Learning How People Respond to Changes in Energy Prices

Ben Lim
bclimate

Jonathan Roth
jmroth

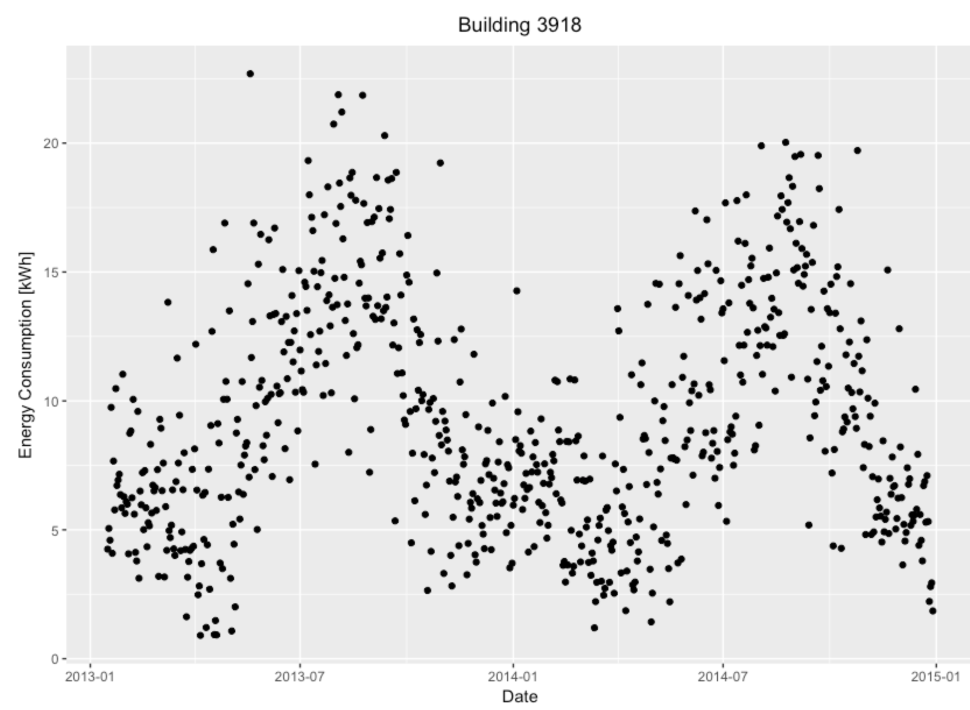
Andrew Sonta
asonta

Overview

Predict: Next-day energy consumption for households after some houses receive notification that energy prices will change

Using: OLS, Random Forests, SVMs, and time-series ARIMA models

Results: Radom forests are found to be the most effective models



Dataset

Our data is collected from the Pecan Street organization, a nonprofit dedicated to using data for modeling of residential energy.
The goal is to forecast the energy consumption of a single household for the hours between 4:00 PM and 7:00 PM after notification that energy prices will change for some households. We use previous days' energy consumption, information about the notification, and weather data in order to make these predictions.

Data Split Chronologically: Train - 76%, Validation - 12%, Test - 12%

Features

- Consumption for 4:00 - 7:00 PM in kWh on the the following days: 1, 2, 7, 8, 14, 15 days before the event
- Type of notification the household received 1 day before the event: nothing, text, email
- Weather data, including temperature, maximum temperature, cooling degree days, and heating degree days in the day of the event

These features all have some impact on building energy consumption, as understood in the civil engineering domain

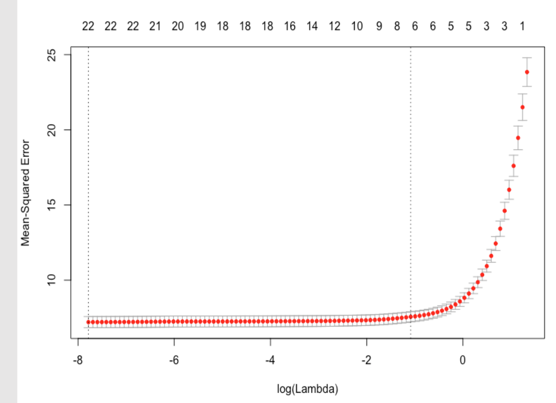
Models

We tested Ordinary Least Squares, Random Forests, Support Vector Machines (regression), and a time-series ARIMA model. Each category of model was tested with 3 alternatives, for a total of 12 models.

Ordinary Least Squares

- Stepwise variable selection performed using AIC and BIC
- Regularization was performed using the LASSO
- One-standard error rules used to select the lambda

$$\beta = argmin_{\beta}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

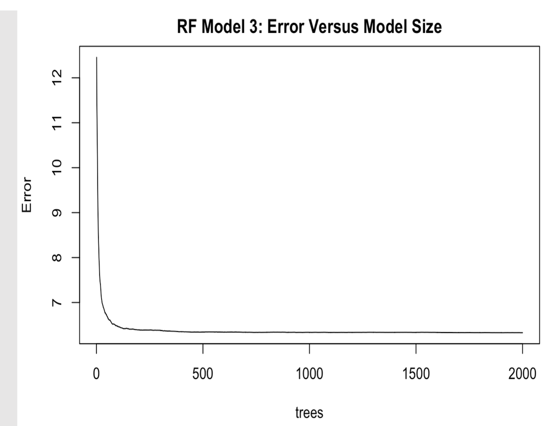


Random Forest

- 3 models of varying size (100 trees or 2000 trees)
- Variable selection performed both manually and automatically
- Number of variables tested at each split varied (4 or 12)

$$\widehat{y_i} = \frac{1}{T}\sum_{t=1}^{T}\widehat{y_{it}}\ , e_i = (y_i - \widehat{y_i})^2 = L(y_i, \widehat{y_i})$$

where y_{it} is the predicted value output by tree t

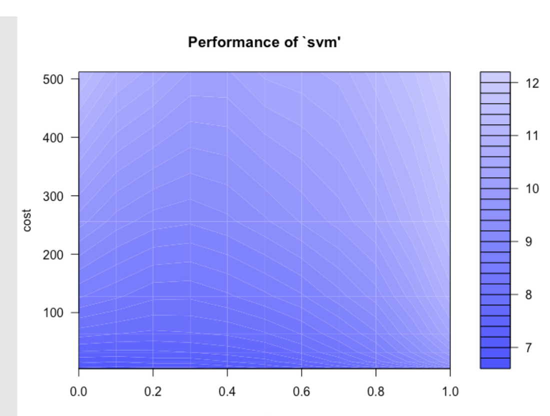


Support Vector Machines

Radial Kernel
$exp(-\gamma|u - v|^2)$
$\gamma = 0.02$

Polynomial Kernel
$(\gamma u'v + coef0)^{degree}$
$\gamma = 0.08; degree = 2$

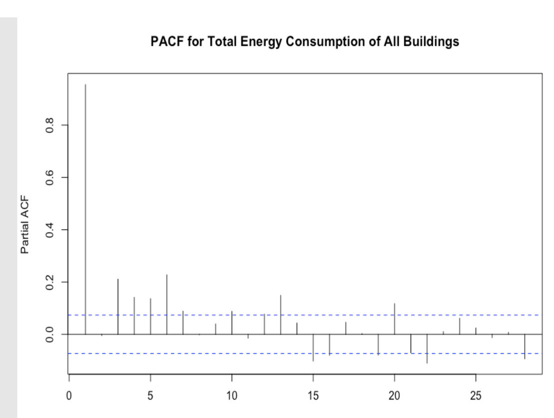
Linear Kernel
$u'v$
No hyperparameters



ARIMA

- Model selection determined through AIC
- Models were constructed for individual houses and aggregate
- Additional exogenous variables were added

$$\phi(B)(1 - B)^d x_t = \theta(B)w_t$$



Training / Validation

The results of our models are shown in the table below. We used percent error of the aggregated building energy consumption as our model selection criterion.

Model Type	Control Group	Event Group	Combined Group
OLS - Regularization	5.26 (5.80) 3.39%	6.42 (7.48) 3.41%	5.59 (6.41) 3.99%
OLS - Stepwise (AIC)	8.63 (5.40) 9.27%	7.35 (7.10) -1.31%	8.51 (6.22) 8.08%
OLS - Stepwise (BIC)	5.25 (5.43) 3.17%	6.77 (7.13) 4.40%	5.58 (6.25) 3.57%
Random Forest (Model 1)	5.38 (1.00) 5.59%	6.39 (1.28) 3.33%	5.69 (1.13) 4.34%
Random Forest (Model 2)	5.35 (1.16) 4.49%	6.34 (1.45) 3.30%	5.64 (1.26) 4.09%
Random Forest (Model 3)	5.93 (1.03) 4.03%	7.33 (1.30) **-0.99%**	6.50 (1.13) 2.83%
SVM (Radial Kernel)	5.35 (5.18) **1.00%**	6.36 (6.85) 3.22%	5.56 (6.08) 2.02%
SVM (Polynomial Kernel)	5.25 (5.48) 1.39%	6.50 (7.05) 3.11%	5.65 (6.26) **1.72%**
SVM (Linear Kernel)	5.23 (5.59) 2.16%	6.27 (7.28) 4.56%	5.61 (6.38) 2.77%
ARIMA (using single building data)	21.10 - -9.98%	12.52 - -1.58%	17.07 - -13.98%
ARIMA (All Buildings summed)	- - 2.76%	- - 3.94%	- - 3.01%
ARIMA (All Buildings summed w/ weather data)	- - 3.36%	- - 3.58%	- - 3.39%

In each box: Dev Set Mean Squared Error
Train Set Mean Squared Error
Percent Error (aggregate of full building portfolio)

Results on Test Set

Test Set Error	Control Group	Event Group	Combined Group
Random Forest (Model 3)	-0.65981%	1.8036%	0.4752%

Discussion

The third random forest model, was chosen as the best model, given its performance on the event group dev set.

Parameters: 2000 Trees, 12 Variables tested at each split, Automated Variable Selection

Out of curiosity, we checked the performance of the SVM models (radial and polynomial kernel) on the test, and found that it performed very poorly. Which led us to believe that these model were overfitting to the training data.

Future Work

Given more time to work on this project, we would like to spend more time thinking about which variables have the most importance in predicting energy consumption.

Furthermore, we would like to find ways to improve the accuracy of our models on the "Event Group," as this is the group that is currently the hardest to predict, since it is difficult to model how people change their behavior after they receive new information.

References

[1]L. Suganthi and A. Samuel, "Energy models for demand forecasting—A review", *Elsevier*, 2017. [Online]. Available: https://doi.org/10.1016/j.rser.2011.08.014. [Accessed: 09- Nov- 2017].
[2]N. Fumo and M. Rafe Biswas, "Regression analysis for prediction of residential energy consumption", *Elsevier*, 2017. [Online]. Available: https://doi.org/10.1016/j.rser.2015.03.035. [Accessed: 15- Nov- 2017].