World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A data-driven approach to gas demand prediction in the USA using machine learning

Nayem Uddin Prince [1,*], Mohd Abdullah Al Mamun [2], Md Mehedi Hassan Melon [3], Anwar Hossain [4], Yasin Arafat [5] and Mohammad Amit Hasan [6]

[1] *Student, Department of Computer Science and Engineering, Daffodil International University, Bangladesh.*
[2] *Student, Department of Business Administration, BRAC University, Bangladesh.*
[3] *Student, Department of Electrical and Automation Engineering, Nanjing Tech University, China.*
[4] *Student, Department Electrical and Electronic Engineering, University of Asia Pacific, Bangladesh.*
[5] *Student, Department of Business Administration, North South University, Bangladesh.*
[6] *Student, Department of Business Administration, University of Liberal Arts, Bangladesh.*

World Journal of Advanced Research and Reviews, 2020, 05(02), 193–203

Publication history: Received on 05 January 2020; revised on 17 February 2020; accepted on 20 February 2020

Article DOI: https://doi.org/10.30574/wjarr.2020.5.2.0002

Abstract

Natural gas is crucial for energy generation, industrial production, and residential heating in the United States. The demand is difficult to forecast due to economic fluctuations, energy price instability, and seasonal temperature variations. Time series analysis and linear regression may fail to account for the nonlinear interactions in gas demand data, potentially resulting in inaccuracies. This work addresses the necessity for precise gas demand projections, which are essential for energy planning and resource management. Natural gas is crucial for electricity production, industrial processes, and residential heating; nevertheless, demand is influenced by weather conditions, economic activity, and energy pricing. Conventional forecasting models are inadequate for capturing these intricate processes, necessitating the employment of more sophisticated predictive methodologies. This research employs decision trees, linear regression, gradient boosting, and random forests to examine data from 10 significant US states spanning 2000 to 2019. The random forest model accurately anticipated demand patterns, achieving an R-squared of 99.67% and a root mean square error (RMSE) of 34.53. These findings demonstrate that machine learning can elucidate nonlinear relationships in gas demand data. The study provides a framework for improved demand forecasting, aiding energy providers and legislators in optimising resource allocation, increasing cost-efficiency, and promoting environmental sustainability.

Keywords: R-squared; RMSE; XGBoost; MLP; Scatter plot; Demand. Null value; Label Encoding; Distribution; State; statistical.

1. Introduction

Anticipating energy demand is crucial for resource allocation and administration in the contemporary economy, especially regarding natural gas. Natural gas's adaptability and diminished environmental impact relative to other fossil fuels have resulted in a significant rise in its usage in the United States. Natural gas is employed in transmission, electricity production, industrial operations, residential heating, and transportation. Accurate demand forecasting improves market efficiency, price stability, and infrastructure planning, thus ensuring a dependable and continuous energy supply. By accurately forecasting gas demand, energy suppliers may fulfil supply requirements, reduce operational expenses, and support sustainability initiatives by eliminating waste and overproduction. Forecasting gas demand is difficult due to its reliance on various intricate and dynamic elements, including but not limited to economic cycles, industrial activity, regional population density, and seasonal fluctuations. Linear regression and time series models are employed in conventional energy demand forecasting. Although effective for identifying seasonal patterns and overarching trends, these methods are inadequate for illustrating the intricate nonlinear dynamics that influence gas demand. Linear models may neglect swift rule alterations or the correlation between temperature and industrial

* Corresponding author: Nayem Uddin Prince

production. In 2019, increased power generation and a favourable pricing landscape propelled natural gas demand to unprecedented levels in the United States. Daily natural gas consumption in the United States rose 3% to 85.0 BCF compared to the prior year. Electric power constituted 31.0 Bcf/d, or 36% of national consumption, reflecting a 7% rise attributed to the sustained reliance on natural gas as the principal fuel source in the United States [1]. A shift towards cleaner, more efficient energy was apparent as natural gas exceeded coal, generating 38% of the nation's electricity instead of 23%. Henry Hub natural gas prices reached a record low that year, averaging $2.57/MMBtu, the lowest since 2016 [2]. The market saw an oversupply, decreasing prices for end-users across many industries due to unprecedented shale play production, especially in the Appalachian region. Consequently, machine learning techniques are utilised to improve the accuracy and adaptability of gas demand forecasting models. Machine learning algorithms can manage extensive and intricate datasets, uncovering complicated links, correlations, and patterns that traditional models overlook. Adaptability is essential due to the numerous elements that must be considered to establish a robust framework for gas demand forecasting. This study utilises multiple machine-learning algorithms, including decision trees, linear regression, gradient boosting, random forests, and MLP regression, to predict gas consumption in the United States, rectifying deficiencies in prior approaches. Each approach has its advantages. Ensemble methods, like gradient boosting and random forest, excel at managing large datasets with intricate, nonlinear interactions. MLP regression, a type of neural network, is advantageous in scenarios with multiple interacting input variables, as it may identify more nuanced relationships. Although seemingly straightforward, decision trees can effectively reveal the determinants of demand in certain areas. This study analyses various algorithms to identify the most precise and efficient gas demand forecasting model for the United States. Our analysis employed gas consumption data from Kaggle spanning 2000 to 2019. This dataset includes gas consumption statistics from Michigan, Ohio, Pennsylvania, Louisiana, Florida, New Jersey, New York, and Pennsylvania, presented monthly or annually. The demographic, industrial, climatic, and energy consumption diversity of the United States is evident in these states. Louisiana is a hub for natural gas processing, and its climate generates a seasonal demand that differs from California's, which has a substantial population and significant industrial requirements. Incorporating diverse states yields a more robust and generalisable model applicable to different domains [3]. This study illustrates the essential significance of sophisticated forecasting in fulfilling the demand for natural gas. An accurate demand forecast is crucial for fulfilling energy requirements, optimising the energy supply chain, lowering costs, and mitigating the environmental effects of overproduction. Comprehending demand trends enables governments and energy companies to formulate improved long-term strategies, adhere to rules, and allocate investments in infrastructure. Precise forecasting can facilitate the responsible and sustainable control of natural gas consumption during the transition to a low-carbon future. Enhancing economic stability and shaping energy policy are advantages of precise demand forecasting for energy suppliers. Reducing energy price volatility safeguards consumers from price surges resulting from supply-demand discrepancies in demand forecasts. Energy infrastructure providers gain advantages from precise forecasting to prepare for unforeseen demand occurrences, such as meteorological irregularities or variations in industrial consumption. Governments can utilise data-driven estimations to foster energy independence and attain environmental goals, including reducing carbon emissions and incorporating renewable energy sources. This study's results indicate that gas demand forecasts in the US are more accurate when employing machine learning techniques. This study addresses the necessity for precise and dependable energy estimations by applying sophisticated models and an extensive dataset. It accounts for the intricacies of natural gas utilisation and geographical variances. Policymakers and industry stakeholders should consider the findings, as they indicate that machine-learning approaches, especially ensemble methods such as random forests, may improve demand forecasts. This study enhances initiatives to develop a robust, efficient, and environmentally sustainable energy infrastructure in response to a dynamic economy.

This paper covers the following points: Part II offers a synopsis of relevant literature. Section III discusses the methodology in depth. Section IV presents the experiment's outcomes, and Section V assesses our model. Section VI covers The Conclusion and Future Work.

2. Literature review

The literature review greatly influences research methods, theoretical framework, and direction. It allows scholars to assess existing knowledge, identify gaps, and build on previous work. A literature review summarises and evaluates past studies to contextualise new research. This will reveal field trends, disagreements, and advances [4].

Istanbul, the biggest natural gas-consuming megacity in Turkey, is located in the province of Istanbul. To accurately predict its future consumption, Beyca et. At. [5] used three different popular machine learning technologies. The MLR, ANN, and SVR tools are part of this set. The results show that the SVR is far better than the ANN approach when forecasting natural gas consumption time series, producing more accurate and dependable outcomes with more minor prediction errors.

Perrotta el. At. [6] employed three machine learning algorithms to model the fuel consumption of articulated trucks using a large dataset. Models, including SVM, RF, and ANN, have been developed and evaluated for this objective. The research indicates that although all three methodologies facilitate extremely accurate model creation, Random Forest slightly outperforms Support Vector Machine and Artificial Neural Network regarding R^2 and fewer error metrics.

Bedi el. At. [7] They propose a deep learning system for forecasting power demand by addressing long-term historical dependencies. Initially, monthly electricity consumption data undergoes cluster analysis to yield seasonally categorised information. Subsequently, trend characterisation is conducted to enhance comprehension of the metadata associated with each cluster. Furthermore, utilising seasonal data, weekday information, and intervals, multi-input, multi-output models based on LSTM Memory networks are employed to forecast electricity consumption.

Seyedzadeh el. At. [8] thoroughly examined four fundamental machine learning methodologies—artificial neural networks, support vector machines, clustering, and Gaussian distribution-based regressions—commonly employed for predicting and enhancing building energy efficiency. In new construction, prioritising energy efficiency is one of the most effective methods to reduce energy use and carbon dioxide emissions. The energy efficiency of the existing infrastructure can be enhanced through intelligent renovations and effective energy management. All of these strategies necessitate accurate energy forecasting to facilitate optimal decision-making.

Laib el. At. [9] This research introduces a novel hybrid forecasting approach that employs an MLP neural network as a nonlinear forecasting monitor to rectify the deficiencies of the two-stage method. Before generating its prediction, this model selects one from various local models to assess the gas consumption profile for the subsequent day. Initially, they examine and categorise daily natural gas consumption patterns; after that, they develop comprehensive LSTM recurrent models predicated on load behaviour. The results are compared with four conventional techniques: MLP neural networks, LSTM, diverse linear regression models, and time series models incorporating external variables STSM.

Čeperić el. At. [10] They introduce specific improvements to the SVR-based forecasting methodology. We present a method for automatically selecting model inputs and their generation utilising feature selection (FS) techniques. The reduction of subjective inputs is accomplished by implementing FS algorithms for automatic model input selection and optimising SVR hyperparameters with PSwarm, a sophisticated global optimisation method. Their findings indicate that published machine learning results often exaggerate the models' effectiveness, as we may notice only negligible improvements relative to time series approaches. Feature selection methods are employed to preselect variables in neural networks and support vector regression, which they identify as beneficial.

Potočnik el. At. [11] The models can forecast gas demand sixty hours ahead with an hourly resolution. The model's predictions are based on historical temperature data, temperature forecasts, and temporal considerations, including holiday and event indicators. Data regarding gas consumption in Ljubljana, Slovenia, was utilised to train and evaluate the models. Several machine-learning techniques were considered, including artificial neural networks, linear regression, and kernel machines. The study of data resulted in the creation of empirical models. Recurrent neural networks and linear regression models were the most precise.

Karadede el. At. [12] The main goal of their study is to show that a nonlinear regression model based on breeder hybrid algorithms can anticipate generic demand for natural gas. One key differentiator between this model and others in the literature is that the proposed model consistently evolves with the best solutions in the breeder genetic algorithm and simulated annealing sections. This is an essential aspect of natural gas demand forecasting. The proposed algorithms far outperform their literature-based counterparts.

Su el. At. [13] They propose a robust hybrid methodology for predicting gas consumption hours in advance by integrating Wavelet Transform, RNN-structured deep learning, and Genetic Algorithm. The Wavelet Transform simplifies forecasting by decomposing the initial gas load series into components. The Genetic Algorithm enhances the RNN-structured deep learning model's performance by improving each layer's configuration. This method utilises dropout technology to prevent overfitting.

Freeman el. At. [14] study initially sought to ascertain the responses of RF and SGB to various tuning parameters. Secondly, it evaluated the performance of the two models by examining the significance and interactions of predictor variables, global accuracy metrics derived from an independent test set, and the visual quality of the resulting tree canopy cover maps. RF and SGB exhibited a notable similarity in their anticipated accuracy across all four test zones. The independent test set mean squared error (MSE) exhibited a three-digit variance across all four research locations, with the most significant divergence observed in Kansas between RF and SGB (0.0113 versus 0.0117, respectively). SGB appeared to prioritise a limited number of variables more heavily than RF in the context of linked predictor factors.

Zhang el. At. [15] Developed three predictive models utilising solar capacity to forecast the hourly load in Southern California 24 hours in advance. The models included multiple linear regression, random forest, and gradient boosting techniques. Air temperature was the most significant meteorological variable, while non-meteorological variables were also crucial, including holiday, month, solar capacity, and the previous week's load. During midday in summer, when demand is elevated, all models exhibited more significant errors. Based on hourly projections, the mean error for RF, GB, and ML was 3.5%, 3.4%, and 3.1%, respectively.

Goliath el. At. [16] This project aims to evaluate the efficacy of gradient-boosting machines in predicting cooling and heating loads for residential buildings. The architectural designs were employed to generate 768 samples, incorporating two thermal output variables and eight geometric input parameters. The parameter selection underwent an exhaustive search with cross-validation. Four statistical indicators and one composite index were employed to evaluate the method's performance. Compared to other machine learning methodologies, such as Support Vector Machines and Random Forests, gradient-boosting machines consistently demonstrate superior performance.

A comprehensive literature review also demonstrates that the researcher is aware of prior work, which helps to avoid repetition of efforts. In addition to proving the study's integrity, it proves the author's command of the subject and supports its aims [17].

2.1. Comparison with other work

Table 1 shows how we compare to the competition. This chart clearly shows that the existing literature lacks algorithms and strong examples of accuracy. Current practices are not implemented. The state-of-the-art methods we used in our investigation were remarkably accurate [18]. Not long ago, cutting-edge methods weren't used. Because of this, our study is exceptional and revolutionary.

Table 1 Comparison Table with Existing Work

Other work		**Our work**
Author Name	Algorithm	Algorithm & Accuracy
Beyca el. At.	SVR, ANN	RF, GB, LG, MLP, and DT & 99.67%
Perrotta el. At.	SVM, RF, and ANN	
Bedi el. At.	LSTM	
Freeman el. At.	RF and SGB	
Zhang el. At.	RF, GB, and ML	

After analysing this table, we can say that our project is rich in algorithms and achieved the best accuracy of existing work. So, our project has performed outstandingly.

3. Material and methods

Figure 1 represents the methodology diagram of our project. Our project had five distinct stages. The collection of preliminary data required for the project is the beginning phase. Between 2000 and 2019, we concentrated on the ten phases of gas demand in the United States. Addressing absent values and converting categorical data into a numerical representation are components of the subsequent data preparation procedure. This section delineates two responsibilities: Remove Null Value: Removing or substituting nonexistent data points. To convert categorical data into numerical format, we utilised level encoding. After preprocessing, the data is systematically stored, facilitating the following operations. Additional statistical analysis or data exploration is conducted to enhance understanding of the dataset's features. This phase employs machine learning models or algorithms on the dataset to identify trends or provide predictions. Finally, we evaluate the model's efficacy.

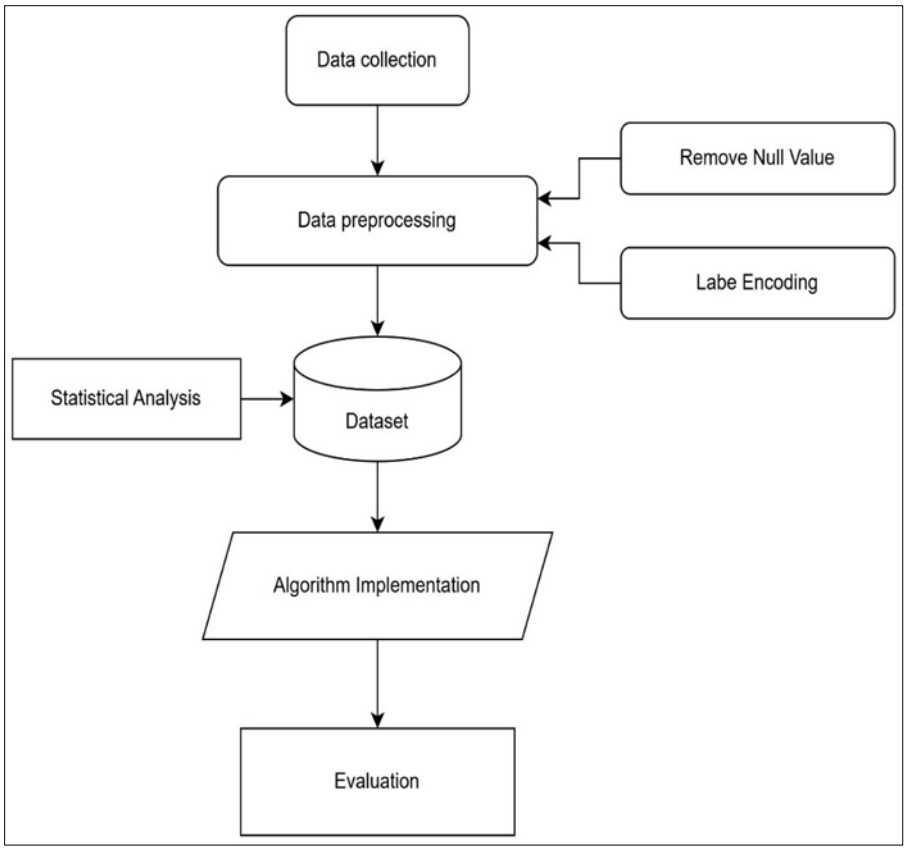

Figure 1 Methodology Diagram

3.1. Data collection

Data from multiple sources is needed to train a machine learning model to anticipate US gas consumption. Michigan, Ohio, Louisiana, Florida, New Jersey, Pennsylvania, New York, and New Jersey are covered by this Kaggle dataset from 2000 to 2019. Data on economic activity, population size, weather, and energy cost might affect gas usage. Primary data sources include utilities' real-time energy use figures and government publications [19]. Pre-processed Kaggle data is utilised as secondary data in research. The Kaggle dataset may include gas demand, geographical consumption, temperature, and seasonality. These data points are used to train trend-predicting algorithms.

Cleanse and prepare data for accurate analysis. Models must address missing values, normalise numerical variables, and structure categorical data [20]. These methods prepare data for our research's machine learning algorithms, such as gradient boosting and random forests.

3.2. Dataset pre-processing and representation

Machine learning pipeline data preparation cleans, standardises, and prepares data for analysis. Gradient boosting and random forests will prepare the dataset for machine learning algorithms in this gas demand prediction project. Label encoding and null removal are crucial preprocessing steps. Not treating missing values in real-world data effectively may affect model training. Missing data can be addressed in many ways: Removing a row or column with multiple missing values may be best [20]. Columns with few missing values can be removed to simplify the dataset without losing information. Imputing values from a scattered dataset with many missing values is possible. Machine learning models need numerical input; thus, state names and seasons must be translated to numbers. Popular label encoding assigns integers to categories. A feature called "State" containing strings like "California," "Florida," "New York," etc., would be transformed to numerical values like 0, 1, 2, etc [21]. This ensures model categorical variable processing accuracy.

3.3. Statistical analysis

We are concerned with some statistical analysis for our project with the existing dataset in this part. Gas demand as a function of state is seen in Figure 2. "Total Demand" is shown in this bar chart broken down by US state. "Total Demand" in millions (1e7 scale) is displayed on the y-axis, while states are listed on the x-axis. A breakdown of the data: The

demand in Texas is approximately 8 million, more than in other states. Roughly six million are needed by California. The demand is small, 3–4 million, in Louisiana, New York, and Illinois. The following states' demand was marginally lower than the previous group: Michigan, Ohio, Pennsylvania, and Florida. With less than 2 million in demand, New Jersey ranks last in this chart. The demand is highest in Texas and California, as seen in the figure, and lowest in the other states.

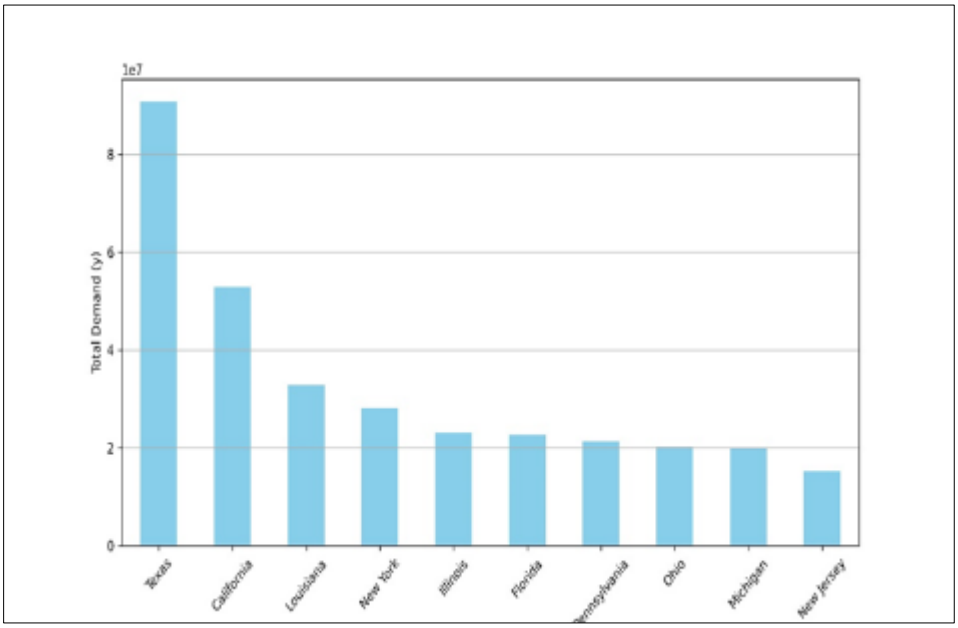

Figure 2 Gas Demand Vs State

Figure 3 shows gas demand vs total year. The "Total Demand" is a line graph from 1995 to 2020. The x-axis represents the years, while the y-axis indicates "Total Demand" in millions, as denoted by the scale of 1e7. The tendency can be encapsulated as follows: Demand exhibited moderate volatility from 1995 to 2008, largely stable within 2.2 to 2.4 million units. An evident upward trend in demand materialised circa 2009. This expansion continued until approximately 2018, with only slight sporadic deviations. A significant rise in demand occurred in recent years, with a sharp increase exceeding 3 million between 2018 and 2020. Initially, demand was somewhat stable but has consistently risen since approximately 2010.

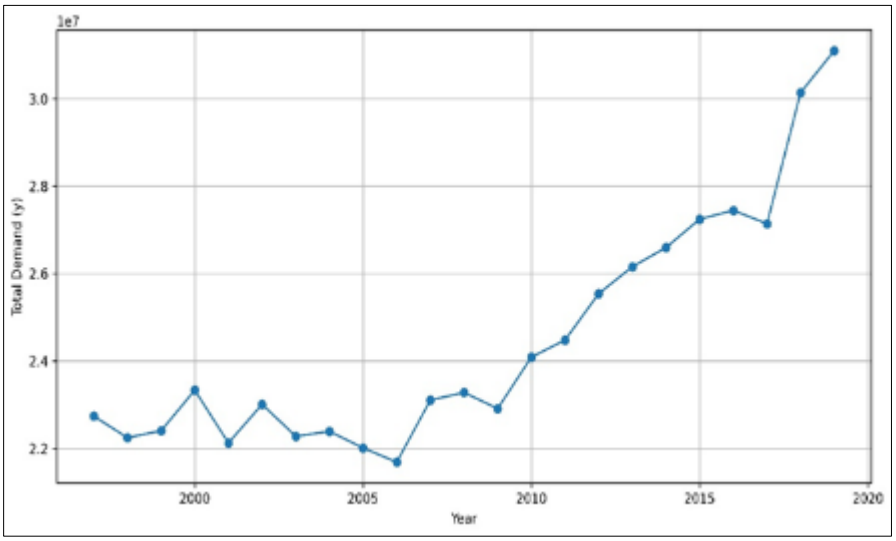

Figure 3 Gas Demand Vs Year

Figure 4 shows a box plot graph depicting the US gas demand by state. All state gas demand distribution medians, IQRs, and outliers are shown in each box figure. Discover features and insights in this full breakdown: The y-axis shows gas demand without units. The axis notation (1e6) allows millions of cubic feet or barrels. X-Axis states: The x-axis shows each state and "Federal Offshore - Gulf of Mexico" for gas demand comparison. The extensive range of box heights shows that gas demands vary by state. Wider boxes indicate demand uncertainty in California, Texas, and Florida, where demand is more robust. Rhode Island and Wyoming's shorter boxes reduce gas demand and fluctuation. Gas demand is highest in California and Texas by box plot height. The Federal Offshore–Gulf of Mexico and Alaska have substantial but fluctuating demand. Demand varies slightly in Rhode Island, Delaware, and Vermont. Gas demand data from several states displays diamond-shaped outliers outside box plots. This shows abnormally high or low demand for that condition. Demand uncertainty widens IQR in California and Texas. The narrow IQR of small-box states shows constant gas use. The graph shows that Texas and California have the most gas demand and volatility. Population, industrial output, and location may be included. Lower population and industrial activity states, like the Northeast, have more consistent gas use.

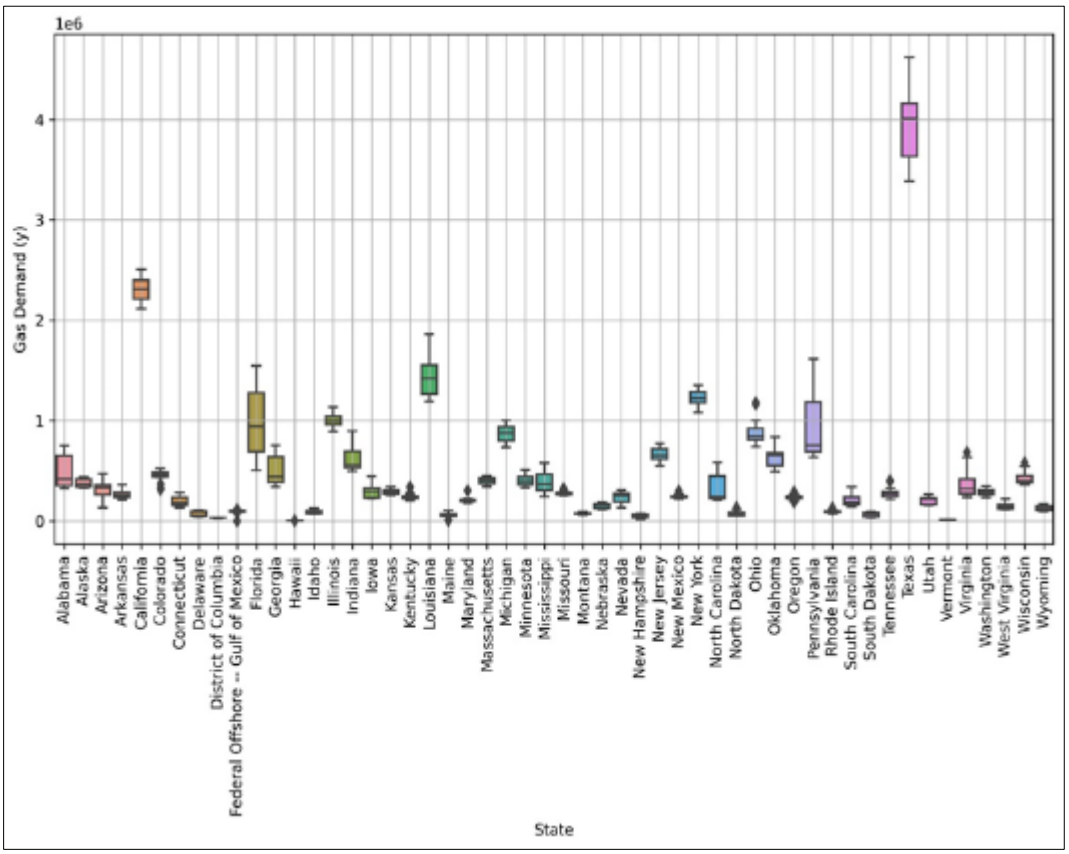

Figure 4 Distribution of Gas Demand Vs State

This scatter plot in Figure 5 shows gas demand along the y-axis and years on the x-axis from the 1990s to 2020. On the right side of the graph, you can see a scale from purple (low demand) to yellow (high demand), representing the degree of gas demand. The colour of each dot reflects this. A yearly pattern follows, and cluster spots are a part of it. According to these numbers, gas demand goes through yearly peaks and dips. The highest points of gas demand are increasing steadily. The yellowest peaks are becoming more frequent, indicating a rising peak demand tendency. The information seems to have been sorted into three primary demand tiers. The smallest group consists of about 1 million units, the medium group of about 2 million, and the biggest group of about 4 million. This trend in gas consumption, represented by this plot with seasonal changes, could be driven by long-term factors such as population growth, industrialisation, or increasing energy demands.

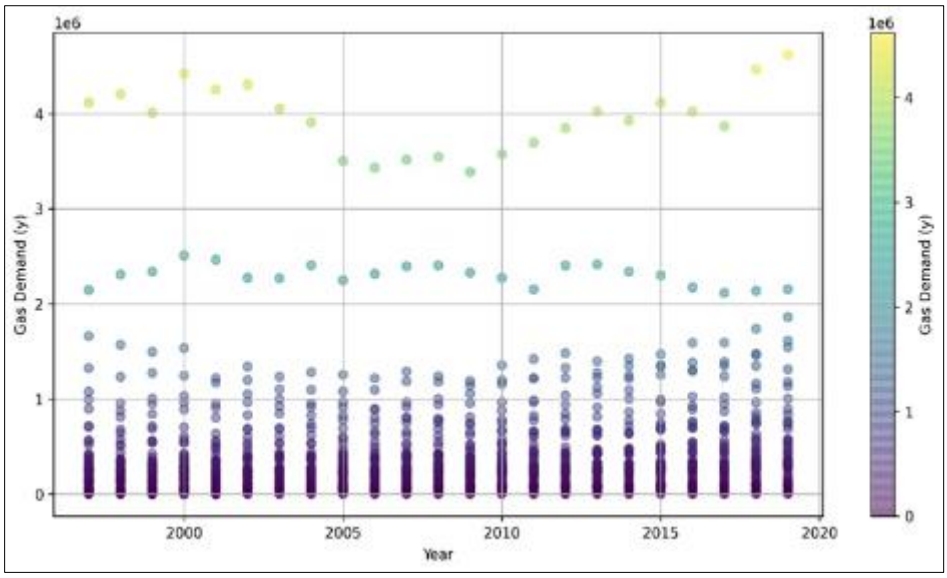

Figure 5 Scatter plot of Gas Demand Vs Year

3.4. Model selection and algorithms

Models and algorithms with excellent accuracy and the ability to handle complicated, nonlinear gas demand data patterns were chosen for this project. Machine learning methods tailored for regression were used. Random forest regression performed best with 99.67% R-squared and 34.53 RMSE. Random forests are helpful for datasets with variable and interacting aspects because they merge numerous decision trees to reduce overfitting and increase forecast accuracy. We also examined gradient-boosting methods like XGBoost to see if they could sequentially fix prediction mistakes and achieve competitive accuracy. Gradient boosting captures gas consumption trends in medium—to large datasets [22]. MLP regression neural networks were tested for complex data connection prediction. MLP can handle very nonlinear patterns; however, hyperparameters must be adjusted. Final tests included decision trees and linear regression to compare baselines. Predictive analytics requires model selection, and each algorithm discusses its pros and cons.

3.5. Evaluation

A critical aspect of this research is the assessment of the models employed to predict gas consumption by machine learning techniques. It involves assessing the model's precision, accuracy, and generalizability to new data. We evaluated their performance using various measures to identify the optimal model for this task. The primary metrics are the Root Mean Squared Error (RMSE) and the R-squared (R^2) value. A superior model's capacity to elucidate the variability of the target variable is signified by an elevated R^2 score. A lower root-mean-squared error (RMSE) signifies superior accuracy in comparing planned and actual values.

4. Results

A variety of test data percentages (ranging from 10% to 30%) are employed to evaluate the five methods listed in Table 1: Decision Tree, Random Forest, Gradient Boosting, Linear Regressor, and MLP Regressor (Multi-Layer Perceptron). Scores around 99.67% are consistently attained by Random Forest and Gradient Boosting, demonstrating robustness and consistency across all data partitions [23]. Both ensemble methods are appropriate for this dataset as they attain the maximum accuracy relative to the alternatives. When evaluated with smaller data partitions (10–25%), the Linear Regressor performs poorly, yielding scores below 1.0. Although it remains significantly inferior to tree-based methods, its performance markedly improves at the 30% test data level, reaching 61.10. The MLP Regressor's scores range from 50.10 to 57.59 over many test data splits, reflecting inconsistent and moderate performance [24]. This inconsistency raises apprehensions regarding the MLP Regressor's efficacy on this dataset.

Ultimately, although Random Forest and Gradient Boosting models marginally surpass the Decision Tree model, the former continuously attains high accuracy, with scores nearing 99.67% across all data percentages. The table indicates that Random Forest and Gradient Boosting are the most effective algorithms for this dataset, yielding consistently favourable outcomes.

Table 2 Accuracy Table

Test Data	Algorithm				
	Random Forest	Gradient Boosting	Linear Regressor	MLP Regressor	Decision Tree
10%	99.67	99.67	00.88	50.10	99.40
15%	99.52	99.14	00.41	57.59	99.26
20%	99.42	99.34	00.50	51.13	99.10
25%	99.41	99.43	00.21	54.22	99.10
30%	99.44	99.50	61.10	50.96	99.33

Table 3 shows the mean absolute, Mean Squared, and Root Mean Squared Error rates. All of the algorithms performed well. The MLP regressor has no far error escape. Our two algorithms, RF and XGB, gained the best results as an accuracy table. So, we applied this error calculation to find out the least root mean squared error. Here, XGB performed 43.90 Root Mean Squared Error rates, while Random Forest gained less than three error metrics with 34.53 root mean squared error.

Table 3 Error Calculate

Algorithm	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Random Forest	22.04	11.92	34.53
Gradient Boosting	26.50	19.27	43.90
Linear Regressor	38.65	35.93	59.94
MLP Regressor	42.97	54.45	73.79
Decision Tree	27.44	21.55	46.42

4.1. Analysis

The graph comparing the two numbers shows how effectively the model anticipates gas demand. The model captures trends well, as seen by the tight alignment of the two lines and its high R-squared value of 99.67% and low RMSE of 34.53 of the Random forest algorithms. This precision helps suppliers improve output, estimate demand, and avoid shortages during high usage periods, making it vital for energy management [25]. The model's ability to predict massive demand peaks at data points 1, 18, and 40 shows its capacity to handle unexpected demand surges. If the actual and predicted values differ little, the prediction error is modest and only minor adjustments may be needed to improve results. This model could improve gas resource planning and help suppliers manage demand changes. Future predictions could be improved by adding real-time data or using hybrid modelling.

5. Discussion

Figure 6 presents a line plot that juxtaposes the projected gas demand values (shown by dashed red lines and blue circles) with the actual gas demand values (represented by solid blue lines) across a sequence of data points. The actual and projected values on the figure are nearly indistinguishable, with the lines exhibiting almost identical trajectories. This signifies that the model effectively captures the peaks and troughs in gas demand. Significant peaks at data points 1, 18, and 40 show increased gas consumption, and the model's predictions align well with these surges. Our study has a high degree of accuracy, evidenced by an R-squared score of 99.67% and a low RMSE of 34.53; yet, minor discrepancies between the lines suggest potential areas for significant forecast deviation. This consistency underscores the model's capacity to reliably anticipate gas consumption and its prospective application in energy management and planning.

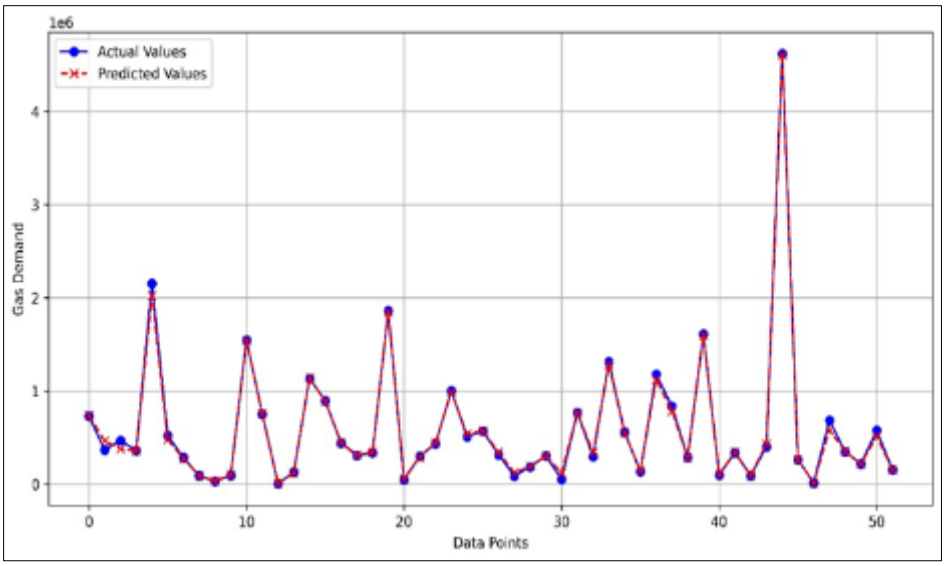

Figure 6 Evaluation Graph

6. Conclusion

The United States depends significantly on natural gas for diverse energy requirements, encompassing electricity generation, industrial activities, and residential heating. Alterations in energy policy, meteorological conditions, economic activity, and population growth are among the factors influencing demand. Gas demand forecasting is essential for sustainable energy planning, resource management, and environmental considerations. Utility providers can enhance operational efficiency through precise demand projections, while informed regulators may achieve a more effective equilibrium between supply and demand. Traditional models struggle to capture complex, nonlinear connections in gas demand, necessitating advanced machine-learning techniques. Gas consumption in eleven US states was forecasted utilising data from 2000 to 2019 with various machine learning methods, including random forest, gradient boosting, MLP regression, decision trees, and linear regression. The random forest model exhibited outstanding performance, evidenced by an R-squared value of 99.67% and a low RMSE of 34.53. The findings of this study offer policymakers and energy suppliers a robust basis for data-driven decision-making by illustrating how machine learning may improve the accuracy and reliability of gas demand forecasts. The subsequent phase in enhancing this project's successful foundation involves the development of a web-based platform and an AI-driven mobile application to deliver real-time gas demand forecasts. To assist consumers in making informed selections, these technologies will integrate real-time data streams, encompassing weather forecasts and market fluctuations. Enhancing model interpretability and including additional variables require further exploration to augment forecast precision and utility. This project represents a significant advancement in energy management efficiency through technological implementation, ensuring sustainable solutions to our energy challenges.

References

[1] https://www.eia.gov/todayinenergy/detail.php?id=43035

[2] https://safety4sea.com/eia-us-natural-gas-consumption-achieves-record-in-2019/

[3] Booth, A., Sutton, A., & Papaioannou, D. (2016). Systematic Approaches to a Successful Literature Review. Sage.

[4] Ridley, D. (2012). The Literature Review: A Step-by-Step Guide for Students. Sage.

[5] Beyca, O. F., Ervural, B. C., Tatoglu, E., Ozuyar, P. G., & Zaim, S. (2019). Using machine learning tools for forecasting natural gas consumption in the province of Istanbul. Energy Economics, 80, 937-949.

[6] Perrotta, F., Parry, T., & Neves, L. C. (2017, December). Application of machine learning for fuel consumption modelling of trucks. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 3810-3815). IEEE.

[7] Bedi, J., & Toshniwal, D. (2019). Deep learning framework to forecast electricity demand. Applied energy, 238, 1312-1326.

[8] Seyedzadeh, S., Rahimian, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. Visualisation in Engineering, 6, 1-20.

[9] Laib, O., Khadir, M. T., & Mihaylova, L. (2019). Toward efficient energy systems based on natural gas consumption prediction with LSTM Recurrent Neural Networks. Energy, 177, 530-542.

[10] Čeperić, E., Žiković, S., & Čeperić, V. (2017). Short-term forecasting of natural gas prices using machine learning and feature selection algorithms. Energy, 140, 893-900.

[11] Potočnik, P., Šilc, J., & Papa, G. (2019). A comparison of models for forecasting the residential natural gas demand of an urban area. Energy, 167, 511-522.

[12] Karadede, Y., Ozdemir, G., & Aydemir, E. (2017). Breeder hybrid algorithm approach for natural gas demand forecasting model. Energy, 141, 1269-1284.

[13] Su, H., Zio, E., Zhang, J., Xu, M., Li, X., & Zhang, Z. (2019). A hybrid hourly natural gas demand forecasting method based on the integration of wavelet transform and enhanced Deep-RNN model. Energy, 178, 585-597.

[14] Freeman, E. A., Moisen, G. G., Coulston, J. W., & Wilson, B. T. (2016). Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. Canadian Journal of Forest Research, 46(3), 323-339.

[15] Zhang, N., Li, Z., Zou, X., & Quiring, S. M. (2019). Comparison of three short-term load forecast models in Southern California. Energy, 189, 116358.

[16] Goliatt, L., Capriles, P. V., & Goulart Tavares, G. (2019, August). Gradient boosting ensembles for predicting heating and cooling loads in building design. In EPIA Conference on Artificial Intelligence (pp. 495-506). Cham: Springer International Publishing.

[17] Kaggle. (n.d.). Energy Consumption Dataset. Retrieved from https://www.kaggle.com/.

[18] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.

[19] Dev, V. A., & Eden, M. R. (2019). Gradient boosted decision trees for lithology classification. In Computer aided chemical engineering (Vol. 47, pp. 113-118). Elsevier.

[20] Qi, C., & Tang, X. (2018). Slope stability prediction using integrated metaheuristic and machine learning approaches: A comparative study. Computers & Industrial Engineering, 118, 112-122.

[21] Zahid, M., Ahmed, F., Javaid, N., Abbasi, R. A., Zainab Kazmi, H. S., Javaid, A., ... & Ilahi, M. (2019). Electricity price and load forecasting using enhanced convolutional neural network and enhanced support vector regression in smart grids. Electronics, 8(2), 122.

[22] Le, L. T., Nguyen, H., Zhou, J., Dou, J., & Moayedi, H. (2019). Estimating the heating load of buildings for smart city planning using a novel artificial intelligence technique PSO-XGBoost. Applied Sciences, 9(13), 2714.

[23] Ali, U., Shamsi, M. H., Nabeel, M., Hoare, C., Alshehri, F., Mangina, E., & ODonnell, J. (2019, November). Comparative analysis of prediction algorithms for building energy usage prediction at an urban scale. In Journal of Physics: Conference Series (Vol. 1343, No. 1, p. 012001). IOP Publishing.

[24] Ma, F., & Yan, X. (2019, October). Research the energy consumption estimation method of pure electric vehicle based on XGBoost. In 2019 3rd international conference on electronic information technology and computer engineering (EITCE) (pp. 1021-1026). IEEE.

[25] Mao, M., Zhang, S., Chang, L., & Hatziargyriou, N. D. (2019). Schedulable capacity forecasting for electric vehicles based on big data analysis. Journal of Modern Power Systems and Clean Energy, 7(6), 1651-1662.