Predicting COVID-19 Risk in Your Community

By: Michael Pozo

Agenda

01 | **Business Problem**

Why run this model?

02 | **Data**

Center for Disease Control (CDC) Studies

03 | **Models**

Iteration Process to Reach Final Model

04 | **Insights**

Model Findings

01
Business Problem

BACKGROUND

Who: New York Department of Health (NYDH)

What: Predicting community COVID-19 (COVID) levels using socioeconomic and demographic population estimates from the the CDC
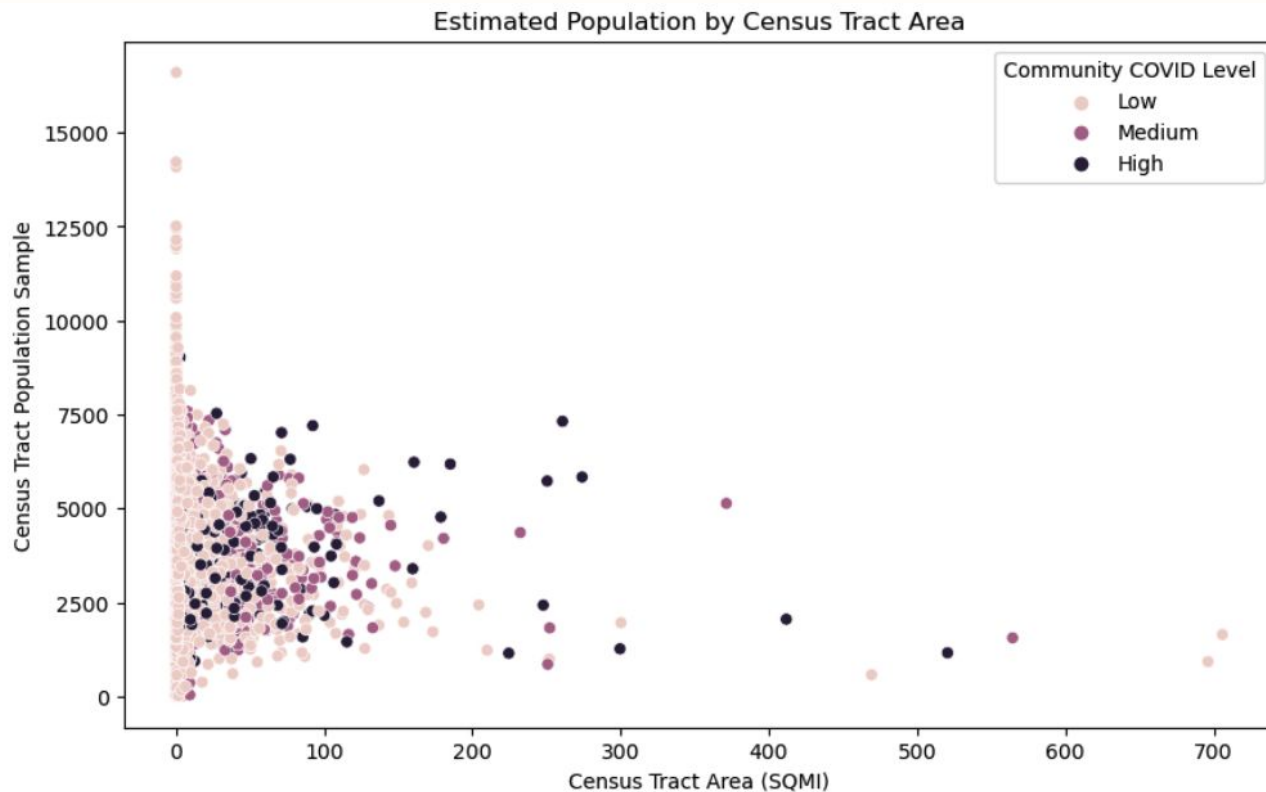
Why: Help the NYDH better plan resource allocations for future emergencies **before** they happen and prevent loss of life/economic damage

How: Train machine learning models using CDC data to identify the key features in determining COVID levels
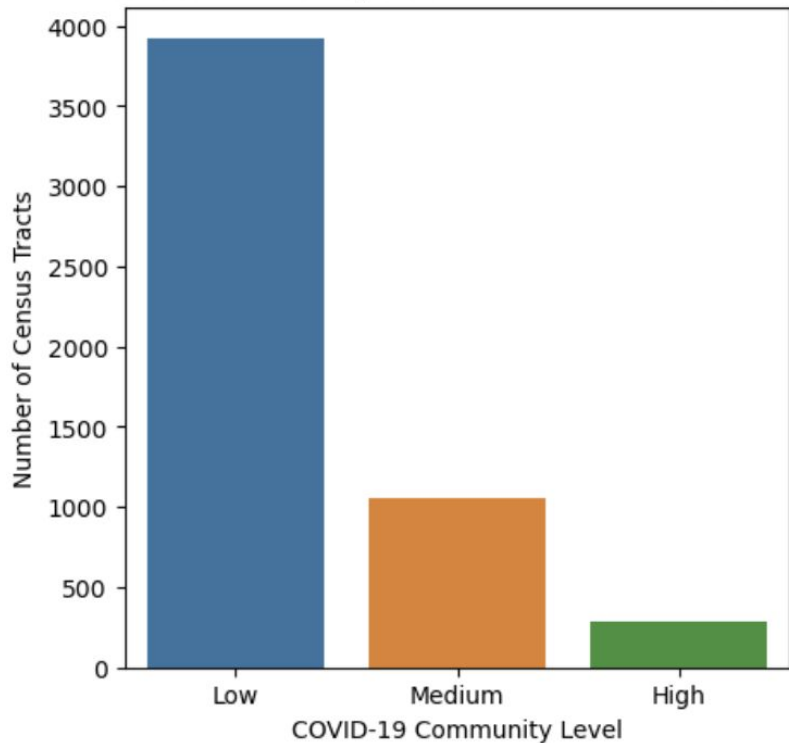
02
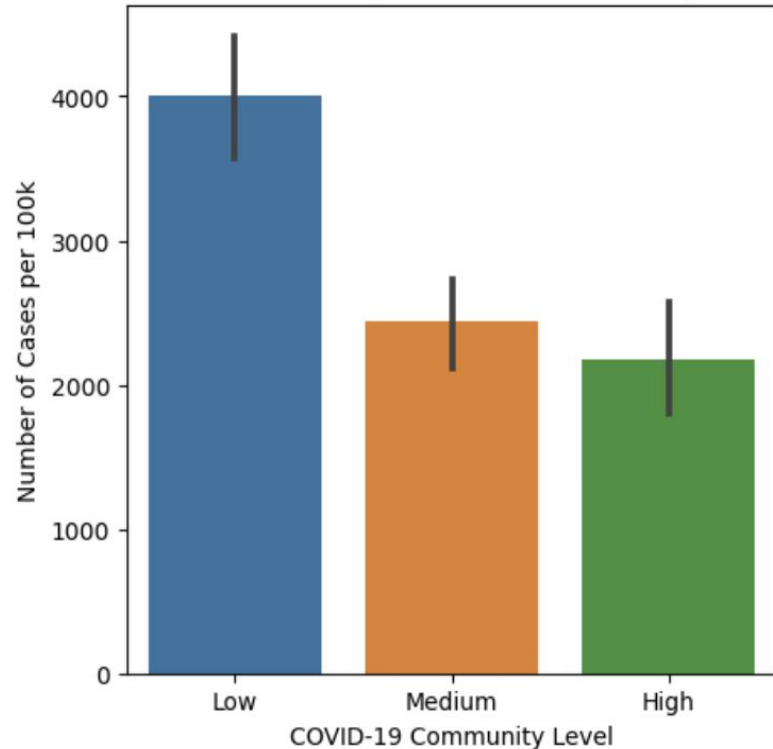## The Data

Population Distribution in New York

Estimated Population by Census Tract Area
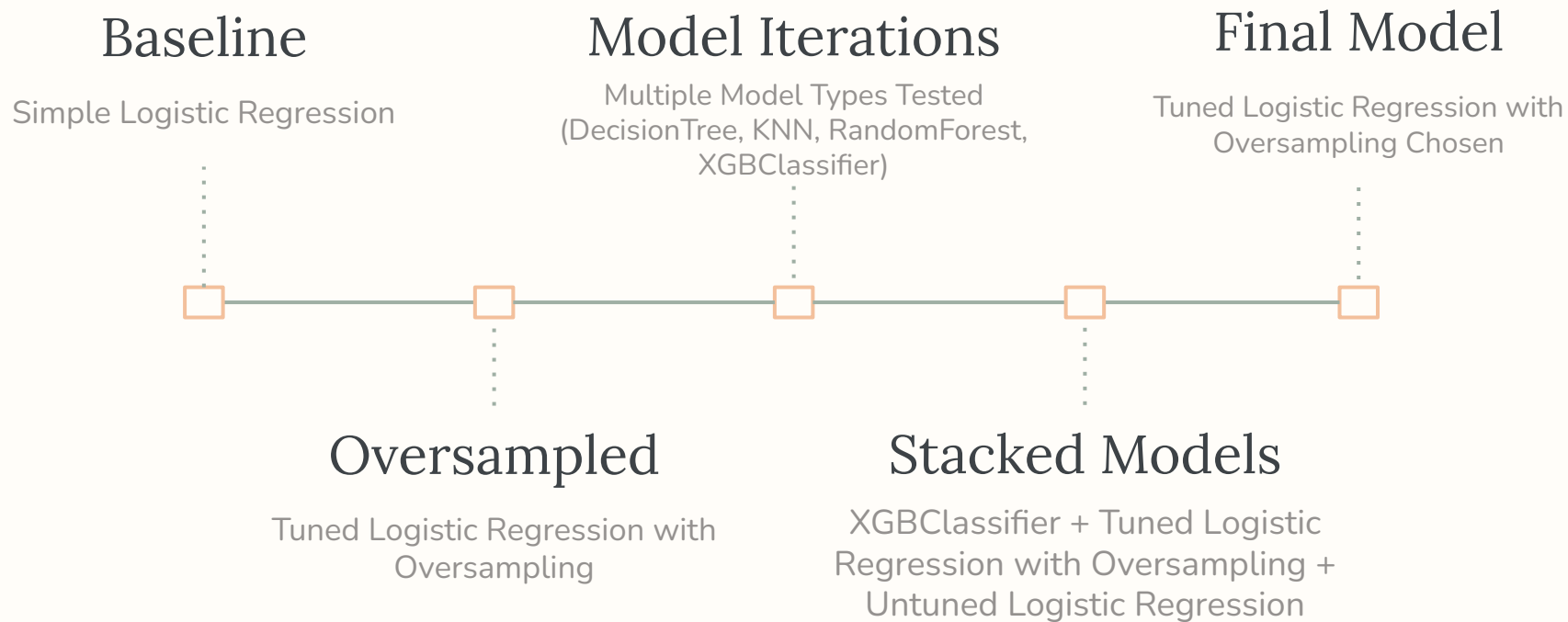
COVID Community Levels in New York

Why This Matters

- High COVID level areas make up only 5% of New York State census tracts, but almost a third of COVID-19 cases

- Classifying high risk areas in the early days of a pandemic can lead to proper resource management and strategic response planning

- To prevent the loss of human life, correctly identifying the most high risk areas is the most important criteria for this model. This is will increase the amount of false positive detection, but this is less of a concern
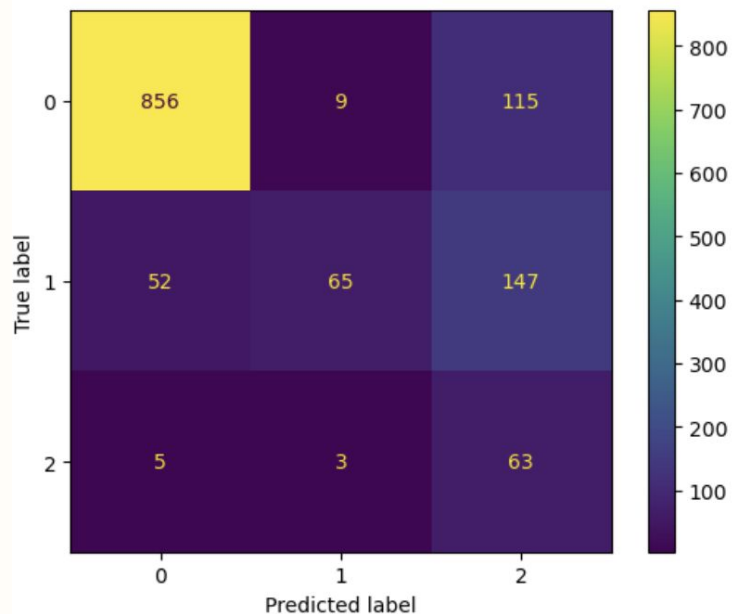
03
Model Progression

Model Iteration

Baseline

Simple Logistic Regression

Model Iterations

Multiple Model Types Tested
(DecisionTree, KNN, RandomForest,
XGBClassifier)

Final Model

Tuned Logistic Regression with
Oversampling Chosen

Oversampled

Tuned Logistic Regression with
Oversampling

Stacked Models

XGBClassifier + Tuned Logistic
Regression with Oversampling +
Untuned Logistic Regression
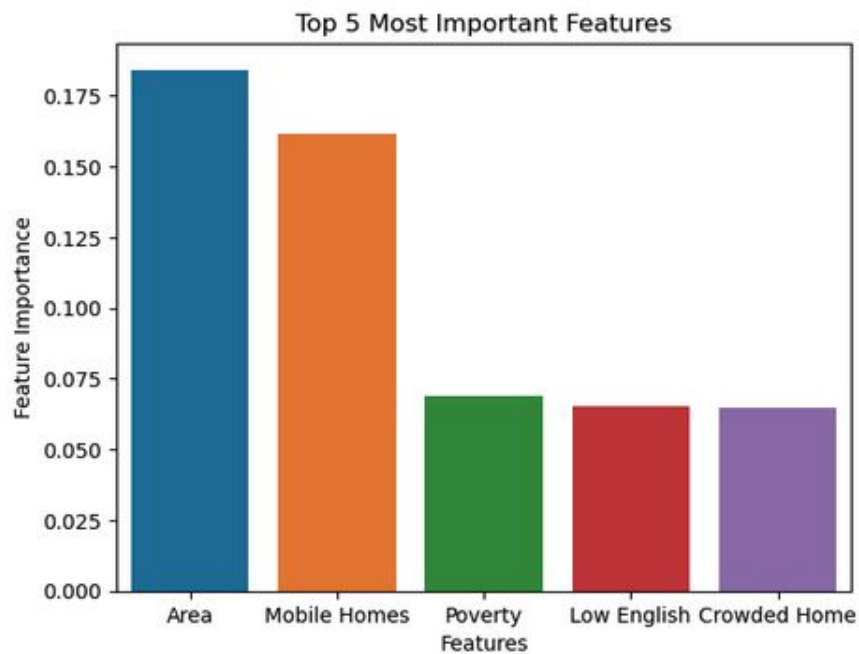
Final Model

- Legend:
  - 0 = Low COVID
  - 1 = Medium COVID
  - 2 = High COVID

- Final model **75%** accurate in predicting community COVID levels

- Model correctly identifies **89%** of all High risk census tracts

- We find the high number of false positives for the High risk class to be acceptable

04
Insights

Feature Importance


Top 5 Most Important Features

- Features like square mileage and crowded homes make sense in the context of a viral pandemic

- Less predictable:
  - Population living below poverty line
  - Population living in mobile homes
  - Population of limited English speakers

- Many of the less predictable features could be generalizable to other types of emergencies

Poverty and COVID

Distribution of Estimated Population Living 150% Below Poverty Level
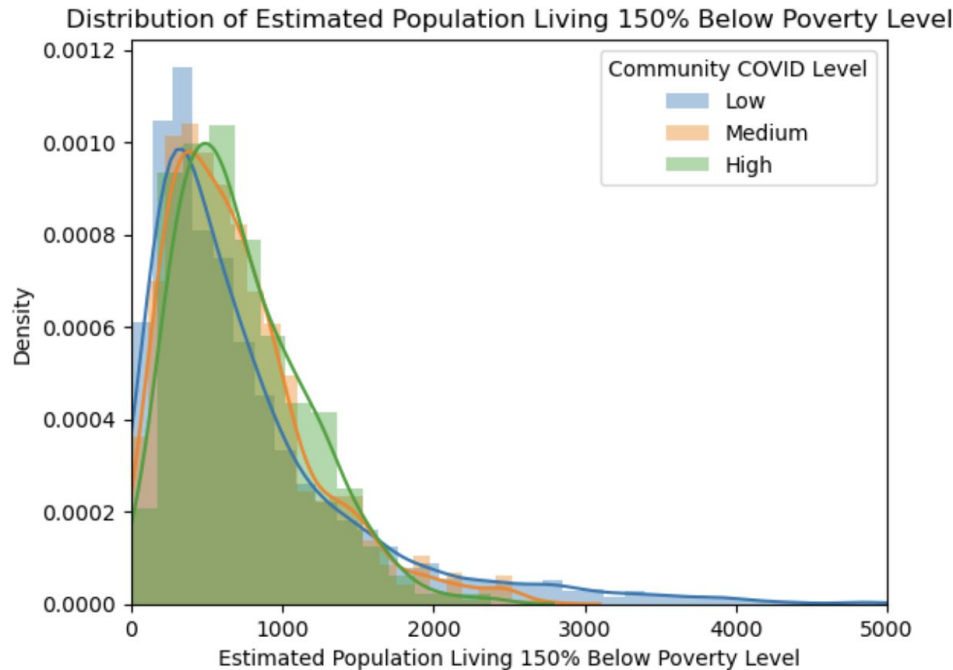
- Community COVID levels showed different distributions of populations living below the poverty line

- Statistical testing (ANOVA, T-test) showed that High COVID Level areas have meaningfully different levels of poverty at the 95% confidence level

- Suggests that poverty levels could be an identifying factor of High COVID level areas

Other Potential Flags - Mobile Homes
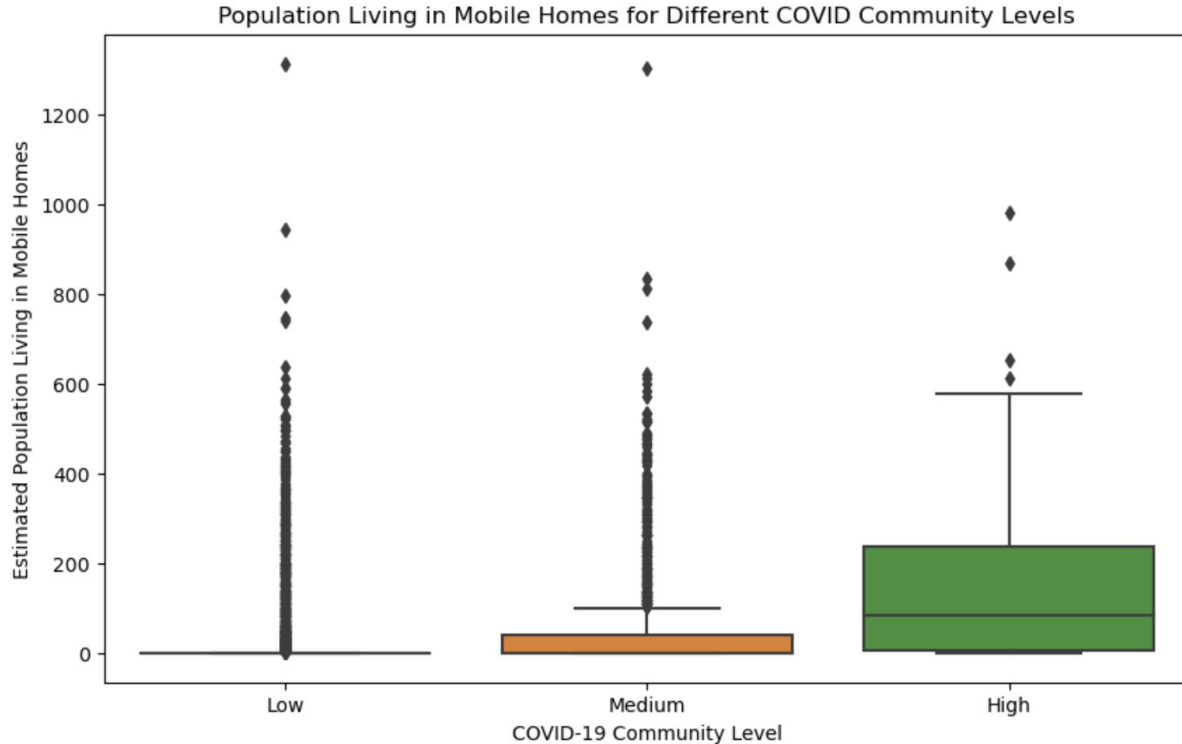
Recommendations

1. Use machine learning models to identify potential High COVID level areas and plan ahead

2. Monitor flags like poverty levels and populations living in mobile homes

3. Continuously update models using the most up to date Census data

Next Steps

1. Bring in additional healthcare information to make more connections

2. Try other targets to see if the model generalizes to other emergency events

Questions?