VISDOM feature categories

By Sam Borgeson

12/15/2016

Although both the features calculated and their groupings are configurable, the “standard features” pre-implemented in VISDOM and used in the usage examples can be placed into a “standard grouping”. In the context of filtering customers, the groupings can help locate the features of greatest interest for filtering or visualization. Usage by non-technical users should be guided by people with technical knowledge of the feature calculations via in-person discussion and configuration changes that remove features from view that are not relevant to a particular task, with features re-grouped into categories that are familiar to the users.

It is suggested that technical readers refer to the VISDOM R documentation and source code to resolve specific implementation questions (https://github.com/ConvergenceDA/visdom/blob/master/R/features-basic.R, https://github.com/ConvergenceDA/visdom/blob/master/R/features-weather.R, https://github.com/ConvergenceDA/visdom/blob/master/R/util-census.R, https://github.com/ConvergenceDA/visdom/blob/master/R/util-regression.R).

Geography

These features relate to where a customer is located and can include zip code, county, city, state, census tract, climate zone, etc. or, where available, features related to grid topology, like upstream substation and feeder. Which geography features are available depends on what data is available with each customer’s account. The default geography is zip code and these are used to identify matching local weather data and US census statistics. Their most obvious usage of geography features is to isolate customers living in specific

Consumption

These features are intended to capture basic means (in kW) and totals (in kWh) of each customer’s consumption. These means and totals are calculated (a) across all available data – all obs (b) for the last full year of data - annual (c) summer months defined as May through September – summer (d) winter months defined as October through April – winter (e) for each month, defined as 30 days of consumption across the average of all observations within each month, which can include readings from multiple years.

Stats

These features capture basic statistical extracts from the meter data, including percentiles, maxima, minima, variance, ratios, simple correlation with other phenomena, like outside temperature.

Seasonal

These features are meant to capture seasonal variability in consumption by computing a suite of features for both August and January. These are all based on metrics drawn from daily load shapes averaged across all days and include average consumption, the daily minimum and maximum demand, the range between those two values, the duration (in hrs) of elevated demand - defined as the time spent higher than ½ way between the daily min and max demand, and the ratios of minimum to maximum demand and of nighttime (2-5am) to day time (4-7pm) consumption.

Hourly

These features are simply the average consumption for each hour of the day across all supplied data. Taken together, they represent the average load shape for each customer, and individually, they can be used to identify the typical timing of each customer's demand, which is a proxy for occupancy and often relevant to grid operational concerns.

Weather

These features are averages from the weather data associated with each customer (typically by zip code). All customers associated with the same weather station will have the same weather features. For these calculations, Summer is May through September and winter is October through April.

Meta

These features provide meta-data for each customer's data. These include the number of meter readings (aka electricity observations) analyzed. For example, one year of hourly data without any missing readings and on a non-leap year would be 365 x 24 = 8760 observations, and the dates of the first and last readings.

Census

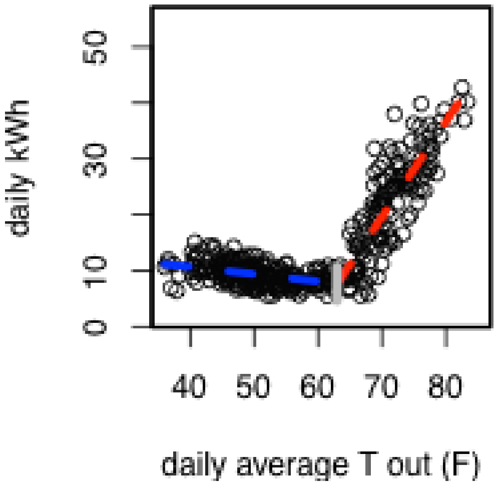
These are census statistics from the American Community Survey (ACS) of the US Census, associated with each customer by geography. The correlation is typically done via zip code, using Zip Code Tabulation Area averages and percentages published by the census bureau in ACS tables DP02-DP05. The ACS data itself is a 3 or 5 year running average looking back from a given year. The default as of 12/15/2016 is a 5 year average back from 2014.

Model

These are the parameter fits from a piecewise temperature regression model using daily average outside temperature to explain daily total electric energy consumption. The default model specification splits the data into heating and cooling temperature regions and fits a temperature response term separately to each. See the figure below for an illustration of a model fit with underlying data for reference. The actual regression equation is:

$$kWh = \beta_0 + \beta_{tout_-}(CP - tout)_+ + \beta_{tout_+}(tout - CP)_+ + \varepsilon$$

The heating data is fit with the blue line, which gradually increases as it gets colder out in the figure (this is consistent with the fan energy required to circulate hot air from a furnace). The cooling data is fit with the red line, which increases quickly beyond the change point in the figure, indicating the presence of a significant AC load). The change point temperature that divides the heating and cooling regions is derived automatically based on the best overall model fit. For a building scientist, the change point can be understood as the balance point of the building.



The model statistics captured as features include:

Sum of squared residuals for daily change point model	The SSR is a metric of how good or bad the model fit is. It is the sum of the square of the difference between model predictions and corresponding underlying data. Larger SSR's mean a worse fit.
Heating and cooling change point	This is the "balance point" temperature that minimizes the SSR for the model. Above this temperature, the cooling slope is used to predict consumption. Below this temperature, the heating slope is used.
Model constant term	This is the non-heating and non-cooling daily energy consumption for the household. Put a different way, it is the expected daily energy consumption at the change/balance point average daily temperature.
Model heating sensitivity	This is the slope of the cold temperature fit (aka heating), in units of kWh/day / °F of daily mean outdoor temperature.
Model cooling sensitivity	This is the slope of the hot temperature fit (aka cooling), in units of kWh/day / °F of daily mean outdoor temperature.
xxx pvalue	The regression model used will always provide a piecewise fit regardless of the nature of the underlying data. For example, many people don't have air conditioning, but there will still be an upper slope. All model coefficients come with a corresponding p-value metric that indicates statistical confidence in the fit. A p-value represents the probability that the corresponding term could actually be zero and its fit value is a statistical anomaly. Smaller p-values indicate better fit.