UNIVERSITY OF TARTU

Faculty of Social Sciences

Johan Skytte Institute of Political Studies

Kaarel Kaasla

FORECASTING THE PARTY SUPPORT IN ESTONIA: COMPARISON OF MACHINE LEARNING REGRESSION ALGORITHMS

MA Thesis

Supervisor: Mihkel Solvak, PhD
Co-supervisor: Kaspar Märtens, MsC

Tartu 2018

I have written this Master's thesis independently. All viewpoints of other authors, literary sources and data from elsewhere used for writing this thesis have been referenced.

....

/ Kaarel Kaasla /

The defence will take place on .... / date / at .... / time / ..../ address / in auditorium number .... / number /

Opponent .... / name / (.............................. / academic degree /), ... / position /

Abstract

Forecasting political behavior using economic indicators is not a very new phenomenon with the earliest literature going back as far as the 1930s. In the present day, there exists a lot of research on the topic, but the majority of these studies have been conducted in the context of a very limited number of countries such as the United States or the Western European ones. By comparison, the research on forecasting the political behavior using economic voting in Estonia is almost non-existent. This thesis will be the first in-depth study conducted at that level and forecasts the party support of the Estonian Reform Party and the Estonian Center Party using economic indicators as the predictor variables. Based on the previous economic voting theory, it has been argued that the theoretically correct model to forecast using these variables is the linear regression due to the expected associations between the economic variables and party support. However, this thesis contests this claim and argues that when analyzing the phenomena of forecasting party support using economic indicators, certain modern machine learning algorithms could be considered as legitimate alternatives to the linear regression, as each of them addresses the different shortcomings of the model. For this reason, this thesis compares the methods of linear regression, regularized linear models, autoregressive integrated moving average, and the decision-tree models to see whether the more modern approaches are able to improve upon the default linear regression model.

Table of Contents

1 Introduction

There are numerous different theories of how people vote, one of the most popular of them being the concept of economic voting. The basic premise of economic voting lies in the idea that the electorate continuously evaluates the performance of their governments and holds them accountable for economic outcomes, rewarding the incumbent party when the economy is moving in an upwards trend and punishing them when the economic indicators are declining. The practice of forecasting elections using macroeconomic indicators is very widespread in the Western countries, especially the United States and in the field of political science as a whole, but at the same time, the more serious country-level research in Estonia is scarce. This thesis will be the first in-depth study conducted at Estonian level that analyses the concept of forecasting party support of the Estonian Reform Party and the Estonian Center Party using time-series aggregated economic indicators and voting data. The reasoning for choosing these two particular parties is two-fold. Firstly, the Reform Party has been the incumbent party of the Estonian government for the whole duration of the data used in the analysis and, on the contrary, the Center Party has been the non-incumbent the whole time. For this reason, it is possible to automatically view these parties in their respective roles when it comes to interpreting the economic effects on their support. The second reasoning is that these two parties have been the most stable ones throughout the last decade and the party support data sample is the largest. Even though this thesis analyzes only these two parties, its broader goal is to find a more universally applicable model that could also be used to forecast the support of other parties in the system.

The second side of the research question deals with the comparison of machine learning regression algorithms. In past, the research has generally viewed a mechanism-based linear regression model as the "theoretically correct" method. However, this thesis contests this argument on the grounds of linear regression having multiple shortfalls that might make it a sub-optimal choice to model party support using macroeconomic variables. For this reason, in addition to the linear regression, this model also models the party support using regularized linear models which are able to perform variable selection, autoregressive integrated moving average model which takes into consideration the time-series specifics of the data, and also the decision-tree models which offer a non-parametric approach with the simple objective of maximizing the prediction accuracy at the cost of model interpretability. The results of all the different models will be compared to the linear regression see whether any of them is able to provide a significant improvement.

The general structure of the main body of the thesis is as follows: theoretical background, research design, results, and discussion. More specifically, the theoretical background section of the work first gives the historical background of the concept of economic voting, theoretical foundations of the economic voting indicators, and also outlines some of the challenges that the economic voting theory suffers from. The theory chapter also gives an overview of how to actually forecast party support using economic indicators and different machine learning methods, meaning that what are the theoretical relationships between different economic variables and party support and additionally, how exactly do all of the machine learning algorithms used in this work differ from one another and how each of them could be expected to improve upon the linear regression model. The third part of the theoretical background chapter gives a theoretical understanding of how these different machine learning methods work, starting with linear regression and moving onto models that build upon it. Theoretical understanding of different methods is necessary to really grasp the differences between each model and what makes them unique from one another. The research design part of the thesis will firstly outline the sources of the data that is used in the analysis, provides the exploratory data analysis of both the party support over time and the economic indicators used in the analysis, and also how the model accuracy is assessed. The last part of the research design is dedicated to optimizing parameters of different machine learning models. Almost all of them have parameters that must be estimated and it is non-trivial in order to achieve the optimal performance for each model. The results part of the thesis offers the results and brief comments on the different models for both parties. Lastly, the discussion part of the thesis takes a more general approach and interprets the results in a wider context, outlining what was expected and unexpected and providing explanations for the latter.

2 Theoretical Background

2.1 Theory of Economic Voting

2.1.1 Historical Background of Economic Voting

The concept of ’economic voting’ is mainly grounded in the ideas of issue and reasoning voting by focusing only on the issue of the economy (Lewis-Beck & Stegmaier 2013). From the theoretical view, it means that electorate evaluates the performance of governments and holds them accountable for economic outcomes by either rewarding or punishing accordingly. The action of holding government accountable is achieved through prospective voting when the performance is perceived as favorable and by not voting when an alternative offers better expectation. For this reason, support of incumbent increases when the economy is improving and decreases when economic conditions in a society worsen.

Even though there are works exploring the relationship between economic and electoral performance going back as far as the 1930s (Tibbits 1931), it was not until second half of the 20th century that the field became a major focus in the field of political science. While some theoretical foundations of economic voting could be viewed to come from classical voting behavior, it is mainly based on the rational choice theory. The rational choice model essentially views electorate as strategic utility maximizers who make their decisions related to electoral action based on what they stand to gain or lose and always choosing the option that is expected to benefit them the most and at the same time cost them the least (Evans 2004). The shift from viewing economic voting through classical voting theories to rational choice theory did not happen overnight and many of the earlier works on the topic intertwined the two approaches. This can be seen in works such as Anthony Downs’s 1957 article which was one of the first articles to approached the process of prospective voting from the perspective of rational choice. Downs’s work postulated that when making informed political decisions, the electorate does not only use the past performance but also use this past behavior to make predictions about future and subsequently make decisions based on the expected outcome (Downs 1957). Similarly, Campbell et al. empirically demonstrated in their work how voter attitudes and different evaluations, including the economic assessment, are a part of the electorate’s decision making (Campbell et al. 1960). Valdimer Key took this idea of voters evaluating the performance of political parties and expanded the argument to retrospective voting or reward-punishment theory. He contended that electorate also evaluates past performance of government and rewards them on the basis of what they deliver or punishes for falling short of the promises (Key 1966). The aforementioned works could be viewed as a mixture of classical approaches and the rational choice theory, laying important groundwork for further developments in the field of studying economic voting.

Economic voting as a field of research attracted attention in the 1970s which saw a number of path-breaking works on the topic. These include articles such as Kramer's which analyzed aggregate-level economic voting in the United States congressional elections and showed that incumbent party support is related to the national economic performance; namely that improvements in macroeconomic indicators such as average income resulted in increased lower chamber of the Congress vote share for the incumbent president's party and against it when macroeconomic indicators moved the other way (Kramer 1971). Other similar works published around that time, which used aggregate-level data, arrived at similar conclusions. This could be exemplified by Edward Tufte's 1978 article which showed similar to Kramer's conclusions that greater the growth in real disposable income, the better the president's party expected to perform in the upcoming elections and further solidifying the link between economic indicators and electoral support (Tufte 1978). Around the same time Morris Fiorina was one of the first to shift the analysis from macro-level to individual level. Drawing inspiration from Key's 1966 article, Fiorina established the theory of retrospective economic voting, arguing that economy is the primary issue that voters evaluate in national elections and do so by evaluating the government's past economic performance (Fiorina 1981). However, at the same time Fiorina also recognized that electorate could act prospectively as proposed by Downs in 1957. To this day there is no clear answer of whether economic voting should be viewed as a retrospective or prospective process, although the dominant belief is that voters mainly make their decisions based on the past economic performance rather than trying to predict the future (Lewis-Beck & Stegmaier 2013). It is also worth mentioning Roderick Kiewlet's 1983 study which added additional important dimension to the theory of economic voting. His work showed that the economic voting could be sociotropic, that is based on the state of the national economy or pocketbook, which could be viewed as an evaluation of personal financial circumstances. Kiewlet's analysis found strong conclusions that economic evaluations dominate personal finances in voters' electoral decision-making and additionally that economic voting is generally incumbency-oriented (Kiewlet 1983). Basically all of the aforementioned early work on economic voting was the United States-centric and it was not until 1988 when Michael Lewis-Beck published his study which took the same concepts and applied them to national surveys in Western Europe. His work concluded that similar to the

United States, both retrospective and prospective economic voting is evident in European nations and furthermore, it is also primarily sociotropic (Lewis-Beck 1988). Ever since then, the field of research is not any more the United States-centric and there is a vast corpus of articles and book on the subject available. However, at the same time, most of that research has been carried out in the Western Europe, focusing on countries such as the United Kingdom, France, Germany, and Italy (Lewis-Beck 1986) and to a lesser degree countries such as the Netherlands (van der Eijk & Niemöller 1987) and Denmark (Nannestad & Paldam 1997). In many other instances, such as Estonia, the country-level research on the topic of economic voting is at best scarce or sometimes non-existent. Even though there does exist some rudimentary research in the Estonian case (Solvak 2015), this thesis will be the first in-depth study at Estonian level that explores forecasting party support using the economic voting indicators as independent variables.

2.1.2 Theoretical Background of Economic Voting Indicators

Most of the early work in the field of economic voting was done examining popularity functions which proposed that government's popularity is determined by macroeconomic conditions and political controls (Lewis-Beck & Stegmaier 2013). Because of the limitations in data availability, most of the early research on the subject of economic voting was done using aggregated time-series data. In the early studies of popularity functions, national-level macroeconomic indicators took the role of independent variables and were analyzed to see how these relate to government support or electoral outcomes (Goodhart & Bhansali 1970; Mueller 1973). Goodhart and Bhansali's article showed the relationship between government popularity and macroeconomic conditions such as unemployment and inflation rate. The article found that when the economy performs well, the electorate favors incumbent and when economy moves in a downward trend, the incumbent support suffers (Goodhart & Bhansali 1970), also known as the responsibility hypothesis (Nannestad & Paldam 1994). Mueller's 1973 work arrived at the similar conclusions with the addition that the impact of the economy is asymmetrical, meaning that the electorate is more likely to punish the incumbents for economic decline than reward them following prosperity (Mueller 1973). In addition to the aforementioned economic indicators influencing the support, the later studies such as Gary Jacobson's 1990 article found that change gross domestic product is also a significant factor (Jacobson 1990). Subsequent studies have also involved variables such as nonfarm payrolls, consumption expenditures, and stock market performance among others

(Silver 2012), but the links between these indicators are generally weaker than the 'big three' and more often than not localized to only certain cases. Similarly, because of their proven universal applicability, unemployment, inflation, and changes in the gross domestic product are the only economic variables used in this present thesis. In addition, studies have found that electorate tends to have a short time horizon when evaluating economic performance and the effects decay fast (Nannestad & Paldam 1994), so it is necessary to include as short lag structure as possible. Seeing as the aggregate macroeconomic data is generally collected monthly or quarterly, the most reasonable course of action would be to use the lag value of $t - 1$ as it minimizes the decay that occurs. While it is not an economic indicator, the models generally also include some kind of seasonality or temporal dimension. Previous studies have argued that retrospective and prospective decision-making do not stay constant throughout time, but rather are influenced by the election cycle (Singer & Carlin 2013). The findings show that if an election cycle starts right after the election and ends right before the next one, at the start of the cycle the support based on economic voting is mainly based on prospective decision-making. This is so because there is a 'honeymoon' period and the electorate does not yet have sufficient information to evaluate the government's actions after the election. However, after the new incumbent's record mounts, the electorate starts to approach the decision-making from more retrospectively based on the events that have happened since the election, but as shown by previous theory, even though retrospective voting rises, prospective still reigns supreme with electorate comparing the incumbent's performance to the expectations for the future. At the end of a cycle, the electorate's focus shifts again to prospective voting as they do not think about what has happened in the past, but rather base their decisions on the expectations related to future. The previous research conducted on Estonian level has also shown that there exists a statistically significant link between the electoral cycle and party support (Solvak 2015).

Inflation signifies an increase of goods and services in an economy over time, meaning that when price level rises, each unit of currency can buy less. Inflation is generally measured using inflation rate which is the annualized percentage change in a general price index such as the consumer price index. The opposite of inflation is deflation, meaning a decrease in the price of goods and services and occurs when the inflation rate falls below 0 %. The best case scenario for the incumbent party is a situation with moderately low inflation because some inflation in an economy is a sign of health. However, on the contrary, too large or negative rate could be interpreted as a negative aspect and therefore it makes sense for the electorate to punish the incumbent party. Economic growth could be viewed as the inflation-adjusted value of goods and services and is measured as the percent rate of increase in real gross domestic product. The relationship between change in gross domestic product and party support could be argued to be linear, meaning that when economic growth is larger, the economy prospers and the support for incumbent party increases. However, when the change in gross domestic product is low or negative, it can be assumed that electorate support for the incumbent party will decrease and voters will start looking for other options in the form of non-incumbent parties. Lastly, unemployment shows the percentage of unemployed workers in the total labor force. Moderately low unemployment is common in societies and a small change in the environment where unemployment is already low could be argued to have a small effect. However, when unemployment increases over some critical limit, the effect it has on the party support changes more radically than with lower unemployment. For that reason, it could be said that the relationship between unemployment and party support is non-linear and asymmetrical.

2.1.3 The Challenges of Economic Voting Theory

Even though the economic voting research published since the 1970s shows a clear link between economic indicators and political support, the field also faces some considerable challenges. One of the main concerns the economic voting theory faces is that the economic effects tend to be conditional and not universal (Anson & Hellwig 2015). For example, most of the research done on the subject has been bounded to political systems such as the United States, the United Kingdom, or France and even though there does exist theoretical overlap, there are also differences among cases which must be considered when conducting an analysis. For this reason, there does not exist a universal theory of economic voting, but rather each country or case should be viewed in a vacuum. More so, the differences in results do not only become prevalent between different countries, but the research has shown that the lack of stability, also known as 'instability dilemma' (Paldam 1991) caused by imprecise modeling, can also happen in a single country over time (Lewis-Beck & Paldam 2000). The authors have suggested that in order to sufficiently account for potential instability in different countries, institutional conditions such as the party system type must be considered in the model (Lewis-Beck & Paldam 2000), but so far, the relevant evidence remains thin. Other authors have argued that the instability can also be caused by of how the dependent variable is operationalized in different studies. Van der Brug et al. believed that in different party system types such as single-party or multi-party systems, the electorate's decision-making process differs because the competition between parties and their alternatives is also different (van der Brug et al. 2007). There is no clear consensus on what is the actual root of this lack of stability among countries and sometimes in a single country over time, but the aspect that the theory is not universal and each case should be viewed separately is something that must be kept in mind when researching economic voting.

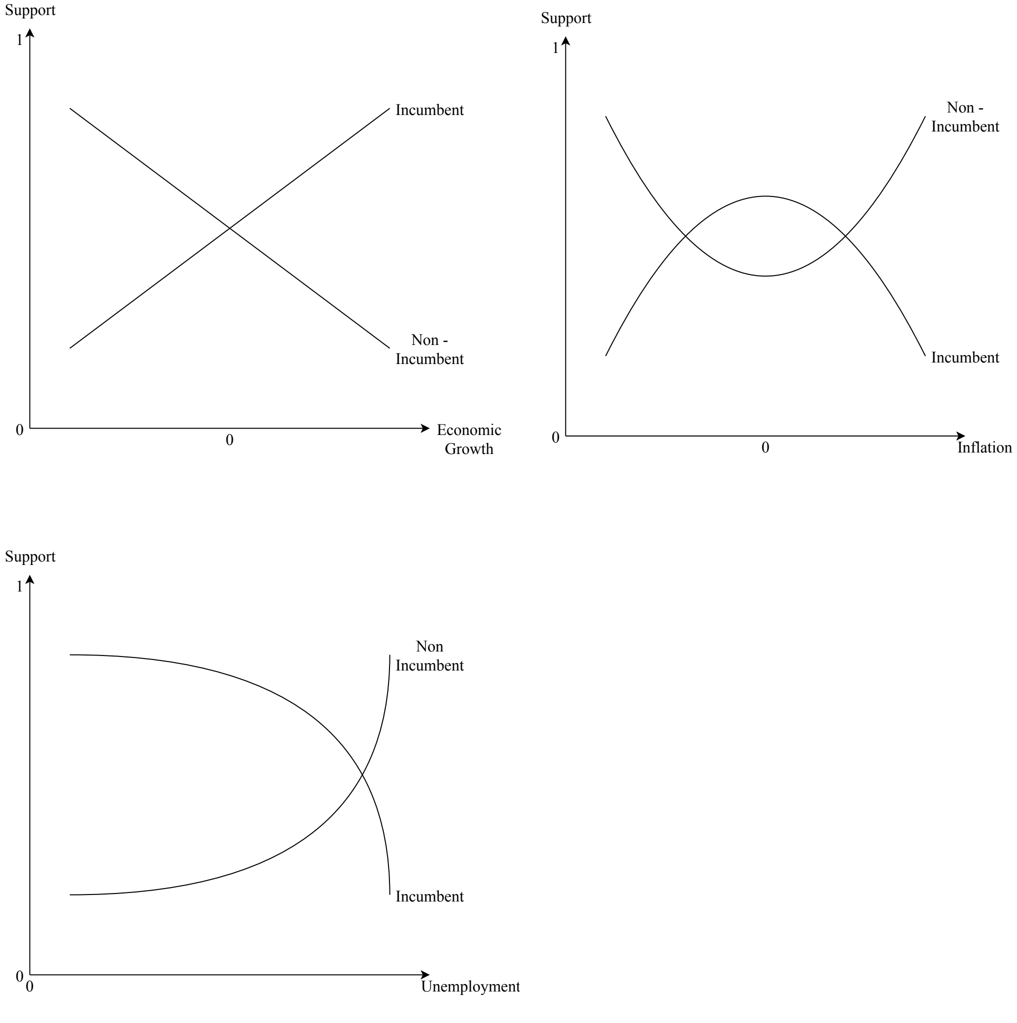
Economic voting also makes the assumption of homogeneity; that the electorate responds to economic stimulus uniformly (Lewis-Beck & Stegmaier 2013). However, in reality, that is not the case and certain groups of the population have their vote probability affected more than others. For example, previous research has found that women may have the different response than men to the economic changes (Welch & Hibbing 1992). Similarly, Duch et al. found that voters with differing levels of voter experience and information exposure can also vary in their response because of being able to assess the quality of economy differently (Duch et al. 2000) and also that the level of political trust can alter perception about national economic conditions (Duch 2001). These studies show that there is some heterogeneity in the economic voting, but it is not clear to what extent. Lewis-Beck and Stegmaier suggest that while heterogeneity exists, with properly controlling for the relevant variables in a correctly specified model of the vote, it is still possible to make meaningful generalizations about the economic effects, meaning that while heterogeneity introduces some noise to the analysis, the main signal still comes from the homogeneous effect (Lewis-Beck & Stegmaier 2013).

Finally, the economic voting also has the concern of endogeneity of economic evaluations. The economy, as shown by a number of studies published on the subject, has a statistically significant effect on the election result (Lewis-Beck & Stegmaier 2013). However, this perspective has received opposition from a perspective which argues that these economic effects are exaggerated, mainly due to the endogeneity problem. Authors such as Kramer have proposed that the effects are inflated because the economic evaluations of voters can be shaped by their pre-existing political preferences (Kramer 1983). For example, the part of the electorate that already supports the incumbent government might view economic effects more positively, while the voters who support non-incumbent parties might be more negative in their evaluation of the same effects towards the incumbent party. For that reason, the differences in responses might not be linked to changes in economic conditions, but rather the subjective judgment of economic conditions (van der Brug et al. 2007). Some authors have claimed that the economic effects in the previous studies have been overstated (Wlezien et al. 1997; Evans & Andersen 2006; Anderson 2007) while the proponents of economic voting have addressed this critique by employing a method of variable exogenisation and have found persistent effects of sociotropic retrospective evaluation on the vote (Lewis-Beck et al. 2008; Fraile & Lewis-Beck 2014).

2.2 Forecasting Party Support Using Economic Indicators

Before modeling the party support using economic indicators, it is necessary to explore the expected relationships between the variables or how exactly is the economy supposed to influence the support. The expected relationships between the indicators and the support for incumbent/non-incumbent party are pictured below in figure 1.

Figure 1. The theoretically expected relationships between economic indicators and both incumbent/non-incumbent parties. Upper left: gross domestic product, upper right: inflation, lower left: unemployment.



As the figure shows, the relationship between change in gross domestic product and party support can be expected to be linear. It means that if the change in gross domestic product is positive, the support rating for incumbent party is also expected to go up at the approximately constant rate and on the contrary, if the change in the economic indicator is negative, the incumbent party support is expected to similarly decrease. For the inflation indicator, the relationship between the variable and party support is expected to be parabola-shaped. It means that for the incumbent party, the support is expected to be at its highest in cases where the inflation is around zero and decreases as inflation approaches either end. The reason for such relationship is that near-zero inflation is the ideal case scenario as some inflation in an economy is to be expected, but it is not supposed to be too high or too low. Very high or negative inflation in an economy is perceived as a problem, so it is logical that the electorate would view such inflation values as the fault of the incumbent party, which in turn leads to lower support. As for the unemployment, there is always some unemployment in every society and for this reason, a small initial increase in unemployment should not, at least theoretically, affect the incumbent party support by a comparable degree. Instead, it could be expected that as the unemployment level initially increases, the decrease in party support is relatively slow. However, as the unemployment level increases over some critical point, the electorate starts to view it as a real problem in the society and the party support starts to decrease more rapidly. It is worth noting that all of these relationships hold true for the incumbent party but not the non-incumbent. For the non-incumbent party, the relationships are expected to be exactly the opposite with low or negative gross domestic product, high inflation or deflation, and high unemployment being positively correlated with high support, as at these points, the non-incumbent becomes a viable alternative in the eyes of the electorate.

The past research on forecasting party support using economic indicators has viewed the linear regression with inflation and unemployment variables linearized as the "theoretically correct" or a mechanism-based model. The reason for viewing linear regression as the correct model, in this case, is that if the predictor variables are linear, the linear regression model is also expected to yield the best results. Similarly, viewing it as a mechanism-based model makes assumptions of how the economic indicators should theoretically influence the economic voting function as we have the theoretical understanding of the underlying associations. However, this thesis contests the notion that linear regression models are the correct way to forecast party support using economic variables, as there are aspects that the default linear model does not take into account, such as variable importance, time-series characteristics, and non-linear associations, all of which might considerably improve the prediction accuracy. This view places a premium on the forecast accuracy which could possibly come at the expense of the model interpretability. As social sciences rely heavily on explaining the causality between variables, the black-box models might not be the ideal option to understand different phenomena, but at the same time, if a theoretically uninformed model produces consistently better forecasts than a theoretically informed one, it raises a question if the theory under the latter model actually holds. To analyze this argument, the thesis will explore three alternative approaches to forecast party support, each of which addresses different shortcomings of the linear regression model: regularized linear models, autoregressive integrated moving average models, and the decision-tree models.

The regularized linear models are viewed as a viable alternative as these models are able to model data more flexibly and avoid over-fitting through variable selection. What variable selection does is that it estimates the importance of each indicator in the model and places greater importance on the ones that are more relevant to the model. Similarly, it shrinks the coefficients of the redundant variables, which do not contribute to the accuracy or might even decrease it, towards (or to exactly) zero. For this reason, in comparison with a default regression model, the regularized linear model allows for better prediction accuracy by being more flexible and in addition, less complex model due to variable selection. Another model which the thesis argues to have improvements over linear regression is the autoregressive integrated moving average model with external economic variables. The justification for considering this model is that since the party support and economic data is in the time-series format, the peculiarities of time-series must be also taken into account when forecasting. Lastly, the thesis also benchmarks the decision-tree methods against the linear regression model. For linear regression or any other parametric model, there is usually some underlying mechanism in the data, such as relationships between different variables, which allow to obtain the results. However, a disadvantage of the parametric approach is that the resulting model will almost never match the true unknown form of the function, leading to poor estimations. This problem can be potentially alleviated by choosing a non-parametric model such as the decision-trees which do not make any explicit assumptions about the functional form of the functions, but rather have the objective of seeking an estimate of the functions that gets as close to all data points as possible without over-fitting. Such approach can have a major advantage over linear regression or other similar models when the objective is to maximize the prediction accuracy since it is not bounded by the same limitations as the parametric models. For this reason, it is also necessary to analyze the data from the methodological perspective that the parametric methods are unable to do by their design.

In order to outline the distinctions that set all of the aforementioned models apart from each other, the thesis will also give an overview of the each of their individual theoretical backgrounds. This is necessary, as having a full understanding of the mechanisms of each model allows to understand how their individual differences affect the results they produce.

2.3 Theoretical Background of Machine Learning Algorithms

2.3.1 Linear Regression

The simplest and most straight-forward approach to supervised machine learning is the linear regression. Even though it may seem overly simplistic compared to some of the more modern approaches such as decision-trees, kernel methods, and neural networks, it can still be used to draw useful inferences off data. Additionally, linear regression can be seen as a starting point for more complicated methods, as many other models can be viewed as generalizations or extensions of it.

The main idea behind simple (single explanatory variable) linear regression is that it predicts a quantitative outcome *Y* on the basis of a variable *X*. As the name of the method suggests, linear regression assumes that there is approximately a linear relationship between *X* and *Y* (James et al. 2017: 61). Mathematically, this linear relationship between these two variables can be written as

$$Y = \beta_0 + \beta_1 X + \epsilon$$

and read as "regressing *Y* on *X*". Coefficients or parameters β_0 and β_1 are unknown constants that represent respectively intercept - that is, the expected value of *Y* when *X* = 0, and slope - the average increase in *Y* associated with a one-unit increase in *X*. ϵ is a mean-zero random error term used as a catch-all for what the model misses as the true relationship is generally not linear or there might be a measurement error. Using training data of *X* and *Y*, the estimates of the coefficients can be produced and furthermore used to predict unknown *Y* values, assuming that *X* values are known.

When it comes to finding the estimations of the β_0 and β_1, the goal is to find an intercept and slope values such that the resulting line is as "close" (meaning minimized distance) to all data points as possible; also known as the least squares method. In order to find this line, *X* and *Y* should be first viewed as *n* (denoting training set sample size) observation pairs

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

where each pair consists of a measurement of *X* and the *Y* value with the corresponding index. Plotting each of these pairs on a two-dimensional Cartesian coordinate system, a straight line with intercept β_0 and slope β_1 can be drawn through them. However, generally the data is scattered and not in a straight line, so the drawn line cannot directly go through all of the data points. As a second-best alternative, it is possible to find a line with some intercept and slope which minimizes the distance between data points and the said line. Mathematically, the prediction for *i*th value of *Y* is based on *i*th value of *X* and can be written as $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ (circumflex indicating predicted values). Using $\hat{y}_i$ value, it is possible to find the difference e_i (or residual) between the *i*th observed response value and the *i*th response value that is predicted by the linear model or using notation, $e_i = y_i - \hat{y}_i$. The minimum distance between all of the data points and the line passing them can thus be calculated by finding the residuals for all indexes, squaring the results (since the distance can be either positive or negative) and summing the squares together (*ibid.*: 62). This is also known as the residual sum of squares (RSS) and is defined as

$$RSS = e_1^2 + e_2^2 + ... + e_n^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Using RSS in the context of the least squares method means that the algorithm iterates over different combinations of $\hat{\beta}_0$ and $\hat{\beta}_1$ and chooses the pair which yields the smallest RSS value. However, in reality, brute-force approximation is not needed as the theory (*ibid.*) shows that the least squares coefficient estimates which minimize distance can be defined as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, or more simply, arithmetic means of the sample.

Simple linear regression approach generally works well when predicting a response on a single predictor variable. However, in practice variance in a variable is explained by more than one predictor. In these cases, fitting separate simple linear regression models would not be a feasible solution as each of the models would ignore the other predictor variables while in reality, there would exist a correlation between them which, in turn, influences the coefficient values. To alleviate this problem, instead of fitting a separate model for each predictor variable would be to extend the simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ so that it can accommodate multiple predictors. This can be accomplished by adding other predictor variables to the model and giving each of them a separate slope coefficient. Suppose that the model includes *p* distinct predictors, mathematically, the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$$

where X_j represents the *j*th predictor and β_j quantifies the association between that variable and the response. The each coefficient in the equation can thus be interpreted as the "average effect on *Y* of a one unit increase in X_j, holding all other predictors fixed" (*ibid.*: 72).

Similar to β_0 and β_1 in the simple linear regression model, the regression coefficients β_0, β_1, ... , β_p in a multiple linear regression model are unknown and must be estimated. Extending $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for the multiple predictor setting, the equation becomes

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$$

and the least squares approach noted before is used to estimate the coefficient values. This means the algorithm chooses β_0, β_1, ... , β_p to minimize the sum of squared residuals which can be written as

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - ... - \hat{\beta}_p x_{ip})^2.$$

where the values of $\hat{\beta}_0$, $\hat{\beta}_1$, ... , $\hat{\beta}_p$ that minimize the equation are the multiple least squares regression coefficient estimates (*ibid.*: 73).

2.3.2 Regularized Linear Methods

As mentioned, more often than not, real-world data does not follow a linear trend. At first glance, this aspect might make it seem like linear models are at a clear disadvantage in relation to non-linear approaches which fit a non-parametric function to data, but empirically, they are actually quite competitive. However, the least squares approach is not the only linear model and for this reason, alternative fitting procedures which extend and improve upon the linear framework should also be explored.

One of the ways to modify linear models is through regularization. Regularization, also known as shrinking, essentially fits the model involving all p predictor variables, but the estimated coefficients are shrunken towards zero (or estimated to be exactly zero) relative to the least squares estimates. Two main reasons to consider for doing so are that the alternative fitting procedures can yield better (1) prediction accuracy, (2) model interpretability (*ibid.*: 203).

For prediction accuracy, if the true relationship between the response and the predictions is approximately linear, the least squares estimates will have a low bias. Additionally, if $n \gg p$, where n is the number of observations and p the number of predictor variables the least squares models also tend to have low variance and will generally perform well on test data (*ibid.*). However, problems arise when n is not much larger than p as it leads to high variance in the fit which can result in over-fitting and thus poor predictions on future data. To alleviate the problem of increased variance as the difference between n and p decreases, regularization (or shrinking) of estimated coefficients can be used. Using this method allows for substantial reduction in variance at the cost of a negligible increase in bias, leading to improvements in model accuracy and therefore better performance on test data (*ibid.*: 204).

In terms of model interpretability, it is often the case that some of the independent variables used in a multiple linear regression model are weakly or not at all related to the response, leading to unnecessary complexity in the model (*ibid.*). Using regularization, it is possible to shrink (or remove) irrelevant variables in the model, leading to a model that could yield a better prediction accuracy or be more easily interpretable. The least squares approach by itself is extremely unlikely to yield and coefficient estimates that are exactly zero, so employing regularization (or alternatively variable selection) is required.

The two best-known methods for regularizing the regression coefficients towards zero are ridge regression and lasso (*ibid.*: 215, 219). The former is very similar to least squares, except that the coefficients are estimated by minimizing different quantity. To show the comparison, the least squares fitting procedure minimizes

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i_j})^2.$$

Ridge regression, on the other hand, could be written as

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i_j})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is a tuning parameter that must be determined separately (*ibid.*: 215). The similarity between the two models is that both seek coefficient values that fit the data well by minimizing RSS. However, the difference lies in the aspect that the ridge model includes the second term $\lambda \sum_j \beta_j^2$, called a shrinkage penalty (*ibid.*). The shrinkage penalty is small when $\beta_1, ..., \beta_p$ are close to zero, so it has the effect of shrinking the estimates of β_j towards zero. The tuning parameter λ controls the relative impact of these two terms on the regression coefficient estimates (*ibid.*). This means that when $\lambda = 0$, the penalty term has no effect and ridge regression will produce the same estimates as the least squares method. On the contrary, as $\lambda \to \infty$, the impact of shrinkage penalty grows, and the ridge regression estimates will approach zero (*ibid.*). Another difference between the two approaches is that while the least squares method produces only a single set of coefficient estimates, ridge regression generates a different set of coefficient estimates for each value of λ. For this reason, selecting a good λ value is critical from the model prediction accuracy point of view as different λ values produce different results. The exact process of estimating the optimal λ value for ridge regression will be further elaborated on later in the thesis.

Even though in many cases ridge regression can be viewed as an improvement over the least squares method, it still has one general disadvantage - including all p predictors in the final model. The shrinkage penalty $\lambda \sum_j \beta_j^2$ shrinks all the coefficients towards zero, but does not set any of them exactly to zero (except when $\lambda = \infty$) (*ibid.*: 219). While this may not be a problem of prediction accuracy, it can make a model more complex and difficult to interpret. The lasso (least absolute shrinkage and selection operator) could be regarded as one alternative to ridge regression and is able to overcome the disadvantage of including all predictors. Similar to the least squares and ridge regression, lasso coefficients minimize the quantity (*ibid.*)

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{i_j})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Comparing lasso with the ridge regression formulation, it can be seen that the only difference is that that β_j^2 term in the ridge regression penalty has been replaced by $|\beta_j|$ in the lasso penalty. In a statistical sense, the lasso uses an ℓ_1 penalty instead of an ℓ_2 penalty and "the ℓ_1 norm of a coefficient vector β is given by $\|\beta\|_1 = \sum|\beta_j|$" (*ibid.*). As is the case with ridge regression, the lasso model also shrinks the coefficient estimates towards zero. However, the difference is that the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly zero when the tuning parameter λ is sufficiently large so in addition to regularization it also performs variable selection (*ibid.*). As the result, the models generated by the lasso approach could theoretically be regarded to have greater prediction accuracy than the least squares method due to λ parameter and additionally, better interpretability than the models generated by ridge regression because of variable selection. Similar to ridge regression, the correct estimation of λ in a lasso model is of utmost importance and the process will be explored in-depth later on.

2.3.3 Autoregressive Integrated Moving Average Method

An autoregressive integrated moving average (ARIMA) model offers an alternative approach to time-series forecasting problems by seeking to describe the autocorrelations in the data. It is a generalization of an autoregressive moving average (ARMA) model and is commonly used in the cases where data is non-stationary or in other words, where the time series properties depend on the time at which the series is observed (Hyndman & Athanasopoulos 2018). The ARIMA model could be specified as three different parameters: AR(p), I(d), and MA(q) or (p, d, q).

The autoregressive component of ARIMA is referring to the notion that in an autoregressive model, the dependent variable is regressed on its own lagged values or could be viewed as a regression of the variable against itself. While in the multiple regression setting the predictions are made using a linear combination of predictor variables, the autoregressive model uses a linear combination of the past values of the variable to make predictions. Based on this, an autoregressive model of order p can be written analogous to multiple linear regression as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-} + \epsilon_t$$

where c is the intercept, ϕ values are coefficients, y the variable, and ϵ_t error term or white noise.

The "integrated" I part of the model represents the degree of differencing the series. For time-series analysis the data should be stationary, but it is not always the case and differencing is the most commonly used approach to transform non-stationary series into a stationary one. What differencing does is that it subtracts the observation in the current period from the previous. Commonly, differencing is used to reduce trend and seasonality by removing the changes in the level of a time-series (*ibid.*). As it is the change between consecutive observations in the series, it could be written as

$$y'_t = y_t - y_{t-1}.$$

Lastly, the moving average, or MA(q) component of ARIMA, represents using a linear combination of past forecast errors in a regression-like model as compared to AR(p) model which uses past values of the forecast variable. It means that the moving average models relate to what happens in period t to only past random errors that occurred in the previous periods. The MA(q) model can, therefore, be written as

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + +\theta_q \epsilon_{t-q}$$

where ϵ_t is white noise. In the MA(q) model, each value of y_t can be viewed as a weighted moving average of the past few forecast errors (*ibid.*).

Combining autoregressive and differencing with moving average model yields a non-seasonal ARIMA(p, d, q) model where the parameter values are non-negative integers. In this model, p is the number of time lags of the autoregressive model, d is the degree of differencing or the number of times the data have had its past values subtracted, and q is the order of moving average model. The full model can thus be written as

$$y'_t = c + \phi_1 y'_{t-1} + ... + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_p \epsilon_{t-q} + \epsilon_t$$

where y'_t is the series differenced t times and the "predictors" on the right hand side include both lagged values of y_t and lagged errors (*ibid.*).

The disadvantage of this model is that while it allows the inclusion of the past values of the dependent variable, it does not allow the other variables that might be relevant. On the contrary, the linear regression models can include these variables but do not allow the ARIMA time-series dynamics. However, by combining these two models it is possible to extend the ARIMA model to allow other independent variables to be included in it. The equation below shows the form of the equation which includes both ARIMA parameters and the linear regression model

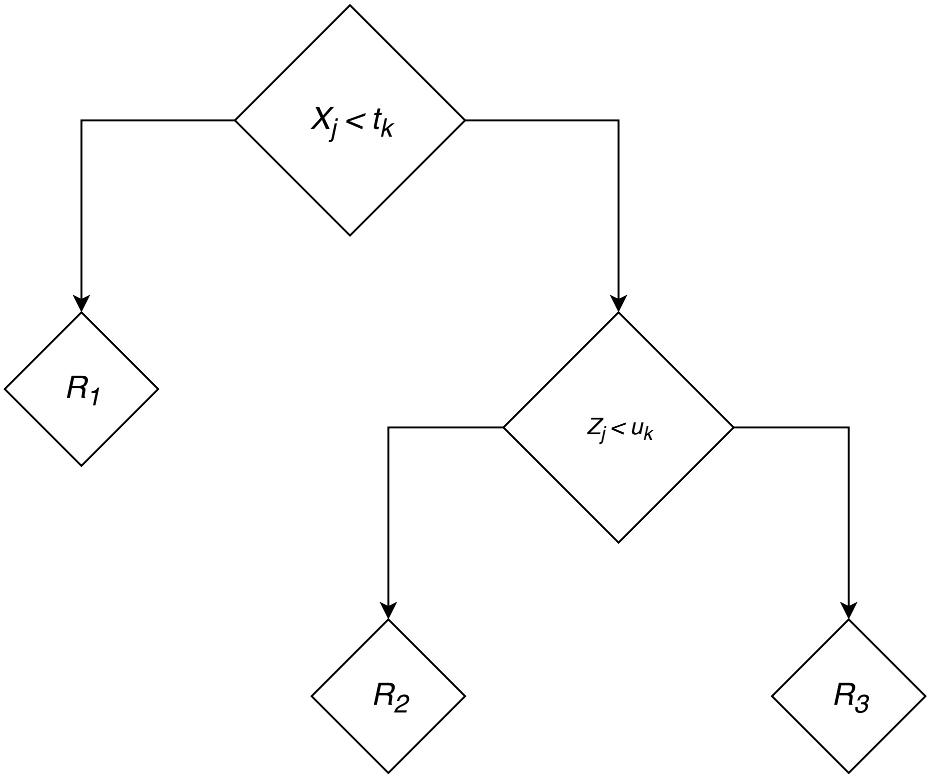
$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \eta_t, \\ \eta'_t &= c + \phi_1 \eta'_{t-1} + ... + \phi_p \eta'_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_p \epsilon_{t-q} + \epsilon_t. \end{aligned}$$

2.3.4 Tree-Based Methods

Another approach to regression analysis is using tree-based methods. These models involve stratifying the predictor space into a smaller number of sub-spaces and then predicting using the mean of the training observations in the region to which it belongs. The process of splitting the predictor space can be graphically visualized in a tree-like fashion, hence the name 'decision trees'. Default decision-tree models by themselves are generally simple and useful for interpretation, but in terms of prediction accuracy, they cannot generally compete with the best supervised learning approaches. To alleviate this disadvantage, the second half of this sub-chapter will look at the approaches such as random forests and gradient boosted trees, both of which involve producing multiple trees which are then combined to yield a single consensus prediction. The reason for employing these methods is that while the resulting models are somewhat more difficult to interpret, there are generally also significant improvements in prediction accuracy (James et al.: 303).

Pictured below in figure 2 is a graphical representation of the general form of a regression tree

Figure 2. Graphical representation of a decision-tree model.



Where first at the top of the tree, the model splits the variable X_j based on the splitting rules with the left-hand side being a sub-region where $X_j < t_k$. Similarly, the right-hand side consists of the sub-region of the data for which $X_j \geq t_k$. However, this sub-region is further divided into two different regions based on the value of variable Z_j compared to u_k. The predictions for each path are made at the bottom of the tree at respective end nodes R_1, R_2, R_3 by calculating their mean values of Y. In other words, it could be said that the decision tree stratifies the data into three regions of prediction space and these regions can in turn be written as $R_1 = \{X|X_j < t_k\}$, $R_2 = \{X|X_j \geq t_k, Z_j < u_k\}$, and $R_3 = \{X|X_j \geq t_k, Z_j \geq u_k\}$.

The process of creating a decision tree model could be viewed as two steps. First, the model divides the predictor space into J distinct and non-overlapping regions $R_1, R_2, ..., R_j$. Second, for every observation that falls into the region R_j, the model makes the prediction which is the mean of the response values for training observations in R_j (James et al. 2017: 306). The first step of the model divides the predictor space into high-dimensional rectangles with the objective of finding rectangles $R_1, ..., R_j$ that minimize the RSS, given by

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

where $\hat{y}_{R_J}$ is the mean response for the training observations within the jth rectangle (*ibid.*). It would be infeasible, and for larger data sets computationally impossible to consider every possible partition of feature space into J sub-spaces and for that reason, decision tree models use a top-down, greedy approach known as recursive binary splitting. The approach is considered top-down because it begins at the top of the tree, splitting each branch of predictor space into two successive branches down on the tree and it is greedy because at each split, it chooses the best split at that one point, rather than looking ahead and picking a split that leads to a better tree in future steps (*ibid.*). To perform recursive binary splitting, the model first selects the predictor X_j and the cutpoint s such that splitting the predictor space into regions $\{X|X_j < s\}$ and $\{X|X_j \geq s\}$ leads to the greatest possible reduction in RSS (*ibid.*: 307). More formally, the process can be viewed as that for any j and s, a pair of half-planes $R_1(j, s) = \{X|X_j < s\}$ and $R_2(j, s) = \{X|X_j \geq s\}$ can be defined and the model seeks the values of j and s that minimize the equation

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

where $\hat{y}_{R_1}$ is the mean response for the training observations in $R_1(j, s)$ and $\hat{y}_{R_2}$ is the mean response for the training observations in $R_2(j, s)$ (*ibid.*). After the model has found the values of j and s it repeats the same process of finding the best predictor and the best cut-point to split the data so as to minimize RSS, but this time it does not pick the whole feature space rather than one of the sub-spaces that resulted from the first split. The algorithm continues to split sub-spaces into two until some stopping criterion, such as a minimum number of samples in a sub-space, is reached. After that condition is met, the model has created sub-spaces $R_1, ..., R_J$ and predicts the response for a given test observation using the mean of training observations in the region to which that test observation belongs to (*ibid.*).

One of the disadvantages of the decision trees is that the model suffers from high variance. For example, if the training data set was split into two at random and a decision tree was fit on both halves, the results could be quite different. One of the ways to reduce the variance in a decision tree (or statistical learning methods in general) is to use bootstrap aggregation. The idea of the method is that by taking repeated samples from a single training data set, building a separate prediction model using each sample, and averaging the results, the variance of the model can be reduced and prediction accuracy increased. In other words, it would be possible to calculate $f^1(x), f^2(x), ..., f^B(x)$ using B separate samples drawn from the training data set, and obtain a single, low-variance statistical learning model by averaging the results, shown as follows (*ibid.*: 316-317)

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x).$$

Another alternative to the different types of tree-based methods is the random forests model. The algorithm itself is relatively similar to bootstrap aggregation, but with the difference that it decorrelates the trees from one another (*ibid.*: 319). This means that just like for bootstrap aggregation, random forests model builds a number of decision trees using bootstrapped training samples. However, when building the trees, each time a split is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors and the split is allowed to use only one of those m predictors (*ibid.*). It is also worth noting that for each subsequent split, the model does not use the m sample already selected but rather takes a new sample of m at each split. Also, the algorithm generally chooses the m value so that the number of predictors considered at each split is approximately equal to the square root of the total number of predictors or $m \approx \sqrt{p}$ (*ibid.*). The rationale for selecting a sample of predictors at each split and not even considering a majority is that it alleviates a potential problem of having one very strong predictor in the data over-influencing the results. For example, having a very strong predictor and using the bootstrap aggregation on the data would yield a set of trees which all, or at least most of them, would have this strong predictor in the top split. This would result in a set of trees which all look very similar to each other and thus the results from these trees would be highly correlated and averaging highly correlated quantities does not lead to as large of a reduction in variance as averaging many uncorrelated quantities (*ibid.*: 320). On the contrary, the random forest model would not consider the strong predictor for on average $(p-m)/p$ of the splits and so the other predictors would have more of a chance. Doing so decorrelates the trees and could be regarded as providing trees that are less variable and hence more reliable (*ibid.*).

A third approach to improve upon the predictions resulting from a decision tree is called "gradient boosting". The difference between gradient boosting and other tree-based methods is mainly that while, for example, bootstrap aggregation involves building each tree on a bootstrap data independently of the other trees and then averaging the results to obtain a single predictive model, but for gradient boosting the trees are grown sequentially. This means that "each tree is grown using information from previously grown trees" (*ibid.*: 321). Additionally, boosting does not use bootstrap sampling like other improvements upon the default decision tree model, but rather each tree is fit on a modified version of the original data set (*ibid.*). Similar to bootstrap aggregation and random forests, gradient boosting involves combining a large number of decision trees $\hat{f}^1, ..., \hat{f}^B$ into a single predictive model, but the algorithm used to obtain the trees is different. The process of gradient boosting for regression trees could be viewed as three separate steps. Firstly, the model sets $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training data set. The second step is that for $b = 1, 2, ..., B$, the model fits a tree $\hat{f}^b$ with d splits to the training data (X, r), updates $\hat{f}$ by adding in a shrunken version of the new tree $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$, and then updates the residuals $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$ (*ibid.*: 323). Lastly, the algorithm outputs the boosted model

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$$

More broadly, it could be said that instead of fitting a single large decision tree to the data and therefore potentially overfitting, the gradient boosting algorithm "learns slowly" (*ibid.*: 321). By this, it is meant that the algorithm fits the tree using the current residuals, rather than the outcome Y as the response and then adds this new decision tree into the fitted function in order to update the residuals. It is also worth noting that each of these trees can be rather small with just a few terminal nodes, the number of which is determined by the parameter d. The advantage of fitting small trees to residuals is that doing so allows to improve $\hat{f}$ in areas where it does not perform well (*ibid.*: 322). Additionally, the shrinkage parameter λ slows the process down even further and allows more and different shaped trees to attack the residuals (*ibid.*). Overall, it could be said that the statistical learning approaches that learn slowly tend to perform well. Mainly because they address the problems occurring in the models which fit the data hard.

As it can be seen from the algorithm, the boosting approach has three tuning parameters: the number of threes B, the shrinkage parameter or learning rate λ, and the number of splits in each tree d. All of these parameters must be tuned optimally in order for a boosted model to be as accurate as possible. The process of estimating these tuning parameters will be explored more in-depth later on in the work, as the values are not constant but rather depend on the particular data at hand.

3 Research Design

3.1 Sources of Data

In order to conduct the empirical analysis, the work uses two different data sources. For the party support, the data published by Kantar Emor is used. Kantar Emor is an Estonian research agency, which specializes in many different expertises such as market intelligence, customer strategies, behavioral economics, and among them, surveys support for Estonian political parties among the electorate on a monthly basis in aggregate time-series format. With regards to the sample size, the monthly surveys conducted by the agency generally include around 1000 respondents and are representative of the voting-eligible part of the population.

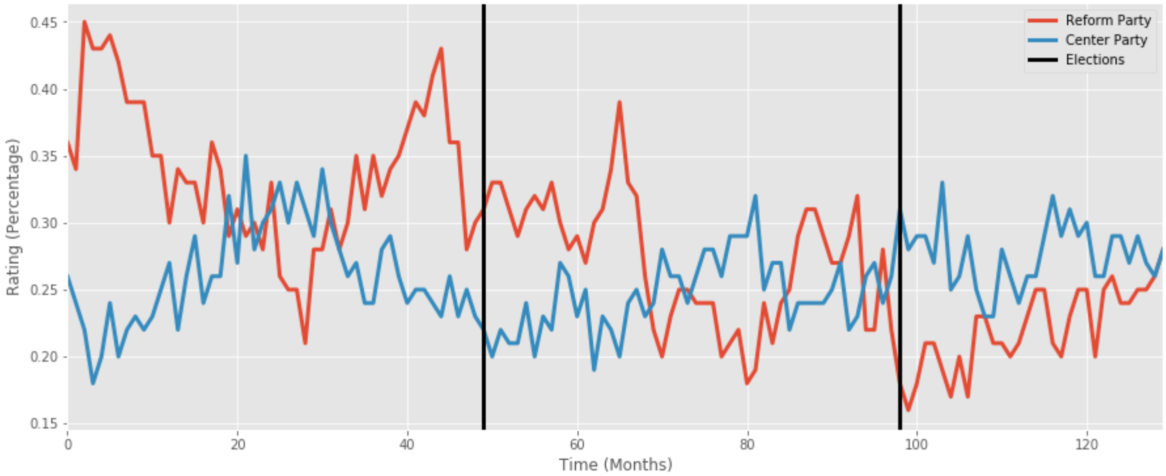
The three macroeconomic indicators used in the analysis, inflation, gross domestic product growth, and unemployment are all from the data published by Statistics Estonia. Statistics Estonia is the Estonian government agency responsible for providing both institutions and individuals with reliable and objective data on a number of areas, such as economic, social, demographic, and environmental. Even though the economic indicators are not all from the same dataset, rather than from different sub-datasets under the economic indicators category, the agency is in compliance with "international classifications and methods and in accordance with the principles of impartiality, reliability, relevancy, profitability, confidentiality, and transparency" (Statistics Estonia 2018). For this reason, even though the indicators are from different datasets of the same agency, they are compatible to be used in an analysis together.

3.2 Exploratory Data Analysis

3.2.1 Party Support

In order to understand party support better, it is important to evaluate why how large of a degree does it fluctuate from one observation to another. The figure 3 below shows the support for both parties over the observable period with the vertical black lines showing the start/end of an electoral cycle.

Figure 3. Support of the Reform Party (red) and the Center Party (blue) from 2007 to 2015 with the black lines indicating elections.

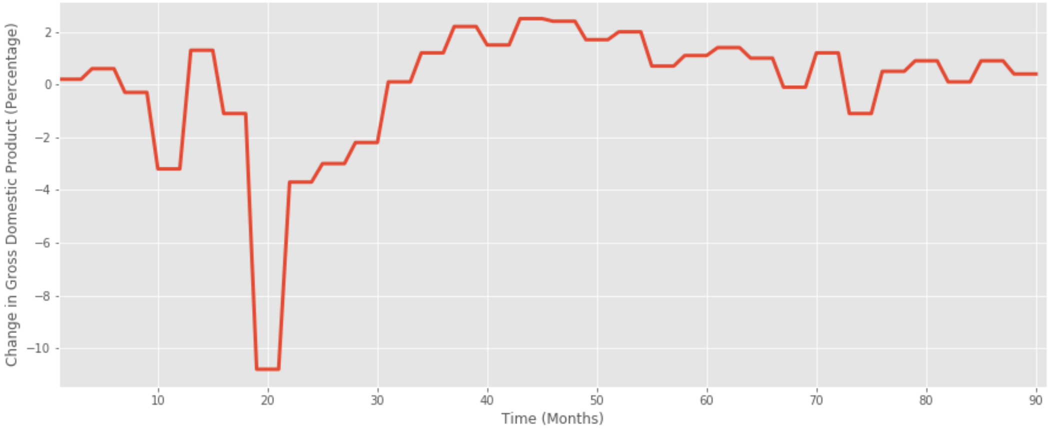


Firstly, it is possible to infer from the plot that the support ratings between consecutive periods are not very volatile. There does seem to be some inertia in the party support and the indicator generally either moves in a clear upward or downward trend or stays approximately at the same level over short-term with the changes becoming more pronounced over mid to long periods. For this reason, the models used in the analysis also incorporate a lagged party support value of $t-1$ since it can be expected that using an observation from a month before to predict the next one can yield accurate forecasts. Secondly, there seem to be clear cyclical trends in the support that correspond to the theory of electoral cycle. As the Reform Party has been in the coalition for the whole duration of the data, in accordance with the theory, the support for the incumbent party is at its highest at the start and end of a cycle and lowest in the middle. On the contrary, for the opposition party, the support is highest at the mid-point in the electoral cycle and it can be clearly seen to be the case for the Center Party with their rating being higher than for the Reform Party during these certain periods in the cycle.

3.2.2 Gross Domestic Product

One of the economic variables used in the analysis is the gross domestic product which measures the change in the market value of all the goods and services produced in Estonia on a quarterly basis. The time-series plot for the variable is pictured below in figure 4.

Figure 4. Change in gross domestic product from 2007 to 2014.

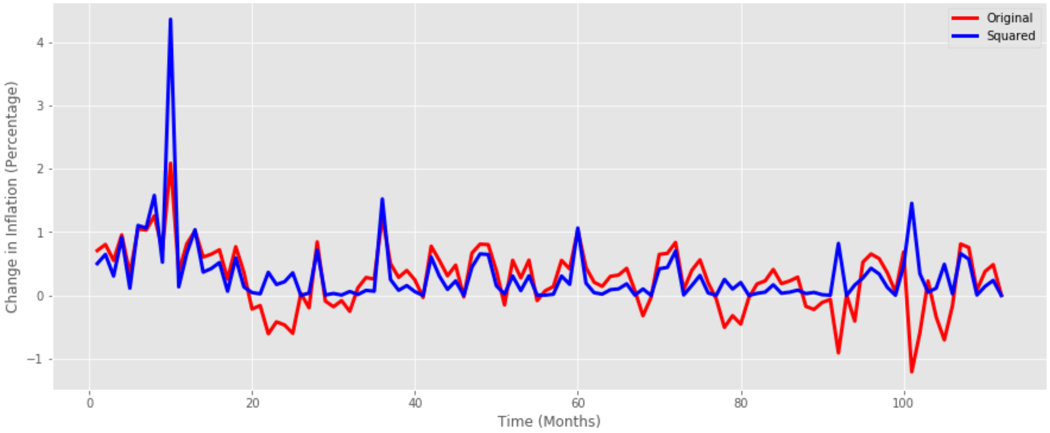


It can be seen from the figure that for the gross domestic product variable, there is not the same amount of data available as for the party support. The reason for it is that from 2007 to 2014, the variable used 2005 as the reference period, based on which the change was calculated. However, from 2014 to present, the Statistics Estonia results have used 2010 as the reference period, making the latter data incompatible with the earlier results. For this reason, the models used in the analysis will use the first 91 observations in the dataset to fit the models. As was explained in the chapter on economic voting theory, the relationship between party support and the change in the gross domestic product can be expected to be linear. It means that if the change in the gross domestic product increases, the incumbent party support is also expected to be higher and vice versa.

3.2.3 Inflation

Second of the economic indicators used in the analysis is inflation which measures the change in the price of goods and services on a monthly basis. The data for inflation is depicted below in figure 5.

Figure 5. Change in inflation from 2007 to 2016.

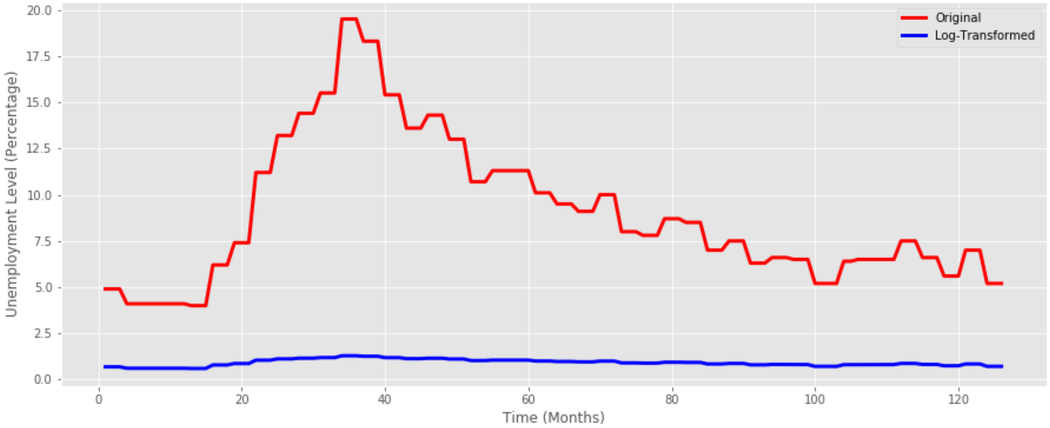


The figure includes the inflation variable both in its original form and as the squared version of the same indicator. The reason for doing so is that as can be recalled from the economic indicators theory, the theoretical relationship between party support and inflation is expected to be parabola-shaped. However, as some of the models used in the analysis are linear, the data must be transformed to better correspond to the ideal data shape that the linear model works best on.

3.2.4 Unemployment

Last of the economic variables used in the analysis is unemployment which is measured on a quarterly basis. Figure 6 shows the change in the value of the variable over time.

Figure 6. Change in unemployment level from 2007 to 2017.

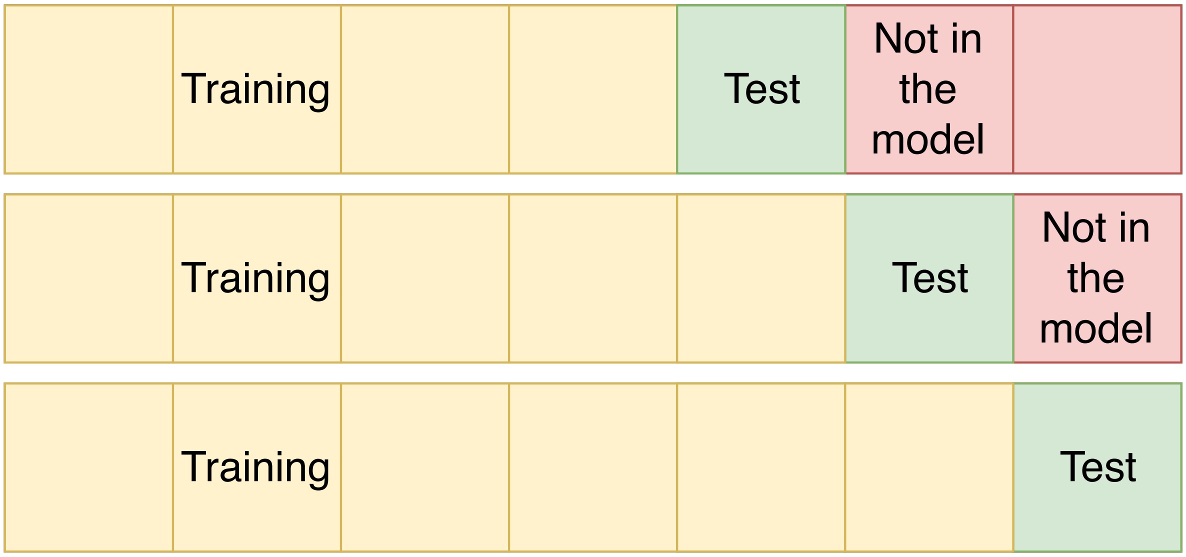


Similar to the inflation variable, the relationship between party support and unemployment is not linear, but rather steeper at some levels than others. As this type of data is not ideal for the linear model, the variable is linearized by logarithmic transformation.

3.3 Assessing Model Accuracy

In order to assess and compare the accuracy of different machine learning algorithms, it is also necessary to fit the models properly. Since the data is in the time-series format, meaning that the order of observations in the model is important, the one-step-ahead prediction method is used to fit the models and obtain the results. While many other fitting methods simply divide the data into training and test sets or alternatively training, validation, and test sets, one-step-ahead divides the dataset into three different parts: training set, test set, and the data which the model does not use. Essentially, one-step-ahead prediction could be viewed as $\hat{y}_{t+1} = f(x_{1:t}, y_{1:t})$ where each prediction is made by fitting the model with all of the data points that precede it. Visually, the process of the prediction method could be viewed as shown in the figure 7 below.

Figure 7. Visualization of one-step ahead prediction.



To compare the results of different models that have been obtained using the one-step-ahead method, the thesis uses the variance explained indicator, also known as the R^2 score to compare the observed test values to the respective predicted values. The way this indicator works is that if $\hat{y}$ is the predicted dependent variable output, y the observed output, and σ^2 the squared standard deviation or variance of the given data set then the explained is estimated as

$$R^2 = \frac{\sigma^2\{y - \hat{y}\}}{\sigma^2\{y\}}$$

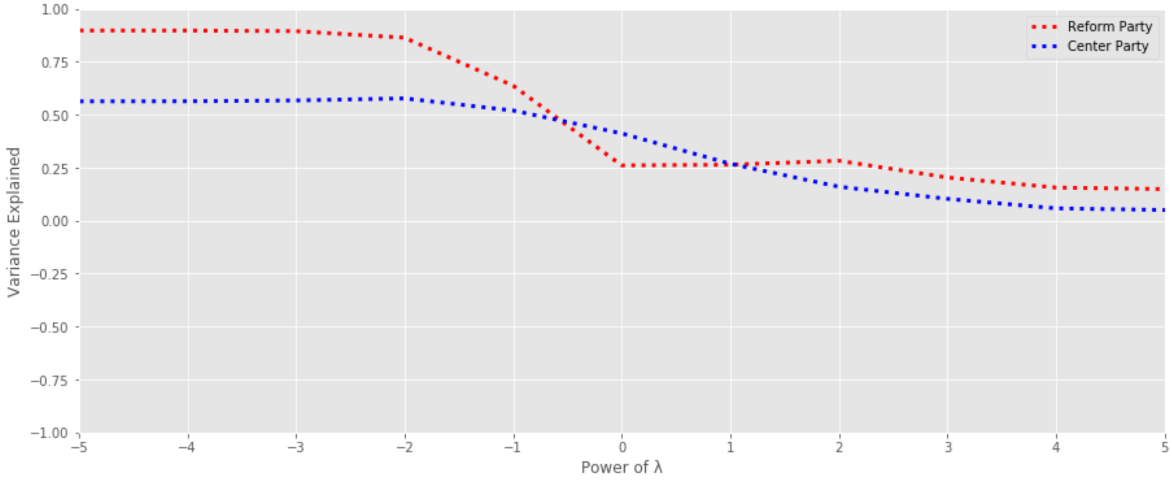
where the best possible score is 1.0 and lower values are worse. In this equation, σ^2 indicates the expectation of how much does the dependent variable deviates from its mean.

3.4 Optimizing Tuning Parameters of the Models

3.4.1 Ridge Regression

As noted before, in order to implement ridge regression and the lasso correctly, it is important to select at least near-optimal value for the tuning parameter λ in their respective equations. In reality, it is often a challenging task as the parameter is difficult to calibrate. Generally, the most common approach to choose the λ value is to use cross-validation, but in the present case, the method cannot be reliably used since the data is in the time-series format and cross-validation might misinterpret the trend. The next most reasonable approach is to fix the λ values as powers of 10 and compare the model performance for the different λ values. There is no clear agreed-upon consensus on the range which to choose the λ value from, but this work limits the range to the integer powers of 10 from 10^{-5} to 10^{5}. The reason for doing so is that even if the values smaller than 10^{-5} or larger than the upper limit were to prove better estimations, the improvement would be marginal and not worth the computation time. Additionally, it is reasonable to use integers as the powers of 10, since it could be said that the results for different integer powers do not generally fluctuate enough to warrant for iterating over rational numbers in small steps. Below in the figure 8 are shown the ridge regression results of macroeconomic indicators and party support for λ values from 10^{-5} to 10^{5} for both the Reform Party and the Center Party.

Figure 8. Ridge regression results with different λ values for both the Reform and the Center Party.

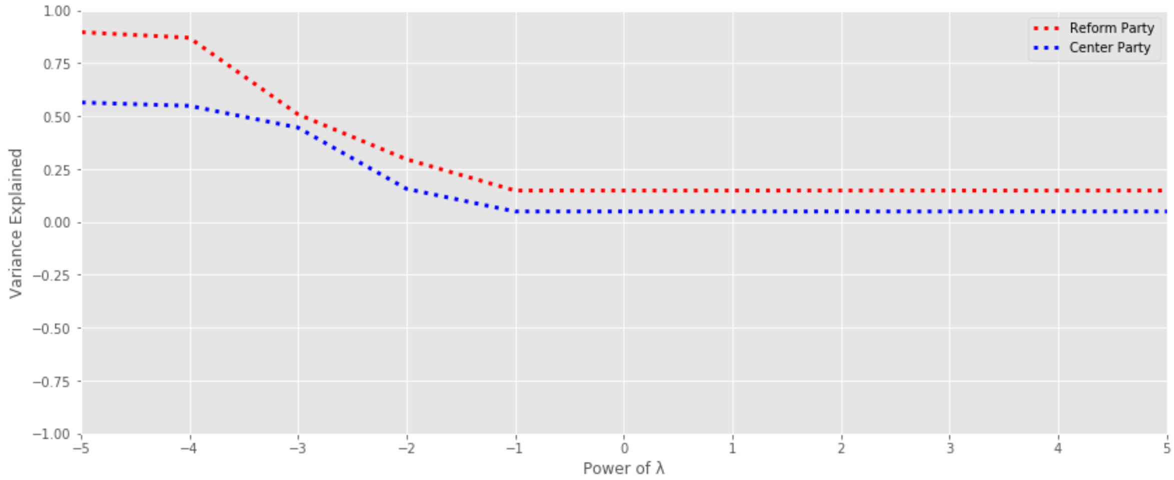


It can be seen from the figure that for both parties the λ values 10^{-3} or smaller provide the best estimation since for these λ values, the models are able to explain variance the best. If the explanation power lies constant like this over multiple different values, there is theoretically no difference which parameter value to choose. For this reason, the ridge regression models for both parties will use the λ value of 10^{-3} as the constant in their respective models.

3.4.2 Lasso Regression

As mentioned, the lasso model works similarly to the ridge model in a way that it also uses the λ parameter that shrinks the coefficient estimates towards zero with the exception that it can also set coefficients exactly to zero, or in other words, perform variable selection. Even though the optimal ridge regression λ value cannot be used for lasso model, the process of estimating the parameter is exactly the same. Below in the figure 9 are plotted lasso regression estimations for λ values from 10^{-5} to 10^{5} for both parties.

Figure 9. Lasso regression results with different λ values for both the Reform and the Center Party.



The figure shows that for both parties, the lowest λ values provide the best estimation. Even though the high values side of the model flattens out at around 10^{-4}, both of the models in the analysis will use $\lambda = 10^{-5}$ as the parameter since it still seems to somewhat improve over 10^{-4}. However, it is not necessary to iterate over even smaller values as the possible improvements in the model performance are very likely to be marginal as the variance explained by both models at 10^{-5} is already very high.

3.4.3 Autoregressive Integrated Moving Average Method

The autoregressive integrated moving average model consists of three parameters that must be tuned: the autoregressive component p, the degree of differencing in the model d, and the moving average q. The optimal combination of the different parameter values can be obtained by looping through the different values of p and q for both $d = 0$ and $d = 1$. Even though both the autoregression and moving average can take positive integer values limited by the size of the dataset, for the present analysis, the values of these both parameters have the upper limit of 3. The reason for doing so is that increasing the range would make the analysis more complex with seldom providing a significant improvement over using the range from 1 to 3. Below in tables 1 and 2 are displayed the ARIMA model R^2 estimations for the Reform Party and the Center Party.

Table 1. Comparison of different ARIMA parameter combinations for the Reform Party.

$d = 0$	$p = 1$	$p = 2$	$p = 3$	$d = 1$	$p = 1$	$p = 2$	$p = 3$
$q = 1$	0.947	0.931	0.834	$q = 1$	0.963	0.954	0.936
$q = 2$	0.919	0.841	0.847	$q = 2$	0.954	0.946	0.907
$q = 3$	0.860	0.860	0.865	$q = 3$	0.918	0.915	0.897

Table 2. Comparison of different ARIMA parameter combinations for the Center Party.

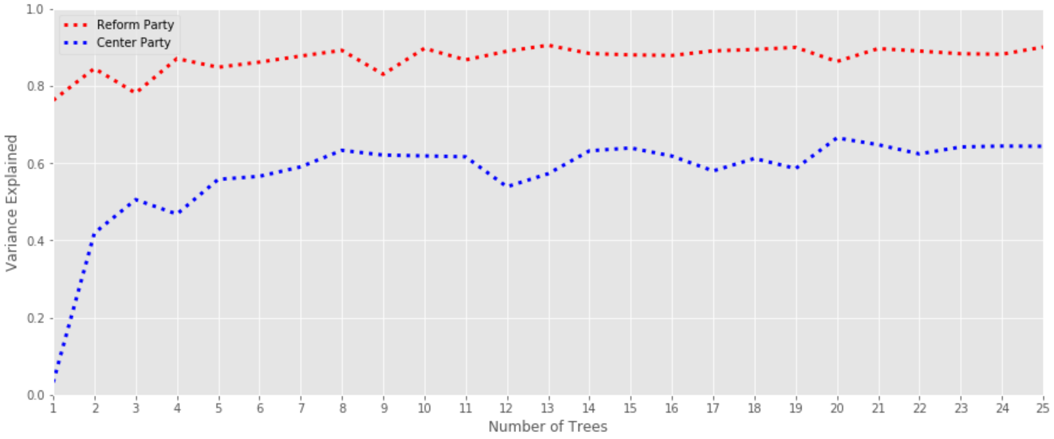
$d = 0$	$p = 1$	$p = 2$	$p = 3$	$d = 1$	$p = 1$	$p = 2$	$p = 3$
$q = 1$	0.724	0.728	0.721	$q = 1$	0.755	0.756	0.756
$q = 2$	0.713	0.758	0.670	$q = 2$	0.708	0.756	0.731
$q = 3$	0.684	0.667	0.678	$q = 3$	0.752	0.713	0.739

It can be seen from the results, that for the Reform Party, the model that is able to explain the variance best is $p = 1$, $d = 1$, and $q = 1$ which yields the value of 0.963. For this reason, this is the parameter configuration that will be used for the Reform Party model that is compared against the other machine learning algorithms. For the Center Party, there best estimation is provided by the model $p = 2$, $d = 0$, $q = 2$ which gives the $R^2 = 0.758$. However, seeing as the broader goal is to provide more universally applicable models that could also be potentially used to forecast the support of the other parties, the Center Party model will use the same $p = 1$, $d = 1$, and $q = 1$ configuration as the Reform Party model. It can be seen from the estimations that the difference between these two models for the Center Party is rather marginal, so the loss of 0.2% variance explained is justifiable.

3.4.4 Random Forest

In order for the random forest and the gradient boosted trees models to provide the best estimations, there are parameters that must be properly tuned beforehand for both models. For the random forests model, there are two parameters that must be estimated: the number of tried attributes and the number of trees. The importance of the former parameter is that it chooses the number of possible predictors considered each split. Even though this parameter allows for different options such as percentage or logarithm of all possible predictors, as the theory already noted, the best estimation is generally provided by $m \approx \sqrt{p}$ where p is the total number of possible predictors. As the current model has five independent variables, the value of this parameter is $m \approx \sqrt{5}$. It is not very important to try to optimize this parameter rather than just choose the value that works generally the best as the random forest model is usually not very sensitive about the value of this parameter. The number of trees parameter in the random forest model must be estimated manually. There is no generally agreed upon optimal number of the number of trees, but at the same time. it is not really possible to overshoot with this parameter as the upper limit is essentially bounded by computational and time constraints. For this reason, the correct way to estimate the optimal number of trees in a random forest model is similar to the regularized linear models in a way that the estimation can be obtained by looping the model for different values of the variable and then analyzing how different values of the parameter compare to one another. The goal is to find the number of trees where the model variance explanation power flattens out toward its upper limit, but at the same time, is as low as possible to minimize the computational time. Below in figure 10 are pictured the estimations for both parties over different numbers of trees.

Figure 10. Random forest regression results with different numbers of trees.



The results of both parties seem to stabilize when the model consists of around 20 trees. However, it can be seen that at this point, the estimation for the Reform Party dips while the Center Party peaks. For this reason, it would be more reasonable to pick a number of trees parameter value where the estimations for both parties is comparably high. Looking at the figure, it can be seen that one such point is while the number of trees in the model is 25 as at that point the variance explained for the Reform Party is at its high and for the Center Party near of its peak.

3.4.5 Gradient Boosted Trees

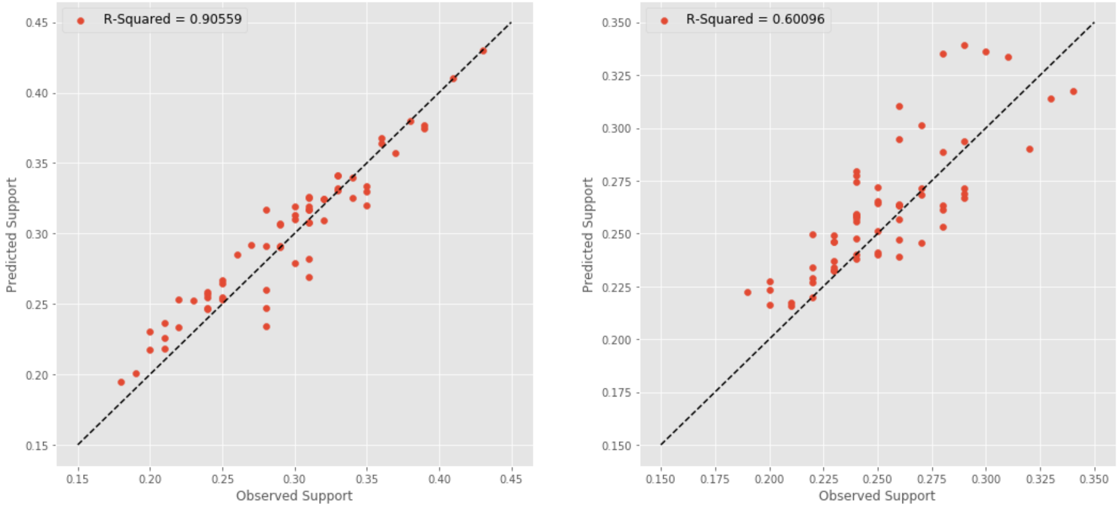
The gradient boosted trees regression model has three parameters that must be tuned for the optimal model performance: the number of trees B, the shrinkage parameter or learning rate λ, and the number of splits in each tree d. Similar to the random forest model, there is no specific number of trees a gradient boosted trees model should include and more is generally better as the model cannot over-fit. However, iterating over different values is not feasible for the gradient boosted trees model as it would require tuning all three parameters simultaneously to obtain the optimal combination. Instead, it is more reasonable to pick "safe" parameter values that are bound to be more or less theoretically correct. For this reason, in the present analysis, both models will use $B = 1000$ as using such value will surely be enough. For the other two parameters, there does exist some previous literature on how to properly tune them. For the λ value, the research states that the best strategy appears to be to set the λ as low (λ < 0.1) as possible as larger shrinkage yields improvements in the model performance (Hastie et al. 2016). Taking this into consideration, the models for both parties will use the shrinkage rate of λ = 0.01 in their algorithms. It would be theoretically possible to set the value even lower, but it is unlikely that the improvements would the significant enough to justify doing so. Lastly, the number of splits in each tree can also be estimated using gradient boosted trees theory. It states that generally, $d = 2$ will be insufficient and at the same time, it is unlikely that $d > 10$ would be required (Hastie et al. 2016). For this reason, it could be said that the values in the range of $4 \leq d \leq 8$ would work relatively well in the context of booting and as the results are relatively insensitive to the different values in this range, it does not really matter which one to choose so the models will simply take the middle value of the range and use $d = 6$.

4 Results

The following chapter gives an overview of the results for all of the models for both the Reform Party and the Center Party. It is worth noting that even though the metric used to measure the model accuracy is the R^2 score, the results are also visualized as scatter plots where on the x-axis are the observed values of the party support and on the y-axis, the predicted values. Additionally, the variance explained values presented in the results are calculated using only the one-step-ahead prediction results from index $t = 24$ onwards. The reason for doing so is that for the predictions, where the training data set is smaller, the model might not fit the function well and the results can vary a lot from one observation to another. However, after a while, as the number of observations in the training data grows, a better fit can also be expected. It is likely that if the predictions, where the training sample size is very low, were included, the R^2 value would be significantly influenced by the outlier data points and not so much representative of the actual value. The $t = 24$ value as the starting point is completely arbitrary and assumes that around this point, the training data size becomes large enough to start producing meaningful results.

4.1 Linear Regression

Figure 11. Linear regression results. Left: the Reform Party, right: the Center Party.

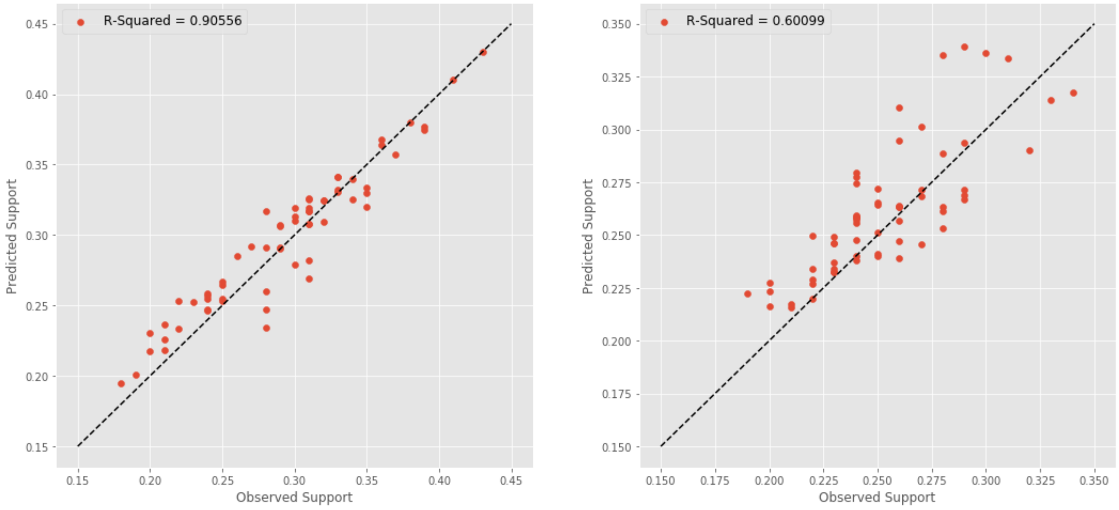


In the figure 11 are shown results of the simple multiple linear regression models for both the Reform Party and the Center Party. As it can be seen, the model for the Reform Party has the R^2 value of 0.90559 or in other words, the model is able to explain 90.559% of the variance between the independent variables and the dependent variable of the party support. However, for the Center Party, the linear regression model is able to forecast less accurately, being able to explain only 60.096 of the variance ($R^2 = 0.60096$), which is still a relatively good result. Based on the associations outlined in the theoretical framework section of the thesis, this model could be regarded as the theoretically correct. For this reason, the results of all of the subsequent models which address the shortfalls of this linear regression will be benchmarked against it.

4.2 Regularized Linear Methods

4.2.1 Ridge Regression

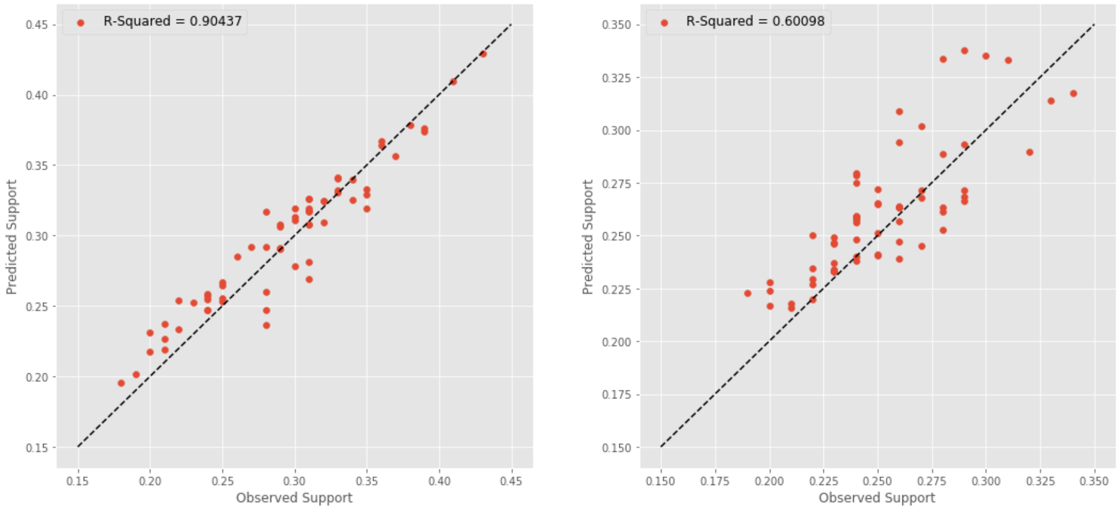
Figure 12. Ridge regression results. Left: the Reform Party, right: the Center Party.



It can be seen from the figure 12 that for the Reform party, the regularized ridge regression shows the R^2 score of 0.90556. This is basically the same result as the simple ordinary least squares model produced, being able to explain 0.003% less variance. For the Center Party, the result is very similar, giving the $R^2 = 0.60099$, a 0.003% increase over the linear regression model. From these results, it could be inferred that there is not much difference whether to use a linear regression or the ridge regression as the results differ only marginally.

4.2.2 Lasso Regression

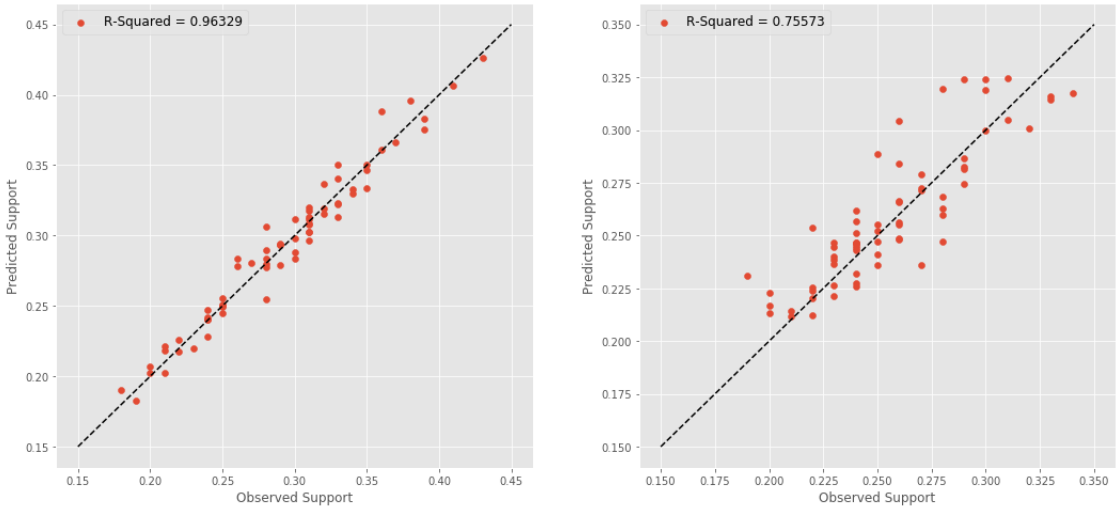
Figure 13. Lasso regression results. Left: the Reform Party, right: the Center Party.



As can be seen from the figure 13, compared to the ridge regression, the lasso regression model also produces relatively similar results. For the Reform Party, the model has the R^2 value of 0.90437, a slight decrease over the ridge regression and the linear model. For the Center Party, the model shows the R^2 value of 0.60098 which is also similar to the previous models. Based on these results, it is possible to say that the lasso regression does not seem to offer any meaningful improvements over the ridge model and overall, variable selection does not seem to play an important role.

4.3 Autoregressive Integrated Moving Average

Figure 14. ARIMA regression results. Left: the Reform Party, right: the Center Party.

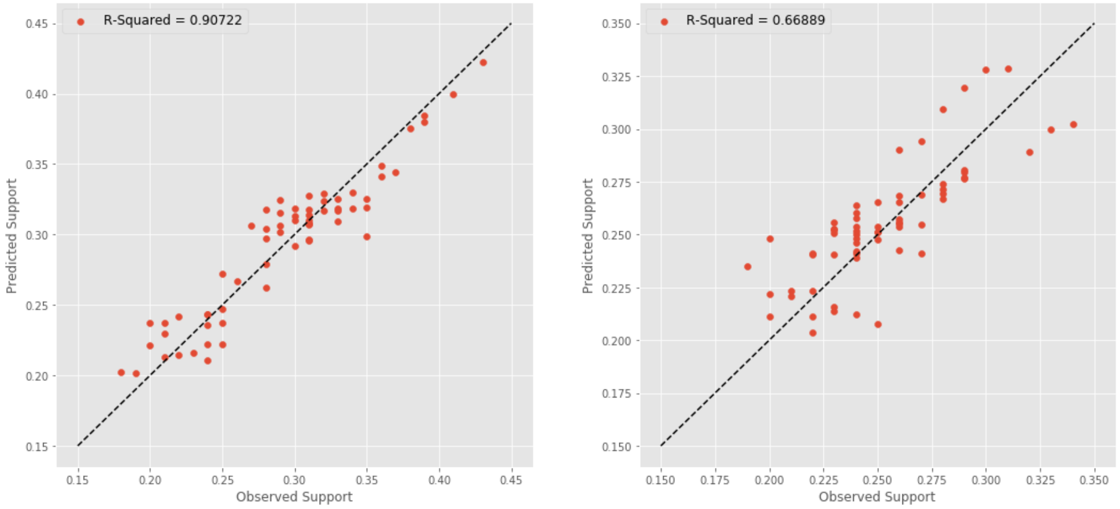


For the autoregressive integrated moving average model, it can be seen from the figure 14 that the results are better for both parties than the other algorithms are able to produce. The R^2 score of the Reform Party is 0.96317 which indicates that the model is able to explain variance extremely well. For the Center Party, the same score is 0.75573 which is also a good result taking into consideration the performance of the linear regression and the regularized models, all of which provide around 15% worse results. More specifically, it seems like both the differencing and the moving average aspect of it play part in the improved results. As shown in the optimal ARIMA parameter choosing section, for both models, moving from $d = 0$ to $d = 1$ with all of the other parameters the same yields around 1.6% increase for the Reform Party and 3.1% increase for the Center Party. The rest of the improvement in the models must be accounted to the moving average parameter q and not split between it and p as the latter indicates the number of the dependent variable lags in the model and every other model already includes a lag variable of $t - 1$. For this reason, it is possible to say that there are time-series aspects for the forecasting of party support using economic voting that the linear model by itself is not able to capture.

4.4 Tree-Based Methods

4.4.1 Random Forest Regression

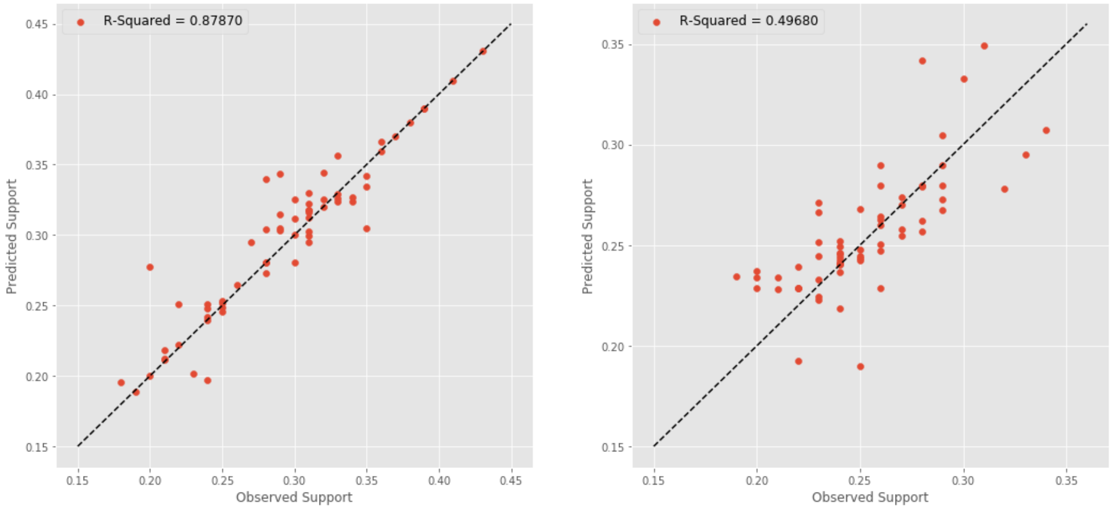
Figure 15. Random forest regression results. Left: the Reform Party, right: the Center Party.



Theoretically, the tree-based methods such as the random forest model strive for achieving the best possible estimation by fitting a non-parametric function to the data and could, therefore, be expected to provide the best forecasts. In the present case, as can be seen from the figure 15, the random forest model performs better than the linear regression and the regularized models, but worse than ARIMA. Looking at the results more closely, it can be seen that for the Reform Party, the R^2 value is 0.90722 which is only a marginal increase over the linear and regularized regression. For this reason, it could be said that in the Reform Party's case it is not necessary to fit a non-parametric function to the data as it does not offer much of an improvement. For the Center Party, the R^2 score is 0.66889. Similar to the Reform Party, it is an improvement over the linear regression and regularized models, but at the same time, performs worse than ARIMA. For this reason, it is likely that even though the model tries to find the maximum estimation, it does not produce better results than ARIMA on its own.

4.4.2 Gradient Boosted Trees Regression

Figure 16. Gradient boosted trees regression results. Left: the Reform Party, right: the Center Party.



The last model used in the analysis was the gradient boosted trees regression. Similar to the other tree-based methods, it is a non-parametric approach that tries to fit a function that maximizes the prediction accuracy and for this reason, is expected to provide highly accurate results. However, as it can be seen, in the present case, as the figure 16 shows, the gradient boosted trees algorithm estimations are the worst out of all models. For the Reform Party, $R^2 = 0.87870$ which is around 2.7% worse than the baseline linear regression model. Similarly, for the Center Party, the R^2 value is 0.49680 which is also by far the worst estimation by any model in the analysis. It is not entirely clear what exactly causes such results in this model, but the most likely explanation is that the model simply is not able to fit a good enough function to the model.

5 Discussion

Even though the results were already separately outlined in the previous section, the following table 3 also gives a concise overview of the prediction accuracy of the different algorithms for both parties.

Table 3. Results of the different machine learning algorithms obtained in section 4 for both parties.

	Reform Party	Center Party
Linear Regression	0.90559	0.60096
Ridge Regression	0.90556	0.60099
Lasso Regression	0.90437	0.60098
ARIMA	0.96317	0.75573
Random Forest	0.90722	0.66889
Gradient Boosted Trees	0.87870	0.49680

The linear regression model, which has previously been the default method when it comes to forecasting party support using economic indicators was used as the benchmark to compare the other models against, all of which could be argued to improve upon it in their own way. However, as it can be seen from the results that most of the models do not provide better estimations for forecasting party support. In terms of the regularized models, neither ridge or lasso regression are able to increase or decrease the prediction accuracy by a very marginal amount. For this reason, it could be said that for the data used in the analysis, variable selection was not needed or was not able to produce better results. For the tree-based models, the random forest regression was able to produce approximately the same results as the linear regression for the Reform Party and somewhat better results for the Center Party. The gradient boosted trees model, however, produced the worst results across the board. As the tree-based models have the primary goal of fitting the data as close to the data points as possible without actually over-fitting the data, such result was unexpected. It was expected that the tree-based models would at least match the prediction accuracy of the baseline model or improve upon it, seeing as these models are not bounded by the same restrictions as the parametric models. It is difficult to evaluate and make an exact judgment on why did these models provide such mediocre results compared to all of the others. It is a possibility that the models simply weren't able to find a function that would work both well on the training and testing data. The only model that provided improvements upon the linear regression was the autoregressive integrated moving average algorithm. As can be seen from the results, the model was able to explain variance much better than any other model in the analysis, improving the variance explained for the Reform Party by around 5.76% and for the Center Party by around 15.47%. These are pretty substantial gains and give an indication that there are time-series aspects about modeling the party support through economic voting that the simple multiple linear regression does not take into account. As already mentioned in the results, the improvements come approximately equally from the inclusion of both differencing and moving average parameters in the model, but not from the autoregressive parameter as it is already included in the other models used. For this reason, the further research on the topic should also consider incorporating these variables into the analysis as these results give a legitimate reason for doing so.

Next, it is possible to see that the forecast accuracy for the Center Party is systematically around $20 - 30\%$ worse than all across the board than for the Reform Party. The exact reasons for why is it so much more difficult to model the party support forecasts for the Center Party than it is for the Reform are difficult to pinpoint. One reason might be that the electorate of the Center Party is more stable and less responsive to the economic effects than the electorate of the Reform Party. It is very much possible that the Reform Party electorate is more in line with the economic voting theory while the Center Party voters are not. This means that when the economic effects become favorable to the Center Party, the voters do not move from the Reform Party to the Center Party, but instead, the other non-incumbent alternatives and the Center Party voters are more reluctant to change their vote based on the economic effects. Overall, it is a question that is difficult to address without a proper analysis and could, therefore, be viewed as a further research topic on studying the party support in Estonia.

Lastly, while the main research question of the thesis was to compare the linear regression forecasting model to the models that address its shortcomings, its secondary goal was to explore using more modern machine learning algorithms as the methods for social sciences research. Even though most of the models, especially the tree-based methods, were unable to improve upon linear regression in the present case, with the rise of big data, social sciences researchers should be more motivated to consider these methods in their research, especially when it comes to forecasting. There is already research published that uses machine learning algorithms as methodology such as identifying behavioral patterns (King et al. 2013; Pierson 2017), measuring ideological and political preferences using big data (Bond & Messing 2015;

Barbera 2014), applying machine learning algorithms in society to study human decisions (Kleinberg et al. 2018). It can only be expected that as the time goes on, machine learning models as legitimate social sciences research method will only become more prevalent so for this reason, the researchers should already start looking in that direction.

6 Conclusion

The main objective of the thesis was to compare machine learning regression algorithms in the context of forecasting party support in Estonia using economic voting variables and see whether more modern methodological approaches would produce improvements over the linear regression model that has been used in the past. To analyze this research question, the thesis compared the regularized linear methods of ridge and lasso regression, autoregressive integrated moving average, and the decision-tree models of random forests and gradient boosted trees, all of which have their separate merits over the linear regression and could therefore potentially improve upon the default model. Even though there have been some rudimentary studies at the Estonian level in the past, the analysis conducted in this thesis was the first in-depth study that forecasts the party support using economic indicators as the predictor variables.

The body of the thesis was divided into four main sections: theoretical background, research design, results, and discussion. While the first of them gave a theoretical understanding of both economic voting theory, the machine learning methods used, and how the economic voting indicators are theoretically related to forecasting the party support, the other parts were more empirical. Research design gave an overview of the sources of the data used in the analysis and visualized the variables in the time-series format. Additionally, it focused on optimizing parameters of the different algorithms as it is necessary to ensure the optimal model prediction accuracy. The results chapter gave an overview of the results of all of the models and benchmarked all of the models against the linear regression. Lastly, the conclusion part of the thesis provided more general comments of the results of the analysis and drew inferences of them.

The results of the analysis showed that while most of the models, such as the decision-trees, were not able to estimate better than the linear regression model or in the case of regularized models were able to do it very marginally, the autoregressive integrated moving average model was able to produce very clear improvements for both parties used in the analysis. For this reason, maybe aspects of the economic voting theory should be rethought to also incorporate the time-series aspects that give the ARIMA model an edge over the linear regression.

Even though the thesis was able to show that the ARIMA model could be regarded as a better method to study this phenomenon, there still remain avenues for further research in economic voting in Estonia. One possibility would be the already mentioned question of why is the prediction accuracy gap between different parties significantly large. Secondly, the present thesis incorporated a small number of independent variables and for this reason, it might also be worth looking into the variables that weren't included in the present analysis but which might influence the party support.

7 Bibliography

Anderson, C. J. 2007. “The End of Economic Voting? Contingency Dilemmas and the Limits of Democratic Accountability.” *Annual Review of Political Science 10*(1): 271–96.

Anson, I. G., and T. T. Hellwig. 2015. “Economic Models of Voting.” *In Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource* edited by R. Scott and S. Kosslyn, 1–14. John Wiley & Sons, Inc.

Barberá, P. 2014. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis 23*(1): 76-91.

Bond, R. and Messing, S. 2015. Quantifying social media’s political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review 109*(1): 62-78.

Campbell, A., P. E. Converse, W. E. Miller, and D. E. Stokes. 1960. *The American Voter*. New York, NY: Wiley.

Downs, A. 1957. *An Economic Theory of Democracy*. 1st edition. New York, NY: Harper.

Duch, R. M., H. D. Palmer, and C. J. Anderson. 2000. “Heterogeneity in Perceptions of National Economic Conditions.” *American Journal of Political Science 44*(4): 635–52.

Duch, R. M. 2001. “A Developmental Model of Heterogeneous Economic Voting in New Democracies.” *The American Political Science Review 95*(4): 895–910.

Evans, J. A. 2004. *Voters and Voting: An Introduction*. 1 edition. London; Thousand Oaks, CA: SAGE.

Evans, G., and R. Andersen. 2006. “The Political Conditioning of Economic Perceptions.” *Journal of Politics 68*(1): 194–207.

Fiorina, M. P. 1981. *Retrospective Voting in American National Elections*. New Haven: Yale University Press.

Fraile, M., and M. S. Lewis-Beck. 2014. “Economic Vote Instability: Endogeneity or Restricted Variance? Spanish Panel Evidence from 2008 and 2011.” *European Journal of Political Research 53*(1): 160–79.

Friedman, J., Hastie, T. and Tibshirani, R. 2001. The Elements of Statistical Learning. New

York: Springer.

Goodhart, C. A. E., and J. Bhansali. 1970. “Political Economy.” *Political Studies 18*(1): 43–106.

Hyndman, R.J. and Athanasopoulos, G. 2014. *Forecasting: Principles And Practice*. OTexts.

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. *An introduction to statistical learning*. New York: Springer.

Key, V. O. 1966. *The Responsible Electorate: Rationality in Presidential Voting, 1936- 1960*. 2nd edition. Cambridge: Belknap Press of Harvard University Press.

Kiewiet, D. R. 1983. *Macroeconomics and Micropolitics: The Electoral Effects of Economic Issues*. Chicago, IL: University of Chicago Press.

King, G., Pan, J. and Roberts, M.E. 2013. How censorship in China allows government criticism but silences collective expression. *American Political Science Review 107*(2): 326-343.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics 133*(1): 237-293.

Kramer, G. H. 1971. “Short-Term Fluctuations in U.S. Voting Behavior, 1896–1964.” *American Political Science Review 65*(1): 131–143.

Kramer, G. H. 1983. “The Ecological Fallacy Revisited: Aggregate- versus Individual Level Findings on Economics and Elections, and Sociotropic Voting.” *American Political Science Review 77*(1): 92–111.

Lewis-Beck, M. S. 1986. “Comparative Economic Voting: Britain, France, Germany, Italy.” *American Journal of Political Science 30*(2): 315–46.

Lewis-Beck, M. S. 1988. *Economics and Elections: The Major Western Democracies.* Ann Arbor, MI: University of Michigan Press.

Lewis-Beck, M. S., and M. Paldam. 2000. “Economic Voting: An Introduction.” *Electoral Studies 19*: 113–21.

Lewis-Beck, M. S., W. G. Jacoby, H. Norpoth, and H. F. Weisberg. 2008. *The American Voter Revisited.* Ann Arbor, MI: University of Michigan Press.

Lewis-Beck M. S., and Stegmaier, M. 2013. “Economic Voting.” *Oxford Bibliographies Online: Political Science.*

Mueller, J. E. 1973. *War, Presidents and Public Opinion*. New York, NY: Wiley.

Nannestad, P., and M. Paldam. 1994. “The VP-Function: A Survey of the Literature on Vote and Popularity Functions after 25 Years.” *Public Choice 79*(3–4): 213–45.

Nannestad, P., and M. Paldam. 1997. “From the Pocketbook of the Welfare Man: A Pooled Cross-Section Study of Economic Voting in Denmark, 1986–92.” *British Journal of Political Science 27*(1): 111–155.

Paldam, M. 1991. “How Robust Is the Vote Function? A Study of Seventeen Nations over Four Decades.” *In Economics and Politics: The Calculus of Support.*, edited by H. Norpoth, M. S. Lewis-Beck, and J. D. Lafay, 9–31. Ann Arbor, MI: Michigan University Press.

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Ramachandran, V., Phillips, C. and Goel, S. 2017. A large-scale analysis of racial disparities in police stops across the United States. arXiv preprint arXiv:1706.05678.

Silver, N. 2012. "*Measuring the Effect of the Economy on Elections*".

Singer, M.M. and Carlin, R.E. 2013. Context counts: The election cycle, development, and the nature of economic voting. *The Journal of Politics, 75*(3): 730-742.

Solvak,M.2015."*Erakonnatoetus perioodil 2007-2015:valimistsükkel ja majanduse käekäik*".

Tibbitts, C., 1931. Majority votes and the business cycle. *American Journal of Sociology, 36*(4), pp.596-606.

Tufte, E. R. 1978. *Political Control of the Economy*. Princeton, NJ: Princeton University Press.

van der Brug, W., C. van der Eijk, and M. Franklin. 2007. *The Economy and the Vote: Economic Conditions and Elections in Fifteen Countries*. New York, NY: Cambridge University Press.

van der Eijk, C., and K. Niemöller. 1987. “Electoral Alignments in the Netherlands.” *Electoral Studies 6*(1): 17–30.

Welch, S., and J. Hibbing. 1992. “Financial Conditions, Gender, and Voting in American National Elections.” *The Journal of Politics 54*(1): 197–213.

Wlezien, C., M. Franklin, and Daniel Twiggs. 1997. "Economic Perceptions and Vote Choice: Disentangling the Endogeneity." *Political Behavior 19*(1): 7–17.

Non-exclusive license

I, Kaarel Kaasla (39201076511),

herewith grant the University of Tartu a free permit (non-exclusive licence) to:

FORECASTING THE PARTY SUPPORT IN ESTONIA: COMPARISON OF MACHINE LEARNING REGRESSION ALGORITHMS

supervised by Mihkel Solvak and Kaspar Märtens.

1. To reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright.
2. To make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright.
3. I am aware that the rights stated in point 1 also remain with the author.
4. I confirm that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 21.05.2018

....