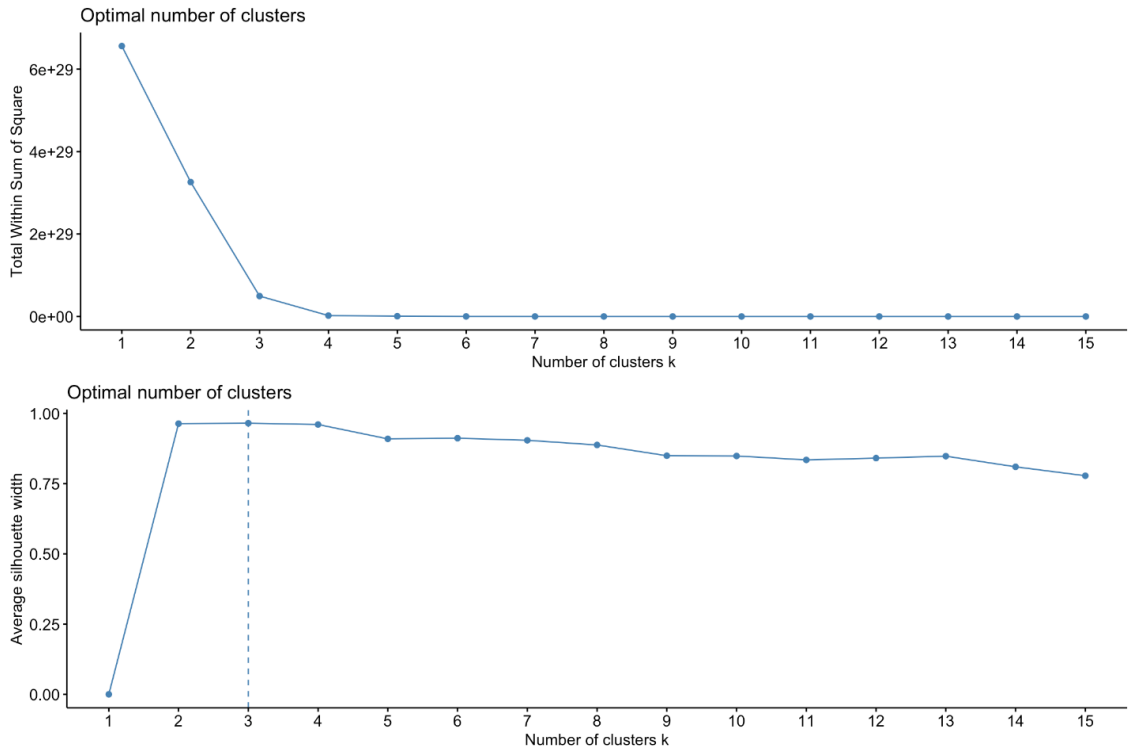

Figure 13: Assessment on the number of clusters. On top, the within sum of square for the total of clusters. Below, the average silhouette width.

As a 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%).

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%	

Figure 14: Profiles' analysis by features.

Based on the clusters' characteristics extracted from the descriptive model's results (shown above in Figure 14), what follows is a statistical translation of the profiles conceptually discussed during section 2.3.

<u>Cluster 1 – Low-income households:</u>

- 100% of consumers have over-indebtedness problems due to causes not related to the crisis.
- Medium-sized families (M = 2.65 people).
- Lowest income per capita (401.94 euros per month, Z-score mean = -0.34).
- Lowest total credit monthly installment (453.65 euros per month, effort rate = 40%, Z-score mean = -0.46).
- Lowest credit card monthly effort rate (149.54 euros per month, effort rate = 12%)
- Lowest housing credit monthly installment (M = 80.21 euros per month, effort rate = 6%)
- Lowest level of unemployment (6.6%) –12.6% below the dataset mean

- Mostly employed in the private sector (51.3% of the consumers, 7% above the dataset mean)
- Main cause of financial difficulty reported: increase in family members (12.8% of the households).

Cluster 2 – Low credit control households:

- Few crisis-related cases (16.04% of the observations)
    - Predominantly due to other causes (83.96%)
- Highest income per capita (686.35 euros per month, Z-score mean = 0.54)
- Lowest mean number of people in the household (M=1.78, Z-score mean = -0.48)
- Highest credit effort rate (M=75%, Z- score mean=0.27)
- Highest personal credit rate (246.00 euros per month, effort rate = 28%)
- Lowest car credit effort rate (19.88 euros per month, effort rate = 2%)
- Lowest household expenses (570 euros per month)

Cluster 3 – Crisis-affected households:

- Mostly crisis-related cases (83.7% of people)
    - Few due to other causes (16.3% of people).
- Low income per capita (413.15 euros per month, Z-score mean = -0.3)
- Largest families (2.76 people in the household)
- Highest household expenses (790.69 euros per month)
- Main causes for over-indebtedness:
    - Unemployment (40.5%), 21.3% higher than dataset mean;
    - Salary cuts (12.2%), 6% higher than dataset mean; and
    - Spouse's unemployment (8.4%), 4% higher than dataset mean. These
- Highest provision with housing (209.63 euros per month, effort rate = 20%, Z-score mean= 0.27)
- Highest provision with other credits or debts (79.54 euros per month, effort rate = 10%, Z-score mean 0.33)

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 statistically significant for education level ($\chi 2(4, 1455) = 0.9608$, $p = 0.9157$) nor years of education ($F(2, 1564) = 0.5813$, $p = 0.5593$, $\eta p2 = 7e\text{-}04$). The total income is also not statistically significant to distinguish the groups ($F(2, 1564) = 0.9568$, $p = 0.3844$, $\eta p2 = 0.0012$), only income per capita is statistically significant ($F(2, 1564) = 162.6146$, $p < 0.001$ $\eta p2 = 0.1721$).

3.3 SUPERVISED ML MODELING

For the goal of building the predictive model, this research followed an Automated Machine Learning (AutoML) approach. AutoML is a broad concept, possibly encompassing all Machine Learning phases, including, for instance, feature engineering. Within the scope of this study, the automated strategy applied focuses mainly on modeling steps – hyperparameter optimization and model selection. The motivation for this approach is to scale a traditionally manual routine of fine-tuning and comparing algorithms, increasing the chances of arriving at a favorable result. Conceptually, it generates a vast number of candidate models from numerous families of algorithms and finishes by selecting the best one among all. The process is detailed at subsection 3.3.2; however, for a better understanding, it is fitting to first introduce the ideas of hyperparameter optimization and cross-validation.

3.3.1 Hyperparameter optimization & Cross-validation

In Machine Learning, hyperparameters are the algorithm's parameters that are not directly learned within estimators and need to be defined prior to training. These actually define the settings for the estimator's learning process – e.g. the number of "trees" in a Decision Tree algorithm. A same algorithm may generalize different data patterns depending on its weights, constrains and learning rate. **Hyperparameter optimization** (also called tuning) is then the problem of choosing a set of optimal values for hyperparameters regarding a specific task; the one that yields the best results on the data being used.

In a manually handled scenario, tuning represents repeatedly selecting a combination of hyperparameter values, training the model and computing its performance. After sufficient examples are collected, the machine learning engineer compares each one and chooses the most appropriate for the task at hand. Needless to say, this approach is hard to scale.

In comparison, (automatic) hyperparameter optimization defines what is known as an objective function, which receives the hyperparameters' values as input and returns the model's associated performance. It then continues to select different combinations and defines as "best model" the one that maximizes the objective function's returned value; or the one that minimizes it, if performance is measured in terms of "least loss" instead of "greater score". In sum, through the concept of an objective function, hyperparameter optimization mimics a manual behavior, but leverages computing power to scale the number of models tested.

One common way to define the "inner-workings" of this objective function is through **cross-validation**, a category of model validation technique for assessing an estimator's generalization ability.

As a general concept, training a predictive model involves splitting the original dataset into two parts so the model can estimate parameters on the usually larger portion (training set) and, afterwards, evaluate its generalization ability on the remaining unseen data (testing set). However, questions such as "which fragment from the original set should be used for testing" may arise. Cross-validation, in turn, splits the entire data into several equal parts (e.g. 4), initially holding one of the fragments as test set (or validation set) and training on the remaining ones combined – e.g. 3 parts combined for training; 1 for validation). It then repeatedly selects a different part to be its testing set and trains on the other ones (figure 16 illustrates the process). It finishes by averaging the test results, deriving a more accurate estimate of model performance. As a side note, this method of splitting the data set into equal parts is known as *K-fold Cross-validation* ("k" representing the number of parts).

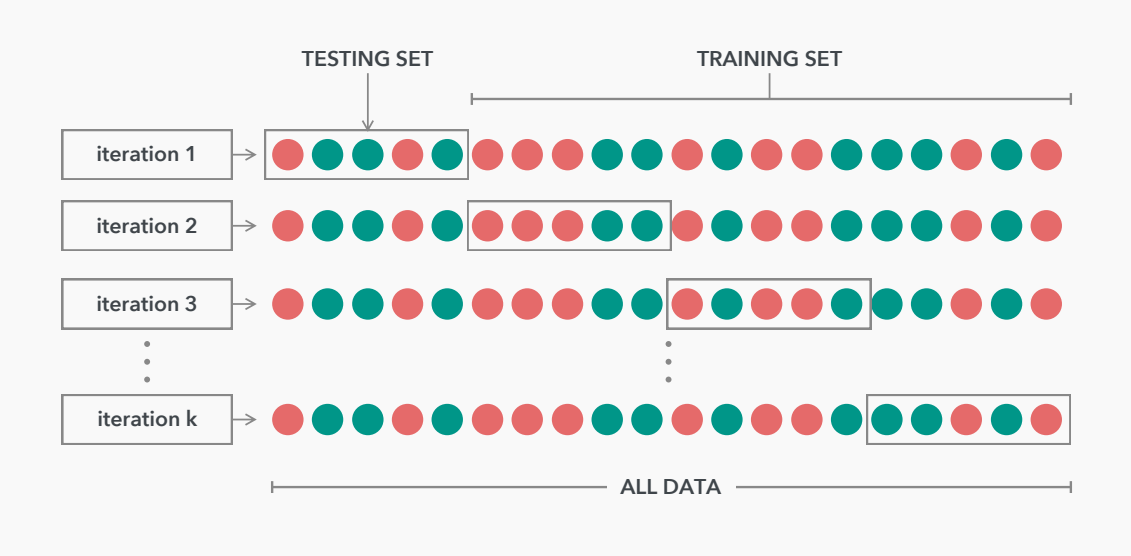

Figure 15: Conceptual example of cross-validation partitioning per iteration.

Therefore, in a hyperparameter scenario, some guiding principle selects several combinations of values and cross-validation returns averaged performances for each combination. By averaging multiple run tests, variability is reduced and, thus, makes for a more reliable method of comparing algorithm's hyperparameters. Also, a variant of K-fold cross-validation improves upon the described behavior, where each split fragment maintains the same distribution as the original dataset. It is called *Stratified K-fold Cross-validation*.

3.3.2 AutoML framework

Now, both concepts are combined to detail the modeling framework implemented in this study. As a refresher, the applied strategy employs an "inter-algorithmic" search – that is to say that it looks for the best model among several algorithms and their different hyperparameter combinations, not only within variations of the same algorithm. Below is a high-level description of the steps carried out during the process.

step 1. It starts by splitting the original data into training and test sets. Specifically, ~20% of the data (n=314) is kept away from the training process to serve later on as an unbiased source of comparison between the generated models.

step 2. For each algorithm in a list of algorithms, the following steps are applied.

step 2.1. Until a termination criterion is reached, a set of hyperparameter values is sampled and served to the objective function.

step 2.1.1. Using the algorithm's input configuration, stratified 5-fold cross-validation is applied on the data's remaining ~80% (n=1253) of observations. Figure 16 represents the updated data's partition during iterations. Each validation step is evaluated with the Log Loss metric, where smaller values are favored (to be further discussed).

step 2.2. After all relevant set of configurations are evaluated, the combination of hyperparameter values producing the best performing model is used to retrain the current algorithm on the entire training set (i.e. the 80% of observations). Thus, it becomes the algorithm's intermediary winner.

step 2.3. The intermediary winner is than evaluated against the unseen data. Again, Log Loss is computed, plus an additional score for empirical comparison (Accuracy Score, also to be detailed later on), and both are stored.

step 3. The process concludes by selecting amongst intermediary models the one that presented the best generalization ability on unseen data. Hence, the ultimate winning model.

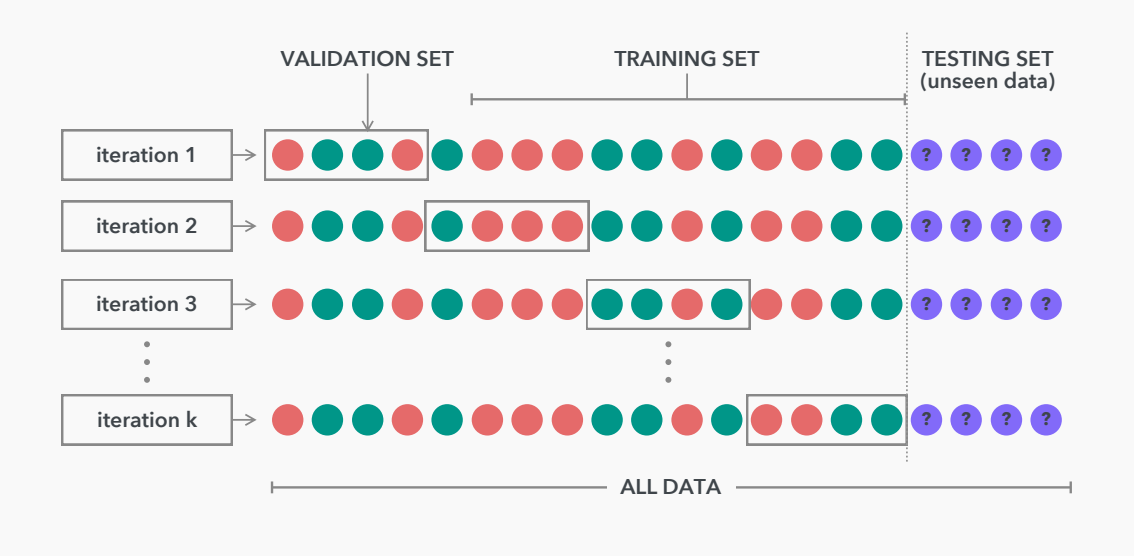

Figure 16: Conceptual example of data partitioning with initial "train / test split", followed by cross-validation.

Now that the general outline of the modeling framework was established, some gaps need to be filled regarding the metrics for model evaluation and the guiding principle for sampling hyperparameter values during training.

Starting with the evaluation of models' performances, two main metrics were used: log loss (logistic loss) and balanced accuracy score. Moreover, log loss was the defining metric throughout the automated selection between models. The rationale for choosing such metrics takes into consideration the metrics' nature and the characteristics of the task at hand.

Log Loss (logistic loss) – which in a multi-class scenario might be referred to as Cross Entropy – provides a probabilistic assessment, so it offers a nuanced view of each model's performance. It considers the uncertainty of a prediction and calculates how much it varies from the actual label. This makes for a good choice of metric when comparing models. Figure 17 illustrates the behavior.

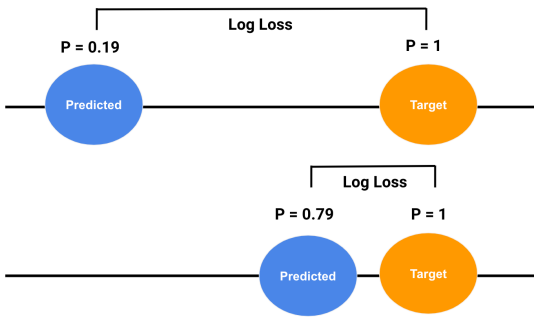

Figure 17: Log loss measuring a prediction's uncertainty based on the distance between the actual label and the predicted probability.

If the predicted probability diverges from the actual label, the log loss value increases, assuming a range between 0 and infinite. The classifier's objective is to minimize the returning log loss value, consequently, a perfect model would have a log loss equal to 0.

On the other hand, since Log Loss does not consider a strict interval, that makes it harder for an empirical assessment of the candidate models' general performance. In response, **Accuracy Score** was included. As it returns the fraction of correct predictions (division of total correct predictions over all observations) it is easy to assess and communicate, allowing the representation of performance in terms of percentage. Also, the target variable's balanced value counts for each cluster provides the freedom for using such simple metrics.

The second gap to be discussed is regarding the guiding principle for selecting hyperparameter values; in other words, the *sampler*. This topic has deliberately been referred to in abstract terms in order to highlight the difference between the following subsections. Subsections 3.3.3 and 3.3.4 describe the project's first results on model performance, garnered from a Grid Search sampling approach towards hyperparameter value selection. The subsequent section interjects a brief consideration on such results and bridges the reader to what came to be a second iteration within the project. To conclude, section 3.3.6 presents the process applied during this second phase – a Bayesian Hyperparameter Optimization – and results are reported in chapter 4 as the final predictive modeling outcome.

3.3.3 First iteration: Grid Search sampler

The hyperparameter optimization strategy used for the initial model was an exhaustive grid search. A brute-force approach where for each compared algorithm a grid of hyperparameters values is manually specified. Then, for each algorithm a model is fitted to the data with every possible combination of hyperparameters defined in the grid.

Grid Search's experimental settings

In the case of having Grid Search as a sampler, the AutoML's termination criterion is reached when all algorithms and all possible combinations created from its configuration options are covered. For the first iteration, the classification algorithms evaluated were:

- Logistic Regression
- Gaussian Naïve Bayes

- K Nearest Neighbors
- Linear Discriminant Analysis
- Decision Trees
- Random Forest
- Extra Trees
- Gradient Boosting
- Support Vector Machine
- Nu-support Vector Machine

As a side note, while the use of Deep Learning was proposed in recent literature on ML for social good (Khatua, Cambria, & Khatua, 2018; Sawhney et al., 2018; Al-Hashedi, Soon, & Goh, 2019), those studies applied it to vast amounts of unstructured data – a scenario where Deep Learning thrives. However, the problem dealt with in this study is characterized by a limited amount of tabular data. Therefore, Deep Learning ends up being a slow algorithm to train that yields poor results; consequently, the technique was not included in the search.

Each classifier has a different set of parameters in accordance with the algorithm's design. Figure 18 shows the algorithms listed above followed by the numerous hyperparameters tested with Grid Search.

Figure 18: Algorithms and hyperparameters evaluated during exhaustive grid search.

3.3.4 ***First iteration: results***

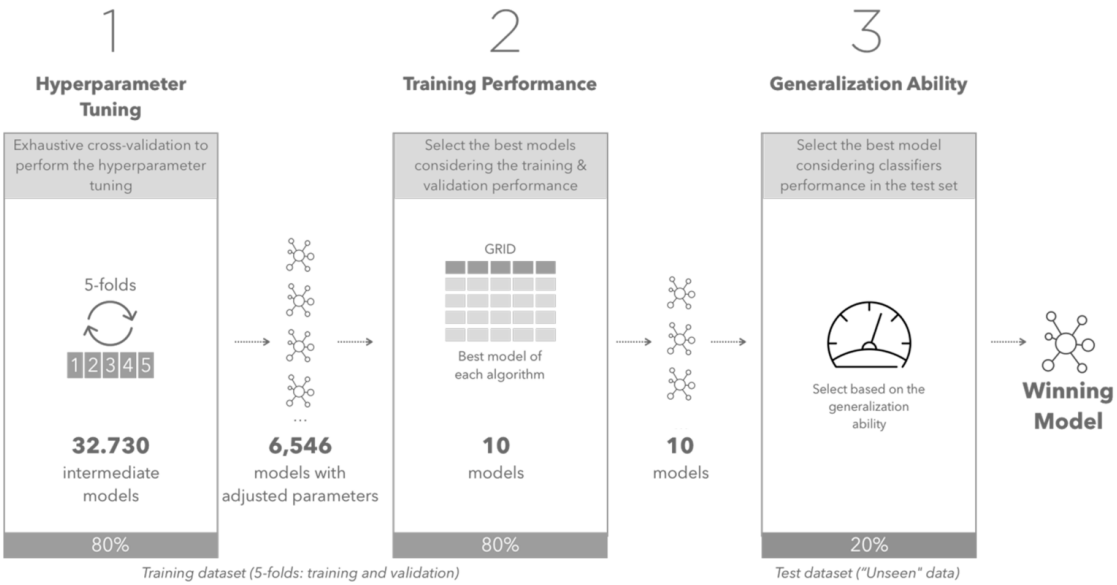

Figure 19: Grid Search Hyperparameters Tuning Process

Figure 19 presents the evolution throughout model evaluation with Grid Search. The 5-fold cross-validation routine within the hyperparameter tuning phase fitted 32,730 intermediate models, creating 6,546 candidate models in total – i.e. for each hyperparameter combination.

The performance of each intermediate winner classifier (i.e. best per algorithm) is presented in Figures 20 and 21. During this analysis, 4 algorithms were identified as poor performers and, as will be discussed, this served as prior knowledge for the second iteration. The methods were: (6) Decision Trees, (7) Gaussian Naïve 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 Decision Trees algorithm to learn and express these features.

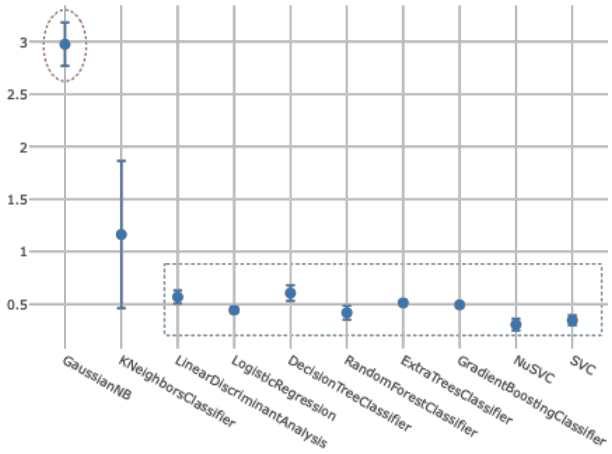

Figure 20: Log Loss results on validation set.

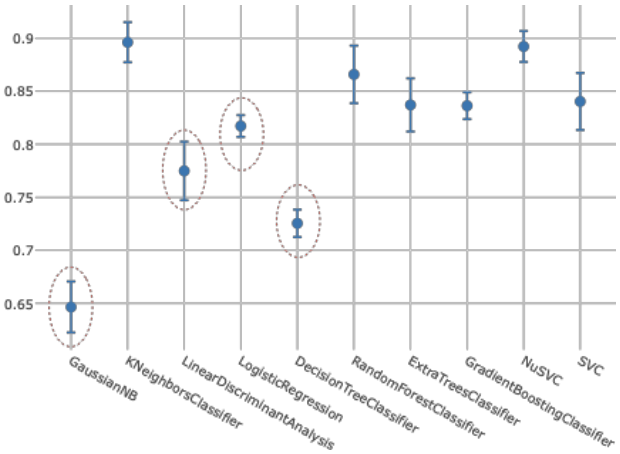

Figure 21: Accuracy score results on validation set.

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. The intermediary winning models were then assessed on their generalization ability by evaluating their performance on the test set (unseen data). Figure 22 shows a comparison of their performance using Accuracy Score and Figure 23 on Log Loss.

Winning model

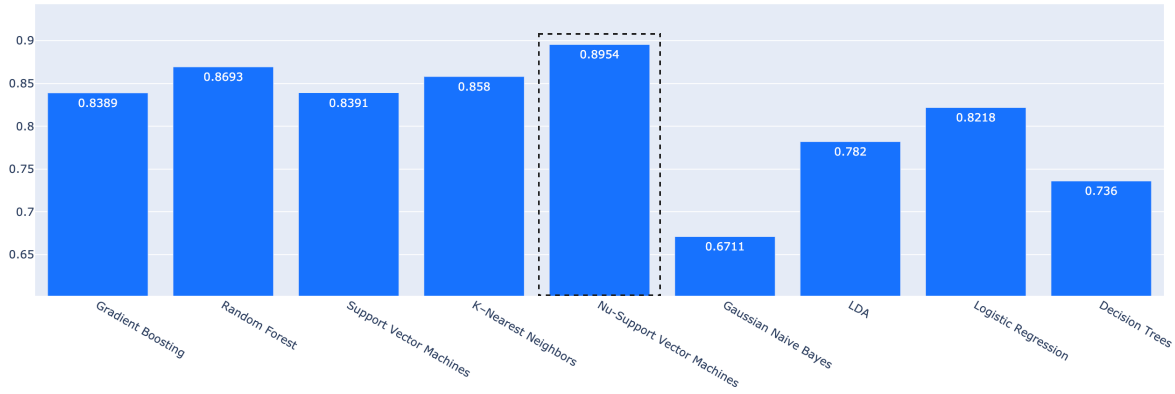

Figure 22: Accuracy score of best models (per algorithm) on unseen data.

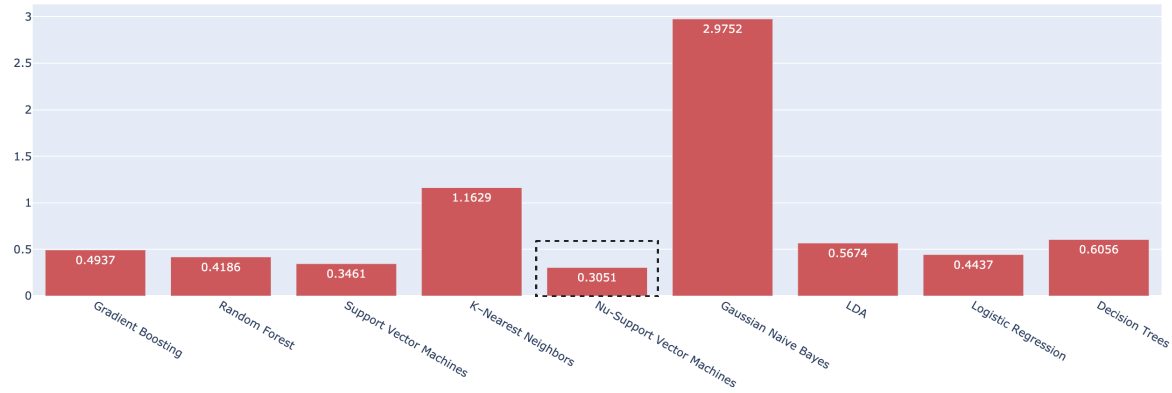

Figure 23: Log Loss of best models (per algorithm) on unseen data.

The classification algorithm that generated the best model after the exhaustive grid search was a variant of the Support Vector Machine (SVC) algorithm, the Nu-SVC. The Nu-SVC is similar to SVC but uses a regularization parameter to control the number of support vectors, which implements a penalty on the misclassifications that are performed while separating the classes. Given Nu-SVC is based on SVC, in the next section SVC's functioning will be described. The best-fitting model of Nu-SVC generated automatically had the following parameters (see Table X 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.

Description of 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 (by

means of 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).

3.3.5 Considerations on first iteration

Each predictive modeling problem is unique and therefore a model's performance is relative. It extensively depends on the data being served to the algorithm, the task's complexity and the model's practical use. Reasoning solely from accuracy-based performance, the results yielded by grid search and its winning model appear satisfactory (89.5%). Especially when compared to other predictive modelling applied to similar business scenarios, as the ones reported in section 2.2.2; Agarwal et al. (2018) with 68.7% accuracy and Alomari (2017) with 71.75% accuracy.

However, when considering a real-world scenario, the concept of error deserves a second look. For instance, in the case of ACP's internal software, if the model predicts "overspending" as the main contributor to over-indebtedness, their policy determines the case should be rejected. Thus, around 10.5% of the households that contact ACP runs the risk of not receiving help or, at least, not the appropriate assistance. In a sample of 1000 cases, for example, that translates to 105 families experiencing financial hardships without much needed guidance.

This alternative perspective on performance, instigated during the software's ideation, made the case for a second iteration on the predictive model in charge of defining consumers' over-indebtedness profiles. A "revision", so to speak, on the process of searching the best model. One that prior evidence has shown to produce better results than the Grid Search method used initially.

3.3.6 Second iteration: Bayesian Optimization sampler

For the second iteration, a Bayesian Optimization method (specifically, a Tree-Structured Parzen Estimator) (Bergstra et al., 2011) was applied for sampling. Differently from the brute-force approach Grid Search follows, Bayesian Optimization tries to balance exploration

(hyperparameters with uncertain outcome) and exploitation (hyperparameters that are expected to provide optimal results). It has been shown to obtain faster and better results than Grid Search due to its ability to reason about the outcomes prior to executing a run (Bergstra et al., 2011). Conceptually, it builds a probabilistic model of a function mapping hyperparameter values to the task's objective. It starts by randomly selecting initial values for hyperparameters and at each iteration (in this scenario, that is, at each cross-validation execution) it updates the probabilistic model.

Bayesian Optimization experimental settings

On this second experiment, the algorithms line-up was updated considering prior knowledge obtained from Grid Search's outcome on the task. First, due to poor performance, 4 algorithms were not included (Gaussian Naïve Bayes, Decision Trees, Linear Discriminant Analysis and Logistic Regression). In this sense, more time can be dedicated to searching hyperparameter values of promising algorithms. Still, in hopes of further reducing execution time and, to some extent, redundancy, a fifth algorithm was excluded. The previous line-up had 3 Decision-Tree-based ensembles – Random Forest, Gradient Boosting and Extra Trees. Since Random Forest was the most performant among the 3 and considerably similar to Extra Trees (Geurts et al., 2006), the latter was removed. If based solely on performance, Extra Trees should have been maintained in place of Gradient Boosting; nonetheless, Gradient Boosting was in fact substituted for a different implementation – i.e. XGBoost, a regularized variant. In sum, the final line-up constitutes of the following algorithms:

- K–Nearest Neighbors
- Random Forest
- Support Vector Machines
- NU-Support Vector Machines
- XGBoost (i.e. Gradient Boosting)

Regarding hyperparameters, the same as Grid Search were tested. Only, instead of manually inputting the static values to be evaluated, intervals were defined for the Bayesian optimizer to sample from.

One aspect of Bayesian Optimization that differs from Grid Search, and should be taken into consideration, is its random initialization. For avoiding the possible negative effect of random initialization leading to local optimal solutions, the Bayesian Hyperparameter sampling was executed 3 times per algorithm. On that note, each algorithm was run 300 times – i.e. 3 x 300 per algorithm, which equals a total of 4500 runs for all five.

4 FINAL RESULTS

This chapter presents the results of the second iteration on the predictive model. It shows performance outcomes for the winning model, the intermediate models and briefly compares to associated results from the first iteration. It completes by exposing the elected hyperparameter values for each wining algorithm.

4.1 GENERAL RESULTS

After 300 runs per algorithm, each run with a different sampled hyperparameter values, Figure 24 exposes the performance of each best model on unseen data. Gradient Boosting came on first, with approximately 92.1% score on balanced accuracy, followed by Nu-Support Vector Machines (~91.5%) and Support Vector Machines (~90.7%).

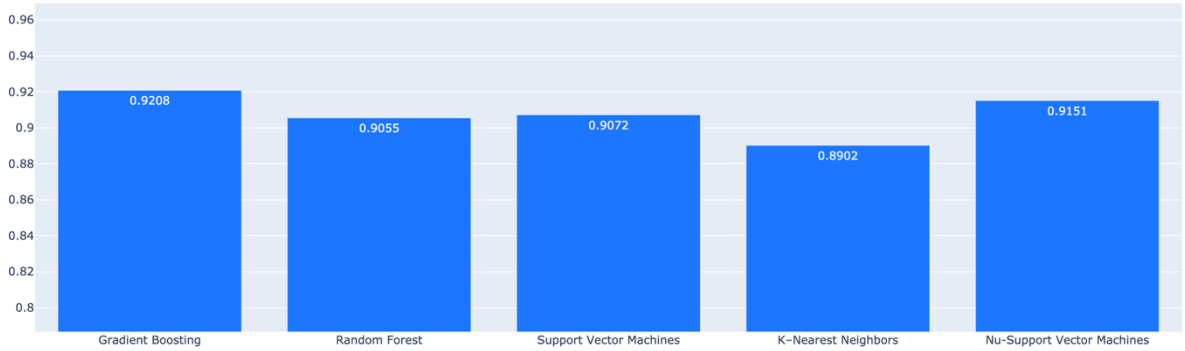

Figure 24: Best models' accuracy score on unseen data.

Log loss (Figure 25) reaffirms the winning model; however, it differs among the other best models. Random Forest now comes in second, followed by Nu-Support Vector Machine. On both metrics Support Vector Machine was surpassed by its alternative implementation, indicating that Nu-Support Vector Machine is a better fit for this specific task. K-Nearest Neighbors lags behind on both accounts.

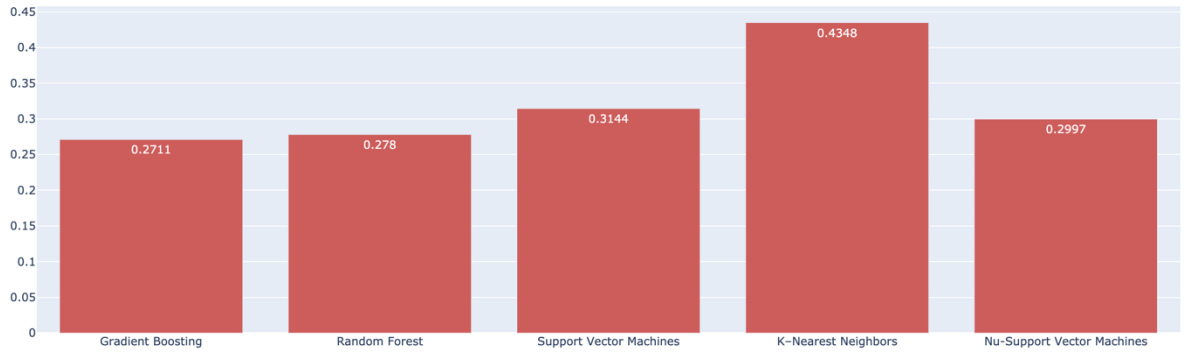

Figure 25: Best models' log loss value on unseen data.

Relating to the previous iteration, all models showed an improvement, as seen in Figure 26. Specifically, a mean increase of 4.76% considering all algorithms.

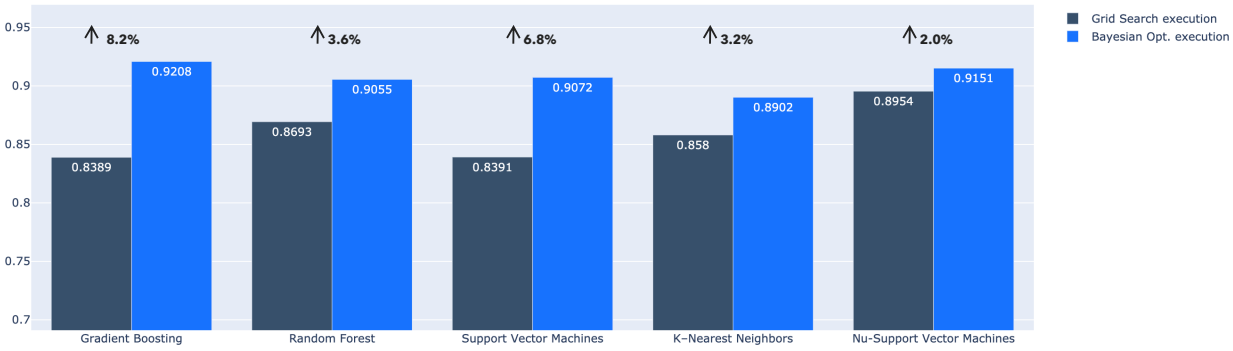

Figure 26: Comparison of associated best models per project iteration. Accuracy score on unseen data.

Similarly, only now in regard to minimizing loss, best models per algorithm also reduced their Log Loss (Figure 27). A total mean reduction of 0.2258 points in Log loss scores.

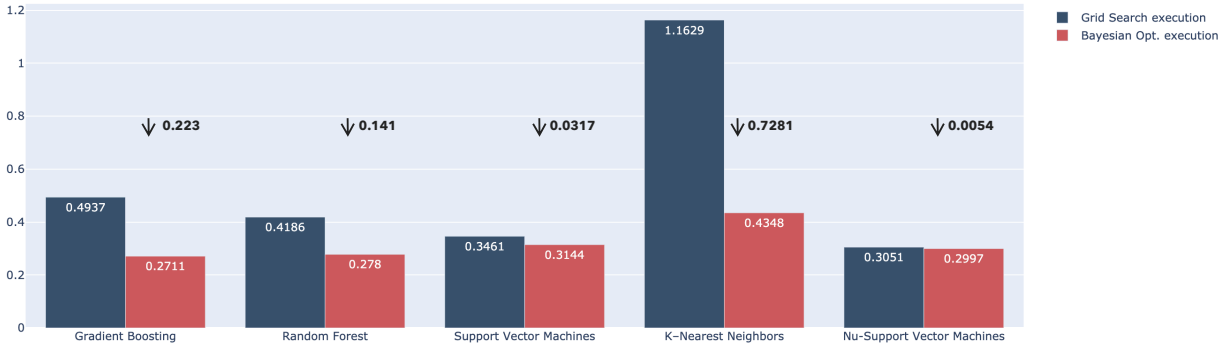

Figure 27: Comparison of associated best models per project iteration. Log Loss score on unseen data.

4.2 WINNING MODEL: GRADIENT BOOSTING (XGBOOST)

Gradient Boosting, specifically the XGBoost implementation, was the best model on unseen data after the second iteration. Gradient Boosting Machines (GBMs) (Friedman, 2001) is a supervised machine learning technique that works as an ensemble of weak prediction models, usually Decision Trees, in order to combine all their results and form a strong model. XGBoost (Chen & Guestrin, 2016), simply put, is a more regularized form of Gradient Boosting, using advanced regularization – i.e. L1 and L2 – to improve the model's generalization capability.

Its best model was achieved at trial 164 of the hyperparameter optimization execution (seen in Figure 28). Below are the sampled hyperparameter values and their relevance.

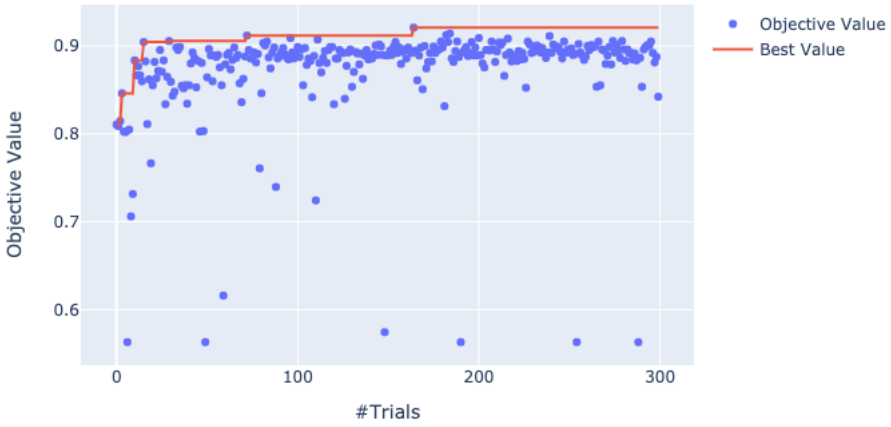

Figure 28: Optimization history plot for Gradient Boosting.

4.2.1 Gradient Boosting (XGBoost) elected hyperparameters values

Table 2 presents Gradient Boosting's optimized hyperparameters and their corresponding values returned by the Bayesian sampler. The algorithm itself and its specific implementation has other hyperparameters that can be found at the library's official documentation website – 'https://xgboost.readthedocs.io/en/latest/parameter.html'. All secondary hyperparameters (i.e. not included below) were defined by their default value.

Name	Value	Description
booster	gbtree	Defines which booster to use. Chooses between *gbtree*, *gblinear* or *dart*. *gbtree* and *dart* use tree-based models while *gblinear* uses linear functions.
eta	0.9675	Step size shrinkage used to prevent overfitting. After each boosting step, the weights of new features can be returned and eta shrinks the feature weights to make the boosting process more conservative. It expects a value between 0 and 1.
gamma	9.443e-08	Defines the minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm will be. It expects a value between 0 and infinity.
max_depth	8	Defines the maximum depth of a tree. Increasing this value makes the model more complex and more likely to overfit. It expects a value between 0 and infinity.

lambda	0.8152	L2 regularization term on weights. The higher the value, the more conservative the model is.
alpha	7.884e-06	L2 regularization term on weights. The higher the value, the more conservative the model is.
grow_policy	depthwise	Controls the way in which new nodes are added to the tree. In this case, *depthwise* defines splits at nodes closest to the root.

Table 2: Selected hyperparameter values for Gradient Boosting after optimization. Descriptions were based on XGBoost's official documentation.

Below, Figure 29 shows the importance of hyperparameters for Gradient Boosting. Only 3 are included – this leads to the assumption that the ones excluded from the graph did not surpass a minimum threshold for its inclusion to be relevant. Yet, between the included, tuning of the *booster* parameter is supposed to affect the most changes in objective value.

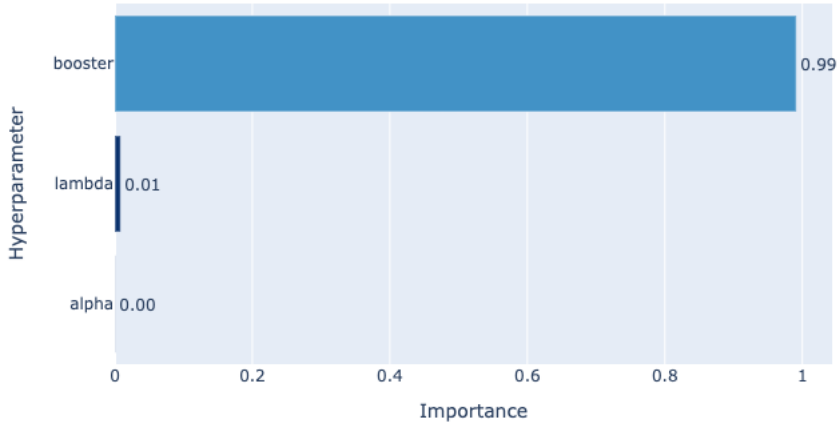

Figure 29: Hyperparameters' importance for Gradient Boosting

Figure 30 presents the hyperparameters individually, relating their values to the objective value yielded on the associated trial. In fact, it shows that, with a few exceptions, most hyperparameters do not seem to influence the objective value. For its most part, hyperparameter values are concentrated in the upper part of the graph. The *booster* parameter, however, does indicate that one of its categorical values (i.e. *gblinear*) is related to a considerable drop on performance. In that respect, one might question if the slight relationship with improving objective values observed in *max_depth*, for instance, could have a greater importance in case the *gblinear* value from parameter *booster* was removed from the options.

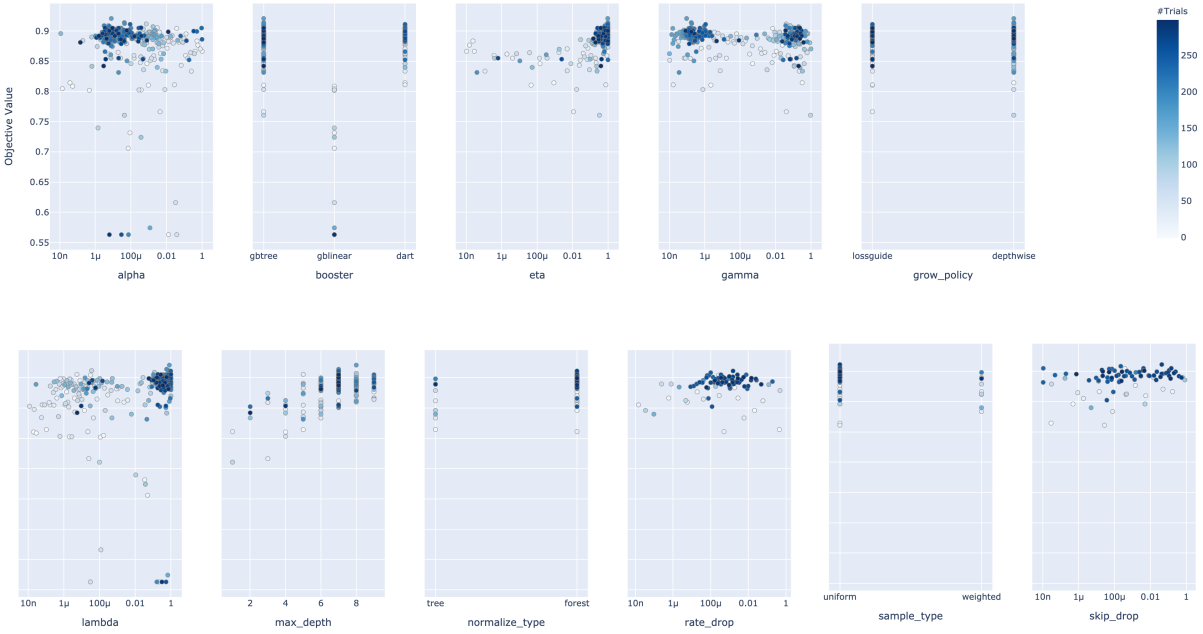

Figure 30: Individual hyperparameters of Gradient Boosting against objective value.

To conclude on the winning model, Figure 31 presents trials' sampled values against the objective value. Admittedly, it is a difficult plot to interpret statically – i.e. without being able to interact with the strings and analyze their connections. Nonetheless, it is presented here to serve a point on showing hyperparameters relationship not only in regard to the objective value, but also among themselves. Even though some hyperparameters seem to concentrate on specific values, others present a myriad of spread out connections.

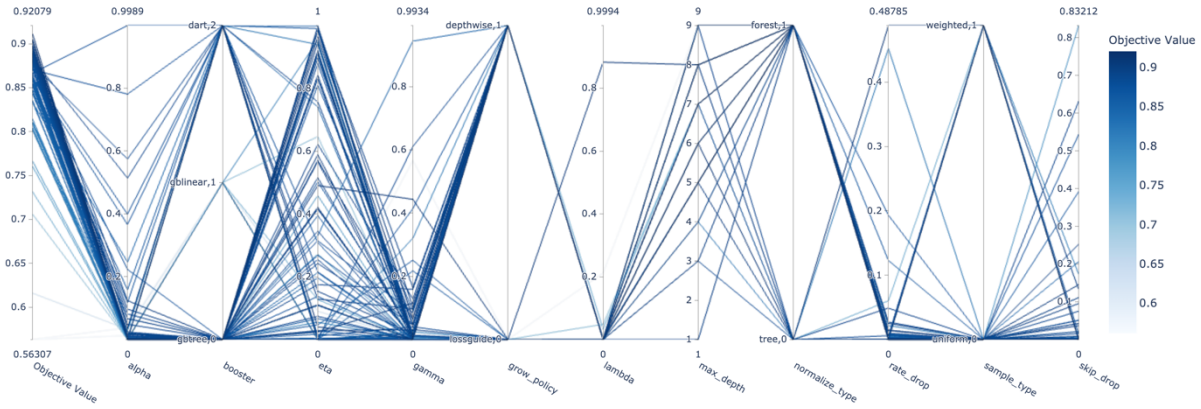

Figure 31: Parallel coordinates of hyperparameters connected by trials exposing the sampled values against the objective value – Gradient Boosting

4.3 INTERMEDIATE WINNING MODELS

This section briefly presents the same hyperparameter analysis of Gradient Boosting for the winning models of the 4 remaining algorithms. They were grouped by the analysis type, in order to focus on a high-level discussion regarding the optimization process as a whole, instead of concentrating on each algorithm's specific details.

4.3.1 Objective values' evolution

As previously mentioned, every algorithm was run 300 times. Random Forest's best model was built on trial 156 (Figure 32). K-Nearest Neighbors' on trial 73 (Figure 34). Support Vector Machine's on trial 219 (Figure 33) and Nu-Support Vector Machine's on trial 203 (Figure 35). With all algorithms, the objective value (in this case, accuracy score) quickly arrives at a high score and plateaus for the remainder of the execution, with only minor improvements.

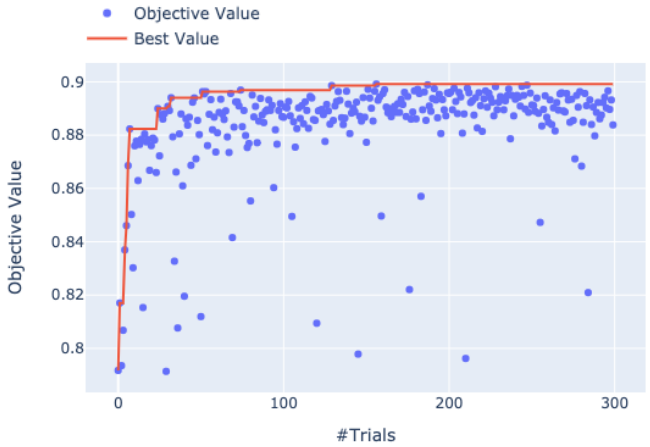

Figure 32: Optimization history plot for Random Forest.

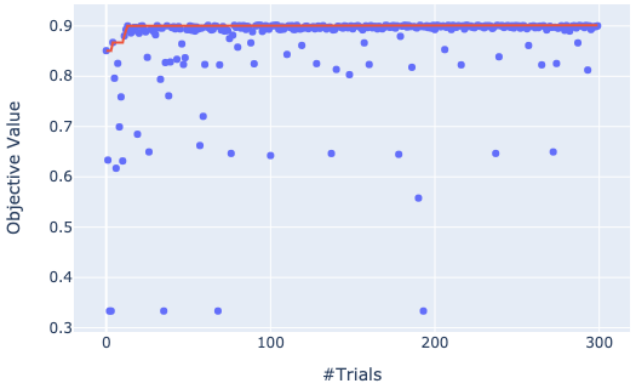

Figure 33: Optimization history plot for Support Vector Machine.

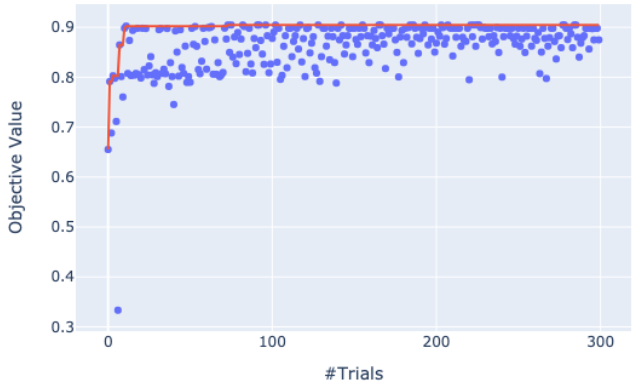

Figure 34: Optimization history plot for K-Nearest Neighbors.

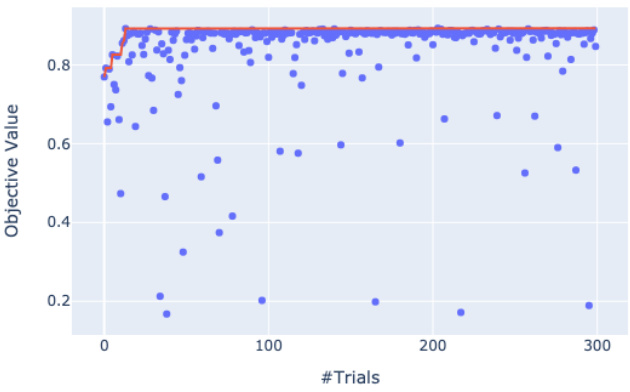

Figure 35: Optimization history plot for Nu-Support Vector Machine.

4.3.2 Hyperparameter values

Tables 3, 4, 5 and 6 present the hyperparameter values elected by the Bayesian sampler for each algorithm. Just as with Gradient Boosting, the algorithms have other hyperparameters possible of tuning; however, the most prominent were chosen for optimization.

RANDOM FOREST

Name	Value	Description
n_estimators	29	Specifies the number of trees in the forest.
criterion	entropy	Defines the function to measure the quality of a split. Chooses between *gini* and *entropy*.
max_depth	41	The maximum depth of the tree.
min_samples_split	2	The minimum number of samples required to split an internal node.
min_samples_leaf	1	The minimum number of samples required to be at a leaf node.
max_features	sqrt	The number of features to consider when looking for the best split. Chooses between *sqrt* (i.e. the square root of the number of features), *log2* (i.e. the $\log_2$ of the number of features), and the actual number of features.
bootstrap	False	Defines whether bootstrap samples are used when building trees. If False, the whole dataset is used to build each tree.

Table 3: Selected hyperparameter values for Random Forest after optimization. Further description and other hyperparameters can be found at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

K-NEAREST NEIGHBORS

Name	Value	Description
n_neighbors	4	Number of neighbors to use as classifying threshold.
weights	distance	The weight function used in prediction. Chooses between *uniform* (i.e. all points in each neighborhood are weighted equally) and *distance* (weight points by the inverse of their distance. In the case of *distance*, closer neighbors of a

		query point will have a greater influence than neighbors which are further away.
algorithm	kd_tree	Algorithm used to compute the nearest neighbors. Chooses between a brute-force algorithm, a Ball Tree algorithm and a KD Tree algorithm.

Table 4: Selected hyperparameter values for K-Nearest Neighbors after optimization. Further description and other hyperparameters can be found at: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

SUPPORT VECTOR MACHINE

Name	Value	Description
C	36.895	Regularization parameter, where the strength of the regularization is inversely proportional to the value of C. Applies a squared l2 penalty.
kernel	rbf	Defines the kernel type to be used in the algorithm. Chooses between *linear*, *poly*, *rbf*, and *sigmoid*.
degree	null	Annulled. Only activated if *kernel's* value is *poly*.
gamma	scale	Specifies the kernel coefficient for *poly*, *rbf*, and *sigmoid*. Chooses between *scale* (i.e. "1÷ number of features × input's variance") and *auto* (i.e. "1 ÷ number of features").
shrinking	True	Whether to use shrinking heuristic.
class_weight	None	If defined, it sets the parameter *C* of class i to "$class\_weight_{\text{i}}$ × *C*". If not given, all classes have weight one.

Table 5: Selected hyperparameter values for Support Vector Machine after optimization. Further description and other hyperparameters can be found at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

NU-SUPPORT VECTOR MACHINE

Name	Value	Description
nu	0.0883	A modified C parameter to create an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors. Its interval is between (0, 1].
kernel	rbf	Defines the kernel type to be used in the algorithm. Chooses between linear, poly, rbf, and sigmoid.

degree	null	Annulled. Only activated if kernel's value is poly.
gamma	scale	Specifies the kernel coefficient for poly, rbf, and sigmoid. Chooses between scale (i.e. "1÷ number of features × input's variance") and auto (i.e. "1 ÷ number of features").
shrinking	True	Whether to use shrinking heuristic.
class_weight	None	If defined, it sets the parameter *nu* of class i to "class_weight$_i$ × *nu*". If not given, all classes have weight one.

Table 6: Selected hyperparameter values for Nu-Support Vector Machine after optimization. Further description and other hyperparameters can be found at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.NuSVC.html#sklearn.svm.NuSVC

4.3.3 Hyperparameters' importances

As with Gradient Boosting, one or two specific hyperparameters dominate in terms of importance for every algorithm, as can be observed in Figures 36, 37, 38, and 39. That is, one hyperparameter accounts for most of the variance in the objective value's outcome. *min_samples_leaf* for Random Forest, *weights* and specially *n_neighbors* for K-Nearest Neighbors, *kernel* for Support Vector Machine, and *kernel* and *nu* for Nu-Support Vector Machine.

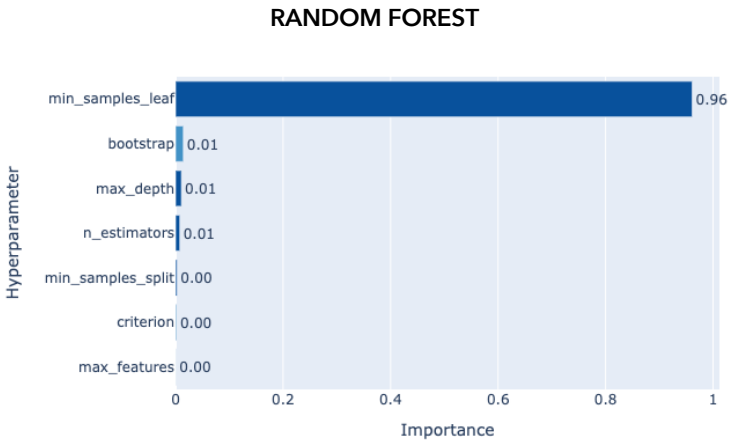

Figure 37: Hyperparameters' importance for Random Forest.

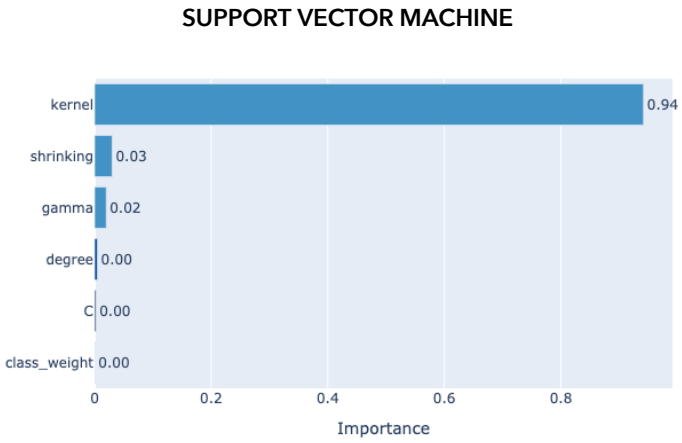

Figure 36: Hyperparameters' importance for Support Vector Machine.

K-NEAREST NEIGHBOR

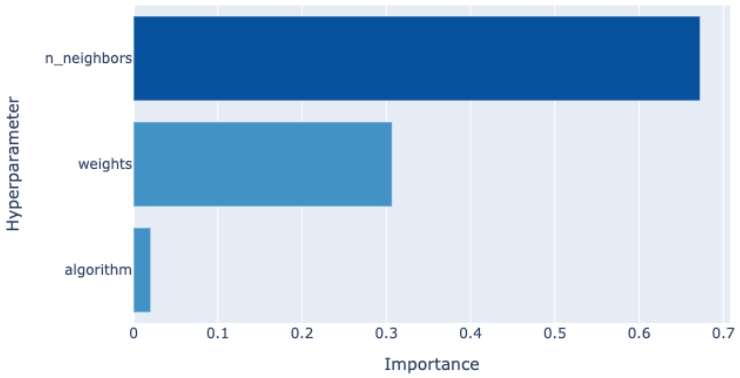

Figure 38: Hyperparameters' importance for K-Nearest Neighbors.

NU-SUPPORT VECTOR MACHINE

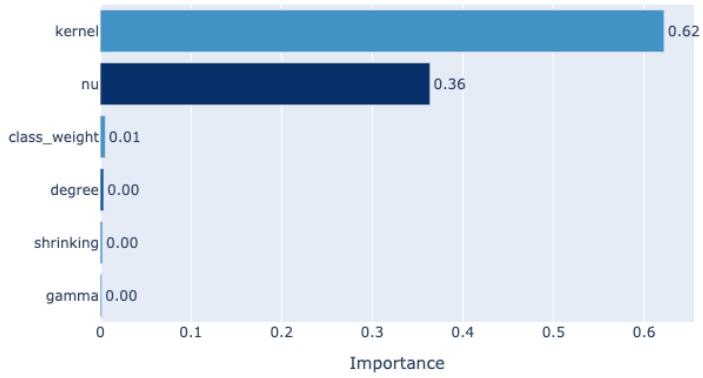

Figure 39: Hyperparameters' importance for Nu-Support Vector Machine.

4.3.4 Relation between individual hyperparameters and objective value

Again, as seen with Gradient Boosting's model, this subsection presents how each individual hyperparameter range of values relates to a trial's produced objective value. Variance is easily identified in the "important" hyperparameters listed in the previous subsection. In Figure 40, *min_samples_leaf* shows a drastic improvement (i.e. on Accuracy Score) towards smaller values. Likewise, *n_neighbors* in Figure 41 also seems to have a positive impact on the model's accuracy when its value drops below 100 neighbors. In Figure 42, the Bayesian sampler seems to identify *rbf* as a preferred *kernel* method and *sigmoid* as having the smallest contribution within the 4 options. For Nu-Support Vector Machine, that is, in Figure 43, *kernel* also indicates *rbf* as a clear winner among the relating methods. However, *sigmoid* is not necessarily the worst candidate, yielding better trials than that of the *linear* method.

Apart from the important ones mentioned above, a relevant observation to make is regarding other less obvious hyperparameters. For instance, *min_samples_split* and *C (*in Figures 40 and 42, respectively) do not, necessarily, present a clear trend regarding best and worst value intervals. However, it is possible to identify a "general preference" towards a specific interval or, in the other hand, a "less favorable" section of values. In *min_samples_split*, there were trials with small values that did produce a low Accuracy score. Nonetheless, a significant majority of high Accuracy trials do concentrate at the hyperparameter's values below 20. In the same respect, *C* exposes a broad range of promising values (e.g. from 1 to 10,000), and further investigation would be required to narrow down the best interval. Still, values below 1 seem to present a downwards trend; the smaller value of *C*, the lower is the trial's Accuracy score. Just as with the more prominent

hyperparameters, the identification of promising intervals (or removal of unpromising values) can help in reducing the optimizer's search space.

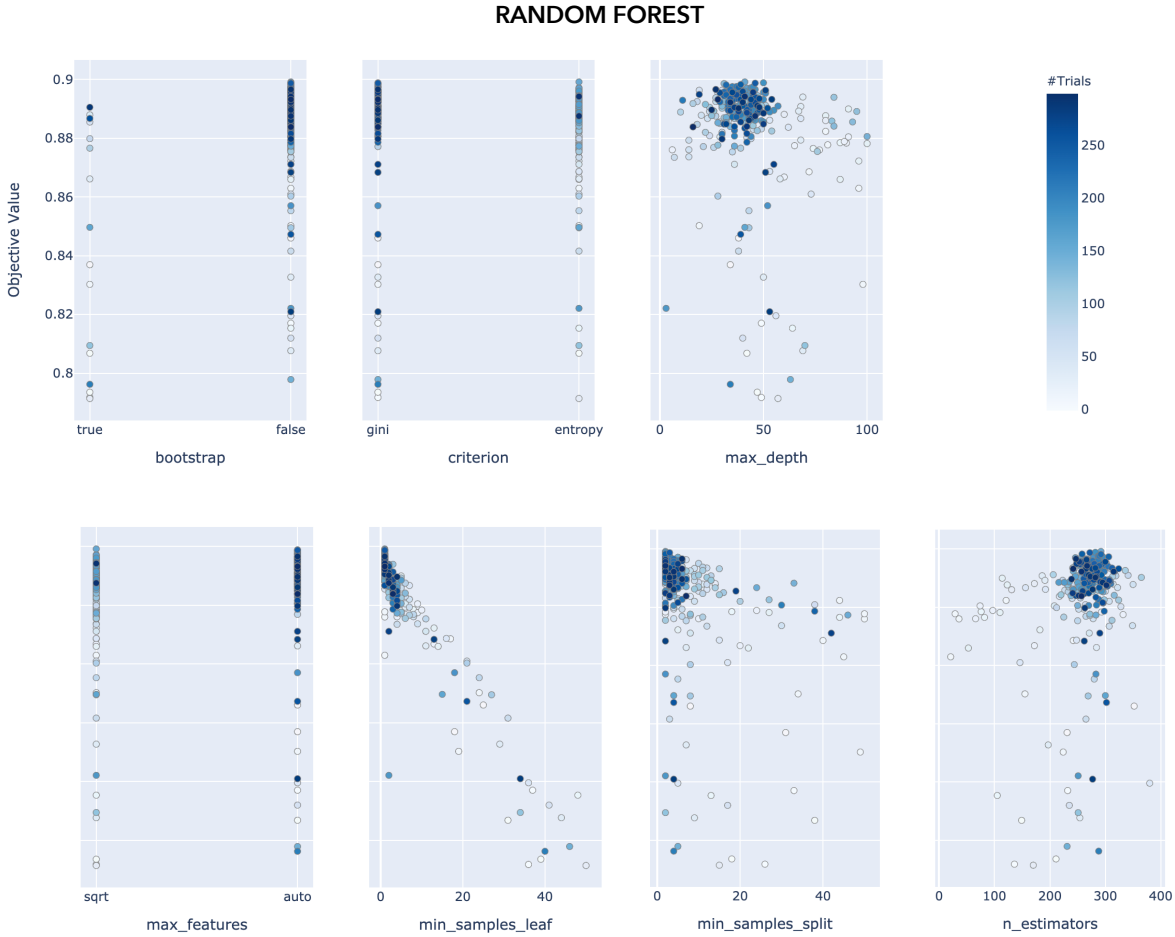

Figure 39: Individual hyperparameters of Random Forest against objective value.

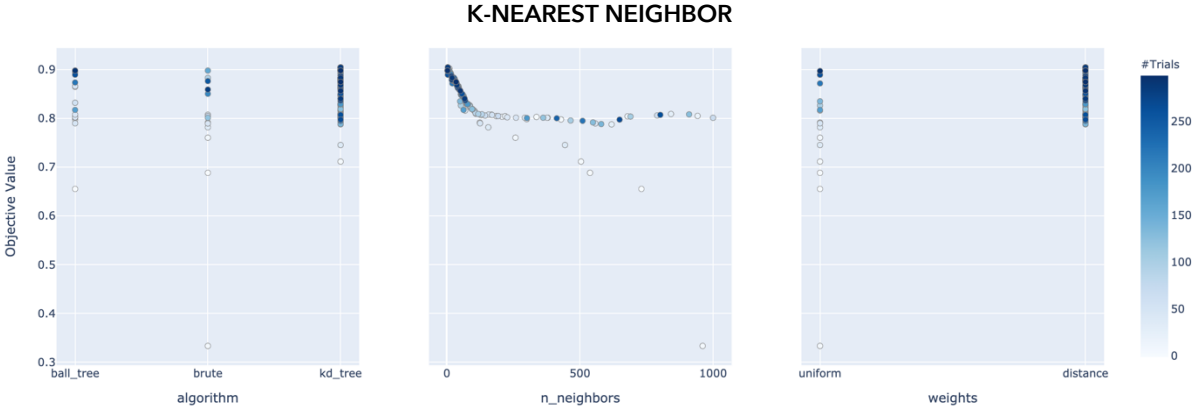

Figure 40: Individual hyperparameters of K-Nearest Neighbors against objective value.

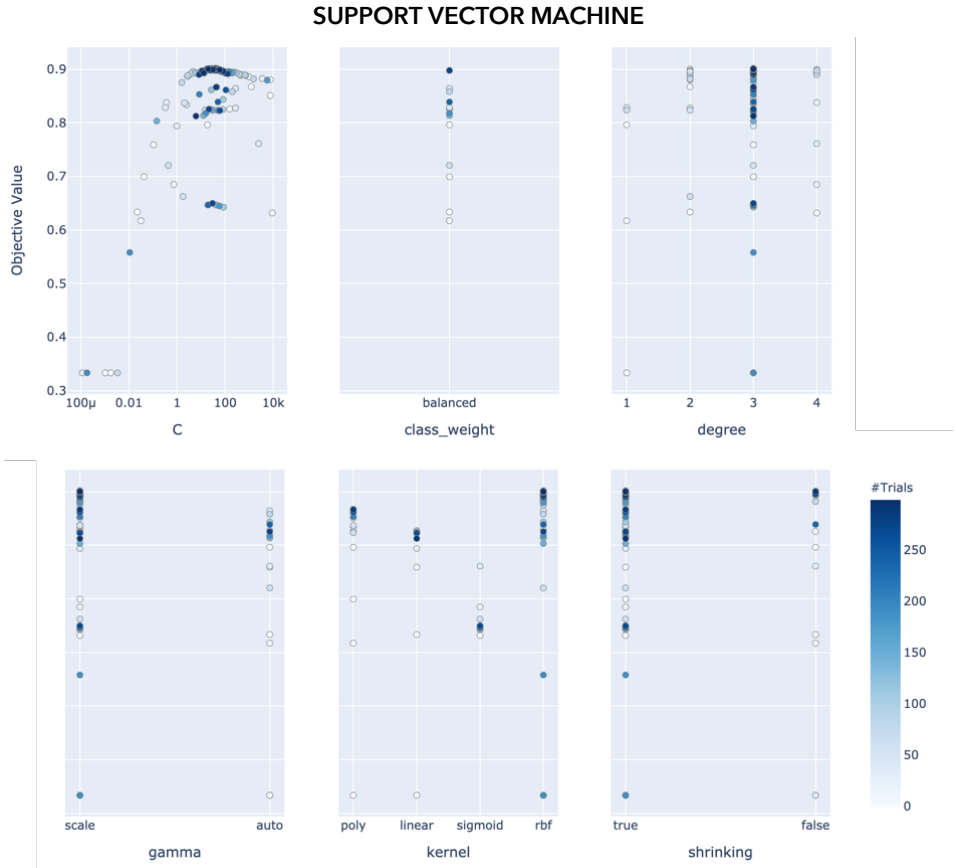

Figure 41: Individual hyperparameters of Support Vector Machine against objective value.

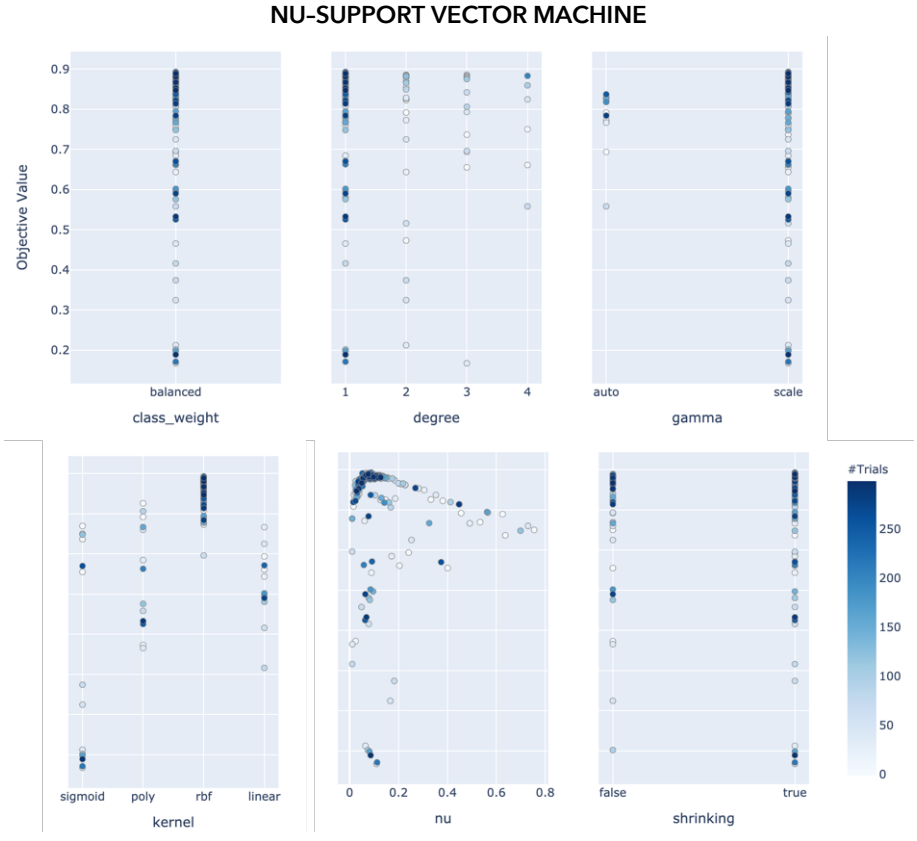

Figure 42: Individual hyperparameters of Nu-Support Vector Machine against objective value.

4.3.5 *Parallel Coordinates of Hyperparameters*

To finish chapter 4, Figures 44, 45, 46, and 47 present, per algorithm, the parallel coordinates connecting hyperparameters' values with the trials' Accuracy score. As with Gradient Boosting, the general objective is to observe the relationship among hyperparameters. The main idea to take from these visualizations is the myriad of possible combinations that lead to a satisfactory objective value. The hyperparameters with reported high importance do seem to concentrate high yielding trials on specific values (or range of values). However, apart from those, the remaining hyperparameters present trials with good Accuracy score coming from a multitude of possible values. Also, the visualization reinforces the perspective of an intertwined relationship among hyperparameters, instead of an independent one.

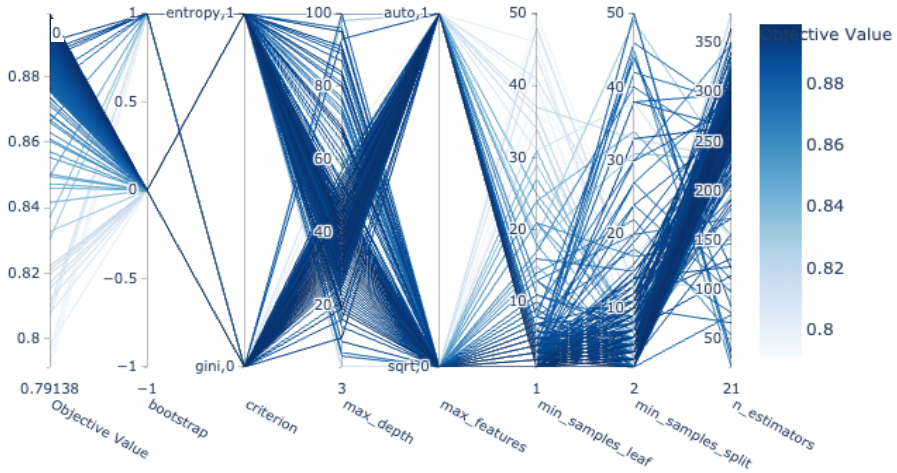

Figure 43: Parallel coordinates of hyperparameters connected by trials exposing the sampled values against the objective value – Random Forest.

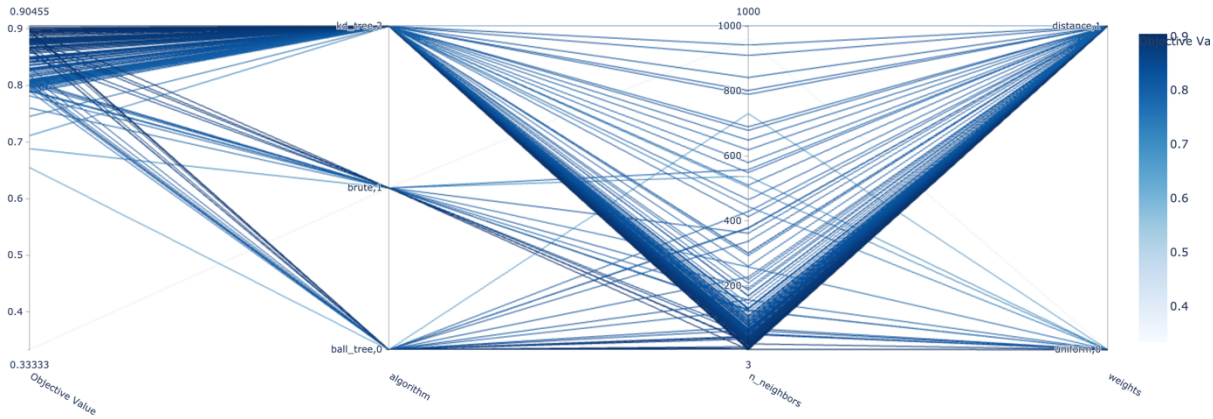

Figure 44: Parallel coordinates of hyperparameters connected by trials exposing the sampled values against the objective value – K-Nearest Neighbors.

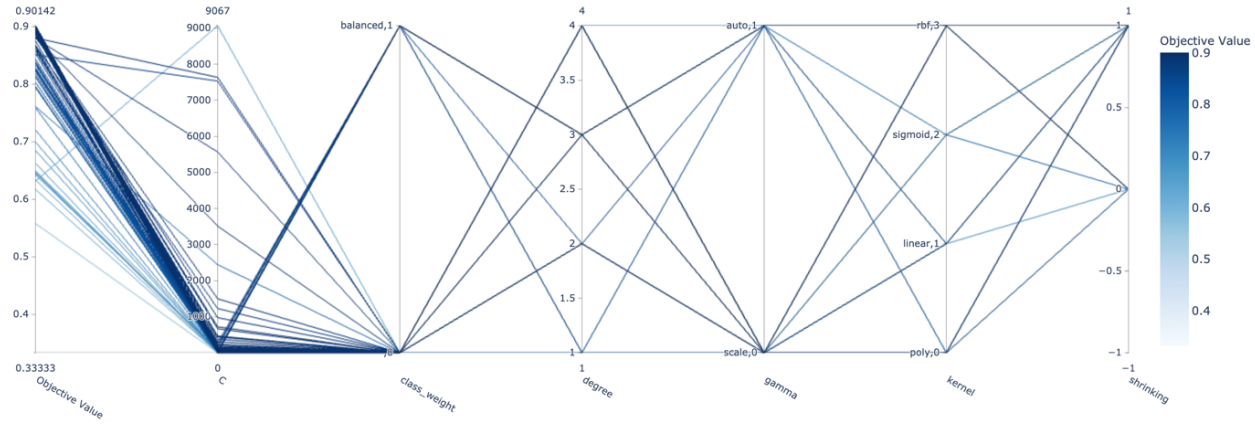

Figure 45: Parallel coordinates of hyperparameters connected by trials exposing the sampled values against the objective value — Support Vector Machine.

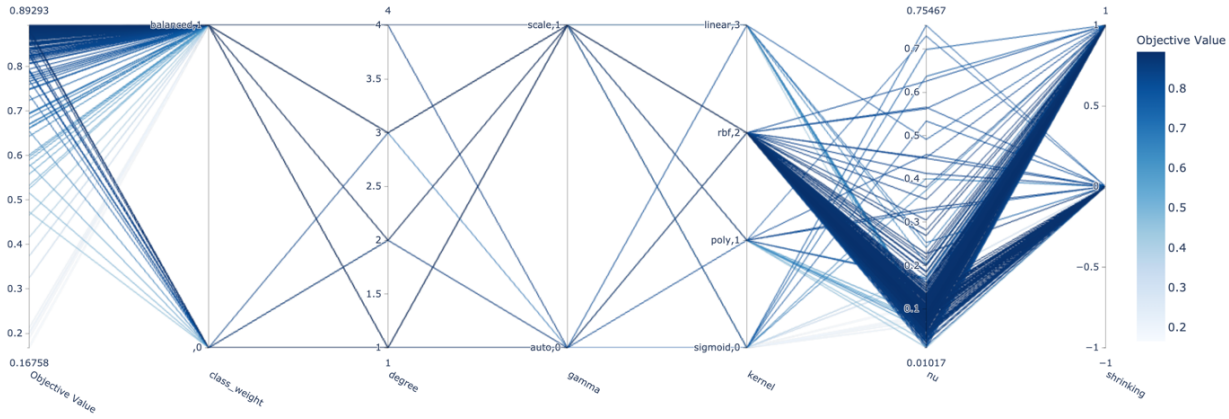

Figure 46: Parallel coordinates of hyperparameters connected by trials exposing the sampled values against the objective value — Nu-Support Vector Machine.

5 DISCUSSION & FUTURE WORK

On final results and future work for the AI system

Ultimately, AutoML was able to produce a model with 92.1% accuracy. A remarkable result, even more so, considering the reduction of time invested on manually configuring and adjusting the algorithms. That is especially true in regard to the Bayesian approach, where one simply defines the intervals to be sampled. It is not necessary to specify each individual value intended for testing.

Following on the different sampling strategies implemented in this project, the second iteration with Bayesian Hyperparameter Optimization was able to surpass the previously generated solution by a significant amount. Comparing best models, Grid Search's Nu-SVC yielded an 89.5% accuracy score – 2.6% below the second iteration's XGBoost. It would be reasonable to state that the comparison was not exactly "one-to-one", since the Bayesian optimization's winning model originated from a newly introduced algorithm. The second iteration, however, enhanced all intermediary models, Nu-SVC including. A mean additional 4.76% points in accuracy score. Therefore, it corroborates with previous research on the superiority of Bayesian methods for hyperparameter optimization over Grid Search (Bergstra et al., 2011). Only, in this case, it was tested on a real-world scenario.

Nonetheless, there's always room for improvement. From Grid Search to Bayesian Optimization what changed was the inclusion of a guided search throughout the sampling process. A "reasoning" on the algorithm's part, that leads to better combinations of hyperparameter values. Following on such enhancement, evolutionary optimization presents a promising alternative capable of escaping local optima solutions. Leveraging what was observed from the algorithms hyperparameters, a more robust search process could be established in order to avoid scenarios such as the many iterations run pointlessly, removing hyperparameters that do not contribute and better account for relationships between the hyperparameter values.

Also, the temporal aspect of the dataset must be taken into consideration. On top of naturally being a time dependent scenario – due to socio-economic factors – it is undeniable that models trained on data prior to COVID-19 will most probably fall short for subsequent times. Therefore, a framework for continuous assessment on performance decay should be structured for the production model. On that note, this is where the change from Grid Search to Bayesian Optimization made a second contribution. By not being fixed values, the latter approach offers the modeling framework a higher flexibility in case of domain or task modification.

Since one of the objectives for the second iteration was to compare results in respect to Grid Search's AutoML, the evaluation metrics were maintained throughout both iterations. However, other useful assessments may be included in the future. For example, considering a crucial motivation for the second iteration was avoiding misclassifications that lead to a detrimental outcome (e.g. a crisis-affected household being identified as low-control), one routine to be implemented is optimizing for Recall per class.

On social contributions

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 and the development of measures to fight such phenomenon 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.

Following from the above, the research was able to generate reliable descriptive and predictive models responding to the established goals. The AI embedded softwares will be able to support ACP in assisting households suffering with over-indebtedness and help guide Portuguese consumers (on a national level) on making better financial decisions, while flagging cases that indicate possible financial hardship. As future work, a translation of the external application was requested by the FPUL project partners so as to expand the software's reaching to all European countries.

Reflecting on the two iterations, apart from the improved accuracy, a point to be made is that the case should not be framed solely from a benchmark perspective. Instead, what the extra 2.6% translates is the number of households benefiting from such improvement. Following on the rationale discussed at section 3.3.5, in a sample of 1000 households, now around 26 families will get a better prediction towards receiving the necessary help. A substantial leap when framed in such terms and, consequently, worthy of applying further iterations. Therefore, this formally states one of the intended future works: continuously iterate over more promising methods in hopes of yielding better (incremental) results.

On methodological contributions to research

Methodologically, the study contributes to business research presenting an AutoML perspective for social good. Inter-algorithmic hyperparameter optimization automates the

configuration and selection of a complex machine learning model of over-indebtedness and fosters the generation of performant models. 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 the author's knowledge, the work here reported represents the first attempt of exploiting AutoML in business research (see Table 1 for details). By doing so, the AutoML approach was able to promptly return accurate predictions on new over-indebted cases. Such results have important practical and social implications.

On theoretical contributions

The study's 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 it 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 to 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 by 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 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 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.

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 level across the three profiles. Considering educational level as a proxy of literacy in general and financial literacy in particular, this suggests that financially 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.

6 CONCLUSIONS

The author's purpose with this study was to contribute on 2 levels:

(1) on a real, globally shared problem; and

(2) on how other research questions and social problems may benefit today from AI.

Since the aftermath of the 2008-2009 financial crisis, Portuguese families, as in many other nations, have been dealing with over-indebtedness at varying levels. An additional strain imposed by the scenario is that numbers on over-indebtedness does not seem to accompany the general socio-economic improvement the country has been experiencing since 2016. Despite the decline in unemployment and progressive removal of cuts in monthly income, the Portuguese households' debt-to-income rate increased from 70.8% in 2017 to 73% in the first semester of 2018 (DECO, 2018). This contradicts accounts on over-indebtedness being the consequence of a single factor – such as the crisis – and hints on a complex, multi-faceted phenomenon.

In response to this nationally experienced issue, Artificial Intelligence models were applied to a dataset of financial and social-demographic information on over-indebted households. First, a descriptive model searched for a multi-dimensional profile analysis of over-indebtedness by clustering the reported cases. 3 profiles were identified and interpreted as "low-income households", "low-credit control households" and "crisis-affected households". These served the considerable contribution on characterizing the interplay of factors leading to over-indebtedness. Consequently, the study comes to the conclusion that the concept of over-indebtedness *per se* might be understood as inadequate, and, in fact, should be interpreted on a case basis.

Following on such results, two applications were conceived and developed to leverage AI and what was learned from the descriptive model. One software focuses on assisting debt advisory services in helping over-indebted households. The other incorporates a far-reaching approach, offering guidance to the general public on financial decisions and interpretation of one's financial scenario. The external application, focused on the Portuguese public is already in a beta version and may be visited at www.saudefinanceira.psicologia.ulisboa.pt.

"Leveraging AI", as mentioned above, translates into building predictive models able to streamline the profiling of over-indebted consumers. For such, an AutoML approach was implemented that proved itself extremely valuable. Avoiding the manual process of fine-tuning and comparing models, 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.

This work opens several possible research opportunities using AutoML in business investigation. Based on the results reached using AutoML to predict over-indebtedness, it is plausible to posit 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. In sum, this proposes a framework for both scaling the modeling process as well as to become a powerful tool for less technical, more business-oriented researchers. Not having to invest considerable amounts of time on understanding algorithms and hyperparameters, allows researchers from different backgrounds to make use of the power of AI in their investigations.

Finally, the combined use of descriptive models and 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 approach followed in this paper appears suitable to study such diversity.

7 APPENDIX A: DATASET'S VARIABLES

Feature	Data Type	Group	Note
Process Number	Categorical	N/A	Unique anonymous Categorical identifier of consumer process
Marital Status	Categorical	Socio-demographic	
People in the household	Numeric	Socio-demographic	
Level of Education	Categorical	Socio-demographic	
Years of study	Numeric	Socio-demographic	
Employment status	Categorical	Socio-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	

8 BIBLIOGRAPHY

Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. 2011. Algorithms for hyper-parameter optimization. In Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11). Curran Associates Inc., Red Hook, NY, USA, 2546–2554.

Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. Mach Learn 63, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Friedman, Jerome H. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 (2001), no. 5, 1189--1232. doi:10.1214/aos/1013203451. https://projecteuclid.org/euclid.aos/1013203451

Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. 2011. Algorithms for hyper-parameter optimization. In Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11). Curran Associates Inc., Red Hook, NY, USA, 2546–2554.

Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. Mach Learn 63, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Friedman, Jerome H. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 (2001), no. 5, 1189--1232. doi:10.1214/aos/1013203451. https://projecteuclid.org/euclid.aos/1013203451

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), doi.org/10.1371/journal.pone.0191863

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.

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 (2).

Angel, S., Einböck, M., & Heitzmann, K. (2009). Politik gegen 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/

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?

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

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.

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

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.

Fondeville, N., Őzdemir, E., & Ward, W., "Over-indebtedness. New evidence from the EU-SILC special module" Research note 4/2010, European Commision

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

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. European Review 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.

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.

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

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), (pp. 160-217) Retrieved from https://www.jstor.org/stable/40969831?seq=1

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.

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

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.),Thescienceofreason:AfestschriftforJonathanSt.B.T.Evans(pp.335-396). New York: Psychology Press.

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

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 (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.