Version of Record: https://www.sciencedirect.com/science/article/pii/S0306261917313429
Manuscript_6970b9ff32e8422e24d480360ef1c366

Machine Learning Approaches for Estimating Commercial Building Energy Consumption

Caleb Robinson[a], Bistra Dilkina[a,*], Jeffrey Hubbs, Wenwen Zhang[b], Subhrajit Guhathakurta[b], Marilyn A. Brown[c], Ram M. Pendyala[d]

[a]*School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332*
[b]*School of City and Regional Planning, Georgia Institute of Technology, Atlanta, GA 30332*
[c]*School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332*
[d]*School of Sustainable Engineering and the Built Environment, Arizona State University, 660 S. College Avenue, Tempe, AZ*

Abstract

Building energy consumption makes up 40% of the total energy consumption in the United States. Given that energy consumption in buildings is influenced by aspects of urban form such as density and floor-area-ratios (FAR), understanding the distribution of energy intensities is critical for city planners. This paper presents a novel technique for estimating commercial building energy consumption from a small number of building features by training machine learning models on national data from the Commercial Buildings Energy Consumption Survey (CBECS). Our results show that gradient boosting regression models perform the best at predicting commercial building energy consumption, and can make predictions that are on average within a factor of 2 from the true energy consumption values (with an r^2 score of 0.82). We validate our models using the New York City Local Law 84 energy consumption dataset, then apply them to the city of Atlanta to create aggregate energy consumption estimates. In general, the models developed only depend on five commonly accessible building and climate features, and can therefore be applied to diverse metropolitan areas in the United States and to other countries through replication of our methodology.

Keywords: Commercial building energy consumption, Modeling, Machine learning, CBECS
2010 MSC: 00-01, 99-00

Highlights

- Machine learning models were used to estimate commercial building energy consumption.
- CBECS was used to train a US-wide model with five commonly available features.
- Validation of the model on city-specific building data was performed for New York City.
- The gradient boosting model performs best compared to Linear, SVM, and other methods.
- Availability of more building features results in more accurate models.

1. Introduction

There is substantial evidence to suggest that different configurations of the built environment are closely associated with variations in energy consumption and climate altering greenhouse gas emissions [1, 2, 3, 4, 5]. While the relationship between urban form and energy use in transportation has been well studied, we know far less about the impact of urban form on residential and commercial energy demands [6, 7, 8, 9]. A 2009 study commissioned by the National Academy concluded that increasing development densities leads to modest savings in energy use in transportation, and by extension, a reduction in greenhouse gas emissions [10]. Yet, if our interest is in building energy efficient communities, a more comprehensive set of attributes of the built environment need to be examined to determine whether increasing development densities actually lead to energy savings. The estimation of building energy consumption at the scale of small urban areas is difficult without building level data and few studies have attempted to provide energy footprints for residential and commercial buildings at neighborhood scales. This paper fills some of that gap by providing a generic technique for estimating commercial building energy from publicly available data in the U.S.

According to the 2015 annual energy consumption data released by the U.S. Energy Information Administration (EIA), residential and commercial buildings consumed 39 quadrillion Btus., representing 40% of total energy con-

*Corresponding author
Email addresses: dcrobins@gatech.edu (Caleb Robinson), bdilkina@cc.gatech.edu (Bistra Dilkina), jeffrey.hubbs@gmail.com (Jeffrey Hubbs), wzhang300@gatech.edu (Wenwen Zhang), subhro.guha@coa.gatech.edu (Subhrajit Guhathakurta), Marilyn.Brown@pubpolicy.gatech.edu (Marilyn A. Brown), ram.pendyala@asu.edu (Ram M. Pendyala)

sumption in the United States [11]. Similarly, according to the European Commission [12], building energy consumption accounts for 40% of the total energy consumption in the EU. Globally, the building sector accounted for approximately 32% of energy consumption in 2010 [13]. Thus, advanced economies spend a particularly large percentage of their energy in buildings compared to developing countries. While the EIA releases highly detailed annual energy consumption estimates by sector for the U.S. as a whole, it is useful for local policy makers to have small area or neighborhood level estimates of energy consumption. Without access to fine scale data on energy use, urban planners will not be able to benchmark the effects of environmental or climate related policies affecting different sections of the urban region or make confident predictions about the outcomes of proposed policies. Machine learning models can also help city and regional planners predict the energy burdens that could result from alternative urban growth patterns and global warming scenarios. Spatial energy consumption information at a granular scale is therefore crucial to fulfilling sustainability goals.

One way of estimating building energy consumption, in the absence of actual sensor data, is to create physical building models with a "template" of representative buildings, then run thermodynamic simulations to estimate the energy demands [14]. These "engineering" models of building energy consumption are computationally expensive and cannot capture the wide variety of different buildings present in cities, as modeling each type of building requires very detailed input data, which is costly to collect. Statistical models can be used to fill the gaps where resources are too limited to use physical models, or the scale of the study area makes physical modeling impractical.

We aim to model commercial building energy consumption at the building level using machine learning models. This statistical approach avoids expensive physical modeling efforts, and is able to provide reasonable estimates that can be validated against existing building level energy consumption databases. Specifically, we train machine learning models on the 2012 Commercial Building Energy Consumption Survey microdata [15], then validate this approach using the Local Law 84 (LL84) dataset from New York City. We show how our models can be used to create comprehensive metropolitan wide commercial energy consumption maps by applying them to 73,388 commercial buildings in Atlanta, GA. These maps will help city planners better understand the relationships between urban form and energy consumption, and plan for the future. Our models purposefully only rely on a limited set of building level features, namely: square footage, principal building activity, number of floors, and heating and cooling degree days, so that they can be applied to *any* metropolitan area in the United States. Furthermore, to facilitate the wider adoption of our methods in other metropolitan areas throughout the US, we have released the code and trained models used in this paper in a public GitHub repository[1]. The code and instructions provided in the GitHub repository can be used to reproduce the modeling and validation results from this paper, and to apply trained models in new settings. In general, the machine learning modeling approach for broad commercial building energy consumption prediction presented in this work is a novel step toward better understanding the energy consumption landscape in the United States.

2. Related Work

Methods for predicting building energy consumption can be categorized into three groups: engineering methods (white-box models), statistical methods (black-box models), and hybrid approaches (grey-box models) [14, 16, 17]. Engineering methods physically model building energy consumption by simulating the laws of thermodynamics using extensive building level data. This method cannot be applied precisely at the urban scale, due to its large data and computational demands, however it is used to estimate the energy consumption of a small typology of buildings which are then aggregated over entire urban areas [18, 16, 17]. Statistical methods for estimating building energy consumption aim to directly regress energy consumption values on associated building and climate variables. In general, machine learning methods (such as the methods used in this study) fall into this category, although Zhao and Magoulès have separated machine learning based studies from linear regression model based studies in their review [14]. Hybrid methods involve a combination of both engineering and statistical models, and use the output from engineering models as an input to statistical models. The purpose of these models is to offset some of the constraints involved with physical modeling (such as the inability to model every building in a district) with the flexibility of statistical approaches [19].

A commonality between the engineering, statistical, and hybrid methods is that they are all limited by the availability of relevant data. Indeed, availability of data is crucial for any statistical modeling approach, but our method lowers the bar for data requirement as discussed later. Mathew et al. discuss big data applications of the US Department of Energy's building performance database (BDP) [20]. The BDP has data for both residential and commercial buildings on a larger scale than either the Commercial Building Energy Consumption Survey (CBECS) or the Residential Building Energy Consumption Survey (RECS), however can only be used in benchmarking applications. Access to fine grained data, such as that collected by BDP, will be crucial for development of more accurate and relevant statistical based models [21].

Linear statistical models have been used in studies for predicting energy consumption at both the building and

[1] The code and trained models are available at: `https://github.com/SEI-ENERGY/Commercial-Energy/`

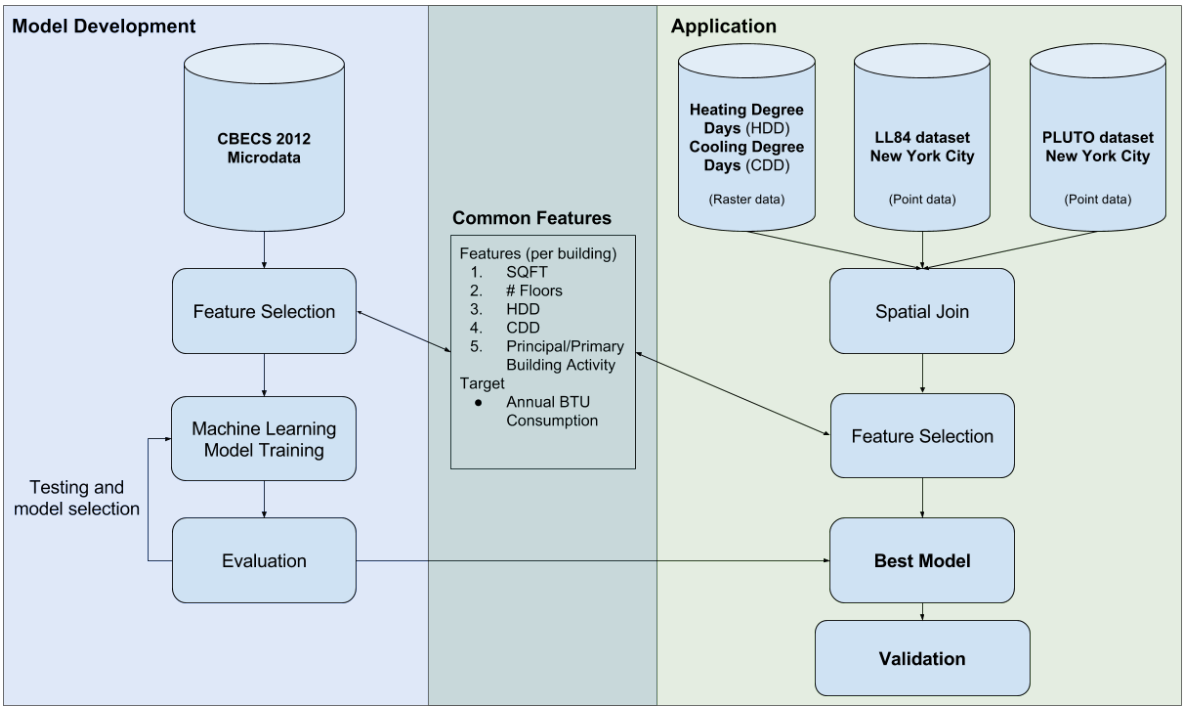

Figure 1: Modeling framework.

Abbreviation	Description
CBECS	Commercial Building Energy Consumption Survey - A national (for the US), statistically representative survey of commercial building characteristics and energy consumption published by the Energy Information Administration.
LL84	New York City Local Law 84 data - A dataset published yearly by New York City on the energy consumption of commercial buildings.
MFBTU	Major fuel consumption in BTUs - This is the total energy consumption of a building, i.e what our models predict.
PBA	Principal building activity - The main use category of a building, e.g. “Office”, “Education”, “Nonrefridgerated warehouse”.
TAZ	Traffic analysis zone - The name given to small geographic areas used for planning in many cities. We spatially aggregate our model’s commercial building energy consumption estimations in Atlanta using these zones.
HDD	Heating degree days - The total number of degrees below 65 Fahrenheit summed for each average daily temperature in a year. This a proxy measure for the total energy needed to heat a building in a year.
CDD	Cooling degree days - The total number of degrees above 65 Fahrenheit summed for each average daily temperature in a year. This a proxy measure for the total energy needed to cool a building in a year.

Table 1: Commonly used abbreviations.

zone level resolutions. Boulaire et al. use robust linear models to model energy consumption at the zone level in NSW, Australia [22]. Kuusela et al. use a lognormal modeling framework to model electricity consumption from aggregate building features at the zone level of a Finnish city [23]. Kontokosta use robust linear models to estimate building energy consumption of residential and commercial buildings using 2011 New York City’s Local Law 84 (LL84) dataset [24]. While these models are easily interpretable, machine learning models are better suited for modeling the complex relationships between building level characteristics and energy consumption since such models have fewer constraints about the statistical relationships among variables.

Previous studies have shown that machine learning models out-perform linear models in modeling building energy consumption. Tso et al. use linear regression models, decision tree models, and neural networks to model residen-

tial electricity consumption at the building level in Hong Kong [25]. The study splits the dataset across the summer and winter seasons and trains models separately for each season.Similarly, Fan, Xiao, and Wang use an ensemble of machine learning models to predict the next-day building energy consumption of the "International Commerce Center" in Hong Kong with good results [26]. Wei et al. use two linear models and four non-parametric machine learning models to estimate gas and electricity consumption at a zone level in London [27]. Similarly, Yalcintas et al. train an artificial neural network and multiple linear regression models to predict the energy use intensity values (kWh per square meter) with the 1999 CBECS data [28]. This study only uses one category of building from the CBECS dataset, and categorizes the target values to convert the problem into an easier classification problem. These three studies all find that machine learning models perform better than linear regression based models, however they are limited both by the few models they consider, and the smaller datasets they use. In our work we consider a broad range of machine learning models, and use as much of the CBECS data as possible with the objective of creating a general model for estimating commercial building energy consumption.

Finally, the most similar work to ours is by Howard et al., which uses the LL84 dataset to estimate robust linear regression models for predicting energy consumption in commercial and residential buildings in New York City [29]. The authors calibrate the linear model using the building function, square footage, and energy use intensity features from the New York specific data, and the final predicted energy consumption is disaggregated into several different end uses based on CBECS and RECS data. The final estimates were aggregated at the block area level to provide spatial energy consumption distributions. Our work is different from this as we train multiple machine learning models on the national CBECS data, which we then apply to specific metropolitan areas. We are not concerned with the specific energy end-uses in commercial buildings, as this would require more detailed data to model and validate, but instead focus on estimating generally capable models that can be used to create acceptable estimates of total energy consumption using as few features as possible.

3. Data and Methods

Our two main objectives are to 1.) train machine learning models to predict the annual major fuel, or the combination of electricity, natural gas, and fuel oil, consumption, of commercial buildings from easily accessible descriptive features of buildings, and 2.) validate the models' ability to be applied to specific metropolitan areas. Specifically, we train and test our models using national survey data from CBECS, then use true energy consumption values from New York City's Local Law 84 (LL84) dataset to validate the ability of the nationwide CBECS-trained models to be applied accurately to a specific metropolitan area. In Section 3.1 we describe the datasets we use in further detail, in Section 3.2 we explain the details of our methods, and in Section 4 we present our results and discussion. We give a listing of commonly used abbreviations in Table 1.A graphical overview of the methodology that we follow in this work is given in Figure 1.

3.1. Data

The CBECS microdata is released by the U.S. Energy Information Administration (EIA) approximately every five years. The latest microdata, from 2012, contains 6,720 rows of data, where each row represents some of the estimated 5.6 million commercial buildings in the U.S. [15]. Specifically, each row contains the features, or information on numerous characteristics, of a particular representative building gathered through the CBECS 'Building Survey' questionnaire. These features include information such as: the square footage of the building, the principal building activity (PBA), heating and cooling degree days, number of employees, etc.

The New York City Benchmarking Law, known as Local Law 84 (LL84), requires buildings that are over 50,000 square feet, or lots with buildings with over 100,000 square feet combined, to report their annual energy and water consumption values in a standardized manner using the EPA's portfolio manager database [30]. This consumption data, along with some of the building characteristics, have been released annually since 2011. We use the most recent "2016 Energy and Water Data Disclosure" data that contains information on energy consumption from 2015. This dataset has 13,223 rows of data, where each row represents a building or collection of buildings on a single lot, in one of the five boroughs of New York City. Each row of data contains feature such as: total square feet, year built, primary building activity, and energy use intensity (kWh/ft^2).

An important component of building energy demand is the amount of energy used to heat and cool the building. Heating and cooling degree day variables have been shown as useful indicators of this demand [31], and are included in the CBECS dataset, however are absent from the LL84 dataset. We augment each row in the LL84 dataset with heating degree day (HDD) and cooling degree day (CDD) features from 2015 CDD and HDD raster maps. These raster maps were calculated according to the methodology found in [32], from the daily average temperatures predicted by an aggregate of 11 climate models run at Oak Ridge National Laboratory [33]. We further join the LL84 dataset to the New York City PLUTO dataset [34] in order to get more information, such as the number of floors, for each building in the LL84 dataset. After this processing, we have information on 2,612 commercial buildings, which we will simply refer to as the LL84 dataset.

For more information about how we clean and process these datasets please see Appendix A.

3.2. Modeling Commercial Building Energy Consumption

We want to predict the 'Annual Major Fuel Consumption' (the `MFBTU` field in CBECS) of commercial buildings by only using some features of the buildings. We express this objective in a machine learning regression format as follows: We are given $\mathbf{X}$, the features for all buildings in the CBECS dataset, and $\mathbf{e}$, the target `MFBTU` (energy) values for all buildings, where a row, $X_{i,:}$, represents the features for building i, and an entry, e_i, is the `MFBTU` value for building i. In the remainder of the paper we focus on predicting the logarithm of the actual `MFBTU` values as we have observed that the `MFBTU` values follow an approximate log-normal distribution, and some machine learning models will be able to better estimate the values in the log-transformed normal distribution [35, 36](see Figure A.6). Specifically, we let $y_i = \log_{10}(e_i)$, and refer to $\mathbf{y}$ as our target values. We want to learn a function, $f(X_{i,:}) = \hat{y}_i$, that takes the features of a building as input, and outputs the estimated log of the energy consumption for that building, $\hat{y}_i$. From this, we predict the `MFBTU` value for a building as $\hat{e_i} = 10^{\hat{y}_i}$. To estimate f we will use machine learning models such as: linear regression, gradient boosting regression models, and random forest regressors. In general, these models attempt to tune their internal parameters, θ, to minimize some loss function, L, between the target values and values predicted by the model, i.e. solving $\min_{\theta} L(\mathbf{y}, f(\mathbf{X}; \theta))$. The loss function will be a function that penalizes inaccurate measurements, for example, mean squared error, $MSE(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Once a model is 'trained', and we are confident in its ability to *generalize* to unseen inputs, it can be used to estimate $\log_{10}$(`MFBTU`) values for buildings that are not in $\mathbf{X}$.

To evaluate the performance of the models on unseen data, i.e. to see how well the model is able to *generalize* to inputs it has not seen, we use stratified k-folds cross validation [37] on the CBECS dataset. This process involves splitting the data into k subsets, where each subset contains an equal portion of each class of building from the CBECS data, training the models on $k-1$ sets, then evaluating their performance on the single remaining testing set. This process is repeated k times so that each of the k sets is used as the testing set once. The evaluation metrics are reported as the average metric over the k iterations. Figure 2 gives a graphical representation of this method. We chose stratified k-folds cross validation over the traditional k-folds cross validation because of the class imbalance in the CBECS dataset. The CBECS dataset contains 20 unique classes of buildings (where classes are defined as PBAs). The number of samples in each class differs from 1044 "Office Buildings" to 14 "Enclosed Malls", where each class has an unique energy consumption distribution, see Figures A.5 and A.6. This cross validation evaluation is the method used in the "Evaluation" step shown in Figure 1. In all of our experiments we have set k equal to 10. During each cross validation split, we center and scale each feature in the training and testing splits based on statistics calculated from the training split (i.e. we subtract the mean and divide by the standard deviation). Finally, without loss of generality, we clip any negative predictions from any model to 0.

We use the following models in our experiments: linear regressor, ridge regressor, RBF kernel support vector regressor (SVR), elastic net regressor, linear kernel support vector regressor (linear SVR), adaboost regressor, bagging regressor, gradient boosting regressor (XGBoost), random forest regressor (RF regressor), extra trees regressor (ET regressor), multi-layer perceptron regressor (MLP regressor), and k-nearest neighbor regressor (KNN regressor). All model implementations are based on the scikit-learn Python library [38] and use the default parameter settings. To evaluate the models we use stratified k-folds cross validation with $k = 10$, and record the cross-validated mean absolute error (mean AE), median absolute error (median AE), and the r^2 between the true and predicted $\log_{10}$(`MFBTU`) values. The r^2 values calculated between the predicted values, $\hat{\mathbf{y}}$, and the actual values, $\mathbf{y}$, is given as $r^2(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean value of $\mathbf{y}$. A model which guesses the mean for every observation will have an r^2 score of 0, therefore any model which performs worse than this is "learning" the wrong relationships and will be useless. We also report the $10^{\text{Mean AE}}$ and $10^{\text{Median AE}}$, which capture the average number of multiples away the model's `MFBTU` estimate is from the true `MFBTU` value.

The CBECS dataset has many features that might not be feasible to collect in localized studies. The purpose of training models on the CBECS dataset is to later use them to estimate the energy consumption of commercial buildings in *any* city where there is building data, but no energy consumption data. Most features in the CBECS dataset such as, 'Number of Employees', 'Number of X-ray machines', or 'Insulation upgraded', are not commonly available, and therefore should not be included when training the models. It is important, however, to determine the influence of each possible feature included in the CBECS data on predicting energy consumption, in order to determine the potential benefits of additional data collection efforts. To this end, we run two sets of experiments using the methodology described in the previous two paragraphs: one that involves training the models using only a set of features that will be commonly available, or easily obtainable, in many cities and one that includes all of the features available in CBECS. We refer to the first group of features as the "common feature set"; it includes the following features: principal building activity, square feet, number of floors, heating degree days, and cooling degree days. We refer to the second group as the "extended feature set". As the "common feature set" is the set we expect to be available when using our models in specific urban areas, Figure 1 shows this set of features as common between the "Model Development" section and "Application" section.

To supplement the previous two experiments, and to

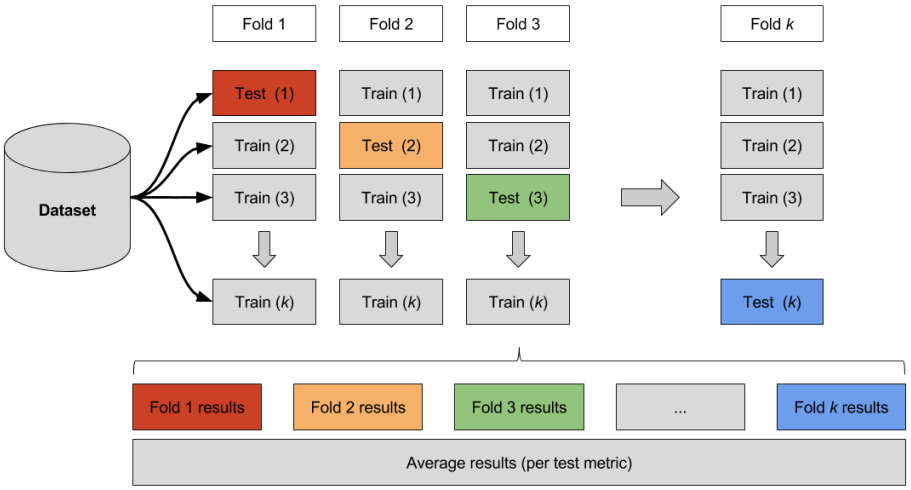

Figure 2: Graphical representation of k-folds cross validation.

aid the interpretability of our modeling process, we determine which features are the most important to the gradient boosting models (which we show are the best performing models). Feature importances in gradient boosting models are calculated as the amount of reduction in Gini impurity each feature causes over all splits for which that feature is present, over all of the trees that make up the model [39]. These values give us the relative importance of each feature included in a model, allowing us to rank the features in terms of "most useful" in the model, and compare the relative importance of features within a model. By performing this step for models trained with both the *common* and *extended* feature sets, we can see which features in the extended feature set that are not included in the common set will be most beneficial to include.

While machine learning models may perform well *within* the CBECS dataset, there is no guarantee from the cross-validated experimental results about the performance of the models on *external* data. Considering this, we validate our models on the augmented LL84 dataset, which describes the characteristics and energy consumption levels of 13,223 buildings in New York City. To do this we choose the best performing models from the first two experiments, train them with the CBECS data, then use them to predict the energy consumption values for each building in the LL84 dataset. Finally, in addition to this experiment, we perform a cross-validated experiment, with the same setup as our initial experiment on the CBECS data, using only the LL84 data (i.e. both training and testing models on the LL84 dataset). The results of this experiment will give us an upper bound on how well we can expect our models to perform on the LL84 data. The difference between these results, and the results of the previous experiment will show us how much information our general models are lacking about specific New York City energy consumption patterns.

4. Results and Discussion

Our results are presented in five parts: 1.) testing the ability of machine learning models to predict commercial energy consumption from CBECS with the *common feature set* in Section 4.1; 2.) testing the same models with the *extended feature set* from CBECS in Section 4.2; 3.) evaluating the most important features from the previous two parts in Section 4.3; 4.) validating our machine learning models on the LL84 dataset in Section 4.4, and 5.) applying our machine learning models to the city of Atlanta in Section 4.5. The methodology for all of these experiments can be found in Section 3.

4.1. Experiments with Common Features

We first experiment with training machine learning models to predict commercial building energy consumption using a common set of features from the CBECS dataset. This *common feature set* contains only the features from CBECS that are also available in the augmented LL84 data, namely: principal building activity, square footage, number of floors, heating degree days, and cooling degree days. In addition to being common with the LL84 dataset (to allow for external validation), this set of features should be widely available for commercial buildings in many metropolitan areas through local or commercial datasets. This modeling choice will let the models we train with the CBECS data be applied to a wide range of metropolitan areas.

In Table 2 we show the cross validated mean absolute error (mean AE), $10^{\text{mean AE}}$, median absolute error (median AE), $10^{\text{median AE}}$, and r^2 score of all models tested on the entire dataset. Because we perform a log_{10} transformation of the `MFBTU` values, $10^{\text{mean AE}}$ should be interpreted as the number of multiples away the predicted energy value would be from the true energy value for a building where

our model exhibits the mean absolute error [2]. The same reasoning applies for the median absolute error. For all evaluation metrics, we report the metric by averaging over the results of k folds, hence each metric also has a standard deviation σ associated with it over the folds, which we show as '+/- σ' in the results tables.

Table 2 shows that over all classes of buildings, the gradient boosting regressor (XGBoost) outperforms the other models in all metrics, with an r^2 value of 0.82 and `MFBTU` predictions that are within 1.99 times of the true `MFBTU` value on average. The linear models (linear regression, ridge regression, etc.) all perform comparatively poorly, with a maximum r^2 score of 0.53.

In Figure 3 we show a comparison of the predictions made by the linear regression model to those of the gradient boosting model. Qualitatively, this figure shows that the gradient boosting model is not systematically over- or under-estimating energy consumption, compared to the linear regression model. The linear regression models overestimate consumption at the low end of the actual energy consumption range, and underestimate at large energy consumption ranges. This unbiased attribute is important for models that will be used to create aggregate summaries of energy consumption. If we use a biased model to create building level predictions, that are then aggregated over some spatial areas (at the zipcode, or county level for example), any bias in the model's predictions will be compounded and will result in less accurate predictions.

Finally, in Table 3 we show the resulting r^2 values for each model described in Section 3.2 for each different class of building in the CBECS dataset. Consistent with the results observed in Table 2, this table shows that the gradient boosting model is the best model over all classes of building, and generally performs better on classes with more training samples. For some classes of commercial buildings, such as "Service", and "Food Service", the per class predictions, using the *common feature set*, are particularly poor. This observation partially motivates our subsequent experiments in Sections 4.2 and 4.3, as we need to determine how the models should be improved to cover the deficiencies. Most of the models are unable to give better than average guesses on the smallest two classes, "Refrigerated warehouse", and "Enclosed mall". This poor performance can possibly be explained by both the small number of samples for these classes, and the complexity of the 'features' that play a role in the determination of the true energy consumption of buildings in these categories. Buildings in the "Refrigerated warehouse" class, for example, will have expensive cooling equipment that make up the majority of their power consumption signal. Similarly, buildings in the "Laboratory" category are likely to have highly equipment with large energy demands, diminishing the impact of square footage as the most useful variable. For this reason the Federal Energy Management Program calculates the energy-savings goals of laboratories independent of square footage, which is the denominator of the goals of other building types [40].

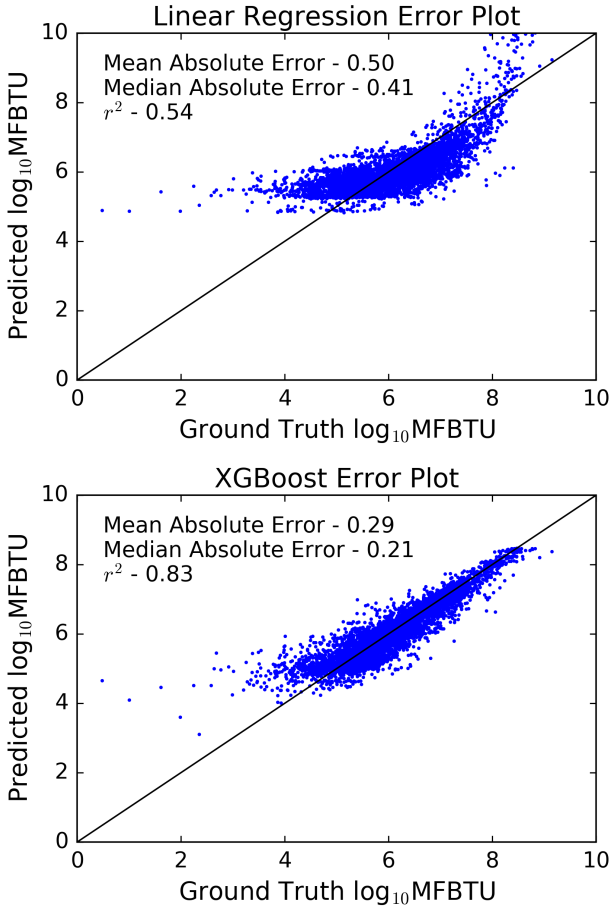

Figure 3: Error plots comparing the predicted log of `MFBTU` values versus the true log values for the Linear Regression and XGBoost models. These models were trained on 9 out of 10 stratified splits, then used to predict log `MFBTU` values for all of the data points.

4.2. Experiments with Extended Features

This section involves repeating the experiments described in Section 4.1 while using *all* of the features available in the CBECS dataset, i.e. the *extended feature set*.

In Table 4 we show how the models perform using the extended feature set, comparable to the results in Table 2. Here we see that the gradient boosting model again performs the best, with a 0.07 increase in r^2 over the same model trained using the common feature set. We also observe that the linear models perform much better, with nearly equal results to those of the gradient boosting regressors. The large gap in performance between the linear regression models trained with the common feature set versus the ones trained with the extended features, compared to the relatively small gap between the gradient boosting models, further reinforces that the gradient boosting models should be used in external applications, where few features are available. This shows that gradient boosting

[2] Given a mean or median absolute error, x, if for a building i, $x = |log_{10}(y_i) - log_{10}(\hat{y}_i)| = |e_i - \hat{e}_i|$, then that means that $\frac{e}{\hat{e}} = 10^{\pm x}$.

Common Features	**Mean Absolute Error**	$10^{\textbf{Mean AE}}$	**Median Absolute Error**	$10^{\textbf{Median AE}}$	r^2
XGBoost	**0.30 +/- 0.01**	**1.99 +/- 0.06**	**0.22 +/- 0.01**	**1.66 +/- 0.03**	**0.82 +/- 0.02**
Bagging	0.33 +/- 0.01	2.13 +/- 0.07	0.24 +/- 0.01	1.73 +/- 0.05	0.78 +/- 0.03
MLP Regressor	0.33 +/- 0.01	2.16 +/- 0.05	0.25 +/- 0.01	1.77 +/- 0.04	0.78 +/- 0.02
Random Forest Regressor	0.33 +/- 0.02	2.13 +/- 0.07	0.24 +/- 0.01	1.73 +/- 0.05	0.78 +/- 0.02
Extra Trees Regressor	0.34 +/- 0.02	2.17 +/- 0.08	0.24 +/- 0.01	1.74 +/- 0.05	0.76 +/- 0.03
SVR	0.39 +/- 0.01	2.44 +/- 0.07	0.29 +/- 0.01	1.95 +/- 0.04	0.70 +/- 0.03
KNN Regressor	0.43 +/- 0.01	2.68 +/- 0.08	0.32 +/- 0.02	2.10 +/- 0.07	0.65 +/- 0.03
AdaBoost	0.43 +/- 0.03	2.71 +/- 0.16	0.36 +/- 0.03	2.29 +/- 0.17	0.68 +/- 0.03
Linear SVR	0.51 +/- 0.02	3.28 +/- 0.15	0.40 +/- 0.02	2.54 +/- 0.11	0.52 +/- 0.04
Linear Regression	0.52 +/- 0.02	3.33 +/- 0.13	0.43 +/- 0.02	2.72 +/- 0.12	0.53 +/- 0.03
Ridge Regressor	0.52 +/- 0.02	3.33 +/- 0.13	0.43 +/- 0.02	2.72 +/- 0.12	0.53 +/- 0.03
ElasticNet	0.76 +/- 0.02	5.75 +/- 0.32	0.67 +/- 0.03	4.67 +/- 0.35	0.09 +/- 0.01
Lasso	0.79 +/- 0.02	6.17 +/- 0.35	0.69 +/- 0.03	4.92 +/- 0.38	0.00 +/- 0.00

Table 2: **Common features**. Results of all machine learning models trained on the common feature set. The mean absolute error (mean AE), median absolute error (median AE), and the r^2 values are calculated in terms of log_{10} `MFBTU` values. The $10^{\text{Mean AE}}$ and $10^{\text{Median AE}}$ columns show the average number of multiples away the model's estimate is from the true value. The values following "+/-" are the standard deviations of each metric calculated over the 10 cross validation folds.

	n	Linear Regression	ET Regressor	RF Regressor	Bagging	MLP Regressor	XGBoost
Office	1044	0.45 +/- 0.05	0.85 +/- 0.03	0.86 +/- 0.03	0.86 +/- 0.03	0.84 +/- 0.03	**0.88 +/- 0.02**
Education	580	0.37 +/- 0.05	0.80 +/- 0.04	0.80 +/- 0.04	0.81 +/- 0.04	0.80 +/- 0.04	**0.84 +/- 0.03**
Nonrefrigerated warehouse	567	0.31 +/- 0.07	0.55 +/- 0.09	0.55 +/- 0.11	0.55 +/- 0.12	0.59 +/- 0.05	**0.63 +/- 0.06**
Service	354	0.08 +/- 0.10	0.12 +/- 0.25	0.22 +/- 0.19	0.25 +/- 0.20	0.31 +/- 0.12	**0.37 +/- 0.16**
Religious worship	322	0.12 +/- 0.09	0.39 +/- 0.29	0.45 +/- 0.21	0.43 +/- 0.27	0.46 +/- 0.09	**0.57 +/- 0.16**
Retail other than mall	316	0.17 +/- 0.11	0.69 +/- 0.12	0.70 +/- 0.11	0.70 +/- 0.11	0.68 +/- 0.11	**0.73 +/- 0.11**
Public assembly	311	0.42 +/- 0.10	0.73 +/- 0.06	0.77 +/- 0.04	0.77 +/- 0.04	0.75 +/- 0.06	**0.81 +/- 0.03**
Food service	306	negative	negative	negative	negative	0.07 +/- 0.08	**0.20 +/- 0.12**
Strip shopping mall	277	0.32 +/- 0.07	0.67 +/- 0.11	0.67 +/- 0.12	0.67 +/- 0.12	0.70 +/- 0.08	**0.73 +/- 0.09**
Lodging	221	0.40 +/- 0.15	0.81 +/- 0.12	0.79 +/- 0.12	0.79 +/- 0.11	0.77 +/- 0.15	**0.83 +/- 0.11**
Inpatient health care	215	negative	**0.81 +/- 0.06**	0.80 +/- 0.07	0.79 +/- 0.07	0.80 +/- 0.07	**0.81 +/- 0.07**
Outpatient health care	131	0.21 +/- 0.24	0.69 +/- 0.24	0.72 +/- 0.20	0.72 +/- 0.20	0.64 +/- 0.23	**0.76 +/- 0.16**
Food sales	111	0.03 +/- 0.17	0.46 +/- 0.24	0.45 +/- 0.21	0.48 +/- 0.19	0.45 +/- 0.22	**0.57 +/- 0.23**
Vacant	101	negative	negative	negative	negative	0.06 +/- 0.48	**0.17 +/- 0.45**
Other	68	negative	negative	negative	negative	negative	**0.23 +/- 0.79**
Nursing	62	**0.31 +/- 0.27**	**0.31 +/- 0.79**	0.20 +/- 1.14	0.21 +/- 1.05	0.26 +/- 1.01	0.25 +/- 1.28
Public order and safety	60	0.24 +/- 0.37	0.58 +/- 0.50	0.58 +/- 0.57	0.56 +/- 0.56	0.65 +/- 0.20	**0.69 +/- 0.34**
Laboratory	23	negative	0.54 +/- 0.43	0.33 +/- 0.55	0.26 +/- 0.64	0.11 +/- 0.98	**0.59 +/- 0.52**
Refrigerated warehouse	16	negative	negative	negative	negative	negative	negative
Enclosed mall	14	negative	negative	negative	0.05 +/- 0.99	negative	**0.17 +/- 0.31**
Total	5099	0.53 +/- 0.03	0.76 +/- 0.03	0.78 +/- 0.02	0.78 +/- 0.03	0.78 +/- 0.02	**0.82 +/- 0.02**

Table 3: **Common features, per PBA**. Prediction accuracy is broken down by PBA. This table shows the r^2 scores of the predicted values by the top 5 performing models and the Linear Regression model trained with the *common feature set*. The values following "+/-" are the standard deviations of each metric calculated over the 10 cross validation folds.

models are able to combine the signals present in the common feature set to make predictions with accuracies that the linear regression models need many extra features to match. From this table we see that the mean and median absolute errors of the gradient boosting model also improve with the extended feature set. Considering the median absolute error, the gradient boosting model makes predictions within 1.48 multiples of the true `MFBTU` value. We discuss which features are most important to the gradient boosting models in Section 4.3. In Table 5 we show the r^2 of all the models per building type, similar to Table 3. The gradient boosting model is still the best performing model in most cases. It is worse for buildings in the 'Food sales' and 'Public order and safety' classes, where the linear regression model is better. Linear regression models also tie for the best model in the 'Nonrefrigerated warehouse', 'Religious worship', 'Public assembly', and 'Retail other than mall' categories. With the common feature set, the gradient boosting model had an overall r^2 of 0.82, however performed very poorly in some classes of buildings, such as "Service", "Food Service", and "Nursing". With access to the *extended feature set*, the gradient boosting models were able to improve the r^2 scores for these classes of buildings by 0.29, 0.5, and 0.53 respectively. This supports our hypothesis that for some classes of buildings, more features than those found in the common feature set are needed to reliably predict a building's energy consumption. Examples might include behavioral variables reflecting how

building equipment is utilized, which has become an increasing focus of energy-efficiency initiatives [41].

In general, these results give a rough upper bound on the ability of machine learning models to predict energy consumption from simple survey data. The gradient boosting model is able to achieve a cross validated r^2 score of 0.89 when fit with the extended feature set of CBECS features, suggesting that it is able to generalize very well to unseen data. As we show in the next section, the models trained with the extended feature set available in CBECS can indicate which features should be prioritized in any data collection efforts, in order to close the performance gap between the models trained with the common feature set.

4.3. Feature Importance

We have shown, without much surprise, that the gradient boosting models will perform better at predicting commercial building energy consumption when trained with the extended feature set, versus when trained with the common feature set. The question of 'which' features in the extended feature set are important motivates this experiment, as the answer to this question can guide future data collection efforts. We purposely keep our common feature set as simple as possible so that the trained models will be applicable to a wide range of metropolitan areas (with the assumption that many cities will be able to get access to these basic features). The most important features in the extended feature set that are not in the common set, should be the focus of data collection efforts, as they will give the largest boost to the models' predictive power.

In Table 6 we show the top 10 most important features from the gradient boosting models trained with the CBECS data using the common feature set, and the extended feature set. From the most important features in the common feature set, we observe that the square footage of a commercial building is almost 3 times important as the next most important feature, and that the number of floors feature is relatively unimportant (possibly because it will be indirectly included in the square footage).

Out of the most important features out of the extended feature set, we observe that the square footage is still the most important feature, although the 'Number of employees' and 'Total hours open per week' are the second and third most important features, both of which are not present in the common feature set. This idea, that building occupancy is an important feature, is supported up by a recent study that shows that occupancy rates of commercial buildings are important to consider in energy savings measures [42]. The climate related features ("heating degree days", "cooling degree days") are relatively less important. Similarly, other features that have shown to be important to energy consumption calculations, such as building envelope insulating characteristics [43](e.g. wall and roof constriction materials), are not included in the model's top ten important features. We note that the important variables are determined based on how much they contribute to the model's decision, and are not necessarily directly related to the actual calculation of building energy consumption. This relationship is illustrated by looking at the most important features for estimated energy consumption within buildings of a given principal building activity. From Table 5 we observed that in the 'Service', 'Religious worship' 'Food Service', 'Vacant', 'Other', and 'Nursing' classes of buildings, adding all the features increased the r^2 score of the gradient boosting model by over 0.2. When we train a gradient boosting regressor on *just* samples from the 'Service' class of buildings we observe that the most important feature is 'Total hours open per week', instead of 'Square footage'. This suggests that for some PBAs, the common feature set does not contain the correct signals to reproduce the `MFBTU` targets. This also suggests that within class consumption differences are explained by features that are not necessarily relevant to all classes. For further results on the features that are most important per PBA, see the"Feature importances" notebook in the accompanying GitHub repository.

4.4. Validation

The CBECS data provides us with a national dataset of commercial buildings that we can train and test our models with, however this does not let us make conclusions about the efficacy of the models for predicting energy consumption in a particular metropolitan area. We validate our models by using them to predict the energy consumption values of buildings in New York City, for which the true consumption values are known.

We train a gradient boosting model on *all* of the CBECS data with the common feature set, then apply that model on the 2,612 commercial buildings from the augmented LL84 dataset and record the same metrics from previous experiments. This results in a mean absolute error of 0.25, median absolute error of 0.15, and r^2 value of 0.50. Consistent with the interpretation given in Section 4.1, the predicted energy consumption values have a mean of 1.78 multiples away from the true value, and a median of 1.41 multiples away from the true value. These mean and median errors are *better* than the best values observed from the CBECS dataset (both in the reduced features, and all features cases), although the r^2 fit is worse.

We further train and test all the machine learning models on solely the LL84 dataset using the same methodology as in the first two experiments. The results from this test are shown in Table 7. The gradient boosting model is again the best performing model, however, surprisingly, when the gradient boosting model is trained on the LL84 data, it only performs slightly better than the model that was trained on the CBECS data. Specifically, the gradient boosting model has an r^2 of 0.54 from training on the LL84 data, compared to an r^2 value of 0.51 from training on the CBECS data. The mean absolute error is 0.01 lower, and the median absolute error is the same. These results provide strong evidence that our model trained with the

Extended Features	**Mean Absolute Error**	$10^{\text{Mean AE}}$	**Median Absolute Error**	$10^{\text{Median AE}}$	r^2
XGBoost with Common Features	0.30 +/- 0.01	1.99 +/- 0.06	0.22 +/- 0.01	1.66 +/- 0.03	0.82 +/- 0.02
XGBoost	**0.23 +/- 0.01**	**1.69 +/- 0.02**	**0.17 +/- 0.01**	**1.48 +/- 0.03**	**0.89 +/- 0.01**
Linear Regression	0.24 +/- 0.01	1.75 +/- 0.02	0.19 +/- 0.01	1.53 +/- 0.04	0.88 +/- 0.01
Ridge Regressor	0.24 +/- 0.01	1.75 +/- 0.02	0.19 +/- 0.01	1.53 +/- 0.04	0.88 +/- 0.01
SVR	0.25 +/- 0.01	1.79 +/- 0.04	0.19 +/- 0.01	1.53 +/- 0.03	0.87 +/- 0.01
Bagging	0.25 +/- 0.01	1.79 +/- 0.04	0.18 +/- 0.01	1.53 +/- 0.04	0.87 +/- 0.02
Random Forest Regressor	0.25 +/- 0.01	1.79 +/- 0.04	0.18 +/- 0.01	1.53 +/- 0.04	0.87 +/- 0.01
Extra Trees Regressor	0.25 +/- 0.01	1.79 +/- 0.04	0.19 +/- 0.01	1.54 +/- 0.03	0.87 +/- 0.01
Linear SVR	0.26 +/- 0.01	1.80 +/- 0.03	0.20 +/- 0.01	1.58 +/- 0.04	0.87 +/- 0.01
AdaBoost	0.32 +/- 0.01	2.07 +/- 0.05	0.26 +/- 0.01	1.80 +/- 0.05	0.82 +/- 0.01
KNN Regressor	0.37 +/- 0.01	2.34 +/- 0.06	0.29 +/- 0.01	1.93 +/- 0.04	0.75 +/- 0.02
MLP Regressor	0.45 +/- 0.02	2.82 +/- 0.11	0.36 +/- 0.02	2.31 +/- 0.10	0.64 +/- 0.03
ElasticNet	0.60 +/- 0.02	4.00 +/- 0.20	0.51 +/- 0.02	3.26 +/- 0.16	0.40 +/- 0.01
Lasso	0.79 +/- 0.02	6.17 +/- 0.35	0.69 +/- 0.03	4.92 +/- 0.38	0.00 +/- 0.00

Table 4: **Extended features**. Results of all machine learning models trained/tested with the extended feature set, compared to the XGBoost model results from Table 2. The mean absolute error (mean AE), median absolute error (median AE), and the r^2 values are calculated in terms of log_{10} `MFBTU` values. The $10^{\text{Mean AE}}$ and $10^{\text{Median AE}}$ columns show the average number of multiples away the model's estimate is from the true value.

	n	Linear Regression	RF Regressor	Bagging	MLP Regressor	XGBoost	XGBoost with Common Features
Office	1044	0.90 +/- 0.01	0.90 +/- 0.02	0.90 +/- 0.02	0.66 +/- 0.07	**0.91 +/- 0.01**	0.88 +/- 0.02
Education	580	0.84 +/- 0.03	0.85 +/- 0.03	0.85 +/- 0.03	0.34 +/- 0.16	**0.87 +/- 0.02**	0.84 +/- 0.03
Nonrefrigerated warehouse	567	**0.81 +/- 0.03**	0.77 +/- 0.06	0.77 +/- 0.06	0.59 +/- 0.10	**0.81 +/- 0.04**	0.63 +/- 0.06
Service	354	0.65 +/- 0.11	0.61 +/- 0.12	0.60 +/- 0.12	0.14 +/- 0.27	**0.66 +/- 0.10**	0.37 +/- 0.16
Religious worship	322	**0.77 +/- 0.07**	0.72 +/- 0.08	0.72 +/- 0.09	0.34 +/- 0.35	**0.77 +/- 0.09**	0.57 +/- 0.16
Retail other than mall	316	**0.86 +/- 0.05**	0.81 +/- 0.08	0.81 +/- 0.08	0.45 +/- 0.19	**0.86 +/- 0.06**	0.73 +/- 0.11
Public assembly	311	**0.89 +/- 0.03**	0.86 +/- 0.04	0.86 +/- 0.03	0.60 +/- 0.15	**0.89 +/- 0.02**	0.81 +/- 0.03
Food service	306	0.66 +/- 0.07	0.56 +/- 0.13	0.56 +/- 0.15	negative	**0.70 +/- 0.09**	0.20 +/- 0.12
Strip shopping mall	277	0.85 +/- 0.06	0.87 +/- 0.05	0.87 +/- 0.04	0.21 +/- 0.27	**0.91 +/- 0.03**	0.73 +/- 0.09
Lodging	221	0.85 +/- 0.07	0.82 +/- 0.11	0.83 +/- 0.11	0.24 +/- 0.27	**0.87 +/- 0.07**	0.83 +/- 0.11
Inpatient health care	215	0.71 +/- 0.12	**0.84 +/- 0.06**	**0.84 +/- 0.07**	negative	**0.84 +/- 0.08**	0.81 +/- 0.07
Outpatient health care	131	0.83 +/- 0.10	0.82 +/- 0.11	0.82 +/- 0.11	0.45 +/- 0.29	**0.84 +/- 0.11**	0.76 +/- 0.16
Food sales	111	**0.74 +/- 0.17**	0.61 +/- 0.24	0.60 +/- 0.26	negative	0.68 +/- 0.19	0.57 +/- 0.23
Vacant	101	0.40 +/- 0.19	0.29 +/- 0.39	0.23 +/- 0.53	negative	**0.48 +/- 0.20**	0.17 +/- 0.45
Other	68	0.61 +/- 0.30	0.52 +/- 0.53	0.54 +/- 0.49	negative	**0.64 +/- 0.28**	0.23 +/- 0.79
Nursing	62	0.71 +/- 0.21	0.56 +/- 0.72	0.55 +/- 0.77	negative	**0.78 +/- 0.22**	0.25 +/- 1.28
Public order and safety	60	**0.83 +/- 0.12**	0.77 +/- 0.24	0.78 +/- 0.23	negative	0.80 +/- 0.21	0.69 +/- 0.34
Laboratory	23	0.58 +/- 0.32	0.16 +/- 0.97	0.17 +/- 0.94	negative	0.45 +/- 0.70	**0.59 +/- 0.52**
Refrigerated warehouse	16	negative	negative	negative	negative	negative	negative
Enclosed mall	14	negative	negative	negative	negative	negative	**0.17 +/- 0.31**
Total	5099	0.88 +/- 0.01	0.87 +/- 0.01	0.87 +/- 0.02	0.64 +/- 0.03	**0.89 +/- 0.01**	0.82 +/- 0.02

Table 5: **Extended features, per PBA**. Prediction accuracy is broken down by PBA. This table shows the r^2 scores of the predicted values by the top 4 performing models and the Linear Regression model, trained/tested with the extended feature set, compared to the XGBoost model results from Table 3.

CBECS data is able to be applied to specific metropolitan areas while maintaining reasonable results.

4.5. Atlanta Case Study

We show how our models can be applied to a large metropolitan area by creating commercial energy consumption summaries for the 20 county Atlanta metropolitan area. To do this, we applied the CBECS-trained gradient boosting regression model to the 73,388 commercial buildings in Atlanta from the CoStar real estate database[3]. We supplement the CoStar data with the 2017 heating and cooling degree day data from the Oak Ridge Climate models (using the same methodology we used to create the Augmented LL84 dataset, see Appendix A). Figure 4 shows a map of the *total estimated energy consumption* values from all commercial buildings aggregated at the Transportation Analysis Zone (TAZ) level, and a map of the *median estimated energy use intensity across commercial buildings* in each TAZ[4]. Both maps are colored according to quantile binning with 10 bins. As we show in Figure 3, the gradient boosting model does not systematically

[3]This dataset is continuously updated, see `http://www.costar.com/`. The dataset that we use was downloaded in March of 2017.

[4]The energy use intensity of a building is its total energy consumption divided by its square footage.

Common Feature Set			Extended Feature Set		
Feature Name	**Feature Description**	**Importance**	**Feature Name**	**Feature Description**	**Importance**
SQFT	Square footage	0.3634	SQFT	Square footage	0.1391
CDD65	Cooling degree days (base 65)	0.1153	NWKER	Number of employees	0.0576
HDD65	Heating degree days (base 65)	0.1125	WKHRS	Total hours open per week	0.0557
PBA 5	Non-refrigerated warehouse	0.0569	ZMFBTU	Imputed major fuels consumption	0.0312
PBA 1	Vacant	0.0524	MONUSE	Months in use	0.0299
PBA 6	Food sales	0.0412	NGUSED	Natural gas used	0.0295
PBA 15	Food service	0.0384	HDD65	Heating degree days (base 65)	0.0293
PBA 23	Strip shopping mall	0.0348	HEATP	Percent heated	0.0278
PBA 12	Religious worship	0.0345	CDD65	Cooling degree days (base 65)	0.0224
PBA 4	Laboratory	0.0282	NWKERC	Number of employees category	0.0221

Table 6: Top 10 important features for the XGBoost model trained with all data using both the common and extended feature sets.

	Mean Absolute Error	$10^{\textbf{Mean AE}}$	**Median Absolute Error**	$10^{\textbf{Median AE}}$	r^2
XGBoost - CBECS	0.25	1.78	0.15	1.41	0.51
XGBoost	**0.24 +/- 0.02**	**1.75 +/- 0.09**	**0.15 +/- 0.01**	**1.40 +/- 0.03**	**0.54 +/- 0.09**
SVR	0.25 +/- 0.02	1.77 +/- 0.10	**0.15 +/- 0.01**	**1.40 +/- 0.03**	0.51 +/- 0.11
Linear SVR	0.28 +/- 0.02	1.92 +/- 0.08	0.17 +/- 0.00	1.50 +/- 0.01	0.42 +/- 0.05
MLP Regressor	0.28 +/- 0.04	1.92 +/- 0.17	0.17 +/- 0.02	1.48 +/- 0.06	0.44 +/- 0.13
Linear Regression	0.29 +/- 0.02	1.96 +/- 0.10	0.19 +/- 0.01	1.56 +/- 0.05	0.44 +/- 0.08
Ridge Regressor	0.29 +/- 0.02	1.96 +/- 0.10	0.19 +/- 0.01	1.56 +/- 0.05	0.44 +/- 0.08
Bagging	0.29 +/- 0.02	1.95 +/- 0.09	0.18 +/- 0.01	1.50 +/- 0.04	0.43 +/- 0.08
Random Forest Regressor	0.29 +/- 0.02	1.95 +/- 0.10	0.18 +/- 0.02	1.51 +/- 0.05	0.43 +/- 0.08
Extra Trees Regressor	0.30 +/- 0.03	2.00 +/- 0.12	0.18 +/- 0.01	1.51 +/- 0.05	0.39 +/- 0.09
KNN Regressor	0.30 +/- 0.03	2.01 +/- 0.15	0.19 +/- 0.02	1.53 +/- 0.06	0.40 +/- 0.12
AdaBoost	0.42 +/- 0.07	2.67 +/- 0.43	0.30 +/- 0.04	2.01 +/- 0.20	0.14 +/- 0.22
Lasso	0.45 +/- 0.01	2.80 +/- 0.04	0.33 +/- 0.01	2.13 +/- 0.06	negative
ElasticNet	0.45 +/- 0.01	2.80 +/- 0.04	0.33 +/- 0.01	2.13 +/- 0.06	negative

Table 7: **LL84 Validation**. Comparison of the best external model tested on the LL84 dataset (out of sample validation result) to all machine learning models trained and tested on the LL84 dataset. The first row, 'XGBoost - CBECS', is the best external model and shows the results from applying the XGBoost model trained on all of the CBECS data, to all of the LL84 data. The remaining rows show the cross validated results on models trained and tested on the LL84 dataset. All results are shown with models using the common feature set. The mean absolute error (mean AE), median absolute error (median AE), and the r^2 values are calculated in terms of log_{10} `MFBTU` values. The $10^{\text{Mean AE}}$ and $10^{\text{Median AE}}$ columns show the average number of multiples away the model's estimate is from the true value.

overestimate or underestimate energy consumption values. Considering this, we expect the modeling errors for individual building's energy consumptions to cancel out in the aggregate energy consumption estimates. From the energy consumption map we observe that the greatest energy consuming TAZs are clustered in the "Downtown" and "Midtown" parts of Atlanta (in the center of the mapped area), as well as the suburban cities surrounding Atlanta. Commercial energy consumption is generally greater in TAZs immediately adjacent to the I-85 and I-75 highways that cut diagonally through the city, from southwest to northeast, and southeast to northwest, respectively. The median energy use intensity map shows that, although TAZs in the "Downtown" and "Midtown" parts of Atlanta consume more energy, they are more energy efficient on average than TAZs in the northern Buckhead suburb.Similarly, these maps show some TAZs in the surrounding suburban cities that have disproportionately high energy consumption compared to their surroundings, indicating locations where energy efficiency building retrofits should be considered [44]. Our total estimated commercial energy consumption for Atlanta is 126.62 billion kBTUs/year, which would make up 0.7% of the total annual commercial energy consumption of the U.S. in 2016 [45].

We note that the city of Atlanta's new energy benchmarking ordinance for commercial buildings may change this geography of energy consumption in commercial buildings. It aims to achieve a 20% reduction of energy consumption in Atlanta's private and City-owned buildings over 25,000 square feet, by 2030 [46]. If successful, this could curb the energy consumption peaks shown in the Downtown and Midtown TAZs.

Predicting the commercial building energy consumption landscape in Atlanta is just a single example of what our models are capable of doing. The results illustrate the ability to use our US-wide CBECS-based commercial building energy consumption models for various kinds of analysis and scenario evaluations that can inform urban planning and policy making. Our models can be applied to other metropolitan areas for which there is building level data available by following the instructions given in

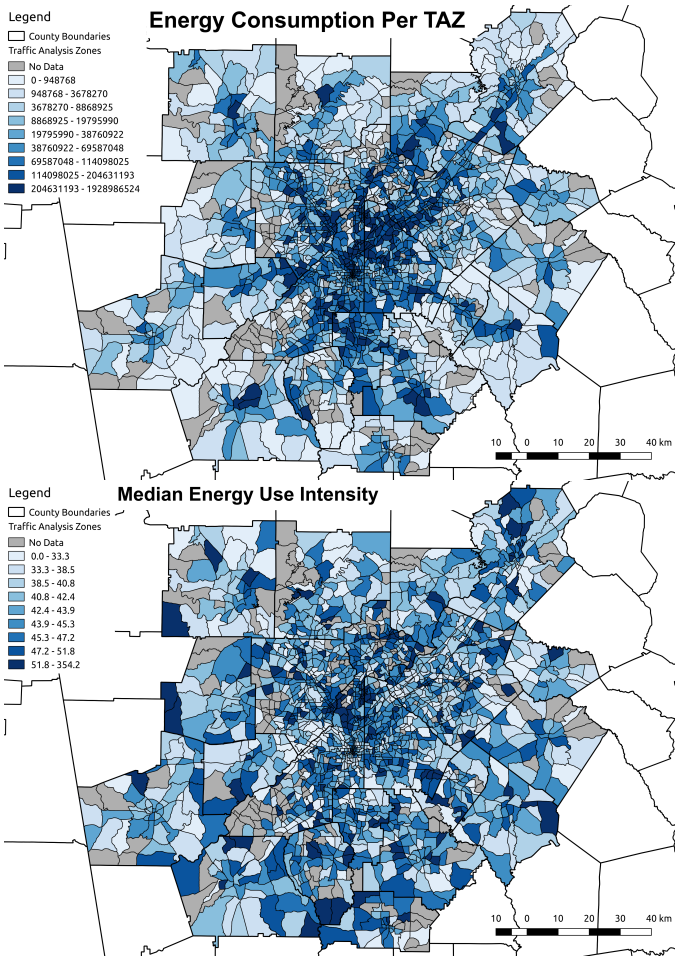

Figure 4: Estimated energy consumption (kBTU/year) of commercial buildings in Atlanta, aggregated per TAZ (**top**). Median estimated energy use intensity of commercial buildings per TAZ (**bottom**).

our GitHub repository[5].

5. Conclusion

We create machine learning models trained on the CBECS dataset for estimating commercial building energy consumption, analyze feature importance in our models, then validate the models on external data from New York City, and create commercial building energy consumption estimates for the Atlanta metropolitan area. An important aspect of our work involves limiting the information about each building used by the machine learning models to five *commonly available* features, so that our models can be used in a wide range of metropolitan areas without requiring expensive data collection efforts. We find that some of the models are able to perform acceptably well under this constraint, and the gradient boosting models are able to make predictions that are on average under 1.78 multiples away from the true value on the external validation dataset. Although this error is too large for analyzing the energy consumption of any specific building, when the models are used to make predictions for all the buildings in entire metropolitan areas (where individual prediction errors will cancel out when aggregated), as we show for Atlanta, they can offer useful insights into a city's commercial energy consumption landscape. Furthermore, our analysis of important features used by the machine learning models will serve to drive future data collection efforts that could help maximize the accuracy of the models.

Admittedly, our modeling approach has several limitations. One limitation is that our models can only be used in the United States, as the CBECS training dataset only provides a validated statistical representation of commercial buildings in the US. When more detailed commercial building datasets become available in other countries, the same methodology we use in this paper will be applicable. Another limitation of our models is the trade-off between accuracy and data requirements which we show with the results comparing the accuracy between models using our *common feature set* and *extended feature set*. Using more building level variables when modeling commercial building energy consumption predictably gives better results, however will limit the usefulness of the models in metropolitan areas that do not have access to such detailed data. While we validate the results of our models trained with "common features" using data from New York City, and show they give reasonable results, the models will not be able to capture the energy consumption patterns of buildings that are extremely efficient or inefficient.

Related works in statistical building energy consumption predictions have not been thoroughly explored using more complicated machine learning models, such as gradient boosting regression models (the best performing model). Our results show that these higher capacity models are more capable of exploiting a limited number of features to achieve better performance than possible with general linear models. Currently, city planners and policy makers will be able to use our models to create summary commercial energy consumption maps to assist with future planning, and achieving sustainable development goals. As more energy consumption data becomes available through crucial data collection efforts such as New York City's Local Law 84, statistical models for estimating building energy consumption will become more powerful, and will be able to give more confident estimates, further improving the capability for understanding our urban environments.

Our work has several opportunities for future development that we are interested in pursuing. One important aspect of machine learning modeling is model selection. Many of the machine learning models that we reference in this paper have hyperparameters that can be tuned to further increase their predictive ability. As maximizing the predictive performance of any specific model was not the focus of this paper, we opted to leave the hyperparameter settings at their default values, and instead focus on the role of available features in model performance. We are in-

[5] `https://github.com/SEI-ENERGY/Commercial-Energy`

terested in performing a broad computational study that focuses on maximizing model performance on the task of predicting building energy consumption. Secondly, we are interested in applying the models developed in this study to all the commercial buildings in major metropolitan areas as part of a summary of total metropolitan energy consumption. Finally, an analysis of the potential impact of future climate scenarios and alternative patterns of urban growth can be informed by the models developed in this paper. We are interested in using climate model projections to evaluate how the energy consumption landscapes of different cities will change under various climate scenarios.

Acknowledgements

We would like to thank Jack Fellows and Deeksha Rastogi from the Climate Change Science Institute at the Oak Ridge National Laboratory for providing the ensemble climate model data needed to estimate the heating and cooling degree day rasters.

All authors were partially supported by the Georgia Tech Strategic Energy Institute. Robinson and Dilkina were also partially supported by BCS-1638268 (CRISP: Sustainable and Resilient Design of Interdependent Water and Energy Systems at the Infrastructure-Human-Resource Nexus) and CCF-1522054 (COMPUSTNET: Expanding Horizons of Computational Sustainability). Brown and Dilkina were partially supported by the Georgia Tech Brook Byers Institute for Sustainable Systems.

[1] S. Hankey, J. D. Marshall, Impacts of urban form on future us passenger-vehicle greenhouse gas emissions, Energy Policy 38 (9) (2010) 4880–4887.

[2] M. A. Brown, F. Southworth, A. Sarzynski, The geography of metropolitan carbon footprints, Policy and Society 27 (4) (2009) 285–304.

[3] J. Norman, H. L. MacLean, C. A. Kennedy, Comparing high and low residential density: life-cycle analysis of energy use and greenhouse gas emissions, Journal of urban planning and development 132 (1) (2006) 10–21.

[4] B. G. Nichols, K. M. Kockelman, Life-cycle energy implications of different residential settings: Recognizing buildings, travel, and public infrastructure, Energy Policy 68 (2014) 232–242.

[5] S. Guhathakurta, E. Williams, Impact of urban form on energy use in central city and suburban neighborhoods: Lessons from the phoenix metropolitan region, Energy Procedia 75 (2015) 2928–2933.

[6] P. G. Newman, J. R. Kenworthy, Cities and automobile dependence: An international sourcebook, 1989.

[7] D. Brownstone, Key relationships between the built environment and vmt, Transportation Research Board 7.

[8] V. M. Garikapati, D. You, W. Zhang, R. M. Pendyala, S. Guhathakurta, M. A. Brown, B. Dilkina, Estimating household travel energy consumption in conjunction with a travel demand forecasting model, Transportation Research Record: Journal of the Transportation Research Board (2017) (forthcoming).

[9] W. Zhang, S. Guhathakurta, C. Ross, Trends in automobile energy use and ghg emissions in suburban and inner city neighborhoods: Lessons from metropolitan phoenix, usa, Energy Procedia 88 (2016) 82–87.

[10] Committee for the Study on the Relationships Among Development Patterns, Vehicle Miles Traveled, and Energy Consumption, Board on Energy and Environmental Systems, Transportation Research Board, National Research Council, Driving and the Built Environment: The Effects of Compact Development on Motorized Travel, Energy Use, and CO2 Emissions – Special Report 298, National Academies Press, 2010.

[11] U.S. Energy Information Administration, U.S. Energy Flow, 2015, `https://www.eia.gov/totalenergy/data/monthly/pdf/flow/total_energy.pdf` (2016).

[12] European Commission Buildings, `https://ec.europa.eu/energy/en/topics/energy-efficiency/buildings`, accessed July 29, 2017.

[13] L. Oswaldo, D. Urge-Vorsatz, A. Z. Ahmed, H. Akbari, P. Bertoldi, L. F. Cabeza, N. Eyre, A. Gadgil, L. D. Harvey, Y. Jiang, E. Liphoto, S. Mirasgedis, S. Murakami, J. Parikh, C. Pyke, M. V. Vilarino, Climate Change 2014: Mitigation of Climate Change: Working Group III Contribution to the IPCC Fifth Assessment Report, Cambridge University Press, 2014, Ch. 9, p. 671–738.

[14] H.-x. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renewable and Sustainable Energy Reviews 16 (6) (2012) 3586–3592.

[15] U.S. Energy Information Administration, Commercial Building Energy Consumption Survey, `https://www.eia.gov/consumption/commercial/` (2012).

[16] Z. Li, Y. Han, P. Xu, Methods for benchmarking building energy consumption against its past or intended performance: An overview, Applied Energy 124 (2014) 325–334.

[17] G. Tardioli, R. Kerrigan, M. Oates, O. James, D. Finn, Data driven approaches for prediction of building energy consumption at urban level, Energy Procedia 78 (2015) 3378–3383.

[18] I. Korolija, L. Marjanovic-Halburd, Y. Zhang, V. I. Hanby, Uk office buildings archetypal model as methodological approach in development of regression models for predicting building energy consumption from heating and cooling demands, Energy and Buildings 60 (2013) 152–162.

[19] H. A. Nielsen, H. Madsen, Modelling the heat consumption in district heating systems using a grey-box approach, Energy and Buildings 38 (1) (2006) 63–71.

[20] P. A. Mathew, L. N. Dunn, M. D. Sohn, A. Mercado, C. Custudio, T. Walter, Big-data for building energy performance: Lessons from assembling a very large national database of building energy use, Applied Energy 140 (2015) 85–93.

[21] R. E. Brown, T. Walter, L. N. Dunn, C. Y. Custodio, P. A. Mathew, L. Berkeley, Getting real with energy data: Using the buildings performance database to support data-driven analyses and decision-making, in: Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings, 2014, pp. 11–49.

[22] F. Boulaire, A. Higgins, G. Foliente, C. McNamara, Statistical modelling of district-level residential electricity use in nsw, australia, Sustainability science 9 (1) (2014) 77–88.

[23] P. Kuusela, I. Norros, R. Weiss, T. Sorasalmi, Practical log-normal framework for household energy consumption modeling, Energy and Buildings 108 (2015) 223–235.

[24] C. E. Kontokosta, Predicting building energy efficiency using new york city benchmarking data, Proceedings of the 2012 ACEEE Summer Study on Energy Efficiency in Buildings, Washington, DC, American Council for an Energy-Efficient Economy.

[25] G. K. Tso, K. K. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy 32 (9) (2007) 1761–1768.

[26] C. Fan, F. Xiao, S. Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Applied Energy 127 (2014) 1–10.

[27] L. Wei, W. Tian, E. A. Silva, R. Choudhary, Q. Meng, S. Yang, Comparative study on machine learning for urban building energy analysis, Procedia Engineering 121 (2015) 285–292.

[28] M. Yalcintas, U. Aytun Ozturk, An energy benchmarking model based on artificial neural network method utilizing us commer-

cial buildings energy consumption survey (cbecs) database, International Journal of Energy Research 31 (4) (2007) 412–421.
[29] B. Howard, L. Parshall, J. Thompson, S. Hammer, J. Dickinson, V. Modi, Spatial distribution of urban building energy consumption by end use, Energy and Buildings 45 (2012) 141–151.
[30] New York City Mayor's Office of Sustainability, Local Law 84 Data Disclosures, http://www.nyc.gov/html/gbee/html/plan/ll84_scores.shtml (2016).
[31] M. Christenson, H. Manz, D. Gyalistras, Climate warming impact on degree-days and building energy demand in switzerland, Energy Conversion and Management 47 (6) (2006) 671–686.
[32] M. A. Brown, M. Cox, B. Staver, P. Baer, Modeling climate-driven changes in us buildings energy demand, Climatic Change 134 (1-2) (2016) 29–44.
[33] M. Ashfaq, D. Rastogi, R. Mei, S.-C. Kao, S. Gangrade, B. S. Naz, D. Touma, High-resolution ensemble projections of near-term regional climate over the continental united states, Journal of Geophysical Research: Atmospheres 121 (17) (2016) 9943–9963, 2016JD025285. doi:10.1002/2016JD025285.
URL http://dx.doi.org/10.1002/2016JD025285
[34] New York City Department of City Planning, PLUTO 16v2, https://www1.nyc.gov/site/planning/data-maps/open-data.page (2016).
[35] O. N. Keene, The log transformation is special, Statistics in medicine 14 (8) (1995) 811–819.
[36] M. Kuhn, K. Johnson, Applied predictive modeling, Vol. 26, Springer, 2013.
[37] R. Kohavi, et al., A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Ijcai, Vol. 14, Stanford, CA, 1995, pp. 1137–1145.
[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[39] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.
[40] United States Department of Energy, Federal Building Energy Use Benchmarking Guidance, August 2014 Update: Use of Energy and Water Efficiency Measures in Federal Buildings (42 U.S.C. § 8253[f]), https://energy.gov/sites/prod/files/2014/09/f18/benchmarking_guidance08-2014.pdf (2014).
[41] P. C. Stern, K. B. Janda, M. A. Brown, L. Steg, E. L. Vine, L. Lutzenhiser, Opportunities and insights for reducing fossil fuel consumption by households and organizations, Nature Energy 1 (2016) 16043.
[42] Y.-S. Kim, J. Srebric, Impact of occupancy rates on the building electricity consumption in commercial buildings, Energy and Buildings.
[43] Y. Huang, J.-l. Niu, T.-m. Chung, Study on performance of energy-efficient retrofitting measures on commercial building external walls in cooling-dominant cities, Applied energy 103 (2013) 97–108.
[44] T. Hong, M. A. Piette, Y. Chen, S. H. Lee, S. C. Taylor-Lange, R. Zhang, K. Sun, P. Price, Commercial building energy saver: an energy retrofit analysis toolkit, Applied Energy 159 (2015) 298–309.
[45] Energy Information Administration, Monthly Energy Review, https://www.eia.gov/totalenergy/data/monthly/pdf/sec2_7.pdf (2017).
[46] City of Atlanta Adopts Progressive Energy Policy to Tackle Commercial Energy Use, http://www.atlantaga.gov/index.aspx?page=672&recordid=3498 (2015).

Appendix A. Data Preprocessing

Commercial Building Energy Consumption Survey (CBECS)

The procedure we use to preprocess the raw 2012 microdata into the format we use in the study is as follows[6]:

- Remove rows where the `ZMFBTU` field is not 2 or 9 (i.e. only keep rows where our target value, `MFBTU` is present).
- Replace values where `NFLOORS` = 994 with 20. Replace values where `NFLOORS` = 995 with 30. The 994 value represents 15 to 25 floors and the 995 value represent greater than 25 floors. These choices of values will let algorithmic approaches treat the `NFLOORS` feature as an integer value.
- Discard all features that start with the letter 'Z' (i.e. features that take the form 'Z*****'). These fields report whether or not another feature is: 'reported', 'imputed', 'estimated', 'missing', or 'inapplicable', and will not be useful for our models.
- Discard all of the `FINALWT` columns.
- Discard features (columns) that have over 25% of missing values, then impute remaining missing values per feature, with the most common value for that feature.
- Perform a one-hot encoding on the `PBA` feature by removing the original feature, and adding 20 new features, where each feature represents a particular PBA. For each row of data, the new feature that represents the original PBA value will be set to 1, and the remaining will be set to 0.

This process will result in a data table, $\mathbf{X}$, containing 179 features for each of the 5099 buildings, $\mathbf{X} \in \mathbb{R}^{5099\times179}$, with a single target vector, Y, representing the `MFBTU` value for each building, $Y \in \mathbb{R}^{5099}$. This is trimmed down from the original dataset that had 6720 rows and 1119 features. Each building in $\mathbf{X}$ falls into one of 20 different classes according to the buildings principal building activity, or `PBA`. The numbers of building per `PBA` can be seen in Figure A.5. The distribution of the `MFBTU` values in Y can be seen in Figure A.6.

Augmented Local Law 84 Dataset (LL84)

We join the latest LL84 dataset from the 2015 calendar year with the latest PLUTO 16v2 dataset, and cooling and heating degree day rasters from climate model outputs. This process will enable to construct the same *common feature set* for NYC buildings that we use in the CBECS dataset.

The steps we follow to get the final LL84 dataset used in the study are as follows:

[6]For code to reproduce this cleaning process, see the GitHub repository at `https://github.com/SEI-ENERGY/Commercial-Energy`

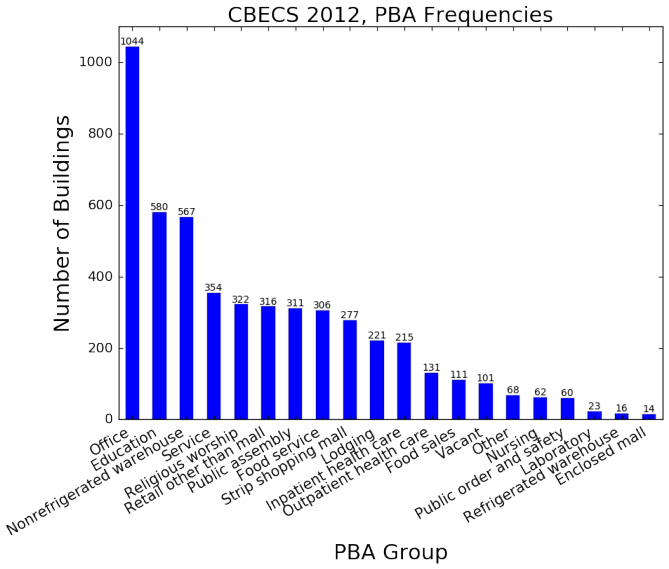

Figure A.5: Number of samples of each class of building (PBA), after preprocessing, in the CBECS dataset.

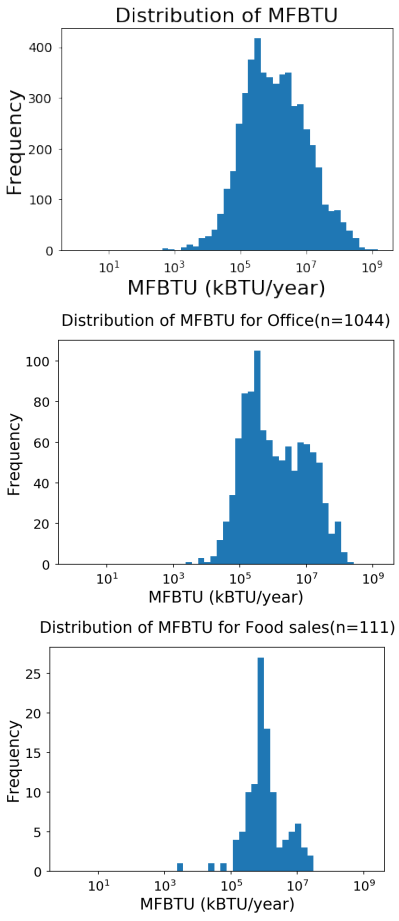

Figure A.6: MFBTU distributions for: all CBECS data (**top**), buildings in the "Office" class (**middle**), and buildings in the "Food Sales" class (**bottom**). The top panel shows how the `MFBTU` target values follow a log-normal distribution, and therefore, how the log transformation of the `MFBTU` values will follow a normal distribution.

1. Each building/tax lot in New York City has an unique identifying number called the Borough, Block, and Lot (BBL) number. We use the BBL field from both the LL84 and Pluto datasets to perform an inner join on the two datasets.
2. We drop all rows for which any of the following fields are missing: "Primary Property Type - Self Selected", "Site EUI (kBtu/ft^2)", "Property GFA - Self-reported (ft^2)", or "NumFloors".
3. We map the "Primary Property Type - Self Selected" field to CBECS "Principal Building Activity" field according to the custom mapping defined in Table A.8. Any rows with a "Primary Property Type - Self Selected" value in the LL84 field that is not present in Table A.8, such as "Multifamily Housing", are dropped from the dataset. We then perform an one-hot encoding of this mapped `PBA` field using the same method used in the CBECS processing steps.
4. The PLUTO dataset comes with shapefiles that have a polygon for each row (MapPLUTO). We calculate the centroid points for each of these shapes, which lets us associate a latitude/longitude point with each row in the LL84 dataset.
5. We use the latitude/longitude points for each row to lookup the cooling and heating degree day values from rasters derived from an average of 11 climate models run at Oak Ridge National Laboratory. The heating and cooling degree values are calculated from the 2015 average daily temperature as reported by the climate models, with a temperature of 65 Fahrenheit used as the base value for the degree day calculation. The rasters are shown in Figure A.7.
6. For each row, we convert the energy use intensity value, "Site EUI (kBtu/ft^2)", to total energy use by multiplying by the total square footage. These values are then used as the target values in the LL84 dataset.

The result of this process is a set of 2,612 commercial buildings from New York City, each with all features in the *common feature set*: square footage, number of floors, cooling degree days, heating degree days, and PBA, as well as the total energy consumption value.

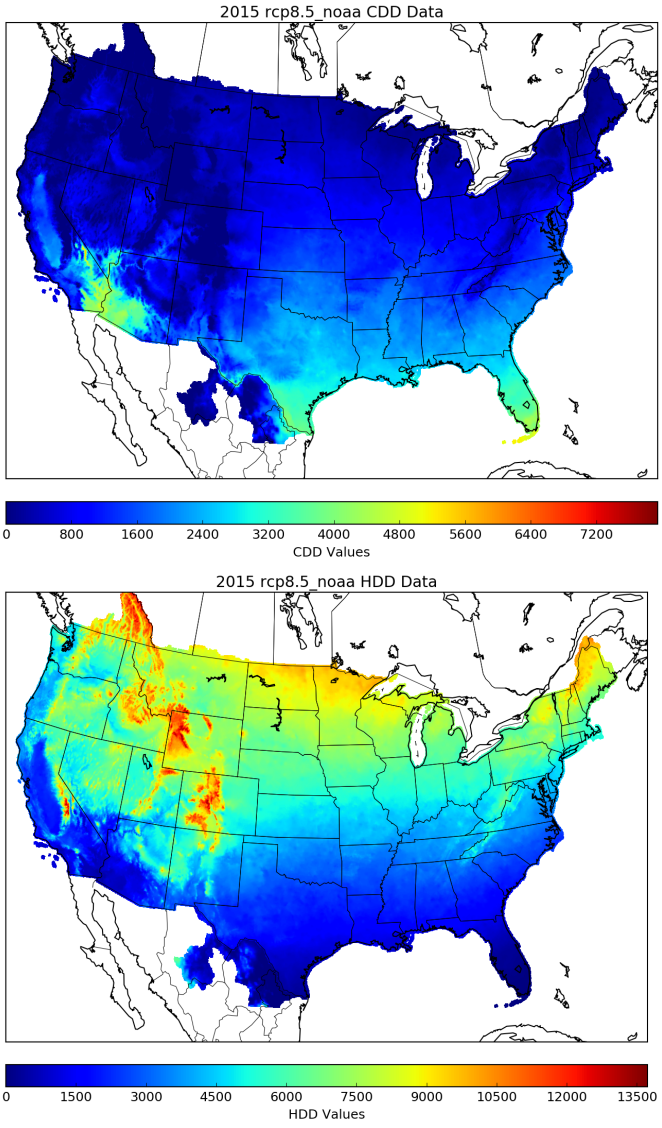

Figure A.7: Cooling and heating degree day rasters for 2015.

Appendix B. Extended Validation

Here we extend our discussion from Section 4.4 on the validity of our models when applied to specific metropolitan areas. Specifically we exam the errors made by XGBoost model, that was trained on all of the CBECS data, and applied to the Augmented Local Law 84 Dataset. Figure B.8 shows a scatter plot of the predicted $\log_{10}$(`MFBTU`) values versus the actual values for all points in the LL84 dataset. From this plot we can see that although the majority of the predicted values are similar to their actual values, there are a handful of points that are badly over- and under-estimated. A common feature among these bad

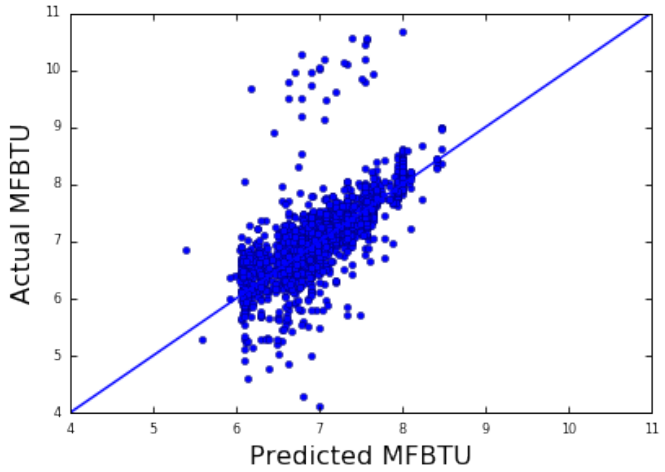

Figure B.8: Error plots comparing the predicted log of `MFBTU` values versus the true log values for the XGBoost model on the LL84 validation dataset.

outlier predictions, is that they all have extreme Energy Star scores. Figure B.9 shows the model's error versus the reported Energy Star score (from the LL84 data), and illustrates that the "mistakes" being made by the XGBoost model can be explained through a building's Energy Star score. An Energy Star score is a value between 1 and 100 that is calculated by the EPA based on the most recent CBECS data, and represents how energy efficient a building is compared to the national average. A score of 50 means that a building is as energy efficient as 50% of other buildings within its category (based on principal building activity) nationally. Similarly, a score of 99 means that a building is more efficient than 99% of other buildings within in its category. Our model, which is trained with the common feature set, makes large errors for buildings that fall on either end of this extreme, over predicting the energy consumption of buildings that are very energy efficient, and under predicting the consumption of buildings that have poor energy efficiency. A building's Energy Star score is calculated using features from the extended feature set, and considering that all of the models we test perform better when using more features, we believe that these features are able to explain the errors observed in the LL84 data. We are unable to test this hypothesis however, as the LL84 data does not include the richer set of features used to calculate the Energy Star scores.

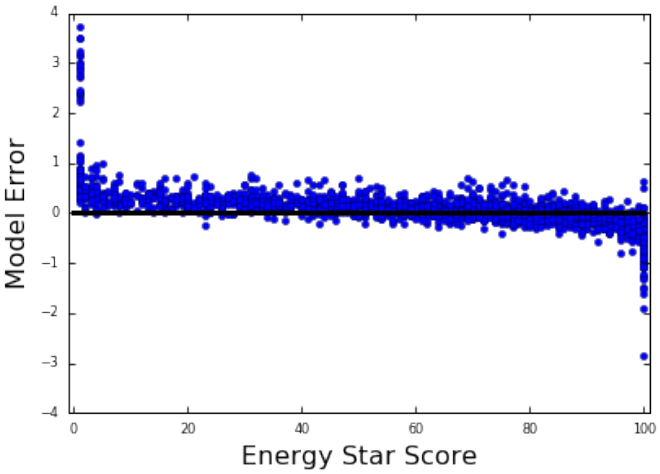

Figure B.9: XGBoost model error versus EnergyStar score.

LL84 PBA	Matched CBECS PBA	CBECS PBA Code
Courthouse	Office	2
Bank Branch	Office	2
Financial Office	Office	2
Medical Office	Office	2
Office	Office	2
Laboratory	Laboratory	4
Self-Storage Facility	Nonrefrigerated warehouse	5
Distribution Center	Nonrefrigerated warehouse	5
Non-Refrigerated Warehouse	Nonrefrigerated warehouse	5
Wholesale Club/Supercenter	Food sales	6
Restaurant	Food sales	6
Supermarket/Grocery Store	Food sales	6
Police Station	Public order and safety	7
Other - Public Services	Public order and safety	7
Urgent Care/Clinic/Other Outpatient	Outpatient health care	8
Outpatient Rehabilitation/Physical Therapy	Outpatient health care	8
Refrigerated Warehouse	Refrigerated warehouse	11
Data Center	Refrigerated warehouse	11
Worship Facility	Religious worship	12
Museum	Public assembly	13
Other - Entertainment/Public Assembly	Public assembly	13
Social/Meeting Hall	Public assembly	13
Fitness Center/Health Club/Gym	Public assembly	13
Senior Care Community	Public assembly	13
Library	Public assembly	13
Movie Theater	Public assembly	13
Lifestyle Center	Public assembly	13
College/University	Education	14
Adult Education	Education	14
Other - Education	Education	14
K-12 School	Education	14
Hospital (General Medical & Surgical)	Inpatient health care	16
Other - Specialty Hospital	Inpatient health care	16
Ambulatory Surgical Center	Inpatient health care	16
Residential Care Facility	Inpatient health care	16
Hotel	Lodging	18
Other - Lodging/Residential	Lodging	18
Residence Hall/Dormitory	Lodging	18
Strip Mall	Strip shopping mall	23
Other - Mall	Enclosed mall	24
Enclosed Mall	Enclosed mall	24
Automobile Dealership	Retail other than mall	25
Retail Store	Retail other than mall	25
Other - Services	Service	26
Repair Services (Vehicle, Shoe, Locksmith, etc.)	Service	26
Personal Services (Health/Beauty, Dry Cleaning, etc.)	Service	26
Performing Arts	Other	91
Other	Other	91
Other - Recreation	Other	91
Other - Utility	Other	91

Table A.8: Mapping between the LL84 'Primary Building Activity' and the CBECS 'Principal Building Activity' fields. We exclude the "Not Available", "Multifamily Housing", "Manufacturing/Industrial Plant", "Parking", and "Mixed Use Property" classes from the LL84 data.