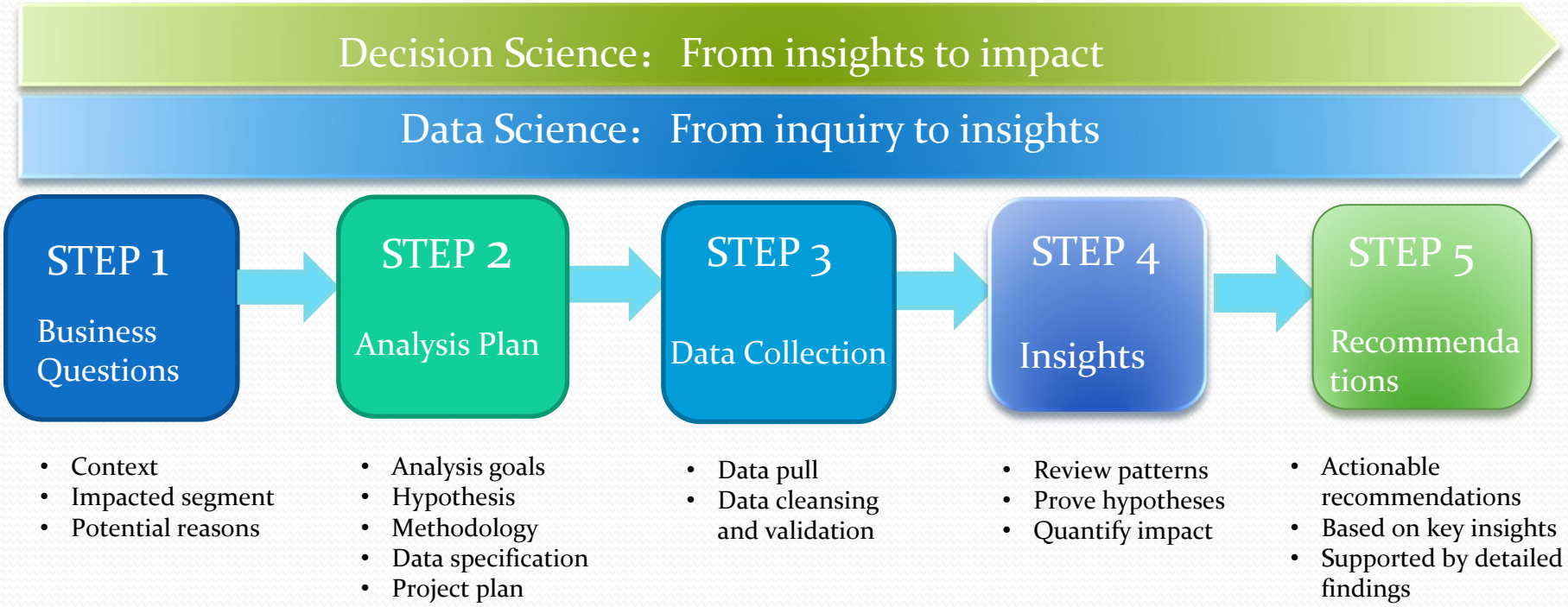
Report on your Data Science process and initial observations

2018/06/05

Effective Predictive Analytics Framework

- An effective predictive analytics process is key to make good business decisions



Background

- Our clients provided us with a very large data set that contains 47 months of energy usage data from electrical sub-metering devices used for power management.

- Our mission is to analyze this data to create analytics and visualizations and help developer gain deep insight into benefits of sub-metering devices. Our report will also help Smart Home owners with greater understanding and control of their power usage.

Objective

Sub-metering devices bring a lot of benefits.

- home-owner:

1. Lower power consumption.
2. Gain power usage analytics
3. Monitor and control over power consumption.
4. Detect appliance problems

- Developers

1. Gain competitive advantage-offering highly efficient Smart Homes.
2. Attract more customers.
3. Brings more profits and grow their businesses.

Data Management

- Obtain the data using SQL query

1. We can connect to the database that contains several annual tables (yr_2006, yr_2007, yr_2008, yr_2009, yr_2010).
2. We pull the data for each year with Date, Time and the 3 sub-meter attributes.
3. We query the database and download tables 2006 through 2010 with the specified attributes.
4. We combine tables (yr_2007, yr_2008, yr_2009) that span an entire year into a primary data frame.

- Preprocessing

We create a DateTime attribute and other new attributes and prepare the data for exploration.

Descriptions and location of data

- Measurements of electric power consumption in one household with a one-minute sampling rate over a period of almost 4 years. This archive contains 2075259 measurements gathered between December 2006 and November 2010 (47 months).

- Location of data: The data for this project is currently stored on a database in several annual tables (yr_2006, yr_2007, yr_2008, yr_2009, yr_2010). The database is stored by the host which address is data-analytics-2018.cbrosir2cswx.us-east-1.rds.amazonaws.com.

- Attribute Information

- date: Date in format dd/mm/yyyy
- time: time in format hh:mm:ss
- sub_metering_1: kitchen containing mainly a dishwasher, an oven and a microwave.
- sub_metering_2: laundry room containing a washing-machine, a tumble-drier, a refrigerator and a light.
- sub_metering_3: an electric water-heater and an air-conditioner.

Issues with the data

- The dataset contains some missing values in the measurements (nearly 1,25% of the rows). All calendar timestamps are present in the dataset but for some timestamps, the measurement values are missing: a missing value is represented by the absence of value between two consecutive semi-colon attribute separators.

- Submetering_3 has 25979 missing values. The dataset shows missing values of Submetering_3 on April 28, 2007, April 29, 2007, April 30, 2007,June 13,2009, Aug 13,2009,etc.

- Missing data is omitted since it can lead to invalid conclusions that can be drawn from the data.

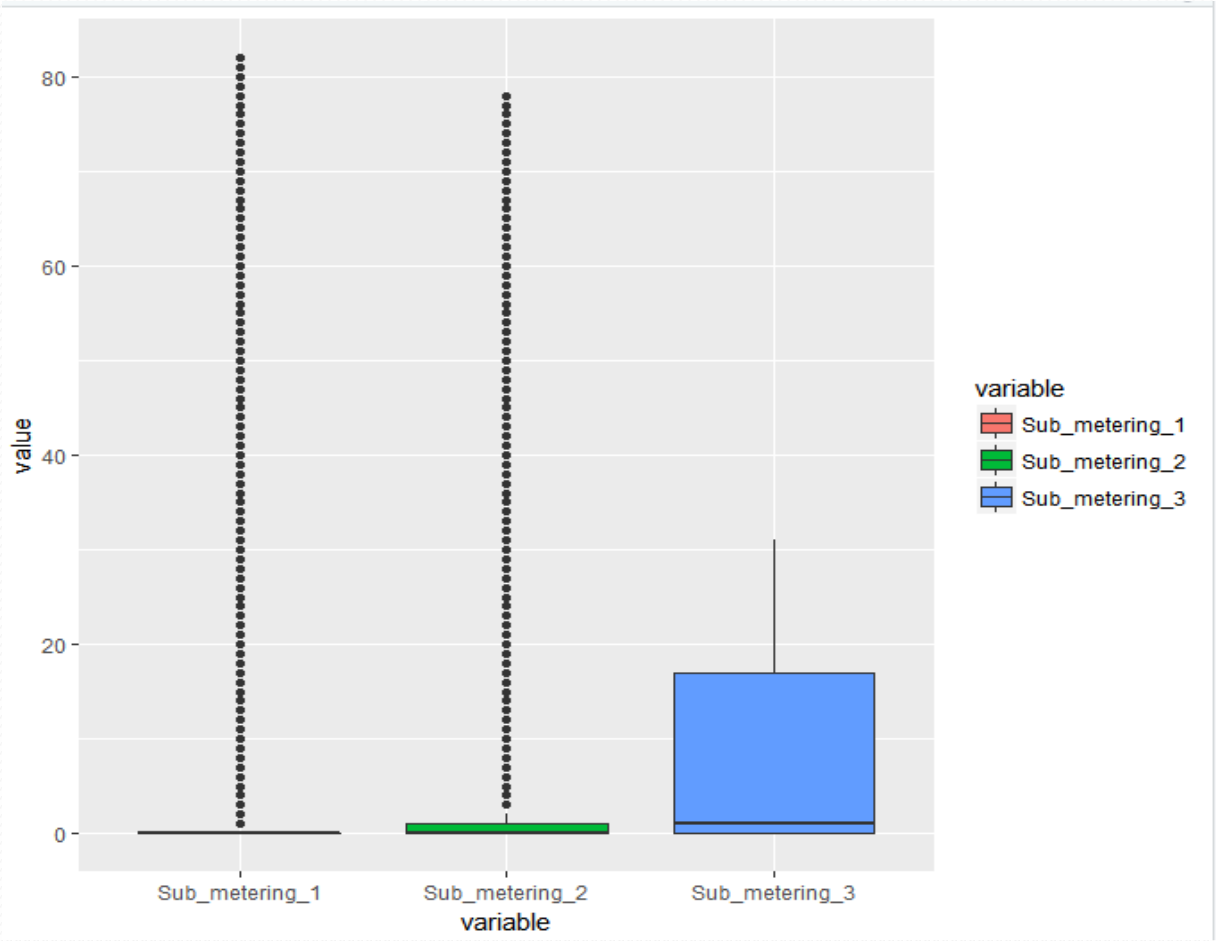
Descriptive statistics

- I listed the summary statistics of dataset below.

	Sub_metering_1	Sub_metering_2	Sub_metering_3
Min.	0.000	0.000	0.000
1st Qu.	0.000	0.000	0.000
Median	0.000	0.000	1.000
Mean	1.159	1.343	6.216
3rd Qu.	0.000	1.000	17.000
Max.	82.000	78.000	31.000

Sub-meter 3 used the most power, while sub-meter 1 used the least power. The min of these three sub-meters is 0 , while the max of sub-meter 1 is 82, the max of sub-meter 2 is 78,and the max of sub-meter 3 is 31, which are far beyond the mean and median of each sub-meter.

Descriptive statistics



The boxplot of sub_metering_1,sub_metering_2,sub_metering_3 of our datasets containing year 2007,2008,2009 indicates there are some outliers in the dataset of each sub-metering columns. There are lots of factors which lead to anomalies, such as aging appliance or appliance problems.

High-Level Recommendations

- There are more recommendations about the existing data. I change the measured scope of sub-metering 1 ,2 3 and add 4 new attribute to the dataset.
- sub_metering_1: kitchen containing mainly a dishwasher, an oven and a microwave, a refrigerator and a light.
- sub_metering_2: laundry room containing a washing-machine, a tumble-drier.
- sub_metering_3: an electric water-heater
- sub_metering_4: an air-conditioner.
- sub_metering_5: TV Stands and Entertainment Centers in the living room
- sub_metering_6: computers in the study room
- Remaining power consumption

- Questions?