WS 2017/18- Assignment Data Mining

Help for the Procurement Department of an Energy Trading Company

Background

Imagine you start working in the Data Mining Department of Energy Trading Company. Most of the customers are households and you have a long term relationships with them. The business model is to procure energy for the next year at the wholesale market and charge your customers once a year for their energy consumption. So normally you need the value “METER_READING” only once a year for billing. But there are also other reasons for meter reading so often you have a few more readings from the metering device of the customer. The difference to the last value is the amount of consumption in the time between the two “DATE(s)_OF_METER_READING”.

If your company buys too much or not enough energy in advance the difference is procured with an additional fee against the spot market. Therefore it is very important to accurately estimate how much energy your customers will require in the future. For existing customers the common understanding is that the expected energy amount for the next year is the energy amount required the customer required the year before. Buying energy is the responsibility of the Procurement Department so they do most of the calculation but sometimes they need your assistance.

Today the head of the procurement department needs your help with a specific question. Although they calculated all the details very carefully one year later they always have a bit too much energy. After reading a newspaper article about energy saving options, he thinks that this might be a reason for their wrong estimation but he is not sure. Furthermore, he wonders if there are regional differences. So please help him answer these questions with an analysis!

Because you are the new guy, your manager wants you to document and explain this analysis extra careful.

The IT department has already extracted the metering data and customer location data from the SAP system. They even pre-processed it and added Latitude and Longitude to the customer address so that you could better focus on the analysis. All data is provided for downloading.

Challenges

The points from a) to c) should be summarized in a 5 – 10 pages paper and 2 or more presentation slides. Code for Point d) should be provided as plain text file (.txt or .r). Upload everything to moodle.

a) Cluster your data to find regional differences. Please explain what you did. How did you select and appropriate method/algorithm? How are your parameters selected? Visualize the results.
b) Compare the energy consumption in the different regions, e.g. in 2015. Use a chart to illustrate you explanation.
c) Is the energy consumption of these long term customers rising or declining? Are there regional differences? Give an recommendation to the Procurement Department (in percent per year) either per region or in general. Illustrate and explain your findings.
d) Document your code and/or Excel files in a way that a colleague could repeat your work.

Data Files

The following files contain data from the “point of deliveries” (pod) which are the location of the metering devices.

meter_readings_export_homework.csv (or rds)

- pod_id: The ID of the point of deliveries (pod) or location of the metering device
- ACTUAL_DATE_OF_METER_READING: The date of the meter reading
- BILLED_METER_READING: Value of the meter reading at given date
- READING_UNIT: The measure of the reading, for energy kilowatthours (KWH)
- METER_READING_TYPE: The meter reading type, see table below
- METER_READING_CATEGORY: The meter reading category, see table below

metering_points_homework.csv (or rds)

- pod_id: Unique identification of the point of delivery
- pod_country: County code (usually DE) for the point of delivery
- pod_postal_code: 5-digits postal code of the point of delivery
- pod_city: Name of the city where the point of delivery is located
- LONG: Longitude of the postal code / city center of the point of delivery
- LAT: Latitude of the postal code / city center of the point of delivery

Comments from the IT Department

We selected the data of customers that had a contract overlapping the time from 2011-01-01 to 2015-12-31.

Furthermore, we found a way to geocode the customers for you. You will find LAT and LONG in the files. Feel free to use it, but think about geo-distances, e.g. using the package “geosphere” could be a good idea.

Although it is probably not needed, the values of METER_READING_CATEGORY and the METER_READING_TYPE are in the data file meter_readings_export_homework.csv. The following tables explain the meaning of the IDs.

METER_READING_TYPE

ID	Text
1	Reading by supplier
2	Reading by customer
3	Machine Estimation
4	Derived meter reading
5	Revaluation after overestimation
6	Internet reading
7	Meter reading - Upload invoice
10	DSO reading by supplier
11	DSO reading by customer
12	DSO reading by estimate
13	Fixed value 0 Removal of empty system
17	Daily average calculation
20	Reading by installer / mechanic
21	Reading by administrator / owner

METER_READING_CATEGORY

ID	Text
1	Reading by supplier
2	Reading by customer
3	Calculated by machine
4	Derived reading

Tipp:

You may come up with a data set with (normed) energy consumption (e.g. cons_365) which has an attribute “valid_from” (as date) and “valid_to” (as date).

If you want to evaluate a consumption by specific dates you can do that by using a DUMMY to join:

```
list_of_months <- data.frame(
             seq( as.Date('2011-01-01'), as.Date('2016-01-31'), by = 'month'))

energy_by_months <- mutate(list_of_months, dummy=TRUE) %>%
                    left_join( mutate( data_set, dummy=TRUE)) %>%
                    filter( valid_from <= con_month, con_month <= valid_to) %>%
                    select(-dummy)
```