Master Thesis

William Möllestam

2022

Predicting Saving Behavior

Artificial Neural Network & Machine Learning Algorithms

Lund University

School of Economics and Management

Department of Economics

Supervisor: Erik Wengström

Abstract

This study aims to predict saving behavior using Artificial Neural Network (ANN), XGBoost, and Support Vector Machine (SVM) algorithms. First, 25 variables were chosen from the original 217 questions asked by the National Financial Capability Well-Being Survey (2018) NFCS, using exploratory data analysis. K-means clustering was applied to determine the optimal number of saving classes (k=5) to include in the final model. Thereafter, a five-fold cross validation (CV) technique was used to tune each model's hyperparameters. Using the optimal hyperparameter configuration and a training set of 70% of the data, prediction models were constructed. The performance of each model was then evaluated using the test set (30% of the data). The precision, recall, and F_1 indexes were used to analyze the prediction performances of each saving class, whereas the accuracy and their macro-average values were applied to evaluate the overall performance of the prediction model. The relative importance of each variable was determined based on the sensitivity analysis of the variables. The financial planning horizon and how long individuals believed they would live had the biggest influence on prediction outcomes. In addition, classical economic methods and other ML algorithms were adopted as comparisons. The results showed that ANN, XGBoost, and SVM algorithm achieved a better comprehensive performance, and their prediction accuracies were 0.85, 0.84, and 0.80, respectively. For questions related to behavioral economics and saving behavior, the presented methodology can serve as a reliable reference.

Keywords: Saving Behavior, Artificial Neural Network, Machine Learning, XGBoost, SVM

Acknowledgement

Thank you to my beautiful wife, Kamelia Möllestam, for her unconditional love, support, and encouragement throughout my academic journey, as well as to my son, Kyler, who is the apple of my eye, and my daughter, Kiara, who is the love of my life – I am so proud and happy to be your daddy.

Without you, none of this would have been possible.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
DT	Decision Tree
ML	Machine Learning
RF	Random Forest
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting

Introduction

Planning for savings remains one of the most crucial decisions for each consumer, with decision making being the most significant aspect of this process. Putting money away now in order to prepare for the future is a complicated process influenced by a variety of circumstances. There are several causes for concern regarding the amount of money that the average individual saves. For instance, saving enables families and individuals to disperse their income over the course of their lives, providing themselves with financial stability for any future challenges and for retirement (Modigliani, 1970). In 2019, 4,9 million or 8.9 percent of senior citizens in the United States were living below the poverty level. Also, we are currently facing a demographic crisis as the older population is increasing relative to the working age group, and this trend is predicted to reach a peak around 2055. Hence, saving is of increasing interest for both economists, researchers, and governments. In response to the challenge of the aging population, in the last decade, many governments in EU have taken steps to address the demographic crisis with reforms of their pension systems. Many governments have raised or are planning to gradually increase the pension age or have taken other parametric measures such as reducing early retirement opportunities or increasing required contribution periods just to mention a few.

A few previous studies have tried to predict savings using a variety of different methods, for instance, Gerhard et al. (2018) used linear predictions. Mahalingam & Vivek (2016) utilized a sigmoid function to predict the maximum amount to save based on current account balance, Banerjee et al. (2011) estimated a 2SLS to predict household savings and (Fisher, 2011) used a logistic regression estimation to predict participation in a savings plan. However, as saving behavior tends to be nonlinear and complex (Jenkins et al., 2017), nonlinear methods like Artificial Neural Network and XGBoost might be more appropriate.

In this paper, I predict saving behavior of 5210 individuals by applying machine learning algorithms and a deep neural network using the 2018 National Financial Capability Study (NFCS) data from the United States. The dataset contained 217 questions regarding the individuals current state of financial well-being of American adults. Based on previous theoretical and empirical findings of saving behavior, I combined psychological characteristics, situational factors, demographic data, and financial literacy to reduce the overall questions to 25.

These questions are then used with the K-means clustering method to categorize the individuals based on their savings level. Following this, the Artificial Neural Network, XGBoost, and Support Vector Machine algorithms are used to predict the class of saving behavior. The results will be compared to Linear Regression, Logistic Regression, Naïve Bayes regression, Decision Tree, Random Forest, and LightGBM as a robustness check. To evaluate the results, I will use evaluation metrics such as Precision, Recall, F_1-score. Finally, I will examine which variables that are the most important for behavior prediction. I found that the Artificial Neural Network outperformed all other methods in the prediction of saving behavior with an accuracy of 0.85. Overall, the three chosen methods could class 4/5 individuals correctly. I observed that saving behavior is a complex question that depends on a variety of background characteristics. From the time an individual's parents introduce them to money and savings to the time they estimate dying. Each period from childhood to retirement has a significant effect on consumers' saving behavior. Hence, time horizon of savings and future utility discount play an important role in predicting saving behavior.

I contribute to the existing literature in mainly two ways. Firstly, to the best of my knowledge, no previous study on savings or saving behavior has applied machine learning methods to predict saving behavior combining demographic, situational and psychological variables. Secondly, to the best of my knowledge, no previous study on the subject has contained such a large dataset in the same paper when trying to predict the saving behavior of individuals or households. To be able to predict saving behavior with the questions used in this paper has a significant impact on the economic literature. Firstly, all individuals will be able to answer the questions regardless of their previous financial literacy. Secondly, instead of nudging individuals to save, which has no empirical estimates on how long it will last, I propose to use these questions as a tool for young adults to learn the different aspects of life that relates to saving so that the individual him/herself can change a pattern/behavior that he/she feels is possible to maintain in the long run.

The structure of the paper is as follows. Section 2 will highlight the background by presenting the theoretical – and empirical framework and previous literature. The data and the data preparation are included in Section 3. Furthermore, Section 4 presents the methodological framework, while Section 5 will describe the results of the prediction performance. The findings of this paper are discussed in Section 6 and the last Section 7 concludes the paper.

Background

This section aims to present the theoretical and empirical framework, covering the basic of saving behavior and possible explanations. This is followed by the related literature and previous findings of machine learning algorithms related to economic research.

Theoretical Frameworks

The Life-Cycle Hypothesis

The life cycle theory is one of the most important hypotheses that attempts to explain individuals' saving decisions. The theory initially appeared in two publications written by Modigliani and Brumberg in the early 1950s (Modigliani & Brumberg, 1954; 1980). The theory is based on the idea that individual consumers try to maximize their utility by smoothing consumption over their expected lifetime. At each stage of life, individuals have their present value of wealth, which is the discounted value of their current and expected future resources. When income is high, there is an increase in savings, and when income is low, there is an increase in credit and borrowing. Thus, according to the theory, the primary purpose of saving is to accumulate resources for consumption after retirement. In other words, the hypothesis argues that individuals earn the most when they are of working age, that their money is diminished during old age, and that those with higher earnings are able to save more and have better financial awareness than those with lower incomes. Individuals' savings are based on their expected average lifetime income rather than their income at any particular period in their lives (Lusardi & Mitchell, 2014).

The Permanent Income Hypothesis

The permanent income hypothesis, proposed by Friedman (1957), is an economic theory similar to the life cycle hypothesis that focuses on smoothing consumption across a lifetime and being prepared for income reduction. However, there is a distinction between the two hypotheses since Friedman's permanent income hypothesis places greater emphasis on expected future income. What households anticipate their future income to be influences their current spending and saving

patterns. This implies that if current income falls below the average expected lifetime income, consumers will reduce their savings and increase their borrowing to fund consumption. However, according to the theory, it is not one's present income that defines how much one spends and saves, but rather the expected average income for any given period. This implies that individuals and households will maintain the same level of consumption even if their current income grows and will save the remaining money instead of spending them in order to prepare for future unexpected income losses (Friedman, 2016, p.29-30).

Permanent income is the expected income over a long-term planning horizon, whereas transitory income is the difference between current income and the expected permanent long-term income (Muradoglu & Taskin, 1996). According to the definition and purpose of the permanent income hypothesis, short-term fluctuations and increases in income should have no effect on spending and consumption, as consumption is at a steady level in relation to permanent income (Friedman, 1957).

The Behavioral Life-Cycle Hypothesis

Shefrin & Thaler's (1988) behavioral life-cycle theory is based on the premise that even individuals who desire smooth spending throughout their entire life cycle, as predicted by the traditional life-cycle theory, find it difficult to avoid cognitive and psychological errors. The theory accounts for three different biases, namely, framing, mental accounting, and self-control. Self-control is used to balance preferences for spending now and preferences for saving for the future. Further, Shefrin and Thaler (1988) propose that wealth is separated into three mental accounts: present income, current assets, and future income, with the temptation to consume being greatest for current income and lowest for future income. In order to manage their impatience, some individuals delay the receiving of income in order to maintain spending discipline. This prediction stands in sharp contrast to the neoclassical assumption that a higher present value corresponds to a higher utility. The life cycle hypothesis, which also includes framing, essentially suggests that the way in which information is presented has a significant impact on individual's investing and saving habits. The framing effect states that consumer choices will be influenced by how information is presented.

Empirical Framework

Demographics

The objective of demographic methods to saving behavior is to comprehend the relationship between microeconomic characteristics and household saving patterns (Walden, 2012). Demographic models relating to savings have often been employed to describe the microeconomic behavior of an individual or a household (Haron et al., 2013). Directly and indirectly, economic, and demographic factors, particularly those related to the life-cycle stage, have been found to impact saving behavior. These include age, income, marital status, education, ethnicity, and gender (Browning & Lusardi, 1996; Gutter et al., 2007; Whitaker et al., 2013). An increasing age, income or a higher education led to more savings (Aktas et al., 2012; Case et al., 2005; Fisher & Montalto, 2011; Juster et al., 2004; Metin-Ozcan et al., 2012; Rha et al., 2006; Yilmazer, 2010). The likelihood of engaging in prudent financial management activities, such as saving money, was higher among those with higher incomes (Chang, 1994; Perry & Morris, 2006; Wakita et al., 2000). In addition, Remble et al. (2014) found that a household's willingness to save was a function of its income and expenditures. Moreover, there were disparities in the rate of household savings throughout the income range. In addition, marital status was shown as a significant variable of saving behavior (Delafrooz & Paim, 2011). The saving behaviors and investing decisions of married and unmarried individuals differ (Chang, 1994; Johannisson, 2008; Sunden & Surette, 1998). The saving preferences of married individuals appeared to be decided at the household level (shared preferences) as opposed to as two separate individuals (Johannisson, 2008). Lastly, ethnicity affects savings, with white households saving more than any other ethnic group (Lee & Hanna, 2015; Rha et al., 2006).

Situational

Situation influences human behavior (Ross et al., 2011), beyond dispositional considerations at times (Darley & Latane, 1968; Darley & Batson, 1973). Literature divides situational factors into three categories: Cues, Characteristics, and Classes. Cues are objective descriptions of the environment and may contain interaction, object, location, and activity descriptions. The subjective perception of events like conflict, pleasantness, favorability, intelligence, and social

interaction are represented by characteristics. Classes categorize the entire circumstance depending on how individuals see it, such as a trading situation, a working situation, a dispute situation, and so on (Rauthmann et al., 2015). I refer to situational elements as cues in the current thesis since this term is less susceptible to interpretational discrepancies.

Shock can affect individuals' savings (Mullainathan & Shafir, 2009), and emergency health situations may cause financial stress that results in a rise in household debt and a decrease in savings (Babiarz & Robb, 2022). Few researchers have investigated the link between financial socialization, the mechanisms by which we learn about money, and financial habits during the past several decades. A concentration on lifestyles financed by debt and a loss in personal savings contribute to the interest in financial socialism (Cho et al., 2012; Guidolin & La Jeunesse, 2007; Lea et al., 1995; Schuchardt et al., 2009). Parents were viewed as the most significant socialization agents for their children. Family conversations, the use of diverse information sources, and parental modeling were utilized to evaluate socialization possibilities (Kim & Chatterjee, 2013; Kim et al., 2011). Cho et al. (2012) found that financial socialization has a substantial impact on the financial management of individuals beyond the youth and college years. Parent–child interactions over money were also connected with student financial behaviors (Kim et al., 2011). Prior research indicated that financial socialization throughout childhood favorably influenced the savings and long-term planning habits of young people. Additionally, having a savings account as a teenager was favorably related to managing one's own money as a young adult (Kim & Chatterjee, 2013).

Psychological

Psychological Economics by Katona (1975) is a classic starting point for current economic psychology methods. He deconstructs social conduct into several components, two of which are people's ability and motivation. In terms of saving, the approach highlights the individual's capacity and willingness to save. Katona proposed, in line with the major economic theories of consumption, that disposable income was a direct indicator of a person's willingness to save. However, he claimed that a person's inclination to save was influenced by his or her economic optimism or pessimism. Traditional economic theories have recognized psychological elements in saving, such as dread of economic instability and economic pessimism (Lunt & Livingstone, 1991). One example is the concept of precautionary saving, in which households'

worry about their economic future pushes them to save money in case their economic situation changes (Walden, 2012). Loibl et al. (2010) examined the significance of psychological disposition for predicting saving behavior. Future-oriented participants were more likely to continue accumulating assets through savings. The "Big Five" model is the prevailing paradigm for personality characteristic assessment. Originating in psychology, its application in economic research has increased. The five personality characteristics are agreeableness, conscientiousness, extraversion, neuroticism, and experience-seeking. Despite the fact that various research has attempted to predict saving behavior using the Big Five, there is no consensus on which personality qualities are most closely associated with saving behavior and how. For this reason, self-efficacy, self-control, planning horizon, and optimism were chosen as explanatory variables of psychological drive for saving in the current study.

Self-control

Prior research suggests that self-control may impact saving behavior over the life cycle via multiple mechanisms (Thaler and Shefrin, 1981). Self-control is the capacity to constrain one's temptations, emotions, impulses, desires, time preferences, and actions in order to maintain a desired outcome (e.g., having a financially secure retirement) or resist a temptation (e.g., spending money on non-essential items). Self-control is the act of self-regulation in circumstances where there is an obvious trade-off between long-term goals and immediate satisfaction (Thaler and Shefrin, 1981; Vohs et al., 2012). Self-control varies significantly across individuals, and those with a lower level are less prone to save for the future and more willing to give in to impulses to spend today (Thaler and Benartzi, 2004). Self-control may influence savings both directly and indirectly by influencing other factors, such as income. Researchers have discovered a correlation between self-control and higher earnings (Haushofer and Fehr, 2014). Self-perception may be defined as the extent to which a person feels responsible for shaping his or her own destiny. Those with an external locus of control have a tendency to assume that luck and chance are the most important components in attaining goals. Individuals with an internal locus of control, in contrast, tend to assume that their actions are accountable for goal accomplishments. Research demonstrates a negative relationship between external locus of control and responsible financial behavior, as measured by savings (Perry & Morris, 2006). Behavioral economists have devoted a significant amount of time to studying customers' time preferences (Grossman, 1972). Typically, multiperiod consumption models are employed,

where savings offer a mechanism for transferring consumption from early periods to later periods in order to maximize life utility. Future utility is discounted to some degree based on an individual's impatience, uncertainty, or perception of their expected lifespan. Among younger individuals, the motivation to save for retirement may be more sensitive to the rate at which future utility is discounted (Epper et al., 2020).

Self-efficacy

As a theoretical construct, self-efficacy refers to an individual's confidence in his or her own abilities to achieve the intended outcomes (Bandura, 1971; Sherer et al., 1982). For example, individuals avoid tasks they believe are above their capabilities (Bandura, 1971). This concept has been demonstrated to be crucial for both the onset and maintenance of behavioral change (Bandura, 1990). Individuals may desire to save but fail to do so because they fear they will fail in this attempt (Lown et al., 2015). This term is interchangeable with self-efficacy (Magendans et al., 2017). Individuals with high self-efficacy are more likely to save, and those who are successful at saving are more likely to have high self-efficacy (Lown et al., 2015). Magendans et al. (2017) observed that financial self-efficacy influenced saving purpose and behavior: individuals with a high financial self-efficacy reported a higher saving intention and more saving behavior.

Optimism

Optimism may be described as a widespread positive outlook on the future (Scheier et al., 1994). Individuals who are optimistic about the future are less likely to feel they need to save for potential unfavorable life occurrences. In fact, Vanden Abeele (1988) demonstrated a negative association among short-term savings and consumer optimism using Katona's indicator of consumer expectations, which was later included into the well-recognized University of Michigan Consumer Sentiment Index. Similarly, van Raaij & Gianotten (1990) evaluated a different data set and found that families with more optimistic financial prospects tend to save less. Lastly, Puri & Robinson (2007) discovered that individuals who are extremely enthusiastic regarding their future prospects may not devote enough resources to precautionary savings since they do not sense the need to save.

Planning Horizon

Time preference (the opportunity cost of sacrificing present value for future utility) has served as a bridge between the economic notion of utility maximization and the psychological concepts of impulsiveness and impatience. Beverly (1997) argued that low-income families may be more patient than high-income households due to their restricted resources. According to research by Chamon et al. (2010), households with younger couples have a longer time horizon to adjust their finances to account for approaching retirement. Both DeVaney et al. (2007) and Lee et al. (2000) determined that saving was associated with a long-term planning horizon.

Related Literature on Machine Learning

Arthur Samuel established the term Machine Learning (ML) in 1959, primarily to represent the pattern recognition tasks that provided the "learning" component of the then-pioneering Artificial Intelligence (AI) systems. The notion of Artificial Intelligence was researched tentatively and theoretically in the 1930s, but it was not studied systematically until the famous Dartmouth Workshop of 1956 (Kline, 2011). The use of Machine Learning to economics questions can be traced as far back as 1974 (Lee & Lee, 1974), although only in the abstract. Wang et al. (1984) published the first paper that applied ML approach entirely to an economics topic; the authors attempted to predict creditworthiness.

In 1988, White published a research paper that utilized Neural Networks (NN) to predict the daily stock returns of IBM. Since then, the prevalence of machine learning in economics has progressively grown. ML approaches have gained popularity in predicting stock prices as well over the past decade (Ghosh et al., 2021). For instance, Gorenc Novak and Veluscek (2016) used ML techniques, such as support vector machine (SVM), to predict stock prices and discovered that ML techniques significantly improve prediction accuracy compared to traditional models. Chen and Ge (2021) used neural networks to implement a learning-based method for optimum investing, whereas Braun et al. (2020) employed ANNs to analyze stock liquidity. Huang et al. (2004) used SVM and neural networks to predict bond ratings in the European debt market.

Hernandez and Wilson (2013) tried to predict company bankruptcy where they found that the application of artificial neural networks (ANN) outperformed all other ML algorithms. Kwak et

al. (2012) applied different ML techniques and found that SVM predicted the bankruptcy of Korean companies best. Prior research has also used artificial intelligence approaches to detect money laundering activities, where they found that ANN outperformed ML algorithms (Garcia-Bedoya et al., 2020). Additionally, ML techniques have been applied to predict the prices of gold and agricultural products (Malliaris & Malliaris, 2013). Finally, Omar et al. (2017) used an ANN to predict fraudulent financial reporting and observed that the model outperformed traditional statistical methods.

Data

This section describes the data gathered by NFCS, as well as the data processing and descriptive statistics.

Original Data

This study is based on data from the National Financial Capability Study Well-Being Survey (2018) NFCS, collected by the Consumer Financial Protection Bureau. The survey was conducted in English and Spanish between October 21, 2016, and December 5, 2016, in the 50 states of the U.S. and Washington D.C. Only adults aged 18 and above were sampled, and only one panelist per household. The final survey includes 6,394 participants who answered a total of 217 questions.

The survey was done by GfK group using KnowledgePanel, the largest U.S. probability-based non-volunteer internet panel, which allows for stratification of the sample with oversampling to subgroups with low representation. The randomization of the sample was done using address-based sampling (ABS). To ensure that the collected data represents U.S. population segments, weights are reported for each group by age, sex, race, poverty, and education. The purpose of the survey was to measure the current state of financial well-being of American adults among subpopulations. The questions in the survey represent information about individuals, households, and families, income, and employment, saving and safety nets, financial experience, financial behavior and attitude, financial knowledge, social context, and personal traits.

Data Preparation

The initial phase of this study was to conduct an *exploratory data analysis* (EDA) (Tukey, 1977). The primary goal of conducting an EDA was to uncover variables, patterns, and correlations within the data set. When preparing the data, I had to consider a variety of factors. Firstly, I applied all conclusions and focus restrictions from the EDA to the dataset such as outliers, deleting missing values and limited the savings prediction to the dependent variable. As

previously mentioned, a fundamental issue with gathering data from a survey whose objective was not to predict saving behavior is the quality of the data, notably the high proportion of missing values. This issue can be addressed in two distinct ways, via imputation or deletion:

- Imputing data replaces missing data with either an approximation or a random value.
- By removing data, the whole data row for rows with missing values is removed.

Since I had access to a big dataset to begin with, I reasoned that removing rows with missing values rather than imputed partially wrong values would have less of an impact on the final model's accuracy. Despite the fact that this is a common step in many data mining projects, there is a chance that a bias will be introduced. This relies greatly on the cause of the missing data. The possibility of creating bias is diminished if missing values are distributed uniformly across the dataset. However, a biased result might be generated if the missing values occurred mostly for a single question. In this case, the missing values was uniformly across the questions

Data Understanding

As the main goal of the paper was to build a model that can predict saving behavior of individuals using demographical-, situational-, and psychological characteristics and combine these with financial literacy to increase the understanding of saving patterns, I focused the EDA on the correlations between different variables rather than finding causalities. I will rely on the literature to understand this behavior. I recognized psychological variables using a combination of answers based on literature and inventories, while demographics and situational factors could be identified by their variable name.

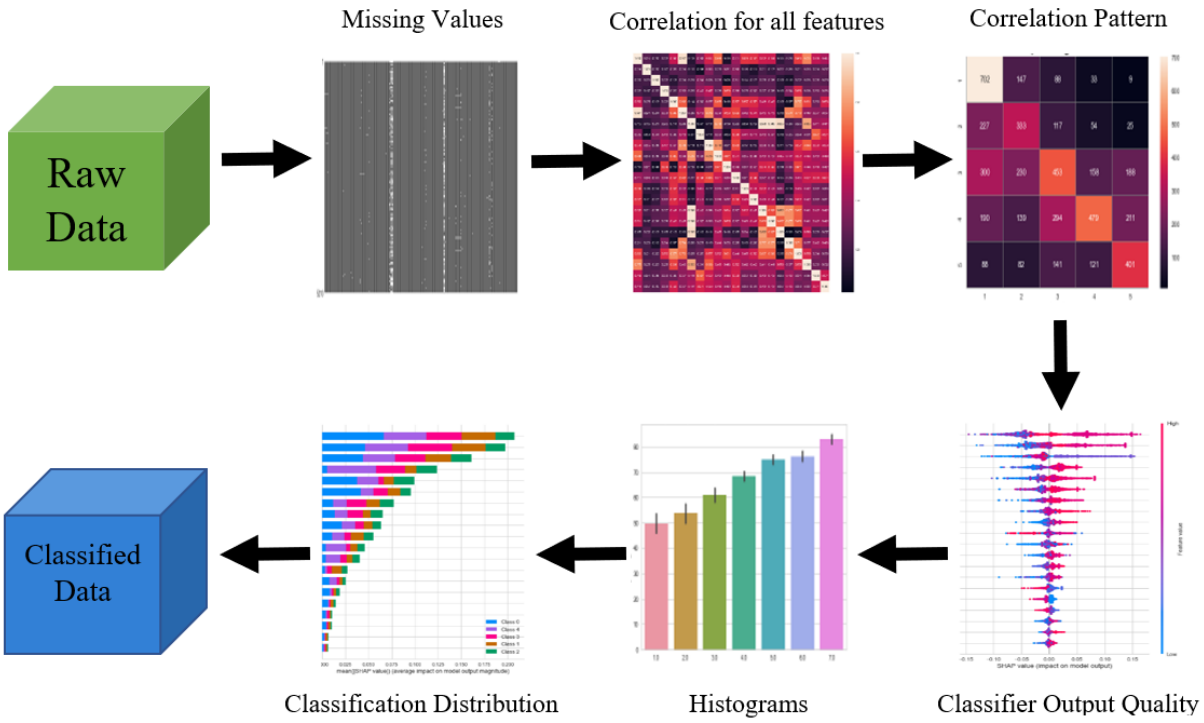
To determine the possible distinctions between classes, I considered numerous aspects[1]:

- Correlation for all variables: to be able to reduce the questions as much as possible for future correspondents, variables were allowed to have a correlation no higher than 0.35 percent. If higher, one got removed.
- Correlation pattern: to recognize trends in different questions to make it easier for the algorithms to separate different classes from each other.

[1] For the full EDA, see Appendix A.

- Classifier Output Quality: was applied to see structure of all variables and to see in what direction different variables will affect the algorithm.
- Histograms: for separation of classes and to see how the average individual in each class had answered different questions.
- Classification distribution: was then conducted to see if the variables were more or less important for the different classes.

Figure 1: Exploratory Data Analysis



Outcome Variable

There are several sorts of savings and ways to analyze it. The conventional definition of savings, for instance, would be the sum that remains after paying for essentials and other expenses. In order to provide a more in-depth explanation of what saving is, we may define saves as money that is not being invested, spent, or otherwise put at risk in any other way. However, I will define

savings in the same way that the question was asked in the survey to not risk any confusion or to not risk for the reader to have a different definition of what saving can be.

The question was asked in the following way:

How much money do you have in savings today (in cash, checking, and saving account balances)

1. $0
2. $1-99
3. $100-999
4. $1,000-4,999
5. $5,000-19,999
6. $20,000-74,999
7. $75,000 or more

Descriptive Statistics

Descriptive statistics of the 25 questions used after the EDA are reported in Table 1. As can be seen in Table 1, there are 5,210 observations for each explanatory variable after the EDA. As can be noted, there are approximately equal observations for men compared to women. This can be seen, where the variable gender is equal to 1.516. Further, one can see that the mean age is 43 years old. As some of the questions asked related to situations such as household size and marital status, a mean age of 43 gives a better distribution of these variables.

Furthermore, the columns Min and Max refer to the number of alternatives one has for each question. For example, one could choose between two different genders whereas the same individual could choose between five different education alternatives. To see all questions and the different answer alternatives, go to Appendix B.

Lastly, Table 1 displays the descriptive statistics of the four different categories used to create the prediction model. There are seven demographic variables, five questions that relate to the individual's financial literacy, seven questions relating to psychological characteristics and finally six questions regarding an individual's situational factors.

Table 1: Descriptive Statistics

Outcome Variable					
Savings Today	Ordinal Variable ranging from 1 to 5. $0-999=1, $1000-4,999=2 $5,000-19,999=3, $20,000-74,999=4 $75,000 or more=5				
Explanatory[2] Variables		Non-missing observations	Mean	Min	Max
Demographic	Age	5,210	43	18	83
	Gender	5,210	1.516	1	2
	Race / Ethnicity	5,210	1.752	1	4
	Education	5,210	2.894	1	5
	Household Income	5,210	5.376	1	9
	Household Size	5,210	2.676	1	5
	Marital Status	5,210	2.205	1	5
Financial Literacy	I know how to make myself save	5,210	3.631	1	5
	I know when I need advice about my money	5,210	3.604	1	5
	Prefers words for expression of probabilities	5,210	3.640	1	6
	How good are you at working with percentages?	5,210	4.204	1	6
	I am able to recognize a good financial investment	5,210	3.010	1	5
Psychological	Everyone has a fair chance at moving up the economic ladder	5,210	4.696	1	7
	Psychological Connectedness	5,210	68.360	0	100
	Financial planning time horizon	5,210	3.033	1	5
	I am Satisfied with my life	5,210	1.234	1	3
	I am good at resisting temptation	5,210	2.860	1	4
	I am optimistic about my future	5,210	5.406	1	7
	If I work hard today, I will be more successful in the future	5,210	5.540	1	7
Situational	In general, would you say your health is…	5,210	3.431	1	5
	Lot of stress in respondent's life	5,210	3.226	1	5
	How likely do you believe it is that you will live beyond age 75?	5,210	71.226	0	100
	About How much do you pay for your home each month	5,210	3,791	1	7
	Did your family educate you about saving, credit, allowance, or other finances when growing up?[3]	5,210	3.789	1	7
	I like to own things that impress people	5,210	2.443	1	5

[2] All questions can be seen in Appendix B.

[3] Calculated as the sum of 7 questions asked regarding family relationship to savings.

Methodology

This chapter provides a brief explanation of several machine learning methods and a justification for their application.

Machine Learning

The use of machine learning for the modeling and handling of large datasets has become widely used in recent years within the context of numerous research fields. It is a category of artificial intelligence that can learn on its own without being trained by an individual. It can be thought of as a machine's ability to make predictions and decisions like a human would. It can be applied in numerous areas, like computer vision, pattern recognition, and big data sets, where it's hard and expensive to make algorithms that can do complicated tasks.

To Choose an Algorithm for Machine Learning

A vast majority of approaches are evaluated in order to select the optimal machine learning model for the specific task. Depending on the nature of the data and the background knowledge already existing, certain models will be more or less appropriate. There are two main types of machine learning: supervised and unsupervised. Supervised learning is advantageous when the task is well-defined and the outputs to the data are known in advance. The goal of supervised learning is to learn a function that maps inputs to outputs. Unsupervised learning, on the other hand, is effective at uncovering hidden data patterns, which is beneficial for exploratory tasks (Tsymbal, 2022). Unsupervised learning tries to learn the underlying structure, without given any value pair examples. Thus, the algorithm tries to learn without explicit feedback.

In this paper, both supervised and unsupervised methods will be used for two reasons. Since the data is large and I want to find out which variables affect the outcome the most, I use unsupervised models. This is because I want to find hidden undetected patterns in the data.

Secondly, I employ supervised algorithms as past research on saving has offered a general idea of which variables have been found to be significant. Supervised approaches deal with labeled data when the machine is informed of the output data patterns. Since different methods perform better on certain challenges, the algorithm that predicts the highest result is not always the same algorithm. In addition, the following algorithms (ANN, XGBoost, and SVM) are utilized since they have established strong performance in the relevant literature regarding economic predictions.

K-means clustering

K-means is a non-parametric unsupervised learning clustering algorithm used to construct clusters from large unlabeled data sets where data with similar properties or patterns are stored together. The aim of K-means clustering is to split data into k clusters such that data points within the same cluster are comparable and data points within separate clusters are more dissimilar. The practitioner determines the number of clusters represented by "K" in K-means. K-mean clustering is one of the most widely used techniques for clustering data objects (Ali et al., 2019). The primary goal is to determine an object's nearest neighbors by minimizing the distance within a cluster and increasing the distance between clusters. Calculating the distance between two locations in n-dimensional space is possible using several distance metrics, in this example the Euclidean distance, which is defined as follows (Ali et al., 2019):

$$d(x,y) = \sqrt{\sum_{t=1}^{n}(x_i - y_i)^2} \quad (1)$$

The choice of k influences the smoothness of the density estimate used to identify class label (Everitt, 2011). This has a significant impact on the clustering outcomes. However, the K-means algorithm cannot identify the number of clusters The metric known as the Elbow Method (EM), which is a visual method for testing the consistency of the optimal number of clusters, is used to evaluate the selection of K. The idea is to determine the number of clusters, then add clusters, calculate the sum squared error (SSE) per cluster until the maximum number of clusters has been

determined. The SSE with the largest difference between two clusters indicates the optimal number of clusters and creating the angle of the elbow. Here is the SSE equation:

$$SSE = \sum_{i=1}^{k} \sum_{xj \in C_i} \left|\left| x_j - \mu_i \right|\right|^2 \quad (2)$$

Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm that uses regression or classification approaches to find hyperplanes in N-dimensional space that can categorize data points into two (or more) unique groups (Hastie et al., 2009). Numerous hyperplanes are able to classify the data, but the goal is to select the hyperplane that represents the largest separation of data points (Goodfellow et al., 2016). This would maximize the margin between the two classes and reduce the generalization error of the classifier, allowing for more accurate classification of future data points. The dimension of the hyperplane depends on the number of variables. In R^2, a hyperplane is simply a line, whereas in R^3, it becomes a two-dimensional plane. Looking at the mathematical representation of the linear SVM hyperplane, it is depicted in the following way (Goodfellow et al., 2016):

Given a training set of n points

$$(x_1, y_1), \ldots, (x_n, y_n) \quad (3)$$

Where $y_1 = -/+1$ denotes the categories to which x_1 belongs. The goal is to identify the maximum margin hyperplane that divides the group of x_1, where $y_1 = -1$, from the group of x_1, where $y_1 = +1$, the most. The definition of the hyperplane is:

$$w^T x + b = 0 \quad (4)$$

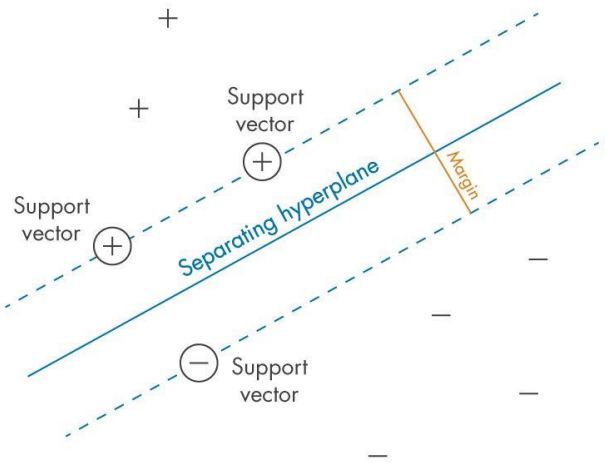
where *w* is the normal vector to the hyperplane and *b* is a real number.

If the training data can be separated linearly, the optimal separation hyperplane is determined by solving the following optimization problem:

$$\min||w|| \ subject\ to\ y_i(w^T x_i - b) \geq 1, \qquad \forall\, i = 1, \dots, n \tag{5}$$

This indicates that *w* and *b* solve the problem. Therefore, the support vectors are the x_i such that $y_i(w^T x_i + b) = 1$, indicating they are on the boundary. Figure 2 illustrates a graphical presentation of support vectors in R^2.

Figure 2: Support Vector Machine | Separating Hyperplane



+ represents data points of type +1 and - indicates data points of type -1.

XGBoost

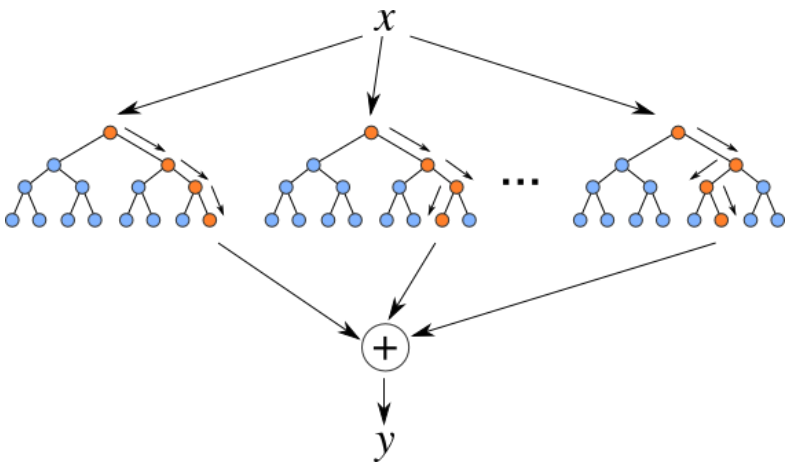
One of the most advanced statistical gradient-boosting decision trees (GBDTs) is known as XGBoost (extreme gradient booster), and it uses an ensemble of binary trees to predict an outcome. Before further explanations, a short description of binary tree model is necessary.

Binary trees are a data structure generally consisting of one root node, many internal nodes, and multiple leaf nodes. Every other node corresponds to a variable test, with the leaf node representing the decision results. The samples inside each node are subdivided into child nodes based on the findings of variable splitting. This process continues until no additional gains can be

made or a predetermined rule is satisfied, such as the tree reaching its maximum depth. The model identifies the optimal variable attributes to separate classes into homogenous groups with minimal impurity or noise. The objective is to generate a tree that can predict samples that have not yet been observed.

The XGBoost builds multiple decision trees in order to predict classes of classifications, where every tree is evaluated using a scoring function. In other words, the gradient booster XGBoost uses the gradient descent optimizer to sequentially add new "weak" models in order to enhance the final "strong" model (See Figure 3). The XGBoost algorithm has an improved scalability, portability, memory-efficiency, and predictability when compared to other binary tree models (Chen & Guestrin, 2016). For a mathematical derivation of the algorithms, go to Appendix C.

Figure 3: XGBoost Trees



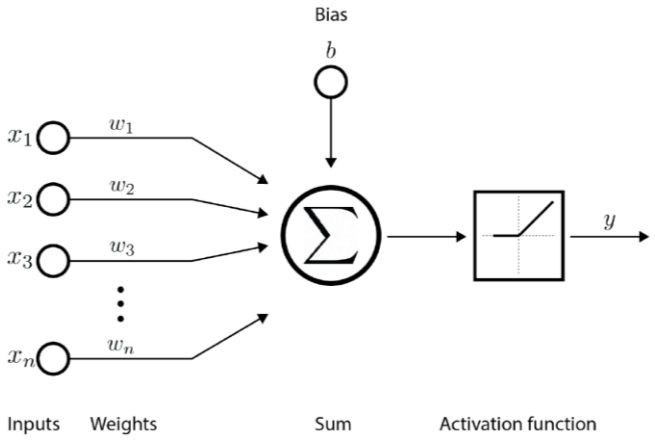
Artificial Neural Network

The concept of Artificial Neural Network (ANN) was inspired by the brain's actual biological system (Schmidhuber, 2015). The brain is a highly complicated system for processing information and is capable of handling multiple tasks simultaneously. Nonetheless, the idea of ANN is not new; it traces back to the mid-1940s, when neurophysiologist McCulloch and mathematician Pitts presented an early model of an artificial neuron (Rojas, 1996). Following, Alan Turing proposed his own concept of an unstructured machine: a simpler form of a binary neural net with all processing units linked to one another (Buccato et al., 2011). The discovery of new applications and flexible implementations have brought ANN to the frontline of ML and AI research (Russell & Norvig, 2009). ANN are created to provide a rational output or conclusion by utilizing many layers of calculation. Each layer includes a variety of artificial neurons. As input, an ANN can accept either raw data or classifications. Unique, observable attributes or variables provided in data are known as characteristics.

The Perceptron

The fundamental component and computational unit of the majority of neural networks is a single neuron classified as the perceptron. Through a sequence of computations, it transfers the input signals $x_1, x_2, \dots, x_n$to the output *y*. Each input x_n is multiplied by its individual weight w_n before being added together. Additionally, an optional bias can be introduced to the sum. It is then passed through an activation function, which converts the data to a nonlinear output range (see Figure 4).

Figure 4: The Perceptron



Each artificial neuron receives one or more input signals $x_1, x_2, x_3 \ldots, x_n$, and transmits a value to the neurons of the following layer. Output *y* is a weighted nonlinear sum of inputs. Passing the linear sum via nonlinear functions known as activation functions creates nonlinearity.

Thus, the output may be written as:

$$y = \sigma(w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n + b) \tag{6}$$

Where σ denotes any activation function. By finding the linear combination between w and x, the equation can be restated in compact matrix form:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

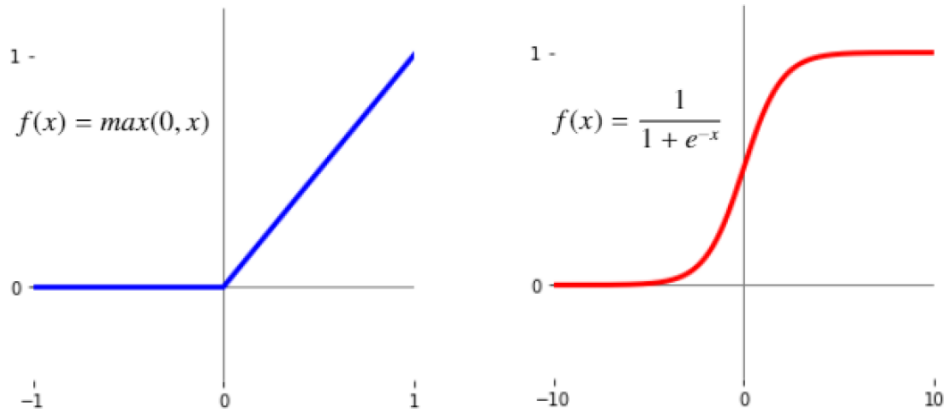
$$y = \sigma(X^T W + b) \tag{7}$$

Activation function

Without the activation function, the relationship between the output and inputs would be linear, as illustrated by Equation (7). Hence, the neuron would only predict and solve linear problems. Non-linearities are incorporated to the algorithm by activation functions (Ali et al., 2021) to

enable it to handle difficult tasks and predict arbitrary functions. The sigmoid function seen on the right of Figure 5 is one of the first activation functions employed. Due to its restricted character, it is ideally suited for probabilistic problems. Nair & Hinton (2010) presented a second activation function (The Relu) seen on the left of Figure 5. This activation function converts ≤ value to zero, whilst values above zero remain unaffected.

Figure 5: Activation function: Relu (left) | Sigmoid (right).



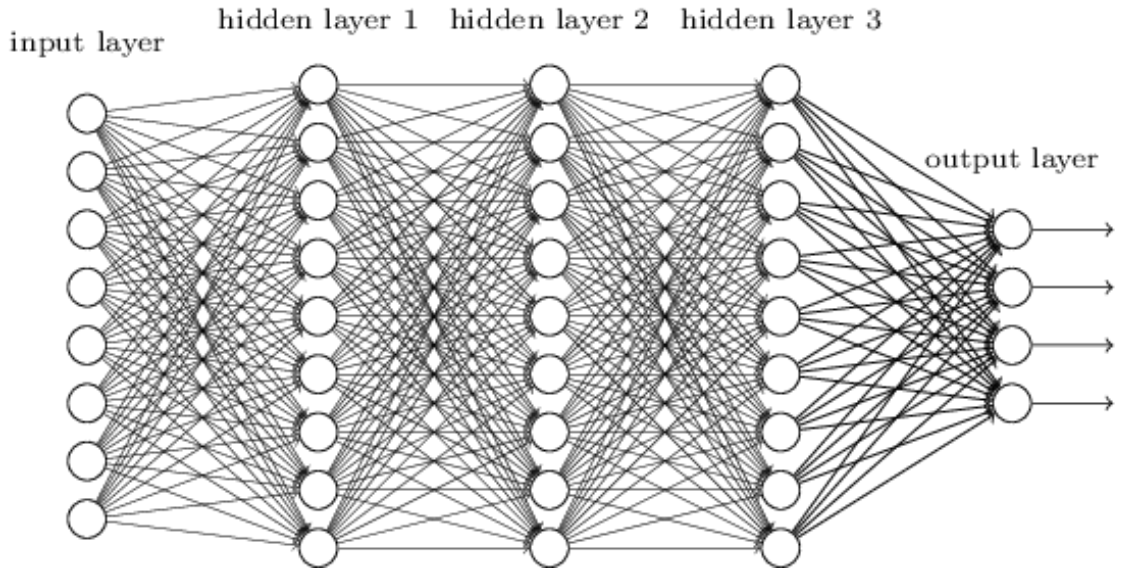
In the last layer of an ANN, a sigmoid activation function is often employed. In classification tasks, the sum of each output value equals 1 and may be represented as a probability indicator:

$$\sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \quad for\ i = 1, \dots, K \tag{8}$$

Deep Learning

As previously stated, ANN are nothing more than layers of linked perceptron/neurons. The first layer is known as the input layer, while the final layer is known as the output layer. Hidden layers are those that appear in the spaces between the input and the output layers. Figure 6 depicts a basic ANN with three hidden layers.

Figure 6: Artificial Neural Network



For each layer, the values can be derived from the preceding layer as follows:

$$\begin{bmatrix} a_0^{(1)} \\ a_1^{(1)} \\ \vdots \\ a_n^{(1)} \end{bmatrix} = \sigma\left(\begin{bmatrix} w_{0,0} & w_{0,1} & \dots & w_{0,n} \\ w_{1,0} & w_{1,1} & \dots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \dots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} \right) \tag{9}$$

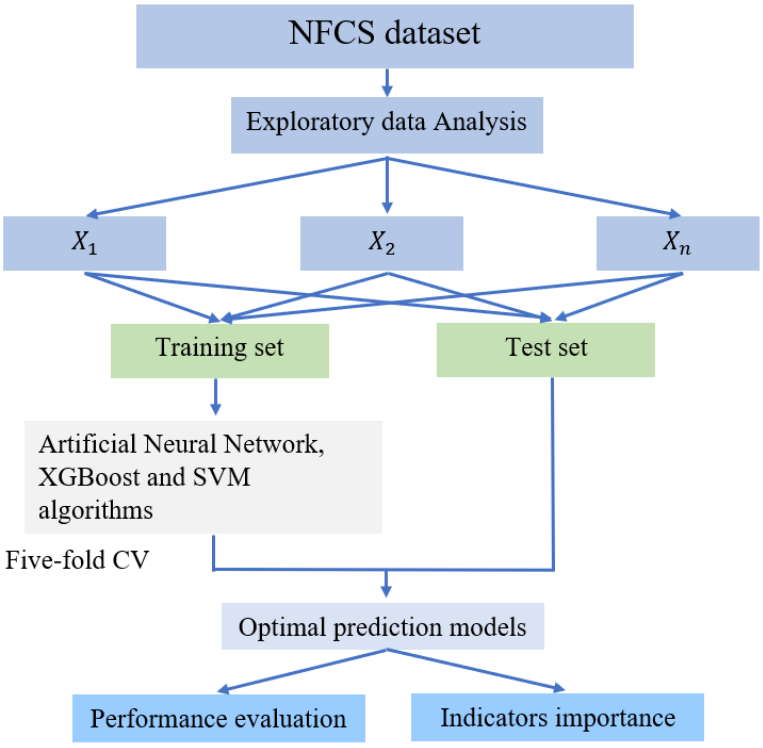
$$a^{(n)} = \sigma(W a^{(n-1)} + b) \tag{10}$$

While the concept of deep learning relates to neural networks with several (hidden) layers (Belotti et al,. 2020), deep learning is frequently associated with a model that obtains "deeper" knowledge through learning directly from data as opposed to constructed characteristics. In conclusion, the values for each layer may be determined using the preceding layer.

Construction of Prediction Model

Figure *7* demonstrates the construction process of the prediction model, which is based on Liang et al. (2020) construction model of predicting stability levels. First, 70 percent and 30 percent of the original NFCS dataset are selected as training and test sets, respectively. The model hyperparameters are then optimized via a five-folder cross validation (CV) approach. Third, the prediction model is calibrated using the appropriate hyperparameter configuration based on the training set. Fourth, the test set is used to evaluate the performance of the model based on the overall prediction results and the prediction ability for each level of savings. The optimal model is then determined by comparing each models' overall performance. If the performance of this model's predictions is good, it can be adopted for implementation. All calculations have been done in Python 3.10 using scikit-learn (Pedregosa et al., 2011). Following is a description of the hyperparameter optimization procedure and model evaluation indices.

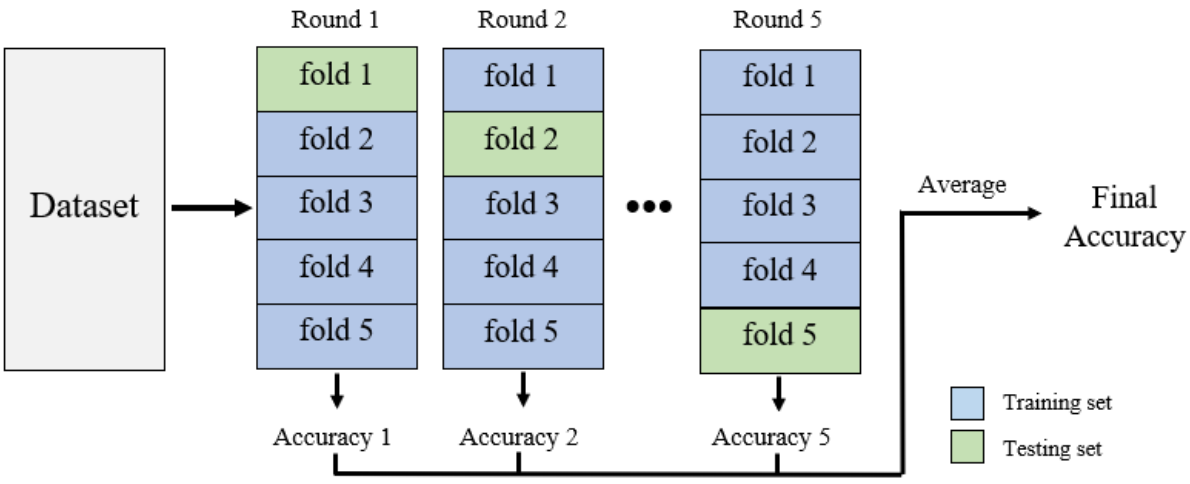
Figure 7: Construction Process



Hyperparameter Optimization

The majority of machine learning algorithms require tuning of their hyperparameters (see Appendix D). These hyperparameters should be changed based on the dataset as opposed to being explicitly specified. Bayesian optimization, Heuristic search, Grid search, and randomized search are the most common hyperparameters search methods (Kumar, 2019). Since the randomized search approach is more effective for simultaneously tuning numerous hyperparameters, it is utilized in this paper to determine the optimal set of hyperparameters. In general, the K-fold Cross Validation (CV) method is used to configure hyperparameters (Jung, 2018). Figure 8 illustrates the five-fold CV method that we employ in our paper. Five subsamples of equal size are randomly split from the original training set. A single subsample is selected as the validation set, while the remaining four are used as the training subsample. This approach is done five times until every subsample has been selected once as a validation set. An optimal set of hyperparameters is then determined by averaging the accuracy of the five validation sets.

Figure 8: Five-Fold Cross Validation



Model Evaluation Indexes

The accuracy, precision, recall, and F_1 metrics have been frequently used to evaluate the performance of machine learning algorithms (Kumar, 2019). Accuracy is the proportion of successfully predicted samples, precision is the ability to accurately predict samples, recall is the capability to correctly predict as many actual samples as possible, and the performance of both recall and precision are measured by the comprehensive metric F_1. Thus, in this paper, like Liang, et al (2020), these indicators are used to evaluate model performance. Consider the confusion metric to be stated as follows:

$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1E} \\ g_{21} & g_{22} & \cdots & g_{2E} \\ \vdots & \vdots & \ddots & \vdots \\ g_{E1} & g_{E2} & \cdots & g_{EE} \end{bmatrix} \tag{11}$$

Where g_{aa} represents the number of samples correctly predicted for level α, g_{ab} is the number of levels a sample assigned to level *b* and *E* represents the number of saving classes. The precision, recall, and F_1 measure for each saving class is determined, based on the confusion matrix, by:

$$Pr = \frac{g_{aa}}{\sum_{a=1}^{E} g_{ab}} \tag{12}$$

$$Re = \frac{g_{aa}}{\sum_{b=1}^{E} g_{ab}} \tag{13}$$

$$F_1 = \frac{2 \cdot Pr \cdot Re}{Pr + Re} \tag{14}$$

To accurately reflect overall prediction performance, the accuracy and macro average of precision, recall, and F_1 are calculated as:

$$Accuracy = \frac{1}{\sum_{a=1}^{E} \sum_{b=1}^{E} g_{ab}} \sum_{a=1}^{E} g_{aa} \tag{15}$$

$$macro - Pr = \left(\sum_{b=1}^{E} \frac{g_{aa}}{\sum_{a=1}^{E} g_{ab}} \right) / E \tag{16}$$

$$macro - Re = \left(\sum_{a=1}^{E} \frac{g_{aa}}{\sum_{b=1}^{E} g_{ab}} \right) / E \tag{17}$$

$$macro - F_1 = \frac{2 \cdot macro - Pr \cdot macro - Re}{macro - Pr + macro - Re} \tag{18}$$

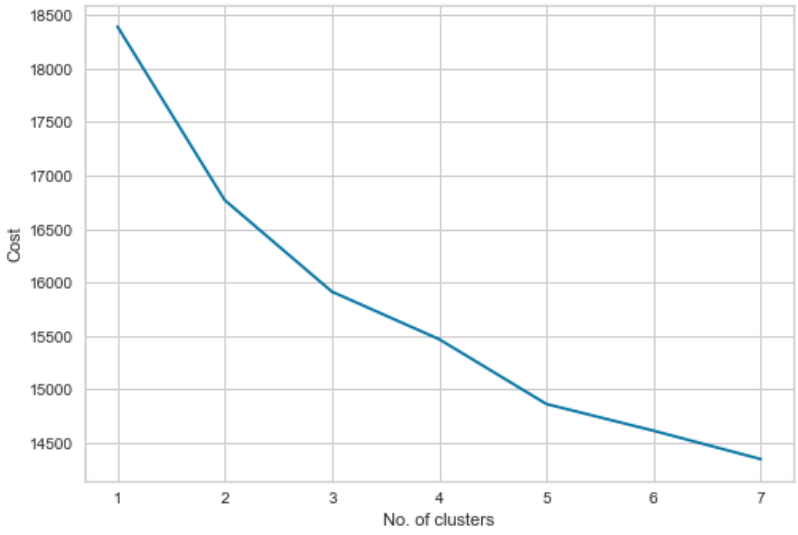
Results

This section aims to provide a descriptive presentation and analysis of the results and try to bridge the findings to the previously established research

Identification of the optimal number of classes

In this test, the performance of each number of clusters adjusted to the range of values for the Elbow method will be determined. The graph contained the SSE value in the experimental number of clusters between 1-7. The results of sum of square Error calculations of each cluster have experienced the greatest decrease in k=2 which can be seen in Figure 9. However, as the number of clusters increases, the SEE start to decrease with the largest value when k=1. When k=5, the graph starts to move almost parallel to the X-axis, thus creating an elbow shape. The k value corresponding to this point is the optimal number of clusters. Furthermore, there is a change at k=3 as well. This point will be used as a robustness check (see Appendix E). If there is a little to no increase in the overall prediction performance, k=5 is to prefer over k=3, as it is harder to predict more classes compared to fewer classes.

Figure 9: Elbow Method for Optimal k



Overall Prediction Results

The prediction results of ANN, XGBoost and SVM were obtained on the test set. Subsequently, the confusion matrix of each algorithm was determined, where the values on the main diagonal indicated the number of samples correctly predicted. As shown in Figure 10. It can be observed that the majority of samples were correctly classified, with the exception being classified with the closest class. Based on the confusion matrix, the *accuracy*, *macro – Pr*, *macro – Re*, *and macro – F_1* were calculated based on Equations (15)-(18), which were listed in table X. The accuracy degree of XGBoost was the largest, with an accuracy of 0.85, followed by XGBoost with an accuracy of 0.84, and lastly SVM with an accuracy of 0.80. Furthermore, comparing their values based on Equation (16)-(18), ANN continued to perform better than the other algorithms.

Figure 10: ANN | Confusion Matrix[4]

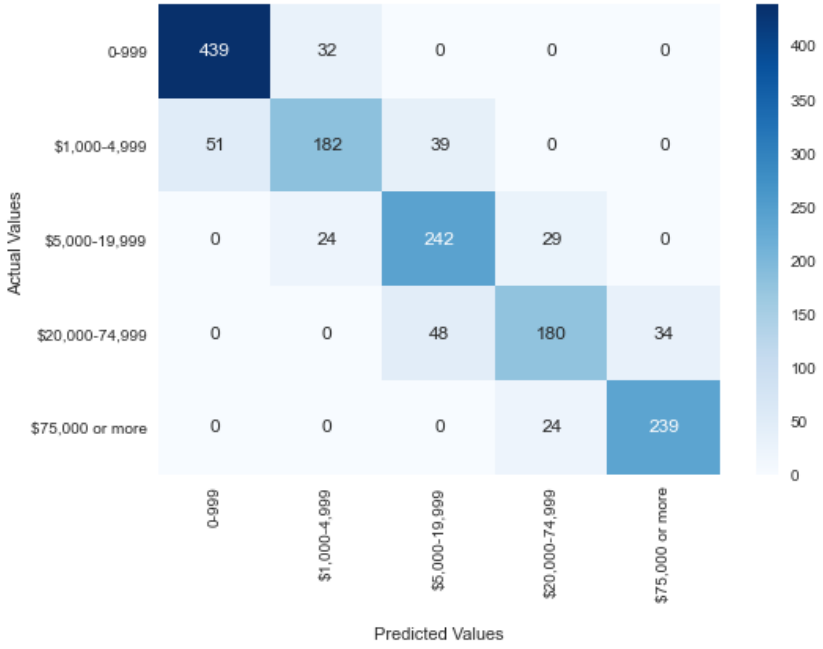


Table 2: Overall prediction results of each algorithm

	Accuracy	Macro – Pr	Macro – Re	Macro F_1
ANN	0.85	0.84	0.84	0.84
XGBoost	0.84	0.83	0.84	0.83
SVM	0.80	0.78	0.78	0.78

Note: Total support for each algorithm was 1563 with the following distribution: $0-999=251, $1,000-4,999=270, $5,000-19,999=312, $20,000-74,999=254 and $75,000 or more=251

[4] Confusion Matrix for XGBoost & SVM can be seen in appendix F.

Prediction Results of Each Class

To analyze the prediction performance of algorithms for each class level, the precision, recall and F_1 indexes were calculated based on Equations (12)-(14), which were shown in Table 3)-(Table 5), respectively. It can be shown that the prediction performance of ANN, XGBoost and SVM for the different saving levels was not the same. For the $0-999 class, ANN achieved the highest precision value (0.93); while XGBoost possessed both the highest recall value (0.92) and the highest F_1 value (0.92). For the $75,000 or more class, ANN outperformed the other algorithms with the highest recall value (0.93), and the highest F_1 value (0.94), while XGBoost performed best with the highest precision value (0.91). In addition, it can be observed that the prediction performance of these algorithms for these two classes was the best, while the algorithms had it harder with the prediction performance for the three middle classes (i.e., $1,000-4,999, $5,000-19,999 and, $20,000,74,999)

	ANN	XGBoost	SVM
$0 – 999	0.93	0.90	0.91
$1,000 - 4,999	0.81	0.79	0.68
$5,000 – 19,999	0.79	0.77	0.77
$20,000 – 74,999	0.79	0.78	0.68
$75,000 or more	0.88	0.91	0.86

Table 3: Precision values of algorithms for each class

	ANN	XGBoost	SVM
$0 – 999	0.89	0.92	0.86
$1,000 - 4,999	0.82	0.83	0.76
$5,000 – 19,999	0.77	0.74	0.58
$20,000 – 74,999	0.79	0.81	0.78
$75,000 or more	0.93	0.91	0.91

Table 4: Recall values of algorithms for each class

	ANN	XGBoost	SVM
$0 – 999	0.91	0.92	0.90
$1,000 - 4,999	0.78	0.79	0.71
$5,000 – 19,999	0.76	0.77	0.66
$20,000 – 74,999	0.80	0.78	0.71
$75,000 or more	0.94	0.90	0.87

Table 5: F_1 value of each algorithm for each class

Prediction Result of Each Category

Table 6 is the result of the prediction performance of each category, separately. It can be seen that ANN performed best in all predictions except situational, where XGBoost performed 0.04 points better, that is 4 percentage points. Overall, all categories separately did not perform well in the prediction performance. Psychological characteristics performed best with ANN accuracy of 0.68. Demographic characteristics performed worse with SVM accuracy of 0.39.

Table 6: Prediction Result | Each Category

		Accuracy	Macro – Pr	Macro – Re	Macro F_1
Demographic	ANN	0.46	0.44	0.44	0.44
	XGBoost	0.45	0.43	0.44	0.41
	SVM	0.39	0.35	0.36	0.35
Financial Literacy	ANN	0.50	0.47	0.49	0.47
	XGBoost	0.47	0.45	0.45	0.44
	SVM	0.48	0.45	0.46	0.45
Psychological	ANN	0.68	0.67	0.68	0.67
	XGBoost	0.67	0.66	0.66	0.66
	SVM	0.66	0.66	0.67	0.66
Situational	ANN	0.49	0.47	0.47	0.46
	XGBoost	0.53	0.52	0.53	0.52
	SVM	0.50	0.49	0.50	0.49

Note: Total support for each algorithm was 1563 with the following distribution: $0-999=251, $1,000-4,999=270, $5,000-19,999=312, $20,000-74,999=254 and $75,000 or more=251

Relative Importance of Indicators

The relative importance of indicators is a valuable reference for increasing individuals' saving behavior. In this study, the importance percentage of each variable was obtained from all three algorithms, separately. Table 7 ranks the variables according to the algorithms from the variable that is the most responsible for the predicted output to the variable that helps the algorithm least in the prediction performance. The higher the value the more important is the variable. The most important factor for the ANN algorithm was "financial planning time horizon" which helped the prediction performance with a score of 7.5. For the XGBoost algorithm, "how likely do you believe it is that you will live beyond age 75?" was the most important variable with a score of 7.1 and for the SVM algorithm, "psychological connectedness" was the most important with a score of 6.8. In table X, one can also see that the three least important indicators for all thee algorithms was "gender", "race / ethnicity" and "marital status" with a score of 1.4, 1.8, and 2.2 respectively. All three algorithms found the variable "Did your family educate you about saving, credit, allowance, or other finances when growing up?" to be more important than age and education when trying to predict savings.

Additionally, each category is tested separately to see the importance of each variable inside its own category (see Appendix G). When testing the importance of each variable separately inside their own category, one can see that each variable is given a similar weight of importance even in their own category, when compared to the overall prediction performance, as seen in Table 7.

Table 7: Classification Variable Importance

		Importance		
	Explanatory Variables	**ANN**	**XGBoost**	**SVM**
		(1)	**(2)**	**(3)**
Demographic	Age	4.4	4.2	3.7
	Gender	1.4	1.5	1.8
	Race / Ethnicity	1.8	1.5	1.8
	Education	3.1	3.9	3.2
	Household Income	5.6	5.1	5.5
	Household Size	2.7	1.6	3.2
	Marital Status	2.2	1.3	2.0
Financial	I know how to make myself save	7.3	5.8	6.2
	I know when I need advice about my money	2.3	2.6	2.0
	Prefers words for expression of probabilities	4.3	6.4	4.0
	How good are you at working with percentages?	6.8	6.2	5.5
	I am able to recognize a good financial investment	2.5	2.9	3.2
Psychological	Everyone has a fair chance at moving up the economic ladder	5.4	4.3	4.7
	Psychological Connectedness	6.5	6.9	6.8
	Financial planning time horizon	7.5	5.7	6.4
	I am Satisfied with my life	3.0	3.5	3.7
	I am good at resisting temptation	2.8	2.9	3.0
	I am optimistic about my future	2.7	3.0	3.1
	If I work hard today, I will be more successful in the future	4.9	4.4	4.2
Situational	In general, would you say your health is…	2.3	3.6	3.0
	Lot of stress in respondent's life	2.6	3.3	3.5
	How likely do you believe it is that you will live beyond age 75?	5.5	7.1	6.7
	About How much do you pay for your home each month	4.3	3.4	4.6
	Did your family educate you about saving, credit, allowance, or other finances when growing up?	4.8	5.0	4.8
	I like to own things that impress people	3.3	3.9	3.4

Note: The relative importance of each variable was determined based on the sensitivity analysis of the variables. Each variable is given a percentage of how important they were to predict the outcome.

Discussion

Although the behavior of saving can be accurately predicted using the Artificial Neural Network, XGBoost, and Support Vector Machine algorithms, the prediction performance for different levels of saving was not the same. There may be two explanations to this. The quantity of samples for the $0-999 class is greater than those of the other saving levels. Another issue is that the boundaries of the different saving levels are not identical, which may impact the data quality. The quantity and quality of supportive data have a significant impact on how well ANN, XGBoost, and SVM algorithms predict outcomes. Consequently, the prediction performance for the categories $1,000 to $4,999, $5,000 to $19,999, and $20,000 to $74,999 was poorer than that of the categories $0-$999 and $75,000 or more. According to the analysis of variable importance, “financial planning time horizon”, “how likely do you believe it is that you will live beyond age 75?”, and “psychological connectedness” were the most important factors for the ANN, XGBoost, and SVM algorithms, respectively. This is consistent with the vast majority of prior empirical findings. It has already been found that the financial planning horizon has a substantial association with saving (Lee et al., 2004). The question regarding one’s belief of living beyond age 75 can be closely related to Modigliani & Bromberg’s (1954; 1980) life-cycle theory. This is due to the fact that individuals who believe that they will live longer, save more money as they will be in retirement for a longer period of time. Even though the theory has met some critical review for its notion that individuals deplete their wealth as they age, the findings in this paper support the life-cycle theory of Modigliani & Bromberg.

All three algorithms ranked psychological connectedness in the top five most important variables, ranking it above age, gender, ethnicity, household income, household size, and marriage. There is a considerable correlation between the quantity of savings and the extent to which individuals assume they will stay unchanged. When individuals are in a better condition, they tend to assume it will continue and resist the impulse to change, but if the situation worsens, they feel compelled to do so. The psychological traits played a significant effect in the prediction performance, with five of the top ten variables relating to psychological characteristics. Moreover, when predicting each category individually, the psychological components performed best with an accuracy of 0.68, 0.67, and 0.66 for the ANN, XGBoost, and SVM algorithms, respectively. The question "If I work hard today, I will be more successful in the future" and the "financial planning time

horizon" are applied as a measure of time preference and function as a discount rate for future utility. Individuals with a longer saving horizon and those who can distinguish between today's work and tomorrow's payoff have greater savings. According to the permanent income theory, this is because individuals with lower discounted future utility are more inclined to divide their money evenly to smooth consumption over a lifetime (Friedman, 1957).

In addition, the demographic variables did not perform well, neither in combination with the other categories nor by themselves. However, this outcome must be interpreted with caution. For example, the explanatory variable gender is a binary variable whose value is either male or female (e.g., 0 or 1). Thus, there is less potential for machine learning algorithms to differentiate between saving levels when there are just two choices, as opposed to when there are nine options, as with household income. However, recent research on gender differences in savings behavior found no statistically significant differences across genders (Whitaker et al., 2013). All algorithms placed a greater emphasis on household income than age, and Table 7 reveals that the XGBoost algorithm gave education a greater weight compared to what the SVM algorithm gave age.

Situational factors showed that stress and health was of lesser importance when predicting saving behavior. Parent-child interactions over money “Did your family educate you about saving, credit, allowance, or other finances when growing up? had a great impact on adult saving behavior. This is consistent with recent studies indicating that early financial socialization influences savings and long-term planning habits (Kim & Chatterjee, 2013). Omitted variable bias may have an upward bias on this question as families with higher savings may introduce saving earlier to encourage saving behavior. Nonetheless, family conversations affect savings.

The category 'financial literacy' contained questions such as 'I know how to make myself to save, which are highly related to savings. This variable had a 0.22 correlation with savings and is a direct measure of an individual's ability to save. This can be interpreted in two distinct ways. As seen in the correlation matrix in Appendix A, the first interpretation is a robustness check to ensure that survey respondents provided truthful information. Second, this highlights the relevance of prediction methods in economics literature. Neoclassical economics assume that individuals are rational and want to maximize their utility with perfect information. Nonetheless, this question reveals that individuals who do not have a lot of savings, do not know how to save

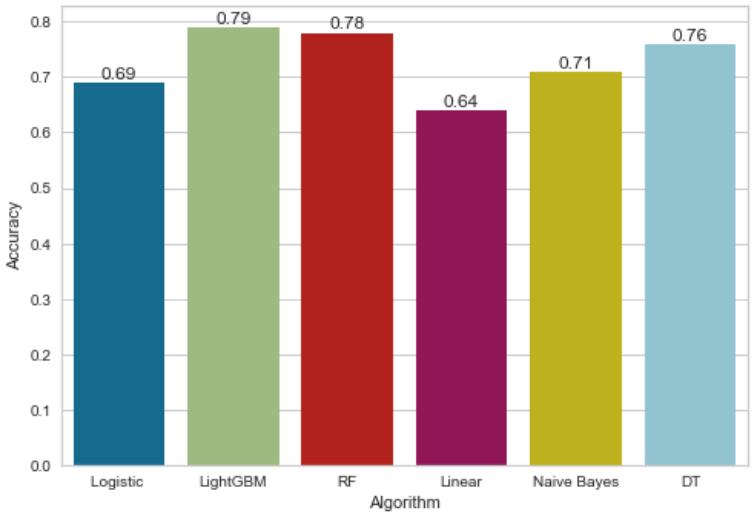
and/or lack the right information, as is often assumed in the literature, to do so. The questions "How good are you at working with percentages?" and "Prefers words for expression of probabilities" both influenced the prediction a lot. In the XGBoost algorithm, these questions were more important than "I know how to make myself save". These two questions are closely related to the core difficulties that a saver must grasp. Knowing how to work with percentages is essential in all aspects of economic challenges and is a daily concern. From purchasing to investing to saving. 83.8 percent of those with incomes of $75,000 or more responded that they are good or extremely good at dealing with percentages, whereas just 18.9 percent of those with incomes of $0-$999 responded in the same way. Consequently, a lack of financial knowledge is a significant obstacle to saving.

As demonstrated by the algorithms in this paper, saving behavior is a complicated topic that requires consideration of the entire individual in question. Previous research has found a significant correlation between age, education, income, and savings, which is supported by the results of this study as well. However, these variables alone do not answer the issue of what determines savings. This study demonstrates that time and time preference play a significant role in determining saving behavior. From the moment an individual was introduced to saving from their parents to the time horizon of their savings and their future utility discount to work hard today in order to receive a payoff tomorrow, to the time they predict passing away. Hence, to understand saving behavior, one must recognize the value of time. This furthermore implies that it is never too late to begin saving.

To further illustrate the effectiveness of ANN, XGBoost, and SVM algorithms, the following ML algorithms and statistical approach were adopted: Linear regression, Logistic regression, Naïve Bayes, Decision Tree, Random Forest, and LightGBM. Figure 11 illustrated the accuracy of each and every algorithm. It can be observed that the accuracy of every method was less than 0.8, however the accuracy of the ANN, XGBoost, and SVM algorithms were all better than 0.8. It demonstrated that ANN, XGBoost, and SVM algorithms were superior to other techniques for predicting saving behavior. The reason for this may be that the chosen methods in this study are nonlinear models. SVM for instance that can both perform as a linear and non-linear method, handles outliers better than the logistic- and linear regression. The logistic regression uses a sigmoid function which has problem with the vanishing/exploding gradient problem which means

that when the derivative becomes negligibly small at either end of the output space, the procedure for updating weights in the function is inefficient which becomes more apparent as the number of observations increase. The ANN algorithms correct for this by using multiple layers before using the sigmoid function in the last layer. As expected, ANN algorithm outperformed the other chosen algorithms since it is an unsupervised learning algorithm that is used to discover hidden patterns in the data.

Figure 11: Accuracy of each Comparison Method



Despite the fact that the suggested technique yields good prediction outcomes, it will be essential in the future to overcome some limitations:

1. The dataset was quite imbalanced, mainly for the $0-999 class relative to the other saving classes. Since the prediction performance of the algorithms is highly dependent on the quantity and quality of datasets, this might have an impact on the middle classes ($1,000-4,999, $5,000-19,999, and $20,000-74,999). Thus, it is important to build a more balanced database.
2. Other variables may also affect the results of the prediction. Numerous variables influence saving behavior, including personal attributes and the external environment. The 25 variables included in this study can characterize the conditions for saving behavior, but additional variables, such as risk aversion, the presence of children, and pension plans like 401(k)s and IRAs, may also have an effect. Consequently, it is essential to investigate the effects of these markers on the prediction outcomes.

Conclusion

One of the most debated topics today, both in the US and Europe, is how to account for the increased amount of elderly. How do public and private policies interact? How much does a state pension replace private retirement savings and if so, to what extent? What effects do alterations in retirement behavior have on the economy? Does social security impact the age at which individuals retire and, by extension, the level of wealth in the economy? More generally, everyone who considers economic development must consider the function of saving in economic growth. For this and related reasons, saving prediction is a crucial task for policy makers in understanding individual and group behavior.

This study investigated the performance of Artificial Neural Network, XGBoost, and Support Vector Machine algorithms for saving behavior prediction. The models were constructed based on a training set (3647) after their hyperparameters were tuned using the five-fold CV method. The test set (1563) was adopted to validate the feasibility of trained models. Overall, the performances of ANN, XGBoost, and SVM algorithms were acceptable, and their prediction accuracies were 0.85, 0.84, and 0.80, respectively. By comprehensively analyzing the accuracy and macro average of precision, recall and F_1, the rank of overall prediction performance stayed the same with ANN being the best and SVM being the worst. According to the precision, recall and F_1 of each class, the prediction performance for *$0-999* and *$75,000 or more* was better than that for the other classes. Based on the importance scores of indicators from the algorithms, XGBoost algorithm, the($1,000-4,999, $5,000-19,999, and $20,000-74,999 classes was not as good. According to the analysis of indicator importance, “financial planning time horizon”, “how likely do you believe it is that you will live beyond age 75?”, and “psychological connectedness” were the most influential indicators on the prediction results. Compared with the linear- and logistic regression and other ML algorithms (Decision Tree, Random Forest, Gaussian naïve Bayes, and LightGBM), the performance of ANN, XGBoost and SVM were better, which further verified that they were reliable and effective for the saving behavior prediction.

I propose to use the 25 questions as a tool for young adults to learn the different aspects of life that relates to saving so that the individual him/herself can change a pattern/behavior that he/she feels is possible to maintain in the long run, even if an individual only improves his/her savings

by one class. The methodology can also be applied in other fields, such as the risk prediction of bankruptcy and assets and liability predictions.

Lastly, I suggest that machine learning techniques can help behavioral economics to become a more predictive research field. By using some of the fundamental concepts and methods of machine learning, I propose that a stronger emphasis on prediction, as opposed to explanation, might eventually lead to a deeper understanding of behavioral economics.

References

Ali, N., Neagu, D. and Trundle, P., (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Applied Sciences, 1(12).

Arya, S., Eckel, C. and Wichman, C., (2013). Anatomy of the credit score. Journal of Economic Behavior & Organization, 95, pp.175-185.

Babiarz, P. and Robb, C., (2013). Financial Literacy and Emergency Saving. Journal of Family and Economic Issues, 35(1), pp.40-50.

Bandura, A., (1990). Perceived self-efficacy in the exercise of personal agency. Journal of Applied Sport Psychology, 2(2), pp.128-163.

Banerjee, S., (2011). How do financial literacy and financial behavior vary by state?. EBRI Notes 32 (11).

Barr, M., (2008). Financial Services, Savings and Borrowing Among Low- and Moderate-Income Households: Evidence from the Detroit Area Household Financial Services Survey. SSRN Electronic Journal,.

Bernheim, B., Ray, D. and Yeltekin, Ş., (2015). Poverty and Self-Control. Econometrica, 83(5), pp.1877-1911.

Beverly, S. G. (1997). How can the poor save? Theory and evidence on saving in low-income households. Working paper number 97-3.Washington University in St. Louis, Center for Social Development.

Bholowalia, P., Kumar, A. (2014). EBK-Means: Clustering Technique Based on Elbow Method and K-Means in WSN. International Journal of Computer Applications. 105(9): 17-24

Boccato, L., Schumacker, E., Fernandes, M., Soriano, D., and Attux, R. (2011). Unorganized machines: From turing's ideas to modern connectionist approaches. International Journal of Natural Computing Research, 2:1, 10.

Bommier, A., Chassagnon, A. and Le Grand, F., (2012). Comparative risk aversion: A formal approach with applications to saving behavior. Journal of Economic Theory, 147(4), pp.1614-1641.

Brown, S., Ghosh, P., Gray, D., Pareek, B. and Roberts, J., (2021). Saving behaviour and health: A high-dimensional Bayesian analysis of British panel data. The European Journal of Finance, 27(16), pp.1581-1603.

Browning, M., Chiappori, P.-A. & Weiss, Y. (2011), `Family economics.

Browning, M., and Lusardi, A. (1996). Household saving: Micro theories and micro facts. Journal of Economic Literature, 34(4),1797–1855

Case, K., Quigley, J. M., and Shiller, R. J. (2005). Comparing wealth effects: The stock market vs. the housing market. Advances in Macroeconomics, 5(1), 1–32.

Chamon, M., Liu, K., and Prasad, E. S. (2010). Income uncertainty and household savings in China. NBER Working paper series. Working paper 16565.

Chang, Y. R. (1994). Saving behavior of US Households in the 1980s:Results from the 1983 and 1986 survey of consumer finance. Financial Counseling and Planning, 5(1), 45–64.

Chaulk, B., Johnson, PJ., and Bulcroft, R., (2003). Effect of Marriage and Children on Financial Risk Tolerance: A Synthesis of Family Development and Prospect Theory. Journal of Family and Economic Issues. 24(3). 257-279

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD x27;16. New York, NY, USA: ACM, pp. 785–794.

Cho, S. H., Gutter, M., Kim, J., and Mauldin, T. (2012). The effect of socialization and information source on financial management behaviors among low-and moderate-income adults. Family and Consumer Sciences Research Journal, 40(4), 417–430.

Darley, J. M., and Batson, C.D., (1973). "From Jerusalem to Jericho": A study of Situational and Dispositional Variables in Helping Behavior". JPSP, 27, 100-108.

Darley, J. M., and Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. Journal of Personality and Social Psychology, 8(4, Pt.1), 377–383

Delafrooz, N., and Paim, L. (2011). Personal saving behavior among Malaysian employees: Demographic comparison. In 2011 Inter-national conference on social science and humanity IPEDR (Vol.5, pp. 361–363)

DeVaney, S. A., Anong, S. T., and Whirl, S. E. (2007). Household savings motives. Journal of Consumer Affairs, 41(1), 174–186.

Epper, T., Fehr, E., Fehr-Duda, H., Kreiner, C. T., Lassen, D. D., Leth-Petersen, S. and Rasmussen, G. N., (2020). Time Discounting and Wealth Inequality. American Economic. 110(4), 1177–1205.

Everitt, B. (2011) “Miscellaneous Clustering Methods”. In: Cluster Analysis. John Wiley & Sons, Ltd, pp. 215–255.

Fisher, P. J., and Montalto, C. P. (2011). Loss aversion and saving behavior: evidence from the 2007 U.S. survey of consumer finances. Journal of Family and Economic Issues, 32(1), 4–14.

Friedman., M. (2016). A theory of consumption function.

Friedman., M. (1957). “The Permanent Income Hypothesis.” National Bureau of Economic Research I: 20–37.

Gelber, A., (2011). How Do 401(k)s Affect Saving? Evidence from Changes in 401(k) Eligibility. SSRN Electronic Journal,.

Gerhard, P., Gladstone, J. J. and Hoffmann, A. O. I., (2018). Psychological characteristics and household savings behavior: The importance of accounting for latent heterogeneity. Journal of Economic Behavior & Organization. 148, 66–82.

Goodfellow, I., Bengio, Y., and Courville, (2016). A. Deep learning. Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press

Grossman, Michael. (1972). The demand for health: a theoretical and empirical investigation. New York: Columbia University Press for the National Bureau of Economic Research

Guidolin, M., and La Jeunesse, E. A. (2007). The decline in the Us personal saving rate: Is it real and is it a puzzle? Federal Reserve Bank of St. Louis Review, 89(6), 491–514.

Gutter, M. S., Hayhoe, C. R., and Wang, L. (2007). Examining participation behavior in defined contribution plans using the transtheoretical model of behavior change. Financial Counseling and Planning, 18(1), 46–60.

Haron, S. A., Sharpe, D. L., Abdel-Ghany, M., and Masud, J.(2013). Moving up the savings hierarchy: Examining savings motives of older Malay Muslim. Journal of Family and Eco-nomic Issues, 34(3), 314–328.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer series in statistics. New York, NY: Springer.

Haushofer, J. and Fehr, E., (2014). On the psychology of poverty. Science, 344(6186), pp.862-867.

Jadhav S, He H, and Jenkins K, (2017). An academic review: applications of data mining techniques in finance industry, International Journal of Soft Computing and Artificial Intelligence, Volume 4, Issue 1, Pages 79 – 95.

Johannisson, I. (2008). Private pension savings: gender, marital status and wealth-evidence from Sweden in 2002. Licentiate thesis, Department of Economics, School of Business, Economics and Law, University of Gothenburg

Jones, I. C. (2014). Macroeconomics, Third Edition. Stanford University, WW Norton Co

Jung, Y. (2018) Multiple predicting K-fold cross-validation for model selection. J. Nonparametric. Stat. 30, 197–215.

Juster, F. T., Lupton, J., Smith, J. P., and Stafford, F. (2004). The decline in household saving and the wealth effect. Review of Economics and Statistics, 88(1), 20–27.

Kahneman, D. (1979), `Prospect theory: An analysis of decisions under risk', Econometrica 47, 278.

Katona, G., (1975). Psychological economics. New York: Elsevier Scientific.

Kevin P. Murphy. (2021). Probabilistic Machine Learning: An introduction. MIT Press

Keynes, J. M. (1936). The general theory of employment, interest and money (pp. 47–68). New York: Harcourt, Brace & Co.

Kim, J., and Chatterjee, S. (2013). Childhood financial socialization andyoung adults' financial management. Journal of Financial Counseling and Planning, 24(1), 61–79

Kim, J., LaTaillade, J., and Kim, H. (2011). Family process and adolescents' financial behaviors. Journal of Family Economic Issues,32(4), 668–679

Kumar, P. (2019) Machine Learning Quick Reference; Packt Publishing Ltd.: Birmingham, UK.

Kytle, J. and Bandura, A., (1978). Social Learning Theory. Contemporary Sociology, 7(1), p.84.

Lea, S. E. G., Webley, P., and Walker, C. M. (1995). Psychological fac-tors in consumer debt: Money management, economic socialization, and credit use. Journal of Economic Psychology, 16(4),681–701

Lee, S., Park, M. H., and Montalto, C. P. (2000). The effect of family lifecycle and financial management practices on household saving patterns. International Journal of Human Ecology, 1(1), 79–93

Lee, MJ., Sherman, DH., (2015) Savings Goals and Saving Behavior from a Perspective of Maslow's Hierarchy of Needs. Journal of Financial Counseling and Planning, 26(2).

Loibl, C., Grinstein-Weiss, M., Zhan, M., and Red Bird, B. (2010). More than a penny saved: Long-term changes in behavior among savings program participants. Journal of Consumer Affairs, 44(1),98–126.

Lown, J., Kim, J., Gutter, M. S., and Hunt, A. T. (2015). Self-efficacy and savings among middle- and low-income households. Journal of Family and Economic Issues, 36(4), 491–502

Lunt, P. K., and Livingstone, S. M. (1991). Psychological, social and economic determinants of saving: Comparing recurrent and total savings. Journal of Economic Psychology, 12(4), 621–641.

Lusardi, A. and Mitchell, O. S., (2013). The Economic Importance of Financial Literacy: Theory and Evidence. SSRN Electronic Journal.

Madhulatha, T. S. (2012). An Overview on Clustering Methods. 2(4): 719–725

Magendans, J., Gutteling, J. M., and Zebel, S. (2017). Psychological determinants of financial buffer saving: The influence of financial risk tolerance and regulatory focus. Journal of Risk Research,20(8), 1076–1093

Mahalingam, P. R. and Vivek, S., (2016). Predicting Financial Savings Decisions Using Sigmoid Function and Information Gain Ratio. Procedia Computer Science [online]. 93, 19–25.

Marblestone, A., Wayne, G. and Kording, K., (2016). Toward an Integration of Deep Learning and Neuroscience. Frontiers in Computational Neuroscience, 10.

Metin-Ozcan, K., Gunay, A., and Ertac, S. (2012). Macro and socio-economic determinants of Turkish private savings. Journal of Economic Cooperation and Development, 33(2), 93–130

Modigliani, F., (1986), Life cycle, individual thrift, and the wealth of nations. AM. Econ, Rev. 76 (3), 297-313

Modigliani, F. and Brumberg, R.H. (1954) Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data. In: Kurihara, K.K., Ed., Post-Keynesian Economics, Rutgers University Press, New Brunswick, 388-436.

Muradoglu Gulnur and Taskin Fatma. (1996). Differences in household savings behavior: Evidence from industrial and developing countries

Mullainathan, S., and Shafir, E. (2013). "Scarcity: Why Having Too Little Means so Much." New York: Henry Holt and Company.

Niculescu-Aron, I. and Mihăescu, C., (2012). Determinants of Household Savings in EU:What Policies for Increasing Savings?. Procedia - Social and Behavioral Sciences, 58, pp.483-492.

Nwankpa, C., Ijomah, W., Gachagan, A., and Stephen Marshall. (2020). Activation functions: Comparison of trends in practice and research for deep learning. 12

Pedregosa, F. Varoquaux, G. Gramfort, A. Michel, V. Thirion, B. Grisel, O. Blondel, M. Prettenhofer, P. Weiss, R. Dubourg, V. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Perry, V. G., and Morris, M. D. (2006). Who is in control? The role of self-perception, knowledge, and income in explaining consumer financial behavior. Journal of Consumer Affairs, 39(2),299–313.

Puri, M. and Robinson, D., (2007). Optimism and economic choice. Journal of Financial Economics, 86(1), pp.71-99.

Rauthmann, J., Sherman, R., Nave, C. and Funder, D., (2015). Personality-driven situation experience, contact, and construal: How people's personality traits predict characteristics of their situations in daily life. Journal of Research in Personality, 55, pp.98-111.

Rha, J. Y., Montalto, C. P., and Hanna, S. D. (2006). The effect of self-control mechanisms on household saving behavior. Financial Counseling and Planning, 17(2), 3–16.

Remble, A. A., Marshall, M. I., and Keeney, R. (2014). Household saving behavior and the influence of family-owned business. Journal of Family and Economic Issues, 35(3), 411–422

Rojas, R. (1996). Neural Networks: A Systematic Introduction. Springer-Verlag, Berlin, Heidelberg,

Ross, L., Nisbett, R. and Gladwell, M., (2011). The person and the situation. 1st ed. Pinter Martin Ltd.

Russell, S., and Norvig, P. (2009). Artificial Intelligence: A Modern Approach. Prentice Hall Press, USA, 3rd edition.

Scheier, M., Carver, C. and Bridges, M., (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. Journal of Personality and Social Psychology, 67(6), pp.1063-1078.

Schmidhuber, J. (2015) Deep Learning in Neural Networks: An Overview. Neural Networks, 61, 85-117.

Schuchardt, J., Hanna, S. D., Hira, T. K., Lyons, A. C., Palmer, L., and Xiao, J. J. (2009). Financial literacy and education research priorities. Journal of Financial Counseling and Planning, 20(1),84–95

Sherer, J. M., Jaddus, J. E., Mercadente, B., Prentice-Dunn, S., Jacobs,B., and Rogers, R. W. (1982). The self-efficacy scale: Construction and validation. Psychological Reports, 51(2), 663–671.

Sundén, A. E., and Surette, B. J. (1998). Gender differences in the allocation of assets in retirement savings plans. The American Eco-nomic Review, 88(2), 207–211

Thaler. R.H., Shefrin, H.M., (1981). The Behavioral Life-Cycle Hypothesis. Economic Inquiry 26(4). 609-643

Thaler. R.H., Shefrin, H.M., (1988). An economic theory of self-control. J. Polit. Econ 89(2), 392-406

Thaler, R. and Benartzi, S., (2004). Save More Tomorrow™: Using Behavioral Economics to Increase Employee Saving. Journal of Political Economy, 112(S1), pp.164-S187.

Tsymbal, O. (2022). Machine Learning Algorithms For Business Applications.

Van Raaij, W. and Gianotten, H., (1990). Consumer confidence, expenditure, saving, and credit. Journal of Economic Psychology, 11(2), pp.269-290.

Vanden Abeele, P. (1988). Economic Agents' Expectations in a Psychological Perspective. In: van Raaij, W.F., van Veldhoven, G.M., Wärneryd, KE. (eds) Handbook of Economic Psychology. Springer, Dordrecht.

Vinod Nair and Geoffrey Hinton. (2010). Rectified linear units improve restricted Boltzmann machines vinod nair. volume 27, pages 807–814, 06.

Vohs, K., Baumeister, R. and Schmeichel, B., (2012). Motivation, personal beliefs, and limited resources all contribute to self-control. Journal of Experimental Social Psychology, 48(4), pp.943-947.

Wakita, S., Fitzsimmons, V. S., and Liao, T. F. (2000). Wealth: Determinants of savings net worth and housing net worth of pre-retired households. Journal of Family and Economic Issues, 21(4),387–418.

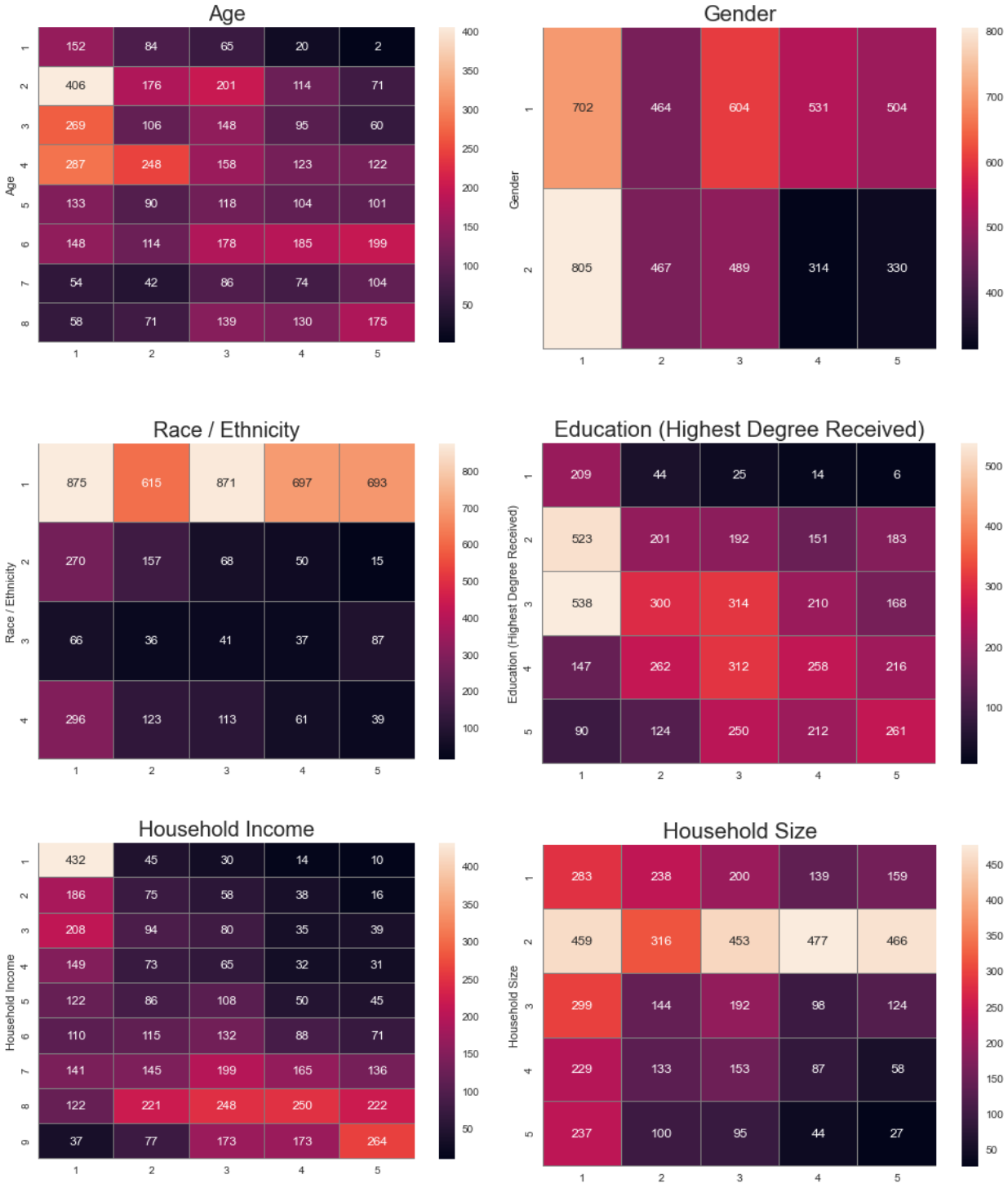
Walden, M. L. (2012). Will households change their saving behavior after the “Great Recession”? The role of human capital. Journal of Consumer Policy, 35(2), 237–254.

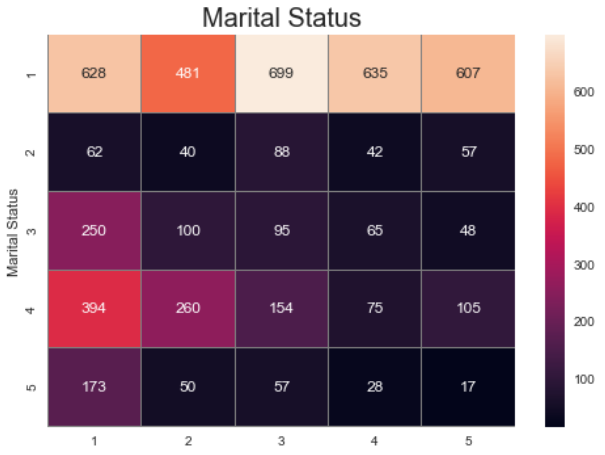
Whitaker, E. A., Bokemeiner, J. L., and Loveridge, S. (2013). International associations of gender on savings behavior: Showing gender’s continued influence on economic action. Journal of Family and Economic Issues, 34(1), 105–119.

Yilmazer, T. (2010). The profile and determinants of household savings. Report for the World Bank.

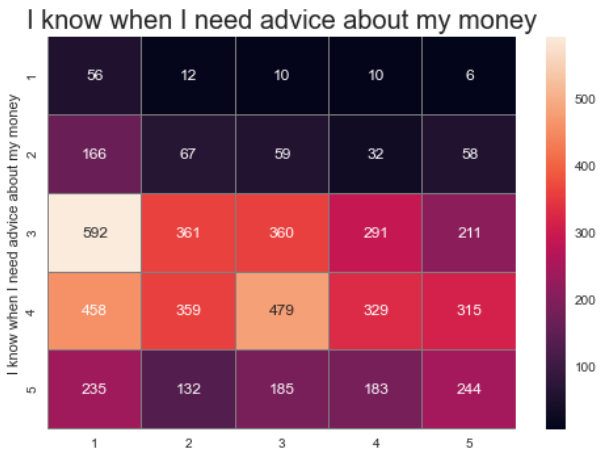
Appendix

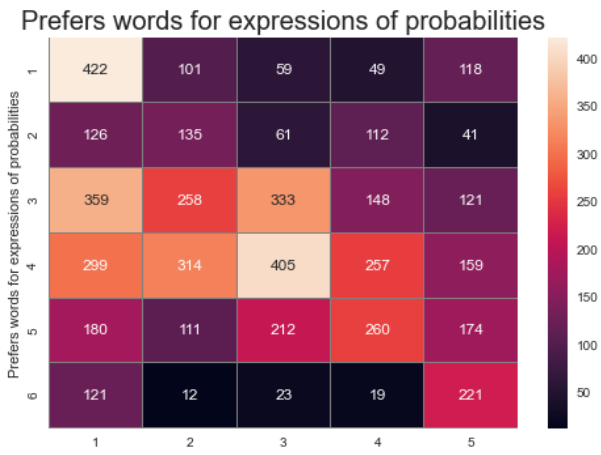
Appendix A. Full EDA

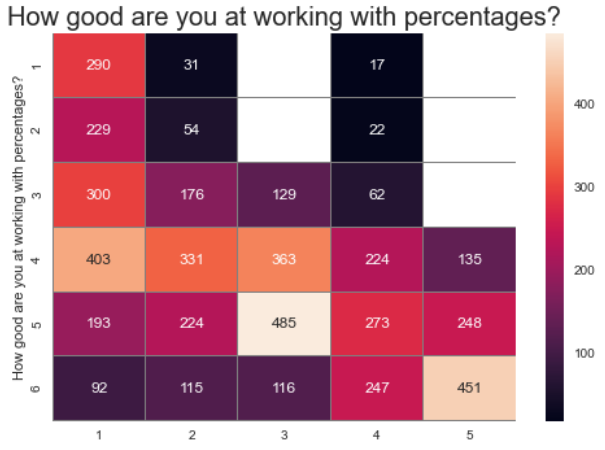


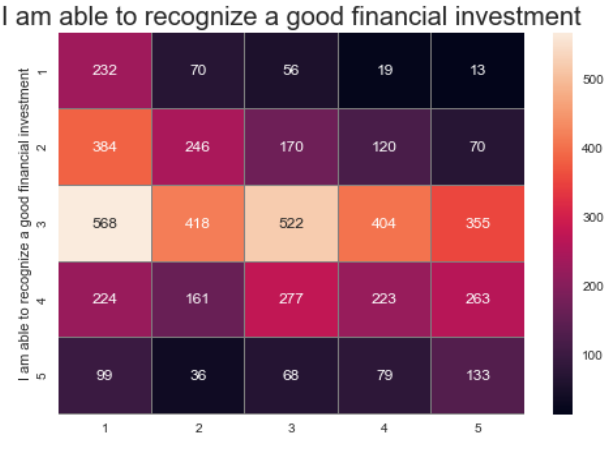

Marital Status
Marital Status
1 2 3 4 5
628 481 699 635 607
62 40 88 42 57
250 100 95 65 48
394 260 154 75 105
173 50 57 28 17
600 500 400 300 200 100

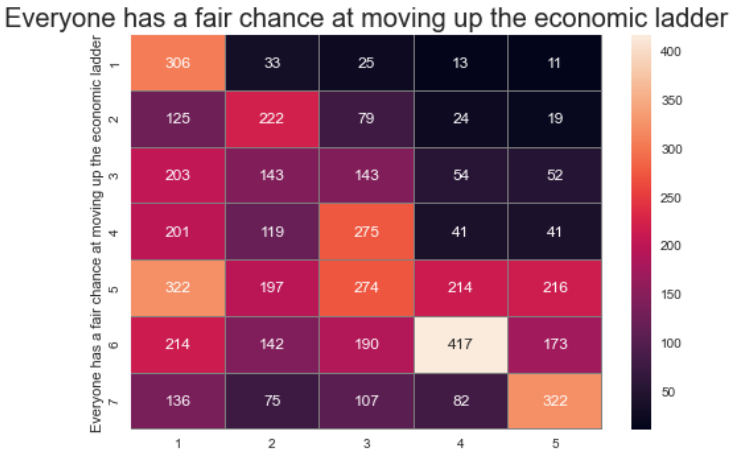

I know how to make myself save
I know how to make myself save
1 2 3 4 5
295 7 4 1 1
261 213 31 7 3
549 363 418 84 34
248 261 443 652 372
154 87 197 101 424
600 500 400 300 200 100

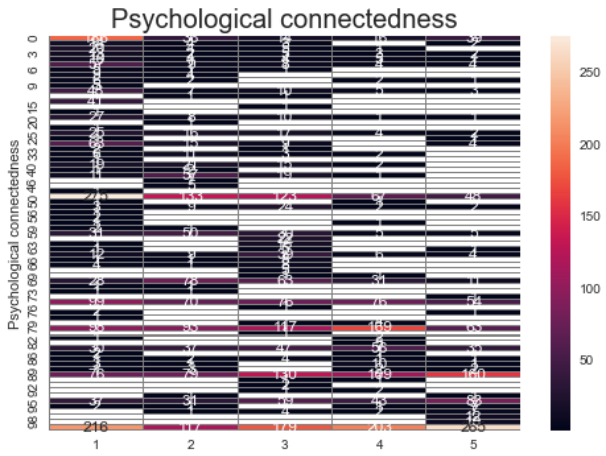

I know when I need advice about my money
I know when I need advice about my money
1 2 3 4 5
56 12 10 10 6
166 67 59 32 58
592 361 360 291 211
458 359 479 329 315
235 132 185 183 244
500 400 300 200 100


Prefers words for expressions of probabilities
Prefers words for expressions of probabilities
1 2 3 4 5 6
422 101 59 49 118
126 135 61 112 41
359 258 333 148 121
299 314 405 257 159
180 111 212 260 174
121 12 23 19 221
400 350 300 250 200 150 100 50

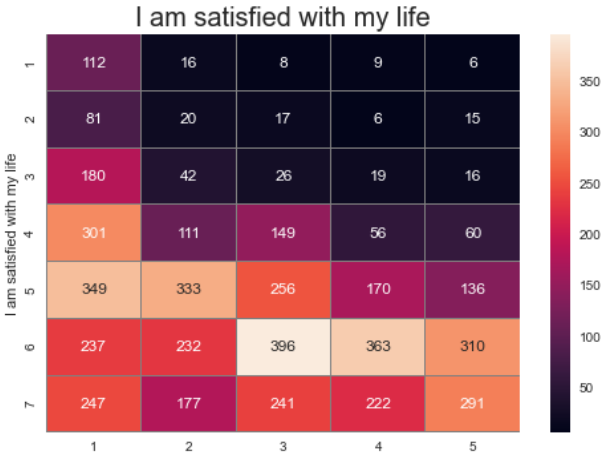

How good are you at working with percentages?
How good are you at working with percentages?
1 2 3 4 5 6
290 31 17
229 54 22
300 176 129 62
403 331 363 224 135
193 224 485 273 248
92 115 116 247 451
400 300 200 100


I am able to recognize a good financial investment
I am able to recognize a good financial investment
1 2 3 4 5
232 70 56 19 13
384 246 170 120 70
568 418 522 404 355
224 161 277 223 263
99 36 68 79 133
500 400 300 200 100

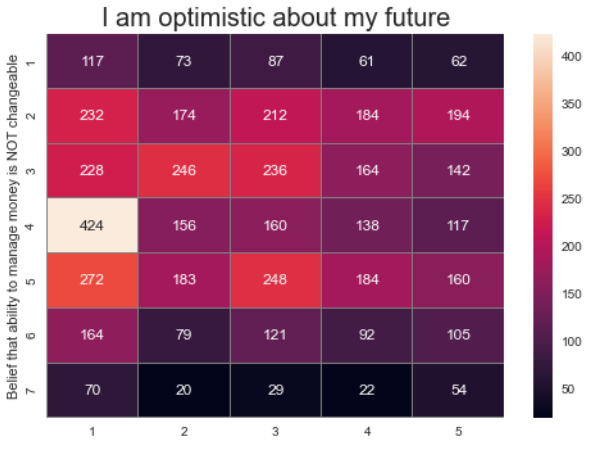

Everyone has a fair chance at moving up the economic ladder
Everyone has a fair chance at moving up the economic ladder
306 33 25 13 11
125 222 79 24 19
203 143 143 54 52
201 119 275 41 41
322 197 274 214 216
214 142 190 417 173
136 75 107 82 322
1 2 3 4 5
400 350 300 250 200 150 100 50


Psychological connectedness
Psychological connectedness
1 2 3 4 5
250 200 150 100 50

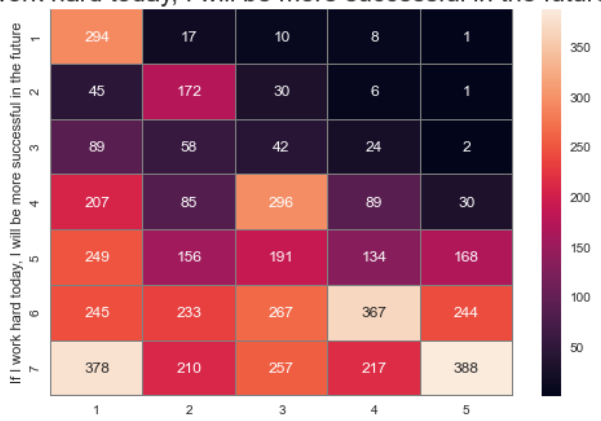

Financial planning time horizon
Financial planning time horizon
702 147 88 33 9
227 333 117 54 25
300 230 453 158 188
190 139 294 479 211
88 82 141 121 401
1 2 3 4 5
700 600 500 400 300 200 100


I am satisfied with my life
I am satisfied with my life
112 16 8 9 6
81 20 17 6 15
180 42 26 19 16
301 111 149 56 60
349 333 256 170 136
237 232 396 363 310
247 177 241 222 291
1 2 3 4 5
350 300 250 200 150 100 50

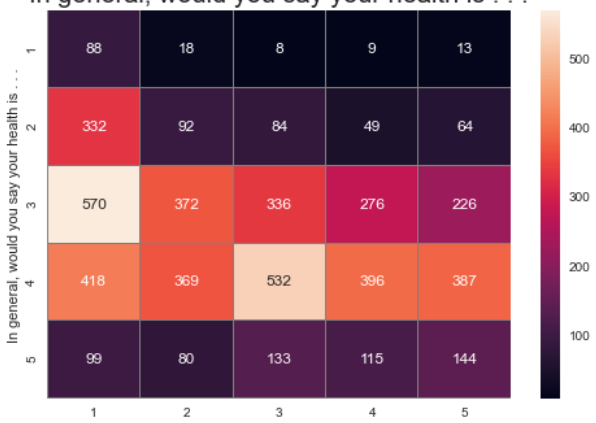

I am good at resisting temptation
I am good at resisting temptation
180 115 14 8 1
459 315 206 122 101
674 401 723 565 521
194 100 150 150 211
1 2 3 4 5
700 600 500 400 300 200 100


I am optimistic about my future
Belief that ability to manage money is NOT changeable
117 73 87 61 62
232 174 212 184 194
228 246 236 164 142
424 156 160 138 117
272 183 248 184 160
164 79 121 92 105
70 20 29 22 54
1 2 3 4 5
400 350 300 250 200 150 100 50

If I work hard today, I will be more successful in the future

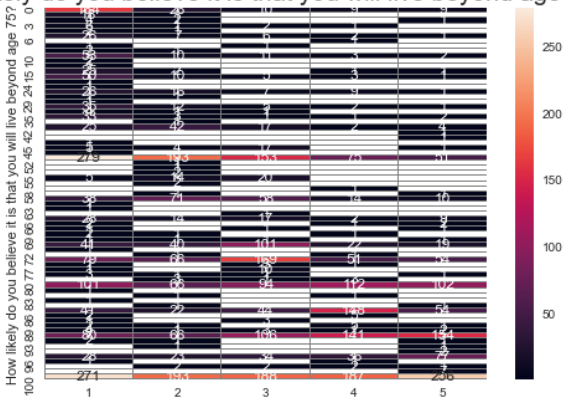


In general, would you say your health is . . .

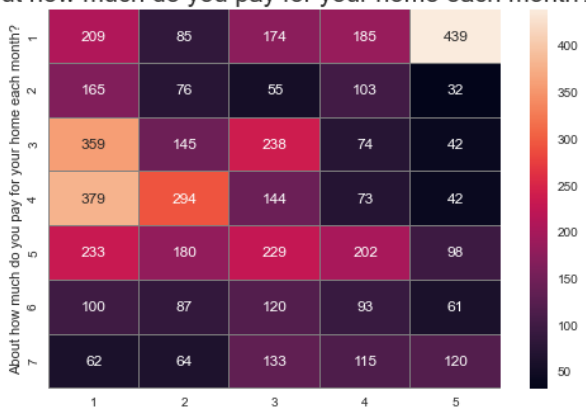


Lot of stress in respondent's life

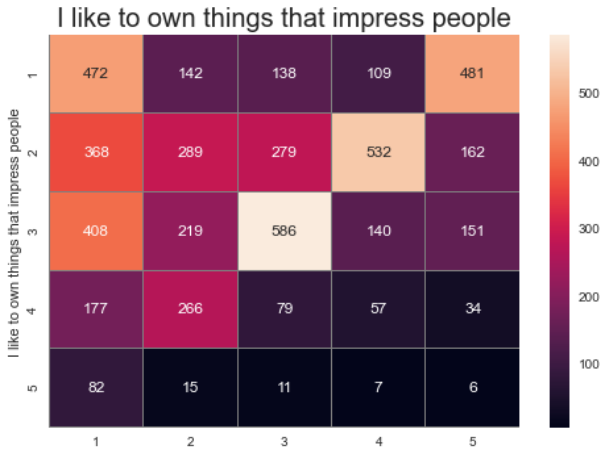
How likely do you believe it is that you will live beyond age 75?

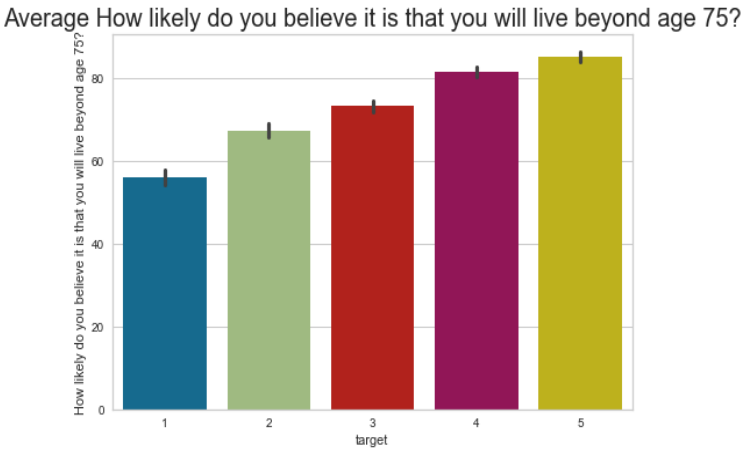


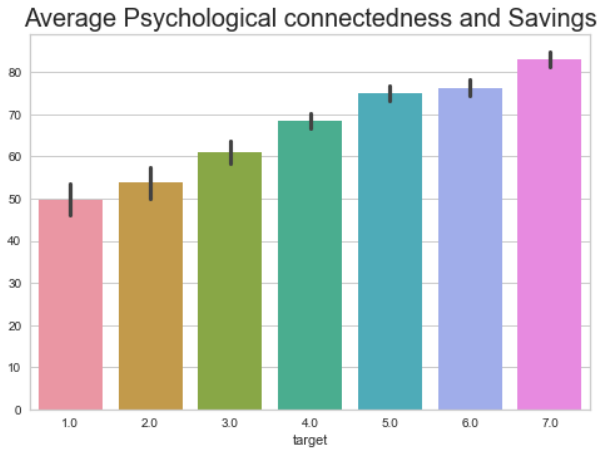
About how much do you pay for your home each month?



While growing up, did your family do any of the following?


I like to own things that impress people
I like to own things that impress people
1
2
3
4
5
472 142 138 109 481
368 289 279 532 162
408 219 586 140 151
177 266 79 57 34
82 15 11 7 6
500
400
300
200
100


Average How likely do you believe it is that you will live beyond age 75?
How likely do you believe it is that you will live beyond age 75?
0
20
40
60
80
1
2
3
4
5
target


Average Psychological connectedness and Savings
0
10
20
30
40
50
60
70
80
1.0
2.0
3.0
4.0
5.0
6.0
7.0
target

Appendix B. List of the 25 questions used after the EDA

Age

1. 18-24
2. 25-34
3. 35-44
4. 45-54
5. 55-61
6. 62-69
7. 70-74
8. 74+

Gender

1. Male
2. Female

Race / Ethnicity

1. White, Non-Hispanic
2. Black, Non-Hispanic
3. Other, Non-Hispanic
4. Hispanic

What is the highest level of education anyone in your household (including yourself) has completed?

1. Less than high school
2. High School degree/GED
3. Some college
4. Associate degree
5. Bachelors' degree
6. Graduate/professional degree

Household Income

1. Less than $20,000
2. $20,000-29,999
3. $30,000-39,999
4. $40,000-49,999
5. $50,000-59,999
6. $60,000-74,999
7. $75,000-99,999
8. $100,000-149,999
9. $150,000 or more

Household Size

1. 1
2. 2
3. 3
4. 4
5. 5+

Marital Status

1. Married
2. Widowed
3. Divorced/Separated
4. Never married
5. Living with partner

I know how to make myself save.

1. Not at all
2. Very little
3. Somewhat
4. Very well
5. Completely

I know when I need advice about mt money

1. Never
2. Rarely
3. Sometimes
4. Often
5. Always

Prefers words for expressions of probabilities

1. Always prefer words
2. -
3. -
4. -
5. -
6. Always prefer numbers

How good are you at working with percentages?

1. Not good at all
2. –
3. –
4. –
5. –
6. Extremely good

I am able to recognize a good financial investment

1. Not at all
2. Very little
3. Somewhat
4. Very well
5. Completely

Everyone has a fair chance at moving up the economic ladder

1. Strongly disagree
2. Disagree
3. Somewhat disagree
4. Neither agree nor disagree
5. Somewhat agree
6. Agree
7. Strongly agree

Psychological Connectedness

Please think about the important characteristics that make you the person you are now – your personality, temperament, major likes and dislikes, beliefs, values, ambitions, life goals, and ideals – and please rate the degree of connectedness between the person you expect to be in 5 years compared to the person you are now, where 0 means "I will be completely different in the future" and 100 means "I will be exactly the same in the future."

ENTER NUMBER_____

In planning your and/or your family's saving and spending, which of the time periods is most important?

1. The next few months
2. The next year
3. The next few years
4. The next 5 to 10 years
5. Longer than 1o years

I am satisfied with my life

1. Strongly disagree
2. –
3. –
4. –
5. –
6. –
7. Strongly agree

I am optimistic about my future

1. Strongly disagree
2. –
3. –
4. –
5. –
6. –
7. Strongly agree

If I work hard today, I will be more successful in the future

1. Strongly disagree
2. –
3. –
4. –
5. –
6. –
7. Strongly agree

I am good at resisting temptation

1. Not at all
2. Not very well
3. Very well
4. Completely well

In general, would you say your health is…

1. Poor
2. Fair
3. Good
4. Very good
5. Excellent

I have a lot of stress in my life

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

How likely do you believe it is that you will live beyond age 75? Use the scale from zero to 100 to indicate your response. 0 would mean not at all likely and 100 would mean that it is certain. You can choose any number between 0 and 100.

ENTER NUMBER_____

About how much do you pay for your home each month

1. Less than $300
2. $300-499
3. $500-749
4. $750-999
5. $1,000-1,499
6. $1,500-1,999
7. $2,000-2,999
8. £3,000-4,999
9. $5,000 or more

Did your family educate you about saving, credit, allowance, or other finances when growing up?

1. Discuss family financial matters with me
2. Spoke to me about the importance of saving
3. Discussed how to establish a good credit rating
4. Taught me how to be a smart shopper
5. Taught me that my actions determine my success in life
6. Provided me with a regular allowance
7. Provided me with a savings account

Answers in column:

My family did ___ out of the 7 previous statements.

I like to own things that impress people

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Appendix C. Mathematical Derivation of the XGBoost algorithm

Given a data collection with n observations and m variables $D = \{x_i, y_i\}|D| = n, x_i \in \mathbb{R}^m, y_i \in \{0,1\}$, *H* additive functions are used to predict the output of a tree ensemble model:

$$p_i = \sum_{h=1}^{H} f_h(x_i), f_h \in S$$

Where *S* is the classification tree space $S = \{f(x) = w_{q(x)}\}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T)$. In the classification tree space, q is defined as the structure of a tree with **T** leaves. f_h represents an independent tree structure q and leaf weights *w*.

The XGBoost objective function includes a training loss term l and a regularization term **Ω**. The training loss term, also known as the loss function, measures the model's fit to the training data, whereas the regularization term measures the trees' complexity. This function is intended to be minimized for learning

$$L(p_i) = \sum_{i=1}^{n} l(y_i, p_i) + \sum_{k=1}^{K} \Omega(f_k) \quad , \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$$

Where K is different tree structures, T is the number of leaves in the tree and w is the leaf weights. *l* is simply any differentiable convex loss function that measures the difference between the prediction $\hat{y}$, and the target y_i (e.g., a Log-loss function). Ω is the so-called regularization function and controls the complexity (to avoid overfitting), with γ and λ being tuning parameters. Complex models with several leaf notes (i.e., larger T) are penalized with λ, but the penalty can be adjusted using γ.

To minimize the objective function, f_t should be added to $p_i^{(t)}$, which represents the prediction of the i^{th} instance at the t^{th} iteration. Hence, it may be derived as

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, p_i^{t-1} + f_t(x_i)\right) + \Omega(f_t)$$

Gradient descent can be used to optimize the objective function. This iterative algorithm minimizes the given function. Calculating the first and second order gradient over the predictive objective p_i at the t^{th} iteration is required for gradient descent optimization. Due to the absence of the objective function's derivative, the objective function is approximated to the second order of Taylor. As a result of this:

$$L^{(t)} \sim \sum_{i=1}^{n} \left[l(y_i, p_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + (f_t)$$

Where g_i and h_i represents the first and the second order derivative (gradients) of the loss function. They can be defined as: $g_i = \partial_{\hat{y}_i{}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}_i{}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$.

The objective function may be rewritten and simplified as follows:

$$\bar{L}^{(t)} = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} gi\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} \mathrm{h_i} + \lambda\right)w_j^2\right] + \gamma T$$

For a given tree structure q(x) , the optimal weight w_j^* of leaf j can be derived as:

$$\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}$$

Thus, the ideal value of the objective function corresponding to the above equation is:

$$\bar{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T$$

To conclude, the optimization of objective function is converted to a problem of determining the minimum of a quadratic function.

Appendix D. Optimized Hyperparameters

Algorithm	Symbol	Hyperparameters	Meanings	Best Scale
Artificial Neural Network	ANN	Hidden layers	Number of hidden layers	3
		Epochs	Number of training epochs	100
		Batch size	Mini-batch training size	500
		η	Learning rate	0.001
Extreme Gradient Boosting	XGBoost	n_estimators	Number of trees	100
		Learning_rate	Shrinkage coefficient of each tree	0.1
		Max_depth	Maximum depth of a tree	5
		Min_samples_leaf	Minimum number of samples for leaf nodes	7
		Min_samples_split	Minimum number of samples for nodes split	7
Support Vector Machine	SVM	Kernel	Radial basis function	Rfb
		C	Cost	0.1
		γ	Gamma	5

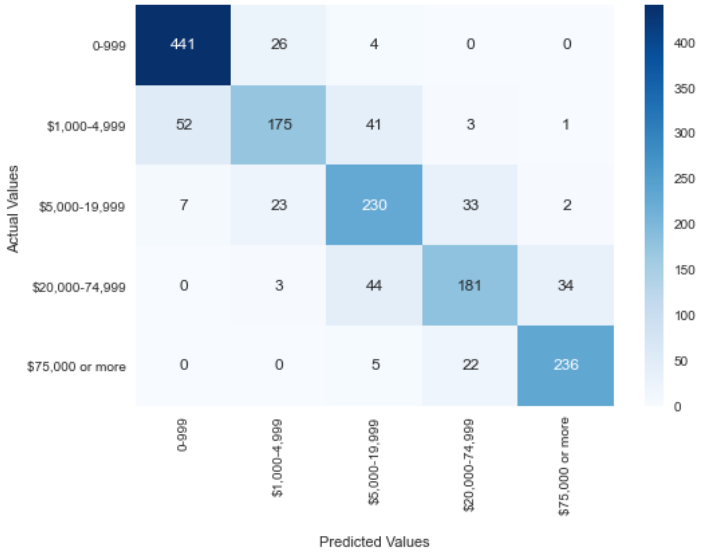
Appendix E. (*k*=3) Robustness Check

	Accuracy	Macro – Pr	Macro – Re	Macro F_1
ANN	0.87	0.86	0.86	0.87
XGBoost	0.85	0.86	0.85	0.85
SVM	0.81	0.80	0.80	0.80

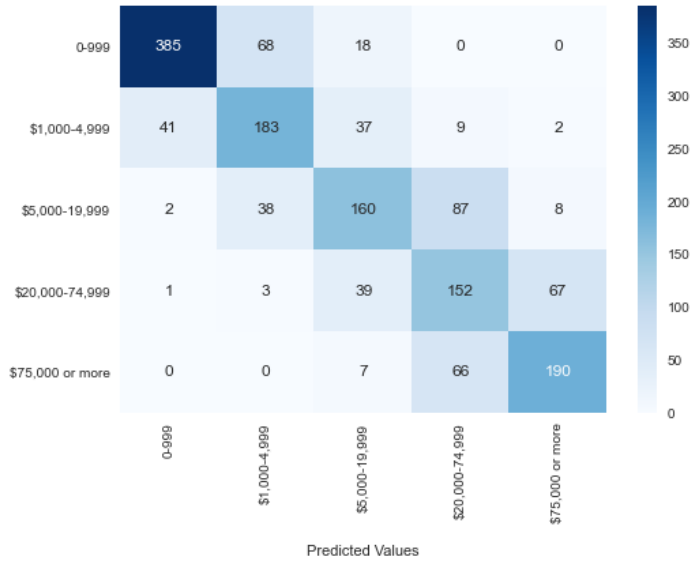
Note: Total support for each algorithm was 1563 with the following distribution: $0-999=251, $1,000-74,999=836 and $75,000 or more=251.

Appendix F. Confusion Matrix

XGBoost



SVM



Appendix G. Classification Variable Importance

	Classification Variable Importance	(1)	(2)	(3)	(4)
Demographic	Age	21.3			
	Gender	6.6			
	Race / Ethnicity	7.2			
	Education	16.1			
	Household Income	26.2			
	Household Size	13.8			
	Marital Status	8.8			
Financial	I know how to make myself save		22.5		
	I know when I need advice about my money		15.7		
	Prefers words for expression of probabilities		27.2		
	How good are you at working with percentages?		26.1		
	I am able to recognize a good financial investment		8.5		
Psychological	Everyone has a fair chance at moving up the economic ladder			15.6	
	Psychological Connectedness			21.4	
	Financial planning time horizon			12.1	
	I am Satisfied with my life			11.4	
	I am good at resisting temptation			16.7	
	I am optimistic about my future			10.8	
	If I work hard today, I will be more successful in the future			12.0	
Situational	In general, would you say your health is…				9.7
	Lot of stress in respondent's life				13.5
	How likely do you believe it is that you will live beyond age 75?				25.6
	About How much do you pay for your home each month				16.8
	Did your family educate you about money, saving, etc growing up??				22.3
	I like to own things that impress people				12.1