Trends and Estimations of US Energy generation and prices

Capstone project for Udacity Data Science Nanodegree.

Author – Aayush Soni. 5th June,2024.

PROJECT OVERVIEW

The quest for sustainable energy sources has never been more critical. However, the transitions towards a greener energy utilization landscape brings forth its own set of challenges, and understanding energy generation and consumption patterns becomes paramount.

In this project, I shall look into the production and utilization of energy across the United States to uncover insights and identify trends.

PROBLEM STATEMENT

The goal of the project is to implement a basic ML pipeline to estimate costs of energy for different sectors based on energy and demographic statistics for the different states. The project deals with the following parts:

1. **Data gathering and data cleaning**: Gathering, preliminary processing and condensing of various datasets
2. **Data Exploratory Analysis**: A deeper look into our data to understand information provided and uncover interesting questions to ask
3. **Machine Learning Pipeline**: Creating a Regression pipeline to process and estimate electricity costs.
4. **Creating a Web App**: Creating a basic web app to highlight our learnings. The app will also feature the trained ML model which accepts user input to generate an estimate of electricity prices.

METRICS

R2 Score

The accuracy of our model is estimated using R2 Score, which is a measure of how much of the variance in the dependent variable is predictable from the independent variable. It is calculated as:

$$R2\ score = 1\ - \frac{variance\ in\ residuals\ of\ observations\ and\ predicted\ data}{variance\ of\ observations}$$

A low R2 score implies a lower fit of the model (the model only accounts for a little of the observed variance in data) whereas an R2 score of 1 implies a perfect fit – the model perfectly encapsulates all the variance in the observations.

RMSE per category

The individual RMSE error is also calculated to compare fit of the model to each individual label.

ANALYSIS

Data Exploration and Pre-Processing

The main dataset being used is the monthly energy production data made available by The US Energy Information Administration. The dataset contains a massive amount of data collected including:

- Fuel receipts and costs
- Generator data including generation, fuel consumption and stocks
- Fossil fuel stocks
- Non-utility source and disposition of electricity
- Environmental data

For the purposes of this project, I have focussed only on the data generation subsection of point-2 above.

Data of price per unit of electricity was downloaded from this Kaggle link by user Alistair King. The data was represented as price per month per state for different end-customers, i.e. *residential, commercial, industrial, transportation, and all sectors.* The dataset included an additional field, i.e. *transportation* however the data was sparsely populated and so the column was dropped.

These values needed to be inflation-adjusted to minimize dependence on time on each individual observation. Inflation adjustment figures were captured from this site and manually converted to show dollar value with respect to 2010 dollar value.

Finally, I wanted to capture state-wise population as a separate datapoint, to compensate for the increase in electricity production/ price fluctuations due to an increase in state populations. Population figures were taken from census data available on Wikipedia and interpolated for each month.

Post cleaning the following fields are kept in the dataset:

- Grouping fields:
  - *date*: Month of record
  - *state*: 2-character representation of state. This
- Metadata
  - *population*: Estimate of population in each state in the given month
- Electricity generation data
  - *fossil*: aggregation MWh of electricity generated by all fossil-fuel sources
  - *nuclear*: MWh of electricity generated by nuclear power
  - *hydro*: MWh of electricity generated by hydro power
  - *solar*: MWh of electricity generated by solar power
  - *wind*: MWh of electricity generated by wind power
  - *secondary*: MWh of electricity generated by secondary carbon-based power sources
  - *geo*: MWh of electricity generated by geothermal power
- Electricity price data.
  - *Commercial:* Cost in cents / kWh of electricity for commercial consumers
  - *Industrial:* Cost in cents / kWh of electricity for industrial consumers
  - *Residential:* Cost in cents / kWh of electricity for residential consumers
  - *All sectors:* Average cost in cents / kWh of electricity across all consumers

A total of 14,076 data points were collected, representing the data of 51 states (incl. DC) across 23 years (from 2001 January till 2023 December). All prices are normalized with respect to 2010 dollar purchasing power.

Exploratory Visualizations

1. *Change in electricity generation over time*

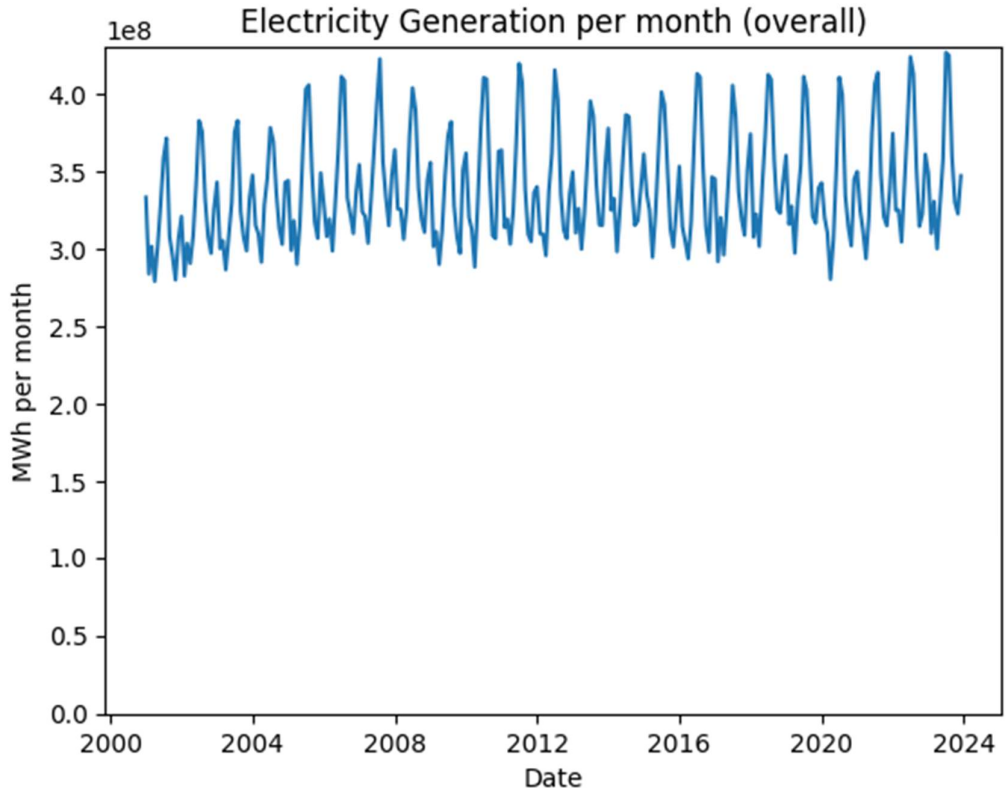

Figure 1: Electricity Generation per month

Electricity generation follows annual peaks and troughs, with peaks around the middle of the year and troughs around the beginning and ends of the year. The total electricity generated is trending upwards since the early 2000's.

2. *Electricity generation distributed across the different generation types*

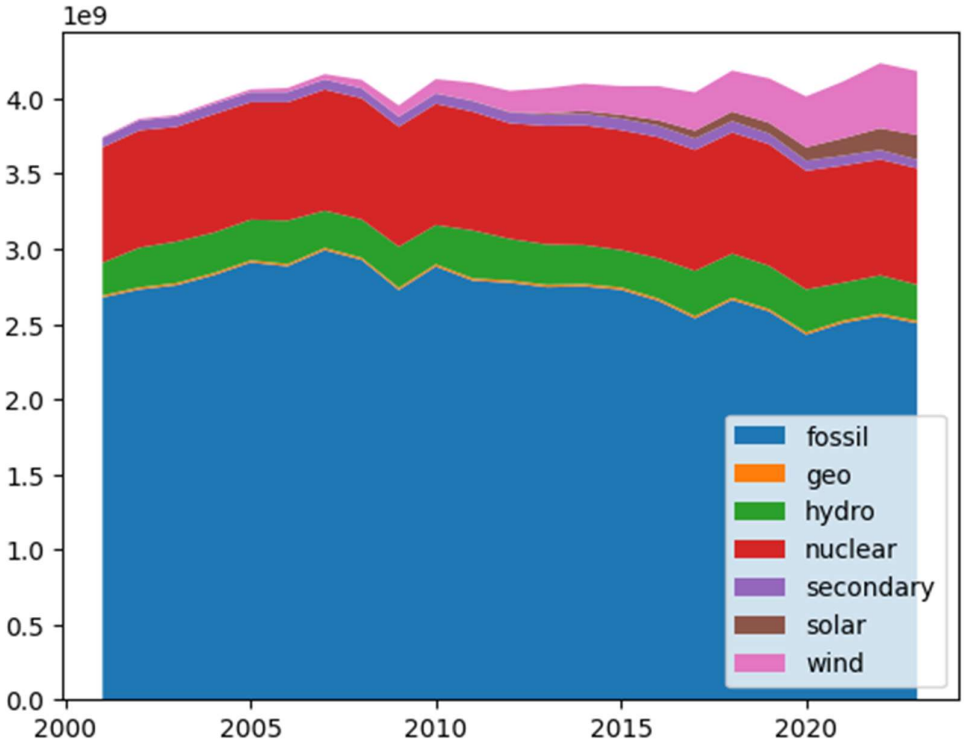

Figure 2: Electricity production divided by generation category

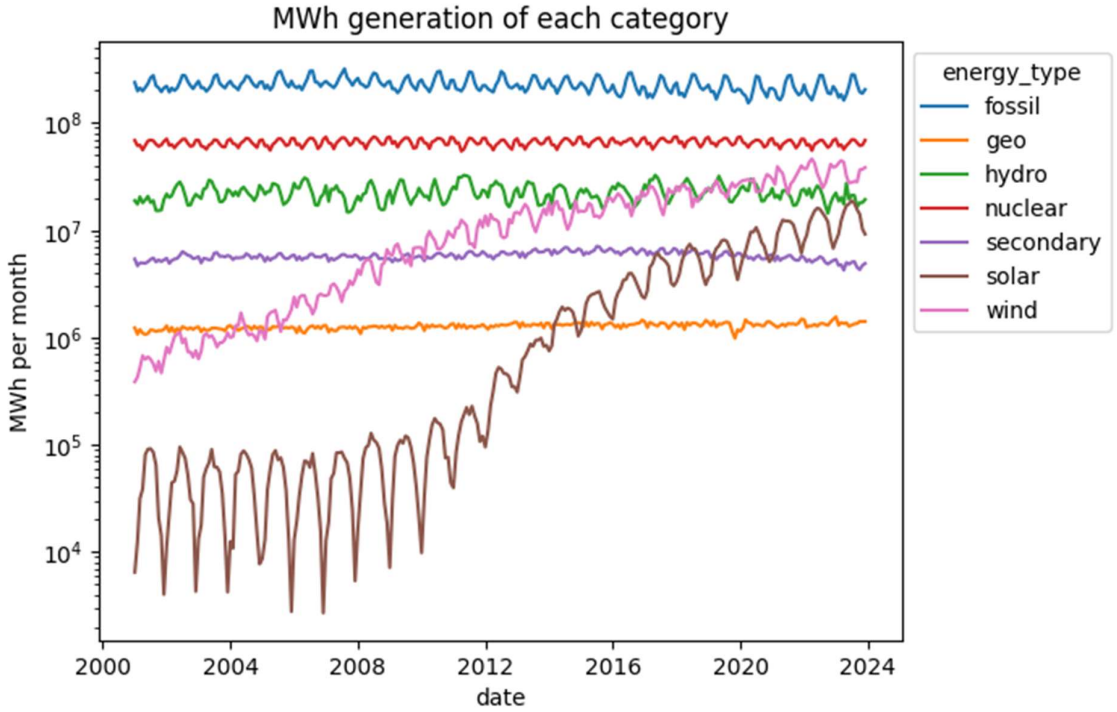

Figure 3: MWh generation per energy category

Fossil Fuels still make up a bulk of the energy generated, with over 60% of the electricity generated in 2023 by fossil fuels. The electricity generation by nuclear is stable throughout this period, at around 0.7GWh produced per year.

The biggest increase has been in wind energy - in the early 2000's it account for a negligible percentage of total electricity production, however by 2023 it has increased to a respectable 10% of total electricity production.

Solar energy is still lagging far behind in terms of electricity generation, however has been gradually increasing since 2015.

3. *Variation in electricity distribution for each category across the year*

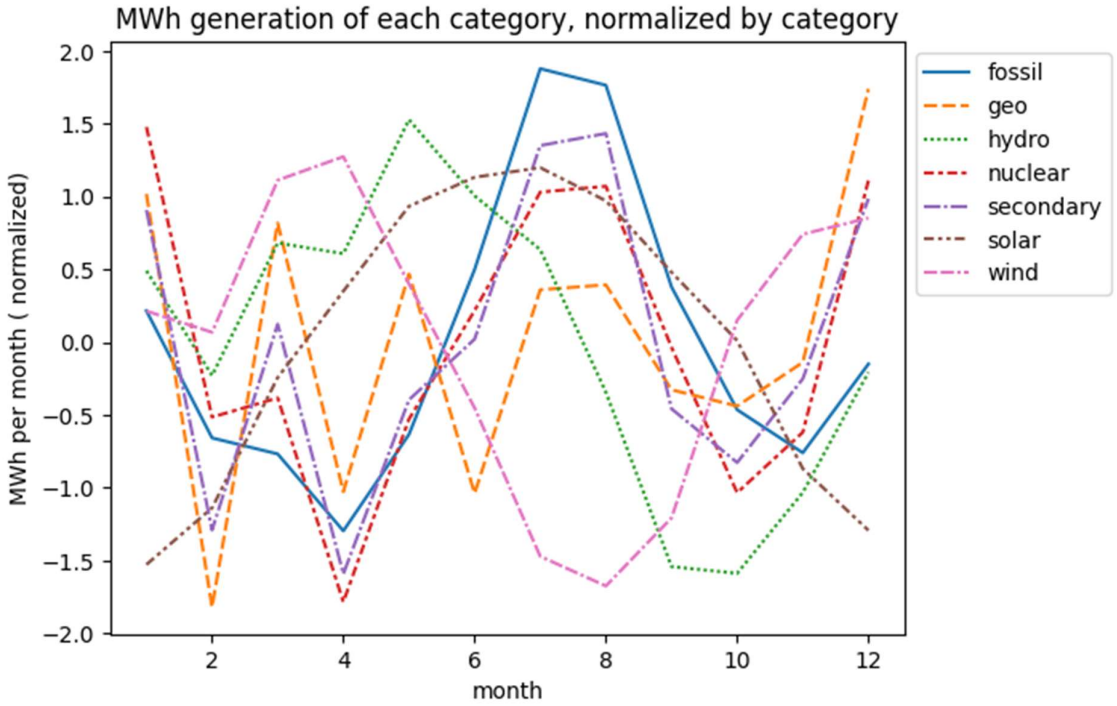

Figure 4: Electricity generation variance over a year, normalized

Most energy has seasonal variation in electricity production.

Fossil fuels and nuclear energy generation both show clear peaks during July and August. Compared to the other (green) energy generation methods - generation by these two categories can scaled up or down based on demand. It is likely that these two months represent peak electricity demand in United States.

Solar energy has a larger spread than the above two categories. It shows a peak in July and smoothly decreases in both directions from July. This corresponds to times when the Sun is over the Tropic of Cancer, due to longer daylight times, as well as the fact that the angle the sunlight makes with United States is nearly perpendicular at this time.

Hydro Energy follows a similar pattern to solar, but with its peak shifted to May. This corresponds to the fact that April is generally the month which brings a lot of rains through the country.

Wind Energy shows a bimodal distribution, with a peak in April and again in December.

4. *Change in electricity generated per capita over time*

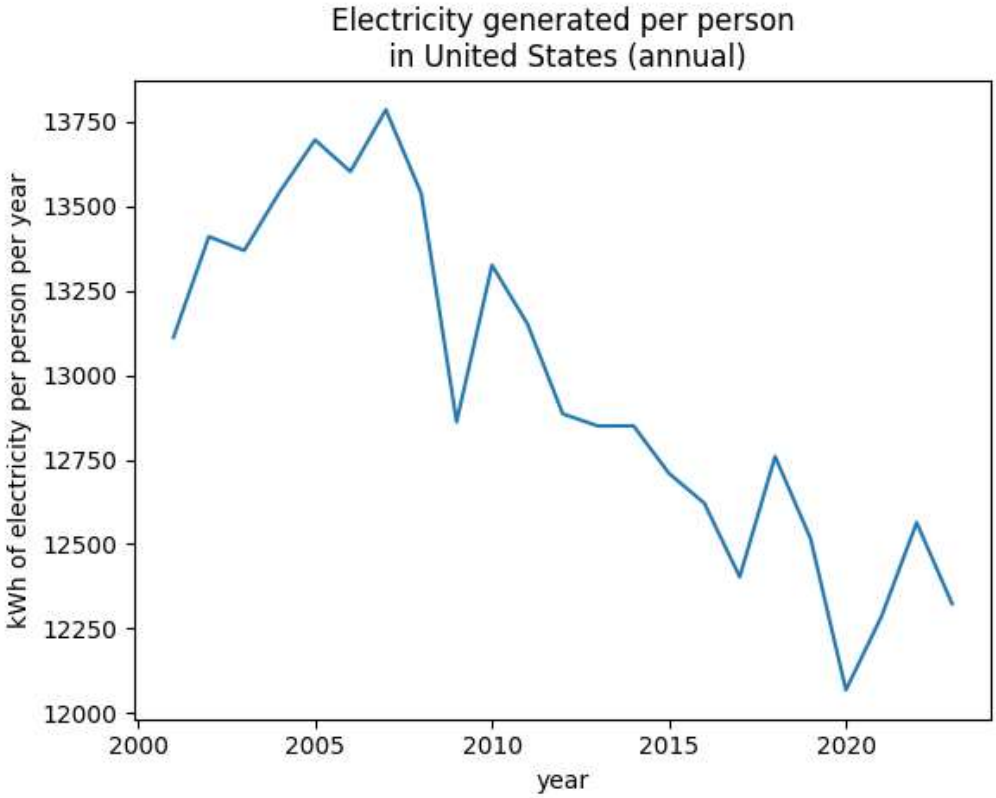

Figure 5: Electricity generation per capita

The electricity consumed per capita was between 13,000 and 13,250 units per person in around 2001 and peaked in 2006 at ~13,750 units per person per year. Since then it has steadily fallen, reaching ~12,250 units per person per year in 2023.

There are exaggerated dips in 2008 (corresponding to the 2008 Financial Crisis) and again in 2020 (corresponding to the Covid-19 pandemic). These would likely be caused to drops in industrial and commercial demands rather than residential electricity demands, however further analysis would need to be done to ascertain this claim.

5. *Change in electricity generation for each state from early 2000's to early 2020's*

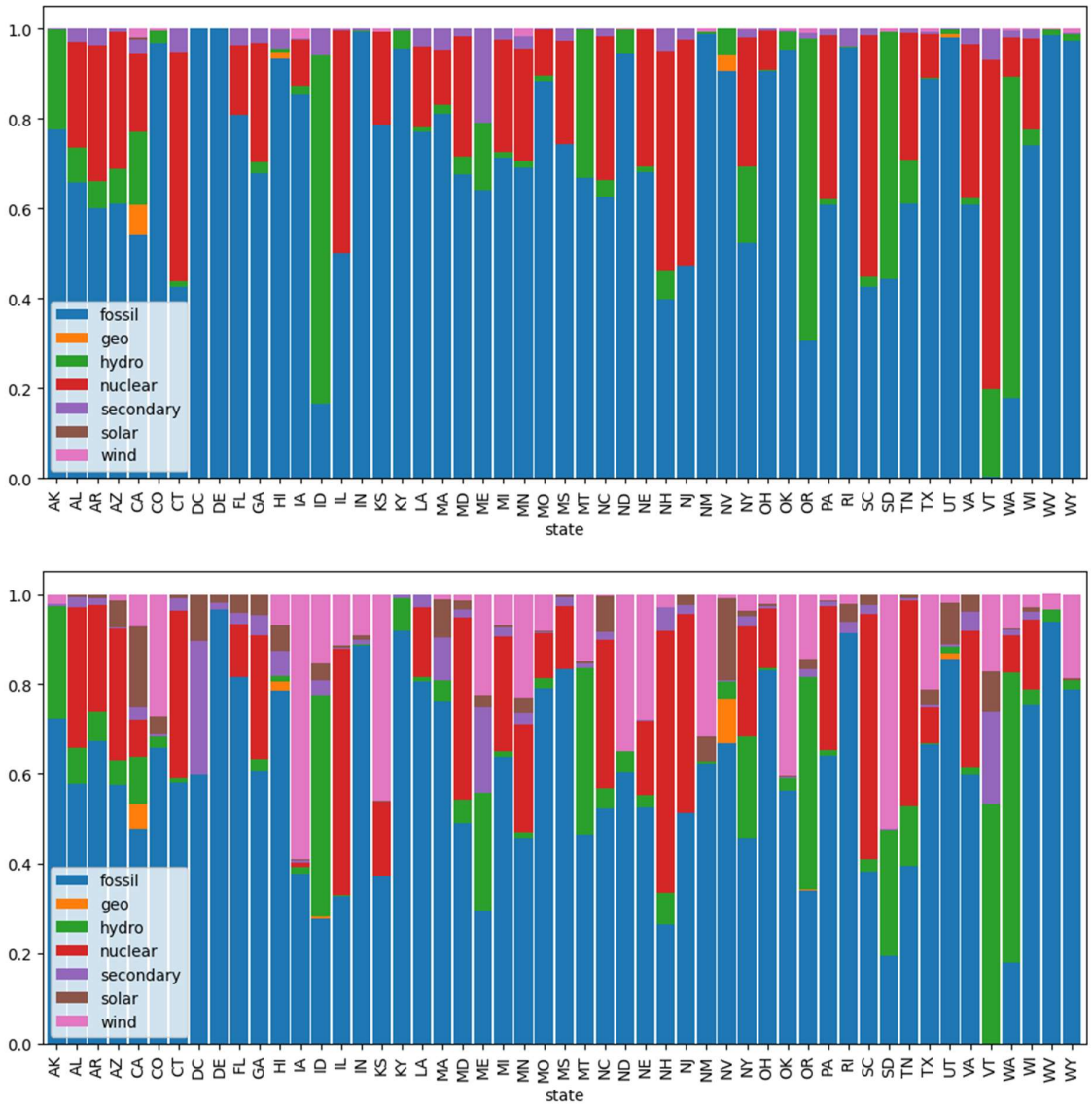

Figure 6: Share of electricity generated by each category, 2000-2005 mean v/s 2020-2023 mean

The biggest change since the early 2000's to the early 2020's is the massive adoption of wind energy across the states, noticeably with it accounting for around half of the demand in Iowa, Kansas and South Dakota. Most states still average around half of their electricity generation from fossil fuels with it accounting for > 90% of the generation in some states (e.g. Delaware, Wyoming, Utah).

Solar adoption is noticeable in many states, however in most of these it contributes very less to the total energy production. By percentage, California and Nevada have the largest percentage of solar energy production, where it accounts for ~20% of the total electricity produced.

6. *Comparison of cost of electricity for each state*

All sectors Residential Consumers

Rank of each State	2000	2010	2020
AK	45	50	48
CA	44	44	50
CT	49	49	49
HI	51	51	51
ID	3	2	2
KY	1,	7	14
MA	48	48	47
ND	7	11	3
NY	50	45	43
WA	8	1	7
WY	2	3	1

Rank of each State	2000	2010	2020
AK	46	49	46
CA	42	43	48
CT	49	50	49
HI	51	51	51
ID	1	3	2
KY	3	7	15
LA	19	2	6
MA	48	47	50
NY	50	48	43
UT	9	8	3
WA	4	1	1
WV	2	11	20

Figure 7: Ranks of each state wrt cost of electricity for (i) all sectors averaged and (ii) residential customers

Overall, Idaho and Wyoming have the lowest costs of electricity per unit across the decades. Kentucky had the lowest average rates in 2000-2010, however has fallen into the middle of the pack by 2020. Massachusetts, Connecticut, California, Alaska and Hawaii consistently have the highest costs for electricity throughout the 2000's.

The states with the lowest domestic costs are Idaho, West Virginia, and Washington. Kentucky had the lowest prices in 2000-2010 but has since had its rank increased. Interestingly Loisiana has reduced its rank considerably, from 19 in 2000-2010 to 6 in the 2020's.

Massachusetts, Connecticut, California, New York and Hawaii consistently have the highests costs for electricity throughout the 2000's. California fares better here compares to overall statistics, possibly due to higher costs to commercial/industrial sector as compared to residential sector.

ML MODEL

Implementation

The regression model use is a Multi-Output Linear Regression model, where each output corresponds to one label (i.e. *commercial, industrial, residential or all_sectors'* cost per unit of electricity). The model outputs the estimated cost per unit of electricity for each label give a vector of inputs, i.e. US State, population of state in given month, and electricity production in MWh of each category of generators. The linear regression model does not have hyper-parameters to tune.

Liner regression by default outputs a single estimate at a time; to convert it into a multi-output regression model, it must be wrapped around the scikit-learn *MultiOutputRegressor* module. Under the hood, this instantiates as many parallel linear regressors as there are labels, and trains each model individually.

Refinement

Multiple alternate regression models were experimented with, including Ridge Regression, Lasso Regression, Stochastic Gradient Descent Regression. The Linear Regression was chosen as the one with the greatest R2 score and minimum RMSE error across categories.

The original implementation of the Linear Regressor included data normalization followed by fitting / estimating. However, considering that a Linear regressor should be independent of linear modifications to input variables, such a transform before feeding to the model should have been irrelevant; indeed, the R2 score with and without the transform step were comparable. And so, the final model excludes data normalization for the regressor.

The original implementation of the linear regressor did not include inflation normalization. Without price normalization the model had an R2 score of 0.798

MODEL EVALUATION AND JUSTIFICATION

The final trained model has an R2 Score of 0.868.

The RMSE across the different categories are:

Consumer	RMSE in cents/kWh
all sectors	1.192
commercial	1.193
industrial	1.287
residential	1.398

Figure 8: RMSE scores for each consumer category

On validating against the entire dataset, the below error-rates were observed:

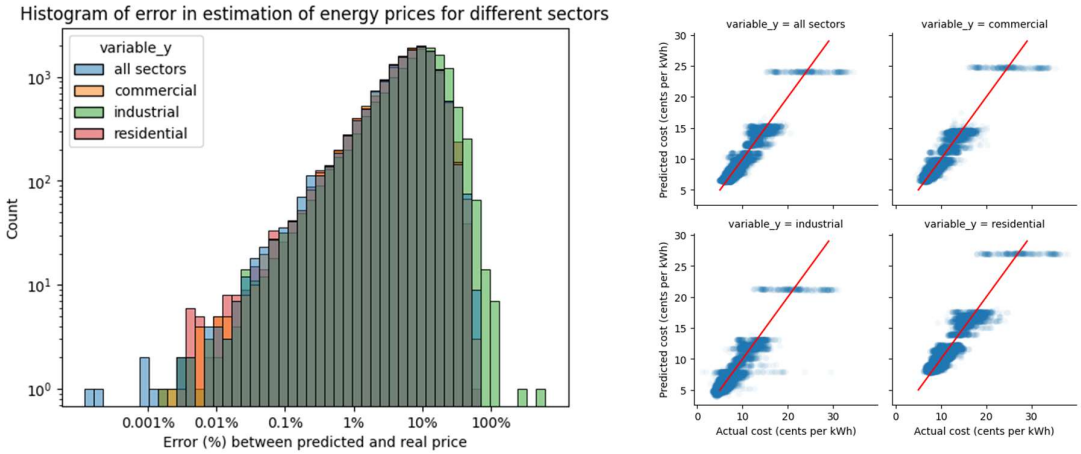

Figure 9: Distribution of errors in estimates (log scale count). (ii) Actual v/s Predicted costs by category

Most predictions lie within 1-10% of the actual value. Industrial predictions have a higher proportion of predictions with errors > 10% with a handful cases crossing 100%.

The model does a really good job estimating for prices <= 15 cents/unit, however fails for prices above this, where it consistently predicts an approximately constant value. This trend is seen across all categories.

CONCLUSION

User-Input Validation

A simple web-interface was created to demonstrate user-input in estimation of energy costs:

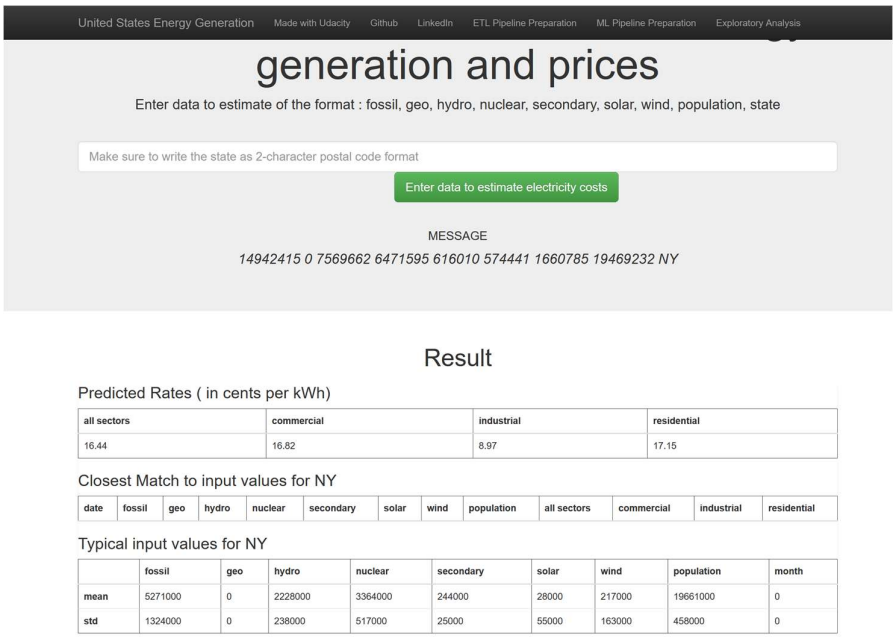

Figure 10: Sample webapp screenshot showing prediction of prices based on user input

The data for the query provided was taken from Jan-2024 energy generation figures, which the model had not encountered yet. The predicted rates were as shown in the image above, while the actual rates for New York for Jan-2024 were as below (inflation-adjusted):

all sectors	commercial	Industrial	residential
13.11	12.54	5.24	16.28

Figure 11: Actual rates for comparison to predicted prices (inflation-adjusted)

The rates for industrial was far removed from the actual data (which we could expect given that estimation for industrial prices frequently had high error values). Residential rates were the most accurate, with an error of ~5%.

Reflection and Improvements

The main challenge in this was to collect the data. The electricity generation data was divided across >20 files, with multiple different formats causing issues in the programmatic extraction of the relevant data. Another major problem was to figuring out further improvements which was a step-by-step process and took multiple iterations; for example, inflation-adjusting price data, and adding population statistics to make the datapoints as time-agnostic as possible.

One major point that our model excludes is capturing information about how the energy is being consumed by the different categories of consumers. I have compensated for growth in population as a placeholder for increase in demands from residential sector, however to fine-tune the model we would need information about energy consumption by commercial and industrial sectors by state as well.

As another point, the data ignores the effect of energy storage, which will play a bigger role in the energy grid as the share of green energies supplying the electricity grid increase.

Finally, the model selected for training (Linear Regression) can be updated for further improvements to the prediction engine. As a future improvement, I can try a different continuous regression model e.g. Random Forest Regressors.

TABLE OF FIGURES

Figure 1: Electricity Generation per month 3
Figure 2: Electricity production divided by generation category 4
Figure 3: MWh generation per energy category 4
Figure 4: Electricity generation variance over a year, normalized 5
Figure 5: Electricity generation per capita 6
Figure 6: Share of electricity generated by each category, 2000-2005 mean v/s 2020-2023 mean 7
Figure 7: Ranks of each state wrt cost of electricity for (i) all sectors averaged and (ii) residential customers 8
Figure 8: RMSE scores for each consumer category 9
Figure 9: Distribution of errors in estimates (log scale count). (ii) Actual v/s Predicted costs by category 9
Figure 10: Sample webapp screenshot showing prediction of prices based on user input 10
Figure 11: Actual rates for comparison to predicted prices (inflation-adjusted) 10