LUDWIG MAXIMILIAN UNIVERSITY OF MUNICH

Interpretting Spain’s Energy Price Prediction

Author:
Konul Baghirova
Pranav Ragupathy

Supervisor:
Gunnar König

March 14, 2021

Contents

List of Figures

List of Abbreviations

NaN	**N**ot **a** **N**umber
TSO	**T**ransmission **S**ervice **O**perator
PDP	**P**artial **D**ependence **P**lot
ICE	**I**ndivdual **C**onditional **E**xpectations
ALE	**A**ccumulated **L**ocal **E**ffects
PFI	**P**ermutation **F**eature **I**mportance
SHAP	**SH**apley **A**dditive ex**P**lanations

Chapter 1

Interpretable Machine Learning

1.1 Introduction

As part of a seminar on Interpretable Machine Learning, we picked a challenging dataset, fit a model on it and explained the predictions made by it. We choose to work in the Energy domain. Of all the industries where Machine Learning methods have been applied to improve efficiency, the Energy industry is one that has benefited immensely and allows us to focus on sustainable and cost-effective solutions. Since energy is one that impacts all our daily lives and the climate in a more insidious sense, it is imperative for us to be able to have efficient, but more importantly, transparent decisions that can be interpretable by those involved in making decisions as much as those affected by it.

Electricity, as a form of energy, is produced from various sources - both renewable and non-renewable. Although we have renewable sources such as hydroelectric and wind energy, a bulk of the electricity demand is met by fossil hard coal and fossil gas. Any form of energy generated from fossil fuels is not sustainable and we must find ways to reduce the fraction of energy generated from these sources. The price of electricity is dynamically set based on various factors of supply and demand. Each country has an organisation that has its own model to set these prices. The power of being able to forecast the energy requirements and price a day in advance can have several benefits from cutting cost and reducing the supply if a low demand is predicted. However, while forecasting we must know why a specific model makes a certain prediction. We might have questions like ***'What are the most important features/factors impacting tomorrow's price?'*** or something like ***'Can we reduce the amount of energy generated from fossil oil tomorrow without the price increasing by more than 2 EUR?'***. These are the kinds of questions we will attempt to answer through our analysis of Spain's energy supply and demand data.

1.2 Dataset

1.2.1 Description

We have chosen to work with the Hourly energy demand generation and weather dataset which records the hourly energy generation, consumption and pricing in Spain over 4 years along with weather-related data in the top 5 cities (in terms of population and size) during the same period(Jhana, 2019). The two raw files we have are energy.csv and weather.csv. Brief description of features in each can be found in table 1.1.

Before we start the Exploratory Data Analysis, we must state our assumptions regarding the two datasets. Firstly, the individual entries for every hour are independent of each other. We also assume that the weather-related features are a comprehensive description of the overall

TABLE 1.1: Datsaset desciption

Filenames	Number of Rows	Number of Columns
energy.csv	35,064	29
weather.csv	178,396	17

weather in Spain. This is mainly backed by the fact that their geographic distribution covers most of the part of Spain's territory uniformly. Moreover, it is useful to note that these 5 cities alone comprise approximately a third of the total population of Spain. Another key feature of the dataset is that it provides predictions of the total load and total price made by the Spanish TSO Red Electric España. This gives us an ideal benchmark to compare our predictions with the official forecast values of the industry. We further motivate our model choice in comparison to the TSO's predictions.

1.2.2 Preparation

Energy Data

This dataset contains hourly information about the generation of energy in Spain. In particular, there is info (in MW) about the amount of electricity generated by the various energy sources, as well as the total load (energy demand) of the national grid and the price of energy (€/MWh).

Note: Since the generation of each energy type is in MW and the time-series contains hourly info, the value of each cell represents MWh (Megawatt hours).

From this naive look at the features in the energy data, we have the choice of modelling our data to predict either the total load or price of unit energy for every hour. For us, price is a more relevant feature to predict than total load since it is affected by several intangible factors and we would like to find the relationship between these features and price.

Time is the only common feature between the two datasets and we will use it as our index which will be convenient while merging the two datasets. We convert this into the python *datetime* format for standardisation.

After a quick check of all features, we dropped the columns which had a majority of NaNs or those that did not contribute to the variance in the data. There were some forecast features such as *'price day ahead'* and *'total load forecast'* which are predictions made by TSO. We could include these in our model as it would give us better accuracy, but it would take away the marginal importance of the other features which we want to know the effect of.

Further, we did some outlier detection and a basic check to see if there are any missing values. Considering that there were no missing values in the target variable, we filled the other missing values with a simple linear interpolation method. At this stage, the energy data was cleaned and ready to be merged with the weather features.

Weather Data

In order to clean the data, we first dropped the features that were not descriptive and the features that had a majority of NaNs. We then performed outlier detection by constructing box-plots for each feature and noticed some very unlikely values. From an initial glance at

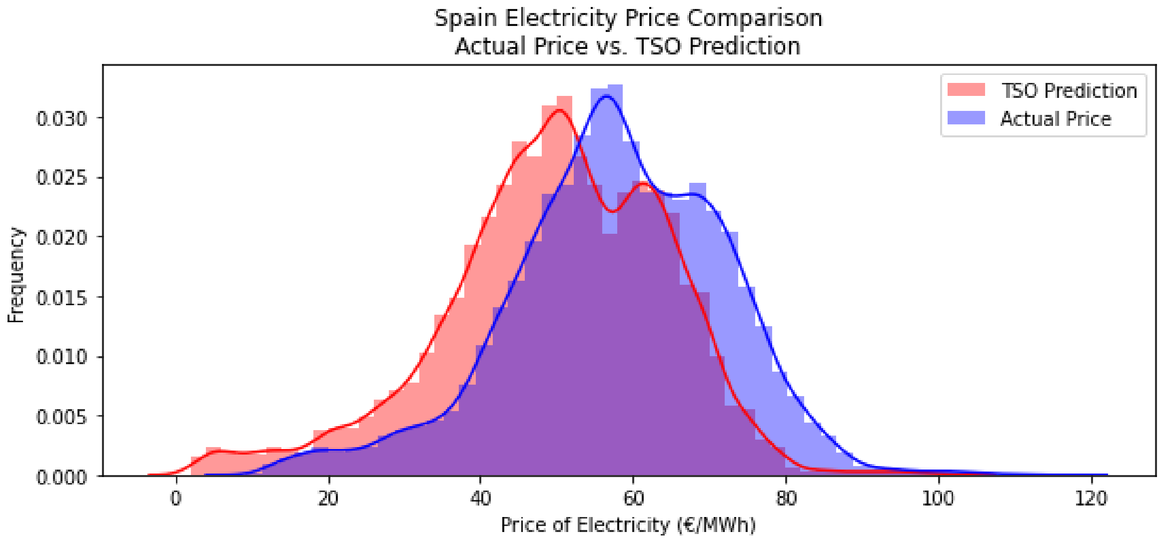

FIGURE 1.1: Spain electricity price comparison: Actual price vs TSO prediction.

all the variables, we saw that *'pressure'* and *'wind_speed'* have very apparent outliers concerning the other data. We changed the specific outlier values to NaN and interpolated new values in its place.

After running a consistency check on whether all cities had the same number of entries we split the data respective to each city and merged them all.

After all the preprocessing, our final merged dataset had a total of 65 feature variables where 15 were energy-related features and 10 weather-related features from each of the 5 cities for each hour. The final list of energy features and weather features are:

Weather: ['temp', 'temp_min', 'temp_max', 'pressure', 'humidity', 'wind_speed', 'wind_deg', 'rain_1h', 'rain_3h', 'snow_3h', 'clouds_all']

Energy: *['generation biomass', 'generation fossil brown coal/lignite', 'generation fossil gas', 'generation fossil hard coal', 'generation fossil oil', 'generation hydro pumped storage consumption', 'generation hydro run-of-river and poundage', 'generation hydro water reservoir', 'generation nuclear', 'generation other', 'generation other renewable', 'generation solar', 'generation waste', 'generation wind onshore', 'total load actual', 'price day ahead']*

1.3 Model Selection

We use the given TSO predictions as our initial benchmark as reflected in figure 1.1. We can see that there is a systematic aberration in the predictions. As our first model, we fit a simple inherently interpretable linear regression model on our pre-processed data and compared the results with the TSO prediction. Even for a simple model, it seemed to have a much lower Mean squared error as compared to the TSO Prediction. However, this model was not enough to provide a good source of interpretation.

TABLE 1.2: Comparison of All Models on Validation Data

Hyperparameter	Value
base_score	0.5
booster	gbtree
importance_type	gain
learning_rate	0.1
max_depth	100
min_child_weight	0.1
n_estimators	100
objective	reg:linear
scale_pos_weight	1

TABLE 1.3: Comparison of All Models on Validation Data

Model	Mean Squared Error	R^2 score
Linear Regression	96.135	0.52
Random Forrest	29.342	0.81
Gradient Boosting Regressor	61.432	0.69
XGBoost	24.563	0.87

Next, we tried fitting a gradient boosting regressor and a random forest regressor and although it performed much better than the Linear regression model, it only provided a best-case mean square error of around 30. It seemed like we could do better and fit a more complex model.

After trying various parameters for the simpler models we arrived at the XGBoost Regressor which performed considerably better than any of the previous models. We could further improve the performance of the model with hyperparameter tuning by testing out different depths and learning rates. We also experimented with the number of estimators and the minimum child weight to eventually give us a fairly precise model with an MSE of around 24.2 and an R2 score of 0.89 which seemed to be our best option. The final set of hyperparameters used for the XGBoost model is reflected in table1.2.

To compare the models, we plot the absolute residuals of all the models we fit and further compared our final chosen model with the TSO predictions and actual price as seen in this figure 1.2. We can also see the absolute values of the performance measures in this table 1.3. We can conclude that our XGBoost model outperforms all other models.

1.4 Model Interpretation

Final Model: XGBoost

1.4.1 Hypothesis 1

Which features are most important for the model?

This is a basic, yet important question to be answered on the model behaviour. In our case it would be useful to know which are the features most useful to the model to predict the

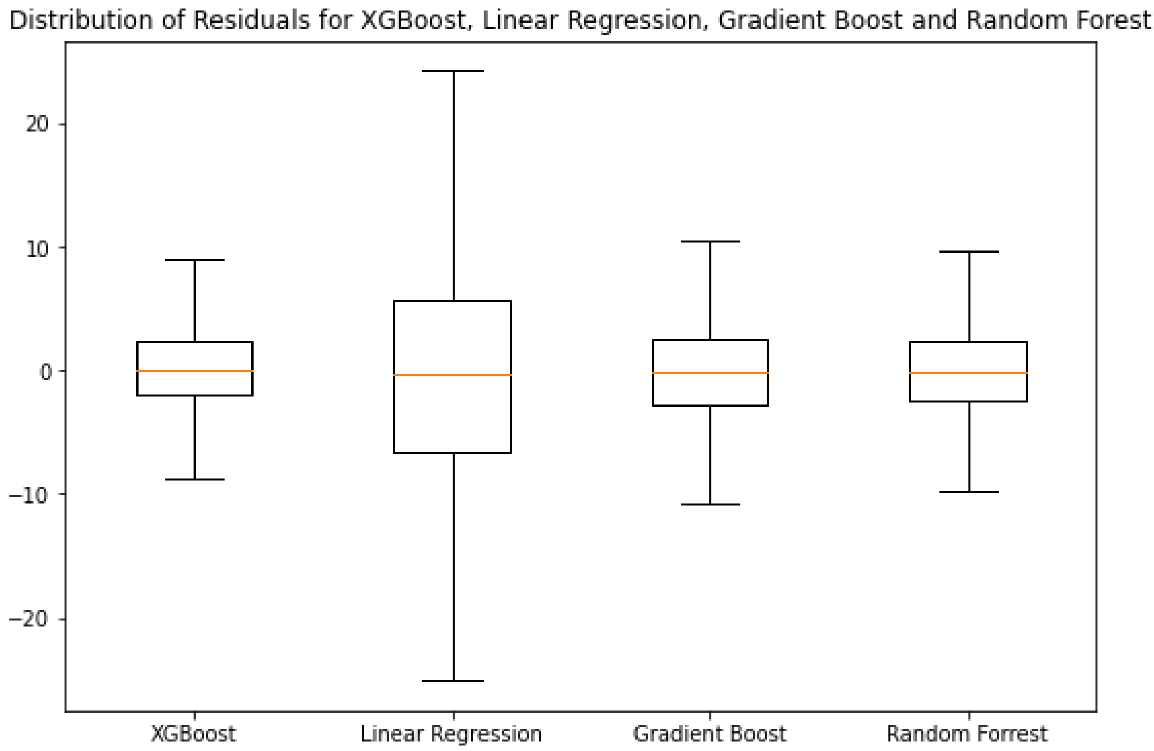

FIGURE 1.2: Distribution of residuals for different models.

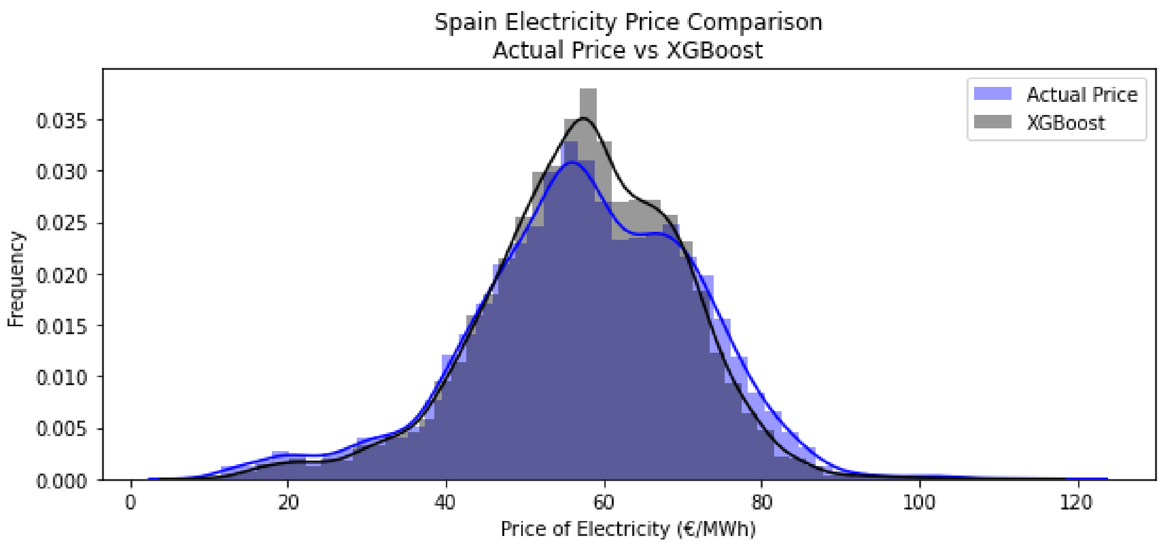

FIGURE 1.3: Spain electricity price comparison: Actual price vs XGBoost prediction.

```
generation fossil gas0.242 +/- 0.005
generation fossil hard coal0.126 +/- 0.003
total load actual0.108 +/- 0.004
generation other renewable0.085 +/- 0.004
generation hydro water reservoir0.059 +/- 0.002
generation waste0.058 +/- 0.002
generation hydro run-of-river and poundage0.057 +/- 0.001
pressure_Barcelona0.056 +/- 0.002
generation solar0.051 +/- 0.003
generation nuclear0.049 +/- 0.002
pressure_Bilbao0.046 +/- 0.003
generation hydro pumped storage consumption0.044 +/- 0.002
generation other0.038 +/- 0.002
temp_max_Seville0.031 +/- 0.002
```

FIGURE 1.4: Results of PFI based on test data.

actual price. This could have further implications (depends on other factors as well) on what sources of energy are vital to the price of electricity. Although it is not possible to make any conclusive inferences on this effect beyond the model, it could provide a useful starting point to further investigate these features. It could also be used for feature selection, but we are doing only a post-hoc analysis.

The methods we will use to measure the importance of each feature is the inbuilt *feature_importances* attribute of the *sklearn xgboost* library and Permutation Feature Importance using the *permutation_importance* module which is a part of the *sklearn.inspection* library. Each method has their respective pitfalls and we will address them while interpreting their results.

PFI based on test data: This gives a fair representation of the feature importances as it is computed on unseen data. We are more interested in how much each feature contributes to the performance of the model on unseen data. Results can be found in 1.4 and 1.5.

PFI based on training data: As mentioned in Molnar, 2020, the PFI score is essentially the increase in model prediction error after permuting the feature's values . Since this depends on the prediction error it does not make complete sense to measure importance on the training data but it gives us an insight into how important certain features were while training. Results can be found in figure 1.7.

Feature importances of XGBoost: This is the most easy and naive way of finding feature importances. The general method of obtaining feature importance scores from tree based models such as this and Random Forests is the mean change in impurity when the particular feature is added or removed. This measure of impurity could be the Gini index or entropy which is essentially the information gain when the feature is included. However, it must be noted that this inbuilt method was found to be biased and unstable as stated by Strobl et al., 2007. This is the reason we trained an *XGboost* model 10 times and averaged the features importance values. Results can be found in figure 1.8.

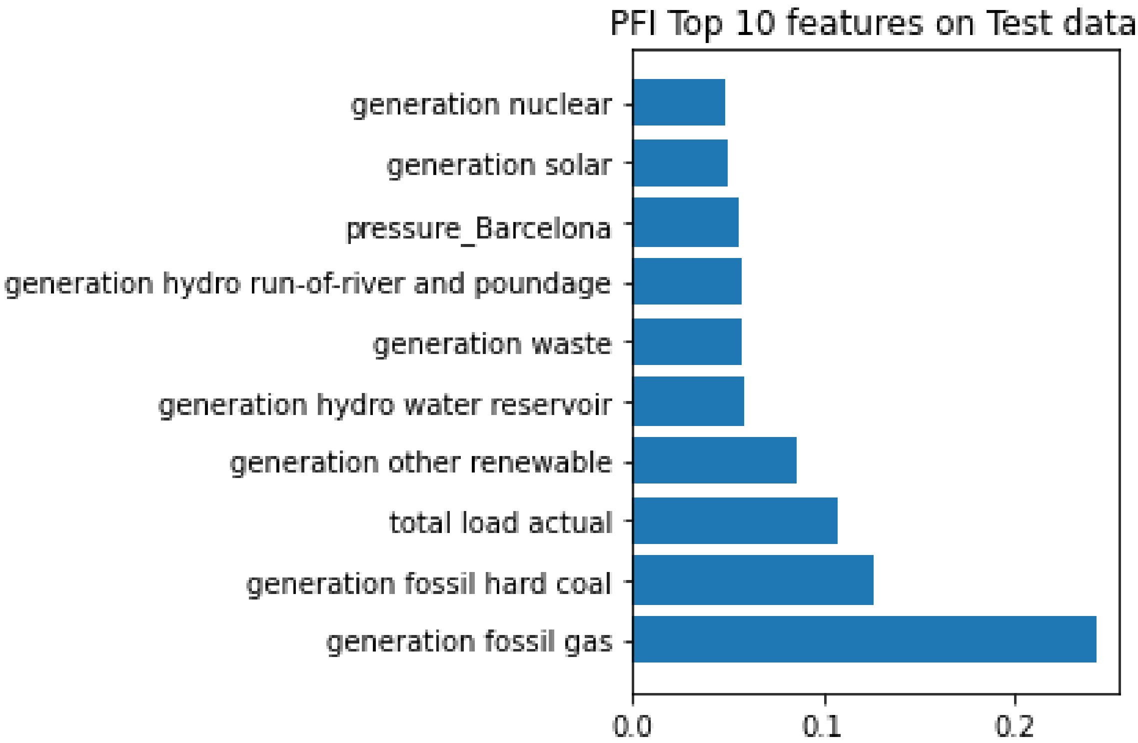

FIGURE 1.5: Top 10 features according to PFI based on test data.

```
generation fossil gas0.261 +/- 0.004
total load actual0.128 +/- 0.001
generation fossil hard coal0.121 +/- 0.001
generation other renewable0.094 +/- 0.001
generation hydro water reservoir0.061 +/- 0.001
pressure_Barcelona0.060 +/- 0.001
generation hydro pumped storage consumption0.054 +/- 0.001
generation hydro run-of-river and poundage0.050 +/- 0.001
generation solar0.045 +/- 0.001
generation waste0.045 +/- 0.001
pressure_Bilbao0.035 +/- 0.001
```

FIGURE 1.6: Results of PFI based on training data.

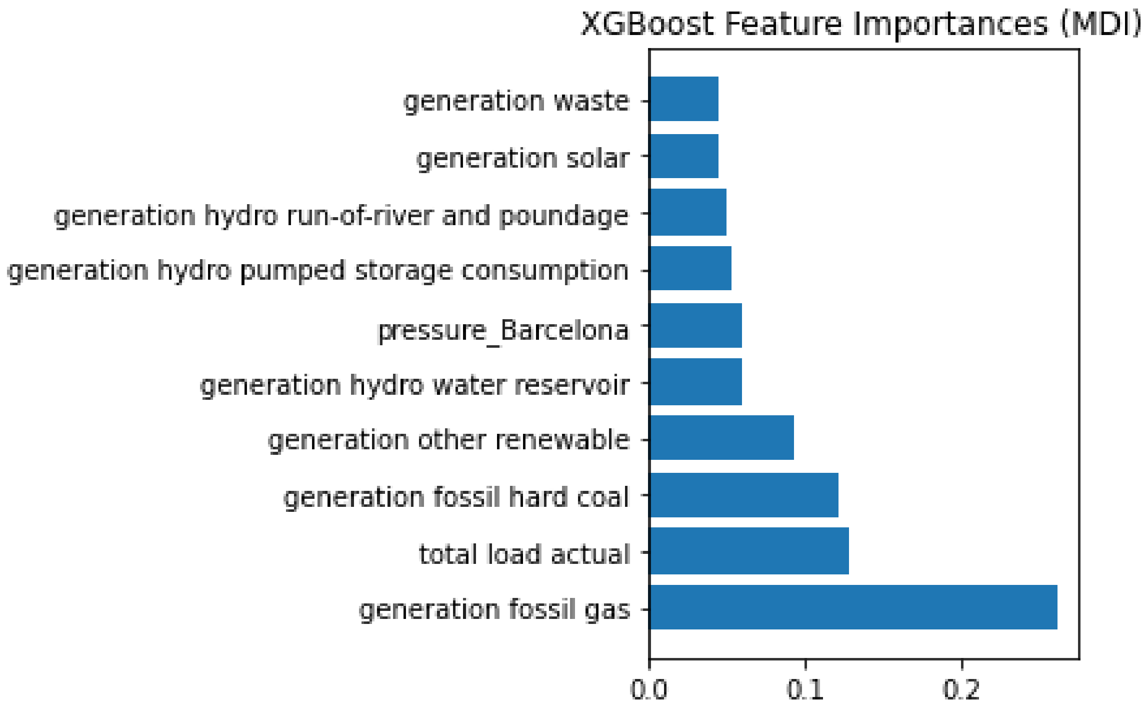

FIGURE 1.7: Top 10 features according to PFI based on training data.

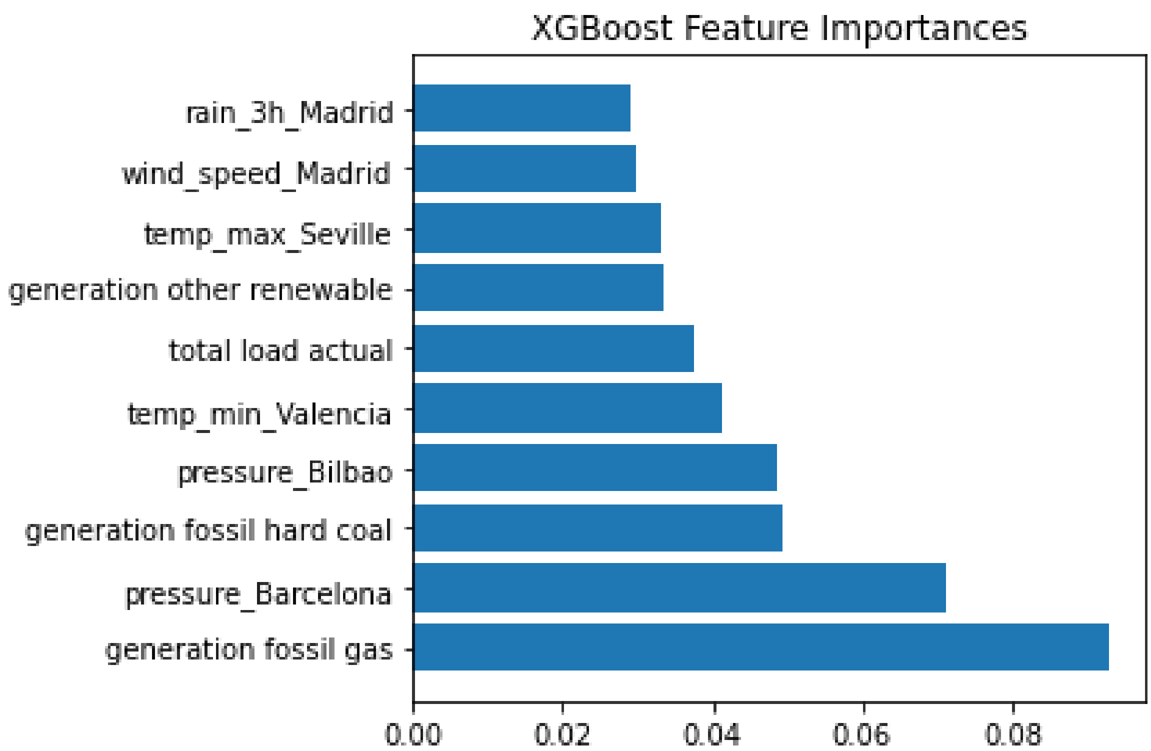

FIGURE 1.8: Top 10 features according to XGBoost feature importances.

Final Interpretation

From the three different measures of feature importances we can see that generation fossil gas has the highest score. This makes logical sense because we must also note that this contributes to the highest mean power generated other than through nuclear sources. It has to be the biggest driver to the actual price of energy.

However, before making any conclusions we must note that these features individually do not contribute very greatly to the overall variance.

Another general trend we can notice is that a majority of the top 10 features are energy related features rather than weather related which makes intuitive sense while predicting price.

1.4.2 Hypothesis 2

Does temperature interact with generation of fossil oil? If so, what is the strength of interaction?

We want to see if the temperature in different cities would interact with the generation of fossil gas. It might be interesting to see if variables that are not very correlated interact with each other for the model prediction.

The various methods that we will use to answer this question are PD interaction plots and Friedman's H-statistic using the pdpbox library in python and calculating SHAP values using the shap and alibi libraries in python.

H-statistic

To get a basic measure of the interaction measure between each pair of features in question we compute Friedman's H-statistic measure. Since there is no direct implementation of the H-statistic measure in Python, we follow an implementation as suggested on *Discovering Interaction Effects in Ensemble Models* which draws from Molnar, 2020.

It is interesting to note that the H-statistic between generation fossil oil and the temperature in each city is different as seen in 1.9. We can clearly see that the temperature variables from Barcelona, Bilbao, and Madrid are far greater than those from Seville and Valencia. Although the individual interaction strengths are not great, it is useful to see the relative change in interaction values in different cities. We can further investigate how these individual interactions look using Partial Dependence Interaction plots.

Partial Dependence Interaction Plots

To further investigate how the interaction looks like we want to compare the measure of strength along with PD interaction plots for each pair of variables in concern.

Before inferring anything from the PDP plots we must see the correlation of the variables in concern with all other variables in the dataset. In the correlation heatmap 1.10, the first 5 rows are the temperature features from the 5 cities and the last row is generation fossil gas. We can see that the temperature variables are highly correlated with other temperature variables from each city and the generation fossil gas is not that correlated with the other variables.

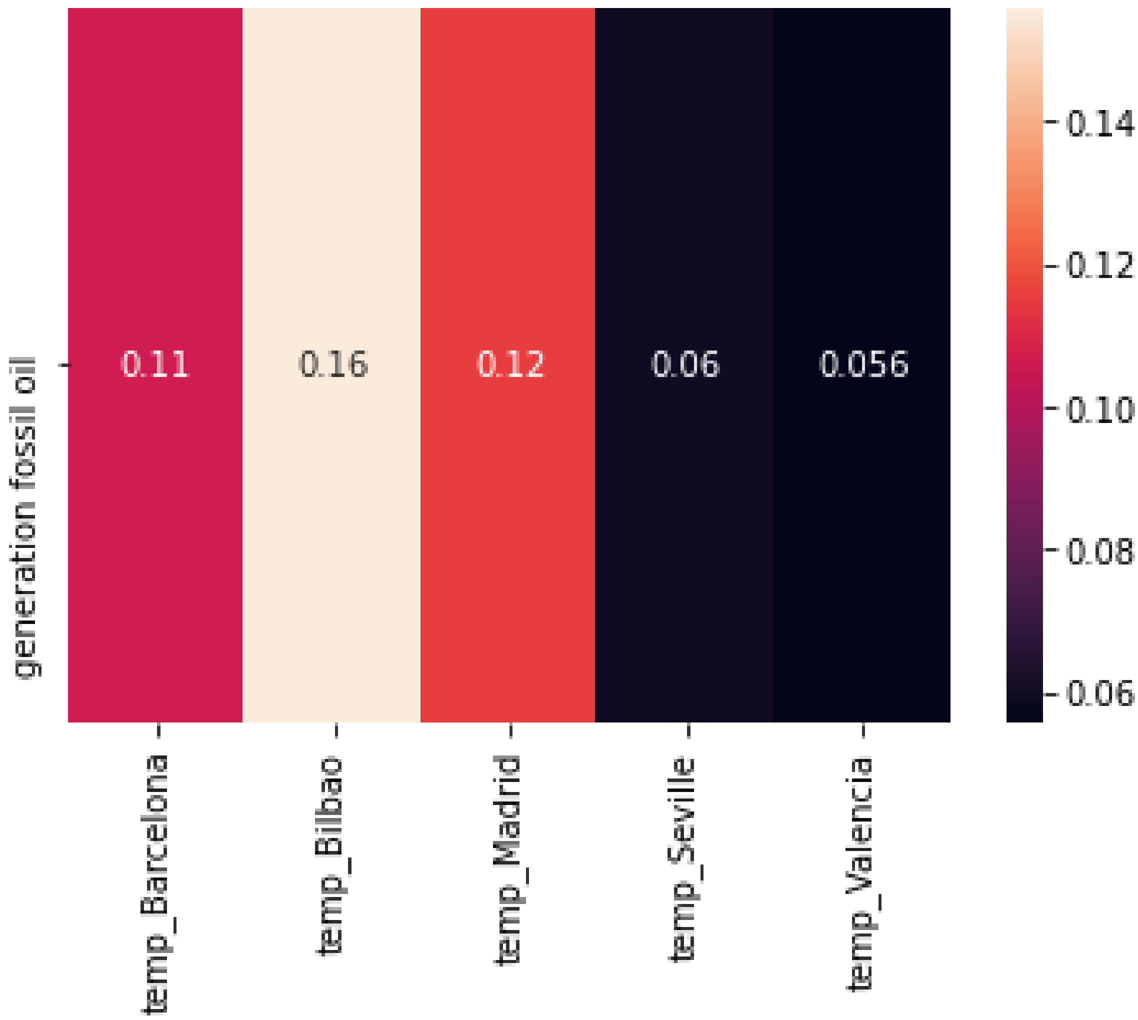

FIGURE 1.9: H-statistics between pair of features in question.

FIGURE 1.10: Correlation heatmap: temperature - other variables, generation fossil gas - other variables.

SHAP

SHAP interaction values provide the most theoretically strong measure of interaction (Molnar, 2020). Although this would give a comprehensive measure of interaction, it was not possible to compute Shapley values due to the high complexity of the data which took over 8 hours to compute. In light of time restrictions, we did not proceed with this method to answer our hypothesis.

Final Interpretation

Considering the correlation between the variable in concern, we cannot make any conclusive inferences from our PD Interaction plots as seen in 1.11. Without complete certainty we can say that the the variables do interact in the prediction of price by the model.

1.4.3 Hypothesis 3

How would the prediction of the actual price be affected on average by the* generation solar *and generation fossil gas?

We are interested to find out if the generation solar and generation fossil gas features have an effect on the model's prediction of the target variable.

We will answer this hypothesis by first investigating the correlation of generation solar with all other variables. Then we will create a Partial Dependence plot using the *pdpbox* library and then create an Accumulated Local Effects plot to compare with in case the PDP is biased by correlated features.

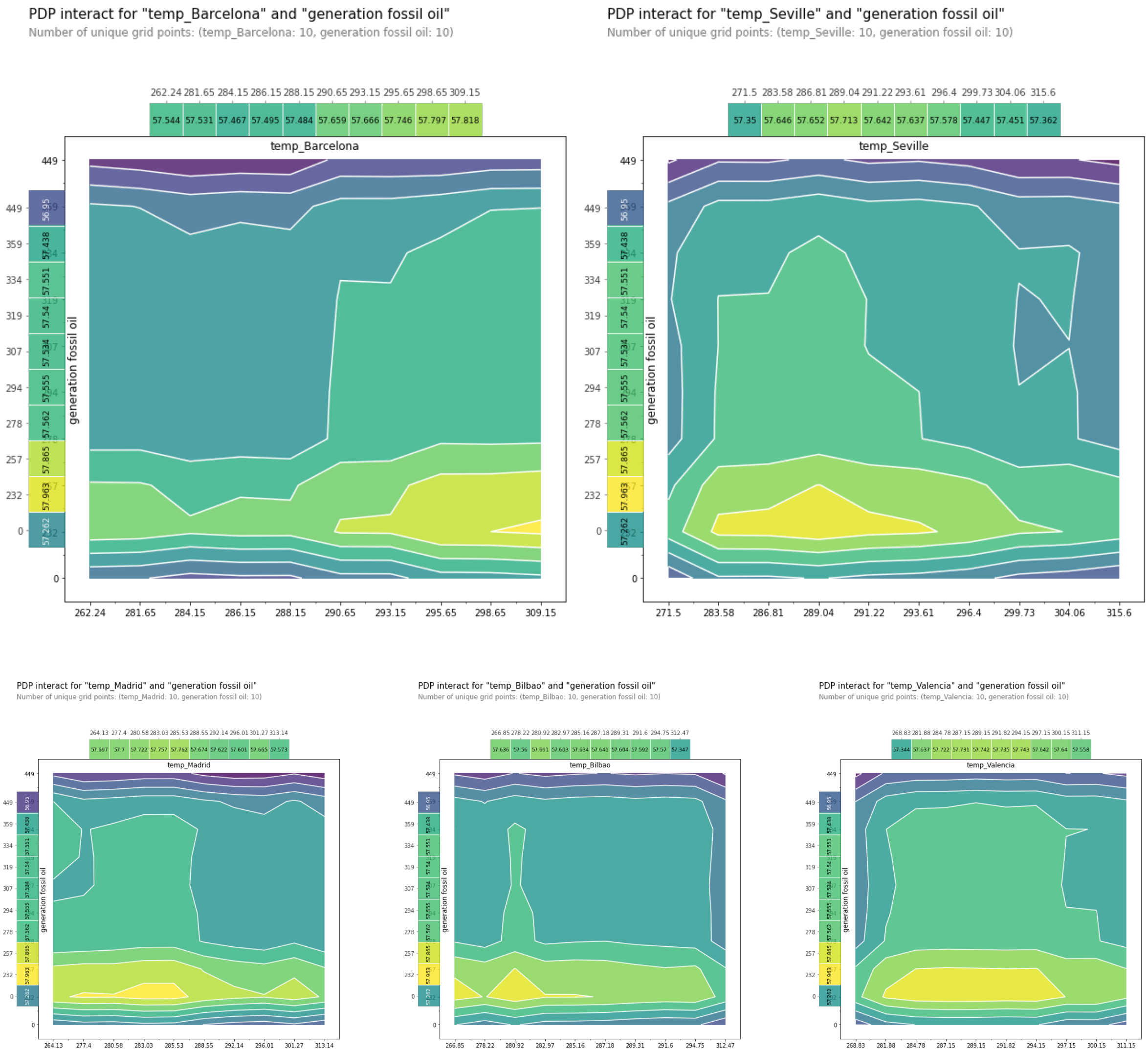

FIGURE 1.11: Partial dependence interaction plot: temperature in each city - generation fossil gas.

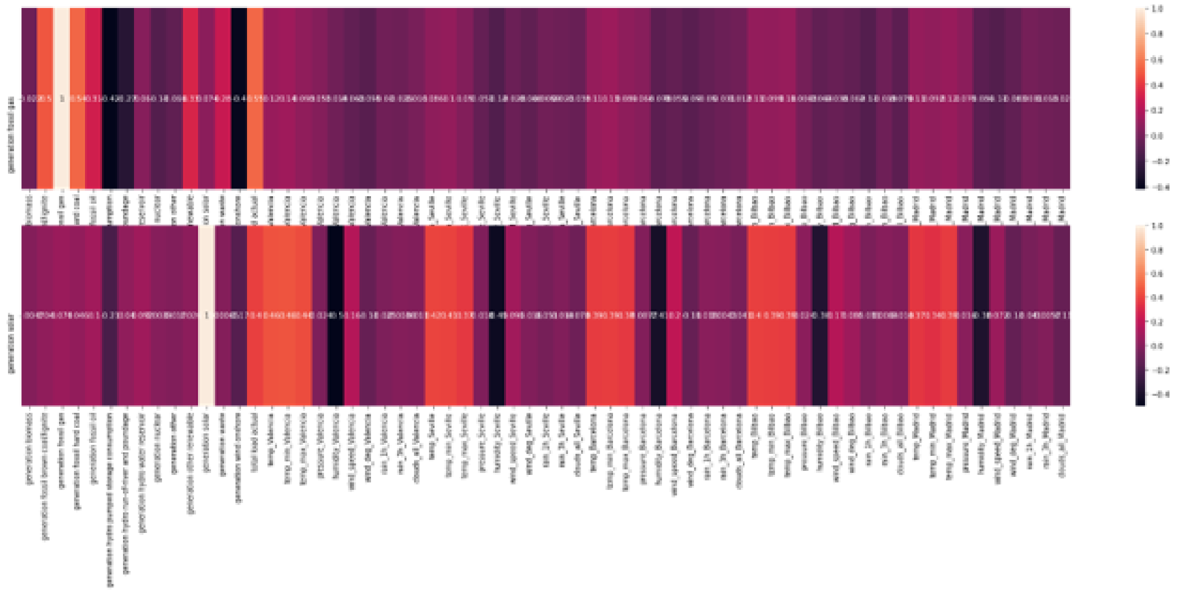

FIGURE 1.12: Correlation heatmap: generation fossil gas - other features, generation solar - other features.

Partial Dependence Plots

If we want to just see how the target is affected on average, the PD plot shows that it is not heavily affected by generation solar as seen in 1.13. However, with the individual lines in the ICE plot we do notice heterogeneous effects of each instance. The effects are diminished by the PD plot due to correlation.

Accumulated Local Effects Plots

We can clearly see that the ALE plot 1.15 shows a very different story from the PDP which is possibly due to the correlation of generation solar with the other variables in the data. It is interesting to see the effect of generating solar energy on the total price of energy. The only similarity between the two plots is that the price fluctuates irregularly upto 1000 MW of energy and constant reduction in predicted price on more generation. It is logical and affirming to see that the model predicts that the price of electricity decreases with the increase of generation from certain renewable sources of energy.

Final Interpretation

Similar to the generation of solar, when we compare the PDP 1.14 vs ALE 1.16 plot for generation of fossil gas, we see that the PDP suppresses the effect of the variable but shows a similar trend overall. It is interesting to note that the predicted value of price substantially increases with the increase in generation from fossil gas. The confidence interval for 95% is also much smaller than that of generation from solar. A simple conclusion from the two variables can be made that the predicted price is directly proportional to generation from non-renewable sources, but inversely proportional to generation from renewable sources. This supports the fact that we must increase generation of energy from renewable sources.

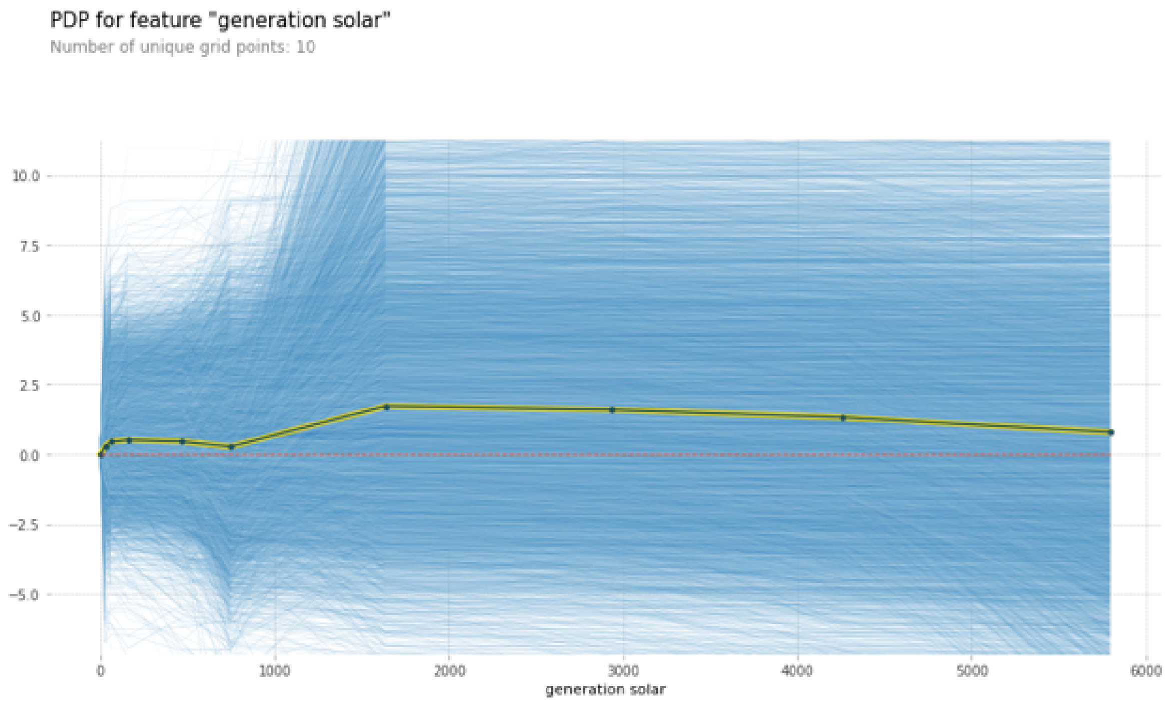

FIGURE 1.13: Partial Dependence Plot for “generation solar” including Individual Conditional Expectation lines.

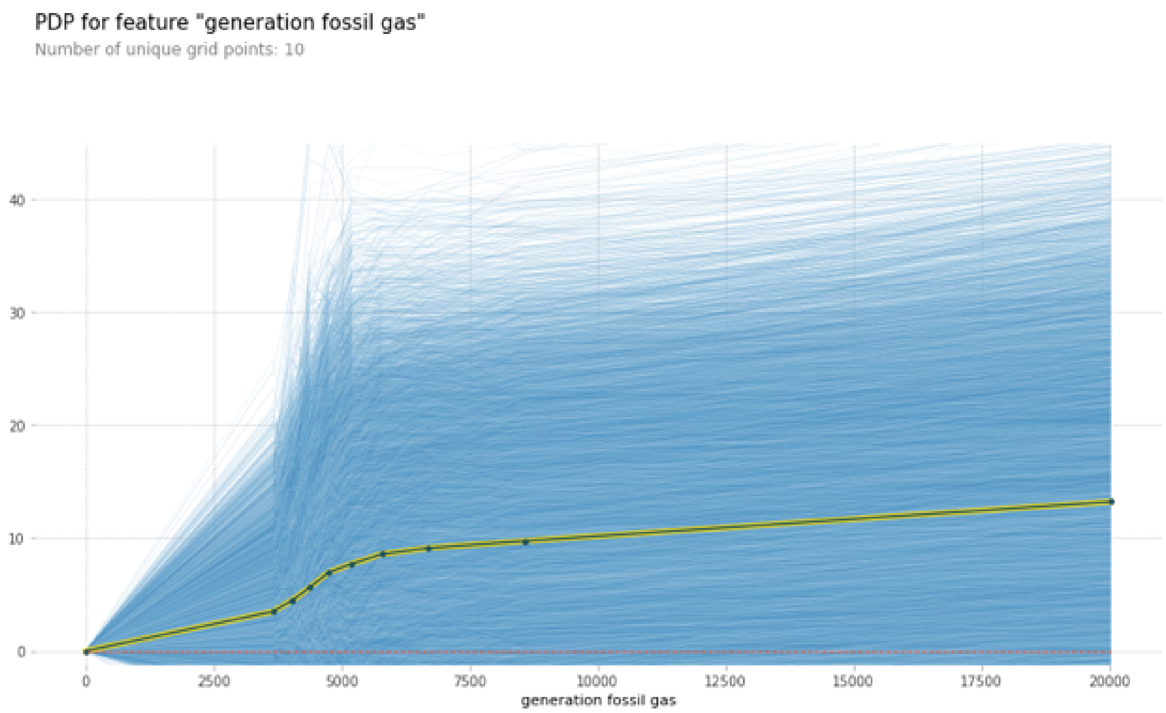

FIGURE 1.14: Partial Dependence Plot for “generation fossil gas” including Individual Conditional Expectation lines.

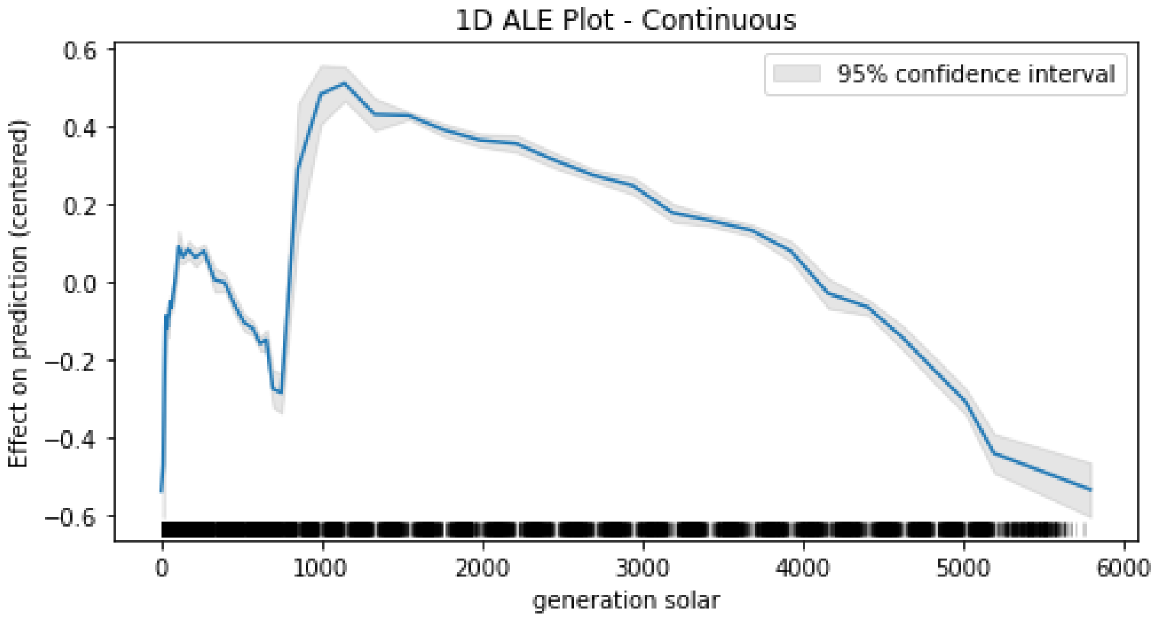

FIGURE 1.15: Accumulated Local Effects Plot for "generation solar".

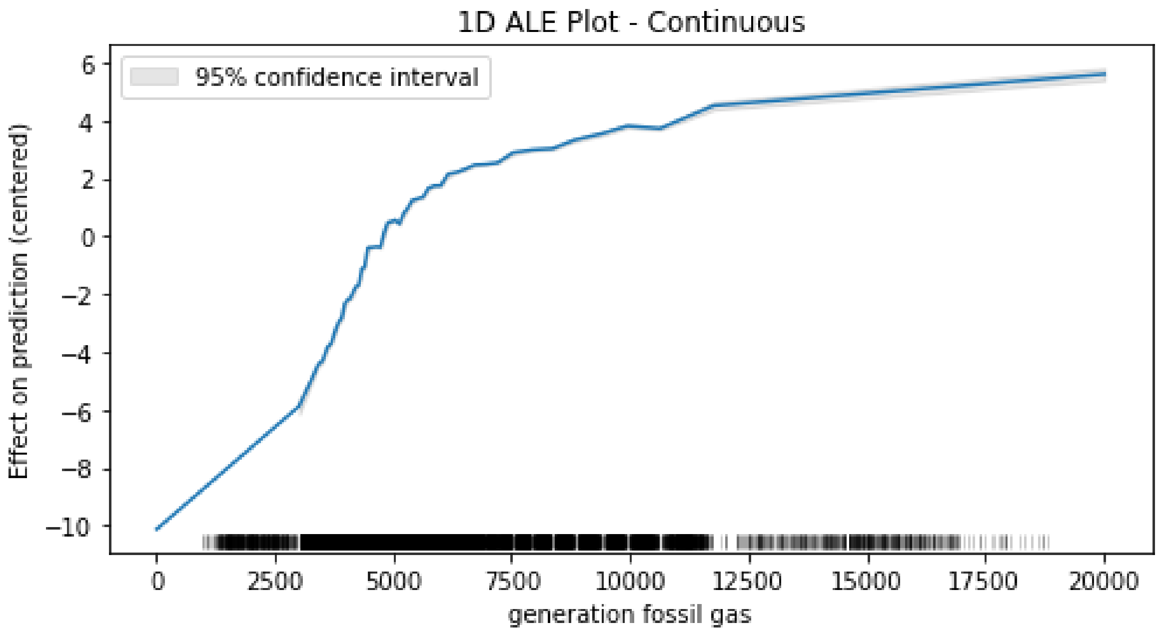

FIGURE 1.16: Accumulated Local Effects Plot for "generation fossil gas".

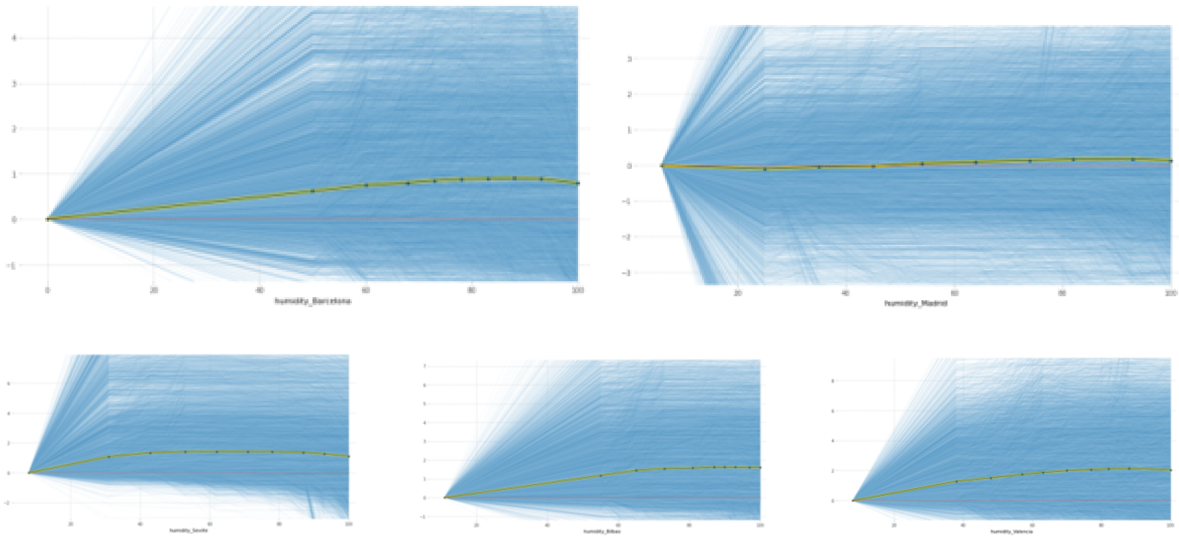

FIGURE 1.17: Individual Conditional Expectation plots for humidity.

1.4.4 Hypothesis 4

How would the individual data instances' prediction of price actual react to change in humidity and wind speed?

The intuition behind such a question was to investigate if humidity and wind speed had direct effects on the price. It is assumed that wind speed has a direct correlation with the amount of electricity generated from onshore wind belts, which is the largest source of renewable energy in Spain.

We will use Individual Conditional Expectation Plots to see if different instances contribute very differently to the model prediction.

Individual Conditional Expectation Plots

As shown in 1.17, the general trend in all five cities is that the price increases slightly as the humidity increase. However, if we zoom in to investigate individual instances, we can see heterogeneity. That is, for example, for some instances in Seville price goes down with an increase in humidity.

On the other hand, if we investigate ICE plots for wind speed in five cities of Spain as demonstrated in 1.18, a general downward trend is noticed in price as the wind speed increases. A close look at the plots helps to see a heterogeneous effect here as well. Some instances show that price may increase as wind speed rises.

One important issue to notice is that the general effects in ICE plots for wind speed, which are PDP lines, should not be trusted. Wind speed can be correlated with "generation wind onshore" which is the amount of energy generated in onshore wind belts. This makes the partial dependence plots biased and that is why we should use ALE plots if we want to see the average effect.

1.4.5 Hypothesis 5

Do the clouds have an impact on the generated solar? Are they useful for predicting actual price?

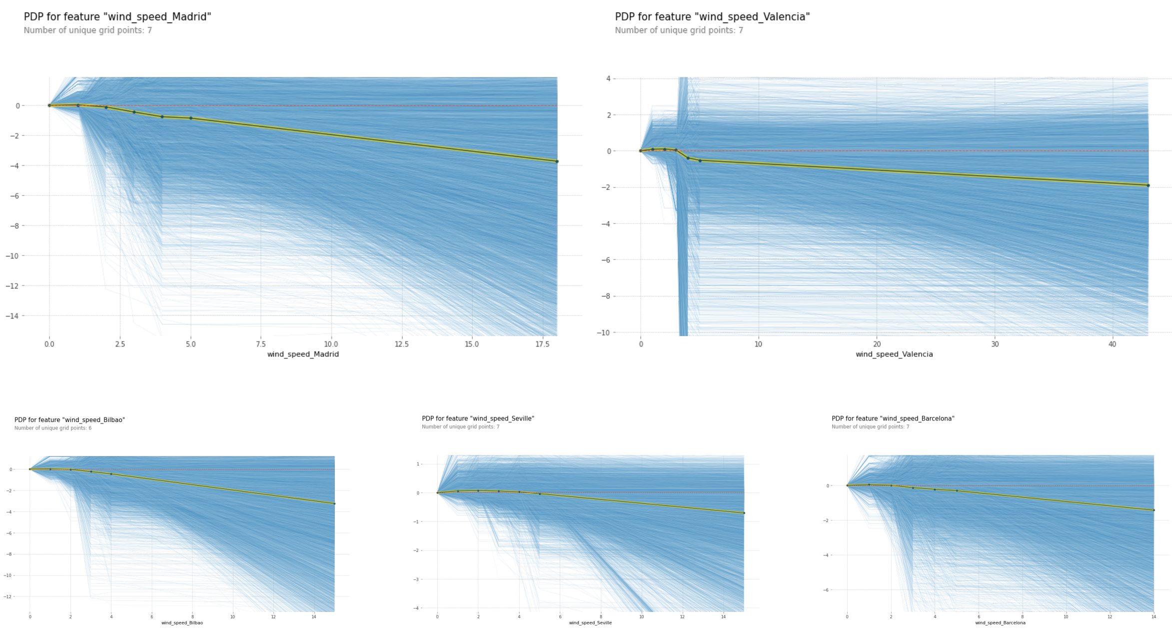

FIGURE 1.18: Individual Conditional Expectation plots for wind speed.

There are two parts of this question. It makes intuitive sense that the amount of clouds would be correlated with the amount of energy generated from solar. We would like to know if these variables interact with each other in the prediction of actual price by our model. The second part of the question is to further investigate if they are important to the model while predicting the target variable.

To get the strength of the interaction we will compute the H-statistics and PD interaction plots for each pair of variables. We will also see the correlation values between the variables to see if they make intuitive sense. Further, to observe the importance of the cloud features from each city we will observe ALE plots.

H-statistic

The heatmap in 1.19 goes against our assumption that the clouds feature from each city would interact with the amount of solar energy generated. The strength of interaction is almost negligible.

Correlation matrix

The first 5 rows of the correlation matrix in 1.20 are the cloud features from each city and generation solar is the last. The lighter the colour, the higher the correlation coefficient. From this we can see that the cloud features are not that correlated with any of the other features. But we can see that the last row has a lighter orange shade with a lot of features which are all temperature related features from each city. Also, we have marked the cloud features from each city with a black box in the last row which highlights the fact that generation solar is not correlated with any of the cloud features from these cities. We still want to see if there is some interaction visible in the PD interaction plots.

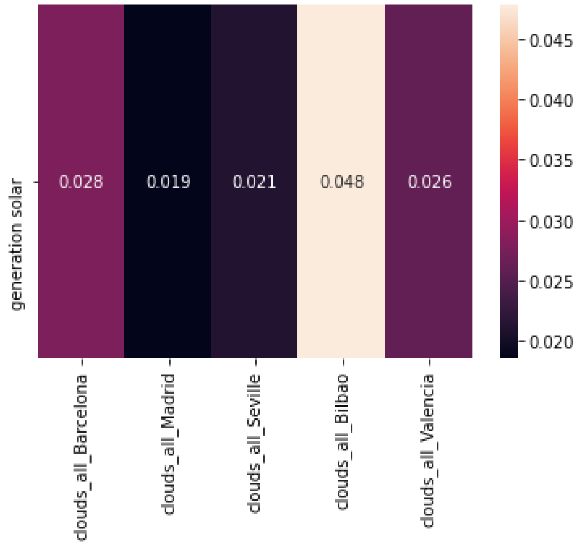

FIGURE 1.19: H-statistic for "generation solar" and clouds in 5 cities.

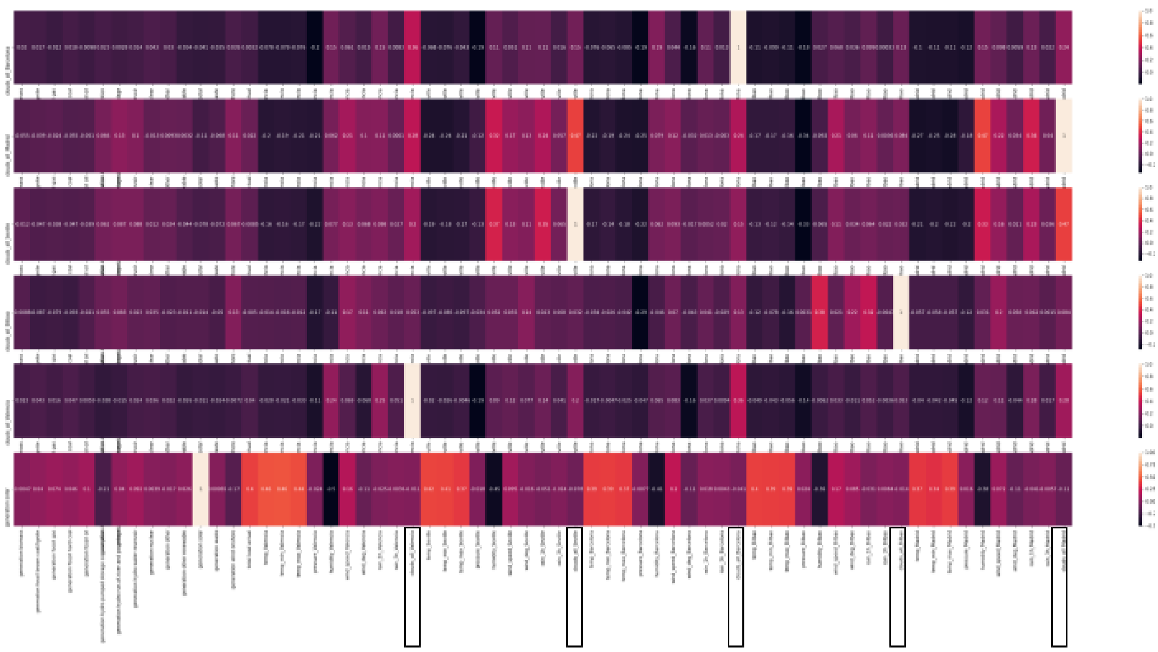

FIGURE 1.20: Correlation matrix: Cloud features from 5 cities - other variables, generation solar - other variables.

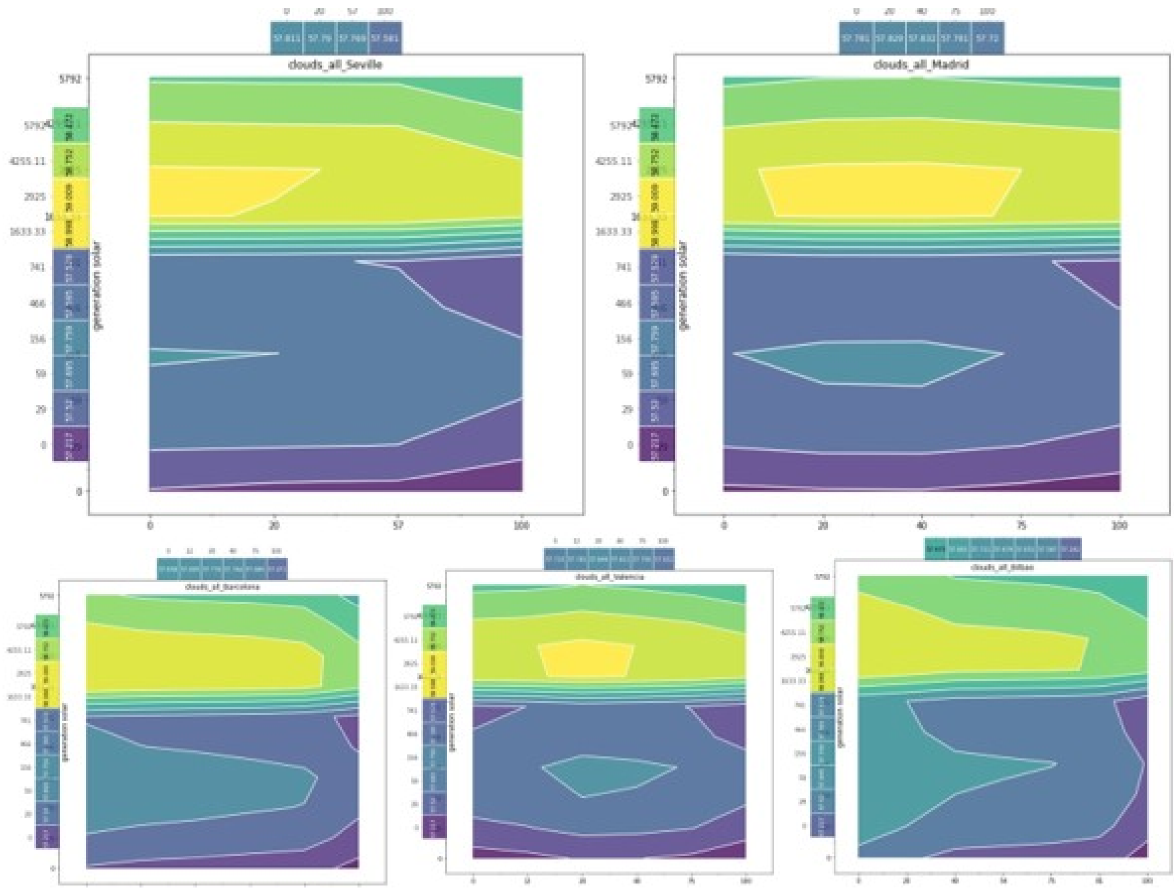

FIGURE 1.21: Partial Dependence Interaction plots for generation solar and clouds in 5 cities.

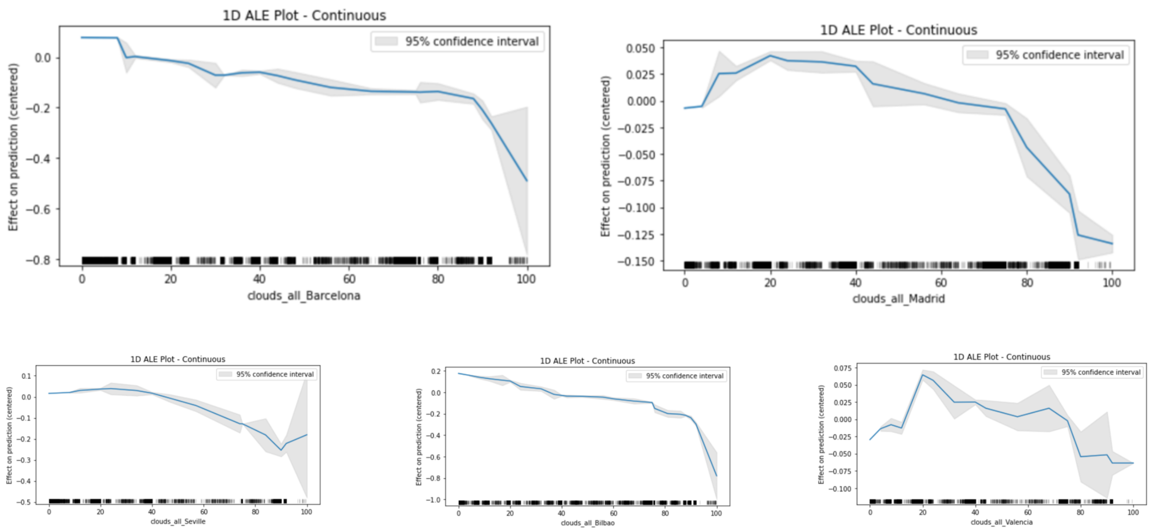

FIGURE 1.22: Accumulated Local Effects clouds in 5 cities.

Partial Dependence Interaction plots

From the plots in 1.21 we can draw an obvious conclusion that the target variable is not affected greatly by the clouds in the various cities. Also, the 95 % Confidence Interval is very wide which does not let us make any meaningful conclusions about these features.

Accumulated Local Effects

Next, we look at the ALE plots in figure 1.22 to see if cloud amounts in different cities play a role in how high the price is. We can see a decreasing trend for very high values of cloud amount. However, since the decrease is not very significant and the variation is too high we cannot make conclusive inference.

Final Interpretation

We observed that the solar energy generation does not interact with the amount of clouds in five cities of Spain. That might be because the solar energy plants are situated outside of these particular 5 cities. When it comes to whether the amount of clouds affect the price we can say that the effect is negligible.

1.4.6 Hypothesis 6

Does the weather in different cities impact total load actual differently?

Overall this may be a very vague question, but we want to see how the weather changes might affect the amount of energy being consumed by people in Spain. The idea behind this was that features like pressure and temp_max which have high permutation feature importance could possibly affect the total load.

To answer this we will consider the top 4 weather features in terms of PFI (pressure, humidity, max temp, wind speed) and calculate the H-statistic measures and also observe their respective PD Interaction plots.

H-statistic

Looking at the figure 1.23 we can see the extent of the interaction between weather features in the given five cities and total load. It is observed that there is a relatively high interaction between total load and wind speed in Bilbao. Also, the maximum temperature in Seville interact slightly with the total load.

Partial Dependence Interaction Plots

Figure 1.24 shows the interaction of total load with wind speed in Bilbao and the figure 1.25 demonstrates the interaction of total load with maximum temperature in Seville. In both plots, we can observe that for only high values of wind speed and maximum temperature some interaction takes place.

Final interpretation

The amount of electricity consumed by the people might be affected by the change in weather conditions but the interaction is minor. Higher temperatures could mean using more cooling systems which leads to more energy consumption.

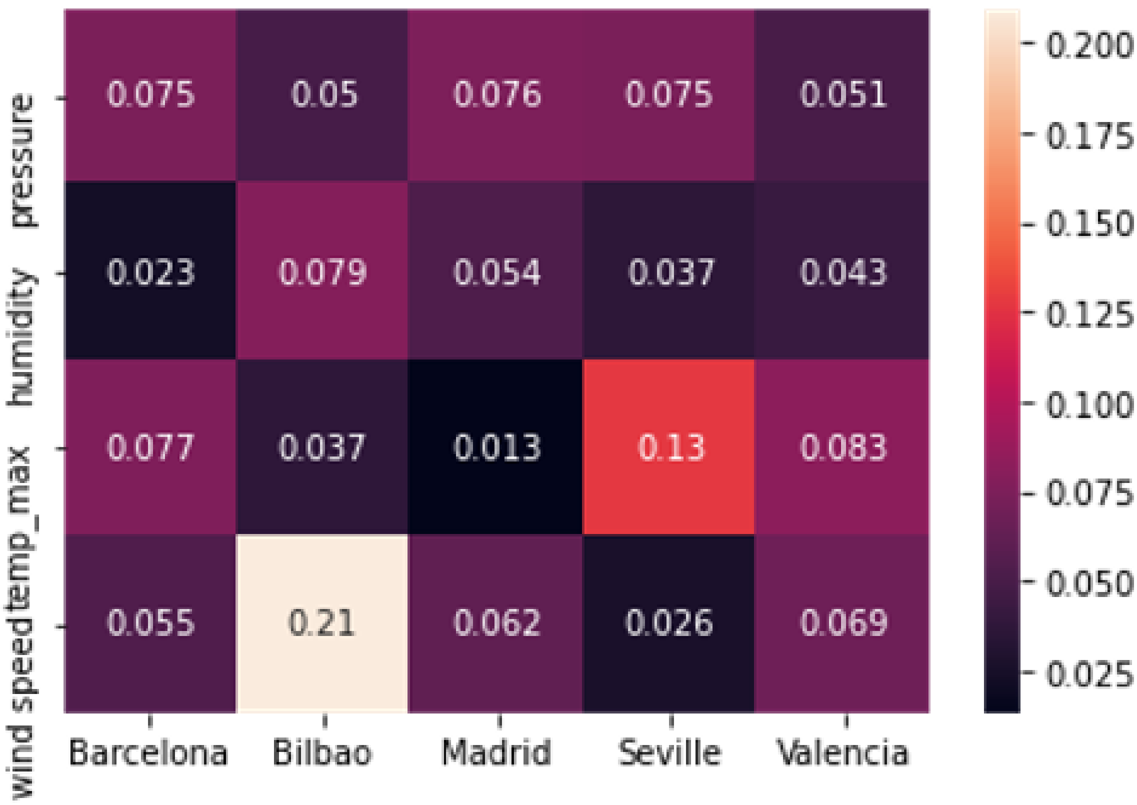

FIGURE 1.23: H-statistic for weather features in 5 cities and total load.

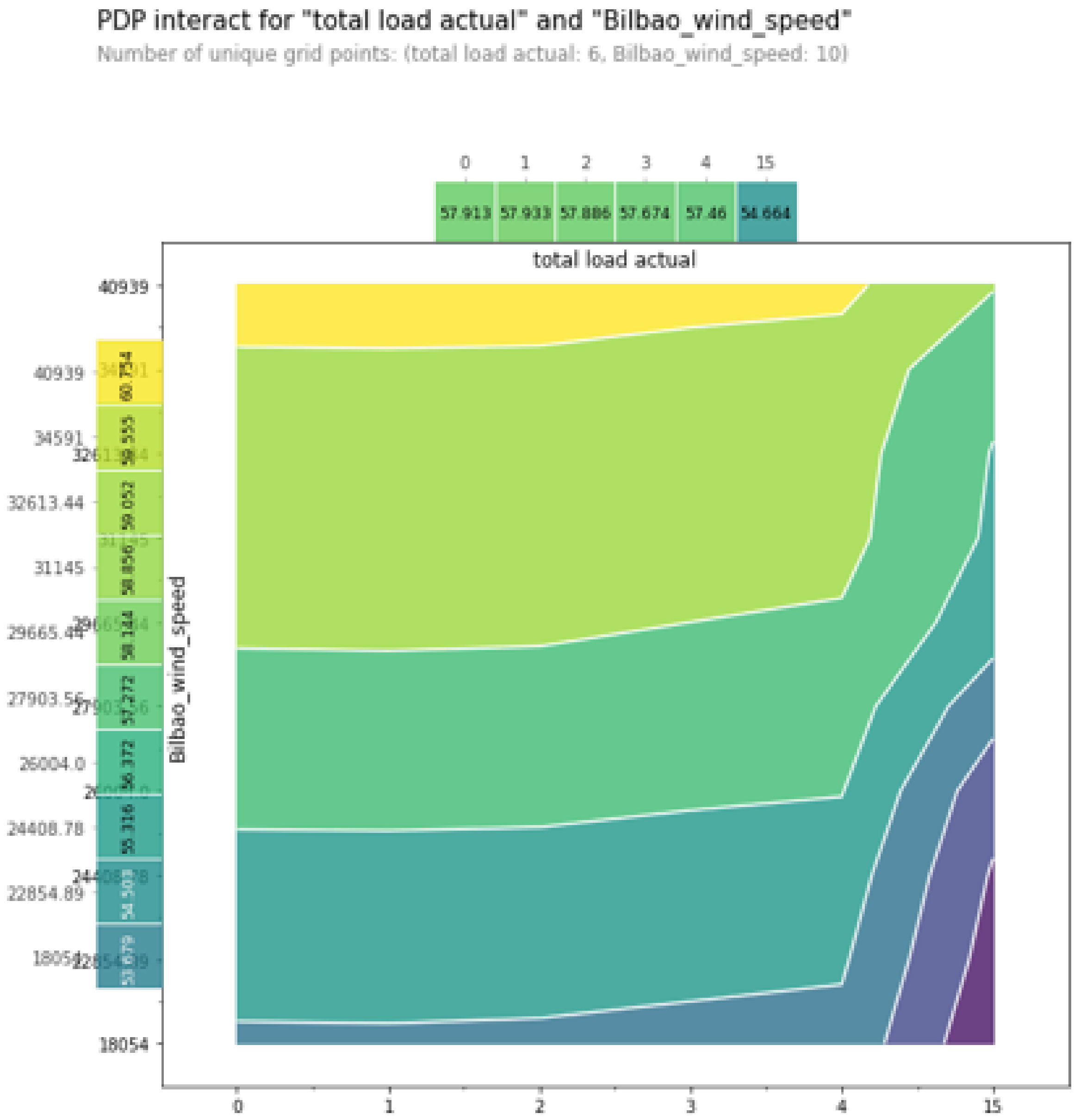

FIGURE 1.24: PD interaction plot for wind speed in Bilbao and total load.

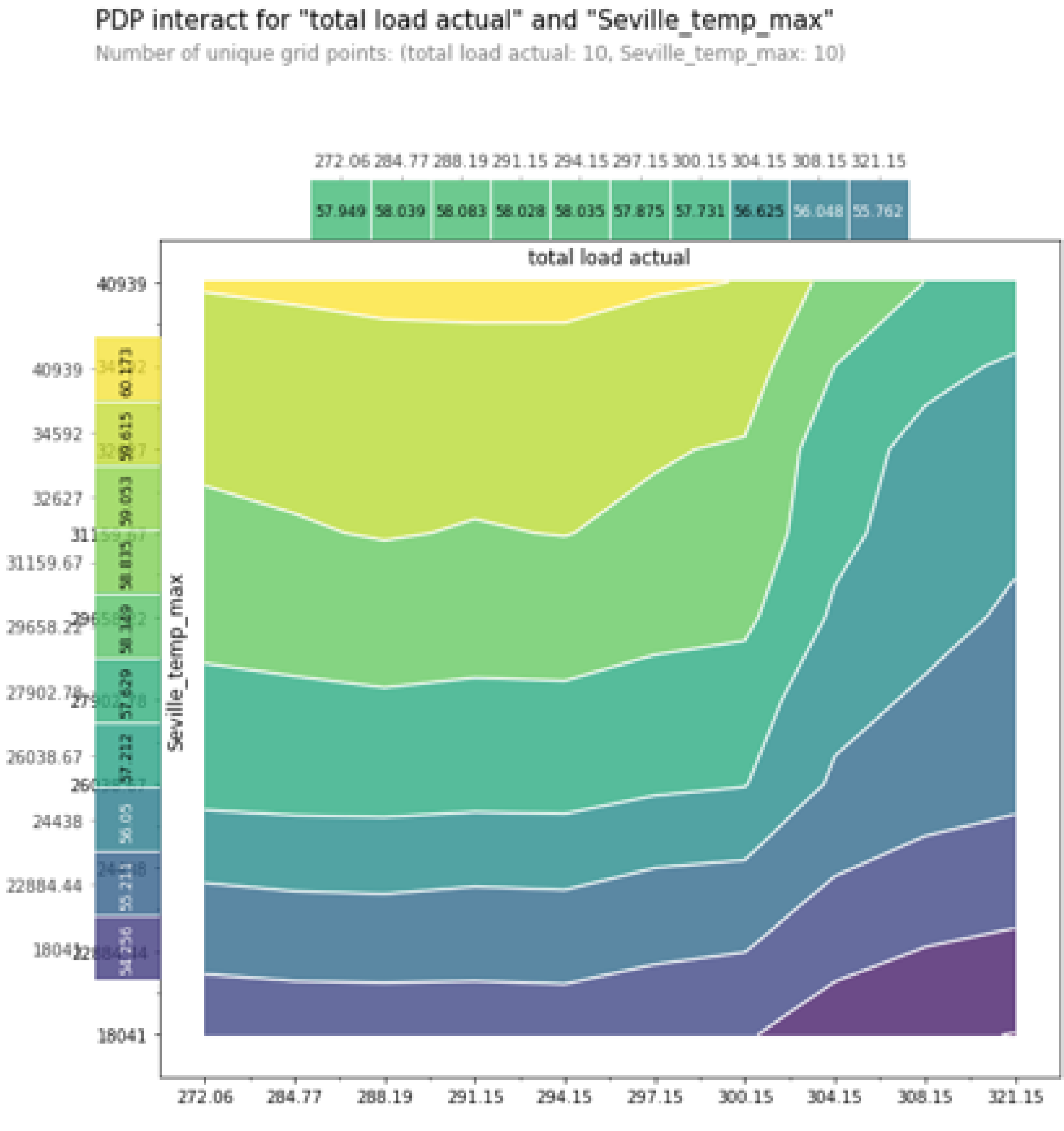

FIGURE 1.25: PD interaction plot for maximum temperature in Seville and total load.

1.5 Unanswered Hypothesis

While defining the scope of our project we wanted to answer a dew more hypotheses which we were unable to due to certain limitations. We discuss these hypotheses and possible implementations in the future.

1.5.1 Hypothesis 7

At least how much should the generation biomass change to have the price actual for observation z below the mean?

Two possible methods to answer this question was producing counterfactuals and anchors. However, there are no perfect implementations of these methods in python for a regression setting.

1.5.2 Hypothesis 8

Can we reduce the generation fossil (all combined) for the price actual to remain the same?

Similar to the previous unanswered hypothesis, using anchors would have been ideal to answer this as we could get a feature range for every particular instance for which the price does not change. Due to the lack of an available library, we tried the following methods to answer it.

We discretized the target variable and placed it in bins to convert our regression problem into a classification setting. Using this, we were able to formulate anchors for individual predictions but we have not published these results as it does not fit in with our framework of model-agnostic interpretation. The results can however be found in the appendix A.

1.6 Future Work

The scope of exploring this Energy prediction dataset is large and this research work is not exhaustive. In light of the possible areas of work, we have listed some of them below:

- Since we are comparing similar features from various cities, it would be interesting to apply the Relative Feature Importance measure as mentioned in "Relative Feature Importance".
- Extend the use of Counterfactual and Anchors to python for regression.
- Generate more relevant hypotheses like comparing effects of non-renewable sources with renewable sources of energy.

Appendix A

Implementation of Project

A.1 Python notebook with code and results

The iPython notebook with reproducible results can be found here or in the following github project.

Bibliography

Brucke, Seppe vanden. *Discovering Interaction Effects in Ensemble Models*. URL: https://blog.macuyiko.com/post/2019/discovering-interaction-effects-in-ensemble-models.html.

Jhana, Nicholas (2019). "https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather". In:

König, Gunnar et al. "Relative Feature Importance". In: *arXiv.org* (). URL: arXiv:2007.08283.

Molnar, Christoph (2020). *Interpretable Machine Learning ; a guide for making black box models explainable*. Leanpub.

Strobl, Carolin et al. (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC Bioinformatics* 8.1. DOI: 10.1186/1471-2105-8-25.