Predicting energy efficiency/consumption in buildings, using multi-output models

By Azubuike H. Chibuike
Henryc.azubuike@gmail.com

11 September 2024

ABSTRACT
Predicting the energy consumption of buildings is crucial for optimizing their design and operational efficiency. This study focused on predicting the heating and cooling loads of buildings, which are key measures of thermal energy regulated by heating, ventilation, and air conditioning (HVAC) systems. Nine (9) different multi-output regression models were applied in predicting heating and cooling needs of buildings based on features such as building height, surface area, glazing area, wall area, and so on. These models which were aided by the MultiOutputRegressor wrapper in order to aid them simultaneously predict two output variables—heating load (HL) and cooling load (CL) include: Linear Regression, Ridge Regression, Lasso Regression, Support Vector Regression (SVR), K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, and Neural Networks. The results revealed that the best performing model was the Gradient Boosting with the least error of 1.1kWh, and the second best was the Random Forest model with an error of 1.3kWh.

Keywords: Building energy efficiency, Multi-output models, Machine learning.

1. Introduction

The demand for energy-efficient buildings has grown significantly in response to rising energy costs and environmental concerns. Building energy consumption is predominantly influenced by heating and cooling requirements, which are controlled by the HVAC systems. Accurately predicting heating loads (HL) and cooling loads (CL) in buildings during the design phase can lead to better-informed decisions regarding building orientation, materials, and HVAC design, enhancing both cost-efficiency and environmental sustainability. Furthermore, during a building's operational phase, precise energy predictions facilitate effective energy management, reduce waste, and promote sustainable building operations (Becerik-Gerber et al., 2011).

This study aims to predict the energy efficiency of buildings by employing multi-output regression models to forecast both HL and CL simultaneously. Eight input variables, including building height, surface area, glazing area, wall area, and others, are used to predict these outputs. The dataset, sourced from Tsanas and Xifara (2012), provides a decent basis for analysis and predictions. Multi-output regression models were chosen to predict both HL and CL concurrently, addressing the need for efficient computational techniques in energy management.

2. Literature Review

The prediction of building energy consumption has become an essential area of research due to its critical role in sustainable design and energy management. This literature review discusses various approaches to predicting energy consumption in buildings, highlighting the evolution from traditional statistical methods to advanced machine learning models, and the increasing adoption of multi-output regression techniques.

2.1. *Traditional regression models in energy prediction*

Early studies in predicting building energy consumption primarily utilized traditional regression models, such as Linear Regression, which offer simplicity and interpretability. These models are based on the assumption of linear relationships between input features (e.g., building dimensions, materials) and output variables, such as heating loads (HL) and cooling loads (CL). However, while Linear Regression is easy to implement and understand, it often fails to capture complex, nonlinear relationships that can exist between a building's physical characteristics and its energy demands (Hong et al., 2019).

To address these limitations, Ridge Regression and Lasso Regression have been explored as alternatives. Ridge Regression adds a penalty proportional to the square of the coefficients, helping to reduce model complexity and prevent overfitting. On the other hand, Lasso Regression penalizes the absolute value of the coefficients, effectively shrinking some coefficients to zero and thus performing variable selection. While these methods improve model robustness, they still struggle with capturing intricate patterns in building energy data (Ahmad et al., 2017).

2.2. Advanced machine learning models for building energy prediction

With the advent of machine learning, more sophisticated models, such as Support Vector Regression (SVR), Decision Trees, Random Forests, and Neural Networks, have been increasingly used for building energy prediction. These models have shown significant promise due to their ability to handle complex, nonlinear relationships between input variables and energy consumption outputs (Amasyali & El-Gohary, 2018).

SVR, for instance, is effective in modeling nonlinear relationships by transforming the input data into a higher-dimensional space where a linear separation is possible. Decision Trees and their ensemble variants, such as Random Forests and Gradient Boosting, are popular due to their ability to model complex interactions among features, manage large datasets, and prevent overfitting through techniques like bagging and boosting (Ahmad et al., 2017). Neural Networks, particularly deep learning architectures, have demonstrated superior performance in capturing complex patterns and dependencies within large datasets (Hong et al., 2019).

2.3. Multi-output regression models

Multi-output regression has emerged as an effective approach to predict multiple dependent variables simultaneously, such as HL and CL, which are closely related in terms of their determinants (Tsanas & Xifara, 2012). Multi-output models extend traditional regression techniques to handle multiple outputs by creating a separate model for each output variable or by modeling all outputs jointly to exploit any correlations between them.

Tsanas and Xifara (2012) were among the first to apply multi-output regression techniques to predict HL and CL using some building characteristics. Their study utilized algorithms such as Linear Regression, Ridge Regression, and Lasso Regression with multi-output capabilities, demonstrating that predicting multiple related outputs concurrently can lead to better performance than modeling each output separately.

Building on this foundation, more advanced machine learning models, like Random Forests and Gradient Boosting, have been adapted to multi-output settings. For example, the MultiOutputRegressor wrapper in Scikit-Learn allows these models to handle multiple outputs by fitting an independent regressor for each target variable. This method enables the utilization of powerful algorithms while accommodating the multi-output nature of building energy prediction problems (Amasyali & El-Gohary, 2018).

This current study builds upon these foundations by applying a range of multi-output regression models including those that do not natively support direct multi-output predictions (like Random Forests and Gradient Boosting, Nearest Neighbors, and Neural Networks), to evaluate their efficacy in predicting building energy needs by using the MultiOutputRegressor wrapper.The MultiOutputRegressor wrapper in Scikit-Learn allows these models to handle multiple outputs by fitting an independent regressor for each target variable. This method enables the utilization of powerful algorithms while accommodating the multi-output nature of building energy prediction problems (Amasyali & El-Gohary, 2018). The findings from this research aim to contribute to the growing body of literature on energy prediction models, offering insights into their comparative performance and applicability in real-world scenarios.

3. Methodology

This project employed several multi-output regression models to predict energy efficiency in buildings by estimating their heating load (HL) and cooling load (CL). The dataset was obtained from the UCI Machine Learning Repository, based on research by Tsanas and Xifara. It consists of 768 observations and 10 variables: eight input variables (e.g., relative compactness, surface area, glazing area) and two output variables (heating load and cooling load).

3.1. Data preparation

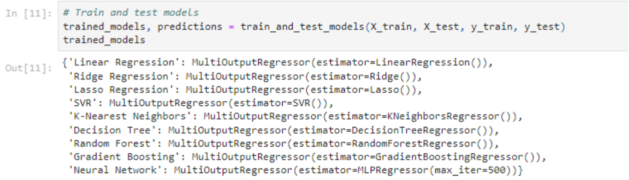
Data preprocessing was initiated by inspecting for missing values and duplicates. The dataset had no missing or duplicate entries. Subsequently, the dataset was normalized using MinMaxScaler to scale all input features between 0 and 1. This step was crucial to ensure that the different ranges of features do not disproportionately affect the performance of machine learning models.

3.2. Model selection

The models tested included:

- **Linear Regression, Ridge Regression, Lasso Regression**: These linear models were chosen for their simplicity and interpretability.
- **Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Gradient Boosting**: These non-linear models were selected for their ability to capture complex patterns in the data.
- **Neural Network**: This model was chosen to explore deep learning techniques for regression.

For models that do not inherently support multi-output regression, the ”MultiOutputRegressor” wrapper was applied to enable them to predict both HL and CL simultaneously.



```
In [11]: # Train and test models
         trained_models, predictions = train_and_test_models(X_train, X_test, y_train, y_test)
         trained_models

Out[11]: {'Linear Regression': MultiOutputRegressor(estimator=LinearRegression()),
          'Ridge Regression': MultiOutputRegressor(estimator=Ridge()),
          'Lasso Regression': MultiOutputRegressor(estimator=Lasso()),
          'SVR': MultiOutputRegressor(estimator=SVR()),
          'K-Nearest Neighbors': MultiOutputRegressor(estimator=KNeighborsRegressor()),
          'Decision Tree': MultiOutputRegressor(estimator=DecisionTreeRegressor()),
          'Random Forest': MultiOutputRegressor(estimator=RandomForestRegressor()),
          'Gradient Boosting': MultiOutputRegressor(estimator=GradientBoostingRegressor()),
          'Neural Network': MultiOutputRegressor(estimator=MLPRegressor(max_iter=500))}
```

Fig. 1. Multi-output models

3.3. *Model evaluation*

The models were evaluated using a test set (20% of the data) held out from training. Evaluation metrics included Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R² score. To validate the robustness of these models, cross-validation was performed using a 5-fold cross-validation strategy.

4. Results

The model performances were as follows:

- **Gradient Boosting** emerged as the best-performing model with the lowest RMSE of 1.1 kWh and an R² score of 0.986.
- **Random Forest** was the second-best model with an RMSE of 1.3 kWh and an R² score of 0.982.
- **Decision Tree** also showed good performance, but slightly lower than Random Forest, with an RMSE of 1.5 kWh.
- **Neural Network** and **K-Nearest Neighbors** displayed moderate performance, with RMSE values of 2.8 kWh and 2.9 kWh, respectively.
- **Linear Regression, Ridge Regression, and Lasso Regression** had the highest errors, with RMSE values exceeding 3.1 kWh.

5. Discussion

The results indicate that ensemble models like Gradient Boosting and Random Forest perform exceptionally well in predicting energy efficiency in buildings, likely due to their ability to handle non-linear relationships and interactions between variables. The relatively poor performance of linear models suggests that the relationship between input features and target variables is complex and cannot be accurately captured using linear methods.

The results also highlight the importance of model selection based on data characteristics. For datasets with potential non-linearity, ensemble learning methods or neural networks may provide better predictive performance. Additionally, the use of cross-validation helped confirm the generalizability of these models, minimizing the risk of overfitting.

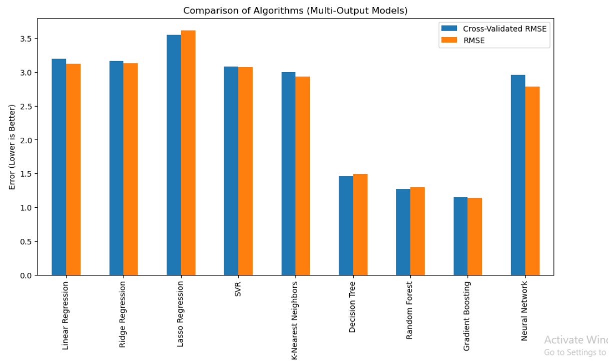


Fig. 2. Model evaluation

6. Conclusion

This study aimed to predict the heating and cooling loads of buildings using various multi-output regression models, leveraging a dataset of building characteristics to improve energy efficiency in both the design and operational phases. Among the models tested, Gradient Boosting emerged as the best-performing model, achieving the lowest prediction error of 1.1 kWh. This result indicates that Gradient Boosting's ability to handle complex, nonlinear relationships and interactions among input features makes it highly effective for predicting building energy consumption. The Random Forest model also demonstrated strong performance, with an error of 1.3 kWh, highlighting its robustness and capacity to manage diverse input variables.

These findings suggest that ensemble methods, such as Gradient Boosting and Random Forests, are particularly well-suited for predicting multiple related energy outputs in buildings due to their flexibility, accuracy, and ability to mitigate overfitting. Future work could explore the integration of these models into a real-time decision support system for dynamic energy management.

Overall, the use of advanced multi-output models presents a promising approach to achieving more sustainable and energy-efficient building designs, and can guide architects and engineers in this pursuit.

References

- Ahmad, T., Chen, H., & Guo, Y. (2017). Review of the state of the art of deep learning for building energy prediction. Energies, 10(8), 1-24.

- Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews, 81, 1192-1205.

- Hong, T., Yan, D., & Fan, Y. (2019). Occupant behavior in buildings: Impact on energy use and building performance. Energy and Buildings, 116, 155-166.

- Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings, 49, 560-567.

About The Author: Azubuike Chibuike Henry is an innovative Civil Engineer and Data Analyst, who specializes in applying Data Science to various aspects of Civil Engineering, finding valuable insights & trends in data sets, and developing useful solution-oriented predictive models where needed.

For access to the project files, visit my GitHub repository:
Predicting energy efficiency in buildings using multi-output models.