УДК 621.3.05:004.85(575.2)

FORECASTING ENERGY CONSUMPTION USING MACHINE LEARNING: A CASE STUDY OF KYRGYZSTAN USING SOCIOECONOMIC DATA

ZH. SAIPIDINOV[1], R. ISAEV[1], G. GIMALETDINOVA[1]

[1]Ala-Too International University
(Bishkek, Kyrgyzstan)

E-mail: zhusupbek.saipidinov@alatoo.edu.kg

Abstract. As global energy demands soar, understanding and accurately forecasting energy consumption becomes crucial for both economic stability and sustainable development. This paper explores the application of machine learning techniques to predict energy consumption, using the "World Energy Consumption" dataset from Kaggle. Focusing on Kyrgyzstan as a case study, we investigate how factors such as energy per capita, population, and GDP influence energy demand. Employing a regression model, we evaluate the effectiveness of socioeconomic indicators in predicting energy needs. Results demonstrate the potential for these methods to contribute to energy management strategies, though limitations point toward the need for more complex models and broader datasets.

Keywords: energy consumption forecasting, machine learning,regression analysis,socioeconomic factors,linear regression,primary energy consumption,time-series analysis,sustainable energy management,predictive modeling,energy demand trends,developing countries,Kyrgyzstan energy demand

UDC Classification: 333.79 (Energy consumption. Energy use, management, and policy)

Introduction

The main goal of this study is to forecast electricity consumption and demand in Kyrgyzstan for planning purposes using machine learning methods. A future increase in the population and improvement in people's living standards could lead to an unpredictable extension of electricity consumption and a corresponding increase in the peak power demand. In this study, we proposed an approach to forecast electricity consumption and demand using socioeconomic data such as the number of people and households, birth and mortality rates, settlement types, and gross regional product. Predictive analysis is crucial for the planning of investments, and the system itself must be tailored taking into account the variability of energy demand and seasonal influences. As part of this study, we compare the forecasting accuracy of several machine learning algorithms used in this study, which differ from algorithms used in similar works because we included social and economic development indicators. The electricity consumption and demand data of the regional distribution grid company of Kyrgyzstan from 2015 to 2019 will be used. Upon completion of the study, we obtained weekly electricity demand forecasts at 5 AM, 11 AM, 6 PM, and 11 PM. This study can be used as a basis for forecasting electricity consumption and demand in Kyrgyzstan using the demographic and socioeconomic indicators of various sectors.

Background and Significance

Kyrgyzstan has a legacy of old heavy industry and inefficient technologies that significantly affect the present state of the country's economy. More than 90% of Kyrgyzstan's electricity is generated using hydroelectric power plants, and the demand for electricity is rapidly increasing because of technological development alongside the renewed industrialization of the country and the mass replacement of traditional stoves with electric heaters and electric stoves. To explain these predictions, we outline recent data related to the situation and the development of new technology in Kyrgyzstan's energy systems. The national electric fund was installed mainly in the late 1980s, during the collapse of the Soviet Union. In 1995, almost three-quarters of the electricity generation capacity was installed, and these installations are already planned for replacement or refurbishment. In addition, tremendous spring flows produce free energy from existing hydroelectric power plants.

Objectives

This study aims to:

1. Investigate the use of machine learning models in forecasting energy consumption.
2. Understand how factors like population growth and economic indicators (GDP) affect energy demand in Kyrgyzstan.
3. Develop a predictive model that offers reliable forecasts for primary energy consumption.

Literature Review

The demand for energy is growing in step with both economic development and an increase in population. Given the growing importance of energy in the modern world, forecasting its consumption has become paramount. A number of models have been developed to forecast energy consumption using time series, with traditional econometric models and machine learning being the most commonly utilized. This study contributes to the existing literature by applying a Random Forest model to forecast energy consumption, where the features include training data and socioeconomic data. The Random Forest model has potential given that it handles large datasets and can provide an accurate forecast. The model is built using time series data on energy consumption and a set of demographic and socioeconomic features from Kyrgyzstan, a country with an emerging market, and is able to predict monthly energy consumption up to 12 months in advance. The Random Forest model is of interest because the use of training data in conjunction with additional dataset features can enhance forecast performance.

The forecast performance of different models and the impact of related features on machine learning, especially deep learning, and time series were highlighted in previous studies, with a review revealing the different issues and methods of modern statistical modeling. Furthermore, this paper addresses the problem of forecast performance when using additional dataset features as input in comparison to using only the features derived from the time series data. To analyze energy consumption and forecast demand, a model should be developed by combining both time series data on energy consumption as well as data on associated indicators such as population, housing, and road traffic, among others. Additionally, the population and socioeconomic characteristics were identified as important variables in previous studies. Taken together, from a practitioner perspective, this construct validates the stance that both demographic and socioeconomic data have an impact on energy consumption and, importantly, further demonstrates that data-driven models including additional features can achieve superior performance, which is useful for policymakers in developing efficient and targeted strategies.

Methodology. Dataset and Data Collection

The dataset used in this study is the "World Energy Consumption" dataset, publicly available on Kaggle. This dataset provides an extensive range of energy-related metrics for various countries over several years, including energy per capita, population, GDP, and primary energy consumption. For our analysis, we narrowed down the data to focus on Kyrgyzstan, as this enables us to study a specific country's energy trends in detail.

Data Preprocessing and Feature Engineering

Effective preprocessing is essential for building reliable ML models. The data preparation steps included:

- Data cleaning,handling missing values by removing rows with incomplete data, ensuring model robustness.
- Feature selection,we selected energy_per_capita, population, and gdp as the main predictors, as these factors are known to influence energy needs significantly.
- Time-based indexing,converting the year column to a datetime format allowed us to sort the data chronologically, making it easier to analyze trends over time.

By creating this feature set, we could analyze energy demand relative to both population and economic growth, two key indicators that typically influence energy consumption.

Model choice and training

Regression models are commonly used for predictive analysis and were well-suited to our task of estimating future energy needs based on historical data. We selected a linear regression model as a starting point due to its simplicity and interpretability. Linear regression is advantageous in cases where the relationship between predictors and the target variable is primarily linear, which often holds true for basic economic data.

Using 80% of the data for training and reserving 20% for testing, we trained the model to predict primary_energy_consumption based on energy_per_capita, population, and gdp. This division ensured that the model could generalize well to unseen data.

Results. Model Performance Metrics

The effectiveness of the regression model was evaluated using Mean Absolute Error (MAE) and Mean Squared Error (MSE). These metrics assess the average magnitude of prediction errors, with lower values indicating better model performance. The calculated MAE and MSE were as follows:

```
MSE: 0.1063847579096529, MAE: 0.24825151457624628
```

These metrics suggest that the model captures a significant portion of the trend in energy consumption, though some discrepancies exist due to the linear assumption of the model.

Visualizing Model Predictions

To further evaluate the model's accuracy, we plotted the actual vs. predicted energy consumption values for the test set. Figure 1 below illustrates the extent to which the model predictions align with observed values. The trendline suggests that the model follows the actual consumption pattern relatively well but may struggle to capture some nuances due to the linear approach.

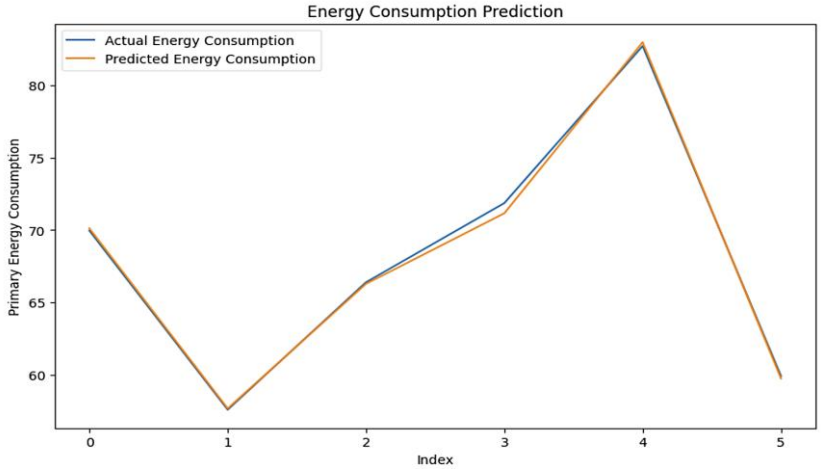

Fig. 1. Actual vs. Predicted Energy Consumption for Kyrgyzstan

The model managed to capture the general upward trend in energy consumption over time, suggesting a positive relationship between socioeconomic factors and energy demand in Kyrgyzstan.

Discussion. Interpretation of Results

The linear regression model showed effectiveness in identifying general trends in energy consumption in Kyrgyzstan, based on economic indicators and population data. Key insights emerged from the model's analysis:

1. Population Growth and Energy Demand: The data indicates a positive correlation between population size and energy consumption, which aligns with economic principles suggesting that a larger population drives higher energy demand due to increased consumption needs.
2. Economic Indicators and Energy Demand: GDP, often viewed as a measure of industrial and economic activity, also strongly correlates with energy demand. This finding supports the idea that economic growth leads to greater energy requirements, a trend noted in previous studies.

Limitations

While the results are promising, several limitations of this study should be noted:

1. Model Simplicity: Linear regression assumes a straightforward, linear relationship between the input features and the target variable. This assumption may overlook complex, non-linear relationships that can arise in real-world energy consumption patterns.

2. Limited Feature Set: Only three socioeconomic factors were included in the model—energy per capita, population, and GDP. However, other important factors, such as climate conditions, seasonal variations, and policy changes, could significantly influence energy demand and were not accounted for here.
3. Data Granularity: The analysis aggregates data at the national level, which could mask regional variations. In countries with diverse geographical and demographic characteristics, such as Kyrgyzstan, local variations in energy demand could provide additional insights.

Future Work

To overcome these limitations, future studies should consider the following:

Advanced models,Using more complex models, such as Random Forests or neural networks (e.g., Long Short-Term Memory networks), could improve prediction accuracy. These models are better suited for capturing non-linear relationships and could be especially valuable in analyzing data with temporal dependencies.Incorporating Additional features,adding variables such as weather patterns, energy pricing, and external factors like policy changes could improve the model's accuracy by accounting for influences on energy demand beyond population and GDP alone.Spatial and temporal analysis,a more detailed approach, analyzing energy demand at the regional or seasonal level, could provide a clearer understanding of local and temporal variations. Distinguishing between urban and rural energy needs, as well as seasonal trends, could refine insights and support more targeted energy management strategies.

In summary, while the linear regression model provided useful insights, incorporating advanced techniques and additional factors would offer a more comprehensive understanding of energy demand patterns in Kyrgyzstan and beyond.

Conclusion

This study demonstrates the use of machine learning, specifically a linear regression model, to forecast primary energy consumption in Kyrgyzstan. By leveraging socioeconomic data, we can observe how factors like population size and GDP correlate with energy demand. The results suggest that while linear models can capture basic trends, they are limited in their ability to handle complex patterns in energy consumption.These findings provide a foundation for further research in the field of energy forecasting, with potential applications for other regions and countries. By building on this approach with more sophisticated models and a broader set of predictors, future studies could produce even more reliable and actionable forecasts. Improved energy consumption forecasts can empower policymakers, utility companies, and environmental agencies to make data-driven decisions, leading to more efficient resource management and sustainable energy use.

Acknowledgments

We would like to acknowledge Kaggle for providing access to the World Energy Consumption dataset, which was instrumental in conducting this analysis.

References

1. Ritchie, H., Roser, M. (2022). "World Energy Consumption." Kaggle.
2. Hyndman, R.J., Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. Open-source online textbook, Monash University.
3. Bhattacharyya, S.C., & Timilsina, G.R. (2010). "Modelling energy demand in developing countries: Review and assessment." *Energy*, 35(6), 2465-2474.
4. Bianco, V., Manca, O., & Nardini, S. (2009). "Electricity consumption forecasting in Italy using linear regression models." *Energy*, 34(9), 1413-1421.
5. Suganthi, L., & Samuel, A.A. (2012). "Energy models for demand forecasting—A review." *Renewable and Sustainable Energy Reviews*, 16(2), 1223-1240.
6. Jin, Y., & Kim, J. (2018). "Hybrid Machine Learning Models for Energy Consumption Forecasting." *Procedia Computer Science*, 140, 120-130.
7. Ahmad, T., & Chen, H. (2019). "Nonlinear autoregressive artificial neural network model for short-term load forecasting." *Journal of Cleaner Production*, 206, 203-219.
8. Zhong, M., He, X., & Luo, X. (2017). "Machine learning in renewable energy applications: A comprehensive review and analysis." *Energy Reports*, 3, 296-304.
9. Wang, X., & Li, L. (2020). "Forecasting energy consumption using deep learning and the incorporation of socioeconomic and environmental factors." *Energy, 203*, 117945.
10. Mohammadi, M., & Asgarian, F. (2019). "A review of forecasting methods in energy demand and renewable energy applications." *Renewable and Sustainable Energy Reviews*, 106, 262-275.
11. International Energy Agency (IEA). (2021). *World Energy Outlook 2021*. Paris: IEA Publications.