Appliance Energy Prediction with Supervised Machine Learning

Abdoulfatah Abdillahi

Stats Learning & Data Mining

May 8th, 2025

PROJECT GOALS

- **Predict energy consumption** for appliances by utilizing indoor sensor data alongside external weather information.
- **Investigate time-based patterns** in energy usage, focusing on variations across hours, days, and weekdays compared to weekends.
- **Clean and preprocess the dataset** by addressing outliers and encoding categorical features.
- **Identify key predictive features** through ANOVA F-test and Variance Inflation Factor (VIF) analysis.
- **Compare various regression models**, including Linear, Ridge, Lasso, Decision Tree, and Random Forest.
- **Select the optimal model based** on metrics such as RMSE, MAE, R^2, and MAPE.
- **Examine the influence:** of weather conditions and room-specific variables on energy consumption.
- **Emphasize the importance** of atmospheric pressure and specific room data (e.g., kitchen, laundry) in making accurate predictions.

PREPROCESSING PIPELINE: CLEANING, TEMPORAL FEATURES & CATEGORIZATION

Step 1: Original Dataset

- **19,735** rows and **29** columns
- No missing values
- No duplicate rows
- Removed the **lights** column as it was an irrelevant feature
- All variables are numeric (except date)

Step 2: Date Column: Time Features Extracted

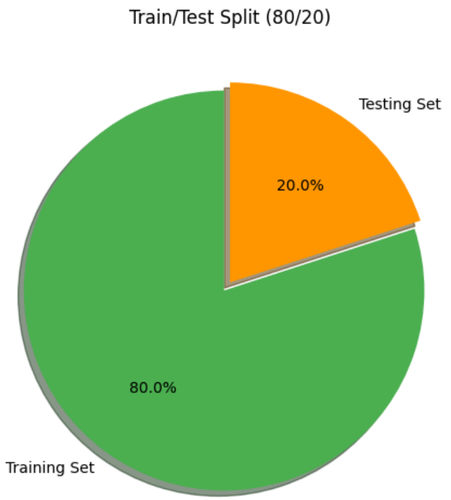
- Hour
- Day of the Month
- Day of the Week
- Month
- Month Name
- Day Name

Step 3: Categorical Time Blocks

- 6:00 AM – 12:00 PM: Morning
- 12:00 PM – 6:00 PM: Afternoon
- 6:00 PM – 12:00 AM: Evening
- 12:00 AM – 6:00 AM: Night

Sensors | Random | Target | Time Features

ıte	T1	RH_1	Appliances	hour	day_name
)16–0-11 17	47.596	47.568	13.275433	17	Monday
)16–0-12 17	45.487	47.436	0	17	Monday
)16–0-12 17	46.333	47.022	18.330886	17	Monday
)16–0-12 17	46.633	46.924	24.646068	17	Monday
)16–0-12 12	47.846	47.284	36.268866	17	Monday
)16–0-19 17	44.518	44.518	10.042097	17	Saturday

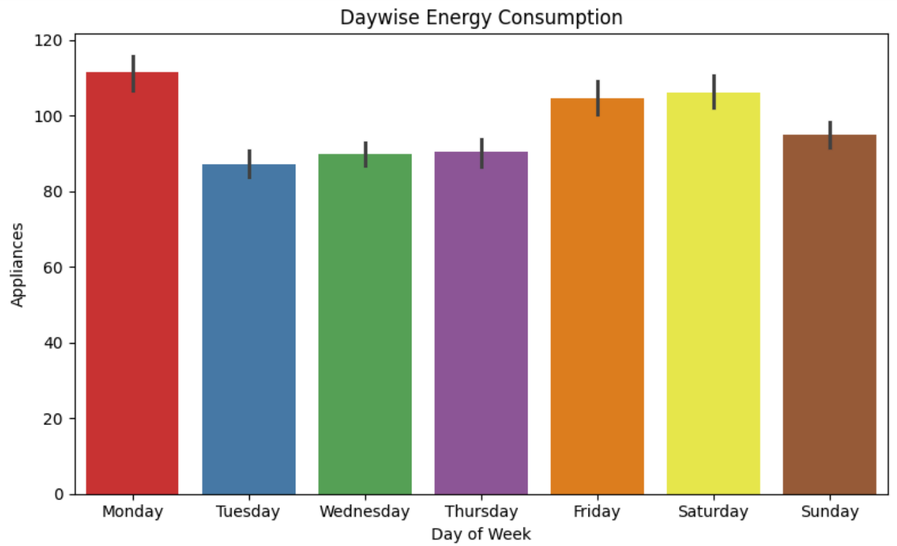


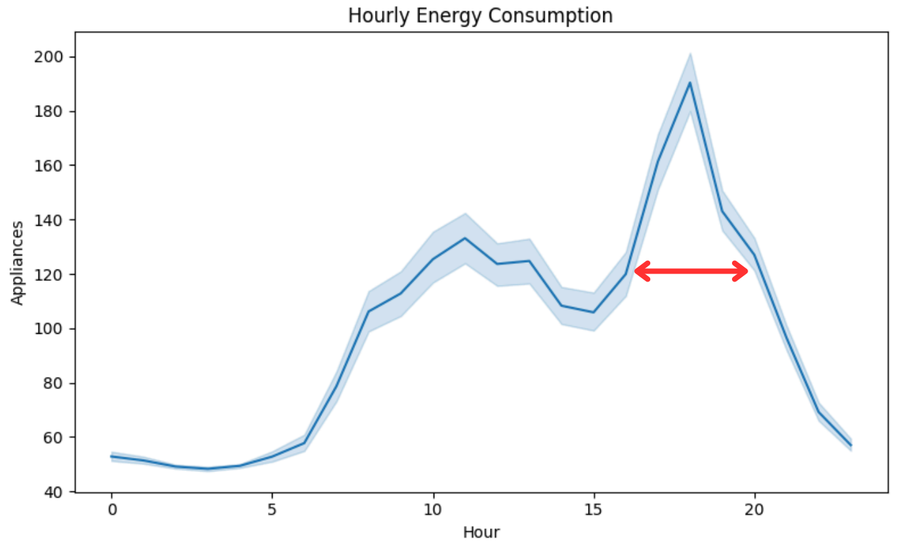
Step 4: Split into 80% training (15,788 samples) and 20% testing (3,947 samples)

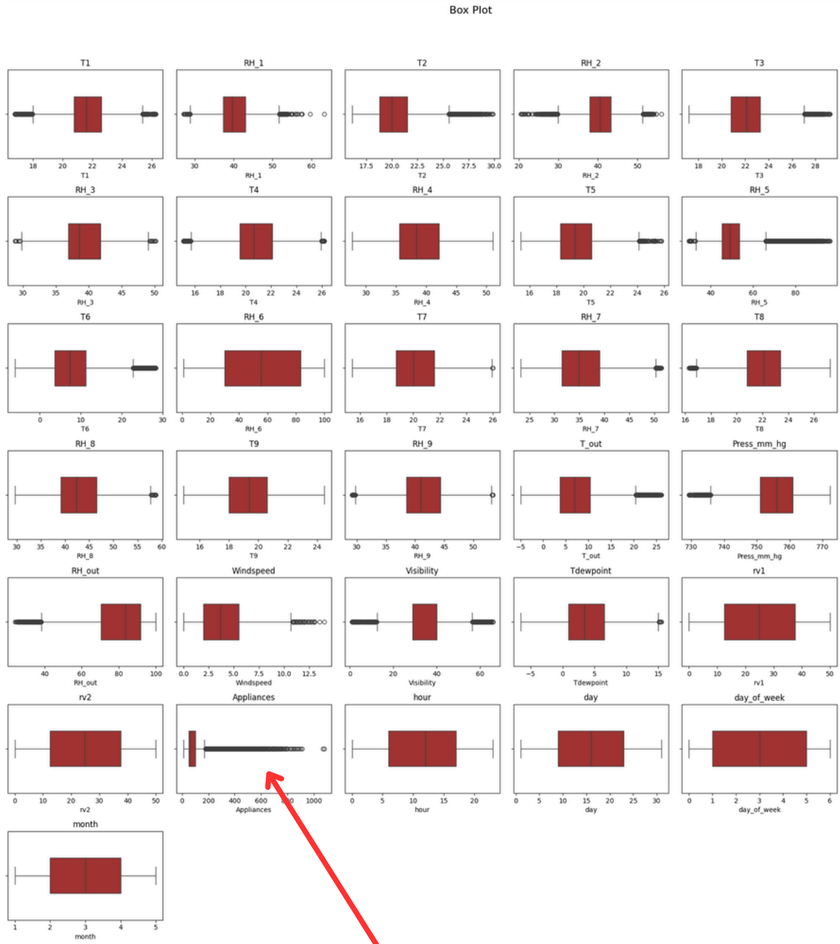
EXPLORATORY DATA ANALYSIS

Data Insights

- **No missing values**; the dataset is clean and prepared for modeling.
- The target variable **Appliances** is right-skewed, as seen in the box plot.
- **Outliers** are visible in many features, especially in sensor readings like Windspeed, T_out, and RH_1.
- Energy consumption **peaks in the evening**, particularly between 18:00–20:00, based on hourly plots.
- Monday shows the **highest average energy usage,** while midweek days like Tuesday–Thursday are lower.
- Most variables deviate from a **normal distribution,** as supported by Q-Q and box plots.







FEATURE SELECTION & DIMENTIONAL REDUCTION

Outlier Treatment Numerical Features

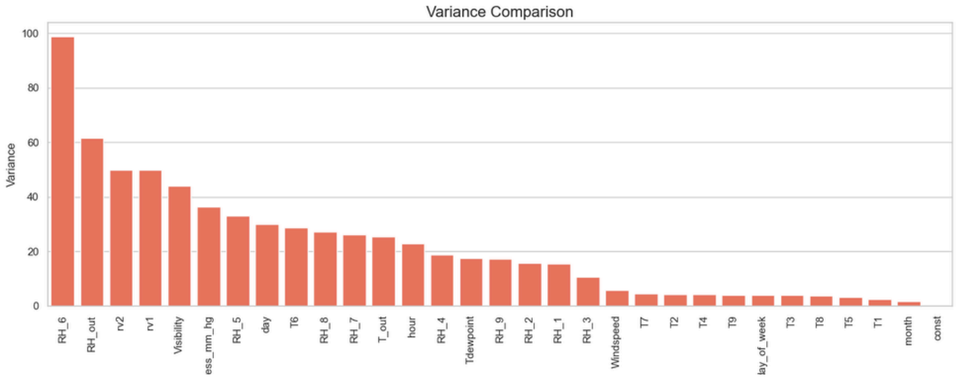
- Applied the **IQR method** to identify and address outliers in numerical features.
- For each feature:
  - Calculated Q1 (25th percentile) and Q3 (75th percentile).
  - Established bounds: [Q1 – 1.5×IQR, Q3 + 1.5×IQR].
- Capped values below the lower bound and above the upper bound.
- Reduced the impact of extreme values on model training.

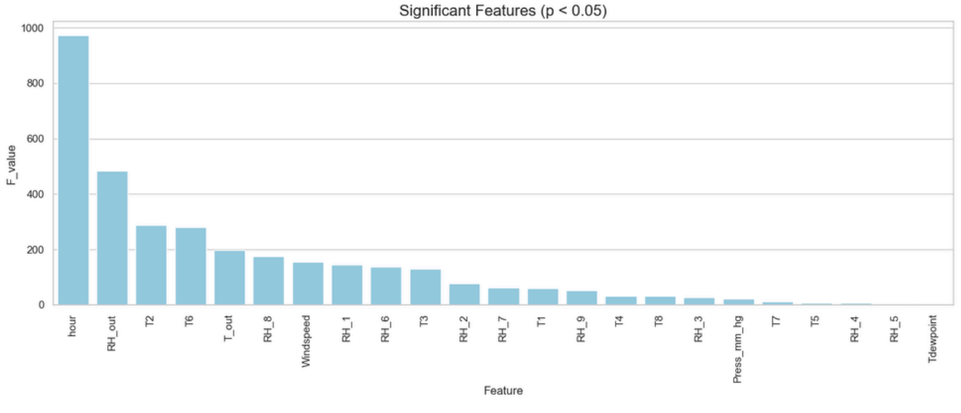
One-Hot Encoding for Categorical Features

- Applied one-hot encoding to convert categorical features to binary.
- Used `pd.get_dummies()` with `drop_first=True` to avoid multicollinearity.
- Created separate 0/1 indicator columns for each category.
- Retained (n−1) columns per feature to prevent the dummy variable trap.
- Enabled direct use of categorical features in regression models.

Dimentional Reduction

- **Variance Inflation Factor (VIF):**
  - Used VIF to measure multicollinearity among independent variables.
  - High VIF values indicate stron correlation between predictors.
  - Helped identify and manage redundant features to improve model stability
- **Annova (F-test):**
  - Applied an ANNOVA-based F-test to assess each feature's relationship with target (Appliances)
  - Retained only features with p-values < 0.05, indicating statistical significance.
  - Helped eliminate irrelevant or weak predictors, reducing model noise.

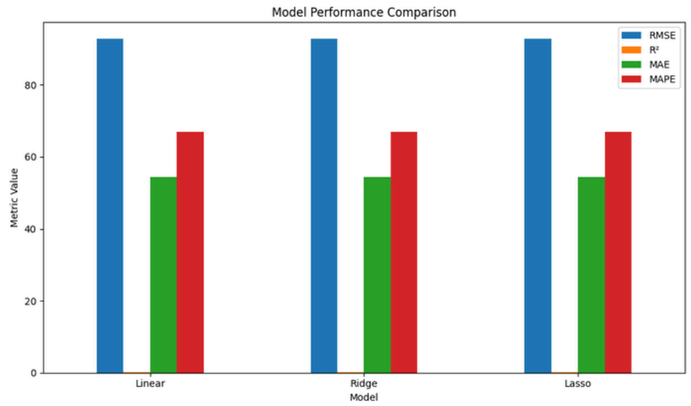




MODEL TRAINING & EVALUATION

Linear Regreesion

- I used Linear Regression as a baseline model to understand general trends in energy consumption.
- I applied Ridge and Lasso techniques to regularize my linear model, helping to avoid overfitting. Although the performance metrics were comparable across all three models, the regularized versions provide greater stability.
- All the regression models were trained with 10-fold cross-validation to select best.



Decision Tree

- I employed a Decision Tree Regressor to explore non-linear relationships.
- Utilized GridSearchCV to optimize the hyperparameters for the Decision Tree Regressor.
- Selected the best combinations: max_depth: [3, 5, 10, None], min_samples_split: [2, 5, 10], and min_samples_leaf: [1, 2, 4].
- Achieved enhanced performance compared to previous models.
- This model successfully captures non-linear patterns and interactions, surpassing linear models.
- To prevent overfitting, the maximum tree depth was limited to 5.

Random Forest

- Implemented to improve prediction accuracy by combining multiple decision trees.
- Used 100 estimators with default depth and parallel processing (n_jobs = -1).
- Achieved the best overall performance across all tested models.
- Among all the models tested, Random Forest yielded the most accurate and reliable results for predicting energy use.

MODEL EVALUATION COMPARISON

Model	RMSE	MAE	R^2	MAPE (%)
Linear Regression	92.75	54.45	0.14	66.96
Ridge Regression	92.75	54.45	0.14	66.95
Lasso Regression	92.74	54.44	0.14	66.95
Decision Tree	90.72	50.0	0.18	58.58
Grid Search Tree	85.04	37.94	0.28	35.65
Random Forest	63.1	29.55	0.6	29.54

CONCLUSION

- Explored 5 regression models: Linear, Lasso, Ridge, Decision Tree, Random Forest.
- Performed data preprocessing including outlier treatment, feature selection, and encoding.
- Compared all models using evaluation metrics: RMSE, MAE, R^2, and MAPE.
- Linear-based models assigned negligible weights to random variables, indicating minimal predictive power.
- Random Forest Regressor achieved the best performance with an R^2 score of 0.64, indicating strong predictive capability.

Thank you!

QUESTIONS ?