SEPTEMBER 2018

THE FUTURE OF SOLAR ADOPTION

PREDICTING ADOPTION TO IMPROVE MARKETING AND SALES FOR MANUFACTURERS AND INSTALLERS

JERRY CHIANG
CARL RIOS
SEBASTIAN SOBOLEV

GEORGETOWN UNIVERSITY

THE PROBLEM

- There has not been much data collected on installation of solar PVs at the national level

- Data on solar installers are collected locally; solar incentives are legislated at the state level

- Based on this bifurcated structure, installations are typically driven by local utility programs, business models, and community awareness.

- Solar adoption is therefore a matter of **consumer choice and preferences**

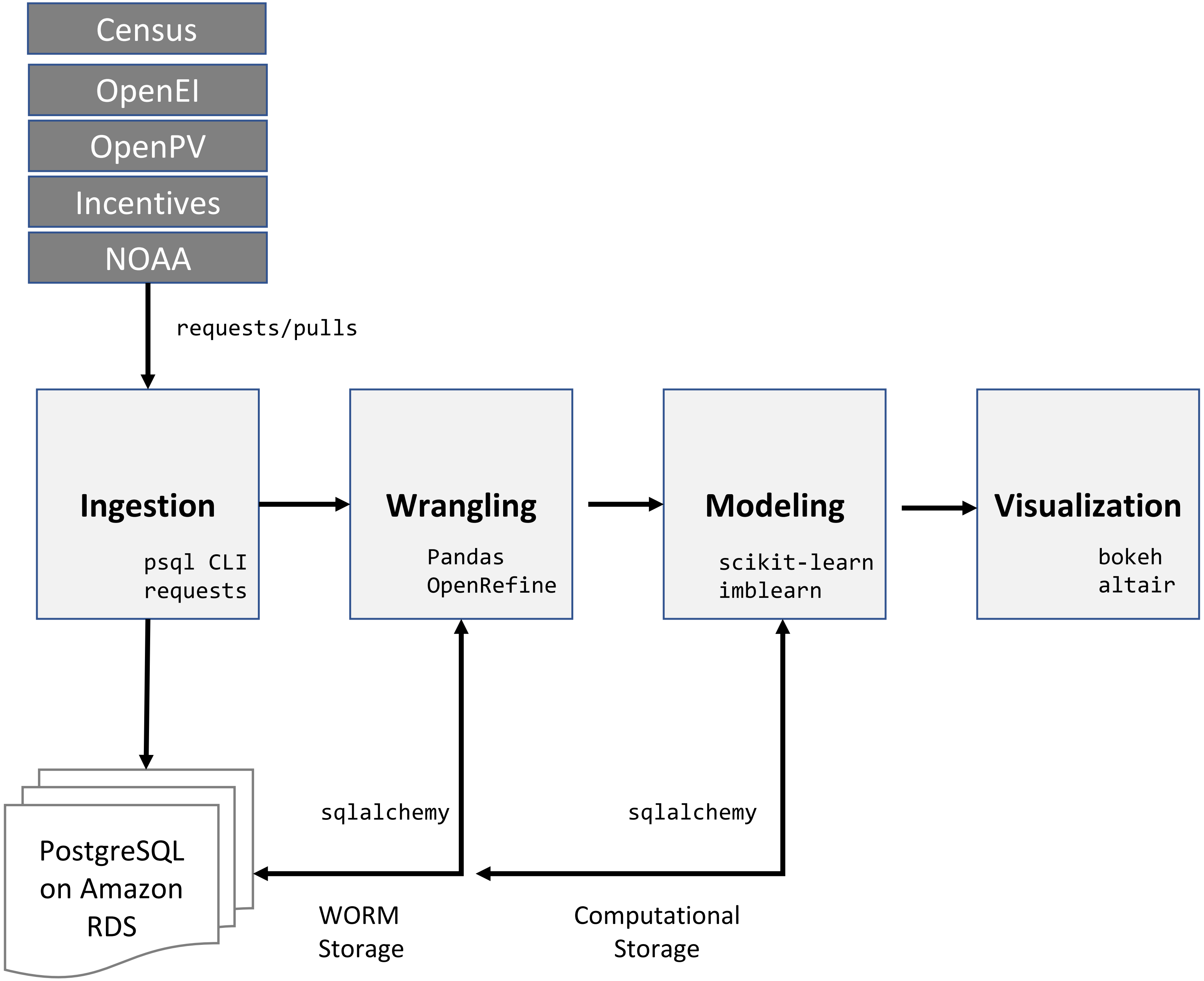
- This begs the question- given available data, can we figure out where consumers are likely to adopt?

- ***Which zip codes are likely to have high, medium and low adoption rates?***

Source: National Renewable Energy Lab data, mapped with altair; each dot represents a zipcode with one or more residential solar installations

HYPOTHESIS

Economic, demographic, and regulatory attributes of a zip code can predict whether consumers in that zip code are likely to switch to solar energy

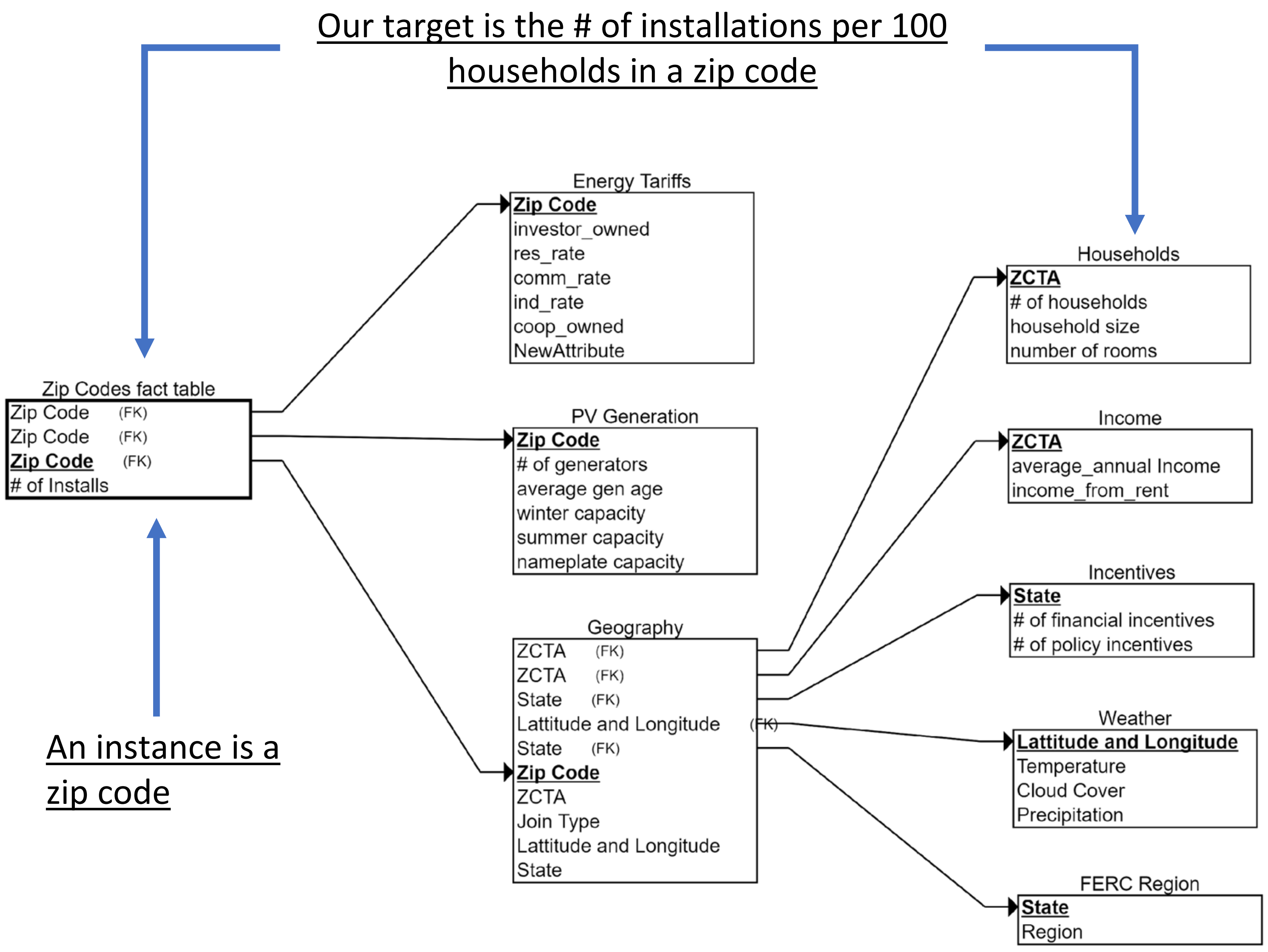


ARCHITECTURE

- Data ingestion methods varied by source – some CSVs, some REST APIs, some text files, etc.
- A PostreSQL instance on Amazon RDS was used for raw and computational storage, given the ease of interaction with sqlalchemy and pandas
- OpenRefine was key in cleaning user-entered data, saving us from having to implement similarity searches manually
- Modeling was done using scikit and imbalance learn APIs
- The final output was generated with altair, which has an intuitive way to map lat/long

Ingestion & Wrangling

- We relied on a variety of government sources
  -NREL
  -EIA
  -Census Bureau
  -NOAA
  -OpenPV
- Our key data source, the openPV database of solar installations across the country, are collected on a voluntary basis
- NREL claims they sanitize, de-duplicate, and quality control this data
- <u>An instance in our dataset is **zip code**</u>, and we aggregated all individual install data to this level, turning ~1m records into ~14k instances

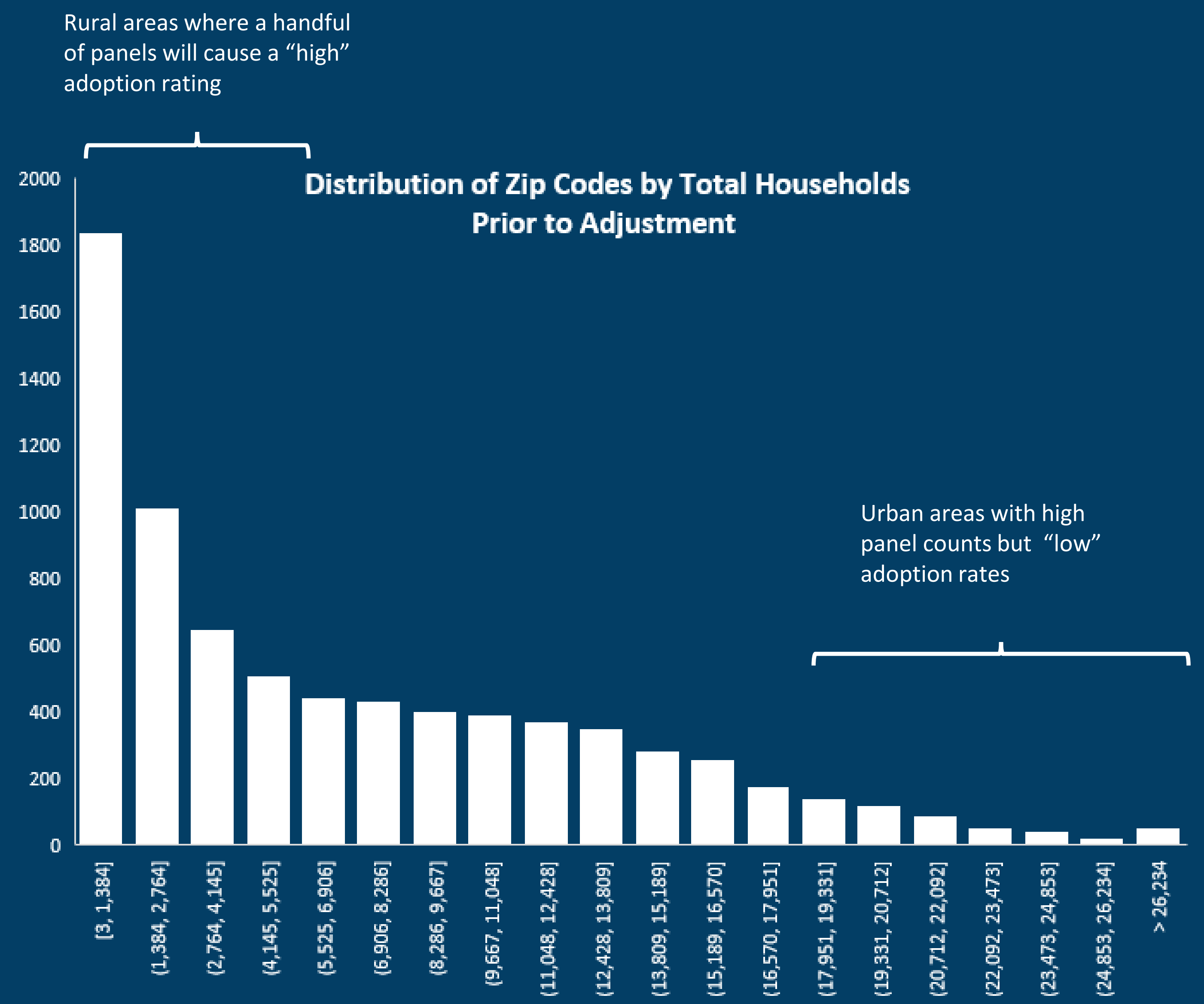


MODELING – ENGINEERING THE TARGET

- Dividing the number of installs by the number of households by Zip Code would overwhelmingly favor zip codes with few houses

- We standardized the calculation to the number of installs per 100 households in a zip code to account for this and used a histogram to eyeball classes (High is anything above 1 in 100)

- We later realized there are “off-the-shelf” solutions to this problem – for example, we may have wanted to use <u>Congressional Districts as an attribute</u>, <u>since the demographics information is more or less fixed</u>



MODELING – CLASS IMBALANCE

The vast majority of our zipcodes wound up in the “Low” adoption category, causing our early machine learning experiments to completely overlook the “high” class

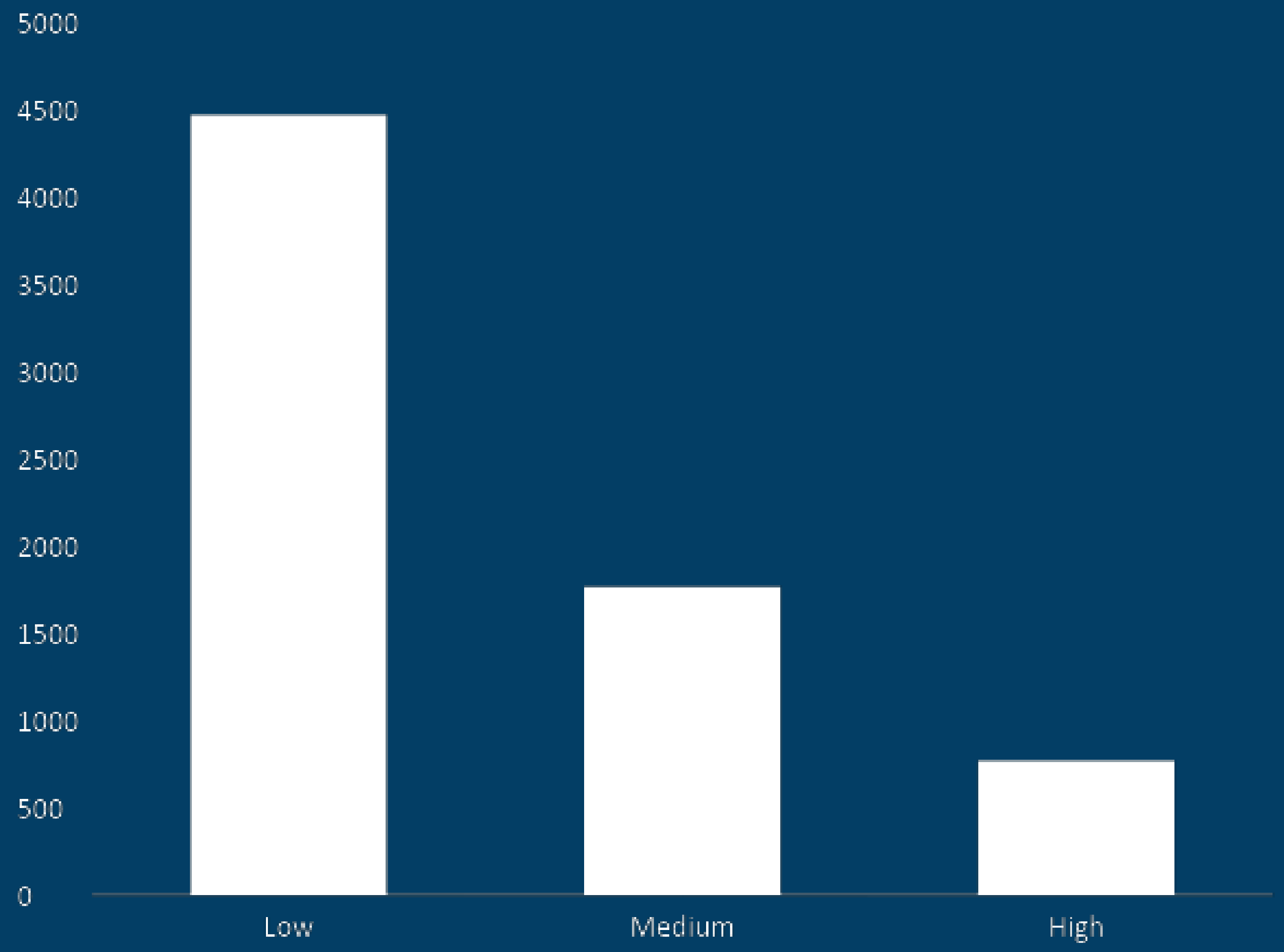
We also made a big assumption: no such thing as a zip code that would have no adoption – there was no data source that described the **absence** of panels

Experimented with:

-Naïve Over-sampling

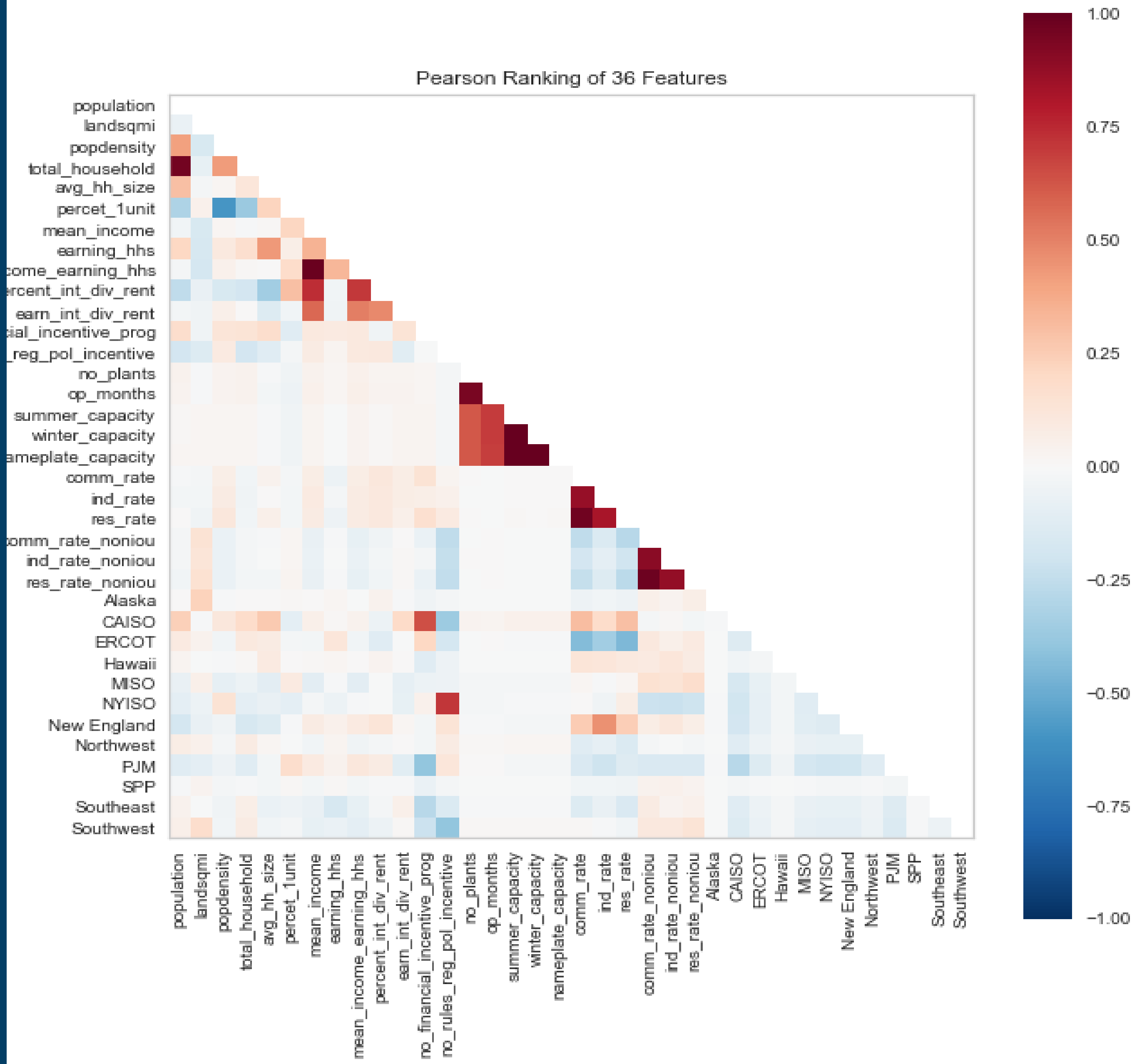
-SMOTE (Synthetic Minority Over-sampling)

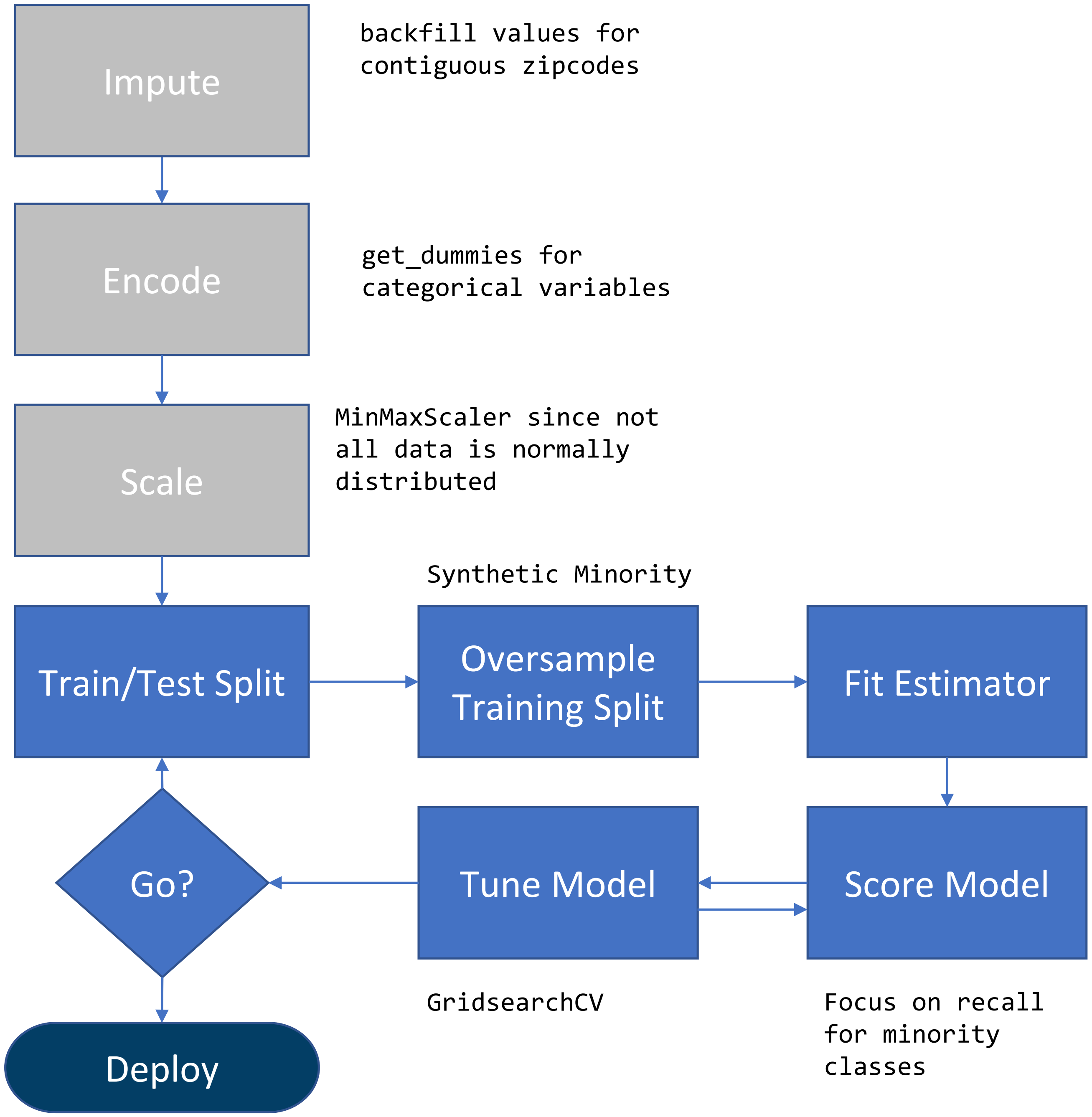
-ADSYN (Adaptive Synthetic)



FEATURE ANALYSIS

- Several features essentially described population in different ways (number of households, etc.) so we dropped those
- Regulatory regions and incentives tracked closely where the regions were small (NYISO, CAISO)
- Bigger PV generators also happen to be bigger across the board, regardless of seasonality
- The distinction between commercial and industrial tariffs also became less and less important to us as we built the model out





MODELING WORKFLOW

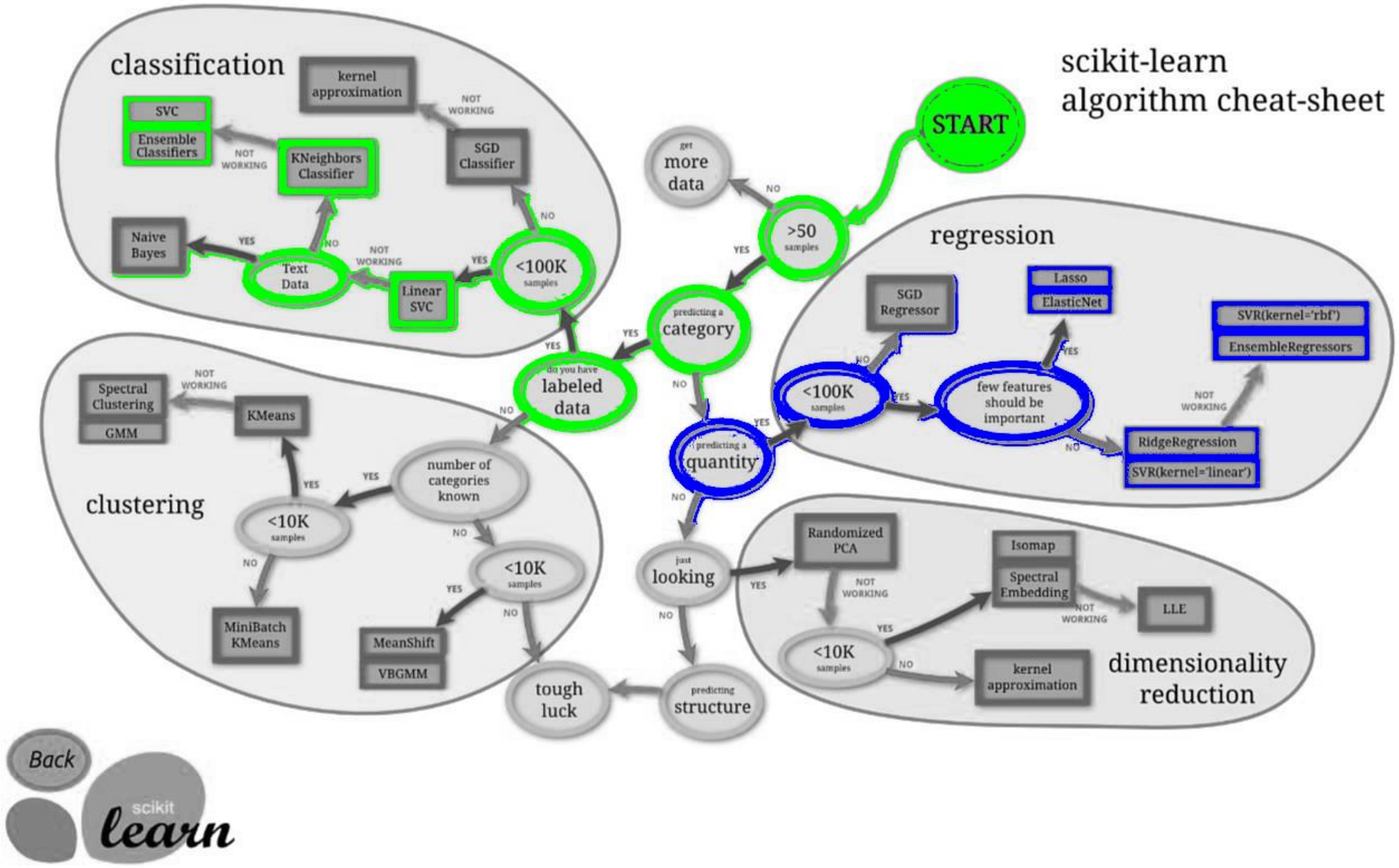
- MinMax Scaler was used to account for the fact that not all variables were (perfectly) normally distributed (income)
- Recall for the ‘middle’ and to a lesser degree ‘high’ class was problematic, so we focused on that in scoring
- GridsearchCV used for tuning – if you don’t know what you’re doing (we didn’t) this can be very resource-intensive

MODEL SELECTION

- We chose to pose this as a classification problem (Low/Med/High -probability of adoption)
- This could have also been a regression problem, if we were trying to predict adoption as a continuous variable
- It’s easy to think of this as a similarity problem – what do High/Low/Med zip codes have in common?
- So we got a lot of mileage out of K-Nearest Neighbors

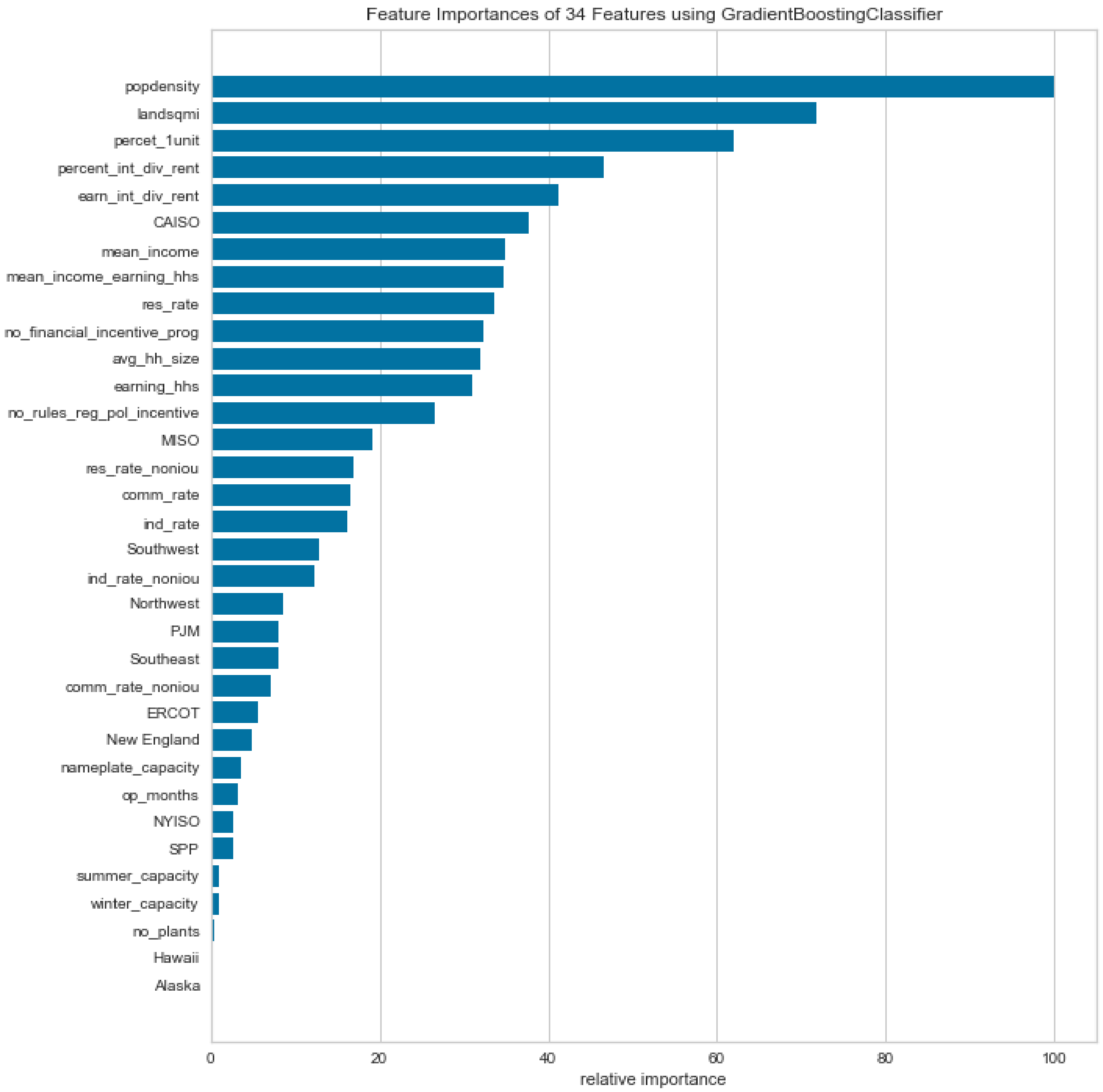
Primary path

Alternate path, if only there were more time



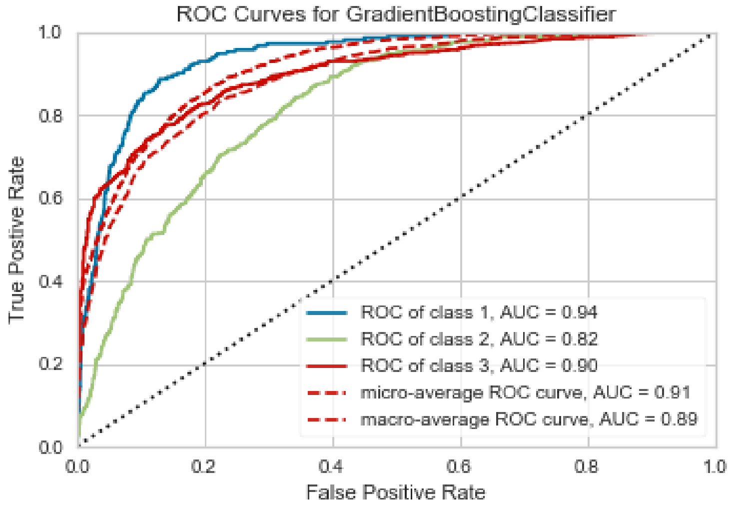
FEATURE IMPORTANCE

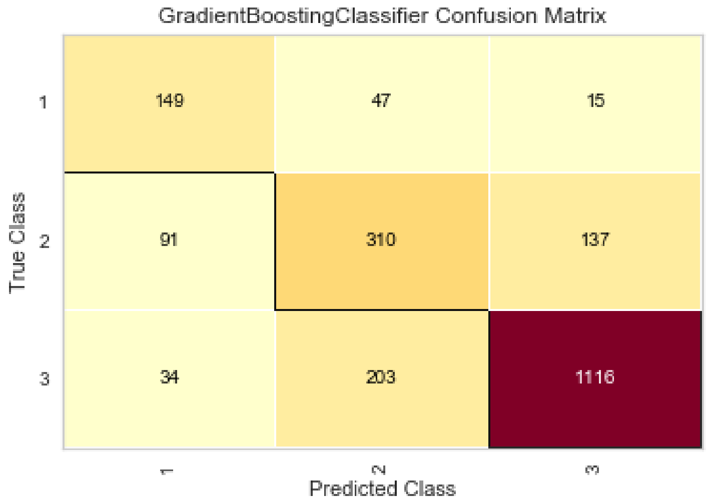
- We relied on domain expertise for our feature selection
- At a glance, PV adoption is still very much for wealthy households with a lot of space (income, density, square miles per zip)
- Incentive programs didn't move the needle as much as we had assumed they would
- Tariff rates seemed totally unimportant, suggesting that adoption is more a matter of customer preference than rational economic choice

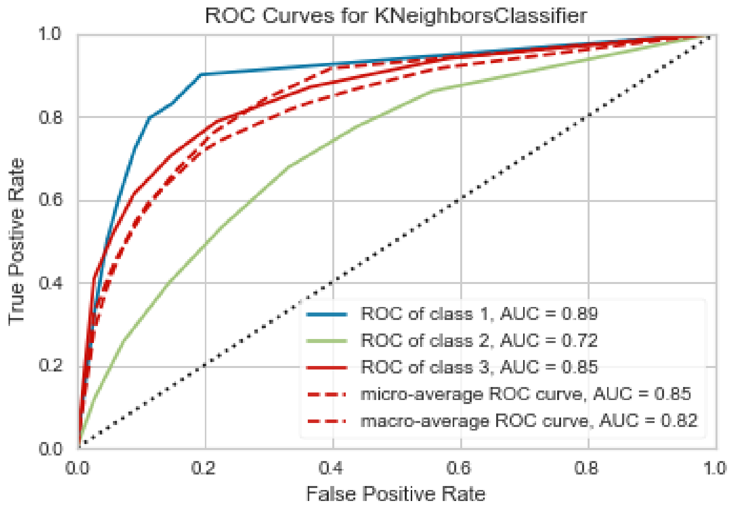


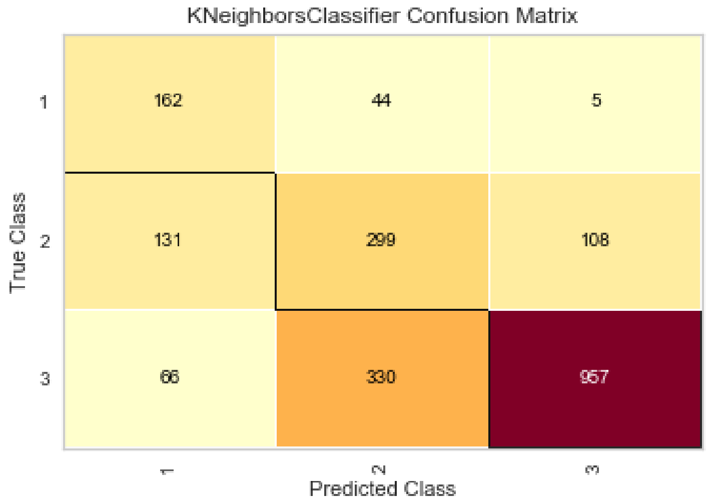
PERFORMANCE - One Estimator

- Best results with Gradient Boosting Classifier and K-Nearest Neighbors Classifier
- Hyperparameter tuning on GBC produced a greater variation in performance that KNN (there are a lot more, too)
- In all, ~.74 F1 score was our maximum
- NuSVC (Nu-Support Vector Classification) gets honorable mention, and a voting role in our ensemble model

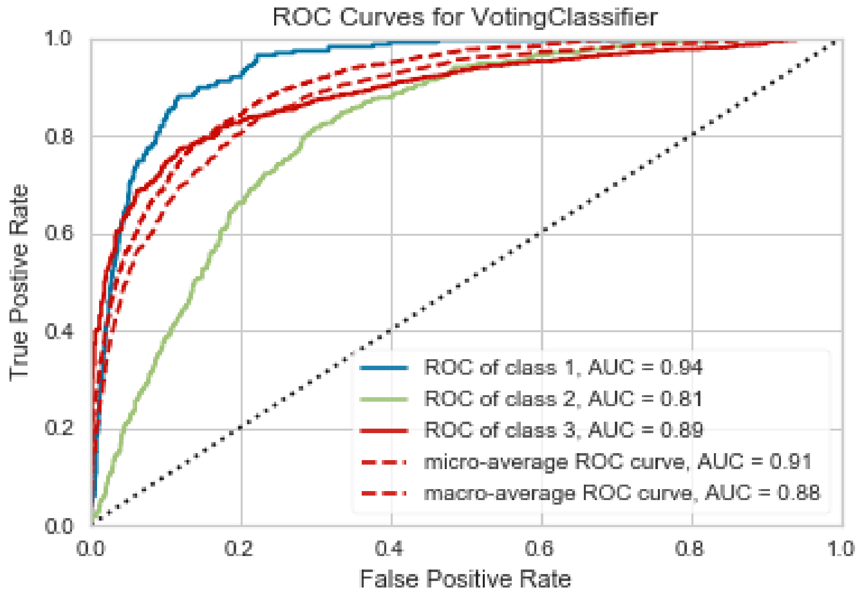


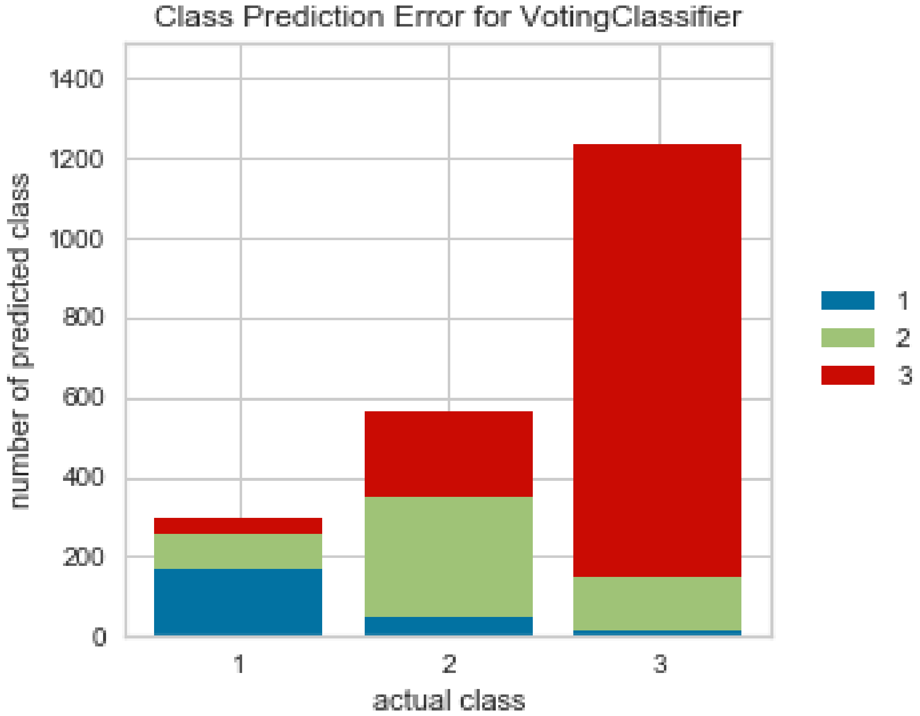




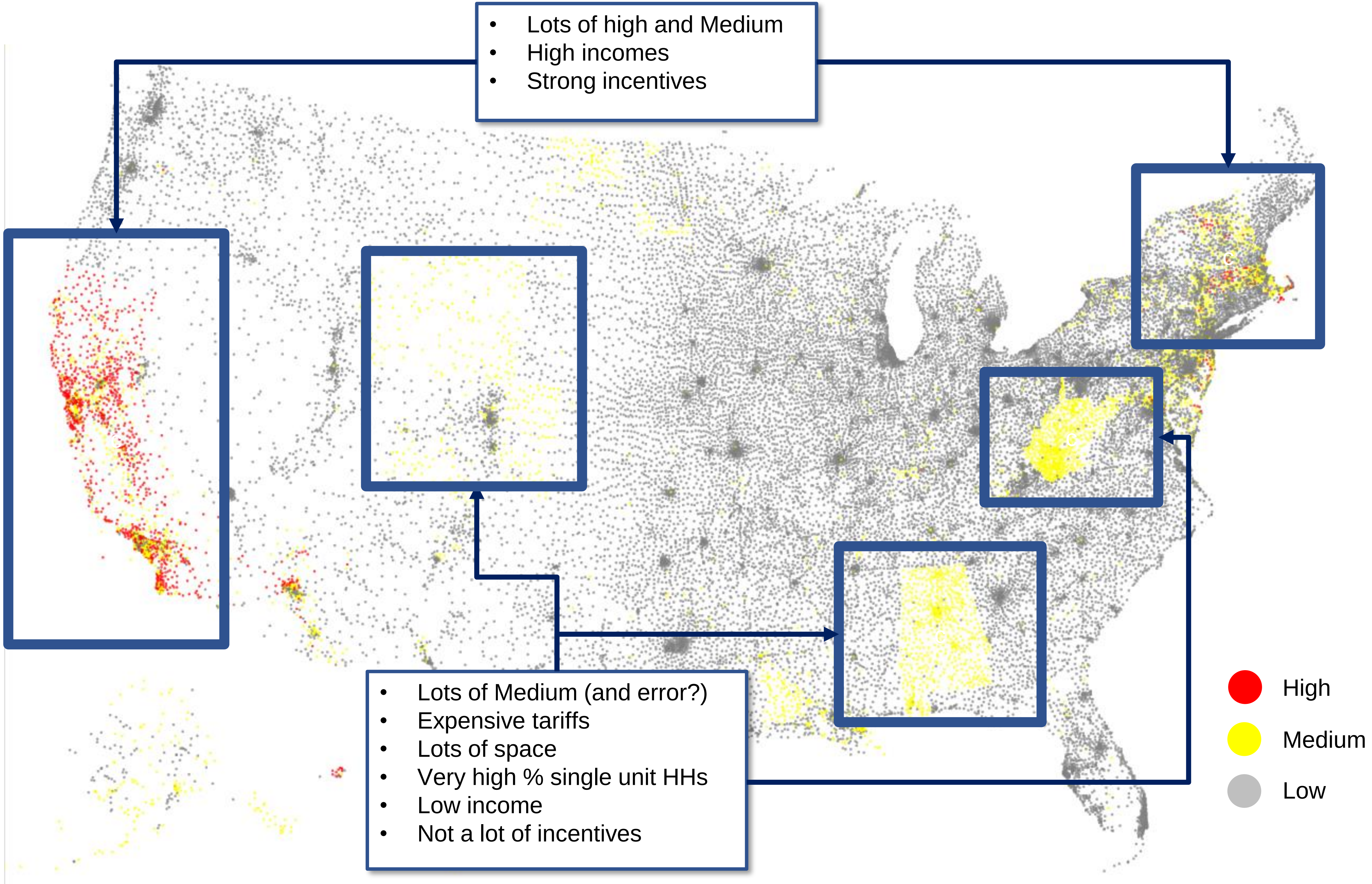


FINAL MODEL – Voting Classifier



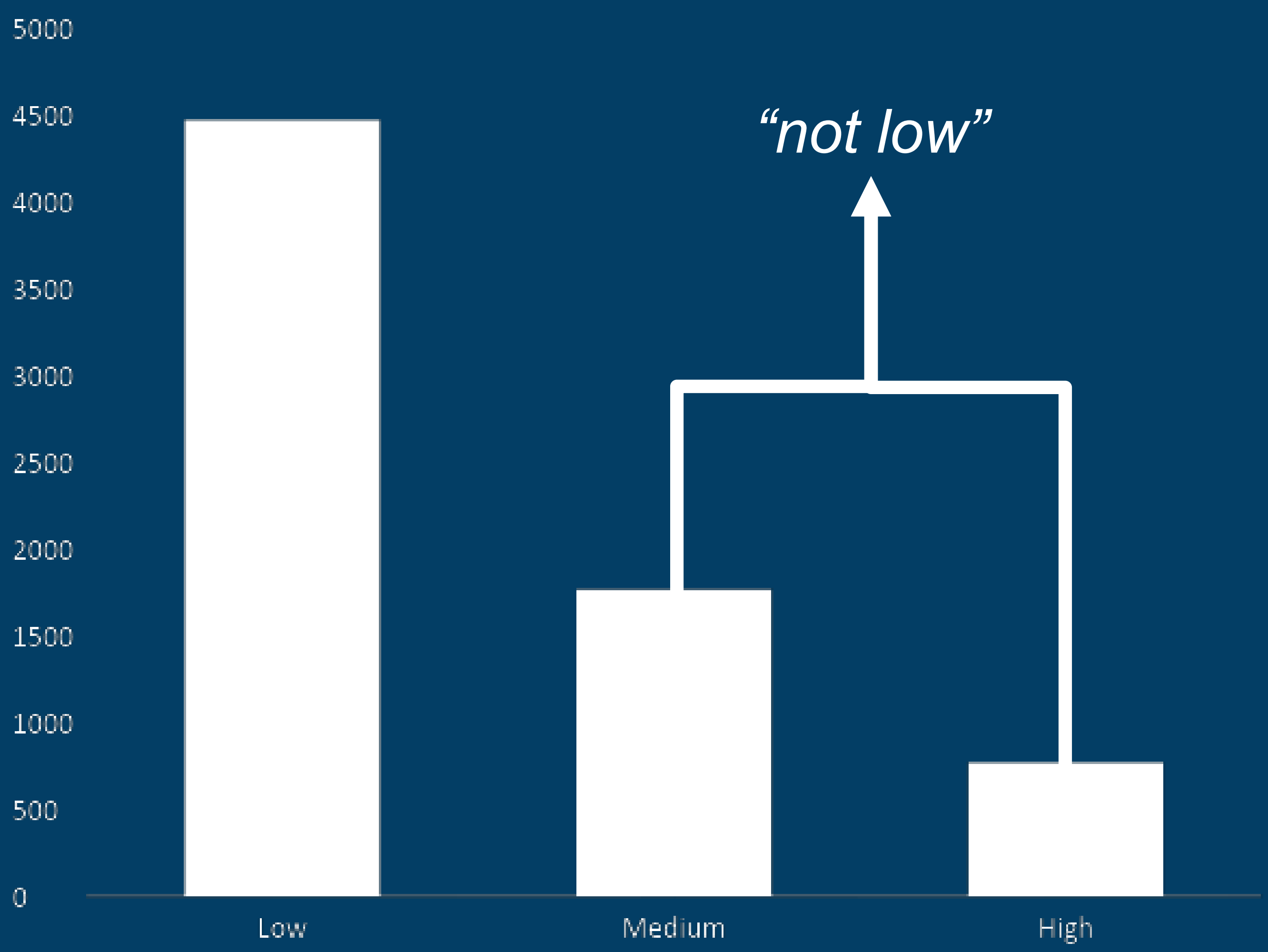


- A voting classifier with soft voting produced our best results
- All of our models struggled with the “medium” class, so we gave the models that struggled the least with it (Random Forests) the greatest weight in voting
- Making the target binary, rather than multiclass, does not take much of the business value out of the model, but improves the accuracy quite a bit – to around .92


• Lots of high and Medium
• High incomes
• Strong incentives
• Lots of Medium (and error?)
• Expensive tariffs
• Lots of space
• Very high % single unit HHs
• Low income
• Not a lot of incentives
High
Medium
Low

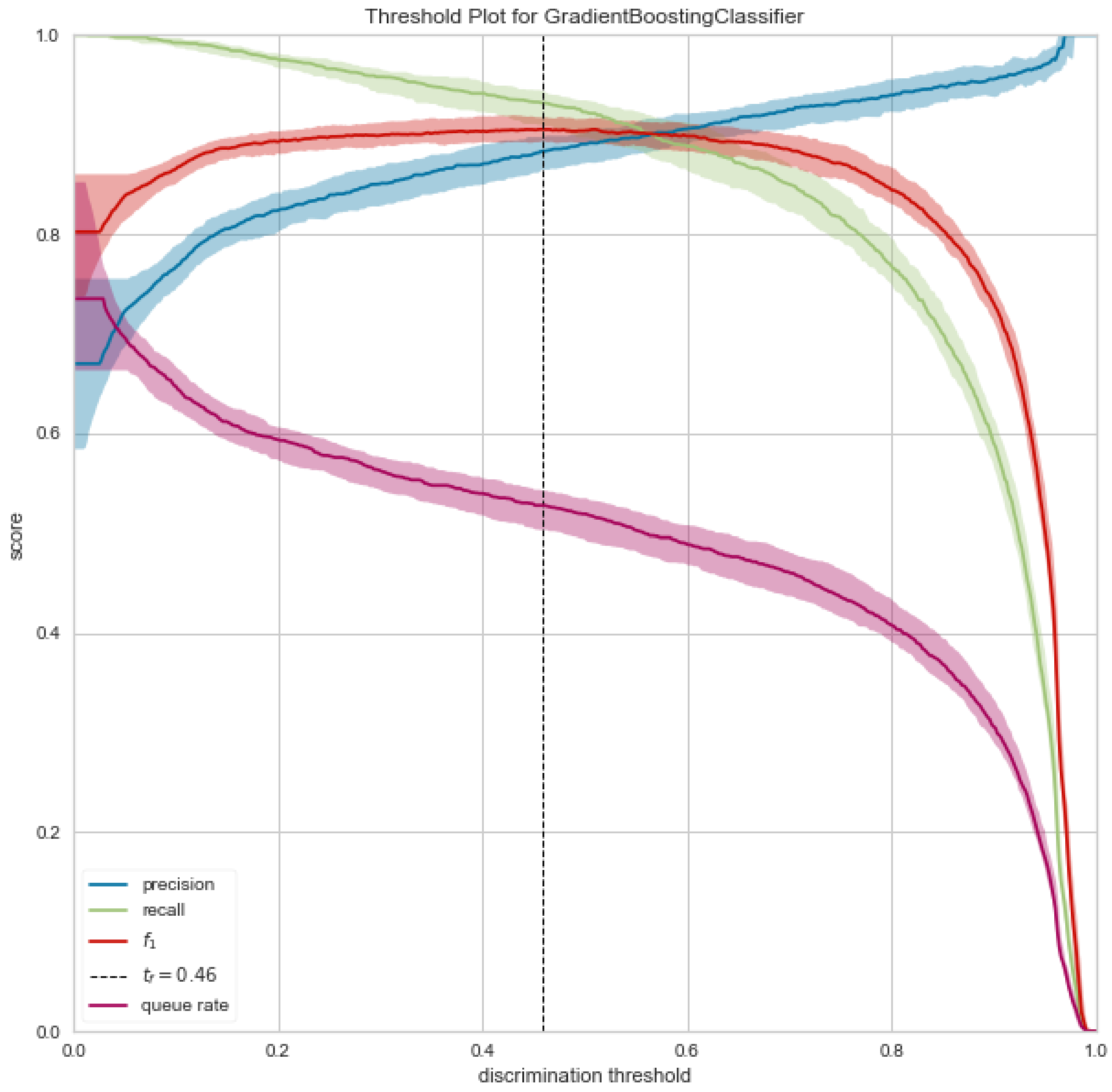
REFRAMING THE PROBLEM – TWO CLASSES

- A quick experiment: same models, same features
- Slight modification of hyperparameters - swapping the loss function to ‘exponential’, which is only supported for two-class problems, improved performance
- Went from .76 score, achieved after considerable effort and tuning, to a .93 score achieved more or less with “out of the box” estimators

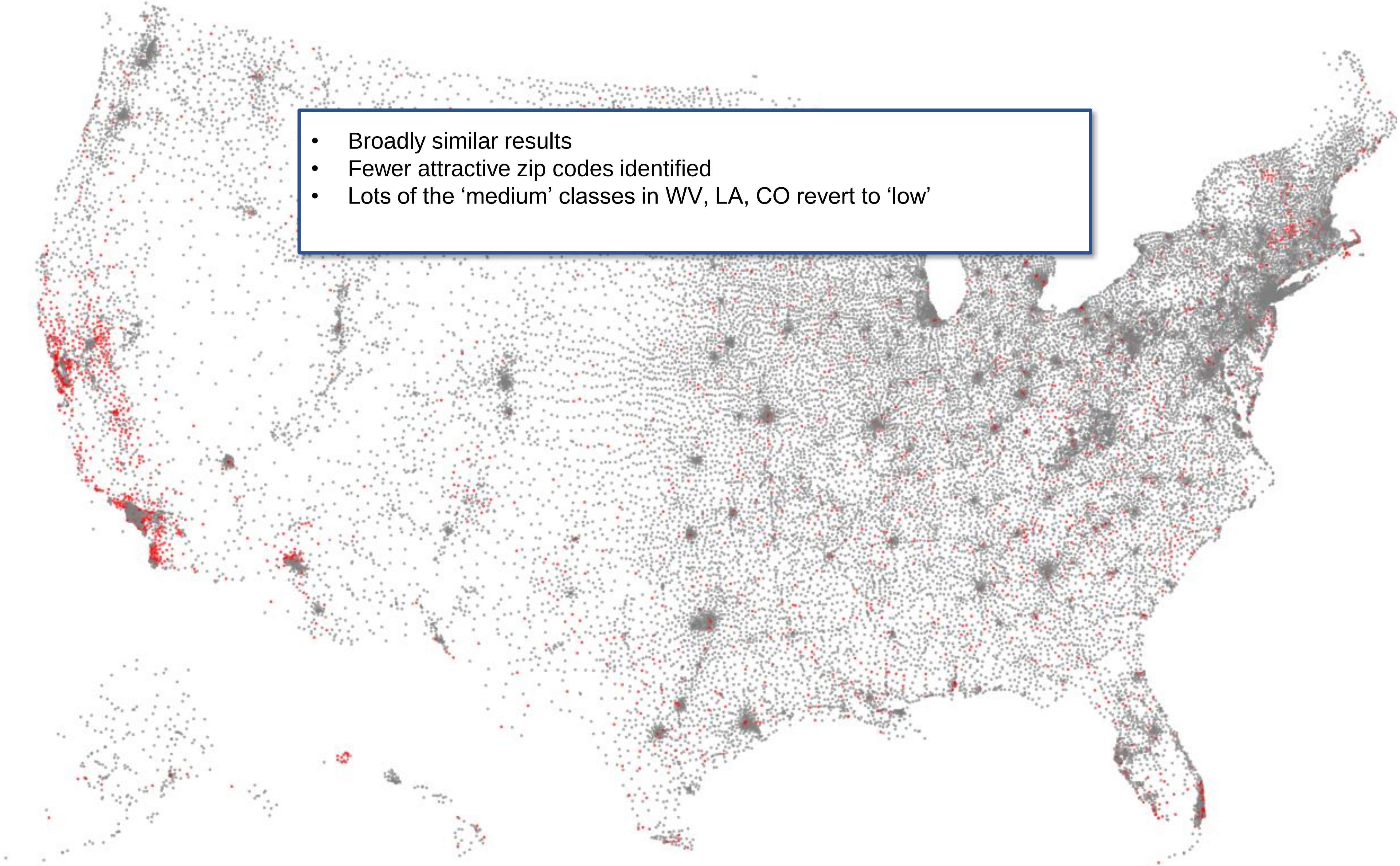


Discrimination Threshold

- Optimal threshold at .46 seemed close enough to the default to leave as is
- ~0.9 – 0.93 score with a good balance between precision and recall gave us higher confidence in this model



- Broadly similar results
- Fewer attractive zip codes identified
- Lots of the ‘medium’ classes in WV, LA, CO revert to ‘low’



BINARY VS. MULTI-CLASS – A matter of expected value?

- What is the cost of taking action on the basis of this information?
- What is the benefit derived from a true positive?
- Cost of a false positive?
- Money left on the table from a false negative?
- Can adding a third dimension to the multi-class view (e.g. income) mitigate risk?
- **All of this can be tuned according to the specifics of a given business or use case**

NEXT STEPS

- **Predict adoption as a continuous variable using regression methods**
  - arguably easier to use in business analysis
  - a little harder to interpret

- **More data, and algorithmic feature selection**
  - our data sources have lots more to offer (age of house, etc.)
  - and there are other data sources that could supplement

- **More robust visualization**
  - more dimensions on the map – size by income, etc.
  - interactivity, at a minimum with tool tips

- **Same data, other applications**
  - time series analysis
  - commercial, industrial, non-profit, and government markets