THE USA JOURNALS
THE AMERICAN JOURNAL OF ENGINEERING AND TECHNOLOGY (ISSN – 2689-0984)
VOLUME 06 ISSUE12

PUBLISHED DATE: - 19-12-2024
DOI: - https://doi.org/10.37547/tajet/Volume06Issue12-11 **PAGE NO.: - 99-118**

RESEARCH ARTICLE **Open Access**

PREDICTIVE MODELING OF HOUSEHOLD ENERGY CONSUMPTION IN THE USA: THE ROLE OF MACHINE LEARNING AND SOCIOECONOMIC FACTORS

Muhammad Shoyaibur Rahman Chowdhury
Information Technology, Gannon University, Erie, PA

Mohammad Saiful Islam
MS, Management - Information Technology Management, St. Francis College

Md Abdullah Al Montaser
Ms-Business Analytics, University of North Texas

Mohammad Abul Basher Rasel
MSc Hospitality & Tourism Data Analytics, University of North Texas

Ayan Barua
MBA in Business Analytics, Trine University: Angola, US

Anchala Chouksey
Masters in financial mathematics, University of North Texas, Denton, Texas

Bivash Ranjan Chowdhury
MBA in Management Information System, International American University, Los Angeles, California, USA

Corresponding Author: Muhammad Shoyaibur Rahman Chowdhury

THE USA JOURNALS
THE AMERICAN JOURNAL OF ENGINEERING AND TECHNOLOGY (ISSN – 2689-0984)
VOLUME 06 ISSUE12

Abstract
Understanding the pattern of energy use at the household level becomes ever more urgent in light of growing concerns about climate change and resource sustainability in the USA. Energy use depends upon various factors, such as climate, household characteristics, and behavior. Of these, income, education, and size of the family are very vital socio-economic factors that depict energy consumption levels and their pattern. The utmost objective of this research project was to develop predictive models using machine learning techniques to analyze household energy consumption trends in the USA, integrating socioeconomic factors such as income, family size, and education. The dataset retrieved from Kaggle integrates detailed weather patterns with energy consumption data, putting into perspective the interaction between climatic variables and household energy use. It included key features such as temperature, humidity, wind speed, and precipitation, along with time-series data on energy consumption metrics like electricity and natural gas usage at the household level. It provided information on several geographic zones across extended periods, so seasonality and regional variations may be studied. It was complemented with metadata that included timestamps, energy pricing, and household attributes and should therefore be a rich resource for predictive modeling and extracting relationships between weather conditions and energy demand. For this research project, three models were selected: Logistic Regression, Random Forest, and Support Vector Machines, each possessing particular strengths for the nature of the problem. This study employed key performance metrics such as precision, recall, F1-Score, and accuracy. The Random Forest model had the highest value for accuracy, similarly, the highest AUC was for the Random Forest with the best AUC. As such, it was concluded that the Random Forest model provided the best trade-off between true positive rate and false positive rate and can be relied upon for this classification task. The machine learning models generate valuable predictions about household energy use. Particularly, Random Forest models, which are trained on socioeconomic and weather data to predict the likelihood of a given household having high energy usage. The predictions by such models can be used to help energy providers determine when to invoke tiered pricing or encourage energy-saving behavior.

Keywords Household Energy Consumption; Socioeconomic Factors; Machine Learning; Random Forest; Predictive Modelling; Energy Efficiency, USA.

INTRODUCTION

Background and Motivation

According to Hasan (2024), household energy consumption encompasses a substantial portion of total energy usage in the United States. According to the United States Energy Information Administration, in recent years, residential energy use has made up about 22% of total energy consumption. This huge share underlines how important households are in shaping the nation's energy profile and environmental footprint. Understanding the pattern of energy use at the household level becomes ever more urgent in light of growing concerns about climate change and resource sustainability. Nasiruddin et al. (2023), reported that Energy use depends upon various factors, such as climate, household characteristics, and behavior. Of these, income, education, and size of the family are very vital socio-economic factors that depict energy consumption levels and their pattern. For example, larger households may have larger homes and a proliferation of energy-using appliances. On the other hand, low-income households may face several barriers, including limited access to energy-efficient technologies that contribute to higher energy inefficiencies. This calls for the need to recognize and address such disparities to achieve equity and efficiency in energy access (Amiri et al., 2023; Chen et al, 2023).

Problem Statement

Karmakar et al. (2024), posited that predicting household energy consumption presents noteworthy challenges because of the diverse array of factors that influence energy use, ranging from interacting physical attributes of the home, like size and age, to behavioral elements of energy conservation practices to broader socioeconomic

determinants. These relations are nonlinear, hence demanding advanced levels of techniques to be captured or modeled. It is in this aspect that Machine Learning can provide promising solutions. In contrast to traditional statistical approaches, Machine Learning models can handle big, heterogeneous datasets and find complex patterns in the data. Alam et al. (2023), asserted that while ML techniques have been applied to perform well in different predictive tasks, incorporating socioeconomic factors into such models remains largely unexplored. This is considered a serious gap in the literature because it reduces the understanding of how variables like income and education contribute to disparities in energy consumption, thus making the development of fair and efficient energy policies difficult (Charfeddine et al., 2023; Goriparthi, 2023).

Research Questions

RQ1: How can machine learning models be used to predict household energy consumption in the USA?

This research question aims to examine the possibility of machine learning techniques to predict household energy consumption. Techniques for machine learning can identify latent patterns and trends by analyzing historical data on energy usage, alongside other related factors such as weather conditions, appliance ownership, and occupancy. This can then be used for prediction in forecasting future demand and may allow utilities to optimize resource allocation and grid management while offering personalized advice to consumers on saving energy.

RQ2: What is the impact of socioeconomic factors such as income, family size, and education on energy usage?

This research question seeks to explore the interrelation between socioeconomic factors and energy consumption. All this information on family income level, family size, and attainment of education would help map changes in energy usage behaviors. From such findings, one can develop a practical and targeted energy efficiency program or policies as well as enable utilities to propose strategies for reducing energy use and promoting sustainability.

Significance of the Study

The findings of this study have large implications for policy, the energy provider, and broader society in general. First is that the forecast of the consumption of energy at an individual household level may underpin intervention designs targeted at bringing about waste reduction and increased efficiency. For example, utilities may use forecast peaks to make dynamic pricing per time to motivate the shift of household usage to off-peak. Second, factoring in socioeconomic variables in these predictive models allows the analysis of patterns in energy consumption inequality at a much deeper level. Gaining this insight is crucial to ensuring fairness and equal access to policies that make energy-efficient technologies and programs available. For instance, identifying communities with the highest energy inefficiencies might guide investments in energy retrofitting and education programs. This case scenario represents the application of ML for energy prediction in the general direction of digital transformation and smart grid technologies that will contribute to more resilient and adaptive energy systems. With these advantages, the United States will be able to take big steps toward its goals of efficient and sustainable energy.

Literature Review

Household Energy Use Patterns in the USA

Hasannuzzaman et al. (2023), indicated that energy consumption by households in the United States represents a major component of national

energy use, contributing about 22% to the overall energy demand according to the U.S. Energy Information Administration. Drivers of residential energy consumption include heating, cooling, water heating, and appliances; among these, space heating is the largest contributor. Over the past years, however, there has been an evident shift towards more energy-efficient appliances and utilization of renewable energy sources, in which solar panels have become most highly instrumental, especially in states that also incentivize the installment through rebates (Shawon et al. 2023a).

Notwithstanding these developments, Kapp et al. (2023), reported that consumption remains unequal across regions and demographics. Spates in colder climes, such as the Northeast and Midwest, tend to consume more energy for heating purposes, while warmer states like those in the South create higher cooling energy demands. Generally, urban households have smaller living spaces and more efficient energy infrastructure, making per capita energy consumption lower compared with rural households (Shil et al., 2024). Besides, emerging trends show the adoption of technologies is restructuring the way energy is consumed. While this can be enabled through increased proliferation of smart home devices and energy management systems, the latter allows households to manage their energy use better and reduce energy waste. However, their diffusion is highly variable and influenced by income and educational level; socioeconomic factors have to be taken into consideration while explaining the consumption pattern (Rahman et al., 2024).

Factors Affecting Energy Use in Households

As per Sumon et al. (2024), household energy consumption is a result of a mix of structural, behavioral, and environmental factors. The size and age of a home, insulation quality, and type of heating and cooling systems all fall into the category of structural factors. Behavioral factors relate to energy use habits, awareness of practices that can save energy, and the willingness to invest in upgrades to efficiency. Other critical environmental factors include regional climate, seasonal variations, and availability of renewable energy sources. Besides, other demographic variables such as household size, composition, and income level shape the pattern of energy use (Kesriklioglu et al., 2023). For example, large households are likely to be the highest energy consumers, whereas the per capita consumption may be lower compared with the smaller households because of certain scale economies. Income levels affect not only the affordability of energy efficiency but also the appliances that will be able to take advantage of energy efficiency improvements and retrofits, producing varying energy efficiencies across various economic groups (Buiya et al., 2023a).

Machine Learning for Energy Consumption Prediction

Debnath et al. (2024), contend that machine learning has emerged as one of the strong analytical and predictive tools for household energy consumption due to its capability to model complex and nonlinear data relationships. Whereas traditional statistical approaches have proven helpful, many-faceted interactions between variables like household characteristics, environmental conditions, and socioeconomic factors often remain beyond the grasp of conventional statistical methods. In contrast, ML techniques can handle large high-dimensional datasets and can come up with patterns not that obvious (Kumar2023; Mukelabai et al., 2023). The various ordinary ML techniques used in energy consumption prediction include:

Regression Models: Some of the generally used algorithms are linear regression, support vector regression, and boosting to predict the continuous

level of energy consumption based on weather, appliance usage, and household demographics (Islam et al., 2024).

Classification Models: These models classify households into pre-defined levels of consumption and, therefore, give insights into high or low users of energy. Decision trees and random forests are some of the popular ones used in this area (Islam et al., 2024).

Neural Networks: Temporal patterns and seasonality in time-series energy consumption data are modeled using deep learning methods, including convolutional and recurrent neural networks (Khan et al., 2024).

Clustering Techniques: Some unsupervised machine learning techniques include clustering, for example, k-means clustering techniques, to segment households based on consumption behavior to enable targeted interventions (Khan et al., 2024).

Several studies have showcased the performance of ML on household energy consumption forecasting. For instance, Hasan (2024) conducted a study where he implemented SVM to forecast residential electricity demand and achieved a high degree of accuracy by incorporating weather data along with household characteristics. Similarly, the integration of ANNs in the work of Al Mukaddim et al. (2023) moderated energy consumption trends to be forecasted with time-series data and gave insight into peak usage periods. Other studies have investigated hybrid models that use a combination of ML algorithms and optimization algorithms. For example, Sumon et al. (2024) developed a model that combined genetic algorithms and gradient boosting for improved accuracy in prediction. These studies therefore provide evidence of an increase in the sophistication of applications of ML in the energy sector, underpinning their potential to further enhance energy efficiency and planning. Notwithstanding these developments, a limiting factor in most of the models is the underrepresentation of socioeconomic factors. In most instances, the models have a predisposition toward technical and environmental variables at the expense of more holistic studies that include income, education, and demographic data which would provide a fuller understanding of household energy consumption (Nguyen et al., 2023; Ohalete et al., 2023).

Socioeconomic Factors and Energy Consumption

According to Sumon et al. (2023), socioeconomic factors play a pivotal role in shaping household energy consumption patterns. Income appears to be the main driver and Influences the type and quantity of energy-consuming appliances the family owns, the adoption of technologies that will help save energy, or even afford renewable energy installation costs. For instance, a high-income household is most likely to possess energy-consuming equipment such as home entertainment systems, which result in a higher general consumption (Shawon et al., 2024b). Another factor that makes a difference in energy consumption is family size and composition. Essentially, larger households tend to use more energy because of the higher demand for lighting, heating, and appliances. However, their per capita energy consumption is usually lower because several people are using the same resources (Palani et al.,2023; Omogoroye, 2023; Sarwar et al., 2024).

Furthermore, households with children may have different consumption patterns than households where all members are older, given that children may use more electricity for laundry and electronic entertainment. The level of education equally affects energy behavior (Nasiruddin et al. 2023). The more educated the people, the higher their awareness of energy-saving practices and the likelihood that they invest in energy-efficient

appliances (Singh et al., 2023; Shi et al., 2023). Indeed, several studies have demonstrated a positive relationship between education and willingness to adopt renewable energy solutions like solar panels, which are bound to drastically reduce dependency on non-renewable sources of energy (Karmakar et al. 2024).

DATA COLLECTION AND PREPROCESSING

Data Sources

The dataset retrieved from Kaggle integrates detailed weather patterns with energy consumption data, putting into perspective the interaction between climatic variables and household energy use. It includes key features such as temperature, humidity, wind speed, and precipitation, along with time-series data on energy consumption metrics like electricity and natural gas usage at the household level. It provided information on several geographic zones across extended periods, so seasonality and regional variations may be studied (Pro-AI-Robikul, 2024). It was complemented with metadata that included timestamps, energy pricing, and household attributes and should therefore be a rich resource for predictive modeling and extracting relationships between weather conditions and energy demand. This dataset was particularly useful for the application of machine learning methods to forecast energy requirements and optimize resource allocation accordingly.

Data-Preprocessing

Preprocessing of data was a very significant step in preparing for the analysis and modeling that followed. Handling missing values involved dealing with gaps in the data that could compromise the accuracy of the predictions. These included mean or median imputation techniques for numerical variables, while mode imputation or predictive imputation would be considered for categorical data. In worst-case scenarios about the quality of the data, a threshold beyond which, according to that threshold, rows or columns with excessive missingness were removed. This was then followed by feature engineering, which transformed raw data into meaningful input for machine learning (Pro-AI-Robikul, 2024). For weather features, temperature and wind speed were numerical variables scaled to standardize the range of the values for the model's stability. Categorical variables were encoded, for instance, one-hot encoding for the conditions, whether sunny or rainy. Finally, data transformation converted the Date column into a Date-Time format, which allowed the extraction of time-based features like day, month, or hour features that are crucial for time-series modeling or finding temporal patterns in energy consumption. Such steps ensure that the dataset is clean and consistent for predictive analysis.

Exploratory Data Analysis (EDA)

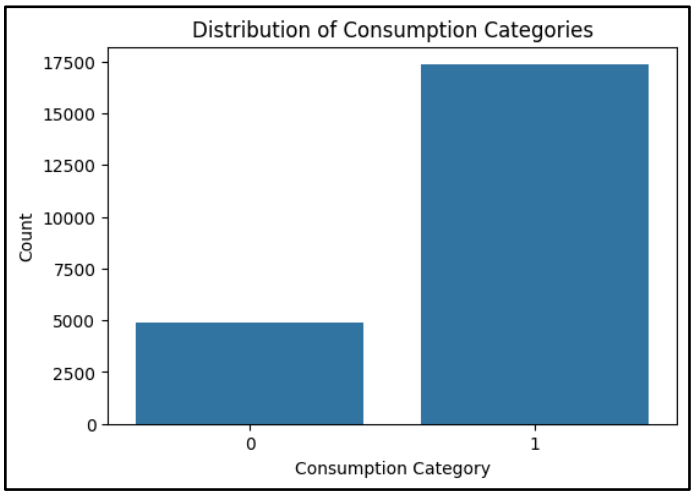

Figure 1: Portrays the Distribution of Consumption Categories

This histogram represents the distribution of two consumption categories, 0 and 1, with the frequency of each category in the dataset. Category 1 has a very high frequency, about 17,500 occurrences, whereas Category 0 has only around 5,000 occurrences. This shows the imbalance in the dataset. This would suggest that more households or records relate to Category 1, the higher-consumption group, and fewer to Category 0, the lower-consumption group. The classes in this data set are fairly imbalanced, and in any predictive modeling work, resampling or weighting would be needed to allow fair representation across both categories.

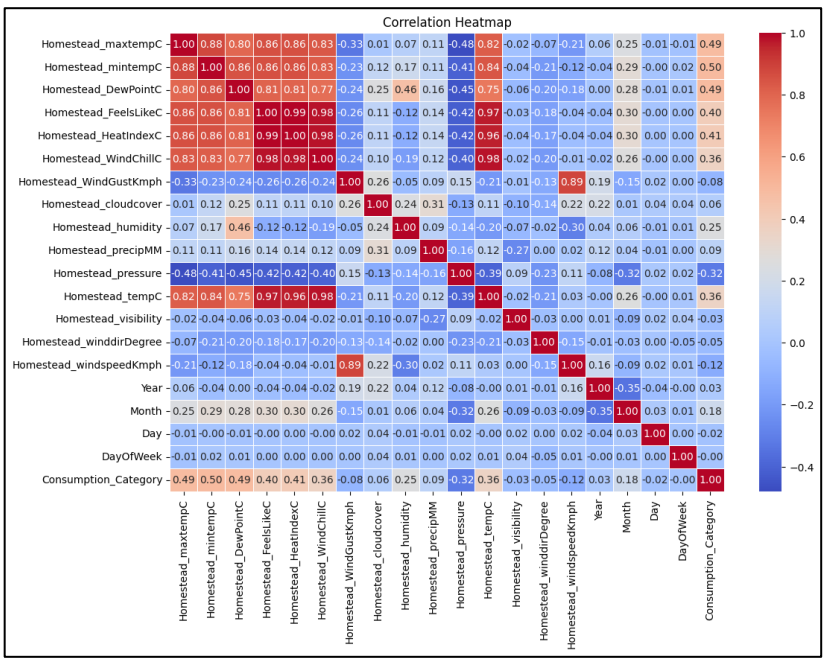

Figure 2 displays the Correlation Heatmap of Selected Features.

The above correlation heatmap describes the relationship of various weather and temporal

variables with energy consumption, represented by the "Consumption-Category" column. Where strong correlations (closer to ±1) are in red (positive) or blue (negative), weak or no correlation is shown in light shades. Some of the key findings are: The positive strong relationship is depicted by Homestead-maxtempC, Homestead_tempC with Consumption-Category at approximately 0.49 and 0.50, indicating that consumption of energy would go up when temperatures increase to a point, logically by using cooling appliances. Likewise, Homestead-Dew-PointC and Homestead-FeelsLikeC indicate a good moderate positive correlation around 0.49. Lastly, on negative notes: There is a negative correlation in the variables such as Homestead-pressure of about -0.48, meaning lesser atmospheric pressure corresponds to a high consumption of electricity. The temporal variables are Year, Month, and Day-related, and thus are less than ±0.20, to present their weak correlations as having minor influences on the consumption pattern. These will give some insights into the most significant predictors of energy use to be considered in the modeling.

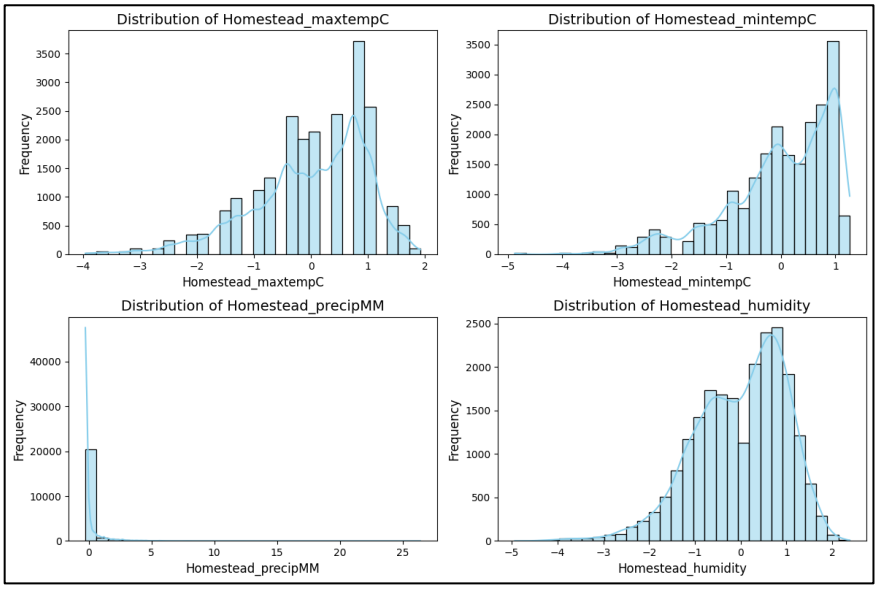

Figure 3: Exhibits Distribution of Key Features

The above histograms represent the distribution for the four important weather-related variables in the data: Homestead-maxtempC, Homestead-mintempC, Homestead-precipMM, and Homestead-humidity. The two temperature variables-maxtempC and mintempC-c are seen to be normally distributed; most of their values are around 0 and taper off symmetrically on both sides, indicating these are standardized data with a mean close to 0. Homestead_precipMM is very right-skewed since most of the values are around 0; therefore, for most of the observations, either a small amount of rainfall happened, or it was dry. However, the long tail of the distribution denotes very rare but most intense rainfall. Homestead humidity, again, is normally distributed as its peak is around 0, and its spread is balanced about the mean. Hence, these distributions are forwarded by a set of nice weather conditions in which temperature and humidity follow expected bell shapes, where precipitation shows up as sparse. Such analysis is needed to understand the variability that there is in data. For subsequent modeling strategies to make sense, these measures must be calculated and appropriate adjustments made.

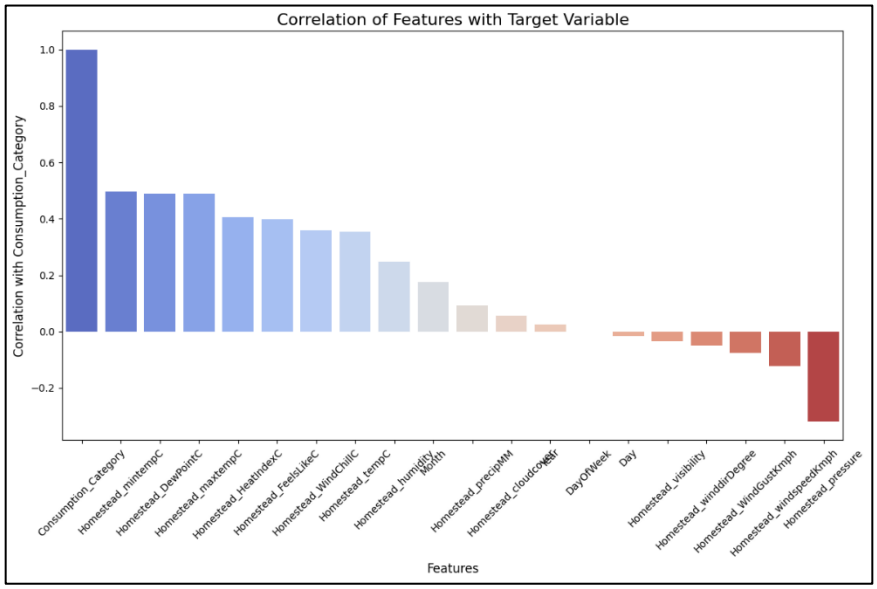

Figure 4: Visualizes the Correlation of Features with Target Variable

The histogram "Correlation of Features with Target Variable" depicts various features against the target variable, "Consumption_Category." It is observed that "Homestead_mintempC" has the highest positive value of 0.93, indicating that it is highly correlated with the target variable. On the other hand, "Homestead_pressure" has the highest negative correlation, amounting to -0.22, thus showing the opposite behavior. Surprisingly, some features like "Month," "Day," and "Homestead_visibility" are almost uncorrelated with the target variable. This analysis suggests that temperature factors, especially minimum temperature, have a very important role in influencing the consumption category, while other weather parameters like atmospheric pressure have less pronounced effects.

METHODOLOGY

Feature Engineering and Selection

Feature engineering involves transforming raw data into formats that will be most meaningful to the machine learning models, hence improving predictability. First, the transformation of weather variables into a format that was relevant to energy use. For example, temperature data at maximum, minimum, and average were transformed into derived metrics such as the range of temperature or deviation from seasonal averages to better capture energy usage patterns driven by heating or cooling needs. Similarly, features like day of the week, month, and season are derived from date-time columns to include temporal energy usage trends. Features like "precipitation intensity" and "humidity index" were engineered by keeping more variables combined to better grasp their respective effects on energy demands. In developing predictive models, feature selection techniques were used to choose those features that are most indicative. This will help thin out the dataset and thereby improve model efficiency. The correlation analysis was therefore performed between the individual variables and the target variable "Consumption_Category." Features highly positively or negatively correlated with the target variable-for example, temperature or humidity-were chosen, whereas any redundant variables showing multicollinearity, for example, Homestead_maxtempC and Homestead_tempC, are removed to avoid overfitting. The present study

leverages advanced RFE, a method applied for feature selection, along with feature importance rankings using tree-based models such as Random Forest. In such a way, all further steps of data focusing have been done on those variables that possess the highest predictive values, thus assuring the top performance of the model at minimal computational complexity.

Model Selection and Justification

The nature of the data and the goals of the study guide the choice of machine learning models to adopt for energy consumption prediction. For this research project, three models were selected: Linear Regression, Random Forest, and Support Vector Machines, each possessing particular strengths for the nature of the problem.

Linear Regression: We used linear regression because this algorithm is both simple and explainable. The performance of this algorithm means that it will serve as an excellent baseline. This algorithm assumes a linear relationship among input features concerning the target variable, hence drawing useful interpretations for understanding how things like temperature or precipitation on their own would relate to energy consumption. However, possible nonlinearities are one limitation of Linear Regression.

Random Forest: This algorithm is considered among the most powerful ensemble methods incorporating decision trees that can nicely handle nonlinear relationships and intricate feature interactions. It is robust concerning outliers and allows extraction of feature importance and relative importance of different independent variables in explaining the data-which can be then used to get a deeper understanding of the underlying drivers of consumption. Random Forest also naturally supports categorical and numerical input data.

The Support Vector Machines: This model was adopted for its capability to capture, through kernel functions, a non-linear relationship between variables. Therefore, SVMs are fitted for classification problems like those of "high" against "low" energy consumption classes, where the decision boundary may not be linear. They also perform well in high-dimensional space, making them ideal in this case since the datasets have many engineered features.

Training and Testing Framework

The dataset was split into training and testing sets for performance evaluation and generalizability. A typical split ratio was used: 80% for training the models and 20% for testing. This split ensured that the models were trained on a substantial portion of the data with a separate dataset to check their predictive power. To further enhance model reliability, k-fold cross-validation was used during training. This technique involved dividing the training set into k subsets, or folds, and iteratively training the model on k-1 folds while validating on the remaining fold. For this study, a 10-fold cross-validation was chosen, which offers a good balance between computational efficiency and robustness of performance. This approach ensures that the model is evaluated on different subsets of the training data, reducing overfitting, and providing a more general estimate of its performance. Hyperparameter tuning was therefore conducted as part of the training framework to optimize performance. Techniques such as grid search and random search were applied to test various combinations of hyperparameters for each model. For instance, Random Forest's number of trees and maximum depth were in turn varied to find the configuration that best suited the data; similarly, the type of kernel and regularization parameter have been fine-tuned in the case of the SVM model. Moreover, all models have in common the use of several performance metrics for their evaluation, such as accuracy, precision, recall, and F1-score, to

comprehensively provide a measure of each model's predictive ability.

Hyperparameter Tuning

One of the important steps to optimize a machine learning model toward its best performance is hyperparameter tuning. The essence of hyperparameter tuning involves the process of choosing the most appropriate configuration of model parameters that cannot directly be learned from the data during training. Among the most widely used techniques for hyperparameter tuning are grid search and random search. Grid search systematically evaluates all possible combinations of pre-defined hyperparameter values. For example, while tuning a random forest model, one might test different values via grid search for n_estimators, max_depth, and min_samples_leaf. While it is exhaustive, performing a grid search can be computationally expensive for large parameter spaces. On the other hand, random search samples random combinations of hyperparameters within the specified range, offering a much more lightweight way of searching through large spaces. For example, if the algorithm is SVM, the random search might navigate through combinations of kernel kinds and regularization parameters (C). Cross-validation is used in both methods to ensure good generalization across different folds of data.

Evaluation Metrics

There are significant key metrics that are needed so the model's performance concerning the prediction, actually going into the real world, is accurately conveyed. In that respect, this study employed key performance metrics such as precision, recall, F1-Score, and accuracy. Accuracy is the ratio of correctly predicted instances concerning the total; it's a measure of overall performance, but it can be misleading in the case of class imbalance in datasets. Precision is the proportion of true positives to those that were predicted as positive, and recall is the proportion of true positives to the actual. F1-Score, revolves around the harmonic mean of precision and recall, offering a balanced measure for classification problems. Model performance will be weighed against the base models, such as linear regression, as well as against those in the literature to ascertain relative performances. Comparisons such as these assure that the approach selected is robust and that much gain was realized through advanced technique applications such as hyperparameter tuning.

Results

Model Performance

a. Logistic Regression

Table 1: Showcases the Logistic Regression Classification Report

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.69	0.54	0.60	980
1	0.88	0.93	0.90	3461
accuracy			0.84	4441
macro avg	0.78	0.73	0.75	4441
weighted avg	0.84	0.84	0.84	4441

Logistic Regression Accuracy Score: 0.8448547624408916

The classification report above shows the results of a logistic regression model on a binary classification problem. Overall, the model's accuracy is 0.84, indicating that the model has correctly predicted the class labels in 84% of cases. In Class: 0, the precision value is 0.69, recall is 0.54, and F1-score is 0.60. The higher precision for class 0 indicates that the model can correctly identify instances of this class but might miss a few true positives. The precision for class 1 is 0.88, the recall is 0.93, and the F1 score is 0.90. This suggests very good performance in identifying instances of class 1 with quite strong precision and recall. Overall measures are given by the macro-average F1-score of 0.75 and weighted-average F1-score of 0.84, considering both classes.

b) Random Forest

Table 2: Depicts the Random Forest Classification Report

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.83	0.73	0.78	980
1	0.93	0.96	0.94	3461
accuracy			0.91	4441
macro avg	0.88	0.85	0.86	4441
weighted avg	0.91	0.91	0.91	4441

Random Forest Accuracy Score: 0.9088043233505967

This classification report presents the performance of a Random Forest model on a binary classification task. Overall, the model is 0.91 accurate, which means it classifies the class labels correctly in 91% of the cases. Class 0 precision is 0.83, recall is 0.73, and F1-score is 0.78. This means that it is good at correctly identifying instances of class 0 but might miss some true positives. In the case of class 1, the precision is 0.93, the recall is 0.96, and the F1 score is 0.94. This represents a good identification of class 1 instances with both high precision and recall. The overall measures of the model performance are the macro-average F1-score of 0.85 and weighted-average F1-score of 0.91, considering both classes.

c) Support Vector Machines

Table 3: Portrays the SVM Classification Report SVM Classification Report:

SVM Classification Report:

	precision	recall	f1-score	support
0	0.75	0.57	0.65	980
1	0.89	0.95	0.92	3461
accuracy			0.86	4441
macro avg	0.82	0.76	0.78	4441
weighted avg	0.86	0.86	0.86	4441

SVM Accuracy Score: 0.8630938977707724

The Support Vector Machine classification report highlights the performances of the model on its assigned binary classification task. Under this model, the mean accuracy is 0.86, which correctly identifies 86% of the real labels. At class 0, precision is 0.75, recall is 0.57, and F1-score is 0.65, which explains how the model would be quite correct in identifying the cases of class 0 yet usually misses some true positives in the process. The class 1 precision is 0.89, recall is 0.95, and F1-score is 0.92, which reflects very good performance concerning identifying instances of class 1 with high precision and recall. The macro-average F1-score of 0.78 and weighted-average F1-score of 0.86 give overall measures for model performance, considering both classes.

Model Accuracy Comparison

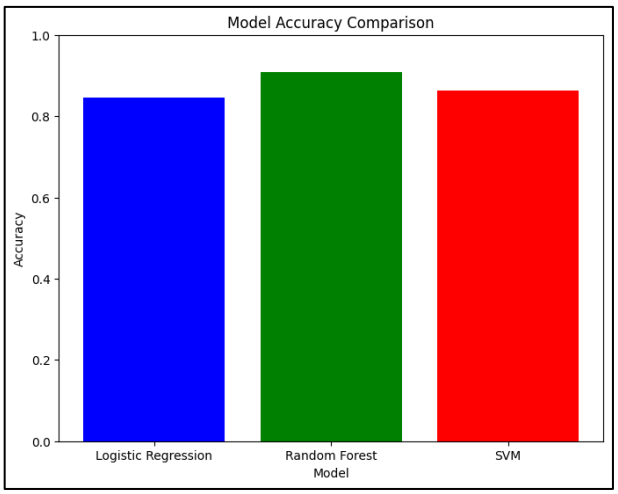

Figure 5: Displays the Model Accuracy Comparison

The bar chart "Model Accuracy Comparison" shows the accuracy score of three classification models: Logistic Regression, Random Forest, and SVM. Of these, the Random Forest model has the highest value for accuracy, about 0.9, while for the SVM model, this measure is about 0.85. The Logistic Regression model is the worst concerning value accuracy, having about 0.83. This suggests that, overall, the random forest model is the best classifier in this comparison, whereas logistic regression performs the worst for this dataset.

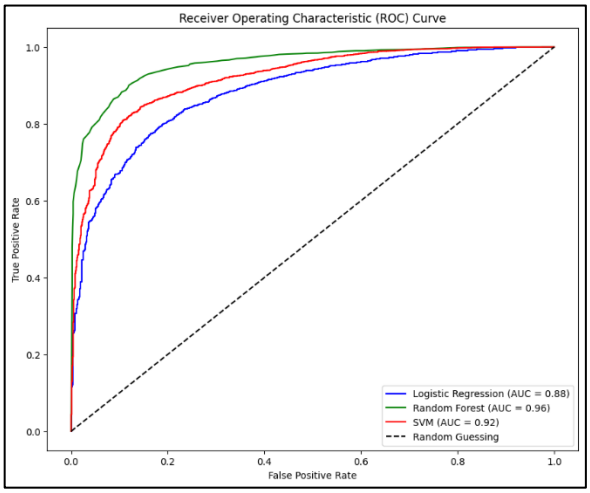

Figure 6: Exhibits the ROC Curve Comparison of Models

The ROC curve above presents the different performances for three different classification models: LR, RF, and SVM. Overall, the performance of the model can be summarized by the AUC of its curve. It is derived that the highest AUC was for the Random Forest with an AUC of 0.96, reflecting very high capability in the separation of the positive and negative classes. The SVM model follows closely with an AUC of 0.92, while the Logistic Regression model has the lowest AUC of 0.88. Based on this comparison, it can be concluded that the Random Forest model provides the best trade-off between true positive rate and false positive rate and can be relied upon for this classification task.

Socioeconomic Factors, Impact

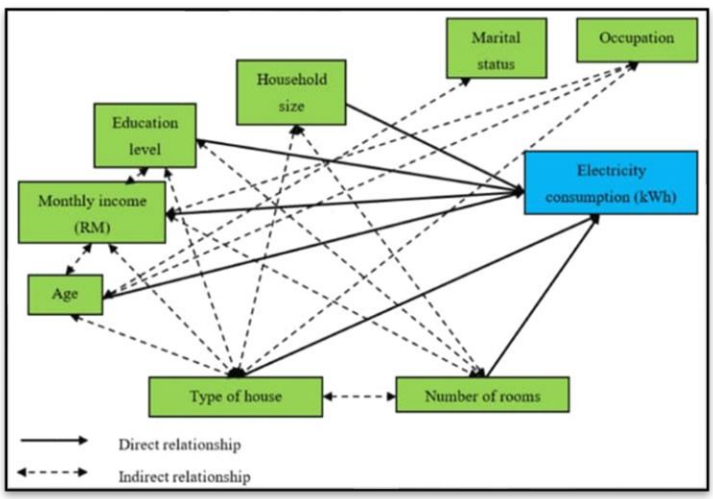

Figure 7: Visualizes the Socioeconomic Factors Impacting Energy Consumption

Socioeconomic factors remain important in determining household energy consumption patterns. Income levels strongly influence energy use because higher-income households tend to have larger homes, more energy-consuming appliances, and more heating or cooling systems. Lower-income households may be more energy-conserving because of budgetary pressures.

Family size is another factor affecting energy demand; large families generally require more energy for cooking, lighting, and heating than small households. Education plays a similar role in affecting energy behavior: better-educated households may have a higher probability of investing in energy-efficient technologies or practices, including programmable thermostats and renewable sources of energy. Other factors at play that may cause the gaps to widen are the place of residence and ownership of the home.

Predictive Insights

The machine learning models generate valuable predictions about household energy use. Examples include Random Forest models, which are trained on socioeconomic and weather data to predict the likelihood of a given household having high energy usage. These predictions can be used to help energy providers determine when to invoke tiered pricing or encourage energy-saving behavior. Case studies by the United Nations have consistently highlighted how these types of predictions can be applied in practice. It encompasses developing, curating, and executing predictive models that can certainly identify low-income households with disproportionately high energy use, on whom assistance program subsidies or the installation of energy-efficient appliances can be concentrated. The insight provides a clear example of how a data-driven approach may reduce energy waste and ensure socioeconomic equity in the pursuit of sustainable energy.

DISCUSSION

Energy Providers' Implications

These developed predictive models in the present study have much relevance to helping energy providers improve efficiency in energy management. Precise prediction of household energy consumption can help the energy provider in operational optimization, cost reduction, and enhancement of consumer satisfaction. For example, such models will help demand-side management strategies by estimating the time of peak consumption and delivering focused interventions in the form of time-of-use pricing or demand response programs. Most crucial for the energy providers, models currently predict future energy demand in trends, which helps Capacity Planners and Infrastructure Investors adapt to the future. Having integrative machine learning insights does not come easy; accordingly, the following recommendations might be put forward:

Data Quality and Collection: Emphasize the collection of complete and quality data, such as comprehensive household energy consumption history, weather data, and factors relating to socio-demography.

Model Development and Validation: Robust model development and validation with the investment of time and computational resources to ensure accuracy in the predictions. Update and retrain models periodically to maintain a good fit for the dynamically changing energy consumption patterns or other external factors.

Collaboration and Partnerships: Collaborate with researchers, and data scientists in advanced Analytics techniques and share best practices across the board.

Customer Engagement and Education: Educate the customer on the benefits of energy efficiency through personalized recommendations and incentives.

Policy Development Implications

These findings have crucial implications for policymakers in the pursuit of increased energy efficiency and sustainability in the USA. With clear knowledge of what determines energy demand at the household level, policymakers can thus institute efficient policies and incentives to stimulate reduced consumption. For instance,

policymakers would be able to structure and implement policies on energy-efficient appliances and building retrofits, renewable energy technologies, or shifting to renewable sources. From this perspective, some strategies or measures that policymakers might enact to address the disparity among socioeconomic groups in energy consumption are the following:

Targeted Subsidies and Incentives: These include financial contributions that are meant for low-income households toward their investment in energy efficiency enhancement and renewable energy systems.

Energy Education and Awareness Campaigns: The Government should embark on public awareness campaigns to educate households about how best they can save energy in the process of realizing the accruable benefits from sustainable use of the same.

Community-Based Energy Programs: Support community-based initiatives that promote energy efficiency and renewable energy, such as community solar projects and energy co-ops.

Limitations and Challenges

While this research project affords valuable insights, it is pivotal to acknowledge the ethical considerations and limitations related to using household data for analysis. First and foremost, such data needs to ensure privacy through anonymity and security. Also, some machine learning models, while powerful in terms of complexity, make it hard to interpret what is going on or which underlying factors influence energy consumption patterns. The generalizability of the findings may also be limited by the specific characteristics of the dataset and the geographical location of the study. Future research should explore the applicability of these models to diverse populations and regions.

Future Research Directions

Future research can advance this understanding of energy household consumption by opening several other avenues:

Data Expansion and Diversity: Larger and more diverse datasets, including those from various geographical regions and different socioeconomic groups, will enhance model performance and generalizability.

Real-time integration: Researchers can merge real-time energy consumption data with weather conditions and energy prices to arrive promptly and even more precise forecasts.

Socioeconomic Factors and Behavioral Insights: Elaborating on socioeconomic factors and psychological biases can bring forth more effective interventions in changing energy consumption behavior.

Advanced Machine Learning Techniques: Scholars can explore how the use of techniques like deep learning and reinforcement learning can open up new paths toward better model performance.

CONCLUSION

The utmost objective of this research project was to develop predictive models using machine learning techniques to analyze household energy consumption trends in the USA, integrating socioeconomic factors such as income, family size, and education. The dataset retrieved from Kaggle integrates detailed weather patterns with energy consumption data, putting into perspective the interaction between climatic variables and household energy use. It included key features such as temperature, humidity, wind speed, and precipitation, along with time-series data on energy consumption metrics like electricity and natural gas usage at the household level. It provided information on several geographic zones across extended periods, so seasonality and regional variations may be studied. It was complemented with metadata that included

timestamps, energy pricing, and household attributes and should therefore be a rich resource for predictive modeling and extracting relationships between weather conditions and energy demand. For this research project, three models were selected: Logistic Regression, Random Forest, and Support Vector Machines, each possessing particular strengths for the nature of the problem. This study employed key performance metrics such as precision, recall, F1-Score, and accuracy. The Random Forest model had the highest value for accuracy, similarly, the highest AUC was for the Random Forest with the best AUC. As such, it was concluded that the Random Forest model provided the best trade-off between true positive rate and false positive rate and can be relied upon for this classification task. The machine learning models generate valuable predictions about household energy use. Particularly, Random Forest models, which are trained on socioeconomic and weather data to predict the likelihood of a given household having high energy usage. The predictions by such models can be used to help energy providers determine when to invoke tiered pricing or encourage energy-saving behavior.

REFERENCES

1. Alam, M., Islam, M. R., & Shil, S. K. (2023). AI-Based Predictive Maintenance for US Manufacturing: Reducing Downtime and Increasing Productivity. International Journal of Advanced Engineering Technologies and Innovations, 1(01), 541-567.
2. Al Mukaddim, A., Nasiruddin, M., & Hider, M. A. (2023). Blockchain Technology for Secure and Transparent Supply Chain Management: A Pathway to Enhanced Trust and Efficiency. International Journal of Advanced Engineering Technologies and Innovations, 1(01), 419-446.
3. Amiri, S. S., Mueller, M., & Hoque, S. (2023). Investigating the application of a commercial and residential energy consumption prediction model for urban Planning scenarios with Machine Learning and Shapley Additive explanation methods. Energy and Buildings, 287, 112965.
4. Buiya, M. R., Laskar, A. N., Islam, M. R., Sawalmeh, S. K. S., Roy, M. S. R. C., Roy, R. E. R. S., & Sumsuzoha, M. (2024). Detecting IoT Cyberattacks: Advanced Machine Learning Models for Enhanced Security in Network Traffic. Journal of Computer Science and Technology Studies, 6(4), 142-152.
5. Buiya, M. R., Alam, M., & Islam, M. R. (2023). Leveraging Big Data Analytics for Advanced Cybersecurity: Proactive Strategies and Solutions. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 882-916.
6. Charfeddine, L., Zaidan, E., Alban, A. Q., Bennasr, H., & Abulibdeh, A. (2023). Modeling and forecasting electricity consumption amid the COVID-19 pandemic: Machine learning vs. nonlinear econometric time series models. Sustainable Cities and Society, 98, 104860.
7. Chen, G., Hu, Q., Wang, J., Wang, X., & Zhu, Y. (2023). Machine-learning-based electric power forecasting. Sustainability, 15(14), 11299.
8. Debnath, P., Karmakar, M., Khan, M. T., Khan, M. A., Al Sayeed, A., Rahman, A., & Sumon, M. F. I. (2024). Seismic Activity Analysis in California: Patterns, Trends, and Predictive Modeling. Journal of Computer Science and Technology Studies, 6(5), 50-60.
9. Gazi, M. S., Nasiruddin, M., Dutta, S., Sikder, R., Huda, C. B., & Islam, M. Z. (2024). Employee Attrition Prediction in the USA: A Machine Learning Approach for HR Analytics and Talent Retention Strategies. Journal of Business and Management Studies, 6(3), 47-59.

10. Goriparthi, R. G. (2024). AI-Driven Predictive Analytics for Autonomous Systems: A Machine Learning Approach. Revista de Inteligencia Artificial en Medicina, 15(1), 843-879.

11. Hasan, M. R. (2024). Revitalizing the electric grid: A machine learning paradigm for ensuring stability in the USA. Journal of Computer Science and Technology Studies, 6(1), 141-154.

12. Hasanuzzaman, M., Hossain, S., & Shil, S. K. (2023). Enhancing Disaster Management through AI-Driven Predictive Analytics: Improving Preparedness and Response. International Journal of Advanced Engineering Technologies and Innovations, 1(01), 533-562.

13. Islam, M. Z., Islam, M. S., Al Montaser, M. A., Rasel, M. A. B., Bhowmik, P. K., & Dalim, H. M. (2024). EVALUATING THE EFFECTIVENESS OF MACHINE LEARNING ALGORITHMS IN PREDICTING CRYPTOCURRENCY PRICES UNDER MARKET VOLATILITY: A STUDY BASED ON THE USA FINANCIAL MARKET. The American Journal of Management and Economics Innovations, 6(12), 15-38.

14. Islam, M. R., Nasiruddin, M., Karmakar, M., Akter, R., Khan, M. T., Sayeed, A. A., & Amin, A. (2024). Leveraging Advanced Machine Learning Algorithms for Enhanced Cyberattack Detection on US Business Networks. Journal of Business and Management Studies, 6(5), 213-224.

15. Kapp, S., Choi, J. K., & Hong, T. (2023). Predicting industrial building energy consumption with statistical and machine-learning models informed by physical system parameters. Renewable and Sustainable Energy Reviews, 172, 113045.

16. Kesriklioğlu, E., Oktay, E., & Karaaslan, A. (2023). Predicting total household energy expenditures using ensemble learning methods. Energy, 276, 127581.

17. Khan, M. T., Akter, R., Dalim, H. M., Sayeed, A. A., Anonna, F. R., Mohaimin, M. R., & Karmakar, M. (2024). Predictive Modeling of US Stock Market and Commodities: Impact of Economic Indicators and Geopolitical Events Using Machine. Journal of Economics, Finance and Accounting Studies, 6(6), 17-33.

18. Karmakar, M., Debnath, P., & Khan, M. A. (2024). AI-Powered Solutions for Traffic Management in US Cities: Reducing Congestion and Emissions. International Journal of Advanced Engineering Technologies and Innovations, 2(1), 194-222.

19. Kumar, S. (2023). A novel hybrid machine learning model for prediction of CO2 using socio-economic and energy attributes for climate change monitoring and mitigation policies. Ecological Informatics, 77, 102253.

20. Mukelabai, M. D., Wijayantha, K. G. U., & Blanchard, R. E. (2023). Using machine learning to expound energy poverty in the global south: Understanding and predicting access to cooking with clean energy. Energy and AI, 14, 100290.

21. Nasiruddin, M., Al Mukaddim, A., & Hider, M. A. (2023). Optimizing Renewable Energy Systems Using Artificial Intelligence: Enhancing Efficiency and Sustainability. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 846-881.

22. Nguyen, V. G., Duong, X. Q., Nguyen, L. H., Nguyen, P. Q. P., Priya, J. C., Truong, T. H., ... & Nguyen, X. P. (2023). An extensive investigation on leveraging machine learning techniques for high-precision predictive modeling of CO2 emission. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 45(3), 9149-9177.

23. Ohalete, N. C., Aderibigbe, A. O., Ani, E. C., Ohenhen, P. E., & Akinoso, A. E. (2023). Data science in energy consumption analysis: a review of AI techniques in identifying patterns and efficiency opportunities. Engineering Science & Technology Journal, 4(6), 357-380.

24. Omogoroye, O. O., Olaniyi, O. O., Adebiyi, O. O., Oladoyinbo, T. O., & Olaniyi, F. G. (2023). Electricity consumption (kW) forecast for a building of interest based on a time series nonlinear regression model. Asian Journal of Economics, Business and Accounting, 23(21), 197-207.

25. Palani, H., Acosta-Sequeda, J., Karatas, A., & Derrible, S. (2023). The role of socio-demographic and economic characteristics on energy-related occupant behavior. Journal of Building Engineering, 75, 106875.

26. Pro-AI-Rokibul. (2024). Machine-Learning-FOr-Optimizing-USA-Household-Energy-Consumption/Models/main.ipynb at main · proAIrokibul/Machine-Learning-FOr-Optimizing-USA-Household-Energy-Consumption. GitHub. https://github.com/proAIrokibul/Machine-Learning-FOr-Optimizing-USA-Household-Energy-Consumption/blob/main/Models/main.ipynb

27. Rahman, M. K., Dalim, H. M., & Hossain, M. S. (2023). AI-Powered Solutions for Enhancing National Cybersecurity: Predictive Analytics and Threat Mitigation. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 1036-1069.

28. Rahman, A., Debnath, P., Ahmed, A., Dalim, H. M., Karmakar, M., Sumon, M. F. I., & Khan, M. A. (2024). Machine learning and network analysis for financial crime detection: Mapping and identifying illicit transaction patterns in global black money transactions. Gulf Journal of Advance Business Research, 2(6), 250-272.

29. Sarwar, S., Aziz, G., & Tiwari, A. K. (2024). Implication of machine learning techniques to forecast the electricity price and carbon emission: Evidence from a hot region. Geoscience Frontiers, 15(3), 101647.

30. Singh, A., Yadav, J., Shrestha, S., & Varde, A. S. (2023). Linking alternative fuel vehicles adoption with socioeconomic status and air quality index. arXiv preprint arXiv:2303.08286.

31. Shi, Z., Wu, L., & Zhou, Y. (2023). Predicting household energy consumption in an aging society. Applied Energy, 352, 121899.

32. Shawon, R. E. R., Chowdhury, M. S. R., & Rahman, T. (2023). Transforming Urban Living in the USA: The Role of IoT in Developing Smart Cities. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 917-953.

33. Shawon, R. E. R., Miah, M. N. I., & Islam, M. Z. (2023). Enhancing US Education Systems with AI: Personalized Learning and Academic Performance Prediction. International Journal of Advanced Engineering Technologies and Innovations, 1(01), 518-540.

34. Shil, S. K., Chowdhury, M. S. R., Tannier, N. R., Tarafder, M. T. R., Akter, R., Gurung, N., & Sizan, M. M. H. (2024). Forecasting Electric Vehicle Adoption in the USA Using Machine Learning Models. Journal of Computer Science and Technology Studies, 6(5), 61-74.

35. Shawon, R. E. R., Rahman, A., Islam, M. R., Debnath, P., Sumon, M. F. I., Khan, M. A., & Miah, M. N. I. (2024). AI-Driven Predictive Modeling of US Economic Trends: Insights and Innovations. Journal of Humanities and Social Sciences Studies, 6(10), 01-15.

36. Sumon, M. F. I., Osiujjaman, M., Khan, M. A.,

Rahman, A., Uddin, M. K., Pant, L., & Debnath, P. (2024). Environmental and Socio-Economic Impact Assessment of Renewable Energy Using Machine Learning Models. Journal of Economics, Finance and Accounting Studies, 6(5), 112-122.

37. Zhussupbekov, M., Memon, S. A., Khawaja, S. A., Nazir, K., & Kim, J. (2023). Forecasting energy demand of PCM integrated residential buildings: A machine learning approach. Journal of Building Engineering, 70, 106335.