# Real Estate Investment

**Which zip codes are primed for an increase in property value?**

Attempting to analyze fluctuations in price per square footage utilizing *neural networks* and *natural language processing*

# The Data

Data is sourced through the commonly used Zillow datasets as well as from social media platforms. There will be three versions of the data and each version will build on the prior.

Strictly Zillow information

Zillow + local business information

Zillow + local business + natural language processing (crowdsourcing)

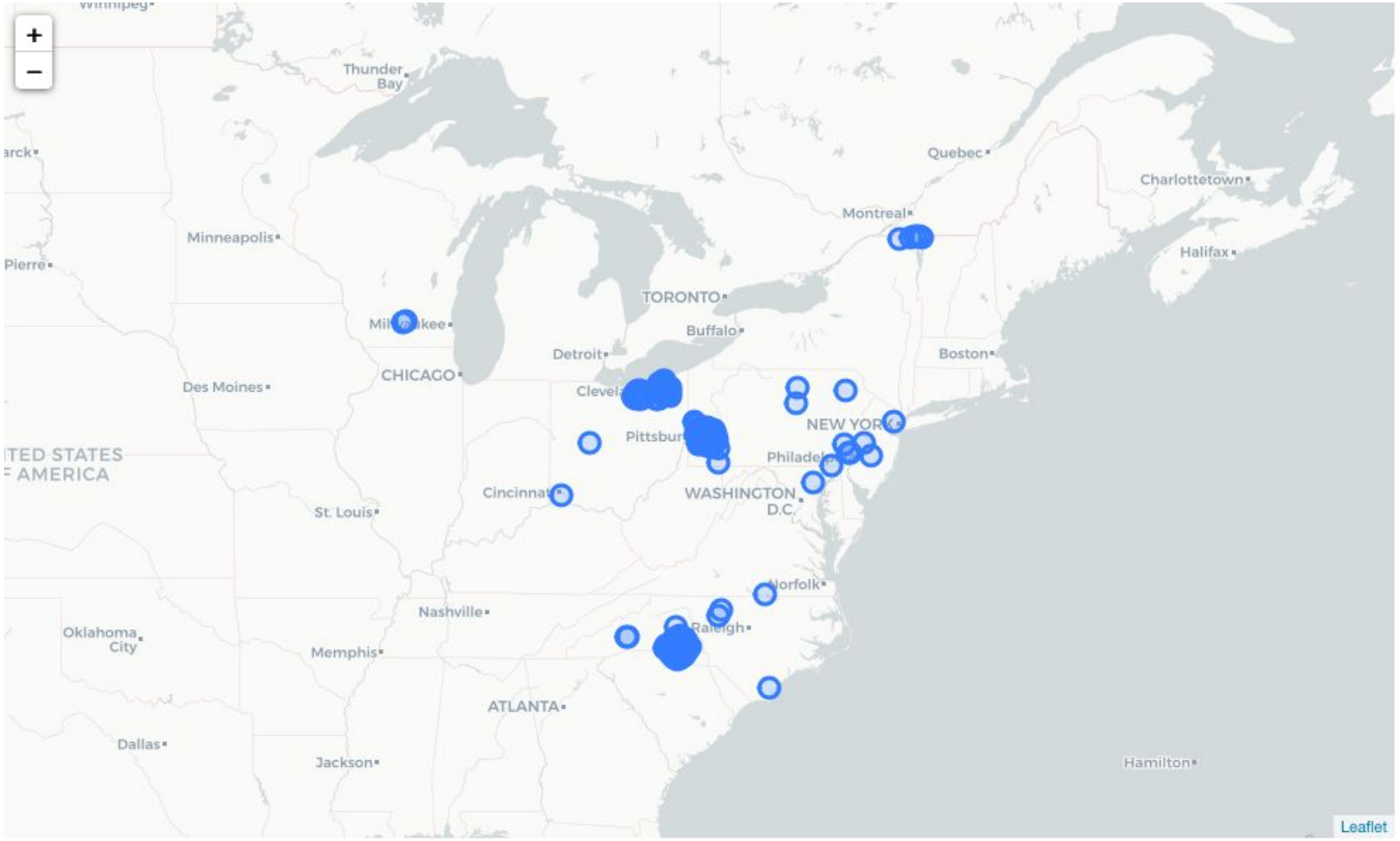
# Data Challenges

**Zillow**:

- Not all files had similar time periods
- Some data files were lacking in volume of zip codes
- Some data files were already aggregated and fit a different format
- Many missing values in data
- Each file was essentially it's own feature

**Business and NLP**:

- Text data for NLP was 5 GB and was very difficult to work with computationally
- Limited number of zip codes due to source of data


Thunder Bay
Quebec
Charlottetown
Montreal
Minneapolis
Halifax
TORONTO
Buffalo
Detroit
Boston
CHICAGO
Des Moines
NEW YORK
Pittsburgh
Philadelphia
WASHINGTON D.C.
Cincinnati
St. Louis
Norfolk
Nashville
Raleigh
Oklahoma City
Memphis
ATLANTA
Dallas
Jackson
Hamilton
Leaflet

# Zillow Feature Examples

- **Median Listing Price for all homes**

- **Price to rent ratio**

- **Median listing price per square foot for single family homes**

- **Percentage of homes increasing in value**

# Example Zillow File

| RegionName | 2010-09 | 2010-10 | 2010-11 | 2010-12 | 2011-01 | 2011-02 | 2011-03 | 2011-04 | 2011-05 | 2011-06 | ... | 2017-03 | 2017-04 | 2017-05 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **99801** | 5.001480 | 12.059667 | 9.721639 | 12.823805 | 12.067908 | 6.735467 | 12.817048 | 5.351705 | 7.022677 | 11.403385 | ... | 8.149935 | 7.839226 | 8.34047 |
| **99709** | 13.332107 | 11.447750 | 20.230705 | 17.569812 | 9.501848 | 16.866147 | 25.336712 | 14.339757 | 14.388552 | 11.076664 | ... | 15.020218 | 15.179652 | 9.60874 |
| **99705** | 14.238275 | 11.455669 | 21.931982 | 19.715816 | 13.152595 | 16.048516 | 16.539511 | 13.047547 | 15.500396 | 18.073011 | ... | 17.252181 | 14.419267 | 13.72524 |
| **99701** | 12.502596 | 3.868248 | 11.226388 | 8.029469 | 11.557336 | 13.020347 | 6.255375 | 7.774582 | 10.831500 | 12.234200 | ... | 11.990818 | 11.121010 | 13.57381 |
| **99669** | 12.342694 | 9.547881 | 9.399721 | 11.789248 | 13.575353 | 8.824402 | 11.448679 | 12.713187 | 12.052649 | 9.105020 | ... | 8.814888 | 10.137303 | 13.93226 |

# Forecasting and Data Generation

## “Now Time”

- Create a point in time and ensure that all training information has occurred beforehand as is that point in time was today

- Test model on how well it is able to predict price per square footage growth in six months later then the “Now Time” date

## Data Generation with “Now Time”

- Generate multiple instances of “Now Time” to add train the model on different time periods and different instances of “NowTime”

- Settled on training the model on the same time period each year as well as the six months preceding it. This was to adjust for seasonality bias

# Business and social media data

## Business:

- Business data is a collection of posts about each business within a certain zip code
- Number of businesses and volume of posts were aggregated to create features
- Volume of businesses compared to volume of reviews taken into consideration

## NLP:

- Social media text information
- Attempting to identify trends in comments about businesses in trending neighborhoods

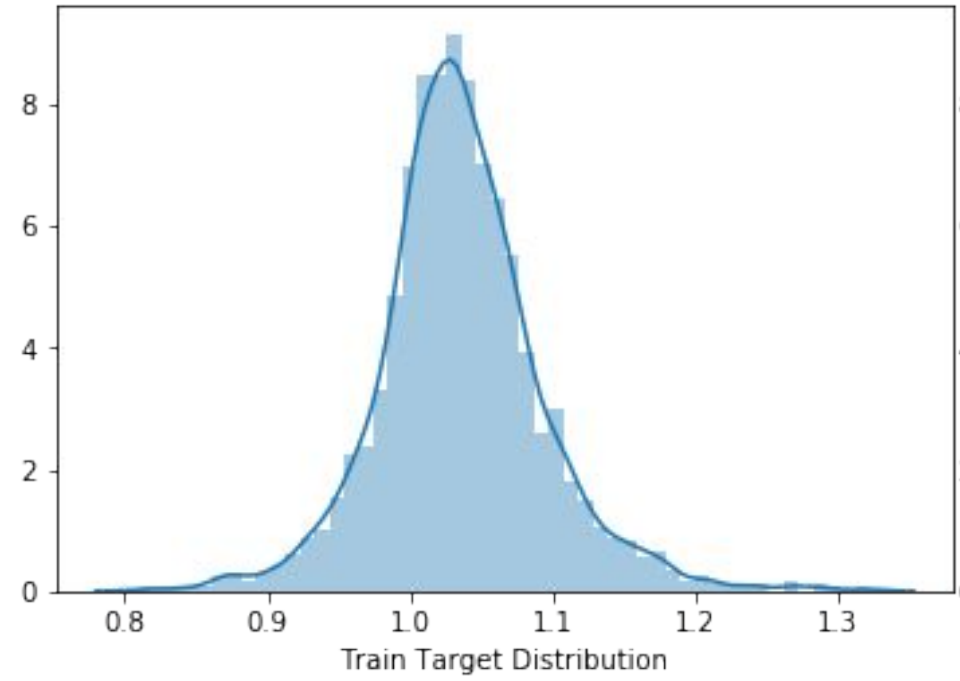
# Condensing Reviews and Businesses

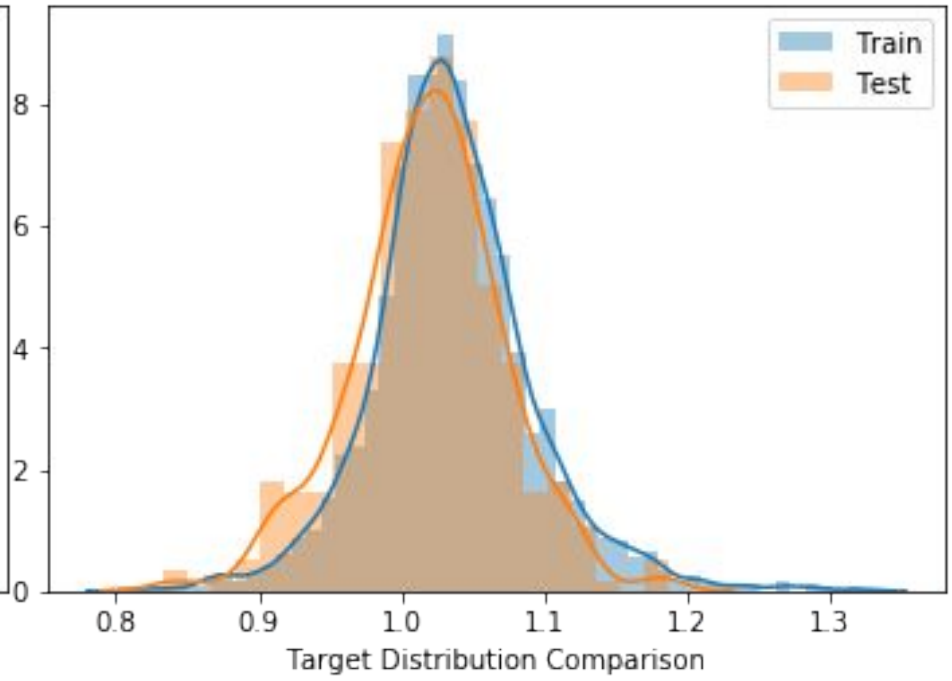
| is_open | latitude | longitude | name | neighborhood | postal_code | review_count | stars |
|---|---|---|---|---|---|---|---|
| 1 | 41.499208 | -81.536689 | McDonald's | | 02224 | 7 | 2.5 |
| 1 | 44.953815 | -73.270500 | Ransom Bay Inn B & B | | 05440 | 5 | 5.0 |
| 1 | 44.974341 | -73.300597 | Island Tree Service | | 05440 | 3 | 3.5 |
| 1 | 40.780821 | -74.150722 | Martone's Market & Café | | 05452 | 49 | 4.0 |
| 1 | 51.083200 | 11.858200 | Landgasthof Gieckau | | 06618 | 8 | 3.5 |
| 1 | 51.212600 | 11.746200 | Mühle | | 06632 | 7 | 4.5 |
| 1 | 51.217606 | 11.767746 | Rotkäppchen-Mumm Sektkellereien | | 06632 | 8 | 5.0 |
| 1 | 51.213900 | 11.733400 | Weingut Bernard Pawis | | 06632 | 7 | 5.0 |
| 1 | 51.212500 | 11.769500 | Eis-Cafe Merle | | 06632 | 4 | 3.5 |
| 1 | 51.208500 | 11.771460 | Restaurant am Unstrut- Wehr Donath Stefan | | 06632 | 8 | 2.0 |

| | postal_code | business_id | date | text |
|---|---|---|---|---|
| **3539486** | 5440 | tJRDll5yqpZwehenzE2cSg | 2012-07-26 | Unbelievable cute nicest owners ever. I got in... |
| **3539487** | 2224 | 2mroQ_qD_5kLTv88zADnXg | 2016-10-12 | The service here is TERRIBLE!! They told me t... |
| **3539488** | 2224 | 2mroQ_qD_5kLTv88zADnXg | 2016-02-01 | Every time I go through the drive through they... |
| **3539489** | 2224 | 2mroQ_qD_5kLTv88zADnXg | 2015-09-21 | Normally I consider eating at a McDonald's a d... |
| **3539490** | 2224 | 2mroQ_qD_5kLTv88zADnXg | 2015-06-05 | One of the better Mickey D's in the area. Fast... |

- Merge reviews on business id

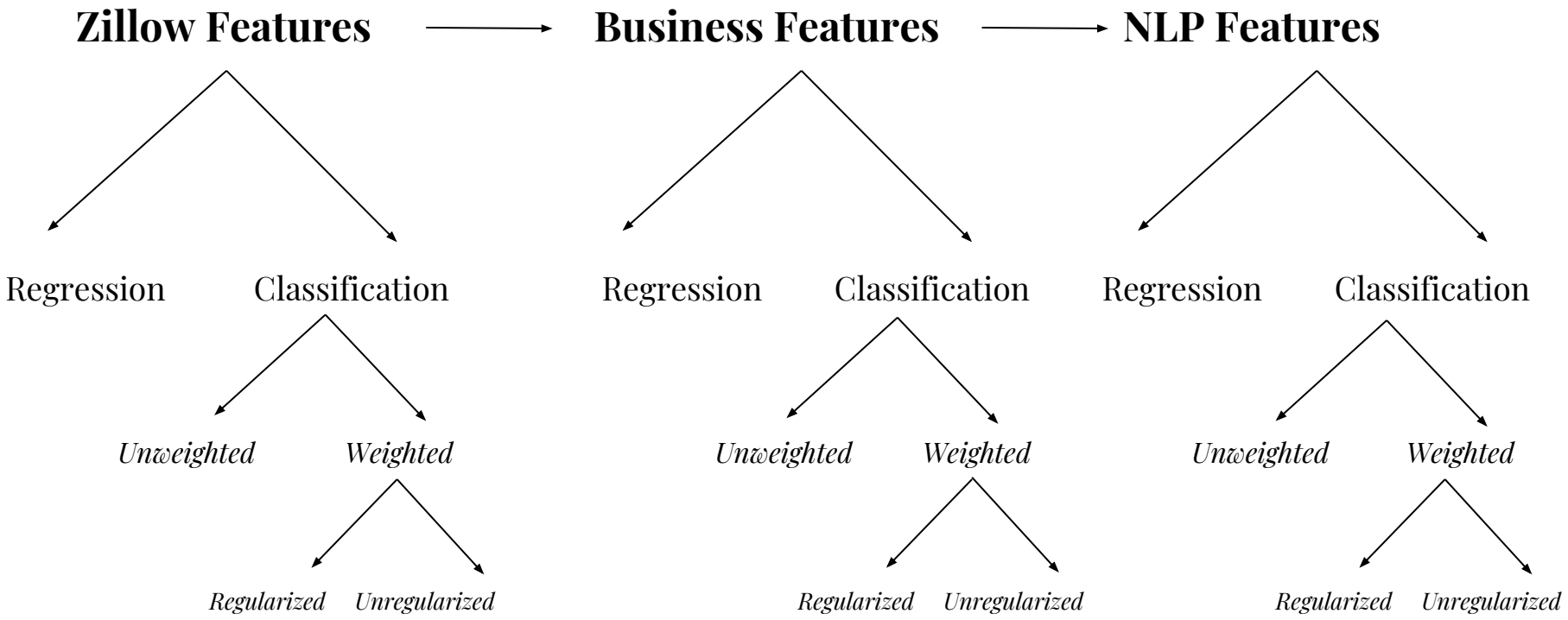
- Merge businesses by zip code


0
2
4
6
8
0.8
0.9
1.0
1.1
1.2
1.3
Train Target Distribution


Train
Test
0
2
4
6
8
0.8
0.9
1.0
1.1
1.2
1.3
Target Distribution Comparison

# Model construction

# Model Workflow



# Categories, Validation Metric, and Class Weights

**Categories**:

- **1** = **Major Loss**: >= 10% loss
- 2 = **Minor Loss:** > 3% loss but not >= 10%
- 3 = **Within Inflation**: < 3% gain and not > 3% loss
- 4 = **Minor Gain**: > 3% gain but not >= 10%
- 5 = **Major Gain**: >= 10% gain

**Weights**:

- Major Loss: **2**
- Minor Loss: **1**
- Within inflation: **0.5**
- Minor Gain: **1**
- Major Gain: **2**

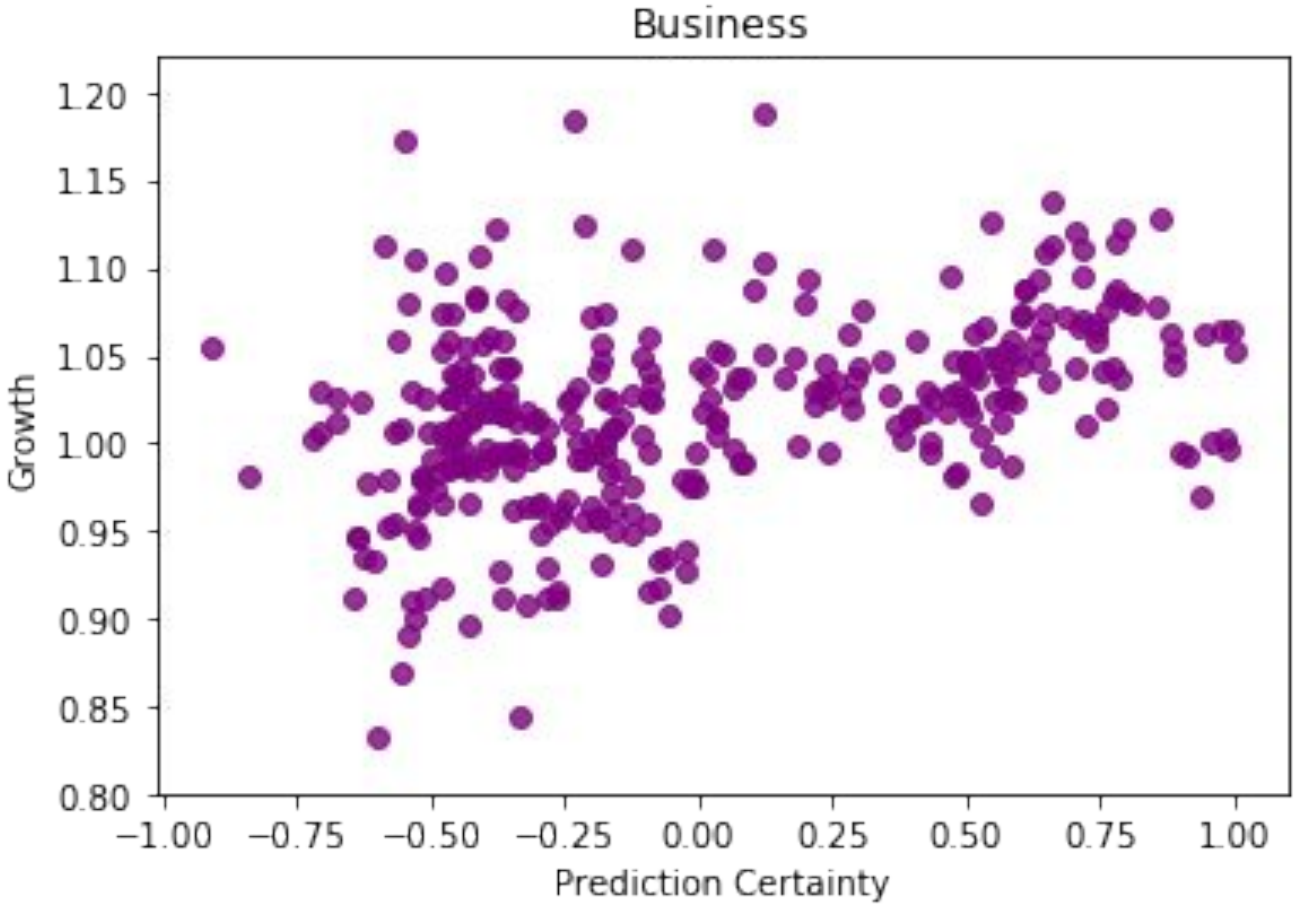
# Performance indicator

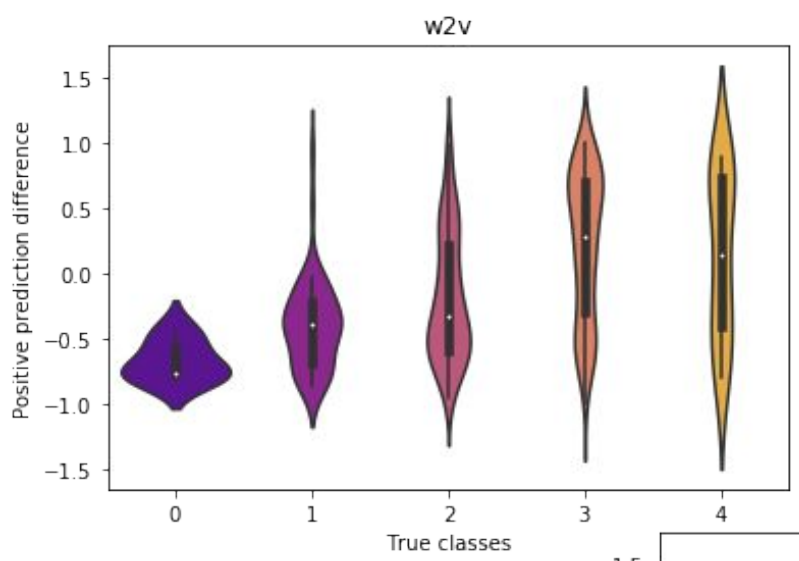
**Indicator** = *(prob(5) + prob(4))* – *(prob(1) + prob(2))*

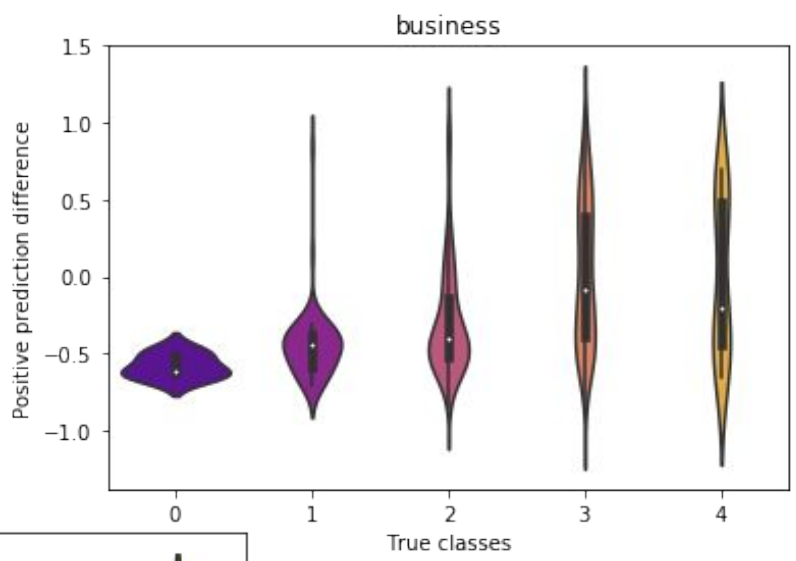
# Model performance

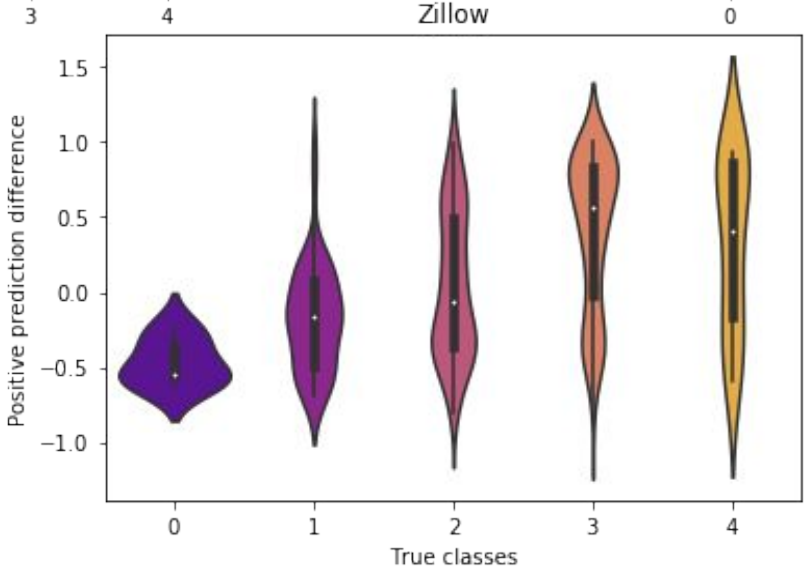
- 10% or greater
- 3% to 9.999%
- Within inflation
- Categorical cross entropy

| | Zillow | business | w2v |
|---|---|---|---|
| top_pred_prob_10% | 0.228000 | 0.228000 | 0.232000 |
| top_pred_prob_3% | 0.796000 | 0.876000 | 0.804000 |
| top_pred_prob_mean | 0.940000 | 0.972000 | 0.960000 |
| acc | 0.504532 | 0.477946 | 0.426061 |


Business
Growth
1.20
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Prediction Certainty


w2v
Positive prediction difference
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
True classes


business
Positive prediction difference
1.5
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
True classes

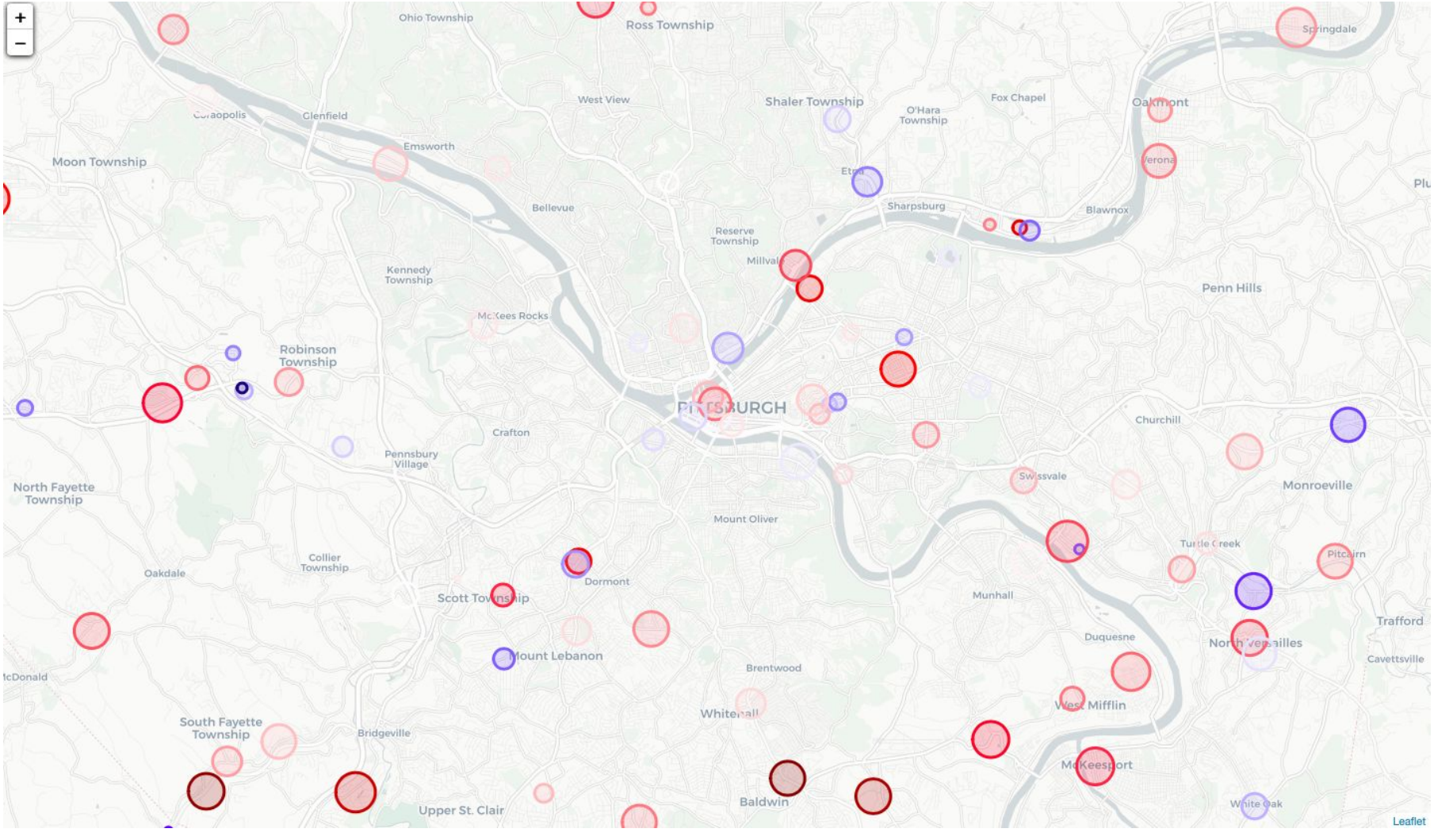

Zillow
Positive prediction difference
1.5
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
True classes

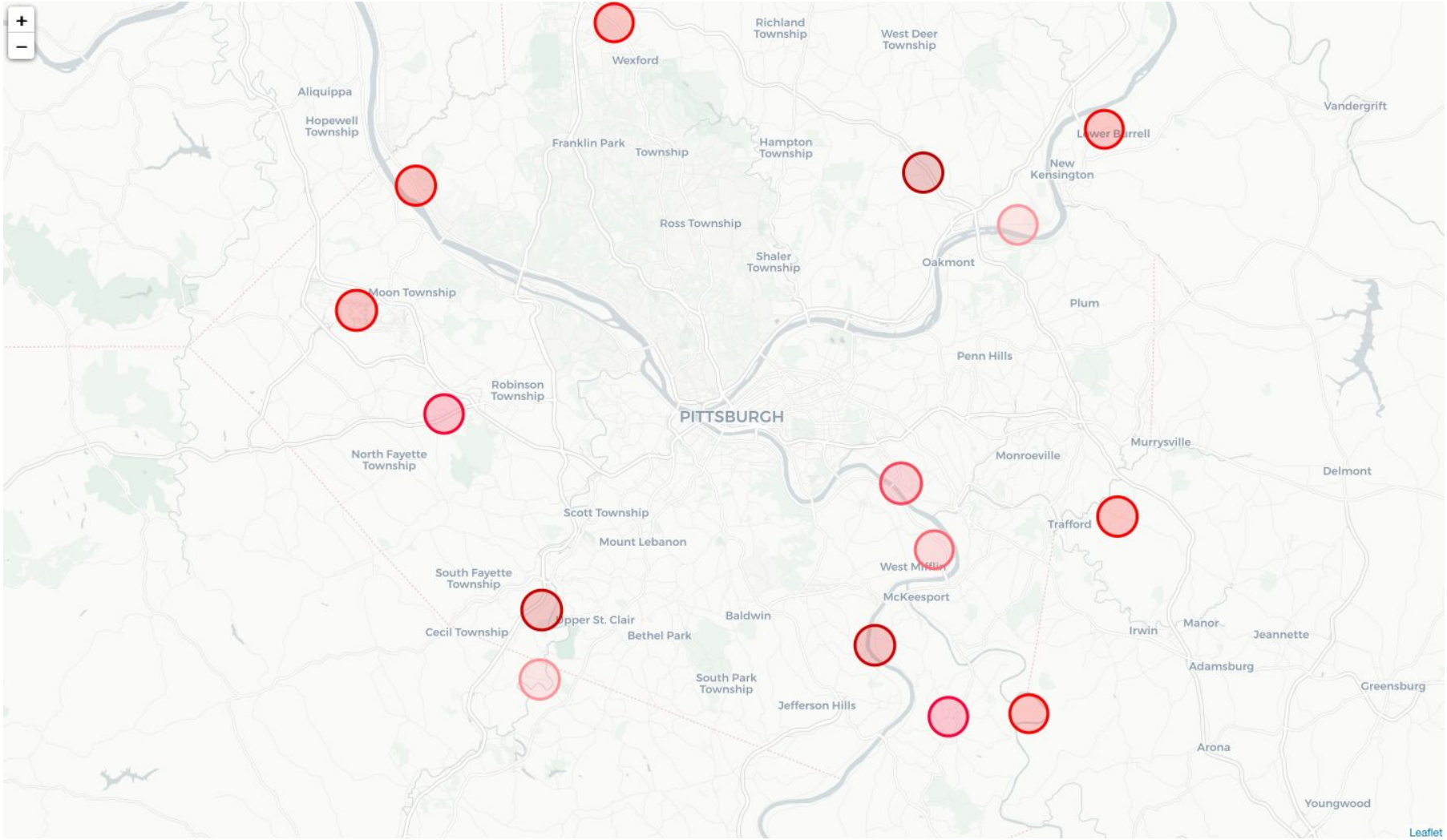
# Investment application

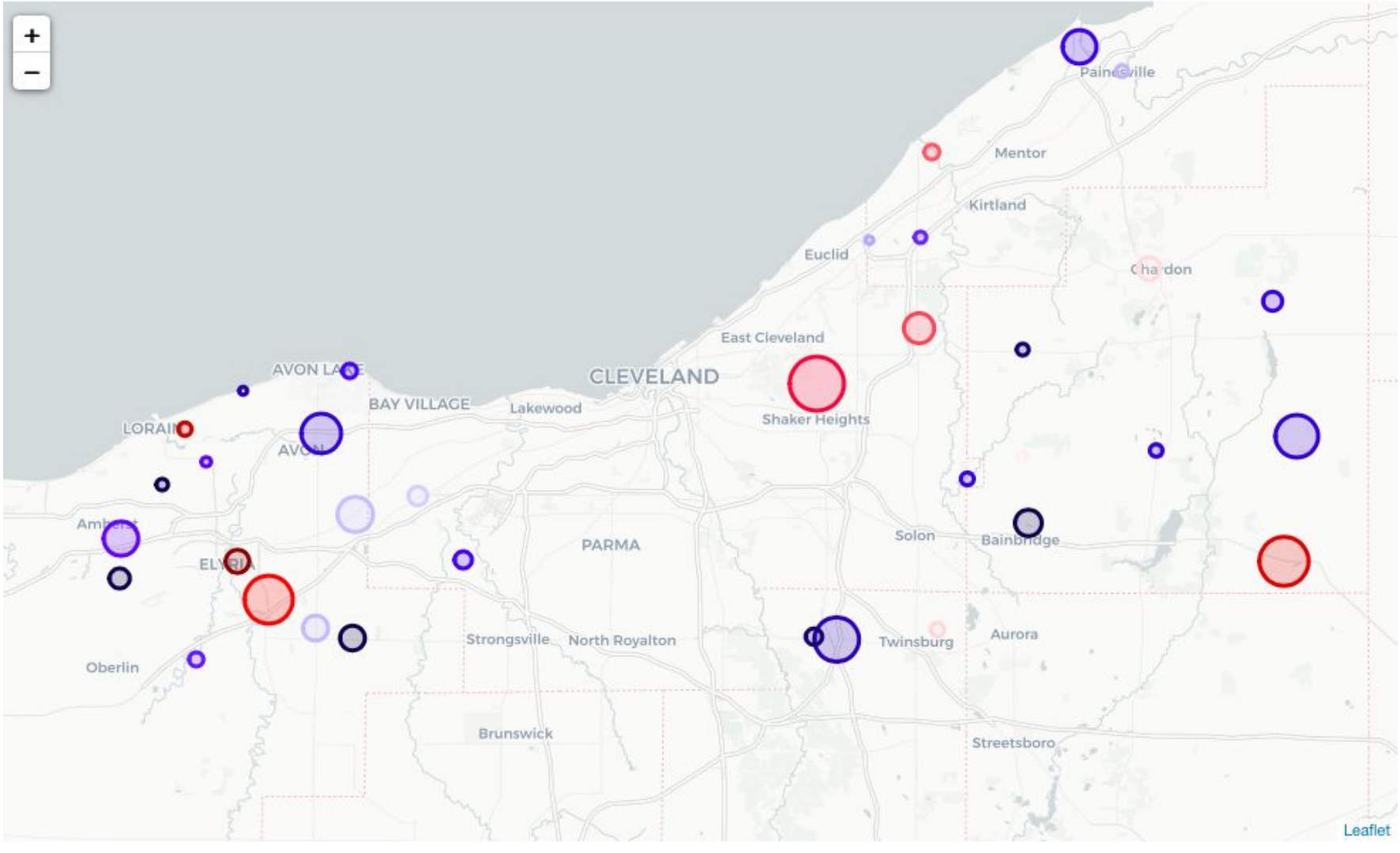
How could we roll out this product for investors?

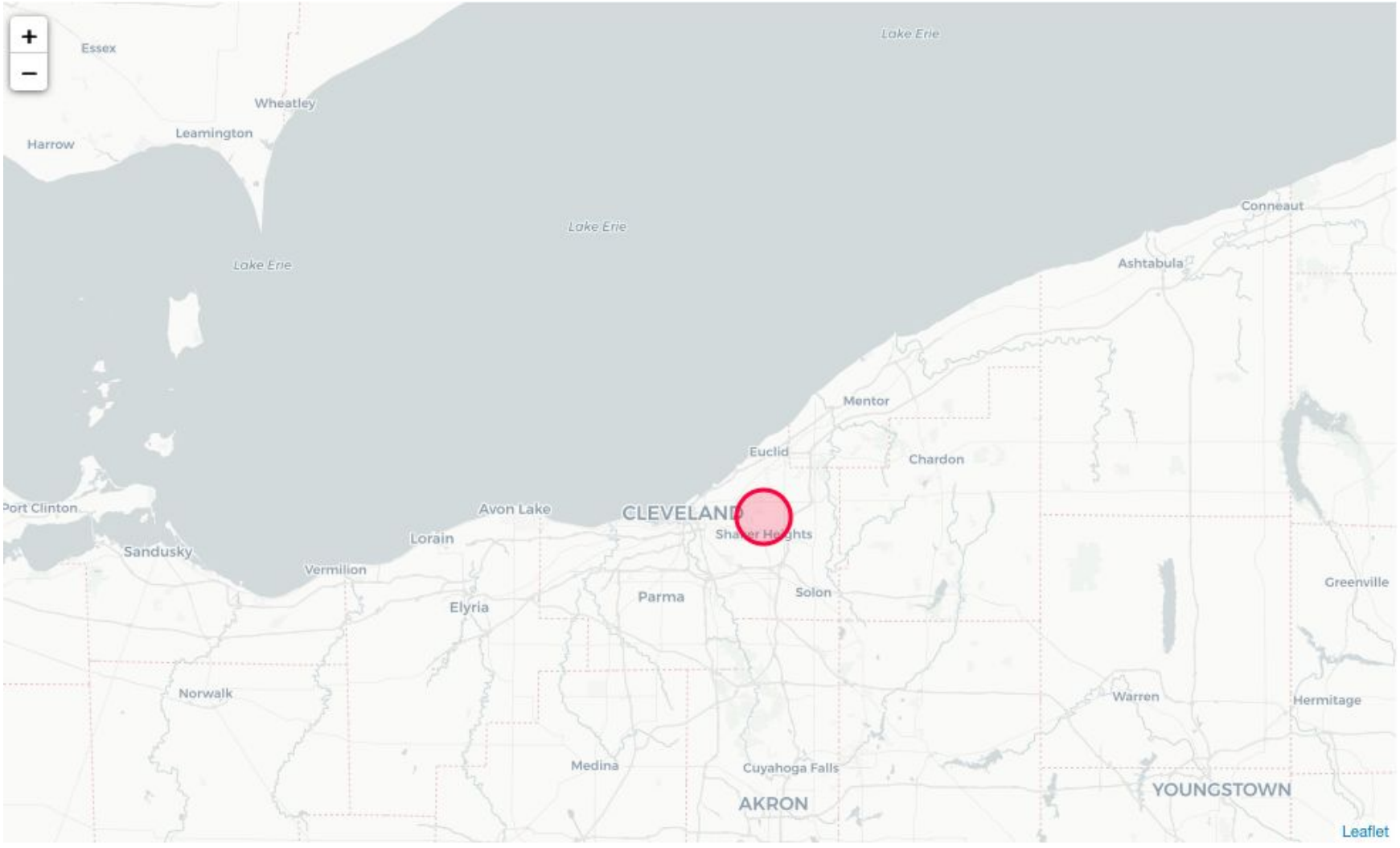
- Initial target: House flippers

- Quantify risk and reward potential for individuals who are looking into short term real estate investments.


Ohio Township
Ross Township
Springdale
West View
Shaler Township
Fox Chapel
Oakmont
O'Hara Township
Glenfield
Emsworth
Moon Township
Bellevue
Sharpsburg
Blawnox
Reserve Township
Millvale
Kennedy Township
Penn Hills
McKees Rocks
Robinson Township
PITTSBURGH
Churchill
Crafton
Pennsbury Village
Swissvale
Monroeville
North Fayette Township
Mount Oliver
Turtle Creek
Pitcairn
Collier Township
Oakdale
Dormont
Scott Township
Munhall
Trafford
Duquesne
Mount Lebanon
Brentwood
Cavettsville
West Mifflin
South Fayette Township
Bridgeville
Whitehall
McKeesport
Baldwin
Upper St. Clair
White Oak
Leaflet


Richland Township
West Deer Township
Wexford
Aliquippa
Hopewell Township
Franklin Park
Township
Hampton Township
Vandergrift
Lower Burrell
New Kensington
Ross Township
Shaler Township
Oakmont
Moon Township
Plum
Penn Hills
Robinson Township
PITTSBURGH
North Fayette Township
Monroeville
Murrysville
Delmont
Scott Township
Trafford
Mount Lebanon
West Mifflin
South Fayette Township
McKeesport
Upper St. Clair
Baldwin
Cecil Township
Bethel Park
Irwin
Manor
Jeannette
Adamsburg
South Park Township
Greensburg
Jefferson Hills
Arona
Youngwood
Leaflet


Mentor
Kirtland
Euclid
East Cleveland
CLEVELAND
Shaker Heights
AVON LAKE
BAY VILLAGE
Lakewood
LORAIN
AVON
PARMA
Solon
Bainbridge
Strongsville
North Royalton
Twinsburg
Aurora
Oberlin
Brunswick
Streetsboro
Leaflet


Lake Erie
Essex
Wheatley
Harrow
Leamington
Lake Erie
Lake Erie
Conneaut
Ashtabula
Mentor
Euclid
Chardon
Port Clinton
Avon Lake
CLEVELAND
Lorain
Sandusky
Vermilion
Elyria
Parma
Solon
Greenville
Norwalk
Warren
Hermitage
Medina
Cuyahoga Falls
AKRON
YOUNGSTOWN
Leaflet

# Next Steps

Ways that we can improve this product

- Continue to work with NLP
- Create models that represent each month of the year
- Engineer features further
- Apply convolutional layers

# Conclusion