Characterizing the Key Predictors of Renewable Energy Penetration for Sustainable and Resilient Communities

Jackson Bennett[1]; Aidan Baker[2]; Emily Johncox[3]; and Roshanak Nateghi, Ph.D.[4]

Abstract: With a mean annual growth rate of roughly 50%, the solar industry has experienced unprecedented growth in the last decade, largely owing to the steadily falling prices of solar installations. Utility-scale energy prices from solar installations are now comparable to all other forms of generation, and the cost of residential system installation has dropped on average by 70%, before incentives. However, the factors explaining this trend extend beyond falling prices. This paper presents a data-centric framework, grounded in machine-learning theory, to estimate solar installations as a function of social, economic, and demographic factors. By doing so, the authors seek to identify the key influencing factors of a community's adoption of renewable energy. To illustrate the applicability of the proposed data-centric framework, the state of California was selected as a case study. Results indicate that differences in population-adjusted adoption rates can be largely explained by variations in key factors such as income, race, political leaning, average electric power consumption, and solar radiation. By analyzing these differences, decision makers can devise effective incentive mechanisms to nudge homeowners toward improved access to renewable technology. **DOI: 10.1061/(ASCE)ME.1943-5479.0000767.**

Introduction

One of the most pervasive global environmental issues is the ongoing energy transition from fossil fuels to renewable resources. This topic has emerged in conversations at every level of society in recent years. As population and quality of life continue to increase in the United States and around the globe, the demand for energy is increasing worldwide. However, while demand is at an all-time high, traditional fossil fuel reserves are being steadily depleted. Before the emergence of extraction techniques such as hydraulic fracturing, petroleum production in the United States had been on a decline for roughly 40 years (Energy Information Administration 2019). Although these technologies have allowed production to increase, they pose serious threats to the environment, such as reduced air and water quality and increased risk of oil spills and earthquakes (Osborn et al. 2011). Looking forward, renewable energy is a natural response to the rising energy demand, which will preserve environmental quality while meeting energy needs.

This is a particularly relevant issue for civil and environmental engineering, a field that is intimately linked to the energy industry. Every energy transition requires new infrastructure, and it is largely the responsibility of civil, environmental, and construction engineers to develop that infrastructure (Hendrickson 2012). As society transitions to renewable energy, it is imperative to retrofit and adapt existing infrastructure and develop new infrastructure to overcome the challenges and leverage the opportunities that come with renewable energy. Of particular importance is energy storage systems and modular microgrid installations that can be easily taken online and offline when necessary. While these challenges are inherently interdisciplinary in nature, the construction of these projects and their successful integration into the existing infrastructure will largely fall in the domain of civil, environmental, and construction engineering (Hendrickson 2012).

Historically, civil and environmental engineering has been significantly impacted by changes in social values (Hendrickson 2012). This is particularly true in areas related to the environment, such as adjusting industry standards to meet new air and water pollution requirements and incorporating principles of sustainability into everyday design. Furthermore, compared with other engineering disciplines, projects in civil, environmental, and construction engineering are typically executed at a large temporal and spatial scale. Thus, these projects are often carried out under the public eye and generally require public engagement and consensus if they are to be ultimately sustainable (Pearce 2003). Consequently, it is critical to understand how different factors can influence public acceptance of civil infrastructure systems to enable effective project planning and implementation, and new developments in energy infrastructure will be no exception. With an increasing concern for global climate change and the adverse effects of the fossil fuel industry on public health, the push for a transition to renewable energy is complex in nature (Hendrickson 2012). Cities and states around the United States are already establishing requirements for renewable energy production.

Beyond sustainability concerns, the transition to renewable energy is a key example of the emerging emphasis on resilient socioenvironmental systems. This is especially true of small-scale solar energy systems that can be integrated into local microgrids (Chen et al. 2016). Rather than relying on a single generation and distribution system whose failures can affect thousands of people, distributed generation can prevent systemwide failure and allow

[1]Research Fellow, School of Industrial Engineering, Purdue Univ., 315 Grant St., West Lafayette, IN 47907 (corresponding author). ORCID: https://orcid.org/0000-0003-4875-930X. Email: benne105@purdue.edu; jsonben@gmail.com

[2]Undergraduate Researcher, School of Industrial Engineering, Purdue Univ., 315 Grant St., West Lafayette, IN 47907. ORCID: https://orcid.org/0000-0001-5965-8542. Email: baker339@purdue.edu

[3]Undergraduate Researcher, Division of Environmental and Ecological Engineering, Purdue Univ., 500 Central Dr., West Lafayette, IN 47907. Email: ejohncox@purdue.edu

[4]Assistant Professor, School of Industrial Engineering, Division of Environment and Ecological Engineering, Purdue Univ., 315 Grant St., West Lafayette, IN 47907. Email: rnateghi@purdue.edu

Note. This manuscript was submitted on June 3, 2019; approved on October 28, 2019; published online on March 19, 2020. Discussion period open until August 19, 2020; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Management in Engineering***, © ASCE, ISSN 0742-597X.

individuals to support each other in the event of a failure as its effects are highly localized (Chen et al. 2016). By incentivizing individuals and homeowners to install small-scale distributed renewable energy systems, energy resilience within a community can be significantly enhanced. Of the variety of renewable energy options available, solar energy is one of the most accessible forms in the residential sector because of its variability in scale. With an average annual growth rate of roughly 50% in the last decade, solar technology is becoming increasingly efficient and widespread (Fu et al. 2016). Improving current understanding of the factors that drive residential solar installations will provide insight into how the technology can be best incentivized in the transition to increasingly sustainable and resilient urban systems. In particular, it is critical for decision makers and urban planners to consider the role of social and environmental factors in motivating individuals and families to install residential solar systems.

This analysis proposes a data-centric and generalizable framework to identify the key influencing factors of solar roof adoption. Contrary to previous approaches, the authors go beyond explanatory modeling with linear model constructs, and aim to develop a rigorously validated and transferable paradigm for modeling residential solar installations. The proposed data-centric framework is leveraged to identify the key predictors of residential solar roof installations, using the principles of statistical learning theory. To demonstrate the applicability of the proposed framework, the state of California is used as a case study. California is particularly well-suited to analysis as the state has the most solar installations in the United States and has gone to great lengths to make data on solar installations publicly available.

This paper continues with a review of existing literature in the "Literature Review" section. In the "Input Data" section, a description of the input data and response variable is presented. The "Methodology" section outlines the model selection and interpretation process, as well as the statistical foundation of the final model, extreme gradient boosting. The major findings from the study are presented in the "Results" section, including key predictors and interpretations of their relationship with the response variable. Finally, the paper concludes with a discussion of the implications of these results for enhancing the resilience of urban sociotechnical systems in the "Discussion" section and with several concluding thoughts in the "Summary and Conclusion" section.

Literature Review

Previous literature analyzing sustainable energy technologies has identified public acceptance as a key factor in attaining energy goals (Liu et al. 2018). This is especially true for technologies that are installed at the individual or home level, such as residential solar systems. Identifying the key factors that motivate public participation in sustainable technology adoption allows researchers and policy makers alike to cultivate increased levels of adoption and support (Liu et al. 2018). Small-scale solar systems make for a particularly interesting study as they are highly modular and have recently begun to attract more attention because of their potential applications in intelligent energy systems such as smart grids, buildings, and cities (Hastak and Koo 2017). These systems can be used to improve building sustainability under Leadership in Energy and Environmental Design (LEED) guidelines, which have attracted increased attention among the public and construction agencies alike (Abdallah and El-Rayes 2016). Furthermore, solar technology has been shown to be a viable option under a variety of future climate scenarios, as opposed to other renewable solutions, such as hydroplants, whose expected performances vary drastically under different climate futures (Kim et al. 2017).

Understanding the strategies available to achieve sustainability goals is critical for the managers of civil infrastructure projects in a time where support for renewable energy projects is at a record high. Studies in other countries have identified policy as a major factor in both encouraging and limiting growth in the solar industry (Chapman et al. 2016). Frequent policy changes in Australia were largely responsible for the erratic growth the country has seen in residential solar capacity. A different study in Malaysia also identified government incentives as being highly influential in encouraging residential solar installations (Muhammad-Sukki et al. 2011). The analysis also determined that a major limiting factor in growth of the industry was a lack of public awareness of available incentives; because homeowners did not realize that there was government support available for residential solar systems, they believed that the technology was not affordable. By studying the impacts and shortcomings of policies aimed at civil infrastructure projects, managers of such projects can improve their understanding of how to most effectively win public support and focus on meeting sustainable energy targets.

As the residential solar industry has only recently become competitive with traditional methods of power generation, the body of knowledge explaining the trends in this sector is nascent. Previous work is primarily grounded in the social sciences (based on survey data at the homeowner level) to understand what motivates people to install residential solar systems (Chen 2014; Schelly 2014). More recently, studies have used statistical techniques to identify the relationship between select demographic variables and the presence of rooftop solar installations (Kwan 2012; Sunter et al. 2019, 2018). The overarching goal of these statistical analyses is to explain trends in solar installations and to understand how a variety of factors account for the current state of the industry. A more detailed account of the existing literature is outlined below.

In one of the earlier studies on the subject, Chen (2014) found that among college students, environmental value was positively associated with the intention to install solar power systems. Environmental value was assessed using the six-item green consumer value scale, which was strongly related to actual green behavior. Intention to install solar power systems was assessed by four questions on a 5-point Likert scale. Based on the survey results, Chen (2014) also found that customer innovativeness is associated with intention to install. It is important to note that this study analyzed *intention* to install rather than actual solar installations. Because of constraints such as limited budget, living situation, or other life factors, it is possible that those who intend to install will not do so for a long time.

In a subsequent study, Schelly (2014) surveyed 48 early adopters of solar technology in 36 households across the state of Wisconsin. Respondents were gathered via mail and email recruitment and represented both urban and rural communities across the state. The authors found that environmental motivations were neither a necessary nor a sufficient factor in installing solar technology. Of all participants, 40% did not identify environmental values as a factor in their decision to install solar energy systems, and no participant identified environmentalism as their sole motivating factor. For many of the participants surveyed, perception of future energy savings was one of the largest motivating factors. Schelly (2014) found that even though solar technology is often identified as a green choice and is associated with political liberalism, many of the survey participants identified as conservative. For this subset of respondents, many of whom negatively regarded concerns about climate change and global warming, the financial savings were the most important factor in motivating their decision.

Many of those surveyed viewed solar technology as a means of significantly reducing their utility expenses, and considered it a smart financial investment. A select few identified religion as a reason for their decision; they saw solar installations as an opportunity to be good stewards of the planet they inhabit. Additionally, upfront discounts (tax credits and rebates) served as stronger motivation for adoption than a short payback period. A surprising connection between those surveyed was an interest in do-it-yourself projects and energy technology. The majority of homeowners designed their own homes, and everyone surveyed was using some sort of alternative technology. These findings very much agree with the survey data collected by Chen (2014) that identified innovative inclinations as a strong indicator of intent to install. Because of the local nature of this case study and the somewhat unique population surveyed, it is difficult to generalize these findings to areas with a developed solar industry and identify what drives installations among the general population. Another survey-based study in California focused on specific events that motivate people to install solar technology (Rai et al. 2016). The analysis identified direct marketing by a solar company, planning for retirement, and an increase in electricity rates as the top three "spark events," or events that motivated homeowners to install residential solar technology. Similar to previous work, Rai et al. (2016) found that perceived financial savings were a significant motivating factor in homeowners' decisions.

To better generalize trends in solar technology adoption, an increasing number of studies based on statistical techniques have begun to emerge. Statistical techniques have been proven to be useful in management applications in recent years (Yi and Chan 2015; Alipour et al. 2019; Mukherjee and Nateghi 2019; Bruss et al. 2019; Raymond et al. 2018; Qiao et al. 2018). By leveraging the increasingly large amount of data available to researchers, analyses can be conducted to deliver novel insights to managers of civil infrastructure projects. One of the earliest efforts in leveraging data-driven methods to identify factors that motivate solar installations was by Kwan (2012). The study used zip code–level data from the 2000 US Census to explore the impact of social, political, environmental, and economic factors on the distribution of residential solar energy systems (Kwan 2012). Corresponding solar instillation data at a census tract–level was collected from the National Renewable Energy Lab's (NREL's) Open PV (Photovoltaic) Project. The goal of Kwan's analysis was to predict the percentage of housing units with solar system installations by zip code. This was done using a zero-inflated negative binomial regression model. The authors found that the percentage of solar installations was positively influenced by the amount of solar radiation, cost of electricity, amount of financial incentives, median home value, proportion of the population with incomes between $25,000 and $100,000, proportion of the population with a college education, proportion of the population that is White or Hispanic Latino, and proportion of the population that is registered Democrats. The most important variable based on this study was solar radiation. Variables that negatively affected the percentage of solar installations were proportion of the population in age groups 25–34 or 55–64, proportion of the population that is Black or Asian, housing density, and classification as a suburban area. Beyond negative and positive associations, details about the relationship between predictors and the response were not provided. The study also fell short of assessing the predictive power of the developed model and focused only on the model's goodness of fit. While such an explanatory model can provide useful insights about the past variability in the data, it cannot be used for predictions as the model cannot be generalized beyond the data used in the analysis (Shmueli 2010).

A more recent study by Sunter et al. (2019) analyzed the impacts of race and ethnicity on solar energy adoption. Solar installation data were collected from Google's Project Sunroof, and demographic information from the American Community Survey. After controlling for differences in household income and home ownership, the authors found that there was a significant disparity in solar installations between White- and Black-majority census tracts. A similar disparity was observed between White- and Hispanic-majority census tracts. These relationships were determined using locally weighted scatterplot smoothing (LOWESS). One identified reason for this difference is the lack of initial deployment of solar technology in Black and Hispanic-majority tracts. Sunter et al. (2018) also hypothesized that the lack of racial diversity in the renewable energy workforce could explain the lack of diffusion to certain communities. In an earlier study, Sunter et al. (2018) used similar methods applied to Democrat versus Republican-majority census tracts. They found that even after controlling for income, Republican-majority census tracts were more likely to install solar energy systems. The research team had a number of hypotheses for this result, but were unable to test any of them with their available data (Sunter et al. 2018). By including a variety of factors in the same model, this study will provide information on the relative importance of the factors identified in previous work in addition to exploring additional variables that may explain the relationship between party affiliation or race and solar installations.

To improve upon previous approaches, the authors went beyond simple explanatory modeling approaches—typically with linear architectures—that primarily focus on explaining the variability in past data. Instead, the authors aimed to develop a rigorously validated and generalizable paradigm for modeling residential solar installations at a systems level and identify its key predictors, using the principles of statistical learning theory. The framework incorporated data about the solar installations (i.e., cost, electricity output, and economic incentive received) as well as a wide variety of demographic, environmental, and social data. The predictive performance of a range of parametric and nonparametric models were then compared to identify the best-performing predictive model. Specifically, the predictive performance of multiple linear regression, regularized linear regression (RLR), generalized additive models (GAM), multivariate adaptive regression splines (MARS), random forest (RF), support vector machines (SVM), and extreme gradient boosting (XGBoost) was tested using the out-of-sample accuracy estimates. The most influential predictors were then identified using the "information gain" metric. Moreover, model inferencing was conducted through plotting the partial dependence between the response variable and the most influential variables.

Input Data

The variables considered for model development were composed of social, economic, and environmental measures as well as specifications on the solar installations. These particular classes of variables were selected as they have been identified in previous studies as factors that motivate or influence a homeowner's decision to install solar technology. Furthermore, the authors were interested in identifying factors beyond economic indicators that influence a homeowner's installation decision (Schelly 2014; Sunter et al. 2019). Unless otherwise stated, all variables are averaged at the zip code level. Data was collected at the zip code level as it is the most granular level offered by the US Census, which was the primary source of data in this analysis. This granularity is important because it allows the study to capture heterogeneity, which would otherwise be masked by data aggregated at a coarser spatial

scale. Table 1 summarizes all of the input data described in the ensuing paragraphs.

Social Data

The social factors incorporated into the model included race breakdown, education level, household size, median age, 2016 election results, Rural Urban Commuting Area (RUCA) rating, and employed population. The results from the 2016 election were based on data collected from Politico, a political journalism company, at the county level (Politico 2016). All zip codes within a county were assumed to have the same voting distribution as the county itself. RUCA rating was collected from the USDA Economic Research Service (USDA 2010). Although there are many measures of an area's urban-ness, RUCA is recommended when analysis is carried out at the zip code level (Washington State Dept. of Health 2008). All other information was collected from the 2016 5-year US Census (US Census Bureau 2016). To reduce the number of variables, education was binned into four categories: those without a high school diploma, those with a high school diploma or some level of college, those with an associate's degree, and those with a bachelor's degree or higher. All variables other than median age and household size were represented as proportions to ensure that information on population was not indirectly present in the data set.

Economic Data

The economic factors considered were median income, solar installation cost, and incentive amount. Similar to the social factors, median income was collected from the 2016 5-year US Census (US Census Bureau 2016). Price and incentive information was collected from the California Solar Initiative Working Dataset (CSI 2018). This is a comprehensive data set that includes information on every solar system installed in California since 2007 that received a cash incentive from the California Solar Initiative (CSI). This is the largest incentive program in the state and collaborates with every major utility company in California. All residents of the state have access to this program regardless of geographic area, and the sample it provides is assumed to be representative. Information from this data set was aggregated at the zip code level to yield the number of solar installations, the cost of system installation, and the average incentive amount received through the CSI program. Rather than predicting the cost and incentive amount, the proportion of costs covered was used as a predictor variable to avoid confounding with the capacity of the installed system.

Environmental Data

Environmental factors included in the model are limited to solar radiation. Solar radiation data were collected from the NREL's National Solar Radiation Database (NSRDB) (NREL 2016). Users can interface with the NSRDB via a simple application programming interface (API) that allows for the specification of a location from which to pull data. Data are available at a 4×4-km resolution and are returned as a set of hourly solar radiation values over the course of a year. For every zip code in the data set, solar radiation data were pulled from 2016 and the average value over the course of the year was recorded. For roughly 10% of the zip codes in the data set, solar radiation data were unavailable. To impute the missing data, the distance-weighted k-nearest-neighbors algorithm was leveraged using the available data (Dudani 1976).

Table 1. Summary of input data

Data	Unit	Source
Number of projects	N/A	CSI (2018)
Incentive application year	N/A	
Solar system output (AC)	kW	
Solar system output (DC)	kW	
Incentive amount	Dollars	
System cost	Dollars	
Cost covered[a]	%	
Median annual income	Dollars/year	American Community Survey 5-year estimates (US Census 2016)
Median age	Years	
Total households	Houses	
Population	People	
Proportion of population employed	%	
Average house size[a]	People per house	
Proportion of population without a high school diploma[a]	%	
Proportion of population with some high school or college education[a]	%	
Proportion of population with an associate's degree[a]	%	
Proportion of population with a bachelor's degree or higher[a]	%	
Proportion of White population	%	
Proportion of Black population	%	
Proportion of Asian population	%	
Proportion of Hispanic population	%	
Proportion of Pacific Islander	%	
Proportion of population with two or more races	%	
Proportion of other race	%	
Trump votes (2016 presidential election)	%	Politico (2016)
Clinton votes (2016 presidential election)	%	
Solar irradiance	$kW \cdot h/m^2$	NREL (2016)
Monthly electricity consumption	$kW \cdot h$/month	Southern California Edison (2016), San Diego Gas & Electric (2016), and Pacific Gas and Electric (2016)

[a]Derived variable.

Other Data

In addition to the aforementioned variables, average monthly electricity consumption, the solar installations output capacity, and installation year were also considered by the model. Electricity consumption data were collected from the three major utilities in California: Pacific Gas and Electric (2016), San Diego Gas & Electric (2016), and Southern California Edison (2016), which collectively serve 84% of the state's population. All data are publicly available at the zip code level on a monthly basis for all types of customers (agricultural, residential, industrial, and commercial) following California Public Utilities Commission Decision 14-05-016. Residential consumption data were collected from the three aforementioned companies and aggregated at the monthly level. Solar output capacity was available in the California Solar Initiative Working Dataset and was aggregated in a similar way to system cost and incentive amount. Installation year was also taken from this data set and aggregated by mode within zip codes.

Response Variable

The purpose of this analysis was to predict a community's willingness to install solar projects. As willingness is difficult to quantify, the authors used the number of residential solar installations in a given area as a proxy. Initial exploration revealed that there was a strong linear correlation between population and the number of projects installed, which makes intuitive sense. However, as the goal of this analysis was to isolate the effects of social, economic, and environmental factors, the number of solar installations was scaled by the population to give a final response variable of

$$y_i = \frac{\text{Project count}_i}{\text{Population}_i} \tag{1}$$

The authors also considered the number of solar installations scaled by the number of housing units as a potential response variable, but based on preliminary results, and determined that scaling installations by population was more effective at detecting signals in the data and improved model performance by close to 3%.

Methodology

The goal of this analysis was to develop a validated and interpretable predictive model that can be used to understand how different socioeconomic factors affect a community's likelihood to increase its residential solar capacity. Considering the trade-off between model complexity and interpretability, the authors elected for models that lend themselves more easily to characterizing the relationship between predictors and response. Specifically, no deep-learning approaches were used for this analysis. Although deep learning can be advantageous for capturing the underlying relationship between variables and response, the results are difficult to interpret (because of the various transformations of the independent variables in the learning process) and would be of little help in explaining the impact different environmental, social, and economic factors have on residential solar installations. The modeling framework is outlined in Fig. 1.

In this analysis, seven types of statistical learning algorithms were considered. These algorithms were selected as they represent distinct classes of supervised machine learning, namely, parametric, semiparametric, and nonparametric methods (Hastie et al. 2005). The included models ranged from simple approaches such as multiple linear regression to recent advances in the field of machine learning such as extreme gradient boosting. Previous studies, particularly in the field of civil infrastructure management, have established comparative assessment of these models as an effective tool for assessing the trade-offs between model interpretability and generalization to identify models most suitable for supporting decision-making (Lokhandwala and Nateghi 2018; Obringer and Nateghi 2018; Mukhopadhyay and Nateghi 2017; Nateghi et al. 2011, 2016). The full list of models considered is as follows: multiple linear regression, ridge regression and lasso regression (Friedman et al. 2010), generalized additive models (Hastie and Tibshirani 1986), multivariate adaptive regression splines (Friedman 1991), support vector machine (Drucker et al. 1997), random forest (Breiman 2001), and extreme gradient boosting (Chen and Guestrin 2016). A brief description of the models considered can be found in Table 2.

Model Assessment and Selection

To assess the predictive performance of the models and select the model with the highest predictive power, the authors used a k-fold cross-validation scheme (Kohavi 1995). In this approach, data are randomly segmented into k mutually exclusive partitions (folds) of approximately equal sizes denoted by $S_1, S_2, \ldots, S_k$. For element i of $\{1, 2, \ldots, k\}$, S_i is withheld from the data used to train the model. The model is then tested on S_i, and a measure of its predictive accuracy is recorded. The reported performance metric is the mean value across all k folds. In this study, the metric used was root-mean-square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\text{Predicted}_i - \text{Actual}_i)^2}{N}} \tag{2}$$

To select the model that best predicted solar installations in a zip code, this analysis employs fivefold cross validation. Additionally, each model was compared to the null (mean only) model as a benchmark for the skill of statistical regression models. The null model does not rely on leveraging a statistical model (based on supervised learning) to estimate the conditional mean of the response variable as a function of the independent variables; instead it assumes that reasonable predictions can be achieved by simply calculating the historical average value of the response variable as follows:

$$\hat{y}_i = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{3}$$

While the null model is not useful for deriving insights on the relationships between predictors and response, it serves as a baseline for assessing the effectiveness of statistical models. If a similar predictive power can be derived from the null model and a candidate model, it can be concluded that the candidate model performs poorly and is of little value for drawing statistical inferences.

Statistical Inferencing

One of the most powerful features of any statistical model, particularly for predictive applications, is its ability to offer useful inferences. Although the theoretical foundations for statistical learning models vary, the approach to developing useful insights follows the same basic procedure and can generally be separated into two steps: (1) identifying key predictors; and (2) characterizing the relationship with the response. Notably, as the model increases in complexity, so too do the evaluation metric for variable (predictor) importance and the tools used to characterize their relationship with the response.

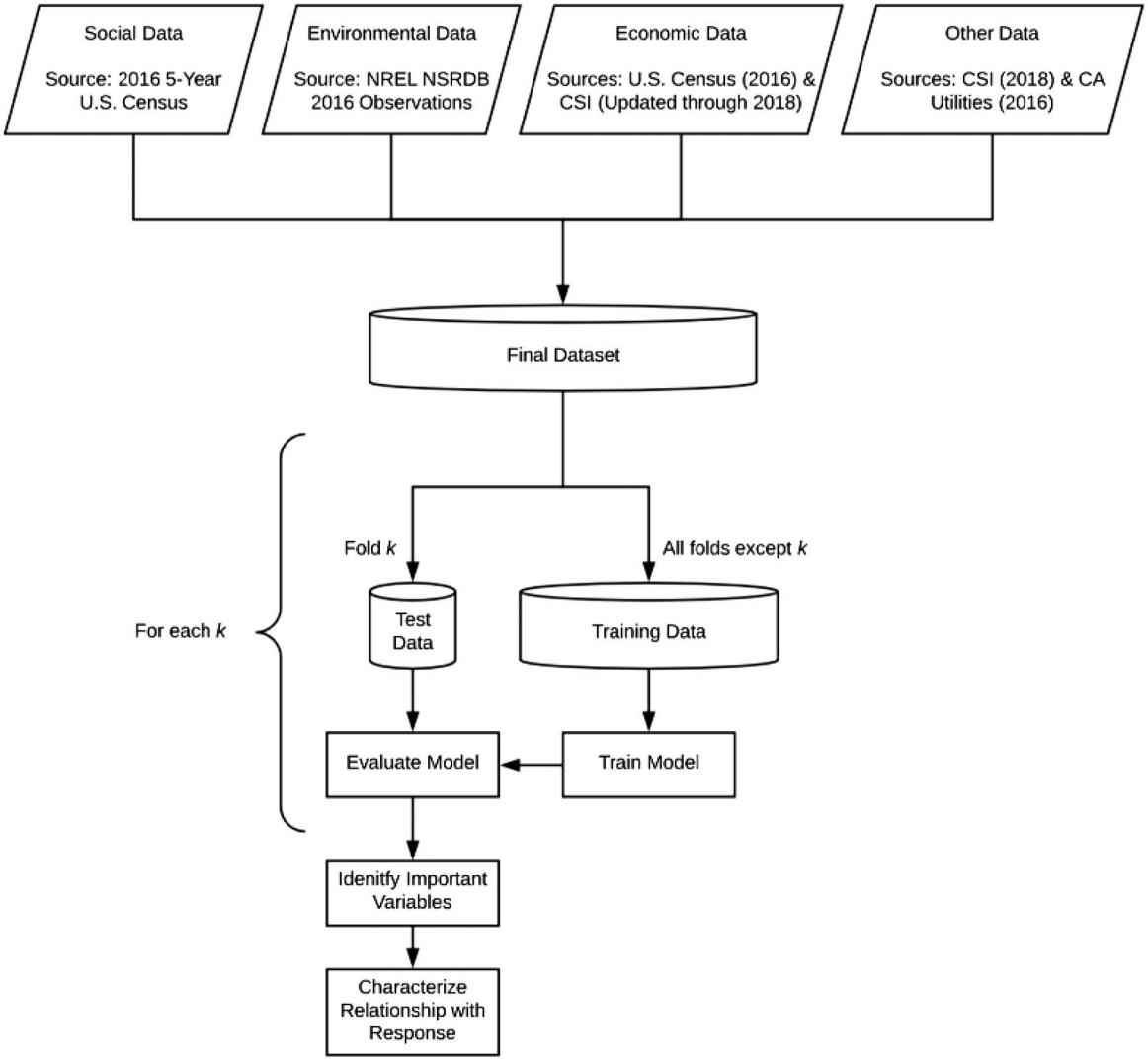

Fig. 1. General modeling framework from data set compilation to model development, testing and validation, and inferencing. CA Utilities (2016) includes Pacific Gas & Electric (2016), San Diego Gas & Electric (2016), and Southern California Edison (2016).

Table 2. Summary of models

Model	Overview
Multiple linear regression	An extension of a simple linear regression that allows for multiple predictor variables. There are several assumptions associated with this model including the linearity of the relationship between the predictor variables and the response, as well as independence and homoskedasticity of the normally distributed residuals.
Ridge and lasso regressions	Shrinkage methods that are based on the linear model. The difference is that the model scoring function (residual sum of squares) includes an additional penalty that rewards simple models to reduce overfitting. Lasso regression uses L1 (absolute magnitude) regularization, while ridge regression employs L2 (squared magnitude) regularization.
Generalized additive models	An extension of the multiple linear regression model that applies smooth, nonlinear functions to each predictor rather than a linear coefficient. Predictions are derived by summing the contributions of each function.
Multivariate adaptive regression splines	An algorithm that essentially generates a set of piecewise linear functions, or splines. Splines are iteratively added to $f(X)$ in the way that yields the most significant decrease in training error. Once the model has been trained, it is generally overfit, so the MARS algorithm employs a pruning technique using generalized cross validation.
Support vector machines	This algorithm uses a kernel function to transform the data being modeled to a higher dimension where it can be modeled using a linear function. The transformation is done in such a way that all data are within ϵ (an error term) of the linear function.
Extreme gradient boosting	An in-depth explanation is provided at the end of the "Methodology" section.
Random forest	An ensemble tree-based method based on bootstrap aggregation, or bagging. The algorithm generates a high number of decision trees using a training set collected by bootstrap sampling. To reduce correlation between trees, each tree is built with only a subset of the predictors in the data set. Final predictions are made by aggregating predictions across all trees.

In statistical model inferencing, there exists a significant distinction between parametric and nonparametric models. Parametric models make assumptions about the distribution of the response variable, as well as the structure of the underlying relationship between predictors and response. Inferencing with such a model can be as simple as evaluating the statistical significance of each variable and interpreting the magnitude and sign of the coefficients in the model. Nonparametric models make no assumptions about distributions or the nature of the response–predictors relationship and are typically more data intensive and complex. Variables are selected based on their importance, which is determined by their contribution to out-of-sample predictive accuracy. To characterize the relationship between predictor and response, one of the most common techniques is plotting partial dependencies. These plots show the effect a predictor variable on the response while the effects of the other variables in the model are accounted for, essentially only allowing the predictor of interest to vary (Friedman 2001). Mathematically, the relationship of the predictor on the response is given as

$$\hat{f}_j(x_j) = 1/n \sum_{i=1}^{n} \hat{f}_j(x_j, x_{-j,i}) \tag{4}$$

where $\hat{f}$ represents the model; n = number of observations in the training set; and x_{-j} = all variables other than x_j in the training set.

Extreme Gradient Boosting

This section outlines the theoretical underpinning of the extreme gradient boosting (XGBoost) algorithm, which outperformed all the other models based on the results summarized in Table 3. Gradient boosting is a machine-learning technique that is used for classification and regression. XGBoost creates a network of decision trees using regression models to optimize an objective function (Chen and Guestrin 2016). While traditional optimization models use Euclidean methods for solving problems, gradient tree boosting creates models in an additive manner. The method aims to minimize the loss function at step t as follows:

$$L^t = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2 f_t(x_i) \right] + \Omega(f_t) \tag{5}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} \omega(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} \omega(y_i, \hat{y}^{(t-1)})$.

The function $f(t)$ is greedily added to the model in a way such that f_t most benefits the model. Each f_t is an independent tree structure at instance t. A second-order approximation is used to accelerate the optimization of the objective. The functions g_i and h_i are first-order and second-order approximations of $\omega(y_i, \hat{y}^{(t-1)})$, respectively. The term ω measures the difference between the prediction and the target (Chen and Guestrin 2016). The final term, $\Omega(f_t)$, is a penalty given to the model for overcomplexity. A weight to each leaf of each tree is given by

$$w_j^* = -\frac{\sum_{i \epsilon I_j} g_i}{\sum_{i \epsilon I_j} h_i + \lambda} \tag{6}$$

The model uses a scoring function for the impurity to evaluate a given tree structure.

Extreme gradient boosting is used over gradient boosting because it is faster and less computationally complex. This is accomplished using shrinkage and column subsampling. Shrinkage scales the weights of each added tree after each step as the model becomes more complex. This decreases the influence of previous trees in favor of possible, future trees to improve the overall model. Column subsampling prevents overfitting and speeds up computations of the parallel algorithm.

Table 3. Out-of-sample RMSE, based on fivefold cross validation, as well as the percentage improvement in accuracy over the null model

Model	RMSE	Improvement over null model (%)
Null (i.e., mean only) model	775.2	N/A
Multiple linear regression	487.4	37.2
Ridge regression	491.7	36.6
Lasso regression	486.1	37.3
Generalized additive models	487.3	37.2
Multivariate adaptive regression splines	438.0	43.5
Support vector machines	408.8	47.3
Extreme gradient boosting	**370.8**	**52.2**
Random forest	399.6	48.5

Note: Bold value indicates that the model has minimum RMSE value.

Results

Model Assessment and Selection

Table 3 summarizes the out-of-sample errors resulting from the candidate models tested including multiple linear regression, ridge regression, lasso regression, generalized additive models, multivariate adaptive regression splines, support vector machines, extreme gradient boosting, and random forest. The table also includes the percentage improvement of each of the models over the null (mean only) model. This benchmark model makes predictions by calculating the average value of the response variable and serves as a baseline of comparison for the other models tested in this analysis.

It is evident from the results that the model based on the extreme gradient boosting algorithm outperformed all other models, with an RMSE of 370.8 and an improvement over the null model of 53%. A plot displaying the fit of the final model can be found in Fig. 2.

Model Inferencing

Variable Importance

The most important variables in the model were identified using the information gain criterion. This is a metric that measures the information gained about the response variable (Chen and Guestrin 2016). In other words, it captures information about the quality of the split in the tree and includes a regularization term that penalizes complexity. Variables with a higher gain are more important in the model. An important note about information gain is that the absolute value of the metric has little meaning. Rather, the comparative values within the training data are what allows one to identify the most important variables and compare their relative importance. Gain values for the ten most important variables after model tuning and variable selection are presented in Table 4.

Partial Dependence Plots

In this study, the top ten most important variables were identified, and their partial dependence plots are presented in Fig. 3. These plots show the marginal effect of a particular variable on the response, denoted as "yhat" in the plots. When reading these plots, it is important to note that the y-axis displays the transformed response variable rather than the number of solar projects. The scale of this axis differs considerably between plots, so it is important to observe the range encompassed. In general, this range decreases with variable importance.

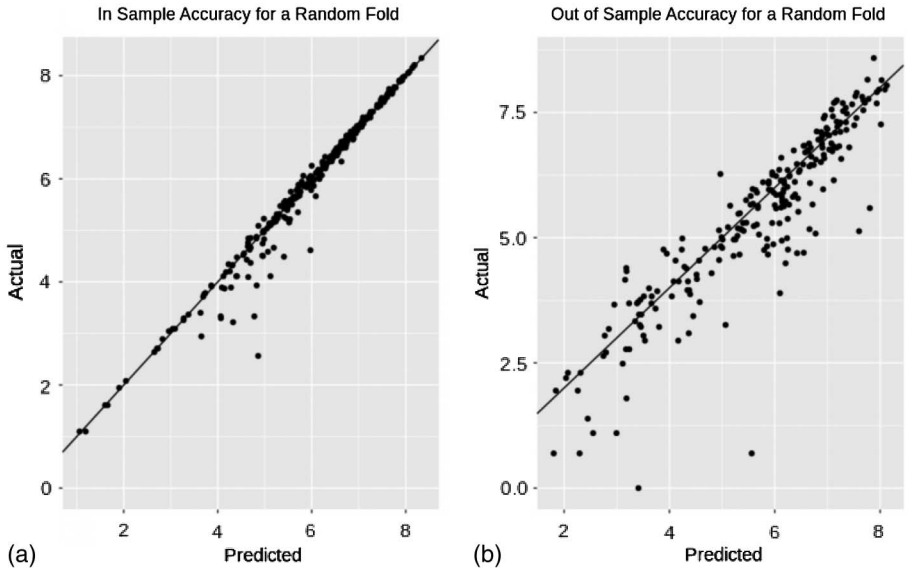

Fig. 2. Model predictions compared to actual observations: (a) in-sample estimates; and (b) out-of-sample estimates. Plots use a log-log scale for clearer visualization. Both plots have a 1-1 line displayed for reference.

Table 4. Information gain for the ten most important predictor variables in the final best model

Variable	Information gain (%)
Average electricity consumption	9.16
Median income	7.66
Median age	6.77
Proportion of White population	5.81
Proportion of Clinton votes	5.57
Cost covered	5.48
Solar radiation	4.21
System output (AC)	4.13
Proportion of Trump votes	4.10
Proportion of population with an associate's degree	3.62

Based on the partial dependence plots, average electricity consumption, median income, median age, White proportion of the population, solar radiation, proportion of the population with an associate's degree, and proportion of the population that voted Republican in 2016 all positively affect a community's willingness to install solar systems. The proportion of the population that voted Democratic in 2016 and solar system output are both inversely related to a community's willingness to install solar, based on the results of the model. The percent of system cost covered by incentives exhibits peak behavior around 5% coverage and declines once this threshold is surpassed. Another interesting observation is that income appears to exhibit saturation behavior with the response variable. For areas with a median annual income less than $130,000, more income is associated with more installations. However, once this threshold is passed, it appears to have little impact on the number of solar installations.

Interpretation

The most important predictor variables, perhaps the ones with the most intuitive relation to the response variable, are electricity consumption and income; both of these appear to exhibit saturation behavior. As average electricity consumption increases, homeowners are more inclined to install solar energy. This is likely explained by the fact that, for low-usage consumers who have a low utility bill, the initial investment in a solar system doesn't make much financial sense. As average consumption increases, so do the benefits of installing a solar energy system. Beyond a threshold of approximately 540 kW · h, increased electricity consumption does not affect the response because homeowners can realize the full benefit of solar energy. With regard to income, a certain amount of initial capital is required to invest in a solar energy system. As disposable income increases, so does the likelihood that a homeowner will install a solar system. The partial dependence plot also suggests the presence of a threshold at roughly $130,000. Once this threshold is surpassed, homeowners have the capital necessary to invest in renewable energy and other factors determine whether or not they will adopt such a system. It is important to note that this threshold occurs at the 90th percentile. As many areas do not have a median income that exceeds this amount and the variable is skewed in nature, the relationship between income and the response is less certain above this range.

Several variables behave as expected based on results from other studies. In general, median age has a positive relationship with the number of installations. This was observed in the published survey of Wisconsin homeowners, in which the average age of survey respondents was 60 years (Schelly 2014). A plausible hypothesis, based on the aforementioned study, is that those who are older generally have more disposable income to use on projects like solar installations. Another conjecture was that older members of the population are more likely to stay in their current home and are willing to invest in it. The model results are also in line with Sunter et al. (2019)'s prior work. Despite controlling for other social and economic factors, the proportion of the population that is White has a positive relationship with the number of solar installations, indicating a racial disparity in solar deployment. The model also identifies areas with more Democratic votes in the 2016 election as less likely to install solar energy systems, while those with more Republican votes are more likely to do so. One possible explanation is identified in the Wisconsin study, which identified financial savings as a strong motivator for conservative homeowners when installing solar capacity (Schelly 2014).

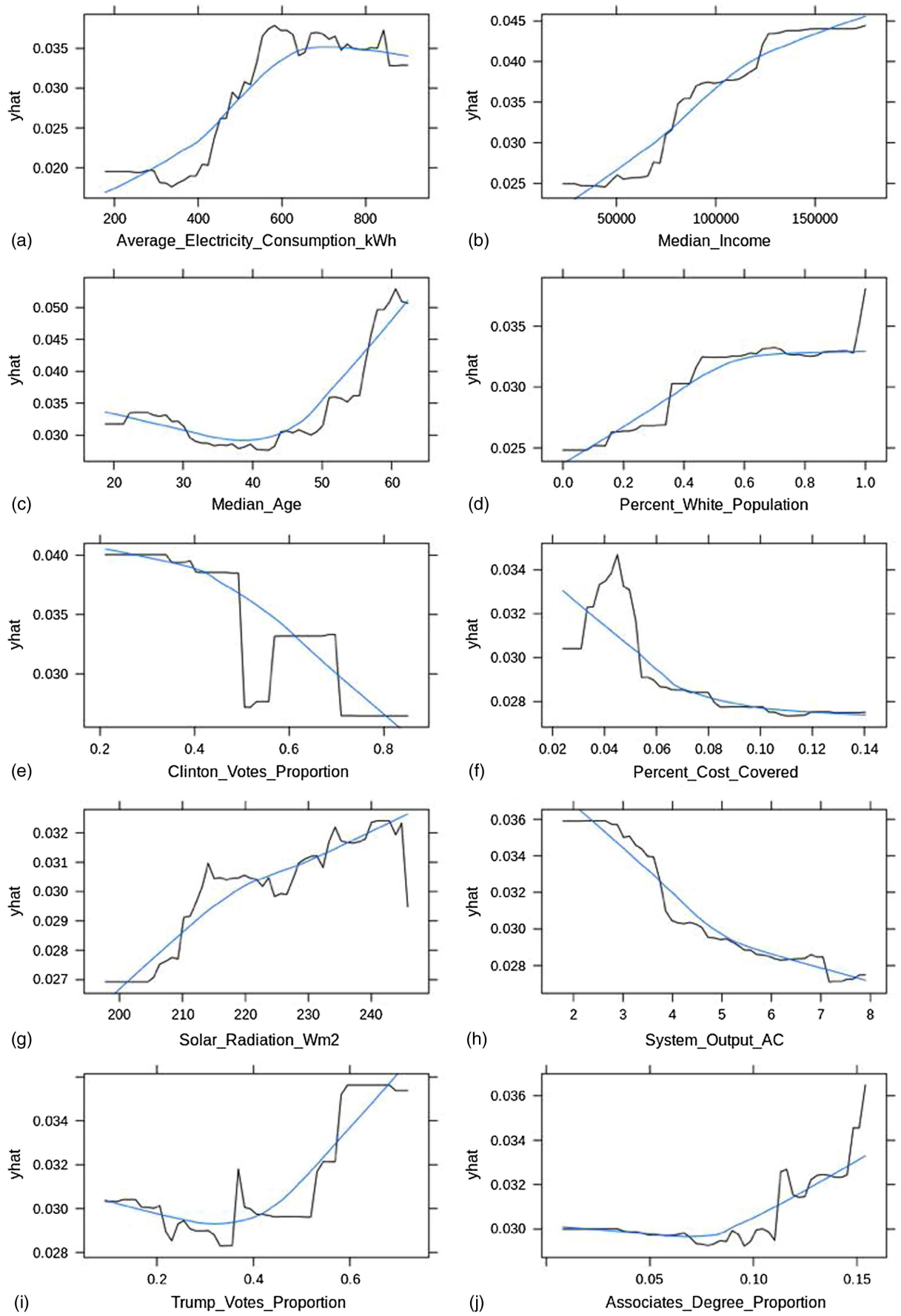

Fig. 3. Partial dependence plots for the ten most important predictors with smoothing curves.

For the remaining variables included in the ten most important, the relationships are fairly intuitive. With a higher abundance of solar energy, homeowners are more likely to take advantage of technology that can harness it. Furthermore, in communities with a higher number of solar installations, the workforce associated with the solar industry is larger. This explains the relationship of those with associate's degrees—the most common qualification for solar technicians—to the response. This is more likely a consequence of a thriving solar industry than a cause of it. Both percent of cost covered and solar system output are highly confounded with cost, which likely explains the negative relationship with the response variable. Financial savings is frequently identified as a primary driver in solar installations, and it follows that more costly systems (e.g., in the case of a high output system) would be less desirable.

Discussion

A notable observation based on these results is the importance of saturation behavior in driving solar installations. Average electricity consumption, White proportion of the population, and median income—three of the four most important variables—all appear to exhibit threshold effects. These thresholds essentially define a space for decision makers and urban planners to focus their policies to most effectively nudge solar installations. Their presence in the model is particularly interesting, as thresholds are a key concept in the discussion on resilience in socioecological systems. While the saturation behavior identified in this study is perhaps intuitive, it is notable as this is the first data-driven analysis of residential solar installations to demonstrate the existence of this type of relationship. Furthermore, recent studies have found that the economy of scale is a significant factor in driving solar installations (Kavlak et al. 2018), suggesting that solar installations will strictly increase with electricity consumption as the technology becomes cheaper on a unit basis. However, this analysis demonstrates otherwise. Because solar technology is relatively new, analyses such as the present study can determine whether or not the installation patterns in residential settings follow or deviate from expected behavior, which is valuable knowledge for policy making. Furthermore, identifying predictors with saturation behaviors can provide decision makers in other regions with valuable insights. While the specific thresholds identified in this analysis are likely unique to the California data set, the saturation behavior is likely not. While this analysis used data from California, the results should be considered in decisions being made in a variety of regions. In particular, future policy efforts should more effectively target majority non-White regions where median income is relatively low. As the expansion of clean and reliable energy sources is paramount in the development of sustainable and resilient communities, it is crucial to understand the barriers in access to renewable energy technology and the ways they can be overcome.

Another notable observation is that the relationship between age and solar installations is similar between California and Wisconsin (Schelly 2014). Despite being in geographically different regions of the country, both states see an increase in solar installations as median age rises. Because this relationship appears to hold true in two distinct regions of the country, the authors hypothesize that age may be a key factor in predicting solar installations in states beyond California and Wisconsin. This could represent an opportunity for policies that target older homeowners, especially in states with high potential for solar installations. Although this study focuses on California, the authors believe that many of the trends identified through the analysis will be similar to those observed in other parts of the country.

A key contribution of this study is the ability to compare the impacts of different predictors on residential solar installations. While some of the most important variables identified by the model used in this study have been highlighted in previous work, there are no analyses to date that have simultaneously considered the impact of this number of predictors. This provides unique insight into the relative importance of specific factors, which makes it possible to identify factors that are less relevant. Notable variables that were not included in the final model are education levels and urban classification of zip code. These results indicate that from an aggregated view, solar installations are well distributed among communities of various education levels and urban classifications. Because neither of these variables emerged as important in the model, one can infer that knowledge about a community's education or urban classification does not significantly help with making accurate predictions, implying that the distribution of solar technology does not depend on education level or how urban or rural a community is.

Civil engineering infrastructure projects and public policy are intimately linked, and the solar energy sector is no exception. To effectively manage the energy transition to renewable sources, it is critical to understand what factors motivate both businesses and people to adopt new technology. Policy can be instrumental in driving major changes in infrastructure, and by studying the factors that are relevant to the behavior it seeks to incentivize, its efficacy can be significantly increased.

Summary and Conclusion

This study aims to fill a gap in the current state of knowledge on the factors that drive solar installations. While previous work has studied the relationship between certain factors, such as race or political affiliation and solar installations, this is the first study to consider a comprehensive set of predictors simultaneously. This provides new insights into the residential solar energy by (1) providing an opportunity to analyze the impact of individual predictors in light of other predictors (e.g., what the impact of political affiliation is, considering that income has already been considered); (2) offering a relatively simple means of comparing the relative importance of different variables; and (3) identifying variables that do not contribute to predictive accuracy and are therefore likely to be unimportant in driving solar installations. Furthermore, this study employs a *predictive*, rather than an explanatory, approach. This includes a rigorous statistical predictive assessment framework, meaning that both out-of-sample and in-sample model performances are evaluated and compared between models. The advantage of a rigorous validation process such as the one presented in this study is that the final model offers a better generalization power (Shmueli 2010; Hastie et al. 2005). This is an important contribution to the literature, because it provides insight into how solar installation can be expected to change if one of the key predictors from the study changes in the coming years. Furthermore, a model such as the one developed in the study can be used to estimate baseline levels of solar installations, based on a variety of economic, environmental, and social factors. By comparing these estimates to observed behavior, regions of California that are "over-" or "underperforming" can be identified. In other words, the model could be used to identify regions with more or fewer installations than expected. This would allow decision makers to analyze the differences in these regions—which could include less quantifiable factors, such as attitude—to understand how other factors may play an important role in driving solar installations.

Specifically, this study presents a rigorous statistical modeling approach to predicting solar installations at the zip code level in California. Of the predictive models considered, the model based on the extreme gradient boosting algorithm yields the strongest performance and is best able to identify the underlying relationship between the key predictors and the response variable. Based on the model, the ten most important predictors are identified, and their relationship with solar installations is characterized. The paper presents a discussion of how these relationships can be leveraged to incentivize solar installations with a focus on the most important predictor variables. A particular focus is placed on the importance of saturation behavior in terms of electricity consumption, median income, and racial composition. By focusing on these areas, policies can more effectively increase solar energy capacity to more equitably enhance the sustainability and resilience of communities.

While the methodology used in this study is transferable to other regions of the country, the case study was restricted to California to ensure that the structure of incentive programs was consistent for all observations in the data set. California also served as an excellent case study because of the amount of relevant data that have been made publicly available. One limitation of this study is that it relies on publicly available data, which can be difficult to acquire depending on local legislation. Another limitation is that political affiliation data were collected at the county level rather than the zip code level. Because of this, the relationship identified between political affiliation and solar installations is only available at a high level, making it difficult to discern its exact nature. While the true nature is likely less "noisy" than the one presented in the partial dependence plots, the general increasing or decreasing trends are expected to hold.

This work focused on predicting solar installations to inform future policy efforts. This analysis has identified key attributes of communities where targeted incentive structures could be especially effective. A natural extension of this work would be to investigate a variety of policy levers, and to develop a model to simulate the effect of implementing them on nudging communities toward higher solar installations. This will identify effective policies for driving the clean energy transition and therefore enhancing the sustainability and resilience of communities.

Data Availability Statement

All data, models, or code generated or used during the study are available from the corresponding author by request.

References

Abdallah, M., and K. El-Rayes. 2016. "Multiobjective optimization model for maximizing sustainability of existing buildings." *J. Manage. Eng.* 32 (4): 04016003. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000425.

Alipour, P., S. Mukherjee, and R. Nateghi. 2019. "Assessing climate sensitivity of peak electricity load for resilient power systems planning and operation: A study applied to the Texas region." *Energy* 185 (Oct): 1143–1153. https://doi.org/10.1016/j.energy.2019.07.074.

Breiman, L. 2001. "Random forests." *Mach. Learn.* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Bruss, C. B., R. Nateghi, and B. F. Zaitchik. 2019. "Explaining national trends in terrestrial water storage." *Front. Environ. Sci.* 7: 85. https://doi.org/10.3389/fenvs.2019.00085.

Chapman, A. J., B. McLellan, and T. Tezuka. 2016. "Residential solar PV policy: An analysis of impacts, successes and failures in the Australian case." *Renewable Energy* 86 (Feb): 1265–1279. https://doi.org/10.1016/j.renene.2015.09.061.

Chen, C., J. Wang, F. Qiu, and D. Zhao. 2016. "Resilient distribution system by microgrids formation after natural disasters." *IEEE Trans. Smart Grid* 7 (2): 958–966. https://doi.org/10.1109/TSG.2015.2429653.

Chen, K. K. 2014. "Assessing the effects of customer innovativeness, environmental value and ecological lifestyles on residential solar power systems install intention." *Energy Policy* 67 (Apr): 951–961. https://doi.org/10.1016/j.enpol.2013.12.005.

Chen, T., and C. Guestrin. 2016. "XGBoost: A scalable tree boosting system." Preprint, submitted March 9, 2016. http://arxiv.org/abs/1603.02754.

CSI (California Solar Initiative). 2018. "California Solar Initiative (CSI)." *Go Solar California*. Accessed November 5, 2018. https://www.gosolarcalifornia.ca.gov/csi/.

Drucker, H., C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. 1997. "Support vector regression machines." In *Advances in neural information processing systems 9*, edited by M. C. Mozer, M. I. Jordan, and T. Petsche, 155–161. Cambridge, MA: MIT Press.

Dudani, S. A. 1976. "The distance-weighted *k*-nearest-neighbor rule." *IEEE Trans. Syst. Man Cybern.* SMC-6 (4): 325–327. https://doi.org/10.1109/TSMC.1976.5408784.

Energy Information Administration. 2019. "Petroleum and other liquids: U.S. field production of crude oil." Accessed April 18, 2019. https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=MCRFPUS2&f=M.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. "Regularization paths for generalized linear models via coordinate descent." *J. Stat. Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Friedman, J. H. 1991. "Multivariate adaptive regression splines." *Anal. Stat.* 19 (1): 1–67. https://doi.org/10.1214/aos/1176347963.

Friedman, J. H. 2001. "Greedy function approximation: A gradient boosting machine." *Anal. Stat.* 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

Fu, R., D. Chung, T. Lowder, D. Feldman, K. Ardani, and R. Margolis. 2016. *U.S. solar photovoltaic system cost benchmark: Q1 2016*. Technical Rep. No. NREL/TP-6A20-66532. Golden, CO: National Renewable Energy Laboratory.

Hastak, M., and C. Koo. 2017. "Theory of an intelligent planning unit for the complex built environment." *J. Manage. Eng.* 33 (3): 04016046. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000486.

Hastie, T., and R. Tibshirani. 1986. "Generalized additive models." *Stat. Sci.* 1 (3): 297–310. https://doi.org/10.1214/ss/1177013604.

Hastie, T., R. Tibshirani, J. Friedman, and J. Franklin. 2005. *The elements of statistical learning: Data mining, inference and prediction*. 2nd ed., 83–85. Berlin: Springer.

Hendrickson, C. 2012. "Sustainable energy challenges for civil engineering management." *J. Manage. Eng.* 28 (1): 2–4. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000074.

Kavlak, G., J. McNerney, and J. E. Trancik. 2018. "Evaluating the causes of cost reduction in photovoltaic modules." *Energy Policy* 123 (Dec): 700–710. https://doi.org/10.1016/j.enpol.2018.08.015.

Kim, K., T. Park, S. Bang, and H. Kim. 2017. "Real options-based framework for hydropower plant adaptation to climate change." *J. Manage. Eng.* 33 (3): 04016049. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000496.

Kohavi, R. 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection." In *Proc., 14th Int. Joint Conf. on Artificial Intelligence—Volume 2, IJCAI'95*, 1137–1143. San Francisco: Morgan Kaufmann.

Kwan, C. L. 2012. "Influence of local environmental, social, economic and political variables on the spatial distribution of residential solar PV arrays across the United States." *Energy Policy* 47 (Aug): 332–344. https://doi.org/10.1016/j.enpol.2012.04.074.

Liu, B., Y. Hu, A. Wang, Z. Yu, J. Yu, and X. Wu. 2018. "Critical factors of effective public participation in sustainable energy projects." *J. Manage. Eng.* 34 (5): 04018029. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000635.

Lokhandwala, M., and R. Nateghi. 2018. "Leveraging advanced predictive analytics to assess commercial cooling load in the US." *Sustainable Prod. Consumption* 14 (Apr): 66–81. https://doi.org/10.1016/j.spc.2018.01.001.

Muhammad-Sukki, F., R. Ramirez-Iniguez, S. H. Abu-Bakar, S. G. McMeekin, and B. G. Stewart. 2011. "An evaluation of the installation of solar photovoltaic in residential houses in Malaysia: Past, present, and future." *Energy Policy* 39 (12): 7975–7987. https://doi.org/10.1016/j.enpol.2011.09.052.

Mukherjee, S., and R. Nateghi. 2019. "A data-driven approach to assessing supply inadequacy risks due to climate-induced shifts in electricity demand." *Risk Anal.* 39 (3): 673–694. https://doi.org/10.1111/risa.13192.

Mukhopadhyay, S., and R. Nateghi. 2017. "Estimating climate—Demand nexus to support long-term adequacy planning in the energy sector." In *Proc., 2017 IEEE Power and Energy Society General Meeting*, 1–5. New York: IEEE.

Nateghi, R., J. D. Bricker, S. D. Guikema, and A. Bessho. 2016. "Statistical analysis of the effectiveness of seawalls and coastal forests in mitigating tsunami impacts in Iwate and Miyagi Prefectures." *PLoS One* 11 (8): e0158375. https://doi.org/10.1371/journal.pone.0158375.

Nateghi, R., S. D. Guikema, and S. M. Quiring. 2011. "Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes." *Risk Anal. Int. J.* 31 (12): 1897–1906. https://doi.org/10.1111/j.1539-6924.2011.01618.x.

NREL (National Renewable Energy Lab). 2016. "What is the NSRDB?" Accessed November 7, 2018. https://nsrdb.nrel.gov/.

Obringer, R., and R. Nateghi. 2018. "Predicting urban reservoir levels using statistical learning techniques." *Sci. Rep.* 8 (1): 5164. https://doi.org/10.1038/s41598-018-23509-w.

Osborn, S. G., A. Vengosh, N. R. Warner, and R. B. Jackson. 2011. "Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing." *Proc. Natl. Acad. Sci.* 108 (20): 8172–8176. https://doi.org/10.1073/pnas.1100682108.

Pacific Gas and Electric. 2016. "PG&E's energy data request portal." Accessed November 12, 2018. https://pge-energydatarequest.com/public_datasets.

Pearce, L. 2003. "Disaster management and community planning, and public participation: How to achieve sustainable hazard mitigation." *Nat. Hazards* 28 (2): 211–228. https://doi.org/10.1023/A:1022917721797.

Politico. 2016. "California election results 2016: President live map by county, real-time voting updates." *Election Hub.* Accessed November 15, 2018. https://www.politico.com/2016-election/results/map/president/california/.

Qiao, Y., T. U. Saeed, S. Chen, R. Nateghi, and S. Labi. 2018. "Acquiring insights into infrastructure repair policy using discrete choice models." *Transp. Res. Part A: Policy Pract.* 113 (Jul): 491–508. https://doi.org/10.1016/j.tra.2018.04.020.

Rai, V., D. C. Reeves, and R. Margolis. 2016. "Overcoming barriers and uncertainties in the adoption of residential solar PV." *Renewable Energy* 89 (Apr): 498–505. https://doi.org/10.1016/j.renene.2015.11.080.

Raymond, L., D. Gotham, W. McClain, S. Mukherjee, R. Nateghi, P. V. Preckel, P. Schubert, S. Singh, and E. Wachs. 2018. "Projected climate change impacts on Indiana's energy demand and supply." *Clim. Change* 1–15. https://doi.org/10.1007/s10584-018-2299-7.

San Diego Gas & Electric. 2016. "Energy data access." Accessed November 12, 2018. https://energydata.sdge.com/.

Schelly, C. 2014. "Residential solar electricity adoption: What motivates, and what matters? A case study of early adopters." *Energy Res. Social Sci.* 2 (Jun): 183–191. https://doi.org/10.1016/j.erss.2014.01.001.

Shmueli, G. 2010. "To explain or to predict?" *Stat. Sci.* 25 (3): 289–310. https://doi.org/10.1214/10-STS330.

Southern California Edison. 2016. "Energy data—Reports and compliance." *SCE.com*. Accessed November 12, 2018. https://www.sce.com/partners/partnerships/access-energy-usage-data.

Sunter, D. A., S. Castellanos, and D. M. Kammen. 2019. "Disparities in rooftop photovoltaics deployment in the United States by race and ethnicity." *Nat. Sustainability* 2 (1): 71–76. https://doi.org/10.1038/s41893-018-0204-z.

Sunter, D. A., J. Dees, S. Castellanos, D. Callaway, and D. M. Kammen. 2018. "Political affiliation and rooftop solar adoption in New York and Texas." In *Proc., 2018 IEEE 7th World Conf. on Photovoltaic Energy Conversion (WCPEC) (A Joint Conf. of 45th IEEE PVSC, 28th PVSEC 34th EU PVSEC)*, 2426–2429. NewYork: IEEE. https://doi.org/10.1109/PVSC.2018.8548257.

US Census Bureau. 2016. "2012–2016 ACS 5-year estimates." Accessed October 25, 2018. https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2016/5-year.html.

USDA. 2010. "USDA ERS—Rural-urban commuting area codes." Accessed November 14, 2019. https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes.aspx.

Washington State Dept. of Health. 2008. *Guidelines: A recap list of acronyms*. Tumwater, WA: Washington State Dept. of Health.

Yi, W., and A. P. C. Chan. 2015. "Which environmental indicator is better able to predict the effects of heat stress on construction workers?" *J. Manage. Eng.* 31 (4): 04014063. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000284.