Predictive Modeling of Energy Consumption for Buildings Based on Their Characteristics and Weather Patterns

Son Minh Duc Nguyen

Received September 20, 2023
Accepted March 14, 2024
Electronic access March 31, 2024

Energy consumption in a building can be affected by various factors such as its design or the surrounding environment, and thorough understanding of the relationship between such features and the energy consumption pattern of the building can lead to more efficient energy management. This paper presents a model to estimate the energy consumption in buildings from a series of features including building metadata such as area, number of floors, primary use and weather data such as air temperature, dew temperature, wind speed. The hypothesis of this project is that there is a predictive relationship between mentioned features in the amount of energy consumed. The LightGBM model is able to achieve an average root mean square error (RMSE) of 2.63 on both the training set and test set. As the RMSE score is expressed in the same unit as the target variable 'meter_reading' which is kilowatt-hour (kWh), this means the model can predict energy consumption with an error of less than 1 kWh, providing reliable evidence for decision-making in smart energy management. The results of this study have implications for the building industry by providing further evidence that there is a direct, predictable relationship between building features, weather patterns and the energy consumption. This study is another step towards developing more accurate and reliable models for estimating the energy consumption of buildings.

Introduction

Energy consumption in buildings is a highly important issue due to its significant impact on the environment and its economic implications. As sustainability and the urgency to address climate change gain traction, there is an increasing need to optimize energy usage and improve efficiency in buildings. Predictive modeling using machine learning has emerged as a promising approach to tackle this challenge. By using data-driven algorithms, machine learning provides valuable insights into energy consumption patterns, empowering stakeholders to make informed decisions and implement effective strategies. These models analyze historical data, real-time sensor readings, and contextual information to predict energy usage, identify inefficiencies, and propose targeted interventions. According to Allouhi et al., 2015[1], architectural design and weather are the main underlying factors impacting energy consumption in residential buildings. Therefore, the model in this paper will utilize a dataset containing building metadata and weather data to predict the energy consumption in buildings of other uses as well. Intuitively, the temperature changes within a day-time and within a year-time or across seasons are expected to impact consumers' behaviors, leading to a potentially predictable relationship between the two most apparent factors - temperature and time. Integrating machine learning with building automation systems and smart technologies enables dynamic optimization, actively managing energy consumption in response to changing conditions and user behavior. The potential benefits are diverse, including reduced carbon emissions, lower utility costs, improved occupant comfort, and enhanced building performance.

Literature Review

Several research papers have been published in this domain, exploring different approaches and methodologies. The following are the papers that would provide valuable insights into predictive modeling of energy consumption for buildings using machine learning. A paper titled "Machine learning for energy consumption prediction and scheduling in smart buildings" by Bourhnane et al., 2020[2] investigates the impact of PV installation and electrical appliances on energy consumption in Smart Buildings which is crucial to the functioning of Energy-efficient Management Systems (EMS). The EMS serves as a foundation for Smart Grid technology, in which energy consumption needs to be predicted accurately in real time to cope with the variance between energy demand and cost in order to maintain its benefits such as improving security and reducing peak demand. The study uses Artificial Neural Networks and Genetic Algorithms to train its model and tests it using a real-word Smart Building testbed. The model eventually yields modest accuracy due to the small size of the dataset. This research aims to determine the energy output of specific electrical appliances using actual data from appliances such as fridges, air conditioners, or microwaves. Our study instead will focus on the total energy usage

of a building using its characteristics such as number of floors, or area and weather factors such as temperature rather than data obtained from electrical appliances, with a much larger dataset consisting of over 20 million rows for each feature compared to 126 rows in the study above.

Another study by Shapi et al., 2021[3] investigates the characteristics of energy consumption of buildings by investigating real-world data from two tenants in a commercial building in Malaysia. By using various algorithms such as Support Vector Machine, Artificial Neural Network, and k-Nearest Neighbor, the research aims to find which method is the most accurate in predicting energy consumption, one of the Building Energy Management Systems (BEMS) important applications to help improve energy efficiency and yield economical savings. The research's findings indicate that each tenant's energy usage has different distribution characteristics. Therefore, this study mainly focuses on the impact of residential behavior on energy consumption and the prediction of energy usage of separate households. Instead, our research aims to predict the energy usage of a whole building considering building characteristics such as area, primary use and external factors, including temperature and humidity.

The paper titled "Multiple Electric Energy Consumption Forecasting Using a Cluster-Based Strategy for Transfer Learning in Smart Building" by Son et al., 2020[4] investigates the impact of various profiles–such as a time series of a whole building or an individual household in a smart building–on the energy consumption of smart buildings rather than the impact of each specific profile which can be predicted using existing approaches. Therefore, this research aims to develop a robust framework to predict the impact of many profiles by clustering them in the training set. Experimental results on two smart buildings in South Korea demonstrate the framework's capability to achieve superior forecasting performance with reduced computational time, making it suitable for intelligent energy management in smart buildings. The energy demand of a profile is also the combination of the total energy usage of specific equipments such as computers, lighting, or air conditioner at a specific time, which makes this study more like an extension to the study by Bourhnane et al., 2020[2] with an alternative approach. Our study, instead of utilizing actual energy usage of appliances used within a building to identify patterns, explores the relationship of a building's overall characteristics such as primary use and external temperature factors to identify any direct impacts if relevant.

These papers provide a comprehensive insight into the effort of predicting energy consumption of buildings. Yet, the intricate interplay of building characteristics - such as when a building was built, number of floors, its area and primary use - and weather patterns remain unexplored in these researches. Therefore, our study is determined to shed light on this multifaceted relationship, ultimately contributing to a more holistic understanding of factors affecting energy consumption prediction in buildings.

Data

Dataset

The data used for this project is the dataset from a Kaggle competition by ASHRAE called Great Energy Predictor III. This dataset includes three years of hourly meter readings from over a thousand buildings at different sites worldwide. It contains five csv files: a weather data file, an energy reading file, their corresponding test files, a building metadata file. The weather file contains data obtained from the closest meteorological station to the site such as air temperature, dew temperature, wind speed. The energy reading file contains readings from four meters: electricity, chilled water, steam and hot water. Not all buildings contain all types of meter as mentioned. These readings represent the actual energy consumption patterns over an extended period, enabling the development of a robust predictive model. The building metadata file consists of several features of a building including primary use, area, number of floors, and the year in which a building was built.

Exploratory Data Analysis

In this section, we conduct preliminary inspection of the ASHRAE dataset of energy readings and other features prior to training the dataset. Before analyzing the data, the separate csv files must be merged together to create a complete file for analysis and later training. The energy reading file is sequentially merged with the building metadata file through the building ID column, and then they are both joined with the weather data file on the site ID and timestamp columns. Before the EDA is carried out, we apply log transformation to the target variable 'meter_reading' to improve its skewness, as seen in Figure 1.

From Figure 2, we can see that energy consumption experiences a decline during the early hours of the day, followed by a noticeable increase after 5am. This early-morning drop and subsequent rise can be attributed to changing human activity levels, where energy demand reduces during the overnight hours due to decreased occupancy and utilization of appliances. As the day progresses, energy usage sharply ascends, reaching a peak during mid-day and gradually declines until the end of the day. This midday peak aligns with the period of heightened activity and occupancy in buildings. Moreover, the observed daily patterns of energy consumption align with expected temperature variations, suggesting a correlation between temperature and energy usage. In a year, energy consumption remains minimal during the period from January to March, coinciding with milder climates and reduced heating needs. The increase from April to August corresponds to warming weather and heightened cool-

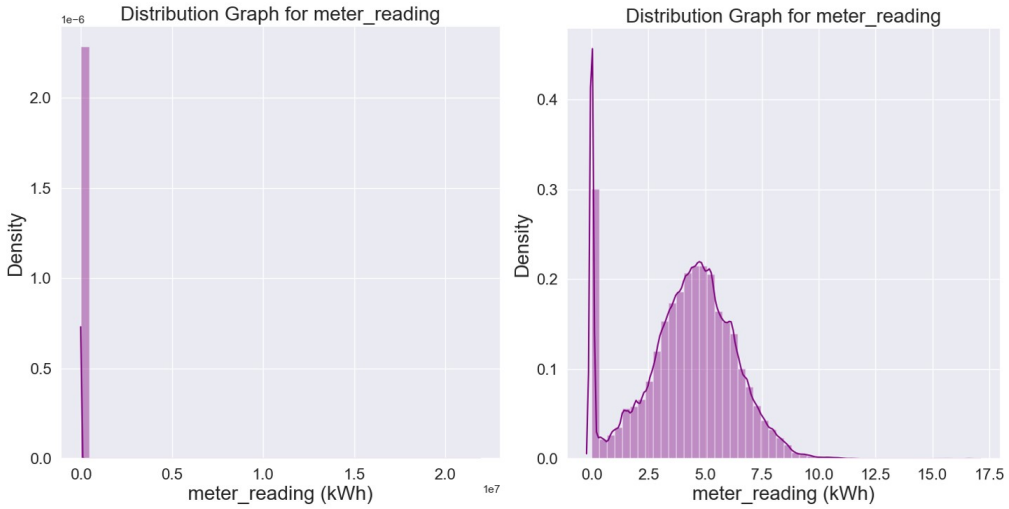

Fig. 1 'meter_reading' distribution before (left) and after (right) log transformation is applied.

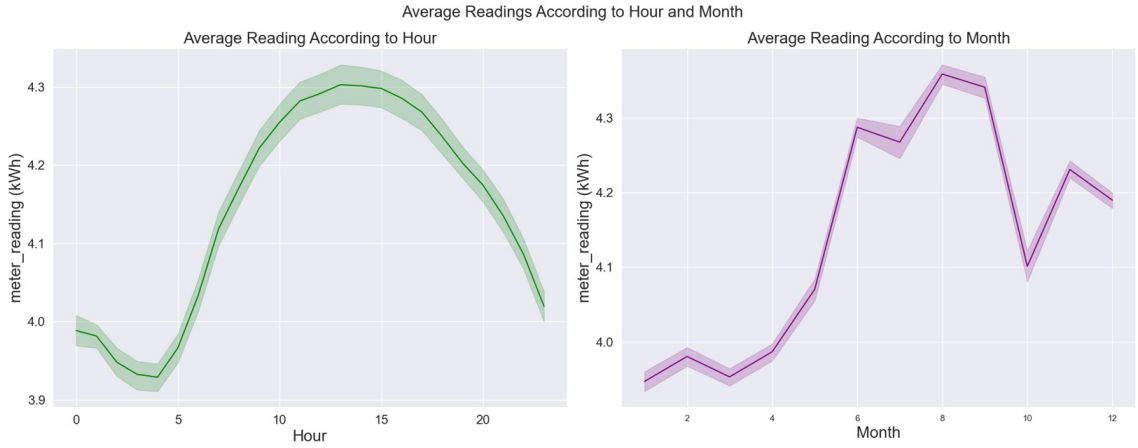

Fig. 2 Average energy reading in hour (left) and month (right).

ing requirements, again reinforcing the predicted correlation between temperature and energy consumption in buildings.

Figure 3 compares the significance of different meter types in shaping energy consumption and efficiency within buildings. It can be seen that electricity is the most popular meter used in buildings, highlighting its pervasive use for various purposes. Intriguingly, despite being the second least popular meter, steam meter exhibits the highest average energy consumption, possibly due to their energy-intensive applications. Hot water is the least popular meter in the dataset but the most energy-efficient meter type.

Figure 4 compares the average energy usage across multiple sites around the world while displaying the composition of various primary uses of the buildings at each site. In this dataset, each site represents a city in order from 0 to 16: Orlando(US), Heathrow(UK), Tempe(US), Washington(US), Berkeley(US), Southampton(UK), Washington(US), Ottawa(Canada), Orlando(US), Austin(US), Saltlake(US), Ottawa(Canada), Dublin(Ireland), Minneapolis(US), Philadelphia(US), Rochester(US). Site 13 exhibits the highest average reading and contains buildings with the most diverse uses while Site 11 experiences the lowest energy usage and consists of buildings of only one primary use which is education. As seen in Figure 4 and 5, education is the most prominent category of primary use as it exists in all sites of the dataset. There is a disparity between the popularity of different primary uses, with education, entertainment/public assembly, residential, office, and public service being much more prevalent compared to

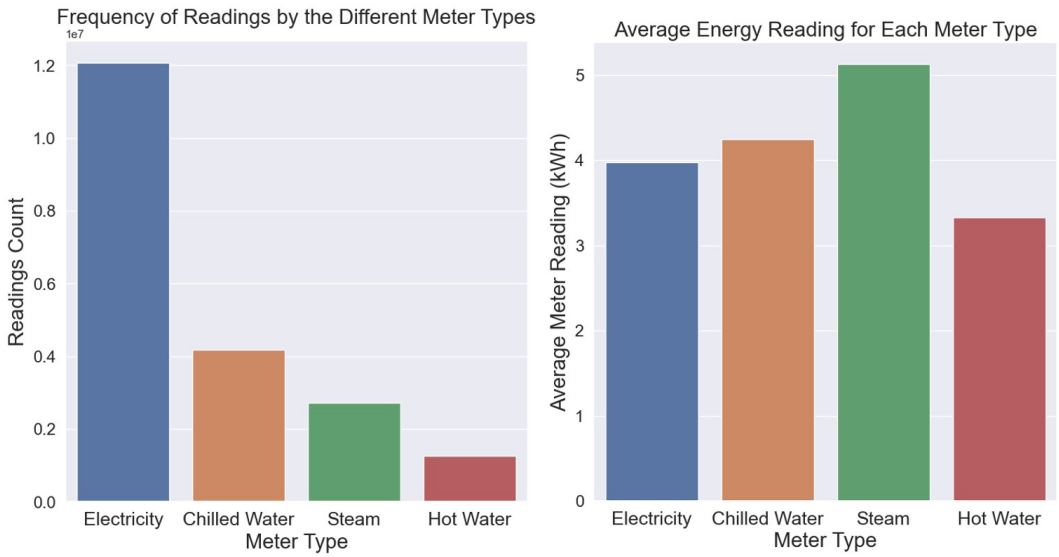

Fig. 3 Frequency of different meters used in buildings (left) and Average reading for each meter (right)

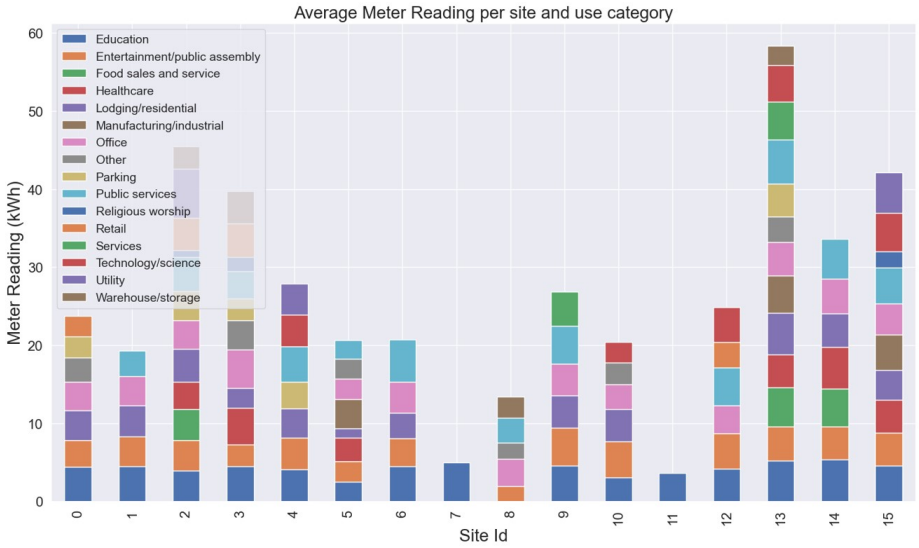

Fig. 4 Average meter_reading per site with the composition of primary uses of buildings at each site displayed within a bar.

other categories.

Examining the correlation plot depicted in Figure 6 reveals a distinct pattern: the lack of pronounced correlations between the various features and the target variable, ”meter_reading.” The highest correlation coefficient observed is a modest 0.36, associated with the ”square_feet” feature. This suggests that relationships between these features and energy consumption are relatively weak. While the ”square_feet” correlation hints at a potential connection between building size and energy usage, it falls short of indicating a robust association, implying that other influential factors are at play.

Curiously, the weather-related features, which are naturally anticipated to significantly impact energy consumption, exhibit correlations below 0.1. In fact, some of these features display correlations as lower than 0.01. This unexpected lack of strong correlations with weather metrics suggests that additional variables and complexities beyond those explicitly considered are contributing significantly to the observed energy consumption patterns.

It is imperative to recognize that the heatmap visualization in

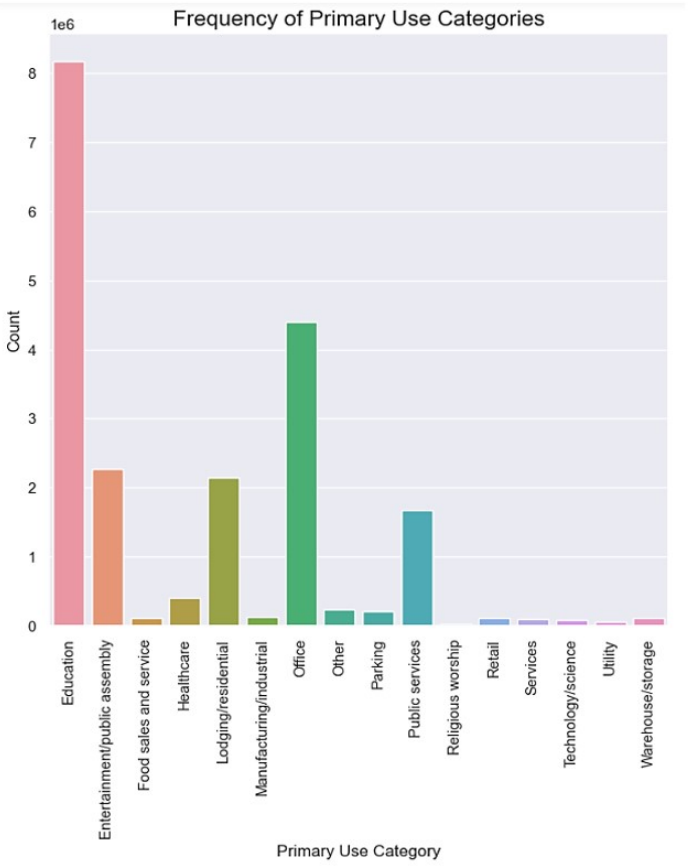

Fig. 5 Frequency of Different Primary Uses of Buildings, displayed on the horizontal axis.

Figure 6 excludes categorical data, specifically the categorical variables associated with primary building use and meter types. These variables can provide invaluable insights into energy usage tendencies and unveil intricate relationships between the intended use of a building and its corresponding energy consumption behaviors. Data encoding will be carried out at a later part, in the preprocessing section to ensure the presence of these variables in training.

Preprocessing

The first step of processing the data is already mentioned in the previous section, where log transformation is applied to the target variable 'meter_reading. Another feature that needs to be logarithmically transformed is 'square_feet' to improve its skewness as seen in Figure 7. Log transformation is used to normalize the distribution of the data, as for energy reading, there are too many zero values, and for square footage, a few buildings may have very high values while the majority have low values. This also helps to increase the interpretability of the relationships between the two features with the target variable when training the model as it calculates the percentage changes between them rather than absolute changes.

Next, we break down the 'timestamp' column into multiple time categories including hour, weekday, and month to understand trends in energy consumption over a specific time period. Then, we decided to drop the 'sea_level_pressure' column for a more interpretable and straightforward model as it is hard to intuitively draw a substantial connection between sea level pressure and building energy consumption.

In this study, for missing values imputation, we leverage the capabilities of the LightGBM machine learning framework, which exhibits robustness in handling missing values within datasets. A prominent example within our dataset is the 'year_built' column that contains approximately 80% of missing values. Unlike some traditional machine learning algorithms that require imputation of missing values prior to training, LightGBM is designed to naturally handle such data discrepancies during its learning process. This advantageous feature is attributed to the algorithm's histogram-based approach to splitting data and constructing decision trees. Given that, LightGBM can effectively partition data points even when they contain

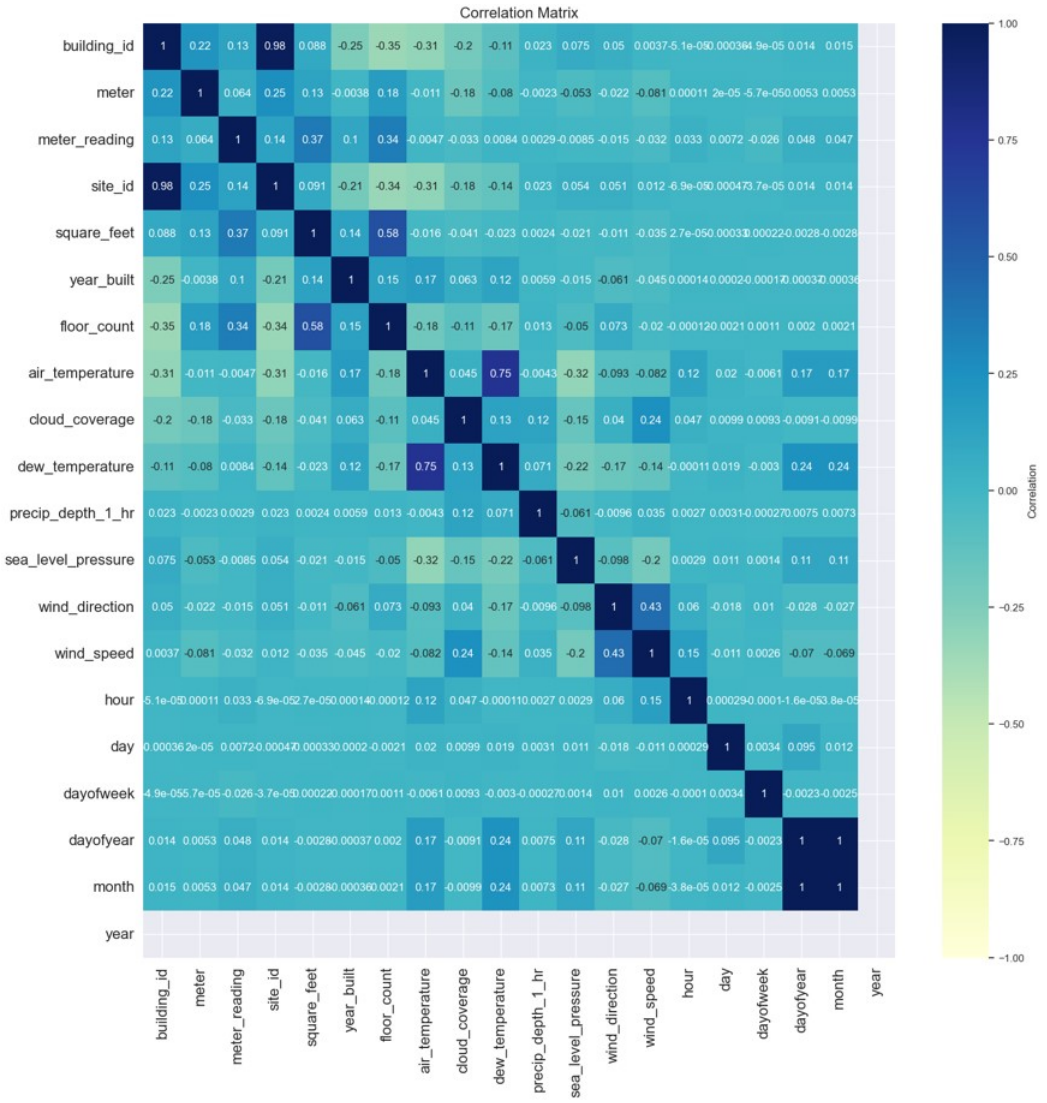

Fig. 6 Correlation heatmap among dataset features

missing values, thereby circumventing the need for explicit imputation techniques that could potentially introduce inaccuracies or distortions[5].

This characteristic not only streamlines the preprocessing phase but also contributes significantly to preserving the integrity of the dataset. However, this potentially leads to reduced feature importance for features with high missingness such as 'year_built', which introduces difficulty in assessing and comparing the impact of each feature later on[5]. Given that, in cases where a substantial proportion of missing values is present, imputation methods might inadvertently introduce bias, skewing the outcomes of subsequent analysis. By employing LightGBM's intrinsic ability to adapt to incomplete data, our study ensures accurate and efficient model training without necessitating additional imputation strategies. This approach is particularly pertinent when dealing with data columns with high percentages of missing values, such as the aforementioned case where 80% of data is absent.

Finally, we convert the 'wind_speed' column using the boundaries of the beaufort scale to assign a beaufort score value to each data point in the column. A beaufort scale value 0 to 12 illustrates the intensity of wind based on its speed in order to make the data in this column more comprehensible. We also reformat the 'wind_direction' column to be a categorical feature representing a 16-wind compass rose. All the categorical variables are then encoded using 'LabelEncoder' from scikit-learn, which converts categorical features into numerical labels. The dataset is prepared for training.

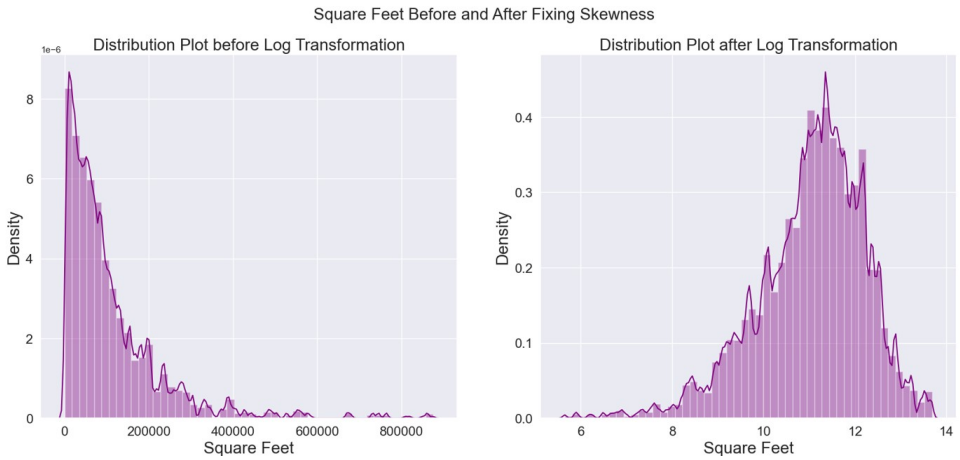

Fig. 7 'Square_feet' distribution before (left) and after (right) log transformation is applied.

Feature	R-squared value with meter_reading
square_feet	5.91e-04
year_built	1.25e-02
floor_count	1.69e-02
air_temperature	1.71e-05
cloud_coverage	3.63e-05
dew_temperature	1.11e-05
precip_depth_1_hr	2.59e-07
wind_direction	6.09e-09
wind_speed	3.69e-05

Table 1 R-squared value of the linear regression between different features and target variable

Methods

This section presents the machine learning methods we used in order to develop a finalized model to accurately predict energy consumption in buildings. Initially, we expected that there would be a straightforward connection between the features in the dataset and the energy reading, especially for weather-related features. Therefore, we tried to test out the performance of a simple linear regression model for each feature in relationship with the target variable 'meter_reading'. As seen in Table 1, we can see that the R-squared value for the relationships between chosen features and the target variable is extremely low, suggesting that regression models that are capable of predicting the energy use pattern are much more complex than just linear models.

To tackle the complexity inherent in the dataset, we employed a decision tree gradient boosting framework called LightGBM, which is a powerful class of techniques well-suited for non-linear regression tasks. By configuring a set of parameters, we ensure an efficient learning process.

We first established the boosting type as 'gbdt', signifying gradient boosting decision trees. This type of boosting constructs an ensemble of decision trees sequentially, refining their performance through gradient-based optimization. The objective function was designated as 'regression', aligning with the study's goal of predicting continuous numerical outcomes. The assessment of model accuracy was based on the root mean squared error (RMSE) metric, a standard metric that quantifies the disparity between predicted values and actual outcomes and penalizes poor predictive performance quadratically. RMSE provides a comprehensive measure of the model's prediction accuracy by penalizing larger errors more heavily than smaller ones, making it particularly suitable for regression tasks like energy consumption prediction. However, we acknowledge that different evaluation metrics could offer additional insights. In the domain of energy consumption prediction, metrics such as Mean Absolute Error (MAE) could also be used alternatively. While these metrics could provide different perspectives on model performance, we opted for RMSE as MAE does not penalize large errors heavily, making it less sensitive to outliers than RMSE[6]. Nevertheless, we acknowledge the importance of considering domain-specific metrics tailored to specific applications or stakeholder requirements in future research endeavors, as they can offer nuanced insights and support more informed decision-making in the building energy sector. For example, Percentage of Energy Savings (PES) could quantify the percentage reduction in energy consumption achieved by implementing energy-saving measures recommended by the model. While these metrics provide valuable perspectives on model performance, we have chosen to focus exclusively on RMSE to maintain clarity and consistency in our evaluation methodology.

In order to enhance the model's ability to generalize patterns within the data, a subsample ratio of 0.25 was employed. This

Validation RMSE	Test RMSE
0.96884	0.974248
0.974681	0.980453
0.970309	0.975502
0.972683	0.979373

Table 2 RMSE values of the training set and validation set for each fold

technique entailed utilizing a random subset of 25% of the training data during each boosting iteration, fostering diversity in the training process and reducing overfitting risks. The subsample frequency, set at 1, indicated that the subsampling occurred in every boosting iteration.

A learning rate of 0.4 was chosen. This parameter determined the step size of each iteration's gradient descent, ensuring a balanced trade-off between swift learning and fine-tuning. To regulate the complexity of individual trees and the ensemble, the maximum number of leaves per tree was confined to 20. This restraint prevented over-segmentation, thereby curtailing the model's potential to memorize noise. These two hyperparameters are tuned by the grid search method using the GridSearchCV function from the scikit-learn library. For the learning rate parameter, four values "0.2, 0.3, 0.4, 0.5" are tested while the range of values for number of leaves is "5, 10, 20, 30".

Feature selection, an integral aspect of the model design, involved the strategic incorporation of informative attributes. A feature fraction of 0.9 was employed, indicating that during each boosting iteration, 90% of the features were considered, promoting diversity and reducing the risk of overfitting. Additionally, L1 and L2 regularization terms were introduced with lambda values set at 1, imposing penalties on the magnitudes of the model's coefficient to reduce model complexity and improve feature selection.

To ensure robustness and reliability, a K-fold cross-validation strategy was adopted. We employed Stratified K-fold with 4 folds, preserving the distribution of the target variable across the dataset. The random seed was set to 666 to facilitate reproducibility. In the implementation process, each fold was used iteratively as a validation set while the remaining folds were merged for training. For every fold iteration, a LightGBM dataset was created for both training and validation data, incorporating categorical feature information. The boosting process was executed with a maximum of 500 rounds, and early stopping was integrated into the training using a patience of 100 rounds.

Results

As seen in Table 2, the training phase consistently yielded a low RMSE of approximately 0.96 to 0.97 kWh, indicating the model's adeptness at fitting to the intricacies of the training data. However, the true test of the model's utility lies in its ability to generalize its predictions to previously unseen data, which is of paramount importance in real-world scenarios. In this regard, the validation phase produced RMSE values ranging from 0.974248 to 0.980453 kWh across different cross-validation folds. This slight but meaningful increase in RMSE from training to validation is indicative of the model's robustness in handling new, unseen instances. Notably, the modest disparity between training and validation RMSE highlights the model's capacity to strike a balance between capturing intricate training data patterns and avoiding overfitting, which often occurs when a model becomes too tailored to the training data. These results hold profound implications for building energy consumption management. The narrow margin between training and validation RMSE signifies that our models excel in making accurate energy consumption predictions even in real-world scenarios.

However, because log transformation is applied to our energy reading, in this scenario, an RMSE of 0.97 on the log-transformed scale implies that, on average, the absolute difference between the logarithm of the predicted energy consumption and the logarithm of the actual energy consumption is 0.97. Therefore, the predictions need to be back-transformed into the original scale from the log scale in order for the RMSE value to be directly interpretable, for which an RMSE value of about 2.63 is obtained. This means that, on average, our model's predictions deviate from the actual energy consumption by 2.63 kWh. In practical terms, achieving this level of accuracy implies that building managers or energy professionals can rely on our model to make reasonably accurate predictions of energy consumption, enabling them to optimize energy usage, plan maintenance schedules, and implement energy efficiency measures more effectively. Moreover, understanding the practical implications of RMSE values involves considering the impact of prediction errors on decision-making and resource allocation. For example, if our model consistently underestimates energy consumption (negative bias), building managers might risk inadequate heating or cooling, leading to discomfort for occupants or equipment failures. Conversely, if the model consistently overestimates energy consumption (positive bias), building managers might overspend on energy bills unnecessarily. Therefore, achieving lower RMSE values signifies higher prediction accuracy, thereby reducing the risk of such adverse outcomes and enabling more informed and efficient energy management decisions.

Figure 8 provides valuable insights into the relative importance of features in our energy consumption prediction model. While 'building_id' and 'site_id' emerge as the top-ranking features in terms of importance scores, it's important to recognize that these identifiers may not directly influence energy consumption patterns but rather serve as categorical markers for different buildings and sites. Consequently, their high importance scores may reflect the inherent variability across buildings rather than

actionable insights for energy management.

On the other hand, the prominence of 'hour' as the second most influential feature underscores the temporal aspect of energy consumption. This finding suggests that energy usage patterns vary significantly throughout the day, likely influenced by factors such as occupancy patterns, work schedules, and climatic conditions. Understanding these temporal dynamics is crucial for implementing effective energy management strategies, such as optimizing HVAC systems, scheduling equipment operations, and implementing demand response measures.

Moreover, the notable impact of weather-related features, particularly 'air_temperature', highlights the strong relationship between ambient conditions and energy consumption. As evidenced by Figure 2, variations in temperature exhibit a pronounced effect on energy usage, with increased heating or cooling demands during extreme weather conditions. Recognizing these dependencies enables stakeholders to anticipate energy demand fluctuations, adjust building operations accordingly, and implement proactive energy conservation measures to mitigate adverse effects on energy costs and environmental sustainability.

Among building metadata features, 'square_feet' emerges as a significant contributor to the model's predictions, albeit with a comparatively lower importance score compared to weather variables. Building area serves as a proxy for the size and scale of the facility, influencing factors such as heating and cooling loads, lighting requirements, and overall energy demand. While 'square_feet' plays a pivotal role in shaping energy consumption patterns, its lower importance score relative to weather features underscores the multifaceted nature of energy usage in buildings, which is influenced by a myriad of factors beyond just physical dimensions.

Furthermore, the identification of meter type as a relatively important feature underscores the heterogeneity in energy consumption behaviors across different building sectors or usage categories. Understanding the distinct energy profiles associated with different meter types facilitates targeted interventions and tailored energy management strategies tailored to specific building types or functions.

After training our model, we applied it to the test dataset and obtained the result shown in Figure 9. We can see that the skewness of the actual energy reading data and the prediction made by our model are very similar. This synchronization between the actual and predicted energy distributions implies that our model has been successful in comprehending the complex interplay between building metadata, weather data, and energy consumption behaviors. This outcome underscores the model's aptitude in not only predicting energy consumption levels but also in reproducing the intricate distributional nuances present in the authentic energy data. As such, our machine learning model emerges not only as a predictive tool but also as a means to gain a deeper understanding of the intricate dynamics driving energy consumption within buildings. This harmonization between the distributions serves as a testament to the model's potential in aiding decision-makers in formulating well-informed energy management strategies and underscores its relevance in advancing sustainable practices within the realm of building energy consumption.

Discussion

From data analysis and predictive modeling using LightGBM, our study has unveiled an intricate relationship between weather patterns and energy consumption. This connection, while undeniably present, defies simplistic linear interpretations, demonstrating a multifaceted interplay that requires deeper exploration.

Interestingly, our examination of building attributes has yielded unexpected findings. While conventional wisdom would suggest that building-specific characteristics significantly influence energy consumption, our results suggest otherwise. Surprisingly, the impact of building attributes appears to be relatively subdued compared to other factors. Notably, the primary use of a building emerged as the second least important feature, as illustrated in Figure 8. Despite assumptions that certain usage types, such as education, would consistently correlate with higher energy usage due to their frequent occupancy within the dataset, our analysis unveils a more intricate scenario where energy consumption is not solely dictated by the popularity of a particular usage. Delving deeper into this aspect could shed light on the underlying dynamics driving energy consumption patterns across different building types and usage categories.

Moreover, another building-specific characteristic, which is 'square_feet', displays its influence to be less pronounced compared to weather variables and temporal factors. Despite its intuitive relevance, the subdued importance of 'square_feet' in the model suggests that other factors, such as building occupancy patterns, operational efficiency, and energy management practices, may play a more substantial role in determining energy consumption levels.

The apparent humble influence of building attributes, despite their intuitive relevance, may stem from several factors that warrant exploration. Firstly, while building-specific characteristics such as size, layout, and construction materials are undoubtedly influential in shaping energy consumption patterns, their impact may be overshadowed by other dynamic factors such as occupant behavior, operational practices, and technological advancements. For instance, advancements in energy-efficient building designs and technologies may mitigate the influence of certain building attributes on overall energy consumption, thereby reducing their relative importance in predictive models. Understanding the underlying mechanisms driving these disparities in feature importance could offer valuable insights into optimizing energy efficiency strategies and resource allocation efforts in building management practices.

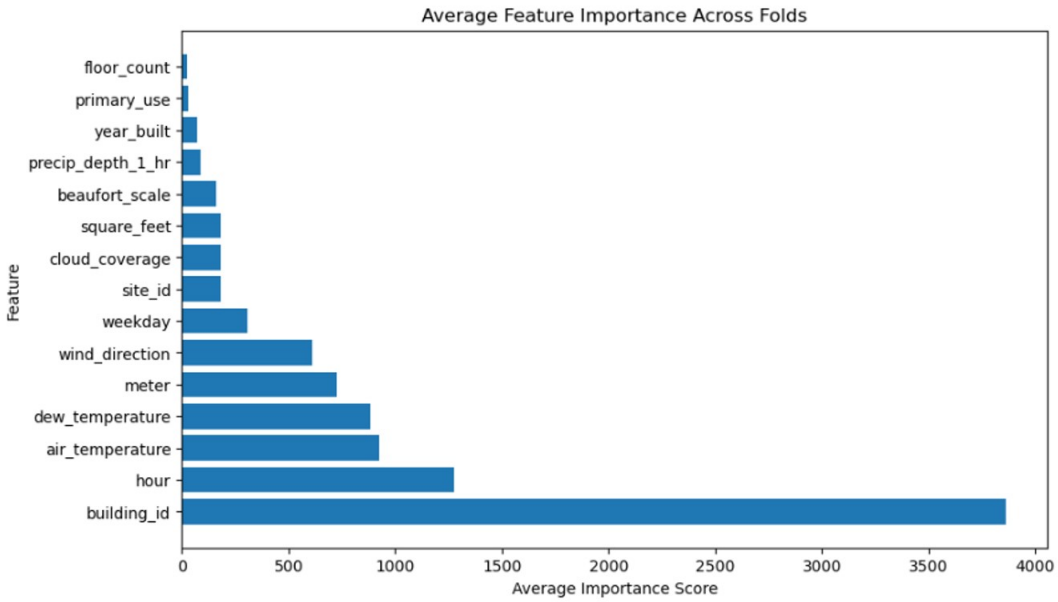

Fig. 8 Average importance score of different features used in training

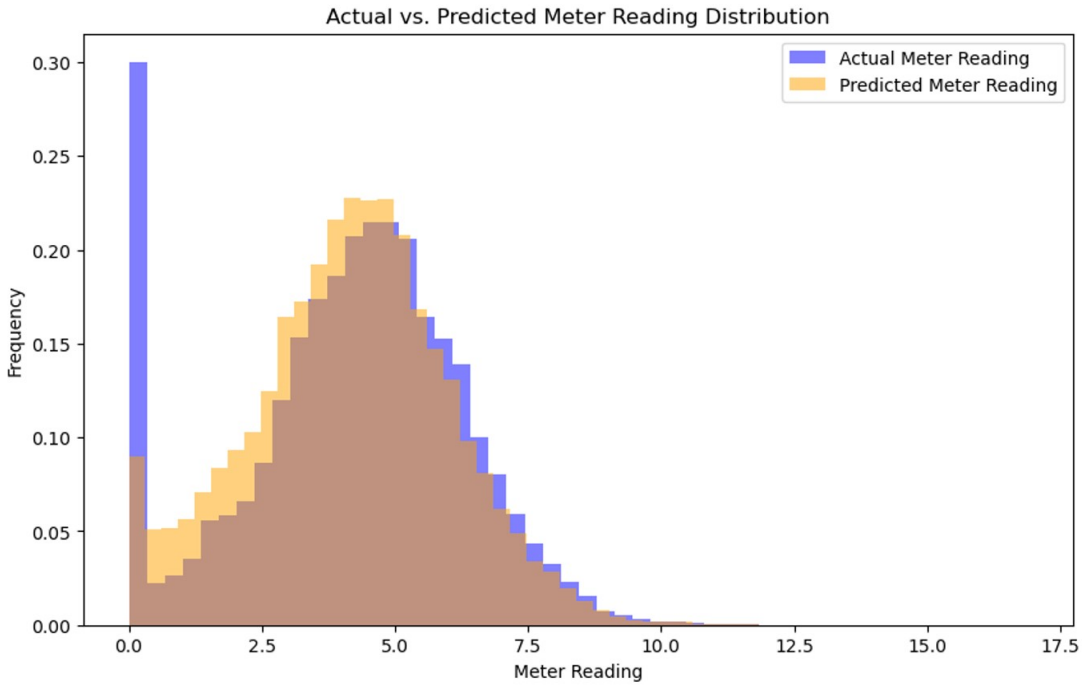

Fig. 9 Overlapping distribution graphs of actual meter_reading and predicted meter_reading, brown indicating overlap

Furthermore, our investigation into temporal dynamics revealed the pivotal role of time in influencing energy consumption patterns. Specifically, the temporal aspect emerged as a crucial factor, with specific hours throughout the day exerting a remarkable influence on overall energy consumption trends, as shown in Figure 8. This temporal dependency underscores the importance of considering not only the physical and functional aspects of buildings but also the dynamic temporal dynamics

that contribute to energy utilization. These findings highlight the need for tailored energy management strategies that account for temporal variations in energy demand and usage patterns.

Understanding these temporal patterns has significant practical implications for energy management practices. For example, during peak hours of energy consumption, such as midday or early evening when occupancy and activity levels are high, building managers can implement load-shifting strategies to redistribute energy-intensive tasks to off-peak hours. This could involve scheduling HVAC systems, lighting, and equipment operations to coincide with periods of lower energy demand, thereby reducing overall energy costs and minimizing strain on the electrical grid.

Additionally, insights into temporal patterns can inform the development of demand response programs, where buildings can adjust their energy usage in response to grid conditions or pricing signals. By leveraging real-time data and predictive models, building managers can proactively manage energy consumption, participating in demand response initiatives to earn incentives or avoid peak demand charges.

Moreover, understanding the temporal dynamics of energy consumption enables better planning and optimization of energy efficiency measures. Building automation systems can be programmed to adjust temperature setpoints, lighting levels, and ventilation rates based on anticipated occupancy patterns and external factors such as weather forecasts. By aligning energy usage with building occupancy and operational schedules, organizations can optimize comfort levels for occupants while minimizing energy waste and operational costs.

Limitation

The identification of zero values in the 'meter_reading' column presents a significant limitation in our study, warranting careful consideration and mitigation strategies. Approximately 9.27% of readings were registered as zeros, posing a challenge in data management and interpretation. To address this issue, we employed a logarithmic transformation after adding 1 to each energy reading. This transformation allowed us to incorporate the zero values into our analytical process while mitigating the impact of skewness and facilitating the modeling of the data.

While the logarithmic transformation provided a mechanism to handle the presence of zero values, it inadvertently introduced a layer of complexity. By fundamentally altering the distribution of the data, the transformation may have influenced the relationships and patterns inherent in the dataset. This alteration could potentially affect the predictive performance of our model, as the transformed data input may not fully capture the original nuances and variability present in the raw data. Additionally, the transformation may have implications for the overall robustness of our study, as the insights derived from the analysis may be influenced by the specific characteristics of the transformed data.

It is crucial to acknowledge the potential unintended consequences of the log transformation and its implications for the interpretation of our study's findings. While the transformation enabled us to address the challenge posed by zero values and proceed with our analysis, it introduces a degree of uncertainty regarding the validity and generalizability of our results[7]. Consequently, careful consideration should be given to the potential implications of the transformation on the outcomes and broader implications drawn from our study's findings. Future research endeavors may benefit from exploring alternative approaches to handling zero values, such as imputation techniques or model-based adjustments, to mitigate the limitations associated with data transformations and ensure the robustness of the analytical process.

Conclusion

In conclusion, this research paper has traversed the domain of predictive modeling for energy consumption in buildings, harnessing metadata and weather patterns through Machine Learning techniques. Leveraging a dataset sourced from the ASHRAE Great Energy Predictor III Kaggle competition, our exploration yielded significant insights. In the Exploratory Data Analysis (EDA) phase, we unearthed several trends and patterns, notably revealing a discernible relationship between temperature fluctuations, temporal variations in hours and months, and shifts in energy usage.

Central to our research was the development of a predictive model using the potent LightGBM framework. The outcome was compelling, with our model achieving a commendable average Root Mean Square Error (RMSE) range of 2.63 on both training and validation cross-validation sets. The close alignment between the skewness of predicted energy consumption and actual data underscores the model's effectiveness in capturing intricate data relationships, suggesting its potential for generalization beyond the training context.

However, it is imperative to acknowledge the inherent complexity of real-world energy dynamics, extending beyond the scope of this study. Factors such as building operations and occupant behaviors contribute to nuanced consumption patterns. Nonetheless, this research underscores the potential of data-driven insights in shaping energy-efficient strategies. By enabling stakeholders to predict consumption patterns, our model empowers them to optimize energy usage effectively, leading to tangible benefits in terms of cost savings, environmental sustainability, and operational efficiency in building energy management.

Moving forward, the practical implications of our research findings for the field of building energy consumption management are substantial. Our model can be applied in real-world scenarios to inform decision-making processes related to en-

ergy management strategies, building retrofitting initiatives, and investment in energy-efficient technologies. By leveraging predictive analytics, building managers can anticipate fluctuations in energy demand, optimize resource allocation, and implement targeted interventions to reduce energy waste and enhance overall sustainability.

Moreover, as the global focus on energy efficiency intensifies, our research paves the way for strategic advancements that steer us towards a more sustainable energy future. Future studies could further explore the performance of different machine learning models, including ensemble methods, to identify the most effective approaches for energy consumption prediction and mitigate the impact of zero values. Additionally, addressing the identified limitations, such as the handling of zero values in the dataset using imputation methods such as regression imputation, and incorporating domain-specific metrics for evaluation could enhance the robustness and applicability of predictive models in real-world settings. By embracing a multidisciplinary approach and leveraging advances in data science and machine learning, we can continue to drive innovation and progress towards a more sustainable built environment.

Future studies could investigate the influence of additional external factors, such as socio-economic indicators, policy interventions, and technological advancements, on building energy consumption. By considering a broader range of contextual factors, researchers can develop more robust and adaptable energy consumption models that account for systemic changes and external influences.Furthermore, spatial analysis techniques, such as geospatial modeling and spatial clustering, can offer valuable insights into spatial variations in energy consumption patterns across different regions or building clusters. Future research could explore the application of spatial modeling approaches to identify spatially-varying factors influencing energy consumption and develop localized energy management strategies tailored to specific geographic areas.

Acknowledgements

I would like to thank Lucien Werner from California Institute of Technology for his mentorship throughout this entire process of exploring machine learning concepts and writing this research paper.

References

1 A. Allouhi, Y. Fouih, T. Kousksou, A. Jamil, Y. Zeraouli and Y. Mourad, *Journal of Cleaner Production*, **109**, 118–130.

2 S. Bourhnane, M. Abid, R. Lghoul, K. Zine-Dine, N. Elkamoun and D. Benhaddou, *SN Applied Sciences*, **2**, year.

3 M. Shapi, N. Ramli and L. Awalin, *Developments in the Built Environment*, **5**, 100037.

4 T. Le, M. T. Vo, T. Kieu, E. Hwang, S. Rho and S. W. Baik, *Sensors*, 2020, **20**, 2668.

5 G. Huang, *Journal of Physics*, **1754**, 012187.

6 T. Chai and R. R. Draxler, *Geoscientific Model Development*, 2014, **7**, 1247–1250.

7 J. Martín-Fernández, J. Palarea-Albaladejo and R. Olea, *Compositional Data Analysis: Theory and Applications*, p. 43–58.