# Commercial Building Energy Optimization A Machine Learning Approach

Aryama Ray
*MS in Data Analytics*
*San Jose State University*
San Jose, Califonia
aryama.ray@sjsu.edu

Ananya Varma Mudunuri
*MS in Data Analytics*
*San Jose State University*
San Jose, Califonia
ananyavarma.mudunuri@sjsu.edu

Shikha Singh
*MS in Data Analytics*
*San Jose State University*
San Jose, Califonia
shikha.singh@sjsu.edu

Venkata Sai Sreelekha Gollu
*MS in Data Analytics*
*San Jose State University*
San Jose, Califonia
venkatasaisreelekha.gollu@sjsu.edu

***Abstract*—Huge energy consumption is one of the many global concerns regarding the environmental impact, this project aims for the optimization of energy management in buildings, to be specific by using an archive dataset that has data collected from over 300 sensors for three years. This data consists of information about indoor environment, outdoor environment, and HVC.The data plays an important role in validating the model's predictions. A combination of linear and nonlinear machine learning models like Multiple Linear Regression, Random Forest, K Nearest Neighbors and Gradient Boosting algorithms are used to carry out this project. This project involves using the already extracted and cleaned data, using the right models to train and evaluate, through which we get important insights that contribute to sustainable building methods. The project aims not just for efficient environmental usage but also to understand the relationship between the energy consumption by buildings and their impact on the environment by accurately predicting the HVAC cooling and heating load. In implementing advanced machine learning methods, the project hopes to make a sustainable statement about the environment, hoping to make the future healthier and greener. The output expected includes predicting HVAC cooling and heating load based on collected indoor and outdoor environment data.**

## I. Introduction

The ever-increasing significance of environmental conservation and sustainability is otherwise a pressing concern in the present-day world. Having understood this importance, we are highly motivated to pursue this project where we target to address inefficiencies in energy management in building operations. Quite notably, the total energy consumption worldwide is significantly contributed by various buildings. This has a direct impact on greenhouse gas emissions. In this project, we aim to develop various insightful models that optimize energy usage in addition to contributing towards a sustainable future using a three-year building operational performance dataset. We believe this project can drive positive changes, which aligns with our values, commitment, and responsibility towards the environment. By considering our knowledge in data analysis and machine learning, we want to handle real-world problems that help a more sustainable future for the upcoming generations. We aim, through this project, our meaningful contribution to encourage data-driven insights for a greener, healthier, and more prosperous tomorrow.

## II. Literature Survey

### A. Maintaining the Integrity of the Specifications

Buildings consume about 40% of the primary energy in the United States and about one-third globally. With modern technologies like temperature sensors and other energy efficiency tools, we can achieve an energy use reduction of about 50%. The dataset includes whole-building and end-use energy consumption, HVAC load, and indoor and outdoor environmental factors. The data was collected for three years from more than 300 sensors and meters on two office floors of the building [1]. Global concerns are growing about energy waste and its adverse impact on the environment. When designing efficient buildings, it is inevitable to calculate their cooling load (CL) and heating load (HL) to specify the required cooling and heating equipment to achieve comfortable indoor air conditions. The environmental characteristics of a building are among the main aspects or conditions that can affect its energy consumption, i.e. contribute to sustainability and energy efficiency [2]. There are several methods available for predicting energy consumption in a commercial building. Multiple Linear regression is a valid approach due to its effectiveness and linearity. Decision trees, neural networks, and linear regression gave good performance with minimal difference in error [3]. To promote building energy efficiency and motivate people to adopt sustainable practices, buildings are assigned ratings and labels [4]. Support Vector machines are highly robust machine learning models. Support Vector Machines (SVM) are used for both classification and regression purposes. For buildings, SVM has been used for forecasting energy consumption [5] and classification of energy usage of the buildings [6]. The main benefit of using random forest as a predictive model is its high generalization ability, which can effectively prevent overfitting in data prediction [7]. Statistical properties of the variables can be first analyzed with visualization such as a histogram of the empirical probability distributions of all the input and output variables e.g. Heating and cooling load [8]. The scope of this project covers the course curriculum covered in the class. The project includes the implementation of supervised machine learning algorithms such as Multiple Linear Regression, Random Forest, Support Vector Machines,

and Naïve Bayes. The project will also cover the probability distribution of the variables.

## III. Methodology

The goal of smart building lies in minimizing the energy consumption by the end user. This project aims at optimizing energy consumption by accurately predicting the HVAC cooling and heating load [9]. The proposed statistical techniques and traditional machine learning algorithms to predict HVAC cooling and heating load are based on collected indoor and outdoor environment data. The hypothesis for this project is that outdoor temperature, humidity, solar radiation indoor CO2 concentration, heating and cooling temperature set points, and interior and exterior zone temperature have an impact on HVAC energy load. Different statistical experiments are designed to identify the features with high predictive power. This will contribute towards achieving high-performance machine learning by adjusting parameters. SelectKBest and mutual information regression are used in this project to assess the relationship between feature variables and analyze the predictive power of the feature variables, respectively. Fig 1 shows the implemented methodology here using a supervised machine learning algorithm.

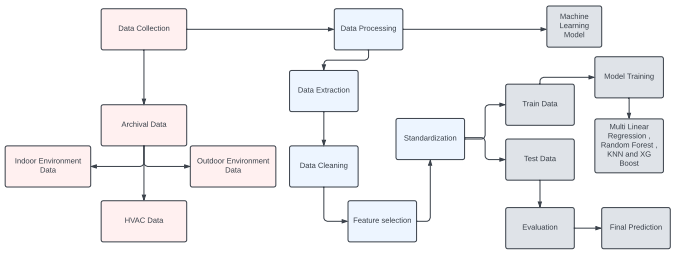


Fig. 1. The process of the methodology.

The overall methodology consists of three stages including 1) data collection 2) data processing and normalization and 3) supervised machine learning model building and prediction of HVAC cooling and heating load attributes. Data is collected from more than 300 sensors and meters installed in a two-floor office building in Berkeley, California [10]. Data Processing steps involve dimensionality reduction technique, data cleaning, and normalization. Multiple traditional machine learning techniques are used in this project such as Multiple linear regression, Random Forest, KNN, and Gradient Boosting. The multiple linear regression technique is extensively used when there exists more than one predictor variable with continuous values. Random forest regression is well-known to capture non-linear relationships between feature variables while increasing generalization ability. The gradient boosting algorithm is another powerful regression technique where multiple weak models are created to create a strong model. The performance of these models is evaluated using cross-validation techniques and using $R^2$, RMSE evaluation metrics. Hypertuning of the models through Grid-search techniques often results in an optimized solution. Comparison of evaluation metrics among multiple modeling approaches leads toward the most optimized solution.

## IV. Data

This data set was collected from an office building in the year 2015 in Berkeley, California. It consists of information regarding the whole building and end-use energy consumption, the HVAC system operating conditions, indoor, and outdoor environmental parameters, and occupant count. The data is collected for three years from over 300 sensors and meters for two office floors (each 2,325 m²) of the building. The data-to-research-grade data transformation followed is summarized in the three-step data curation strategy: outlier value and data gap identification and adjustment in the raw data; model of the building systems and data points' metadata using the Brick schema; description of the metadata of the dataset with a semantic JSON schema. This dataset has the flexibility to be used for various application types, among them building energy benchmarking, load shape analysis, energy and occupancy forecasting, and HVAC control [1]. This data consists of outdoor environmental data, indoor environmental data, and target variables regarding energy usage. The datasets are important in the development of prediction models for the HVAC (Heating, Ventilation, and Air Conditioning) load.

### A. Outdoor Environmental Data

This data contains the from the following outdoor-sited sensors, captured in file `site_weather.csv`: air_temp_set_1: The sensor will measure the outside air temperature at the. air_temp_set_2: External air temperature measured by sensor 2. dew_point_temperature: The dew point temperature of the second sensor indicates how humid the air is.

relative_humidity_set_1: Set relative humidity from the set of sensors 1. Solar_radiation_set_1 is the set of data that describes the solar radiation recorded from sensor 1. These are very highly important features because of the contribution they make to indoor temperatures hence solar gain.

### B. Indoor Environmental Data

The data set comprises the temperature set points and other internal conditions of the building stored in separate files. `zone_temp_sp_c.csv`, `zone_temp_sp_h.csv`: These files contain the setpoints for the cooling temperature and heating temperature, respectively, for various zones of a building. Their file names correspondingly are zone_cooling_sp and zone_heating_sp. `zone_temp_interior.csv`: Temp readings from internal zones. The file `zone_temp_exterior.csv` provided

the required temperatures in the respective exterior zone (zone_*_temp) to infer the conditions of the boundary that affect the HVAC needs of the building. `zone_co2.csv`: CO2 concentration within individual zones (zone_*co2) can impact air quality.

### C. Target variable:ele.csv

Data specific for Energy Usage by HVAC systems: South Wing and North Wing. hvac_S is for the South Wing and hvac_N is for the North Wing. It is an important variable to be realized and taken care of at the time of modeling and forecasting energy demand.

## V. Exploratory Data Analysis

### A. Data Preparation

Multiple datasets from CSV files are loaded into the Panda data frame. Basically, the datasets loaded are energy-used data, indoor environmental data, outdoor environmental data, CO2 concentrations, occupancy data, and WiFi data, all of which are essentially used in understanding building dynamics. Basic commands, such as head () and info (), can be used to notify the user of the construction of the data and the existence of missing values. Multiple datasets from CSV files are loaded into the Panda data frame. Basically, the datasets loaded are energy-used data, indoor environmental data, outdoor environmental data, CO2 concentrations, occupancy data, and WiFi data, all of which are essentially used in understanding building dynamics. Basic commands, such as head () and info (), can be used to notify the user of the construction of the data and the existence of missing values.

*1) Data Cleaning:* Date Format Standardization: Since the input files had different date formats, to allow the joining from timestamps, a common date format (% Y-% m-% d % H: % M:% S ) was applied over the datasets to allow join keys.

Missing Data Handling: Removed all fields that have the string 'unnamed' in them and play important roles in holding nan's, in the datasets regarding the cooling and heating set points. Iterative imputers are used to impute missing values, which is effective when a bunch of variables are interacting with each other in this manner. In this way, the missing values are estimated without overfitting by the Bayesian ridge regression type of regularized linear regression.

## VI. Feature selection and Data Transformation

### A. Dimensionality Reduction:

In this project, we then used Principal Component Analysis (PCA) to determine how many features could be used to pick a subset for the purpose of model building. Our merged data set with 168 columns is in a high-dimensional feature space that might be a limitation in achieving model accuracy. That is when we needed to evaluate the minimum number of features that would take care of data explanation before proceeding. Our results with PCA show that close to 100% data interpretability was achieved with about 40 components. Note, however, that in general, PCA is a very useful technique for dimensionality reduction but not quite useful for feature selection in a numerical dataset. In our case, it led to a loss of dataset interpretability, indicating that the interpretability of the original feature set was not kept by the technique.

### B. Feature selection:

SelectKBest is one of the univariate feature selection algorithms that pick the top k influential features according to a defined scoring function. For feature selection, we performed the selectKbest univariate feature selection algorithm. This project considered two different scoring methodologies, including f_regression and mutual information. We have chosen these methodologies as they give an idea about the strength and relevance of the relationship of the variables to the target variables. The variables used to perform feature selection were independent of each target variable, hvac_S, and hvac_N, to fine-tune the feature set specifically to the predictive needs of each HVAC system wing.

*1) :* The scoring of the f_regression is a two-step means by which the scoring used for the prediction of the feature's importance is done. It is important for the selection of the feature components in regression models and is done by the following:

Calculate cross-correlation: The procedure starts with the testing of the relationship of each of the predictor variables to the target variable. This would be done by checking how much each feature and target variable move together, hence providing a linear indication. This is an important step in that it quantifies the strength of the association of each feature individually with the outcome of interest.

$$r_{\text{regression}} = \frac{E\left[(X[:,i] - \text{mean}(X[:,i])) \cdot (y - \text{mean}(y))\right]}{\text{std}(X[:,i]) \cdot \text{std}(y)} \tag{1}$$

where $X[:,i]$ is the $i$-th regressor, $y$ is the target variable, $\text{mean}(X[:,i])$ and $\text{mean}(y)$ are the means of the $i$-th regressor and target variable respectively, and $\text{std}(X[:,i])$ and $\text{std}(y)$ are their standard deviations.

Transformation to F-score and p-value: The correlation is then transformed to F-scores. The F-scores can be applied in making statistical determinations regarding the significance of the correlations observed. P-values are then obtained from the F-scores, showing the measure of the probability that the observed correlations could occur by chance.

The method ranks the features based on a linear relationship between the feature and the target variable. Out of these, the positive linear relationships are given more importance, as they stand more chances of affecting the target variable considerably. In this way, the method ensures that the selected features derive the maximum influence on the predictive accuracy of the model; thus, it is one of the very basic and important steps of feature selection.

*2) :* In mutual information regression, it returns scores based on mutual information of the continuous variables. These values are non-negative and measure the dependencies between the feature variable and the target variable. A higher value indicates high dependency. The current feature set shows mutual information scores ranging from 0 to 0.6.

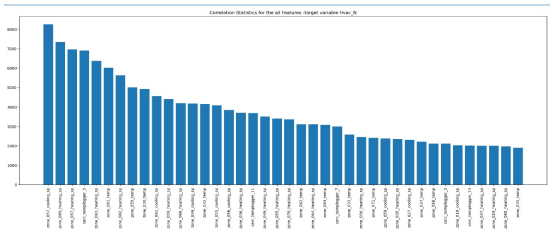

Fig. 2. Top 40 features selected based on f regression score for Hvac_N target variable

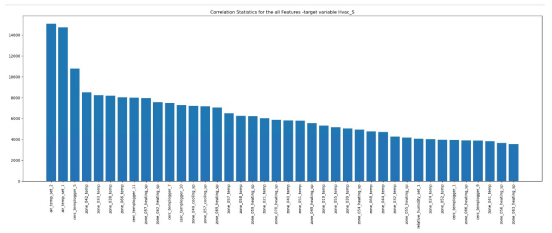

Fig. 3. Top 40 features selected based on f regression score for Hvac_S target variable

We selected 4 sets of 40 features for each of the target variables hvac_S and hvac_N with the highest f_regression scores and with the highest mutual information regression score. As PCA gave us the estimate of 40 features, based on that for further modeling we are considering the top 40 features based on the selection of the best algorithm.

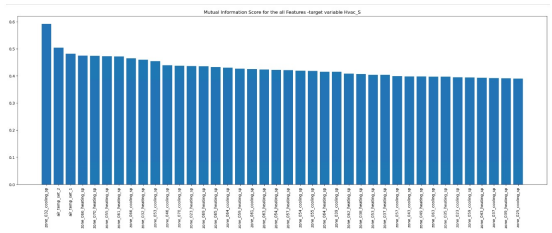

Fig. 4. Top 40 features selected based on mutual information score for Hvac_S target variable

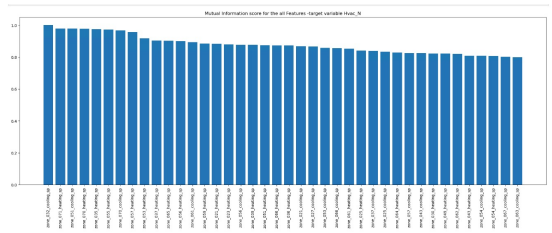

Fig. 5. Top 40 features selected based on mutual information score for Hvac_N target variable

## VII. Data Visualisations

Visualization helps explore and understand patterns, trends, and possible anomalies lying behind the data. More complex information becomes easier to digest in graphical form, and this can be a guide for feature selection and the initial steps of model design. These findings can be communicated to multiple stakeholders with different levels of technical background through visual tools. Results from complex models can be shared by using charts, graphs, and heat maps, hence guiding informed decisions and translating the technical results into business strategies.

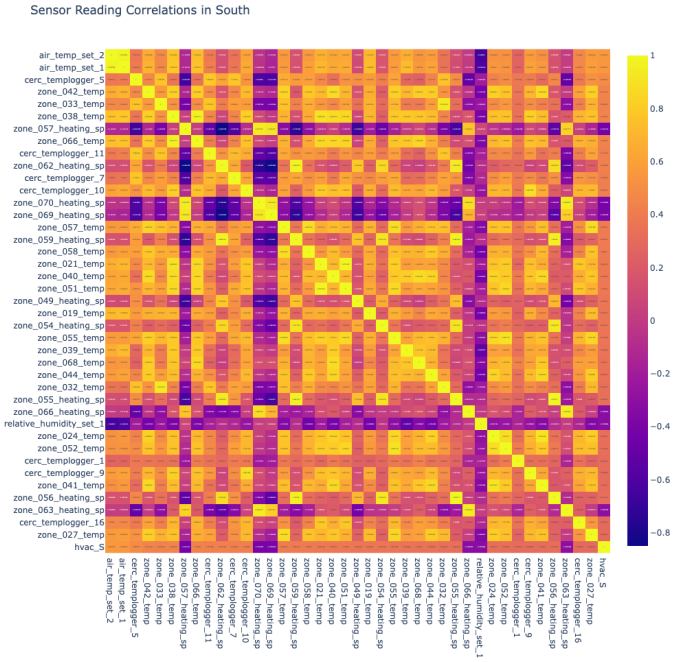


Fig. 6. Heatmap showing the correlation of the South sensors and HVAC

This heat map represents sensor reading correlation, where the colors close to yellow will have a high or good positive correlation, dark purple indicates a strong negative correlation, and red represents very little to no correlation. Each colored cell shows how two variables are related and would help in picking out patterns that can give insight into strategies for energy management. The labels note that the data includes temperature, humidity, and HVAC system metrics.

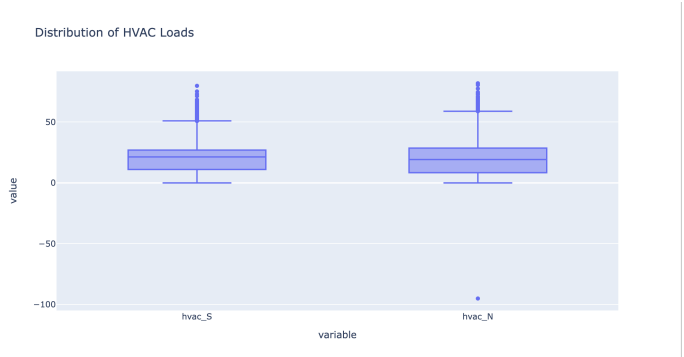


Fig. 7. Box plot of North and South HVAC

Box plots give an indication of the spread of HVAC loads, to the extent it could be applied to zone_temp_sp for cooling and zone_temp_sp_h for heating to determine optimal temperature settings. From this plot, we can notice that there are very minimal outliers. The middle line in each box represents the median of the dataset. The median appears larger for hvac_N compared to hvac_S, suggesting that on average the North HVAC system is busier.

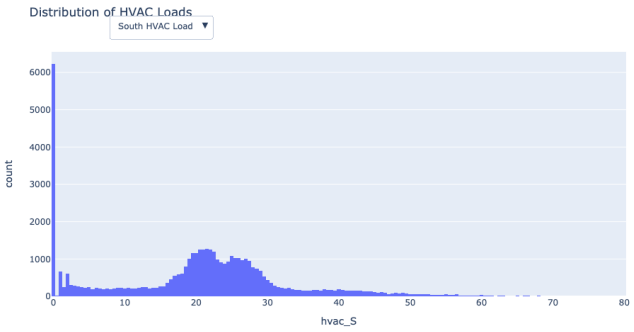


Fig. 8. Interactive plot for the distribution of hvac_N and hvac_S

The histogram indicates a clustering of most of the HVAC load values towards the left side of the graph, i.e., more frequent lower load values. The right skew of the distribution, as seen with the tail extending towards the right, means that it is a right-skewed distribution. This generally means that once in a while, there are likely to be instances where the HVAC load can be significantly higher than the usual set of values. The detached bars at the extreme right end of the histogram show the existence of possible outliers. These are the higher-than-the-rest values of HVAC load, which could stand for some unusual or exceptional conditions, such as severe weather or malfunctioning systems.

## VIII. Modeling

Various models like Multilinear regression, K nearest Neighbors, Random Forest, and XG boost are used, and subsequently, their accuracy is compared.

### A. Multiple Linear Regression

In this project, multiple linear regression models(MLR) are constructed to predict the HVAC loads for the South Wing (hvac_S) and North Wing (hvac_N) of a building. Feature selection is executed using both f-regression and mutual information scores, determining which features from the dataset are most predictive of the target variables. The pipeline is built to standardize the features and train the model with a regularized multiple linear regression algorithm. For each combination of selected features 3 different pipelines are built Standard Scalar, Robust Scalar techniques are used to standardize the features. Regularization techniques - Ridge and Lasso are incorporated to prevent overfitting and improve model reliability. Models are tuned with the application of different scaling techniques, and alpha parameters in regularization terms.

### B. K nearest Neighbors

K-nearest neighbors' algorithm is applied to the north and south wing HVAC load prediction within a building. It seems very appropriate for the project's application in that, during testing, data is not processed unless new data has been given; the processing of data is done only at that time. This approach computes the distances between new and existing data points to detect the nearest neighbors through various distance metrics like Euclidean and Chebyshev. A great deal of accuracy of the model was assessed and tuned very minutely via feature selection by using the SelectKBest method of the sci-kit-learn library for extracting the most significant features with respect to their correlation to the target variables, hvac_S, and hvac_N.

The k-NN model is first built with a set of five neighbors. GridSearchCV along with methods for optimization such as k-fold cross-validation was set to sweep the neighbor values between 1 to 25 for an effective setup. This would validate the initial setup; thus, no further tuning in the number of neighbors is required. This validation underpins the robustness of the model and the effectiveness of our strategies related to feature selection and parameter optimization, which gave rise to an accurate and efficient setup for HVAC load forecasting.

The scatter plot shows the comparison between predicted and actual values for hvac_N and hvac_S. This gives an indication that our predictions are accurate to a great extent. The predicted values lie across and near to diagonal straight line. The straight line represents actual target values in the dataset.

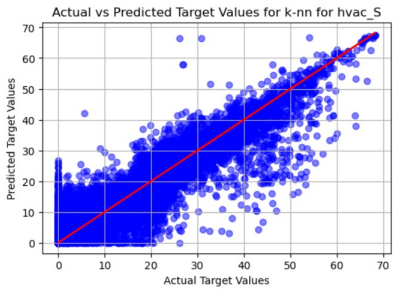


Fig. 9. Actual vs Predicted Target Values for hvac_S for KNN

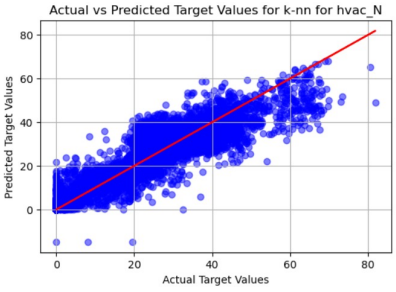


Fig. 10. Actual vs Pedicted Target Values for hvac_N for KNN

### C. Random Forest

First, decision tree construction and random sampling are used. Multiple decision trees are constructed to train the Random Forest; in this instance, each tree is constructed by randomly selecting subsets of the training sets data together with an extra random feature selection. Subsequently, a decision tree is constructed, selecting some samples from a training dataset through additional recursive partitioning to reduce impurity (mean squared error for regression). The second is averaging, which indicates that the forecast in a regression will be the mean of the tree models. After the pre-processing steps like cleaning the data, the dimensionality reduction of the dataset is prepared after splitting the data into 90 and 10 percent at first for testing the purpose. After that split, we took 90 percent of the data and built the models and later split them into 80:20 for training and validation. Loaded the data into pandas Data Frame and Split it into features X and the target variable y, which are the hvac_N and the hvac_S. Some of the features include dew point temperature, solar radiation, air temperature, zone temperature, zone CO2, humidity, etc. Implementing the standardization using the scaling features MinMaxScaler. MinMaxScaler scales and translates each feature to [0, 1]. It's used when scaling all features before training machine learning models to give them equal scales and prevent features with a larger magnitude from dominating. Next, an instance of the Random Forest regression is created with the specific parameters: number of trees – n_estimators, random state, etc. The model is then trained on the training data X_train, y_train using the fit method. During the fulfillment of this method, lots of decision trees are built on random details of data and properties. After that, the model predicts the target variable for the test set, X_test, using the predict method. The entire code demonstrates the use of a Random Forest regression for predicting HVAC (Heating, Ventilation, and Air Conditioning) energy consumption in commercial buildings.

This chart is generated using DataRobot which clearly shows that the prediction model has a good consistency in the range of its predictions and is able to give very close tracking to the real data, which may suggest that this is a well-performing model for the task. It is able to sort and drill down on the data, which hints at an interactive feature that could give greater insights into specific segments within the data.

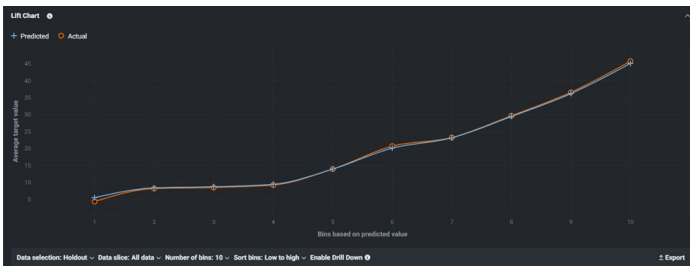


Fig. 11. Bins based on predicted values

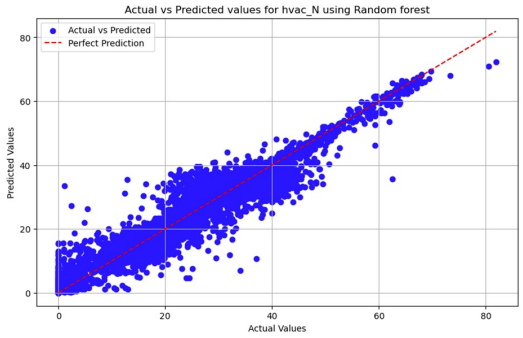


Fig. 12. Actual vs Pedicted Target Values for hvac_N for Random forest

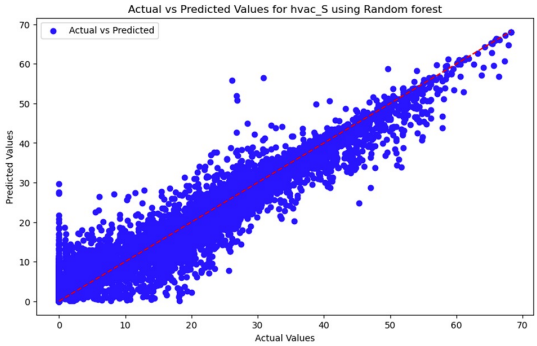


Fig. 13. Actual vs Pedicted Target Values for hvac_S for Random forest

### D. XGBoost Regression

We also employ the Xtreme Gradient Boosting (XGBoost) algorithm, a sophisticated decision-tree-based ensemble technique that utilizes the principles of gradient boosting frameworks. This model is particularly advantageous for its efficiency and effectiveness in handling various types of predictive modeling scenarios. The XGBoost models are separately trained to predict hvac_S and hvac_N using features selected through f regression. These models undergo extensive cross-validation through GridSearchCV to fine-tune the parameters, thereby enhancing the prediction accuracy

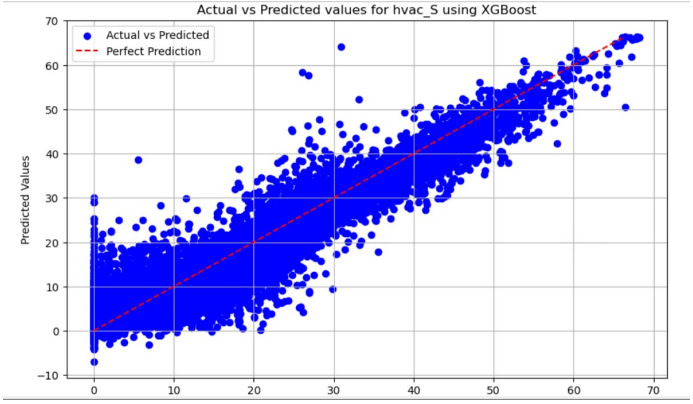


Fig. 14. Actual vs Pedicted Target Values for hvac_S for XGBoost Regression

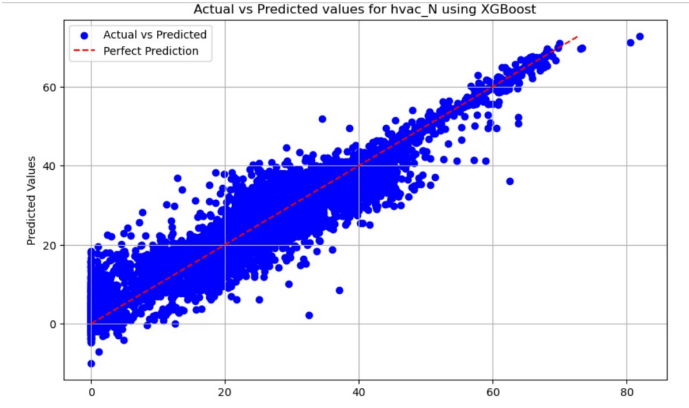


Fig. 15. Actual vs Pedicted Target Values for hvac_N for XGBoost Regression

## IX. Evaluation

Evaluation metrics are useful for evaluating how well machine learning models perform. Several metrics employed in the model to assess the performance include MSE or mean squared error: The average squared difference between the actual and anticipated numbers is what it calculates. Better model performance is indicated by a lower MSE. The average absolute difference between the expected and actual values is measured by the Mean Absolute Error (MAE). Though it may not penalize huge errors as much as MSE, MAE is straightforward to grasp. The percentage of the dependent variable's variance that can be predicted from the independent variables is shown by the R-squared (R2) score. Higher values of the R2 score, which goes from 0 to 1, indicate a better model fit. Mean Squared Logarithmic Error is another metric that is just like MSE, except it takes the Mean of the squared difference between the natural logarithm of predicted and true values. It should be used when the target variable has exponential growth. Median Absolute Error (MedAE) is the value that shows the median absolute difference from the prediction. It is less susceptible to outliers compared to MAE. Mean Absolute Percentage Error is a measure of the average percentage difference from the actual value. It's beneficial when the errors need to be shown up regarding the actual value it compares. Explained Variance Score is a quality measure of how much variance in the target variable can be understood from the independent variables. Explained Variance Score can have any value between over negative infinity to 1, with higher values indicating better explanatory power.

## X. Results

### A. This is for the North

XGBoost with k-Fold Achieved the loftiest R2 score of0.943, indicating that this model explains roughly 94.3 of the friction in the target variable. It also demonstrated the smallest mean absolute error( MAE) and mean squared error( MSE) indicating better accuracy and lower prediction errors compared to other models where as Multiple Linear Regression with StandardScalar and Ridge Regularization with an R2 score of 0.774. still, it displayed advanced MAE and MSE compared to more complex models, indicating less accurate predictions and advanced errors.k- Nearest Neighbour( k- NN) Achieved a moderate R2 score of0.893, indicating reasonable predictive performance. still, it displayed advanced errors compared to XGBoost. Random Forest Demonstrated the loftiest overall performance with an R2 score of0.949, indicating that it explains roughly 95.0 of the variance in the target variable. It also displayed the smallest MAE, MSE, and RMSE, indicating superior accuracy and lower forecast errors compared to other models.

| Model | R2 Score | MAE | MSE | RMSE | MSLE | MedAE | MAPE | EVS | F-statistic |
|---|---|---|---|---|---|---|---|---|---|
| XGBoost | 0.943 | 1.897 | 10.644 | 3.263 | - | - | 5.3225E+14 | - | - |
| Multiple Linear Regression - StandardScalar | 0.774 | 4.404 | 39.892 | 6.316 | - | - | 2.5989E+15 | - | - |
| k-Nearest Neighbour k-NN | 0.893 | 2.474 | 20.095 | 4.483 | - | - | - | - | - |
| Random forest | 0.949 | 1.656 | 9.298 | 3.049 | 0.087 | 0.577 | 4.0411E+14 | 0.950 | 1585.49 |

Fig. 16. Evaluation metrics of North

### B. This is for the South

| Model | R2 Score | MAE | MSE | RMSE | MSLE | MedAE | MAPE | EVS | F-statistic |
|---|---|---|---|---|---|---|---|---|---|
| XGBoost | 0.892 | 2.939 | 18.071 | 4.251 | - | - | 2.186e+15 | - | - |
| Multiple Linear Regression | 0.774 | 4.404 | 39.892 | 6.316 | - | - | 2.599e+15 | - | - |
| k-Nearest Neighbors (k-NN) | 0.839 | 3.230 | 27.063 | 5.202 | - | - | - | - | - |
| Random Forest | 0.915 | 2.419 | 14.210 | 3.770 | 0.340 | 1.376 | 1.873e+15 | 0.915 | 1484.644 |

Fig. 17. Evaluation metrics of South

XGBoost's k-Fold Cross Validation. Even though all four errors are minimized and the R2 score of 0.89 is robust, the MAPE confirms that fair chance errors still exist. Standard Scaler and Ridge Regularization in Multiple Linear Regression. The model is respectable overall with an R2 value of 0.77 and has more complex RMSE and MAE mistakes than XGBoost. The MAPE is also very high. k-NN algorithm. The k-NN algorithm is closing in on competitive excellence because of its R2 score of 0.84, low MSE, MAE, and RMSE errors, as well as the high RMSE for the other three errors. Random Forest has a high performance by the high R2 score of 0.91 and the low error rate across the metric parameters it requires. The Random Forest demonstrates its resistance to outliers by the low MSLE and MedAE it needs.

In conclusion, the Random Forest model outperformed other models in predicting HVAC energy consumption in commercial buildings, achieving the highest accuracy and explaining the most variance in the target variable.

## XI. Learning Outcomes

We learned quite a good number of lessons from the project, one thing is as discussed in class, preprocessing itself is very important for the better performance of the model and it takes 70 percent of the time. Effective planning according to the deadlines and understanding the importance of communication and equal work distribution among the team members. Exploring multiple papers and exploring new methodologies improves the understanding of the topic and brings great command over the topic. By embracing change and adjusting our strategies as needed, we were able to overcome obstacles and achieve our goals more effectively. Learned the importance of dimensionality reduction, feature selection, tuning, etc. in model building. Another one is the usage of multiple evaluation metrics by considering multiple metrics, we ensure a more thorough understanding of the model's strengths and weaknesses, facilitating informed decision-making and improving the quality of machine learning solutions. Another important concept learned is using a data robot or lazy prediction of things prior to implementing the models is a good approach. Here we implemented a data robot after implementing the models in our project then we got to know that light gradient boosted trees regressor with early stopping performed better and achieved an accuracy of 95.8 percent for hvac_N and 92.3 percent for hvac_S if we had known it earlier we could have picked it initially only these little percent increase in accuracy in real time leads to lots of difference in many fields like healthcare and many more.

## References


[1] Na Luo, Zhe Wang, David Blum, Christopher Weyandt, Norman Bourassa, Mary Ann Piette and Tianzhen Hong , 'A three-year dataset supporting research on building energy management and occupancy analytics' 2022.

[2] Dina M. Ibrahim,Abdulbasit Almhafdy,Amal A. Al-Shargabi, Manal Alghieth,Ahmed Elragi and Francisco Chiclana , 'The use of statistical and machine learning tools to accurately quantify the energy performance of residential buildings' 2022.

[3] Geoffrey K.F. Tso and Kelvin K.W. Yau,'Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks' 2007.

[4] William Chung , 'Review of building energy-use performance benchmarking methodologies' 2011.

[5] Florence Lai, Frédéric Magoulès and Fred Lherminier , ' Vapnik's learning theory applied to energy consumption forecasts in residential buildings' 2007.

[6] Xiaoli Li ,Chris P. Bowers and Thorsten Schnier , 'Classification of Energy Consumption in Buildings With Outlier Detection' 2007.

[7] Wang Zeyu , Wang Yueren , Zeng Ruochen , Ravi S. Srinivasan and Sherry Ahrentzen , 'Random Forest based hourly building energy prediction' 2018.

[8] Athanasios Tsanas and Angeliki Xifara , 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools' 2018.

[9] Saleh Seyedzadeh, Farzad Pour Rahimian,Ivan Glesk and Marc Roper 'Machine learning for estimation of building energy consumption and performance: a review'2018.

[10] Hong and Tianzhen, 'A three-year building operational performance dataset for informing energy efficiency' 2018.