Spreading the rooftop revolution

Understanding the residential solar energy sector using machine learning: who is buying solar panels and where is the potential?

Maurice Bosma

Abstract

To help policy makers and businesses active in the residential solar energy sector understand who is buying solar panels and where the potential is, this study aims to provide a systematic understanding of the interplay between socioeconomic factors and the photovoltaic (PV) adoption rate. A large data set about solar energy generation and a variety of socioeconomic factors about wealth, demographics and way of living is used including information for every neighborhood in the Netherlands of 2019. The research framework contains the machine learning methods generalized linear model (GLM), decision tree and random forest and test RMSE-scores are compared to measure performance. The best performing model is a random forest. Moreover, sophisticated interpretation methods such as partial dependence plots and decision tree-based clustering are used to interpret the models and uncover regional differences for PV adoption. The results show that 12 out of 16 socioeconomic factors have a significant estimate and the key drivers of PV adoption are the absence of old houses, the presence of single-family homes, the education rate and electricity use. Moreover, this paper proposes a decision tree-based clustering of neighborhoods based on similar socioeconomic characteristics and PV adoption rates. The 12 clusters of similar neighborhoods provide a good overview of the different PV adoption rates by looking at the socioeconomic characteristics which can be used for policy making and marketing purposes. Especially the 4 clusters with the lowest average PV adoption rate are of interest to stakeholders because the specific characteristics of these neighborhoods can be used to target groups effectively and increase PV adoption. Finally, one of the most important results is a better understanding of the regional disparity for PV adoption in the Netherlands by combining machine learning model output and geographical data.

ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Master Thesis Data Science and Marketing Analytics
Supervisor: O. Karabag
Second assessor: B.G.C. Dellaert
Date final version: July 17, 2022
Student number: 454492

Table of contents

1. Introduction

The coming decades, our society is facing one of the biggest challenges since human existence. The world population is increasing rapidly and individual consumption keeps growing which leads to a higher footprint on our planet. Using unlimited resources on a limited planet takes its toll and if we don't act quickly, global warming will lead to drastic problems which will worsen in the decades to come (IPCC, 2021).

Fortunately there is still time to face this challenge. Both governments and corporations are taking actions to limit human-caused effects on climate change by working together and reducing our footprint. The strongest cooperation to fight global heating is the Paris Agreement adopted at the Paris Climate Conference (COP21) in December 2015. The Paris Agreement provides a global framework to avoid irreversible and dangerous climate change by limiting global warming to well below 2°C making sure our planet is preserved in the long run.

One of the fundamental transformations in this process is related to the energy supply system. Today's energy supply system is heavily dependent on fossil fuels which are both non-sustainable and very polluting resources. To limit global warming the use of fossil fuels should be replaced by low greenhouse gas alternatives such as renewable energy. In a broad sense, renewable energy sources refer to biomass energy, hydro energy, tidal energy, wind energy, geothermal energy and solar energy. Of these renewable energy sources, solar energy, also called photovoltaic (PV) energy, is amongst the highest potential sources in both efficiency and cost reduction potential (Panwar et al., 2011).

To ensure the energy transition from fossil fuels to renewables goes well, governmental policies are necessary in order to overcome barriers, allow the technology to mature and accelerate the transition. Renewables have long occupied energy policies in a wide range of countries especially focussing on the adoption of PV within the residential sector. Support from local governments largely consisted of financial assistance by offering subsidies for PV installations and establishing Feed-in Tariffs (FiT), which is the uptake by energy suppliers of generated renewable energy by residents for a fixed price.

Although energy policies have led to a significant increase in PV adoption among many countries, the potential in increasing residential PV adoption is still enormous. In the Netherlands for example, solar panels generate 9.6% of the total electricity supply in 2021 (DNE research, 2022). Research conducted by Deloitte (2018) shows that supplying every suitable roof in the Netherlands with solar panels can supply 50% of the Dutch energy usage. The large potential in PV adoption leads to new policies implemented every year. This raises the question how such policies should look like to reach their goal effectively. It is important to develop sustainable and effective policies to make sure the

goal of introducing such policy is reached sufficiently. Especially targeting the right people and reaching those people can be of challenge when a new policy is implemented (Lu et al., 2020). A better understanding of who is already buying solar panels and who is not can be of great help for policy makers to define their policy. Moreover, knowing where the potential is can be used for implementing effective policies and targeting the right people and can be of great benefit for spreading the rooftop revolution.

With this thesis, I aim to give a better understanding of who is buying solar panels and where the potential is. A variety of socioeconomic factors about wealth, demographics and way of living are used to better understand which characteristics of a neighborhood are associated with residential solar adoption. The study is conducted within the Dutch residential sector and uses solar energy and socioeconomics data of 2019. The research questions is as follows:

To what extent can socioeconomic factors be used to predict and understand the residential PV adoption in the Netherlands using machine learning models?

The thesis is structured by using several sub-questions. First, previous literature on residential solar energy is reviewed to find out what drives PV adoption. Then, research is conducted on the key socioeconomic factors that explain the variance of solar adoption in the Netherlands using machine learning. Finally, the results are interpreted to find out how they can contribute to effective policy making to accelerate PV diffusion in the Netherlands.

The above combined leads to the formulation of the following sub-questions:

- **Which socioeconomic factors drive PV adoption according to previous literature?**

This sub-question helps to address the main research question by understanding which socioeconomic factors are related to PV adoption according to existing literature. The literature is reviewed and provides the fundamental context for this analysis.

- **What are the key socioeconomic factors for explaining the variance of residential PV adoption across different neighborhoods of the Netherlands?**

This sub-question comprises the main analysis of this thesis. Machine learning models are applied and interpreted to understand which socioeconomic factors are related to PV adoption in the Netherlands. Moreover, this question focuses on regional disparity within the country.

- **How can these key socioeconomic factors contribute to effective policy making to accelerate PV diffusion in the Netherlands?**

Within machine learning, one should pay attention to the goal of applying machine learning techniques without getting lost in the details. Therefore, the last sub-question focuses on the main findings derived from the analysis and how they are relevant for policy making. Eventually this analysis should give a better understanding on PV adoption and how to accelerate the rooftop revolution.

To address the research questions at hand, a large dataset is analyzed containing solar energy generation and socioeconomic factors for every neighborhood in the Netherlands of 2019. The research framework contains the machine learning methods generalized linear model (GLM), decision tree and random forest and test RMSE-scores are compared to measure performance. Moreover, sophisticated interpretation methods such as partial dependence plots and decision tree-based clustering are used to interpret the models and uncover regional differences for PV adoption. Concerning the methods, the best performing model is a random forest. However, the generalized linear model and decision tree also predict well and therefore all three models are interpreted to strengthen the findings of this analysis. The results show that 12 out of 16 socioeconomic factors have a significant estimate and the key drivers of PV adoption are the absence of old houses, the presence of single-family homes, the education rate and electricity use. The presence of old houses and electricity use have a negative effect on PV adoption whereas the presence of single-family homes and education rate have a positive effect on PV adoption. Moreover, this paper proposes a decision tree-based clustering of neighborhoods based on similar socioeconomic characteristics and PV adoption rates. These groups of similar neighborhoods provide a good overview of the different PV adoption rates by looking at the socioeconomic characteristics and can be used for policy making and marketing purposes. Of the 12 neighborhood clusters, especially the 4 clusters with the lowest average PV adoption rate are of interest to stakeholders. Policy makers could target the residents within these clusters with different strategies that respond to the characteristics in these neighborhoods. Policy makers could, for example, highlight the opportunity to loan money for green investments and the cost benefit of having solar panels to neighborhoods that are relatively poor. For another cluster of neighborhoods, policy makers could focus more on providing information in multiple languages because there is a relatively high percentage of residents having a migration background in this cluster. As a final example, policy makers could focus more on raising awareness about climate change within the cluster of neighborhoods with relatively many elderly that might be unaware of the challenge we are facing. Finally, one of the most important results is a better understanding of the regional disparity for PV adoption in the Netherlands by combining machine learning model output and geographical data. This information can be used to understand which clusters of neighborhoods are present in which provinces to even better target neighborhoods with a low PV adoption rate.

The academic contribution of this research is threefold. First, to the best of my knowledge, state of the art machine learning methods have barely been used before within this research topic and can provide relevant new insights. Using state of the art machine learning methods have great advantages in dealing with non-linear relationships and unknown interactions, which allows for better explanatory power compared to traditional methods such as linear regression models. After thoroughly studying existing literature on PV adoption, there are few papers that do make use of machine learning already (Lan, Gou & Lu, 2021; Sommerfeld et al., 2017; Sunter et al.,2019). However, these papers assess different methods and there is still room to explore other modern machine learning methods within this topic. Second, this type of research has not been conducted on the Dutch residential PV market yet. Therefore, this research contributes to a better understanding of the relationship between socioeconomic attributes and PV adoption in the Netherlands which can be of great benefit for Dutch policy makers to understand who is buying solar panels and where the potential is. Third, the non-examined factor low literacy is researched in the context of PV adoption. It is expected that individuals with low literacy skills tend to need more time to read and understand information than others which usually leads to a slower adoption of new information, like the urge of climate change in this context. Although existing literature about PV adoption did focus on education levels, low literacy was not examined before and could be a potential driver for PV adoption.

Besides academic contribution, conducting this research is also of great relevance for marketing strategies and policy making. A better understanding of who is buying solar panels and who is not (yet) can be used for multiple purposes. First, governmental organizations can utilize this information to better understand if previous energy policies have worked effectively and if the end users of the policy have been reached accordingly. Second, this information can be used to develop new effective policies by focussing on groups that have a relatively low PV adoption. Third, policy makers can leverage this analysis to start targeting specific groups or neighborhoods that are not part of the rooftop revolution yet. Also businesses active in the PV market can utilize this information in the same way. By understanding who is not buying solar panels yet, a company or the government can adjust its marketing strategy accordingly to effectively reach those groups. For the sake of brevity, this analysis mainly focuses on relevant implications for targeting groups and neighborhoods from a policy maker perspective. However, the same implications can be used for commercial marketing strategies and new policy making.

Last, but certainly not least, conducting this research can be of great societal and environmental relevance. In the end, society as a whole is depending on a healthy and sustainable planet for this generation and generations to come. To assure our planet is preserved, global warming should be limited and the transition of our energy supply system to renewable energy sources plays a crucial role in this process. Although governments are already taking up this challenge, the transition goes slow

and there is a big urge to accelerate. Understanding who is buying solar panels and, more important, who is not can be of great contribution in accelerating PV adoption. By using the results obtained in this study, policy makers have a better understanding of who is buying solar panels and this information can be leveraged to arrange policies more effectively and make the adoption even faster.

The remainder of this paper is structured as follows. Chapter 2 consists of the theoretical background including an extensive overview of the Dutch residential PV market and previous studies focussing on PV adoption and socioeconomic attributes. The data used in this research is further discussed in Chapter 3. In Chapter 4 the methods are explained including machine learning techniques and interpretation methods. Moreover, Chapter 5 reviews the methods and model outcomes are compared. Chapter 6 contains a discussion and Chapter 7 consists of a conclusion including limitations and recommendations for future research.

2. Theoretical background

The following subsections summarize and discuss the existing literature on PV adoption including the academic contribution of this paper. First, PV adoption is modeled by looking at consumer behavior and decision making. Second, PV adoption in the Netherlands is discussed to give more context on where the country stands today and the potential of PV adoption. Moreover, the current policies for residential solar adoption are summarized. Next, the relationship between socioeconomic attributes and PV adoption are discussed globally and locally. The non-examined factor low literacy is also discussed and this section ends with the academic contribution of this paper.

2.1 Modeling residential PV adoption

The main goal of this thesis is to better understand who is buying solar panels using socioeconomic factors. This information can be used to help policy makers and businesses active in the residential solar sector to accelerate PV adoption. Before analyzing which socioeconomic factors are related to a higher PV adoption, it is important to understand what factors influence the decision making of individuals related to PV diffusion in the first place. Therefore, this subchapter aims to better understand the relationships between individuals' adoption decision making and diffusion of innovations. Innovation diffusion theory and literature review on adoption drivers are combined to provide a framework on residential PV adoption.

Solar photovoltaic systems are able to convert sunlight into electricity and can easily be installed at any place with a decent amount of sunlight. For an extensive overview about PV technology, have a look at the article of Green (2000). Because electricity is generated using a renewable energy source rather than fossil fuel, this new technology is considered as a disruptive innovation (Sood & Tellis, 2011). A disruptive innovation is an innovation that usually creates a new market or enters an existing market eventually replacing market-leading firms or products (Bower & Christensen, 1995). Identifying solar PV as a disruptive innovation allows the use of innovation diffusion theory to better understand individual adoption decisions. A relevant theory in the context of PV adoption is the diffusion of innovations theory by Rogers (2003) that seeks to explain how and why new technologies are spread. Potential adopters of a new technology evaluate relative advantage, compatibility, testability, complexity and observability. These attributes interact and are judged as a whole. The five relative attributes of this theory provide a global framework to describe the adoption of innovations and within this framework multiple drivers for PV adoption are identified.

From a consumer behavior perspective, adoption of PV is driven by a combination of endogenous and exogenous factors. Endogenous factors could for example be knowledge of the technology and environmental awareness, whereas exogenous factors could be investment costs and characteristics of

the technology. Based on literature review, the main drivers for PV adoption by consumers can be categorized in the following factors: (1) problem awareness, (2) financial barriers, (3) non financial barriers and (4) social influence and (5) socioeconomic factors. These factors are shortly introduced and the following sections of this chapter will go into more detail about the socioeconomic drivers behind PV adoption. Problem awareness could influence the adoption decision of a household due to ignorance of the climate change challenge and the need to generate renewable electricity. If consumers are not aware of the environmental benefit of having solar panels, they are less likely to buy them (Vasseur & Kemp, 2015). Moreover, financial barriers such as investment cost and payback periode have a large influence on PV diffusion (Palmer et al., 2015). Also government support, feed-in tariffs and electricity price are considered to influence individuals' adoption decisions. Next, non financial barriers, such as knowledge about PV and relevant subsidies and policies are of importance within consumer decision making (Vasseur & Kemp, 2015). Consumers that are less aware about this technology and its benefits are less likely to invest in PV. Also the knowledge of relevant subsidies and policies is of importance and are found to be positive predictors in the willingness to adopt (Vasseur & Kemp, 2015). Fourth, social influence is of importance and previous research shows that social interaction effects are recognized as an important factor in the diffusion of new technologies (Bollinger & Gillingham, 2012). Consumers that are surrounded by other consumers that have adopted PV are more likely to invest in PV systems. Finally, many socioeconomic factors are linked to PV adoption and are further discussed in the next sections.

Above framework provides an overview of factors that drive PV adoption from a consumer behavior perspective. The main focus of this thesis is on the socioeconomic factors and their relationship with PV adoption. In order to increase PV diffusion, policy makers or businesses active in the solar energy sector could try to steer consumer behavior in the right direction by influencing the relative attributes of this innovation. To give an example, policy makers could try to overcome financial barriers by providing more subsidies on green investments. Moreover, policy makers could inform consumers about environmental awareness and the risk of climate change. To even better influence decision making of consumers, policy makers could benefit from a better understanding of which consumers value which factor as important. The analysis within this thesis can be used to better understand different types of consumers by looking at socioeconomic differences. This information can be used to target consumers with different strategies that respond to the characteristics of these groups.

2.2 PV adoption in the Netherlands

The last decades there has been raising awareness on the effect of humans on climate change. In order to overcome the great challenges that are ahead of us, every country has to contribute towards a more sustainable future. Despite the Netherlands being a relatively small country responsible for only 0.4% of the worldwide CO_2 pollution, it wants to take responsibility and be at the forefront of the energy

transition (European Commission, 2020). In order to make sure the Netherlands contribute and reach the Paris Agreements by 2030, the Dutch government makes a great effort to ensure the transition goes well by subsidizing the purchase of solar panels and introducing favorable renewable energy policies. This effort has paid off and the Netherlands has the highest number of solar panels per resident in Europe having more than two solar panels on average (DNE Research, 2022). In a global context only Australia has more solar panels per resident.

The Netherlands has reached this position by having a yearly PV growth rate between 30 to 60% for the last 10 years (DNE Research, 2022). By the end of 2021 1.5 million households had solar panels on their rooftops compared to 1 million a year earlier. These solar panels, combined with non-residential solar panels, have generated 11.2 TWh of solar energy and is therefore the biggest electricity source within renewable energy. Electricity generated from solar panels is responsible for 9.6% of the Dutch electricity supply.

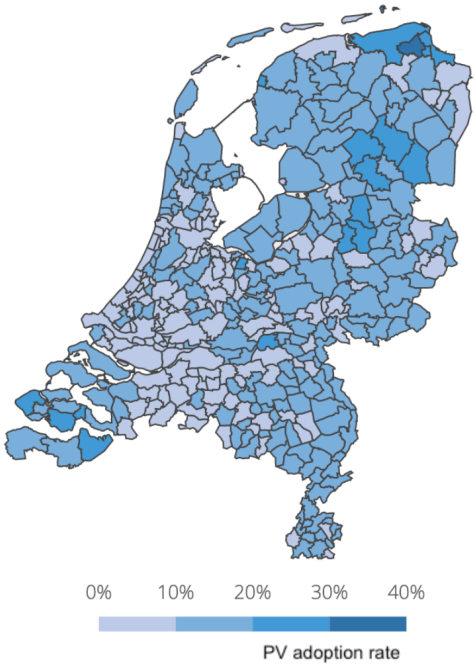

Figure 1. (left) Map of Dutch residential PV installation rates per township (Vattenfall, 2018)

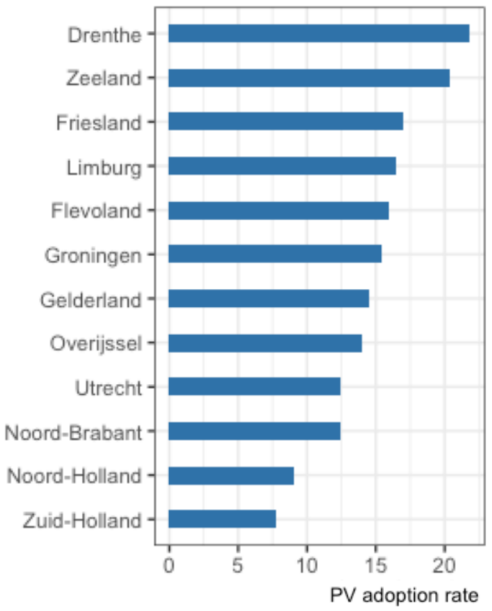

Figure 2. (right) PV adoption rate of Dutch provinces

Although millions of solar panels have already been installed on Dutch houses, there is still a long way to go before the country depends on sustainable energy sources only. The potential of PV is still very large as Deloitte (2018) conducted research showing that supplying every suitable roof with solar panels can supply 50% of the Dutch energy usage. Moreover, there is still disparity of PV adoption among different regions. Figure 1 shows the residential PV adoption rates for all the Dutch townships in 2018. Although the Dutch total adoption rate is relatively high, there is still a significant number of

townships with potential to increase PV adoption. Rotterdam and Amsterdam, the largest cities in the Netherlands, only have an adoption rate of 2% whereas smaller townships in less urban areas reach adoption rates up to 34%. Figure 2 shows the solar panel adoption rate for all 12 provinces. Drenthe and Zeeland are among the regions with the highest adoption rates above 20% whereas Noord- and Zuid-Holland are well below 10%. The average adoption rate for the Netherlands was 9% in 2018. Besides understanding which socioeconomic factors help explain PV adoption in the Netherlands, this paper also aims to better understand the regional disparity of residential solar adoption.

2.3 Government policies on PV adoption

The adoption of renewable energy sources is of interest to policy makers and national agencies that want to tackle global climate change. Policies aim to affect the reduction of greenhouse gas (GHG) emissions by aiding the adoption of renewable energy. Policies are implemented in a variety of ways like subsidies or legislation and can target different end users such as corporations or individuals. In the Netherlands, energy policies have been introduced to achieve the long term sustainability goals that are set by the government.

In 2007, the Dutch government had set ambitious goals to reduce 30% GHG emissions in 2020 compared to the levels of 1990 and a 20% energy coverage by renewable energy sources (VROM, 2007). After a few years, these targets have changed to 25% reduction of GHG emissions and only 14% of renewable energy due to infeasibility of the targets. A policy program called “Clean and Economical” (in Dutch: “Schoon en Zuinig”) was launched to reach the targets introducing a variety of policies focussing on both individuals and businesses. Although these policies did contribute to the energy transition, both targets have not been met. The reduction of GHG emissions in 2020 is 24.5% and the share of renewable energy is 11.1% (CBS, 2021a; CBS, 2021b). Moreover, the Covid pandemic contributed to the reduction of emissions due to strict lockdowns and less mobility and without this pandemic the GHG emission reduction would have been even less.

As 2020 has passed, the Dutch government has set new ambitious targets for 2030 and 2050 in accordance with the Paris Agreement. In 2050, at least 95% of the GHG emissions should be reduced compared to 1990. To ensure reaching this goal an intermediate goal has been set to reduce 50% of GHG emissions by 2030. These sustainability goals are very ambitious and require large transformations within almost every sector of the country. The last few years many policies have been introduced to ensure reaching these goals.

Policies introduced by the Dutch government to accelerate the adoption of renewable energy sources are focussing on both corporations and individuals. One of the major policies introduced to encourage residential renewable energy generation is a Feed-in Tariff program (in Dutch: salderingsregeling).

The Feed-in Tariff program ensures the uptake by energy suppliers of generated renewable electricity by residents for the same price electricity is bought from the supplier. This program ensures that residents can generate electricity for a reasonable price and can depend on the network of energy suppliers. The Feed-in Tariff program remains active until 2023 and is phased out until 2030 (Rijksoverheid, 2022). The phasing out of this policy is deployed because policymakers think it is no longer necessary due to cost reductions in the purchase of solar panels.

Another large policy introduced by the government to ensure PV adoption is the VAT discount on purchased renewable energy technology such as solar panels. This discount makes installation costs cheaper as the 21% VAT tariff is not applied on these products. A policy like the VAT discount gives an equal discount to everyone and thus therefore not specifically target groups that should be encouraged for buying solar panels. Other policies introduced are cheap loans for renewable investments and policies focussing on businesses and groups rather than individuals. Lastly, local governments are also able to subsidize 'green' investments and can introduce supplemental subsidies using their local budget. In the context of this research it is important to keep in mind that regional differences in extra subsidies can also be a driver for regional disparity of PV adoption and therefore be a limitation of the results. However, local subsidies have barely been issued the last couple of years due to extensive global policies.

The policies mentioned above all contributed to the adoption of PV in the Netherlands. According to Energie Centrum Nederland, a research institute focussing on renewable energy commissioned by the Dutch government, the policies introduced in the recent years have effectively contributed to a more lucrative cost-benefit when considering the purchase of solar panels (ECN, 2017). Especially the Feed-in Tariff and VAT discount policies have given higher incentives to purchase solar panels by reducing the payback period of these investments significantly. Although these policies have introduced a discount on PV investments for everyone, not every resident has made the decision to actually purchase solar panels for their dwelling. Moreover, there is still a large regional disparity among regions as shown in Figure 1. This raises the question who did buy solar panels and who did not. Relevant insights on the understanding of PV adoption can contribute to effective policies and targeting the right people to ensure PV diffusion keeps growing.

2.4 The socioeconomic factors behind PV adoption

Many studies have researched the explanatory role of socioeconomic attributes behind PV adoption and have demonstrated that there are differences among groups of people and regions. Previous conducted literature is examined below and will provide a solid foundation for which socioeconomic attributes should be included in this analysis and what relationships can be expected.

Raising awareness about climate concerns has emerged only in the last few decades and therefore research on the adoption of PV has started recently. One of the first large studies focussing on PV diffusion is conducted by Zhang et al. (2011) based on data of all prefectures in Japan between 1996 and 2006, considering a variety of attributes such as investment cost, regional policies, income and environmental awareness. This research concludes that governmental subsidies, environmental awareness and housing investments positively contribute to PV adoption whereas high investment costs are a barrier to purchasing solar panels. Sardianou & Genoudi (2013) analyzed the influence of characteristics such as gender, age, marital status and income on the adoption of solar panels in Greece. By using a probabilistic regression model the researchers found that well educated and middle-aged people are more likely to adopt energy sources that are renewable instead of fossil energy. In addition, marital status and gender are not significant attributes for explaining the adoption of renewable energy. Davidson et al. (2014) conducted research in California using a linear regression to understand which socioeconomic attributes predict PV diffusion. This study focussed on variables that had not been explored in previous literature and found that especially the number of rooms, heating source and house age explain PV adoption.

In 2014, Ameli & Brandt analyzed the determinants of households' investments in energy efficiency and renewables by using the OECD survey on household environmental behavior and attitudes. By performing binary logit regression models on this large survey data the researchers found that investments in clean energy mainly depend on home ownership, income, social context and households' energy practices. This study confirms previous research that homeowners and high-income households are more likely to adopt renewable energy sources than renters and low-income households. Balta-Ozkan, Yildirim & Connor (2015) have conducted research in the UK using spatial econometric methods in which cross-sectional data relating to geocodes is analyzed. This paper shows that PV adoption in a region is negatively correlated with housing density, the average number of households and the share of homeowners. Housing density as well as home ownership can make it more difficult to purchase solar panels as there is a lack of rooftop space or the rooftops are not owned by the residents themselves. On the other hand, the share of detached homes and education level are positively related to PV adoption.

Sommerfeld et al. (2017) are one of the first to analyze complex interactions between socioeconomic attributes by using classification and regression trees. The research is conducted in Queensland, Australia, and shows that household size and postal area are the most important factors for explaining PV diffusion. Moreover, the proportion of people aged over 55 and the proportion of homeowners are positively related to the purchase of solar panels. In the same year, Dharsing (2017) performed a spatial econometric analysis in Germany revealing that return on investment, income, education and thoughts on environmental change positively impact the adoption of solar panels. Attributes such as

unemployment rate and the construction of new buildings explain a lower adoption rate of solar panels in German counties. Sunter et al. (2019) combined satellite data on PV installations with American Community Survey data in the United States. By applying locally weighted scatterplot smoothing this research focussed on racial and ethnic differences and found that dominantly black and hispanic regions have a lower PV adoption rate. However, this disparity is often attributed to differences in household income and homeownership between ethnic groups and should be further researched to give more relevant insights on renewable energy adoption and ethnicity.

A recent study from Lan, Gou and Lu (2021) decided to use modern machine learning techniques to better understand unknown interactions and non-linear effects within PV adoption. Their research is conducted in Australia and makes use of conditional inference trees to identify different scenarios in which solar panels have been adapted or not. This paper found that regions with a low housing density are not attractive for PV promoters and therefore the adoption is low. Moreover, high density regions with high incomes also resulted in lower PV adoption as many residents were living in apartments and had a lack of rooftop space. On the other side, medium density regions with a middle level of household income have a higher PV adoption rate.

2.5 The socioeconomic factors behind PV adoption in the Netherlands

In the Netherlands less research is done on the explanation of PV adoption using socioeconomic attributes. In 2015, Vasseur & Kemp researched the adoption of PV in the Netherlands from a user perspective conducting a large survey. This study focuses on the adoption factors such as complexity of the innovation, social influence and grants and costs by constructing a theoretical framework. The researchers conclude that the most determining factor for PV adoption is the cost-reduction benefit of such investment. Moreover, adopters of PV are on average middle-aged, well educated, able to make their own decisions and think a good environment is important. On the other hand, PV rejecters on average have a lower income, are more dependent on others to make decisions and need more time for decision making. A second research on PV adoption in the Netherlands is conducted by Kausika et al. (2017). This research focuses on the Postal Code Rose Policy which enables residents to have solar panels placed on offices or other large rooftops instead of their own dwellings. The most important factors for PV adoption are income, house value, electricity consumption and neighbors with PV installations. Especially the last factor contributes to existing literature as social factors play an important role in considering PV adoption as well.

2.6 Low literacy and PV adoption

Although extensive research is done in the relationship between socioeconomic factors and PV adoption, there are still non-examined factors to further explore such as low literacy skills. Individuals with low literacy skills tend to need more time to read and understand information than others. These

individuals also have trouble with things like filling out forms, handling a smartphone and spelling. Despite the fact that this has not been researched before, there are multiple arguments that can substantiate a potential relationship between low literacy and PV adoption.

First, climate change is a recent phenomenon that involves a lot of scientific information. A thorough understanding of what climate change is and how this affects the world we live in today is reserved to only a small group of scientists involved in this topic. Moreover, in today's social media misinformation is a big challenge and also misinformation about climate change is often spread (Treen et al., 2020). Keeping up with the latest news and challenges about climate change already takes a lot of time for people having average or good literacy skills. Individuals having below average literacy skills might therefore have a hard time understanding what climate change is and why it should be prevented. Second, individuals having low literacy skills could have a lack of understanding of which policies are in place for PV adoption. Usually policies are difficult to understand because of the many exemptions and rules that are included. Although there are often user-friendly explanations of such policies from third parties, it is still important to find those explanations and knowing how to handle a smartphone or computer is crucial to obtain this information. This is often difficult for someone having low literacy skills. Both arguments could result in a lower PV adoption among low literacy groups and this relationship is further explored in this paper.

2.7 This paper

In subsection 2.4 examined research about the Netherlands already provides some understanding of the PV adoption in the Dutch residential housing sector. However, there is still potential in better understanding residential PV adoption using more data and more sophisticated methods to explore socioeconomic effects. Both the research conducted by Vasseur & Kemp (2015) and Kausika et al. (2017) have some limitations. The research of Vasseur & Kemp is performed on survey data of 817 respondents focussing on a wide variety of factors that are related to resident behavior. There are some disadvantages to using survey data like representativeness and uncertainty if the answers of respondents correctly represent their true behavior. The study conducted in the Netherlands by Kausika et al. does make use of a large dataset combining socioeconomic factors with spatial data to understand PV adoption. However, the scope of this research consists of one region in the Netherlands and focuses on the PV adoption by one policy of the many policies introduced.

Previous studies have shown that socioeconomic attributes do play an important explanatory role in understanding residential PV diffusion across the world. Although many researchers have done comprehensive research on socioeconomic factors, modern machine learning techniques have not been applied on a regular basis. The research conducted by Lan, Gou and Lu (2021) did make use of more sophisticated methods and resulted in interesting interaction effects between factors in order to

understand PV adoption in Australia. Although my paper has a relatively similar approach like the research conducted by Lan, Gou and Lu, there are several academic contributions compared to their research. My paper, unlike the paper of Lan et al., will make use of a numeric target variable for PV adoption which results in less loss of data and more accurate estimations. Moreover, my research includes the non-examined feature low literacy skills. The most important academic contribution of my paper is the data and region that are examined. In the Netherlands, research on PV adoption is scarce and is not yet conducted on the large amount of data available for this research topic. Within my paper research will be conducted on the Dutch residential PV market using state of the art machine learning. My study will also provide a constructive and systematic understanding of the interplay between a large number of socioeconomic attributes and PV adoption to inform energy policy makers on how to implement successful policies and who should be targeted to increase PV adoption.

3. Data

The next subsections contain the data collection, data pre-processing and an exploratory analysis. First, the different data sources as well as the selection of the final dataset are laid out. Second, the pre-processing stage of the data is further explained to prepare the data for analysis. The last subsection aims to explore the data and the underlying relationships independent of model choice.

3.1 Data description

The data at hand is a combination of three publicly available datasets provided by Statistics Netherlands (in Dutch: CBS) and Geletterdheid InZicht. Statistics Netherlands collects, processes and publishes a large number of reliable data sources that give more information about the country on a variety of topics, including energy usage and socioeconomic information. Geletterdheid InZicht is a research and expert center on the topic of literacy skills in the Netherlands. The first dataset contains information on PV installations and generated solar power from these installations for all neighborhoods in the Netherlands for 2019 (CBS, 2019a). This data is used to construct the target variable PV adoption rate, which is the percentage of households having solar panels installed. The second dataset provides a large number of socioeconomic attributes for every neighborhood in the Netherlands (CBS, 2019b). The data includes more than 100 socioeconomic attributes about the population, dwellings, energy, education, income and environment of these regions. Some examples of social and economic factors included in the data are gender, age, origin, household, built year of homes, ownership, social security, neighborhood, housing density and urbanity. However, due to privacy restrictions certain socioeconomic attributes are not provided for the smaller neighborhoods. This results in some variables having limited data available. These variables can not be included for analysis and are therefore not considered in the final dataset. The third dataset contains information about low literacy skills for every township in the Netherlands (Geletterdheid InZicht, 2020).

The datasets from Statistics Netherlands on socioeconomic and solar information both contain data on neighborhood level and are therefore exactly matched on unique neighborhood codes. Moreover, the third dataset about literacy skills is added to the data. This data is not available on neighborhood level but only for all 352 townships in the Netherlands. Therefore, multiple neighborhoods in the same township will have the same low literacy value. The merged datasets contain all neighborhoods in the Netherlands for 2019, which are more than 13.000 observations.

The dependent variable within this analysis is the PV adoption rate, which is the percentage of households having solar panels. This variable is constructed by taking the number of official registration of PV installations and dividing this by the number of dwellings in a certain neighborhood. The registration of solar panels is required by law and is also needed to make use of

the Feed-in Tariff. Not registering your solar panels leads to redelivery of solar energy to the energy provider without getting paid for it. Because of the legal requirement and the need for registering to redeliver solar energy, it is assumed that the number of registered solar panels approaches the actual number of total solar panels in the Netherlands.

Variable	Description	Mean	Range
PV adoption - **target**	Percentage of households having PV installed	0.126	0.001 ~ 1.000
People	Total number of people within neighborhood	1597	105 - 28750
Gender	Percentage male within neighborhood		
Elderly	Percentage 65 and older within neighborhood	0.192	0.000 ~ 0.879
Migration background	Percentage with migration background	0.231	0.000 ~ 0.935
Education rate	The rate of persons that received higher education at a university (in Dutch: universiteit & hogeschool)	0.229	0.000 ~ 0.856
Low literacy	Percentage of people with low literacy skills	0.115	0.010 ~ 0.320
Labour participation	Net employment rate	0.690	0.160 ~ 0.920
High income	Percentage of neighborhood receiving a high income	0.212	0.000 ~ 0.807
Social security	Percentage receiving social security	0.025	0.000 ~ 0.221
Urbanity	Urbanity score from 1 to 5 (1 = very urban, 2 = urban, 3 = moderate urban, 4 = little urban 5 = not urban)		
Household size	Average household size	2.175	1.100 ~ 4.000
Single-family homes	Percentage single family homes	0.656	0.000 ~ 1.000
Home value	Average home value (in thousands of €)	250	77 ~ 1985
Owned houses	Percentage of houses that is owned by the resident	0.582	0.000 ~ 1.000
Old houses	Percentage of houses that is built before 2000	0.839	0.000 ~ 1.000
Electricity use	Average electricity use (in kWh)	2754	810 ~ 6700

Table 1. Dependent and socioeconomic variables included for analysis

The independent variables included for analysis are carefully selected out of the large socioeconomic dataset. A couple of things are taken into account during the selection process. First, attributes are selected based on the review of residential PV studies in the literature review section to understand which factors may explain PV adoption. Second, the large dataset is reviewed to find additional attributes that could help explain PV adoption. However, many of the socioeconomic factors are closely related to each other and therefore strong correlations exist. Strong correlations among independent variables, also called multicollinearity, can be a problem for certain machine learning models as results are less reliable due to high standard errors. To ensure the reliability of this analysis the inclusion of highly correlated socioeconomic variables is avoided in the final dataset. Third, the non-examined factor low literacy is added to the subset of attributes. Note that low literacy skills is not available for every neighborhood but on the level of township. Less granular data could be a limitation in finding a relationship between PV adoption and literacy skills. Moreover, less granular

data could underestimate the actual effect of literacy skills if this effect is demonstrated. However, a proven relationship between literacy skills and the target variable can give relevant insights for policy makers and PV adoption. The socioeconomic factors and target variable included for analysis are displayed in Table 1.

3.2 Data pre-processing

After selecting the final set of variables, some transformations are done to prepare the data for analysis. The data at hand contains missing values due to privacy restrictions of certain socioeconomic attributes. Smaller neighborhoods have almost no data available and therefore neighborhoods with less than 100 residents are excluded from the data. After exclusion of these neighborhoods, there are still some smaller neighborhoods with many missing socioeconomic attributes. Neighborhoods with more than 8 missing attributes are excluded from the dataset because there is too little data available. Moreover, some neighborhoods have missing adoption rates and are also excluded from the data. After excluding these observations there are still a few attributes, mainly related to income, that have missing values. These missing values are replaced by the mean value of the attribute to ensure the machine learning methods are not affected by these values. After the data transformations the number of neighborhoods has decreased from 13.594 to 10.336, a reduction of 24.0%.

Moreover, the data contains a few outliers that should be reviewed before further analyzing the data. Outliers can occur due to variability in the data but can also be caused due to a measurement error. In the latter case, it is preferred to remove the outlier from the data. Outliers in the data can cause serious problems in statistical analysis and should be taken into account while deciding which statistical models are applied on the data. The data at hand contains many values that are considered as outliers as they are distant from the other observations. Figure 3a shows a boxplot of gender within the data. Most of the neighborhoods have a male/female ratio around 0.5 whereas a few neighborhoods are dominantly inhabited by either males or females. However, these observations occur due to variability in the data and are therefore true values. A closer look at the data tells us that neighborhoods dominantly inhabited by women are often small neighborhoods with a lot of elderly and an average household size between 1 and 2. Adding the fact that women live on average 3 years longer than men to this information could imply that these neighborhoods are inhabited by many widows and therefore results in this distribution. On the other hand, neighborhoods

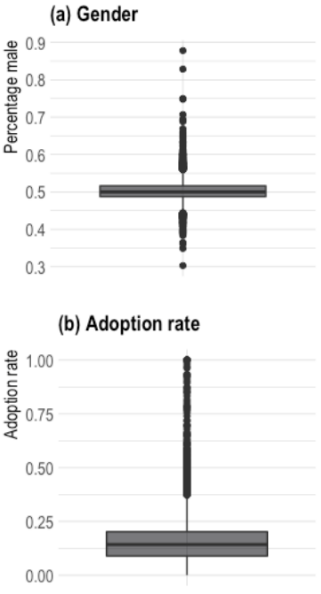

Figure 3. Boxplots for gender and adoption rate

dominantly inhabited by males tend to be small neighborhoods with a high percentage of people having a migration background and a low percentage of elderly. A possible reason for these neighborhoods having dominantly males could be immigrant workers living together, separate from their families. Because these outliers are true values, they are not excluded from the data and the outliers should be taken into account during the analysis. Many of the socioeconomic attributes included in the data have distant observations that occur due to variability in the data and are therefore not excluded.

The data also include outliers that are caused due to a measurement error. The target variable PV adoption contains some observations where the adoption rate is higher than 100%. The adoption rate should not exceed 100% and these errors occur due to the way this data is collected. Statistics Netherlands counts the number of official registrations of PV installations by adding up the registrations from multiple sources. This, in some cases, leads to double counts of PV installations and can not be avoided as the double count is already included in the primary data source. The neighborhoods with an adoption rate higher than 100% are therefore set to 100%. To ensure the data at hand is still reliable and fit for the analysis, multiple checks have been done on an aggregated level. The aggregated adoption rates are compared with reportings from Statistics Netherlands and external energy provider Vattenfall and both checks confirm that the data is accurate (CBS, 2019c; Vattenfall, 2022). Figure 3b shows the distribution of the adoption rate. The average adoption rate is 15.7% and only 1% of the neighborhoods have an PV adoption rate higher than 50%.

3.3 Data exploration

Table 1 in the data description section summarizes the dependent and independent variables included for analysis. Besides the explanation of the variables, the mean and the range of the attributes are included and give a first glance of the data. Most of the variables are included as a ratio between 0 and 1. The attributes number of people, average home value and average electricity use are numeric values and urbanity is an ordinal variable representing the urbanity of a neighborhood. Figure 4 shows a histogram of how the different levels of urbanity are represented in the data. 39.5% of the neighborhoods are classified as not urban, whereas the other categories are somewhat equally distributed. Figure 5 shows a heatmap displaying the correlation between all included variables. The categorical variable urbanity is added as a continuous variable as the categories are ordered. The dark blue squares between a row and column indicate a strong positive correlation between these variables whereas the dark red squares indicate a strong negative correlation.

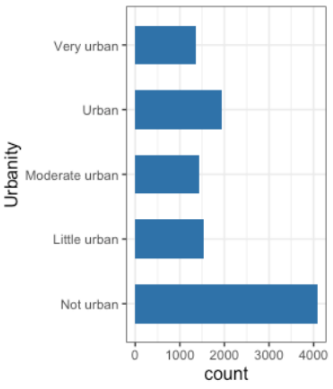

Figure 4. Histogram for urbanity

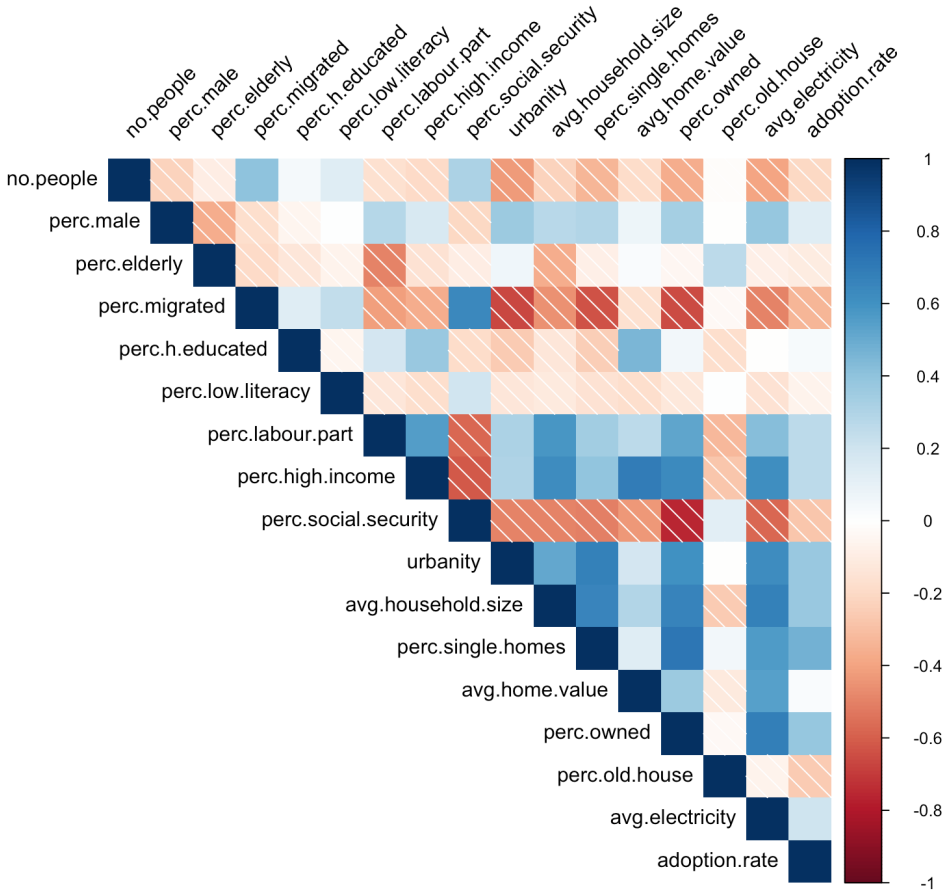

Figure 5. Heat map of correlations between variables including correlation color scale

Unsurprisingly many correlations are found among the socioeconomic attributes. Attributes that indirectly say something about wealth in a certain neighborhood are positively correlated with each other and are negatively correlated with variables that could imply poorness. Neighborhoods that are, for example, receiving a higher income are also more likely to have a higher labour participation, a higher percentage of single homes, a higher use of electricity and more owned dwellings. These neighborhoods are also more likely to have less people receiving social security and less people that have a migration background. Some strong positive and negative correlations are also found among socioeconomic attributes. The percentage of people having a migration background has a strong negative correlation with urbanity, the percentage of single-homes and the percentage of dwellings owned by the resident. Moreover, neighborhoods with many social securities are less likely to have a lot of homeowners so people are renting instead. Another example is the positive correlation between high income and home value, or homeowners and the percentage of single homes. The high correlations between independent variables might lead to collinearity problems. Although high correlations are avoided while selecting the attributes for this analysis, there are still a few relatively strong correlations between variables as described above. However, avoiding every correlation among

the independent variables is not desirable as this comes with a loss in predicting PV adoption using these attributes. The fact that a few independent variables are strongly correlated with each other is taken into account during the analysis.

Figure 5 also shows the correlation between socioeconomic attributes and PV adoption. Most of the included attributes tend to have a correlation with the target variable. Attributes like the percentage of homeowners, the number of single homes, less urban areas and larger households are positively correlated with a higher adoption rate. These positive relationships with the PV adoption rate are confirmed by previous literature and are also intuitive because neighborhoods having these characteristics have better circumstances for having solar panels (Balta-Ozkan et al., 2015; Dharsing 2017). Homeowners can more easily install solar panels on their rooftops compared to residents that are renting. Moreover, residents living in single homes have more rooftop space available than residents living in apartments and less urban areas tend to have more rooftop space available as well. Larger households could also be a proxy for house size and therefore be positively correlated with PV adoption. Next, income and labor participation have a positive relationship with PV adoption. Both implicate prosperity within a certain neighborhood and existing literature confirms that wealthy people tend to have solar panels installed more often (Sardianou & Genoudi, 2013). On the other hand, neighborhoods with a high percentage of people receiving social security, with a larger migration background and a larger number of houses built before 2000 are less likely to have adopted PV on a large scale. Neighborhoods with many social securities could display poverty and therefore have less solar panels. Moreover, research of Sunter et al. (2019) confirms that having a migration background is negatively correlated with PV adoption. Although this disparity is often attributed to ethnic differences in household income and home ownership, this study shows that after accounting for these differences a difference in PV adoption still remains. Within this study, the social dispersion effect is considered, meaning that people take less responsibility for their acts when they are in an environment where this is accepted. Last, older houses tend to have less solar panels installed. A reasonable explanation for this phenomenon could be that older houses are less isolated and therefore other options to increase sustainability are considered first before purchasing solar panels.

To further explore the underlying relationship between socioeconomic attributes and the adoption of solar panels, Figure 6 shows distributions of a few attributes for different adoption rates. For ease of interpretation, the adoption rates are categorized as low (below 10 %), medium (between 10-30%) and high (above 30%) PV adoption. Figure 6a shows that the percentage of owned houses is a lot lower for neighborhoods with a low PV adoption, whereas neighborhoods with a medium or high PV adoption have a similar percentage of house owners on average. The same relationship is found for the number of single homes in Figure 6b. Neighborhoods with a low PV adoption tend to have a much lower share of single homes compared to neighborhoods with medium or high PV adoption. The

difference among adoption rates is a lot smaller for the percentage of elderly and the percentage having low literacy skills in a neighborhood (see Figure 6c and 6d). Figure 13 in appendix A1 shows the distributions of all the other attributes for different adoption rates and provides a good overview of potential relationships with the target variable. For the sake of brevity, these relationships are further analyzed in the results section, after machine learning models have been applied to the data.

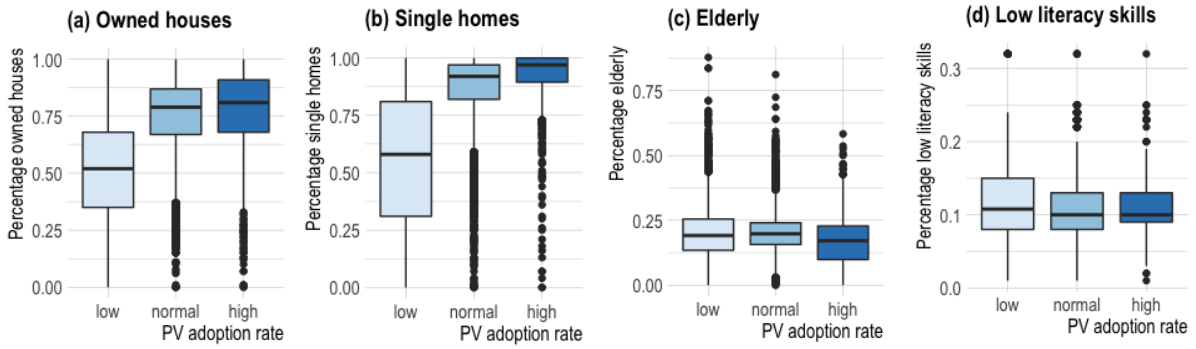

Figure 6. Comparison of socioeconomic variables among different adoption rates (low <= 10%, 10% < normal < 30%, high >= 30%).

4. Methodology

This section explains which methods are used to answer the research question. A large variety of methods is available as machine learning is rising in popularity among practitioners and researchers. To ensure the right machine learning models are applied to the data it is important to keep in mind the main goal of the analysis. To best answer the research question, interpretability is of great importance to understand the relationships between socioeconomic attributes and PV adoption. Second, the quality of prediction is important to ensure the relationships are as accurate as possible. Therefore, the main focus lies on interpretable machine learning techniques that fit the data well. In the next subsections, model selection, model explanation and model evaluation are discussed.

4.1 Model selection

The baseline method is a multiple linear regression due to simplicity and ease of interpretation. Because the target variable is a ratio between 0 and 1 and therefore does not follow a gaussian distribution, the linear model is extended to a generalized linear model (GLM). However, using a linear model brings along strong assumptions on the data related to linearity, variance, outliers and multicollinearity. (James et al., 2013, p. 90-102). If these assumptions are not met the prediction quality can be poor and interpretation of the models might be biased. Moreover, the generalized linear model does not take into account non-linear relationships and interaction effects between variables. Therefore, the second method introduced is a decision tree. Decision trees can be used for both classification and regression problems and allow the model to include nonlinear features and interaction effects. Moreover, decision trees are non-parametric and therefore, unlike linear models, do not make any assumptions about the characteristics of the data. Both linear models and decision trees are simple interpretable machine learning models. However, if the data structure is more complex, these models are less suitable and tend to underperform compared to advanced machine learning models. To account for this, the third method introduced is random forest. A random forest is an ensemble learning method that constructs many trees and takes an average prediction of these trees. Using average prediction often leads to a great reduction of variance in comparison to above models and is great in dealing with noisy data. Unfortunately, the increase in prediction power comes along with a decrease in interpretation of the model. Interpretation methods can be applied to the model to increase interpretability but offer less insights compared to simple models such as linear regression and decision trees.

Besides the models introduced above, there are many more machine learning techniques available. Preliminary analysis is done on a variety of techniques to check whether they could improve interpretation and/or prediction on the data at hand. Related to interpretable machine learning, sparse linear models and generalized additive models are fitted on the data (Friedman & Popescu, 2008;

Hastie & Tibshirani, 2017). However, these models did not have added a significant value in terms of explainability and are therefore not further examined. Related to more complex machine learning models, gradient boosting and an artificial neural network are fitted on the data to find out if prediction accuracy could be increased. Although just a preliminary analysis is performed, these models did not outperform the well performing random forest technique and are therefore not included in this analysis.

4.2 Generalized linear model (GLM)

A generalized linear model is an extension of a simple linear regression. A linear regression estimates a relationship between one or more variables and the target variable. The ordinary least squares (OLS) method is used to find the coefficients that minimize the sum of squared residuals (RSS). The learned relationships are linear and the OLS regression looks as follows:

$$\hat{\gamma} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \ldots + \widehat{\beta_p} x_p \quad (1)$$

with $\hat{\gamma}$ being the estimation of the target variable, the betas $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$, ..., $\widehat{\beta_p}$ represent the estimated coefficients and x_1, x_2, ..., x_p being the predictors. The residual sum of squares is calculated as follows:

$$RSS = \sum_{i=1}^{n} (\widehat{\gamma_i} - \hat{\gamma})^2 = \sum_{i=1}^{n} (\gamma_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \widehat{\beta_2} x_{i2} - \ldots - \widehat{\beta_p} x_{ip})^2. \quad (2)$$

The linear regression model assumes that the target variable given the input features follows a Gaussian distribution. In this analysis, the target variable does not follow a Gaussian distribution as the target variable is a ratio between 0 and 1. Using a linear regression could therefore result in predictions that are negative or above 1 and this limits the interpretation of the model. By extending the model to a GLM, the weighted sum of features are still used, but the outcome allows other distributions including one that meets the requirement of predicting between 0 and 1 (Faraway, 2016). GLMs consist of three components: a link function, the OLS regression and a particular distribution from the exponential family (Molnar, 2020). The use of an alternative distribution is connected through a link function and the GLM looks as follows:

$$g(E(\gamma|x)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \quad (3)$$

$E(\gamma|x)$ is the expected value of γ using a particular distribution from the exponential family (Molnar, 2020). The exponential family is a parametric set of distributions that can be applied on the same formula and allow the outcome to have a certain type of distribution depending on what is desired.

Whereas the Gaussian distribution is said to be normally distributed, other distributions can be used for mathematical convenience. The Bernoulli distribution, for example, is a discrete probability distribution which is used when the outcome variable should simply be yes or no. For the GLM fitted in this paper a beta distribution is preferred as this family follows a continuous probability distribution (Gupta & Nadarajah, 2004). A GLM with a beta distribution (also called: beta regression) is used when the desired outcome should be within a certain interval, such as a ratio or percentage. In this analysis the target variable is the PV adoption rate and therefore the beta distribution is suited. The beta distribution is defined on a certain interval and shaped by the two positive parameters s and q. In this GLM, the predictions on the target variable should be between 0 and 1 and therefore the beta distribution is set to $[0, 1]$. The formula of the beta distribution is as follows:

$$\frac{x^{s-1}(1-x)^{q-1}}{B(s,q)} \qquad 0 \leq x \leq 1;\ s, q > 0 \tag{4}$$

where $B(s, q) = \frac{\Gamma(s)\Gamma(q)}{\Gamma(s+q)}$ and $\Gamma()$ is the complete gamma function (Artin, 2015). The gamma function shapes the beta distribution within the defined interval. The parameters s and q are optimized by performing a maximum likelihood which is included in the R package that is used for estimation.

The corresponding link function $g(.)$ of this probability distribution is the logit function (Cribari-Neto & Zeileis, 2010). The logit function is used to model the outcome as a function of covariates. The purpose of this function is to take a linear combination of covariates and convert these values to a ratio between 0 and 1. Therefore the logit link function is suited to link the beta distribution and the OLS regression. The logit function with p being the expected value $E(\gamma)$ is defined as:

$$logit(p) = \sigma^{-1}(p) = ln(\frac{p}{1-p})\ for\ p \in (0, 1) \tag{5}$$

Another challenge that is addressed besides the distribution of the target variable is the selection of independent variables included in the model. To prevent multicollinearity and noise in the output, redundant variables are left out of the GLM. To do so, backward selection is performed using the Akaike Information Criteria (AIC) (Sakamoto et al., 1986). The AIC deals with the trade-off between the goodness of fit and the simplicity of a model and takes into account both the risk of overfitting and the risk of underfitting. After the full model is estimated, the variable that leads to the biggest decrease in AIC is removed. Variables are removed until the AIC is minimized resulting in a model with only relevant attributes.

The generalized linear model performs very well if the relationships between the variables are linear. However, if the relationships are more complicated, non-linear or noisy, performance can decrease drastically. More specifically, a number of assumptions should be met before this model is a good fit on the data (James et al., 2013, p. 90-102). First, a linear relationship is assumed between the predictors and a target variable. Second, the variance of the error terms should be constant. However, some variance within the error terms is inevitable because this is commonly observed in data taking values in a standard unit interval, such as proportions or rates. Often more variation is observed around the mean and less variation is observed around the upper and lower limits of the unit interval, which is between 0 and 1 in this analysis. To prevent variance in error terms a logit link function is often used and ensures this assumption holds (Cribari-Neto & Zeileis, 2010). Third, the error terms should not be correlated and fourth, the OLS is very sensitive to outliers and high-leverage points. Finally, multicollinearity can be a problem for finding the right coefficients and therefore interpretation of the model could be limited. If all these assumptions are met, the model is accurate and the variables can be interpreted (Poole & O'Farrell, 1971).

4.3 Decision tree

The GLM is a simple, yet effective way to model the relationship. However, many assumptions should hold and interaction effects as well as nonlinear effects are not included. If either the assumptions don't hold or the relationships among variables are complex, the use of a non-parametric model is often preferred. A simple and well performing non-parametric model is the decision tree (Myles et al., 2004). A decision tree is a tree-based method that segments the predictor space in regions which are distinct, non-overlapping and high-dimensional rectangles. These regions are denoted as X_1, X_2, …, X_p and defined by a set of splitting rules R_1, …, R_J. All observations in these regions are given the same prediction. In this dataset the most important variables for predicting PV adoption could make up the splitting rules that define the prediction of PV adoption. To give an example, if the percentage of house owners in a neighborhood is above a certain threshold, the prediction outcome on PV adoption will be higher. Because the decision tree makes use of splitting rules instead of linear relationships, nonlinear features are introduced in modeling the data.

More technically, decision trees are fitted by using recursive binary splitting. Because it is computationally infeasible to consider every region partition, recursive binary splitting takes a top-down and greedy approach to find the best splits minimizing the sum of squared residuals (RSS) (James et al., 2013, p.306). This approach is top-down because only one split at the time is considered and greedy since it does not consider future divisions of the predictors' space. Moreover, the tree is pruned to avoid overfitting meaning that a smaller subtree is considered to find the regions with the

best quality of split. This is done using a complexity parameter cp, which controls how many splits are included. The equation to find the regions that minimize the RSS is defined by:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (\gamma_i - \hat{\gamma} R_j)^2 . \quad (6)$$

Decision trees take into account non-linear and interaction effects while still being very easy to interpret (Myles et al., 2004). If the relationships between variables are not too complex, decision trees tend to perform very well and often outperform linear regression models. However, if the data structure is complex and noisy, decision trees are less capable of making well predictions. Moreover, decision trees generally have high variance (James et al., 2013, p.315). Therefore, an extension of a simple decision tree is included, called random forests (Breiman, 2001).

4.4 Random forests

The third method applied is a random forest (Breiman, 2001). A random forest grows many decision trees, aggregates the trees and makes an average prediction. This is done by building a number of unpruned trees B that are grown out of bootstrap samples from the data (James et al., 2013, p. 316-322). Bootstrap aggregation, also called bagging, is a statistical learning method that reduces overfitting and variance by drawing samples with replacement. Random forests are similar to bagging except that a subset m of p predictors are taken along per bagged tree. The advantage of using a subset of predictors is that this decorrelates the trees which further reduces variance in the aggregated prediction. If, for example, a variable has a few large outliers, these outliers are not taken along in every subtree because a subset is taken. This makes the model less sensitive to outliers which reduces the variance.

Two parameters should be tuned while fitting a random forest. The first parameter B is the number of trees grown and included in the random forest. When B is equal to 1, the random forest is similar to a simple decision tree and works the same way. If the number of trees increases, the random forest takes an average prediction of the grown trees. Fortunately, increasing the number of trees does not lead to overfitting and therefore it is only important to set the number of trees sufficiently high. Second, the number of parameters considered for every split, denoted as m, should be defined. The out of bag error is used to validate which m gives the best predictions out of 1 to m possibilities. In this analysis 16 parameters are included and therefore the out of bag error is used to validate how many parameters should be included for every subset of which a tree is grown.

As mentioned above, random forests often lead to a great reduction of variance compared to decision trees or linear regression models and are great with handling noisy data (James et al., 2013, p.

319-321). Influential outliers in the data are not considered in every tree because a subset is used. Moreover, multicollinearity is reduced, because only a subset of predictors is considered for every grown tree. A disadvantage of using random forests is interpretability. Because many trees are grown, there is not just a set of decision rules to interpret or a simple decision tree visualization. However, there are some interpretation methods that could be used such as feature importance and partial dependence plots (Fisher et al., 2019; Friedman, 2001; Molnar, 2020).

The feature importance is calculated by permuting a feature from the model and taking the average decrease in accuracy over all the trees (Fisher et al., 2019). A feature is defined as "important" if permuting the values increases the model error because the model relies on this feature. However, the feature is defined as "unimportant" if permuting the feature barely affects the model error. Moreover, partial dependence plots illustrate the marginal effects of a certain feature x_s in the model (Molnar, 2020). A partial dependence plot shows whether the relationship between the feature and the target variable is linear or more complex. Partial dependence plots are calculated as follows:

$$\widehat{f}_s(x_s) = EX_C[\widehat{f}(x_s, X_C)] = \int \widehat{f}(x_s, X_C) dP(X_C). \quad (7)$$

Vector x_s is the plotted feature for the dependence function and X_C the other features of model $\widehat{f}$. The feature vectors x_s and X_C combined are the total feature space x (Friedman, 2001, p. 29-36). For more details, have a look at the research of Friedman (2001). All in all, random forest is a strong predictive model being more difficult to interpret (Breiman, 2001; Molnar, 2020).

4.5 Model evaluation

Model evaluation is important to understand the performance of the models and compare them. First, the data is split up into a train (80%) and test (20%) set whereas the test set is used to evaluate the model performance. In this analysis, a regression problem is solved and a sufficient metric to evaluate is the root mean square error (RMSE). The mean square error is calculated by the sum of square prediction error and gives an absolute number on how much the predicted results deviate from the actual numbers. The lower the RMSE, the better the performance of the model. RSME is calculated as follows:

$$RMSE = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (\gamma_i - \widehat{f}(x_i))^2} \quad (8)$$

with n observations and $\widehat{f}(x_i)$ being the prediction of $\widehat{f}$ for observation i.

5. Results

In the results section, different outputs of the machine learning models are interpreted and analyzed. First, the models are estimated and their performance is compared. Second, the output of the models are analyzed to understand predictor influence, including nonlinearity and interaction effects. Third, regional differences are compared and finally the stability, reliability and robustness of the results are discussed. Although this analysis focuses on the relationships between attributes rather than why these relationships exist, some suggestions are done based on existing literature and logical reasoning.

5.1 Model estimation & performance

A total of three models is estimated using *R* statistical software. Table 2 shows the properties and test RMSE scores of the estimated models. The generalized linear model is estimated using the $betareg$ package (Cribari-Neto & Zeileis, 2010). Before the GLM is estimated, the data is standardized to ensure the features have the same scale and contribute equally. Standardizing the data is important for regression models to avoid bias due to difference in scale. This model follows a beta distribution and uses a logit function. Moreover, AIC stepwise selection is performed and in the final model 14 of the 16 attributes are included. The attributes percentage male and percentage high income are excluded from the model because they add no significant value according to the AIC. The decision tree model is estimated using the $rpart$ package (Therneau et al., 2015). The decision tree is pruned with cp is 0.006 as complexity parameter which results in a tree with 12 final nodes. The last model estimated is a random forest using the $randomForest$ package (Liaw & Wiener, 2002). The number of trees B is set using the out-of-bag error displayed in Figure 14 in appendix A2. The out-of-bag error of the random forest becomes stable after growing a few hundred trees and therefore setting $B = 1000$ is more than sufficient. Second, parameter m is set using the out-of-bag error for different values of m and is displayed in Figure 15 in appendix A2. m is validated out of 1 to p predictors and $m = 7$ leads to the lowest out-of-bag error.

Model	Properties	RSME
Generalized linear model	dist = β, link = logit, AIC stepwise selection, p = 14	.0859
Decision tree	cp = 0.006, nodes = 12	.0910
Random forest	B = 1000, m = 7	.0795

Table 2. Properties and test RSME scores of the GLM, decision tree and random forest.

The best way to assess the predictive power of a model is to put it to a test. In this analysis, 20% of the data is not used for model estimation leaving 2067 observations for testing the models. For each method, the root mean square error (RMSE) is calculated over the test data set and these estimations

are displayed in Table 2. The lower the root mean square error the better the prediction power of the model. The model with the highest RMSE score is the decision tree being 0.0910. The GLM model outperforms the decision tree with about 5% having an RMSE of 0.0859. The random forest proves to generate the strongest prediction having an RMSE of 0.795 and outperforms the decision tree with about 13%. Although there is some difference between model performance, all three models tend to show relatively similar performance. To best answer the research question, interpretation of the results is most important and therefore interpreting all three models is preferred to get the most relevant findings. Moreover, interpreting the models all together can strengthen the findings of this analysis by doing a double check and provide more robust results.

5.2 Predictor influence

This paragraph reviews the output of the models to understand predictor influence, including nonlinearity and interaction effects. First, the generalized linear model is reviewed, followed by the decision tree and the random forest.

5.2.1 Generalized linear model results

Table 3 shows the estimations of the generalized linear model. Because the data is standardized before fitting the model, the units of measurement are eliminated making it possible to compare the estimates and get a sense of the importance of these variables. The most important variables will have the highest absolute values of the standardized coefficients. Unfortunately, using standardized data makes interpretation of single estimates more difficult because it is less intuitive. The standardized coefficients are measured in units of standard deviation. The beta value of 0.085 of the feature elderly, for example, indicates that a change of one standard deviation in the independent variable results in a 0.085 standard deviation increase in the target variable. Thus, the actual effect of a single estimate depends on the standard deviation of both the feature and the target variable. An attribute with a positive estimate leads to a higher PV adoption rate when there is an increase, whereas an attribute with a negative estimate leads to a lower PV adoption rate when there is an increase.

The most important predictors leading to a higher PV adoption rate when they increase are the percentage of single family homes, urbanity of the neighborhood, education rate and household size. Please note that an increase in urbanity score means that the area is less urban. The effect of these attributes is intuitive and in line with existing literature on PV adoption examined in other countries (Balta-Ozkan et al., 2015; Sommerfeld et al., 2017). The positive effect of single-family homes makes sense as this factor could imply that there is more rooftop space available to install solar panels because less people live under the same roof. Moreover, less urban neighborhoods might also have more rooftop space available due to a lower housing density. The positive relationship of household size and PV adoption could also imply that larger households have a larger house and therefore more

rooftop space available. Education rate is also a strong predictor for PV adoption. A possible cause for this relationship could be that higher educated people are better informed about climate change and therefore invest in sustainable energy (Treen et al., 2020). Moreover, higher educated people tend to have a higher income and are therefore able to invest in solar panels (Griliches & Mason, 1972). All in all, the relationships of these predictors are in line with existing literature and this analysis shows that the same relationships occur in the Netherlands.

Model output

Predictor	Estimate
Intercept	-1.725***
People	-0.024**
Elderly	0.085***
Migration background	-0.018
Education rate	0.219***
Low literacy	-0.010
Labour participation	0.040***
Social security	0.065***
Urbanity	0.138***
Household size	0.123***
Single-family homes	0.450***
Home value	-0.139***
Owned houses	0.067***
Old houses	-0.187***
Electricity use	-0.159***

Note. * p < 0.05, ** p < 0.01, *** p < 0.001

Table 3. (left) Model output of the GLM

Feature importance

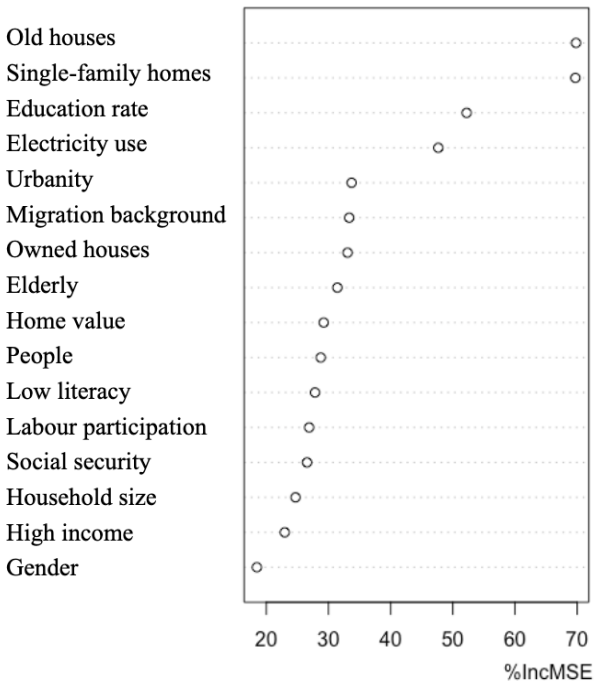

Figure 7. (right) Feature importance of random forest including all features

The most important predictors leading to a lower PV adoption rate when they increase are percentage of old houses, electricity use and home value. The negative relationship between the percentage of old houses in a neighborhood and PV adoption supports existing literature on this occurrence (Davidson et al., 2014). Residents having a relatively old house might spend their money on alternative ways to decrease energy consumption such as isolation before considering solar panels. The negative relationship of electricity use and PV adoption could be explained by the fact that a resident with lower electricity usage already has solar panels installed and therefore needs less electricity from the energy supplier. The negative relationship between home value and PV adoption has not been examined in existing literature and seems less intuitive. A possible reason for this effect could be that neighborhoods with a higher average home value are more urban neighborhoods in which house

prices are higher and rooftop space to install solar panels is more scarce. However, further analysis should be done to better understand if and why this relationship exists.

The model output of the GLM also shows that there are some attributes without a significant estimate. Besides the attributes high income and gender that are left out of the model during model selection, the percentage of residents having a migration background and the percentage of residents having low literacy skills do not affect the PV adoption rate within this model. The relationship between ethnicity and PV adoption in the United States that is examined in previous literature is not found in this analysis and could imply that there is no such effect in the Netherlands (Sunter et al., 2019). The absence of such an effect could be, for example, due to the fact that the Netherlands welcomes a variety of immigrants for both low-skilled and high-skilled labour. High skilled immigrants might be more aware of climate change and have a higher income which could both stimulate PV adoption. Moreover, the non-examined feature low literacy skills does not affect PV adoption in this analysis. There could be a few reasons why this effect is not found. First, it could simply be that such an effect does not exist in reality. The arguments specified in Chapter 2.5 are mostly based on logical reasoning and have not been scientifically proven before. Second, it could be that a relationship between literacy skills and PV adoption does exist but is not found in this analysis. The data for this feature was only available on township level instead of neighborhood level and this could limit finding the actual relationship between literacy skills and PV adoption. Further research should be done to better understand the relationship between literacy skills and PV adoption.

5.2.2 Random forest results

Figure 7 shows the feature importance calculations of the random forest model including all variables. Because the random forest model generates the best predictions on the data, these feature importance measures are used to confirm the relationships and the importance of these relationships between the attributes and PV adoption. Similar to the output of the generalized linear model, it suggests that house age, single-family homes, education rate and electricity use are the most important features. Moreover, the least important features are gender and income which is also supported by the GLM as they are not considered due to irrelevance of these features. The generalized linear model and random forest confirm the robustness of the results as these models consider the same features as important and the same features as unimportant.

The generalized linear model performs well on the test data and provides relevant information. However, the GLM does not take into account nonlinearity. The random forest fitted on the data does allow nonlinear relationships and the fact that this model performs better on the test data could imply nonlinear relationships exist. To move beyond linearity, the random forest model is further explored using partial dependence plots. Figure 8 shows the partial dependence plots of the four most important

predictors. Partial dependence plots illustrate the marginal effects of a certain feature in the model (Molnar, 2020). These plots can be used to derive the expected PV adoption rate for different values of a certain feature. Note that partial dependence plots should be interpreted with care. These plots only capture the main effect of a feature and ignore possible interaction between features. Moreover, regions of the plotted feature with almost no data should not be interpreted and are therefore not included in the models. For example, in Figure 8a only values of the percentage of old houses between 50% and 100% are plotted. For this feature there are very few neighborhoods having less than 50% of old houses and therefore this part is not interpreted.

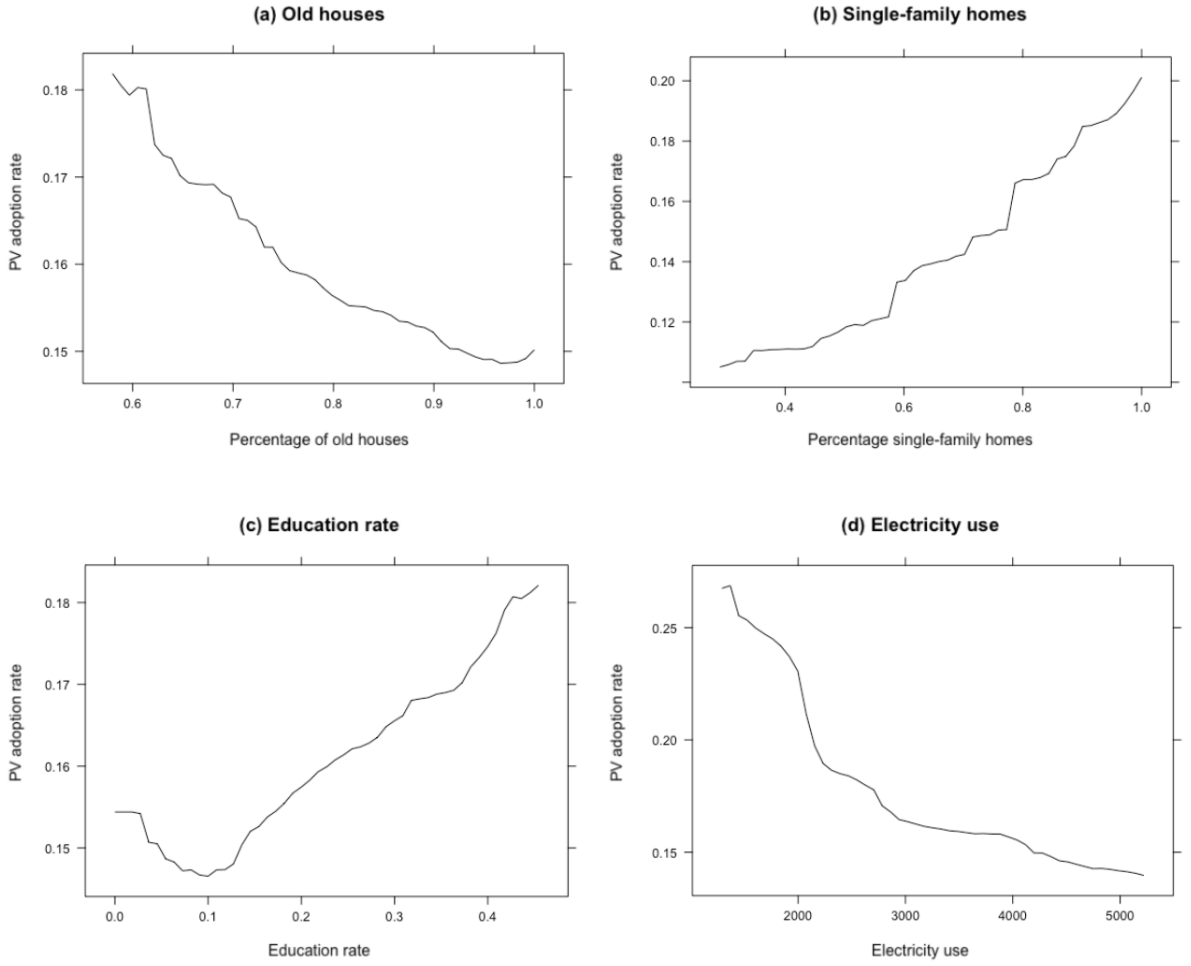

Figure 8. Partial dependence plots for the four most influential predictors of the random forest model

In Figure 8a the percentage of old houses is plotted and shows the PV adoption rate for different values of this feature. Most observations of this feature are between 50% and 100% and therefore only this part is interpreted. A higher percentage of old houses is correlated with a lower PV adoption rate and this relationship tends to be almost linear according to the partial dependence plot. Figure 8b shows the percentage of single-family homes and should not be interpreted for values below 30%. The feature single-family homes and PV adoption rate have a linear positive relationship.

The education rate, presented in Figure 8c tends to have a U-shaped relationship with PV adoption at first glance. However, the values of education rate between 0% and 10% should not be interpreted because there are too few observations within this range. Therefore, the education rate also tends to have a linear relationship with PV adoption for values between 10% and 50%. Figure 8d shows the partial dependence plot for electricity use and this feature has a somewhat convex relationship with PV adoption. Nevertheless, the curvature is small and the relationship is close to a linear relationship. All in all, these plots show that there exists some nonlinearity in the data, but not much. The fact that there is some nonlinearity could explain why the random forest outperforms the generalized model with just a few percent. However, the nonlinearity is marginal and this substantiates that interpreting the coefficients of the GLM is reasonable because the features tend to have a somewhat linear relationship with the target variable.

5.2.3 Decision tree results

Next the data is further analyzed to find out if interaction effects between features exist. The generalized linear model does not take into account interaction effects. The random forest model does allow interaction effects to exist, but these effects are difficult to extract from the data as many trees are summarized. The decision tree is considered to explore interaction between features. Figure 9 shows that the optimally pruned tree using recursive binary splitting has 12 final nodes (James et al., 2013, p.302-311). The plot shows the decision rules of the tree and the final nodes including the PV adoption rate (top number) and what % of all the neighborhoods are within each node (bottom number). The first and therefore most important decision rule is the percentage of single-family homes being above or below 79%. The next two splits are about the percentage of single-family homes and the percentage of old houses in a neighborhood. Once again, the decision tree confirms that these two attributes are most important for explaining PV adoption. Moreover, the decision tree plot shows that there are multiple interactions between these two attributes resulting in different predicted values for PV adoption. On the left side of the decision tree plot, the neighborhoods are splitted by a percentage of 51% for single homes and a percentage of 48% for old houses. On the right side of the plot these attributes have splitting values of 42% for old houses and 94% for single homes. Generally speaking, the prediction for PV adoption is higher for neighborhoods with more single-family homes and less old houses. The final set of decision rules leads to an uneven distribution of the neighborhoods over the 12 final nodes. Node 5, for example, contains 33% of the observations whereas Node 12 contains less than 1% of the observations. Especially the nodes with a high predicted adoption rate have few classified neighborhoods. This occurrence does not limit interpretation of the model but does provide information on which nodes are more relevant for interpretation.

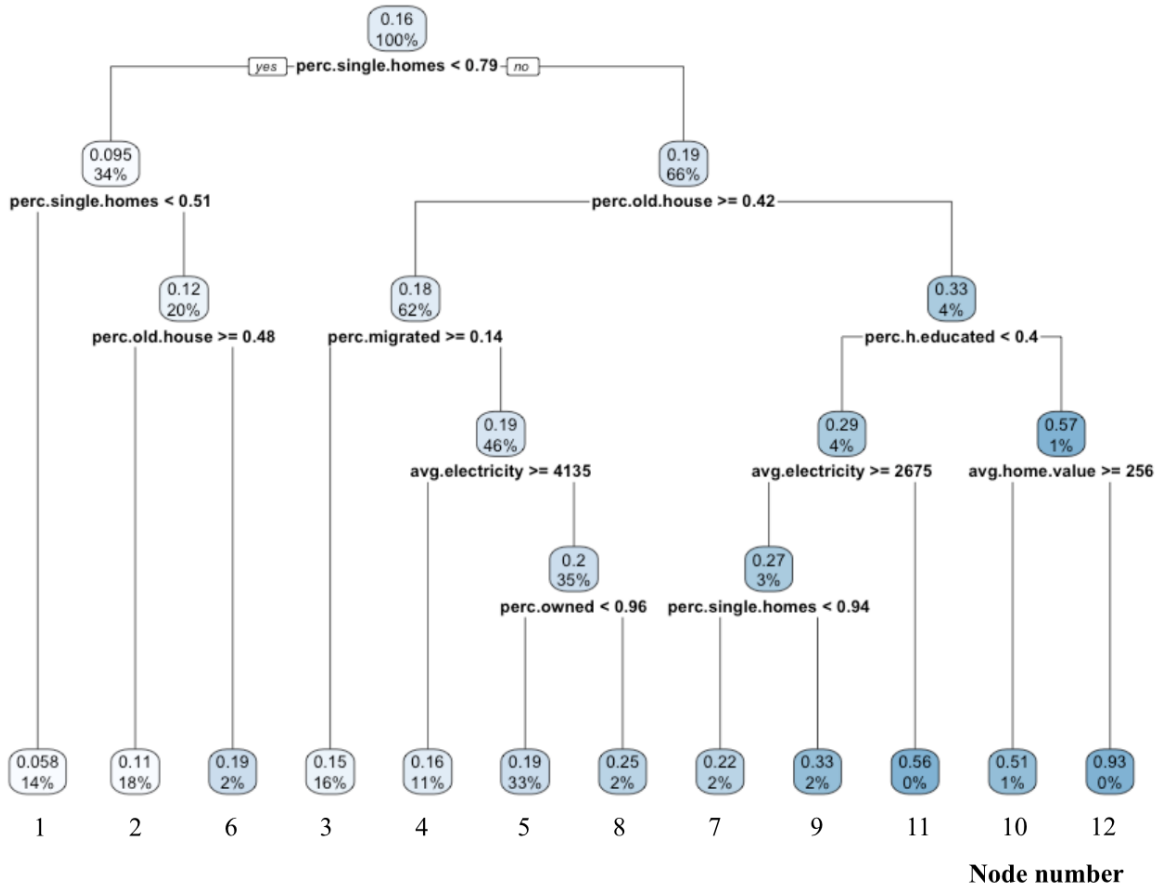

Figure 9. Decision tree plot with 12 final nodes which are ordered on PV adoption rate

Another advantage of interpreting the decision tree is that this model groups the neighborhoods based on a set of decision rules and provides a predicted value for these groups. This makes it possible to compare the difference between groups and find out how this is related to a different PV adoption rate. Moreover, these insights provide relevant information for policy makers on who is buying solar panels and who is not, which is further discussed in the next chapter. Figure 10 shows a heatmap considering all numeric attributes and the 12 final nodes from the decision tree. The heatmap provides a color scale for every attribute to visualize the difference in value for the different nodes. A dark blue square presents a relatively high value for the attribute whereas a light blue square presents a relatively low value for this attribute. Moreover, the nodes are ordered on PV adoption rate and the rates are given on the right side of the figure. Also the node size is included which represents all neighborhoods classified to their corresponding nodes according to the decision tree model. Some of the nodes contain considerably less observations and are therefore less relevant to interpret compared to nodes containing a large number of neighborhoods. Interpretation of the heatmap should be done with care because the color-scale provides relative difference of an attribute and absolute difference could therefore still be low.

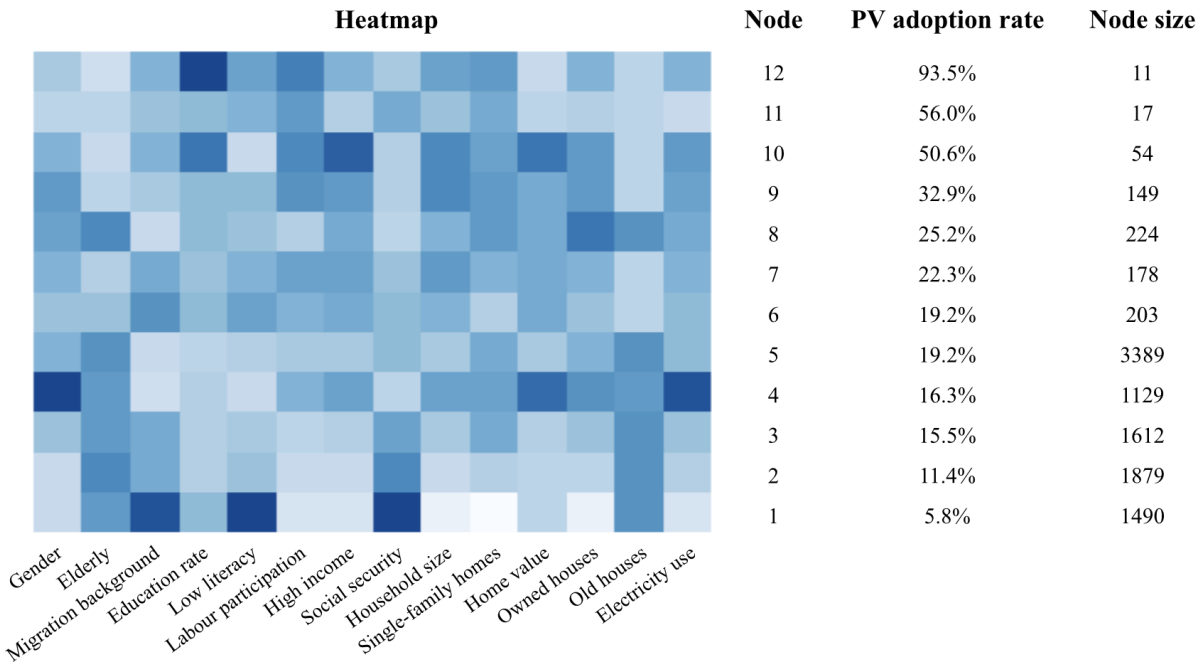

Figure 10. Heatmap showing the relative values of the attributes for the different nodes, the corresponding PV adoption rate and the node size for all neighborhoods.

The node with the lowest PV adoption rate (Node 1) is characterized by a relatively large proportion of people having a migration background, low literacy skills and more social security is provided. Moreover, these neighborhoods consist of relatively few single-family homes and few owned houses. These characteristics could reflect poorness in these neighborhoods. Moreover, residents might not speak the Dutch language well or have below average literacy skills. Node 2, with a predicted PV adoption rate of 11.4% consists of neighborhoods with a relatively high percentage of elderly, relatively low labour participation and residents living in older houses compared to other nodes. These characteristics could reflect neighborhoods with dominantly elderly and retired residents. A reason for the relatively low adoption rate could be that older people are less informed about climate change and do not decide to invest in a sustainable energy source such as solar energy. The characteristics of Node 3 are relatively similar to the characteristics of Node 2 and the same inference holds. The neighborhoods of Node 4 tend to have more expensive homes and use more electricity. However, these neighborhoods still have a relatively low PV adoption rate and purchasing solar panels could be a good option to save money. Informing these specific neighborhoods about cost savings could increase PV adoption and could be an important lead for policy makers. Node 5 to 9 have less distinctive characteristics which makes it more difficult to understand why these neighborhoods have an average PV adoption rate. This is a limitation of wanting to group all the neighborhoods within just 12 clusters and shows that reality is sometimes more complicated. A possible way to have better defined clusters is to increase the number of clusters but this is not beneficial for interpretability.

Therefore, only the nodes that have clear characteristics are interpreted. Node 10 contains neighborhoods with a relatively high education rate, a high income and high home values. These characteristics could reflect more wealthy neighborhoods in which residents have the financial capacity to invest in sustainable energy. Node 11 has less distinctive characteristics while the PV adoption rate is relatively high. Moreover, Node 12 consists of a few neighborhoods having the highest education rate on average. These neighborhoods have the highest average PV adoption rate and a possible reason could be that the residents are intelligent, well informed people that understand the need for PV adoption.

5.3 Regional disparity of PV adoption

Besides understanding who is buying solar panels and who is not, it is also interesting to understand the regional difference for PV adoption. Figure 11 shows a scatter plot including all neighborhoods of the Netherlands classified to their corresponding nodes according to the decision tree model. Although this scatterplot provides just a glance of regional differences, this could already give policymakers interesting insights. According to Figure 2 in the theoretical framework section, Zuid-Holland and Noord-Holland have the lowest PV adoption rate of 7.8% and 9.0% respectively. Figure 11 shows that both provinces are overrepresented by nodes with a low PV adoption rate. Also Gelderland, Limburg, Noord-Brabant, Overijssel and Utrecht are provinces with many neighborhoods of which a large part are neighborhoods with a low PV adoption rate. The scatterplot shows that there is not a large geographical difference between these provinces merely looking at the clusters of neighborhoods. Still, this information can be used by policy makers to understand which provinces might need more attention in accelerating PV adoption. Drenthe and Zeeland, the regions with the highest PV adoption rate, have fewer neighborhoods with low PV adoption. Moreover, Drenthe, Friesland and Groningen have a dominantly high presence of Node 5 neighborhoods. Node 5 neighborhoods are characterized as having relatively few immigrants, a low education rate and few houses are owned by the residents. This information can be utilized by policy makers to specifically target certain neighborhoods depending on their characteristics and PV adoption rate and is discussed in more detail in the discussion section. Moreover, the province Flevoland follows a different distribution than other provinces and especially neighborhoods classified as Node 3 are present. Node 3 is characterized as having relatively many elderly and old houses including a low PV adoption rate. Also this information as well as other province specific characteristics can be beneficial for policy makers to increase PV adoption.

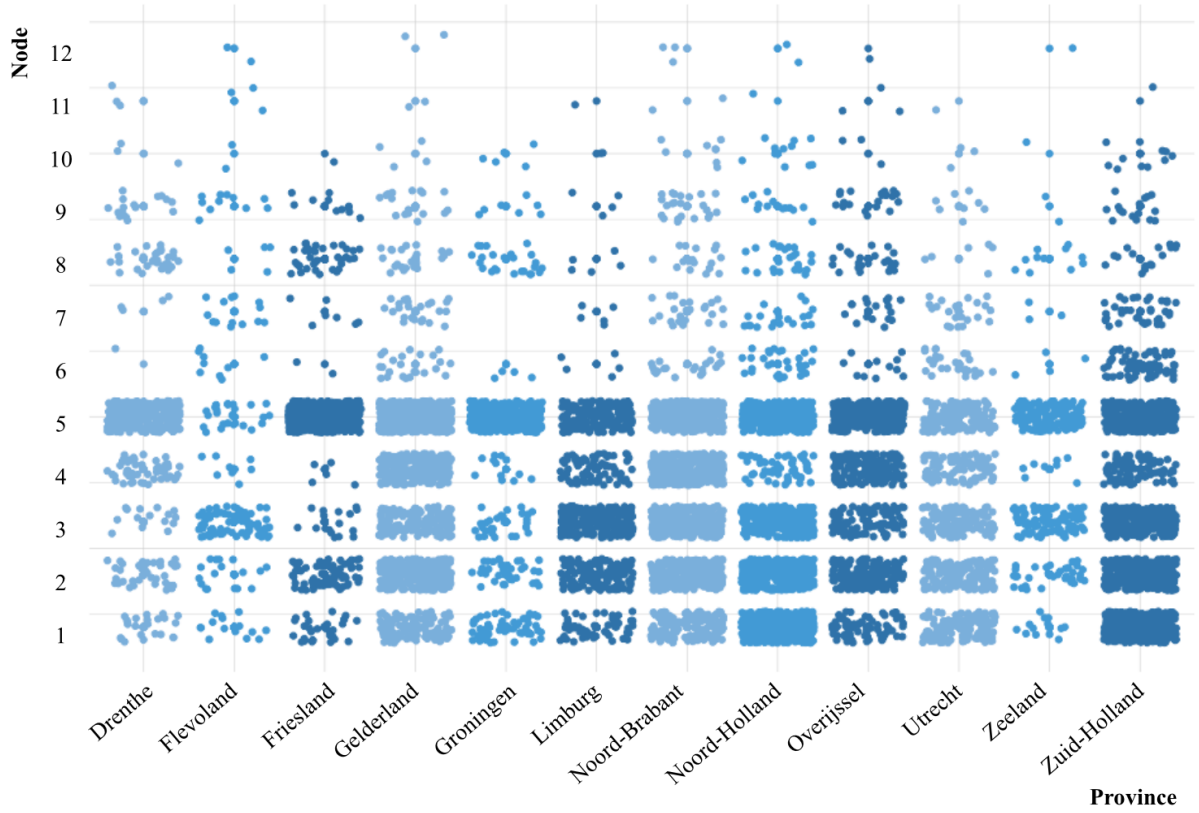

Figure 11. Scatterplot showing the distribution of nodes and provinces

5.4 Assumptions and robustness

To ensure the models are eligible for interpretation, the assumptions and robustness of the results are evaluated. Regarding the parametric generalized linear model, none of the assumptions seems to be grossly violated. The residual plot in Figure 16 in appendix A3 shows that the relationship between the target variable and attributes seems roughly linear and the error terms are uncorrelated. A few outliers can be detected in the data but compared to the large sample size these outliers have no effect on the performance of the model. The model shows a little heteroskedasticity but this is reduced by using the logit link function. Moreover, the model has few outliers and high leverage points which do not affect the model significantly. Figure 17 in appendix A3 shows the Cook's distance plot which is a commonly used estimate to detect outliers (Cook, 1977). An often used cut off value for detecting outliers is 1 on Cook's distance for datasets with a relatively large number of observations (Bollen & Jackman, 1985). Within the generalized linear model, the largest Cook's distance is 0.13 and therefore the model does not contain significant outliers. Figure 18 in appendix A3 shows the generalized leverage for the predicted values. The generalized leverage of an estimator is defined as a measure of importance of individual observations (Wei et al., 1998). A few observations have a relatively high generalized value which could potentially influence the performance of the model. A double check is done for these observations and confirms that this is actual data without any mistakes being made.

Moreover, the generalized linear model is fitted without these data points and performance scores are similar to the model with these data points included. Because these potential high leverage points are true observations of the model and their presence does not influence model performance, these observations are included in the final model. The last assumption is that the model does not suffer from multicollinearity. Multicollinearity occurs when correlations among independent variables exist and limits interpretation of the model. Figure 5 in the data section shows that there are many correlations between the independent variables and therefore multicollinearity is further examined to ensure this assumption holds.

The extent of multicollinearity is commonly measured using variance inflation factors (VIF). The VIF is a ratio of the variance of the estimated coefficient of a certain predictor of the full model and the variance of the estimated coefficient of this predictor when estimating the model with this predictor only (James et al., 2013). A value of 1 for the VIF indicates a complete absence for collinearity. However, as a rule of thumb, if the VIF exceeds 5 or 10 this indicates a problematic amount of collinearity (Craney & Surles, 2002). Figure 12 shows the variance inflation factors for all the attributes included and are calculated using the *car* package (Fox and Weisberg, 2019). None of the attributes exceeds the maximum VIF value and the variable Owned houses has the highest VIF being 4.02. Therefore, multicollinearity is limited and not an issue while interpreting the coefficients. All the assumptions of the generalized linear model hold and therefore interpretation is possible.

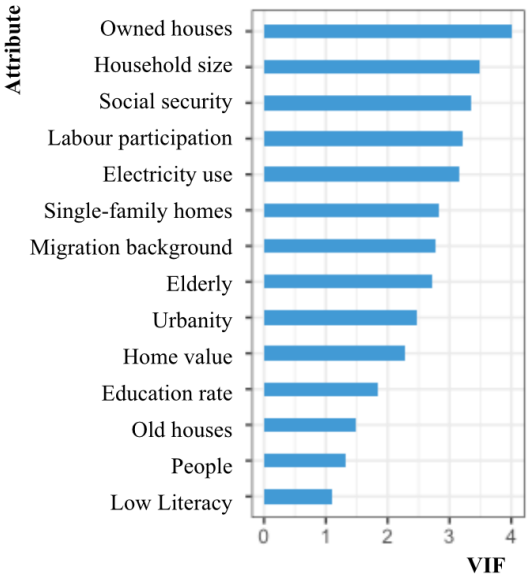

Figure 12. VIF values for all attributes

Regarding the non-parametric models there are no formal distribution assumptions and therefore these models can be interpreted. However, more sophisticated interpretation methods such as partial dependence plots should be evaluated with care. Partial dependence plots measure the marginal effect of a certain feature in the model while all the other features are held constant. However, if correlation between features exist, it is in reality not possible that a certain feature changes without everything else to change. Therefore, multicollinearity is also a problem for interpreting partial dependence plots (Friedman, 2001). Fortunately the VIF values do not exceed the maximum VIF value and therefore interpretation is possible, but should be done with care.

6. Discussion

The main implication of this research is to provide a constructive and systematic understanding of the interplay between a large number of socioeconomic attributes to inform stakeholders on who is buying solar panels and who is not. While the results section outlines the different machine learning models and their output, this section further examines how stakeholders can leverage this information to accelerate PV adoption. Although both, businesses active in the PV market and policy makers responsible for climate change prevention, are important stakeholders, the focus in this thesis will be on the latter group for ease of interpretation. However, the implications provided in the discussion are relevant for commercial businesses as well. Policy makers can leverage the results from this analysis by creating new efficient policies or targeting specific groups with existing policies.

The most important predictors for PV adoption are the percentage of old houses, the percentage of single-family homes, the education rate and the electricity usage. The presence of old houses and electricity use are negatively correlated with PV adoption whereas the presence of single-family homes and education rate are positively correlated with PV adoption. Especially the positive relationship between education rate and PV adoption rate can be of interest to policy makers. A possible cause could be that higher educated people better understand why climate change should be prevented and are better able to filter out misinformation on the topic (Treen et al., 2020). Informing neighborhoods with a relatively low education rate about the climate challenge that is ahead of us could stimulate PV adoption in these neighborhoods. Moreover, higher educated people tend to have a higher income and another possible reason for PV adoption among this group could be that these residents are able to afford the purchase of solar panels compared to residents having a lower income (Griliches & Mason, 1972). Although current policies aim to make investments in renewable energy affordable for everyone, a review of the current policies could clarify if this is also really the case in practice. New policies that lower the investment burden of having solar energy could therefore stimulate PV adoption, especially within neighborhoods that have less money to spend. Another actionable insight for policy makers is the negative relationship between electricity use and PV adoption. Informing neighborhoods with relatively high energy usage about cost reduction by purchasing solar panels could also stimulate PV adoption. The attributes percentage of old houses and single-family homes seem to provide less actionable insights without further research. As mentioned in the results section, a negative relationship between the presence of old houses and PV adoption could be due to the fact that residents have more efficient energy saving options before considering solar panels, like isolation. Moreover, neighborhoods with a relatively low percentage of single-family homes are simply not able to increase PV adoption because there is less rooftop space available.

Even more interesting for policy makers is the grouping of neighborhoods after fitting the decision tree model and classifying the neighborhoods to their corresponding nodes. Policy makers could utilize this information to make tailor made campaigns for different types of neighborhoods based on their characteristics. Especially Node 1, 2, 3 and 4 displayed in Figure 10 are of interest to policy makers because these nodes have distinctive characteristics and contain neighborhoods with a relatively low predicted PV adoption rate. As mentioned in the results section, Node 1 is characterized by a relatively large proportion of people having a migration background, low literacy skills and more social security is provided. Moreover, these neighborhoods consist of relatively few single-family homes and few owned houses. These characteristics could reflect poorness in these neighborhoods. Moreover, residents might not speak the Dutch language well or have below average literacy skills. From a policymaker's perspective, these neighborhoods could be a point of focus for both new policies and raising awareness of existing policies among the population. One of the reasons for having a low PV adoption rate could be that residents are less informed on buying solar panels because they either don't speak the Dutch language well or because their literacy skills are below average. Targeting these neighborhoods to provide better information about the importance of PV adoption could stimulate PV adoption. Moreover, information could be provided in multiple languages to also address residents that don't speak the Dutch language well. The relatively similar nodes 2 and 3 can also be of interest to accelerate the purchase of solar panels. Within these neighborhoods, the percentage of elderly is relatively high, labour participation is relatively low and residents live in older houses compared to other nodes. A possible reason behind a relatively low PV adoption rate in these neighborhoods could be that elderly are less informed about climate change and the importance of renewable energy sources. Policy makers could focus on tailor made informing of these neighborhoods with a focus on elderly that are less aware of today's climate change challenge. Also the characteristics of Node 4 could provide interesting actionable insights for accelerating PV adoption. Neighborhoods within this group tend to have more expensive homes that also use more electricity on average. Informing the residents in these neighborhoods about cost reduction that goes in hand with adopting solar panels could therefore be an effective way to increase PV adoption.

Finally the regional differences between PV adoption rates could also be a point of attention for policy makers. By understanding which provinces have a low PV adoption rate and which type of neighborhoods are dominantly represented in these provinces, policy makers can specifically target these provinces. Figure 11 shows that, for example, Zuid-Holland and Noord-Holland are overrepresented by nodes with a low PV adoption rate. However, these provinces don't have a specific node that stands out. On the other hand, the provinces Drenthe, Friesland and Groningen have a dominantly high presence of Node 5 neighborhoods. Within these provinces, a campaign could be launched that anticipates based on the characteristics of Node 5. Node 5 neighborhoods are characterized as having relatively few immigrants, a low education rate and few houses are owned by

the residents. As mentioned in the results section, Flevoland follows a different distribution than other provinces and especially neighborhoods classified as Node 3 are present. Node 3 is characterized as having relatively many elderly and old houses including a low PV adoption rate. With this information, policy makers could start a campaign focusing on informing these neighborhoods about climate change while keeping in mind that their target group consists of relatively many elderly.

Besides the examples provided above, many more interesting insights can be derived from the data. Policymakers could use both socioeconomic and regional differences to specifically adjust policies or target groups with a low PV adoption rate depending on their characteristics. Also businesses can leverage this information to adjust marketing strategies and target certain groups of people or neighborhoods. Businesses could, for example, make marketing content related to what type of people live in a certain neighborhood. Content within neighborhoods that are relatively poor could focus on the green investment loan made available by the government and the cost benefit of having solar panels. Content within neighborhoods with relatively many people having a migration background could be in multiple languages to make sure everyone understands it. As a last example, to target neighborhoods with relatively many elderly, businesses active in the PV market should make sure the marketing content is easy to understand for older people.

7. Conclusion

Within the conclusion section, the main findings of this paper are summarized in order to answer the main question in this analysis. Moreover, limitations and options for future research are provided.

7.1 Main findings

The main goal of this analysis is to better understand who is buying solar panels and who is not. By looking at socioeconomic factors the residential PV adoption in the Netherlands is closely examined and regional differences are explained. The data used for research contains solar energy generation and socioeconomic factors for every neighborhood in the Netherlands of 2019. The research framework contains the machine learning methods generalized linear model (GLM), decision tree and random forest and test RMSE-scores are compared to measure performance. Moreover, sophisticated interpretation methods such as partial dependence plots and decision tree-based clustering are used to interpret the models and uncover regional differences for PV adoption. Concerning the methods, the best performing model is a random forest. However, the generalized linear model and decision tree also predict well on the test data and therefore all three models are used for interpretation due to ease of interpretability and simplicity. Interpreting the models all together strengthen the findings of this analysis and provide robust results that help to answer the research question at hand.

The key socioeconomic factors that drive PV adoption in the Netherlands are the abscence of old houses, the presence of single-family homes, the education rate and electricity use. The presence of old houses and electricity use have a negative effect on PV adoption whereas the presence of single-family homes and education rate have a positive effect on PV adoption. These findings are in line with previous research on PV adoption in other countries and confirm the same relationships exist in the Netherlands (Ameli & Brandt, 2014; Balta-Ozkan et al., 2015; Davidson et al., 2014; Sommerfeld et al., 2017). Whilst these socioeconomic factors have the strongest relationship with PV adoption, 12 out of 16 features have a significant estimate within the generalized linear model. The effect of the other features are discussed in more detail in the data section. Moreover, the relationships of the most important predictors and PV adoption are roughly linear. The models show that the relationships within the data are not very complex and nonlinear relationships and interaction effects among features and PV adoption have a marginal presence. Another important finding is that this paper proposes a decision tree-based clustering of neighborhoods based on similar socioeconomic characteristics and PV adoption rates. These groups of similar neighborhoods provide a good overview of the different PV adoption rates by looking at the socioeconomic characteristics and can be used for policy making and marketing purposes. Of the 12 neighborhood clusters, especially the 4 clusters with the lowest average PV adoption rate are of interest to stakeholders. Policy makers could target the residents within these clusters with different strategies that respond to the characteristics in

these neighborhoods. Policy makers could, for example, highlight the opportunity to loan money for green investments and the cost benefit of having solar panels to neighborhoods that are relatively poor. For another cluster of neighborhoods, policy makers could focus more on providing information in multiple languages because there is a relatively high percentage of residents having a migration background in this cluster. As a final example, policy makers could focus more on raising awareness about climate change within the cluster of neighborhoods with relatively many elderly that might be unaware of the challenge we are facing. Finally, one of the most important results is a better understanding of the regional disparity for PV adoption in the Netherlands by combining machine learning model output and geographical data. This information can be used to understand which clusters of neighborhoods are present in which provinces to even better target neighborhoods with a low PV adoption rate.

The findings in this paper have important academic and practical contributions. First, to the best of my knowledge, state of the art machine learning methods have barely been used before within this research topic. Second, this type of research has not been conducted on the Dutch residential PV market yet and confirms similar relationships found in other countries also exist in the Netherlands. Third, the non-examined factor low literacy is researched in the context of PV adoption. Related to the practical contribution of this paper, policy makers and businesses active in the PV market can utilize these results to better understand who is buying solar panels and who is not. This information can be used to specifically adjust policies or target groups with a low PV adoption rate depending on their characteristics. Concrete examples for this are included in the main findings above and the results section provides a more detailed overview. Last but not least, society at large can benefit from an accelerating rooftop revolution preventing climate change and preserving our planet in the long run.

7.2 Limitations

As with all research, this paper has its limitations. A few limitations are related to data availability within this research topic. The main goal of this thesis is to understand which socioeconomic factors drive PV adoption in the Netherlands. Socioeconomic factors consist of a wide range of variables about wealth, demographics and way of living. Within this paper, only the socioeconomic factors that are publicly available are analyzed. Although the socioeconomic dataset made publicly available by Statistics Netherlands (in Dutch: CBS) contains a wide variety of variables, not everything that could be of interest for this research is included. Factors such as direct income, political preference and health are not included but could be of interest to provide even better results in understanding the interplay between socioeconomic factors and PV adoption. Moreover, some of the variables included in this research contain missing values due to the privacy of residents living in these neighborhoods. Especially the variables labour participation, high income, home value and low literacy skills contain missing values which have been provided the mean value in the data section. A complete dataset

could contribute to even better results. Another data limitation is the availability of data for examining the non-examined feature low literacy skills. The data for this attribute is less granular and only available on township level rather than neighborhood level. Because all the neighborhoods in a certain township have the same low literacy value, it is more complex to really estimate the effect of low literacy on PV adoption. Although such an effect could exist in reality, this paper is not able to discover the potential relationship due to data limitations.

Also the use of modern machine learning and sophisticated interpretation techniques come with some flaws. The interpretation techniques that are used, such as the partial dependence plots, decision tree clustering and the heatmap with neighborhood characteristics, make it possible to better interpret the data and generalize different types of neighborhoods. However, decision making based on these outputs should be done with care because the results provide a generic overview rather than a detailed analysis. Neighborhoods are grouped based on similar characteristics and PV adoption rates but the characteristics of two neighborhoods within the same grouping could still be very diverse. The machine learning models try to divide over 10.000 neighborhoods within 12 groups. Policy makers and marketing departments should treat this analysis as a generic overview which could be used to have a rough understanding of residential PV adoption in the Netherlands. Regarding the parametric generalized linear model, a variable selection procedure is used to remove redundant variables from the model. Although this technique is quite effective to prevent multicollinearity and noise in the output, the approach is greedy and naive. The stepwise selection procedure only considers a set of variables at a certain time and permanently deletes or adds variables based on the dataset that is provided. Removed variables at the start of the procedure might actually add value in the final model but are not considered anymore. This flaw often leads to standard errors and p-values that are biased towards zero, overcomplexity of a model and inflation of coefficients (Harrel, 2001).

Finally, it is important to mention that this paper examines correlations between variables rather than causal relationships. Causality implies that the change of a certain feature causes a change of the target variable. Assuming causal relationships can only be done after running robust experiments that determine causation and is out of scope for this paper. Therefore, the inferences about why certain relationships exist should be interpreted with care and are mostly suggestions based on existing literature and logical reasoning.

7.3 Future research

The significant effect of many socioeconomic features is the main starting point for further research. Although the three models within this analysis provide consistent results that substantiate the existence of certain relationships between socioeconomic attributes and PV adoption, further analysis could be done to make the results more robust. Policy making decisions can have big consequences

for the public treasury and are therefore only done after a thorough analysis of the policy. Therefore, a redo of the study with more data or updated data could strengthen the results in this analysis. The same study could be done for previous years than 2019 or more recent years when data is available. Another advantage of a multi-year analysis is a better understanding of PV adoption development over time. This analysis could have valuable information on which direction PV adoption is moving and how fast. Policy makers and businesses active in the PV market would better understand which types of neighborhoods are already adopting over time and which are not and this information can be used to accelerate the adoption of residential PV. Also the extension of the current research with more socioeconomic variables could be a valuable addition. Including factors such as direct income, political preference and health could provide even better results in understanding who is buying solar panels and who is not. Especially the relationship between a socioeconomic feature like political preference and PV adoption could provide interesting and actionable insights. If such a relationship exists, policy makers could target neighborhoods with tailor made information about solar panels that takes into account the dominant political preference in a certain neighborhood. To give an example based on US politics, democrats might be more encouraged to purchase solar panels when they are educated about the prevention of climate change whereas republicans might be more encouraged to purchase solar panels when they hear about the return on investment. Moreover, the new feature of low literacy could be further researched using more granular data. Understanding the relationship between literacy skills and PV adoption could provide actionable insights to policy makers as they are more informed about what effect literacy skills have on PV adoption.

Further research on causality between variables could also be an interesting contribution to existing literature. This paper focussed on the interplay between socioeconomic factors and PV adoption and proves certain relationships exist. However, policy makers and businesses active in the PV market would benefit from knowing whether these relationships are also causal to have an even better understanding of the residential PV market. Further research could consist of experiments that determine causation. Also the inferences about why certain relationships exist could be further researched to really understand why residents buy solar panels or not. Conducting a large survey, for example, could provide insights about behavior and decision making of residents.

On the methodology side there is still room to use other machine learning models within this research topic. Although preliminary analysis has been done one a variety of machine learning models and the GLM, decision tree and random forest have been extensively applied, other machine learning models might be able to even better predict or interpret PV adoption. It would be interesting to see whether different extensions could lead to an increase in prediction accuracy or a better understanding of the interplay between socioeconomic factors and PV adoption.

8. Appendix

Appendix A1: Data description

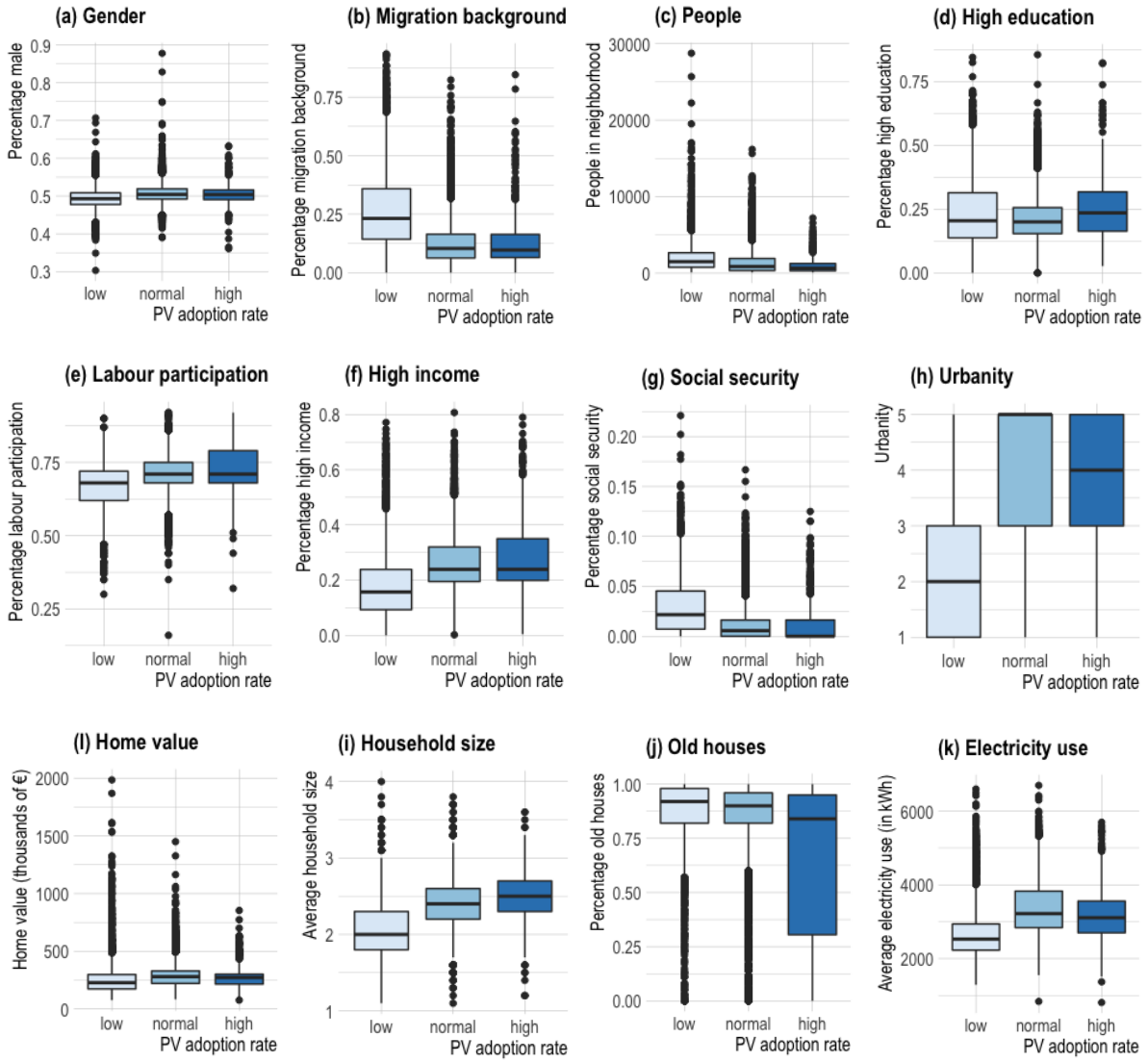

Figure 13. Comparison of socioeconomic variables among different adoption rates (low <= 10%, 10% < normal < 30%, high >= 30%).

Appendix A2: Parameter testing

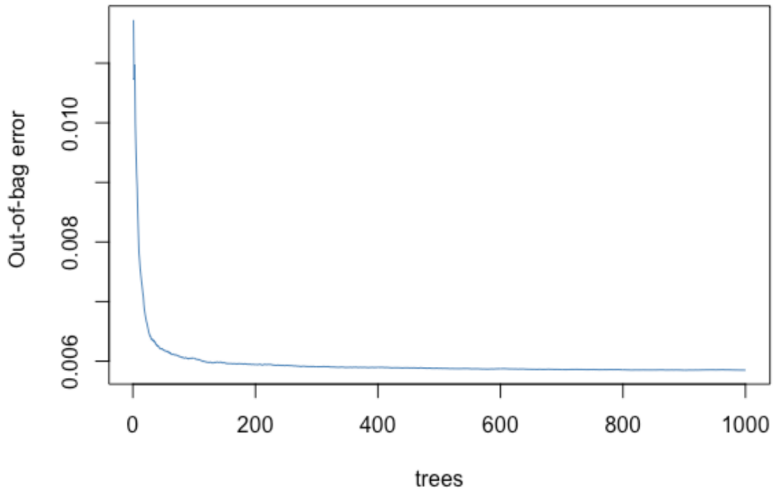

Figure 14. The out-of-bag train error for estimating the number of trees B in random forest

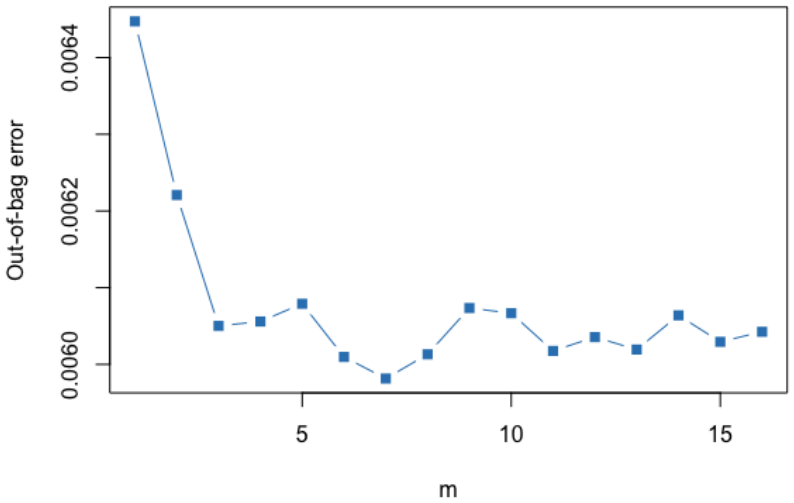

Figure 15. The out-of-bag train error for estimating subset m in random forest

Appendix A3: Robustness

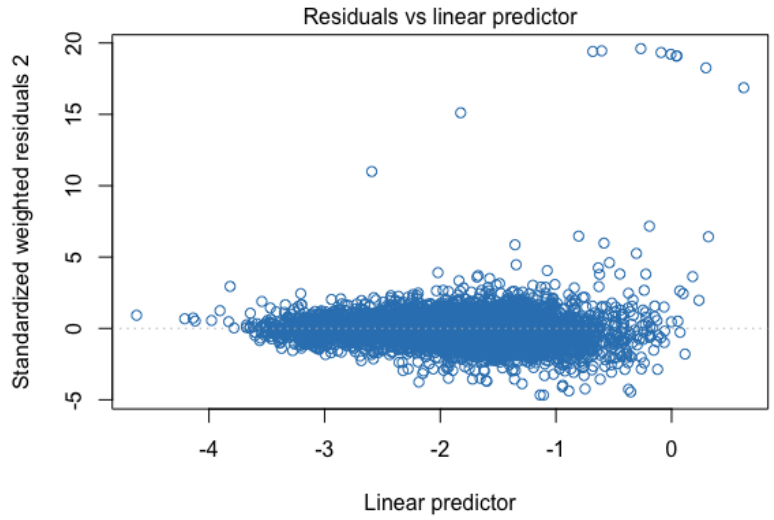

Figure 16. Residual vs linear predictor plot to detect linearity of the data

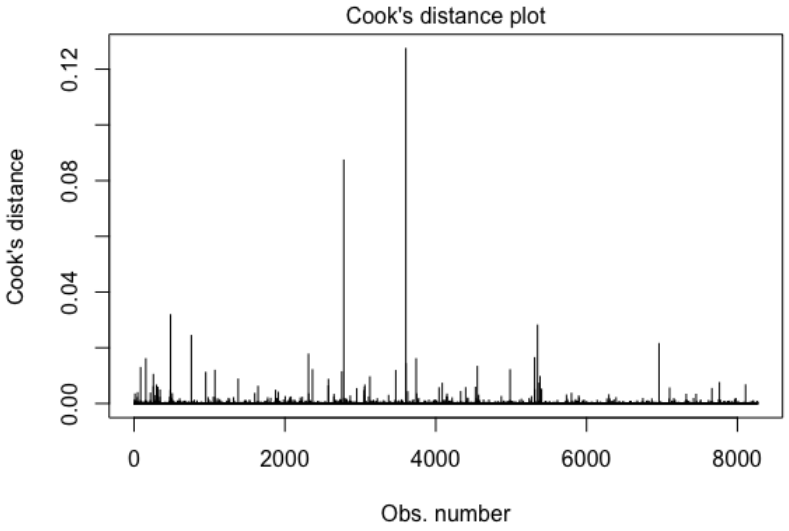

Figure 17. Cook's distance plot for detecting outliers

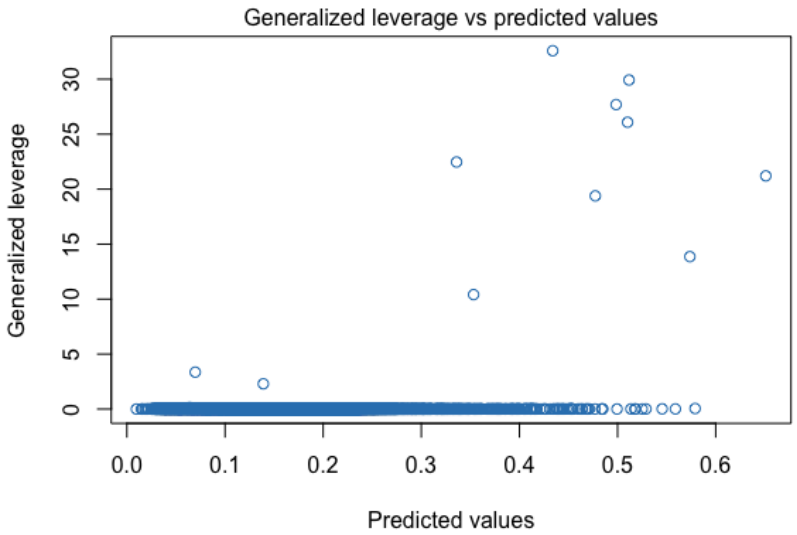

Figure 18. Plot for detecting high leverage points

9. References

Ameli, N., Brandt, N., 2014. Determinants of Households' Investment in Energy Efficiency and Renewables: Evidence from the OECD Survey on Household Environmental Behaviour and Attitudes. *Organization for Economic Co-operation and Development (OECD) Economics Department Working Papers, No. 1165 OECD Publishing*, Paris.

Artin, E. (2015). The gamma function. *Courier Dover Publications*.

Balta-Ozkan, N., Yildirim, J., & Connor, P. M. (2015). Regional distribution of photovoltaic deployment in the UK and its determinants: A spatial econometric approach. *Energy Economics*, 51, 417-429.

Bollen, K. A., & Jackman, R. W. (1985). Regression diagnostics: An expository treatment of outliers and influential cases. *Sociological Methods & Research*, 13(4), 510-542.

Bollinger, B., & Gillingham, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science,* 31(6), 900-912.

Bower, J. L., & Christensen, C. M. (1995). Disruptive technologies: catching the wave.

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

CBS. (2019a). Zonnestroom; vermogen zonnepanelen woningen, wijken en buurten, 2019. Retrieved from https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=85010NED&_theme=278

CBS. (2019b, July 30). Kerncijfers wijken en buurten 2019. Retrieved from https://www.cbs.nl/nl-nl/maatwerk/2019/31/kerncijfers-wijken-en-buurten-2019

CBS. (2019c). Hernieuwbare energie in Nederland 2019. Retrieved from https://longreads.cbs.nl/hernieuwbare-energie-in-nederland-2019/zonne-energie/#:~:text=Het%20opgesteld%20vermogen%20voor%20en,Nederland%20is%20ruim%2010%20procent.

CBS. (2021a, March 12). Uitstoot broeikasgassen 8 procent lager in 2020. Retrieved from https://www.cbs.nl/nl-nl/nieuws/2021/10/uitstoot-broeikasgassen-8-procent-lager-in-2020

CBS. (2021b, September 30). Hernieuwbare Energie in Nederland 2020. Retrieved from https://www.cbs.nl/nl-nl/longread/aanvullende-statistische-diensten/2021/hernieuwbare-energie-in-nederland-2020

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics,* 19(1), 15-18.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering,* 14(3), 391-403.

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, 34, 1-24.

Dharshing, S. (2017). Household dynamics of technology adoption: A spatial econometric analysis of residential solar photovoltaic (PV) systems in Germany. *Energy Research & Social Science*, 23, 113-124.

Davidson, C., Drury, E., Lopez, A., Elmore, R., & Margolis, R. (2014). Modeling photovoltaic diffusion: an analysis of geospatial datasets. *Environmental Research Letters*, 9(7), 074009.

Deloitte. (2018). State of the State-onderzoek: Zonnepanelen kunnen de helft van de Nederlandse elektriciteitsbehoefte opwekken. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/deloitte-analytics/deloitte-nl-data-analytics-state-of-the-state-zonnepanelen.pdf

DNE Research. (2022). Het Nationaal Solar Trendrapport 2022. Retrieved from https://www.solarsolutions.nl/trendrapport/

European Commission. (2020). EDGAR - Emissions Database for Global Atmospheric Research. Retrieved from https://edgar.jrc.ec.europa.eu/country_profile/NLD

Faraway, J. J. (2016). Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. *Chapman and Hall/CRC*.

Fox, J. and Weisberg, S. (2019). An R Companion to Applied Regression. Sage, *Thousand Oaks CA,* third edition.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J*ournal of Machine Learning Research,* 20(177), 1-81.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics,* 1189–1232.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics,* 2(3), 916-954.

Geletterdheid InZicht. (2020). Aandeel laaggeletterden naar regio. Retrieved from https://geletterdheidinzicht.nl/

Green, M. A. (2000). Photovoltaics: technology overview. *Energy Policy,* 28(14), 989-998.

Griliches, Z., & Mason, W. M. (1972). Education, income, and ability. *Journal of Political Economy, 80(3, Part 2)*, S74-S103.

Gupta, A. K., & Nadarajah, S. (2004). Handbook of beta distribution and its applications. *CRC press.*

Harrell, F. E. (2001). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. *Springer Science & Business Media.*

Hastie, T. J., & Tibshirani, R. J. (2017). Generalized additive models. *Routledge*.

IPCC. (2021). Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Retrieved from https://www.ipcc.ch/report/ar6/wg1/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. *New York: Springer,* 112, 18

Kausika, B. B., Dolla, O., & Van Sark, W. G. J. H. M. (2017). Assessment of policy based residential solar PV potential using GIS-based multicriteria decision analysis: A case study of Apeldoorn, The Netherlands. *Energy Procedia*, 134, 110-120.

Lan, H., Gou, Z., & Lu, Y. (2021). Machine learning approach to understand regional disparity of residential solar adoption in Australia. *Renewable and Sustainable Energy Reviews*, 136, 110458.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News,* 2(3):18–22.

Londo, H. M., Matton, R., Usmani, O., van Klaveren, M., & Tigchelaar, C. (2017). De salderingsregeling: Effecten van een aantal hervormingsopties. *Petten: ECN*.

Lu, Y., Khan, Z. A., Alvarez-Alvarado, M. S., Zhang, Y., Huang, Z., & Imran, M. (2020). A critical review of sustainable energy policies for the promotion of renewable energy sources. *Sustainability*, 12(12), 5078.

Molnar, C. (2020). Interpretable machine learning. *Lulu.com*

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 18(6), 275-285.

Palmer, J., Sorda, G., & Madlener, R. (2015). Modeling the diffusion of residential photovoltaic systems in Italy: An agent-based simulation. *Technological Forecasting and Social Change,* 99, 106-131.

Panwar, N. L., Kaushik, S. C., & Kothari, S. (2011). Role of renewable energy sources in environmental protection: A review. *Renewable and Sustainable Energy Reviews*, 15(3), 1513-1524.

Paris Agreement. (2015, December). Paris agreement. *In Report of the Conference of the Parties to the United Nations Framework Convention on Climate Change*, 21st Session, 2015. Paris.

Poole, M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, 145-158.

Rijksoverheid. (2022). Overheid bevordert groei zonne-energie. Retrieved from https://www.rijksoverheid.nl/onderwerpen/duurzame-energie/zonne-energie

Rogers, E. M., Singhal, A., & Quinlan, M. M. (2014). Diffusion of innovations. In An integrated approach to communication theory and research (pp. 432-448). *Routledge.*

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel,* 81(10.5555), 26853.

Sardianou, E., & Genoudi, P. (2013). Which factors affect the willingness of consumers to adopt renewable energies?. *Renewable Energy*, 57, 1-4.

Sommerfeld, J., Buys, L., Mengersen, K., & Vine, D. (2017). Influence of demographic variables on uptake of domestic solar photovoltaic technology. *Renewable and Sustainable Energy Reviews*, 67, 315-323.

Sood, A., & Tellis, G. J. (2011). Demystifying disruption: A new model for understanding and predicting disruptive technologies. *Marketing Science,* 30(2), 339-354.

Sunter, D. A., Castellanos, S., & Kammen, D. M. (2019). Disparities in rooftop photovoltaics deployment in the United States by race and ethnicity. *Nature Sustainability*, 2(1), 71-76.

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2015). Package 'rpart'. Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf.

Treen, K. M. D. I., Williams, H. T., & O'Neill, S. J. (2020). Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change, 11(5), e665*.

Vasseur, V., & Kemp, R. (2015). The adoption of PV in the Netherlands: A statistical analysis of adoption factors. *Renewable and Sustainable Energy Reviews*, Vol. 41, 483-494.

Vattenfall. (2022). Duurzaamheidsindex van Nederland (DiNG). Retrieved from https://www.vattenfall.nl/producten/energie/duurzaamheidsindex/

VROM. (2007). Nieuwe energie voor het klimaat - Werkprogramma schoon en zuinig. Retrieved from https://europadecentraal.nl/wp-content/uploads/2013/01/Werkprogramma-Schoon-en-Zuinig.pdf

Wei, B. C., Hu, Y. Q., & Fung, W. K. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistics*, 25(1), 25-37.

Zhang, Y., Song, J., & Hamori, S. (2011). Impact of subsidy policies on diffusion of photovoltaic power generation. *Energy Policy*, 39(4), 1958-1964.