Predictive Analysis of Building Energy Needs

Jules Prod'homme

Data Scientist

Project Summary

Objective

Predict the total energy consumption of non-residential buildings using only available structural data.

This project aims to utilize various data science techniques to build accurate predictive models.

Additionally, it involves evaluating the effectiveness of the ENERGY STAR Score as a metric for these predictions.

The secondary goal is to determine how well the ENERGY STAR Score correlates with the actual energy consumption and CO2 emissions, and to assess its value in predicting these factors.

Data used

Detailed Records:

- **CO2 Emissions:** Comprehensive data on carbon dioxide emissions from buildings in Seattle, collected in 2016.
- **Energy Consumption:** Detailed information on the total energy consumption (electricity, gas, steam) of these buildings for the year 2016.
- **Source:** These records were obtained from official municipal or environmental databases, ensuring accuracy and reliability.

Structural Data of Buildings:

- **Size:** Information on the square footage and overall dimensions of the buildings.
- **Usage:** Details on the type of usage (e.g., office, retail, industrial) for each building.
- **Construction Date:** The year when each building was constructed, providing insights into the age and potentially the energy efficiency of the structures.
- **Geographical Location:** Data on the location of each building, including coordinates and neighborhood information, which can influence energy consumption patterns.
- **Additional Features:** Other relevant structural characteristics such as the number of floors, the presence of energy-saving installations, and building materials.

Tasks

Exploratory Data Analysis (EDA):

- **Understanding and Visualizing Distributions and Relationships:**
  - Examine the distributions of key variables to understand their central tendencies, variability, and outliers.
  - Use visualizations such as histograms, box plots, scatter plots, and correlation matrices to explore relationships between variables.

- **Identifying and Handling Anomalies and Missing Data:**
  - Detect anomalies and outliers that may skew the analysis and determine appropriate strategies for handling them (e.g., removal, transformation).
  - Identify missing data points and decide on suitable imputation methods or whether to exclude them from the dataset.

- **Selecting Relevant Variables:**
  - Perform feature selection to identify the most important variables that influence energy consumption and CO2 emissions.
  - Use techniques such as correlation analysis, mutual information, and domain knowledge to choose variables that provide the most predictive power.

Modeling and Prediction:

- **Testing Different Predictive Models:**
  - Evaluate various models to identify the most effective for predicting CO2 emissions and energy consumption.
  - Compare model performance using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) to assess accuracy and reliability.

- **Tested Models Include:**
  - **Random Forest:** Ensemble method that improves accuracy and controls over-fitting by combining multiple decision trees.
  - **Extreme Gradient Boosting (XGBoost):** Efficient gradient boosting implementation for structured data.
  - **Support Vector Machines (SVM):** Finds the optimal hyperplane for prediction.
  - **ElasticNet:** Combines L1 and L2 regularization for improved generalization.

- **Integrating the ENERGY STAR Score:**
  - Incorporate the ENERGY STAR Score into the predictive models to analyze its impact and usefulness.
  - Evaluate how the inclusion of the ENERGY STAR Score influences the model’s performance and predictive accuracy.

Evaluation

Cross-Validation and Testing:

- Perform cross-validation to ensure robust predictions and avoid overfitting.
- Test the models on a separate test set to validate their performance.

Hyperparameter Optimization:

- Optimize model hyperparameters to enhance performance and accuracy.

This evaluation process ensures the models are reliable and well-tuned for making accurate predictions.

Dataset Presentation

Dataset

FEATURES

Building Location:

- **Neighborhood:** The specific neighborhood where each building is located.
- **Coordinates:** Longitude and latitude for precise geographical positioning.

Year of Construction:

- The year each building was constructed, providing insights into the age and potential energy efficiency of the structures.

Building Characteristics:

- **Floors:** The number of floors in each building.
- **Surface Area:** The total floor area of each building.
- **Usage Type:** The primary use of the building (e.g., office, retail, industrial).

Number of Buildings:

- The total number of buildings included in the dataset.

Type of Buildings:

- **Categories:** Different types of buildings such as residential, offices, schools, etcs

TARGETS

Energy Consumption:

- The total energy usage of each building, measured in relevant units .

CO2 Emissions:

- The amount of carbon dioxide emissions produced by each building, measured in relevant units.

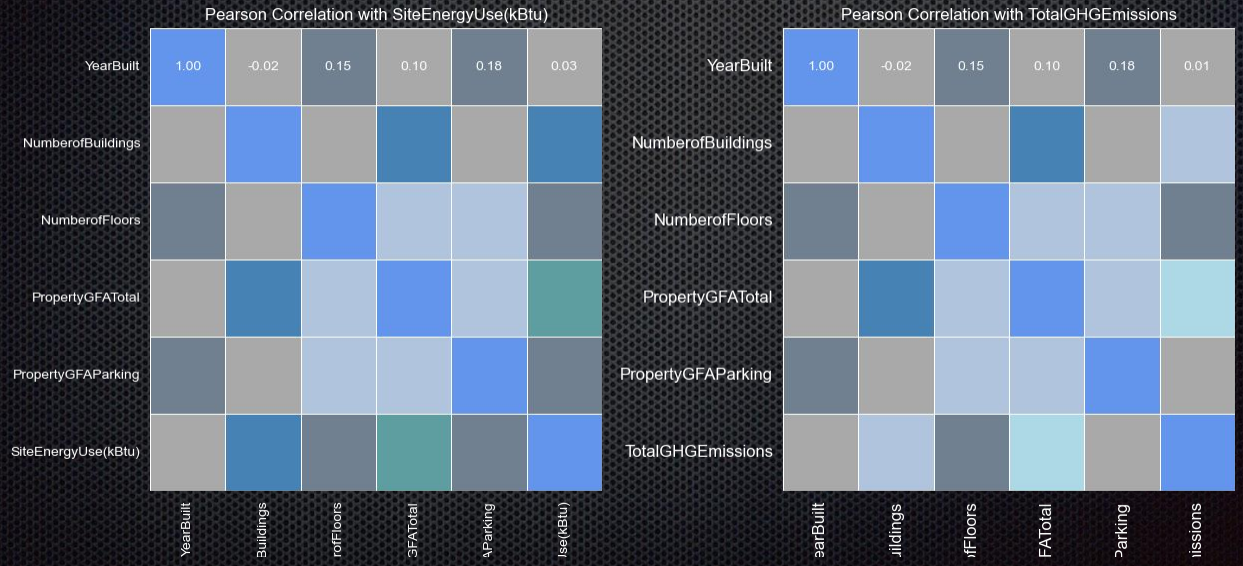
Cleaning

Cleaning steps	Variables involved
1. Selection of Non-Residential Buildings	Building Type
2. Compliance Status of Buildings	Compliance Status
3. Removal of Buildings with Outlier Values	Outliers Compliance Status
4. Removal of Buildings with Negative CO2 Emission Values	TotalGHGEmissions

Exploratory data analysis

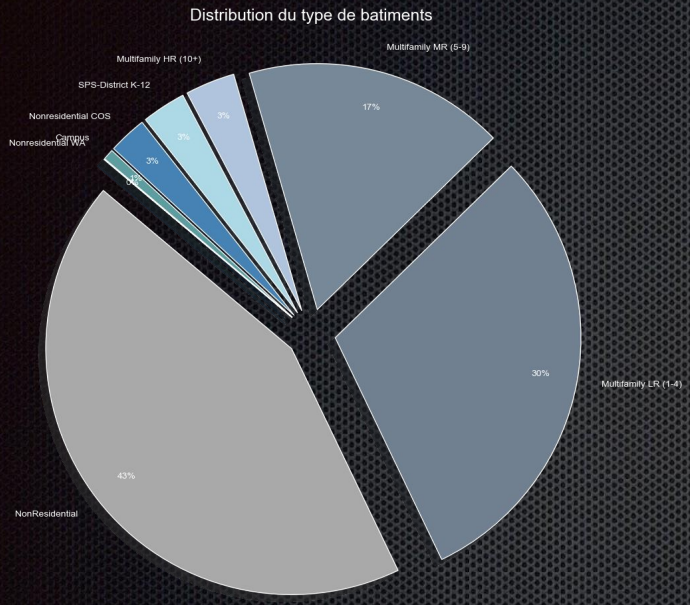
Displaying a Pearson Correlation Matrix with Target Variables:

- To understand the relationships between different variables in the dataset and the target variables (energy consumption and CO2 emissions), we will generate and display a Pearson correlation matrix.
- **Pearson Correlation Coefficient:** This coefficient measures the linear correlation between two variables, providing a value between -1 and 1.

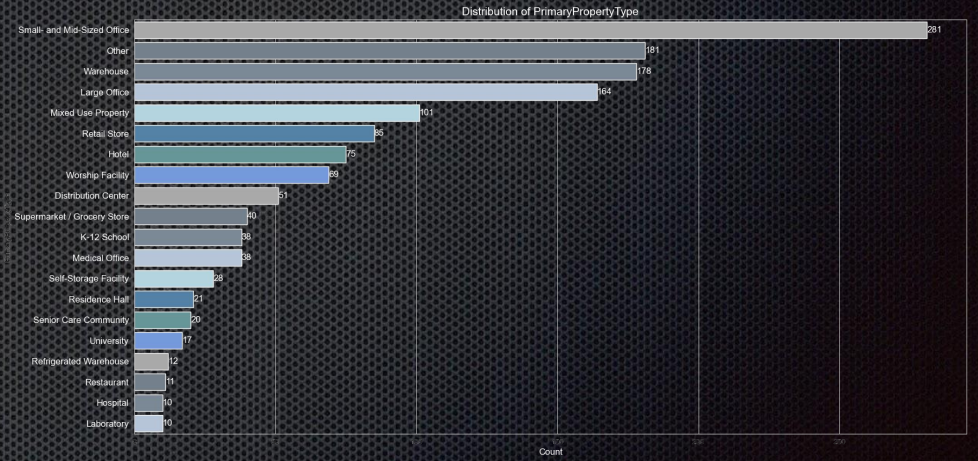


Building type distributions

- All building types



- Non residential building types



APPROACH TO MODELING AND PRESENTATION OF RESULTS

Modeling approach

Iterations between feature engineering and model training.

1. Trials using raw data

Trials of the 4 different models on the selected variables without any data transformations.

The metrics used are RMSE, MSE, R^2, and computation time.

We compare the results with a naive approach.

2. Trials on transformed data

Logarithmic transformation of variables

Cap and flooring on the targets

Binning into decades on the year of construction

3. Trials on data with dimensionality reduction

Principal Component Analysis (PCA) implementation

Application of dimensionality reduction on the data

4. Addition of the EnergyStarScore and comparison

Integration of the EnergyStarScore

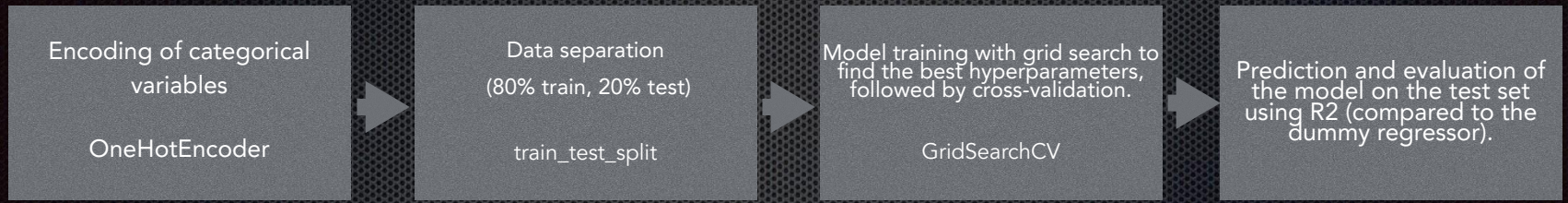
Reassessment on transformed data and on data with dimensionality reduction

Comparison of results

Modeling steps

➔ Targets:

- SiteEnergyUse (KBtu) (energy consumption)
- TotalGHGEmissions (co2 emission)



➔ Models used :

- Dummy regressor (to establish a baseline)
- Randomforest regressor
- XGBoost regressor
- Elasticnet
- SVM

➔ Metrics used :

- RMSE
- MSE
- R^2
- Temps de calcul

1st trial : raw data

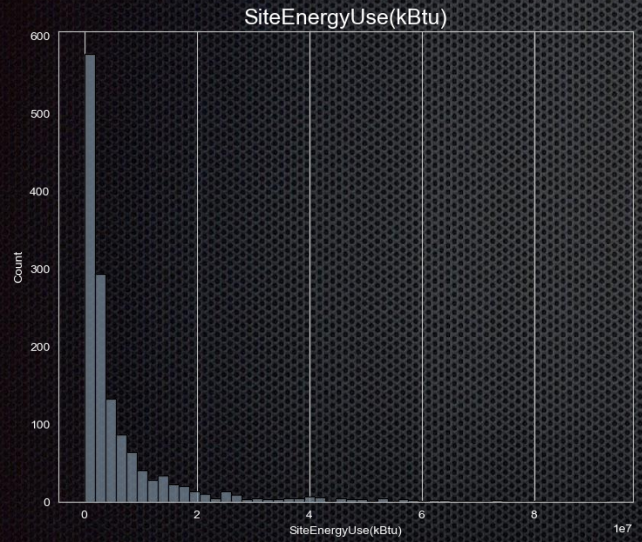
➔Trial on raw Data :

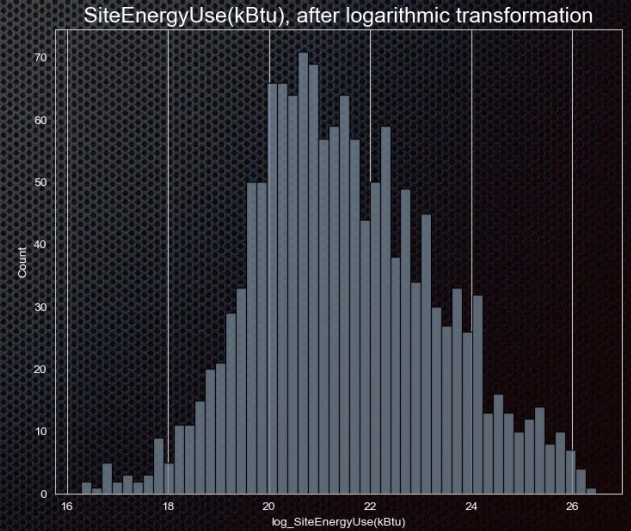
- This allows for a comparison baseline.
- A grid search is performed to optimize the models.
- We test the results solely by cross-validation.

2nd trial: Data transformation (log)

Logarithmic transformation of numerical variables:

- The data is left-skewed.
- Logarithmic transformation helps to center the distribution
- This process is applied to all quantitative variables that have a skew.





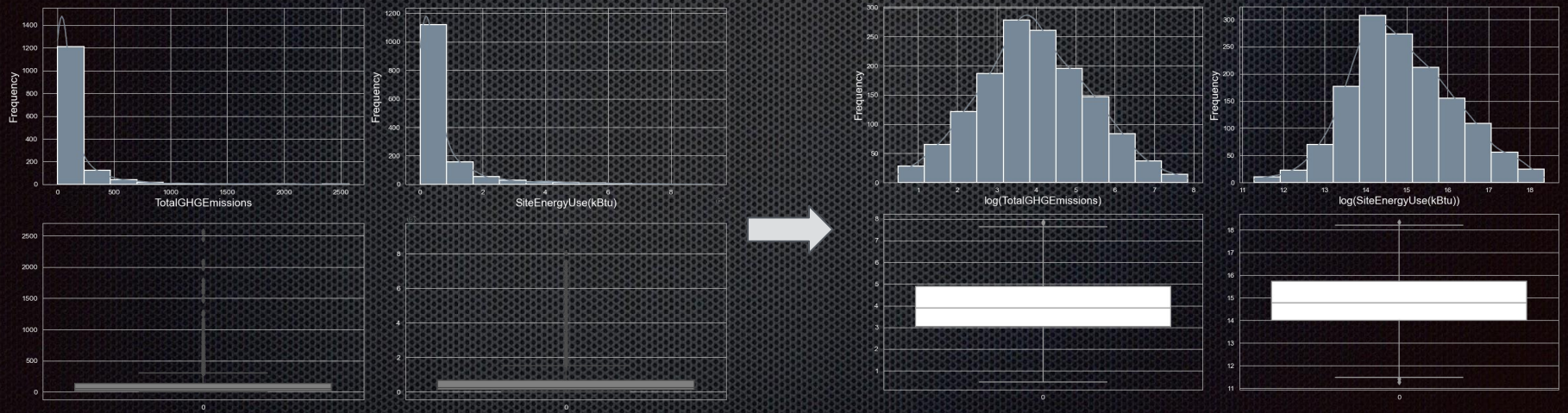
2nd trial: Data transformation (cap and floor)

➔ Cap and flooring on targets :

- We use the method of interquartile ranges to cap certain extreme values in the targets.

```
Q1 = round(np.percentile(ds_w_features["SiteEnergyUse(kBtu)"].apply(lambda x : np.log(1 + x)), 25))
Q3 = round(np.percentile(ds_w_features["SiteEnergyUse(kBtu)"].apply(lambda x : np.log(1 + x)), 75))

born_sup_energy = Q3 + 1.5*(Q3-Q1)
born_inf_energy = Q1 - 1.5*(Q3-Q1)
✓ 0.0s
```



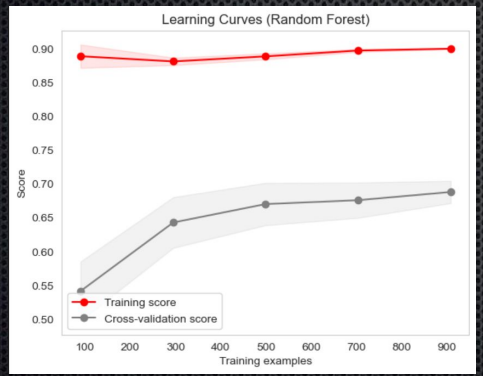
Random forest model

➔ The Random Forest method is an ensemble technique that relies on combining multiple decision trees. It uses both tree bagging (bootstrap aggregating) and feature sampling to enhance the model's performance.

➔ Hyperparameter grid search through cross-validation with GridSearchCV:

- Max depth [None, 10 , 20]
- N estimators [100, 200, 300]
- Min sample split [2, 5 , 10]
- Min sample leaf [1, 2 , 4]

Learning curve :



Prediction on energy consumption	Random Forest (scores on test set)
Best hyperparameters	Max_depth : 10, N_estimators : 200 Min_Sample_Leaf :4 ,min_sample split : 10
R^2	0.68
RMSE	1.02
MSE	1.04
Processing time (s)	0.89

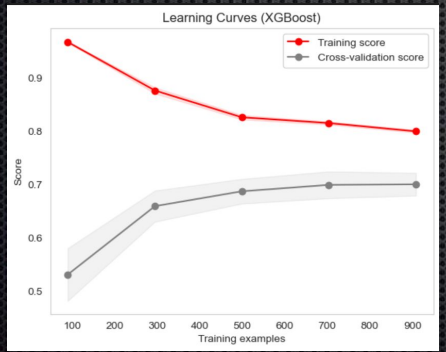
Xgboost model

➔ The XGBoost model is an ensemble method based on decision trees, optimized for speed and performance through regularization techniques and parallel processing.

➔ Hyperparameter grid search through cross-validation with GridSearchCV :

- Max depth : [3, 5, 7, 9, 11]
- N estimators : [100, 200, 300, 400, 500]
- Learning rate : [0.01, 0.05, 0.1, 0.2]
-

Learning curve :



Prediction on energy consumption	XGBOOST (scores on test set)
Best hyperparameters	max depth : 3, n estimators = 100, learning rate = 0.01
R^2	0.69
RMSE	0.99
MSE	0.99
Processing time (s)	0.069

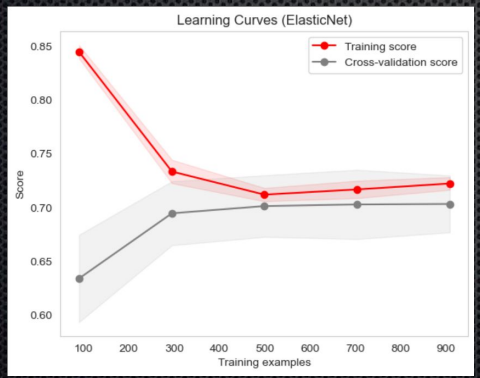
Elasticnet model

➔ The ElasticNet regression model combines L1 (Lasso) and L2 (Ridge) regularizations to enhance the performance of linear regression in the presence of collinear data.

➔ Hyperparameter grid search through cross-validation with GridSearchCV :

- alpha : [0.01 , 0.1 , 1.0]
- l1_ratio : [0.1 , 0.5 , 0.9]

Learning curve :



Prediction on energy consumption	ELASTICNET (scores on test set)
Best hyperparameterss	Alpha : 0.1, l1 ratio : 0.5
R^2	0.69
RMSE	0.99
MSE	0.98
Temps de calcul (s)	0.01

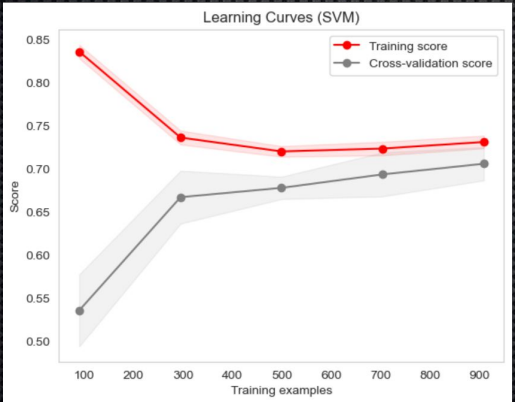
SVM model

➔ The Support Vector Machine (SVM) model is a supervised machine learning algorithm known for its ability to find the optimal hyperplane that separates classes in a high-dimensional space.

➔ Hyperparameter grid search through cross-validation with GridSearchCV :

- Kernel [linear, rbf, poly]
- C [0.01, 0.1, 10]
- Gamma [scale , auto]

Learning curve :



Prediction on energy consumption	Random Forest (scores on test set)
Best hyperparameters	Max_depth : 20, N_estimators : 300 Min_Sample_Leaf :4 ,min_sample split : 2
R^2	0.7
RMSE	0.97
MSE	0.93
Temps de calcul	0.12

Results of predictions on energy consumption

Predictions on energy consumption	Model							
	ElasticNet		Random Forest		XGBOOST		SVM	
	Transformed data	ACP	Transformed data	ACP	Transformed data	ACP	Transformed data	ACP
MSE	0.98	0.98	1.04	1.07	0.99	1.03	0.93	0.93
RMSE	0.99	0.99	1.02	1.04	0.99	0.99	0.97	0.97
R2	0.69	0.7	0.675	0.66	0.69	0.68	0.7	0.7
Processing time (s)	0.01	0.01	0.08	0.99	0.06	0.08	0.52	0.12

Results of predictions on energy consumption

Results of predictions on CO2 emission

Prédiction des EMISSIONS	Model							
	ElasticNet		Random Forest		XGBOOST		SVM	
	Transformed data	ACP	Transformed data	ACP	Transformed data	ACP	Transformed data	ACP
MSE	1.86	1.87	2.03	2.04	1.94	1.98	1.85	1.86
RMSE	1.36	1.37	1.42	1.43	1.39	1.39	1.36	1.37
R2	0.48	0.48	0.43	0.43	0.46	0.45	0.48	0.49
Processing time (s)	0.01	0.01	0.85	1.01	0.03	0.07	0.12	0.1

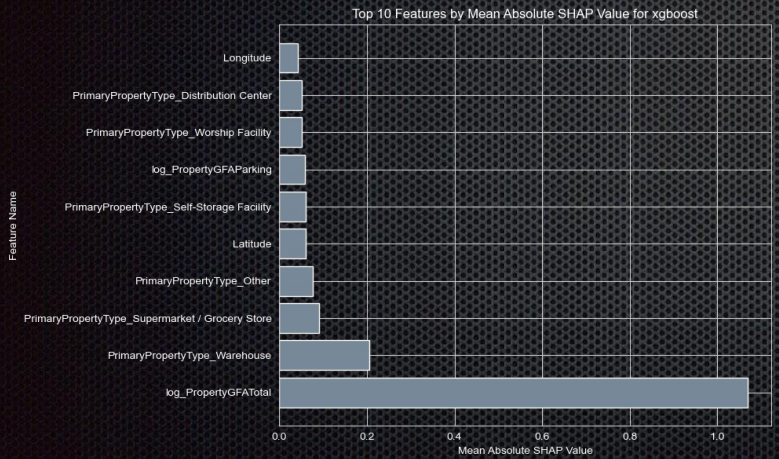
Transformed data - feature importance for energy consumption

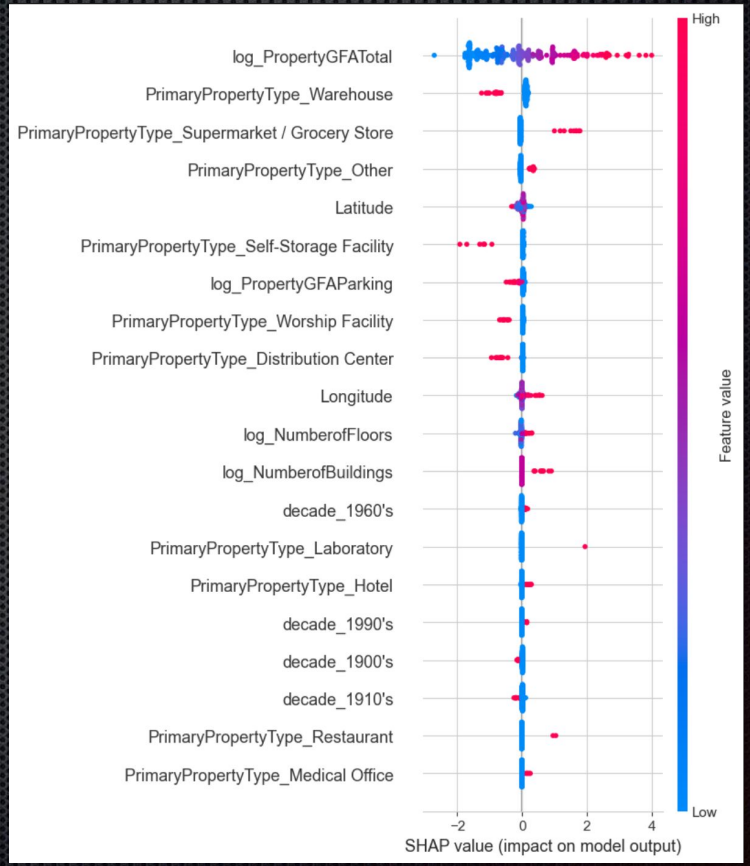
To explain the results, we use the SHAP method (SHapley Additive exPlanations).

It allows us to identify the variables that have had the most impact on the model predictions.

In this case, we visualize the results on the XGBOOST model as it has performed the best.

It is clear that the total surface area has the most significant impact on predicting energy consumption.





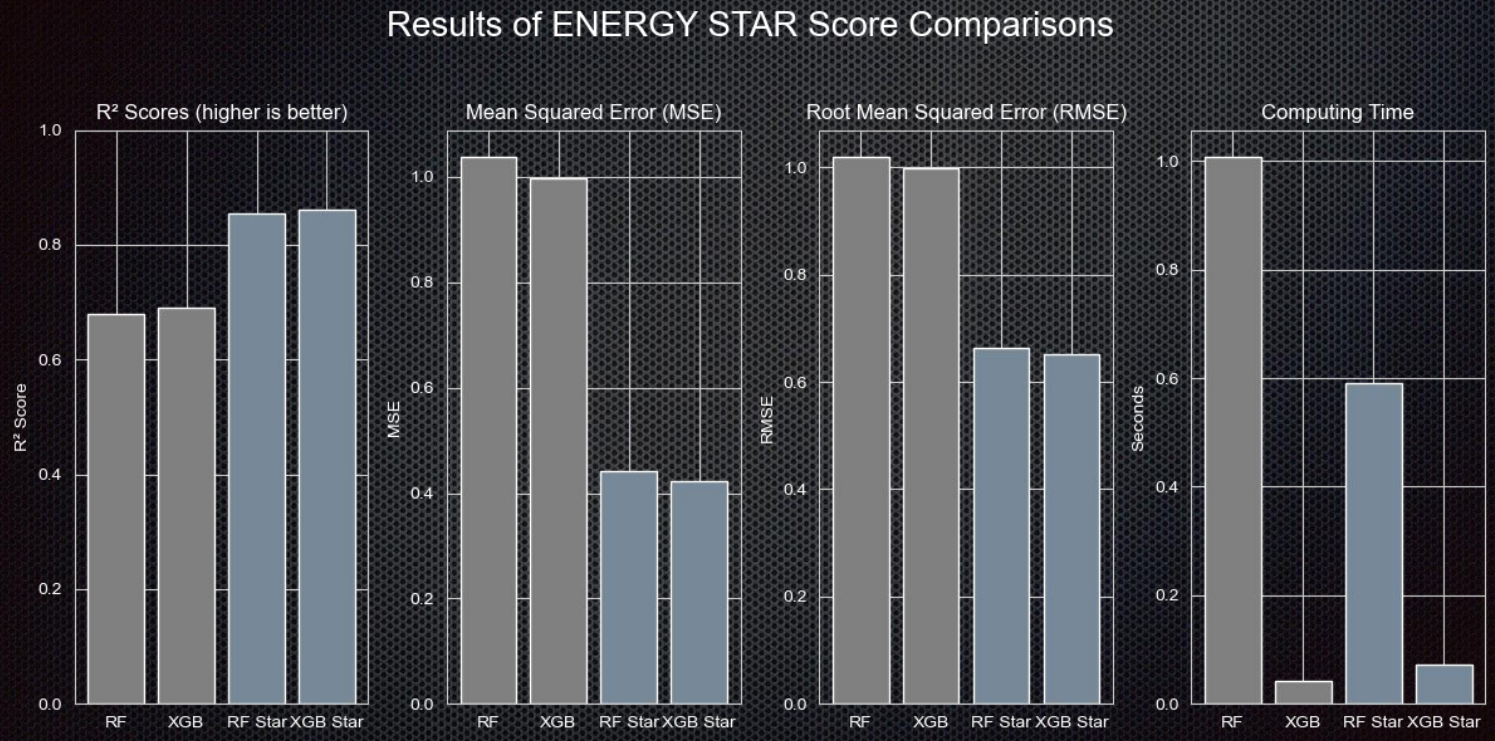
Energy star score addition

The ENERGY STAR Score ranges from 1 to 100 and assesses the energy performance of a building, considering its physical characteristics, usage, and occupants' behavior. Calculating this score involves a complex process. Therefore, we aim to evaluate its usefulness in predicting emissions.

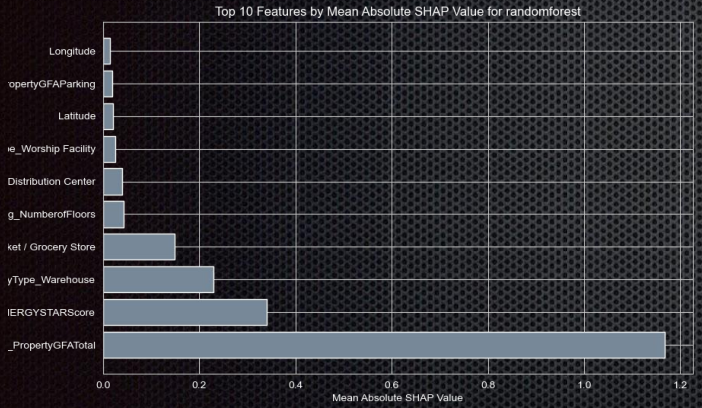
We will follow a similar methodology to our previous approach, incorporating the ENERGY STAR Score as an additional variable. Subsequently, we will compare the model's performances with and without this variable, both on transformed data and data subjected to dimensionality reduction. This comparison will allow us to assess the impact of adding the ENERGY STAR Score on prediction accuracy and its contribution to enhancing the prediction model.

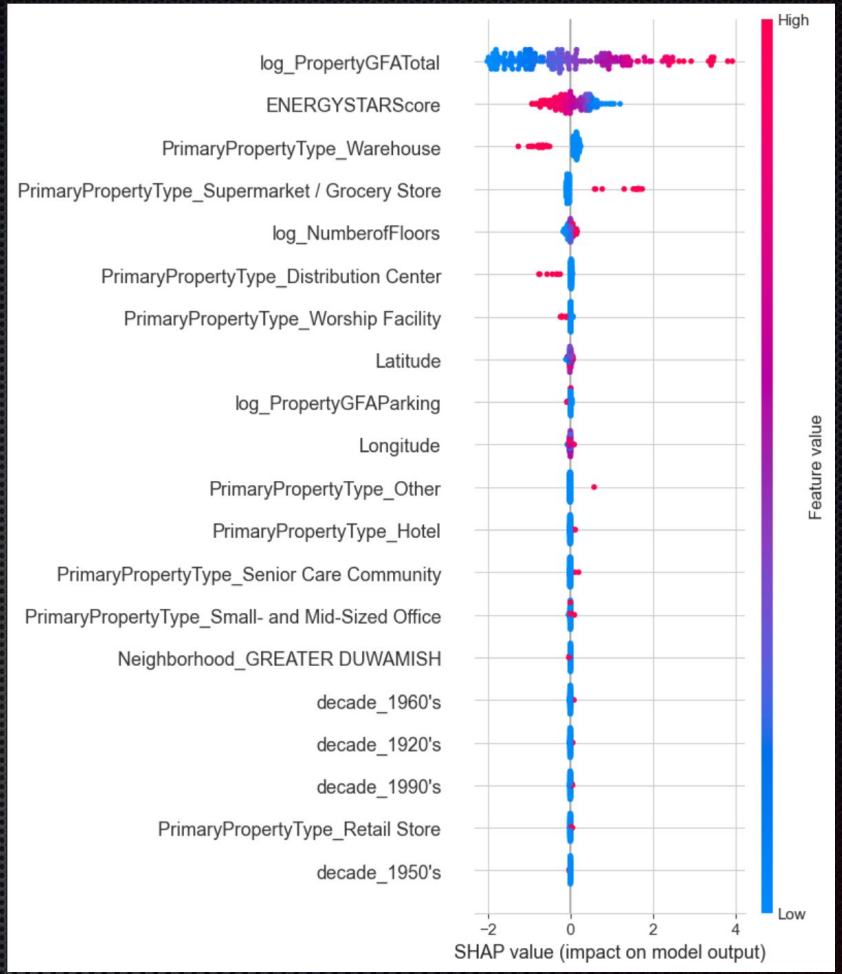
Results (on transformed data)

We observe a significant improvement in prediction performance, with a shorter computing time.



When we observe the feature importance to explain the model results, we notice that the Energy Star Score is in the 2nd position after the total surface area. We can conclude that it is a relevant variable.





In this analysis, we explored various models to predict energy consumption and CO2 emissions based on building characteristics. Our findings revealed significant insights into the predictive performance of different models and the impact of data transformations on model accuracy and computational efficiency.

Key Findings:

Transformation Effectiveness: The most notable improvement in prediction accuracy was observed when using transformed data. Raw data resulted in average performance, while data processed through principal component analysis (PCA) yielded predictions similar to those with transformed data. However, the computational time for PCA-transformed data was comparatively longer

Model Performance: Among the models tested, XGBoost demonstrated the best balance between computational efficiency and prediction accuracy. Its superior performance underscores its effectiveness in capturing complex relationships within the dataset.

Feature Importance: Analysis of feature importance revealed that floor area had the most significant impact on predictions. Additionally, building location emerged as a relevant factor, reflecting the influence of geographic variables such as proximity to water bodies and variations in temperature across different regions of Seattle.

Prediction Quality: While predictions for CO2 emissions were slightly less accurate than those for energy consumption, they still outperformed simple mean-based predictions. The variability in prediction quality underscores the importance of setting appropriate accuracy thresholds for evaluating model performance.

Energy Star Score Integration: Integrating the Energy Star score substantially improved prediction quality and computational efficiency for both CO2 emissions and energy consumption. However, the decision to implement this feature should be based on a careful assessment of the benefit-to-cost ratio.

In conclusion, our analysis demonstrates the effectiveness of advanced modeling techniques in predicting building energy consumption and CO2 emissions. By leveraging data transformations and integrating relevant features such as the Energy Star score, we can enhance prediction accuracy and computational efficiency. Moving forward, stakeholders should carefully consider the trade-offs between model complexity, computational resources, and predictive performance to optimize decision-making processes in energy management and environmental sustainability initiatives.