Data -245 Machine Learning Technologies

Commercial building Energy Optimization

A Machine Learning Approach

Group 8

Introduction

- Project Overview: Focus on optimizing building energy management using data from over 300 sensors collected over three years.
- Data Utilization: Comprehensive dataset includes indoor and outdoor environmental data, and HVAC system metrics.
- Machine Learning Integration: Employs a mix of linear and nonlinear models:
  - Multiple Linear Regression
  - Random Forest
  - Support Vector Machines
  - Gradient Boosting
- Goals and Insights: Aims to enhance energy efficiency and deepen understanding of the impact of building energy consumption on the environment.
- Predictive Analysis: Key objective to accurately predict HVAC energy consumption, driving sustainable building practices.
- Sustainability Commitment: Leveraging advanced ML techniques to contribute to a healthier, greener future through improved energy management.

Data for the project

The office building in Berkeley, California.

Where is the data from?

- **Dataset Overview**: Collected from an office building in Berkeley, California, spanning three years (2015 onward).
- **Building Specifics**: Data gathered from two office floors, each 2,325 m^2, using over 300 sensors and meters.
- **Data Composition**: Includes comprehensive details on whole building and end-use energy consumption, HVAC system operations, indoor and outdoor environmental conditions, and occupant count.
- **Data Curation Strategy**:
  - **Step 1**: Identification and adjustment of outliers and data gaps in the raw data.
  - **Step 2**: Application of the Brick schema for modeling building systems and data points' metadata.
  - **Step 3**: Utilization of a semantic JSON schema to describe the dataset's metadata.
- **Application Potential**: Suitable for various uses such as building energy benchmarking, load shape analysis, energy and occupancy forecasting, and HVAC control systems optimization.
- Link to the paper

What does this data consist of?

Outdoor environmental data	site_weather.csv	air_temp_set_1	Outdoor air temperature from sensor 1
		air_temp_set_2	Outdoor air temperature from sensor 2
		dew_point_temperature	Outdoor air dew temperature of sensor 2
		relative_humidity_set_1	Outdoor air relative humidity from sensor 1
		solar_radiation_set_1	Outdoor solar radiation from sensor 1
Indoor environmental data	zone_temp_sp_c.csv	zone_*_cooling_sp	Cooling temperature setpoint of Zone *
	zone_temp_sp_h.csv	zone_*_heating_sp	Heating temperature setpoint of Zone *
	zone_temp_interior.csv	cerc_templogger_*	Zone temperature of interior zone
	zone_temp_exterior.csv	zone_*_temp	Zone temperature of exterior zone
	zone_co2.csv	zone_*_co2	CO2 concentration of each zone

Target Variable:

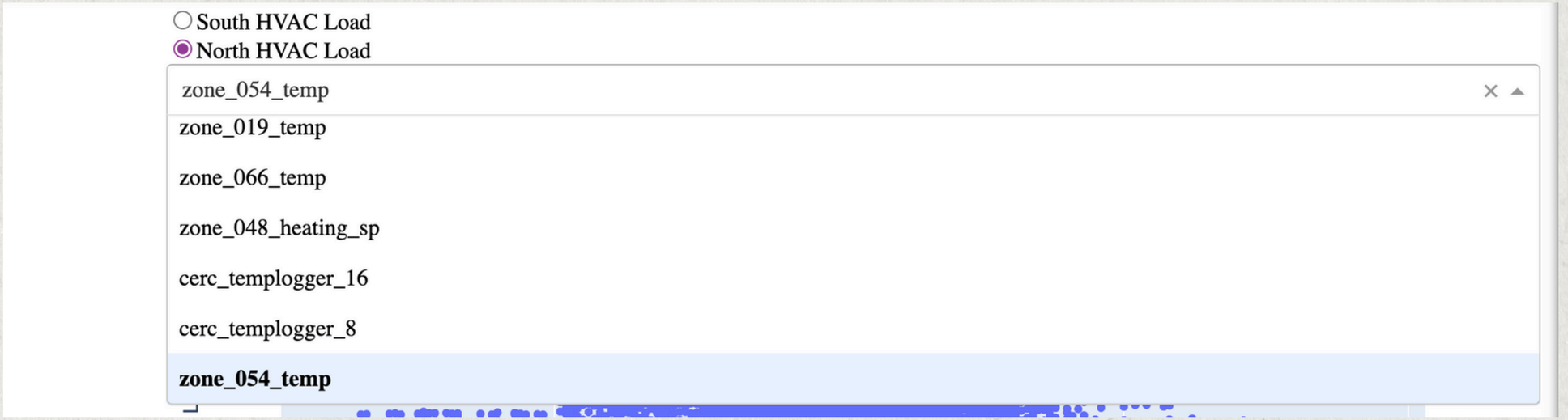
Energy use data	ele.csv	hvac_S	Heating Ventilation and Air Conditioning load for the South Wing
		hvac_N	Heating Ventilation and Air Conditioning load for the Nouth Wing

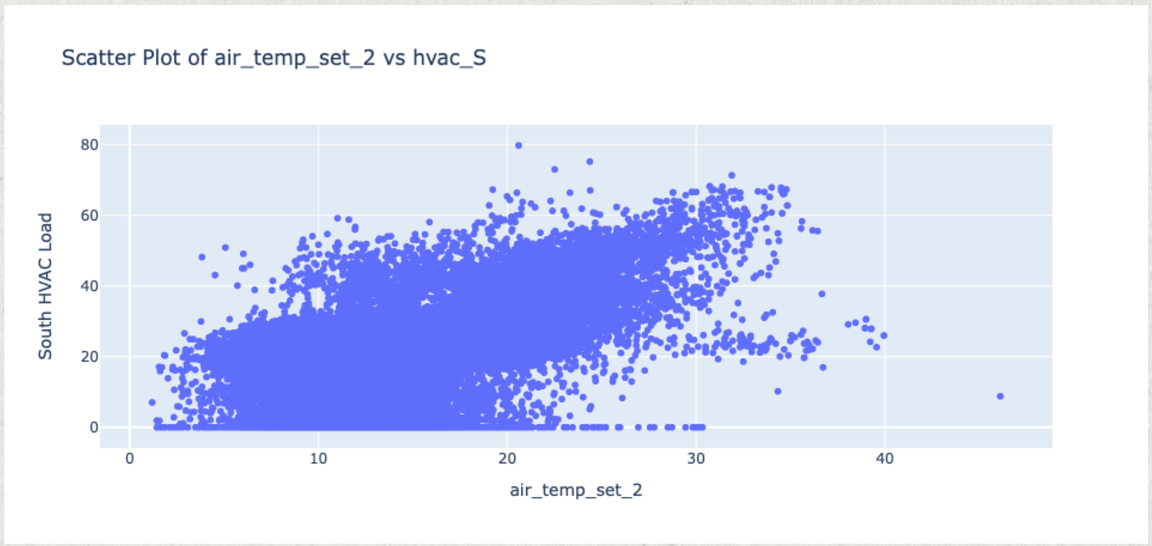
Exploratory Data Analysis

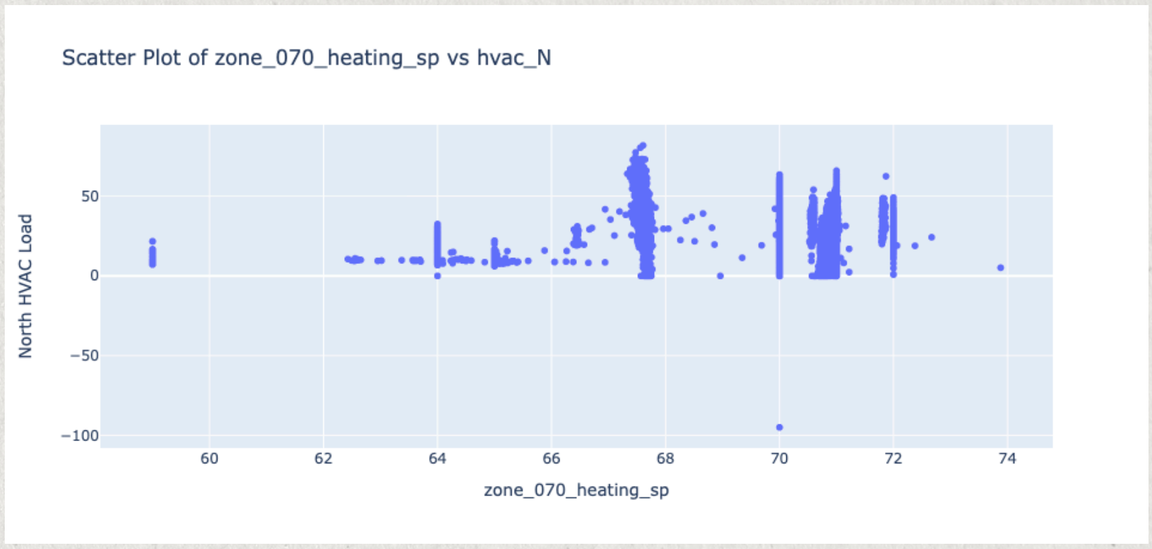
Preprocessing is a critical step to prepare the dataset for effective machine learning modeling. For this project, this involves several key actions:

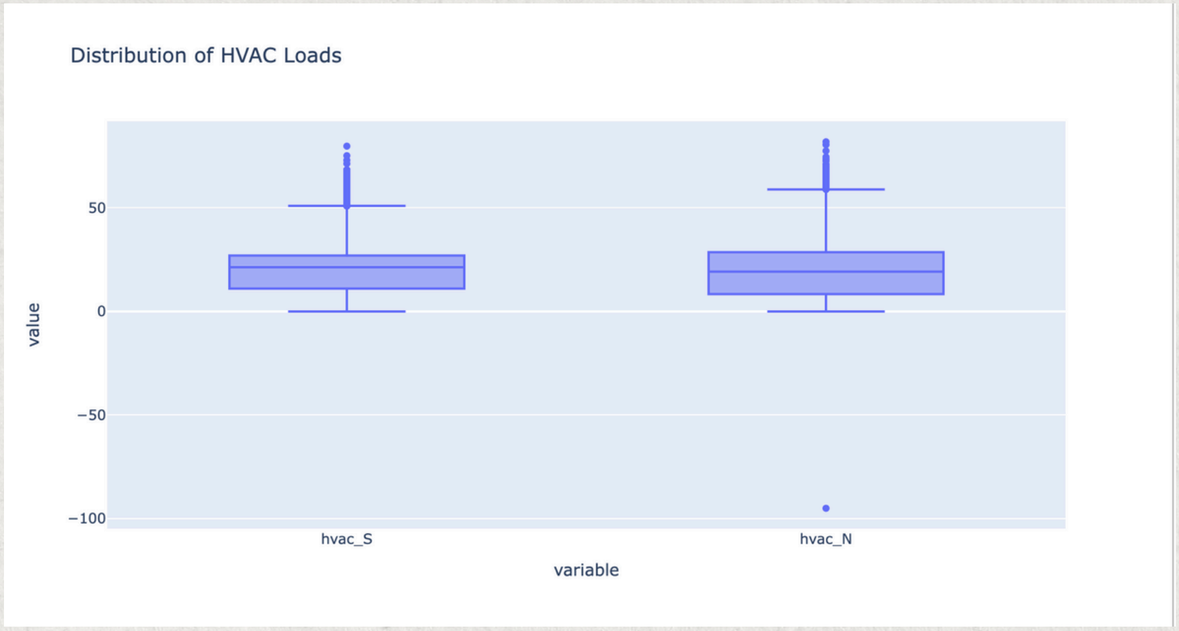
- Date Format Conversion - Converted all date fields to a uniform format '%Y-%m-%d %H:%M:%S' for accurate dataset merging.
- Dropping Unnecessary Data-Removed 'unnamed' fields containing NaN values from the cooling and heating datasets.
- Missing Value Imputation-Employed Iterative Imputer for missing value imputation, using multivariate imputation by chained equations (MICE).Used Bayesian Ridge Regression within the Iterative Imputer to predict missing values, ensuring regularization to prevent overfitting.

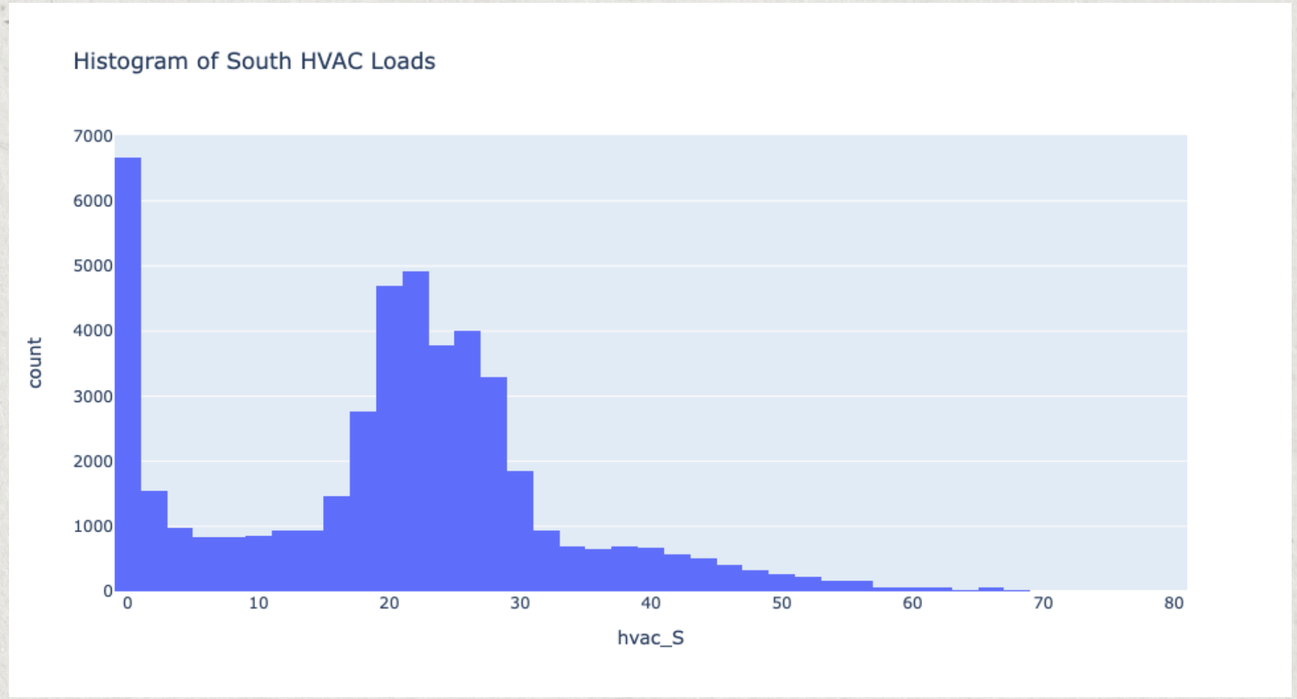
Visualizations

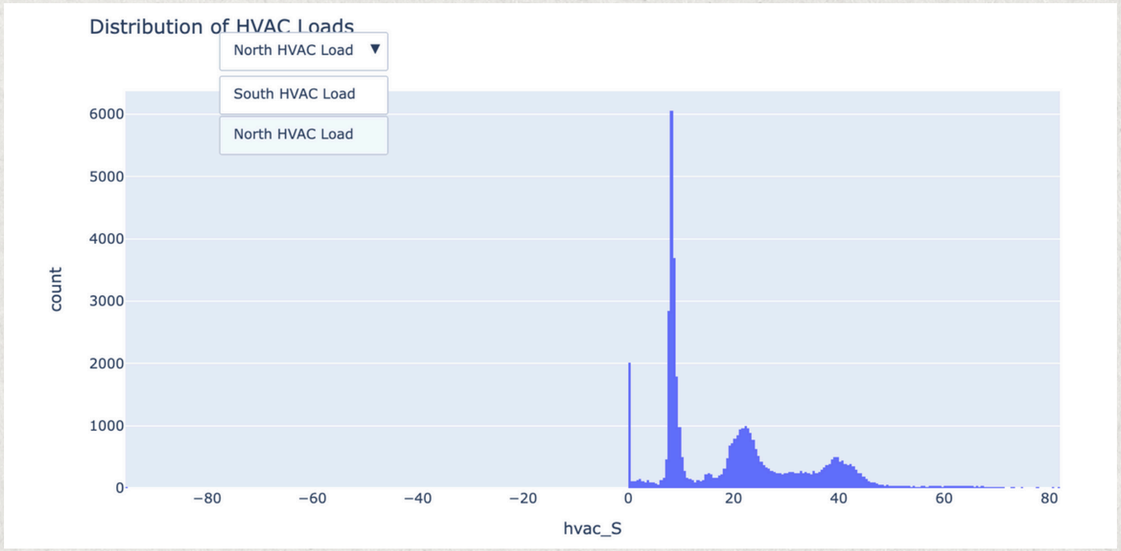


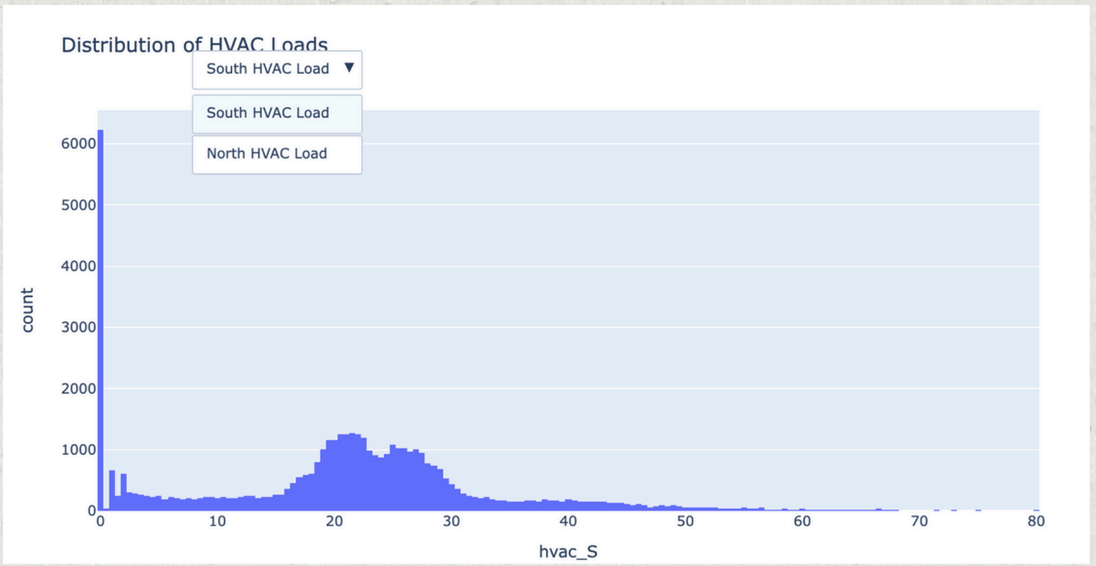





Distribution of HVAC Loads
value
50
0
−50
−100
hvac_S
hvac_N
variable


Histogram of South HVAC Loads
count
7000
6000
5000
4000
3000
2000
1000
0
0
10
20
30
40
50
60
70
80
hvac_S


Distribution of HVAC Loads
North HVAC Load
South HVAC Load
North HVAC Load
count
6000
5000
4000
3000
2000
1000
0
−80
−60
−40
−20
0
20
40
60
80
hvac_S


Distribution of HVAC Loads
South HVAC Load
South HVAC Load
North HVAC Load
count
6000
5000
4000
3000
2000
1000
0
0
10
20
30
40
50
60
70
80
hvac_S

Feature Selection

For feature selection we performed selectKbest univariate feature selection algorithm. It selects k-highest scored features based on pre-defined scoring function.

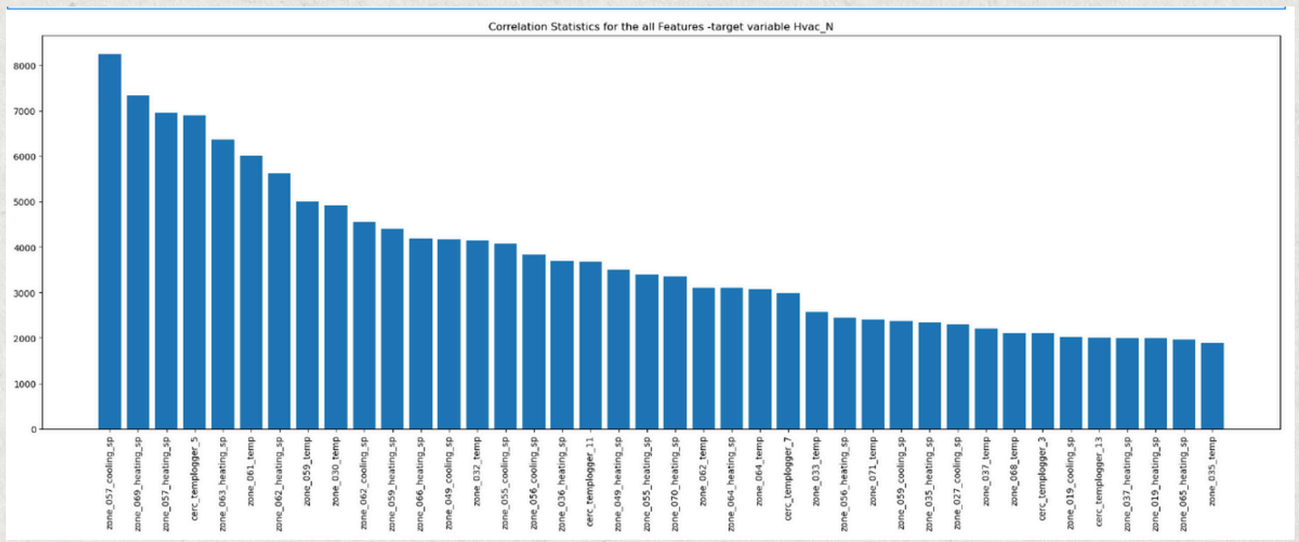
We adapted two scoring techniques for this project –f_regression and mutual information.

1:In f_regression scoring returns F-statistics and corresponding p-values in 2 steps process.

i) The cross correlation between each regressor and the target is computed using r_regression as:
E[(X[:, i] - mean(X[:, i])) * (y - mean(y))] / (std(X[:, i]) * std(y))

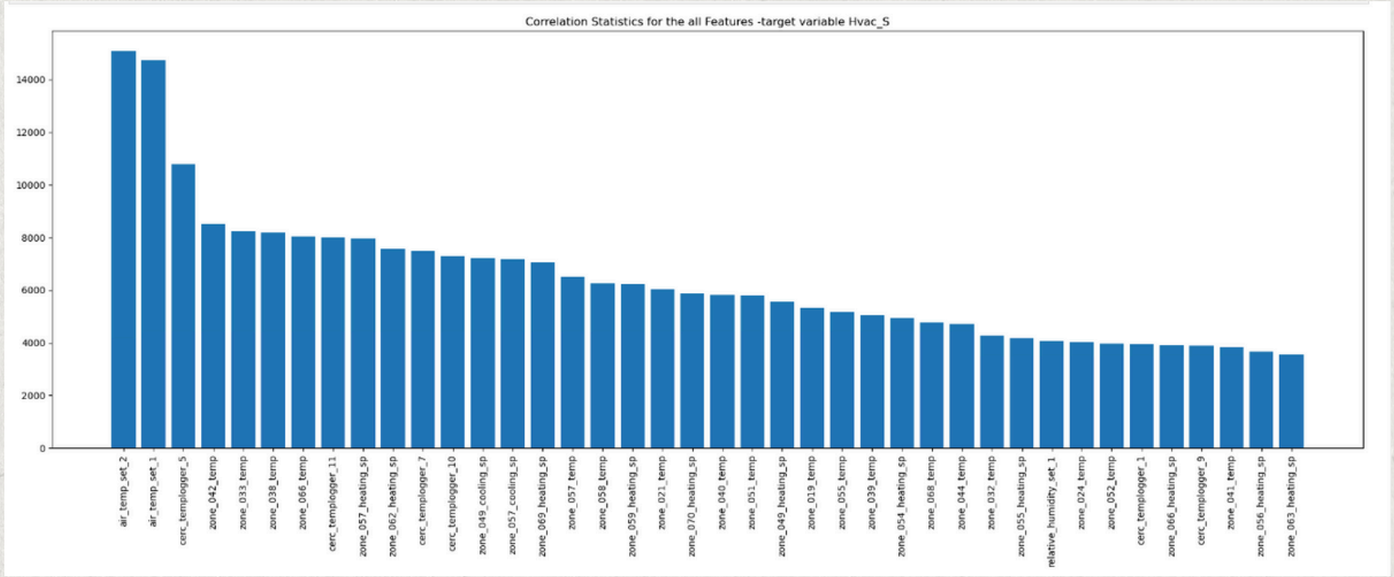
ii) In second step , this r_regression is converted to f-score and then to a p-value.
Thus, derived f-regression ranks the features in the same order as the positive correlation order with all the features to the target variable.
F-regression score statistics generate high positive score for each features ranging from 0 to 8000.

Feature Selection



Top 40 features based on f_score computed for Hvac_N target variable

Top 40 features based on f_score computed for Hvac_S target variable

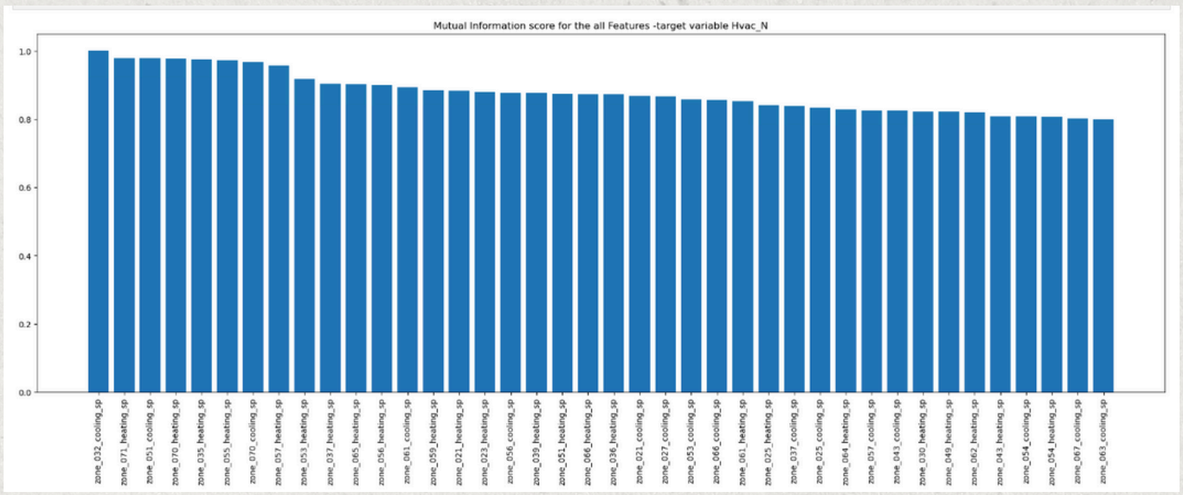


Feature Selection

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 target variable. Higher value indicates high dependency. The current feature set shows mutual information scores ranging from 0 to 0.6.

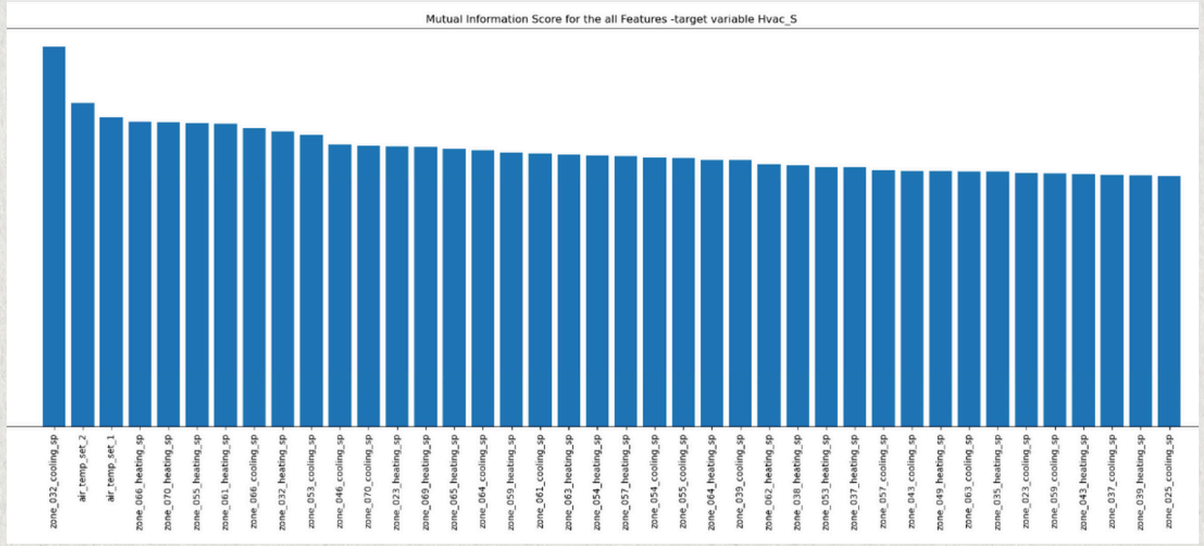
We selected 4-sets of 40 features for each of the target variables hvac_S and hvac_N with highest f_regression scores and with highest mutual information regression score. As PCA gave us the estimate of 40 features, based on that for further modeling we are considering the top 40-features based on selectKbest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

Feature Selection



Top 40 features selected based on Mutual Information score computed for Hvac_N target variable

Top 40 features selected based on Mutual Information score computed for Hvac _S target variable



Machine Learning Models

Multi Linear Regression

K Nearest Neighbors

Random Forest

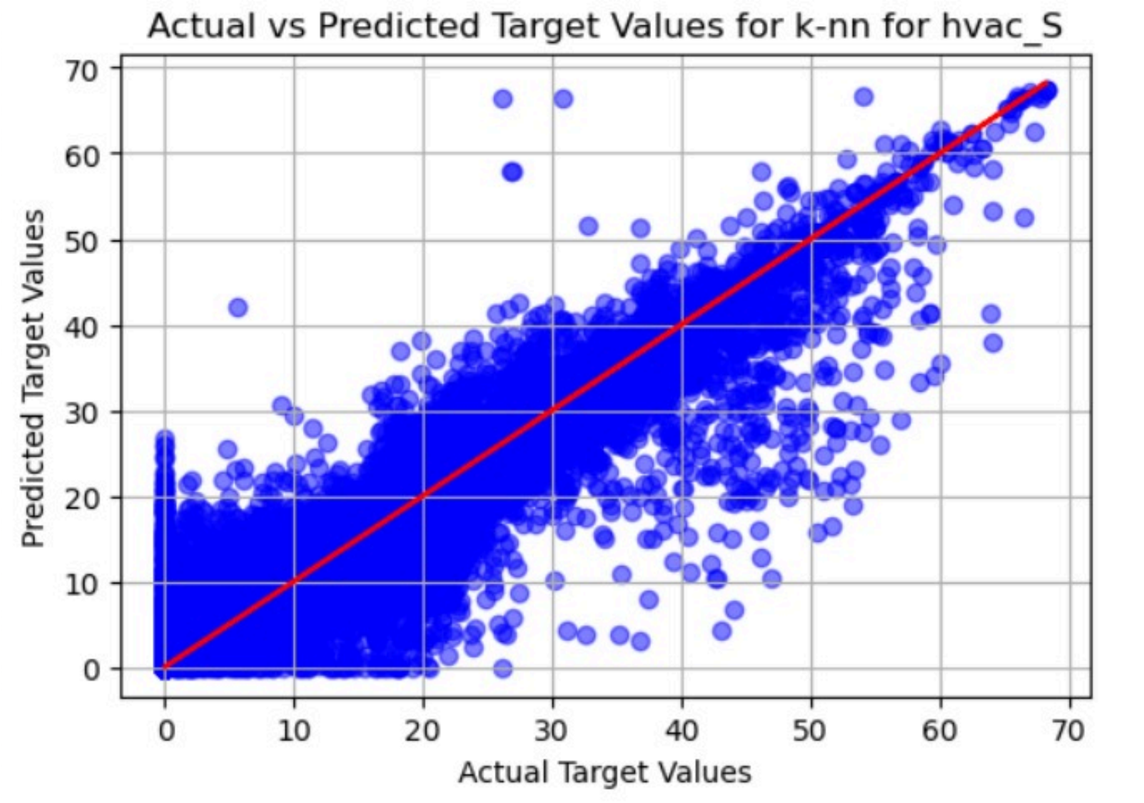
XG Boost

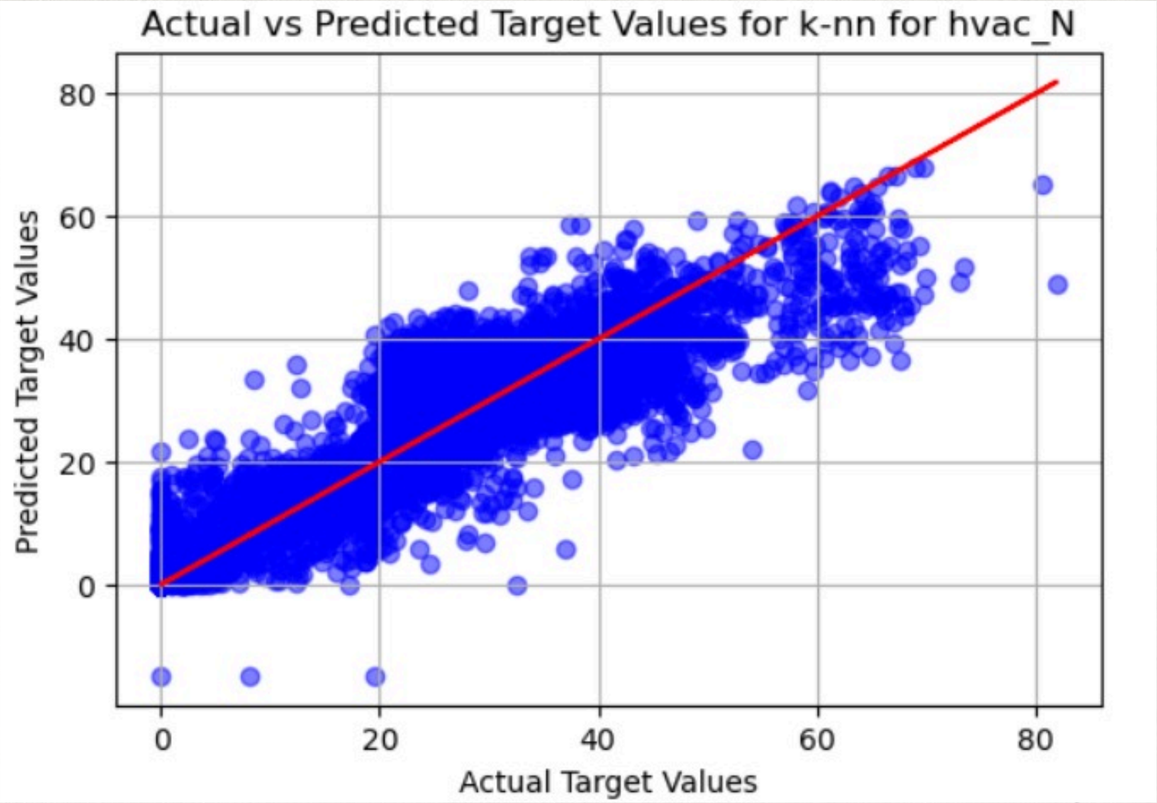
Multi Linear Regression

- MLR models are built to predict the HVAC loads for the South and North Wing
- Multiple models are built based on each set of selected features.
- Pipeline is built to standardize the features and train the model with regularized MLR algorithm
- Standard Scalar, Robust Scalar techniques are used to standardize the features.
- Regularization techniques - Ridge and Lasso are incorporated to prevent overfitting

K Nearest Neighbors

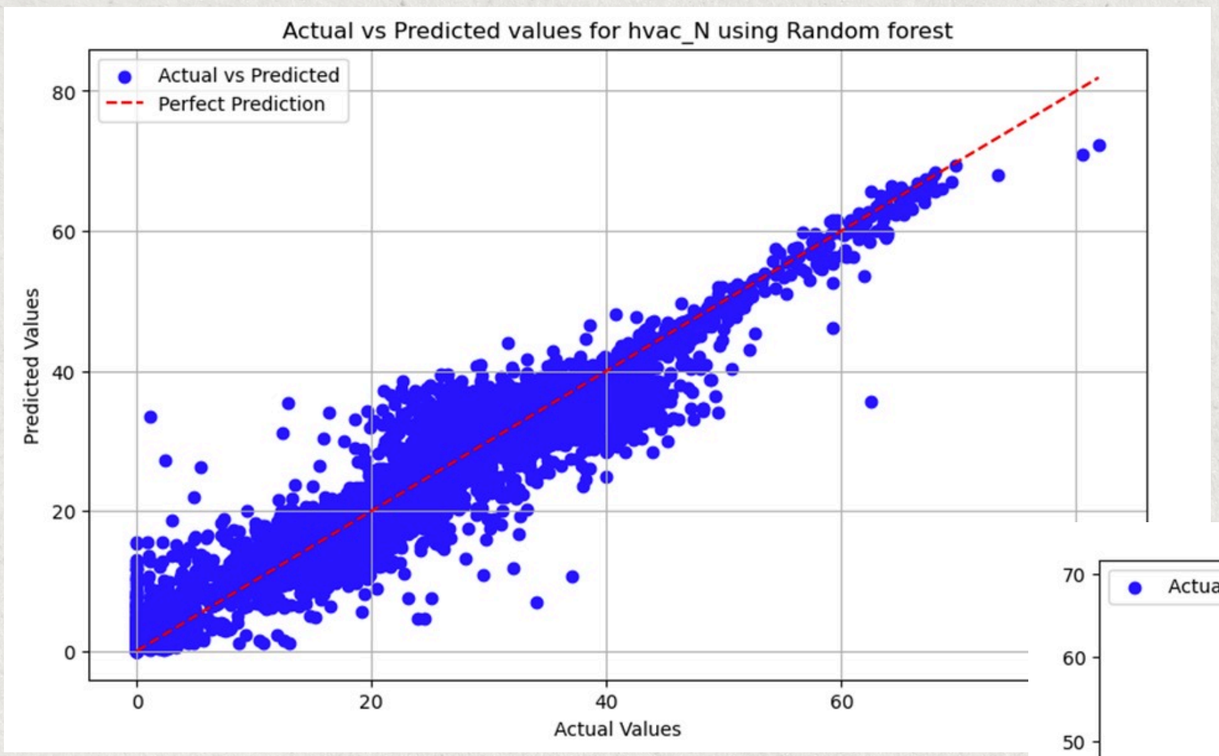
- k-NN is a supervised machine learning algorithm which is used for regression in our project. K-NN is also called lazy learning because until test data arrives it won’t do anything.
- K-NN is implemented with distance metrics such as Euclidean, Minkowski, Chebyshev, Manhattan.
- Cross Validation technique Grid-SearchCV is utilized to arrive at best hyperparameters so that high accuracy is achieved.
- SelectKBest method is used to find the features which are highly correlated to target variable hvac_N and hvac_S.

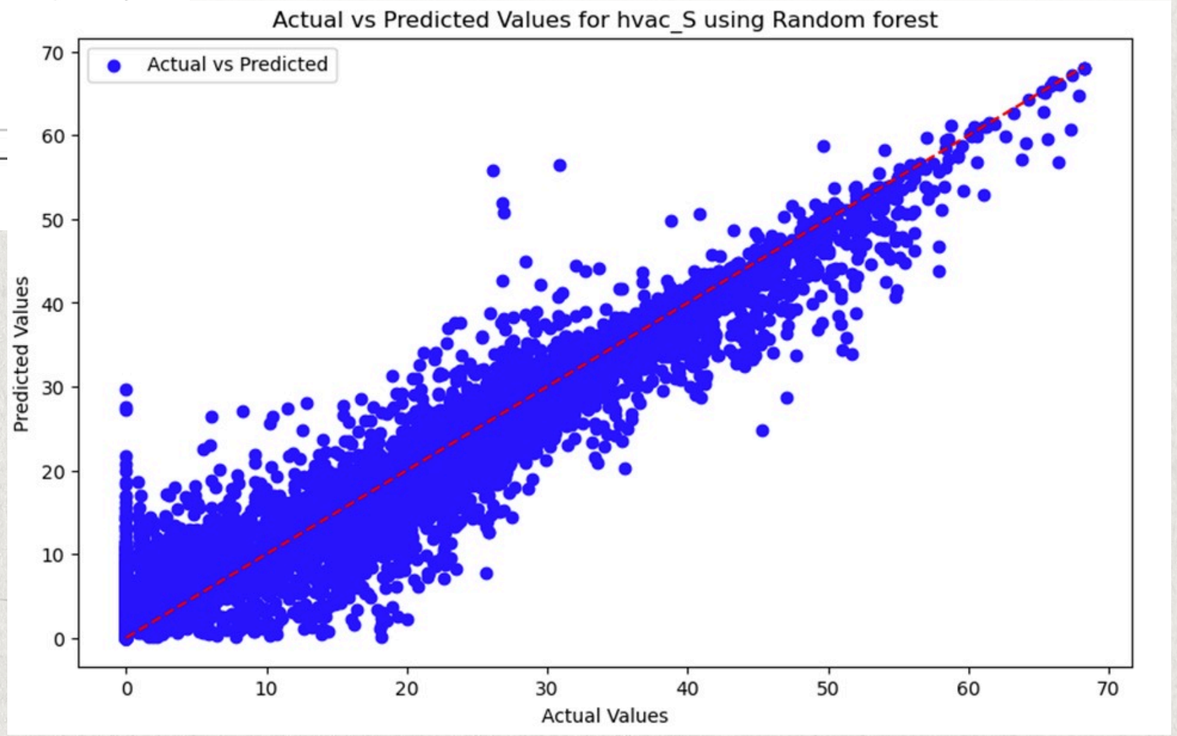

Actual vs Predicted Target Values for k-nn for hvac_S
Predicted Target Values
Actual Target Values
0
10
20
30
40
50
60
70


Actual vs Predicted Target Values for k-nn for hvac_N
Predicted Target Values
Actual Target Values
0
20
40
60
80

Random Forest

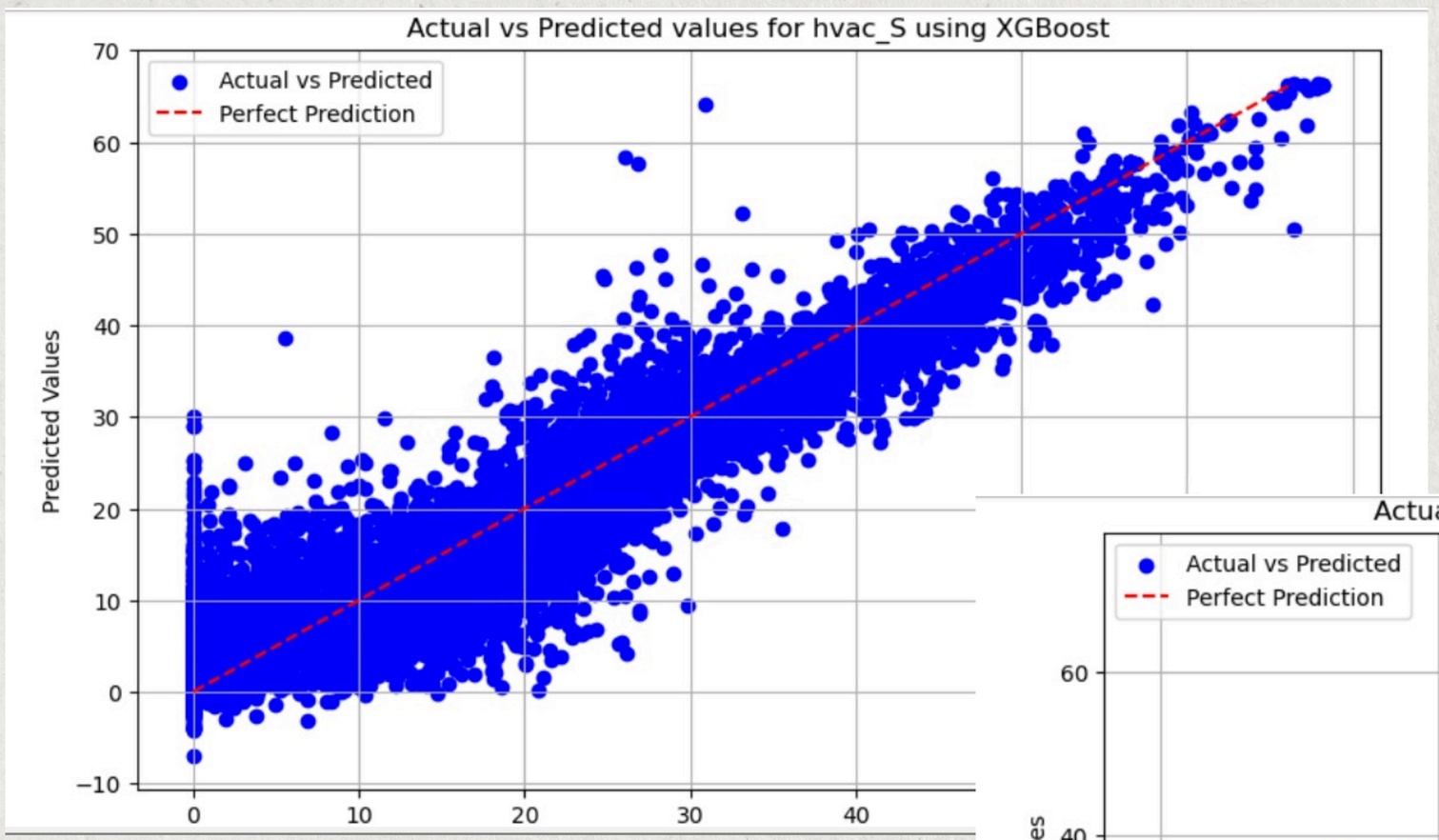
- The Random Forest Regressor is a learning method that enhances predictive accuracy and stability by constructing numerous decision trees during training. Each tree is trained on a random subset of the dataset, making decisions based on features.
- The dataset is split into training and testing sets. A Random Forest Regressor is instantiated with 100 decision trees and fitted to the training data. The final prediction is often the average (or mean) of the predictions made by individual trees.
- By analyzing the most important features as part of feature selection using f regression, we identified the most influential factors driving energy consumption in commercial buildings and built the random forest model after normalizing the features.
- Random Forest is known for its high accuracy due to the aggregation of multiple decision trees and robust to outliers and noisy data. It also handles datasets with many features without overfitting.

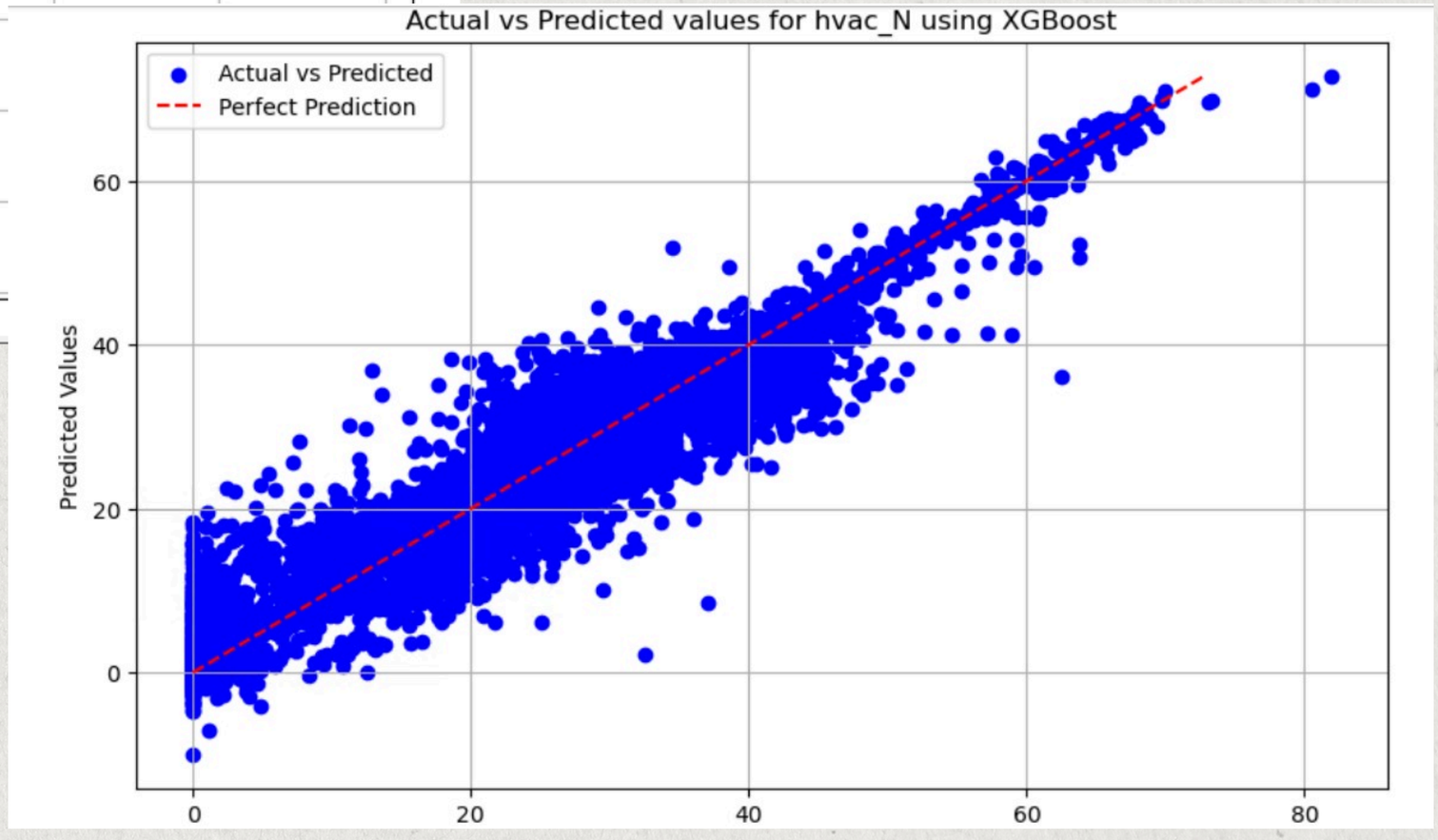

Actual vs Predicted values for hvac_N using Random forest
Actual vs Predicted
Perfect Prediction
Predicted Values
Actual Values
0
20
40
60
80


Actual vs Predicted Values for hvac_S using Random forest
Actual vs Predicted
Predicted Values
Actual Values
0
10
20
30
40
50
60
70

XG Boost

- An decision tree-based ensemble tree methods which uses principle of gradient boosting framework
- Here this algorithm predicts Hvac_S and Hvac_N in separate models.
- Exhaustive cross-validation technique Grid-SearchCV is used to achieve highest performance with best tuned parameters.
- With feature selected through f_regressor for hvac_N, XGBoost achieved around 94% accuracy to predict HVAC load in the North wing.


Actual vs Predicted values for hvac_S using XGBoost
Actual vs Predicted
Perfect Prediction
Predicted Values
70
60
50
40
30
20
10
0
−10
0
10
20
30
40


Actual vs Predicted values for hvac_N using XGBoost
Actual vs Predicted
Perfect Prediction
Predicted Values
60
40
20
0
0
20
40
60
80

Evaluation

Mean Squared Error (MSE): an average squared difference between the true and the predicted values, where smaller values are desirable.

MAE (Mean Absolute Error): This measures an average of the absolute difference between expected and actual values. It is not very strict because errors of high magnitude are not penalized as severely in comparison to in the case of MSE.

R-squared (R^2) Score: This is the amount of the variance in the dependent variable predictable from the independent variables. It has values ranging between 0 and 1; the higher the value, the better the fit indicated by the model.

MSLE (Mean Squared Logarithmic Error): The mean of the square of natural logarithm differences of the predicted value and true value is taken. Appropriate for targets showing exponential growth.

MedAE: Median absolute error. It shows the median absolute difference between the predicted and the real value. It is less sensitive to outliers than the MAE.

MAPE (Mean Absolute Percentage Error): This will provide the average percentage difference from the real values, hence useful in comparing errors with respect to actual values.

Explained Variance Score: A scoring of how much variance in the target variable is explained by the independent variables. Greater values have better explanatory power, but scores go from negative infinity to 1.

Results

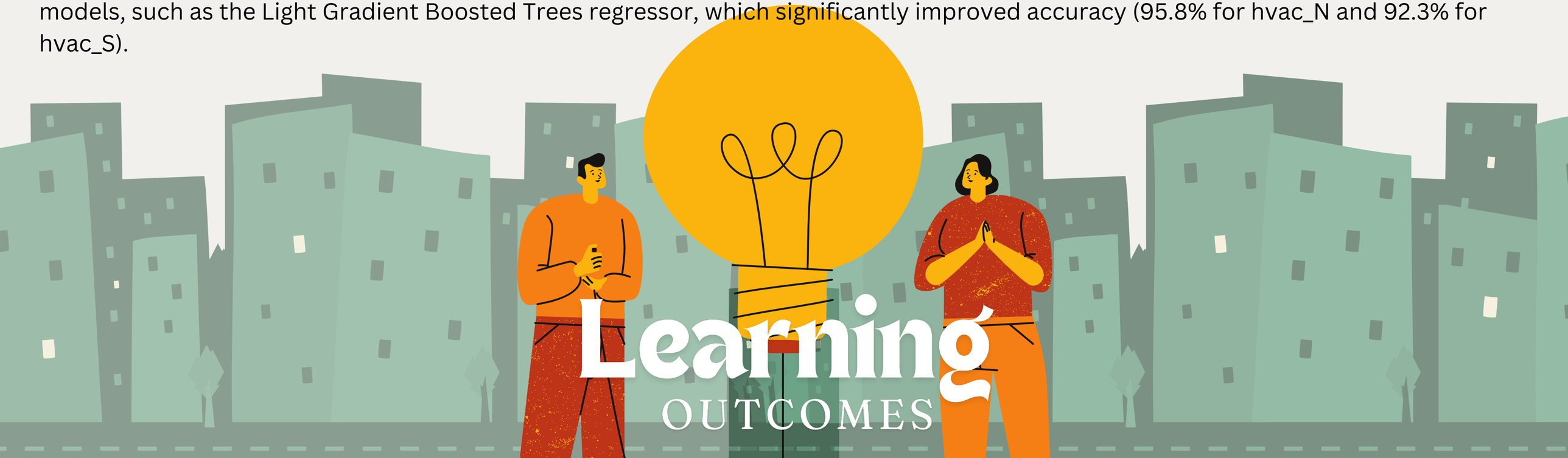
South

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

North

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

1. Preprocessing is Crucial: Preprocessing is vital for model performance, consuming about 70% of the project time.
2. Effective Planning and Teamwork: Proper planning, clear communication, and equitable work distribution among team members are essential for meeting deadlines and achieving goals.
3. Research and Methodology Exploration: Reviewing multiple papers and exploring new methodologies deepens topic understanding and enhances command over the subject.
4. Adaptability in Strategy: Embracing change and adjusting strategies as needed helps overcome obstacles and achieve objectives more effectively.
5. Importance of Dimensionality Reduction and Feature Selection: These techniques are critical in model building for improving machine learning outcomes.
6. Using Multiple Evaluation Metrics: Considering various metrics ensures a thorough understanding of a model's strengths and weaknesses, facilitating informed decision-making.
7. Early Implementation of Advanced Tools: Using tools like DataRobot or LazyPredict early in the process can identify more effective models, such as the Light Gradient Boosted Trees regressor, which significantly improved accuracy (95.8% for hvac_N and 92.3% for hvac_S).



Q&A?

Presented by Sandra Haro

Thank you very much!