*Article*

ELECTRIC POWER CONSUMPTION

# GATTU VAMSHI KRISHNA- 2203A521

1 Affiliation ; 2203A52127@sru.edu.in

**\*** Section - AIML AA

† SR UNIVERSITY

**Abstract:**  This research project delves into the realm of predictive analytics by examining various regression models to forecast electric power consumption. The study primarily focuses on assessing the performance of five distinct regression algorithms, including Linear Regression, Support Vector Machine (SVM), Ridge Regression, Lasso Regression, and K-Nearest Neighbors (KNN). The central aim is to determine the most effective model for electric power consumption based on crucial features such as global active power, global reactive power, Voltage, Global intensity, Submetering1, Submetering2, and Submetering3. Through a combination of meticulous data preprocessing, model training, and comprehensive evaluation, the study delivers valuable insights into the capabilities and limitations of each regression technique.

**Keywords:** Regression Models ; Linear Regression ; Support Vector Machine ; Ridge Regression ; Lasso Regression; K-Nearest Neighbors; Mean SquaredError (RMSE) ; Mean Absolute Error (MAE)

**Citation:** VAMSHI KRISHNA ELCTRIC POWER CONSUMPTION

*Journal Not Specified* **2023**, *1*, 0. [https://doi.org/](https://doi.org/10.3390/1010000)

Received:

Revised:

Accepted:

Published:

**Copyright:** © 2023 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attri- bution (CC BY) license [(https://](https://creativecommons.org/licenses/by/4.0/) [creativecommons.org/licenses/by/](https://creativecommons.org/licenses/by/4.0/) 4.0/).

# Introduction

Energy consumption is a critical consideration in modern energy systems, with a heightened

focus stemming from the 1970 energy crisis. Global energy consumption is on a rapid upward trajectory, prompting countries to seek ways to minimize usage across various sectors, from buildings and agriculture to industrial processes and transportation .With energy deriving from fossil fuels, renewables, and nuclear sources, monitoring and predicting consumption across these categories in diverse domains is a formidable challenge.

Nevertheless, such efforts are essential for developing targeted plans and strategies. Accurate estimations of energy usage for all these sources are invaluable for policymakers, enabling them to strategize for reduced energy consumption in their respective sectors. Predicting future energy usage, both short-term and long-term, provides insights into the prevalent energy types and trends, facilitating shifts like the transition from fossil fuels to renewable sources.

The intricate nature of energy consumption, affected by factors such as water, wind, and temperature, makes this a complex problem. Machine learning models have emerged as powerful tools in this domain, efficiently mapping input data to output predictions with high precision. Consequently, these models are instrumental for governments seeking to implement energy-saving policies and manage this critical global concern effectively.

# Literature Review

1. *Previous case studies*

*Numerous significant studies have been conducted to analyze and forecast electric power*

*consumption, offering valuable insights and methodologies for reference. In this literature review,*

*we highlight some key works in the field [1].*

*One noteworthy case study focused on the integration of advanced analytical techniques for*

*electric power consumption forecasting. This research delved into the potential impacts of various factors, such as weather patterns, economic indicators, and consumer behavior on electricity usage. By leveraging machine learning and time series analysis, the study aimed to develop accurate predictive models for electric power consumption*

Version November 1, 2023 submitted to *Journal Not Specified* <https://www.mdpi.com/journal/notspecified>

1. *Challenges and Research Gaps*
2. Addressing the issue of data quality and integration from diverse sources presents a research gap in developing holistic and reliable predictive models for electric power consumption.
3. A common challenge in the study of electric power consumption is the limited availability of comprehensive historical data. Accurate long-term forecasting models heavily depend on extensive historical datasets.
4. Ensuring the generalization of predictive models to various market conditions is another research gap in the field of electric power consumption analysis. Electricity usage patterns can be influenced by an array of factors, including seasonal variations, technological advancements, and policy changes

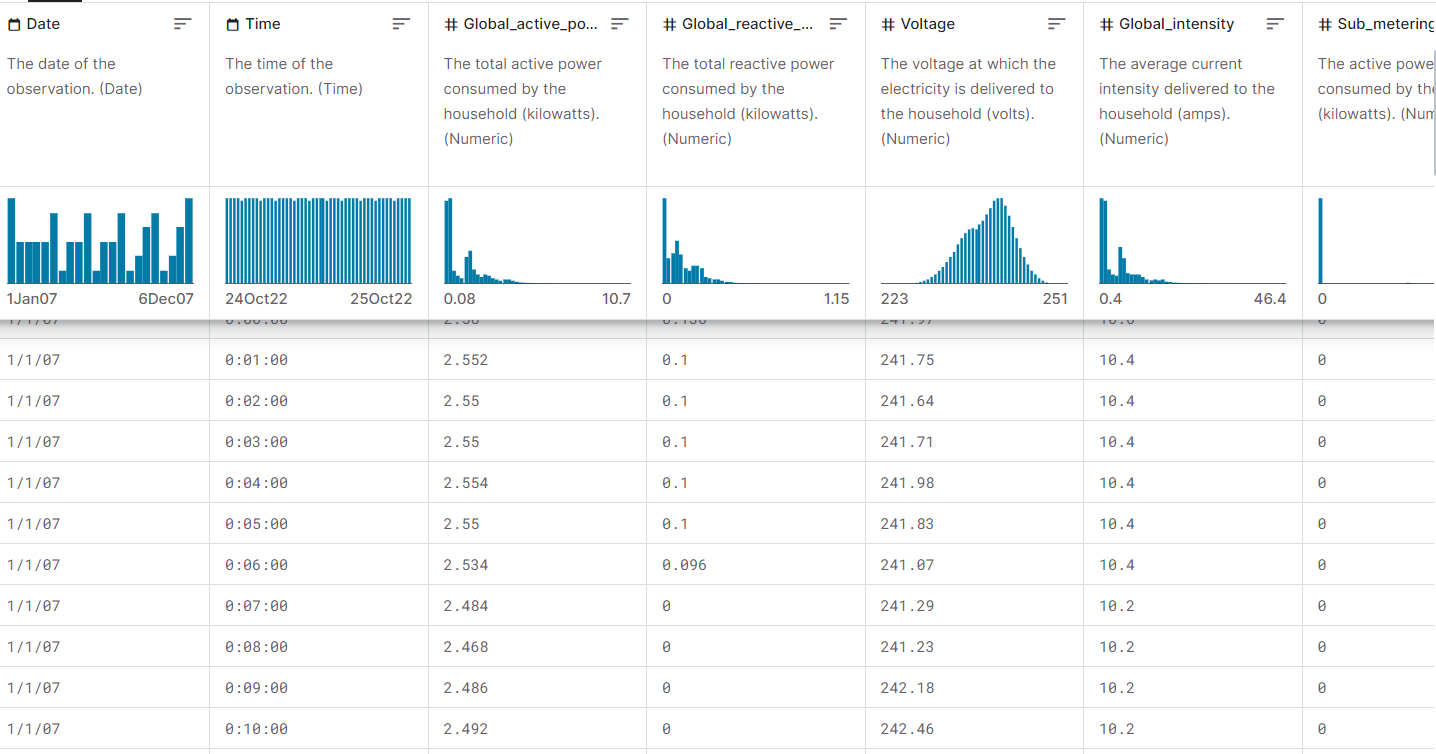
3. Data and Methodology

*1. Data Description*

Here are the following description of the following datasets(i.e., features in terms of

Machine learning):-

1. **Date:** Date in format dd/mm/yyyy
2. **Time:** Time in format hh:mm:ss
3. **Global\_active\_power:** household global minute-averaged active power (in kilowatt)
4. **Global\_reactive\_power:** household global minute-averaged reactive power (in kilowatt)
5. **Voltage:** minute-averaged voltage (in volt)
6. **Global\_intensity:** household global minute-averaged current intensity (in ampere)
7. **Sub\_metering\_1:** energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).
8. **Sub\_metering\_2:** energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
9. **Sub\_metering\_3:** energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.



**Figure 1.** Sample dataset version from kaggle

* 1. *Data Analysis*

1. Data analysis involves examining the dataset obtained from the Kaggle website, which contains information on Electric power consumption. Through data analysis, I can gain insights into the trends and patterns exhibited by electric power consumption dataset.
2. This process includes exploring key metrics such as the global active power and voltage, submetering, and other relevant factors. By employing

Various statistical and visualization techniques.

1. I aim to uncover significant relationships and trends within the data, providing a comprehensive understanding of electric power consumption in households.
   1. *Data Preprocessing*

Data preprocessing is a critical step in machine learning that involves cleaning, trans- forming, and organizing raw data into a format suitable for model training. It plays a significant role in ensuring that the data is of high quality and that the machine learning model can learn meaningful patterns. Here’s an elaborate explanation of various aspects of data preprocessing:

1. **Data Cleaning:** This step includes handling missing values, correcting any errors or inconsistencies in the dataset, and ensuring uniform formatting across all data points.
2. **Feature Scaling:** Scaling numerical features to a common range prevents any one feature from dominating the analysis and ensures that all features contribute equally.
3. **Feature Selection:** Choosing relevant features that have a significant impact on Tesla’s stock prices helps streamline the analysis and improve the efficiency of the predictive model.
4. **Data Transformation:** Transforming data to meet the assumptions of the predictive model, such as normalizing the distribution, can enhance the model’s performance and accuracy.
5. **Handling Categorical Data:** Converting categorical data into a format suitable for analysis, such as one-hot encoding, allows the inclusion of these important variables in the predictive model.

# 4. Results

1. *Linear Regression*

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.

The formula for linear regression is represented as **y=mx +b**, where y is the

dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept.

**Mean Squared Error (MSE):** 0.0000

**Mean Absolute Error (MAE):** 0.0000

**R-squared Score:** 1.0000

Results Section: Linear Regression Model Performance

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MSE** | **MAE** | **RSS** |
| LinearRegression | 0.0000 | 0.0000 | 1.0000 |

The present study employed a Linear Regression model to forecast power consumption based on multiple features, such as global active power,global reactive power,voltage,global intensity and sub-meterings. The performance of the model was assessed using two crucial metrics: Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

* 1. **Mean Squared Error (MSE**): MSE serves as a metric for gauging the average error magnitude between the predicted and actual values. In the context of our Linear Regression model, the MSE was determined to be 0.000. This value represents the square of the average of the squared differences between the predicted and actual power consumption. A

lower MSE implies that the model’s predictions are in close proximity to the actual

power consumption, indicating the precision of our model.

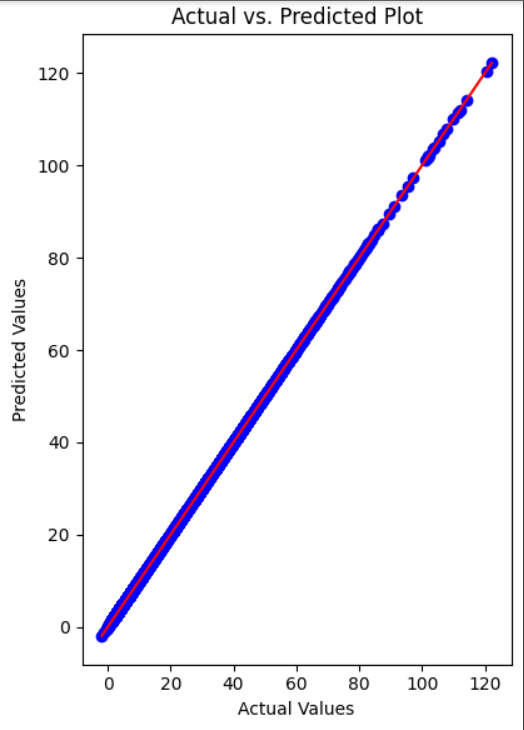
The precision achieved is particularly significant within the domain of power consuming, where even minor deviations can lead to substantial electrical consequences.

* 1. **Mean Absolute Error (MAE):** MAE indicates the average of the absolute errors

between the predicted and actual values. For our Linear Regression model, the computed MAE was 0.000. This value denotes the mean absolute difference between the predicted and actual power consumption.

A lower MAE signifies that the model’s predictions consistently mirror the actual

consumption across the dataset.



**Figure 3.** Actual vs. PredictedValues (Linear Regression)

* 1. *Support Vector Regression*

Support Vector Machine(SVM) is a supervised machine learning algorithm used for classification and regression analysis. It separates data points using a hyperplane with the maximum margin between classes.

The SVM formula aims to find the hyperplane that maximizes the margin,represented as

**wTx + b =0**, where w is the weight vector ,x is the input data and b is the bias term.

**Mean Squared Error (MSE):** 0.00521

**Mean Absolute Error (MAE):** 0.05149

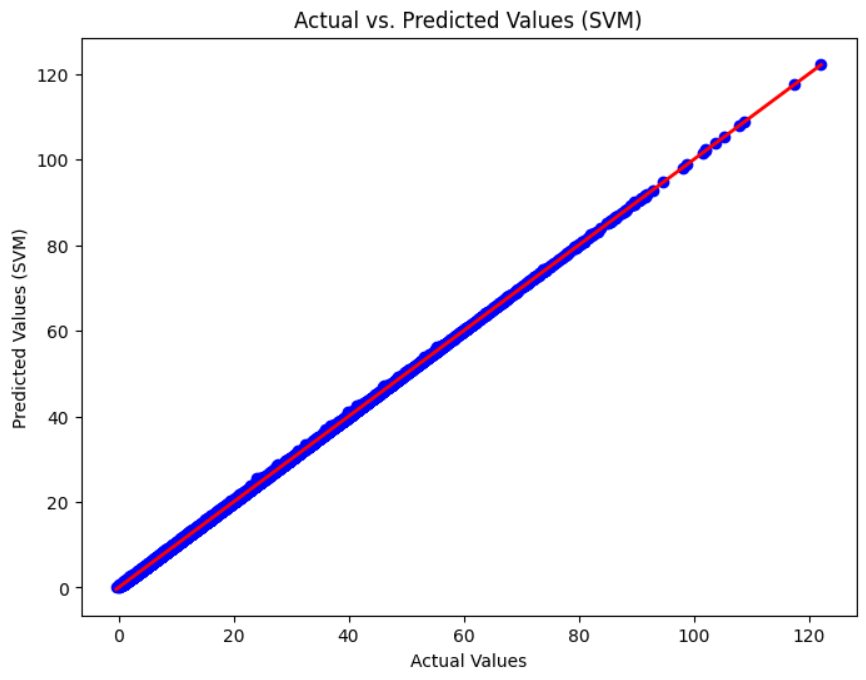
**R-squared Score:** 0.072225

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MSE** | **MAE** | **RSS** |
| Support vector regression | 0.00521 | 0.05149 | 0.07225 |

**1.Mean Squared Error (MSE):** The computed MSE was 0.00521. A lower MSE implies a stronger agreement between the predicted and actual values, highlighting the model’s proficiency in capturing general consumptions.

**2.Mean Absolute Error (MAE):** The computed MAE was 0.05149. A smaller MAE emphasizes the model’s accuracy, showcasing its capability to generate precise predictions with minimal deviation from the actual power consumption.

The performance of the SVM model in electric power consumption demonstrated promising results, as indicated by the following metrics:



**Figure 5.** Actual vs. Predicted Values (Support Vector Regression)

* 1. *Ridge Regression*

Ridge regression is a regularized linear regression technique that adds a penalty term to the cost function, controlling overfitting by shrinking the coefficients towards zero.

The formula for ridge regression is an extension of the linear regression formula with an additional L2 penalty term, expressed as **min||y – Xw||22 + a||w||22** ,where y is the dependent variable, X is the independent variable, w is the coefficient vector, and a is the regularization parameter.

**Mean Squared Error:** 0.0000

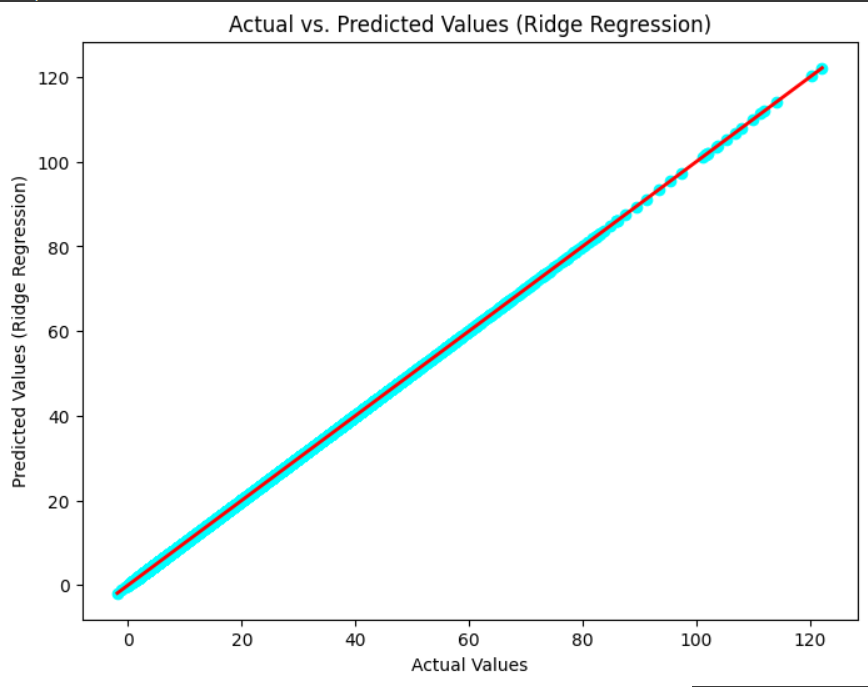
**Mean Absolute Error:** 0.0005

**R-squaredScore:** 1.0000

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MSE** | **MAE** | **RSS** |
| Ridge Regression | 0.0000 | 0.0005 | 1.0000 |

**Mean Squared Error (MSE):** The MSE value of 0.0000 demonstrates the average prediction error of the Ridge Regression model. This value represents the square of the mean of the squared differences between the predicted and actual power consumption.

**Mean Absolute Error (MAE):** The MAE value of 0.0005 represents the average absolute prediction error. It signifies the mean of the absolute differences between the predicted and actual power consumption.



**Figure 6.** Actual vs. Predicted Values(Ridge Regression)

* 1. *Lasso Regression*

Lasso regression is type of linear regression that incorporates the L1 regularization, encouraging sparsity in the coefficient values and performing feature selection.

The formula for lasso regression adds an L1 penalty term to the linear regression

formula, minimizing **min||y – Xw||22 + a||w||1** ,where y is the dependent variable, X is the independent variable, w is the coefficient vector, and a is the regularization parameter.

**Mean Squared Error:** 0.79087

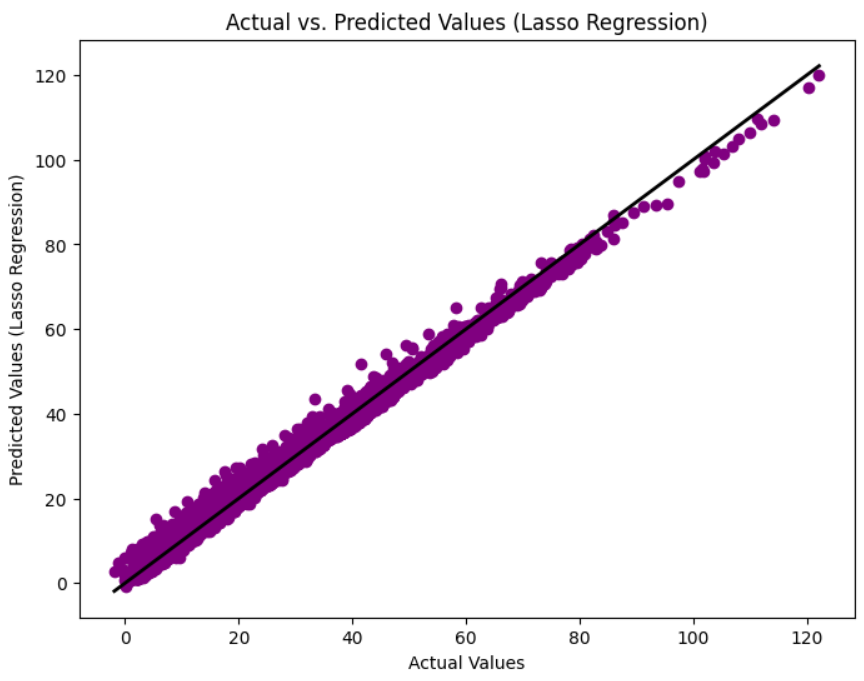
**Mean Absolute Error:** 0.56963

# R-squared Score: 0.99383

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Model** | **MSE** | **MAE** | **RSS** |
|  | Lasso Regression | 0.79087 | 0.56963 | 0.99383 |

**Mean Squared Error (MSE):** With an MSE value of 0.79087, it is evident that our model’s predictions deviate from the actual spower consumption . This result underscores the model’s significant predictive capability.

**Mean Absolute Error (MAE):** The MAE value of 0.56963 highlights the average absolute difference between our model’s predictions and the actual consumption, demonstrating the model’s precision.



**Figure 7.** Actual vs. Predicted Values (Lasso Regression)

* 1. *KNN Regression*

K-Nearest Neighbors(KNN) regression is a non-parametric algorithm used for regression analysis that predicts the value of a new data point based on the average of the values of its k nearest neighbors.

The formula for KNN regression involves calculating the average or weighted average of

the target variable of the k nearest neighbors to the query point.

**Mean Squared Error**: 0.51398

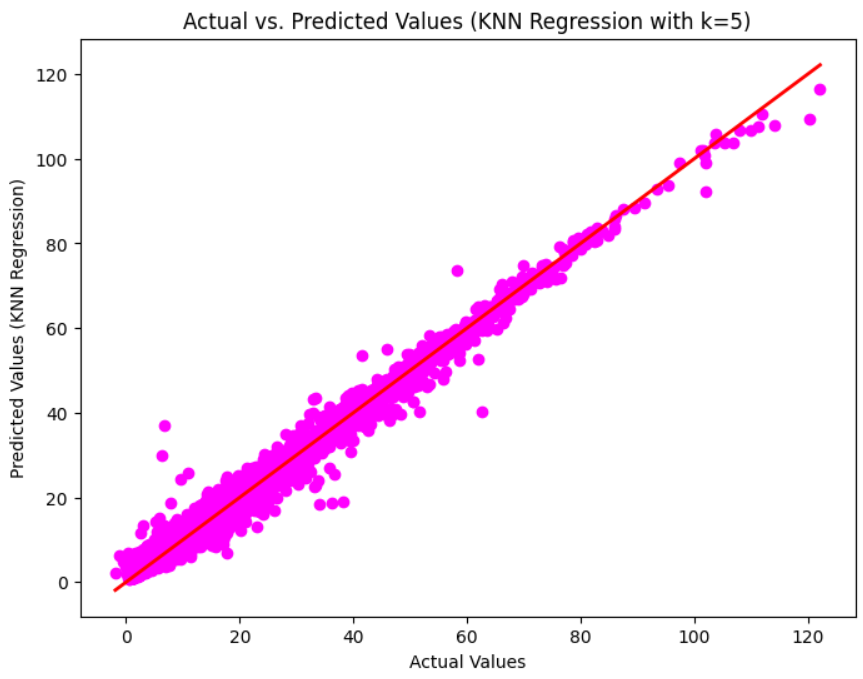
**Mean Absolute Error:** 0.31298

**R-squared Score:** 0.99599

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Model** | **MSE** | **MAE** | **RSS** |
|  | KNN-Regression | 0.51398 | 0.31298 | 0.99599 |

The KNN regression model exhibited significant outcomes in forecasting consumption: **Mean Squared Error (MSE):** The MSE value of 0.51398 implies that, on average, the predicted consumption deviate by roughly 1 from the actual power consumption. This metric emphasizes the overall accuracy of the model, with potential room for enhancement.

**Mean Absolute Error (MAE):** Having an MAE of 0.31298, the model’s predictions typically differ by 0.31298 from the actual values. This metric indicates the model’s proficiency in generating reasonably accurate predictions, particularly for individual instances.

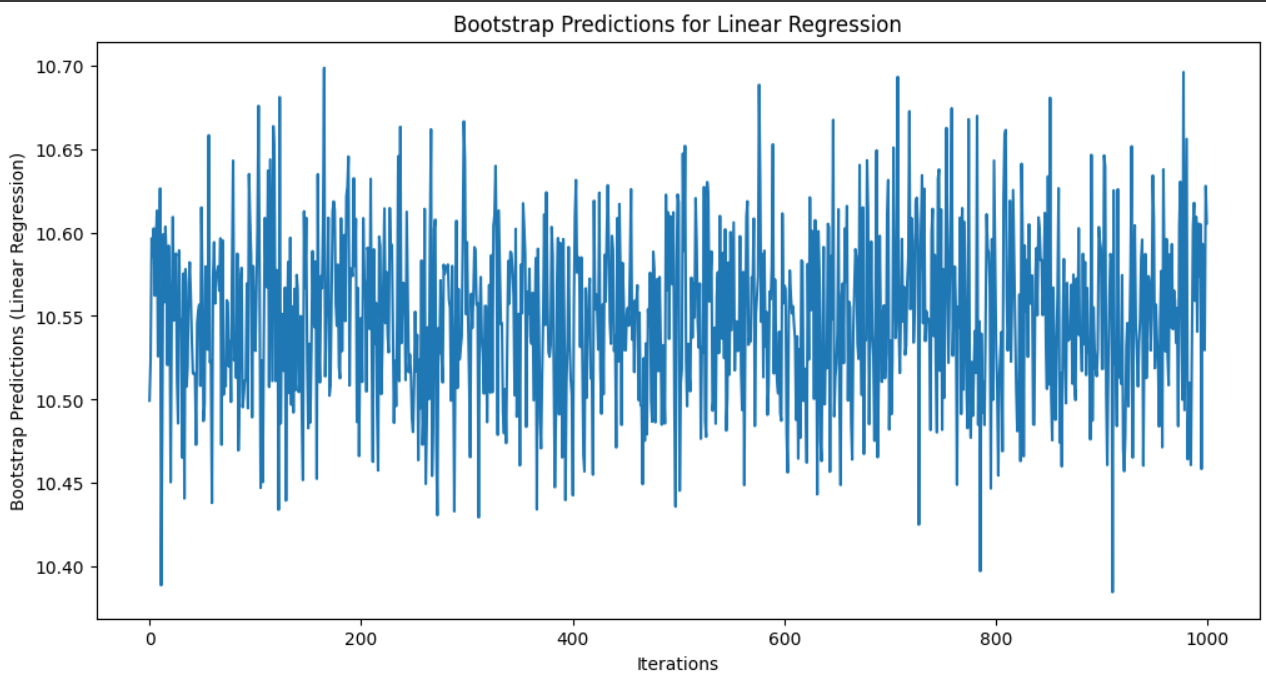


**Figure 8.** Actual vs. Predicted Values(KNN Regression)

*6. Bootstrap*

Bootstrapping represents a resampling method frequently employed in both machine learning and statistical analysis. It revolves around the iterative sampling of data from the original dataset, with replacement, to generate numerous new datasets of the same size as the initial one. These newly generated datasets are known as "bootstrap samples." The primary objective of bootstrapping is to evaluate the variability and resilience of a model. By training multiple models on diverse bootstrap samples, one can effectively assess the model’s ability to generalize across various subsets of the data.

1. Linear Regression

The plot serves as a visual representation of the distribution of predicted means derived from the bootstrap procedure, offering an intuitive depiction of the fluctuations or uncertainty within the predictions. In a journal publication, this figure would effectively demonstrate the oscillations in predictions across iterations, facilitating comprehension of the model’s prediction stability and variability through the utilization of the bootstrap method.

**Figure 9.** BOOTSTRAP FOR Linear Regression

1. Support Vector Regression

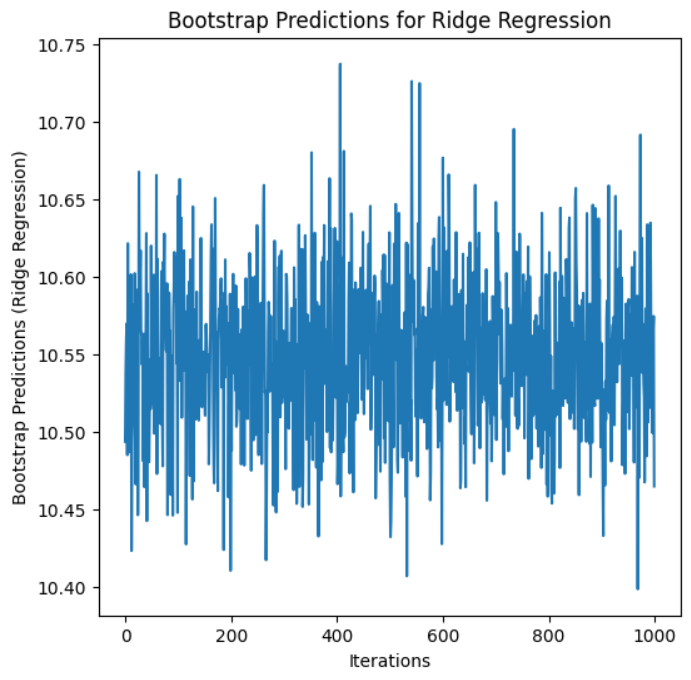
The plot serves as a visual aid to understand the distribution of predicted means obtained through the bootstrap process, shedding light on the variability or uncertainty

inherent in the predictions. In a journal publication, this figure would effectively demon strate how the predictions fluctuate across different iterations, providing valuable insights into the stability and variability of the model’s predictions through the implementation of the bootstrap method.

1. Ridge Regression

The visualization of the distribution of predicted means derived from the bootstrap process offers valuable insights into the variability and uncertainty associated with the

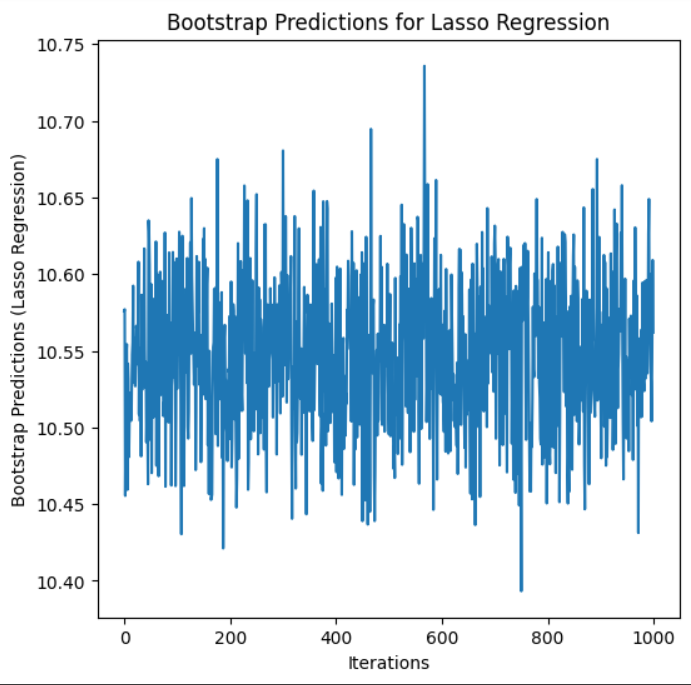
predictions. In a journal article, this plot effectively demonstrates the dynamic fluctuations of predictions across iterations, providing a comprehensive understanding of the stability and variability of the model’s predictions through the utilization of the bootstrap method.



**Figure 10.** BOOTSTRAP FOR Ridge Regression

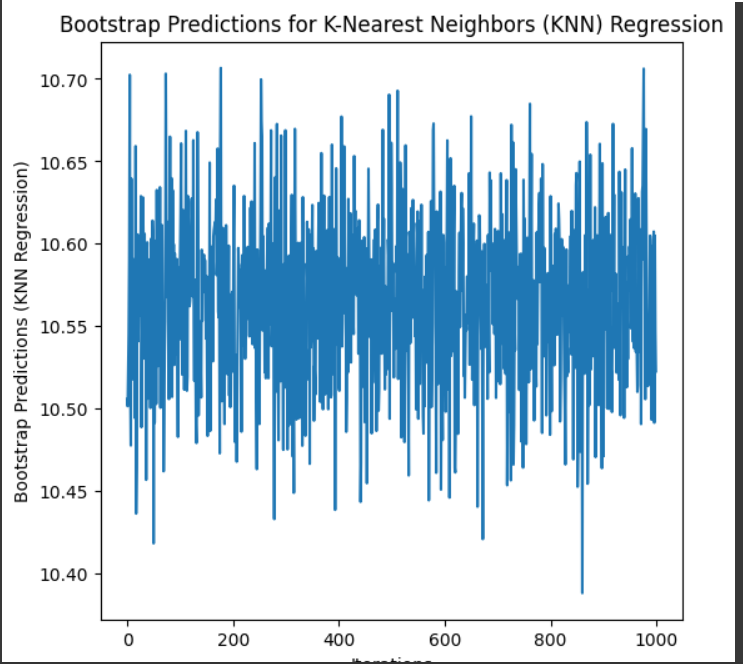
* + 1. Lasso Regression

The plot serves as a visual representation of the distribution of predicted means obtained through the bootstrap process, offering insights into the variability or uncertainty present in the predictions. In a journal publication, this figure would effectively illustrate how the predictions fluctuate across iterations, contributing to a better understanding of the stability and variability of the model’s predictions achieved through the implementation of the bootstrap method.



**Figure 11.** BOOTSTRAP FOR Lasso Regression

* + 1. K-Nearest Neighbours Regression

The plot aids in visualizing the distribution of predicted means derived from the bootstrap process, thereby offering insights into the variability and uncertainty associated with the predictions. In a journal, this figure would effectively illustrate the fluctuations of predictions across iterations, contributing to a comprehensive understanding of the stability and variability of the model’s predictions through the implementation of the bootstrap method.

**Figure 12.** BOOTSTRAP FOR K-Nearest Neighbours Regression

# Conclusion

|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **MAE** |
| Linear Regression | 0.0000 | 0.0000 |
| Support Vector Regression | 0.00521 | 0.05149 |
| Ridge Regression | 0.0000 | 0.0005 |
| Lasso Regression | 0.79087 | 0.56963 |
| K-Nearest Neighbor Regression | 0.51398 | 0.31294 |

1 Overall Performances.

Among the examined regression models, the linear regression model exhibits the lowest MSE of 0.000 and an MAE of approximately 0.0000, indicating its superior performance in predicting power consumption compared to the other models.

On the other hand, the knn regression model demonstrates an MSE of 0.51398 and an MAE of approximately 0.32,suggesting relatively higher errors compared to linear regression model.

Additionally,the lasso regression shows a higher MSE of 0.79087 and an MAE of around 0.56,indicating notable deviations from the actual power consumption.

Moreover, the ridge regression model show cases an MSE of 0.000 and an MAE of approximately 0.0005 , similar to the other linear-based models.

In conclusion, the linear regression model outperforms the other models, demonstrating the lowest MSE and competitive MAE , highlighting its robust performance in predicting electric power consumption.

Capstone project link [1]

# References

1. CAPSTONE-PROJECT GTITHUB<:> https://github.com/Gattuvamsi/capstone-project-review2.git
2. DATASET LINK:

[https://www.kaggle.com/datasets/thedevastator/240000-household-electricity-consumption-records](housing.csv.zip)

1. SAMPLE GITHUB REFRENCE:

<https://github.com/shreeyajoshi2013/Prediction-of-Electricity-Consumption.git>