Energy & Buildings 287 (2023) 112965

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

Investigating the application of a commercial and residential energy consumption prediction model for urban Planning scenarios with Machine Learning and Shapley Additive explanation methods

Shideh Shams Amiri, Maya Mueller, Simi Hoque *

Drexel University, Philadelphia 19104, USA

ARTICLE INFO

Article history:
Received 17 November 2022
Revised 11 February 2023
Accepted 6 March 2023
Available online 9 March 2023

Keywords:
Energy consumption
Machine learning
Scenario planning
Bottom-up modeling, Commercial and residential building energy model
Residential energy
Commercial energy

ABSTRACT

Building energy forecasting methodologies utilized by municipal governments tend to be geared heavily towards depicting broader qualitative representations of regional change and are in need of complementary data-driven models that can produce quantitatively reliable depictions of future energy consumption at the neighborhood-level. The current research demonstrates an application of a Machine Learning (ML) model in the form of an Extreme Gradient Boosting (XGBoost) algorithm for forecasting the energy use of commercial and residential buildings. The methodology serves to improve on municipal scenario planning by providing a more spatially granular representation of future energy use. In this way, city government and urban planners can more accurately set carbon emission benchmarks and target specific locales for sustainability initiatives. The second major contribution of the study is to demonstrate how scenario planning approaches can utilize existing Machine Learning techniques to compensate for gaps in the data. This work is developed through a case study of Philadelphia. The study begins with the construction of residential and commercial energy models for the year 2015 and corresponding models for the year 2045. The forecast models integrate regional socioeconomic trends from the Delaware Valley Regional Planning Commission (DVRPC) scenario of Enduring Urbanism. The commercial energy use model is developed from the DVRPC's open-source Geographic Information System (GIS) datasets, the Commercial Buildings Energy Consumption Survey (CBECS), and CoStar commercial real estate data. The residential model applies the Residential Energy Consumption Survey (RECS), the Public Use Microdata Sample (PUMS), and Census Bureau American Community Survey (ACS) estimates. A corresponding SHAP (SHapley Additive exPlanations) analysis is implemented to pinpoint feature contributions to the model's energy estimates. By using the PopGen software, the model's energy estimates could be analyzed at the household level, the smallest possible scale. To provide a useful resource for key stakeholders, the study aggregates model output by Traffic Analysis Zone (TAZ) and Public Use Microdata Area (PUMA) to display a detailed forecast of energy use. The results indicate that the DVRPC Enduring Urbanism trends in income and employment do not significantly affect energy consumption for the study area. However, features related to lower building intensity (e.g., lower square footage, fewer floors per building) were associated with reduced energy use in both models. Additionally, the study found residential buildings under the "single-family attached" zoning designation to correspond with higher energy estimates.

1. Introduction

With the intention of future carbon neutrality, the City of Philadelphia set the initiative to halve Greenhouse Gas (GHG) emissions from the built environment by the year 2030 [22]. Targeting carbon reduction programs for the commercial and residential building sectors are of high priority, with commercial buildings and facilities currently being the single highest contributor to building-derived GHG emissions for the region and the residential sector the second highest (City of Philadelphia Office of Sustainability; 2019 Greenhouse Gas Inventory).

Predicting and mapping commercial building energy consumption remains a challenge, due in part to the lack of consistent data on built features and the dearth of information on energy use at spatially granular scales (i.e., the energy consumption of a single

* Corresponding author.
E-mail address: sth55@drexel.edu (S. Hoque).

https://doi.org/10.1016/j.enbuild.2023.112965
0378-7788/

building rather than a geographic aggregate). Moreover, building energy use tends to follow dynamic and non-linear relationships with influential variables and therefore developing accurate model predictions can rely heavily on data availability and quality [13]. Thus, researchers must customize a modeling approach depending on the nature of the existing constraints. Some important considerations include the level of data access, the geographic dimensions of input and output features, the temporal length of the forecast, and the contextual motivation for the research – i.e., whether there is more weight given to model development or prediction reliability.

Physical energy models, or white box models, utilize set equations to simulate the building load. These models are optimal for conditions where there exists extensive detail on the thermal performance of the building envelope, surrounding climate conditions, and building occupancy schedules [7,5]. As such, white box models are best for producing detailed results for a given building. Compared to other modeling approaches, white box models are notable for their capacity to perform reliably in the absence of historical data and their ability to achieve high prediction accuracy [12]. Due to the complexity and labor-intensive nature of the approach, white box models tend to be applied to one building or a small subset of the building stock, and are not ideal for depicting the spatial variability of energy use for a neighborhood, block, or broader region.

For the output energy prediction to be mapped in terms of larger spatial scales, statistical (i.e., non-ML) regression models can be a useful alternative for forecasting energy consumption given the presence of quality data on the building stock (e.g., [8,17,15]. Importantly, statistical energy prediction models require consistent building and energy use data such that input features are mapped to energy demand for each building within the dataset. As statistical models infer the relation between input and output factors, they are commonly applied in the context of developing a model that is applicable to a wide array of buildings and furthering an understanding of how input features influence energy use for a general building type [11]. If there is more weight given to prediction accuracy over model understanding, Machine Learning (ML) black box models (e.g., Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Neural Network (NN) algorithms) can serve as a powerful tool for reliable forecasts over a range of geographic scales (see literature reviews, [1,3].

Recently, Machine Learning (ML) has been shown to be particularly useful in the context of commercial and residential building energy modeling. ML is powerful in its capacity to compensate for unlabeled feature datasets. In this context, the term "unlabeled" refers to when data samples lack consistent information on a desired feature (say, the "label" of energy use, the desired feature, is absent in a dataset of buildings).

Often, researchers have labeled datasets aggregated for larger geographic areas, but lack consistently labeled data at the smaller geographic scale. In this case, ML can utilize common variables that exist across the labeled and unlabeled data (e.g., an assortment of common building features) and produce "labels" (the energy estimates) for the samples at the desired spatial scale. This technique is performed by training the ML model with a labeled dataset mapped to common explanatory features and then testing the trained ML model on unlabeled data.

For example, Jiang et al. [9] successfully applied a semi-supervised deep learning approach to predict commercial and residential building energy use intensity (EUI) in NYC. The model was trained with sociodemographic census data (e.g., population, income, employment) and building features (e.g., built year, land use, total floor square footage, number of floors). For the same region, Kontokosta (n.d.) tested three ML models (i.e., Ordinary Least Squares, Random Forest, and Support Vector Regression) to predict annual commercial and residential energy use at the building level, district level, and for the entire municipal area.

Zhang et al. [20] applied data from the Residential Energy Consumption Survey (RECS), Public Use Microdata Sample (PUMS), and American Community Survey (ACS) to predict residential energy demand. The RECS and PUMS datasets were statistically matched to develop a synthetic population of household residential energy use for the entire Atlanta Metropolitan region. The paper examined the ML models of Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression, Bagging, Random Forest, Support Vector Machine, AdaBoost, Gradient Boosting, and Extra Trees. The model output was found to be largely consistent with Georgia Power and Atlanta Gas Light electricity and natural gas use estimates.

Robinson et al. [16] trained and validated a series of ML algorithms (e.g., Linear Regression, Gradient Boosting, Random Forest, and more) on a commercial building energy estimation model for New York City, with Extreme Gradient Boosting (XGBoost) demonstrating superior performance in comparison to other Machine Learning techniques. The model was further tested on the Atlanta metropolitan area to develop building-level estimates for energy use aggregated by Traffic Analysis Zone (TAZ). The Commercial Buildings Energy Consumption Survey (CBECS) contains nationwide estimates for aggregated commercial building energy use by region and was thus utilized to train the prediction model.

The present work will integrate part of the methodology from Zhang et al. [20] for residential buildings and Robinson et al. [16] for the commercial building stock, with the novel contribution of using different features to develop an interpretable ML model and applying ML approaches for long-term prediction scenario planning. The current research aims to demonstrate how recent advances in ML can create more geospatially refined, quantitatively detailed forecasts of energy use. This methodology will be tested for data specific to the City of Philadelphia.

2. Methodology

The current research will apply an Extreme Gradient Boosting (XGBoost) Machine Learning (ML) model to predict building energy use for the City of Philadelphia's residential and commercial buildings. XGBoost was selected as the ML approach of choice due to its ability to perform well despite spurious (or non-causal) relationships between dependent and independent variables [19]. Building energy outcomes are associated with a wide array of influential features, such as climate conditions, physical building features (e.g., building site orientation and location, the building envelope, ventilation and insulation), and even broader economic and sociodemographic trends. Thus, it is critical that the ML model can perform without necessitating a correlational relationship between each input and output factor. As the learning algorithm combines the predictions of multiple base learners (i.e., learns from an ensemble), XGBoost can take into consideration complex interactions between input variables and the energy prediction, in addition to performing well on heterogeneous datasets. XGBoost is also beneficial in its ability to employ an ensemble sequential learning process where new models are adapted to take into consideration the residual error of prior model iterations, thus improving the accuracy and generalizability of the model output.

For the residential model, Public Use Microdata Sample (PUMS) and Residential Energy Consumption Survey (RECS) datasets are statistically matched according to the methodology proposed by Zhang et al. [20] and incorporated with Census Bureau American Community Survey (ACS) microdata to create an initial sample of household electricity energy use and demographic and socioeconomic characteristics (See the "Data Preparation" portion of

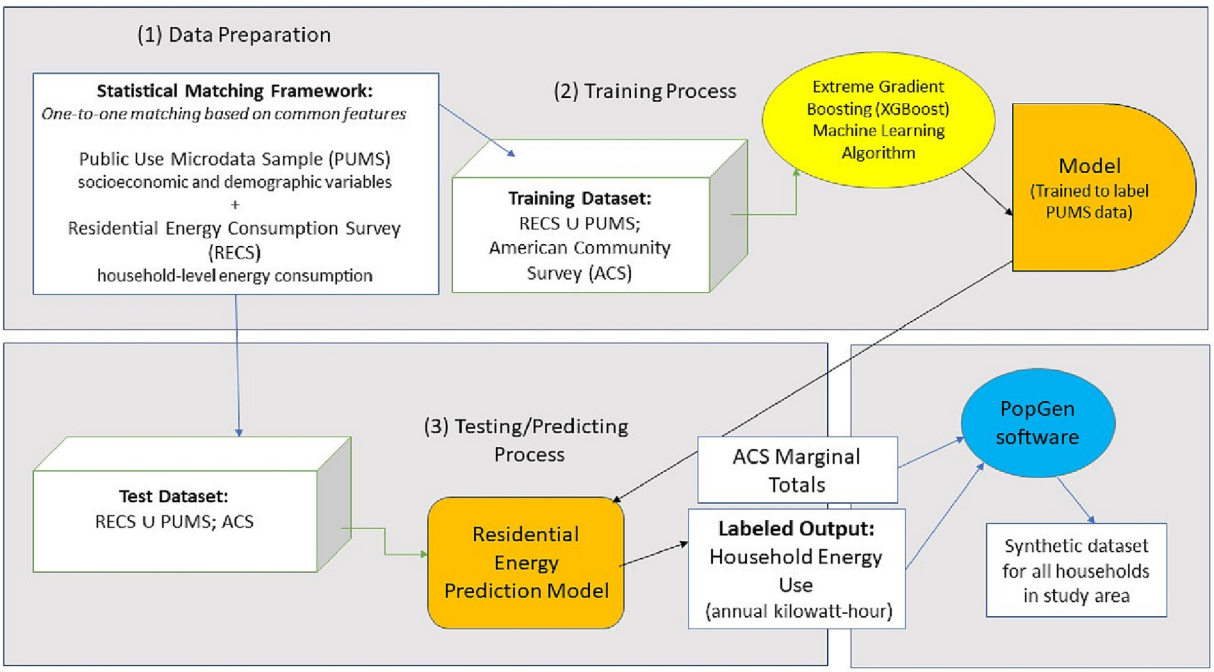

Fig. 1. Flowchart for residential model development with Public Use Microdata Sample (PUMS), Residential Energy Consumption Survey (RECS), and American Community Survey (ACS) estimates as data inputs. Labeled output in annual kilowatt-hour.

Fig. 1). These data inputs are then utilized for the training and validation of the Extreme Gradient Boosting (XGBoost) model.

With the household samples for predicted energy consumption, the PopGen software incorporates the XGBoost model output in conjunction with American Community Survey (ACS) marginal totals to develop a corresponding synthetic dataset representative of energy use for all households in the region (the final step to the right of the "Testing/Predicting Process" in Fig. 1). Household energy consumption data is mapped by the geographical scale of a Public Use Microdata Area (PUMA) with 12 total PUMAs within Philadelphia County and converted to annual kilowatt-hour (kWh) for the study area. See section 2.1.1 and 2.1.2 for further explanation on residential model data processing and residential model development.

With respect to the commercial model, the Commercial Buildings Energy Consumption Survey (CBECS) dataset is utilized to train an Extreme Gradient Boosting (XGBoost) Machine Learning model on regional energy use aggregates (See the "Data Preparation" and "Training Process" section of Fig. 2). The trained commercial model is then tested on unlabeled building-level data from the CoStar real estate database. As the CoStar database contains comprehensive information on commercial buildings for the region, it is not necessary to apply PopGen software on the model output. Building-level data is aggregated by Traffic Analysis Zone (TAZ) and Public Use Microdata Area (PUMA) with energy use in annual kilowatt-hour (kWh). See section 2.2.1 and 2.2.2 for further explanation on commercial model data processing and commercial model development.

Using Shapley, an Explainable Artificial Intelligence (XAI) tool, the output is analyzed to gain a better understanding of the model and to assess the validity of individual predictions for both the residential and commercial model. Importantly, the intended purpose of a SHAP analysis is to provide a *diagnostic* tool for the Machine Learning Extreme Gradient Boosting (XGBoost) model in terms of how the model output interacted with data inputs. For example, SHAP results that show a seemingly insignificant input feature having great predictive power on the model output would provide an indicator that the model performed unreliably. For this reason, SHAP results are not intended to provide causality between household and building characteristics and energy estimates. However, notable and strong trends across genres of data variables (say, bundles of related building characteristics or household features) in the SHAP results can still provide a valid barometer for identifying trends of interest and potential areas that warrant future research.

The Delaware Valley Regional Planning Commission (DVRPC) Enduring Urbanism scenario stipulates a selection of regional trends in real estate development and socioeconomic growth patterns across residential, transportation, and commercial sectors for the year 2045. As demonstrated in Fig. 3, Delaware Valley Regional Planning Commission (DVRPC) projections for the residential sector are applied to the residential energy model output in terms of forecasted shifts in household income and corresponding commercial sector forecasts are applied to the commercial model for employment trends. The scenario outputs for residential household and commercial building energy use are compared to the base case, or the previous Machine Learning outputs ("Labeled Output" from Fig. 3 for commercial, "Synthetic dataset for all households in study area" from Fig. 3 for residential). Shapley analyses for feature importance are performed for the corresponding scenario datasets. See Section 2.3.1 for more information on residential model scenario development and Section 2.3.2 for the commercial model scenario development.

2.1. Residential energy model

2.1.1. Residential energy model data and data processing

The Residential Energy Consumption Survey (RECS), the Public Use Microdata Sample (PUMS), and the American Community Survey (ACS) datasets were applied to develop the residential energy use model. The Residential Energy Consumption Survey (RECS) contains information on household-level energy consumption, demographic, and socioeconomic characteristics. Since the present analysis focuses on the City of Philadelphia, RECS records are fil-

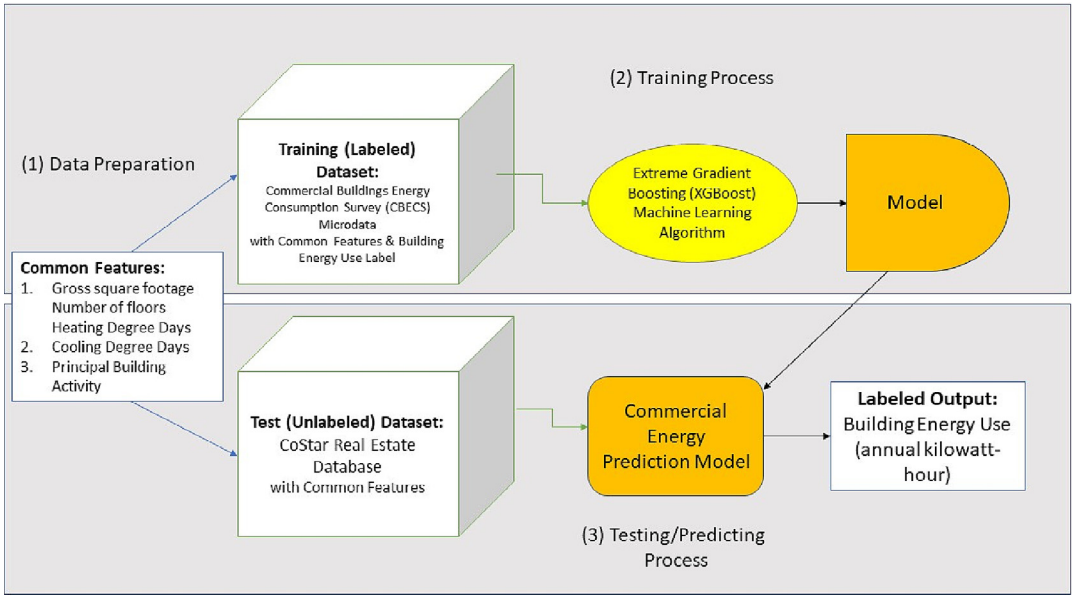

Fig. 2. Flowchart for commercial model development with the Commercial Buildings Energy Consumption Survey (CBECS) dataset and CoStar database as data inputs. Labeled output in annual kilowatt-hour (kWh).

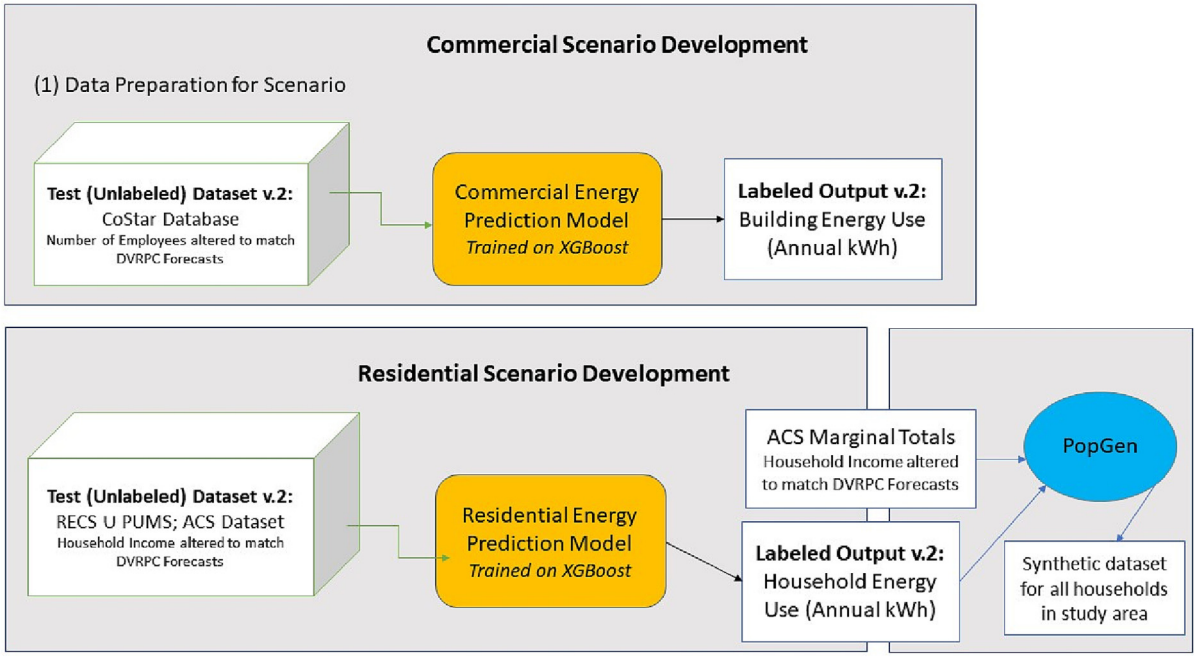

Fig. 3. Flowchart for the residential and commercial scenario model. Both forecast scenarios are constructed according to regional trends from the Delaware Valley Regional Planning Commission (DVRPC). Model output in annual kilowatt-hour (kWh).

tered for the Northeast regions and Middle Atlantic Census Divisions for the energy consumption model estimation.

The Public Use Microdata Sample (PUMS) dataset possesses a larger sample size relative to the Residential Energy Consumption Survey (RECS) dataset for the state of Pennsylvania and additionally contains population and housing unit features with individual data surveyed from the American Community Survey (ACS). While the RECS data contains more housing unit features, the PUMS data includes more demographic and socioeconomic features. Energy consumption data is not available in PUMS, but the reported energy bills and expenses are formulated to be representative of residential energy consumption.

The Census Bureau American Community Survey (ACS) data is used as a marginal distribution source to synthesize a complete population of households in the Philadelphia region. The ACS contains aggregated statistical data related to person-level, household, and housing information by different geographic levels (i.e., census tract, block group, and blocks).

2.1.2. Residential model development

The residential energy consumption model is generated in four steps. First, the Residential Energy Consumption Survey (RECS) and Public Use Microdata Sample (PUMS) datasets are joined using the STATMatch package provided by the R programming software. This

nearest neighbor distance hot deck macro-matching method searches in the Public Use Microdata Sample (PUMS) datasets for the nearest neighbor of each record in the Residential Energy Consumption Survey (RECS) dataset based on the distance measured on the numerical matching variables. The distance d (Equation (1) is measured as Manhattan Distance in this study – i.e., the default measurement in STATMatch.

Manhattan Distance is a desirable choice for one-to-one matching since the formula reduces the computational cost of the matching process and standardizes small numerical features ([26], [20]).

$$d = \sum_{i=1}^{n} \left| X_{M_i}^{RECS} - X_{M_i}^{PUMS} \right|$$

$$n = \text{total number of matching variables}$$

$$X_{M_i}^{RECS} = i^{th} \text{ matching variable from the RECS dataset}$$

$$X_{M_i}^{PUMS} = i^{th} \text{ matching variables from the PUMS dataset} \quad (1)$$

Statistical matching methods are used to integrate the Residential Energy Consumption Survey (RECS) data in Region 1 (the Northeast region) and Public Use Microdata Sample (PUMS) data for Pennsylvania to present a full set of variables containing residential energy consumption in addition to a wide range of household- and population-level socioeconomic variables in the same dataset [23].

The sample households in the Residential Energy Consumption Survey (RECS) data are not geolocated; Therefore, households in RECS and their corresponding electricity in kilowatt-hour (kWh) for Region 1 are first statistically matched with PUMS data (RECS ∪ PUMS) based on common variables. The shared features contain information on energy bills, housing units, unit structure, year-built, household income, and the number of rooms in both datasets.

The study applies the PopGen 2.0 software package for producing a synthetic dataset of households. The PopGen output serves as a comprehensive representation of residential electricity energy use for the study area. The household samples with labeled energy estimates and American Community Survey (ACS) marginal totals of common model features serve as the input dataset for the PopGen algorithm.

The PopGen software utilizes the matched Public Use Microdata Sample (PUMS) data from the Machine Learning (ML) model output. PopGen then constructs a seed matrix based on the joint distributions of households from PUMS. Marginal controls derived from the American Community Survey (ACS) dataset attribute weights to each household in the PUMS dataset.

These weights provide a basis for expanding the Public Use Microdata Sample (PUMS) data into a fully comprehensive, synthetic population for the region. Both household-level and person-level variables are controlled and matched in the synthesis process through the application of an iterative proportional updating (IPU) algorithm embedded within PopGen. The control variables are selected based on its correlational relationship with the target variable of energy use.

To finalize the matching variables, a Spearman rho correlation is applied. After determining which variables correlate highly with energy use, the Residential Energy Consumption Survey (RECS) and Public Use Microdata Sample (PUMS) variables are joined. With respect to synthesizing households in PopGen, the Machine Learning output of the matched Public Use Microdata Sample (PUMS) serves as the data for the sample inputs. American Community Survey (ACS) marginal totals are drawn in the geographic scale of the Municipal County District (MCD) and Traffic Analysis Zone (TAZ). Further information on the shared RECS and PUMS variables is available in Table S1 of the Supplementary Materials.

2.2. Commercial building energy model

2.2.1. Commercial building model data and data processing

Approximately every five years, the U.S. Energy Information Administration (EIA) releases Commercial Building Energy Consumption Survey (CBECS) microdata. The most recent dataset released from 2015 has 6,720 building samples where each row corresponds to all the commercial buildings estimated to exist in the United States ("Energy Information Administration (EIA)- Commercial Buildings Energy Consumption Survey (CBECS)" n.d.).

Commercial Building Energy Consumption Survey (CBECS) serves as the study's labeled dataset for the training model such that common features are mapped to aggregate building energy use. Each row contains details on building attributes from the CBECS Building Survey questionnaire. Notably, several CBECS features are difficult to apply for localized research due to having an inconsistent presence in the dataset and therefore future studies that utilize CBECS should exercise caution before determining model input features.

The following features are used in the training model: gross square footage (SQFT), Principal Building Activity (PBA), number of employees (NWKR), Heating Degree Days (HDD65), and Cooling Degree Days (CDD65). We applied the CBECS-trained model to the 8,649 commercial buildings in Philadelphia as downloaded from the CoStar real estate database. A key component of a building's energy consumption is the amount of energy used for heating and cooling. Heating and Cooling Degree Day (HDD65 and CDD65) variables, which have been shown to be useful indicators of energy demand [6], are present in the CBECS dataset. These variables are absent from the Costar dataset and are thus appended manually from the Oak Ridge Climate Models for the year 2019.

2.2.2. Commercial model development

For training the Machine Learning (ML) model, Commercial Building Energy Consumption Survey (CBECS) categorical and numerical building characteristics are used as input features and CBECS building energy consumption estimates in Major Fuel British Thermal Units (MFBTU) are applied as the output feature. All energy use values are converted from MFBTU to annual kilowatt-hour (kWh).

For the current study, the commercial building energy consumption prediction model is reformulated as an Extreme Gradient Boosting (XGBoost) problem along with a SHapley Additive exPlanations (SHAP) analysis of the output to interpret feature influence and assess the validity of individual predictions.

K-fold cross-validation was used for training the dataset to find the best model parameters and to avoid over-fitting. Grid search algorithms were used for hyper-parameter optimization by searching through subsets of the hyper-parameter space to find the sequence to the lowest cross-validation error.

While the original energy variable is highly right skewed, the log transformation produces a distribution much closer to normal. There are 20 distinct Principal Building Activity (PBA) designations in the CBECS dataset. To evaluate the models, the current research records the 10 k-fold cross-validated Mean Absolute Error (Mean AE), Median Absolute Error (Median AE), and the R^2 between the true and predicted energy consumption values. To supplement the prediction analysis and to aid the interpretability of our modeling process, the current study applies a Shapley analysis on the Extreme Gradient Boosting (XGBoost) Machine Learning model. The XGBoost output is then integrated with the marginal sums to estimate total energy consumption values for all commercial buildings aggregated at the Traffic Analysis Zone (TAZ) geographic scale.

2.3. Forecast scenarios

The Delaware Valley Regional Planning Commission (DVRPC) Greater Philadelphia Enduring Urbanism scenario supplies a series of potential outcomes for the Delaware Valley Region based on the interaction between population, employment, urban and transportation infrastructure development, income per capita, and other variables of urban phenomena [24]. These "what-if" scenarios are exploratory and mostly qualitative possibilities for the region.

One of the potential forecasts for Philadelphia, called the "Enduring Urbanism" scenario, describes an acceleration of urbanism into 2045. With respect to the residential sector, Philadelphia experiences drastic capital inflows towards the urban core and Central Business Districts (CBDs) followed by corresponding disinvestment in suburban and rural areas. For the commercial sector, Philadelphia undergoes shifts in employment specific to geographic area in terms of land use and zoning specifications.

From the Enduring Urbanism forecast scenario, the study reconstructs spatial shifts in income distribution as forecasted for the residential sector (see section 2.3.1) for the residential scenario model and in employment distribution for the commercial sector (section 2.3.2) for the commercial scenario model.

2.3.1. Scenario residential model

In accordance with the Delaware Valley Regional Planning Commission (DVRPC) forecast predictions, the population increases by 17% for the year 2045 for a total of 1.82 million persons for the study area. As the average household size for Philadelphia County is 2.55 persons per household [25], the study assumes this ratio will stay consistent in the base case and the number of households will also increase by 17% for a total of 713,725 households in the region. As the household residential energy prediction output is geolocated by Public Use Microdata Area (PUMA), income shifts are also analyzed at this spatial scale.

The Delaware Valley Regional Planning Commission (DVRPC) Enduring Urbanism scenario for the residential sector predicts middle- and lower-income households will shift out of core urban areas in the Delaware Valley region to outlying suburbs. These socioeconomic shifts as stipulated by the DVRPC are akin to an extreme case of classic gentrification in that capital inflows to the City of Philadelphia result in massive displacement of low- to middle- income households and an overall stratification of spatial income distribution for the region (see [18,4] on the features of classic gentrification). For more detailed information on residential scenario development, refer to [2], the complement to the current paper, where the following income shifts were applied on a transportation energy model.

To replicate the forecasted income shifts, income status (low-, mixed-, or high-income) is found for each Public Use Microdata Area (PUMA) region based on Delaware Valley Regional Planning Commission (DVRPC) 2018 median income data for the 11 PUMAs contained in Philadelphia County.

The study applies the following definitions for classifying the income groups of Public Use Microdata Areas (PUMAs) in Philadelphia County such that households are divided into three groups: those below 80%, between 80% and 120%, and above 120% of the Area Middle Income (AMI) [10].

If each of these groups make up to 20% but no more than 50% of total households in the region, then the Public Use Microdata Area (PUMA) is classified as a mixed-income area. A high-income PUMA will have over 50% of households that have median incomes greater than 120% of the Area Middle Income (AMI), and a low-income PUMA will have over 50% of households that have median incomes less than 80% of the AMI. In this way, a mixed-income PUMA demonstrates that a wide variety of income groups are represented in the neighborhood and no one group has a dominant presence. A high-income PUMA will have over 50% of households that have median incomes greater than 120% of the AMI, and a low-income PUMA will have over 50% of households that have median incomes less than 80% of the AMI.

The gentrification process is currently understood to function in relation to micro- and macro-economic trends underlying the real estate market in addition to regionally specific patterns of urban investment and disinvestment. A robust representation of future gentrification would be outside the scope of the current study, and thus the research aims to propose a method which – on the simplest terms – demonstrates how quantitative methods can augment analyses of more qualitatively defined trends of urban change. For this reason, the scenario is developed to simulate a simplified extreme case of income stratification where household income is redistributed such that only high- and low-income Public Use Microdata Areas (PUMAs) are contained in Philadelphia County.

To implement this, we take the previously mixed-income Public Use Microdata Areas (PUMAs) in the base case and change the distribution of household sample median income to match that of a neighboring high-income PUMA that contains a relatively high number of intersecting Traffic Analysis Zones (TAZs). Both low- and middle-income households are removed, or "displaced", from the altered PUMAs to simulate the concentration of high-income newcomer households into the area. Under the assumption that disinvestment towards low-income neighborhoods will persist, low-income PUMAs are left as is to replicate the effect of a widening income gap within the city.

For example, PUMA 3209 is high-income and PUMA 3211 is mixed-income in the base case, and both PUMAs share a border with overlapping income data at the TAZ scale (see Fig. 4 for a visual representation). For the scenario, the distribution of household income in the energy consumption sample is manually altered such that PUMA 3211 has an identical distribution of income groups as high-income PUMA 3209. The new PUMA 3211 will have 21% of households earning less than 80% of the AMI, 11% earning between 80% and 120% of the AMI, and 68% earnings over 80% of the AMI.

For the residential scenario, the corresponding categorical income variable is changed in the matched Public Use Microdata Sample (PUMS) dataset (see Table S1 in the Supplementary Materials for a variable description). The geospatial QGIS software tool is utilized for developing an approximate one-to-one mapping from Public Use Microdata Area (PUMA) to Traffic Analysis Zone (TAZ) for the purposes of implementing the residential scenario income level designations. Similar to the prior residential model process, the revised inputs for the household sample are put through the previously trained Extreme Gradient Boosting (XGBoost) Machine Learning model to impute energy use for study area households.

The PopGen software is then utilized to produce a representative dataset of households and then followed by a SHAP analysis for model interpretability is performed to better understand feature influence on the model output. Notable trends in influential SHAP features will provide insight to areas that warrant future research. For a visualization of the residential scenario development process, see Fig. 3.

2.3.2. Scenario commercial model

The Delaware Valley Regional Planning Commission (DVRPC) forecast for Enduring Urbanism applies scenario planning to guide regional development initiatives in accordance with carbon reduction benchmarks. The commercial scenario employs the employment projections set by land use type for the year 2045. The commercial scenario increases total employee estimates in the CoStar real estate dataset by Traffic Analysis Zone (TAZ) according to the Enduring Urbanism employment forecast.

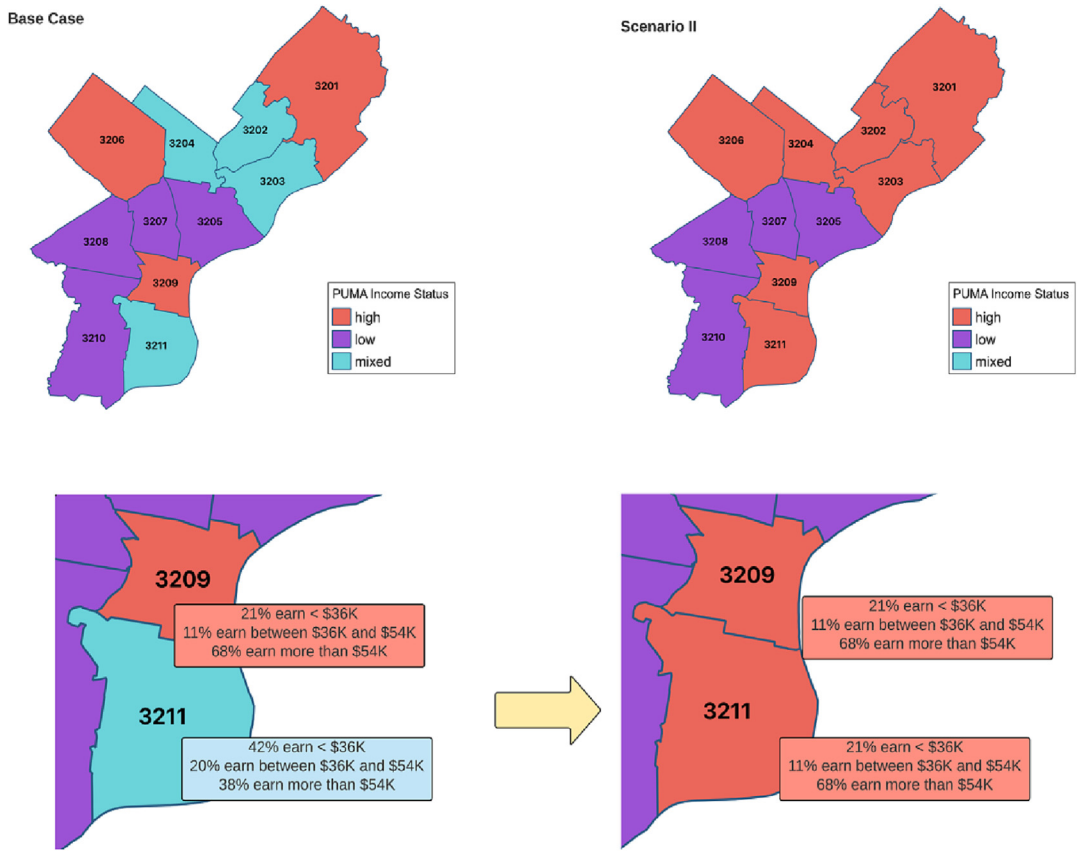

Fig. 4. Map of Philadelphia County Public Use Microdata Area (PUMA) income status with comparison between pre- and post-scenario income group distributions at an Area Middle Income (AMI) of $45,000.

First, employment counts are manually altered in the dataset to correspond to DVRPC employment projections. The DVRPC's Rapid Policy Assessment Tool (RPAT) for scenario planning provides forecasts by "Place Type", or zones with shared regional roles and development types. Thus, the proportion of employment increase per building is determined by their corresponding Place Type-approximate location in the CoStar database. For example, employment increases vary across Central Business District (CBD) zones and non-CBD zones, Commercial Mixed-Use zones and Residential Mixed-Use zones, and so on. For more information on RPAT Place Types refer to the City of Philadelphia Zoning Code Information Manual (2020).

With the updated data inputs, we run the XGBoost model to produce the energy prediction for the year 2045. The new, unlabeled testing data, as gathered from the CoStar inputs, is then processed through the commercial energy prediction model which was previously trained for the year 2015.

As the CoStar real estate database provides a comprehensive account of commercial buildings in the study area, there is no need to develop a synthetic dataset with PopGen. Building energy use is evaluated according to annual kilowatt-hour (kWh) per Traffic Analysis Zone (TAZ). For a visualization of the commercial scenario development process, see Fig. 3.

3. Results and discussion

3.1. Residential model results

Before developing the prediction model, the variables in the matched Public Use Microdata Sample (PUMS) dataset are processed such that there are twenty-six continuous standardized variables and 37 categorical variables converted into binary variables. The variables with missing values are removed from the dataset, resulting in a total of 187 variables in the residential energy model. The Extreme Gradient Boosting (XGBoost) Machine Learning model is implemented using the programming software Python 3 scikit-learn function according to the same methodology as applied to the Household Transportation Energy (HTE) model in Amiri et al. (2020). The results show that the model performed well, with a Median Absolute Error (MAE) of 6,695 for the training dataset and 7,373 for the testing dataset. The Mean Square Error (MSE) is 1,156 and 1,373 for the training and testing datasets, respectively. The R^2 value is significant at 0.79 for the training dataset and 0.76 for the testing dataset.

In addition to training time and accuracy, interpretability is critical for the transient stability prediction model. Extreme Gradient Boosting (XGBoost) is a popular Machine Learning (ML) model because the algorithm performs well, but the decision boundary in an XGBoost problem can be challenging to identify when the model contains many features. An alternative solution is to extract explanations for individual predictions, which can be accomplished using Explainable Artificial Intelligence (XAI) approaches such as SHapley Additive exPlanations (SHAP). Lundberg and Lee (2017) provide a detailed description of SHAP. SHAP techniques provide global and local interpretability, allowing urban planners and engineers to understand and trust the results of ML-based energy predictions.

The color of each dot in the SHAP plot shows how changes in a feature's value affect the change in energy consumption for the buildings based on the gradient scale provided on the y-axis. The effect of a feature on the building energy prediction model is represented by the SHAP x-value for that feature. The x-axis units

describe the XGBoost model's change in margin output in the unit of log-odds. The distribution of SHAP values for each feature is provided by overlapping points scattered vertically.

A SHAP summary plot was generated to understand which features are important in the residential energy model (Fig. 5). As expected, Fig. 5 demonstrates that high monthly electricity utility costs (ELEP) increase the model's prediction of residential energy consumption; The fact that the electricity cost (ELEP) variable is the most influential gives some assurance that the XGBoost model is performing correctly.

In Fig. 5, other variables are presented according to their predictive power in descending order, with a higher presence of single-family attached buildings (BLD_3) and lot sizes less than one acre (ACR_0) associated with lower energy use predictions. A lower number of rooms per building (RMSP) is also associated with lower energy use predictions, the fourth most influential feature. As the variables for single-family detached units (BLD_2) and property value (VALP) are more clustered around the axis for SHAP values, their influence on model predictions are more difficult to parse out. Finally, lower monthly natural gas costs are associated with lower energy use estimates, as expected. For a more detailed description of the variables, refer to Table S1 in the Supplementary Materials.

A SHAP partial dependence plot is implemented to understand the marginal influence of monthly electricity cost (ELEP) on the model output and the interaction effect of the number of rooms (RMSP) on the model trends (Fig. 6). For households with monthly electricity costs approximately under $300, Fig. 6 shows a positive and mostly linear relationship between the selected feature and the model prediction; In other words, a unit higher value of electricity cost for a selected housing unit would result in a proportionally higher jump in energy consumption until the electricity bill exceeds $300.

Additionally, the number of rooms (RMSP) feature interacts strongly with the monthly electricity cost (ELEP) for housing units with a greater number of rooms, as demonstrated by the cluster of purple to red to data points on the top right-hand side of Fig. 6. This result is as expected given that a higher number of rooms means more divisions of the residential unit in need of electricity, and further demonstrates that the model is performing as expected. If a reversal of the aforementioned relationship was presented in the Shapley output (i.e., a smaller number of rooms associated with higher electricity cost or vice versa), this would give appropriate reason to adjust and rerun the model.

Another SHAP partial dependence plot is constructed to visualize the marginal influence between unit property value (VALP) and monthly electricity cost (ELEP) on the residential model output (Fig. 7). Fig. 7 demonstrates that most households have property values (VALP) below $500,000 (USD) for the study area. The relationship between the selected and target feature is difficult to parse as the data points are clustered around the y-axis (VALP = $0), thus the interpretability of marginal feature influence is more speculative. However, the dependence plot does suggest that housing units with low property values (VALP, left-hand side of horizontal axis) strongly influence the model energy prediction as indicated by the relative lack of points on the right-hand side of the plot.

The synthetic population for residential energy consumption is created with the Machine Learning (ML) output as household samples for the Public Use Microdata Area (PUMA) regions in Philadelphia County, and Census Bureau American Community Survey (ACS) data as the marginal controls. For this scenario, household income, total households, and total population comprise the marginal sums at the Traffic Analysis Zone (TAZ) geographic scale.

Base case residential electricity energy consumption in terms of annual kilowatt-hour (kWh) is appended to the household sample. The resulting synthesized dataset represents all households in Philadelphia County, and the corresponding Public Use Microdata Area (PUMA) each household is sampled from. Categorical data on income is also available at the Traffic Analysis Zone (TAZ) level. The results of section 3.3.4 further analyze the residential energy consumption results concerning Public Use Microdata Area (PUMA) region and income status.

3.2. Commercial Model results

We experiment with training the Extreme Gradient Boosting (XGBoost) Machine Learning model to predict commercial building energy consumption using six primary energy-related features from the Commercial Buildings Energy Consumption Survey (CBECS) dataset. This feature set contains only the features from CBECS that are also available in the Costar real estate database: Principal Building Activity (PBA), square footage (SQFT), number of employees (NWKER), number of floors (NFLOOR), Heating Degree Days (HDD65), and Cooling Degree Days (CDD65). The Principal Building Activity (PBA) types, describing the primary building function across different categories in the datasets, are available in Table 2. The XGBoost model is applied to predict Philadelphia

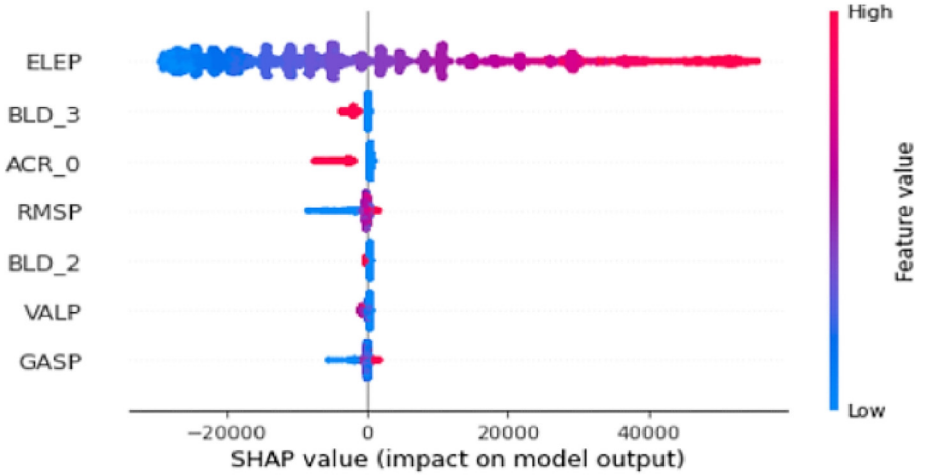

Fig. 5. Shapley summary plot for feature importance in the residential energy model with top seven variables that display predictive power on the model energy estimate. Variables are sorted according to model influence: monthly electricity utility cost (ELEP), single-family attached buildings (BLD_3), housing lot sizes less than one acre (ACR_0), number of rooms per building (RMSP), single-family detached buildings (BLD_2), property value of housing unit (VALP), and monthly natural gas cost (GASP).

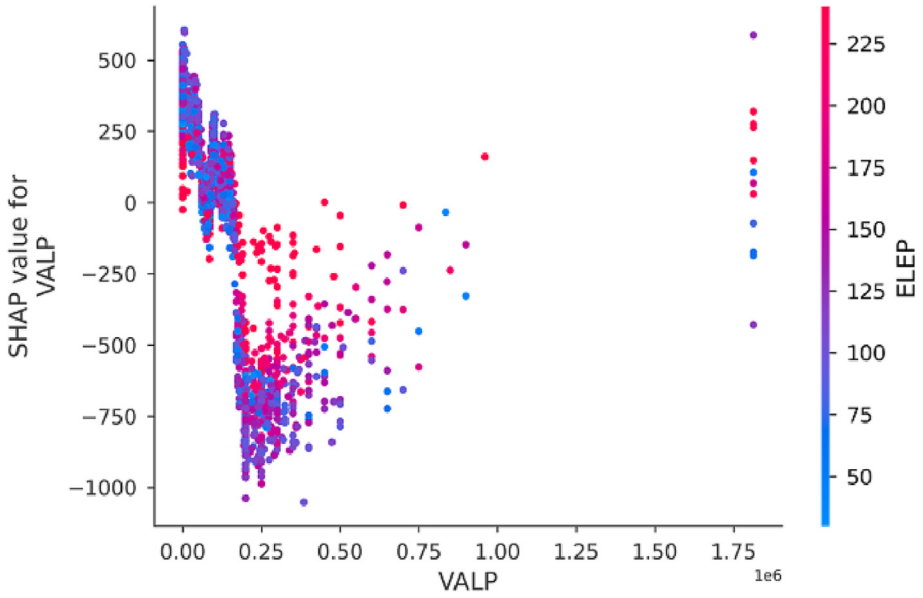

Fig. 6. SHAP partial dependence plot demonstrating the marginal effect of the monthly electricity cost (ELEP) and number of rooms (RMSP) on the outcome of the residential energy prediction model.

$$d = \sum_{i=1}^{n} \left| X_{M_i}^{RECS} - X_{M_i}^{PUMS} \right|$$

n = total number of matching variables

$X_{M_i}^{RECS}$ = i^{th} matching variable from the RECS dataset

$X_{M_i}^{PUMS}$ = i^{th} matching variables from the PUMS dataset

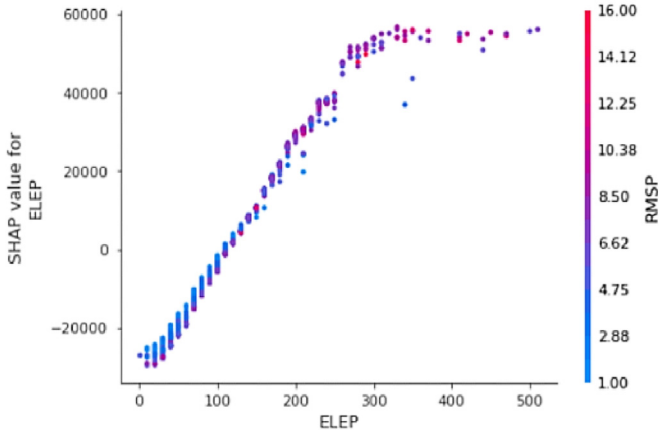

Fig. 7. SHAP partial dependence plot demonstrating the marginal effect of the property value (VALP) and monthly electricity cost (ELEP) on the outcome of the residential energy prediction model.

commercial energy consumption and then tested with the above scenarios.

The XGBoost Machine Learning model is built with the Python 3 programming software scikit-learn function. The hyperparameters are optimized using grid-search. In order to evaluate the effectiveness of the model, the cross-validated Mean Absolute Error (Mean AE), 10 $^{mean\ AE}$, Median Absolute Error (Median AE), 10 $^{median\ AE}$, and R^2 score is tested on the entire dataset (Table 1). For all evaluation metrics in Table 1, the average accuracy of each fold is reported with the standard deviation '±σ'.

When the output of a biased model is aggregated to produce estimates for larger spatial areas (e.g., the zip code or county level), the bias within the model prediction compounds and results in increasingly less reliable predictions. The error distribution of our model shows that it neither systematically overestimates nor underestimates the energy consumption values (Fig. 8). Given this, we anticipate that modeling errors will be canceled out in aggregated energy consumption estimates.

Table 2 displays the R^2 value for the XGBoost model in regard to every Principal Building Activity (PBA) class in the Commercial Buildings Energy Consumption Survey (CBECS) dataset. According to the results, the XGBoost model performs better in PBA categories with larger sample sizes, such as the case with the PBA "office" category (n = 1,044).

Most models cannot provide robust predictions for the two smallest PBA classes, "refrigerated warehouse" and "enclosed mall." The limited sample sizes (n = 16 and 14, respectively) may explain the poor performance of these classes. Additionally, buildings classified as "refrigerated warehouses" will have higher capacity cooling systems that account for the majority of their power consumption signal. This phenomenon occurs similarly with the energy use of "laboratory" class buildings, which are more likely to include equipment with high energy requirements. For these unique building usages, the variable of total building square footage is less helpful in contributing to the energy prediction of the model. For the aforementioned reasons, the Federal Energy Management Program does not use square footage as the denominator for calculating the energy-saving goals for laboratories.

Fig. 9 shows the SHAP summary plot for the commercial energy model which describes the importance of model features in terms of the range of their effect on the dataset. The model features are sorted according to their predictive power. The two highest features in Fig. 9, total building square footage (SQFT) and the number of employees (NWKER), are the most important predictors of commercial energy consumption.

As expected, the third and fourth highest predicting features are Heating Degree Days (HDD65) and Cooling Degree Days (CDD65) as these climate-related variables are known to influence building energy consumption. The Shapley results give good indication that the commercial XGBoost model is performing properly as there are

Table 1
K-fold cross validation metrics for Extreme Gradient Boosting (XGBoost) model.

Model	Mean Absolute Error (Mean AE)	$10^{\text{Mean AE}}$	Median Absolute Error (Median AE)	$10^{\text{Median AE}}$	R^2
XGBoost	0.28 +/- 0.01	1.90+/- 0.05	0.20 +/- 0.01	1.60+/- 0.03	0.84+/- 0.02

Table 2
Extreme Gradient Boosting (XGBoost) R^2 value per Commercial Buildings Energy Consumption Survey (CBECS) Principal Building Activity (PBA) Group.

Sample Size (n)	PBA	Abbreviation	R^2 value
1044	Office	PBA_2	0.89 +/- 0.01
580	Education	PBA_14	0.84 +/- 0.03
567	Non-Refrigerated warehouse	PBA_5	0.72 +/- 0.07
354	Service	PBA_26	0.43 +/- 0.13
322	Religious worship	PBA_2	0.59 +/- 0.11
316	Retail other than mall	PBA_25	0.79 +/- 0.08
311	Public assembly	PBA_7	0.82 +/- 0.03
306	Food service	PBA_15	0.36 +/- 0.15
277	Strip shopping mall	PBA_23	0.80 +/- 0.08
221	Lodging	PBA_18	0.86 +/- 0.10
215	Inpatient health care	PBA_16	0.83 +/- 0.07
131	Outpatient health care	PBA_8	0.78 +/- 0.16
111	Food sales	PBA_6	0.59 +/- 0.19
101	Vacant	PBA_1	0.17 +/- 0.49
68	Other	PBA_91	0.31 +/- 0.56
62	Nursing	PBA_17	0.29 +/- 0.97
60	Public order and safety	PBA_7	0.67 +/- 0.35
23	Laboratory	PBA_4	0.42 +/- 0.79
16	Refrigerated warehouse	PBA_11	Negative
14	Enclosed mall	PBA_24	Negative
5099	Total		0.87 ± 0.01

no unexpected, innocuous features causing undue effects on the model output.

From the feature value gradient in Fig. 9, the Shapley plot points indicate that lower values of square footage (SQFT) and employee totals (NWKER) are associated with reduced building energy use estimates. The Principal Building Activity (PBA) type "non-refrigerated warehouse" (PBA_5) has an inverse impact on the model prediction (i.e., a higher number of "non-refrigerated warehouse" building types increases the chances of a lower energy use prediction in the model). Alternatively, a greater presence of the PBA type "food service" (PBA_15) increases the energy prediction.

The Principal Building Activity (PBA) of "non-refrigerated warehouse" and "food service" are generally considered more 'tricky' variables to wield in energy consumption models due to their smaller sample sizes and particular usages of building energy compared to more standard building activity types, such as "office" or "education" building classes. However, the results of the Shapley are still aligned with the understood effects of these PBA types on energy consumption: buildings primarily functioning to carry non-refrigerated warehouses would utilize less energy and food service-type commercial buildings would utilize more. These results, again, demonstrate that the model is performing as expected, and that researchers can better trust the validity of the Machine Learning model's energy estimates.

The seventh most influential feature in the Shapley plot is the number of floors per building (NFLOOR); The feature influence of number of floors (NFLOOR) is more difficult to parse as the distribution is clustered around the vertical axis for SHAP values, but there does appear to be an association with a lower number of floors and low energy estimates in the model.

A SHAP partial dependence plot is implemented to understand the marginal influence of the square footage (SQFT) on the model output and the interaction effect of the number of employees (NWKER) on the model trends (Fig. 10). For buildings approximately under 100,000 square feet (9290.30 square meters), the building square footage does not have a strong effect on building energy consumption even when the number of total employees per building increases. Fig. 10 additionally demonstrates that buildings on the high end of total square footage (SQFT) and employee count (NWKER) contribute to increased energy use estimates in the model.

3.3. Scenario results

3.3.1. Residential model scenario results

The residential scenario analyzes the effect of forecasted shifts in income distribution on residential electricity energy consumption. The energy prediction in terms of per capita annual kilowatt-hour (kWh) is aggregated for each Public Use Microdata Area (PUMA) in Philadelphia County.

A one-way analysis of variance (ANOVA) demonstrates that there is no significant difference in energy usage across the base case and scenario treatment groups, $F(1, 21) = 0.0001$, $p = 0.99$. There was a minimal increase from the base case ($M = 17.65$, $SD = 16.58$) and the scenario ($M = 17.66$, $SD = 16.60$) with respect to energy use at the Public Use Microdata Area (PUMA) level.

For the base case and the scenario, the highest energy consumption comes from mixed-income PUMAs with PUMA Income

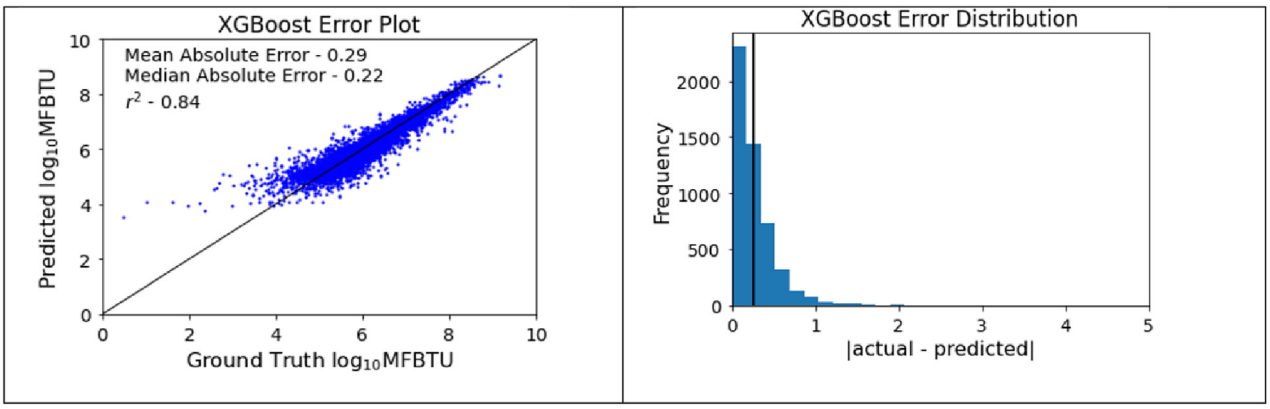

Fig. 8. Extreme Gradient Boosting (XGBoost) model error plots comparing the predicted log of Major Fuel British Thermal Unit (MFBTU) values versus the actual log values.

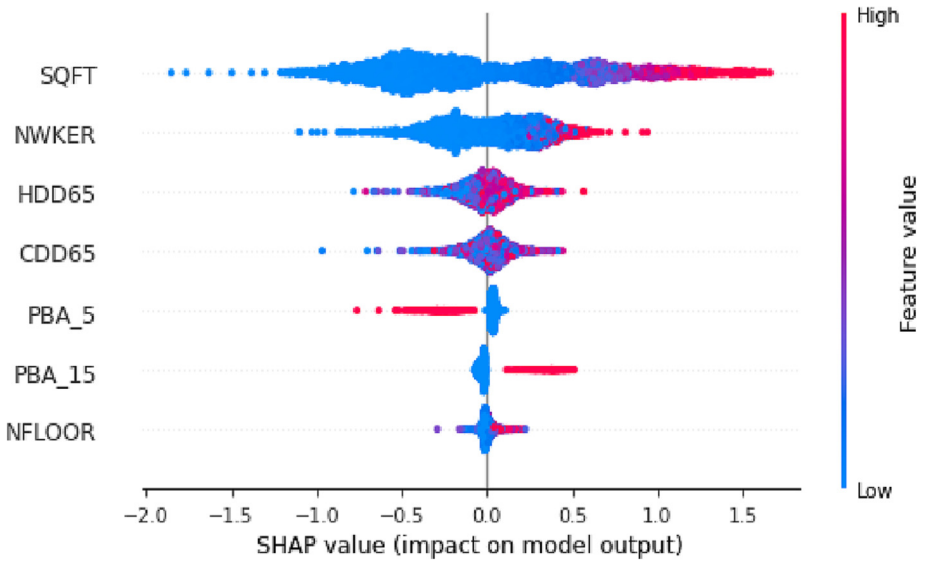

Fig. 9. Shapley summary plot for feature importance in the commercial energy model with top seven variables that display predictive power on the model estimate. Variables are sorted according to model influence: building square footage (SQFT), number of employees (NWKER), Heating Degree Days (HDD65), Cooling Degree Days (CDD65), Principal Building Activity "non-refrigerated warehouse" (PBA_5), Principal Building Activity "food service" (PBA_15), and number of floors (NFLOOR).

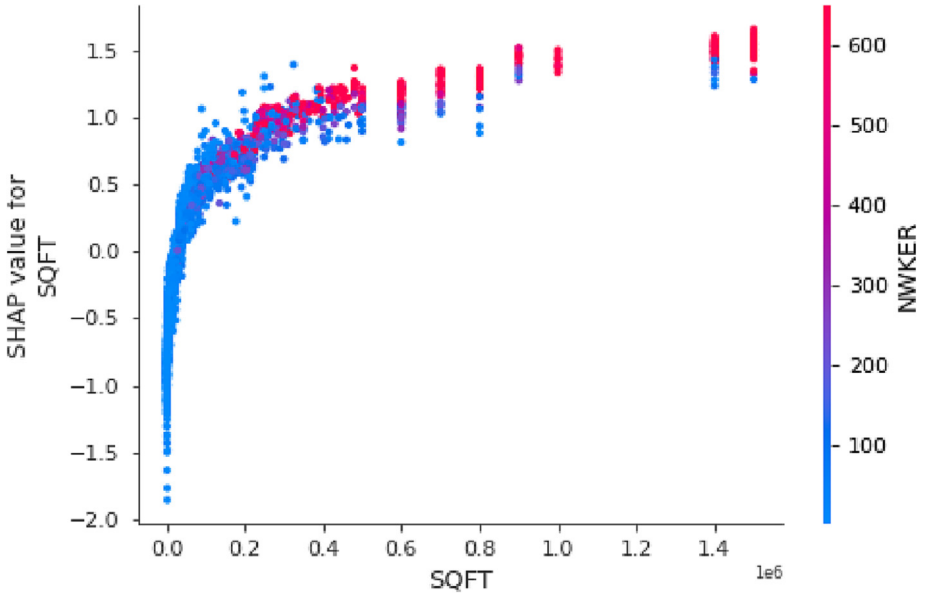

Fig. 10. SHAP partial dependence plot demonstrating the marginal effect of the square footage of building (SQFT) and number of employees (NWKER) on the outcome of the commercial energy prediction model.

Status "M" and formerly mixed-income PUMAs (i.e., PUMAs changed from mixed-income to high-income in the scenario) with PUMA Income Status "H^M". See Table 3 for PUMA energy values.

At the Public Use Microdata Area (PUMA) level, 6 out of 11 PUMAs decreased in energy use, with low-income PUMA 3207 making up 31% of the decrease in energy.

PUMA 3204, a mixed-income area, had the highest energy use with a z-score of 2.34 for the residential scenario distribution and contributed to 13% of the total per capita residential energy consumption for Philadelphia County. PUMA 3204 encompasses the Olney-Oak Lane section of Philadelphia, north of Upper North Philadelphia and south of Cheltenham, and is bounded by Roosevelt Boulevard to the south and Godfrey Avenue to the north. See Fig. 4 for a map of Philadelphia County PUMAs.

PUMA 3204 is a region of interest given its high residential energy consumption for both the base case and the scenario relative to other sampling areas. The second highest energy consumer (for both the base case and scenario) is PUMA 3206, encompassing the sections of Germantown/Chestnut Hill and Roxborough/Manayunk, and closely followed by PUMA 3209, in Center City.

Fig. 11 demonstrates the changes in energy use from the base case to the scenario, with the legend in the scale of a 0.45% decrease to a 0.42% increase in household energy consumption. In summary, the scenario changes in median income distribution did not significantly alter the overall residential electricity energy consumption for the study area.

Fig. 12 demonstrates the Shapley results for the residential scenario model. Monthly cost of electricity has the highest impact on residential energy consumption in the model. Household income

Table 3
Base case and scenario II residential energy totals and per capita use with respect to Public Use Microdata Area (PUMA) income status. Energy use in annual kilowatt-hour (kWh).

PUMA Income Status	Base Case Energy	Base Case Energy Per Capita	PUMA Income Status	Scenario Energy	Scenario Energy Per Capita
L	39,144,912	14,002	L	39,146,295	14,002
M	41,284,641	17,142	H^M	41,304,333	17,154
H	33,918,618	10,371	H	33,970,778	10,392

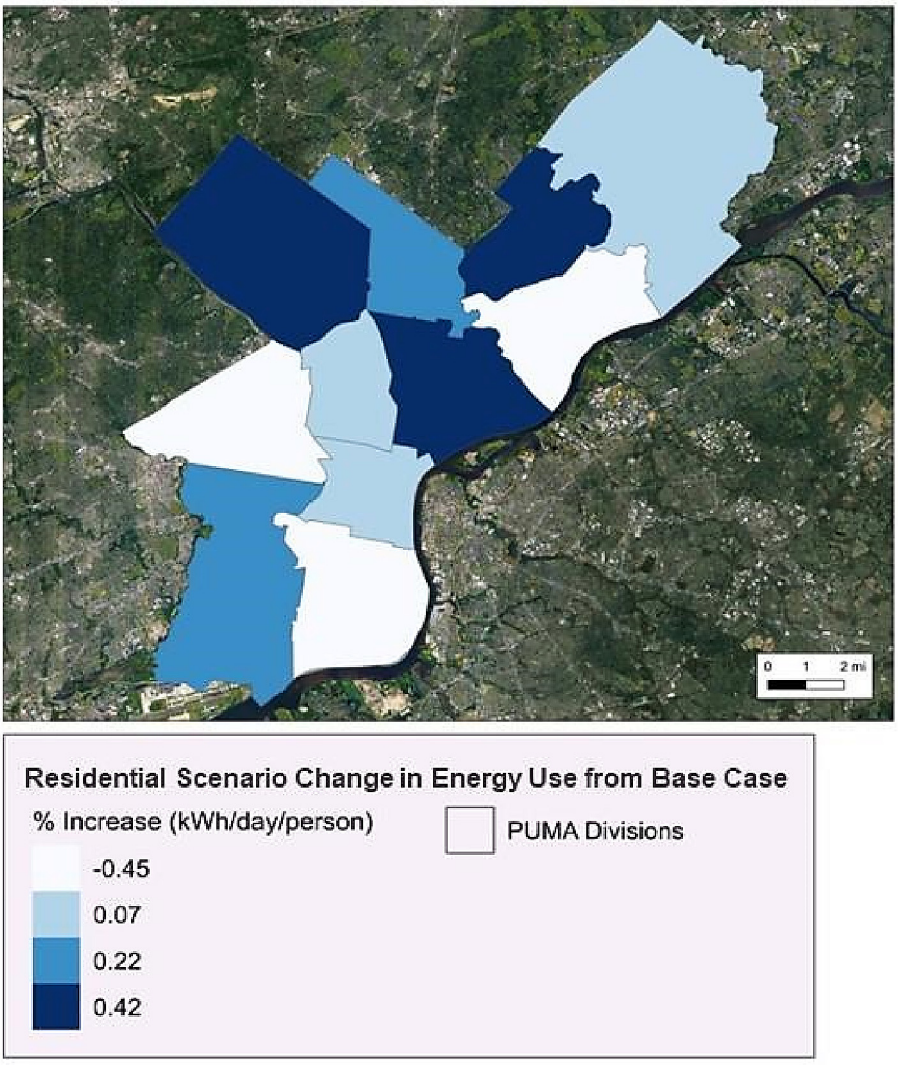

Fig. 11. Scenario change in residential energy consumption heat map for Philadelphia County in terms of percentage increase in per capita energy consumption by kilowatt-hour (kWh) and energy aggregated at the geographic scale of Public Use Microdata Area (PUMA).

(HINCP) is the ninth most influential feature for the energy prediction model. However, the Shapley results suggest that low-income households decrease predicted energy use (see Fig. 12). Relatively, higher-income households either have a minimal impact on the model or, alternatively, have an unclear distribution in terms of their effect on the model prediction. As all mixed-income PUMAs in the base case were altered to high-income PUMAs in the scenario (i.e., only high-income households were introduced to the region), the SHAP results are consistent with the lack of significant difference between treatment groups. In other words, if a greater presence of high-income households does not substantially affect the energy use estimate of the model, introducing a large number of high-income households into mixed-income neighborhoods would not produce a substantial change in energy use.

Unsurprisingly, monthly electricity cost (ELEP) demonstrates the most influence on the model prediction. In descending order of feature influence, the other features in the top ten are the presence of single-family attached units (BLD_3), lot sizes less than one acre (ACR_0), number of rooms per housing unit (RMSP), the presence of single-family detached units (BLD_2), property value (VALP), monthly natural gas cost (GASP), fire hazard/flood insurance cost (INSP), household income (HINCP), and number of bedrooms (BDSP).

For the second half of influential features, gross monthly rent (GRNTP), presence of a business or medical office on property (BUS_0), yearly water cost (WATP), three-people families (NPF_3), monthly rent (RNTP), household ownership of handheld computer (HANDHELD_1), household subscription to fiber-optic

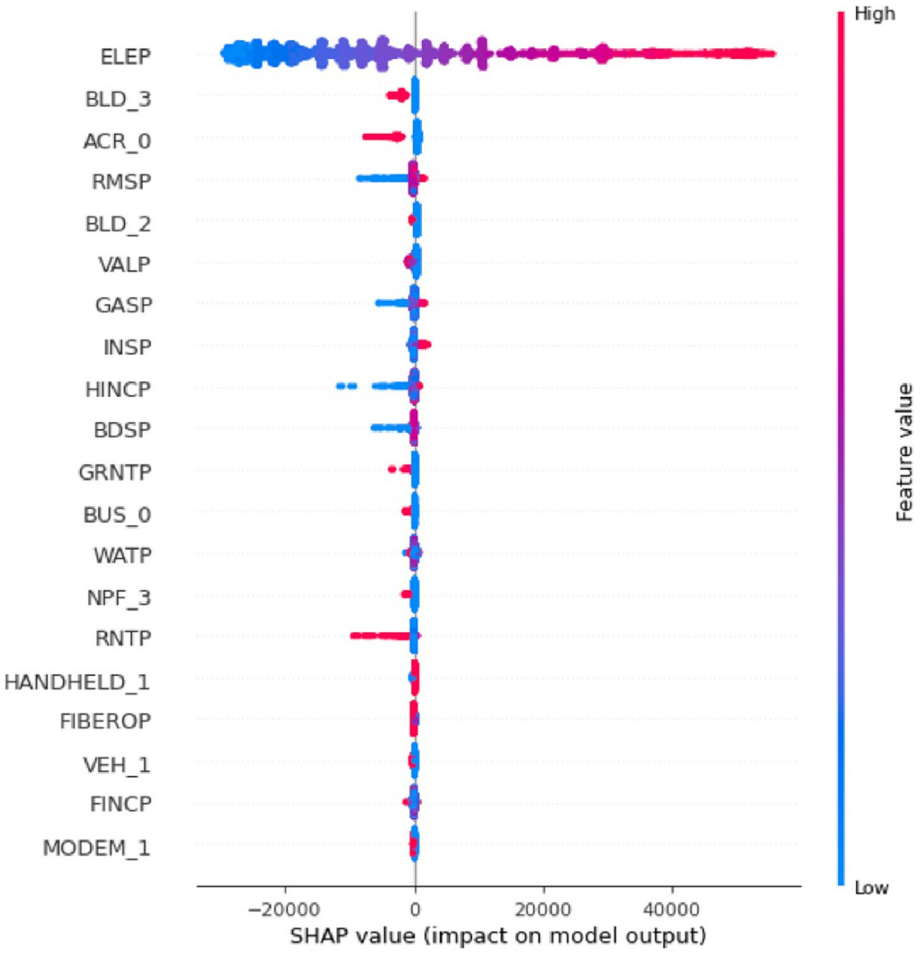

Fig. 12. SHAP for residential scenario describing feature influence on energy prediction. Variables are sorted according to model influence: monthly electricity utility cost (ELEP), single-family attached buildings (BLD_3), housing lot sizes less than one acre (ACR_0), number of rooms per building (RMSP), single-family detached buildings (BLD_2), property value of housing unit (VALP), and monthly natural gas cost (GASP), and others.

internet service (FIBEROP), presence of one vehicle per household (VEH_1), family income in the past 12 months (FINCP), and cable internet service (MODEM_1) were found to hold notable predictive power on the model.

Compared to the Shapley results of the residential base case (refer to Fig. 5), there is an introduction of new household-related features in Fig. 12 due to the scenario changes in income distribution. For this reason, the Shapley features are likely related to higher-income households, such as the case with the presence of technology-related variables specific to the PUMS household-feature data (e.g., fiber-optic internet service, cable internet service) and the greater frequency of income related features (e.g., family income in the past 12 months, gross rent). For the variable description of common features in PUMS and RECS, refer to Table S1 in the Supplementary Materials. For variable descriptions of other PUMS-specific household features, refers to the ACS PUMS dictionary codebook (2015).

The results do not indicate any confluence with the capacity to reach energy benchmarks and the increased income stratification as predicted by the DVRPC 2045 regional forecast [24]. However, as low-income households are associated with lower energy usage in the model prediction, the built characteristics of low-income housing and/or the household behaviors of low-income residents may warrant further research and demonstrates a potential intersection between topics of income representation and carbon reduction programs.

Irrespective of the results, the decreased representation of mixed-income neighborhoods represents a socially concerning trend for urban planners and policy makers alike [21].

When considering the interaction between anti-gentrification policy and sustainable design, attention may be better placed in identifying energy-saving features of subsidized affordable housing and rehabilitated low-cost units. For example, municipal zoning bonuses under the Philadelphia Planning Commission (PCPC) provide financial incentives for developers to build mixed-income housing by permitting upzoning (e.g., greater density of dwelling units, additional building height, additional gross floor area and other factors that change zoning code to increase the amount of future development) given developers reserve a proportion of the newly built units for moderate- to low-income tenants [14]. Such programs may not be conducive for reaching energy benchmarks, as the results of the current research repeatedly point to an association between lower development intensity and lower energy consumption predictions within the energy model.

For example, the top three features below electricity cost (ELEP) that reduced the energy output were the presence of single-family attached units (BLD_3), lot sizes less than one acre (ACR_0), and lower number of rooms per housing unit (RMSP) as demonstrated in Fig. 12. Although the Shapley results do not imply causality, they nonetheless provide insight to the importance of building intensity characteristics in the model prediction and give strong motivation

for further investigation in the relationship between upzoning policies and increased energy use.

As single-family attached and detached building types were found to be significant features of influence in both the base case and the scenario for residential energy use estimates, the current research provides a map of residential zoning district codes for parcels along Public Use Microdata Area (PUMA) geographic lines (see Fig. 13). The analyses corresponding to Fig. 13 are intended as a resource for researchers and policy makers in identifying local study areas of interest.

From Fig. 13, the three PUMAs that consume the most aggregate residential energy across the base case and scenario are presented in a more distinct outline and darker gray shade (PUMA 3204, 3206, and 3209). From the legend of Fig. 13, residential single-family attached zones are given in green and refer to variable code "BLD_3" in the Shapley model, the second most influential feature on energy use estimates, and residential single-family detached zones are given in blue and refer to variable code "BLD_2" in the same model, the 5th most influential feature.

Interestingly, PUMA 3204, the highest energy use contender, demonstrates a high proportion of residential single-family attached (RSA) building square footage (68% RSA square footage out of the total square feet in PUMA 3204) compared to the proportion of detached building square footage area (at 4%). Although a greater presence of single-family attached building units is associated with lower energy use predictions in the model, there appears to be other factors at play in the above-average energy expenses of PUMA 3204.

Corresponding with the Northwest neighborhoods of Philadelphia, PUMA 3206 demonstrates a relatively lower proportion of single-family attached units (approximately 29%) compared to detached units (approximately 32%). This is as expected, given the high presence of Victorian homes and the generally more suburban layout of the neighborhoods encompassing these neighborhoods. The energy consumption of single-family detached versus attached zoning areas in Northwest Philadelphia may be a topic of interest for future research, particularly given the relatively high energy consumption of this geographic area.

The outcomes of PUMA 3209 are less surprising, given the disproportionate level of commercial and industrial zoning sections within Center City. Around 18% of the zones in PUMA 3209 are residential single-family attached, with a negligible quantity pertaining to single-family detached zones.

3.3.2. Commercial model scenario results

A one-way analysis of variance (ANOVA) was performed for the base case and scenario groups with respect to building energy consumption (annual kilowatt-hour) aggregated by Traffic Analysis Zone (TAZ). After removing outliers, there were 589 TAZs for the base case and 590 TAZs for the scenario.

Results suggest that there were no significant differences between base case building energy use (M = 5581.13, SD = 4419.55) and commercial scenario building energy use (M = 5609.54, SD = 4450.47), $F(1,1177) = 0.012$, p greater than 0.5.

Fig. 14 demonstrates the distribution of commercial energy consumption for the base case Traffic Analysis Zones (TAZs). As expected, the greatest density of energy use is clustered around the Central Municipal County District (Central MCD) where there exists the highest concentration of Central Business District (CBD) zones.

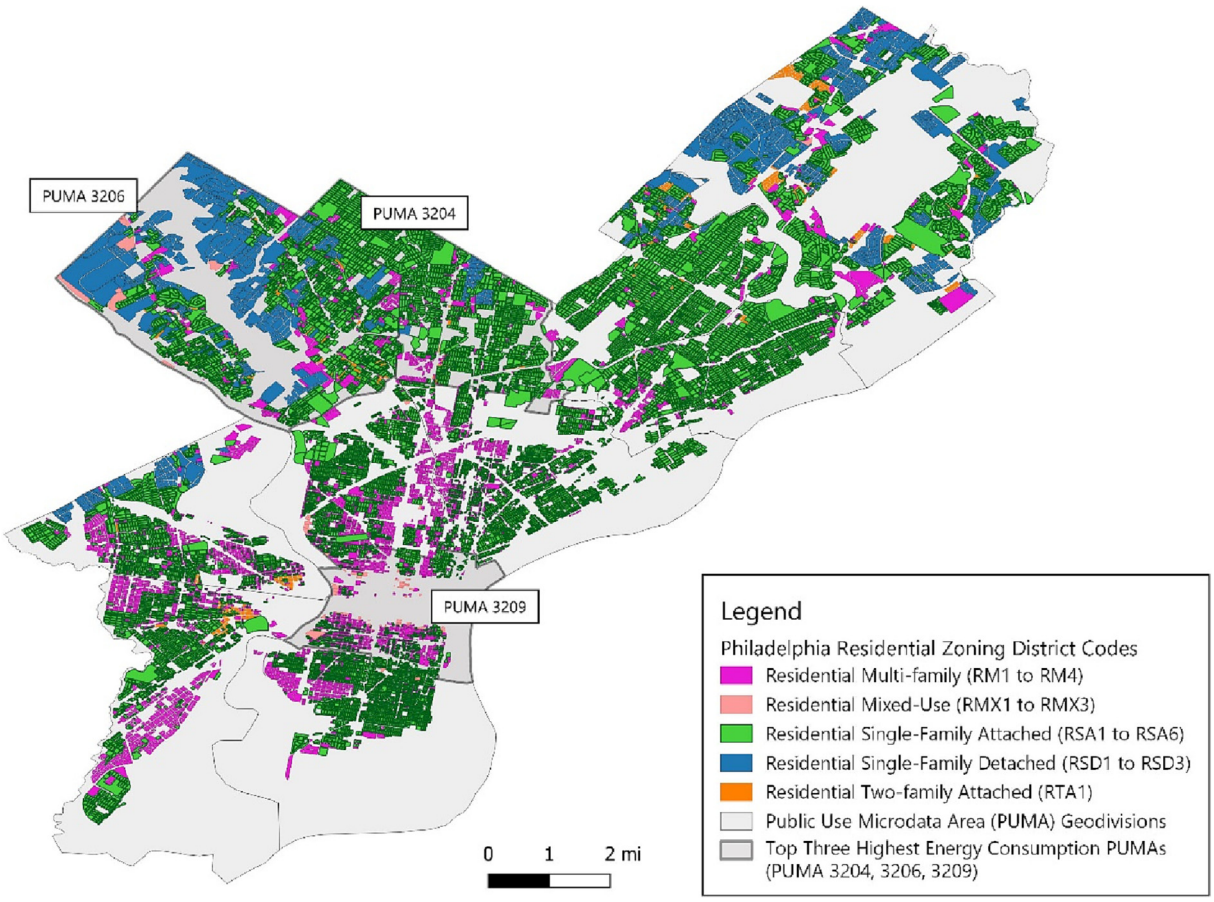

Fig. 13. Map of Philadelphia Residential Zoning District Codes overlaid with Public Use Microdata Area (PUMA) geographic boundaries.

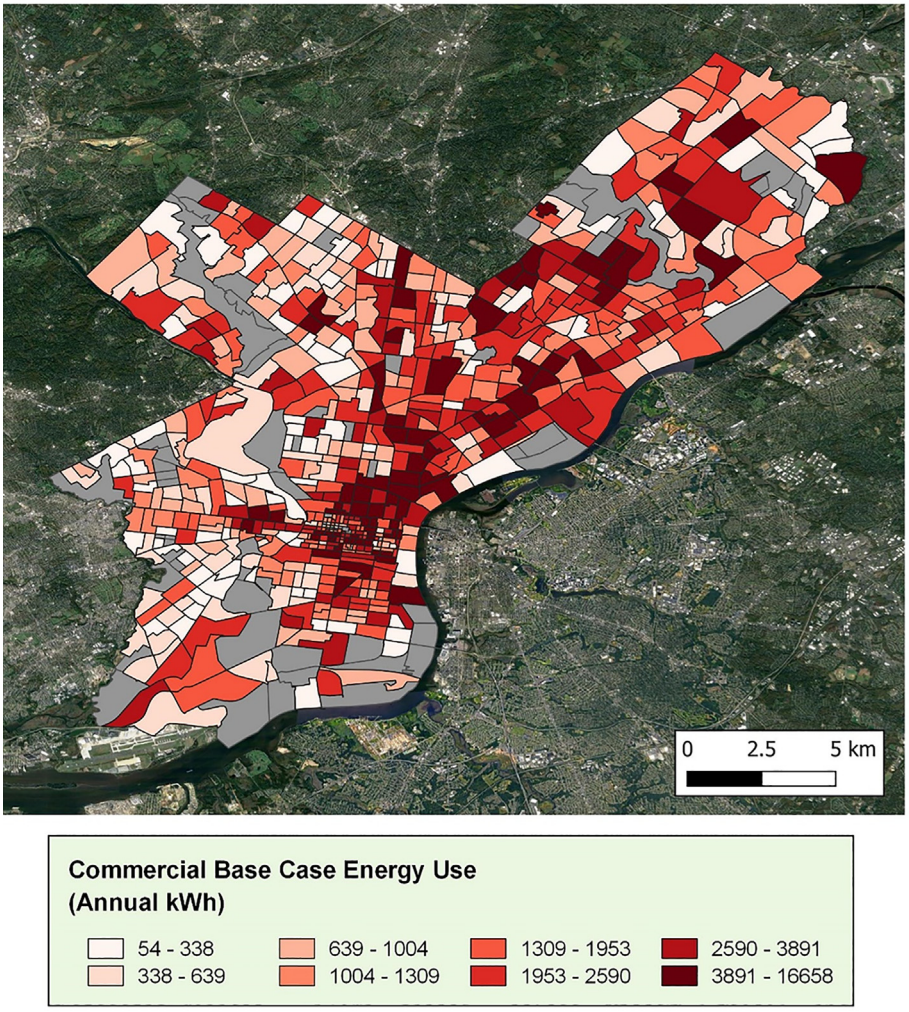

Fig. 14. Heatmap of commercial energy use for study area Traffic Analysis Zones (TAZs) in terms of annual kilowatt-hour (kWh).

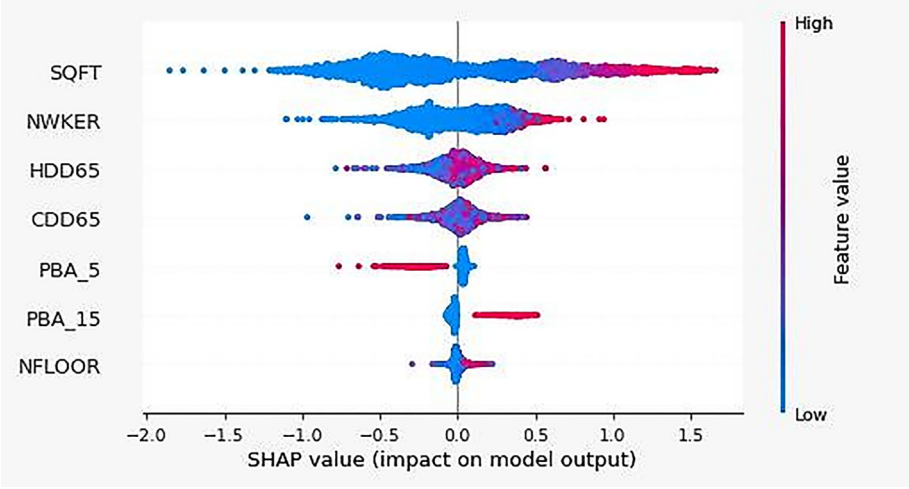

Fig. 15. SHAP for commercial scenario describing feature influence on energy prediction with variables of building square footage (SQFT), number of total employees per building (NWKER), heating degree days (HDD65), cooling degree days (CDD65), Principal Building Activity of non-refrigerated warehouse and food service (PBA_5 and PBA_15, respectively), and number of floors (NFLOOR) in descending order of model impact.

The output is put through SHAP for the Traffic Analysis Zones (TAZs) in the top 90th and lowest 10th quantile of energy consumption (Fig. 15). The variable of gross building square footage (SQFT) is the strongest explanatory feature for the energy use prediction of commercial buildings. Although the number of employees (NWKER) is an influential feature, the feature value is low on

the left-hand side of the SHAP plot, suggesting that the number of employees has a weak effect on building energy use unless the building contains an employee count significantly above the mean. This result can explain why the employment changes defined under the commercial scenario resulted in negligible effects on regional energy consumption. (see Fig. 14)

Although over 400 million employees were appended to the dataset (approximately 945 million employees are present in the base case), employment growth never exceeded 5% per Traffic Analysis Zone (TAZ) except for residential zones where employment increased by 12% Given that residential zone TAZs contain relatively low counts of commercial buildings to begin with, this would not provide large enough contributions to the model prediction for scenario energy use to be significantly different from the base case.

As the number of floors per building (NFLOOR) feature has a relatively lower impact on the model prediction, the results suggest that building gross square footage (SQFT) should be prioritized as a variable of interest over Floor Area Ratio (FAR) in carbon reduction programs.

Fig. 15 additionally demonstrates that samples with the Principal Building Area (PBA) type of "nonrefrigerated warehouse" (PBA_5) and "food service" (PBA_15) influence the model prediction more than other principal building area categories. Unsurprisingly, the results of the Shapley are highly similar to the plot for the commercial base case (see Fig. 9 for the commercial base case Shapley plot). For more information on common feature variables and their abbreviations, refer to Table 2.

The commercial scenario suggests that employment growth under the Delaware Valley Regional Planning Commission (DVRPC) Enduring Urbanism scenario may not be a significant concern with respect to reducing carbon emissions, particularly if the building has a low to average employee capacity level. However, larger commercial buildings with greater gross square footage and employment capacity should be primary targets for carbon reduction initiatives. Furthermore, buildings in the top quantile of square footage can be good target subjects for physical white-box research on energy demand.

4. Conclusion

The current research provides a novel method for supplementing knowledge of regional sociodemographic and economic trends with existing Machine Learning techniques in order to develop more spatially granular and quantitatively detailed information on energy use predictions. The residential and commercial scenarios are constructed from the Delaware Valley Regional Planning Commission (DVRPC) Enduring Urbanism forecasts of income and employment trends. The study aims to provide more actionable detail on household- and building-level characteristics and their corresponding energy demands at the Traffic Analysis Zone (TAZ) and Public Use Microdata Area (PUMA) geographic scale. The study provides reliable energy estimates for the year 2015, the base case, and for the year 2045, the scenario, which is constructed in consideration of accelerated urbanism of the county as forecasted by the Delaware Valley Regional Planning Commission (DVRPC) Enduring Urbanism report.

The study applied Extreme Gradient Boosting (XGBoost), a Machine Learning (ML) algorithm, to train and validate the energy models with a Shapley analysis of feature influence and model validity. Through Shapley, the research provides insight on the underlying mechanisms of the model and provides a basis on which researchers can trust the model's energy estimates. Through the application of Machine Learning, the current research could compensate for gaps in data availability in the residential and commercial datasets, gaps that had previously prevented researchers from utilizing the wealth of household and building data available to them in the construction of an energy prediction model. These energy estimates provide a promising framework to make more informed decisions on developing carbon reduction policies on energy hotspot neighborhoods and in reducing the presence of energy expensive features in the building stock.

From the Machine Learning output, the research was able to identify an interplay between certain residential building zoning codes (e.g., residential single-family attached zones) and reduced energy consumption. Although Machine Learning models are constructed for predictability over causality, the spatially granular level of the data output in relation to certain building classifications allows stakeholders to go by more than just a generalization of potential energy-expensive features; The model output provides a highly detailed map of potential energy hotspots in relation to plausible energy expensive features, and a productive base on which to base future research. Overall, the residential energy prediction model finds that features related to lower building intensity relate to lower energy consumption estimates in the model (e.g., lower lot acreage, lower number of rooms per unit). These results give reason to reinvestigate the effects of upzoning policies, commonly present as an affordable housing solution in Philadelphia and other cities across the U.S., and subsequent changes in energy use for these areas.

With respect to the commercial sector, the study suggests that commercial buildings in the top quantiles of square footage and employee count should be the primary targets for energy reduction programs. The research posits an approximate threshold of 10,000 square feet of total building area, with buildings over that marker being prioritized due to their disproportionate influence on the energy prediction of the model. Buildings with a primary function of providing food services should also be taken into consideration for future energy use research, as this building activity type had higher predictive power on the energy prediction model relative to other building categories.

With respect to forecasted energy for 2045, the changes in income distribution did not significantly alter model behavior for residential energy, nor did the increase in employment change the outcomes for the commercial energy model. For the residential scenario, the study finds that high-income households have a negligible effect on the energy prediction with low-income households associated with reduced energy estimates. Income diversity stands as an important quality to uphold for the municipality in the future, regardless of the lack of confluence with energy estimates, but the model results generally place more emphasis on the impact of building characteristics over household socioeconomic features. Notably, an influx of high-income residents into a neighborhood would likely result in a corresponding reshaping of the built environment, but the model did not integrate any prediction of such housing rehabilitation and development as it was outside the scope of the study.

For the commercial scenario, employment increases have a negligible effect on commercial building energy consumption, even when employee counts are nearly doubled. Thus, increased employee totals overall are not necessarily a cause for concern, but rather a high concentration of employees within a single building may increase energy use for the region. For this reason, commercial buildings at the top quantiles of employee counts should be a priority for sustainability initiatives and further research. Similar to the residential model findings, built features such as gross square footage are notably influential for commercial energy use.

The current research was written as a complement and follow-up to Amiri et al. (2020) which implemented the same methodological structure on DVRPC Enduring Urbanism forecasts for the transportation sector. Thus, both papers provide a holistic

framework of transportation (Amiri et al. 2020) and residential and commercial (current paper) sector energy consumption for both the year 2015 and the forecasted year 2045. Beyond demonstrating novel uses for Machine Learning and statistical techniques in scenario planning contexts, the research also aims to provide a diverse set of resources to aid in municipal planning initiatives and a rich map of target areas and features of interest for future research in energy modeling.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Shideh Shams Amiri reports financial support was provided by National Science Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.enbuild.2023.112965.

References

[1] Amasyali, Kadir, and Nora El-Gohary. 2016. "Building Lighting Energy Consumption Prediction for Supporting Energy Data Analytics." *Procedia Engineering*, ICSDEC 2016 – Integrating Data Science, Construction and Sustainability, 145 (January): 511–17. 10.1016/j.proeng.2016.04.036.

[2] Amiri, Shideh Shams, Maya Mueller, and Simi Hoque. 2022. "Investigating the Application of a Transportation Energy Consumption Prediction Model for Urban Planning Scenarios in Machine Learning and Shapley Additive Explanations Method." *Journal of Sustainability Research* 4 (1).

[3] Bourdeau, Mathieu, Xiao qiang Zhai, Elyes Nefzaoui, Xiaofeng Guo, and Patrice Chatellier. 2019. "Modeling and Forecasting Building Energy Consumption: A Review of Data-Driven Techniques." *Sustainable Cities and Society* 48 (July): 101533. 10.1016/j.scs.2019.101533.

[4] K. Chapple, Income Inequality and Urban Displacement: The New Gentrification, New Labor Forum 26 (1) (2017) 84–93, https://doi.org/10.1177/1095796016682018.

[5] Y. Chen, M. Guo, Z. Chen, Z. Chen, Y. Ji, Physical Energy and Data-Driven Models in Building Energy Prediction: A Review, Energy Rep. 8 (November) (2022) 2656–2671, https://doi.org/10.1016/j.egyr.2022.01.162.

[6] M. Christenson, H. Manz, D. Gyalistras, Climate warming impact on degree-days and building energy demand in Switzerland, Energy Convers Manage 47 (6) (2006) 671–686.

[7] D.B. Crawley, J.W. Hand, M. Kummert, B.T. Griffith, Contrasting the Capabilities of Building Energy Performance Simulation Programs, Build. Environ., Part Special: Building Performance Simulation 43 (4) (2008) 661–673, https://doi.org/10.1016/j.buildenv.2006.10.027.

[8] N. Djuric, V. Novakovic, Identifying Important Variables of Energy Use in Low Energy Office Building by Using Multivariate Analysis, Energ. Buildings 45 (February) (2012) 91–98, https://doi.org/10.1016/j.enbuild.2011.10.031.

[9] F. Jiang, J. Ma, Z. Li, Y. Ding, Prediction of Energy Use Intensity of Urban Buildings Using the Semi-Supervised Deep Learning Model, Energy 249 (June) (2022), https://doi.org/10.1016/j.energy.2022.123631.

[10] Kneebone, Elizabeth. 2019. "How Housing Supply Shapes Access to Entry-Level Homeownership," 25.

[11] X.J. Luo, L.O. Oyedele, A.O. Ajayi, O.O. Akinade, Comparative Study of Machine Learning-Based Multi-Objective Prediction Framework for Multiple Building Energy Loads, Sustain. Cities Soc. 61 (October) (2020), https://doi.org/10.1016/j.scs.2020.102283.

[12] D. Mazzeo, N. Matera, C. Cornaro, G. Oliveti, P. Romagnoni, L. De Santoli, EnergyPlus, IDA ICE and TRNSYS Predictive Simulation Accuracy for Building Thermal Behaviour Evaluation by Using an Experimental Campaign in Solar Test Boxes with and without a PCM Module, Energ. Buildings 212 (April) (2020), https://doi.org/10.1016/j.enbuild.2020.109812.

[13] D. Monfet, M. Corsi, D. Choinière, E. Arkhipova, Development of an Energy Prediction Tool for Commercial Buildings Using Case-Based Reasoning, Energ. Buildings 81 (October) (2014) 152–160, https://doi.org/10.1016/j.enbuild.2014.06.017.

[14] Philadelphia Department of Planning and Development, Mixed Income Bonus Fact Sheet. In,editor. 2022. https://www.phila.gov/media/20220713155329/mixed-income-housing-zoning-bonus-fact sheet_7.2022.pdf.

[15] T.A. Reddy, D.E. Claridge, Uncertainty of 'Measured' Energy Savings from Statistical Baseline Models, HVAC&R Research 6 (1) (2000) 3–20, https://doi.org/10.1080/10789669.2000.10391247.

[16] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M.A. Brown, R.M. Pendyala, Machine Learning Approaches for Estimating Commercial Building Energy Consumption, Appl. Energy 208 (December) (2017) 889–904, https://doi.org/10.1016/j.apenergy.2017.09.060.

[17] M. Shin, S.L. Do, Prediction of Cooling Energy Use in Buildings Using an Enthalpy-Based Cooling Degree Days Method in a Hot and Humid Climate, Energ. Buildings 110 (January) (2016) 57–70, https://doi.org/10.1016/j.enbuild.2015.10.035.

[18] N. Smith, Toward a Theory of Gentrification A Back to the City Movement by Capital, Not People, J. Am. Plann. Assoc. 45 (4) (1979) 538–548, https://doi.org/10.1080/01944367908977002.

[19] R. Wang, L.u. Shilei, W. Feng, A Novel Improved Model for Building Energy Consumption Prediction Based on Model Integration, Appl. Energy 262 (March) (2020), https://doi.org/10.1016/j.apenergy.2020.114561.

[20] W. Zhang, C. Robinson, S. Guhathakurta, V.M. Garikapati, B. Dilkina, M.A. Brown, R.M. Pendyala, Estimating Residential Energy Consumption in Metropolitan Areas: A Microsimulation Approach, Energy 155 (July) (2018) 162–173, https://doi.org/10.1016/j.energy.2018.04.161.

[21] M. Zuk, A.H. Bierbaum, K. Chapple, K. Gorska, A. Loukaitou-Sideris, Gentrification, displacement, and the role of public investment, J. Plan. Lit. 33 (1) (2018) 31–44.

[22] *2021* Greenworks Initiative Update. (2021). Retrieved from https://www.phila.gov/media/20210420095452/2021-Greenworks-Initiatives-Update.pdf

[23] S Rässler, Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches, Springer Science & Business Media, 2012.

[24] DVRPC Report. (2016). Retrieved from Philadelphia, PA: https://www.dvrpc.org/reports/16007a.pdf

[25] Census. U.S. Census Bureau, Selected Housing Characteristics, 2015-2019 American Community Survey 5-year estimates. 2019. URL http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_19_5YR_DP04.

[26] Marcello D'Orazio. StatMatch: Statistical Matching or Data Fusion, 2019. URL https://CRAN.R-project.org/package=StatMatch. R package version 1.3.0.