Business Problem:

A solar power company from out of town is planning on creating a location in Mesa, Arizona. To do this they need to show their stakeholders that there is plenty of business opportunity for them in this new location. The company will need models showing electrical usage over time and solar usage overtime. If solar electric is increasing this may mean that there is a want for solar power. Therefore, the company wants to know, is energy usage expected to increase? How is energy usage for properties with solar power changing overtime? Are there any trends where electrical usage increases? What would be the best model to show energy consumption forecasting?

Background/History:

Everything seems to be increasing in price these days and the cost of electricity is no different. 12 News reports an increase up to 39% for the cost of electricity as of December 2024. Currently, the average cost of electricity is 15 cents per kWh leading the average family to have an electric bill of about $220 a month (Parker & Sons, 2025). If the cost for solar power comes in lower than the cost of electricity per month while covering all the consumers electric needs, then they may want to consider switching. Especially, since Arizona wants to increase their renewal energy usage by 15% by the end of 2025 (Parker & Sons, 2025). Based on the Work Population Report the population of Mesa has been increasing significantly over the past couple of years. With this information it seems reasonable that a new solar company moving into the area would do well. Also, as shown below, the use of solar power has still not been fully adopted.

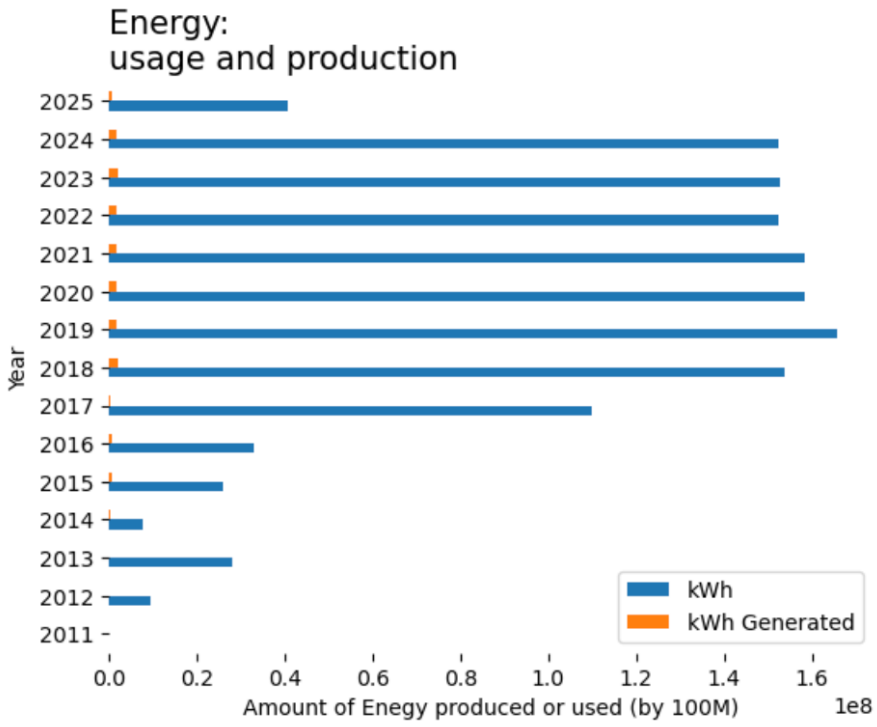


Fig 1: Energy Usage versus the amount pf energy produced by solar power.

Data Explanation:

The Data set I've obtained is from the City of Mesa website, which provides open data to the public. The dataset can be pulled from the link below:

https://data.mesaaz.gov/Environmental-and-Sustainability/City-Energy-Usage/ksen-g4gs/about_data

This data set consists of 22 columns:

Row ID: This is a unique row identifier.

Site Type: This is the general name for each site.

Site Sub Type: This is a name given to the sites that's more specific to what they are.

Site: This is the actual name or description of the site.

Category: This is the electrical provider. Either the City of Mesa (COM) or Salt River Project (SRP).

Address: The address to the site.

Latitude: The latitude of the site.

Longitude: The longitude of the site.

Geolocation: The geolocation of the site.

Premises Code Account Number: The account number for the site with either COM or SRP.

EMS Flag: Indicates if the site has an emergency management system (EMS).

EMS Install Date: The date that the EMS system was installed.

Month Date: The month and year of the billing cycle.

Year: The year of the billing cycle.

Month: The month of the billing cycle.

kWh: The number of kilowatt-Hours used.

kWh Generated: The amount of kWh generated by solar panels.

kW: The maximum usage of kWh in period of an hour.

Solar Flag: Indicates if the location generates solar electricity.

Solar Install Date: The day the solar panels were installed.

Full Month: Indicates if the site was using electricity for the full month.

Site Type Simplified: A simplified site name.

For the initial pass of data processing, I dropped multiple columns, and I only kept 'kWh', 'Month Date', 'Solar Flag'. Next the data frame was split into two as energy consumption with solar and without out. At this point I checked for missing values and since the only missing values were in the 'kWh' column, I decided to drop the rows. Next, I took the mean for each date of kWh used

for that month and conducted a time conversion of the date. I then made box plots of the data frames to analyze for outliers. I then removed the outliers using quartile ranges. Following this I made histograms of the data. The non solar data set was left skewed, and the solar data set had many values about the mean. To fix these I applied a Lamba transformation to the data and replaced the 'kWh' columns with this data. Finally, I set the date as the index.

Methods:

As we're working with time series modeling for electrical usage in the desert, I'm already assuming that seasonality will be present in the model. While I've tried multiple models, the two with the best results were Facebook's Profit as well as XGBoost regression with elastic net features. I chose the XGBoost model since this is a large dataset consisting of 135,710 rows of data and XGBoost is made to handle large datasets. XGBoost is also meant to have high accuracy and performance. This will be important to assist the solar companies pitch to the stake holders. I chose to use profit because most other methods for modeling time series were failing for me, but profit is supposed to be more intuitive to use.

Analysis:

Both Models were run twice, once on the energy usage with solar panels and once for energy usage without solar panels. The Profit models were difficult to interpret. However, overall, this model shows an increase in energy usage for buildings with Solar power and a slight decrease for buildings without solar usage. Also, these models had a higher, MAE MSE, RMSE than the XGBoost models. This indicates that the model wasn't performing as well and had high error. This model also appeared to be overfit, and the increases may have been exaggerated in the model.

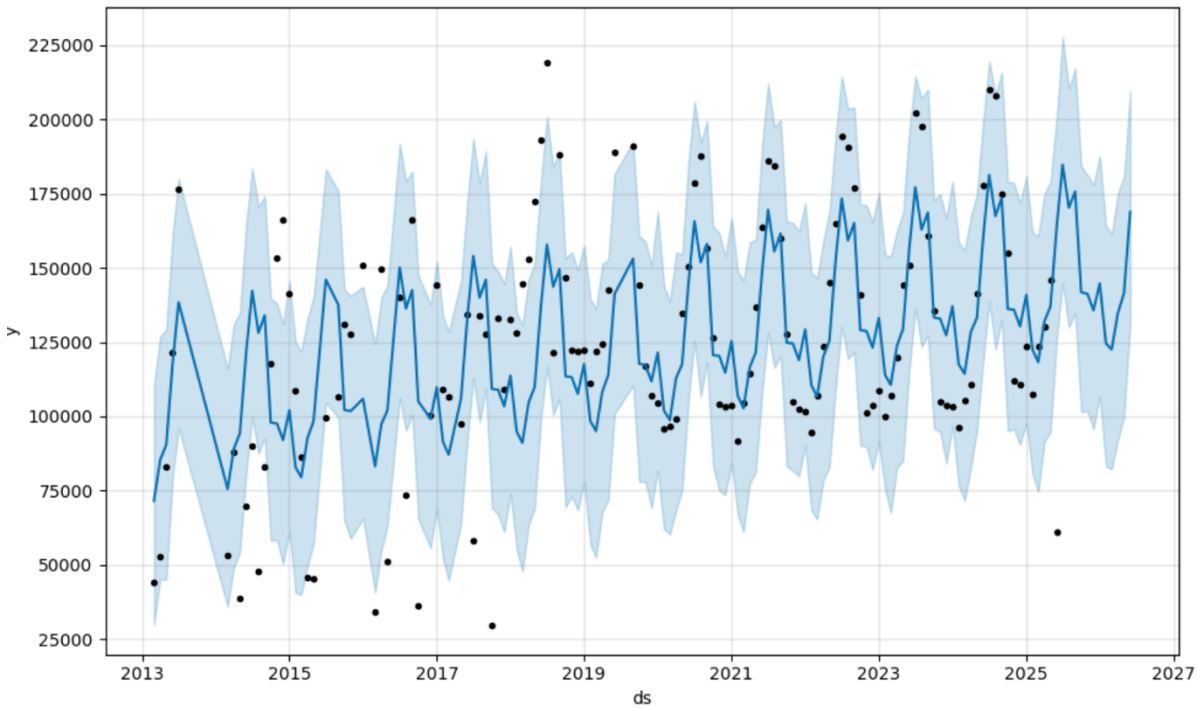


Fig2: Facebook Profit model for energy use with solar power.

The XGBoost model performed well, but I added elastic net features to prevent overfitting. This model fitted well to the seasonality changes but underperformed slightly for the summertime increase. However, this can be attributed to external factors such as increase in temperatures, changes in population and new infrastructure/businesses. This model had an MAE of 15, 526.35, MSE of 20,825.82 and a MAPE of 10.83% for the buildings with solar, The MAPE is in a good range for the model. The other values may seem large but are perfectly acceptable as the data ranges upwards to 225,000 kWh.

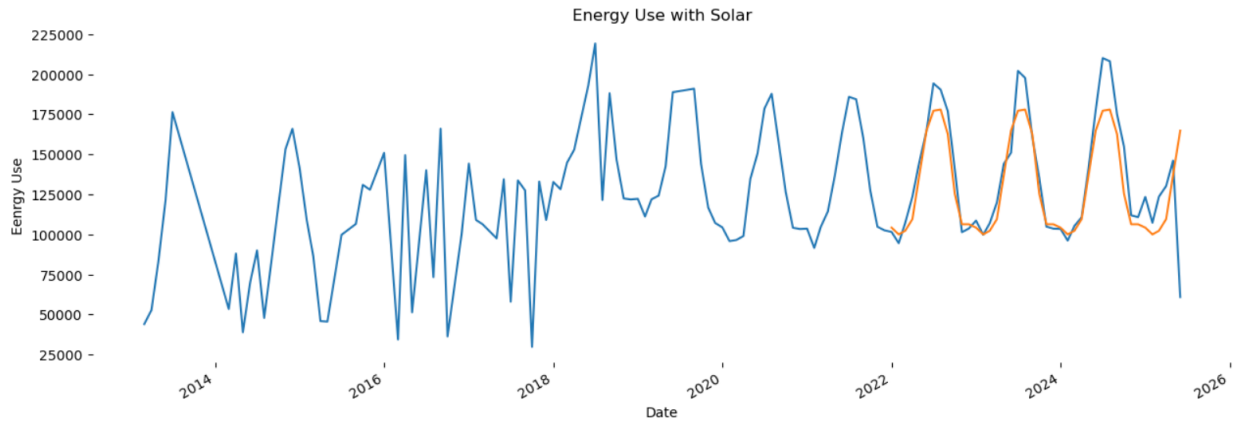


Fig 3: XGBOOST model for buildings energy use with solar power.

The model for the non-solar energy consumption had a higher accuracy close to the solar model for the MAE and MSE. The MAPE for this model was around 12%

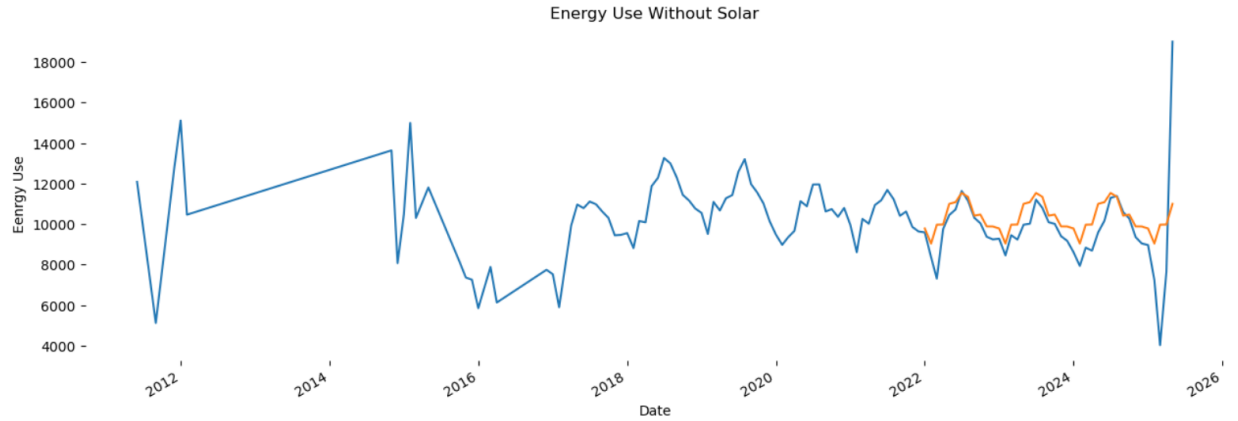


Fig 4: XGBOOST model for buildings energy use without solar power.

Conclusion:

With these models we show that there is an increase in energy consumption for buildings with and without solar production. There is a slow increase in buildings with solar usage which can imply that people are converting to solar to save money on their energy bill. We do see a trend in energy usage especially in the summertime. This is most likely due to the increased temperatures, the energy usage then drops during the winter months. The best model to use for energy forecasting in our scenario is the XGBoost regression model. This model is easier to interpret for

those who may not be experienced in interpreting models. In all, this data shows that there is a growing market demand to switch to solar power.

Assumptions:

I assumed that the data contains residential properties as well as commercial businesses. This stemmed from conducting unique searches on the 'Site' column and returning a line for residential property. I had also assumed that the XGBoost model for properties with solar was overfitting, to fix this I employed the best model method to obtain elastic net features to prevent the model from overfitting. The last assumption I made is that I only needed the "Date" and "kWh" columns to model the data as the rest of the columns didn't seem like it would attribute much to the model.

Limitations:

There were some data quality issues that may have affected the model performance. This was that there missing or unreported values in the kWh used column. I also have no previous experience using Profit so, the only result I can really pull from this model would be the predicted increase in energy consumption.

Challenges:

I've used multiple methods to model the data, but the results were so poor that I removed them from the code and moved onto the next model. These would consist of rolling means, SVR, LTSM, ARIMA and SARIMA. The SVR, ARIMA and SARIMA models weren't capturing the seasonality and were just creating straight lines for predictions as a constant value. Rolling means and LTSM results had a higher error rate than the XGBoost model. Also, when the predictions were plotted against the actual values, the predictions were significantly higher than the expected values.

Future Uses/Additional Applications:

Other companies can use this data to see if the current electrical infrastructure can handle the loads that they would require to run their business. Also, the city itself can use these models to monitor their infrastructure to make sure that it can handle the needs of the people and plan upgrades if needed. The city can also use these models to plan for maintenance. They can predict the time frame with the least energy draw to plan for this maintenance.

Recommendations:

The company should complete some research into reasons why the use of solar power isn't growing faster. A concern would be if HOAs have restrictions on homes getting solar panels. Model performance might be improved if we add three more features. This being population changes, weather patterns, and if we had some information on new business locations. Adding this might contribute to an increase in accuracy for the summertime energy use.

Implementation Plan:

Any new data obtained would need to be cleaned, removing any columns that may contain personal identifying information (PII). Once cleaned this should go through the same transformations that the data for this model went through like removing unnecessary columns, removing outliers, transforming and splitting the data. The data will need to be stored properly and secure as it may contain PII. Preferably, we would need to ensure this information doesn't get saved or is anonymized. Since the data only becomes available once a month an analyst may need to pull the new data and update the model appropriately. Once updated the analyst can complete further exploratory data analysis to locate other items of interest for the company.

Ethical Assessment:

The dataset appeared to contain PII. This data could be primarily for business, but to be safe it's better to remove this information so that it's not leaked out. The data itself shouldn't contain bias since it was localized. If the data for energy consumption was broader and more expanded to the entire state, then it may contain some societal bias. In my opinion, the prophet model isn't very transparent. Yes, it does show an upward trend, but there are some aspects to the plot that I'm not familiar with like the black dots. With this model it could create a misrepresentation of the model results. Since the data is being provided publicly, I did read the policy on its use, and we do have free use of the data to use and share.

References

Mesa, Arizona population 2025. (2025). Retrieved from https://worldpopulationreview.com/us-cities/arizona/mesa

Parker & Sons, P. &. (2025). Understanding the Rising Cost of Today's Electrical Prices. Retrieved from https://www.parkerandsons.com/blog/understanding-the-rising-cost-of-of-todays-electrical-prices

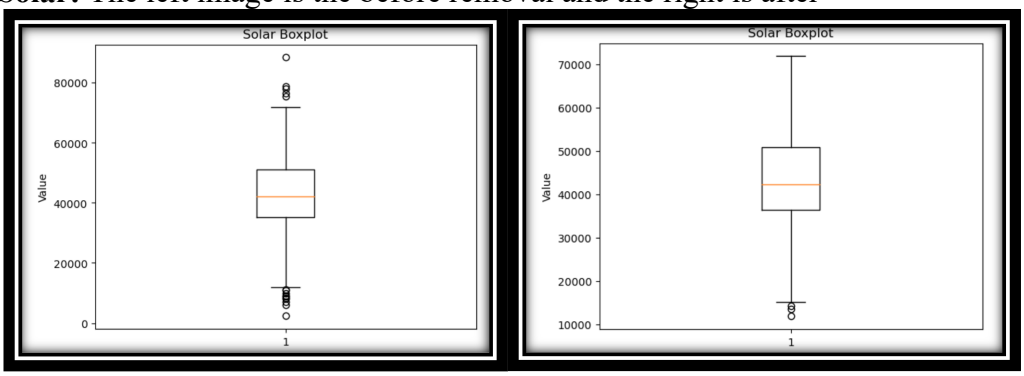
Reagan, K. (2024). Mesa City Council pass utility rate hike despite public objections | 12news.com. Retrieved from https://www.12news.com/article/news/local/valley/mesa-city-council-pass-utility-rate-hike-despite-objections-from-residents-arizona-mayor-john-giles/75-f80084ec-e6bf-4584-9704-618f764ad63f

Sustainability, E. and. (2025). City Energy Usage: City of Mesa Data Hub. Retrieved from https://data.mesaaz.gov/Environmental-and-Sustainability/City-Energy-Usage/ksen-g4gs/about_data

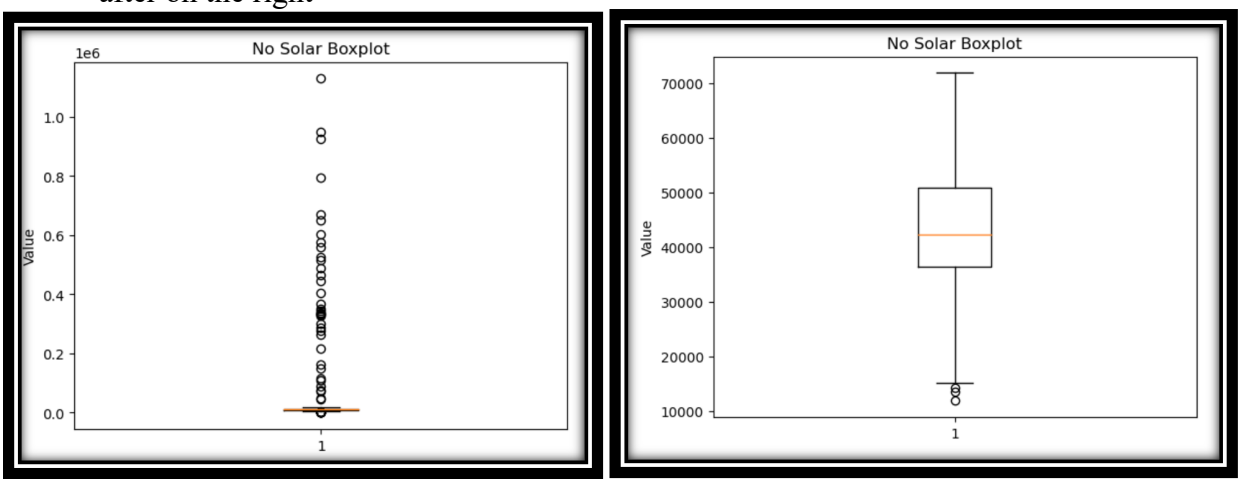
Appendix

Before and after box plot for removing outliers in the data:

Solar: The left image is the before removal and the right is after

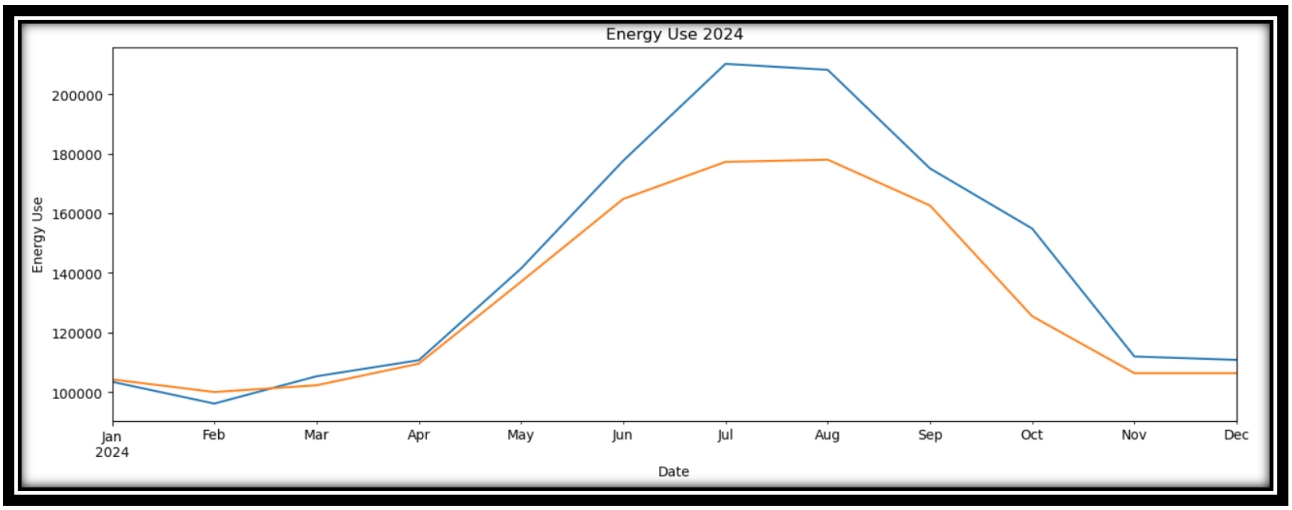


No solar: There appears to be no change due to the size of the data. The before is on the left and after on the right



XGBoost regression model:

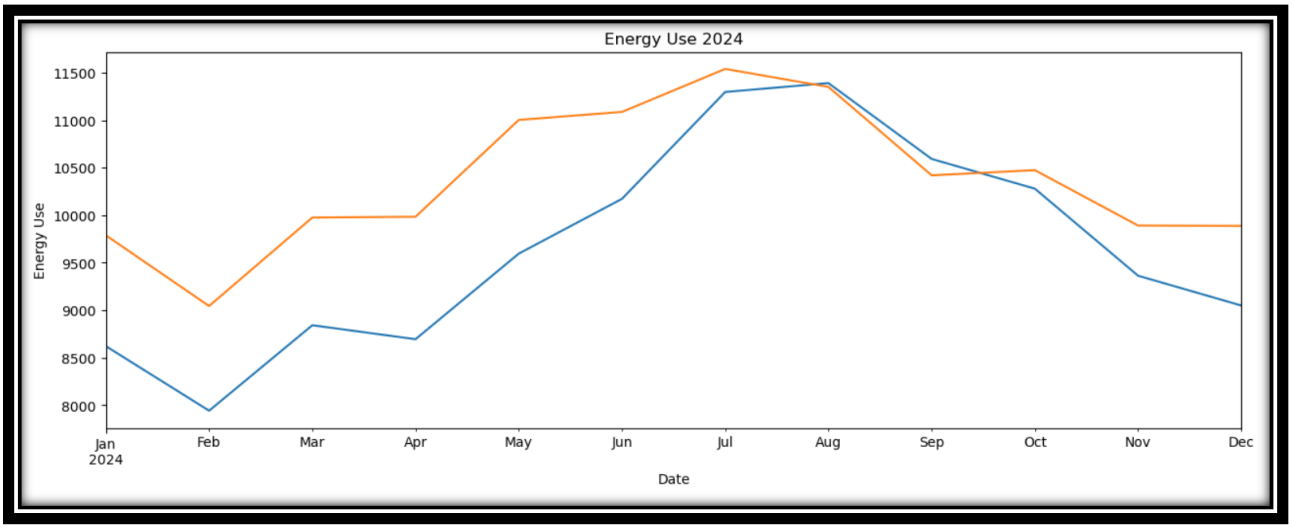
One year of predictions for energy usage with Solar:



Solar model accuracy:

```
RMSE: 20825.818737943897
MAE: 12526.350410508907
MAPE: 0.10826551421502704
```

One year of predictions for energy usage with without solar:

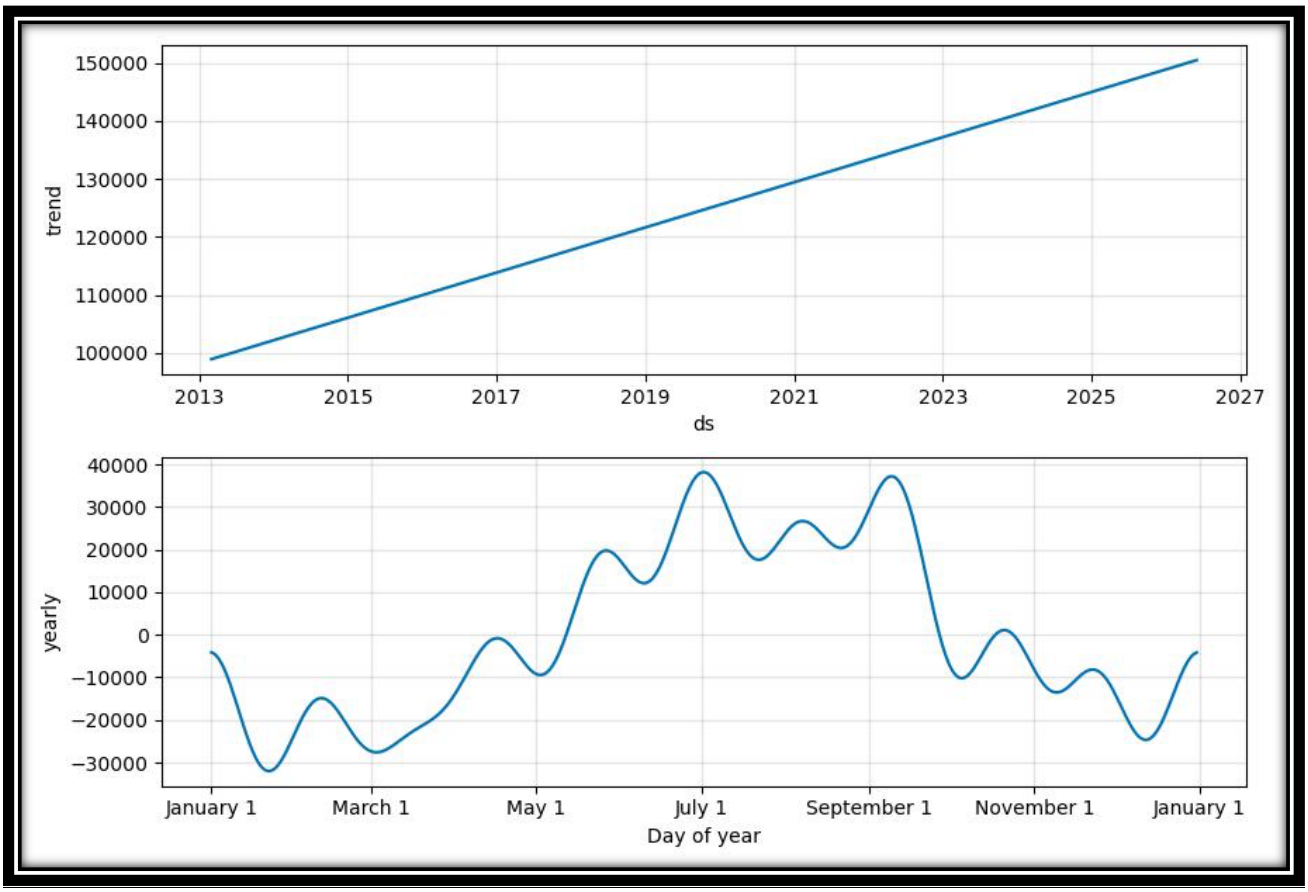


No Solar model accuracy:

```
RMSE: 1795.4131174629679
MAE: 1020.5826200976732
MAPE: 0.12353823919799298
```

Profit model:

Trend and year view for energy usage with solar power:

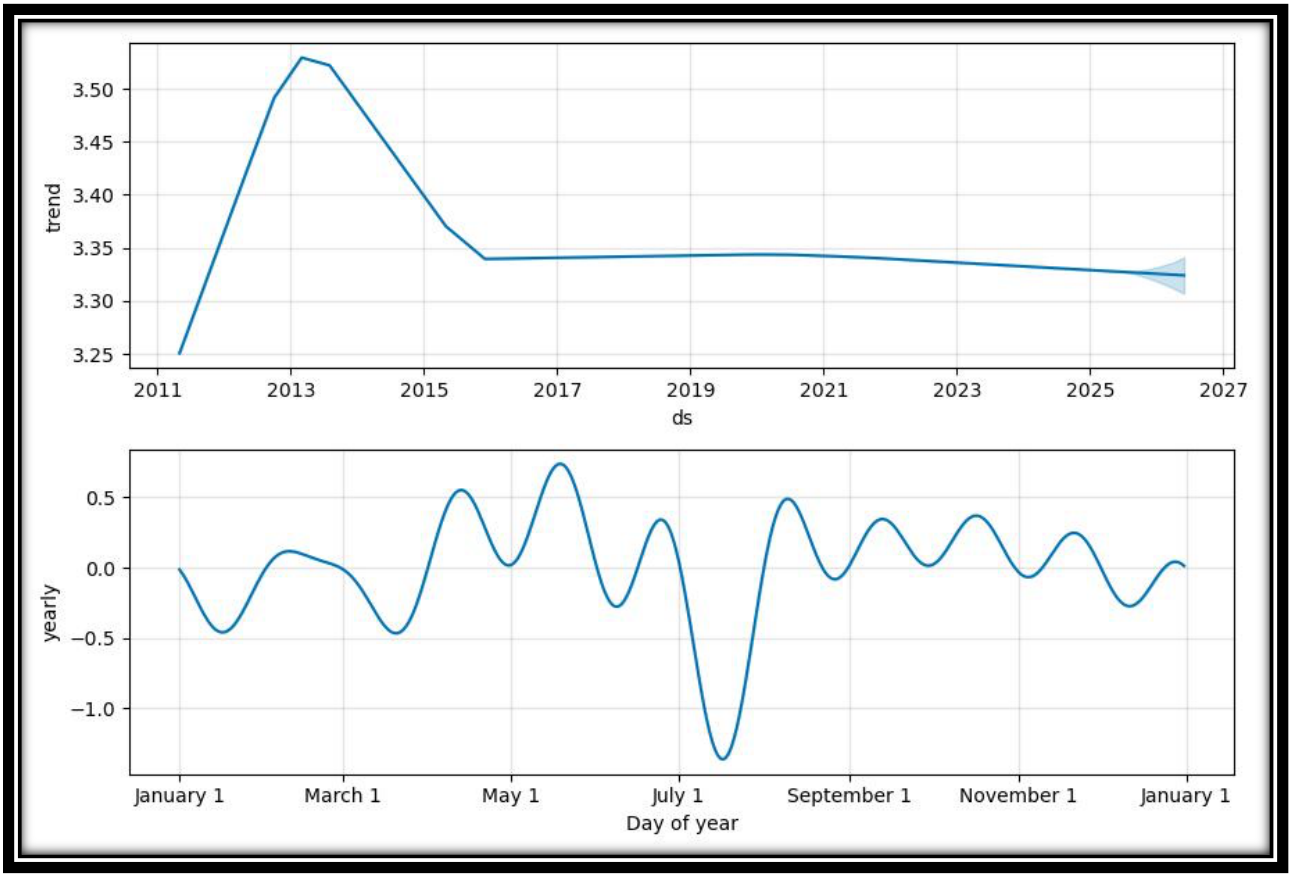


Profit model accuracy results:

	<bound method NDFrame.head of	horizon	mse	rmse	mae	mape	mdape	\
0	44 days	2.991062e+09	54690.603506	43938.129289	0.414531	0.312061		
1	45 days	2.982698e+09	54614.086035	43846.659535	0.407903	0.269462		
2	46 days	2.944785e+09	54265.876040	43541.533054	0.459668	0.269462		
3	49 days	3.021773e+09	54970.653746	45397.473210	0.479595	0.312061		
4	50 days	3.713868e+09	60941.509541	50089.796632	0.492739	0.312061		
..	...	...	...	...	...	...	...	
161	358 days	1.889674e+09	43470.383914	34817.936644	0.299717	0.307832		
162	360 days	1.802595e+09	42456.981652	32818.228955	0.280507	0.303019		
163	362 days	1.634033e+09	40423.179510	30781.973365	0.272817	0.163697		
164	363 days	1.673177e+09	40904.490810	32060.601973	0.284531	0.259293		
165	365 days	2.325008e+09	48218.340733	36897.809130	0.370415	0.303019		

	smape	coverage
0	0.371147	0.380952
1	0.369031	0.380952
2	0.441575	0.380952
3	0.457949	0.380952
4	0.508001	0.333333
..	...	...
161	0.261014	0.571429
162	0.245093	0.619048
163	0.233807	0.666667
164	0.244099	0.666667
165	0.284921	0.619048

Trend and year view for energy usage with no solar power:



Profit model accuracy results:

```
<bound method NDFrame.head of      horizon       mse      rmse       mae      mape     mdape     smape  \
0     35 days  0.007458  0.086358  0.064955  0.019197  0.018595  0.019203
1     38 days  0.007529  0.086773  0.066195  0.019569  0.018595  0.019579
2     39 days  0.007038  0.083892  0.064176  0.018952  0.018595  0.018989
3     40 days  0.006451  0.080318  0.061595  0.018217  0.014510  0.018268
4     44 days  0.005788  0.076076  0.058300  0.017454  0.014510  0.017417
..        ...       ...       ...       ...       ...       ...       ...
191 358 days  0.011881  0.109000  0.071884  0.021205  0.013041  0.021486
192 360 days  0.011809  0.108667  0.071017  0.020946  0.012784  0.021223
193 362 days  0.011737  0.108337  0.070170  0.020694  0.012784  0.020968
194 363 days  0.012322  0.111004  0.072820  0.021476  0.012784  0.021725
195 365 days  0.013796  0.117458  0.078762  0.023382  0.012784  0.023559

     coverage
0    0.592593
1    0.592593
2    0.592593
3    0.648148
4    0.666667
..        ...
191  0.851852
192  0.851852
193  0.851852
194  0.814815
195  0.777778
```