SimBuild
2024

Eleventh National Conference of IBPSA-USA
Denver, Colorado
May 21 – 23, 2024

Machine learning for determining building type

Shovan Chowdhury[2], Fengqi Li[1], Avery Stubbings,[3], Joshua New[1]
Ankur Garg[4], Kevin Bacabac[4], Santiago Correa[4]
[1]Oak Ridge National Laboratory, Oak Ridge, TN
[2]The University of Tennessee, Knoxville, Knoxville, TN
[3]Illinois Institute of Technology, Chicago, IL
[4]BlocPower, Brooklyn, NY

Abstract

Insufficient building information, including footprint, conditioned area, age, and type, hinders urban-scale energy modeling. These parameters are crucial inputs for the simulation and optimization processes integral to the modeling. Prototypical building energy models, based on building surveys and code requirements at the time of construction, are frequently used when audit-quality data is unavailable. This helps to infer internal building characteristics. However, even local data sources like tax assessors' data contain unique land use or parcel codes that can be challenging to map to these prototypical buildings. This information does not directly correlate with the standard building type used to perform energy simulations. In this study, we apply and cross-validate several machine learning algorithms to automate the mapping from general building descriptions to standardized building types, as defined by the U.S. Department of Energy (DOE), a key component to accurately estimate building energy profiles at scale. The XGBoost algorithm outperformed others, achieving an F1 score, precision, and recall of 92.8%, 93.4%, and 93.0%, respectively. These results highlight the potential of advanced machine learning techniques in bridging the data gap for urban-scale energy modeling and suggest a path forward for enhancing the resolution and accuracy of large building energy datasets.

Introduction

Building Type Prediction

In recent years, building information technology has experienced significant advancements facilitated by many innovative technologies that empower the collection and translation of building-related data into valuable datasets. Microsoft and Google have developed frameworks to capture comprehensive building footprint information. These footprints are then processed by advanced machine learning algorithms to convert them into practical building characteristics (Wei, Ji, and Lu 2020). These details are used as inputs to Urban Building Energy Models (UBEMs), which in turn reduces the uncertainty of the estimates. Researchers at Oak Ridge National Laboratory have developed building models for every building in the US (New et al. 2021). Internet of Things has also begun to be incorporated into some of these models (Tang et al. 2019).

The development of many of these building data sets has become the cornerstone for the emergence of the UBEMs, which are dedicated to comprehensively analyzing and evaluating buildings' energy performance across various scales. Within UBEMs, the proper collection and processing of the building's physical characteristics are crucial steps to guarantee the high reliability of the outputs and meet the specific stakeholders' requirements. Many different approaches have been compared to see how they meet the stakeholders' requirements (Sun, Haghighat, and Fung 2020). Two common approaches include the use of the energy plus simulation software (Zhang et al. 2019), as well as the use of Artificial Intelligence (AI) (Himeur et al. 2021) and Machine Learning (ML)(Bourdeau et al. 2019). Since nearly all methodologies rely on quality input data, methods must be developed to collect required feature information. To assess an ad valorem tax on a land parcel, tax assessors must gather and leverage pertinent building information. However, this data varies by county and often exhibits conspicuous gaps, particularly regarding critical building information such as the building's type or function (New et al. 2020). These gaps can be attributed to various factors, including technological limitations and socioeconomic concerns. Consequently, the absence of critical building type information can significantly impact the assessment and analysis of a building's energy performance when utilizing these data sets.

Unfortunately, the definition, description, and data format of a building type will vary across different data sources, and this divergence can lead to inconsistencies when multiple data sources are integrated into a single

workflow for assessing the energy performance of buildings. The ability to harmonize building descriptions and convert them into standardized building types, as defined by the U.S. Department of Energy (DOE) based on the ASHRAE 90.1 Standard (The U.S. DOE-EERE 2021), which encompasses the building's function from an energy perspective, holds significant potential for advancing research within the domain of UBEMs. Alignment with such authoritative sources not only promotes consistency and comparability across various data sets, but also streamlines the process of categorizing and analyzing building energy performance, mainly when dealing with buildings with similar idiosyncrasies. By aligning building types with these standardized classifications, researchers can draw meaningful and consistent insights, facilitating more robust and reliable research insights. Ultimately, these efforts streamline data integration and enhance the broader applicability and effectiveness of energy modeling methodologies, contributing to more resilient and energy-efficient built environments.

In light of these considerations, developing a straightforward methodology for ascertaining the building type based on available building information becomes imperative, especially within building energy modeling. For energy evaluation and analysis, the building type assumes a central role as it significantly influences a building's energy performance. Therefore, there is a pressing need to develop a reliable model for determining the building type based on available information. Accurately characterizing building types can significantly enhance the quality and reliability of energy performance assessments and simulation result analyses.

Researchers have explored many data-driven approaches to predict building energy use and efficiency, including structural design, the Internet of Things, and geospatial data integration. However, there is still a need to categorize buildings effectively. A comprehensive literature review on building type prediction models reveals a growing need for compelling building type predictions. Machine learning models have been developed to capture and predict building energy loads and demand (Zhang et al. 2021) (Wang, Hong, and Piette 2020), in addition to models that focus on energy-efficient designs (Fathi et al. 2020). Models have also been developed for energy prediction of groups of buildings (Xu et al. 2019). Outside of the building energy, machine learning models have also been created to outline structural design (Sun, Burton, and Huang 2021). These various types of building information modeling have garnered substantial attention due to their pivotal role in urban planning, energy efficiency assessments, and real estate analysis.

Using machine learning in building modeling makes building type prediction an ideal candidate for the machine learning approach. The continued growth of UBEM field showcases a growing body of research that employs diverse methodologies and data sources to address the critical issue of building type determination. These models hold significant promise for enhancing the accuracy of urban planning, energy efficiency assessments, and real estate analysis, ultimately contributing to more sustainable and informed decision-making in the built environment.

In this paper, we have developed a machine learning model that can translate and interpret the tax assessor data containing building information and predict building types of selected buildings in selected cities in the United States. The predicted building types will be converted to DOE-referenced building types through a direct mapping technique using a predefined Python function.

Data Preparation and Explanation

The foundation of our predictive model is a comprehensive dataset obtained from tax assessors in New York, which provides a detailed landscape of building attributes within the urban context. The data (Energy and Water Data Disclosure for Local Law 84 2022) is provided by New York City's Mayor's Office of Climate and Environmental Justice (MOCEJ) (NYC Open Data 2022).The original dataset comprises 29.8K buildings in New York City, with each building represented by a row containing 249 columns of property information. For our current model development, we have pre-processed the data as outlined below.

Data Labeling

Creating a reliable and accurately labeled dataset forms the cornerstone of any supervised machine learning endeavor. In our study, we tackled the challenging task of annotating an unlabeled dataset with building types, a process critical for the subsequent predictive modeling.

For the current stage, our approach to data labeling involved a domain-driven classification system design, which drew on a comprehensive understanding of the functional and physical attributes of buildings. We implemented a Python function, which operated on the principles of building use and size to assign building types. This function utilized the 'LargestPropertyUseType' attribute, which denotes the primary use of a property as a primary discriminant. The area of each property was used as a secondary measure to refine the categorization further. For instance, Residential properties such as 'Single Family Home', 'Multifamily Housing',

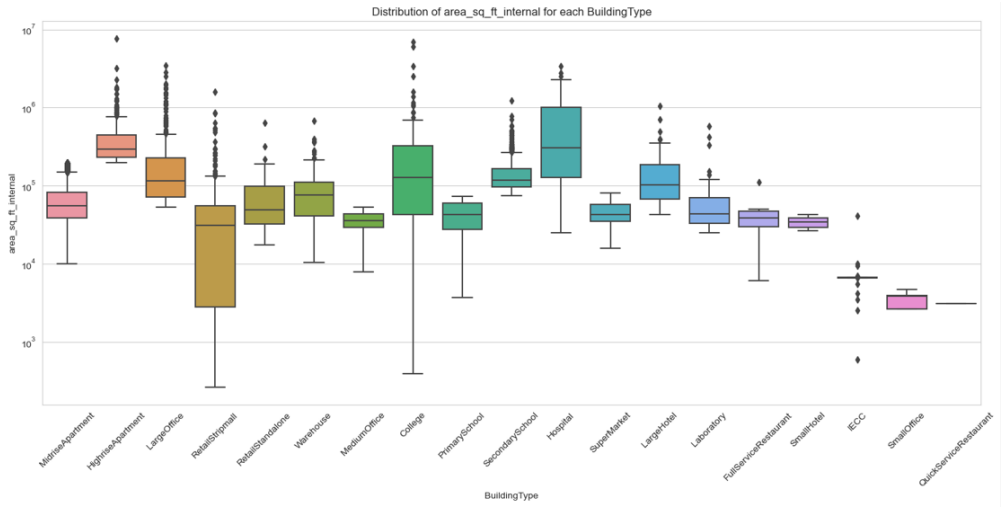

Figure 1: Boxplot of Internal Area by DOE Prototype Building Type providing a visual comparison of internal floor area distributions across different types of buildings.

Table 1: Selected Features for Building Type Classification Model

Feature	**Description**
LargestPropertyUseType	A categorical variable (53 types) that signifies the primary function of the building.
Largest Property Use Type - Gross Floor Area (ft^2)	A quantitative measure of the primary occupancy space.
area_sq_ft_internal	The total internal floor area, providing a scale of the building size.
EUI_elec	The Electricity Use Intensity, an efficiency metric representing electrical energy consumption per square foot.
Number of Buildings	Indicates whether the property is a single building or part of a larger complex (Range from 1 to 161).

'Senior Living Community', and other residential lodgings were classified into 'IECC', 'MidriseApartment', or 'HighriseApartment' based on their floor area, with specific square footage thresholds acting as decision boundaries. In this study, we defined one building type as 'IECC', which originally stands for 'International Energy Conservation Code'. This building type was specifically developed to classify residential buildings within our dataset since the ASHRAE 90.1 Standard had only covered the commercial building types. This "IECC" category represents both single-family and multi-family building types, separating from other residential types such as mid-rise and high-rise apartment building types. The resultant dataset, now structured and enriched with meaningful labels, provided a robust platform for deploying machine learning techniques to predict DOE building types. This manual rule-based labeling process serves as an additional step to address the complexity of building function descriptions. This rule-based labeling process is employed as an extra measure only when the dataset lacks predefined DOE building types. In cases where DOE building types are already defined within the dataset, this additional step of rule-based labeling is not necessary. However, this process allows the framework to perform efficaciously.

While rule-based models offer simplicity and interpretability, their application to the entire dataset is limited by several factors. Rule-based systems can become exceedingly complex and difficult to manage as the number of rules grows to cover the diversity within a large dataset. Our dataset encompasses a wide variety of building types with complex and non-linear relationships between features, which makes the creation of a comprehensive set of rules challenging and prone to human error. So the adoption of machine learning in this context is important. Machine learning algorithms,

particularly ensemble methods like XGBoost, are adept at capturing non-linear interactions and subtle patterns in high-dimensional data that rule-based models might miss. These complex relationships are essential for accurate classification and can significantly impact model performance.

In the future, the labeling process is likely to be replaced by advanced machine learning components. With this improvement, the developed prediction framework will be used to handle more intricate building description systems overall.

Feature Selection and Visualization

Our comprehensive dataset comprises 9,332 individual records, each described by 25 features. In the realm of feature selection, we embarked on a judicious process to identify the predictors most salient for our building type classification model. The original dataset included a diverse range of features: 'bp_building_id', 'NYC Building Identification Number (BIN)', 'Address', 'City', 'building_subtype_internal', 'LargestPropertyUseType', 'Largest Property Use Type - Gross Floor Area (ft^2)', '2nd Largest Property Use Type', '2nd Largest Property Use - Gross Floor Area (ft^2)', '3rd Largest Property Use Type', '3rd Largest Property Use Type - Gross Floor Area (ft^2)', 'area_sq_ft_internal', 'Property GFA - Calculated (Buildings) (ft^2)', 'EUI_elec', 'Site Energy Use (kBtu)', 'total_bldg_annual_consumption_internal', 'year_built_internal', 'Construction Status', 'Number of Buildings', 'Occupancy', 'latitude_internal', 'longitude_internal', 'BuildingType', 'Standard', and 'EUI_per_occupancy'. Out of the original 25 features, 5 were selected based on their relevance and potential to improve model accuracy (See Table 1). These features underwent a rigorous selection process, underpinned by the hypothesis that they hold the most significant information regarding the building type. This hypothesis stems from both statistical evidence and domain expertise, ensuring that each feature plays a pivotal role in the predictive power of our model. To gain deeper insights into the relationship between the building features and the building types, we conducted an exploratory data analysis through various visualization techniques. For instance, figure 1 represents the distribution and variance of area_sq_ft_internal across different building types. These visual representations were instrumental in understanding how each feature contributes to the identification of building categories.

A critical aspect of our dataset that required special attention was the imbalance present in the distribution of the response variable, 'BuildingType' (see figure 2). Class imbalance is a prevalent issue in machine learning, where some classes are over-represented in the dataset while others are under-represented. This imbalance can lead to biased models that favor the majority class, often at the expense of minority class prediction accuracy.

Our dataset exhibited a significant skew in the distribution of building types, with 'MidriseApartment' buildings being the most prevalent with 5,685 instances. In stark contrast, 'Small Office' buildings represented the smallest group, with a mere 5 instances. This discrepancy presents a substantial challenge as it can cause a model to perform well on majority classes while failing to accurately identify minority classes. We excluded 'QuickServiceRestaurant' building type from our model because it is represented by only a single instance in our dataset, making it statistically insignificant for our analysis.

Our dataset features the LargestPropertyUseType as a pivotal categorical variable, which denotes the primary function of a building. This variable comprises a wide array of categories, with 53 distinct types of large building uses represented. To prepare this categorical data for our machine learning algorithms, which necessitate numerical input, we utilized label encoding. Label encoding is a process where each unique category is systematically assigned a numerical value. (Mottini and Acuna-Agost 2016)

Methodology

Overall Workflow

Given that DOE-referenced building prototypes (The U.S. DOE-EERE 2021) encompass a wide range of sub-building types from an energy efficiency standpoint, we propose an approach that involves creating a predictive model to determine building types based on common tax assessor data attributes. Subsequently, we map these determined types to more generalized DOE building categories. This methodology leverages insights obtained from an extensive analysis of tax assessor data across various cities, revealing a common pattern in the way building functions are described.

This strategy involves training the model to assimilate and interpret these property descriptions as informative learning resources, thereby enhancing its comprehension of building stock information. For instance, within tax assessor data, keywords such as "Residence Hall/Dormitory," "Multifamily Housing," and "Other - Lodging/Residential" can be grouped into categories like single/multi-family residential units, mid-rise apartments, or high-rise apartments based on physical attributes like the total floor area of a building. If the

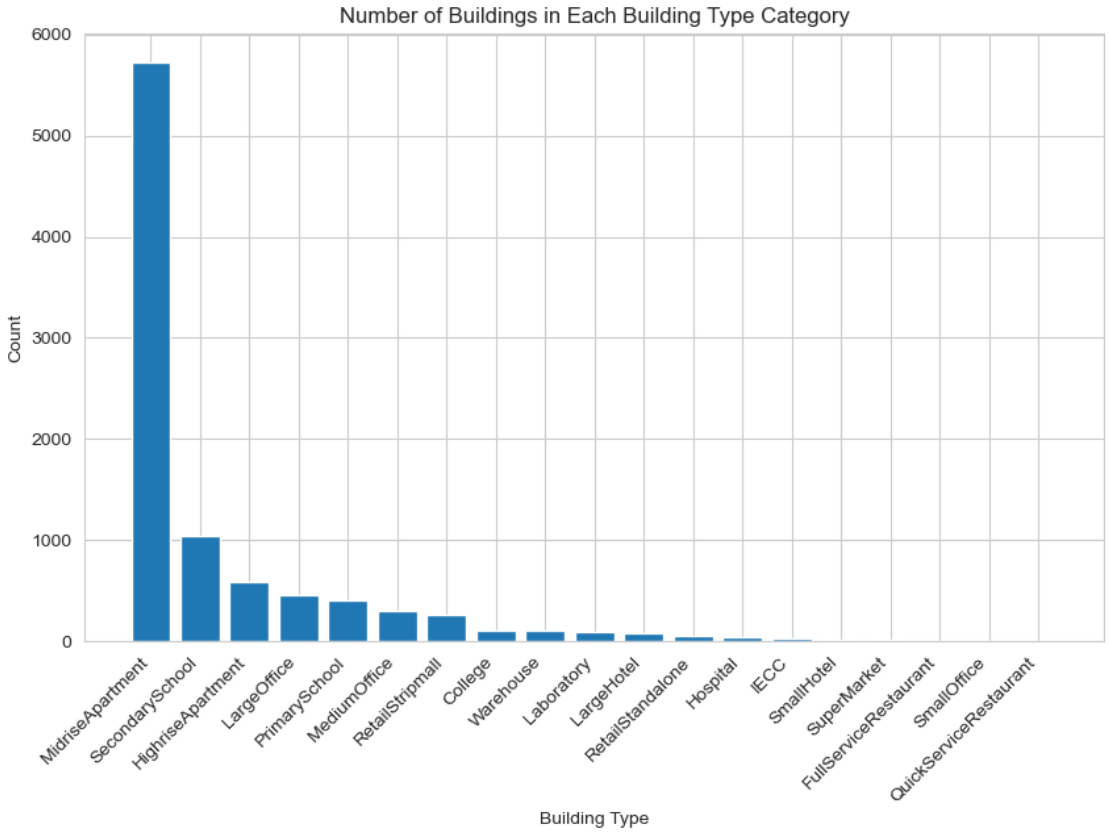

Figure 2: The bar chart illustrates the frequency of each building type within our dataset, highlighting the significant imbalance present. 'Midrise Apartment' buildings dominate the dataset with 5,685 instances, reflecting the commonality of this type in urban areas. In sharp contrast, 'Small Office' buildings are substantially underrepresented with only 5 instances, indicating the rarity of this building type within the scope of our data.

model can discern distinctions based on selected features among buildings and apply the mapping function, it can readily translate these identified building descriptions into DOE building types, such as "Midrise Apartment" or "Highrise Apartment."

This approach is equally applicable to commercial and manufacturing building types. In summary, the critical aspect of classifying a building into a DOE building type involves training a model to continuously learn from tax assessor data's building property information and convert the building descriptions into predefined DOE building categories.

Machine Learning Model Development

Our study implements a suite of machine learning algorithms, each offering distinct mechanisms for pattern recognition and decision-making. We selected four algorithms renowned for their efficacy in classification tasks: Random Forest, Gradient Boosting (via XGBoost), Support Vector Machine (SVM), and Logistic Regression. The choice of algorithms spans ensemble methods, gradient boosting, and linear models to ensure a comprehensive analysis through various statistical learning perspectives. Each algorithm is capable of handling the complexities of our imbalanced multiclass dataset. The detailed algorithms' hyperparameter selections and comparisons are discussed below.

Random Forest is an ensemble learning method based on decision tree classifiers. It operates by constructing a multitude of decision trees during training time and outputting the class that is the mode of the classes of the individual trees (Liaw et al. 2002). This method is robust to overfitting and is capable of capturing complex structures in the data. In our implementation, we first initialized the Random Forest Classifier with a fixed random state to ensure reproducibility. The model was

then trained on the preprocessed training dataset. Post-training, we performed predictions on the test set and evaluated the model using a classification report and a confusion matrix to assess its performance.

XGBoost stands for extreme Gradient Boosting, an advanced implementation of gradient boosting algorithms known for its speed and performance. It builds sequential trees where each tree attempts to correct the errors of the previous one (Ramraj et al. 2016). In our application, XGBoost was configured for multi-class classification, with the objective set to 'multi:softprob' to output predicted probabilities for each building type, addressing our multi-class problem directly. This approach is particularly beneficial in scenarios where the classification involves more than two classes, as in our case with numerous building types. By using these probabilities, the model offers a nuanced perspective on its confidence across the multiple categories, aiding in the interpretability of the predictions.

The Support Vector Machine is a powerful classifier that works by finding the hyperplane that best divides a dataset into classes (Yu and Kim 2012). SVM is effective in high-dimensional spaces and with datasets where the number of dimensions exceeds the number of samples. In this study, an SVM classifier with a linear kernel was utilized, benefiting from its ability to handle high-dimensional data effectively. Given the imbalance in our data, the 'class_weight' parameter was set to 'balanced' to adjust weights inversely proportional to class frequencies, thus emphasizing the minority classes.

Logistic Regression (LaValley 2008) is a linear classifier and thus particularly well-suited to binary classification problems. For our multiclass classification task, we employed Logistic Regression with balanced class weights to address the issue of class imbalance.

Feature Scaling and Hyperparameter Tuning

Before training the SVM and Logistic Regression models, numeric features were standardized to have a mean of zero and a standard deviation of one, ensuring that all features contributed equally to the result without bias from differing scales. This scaling was critical for methods sensitive to feature magnitude, such as SVM and Logistic Regression (Grandvalet and Canu 2002).

For both the Random Forest and XGBoost classifiers, hyperparameter tuning was conducted via grid search to identify the optimal settings that would yield the best cross-validated F1 macro score. This involved iterating over a predefined range of values for parameters such as the number of estimators, maximum tree depth, and minimum samples per leaf.

Evaluation Metrics

The models' performance was assessed using a classification report, which provided a detailed view of the precision, recall, and F1 scores for each class. Detailed description of these performance metrics can be found at (Chowdhury and Schoen 2020). Additionally, a confusion matrix was generated for each model, giving insight into the types of errors made and areas where the model may require further refinement.

Results

The performance of the four machine learning models was evaluated using 10-fold cross-validation to ensure the robustness and reliability of our results. Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used to compare and select a model for a given predictive modeling problem because it ensures that every observation from the original dataset has the chance of appearing in the training and test set. This is particularly valuable when the dataset is imbalanced, as it provides a more comprehensive assessment of the model's performance.

The precision, recall, and F1 scores—critical metrics for classification tasks, especially in imbalanced datasets—were calculated for each model. Precision measures the proportion of true positives among all positive predictions, recall (also known as sensitivity) measures the proportion of true positives identified among all actual positives, and the F1 score is the harmonic mean of precision and recall, providing a single score that balances both concerns.

The results are summarized in the table 2, which shows the cross-validated performance metrics for each algorithm

Table 2: 10-Fold Cross-Validated Model Performance Metrics

Algorithm	Precision	Recall	F1 Score
Random Forest	0.83	0.82	0.82
XGBoost	0.93	0.93	0.92
SVM	0.57	0.72	0.60
Logistic Regression	0.32	0.43	0.31

The XGBoost model exhibited superior performance across all metrics, achieving nearly 0.93 in both pre-

cision and recall, which resulted in an F1 score of 0.92. These results underscore XGBoost's effectiveness in handling multi-class classification problems, even in the presence of class imbalance.

Conversely, the SVM and Logistic Regression models showed lower precision and F1 scores. The SVM model's relatively higher recall indicates its ability to identify most of the positive instances but at the expense of a larger number of false positives, as evidenced by its lower precision. Logistic Regression, while generally robust and effective for binary classification, did not perform as well in this multi-class, imbalanced context, leading to the lowest scores across all metrics.

In addition to the cross-validated performance metrics, we examined the convergence of the XGBoost model by plotting the log loss (see figure 3). The log loss plot demonstrates how the model's performance improved as the number of boosting iterations increased. The declining trend of the log loss value indicates the model's increasing accuracy in predicting the correct classes over iterations. A sharp decrease in log loss early in the training process signifies rapid learning, while a plateau suggests that subsequent iterations provide marginal gains, which can inform decisions on the appropriate number of boosting rounds to prevent overfitting.

The performance of the XGBoost model was further analyzed to understand its precision, recall, and F1 score on a per-class basis. This detailed breakdown is crucial for multi-class classification problems, especially when dealing with imbalanced datasets. It provides insights into how well the model can identify each specific class, which is essential for practical applications where certain classes may be more critical than others.

The class-wise performance metrics, presented in Table 3, show that the model achieved high precision and recall for most of the building types, with corresponding F1 scores that indicate a well-balanced prediction capability. Notably, 'HighriseApartment', 'LargeHotel', 'LargeOffice', 'MidriseApartment', and 'SecondarySchool' classes have F1 scores above 0.99, reflecting the model's excellent ability to classify these building types accurately. These results are particularly impressive given the challenges posed by the imbalanced nature of the dataset.

On the other hand, the 'FullServiceRestaurant' class has the lowest F1 score at 0.43, which suggests that this particular class is more challenging for the model to predict accurately. This could be due to a smaller representation in the dataset or higher variability within the features of this class. In general, the model demonstrates exceptional performance across most classes, indicating its robustness and reliability as a predictive tool for building type classification.

Feature importance is a technique used to identify which features have the most influence on the predictive power of a model. In the XGBoost algorithm, feature importance is often represented by the F score, also known as the "feature score," which quantifies the number of times a feature is used to split the data across all decision trees within the model. A higher F score indicates a greater impact on the model's decisions, reflecting the relative importance of each feature in the classification process.

Table 3: Class-wise 10 fold cross-validated Performance Metrics of the XGBoost Model

Building Type	**Precision**	**Recall**	**F1**
College	0.99	0.94	0.96
HighriseApartment	0.99	0.99	0.99
Hospital	0.92	0.90	0.90
IECC	0.88	0.88	0.88
Laboratory	0.97	1.00	0.98
LargeHotel	1.00	1.00	1.00
LargeOffice	0.99	0.99	0.99
MediumOffice	0.96	0.96	0.96
MidriseApartment	1.00	1.00	1.00
PrimarySchool	0.98	1.00	0.99
RetailStandalone	0.94	0.98	0.96
RetailStripmall	0.97	0.91	0.94
SecondarySchool	1.00	0.99	1.00
SmallHotel	0.97	0.95	0.95
SmallOffice	0.80	0.80	0.80
SuperMarket	0.90	1.00	0.93
Warehouse	0.99	0.99	0.99
FullServiceRestaurant	0.43	0.43	0.43

The feature importance graph for our XGBoost model, as illustrated in Figure 4, reveals that 'area_sq_ft_internal' has the highest F score, indicating it is the most frequently used feature in tree splits and thus the most significant predictor of building type. This feature's F score is notably higher, close to 20, suggesting its pivotal role in the model's classification decisions. The second most important feature, 'LargestPropertyUseType', has an F score close to 14, which, while substantially less than 'area_sq_ft_internal', still signifies a considerable influence on the model's outcomes.

The other features in the analysis exhibit F scores lower than 1, suggesting they contribute less to the model's predictive capability. This disparity in F scores highlights the varying degrees of relevance each feature has in determining building types. The high F score of

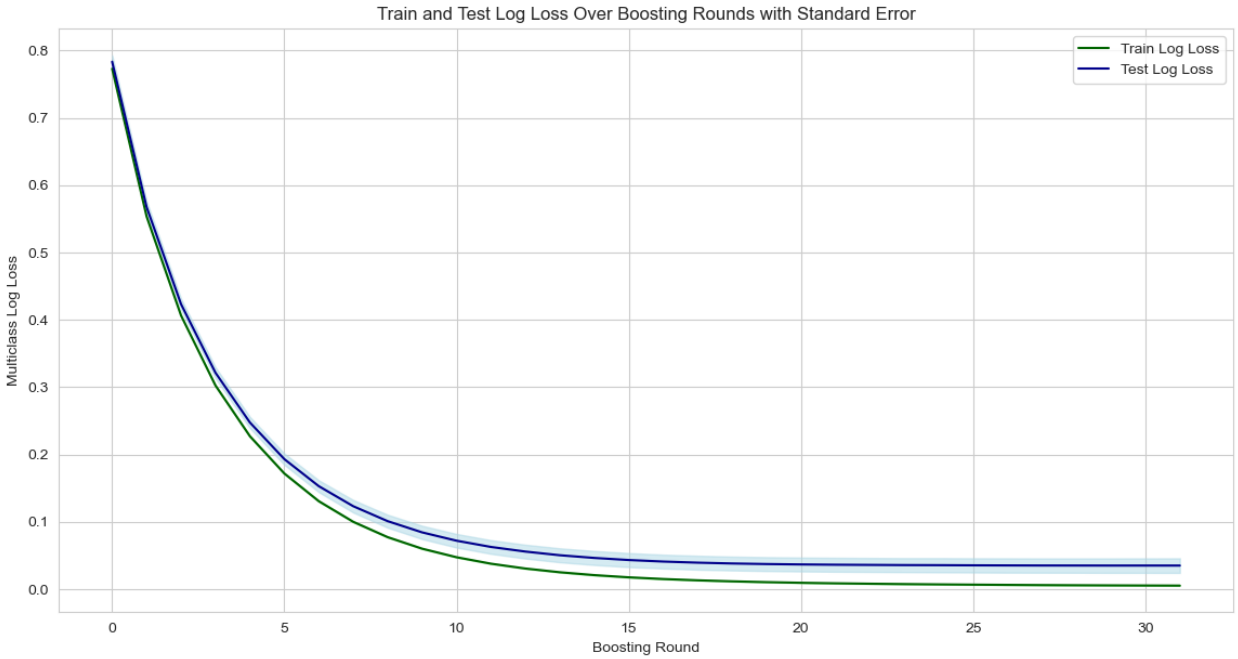

Figure 3: Log loss of the XGBoost model over boosting iterations. The plot shows a decreasing trend, indicating an improvement in the model's predictive accuracy with additional iterations

'area_sq_ft_internal' underscores the critical role that the size of the property plays in distinguishing between different types of buildings.

Understanding why certain features are more important than others can inform domain-specific strategies for data collection and feature engineering. In the context of building classification, the results suggest that a building's internal area and primary use type are critical factors to consider, while other characteristics may have a more marginal impact.

Discussion

While the results are encouraging, we acknowledge certain potential and limitations. The existing methodology involves the integration of specific features sourced from New York City's MOCEJ dataset (NYC Open Data 2022). However, there exists untapped potential to refine the prediction model further by incorporating additional detailed building features found in the original dataset. This expansion aims to comprehensively evaluate and potentially improve the accuracy of the model. Should noticeable enhancements in accuracy occur, the initially selected features could serve as baseline parameters for adoption by other cities. This approach bears considerable significance, particularly if the model is intended for predicting the DOE's referenced building types across a spectrum of cities in the United States.

In the course of scrutinizing tax assessor data from diverse urban areas, it becomes evident that the model's accuracy is contingent upon the quality and comprehensiveness of the underlying tax assessor data. Any inaccuracies inherent in the source data have the potential to permeate the predictive outcomes. Notably, variations in the formats employed for describing primary building types of buildings across different cities were observed. To mitigate this concern, further steps involve the implementation of a standardized description format, coupled with the establishment of a dynamic database capable of assimilating emerging building description categories. This approach is poised to be a noteworthy undertaking, with the potential to significantly ameliorate the overall quality of building data, particularly tax assessor data.

Moreover, the study's scope was currently limited to New York City, and further research is needed to validate the approach across different urban contexts with varying building typologies and data quality. Future work could focus on integrating additional datasets, such as remote sensing data or newly developed data characteristics, to refine the predictions. Additionally, exploring the impact of temporal changes in building usage and renova-

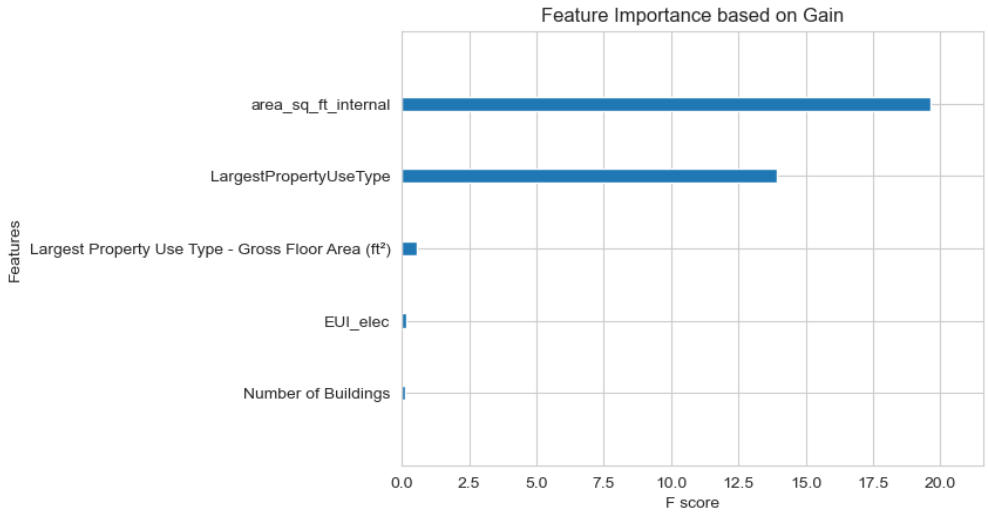

Figure 4: Relative Importance of Predictive Features in Building Type Classification Using XGBoost

tions on the model's performance could yield insights for maintaining the accuracy of building type classifications over time.

Conclusion

In conclusion, this paper has provided a method using machine learning to address the gap of building information to building type for assessing internal details and building-specific energy use scalably at urban- to nation-wide geographical areas. This study underscores the importance of connecting tax assessor data to DOE building types for enhanced digital twins and future predictions. Through extensive training and validation processes, key algorithms and metaparameters were selected to optimize performance according to cross-validated evaluation metrics. Finally, this research has consolidated insights into a comprehensive and scalable machine learning approach for building type prediction. The authors hope variants of this method could have significant implications for urban planning, urban building energy modeling, resource allocation, and simulation-informed decision-making towards a more sustainable built environment.

Acknowledgment

Notice of Copyright. This work was funded by field work proposal CEBT105 under US Department of Energy Building Technology Office Activity Number BT0305000, as well as Office of Electricity Activity Number TE1103000. The authors would like to thank Amir Roth for his support and review of this project. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725. This manuscript has been authored by UT-Battelle, LLC, under Contract Number DE-AC05-00OR22725 with the U.S. Department of Energy (DOE). The U.S. government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. government retains, a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

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* 48:101533.

Chowdhury, Shovan, and Marco P. Schoen. 2020. "Research Paper Classification using Supervised Machine Learning Techniques." *2020 Intermountain Engineering, Technology and Computing (IETC)*. 1–6.

Fathi, Soheil, Ravi Srinivasan, Andriel Fenner, and Sahand Fathi. 2020. "Machine learning applications in urban building energy performance forecasting: A systematic review." *Renewable and Sustainable Energy Reviews* 133:110287.

Grandvalet, Yves, and Stéphane Canu. 2002. "Adaptive scaling for feature selection in SVMs." *Advances in neural information processing systems*, vol. 15.

Himeur, Yassine, Khalida Ghanem, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. 2021. "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives." *Applied Energy* 287:116601.

LaValley, Michael P. 2008. "Logistic regression." *Circulation* 117 (18): 2395–2399.

Liaw, Andy, Matthew Wiener, et al. 2002. "Classification and regression by randomForest." *R news* 2 (3): 18–22.

Mottini, Alejandro, and Rodrigo Acuna-Agost. 2016. "Relative label encoding for the prediction of airline passenger nationality." *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 671–676.

New, Joshua, Mark Adams, Anne Berres, Brett Bass, and Nicholas Clinton. 2021. "Model America – data and models of every U.S. building." 4.

New, JR, MB Adams, E Garrison, Brett Bass, and Tianjing Guo. 2020. "Scaling beyond tax assessor data." *Proceedings of the ASHRAE/IBPSAUSA 2020 Building Performance Analysis Conference & SimBuild (BPACS), Chicago, IL, USA*, Volume 29.

NYC Open Data. 2022. Energy and Water Data Disclosure for Local Law 84 2022 (Data for Calendar Year 2021). https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/7x5e-2fxh.

Ramraj, Santhanam, Nishant Uzir, R Sunil, and Shatadeep Banerjee. 2016. "Experimenting XGBoost algorithm for prediction and classification of different datasets." *International Journal of Control Theory and Applications* 9 (40): 651–662.

Sun, Han, Henry V. Burton, and Honglan Huang. 2021. "Machine learning applications for building structural design and performance assessment: State-of-the-art review." *Journal of Building Engineering* 33:101816.

Sun, Ying, Fariborz Haghighat, and Benjamin C.M. Fung. 2020. "A review of the-state-of-the-art in data-driven approaches for building energy prediction." *Energy and Buildings* 221:110022.

Tang, Shu, Dennis R. Shelden, Charles M. Eastman, Pardis Pishdad-Bozorgi, and Xinghua Gao. 2019. "A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends." *Automation in Construction* 101:127–139.

The U.S. DOE-EERE, The U.S. Department of Energy (DOE) Office of Energy Efficiency Renewable Energy. 2021. Prototype Building Models.

Wang, Zhe, Tianzhen Hong, and Mary Ann Piette. 2020. "Building thermal load prediction through shallow machine learning and deep learning." *Applied Energy* 263:114683.

Wei, Shiqing, Shunping Ji, and Meng Lu. 2020. "Toward Automatic Building Footprint Delineation From Aerial Images Using CNN and Regularization." *IEEE Transactions on Geoscience and Remote Sensing* 58 (3): 2178–2189.

Xu, Xiaodong, Wei Wang, Tianzhen Hong, and Jiayu Chen. 2019. "Incorporating machine learning with building network analysis to predict multi-building energy use." *Energy and Buildings* 186:80–97.

Yu, Hwanjo, and Sungchul Kim. 2012. "SVM Tutorial-Classification, Regression and Ranking." *Handbook of Natural computing* 1:479–506.

Zhang, Liang, Jin Wen, Yanfei Li, Jianli Chen, Yunyang Ye, Yangyang Fu, and William Livingood. 2021. "A review of machine learning in building load prediction." *Applied Energy* 285:116452.

Zhang, Zhiang, Adrian Chong, Yuqi Pan, Chenlu Zhang, and Khee Poh Lam. 2019. "Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning." *Energy and Buildings* 199:472–490.