Effect of data imbalance in Machine Learning Models for building energy performance prediction

# Effect of Data Imbalance in Machine Learning Models for Building Energy Performance Prediction

Hamidreza Seraj[1], Ali Bahadori-Jahromi[1], Hooman Tahayori[2]
[1] *University of West London, London, UK; hamidreza.seraj@uwl.ac.uk*
[2]*Shiraz University, Shiraz, Iran*

***Abstract*— One of the promising methods that has recently gained attention for investigating building energy performance is utilisation of AI data-driven approaches, such as machine learning (ML) models. These methods, despite their advantages over physics-based models—such as faster prediction time and simplicity of application to case study buildings—have challenges during the model development stage, including the need for large datasets, potential overfitting, and the difficulty of capturing complex physical interactions within the building energy systems. As a result, this research aims to address data imbalance issue in developing a ML model for predicting the Energy Performance Certificate (EPC) rating of residential buildings. On this context, two resampling methods, including SMOTE (Synthetic Minority Over-sampling Technique) and SMOTE-Tomek were applied to XGBoost ML model to improve its accuracy. The results of this study showed that although applying data resampling methods slightly reduced the model's overall accuracy score (by less than 2%), it significantly enhanced the model's ability to predict minority classes. Specifically, the model's performance in predicting labels B, F, and E improved by more than 7%, 10%, and 6% points, respectively. This highlights how class imbalance in EPC labels can distort evaluation metrics like accuracy, potentially masking poor performance in minority classes. Addressing this imbalance is crucial for effectively integrating ML models into more advanced AI tools and smart systems for comprehensive building performance analysis.**

Keywords—Imbalanced data, machine learning, SMOTE, EPC rating, Building energy performance

## I. INTRODUCTION

The building sector accounts for more than 35 percent of global total energy consumption [1]. In response, many energy conservation plans have been developed to stop increasing buildings' energy demand. One of the key challenges in developing efficiency policy in this sector is predicting building energy performance. Generally, energy performance prediction models can be classified into two main categories: 1- parametric simulation tools which utilise physics based equations (e.g., heat transfer, electrical) 2- data-driven models which utilise artificial intelligence algorithm (e.g., machine learning algorithms) to discover non-linear relationships between inputs (e.g., building features) and outputs (e.g., annual energy consumption).

Although some research has been conducted to verify the applicability of parametric simulation tools, such as EnergyPlus, for building energy retrofits and the integration of renewable energy into building energy systems [2], utilisation of ML models to predict building energy performance metrics has recently gained attraction. This shift is driven by their faster prediction times, reduced dependency on detailed building information, and their integration with IoT and smart building systems. On this context, many studies have conducted to assess the accuracy and effectiveness of these models.

Seraj et al. [3] developed an AI tool to analyse effectiveness of machine learning models including ensemble learning algorithms and artificial neural network (ANN-MLP) in predicting residential buildings annual energy consumption. The results of this study highlighted that XGBoost (XGB) model outperformed others, however the $R^2$ value did not exceed 0.8 in the developed models. Furthermore, Razak et al. [4] presented utilisation of nine ML techniques including support vector machine (SVM) and Deep neural network (DNN) to predict the same metric in residential buildings. They also examined the effect of building clusters on model performance. The results of their study revealed the effectiveness of the DNN model, with an $R^2$ score of over 0.9. More studies about the effectiveness of ML models in predicting energy performance of different types of buildings can be found in [5].

One of the issues that can affect the accuracy of a model is imbalanced distribution of classes in the dataset; meaning that one class has significantly more instances than others. This can negatively impact models' training procedure, as the model tends to fit toward majority classes which leads to biased prediction and unreliable accuracy metrics. On this context, Zhang et al. [6] conducted a research to highlight the effect of data imbalance in building energy performance prediction. They showed that addressing data imbalance issue in a ML model can decrease building energy load prediction mean absolute error by 12% and enhance $R^2$ score up to 14%.

SMOTE (Synthetic Minority Over-sampling Technique) and SMOTE-Tomek are popular methods used to address the problem of imbalanced datasets by generating synthetic examples for the minority classes. In these techniques new instances will be created using interpolation between existing datapoints and it will populate the minority classes. Swana et al. [7] investigated integrating SMOTE technique with Naïve Bayes (NB), SVM, and K nearest neighbors (KNN) ML algorithms to improve fault detection accuracy in a wound-rotor induction generator. The results of this study revealed that SMOTE integrated with KNN outperformed other models that could increase the accuracy score almost 30 points.

As a result, this study aims to overcome the challenge of data imbalance in developing ML models to predict energy performance certificate (EPC) rating of residential buildings in the UK. In this regard, SMOTE and SMOTE-Tomek methods will be integrated with XGBoost ML model to answer following questions in this study: 1-How These techniques can improve the overall accuracy of the classifier model 2- its effectiveness in predicting minority classes of EPC rating in the case study area 3-which resampling model is more reliable to enhance model's performance.

## II. METHODOLOGY

### A. Dataset

ML based data-driven methods are typically trained on datasets reflecting a wide range of conditions from the case study. So, this study has utilised the EPC dataset for residential buildings in the UK, published by the Department for Levelling Up, Housing and Communities [8]. This dataset contains detailed numerical and categorical information on building envelope characteristics, energy systems, and estimated annual energy consumption.

### B. Data Pre-processing

Raw data often includes various irregularities, such as missing values, noise, inconsistencies, and redundancies. In this research, the following methods were applied to the raw dataset using the "Scikit-learn" Python package to enhance the accuracy and efficiency of the ML model [9]. First of all, common data irregularities were addressed by excluding case studies with outliers and missing values, ensuring more accurate observations for model training. Furthermore, Categorical features were handled using one-hot encoder, while numerical features were normalised to prevent any single feature from dominating the model.

Besides, new features were generated based on the provided building descriptions, such as calculating U-values for external walls from building characteristics and the UK standard assessment procedure (SAP) regulations. These steps ensured a well-balanced dataset with a mix of categorical and numerical features. Further details about the selected and designed features can be found in Table 1.

TABLE I. LIST OF FEATURES FOR MODEL DEVELOPMENT

| ***General details*** | ***Building envelop*** | ***Energy system*** |
|---|---|---|
| Property type<br>Built form<br>Total Floor area<br>Floor to ceiling height<br>Construction year | Glazing type<br>Glazing area<br>Floor type<br>Floor insulation<br>External wall U-value<br>External wall type<br>External wall insulation<br>Roof type<br>Roof insulation | Main heating system<br>DHW supply system<br>Secondary heating system<br>Main fuel type<br>Type of ventilation system<br>Lighting type<br>Installed PV capacity |

### C. Model Selection

For model selection, XGB was chosen due to its strong performance in existing literature, particularly for building energy performance and EPC prediction [10]. XGB not only offers high predictive accuracy but also delivers faster computation compared to other ensemble learning models, such as Random Forest and ANN-MLP. In this context, its scalability is a key advantage, as it runs up to ten times faster than many popular algorithms on a single machine [11]. Highlight of the key features and algorithms employed in the XGB model can be found in [3], [11].

### D. Data Imbalance

As mentioned earlier, imbalanced datasets can negatively affect the performance of ML models (particularly classifiers) in a way that the classifier system tends to be biased in favor of Majority classes. The classifier also tends to ignore the minority instances and detects them as noise [12]. To address this, three types of data resampling techniques including over-sampling, under-sampling, and combine sampling have been developed to balance the dataset. over-sampling is a technique to duplicate or produce new synthetic data within the minority class, whereas under-sampling is a technique to delete or merge data within the majority category. Combine-sampling is one method that combines over-sampling and under-sampling [13].

On this context, an over-sampling method SMOTE (synthetic minority over-sampling technique) and a combine-sampling method, SMOTE-Tomek, have been utilised in this research to help balance the dataset, and give the model more diverse training data for minority classes to improve its performance.

SMOTE is based on a k-nearest neighbor to generate new synthetic sampling in feature space based on a certain percentage for the minority classes. Moreover, SMOTE-Tomek is a hybrid method that combines SMOTE's synthetic oversampling with Tomek Links, a technique for cleaning data. SMOTE generates new synthetic samples for the minority class by creating data points between a minority instance and its nearest neighbors, helping balance the dataset. This synthetically generated data can be formulated as shown in equation 1 [7]:

$$S_{syn} = r\left(S_{knn} - S_f\right) + S_f \quad (1)$$

Where $S_{syn}$ is generated synthetic samples; $S_f$ feature samples; $S_{knn}$ considered feature sample k-nearest neighbor; and r is a random number between 0 and 1.

Moreover, Tomek Links are applied to remove noisy examples, specifically those that are difficult to classify because they are too close to instances of the opposite class [14].

### E. Model Performance Assessment

In order to assess the performance of the model, k-fold cross validation was utilised. In this approach, the dataset is divided into k=5 subsets and the model is trained and validated k times; each time k-1 subsets are used for training and remaining subset for validation. This method prevents overfitting and ensures that the models' performance is tested across different data subsets. The overall performance is averaged across all fields, and the "accuracy score" was used to assess the model's classification performance.

Additionally, a confusion matrix was employed to provide more detailed insights into model's performance, particularly for minority classes. This is a table that compares the predicted class labels with the actual class labels, showing the number of correct and incorrect predictions for each class.

## III. RESULTS AND DISCUSSION

Figure 1 shows share of each EPC label, which is target variable in ML model, and how using SMOTE and SMOTE-tomek has made it more balanced. In part (A), the imbalanced dataset shows certain EPC labels, such as label C, dominating the distribution, while other labels like F and G are significantly underrepresented. Parts (B) and (C) illustrate the application of SMOTE and SMOTE-Tomek, where synthetic samples have been added to the minority classes (F, G), resulting in a more balanced class distribution. For instance share of case studies with EPC label F has increased from 2% to 16.66% and 17.35% using SOMTE and SMOTE-Tomek, respectively.

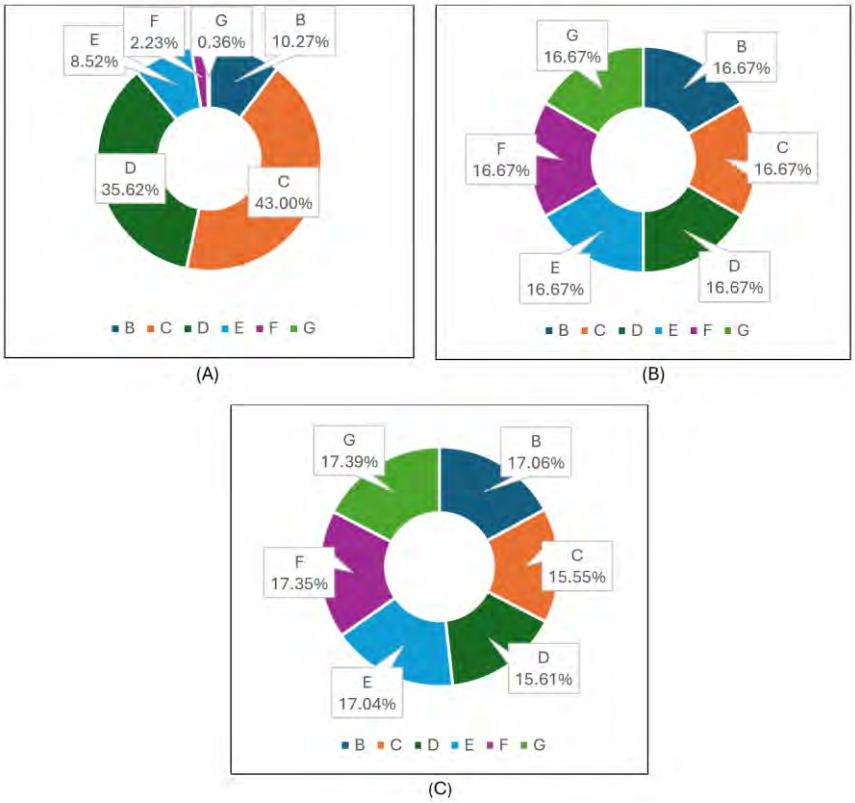


Fig. 1. Effect of resampling techniques on target variable distribution: (A) Imbalanced dataset (B) SMOTE applied (C) SMOTE-Tomek applied

Similarly, before applying resampling techniques, EPC label C dominated the model's target variable. After applying SMOTE and SMOTE-Tomek, its share was reduced to 16.5% and 15.5%, respectively.

Table 2 shows the accuracy scores of the developed model after applying different resampling techniques in comparison to the accuracy obtained on the imbalanced dataset. The model was evaluated using 5-fold cross-validation, and the accuracy score for each fold is shown across the three scenarios: (1) Imbalanced dataset, (2) SMOTE applied, and (3) SMOTE-Tomek applied.

For the imbalanced dataset, accuracy scores ranged from 0.780 to 0.800 across the five folds, with a mean accuracy of 0.792. After applying SMOTE, the accuracy scores slightly decreased across some folds, with the mean accuracy dropping to 0.779. The scores ranged from 0.772 to 0.787, indicating a slight reduction in performance. Similarly, with SMOTE-Tomek, the mean accuracy was 0.780, showing a small improvement over SMOTE but still lower than the accuracy achieved with the imbalanced dataset. Individual fold scores with SMOTE-Tomek varied between 0.769 and 0.789.

However, it is important to note that a slight drop in accuracy does not necessarily indicate a decline in the model's overall performance. In the case of imbalanced datasets, accuracy can be misleading, as the model may primarily fit the majority classes, and most test data points belong to these classes. As a result, further analysis is conducted to assess the model's performance using confusion matrix, as presented in Figure 2.

TABLE II. DEVELOPED MODEL ACCURACY SCORE AFTER APPLYING RESAMPLING TECHNIQUES

| | ***Imbalanced dataset*** | ***SMOTE applied*** | ***SMOTE-Tomek applied*** |
|---|---|---|---|
| Fold-1 | 0.792 | 0.784 | 0.785 |
| Fold-2 | 0.790 | 0.773 | 0.778 |
| Fold-3 | 0.800 | 0.787 | 0.789 |
| Fold-4 | 0.798 | 0.780 | 0.780 |
| Fold-5 | 0.780 | 0.772 | 0.769 |
| Mean | 0.792 | 0.779 | 0.780 |

Figure 2 provides detailed information on the performance of the model in predicting the EPC label classes using confusion matrices for three scenarios: (A) after applying SMOTE, (B) after applying SMOTE-Tomek, and (C) on the imbalanced dataset. Each matrix shows the true labels on the y-axis, the predicted labels on the x-axis, and the number of instances classified correctly or incorrectly.

Before applying any resampling techniques, the model struggles with minority classes, such as labels E, F, and G. For instance, there are many misclassifications in label E, while label C dominating the model, indicating a bias towards majority classes. Minority classes F and G are barely classified correctly, with only 39% and 25% accurately predicted, respectively.

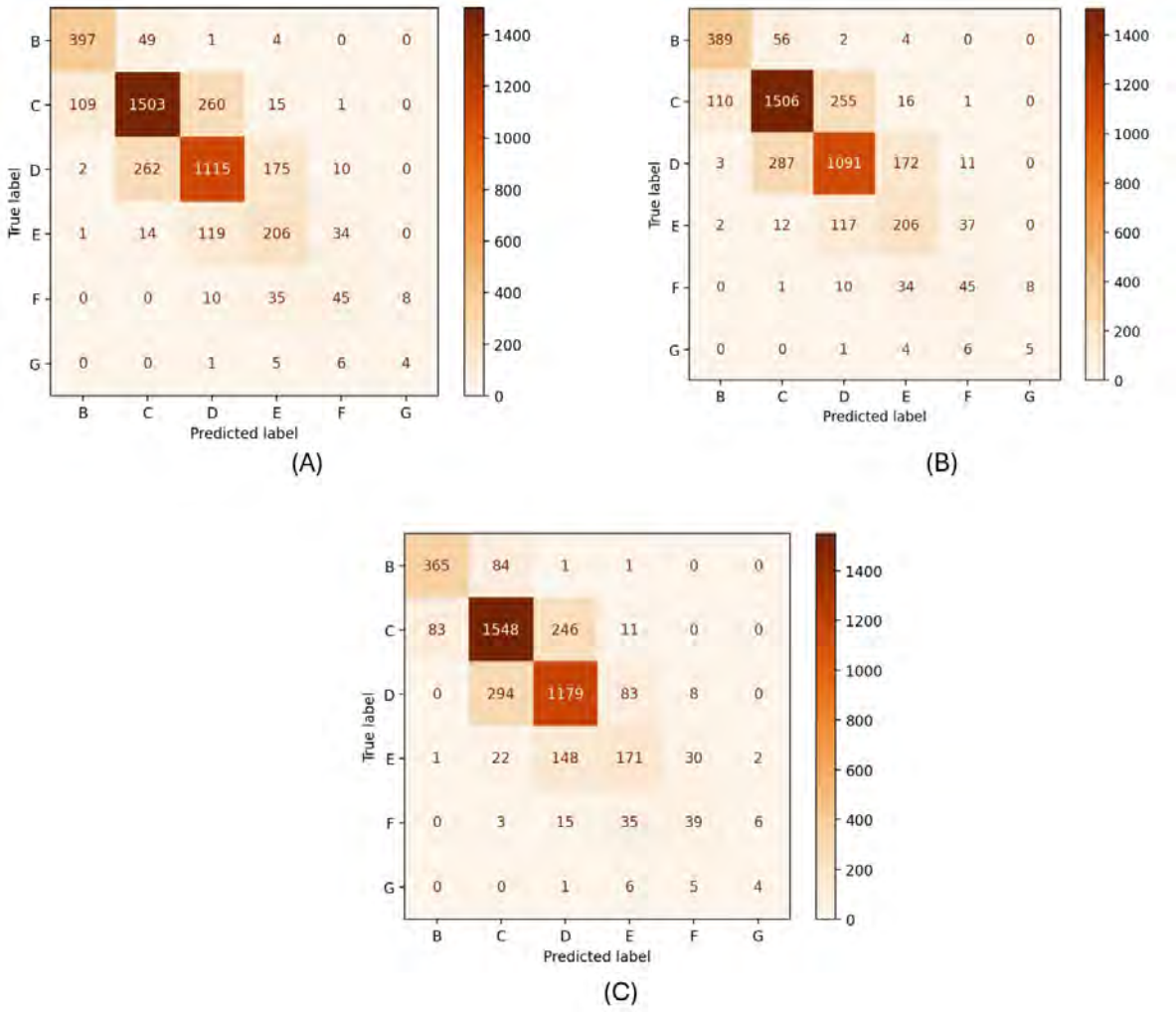


Fig. 2. Details of SMOTE and SMOTE-Tomek effect of on EPC label classes: (A) SMOTE applied, (B) SMOTE-Tomek applied, and (C) Imbalanced dataset

After applying SMOTE, the confusion matrix shows improved classification for minority classes like E and F. The percentage of correctly predicted instances in E increased by 10 points, but there are still some misclassifications, particularly between labels D and C. Nevertheless, the overall balance across the labels is slightly improved.

With SMOTE-Tomek, the model does not demonstrate significant improvement over SMOTE for either minority or majority classes. While the accuracy for label G increased by over 6 points, the model's performance slightly declined in predicting labels B and D.

## IV. Conclusion

This research aimed to address one of the key challenges in applying machine learning models for building energy performance prediction: data imbalance in predicting EPC ratings. The study investigated this issue by implementing two data resampling techniques, SMOTE and SMOTE-Tomek, to improve model accuracy. The results of this study showed that applying resampling techniques significantly improved the prediction accuracy for minority EPC classes (B, E, and F), over 10%, 7%, and 6% points, respectively. Both SMOTE and SMOTE-Tomek yielded relatively similar improvements in accuracy across these minority classes.

These findings suggest that addressing data imbalance through resampling techniques enhances the ML models' accuracy to predict minority EPC labels, which are often underrepresented in datasets. This improvement is crucial for ensuring more reliable predictions, particularly in real-world applications where accurate EPC ratings are essential for energy policy, compliance, and building efficiency assessments. By mitigating the impact of imbalance, these techniques enable more robust ML models that can be integrated into smart systems and AI data-driven tools for building energy performance analysis.

## References

[1] UNEP, "Towards a zero-emissions, efficient and resilient buildings and construction sector," *Global Status Report for Buildings and Construction 2020*, pp. 9–10, 2020, [Online]. Available: https://globalabc.org/sites/default/files/inline-files/2020 Buildings GSR_FULL REPORT.pdf

[2] A. Abbaspour, H. Yousefi, A. Aslani, and Y. Noorollahi, "Economic and Environmental Analysis of Incorporating Geothermal District Heating System Combined with Radiant Floor

Heating for Building Heat Supply in Sarein, Iran Using Building Information Modeling (BIM)," *Energies (Basel)*, vol. 15, no. 23, Dec. 2022, doi: 10.3390/en15238914.

[3] H. Seraj, A. Bahadori-Jahromi, and S. Amirkhani, "Developing a Data-Driven AI Model to Enhance Energy Efficiency in UK Residential Buildings," *Sustainability (Switzerland)* , vol. 16, no. 8, Apr. 2024, doi: 10.3390/su16083151.

[4] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F. Sunmola, and S. Ajayi, "Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques," *Journal of Building Engineering*, vol. 45, Jan. 2022, doi: 10.1016/j.jobe.2021.103406.

[5] R. Olu-Ajayi, H. Alaka, H. Owolabi, L. Akanbi, and S. Ganiyu, "Data-Driven Tools for Building Energy Consumption Prediction: A Review," Mar. 01, 2023, *MDPI*. doi: 10.3390/en16062574.

[6] C. Zhang *et al.*, "Problem of data imbalance in building energy load prediction: Concept, influence, and solution," *Appl Energy*, vol. 297, Sep. 2021, doi: 10.1016/j.apenergy.2021.117139.

[7] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, May 2022, doi: 10.3390/s22093246.

[8] H. & C. Department for Levelling Up, "Energy Performance Certificates and Display Energy Certificates data for buildings in England and Wales," Department for Levelling Up, Housing & Communities. [Online]. Available: https://epc.opendatacommunities.org/

[9] F. Pedregosa, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, p. 2825, 2011.

[10] G. R. Araújo, R. Gomes, P. Ferrão, and M. G. Gomes, "Optimizing building retrofit through data analytics: A study of multi-objective optimization and surrogate models derived from energy performance certificates," *Energy and Built Environment*, no. April, 2023, doi: 10.1016/j.enbenv.2023.07.002.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[12] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *J Intell Inf Syst*, vol. 46, no. 3, pp. 563–597, Jun. 2016, doi: 10.1007/s10844-015-0368-1.

[13] *Proceedings, the 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology : July 7-8, 2020, Bali, Indonesia*. IEEE, 2020.

[14] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data."