Understanding Urban Depopulation:

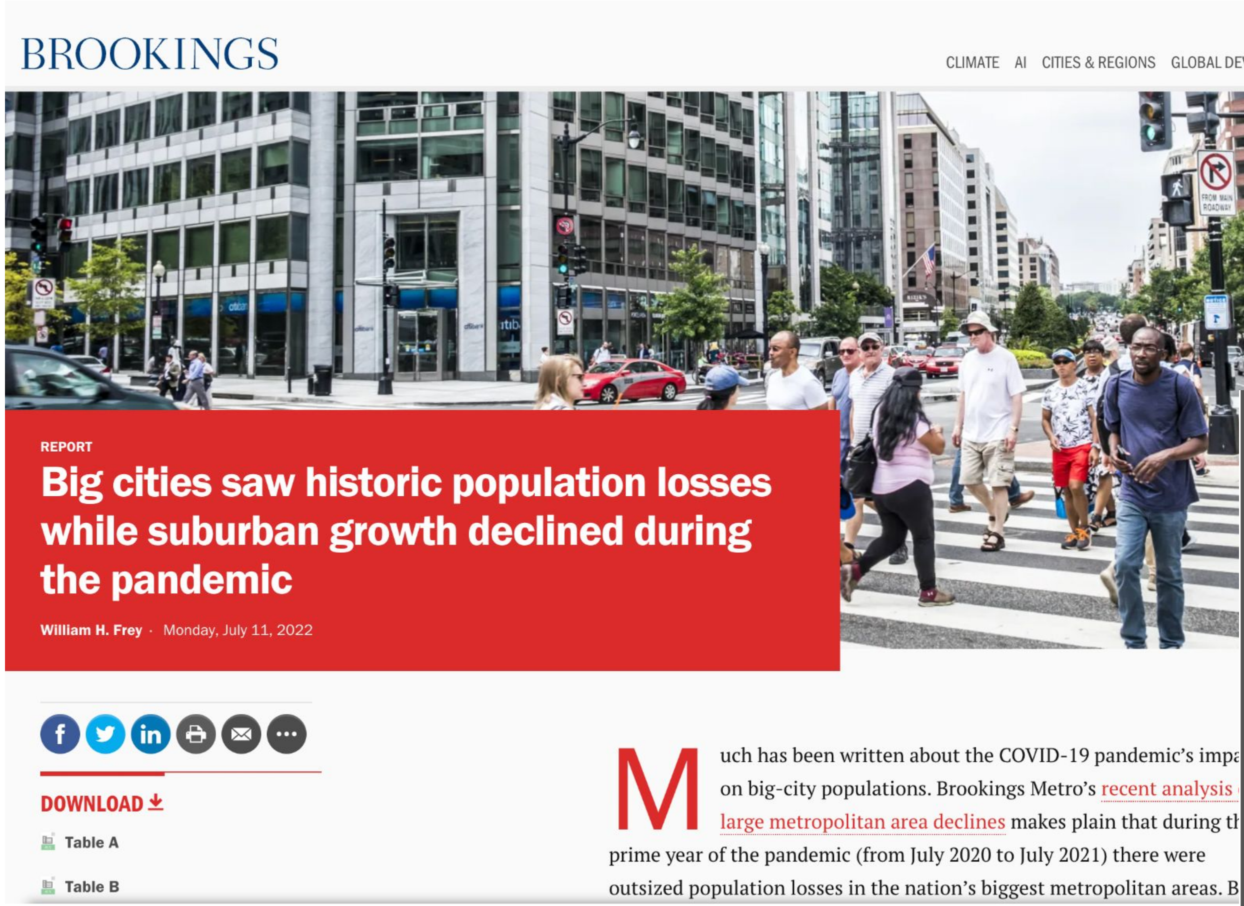
An analysis of shrinking cities in the US with machine learning approaches

Colton Lapp | Chi-Shiun Tsai

Introduction

Introduction

- Motivation:
  - Population growth important for city health
  - Covid 19 caused migration patterns to change: many big cities experienced population loss


BROOKINGS

CLIMATE AI CITIES & REGIONS GLOBAL DEV INTL AFFAIRS U.S. ECONOMY U.S. POLITICS & GOVT

REPORT

Big cities saw historic population losses while suburban growth declined during the pandemic

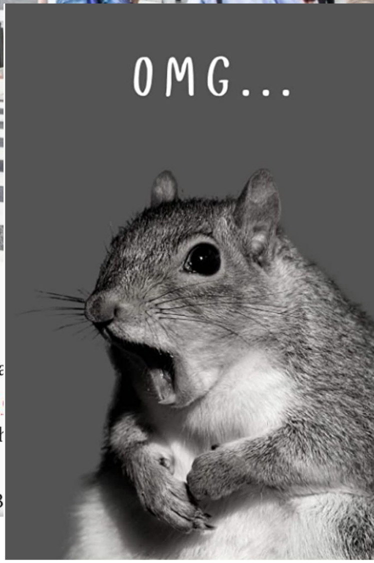
William H. Frey · Monday, July 11, 2022

DOWNLOAD

Table A

Table B

Much has been written about the COVID-19 pandemic's impa
on big-city populations. Brookings Metro's recent analysis
large metropolitan area declines makes plain that during th
prime year of the pandemic (from July 2020 to July 2021) there were
outsized population losses in the nation's biggest metropolitan areas. B



Introduction

- Questions:
  - Can we predict which cities will experience population growth/losses?
    - Help city leaders focus on key drivers of population change.
  - Can we cluster cities to help detect similarity beyond population size and geographical closeness?
    - Useful for people looking for areas to move or policy makers trying to connect with leaders from other cities.

Introduction

- Data Sources:
  - American Community Survey (ACS)
    - 40 variables - Income, Race, Transportation patterns, Home values, etc
  - Covid
    - NYTimes data for cases/deaths
  - FBI Crime data
    - Cleaned, but sadly too much missing data to use as input feature

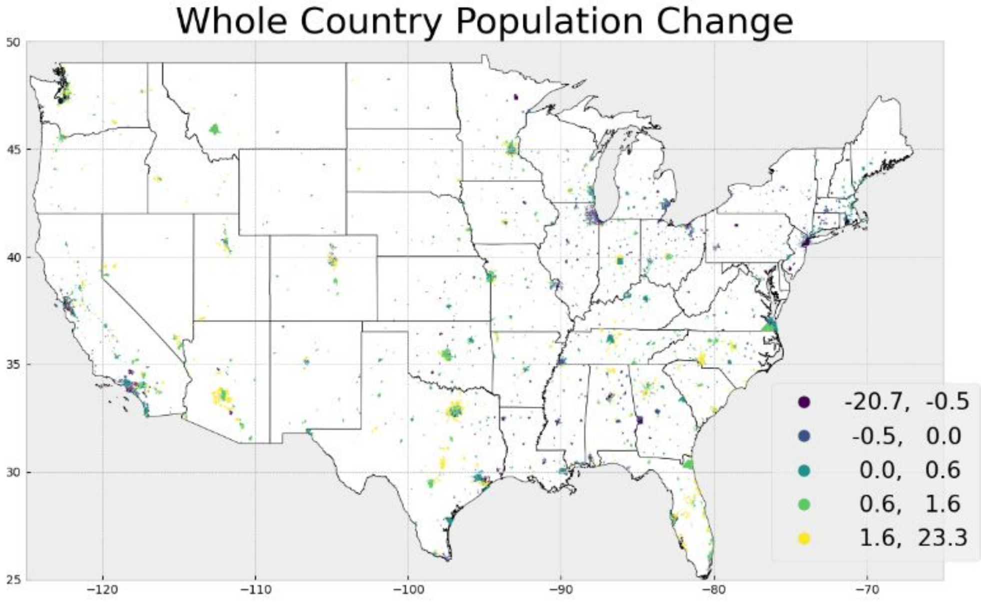
EDA

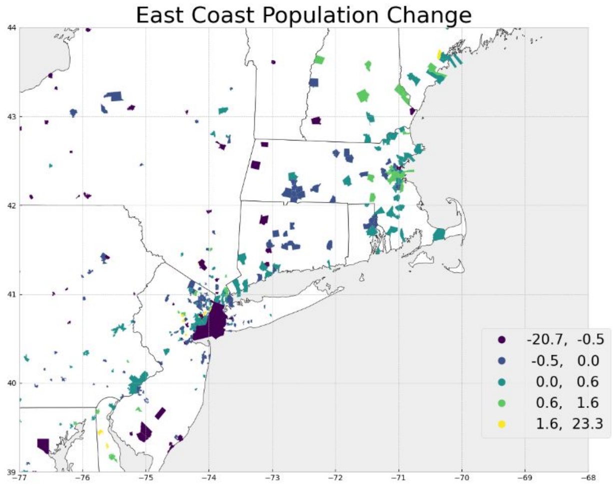
EDA: Data Coverage

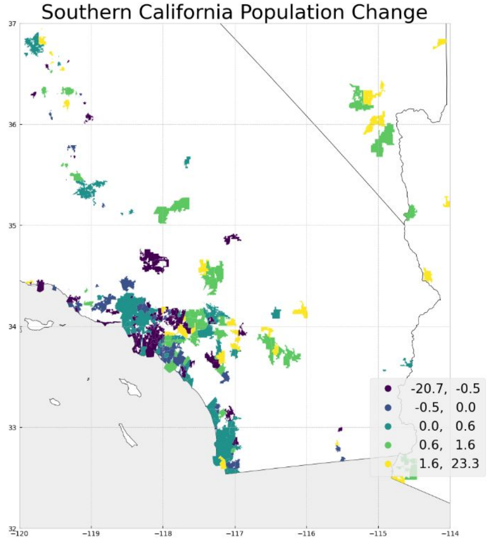
```
How many cities fall into different population levels?:
-----------------------------------------------
Number of cities w/ >100K Pop:  308   ... 1.848%
Number of cities w/ >50K Pop:   775   ... 4.65%
Number of cities w/ >20K Pop:   1516  ... 9.097%
Number of cities w/ >10K Pop:   3092  ... 18.554%
Number of cities w/ >5K Pop:    4745  ... 28.473%
Number of cities w/ >1K Pop:    10427 ... 62.568%
-----------------------------------------------
```

Data covers over 16.6K “cities” in the US

Many are very small towns/”places” etc





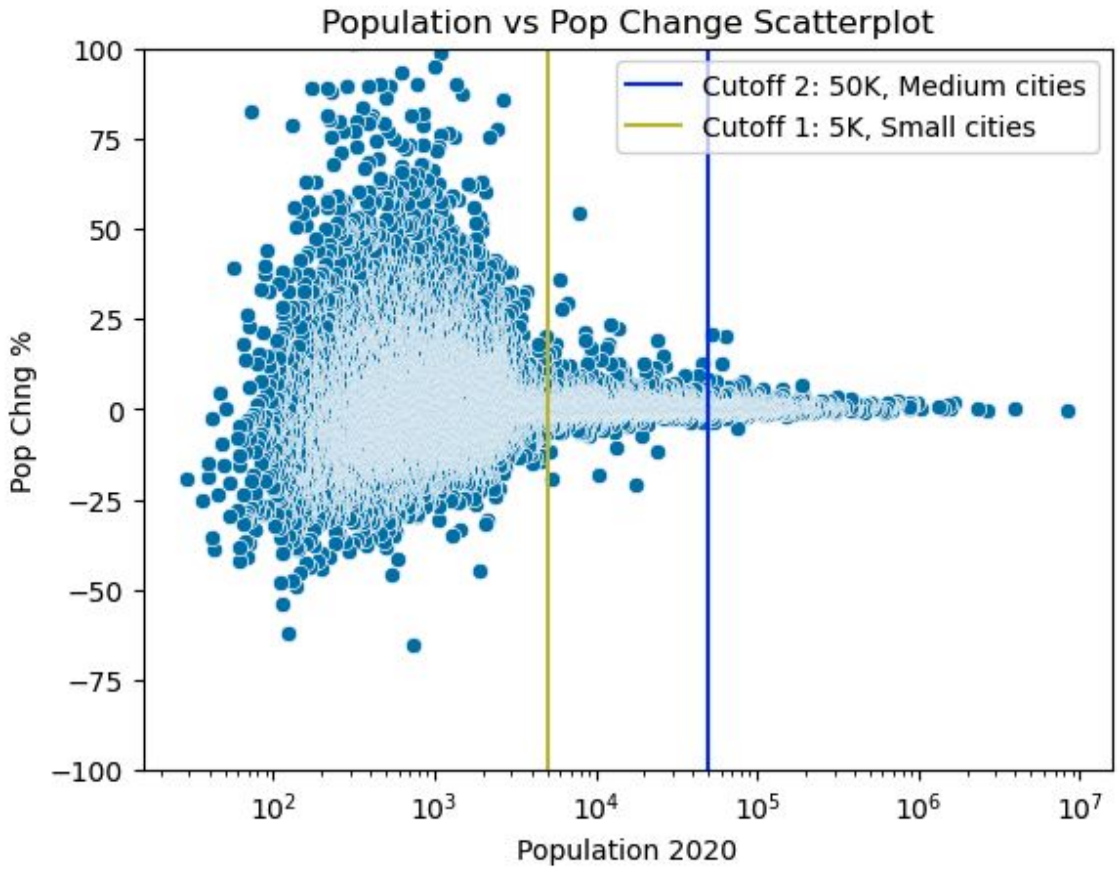


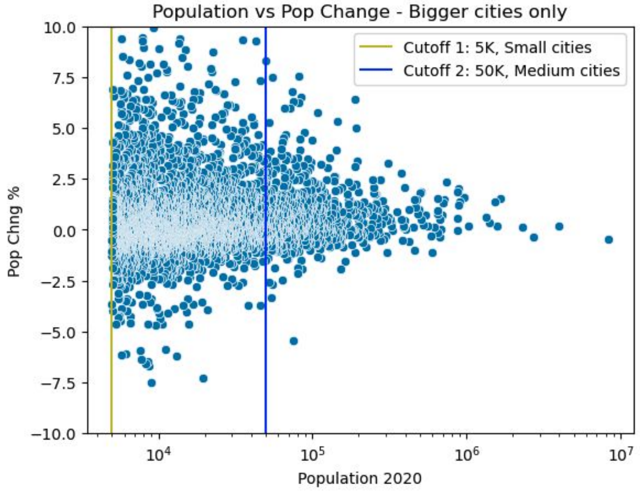
EDA: Population Growth Rates

How to classify growth rates?

Different "cutoff" levels for
1. Shrinking
2. Neutral
3. Growing

Depending on city size
Small: +/- 3%
Medium: +/-0.75%
Big: +/- 0.25%





Q1: Classification

Shrinking / Neutral / Growing

Methodology

Classifying Cities by Population Change:

- **Feature classes?**
  - *Binary*: Shrinking/Growing
  - *Multiclass*: Shrinking/Neutral/Growing
- **Feature transformation:**
  - Binary classification/Multinomial
  - Log skewed variables, scale features
- **Separate Models for small/medium/big cities**
  - Capture different causal mechanisms

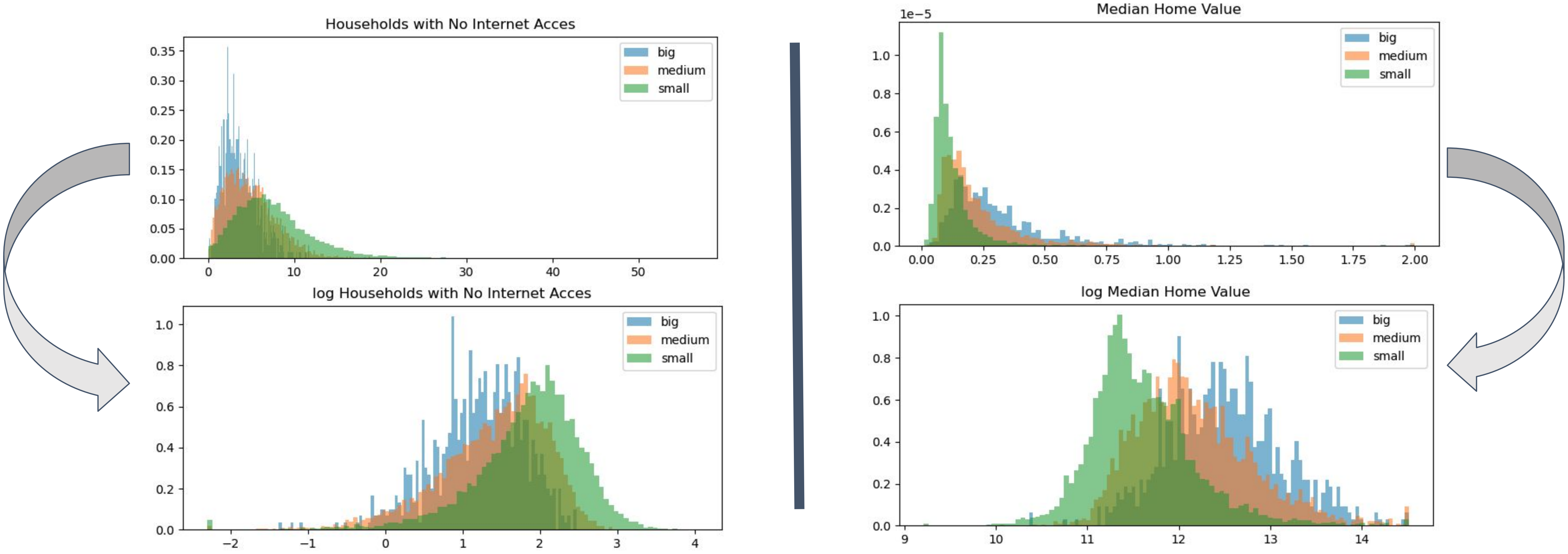
Models Tried:

1. Logistic Regression
2. Naive Bayes
3. KNN
4. **SVM** (new)
5. RF

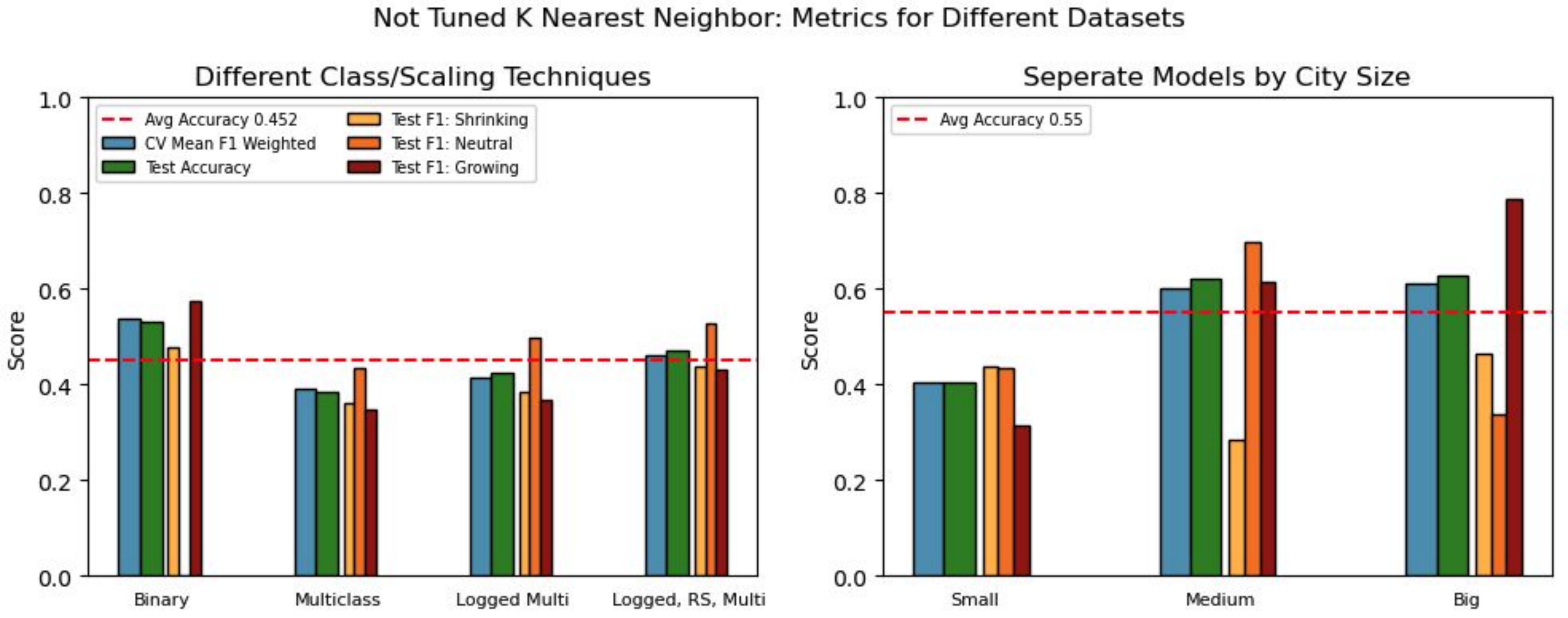
Tuned with GridSearchCV

Methodology

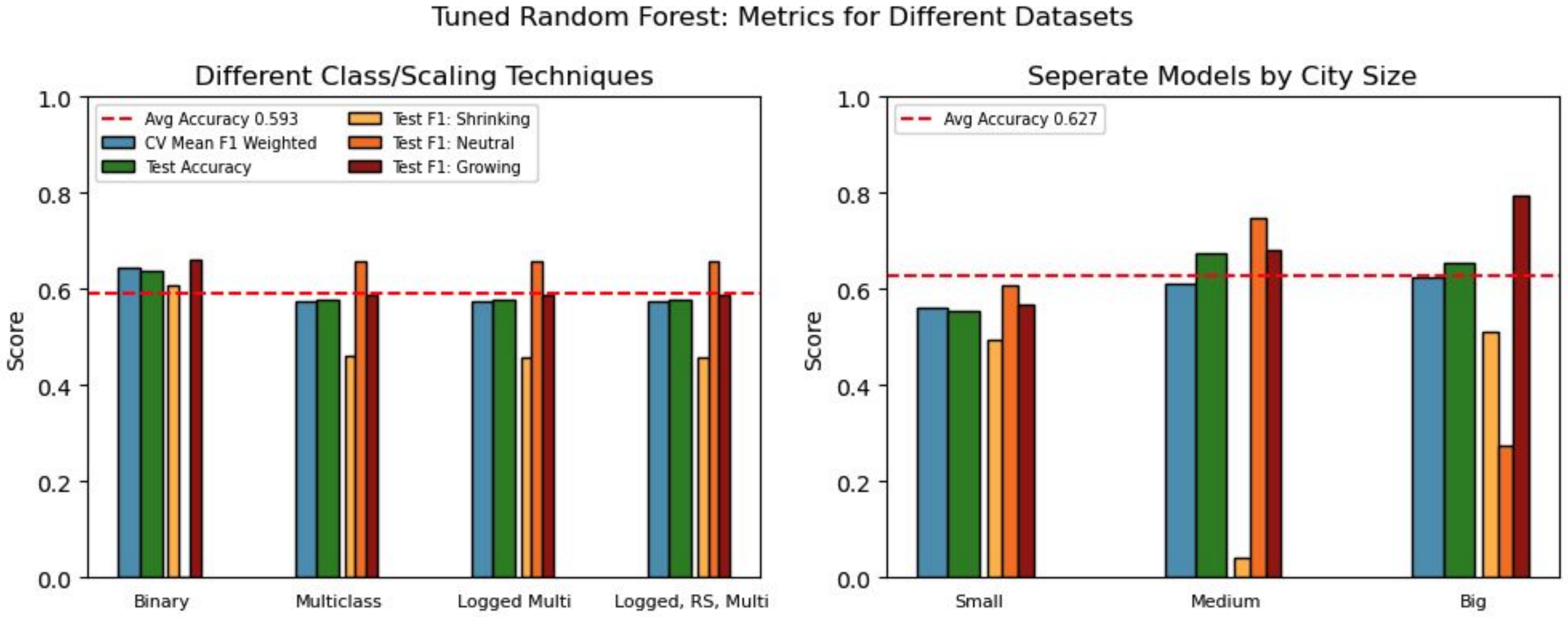
Dealing with skewed variables with Logs



Results: worst model



Results: best model



Results: All Validation Results

Weighted F1 Score for Test Dataset

	Binary Classification	MultiClass, Logged, Scaled	Small Cities	Medium Cities	Big Cities
RF	0.636000	0.576000	0.556000	0.621000	0.616000
SVM	0.365000	0.542000	0.510000	0.636000	0.620000
LR	0.588000	0.504000	0.513000	0.629000	0.640000
KNN	0.531000	0.495000	0.447000	0.626000	0.623000
GNB	0.365000	0.463000	0.472000	0.580000	0.606000

Q2. Clustering

Can we cluster cities to help detect similarity beyond population size and geographical closeness?

Methodology

- Using the same logged data for Question 1
- Robust Scaler before clustering
- Separate models for small/medium/big cities

Methodology

- **Models**
  - K-means clustering
    - find number of k with elbow method
  - Hierarchical clustering
    - Height-based cut
  - DBSCAN
    - Find epsilon with elbow method
  - Gaussian Mixture Model
    - Using BIC to find optimal number of clusters

- Use <u>**silhouette score**</u> to evaluate the final model performance

Model performance

K-Means

	Models	Number of clusters	Size of each cluster	Silhouette Score
0	for small cities	4	[5940 660 1325 3995]	0.094727
1	for medium cities	3	[2116 449 1405]	0.194146
2	for big cities	3	[319 175 281]	0.203877

Hierarchical clustering

	Models	Number of clusters	Size of each cluster	Silhouette Score
0	for small cities	5	[4558 4761 356 830 1415]	0.048559
1	for medium cities	3	[481 1645 1844]	0.155736
2	for big cities	4	[268 194 222 91]	0.138293

Model performance

DBSCAN

DBSCAN is not finding any clusters.

GMM

	Models	Number of clusters	Size of each cluster	Silhouette Score
0	for small cities	9	[4569 282 133 560 3615 480 738 947 596]	-0.054288
1	for medium cities	4	[245 1377 884 1464]	0.155736
2	for big cities	1	[775]	not enough clusters

Model performance

DBSCAN

DBSCAN is not finding any clusters.

GMM

	Models	Number of clusters	Size of each cluster	Silhouette Score
0	for small cities	9	[4569 282 133 560 3615 480 738 947 596]	-0.054288
1	for medium cities	4	[245 1377 884 1464]	0.155736
2	for big cities	1	[775]	not enough clusters

Results

Based on the silhouette score, the best model is K-means.

	Models for	K-means	Hierarchical clustering	GMM
0	small cities	0.094727	0.048559	-0.054288
1	medium cities	0.194146	0.155736	0.071209
2	big cities	0.203877	0.138293	not enough clusters

Results

Cities in each cluster based on k-means clustering:

Big cities:

- Cluster 1: Cape Coral city, Florida/ Bossier City city, Louisiana/ Waukesha city, Wisconsin
- Cluster 2: Las Cruces city, New Mexico / Lodi city, California/ Harlingen city, Texas
- Cluster 3: Round Rock city, Texas/ Redwood City city, California/ Milpitas city, California

How about Pittsburgh?

New York city, New York / Chicago city, Illinois / Philadelphia city, Pennsylvania / Jacksonville city, Florida/ Columbus city, Ohio

Conclusions

Conclusions:

- Classification:
  - Hard to predict city growth but our data was much better than naively guessing
  - Moderate difference in accuracy across models
  - Tuning and data transformations can be very important depending on the model
- Clustering:
  - Challenging to cluster cities based on other variables besides population size and geographical closeness but it still gives us a sense of the similarity between cities
  - K-means clustering dominates across models in terms of silhouette score

Future Work

Future Work:

- Modeling choices:
  - Look at population growth over longer time horizon
  - Interaction terms between variables to capture non-linearity
  - Weighting feature importance for clustering algorithms
- Additional analysis:
  - More work understanding feature importance for classification
- Other outcome variables:
  - Crime outcomes, such as crime rate
  - Political outcomes, such as voting patterns

References

Büchler, S., Niu, D., & Kinsella Thompson, A. (2021). Predicting urban growth with machine learning. *MIT Center for Real Estate Research Paper*, (21/06).

Jato-Espino, D., & Mayor-Vitoria, F. (2023). A statistical and machine learning methodology to model rural depopulation risk and explore its attenuation through agricultural land use management. *Applied Geography*, *152*, 102870.

Johnson, K., & Lichter, D. (2019). Rural depopulation in a rapidly urbanizing America.

Kim, Y., Safikhani, A., & Tepe, E. (2022). Machine learning application to spatio-temporal modeling of urban growth. *Computers, Environment and Urban Systems*, *94*, 101801.

Shafizadeh-Moghadam, H., Asghari, A., Tayyebi, A., & Taleai, M. (2017). Coupling machine learning, tree-based and statistical models with cellular automata to simulate urban growth. *Computers, Environment and Urban Systems*, *64*, 297-308.

Rahmah, N., & Sitanggang, I. S. (2016). Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra. In *IOP conference series: earth and environmental science* (Vol. 31, No. 1, p. 012012). IoP Publishing.

Thank You