Predictive Analytics for Residential Energy Efficiency: Evaluating Machine Learning Models and Key Determinants

Introduction

In pursuit of climate change mitigation, comprehensive understanding and reduction of energy consumption stand as critical imperatives of our time [1-3]. In the United States, residential buildings account for a substantial proportion of greenhouse gas emissions, primarily attributed to their energy consumption [4, 5]. With the prominence of environmental sustainability in global policy discussions, it is critical to pursue the development and assessment of innovative strategies for energy savings and to document their performance [6]. This research aims to examine the application of machine learning techniques in the prediction and improvement of energy efficiency within these buildings, emphasizing the integration of clean energy technologies.

The imperative to address the environmental ramifications of residential energy consumption is underscored by escalating ecological vulnerability [7]. Sophisticated machine learning approaches hold the potential to revolutionize how we understand and predict energy consumption behaviours [8]. The specific objectives of this study are to identify and evaluate the most effective machine learning models for predicting energy consumption in residential environments using clean energy and to find features that have a greater impact on energy consumption, to determine which factors have a greater impact on energy consumption.

To lay the foundation for this investigation, an understanding of the fundamental principles of machine learning and its relevance to data on energy consumption is essential. Machine learning falls under the umbrella of artificial intelligence and provides systems with the capability to learn from data and make informed predictions or decisions [8]. In this framework, Energy Use Intensity (EUI)—which correlates the energy consumed by a building with its size—serves as the primary indicator for evaluating energy efficiency [9].

This study is based on data sourced from the Residential Energy Consumption Survey (RECS), conducted by the Energy Information Administration (EIA) of the U.S. The RECS dataset provides a broad perspective on the energy characteristics of residential buildings in the U.S. The 2015 update expanded the granularity and scope of the dataset, thus augmenting the robustness of the dataset while simultaneously introducing greater complexity in data preprocessing.

The introductory phase of this research involved the application of seven distinct regression and machine learning methodologies to ascertain their predictive accuracies. Linear regression, renowned for its clarity and

interpretability [10], has exhibited substantial precision and reliability. Concurrently, Artificial Neural Networks (ANN) have proven to be effective in clarifying the impact of diverse feature variables on energy consumption patterns. In addition, evaluations of feature importance have revealed the proficiency of the Random Forest algorithm in pinpointing the principal factors that affect energy consumption within residential settings [11].

The knowledge gained from this investigation is anticipated to be instrumental in gauging the success of clean energy implementations in domestic environments. By isolating the most pivotal elements that drive energy consumption, decision-makers—including policymakers, industry representatives, and homeowners—can execute calculated adjustments to optimize energy usage. Such actions are in concert with the wider goals of advancing environmental sustainability.

Methodology

Raw Data Summary

The 2015 Residential Energy Consumption Survey (RECS) is a major survey conducted by the U.S. Energy Information Administration (EIA) to collect and analyze data on national household energy use and consumption in the U.S. The survey collected residential energy consumption from over 5,600 housing units at random using a complex multistage, area-probability sample design, representing 118.2 million U.S. households. The survey covered data 758 features, including the physical characteristics, geographical location and climate conditions of the houses, the types of equipment and occupancy patterns of the houses, energy expenditure data, and clean and dirty energy usage habits (Table 1).

Data Pre-processing

However, there is much data not relevant to our topic and flawed variables in the data. Also, some features in the dataset are not relevant to the topic of energy consumption analysis, or some features themself shows direct relation with the target prediction value which could affect the accuracy of our models. Hence, a pre-processing is needed for the dataset before analyzing.

One widely used metric to assess building energy performance in the U.S. is Energy Use Intensity (EUI) [12], which is calculated by dividing the total Btu used in a residential house by the total square footage of the house. Hence, we used EUI generated from the primary clean energy source as the response variable for the analysis and created a new column in the dataset to store the EUI data.

After obtaining a list of targeted energy-related variables, we conducted a multi-step screening process under the instruction of the Codebook[9]. Initially, we handled variables that were either unlabeled or non-numeric, rendering

them unsuitable for regression analysis. Subsequently, we referred to the features USENG, USELP, USEFO, USEWOOD and deleted all the label 0 in USENG and all the label 1 in the other three features, reducing the size of the dataset to around 3,000 houses. Also, based on whether heating equipment is used, we deleted around 1,500 variables since heaters are a major source of energy. Next, to mitigate the impact on the fitting performance, we excluded 169 features that showed excessively high correlations with the target variable. Meanwhile, 97 variables which showed no significant relevance to the analytical objective were also removed. Additionally, we computed the variance of each feature and excluded 131 variables with variance below 5%. Moreover, we identified variables in the dataset with excessively large variance, often containing missing or abnormal data. To ensure both the accuracy and quality of the data, for the more significant features, we removed merely the samples with abnormal data, while for less significant ones, we excluded the entire variable, hence 358 features and 987 samples remained. Finally, we subjected the coarsely processed dataset to LASSO for regression analysis of importance, thereby eliminating a substantial number of variables with lower importance, resulting in a final set of 19 continuous variables and 52 categorical variables. The purpose of this data processing procedure is to select variables that are properly related to the response variable while retaining an adequate sample size to prevent underfitting and overfitting, thus obtain the most accurate regression analysis results.

Analytical methods for regression

This project uses several methods applicable to regression problems to construct models for prediction.

1. Linear regression

Linear regression is a common method for regression with the function: $y = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r + \varepsilon$. It uses a linear combination of a dependent variable and a set of independent variables to explain the relationship between the latter.

2. Lasso Regression

The least absolute shrinkage and selection operator (LASSO) is a penalized regression which applied as an extension for linear regression. The formula for estimating weight is: $\hat{\beta}^{lasso} = arg \min_{\beta \in \mathbb{R}^p} ||y - Z\beta||^2 + \lambda \sum_{i=1}^{r} |\beta_i|$. The tuning parameter λ controls the strength of the penalty. When it tends to infinity, $\beta lasso$tends to zero and means the weak effect of the corresponding variable. As mentioned before, we excluded features with a weight of 0 to optimize model performance in the final step of data cleaning.

3. Linear support vector machine

Support Vector Machine (SVM) is a machine learning algorithm that aims to find the boundaries defined by some

vectors that describe the data. In regression problems, it aims to find the optimal generalization bounds for regression [9]. Our project uses linear support vector machine as a machine learning method.

4. Kernel support vector machine

To determine the nonlinear boundaries and thus improve the accuracy of the prediction, SVM can be applied with different kernel functions to map the input variables to a high-dimensional feature space [9]. Kernel SVM with Gaussian kernels applied provides better predictions [13], and therefore is also used in this project: $K\ (x_i, x_j) = exp\ (-\frac{||x_i - x_j||^2}{2\sigma^2})$.

5. Gradient boosting decision tree

Some decision tree methods have a good performance on regression problems. The idea of the algorithm is to successively split the classes of predicted values until it is obtained.

Gradient boosting decision tree is a tree ensemble method. Each tree is constructed based on the performance of the previous tree, achieving a higher model fit.

6. Random forest

Random forest is another tree integration method that contains many parallel trees. The best tree is selected by comparing the performance of each tree and its prediction is obtained as the result.

7. ANN

The structure of ANN is divided into input layer, hidden layer, and output layer. Each node in the input layer represents a feature or attribute of the input data based on the linear combination of the input data$f(x) = W \cdot x + b$ where **W** represents the weight and **b** represents the bias for each neuron. The weighted input will then be processed by the non-linear activation functions in the hidden layers for regression [14]. Finally, it is passed to the output layer and a continuous outcome will be displayed. After an epoch, the model will conduct gradient descent optimization. By evaluating the partial derivative of the loss function with respect to each node, the weights would be adjusted, and the accuracy of the model would be improved. Owing to the suitability for complex nonlinear relationship modeling and strong flexibility and approximation capabilities, it can handle diverse regression analysis problems.

Analytical methods for the importance of features

A more precise evaluation of variable importance can be applied in Random Forests [15]. The principle is to calculate how much each feature contributes to each tree in the forest, and then take an average to compare the contributions between features. The evaluation indicator for measuring contribution is the change in the Gini index,

which represents the probability that two randomly selected samples will have inconsistent category labeling at a node: $Gini(t) = 1 - \sum_{i=1}^{c}[p(i|t)]^2$. Importance of feature j at node q of the tree i is the change in Gini index before and after branches [16]: $I_{jq}^{(Gini)(i)} = Gini_q^{(i)} - Gini_{left}^{(i)} - Gini_{right}^{(i)}$. The variable importance measures (VIM) for feature j is the sum of the importance of the nodes in every tree of the forest after normalization [16]: $VIM_j^{(Gini)} = \frac{\sum_{i=1}^{I}\sum_{q\in Q} I_{jq}^{(Gini)(i)}}{\sum_{j'}^{J}\sum_{i=1}^{I}\sum_{q\in Q} I_{j'q}^{(Gini)(i)}}$.

Evaluation methods

The project used cross-validation to assess the model performance of analysis methods, which ensures all records have been used for training and testing. K- fold cross-validation is used for this project and the K is 5. For the test of each model, the data is first divided into 5 partitions. During 5 repetitions, one partition will be used as a test set and the rest will be used as a training set.

The assessment of model accuracy is based on the values of absolute error (MAE): $MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - \hat{X}_i|$, and mean square error (MSE): $MSE = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_i)^2}{n}$. MSE and MAE will be calculated for each training and test set, which helps to analyze the stability and fit of the model.

Results

In our meticulous analysis, a septet of machine learning algorithms was meticulously applied to a dataset with the objective of prognosticating outcomes utilizing a duo of evaluative metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE) (Table 2, Figure 1, Figure 2). These metrics quantitatively assess the fidelity of predictions rendered by the models. The algorithms were dichotomized into two categories predicated on their mathematical underpinnings: linear algorithms (inclusive of Linear Regression, Lasso Regression, and Linear SVM) (Figure 1) and nonlinear algorithms (comprising Kernel SVM, Boosting Tree, Random Forest, and Artificial Neural Network) (Figure 2).

The efficacy of linear algorithms underwent a rigorous examination initially. Both Linear Regression and Lasso Regression yielded results in proximity about MAE and MSE. This phenomenon leads to the postulation that the regularization wrought by Lasso Regression exerts a marginal influence on the outcomes for the dataset under scrutiny. In contrast, Linear SVM manifested a discernible decline in performance—a phenomenon posited to be a consequence of Linear SVM's augmented sensitivity to the calibration of hyperparameters, as corroborated by extant scholarly discussions. A scrupulous cross-validation procedure was implemented to verify the stability of these

models. This procedure unveiled a laudable consistency in the predictive precision across the training and validation phases, signalling a formidable defence against the pitfalls of both overfitting and underfitting.

Conversely, a different narrative was painted by the nonlinear algorithms. Kernel SVM, Boosting Tree, and Random Forest exhibited adeptness in fitting the training data with high precision, but when their predictions were tested against the test set, they succumbed to increased MAE and MSE values—a clear indication of overfitting. Kernel SVM was particularly prone to fluctuations in stability, more so than Random Forest. While the Boosting Tree algorithm maintained some level of stability and accuracy, the threat of overfitting persisted, although it was not entirely unconstrained. The Artificial Neural Network, distinguished by a distinctive architecture that included two hidden layers, stood out as the single nonlinear model that successfully circumvented the common issue of overfitting. Despite this, it registered as the least precise and stable, a shortcoming that intimates the possibility of improvement through an augmentation in the number of hidden layers.

Upon comparison of the models, it became conspicuously clear that linear algorithms outstripped their nonlinear counterparts in melding stability with precision for the datasets employed in this investigation (Figure 1, Figure 2). Linear Regression and Lasso Regression were heralded for their consummate performance, which was in tight competition. Nonetheless, owing to Linear Regression's more streamlined constitution and diminished computational requisites, it was adjudged the quintessential machine learning stratagem for this scholarly exploration.

In the quest to elucidate the variables that significantly influence Energy Use Intensity (EUI), the Random Forest algorithm, renowned for its precision in assessing variable importance, was engaged. This algorithm's application provided deep insights into which variables most heavily bear upon EUI (Figure 3). The size of the domicile stood out as the preeminent factor, accounting for an estimated 35% of the model's predictive accuracy. Additional variables of consequence included the area of the domicile, the utilization of an oven, and the ambient temperature at night, with each contributing to the model's efficacy within a 4% to 5% range (Figure 3). The Random Forest model's robustness underwent validation via a rigorous cross-validation methodology, ensuring that the consistency of variable importance metrics across multiple iterations stands as a robust indicator of the model’s stability[17]. The graphical figures provided herein encapsulate these metrics. Articulating the significance of these critical variables holds immense value, for it lays the foundation for imminent scholarly endeavours aimed at propelling the design of edifices towards greater energy efficiency, harmoniously aligning with clean energy imperatives.

Discussion

In this investigation, empirical evidence strongly suggests that Linear Regression constitutes the superlative machine learning paradigm for our dataset, outperforming its non-linear counterparts substantially in terms of model stability, predictive accuracy, and goodness of fit. It is inferred that the dataset under examination predominantly exhibits linear correlations.

Through a meticulous examination of diverse machine learning methodologies, a spectrum of inherent strengths and weaknesses was brought to light. Linear Regression, while efficacious in its domain, is intrinsically limited by suitability that is exclusive to datasets with linear characteristics, revealing its constraints despite commendable performance. In contrast, Kernel SVM, with its adaptability and increased complexity, shows a performance acutely sensitive to hyperparameter tuning, a susceptibility that can lead to suboptimal results, especially when compared with Linear Regression. The performance of non-linear algorithms, save for Artificial Neural Networks (ANNs), was exceptional in the training phase but diminished in the testing phase, suggestive of an inclination towards overfitting and model instability.

Artificial Neural Networks, proficient in averting the overfitting dilemma through their complex network architectures, nonetheless demand considerable computational investment. The performance of ANNs is critically dependent on the depth of their architecture; a dearth of hidden layers can lead to subpar performance metrics.

The role of dataset preprocessing and cross-validation asserted itself as pivotal throughout our algorithmic exploration. The initial phase of preprocessing uncovered a tendency towards overfitting, particularly when the selection of the dependent variable was not sufficiently rigorous, a complication often compounded by the ubiquitous presence of derivative variables in publicly available datasets. Furthermore, the accuracy of the models was significantly influenced by the occurrence of missing values. Cross-validation techniques were introduced in response to the observed fluctuations in model performance attributable to variations in the testing subset. These techniques not only buttressed the assessment of model stability but also affirmed the integrity of the fit. In the realm of comparative analysis, cross-validation proved to be a quintessential tool, enabling a quantitative juxtaposition of the stability and accuracy across the seven distinct machine learning methods we evaluated.

Conclusion

In conclusion, this study evaluated the performance of seven distinct machine learning models in predicting the energy consumption of residential buildings powered by clean energy. The assessment revealed that linear regression models outperformed artificial neural networks in terms of error rates and consistency. Utilizing Random Forest algorithms for feature importance analysis, our research corroborated the predominant influence of a

building's total square footage on Energy Use Intensity (EUI), with heating degree days and geographic location also being significant. The oven emerged as a notable electrical appliance impacting energy use. Conversely, the influence of building materials and household size on energy consumption was comparatively marginal.

The implications of our findings suggest a strategic focus on enhancing heating efficiency and optimizing the use of key electrical appliances, particularly ovens, to augment clean energy efficiency in residential buildings. These recommendations could inform targeted interventions to leverage clean energy more effectively, contributing to the broader effort of combating climate change. Future research should consider these insights to refine energy consumption models and support the sustainable use of clean energy resources.

Appendix

Table 1. Variables after Pre-processing

Abbreviation	**Full name**
ADQINSUL	Level of insulation
AGEFRZR	Age of most-used freezer
AGERFRI2	Age of second most-used refrigerator
COMBODVR	Number of cable or satellite boxes with DVR
DESKTOP	Number of desktop computers
DIVISION	Census Division
DOOR1SUM	Number of sliding glass doors
DRAFTY	Frequency of draft
ELWARM	Electricity used for space heating
ENERGYASSTOTH	Received home energy assistance in some other year
EQUIPMUSE	Main heating equipment household behavior
ESLIGHT	Energy Star qualified lightbulbs
FOODPROC	Food processor used
FUELAUX	Secondary space heating fuel
FUELH2O	Fuel used by main water heater
FUELH2O2	Fuel used by secondary water heater
FUELHEAT	Main space heating fuel
FUELTUB	Fuel used for heating hot tub
GNDHDD65	Heating degree days of ground temperature in 2015, base temperature 65F
HDD65	Heating degree days in 2015, base temperature 65F
ICE	Through-the-door ice on most-used refrigerator
METROMICRO	Housing unit in Census Metropolitan Statistical Area or Micropolitan Statistical Area
MICRO	Microwave oven used
MONTUB	Months hot tub used in the last year

NHAFBATH	Number of half bathrooms
NHSLDMEM	Number of household members
NUMATTICFAN	Number of attic fans used
NUMFLOORFAN	Number of floor, window, or table fans used
NUMMEAL	Frequency hot meals are cooked
NUMTABLET	Number of tablet computers or e-readers
OUTGRILL	Outdoor grill used
OVENFUEL	Fuel used by separate oven
OVENUSE	Frequency of use of oven part of stove
PAYHELP	Received energy assistance to help pay energy bills after disconnect notice
PERIODNG	Number of days covered by Energy Supplier Survey natural gas billing data and used to calculate annual consumption and expenditures
POOL	Heated swimming pool
ROOFTYPE	Major roofing material
SCALEB	Frequency of reducing or forgoing basic necessities due to home energy bill
SEPOVENUSE	Frequency of separate oven use
STORIES	Number of stories in a single-family home
STOVEFUEL	Fuel used by separate cooktop
TEMPHOME	Winter temperature when someone is home during the day
TEMPNITE	Winter temperature at night
THERMAIN	Any thermostats
TOASTOVN	Toaster oven used
TOTSQFT_EN	Total square footage (used for publication)
TYPEHUQ	Type of housing unit
UATYP10	Census 2010 Urban Type
UGOTH	Natural gas used, other than for space heating, water heating, or cooking
WINDOWS	Number of windows
WINFRAME	Window frame material
WWACAGE	Age of most-used individual air conditioning unit
YEARMADERANGE	Range when housing unit was built
ZAGECENAC	Imputation flag for AGECENAC
ZBASEHEAT	Imputation flag for BASEHEAT
ZDWASHUSE	Imputation flag for DWASHUSE
ZFUELPOOL	Imputation flag for FUELPOOL
ZHOUSEHOLDER_RACE	Imputation flag for HOUSEHOLDER_RACE
ZMORETHAN1H2O	Imputation flag for MORETHAN1H2O
ZNOHEATHELP	Imputation flag for NOHEATHELP
ZSEPDVR	Imputation flag for SEPDVR
ZTEMPGONE	Imputation flag for TEMPGONE
ZTEMPHOMEAC	Imputation flag for TEMPHOMEAC
ZTOTSQFT_EN	Imputation flag for TOTSQFT_EN

ZTVONWE1	Imputation flag for TVONWE1
ZTYPERFR2	Imputation flag for TYPERFR2
ZUPRTFRZR	Imputation flag for UPRTFRZR
ZUSENOTMOIST	Imputation flag for USENOTMOIST
ZWINFRAME	Imputation flag for WINFRAME
ZWWACAGE	Imputation flag for WWACAGE

Table 2. MAE of Different Methods

MAE on Total EUI (kBtu/ft2)	**Training Set**	**Testing Set**
Linear Regression	0.46175±0.00589	0.51586±0.01778
Lasso Regression	0.45934±0.00639	0.50799±0.02144
Linear SVM	0.44433±0.00763	0.51188±0.03244
Kernal SVM	0.09602±0.00018	0.73101±0.03263
Boosting Tree	0.18320±0.00235	0.06808±0.00302
Random Forest	0.10920±0.00436	0.02122±0.00210
ANN	0.63501±0.12337	0.81968±0.19656

Table 3. MSE of Different Methods

MSE on Total EUI (kBtu/ft2)	**Training Set**	**Testing Set**
Linear Regression	0.40272±0.00974	0.50119±0.03937
Lasso Regression	0.40662±0.00977	0.49524±0.04147
Linear SVM	0.44606±0.00906	0.53204±0.07263
Kernal SVM	0.00946±2.36009e-05	0.97403±0.10960
Boosting Tree	0.49774±0.03932	0.49386±0.07618
Random Forest	0.48469±0.01924	0.46081±0.02926
ANN	0.64296±0.14326	0.85177±0.32543

Figure 1. Linear Methods MSE&MAE

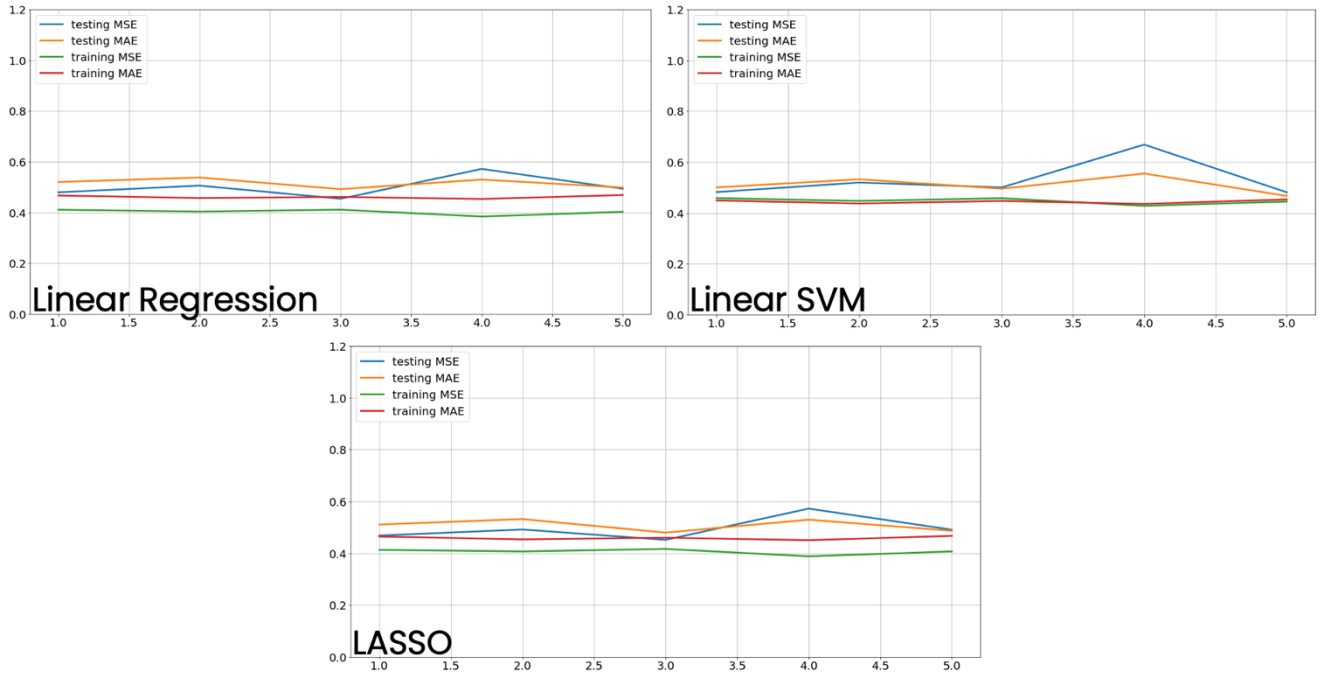


Figure 2. Non-Linear Methods MSE&MAE

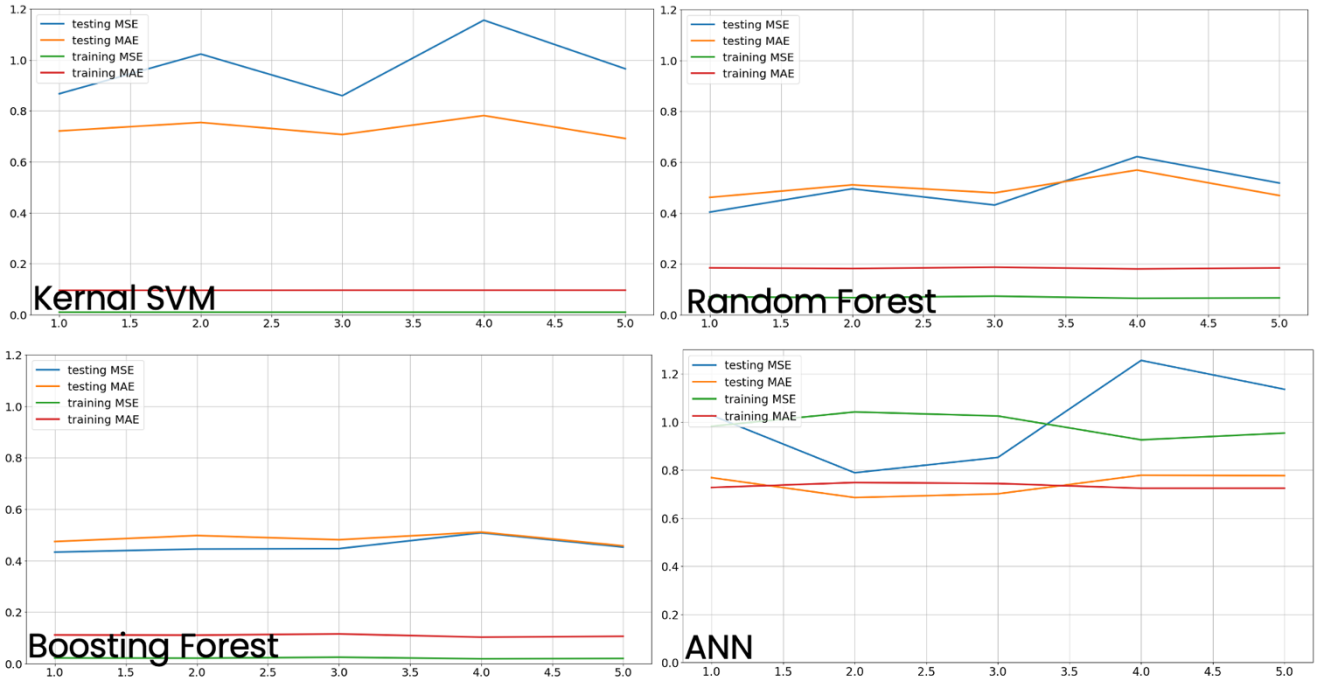
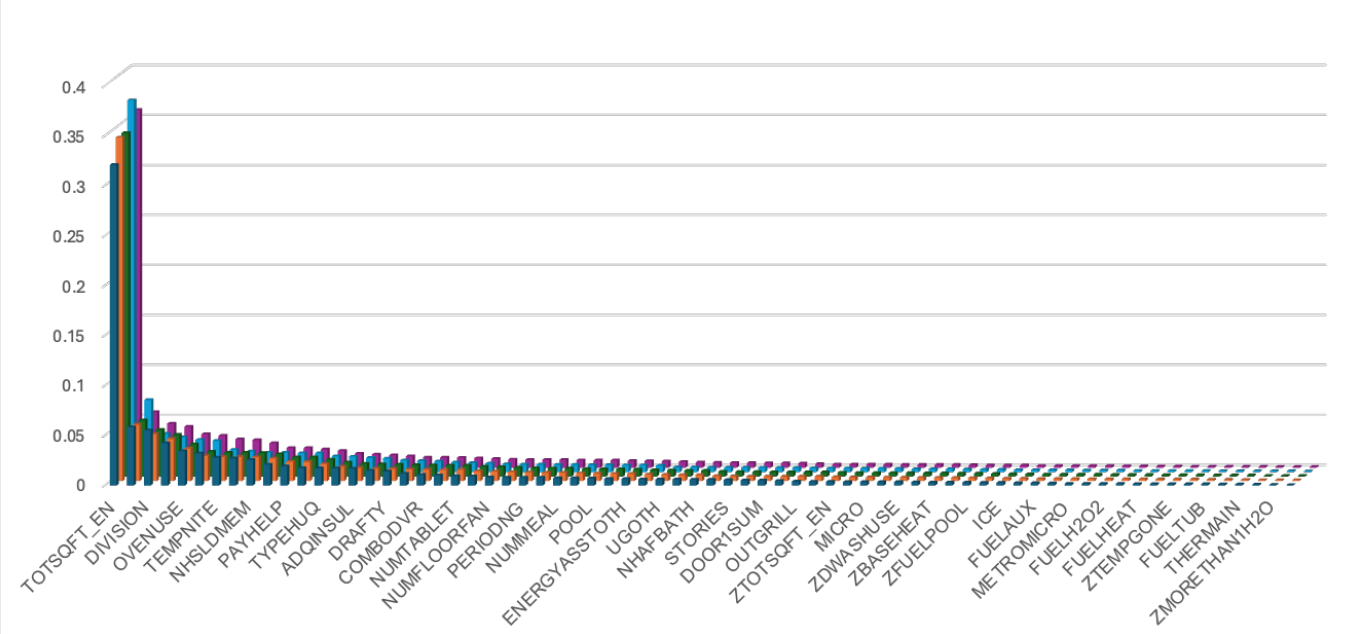


Figure 3. Weight Result



References

1. Olabi, A. and M.A. Abdelkareem, *Renewable energy and climate change.* Renewable and Sustainable Energy Reviews, 2022. **158**: p. 112111.
2. Pryor, S.C. and R.J. Barthelmie, *Climate change impacts on wind energy: A review.* Renewable and sustainable energy reviews, 2010. **14**(1): p. 430-437.
3. Sims, R.E., *Renewable energy: a response to climate change.* Solar energy, 2004. **76**(1-3): p. 9-17.
4. Rockett, P. and E.A. Hathway, *Model-predictive control for non-domestic buildings: a critical review and prospects.* Building Research & Information, 2017. **45**(5): p. 556-571.
5. O'Neill, B.C. and M. Desai, *Accuracy of past projections of US energy consumption.* Energy Policy, 2005. **33**(8): p. 979-993.
6. Howes, M., et al., *Environmental sustainability: a case of policy implementation failure?* Sustainability, 2017. **9**(2): p. 165.
7. Willow, A.J., *The new politics of environmental degradation: un/expected landscapes of disempowerment and vulnerability.* Journal of political Ecology, 2014. **21**(1): p. 237-257.
8. Al-Garadi, M.A., et al., *A survey of machine and deep learning methods for internet of things (IoT) security.* IEEE communications surveys & tutorials, 2020. **22**(3): p. 1646-1685.
9. Deng, H., D. Fannon, and M.J. Eckelman, *Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata.* Energy and Buildings, 2018. **163**: p. 34-43.
10. Kim, H., et al., *Visualizable and interpretable regression models with good prediction power.* IIE Transactions, 2007. **39**(6): p. 565-579.
11. Shine, P., et al., *Machine-learning algorithms for predicting on-farm direct water and electricity consumption on pasture based dairy farms.* Computers and electronics in agriculture, 2018. **150**: p. 74-87.
12. Chung, W., Y. Hui, and Y.M. Lam, *Benchmarking the energy efficiency of commercial buildings.* Applied energy, 2006. **83**(1): p. 1-14.
13. Wang, H. and D. Hu. *Comparison of SVM and LS-SVM for regression*. in *2005 International conference on neural networks and brain*. 2005. IEEE.
14. Wu, W., G.C. Dandy, and H.R. Maier, *Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling.* Environmental Modelling & Software, 2014. **54**: p. 108-127.
15. Chan, J.C.-W. and D. Paelinckx, *Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery.* Remote Sensing of Environment,

2008. **112**(6): p. 2999-3011.
16. Raschka, S., *Python machine learning*. 2015: Packt publishing ltd.
17. Houborg, R. and M.F. McCabe, *A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning.* ISPRS Journal of Photogrammetry and Remote Sensing, 2018. **135**: p. 173-188.