Baseline in Streaming Data Analysis: What, Why, and How?

Alex Sim, John Wu
Scientific Data Management, LBNL

Sam Borgeson, Ling Lin, Anna Spurlock, Annika Todd
Electricity Markets and Policy, LBNL

Taehoon Kim, Dongeun Lee, Jaesik Choi
Computer Science, UNIST, South Korea

Data “explosion” in energy

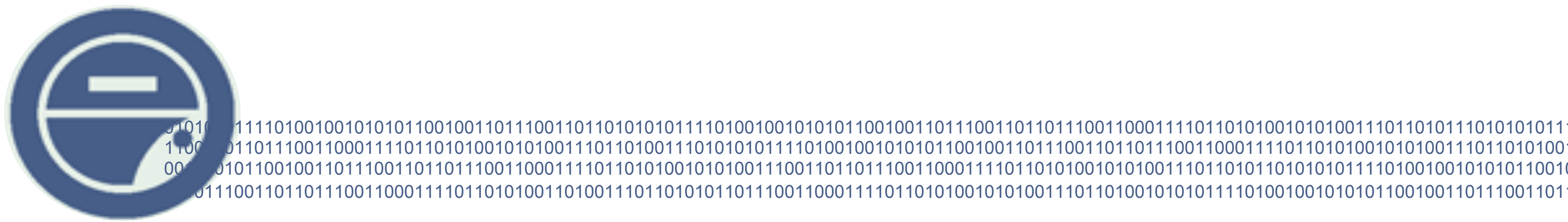
Smart meters, thermostats, appliances, cars

Linked to other time and location-specific information (temperature, census, satellite)

Provide vast, constantly growing streams of rich data

Smart meter data enables many possibilities for cutting edge analyses

What can we do with all of this data?



Many possibilities!

Insights from the data → tremendous potential value for a wide range of energy policies.

Our solution - Behavior Analytics:

combine behavioral theories with cutting-edge data science analytics

Novel insights into patterns of electricity consumption, underlying human drivers, and the way that people make energy decisions

Goal: **shed light on key policy issues**, including improving program effectiveness, increasing the accuracy of load forecasting, and creating better EM&V methods

Using only **readily available data** from smart meters and other sources

Our current test dataset

- Residential electricity records: 100,000 households from a region where electricity usage peaks in the summer time
- Data collected hourly over three years, one pre-rate, two after
- Randomized controlled trial of new rate structures - households are randomly placed in different groups
  - Control: using existing fixed rate for electricity
  - Active1: households opted in to Critical Peak Pricing
  - Active 2: households opted in to Time-of-Use Pricing
  - Passive1: households defaulted in to Critical Peak Pricing
  - Passive2: households defaulted in to Time-of-Use Pricing

A few examples of our research

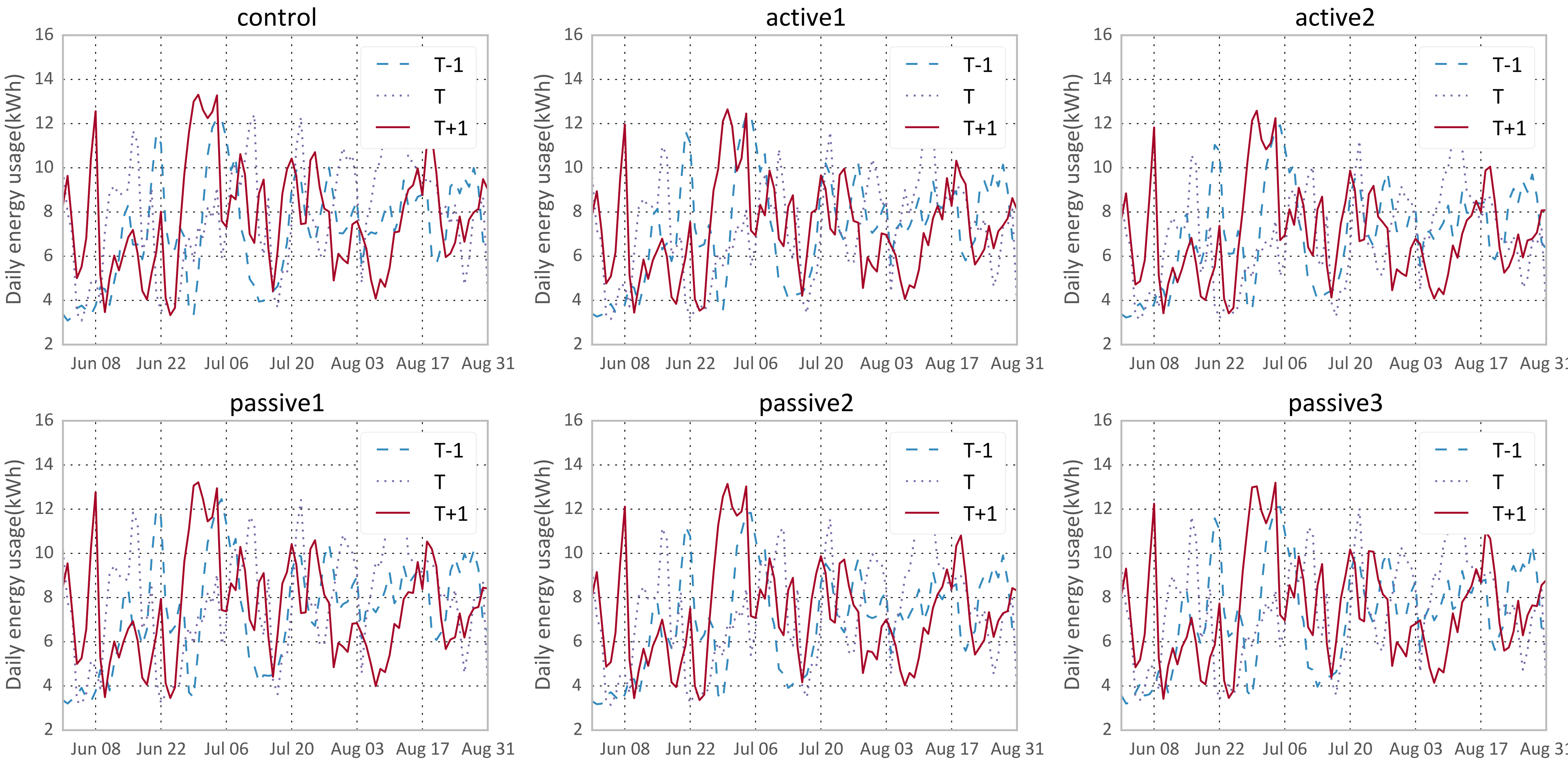
Goal	Method	Policy Implication
Better baseline models of energy use	Gradient tree boosting	Better program evaluation
Define relevant household characteristics:		Classify and segment households using easily accessible data
• Define representative load shapes	Adaptive K-means clustering	
• Estimate household-specific cooling change points (AC set point)	Piecewise linear regression, bootstrapping	
• Characterize customers into "Lifestyle Groups"	Blend behavioral theory with machine learning techniques	
• Define relevant household energy characteristics	Simple feature algorithms (e.g., mean, min, max, peak usage; variance; entropy, etc.)	

A few examples of our research

Goal	Method	Policy Implication
Determine what types of households are more likely to enroll and respond to programs	Instrumental variable regression using household classifications developed above	• Targeting to improve program uptake and response • Optimize program deployment by tailoring specific programs to specific households • Increase demand side flexibility by making peak-load reductions more readily available and on-demand
Develop a proof-of-concept for customer load prediction and forecasting models	Prediction algorithms using household characteristics and classifications developed above	• Create tools for distributed resource planning • Enable DR as a dependable system resource

Test Data

- Residential electricity records: 100,000 households from a region where electricity usage peaks in the summer time
- Data collected over three years
  - Year T-1: the year before the introduction of new rate structures
  - Year T: new rate structure introduced at the beginning of the summer of the treatment year
  - Year T+1: the year following Year T
- Households are randomly placed in different groups
  - Control: using existing fixed rate for electricity
  - Active1: households opted in to Critical Peak Pricing
  - Active2: households opted in to Time-of-Use Pricing
  - Passive1: households defaulted in to Critical Peak Pricing
  - Passive2: households defaulted in to Time-of-Use Pricing
  - Passive3: households defaulted in to both new rate structures



Electricity usage data from three consecutive nears (labeled as T-1, T, and T+1) with carefully designed groups (labeled control, active1, active2, passive1, passive2, and passive3), however, usage curves are nearly identical because the common factors such as temperature dominate the typical electricity usage.

“Gold Standard” for Establishing Baselines

Randomized Controlled Trials (RCT)

- Well tested both in practice and theory
  - First reported use by James Lind in 1747 to identify treatment for scurvy
- However, a properly designed RCT has to get many things right (Anna Karenina principle -- Happy families are all alike; every unhappy family is unhappy in its own way)
  - Randomization – bias (Berger and Berry 1988)
  - Concurrent control group – blind data collection (Horwitz and Feinstein 1979)
- Other concerns with RCT:
  - RCTs are not always possible (Liddle, Williamson and Irwig 1996)
  - Inappropriate surrogate outcome measures, so that what is said to have been measured is in fact not (see Gøtzsche et al. 1996; Jaeschke and Sackett 1989);
  - Incomplete intention-to-treat analysis, resulting in inaccurate statistical analysis (Feinstein and Horwitz 1997; Evidence-Based Medicine Working Group 2002);
  - Problems with blinding and use of suitable placebo, resulting in inade- quate controls on the effects of bias (Hensley and Gibson 1998);
  - Inability to generalize from the trial’s results, due to ambiguous objec- tives or failure to report inclusion/exclusion criteria (Jaeschke and Sackett 1989);
  - False negatives, resulting from small numbers or insensitive outcome measures (Jaeschke and Sackett 1989).

Control Group Did Not Do Its Job? Or Signals Too Weak?

Average electricity usage during the summer time (**averaged over all hours**)			
	Year T-1	Year T	Year T+1
Control	1.128	1.205	1.197
Active1	1.136	1.163	1.161
Active2	1.125	1.160	1.173
Passive1	1.157	1.206	1.181
Passive2	**1.100**	1.152	1.154
Passive3	1.174	1.216	**1.228**

Comparisons again the Control group e.g., Active1 - Control			
	Year T-1	Year T	Year T+1
Control	-	-	-
Active1	0.008	-0.042	-0.036
Active2	-0.003	-0.045	-0.024
Passive1	0.029	0.001	-0.016
Passive2	-0.028	**-0.053**	-0.043
Passive3	**0.046**	0.009	0.031

- In the tables: **red number** is the largest in the table, **blue number** is the smallest values
- During Year T-1, the average usage of control group should be the same as those of the passive participant groups, but they are quite different
- Question: badly designed groups?

Control Group Did Not Do Its Job? Or Signals Too Weak?

Average electricity usage during the **peak demand hours** of each day			
	Year T-1	Year T	Year T+1
Control	1.790	**1.973**	1.937
Active1	1.805	1.796	1.806
Active2	1.752	**1.696**	1.739
Passive1	1.853	1.952	1.877
Passive2	1.742	1.822	1.818
Passive3	1.809	1.870	1.854

Comparisons again the Control group e.g., Active1 - Control			
	Year T-1	Year T	Year T+1
Control	-	-	-
Active1	0.015	-0.177	-0.131
Active2	-0.038	**-0.277**	-0.234
Passive1	**0.063**	-0.021	-0.060
Passive2	-0.048	-0.151	-0.119
Passive3	0.019	-0.103	-0.083

- Good: usage during peak demand hours are all less than the corresponding values of Control group
- Bad: the maximum difference still appears in Year T-1
- Question: badly designed groups?

Gradient tree boosting

Gradient Boosting = **Gradient Descent** + Boosting
Similar to **Adaboost** : Ensemble method that a set of weak learner to compensate the shortcomings of existing weak learners

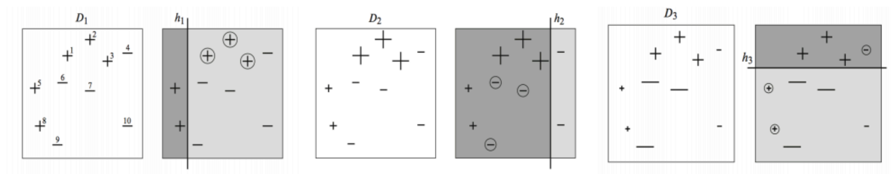
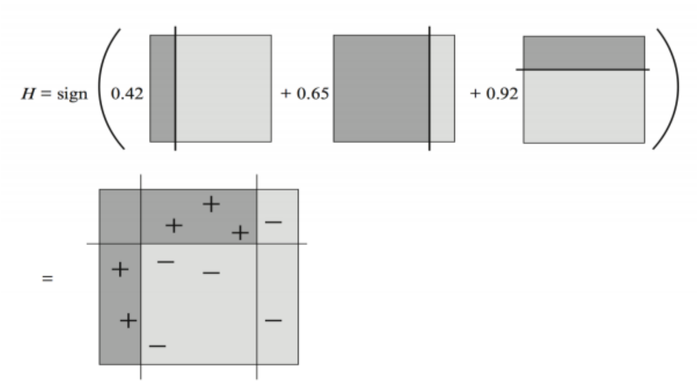


Figure : AdaBoost. Source: Figure 1.1 of [Schapire and Freund, 2012]

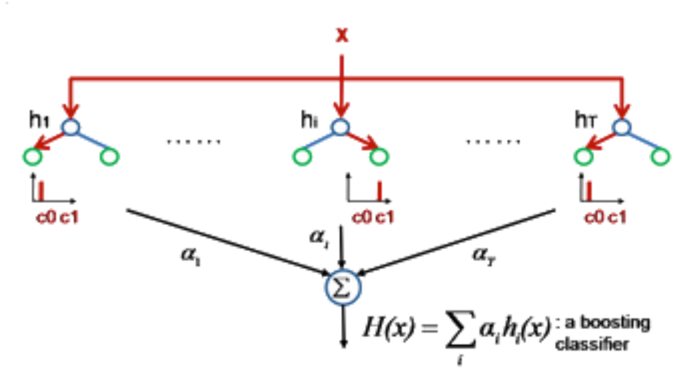


$$H(x) = \sum_t \rho_t h_t(x)$$

reference : http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf

Gradient tree boosting

Build a decision tree continuously to fit a set of weak decision trees to the dataset.



The reason why “**Gradient**” is that fitting a *residual* $y - F(x)$ can be considered as adding negative gradients of the squared error loss function, which is a gradient descent.

reference : http://www.52ml.net/8555.html, https://en.wikipedia.org/wiki/Gradient_boosting

Gradient tree boosting

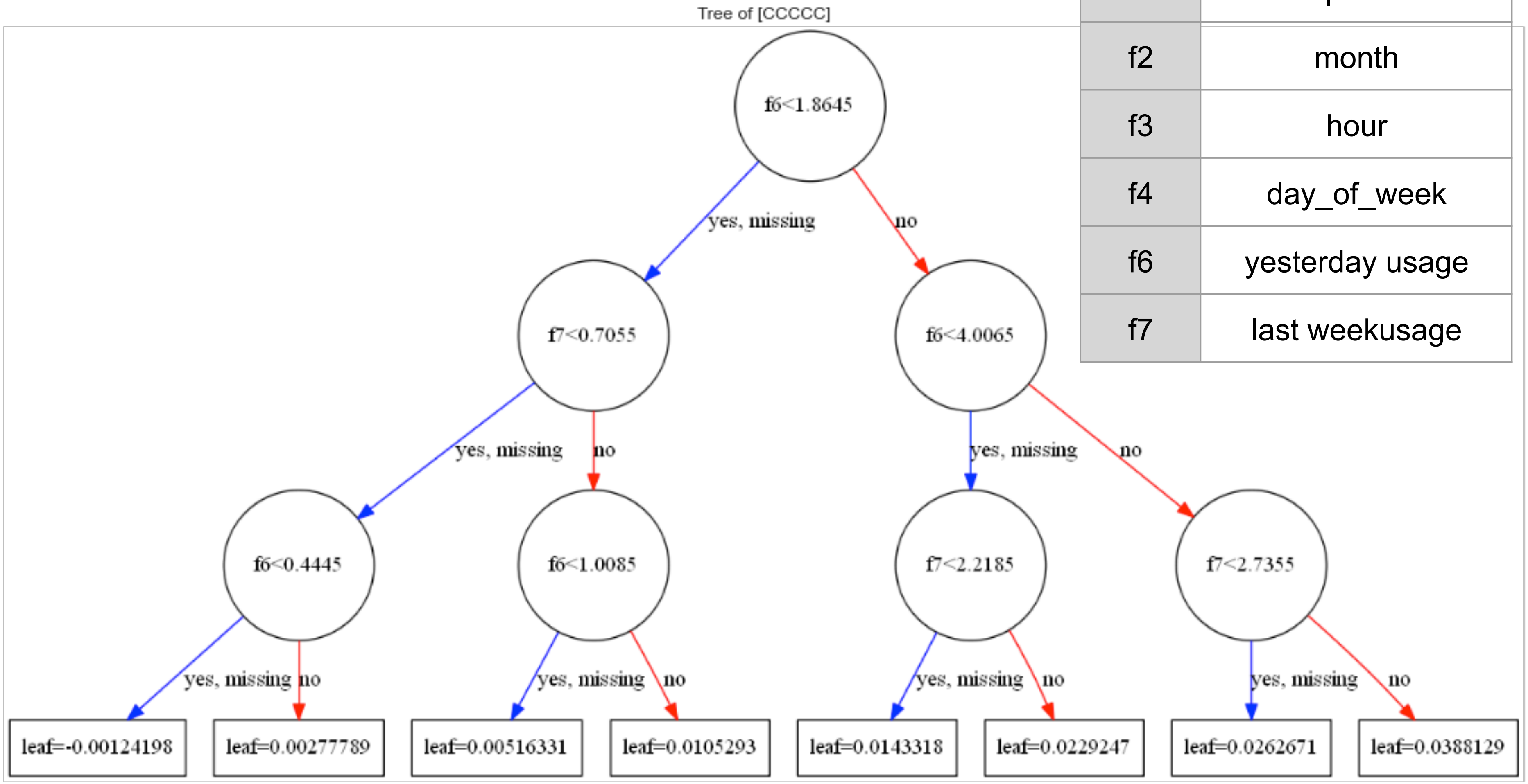
1. Training for each household
2. Training for each group (ex. RIXXXX)

Features

- temperature
- month
- hour
- day_of_week
- usage of yesterday (for group training)
- usage of last week (for group training)
- Before and After treatment
- RITTNE things

Build **1000 decision trees** for boosting and vary a max_depth (=3, 5, 18) for a single tree

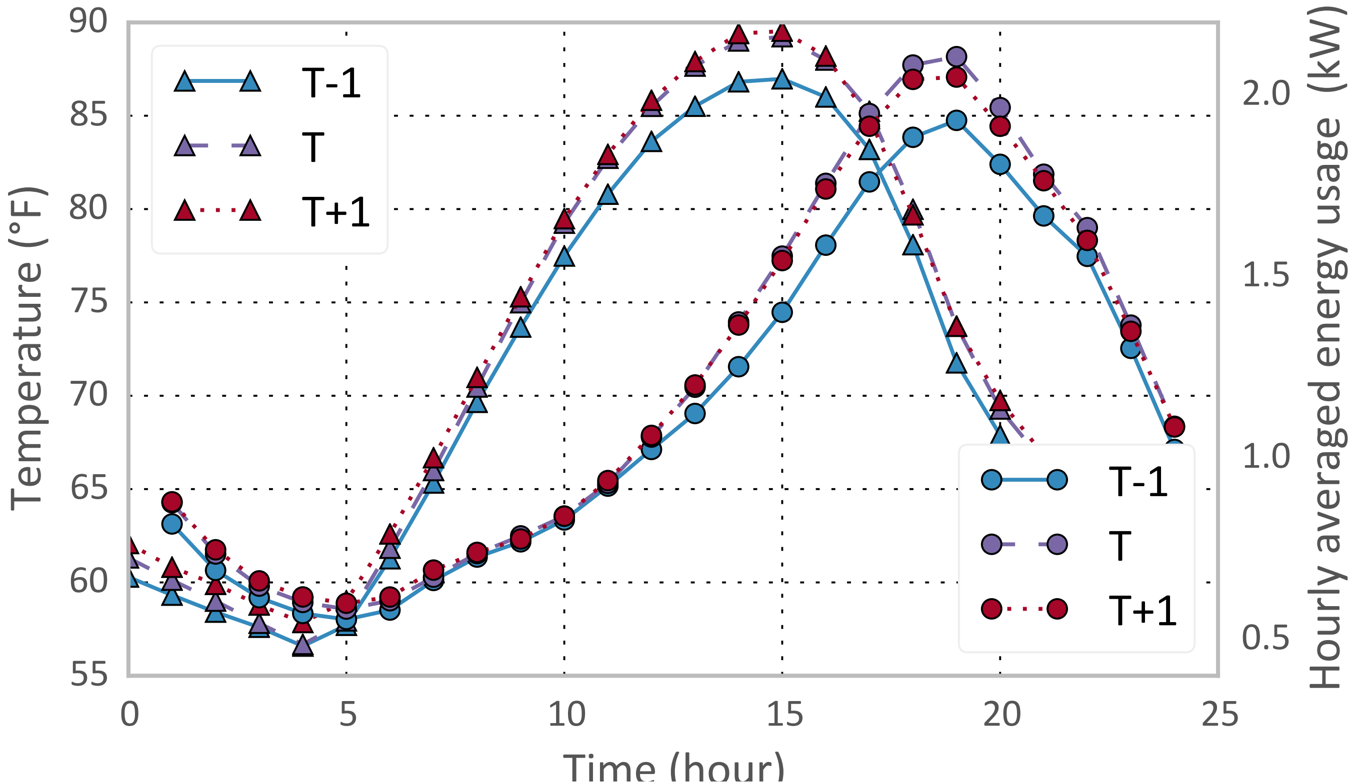
Decision tree of Gradient Tree Boosting model for Control Group with max_depth=3



f0	tempearture
f2	month
f3	hour
f4	day_of_week
f6	yesterday usage
f7	last weekusage

The **weighted sum** of leaf value will be a predicted value of energy usage.
F score : addition of reference value for each node per feature

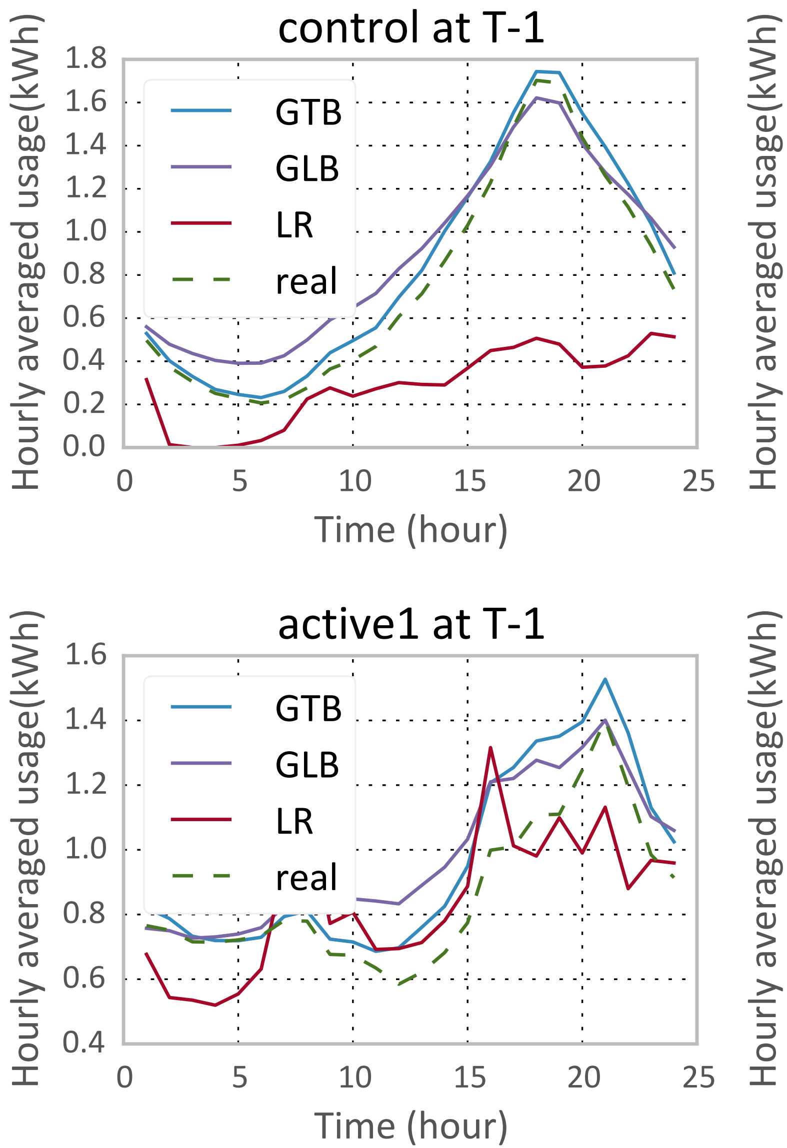
GTB Produces Baselines with Correct Delay in Daily Peak Usage



Observation 1: GTB produced baselines are promising because they have the right delays for daily peak usage

GTB Produced Baselines Match Observations Better than Others

- The right figure shows the comparison o baselines produced by three popular machine learning methods: GTB, GLB and LR
- GTB == Gradient Tree Boosting
- GLB == Gradient Linear Boosting
- LR == Linear Regression



Summary & Planned Work

- Problem: Randomized Controlled Trials (RCT) are hard to design, the data set we have contains strong indication that the behavior of control group does not match treatment groups prior to the application of treatments (year T-1)
- Solution: Explore a different strategy for producing baselines with machine learning
- Preliminary tests shown Gradient Tree Boosting is promising because the baselines it produced have expected shapes and relatively small errors
- To Do: more rigorous study of machine learning options