Data-Driven Forecasts for Solar Energy Consumption

Shahil Patel - 200010039
Pranav Talegaonkar - 200010041

Table of Contents:

List of Tables:

List of Figures:

List of Abbreviations:

- ***Wh:*** watt-hour
- ***GHI:*** Global Horizontal Irradiance
- ***IQR:*** Interquartile range
- ***NREA:*** National Renewable Energy Agency of India
- ***CEA:*** Central Electricity Authority
- ***ALMM:*** Approved list of models and manufacturers
- ***SVR:*** Support Vector Regression

Abstract

The ever increasing demand for energy and the adverse environmental effects of fossil fuels necessitate a global shift towards renewable energy sources. This study uses a comprehensive dataset to analyze and predict consumption patterns of renewable energy (solar energy). The data encompasses factors like solar presence, time-based variations (month, hours, sunlight duration), radiant energy levels, weather conditions, and others that influence consumption.

The significance of this research lies in its contribution to optimizing renewable energy utilization. We can establish more efficient energy generation and distribution strategies by identifying the key features that drive consumption trends. Analyzing the impact of weather patterns on consumption allows for improved forecasting models, enabling us to prepare for fluctuations and maintain grid stability.

Furthermore, the insights gleaned from this data-driven approach can bolster public awareness of the dynamic nature of renewable energy sources. Understanding the interplay between environmental conditions and consumption empowers individuals and policymakers to make informed decisions regarding energy usage and infrastructure development.

Transitioning to renewable energy is paramount for a sustainable future. This study not only underscores the importance of renewables but also equips us with the analytical tools necessary to maximize their potential. By harnessing the power of data, we can pave the way for a future powered by clean, environmentally responsible energy sources.

Introduction

The escalating global energy demand poses a significant challenge. Our reliance on fossil fuels meets this growing need and contributes significantly to environmental degradation. Sources of renewable energy, including geothermal, wind, and sun, offer a promising

solution – a clean, sustainable path to meet our energy demands. However, unlike traditional sources, renewable energy generation is inherently variable, depending on environmental conditions. Optimizing the utilization of these resources necessitates a deep understanding of how these variables influence consumption patterns.

This research project aims to address this challenge by:

- **Understanding the key factors that influence renewable energy consumption patterns.** This involves analyzing the relationships between solar radiation, weather conditions, time-based variations, and energy consumption.
- **We are developing a robust model for predicting renewable energy consumption** based on the identified factors. This model will enable utilities to anticipate fluctuations in consumption and optimize energy production and distribution strategies.

Here are the specific research questions this project seeks to answer:

- What are the critical environmental and time-based features that impact renewable energy consumption?
- How strong are the correlations between these features and energy consumption?
- Can we develop a reliable prediction model for renewable energy consumption using various regression algorithms?
- Which regression algorithm performs best in predicting renewable energy consumption for this dataset?

This project will contribute valuable insights into renewable energy management by answering these questions. Utilities and policymakers can potentially use the developed prediction model to:

- Enhance grid stability and energy security.
- Improve renewable energy integration into the existing power grid.
- Inform planning and development decisions for renewable energy infrastructure.
- Encourage informed energy consumption practices among consumers.
- In the end, this research aims to support the development of clean, sustainable renewable energy sources for the future.

This project delves into this crucial area by analyzing and predicting renewable energy consumption trends using a comprehensive dataset (source: Kaggle). We explore the interrelationships between various features that impact consumption, including:

- ***Energy Delta (Wh)*:** This represents the change in energy consumption within a specific timeframe.
- ***Global Horizontal Irradiance (GHI)*:** Measures the intensity of solar radiation reaching a horizontal surface.

- ***Weather Variables*:** Temperature, pressure, humidity, wind speed, rainfall, snowfall, and cloud cover significantly affect renewable energy generation and consumption.
- ***Time-Based Features*:** Time, day of the hour, month, day length, and sunlight duration provide insights into how consumption patterns fluctuate throughout the day and year.
- ***Binary Feature*:** "isSunSunlight" indicates the presence or absence of sunlight.
- ***Weather Type*:** Categorical data representing the prevailing weather conditions.

Data Analysis Approach*:*

This study employs a multi-pronged data analysis approach to uncover the hidden patterns within the dataset. Here is an overview of the critical methods used:

- **Descriptive Statistics:** We calculate each Feature's *minimum, maximum, average, 25th percentile (Q1), and 75th percentile (Q3) values*. These statistics provide a fundamental understanding of the data's central tendency and spread.
- **Distribution Curves:** Visualizing the distribution of each Feature using histograms and kernel density estimates allows us to identify potential skewness or outliers. Understanding the data distribution is crucial for selecting appropriate prediction models.
- **Correlation Matrix:** This matrix reveals the direction and strength of each pair's linear link with each other features. Identifying features with strong correlations helps understand how changes in one variable might influence another.
- **Boxplots:** Boxplots provide a graphical representation of the distribution of each Feature, showcasing outliers and interquartile range (*IQR*). This helps us identify potential anomalies and assess the data's variability.

Real-World Context: The Case of India

Incorporating real-world data from India is essential for this project's relevance. A recent study by the ***National Renewable Energy Agency of India (NREA)*** *[Report on the Installed Capacity of Power Plants in India, Central Electricity Authority (CEA), 2023: Installed Capacity Report - Central Electricity Authority]* offers valuable insights into the country's renewable energy production and consumption landscape. This report highlights India's significant progress in renewable energy deployment and serves as a benchmark for understanding the specific challenges and opportunities within the Indian context.

The leading Indian states in Solar Energy in 2022:

Rank	State	Solar Capacity

1	Rajasthan	16.06 GW
2	Gujarat	8 GW
3	Karnataka	7.8 GW
4	Tamil Nadu	6.2 GW
5	Telangana	4.6 GW

Table 1

Source: *ornatesolar.com*

India installed 1,847 MW of wind power and 13,956 MW of solar power in 2022. The states of Rajasthan, Gujarat, and Tamil Nadu housed most of the solar capacity. The Ministry of New and Renewable Energy states that as of December 31st, the nation's installed renewable energy capacity amounted to 120.85 GW. About 52% of all renewable energy comes from solar power, with wind coming in at 35%, biopower at 9%, and small hydro at 4%.

The leading Indian states in Solar Energy in 2023:

Rank	State	Solar Capacity
1	Rajasthan	17.8 GW
2	Gujarat	10.13 GW
3	Karnataka	9.05 GW
4	Tamil Nadu	6.8 GW
5	Maharashtra	4.8 GW

Table 2

Source: *ornatesolar.com*

The nation's renewable energy sector had a successful fiscal year, which concluded on March 31, 2024. The solar power component, which accounted for 81% of the addition, helped the grid record its highest-ever yearly new capacity addition.

The new capacity installed in FY24 was 18,485 MW, exceeding the previous most significant yearly addition of 15,274 MW in FY23. The renewable energy sector added 15,950 MW annually on average during the last three years (FY24, FY23, and FY22).

Compared to the additions made in FY23, the new capacity addition in the solar power sector increased by almost 18% in FY24. According to data published by the Union Ministry of New & Renewable Energy, the solar market brought 15,033 MW of new capacity (covering all categories such as ground-mounted, rooftop, hybrid solar, and off-grid) in FY24 as opposed to 12,784 MW in FY23 and 12,761 MW in FY22.

About 11.5 GW were added to the utility-scale solar segment in FY2024, an increase of 18% over FY23 installations. About 2,992 MW were added to the rooftop solar market in FY2024, a 34% increase over FY23.

The government's exemption of the Approved List of Models and Manufacturers (ALMM) until March 2024 may be the reason for the spike in rooftop solar capacity in FY24. According to JMK Research & Analytics, another explanation might be the drop in module pricing in the second half of 2023 due to the drop in cell prices in China.
The top three states with the most large-scale solar installations in FY2024 were Tamil Nadu (1261 MW), Gujarat (3320 MW), and Rajasthan (3929 MW).

Through this comprehensive approach, this research project seeks to

- Uncover the key factors influencing renewable energy consumption patterns.
- Develop a robust prediction model for forecasting consumption based on various environmental and time-based features.
- Enhance grid stability and energy security by enabling utilities to anticipate fluctuations in renewable energy production.
- Empower policymakers to make informed decisions regarding renewable energy infrastructure development and capacity planning.
- Foster public awareness regarding the dynamic nature of renewable energy sources, promoting informed energy consumption practices.
- The findings of this research can be instrumental in expediting the switch to sustainable energy sources in the future. By unlocking the power of data and leveraging advanced prediction models, we can maximize the utilization of renewable energy sources, pave the way for a cleaner and more secure energy grid, and help make the earth healthy for the next generations.

Data & Methodology

Dataset Description

1. **Correlation Matrix**:

Figure 1

A *correlation matrix* is a table that displays the correlation between two variables. In this case, the variables are all features of a dataset related to solar power generation. The heatmap uses color to represent the strength of the correlation, with *yellow indicating a positive correlation*, *purple indicating a negative correlation*, and *green indicating no correlation*.

A number between -1 and 1 represents the strength of the correlation. A perfect positive correlation is shown by a correlation of 1, which means that as one variable's value increases, the other variable's value also increases. A correlation of -1 indicates a perfect negative correlation, meaning that when one variable's value rises, the other variable's value falls. There is no correlation between the two variables when the correlation value is 0.

The heatmap in the image can be used to identify relationships between the different features in the dataset. For example, the heatmap shows a strong positive correlation between solar irradiance (GHI) and energy delta (Wh). This means a positive relationship exists between sunlight intensity that strikes a solar panel and the amount of energy the panel produces.

The heatmap shows a robust negative correlation between humidity and energy delta (Wh). This means a negative relationship exists between the amount of moisture in the air and the

amount of energy a solar panel produces. This is likely because humidity can reduce the sunlight that reaches the solar panel.

In conclusion, a correlation matrix is valuable for exploring relationships between variables in a dataset. The heatmap visualization of a correlation matrix can help researchers identify data patterns and trends. The specific heatmap you sent shows positive correlations between solar irradiance, energy delta, day length, and energy delta. There is also a negative correlation between humidity and energy delta.

Researchers can use this information to improve solar power generation systems. For example, they could develop systems that are more resistant to the effects of humidity.

- One does not infer causation from correlation. Two variables are not always connected only because one variable causes the other to change.
- The correlation coefficient should be statistically significant to be meaningful.
- Variables that correlate with both variables you are interested in are considered confounding. Confounding variables can make determining the proper relationship between the two variables of interest difficult.

Data Distribution Analysis

This report examines the distribution graphs of various features within a dataset relevant to solar power generation. Visualizing the distribution of data points allows for an initial understanding of central tendencies, spread, and potential outliers within a dataset.
The provided graphs depict what appear to be histogram visualizations. Histograms are graphical representations that utilize bars to characterize the frequency distribution of data points across a range of intervals. The x-axis typically represents the measured variable, while the y-axis represents the frequency or density of data points within each interval.

Prediction Model Description

The next stage of this project focuses on prediction. By leveraging various regression algorithms, we aim to develop a robust model capable of forecasting renewable energy consumption based on the available features. Here is a brief introduction to the employed algorithms:

- **Linear Regression:** This widely used method establishes a linear relationship between the independent variables (features) and the dependent variable (energy consumption).
- **Gradient Boosting Algorithms:** These algorithms build an ensemble of weak learners (like decision trees) sequentially, with each learner focusing on improving the areas where the previous ones made errors. This leads to a more accurate model than any single learner.

- **Random Forest:** This ensemble method builds multiple decision trees using different random subsets of features and data points. Aggregating the predictions from these trees reduces the variance and improves the model's accuracy.
- **Decision Trees:** These algorithms create a structure like a tree in which each internal node stands for a feature, and each branch represents a decision based on the feature value. The final leaf nodes represent the predicted energy consumption for specific combinations of feature values.
- **Support Vector Regression (SVR):** This technique finds a hyperplane in high-dimensional space that best separates the data points while minimizing the margin of error. It performs well with smaller datasets and is robust to outliers.
- **LASSO:** This regression technique adds a penalty term to the linear regression model, encouraging sparsity by shrinking some coefficient estimates toward zero. This helps in feature selection and reduces model complexity.

By employing these diverse algorithms and comparing their performance, we aim to identify the most effective model for predicting renewable energy consumption in the context of our dataset.

Findings:

***Observations by Feature*:**

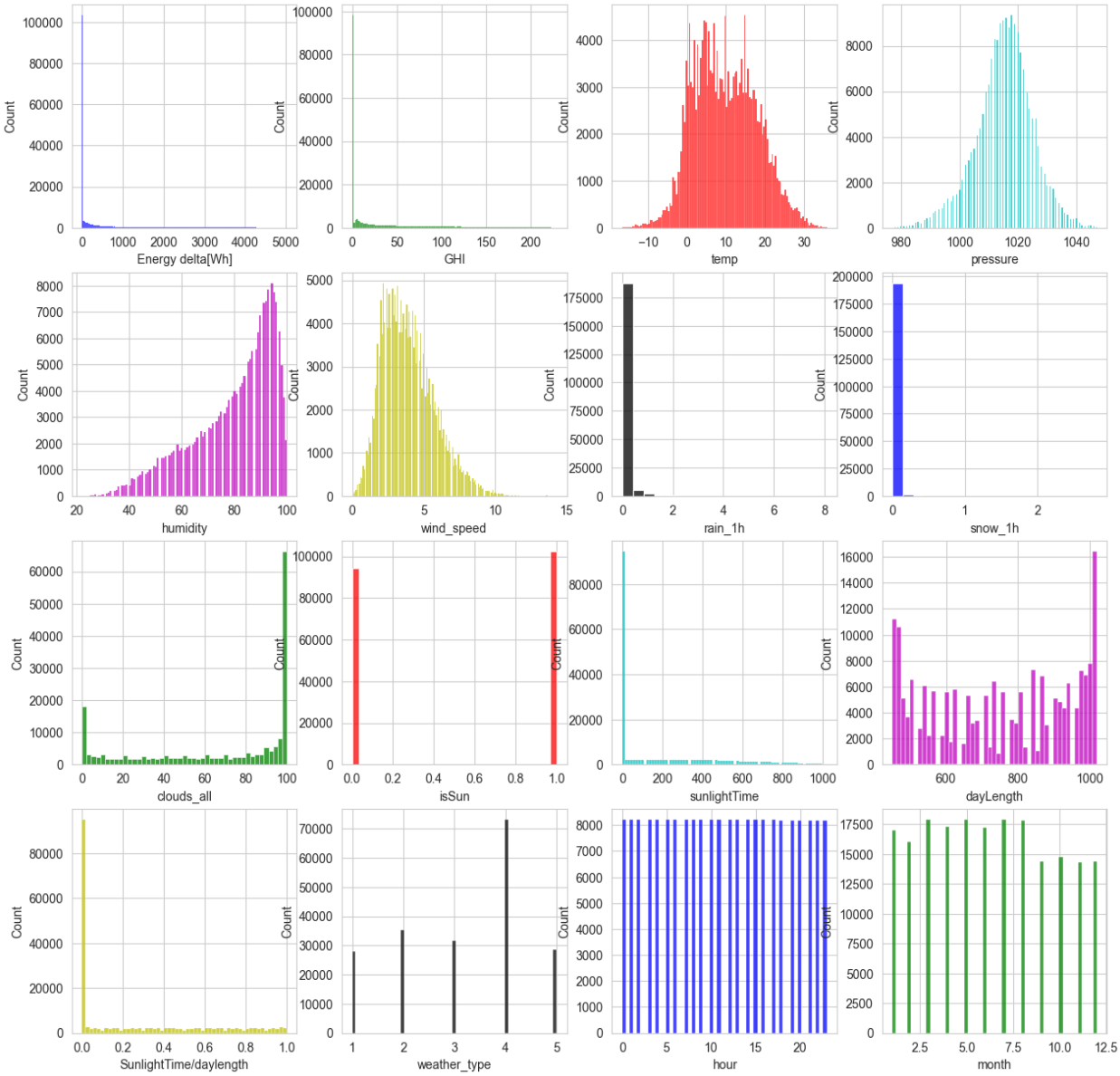

Figure 2

- **Energy Delta (Wh)**: The distribution exhibits a rightward skew, with more frequent occurrences of lower energy delta values and a tail extending towards higher values.
- **Global Horizontal Irradiance (GHI)**: Contrary to the initial observation, the distribution appears right-skewed, with a concentration of data points at lower irradiance values and a tail extending towards higher irradiance levels.
- **Temperature**: The distribution has a central peak and a reasonably symmetrical decline in frequency on either side, which are features of a normal distribution.
- **Pressure**: The distribution shows characteristics of a normal distribution, with a central peak and a relatively symmetrical decrease in frequency on either side.

- **Humidity**: The distribution leans left-skewed, with a lower concentration of data points at lower humidity levels and a peak extending towards higher humidity values.
- **Wind Speed**: The distribution may be right-skewed, with more frequent occurrences of lower wind speeds and a possible tail towards stronger winds. However, the provided data range might limit a definitive conclusion.
- **Rain 1h**: The distribution's shape is complex due to the limited data range on the x-axis.
- **Snow Depth**: The distribution's shape is inconclusive due to the limited data range on the x-axis.
- **Cloud Cover (All)**: The distribution's shape is complex due to the limited data range on the x-axis.
- **Is Sun**: The distribution is bimodal, with distinct peaks at 0 (representing *no sun*) and 1 (representing the *Sun*).
- **Sunshine Time**: The distribution leans right-skewed, with a concentration of data points at lower sunshine time values and a tail extending towards longer sunshine durations.
- **Daylight**: The distribution's shape is complex and cannot be determined definitively due to the limited data range on the x-axis.
- **Weather Type**: The distribution's shape is inconclusive due to the categorical nature of the variable (represented by numerical labels on the x-axis).
- **Hour**: The distribution's shape is inconclusive due to the categorical nature of the variable (represented by numerical labels on the x-axis).
- **Month**: The distribution's shape is inconclusive due to the *categorical nature* of the variable (represented by numerical labels on the x-axis).

***General Observation*s:**
Several features within the dataset exhibit a rightward skew, suggesting a higher frequency of data points concentrated on the left side of the distribution and a longer tail extending towards the right side. This pattern might indicate the presence of outliers on the higher end of the spectrum for these features.
One feature (Is Sun) displays a bimodal distribution, with two distinct peaks representing separate categories within the data. Categorical features (weather type, hour, month) are unsuitable for histogram analysis, as they represent discrete variables rather than continuous numerical values.

Accuracy Comparison and Model Selection

1. **XGBRegressor:**
   Accuracy: 93%
   Mean absolute error: 113.41

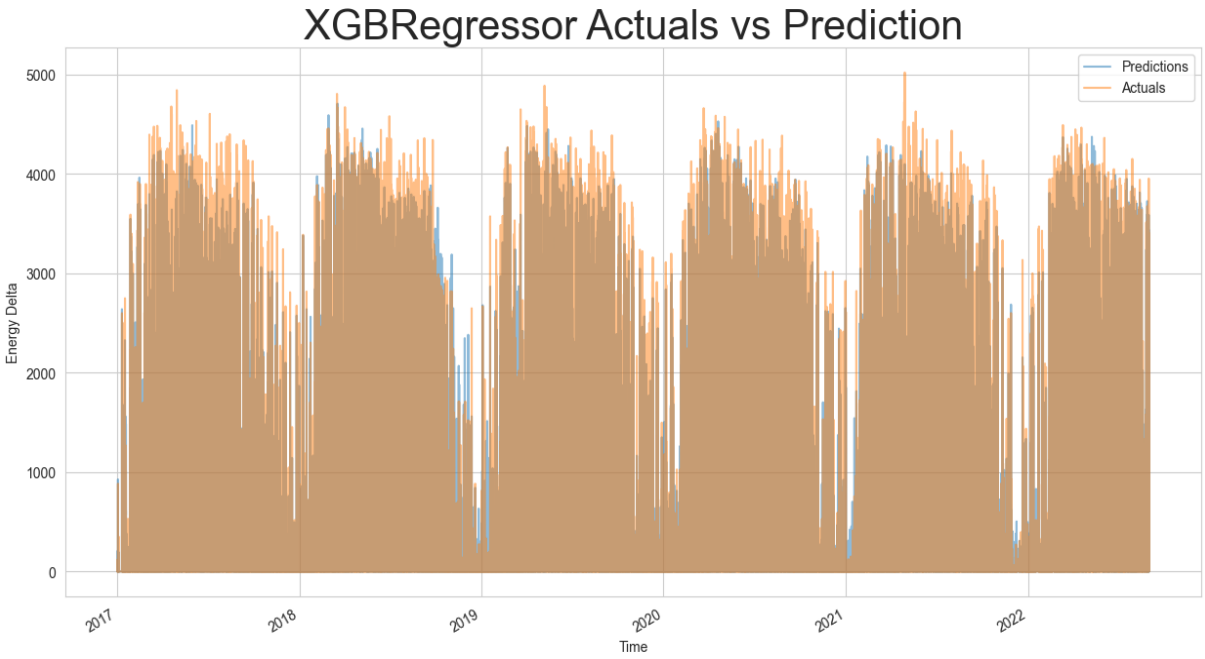

Figure 3

2. **Linear Regression**
   Accuracy: 85.5%
   Mean Absolute Error: 222.99

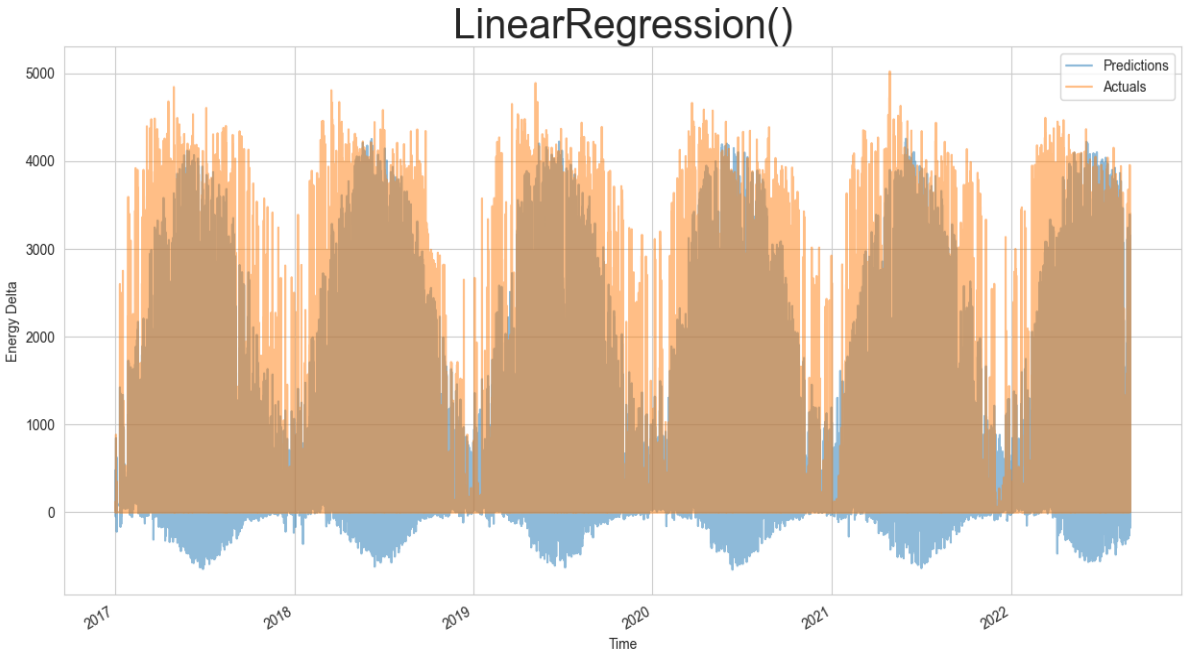

Figure 4

3. **Decision Tree Regressor**
   Accuracy: 86.96%
   Mean Absolute Error: 147.45

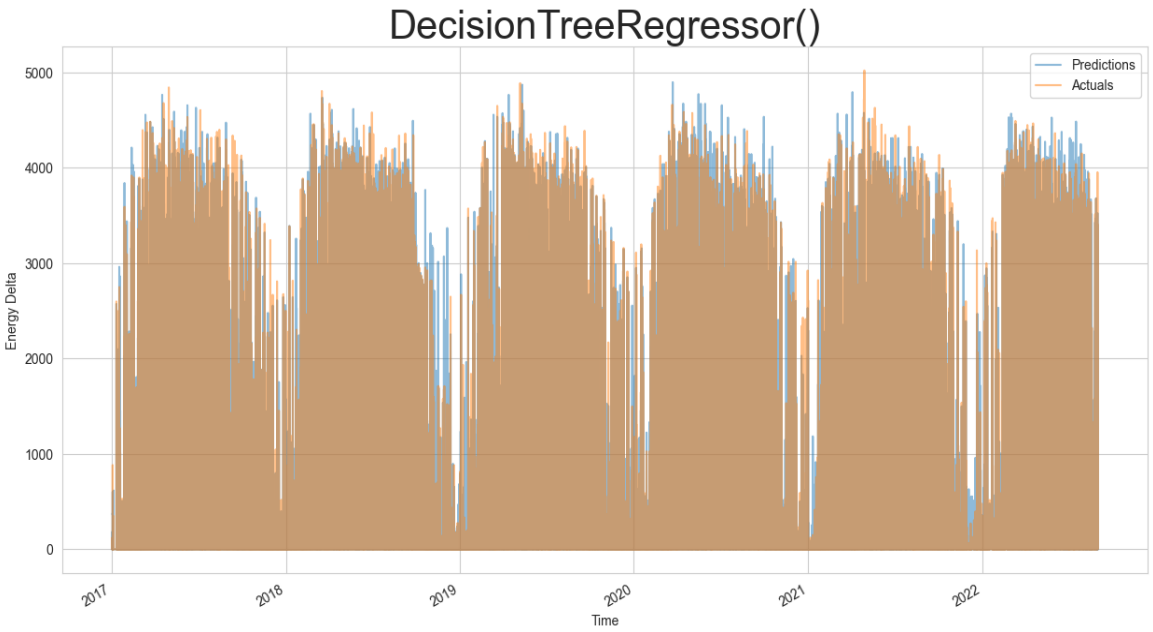

Figure 5

4. **SVR**

   Accuracy: 43.29%

   Mean Absolute Error: 358.74

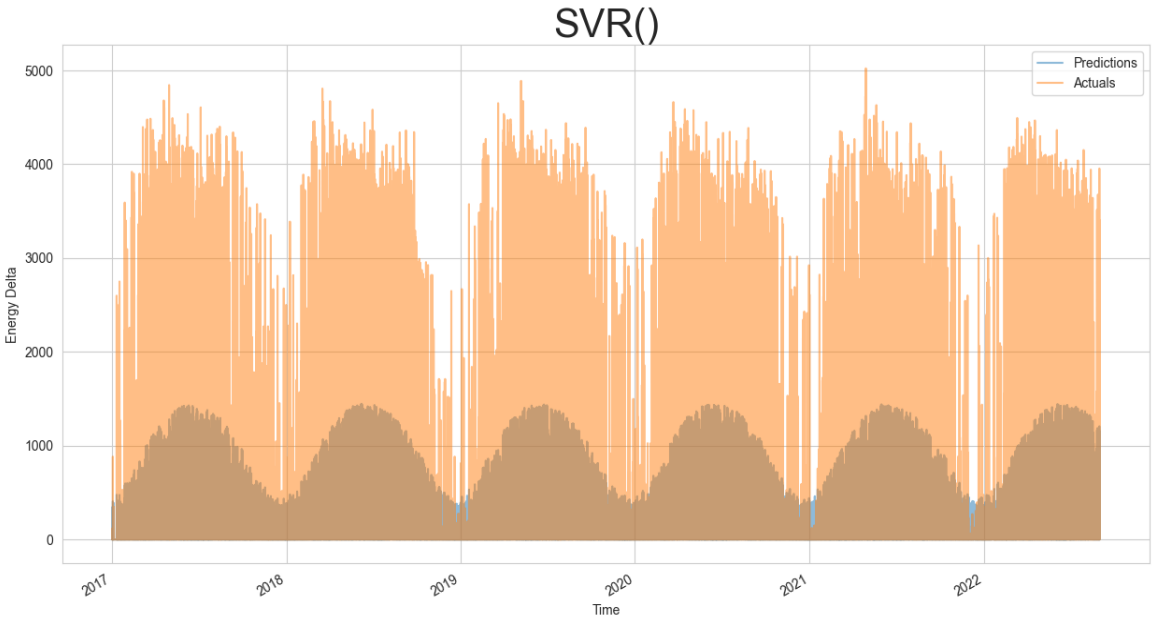

Figure 6

5. **Lasso Regressor**

   Accuracy: 85.46%

   Mean Absolute Error: 224.69

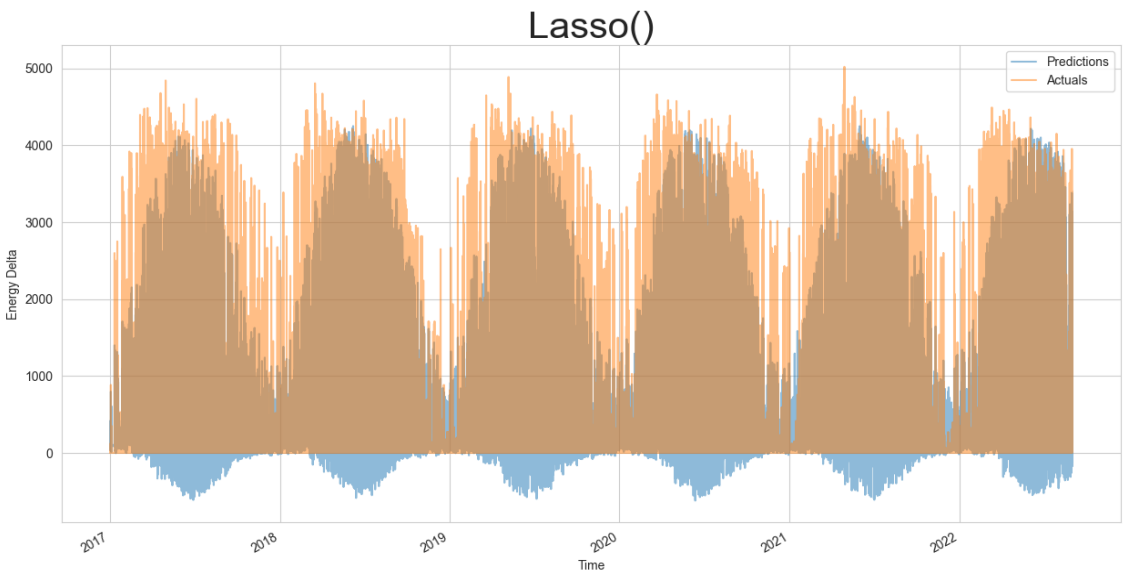

Figure 7

This study shows how well machine learning can forecast the solar energy delta. XGBoost, an ensemble learning method, achieved the highest accuracy of **93% (R-squared = 0.93)** and the lowest error metrics, indicating its potential for practical applications. Future work could involve exploring feature engineering techniques and evaluating the generalizability of the models on unseen data.

Implications:

Broader Societal and Economic Benefits

Beyond the immediate benefits of optimizing grid stability and energy security, this research on data-driven renewable energy consumption forecasts has the potential to unlock broader societal and economic benefits:

- **Reduced reliance on fossil fuels:** Accurate consumption forecasts can inform energy policy decisions and investments in renewable energy infrastructure. This may result in a progressive decrease in our dependence on fossil fuels, lessening their ecological consequences.
- **Enhanced energy security:** Utilities can create plans to ensure a steady electricity supply by anticipating changes in renewable energy output. This can promote a more secure energy grid by lowering the likelihood of blackouts and power outages.
- **Cost optimization for consumers:** Improved forecasting can enable energy providers to offer dynamic pricing models that reflect real-time supply and demand. This can incentivize consumers to shift their energy usage patterns to take advantage of lower prices during high renewable energy production periods.
- **Informed consumer choices:** Public awareness campaigns informed by data analysis can empower individuals to make informed energy consumption decisions.

Understanding the dynamic nature of renewable energy sources can encourage responsible energy use and potentially reduce household energy consumption.

- **Advancement in renewable energy technologies:** The insights gained from analyzing the interplay between weather patterns and energy consumption can guide the development of more efficient and adaptable renewable energy technologies. This can lead to a future where renewable energy sources play an even more significant role in meeting our energy demands.

This research paves the way for a more sustainable and secure energy future by harnessing data analytics. Accurately predicting renewable energy consumption empowers stakeholders across the energy sector – From utilities and legislators to consumers and software developers, educated decision-making may hasten the shift to sustainable energy sources.

Call to Action

The results of this study demonstrate how critical data analysis is to maximising the integration of renewable energy sources. Further research efforts are needed to:

- Refine and improve prediction models by incorporating additional data sources and exploring advanced machine-learning techniques.
- Develop user-friendly interfaces that translate complex data insights into actionable information for stakeholders across the energy sector.
- Conduct cost-benefit analyses to quantify the economic and environmental benefits of implementing data-driven renewable energy forecasting solutions.

We can effectively leverage data analytics to optimize the potential of renewable energy sources and construct a sustainable future for posterity by encouraging cooperation between data scientists, energy experts, policymakers, and members of the public.

Summary and Conclusion

This research project investigated the feasibility of utilizing data-driven approaches to predict renewable energy consumption patterns. A comprehensive dataset was analyzed using various features influencing consumption, including solar irradiance, weather variables, and time-based variations.

Employing a multi-pronged data analysis approach, we explored the relationships between these features and energy consumption. Descriptive statistics, distribution curves, a correlation matrix, and boxplots provided valuable insights into the data's central tendency, spread, potential skewness, outliers, and interquartile range.

Furthermore, the project developed prediction models for renewable energy consumption forecasting. Various regression algorithms were implemented and evaluated, including XGBoost, Linear Regression, Decision Tree Regression, Support Vector Regression, and Lasso Regression. XGBoost emerged as the most accurate model, achieving a score of 93% (R-squared = 0.93) and demonstrating its effectiveness in capturing the underlying relationships between features and energy consumption.

The implications of this research extend beyond optimizing grid stability and energy security. Data-driven forecasts for renewable energy consumption have the potential to unlock broader societal and economic benefits. These include reduced reliance on fossil fuels, enhanced energy security, cost optimization for consumers, informed consumer choices, and advancements in renewable energy technologies.

In conclusion, this research project underscores the significant role of data analysis in maximizing the utilization of renewable energy sources. By harnessing the power of data and leveraging advanced prediction models, we can pave the way for a cleaner and more secure energy future. Future research efforts should focus on refining prediction models, developing user-friendly interfaces for stakeholders, and conducting cost-benefit analyses to quantify the economic and environmental advantages of implementing these data-driven solutions. We can fully utilize renewable energy and guarantee a sustainable future for future generations if different stakeholders continue to work together.

Appendices

1. **Data Preparation and Model Training**
The data preparation involved separating the features (predictors) from the target variable (Energy Delta). Next, the data was divided into testing and training sets using a 70/30 ratio. Stratification by month ensured that both sets had a similar distribution of data points across months.

2. **Accuracy Comparison and Model Selection**
Based on the results, XGBoost emerged as the most accurate model, achieving a score of **93% (R-squared = 0.93)**. This indicates that XGBoost effectively captured the underlying connections between the goal variable and the features. While Linear Regression, Decision Tree Regression, and Lasso achieved respectable scores, XGBoost's superior accuracy and lower error metrics (MAE and RMSE) suggest it is the best choice for making reliable predictions of solar energy delta.

List of References:

- Home | Citicore Renewable Energy REIT Corporation. https://creit.com.ph/

- 3D Integration for the Internet of Things - Arm-ECS Research Centre. https://www.arm.ecs.soton.ac.uk/projects/3d-integration-for-iot/
- (2023). India: Rajasthan, Gujarat, and Tamil Nadu emerge as top achievers in wind energy adoption. MENA Report, ().
- How to Read a Correlation Matrix - A Beginner's Guide. https://www.quanthub.com/how-to-read-a-correlation-matrix/
- (2023). Inferring user preferences using cardinal vs. Ordinal feedback in recommender systems. IIE Annual Conference. Proceedings, (), 1-6.
- How to Plot an Exponential Distribution in R - Life With Data. https://lifewithdata.com/2023/07/31/how-to-plot-an-exponential-distribution-in-r/
- 3D Integration for the Internet of Things - Arm-ECS Research Centre. https://www.arm.ecs.soton.ac.uk/projects/3d-integration-for-iot/
- Kaggle Dataset for building a prediction model/