Lappeenranta-Lahti University of Technology LUT
School of Business and Management
Strategy, Innovation, and Sustainability (MSIS)

APPLYING MACHINE LEARNING IN PREDICTING GROSS DOMESTIC SAVINGS THE CASE OF FINLAND

Master's Thesis

1st supervisor: Kaisu Puumalainen

2nd supervisor: Pontus Huotari

Doan Thi Thanh Thanh

2020

ABSTRACT

Author	Doan Thi Thanh Thanh
Title	Applying Machine Learning in Predicting Gross Domestic Savings. The case of Finland
Faculty	School of Business and Management
Degree programme	Strategy, Innovation, and Sustainability
Year of completion	2020
Master's Thesis	Lappeenranta-Lahti University of Technology LUT 86 pages, 16 figures, 4 tables, and 3 appendixes
Supervisor	Kaisu Puumalainen, Pontus Huotari
Keywords	*gross domestic savings; GDS; saving determinants; life cycle model; machine learning; prediction*

The main objective of this quantitative study is to predict gross domestic savings in Finland with the help of machine learning algorithms. Machine learning-based models and the traditional time-series model were assessed by model estimation and model performance evaluation to answer the question of whether machine learning models can yield more accurate forecasts than the simpler time-series models.

The thesis was conducted in Helsinki, Finland, during winter 2019 and spring 2020. Secondary data was gathered and collected from three different statistic sources in Finland.

The empirical analysis identified that the critical factors of the domestic saving rate in Finland during the period of Q1 1995-Q4 2019 are financial and income variables. All machine learning models produced forecasts that had a lower prediction error than the traditional linear model. In short, the empirical results indicated that by applying machine learning models, better predicting performance could be achieved in terms of gross domestic saving.

ACKNOWLEDGMENTS

I would like to extend my sincere appreciation to my supervisors, Kaisu Puumalainen and Pontus Huotari, who help me a lot since the beginning of the thesis writing process. Thank you very much for always answering my questions and concerns throughout the preparation and finalization of my thesis. I can overcome the confusedness during the theoretical review and methodology development stages, all thanks to your prompt guidance and comprehensive advice.

I also want to thank all my dear friends, Chi Tran, Thao Le, Oanh Truong, and Veronika Piontek, whose assistance and academic supports are always valuable to me.

I would also want to show my gratitude to the professors and the members of staff at the Lappeenranta-Lahti University of Technology LUT for always being so supportive and helpful to all international students and me.

10.6.2020
Doan Thi Thanh Thanh

TABLE OF CONTENTS

List of tables

List of figures

List of Abbreviations

GDS – GDS

GDP – Gross Domestic Products

SVM – Support Vector Machine - Regression

RSS – Residual Sum of Squares

MAD – Mean Absolute Deviation

RMSE – Root Mean Square Error

1 INTRODUCTION

1.1 Background of the study

Gross domestic savings (GDS) is a critical indicator of the economic growth and sustainable development of a country (Duran et al., 2017, 45). To elaborate, it is a crucial macroeconomic indicator that helps finance investments, creates more job opportunities, enhances the level of productivity, and thus reinforces the country's economic growth (Ahmed et al., 2015, 63-71). According to Bebczuk (2000), an increase in domestic savings can lead to an enhancement in the average future economic growth rate of a country. On the other hand, low domestic savings may intensify the dependence on foreign capital inflows, constitute more burdens for the economy with foreign exchange liabilities, create fiscal vulnerability, and endanger the sustainability growth. (Houérou, 2011).

GDS rate is an interesting variable to predict because different groups of variables determine it. Firstly, GDS is equal to gross domestic products (GDP) subtract final consumption spending. In other words, GDS is considered as the total of public savings and private savings, which includes both the numbers of households and firms in the country. (Teng et al., 2018,12). Besides, total domestic savings receive excellent attention not only from researchers but also policy-makers who need to make decisions that build on the incomplete information and vast database about the current economic situations. Previous empirical studies about GDS mostly focus on analyzing the determinants of saving rates. Some empirical studies about GDS such as Metin Ozcan et al. (2003, 1405-1416), Hammad et al. (2010, 23-34), Imran et al. (2010), Jilani et al. (2013), Khan et al. (2017) mostly focus on analyzing the association of GDS determinants. The standard research methods are cross-country model estimation and time series analysis on the national level. Even though the association of GDS determinants has been studied a lot, the future projections of this indicator have received less attention. GDS prediction is a challenging task due to the vast number of influences. In this thesis, the machine learning approach is applied to check whether it is an optimal predicting model for GDS.

Machine learning, which is a separate study area under the umbrella of artificial intelligence, can forecast the future outcomes of events by studying patterns in data with the help of algorithms (Bergstra & Bengio, 2012, 281-305). Motived by relentless improvements in computational power and data availability, machine learning models have gained prominence due to the extensible ability to learn and understand the diversified and complex data sets (Richardson et al., 2018). There is a growing literature trend that assesses the relative success of the machine learning models in macroeconomic prediction over the traditional time-series techniques, which are widely used in quantitative scientific analysis. The traditional predicting approach for time-series data models depends upon a set of different choices that influence model complexity and predicting performance. By utilizing intensive computation models, machine learning approaches can simplify the mentioned set of options efficiently, detect the optimal model complexity and discover complicated hidden correlations faster than the conventional statistical models (Julien & Etienne, 2018, 475). Their advantages raise the question of whether machine learning models can yield more accurate forecasts of macroeconomic indicators.

1.2 Research gap and research questions

The study's objective is to clarify further the future prediction of GDS, one of the essential macroeconomics indicators. The empirical research using machine learning algorithms and some prediction models may be conducted in this study to predict the value of domestic savings of the Finnish economy. The study aims to answer these questions as follows:

RQ 1. What are the main determinants that account for the GDS in Finland?
RQ2. How to conduct the prediction of domestic savings by utilizing the traditional regression models in Finland?
RQ3. How to apply the machine learning model in the prediction of domestic savings in Finland?

RQ4. Do machine learning-based models provide more accurate predictions than the traditional econometric models?

1.3 Scope of the study

It appears that numerous machine learning-based models could be used to predict the gross domestic saving in Finland. As a result, three standard algorithms selected to answer the research questions as follows:

- Regularization Linear Regression (Lasso, Ridge and Elastic Net)
- Support Vector Machine - Regression (SVM)
- Decision Tree Regression

This master's thesis has employed secondary data regarding the main determinants that account for the GDS in Finland across the period of Q1 1995 – Q4 2019. The data may be collected from several websites and reported every quarter. It is crucial to analyze how the prediction models work on the new dataset. Hence the empirical test may be conducted based on not only in – sample but also out – of – sample approach.

1.4 Structure of the study

The empirical study contains six chapters. The first chapter may introduce the background of the study, present research questions, the scope, and the overall structure of the study. The following section mentions the theoretical parts of this study, includes a systematic literature review on the previous empirical studies about the determinants of domestic savings. In chapter 3, the background of machine learning and machine learning algorithms used in this thesis is described. After that, data description and research methodology may be described in chapter 4. Chapter 5 may discuss the empirical results. The final chapter may sum up the findings and answer the four research questions, and then presents the limitations of the thesis and implications for further research in the future.

2 GROSS DOMESTIC SAVINGS

2.1 Introduction of gross domestic savings

Gross domestic savings is a crucial factor that links the economic connection between the preceding and expected growth of a nation. The available amount of GDS in society represents the availability of gross investment and thus create a base for growth rate (Hammad et al., 2010, 23-34). Consistent with the World bank's definition, GDS is calculated as gross domestic product subtract total consumption. The website of the United Nations stated that gross saving is an indicator of economic development or macroeconomic performance, which is briefly defined as disposable income less consumption. This indicator can be calculated for each institutional sector and the entire economy.

Gale et al. (2004, 102-210) showed that this indicator is considered as an essential aspect of national economic development. This economic factor is widely known as a base component that funds domestic investment to accomplish economic growth. Empirical research on cross-country data from Hussain & Brookins (2001, 150-174) proposed an enhancement association of savings, finance, and economic growth, especially emerging nations. Higher savings implies that the country can reserve available funds for investment opportunities, thus can promote economic growth. Shortage of domestic savings can increase the need for foreign funding to support national investments, deepen the external current account deficit, endanger vulnerabilities for the economy, and jeopardize the sustainability of domestic growth. The significant dependence on foreign funding can threaten the country to the jeopardy of a sudden reversal of capital flows, create an adverse impact on the economy. Low levels of savings can not only hinder the country's economic development but also can lead to the risk of economic recession (Oladipo, 2010).

During recent decades, there have been many research papers that investigate the factors that affect domestic savings. The typical time series analysis methods such as VEC, VAR, ARDL, and OLS estimators were used in various empirical studies of macroeconomic

indicators (Narayan, 2006). Khan et al., (2017) reported that the interest rate, income, trade fluctuations, money supply growth, government spending, and financial market development to understand the association between elements of savings rate. According to Metin Ozcan et al. (2003), researchers made great efforts to analyze and understand the determinants that account for GDS.

Most studies about GDS aim to examine the relationship of savings determinants in a particular country. However, there are still a few works on the application of machine learning to the estimation of domestic savings. An accurate prediction of the total domestic savings may provide a reasonable economic development orientation as well as contribute to a proper policy that promotes the sustained economic growth of the country. Therefore, estimating domestic savings is a significant concern in this thesis. With the benefits of domestic saving established, the government needs to evaluate the past trend, understand the main determinants, have an accurate future trend prediction, and adjust policy accordingly to promote this indicator. An accurate forecast of GDS by utilizing machine learning computation power can offer a dramatic difference in terms of timewise. A precise prediction enables policymakers to make adequate decisions or to generate appropriate economic development strategies. Moreover, policymakers may consider different scenarios in the policy stimulation to maximize the benefits of flexible machine learning models. It helps to detect the optimal level of each primary determinant, thus adjust macroeconomic policy accordingly to encourage savings and stimulate investment, eventually increase economic growth (Nagi & Kostoglou, 2010).

Horioka and Hagiwara (2010) conducted an empirical study on the gross domestic savings' prediction in Asia between 1960 and 2007. They forecasted future tendency in household saving rates in the Asian area between 2011 and 2030. Their empirical findings indicated that the age structure of the Asian regional population could play the role of a critical determinant of domestic saving rates in this region in the future.

2.2 Contemporary theories on savings

Since a sound understanding of the factors that account for savings plays an integral part in designing predictive models, the summary of economic theories associated with savings is the first step to help researchers detect determinants essential to savings.

The two contemporary theories that explain the connection between saving and consumption behavior in the private saving sector are the permanent income hypothesis introduced by Milton Friedman in 1957 and the life-cycle hypothesis established by Modigliani in 1954. Their studies connected savings behavior to different lifetime periods of a person, such as the stages of schooling, employment, and retirement. These two hypotheses implied that the consumer behavior of utility-maximizing sense assigned their incomes to level consumption over their different periods of life and hence fluctuated their savings and spending to attain the utility-maximizing goal (Friedman,1957) (Modigliani, 1986, 297-313).

2.2.1 The Life-Cycle Hypothesis

Franco Modigliani developed this theory in 1950, mainly focus on private savings behavior. In line with this hypothesis, people allocated their consumption over their lifetime periods by collecting savings during employment periods and keeping their consumption behavior during the retirement period. Besides, Richard Brumberg established a theory that was formed on the consumption decision of people. He observed that the consumption decision mainly depends on their available assets during their lifetime. Both Brumberg and Modigliani agreed that people achieve assets during their employment period and use them after their retirements. The life-cycle hypothesis was considered as the principal theory that connects people's consumption tendency to their expected personal income (Modigliani and Brumberg, 1954, 24-55) and (Friedman, 1957). The savings decision is the fundamental idea of this theory and inspires many empirical studies on savings behavior.

2.2.2 The Relative Income Hypothesis

James Duesenberry established this hypothesis during the last century. According to Duesenberry (1949), “for any given relative income distribution, the percentage of income saved by a family may tend to be a unique, invariant, and increasing function of its percentile position in the income distribution. The savings may be independent of the absolute level of income. Consequently, the collective savings ratio plays an independent role in the gross amount of income.

2.2.3 The Permanent Income Hypothesis

Milton Friedman developed and introduced the theory of permanent income. His explanation on the cross-sectional association of savings and income said that these two factors were determined by transitory discrepancy within permanent income. In the accumulated case, most transitory elements neutralized each other, causing the close correlation between income and consumption detected in time series data. (Friedman, 1957). In short, the permanent income hypothesis separated transitory income and permanent income and suggested their association as savings determinants. Permanent income was considered concerning the steady income expectation throughout a long-time period. Temporary income was presented as the deviation between permanent income and actual income. The hypothesis indicated that people could save more as long as their current income amount was more significant than the expected amount of permanent income. It was to secure themselves from income uncertainties in the future. Due to the practical effect on real-life scenarios, this theory had enormous implications concerning economic policies (Singh, 2009).

2.3 Potential determinants of GDS

The following section offers an extensive overview of the previous theoretical along with empirical studies concerning the correlation between savings and determinants that account for savings.

Theoretically, not any single element can affect GDS. However, a combination of different macroeconomic factors collectively likes GDP, interest rates, dependency ratio, financial

liberalization, and economic policies that might potentially account for domestic savings. There have been vast experimental studies on the saving behaviors and savings determinants in different nations, such as the study of Metin Ozcan et al. (2003, 1405-1416), Vincelette (2006), Imran, et al., (2010). Empirical research concerning savings has classified the vital potential factors of savings into the following categories: economic growth and income variables, demographic variables, financial variables, government policy proxy variables, and other macroeconomic variables.

2.3.1 Economic growth and income variables

The relation of savings and income or can be considered as the alliance between savings and growth. The literature on subsistence consumption proposed that a nation with higher income per capital potentially has more savings (Houérou, 2011). A previous empirical study from Metin Ozcan (2003, 1405-1416) firmly supports this suggestion. Researchers can explain this direct correlation of savings and growth via the life cycle hypothesis developed from growth models. That life cycle hypothesis expects that improved savings may encourage economic growth through the help of the more available investment (Bebczuk, 2000). This assumption is formed based on Harrod (1939), Domar (1946) and Solow (1956) growth models. Studies conducted by Paxson & Angus (1993), Oladipo (2010), and Singh (2009), Jilani et all. (2013) provide further empirical evidence that supports the life cycle hypothesis of more significant savings to improve economic growth.

Regarding income development, the life-cycle model suggests the accumulative savings tend to grow for a rise in income development. Modigliani (1986, 297-313) indicated that fast-growth nations should have more considerable GDS. Jappelli (1994, 83-109) also observed that income improvement enhances personal savings. Giving the fact of rapid increasing income makes room for saving more comfortable, and higher saving continues the virtual cycle to encourage further economic growth. Previous empirical studies from Bosworth & Chodorow-Reich (2007), and Park & Shin (2009) supported the existence of a virtual circle. Enormous growth attributed to more considerable savings and continuously supported more significant growth. They agreed that the pair of contemporary and persistent per capita GDP development enhance savings.

2.3.2 Demographic variables

The category of demographic variables includes life expectancy, age dependency ratio, the age distribution typically. Many researchers projected a descending tendency in the saving rate shortly from now because of the aging population, reduced birth rates, and longer life expectancy.

Studies from Higgins & Silberman (1998, 78-113), Bosworth & Chodorow-Reich (2007), and Park & Shin (2009) pointed out that demographic variables play a significant role in savings. Their empirical results were backed up by the life cycle model. For example, the influence of the age dependency ratio (the percentage of the elderly population over the employment-population) generated a counter effect on the savings rate. In the case of children, this group of people usually consumed and did not earn money. Consequently, the ratio of child dependency (the percentage of kids to the employment-population) might generate an unsupportive effect on savings. Furthermore, the higher child dependency ratio could lead to a downward influence on savings. Park & Shin (2009) and other experimental research, agreed that the age dependency ratio, along with the adolescence dependency ratio, reduces savings.

The age distribution also plays the central part in savings because people save more when the income drop and save less when they foresee their income improvement. As a result, people tended to keep less at a younger age. They could collect more during their productive years and save less when they are aged and retired (Modigliani, 1970, 197-225).

Life expectancy posed an enhancement on the savings because a more extended life's duration prolonged retirement durations and required an enormous amount of savings in their retirement period. Besides, the employment participation rate of the elder could hurt savings because growth in the employment participation of the elder decreased the retirement period, and cut down the amount of savings needed for retirement (Horioka and Hagiwara, 2010).

2.3.3 Financial variables

The financial variables that potentially affect saving are factors that indicate the development level of the monetary market. Interest rates, one of the economic variables, is known as a critical element that accounts for saving rates. However, the association tendency is quite unclear, according to experimental studies. Khan & Sarker (2016) noticed that interest rate has a significant beneficial influence on savings across developed countries and a harmful yet insignificant impact in emerging nations. Metin Ozcan et al. (2003, 1405-1416), Vincelette (2006), all reported the significant, affirmative influence of interest rate on household savings. However, empirical studies from Hammad et al. (2010, 23-34), Khan et al. (2017) showed an insignificant relation of interest rate and household savings, suggesting the ambiguous connection between interests and savings. Basely et al. (1998), Jilani et all. (2013) revealed savings rates had an optimistic link or relationship with interest rates.

Another financial variable that might account for savings is economic market development indicators, demonstrated by the monetization level of the economy. Empirical studies from Metin Ozcan (2003, 1405-1416) and Khan et al. (2017) pointed out that money supply M2 had a considerable beneficial influence on savings in the long-term period.

2.3.4 Government policy proxy variables

Different government actions might create an impact on saving. Among these, the influence of government spending on public services and pensions has been considered as a factor that increases precautionary savings because if government spending is insufficient, people may be worried about their living in the future. Empirical research based on the lifecycle model from Horioka and Hagiwara (2010) showed that when the social security benefits are distributed generously, personal savings could have a decreasing tendency, fundamentally because of the weak motive for retirement planning and self-insurance savings. Hubbard & Zeldes (1995, 360-399) reported a considerable adverse influence of pensions structure on private savings. Public expenditure on social security benefits generated a substantial impact on savings, especially as a result of

experimental studies in emerging countries. For example, empirical works from Imran et al. (2010) reported government expenditure likely enhances the rates of national savings.

Another factor that can affect domestic saving is public investment. Theoretical research suggested different views on the tie between public investment and savings (Imran et al., 2010). The life-cycle model explains that a lower level of public investment may promote consumption and reduce saving because of the reallocation of tax amounts from current to future populations. As a result, the life-cycle model suggests that a drop in public investment may generate a reduction in domestic savings. The quality of public finance is essential here. Empirical results from Houérou (2011) revealed that a higher level of public investment under the condition of productive spending could help to increase national savings.

2.3.5 Other macroeconomic variables

Other macroeconomic variables that can account for domestic savings are trade measurement and the current account imbalance, household saving rate, and unemployment rate. Experimental research conducted in conjunction with an open economy model from Chaudhry et al. (2015), Khan & Sarker (2016) indicated that trading plays a role in national saving. The results clarified that the trade surplus encouraged saving due to the beneficial impact on the growth of wealth and income.

The general view on the association between savings and the current account balance is an unsupportive tie. Loayza et al. (2000, 165-181) reported that development in the current account shortfall could lead to a deterioration in private saving because external saving could play the role of an alternate element to domestic private saving.

The household savings ratio is perhaps the most widely discussed savings indicator. Household savings, a component of private savings, create a holistic view of total saving rates. It implies a direct and considerable correlation with GDS. Empirical analysis from Komicha (2007) and Kulikov et al. (2007) provided evidence that household saving created a negative influence on overall savings. It did not mean that encourage household

savings could reduce total national savings. This result merely summarized the relationships between the data over the research period.

According to Houérou (2011), the influence of the unemployment rate on savings might be unclear. In case of a higher unemployment rate, more people received less income than average. Concerning the normal consumption behavior, people could expect a drop in personal savings as a result. Cross-country studies provided evidence on the beneficial relationship between the employment rate and saving rates. According to life cycle theory, a high unemployment rate also involved sharp uncertainty in the future, which might encourage households to increase their precautionary savings, at least in the short term. The rapid downward trend in employment conditions, along with a higher unemployment rate, combined with noticeable falls in house prices, might push households to adjust saving levels significantly. Relevant published literature on domestic savings determinants used in this thesis is summarized as follows:

Table 1 Literature summary on savings determinants

Study	**Data**		**Model & statistical techniques**	**Saving determinants**
	Countries	**Period**		
Metin Ozcan et al., 2003	Turkey	1968-1994	OLS estimation	• income variable • public savings • the ratio of M2 to gross national product (GNP) • current account deficit and the terms of trade • inflation rate • the real interest rate on saving deposits • demographic factors such as youth dependency ratio, urbanization ratio, life expectancy ratio, and old dependency ratio
Narayan et al., 2006	Fiji	1968-2000	ARDL method	• interest rate • the deficit of current account, • dependency ratio of age
Vincelette, 2006	Pakistan	1973-2005	OLS regression	• financial development • rate of interest • financial policy • demographic factors
Singh, 2009	India	Annual data from 1950–1951 to 2001–2002	• Standard OLSEG estimates: the OLS-based two-step cointegration estimator of Engle	• GDS

			and Granger (1987) (OLSEG) • Optimal DOLS, FMOLS and NLLS estimates • The estimation of conditional error-correction model (ECM) based on the autoregressive distributed lag (ARDL) model • ML system estimates: Vector autoregressive (VAR) • Short-run dynamics: The error-correction model (ECM) &Monte Carlo simulations	
Hammad et al., 2010	Malay	1978 to 2007	Error correction model (ECM)	• government fiscal balance • per capita income • inflation • rate of return on savings deposit • age dependency ratio
Imran et al., 2010	Pakistan	Annual data from 1972 to 2008	Error Correction Model	• consumer price inflation • public loans, • interest rates, • government consumption • remittances
Chaudhry et al., 2015	Pakistan	1972-2008	• Johansson Cointegration technique • vector error correction model (VECM)	• government spending • workers remittance • consumer price index • public loans • interest rate • exports
Horioka and Hagiwara, 2010	12 economies in developing Asia	during 1966–2007	• The random-effects model with established standard errors • Fixed effects model	• age • age dependency ratio • per capita real GDP • credit • per capita income growth • real GDP • the inflation rates • the real interest rate • government expenditure on social services and pensions
Botha et al., 2011	South Africa	1981Q1 to 2009Q4	Vector error-correction model (VECM)	• short-term interest rates
Houérou, 2011	Turkey	1975-2008	• Time series analysis • OLS regression	• interest rate • gross private disposable income • the young-age dependency ratio • inflation rate • public investment expenditure • employment rate
Mishi, 2012	South Africa	from the year 1963 and 2011	VECM method	• public saving (government debt to GDP) • the real disposable income (GDP per capita) • interest rate

				• the percentage of household saving to household disposable income • the portion of M2 to GDP
Jilani et al., 2013	Pakistan	Annual data 1973-2011	• Augmented Dickey-Fuller (ADF) test • Johansen Co-integration test • Regression estimation • Error correction model (ECM)	• inflation • GDP • interest rate • fiscal deficit • the ratio of Age dependency
Khan & Sarker, 2016	Bangladesh	1983 to 2013	Vector error correction model	• inflation rate • gross domestic income • exports • deposit interest rate
Khan et al., 2017	Pakistan, China, Singapore, Japan Turkey, and Russia	1995-2016	• Fixed effects model • Random Effect Model • Regression Model	• gross domestic product • foreign direct investment (FDI) • money supply (M2) • inflation • per capita income • the ratio of age dependency

2.4 Summary

A sound understanding of savings determinants is fundamental in terms of building up active policy intrusions. Previous research that studied factors that account for domestic savings in different countries of the world showed five essential groups of determinants. They included economic growth and income, demographic, finance, government policy proxy, and other macroeconomic determinants (Metin Ozcan et al.,2003). Overall, growth and financial variables such as gross domestic product, interest rates have considerable influence while other macroeconomic variables, including money supply, per capita income, employment rate, generate insignificant enhancement on total domestic savings. There are many approaches to assess the association between GDS and its determinants, such as using the ARDL model or vector error correction model.

3 MACHINE LEARNING

3.1 Machine learning Introduction

This part may introduce briefly about machine learning, which is known as a crossover between computer science, statistics, and disciplines. It is based on the mathematical model but can explain the underlying structure of data which cannot be captured by the econometric model. Moreover, a complex optimization problem is possibly solved by a machine learning algorithm, so it is considered as the most common form of artificial intelligence (Wang et al., 2014).

Time-series regression models are based on a lot of strict assumptions that make them inefficiently in practice, especially for predicting vital macroeconomic indicators. Machine learning with the advantage of computing power has become an alternative for the traditional model because it can work with a massive dataset (the amount of observations is limited and less than the number of potential regressors) (Aaron, 2018). However, due to the priority to understand the association of gross savings determinants, most studies about savings applied to the traditional regression model in the methodology. As a result, little research has been conducted on total savings prediction using machine learning methods.

3.1.1 Learning paradigms

Machine learning is increasingly well-known and applied in many different industries. Machine learning consists of three different categories, which are unsupervised learning, supervised learning, and reinforcement learning.

Unsupervised learning

Unsupervised learning is frequently explained as “learning without a teacher.” In this category of machine learning, the critical issue that needs to be solved is to discover undetected data structures or data patterns. As there are no y values to predict, the

objective is to identify the consistency in a set of x variables (Julien & Etienne, 2018, 475). The learner does not have the correct answer to start the learning process. Specifically, the learner just has access to the input, while the output is unidentified and must be reconstructed. To sort out the uncertainty of data output, the learner has to explore the uncover patterns, associations, or groups to categorize the vast pool of data without knowledge on the precise answers in advance. The objective is commonly completed via dimensionality reduction, cluster analysis, and association standard of learning (Hastie et al., 2009).

Reinforcement learning

Reinforcement learning is the category of machine learning concerning how software learner decides to maximize the accumulative outcome. First of all, the reinforcing learner does not have any outcome expectation to assess the result. Throughout the learning stages, a modest assessment result, which is known as a reinforcement signal, is provided to the software. The reinforcement signal is an estimation of the correct decision or performance measurement. The feedback is provided after the software gives the decision to measure how well the software performs. Following by subsequent training stage, the software can include this signal to develop from it and provide a more accurate answer. After each step, the software collects the correct behavior benefits, so enable it to give a more precise result. Samuel developed the basis machine learning algorithm in 1953. It was a computing software created to perform checkers. After each stage, the software develops its knowledge to decide if the solution leads to a good or bad outcome. Eventually, the software can make a decision among right and wrong answers, develop the playing ability, and wins its maker, a famous checker (Bishop, 2006).

Supervised learning

Supervised learning's issues are very prevailing in machine learning studies. In supervised learning, the correct outcomes are provided primarily to the software to develop its knowledge from that. If we consider software as a student, materials with the proper results may be provided as a mentor to the student. For example, given a provided database, the entire observations may be categorized using a label that associates with

the category of the representation. The software has to find the line, which is also known as the decision boundary. The line best splits the two classes and uses the splits for predicting outcomes of new samples. A useful application of supervised learning can be value prediction. The objective is to forecast the values of an outcome variable y according to the values of a set of predictor variables x. The algorithm has to figure out adequate estimations to the unidentified association of the predictor variables and the outcome variable to get a precise prediction result. (Julien & Etienne, 2018, 475).

3.1.2 Regression

The research goal in this thesis is to forecast the GDS rate of Finland. This problem is called a regression problem, where the models are used to predict a continuous value. As later described in chapter 4, the amount of domestic savings in the future may be estimated based on analyzing the relationship between a broad set of features (GDS determinants) and the real-valued response (GDS). In this thesis, the empirical test aims to sort out a supervised learning regression problem, not the classification problem, where the objective is a discrete value.

3.2 Learning algorithms

3.2.1 Regularized Linear Regression

In general, the association between a dependent variable Y and many independent variables X_1, X_2,……, X_p can be described by the simple linear regression below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (3.1)$$

where β_0, β_1, β_2,…, β_p are the regression coefficients, and ε represents the random noise or error term.

The regression coefficients are commonly estimated by least-squares estimation, which requires the optimization (minimization) of the loss function, which is called the Residual Sum of Squares (RSS) (Ng, 2015):

$$RSS = \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} \beta_j X_{ij})^2 \quad (3.2)$$

The simple linear regression faces some problems, such as over-fitting and non-generalization on future data (Ng, 2015). The regularization technique is proposed as a well-known method to deal with them because it can shrink the coefficient estimates towards zero (Ng, 2015). Hence, the regularized linear regression model is defined as a particular category of the linear regression model, and it has three types: Lasso, Ridge, and Elastic Net.

Ridge Regression

In 1988, Hoerl and Kennard introduced the Ridge regression model to resolve the challenge of the standard least square estimator. Ridge regression model was used wisely in research to sort out the problem of very high intercorrelations or inter-association among the independent variables, which might cause more variation in future results. Since the independent variables can be highly correlated to each other, the individual regression coefficient might be impacted by either the different independent variables which are incorporated into the model or the ones that are left out. In these cases, a small modification in the model likely created a difference in the coefficient estimations and eventually generated the impracticable result of the model.

In the Ridge model, the RSS was formulated by changing the RSS in the simple linear model by accumulating an enhancement constant (penalty) to decrease the variance of the model and make the regression coefficients smaller. The loss function RSS was defined as the following:

$$RSS_{ridge}(\beta) = \sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2 = RSS + \lambda\sum_{j=1}^{p}{\beta_j}^2 \quad (3.3)$$

In the above equation (3.3), λ expresses the complexity parameter used to control the amount of shrinkage. If the value of λ is more prominent, as a result, the amount of reduction which is imposed on the regression parameters may higher. In the equation, $\lambda\sum_{j=1}^{p}{\beta_j}^2$ defines the L2 penalty. It is useful to maintain the coefficients small. Generally, the Ridge regression tends to choose the linear model with the most insignificant coefficients to value overall sums of squared weights. The insignificant coefficients can help the model to generalize the data better in solving the over-fitting problem.

The parameter λ may be required to be selected in advance, and cross-validation is the most useful method because cross-validation illustrates the regularization strength. In the case of the too-large value, the coefficients can be reduced to 0 and therefore create an under-fitting situation. Figure 1 provides more detail about the way parameter λ disturbs the value of coefficient estimations, in the case of the regression model with two coefficients. The graphic presentation of equation (3.3) is offered in the below:

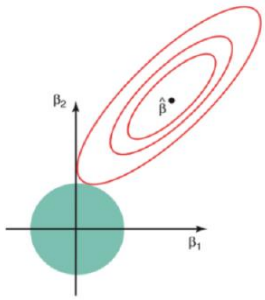


Figure 1. The Ridge regression estimation graph of RRS.
The green zone specifies the regularization term area while the red ellipse zone interprets the contour of the least-squares error function. (Hastie et al., 2009)

According to previous researches, Ridge regression can attain more stable results and enhance predictive performance compared to simple linear regression. Ridge regression has some advantages such as avoid the significant value of λ, decrease their variation, gain better predictive performance, and execute a robust approach to minimize unwanted problems as it penalizes the value of λ. Ridge regression achieves L2 regularization, in which the penalty occurs on the squared value and the magnitude of the coefficients. As a result, the coefficients estimated by Ridge regression may not be reduced to precisely zero. However, the disadvantages of Ridge regression appear when it cannot resolve the following problems: the interpretable option of the coefficients, the meanness situation of the model (simple model but provide reliable explanatory predictive power), or feature selection option. In Ridge regression, the coefficient values are controlled, but they don't equal to zero; therefore, the least-squares error function (RSS) norm doesn't reassure sparsity or feature selection (Hoerl & Kennard, 1988).

Lasso regression

The Lasso regression model represented by Robert Tibshirani in 1996 is similar to the traditional ordinary least squares (OLS) but incorporates diverse kinds of reduction to generate parsimonious models in the situation of a large feature number. It required to include a penalty equalling to the absolute value of the magnitude of the coefficients to achieve L1 regularization in the model and eventually reducing some of them to zero (Tibshirani, 1996). It is a popular method for feature selection and model regularization. Moreover, it is better to interpret the model to compare to the Ridge model and the simple linear model. Another advantage of Lasso regression is to produce better performance than the Ridge model (Trevor et al., 2009). The Lasso model has the loss function (according to OLS regression) which is described as below:

$$RSS_{lasso}(\beta) = \sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (3.4)$$

The regularization term in RSS is the total of different absolute values of the model parameters. It is necessary to identify the parameter λ, which controls the strength of the

penalty in advance. The constraint zone in the Lasso regression model is graphically demonstrated as a diamond shape (Figure 2), which is different from a disk shape in the case of Ridge regression.

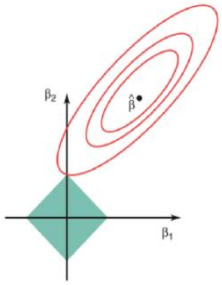


Figure 2 The Lasso regression estimation graph of RSS. The green zone represents the constraint region, while the red ellipse area represents the least-squares error function. (Hastie et al., 2009)

When an ellipse area encounters the constraint zone at the first point, the Lasso coefficients are determined. The corners of the constraint zone are at each of the axes, and so the ellipse area may meet the constraint zone is more than two. That's why the coefficients of the Lasso model can be equal to zero that makes the model become a standard approach for variable selection and easier for understanding (Zou & Hastie, 2005).

The elastic net regression

The elastic net regression, established by Zou and Hastie (2005), is a combination of the Ridge and Lasso regressions. It can shrink the coefficients of the features and eliminate them (a coefficient of zero) at the same time. The penalty term, in this case, is defined as a convex sum of the ridge (L2 term) and lasso penalties (L1 term). The elastic net method combines both continuous shrinkage and variable selection, and the loss function is expressed as follows:

$$RSS_{elastic}(\beta) = \sum_{i=1}^{n}(Y_i - \sum_{j=0}^{p}\beta_j X_{ij})^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}{\beta_j}^2 \quad (3.5)$$

3.2.2 Support Vector Regression

Support Vector Machine

Cortes and Vapnik introduced Support Vector Machines (SVMs) in 1995 with the usage in the binary classification of data points. There are a variety of applications of SVMs in different fields, but few works have reviews of SVMs to predict macroeconomic indicators. SVMs have two alternative types: linear versus non-linear (Awad & Khanna, 2015).

Linear Case

The fundamental idea of linear SVMs is to identify an optimal function (classification rule) that assigns given data points ($\mathbf{x}_l$) with class labels (y_l) with the least possible amount of error or with the most significant reasonable margin. The training set is linearly separable and defined as:

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_L, y_L)$$
$$x_l = (x_{1,l}, \dots, x_{N,l})$$
$$y_l \in \{-1,1\}$$

The classification rule, also known as hyperplane, is formulated as:

$$f(x) = sign\left(\sum_{i=1}^{N} w_i x_i + b\right)$$
$$sign(\alpha) = \begin{cases} 1 \;\; if \;\; \alpha > 0 \\ 0 \;\; if \;\; \alpha = 0 \\ -1 \; if \; \alpha < 0 \end{cases} \tag{3.6}$$

where the constant$b$and the weights vector $\mathbf{w} = (\mathrm{w}_\mathbf{1}, \dots, \mathrm{w}_N)$ are estimated by a learning algorithm.

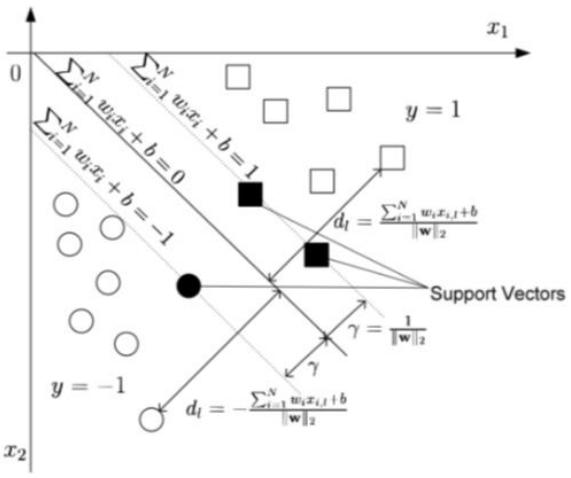


Figure 3 Optimal Separating Hyperplane and Support Vectors (Cao et al., 2013, 283)

The classification rule (hyperplane) aims to separate data points into two classes {-1, 1}. However, there may exist more than one hyperplane going through given data points. Therefore, the hyperplane which needs to be found to separate the data optimally is one creating two clusters with the most substantial distance (maximal margin γ). The optimal separating hyperplane can be computed to solve the optimization problem (Trustorff et al., 2011, 565-581).

$$\begin{aligned} \underset{\mathbf{w}, b}{\text{Minimize}} \quad & \gamma \\ \text{subject to:} \quad & y_l\left(\sum_{i=1}^{N} w_i x_{i,l} + b\right) \geq \gamma \\ & \|\boldsymbol{w}\|_2 = 1 \\ & l = 1, \ldots, L \end{aligned} \tag{3.7}$$

where γ is the margin, and 2-norm is defined as $\|\boldsymbol{w}\|_2 = \sqrt{\sum_{i=1}^{N} w_i^2}$

The optimization problem (3.7) is a quadratic programming problem and can be solved by fixing the equation $\gamma\|\boldsymbol{w}\|_2 = 1$ instead of $\|\boldsymbol{w}\|_2$. The maximization of margin γ is equal to the minimization of $\|\boldsymbol{w}\|_2$. Therefore, the equation (3.7) can be rewritten as:

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{Minimize}} && \tfrac{1}{2}\|\boldsymbol{w}\|_2^2 \\ &\text{subject to:} && y_l\left(\textstyle\sum_{i=1}^{N} w_i x_{i,l} + b\right) \geq 1 \\ & && l = 1, \ldots, L \end{aligned} \tag{3.8}$$

Nonlinear Case

The most advantage of linear SVMs is quick and straightforward training. In contrast, this model can not work effectively with too many features or complex datasets. There are cases where the optimal separating hyperplane is not possible to find as the data is not linearly separable enough. Nonlinear SVMs can solve this problem because of its consistency and explanatory power. This approach introduces errors ξ_l into the optimization problem (3.8), which leads to the following equation.

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{Minimize}} && \tfrac{1}{2}\|\boldsymbol{w}\|_2^2 + C\textstyle\sum_{l=1}^{L}\xi_l \\ &\text{subject to:} && y_l\left(\textstyle\sum_{i=1}^{N} w_i x_{i,l} + b\right) \geq 1 - \xi_l \\ & && \xi_l \geq 0 \\ & && l = 1, \ldots, L \end{aligned} \tag{3.9}$$

Where C is called the "cost" parameter to scale the error, the value of C indicates how close the model can be fitted to the training set. A low value of C results to small effect from errors on the solution while a high value of C becomes closer to the initial problem, but overfitting maybe occur. If C = 0 the weighs vector may be zero vector $\mathbf{w}'$ = (0, ... ,0). When we can get the high enough value of C, two clusters are linearly separate, and then equation (3.8) and (3.9) may be the same.

To solve optimization (3.10), Vapnik and Chapelle (2000, 2013–2036) used the Lagrange function with Lagrange multipliers $\alpha_l, \beta_l \geq 0$.

$$L = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{l=1}^{L}\xi_i - \sum_{l=1}^{L}\alpha_l[y_l(\sum_{i=1}^{N} w_i\, x_i + b) - 1 + \xi_l] - \sum_{l=1}^{L}\beta_l\xi_l \tag{3.10}$$

The saddle points are determined by setting the derivative of w, b and ξ_l to zero

$$\frac{\delta L}{\delta w_i} = w_i - \sum_{l=1}^{L} \alpha_l y_l x_{i.l} = 0 \Leftrightarrow w_i = \sum_{l=1}^{L} \alpha_l y_l x_{i.l}$$

$$\frac{\delta L}{\delta b} = \sum_{l=1}^{L} \alpha_l y_l = 0 \tag{3.11}$$

$$\frac{\delta L}{\delta \xi} = C - \alpha_l - \beta_l = 0 \Leftrightarrow C = \alpha_l + \beta_l \tag{3.12}$$

When to substitute (3.10), (3.11) and (3.12) into (3.13) the Lagrange function may be formulated as:

$$\tilde{L} = \sum_{l=1}^{L} \alpha_l - \frac{1}{2} \sum_{l=1}^{L} \sum_{k=1}^{L} \alpha_l \, \alpha_k y_l y_k \sum_{i=1}^{N} x_{i,l} \, x_{i,k} \tag{3.13}$$

which is maximized on the condition that the constraints

$$0 \leq \alpha_l \leq C$$

$$\sum_{l=1}^{L} \alpha_l y_l = 0$$

Kernels

The nonlinear SVMs may work effectively on non-linearly separable data, but it can underperform in the case of a significantly non-separable dataset with complex non-linearity. Therefore, the data is likely to transform from input space into feature space via the function $\phi(\mathbf{x})$. The feature space is a place where the non-separable data can be separated linearly (Cherkassky and Mulier, 2007). The feature function $\phi(\mathbf{x})$ assigns all data points x_l in the input space to the feature space:

$$(x_l . x_k) = \sum_{i=1}^{N} x_{i,l} x_{i,k} \Rightarrow \big(\phi(x_l) . \phi(x_k)\big) = \sum_{i=1}^{N} \phi\big(x_{i,l}\big) . \phi\left(x_{i,k}\right) = K(x_l . x_k) \tag{3.14}$$

where $K(x_l . x_k)$ is called the kernel function. The transformation of data from input into the feature space can be seen in Figure 4, which shows a non-linear separating hyperplane in input space is converted to a clear separating in the feature space.

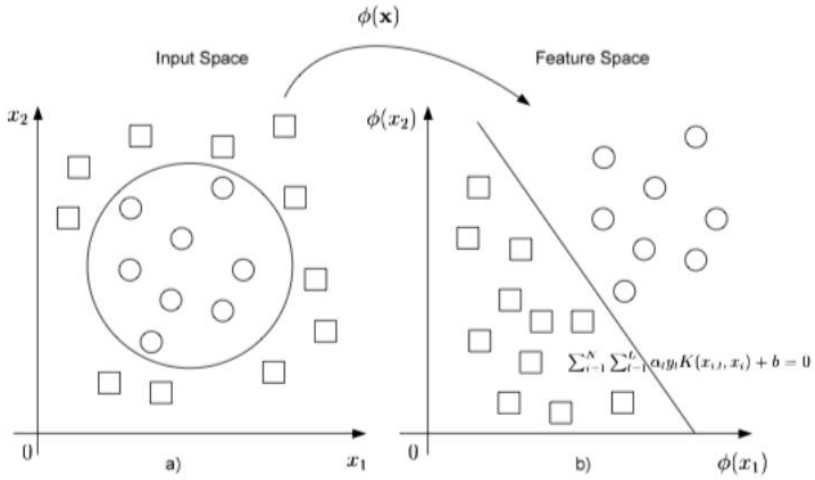


Figure 4 Transformation data from input space into feature space (Cherkassky & Mulier, 2007)

Because of adding the kernel function, the optimal separating hyperplane in the feature space may be determined by rewriting the optimization problem (3.13) as follows:

$$\underset{\alpha}{\text{Minimize}} \quad \tilde{L} = \sum_{l=1}^{L} \alpha_l - \frac{1}{2}\sum_{l=1}^{L}\sum_{k=1}^{L} \alpha_l \, \alpha_k y_l y_k K(x_l . x_k) \tag{3.15}$$

On condition that:

$$0 \le \alpha_l \le C$$

$$\sum_{l=1}^{L} \alpha_l y_l = 0$$

$$i = 1, \dots, N \qquad l = 1, \dots, L$$

The Kernel function has got two vectors as inputs and creates a single scalar value. Due to an inner product, the kernel function should follow Mercer's theorem, which produces a variety of alternatives to kernels. For instance:

Gaussian kernel (Radial basis function)

$$K(x_l . x_k) = e^{[-\gamma \|x_l - k_l\|^2]} \tag{3.16}$$

Hyperbolic tangent

$$K(x_l . x_k) = tanh(k_1(x_l . x_k) + k_2) \quad for\ k_1 > 0 > k_2 \tag{3.17}$$

Polynomial kernel

$$K(x_l . x_k) = [(x_l . x_k) + 1]^d \tag{3.18}$$

In practice, the Gaussian kernel function is widely used since it performed better than polynomial and Hyperbolic tangent kernel function (Friedman et al. 2001). Many empirical studies concerning the application of machine learning in financial data consider the Gaussian kernel function as the basis of SVM – based model, why this kernel is also applied for extending to the regression model in this thesis.

Support Vector Regression

Vapnik (2000) implied that Support Vector Regression (SVM) is a particular case of SVMs. SVM goals to find the function $f(x)$ expressing the relationship between input data points (vectors) $x_l$with known outputs $y_l$by estimating the weights vector w and bias b.

$$y_l = f(x) = \sum_{i=1}^{N} w_i x_{i,l} + b + u_l \quad l = 1, \ldots, L \tag{3.19}$$

u_l is the error term, which indicates the discrepancies between the outputs and the functional approximation.

Support vector regression is also known as linear epsilon – insensitive SVM regression (ε-SVM regression) because it is formed on ε -insensitive loss function. There are three different types of loss functions (Figure below)

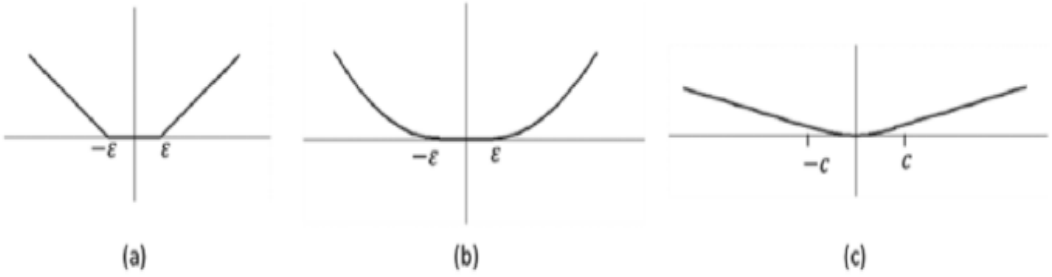


Figure 5 Loss function: (a) linear, (b) quadratic, and (c) Huber (Vapnik, 2000)

Therefore, the derivations in this section may follow the linear loss function of the following:

$$|u_l|_\varepsilon = max\ (0, |u_l| - \varepsilon) \tag{3.20}$$

It is easy to understand that the loss equals zero if deviations are between the parameter – ε and ε. For those deviations with the loss more significant than zero, error slopes ξ_l,ξ_l^* ≥ 0 are introduced to make loss function asymmetrical. These error slopes indicating how many deviations outside ε-tube are formulated as follows:

$$\xi_l = \begin{cases} u_l - \varepsilon & if\ u_l - \varepsilon \geq 0 \\ 0 & otherwise \end{cases}$$

$$\xi_l{}^* = \begin{cases} -(u_l - \varepsilon) & if\ u_l - \varepsilon \leq 0 \\ 0 & otherwise \end{cases}$$

Based on SVMs equation above and equation (3.20), the optimization problem for support vector regression can be written as:

$$\underset{\mathbf{w},b}{\text{Minimize}} \quad \frac{1}{2}||w||_2^2 + C\sum_{l=1}^{L}\xi_l$$

$$\text{subject to:} \quad y_l - \sum_{i=1}^{N} w_i x_{i,l} - b \leq \varepsilon + \xi_l$$

$$\sum_{i=1}^{N} w_i x_{i,l} + b - y_l \leq \varepsilon + \xi_l^* \qquad (3.21)$$

$$\xi_l, \xi_l^* \geq 0$$

$$l = 1, \ldots, L$$

Where the insensitive region is controlled by the trade-off parameter C and parameter ε, both of the parameters need to be pre-determined, and some algorithms are run until the best set is found.

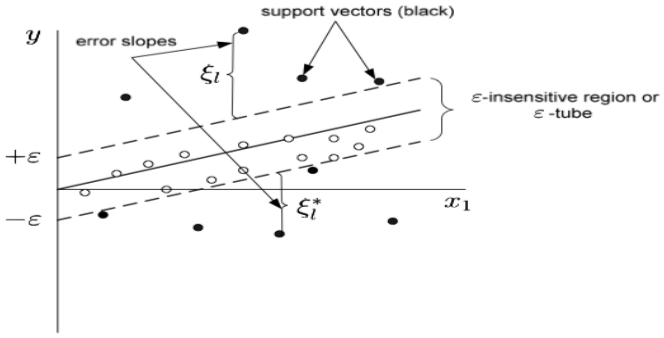


Figure 6 Support Vector Regression (Vapnik, 2000)

To find out the best-fitted function $f(x)$, it is necessary to minimize the ε -insensitive loss function subject to its constraints. The size of the insensitive region has a considerable influence on how well the model may perform. According to Vapnik (2000), the approximation error can be improved by decreasing the region's size and then leads to overfitting. Otherwise, the larger size possibly results in under-fitting. It explains the importance of selecting these parameters. They are formed based on past performance on the training set, which means that running the system today gives the history of different pairs C and ε using the best set to train the model in the future.

3.2.3 Decision tree

The decision tree is under the classification of supervised machine learning, and it is a nonlinear model. The decision tree belongs to a particular category of algorithms used for predictive modeling machine learning. The algorithm performs throughout, so the data is divided continuously set up on a specific predetermined parameter. There are two behaviors to explain the decision tree performance, categorized by the leaves or decision nodes. The leaves characterize the decision made in that particular step or the concluding outcome of the tree. The decision nodes signify the point where the data is divided. In the visual illustration form, decision trees appear like an upside-down tree, with the root at the top (Mu-Yen, 2011). Decision tree algorithms are divided into two types of trees, including classification trees and regression trees.

These two types of decision tree algorithms are categorized together under the name of CART, which stands for Classification and Regression Trees. In a classification tree, the result or decision variable is always definite, for example, Yes or No. In a regression tree, the outcome variable is continuous, for example, a number 156 (Song & Lu, 2015, 130).

Regression trees are constructed through binary recursive partitioning. In a regression tree, the output variable is numerical, and input variables might be a combination of continuous and categorical variables. The tree is shaped when each decision node comprises a test on some input variables value. The terminal nodes (where the tree ends) signify the predictive output variable values. The binary recursive partitioning process begins with the training set. After that, the algorithm starts by assigning data to the two first branches, using binary splits in every field possible. This process is completed by minimizing the total of the squared deviations from the mean in the two distinct branches. Subsequently, in every branch, data is splitting eventually until each node passes the detailed minimum node size and forms the terminal node. Pruning of the tree is finished by using the validation set, which is also called the test set.

Below is a demonstration of a regression tree that has been visualized for more natural understanding. A regression tree is used to predict highway miles per gallon of cars; pre-deciding variables are weight, horsepower, wheelbase, and type of car. In the terminal nodes, the miles per gallon set up on the kind of cars.

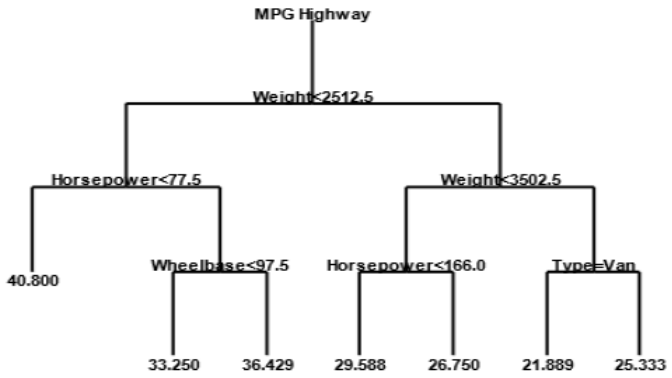


Figure 7 Demonstration of a regression tree (Song & Lu, 2015, 130)

3.3 Summary

In recent years, machine learning is gaining attention from both professional and individual person. It has arisen a question of whether algorithms can be used to employ a big-scale data and then forecast the future outcomes by finding the knowledge from it. When the machine learning model handles diverse and complicated dataset, the model can outperform the traditional models from different perspectives by optimizing model complexity, and that can help predict a variable more adequately.

The GDS rate is a challenging variable to predict because a wide variety of other variables can influence it. Literature has strong support for using a linear regression-based model to analyze the empirical determinants of GDS, but not to predict the value of total domestic savings. The lacking of research on the application of machine learning to this area provides room for further exploration of this thesis. This thesis selected the most popular regression models to predict the GDS, including Linear Regression, Regularized Linear Regression, Support Vector Regression, and Decision tree.

4 METHODOLOGY

This chapter has the objective of explaining and discussing the usage of various empirical methods in this thesis. There are three sections in chapter methodology. The first part may provide information related to the data source, selection of the variables. The second part implies how to conduct the prediction models of gross domestic saving. Furthermore, the second section describes different models based on machine learning algorithms. The last part may discuss some evaluation measures to assess the performance of prediction models.

4.1 Data & variables

This section provides a descriptive summary of the data and variables used in the empirical experiment. Quarterly time series data on Finnish macroeconomic indicators covering the 1995 Q1 – 2019 Q4 period have been chosen in this thesis. The study data set was collected through the following sources:

1. OECD Statistics
2. Statistics Finland
3. Bank of Finland

Data universe

The literature review on the gross saving rate in section 2.3 implies that studies often use the annual data to analyze this indicator because the statistic agencies tend to report macroeconomic indicators every year. The availability of data concerning some economic variables in Finland is only available since 1995. The database is also required to be big enough to produce a reliable estimation. Therefore, the empirical experiment selects quarterly data and the period of 1995 – 2019 to evaluate the predictive model performance.

The data set consists of 11 variables coming with 100 observations and no missing data. The dependent variable is Gross Domestic Saving expressed in percentage of Gross

Domestic Products (GDP). This thesis aims to predict GDS, so the models need to contain independent variables that affect GDS. The Life Cycle Model's theoretical framework and previous researches on the analysis of GDS determinants (Appendix 1) guide the choice of predictors in this thesis. There are ten independent variables, including economic growth, government policy, and financial variables, as well as external factors.

The summary description of all variables (11 variables) used in the empirical test is as below:

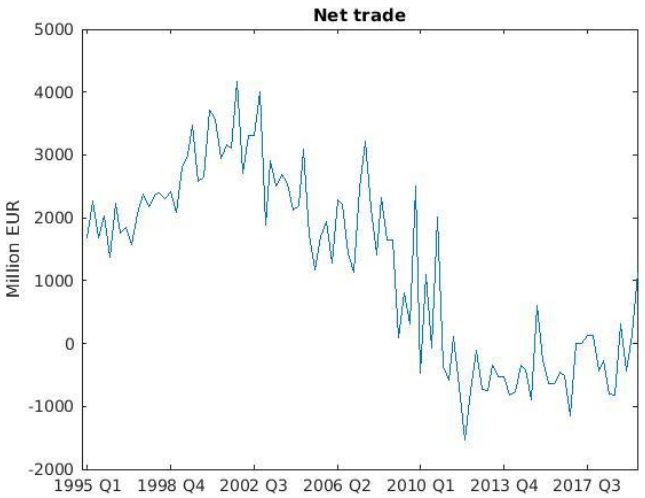
- Gross Domestic Saving $(Y;\ \%GDP)$ refers to the total of private savings and public savings in the form of liquid assets in a country. It is the difference between GDP and total consumption.
- Current Account $(X_1;\ \%GDP)$ reflects the flows of goods and services, primary income and secondary income between non-residents and residents. It is the value of gross national savings subtract gross fixed capital formation. Studies from Metin Ozcan et al. (2003) and Narayan et al., 2006 used it as a predictor of GDS.
- GDP per capita $(X_2;\ EUR)$ refers to the value of all finished domestic commodities or products and services produced in a region or nation at a particular period (i.e., one year) per capita. Metin Ozcan et al. (2003), Horioka and Hagiwara (2010), Mishi (2012), Khan et al. (2017) have used this determinant in their GDS forecast in the Asian region.
- Government spending $(X_3\ ;\ million\ EUR)$ is the amount of money that the public sector spends on the provision of services and acquisition of goods such as healthcare, defense, education, and social protection. Since savings is disposable income minus spending, government spending is a component of GDS. Imran et al. (2010) selected this indicator in the investigation of GDS determinants in Pakistan.
- Net trade $(X_4;\ million\ EUR)$ represents the discrepancy between the sum of exported goods and services and the sum of imported ones. Chaudhry et al. (2010) choose the terms of trade in their empirical study about GDS.

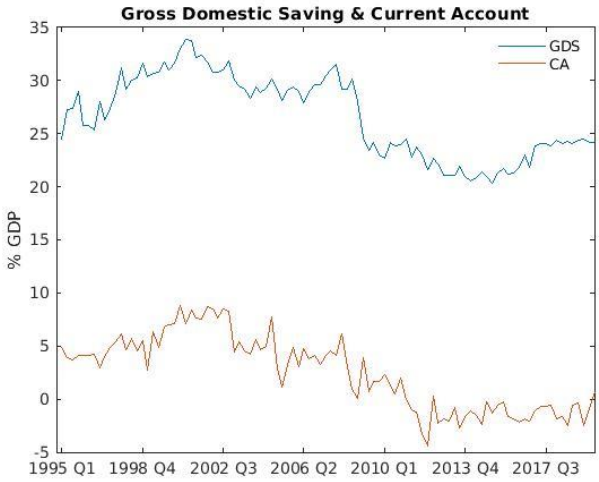
- Unemployment rate (X_5; %) refers to the unemployed population (labor force) divided by the active group of the same age. Houérou (2011) studied the influence of employment rate on household savings behaviors in Turkey.
- Public investment (X_6; $million\ EUR$) is related to government spending on public infrastructure, including into two types: i) economic infrastructures such as airports, roads, railways, ports, water and sewage, power, gas, and telecommunication; and ii) social infrastructures such as schools and hospitals. Houérou (2011) discussed the impact of efficient public investment on the public savings sector and economic growth in Turkey.
- Household saving rate (X_7; %) equals to household net disposable income subtract the household consumption spending or savings divided by net disposable income. Mishi (2012) used this indicator to predict household savings rate in South Africa.
- Short-term interest rates (X_8; %) is the daily rate on average. It is generally associated with the interest rate that short-term government papers are issued or traded in the market or the borrowing rate between financial institutions in the short-term. In some cases, three-month money market rates can be considered as short-term interest rate if it is available. Hence, it is likely called "money market rate" and "treasury bill rate." Botha et al. (2011) from South Africa evaluate GDS growth by using short term interest rates.
- Long term interest rates (X_9; %) is related to 10-year government bonds. There are numerous critical drivers of long-term interest rates, including the decline in the capital value, the change in price by lenders, and the risk from borrowers. Hammad et al. (2010) selected long term interest rates as a determinant of savings in Malaysia.
- Money supply M2 (X_{10}; $million\ EUR$) has two main components. The first one refers to M1, which is the sum of current account deposits and cash outside of the private banking system. The second one is the capital in retail mutual funds and money accounts, saving accounts, and time deposits of under $100,000. M2 is mostly used as a classification for money supply in America and the eurozone. Khan et al. (2017) assessed the role of money supply M2 to promote GDS in emerging countries.

4.2 Data Description

The past trends of selected variables are illustrated in the below graphs. The value of GDP per capita, government spending, and money supply (M2) rose consistency from 1995 through 2019. Public investment also increased over that period, but they grew steadily in a stable manner. GDS (% GDP) reached a peak in 2001 before falling gradually from 2007 to 2010 and starting to recover the increasing speed by 2014. By contrast, short-term interest rates and long-term interest rates steadily decreased over the period, and both figures had a similar pattern of change.

Net trade showed fluctuation across the period with a decreasing trend. Similarly, the current account depicted a declining pattern and hit the lowest point in 2013. Both the unemployment rate and household savings rate fluctuated for the whole period and had the same decreasing trend. The line of household savings rate represented the strong fluctuation each year; this figure continued increasing until 1995; it started to drop slightly.





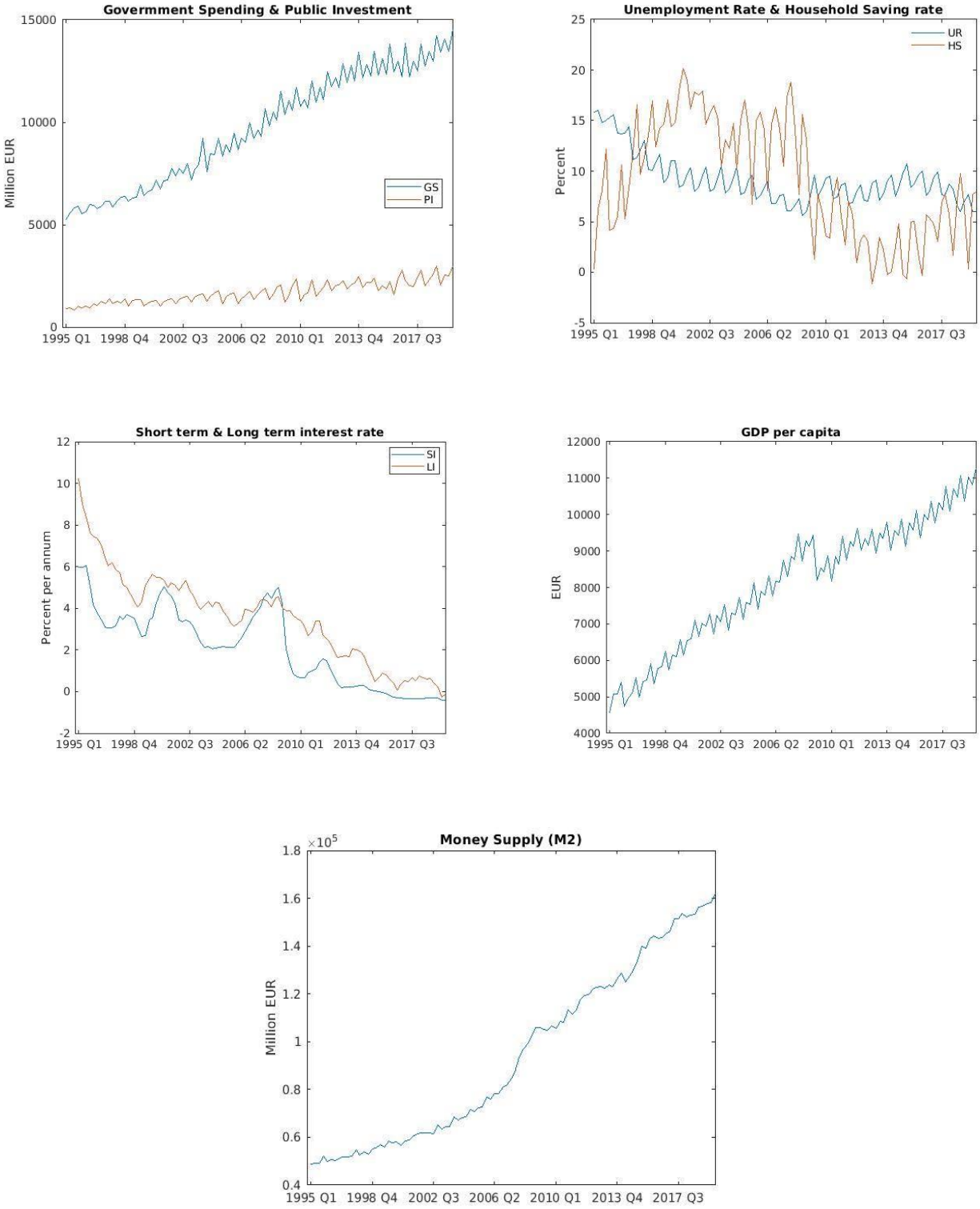


Figure 8 Previous trend of selected variables in Finland (1995Q1-2019Q4)

Table 2 provides a summary of the descriptive statistics for 11 variables. There are many categories of variables, such as absolute value, percentage value. Gross domestics saving in Finland had accounted for 26.63% GDP on average, reaching the highest level at 33.9%. The short-term interest rate had the smallest mean value of 2.1148% for the variable group of a percentage value.

Money supply (M2) appeared to have the highest value of standard deviation within the group of absolute value. The result was consistent with the above graph illustration of money supply (M2), where the line of progression during the period of 1995-2019 was continuously extending its value each year. Within the group of a percentage value, the household savings rate took the first rank in terms of standard deviation. It was in alignment with the illustrated graph as well.

Table 2 Summary statistics of the selected variables

Variable	**Code**	**Obs**	**Mean**	**Median**	**Std Dev.**	**Min**	**Max**
GDS (%GDP)	Y	100	26.630	27.320	3.844	20.310	33.900
Current Account (%GDP)	X_1	100	2.476	3.009	3.475	-4.444	8.861
GDP per capita (EUR)	X_2	100	8,119	8,370	1,744	4,560	11,280
Government spending (million EUR)	X_3	100	9,723	9,715	2,779	5,253	14,535
Net trade (million EUR)	X_4	100	1,250	1,646	1,469	-1,528	4,164
Unemployment rate (%)	X_5	100	9.165	8.633	2.401	5.567	16.000
Public investment (million EUR)	X_6	100	1,678	1,582	513	803	2,996

Household savings rate (%)	X_7	100	9.100	8.056	5.842	-1.124	20.159
Short-term interest rates (%)	X_8	100	2.115	2.140	1.860	-0.403	6.053
Long term interest rates (%)	X_9	100	3.525	3.880	2.195	-0.247	10.213
Money supply (million EUR)	X_{10}	100	94,670	85,812	36,762	48,731	162,250

Appendix 2 indicates the correlation matrix of dependent and independent variables during the period from 1995 to 2019 in Finland. The GDS rate had an unsupportive correlation with the gross domestic product per capita, government expenditure, public investment, and money supply (M2). Additionally, it had a positive association with the current account, net trade, unemployment rate, household saving, long and short-term interest rates. According to the values of correlation among the variables, most of them had the value of correlation above 0.6, which represented the moderate relationship among selected variables.

4.3 Model construction

The thesis aims to assess how well the traditional model and machine learning models perform to predict the gross domestic saving in Finland. It is essential to analyze how good the predictions are and how the different models compare to each other concerning future predictions. The empirical analysis was conducted based on both in-sample and out-of-sample prediction approaches to access the prediction error of the model.

- In – sample: the models were trained and evaluated in the same dataset, the entire data from 1995 to 2019.

- Out-of-sample: the models were trained and tested in different datasets. The holdout method, known as simple validation, was applied with a splitting ratio of 0.7. The parameters of models were estimated on the training set (70% of the dataset). The

testing was implemented on the remaining. This method was widely used due to the better accuracy of the predictive power of models when forecasting the future than in-sample predictions (Aaron, 2018).

In literature, it was prevalent to involve some lagged predictors in the economic model because they can claim whether a commercial pattern is occurring. The previous values of GDS may also impact on the current level. All models, both traditional and machine learning ones, were conducted with the help of the lagged dependent and independent variables as well. The data in this thesis were collected quarterly while the economic data are reported annually, and economic policies tend to be made every year. To predict gross domestic saving one year ahead, the minimum number of lags in all models was 4. There were four prediction models in this thesis, including one the traditional model and three based on the machine learning algorithms (described in Section 3.2). The following section provides details on the construction of prediction models.

4.3.1 The traditional model

In literature regarding macroeconomic prediction, empirical studies utilized the OLS method to study the association between selected variables, particularly in the gross domestic saving analysis. Accordingly, in this thesis, the OLS method was used to estimate the coefficients of a linear regression model of gross domestic saving in Finland, which can be formulated as follows:

$$Y_t = \alpha + \beta_1 Y_{t-4} + \sum_{i=1}^{10} \beta_{1+i} X_{i,t-4} + u_t \quad (4.1)$$

where $\alpha, \beta_1, \beta_2, \ldots, \beta_{11}$ are the regression parameters and u_t is the error term or residuals which indicate the discrepancy between the actual (Y_i) and estimated $(\hat{Y}_t)$ values.

The model can explain one complex behavior in an easily-understanding and straightforward way. Conducting models required less time and effort in computation than other models (non-linear regression). The regression coefficients were estimated by OLS

approach to minimize the loss function, which was the Residual Sum of Squares (RSS) (Ng, 2015):

$$RSS = \sum_{t=5}^{T} (Y_t - \alpha - \beta_1 Y_{t-4} - \sum_{i=1}^{10} \beta_{1+i} X_{i,t-4}) \quad (4.2)$$

According to Gauss – Markow theory, the optimal coefficients likely had the smallest variance among all linear unbiased estimations under certain assumptions. The training data set was used to compute the coefficients. Still, if the noise in training data exists, and including many independent variables, some of the assumptions are violated. It leads to an increase in predictive error when applying the model in the new data set.

4.3.2 Machine learning models

During recent years, many research papers have favored machine learning models in prediction over the more the traditional time-series approach. Four popular machine learning algorithms were selected to train on the same window data and replicate the actual gross domestic saving in Finland.

Regularized Linear Regression

The association between gross domestic saving and other macroeconomic indicators, represented in equation (4.1), was estimated by the regularized technique. They included Lasso, Ridge, and Elastic Net that minimized the loss function of regression with penalty terms (Equation 3.3, 3.4, 3.5). For each regularization technique, the parameter λ, which controlled the penalty strength, was chosen by the 5-fold cross-validation method. The training data was split into five subset data (folds). For each fold, the models were trained with different values of parameters on the prevailing four-folds, and then the model performance was evaluated on it. The parameter which produced the lowest level of average prediction error was chosen. After selecting the parameters, the regularized model coefficients were estimated on the whole training set.

Support Vector Regression

The set of regressors in support vector regression is similar to the set-in linear regression model, which means the study analyzes not only the relationship between Y and X_i but also the lagged value of regressors and predictors.

$$Y_t = f(Y_{t-4}, X_{i,t-4}) \tag{4.3}$$

where i = 1, …, 10.

As mentioned in Section 3.2.2, the support vector regression hyper-parameters, C, and ε, needed to be determined before fitting the model. The best model selected the pair of hyper-parameters by using the 5-fold cross-validation approach. The candidates for C and ε were limited in the range [0.001 1000]. Along with this, the support vector regression models with different kernel functions were trained in respect of optimizing the hyper-parameters. There are three kinds of Kernel functions, including the linear, Gaussian, and polynomial. Among searches, the best-fitted model was one that generated the minimum estimated cross-validation loss.

Decision Tree Regression

Equation (4.3) was estimated by fitting the decision regression tree. The study developed a complete tree on the training set. The hyperparameters were optimized by using the 5-folds cross-validation method. The training process followed section 3.2.3, such that the loss of the regression problem was minimized for each leaf. The model with the lowest cross-validation loss is selected.

4.4 Model evaluation

As the prediction model cannot forecast the gross domestic saving correctly, it seems necessary to analyze how well the model can perform in this context. The standard of model performance is related to prediction error or residuals, which is the deviation between the predicted values and the actual values of gross domestic saving.

There are a variety of different groups of measures, such as quadratic measures like mean squared error (MSE), and the linear measures group, including mean absolute deviation (MAD), root mean square error (RMSE), mean absolute deviation (MAD) and so on. However, the study selected two most popular measures, RMSE and MAD, to evaluate the model performance.

RMSE refers to the square root of MSE; it is the mean of the squared deviations between the actual and forecast values of GDS during a given period:

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \hat{Y}_t\right)^2 \tag{4.4}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \hat{Y}_t\right)^2} \tag{4.5}$$

where Y_t are the actual values of GDS and $\hat{Y}_t$ indicates the forecast values, t = 1, …, T periods. MAD represents the average of the absolute deviations amidst the actual values and the predicted values of gross domestic saving across a selected period:

$$MAD = \frac{1}{T}\sum_{t=1}^{T}\left|Y_t - \hat{Y}_t\right| \tag{4.6}$$

5 RESULTS

The study was conducted by using MATLAB R2020a. Overall, the traditional model was estimated by linear regression programming, while the Lasso and Elastic Net model were achieved by running lasso regularization programming. In the case of the Ridge regression model, the functions in the Global Optimization Toolbox (GlobalSearch) solving the non-convex minimization problem were used to generate the coefficients. Both support vector regression and decision tree regression, trained by using the Regression Learner app, were models subject to optimizing hyper-parameters.

The 5-fold cross-validation approach was implemented to select the parameters for all machine learning-based models. Lasso, Ridge, and Elastic Net model, SVM model, and Decision tree regression model were trained and tested on various values of parameters. The performance of the machine learning-based model could be used for further comparison with the traditional model (Linear model).

The in-sample and out-of-sample methods were applied to each model to estimate the overall performance of the models. Each model was estimated for a given period, the period from 1995 – 2019 for in-sample and period from 1995 – 2012 for out-of-sample analysis. The evaluation of all models was based on prediction performance or prediction errors, which were measured by RMSE and MAD. Lastly, a comparison between the six models (the traditional and machine learning-based model) was conducted.

5.1 In-sample analysis

In this section, all six models were trained and tested on the whole dataset. In the first subsection, estimated models not only displayed the relationship between variables over the period 1995-2019 but also provided the prediction on the value of GDS. The second subsection assessesed the performance of six models using standard model evaluation metrics, including R-squared (R^2), Root Mean Squared Error (RMSE) and Mean Absolute Deviation (MAD).

5.1.1 Model estimation

The traditional model (Linear Model)

$$Y_t = 15.05 + 0.58601 * Y_{t-4} + 0.013623X_{1,t-4} + 0.00049212X_{2,t-4} - 0.00080685X_{3,t-4} + 0.0001373X_{4,t-4} + 0.28961X_{5,t-4} + 0.00022769X_{6,t-4} + 0.12411X_{7,t-4} - 0.75305X_{8,t-4} - 0.28506X_{9,t-4} - 0.000022119X_{10,t-4}$$

If the p-value of a variable generates a significant result, it implies that the particular variable plays a substantial role in increasing GDS. The p-value of the previous value of GDS, unemployment rate, and short-term interest rates were 0.026958, 0.056435, and 0.008144, respectively. Their p-values were less than or equal to 0.05, which means those variables were significant at 5%. Thus, they generated a considerable influence on GDS. Among them, the previous value of GDS and unemployment rate were related positively, and short-term interest rates had a negative relation to domestic savings.

Regularization Models

For each regularization model in this section, the hyper-parameter λ that controlled the penalty strength was selected by the 5-fold cross-validation method. The optimal value of λ, which selected for Lasso, Ridge, and Elastic net model, might depend on the mean squared error (MSE) of the model. The optimal λ for each model was 0.0332; 9 and 0.0346, respectively. The decrease in the value of λ can lead to a decline in the value of MSE.

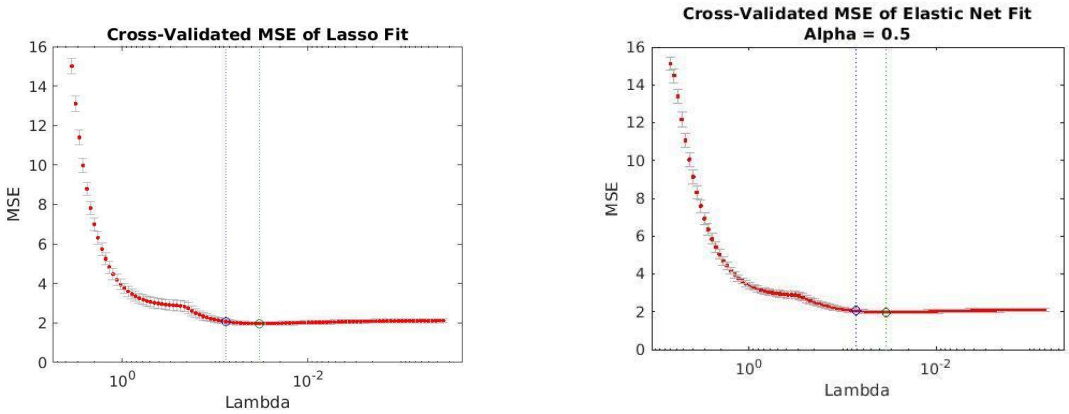


Figure 9 Parameter λ selection for Lasso and Elastic Net model (in-sample analysis)

Lasso

$$Y_t = 13.1639 + 0.712 * Y_{t-4} - 0.0005X_{3,t-4} + 0.0002X_{4,t-4} + 0.1898X_{5,t-4} + 0.107X_{7,t-4} - 0.7714X_{8,t-4} - 0.0743X_{9,t-4} - 0.00000731X_{10,t-4}$$

Ridge

$$Y_t = 21.7266 + 0.3475 * Y_{t-4} + 0.0926X_{1,t-4} - 0.0002X_{2,t-4} - 0.0003X_{3,t-4} + 0.0004X_{4,t-4} + 0.1648X_{5,t-4} - 0.0002X_{6,t-4} + 0.1683X_{7,t-4} - 0.3759X_{8,t-4} - 0.3145X_{9,t-4} - 0.0000133X_{10,t-4}$$

Elastic Net

$$Y_t = 17.337 + 0.5192 * Y_{t-4} + 0.0283X_{1,t-4} - 0.0001X_{2,t-4} - 0.0004X_{3,t-4} + 0.0003X_{4,t-4} + 0.2178X_{5,t-4} + 0.1475X_{7,t-4} - 0.5801X_{8,t-4} - 0.2807X_{9,t-4} - 0.000014X_{10,t-4}$$

In general, the result showed that essential variables were the previous value of GDS, short-term interest rates, unemployment rate, and household saving rate. Money supply (M2) did not generate an influence on gross saving. In all three regularization models, the previous value of GDS illustrated beneficial and robust importance on GDS.

The unemployment rate and household saving rate demonstrated a mild and positive sign, while the short term interest rate indicated a considerable and detrimental alliance with GDS. The unemployment rate represented the situation of the labor market. The higher the unemployment rate is, the higher the growth of domestic savings it can be (a positive sign of the coefficient). This slightly strange impact might be explained by the increasing of precautionary motive savings to handle risky prospects. The short-term interest rate was an indicator of the financial market. The coefficient value reflected the importance of monetary policy in industrial countries. A higher interest rate increased borrowing costs for consumers and added more pressure on disposable income.

Both Ridge and Elastic net models found that long term interest rates had a strong and negative effect. The current account had a slightly significant and beneficial influence on domestic saving, which illustrated the economy's ability to withstand unexpected

economic shocks. The larger the current account surplus, the more exports than imports occur, and thus increase savings.

Support Vector Regression

The optimal support vector regression model had the Kernel function type of linear, as illustrated in figure 10. C and ε (hyper-parameters of the support vector regression model) were selected using the 5-fold cross-validation approach. The limitation of optimal values for C and ε was in the range of [0.001 1000]. It generated the minimum estimated cross-validation loss. All data had been standardized.

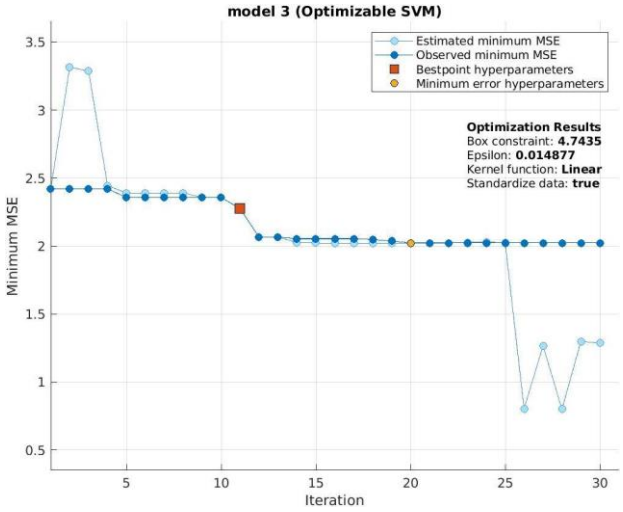


Figure 10 SVM optimization (in-sample analysis)

Decision Tree Regression

The hyper-parameters were optimized by applying 5-folds cross-validation method. The optimal model, which had the lowest cross-validation loss, had the number of nodes is 33. The decision tree regression result was illustrated in the below graph. The repetition of variables in the tree model will emphasize the importance of these variables in GDS prediction. In this study, 9 variables included money supply M2, long term interest rate, the previous value of domestic saving, short term interest rate, GDP per capita, government spending, public investment, net trade, and current account constructed the tree. Among them, money supply (M2) was the most important independent variable used

for the primary split to split five nodes, and the long-term interest rate was used to split three nodes. Short term interest rate was used to split two nodes. The previous value of domestic saving, GDP per capita, government spending, public investment, net trade, and current account were only used once to split the nodes.

```
 1  if money supply (M2)<95246.5 then node 2 else if money supply (M2)>=95246.5 then node 3 else 26.6143
 2  if long-term interest rate <7.19833 then node 4 else if long-term interest rate >=7.19833 then node 5 else 29.8525
 3  if the previous value of domestic saving <21.6613 then node 6 else if the previous value of domestic saving >=21.6613 then
node 7 else 22.7873
 4  if money supply (M2)<61895.5 then node 8 else if money supply (M2)>=61895.5 then node 9 else 30.2999
 5  fit = 26.423
 6  if money supply (M2)<131998 then node 10 else if money supply (M2)>=131998 then node 11 else 21.371
 7  if Short-term interest rates <-0.14295 then node 12 else if Short-term interest rates >=-0.14295 then node 13 else 23.3812
 8  if GDP per capita <5500 then node 14 else if GDP per capita >=5500 then node 15 else 31.1683
 9  if Government spending <8962.23 then node 16 else if Government spending >=8962.23 then node 17 else 29.3525
10  fit = 21.0047
11  fit = 22.1954
12  if long-term interest rate <0.493333 then node 18 else if long-term interest rate >=0.493333 then node 19 else 24.1821
13  if money supply (M2)<108344 then node 20 else if money supply (M2)>=108344 then node 21 else 22.9408
14  fit = 29.7222
15  if long-term interest rate <5.27333 then node 22 else if long-term interest rate >=5.27333 then node 23 else 31.5489
16  if Public investment <1396.5 then node 24 else if Public investment >=1396.5 then node 25 else 28.9287
17  if Short-term interest rates <3.7074 then node 26 else if Short-term interest rates >=3.7074 then node 27 else 29.8611
18  fit = 24.0136
19  fit = 24.2784
20  if Money supply (M2) <105540 then node 28 else if Money supply (M2) >=105540 then node 29 else 23.8104
21  if Net trade <-315.5 then node 30 else if Net trade >=-315.5 then node 31 else 22.0711
22  if Current Account <6.91828 then node 32 else if Current Account >=6.91828 then node 33 else 31.0443
23  fit = 32.4138
24  fit = 29.3493
25  fit = 28.6282
26  fit = 30.3531
27  fit = 29.123
28  fit = 24.1188
29  fit = 23.348
30  fit = 21.5746
31  fit = 22.5676
32  fit = 31.4808
33  fit = 30.6078
```

Figure 11 Decision tree regression model (in-sample analysis)

5.1.2 Model evaluation

This section aims to identify the best performance model that represents the research data and how well the selected model may work in the future. The predictive results from six models were compared with each other by ranking them individually according to their predictive accuracy.

Figure 1 presents the actual value of quarterly gross domestic saving and its forecast results obtained from each model over the 1995 Q1 – 2019 Q4 period. It can be seen that all six models have successfully shown the trend of GDS over the period. The illustrative line of each model represented a similar result in capturing the growth and decline tendency of actual data with a small deviation. However, decision tree performance stood out since it provided the most accurate forecast in terms of predicting the peak and dip of the data. The decision tree model could successfully capture the significant upturns and downturns in the GDS data. In contrast, other models could not demonstrate the highest point and the lowest point as similar to the decision tree. It indicated that the decision tree model effectively learned the underlying relationships between selected macroeconomic variables.

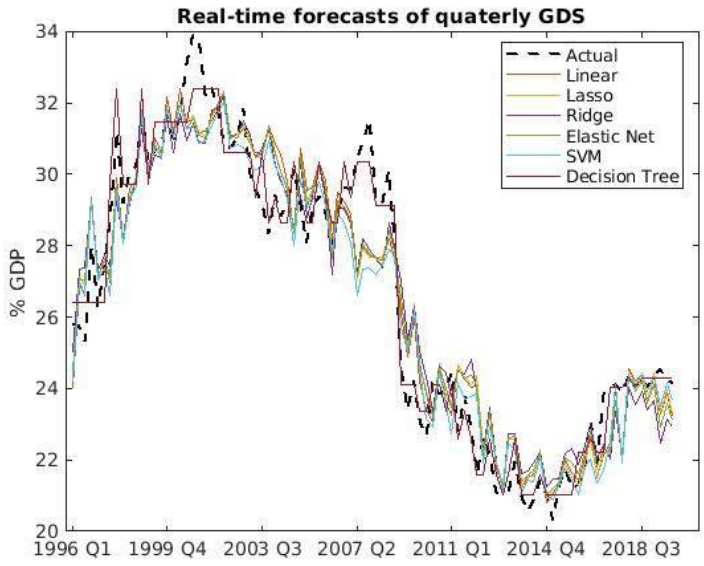


Figure 12 Real-time forecasts of quarterly GDS (1996Q1-2019Q4)

In the below section, 3 model evaluation metrics, including R-squared (R^2), Root Mean Squared Error (RMSE) and Mean Absolute Deviation (MAD), were used to quantify the predictive model performance.

Table 3 Model performances, 1995Q1-2019Q4 (in-sample analysis)

	RMSE	RMSE (Rel. to Linear)	MAD	MAD (Rel. to Linear)
Model 1 - Linear	1.3348		0.9769	
Model 2 – Lasso	1.2634	0.9465	0.9677	0.9905
Model 3 – Ridge	1.3237	0.9916	1.0598	1.0848
Model 4 – Elastic net	1.2666	0.9489	0.9860	1.0093
Model 5 – SVM	1.2950	0.9701	0.9273	0.9492
Model 6 – Decision tree	0.6489	0.4861	0.4914	0.5030

Metric R^2 evaluated the strength of the association between the selected model and the dependent variables on a convenient 0 – 100% scale. If the regression model had a perfect fit for the total observations, R^2 is 100%. According to table 2, all six models had a relatively high value of R^2. There was not a big difference between the traditional model and machine learning-based model regarding the R^2 value. Nevertheless, the decision tree model had the highest R^2 value of 97.31%. This highest value of R^2 indicated that the decision tree was the best performing model in terms of precise predictions among the six models.

Metric RMSE was a standard metric to measure the error rate of a regression model. RMSE was a good measure of how accurately the model forecasts the outcome, and it was the most critical criterion in terms of data fitting for prediction. In this in-sample analysis, all machine learning-based models produced forecasts that had RMSEs higher than the linear benchmark. The top three performance models were decision tree, Lasso, and Elastic Net. Among the six models, the traditional linear model had the highest RMSE

value of 1.3348, while the decision tree model had the smallest RMSE value of 0.6489. It indicated that the decision tree could reduce almost half of the forecast errors. As measured by RMSE, the decision tree model completely outperformed and two times more accuracy than the traditional linear model.

For measure MAD, the top performance model, decision tree with the lowest MAD value of 0.4914, might reduce the forecasting error calculation by approximately 50% compared to the linear model benchmark. The second and third best performance models, which are SVM and Lasso, had quite similar values of MADs, which were lower than the traditional linear model. With the MAD value of 1.0598, the Ridge model had the highest MAD value, double the MAD value than the decision tree model, and was the least accurate model as measured by MAD.

The assessment results through standard model evaluation metrics indicated that the vast majority of machine learning models produce forecasts that have RMSEs and MADs lower than the traditional linear model. As a result, machine learning models likely generated more accurate results than the conventional linear model. Among other machine learning models, the decision tree was the best model for prediction with the highest value of R-squared, the lowest value of RMSE, and MAD.

5.2 Out-of-sample analysis

5.2.1 Model estimation

The traditional Model (Linear Model)

Following the methodology of the out-of-sample study illustrated in Section 4.2, all six models were estimated on the same training set, which accounted for 70% of the dataset). The OLS regression traditional model used the lagged variables, was applied to predict the gross domestic saving (% GDP) in Finland. The estimates of the linear model were:

$$Y_t = 33.564 + 0.242\,Y_{t-4} - 0.029\,X_{1,t-4} + 0.0007\,X_{2,t-4} - 0.0004\,X_{3,t-4} + 0.00008\,X_{4,t-4} - 0.215\,X_{5,t-4} - 0.0005\,X_{6,t-4} + 0.088\,X_{7,t-4} - 0.543\,X_{8,t-4} - 0.155\,X_{9,t-4} - 0.0001\,X_{10,t-4}$$

The model trained on the level data generates robust-looking results with quite a high adjusted R^2 value (80%) and significant F-stat at 1% (p-value = 3.24e-17). The Linear model could explain 83.4% of the real value of GDS in Finland, but only two estimates were being significant at 1%. The intercept played an essential role in prediction while money supply (M2) was the only indicator that had a slightly negative influence on gross domestic saving rate. These showed worsen results than the Linear model from in-sample analysis, which produced five significant estimates. They also consisted of money supply (M2) that implied the critical role of monetary policies or Central Bank in gross domestic saving in Finland.

Regularization Models

The study used the 5-fold cross-validation method to determine the optimal parameters λ in Lasso, Ridge, and Elastic Net models. Each model was formed on the average estimated prediction errors, which are measured by MSE. The results of the selection are shown in Figure 13.

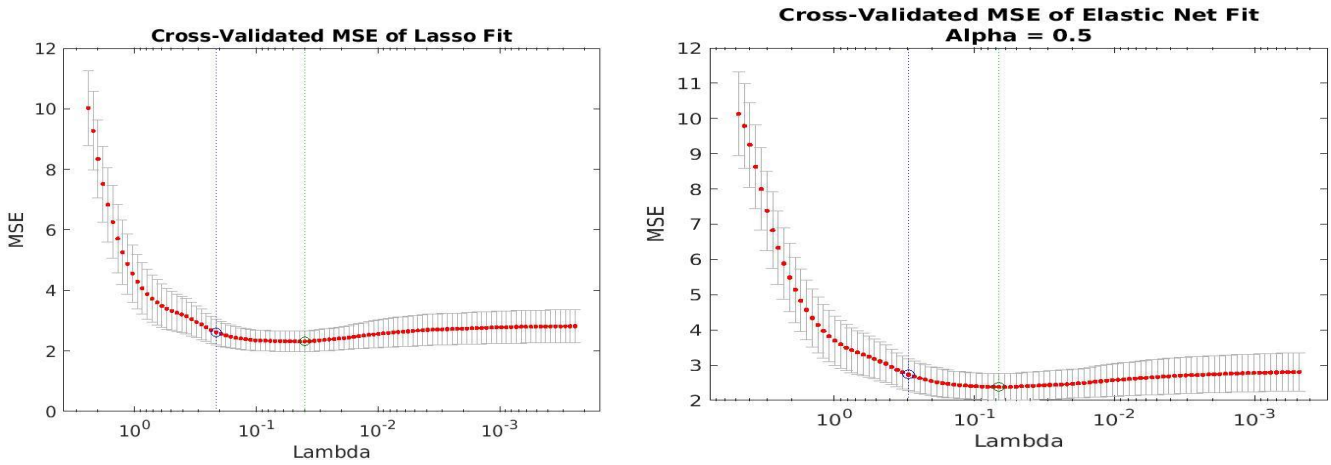


Figure 13 Parameter λ selection for Lasso & Elastic Net model (Out-of-sample analysis)

It showed that Lasso and Elastic Net models had the same pattern of the estimated prediction errors (MSE). MSE initially experienced a sharp decrease and later started to improve as the value of Lambda reduced. It meant that the model performance could be improved if the parameter's value was reduced. Also, the variation of prediction errors for the Elastic Net model was more significant than the Lasso model. Parameter λ = 0.0401, and 0.0666 produced the best fitted Lasso, Ridge, and Elastic Net model, respectively. As a result, the lowest RMSE on average for each model equaled 1.5216 for Lasso, 2.1456 for Ridge, and 1.5430 for Elastic Net. At the optimal value of Lambda, three regularization models were estimated and expressed as follows:

Lasso

$$Y_t = 28.2889 + 0.3827\, Y_{t-4} - 0.1180\, X_{5,t-4} + 0.0816\, X_{7,t-4} - 0.4495\, X_{8,t-4} - 0.2157\, X_{9,t-4} - 0.0001\, X_{10,t-4}$$

Ridge

$$Y_t = 30.3289 + 0.2712\, Y_{t-4} + 0.0402\, X_{1,t-4} - 0.0002\, X_{2,t-4} - 0.0003\, X_{3,t-4} + 0.0003\, X_{4,t-4} - 0.0869\, X_{5,t-4} - 0.0006\, X_{6,t-4} + 0.1237\, X_{7,t-4} - 0.3728\, X_{8,t-4} - 0.3075\, X_{9,t-4} - 0.0001\, X_{10,t-4}$$

Elastic Net

$$Y_t = 26.3591 + 0.3853\, Y_{t-4} + 0.00008\, X_{4,t-4} - 0.0396\, X_{5,t-4} + 0.1012\, X_{7,t-4} - 0.4205\, X_{8,t-4} - 0.2205\, X_{9,t-4} - 0.0001\, X_{10,t-4}$$

Regularization models could explain the real data quite accurately due to the high R-square (83.4%, 82.26%, and 82.9%, respectively). Furthermore, three models showed consistent results related to the significant effects of the previous value of gross domestic saving, unemployment rate, household saving rate, and interest rate (both short-term and long-term) on GDS in Finland. While the previous values of GDS and household saving had a positive impact, the interest and unemployment rate brought a negative relationship. The current account, government spending, and GDP per capita did not affect GDS suggested by both Lasso and Elastic Net, but the Ridge model impliedthe small impact.

The estimation of regularization models emphasized the role of Central Bank again in forecasting GDS in Finland. In terms of analyzing the critical drivers of GDS, regularization models produced similar results as those models estimated through the in-sample approach.

Support Vector Regression

The hyper-parameters of the SVM model were determined by 5-fold cross-validation to optimize the best option. The optimal model was one that has the best prediction performance (MSE) and trained with different types of Kernel functions and standardized data. The final SVM came with the cubic Kernel function, which was displayed in Figure 5. This model obtained a higher R-square value of 84.13% than regularization models and Linear model as well but did not represent a clear relationship between GDS and other indicators.

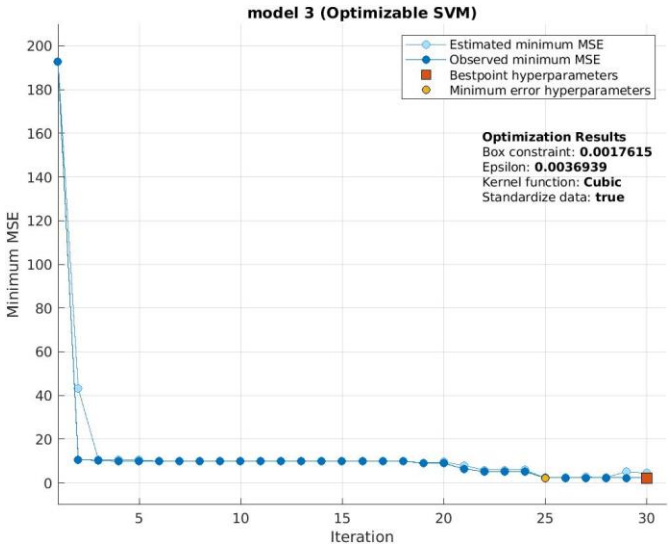


Figure 14 SVM optimization (out-of-sample analysis)

Decision Tree Regression

1 if Money Supply (M2) <95246.5 then node 2 else if Money Supply (M2) >=95246.5 then node 3 else 28.4915
2 if Unemployment rate <14.5667 then node 4 else if Unemployment rate >=14.5667 then node 5 else 29.8525
3 if Money Supply (M2) <108344 then node 6 else if Money Supply (M2) >=108344 then node 7 else 23.4364
4 if Money Supply (M2) <61895.5 then node 8 else if Money Supply (M2) >=61895.5 then node 9 else 30.2999

```
 5  fit = 26.423
 6  if Money Supply (M2) <105540 then node 10 else if Money Supply (M2) >=105540 then node 11 else 23.8104
 7  fit = 22.5013
 8  if GDP per capita <5500 then node 12 else if GDP per capita >=5500 then node 13 else 31.1683
 9  if Government spending <8962.23 then node 14 else if Government spending >=8962.23 then node 15 else 29.3525
10  fit = 24.1188
11  fit = 23.348
12  fit = 29.7222
13  if Long term interest rates <5.27333 then node 16 else if Long term interest rates >=5.27333 then node 17 else 31.5489
14  if Household savings rate <11.818 then node 18 else if Household savings rate >=11.818 then node 19 else 28.9287
15  if Short-term interest rates <3.7074 then node 20 else if Short-term interest rates >=3.7074 then node 21 else 29.8611
16  if Short-term interest rates <3.22403 then node 22 else if Short-term interest rates >=3.22403 then node 23 else 31.0443
17  fit = 32.4138
18  fit = 28.2591
19  fit = 29.1518
20  fit = 30.3531
21  fit = 29.123
22  fit = 31.9103
23  fit = 30.7556
```

Figure 15 Decision tree regression model (Out-of-sample analysis)

The decision tree model with the best performance was generated from a parameter note of 23, explaining up to 95.16% real data, which indicated the spurious relation. Similar to other machine learning-based models, this model presented that financial variables and household saving rates had an impact on GDS. The reason behind that is they were used as a split rule at each node and used many times, especially the Money Supply (M2), affecting GDS mostly. Besides, the unemployment rate, GDP per capita, and government spending were also factors driving GDS in Finland. According to the decision tree estimated with the help of the in-sample method, previous values of GDS, net trade, public investment, and the current account created a small impact on GDS but not the results from model estimated through the out-of-sample approach.

In short, all estimated models could explain the majority of the GDS real values in Finland during the training period from 1995 – 2012 because the R-squared values were higher than 82%. The decision tree regression model produced the highest value, while the remaining models had the values which were not much different from each other. This

result was the same as the results of the in-sample analysis. Moreover, they also displayed the same vital drivers of GDS in Finland, which focused on the financial indicators.

5.2.2 Model evaluation

After training and choosing the optimal models, all six models, including both the traditional model and machine learning-based models, were evaluated on the same test set. The data from the test set accounted for 30% of the dataset containing the data from 2013 – 2019. The models' performance (RMSE and MAD) were analyzed not only for the test period but also for the training period.

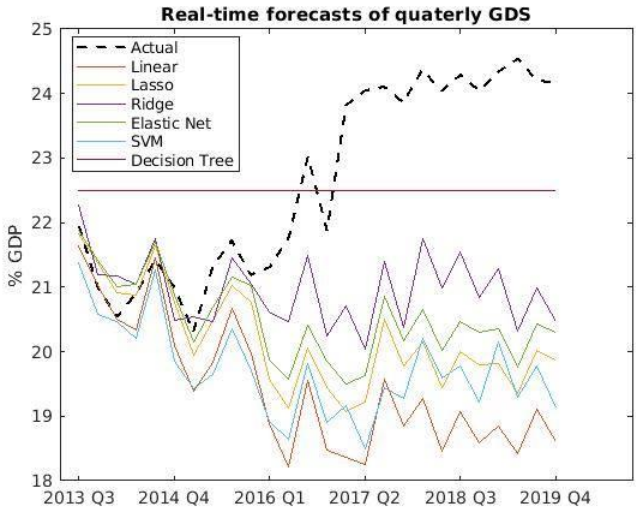


Figure 16 Real-time forecasts of quarterly GDS (2013Q3-2019Q4)

Figure 16 represented the forecasts of quarterly GDS (% GDP) in Finland obtained from each model for the test period 2013 – 2019. The graph indicated that all machine learning models predicted GDS values quite closed to the actual values. They successfully predicted the sharp downturn of GDS, which occurred from 2014 to 2015 but could not capture the dramatic increase tendency after that. Additionally, it could be seen that the

gap between the linear model and the actual line was the biggest, and the decision tree model produces unchanged GDS during the whole period.

Table 4 Model performances, 1995Q1-2019Q4 (out-of-sample analysis)

	Training Set				Test Set			
	RMSE	RMSE (Rel. to Linear)	MAD	MAD (Rel. to Linear)	RMSE	RMSE (Rel. to Linear)	MAD	MAD (Rel. to Linear)
Model 1 - Linear	1.4088		1.0387		3.9057		3.2227	
Model 2 – Lasso	1.2880	0.9142	1.0442	1.0053	3.1678	0.8111	2.5483	0.7907
Model 3 – Ridge	1.3345	0.9473	1.0708	1.0309	2.3155	0.5929	1.8438	0.5721
Model 4 – Net	1.3020	0.9242	1.0466	1.0076	2.8272	0.7239	2.2556	0.6999
Model 5 – SVM	1.2544	0.8904	0.7654	0.7369	3.4587	0.8855	2.9416	0.9127
Model 6 – Decision tree	0.6926	0.4916	0.5496	0.5291	1.4777	0.3783	1.4060	0.4363

The performance of models on both training and test set measured by RMSE and MAD were presented in Table 5. The results showed that all machine learning-based models were trained better than the traditional model because they had lower RMSE. The top three trained models were Decision Tree, SVM, and Lasso, among which SVM and Lasso could reduce the prediction errors by approximately 8% - 10% relative to the Linear model. At the same time, it could be around 50% for the Decision Tree model. In terms of MAD, only SVM and Decision Tree also overperformed the traditional model on the training set and the most fitted model that was the Decision tree model with MAD = 0.5469. The regularization models could produce prediction errors quite close to the Linear model that was demonstrated via relative values of MAD to the Linear model (around 1). Among ML

models, the Ridge model had the highest prediction errors for both cases, with RMSE = 1.3345 and MAD = 1.0708.

During the test period, all machine learning models outperformed the traditional model, which applied for both measures of prediction accuracy (RMSE and MAD). In particular, among machine learning models, Decision Tree generated the lowest prediction errors with RMSE of 1.4777 and MAD of 1.4060. At the same time, SVM had the highest value, which only reduced the prediction error by 9-12 percentages relative to the traditional model. For both cases (RMSE and MAD), three regularization models could decrease the prediction errors by approximately 20% -40% relative to the Linear model. Among them, the Ridge model likely forecasted the GDS values that have the highest accuracy.

On both the training and test set, the Decision Tree Regression model was the most fitted model, which produced the lowest prediction errors. It applied to both RMSE and MAD; hence the use of the Decision Tree model to predict GDS in Finland could bring more efficiency and accuracy in terms of the policymaking process. Besides, all machine learning models performed better in both training and test set than the traditional model. The Regularization models produced stable results in the new dataset (test period) as the RMSE and MAD were not much different relatively compared to the Linear model. In short, by applying machine learning models, we could achieve better predictive performance, despite the measures of prediction errors.

5.3 Summary

Both the traditional and machine learning-based model can not only analyse the relationship between the determinants of domestic savings but also can predict the value of gross domestic saving.

Regardless of in-sample and out-of-sample analysis, all models were trained successfully with R^2 higher than 80 percentages. Among them, the decision tree model proved to be the best-fitted model in terms of model estimation because the value of R^2 were around 97

percent in both cases. Most models, except SVM and decision tree, did not show the relationship among variables. They emphasized the essential roles of financial variables such as money supply M2, short-term interest, long term interest rate as well as previous value of gross domestic saving on GDS in Finland. Mainly, in-sample and out-of-sample analysis of regularized linear regression models showed similar results. The models suggested that financial variables had an unsupportive correlation with savings.

In contrast, the previous value of savings had a beneficial relationship with gross saving. Furthermore, the in-sample analysis of the traditional model specified the relationship between total saving and various determinants. In contrast, in the out of sample analysis, money supply M2 was the only determinant that had a negative influence on savings.

All the machine learning models outperformed the traditional linear model by reducing the forecasting error calculation. The assessment results through standard model evaluation metrics indicated that the vast majority of machine learning models gave the predictions with MADs and RMSEs lower than the traditional linear model. Among all six models, the decision tree had the best prediction performance with the highest value of R-squared, the lowest value of RMSE, and MAD in both in-sample and out-of-sample analysis. Regularized linear regression models generated stable predictive results in both in-sample and out-of-sample analysis. For out-of-sample analysis, regularized linear models provided forecasts with not much difference between the training set and test set.

6 CONCLUSIONS

The conclusion chapter includes two parts. The first part may conclude the findings and answers to the research questions. The last part may present the limitations of the study and propose implications for further studies.

6.1 Summary of the findings

RQ 1. What are the main determinants that account for the GDS in Finland?

In the empirical overview chapter, a literature analysis was conducted to investigate factors that affected GDS in Finland over the period 1995-2019. The investigation was guided by the Life Cycle Model's theoretical framework and previous empirical studies. Following research theory and data availability, the determinants to be used in the empirical specification were 11 variables that were under the group of economic growth, financial variables, government policy, and other macroeconomic variables. Through the model estimation process, this thesis identified that the critical factors of the domestic saving rate in Finland throughout this period were financial and income variables. Financial considerations such as money supply, short-term interest rates, and long-term interest rates had an unsupportive effect on gross domestic saving.

In contrast, the previous value of GDS had a beneficial effect. This finding was aligned with the empirical outcomes of cross-country studies from Metin Ozcan (2003, 1405-1416), Imran et al. (2010), Houérou (2011), which pointed out that the financial system might improve the private saving rates. Furthermore, the unemployment rate and household saving rate had an enhancing influence on the savings of Finland, which indicated the precautionary motive for savings. These empirical results were consistent with the findings in some research papers such as Houérou (2011), and Mishi (2012). This understanding re-established the importance of the life-cycle approach (Modigliani, 1986), which was the most commonly used model regarding the explanation of savings behavior.

From a policy development point of view, in the context of designing effective policy interventions, this study hopes to bring attention to the policymaker that financial determinants are the main factors that account for Finland's domestic savings. Since savings is essential to achieve economic growth, policies promoting savings can be used to encourage the economic growth of Finland. Sound macro policies and government measures can control the excessive growth of consumer loans. The decrease in financial vulnerabilities can be sufficient to promote growth in savings. The rate of return is another beneficial influence. Maximizing the profits on savers and reducing the gap between the lending and borrowing interest rates can make the returns on savings more attractive to savers.

RQ2. How to conduct the prediction of domestic savings by utilizing the traditional regression models in Finland?

Many previous studies used OLS estimations technique to study the influence of major potential savings determinants through cross-section data or international panel data, or time-series data on different nations (Appendix 1). Although considerable studies devoted to the association of GDS and its determinants in conjunction with the usage of OLS estimation results, less research explored in the prediction area.

In this thesis, OLS method was used to estimate the coefficients of a linear regression model of GDS in Finland. The practical experiment was deployed in MATLAB R2020a, and the linear regression program determined the traditional model.

Model estimation and forecast performance evaluation were conducted per in-sample and an out-of-sample approach. The study adopted the holdout evaluation technique with a splitting ratio of 0.7 to identify the transformation of the model generalizes on out-of-sample predictions. The economic policies are usually developed yearly, whereas the financial data in this thesis were collected quarterly. As a result, the minimum number of lags used in all models is 4 to predict the GDS one year ahead.

RQ3. How to apply the machine learning model in the prediction of domestic savings in Finland?

Recently, the relative success of the machine learning models in macroeconomic forecasting over the traditional time-series techniques has become a favorite topic for analysis (Aaron, 2018). Major statistical models in time series forecasting, including linear regression and non-linear methods, depend upon a set of different choices that influence model complexity and predicting performance. By utilizing intensive computation models, machine learning approaches can simplify the mentioned set of options efficiently, detect the optimal model complexity, and discover complicated hidden correlations faster than the conventional statistical models.

This thesis aims to examine whether machine learning models can yield a more accurate prediction of GDS than the traditional regression-based model. The application of machine learning in the empirical specification seems to be the most popular approach as it has been utilized by many researchers such as Richardson et al. (2018), Aaron (2018). Some typical learning machine models are Lasso, Ridge, and Elastic Net model, the SVM model, and the Decision Tree Regression model (Wang et al., 2014).

The empirical study was implemented using MATLAB R2020a. Lasso and Elastic Net models were estimated by running Lasso regularization programming. In the case of the Ridge regression model, the functions in the Global Optimization Toolbox were used to generate the coefficients. Both SVM and Decision Tree regression were trained with the Regression Learner app.

The study conducted an in-sample and out-of-sample analysis. The model was first estimated by selecting hyper-parameter. All the hyper-parameters of machine learning models were chosen by using the 5-fold cross-validation method. Regarding regularized linear regression models, parameter λ was used in Lasso, Ridge, and Elastic Net models. Concerning SVM, the pair of hyper-parameters were C and ε. Lastly, the hyper-parameter of the decision tree regression model was the number of nodes or the number of leaves.

RQ4. Do machine learning-based models provide more accurate predictions than the traditional econometric models?

The assessment results based on model evaluation metric RMSE indicated that all machine learning models produced forecasts that had lower RMSEs than the traditional linear model. In terms of metric MAD, the vast majority of the machine learning models provided estimations that had MADs lower than the traditional one. Specifically, the decision tree could reduce almost half of the forecast errors. As measured by RMSE, the decision tree model completely outperformed and two times more accuracy than the traditional linear model. Overall, machine learning models can achieve more accurate predictions than the conventional linear model.

Among five machine learning models, the decision tree regression model was the most fitted one, which produced the lowest prediction errors given the lowest value of RMSE and MAD in both in-sample and out-of-sample analysis. Hence, the usage of the decision tree model to predict GDS in Finland could bring more efficiency and accuracy in terms of the policymaking process. The regularization models produced stable results in the new dataset (test period) as the RMSE and MAD were not much different relatively compared to the Linear model. In short, the predictors can get better predictive performance by applying machine learning models, despite the measures of prediction errors.

6.2 Limitations and recommendations for future studies

Based on the literature review concerning the life cycle model, the demographic factors also play a considerable role that accounts for savings. For instance, previous empirical studies from Narayan et al. (2006), Horioka and Hagiwara (2010), Houérou (2011), Jilani et al. (2013), Khan et al. (2017) found a possible connection between the high value of age dependency ratio and the youth dependency ratio with the lower private savings. Metin Ozcan (2003, 1405-1416) also found that life expectancy tends to enhance private savings. Due to data availability and limited source of data collection, this thesis could not

include the set of variables under the category of demographic variables, which include the age distribution of the population, the urbanization ratio, and life expectancy. It is suggested that future studies may contain primary data of demographic variables on the topic of domestic saving to achieve better prediction. Since machine learning models have advantages in handling large and diverse data, increasing independent variables under the category of demographic variables may generate more valuable and accurate results, which may improve understanding of the subject as well as the quality of future research studies.

Because of time-wise limitations, this thesis selected the most popular machine learning models of Regularized Linear Regression, Support Vector Machine Regression, Decision Tree Regression. However, it could have been interesting to examine other machine learning models, which are also quite popularly used in time series prediction in future studies. For example, Richardson et al. (2018) showed that K-nearest neighbor (KNN) and neural network (NN) models are also able to reduce the average forecast errors relative to the traditional linear model. Furthermore, combining the prediction of the machine learning models using various weighting schemes leads to further improvements in predictive performance. Another interesting comparison research, as discussed by Ahmed (2010), who applied the K-nearest neighbor (KNN) and neural network (NN) learning algorithm on the monthly M3 time series competition data. He found that neural networks gave the best performance, while K-nearest neighbor regression provided an average result.

The empirical findings in this thesis indicated several determinants that account for gross savings in Finland. This thesis studied the relationship between gross savings and other variables using both traditional and machine learning models. These variables visibly presented the role of monetary policies and government measures that affect savings. When designing a strategy, the use of the machine learning model is not only limited to providing forecasting value of gross domestic saving but also gives a chance to stimulate different scenarios to tackle the complex problem of uncertainties and assumptions. For example, to develop government response to Coronavirus pandemic, policymakers might

consider using suitable machine learning models in Scenario Plan Analysis to estimate the optimal government spending to encourage savings, thus recover economic growth. By reviewing different scenarios, the policymaker can implement different strategies for each of the context variables. The study's results, therefore, recommend the usage of machine learning algorithms along with the traditional regression model as an additional tool of policy planning.

REFERENCES

Aaron, H. (2018). Machine Learning Approaches to Macroeconomic Forecasting. Federal Reserve Bank of Kansas City. [www document]. [Accessed 15 March 2020]. DOI: 10.18651/ER/4q18SmalterHall

Ahmed, N. & Atiya, Amir & Gayar, N. & El-Shishiny, H. (2010). An Empirical Comparison of Machine Learning Models for Time Series Forecasting. *Econometric Reviews*. 29. 594-621. 10.1080/07474938.2010.481556.

Ahmed, S. & Aleemi, Abdur R. & Tariq, M. (2015). The Determinants of Savings: Empirical Evidence from Pakistan. *International Journal of Management Sciences and Business Research*. Volume 4. 63-71.

Alguacil, M., Cuadros, A., Orts, V. (2004). Does Saving Matter for Growth? Mexico (1970-2000), *Journal of International Development*, 16, Issue 2.

Awad, M., & Khanna, R. (2015). Support Vector Regression. In: Efficient Learning Machines. CA: Berkeley.

Basely, T., & Costas, M. (1998). Do tax incentives raise private saving? World Bank Document.

Bebczuk, R. N. (2000). Productivity and Saving Channels of Economic Growth as Latent Variables: An Application of Confirmatory Factor Analysis, *Estudios de Economia*, Vol. 27, Nr 2.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Blomström, M., Lipsey, R. E., & Zejan, M. (1996). Fixed Investment the Key to Economic Growth. *Quarterly Journal of Economics*, 111.

Board of Governors of the federal reserve system website. [Accessed 21 Jan 2020]. Available https://www.federalreserve.gov/faqs/money_12845.htm

Bonham, C., & Wiemer, C. (2012). Chinese saving dynamics: the impact of GDP growth and the dependent share. *Oxford Economic Papers*, 65(1), 173-196.

Bosworth, B. & Chodorow-Reich, G. (2007). Saving and Demographic Change: The Global Dimension. *SSRN Electronic Journal*. 10.2139/ssrn.1299702.

Bosworth, B. P. (1993). Saving and investment in a global economy.

Botha, F., Simleit, C., & Keeton, G. (2011). The determinants of household savings in South Africa. *Studies in Economics and Econometrics*, 35(3), 1-20.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Cao, P., Zhao, D., and Zaiane, O. (2013). An optimized cost-sensitive SVM for imbalanced data learning. *In Advances in Knowledge Discovery and Data Mining*, Pp. 280–292. Springer.

Carroll, C. D., & Weil, D. N. (1994). Saving and growth: a reinterpretation. In the Carnegie Rochester conference series on public policy (Vol. 40, pp. 133-192). North-Holland.

Chaudhry, I., Faridi, M. Z., Abbas, M., & Bashir, D. (2015). Short-run and long-run saving behavior in Pakistan: An empirical investigation.

Cherkassky, V. & Mulier, F. (2007). Learning from Data: Concepts, Theory, and Methods, 2nd Edition. Wiley.

Corbo, V., & Schmidt-Hebbel, K. (1991). Public policies and saving in developing countries. *World Bank Publications*.

Coulombe, P., Leroux, M., Stevanovic, D. (2019) How is Machine Learning Useful for Macroeconomic Forecasting? CIRANO Working Papers 2019, 22, CIRANO.

Denizer, C., & Wolf, H. C. (2000). The savings collapse during the transition in Eastern Europe. The World Bank.

Domar, E. D. (1946). Capital Expansion, Rate of Growth, and Employment, Econometrica, Nr 14.

Duesenberry, J.S., (1949). Income, Savings and theory of Consumer Behavior, Harvard University Press, Cambridge, Mass.

Duran, E., Uzgur Duran, B., Akay, D., and Boran, F.E. (2017). Grey relational analysis between Turkey's macroeconomic indicators and domestic savings. Grey Systems: Theory and Application 7(1), pp. 45–59.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96 (456), 1348-1360.

Freidman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning, 1.

Friedman, C. (1984). Classification and regression trees. CRC press.

Friedman, M. (1957). A theory of the consumption function: A study by the National Bureau of Economic Research. Princeton University Press.

Fuller, Kathleen, and Benjamin M. Blau. (2010). Signaling, free cash flow, and nonmonotonic dividends. *Financial Review*, 45, no. 1 21-56.

Fuller, R., Han, B., & Tung, Y. (2010). Thinking about Indices and Passive versus Active Management. *The Journal of Portfolio Management*, 36 (4), 35-47.

Gale, Mayiam G., and Peter R. Orszag. (2004). Budget deficits, national saving, and interest rates. Brookings Papers on Economic Activity 2004, no. 2 101-210.

Gavin, M., Hausmann, R., and E. Talvi (1997). Saving behavior in Latin America: Overview and Policy issues.

Giovannini, A. (1985). Saving and the real interest rate in LDCs. *Journal of Development Economics* 18, no. 2-3: 197-217.

Gujarati, Damodar N. (2012). Basic econometrics. Tata McGraw-Hill Education.

Hammad, Ahmad K., Hafizah & Abdullah, H. (2010). Saving Determinants in Malaysia. Jurnal Ekonomi Malaysia. 44. 23-34.

Harrod, R. (1939). An Essay in Dynamic Theory, *Economic Journal*, Nr 49.

Haruna, I. (2011). Determinants of saving and investment in deprived district capitals in Ghana: a case study of Nadowli in the upper west region. JCSS. 4. 1-12.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer.

Hausman, Jerry A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society* 1251-1271.

Hausmann, R., Reisen, H., Promoting Savings in Latin America. Organization for Economic Cooperation and Development and Inter-America Development Bank, Paris.

Heer, B., and Bernd, S. (2009). The savings inflation puzzle. Applied Economics Letters 16, no. 6 615-617.

Higgins, E. T., & Silberman, I. (1998). Development of regulatory focus: Promotion and prevention as ways of living. In: J. Heckhausen & C. S. Dweck (Eds.), Motivation and self-regulation across the life span (p. 78–113). Cambridge University.

Hoerl, A., & Kennard, R. (1988). Ridge regression. In Encyclopedia of statistical sciences (Vol. 8). New York: Wiley.

Horioka, C., and Hagiwara, A. (2010). Determinants and long-term projections of saving rates in developing Asia. *Asian Development Bank Economics Working Paper Series No. 228*.

Houérou, P. (2011). Sustaining High Growth: The Role of Domestic Savings. Turkey Country Economic Memorandum. Synthesis Report, conference version. Report No. 66301-TR. The ministry of development of Turkey and the World Bank.

Hubbard, R., Zeldes, P. (1995). Precautionary Saving and Social Insurance. *Journal of Political Economy*. Vol. 103, No. 2, pp. 360-399

Hussain, M. & Brookins, Oscar T. (2001). On the determinants of national saving: An extreme bounds analysis. Review of World Economics 137, no. 1, 150-174.

Imran, Z. M., Abbas, M., & Bashir, F. (2010). Short-run and long-run saving behavior in Pakistan: An empirical investigation. *Eurojournals*.

Irshad, A., Ibrahim & Owais, A. (2014). Effect of tax revenue on the national saving of Pakistan. *International Journal of Economics*.

Jappelli, T., and Pagano, M. (1994). Savings, Growth and Liquidity Constraints, *Quarterly Journal of Economics*, 109: 83-109.

Jilani, S., Sheikh, S., Cheema, F., Shaik, A. (2013). Determinants of National Savings in Pakistan: an exploratory study. *Asian Social Science*, Vol. 9, No. 5.

Julien, B. and Etienne, O. (2018). The Great Regression. Machine Learning, Econometrics, and the Future of Quantitative Social Sciences. *Revue française de sociologie, Centre National de la Recherche Scientifique*, 2018/3 (59), pp.475-506.

Kauffmann, A. (2005). Structural Change and Economic Dynamics in Transition Economies. 101-115. 10.1007/3-540-28526-1_8.

Khan, A. & Sarker, S. (2016). Determinants of GDS in Bangladesh: Evidence from Time Series Analysis.

Khan, M. & Teng, J. & Khan, M. & Jadoon, A. & Rehan, M. (2017). Factors Affecting the Rate of Gross Domestic Saving in Different Countries. *European Academic Research*. 5. 4261-4291.

Kim, Myeong Hwan. (2010). The Determinants of Personal Saving in the US. *Journal of Applied Business Research (JABR*) 26, no. 5.

Komicha, Hussien, H. (2007). Farm household economic behavior in imperfect financial markets. Vol. 2007, no. 78.

Kuhn, M., & Johnson, K. (2018). Applied Predictive Modeling. Springer.

Kulikov, Dmitry, Annika, P., and Karsten, S. (2007). A Micro econometric Analysis of Household Saving in Estonia: Income, Wealth, and Financial Exposure.

Loayza, N., Klaus, S, and Luis, S. (2000). What drives private savings across the world? *The Review of Economics and Statistics* 82, no. 2 165-181.

Lucas, Robert E. (1988). On the mechanics of economic development. *Journal of monetary economics 22,* no. 1 3-42.

McCarthy, J., & Feigenbaum, E. (1990). In Memoriam Arthur Samuel: Pioneer in Machine Learning. AI Magazine, 11 (3).

McKinnon, R. (2010). Money and capital in economic development. Brookings Institution Press.

Metin Ozcan, K., Gunay, A. & Ertac, S. (2003). Determinants of private savings behavior in Turkey. Applied Economics, 35:12, 1405-1416. DOI: 10.1080/0003684032000100373.

Mishi, S. (2012). Trends and determinants of household saving in South Africa. Economic Affairs.

Modigliani, F. (1970). The life cycle hypothesis of saving and intercountry differences in the saving ratio. Induction, growth, and trade: 197-225.

Modigliani, F. (1986). Life cycle, individual thrift, and the wealth of nations. *The American Economic Review* 76, no. 3, 297-313.

Modigliani, F., and Brumberg, A. (1983). Determinants of private savings with special reference to the role of social security, cross-country tests. In: The Determinants of National Savings and Wealth (eds Modigliani, F. and R. Hemming), pp. 24–55. St. Martins Press, New York.

Modigliani, F., and Cao, S. L. (2004). The Chinese savings puzzle and the life-cycle hypothesis. *Journal of Economic Literature*, 42, pp. 145–70.

Muradoglu, Gulnur, and Fatma, T. (1996). Differences in household savings behavior: evidence from industrial and developing countries. *The Developing Economies* 34, no. 2 (1996): 138-153.

Mu-Yen, C. (2011). Predicting corporate financial distress based on the integration of decision tree classification and logistic regression. *Expert Systems with Applications*, Volume 38, Issue 9, Pages 11261-11272.

Nagi, B. & Kostoglou, V. (2010). The Role of Savings in the Economic Development of the Republic of Azerbaijan. *International Journal of Economic Sciences and Applied Research*. 3.

Narayan, P. & Narayan, S. (2006). Savings behavior in Fiji: An empirical assessment using the ARDL approach to cointegration. *International Journal of Social Economics*.

Narayan, P., and Saud, A. L. (2005). An empirical investigation of the determinants of Oman's national savings. *Economics Bulletin* 3, no. 51: 1-7.

Newman, C., Tarp, F., Van, K., Quang, C., and Khai, L. D. (2008). Household savings in Vietnam: Insights from a 2006 rural household survey. *Vietnam Economic Management Review*, Vol 3, No 1.

Ng, A. (2015). Linear regression with one variable - model representation. Coursera Stanford Machine Learning course lecture 6. [www document]. [Accessed 10 Feb, 2019]. Available https://d396qusza40orc.cloudfront.net/ml/docs/slides/Lecture2.pdf

Ng, A. (2015). Regularization - the problem of overfitting. Coursera Stanford Machine Learning course lecture 7. [www document]. [Accessed 10 Feb, 2019]. Available https://d396qusza40orc.cloudfront.net/ml/docs/slides/Lecture7.pdf.

OECD iLibrary. [www document]. [Accessed 21 Jan, 2020]. Available https://www.oecd-ilibrary.org/finance-and-investment/short-term-interest-rates/indicator/english_2cc37d77-en

OECD iLibrary. [www document]. [Accessed 21 Jan, 2020]. Available https://www.oecd-ilibrary.org/finance-and-investment/long-term-interest-rates/indicator/english_662d712c-en

Oladipo, O. (2010). Does saving matter for growth in developing countries? The case of a small open economy. *International business & economics research journal (IBER)*.

Park, D., and Shin, K. (2009). Saving, Investment, and Current Account Surplus in Developing Asia, *Economics Working Paper* Series 158, Asian Development Bank.

Paxson, H., and Angus, D. (1993). Saving, growth, and aging in Taiwan. National Bureau of Economic Research.

Rehman, H., Faridi, M. & Bashir, F. (2010). Households Saving Behavior in Pakistan: A Case of Multan District. *Pakistan Journal of Social Sciences PJSS*. 30. 17-29.

Richardson, A., Thomas van, F.M. & Vehbi, T. (2018), Nowcasting New Zealand GDP using machine learning algorithms, Federal Reserve Bank of St Louis, St. Louis.

Samantaraya, A., and Patra, S. (2014). Determinants of Household Savings in India: An Empirical Analysis Using ARDL Approach. Economics Research International.

Samuel, A. (1953). Computing bit by bit or digital computers made easy. Proceedings of the IRE, 10(41):1223– 1230.

Singh, T. (2009). Does Domestic Saving Cause Economic Growth? A Time-Series Evidence from India, *Journal of Policy Modeling*, Volume 32, Issue 2.

Solow, R. M. (1956). A Contribution to the theory of Economic Growth, *Quarterly Journal of Economics*, Nr 70.

Song, Y. & Lu, Y. (2015). Decision tree methods: Applications for classification and Prediction. Shanghai Archives of Psychiatry, 27 (2), pp. 130.

Teng, Jian Z. & Khan, M. & Rehan, M. & Abasimi, I. (2018). Determinants of GDS: An Evidence from Asian Countries. *Journal of Business Management and Economic Research*. 2. 1-13. 10.29226/TR1001.2018.66.

Teshome, G., Belay, K., Bezabih, E., and Jema, H. (2014). Saving patterns of rural households in the zone of Oromia National Regional State, Ethiopia. *Journal of Development and Agricultural Economics* 6, no. 4 (2014): 177-183.

The Findicator website. [www document]. [Accessed 21 Jan 2020]. Available https://findikaattori.fi/en/115

The World Bank Data Catalog. [www document]. [Accessed 15 August 2019]. Available https://datacatalog.worldbank.org/gross-domestic-savings-current-us-1

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 267-288.

Trevor, H., Robert, T., & JH, F. (2009). The elements of statistical learning: data mining, inference, and prediction. New York.

Trustorff, J., Konrad, P., & Leker, J. (2011). Credit risk prediction using support vector machines. *Review of Quantitative Finance and Accounting*, 36(4), 565-581.

Vapnik, V., & Chapelle, O. (2000). Bounds on error expectation for support vector machines. *Neural Computation*, 12(9), 2013–2036.

Turner, G., and Richardson, P. (1998). The Macroeconomic Implications of Ageing in a Global Context, OECD Economics Working Paper No. 193.

United Nations Statistics Division. [www document]. [Accessed 21 Jan, 2020]. Available https://unstats.un.org/UNSD/nationalaccount/glossresults.asp?gID=231

Vincelette, Gallina A. (2006). Determinants of saving in Pakistan. South Asia Region PREM working paper series; no. SASPR-10. Washington, DC: World Bank.

Wang, S., Luo, J. Jussa, A. Wang, G. Rohal, and D. Elledge December 2014. Signal processing: The rise of the machines. *Deutsche Bank Quantitative Strategy*.

Weller, E. and Manita, R. (2010) Progressive tax policy and economic stability. *Journal of Economic Issues 44, no. 3*: 629-659.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*: Series B (Statistical Methodology), 67 (2).

APPENDICES

Appendix 1

Summary of literature review of gross domestic saving's determinants (full table)

Study	Data		Methods		Dependent variables	Independent variables
	Countries	Period	Model and statistical techniques	Analysis result		
Metin Ozcan et al. (2003)	Turkey	1968-1994	• OLS estimation	• Household saving in Turkey is affected dramatically by macroeconomic stability, external factors, the development of the financial market, life expectancy and economic crisis	• Private savings	• income variable • public savings • the ratio of M2 to gross national product (GNP) • current account deficit and the terms of trade • inflation rate • the real interest rate on saving deposits • demographic factors such as youth dependency ratio, urbanization ratio, life expectancy ratio, and old dependency ratio
Singh, 2009	India	Annual data from 1950–1951 to 2001–2002	• Optimal DOLS, FMOLS and NLLS estimates • Standard OLSEG estimates: the OLS-based two-step cointegration estimator of Engle and Granger (1987) (OLSEG) • The estimation of conditional error-correction model (ECM) as a result of autoregressive distributed lag (ARDL) model • ML system estimates: Vector autoregressive (VAR) • Monte Carlo simulations	• The bidirectional causality between savings and growth • The significant long-run impact of savings on income • The positive effects of interest on savings • The decrease in inflation likely improves savings by encouraging the real interest rate • The interest rate has an ambiguous impact on savings • The increase in productivity could also contribute to savings	• Growth (GDP) • Income	GDS
Jilani et al., 2013	Pakistan	Annual data 1973-2011	• Augmented Dickey-Fuller (ADF) test • Johansen Co-integration test	• Savings and Growth have the same trend • Significant and beneficial	National savings	• inflation • GDP • interest rate • fiscal deficit

			• Regression estimation • Error correction model (ECM)	association between fiscal deficit and national savings • Negative and significant association of inflation and nationwide savings • The interest rate has a harmful and insignificant influence on national savings • Independent variables, i.e., GDP, inflation, fiscal deficit and rate of interest possess long term equilibrium with nationwide savings		• the ratio of age dependency
Khan et al., 2017	Pakistan, China, Singapore, Japan Turkey, and Russia	1995-2016	• Fixed effects model • Random Effect Model • Regression Model	• The age dependency ratio, FDI and inflation is related negatively to GDS • GDP, per capita income, and money supply (M2) have a constructive effect on GDS	Gross domestic saving	• age dependency ratio • foreign direct investment (FDI) • money supply (M2) • inflation • per capita income • GDP
Hammad et al., 2010	Malay	1978 to 2007	Error correction model (ECM)	• per capita income has an unsupportive effect on the savings rate • government fiscal balance, per capita income and young age dependency have a considerable impact on national saving in a short term • Inflation rate brings a negative relation in short-run	Gross domestic saving	• government fiscal balance • per capita income • inflation • rate of return on savings deposit • age dependency ratio
Narayan et al., 2006	Fiji	1968-2000	ARDL method	Real Interest rate, per capita income, the ratio of age dependency have a positive influence on the savings rate	Aggregated savings	• interest rate • the deficit of current account, • dependency ratio of age
Imran et al. (2010)	Pakistan	Annual data from 1972 to 2008	Error Correction Model	• interest rate is a crucial determinant of national savings • Government consumption has a considerable and positive effect on nationwide savings	National savings	• consumer price inflation • public loans, • interest rates, • government consumption • remittances

Vincelette, (2006)	Pakistan	1973-2005	OLS regression	a cynical and essential connection among the development of the financial sector and aggregate savings	Rate of savings	• the income of commercial development, rate of interest, monetary policy and factors of demography
Khan & Sarker, 2016	Banglades h	1983 to 2013	Vector error correction model	• in the long run, deposit interest rate, exports, price index, and gross domestic income have a considerable causality with GDS • in short-run, total household income and the deposit interest rate has a causal connection with GDS	GDS	• Inflation rate • gross domestic income • exports • deposit interest rate
Mishi (2012)	South Africa	from the year 1963 and 2011	VECM method	Public sector savings, level of income, uncertainty (expected inflation), and financial development affect significantly on household saving rate	Household saving	• public saving • real disposable income growth (GDP per capita) • interest rate • percentage of household saving on household disposable income • the ratio of M2 to GDP
Chaudhry et al. (2015)	Pakistan	1972-2008	Johansson Cointegration technique vector error correction model (VECM)	• the interest rate has a constructive influence on saving in short-run • exports value, interest rates, government expenditure, and inflation are related significantly and beneficially while public loans are harmful related to saving rates	National savings	• government spending • workers remittance • inflation • public loans • interest rate • exports
Horioka and Hagiwara, 2010	12 economies in developing Asia	during 1966–2007	Fixed effects model and a random-effects model with robust standard errors	• GDP-related variables have an unsupportive and significant coefficient with DSR • Inflation rate & interest rates: not significant coefficient	Domestic Saving Rates	• Age • Age dependency ratio • per capita real GDP • CREDIT • the growth rate of per capita • real GDP • the inflation rates • the real interest rate • government expenditure on social services and pensions
Botha et al., 2011	South Africa	1981Q1 to 2009Q4	Vector error-correction model (VECM).	• Short-term interest rates have a negative and	Saving rates	• short-term interest rates

				considerable impact on household savings.		
Houérou, 2011	Turkey	1975-2008	Time series analysis OLS regression	• Income, interest rate, youth dependency, inflation are significant determinants of private saving • Not possible to connect the effect of expenditure policies in the past to savings improvement. • Government spending on public services like education and health may have an unsupportive effect on private saving. • To select productive public investment is essential since it might create extra public saving and increase domestic saving • The highly positive association between employment rate and household saving rates • A constructive and considerable relationship between household saving and domestic savings	• Private saving • Public saving • Household saving • Corporate saving	• Interest rate • Gross private disposable income • The young age dependency ratio • Inflation rate • Public Investment Expenditure • Employment rate

Appendix 2

The correlation matrix of dependent and independent variables during the period from 1995 to 2019 in Finland

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Y	1.0000										
X_1	0.8833	1.0000									
X_2	-0.5973	-0.7283	1.0000								
X_3	-0.7573	-0.8250	0.9724	1.0000							
X_4	0.9001	0.9245	-0.6717	-0.7874	1.0000						
X_5	0.1232	0.3214	-0.7649	-0.6512	0.2598	1.0000					
X_6	-0.6135	-0.7186	0.8945	0.8957	-0.6233	-0.6487	1.0000				
X_7	0.9219	0.7976	-0.4185	-0.5857	0.8049	-0.1171	-0.4150	1.0000			
X_8	0.7920	0.7789	-0.7615	-0.8358	0.7744	0.4168	-0.7694	0.6451	1.0000		
X_9	0.6297	0.7422	-0.9044	-0.9064	0.7116	0.6523	-0.8396	0.4552	0.9010	1.0000	
X_{10}	-0.7757	-0.8510	0.9355	0.9703	-0.8214	-0.5586	0.8736	-0.6299	-0.8528	-0.9100	1.0000

Appendix 3

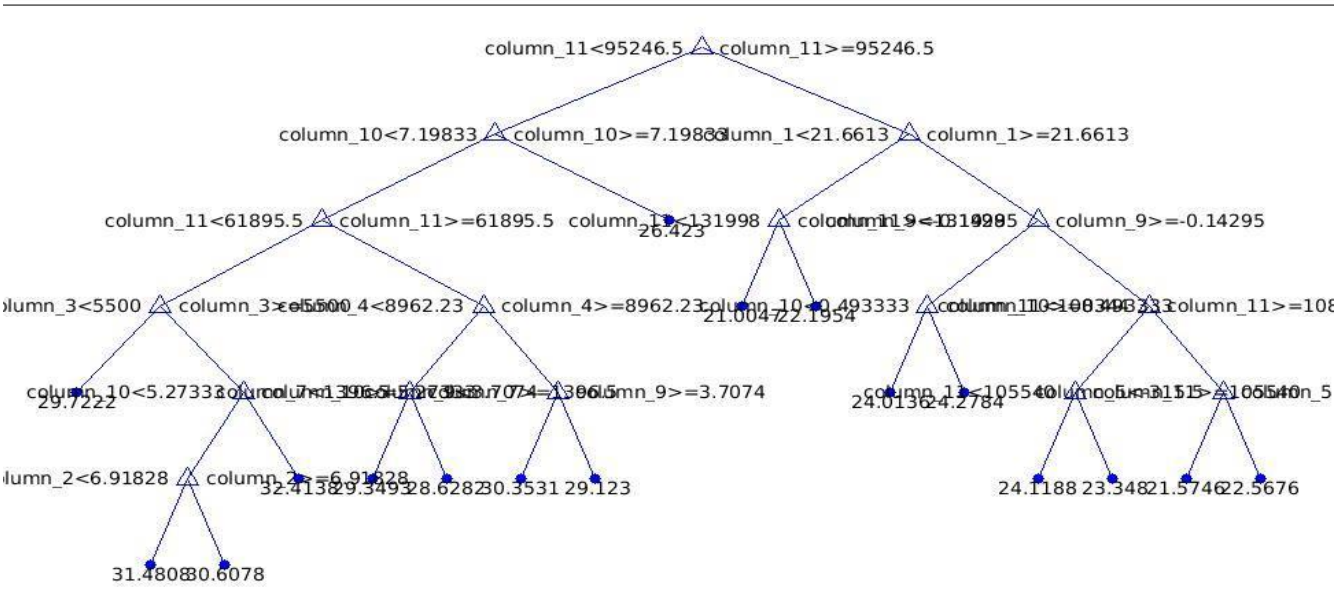
Model estimation

3.1 In-sample analysis

Linear Model

	Estimate	SE	tStat	pValue
(Intercept)	15.05	6.9944	2.1518	0.034283
GDS	0.58601	0.26027	2.2516	0.026958
Current Account	0.013623	0.13128	0.10377	0.9176
GDP per capita	0.00049212	0.00099668	0.49376	0.62276
Government spending	-0.00080685	0.00069576	-1.1597	0.24947
Net trade	0.0001373	0.00031537	0.43536	0.66442
Unemployment rate	0.28961	0.14972	1.9344	0.056435
Public investment	0.00022769	0.00074538	0.30547	0.76076
Household savings rate	0.12411	0.0941	1.3189	0.19079
Short-term interest rates	-0.75305	0.27783	-2.7105	0.008144
Long term interest rates	-0.28506	0.29079	-0.9803	0.32976
Money supply	-2.2119e-05	2.1742e-05	-1.0173	0.31192

Decision Tree Regression



3.2 Out-of-sample Analysis

Linear Model

	Estimate	SE	tStat	pValue
(Intercept)	33.564	11.302	2.9697	0.0044403
GDS	0.24139	0.37399	0.64546	0.52136
Current Account	-0.029787	0.16642	-0.17899	0.85861
GDP per capita	0.00071652	0.0012273	0.58384	0.56176
Government spending	-0.00041879	0.0010172	-0.41172	0.68217
Net trade	8.1948e-05	0.0004512	0.18162	0.85656
Unemployment rate	-0.21471	0.2196	-0.97776	0.33255
Public investment	-0.00055636	0.0013117	-0.42414	0.67315
Household savings rate	0.087581	0.12163	0.72008	0.47458
Short-term interest rates	-0.5434	0.42496	-1.2787	0.20647
Long term interest rates	-0.15547	0.49562	-0.31369	0.75496
Money supply	-0.00013216	4.3119e-05	-3.0652	0.0033931

Number of observations: 66, Error degrees of freedom: 54
Root Mean Squared Error: 1.41
R-squared: 0.834, Adjusted R-Squared: 0.8
F-statistic vs. constant model: 24.6, p-value = 3.42e-17

Decision tree regression model