Residential power usage prediction - regression methods

Boxi Lin, June 20

1. Notation

- $t \in \{2003, 2004, \cdots, 2016\}$: year;
- $w \in \{1, 2, \cdots, 7\}$: weekday;
- $m \in \{1, 2, \cdots, 12\}$: month;
- $d \in \{1, 2, \cdots, 31\}$: day;
- $h \in \{0, 1, \cdots, 24\}$: hour;
- $\mathbf{X}_{\{t,m,d,h\}}$: vector of weather information at year t - month m - day d - hour h, including precipitation, snowfall, snow mass, air density, ground-level solar irradiation, top of atmosphere solar irradiation, cloud cover fraction.
- $Y_{\{t,m,d,h\}}$: Residential power usage at year t - month m - day d - hour h;
- y_t : annual residential power usage in year t
- $u_{\{t,m,d,h\}}$: Power usage of all sectors at year t - month m - day d - hour h;
- u_t : annual power usage for all sectors in year t

2. Problem

- Given $\mathbf{X}_{\{t,m,d,h\}}$, y_t and $u_{\{t,m,d,h\}}$, predict $Y_{\{t,m,d,h\}}$, say $\hat{y}_{\{t,m,d,h\}}$.
- SSC evaluation metric: $\frac{1}{n}\sum_t |y_t - \sum_{m,d,h} \hat{y}_{\{(t),m,d,h\}}|$, where n is the number of years (14), $\hat{y}_{\{(t),m,d,h\}}$ is the above prediction without using data in year t.

3. Issues and assumption

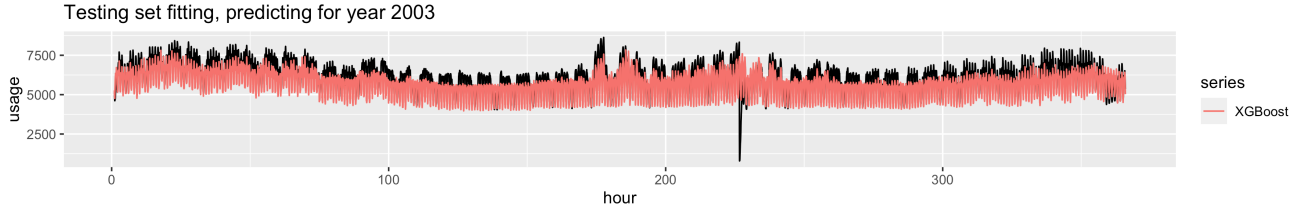
- We assume for each year t, $Y_{\{t,m,d,h\}}/u_{\{t,m,d,h\}} = y_t/u_t$, i.e. the propotion of hourly residential power usage to the total usage is constant within year t.
- We think the above SSC metric makes almost 0 sense so I used my own metric, let's call it mean absolute error (MAE) at hourly resolution: For each year t, $\frac{1}{M(t)}\sum_{m,d,h} |y_{\{t,m,d,h\}} - \hat{y}_{\{(t),m,d,h\}}|$, where $M(t)$ is the number of hourly records in year t (usually $M = 24 * 365 = 8760$ for non-leap year and 8784 for leap year). But we'll still calculate the SSC metric.
- Suppose we are at 2020 now and want to build predictive model for future power usage. Pratically speaking, we would only have historical data at and before current time point and shouldn't have assumed that hourly weather in 2021is available. But for machine learning (regression) setting here, we assume that we have the access to any (future) weather information.

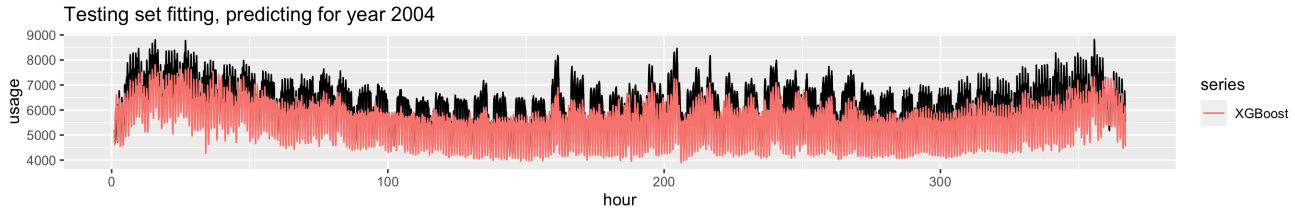
3. Regression setting

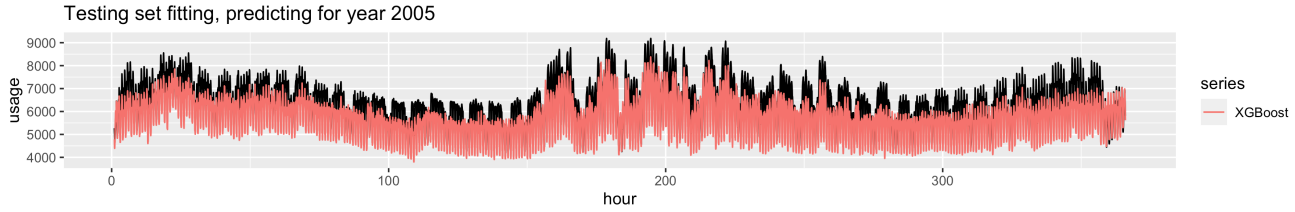
- Response variable: $y_{\{t,m,d,h\}} = u_{\{t,m,d,h\}} y_t / u_t$
- Features ($p = 49$, all numerical):
    1. Year
    2. Dummy variables for weekdays
    3. Dummy variables months
    4. Dummy variables for hours
    5. Weather information
- Methods: XGBoost (https://xgboost.readthedocs.io/en/latest/) in R
- Steps:
    - Drop data in year 2003, details see the notice below;
    - For each year to validate i in $\{2004, \cdots, 2016\}$,
        1. Leave out data in year i; use data in year $(i-1)$ and $(i+1)$ as testing set; use all the rest years as training set;
        2. Training and testing model; objective function is Relative Mean Squared Error; parameter setting: (eta = 0.5, max.depth = 3, early.stop =50, max.iter=800, all others are by default)
        3. Predict model with data in validation dataset, calculate MAE at hourly resolution;
    - Summarize the results and calculate SSC metric.
- Notes:
    1. In above steps, year 2003 is excluded from the analysis. According to the exploratory data analysis there are outliers points in August 13-14, 2003 due to Northeast blackout of 2003 (https://en.wikipedia.org/wiki/Northeast_blackout_of_2003).
    2. For validation of year 2003, use 2004 and 2005 for testing; for year 2004, use year 2005 and 2005 for testing; for year 2016, use year 2014 and 2015 for testing.
    3. The multi-level seasonality of data is (hopefully) captured by the dummy variables for date.

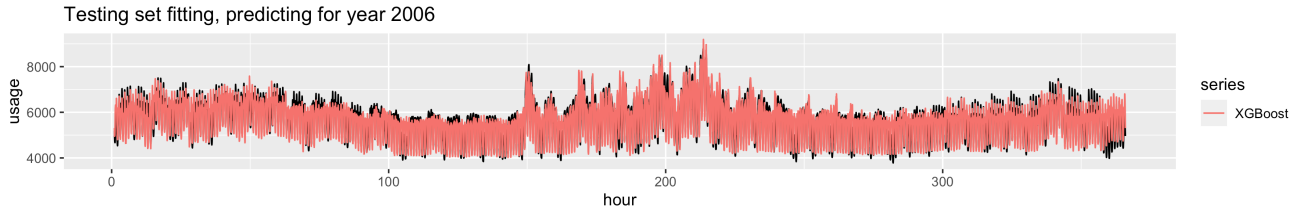
4. Results

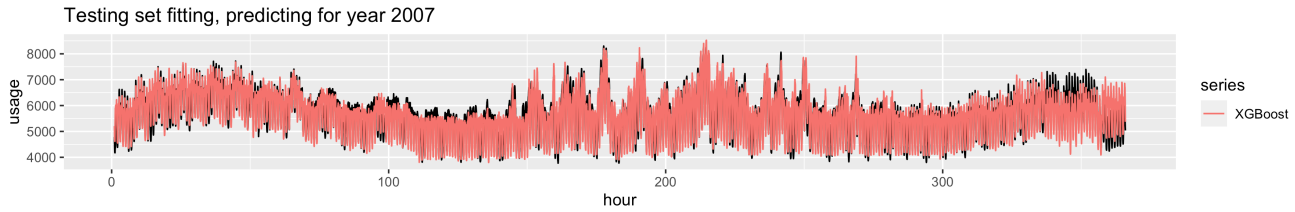
4.1 Graphical check

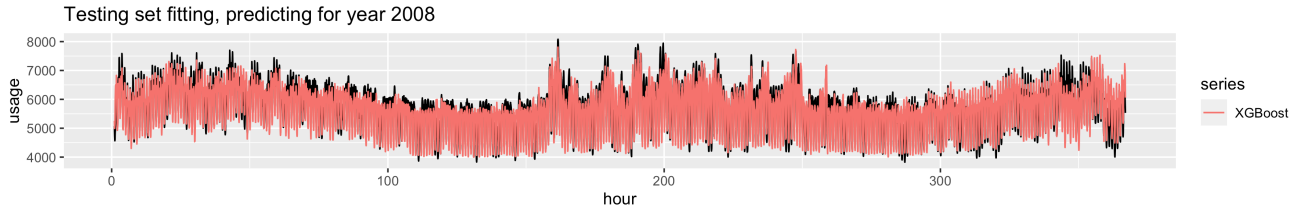


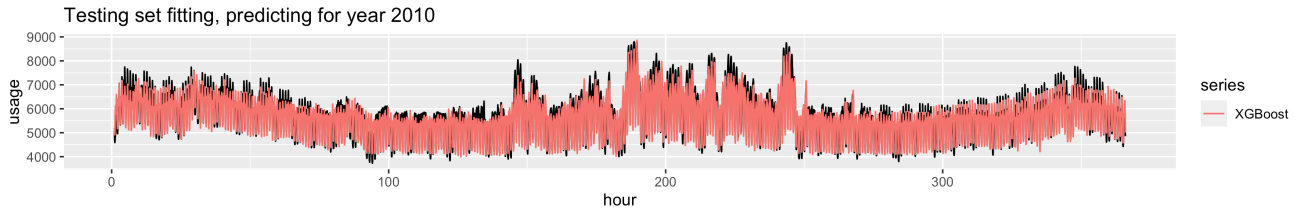

Testing set fitting, predicting for year 2004
usage
hour
series
XGBoost

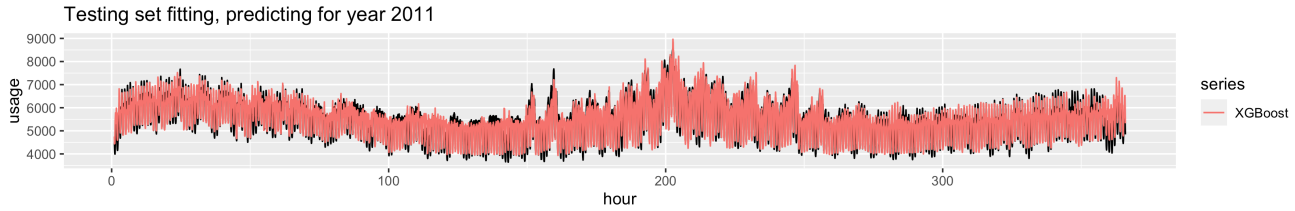

Testing set fitting, predicting for year 2005
usage
hour
series
XGBoost

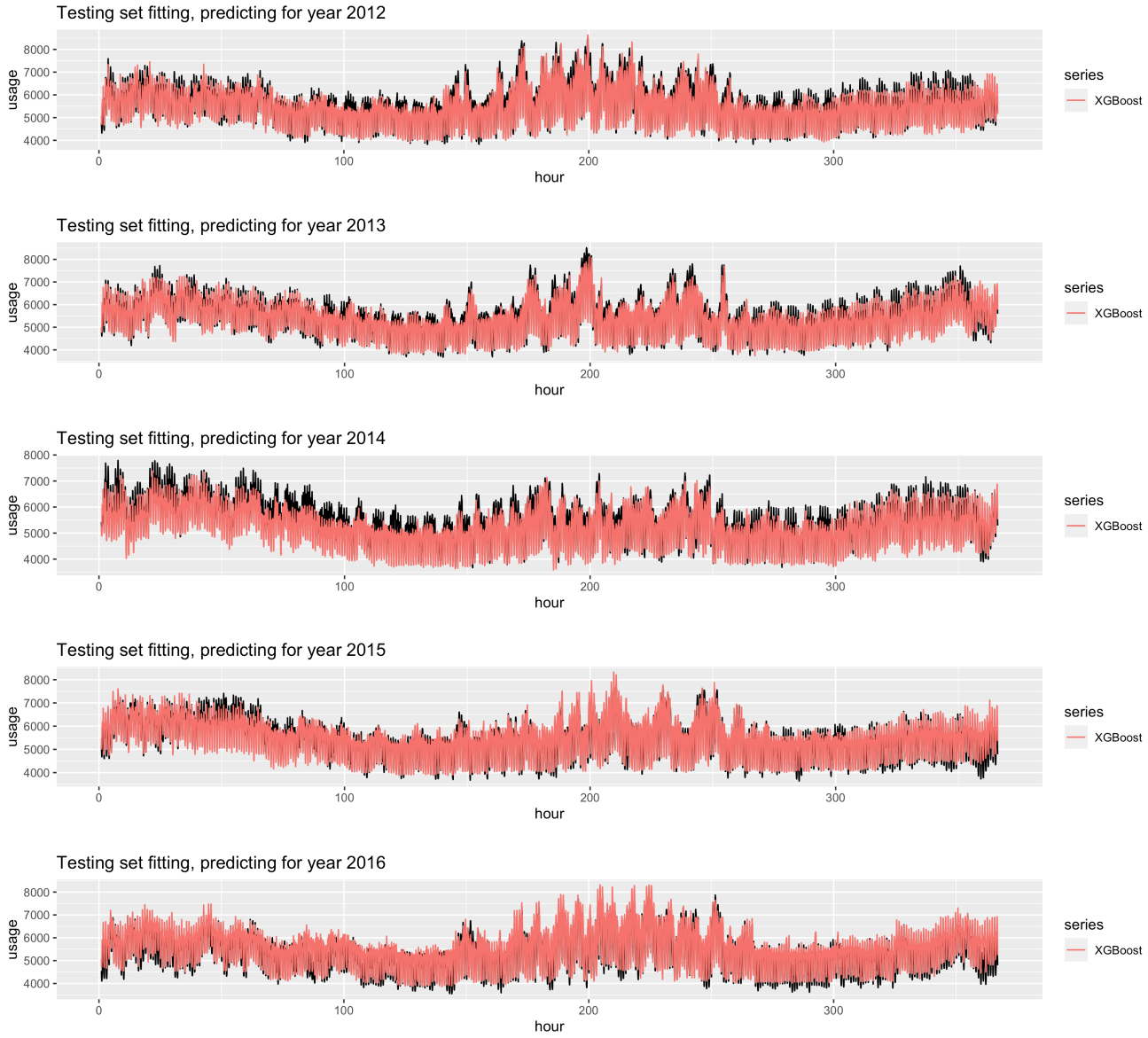

Testing set fitting, predicting for year 2006
usage
hour
series
XGBoost


Testing set fitting, predicting for year 2007
usage
hour
series
XGBoost

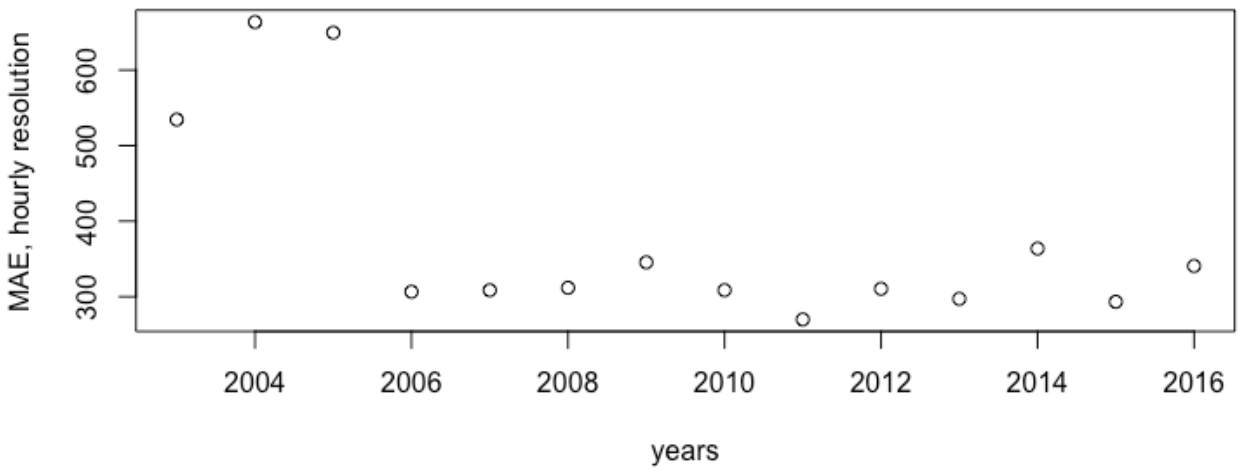

Testing set fitting, predicting for year 2008
usage
hour
series
XGBoost


Testing set fitting, predicting for year 2010
usage
hour
series
XGBoost

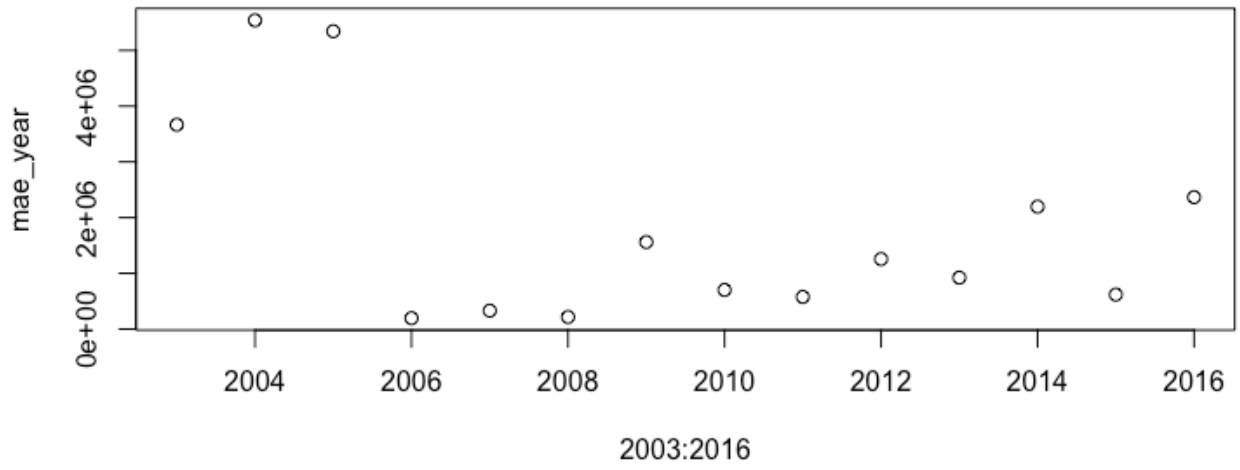

Testing set fitting, predicting for year 2011
usage
hour
series
XGBoost



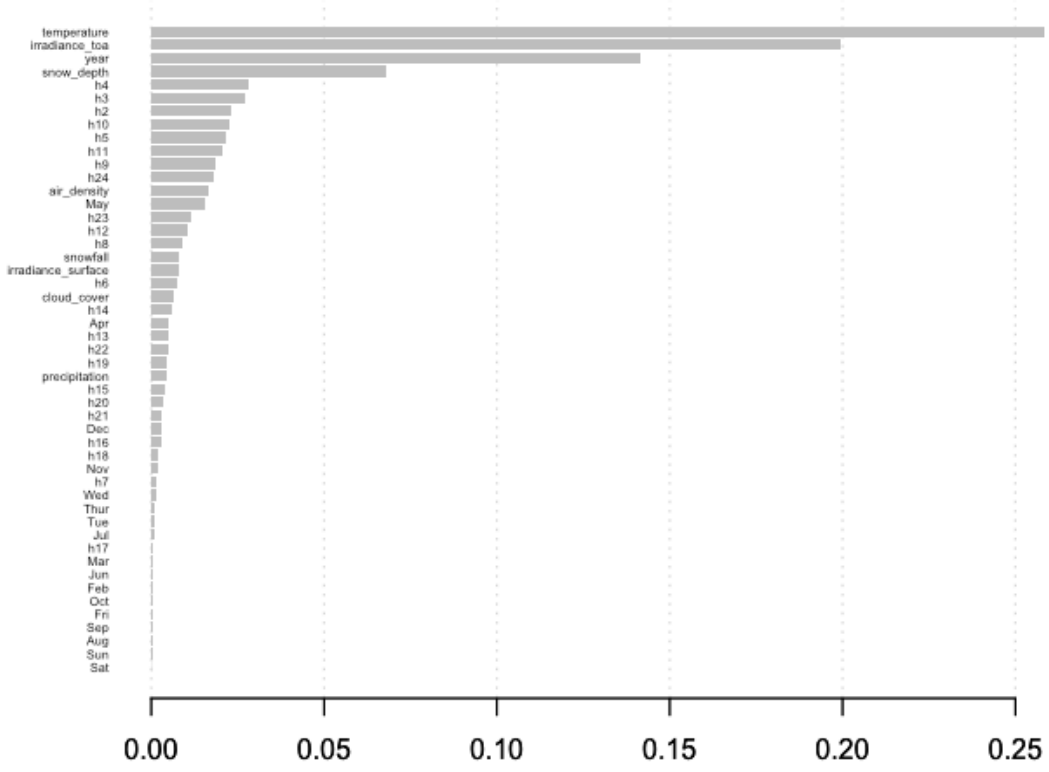
4.2 MAE at hourly resolution



4.3 SSC metric



4.4 Features importance



4.5 Summary

1. The prediction is generally good for year between 2006 to 2016; both the general trend throughout the year and seasonality (hourly, month) are captured;
2. The prediction is pool for year 2003 to 2005 and clearly underestimate the usage. Multiple possible reasons:
    - For 2003, due to the outliers in August, we expect to see failure of fitting;
    - The power usage in these years are possibly systematically different from the other years (policy change, energy/indstrial upgrading);
    - The $Y_{\{t,m,d,h\}}/u_{\{t,m,d,h\}}$ may be lower than we thought;
    - measurement error
3. According to feature importance plot
    - "temperature", "irradiance_toa" and "snow" are important features that are predictive for the usage of residential power;
    - "year" is important, indicating the heterogeneity of power usage over different years;
    - there exits clear hourly seasonality; some signals for monthly seasonality, although it may be weaken by weather due to their strong correlation; there are no clear pattern of weekly seasonality

5. Discussion

1. Changing point analysis/feature before and after year 2005;
2. Outlier analysis to include year 2003;
3. Randomly sample training and testing;
4. More practical and realistic forecasting manner, i.e. avoid use future data to predict past
5. Time series analysis