United States Energy Consumption Analysis and Modeling

Harbaksh Kaur
University of California Los Angeles

AOS C111: Introduction to Machine Learning for Physical Science

Introduction

Energy consumption is a pressing topic globally as most of the world's energy comes from fossil fuels and rising consumption levels means a negative impact on the environment. More energy consumption will contribute to climate change from the greenhouse gasses released from the production process. Energy is essential for an industrialized society as it powers entire industries, cities, and homes. Energy is used for air conditioning and heating, powering medical equipment, powering our electronics, machinery, and even public transportation systems. According to the Environmental Protection Agency, in 2022 25% of the United States's greenhouse gas emissions came from electricity use. Understanding energy consumption patterns for electricity usage would be beneficial for policymakers to create policy towards shifting to renewable energy.

Data

The data will be provided by a Kraggle data set that looks at global energy consumption trends from 2000-2020.
https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy

The variables used:

- Electricity from fossil fuels (TWh): Electricity generated from fossil fuels in terawatt hours.
- Electricity from renewables (TWh): Electricity generated from renewable sources in terawatt hours.
- Primary energy consumption per capita (kWh/person): Energy consumption per person in kilowatt-hours.
- Access to electricity (% of population): What percentage of the population has access to electricity.
- Renewable energy share in the total final energy consumption (%): How much of the total energy consumption was energy created from renewable sources.
- Year: Time is important to look at energy consumption over time.

The pre-processing steps included removing all invalid data values or missing values as well as removing the data of any NaN values if present. The dataset was further narrowed by focusing on the United States.

Modeling

The modeling for this project focused on finding the best machine learning algorithm that provides the best fit for its ability to show the relationship between renewable and nonrenewable energy as well as train a linear regression model that would accurately predict energy consumption levels from renewable energy and fossil fuels.

I first isolated the variable energy consumption per capita, the target variable, and normalized the data. The data was normalized to prepare it for linear regression modeling and it was then split with a train_test_split function with a 40% test size and a random state of 42. A 20% test size resulted in less data points making the correlation between the data harder to relate. I modeled a linear regression to determine fit and the results yielded a 96.3% accuracy rate which is very high meaning a linear regression model would provide an accurate fit. The RMSE was 1075.98 meaning the model will be off around 1076 kWh/person which is reasonable given that the average American's energy consumption ranges from 70,000 kWh/person to 90,000 kWh/person. While the model fits the data well, there are some outliers once the energy consumption gets very high to levels of around 90,000 kWh/person, however, it is unlikely that it would impact the model severely due to the high accuracy rate. To further check the accuracy and fit of the linear regression model, I also plotted the residuals. There was no trend to the residuals further indicating that a linear regression model was a good fit at looking at the relationship between the variables.

A decision tree model was conducted to further test the relationship between the variables as it will show if the relationship is non-linear. The model produced an 87.9% accuracy rate which is high but not as accurate as the linear regression model, indicating that it could be useful for analyzing the relationship between the variables but a linear regression model would be the most accurate in predicting trends and the relationship between the variables is mainly linear.

A scatterplot showing the relationship between renewable energy and fossil fuels in energy consumption in Figure 4 shows that there is a negative correlation between the two variables with more fossil fuel usage resulting in less renewable energy usage and vice versa. The correlation between renewables and energy consumption was -0.84 and the correlation between fossil fuels and energy consumption was 0.44 showing that renewable energy results in less energy consumption.

Results

Figure 1. Linear Regression of Expected Energy Consumption (kWh/person)

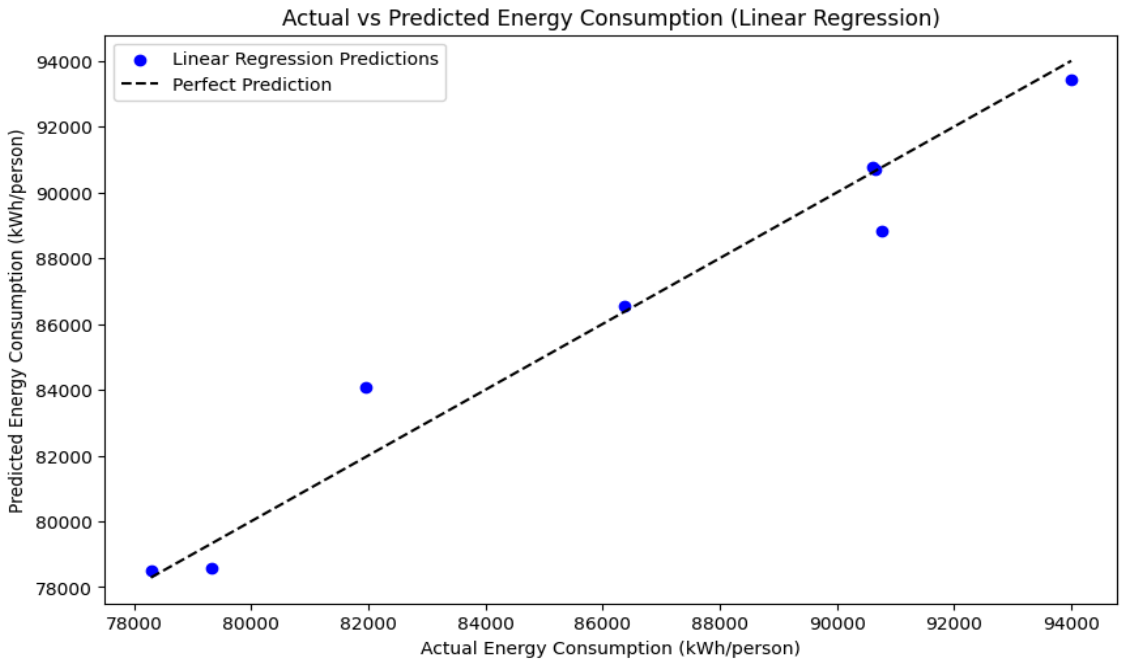


Figure 2. Residuals Plot

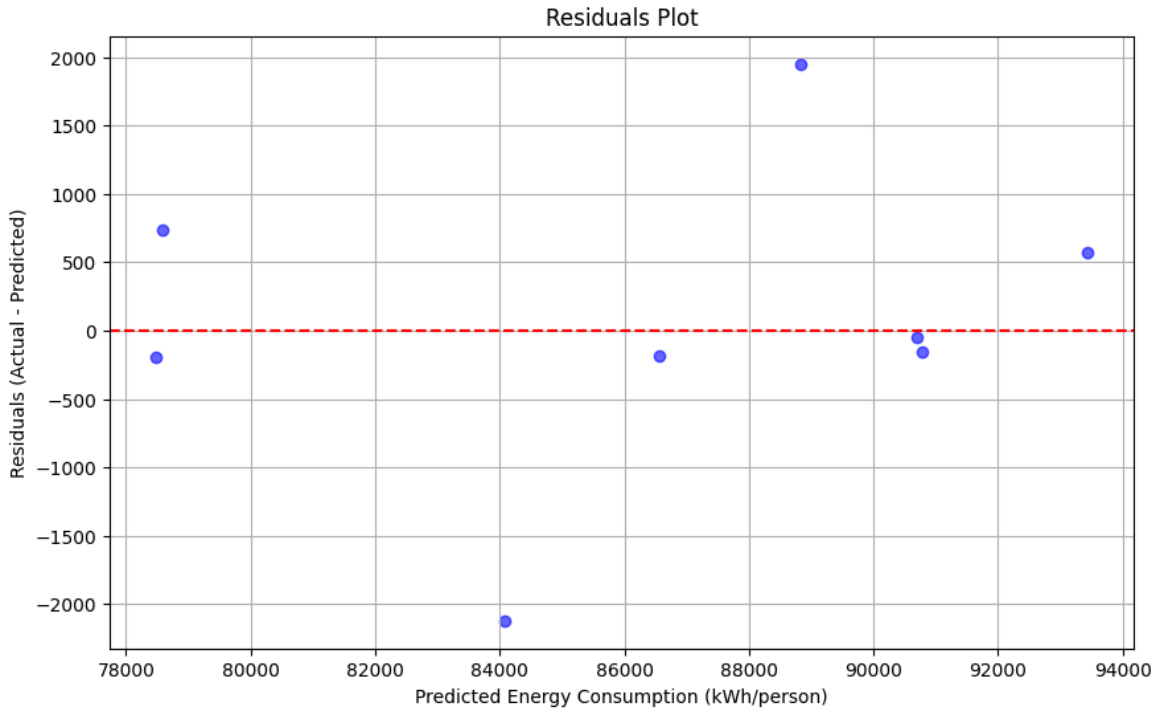


Figure 3. Decision Tree Model for Renewable vs. Fossil Fuel

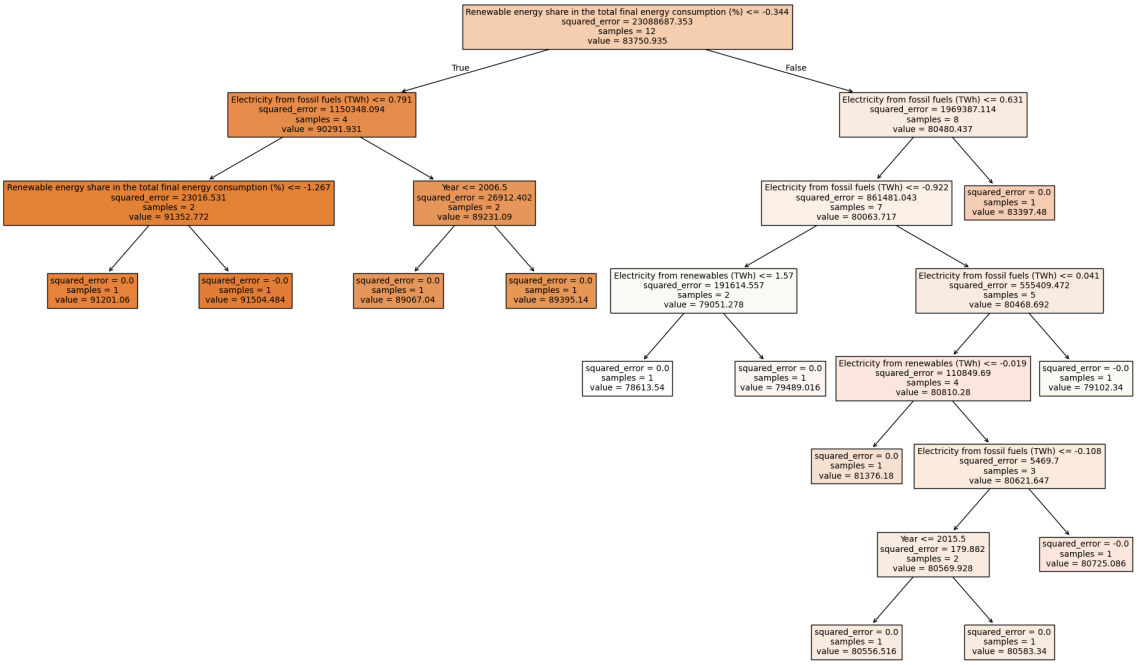


Figure 4. Scatter Plot for Fossil Fuels vs. Renewable Sources

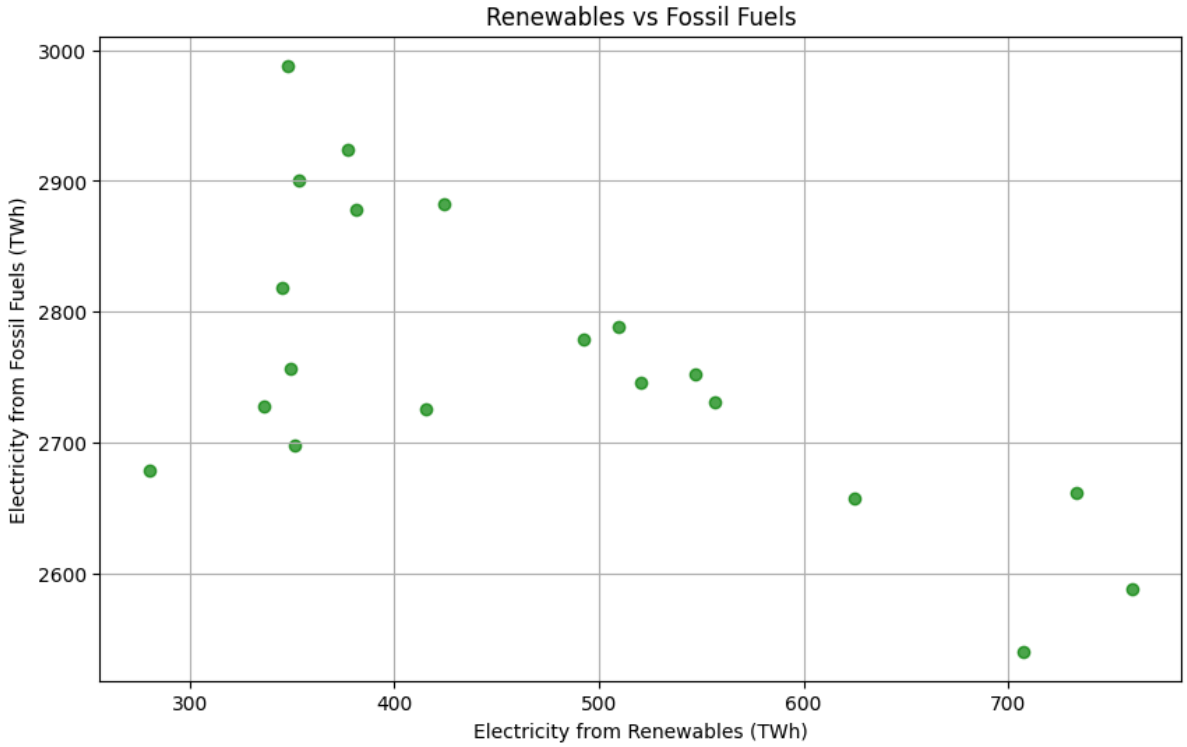


Figure 5. Prediction of Future Energy Consumption for Renewable Energy vs. Fossil Fuels

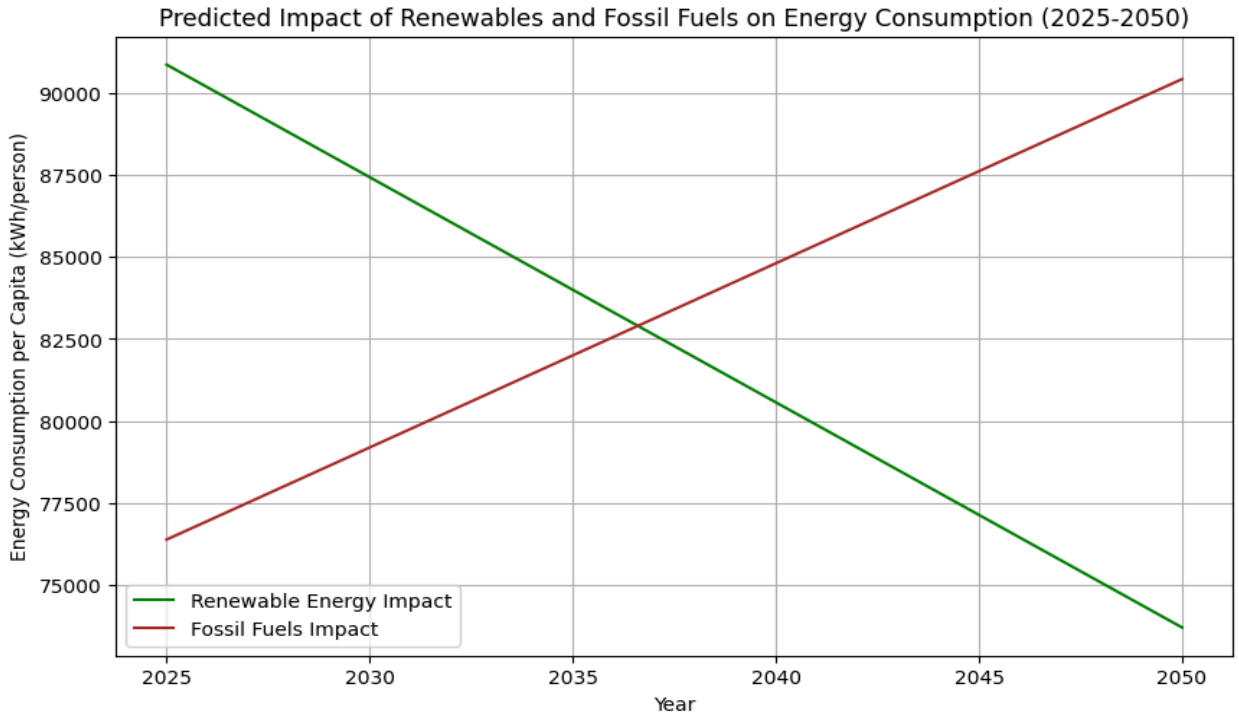
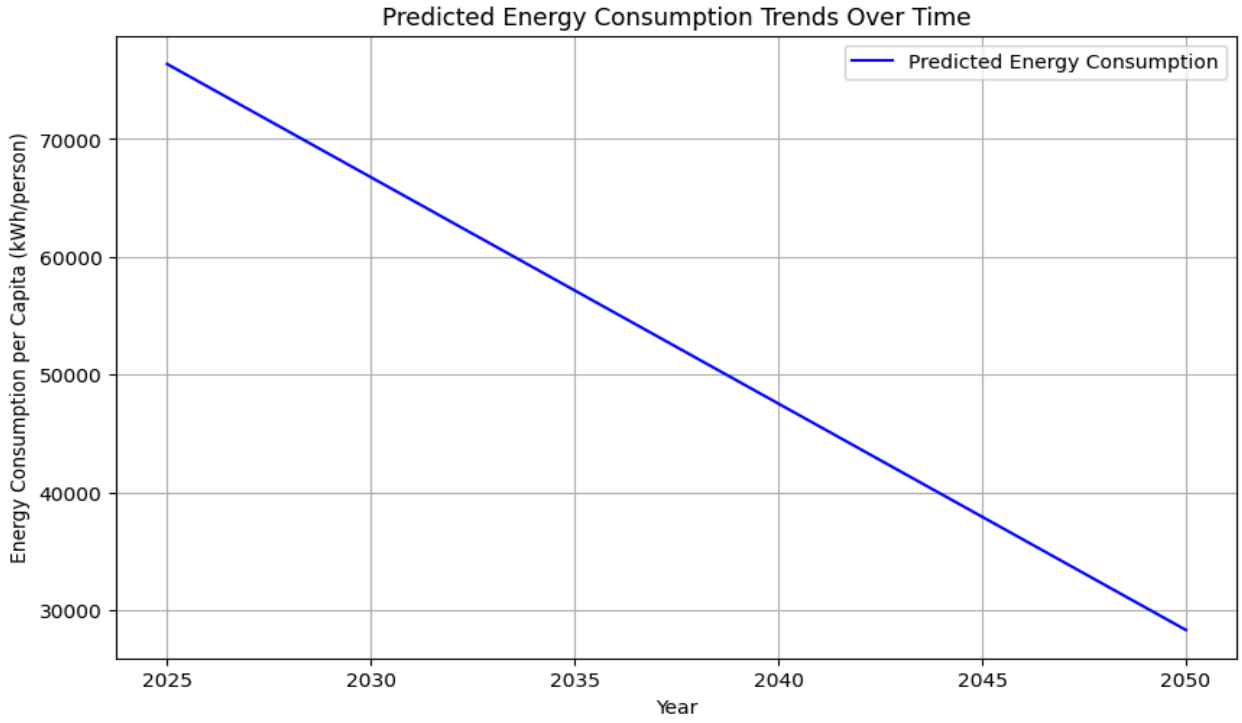


Figure 6. Future Energy Consumption Over Time



Discussion

The linear regression model in Figure 1 shows the model's ability to accurately predict energy consumption with the data provided. The model is plotting the actual energy consumption values from the dataset against the predicted values from the model. A perfect prediction would have all the points aligned at the dotted diagonal line and the points are very closely aligned to the line, indicating that the model is able to predict the energy consumption values fairly accurately. The accuracy rate for this model is 96.3% indicating that the relationship between the input variables and the target variable is linear. It also provides insight into how different shifts in energy source can affect overall energy consumption. The residuals plot from Figure 2 shows the data points do not have a trend or pattern to them further indicating a good fit. Figure 3 shows the relationship between the variables to also have a nonlinear pattern as the accuracy for the decision tree model was 87.9%. The lower accuracy could be due to decision tree models being more prone to overfitting and modeling more noise however because of the high accuracy it can be concluded that the model was able to depict the nonlinear aspect of the dataset and show the nonlinear relationship between the input features and the target variable. Tuning the decision tree model did not present more accurate results. Figure 4 shows a negative correlation between fossil fuel created electricity and renewable sources energy production indicating that as the United States invests in more renewable energy sources, its reliance on fossil fuels decreases. However, the gap is still large indicating that the United States is still reliant on fossil fuels and uses it as a main source of electricity.

Moreover, Figure 6 shows future energy consumption trends over time, until 2050, will decrease and the models have shown that increased renewable energy would result in lower energy consumption over time. This shows that the United States is taking steps to decrease their carbon emissions and that it is working to some degree if the pattern continues.

As 25% of the United States' greenhouse gas emissions comes from electricity usage, it is imperative that the nation focuses on investing in renewable energy resources to offset those emissions as much as it can as global temperatures continue to rise and climate change continues to worsen. It is beneficial to see the future energy consumption trends and the decline in energy consumption over time infers that continued offsetting of fossil fuel energy with renewable energy is a good investment for the United States. Unfortunately, drastic changes to legislation have not been made to create a large substantial change in the United States' climate policy and the continued denial of climate change further exacerbates the issue. Using the models to create policy surrounding energy consumption and fossil fuel usage can help create a more sustainable future and help mitigate the risks of climate change. Using machine learning to accurately

predict what future energy consumption trends could result in will allow policymakers to have access to accurate information about what the future would look like and create better legislation regarding energy efficiency and climate change.

Further research in how different energy sources affect greenhouse gas emissions would be useful in determining what energy sources would be most efficient and also most cost effective in mitigating the effects of climate change. Using more machine learning models to accurately predict information surrounding climate change could be beneficial in garnering more public support in the urgency of climate change.

Conclusion

From this report, the following conclusions can be drawn:

- Linear regression was the best fit for this model and data.
- The relationship between the features was linear.
- The model predicts energy consumption in the United States will decline in the future.
- There is a negative correlation between renewable energy and fossil fuels in energy consumption levels.

References

https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks