



Energy Consumption and Generation Prediction Based on Weather Conditions in Spain

Anatasia Kurakova

Alexandra Nguyen

Sabnam Pandit

Yu-Chieh (Jason) Tu

Wanjing (Anna) Yan

San Diego State University

BDA 602: Machine Learning Engineering

Abstract

One of the main questions of today's world is how to reduce our carbon footprint and preserve the planet for future generations. One of the factors that has an impact on modern ecology is fossil fuel extraction and processing. Renewable energy generation can be a cheaper and greener alternative that will help reduce carbon emissions. Our team analyzed the Kaggle dataset *Hourly energy demand generation and weather*, which compiled weather and temporal data from Spain. We implemented machine learning algorithms to predict the consumption and generation of renewable energy, along with the energy prices based on the weather and historical data. In addition, we wanted to compare the results with nonrenewable energy generation. We were able to achieve high model performance in our prediction using a combination of weather and temporal factors. At the same time, we attempted to predict total renewable energy generation based only on the historical data and observed that using time data alone is insufficient for a high-quality prediction.

Energy Consumption and Generation Prediction Based on Weather Conditions in Spain

The movement towards a greener planet has become a crucial part of world development in the past decades. With climate change accepted and recognized, researchers have started looking for solutions that reduce nations' carbon footprint. One of the most effective solutions for reducing carbon emissions is switching to renewable sources of energy (Pata, 2021). In recent years, renewable energy use has become more widespread and affordable than ever before, with a 14-times increase in wind generation and 700-times increase in solar generation across European Union countries alone from 2010 to 2019 (Kolosok, Saher, Kovalenko, & Delibasic, 2022; Timilsina, 2021). However, there are some complications related to switching to renewable energy, including seasonality patterns that affect the amount of energy produced and the ability to store the energy during peak seasons for later use (Brey, 2021). These factors affect energy price fluctuations and have an impact on the energy market. The ability to predict how much renewable energy is going to be generated and how energy prices will be affected are invaluable sources of information for governments and policy makers that want to strategically plan further development (Ding and Dang, 2023). Our team explored the generation and consumption of renewable energy in Spain across a four-year span (2015-2018) to see which factors had the most profound impact on the amount of energy generated. In addition, we also explored energy prices of nonrenewable energy to study if they provided additional insights to our research.

Exploratory Data Analysis

Data Acquisition

We acquired the dataset [Hourly energy demand generation and weather](#) from Kaggle, which contains hourly weather data in Spain from 1/1/2015 to 12/31/2018.

Our data consists of two separate tables:

- 1) The *Energy* dataset contains 29 columns and 35075 rows. The data includes information on data generated by a variety of renewable and nonrenewable energy sources such as fossil fuels, solar, hydro, and nuclear energy. In addition, there is data on actual and forecasted consumption and energy prices.
- 2) The *Weather* dataset contains 17 columns and 178397 rows. The data includes hourly weather conditions in five cities in Spain: Barcelona, Madrid, Valencia, Bilbao, and Seville.

Descriptive Statistics

The total consumption of energy is highly correlated with *hour*, *previous_day_averages* and weather conditions such as *temperature* and *humidity*. There is almost zero correlation of pressure, snow, clouds, and rain with consumption (*total_actual_load*).

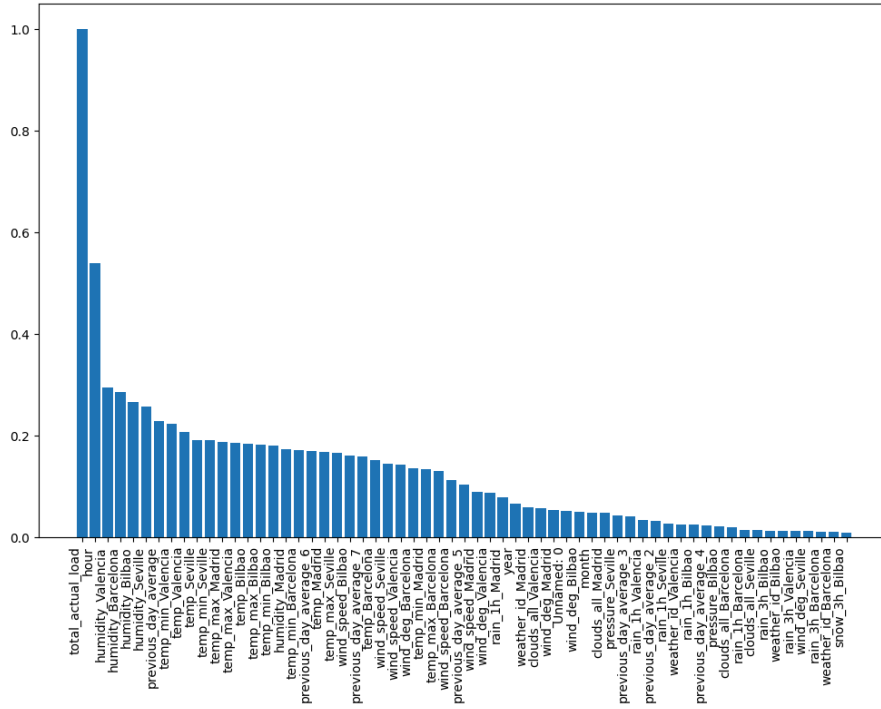


Figure 1. Correlation between Consumption and other variables

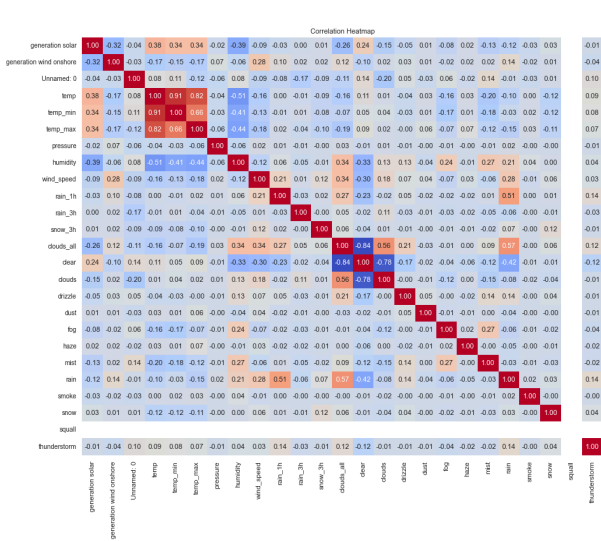


Figure 2. Weather Covariance Matrix

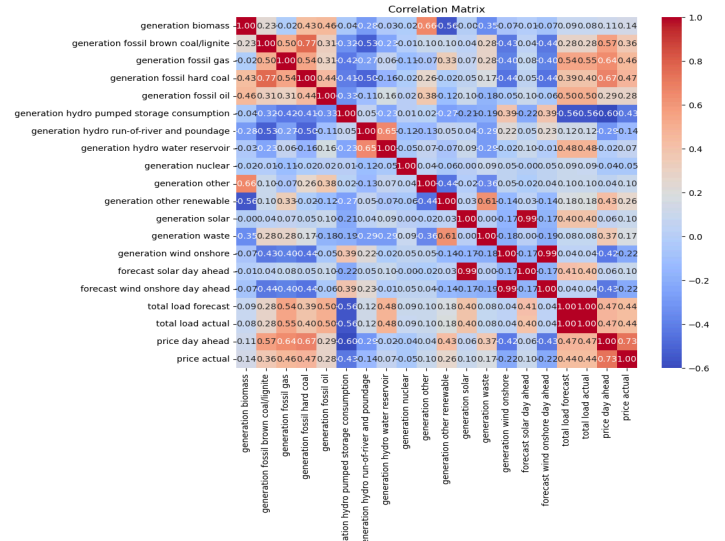


Figure 3. Energy Correlation Matrix

Understanding the intricate relationship between weather conditions and energy generation is paramount, particularly in the realm of renewable energy (Figure 2). A higher correlation coefficient of 0.38 between solar energy generation and temperature indicates that

solar generation tends to increase on warmer, sunnier days. Similarly with wind energy generation, a correlation coefficient of 0.28 with wind speed suggests a positive correlation, implying that wind energy generation correlates with wind speed. There are a few things that can be observed from the Energy Correlation Matrix (Figure 3). First, there is a positive correlation between prices and nonrenewable energy generation and negative correlation between prices and renewable energy generation, which indicates that, in general, renewable energy is cheaper for customers. A negative correlation can also be observed between renewable and nonrenewable energy generation, suggesting that as one type of energy generation decreases, the suppliers compensate with the other type.

Methodology

Data Pre-processing

The first step was to merge the two datasets based on date and time. The second step was to remove unused columns and check for missing values. Next, we tried several approaches, outlined below, to see which performed better in our analysis.

- 1) Grouped weather features and time across five cities to find out average weather conditions across the country
- 2) Merged two datasets without averaging the cities' weather data to see if weather conditions in one of the regions had a stronger influence on energy generation and consumption

Feature Engineering

Our original datasets were the *weather* and *energy* datasets, which were highly influenced by a temporal pattern. We had the date and time, but not the temporal change on features, which would show the influence or impact on energy consumption. We performed feature engineering with the temporal data to extract the previous day's average energy. Since our data contained hourly information, we averaged the previous day's consumption, which was used as an additional feature. Along with that, we extracted the average for the previous seven days as a new feature. The date/time stamp was converted to extract year, month, day, and hour, which helped demonstrate the correlation between variables.

Pipelines

Categorical Pipeline

The one-hot-encoding was performed on the categorical variables listed below to convert categorical data into one hot vector. Then, we created the categorical pipeline for seamless processing and passed the one-hot-encoded categorical features through the pipeline. For the categorical pipeline, we imported the packages *Pipeline* and *OneHotEncoder* from *sci-kit learn*.

- Consumption: weather descriptions and weather icons

- Renewable: *time, clear, clouds, drizzle, dust, fog, haze, mist, rain, smoke, snow, squall, thunderstorm*

Numerical Pipeline

Inside the numerical pipeline, we passed the numerical features listed below. For the numerical pipeline, we imported the packages *Pipeline*, *KNNImputer*, *MinMaxScaler*, and *Standard Scalar* from *sci-kit learn*.

- Consumption: *temp* and *humidity* using *KNNImputer* for interpolation and *MinMaxScaler* for standardization
- Renewable: *temp, pressure, humidity, wind_speed* using *StandardScalar* for standardization

Using the imported *ColumnTransformer* package, we combined the numerical and categorical pipelines to prepare them for training.

Libraries

Along with the libraries stated above, popular ones used were *pandas*, *seaborn*, *numpy*, and *matplotlib.pyplot*. In addition, *sklearn*'s *train_test_split*, *GridSearchCV*, *r2_score*, *mean_squared_error*, *mean_absolute_error* and models noted in the following section were imported.

Models

Our selected machine-learning models and their respective topics are listed below.

- *Gradient Boosting Regressor*: consumption and renewable energy sources
- *Elastic Net Regression*: consumption
- *Random Forest Regressor*: consumption, renewable energy sources, nonrenewable energy sources, and price prediction
- *Voting Regressor*: renewable energy sources
- *Linear Regression*: nonrenewable energy sources and price prediction
- *Extreme Gradient Boosting (XGB) Regressor*: nonrenewable energy sources and time series prediction
- *Support Vector Regression*: nonrenewable energy
- *Prophet*: time series prediction
- *SARIMAX*: time series prediction

Analysis

Consumption

To observe the influence of energy consumption based on weather conditions, we experimented with a few machine learning models. We trained linear regression models such as

Elastic Net, non-linear models such as *Random Forest* model, which comes from the *Bagging* method, and *Gradient Boosting* model from the ensemble method. *Elastic Net* was our baseline model. We split data into training and testing sets, the latter of which was 15% of the total dataset, and employed *scikit-learn* libraries to train the models.

We used *randomizedSearchCV* to achieve the best hyperparameters in the *Gradient Boosting* model. Our hyperparameters were *n_estimator* and *max_depth*. Similarly, for the *Random Forest* model, we fine-tuned the *n_estimator* parameter. The evaluation metric used to compare different methods was the R^2 score, which was agreed to be the most suitable metric for the comparison of regression models.

Renewable Energy Sources

A *train_test_split* was performed with the aggregation of the *weather* and *energy* data sets, focusing on solar generation. The dataset was split into an 80% training set and 20% testing set. The *Random Forest* model was chosen as the baseline model. The chosen hyperparameters to be tuned are *n_estimator* (50, 100, and 150) and *max_depth* (10, 20, and None) for the first attempt and *n_estimator* (100 and 200), *max_depth* (10 and 20), and *min_samples_split* (2 and 5) for the second attempt. The R^2 score and root mean squared error (*RMSE*) were compared to evaluate if a better model could be produced.

For visualizations, we analyzed solar generation with respect to temperature and by weekly and monthly averages.

Nonrenewable Energy Sources

To explore the relationship between weather and nonrenewable energy sources, we combined the *weather* and *energy* datasets to predict nonrenewable energy generation using simultaneous weather data. We utilized different models for prediction, including linear models such as *Linear Regression* and *Support Vector Machine Regression* and ensemble models such as *Random Forest Regression* and *Extreme Gradient Boosting Regression*. We examined two different approaches to split the training and testing datasets and found a significant difference. We used grid search for hypertuning and cross-validation to make the result more reliable.

Price Prediction

The *price day ahead* and *price actual* variables were dropped. The dataset was standardized using *StandardScalar* and a *train_test_split* was performed with 75% as the training set and 25% as the testing set. The baseline model was *Linear Regression*, which was compared to *Random Forest Regression*. For the latter, hyperparameter values were chosen as *n_estimators* (50, 100, and 200), *max_features* (auto and sqrt), and *max_depth* (10, 20, 30, and None). The best parameters were found to be 200, auto, and 10, respectively. The algorithms' performances were evaluated to one another by comparing their R^2 score and root mean squared error (*RMSE*).

Time Series Analysis

We attempted to predict the amount of generated renewable energy using time series analysis to see if making the prediction relying solely on historical data was possible. To perform the analysis, we used the *energy* dataset to create a new variable (*total renewable energy*) that combined energy generated using renewable energy sources and timestamp data. To preprocess the data, we removed empty/irrelevant columns and used *Iterative Imputer* to impute missing values. We utilized three models: *XGBoostRegressor*, *Prophet*, and *SARIMAX* to perform the analysis. We used parameter grid search to fine-tune our models, which allowed us to improve the results and evaluate model performance using the statistical metrics R^2 score, *RMSE*, and mean absolute error (*MAE*).

Results

Consumption

We experimented with a few hyperparameters to see their importance in predicting the consumption. Overall, we saw that the *Gradient Boosting* and *Random Forest* models performed better than our baseline model *Elastic Net Regression*. The finding demonstrated that we need to capture non-linear relationships to adequately predict consumption from multiple features.

Furthermore, in both *Gradient Boosting* and *Random Forest Regressor*, we observed that the inclusion of feature engineered temporal information played a significant role in improving the model's performance. The evaluated metric, R^2 score, was similar for the *Gradient Boosting* and *Random Forest Regressor* models. We know that dimensionality reduction plays an important role if we have a large number of feature columns. Although we had many feature columns, the number of feature columns was still not high enough to benefit from dimensionality reduction. Therefore we observed that adding principal component analysis (*PCA*) led to a drop in the model's performance. The effect of the other parameters such as *n_estimator*, *max_depth*, and *LI_ratio* are shown below.

Table 1. Impact of Hyperparameters in *Gradient Boosting* Performance

Model	<i>n_estimator</i>	<i>max_depth</i>	<i>PCA</i>	Time Information	R^2 Score
<i>Gradient Boosting</i>	300	8	No	1 previous day avg	0.86
	200	8	No	1 previous day avg	0.806
	300	8	Yes	A previous day avg	0.336

	300	8	No	No	0.786
	300	8	No	7 previous days avg	0.924
	300	5	No	1 previous day	0.785
After hyperparameter optimization	415	10	No	7 previous days avg	0.93

Table 2. Impact of Hyperparameters in *Random Forest* Performance

Model	$n_estimator$	Time Information	R^2 Score
<i>Random Forest Regressor</i>	100	7 previous day avg	0.929
	50	7 previous day avg	0.927
	100	No previous day avg	0.762
After hyperparameter optimization	158	7 previous day avg	0.93

Table 3. Impact of Hyperparameters in *Elastic Net* Performance

Model	$L1_ratio$	R^2 Score
<i>Elastic Net</i>	0.5	0.255
	0.2	0.2174
	0.4	0.2408

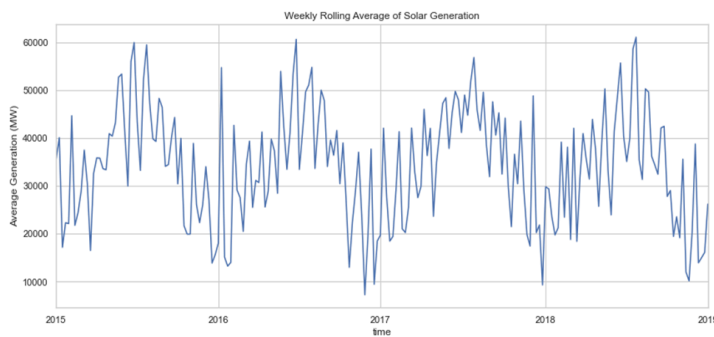
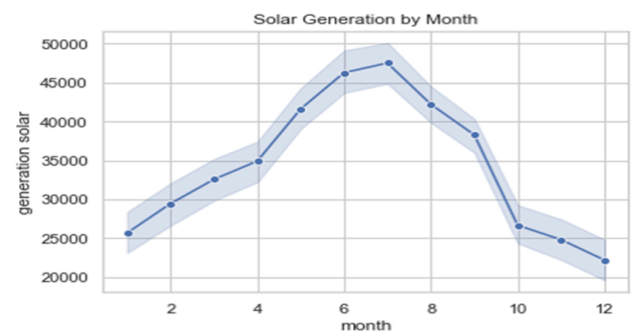
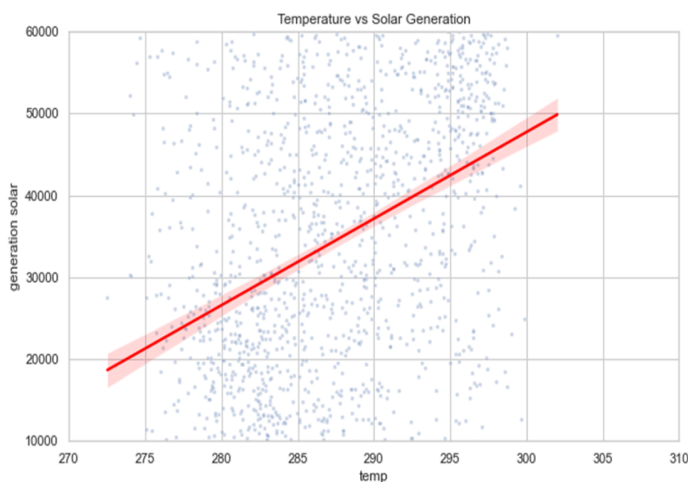
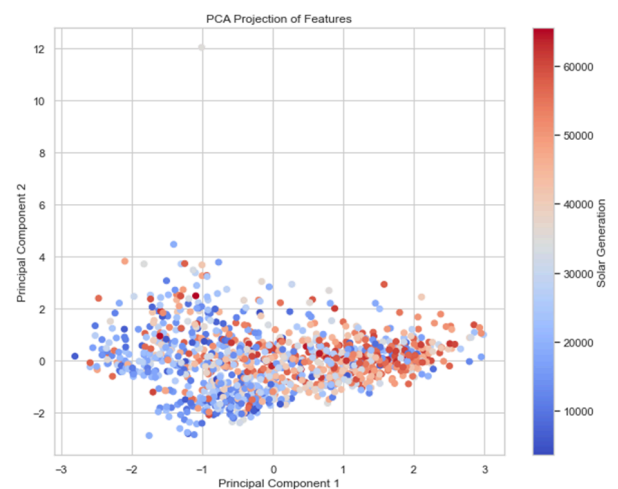
Renewable Energy Sources

The initial values were an R^2 score of 0.3518 and $RMSE$ of 1354.84. After hypertuning the parameters, the best parameters were found to be an $n_estimator$ of 150 and max_depth of 10 (first attempt) and $n_estimator$ of 200, max_depth of 20, and $min_samples_split$ of 5 (second attempt). In addition, the R^2 score increased to 0.9886 and $RMSE$ decreased to 179.71.

Table 4. Performance Comparison of Regression Models

Model	R^2 Score	$RMSE$
<i>Linear Regression</i>	0.3518	1354.84
<i>Ridge Regression</i>	0.3500	1354.84
<i>Random Forest Regressor</i>	0.9885	180.34
<i>Gradient Boosting Regressor</i>	0.9880	184.17
<i>Voting Regressor</i>	0.9886	179.71

Unsurprisingly, we found that as temperature increases, so does solar generation. Generation peaks during the middle weeks/months (summer season) and plummets during the last weeks/months (winter season).

**Figure 4. Weekly Rolling Average of Solar Generation****Figure 5. Solar Generation by Month****Figure 6. Temperature vs Solar Generation****Figure 7. PCA of Solar Generation**

Nonrenewable Energy Sources

We utilized two approaches to split the dataset:

1. Randomly split the training data (80%) and testing data (20%)
2. Take the earliest 80% data as training data and remaining 20% as testing data

We attempted to predict the summation of all nonrenewable energy generation solely by the weather data. The results are shown below.

Table 5. Results of Nonrenewable Energy Generation Prediction

Model	R^2 Score (Approach 1)	R^2 Score (Approach 2)
<i>Linear Regression</i>	0.042	-0.126
<i>Support Vector Regressor</i>	-0.001	-0.680
<i>Random Forest Regressor</i>	0.824	-0.501
<i>XGB Regressor</i>	0.625	-0.084

We observed that *Random Forest Regressor* has the best performance in Approach 1 and the models performed poorly in Approach 2. The results indicated that predicting future nonrenewable energy generation is more challenging, especially for those with larger time differences from the training data. In Approach 1, models can infer easily from nearby data points, whereas in Approach 2, there are fewer clues for inference.

Price Prediction

The *Linear Regression* model produced an R^2 score of 0.8116 and $RMSE$ of 5.252. The *Random Forest Regressor* model produced an R^2 score of 0.8137 and MSE of 5.222. Thus, the *Random Forest Regressor* model was a better model for price prediction. The latter's prediction of the actual prices based on the day-ahead prices were, on average, off by approximately 5.222 (the $RMSE$).

Table 6. Results of Price Prediction

Model	R^2 Score	$RMSE$
<i>Linear Regression</i>	0.8116	5.252
<i>Random Forest Regressor</i>	0.8137	5.222

Time Series Analysis for Generation of Renewable Energy

Predicting the amount of energy generated from renewable sources is crucial for balancing the generation of nonrenewable energy and stabilizing energy prices on the market (Ciarreta, Pizarro-Irizar, & Zarragam, 2020). For this analysis, renewable sources that most contributed to the total amount of energy included solar, wind, and hydro energy.

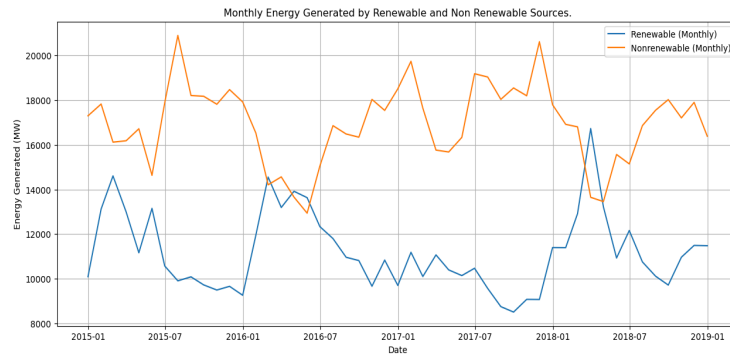


Figure 8. Seasonal Fluctuations of Renewable/Nonrenewable Energy Generation

Figure 8 demonstrates the seasonality pattern in energy generation. In the spring/summer months, the amount of energy obtained from renewable sources goes up, whereas during fall/winter, nonrenewable energy generation compensates for the decreased energy production from renewable sources. The pattern aligns with the findings of Ciarreta, Pizarro-Irizar, & Zarragam (2020).

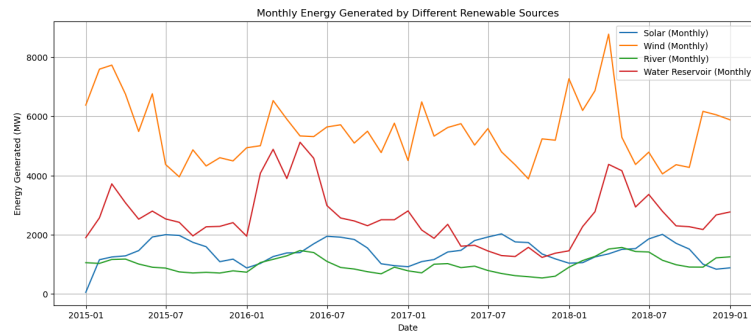


Fig 9. Components of the Total Renewable Energy

Figure 9 indicates that wind and hydro energy generation peak in the spring months, while solar energy increases during the summer.

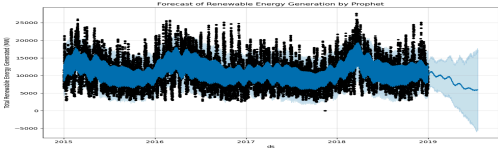
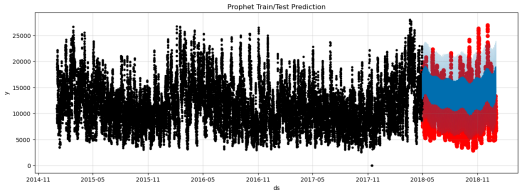
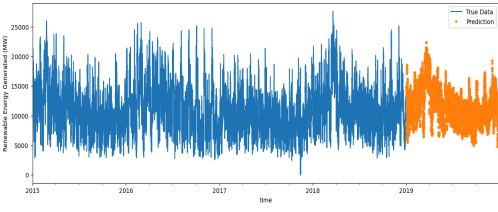
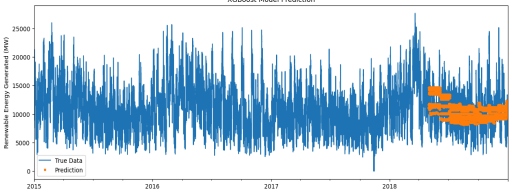
We used three models in our time series analysis: *XGBoost Regressor*, *Prophet*, and *SARIMAX*. All three models used time and total renewable energy data to make their predictions.

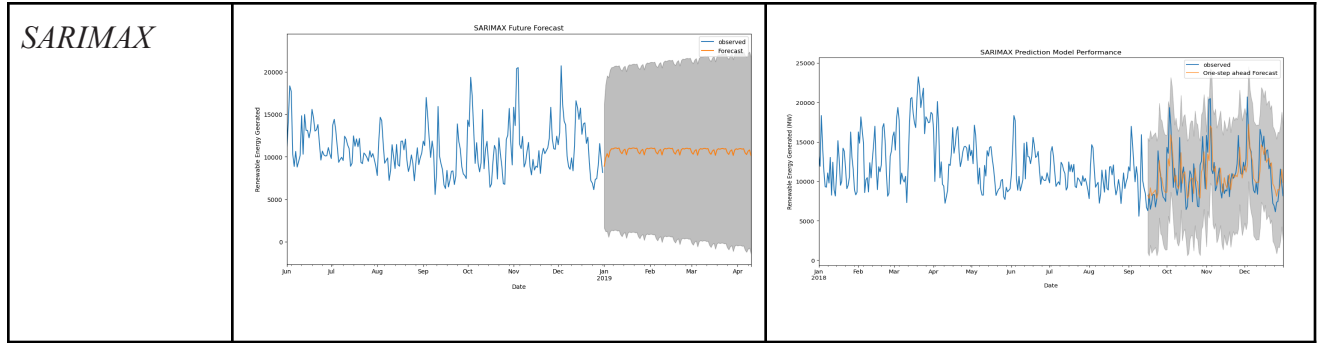
Table 7. Evaluation of Time Series Prediction

	<i>XGBoost</i>	<i>Prophet</i>	<i>SARIMAX</i>
<i>MAE</i>	2383	2831	1928
<i>RMSE</i>	3095	3614	2599
R^2 score	0.22	0.21	-0.55

The *SARIMAX* model outperformed both *XGBoost* and *Prophet* models on *MAE* and *RMSE*, which indicated that *SARIMAX* better represents the overall tendency of the data. However, *SARIMAX* demonstrated the worst R^2 score of the three models despite hyperparameter tuning. *XGBoost* showed second-best results, including the best R^2 score among the three models. The *Prophet* model, despite grid search and hyperparameters fine-tuning, did not perform well. Overall, we concluded that the amount of renewable energy can not be predicted solely on historical data. There are many external factors that can have an impact on results, such as temperature, seasonality, and the amount of non-renewable energy available.

Table 8. Train/Test vs Forecast Model Performance

	Model Future Forecast	Model Train/Test Performance
<i>Prophet</i>		
<i>XGBoost</i>		



One of the possible reasons behind unsatisfactory time series predictions is that total renewable energy consists of the renewable energy generated by a variety of sources, each of them with different seasonal patterns and external factors affecting their performance. To test this hypothesis, we chose to perform additional testing on a single source of renewable energy (solar) using the same three models. Restricting ourselves to only one source of energy and aggregating the dates by month, we were able to obtain R^2 scores of **0.78** from the *XGBoost* model, which demonstrated a high quality of the prediction. It should be noted that changing the aggregation to Daily reduced the quality of our prediction significantly: an R^2 score of 0.47 using the same model, compared to a Weekly R^2 score of 0.53.

Discussion

The major limitation that can be addressed in future research is our use of weather data at the core of our analysis. Since we did not have an exact geographical location of the renewable energy generation sources and the weather stations were located in the five main cities of Spain, we could not match with precision the exact location of the solar panels or wind turbines with given weather conditions. Energy consumption in particular locations is also dependent on other variables such as population density and level of urbanization. These similar factors limit the model's performance to make precise predictions.

Conclusion

Our research demonstrates that when combined, weather conditions and historical data strongly predict energy generation and consumption. In addition, we were able to accurately predict price and nonrenewable energy generation from weather conditions. Our project achieved high accuracy when trying to predict renewable energy generation. At the same time, we attempted to predict total renewable energy generation based solely on the historic data, which we found difficult due to the number of external factors that have an effect on the renewable energy sources. However, when we isolated an energy source (e.g., solar energy) and aggregated the source by month, we were able to predict the renewable energy generation with high accuracy.

About the Authors

Anastasia Kurakova is a first-year graduate student in the Big Data Analytics program at San Diego State University. She holds a Bachelor's degree in Biology from UW-Madison. She is currently a Research Assistant in the SDSU Data Science lab working on the evaluation framework for synthetic data generated using ML algorithms. In addition, she is a grading assistant for the Data Science Principles and Techniques course at SDSU. She is passionate about helping animals, spends her free time volunteering at animal shelters, and hopes to be able to continue to contribute to the field of animal welfare.

Email: kurakova@hotmail.com

Alexandra Nguyen is a first-year graduate student in the Big Data Analytics program at San Diego State University. She did her undergraduate education in Computer Science and Sociology at California State University, Long Beach. In her free time, she likes to spend time with her cats, read classic literature, and volunteer doing tax preparation for low-income families. She hopes to use her education to further help those in need.

Email: alexandra.ht.nguyen@gmail.com

Sabnam Pandit is a first-year graduate student in the Big Data Analytics program at San Diego State University. She holds a Masters in Business Administration-Information Technology from Tribhuvan University, Nepal and Bachelor's of Engineering in Computer Science from Visvesvaraya Technological University, India. Her interests include machine learning, deep learning, and data science. She is currently working on Explainable AI using a large language model in the Data Science Lab.

Email: spandit6542@sdsu.edu

Wanjing (Anna) Yan is a first-year graduate student in the Big Data Analytics program at San Diego State University. She holds a Bachelor's degree in Applied Mathematics from UC San Diego. She is actively involved in extracurricular activities and research projects and is currently exploring research opportunities in the field of data science with a focus on areas such as natural language processing and predictive modeling.

Email: wyan4786@sdsu.edu

Yu-Chieh (Jason) Tu is a fourth-year undergraduate student in Computer Science and Information Engineering at National Taiwan University. His research interests include machine learning, natural language processing, and computer vision. His related experiences include a research internship at Academia Sinica about question generation and answering and working as a research assistant at the Language Training and Learning Center.

Email: b09902138@csie.ntu.edu.tw

References

- Brey, J. J. (2021). Use of hydrogen as a seasonal energy storage system to manage renewable power deployment in Spain by 2030. *International Journal of Hydrogen Energy*, 46(33), 17447–17457. <https://doi.org/10.1016/j.ijhydene.2020.04.089>
- Ciarreta, A., Pizarro-Irizar, C., & Zarraga, A. (2020). Renewable energy regulation and structural breaks: An empirical analysis of Spanish electricity price volatility. *Energy Economics*, 88, 104749. <https://doi.org/10.1016/j.eneco.2020.104749>
- Ding, Y., & Dang, Y. (2023). Forecasting renewable energy generation with a novel flexible nonlinear multivariable discrete grey prediction model. *Energy*, 277, 127664. <https://doi.org/10.1016/j.energy.2023.127664>
- Kolasniwash. (2019, October 10). *Hourly Energy Demand Generation and Weather*. Kaggle. <https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather>
- Kolosok, S., Saher, L., Kovalenko, Y., & Delibasic, M. (2022). Renewable Energy and Energy Innovations: Examining relationships using Markov switching regression model. *Marketing and Management of Innovations*, 2(1), 151–160. <https://doi.org/10.21272/mmi.2022.2-14>
- Pata, U. K. (2021). Linking renewable energy, Globalization, agriculture, CO2 emissions and ecological footprint in BRIC countries: A sustainability perspective. *Renewable Energy*, 173, 197–208. <https://doi.org/10.1016/j.renene.2021.03.125>
- Timilsina, G. R. (2021). Are renewable energy technologies cost-competitive for electricity generation? *Renewable Energy*, 180, 658–672. <https://doi.org/10.1016/j.renene.2021.08.088>

Appendix

Dataset:

<https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather/data>

GitHub Repository: <https://github.com/Sabnam-Pandit/Green-Predictors>