Predicting Spanish Energy Production and Consumption - Write-up

Mahlon Page

For my project I wanted to use weather data to predict energy production of different types and energy prices in the coming hours/days. I've taken two EEPS classes at Brown including one called EEPS 0850: Weather and Climate where we talked about modeling and predicting future weather as well as the weather's impacts on various aspects of the world like energy use and consumption. In my other class EEPS 1320: Intro to GIS with Environmental Applications we've talked about modeling energy use on maps and the impact of energy on the climate. I wanted to bridge my interest in the two topics of machine learning and weather forecasting to predict future energy production and consumption. Similarly, I worked at an internship last summer where some of my coworkers modeled dam energy production using rainfall data and this felt like a cool and similar project.

I found a dataset with hourly weather data in Spain as well as energy data for the same time. I have access to detailed weather information such as wind speed and rainfall in 5 major Spanish cities: Seville, Madrid, Barcelona, Bilbao, and Valencia. I also have detailed energy production data divided into categories, most importantly, solar, wind, and hydro. Similarly we have overall energy price data. The problem at hand is using this data to accurately predict these types of energy production as well as the overall energy price.

Initially I tried to tackle the problem by looking at individual linear regressions of features versus energy production. I looked at a correlation matrix and tried to run linear regressions with several features to see if I could predict the price accurately. Linear regressions made some progress. They helped me see which features might be good predictors of energy production. Primarily, when I started, I was working with wind energy production, so let's talk about that for now.

In my initial searches I found mostly what I suspected, weather data telling us about the wind speed in each of our five cities were some of the best predictors of wind energy production. With all of these features I was able to achieve an R^2 of around .5 meaning we could explain around 50% of the variance. Surely we can do better than that though. I thought that this could surely be improved upon.

My first update was to move from a simple linear regression to a more complex classifier. I eventually settled on decision trees, and more importantly an ensemble of them, a random forest. I saw some improvements in R^2 when switching to this model but it still felt like there was more to be learned.

The next improvement was sure to be made in feature engineering. The current wind speed is useful but information about the past wind speed, how it's changing, and other longer time frame features might be more useful. I made a major improvement here in creating lag features for the dataset and rolling means for the dataset. I added features for 1hr, 3hr, 6hr, 12hr, and 24hr lagged data. I also created features for rolling averages across the last day, 3 days, 7 days, and month. These new features should help our model understand not just the current situation but also what has been happening and what has been happening in the other cities across Spain. Once we have lag features and rolling data, information from other weather stations in different cities can become relevant to weather data in other cities. For example, consider if wind farms are 100 miles from any city. Having data from the past can be more helpful in predicting current weather more so than the current weather in these locations.

These new features brought our model great performance, but now our model has far too many features to work with. It overfits with all of this data to work with. I started to achieve higher and higher training R^2 and lower test R^2. My original 20 or so features had become around 200 features. I had trouble in deciding how to narrow down on features.

I looked at some of the weights for the models I was training and realized how large they were. I decided to scale down all of these features to be in the same range. This improved model performance a decent bit and also made clearer which of the weights mattered more.

After scaling I decided to use a Lasso Regression to determine which features were most important in predicting wind energy generation. I found something interesting here which was that wind speed data was not nearly as good of a predictor as some of my other features in determining wind generation. All of the best features in predicting wind energy generation were features tracking the rolling average of temperature over the last week and month. Tapping into my knowledge from my EEPS class: Weather and Climate. Wind is primarily if not exclusively caused by temperature differences in locations. When it is warm in Boston and cold in Providence a strong wind goes from Boston to Providence to try to stabilize the temperature and reach equilibrium. It seemed like the model was learning this as it found temperature differences between the different cities across the countries. The wind farms between them were surely most impacted by these temperature differences. A random forest predictor with just six features, all related to average temperature in these different cities gives us an R^2 of .95! Almost all of the variance in the wind energy production can be explained by these rolling temperature averages through this model.

For solar and hydro energy I took a similar approach. I standardized and scaled the data, I created lag features and rolling means for each feature and then I used a Lasso Regression to select the best features for prediction in these tasks.

Solar energy had a bit of a different outcome though. The best features for predicting solar energy generation were almost all temperature related, which is interesting because we have percentage of cloud cover as a feature. I would have expected less cloud cover to equate to more solar energy production rather than a higher temperature meaning that. This model also preferred current data and data from the last few hours more so than longer averages. This makes sense though, the current sunlight affects most of the country the same while wind is more of a local phenomenon. Also sunlight converts into solar energy rather quickly while for wind the temperature changes then wind occurs which more slowly produces energy. I think temperature can be a fairly good indicator of direct sunlight however and the model is likely capturing that fact rather than cloud cover as even when cloud cover is low we can have low solar generation.

This model was able to perform fairly well, with just temperature data and some lagged temperature data our model achieves an R^2 of .77 which means we can explain about 77% of the variance with just these details. This seems reasonable, although perhaps could be improved upon by factoring in the season of the year or the angle of the sun on a given day. I imagine a large factor at play will be this for solar generation as opposed to wind where season does not matter as much.

Hydro energy had a stronger result and a more predictable relationship. The best features for predicting hydro energy production were rolling temperatures and rolling precipitation information. Precipitation makes a lot of sense, when it has rained a lot over the last month we expect rivers to flow more and dams to generate more electricity. When it rains less we expect the inverse. Temperature also makes a lot of sense. Hot weather can mean the evaporation of water, leaving less of it, alternatively it can also mean snow melt, adding more water to the rivers. Temperature surely has a strong effect on water quantity nonetheless. Our model here performs fairly well. We achieve an R^2 of .90 which means we explain about 90% of the variance with our model.

Finally I wanted to predict energy prices based on this information. Energy prices are a bit trickier. Weather data can help predict renewable production but does not help forecast non-renewable production very well. Currently around ½ of Spanish energy is renewable so we should have information to predict about half of the production. However, weather conditions can also affect how much energy people use and this is arguably more important in determining energy price. When using our same pipeline, I was able to achieve an R^2 of .90. A large collection of features were useful here. Temperature, wind speed, and precipitation are all helpful in predicting the renewable energy production from before, but also, they all affect how much people are inside, using energy, which increases the price of energy.

All in all, I was pleased with my results. My models were able to achieve the following R^2 values across the various prediction tasks.

Prediction Task	R Squared Value Achieved
Solar Energy Production	.78
Hydro Energy Production	.91
Wind Energy Production	.95
Energy Price Prediction	.90

The biggest improvement to be made on this project is in solar energy production prediction. I think all in all the data pipeline and model training process works well for this process. A random forest is an effective classifier for the task at hand given the simplicity of the way these features often relate to the outcomes. More rain means we will have more hydro production and other things like that. This type of problem is perfect for a simple model in ensemble to combine weak learners and perform stronger at our prediction task where the overall trends aren't that complex but can still be expressed in our ensemble.

To tackle solar energy production prediction I think a big change would be to take into account when in the year the model is predicting. I would pass the model information about the current season, or perhaps the angle of the sun. This data should be fairly easy to get and should give our model the ability to understand the more nuanced trends throughout the year. It is well known that solar energy production drastically decreases in the winter and having information for the datapoint about when in the year it actually is would surely give this model stronger predictive power. Our current solar model is using temperature information, I imagine it is using this as a proxy for trying to learn the season. Providing the season directly would help our model get the rest of the way to a conclusion.

A further extension of this project could include predicting other types of energy production. The dataset had other types of production but they seemed less relevant to the weather so I focused less on predicting them. I think however that these energy productions could be predicted in a similar vein to price in that they are produced as needed. Another interesting task could be to simply predict overall energy production and consumption rather than price. Surely these values are fairly intertwined with price and each other but they could both be predicted to let the model have a more direct value to predict rather than the resulting price.

My takeaway from this project is that weather can predict energy production and energy use fairly well. They have a clear logical connection and the data backs this up. I was successfully able to use my knowledge from Machine Learning to tackle this prediction task and my knowledge from my EEPS classes to rationalize the results.