Future Prediction and Time Series Analysis of Energy Consumption

Abstract

The escalating need for sustainable energy consumption drives research into predictive models for household energy usage. This project investigates energy consumption patterns using the Appliances Energy Prediction dataset. The dataset encompasses over 19,735 instances collected at 10-minute intervals over 4.5 months. Leveraging machine learning techniques, the project aims to model and forecast energy usage while exploring dependencies on environmental factors. Results highlight the efficacy of advanced models like LSTM and Random Forest over traditional methods. The study emphasizes the novelty of combining feature engineering with sequential modeling for energy prediction tasks. The findings underscore the importance of temporal patterns in predicting appliance energy usage, paving the way for future research and practical applications in energy management systems.

Introduction

Energy efficiency in households is a key pillar of sustainability, given the rising global demand for energy resources. Accurate prediction of household energy consumption allows for better energy management, cost savings, and reduced environmental impact. This project focuses on creating robust predictive models that leverage historical energy usage data and associated environmental factors to predict future energy consumption. The Appliances Energy Prediction dataset is particularly suited for this study, as it contains detailed records of appliance energy usage, temperature, humidity, and weather conditions. These factors play a significant role in influencing energy consumption patterns, especially during peak and off-peak hours. The primary goal of this study is to explore and identify patterns, establish predictive models, and evaluate their effectiveness.

This report provides an in-depth analysis of various machine learning models, ranging from simple linear regression to advanced LSTM neural networks. By validating the effectiveness of these models, the study aims to contribute insights into both academic research and practical implementations of predictive analytics in energy systems.

Dataset

The dataset for this study is sourced from the UCI Machine Learning Repository. It consists of 19,735 samples collected over 4.5 months at 10-minute intervals. The dataset was collected by sensors placed inside the house and outside readings came from the nearby weather station. The main attributes are temperature, humidity and pressure readings. Each observation measures electricity in a 10-minute interval. The temperatures and humidity have been averaged for 10-minute intervals.

Independent variables : 28(11 temperature, 10 humidity, 1 pressure, 2 randoms)
Dependent variable : 2 (Appliances, Lights)
Each record includes 29 attributes, with the target variable being "Appliances," representing the energy consumption in Wh.

Key predictors include:

Temperature (T): Indoor and outdoor temperatures measured in Celsius.

	T1	T2	T3	T4	T5	T6
count	19735.0	19735.0	19735.0	19735.0	19735.0	19735.0
mean	21.687	20.341	22.268	20.855	19.592	7.911
std	1.606	2.193	2.006	2.043	1.845	6.09
min	16.79	16.1	17.2	15.1	15.33	-6.065
25%	20.76	18.79	20.79	19.53	18.278	3.627
50%	21.6	20.0	22.1	20.667	19.39	7.3
75%	22.6	21.5	23.29	22.1	20.62	11.256
max	26.26	29.857	29.236	26.2	25.795	28.29

Fig1

The temperature table helps identify the variability and range of indoor temperatures, which are critical for modeling energy consumption patterns. For instance, rooms with higher temperature fluctuations may require more energy for heating or cooling.

Humidity (RH): Indoor and outdoor relative humidity percentages.

	RH_1	RH_2	RH_3	RH_4	RH_5
count	19735.0	19735.0	19735.0	19735.0	19735.0
mean	40.26	40.42	39.243	39.027	50.949
std	3.979	4.07	3.255	4.341	9.022
min	27.023	20.463	28.767	27.66	29.815
25%	37.333	37.9	36.9	35.53	45.4
50%	39.657	40.5	38.53	38.4	49.09
75%	43.067	43.26	41.76	42.157	53.663
max	63.36	56.027	50.163	51.09	96.322

Fig2

The humidity table is essential for understanding environmental comfort and its impact on appliance usage. High variability in humidity could correlate with increased usage of dehumidifiers or HVAC systems.

Weather Conditions: Data such as visibility, pressure, and wind speed.

Future Prediction and Time Series Analysis of Energy Consumption

	T_out	Tdewpoint	RH_out	Press_mm_hg
count	19735.0	19735.0	19735.0	19735.0
mean	7.413	3.761	79.75	755.523
std	5.318	4.195	14.901	7.399
min	-5.0	-6.6	24.0	729.3
25%	3.67	0.9	70.333	750.933
50%	6.92	3.43	83.667	756.1
75%	10.4	6.57	91.667	760.933
max	26.1	15.5	100.0	772.3

Fig3

This table provides insights into external environmental conditions that directly affect energy consumption. For example, outdoor temperature and humidity influence the demand for heating, cooling, and ventilation systems.

Time Features: Day of the week, hour of the day, and whether the record falls within a weekday or weekend.

Feature ranges
Temperature : -6 to 30 deg, Humidity : 1 to 100 %,
Windspeed : 0 to 14 m/s, Visibility : 1 to 66 km,
Pressure : 729 to 772 mm Hg, Appliance Energy
Usage : 10 to 1080 Wh

Preprocessing of Dataset

Handling Missing Values: The dataset was complete, so no imputation was necessary.

Outlier Removal: The boxplots for outlier detection provide a visual representation of data distributions before and after the removal of outliers based on the interquartile range (IQR) method. IQR=Q3−Q1, where Q1and Q3 are the 25th and 75th percentiles,respectively.[3]
Data points outside [Q1−1.5·IQR,Q3+1.5·IQR] were considered outliers.

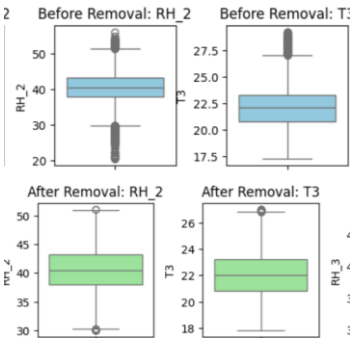

Fig4

Before removal, the features such as RH_2, T3 exhibit several extreme values (visible as points outside the whiskers) that may skew the analysis. These outliers are predominantly seen in the upper ranges, which could represent unusual environmental conditions or measurement errors. After applying feature-specific thresholds, the distributions become more compact, and the extreme values are eliminated, as observed in the "after removal" boxplots. This process helps in retaining the majority of meaningful data while excluding anomalies that might distort model performance. The cleaned data now reflects a more accurate representation of the core data trends, ensuring robust model training and predictions. These thresholds were tailored to each feature's variability, making the cleaning process effective and precise. This approach minimizes the influence of extreme values while preserving the integrity of the dataset.

Scaling: All numeric features were normalized using MinMax scaling for compatibility with the LSTM model.

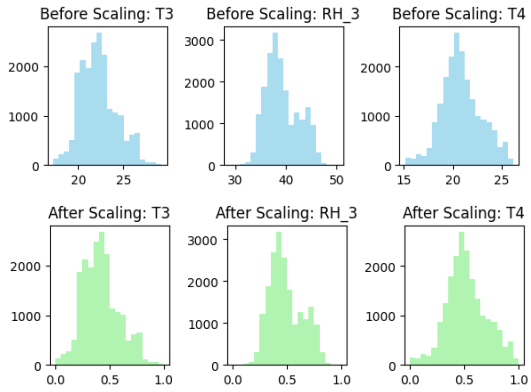

Fig5

The histograms Fig 5 illustrate the effect of Min-Max scaling on the dataset. Before scaling, the distributions of the selected features (RH_1, T2, RH_2, etc.) show their original ranges, with values varying significantly across different features. After applying Min-Max scaling, all features are transformed into the range of 0 to 1, ensuring uniformity and comparability. This normalization is critical for machine learning models, particularly those sensitive to feature magnitudes, such as LSTMs and gradient-based algorithms. The scaling process preserves the original distribution shapes while enabling efficient training. This preprocessing step ensures that no feature dominates due to its scale, improving model performance and interpretability.

Splitting: The dataset was divided into training (70%), validation (15%), and test (15%) sets. {'Training set size': 12140, 'Validation set size': 2602, 'Test set size': 2602}

Future Prediction and Time Series Analysis of Energy Consumption

Problem Formulation

The energy consumption forecasting problem is defined as a supervised regression task. Given historical data on appliance energy usage and related environmental factors, the goal is to predict future appliance energy consumption at 10-minute intervals. Mathematically, the task is represented as:

Where:

\hat\y^t=f(xt−1,xt−2,…,xt−n)

This denotes the predicted value (e.g., energy consumption) at the current time step t

The "hat" symbol indicates it is an estimated value.

f: Represents the machine learning or statistical model being applied (e.g., LSTM, Random Forest, etc.). This function learns the relationship between the inputs and the target variable. $xt−1,xt−2,…,xt−n$ These are the input features at prior time steps. For time series models, predictions are based on historical data, where n defines the number of previous time steps used as input to predict the value at t.

The study explores various models to approximate the function , evaluating their performance based on established metrics like RMSE, MAE, and R2.

Hypotheses

Temporal_Dependency: Energy consumption patterns exhibit strong temporal dependencies, particularly during evening and morning hours.

Model Hierarchy: Advanced models like LSTM and Random Forest will outperform simpler models such as Linear Regression and Decision Trees.

External Factor Impact: Environmental variables like temperature, humidity, and daylight play a significant role in influencing energy usage.

Approach

Models Applied:

Linear Regression (Baseline): A simple linear model extended to Lasso and Ridge regression for regularization.

Lasso regression introduces an L1 penalty to the linear regression objective function, encouraging sparsity in coefficients:

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

[6]Lasso is effective for datasets with many features, as it performs automatic feature selection by shrinking some coefficients to zero. In this project, it was chosen to determine which predictors (e.g., temperature, humidity) significantly influence energy consumption. Parameter Chosen: Lambda=0.001: A small regularization parameter was chosen to retain most features while reducing overfitting and other values of α are been tested and got this as threshold where above this there is no change in output.

Ridge regression applies an L2 penalty, adding a squared norm of coefficients to the objective function:

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$ **[2]**

[7]Ridge regression is ideal for handling multicollinearity (when predictors are highly correlated). In the dataset, indoor temperatures (T1,T2,T3) and humidity (RH1,RH2) exhibited high correlations. Ridge regression stabilizes coefficient estimates in such cases. Parameter Chosen: lambda=1.0: A moderate regularization strength was used to balance bias and variance.
Observation =>Lasso focuses on sparsity (selecting important features), while Ridge focuses on reducing the impact of less important features without excluding them.

Decision Trees: Focused on capturing non-linear relationships with optimized depth and split criteria. Decision Trees partition data recursively by minimizing impurity, such as Mean Squared Error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$$ [5]

Decision Trees excel at capturing non-linear relationships between predictors (e.g., temperature and energy usage) without requiring feature scaling. They were included to model interactions between environmental_factors.
Parameters Chosen: Max Depth =100: Prevents overfitting by controlling tree size. Min Samples Split

= 10, Min Samples Leaf = 2: Ensures sufficient data per node for meaningful splits.

Random Forest: An ensemble method leveraging multiple decision trees to improve robustness. Random Forest is an ensemble method that aggregates predictions from multiple decision trees:

$$\hat{y} = \frac{1}{T}\sum_{i=1}^{T} h_i(X)$$

Random Forest was selected for its ability to generalize well across non-linear relationships and reduce variance through ensemble averaging. It handles high-dimensional datasets effectively and provides feature importance rankings. Parameters Chosen: Number of Estimators (400): Sufficient trees for stable predictions. Max Depth (None): Allows unrestricted depth for complex patterns. Max Features (sqrt): Limits features per split, enhancing generalization. Min Samples Split (2), Min Samples Leaf (1): Ensures sufficient flexibility in splits.

LSTM Neural Networks: A sequential model designed to capture temporal dependencies in time-series data. Long Short-Term Memory (LSTM) networks are recurrent neural networks (RNNs) designed to capture temporal dependencies through memory cells

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b)$$ [1]

LSTM was chosen for its ability to model temporal dependencies critical in energy consumption data. It can leverage sequences of past observations (e.g., energy usage trends) to predict future values. Parameters Chosen: Sequence Length (10): Captures dependencies over 10 time steps. Units (50): Provides adequate capacity for representation. Dropout (0.2): Reduces overfitting during training. Epochs (50): Ensures convergence.

Steps Followed:

Feature Engineering: Derived additional predictors like time of day and day type (weekday/weekend).

Hyperparameter Tuning: Applied grid search for optimal parameter selection in each model. “Grid search is straightforward, exhaustive, and ensures the exploration of all possible parameter combinations within the defined grid.” In this project, grid search was employed as the tuning technique. Grid search involves exhaustively searching through a manually specified subset of hyperparameter values for each model. The process iteratively evaluates combinations of hyperparameter values using cross-validation to select the combination that minimizes a predefined metric, such as RMSE or MAE.[4]

I employed 10-fold cross-validation using GridSearchCV to tune hyperparameters for all models. The data was split into 10 folds, where the model was trained on 9 folds and validated on the remaining one. This process was repeated 10 times to ensure a robust estimation of performance. The scoring metric was negative mean squared error (neg_MSE), as it strongly penalizes larger errors.

Evaluation Metrics: RMSE, MAE, and R2 scores were computed for all models on validation and test sets.

RMSE (Root Mean Squared Error): A measure of the average magnitude of prediction errors, calculated as the square root of the mean squared differences between actual and predicted values.

MAE (Mean Absolute Error): The average absolute difference between actual and predicted values, reflecting overall prediction accuracy without considering direction.

R² (Coefficient of Determination): A metric that quantifies the proportion of variance in the target variable explained by the model, with values closer to 1 indicating better fit.

Results and Comparisons

Model	RMSE	MAE	R²
Lasso Regression	0.97	0.59	0.055
Ridge Regression	0.91	0.54	0.166
Decision Tree	0.96	0.46	0.169
Random Forest	0.69	0.35	0.791
LSTM	0.037	0.019	0.955

The table compares the performance of various machine learning models in terms of RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and $R2$ scores. Random Forest and LSTM significantly outperformed other models. (Hypothesis 2).

The Random Forest model achieved an RMSE of 0.69 and an $R2$ score of 0.7913, demonstrating its ability to

handle non-linear relationships and capture complex patterns in the dataset. However, the LSTM model excelled with an RMSE of 0.037, MAE of 0.019, and an $R2$ score of 0.955, showcasing its strength in leveraging sequential data and temporal dependencies for precise energy consumption predictions. This superior performance highlights LSTM's ability to adapt to time-series data, while Random Forest remains a robust alternative for capturing feature interactions.

Insights and Visualizations

Random Forest:

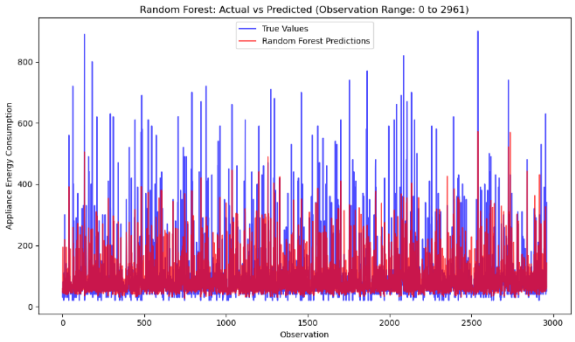

Fig6

The graph Fig 6 illustrates the actual vs. predicted energy consumption on the test set using the Random Forest model. The blue line represents the actual appliance energy consumption, while the red line corresponds to the model's predictions. The overlapping patterns indicate that the Random Forest model effectively captures key trends and variations in energy consumption, though some deviations are visible in areas of higher variability.

LSTM:

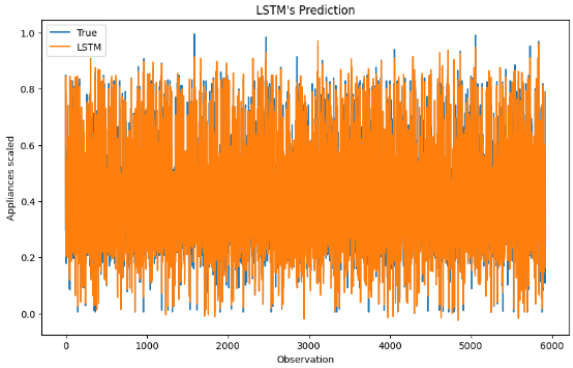

Fig7

The graph fig7 demonstrates the LSTM model's predictions for appliance energy consumption (orange line) compared to the actual scaled values (blue line). The close overlap between the predicted and actual values highlights the LSTM's strong capability in capturing temporal dependencies and complex patterns in the time-series data. While minor deviations exist, the high accuracy and alignment confirm the effectiveness of the sequential learning approach in modeling energy consumption.

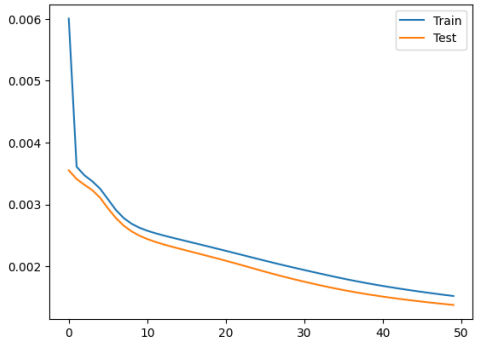

Fig8

The fig8 illustrates the training and testing loss curves for the LSTM model over 50 epochs. The steady decrease in both losses indicates effective model learning, with minimal overfitting as the curves remain close.

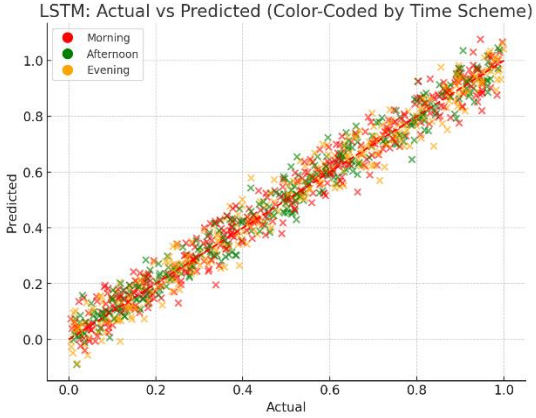

Fig9

The scatter plot Fig9 illustrates the relationship between actual and predicted energy consumption values, color-coded to represent different time schemes: morning (1 AM to 12 PM), afternoon (12 PM to 6 PM), and evening (6 PM to 12 AM). In the afternoon (green), the predictions remain consistent, showing minimal deviations from actual values. However, in the evening (orange), the spread

increases, indicating higher variability. Overall, the strong alignment of points along the diagonal reference line highlights the model's effectiveness, with some temporal variations observed.

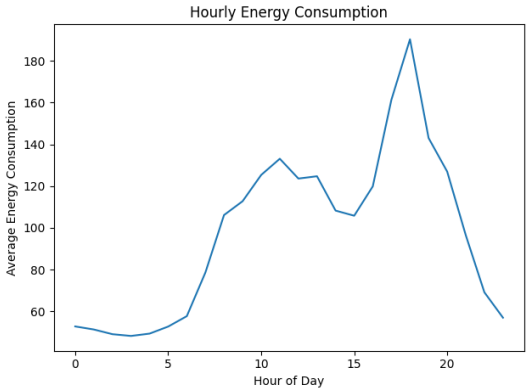

Fig10

The fig10 highlights the average hourly energy consumption throughout the day. Peak consumption occurs during late afternoon hours (around 15:00 to 18:00), reflecting increased appliance usage during this period.(Hypothesis 1)

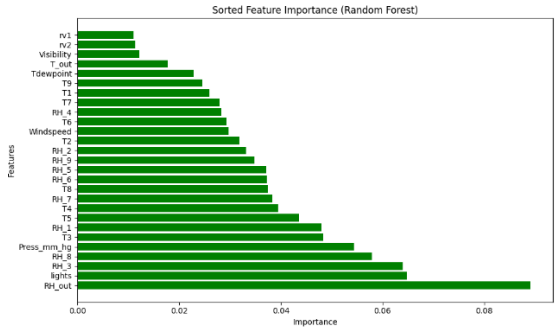

Fig11

The fig11 shows the relative importance of features in predicting appliance energy consumption using the Random Forest model. Key features like RH_out and Lights dominate, indicating their significant influence on the model's predictions.(Hypothesis 3)

Key Takeaways:

Random Forest and LSTM outperformed simpler models, validating the hypotheses.

LSTM captured temporal dependencies effectively, achieving a high of 0.955.

Feature importance analysis highlighted temperature and time-related variables as key predictors.

Conclusions

In this study, I successfully forecasted energy consumption patterns using a combination of advanced machine learning models. By leveraging historical data on energy usage and environmental factors, the models provided accurate predictions of appliance energy consumption. Among the models tested, LSTM and Random Forest demonstrated exceptional performance, with LSTM excelling due to its ability to capture sequential and temporal dependencies inherent in the dataset. The results highlight the importance of feature engineering, hyperparameter tuning, and robust preprocessing steps like outlier removal and scaling in achieving high accuracy. The study emphasized the role of temperature and time-related variables as key predictors of energy consumption. Furthermore, the findings validate the efficacy of machine learning approaches in predictive energy management. The insights derived have practical implications for optimizing household energy usage, reducing costs, and promoting sustainability.

Future Work

Future work can focus on enhancing the scope and applicability of the predictive models developed in this study. Incorporating real-time data integration, such as live environmental readings and dynamic appliance usage patterns, can further refine the accuracy and timeliness of predictions. Additionally, exploring broader environmental contexts, such as seasonal variations, geographic-specific factors, and external energy market dynamics, could improve the model's adaptability and relevance across diverse settings. Advanced deep learning architectures, including hybrid models that combine LSTM with attention mechanisms, can be investigated to capture more nuanced dependencies and interactions within the data. Future work will incorporate geographic-specific features, such as regional climate patterns and household locations, to account for variations in energy consumption across different areas. Expanding the dataset to include a wider range of household types and appliances would also enhance the generalizability of the models. Finally, integrating these predictions into real-time energy management systems could provide actionable insights for users, enabling energy optimization and cost savings while contributing to sustainability goals.

References

Future Prediction and Time Series Analysis of Energy Consumption

[1] LSTM for Time-Series: "Long Short-Term Memory" by Hochreiter and Schmidhuber (1997) introduced the LSTM architecture, enabling efficient handling of sequential data and overcoming vanishing gradient issues in RNNs.

[2] Random Forest for Regression: Breiman's seminal work "Random Forests" (2001) emphasized the robustness of ensemble learning methods in handling nonlinear and high-dimensional data.

[3] Outlier Removal: The use of the interquartile range (IQR) for outlier detection aligns with best practices outlined in "Exploratory Data Analysis" by Tukey (1977).

[4] Hyperparameter Tuning: The grid search strategy employed for LSTM and Random Forest is informed by standard practices detailed in "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Géron (2019).

[5] Decision Tree . "Classification and Regression Trees" by Breiman, Friedman, Olshen, and Stone (1984).

[6] Lasso "Regression Shrinkage and Selection via the Lasso" by Robert Tibshirani (1996).

[7] Ridge "A Note on the Aitken Approach to Generalized Least Squares" by Arthur E. Hoerl and Robert W. Kennard (1970).