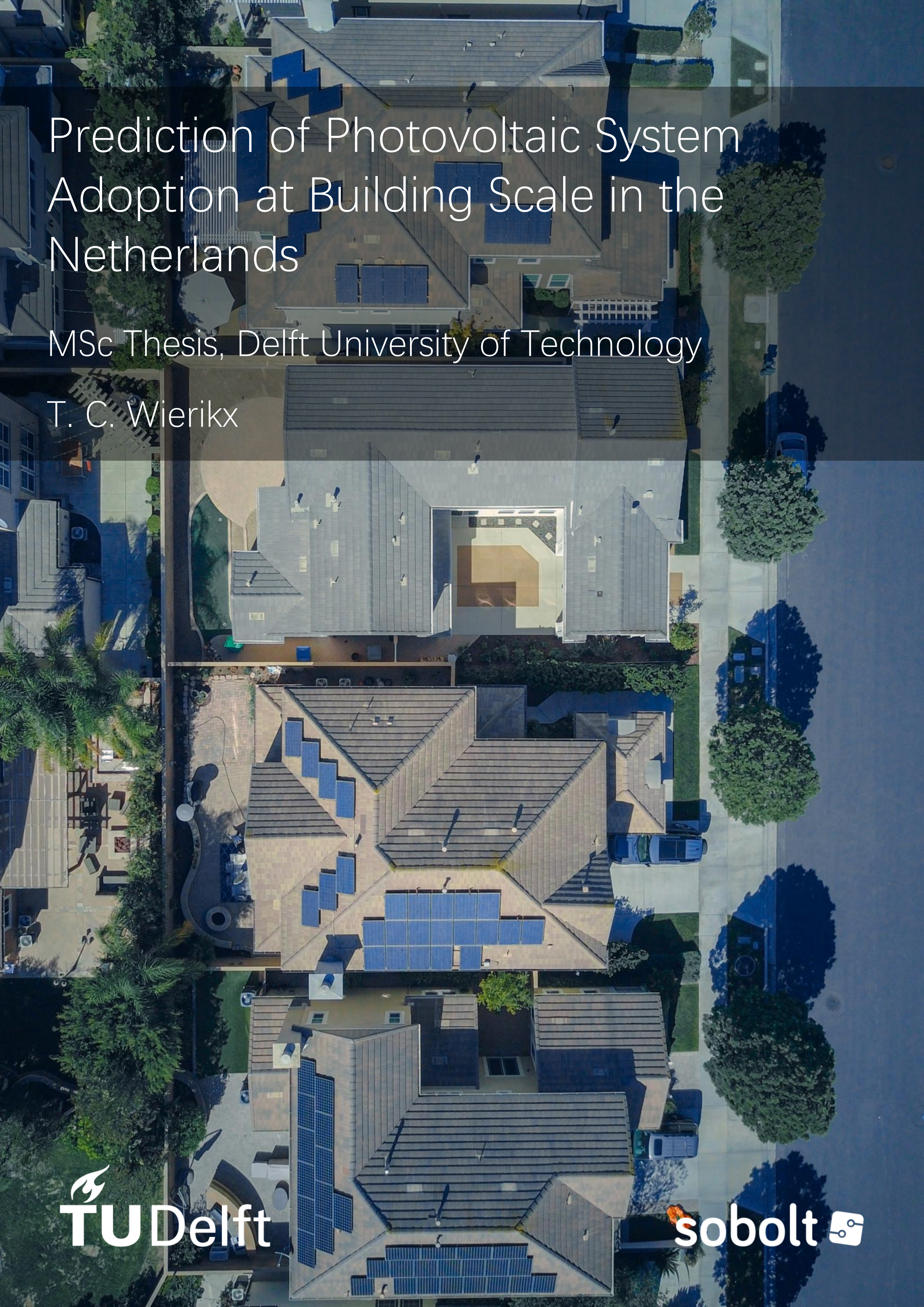


Prediction of Photovoltaic System Adoption at Building Scale in the Netherlands

MSc Thesis, Delft University of Technology

T. C. Wierikx

Prediction of Photovoltaic System Adoption at Building Scale in the Netherlands

Master Thesis

by

T.C. Wierikx

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday Nov 16, 2022 at 2:15 PM.

Student number:	4460510	
Project duration:	February 1, 2022 – Nov 16, 2022	
Thesis committee:	Dr. R.C. Lindenbergh,	TU Delft, supervisor
	Dr. M.A. Schleiss,	TU Delft, supervisor
	Ir. E. Verbree,	TU Delft, supervisor
	MSc. O. Fabius,	Sobolt, supervisor

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Cover image: Kindel Media (2021)

Abstract

Understanding the characteristics of photovoltaic system (PV) adopters can help policymakers realise energy transition more effectively. In this study, we developed a model that predicts PV adoption per building using geometric and socioeconomic variables. Seven geometric variables were created by processing building registration data, airborne laser scanning data and 3D building models based on airborne laser scanning data. Additionally, eight socioeconomic variables were created from building registration data and socioeconomic postal code statistics. The random forest machine learning model, which was trained and evaluated on 646 000 buildings in the province of Overijssel, The Netherlands, displays good overall performance with an AUC of 0.77. Moreover, the model demonstrates that buildings have an increased probability of PV adoption if they (i) have a suitable area above 30 m^2, (ii) have a rooftop higher than 6 m, (iii) have a non-flat roof, (iv) were built after 1970, (v) only have one address registered and (vi) are used for residence. Similar experiments involving a different type of machine learning model (i.e., a neural network) and province (i.e. North Holland) yield similar results. Future improvements could focus on increasing model performance in residential areas and studying the effect of PV stimulation by including a temporal component.

Glossary

α	Aspect
A	Area
fpr	False Positive Rate
k	Number of folds
k_0	Number of folds in outer loop
k_i	Number of folds in inner loop
p	Two-sided p-value for a hypothesis test whose null hypothesis is that the slope is zero
S	Slope
tnr	True Negative Rate
tpr	True Positive Rate
AHN	Actueel Hoogtemodel Nederland
AUC	Area Under the ROC Curve
BAG	Basisregistratie Adressen en Gebouwen
CBS	Centraal Bureau voor de Statistiek
CV	K-fold Cross-Validation
EU	European Union
i.d.d.	Independent and Identically Distributed
LoD	Level of Detail
LR	Logistic Regression
MLM	Machine Learning Model
MLP	Multi-Layer Perceptron
NeCV	Nested K-fold Cross-Validation
NH	North Holland
NN	Neural Network
PC4	Partial Postal Code, 4 digits
PC6	Full Postal Code, 4 digits and two letters
PCA	Principal Component Analysis
RF	Random Forest
ROC	Receiver-Operating Characteristic
VIF	Variance Inflation Factor

Preface

The report you are about to read is the result of my Master's in Geoscience and Remote Sensing at the Delft University of Technology. In the course of nine months, I have carried out research on predicting solar panel prediction for buildings in Overijssel. I was happy to collaborate with Sobolt, a sustainability-driven AI company based in Rotterdam, which provided me with valuable data on solar panel adoption in the Netherlands. Using their data, various other data sources and several processing steps, I developed a model that achieves decent performance. It has contributed to research by proposing a method for creating a PV adoption prediction model on a building scale, as well as an assessment of its advantages and disadvantages.

My main motivator for doing research has always been to contribute to society, and in recent years, my focus has turned towards sustainability and energy transition. During my Master's in Delft and my time at Sobolt, I have learned a ton about how to (and how not to) create more impact, and I am committed to keep doing this. I sincerely hope - and believe - that my study on solar panel adoption has solved a tiny extra part of the enormous climate change puzzle we inevitably have to solve.

Carrying out a full study is rarely a one-person job, and neither was the case this time. I want to express a special thanks to my supervisors. First of all, Roderik, despite being one of the department's most chosen (and rightfully so) supervisors, always finds time to answer emails within an hour and provide feedback with never-ending enthusiasm. Marc has also been essential in my research; several times, I have bragged to others about the level of detail in his feedback and preciseness in reading, which mostly resulted in other students saying they'd wish to have a supervisor like Marc. With Edward coming from a different department, he was able to provide suggestions and comments that were a great addition to Roderik's and Marc's views. Lastly, I thank Otto for raising my thesis to a higher level with sharp analyses and making my graduation internship at Sobolt a more educational and above all enjoyable experience. Moreover, some friends and family also deserve credit for helping me with lots of proofreading and great suggestions, with special thanks to my father, Margot and Willem.

T.C. Wierikx
Delft, November 2022

Contents

List of Figures

List of Tables

1 Introduction

This chapter provides an introduction to this thesis. The first section, Section 1.1, gives some context on the topic of photovoltaic system (PV) adoption and Subsection 1.2.1 provides an overview of other research in this field. Then, Section 1.3 defines the scope of this thesis by establishing the research questions. Last, Section 1.4 presents a brief overview of the structure of this report.

1.1. Context

With an increased focus on energy transition, deploying photovoltaic systems on existing building rooftops is becoming increasingly popular. Furthermore, the recent Russian invasion of Ukraine has given our sustainable transition an added sense of urgency (Aïda Brands, 2022). For example, the European Union (EU) has increased its targets and aims to generate at least 45% of the energy by 2030 sustainably. An integral part of this ambition is increasing the share of energy produced by PV installation. These high ambitions are quantified in a comprehensive plan presented in May 2022. For example, the EU is committed to doubling the number of solar panels by 2025 to almost double that again by 2030. (Thom Opheikens, 2022).

These high ambitions are also present in Dutch policy plans. In the *Klimaatakkord*, the goals of the Dutch government are set out. It explains that, in addition to more large-scale electricity production on land, small-scale production of solar energy is also important for realising the climate challenge (Directoraat-generaal Klimaat en Energie, 2019). Solar energy generation by households also helps improve citizens' involvement and acceptance of the energy transition. Policymakers can help accelerate the growth in PV adoption by promoting PV among potential adopters. Knowing the characteristics of PV adopters helps policymakers target areas and buildings that are either most promising or need extra attention. Therefore, understanding spatial patterns in PV adoption is important not only from a scientific perspective but also from a policy perspective (Graziano & Gillingham, 2015).

1.2. Related Work on PV Adoption

This section describes in more detail what is already known about the characteristics of PV adopters and which factors correlate to the decision to adopt PV. Furthermore, PV data availability and related work on geometric building analysis are discussed.

1.2.1. Influential Factors

Different factors can influence the decision to adopt PV. The factors in this research are grouped into two main categories; socioeconomic and geometric. In this study, the socioeconomic factors refer to features related to the residents of the building and neighbourhood. Examples of socioeconomic factors are the inhabitant's age and income, the number of adopters within the neighbourhood and the local population density. The geometric factors comprise the physical properties of the building and surroundings. Examples are rooftop type and area suitable for PV. Please note that these spatial factors can be generated from both 2D and 3D datasets.

1.2.2. Questionnaire-Based Research on PV Adoption

Most research on factors possibly influencing PV adoption is done using questionnaires. Studies that use this method gather information by directly asking participants questions. The number of participants is usually a few hundred, up to a few thousand for larger studies. Questionnaires provide reliable data on whether the respondents have adopted PV because no remote sensing is needed. Furthermore, detailed questions can be asked, and much information can be gathered. The questionnaire-based studies on PV adoption mostly focus on finding the barriers to adoption. A possible downside of these studies is the limited size compared to readily available datasets.

Questionnaire-based research in the Netherlands concludes that the cost of adoption is an important element; people who consider solar panels affordable are more likely to adopt (Vasseur & Kemp, 2014). Financial factors, such as investment costs or potential energy savings, also have a strong influence. The reduced environmental impact is another motivator but is less significant than the financial factors. A similar study in Pakistan shows a somewhat different conclusion; the environmental impact is the most important driver (Qureshi, Ullah, & Arentsen, 2017). It does agree that financial factors are the biggest barriers. Bashiri and Alizadeh (2017) even find a negative effect of income on adoption probability in the region of Tehran. Their study shows a positive correlation between PV adoption with environmental concerns, knowledge of renewable energies, innovativeness and household size.

The disagreement on the dominant factors influencing PV adoption is often seen within literature (Schulte, Scheller, Sloot, & Bruckner, 2022). There is a large variety of the estimated effect of especially income on the (intention of) adoption. There are several possible explanations for these differences. Firstly, some studies look at the intention of adopting PV, while others look at the actual action. It is hypothesised that the intention is less dependent on income than the installation, which could cause this difference. Another explanation for the difference between research outcomes is the large difference between the adoption rates in research areas. Countries with a lower adoption rate are still in a different phase of the diffusion of adoption, which can attract a different audience for adoption compared to countries where PV adoption is already widely spread (Schulte et al., 2022). Lastly, research also suggests a complex interplay between independent variables, making it harder to determine the effect of every individual variable (Lan, Gou, & Lu, 2020).

1.2.3. Dataset-based Research on PV Adoption

A different approach to researching PV adoption is using readily available datasets, which will be referred to as dataset-based. Studies using a dataset-based approach for research on PV adoption generally look for correlations to explain, e.g. geographic patterns. A large advantage of this method is the scale of available data, and it is often easier to acquire those datasets than to gather a significant number of respondents. Another advantage is the possibility to extrapolate the results if desired. If the model uses widely available data, it could be applied to new areas without the need to acquire new data. One should note, however, that the reliability of models in a new area cannot be guaranteed.

Dataset-based research also has some downsides compared to questionnaire-based. Data reliability can be lower when techniques such as remote sensing are necessary to determine I nputs. Furthermore, the researchers are limited to data that is already available or can easily be determined on a large scale, which means that the data cannot be as specific as could be reached with questions in questionnaires. Lastly, datasets are often incomplete or on a coarser spatial resolution, possibly decreasing accuracy because of the need for interpolation or leaving out data.

In the Netherlands, the government does not have access to PV data at a building level. However, there is data available on a coarse level. Van der Kam, Meelen, van Sark, and Alkemade (2018) use this data to point out correlations between PV adoption and several factors in the Netherlands. The resolution of the data is all on a level of Dutch postal zip code PC4, i.e. the first four digits of the postal code. The PV adoption data is obtained from the Production Installation Register. Among other factors, they show a positive correlation between PV adoption and household size, percentage of people aged 45-65, building footprint area and percentage GroenLinks voters (a left-wing party focused on energy transition). A negative correlation is found for lower education levels and high population density.

A different study, carried out in the city of Apeldoorn, The Netherlands, concludes income to be the dominant driver when compared to house value, electricity consumption and PV adoption rates of the neighbours (Kausika, Dolla, & Van Sark, 2017). Their data is also on a PC4 resolution. The difference in results between Kausika et al. (2017) and Van der Kam et al. (2018) confirms the conclusion of Kausika et al. (2017) that dominant factors of PV adoption are not easily determined and show a large variety.

Several comparable papers also find important motivators for PV adoption based on socioeconomic datasets.

A study that uses logistic regression in a neighbourhood in the Gangnam district, Seoul, South Korea, finds a significant influence of the building type, the number of neighbouring adopters, household density and land price, where the latter two have a negative effect on PV adoption (Lee & Hong, 2019). A negative land price does not agree with the majority of other studies. The authors hypothesise that once a region reaches a certain level of wealth at which PV can be considered affordable, individuals with lower incomes are more motivated to adopt PV to save on their energy bills. Effects of rooftop area, year of construction and age are deemed insignificant in this study area. Their research is one of the few studies that combines some socioeconomic and geometric factors using a dataset-based approach. However, with only 99 buildings that adopted PV, the study size is relatively small when considering the possibilities of dataset-based research. Furthermore, the study took place in the Gangnam district, which has the highest energy consumption and household income in South Korea, which raises doubts as to whether the conclusions drawn from this study are applicable in other areas.

A large-scale study in Australia by Lan et al. (2020) on postal code resolution uses a machine learning approach for predicting PV adoption and finds a non-linear dependence on population density and average income. Furthermore, a study in Japan at a prefecture (i.e. regional) level finds a clear positive effect of regional policy subsidies and, to a lesser extent, housing investment and environmental awareness (Zhang, Song, & Hamori, 2011). They also show a strong negative influence on installation costs. Research by Mendieta and Sarker (2018) focussed on predicting probabilities of PV adoption in New York State. They divide the state into regions of approximately 4000 people and use logistic regression and random forest to predict probabilities per building group. They group buildings by region and PV adoption, so two samples represent each region. For example, a region of 1000 buildings with 30 adopters will create one sample of *PV-adoption = False* with weight 970 and one sample of *PV-adoption = True* with weight 30. They assess the models using the AUC metric (which will be discussed in Subsection 2.1.3) and find that random forest and gradient boosting outperform the logistic regression with AUC scores of 0.76, 0.79 and 0.45, respectively.

1.2.4. PV Adoption Data Availability

Research on PV adoption inevitably involves gathering PV adoption data. The method of obtaining these datasets differs per study and research area. In questionnaire-based research, the researchers can ask participants where they installed PV. In dataset-based studies, a different source is needed. Lan et al. (2020) used a readily available dataset containing PV adoption per postal code in Australia, provided by the APVI (2022). Zhang et al. (2011) also used a dataset containing regional PV data, collected from NEF (2022).

Not all dataset-based studies had access to PV adoption datasets. For example, (Lee & Hong, 2019) did a manual collection of data. Using high-resolution imagery, they determined which roofs had PV installed in a neighbourhood within Gangnam, Seoul, South Korea. Such a method is possible for a relatively small study size of 5000 buildings. However, for any study that investigates buildings on a national or province scale, collecting PV adoption data manually is not a preferred option as this would take a significant amount of time.

PV installations in the Netherlands were partly registered in the production installation register (PIR). This dataset was used by Van der Kam et al. (2018). Registration in the PIR was voluntary; therefore, one can regard it as incomplete because active installations might not have been adequately registered in the PIR (Aarsen et al., 2015). In 2020, the PIR was replaced by the Central Registration of System Elements (CERES) (CBS, 2021). The Dutch Central Agency for Statistics (CBS) uses the data from PIR and CERES. Kadaster, a Dutch organisation responsible for keeping public registers, also has national data on PV adoption (Kadaster, 2022). A recent evaluation of the PV adoption datasets from the CBS and Kadaster demonstrated that they have significant differences (CBS and Kadaster, 2021). A follow-up research carried out a quantitative study and found that the CBS and Kadaster datasets have a false positive rate of 4% and 8% respectively, and false negative rates of 14% and 11% respectively (false positive and false negative rates are further explained in Subsection 2.1.3). Furthermore, they demonstrated that it is currently impossible to combine these two datasets automatically (CBS and Kadaster, 2022).

Aarsen et al. (2015) examine the different possibilities for PV registration in the Netherlands. One of their conclusions is that a reliable database could be set up with the help of aerial imagery.

Several companies in the Netherlands, such as Sobolt (2022a) and NEO (2022), already provide PV adoption data based on aerial imagery. Both use machine learning algorithms for the detection of PV. Without manual verification, NEO reaches a minimum accuracy of 80% and with manual verification, a minimum accuracy of 95% (NEO, 2022). Sobolt manually verifies all data and guarantees a true positive rate of over 95% and a false positive rate of at most 1% (O. Fabius, personal communication, October 27, 2022).

1.2.5. Geometric Building Analysis

Apart from predicting PV adoptions, other studies focus on analysing building geometries to generate characteristics useful for PV prediction. For example, Mohajeri et al. (2017) provide a method for rooftop classification into six different classes based on a total of 35 attributes. They analyse approximately 10 000 rooftops in the Geneva region, Switzerland. 6% of the roofs were labelled manually, and within this group, 66% of the roof shapes were identified correctly within six classes using a support vector machine.

A similar study in Switzerland by Assouline, Mohajeri, and Scartezzini (2017) uses roughly the same methodology for creating attributes; the variables of slope and aspect (i.e. azimuth) are binned first, and the bin values are then used as input for classification. In the case of the latter study, a random forest algorithm is used instead of SVM. A similar accuracy of 67% is reached. Gooding, Crook, and Tomlin (2015) also use lidar data for classifying roof shapes and are able to reach an accuracy of 87%. They also estimate the roof area, orientation, and slope.

A different way of assessing PV potential is to detect obstacles on roofs. Apra et al. (2021) propose three different approaches. The first approach combines airborne laser scan data and 3D building models. The 3D building models are simplified models. Any points in the obtained point cloud from laser scans that deviate significantly from the planes of the 3D building model can be considered part of an obstacle. The second approach applies k-means clustering, an unsupervised machine learning algorithm, to aerial images to separate roof surfaces from obstacles. The third approach also processes aerial images but uses a supervised learning algorithm, a convolutional neural network. The three approaches were combined to obtain a boolean value for each pixel. The overall accuracy of obstacle detection changes between 2% and 60%, and the detection of the main-roof area showed an accuracy of 70% to 90%.

Song et al. (2018) also combine aerial images and airborne laser scan data. First, they classify the roofs into five categories and then estimate rooftop PV potential. Similarly to Mohajeri et al. (2017), Assouline et al. (2017) and Gooding et al. (2015), they use a height raster dataset with a pixel size of 2 x 2 m. On the contrary, Apra et al. (2021) used point cloud data with an average density of 8 points per square meter, which allows for substantially more accurate height data. Apra et al. (2021) is the only one of these studies that proposes a methodology for obstacle detection. The difference in resolution impacts the quality of the roof classification. Moreover, a more detailed mesh generally results in a more precise estimation of PV potential (Alam, Coors, Zlatanova, & Oosterom, 2016).

1.2.6. Research Gap

This section so far has highlighted some conclusions concerning factors that influence the adoption of PV. Table 1.1 provides an overview of the discussed studies and their findings relevant to this thesis. One of the major conclusions is that the literature generally does not agree on which factors are dominant in most regions or situations. Differing methods reach various conclusions with different accuracies and within different regions.

Some studies have data per household but are limited to a few thousand respondents. In contrast, others have nationwide data but can only analyse a broad level, like PC4 or prefecture. In the Netherlands, no research has been done either on a household or building level. Several socioeconomic datasets are openly available on PC4, PC5 and PC6 levels, as will be elaborated on more in Chapter 3 (as explained previously, PC4 indicates the numerical part of the postal code, PC5 indicates only one letter is included and PC6 indicates the full postal code, e.g. 2628 CD). However, these have not yet been combined with a large-scale, high-resolution dataset on PV adoption, possibly because centralised PV administration is unavailable on a PC6 level.

Furthermore, current research mainly focuses on socioeconomic factors. The dataset-based studies do not often take geometric factors into account, and in questionnaire-based studies, this has not been researched either. There is only a limited amount of studies concerning the influence of building shape, type and location, even though some studies show a significant influence of building type. The combination of these four factors could potentially yield increased knowledge on important factors, but this is not known yet.

These two gaps in current studies form the basis of this study. Combined study of the influence of geometric and socioeconomic building factors still has not been carried out in a big data-based approach to building resolution.

Table 1.1: An overview of discussed studies.

Method	Authors	Study Area	Study Size	Main conclusions
Questionnaire	Vasseur and Kemp (2014)	Netherlands	817 respondents	Positive influence of affordability and potential energy savings
Questionnaire	Qureshi et al. (2017)	Pakistan	36 respondents	Positive influence of affordability and environmental impact
Questionnaire	Bashiri and Alizadeh (2017)	Tehran, Iran	345 respondents	Negative influence of income, positive influence of environmental concerns, knowledge of renewable energies, innovativeness and household size
Meta-analysis	Schulte et al. (2022)	-	24 studies	Perceived benefits are the strongest determinant of adoption intention
Dataset	Lan et al. (2020)	Australia	2658 postal code areas	Influence of income is dependent on the population density
Dataset	Van der Kam et al. (2018)	Netherlands	4052 postal code areas	Positive correlation between PV adoption and household size, percentage of people aged 45-65, building footprint area and percentage GroenLinks voters
Dataset	Kausika et al. (2017)	Apeldoorn, Netherlands	22 postal code areas	Positive correlation between PV adoption and income
Dataset	Lee and Hong (2019)	Gangnam, Korea	5000 buildings	Correlation with the building type, the number of neighbouring adopters (positive), household density (negative) and land price (negative)
Dataset	Zhang et al. (2011)	Japan	47 prefectures	Positive effect of regional policy subsidies and to a lesser extent housing investment and environmental awareness
Dataset	Mendieta and Sarker (2018)	New York, United States	5000 regions	Random Forest and gradient boosting significantly outperform logistic regression, maximum AUC score of 0.79

1.3. Research Questions

The previous section highlighted the research gap; no study on geometric building factors and socioeconomic factors has been carried out in a dataset-based approach on a building scale. It is of great use to have more knowledge of the characteristics of adopters to target energy transition campaigns better. Although studies have already found some primary factors that influence the decision, the results differ greatly per research and research area. Therefore, additional research using larger datasets of better spatial resolution can lead to more accurate decision-making. The proposed research aims to do this by developing a model that predicts the probability of PV adoption during the study period by analysing both socioeconomic and geometric characteristics of the considered rooftop and accompanying building.

The research question is formulated as "*To what extent can we predict PV adoption at building scale using socioeconomic and geometric features?.*"

Four sub-questions have been developed to aid in answering this research question:

1. *Which features can we generate from geometric and socioeconomic data to predict PV adoption?*
2. *How can we accurately create and assess PV adoption prediction models?*
3. *Which geometric and socioeconomic features are important to the model when predicting PV adoption?*
4. *Which areas within the study area have a high chance of PV adoption according to the used model?*

The study will be carried out with data from Dutch buildings. Sobolt has manually verified the data for certain areas, as explained in Section 2.2. This manual verification increases the quality of the data in these areas. Therefore, the research will focus mainly on the provinces of Overijssel and Noord-Holland. The research will focus on the period between 2019 and 2021. For most areas in the research area, PV presence data is available at least two moments within this period.

1.4. Reading Guide

This report consists of five chapters. The research context and questions have been explained in the first chapter, Chapter 1. Chapter 2 provides background information relevant to predicting PV adoption by describing relevant literature and data sources. Then, Chapter 3 explains the method in four steps; data preprocessing, feature generation, model training and performance assessment. Chapter 4 presents the results obtained in this study. Furthermore, the results are put into context, and their applicability is also discussed in this chapter. Finally, the last chapter, Chapter 5, contains the key points, main conclusions, and recommendations for further research.

2
Predicting PV Adoption

This chapter provides background information that is relevant to the topic of predicting photovoltaic system (PV) adoption. Section 2.1 describes literature relevant to this study in terms of methods. Section 2.2 gives an overview of the data used in this study and its sources.

2.1. Related Work; Methods for Predicting PV Adoption

The following section provides an overview of techniques that have been used in other studies that are relevant to this study. Section 1.2 already provided an overview of related work that researched PV adoption, PV data availability and geometric building analysis. This section elaborates on related work about potentially applicable methods. It explains theoretical background information for two machine learning models, outlier detection, performance metrics, validation methods, feature importances and the variance inflation factor. Please note the difference between these this section and Chapter 3; this section discusses the theoretical background of already existing methods not developed by the author. Chapter 3 will discuss the implementation of these methods in this study and elaborate on methods developed specifically for this thesis.

2.1.1. Machine Learning Models for Probability Prediction

Biau and Scornet (2016) explain that "to take advantage of the sheer size of modern data sets, we now need learning algorithms that scale with the volume of information while maintaining sufficient statistical efficiency." With a dataset of over 640 000 buildings, a machine-learning model has the potential to reflect complex data structures. Moreover, Lan et al. (2020) showed that non-linear effects play a role in PV adoption prediction. Because conventional regression models cannot take non-linear effects into account, and because of the dataset size, this research chose to use a machine learning model for predicting PV adoption. This research used two machine learning models, and these two paragraphs provide a theoretical description of both.

Random Forest

The random forest (RF) algorithm is based on decision trees (Breiman, 2001). RFs are proven successful as a general-purpose and regression method (Biau & Scornet, 2016). Furthermore, the dataset in this study is imbalanced (which will be further explained in Subsection 2.1.3), and RFs are shown to provide superior results on imbalanced data when compared with other learners (Khoshgoftaar, Golawala, & Van Hulse, 2007).

For creating one decision tree, features are used to split data into increasingly homogeneous groups with respect to the output feature. In the case of growing a regression tree, the algorithm searches for a split criterion that results in the lowest mean squared error (MSE) of the two groups. If the groups are of unequal size, the weighted average is used.

This process will be explained by using a simplified example. Figure 2.1 shows an example of a regression tree built with eight samples. The output feature here is PV. If a sample has adopted PV, the value for PV equals 1; otherwise, it equals 0. In this example, every sample has one feature, the roof area A. The top box is the so-called root node, which contains all samples. The algorithm will look for a split criterion that causes the largest decrease in MSE. The split criterion, in this case, is $A > 45$ m^2. The left box, called a node, contains all samples with a roof area smaller or equal to 45 m^2. The mean value of PV, denoted by $\overline{PV}$, and the MSE of the group can now be calculated:

$$\overline{PV} = \frac{1}{5}(0+0+0+0+1) = 0.2 \tag{2.1}$$

$$MSE = \frac{1}{5}\left((0-0.2)^2+(0-0.2)^2+(0-0.2)^2+(0-0.2)^2+(1-0.2)^2\right) = 0.16 \tag{2.2}$$

Similarly, the values of $\overline{PV}$ and MSE for the right node can be computed to be 0.67 and 0.22, respectively. The (weighted average) MSE, after the split, is

$$MSE = \frac{5}{8}\cdot 0.16 + \frac{3}{8}\cdot 0.22 = 0.18 \tag{2.3}$$

so this split has decreased the MSE from 0.23 to 0.18. Now consider Figure 2.1 to be the entire tree that predicts the probability of PV adoption based on the roof area. A building with a roof area of 30 m^2 will end up in the left node and will, therefore, have a predicted PV adoption probability of 0.2. Likewise, a building with a roof area of 60 m^2 will have a predicted PV adoption probability of 0.67.

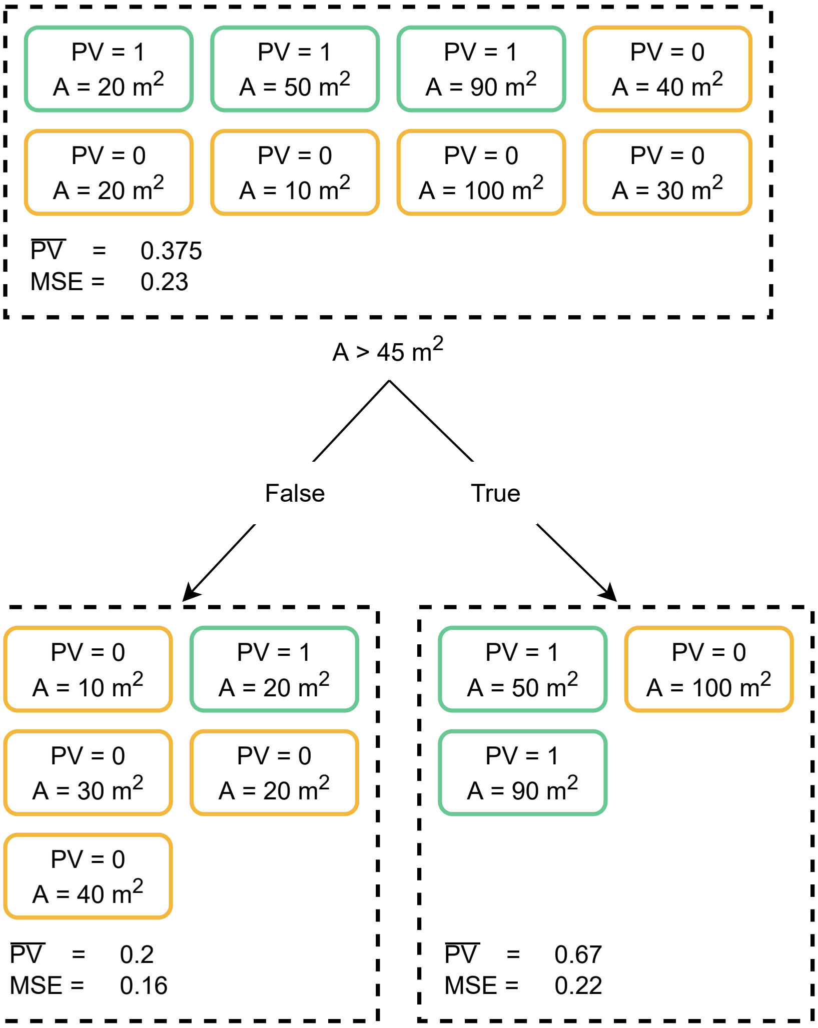


Figure 2.1: An illustration of a random forest regression With a depth of 1

The training process of a decision tree focuses on finding the best split possible by calculating which feature provides the best split and which split value is optimal for this node. This process is repeated until the maximum allowed tree depth is reached. The tree depth is the number of splits before a final node, the so-called leaf node, is reached. The example in Figure 2.1 has a depth of 1.

The final size of the tree is important. The larger the tree, the more difficult the results are to interpret. Smaller trees are easier to understand but might not adequately reflect complex data structures. Increasing the tree size increases performance on the training dataset. However, the goal is to maximise its predictive accuracy on a new dataset rather than on the training data. If the tree size becomes too large, there is a risk that it will fit the noise in the data by memorising peculiarities instead of finding a general rule. This phenomenon is called "overfitting" (Dietterich, 1995; Kruppa, Ziegler, & König, 2012).

Although using a single tree is possible, it generally provides poor estimates (Kruppa, Schwarz, Arminger, & Ziegler, 2013). In a random forest, multiple decision trees are trained. To ensure diverse decision trees, bootstrapping can be used; a fraction of the data is held back when training each decision tree, and the missing data is different for each tree. Similarly, it is possible to include only a limited number of features for a split when choosing which feature to use to split the data.

In the case of classification, the output is determined by majority voting. In the case of regression, the output is determined by taking the average output value of each tree. Using multiple trees decreases the risk of overfitting, and using more trees increases the accuracy (Breiman, 2001). Figure 2.2 illustrates random forest regression with 600 decision trees.

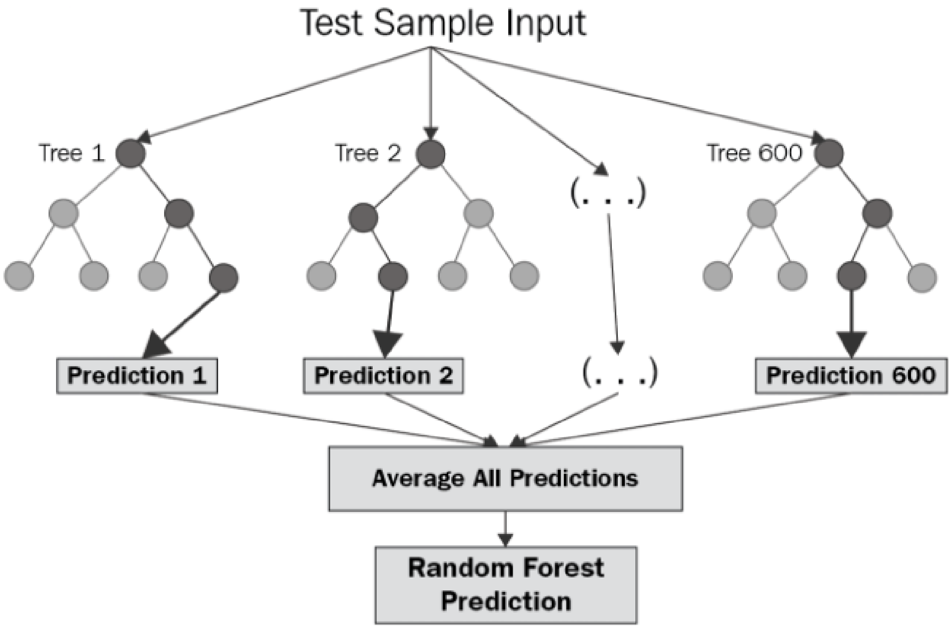


Figure 2.2: An illustration of random forest regression with 600 decision trees(Bakshi, 2020).

Neural Network

An artificial neural network (NN) is a machine learning method based on nodes and layers. One form of an NN is the Multi-Layer Perceptron (MLP). The MLP consists of several layers, where each layer is built up of one or several nodes (Taud & Mas, 2018). The layers can be summarised as:

1. One input layer
2. One or several hidden layer(s)
3. One output layer

The input layer consists of the input variables, and the number of nodes in the layer is, therefore, equal to the number of features. Figure 2.3 shows an example of an MLP with a total of *n* input features. The second layer in the example is the only hidden layer in this case, with nodes indicated by *a*. Each node obtains a value by summing the values from the input layer, each with a different weight. The value of a_1 can be calculated as:

$$a_1 = \sum_{n=1}^{N} w_n \cdot x_n \tag{2.4}$$

where w indicates the weight and x is the value of an input node. x_0 always has a value of 1, and w_0 is called the bias. This procedure is repeated for each of the k nodes in the hidden layer. The output in the example is calculated by summing the values of all nodes in the hidden layer and applying an activation function. This activation function is often non-linear, e.g. the hyperbolic tangent, unit step or logistic sigmoid function. Although the example in Figure 2.3 contains only a single hidden layer, an MLP can have multiple hidden layers consisting of a variable number of nodes.

The output can be calculated and compared to the true training value for a certain set of weights. MLP weights can be corrected by propagating the errors from layer to layer, starting with the output layer and working backwards, hence the name backpropagation (Taud & Mas, 2018).

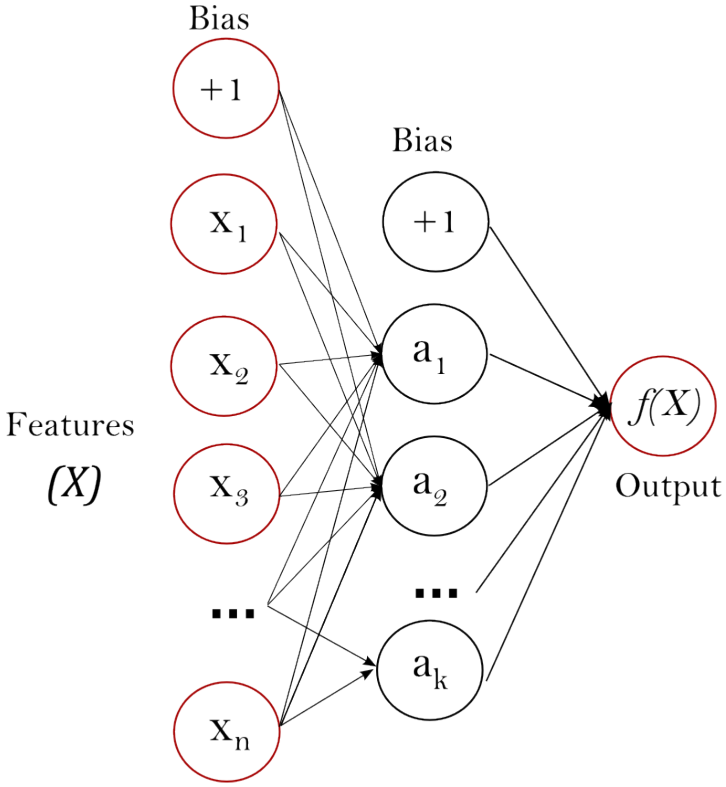


Figure 2.3: An MLP with one hidden layer. (Scikit-learn developers, 2022a)

2.1.2. Local Outlier Factor

The dataset used for this study uses a total of 640 000 samples and 15 features. The size of the dataset complicates finding outliers in the data by manual inspection. However, it can be useful to scan the data for potential anomalies that could have been caused by errors in input data or data processing. Two well-known outlier detection algorithms are Local Outlier Factor (LOF) and Isolation Forest (IF). IF only considers global outliers, unlike LOF, which also performs well at detecting local outliers (Cheng, Zou, & Dong, 2019). The downside of LOF is its computational complexity. However, for the dataset in this study, applying an LOF algorithm required less than 5 minutes of runtime; hence, the decision was made to favour LOF over IF.

The LOF of an object can be viewed as the degree of being an outlier (Breunig, Kriegel, Ng, & Sander, 2000). The degree expresses how isolated the sample is compared to its neighbourhood. More specifically, the LOF compares the density at the sample point to the density of the n nearest neighbours. A larger difference in density leads to a higher LOF. A 'normal' non-outlier sample in a cluster will have an LOF of approximately 1 since the density of the sample is roughly equal to that of its neighbours. An isolated sample will have a substantially lower density than its neighbours, causing a larger ratio between the densities and, thus, a larger LOF.

2.1.3. Model Performance Metrics

Probability prediction models predict a value in the range [0,1]. The ground truth is a boolean, true or false, which can also be expressed as a 0 or 1. For models with a scalar output but a boolean ground truth, it is, therefore, not always trivial to assess the model's performance. Furthermore, the dataset, which will be further explained in Section 2.2, is imbalanced. A data set is considered imbalanced if one class (the minority class) contains significantly fewer samples than the other class or classes (majority class) (Bekkar, Kheliouane Djemaa, & Akrouf Alitouche, 2013). This can lead to poor results of learning algorithms due to low scores for classes underrepresented in the training data. Consider a neighbourhood where 1% of the roofs have PV adopted and 99% do not. A naive model that will predict all roofs not to have PV adopted will have an accuracy of 99/100=99%, which is excellent as an evaluation measure, but a useless model in a real application. Different performance metrics exist for imbalanced datasets in probabilistic classification methods. This subsection explains two often used in other studies; the Area Under Curve and Brier score.

Area Under Curve

An often used performance metric for the classification of imbalanced data sets is the area under the curve (AUC) of the receiver-operating characteristic (ROC) curve, which is also recommended by Bradley (1997). For example, the AUC measure is used to evaluate RF accuracy in forest fire probability mappings (Milanovic et al., 2020) and consumer credit risk probabilities (Kruppa et al., 2013).

The ROC curve, of which Figure 2.4 presents an example, typically shows the relationship between the sensitivity and specificity of a binary classifier. The sensitivity or true positive rate (tpr) measures the proportion of positives correctly classified; specificity or true negative rate (tnr) measures the proportion of negatives correctly classified. Conventionally, the true positive rate is plotted against the false positive rate (fpr), which is one minus true negative rate (Flach, 2016). A confusion matrix as shown in Table 2.1 illustrates this more clearly in combination with equations Equation 2.5, 2.6 and 2.7.

Table 2.1: A confusion matrix.

True Class	Predicted Class	
	Negative	Positive
Negative	T_n	F_p
Positive	F_n	T_p

$$\text{tpr} = \frac{T_p}{F_n + T_p} \tag{2.5}$$

$$\text{tnr} = \frac{T_n}{F_p + T_n} \tag{2.6}$$

$$\text{fpr} = \frac{F_p}{F_p + T_n} \tag{2.7}$$

Predicted probabilities can be converted to booleans by choosing a threshold; anything above this threshold is predicted as true, and everything below the threshold is negative. A low threshold causes a higher fpr, and a high threshold causes a higher fnr. This dependency becomes clear from the ROC-curve, of which Figure 2.4 is an example. The orange line depicts the curve for a model that predicts equal probabilities for all samples or a model that predicts probabilities that are uncorrelated to the true class. The green line shows an example ROC-Curve. The area underneath this curve, the aforementioned AUC, is 0.71 in this case. A perfect model has an AUC of 1, and a poor model with an AUC of 0.5. In general, an AUC of 0.5 suggests no discrimination (i.e., ability to distinguish roofs with and without PV), 0.6 to 0.7 is considered fair, 0.7 to 0.8 is considered good, 0.8 to 0.9 is considered very good, and more than 0.9 is considered excellent (Bekkar et al., 2013).

The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Brahmachari, Jain, & Kimmig, 2013). In other words, for a model with an AUC of 0.75, the following can be said: if the model predicts the probability of PV for a random instance without PV and for a random instance with PV, then a 75% chance exists that the predicted probability is higher for instance with PV. The AUC is a measure of correct ranking, and it does not directly evaluate the correctness of the predicted probabilities but only the order in which the instances are ranked. If, for instance, all predicted probabilities are divided by 10, the AUC would remain the same. Therefore, another scoring metric is used to evaluate the predicted probability: the Brier score.

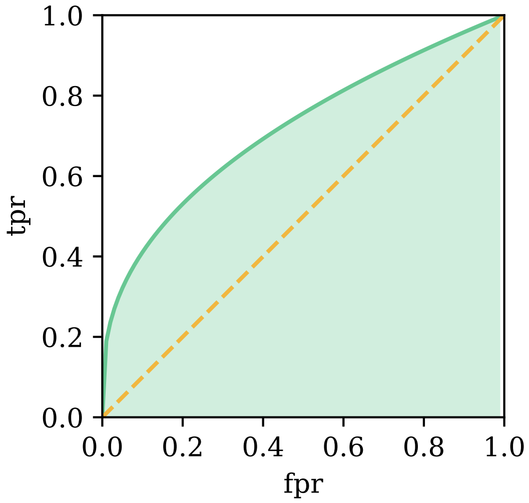


Figure 2.4: An ROC-curve example; the true positive rate plotted against the false positive rate. The orange line shows the curve for a poor model; the green curve is an example of a curve for a model with an AUC value of 0.71.

Brier Score

The Brier score can be explained as the predicted chance of an event (f_t) minus the actual event (o_t), squared. The actual event is represented as a 1 if it is true and 0 if false. For example, if the model predicts a 70% chance that a roof has PV and the roof indeed has PV, the brier score is $(0.7-1)^2 = 0.09$. If the model predicts a 40% chance and the roof has no PV, the Brier score is $(0.4-0)^2 = 0.16$. The Brier score on a test dataset is calculated by averaging the values of each observation. The Brier score (BS) can thus be formulated as

$$\mathrm{BS} = \frac{1}{N}\sum_{t=1}^{N}\left(f_t - o_t\right)^2 \tag{2.8}$$

where N is the number of observations (Rufibach, 2010). A lower Brier score indicates a better-performing model, with a score of 0 indicating a perfect model and 1 indicating the worst model possible. The Brier score is mathematically the same as the mean squared error. This property makes it easier to train models based on this criterion, which is harder to implement for AUC. The Brier score is not often used as an absolute score but can be used to evaluate relative model performance.

2.1.4. Model Validation

To assess the performance of a model, its performance on a test dataset can be computed. It is important that the test dataset and train dataset are independent and identically distributed (i.d.d.). This ensures that the performance measures how well the model performs on new data and not how well it is trained, which can be a case of overfitting.

K-fold Cross Validation

In k-fold cross-validation (CV), the data is randomly split up into k groups of approximately the same size. First, group 1 is left out as a test dataset, and the model is trained using the other k-1 groups. The model is then scored using the data in group 1. Then, group 2 is left out and the process is repeated. The final score of the model is obtained by averaging all the scores from the different runs. Figure 2.5 shows an example of CV for $k = 5$. The advantage of CV is that all data is used for testing and training. The measured model performance can depend on how the data is split, and this effect is reduced by using CV. Furthermore, comparing the performance scores from the different folds can indicate the standard deviation of the model performance. A downside of CV is the higher computation time; using CV takes k times as much time compared to splitting the data into testing and training data only once.

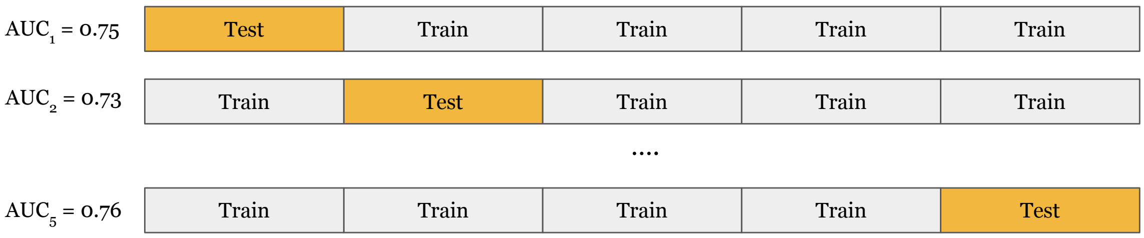


Figure 2.5: An example of CV.

Nested k-fold Cross Validation

Most machine learning algorithms have so-called hyperparameters. These are parameters that the model does not tune but are set beforehand. For example, in the case of RF, these can be the tree depth, the number of trees, the number of samples used for each split and the number of features considered for each split. If these parameters are optimised using CV, the best model performance is found for the training dataset. There is, however, no guarantee that these hyperparameters provide the same performance on another dataset. In other words, optimising hyperparameters using CV can lead to overfitting and provide a too-optimistic performance measure of the model (Cawley & Talbot, 2010). To prevent the overfitting of hyperparameters, nested k-fold cross-validation (NeCV) can be used.

In NeCV, the data is again split into k_o folds for the so-called outer loop. In this loop, folds are left out one by one as the test dataset. The other folds are then split into k_i folds for the inner loop, where CV with k_i folds is used to find the optimal hyperparameters. Once the inner loop is completed, the model is retrained on all data present in the inner loop using the optimal hyperparameters. The model's performance is then scored based on the test dataset from the outer loop. Then the next fold of the outer loop is held back for testing and the process repeats. The final score of the model is found by averaging the scores from the outer loop. An example of NeCV is shown in Figure 2.6, for $k_o = 3$ and $k_i = 5$. This technique separates the data used for optimising the hyperparameters from the data used for performance evaluation. Therefore, it prevents the overfitting of hyperparameters. Similarly to CV, another advantage is that multiple splits in the data are used, which allows a more reliable assessment of the model. The downside of NeCV is the computation time. The example in Figure 2.6, with $k_o = 3$ and $k_i = 5$, evaluates 100 parameters in 5 inner folds for each of the 3 outer folds, which requires $100 \cdot 5 \cdot 3 = 1500$ times as much time compared to the method of splitting the data into testing and training data only once.

2.1.5. Feature Importance Determination

Several methods exist to determine how important each feature is for the MLM. Every method has advantages and disadvantages; therefore, multiple metrics will be used. A popular measure is a decrease in impurity. However, this method can be misleading for high cardinality features, i.e. features with many unique values

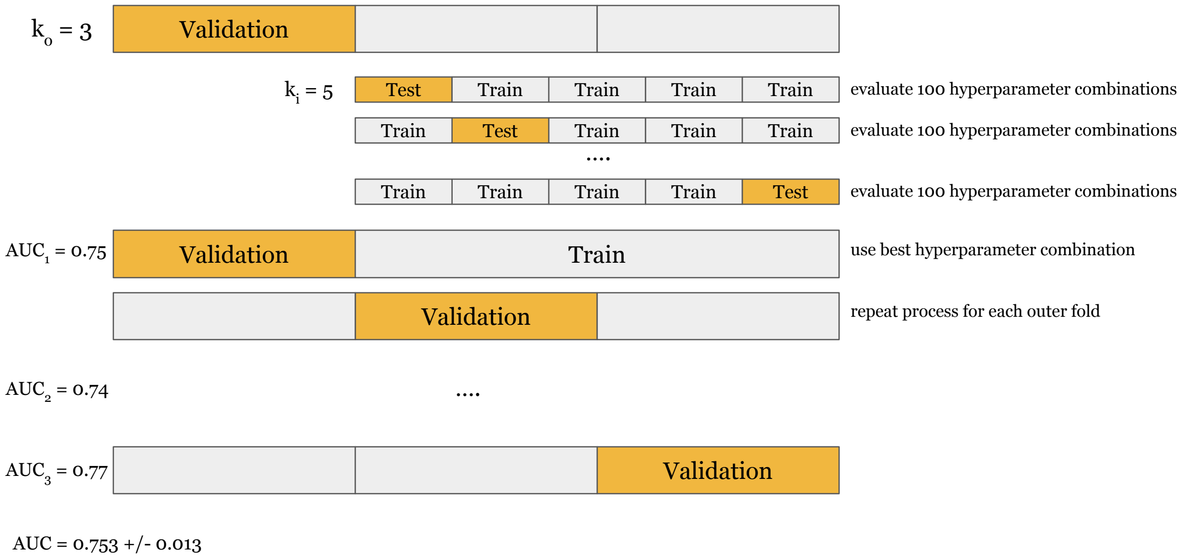

$k_o = 3$
Validation
$k_i = 5$
Test Train Train Train Train
evaluate 100 hyperparameter combinations
Train Test Train Train Train
evaluate 100 hyperparameter combinations
....
Train Train Train Train Test
evaluate 100 hyperparameter combinations
$AUC_1 = 0.75$
Validation
Train
use best hyperparameter combination
Validation
repeat process for each outer fold
$AUC_2 = 0.74$
....
$AUC_3 = 0.77$
Validation
AUC = 0.753 +/- 0.013

Figure 2.6: An example of nested k-fold cross-validation (NeCV).

(Scikit-learn Developers, 2022). The method is biased in favour of variables with many possible split points (Nembrini, Konig, & Wright, 2018). Many features in this research have high cardinality, e.g. roof area, while others are limited to only two values. Therefore, this method was not used in this research.

A method that is considered more accurate than the decrease in impurity is permutation importance. This method shuffles the data of each feature one at a time and then measures the decrease in model performance. If a feature is of high importance for the model, then shuffling the input values of only this feature should lead to a significant decrease in model accuracy. High multicollinearity between the independent variables can negatively impact the accuracy of variable importance analysis. Multicollinearity means that a correlation exists between the predictor variables. The effect of multicollinearity on permutation importance can be explained using the following example: household income and property value are likely to have a high correlation. Residents with higher incomes tend to live in more expensive houses. Assume both are good predictors of PV adoption, and a model trained on these features performs well. If only the household income is permuted, then the model still has the correct values for property value. Because property value and household are correlated, most information that used to be in the model is still present, and the performance of the model with the shuffled household income values will be nearly identical to the original model. Based on the permutation feature importance score, this would imply that household income is relatively unimportant for the model.

Similarly, the property value can also be found to be unimportant. This is a downside of the permutation feature importance method. Therefore, it is important to investigate the correlation between the input features beforehand and assess the feature importance in multiple ways.

Another way to investigate a feature's importance is by investigating its performance on its own. The AUC or Brier score can also be computed from a model trained on only a single feature. The disadvantage of considering only one feature is that this method does not consider the interplay between features. For instance, Lan et al. (2020) found that the influence of income on the PV ratio is not linear but dependent on the population density in the area. In this way, population density is a valuable feature but has low predictive performance when considered as a single feature. These kinds of feature influences cannot be found using single feature analysis but require a method such as permutation importance.

2.1.6. Variance Inflation Factor

The previous subsection explained how high multicollinearity between the independent variables could negatively impact the accuracy of variable importance analysis. Multicollinearity can be quantified using the variance inflation factor (VIF). To calculate the VIF, one needs to compute how well the other independent variables explain the variance of a certain variable. This is quantified by doing regression analyses of one variable on all other variables in the analysis to obtain an R^2 value. The formula for the VIF then is

$$\text{VIF} = \frac{1}{1 - R^2} \tag{2.9}$$

VIF has a minimum value of 1 and no maximum. If a variable is completely uncorrelated to the other variables, R^2 equals 0, resulting in a VIF of 1. As a rule of thumb, the VIF should not exceed a value of 10 (Miles, 2014)

2.2. Data for Predicting PV Adoption

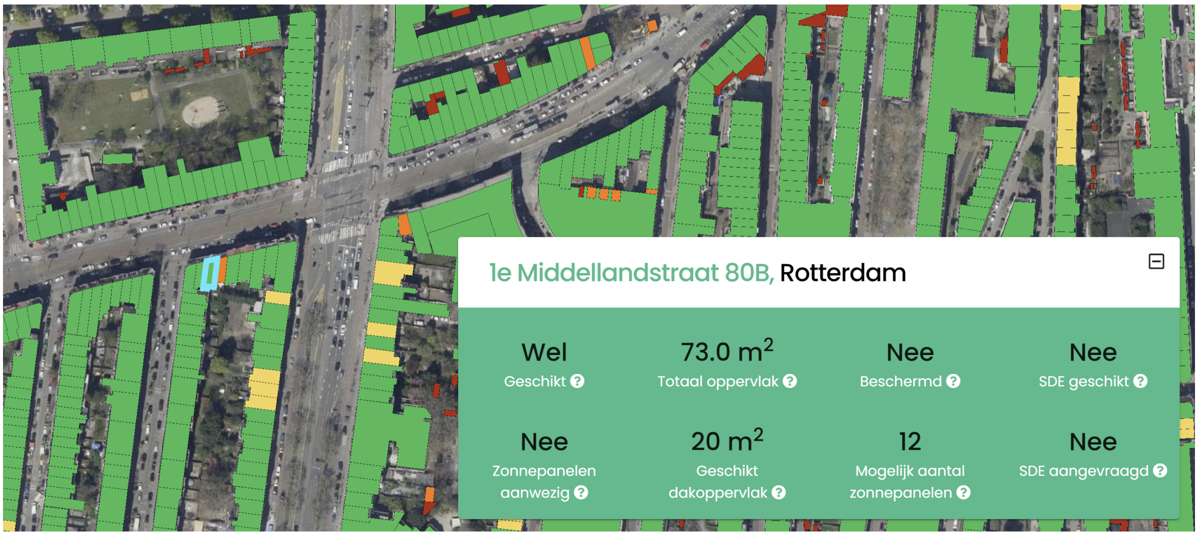
This section provides an overview of all datasets used in this research and the preprocessing steps. It can be divided into three main categories; PV adoption data, socioeconomic data and geometric data, as Table 2.2 points out. Each category will be elaborated upon in a different subsection in this chapter.

Table 2.2: An overview of data sources used for this study.

Category	Source	Contains	Spatial Resultion	Coverage
Geometric	BAG	Building Registration	Building m	Netherlands
Geometric	AHN	Height Raster	0.5 m	Netherlands
Geometric	3D BAG	Rooftop geometries	Building	Netherlands
PV Adoption	Sobolt	Verified Rooftop PV Presence	Building	Overijssel, Noord-Holland
Socioeconomic	CBS	Socioeconomic statistics	PC6, PC4	Netherlands

2.2.1. PV Adoption Data

Sobolt provides data on PV adoption. Sobolt is a company in Rotterdam, The Netherlands, that creates artificial intelligence (AI) based solutions for various problems related to a sustainable future (*Home - Sobolt*, 2022). The company has developed an algorithm to detect solar panels based on aerial images. After classification, the data is manually checked to increase accuracy. Sobolt only has data for municipalities and provinces that are clients of their software product, called Zonnedakje (Sobolt, 2022b). An example of a Zonnedakje visualisation is shown in Figure 2.7. The provinces of North Holland and Overijssel are fully available, as well as several municipalities. The dataset contains approximately two million buildings. Depending on the region, PV adoption data is available for one or several years. For Overijssel, both 2019 and 2021 are present in the dataset.



1e Middellandstraat 80B, Rotterdam

Wel
Geschikt

73.0 m^2
Totaal oppervlak

Nee
Beschermd

Nee
SDE geschikt

Nee
Zonnepanelen aanwezig

20 m^2
Geschikt dakoppervlak

12
Mogelijk aantal zonnepanelen

Nee
SDE aangevraagd

Figure 2.7: A Screenshot of the online Zonnedakje tool. Yellow indicates PV presence on a rooftop, red non-suitable roofs, orange possibly suitable roofs and green suitable roofs.

Data Quality

Subsection 1.2.4 addressed various PV adoption data sources in the Netherlands, including their advantages and disadvantages. Sobolt guarantees a true positive rate of over 95% and a false positive rate of at most 1% (O. Fabius, personal communication, October 27, 2022).

2.2.2. Socioeconomic Data

A publicly available dataset from the Dutch Central Agency for Statistics (Centraal Bureau voor de Statistiek or CBS in Dutch) provides socioeconomic and demographic data on postal code level (CBS, 2022b). The dataset is available every year from 2015 until 2020 and aggregates by either PC4, PC5 or PC6. PC4 indicates that only the numerical part of the postal code is used, e.g. 2628. PC5 indicates only one letter is included, and PC6 indicates the full postal code, e.g. 2628 CD.

Due to privacy laws and assurance of quality, some data is made unavailable for areas with a low population and therefore redacted in the dataset. The threshold for providing data differs per category, but often data is available if more than five residents or houses are present in the aggregation area. Additionally, more recent data has a higher threshold than older data, so the 2020 dataset has more redacted fields than the 2015 dataset. A total of 131 attributes per postal code is available. The CBS splits the dataset up into seven categories:

- Residents
- Housing
- Income
- Energy usage
- Social security
- Services
- Population density

This study creates features based on data from the residents, housing, income and population density categories.

CBS also provides the same data on other regional levels. These are municipalities, districts and neighbourhoods. Of these, the largest is a municipality, "gemeente" in Dutch. A municipality consists of districts ("wijken"), and districts consist of neighbourhoods ("buurten"). In an openly available dataset provided by the CBS, the aforementioned data is also available on these spatial scales, as well as the geometries of these regions, available as shapefiles (CBS, 2022c).

Data Quatlity

The CBS does not provide a separate analysis of the error margins in their data. A confidence interval is provided only for the median income, which will be further explained in Section 3.2. An anticipated source of errors is the time between measurement and data usage. Section 3.2 provides an analysis of the rate of change of the socioeconomic data that are eventually used.

2.2.3. Geometric Data

Spatial data is required to generate building attributes. This research used BAG, AHN and 3D BAG as input data. The 3D BAG dataset is created by combining BAG and AHN data, so one could argue that using the 3D BAG dataset is redundant. However, 3D BAG provides ready-to-use 3D building models which require less processing than a point cloud or height raster. Although these building models are processed data and therefore cause a loss of information, using them reduces the amount of data processing needed for this thesis.

BAG

BAG is the Register of Buildings and Addresses (or Basisregistratie Adressen en Gebouwen in Dutch), which contains data on buildings and addresses in the Netherlands (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2018). There is a distinction between addresses and buildings; some buildings, such as garages or pumping stations, do not have an address. Moreover, some buildings can have multiple addresses, e.g. porch houses or flats. The BAG dataset specifies the building footprint, status and address among several attributes. The data is gathered and managed by municipalities and nationally centralised by Kadaster. The BAG dataset is used directly in this study for building features such as building usage. Furthermore, it is also used for the BAG 3D dataset. 3D BAG uses BAG 2.0 Data.

AHN

AHN is the Up-to-date Height Model of the Netherlands (or Actueel Hoogtebestand Nederland in Dutch), acquired by airborne LiDAR (*Kwaliteitsbeschrijving | AHN*, 2022). The third version, AHN3, was collected between 2014 and 2019. This dataset was used for determining an estimated suitable PV area in Overijssel. Furthermore, the dataset is the input for creating the 3D buildings in the 3D BAG dataset. A newer version, AHN4, is also available for some locations in the Netherlands but has yet to be used for 3D BAG. However, AHN4 is available in North Holland. As a result, Sobolt uses this version to compute the suitable area. Ideally, the same version of AHN should be used for Overijssel and North Holland. However, the calculations of solar potential are computationally intensive, so it was decided to run the potential for the most recently available version only.

3D BAG

3D BAG is an open-source dataset developed by 3D geoinformation research group (2022). It merges AHN and BAG. Point cloud data from AHN is used to create a 3D model of each building by fitting planes. The dataset has three levels of detail (LoD), specified according to the so-called improved LoD specification defined by Biljecki, Ledoux, and Stoter (2016). This research uses the most detailed one available in 3D BAG, LoD 2.2. At this level, building parts also have semantics, i.e. they are classified into categories. The three categories used are ground surface, wall surface and roof surface. Figure 2.8 shows a visualization of BAG 3D in their online tool.

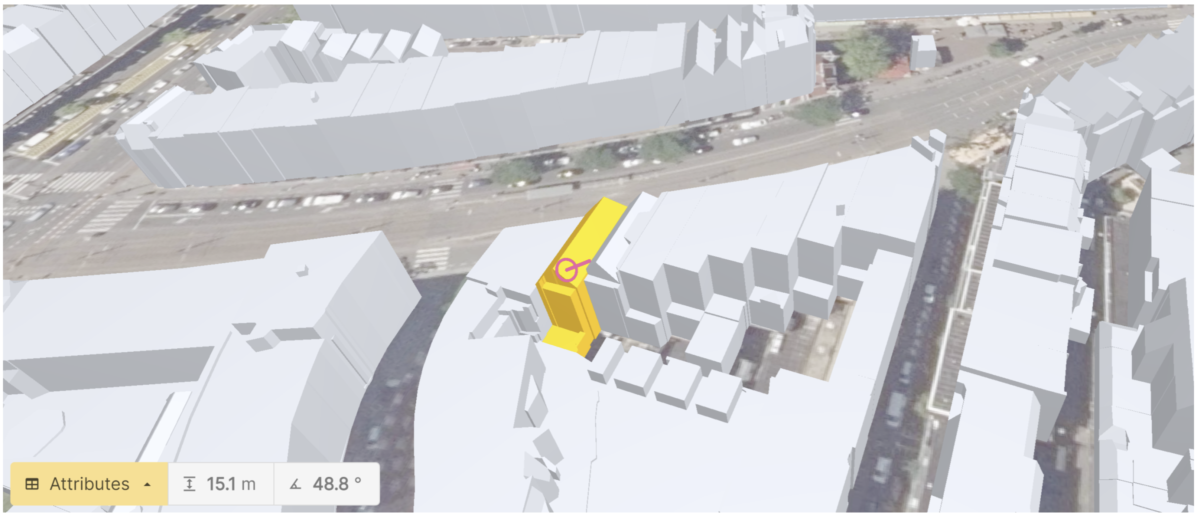


Figure 2.8: An Example of the Online 3DBAG Visualization Tool

Data Quality

Municipalities are responsible for collecting data for BAG. One of the error sources in this dataset is the delay in registration. Changes occur (new construction, conversion and demolition) of which the municipality is not or not immediately aware because there is illegal (re)construction or demolition or because the construction or demolition activities are permit-free (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2018). Furthermore, the registered year of construction has a maximum allowed deviation of one year for buildings built after 1991. For buildings built after 1900 or 1950, the maximum allowed deviations are two and five years, respectively, and these allowed deviations are increasingly larger for older buildings. However, 98% of the buildings used in this study were built after 1900, so the maximum allowed deviations in the year of construction are mostly five years or less. Although the allowances are known for BAG data, no validation data was found that assesses the actuality and reliability of the dataset.

AHN3 and AHN4 have average densities of 8 and 12 points per square meter, although variation exists between regions and buildings. The maximum systematic standard error for both is 5 cm, and the maximum stochastic standard error is 5 cm. This means that 68 % of the points in the point cloud have a height accuracy of 10 cm and 97% of the points have an accuracy of 20 cm (*Kwaliteitsbeschrijving | AHN*, 2022).

The LoD 2.2 data in 3D BAG has an RMS of 0.04 m, i.e. the error between the fitted planes and the AHN point cloud (Dukai, Peters, Vitalis, Van Liempt, & Stoter, 2021). The downside of using AHN for 3D reconstruction is that it is deemed outdated by design due to the length between scans for different regions. The area of interest was scanned between 2016 and 2019, meaning that both the raw LiDAR data and the building geometries are at most six years outdated (*Historie | AHN*, 2022). However, the building stock changes at a relatively slow pace, higher in urban regions and lower in remote areas of the country. The 3D geoinformation research group (2022) estimates that about 95% of the measured building heights are still valid (estimated for the 3D BAG generated in March 2021).

3 Methods

The research question, "*To what extent can we predict photovoltaic system (PV) adoption at building scale using socioeconomic and geometric features?*" was answered by creating multiple models. For each building in the dataset, the models predicted a probability that PV was adopted within the study period. More specifically, the dependent variable in the models was whether a building adopted PV between 2019 and 2021. Based on several independent variables, i.e. the input features, the model predicted this probability for every building. The steps to achieve this can be summarised as follows:

1. Preprocess and clean up the data
2. Generate socioeconomic and geometric features
3. Train a machine learning model (MLM) using the PV adoption data and the generated socioeconomic and geometric features
4. Assess the performance of the model and the importance of the individual features

This chapter has the following structure; Section 3.1 explains the data preprocessing and cleaning steps. Then, Section 3.2 and 3.3 elaborate on the feature generation from socioeconomic and geometric data, respectively. Section 3.4 explains the steps taken to obtain the suitable area. All features are summarised in Section 3.5 and Section 3.6 explains the method used for outlier analysis. Section 3.7 provides more detail about the implementation of the two MLMs, and, last, Section 3.8 explains how the model evaluation was done in this study. A visual summary of the method is given at the end of the chapter, in Section 3.9.

3.1. Data Preprocessing and Cleaning

Several data preprocessing actions were performed on the data before it was suitable to train and evaluate MLMs. This section elaborates on these three steps.

1. **Combining Data Sources:** Firstly, the data sources needed to be combined. Every building in the Netherlands is registered using a unique BAG-id identification. Some buildings were built or demolished after 2016, which causes differences between the datasets from BAG, 3D BAG, and Sobolt. To prevent having to deal with incomplete data, the final dataset only included buildings that were present in each dataset.

2. **Removing Protected Buildings:** Secondly, some Dutch buildings have a protected status and are classified as protected cityscape ("beschermd stadsgezicht") or as a national monument ("nationaal monument"). This protected status often means installing PV on the building roof is not or only partially allowed. The dataset contained 37405 (2.1% of total) buildings with the protected cityscape status and 7022 (0.4% of total) buildings with the national monument status, which had PV adopted on 2.4% and 0.4% of the buildings, respectively, as opposed to the dataset mean of 7.6%. These buildings only make up a small fraction of the dataset but are likely to show deviating behaviour due to their protected status. Therefore, these were not included in the analysis. Another protected status exists, called "gemeentelijk monument". However, data on buildings classified as "gemeentelijk monument" is not always available or centrally stored.

3. **Removing Buildings with Invalid Geometries:** Furthermore, in Overijssel, 2.4% of the buildings were built in 2019 or later, so these buildings do not have a valid geometry in 3D BAG. 2.7% of the buildings were built between 2016 and 2018, so for these buildings, the point cloud was collected in the same years as the year of the construction. Therefore, it is uncertain whether the geometry is valid for these buildings. To decrease the chance for errors due to this temporal mismatch, all buildings built after 2015 were not considered in this analysis, which left 94.9% of the buildings.

4. **Removing Buildings that Adopted PV Before 2019:** Lastly, buildings that had PV adopted before 2019 are classified as not having a change in PV between 2019 and 2021. The change in PV is the dependent variable in this study. However, the features of these buildings are likely to be more similar to buildings with a change in PV than to buildings without a change in PV. Including these buildings in the dataset could lead to a loss in model performance. Therefore, these buildings were excluded from the dataset. Approximately 4.7% of the buildings already had PV adopted before 2019.

5. **Calculating The Change in PV:** The dataset provided by Sobolt indicates whether PV is present on the roof for each building. For most buildings, this has been known for multiple years. This allowed investigation of the change in PV, i.e. whether a building has adopted PV between 2019 and 2021. On average, approximately 5.7% of the buildings have adopted PV in Overijssel between 2019 and 2021.

After these preprocessing steps, the dataset of the Province of Overijssel consisted of 645 681 buildings.

3.2. Generation of Socioeconomic Features

This section describes how seven socioeconomic features were generated. The first subsection, Subsection 3.2.1, describes each feature and how it was generated, and Subsection 3.2.2 explains how missing data was filled up.

3.2.1. Socioeconomic Features

This subsection provides a brief description of each feature. The socioeconomic features are surrounding PV ratio, house ownership, income, address density, number of registered addresses, year of construction and building usage.

Surrounding PV ratio

The PV ratio, i.e. the number of buildings that adopted PV divided by the total number of buildings, was computed for each neighbourhood in the area of interest. In the case of Overijssel, this considered the PV ratio in 2019.

Population 25-44

The feature *Residents aged 25 to 45 years* from the CBS indicates the number of residents aged between 25 and 44 years old in an area. This feature was normalised by dividing by the total number of residents in the area. In the Overijssel dataset, this feature was known for 80% of the buildings on a PC6 level.

Home Ownership

The feature *"Percentage of Owner-Occupied Houses"* from the CBS indicates the number of owner-occupied houses as a percentage of the total number of houses. It is redacted if there are less than 10 owner-occupied homes per area and when the share of homes with unknown owners is less than 50%. The percentages provided by the CBS are rounded to multiples of 10. The feature *"Percentage of Rental Houses"* is built up similarly for rental houses.

These two attributes are highly correlated (r = -0.61). As a result, including both in the model led to high values of the Variable Inflation Factor (explained in Subsection 2.1.6 and was therefore not done, as this could negatively impact the quality of the feature importance determination. Instead, the ratio between the two was used as an input feature. The maximum value was set to 10, so neighbourhoods with a (rounded) percentage of 0 rental houses do not have an infinitely high ratio. This combined feature was known for 24% of the buildings on a PC6 level and 99% on at least a PC4 level.

Income
For each area, the CBS compares the median standardised income per household with the income distribution for all households in the country. This is effectively a percentile score; for example, a value of 70% for a certain PC6 means that the median of this area is higher than 70% of the other households in the Netherlands. CBS does not provide this data as a scalar but as categorical. The categories are low, below middle, middle, above middle, and high. Because the number of households per zip code area is often small, the CBS considers sources of inaccuracy in recording income. For each area, the 99% confidence interval of median income is determined. When the medium income falls into one class, it is classified as such with certainty (e.g., "middle" or "high"). When the interval includes multiple classes, then this range is reported (e.g., "low to below middle" or "middle to above middle").

Categorical values are not easily inputted into an MLM. Therefore, the values were converted to numerical values, using the transformation in Table 3.1

Table 3.1: Conversion of median income to numerical values.

CBS Dataset Value	Used Feature Value
Low	10
Below Average	30
Average	50
Above Average	70
High	90
Low to Below Average	20
Below Average to Average	40
Average to Above Average	60
Above Average to High	80

The median income is the only feature for which the CBS provides uncertainty. Most postal codes have a confidence interval of 40% for the median income percentile. More specifically, the confidence interval was 20% for 30% of the buildings, 40% for 69% of the buildings and 60% for 1% of the buildings.

Address Density
This feature from the CBS is the average number of registered BAG addresses per square kilometre. This feature is only available on a PC4 level and is known for all buildings.

Number of Registered Addresses
In the BAG dataset, all addresses are linked to a building. Reversing this relationship allowed to compute the number of addresses linked to one building, and buildings with no addresses linked to them had a value of 0 for this feature.

Year of Construction
In the BAG dataset, the year of construction is known for each building.

Building Usage
In the BAG dataset, the usage of the building is known for each building. Please note that one building can have multiple uses. The usage functions are community, healthcare, industrial, office, shopping, residential, sports, accommodation, educational, cellular, and other. Most MLMs cannot deal directly with categorical values. Therefore, so-called dummies often have to be used; each of these usages was used as a separate feature. If the building has a certain function, the value equals 1; otherwise, it remains 0. The most frequent building usage in Overijssel is a residential function, for 52% of the buildings. None of the other building usages occurs in more than 5% of the cases. Therefore, only the residential function is included as a feature.

3.2.2. Dealing with Missing CBS Data

The dataset from the CBS has redacted data for some buildings, and most MLMs cannot deal with missing data. A solution would be to omit buildings for which the data is redacted, but this greatly reduces the model's applicability. Furthermore, it is likely to create a model biased towards patterns in densely populated areas since these areas have a lower chance of having redacted data.

This research replaced the missing data with data of a higher aggregation level. For example, when the income of a building is redacted on PC6 level, it can be replaced by the income on PC4 level. If the PC4 level is also redacted, the value is replaced by the median value of the entire dataset. Table 3.2 shows which features contain replaced.

A higher aggregation level decreases the uncertainty because it is less specific. To include this increased uncertainty in the model, an "aggregation feature" was added for every feature where replacement took place. This feature indicates the aggregation level. The assigned values are shown in Table 3.3. Adding an extra column also influenced the way feature importances were determined. In the case of permutation importance, both the feature column and the corresponding aggregation level column are shuffled. In the case of feature importance determination using single feature models, both columns were also considered.

Table 3.2: Overview of replacement percentages for CBS features.

Feature	PC6 Data Used [%]	PC4 Data Used [%]	Dataset Median Used [%]
Residents 25-44	78.1	21.8	0.1
House Ownership	23.6	76.3	0.1
Income	83.1	16.9	0.0

Table 3.3: Aggregation Feature Values

Aggregation Level	Value
PC6	0
PC4	1
Dataset Median	2

3.2.3. Temporal Changes in CBS Data

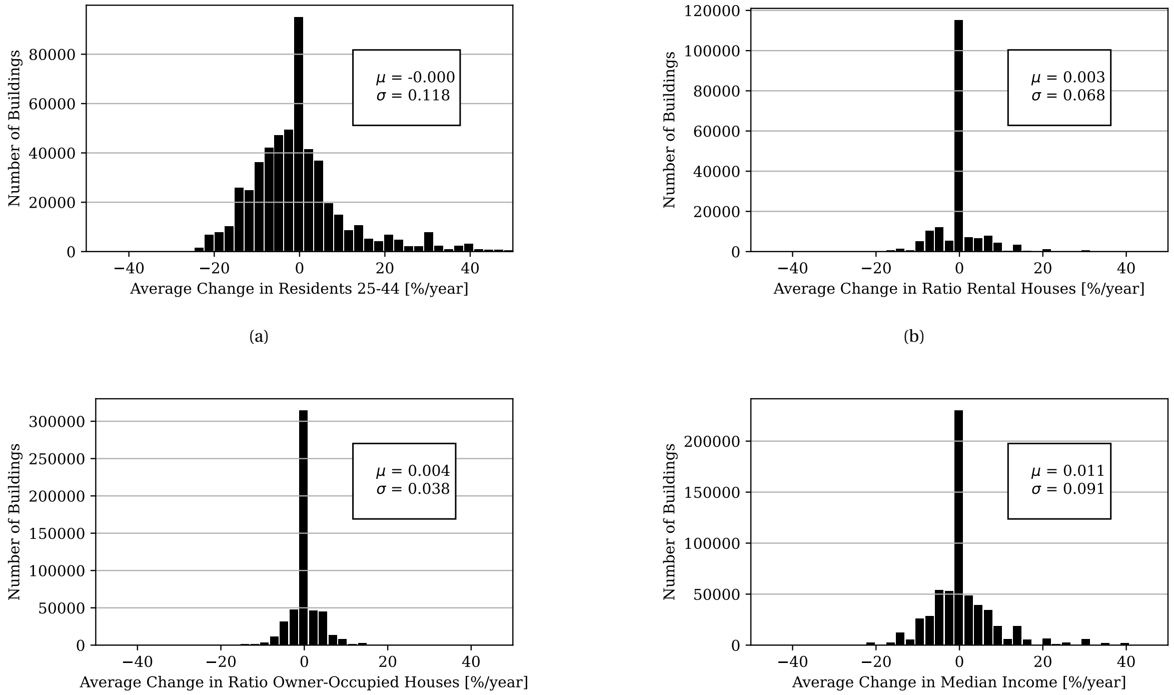
Socioeconomic feature values can change each year. Four of the features provided by the CBS were examined to investigate to what extent these can be assumed to be constant over time. Data from four years, i.e. 2015 to 2018, was used to determine the average change in feature value (using linear regression) for each postal code. Figure 3.1 displays the result of this analysis. Each histogram bin represents the number of buildings in Overijssel that were found to have the same average change in feature value. For example, the highest bin in Figure 3.1a shows that approximately 95 000 buildings in Overijssel were located in a postal code that saw an average change of 0%/year in the number of residents aged 25-44. The mean change over all buildings is denoted by μ, and the standard deviation of the change is denoted by σ. If one can assume a normal distribution, Figure 3.1a can be interpreted as follows; the value for residents aged between 25 and 44 changes less than 12%/year (=1σ) for 68% of the buildings. Similarly, Figure 3.1c shows that the ratio of owner-occupied houses in the postal code changed more than 8%/year (=2σ) for 5% of the buildings.

3.3. Generation of Geometric Features

This section describes the generation of geometric features. The method for generating features from 3D BAG data is explained, which was inspired by the work from Assouline et al. (2017). The calculation of the suitable roof area is a more elaborate method, and therefore, it is treated in a separate section, Section 3.4

3.3.1. Processing 3D BAG data

From the 3D BAG dataset, the roof surfaces of each building were extracted. These are described by the corner points of each fitted plane in the building, i.e. every surface is described by a set of co-planar points. The direction of least variation was determined using Principal Component Analysis (PCA), which is the direction normal to the surface plane. This technique was applied to each surface labelled as a roof surface according to the 3D BAG semantics.



(a)

(b)

(c)

(d)

Figure 3.1: Histograms of the average yearly change for four socioeconomic features.

A roof surface's slope (S) is defined as the angle between the horizontal plane and the normal vector. The aspect angle (α) is defined as the angle between the normal vector and a north-pointing vector, where eastwards is positive and westwards is negative (please note that often A is used for denoting aspect, but in this case α is chosen to prevent confusion with the area). As a result, α is always between -180 and +180 °. For each building, S, α and the area (A) were computed for every roof surface. These values were then aggregated to create attributes for the entire building by binning surface areas by inclination and azimuth. To create dimensionless features, these are then normalised by the total roof area. The aspect visualised as a direction is often more intuitive; a value for α of 0 corresponds to orientation to the North, while ± 180 corresponds to the South. The aspect is binned into eight bins, and the slope into six bins. As a result, each roof had a total of 14 roof surface attributes. An example of this method is shown in Figure 3.2; a building with a simple slanted roof geometry. This roof exists of 3 surfaces in the 3D BAG dataset. The described methodology yielded values as shown in Table 3.4. Four features were derived from this analysis; roof complexity, flatness, area per surface and irradiance. Two other features were not taken directly from this analysis but also originated from the 3D BAG dataset; the number of neighbouring buildings and building height.

Table 3.4: Derived attributes of roof surfaces of a single building registration in 3D BAG.

	A [m^2]	S [°]	α [°]
Surface 1	50.1	44.6	14.7
Surface 2	31.0	44.4	-163.2
Surface 3	6.2	8.5	74.9

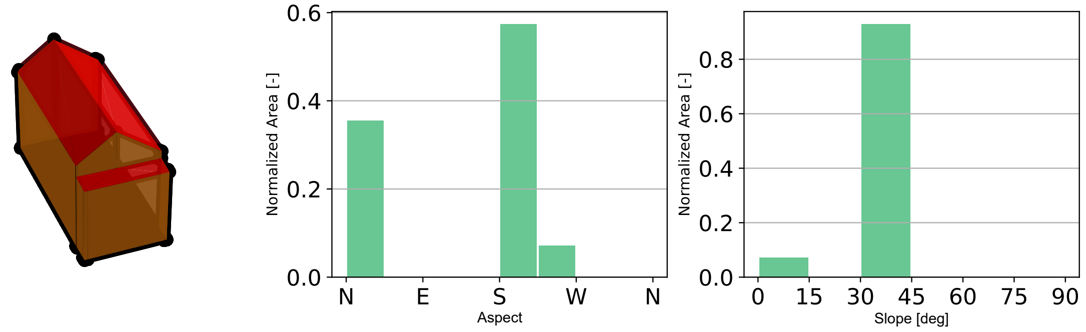


Figure 3.2: A 3D visualisation of a building registration in 3D BAG, where the roof surface is coloured red (left). The aspect distribution of the surfaces of the roof (middle) and the distribution of the inclination (right).

Roof Complexity

Roof complexity was defined as the number of non-zero bins in the aspect and slope histograms. For a roof with only one single surface, this value is 2, and for a complex roof with surfaces of all different aspects and slopes, this value is 14. The example shown in Figure 3.2 has 3 and 2 non-zero bins in the aspect and slope histograms, respectively. The value for the roof complexity, therefore, is $3+2=5$.

Flatness

Flatness was defined as the fraction of roof area with a maximum slope of 15°. This equals the value of the left-most bin in the slope histogram. For the example shown in Figure 3.2, the value is 0.071.

Area Per Surface

Area Per surface was defined as the sum of the roof areas divided by the number of surfaces. For the example shown in Figure 3.2, the value is $50+31+6.2)/3 = 29.1$ m^2.

Average Irradiance

The relative solar irradiance was computed for each surface based on aspect and slope. Figure 3.3 shows the relative irradiance for a surface in the Netherlands, where the maximum of 100% occurs for a South-facing surface with a slope of 35°. The final irradiance value is obtained by taking the weighted average of all irradiances of each surface, with the surface areas as weights. For the example shown in Figure 3.2, the value is 73.5%.

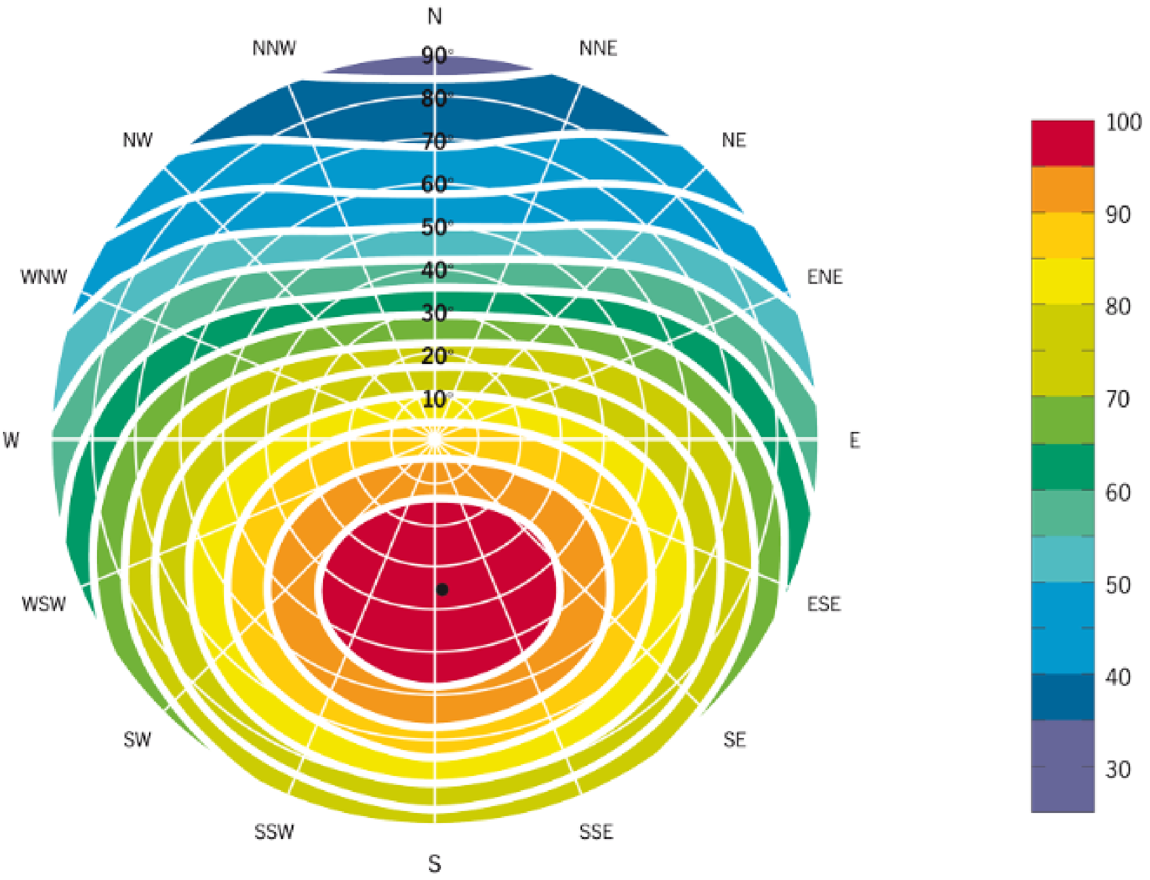


Figure 3.3: Relative solar irradiance as a function of aspect and slope (Induurzaam, n.d.).

Number of Neighbouring Buildings
The number of neighbouring buildings is another feature computed based on data in the 3D BAG dataset. Each building has a building footprint defined by a number of points. Figure 3.4 shows an example of two building outlines. In some cases, buildings are very near to each other or even touch, but this is not always exactly represented in the dataset. Therefore, a buffer of 10 cm is taken around each building outline before looking for intersecting building outlines.

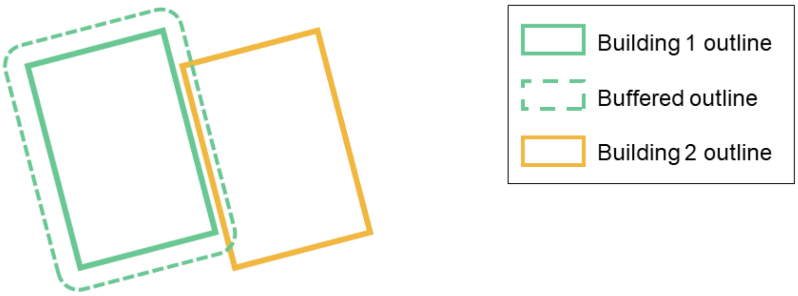


Figure 3.4: Building outlines including one buffered outline.

Building Height
In the 3D BAG dataset, the roof height is included as a feature. For each roof, the feature *h_dak_max* represents the elevation above sea level of the roof's highest point. In this study, the relative height was used, which was obtained by subtracting the elevation above sea level from the ground level.

Suitable Roof Area
The suitable area was determined using AHN and BAG datasets. This required a number of processing steps, which are further explained in Section 3.4.

3.4. Suitable Roof Area

The roof area suitable for PV was estimated by processing roof geometries. These geometries were obtained from the AHN and BAG datasets. Please note that 3D BAG also contains height data of buildings. However, these geometries are simplified, and some smaller obstacles, such as chimneys, are not included in the 3D building models. Therefore, the AHN raster data was more suited for analysing and detecting obstacles. A summary of the processing steps is as follows:

1. Get the building footprint(s) from the BAG dataset
2. Download the AHN raster data
3. Assess which pixels receive enough solar power
   (a) Calculate the aspect and slope for each pixel of the AHN raster
   (b) Determine the relative irradiance of each pixel
   (c) Determine the average yearly solar power of each pixel
   (d) Combine relative irradiance and yearly average solar power
   (e) Select all pixels above a minimum solar power
4. Assess which areas are obstacle free
   (a) Calculate the change in slope for each pixel of the AHN raster
   (b) Apply a median kernel to the obtained slope change raster
   (c) Select all pixels within the footprint area with a low change in slope
5. Combine the two criteria
   (a) Select all pixels that fulfil both criteria
   (b) Convert the selected pixels to polygons
   (c) Take a buffer around the polygons
   (d) Take a negative buffer from the building outline and intersect it with the polygons
   (e) Simplify the polygons
6. Sum the areas of the polygons to find the total suitable area

Steps 3, 4 and 5 are further elaborated upon in this section, in subsections 3.4.1, 3.4.2 and 3.4.3, respectively. Please note that not the entire method was developed for this thesis. Steps 1, 2, 3c and 3e were already created and used by Sobolt (2022a) and thus not developed by the author. Steps 3a, 3b and 3d were developed in collaboration with Sobolt but not solely by the author.

An example further illustrates the processing steps. An aerial image of the example building is shown in Figure 3.5. The AHN3 height data of the example is shown in Figure 3.6. The example roof has a South-facing side and a North-facing side, with a dormer present on the North-facing side. Also, note that the roof to the right (East) of the example building is slightly higher.

The actual method executes processing steps on large areas of AHN3 data. Only in the final processing steps are the results separated for each building. However, in this example, every image will only contain the raster for a single building since this provides more insight into the method used.

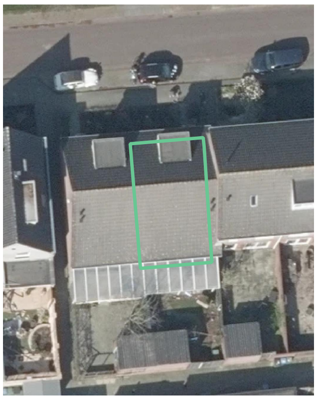


Figure 3.5: Aerial image of example building.

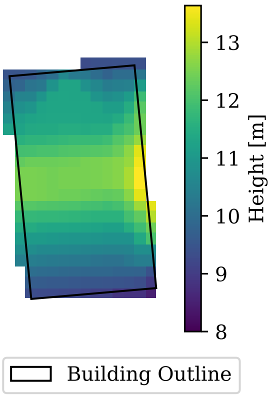


Figure 3.6: AHN3 height raster of example building.

3.4.1. Enough Solar Power

In step 3a, the aspect and slope for each pixel of the height raster were calculated. First, the gradient in the x- and y-direction needed to be computed. If h_0 is the height of a certain pixel, h_{-1} the height of a pixel to the left and h_1 the height of a pixel to the right, then an approximation of the slope in the x-direction $\frac{\partial h}{\partial x}$ can be given by

$$\frac{\partial h}{\partial x} = \frac{h_1 - h_{-1}}{2} \tag{3.1}$$

which equals the operation carried out by the kernel in Equation 3.2. Similarly, Equation 3.3 shows kernel for approximating the slope in y-direction. Such a kernel yields a gradient in m/pixel.

$$\begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \tag{3.2}$$

$$\begin{bmatrix} -\frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix} \tag{3.3}$$

Then the slope S equals

$$S = \sqrt{\left(\frac{\partial h}{\partial x}\right)^2 + \left(\frac{\partial h}{\partial y}\right)^2} \tag{3.4}$$

where the resultant slope is in the units of m/pixel and $\frac{\partial h}{\partial y}$ is the gradient in the y-direction. To convert this to units of m/m, one needs to divide by the pixel size of 0.5 m. The aspect, α, equals

$$\alpha = \arctan 2\left(\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y}\right) \tag{3.5}$$

The same coordinate system as in Section 3.3 is used, meaning that a value for α of 0 corresponds to orientation to the North, while ± 180 corresponds to the South. Figure 3.7 and 3.8 show the result of these operations on the example building. In the figure, the slope is displayed in degrees instead of m/m, as this could be considered more intuitive. In Step 3b, the aspect and slope of each pixel are used to compute the relative irradiance from Figure 3.3. The result of this operation is shown in Figure 3.9.

In step 3c, the average yearly solar power was computed. Assessing the power for an entire year is computationally intensive, so simplification is used. The irradiance was computed for each hour of the day for four days; March 21st, June 21st, September 21st and December 21st, and these were then averaged. This analysis accounts for the shadows cast by other, higher, geometries. The raster was then normalised such that the maximum value in each raster equals 1. Figure 3.10 shows the result of this analysis.

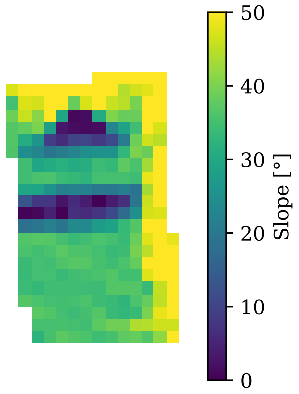


Figure 3.7: The estimated slope S of the example building (step 3a).

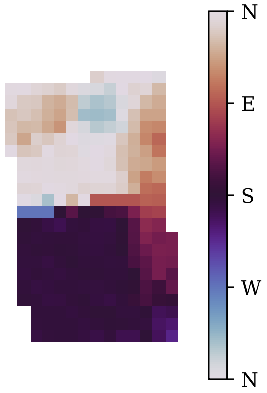


Figure 3.8: The estimated aspect of the example building (step 3a).

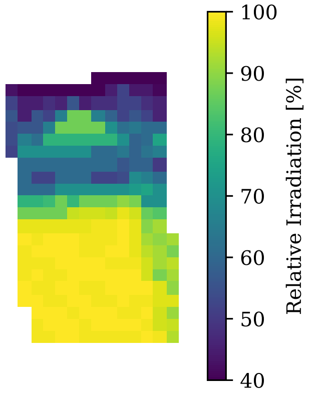


Figure 3.9: The estimated relative irradiance of the example building (step 3b).

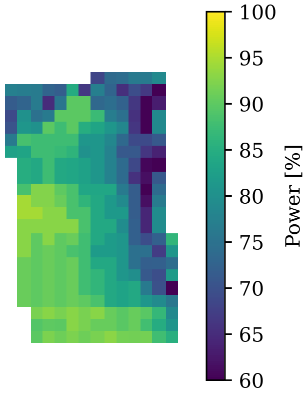
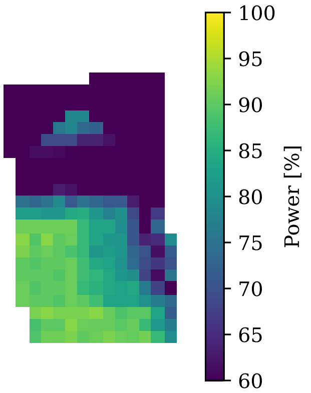


Figure 3.10: The normalised estimated received power of the example building (step 3c).

The raster shown in Figure 3.9 accounts for the roof's orientation but does not account for shadows cast by other geometries. On the contrary, Figure 3.10 only accounts for shadows. Therefore, these masks were combined for step 3d, which resulted in Figure 3.11. A threshold was set for the suitability, and any pixels that received at least 70% of the power of a flat surface were deemed suitable based on the power criterion. A flat surface has a relative irradiance of 87% (see Figure 3.3), so the threshold for the combined mask was set at $0.87 \cdot 0.7 = 0.61$, i.e. 61%. Figure 3.12 shows the result of this step.

Based on the power criterion, the South-facing part of the roof is considered suitable, except for some pixels on the left (East), which is caused by the shadow cast by the building on the left. The North-facing part does not receive enough power to be considered suitable. However, the top part of the dormer is slightly higher and oriented more towards the Sun, which causes some pixels on the North-facing side to be also deemed suitable.



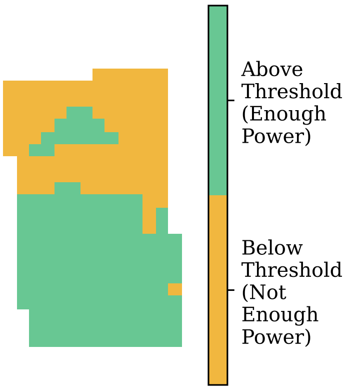


Figure 3.11: The estimated received power of the example building corrected for slope and aspect (step 3d).

Figure 3.12: Pixel suitability of the example building based on high received power (step 3e).

3.4.2. Obstacle Free

Solar panels require an area of the roof with a relatively constant slope. Therefore, this study defines the obstacle-free parts of the rooftop as areas where the change in slope is below a certain threshold.

For step 4a, the changes in horizontal (x-direction) and vertical (y-direction) slopes were computed. If h_0 again represents the height of a certain pixel, h_{-1} the height of a pixel to the left and h_1 the height of a pixel to the right, then an approximation of the slope to the right and the left of this pixel in the x-direction $\frac{\partial h}{\partial x}$ can be given by Equation 3.6 and 3.7 respectively.

$$\frac{\partial h}{\partial x} = h_0 - h_{-1} \tag{3.6}$$

$$\frac{\partial h}{\partial x} = h_1 - h_0 \tag{3.7}$$

The change in slope in the x-direction is then calculated by

$$\begin{aligned} \frac{\partial^2 h}{\partial x^2} &= (h_1 - h_0) - (h_0 - h_{-1}) \\ &= h_1 - 2h_0 + h_{-1} \end{aligned} \tag{3.8}$$

This operation is performed by applying the kernel in Equation 3.9. Likewise, the change in slope in the y-direction is computed using the kernel in Equation 3.10.

$$\begin{bmatrix} -1 & 2 & -1 \end{bmatrix} \tag{3.9}$$

$$\begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix} \tag{3.10}$$

The change in slope, h'', was then obtained by combining these values. This can be summarised as

$$S' = h'' = \sqrt{\left(\frac{\partial^2 h}{\partial x^2}\right)^2 + \left(\frac{\partial^2 h}{\partial y^2}\right)^2} \tag{3.11}$$

where $\frac{\partial^2 h}{\partial x^2}$ and $\frac{\partial^2 h}{\partial y^2}$ indicate the gradient changes in x-direction and y-direction, respectively, both in m/pixel^2. A small kernel was chosen deliberately since larger kernels tend to produce undesired artefacts around obstacles, which caused problems for smaller roofs. However, small kernels are more susceptible to noise, so step 4b applied a 3 x 3 median kernel to the grid containing h'' values to counteract this. The computed change in slope is displayed in Figure 3.13 and Figure 3.14 shows the result after applying the median kernel.

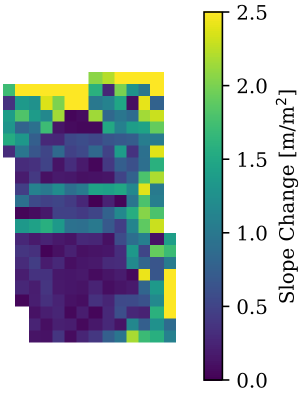


Figure 3.13: The Estimated Change in Slope of the Example Building (Step 4a)

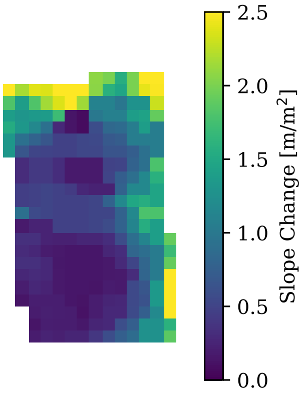


Figure 3.14: The estimated change in slope of the example building after applying a 3 x 3 median kernel (step 4b).

For the next step, 4c, pixels with a low change in slope are selected. All pixels with h'' below 0.12 m/pixel^2 were considered suitable. Since the AHN pixels measure 0.5 x 0.5 m, this equals 0.12 m / 0.5 m^2 or 0.48 m^{-1}. This threshold was found to work best after trying this method for various thresholds and comparing the resulting suitable areas to aerial images of the buildings. Figure 3.15 shows the pixels below the threshold, thus deemed suitable.

Based on the results of step 4, most of the South-facing part of the roof is deemed suitable. The part located close to the border with the Eastern neighbours is deemed unsuitable due to the height difference between the roofs. The North-facing part is mostly deemed unsuitable due to the presence of the dormer. Only the top part of the dormer and the part between the top of the dormer and the highest point of the roof are far away from obstacles to be considered suitable.

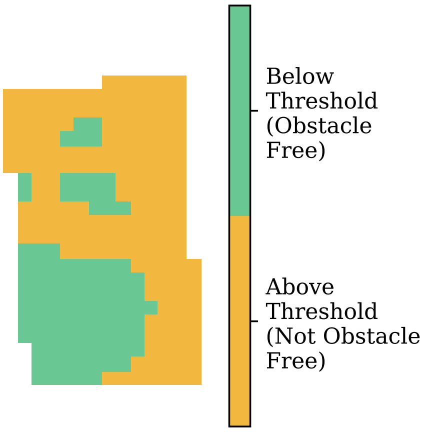
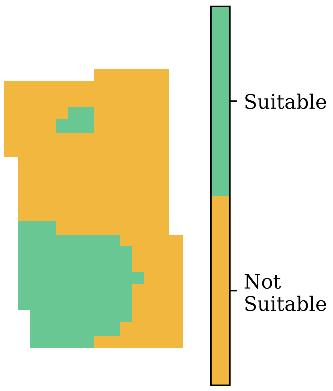


Figure 3.15: Pixels suitability based on the low change in slope of the example building (step 4c).

3.4.3. Combining the Criteria

Step 5a combined the two criteria from Figure 3.12 and 3.15, resulting in Figure 3.16. Then, the suitable pixels are grouped into polygons for step 5b. From some first results, it became clear that these polygons were often slightly smaller than the actual obstacle-free area on roofs. A possible reason for this is the effect of the slope change kernels and median kernel; a large change in slope between two pixels still leads to unsuitable pixels for the adjacent pixels. To partly compensate for this effect, step 5c takes a buffer around the suitable polygons of half a pixel's width, i.e. 0.25. Figure 3.17 shows the suitable polygons, including their buffer.



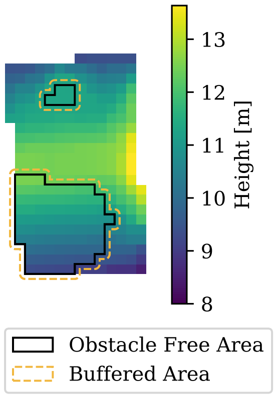
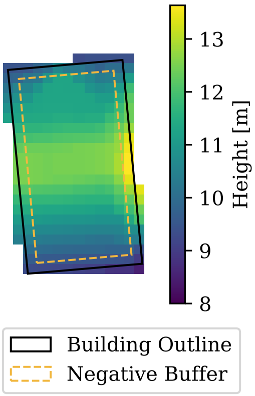


Figure 3.16: Pixels suitability based on both Criteria (step 5a).

Figure 3.17: Polygons of the area deemed suitable, including a buffer of 0.25 m (steps 5b and 5c).

PV installations are often placed slightly from the edge of the roof. Therefore, a margin of 0.5 m from the edge was taken to determine the suitable area, which is equal to taking a negative buffer of the building outline. Figure 3.18 shows this negative buffer. This area was intersected with the suitable polygons for step 5d. Polygons can require significant memory space, especially when buffers are applied. Therefore, the final step simplified the polygons with a tolerance of 0.1 cm, i.e. the simplified shapes could deviate at most 0.1 m from their original shapes. Figure 3.19 shows the final result. Two areas are identified as suitable for PV; a small area on top of the dormer and a larger area on the South-facing side. The former has an area of 2.35 m^2 and the latter of 15.76 m^2. Therefore, the suitable area of this roof is estimated to be 18.1 m^2.

Several parameters in this analysis were chosen by trial and error; the threshold values for slope change and power, the polygon buffer size, the margin from the roof edge and the simplification tolerance. These were varied, and the results of several municipalities were inspected and compared to aerial images. The example shown in this method yields results close to what one would expect intuitively. However, not all roofs yield such results. This will be further elaborated upon in Chapter 4.



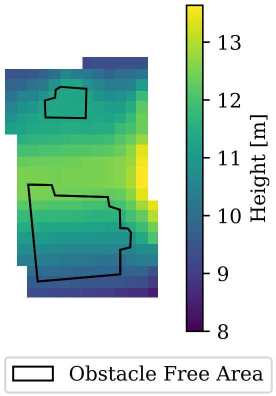


Figure 3.18: A negative buffer from the building outline of the example building.

Figure 3.19: Intersection of the suitable polygons and the negative buffer from the building outline.

3.5. Feature Summary

Table 3.5 provides an overview of the features used as independent variables. A test for multicollinearity was carried out by calculating the VIF as described in Subsection 2.1.6. The features with the highest VIF were "building usage: residence" and building height, with VIFs of 3.50 and 2.89, respectively. These are still below the maximum allowed value of 10, so none of the features had to be discarded. The full results of the VIF analysis are presented in Table A.1.

Table 3.5: A summary of the 15 independent variables used for the model.

Feature	**Type**	**Range**	**Mean**	**Median**	**Unit**
Socioeconomic					
Population 25 - 44	float	0.016 - 1	0.26	0.24	-
Year of Construction	integer	1073 - 2015	1971.2	1974	year
House Ownership	float	0 - 10	2.19	2	-
Median Income	float	10 - 90	49.9	50	%
Address Density	float	4 - 3877	1118	993	addresses/km^2
Surrounding PV Ratio	float	0 - 0.43	0.045	0.039	-
Building Usage: Residence	binary	0 - 1	0.52	1	-
Registered Addresses	integer	0 - 245	0.76	1	addresses
Geometric					
Roof Complexity	integer	2 - 14	3.82	3	-
Flatness	float	0 - 1	0.40	0.21	-
Area Per Surface	float	1.1 - 132573	42.6	21.4	m^2
Average Irradiance	float	0 - 1	0.81	0.82	-
Neighbouring Buildings	integer	0 - 69	0.91	1	buildings
Building Height	float	0.5 - 136	6.4	7.12	m
Suitable Roof Area	float	0 - 95022	27.7	79.9	m^2

3.6. Outlier Analysis

Subsection 2.1.2 explained how outliers in large, multidimensional datasets can be found using the local outlier factor (LOF). This study used the LOF implementation from Skikit Learn (Pedregosa et al., 2011) and set the model up to consider the 20 nearest neighbours for each sample.

The nine largest outliers, i.e. with the highest LOF scores, were investigated manually. Table A.3 shows the feature values of these buildings in Appendix A. None of the buildings showed signs of errors in the input data or data processing. In addition to LOF analysis, each building was manually examined to determine if it had the smallest or largest value for any feature. These two tests cannot guarantee that the rest of the samples are error-free, but it does indicate that the chances of errors in the data are limited.

3.7. Machine Learning Models

Subsection 2.1.1 explained the theoretical background of random forests and neural networks. This section explains how they were implemented in this study.

3.7.1. Random Forest

This study used the random forest regression implementation from the Python package Scikit-Learn (Pedregosa et al., 2011). Parameters deemed to have the most significant influence on model performance were varied during nested k-fold analysis and are shown in Table 3.6. The values for a maximum number of samples and features are given as a fraction. For example, a value of 0.4 for a maximum number of features means that 40% of the features are considered for each split. For a model with 15 features, this results in $0.4 \cdot 15 = 6$ features. The training criterion chosen is *squared error*, so the model is optimised on the Brier score. Other parameters were left to the default values as set by Scikit-learn developers (2022b).

Table 3.6: Parameters varied for training random forest regressors.

Parameter	Values considered
Maximum Number of Samples	0.2, 0.4, 0.6, 0.8, 1
Maximum Tree Depth	1, 5, 10, 15, 20, 50, 100
Maximum Number of Features	0.2, 0.4, 0.6, 0.8, 1
Number of Trees	1, 5, 10, 15, 20, 50, 100

3.7.2. Neural Network

The neural network implementation in this study was from the same Python package as the random forest implementation; Scikit-Learn. Parameters deemed to have the most significant influence on model performance were varied during nested k-fold analysis and are shown in Table 3.7. The activation "logistic" refers to the logistic sigmoid function, which returns $f(x) = 1/(1+e^{-x})$ and "relu" refers to the rectified linear unit function, which returns $f(x) = \max(0, x)$. The training criterion chosen is *squared error*, so the model is optimised on the Brier score. Other parameters were left to the default values as set by Scikit-learn developers (2022a). Neural Networks are sensitive to feature scaling, so before training the network, all features were scaled to have a zero mean and unit variance. Please note that the scaling factors were obtained by only considering the training dataset and not the test dataset to maintain the independence between these two.

Table 3.7: Parameters varied for training neural networks.

Parameter	Values considered
Activation	Logistic, relu
First Hidden Layer Size	5, 10, 20, 50, 100
Second Hidden Layer Size	None, 2, 5, 10

3.8. Model and Feature Evaluation

This section first describes the methods used for model evaluation. Then, the methods for feature importance are explained.

3.8.1. Model Evaluation

The model validation was executed using nested k-fold cross-validation (NeCV) with $k_i = 5$ inner loops and $k_o = 3$ outer loops. The splitting of the data into folds was not done completely randomly. Some features were known only on a PC4 level, so all buildings in that area have the same values for those features. To ensure folds that are as independent as possible, it was decided not to split samples from one PC4 area into different folds. For example, if $k = 5$ folds have to be created, the postal codes present in the dataset are randomly split up into five groups, and these groups are then used to split the data into folds. This method is used for the random forest and neural network models. Moreover, another model used only geometric features and another used only socioeconomic features.

As an additional independent means of verification, the model was tested in a new area, North Holland. This dataset was preprocessed similarly as explained in sections 3.1 up to 3.3. There is one significant difference, however; the PV data for North Holland is available for the years 2020 and 2021, as opposed to 2019 and 2021 for Overijssel. This had two main consequences. Firstly, 4.5% of the buildings adopted PV between 2020 and 2021, which is lower than the 5.7% for Overijssel between 2019 and 2021. As a result, the dataset is even more unbalanced. The second consequence is that the feature "surrounding PV ratio" will be higher since the number of PV installations has increased between 2019 and 2020.

The additional dataset was used for validation in two different ways. First, the model was trained on the Overijssel dataset using the optimal hyperparameters. This model was then used to predict PV adoption in North Holland.

Secondly, the model was retrained on North Holland data and then used for prediction on the North Holland data. To ensure a model evaluation that is as independent as possible, this was again done using NeCV.

3.8.2. Feature Evaluation

The importance of the used features was computed in three different ways; by permutation importance, statistical testing and single-feature models. The theoretic background of permutation importance was already discussed in Subsection 2.1.5; it is a method that indicates how important a certain feature is in a certain model. Its importance also depends on the other features used in the model. However, it is also important to assess the isolated feature because this indicates how well it could work in different models or different feature combinations.

Statistical Testing

Statistical testing is commonly used to quantify the difference between two groups and test whether this difference is significant. Such tests often yield a p-value. A p-value is the probability that the two groups were sampled from the same population. Most features do not have normally distributed data, so applying the commonly used z-test would probably not lead to a reliable result. Instead, the non-parametric Mann-Whitney U test was applied to the dataset (McKnight & Najab, 2010). This test is based on ranking the values of two populations, which allows computing p-values without assuming normally distributed data. However, on large datasets such as this study, statistical tests often yield extremely low p-values. They will often point out a significant difference between the two populations, even if the difference is very small. The results can, therefore, not be interpreted reliably and are not included in Chapter 4.

Single-Feature Models

As an alternative quantification method, simple single-feature models were created; a model was trained and tested using only one feature. The results of this analysis can indicate how well an isolated feature can separate buildings that adopted PV from those that have not. In that regard, it is the opposite of analysis using permutation importance, where the performance of the feature is analysed in combination with other features.

The AUC score (Area under the ROC curve, explained in Section 3.8) was computed for this model using k-fold cross-validation using 5 folds. The AUC was scaled to a chance-standardised variant given by the Gini coefficient for this method. This takes values between 0 and 1. It is calculated as

$$G = 2 \times AUC - 1 \tag{3.12}$$

After converting to the Gini coefficient, all scores were normalised to add up to 1.

3.9. Summary of Methods

The first step of predicting PV adoption consisted of generating useful features. Socioeconomic features were obtained from the CBS, BAG, and Sobolt PV datasets. BAG features were on a building level; year of construction, building usage and number of registered addresses. CBS features, being resident age, house ownership, income and address density, were only available on a postal code level. The surrounding PV ratio, extracted from the Sobolt dataset, was only available on a neighbourhood level. Any redacted CBS data was filled up using the PC4 data or the dataset median.

The geometric features were derived from 3DBAG, BAG and AHN datasets. 3D BAG data was first processed to obtain slope and aspect histograms of each building's roof area. These histograms were then used to create four features; roof complexity, flatness, area per surface and average irradiance. The number of neighbouring buildings and building height were obtained using the BAG dataset. Lastly, AHN data was used to determine the suitable roof area by analysing which parts receive enough solar power and are obstacle free. All steps are summarised in a data flow diagram in Figure 3.20.

In conclusion, this study processed five datasets to create a total of 7 socioeconomic features and eight geometric features.

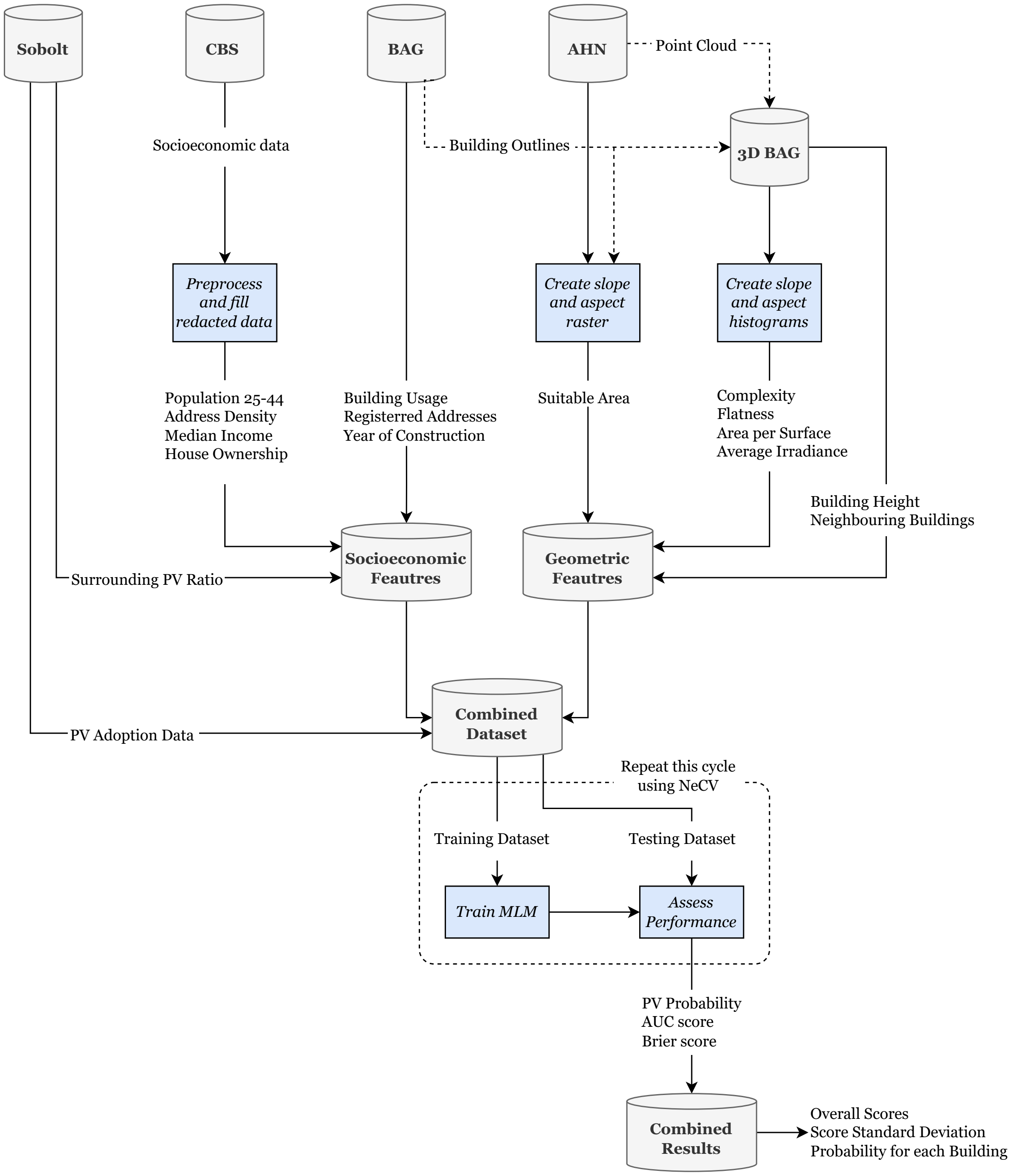


Figure 3.20: A data flow diagram showing the main steps of creating the model.

4
Results and Discussion

This chapter contains all results obtained in this study. Furthermore, the results are put into context, and their applicability is also discussed in this chapter. The results and following discussion will answer research sub-questions (SQ) 2, 3 and 4;

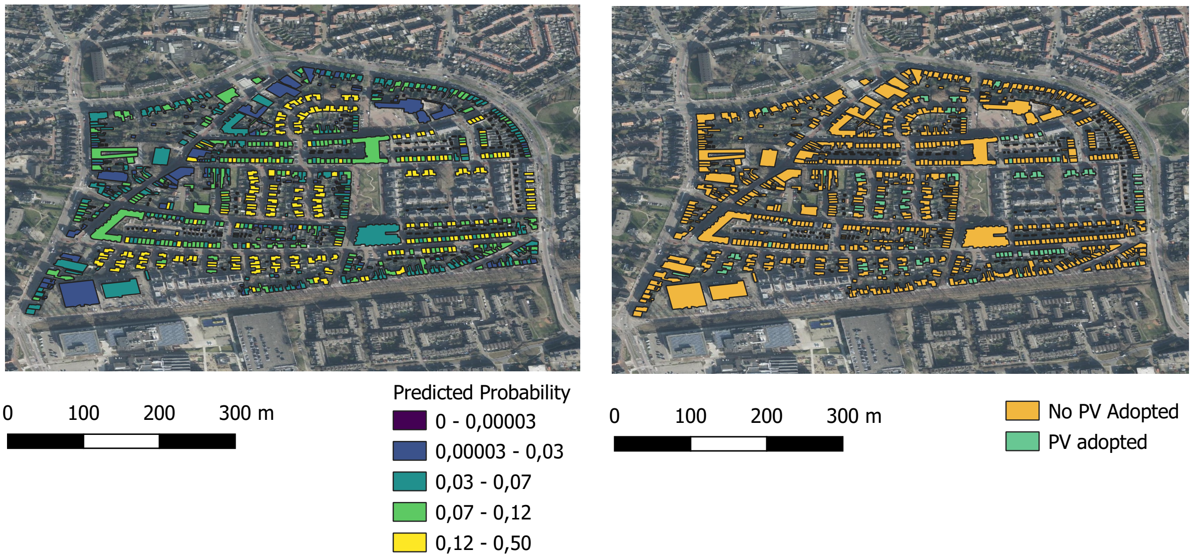
2. *How can we accurately create and assess photovoltaic system (PV) adoption prediction models?*
3. *Which geometric and socioeconomic features are important to the model when predicting PV adoption?*
4. *Which areas within the study area have a high chance of PV adoption according to the used model?*

The questions are answered by examining the machine learning models' (MLMs) outputs, i.e. the probability of PV adoption per building. First, some aggregated results are presented in Section 4.1. Then, this chapter will discuss the main model and its performance in Section 4.2 to answer SQ2. Section 4.3 discusses the importance of the individual geometric and socioeconomic features, which answers SQ3. Section 4.4 elaborates more on the model's performance in specific situations to add more context to the answer to SQ2. Additionally, the results are presented in three case studies in Section 4.5 as concrete examples of the output. This section, together with the results of the other sections, provides an answer to SQ4. Lastly, the results of the suitable area analysis are analysed separately in Section 4.6

4.1. Aggregated Results

This section shows several results visualisations, with the main goal of providing a first, qualitative, view of the results and the way they are presented. Sections 4.2 up to 4.4 will dive more into putting the results in context and discussing their quality and usability, so this section will be limited to explorative aggregated results. Figure 4.1 shows the results for a single district ("Buurt" in Dutch) in Enschede, The Netherlands. The left map shows the predicted probabilities in five colours. The colours are chosen such that each represents a quantile, i.e. 20% of the buildings in this district.

In practice, a predicted probability of 12% indicates a 12% chance that the building adopted PV during the study period (i.e. between 2019 and 2021). In other words, if eight buildings with such a probability of 12% are selected, then it can be expected that one of these buildings adopted PV.



(a) Predicted probability

(b) True PV adoption

Figure 4.1: Predicted probability on a building scale (left) and PV adoption data (right) in De Laares, a district in Enschede, The Netherlands.

The results in Figure 4.1 are on building scale, as was the aim of this thesis, but the predicted probabilities can also be aggregated per region. For example, Figure 4.2 and 4.3 show the results aggregated by district and municipalities, respectively. The colours again represent quantiles within the areas. Note that a larger aggregation region leads to less extreme values.

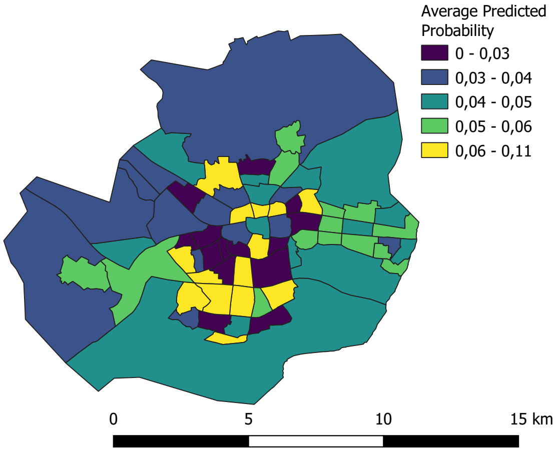


Figure 4.2: Predicted probability per building averaged per district ("buurt") in Enschede, overijssel.

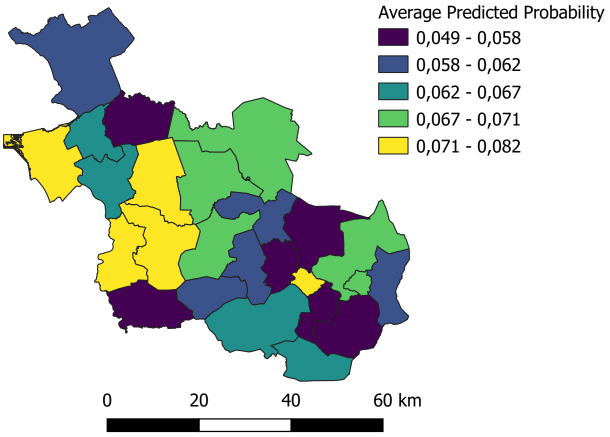
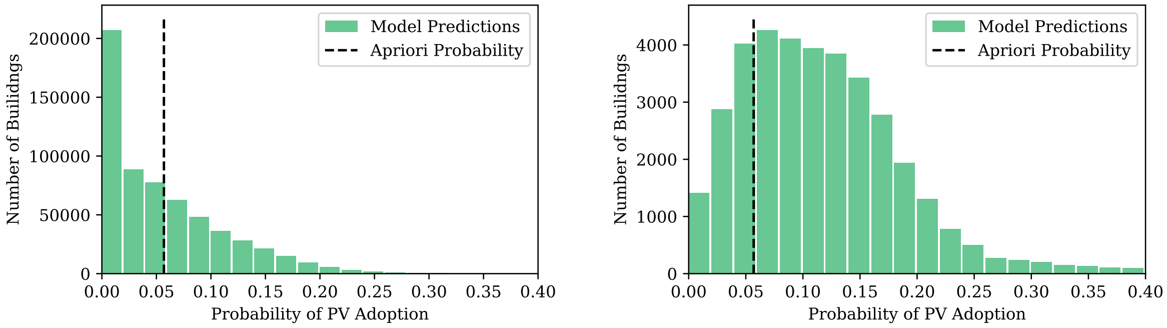


Figure 4.3: Predicted probability per building averaged per municipality in Overijssel.

All results of the model can be visualised in histograms as well. Figure 4.4a shows the predicted probabilities for all buildings that did not adopt PV. The black dashed line indicates the apriori probability of 0.057. A very simple model that predicts the dataset averaged probability would predict 0.057 as 5.7% of the buildings in the dataset adopted PV. If the model in this study predicts a probability less than 0.057, it indicates that the building has a less-than-average probability of PV adoption. Figure 4.4b shows the predicted probabilities for all buildings that adopted PV.



(a) No PV Adopted

(b) PV Adopted

Figure 4.4: Histogram of predicted probabilities for all buildings in Overijssel that did not adopt PV (left) and that adopted PV (right).

.

4.2. Models of PV Adoption

Several different models were created to predict PV adoption. One model, the base model, uses all features and is based on random forest (RF). Other models use different features, MLMs or datasets than the base model. The performance of these models is interesting to compare to the base model because it provides information on the robustness of the base model and indicates how well it could perform in different contexts.

4.2.1. Performance of the Base Model

First, the performance of the base model was assessed. The base model was evaluated using nested k-fold cross-validation (NeCV), a method explained in Subsection 3.8.1.

Results

Table 4.1 shows the results and configurations of the NeCV run for the base model. The AUC and Brier scores were 0.767 ± 0.003 and 0.0508 ± 0.0008, respectively. The error bars indicate the standard deviation of the results of the outer folds. For each fold, the configuration used results from finding the optimal hyperparameters of the other two folds.

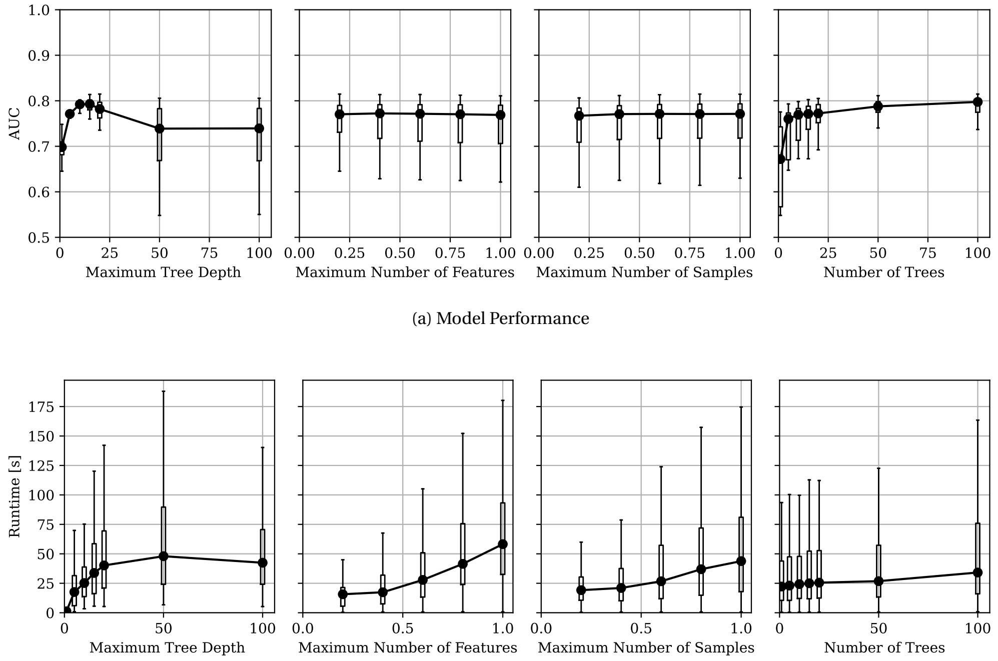
Table 4.1 shows that the model configuration is the same for each fold. This means optimal hyperparameters are not dependent on the sample that is taken from the dataset. Figure 4.5 shows the effect of varying hyperparameters visualised as boxplots for each hyperparameter configuration.

Table 4.1: Model configuration and results for each outer fold.

	Fold 1	**Fold 2**	**Fold 3**
Model Configuration			
Maximum Tree Depth	20	20	20
Maximum Number of Features	0.2	0.2	0.2
Maximum Number of Samples	1	1	1
Number of Trees	100	100	100
Results			
AUC	0.771	0.766	0.764
Brier	0.052	0.050	0.051

Discussion

An AUC score of 0.767 indicates that the model can be considered a good model, according to the interpretations of Bekkar et al. (2013). This is, however, a general interpretation, and this number should be interpreted within the given context. This AUC score also means that if the model predicts the probability of PV for a random instance with PV and one without PV, a 76.7% chance exists that the predicted probability is higher for an instance with PV.



(a) Model Performance

(b) Runtime

Figure 4.5: Box plots of model performance and runtime as a function of different hyperparameters.

The Brier score of 0.0508 means that the predicted probability has a mean squared error of 0.0508 (if no PV adopted equals 0 and PV adopted equals 1). It is interesting to compare this to the prior probability of PV within the dataset, which is 5.7%. This implies that predicting a probability of 0 for each building would yield a Brier score of 0.057. This shows that the model is only marginally better than a naive model (based on Brier score) and also points out why using AUC as a metric is preferred in imbalanced datasets. The average predicted probability for buildings that have adopted PV is 11.8%, compared to 5.6% for buildings that have not adopted PV. This average of 11.8% shows that the actual values of the probabilities are still quite low. This can be expected; for an imbalanced dataset, it is rewarding for an MLM to predict low probabilities as most validation scores are 0.

Discussion on Comparison to Related Work

Other dataset-based studies use a variety of metrics to assess model performance, of which R^2 is the most common. The PV prediction model from Zhang et al. (2011) has an R^2 of 0.52, and Van der Kam et al. (2018) developed a model with an R^2 of 0.45. However, these models did not predict PV adoption on a building level but predicted the PV ratio per region. To compute an R^2 value for this study for comparison, the results were averaged for each PC4 region. Averaging results over multiple buildings yields a prediction of the PV ratio. Comparing these results to the PV ratios per PC4 ratio yielded an R^2 of 0.14 ± 0.11, which is substantially lower than the results from Zhang et al. (2011) and Van der Kam et al. (2018). A few reasons can be thought of to explain this difference.

Firstly, the model in this study was not designed to predict PV ratios over regions but to predict probabilities on a building level. Training an MLM on a different criterion than it is tested on can lead to lower performance than when it was tested on the training criterion. In this case, the approach should have been different if the goal was to predict PV ratios per region. In that case, first, the average of all features should be taken per region, and then the model should be trained on these average features and regional PV ratios. To test this hypothesis, a simple linear regression model was created that predicts the PV ratio for a PC4 region based on the average feature values for each region. This regression results in an R^2 of 0.45, the same as the model from Van der Kam et al. (2018). This indicates that the performance difference between the RF model and other studies can, at least partly, be attributed to whether models are aggregated over regions or predict probabilities per building.

Comparing related work on a regional scale and this study on a building scale illustrates that a trade-off exists between accuracy and performance. This building scale model provides much detail and can predict the probability of PV better for a single building than a regional model because the inputs are of higher detail as well. On the contrary, a regional model has smaller errors when training and evaluation are done on a regional scale. One can argue that policymakers could prefer higher accuracy over higher spatial detail in most applications. For example, if the municipality wants to target an area for a PV stimulation campaign, they do not require the spatial detail of the model proposed in this study. However, some cases can be thought of when spatial detail has added value, e.g. when a municipality has limited resources and can only target a group of 500 households. In that case, a model on a building scale could prove to be useful. Nevertheless, most municipalities target households only on a neighbourhood level, decreasing the need for more spatial detail.

Secondly, next to region aggregation, an explanation for the difference in model performance can be a different setup of the studies. For example, Zhang et al. (2011) used regional policy subsidies as an input feature, and Van der Kam et al. (2018) used the number of GroenLinks voters as an input feature. Both were found to have a significant influence, and the latter was even found to have the strongest correlation with PV adoption. Therefore, not including these features in this study can have decreased model performance. Furthermore, (Vasseur & Kemp, 2014) showed the influence of potential energy savings. Energy usage data is available in the CBS dataset on a PC6 level. A preliminary examination for this thesis showed no clear correlation between energy usage and PV adoption, but it could be interesting to include it as a feature in future models.

Thirdly, the method of aggregation is also important. In this analysis, all buildings were averaged with equal weights. However, one can argue that not all predictions have equal certainties. For an improved aggregation, it is recommended to estimate the certainties of the predictions and use these for a weighted average per area.

One other study discussed in Chapter 1, by Mendieta and Sarker (2018), used AUC as a performance metric. They grouped buildings by area and PV adoption and used gradient boosting, random forest and logistic regression for predicting probabilities. The latter performed significantly worse, with an AUC score of 0.45. The random forest model achieved an AUC score of 0.76, similar to this thesis. This performance indicates that the performance of the model is on a comparable level to similar studies and that a large increase in performance for such a model is unlikely. The gradient boosting from Mendieta and Sarker (2018) yielded an AUC score of 0.79, which suggests that some gain in performance can be obtained by using gradient-boosted trees instead of a random forest.

Discussion on Model Configuration

Figure 4.5a shows that model performance depended on the maximum tree depth and the number of trees. Varying the number of features and the number of samples did not significantly impact the model's predictive performance.

Increasing the number of trees led to an improved performance, which can also be explained by the theory in Subsection 2.1.1; more trees counteract overfitting, provided that they are trained such that they are not similar. Increasing the number of trees above 100 could have increased the model performance even further. However, Probst and Boulesteix (2017) find that the biggest gain is generally achieved when training the first 100 trees.

Figure 4.5b also shows that a higher number of trees leads to longer runtimes. One would expect the runtime to be linearly dependent on the number of trees; training ten trees intuitively takes twice as long as training five trees. However, training multiple trees can be done in parallel on multiple central processing units (CPUs), which is why the number of trees does not severely impact the runtime for lower numbers. This study was carried out on a machine with 96 CPUs, so the effect of training more trees will probably be more present when training more than 100 trees.

Subsection 2.1.1 explained that increasing the tree depth can more easily represent complex data structures, but a too-large tree depth can lead to overfitting. This effect is present in the model, as illustrated by the peak performance at a tree depth of approximately 20 trees. Nevertheless, a tree depth of 20 could still mean that the model is overfitted. For each increment in tree depth, the number of leaf nodes doubles, so a depth of 20 means that each tree can have up to 2^{20} = 1 048 576 leaf nodes. An imbalanced dataset (5.7% minority samples, i.e. buildings that adopted PV) and NeCV with three outer folds means that each tree only had 24490 minority samples for training. This suggests that most branches are expanded until the final leaves

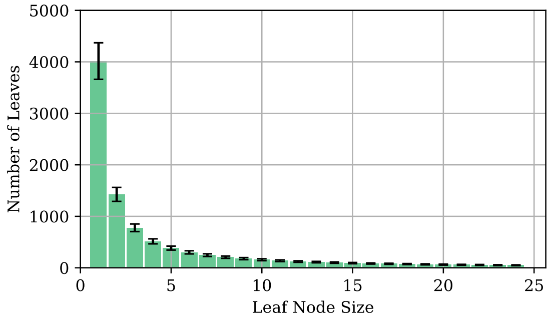


Figure 4.6: Histogram of the leaf node sizes in the RF. The error bars indicate the standard deviation between the 300 trees in the model.

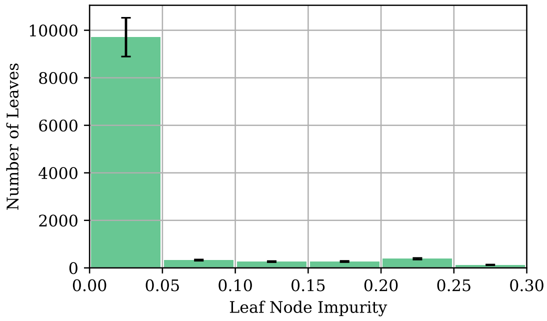


Figure 4.7: Histogram of the leaf impurities in the RF. The error bars indicate the standard deviation between the 300 trees in the model.

are completely pure, i.e. they exist of only samples that adopted PV or only samples that did not adopt PV. This was investigated by examining the leaf nodes.

Figure 4.6 shows the size of the leaf nodes in the models. Note that a total of 300 trees were trained, 100 trees for 3 folds. The most common leaf node size is 1, indicating that the trees are often expanded until the leaves are pure. This is also confirmed by Figure 4.7; approximately 85% of all leaf nodes had an impurity of 0, meaning they were pure leaves. Small leaf nodes with impurities of 0 suggest that the individual trees are overfitted.

Nonetheless, this does not necessarily impact the performance of the forest. Using bootstrapping and not considering all features ensures differing and less correlated trees, which can still yield stability when aggregating (Probst, Wright, & Boulesteix, 2019). Typical values for minimum leaf node size are 1 for classification and 5 for regression (Pedregosa et al., 2011; Probst et al., 2019). Although the algorithm was set up as a regression algorithm to predict probabilities, one could argue that this study is equal to a binary classification problem, and leaf node sizes of 1 should, therefore, not necessarily pose a problem. Nevertheless, it is recommended to further optimise the stopping criterion for tree growing. This study only considered the tree depth, but one can also look at e.g. parent node size, leaf node size or minimum decrease in impurity/MSE. Little, Rosenberg, and Arsham (2022) even show that more complicated stopping criteria could potentially increase performance for certain datasets. Probst et al. (2019) also point out that computation time decreases approximately exponentially with increasing node size and that this often does not lead to a significant decrease in performance. However, the NeCV approach prevents one from simply changing the hyperparameter values and implies that the chosen hyperparameters were optimal for the dataset. If the tree depth is decreased manually and NeCV is omitted, then overfitting the hyperparameters poses a risk; the performance on the test dataset could increase, but this would not guarantee better performance on new data.

In conclusion, both the number of trees and leaf node size could be tuned better as part of an improved model. For the number of trees, one could increase the number of considered values (in Table 3.6) to evaluate RF models with up to 1000 or more trees. For the leaf node size, the hyperparameter needs to be reconsidered since optimising on tree depth leads to overfitted trees while also increasing runtime severely.

Discussion on Imbalanced Dataset

Although random forests are shown to provide superior results on imbalanced data when compared with other learners (Khoshgoftaar et al., 2007), it is recommended to further examine the effect of the imbalanced data on the model used in this study. Imbalanced datasets are a common problem, so multiple methods exist for counteracting this (Kanellopoulos, Kotsiantis, & Pintelas, 2006). One could, for example, consider methods such as synthetic minority oversampling (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to balance the input data. In addition, one could create a costs function such that the costs for misclassifying a positive sample are higher (Pedro Domingos, 1999).

4.2.2. Performance of Different Models

Chapter 1 explained that the research gap lies in the current lack of models that combine socioeconomic features with geometric features. To investigate the added value, the base model was compared to models that use only the socioeconomic features or only the geometric features. Moreover, a neural network was used to create another model.

Results

Figure 4.8 shows the resulting AUC scores of each of these models. The base model has the highest average score measured over the three folds compared to the alternative models. The socioeconomic, geometric and neural network (NN) models have AUC scores of 0.747, 0.743 and 0.766, respectively. Brier scores have also been computed for each model, and are included in Appendix A in Figure A.1. Analysis of model performance as a function of NN model hyperparameters has also been carried out. The NN model performance was only slightly dependent on the layer sizes, especially above ten nodes in the first layer. The full analysis, including runtime, has been included in the Appendix A in Figure A.3.

Discussion

It is a logical result that the base model has the highest score compared to the socioeconomic and geometric models. It has more input features and can thus use more information to compute probabilities. This effect is discussed more elaborately in Section 4.3. However, the two alternative RF models are still relatively close to the base model in terms of performance. This indicates that combining socioeconomic and geometric models only yields a slight gain in performance.

By comparing the model performances, an indication can be given whether the proposed methods work for RF only or can be applied in a broader context. A significant difference between RF and NN is that RF is based on splits and therefore looks at the relative feature values of buildings. NN models, however, use the feature values and are, therefore, sensitive to feature scaling. For example, if a building year of construction of 500 is present in the data, this could potentially lead to (unrealistic) probabilities below 0 in an NN model. An RF model will predict the same value for a building from 1073, since 1073 is the lowest possible feature value and will thus end up at the same node. This is one of the potential reasons why certain model setups and features might work for RF models but not for NN models. Nonetheless, this analysis has shown that the performance of the used model setup is not only valid for an RF model.

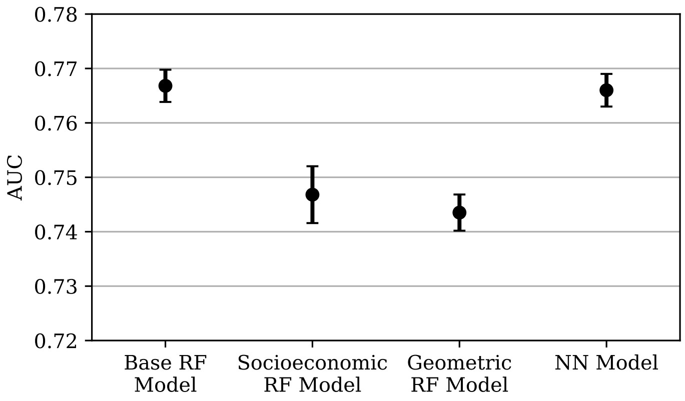


Figure 4.8: Nested k-fold AUC score of the full model compared to the socioeconomic model, the geometric model and the NN model. The error bars indicate the standard deviation between the three outer folds.

4.2.3. Performance on Another Area

The base model was also applied to a different area, namely North Holland. This was done in two different ways. First, the model was trained again on the Overijssel data and then used for prediction on North Holland. Secondly, the model was trained on North Holland and then used for prediction on North Holland while applying NeCV to ensure a proper split between train and test data.

Results

Figure 4.9 shows the performance of the models in North Holland. The figure points out that the model's performance is marginally lower in North Holland compared to the base model used in Overijssel. The model's performance that was only used for prediction on North Holland does not show an error margin. In other cases, the error margin was computed by comparing results from different folds. In this case, the training and testing data are already split, so no splitting into folds was needed, and no error margin was computed.

Discussion

It is to be expected that the model performs better in Overijssel. The base model is trained on Overijssel; it learns to recognise the characteristics of PV-adopting buildings in Overijssel. A different area, such as North Holland, might have different characteristics of PV adopters, which decreases model performance.

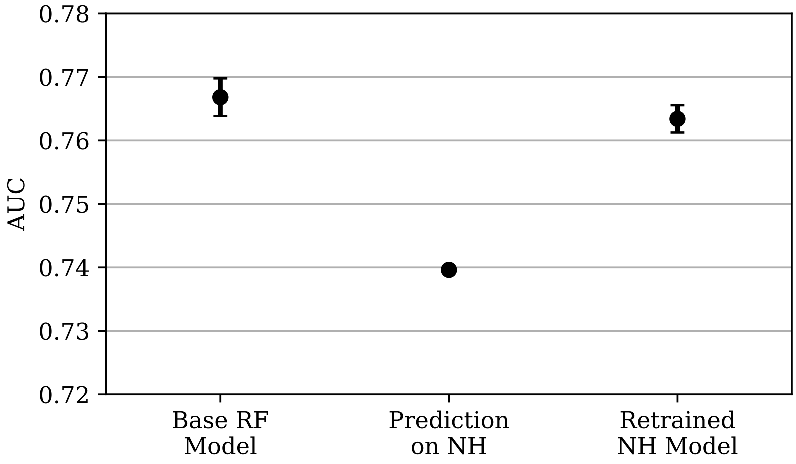


Figure 4.9: A comparison between the performance of the base RF model, prediction on North Holland and a retrained model on North Holland. The error bars indicate the standard deviation between the three outer folds.

Retraining the model on North Holland allowed the model to learn the characteristics of the new area, which increased the model's performance. Nevertheless, the performance is still slightly lower than the base model. A possible reason might be the different type of area. For example, North Holland contains more urban areas, and the urban areas are larger as well. The largest city in Overijssel, Zwolle, has 130 668 inhabitants and the largest city of North Holland, Amsterdam, has 882 633 inhabitants (CBS, 2022a). Section 4.4 will later show that, indeed, the model's performance depends on area characteristics.

Moreover, the study period is different for North Holland, as Subsection 3.8.1 explained. This led to different surrounding PV ratio values and made the dataset more imbalanced. One should also note that different versions of AHN were used for the two provinces; Subsection 2.2.3 explained that the Overijssel dataset used AHN3 and North Holland AHN4. As a result, the height raster data for computing the suitable area is more up-to-date for North Holland. Therefore, one would expect the model's performance to be higher in North Holland. Nevertheless, this effect could have been partly cancelled by the hypothesised decrease in performance due to a different study area and period.

The AUC scores were only 0.005 and 0.03 lower for the retrained and not-retrained models, despite the different area characteristics and study periods. This is a promising sign that the model's performance is not limited to Overijssel but could also apply to the rest of the Netherlands. Even when no PV adoption data is available for a new region, the model trained on Overijssel has the potential to achieve decent performance. Furthermore, it takes less than 30 seconds to train the model on Overijssel and predict probabilities for North Holland in its current configuration, and this increases the usability of the model.

4.2.4. Conclusions on Models of PV Adoption

Several MLMs have been created to predict PV adoption. The first second sub-question, *How can we accurately create and assess PV adoption prediction models?*, can already be partially answered by the results from this section. First, it can be concluded that predicting PV adoption is possible using an RF model and that this model reaches a good overall performance (i.e. an AUC between 0.7 and 0.8 (Bekkar et al., 2013)) on a building level. The Brier score showed only marginal differences from a naive model. Although the model setup was not created for prediction on a regional level, it showed similar performance to other studies. One should note that a trade-off exists between spatial detail and model performance; the model proposed in this study could provide added value if high spatial detail is required, but models trained on regional scale data provide lower errors.

Furthermore, this section showed that the model's performance is not limited to one MLM. Combining socioeconomic and geometric features increased model performance, but only slightly. Lastly, applying the model to a different area (i.e. North Holland) yielded similar results, which indicates that it applies to areas other than the study area. Improvements in model training strategies could focus on creating more, shallower trees and balancing the input data.

4.3. Feature Importances

This section presents the importance of each feature in the used model. First, it assesses the predictive qualities of the features when used individually. Subsection 4.3.1 shows the frequency distribution of PV for different features. Furthermore, it also points out the single-feature model scores and the correlation between features. Then, this section presents the importance of the features when used in the model, in Subsection 4.3.2. These analyses are combined and compared in Subsection 4.3.3.

4.3.1. PV Adoption Frequency

The PV adoption difference per feature value is an important indicator of the usefulness of a feature. This subsection presents the PV adoption frequency and discusses the correlation between input features.

Results

Figure 4.11 and Figure 4.12 show the change in PV across the independent variables, i.e. which buildings have changed from no PV in 2019 to PV in 2021. The figures are normalised histograms, such that the total area per category equals 1, resembling a discretised probability density function (PDF). If the PDFs of the dependent variable are more distinct, then the accuracy of MLMs is likely to increase. In other words, if the distribution of "PV adopted" and "No PV adopted" show very similar frequency distributions, it is harder for the model to use the feature for probability predictions.

In Subsection 3.8.1, it was explained that quantifying the difference in distribution is hard to do using statistical testing. For completeness, the results of the Mann-Whitney U test have been included in Table A.2. The alternative method computed scaled AUC values for single-feature models. Figure 4.10 shows these results. The scaled AUC scores in Figure 4.10 can be regarded as a quantification of the difference in the distributions shown in Figure 4.11 and 4.12.

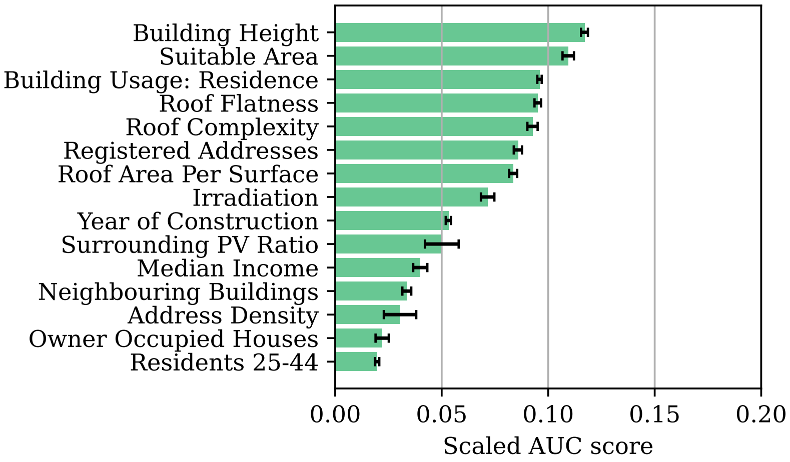


Figure 4.10: Single feature model scores. The error bars indicate the standard deviation between the five CV folds.

Discussion

The two features scoring the highest are building height and suitable area. Looking at the distributions of these two features in Figure 4.11 clearly shows that there indeed is a substantial difference in distribution for buildings that have PV adopted and buildings that have not. Buildings with suitable areas below 20 m^2 rarely adopted PV, and the same can be said for buildings less than 5 m high. This clear distinction causes high scores for the single-feature model score, and these features are likely to be of higher importance when used in predictive models.

It is important to note that these metrics should be interpreted with care. They only indicate how well the feature can separate the groups that have and have not adopted PV. However, this does not take into account any multivariate effects. Some features, such as address density, could contain important information that increases model performance only when combined with other features. These effects are better quantified using permutation importance, of which the results are presented in Subsection 4.3.2.

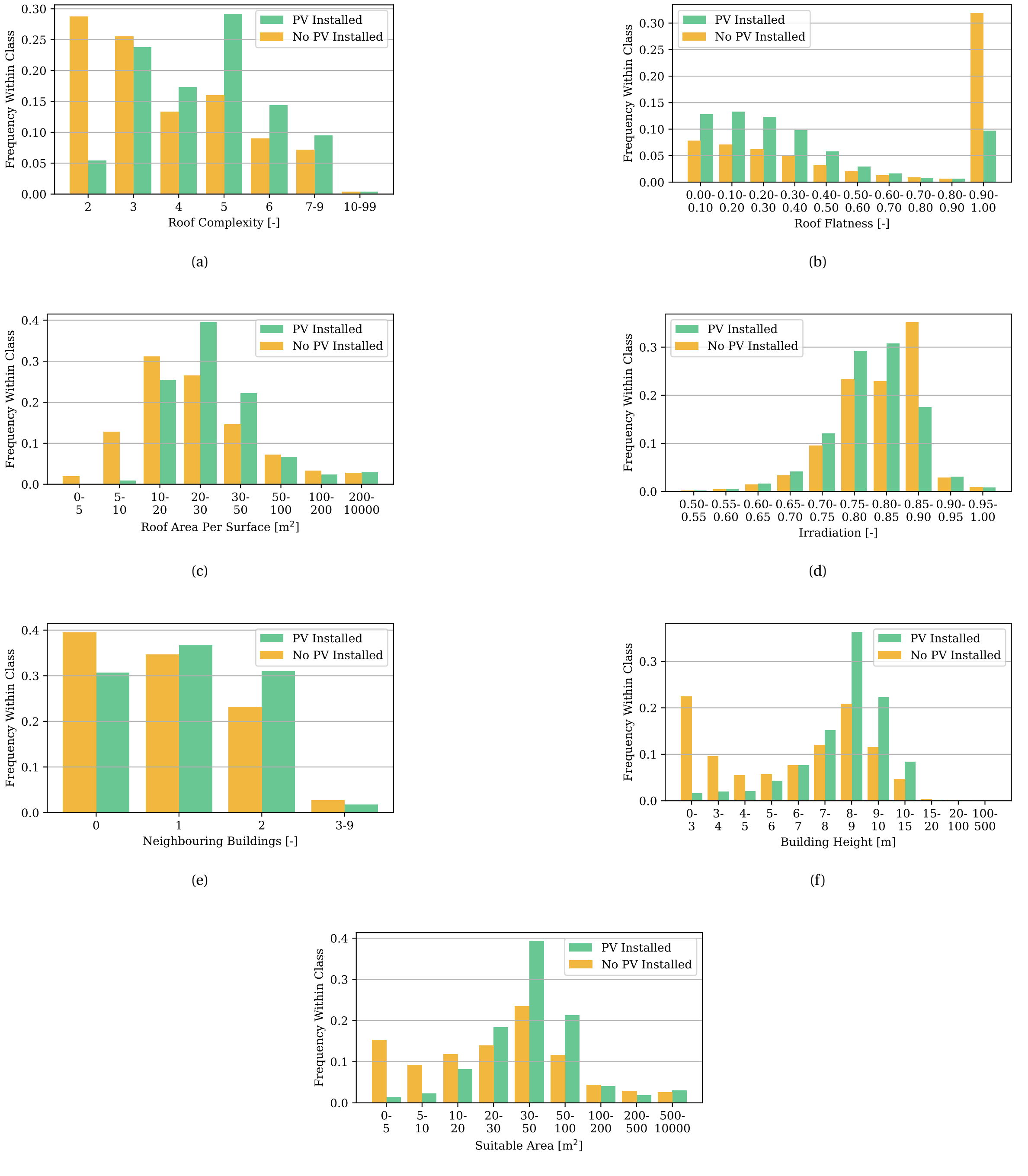
Another factor to take into account is the correlation between features. Table 4.2 shows the correlation between all input features and the dependent variable. The strongest correlations between the input features occur for the feature "building usage: residence". This gives insight into the typical characteristics of residence buildings compared to other buildings and matches what can be intuitively thought of; buildings used for residence are higher and have more complex, less flat roofs than other buildings.

Generated geometric features also show some expected correlation; buildings with a large suitable area also have a larger area per roof surface, and non-complex or flat roofs have higher values for irradiation. The bottom row shows the correlations between input features and "PV adopted, " indicating how useful a feature can predict PV adoption. Despite these correlations, all variance inflation factors (as explained in Subsection 2.1.6 are below 10. Therefore, feature importances computed by permutation can be assumed not to be influenced significantly by multicollinearity.

Nevertheless, these correlations can lead to counterintuitive results. For example, one could expect that complex roofs, with dormers and chimneys, are less suitable for PV installation. However, Figure 4.11a proves the opposite: buildings that adopted PV often have more complex roofs. A likely explanation is that building usage is a confounding factor for the relationship between roof complexity and PV adoption. Residential buildings are probably more likely to adopt PV (as proven by Lee and Hong (2019)) and residential buildings are more likely to have complex roofs. As a result, a positive correlation exists between roof complexity and PV adoption. Assuming this hypothesis is true, this analogy could also apply to other features. The distributions from Figure 4.11 and Figure 4.12 show that the distributions of "PV Adopted" and "No PV Adopted" differ more for building usage than for most other features.

As a result, all features that indicate that a building is used for residence might show a correlation with PV adoption, which is the result of the confounding factor of building usage. For example, one would not expect that having a higher or non-flat roof would be a direct reason for PV adoption, but the distributions suggest this. Creating separate models for residential and non-residential buildings would solve this problem. It is, however, not unlikely that other confounding factors also exist. For example, the percentage of owner-occupied houses and median income are correlated, and both show a positive correlation with PV adoption. From this dataset, it is hard to tell if one is the only cause of the PV adoption and, thus, a confounding factor for the other.

Nevertheless, this study is limited to investigating the correlations. Therefore, all conclusions in this report should not be interpreted as the influence of features on PV adoption but rather as a correlation with PV adoption. To obtain more reliable data on the influence of features on PV adoption, the research setup should be changed, for example, by including a temporal component or using input features that are proven to be completely independent.



(a)

(b)

(c)

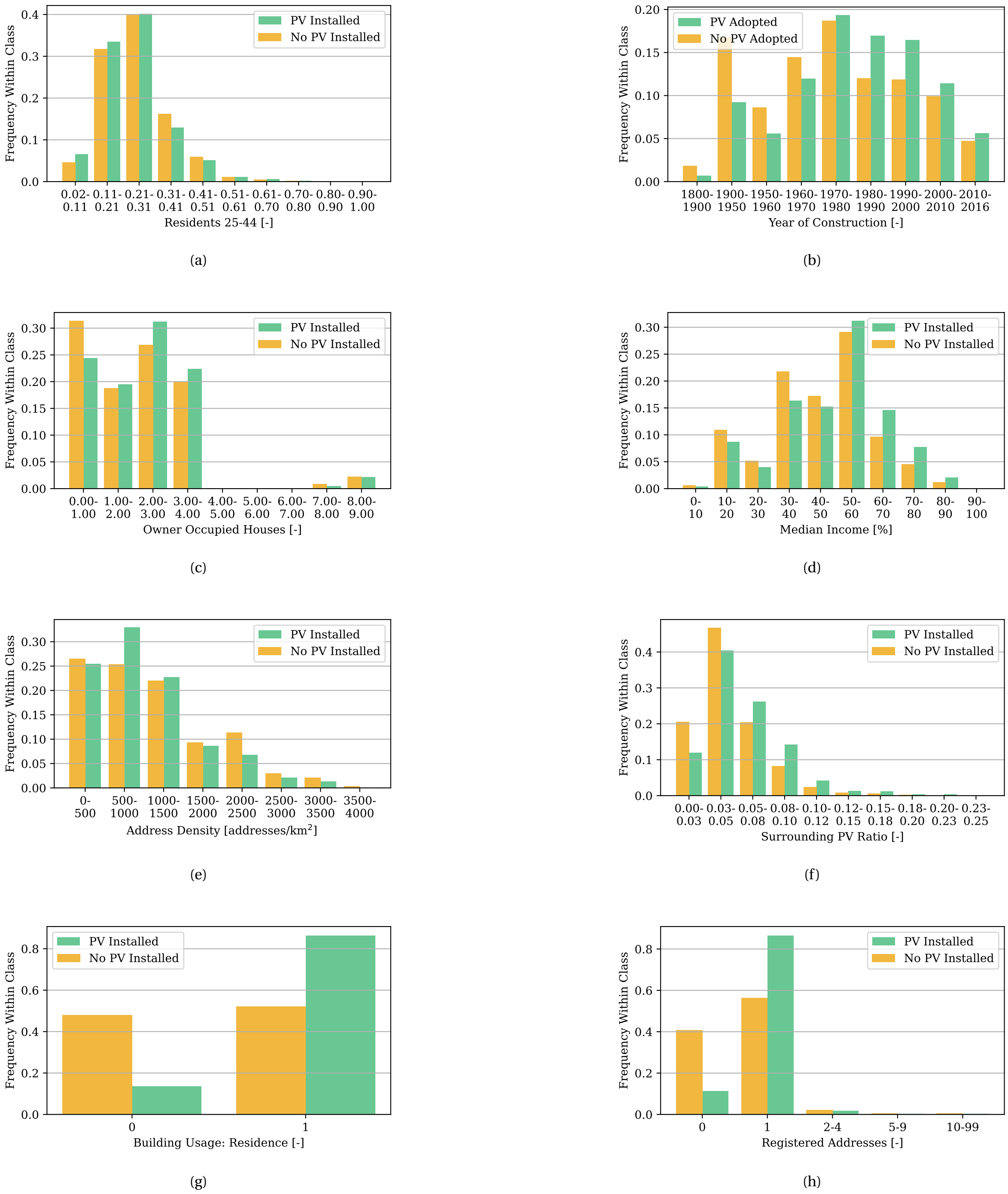
(d)

(e)

(f)

(g)

Figure 4.11: Frequency distribution of the adoption of PV for different geometric features.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 4.12: Frequency distribution of the adoption of PV for different socioeconomic features.

Table 4.2: Correlation coefficients for all features. Red indicates a negative correlation and blue indicates a positive correlation.

	Residents 25-44	Address Density	Median Income	Owner Occupied Houses	Building Usage: Residence	Registered Addresses	Surrounding PV Ratio	Year of Construction	Building Height	Neighbouring Buildings	Suitable Area	Roof Complexity	Roof Flatness	Roof Area Per Surface	Irradiation	PV Adopted
Residents 25-44	1.00	0.27	-0.18	-0.13	0.03	0.04	0.04	0.06	0.02	0.20	-0.03	-0.07	0.13	-0.02	0.07	-0.02
Address Density	0.27	1.00	-0.26	-0.49	0.07	0.07	-0.04	-0.13	0.00	0.34	-0.05	-0.07	0.25	-0.03	0.14	-0.03
Median Income	-0.18	-0.26	1.00	0.33	0.03	-0.05	0.16	0.12	0.09	-0.22	0.02	0.17	-0.11	0.02	-0.09	0.06
Owner Occupied Houses	-0.13	-0.49	0.33	1.00	-0.04	-0.04	0.04	0.08	0.03	-0.24	0.04	0.08	-0.14	0.02	-0.09	0.02
Building Usage: Residence	0.03	0.07	0.03	-0.04	1.00	0.22	0.07	-0.08	0.76	0.31	-0.06	0.56	-0.42	-0.04	-0.31	0.16
Registered Addresses	0.04	0.07	-0.05	-0.04	0.22	1.00	0.00	0.01	0.34	0.05	0.14	0.19	0.00	0.02	-0.01	0.02
Surrounding PV Ratio	0.04	-0.04	0.16	0.04	0.07	0.00	1.00	0.22	0.04	0.05	-0.01	0.00	0.07	-0.01	0.03	0.08
Year of Construction	0.06	-0.13	0.12	0.08	-0.08	0.01	0.22	1.00	-0.01	-0.01	0.04	-0.15	0.10	0.02	0.08	0.07
Building Height	0.02	0.00	0.09	0.03	0.76	0.34	0.04	-0.01	1.00	0.20	0.16	0.62	-0.49	0.03	-0.35	0.15
Neighbouring Buildings	0.20	0.34	-0.22	-0.24	0.31	0.05	0.05	-0.01	0.20	1.00	-0.07	0.00	0.08	-0.04	0.06	0.04
Suitable Area	-0.03	-0.05	0.02	0.04	-0.06	0.14	-0.01	0.04	0.16	-0.07	1.00	0.11	0.04	0.35	0.05	0.01
Roof Complexity	-0.07	-0.07	0.17	0.08	0.56	0.19	0.00	-0.15	0.62	0.00	0.11	1.00	-0.43	-0.02	-0.34	0.11
Roof Flatness	0.13	0.25	-0.11	-0.14	-0.42	0.00	0.07	0.10	-0.49	0.08	0.04	-0.43	1.00	0.02	0.58	-0.09
Roof Area Per Surface	-0.02	-0.03	0.02	0.02	-0.04	0.02	-0.01	0.02	0.03	-0.04	0.35	-0.02	0.02	1.00	0.03	0.00
Irradiation	0.07	0.14	-0.09	-0.09	-0.31	-0.01	0.03	0.08	-0.35	0.06	0.05	-0.34	0.58	0.03	1.00	-0.05
PV Adopted	-0.02	-0.03	0.06	0.02	0.16	0.02	0.08	0.07	0.15	0.04	0.01	0.11	-0.09	0.00	-0.05	1.00

4.3.2. Permutation Importance

The permutation importance of features was obtained by measuring the decrease in model performance after shuffling the values of a feature, as was explained in Subsection 2.1.5. 5 folds were used in CV.

Results

Figure 4.13 shows the feature importances based on the permutation importance method; the decrease in model AUC score is measured after shuffling the values of a feature. Five folds were used in the NeCV, and the error bars represent each fold's standard deviation of the feature importance. A similar plot can be created for the increase in the Brier score and has been included in Figure A.2 in Appendix A.

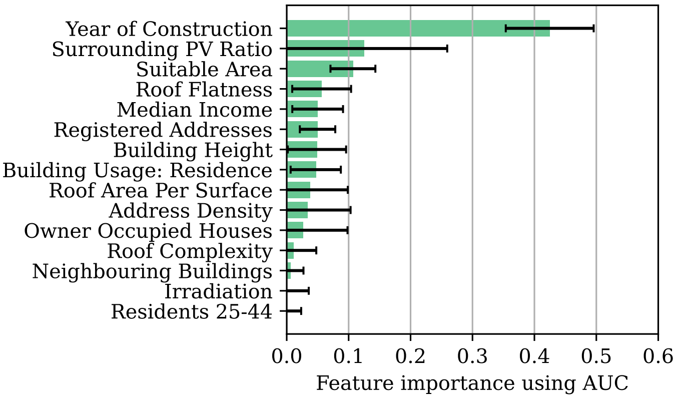


Figure 4.13: Normalised feature importance using permutation importance, based on a decrease in model AUC score. The error bars indicate the standard deviation between the five CV folds.

Discussion

Figure 4.13 points out that year of construction, suitable area and surrounding PV ratio are the most important features according to the permutation importance method. These results differ greatly from Figure 4.10, which will be further discussed in Subsection 4.3.3. To examine the effect of leaving features out, the model performance as a function of the number of features was computed, as shown in Figure 4.14. This was computed by taking 210 random feature samples, such that each number of features from 1 to 14 was assessed 15 times. The error bars represent the standard deviation in model performance between the different assessments. This analysis shows that adding features increases model performance, especially up to 9 features. Adding more than ten features only slightly increases model performance. This is a possible explanation for the small difference in performance between the socioeconomic, geometric and full models, which was displayed in Figure 4.8. The results from Figure 4.14 imply that using a simpler model with fewer features might be able to match or nearly match the performance.

Furthermore, analysing the performance as a function of the number of features also points out the limitation of the method of permutation importance. Taking one feature out barely decreases model performance, so quantifying feature importance based on the decrease in model performance might yield unreliable results. This is probably the reason for the high standard deviation for the feature importances displayed in Figure 4.13, especially for the features with lower importance. For those features, the error in importance often exceeds the importance itself, making the analysis less reliable. Nevertheless, the most important features showed a constant decrease in model performance across each fold, so their relative importance can still be considered valid.

4.3.3. Combined Feature Importance Metrics

This section has discussed two different feature importance metrics, and this subsection combines and compares the different methods.

Results

Table 4.3 shows the results from the single feature models and permutation importance, using AUC and Brier as metrics. The features are ranked from 1 to 15, where 1 represents the most important feature and 15 the least important feature. The first conclusion that can be drawn from Table 4.3 is that the AUC and Brier scores

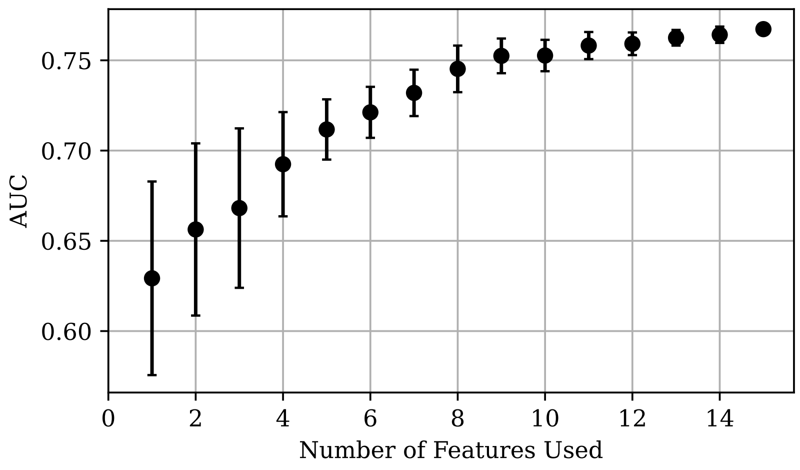


Figure 4.14: Model performance as a function of the number of features. The error bars indicate the standard deviation of the fifteen runs.

are relatively consistent when used for determining the relative importance of features using single feature models. The AUC and Brier scores differ more for permutation importance, again pointing out the limitation of permutation importances which was already discussed in Subsection 4.3.2. Furthermore, comparing the scores for single feature models and permutation importance provides an important indicator of the added value of a feature in the model. A high single-feature model score implies that, in general, a feature performs well at separating buildings that have adopted PV from buildings that have not. A high permutation importance score implies that the feature provides added value to the predictive model used in this study, combined with the other features and the chosen model setup.

Table 4.3: A comparison of different feature importance metrics.

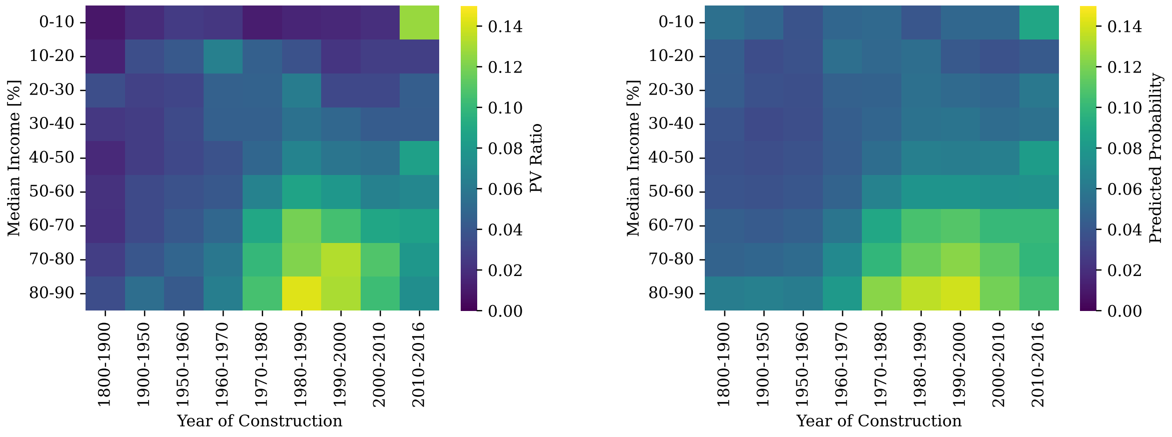
	Single Feature Model Rank		Permutation Importance Rank		
	AUC	Brier	AUC	Brier	Mean Rank
Suitable Area	2	3	3	2	2.5
Building Height	1	1	7	8	4.25
Roof Flatness	4	4	4	7	4.75
Year of Construction	9	9	1	1	5
Registered Addresses	6	5	6	5	5.5
Building Usage: Residence	3	2	8	11	6
Surrounding PV Ratio	10	11	2	4	6.75
Roof Complexity	5	6	12	6	7.25
Roof Area Per Surface	7	7	9	10	8.25
Median Income	11	10	5	13	9.75
Address Density	13	15	10	3	10.25
Irradiation	8	8	14	14	11
Owner Occupied Houses	14	14	11	9	12
Neighbouring Buildings	12	12	13	12	12.25
Residents 25-44	15	13	15	15	14.5

Discussion on Comparison to Related Work

Suitable area scores relatively high in each metric and can, therefore, be considered to be important in this model and are likely to be useful in other models as well. Furthermore, building height, flatness, year of construction, number of registered addresses and building use can all be considered important features. This conclusion is partly in line with the findings of other studies. Lee and Hong (2019) find residence building usage as the most important feature in their logistic regression (LR) model compared to other features such as surrounding PV ratio, resident age and household density. They also used a number of geometric features similar to this study; the total rooftop area and rooftop solar potential (based on irradiance, measured in kWh/m^2/yr). These were found to be insignificant by a $p < 0.1$ criterion. This is in contrast to the findings from this study, where the suitable area is found to be one of the best predictors. A possible explanation is a difference in feature generation; this study uses both irradiance and obstacle detection to determine the absolute suitable area, while Lee and Hong (2019) only use irradiance for PV potential and normalises this by dividing by the roof area. It is, however, still interesting to see that the total rooftop area was also not found to be significant. The total area and suitable area of a roof are positively correlated (i.e. a larger total area often corresponds to a larger suitable area), so based on the feature importance of suitable roof area, one would not expect the total roof area to be found to be insignificant.

Another explanation for the differences is the difference in the study area. Lee and Hong (2019) looked at buildings in one single district in the Gangnam district, which has the highest energy consumption and household income in South Korea. This indicates that conclusions drawn from this study might not apply in other areas. Lastly, the model setup is another possible cause for the difference in outcomes. In an LR model, complex relations between the input features and the dependent variable are not as easily captured as with RF. For example, the year of construction was found to be insignificant by Lee and Hong (2019). The single feature model score for year of construction is relatively low, as can be seen from Table 4.3, which implies that on its own, the feature only has limited capabilities of separating adopters from non-adopters. However, the year of construction was the most important feature of the RF model. This could mean that the year of construction adds important context to the model, such that the model performance increases in combination with other features.

The non-linear effect of features can be illustrated by the interplay between median income and year of construction. Figure 4.15a shows that buildings with a higher median income have, on average, adopted PV more often than buildings with a lower median income. This effect is strongest for buildings built between 1970 and 2010. One could, for example, hypothesise that residents living in older buildings with a higher median income would prefer investing in insulation and other energy-saving measures before investing in PV. In comparison, residents in newer buildings are more likely to have more energy-efficient houses already. Figure 4.15b shows the model predictions reflect the pattern seen in Figure 4.15a. An RF model can capture such non-linear effects, but linear models like LR models cannot capture these effects as easily. Therefore, this can be a possible explanation for why Lee and Hong (2019) found the year of construction to be non-significant while this study concludes it to be one of the most important features. Future improvements could focus on accounting for confounding factors and including a temporal component of features.



(a) The Measured PV Ratio as a Function of Median Income and Year of Construction

(b) The Average Predicted Probability of PV Adoption as a Function of Median Income and Year of Construction

Figure 4.15: Two heatmaps displaying the average PV ratio (model input) and predicted probability of PV (model output) as a function of two features: median income and year of construction.

Building height and the number of registered addresses are the only features ranked in the top 5 that have not been discussed in other literature. The high importance of building height can partly be attributed to its high correlation with residence buildings; high buildings are often used for residence and, as other literature has shown (e.g. Lee and Hong (2019)), residence buildings show a higher probability of PV adoption compared to other buildings. In addition, low buildings might receive less solar energy, which could decrease the probability of PV adoption for these buildings. These correlations cause the building height to have the highest VIF score (3.50) of all features. However, despite these correlations, it still ranks fourth in permutation importance, which points out that leaving this feature out does lead to model performance loss. Therefore, it can be concluded that it has added value to consider building height for PV adoption prediction.

Another interesting conclusion is the difference in performance between features on a building level and features on a postal code level. None of the six highest-scoring features is on a postal code level, while 4 of the six lowest features are on a postal code level; median income, owner-occupied houses, address density and residents. This result can be expected because postal code-level features are averaged over a group of buildings and can, therefore, never be as specific as building-level features. As a result, they are less likely to be able to explain the variance in PV adoption and add less to the performance of a model. However, the surrounding PV ratio and median income are shown to be of value to the model, albeit less than some building-level features. They both show a lower permutation importance rank than the single-feature model rank. An interpretation of these results is that these postal code-level features provide a general context for the model and that the building-level features are used to predict the probability, taking this context into account.

Discussion on Temporal Changes in Data and Models

Subsection 2.2.1 described the data sources of this study, including their quality. A potential major issue in data quality is data being outdated. 3D BAG relies on AHN, so geometric features derived from 3D BAG or the AHN raster are, at most, six years outdated. The BAG dataset is updated more frequently. Although registration delays cause errors in this dataset, one can assume that the number of registered addresses and the building usage are likely more up-to-date than geometric features.

The largest changes in input data presumably occur in the socioeconomic data. Especially income and resident age showed high changes over time; in roughly 30% of the cases, the change in income or resident age was larger than 10% on average, measured over three years (see Subsection 3.2.3). Recent data from the CBS is still redacted, so a study such as this thesis has to use data that is outdated by at least a few years. The dynamic nature of the socioeconomic data has two main consequences.

First, the models use outdated data for training, which could decrease performance. Resident age, income and owner-occupied houses all rank low in feature importance. The previous paragraph already provided a possible explanation for this (i.e. postal code-level features are averaged over a group of buildings and can, therefore, never be as specific as building-level features). However, the dynamic nature of socioeconomic features could be an additional explanation. Geometric, building-level features are outdated by up to six years, but Dukai et al. (2021) estimate that still, 95% of the building geometries are valid. When comparing this to socioeconomic features that change in the order of 10% yearly, one can argue that the quality of the socioeconomic data is lower because it is likely to have changed substantially in the meantime. Moreover, buildings built after 2015 were not considered in this study because a likelihood existed that these would not have valid geometries in the AHN (and thus 3D BAG) dataset. This decreased the chance of outdated geometric features. Excluding buildings with a high chance of outdated socioeconomic features was not done because it was not considered to be implemented easily. A possible method that can be considered for an improved model is assessing which areas show high changes in socioeconomic data in previous years and account for this dynamic nature, e.g. by assigning lower weights in training data. Moreover, one could test this hypothesis by investigating whether a predictive model performs worse in areas with fast-changing socioeconomic features.

Second, it cannot be guaranteed that the effect of socioeconomic features will remain constant in the future. For example, the current rise in energy prices could impact the effect of income on the willingness to PV adoption. A study by Bashiri and Alizadeh (2017) strengthens this hypothesis. Most research discussed in Section 1.2 shows a positive influence of income on PV adoption. However, Bashiri and Alizadeh (2017) show a negative influence and explain that this can be attributed to the profitability and income derived from photovoltaic systems. This led to low-income residents becoming more likely to adopt PV. Such effects decrease the usability of models based on past socioeconomic data. This effect could be incorporated into a future study by including temporal data on energy prices and building-scale energy consumption.

4.3.4. Conclusions on Feature Importances

This section analysed the importance of the features in the model. The third research question, *Which geometric and socioeconomic features are important to the model when predicting PV adoption?*, can already be partially answered by this section's conclusions. Different methods and metrics exist for determining the importance of features, each with its advantages and disadvantages. As a summarising answer to the question, one can say that the results indicate that buildings with the following characteristics are more likely to adopt PV when compared to buildings with opposing characteristics: (i) buildings with a larger suitable area, (ii) buildings with a higher rooftop, (iii) buildings with non-flat roofs, (iv) buildings that were built more recently, (v) buildings with a lower number of registered addresses and (vi) buildings used for residence. Furthermore, building-level features tend to add more value than regional features.

4.4. Performance Evaluation

It is useful to know where the model performs best and worse. This is important information for the possible applicability of the model. This section first investigates which building characteristics lead to the largest model mistakes. Secondly, it shows for which characteristics the model performs best and worst.

4.4.1. Largest Mistakes

In classification, so-called false positive and false negative rates (explained in Table 2.1) are often used for model assessment. This cannot be used directly for this study because the model's output used a probability. However, a similar analogy can be used; a high predicted probability for a building with no PV adopted is similar to a false positive (also called a type-I error), and a low predicted probability for a building that adopted PV is similar to a false negative (type-II error). The 1% highest predicted probabilities for buildings with no PV adopted (and vice versa) were considered the worst mistakes of the model. One can visualise this as the buildings in the right bins of Figure 4.4a and the left bin of Figure 4.4b.

The most interesting characteristics of the buildings with the largest mistakes are shown in Table 4.4 and 4.5. In Appendix A, Table A.4 and A.5 show these, and more, values for each feature.

Table 4.4: Characteristics of buildings with highest predicted probabilities for no PV adopted.

	Dataset Mean	Largest Mistakes Mean	Scaled Difference
Surrounding PV Ratio [-]	0.046	0.095	+108%
Building Usage: Residence [-]	0.520	0.976	+88%
Suitable Area [m^2]	78.2	179.9	+130%

Table 4.5: Characteristics of buildings with lowest predicted probabilities for PV adopted.

	Dataset Mean	Largest Mistakes Mean	Scaled Difference
Registered Addresses [-]	1.01	0.139	-86%
Building Usage: Residence [-]	0.864	0.041	-95%
Suitable Area [m^2]	108.4	11.2	-90%

Table 4.4 shows that large, residential buildings in an area with high historical PV adoption had the most "false positives". This can possibly be explained by looking at the probabilities of PV adoption per feature in Figure 4.11 and 4.12. The model was trained on data that showed that large, residential buildings in an area with high historical PV adoption usually have a higher chance of PV adoption. Therefore, it could have had more difficulty detecting the buildings that have not adopted PV in these areas. Table 4.5 points out a similar pattern. The most "false negatives" occur in areas with a relatively low PV ratio. The model had learned to predict low values for these characteristics, so buildings with these characteristics that did install PV are not predicted well by the model.

These errors are undesirable but also hard to eliminate. It is possible that two buildings with the same feature values still show a difference in PV adoption. The model in this study only considers a limited amount of information; many factors influencing the decision to adopt PV are not included in the features, making it impossible to obtain a perfect model with an AUC of 1.0.

The aggregated results can also be presented differently to understand the false positives and negatives further; Figure 4.16 shows the average frequency of PV adoption of the results aggregated by predicted probability. Binning was done for each incremental 0.02. For example, the black dot at the top right shows represents the buildings that were predicted a probability between 0.38 and 0.40. These buildings have an average frequency of PV adoption of 0.45, i.e. approximately 45% of these buildings adopted PV. Moreover, this implies that 55% of the buildings in this group did not adopt PV, despite the high predicted probability. These are the buildings that are labelled as false positives. Likewise, the leftmost point in the graph indicates that buildings with a predicted probability between 0.00 and 0.02 have, on average, adopted PV in 0.7% of the cases. This implies that for 1000 buildings with a predicted probability below 0.02, 7 buildings adopted PV and 993 did not.

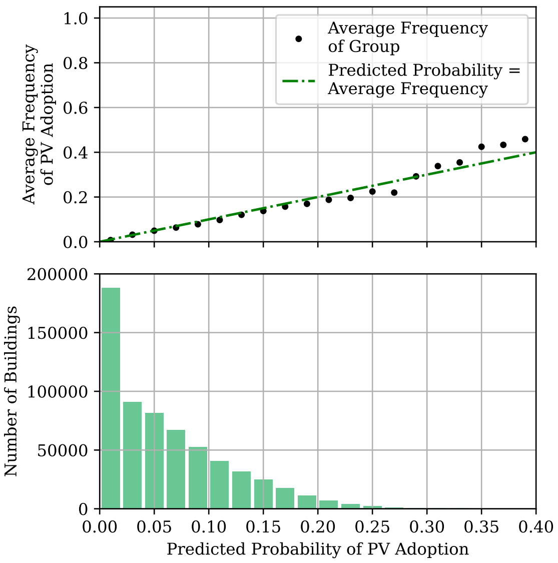


Figure 4.16: Average frequency of PV adoption as a function of the predicted probability of PV adoption.

However, this would not necessarily mean that it results from overfitting. This can be illustrated by taking the example of probabilities between 0.38 and 0.40. On the training data, it learned that buildings with certain characteristics adopted PV in approximately 39% of the case and, therefore, predicted a probability of 0.39 on buildings with these characteristics in the testing data. Although this is not equal to the true frequency of PV of 45%, it is still relatively close. One can observe in Figure 4.16 that the predicted probability often matches the frequency of PV adoption of the group, indicating that the model is, on average, relatively unbiased in predicting probabilities. Also, note that the points start to deviate from the dash-dotted line more for higher predicted probabilities. This can presumably be attributed to the decrease in the number of buildings in these groups. The groups of buildings with probabilities above 0.30 consist of 200 to 600 buildings as opposed to groups of over 10 000 buildings for lower probabilities.

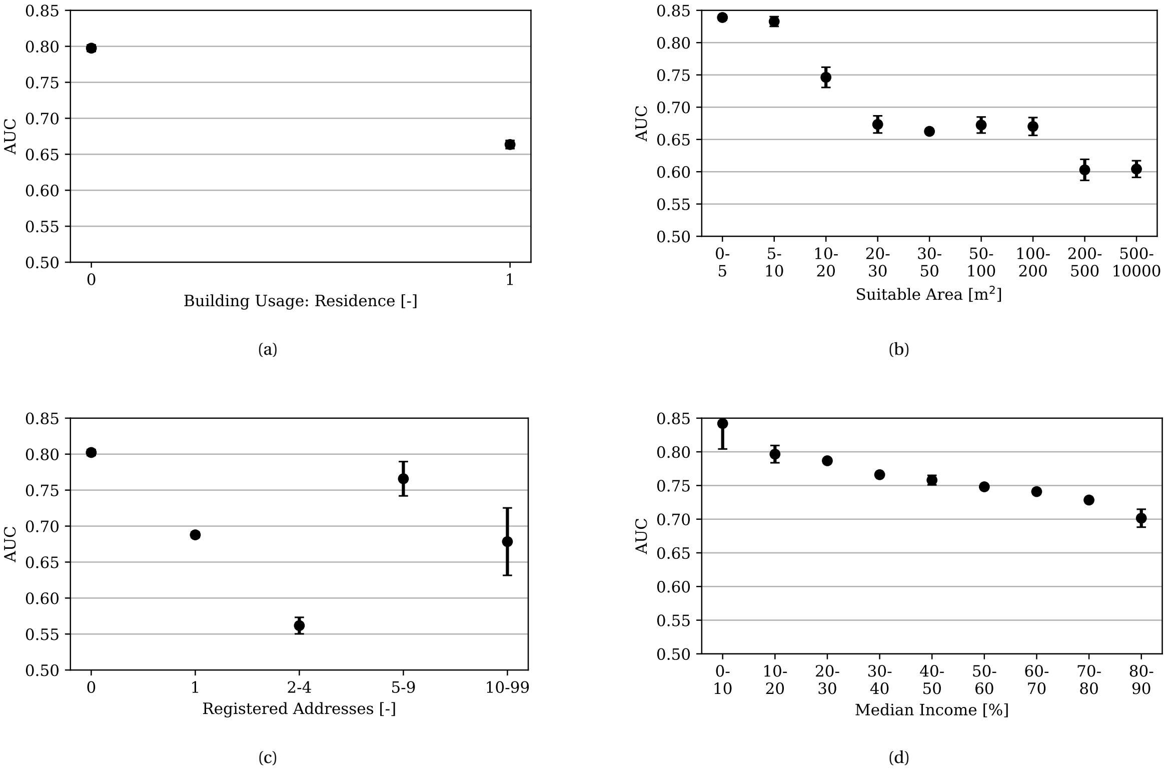
4.4.2. Best and Worst Model Performance

An Investigation of the largest model mistakes only considers the mistakes on a building level. There is another method to investigate where the model performs best and worst; subdividing the dataset results into bins based on feature values and looking at model performance for each feature value. This considers a group of results; therefore, the AUC can be computed for each. This analysis was carried out for each feature, which is shown in Appendix A in Figure A.4 and A.5. Figure 4.17 shows some of the most interesting results.

Firstly, Figure 4.17a shows that the model performed significantly better for non-residential buildings than residential ones. A possible reason is the relatively high number of PV installations on residential buildings. Non-residential buildings have PV installations less often, so the model could predict low probabilities if buildings are non-residential, which will often be correct. On the other hand, residential buildings have more variability within their group and are, therefore, more difficult to predict.

Furthermore, Figure 4.17b depicts that the model performs best for buildings with a low suitable area. This is a logical result; small buildings with insufficient space for PV will rarely adopt PV, regardless of other features such as the owners' income. A model can, therefore, be relatively sure that no PV was adopted on buildings with a suitable area below 10 m^2 and performed well in these areas. If the roof area is large enough for PV, other factors will also play a role for the building owner or stakeholder. These are harder to predict for models, and, as a result, the performance is lower for higher suitable areas. Buildings with a very small suitable area are often buildings in backyards, like sheds or garages, and these are not considered residential buildings and have no registered address. The high predictability of these buildings could also play a factor in the large difference in model performance between residential and non-residential buildings. This could also explain why the model performs best for buildings with no registered addresses, as seen in Figure 4.17c.

In addition, Figure 4.17d shows a similar pattern; the model performs best for buildings with a lower a priori change of PV installation. Figure 4.15 provides a good example. For lower median incomes, the PV probability generally is lower, regardless of the year of construction. As a result, the model had more accurate predictions for buildings with a low median income value.



(a)

(b)

(c)

(d)

Figure 4.17: Model performance as a function of different values of socioeconomic and geometric features. The error bars indicate the standard deviation of the three outer folds.

4.4.3. Conclusions on Performance Evaluation

This section has shown that the largest absolute errors occur in two regions; false positives occur in regions with high PV ratios, and false negatives occur in regions with low PV ratios. The model had learned to predict high and low values for these regions, respectively, so buildings deviating buildings in these areas are the cause of the larger errors.

Furthermore, Subsection 4.4.2 showed that the model performed better in areas with lower PV rates, and lower variability made it easier to predict. One could say that the developed model performed well at predicting where PV was not adopted but was less accurate in predicting where PV was adopted. Although this can be explained, it can be considered undesirable for most applications. Furthermore, the model performs poorer for residential buildings. These two model characteristics make this model less applicable for aiding policymakers in PV stimulation, despite the relatively good overall model AUC score of 0.767.

4.5. Case Studies

This section presents case studies of three different areas to put the performance and results of the model into more context. The first two case studies concern urban residential areas, and the third examines a rural region.

4.5.1. Residential Urban Area - Terraced Houses

Figure 4.18 displays part of a residential neighbourhood in the North of Zwolle, Overijssel, where most buildings consist of terraced houses. The colours of the houses correspond to the predicted probabilities. Five colours are used, corresponding to the five quantiles. For example, "Low" corresponds to the lowest 20% predicted probabilities, and "High" corresponds to the highest 20%. Buildings of each quantile are present in this figure. Several elements in this figure are worth discussing.

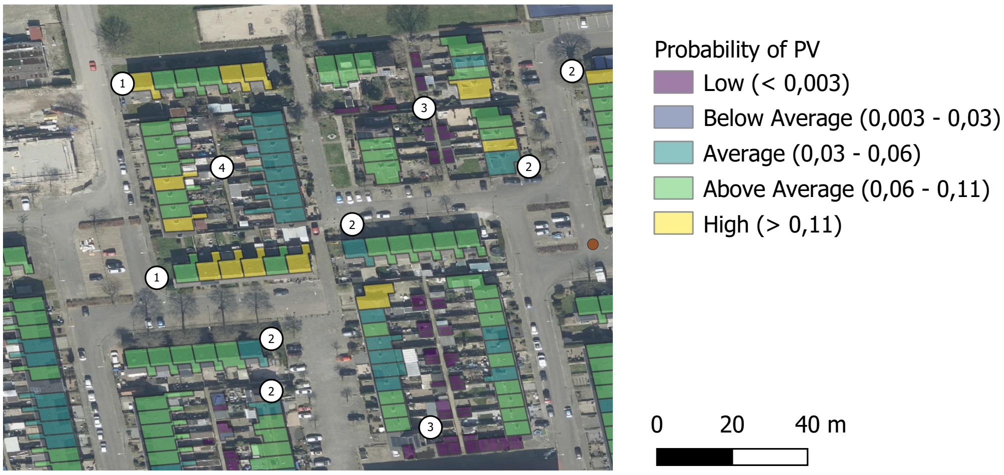


Figure 4.18: Probability of PV adopted plotted over an aerial image in a residential urban area with terraced houses in Zwolle, Overijssel.

First, two rows of buildings in the top left (both indicated by the "1") receive relatively high predicted probability. The largest difference between those buildings and the other buildings is the suitable area. The suitable area of these buildings is approximately 30 to 40 m^2, which is more than other buildings in the figure, whose suitable areas are 10 to 30 m^2. Section 4.3 confirmed that buildings with suitable areas above 30 m^2 indeed more often have adopted PV.

Moreover, the figure shows that corner houses often show deviating behaviour. For some corner houses, the suitable area is significantly lower or higher than the neighbouring house due to differing geometries. Furthermore, these corner houses have one neighbouring building, while the middle houses all have two neighbouring buildings. Figure 4.11 showed that houses with two neighbouring buildings have adopted PV more often, so this can explain why the model predicts a lower probability for corner houses. However, more variability exists between houses within one row. The postal code level features are the same for the buildings, so any differences in predicted probability have to be the result of different building-level features. In the area displayed in Figure 4.18, the biggest differences are roof complexity, suitable area, and area per surface. Any differences in predicted probability likely result from those features differing. Nevertheless, it is not always clear what causes the differences in predictions between houses. This is a big disadvantage of using an MLM such as RF; the decisions are based on, in this case, 100 decision trees of 7 steps, so every prediction results from a complex interplay between feature values.

Furthermore, several small buildings near the number "3" all receive a low probability. These shed-like buildings all have a height below 3 m, a suitable area below 10 m^2 and are non-residential buildings. These three feature values indicate that these buildings are unlikely to have adopted PV; hence the model predicts a low probability. Unlike the residential buildings, these sheds are all assigned to the same class, which shows that the model can easily distinguish these small buildings.

Lastly, an interesting difference can be observed between the building rows left and right of the number "4". The roofs have similar geometries, but the left row of buildings receives significantly higher predicted probabilities. The only difference in feature values between the two rows is the value of residents aged 25-44. The value for the left row is 0.13, meaning that 13% of the residents in those buildings are aged between 25 and 44 years old. For the right row, this value is 0.35. Figure 4.12 displays that for a value of 0.13, PV adoption is relatively high, while for a value of 0.35, PV adoption is relatively low.

In conclusion, one can observe patterns in predicted probabilities for buildings in a residential neighbourhood. Larger houses receive high probabilities; small sheds receive low probabilities. Probabilities between rows of houses can sometimes be attributed to differences in postal code level features. However, some variability exists between houses that cannot always be explained.

4.5.2. Residential Urban Area - Detached Houses

Figure 4.19 shows an area in Rijssen, Overijssel, where most buildings consist of detached houses. The characteristics of this area differ significantly from the area shown in Figure 4.18. All houses in this area are owner-occupied, the neighbourhood is richer, the buildings were built more recently, and the area is less densely populated. The values are shown in Table 4.6.

In this area, a significantly larger number of buildings has a high predicted probability. This can partly be attributed to the larger values for suitable area in this neighbourhood. However, the buildings indicated by a "1" all have suitable areas below 30 m^2 and are still predicted to have a high probability of PV adoption. Buildings with suitable areas in the discussed Zwolle neighbourhood were usually predicted to have average or above-average probabilities. The characteristics of this neighbourhood (owner occupied, richer, recently built and sparsely populated) are characteristics of buildings that have adopted PV more frequently, as Section 4.3 pointed out. Herefore, the difference in prediction behaviour between the two areas is likely due to the difference in the values for postal code level features.

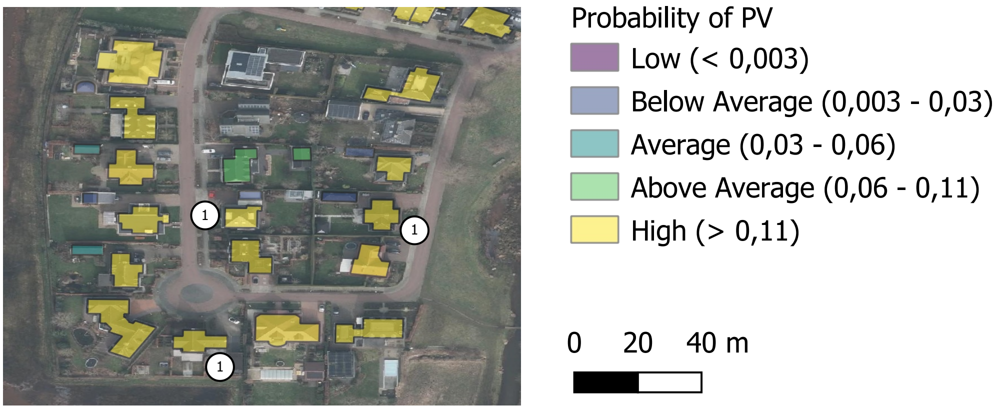


Figure 4.19: Probability of PV adopted plotted over an aerial image in a residential urban area with detached houses in Rijssen, Overijssel.

Table 4.6: A comparison of some postal code level features between to case study areas.

Feature	Average Feature Value	
	Study Area in Zwolle	Study Area in Rijssen
House Ownership	0.67	9
Median Income	25	75
Year of Construction	1960	2002
Address Density	2294	965

The probability distributions of the two areas are also shown in Figure 4.20. The figure presents normalised histograms of the predicted probability. A clear difference between the two is visible; the neighbourhood in Zwolle, with the detached houses, has more buildings with probabilities below 0.10 and barely any with probabilities above 0.15. In contrast, the probabilities of the neighbourhood in Rijssen often exceed 0.15.

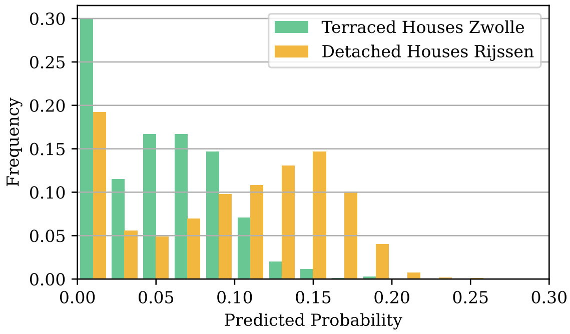


Figure 4.20: Comparison between the probability distributions of the two neighbourhoods discussed in the case studies.

4.5.3. Rural Area

Figure 4.21 shows buildings in a rural region with a similar colour scale as Figure 4.18. The figure shows high variability in predicted probabilities between buildings.

The highest predicted probabilities occur at the buildings indicated by "1". Interestingly, the suitable area of these buildings is relatively low compared to other buildings in the area. Figure 4.11g showed that buildings that adopted PV often have a larger suitable area, but only up to 100 m^2. Between 100 and 500 m^2, the average PV adoption decreases for increasing suitable area. This relation is a possible explanation for the high predicted probabilities for the smaller buildings in Figure 4.18. However, the building indicated by "2" has an above-average predicted probability, despite being the largest building in the figure. The suitable area of 2446 m^2 is significantly higher than that of the other buildings. A reason could be that a very large suitable area could lead to a higher predicted probability. An inspection of the results shows that in rural areas, either the smaller or the largest buildings of a building group have PV adopted. One could hypothesise that there are two types of PV installation in rural areas. Some rural building owners would like to install PV on a small scale to prevent high start-up costs. On the other side, some rural building owners opt for a large-scale PV installation as part of an investment. In the second case, the largest roofs are preferred. This could explain why relatively small and relatively large buildings are used for PV installation more often than the middle group. However, this is only a hypothesis that would need to be tested in a different model. The difference between the probabilities points out the same problem addressed in the previous section; RF model decisions cannot always be explained easily.

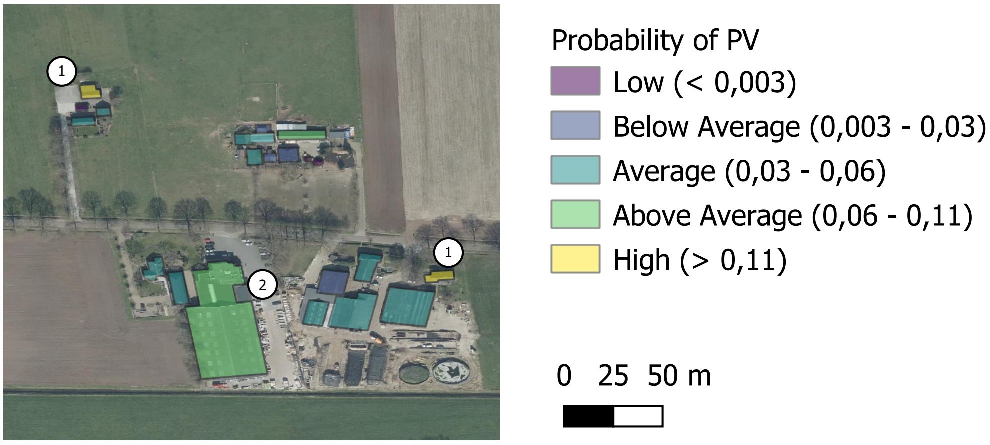


Figure 4.21: Probability of PV adopted plotted over an aerial image in a rural area near Zwolle, Overijssel.

4.5.4. Conclusions on Case Studies

The three case studies from the past sections spatially demonstrated the results. In urban areas, very small buildings such as sheds are easily distinguished and predicted to have a low probability of PV adoption. Residential buildings show predicted probabilities varying from below average to high. These variations can often be attributed to differences in suitable area and sometimes to differences in postal code level features. In some cases, the variety in predicted probability cannot be explained easily, which is a drawback of using an MLM such as RF.

4.6. Suitable Area

Section 3.4 explained how the suitable area was computed for each roof. The main purpose of this computation was to create a feature to be used in the PV adoption prediction model. Nevertheless, the calculations form a significant part of the method, and therefore, this section evaluates the results of the suitable area computation separately.

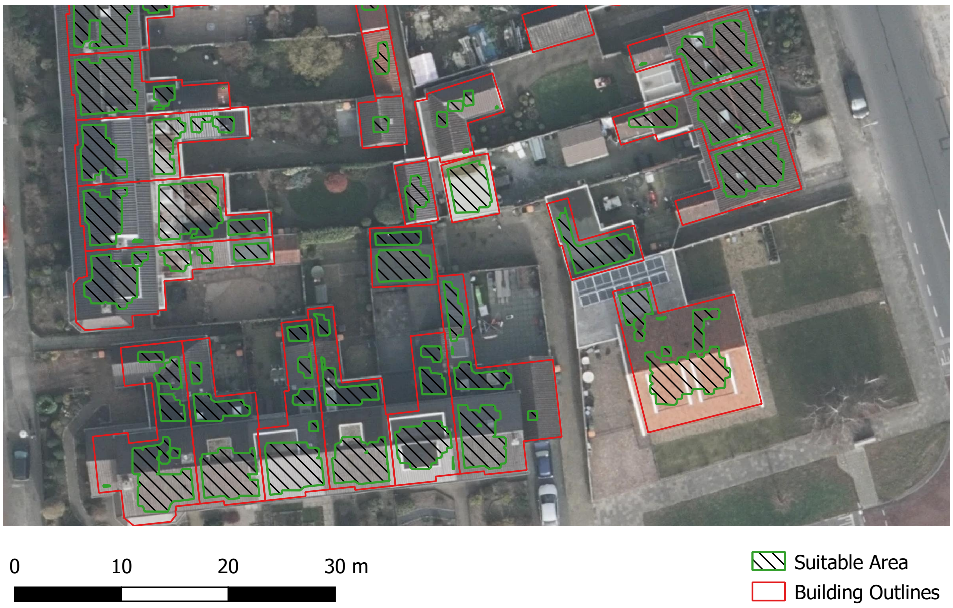


Figure 4.22: A map showing the computed suitable area for buildings in Oldenzaal.

4.6.1. Area Used For PV

None of the available datasets contained validated suitable areas per building, so a different dataset was chosen for evaluating the suitable area. The values of the suitable area were compared to the area used for PV. The PV area per roof dataset is also provided by Sobolt (2022a), who have developed an algorithm based on a convolutional neural network (CNN) that estimates the area used for PV per building. Furthermore, they also have a dataset of 600 buildings, of which the PV area was determined manually to validate the accuracy of the CNN.

Figure 4.23 shows the performance of the CNN compared to the manually verified PV data. A clear correlation is visible, which is confirmed by the R^2 value of 0.64. The standard deviation of the error is 11.1 m^2 and the bias equals 2.65 m^2. The positive bias indicates that, on average, the CNN overestimates the PV area by 2.65 m^2. If, for example, the true PV area equals 50 m^2, then there is a 68% chance that the CNN predicts a PV area between 41 m^2 and 64 m^2. The average roof area used for PV is 23 m^2. Te

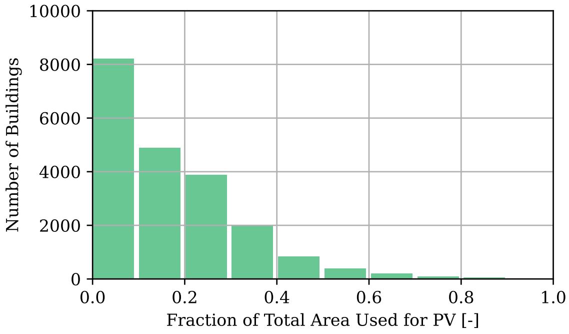


Figure 4.24: PV area as a fraction of total roof area for buildings with a total area smaller than 100 m^2.

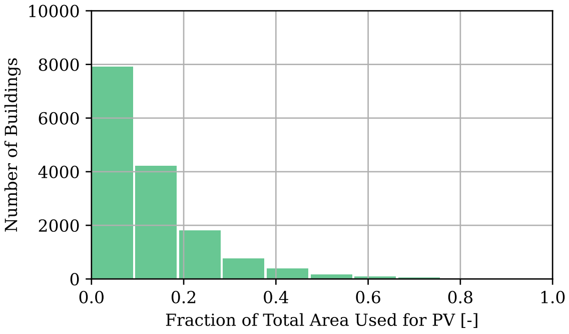


Figure 4.25: PV area as a fraction of total roof area for buildings with a total area larger than 100 m^2

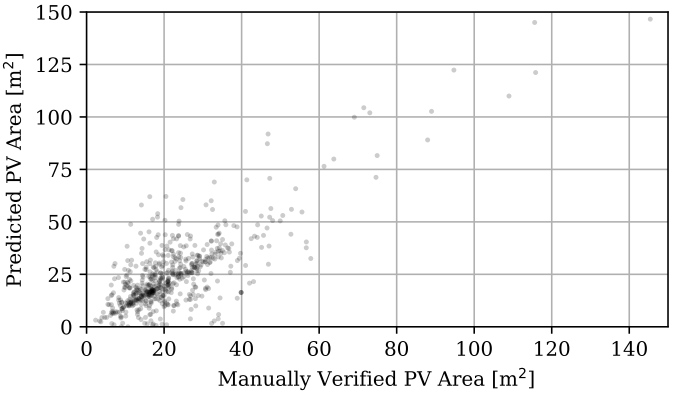


Figure 4.23: The estimated PV area based on a CNN model plotted against manually verified PV area for 600 buildings.

The data obtained by the CNN model can be compared to the total area of the roof to investigate how much of the roof is used for PV. Figure 4.24 and 4.25 show histograms of the fraction of the roof area used for PV area, for small buildings and large buildings, respectively. Only buildings that installed PV in 2019 or earlier have been included. This is the first year PV adoption data in Overijssel was available. Only including the oldest year decreases the chance of analysing buildings for which the geometry has changed between the moment of AHN3 acquisition and the computation of the area used for PV. The analysis included a total of 35924 buildings. The figures point out that buildings rarely use the entire roof area for PV. The distributions of large and small buildings are shaped similarly, although the larger buildings use a smaller fraction of the total roof area for PV.

4.6.2. Results

Figure 4.26 and 4.27 show the predicted suitable area as a function of the PV panel area. The figures point out that a positive correlation exists between suitable area and PV area, which is to be expected. A comparison between the predicted suitable area and the PV area yields an R^2 of 0.694. Furthermore, the figures show that the predicted suitable area far exceeds the PV area. This is confirmed by the histograms shown in Figure 4.28 and 4.29. These figures show a similar pattern as Figure 4.24 and 4.25; the distributions are roughly similar, but large roofs use a smaller percentage of the suitable area.

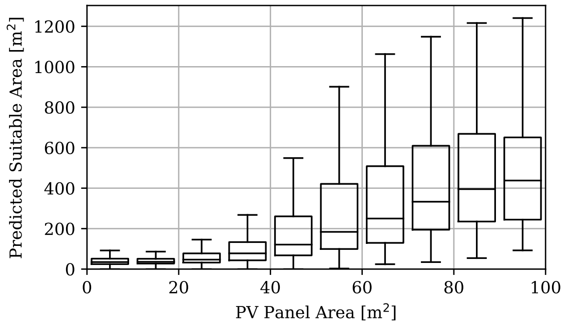


Figure 4.26: Box plots of the predicted suitable area for buildings with a pv area smaller than 100 m^2.

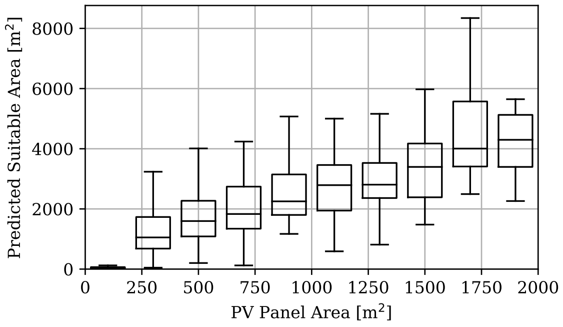


Figure 4.27: Box plots of the predicted suitable area for buildings with a pv area smaller than 2000 m^2

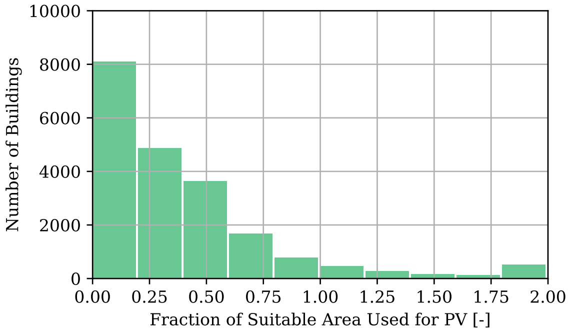


Figure 4.28: Histogram of the PV area as a fraction of the suitable area for buildings with a total roof smaller than 100 m^2

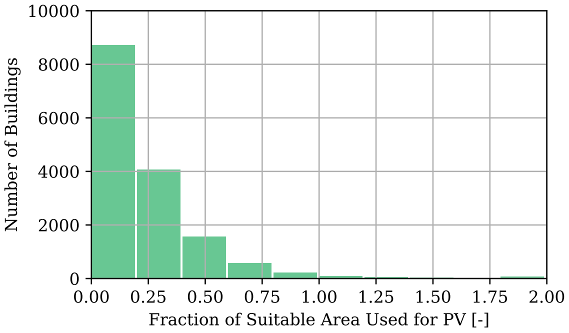


Figure 4.29: Histogram of the PV area as a fraction of the suitable area for buildings with a total roof larger than 100 m^2

4.6.3. Discussion

The figures presented in the previous subsections show a correlation between the PV area and the predicted suitable area. However, a large variability exists between the suitable area and the PV area; an R^2 value of 0.694 indicates that the suitable area can only explain 69% of the variance in PV area. A part of the remaining variance can be the cause of the fact that only some buildings use all available area on the roof when installing PV. Because some buildings use all available space and others do not, the suitable area will never be able to explain all variance in PV area.

Nevertheless, the results point out some of the shortcomings in the model. For example, Figure 4.28 and 4.29 show that the PV Area sometimes exceeds the suitable area. For small buildings (i.e. a total roof area smaller than 100 m^2), this happened in 7.3% of the cases and for large buildings in 1.5% of the cases. This means that in these cases, the suitable area was incorrect; some parts of the roof were deemed unsuitable, but PV was still installed there. This could be interpreted as similar to a false negative or type-II error.

To investigate the model's added value, one can compare it to a simpler model. If the suitable area is estimated by taking a constant fraction of the total area, an R^2 value of 0.656 is obtained. This is 0.038 lower, which means that the model in this study can explain approximately 3.8 percentage points more than a simple model. This shows that the computation of the suitable area as described in Section 3.4 can predict PV area better than a simple model, but the gain can be considered marginal. Based on this small difference, one can question the added value of such a complicated method.

However, it is important to note that the evaluation dataset is not the desired dataset for a proper evaluation of the method. It was already mentioned that the suitable area would never be able to explain all variance in PV area. In addition, the dataset only includes buildings that adopted PV, which could increase a bias towards the simple model. The following example can illustrate this; consider a building with a total area of 20 m^2 and a PV area of 15 m^2. If the proposed method would, wrongly, predict a suitable area of 5 m^2, then the simple model would be more correct. On the contrary, consider a building w with a total area of 20 m^2 without PV installed, with a predicted suitable area of 5 m^2. In this case, the more elaborate method is more correct. The second case is more likely to happen, considering that buildings with a suitable area of 5 m^2 rarely adopt PV (see Figure 4.11g). Nonetheless, only the first case is included in this analysis since only buildings that adopted PV were considered. As a result, the actual difference in performance between the simple and elaborate models could be larger.

This study used the AHN raster to determine the suitable area. One could argue that using point raw cloud data instead of processed data could improve the model's accuracy since averaging over a raster leads to a loss of information. Furthermore, aerial images could also be used to aid in detecting obstacles. Apra et al. (2021) explain that shadows cast by obstacles on the roof complicate the detection of obstacles since these shadows are also identified as obstacles. Although this is indeed a potential problem, one can argue that these locations already have a decreased suitability for PV due to the shadow, so this is not as problematic when considering suitability for PV.

A different evaluation dataset is needed to assess the proposed method's performance better. A similar method as used by Apra et al. (2021) is likely to provide a better understanding of the model's performance; they had a dataset of pixels with manually drawn obstacles. Such a dataset can then estimate whether obstacles are correctly classified as obstacles. Furthermore, one would also need to exclude pixels that receive too little solar energy. This is still not trivial since a certain threshold would have to be chosen.

Nevertheless, it would already substantially help in assessing the performance. Furthermore, the analysis only considered the scalar value of suitable area and PV area. However, both can be visualised as polygons, and their locations can be compared. This could also provide more insight into the model's most common mistakes.

A final suggestion for future research would be to perform a sensitivity analysis on the input parameters of this model. A number of parameters were chosen manually; threshold values for slope change and power, the polygon buffer size, the margin from the roof edge and the simplification tolerance. One could study how changing these parameters influences the results and the accuracy.

4.6.4. Conclusions on Suitable Area

This section described the performance of the suitable area estimation. It is hard to evaluate the suitable area properly because no ground truth dataset was available. A comparison pointed out that the suitable area shows a positive correlation with the PV area (R^2 = 0.694), but this is only a marginal gain with respect to a simple model (R^2 = 0.656), i.e. using the total area. The main purpose of the suitable area estimation was to use this as a feature in PV prediction. Considering that suitable area was found to be the most important feature in PV adoption prediction, one can argue that this main purpose is, at least partially, fulfilled. However, improvements in both the method and result assessment are recommended if this model were to be applied in a different study.

5 Conclusions and Recommendations

This thesis presented how one can predict photovoltaic system (PV) adoption per building using socioeconomic and geometric features. This chapter contains the key points, main conclusions, and recommendations for further research.

5.1. Conclusions

This section first presents the key points of this study by providing answers to the four research sub-questions. Then, the general conclusion of this thesis is discussed.

Which features can we generate from geometric and socioeconomic data to predict PV adoption?
Seven geometric features were generated from 3D building data, building registrations and airborne laser scans. 3D building data was first processed to calculate slope and aspect histograms of each building's roof area. These histograms were then used to generate four features; roof complexity, flatness, area per surface and average irradiance. The number of neighbouring buildings and building height were calculated using registration data. Lastly, airborne laser scanning data was used to determine the suitable roof area by analysing which parts receive enough solar power and are obstacle free.

Eight socioeconomic features were generated from building registrations and socioeconomic postal code statistics. The year of construction, building usage and the number of registered addresses were directly taken from building registration data and required no further processing. House ownership, income, resident age and address density were mostly available on postal code resolution, but some data was classified. This data was filled up using the PC4 data or the dataset median. A remote sensing-based dataset on PV adoption was used to calculate each neighbourhood's average PV adoption ratio in 2019.

How can we accurately create and assess PV adoption prediction models?
The fifteen generated features were used to train a machine-learning algorithm. The dependent variable for the model was the change in PV adoption between 2019 and 2021, where adoption is inputted as a 1 and no adoption as a 0. If regression is applied to such a model, the output estimates the probability of PV adoption. Two metrics were used for assessing performance; AUC and Brier. This study has demonstrated that a random forest regression model and a neural network model can predict the probability of PV adoption, both achieving an AUC score of 0.77. Further analysis of the performance in certain areas indicated that most performance could be attributed to predicting where no PV adoption occurred.

It was also investigated whether combining geometric and socioeconomic features has led to a more accurate prediction of PV adoption probability. This was found to be true, but the effect is only a marginal increment of 0.03 in the AUC score; models with only geometric or only socioeconomic features had AUC scores of 0.75 and 0.74, respectively. The Brier score of the models was only marginally better than a naive model, indicating the importance of using a different metric for evaluating the performance of imbalanced datasets.

Which geometric and socioeconomic features are important to the model when predicting PV adoption?
This study used single feature models and permutation importance to determine feature importance, where the former indicates the predictive performance of an isolated feature and the latter measures the added

value to the model. The combination of these two assessments has shown that the most important features for PV adoption prediction in this model are; suitable area, building height, roof flatness, year of construction, number of registered addresses and building usage function. Furthermore, the results highlight the added value of using building-level features over postal code features when predicting PV adoption on a building level.

Which areas within the study area have a high chance of PV adoption according to the used model?
This question can partially be answered by examining the frequency distribution of the adoption of PV for different features and investigating model feature importances. Based on those findings, one can conclude that areas have a high chance of PV adoption if the buildings (i) have a relatively large suitable area, (ii) have a high rooftop, (iii) have a non-flat roof, (iv) were built recently, (v) only have one address registered and (vi) are used for residence.

Furthermore, the model results and the case studies point out that richer, sparsely populated urban areas also show higher probabilities of PV adoption. Nonetheless, the model evaluation showed that the performance is the poorest for residential areas, decreasing the applicability for the areas with higher chances of PV adoption. Additionally, the case studies showed that not all differences in predicted probability could easily be explained.

Main Conclusion
The main research question of this thesis was: *To what extent can we predict PV adoption at building scale using socioeconomic and geometric features?*. Section 4.2 demonstrated that models were able to predict PV adoption and achieve AUC scores up to 0.77, which means that, overall, it is possible to a good extent to separate PV adopters from non-adopters. However, in Section 4.4, it was pointed out that a large part of this performance can be attributed to predicting where no PV was adopted, e.g. on too small roofs. The decision to adopt PV likely depends on factors not included in the model and presumably on factors that are hard to include in any model (such as the exact decision-making process of individuals). This complicates predicting PV adoption compared to predicting non-adoption.

Some gains in model performance can possibly be achieved by optimising the model training strategy and adding other features. However, the unknowns in the decision-making process of adopters mean that a big gain in performance is not likely, also considering that performance gains were less than 0.02 in AUC score when adding more than eight features as a result of redundancy (see Section 4.3). If a high spatial resolution is not required, then predicting PV adoption on a regional scale (e.g. on a PC4 scale) is recommended, as this is likely to be more reliable. Nonetheless, in cases where spatial resolution is considered important, the proposed model is likely to be valuable; analyses in Section 4.3 showed that building-level features, especially suitable area, proved to be important predictors of PV adoption.

In conclusion, the models shown in this thesis have proven that building-level prediction is possible to a certain extent. However, one can question the model's additional value for policymakers in its current state, especially considering the low performance for residential buildings (see Section 4.4) and the lack of explainability of model decisions (see Section 4.5). Despite its current limitations, an improved model setup has the potential to help understand patterns in PV adoption as an addition to the readily-existing regional models. It could therefore aid policymakers in stimulating PV and contribute to realising energy transition.

5.2. Recommendations

The results and discussion point out several areas where the model needs improvement. This paragraph presents four suggestions for a further improved model and three directions for future research in predicting PV adoption.

Model Improvements
The first recommendation for increasing model performance is to create separate models for each building usage. The results proved a significant difference between residential and non-residential buildings and showed a lower model performance for residential buildings. Creating a separate model for residential buildings can make the model decision-making more explainable, as this removes the possible confounding factor of building usage. Furthermore, the assessment of feature importance becomes more reliable.

Second, creating a simplified model could increase explainability. This thesis has demonstrated which features have the highest predictive power. In addition, it shows that a reduction in the number of features from 15 to 7 only marginally impacts performance. A linear regression model or logistic regression model would make decisions more explainable. However, they are not capable of capturing non-linear effects. A trade-off could exist between explainability and performance. A suggestion for an improved, simplified model is to create a random forest model using, e.g. five features that are shown to be important for PV adoption prediction and are verified to be independent. This model is likely to perform slightly worse, but results can be explained more easily. If more explainability is required, logistic regression could be studied as well.

The third suggestion for model improvement is to optimise the machine-learning training strategy. For example, the random forest algorithm showed signs that the performance could be improved if more than 100 trees were used. Additionally, one could investigate training shallower trees and a different stopping criterion, as this will decrease the complexity of the model and possibly lead to performance gains simultaneously. Moreover, the neural network nearly matched the performance of the random forest, even though little optimisation was carried out. More research on optimising the neural network for this model could lead to gains in performance. In addition, the performance and mistakes of both models could be compared. Combining the two models could decrease the error if the algorithms perform worse for buildings with different characteristics.

Lastly, techniques for balancing the dataset can be implemented in the model. This decreases the risk of overfitting and presumably increases model performance. Section 4.2 discussed two possibilities; synthetic minority oversampling and implementing a costs function.

Future Research

Future research could focus on the effect of PV promotion campaigns. This thesis used changes in PV adoption as the dependent variable and included all buildings in the study area. A new research setup could only include buildings or areas where active PV stimulation occurred during the research period. In this way, the model predicts the probability that PV stimulation will result in PV adoption. As a result, a model can yield characteristics of buildings most susceptible to any PV campaigns. This information can help policymakers execute PV adoption more effectively by targeting those buildings with the highest probability of adoption. One could also consider a longer study period and coarser resolution in such a study, as this presumably increases the model's reliability.

Additionally, new features could have added value to the model. Section 4.3 explained that adding more geometric and socioeconomic features is not likely to lead to a gain in performance due to redundancy. Nonetheless, some features used in other literature could prove valuable to this model, provided that their correlation with the already-used features is low. For example, the number of GroenLinks voters proved to be an important predictor of PV adoption. Furthermore, in light of the recent energy crisis, features such as energy usage would be more interesting to incorporate into the model. Moreover, adding the change in features over the years is another way to improve the information contained in the model.

Lastly, a new model could be created that predicts PV area. The current model only predicts whether a building has adopted PV but does not consider the potential panel area. For example, large buildings show lower chances of PV adoption but are likely to have larger PV areas when they do adopt PV. If the goal is to maximise the share of sustainably generated energy, then the most interesting buildings for PV stimulation are those with the highest expected PV area.

References

3D geoinformation research group, T. D. (2022, 2). *3D BAG.* Retrieved from https://docs.3dbag.nl/en/copyright

Aarsen, R., Janssen, M., Ramkisoen, M., Biljecki, F., Quak, W., & Verbree, E. (2015). INSTALLED BASE REGISTRATION OF DECENTRALISED SOLAR PANELS WITH APPLICATIONS IN CRISIS MANAGEMENT. *ISPRS.* doi: 10.5194/isprsarchives-XL-3-W3-219-2015

Aïda Brands. (2022). *Strenger, groener, zuiniger: zo wil de Europese Commissie af van Russisch gas.* Retrieved from https://nos.nl/artikel/2429319-strenger-groener-zuiniger-zo-wil-de-europese-commissie-af-van-russisch-gas

Alam, N., Coors, V., Zlatanova, S., & Oosterom, P. J. M. (2016). RESOLUTION IN PHOTOVOLTAIC POTENTIAL COMPUTATION. *ISPRS.* doi: 10.5194/isprs-annals-IV-4-W1-89-2016

Apra, I., Bachert, C., Cáceres Tocora, C., Tufan, Ü., Veselý, O., & Verbree, E. (2021). INFERRING ROOF SEMANTICS FOR MORE ACCURATE SOLAR POTENTIAL ASSESSMENT. *ISPRS.* Retrieved from https://doi.org/10.5194/isprs-archives-XLVI-4-W4-2021-33-2021 doi: 10.5194/isprs-archives-XLVI-4-W4-2021-33-2021

APVI. (2022). *Australian Photovoltaic Institute • PV Postcode Data.* Retrieved from https://pv-map.apvi.org.au/postcode

Assouline, D., Mohajeri, N., & Scartezzini, J.-L. (2017, 10). Building rooftop classification using random forests for large-scale PV deployment. In U. Michel & K. Schulz (Eds.), *Earth resources and environmental remote sensing/gis applications viii* (p. 5). SPIE. doi: 10.1117/12.2277692

Bakshi, C. (2020, 6). *Random Forest Regression.* Retrieved from https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Bashiri, A., & Alizadeh, S. H. (2017). The analysis of demographics, environmental and knowledge factors affecting prospective residential PV system adoption: A study in Tehran. *Renewable and Sustainable Energy Reviews.* Retrieved from http://dx.doi.org/10.1016/j.rser.2017.08.093 doi: 10.1016/j.rser.2017.08.093

Bekkar, M., Kheliouane Djemaa, D., & Akrouf Alitouche, D. (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. , *3*(10). Retrieved from www.iiste.org

Biau, G., & Scornet, E. (2016, 6). A random forest guided tour. *Test*, *25*(2), 197–227. Retrieved from https://link.springer.com/article/10.1007/s11749-016-0481-7 doi: 10.1007/S11749-016-0481-7/FIGURES/4

Biljecki, F., Ledoux, H., & Stoter, J. (2016, 9). An improved LOD specification for 3D building models. *Computers, Environment and Urban Systems*, *59*, 25–37. doi: 10.1016/J.COMPENVURBSYS.2016.04.005

Bradley, A. E. (1997). THE EVALUATION OF MACHINE LEARNING ALGORITHMS. *Pattern Recognition*, *30*(7), 1145–1159.

Brahmachari, V., Jain, S., & Kimmig, A. (2013). Area under the ROC Curve. *Encyclopedia of Systems Biology*, 38–39. Retrieved from https://link-springer-com.tudelft.idm.oclc.org/referenceworkentry/10.1007/978-1-4419-9863-7_209 doi: 10.1007/978-1-4419-9863-7{_}209

Breiman, L. (2001). Random Forests. , *45*, 5–32.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers.

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

CBS. (2021). *Zonnestroom; vermogen bedrijven en woningen, regio (indeling 2019).* Retrieved from `https://www.cbs.nl/nl-nl/cijfers/detail/84783NED`

CBS. (2022a). *Inwoners per gemeente.* Retrieved from `https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/regionaal/inwoners`

CBS. (2022b). *Kerncijfers per postcode.* Retrieved from `https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode`

CBS. (2022c). *Wijk- en buurtkaart 2020.* Retrieved from `https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2020`

CBS and Kadaster. (2021, 1). Verkenning samenhang cijfers benut zonnestroomop regionaal niveau.

CBS and Kadaster. (2022, 8). *Detectie zonnepanelen regionaal* (Tech. Rep.).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique* (Vol. 16; Tech. Rep.).

Cheng, Z., Zou, C., & Dong, J. (2019, 9). Outlier detection using isolation forest and local outlier. *Proceedings of the 2019 Research in Adaptive and Convergent Systems, RACS 2019*, 161–168. doi: 10.1145/3338840.3355641

Dietterich, T. (1995). *Overfitting and Undercomputing in Machine Learning* (Tech. Rep.).

Directoraat-generaal Klimaat en Energie. (2019). *Klimaatakkord - C5 Elektriciteit* (Tech. Rep.).

Dukai, B., Peters, R., Vitalis, S., Van Liempt, J., & Stoter, J. (2021, 10). QUALITY ASSESSMENT of A NATIONWIDE DATA SET CONTAINING AUTOMATICALLY RECONSTRUCTED 3D BUILDING MODELS. In *International archives of the photogrammetry, remote sensing and spatial information sciences - isprs archives* (Vol. 46, pp. 17–24). International Society for Photogrammetry and Remote Sensing. doi: 10.5194/isprs-archives-XLVI-4-W4-2021-17-2021

Flach, P. (2016). ROC Analysis Motivation and Background.

Gooding, J., Crook, R., & Tomlin, A. S. (2015, 6). Modelling of roof geometries from low-resolution LiDAR data for city-scale solar energy applications using a neighbouring buildings method. *Applied Energy*, *148*, 93–104. doi: 10.1016/j.apenergy.2015.03.013

Graziano, M., & Gillingham, K. (2015, 7). Spatial patterns of solar photovoltaic system adoption: The influence of neighbors and the built environment. *Journal of Economic Geography*, *15*(4), 815–839. Retrieved from `https://academic.oup.com/joeg/article/15/4/815/2412599` doi: 10.1093/JEG/LBU036

Historie | AHN. (2022). Retrieved from `https://www.ahn.nl/historie`

Home - Sobolt. (2022). Retrieved from `https://www.sobolt.com/`

Induurzaam. (n.d.). *Zoninstraling en Oriëntatie.* Retrieved from `http://www.induurzaam.nl/2-energie-opwekken/zonnepanelen/zoninstraling-en-orientatie`

Kadaster. (2022). *Completer inzicht in locatie van zonnepanelen - Kadaster.nl zakelijk.* Retrieved from `https://www.kadaster.nl/nl/-/completer-inzicht-in-locatie-van-zonnepanelen?redirect=%2Fnl%2Fzakelijk%2Finformatie-per-sector%2Fstartpagina-deurwaarders`

Kanellopoulos, D., Kotsiantis, S., & Pintelas, P. (2006). Handling imbalanced datasets: A review Cite this paper Related papers Handling imbalanced datasets: A review. *International Transactions on Computer Science and Engineering*, *30*.

Kausika, B. B., Dolla, O., & Van Sark, W. G. (2017, 10). 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. doi: 10.1016/J.EGYPRO.2017.09.544

Khoshgoftaar, T. M., Golawala, M., & Van Hulse, J. (2007). An empirical study of learning from imbalanced data using random forest. In *Proceedings - international conference on tools with artificial intelligence, ictai* (Vol. 2, pp. 310–317). doi: 10.1109/ICTAI.2007.46

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*. Retrieved from `http://dx.doi.org/10.1016/j.eswa.2013.03.019` doi: 10.1016/j.eswa.2013.03.019

Kruppa, J., Ziegler, A., & König, I. R. (2012). Risk estimation and risk prediction using machine-learning methods. Retrieved from `http://www.cdc.gov/genomics/gtesting/ACCE/` doi: 10.1007/s00439-012-1194-y

Kwaliteitsbeschrijving | AHN. (2022). Retrieved from `https://www.ahn.nl/kwaliteitsbeschrijving`

Lan, H., Gou, Z., & Lu, Y. (2020). Machine learning approach to understand regional disparity of residential solar adoption in Australia. Retrieved from `https://doi.org/10.1016/j.rser.2020.110458` doi: 10.1016/j.rser.2020.110458

Lee, M., & Hong, T. (2019, 3). Hybrid agent-based modeling of rooftop solar photovoltaic adoption by integrating the geographic information system and data mining technique. *Energy Conversion and Management*, *183*, 266–279. doi: 10.1016/j.enconman.2018.12.096

Little, M. P., Rosenberg, P. S., & Arsham, A. (2022, 9). Alternative stopping rules to limit tree expansion for random forest models. *Scientific Reports 2022 12:1*, *12*(1), 1–6. Retrieved from `https://www.nature.com/articles/s41598-022-19281-7` doi: 10.1038/s41598-022-19281-7

McKnight, P. E., & Najab, J. (2010, 1). Mann-Whitney U Test. *The Corsini Encyclopedia of Psychology*, 1–1. Retrieved from `https://onlinelibrary.wiley.com/doi/full/10.1002/9780470479216.corpsy0524https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470479216.corpsy0524https://onlinelibrary.wiley.com/doi/10.1002/9780470479216.corpsy0524` doi: 10.1002/9780470479216.CORPSY0524

Mendieta, M. P., & Sarker, M. R. (2018). *Gradient Boosting Algorithm to Identify Markets for Residential Solar: New York State Case Study.*

Milanovic, S. M., Markovic, N. M., Pamučar, D., Gigovi´cgigovi´c, L., Kosti´c, P. K., & Milanovi´c, S. D. (2020). Forest Fire Probability Mapping in Eastern Serbia: Logistic Regression versus Random Forest Method. Retrieved from `https://dx.doi.org/10.3390/f12010005` doi: 10.3390/f12010005

Miles, J. (2014). *Tolerance and Variance Inflation Factor* (Tech. Rep.).

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2018). *Catalogus BAG 2018*. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

Mohajeri, N., Assouline, D., Guiboud, B., Bill, A., Gudmundsson, A., & Scartezzini, J.-L. (2017). A city-scale roof shape classification using machine learning for solar energy applications. Retrieved from `https://doi.org/10.1016/j.renene.2017.12.096` doi: 10.1016/j.renene.2017.12.096

NEF. (2022). *About NEF | NEF(New Energy Foundation).* Retrieved from `https://www.nef.or.jp/english/aboutnef.html/`

Nembrini, S., Konig, I. R., & Wright, M. N. (2018). *Supplementary material for "The revival of the Gini importance?"* (Tech. Rep.).

NEO. (2022). *Neo.nl*. Retrieved from `https://www.neo.nl/zonnepanelen-op-de-kaart/`

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. Retrieved from `http://jmlr.org/papers/v12/pedregosa11a.html`

Pedro Domingos. (1999). A general method for making classifiers cost-sensitive.

Probst, P., & Boulesteix, A.-L. (2017, 5). To tune or not to tune the number of trees in random forest? Retrieved from http://arxiv.org/abs/1705.05654

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019, 5). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3). doi: 10.1002/WIDM.1301

Qureshi, T. M., Ullah, K., & Arentsen, M. J. (2017, 10). Factors responsible for solar PV adoption at household level: A case of Lahore, Pakistan. *Renewable and Sustainable Energy Reviews*, *78*, 754–763. doi: 10.1016/J.RSER.2017.04.020

Rufibach, K. (2010). LETTERS TO THE EDITOR Use of Brier score to assess binary predictions. *Journal of Clinical Epidemiology*, *63*, 938–939. doi: 10.1016/j.jclinepi.2009.11.009

Schulte, E., Scheller, F., Sloot, D., & Bruckner, T. (2022). A meta-analysis of residential PV adoption: the important role of perceived benefits, intentions and antecedents in solar energy acceptance. *Energy Research & Social Science*, *84*, 102339. Retrieved from https://doi.org/10.1016/j.erss.2021.102339 doi: 10.1016/j.erss.2021.102339

Scikit-learn Developers. (2022). *Feature importances with a forest of trees — scikit-learn 1.1.1 documentation.* Retrieved from https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

Scikit-learn developers. (2022a). *MLPRegressor.* Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

Scikit-learn developers. (2022b). *RandomForestRegressor.* Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Sobolt. (2022a). *Sobolt.com.* Retrieved from https://www.sobolt.com/solarsearch/

Sobolt. (2022b). *Zonnedakje.* Retrieved from https://zonnedakje.nl/voor-professionals

Song, X., Huang, Y., Zhao, C., Liu, Y., Lu, Y., Chang, Y., & Yang, J. (2018, 11). An approach for estimating solar photovoltaic potential based on rooftop retrieval from remote sensing images. *Energies*, *11*(11). doi: 10.3390/en11113172

Taud, H., & Mas, J. (2018). Multilayer Perceptron (MLP)., 451–455. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-60801-3_27 doi: 10.1007/978-3-319-60801-3{_}27

Thom Opheikens. (2022). *Grote ambities zonne-energie, liefst met zonnepanelen 'made in Europe' | NOS.* Retrieved from https://nos.nl/artikel/2429351-grote-ambities-zonne-energie-liefst-met-zonnepanelen-made-in-europe

Van der Kam, M. J., Meelen, A. A., van Sark, W. G., & Alkemade, F. (2018, 12). Diffusion of solar photovoltaic systems and electric vehicles among Dutch consumers: Implications for the energy transition. *Energy Research & Social Science*, *46*, 68–85. doi: 10.1016/J.ERSS.2018.06.003

Vasseur, V., & Kemp, R. (2014). The adoption of PV in the Netherlands: A statistical analysis of adoption factors. Retrieved from http://dx.doi.org/10.1016/j.rser.2014.08.020 doi: 10.1016/j.rser.2014.08.020

Zhang, Y., Song, J., & Hamori, S. (2011). Impact of subsidy policies on diffusion of photovoltaic power generation. Retrieved from http://trendy.nikkeibp.co.jp/article/special/20090304/1024270/ doi: 10.1016/j.enpol.2011.01.021

A

Additional Data

A.1. Extra Tables

Table A.1: Multicolinearity test for three models.

	Variance Inflation Factor (VIF)		
	Full Model	Socioeconomic Model	Geometric Model
Residents 25-44	1.13	1.11	-
Address Density	1.57	1.44	-
Median Income	1.25	1.21	-
Owner Occupied Houses	1.40	1.40	-
Building Usage: Residence	2.89	1.07	-
Registered Addresses	1.20	1.05	-
Surrounding PV Ratio	1.10	1.08	-
Year of Construction	1.14	1.09	-
Building Height	3.50	-	2.06
Neighbouring Buildings	1.37	-	1.13
Suitable Area	1.29	-	1.21
Roof Complexity	1.90	-	1.74
Roof Flatness	2.02	-	1.84
Roof Area Per Surface	1.15	-	1.15
Irradiation	1.54	-	1.54

Table A.2: Results of the statistical Mann-Withney U test per feature.

Feature	Test Statistic U	P-value
Building Height	6.91E+09	0
Suitable Area	7.55E+09	0
Building Usage: Residence	7.69E+09	0
Roof Complexity	8.12E+09	0
Registered Addresses	8.42E+09	0
Roof Area Per Surface	9.17E+09	0
Surrounding PV Ratio	9.28E+09	0
Year of Construction	9.64E+09	0
Median Income	1.00E+10	0
Neighbouring Buildings	1.06E+10	6E-275
Owner Occupied Houses	1.08E+10	3.9E-157
Address Density	1.26E+10	1.8E-110
Residents 25-44	1.26E+10	3.4E-123
Roof Flatness	1.32E+10	0
Irradiation	1.37E+10	0

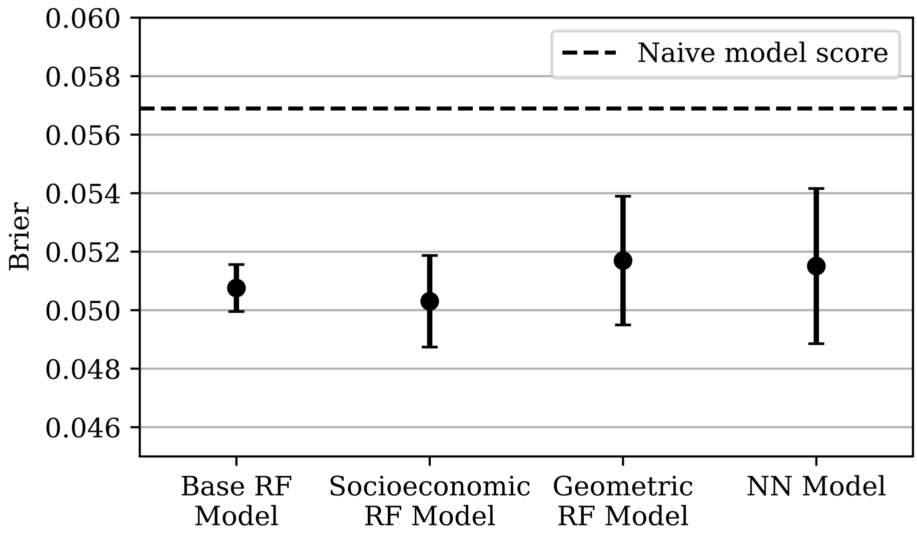


Figure A.1: Nested k-fold Brier score of the full model compared to the socioeconomic model, the geometric model and the NN model. The error bars indicate the standard deviation between the three outer folds

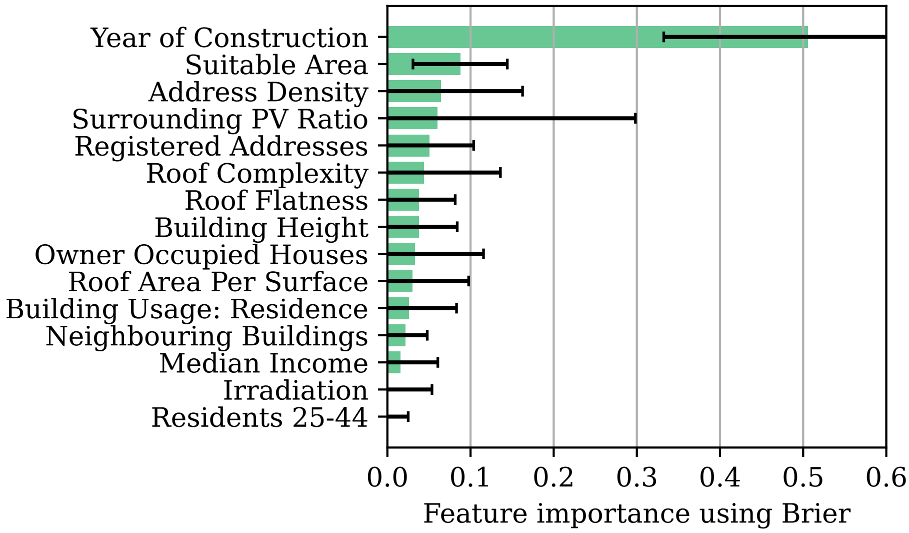


Figure A.2: Normalised feature importance using permutation importance, based on decrease in model AUC score.

Table A.3: Feature values for the nine largest outliers determined by LOF score.

Outlier number	1	2	3	4	5	6	7	8	9
Residents 25-44	0.21	0.21	0.21	0.21	0.17	0.29	0.21	0.13	0.40
Address Density	463	463	463	463	874	2476	463	884	874
Median Income	50	50	50	50	60	20	50	80	70
Owner Occupied Houses	2.3	2.3	2.3	2.3	4.0	0.7	2.3	1.3	4.0
Building Usage: Residence	0	0	0	0	1	0	0	0	1
Registered Addresses	0	0	0	0	1	0	0	0	1
Surrounding PV Ratio	0.08	0.03	0.03	0.08	0.06	0.01	0.03	0.03	0.06
Year of Construction	2000	2000	2005	1999	2009	2008	2003	2006	2006
Building Height	3.2	3.6	4.2	4.1	9.3	2.4	3.1	4.0	9.2
Neighbouring Buildings	1	0	0	0	2	0	0	0	2
Suitable Area	3	0	0	1	36	18	0	35	35
Roof Complexity	3	4	4	3	3	2	3	2	2
Roof Flatness	0.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	1.00
Roof Area Per Surface	4.7	4.7	5.3	5.8	50.6	28.9	4.4	50.1	56.6
Irradiation	0.80	0.76	0.69	0.61	0.87	0.87	0.82	0.87	0.87
LOF	54.4	50.3	48.4	43.2	32.2	31.3	31.0	27.9	25.4

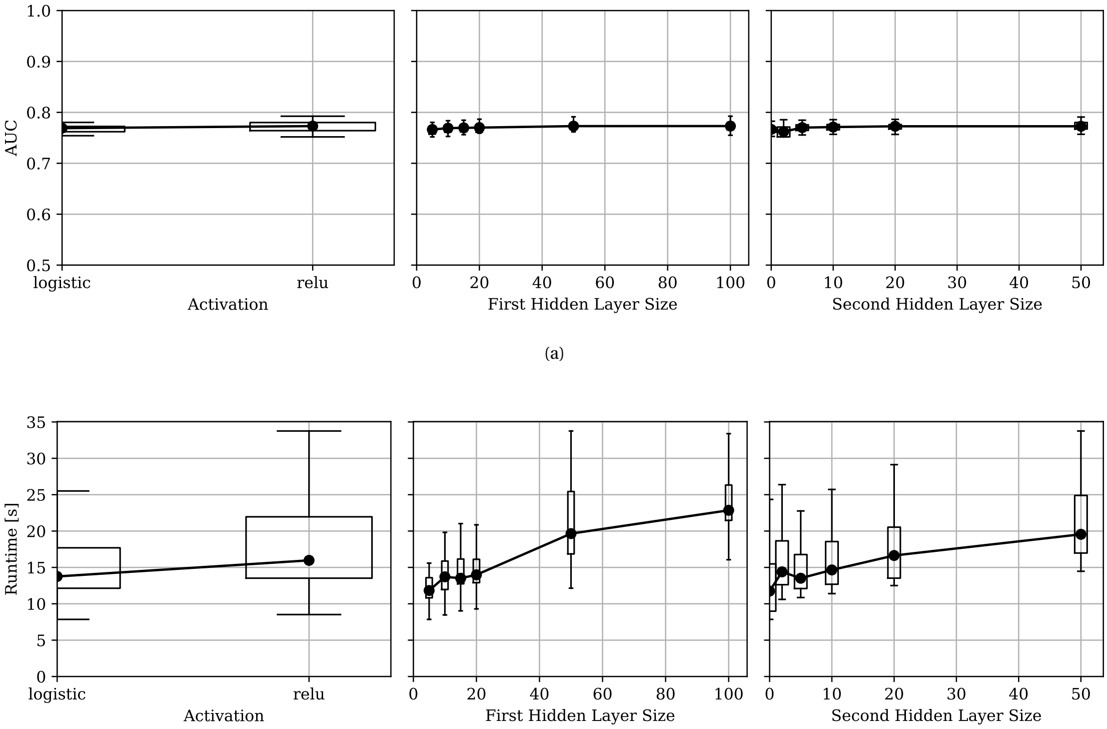
Table A.4: Characteristics of The worst 1% predictions With no PV installed.

	Full dataset		Largest Mistakes		Difference		
	Mean	Std.	Mean	Std.	Absolute	Relative	Norm. by Std.
Surrounding PV Ratio [-]	0.05	0.03	0.09	0.05	0.05	104.1%	1.72
Building Height [m]	6	3	9	2	3	46.5%	1.00
Building Usage: Residence [-]	0.52	0.50	0.97	0.16	0.45	87.4%	0.91
Year of Construction	1971.43	29.60	1998.27	18.53	26.84	1.4%	0.91
Roof Complexity [-]	3.78	1.71	4.80	1.56	1.02	27.0%	0.60
Median Income [%]	49.57	16.77	57.62	19.86	8.05	16.2%	0.48
Suitable Area [m^2]	94.53	547.89	262.51	2502.79	167.98	177.7%	0.31
Neighbouring Buildings [-]	0.91	0.89	1.14	0.87	0.23	25.4%	0.26
Residents 25-44 [-]	0.26	0.10	0.28	0.16	0.02	7.0%	0.18
Registered Addresses [-]	0.76	2.56	1.16	5.66	0.40	53.2%	0.16
Owner Occupied Houses [-]	2.18	1.63	2.39	1.55	0.21	9.6%	0.13
Roof Area Per Surface [m^2]	42.29	384.03	79.88	1485.51	37.59	88.9%	0.10
Irradiation [-]	0.81	0.07	0.81	0.08	-0.01	-1.0%	-0.12
Address Density [addresses/km^2]	1128.77	801.67	966.12	614.36	-162.64	-14.4%	-0.20
Roof Flatness [-]	0.41	0.43	0.28	0.29	-0.14	-33.4%	-0.32

Table A.5: Characteristics of The worst 1% predictions with PV installed.

	Full dataset		Largest Mistakes		Difference		
	Mean	Std.	Mean	Std.	Absolute	Relative	Norm. by Std.
Roof Flatness [-]	0.25	0.31	0.68	0.45	0.43	168.1%	1.40
Irradiation [-]	0.80	0.065	0.85	0.056	0.047	5.8%	0.72
Residents 25-44 [-]	0.25	0.10	0.26	0.10	0.01	4.9%	0.12
Address Density [addresses/km^2]	1013.16	689.17	1088.58	684.03	75.43	7.4%	0.11
Owner Occupied Houses [-]	2.32	1.55	2.26	1.57	-0.05	-2.4%	-0.04
Median Income [%]	53.88	17.14	51.69	14.96	-2.19	-4.1%	-0.13
Roof Area Per Surface [m^2]	47.15	228.69	17.48	10.59	-29.66	-62.9%	-0.13
Suitable Area [m^2]	131.81	655.16	16.40	12.33	-115.41	-87.6%	-0.18
Year of Construction	1980.04	24.77	1973.96	25.51	-6.08	-0.3%	-0.25
Registered Addresses [-]	1.01	2.52	0.12	0.43	-0.90	-88.6%	-0.36
Surrounding PV Ratio [-]	0.06	0.03	0.04	0.02	-0.01	-20.7%	-0.36
Neighbouring Buildings [-]	1.05	0.85	0.69	0.78	-0.36	-34.2%	-0.42
Roof Complexity [-]	4.59	1.51	2.56	0.91	-2.03	-44.2%	-1.34
Building Usage: Residence [-]	0.86	0.34	0.02	0.15	-0.84	-97.5%	-2.46
Building Height [m]	8.25	1.83	3.33	1.05	-4.92	-59.6%	-2.69

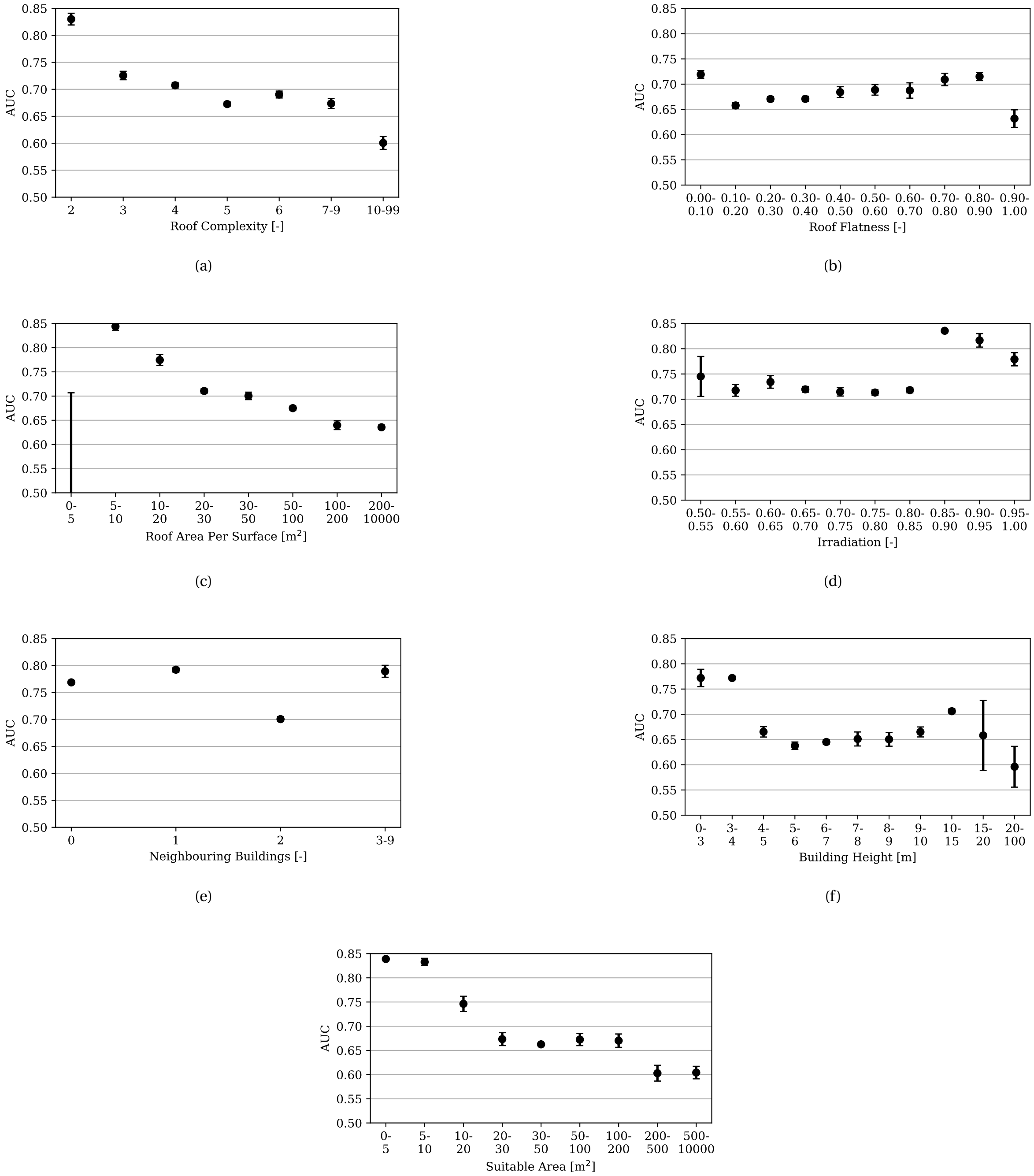
A.2. Additional Figures



(a)

(b)

Figure A.3: MLP model performance and runtime as a function of different hyperparameters.



(a)

(b)

(c)

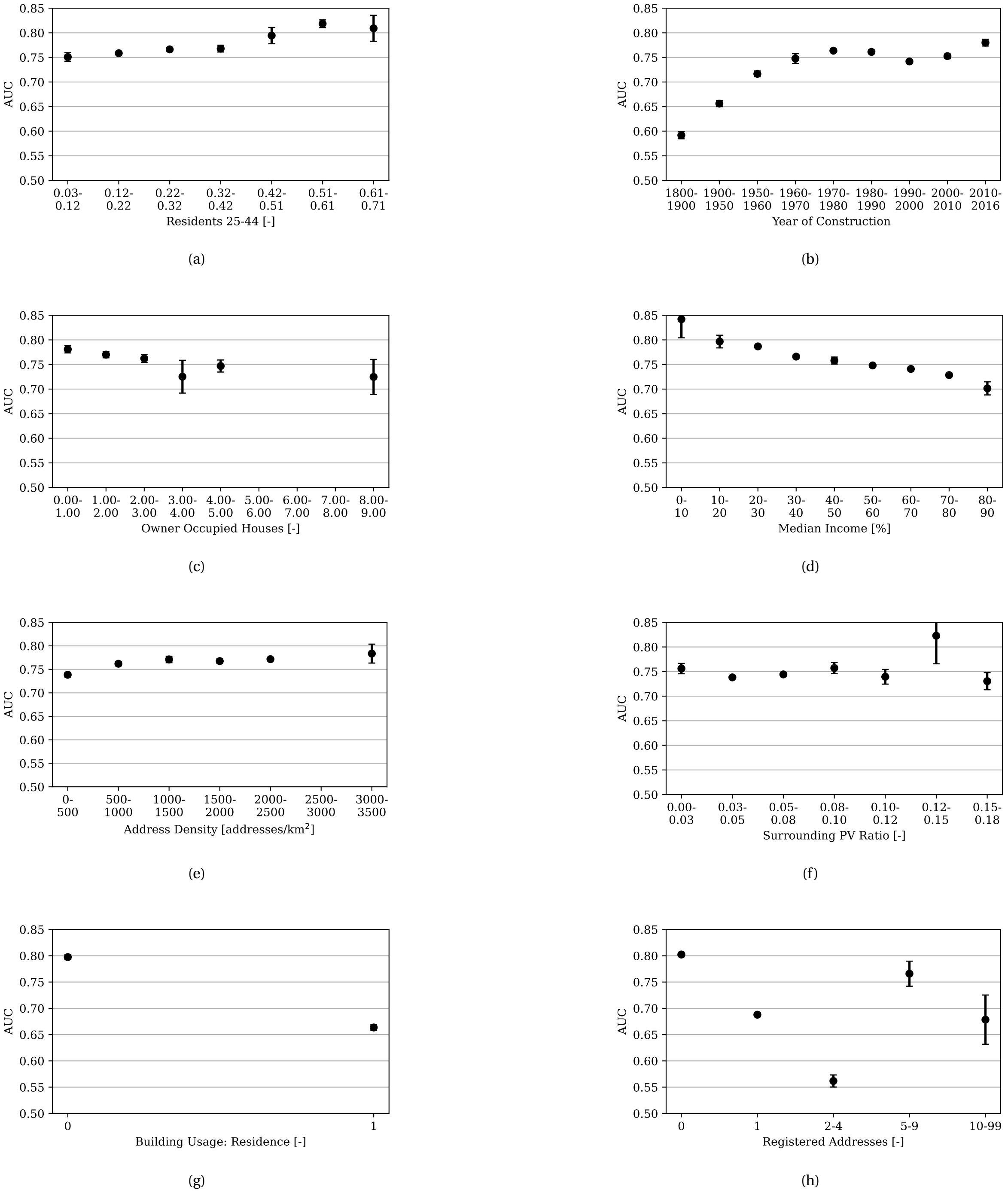
(d)

(e)

(f)

(g)

Figure A.4: Model performance as a function of different values of geometric features.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure A.5: Model performance as a function of different values of socioeconomic features.