Predicting Real Estate Tax Assessment Increases in NYC

11/23/2016

What we're talking about…

- Intro
- Summary of Findings
- Overview of Data
- Baseline Model
- Feature Engineering
- Conclusion

Intro

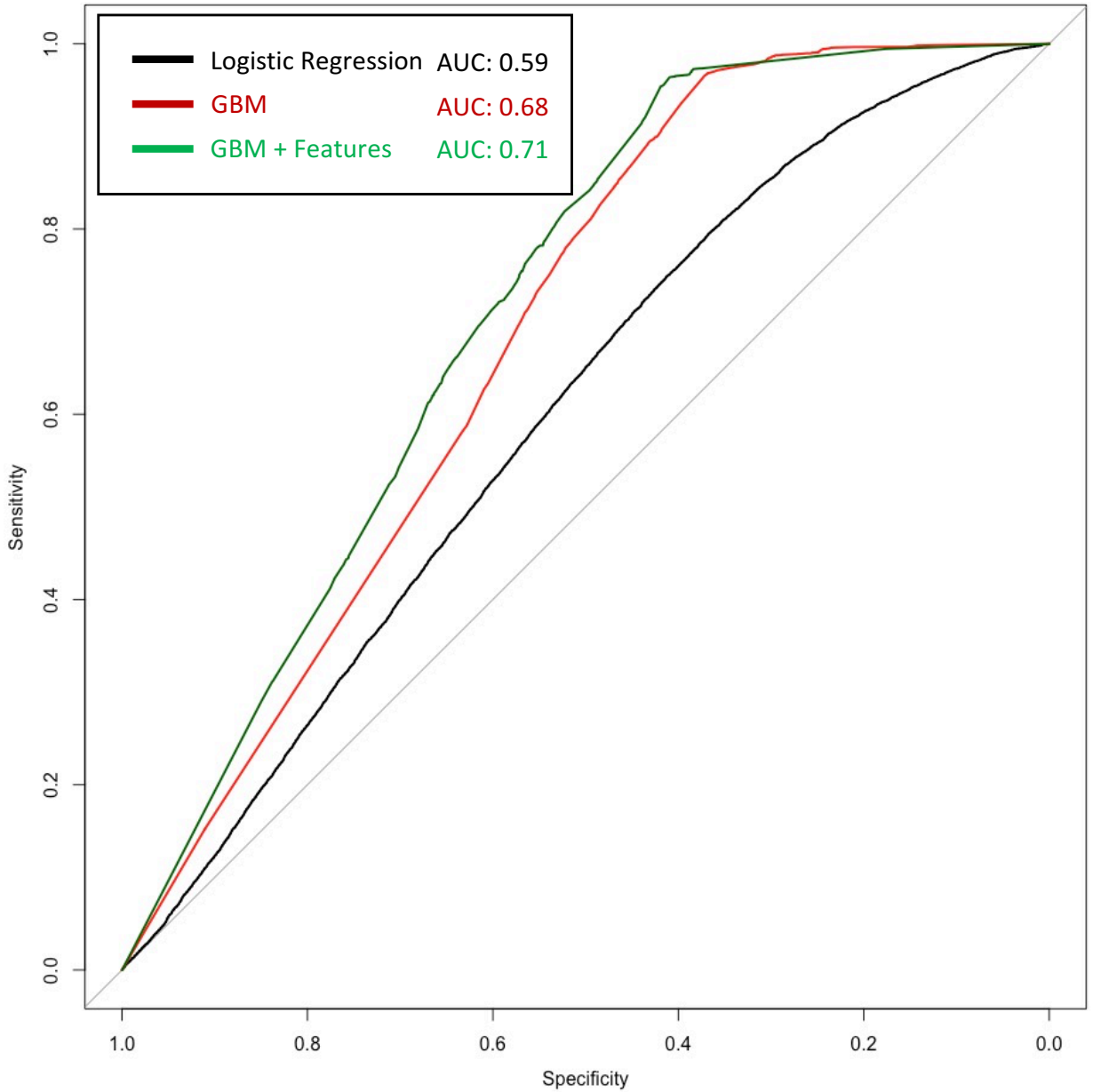
- Real Estate Taxes in NYC represent a major cost to landowners, especially in commercial real estate
- How much is owed in taxes is based primarily on the *Assessed Value* of the building
- Assessed values typically go up slowly over time, however, each year the city chooses a handful of properties for an in-depth re-assessment
- If you get re-assessed, chances are, you're going to pay a lot more in taxes
- ***How* the city chooses which buildings to re-assess is a bit of a mystery**
  - *Some possible considerations are: the gentrification of your neighborhood, new construction nearby, whether the city thinks you can "afford it", etc.*

Can We Predict When an Assessment Will Happen?

- Short answer: YES! (kind of…)
- NYC releases the PRIMARY LAND USE TAX LOT OUTPUT dataset, aka, **PLUTO**
- PLUTO contains information about every building in NYC, including assessed values, # of Residential Units, Year Built, Square Footage and much more
- Our goal: predict the probability that a building will receive a Tax Assessment in a given year

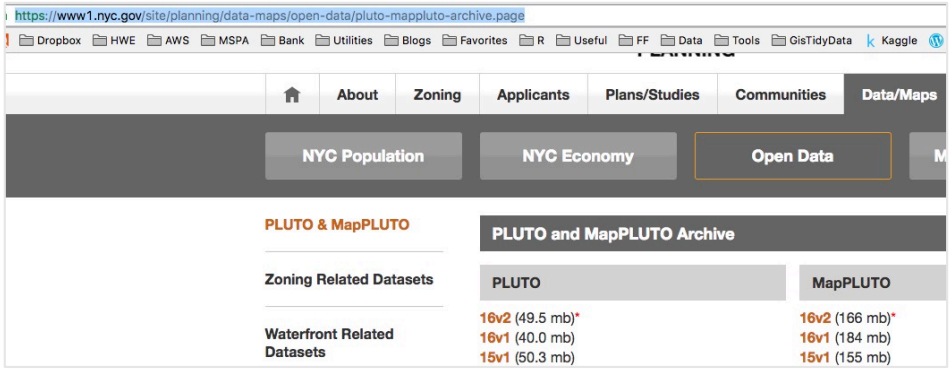
Summary of Findings

- Baseline predictive power of the dataset is relatively high
- GBM outperforms logistic by a wide margin
- Feature Engineering proved valuable, and could likely be pushed much further
- In particular, geo-spatial kernel features show great promise and merit further investigation



Overview of PLUTO Data

- NYC makes it's Tax Lot available online:
  - https://www1.nyc.gov/site/planning/data-maps/open-data/pluto-mappluto-archive.page



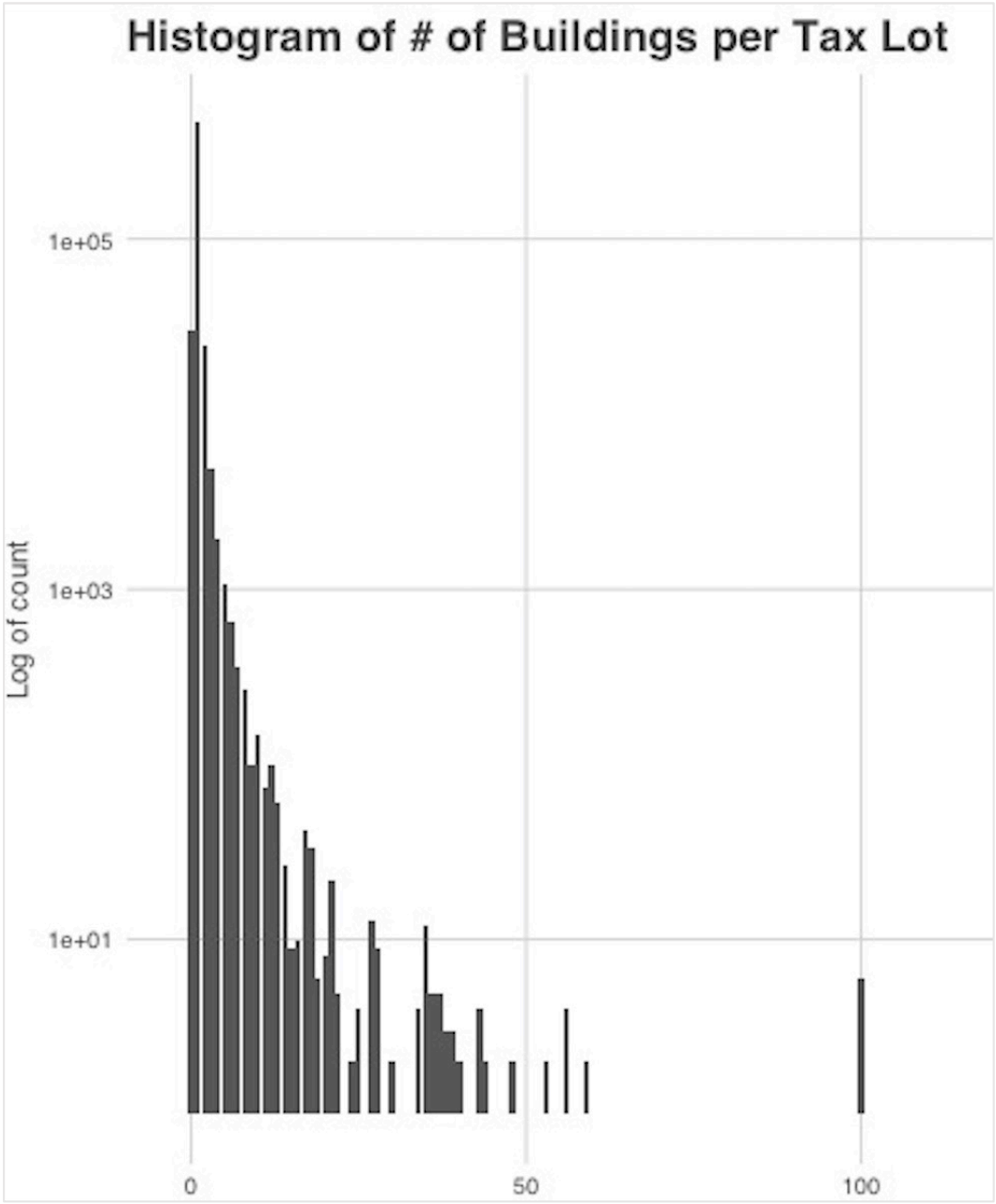
- Contains building characteristic and tax data
- A Tax Lot is, more or less, 1 building
  - Many examples of tax lots having more than one building, for example: the World Trade Center
- 15 years of data, 1-2 versions per year since 2002
- >10 Million rows, 100+ columns

Our Modeling Data

Filtering the entire 10 Million row dataset for:

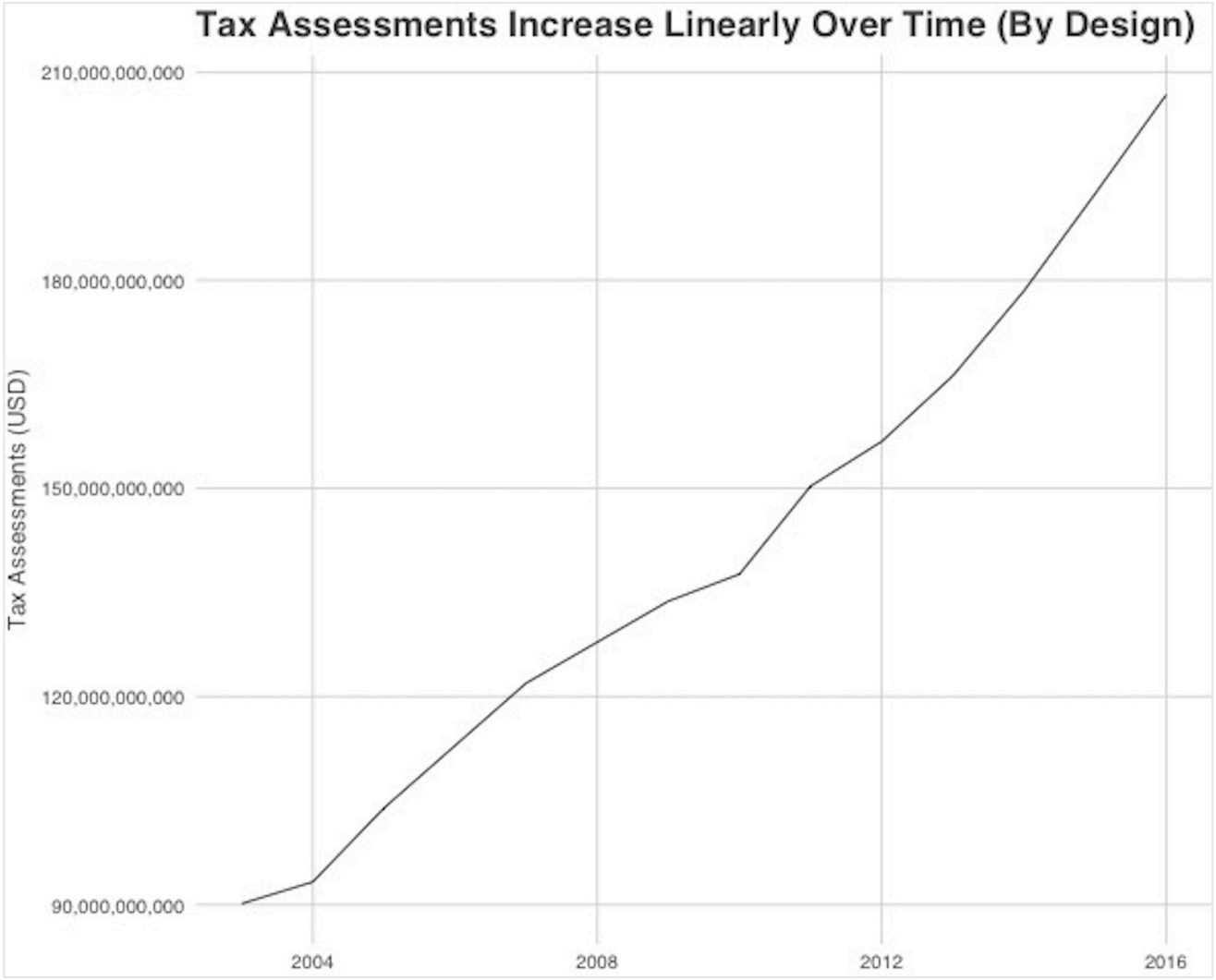
- Only Manhattan
- Only tax lots with 1 building
- 2010-2016

Total rows: 265,911



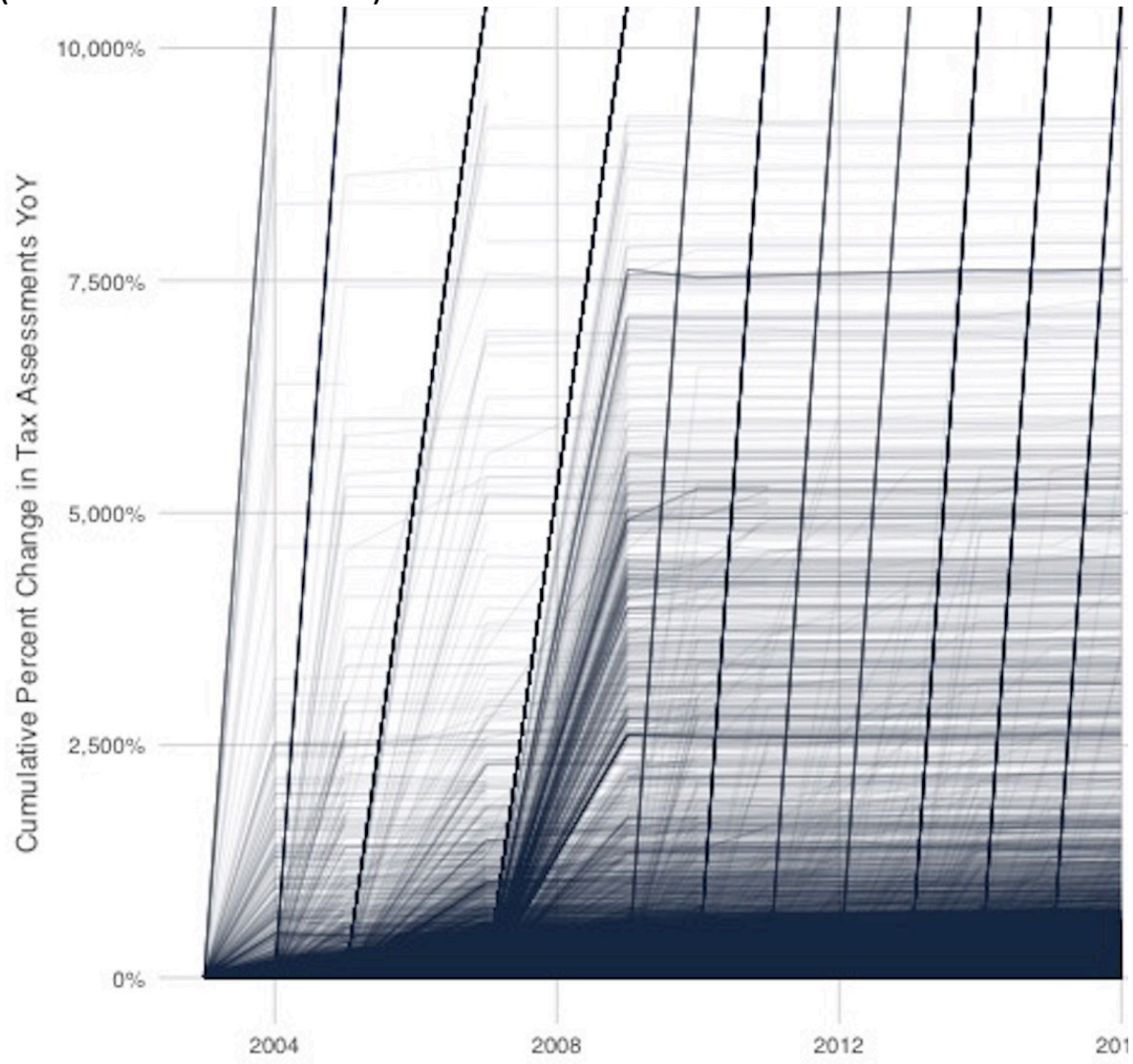
What is Our Target Variable?

- “Re-assessments” are not marked in the data
- Assessed Values increase annually between 4-6% on their own
- What we’re interested in: sudden, non-linear increases in assesed value



Defining a “Sudden, Non-Linear Increase” in Assessed Tax Value

Cumulative YoY Change in Assessment Value:
(Each Line is 1 Tax Lot)

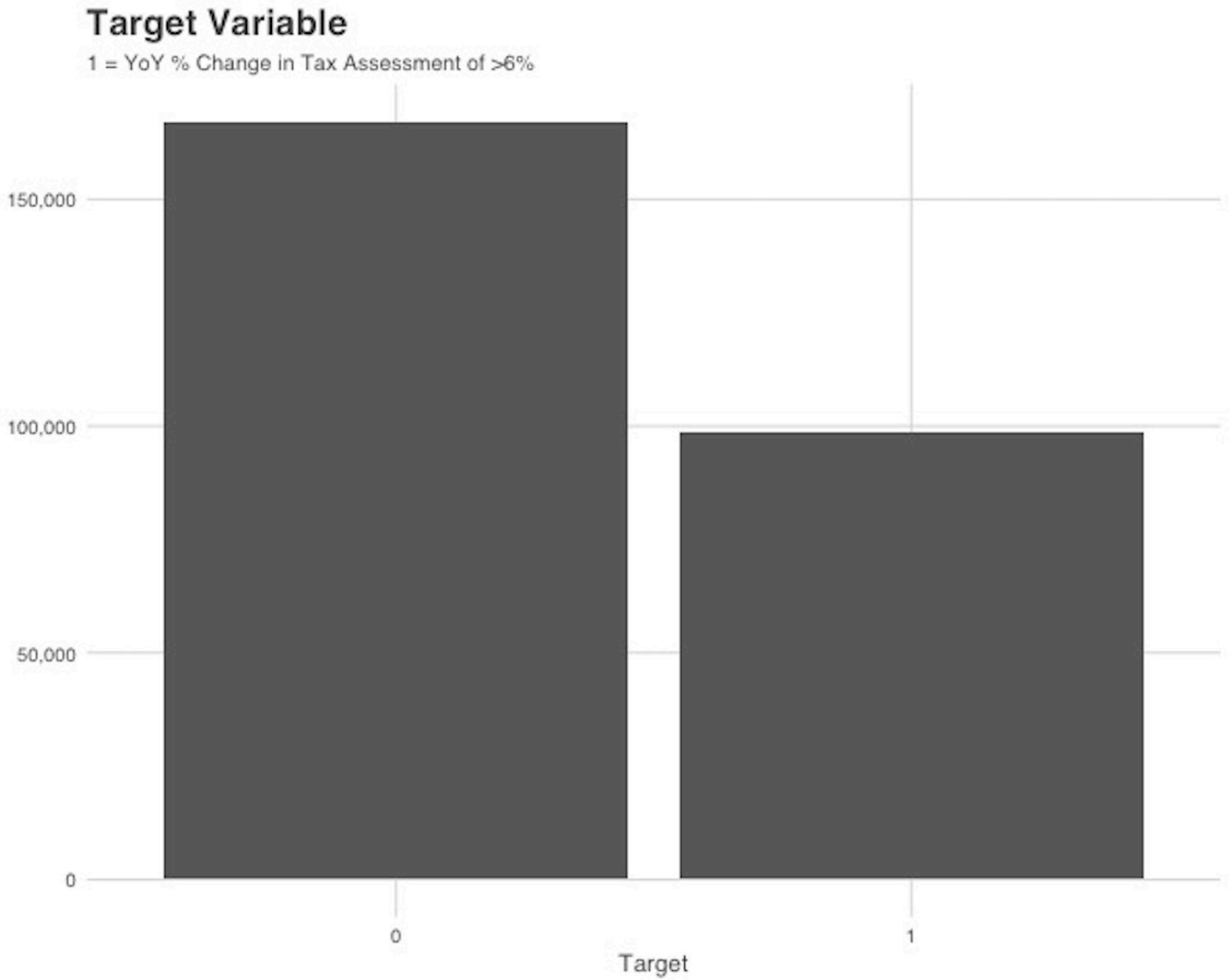


Median Change By Year:

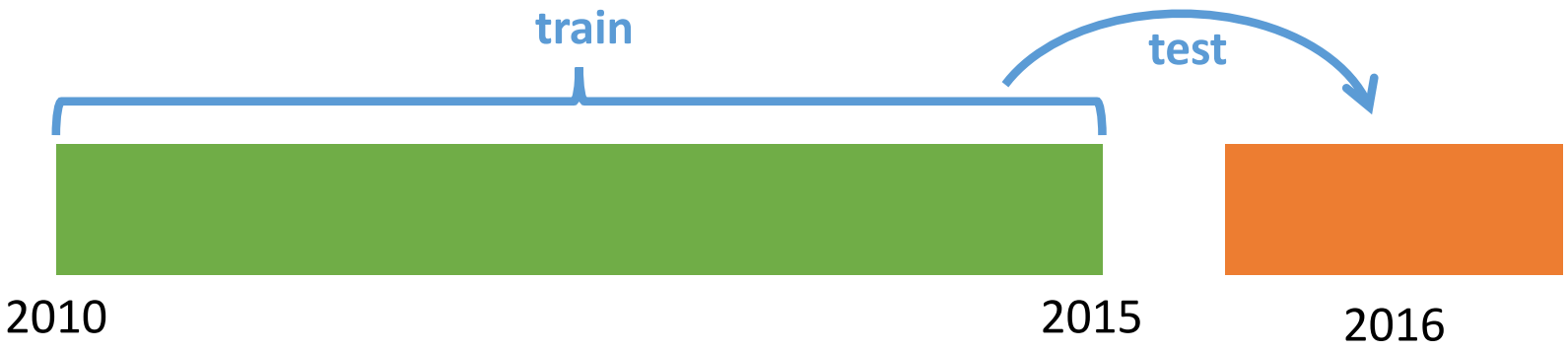
Year	Median YoY Assessment Change
2011	6.0%
2012	3.8%
2013	4.8%
2014	6.0%
2015	6.2%
2016	5.8%

- **Notice in the graph: a handful of tax lots increase dramatically every year**
- **A change of >6% YoY is well above average**

Our Target Variable: >6% Increase YoY



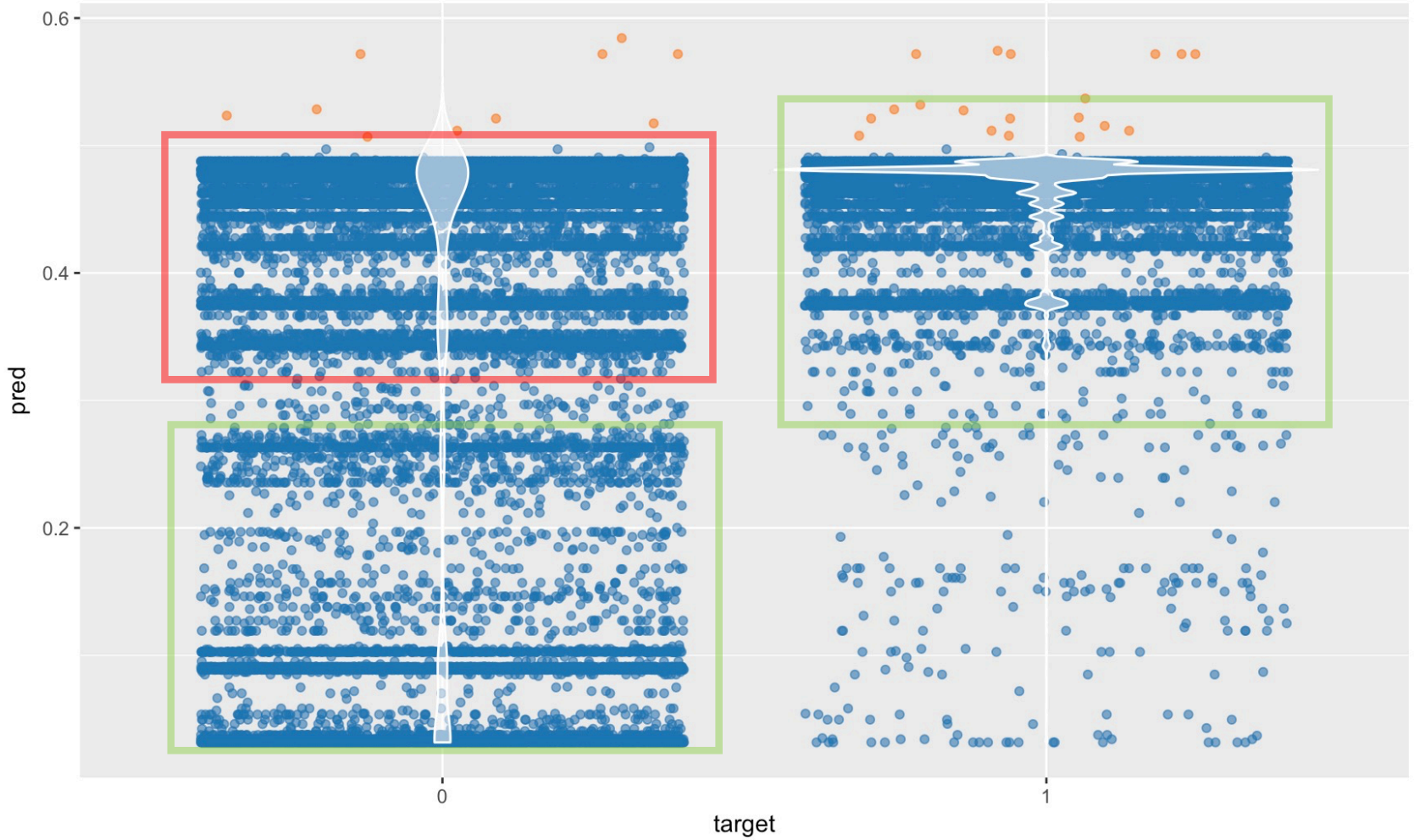
Using Out-of-time Validation



- Since this is a time-series, we want to be able to predict one year into the future
- Training data: 2010-2015
- Testing data: 2016
- This ensures that our model generalizes well into the near-future

Baseline Model

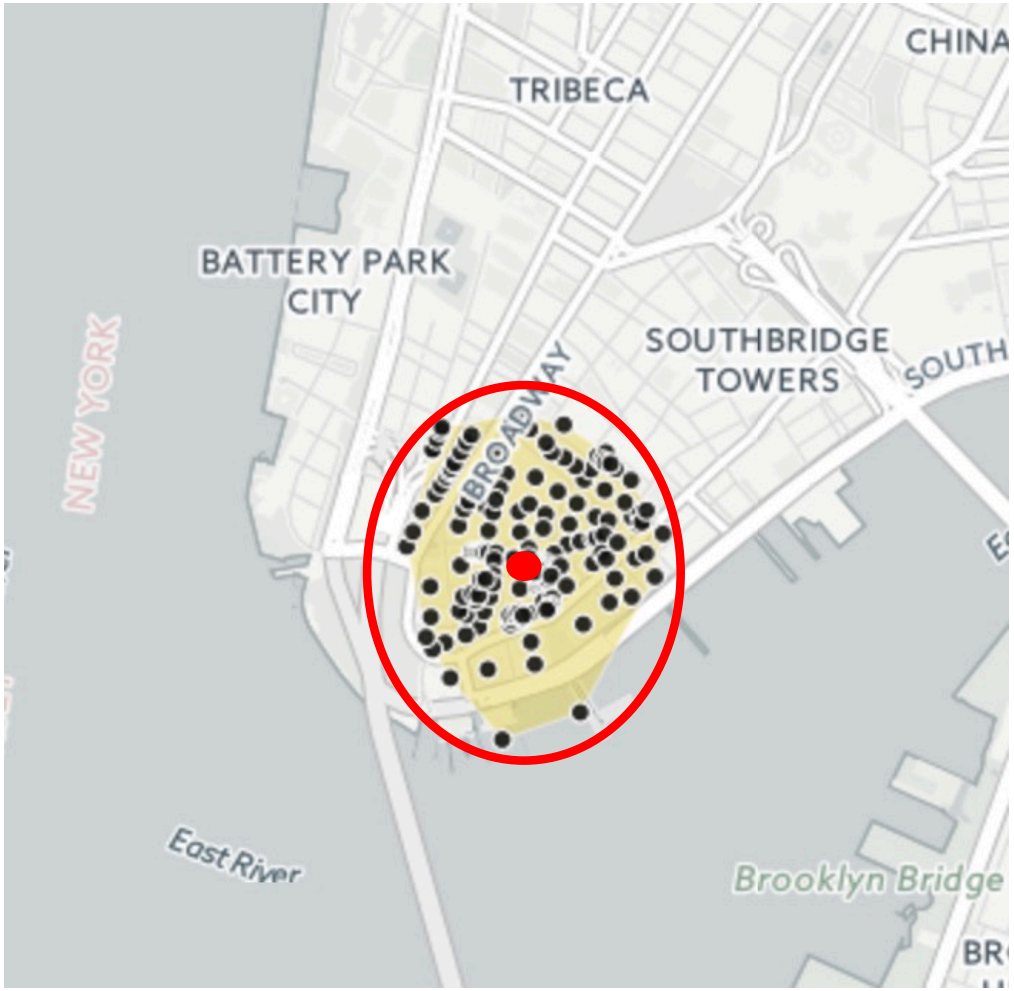
- GBM implemented using XGBoost (in R)
- All numeric variables plus some categorical vars
- (Red box below) Baseline model has a false positive problem



Improving the Model Using Geo-Spatial Kernel Features

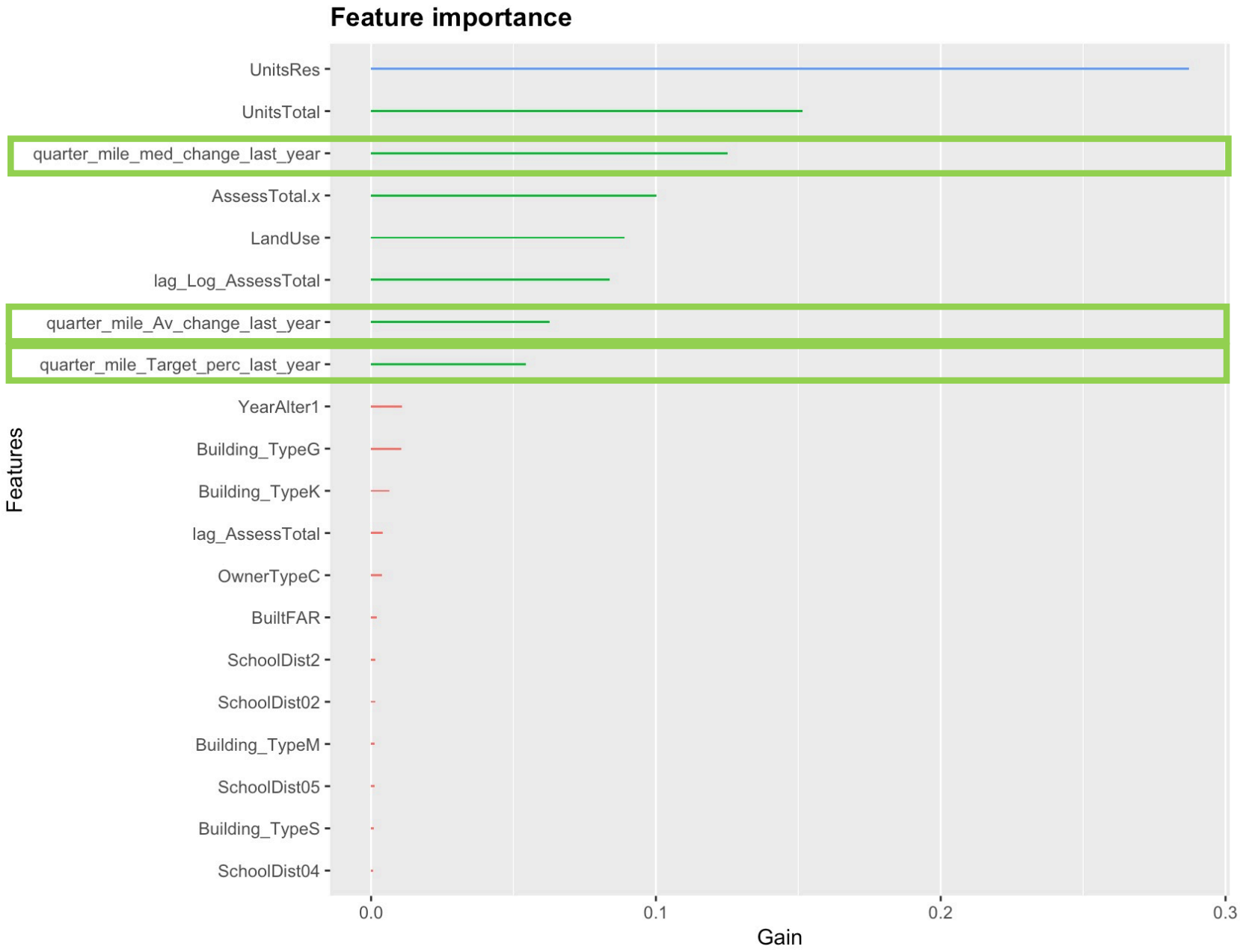
Location, Location, Location!

- For each property, identify all other properties in a given radius (e.g., 0.25 miles)
- Calculate metrics from within radius, e.g., moving averages in historical YoY assessment changes
- Very time consuming (108 user hours for modeling data)
- Technique could be extended for other metrics, different kernel shapes, etc.



Radial Metrics Were Some of the Most Influential Variables

Final Model AUC: 0.71



Future Research / Improvements

- Rather than excluding incomplete observations, interpolate them
- Expand to outer Boroughs
- Use all years of data
- Refine and add more geo-spatial kernels (and optimize!)
- Split models by asset type (Office, Apartment, etc)
- Use address-level data rather than tax-lot

Resources

- Github link for project: https://github.com/timkiely/predicting-nyc-tax-assessments
- NYC PLUTO Data: https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page
- Inspiration: http://www1.nyc.gov/assets/finance/jump/property-data-maps/docs/quintos_latenteffect.pdf