Project ML

Myroslava Hrechyn

1 Introduction

The importance of the power consumption forecasting can not be underestimated in the modern world. From government authorities to private companies to small consumers, it is crucial to predict the hourly power consumption for supplying the demand. In this project we are working with electricity consumption data in Slovakia from 2023 to the last full currently available week from the ENTSO-E Transparency Platform, which provides data about power generation and consumption in Europe. Data and project's code can be found on the GitHub repository. The power consumption has both daily and weekly seasonal patterns (figure 1).

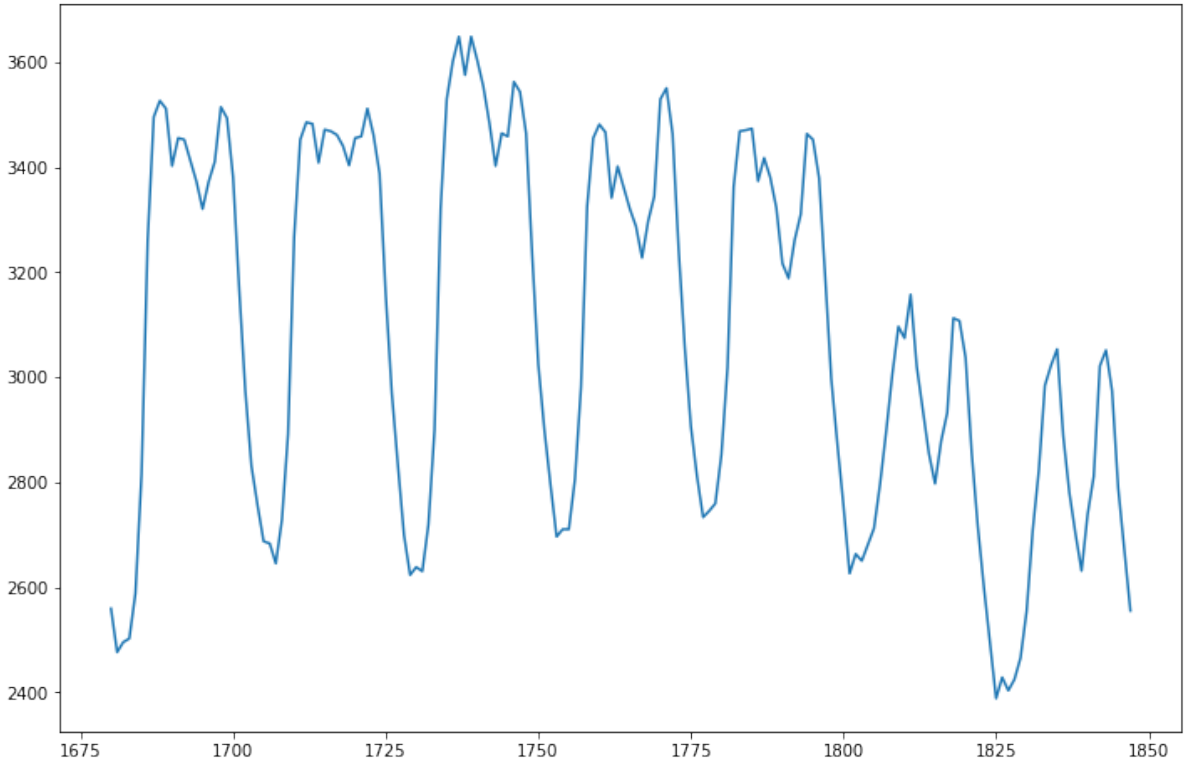

Figure 1: The hourly power consumption over the course of one week

In addition, the profile is highly affected by public holidays. In this project we cover all these effects by adding variables that indicate workday and weekend, public holiday,

hour of the day and consumption from the previous day at this time, as these values have a significant importance for predicting today's consumption. However, this approach is suitable only for the short-term predictions, as in the giving moment we will only know yesterday consumption and for the later days will have to use our predicted value due to the lack of real measurements, which can cause notably worse prediction for the days that are more distant in time.

2 Methods

We have divided our dataset into 3 parts. The last available week, 22 to 28 of January, is used measuring the quality of the prediction of the trained model. The rest of the data is separated to training (80%) and validation dataset (20%) without any shuffle, since we are working with the time series and want to prevent data leaking.

For the purpose of prediction of the power consumption we will use eXtreme Gradient Boosting (XGBoost). It is the method that utilizes parallel tree boosting, supports setting of various parameters and can be used for either regression or classification problems. It starts with the weak model and gradually improves it using wrongly predicted values.

As we are working with the regression task, we have chosen root mean squared error (RMSE) as our objective function. Here we first train a model with the default parameters, then we explore adding regularization terms to improve model's generalization ability. We have compared L1 regularization with the values for the parameter alpha of 2, 10, 100 and L2 regularization with the same values for the parameter lambda.

Regularization	**Parameter value**	**MAPE**
L1 regularization	2	13.224
L1 regularization	10	13.232
L1 regularization	100	13.158
L2 regularization	2	13.297
L2 regularization	10	13.212
L2 regularization	100	12.968

Table 1: Comparison of the trained models using validation data

We have compared trained models using mean absolute percentage error (MAPE) on the validation data. As we can see from the table 1, the best model was with L2 regularization with lambda set to 100.

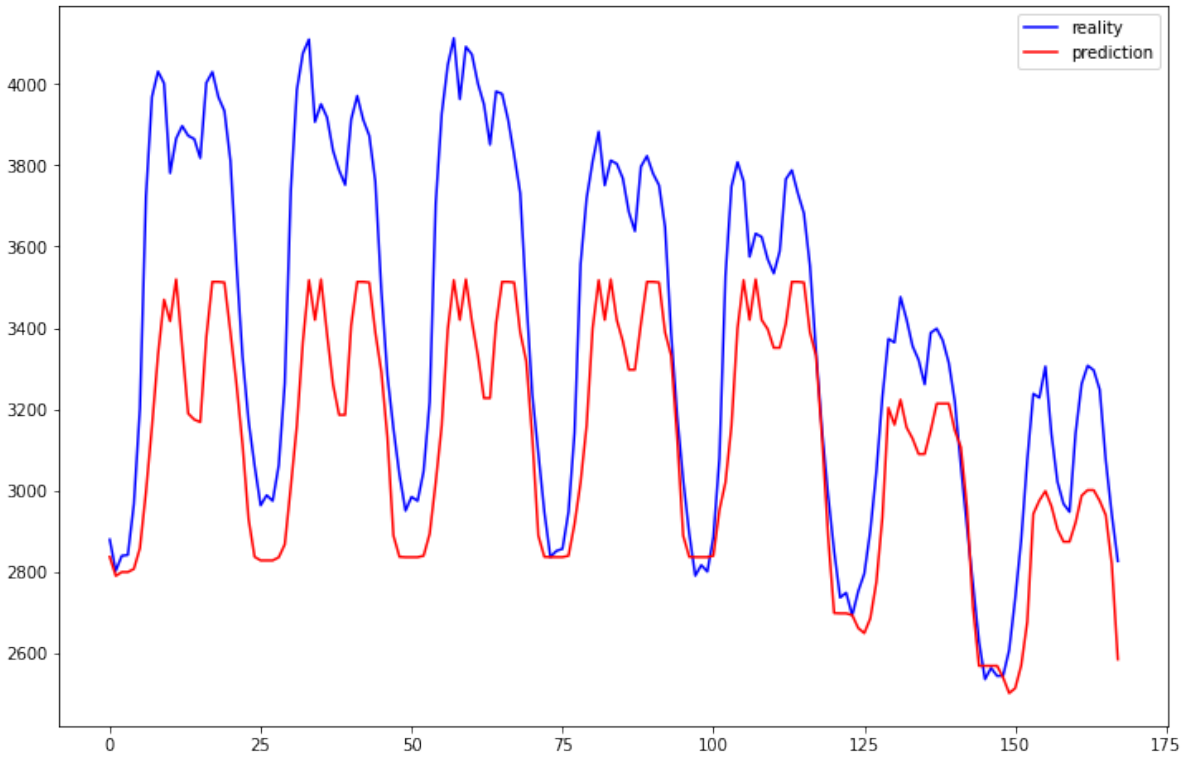

Figure 2: The result on the testing data. Red curve is predicted, blue curve is reality.

3 Results

Finally, we evaluate the quality of the prediction on the unseen testing data from 22.01.2024 to 28.01.2024. The comparison of the real power consumption and our prediction is shown in the figure 2. As we can see, the model has captured both seasonal effects, however, the predictions for the workdays are quite lower than reality. The MAPE of the final result is 8.03%.

4 Discussion

As we are working with the time series, using statistical models such as SARIMA, Holt-Winters method or BATS/TBATS, which is a combination of Box-Cox transformation, ARIMA and exponential smoothing and can model time-series with multiple seasonal pattern as in our case, can also be an object of exploration. Given the fact that power consumption by companies and households is highly correlated with the temperature outside, adding this parameter could also improve the quality of our model. In that case statistical models mentioned above would not be useful, as they can not work with explanatory variables, however, we can easily add another variables to the XGBoost. The further work also includes more detailed regularization parameter selection.