Predicting Building Energy Consumption

Pranav Venkatesh[1], Aadarsh K Narayan[1], Jeriah Yu[1], Jose L Hernandez-Mejia[2], Michael Pyrcz[2,3].

[1]College of Natural Sciences, The University of Texas at Austin, 120 Inner Campus Drive, Stop G2500 Austin, Texas 78712

[2]Cockrell School of Engineering, The University of Texas at Austin, 200 East Dean Keeton Street, Stop C0300 Austin, Texas 78712

[3]Jackson School of Geosciences, The University of Texas at Austin, 2305 Speedway, Stop C1160 Austin, Texas 78712

CONTENTS

ABSTRACT

Construction of large new buildings needs to take into consideration their additional energy usage. Current models for predicting that usage leaves a “performance gap” that prevents an accurate estimation of energy use once the building is in use, leading to unexpected energy costs and additional load on utility grids. It is also imperative for grid operators to have higher-resolution details on the load of their grid’s buildings to generate power at peak efficiency. To solve this necessity, the authors have developed a model to better forecast predicted energy demand precisely, taking into consideration the local weather patterns and seasonal climates through meteorological time-series data, and the characteristics of the constructed building. It is discovered that employing an extensive data imputation process to fill in data gaps - building floor count, year built, seasonal weather data - significantly reduces the error rate for LGBM modeling.

1 INTRODUCTION

Large commercial buildings use substantial amounts of energy per floor space for lighting, heating, and ventilation, and the energy used to power them ranges from electricity to district energy to chilled water (US Energy Information Administration, 2018). Before constructing new buildings, developers need to consider the amount of energy that they would consume when in use. However, current methods of prediction leave a "performance gap" between the predicted usage and actual usage (Abdellatif and Brady, 2017). This presents an issue for measuring efficiency and energy costs for large complexes like university campuses and especially presents a challenge for independent microgrids that generate their own energy such as the University of Texas at Austin. The campus generates all the energy required for the campus's electricity, heating, and cooling using district energy as a cost-saving and efficiency measure (Malewitz, 2016), so it requires precise metrics of the maximum, minimum, and average energy load of every building in the campus down to every hour to avoid potentially ruinous energy deficits and blackouts (Cooper, 2018).

The Great Energy Predictor, a machine learning competition sponsored by ASHRAE, has an extensive compilation of literature that covers successful approaches to the problem from various years. The most common methodology for solving this problem extensively employs tree-based models, specifically gradient boosting machine (GBM), to make effective time-series forecasts. GBM approaches have been more successful for this problem due to the datasets being tabular in nature (Miller, Hao, & Fu, 2022). The biggest deficiency with pre-existing methodologies is the lack of sufficient focus on data pre-processing, specifically for the building metadata.

A simple correlation analysis on this dataset displays the importance of certain attributes — floor count and year built — in predicting energy consumption. This solution is novel due to the extensive data imputation process throughout the weather and building metadata. To solve the problem of accurately predicting energy loads on power suppliers, the investigators use weather and temperature data collected at various sites to predict the precise energy usage of these large buildings based on their characteristics.

2 METHODS

2.1 DATA

The publicly available dataset (ASHRAE, 2019) contains 3 different spreadsheets: building metadata, weather data, and meter readings. The building metadata consists of year built, floor plan, sq ft, and site location. The weather data is hourly, and on a per-site basis. It contains precipitation, cloud coverage, air temperature, sea level, and wind speed. The meter readings include information about the type of meter, as well as the value for every hour during the year of 2016. The dataset package (ASHRAE, 2019) had several instances of missing data across its weather data and building metadata. To overcome this, the investigators extensively utilized data imputation to fill in data gaps and exploited bit memory properties to reduce the overall memory load of the dataset package to optimize performance. Fig. 1 illustrates the degree of inconsistency present in our training data. A large percentage of readings are zero or contain extraneous spikes.

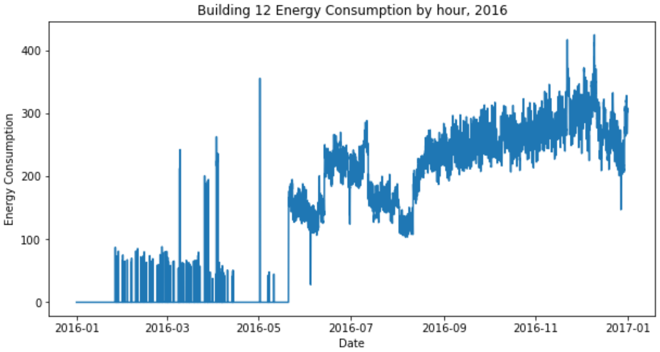

Fig. 1 Example of Meter Reading with outliers, in this from case Building 12

Referencing Sandeep Kumar's weather cleaning methodology (Kumar, 2020), the investigators systematically (1) broke up each timestamp into their respective year, month, day, and hour — a necessary step for LGBM — and (2) imputed missing air temperature, cloud coverage, due temperature, sea level, and wind speed/direction by taking pertinent means. Some attributes were dropped entirely, like precipitation depth, due to their typically low correlation with the response variable according to relevant literature. Past studies typically agree that temperature is the most important factor regarding energy demand, while factors like precipitation are relatively less important - even if a dependence might be present (Fikru & Gautier, 2015).

The building metadata had a noticeable lack of year-built and floor count properties for a considerable number of buildings. Missing year-built data was imputed by using the mean of the buildings on the same site, and correlation was considered between certain building metadata properties (square footage, building type) and energy properties (average energy consumption vs. number of floors). A simple multivariate regression model, using the previously noted attributes as input, predicted the number of floors in a building. It yielded a relatively low RMSE of 2.36, which was satisfactory for use, though a possible optimization would be to reduce this. This regression model then systematically imputed the missing floor count values with a low error. The results are also visually apparent - Fig. 2 illustrates that the overall distribution became a lot less skewed following our imputation.

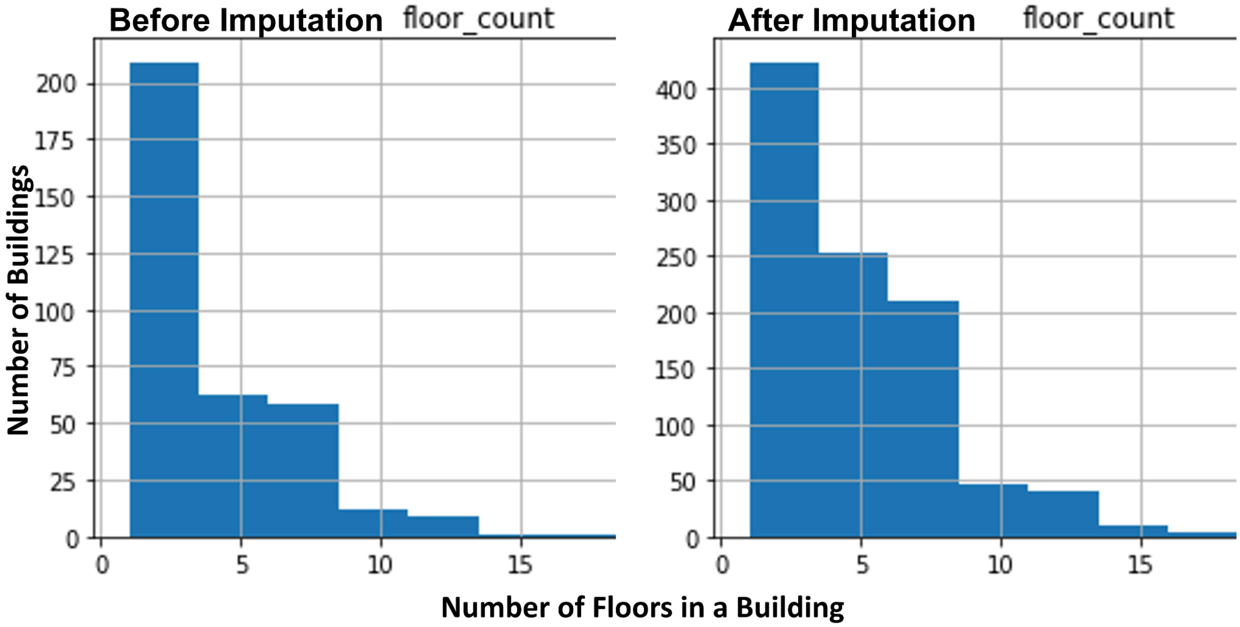

Fig. 2 Floor Count distribution before and after imputation

2.2 MODELS

Several approaches were taken to forecast energy consumption by hour for comparison. Various gradient boosting models — CatBoost, LightGBM, and XGBoost — and time series prediction models — Prophet and ARIMA — were considered. CatBoost, on average, is slower than LightGBM while XGBoost is less intuitive than LightGBM. LightGBM is the kaggle boosting framework of choice — due to its time complexity of *O(0.5 * #feature * #bin)* - which, along with its incredibly fast training turn-around given computing constraints, made it the clear choice for the multivariate approach. Running with this idea of a quick training turn-around, it made sense to choose Prophet due to its flexibility with hyperparameter tuning, its pattern-first design, and its time complexity of *O*(# simulations * length of forecast). Therefore, the final selections were Facebook's Prophet Model as a univariate linear approach and Light Gradient Boosting as a multivariate approach. Facebook Prophet was selected to compare the results from a more traditional time series approach that depends primarily on periodic information with final results. Light Gradient Boosting was selected due to its effectiveness at detecting trends in complex multivariate data and ability to make complex connections with its estimator terms.

The Facebook Prophet Model (Taylor & Letham, 2017) is comparable to other time series models like SARIMA in that it relies on deriving trends from past values and seasonalities. At its core is the following Equation 1, where y(t) is the forecasted output, g(t) is growth, s(t) is seasonality, h(t) adjusts for holiday, and ϵ_t is error.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \qquad (1)$$

Overall, the model was limited by the dataset - many meter readings were unavailable, which made it difficult for a model like Prophet to determine seasonalities. This is because previous values over periodic intervals are the primary estimator. With many such values missing, it follows that the model did not have sufficient information to generate accurate predictions. Fig. 3 shows predictions using the Prophet model.

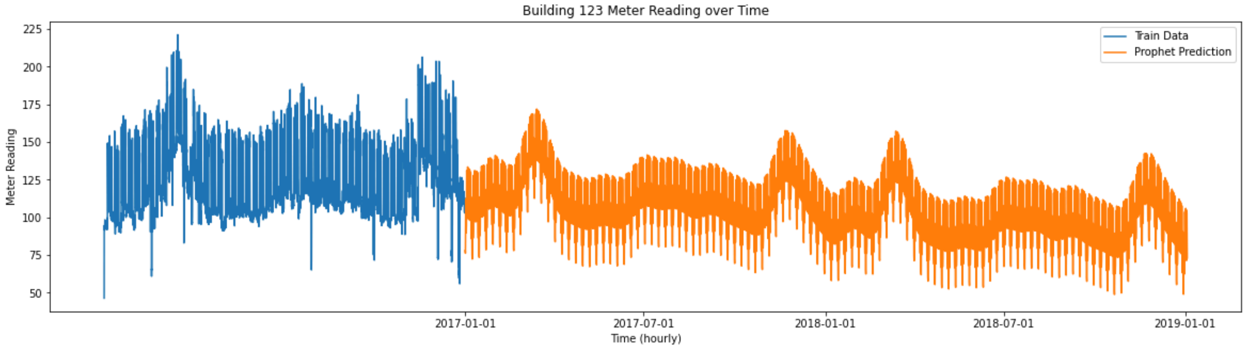

Fig. 3 Prediction for energy consumed over time using Prophet.

The LGBM model is a modified version of gradient boosting to reduce memory usage (Ke, 2017). The investigators followed an approach outlined in a paper by Banerjee on Kaggle. The core idea is to iteratively improve the model according to an error metric by adding an estimator term. For the LGBM model, the timestamps were decomposed, turning the year, month, day, and hour into additional vectors. This also made it possible to predict useful meter readings for buildings which did not have a meter reading on a given day during the year of provided training data. A randomized 80/20 train-test split was used. Fig. 4 shows predictions using the LGBM model after including data imputations and removing outliers, which is discussed in greater detail in the results section.

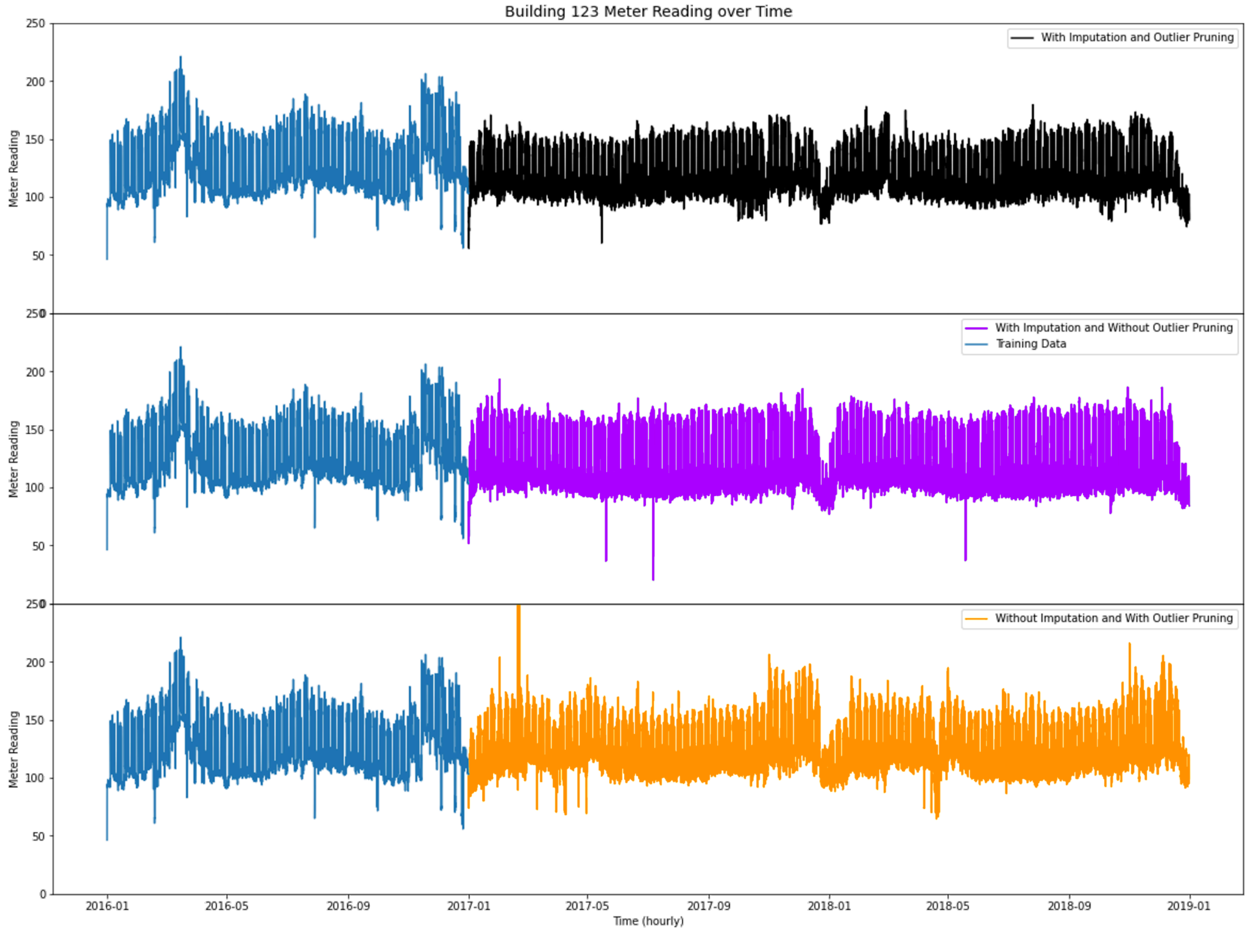

Fig. 4 Prediction for energy consumed over time using LGBM.

3 RESULTS

Prophet RMSE With Imputation and Outlier Pruning on Testing Data	LGBM RMSE With Imputation and Outlier Pruning on Our Testing Data	LGBM RMSE With Imputation and without Outlier Pruning on Testing Data	LGBM RMSE Without Imputation and with Outlier Pruning on Testing Data
34.1757	0.1456	0.3579	0.3755

Fig. 5 RMSE on our testing data for each model

The Prophet model framework handles hyperparameter selection automatically, but its accuracy was quite low with a total RMSE of 34.18. It utilizes a complex L-BFGS optimization, which itself has a non-trivial time complexity and constant memory. The LGBM model was trained with a learning rate of 0.1, 2000 boost rounds, and 20 early stopping rounds. The RMSE was 0.3755 with just outlier removal, 0.3579 with just imputation, and 0.1456 with both. The training of the model with no outlier removal involved not removing zero-values when feeding data into the model. The model with no imputation involved not predicting missing values (as described in the method section above). By running the model with and without these features, a point of comparison and improval was able to be made for the complete model.

LGBM unsurprisingly outperformed Prophet due to the nature of the problem and the models. The problem necessitated a multivariate analysis capable of working with a constrained dataset due to most buildings having long periods of time where their meters were not in operation. But Prophet, which uses past inputs from the same series based on periodic intervals to perform predictions, could only perform well when data was present for the whole year. Prophet also could not make adequate use of other data sources: it needs data over the period of prediction and would not be able to quickly use a large number of added regressors. LGBM, on the other hand, could quickly learn from the entire dataset due to its light decision tree design.

The LGBM predictions were submitted to the ASHRAE Great Energy Prediction competition hosted on Kaggle. Overall, the model with outlier pruning and imputations greatly outperformed submissions that only implemented one method. The total score came in at 1.283, which is in the 83rd percentile of participants with a rank of 447/3592. Without imputation, the score dropped to 1.287, and without pruning, it dropped to 1.474. While it is clear that many improvements can be made, it is also clear that outlier pruning and data imputation provide a significant improvement to overall prediction performance.

4 DISCUSSION

Regarding results from the Prophet model, the main takeaway is that the missing values in the time series significantly reduces the predictive power of the model. This is because Prophet is a more traditional time series, making future predictions based on the same value at the previous interval. Being denied this data without having access to additional regressors like weather makes it incredibly difficult to predict with reasonable accuracy. For LGBM, the main improvement could be made in processing invalid readings through further statistical processing to remove outliers rather than just detecting and removing missing values and zeros.

5 CONCLUSION

Imputation alone results in an improvement of the RMSE from 0.3755 to 0.3579. Combining outlier pruning and data imputation improved results to yield an RMSE of 0.1456. The LGBM model with our data processing regime achieves an RMSE of 0.1456 for forecasting, outperforming the Prophet model which has an RMSE 34.18. The models can be improved with focused feature cleaning and enhanced feature imputation. Further work remains to test the models on sourced real-world non-filtered data from UT and to improve the efficiency of the models and processing algorithms in training and validation.

6 REFERENCES

"2018 Building Characteristics Flipbook." 2021. US Energy Information Administration.

"Ashrae - Great Energy Predictor III." 2019. Kaggle. ASHRAE. https://www.kaggle.com/c/ashrae-energy-prediction (Accessed January 28, 2022).

Banerjee, Koustav. 2020. "Beginner-Friendly Great Energy Predictor with LGBM." Kaggle. https://www.kaggle.com/code/kens3i/beginner-freindly-great-energy-predictor-with-lgbm/notebook#Merging-the-features-of-building_meta_df-with-train_df-and-test_df (Accessed March 1, 2022).

Brady, Laurence, and Mawada Abdellatif. 2017. "Assessment of Energy Consumption in Existing Buildings." *Energy and Buildings* 149 (August 15, 2017): 142–50. https://doi.org/10.1016/j.enbuild.2017.05.051.

Cooper, Rachel. 2018. "How One Campus Complex Powers the Entire Forty Acres." The Alcalde, Texas Exes. https://alcalde.texasexes.org/2018/06/how-ut-runs-the-largest-microgrid-in-the-us/.

Ke, G., et al, 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". *Advances in neural information processing systems*, 30.

Kumar, Sandeep. 2019. "ASHRAE - Missing Weather Data Handling." Kaggle. https://www.kaggle.com/code/aitude/ashrae-missing-weather-data-handling/notebook. (Accessed March 4, 2022).

Mahelet G. Fikru, Luis Gautier. 2015. "The impact of weather variation on energy consumption in residential houses". *Applied Energy*, Vol 144. 19-30. ISSN 0306-2619. https://doi.org/10.1016/j.apenergy.2015.01.040.

Malewitz, Jim. 2016. "UT-Austin Gets Bigger, but Its Energy Bills and Emissions Are Shrinking." The Texas Tribune. The Texas Tribune, November 23, 2016. https://www.texastribune.org/2016/11/23/ut-austin-keeps-getting-bigger-its-energy-bills-an/. (Accessed March 9, 2022).

Miller, C., Hao, L., & Fu, C. (2022). "Gradient boosting machines and careful pre-processing work best: ASHRAE Great Energy Predictor III lessons learned". *arXiv preprint arXiv:2202.02898.*

Taylor SJ, Letham B. 2017. Forecasting at scale. *PeerJ* Preprints. https://doi.org/10.7287/peerj.preprints.3190v2.