WOMAN IN DATA SCIENCE DATATHON

February 2022

TALITA COLL CÁRDENAS
MARÍA MELLADO PALACIOS
ISABEL RODRIGUEZ ROBLEDO

INDEX

Page

1) Main Goals……………………………………………. 2
2) Background……………………………………………. 2
3) Overview: the dataset and challenge…………………... 4
4) Exploratory Data Analysis (EDA)……………………. 6
5) Principal Component Analysis (PCA)………………… 12
6) Regression models……………………………………. 14
7) Hyperparameter tuning and cross validation…………... 14
8) Implementation……………………………………… 15
9) Conclusions……………………………………………. 17

Main Goals

Creation of a regression model that can predict building energy consumption. Accurate predictions of energy consumption can help policy makers target retrofitting efforts to maximize emissions reductions.

Background

Climate change is a globally relevant, urgent, and multi-faceted issue heavily impacted by energy policy and infrastructure. Addressing climate change involves mitigation (i.e. mitigating greenhouse gas emissions) and adaptation (i.e. preparing for unavoidable consequences). Mitigation of Greenhouse gas emissions (GHG) requires changes to electricity systems, transportation, buildings, industry, and land use.

According to a report issued by the International Energy Agency (IEA), the lifecycle of buildings from construction to demolition were responsible for 37% of global energy-related and process-related CO_2 emissions in 2020. Yet it is possible to drastically reduce the energy consumption of buildings by a combination of easy-to-implement fixes and state-of-the-art strategies. For example, retrofitted buildings can reduce heating and cooling energy requirements by 50-90 percent. Many of these energy efficiency measures also result in overall cost savings and yield other benefits, such as cleaner air for occupants. This potential can be achieved while maintaining the services that buildings provide.

<u>Site and Source EUI</u>

Site Energy Use Intensity (EUI) is an indicator of heat and electricity consumed by a building as reflected in a building's utility bills. Site EUI is a mixture of primary energy (such as fuel or gas) and secondary energy (electricity or steam). Source Energy is the total amount of raw fuel required to operate the building; basically, site energy plus all transmission, delivery, and production losses. Source EUI is the parameter used for Energy Star Rating calculation.

It can be calculated by the following equation:

EUI = Annual Energy Use / Area

Where:

EUI units: kbtu/sf/year*

Annual Energy Use units: kbtu/year*

Area: square feet

(Kbtu: kilo-British thermal unit)

*It can be also expressed in kWh/m^2

Energy Star Rating

Energy Star Rating (1-100 scale) provides information about your building´s energy performance, taking into consideration the building's physical assets, operations, and occupant behaviour. This score can be used to compare buildings or indicate the level of energy performance. Nevertheless, it cannot explain the poor or wellness of building performances

Overview: the dataset and challenge

The WiDS Datathon dataset was created in collaboration with Climate Change AI (CCAI) and Lawrence Berkeley National Laboratory (Berkeley Lab). The dataset consists of variables that describe building characteristics and climate and weather variables for the regions in which the buildings are located. The data is divided in two csv files, labelled as train and test. There are eight variables related to building characteristics that include year factor, state factor, building class, facility type, floor area, year built, energy star rating and elevation. Another 36 variables correspond to maximum, minimum and average temperature in each month and another variable for average temperature. Furthermore, five variables include cooling degree days, heating degree days, precipitation inches, snowfall inches and snow depth inches. Another eight variables include temperature below 0,10,20 and 30 F and above 80, 90, 100 and 110 F. To conclude, the last six variable are related to wind and fog and are labelled as direction maximum wind speed, direction peak wind speed, maximum wind speed, days with fog, site energy use intensity and id.

	Feature	Type
0	Year_Factor	int64
1	State_Factor	object
2	building_class	object
3	facility_type	object
4	floor_area	float64
5	year_built	float64
6	energy_star_rating	float64
7	ELEVATION	float64
8	january_min_temp	int64
9	january_avg_temp	float64
10	january_max_temp	int64
11	february_min_temp	int64
12	february_avg_temp	float64
13	february_max_temp	int64
14	march_min_temp	int64
15	march_avg_temp	float64
16	march_max_temp	int64
17	april_min_temp	int64

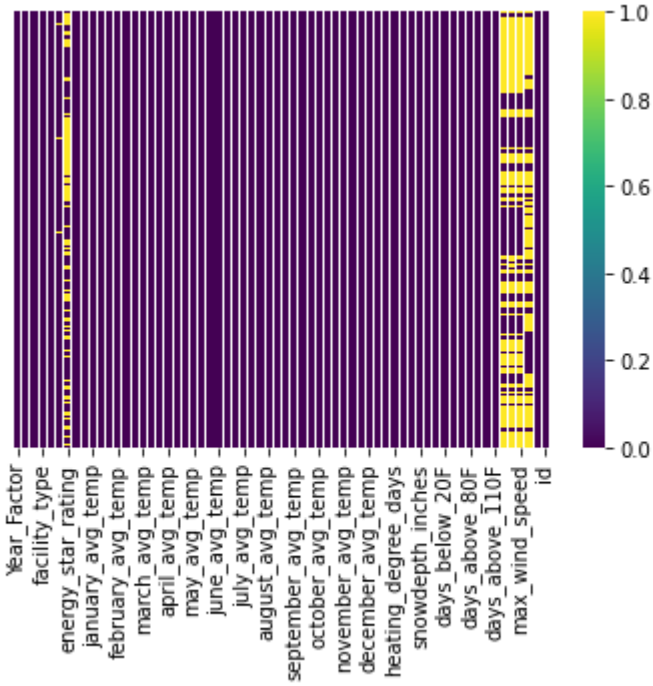
18	april_avg_temp	float64
19	april_max_temp	int64
20	may_min_temp	int64
21	may_avg_temp	float64
22	may_max_temp	int64
23	june_min_temp	int64
24	june_avg_temp	float64
25	june_max_temp	int64
26	july_min_temp	int64
27	july_avg_temp	float64
28	july_max_temp	int64
29	august_min_temp	int64
30	august_avg_temp	float64
31	august_max_temp	int64
32	september_min_temp	int64
33	september_avg_temp	float64
34	september_max_temp	int64
35	october_min_temp	int64
36	october_avg_temp	float64
37	october_max_temp	int64
38	november_min_temp	int64
39	november_avg_temp	float64
40	november_max_temp	int64
1	december_min_temp	int64
42	december_avg_temp	float64
43	december_max_temp	int64
44	cooling_degree_days	int64
45	heating_degree_days	int64
46	precipitation_inches	float64
47	snowfall_inches	float64
48	snowdepth_inches	int64
49	avg_temp	float64
50	days_below_30F	int64
51	days_below_20F	int64
52	days_below_10F	int64
53	days_below_0F	int64
54	days_above_80F	int64
55	days_above_90F	int64
56	days_above_100F	int64
57	days_above_110F	int64
58	direction_max_wind_speed	float64
59	direction_peak_wind_speed	float64
60	max_wind_speed	float64

61	days_with_fog	float64
62	site_eui	float64
63	id	int64

Exploratory Data Analysis (EDA)

<u>Missing Values imputation</u>

By inspecting the data structure, we can visualize via Heatmap of there are several columns with missing values.



<u>Figure 1:</u> Heatmap of Data Frame´s missing values.

Feature	Number of missing values	Percentage of missing values (%)
energy star rating	26709	35.63
year built	1837	2.42
direction max wind speed	41082	54.23
direction peak wind speed	41811	55.19
max wind speed	16488	71.87
days with fog	45796	60.45

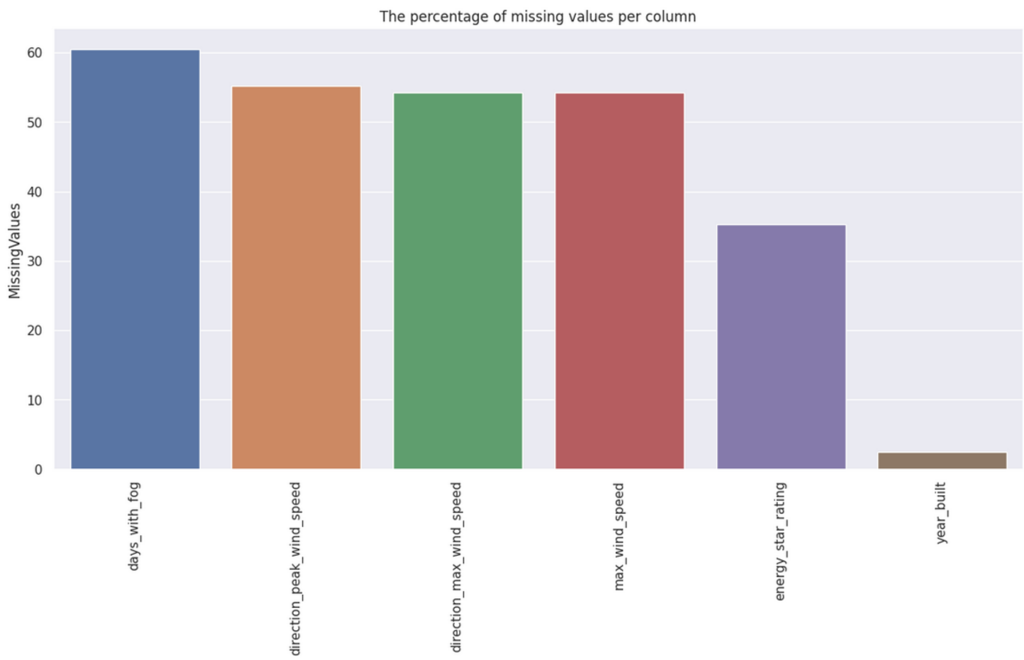


Figure 2: Percentage of missing values per column.

A closer look into the missing values has given information about ten features that contain missing values, in several of them these values are higher than the 50 % of the column values. The first approach for missing values imputation was eliminating columns with a percentage of missing values higher than 40 percent.

The second approach for missing values imputation has been a widely used method such as KNNImputer. The idea of this method is to identify 'k' samples to estimate the value of the missing data that are close in space. For this purpose, the Data Frame was first normalize using Min Max Scaler library from sklearn. Afterwards, KNNImputer was applied using 5 neighbours, uniform weight, and nan Euclidean metric.

Outlier imputation

Outlier analysis was first performed by looking into boxplot of the features. For simplicity, features are depicted in different figures.

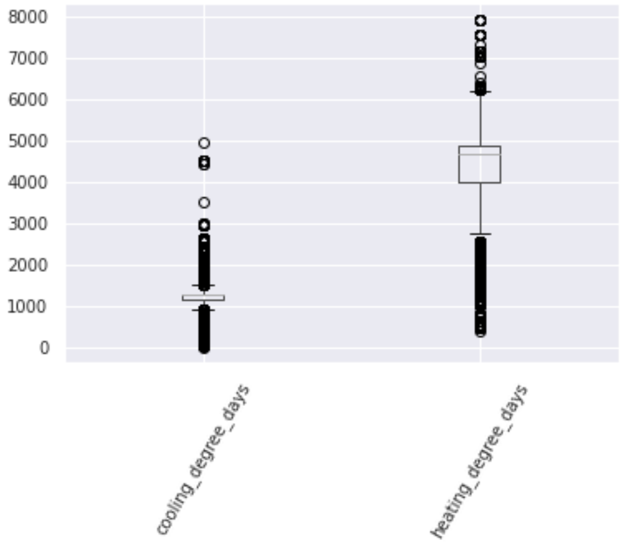


Figure 3: Boxplot representation of the data frame features cooling degree days and heating degree days.

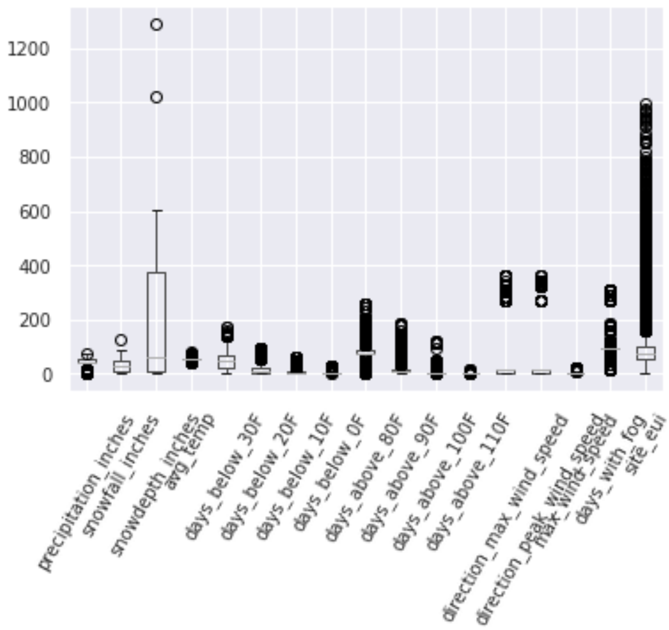


Figure 4: Boxplot representation of the data frame features related to direction of the wing, fog, precipitation, and snow.

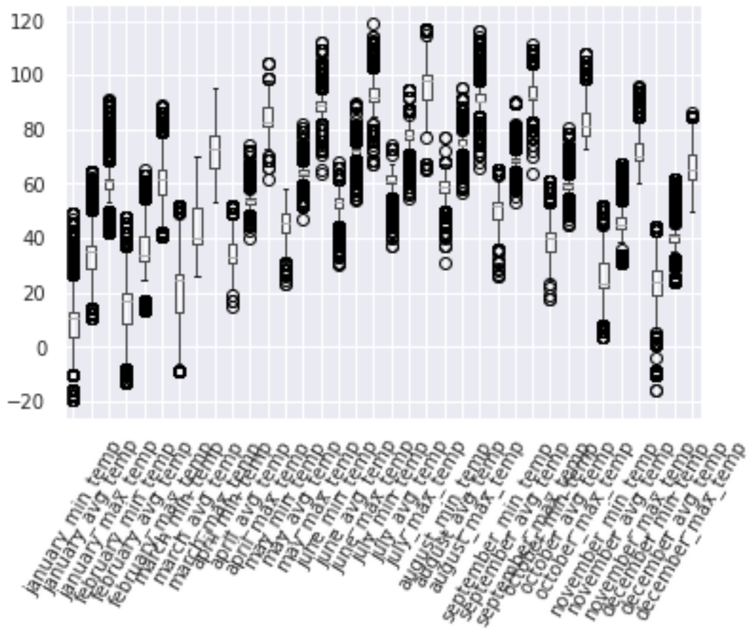


Figure 5: Boxplot representation of the data frame features related to average, minimum and maximum temperature in each month.

Discovering outliers using z-score

By this approach, the first step was calculating z-score for each column. Afterwards, outliers (with z-score lower than 10) where removed on a temporary data frame. The data frame shape changed by this approach as in:

Before removing outliers:(75000, 59)
After removing outliers:(74652, 59)

Categorical features

For further analysis of features, categorical features were represented in bar diagrams as followed in figures 6 to 9.

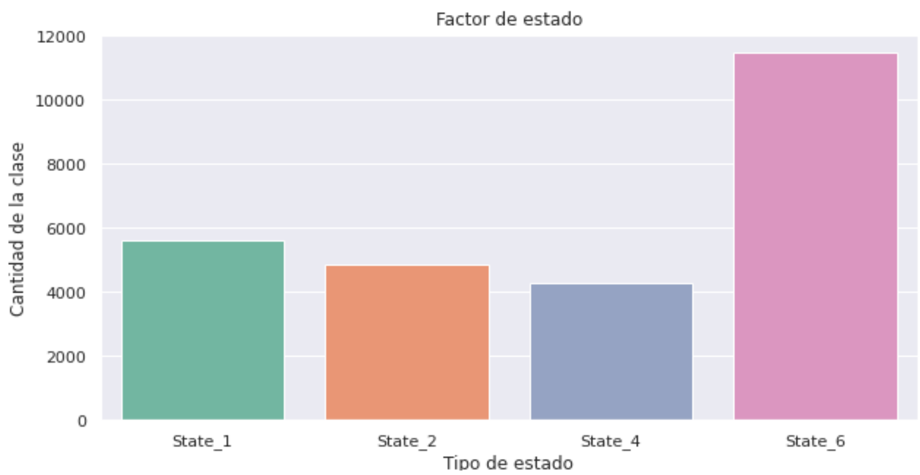
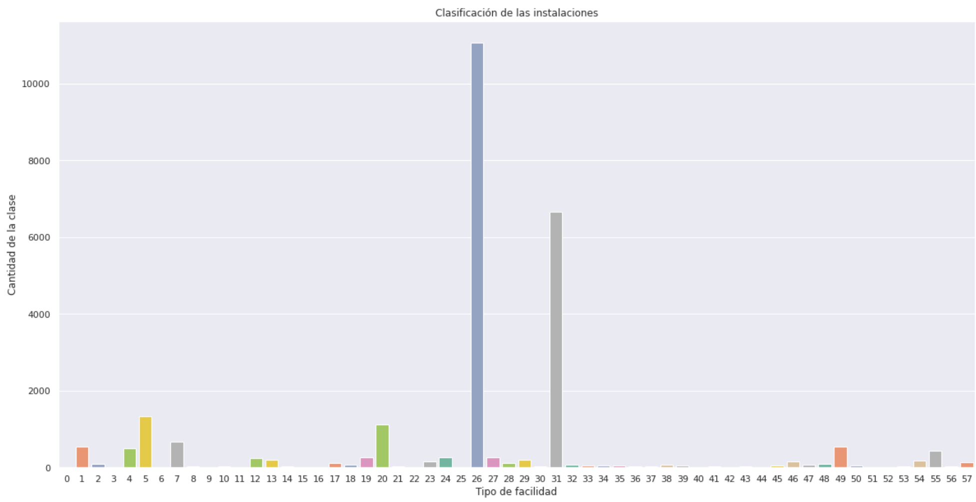
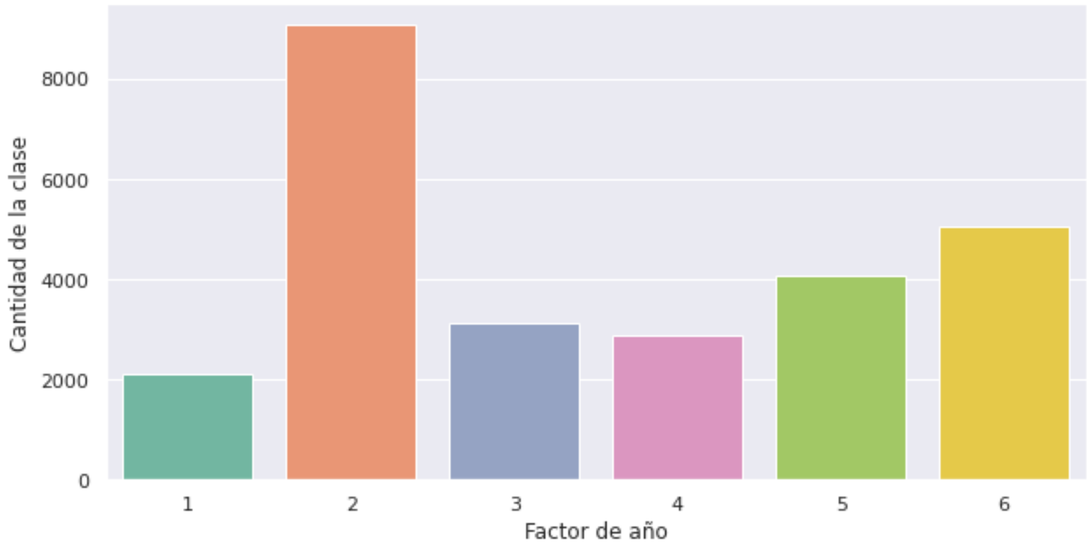


Figure 6: bar diagram of feature State Factor.

<u>Figure 7:</u> bar diagram of feature building use.



<u>Figure 8:</u> bar diagram of feature Facility Type.



<u>Figure 9:</u> bar diagram of feature Factor Year.

In order to perform future modelling, it was necessary to encode the following categorical features:

1) Facility type
2) State Factor
3) Building class

Feature selection

It is already well known that feature selection plays a huge role in machine learning. By studying correlation, it can be determined the linear relationship between two features. In the following figure a heatmap of the correlation matrix is depicted for correlation visualization.

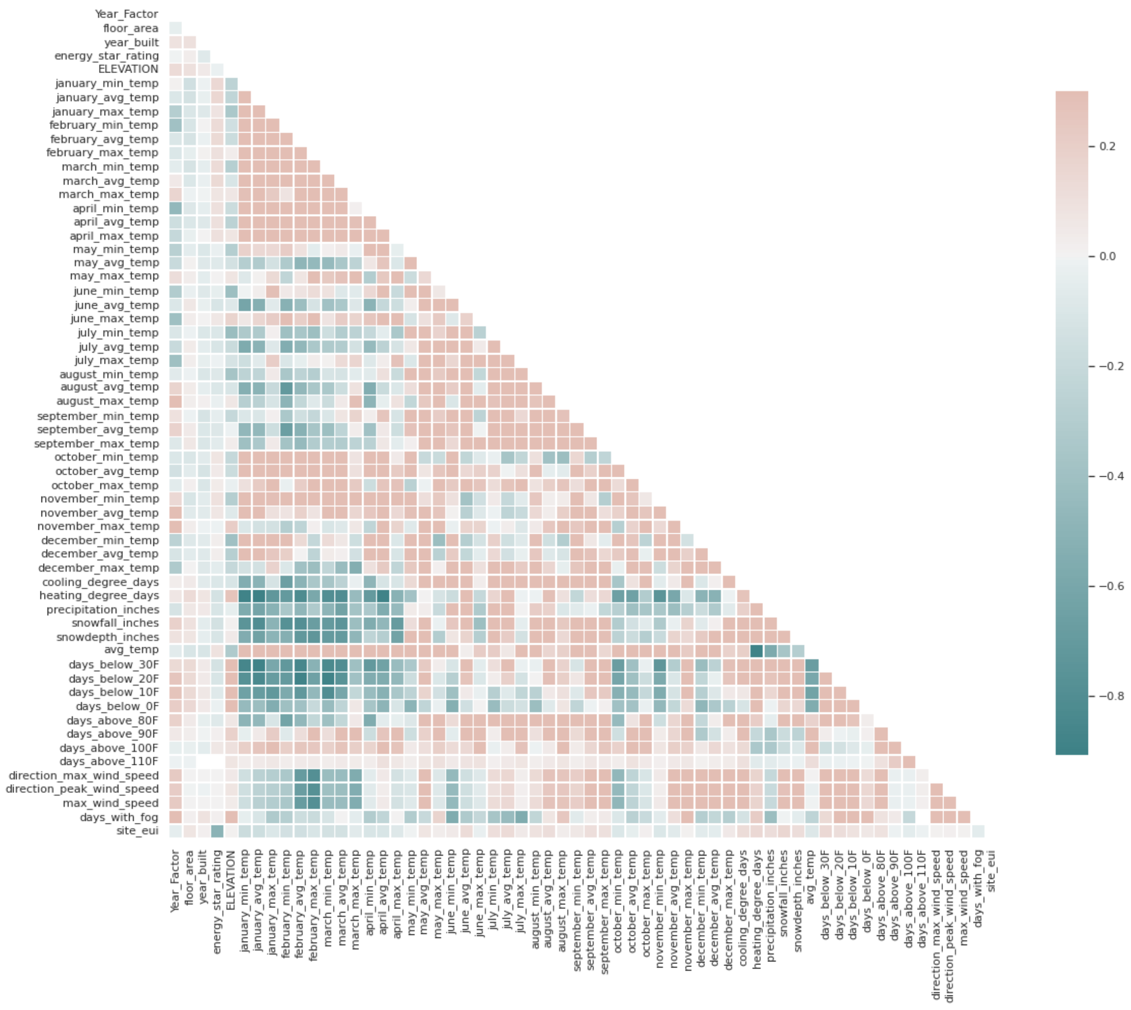


Figure 10: Correlation matrix of Data Frame features.

Feature importance using Random Forest Regressor

For feature selection, random forest regressor was used with 100 estimators. As a result feature importance is depicted in the following bar plot:

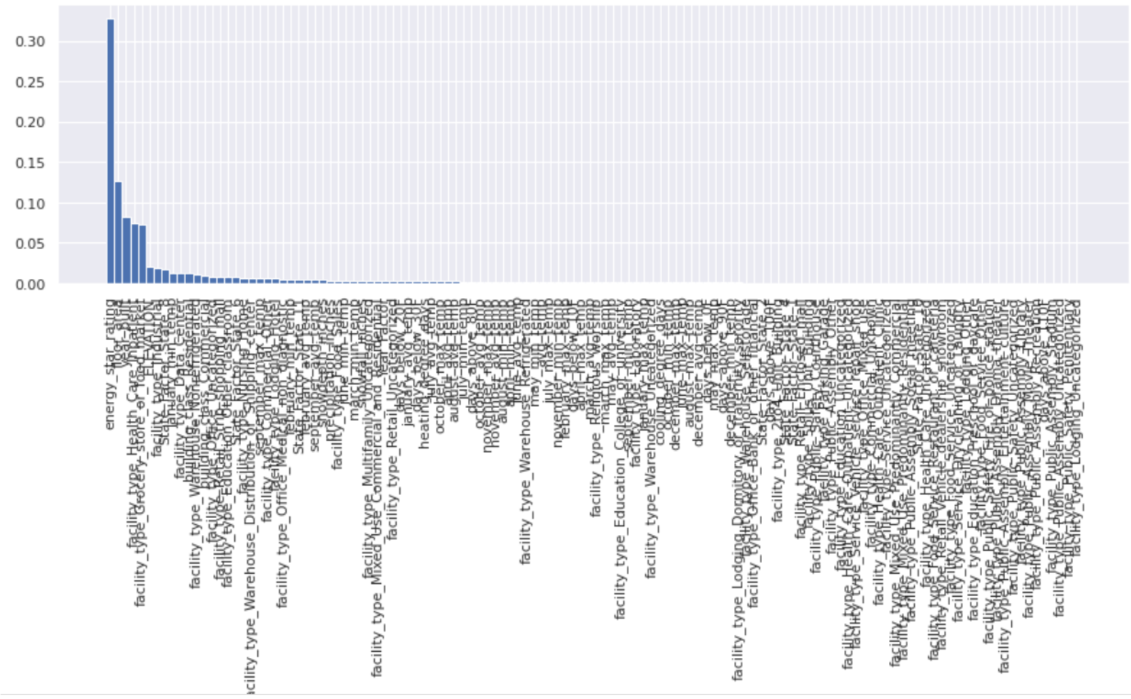


Figure 11: Features Bar plot arranged according to feature importance.

More important features:

- Energy star rating
- Year built
- Facility type
- Floor area

Principal Component Analysis (PCA)

Principal component Analysis (PCA) is often used for dimensionality reduction. It chooses the data in the direction of maximum variance. To perform this task, a pipeline has created, where the first step was the normalization of the data and the second step it is the PCA itself. Afterwards, a new data frame that contains the 64 principal components is created. For further understanding, as seen in the following figure, the heatmap of features and principal components is depicted.

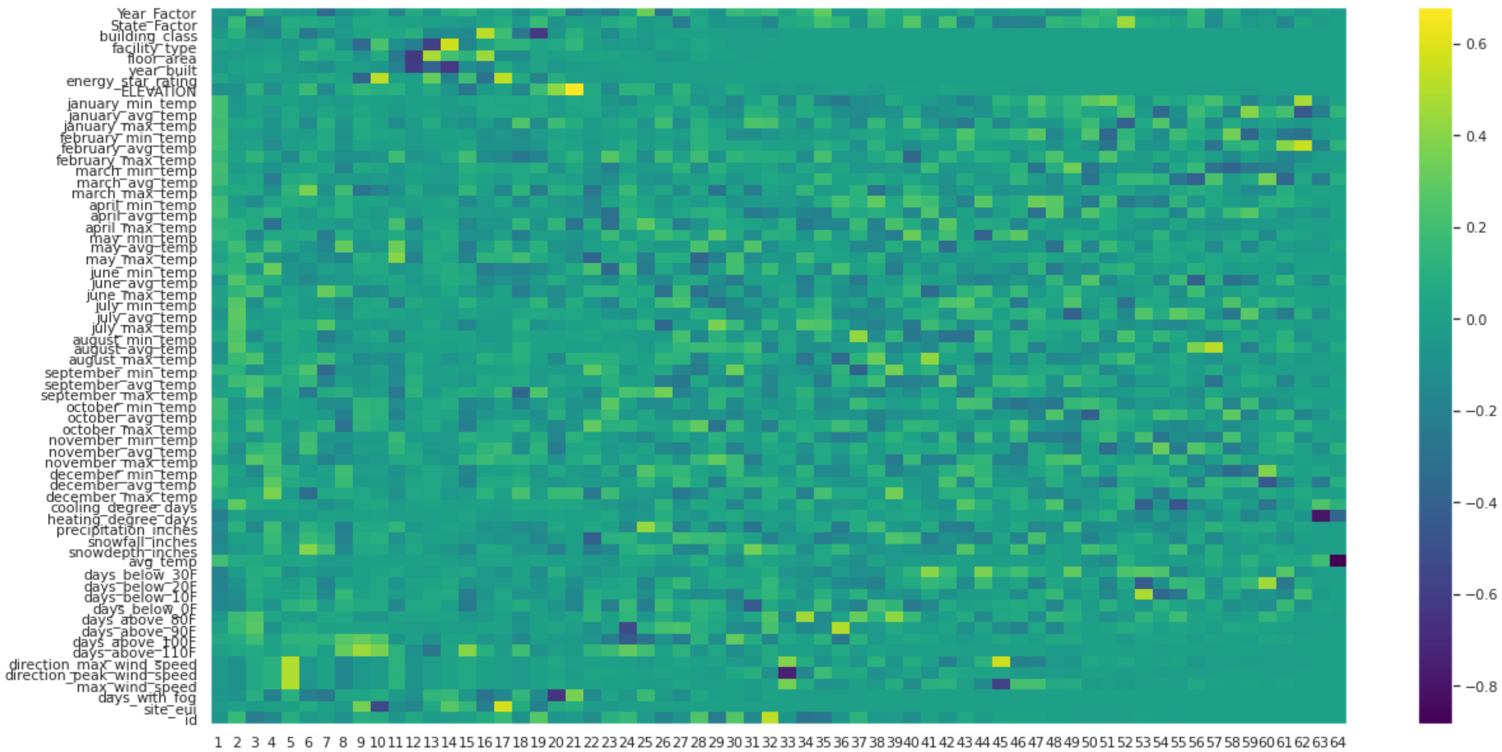


Figure 12: Principal Components heatmap of data frame features

In the following figure the contribution of each component to the variance is shown. From this figure, it can be concluded that 34 % of the variance can be explained with principal component 1 and 16 % with component 2. I the next figure, from accumulated variance, it can be concluded that 99 % of variance can be explained with 23 principal components.

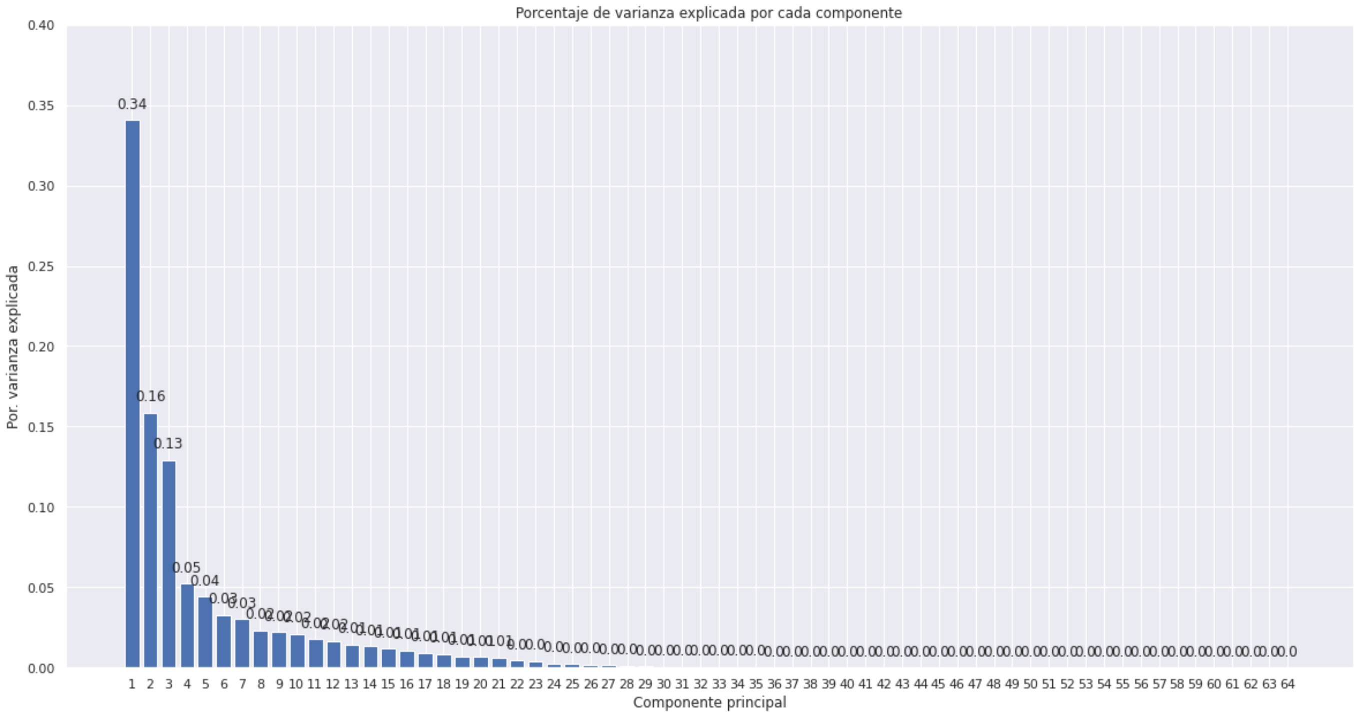
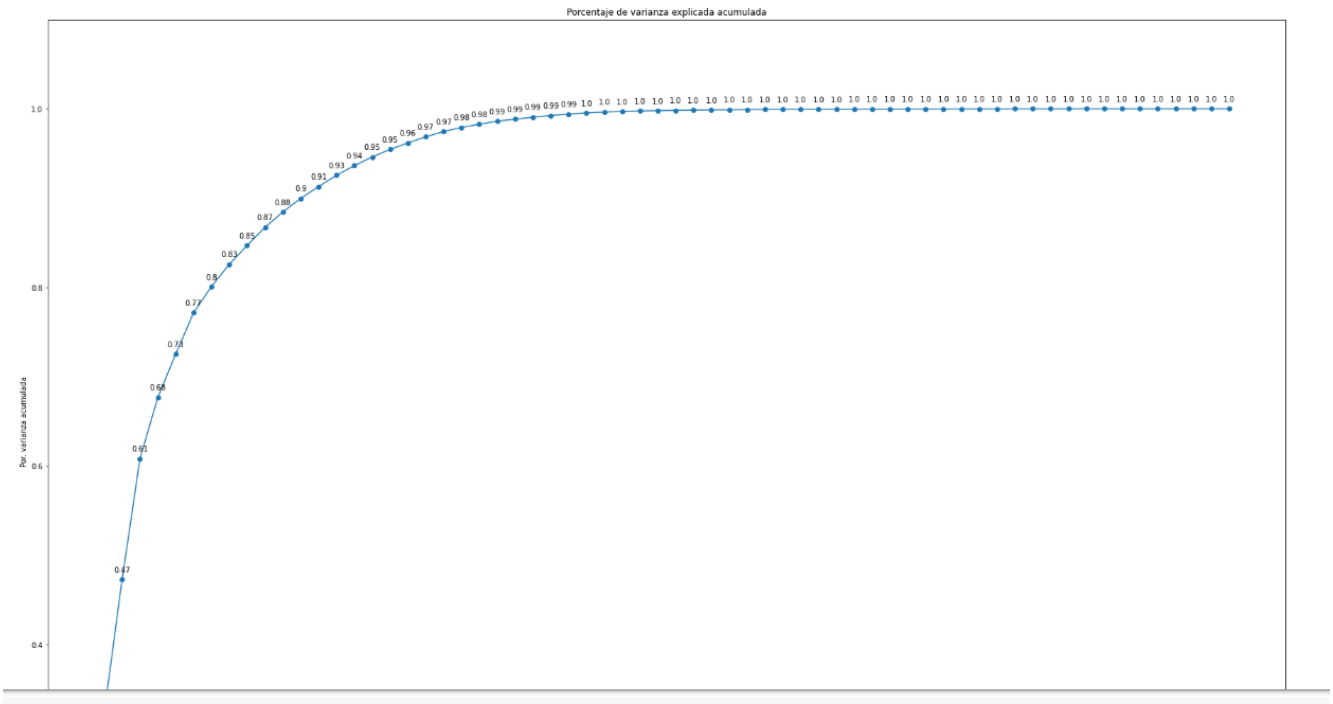


Figure 13: PCA principal components.



<u>Figure 14:</u> accumulated variance

Regression models

The following algorithms were used to analyse the data, followed by the error obtained with each one:

Model	RMSE
Random Forest Regressor	36.01
XGB Regressor	33.64
Lasso	35.94
Support Vector Regressor	35.45
Extra Tree Regressor	34.33
Support Vector Regressor	35.65
Cat Boost Regressor	33.49
Gradient Boosting Regressor	33.59

Hyperparameter tuning and cross validation

Hyperparameter tunning of XGB Regressor. Parameters tunned:

nthread:4
objectiv: reg:linear
learning rate: .03, 0.05, .07
max depth: 5, 6, 7
min child weight: 4
silent: 1
subsample: 0.7
col sample by tree: 0.7
n_estimators:500

Grid Search CV was performed with this parameter, cv= 2 and n jobs = 5. Two folds of 9 candidates results in a total of 18 fits. The best parameters were:

Col sample by tree: 0.7
Learning rate: 0.03,
Max depth: 5
Min child weight: 4
N estimators: 500
N thread: 4
objective: reg: linear
silent: 1
subsample: 0.7

Implementation

Metrics

Mean Absolute Error (MAE): is the defined as the average of the sum of absolute difference between the actual values and the predicted values. This type of metric is not sensitive to outliers.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|yi - \widehat{yi}|$$

Where:

y_i: predicted value

y_i ^: actual value

Mean Square Error (MSE): average of the sum of square of the difference between actual and predicted values. Useful when dataset contains outliers.

$$MSE = \frac{1}{n}\sum(y - \widehat{y})^2$$

Root Mean Square Error (RMSE): defined as the root of the MSE. RMSE is more sensitive to the presence of outliers. Unlike MSE, Root Mean Square Error has the same unit of quantity plotted on vertical axis or y-axis.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(y - \widehat{y})^2}{n}}$$

R^2 score: coefficient of determination. It indicates how closer are the predicted values to the actual values.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y - \widehat{y})^2}{\sum_i (y - \bar{y})^2}$$

Where:

SS_{res}: Sum of Square of Residuals

SS_{tot}: Total Sum of Squares

The R^2 value ranges from $-\infty$ to 1. A model with negative R^2 value indicates that the best fit line is performing worse than the average fit line.

Adjusted R^2: modified form of R^2 that penalizes the addition of new independent variable only increases if the new independent variable or predictor enhances the model performance.

$$Adjusted\ R^2 = 1 - (1 - R^2) * \frac{n-1}{n-k-1}$$

R^2: R^2 Score

n: Number of Samples in Dataset

k: Number of Predictors

In this work, prediction was made on file "test.csv". In the data frame created, the same categorical features as in train file were encoded. Columns where missing values were deleted except energy star rating (missing values were filled randomly) and year built (they were filled with the mode). The regression algorithm used was Cat Boost Regressor with 50 iterations, depth of 3, learning rate of 1 and loss function RSME. The resulting prediction was saved in a csv file and submitted to Kaggle with two columns, "id" and "site_eui". Attending to the Root Mean Square Value (RMSE) obtained by this method, which was 33.49, is the reason for the algorithm election. By Cat Boost Regression the lowest value of RMSE was obtained of all the regression algorithms tried.

Conclusions and Future Perspectives

Studies on building energy consumption and its characteristics is essential for carrying out benchmarking processes and for decision making. To facilitate decision making around this topic, in this work, machine learning models were created to predict buildings energy consumption. For this purpose, it was observed that missing value treatment changed enormously the outcome RSME of the model. Furthermore, model performance is highly determined by feature selection. Random Forest, XG Boost and Cat Boost were the regression algorithms that lead to lowest RMSE. The lowest RSME was obtained with Cat Boost regressor and by imputing missing values with KKNImputer, obtaining a RSME value of 33.49.

Bibliography

- https://www.energystar.gov/buildings/benchmark/analyze_benchmarking_results
- Mel Keytingan M. Shapi, Nor Azuana Ramli, Lilik J. Awalin, Energy consumption prediction by using machine learning for smart building: Case study in Malaysia, Developments in the Built Environment, Volume 5,2021,100037, ISSN 2666-1659.
- Mohammadiziazi, R.; Bilec, M.M. Application of Machine Learning for Predicting Building Energy Use at Different Temporal and Spatial Resolution under Climate Change in USA. *Buildings* **2020**, *10*, 139. https://doi.org/10.3390/buildings10080139