Assignment

Devosmita Chatterjee [*]

Department of Mathematical Sciences, Chalmers University of Technology

[*] Correspondence:tel: +46 709669519, E.mail: chatterjeedevosmita267@gmail.com

Discussion Task 1

Describe how you'd design a more comprehensive system for estimating energy in the case when input data was missing.

The given code task represents a quick initial ingestion of data into a larger system. If in this larger system, I have a large amount of missing data, then I would avoid using simple imputation techniques such as mean, median, mode and try to use model based imputation technique such as missForest method. MissForest algorithm works best with large datasets unlike mean median and mode imputation methods.

MissForest Method is a missing data imputation method with random forests. It can perform imputation for large amount of missing observations in the data. It is less biased than other imputation methods since it is based on random forests. It can handle both continuous and categorical data simultaneously that works well with both data missing at random and not missing at random.

In my Master thesis, the performance of the missForest imputation method is evaluated using error measures (normalized root mean squared error NRSME for continuous data and percentage of erroneous categorical entries PEC for categorical data) as shown in figures 1 and 2.

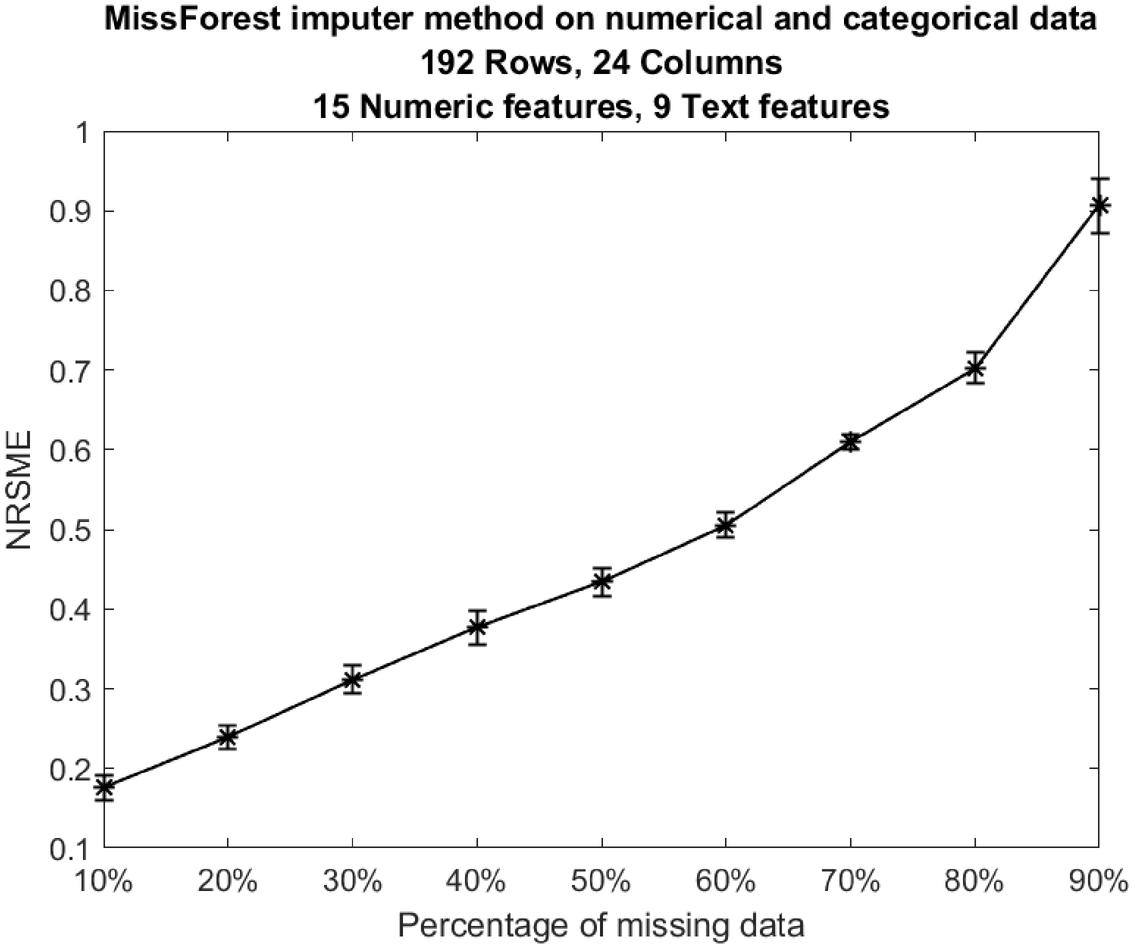

Figure 1

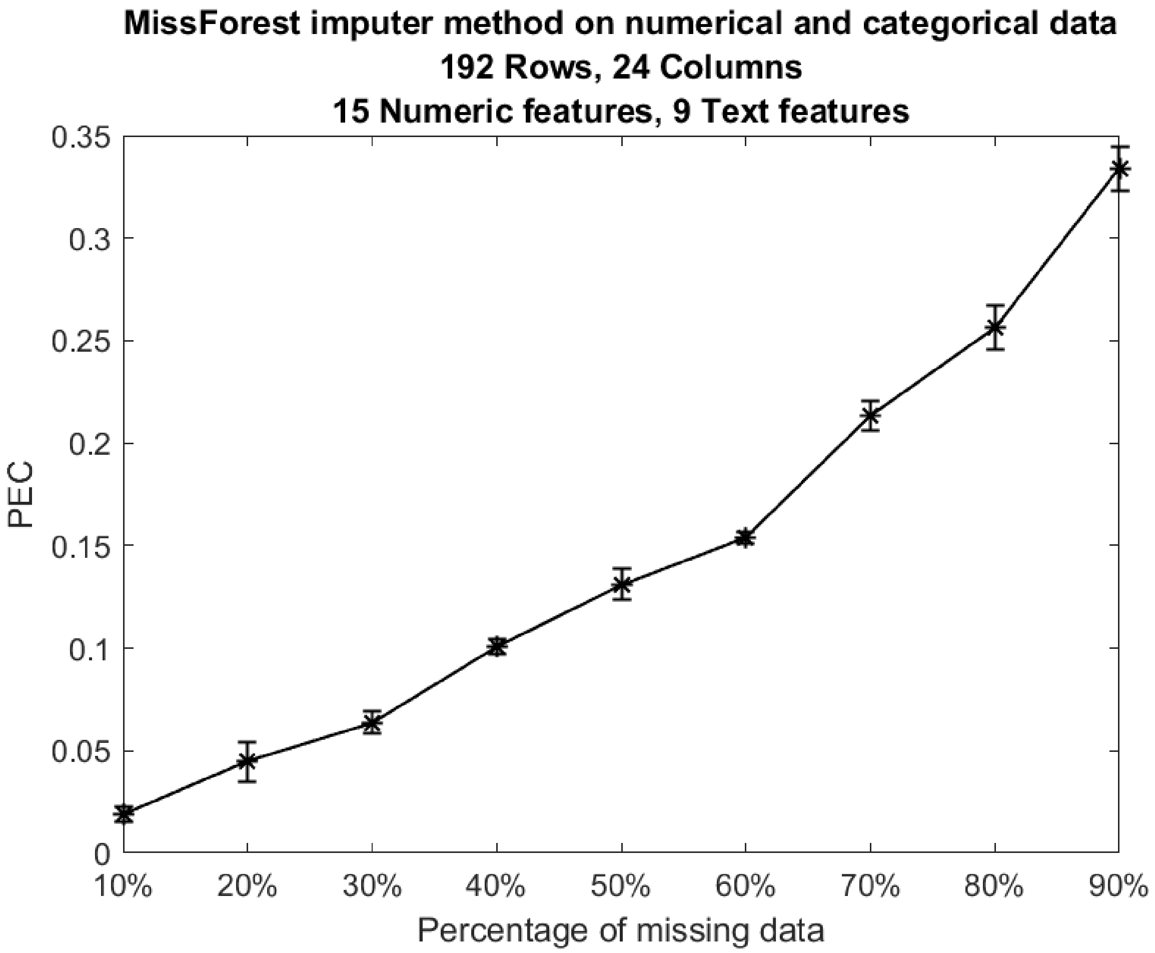

Figure 2

How would you change the design of your routine (and its output) if you knew its output was going to be sent to a dedicated system for estimation of missing data?

If I know the output was going to be sent to a dedicated system for estimation of missing data, then I would concentrate more on data cleaning to solve the data quality issues as well as detect outliers. I have found univariate outliers in given code task. In this case, I would also focus on multivariate outlier detection such as leverage statistics, local outlier factor, DBSCAN.

Discussion Task 2

1. Meter-resets. State meters are occasionally ”reset” down to 0, so the data series will look something like ‘9837.1, 9837.4, 0.2, 0.7‘. This is not an error, the meter will continue ticking upwards normally after the reset.

The state meters data series like ‘9837.1, 9837.4, 0.2, 0.7‘ is not a error. It is a seasonal pattern or seasonality. An example of seasonality (energy_consumption versus datetime plot using output.csv) can be seen in figure 3.

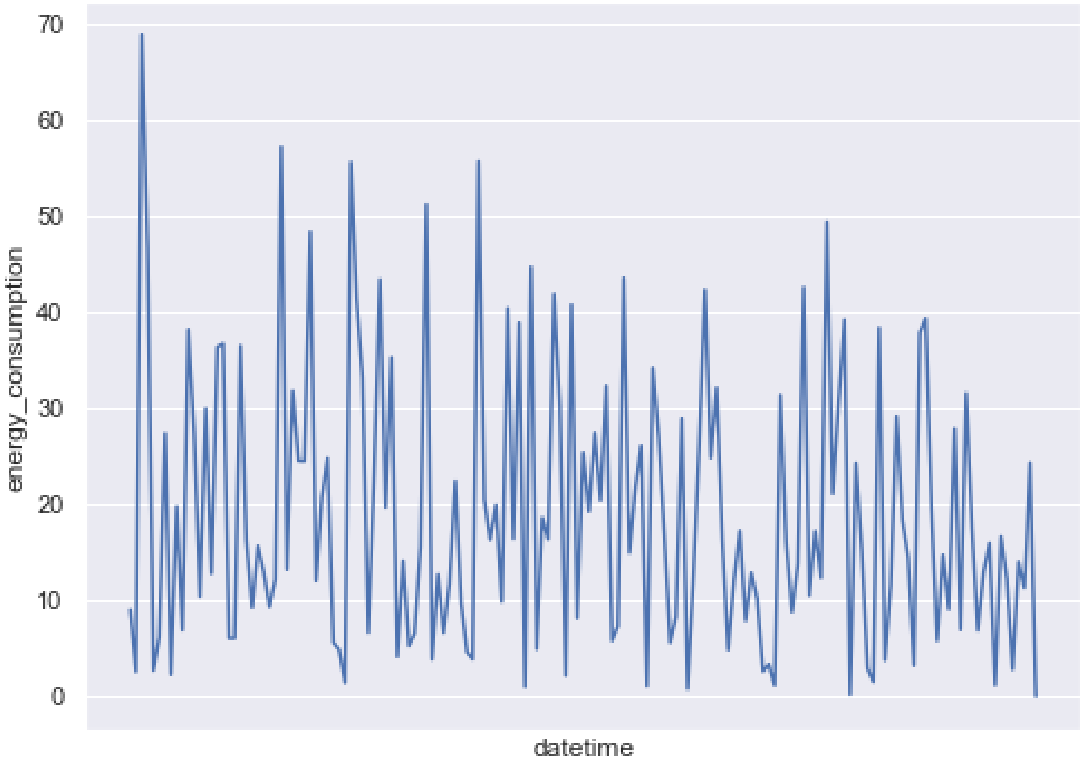

Figure 3

Possible approaches to handle seasonality:

1. One seasonal pattern is detected using fourier transform.
2. Multiple seasonal patterns are detected using autocorrelation function.

2. Unit changes. Some meters occasionally change what unit they report in, meaning the meter-state can suddenly go up/down by a factor 1000.

Possible approach:

Multiple seasonal patterns are detected using autocorrelation function.

3. Long series of missing state data. Sometimes a state meter fails to send data for over a week, despite recording updates to the state locally, then starts sending data again. When this happens we know exactly how much was consumed over the entire period of missing data (by subtracting the first datapoint after the gap from the last datapoint before the gap), but not during which hours consumption happened.

Possible approach:

> We can build a regression based model such a missForest method on the existing data to predict the missing values (which hours consumption happened).