Kim et al. *EPJ Data Science* (2023) 12:25
https://doi.org/10.1140/epjds/s13688-023-00399-1

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

# Spatial distribution of solar PV deployment: an application of the region-based convolutional neural network

Serena Y. Kim[1,2,3*], Koushik Ganesan[4], Crystal Soderman[3] and Raven O'Rourke[5]

*Correspondence:
serena_kim@ncsu.edu
[1]School of Public and International Affairs, North Carolina State University, 2221 Hillsborough St, Raleigh, NC 27607, USA
[2]College of Engineering, Design and Computing, University of Colorado Denver, 1200 Larimer St, Denver, CO 80204, USA
Full list of author information is available at the end of the article

**Abstract**

Solar photovoltaic (PV) deployment plays a crucial role in the transition to renewable energy. However, comprehensive models that can effectively explain the variations in solar PV deployment are lacking. This study aims to address this gap by introducing two innovative models: (i) a computer vision model that can estimate spatial distribution of solar PV deployment across neighborhoods using satellite images and (ii) a machine learning (ML) model predicting such distribution based on 43 factors. Our computer vision model using Faster Regions with Convolutional Neural Network (Faster RCNN) achieved a mean Average Precision (mAP) of 81% for identifying solar panels and 95% for identifying roofs. Using this model, we analyzed 652,795 satellite images from Colorado, USA, and found that approximately 7% of households in Colorado have rooftop PV systems, while solar panels cover around 2.5% of roof areas in the state as of early 2021. Of our 16 predictive models, the XGBoost models performed the best, explaining approximately 70% of the variance in rooftop solar deployment. We also found that the share of Democratic party votes, hail and strong wind risks, median home value, the percentage of renters, and solar PV permitting timelines are the key predictors of rooftop solar deployment in Colorado. This study provides insights for business and policy decision making to support more efficient and equitable grid infrastructure investment and distributed energy resource management.

**Keywords:** Solar PV; Data mining; Computer vision; Region-based convolutional neural network; Energy transition; Renewable energy; Energy justice

## 1 Introduction

In the past decade, solar energy technology has advanced rapidly to the point where the 20-year Levelized Cost of Energy (LCOE) for solar is now lower than that of coal in many countries around the world [1]. However, despite these advancements, the average payback period for a residential solar project still ranges from 7 to 10 years as of 2022 [2]. This means that the upfront costs associated with installing solar panels remain substantial barriers for low- and moderate-income (LMI) households interested in investing in solar energy. Consequently, it is not surprising that individuals with higher incomes are more likely to adopt solar energy compared to those with lower incomes [3]. While more



Springer

affluent communities have been able to take advantage of government subsidies and incorporate solar energy into their energy portfolio, LMI communities have been left behind, lacking access to the benefits of solar energy while being more vulnerable to increasing energy costs.

The deployment of rooftop solar photovoltaic (PV) systems has been uneven across neighborhoods with varying natural and socioeconomic characteristics. Racially diverse communities [4] and lower-income neighborhoods [5] generally have lower rates of solar energy deployment. Additionally, areas with higher solar radiation [6] and government incentives for renewable energy, such as renewable energy portfolio standards (RPS) and net metering (NEM), tend to exhibit higher levels of solar PV adoption [7]. However, the collective influence of socioeconomic and environmental factors on the uneven distribution of rooftop solar deployments remains unclear.

This study utilizes satellite imagery from the state of Colorado, USA in 2021 to model and predict disparities in solar PV deployment at the US Census block group level. Using Geographical Information Systems (GIS), we aggregated 43 layers of data related to the natural and built environment, social and economic factors, as well as energy policy information. Our approach involves building a machine learning (ML) model that predicts two measures of solar PV deployment: the count of solar PV installations per household (henceforth PV Count Per Household (HH)) and the PV-to-roof area ratio (henceforth PV-to-Roof Ratio). These predictions are based on the 43 input features included in the model. The state of Colorado serves as an ideal location for this investigation due to its high solar radiation, diverse natural landscapes, and varying socioeconomic characteristics across neighborhoods within the state. The total area of Colorado is 269,837 $km^2$. We identify the following four unique contributions of this study.

First, our study is one of the first attempts to measure PV-to-roof ratio, or the proportion of roofs covered by solar panels. To achieve this, we used Faster Regions with Convolutional Neural Network (RCNN) [8], a deep convolutional neural network trained end-to-end to generate high-quality region proposals and used for object detection in computer vision. Unlike previous approaches that focus on solar system counts per capita/household [9] or the size of solar panels [6], our study pioneers the use of the PV-to-roof ratio, which provides valuable information for utility infrastructure planning and energy policy design, indicating the availability of underutilized roof areas for future PV installations. Communities aiming to enhance their distributed solar energy generation can use this ratio to identify areas with a low PV-to-roof ratio for potential deployment of additional solar panels. Furthermore, aggregated PV areas have implications for utility infrastructure planning, as utilities may need to adjust procurement strategies and upgrade grid infrastructure to ensure the reliability of the electrical grid system.

Second, our study provides one of the most detailed solar deployment datasets to date. Existing studies examining neighborhood-level solar deployment often rely on US Census tract-level data (with an average size of 4000 residents per tract) or zip code-level data (with an average of 8000 residents per zip code). However, these spatial units may not capture the spatial disparities in PV deployment adequately. Neighborhoods within a tract or zip code can exhibit significant variations in residents' socioeconomic status and living arrangements. In contrast, our study estimates solar deployment at the US Census block group level, which has an average size of 1500 residents. By focusing on a

smaller spatial unit, our data provide more targeted and nuanced results, as the variation in neighborhood-level measures tends to increase with decreasing spatial scale [10].

Third, to the best of our knowledge, this study is the first attempt to explore the relationship between natural disaster vulnerabilities and residents' decisions to deploy rooftop solar. Different types of natural disasters may influence people's willingness to invest in solar energy differently. For instance, individuals residing in areas prone to frequent power outages due to extreme cold or heat might be more inclined to install rooftop solar with storage systems to ensure a continuous power supply during outages. On the other hand, residents experiencing frequent hailstorms, hurricanes, or tornadoes may be hesitant to invest in rooftop solar due to concerns about potential damage to the panels. In this study, we examine the correlation between six common natural disasters in Colorado and rooftop solar deployment.

Lastly, by leveraging recent developments in tree-based algorithms and the SHapley Additive exPlanations (SHAP) method for interpretable machine learning, our model, which includes 43 predictors, is one of the most comprehensive and precise predictive models of rooftop solar deployment to date. Our best model, developed using XGBoost, explains about 70% of the variation in PV deployment. Existing solar deployment prediction models can potentially improve their performance by incorporating the variables identified as important in our analysis, such as political ideology, types and frequency of natural disasters, and local government rooftop solar permitting rules.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related work that underpins the empirical analysis of our study. Section 3 describes the data sources, GIS data processing, computer vision methods, and machine learning models employed in this research. Section 4 presents the results and identifies the significant predictors of rooftop solar deployment. Section 5 discusses the implications and limitations of our study. Finally, in Sect. 6, we summarize our findings and outline potential avenues for future research.

## 2 Related work

Existing studies on solar PV deployment often take one of the three approaches. The first approach involves individual or household surveys [11, 12], which allow for targeted data collection at the individual level. However, this approach can be costly and time-consuming to obtain a geographically representative and comprehensive dataset. The second approach utilizes datasets collected by governments and utility service providers who manage interconnection processes, which determine how rooftop solar connects to the grid [13]. While this approach provides accurate data on solar adopters, address-level solar deployment or interconnection data are often protected and difficult to access for research purposes. Therefore, we have chosen the third approach, utilizing satellite imagery and computer vision models to detect the existence and size of rooftop solar PV installations [6]. Although this approach requires more time and computing resources for collecting and processing large satellite imagery datasets, it allows us to obtain uncensored and comprehensive data on the spatial distribution of solar deployment.

Several rooftop solar deployment datasets have been created using satellite images. Google Project Sunroof [14] estimates the number and sizes of solar PV installations using satellite images since 2013. However, approximately 25% of areas, mainly rural areas, are not included in their dataset. DeepSolar [6] developed a comprehensive database of solar installations for the contiguous United States. However, the data are from 2017 or

earlier and may not capture the recent sharp increase in solar deployment. The National Renewable Energy Laboratory's (NREL) Distributed Generation Market Demand (dGEN) [15] model provides up-to-date solar deployment data, but the most granular spatial resolution available to the public is county-level data, which is insufficient for capturing neighborhood-level spatial disparities. Therefore, we have created our solar deployment dataset from scratch, collecting satellite images of Colorado, our area of interest.

The predictors of solar PV deployment have been extensively studied in the renewable energy literature. Previous studies have found positive correlations between solar PV deployment and residents' income [6], education [16], and age [11]. Census tracts with Black or Hispanic majorities have installed fewer rooftop PV systems compared to tracts without a racial/ethnic majority, even after controlling for median household income [17]. Solar radiation has been found to positively predict PV deployment [6, 18]. More rooftop solar PVs exist in areas with a smaller proportion of renter-occupied homes [13], higher median home value [19], and newer buildings. Studies from various countries have identified an urban-rural divide in solar PV deployment, although the direction of this divide varies depending on the study context. For instance, rural municipalities have more solar PV projects per capita in Switzerland [20], while urban counties have more solar installations per capita in Georgia, USA [21]. Recent research has also found that solar deployment occurs in many Republican households, but to a greater extent in Democratic households [22]. Disadvantaged communities facing socioeconomic, health, and environmental burdens are significantly less likely to adopt solar PV compared to more advantaged communities, even after controlling for median household income [5].

We have identified four significant gaps in the literature on the predictors of solar PV deployment. First, the relationship between tree canopy cover and residents' decisions to deploy PV systems has not been examined. Second, while the proximity to transmission lines can impact the performance of distributed solar PV installations, no studies have investigated the influence of the existence or size of transmission infrastructure on building owners' decisions to adopt solar. Third, although existing literature has examined how solar PV deployment relates to state-level policies, little is known about the impact of local-level policies and rules, such as solar mandates and Solar Permitting, Inspection, and Interconnection (PII) rules. Lastly, there is limited empirical research on how the types and intensity of natural disasters influence solar PV adoption and deployment size.

## 3 Data and materials

### 3.1 Satellite imagery data collection

To ensure that our data collection focused on the geographical areas of interest, we excluded census blocks that had no residents. This exclusion was necessary because our study specifically targeted rooftop solar installations, rather than large-scale ground-mounted solar projects. Out of the 201,062 blocks in Colorado, we removed 55,258 blocks with zero residents from our data collection. Figure 1 visually represents the 145,804 blocks that were included in our analysis.

We developed two measures of rooftop solar deployment: (i) the number of rooftop solar panels per household (PV count per HH) and (ii) the area of solar panels to roofs (PV-to-roof ratio). To obtain comprehensive satellite images encompassing all rooftops in Colorado, we utilized the Google Maps JavaScript API for downloading Google Earth satellite images. The polygons representing the census blocks were subdivided into smaller polygons to acquire high-resolution images. This process resulted in a collection of 652,795

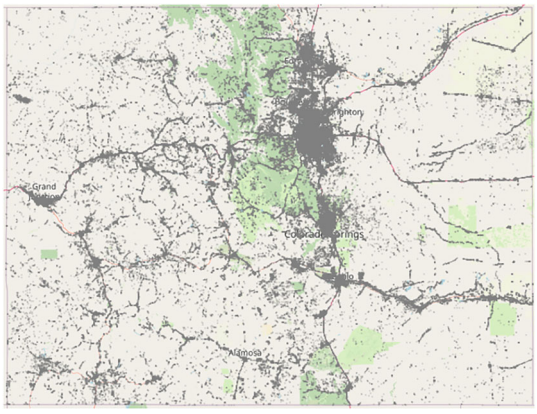

**Figure 1** Gray areas are the selected areas for satellite imagery collection

image tiles, each with dimensions of 640 × 640 pixels at zoom level 20. The latitude and longitude coordinates of each polygon's center and boundary were utilized to retrieve the corresponding images for the respective census blocks. We employed the World Geodetic System WGS84 format to retrieve the spatial information of the polygons. Our data collection took place between January 2021 and February 2021. The satellite images used for Colorado were captured by Google Maps from September 2020 to February 2021.

### 3.2 Computer vision frameworks

The advancements in deep learning have revolutionized various applications, including computer vision tasks. Significant progress has been made in recent years, starting with the introduction of Regions with Convolutional Neural Network (RCNN) and region proposal methods [23, 24]. A more efficient version of the original RCNN was developed to address the computational expense and training time, known as Fast RCNN [25]. The Fast RCNN replaced max pooling with ROI pooling, allowing for the generation of a convolutional feature map by processing the entire image once. While this improvement was substantial, the bottleneck remained in the region proposal step for Fast RCNNs.

To overcome the computational challenges associated with region proposals in Fast RCNN, the Faster RCNN framework was introduced as a more widely used version [8]. The Faster RCNN addresses the performance issues by incorporating a separate convolutional network called the region proposal network (RPN) for generating region proposals. Unlike the previous versions that utilized a CPU-based selective search algorithm, the RPN significantly improves the speed of region proposals and enables the sharing of layers between proposal generation and object detection. This sharing of layers enhances the overall efficiency of the network. The Faster RCNN model has emerged as one of the state-of-the-art object detectors, surpassing the performance of other traditional models such as YOLO, SSD, and other traditional models on several key metrics [26, 27].

### 3.3 Satellite imagery data processing: faster RCNN

To detect solar panels and roofs in satellite images, we employed the Faster RCNN model. For the backbone network, we utilized ResNet-50, which was pretrained on COCO train2017. We replaced the pretrained head with a FastRCNNPredictor to optimize the performance of the image detection.

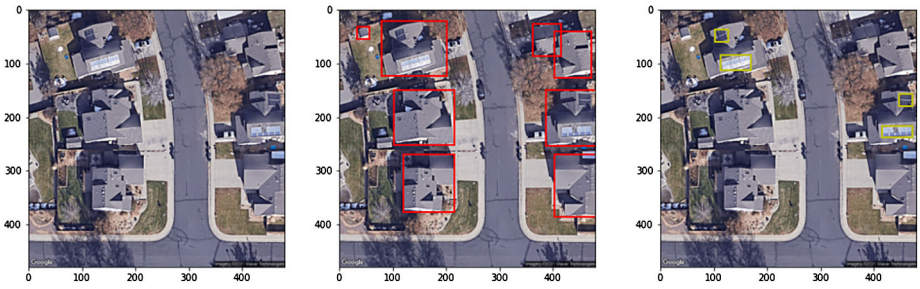

**Figure 2** An example of an annotated image: The leftmost image represents the original image, the center image displays the bounding boxes that indicate the location of roofs, and the rightmost image highlights the solar panels within the detected roofs

To enhance the accuracy of the detection, we created two versions of the pretrained Faster RCNN model: one for identifying roofs and another for identifying PVs. We trained the model using a dataset of 367 annotated images, each of size $480 \times 480$. The dataset was divided into train (80%), development (10%), and test (10%) sets. An example of such an annotated image is depicted in Fig. 2.

We used the TorchVision library to import the model and employed the AdamW optimizer [28] to minimize cross-entropy loss. To fine-tune the other hyper-parameters, we performed a number of experimental tests on the development set. For the final model, we used a learning rate of $2 \times 10^{-4}$, a weight decay of 0.001. We also used a StepLR as our scheduler with decaying the learning every epoch by gamma of 0.7 along with a batch size of 8 and trained for 4 epochs on Tesla K10 GPUs.

After obtaining the predictions, we applied non-maximum suppression (NMS) with Intersection over Union (IoU) thresholds of 0.2 for roofs and 0.1 for PVs. This process helped in eliminating redundant bounding boxes and selecting the most accurate predictions. To assess the prediction accuracy, we employed the mean Average Precision (mAP) metric with an IoU threshold of 0.5 for all object sizes. The IoU is given by

$$\text{IoU} = \frac{\text{area}(A_p \cap A_{gt})}{\text{area}(A_p \cup A_{gt})}, \tag{1}$$

where $A_p$ is the predicted frame and $A_{gt}$ is the ground truth frame. Our model achieved an mAP of 0.95 and 0.81 for detecting roofs and PVs, respectively.

### 3.4 Two measures of rooftop solar deployment

Two measures, *PV Count Per Household* and *PV-to-Roof Ratio*, of rooftop solar PV deployment at the block group level by running our Faster RCNN model on 652,795 satellite imagery data.

$$\text{PV Count Per Household (HH)} = \frac{\sum_{n=1}^{k} \text{Number of Solar PV Systems}_k}{\sum_{n=1}^{k} \text{Number of Households}_k}, \tag{2}$$

$$\text{PV-to-Roof Ratio} = \frac{\sum_{n=1}^{k} \text{Solar PV Area}_k}{\sum_{n=1}^{k} \text{Roof Area}_k}, \tag{3}$$

where $k$ is the number of images for a block group. On average, 189 images correspond to a block group.

### 3.5 Predictors of solar PV deployment

The second objective of our study is to predict rooftop solar deployment using various social and environmental features. To build a comprehensive and accurate predictive model of rooftop solar deployment, we considered four groups of predictors: (1) natural environment, (2) demographics and built environment, (3) energy infrastructure, market, and policy, and (4) social and natural disaster vulnerabilities.

#### *3.5.1 Natural environment*

*Solar radiation* Solar radiation data were obtained from the national solar radiation database (NSRDB). We focused on the Direct Normal Irradiance (DNI), which measures the amount of solar energy received per unit area by a surface held perpendicular to the sun's rays. In our analysis, we calculated the average DNI by taking the mean of the maximum and minimum DNI values measured at the centroid of each block. These block-level DNI measures were then averaged over all blocks within a block group.

*Tree canopy cover* The presence of trees in an area can impact the energy output of solar panels and potentially deter residential solar deployment in regions with significant tree coverage. Tree canopy cover has been shown to substantially reduce the energy output of rooftop solar PV systems [29]. Additionally, in urban areas where tree coverage helps mitigate the urban heat island effect and regulate stormwater runoff [30], the competition for space between solar panels and trees may lead to lower residential and small-scale solar deployment [31]. To assess tree canopy cover, we utilized data from the United States Forest Service (USFS) and created the variable "% of tree-to-land area," which measures the proportion of land area covered by trees within each block group.

#### *3.5.2 Demographics and built environment*

*Demographics* We included various demographic variables in our model, such as median household income, race and ethnicity distribution (i.e., % White, % Hispanic, % African American, % Asian, % Other race), median age, the proportion of individuals with a bachelor's degree or higher, and the proportion of households with at least one member aged 65 or older. These demographic data were obtained from the 2019 American Community Survey (ACS).

*Housing characteristics* Our model incorporated housing characteristics at the block group level, utilizing data from the ACS. These characteristics include the proportion of renter-occupied housing units, median home values, and the median year of construction for residential structures. For approximately 2% of missing estimates in median home values and median year of construction, we imputed the average values from adjacent block groups.

*Rural-urban classification* To account for the distinctions between rural and urban areas, we incorporated rural-urban continuum codes from the U.S. Department of Agriculture (USDA) [32] into our analysis. These codes range from 1 to 9, with higher values indicating greater rurality.

*Political ideology* We included the percentage of votes for the 2020 Democratic presidential candidate at the county level in our models to capture the heterogeneity in political

orientation among residents. This data was sourced from the MIT Election Data and Science Lab [33].

#### *3.5.3 Energy infrastructure, market, and policy*

*Transmission lines* Building owners' decision to install rooftop solar could be influenced by the existence of large grid infrastructure such as a high-power electrical transmission tower in their neighborhood. Thus, we included transmission data from Homeland Infrastructure Foundation-Level Data (HIFLD) by aggregating the length and voltage of the transmission lines within each block group level.

*Utility ownership* Since utilities generally develop and implement interconnection standards which define how rooftop PVs can connect to the grid, our analysis includes three types of utility ownership: investor-owned utilities (IOUs), Municipal utilities (MOUs), and rural electric cooperatives (co-ops). Utility ownership data are obtained by spatially merging the map of block groups and the electric utility territory map.

*Electricity price* Using zip-code level utility rate data from the Utility Rate Database (URDB) [34], we obtained residential, commercial, and industrial utility rates for each block group. When a block group has multiple zip codes, the most populous zip code within the block group was used to extract the utility rates for the block group.

*Solar mandates* Solar mandates, a building code that requires new construction homes to be solar-ready or to have a PV system installed, can address the cost-prohibitive barriers of retrofitting a roof or removing shade obstructions associated with solar PV adoption [35]. We obtained local-level solar mandate information from municipal building codes publicly available on municipality web pages.

*Solar permitting, inspection, and interconnection (PII) rules* Local (i.e., city, town, county) PII requirements can affect the duration of solar PV installations significantly [36]. Thus, our models incorporate four PII rule variables: (i) SolSmart Awardee (1 for a local jurisdiction awarded by the SolSmart award for improvements to local permitting, inspection, planning, zoning, and/or market development to facilitate solar installs and mitigate associated soft costs, or 0 otherwise), (ii) Online Permit (1 for a local jurisdiction accepting permit submissions through an online portal or email, or 0 otherwise), (iii) Same-day In-person Permit (1 for a local jurisdiction offering over-the-counter permit submission and approval, or 0 otherwise), and (iv) Permit & Pre-Install Days (median business days between first permit submission to the local jurisdiction and approval). Data are from NREL's SolarTrace [37].

#### *3.5.4 Social and natural disaster vulnerabilities*

*Social vulnerability* Utilizing the Social Vulnerability Index (SVI) from the Centers for Disease Control and Prevention (CDC) [38], we specifically consider nine social vulnerability variables: (i) % Below Poverty (percentage of persons below poverty), (ii) % Disability (percentage of civilian non-institutionalized population with a disability), (iii) % Single Parent (percentage of single parent households with children under 18), (iv) % Limited English (percentage of persons (age 5+) who speak English "less than well"), (v) % 10+ Unit

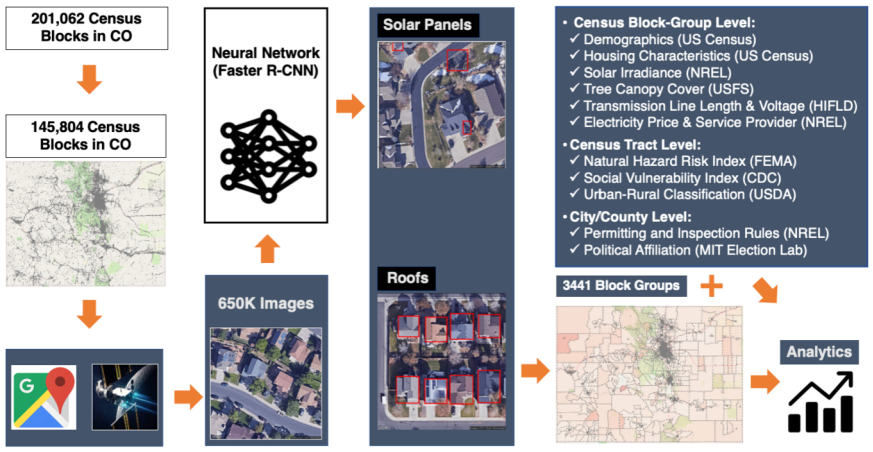


**Figure 3** Data processing workflow

Housing (percentage of housing in structures with 10 or more units), (vi) % Mobile Home (percentage of mobile homes), (vii) % People > Rooms (percentage of occupied housing units with more people than bedrooms), (viii) % No Vehicles (percentage of households with no vehicle), (ix) % Unemployed (unemployment rate). We merged our block-group data with the census-tract level SVI data which is the most granular spatial resolution available.

*Natural disaster vulnerability* Considering the resilience value of rooftop solar, we also examined the impact of natural disasters on PV deployments. While rooftop solar can provide electricity during power outages caused by extreme weather events, certain types of natural disasters, such as hailstorms or tornadoes, can damage roofs and PV systems, discouraging investment in rooftop solar. Therefore, we incorporated the expected annual loss scores (EALS) of six of the most frequent natural disasters in Colorado from the National Risk Index by the Federal Emergency Management Agency (FEMA) [39]. These natural disaster risks include drought, wildfire, hail, winter weather, strong wind, and tornado risks.

These four groups of determinants – (i) natural environment, (ii) demographics and built environment, (iii) energy infrastructure, market, and policy, and (iv) social and natural disaster vulnerabilities – were merged with the two measures of rooftop solar deployment at the block group level. A snapshot of the satellite and geospatial data processing workflow is provided in Fig. 3. The energy infrastructure, market, and policy features corresponding to the boundaries of local jurisdictions, utility service areas, or zip code tabulated areas were spatially merged with the block group map using the GeoPandas Python library [40]. In cases where a block group crossed multiple jurisdictional boundaries (which accounted for less than 5% of all block groups), the jurisdiction with the largest share of the area was selected to represent the block group. Descriptive statistics of the variables are presented in Appendix A.

### 3.6 ML models

In order to cross-validate the prediction results and achieve higher performance, we deployed four ensemble ML algorithms:

1. *Random Forest* is a supervised machine learning algorithm developed by Breiman [41] for both classification and regression tasks. This algorithm utilizes feature bagging, where strong predictors for the target variable are selected in multiple samples or trees. Random Forest was chosen as our baseline algorithm due to its robustness against overfitting compared to other algorithms.
2. *CATBoost* is another supervised machine learning algorithm developed by Dorogush et al. [42] in 2017. It is designed for classification and regression tasks and employs boosting techniques instead of bagging. CATBoost sequentially generates decision trees, utilizing the results from existing trees to enhance the predictions of subsequent trees. It is particularly effective in handling categorical features, making it suitable for our dataset that includes categorical data.
3. *LightGBM*, short for light gradient-boosting machine, is a gradient boosting framework introduced by Microsoft in 2016. It is based on decision tree architectures and boosting techniques, and it offers capabilities for ranking, regression, classification, and other machine learning tasks. We opted for LightGBM due to its faster training speed compared to XGBoost, while still achieving comparable performance [43].
4. *XGBoost*, which stands for Extreme Gradient Boosting, is a powerful machine learning algorithm developed by Chen and Guestrin [44] in 2014. It is widely used for regression, classification, and ranking problems. XGBoost implements the gradient boosted trees algorithm, combining the estimates of weak learners to make accurate predictions. We deployed XGBoost because of its speed, flexibility, and ability to produce high-performing models for prediction tasks.

For regression tasks, the mean or average prediction of and other Four ML models are estimated on four datasets: (i) PV count dataset without energy policy variables ($n$ = 3441), (ii) PV count dataset with energy policy variables ($n$ = 2328), (iii) PV-to-roof ratio dataset without energy policy variables ($n$ = 3441), (iv) PV-to-roof ratio dataset with energy policy variables ($n$ = 2328), yielding a total of 16 models. We separated models without policy variables because policy variables were not available for all cities and towns in Colorado, and some block groups are in unincorporated communities, which are not considered to be municipal areas of their own accord.

*Metrics* We evaluate the performance of all models based on three measures: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination ($R^2$).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{6}$$

where $n$ is the number of sample (observations) in the dataset, $y_i$ is observed (true) value of the target variable, $\hat{y}_i$ is the estimated (predicted) value of the target variable, and $\bar{y}$ is

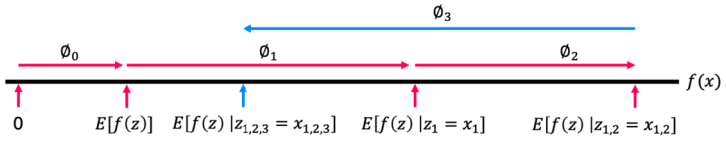


**Figure 4** SHAP values assuming a single ordering (123) of input features (adapted from [48])

the mean value of the target variable. The optimization of hyper-parameters are provided in Table A1 for each algorithm.

*Shapely values (SHAP)* SHAP is an approach for interpretable ML. The SHAP approach evaluates the average marginal contribution of a feature value across all possible coalitions (orderings) of all feature values. The concept of Shapley value, originally introduced by Shapley [45] in 2016, was initially used to assign payouts in cooperative game theory based on each player's contribution to the overall payout. Building upon this concept and other additive feature attribution methods like LIME (Local Interpretable Model-agnostic Explanations) [46] and DeepLIFT [47], Lundberg and Lee [48] developed a fast algorithm in 2017 for explaining the outputs of ML models. The SHAP value represents the importance of each input feature in the expected model prediction and is obtained by aggregating a set of values. These values quantify the attribution or impact of each feature on the change in the expected model prediction. Figure 4 is a simplified illustration of how SHAP values ($\phi_i$) are obtained by attributing the change in the expected model prediction ($f(x)$) by conditioning on each feature ($x$).

The order of input features and their interdependencies can impact the expected model prediction. However, SHAP values are derived by averaging over all possible orderings of input features, providing a unified measure of feature importance. The SHAP value $\phi_i$ is obtained as a unified measure of feature importance is the solution to the following equation (7):

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}\left[f_x(z') - f_x(z' \setminus i)\right] \tag{7}$$

In this equation, $x$ represents the input features, $|z'|$ denotes the number of non-zero entries in $z'$, and $z' \subseteq x'$ encompasses all $z'$ where the non-zero entries in $z'$ are a subset of the non-zero entries in $x'$. The term $f_x(z') - f_x(z' \setminus i)$ represents the marginal contribution of each input feature, capturing the change in the expected model prediction when considering or excluding a specific feature. The weight $\frac{|z'|!(M-|z'|-1)!}{M!}$ accounts for the different ways each input feature can be added to the ordering. Here, $z' \in 0, 1^M$, and $M$ denotes the number of input features.

## 4 Results

### 4.1 Rooftop solar PV deployment in Colorado

At the block group level ($n$ = 3441), approximately 7% of households in Colorado are observed to have solar panels installed on their roofs, while on average, 2.5% of roofs in the state are covered by solar panels (refer to Fig. 5). The percentage of households with solar panels per household ranges from 0% to 78.4%, indicating significant variability. Similarly,

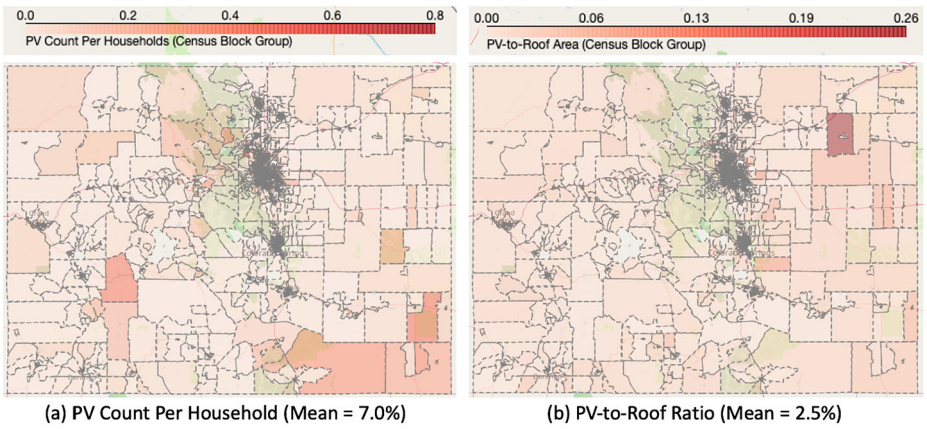


**Figure 5** Rooftop solar PV deployment in Colorado

**Table 1** Model Performance Comparison

| Model | Dataset | Algorithm | MAE | RMSE | $R^2$ |
|---|---|---|---|---|---|
| M1 | PV count per HH | XGBoost | 0.038 | 0.003 | 62.2% |
| M2 | | CATBoost | 0.039 | 0.003 | 57.4% |
| M3 | | LightGBM | 0.038 | 0.003 | 60.4% |
| M4 | | Random Forest | 0.039 | 0.003 | 60.9% |
| M5 | PV count per HH *with energy policy predictors* | XGBoost | 0.038 | 0.003 | **68.5%** |
| M6 | | CATBoost | 0.008 | 0.0001 | 66.0% |
| M7 | | LightGBM | 0.008 | 0.0001 | 66.8% |
| M8 | | Random Forest | 0.009 | 0.001 | 61.8% |
| M9 | PV-to-roof ratio | XGBoost | 0.009 | 0.0002 | 55.7% |
| M10 | | CATBoost | 0.009 | 0.0002 | 56.0% |
| M11 | | LightGBM | 0.009 | 0.0001 | 59.2% |
| M12 | | Random Forest | 0.009 | 0.0002 | 56.0% |
| M13 | PV-to-roof ratio *with energy policy predictors* | XGBoost | 0.008 | 0.0001 | **71.1%** |
| M14 | | CATBoost | 0.008 | 0.0001 | 66.0% |
| M15 | | LightGBM | 0.008 | 0.0001 | 66.0% |
| M16 | | Random Forest | 0.009 | 0.0001 | 61.8% |

the ratio of solar panels to roof area ranges from 0% to 26%. It is important to note that any solar panels detected in images that do not include roofs are not considered in the estimation process. This study focuses exclusively on the deployment of rooftop solar PV systems on buildings, excluding ground-mounted and utility-scale solar installations.

Table 1 presents the predictive performance of the 16 ML models. The hyperparmeters used for each model is presented in Appendix Table B1. Among the models predicting the PV count per household, Model 5, which incorporates energy policy factors and is trained on XGBoost, achieves the highest $R^2$ value of 68.5%. Regarding the models predicting the PV-to-roof ratio, Model 13, also including energy policy factors and trained on XGBoost, achieves the best performance in terms of $R^2$. This XGBoost model, with energy policy factors as features, explains approximately 71.1% of the variance in the PV-to-roof ratio. To visualize the impact of the number of iterations (i.e., the number of trees in the random forest) on the performance of the best-performing XGBoost models (Model 5 and Model 13), Fig. 6 displays the learning curves in terms of Root Mean Square Error (RMSE) for M5 (PV-count-per-HH) (Figure 6a and M13 (PV-to-roof ratio model) (Figure 6b).

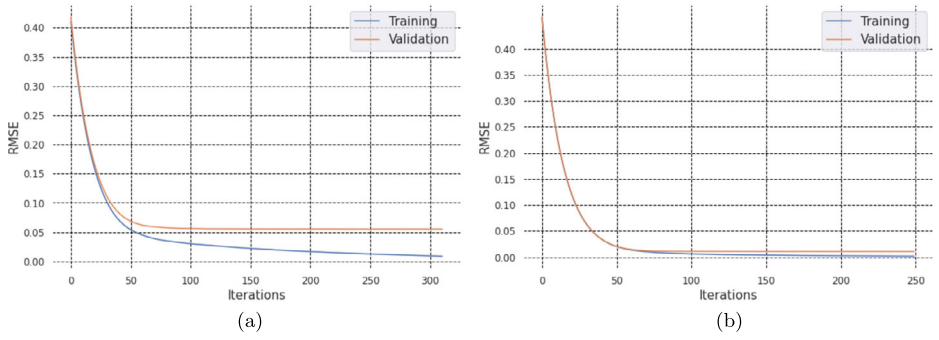


**Figure 6** Learning curves

#### *4.1.1 PV count per household prediction*

Figure 7 displays the ranking of the aggregated standardized feature importance scores (FIS) from the eight models that predict solar PV count per household. Within each model, the FIS values were standardized to a range between 0 and 1. The aggregation of FIS scores was weighted by the $R^2$ value of each model. Additionally, the bivariate correlation coefficient between each predictor and the target variable is indicated with colors to denote statistical significance.

Based on the aggregated FIS (Feature Importance Scores), the percentage of Democratic presidential voters (% Dem. Votes) emerges as the most influential predictor of PV count per household. Block groups with a higher percentage of Democratic voters tend to have a greater number of PVs per household. The second most significant predictor is hail risks, where block groups with a higher hail risk tend to have a lower number of PVs per household. The third most important predictor is the timeline for obtaining solar PV permits and completing the pre-installation process. Block groups with longer expected durations between the first permit submission and pre-installation completion typically exhibit a smaller PV count per household.

Figure 8 displays the SHAP values for the 20 most important features in predicting PV count per household, obtained from XGBoost and LightGBM models. SHAP values quantify the impact of each feature on the target value. A positive SHAP value indicates a positive impact on the target value, while a negative SHAP value indicates the opposite. The colors in the figure represent the relative contribution of each data point (observation) in predicting the outcome. Red pixels indicate higher importance, indicating a stronger impact of the feature, while blue pixels indicate the opposite.

#### *4.1.2 PV-to-roof ratio prediction*

Figure 9 displays the ranking of aggregated standardized FIS from the eight models predicting PV-to-roof area ratio in each block group. The standardized FIS is weighted by the explained variance ($R^2$) of each model. The bivariate correlation coefficient between each predictor and the target variable is annotated with different colors to indicate the level of statistical significance.

The average number of bedrooms, serving as a proxy for the average size of houses, emerges as the most influential predictor of the PV-to-roof ratio. As expected, a higher average number of bedrooms is negatively correlated with the PV-to-roof ratio since larger houses tend to have more roof area available for solar panels. The second most significant

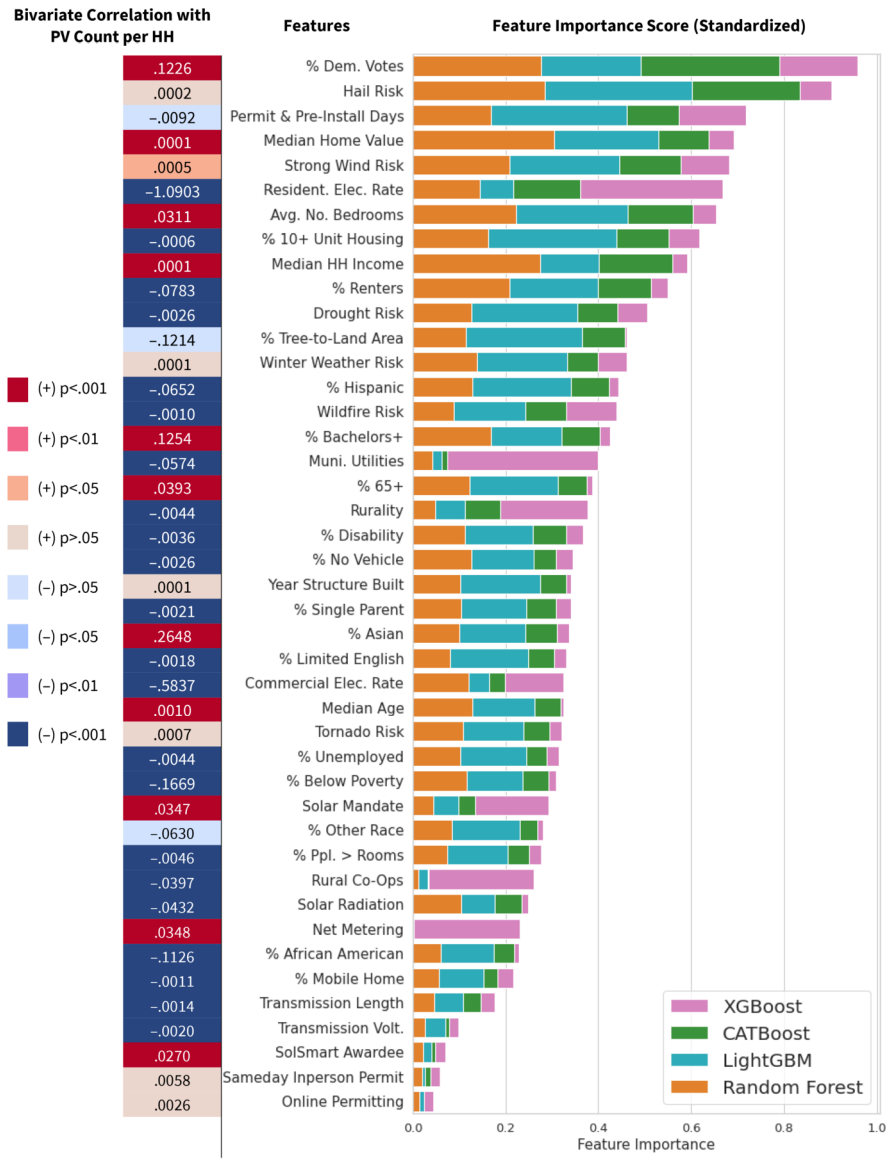


**Figure 7** Aggregated Feature Importance Scores (FIS) for 43 Factors predicting PV count per household

predictor is the duration of the permit and pre-installation process, meaning the PV-to-roof ratio is smaller in block groups which have longer timelines for getting the permits and pre-installation of PV projects. The percentage of renters constitutes the third most important predictor. A higher percentage of renters is associated with a larger PV-to-roof ratio. This finding aligns with expectations as neighborhoods with a greater proportion of rental units often have smaller roof areas available for solar panel installation.

Figure 10 reports SHAP values for 20 of the most important features in predicting PV-to-roof ratio from XGBoost and LightGBM models, respectively. The relative importance and the impact of each predictor are explained below.

Combining the results from the aggregated FIS (Figs. 7 and 9) and the SHAP values (Figs. 8 and 10), the following Sects. 4.2–4.5 summarize how the four groups of input features – (1) natural environment, (2) demographics and built environment, (3) energy

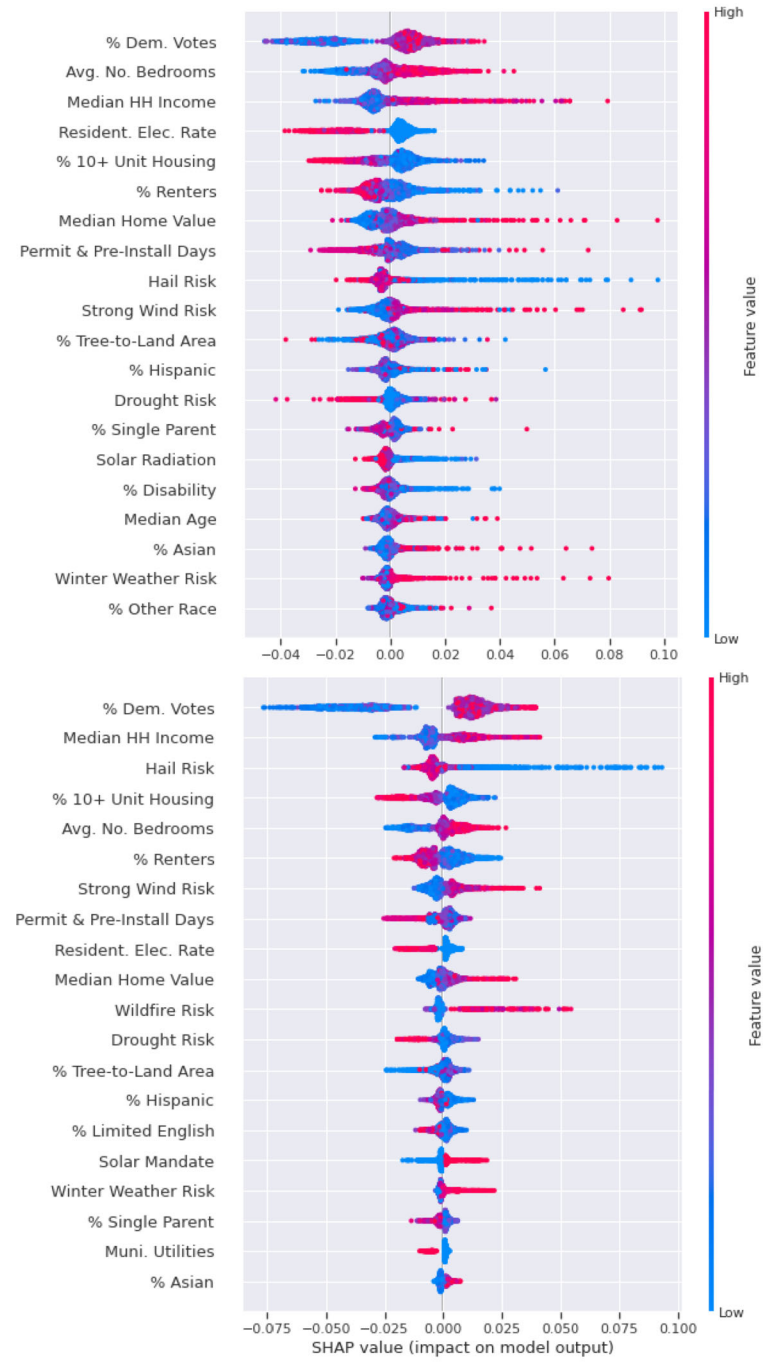


**Figure 8** SHAP Values of Most Important Features Predicting PV Count per Household from XGBoost (upper) and LightGBM (lower) Models (Models 5 and 7)

infrastructure, market, and policy, and (4) social and natural disaster vulnerabilities – predict the two measures of solar PV deployment. Additionally, Appendix Figure S1 compares the aggregated FIS between PV-count-per-HH models and PV-to-roof-ratio models across all 43 predictors.

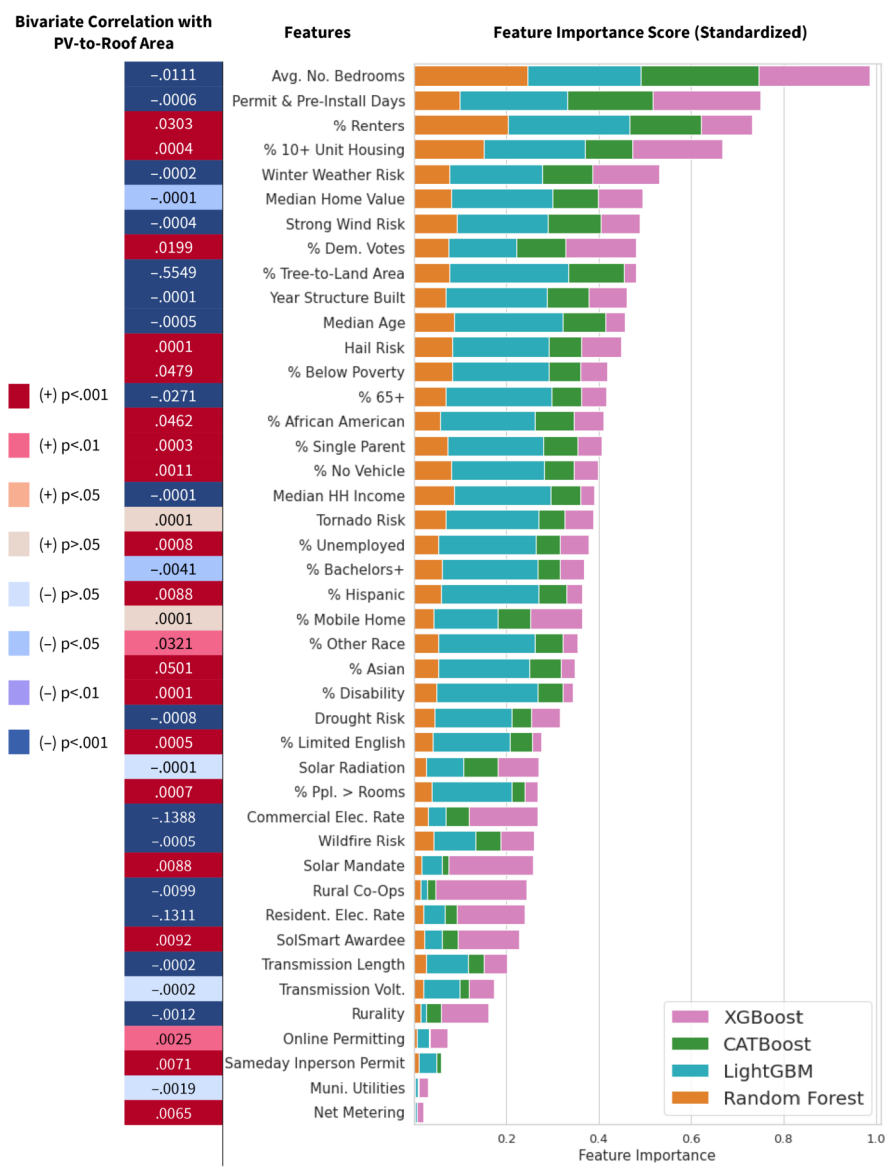


**Figure 9** Aggregated Feature Importance Scores (FIS) for 43 factors predicting PV-to-Roof ratio

## 4.2 Natural environment

*PV count per household* The proportion of tree-to-land area has a negative relationship with both PV count per household and the PV-to-roof ratio. This implies that block groups with greater tree coverage are less likely to adopt PV systems. Solar radiation shows a slight negative and nonlinear association with PV count per household (Appendix Table C1). However, it is important to note that this finding should not be generalized beyond the context of Colorado. The majority of areas within the state boast high solar potential, with annual average direct normal irradiance (DNI) ranging from 4.5 to 7.5. These values are 2 to 3 times higher than the world average DNI. Consequently, any location in Colorado is considered suitable for solar PV deployment in terms of solar radiation. The areas with extremely high DNI levels are typically arid and have relatively fewer residential

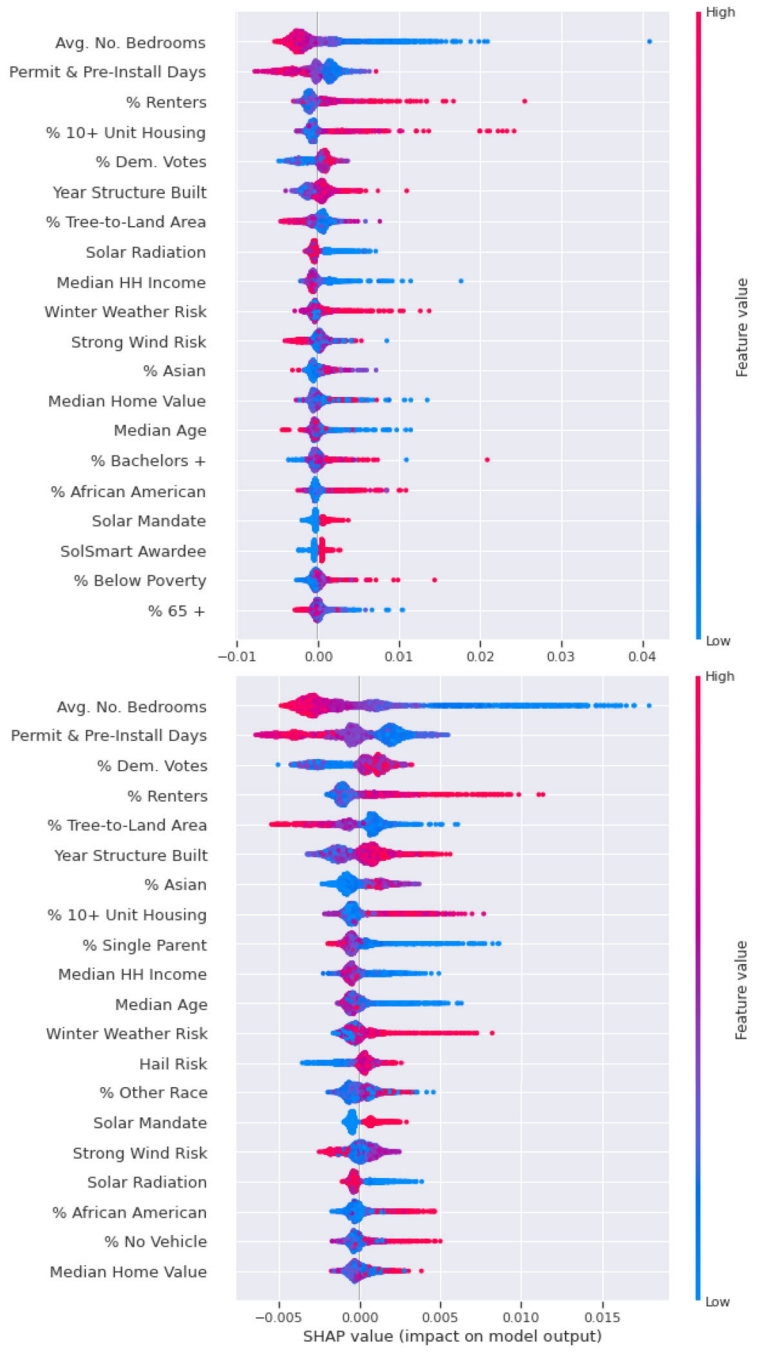


**Figure 10** SHAP Values of Most Important Features Predicting PV-to-Roof Ratio from XGBoost (upper) and LightGBM (lower) Models (Models 13 and 15)

properties. This circumstance may help explain the slight negative association observed between PV count per household and solar radiation in these particular areas.

*PV-to-roof ratio* Overall, the findings regarding natural environment features are consistent with the analysis of PV count per household. The proportion of tree-to-land area exhibits a negative association with the PV-to-roof ratio, indicating that block groups with

greater tree coverage tend to have smaller PV systems relative to the size of their roofs. This negative relationship remains statistically significant even when accounting for other factors (Model 34 in Appendix Table C5). Solar radiation also demonstrates a negative predictive relationship with the PV-to-roof ratio. However, it is important to note that this finding should not be generalized beyond Colorado. The state's abundant solar radiation levels make it an exceptional case, ensuring highly effective energy generation in most areas of the state.

### 4.3 Demographics and built environments

*PV count per household* As depicted in Fig. 7, the analysis reveals that median household income holds the 9th position as a significant feature and displays a positive association with PV count per household. This observation aligns with previous studies indicating that solar deployment tends to be more prevalent in neighborhoods characterized by higher income levels [5]. While median household income does not emerge as the most influential predictor among the 43 factors considered, other income-related variables, such as median home value and the number of bedrooms, rank prominently in the comprehensive feature importance scheme and exhibit higher SHAP values. Consistently across all models, median home value and the number of bedrooms exhibit a positive impact on PV count per household. These findings are consistent with prior research, which indicates that median home values positively predict solar PV installations at the neighborhood level [49].

There are significant disparities in solar deployment across block groups concerning race and ethnicity. The proportion of Asians demonstrates a positive correlation with PV count per household, whereas the proportion of Hispanics exhibits a negative association. Our supplementary statistical analysis yields similar findings (see Appendix Table C2). Even when accounting for median household income, home value, and house sizes, block groups with a higher proportion of African Americans tend to have a lower PV count per household. Moreover, the results from the supplementary analysis, utilizing Model 20 (see Appendix Table C2), which incorporates interaction terms between race and median household income, indicate varying marginal effects of income on PV count per household across neighborhoods with differing racial compositions. Figure 11 visually represents that an increase in the Asian and White populations correlates with positive average marginal

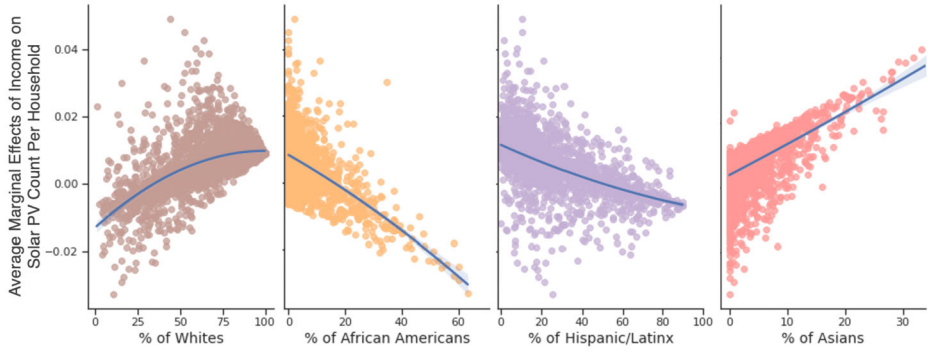


**Figure 11** The average marginal effects of median household income on PV count per household across neighborhoods with varying racial compositions

effects of income on PV count per household, while an increase in the Black and Hispanic populations exhibits a distinct trend.

Political ideology, as measured by the percentage of Democratic voters, exhibits a positive correlation with PV count per household across all four ML models. The share of Democratic party votes consistently emerges as the most impactful or second-most impactful feature across all models. This discovery aligns with prior research indicating that households with rooftop solar installations are more likely to align with the Democratic party [50]. Furthermore, residents' education level, measured by the proportion of individuals with bachelor's degrees or higher, positively predicts PV count per household, which is in line with existing literature [5]. Additionally, our analysis reveals that rural areas tend to exhibit lower PV count per household. However, when considering the aggregated importance of the rurality variable, it ranks 19th out of 43 predictors, indicating a medium influence on the models' predictions (see Fig. 7).

*PV-to-roof ratio* The findings related to demographics and the built environment should be interpreted cautiously due to the potential association between predictors and both PV installations and roof areas. In such cases, the direction of SHAP values for a predictor is determined by which of the two factors, PV or roof areas, is more influenced by the predictor. To illustrate, consider the scenario where median home value positively predicts PV areas in a neighborhood. The PV-to-roof ratio may exhibit a negative association with median home value because areas with higher median home values tend to have significantly larger roof areas. Therefore, it is important to note that the impact of predictors can be influenced by the relative influence they have on PV installations versus roof areas.

The proportions of renters and multi-dwelling unit houses emerge as the third and fourth most important and positive predictors of the PV-to-roof ratio. This suggests that neighborhoods with a higher concentration of renters and multi-dwelling units tend to have a larger proportion of roof areas covered by PV installations. Conversely, median household income and home values exhibit a negative association with the PV-to-roof ratio. This implies that the positive correlation between income or home values and roof sizes outweighs the negative association between these factors and the PV-to-roof ratio.

The proportions of Black and Asian populations within neighborhoods explains some of the variability in the PV-to-roof ratio. Analysis using SHAP values from the XGBoost model reveals a bimodal pattern in the PV-to-roof ratio among block groups with varying proportions of Black residents. Neighborhoods with high or low shares of Black residents tend to exhibit a higher PV-to-roof ratio, whereas neighborhoods with a moderate share of Black residents tend to have a lower ratio. Similarly, both the SHAP values and statistical analyses indicate that block groups with a higher proportion of Asians tend to display a greater PV-to-roof ratio. In contrast to the findings from the PV-count-per-HH analysis, the interaction between income and racial composition appears to have less influence in this particular context.

Consistent with the earlier findings concerning PV count per household, block groups with a higher proportion of Democratic residents also demonstrate a higher PV-to-roof ratio. This alignment suggests that neighborhoods with a greater concentration of Democratic residents tend to have a larger proportion of roof areas covered by PV installations. Additionally, both the ML models and supplementary statistical analysis (see Appendix Table C2) indicate that the PV-to-roof ratio is higher in areas with newer homes or a higher

average year of construction. This implies that neighborhoods with more recently built structures tend to exhibit a greater proportion of roof areas utilized for PV installations.

### 4.4 Energy infrastructure, market, and policy

*PV count per household* While the lengths and aggregated voltage of transmission lines in block groups show a negative relationship with PV count per household (as indicated in Appendix Table C3), the feature importance scores for these factors are relatively low across all ML models. Transmission length ranks 39th out of 43 features, while voltage ranks 40th out of 43 features. This suggests that while there may be a negative correlation between transmission line characteristics and PV count per household as shown in the statistical models (Appendix Table C3), these factors do not significantly contribute to the overall predictive power of the ML models.

Based on the results from statistical analysis and SHAP analysis, we find that neighborhoods located within municipal utility service areas and rural co-ops tend to have a lower PV count per household compared to those in investor-owned utility service areas. However, when energy policy variables are taken into account, this disparity across utility service areas becomes statistically insignificant in the models (as indicated in Appendix Table C3), and the feature importance scores associated with utility ownership types are relatively low Fig. 8. These findings suggest that the disparities observed in PV deployment are more likely influenced by utility policies and soft costs, such as PPI rules, rather than the inherent differences between the types of utilities themselves.

Among the energy policy variables examined, the duration of permit and pre-install days, as well as the presence of a solar mandate, emerge as the two most influential predictors of rooftop solar adoption. The median number of business days between the initial permit submission to the local jurisdiction and the pre-install interconnection is found to have a negative impact on PV count per household. This suggests that longer timelines for solar permit approval may discourage solar adoption in those areas. On the other hand, the results provide strong evidence that neighborhoods located in cities or counties with solar mandates exhibit a higher number of PV installations per household. Net metering is found to have a statistically significant and positive effect on solar deployment. However, the relative importance of net metering as a predictor is relatively low, likely because most utilities in Colorado already offer net metering programs. These findings on energy policy variables align with previous literature emphasizing the significance of soft costs associated with solar PV installations, such as the solar approval process, on the adoption of solar PV installations [51].

*PV-to-roof area* The average timeline for permits and pre-installation activities emerges as the second most important and negative predictor of the PV-to-roof ratio. Neighborhoods with shorter permit and pre-install timelines tend to have a higher proportion of their roof areas covered by PV systems. This suggests that not only do PV permitting and inspection timelines influence the decision to adopt rooftop solar, but they also affect the size of the installed PV systems.

Although solar mandates do not exhibit as high of a feature importance score as other predictors according to the SHAP analysis (Fig. 10), their positive impact on the PV-to-roof ratio remains evident. Other variables related to energy infrastructure, market dynamics, and policy factors (such as net metering, online permitting, transmission voltage and length, and utility ownership) rank lower in the feature importance spectrum.

However, comparatively speaking, these energy-related factors seem to have a more pronounced impact on PV system adoption compared to the PV-to-roof ratio.

### 4.5 Social and natural disaster vulnerabilities

*PV count per household* Several social vulnerability variables have a significant impact on predicting solar PV count per household. The percentage of housing in structures with 10 or more units (% 10+ Unit Housing) emerges as one of the most important predictors according to both the SHAP analysis (Fig. 8) and the aggregated feature importance score (FIS) analysis (Fig. 7). In addition, the share of single parents negatively predicts PV count per household based on the SHAP analysis (Fig. 8). The proportion of persons with disabilities also shows a negative association with PV count per household in the SHAP analysis, specifically based on the LightGBM model. Other social vulnerability factors, such as the proportion of households below the poverty level and households without vehicles, appear to be less influential compared to the aforementioned variables.

Interestingly, natural disaster vulnerabilities emerge as one of the most important group of predictors across all ML and statistical models and analyses. However, the direction of effects varies depending on the type of natural disaster. Based on the results from both machine learning (ML) models and statistical models (Appendix Table C4), areas with higher risks of strong wind and winter weather positively predict PV count per household. This suggests that neighborhoods with more frequent power outages caused by wind and snowstorms are more likely to adopt PV systems. This finding is supported by both the aggregated FIS (Fig. 7) and the SHAP analysis (Fig. 8). On the other hand, block groups with higher hail risks are less likely to deploy PV systems, indicating concerns over potential hail damage to solar panels may act as a barrier to rooftop PV adoption. The direction of effects for drought and wildfire risks is inconclusive, but both factors rank higher than the median in terms of the feature importance.

*PV-to-roof area* In contrast to the analysis of PV count per household, our findings do not provide strong evidence to suggest that social vulnerabilities have a significant impact on PV sizes. The only consistent and notable findings are that block groups with a higher proportion of households below the poverty line and those residing in 10+ unit housing tend to have a higher PV-to-roof ratio. This could be attributed to the fact that these two features are strongly correlated with the size of the houses themselves. According to the ML models and SHAP analysis, block groups with a greater share of single-parent households exhibit a lower PV-to-roof ratio. However, this finding is not supported by the supplementary statistical analysis (Appendix Table C4). The findings for other social vulnerability variables, such as the proportion of individuals with disabilities, limited English proficiency, no vehicles, households with more people than rooms, unemployed individuals, are inconclusive based on the aggregated feature importance score (FIS) analysis (Fig. 9), SHAP analysis (Fig. 10), and supplementary statistical analysis (Appendix Table C4)

Regarding natural disaster vulnerability features, winter weather risk and wind risk emerge as important predictors of the PV-to-roof ratio. Block groups with a higher winter weather risk tend to have a higher PV-to-roof ratio, indicating that residents in areas prone to frequent snowstorms are more likely to deploy larger PV systems in relation to the size of their roofs. The second most important natural disaster vulnerability feature is

strong wind risk, although the direction of its effect remains inconclusive. Based on the SHAP analysis (Fig. 10), neighborhoods with medium wind risk display a lower PV-to-roof ratio compared to those with high and low wind risks. The supplementary analysis in Appendix Table C4 also suggests that wind risk is associated with a lower proportion of PV to roof area. Finally, the findings for other natural disaster vulnerability factors, including drought, wildfire, hail, and tornado risks, remain inconclusive.

## 5 Discussion

The advancement of real-time data flow has created numerous opportunities for innovative research addressing important social issues. These issues include the development of sustainable and healthy cities [52–54], the promotion of energy efficiency [55], the enhancement of emergency management [56], and the fostering of social justice [57]. Notably, the availability and quality of satellite imagery have significantly improved, enabling the detection and analysis of distributed energy resources (DERs) such as solar PV, wind generating units, and outdoor battery storage in near real-time.

By leveraging the recent advancements in ML, computer vision, and public satellite imagery services, this study demonstrates how data science can contribute to developing a deeper understanding of household-level renewable energy deployment. Our novel model predicting neighborhood-level (block group-level) solar PV deployment achieves around a 70% $R^2$. Through the analysis of spatial data and the application of advanced algorithms, this study provides implications for developing sustainable, resilient, and equitable energy systems that align with the broader goals of creating environmentally friendly and socially just communities.

Our findings highlight the importance of local-level energy policies in promoting rooftop solar adoption. Across all models, our results consistently indicate that reducing permit and pre-installation timelines significantly enhances rooftop solar deployment. On the other hand, online and same-day in-person permits are relatively less influential. Soft costs, which encompass non-hardware expenses such as permitting, financing, and knowledge acquisition, play a critical role as they constitute a substantial portion (50 to 70 %) of the total installed PV system price [58]. Therefore, city and county governments seeking to encourage rooftop solar adoption should focus on reducing barriers in the permitting and interconnection processes, particularly by addressing time availability and information gaps that may vary across different social groups and geographic and socioeconomic circumstances.

This study also offers practical implications for energy justice and social equity. State and local government energy regulations often promote rooftop solar as a means to increase the use of DERs, improve grid reliability, lower utility bills, and yield associated health and social benefits. However, in line with previous research [17, 19], we find disparities in solar deployment rates among individuals with limited English proficiency, African American or Hispanic backgrounds, and lower incomes compared to their English-speaking, Asian, White, and higher-income counterparts. These disparities can be attributed to various interconnected factors, including homeownership rates, internet access, information and knowledge gaps, and the ability to participate in environmental decision-making processes. Considering that low-income households, people of color, and renters are disproportionately vulnerable to energy-related challenges [59, 60], our findings emphasize the need for state and local governments to recognize underlying social and demographic conditions that may hinder rooftop solar adoption.

Furthermore, as cities are increasingly adopting new building codes and permitting regulations to encourage rooftop solar systems, the findings of this study have implications for understanding the beneficiaries and potential inequities that may emerge during the transition to renewable energy. Present solar policies generally favor individuals and groups who can afford solar, own their homes, and possess middle- to upper-income levels. For example, solar mandates have the potential to reduce the costs of PV system installation by incorporating solar considerations into building design or construction.

Our study also provides insights into climate- and environment-specific strategies for the development of DERs. Based on our findings, several natural factors emerge as significant predictors of PV deployment in Colorado, namely tree canopy cover, hail risks, winter weather risks, strong wind risks, and tornado risks. While these conditions are difficult to change, they should be considered for making grid infrastructure investments to enhance community resilience and integrate diverse renewable resources into the energy mix. For instance, regions prone to power outages caused by heavy snowfall may have a greater interest in solar systems, and investing in large-scale battery storage could prove advantageous for bolstering power grid resilience in such areas. On the other hand, locations with high risks of hail or tornadoes may be less suitable for rooftop solar installations, necessitating the exploration of alternative DER options. As advancements in solar panel technology have made solar panels more resilient to hailstorms and high winds, educating potential PV adopters about the durability and safety of solar panels could help promote PV deployment in these regions. By considering the specific environmental conditions and tailoring DER strategies accordingly, policymakers and stakeholders can optimize the deployment of DERs, taking into account factors that affect system performance, reliability, and long-term viability.

Our study has a few limitations. Although the models we developed in this study can be applied to most areas in the United States and beyond, the empirical application presented in this study is specific to Colorado, USA in the year 2021. As such, direct applicability of the findings to other states or countries may require appropriate modifications and adjustments. For instance, when applying our ML models to states in the southern regions, it may be important to incorporate data related to hurricanes and floods, as they may have a significant impact on solar PV deployment patterns and outcomes. By incorporating such specific regional data, the models can be tailored to better capture the unique dynamics and factors influencing solar PV deployment in those areas.

Furthermore, to investigate the effects of input features on solar deployment over time, multiple years of satellite imagery data reflecting the evolving patterns of PV deployment diffusion would be necessary. Future studies should aim to measure spatial disparities in solar deployment at the block or block-group level over multiple years to fully understand the underlying dynamics of the spatial distribution of solar PV deployment. Although our XGBoost model achieved higher performance, some feature importance scores differ from the other three models. To further validate the robustness of our predictive models, future studies should expand the spatio-temporal coverage of the solar deployment data.

It is worth noting that we encountered several challenges during the training of Faster-RCNN models [61, 62]. The RPN layer exhibited poor performance in localizing small-sized objects due to the coarse-grained nature of satellite images. Our model experienced similar behavior, as shown in Appendix E Figure S2, which provides an example of training to identify roofs. Additionally, the RPN layer faced difficulties in distinguishing PVs from

background effects such as shadows and contrast changes on roofs, as there is limited color variation, resulting in some false detections. Furthermore, the choice of Intersection over Union (IoU) threshold could potentially impact the model's performance [63]. The available manually annotated data for satellite detection was relatively small, which could potentially lead to overfitting issues. Lastly, our model outputs rectangular frames around the objects of interest, while our objective is to calculate the ratio of the object's area to that of the entire image. This introduces another source of error, as the object of interest may not occupy the predicted frame entirely. Some studies [64, 65] have attempted to address issues related to dense, small, and arbitrarily rotated objects. Future research could focus on addressing these challenges to enhance the performance of computer vision models in detecting roof and PV areas.

Lastly, we do not claim that we have captured the true and complete effects of all input features on solar PV deployment patterns. Our primary objective was to develop a high-performing predictive model based on the 43 predictors, rather than providing a comprehensive explanation of how each predictor relates to solar PV deployment and to what extent. To uncover the genuine underlying effects of all input features, we recognize the need for a more sophisticated analysis that incorporates various interactions between these features. For instance, the impact of median household income on solar PV deployment could be influenced by specific policies or regulations, such as PPI rules or net metering, as well as factors like tree canopy coverage. Similarly, the influence of electricity prices on solar PV deployment might be contingent upon the proportion of renters or the age of houses. Therefore, future studies that aim to explore the complex relationships between predictors, rather than solely focusing on building precise predictive models, should consider incorporating interaction terms into their analyses. This will enable a deeper understanding of these intricate relationships and provide insights into the true effects of the input features on solar PV deployment patterns.

## 6 Conclusion

Using the RCNN computer vision algorithm, this study developed a model that can identify solar panels and roofs in satellite imagery. By analyzing 652,795 satellite images containing rooftops and solar panels in Colorado, we estimated the proportion of households that had installed at least one solar panel and the extent of roof area covered by solar panels at the US Census block and block-group levels. Our RCNN model achieved a mAP of 0.95 for measuring roof areas in the satellite images and an mAP of 0.81 for measuring solar panel areas.

Our computer vision model developed in this study has the potential to generate one of the most detailed solar deployment datasets across the United States. It can provide block-level (typically less than 100 residents) aggregated information on the rate of solar PV deployment, measured by two key metrics: PV Count per Household and PV-to-Roof ratio. Thus, our model enables a fine-grained analysis of solar deployment patterns, taking into account the characteristics of individual blocks.

The other aim of the study was to develop a highly accurate ML model capable of predicting the deployment of rooftop solar based on demographic, environmental, and energy/policy characteristics of neighborhoods. To accomplish this, two solar PV deployment metrics (PV Count per Household and PV-to-Roof ratio) were aggregated at the block group level, which typically consists of 600 to 3000 residents. A total of 43 input features were collected and predictive models for solar PV deployment were built using four

ML algorithms: Random Forest, CATBoost, LightGBM, and XGBoost. The ML model that performed the best was XGBoost, achieving an $R^2$ value of 68.5% for the PV Count per Household model and 71.1% for the PV-to-Roof ratio model. To the best of our knowledge, our novel ML model achieves a higher explained variance ($R^2$) than any existing solar deployment prediction models that utilize statistical and/or ML methods. Our ML model can be easily customized and deployed in other states within the United States as well as potentially in other countries around the world.

We found that lower hail risk, a higher share of Democratic party votes, stronger wind risks, higher median home value, and shorter rooftop solar permitting timelines were the five most significant predictors of a higher PV Count per Household. Regarding the PV-to-Roof ratio, we discovered that shorter rooftop solar permitting timelines, a lower proportion of renters and multifamily housing, and higher winter weather risks were the most influential features predicting a higher PV-to-Roof ratio.

Other ML models aiming to predict solar PV deployment may enhance their performance by incorporating the features that have proven particularly impactful in our model. For instance, our model uniquely considers vulnerabilities to natural disasters, such as hail risks, high winds, and wildfires, as predictors of rooftop PV deployment, which emerged as crucial factors in our analysis. Furthermore, political ideology emerged as an important predictor of PV adoption. Policymakers, researchers, and planners interested in constructing highly predictive models to forecast future solar deployment could consider including the input features that performed well in our model.

In summary, this work offers three key contributions:

- *Introducing a robust RCNN-based computer vision model for measuring roof and solar panel sizes*. This model can be deployed to assess solar PV deployment in any location worldwide.
- *Presenting a novel, precise model for forecasting future solar PV deployment based on 43 predictors in Colorado and beyond*. Our model explains about 70% of the variation in solar PV deployment. This makes it a valuable tool for forecasting future solar PV deployment in the United States and beyond. The model can be used to support more efficient and equitable grid infrastructure investment and distributed energy resource management.
- *Providing insights for business and policy decision making*. The insights from this research can be used to develop targeted marketing campaigns to promote solar PV deployment, as well as policies that encourage solar PV deployment and mitigate disparities in deployment. These insights can also be used to improve the efficiency of solar PV deployment and to make it more accessible to a wider range of people.

## Appendix A: Summary statistics

**Table A1** Summary statistics

| | Count | Mean | STD | Min | Max |
|---|---|---|---|---|---|
| PV Count per HH | 3441 | 0.070 | 0.093 | 0.001 | 0.784 |
| PV-to-Roof Ratio | 3441 | 0.025 | 0.020 | 0.001 | 0.259 |
| % Tree-to-Land Area | 3441 | 0.006 | 0.007 | 0.001 | 0.052 |
| Solar Radiation | 3441 | 5.851 | 0.430 | 4.712 | 7.236 |
| Median HH Income | 3441 | 77,571 | 36,263 | 14,145 | 250,000 |
| Median Age | 3441 | 39.608 | 9.067 | 17.7 | 84.6 |
| % 65 + | 3441 | 0.268 | 0.133 | 0.001 | 1.000 |
| % Bachelors + | 3441 | 0.283 | 0.164 | 0.001 | 0.855 |
| % Renters | 3441 | 0.333 | 0.253 | 0.001 | 1.000 |
| Year Structure Built | 3441 | 1977 | 17 | 1939 | 2014 |
| Avg. No. Bedrooms | 3441 | 2.872 | 0.664 | 0.466 | 4.524 |
| Median Home Value | 3441 | 354,098 | 200,610 | 10,000 | 2,000,000 |
| Rurality | 3441 | 2.098 | 2.429 | 1.000 | 9.000 |
| % Dem. Votes | 3441 | 0.553 | 0.154 | 0.109 | 0.796 |
| % African American | 3441 | 0.035 | 0.073 | 0.000 | 0.633 |
| % Hispanic | 3441 | 0.213 | 0.197 | 0.000 | 0.923 |
| % Asian | 3441 | 0.027 | 0.043 | 0.000 | 0.422 |
| % Other Race | 3441 | 0.032 | 0.042 | 0.000 | 0.973 |
| Transmission Volt. | 3441 | 1.952 | 2.724 | 0.000 | 8.166 |
| Transmission Length | 3441 | 4.739 | 5.791 | 0.000 | 14.455 |
| Muni. Utilities | 3441 | 0.096 | 0.295 | 0.000 | 1.000 |
| Rural Co-Ops | 3441 | 0.215 | 0.411 | 0.000 | 1.000 |
| Resident. Elec. Rate | 3441 | 0.122 | 0.019 | 0.062 | 0.212 |
| Commercial Elec. Rate | 3441 | 0.103 | 0.017 | 0.076 | 0.260 |
| Solar Mandate | 3076 | 0.271 | 0.444 | 0.000 | 1.000 |
| Net Metering | 3441 | 0.919 | 0.273 | 0.000 | 1.000 |
| SolSmart Awardee | 2339 | 0.501 | 0.500 | 0.000 | 1.000 |
| Online Permit | 2339 | 0.754 | 0.431 | 0.000 | 1.000 |
| Sameday InPerson Permit | 2339 | 0.435 | 0.496 | 0.000 | 1.000 |
| Permit & Pre-Install Days | 2339 | 16.267 | 6.002 | 8.000 | 35 |
| Drought Risk | 3441 | 1.756 | 3.285 | 0.000 | 29.35 |
| Wildfire Risk | 3441 | 4.712 | 8.509 | 0.000 | 48.921 |
| Hail Risk | 3441 | 28.292 | 13.027 | 2.583 | 64.351 |
| Winter Weather Risk | 3441 | 13.037 | 9.868 | 0.000 | 62.890 |
| Strong Wind Risk | 3441 | 17.077 | 7.332 | 4.107 | 59.476 |
| Tornado Risk | 3441 | 30.703 | 9.831 | 5.345 | 56.195 |
| % Below Poverty | 3441 | 0.105 | 0.099 | 0.000 | 0.847 |
| % Disability | 3441 | 11.281 | 5.050 | 0.400 | 44.6 |
| % Single Parent | 3441 | 7.884 | 4.738 | 0.000 | 27.600 |
| % Limited English | 3441 | 3.028 | 4.317 | 0.000 | 37.700 |
| % 10+ Unit Housing | 3441 | 14.635 | 18.589 | 0.000 | 98.900 |
| % Mobile Homes | 3441 | 4.329 | 8.638 | 0.000 | 79.100 |
| % Ppl. > Rooms | 3441 | 2.767 | 3.303 | 0.000 | 24.800 |
| % No Vehicle | 3441 | 5.217 | 5.310 | 0.000 | 43.300 |
| % Unemployed | 3441 | 4.861 | 2.995 | 0.000 | 28.400 |

## Appendix B: Hyperparameters used in predictive models

**Table B1** Hyperparameters used in each ML algorithm

| Model | Dataset | Algorithm | Hyperparameters |
|---|---|---|---|
| M1 | PV Count per HH | XGBoost | 'gamma': 0, 'alpha': 12, 'learning_rate': 0.027, 'seed': 712 'colsample_bytree': 0.3, 'reg_lambda': 1,'random_state': 700, 'n_estimators': 299, 'base_score': 0.29, 'max_depth': 7 |
| M2 | | CATBoost | 'l2_leaf_reg': 2, 'learning_rate': 0.1, 'depth': 9, 'iterations': 150 |
| M3 | | LightGBM | 'objective': 'regression', 'metric': 'rmse','is_unbalance': 'true', 'is_training_metric': 'true', 'boosting': 'gbdt', 'num_leaves': 36, 'feature_fraction': 0.99, 'bagging_fraction': 0.69, 'bagging_freq': 4, 'learning_rate': 0.01, 'max_depth': 15, 'max_bin': 23 |
| M4 | | RandomForest | 'n_estimators': 19, 'max_depth': 150, 'min_samples_split': 2, 'max_features': "sqrt",'min_samples_leaf': 2, 'random_state': 531 |
| M5 | PV Count per HH + Energy Policy | XGBoost | 'gamma': 0, 'alpha': 5, 'learning_rate': 0.05, 'random_state': 185, 'colsample_bytree': 0.5, 'reg_lambda': 0, 'n_estimators': 311, 'base_score': 0.5, 'max_depth': 7, 'seed': 855 |
| M6 | | CATBoost | 'l2_leaf_reg': 2, 'learning_rate': 0.1, 'depth': 6, 'iterations': 200 |
| M7 | | LightGBM | 'objective': 'regression', 'metric': 'rmse','is_unbalance': 'true', 'is_training_metric': 'true', 'boosting': 'gbdt', 'num_leaves': 36, 'feature_fraction': 0.81, 'bagging_fraction': 0.91, 'bagging_freq': 20, 'learning_rate': 0.021, 'max_depth': 14, 'max_bin': 23 |
| M8 | | RandomForest | 'n_estimators': 700, 'max_depth': 150, 'min_samples_split': 2, 'max_features': "sqrt",'min_samples_leaf': 2, 'random_state': 372 |
| M9 | PV-to-Roof Ratio | XGBoost | 'gamma': 0, 'alpha': 12, 'learning_rate': 0.025, 'seed':712 'colsample_bytree': 0.35, 'reg_lambda': 1, 'random_state': 789, 'n_estimators':300, 'base_score': 0.5, 'max_depth': 8 |
| M10 | | CATBoost | 'l2_leaf_reg': 1, 'learning_rate': 0.09, 'depth': 10, 'iterations': 200 |
| M11 | | LightGBM | 'objective': 'regression', 'metric': 'rmse','is_unbalance': 'true', 'is_training_metric': 'true', 'boosting': 'gbdt', 'num_leaves': 45, 'feature_fraction': 0.25, 'bagging_fraction': 0.75, 'bagging_freq': 4, 'learning_rate': 0.01, 'max_depth': 15, 'max_bin': 52 |
| M12 | | RandomForest | 'n_estimators': 300, 'max_depth': 64, 'min_samples_split': 3, 'max_features': sqrt, 'min_samples_leaf': 2, 'random_state': 435 |
| M13 | PV-to-Roof Ratio + Energy Policy | XGBoost | 'gamma': 0, 'alpha': 5, 'learning_rate': 0.05, 'seed': 1164 'colsample_bytree': 0.5, 'reg_lambda': 0,'random_state': 185, 'n_estimators': 500, 'base_score': 0.52, 'max_depth': 9 |
| M14 | | CATBoost | 'l2_leaf_reg': 1, 'learning_rate': 0.09, 'depth': 6, 'iterations': 150 |
| M15 | | LightGBM | 'objective': 'regression', 'metric': 'rmse','is_unbalance': 'true', 'is_training_metric': 'true', 'boosting': 'gbdt', 'num_leaves': 36, 'feature_fraction': 0.34, 'bagging_fraction': 0.75, 'bagging_freq': 4, 'learning_rate': 0.01, 'max_depth': 15, 'max_bin': 23 |
| M16 | | RandomForest | 'n_estimators': 300, 'max_depth': 280, 'min_samples_split': 2, 'max_features': sqrt, 'min_samples_leaf': 2, 'random_state': 42 |

## Appendix C: Sub-analysis linear regression models

**Table C1** Statistical analysis: Tree canopy and solar radiation

| | PV Count per HH | PV-to-Roof Ratio |
|---|---|---|
| Predictors | Model 17<br>Coef. (SE) | Model 18<br>Coef. (SE) |
| % Tree-to-Land Area | 2.636 (0.523)*** | −0.537(0.111)*** |
| % Tree-to-Land Area$^2$ | −116 (17.59)*** | −1.111 (3.744) |
| Solar Radiation (log) | −0.053 (0.004)*** | −0.001 (0.001)* |
| Solar Radiation (log)$^2$ | −0.106 (0.008)*** | −0.004 (0.002)* |
| Number of Block Groups | 3441 | 3441 |
| $R^2$ | 0.052 | 0.041 |

*** $p < 0.001$. Robust standard errors (SE) are in parentheses.

**Table C2** Statistical analysis: Demographics and built environments

| | PV Count per Household | | PV-to-Roof Ratio | |
|---|---|---|---|---|
| Predictors | Model 19<br>Coef. (SE) | Model 20<br>Coef. (SE) | Model 21<br>Coef. (SE) | Model 22<br>Coef. (SE) |
| Median HH Income | 0.001 (0.001) | 0.001 (0.001) | −0.001 (0.001)*** | −0.001 (0.001) |
| Median HH Income$^2$ | −0.001 (0.001) | | 0.001 (0.001)*** | |
| Median Age | −0.001 (0.001) | −0.001 (0.001) | −0.001 (0.001)*** | −0.001 (0.001)*** |
| % 65+ | 0.010 (0.018) | 0.009 (0.018) | 0.004 (0.004) | 0.005 (0.004) |
| % Bachelor+ | −0.014 (0.018) | −0.015 (0.018) | 0.004 (0.004) | 0.001 (0.004) |
| % Renters | −0.049 (0.012)*** | 0.049 (0.012)*** | 0.004 (0.003)$^\dagger$ | 0.006 (0.003)* |
| Year Structure Built | 0.100 (0.017)*** | 0.105 (0.017)*** | 0.011 (0.004)*** | 0.011 (0.004)*** |
| Year Structure Built$^2$ | −0.001 (0.001)*** | −0.001 (0.001)*** | −0.001 (0.001)*** | −0.001 (0.001)*** |
| Avg. No. Bedrooms | 0.011 (0.005)* | 0.011 (0.005)* | −0.009 (0.001)*** | −0.009 (0.001)*** |
| Median Home Value | 0.001 (0.001)*** | 0.001 (0.001)*** | 0.001 (0.001)** | 0.001 (0.001)** |
| Rurality | −0.001 (0.001)$^\dagger$ | −0.001 (0.001)$^\dagger$ | −0.001 (0.001)*** | −0.001 (0.001)*** |
| % Dem. Voters | 0.107 (0.014)*** | 0.115 (0.014)*** | 0.002 (0.003) | 0.001 (0.003) |
| % African American | −0.101 (0.022)*** | 0.420 (0.561) | 0.014 (0.005)** | 0.083 (0.115) |
| % Hispanic | −0.026 (0.011)* | 0.409 (0.179)* | −0.002 (0.002) | 0.073 (0.037)* |
| % Asian | 0.220 (0.037)*** | −3.245 (0.818)*** | 0.021 (0.008)** | 0.036 (0.168) |
| % Other Race | 0.008 (0.037) | −0.575 (0.657) | 0.010 (0.008) | 0.001 (0.135) |
| Income x Afri. Ameri. | | −0.048 (0.051) | | −0.006 (0.010) |
| Income x Hispanic | | −0.041 (0.017)* | | −0.007 (0.003)* |
| Income x Asian | | 0.310 (0.073)*** | | −0.001 (0.015) |
| Income x Other Race | | 0.053 (0.061) | | 0.001 (0.013) |
| No. of block groups | 3441 | 3441 | 3441 | 3441 |
| $R^2$ | 0.143 | 0.149 | 0.202 | 0.198 |

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^\dagger p < 0.1$. Robust Standard Errors are in parentheses.

**Table C3** Statistical analysis: Energy infrastructure, market, and policy

| | Solar PV Count per Household | | PV-to-Roof Ratio | |
|---|---|---|---|---|
| Predictors | Model 23<br>Coef. (SE) | Model 24<br>Coef. (SE) | Model 25<br>Coef. (SE) | Model 26<br>Coef. (SE) |
| Transmission Volt. | −0.003 (0.001)** | 0.001 (0.003) | 0.001 (0.001) | −0.001 (0.001) |
| Transmission Length | 0.001 (0.001) | −0.001 (0.001) | −0.001 (0.001) | 0.001 (0.001) |
| Muni. Utilities | −0.048 (0.005)*** | −0.068 (0.009)*** | −0.005 (0.001)*** | 0.004 (0.003) |
| Rural Co-ops. | −0.046 (0.005)*** | −0.005 (0.011) | −0.008 (0.001)*** | 0.002 (0.004) |
| Resident. Elec. Rate | −0.839 (0.140)*** | 0.150 (0.348) | 0.003 (0.029) | 0.582 (0.106)*** |
| Commerc. Elec. Rate | 1.067 (0.165)*** | −2.06 (0.648)** | 0.043 (0.034) | −0.123 (0.139) |
| Solar Mandate | | 0.001 (0.009) | | 0.006 (0.002)** |
| Net Metering | | 0.027 (0.012)* | | 0.008 (0.008)** |
| SolSmart Awardee | | 0.003 (0.007) | | −0.001 (0.001) |
| Online Permitting | | 0.004 (0.005) | | −0.001 (0.001) |
| Sameday Inperson Per. | | 0.002 (0.007) | | −0.004 (0.001) |
| Per. & Pre-Install Days | | −0.001 (0.004)* | | −0.001 (0.001)*** |
| CONTROLS | YES | YES | YES | YES |
| No. of block groups | 3441 | 2328 | 3441 | 2328 |
| $R^2$ | 0.216 | 0.277 | 0.234 | 0.325 |

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{\dagger} p < 0.1$. Robust Standard Errors are in parentheses.

CONTROLS include Income, Median Age, % 65+, % Bachelor+, % Renters, Year Structure Built, Avg. No. Bedrooms, Median Home Value, Rurality, % Dem. Voters, % African American, % Hispanic, % Asian, % Other Race, % Tree-to-Land Area, Solar Radiation.

**Table C4** Statistical analysis: Social and natural disaster vulnerabilities

| | PV Count Per Household | | PV-to-Roof Ratio | |
|---|---|---|---|---|
| Predictors | Model 27<br>Coef. (SE) | Model 28<br>Coef. (SE) | Model 29<br>Coef. (SE) | Model 30<br>Coef. (SE) |
| Drought Risk | 0.001 (0.001) | | 0.001 (0.001) | |
| Wildfire Risk | −0.001 (0.001)*** | | −0.001 (0.001)* | |
| Hail Risk | −0.001 (0.001)$^{\dagger}$ | | 0.001 (0.001) | |
| Winter Weather Risk | 0.001 (0.001) | | 0.001 (0.001)*** | |
| Strong Wind Risk | 0.002 (0.001)*** | | −0.002 (0.001)*** | |
| Tornado Risk | −0.002 (0.001)*** | | −0.003 (0.001)*** | |
| % Below Poverty | | 0.001 (0.001) | | 0.015 (0.001)* |
| % Disability | | −0.001 (0.001) | | −0.001(0.001) |
| % Single Parent | | 0.001 (0.001) | | −0.001(0.001) |
| % Limit. English | | 0.001 (0.001) | | 0.001 (0.001) |
| % 10+ Unit Housing | | −0.001 (0.001)*** | | 0.001 (0.001)*** |
| % Mobile Home | | 0.001 (0.001) | | 0.001 (0.001)** |
| % Ppl. > Rooms | | −0.002 (0.001)*** | | −0.001 (0.001)*** |
| % No Vehicles | | −0.001 (0.001) | | 0.001 (0.001) |
| % Unemployed | | 0.001 (0.001) | | 0.001 (0.001)*** |
| CONTROLS | YES | YES | YES | YES |
| No. of block groups | 3441 | 3441 | 3441 | 3441 |
| $R^2$ | 0.181 | 0.174 | 0.231 | 0.242 |

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{\dagger} p < 0.1$. Robust Standard Errors are in parentheses.

CONTROLS include Income, Median Age, % 65+, % Bachelor+, % Renters, Year Structure Built, Avg. No. Bedrooms, Median Home Value, Rurality, % Dem. Voters, % African American, % Hispanic, % Asian, % Other Race, % Tree-to-Land Area, Solar Radiation.

**Table C5** Statistical analysis: All predictors

| | PV Count Per Household | | PV-to-Roof Ratio | |
|---|---|---|---|---|
| Predictors | Model 31<br>Coef. (SE) | Model 32<br>Coef. (SE) | Model 33<br>Coef. (SE) | Model 34<br>Coef. (SE) |
| % Tree-to-Land Area | 0.389 (0.545) | 2.057 (0.878)* | −0.550 (0.115)*** | −0.476 (0.175)** |
| % Tree-to-Land Area$^2$ | −51.098 (17.448)* | −145.512 (36.07)** | 2.666 (3.669) | 1.140 (7.177) |
| Solar Radiation | 0.068 (0.095) | 0.207 (0.224) | 0.005 (0.020) | −0.054 (0.045) |
| Solar Radiation$^2$ | −0.006 (0.008) | −0.018 (0.019) | 0.001 (0.002) | 0.004 (0.004) |
| Median HH Income | 0.001 (0.001) | 0.001 (0.001)** | 0.001 (0.001) | −0.001 (0.001) |
| Median HH Income$^2$ | −0.001 (0.001) | −0.001 (0.001)* | 0.001 (0.001)* | 0.001 (0.001) |
| Median Age | −0.001 (0.001) | 0.001 (0.001) | −0.001 (0.001)*** | −0.001 (0.001)** |
| % 65+ | 0.001 (0.017) | −0.017 (0.021) | 0.001 (0.004) | 0.003 (0.004) |
| % Bachelor+ | 0.032 (0.018)$^\dagger$ | −0.023 (0.024) | 0.004 (0.004) | −0.003 (0.005) |
| % Renters | −0.026 (0.012)* | −0.039 (0.015)** | 0.003 (0.003) | 0.001 (0.003) |
| Year Structure Built | 0.120 (0.017)*** | 0.089 (0.022)*** | 0.017 (0.004)*** | 0.010 (0.004)* |
| Year Structure Built$^2$ | −0.001 (0.001)*** | −0.001 (0.001)*** | −0.001 (0.001)*** | −0.001 (0.001)* |
| Avg. No. Bedrooms | 0.002 (0.005) | 0.001 (0.006) | −0.007 (0.001)*** | −0.009 (0.001)*** |
| Median Home Value | 0.001 (0.001)*** | 0.001 (0.001)*** | 0.001 (0.001)*** | 0.001 (0.001)*** |
| Rurality | −0.004 (0.001)** | 0.004 (0.003) | −0.001 (0.001)*** | −0.001 (0.001) |
| % Dem. Voters | 0.082 (0.016)*** | 0.166 (0.029)*** | 0.010 (0.003)** | 0.013 (0.006)* |
| % African American | −0.019 (0.023) | −0.041 (0.026) | 0.017 (0.005)*** | 0.007 (0.005) |
| % Hispanic | −0.0145 (0.013) | −0.012 (0.018) | −0.004 (0.003) | −0.004 (0.004) |
| % Asian | 0.205 (0.036)*** | 0.140 (0.041)*** | 0.017 (0.008)* | 0.005 (0.009) |
| % Other Race | 0.076 (0.036)* | 0.068 (0.052) | 011 (0.007) | 0.017 (0.012) |
| Transmission Volt. | −0.004 (0.001)*** | −0.001 (0.003) | 0.001 (0.001) | 0.001 (0.001) |
| Transmission Length | 0.001 (0.001) | 0.001 (0.002) | −0.001 (0.001) | −0.001 (0.001) |
| Muni. Utilities | −0.055 (0.006)*** | −0.061 (0.009)** | −0.004 (0.001)*** | −0.004 (0.002)$^\dagger$ |
| Rural Co−ops. | −0.046 (0.005)*** | −0.013 (0.011) | −0.007 (0.001)*** | −0.006 (0.002)** |
| Resident. Elec. Rate | −0.921 (0.142)*** | 0.429 (0.351) | 0.013 (0.030) | 0.235 (0.070)*** |
| Commerc. Elec. Rate | 0.975 (0.167)*** | −1.246 (0.644)$^\dagger$ | 0.015 (0.035) | −0.351 (0.128)** |
| Solar Mandate | | 0.015 (0.009)$^\dagger$ | | 0.005(0.002)** |
| Net Metering | | 0.033 (0.011)** | | 0.006 (0.001)* |
| SolSmart Awardee | | −0.004 (0.007) | | −0.001 (0.001) |
| Online Permitting | | -0.001 (0.006)* | | −0.001 (0.001) |
| Sameday Inp. Per. | | −0.008 (0.008) | | −0.001 (0.002) |
| Per. & Pre−Inst. Days | | −0.002 (0.001)*** | | −0.001 (0.001)*** |
| Drought Risk | −0.001 (0.001) | −0.011 (0.002)*** | 0.001 (0.001) | −0.001 (0.001)** |
| Wildfire Risk | −0.001 (0.001)*** | 0.001 (0.001)$^\dagger$ | −0.001 (0.001) | 0.001 (0.001)*** |
| Hail Risk | −0.001 (0.001)*** | −0.001 (0.001)** | 0.001 (0.001) | 0.001 (0.001) |
| Winter Weather Risk | 0.001 (0.001)* | −0.001 (0.001) | 0.001 (0.001)*** | 0.001 (0.001) |
| Strong Wind Risk | 0.002 (0.001)*** | −0.004 (0.001)*** | 0.001 (0.001) | 0.001 (0.001) |
| Tornado Risk | −0.002 (0.001)*** | −0.003 (0.001)*** | −0.001 (0.001)*** | 0.001 (0.001) |
| % Below Poverty | −0.001 (0.021) | −0.007 (0.027) | 0.014 (0.004)** | 0.015 (0.005) |
| % Disability | 0.001 (0.001) | −0.001 (0.001) | −0.001(0.001)$^\dagger$ | −0.001 (0.001)* |
| % Single Parent | 0.001 (0.001) | 0.001 (0.001) | -0.001(0.001) | −0.001 (0.001) |
| % Limit. English | −0.001 (0.001) | −0.001 (0.001) | 0.001 (0.001) | 0.001 (0.001) |
| % 10+ Unit Housing | −0.001 (0.001)*** | −0.001 (0.001)*** | 0.001 (0.001)*** | 0.001 (0.001)*** |
| % Mobile Homes | −0.001 (0.001) | 0.001 (0.001) | 0.001 (0.001)*** | −0.001 (0.001)*** |
| % Ppl. > Rooms | −0.001 (0.001) | 0.001 (0.001) | 0.001 (0.001)** | 0.001 (0.001)*** |
| % No Vehicle | −0.001 (0.001) | −0.001 (0.001) | 0.001 (0.001) | −0.001 (0.001) |
| % Unemployed | 0.001 (0.001) | 0.001 (0.001) | 0.001 (0.001)*** | 0.001 (0.001)* |
| No. of block groups | 3441 | 2328 | 3441 | 2328 |
| $R^2$ | 0.262 | 0.329 | 0.271 | 0.348 |

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^\dagger p < 0.1$. Robust Standard Errors are in parentheses.

## Appendix D: Aggregated FIS: PV count per household models vs. PV-to-roof ratio models

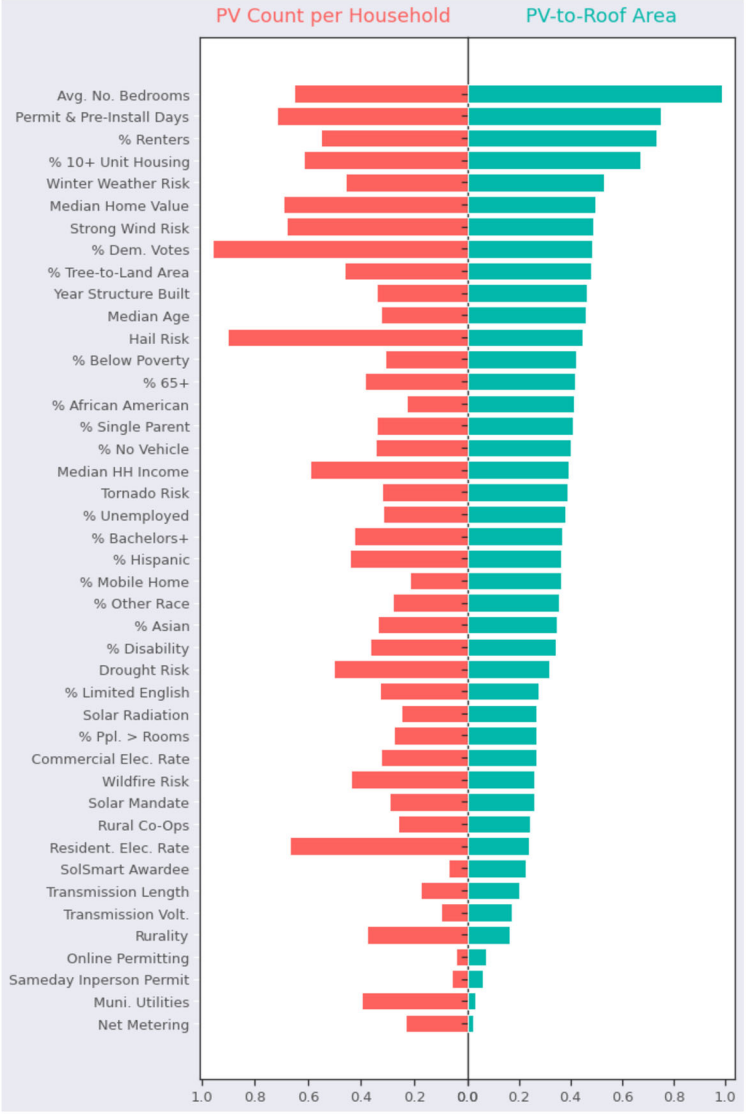


**Figure S1** Comparing FIS between the PV-count-per-household and PV-to-roof-ratio models

## Appendix E: A trial run illustrating accuracy obtained for identifying roofs

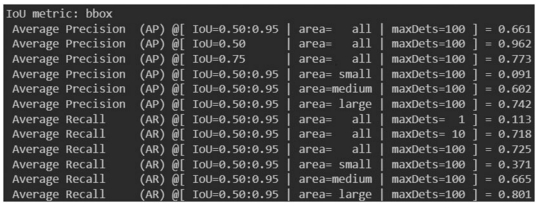


**Figure S2** A trial run illustrating accuracy obtained for identifying roofs

**Acknowledgements**
Most work of Koushik Ganesan was done before his employment at BestEx Research, Stamford, CT. Major work of Raven O'Rourke was done at the University of Colorado Denver. Authors thank the anonymous reviewers for their meticulous reading of our article and their invaluable suggestions, which have improved the quality and clarity of our work. We extend our appreciation to Lynne Kiesling and William Swann for their insightful feedback, which has enriched our manuscript. Any remaining errors in the paper are solely the responsibility of the authors.

**Funding**
This research was supported by the Presidential Initiative on Urban and Place-Based Research and the Education Through Undergraduate Research and Creative Activities program at the University of Colorado Denver.

**Abbreviations**
CDC, Centers for Disease Control and Prevention; DER, Distributed Energy Resources; dGEN, Distributed Generation Market Demand; DNI, Direct Normal Irradiance; FEMA, Federal Emergency Management Agency; FIS, Feature Importance Score; GIS, Geographical Information Systems; HH, Household; IoU, Intersection over Union; mAP, Mean Average Precision; ML, Machine Learning; NEM, Net Metering; NREL, National Renewable Energy Laboratory; PII, Permitting, Inspection, and Interconnection; PV, Photovoltaic; RPS, Renewable Energy Portfolio Standards; SHAP, SHapley Additive exPlanations; SVI, Social Vulnerability Index.

**Availability of data and materials**
The datasets and Python code used in data analysis and visualization in this research are available in a GitHub repository, https://github.com/SerenaYKim/SolarDeployment.

## Declarations

**Competing interests**
The authors declare that they have no competing interests.

**Author contributions**
SYK conceived the idea, secured the funding, collected data, performed data analysis, and wrote the manuscript. KG designed and implemented the computer vision model and wrote the manuscript. CS collected data and wrote the manuscript. RO implemented computer vision model and performed ML model experiments. All authors read and approved the manuscript.

**Author details**
[1]School of Public and International Affairs, North Carolina State University, 2221 Hillsborough St, Raleigh, NC 27607, USA. [2]College of Engineering, Design and Computing, University of Colorado Denver, 1200 Larimer St, Denver, CO 80204, USA. [3]School of Public Affairs, University of Colorado Denver, 1380 Lawrence St, Denver, CO 80204, USA. [4]Physics, University of Colorado Boulder, 2000 Colorado Ave, Boulder, CO 80309, USA. [5]RadiaSoft, 6525 Gunpark Dr, Suite 370-411, Boulder, CO 80301, USA.

Received: 17 July 2022 Accepted: 16 June 2023 Published online: 14 July 2023

**References**
1. International Energy Agency (2020) World Energy Outlook. https://iea.blob.core.windows.net/assets/a72d8abf-de08-4385-8711-b8a062d6124a/WEO2020.pdf
2. Solar.com What Is the Average Payback Period for Solar Panels? https://www.solar.com/learn/what-is-the-average-payback-period-for-solar-panels/. Published: 2023-01-20, Accessed: 2023-02-22
3. Barbose GL, Forrester S, O'Shaughnessy E, Darghouth NR (2021) Residential solar-adopter income and demographic trends: 2021 update. Technical report, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States)

4. O'Shaughnessy E, Barbose G, Wiser R, Forrester S, Darghouth N (2021) The impact of policies and business models on income equity in rooftop solar adoption. Nat Energy 6(1):84–91
5. Lukanov BR, Krieger EM (2019) Distributed solar and environmental justice: exploring the demographic and socio-economic trends of residential pv adoption in California. Energy Policy 134:110935
6. Yu J, Wang Z, Majumdar A, Rajagopal R (2018) Deepsolar: a machine learning framework to efficiently construct a solar deployment database in the United States. Joule 2(12):2605–2617
7. Ryan AJ, Donou-Adonsou F, Calkins LN (2019) Subsidizing the sun: the impact of state policies on electricity generated from solar photovoltaic. Econ Anal Policy 63:1–10
8. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. Advances in neural information processing systems 28
9. Saurav K, Jung S (2019) Robust detection of the coordinate of a solar panel using deep learning algorithm. In: 2019 19th international conference on control, automation and systems (ICCAS). IEEE, Los Alamitos, pp 1116–1119
10. Gupta J, Long A, Xu CK, Tang T, Shekhar S (2021) Spatial dimensions of algorithmic transparency: a summary. In: 17th international symposium on spatial and temporal databases, pp 116–125
11. Schelly C, Letzelter JC (2020) Examining the key drivers of residential solar adoption in upstate New York. Sustainability 12(6):2552
12. Korcaj L, Hahnel UJ, Spada H (2015) Intentions to adopt photovoltaic systems depend on homeowners' expected personal gains and behavior of peers. Renew Energy 75:407–415
13. Graziano M, Gillingham K (2015) Spatial patterns of solar photovoltaic system adoption: the influence of neighbors and the built environment. J Econ Geogr 15(4):815–839
14. Google LLC (2021) Project sunroof. https://sunroof.withgoogle.com
15. Sigrin B, Gleason M, Preus R, Baring-Gould I, Margolis R (2016) Distributed generation market demand model (dgen): documentation. Technical report, National Renewable Energy Lab. (NREL), Golden, CO (United States)
16. Schunder T, Yin D, Bagchi-Sen S, Rajan K (2020) A spatial analysis of the development potential of rooftop and community solar energy. Remote Sens Appl Soc Environ 19:100355
17. Sunter DA, Castellanos S, Kammen DM (2019) Disparities in rooftop photovoltaics deployment in the United States by race and ethnicity. Nat Sustain 2(1):71–76
18. Crago CL, Chernyakhovskiy I (2017) Are policy incentives for solar power effective? Evidence from residential installations in the northeast. J Environ Econ Manag 81:132–151
19. Reames TG (2020) Distributional disparities in residential rooftop solar potential and penetration in four cities in the United States. Energy Res Soc Sci 69:1–12
20. Thormeyer C, Sasse J-P, Trutnevyte E (2020) Spatially-explicit models should consider real-world diffusion of renewable electricity: solar pv example in Switzerland. Renew Energy 145:363–374
21. Tidwell JH, Tidwell A (2021) Decarbonizing via disparities: problematizing the relationship between social identity and solar energy transitions in the United States. Energy Res Soc Sci 77:102099
22. Mildenberger M, Howe PD, Miljanich C (2019) Households with solar installations are ideologically diverse and more politically active than their neighbours. Nat Energy 4(12):1033–1039
23. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
24. Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171
25. Girshick R (2015) Fast R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448. https://doi.org/10.48550/ARXIV.1504.08083
26. Gokul R, Nirmal A, Bharath K, Pranesh M, Karthika R (2020) A comparative study between state-of-the-art object detectors for traffic light detection. In: 2020 international conference on emerging trends in information technology and engineering (ic-ETITE). IEEE, Los Alamitos, pp 1–6
27. Alganci U, Soydas M, Sertel E (2020) Comparative research on deep learning approaches for airplane detection from very high-resolution satellite images. Remote Sens 12(3):458
28. Kingma DP, Ba J (2017) Adam: a method for stochastic optimization
29. Berardi U, Graham J (2020) Investigation of the impacts of microclimate on pv energy efficiency and outdoor thermal comfort. Sustain Cities Soc 62:102402
30. Staley DC (2013) Urban forests and solar power generation: partners in urban heat island mitigation. Int J Low Carbon Technol 10(1):78–86
31. Choia C, Berrya P, Smitha A (2021) The climate benefits, co-benefits, and trade-offs of green infrastructure: a systematic. J Environ Manag 291:112583
32. USD of Agriculture (USDA) (2013) Rural-urban continuum codes. Economic research service, US department of agriculture. Washington
33. MIT Election Data and Science Lab (2018) County presidential election returns 2000-2020. Harvard dataverse. https://doi.org/10.7910/DVN/VOQCHQ
34. NREL The utility rate database. National Renewable Energy Laboratory. Accessed 2021-11-23
35. Watson A, Guidice L, Lisell L, Doris L, Busche S (2012) Solar ready: an overview of implementation practices
36. O'Shaughnessy E, Dong S, Cook JJ, Cruce J, Ardani K, Fekete E, Margolis R (2022) Effects of local permitting and interconnection requirements on solar pv installation durations. Energy Policy 161:112734
37. Cook J, Cruce J, Fekete E, Dong S (2021) Solar permitting, inspection, and interconnection cycletimes and requirements. National Renewable Energy Laboratory. https://solarapp.nrel.gov/solarTRACE
38. For Disease Control, CC, Prevention (2021) CDC/ATSDR social vulnerability index. Agency for toxic substances and disease registry. https://www.atsdr.cdc.gov/placeandhealth/svi/index.html
39. FFEM Agency (2021) National Risk Index (NRI). FEMA. https://hazards.fema.gov/nri/
40. Jordahl K (2020) Geopandas: v0.8.1. Zenodo. https://doi.org/10.5281/zenodo.3946761
41. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
42. Dorogush AV, Ershov V, Gulin A (2018) Catboost: gradient boosting with categorical features support. arXiv preprint. arXiv:1810.11363

43. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: a highly efficient gradient boosting decision tree. Advances in neural information processing systems 30
44. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794
45. Shapley LS (1953) A value for n-person games. Princeton University Press, Princeton
46. Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144
47. Shrikumar A, Greenside P, Shcherbina A, Kundaje A (2016) Not just a black box: learning important features through propagating activation differences. arXiv preprint. arXiv:1605.01713
48. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30
49. Kwan CL (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 47:332–344
50. Mildenberger M, Howe PD, Miljanich C (2019) Households with solar installations are ideologically diverse and more politically active than their neighbours. Nat Energy 4(12):1033–1039
51. Hsu JH-Y (2018) Predictors for adoption of local solar approval processes and impact on residential solar installations in California cities. Energy Policy 117:463–472
52. Alshamsi A, Awad E, Almehrezi M, Babushkin V, Chang P-J, Shoroye Z, Toth A-P, Rahwan I (2015) Misery loves company: happiness and communication in the city. EPJ Data Sci 4(1):1
53. Lepri B, Antonelli F, Pianesi F, Pentland A (2015) Making big data work: smart, sustainable, and safe cities. Springer, Berlin
54. Daraei S, Pelechrinis K, Quercia D (2021) A data-driven approach for assessing biking safety in cities. EPJ Data Sci 10(1):11
55. Bogomolov A, Lepri B, Larcher R, Antonelli F, Pianesi F, Pentland A (2016) Energy consumption prediction using people dynamics derived from cellular network data. EPJ Data Sci 5:1
56. Hong L, Frias-Martinez V (2020) Modeling and predicting evacuation flows during hurricane irma. EPJ Data Sci 9(1):29
57. Tsvetkova M, Nilsson O, Öhman C, Sumpter L, Sumpter D (2016) An experimental study of segregation mechanisms. EPJ Data Sci 5(1):1
58. Ardani K, Seif D, Margolis R, Morris J, Davidson C, Truitt S, Torbert R (2013) Non-hardware ("soft") cost-reduction roadmap for residential and small commercial solar photovoltaics 2013–2020
59. Keady W, Panikkar B, Nelson IL, Zia A (2021) Energy justice gaps in renewable energy transition policy initiatives in Vermont. Energy Policy 159(2021):1–11
60. Hanke F, Guyet R, Feenstra M (2021) Do renewable energy communities deliver energy justice? Exploring insights from 71 European cases. Energy Res Soc Sci 80(2021):1–10
61. Deng Z, Sun H, Zhou S, Zhao J, Lei L, Zou H (2018) Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS J Photogramm Remote Sens 145:3–22
62. Tang T, Zhou S, Deng Z, Zou H, Lei L (2017) Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. Sensors 17(2):336
63. He Z, Zhang L (2019) Multi-adversarial faster-rcnn for unrestricted object detection. In: Proceedings of the IEEE CVF international conference on computer vision, pp 6668–6677
64. Yang X, Yang J, Yan J, Zhang Y, Zhang T, Guo Z, Sun X, Fu K (2019) Scrdet: towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE CVF international conference on computer vision, pp 8232–8241
65. Han X, Zhong Y, Zhang L (2017) An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. Remote Sens 9(7):666


## Publisher's Note