Using Open Data to Predict Energy Usage

What tax lot data can tell us about energy usage intensity in New York City

Theo Love

Green Energy Economics Group
What can all this data tell us about how we use energy?
It’s the Usual Suspects

WANTED
For Explaining Energy Usage Intensity

Building Age
Land Usage/Building Type
Building Value

But that’s only part of the story...
Existing Benchmarks

ENERGY STAR Score vs. EUI for All Buildings and Years

Interesting Outliers

Flattens in the middle
Going Further with New Data

- **PLUTO**: 200 features
- **Code Violations**: 190 features
- **Local Law 84**: 15 features

**Supervised Learning**

- **Linear Regressions**
- **Decision Trees**
- **PCA**

**Socrata**
What Was Found?

**Tier 1 Features**
- Is it a hospital?

**Tier 2 Features**
- Is it an office?
- Year built

**Tier 3 Features**
- Value of building/land
- Year of last renovation
- Building size/usage type

**Best Model:** Random Forest Regression

$R^2$ from 23% to 29% with standard deviation ~3%
Hints in the Data

Residuals

» Clear linear relationship
» Further classification striations

Primary Component Analysis

» Collapses to two main features
» First component explains vast majority of variance
What Does it Mean?

» Around 30% of variance explained by features in PLUTO

» **Sub-sector** analysis **crucial**

» Type > Age > Value > Size
What Now?

» **White paper** with additional details
» **Predicting** out of sample (new years)
» Bring in **more data** to improve model
» Applying and examining in **new areas**
“Large repositories of public data can help improve our understanding of energy usage.”

» Digging for efficiency opportunities
  • Significant features could show hidden usage drivers
  • Outliers may mean opportunity
I used to think correlation implied causation.

Then I took a statistics class. Now I don’t.

Sounds like the class helped.

Well, maybe.

© xkcd.com

Theo Love

» tlove@greenenergyeconomics.com

Green Energy Economics Group