SMART HOME ENERGY CONSUMPTION PREDICTION USING MACHINE LEARNING AND TIME-SERIES MODELS

Course Name: Computer science Project

Course Of Study : M.Sc. Computer Science

Course Id : CSEMCSPCSP01

Student Name : LAKSHMI LAVANYA SYAMAKURI

Matriculation Number :4252231

Tutor Name :Harsha Raju

Table of Contents

List of Figures

1. Introduction

The growing development of smart home technologies has made a tremendous impact on the way the residential energy consumption is monitored and handled (Shareef et al., 2018). Combined with the use of IoT devices, smart meters will constantly capture high-resolution energy consumption information, providing the prospect of smart energy optimisation. Proper forecasting of the household energy usage is important in curbing the cost of electricity, enhancing grid stability as well as supporting sustainability efforts as it allows individuals and energy suppliers to make well-informed decisions (Manjula et al., 2025). Human behaviour, environmental conditions and patterns of use of appliances are dynamic and stochastic making energy consumption particularly challenging to predict (Chen et al., 2020). The old-fashioned energy management methods tend to use fixed rules or past averages, which cannot keep up with the short-term variability and very intricate time relationships that smart meter data entails. Thus, machine learning and time-series analysis computational methods have become more topical and applicable to this issue. In a practical sense, predicting the energy with certain reliability can allow such uses as load balancing, reducing the peak demand, detecting anomalies, and providing individualised energy-saving suggestions. To utility providers, precise predictions can aid in managing demand-side and infrastructure, and to households, they assist in saving of money and greater energy awareness. Thus, the issue of energy consumption prediction is valuable and influential in the applied computer science domain.

The purpose of this project is to design, implement and test a smart home energy consumption prediction system with the use of various computational models. The main research task is to examine the performance of various categories of models, that is, machine learning models, statistical time-series models, and deep learning architectures when applied to the same set of data on the same evaluation grounds. Through the comparison of these methods, the project aims at determining their relativity, weaknesses and how they can be applied in smart homes. The current work scope will be limited to one of the short-term energy consumption forecasting based on the historical smart meter data. Instead of offering one optimal model, the project focuses on comparative distribution and methodological transparency, thus giving a hint on the model selection and system design in real-life energy prediction usage.

2. Related Work

The forecasting of energy consumption has been largely researched as applied to power systems, smart grids, and smart homes. The initial methods were largely based on statistical prediction methods including autoregressive models, exponential smoothing models (Attanayake et al., 2019). The use of Autoregressive Integrated Moving average (ARIMA) models has been popular since they have a solid theoretical basis and can be able to model linear time dependencies (Arumugam et al., 2023). Yet, ARIMA models presuppose the stationarity and have a drawback in terms of their inability to describe non-linear trends that are prevalent in residential energy data. Machine learning methods have become popular in energy prediction tasks with the development of computational resources (He et al., 2020). Linear regression models offer an easy and understandable baseline whereas tree-based models, e.g., Decision Trees and Random Forests, are more flexible in the modelling of non-linear relationships. It has been identified that the methods should work well in contrast to pure statistical models when other contextual features, including weather or past averages, are present. However, the classic machine learning algorithms need a lot of feature engineering and can be prone to long-term temporal correlations.

Lately, deep learning models have been recognised as an effective substitute to time-series prediction. RNNs (recurrent neural networks) and specifically LSTM networks have been shown to be very effective in sequential data modelling through the preservation of internal memory states. Several studies have indicated that LSTM-based models are found to be better than the classical methods in short-term load forecasting tasks, particularly with highly volatile and complex consumption pattern. Although the accuracy of deep learning models is high, they are computationally demanding and, in some cases, they are not interpretable, thereby restricting their use in specific applications. The proposed project is based on the existing research, as it provides a systematic comparison of statistical, machine learning, and deep learning methods on a common experimental setup. The work will offer a balanced performance evaluation, complexity and practical applicability of several models in smart home energy forecasting contexts by comparing models based on similar metrics on the same dataset.

3. Technical Background

3.1 Smart Meter Data Characteristics

The data of smart meters is a time-series by nature, which is a sequence of measurements of energy consumed at a given time interval (Rashid et al., 2018). The data set used in this project consists of about 2,400 timestamped observations where a record is a compiled household energy consumption data with time. The data is highly variant due to the daily habits, usage patterns of appliances, and other factors in time of day, like in the real-life smart meter data, there are missing points and anomalies. These features require a special preprocessing, time-conscious interpolation, and strict chronological order to guarantee the appropriateness of the information to the forecasting models.

3.2 Regression Models

One of the most basic predictive models in machine learning is the Linear Regression, which was used in this project as a baseline model to determine a reference level of performance. It also assumes a linear correlation between the input features and the target variable, thus it is simple to compute and interpret. Nevertheless, its powerful assumptions restrict its capability to model difficult, non-linear energy consumption patterns. Decision Trees are regressions which expand the regression ability by the recursive partitioning of the feature space in regions sharing the same target. The Decision Tree Regression applied in this project also helped in capturing possible non-linear relationship between engineered numerical features and energy consumption when it was not necessary to carry out intricate transformations of features.

3.3 Time-Series Models

ARIMA models are classical statistical tools that are specifically used to forecast univariate time series and were added to this project because it is used as a benchmark statistic. The model was implemented only on the energy consumption series since the ARIMA models only require the use of a single dependent variable. The model internally applied differencing so that it could meet the stationarity assumption. Although ARIMA is an effective model in capturing linear temporal dependencies, its restricted ability to capture non-linear behaviour, which is often complex, makes it a useful comparison to machine learning and deep learning methods (Hossain et al., 2025).

3.4 Evaluation Metrics

Root Mean Squared Error (RMSE) and Means Absolute error (MAE) are generally the metrics used to assess the regression models. MAE is an easy-to-use indicator of average prediction error, whereas RMSE is more severe towards bigger errors. Combined, these measures are a complete evaluation of model accuracy and robustness. These measures are especially appropriate to energy forecasting exercises, because they represent absolute consumption errors in units that are easily interpretable and are also sensitive to large errors in prediction.

4. Method

The proposed project approach is a systematic pipeline that deals with information preprocessing, model training, and evaluation. This dataset is initially cleaned and ready to be compatible in terms of time across as well as across models numerically. The chronological train-test split of 80/20 is used, which guarantees that all models have been trained based on the historical data only and tested on previously unknown observations, which prohibits the leakage of time-related information. Different models are chosen to represent various paradigms of forecasting to facilitate a systematic comparison of traditional statistical techniques, classical machine learning models, and deep learning approaches. Linear Regression and Decision Tree are selected as representative machine learning baselines because they are easy to understand and are frequently used. ARIMA model is also mentioned as a statistical benchmark that is specifically designed to perform the time-series prediction. The reason why the LSTM model is selected is because it has the capability of storing long-term temporal relationships with gated memory processes that are especially important in sequential energy consumption data.

All the models are trained on the same training subset to make them comparable. The test subset gets predicted and assessed with the help of MAE and RMSE. Such a comparative methodology enables the systematic analysis of the difference in the performance between different types of models with the variation of data-related variables. The decision to use Python as the implementation language is influenced by the wide range of data science libraries it has, such as scikit-learn, a machine learning library, and statsmodels, a time-series analysis library, and TensorFlow, a deep learning library. This combination allows fast development and at the same time maintains the methodological rigor and reproducibility.

5. Implementation

In this section, the actual implementation of the proposed energy consumption prediction system is presented, emphasizing the selected technologies, software architecture, preliminary data processing, and model training. Its implementation will be based on the principles of established software engineering and the focus on modularity, reproducibility, and clarity.

5.1 Technology Stack

The complete application was developed in Python, which is a good language to use in data-driven apps because it has an extensive scientific computing environment. Python features high-level abstractions and flexibility to enable quick prototyping and experimentation and is, therefore, suitable in machine learning and time-series analysis assignments. Several libraries were used in supporting various pipeline phases. Data manipulation, numerical operations and indexing of a time-series were done using Pandas and NumPy, whereas scikit-learn was adopted to implement the Linear Regression and Decision Tree models, and evaluation metrics because it has a rich time-series statistical support. To apply deep learning, TensorFlow/Keras was employed in the implementation of the LSTM model, which was trained in Google Colab to deal with the local hardware and dependency limitations. Matplotlib was used in visualisation. GitHub was used to control version controls and share code, with the entire project repository, including source code, scripts to process data and documentation, publicly available. This makes it transparent, reproducible and in line with the assignment requirements.

5.2 Software Architecture

The system has a modular software structure that isolates concerns within a number of files and directories. This design enhances maintainability, readability and extensibility. The basic elements comprise preprocessing, feature engineering, model training, and evaluation, each of which is realised as an independent module. The preprocessing module involves ingestion, parsing of timestamps and cleaning of raw data. The scripts of each type of algorithm contain model-specific training logic. The training process is organised by a centralised training script, and a special evaluation module calculates performance metrics and creates visualisations. The structure is consistent with commonplace procedures in professional machine learning workflow.

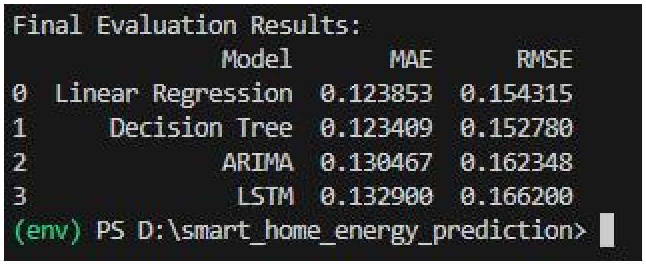
5.3 Data Preprocessing

Smart meter data is highly irregular and noisy data for preprocessing which is an important step. The timestamp column was directly interpreted and made the index of the DataFrame to allow time sensitive operations like resampling and interpolation. Interpolated values were treated as missing values and therefore, time continuity was maintained and artificial discontinuities were not introduced. Selecting of features was done by choosing only numerical features that will be used in prediction. Categorical labels that were not numbers like the indicators of anomalies were eliminated before the training of the models to make them compatible with regression-based algorithms. All the features were confirmed to have similar consistent data type and indexing all the data set.

5.4 Model Training

It has four predictive models that were trained. Linear Regression was employed as a control since it is simple and can be interpreted. Decision Tree Regression was also added to capture the non-linear relationships without much feature engineering. The ARIMA model was fitted using the univariate series of energy consumption only because that is what the statistical formulation of the model requires. Correct approach to differencing and indexing made valid predictions in line with the period of testing. The LSTM model was then trained in Google Colab where it was easy to get access to both GPU acceleration and compatibility with TensorFlow. The evaluation metrics of the trained model were copied to the local pipeline so that they could be compared.

6. Testing and Evaluation



```
Final Evaluation Results:
               Model       MAE      RMSE
0  Linear Regression  0.123853  0.154315
1      Decision Tree  0.123409  0.152780
2              ARIMA  0.130467  0.162348
3               LSTM  0.132900  0.166200
(env) PS D:\smart_home_energy_prediction>
```

Figure 1: Quantitative comparison of forecasting models using MAE and RMSE on the test dataset.

The models that were implemented were tested and assessed to ensure that they were correct and predictive. The main testing approach was a temporal train-test split with 80% of the data being used in training and the rest in testing. This method corresponds to real-life forecasting problems, where the values in the future are to be forecasted based on past observations. The validation was initially correct with the help of intermediate validation tests that consist of checking data integrity, dimensional consistency and successful model convergence. Model predictions were visually checked by having the eyes on them to ascertain that they were visually aligned and plausible in terms of numbers. Mistakes on issues of data leakage or misalignment of indexes were strictly avoided. Mean Absolute Error (MAE) and root mean squared error (RMSE) were used to perform the evaluation of performance. The reason is that MAE is a more appropriate metric of average prediction deviation, whereas RMSE is more severe on the large deviations, so it is more sensitive to extreme deviations. These measures are properly established in the literature of energy forecasting and can be compared fairly across the models.

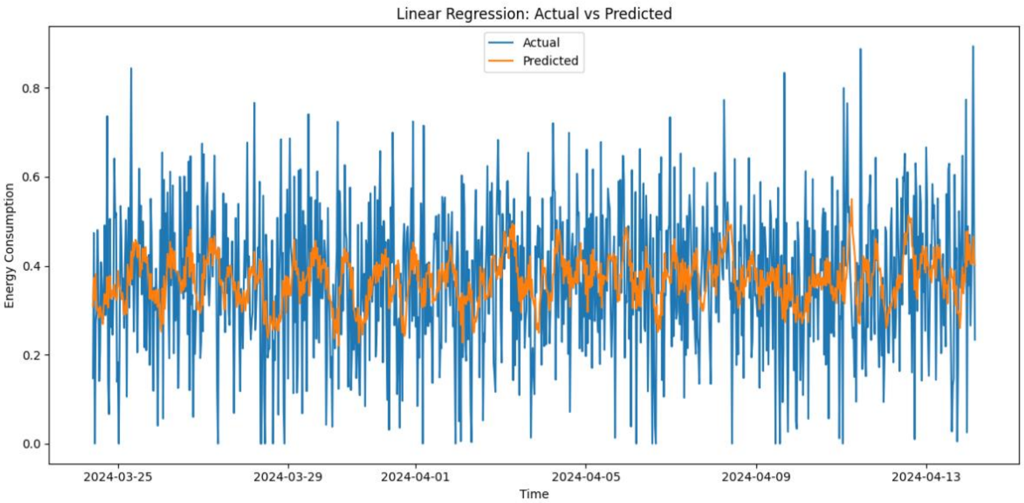


Figure 2: Actual versus predicted energy consumption using a machine learning regression model on the test dataset.

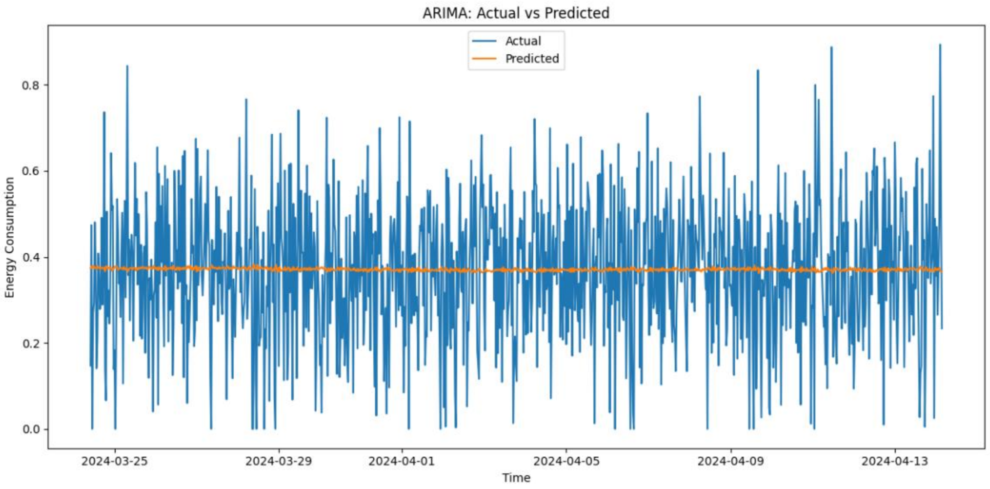


Figure 3: Actual versus predicted energy consumption using the ARIMA model. The smooth prediction curve reflects the model's reliance on averaged temporal trends.

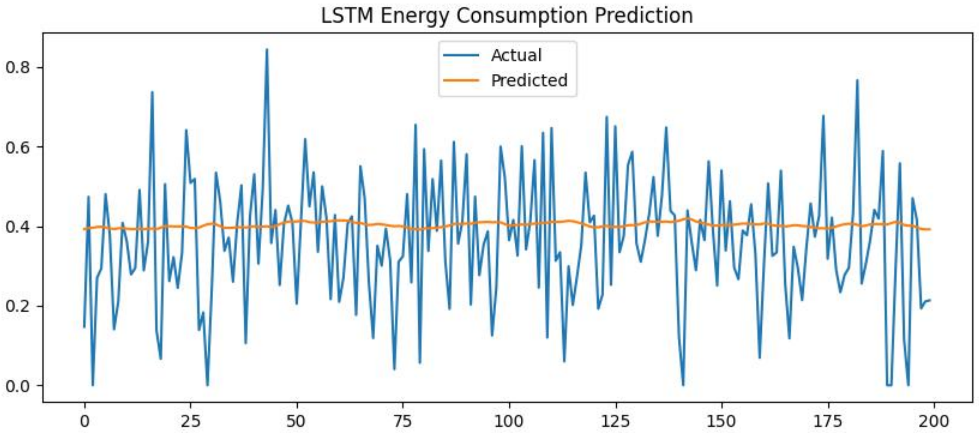


Figure 4: Actual versus predicted energy consumption using an LSTM model, demonstrating improved tracking of short-term fluctuations.

The results of the evaluation were presented in a comparative table, with the differences in the accuracy of predictions. The actual and predicted energy consumption were visually produced in the form of each model. These plots give qualitative results on model behaviour with time. It is important to note that ARIMA forecasts seem smoother than the actual ones, which are a result of the fact that the model makes use of linear time dependencies and that it averages high-frequency variations. The LSTM model had the highest overall performance, and this indicates that it can encode non-linear and complex temporal patterns. This high performance can be explained by the fact that this has a higher internal memory structure that allows it to learn long term dependencies that cannot be captured using the simpler models. To achieve reproducibility, data splits were addressed, preprocessing steps were recorded, and a standard evaluation pipeline was used. Code and configuration information is made publicly available on GitHub, and the results can be independently verified.

7. Discussion

The results of the experiment indicate that there is evident performance disparity between the considered models, presenting significant trade-offs between precision, sophistication, and explainability. Linear Regression, though easy and computationally inexpensive, failed to explain non-linear trends of household energy consumption data. Decision Trees built on this baseline but had the problem of sensitivity to noise. The ARIMA model gave predictable and understandable forecasts although it generated excessively smooth forecast that was not responsive to the volatility in the short-term. Such behaviour corresponds to the assumptions of the model and underlines its drawbacks in the highly dynamic settings (smart home). LSTM model has proven to be better in all the other methods with least error metrics. This enables it to capture long-term dependencies and non-linear response to time, and thereby it is especially appropriate to sub-tasks of energy forecasting. This performance is however achieved at the expense of increased computational complexity, interpretability, and an implementation overhead. The dataset covers a narrow time frame and fails to include the external contextual details like weather or occupancy statistics. Moreover, the optimisation of hyperparameters was limited in order to keep the project within its limits. Nevertheless, the findings indicate the potential usefulness of state-of-the-art machine learning algorithms in smart home energy prediction and prove that improvements are possible in the future.

8. Conclusion

This project explored the issue of predicting the short-term energy consumption of homes in smart homes through a blend of the traditional statistical techniques and the current machine learning approaches. The work was inspired by both the growing access to smart meter data and the imminent necessity to have a more efficient approach towards managing energy, and the specific research question explored whether the lightweight predictive models would be suitable to be implemented practically. Four forecasting techniques were used and contrasted, Linear Regression, Decision Tree Regression, ARIMA, and an LSTM-based neural network. The findings indicate that more complex models like Linear Regression and Decision Trees, although computationally efficient and easy to understand, can be used to describe only basic temporal behaviour. The ARIMA model was stable in its predictions and gave an over-smoothed prediction, which demonstrated the weaknesses of using univariate statistical models in prediction using highly varying data of smart meters.

The LSTM model had the highest predictive performance which validates the fact that deep learning models are appropriate in modelling non-linear temporal dependencies. Nonetheless, the enhanced precision comes at the reduced interpretability and complexity of computations, making it important to consider the trade-offs when choosing the models to be used in the real-world smart home systems. All in all, the project proves that methods of machine learning can be used to a great effect to enhance the short-term forecasting of energy consumption. The adopted framework offers a modular and reproducible platform to build upon by creating more energy-efficient and intelligent smart home surroundings.

9. Future Work

The proposed framework can be expanded in several ways to facilitate the predictive abilities and usability of the proposed framework. To start with, it would enhance performance by including seasonal elements by using models like SARIMA to explicitly capture daily or weekly consumption cycles, which are commonly seen in household energy data. Second, it would be possible to expand the framework to facilitate multivariate prediction to enable the use of exogenous variables like outdoor temperature, humidity, occupancy patterns, or appliance use. These context characteristics are known to have an impact on energy use and may contribute greatly to better prediction. By methodically optimising the hyperparameters, such as grid search or Bayesian optimisation algorithms, systematic hyperparameter optimisation, which will especially be effective with Decision Tree and LSTM models, can be further improved. As well, discussing model explainability methods, including feature importance analysis or SHAP values, would raise the level of transparency and user confidence, which is a key factor to consider in smart home applications. One possible direction of future work is to consider deployment, such as real-time prediction pipelines, model compression to low-power devices, and interoperability with home energy management systems. These additions would take the project one step further into a deployable smart home system.

References

Shareef, H., Ahmed, M. S., Mohamed, A., & Al Hassan, E. (2018). Review on home energy management system considering demand responses, smart technologies, and intelligent controllers. *Ieee Access*, *6*, 24498-24509.

Manjula, A., Niraimathi, R., Rajarajeswari, M., & Devi, S. C. (2025). Grid integration of renewable energy sources: challenges and solutions. In *Green Machine Learning and Big Data for Smart Grids* (pp. 263-286). Elsevier.

Chen, S., Wu, J., Pan, Y., Ge, J., & Huang, Z. (2020). Simulation and case study on residential stochastic energy use behaviors based on human dynamics. *Energy and Buildings*, *223*, 110182.

Attanayake, A. M. C. H., Perera, S. S. N., & Liyanage, U. P. (2019). Combining forecasts of ARIMA and exponential smoothing models. *Advances and Applications in Statistics*, *59*(2), 199-208.

Arumugam, V., & Natarajan, V. (2023). Time Series Modeling and Forecasting Using Autoregressive Integrated Moving Average and Seasonal Autoregressive Integrated Moving Average Models. *Instrumentation, Mesures, Métrologies*, *22*(4).

He, Y., Wu, P., Li, Y., Wang, Y., Tao, F., & Wang, Y. (2020). A generic energy prediction model of machine tools using deep learning algorithms. *Applied Energy*, *275*, 115402.

Rashid, M. H. (2018, August). AMI smart meter big data analytics for time series of electricity consumption. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)* (pp. 1771-1776). IEEE.

Hossain, M. L., Shams, S. N., & Ullah, S. M. (2025). Time-series and deep learning approaches for renewable energy forecasting in Dhaka: a comparative study of ARIMA, SARIMA, and LSTM models. *Discover Sustainability*, *6*(1), 775.

GithubLink : https://github.com/Lakshmi4252231/smart-home-energy-consumption-prediction-using-machine-learning-and-time-series-models.git