IST 687 final project report

Francisco Franco Arenas

2023-11-27

1. introduction

2. Data reading and merging approach

3. Building a Model for Prediction

The assignment consisted of predicting hourly energy consumption for each county in the month of July, assuming a temperature increase of 5 degrees Celsius. To this end, we first aggregated the individual household energy consumption at the county level.

The initial approach treated this problem as a panel data regression with fixed effects, considering counties as the cross-sectional variable and the hour as the time variable. However, this approach yielded an R-squared of 0.13, resulting in an ineffective model.

The second approach involved modeling a linear regression, where we extracted the hour and the day of the week as independent features. The regression model is as follows:

$y = a + bX + e$

where y represents the electricity consumption, and X includes the temperature, the hour, and the day of the week. This regression produced the following output:

To compare the predictive capability with other models, we conducted a train-test split and computed performance metrics from 10-fold cross-validation:

The metrics indicate that this is a suitable model. Nevertheless, we postulated that there might be non-linearities associated with temperature changes, as air conditioners, for instance, switch off below a certain temperature and on above a certain threshold. Consequently, we retried the regression with a quadratic term for temperature. The results were as follows:

Some metrics suggest a marginally better model; however, the improvement is minimal and does not justify the increased complexity of the model.

Another potential approach is using a tree-based method, such as random forest regression. However, we do not prefer this method for our task, which involves extrapolation. A regression tree may struggle with extrapolation, as it can only make predictions within the range of the original data.