COVID-19 US Nursing Home Forecasting and Visualizing

Team 88: Yilin Cao Xinrui Yuan Runbai Wan Yumei Wang He Sun
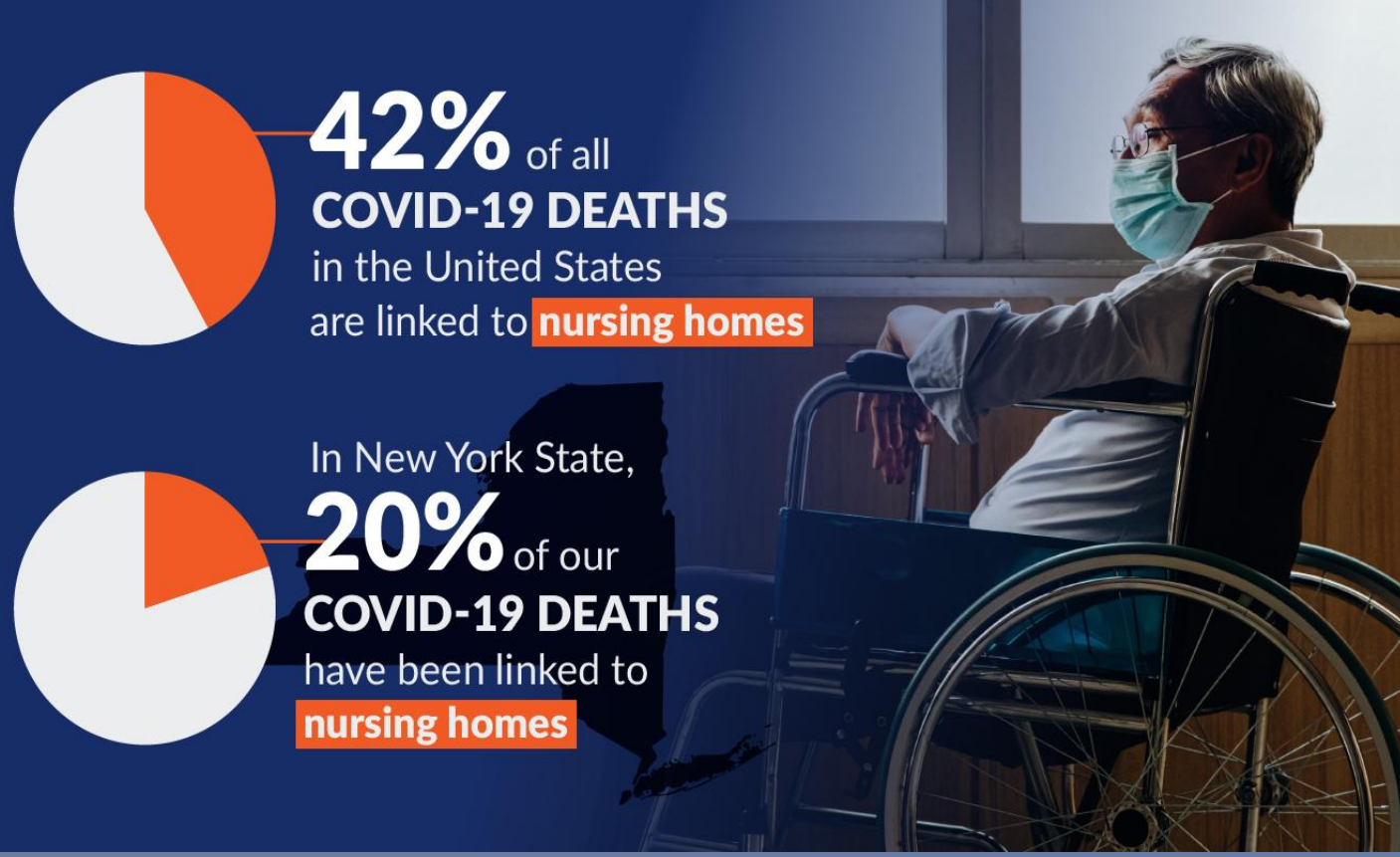
Motivation/Introductions

What is the problem:
- The COVID-19 pandemic has wreaked havoc across the globe. Currently, the number of COVID-19 cases and deaths among U.S. nursing home residents has been always above the average
- While various COVID-19 prediction models and research papers on nursing home risk factors are reported, no tools are available to predict the future outbreaks for nursing homes

Why is it important and why should we care:
- Predict the COVID-19 infection risk for each individual nursing home
- Help nursing homes to act swiftly on future outbreaks and health policymakers to analyze the facility's responses to infections and mortality

42% of all COVID-19 DEATHS in the United States are linked to nursing homes

In New York State, **20%** of our COVID-19 DEATHS have been linked to nursing homes

Reference: mainetti.law

Risk factors → Datasets → ML model (LightGBM) → Visualization

Multiple literatures

CMS government LTCfocus.org NYTimes

Google colab

tableau

Our approaches

What are they:
- Find the infection and mortality risk factors and collect the data
- Build a model to predict infection risk
- Visualize the historical time series data and model predicted data

How do they work:
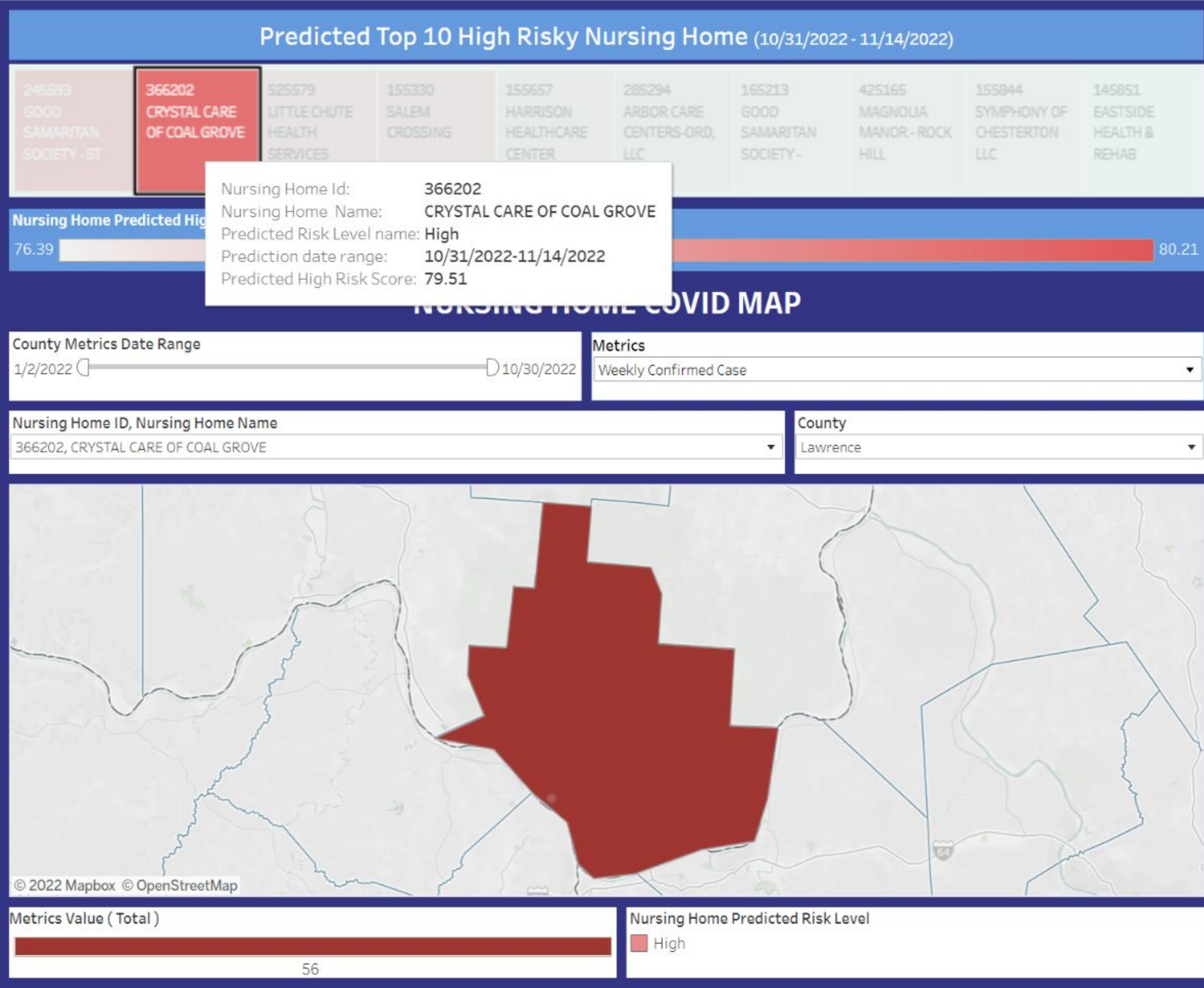- Summarize risk factors reported from multiple literatures
- Gather and Integrate data from CMS government, LTCfocus.org and NYTimes for model consumption
- Build a Machine Learning (ML) based model (LightGBM) to predict short term nursing home COVID-19 risk level using Google Colab
- Create final visualization input data with model prediction result and historical COVID trend, build interactive Tableau Dashboard map and publish to Tableau public server

Why can they effectively solve the problem:
- The model predicts the COVID-19 risk level (based on confirmed cases in the next two weeks and size of the nursing home) for each individual nursing home
- A published interactive Tableau Dashboard COVID-19 map provides nursing homeowners an effective tool to quick check their COVID-19 trend and risk

What is new in our approaches:
- Reliable and high-quality data source retrieved from CMS government, which follows federal reporting guidelines and is updated weekly
- Nursing home level COVID-19 historical and predicted data visualization
- Combined risk factors from various literatures for model prediction
- Use Machine Learning Algorithm (LightGBM model) to predict the COVID-19 infection risks for each nursing home

Data

How to get it:
- Download nursing home COVID-19 data from CMS government, LTCfocus.org and NYTimes
- Convert nursing home locations to latitude and longitude for map visualization by Google geocoding API
- Combine datasets from different data sources

What are its characteristics:
- CMS government data (752 MB, 119 Columns, 1,997,813 Rows): facility characteristics, staff-related factors, shortage, PPE, facility size, vaccination, treatment
- LTCfocus.org data (7.1 MB, 79 Columns, 14,441 Rows): Residents related factors
- NYTimes data (13 MB, 6 Columns, 416,024 Rows): County level COVID-19 data

Top of the dashboard, a tree map highlights model predicted top 10 risky nursing homes in next 2 weeks, it also set as a filter-to-filter COVID-19 Map below
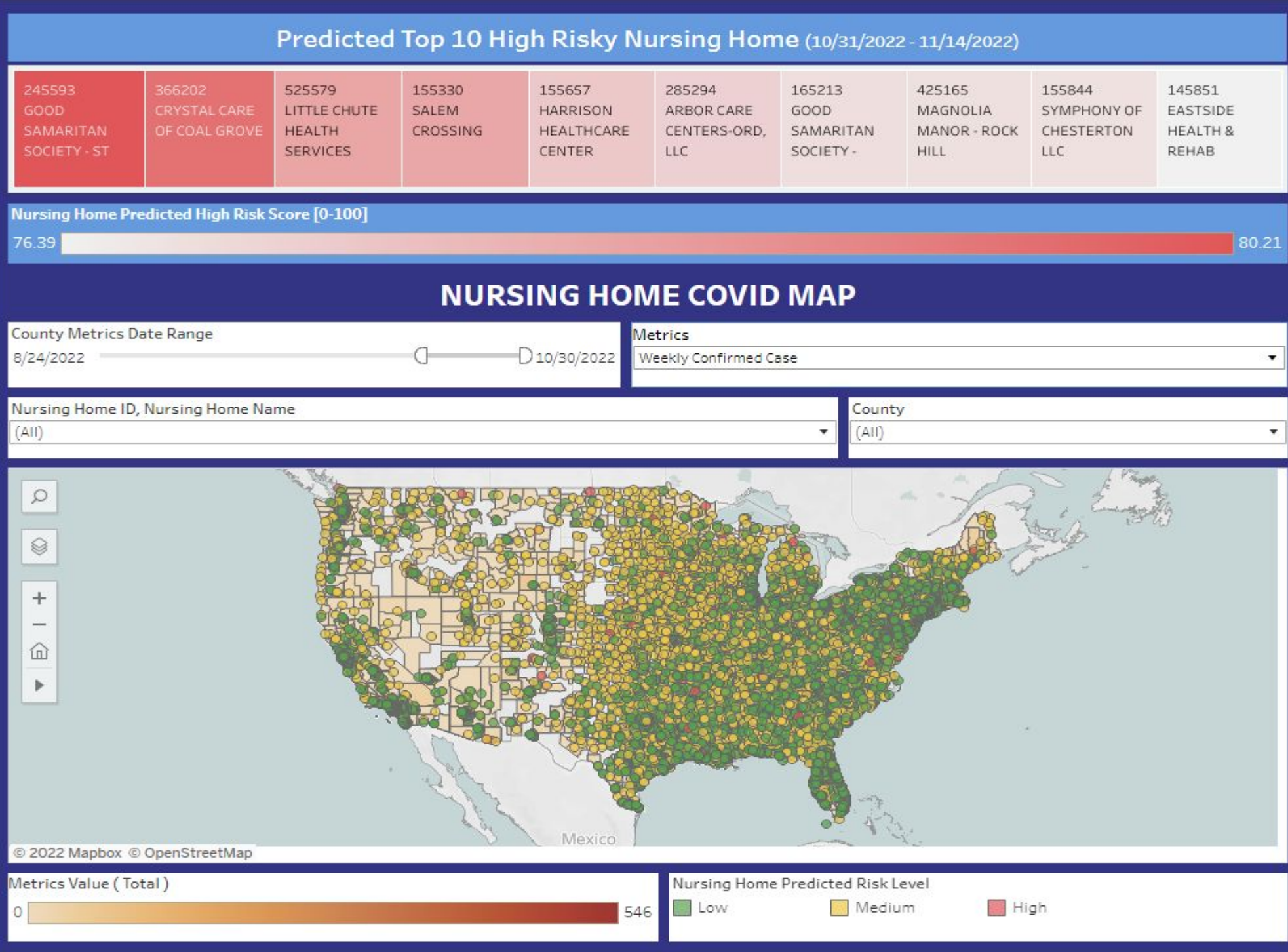
Bottom of the dashboard, a two-layer COVID-19 map (Nursing home layer and County layer) combined historical data and predicted risk level
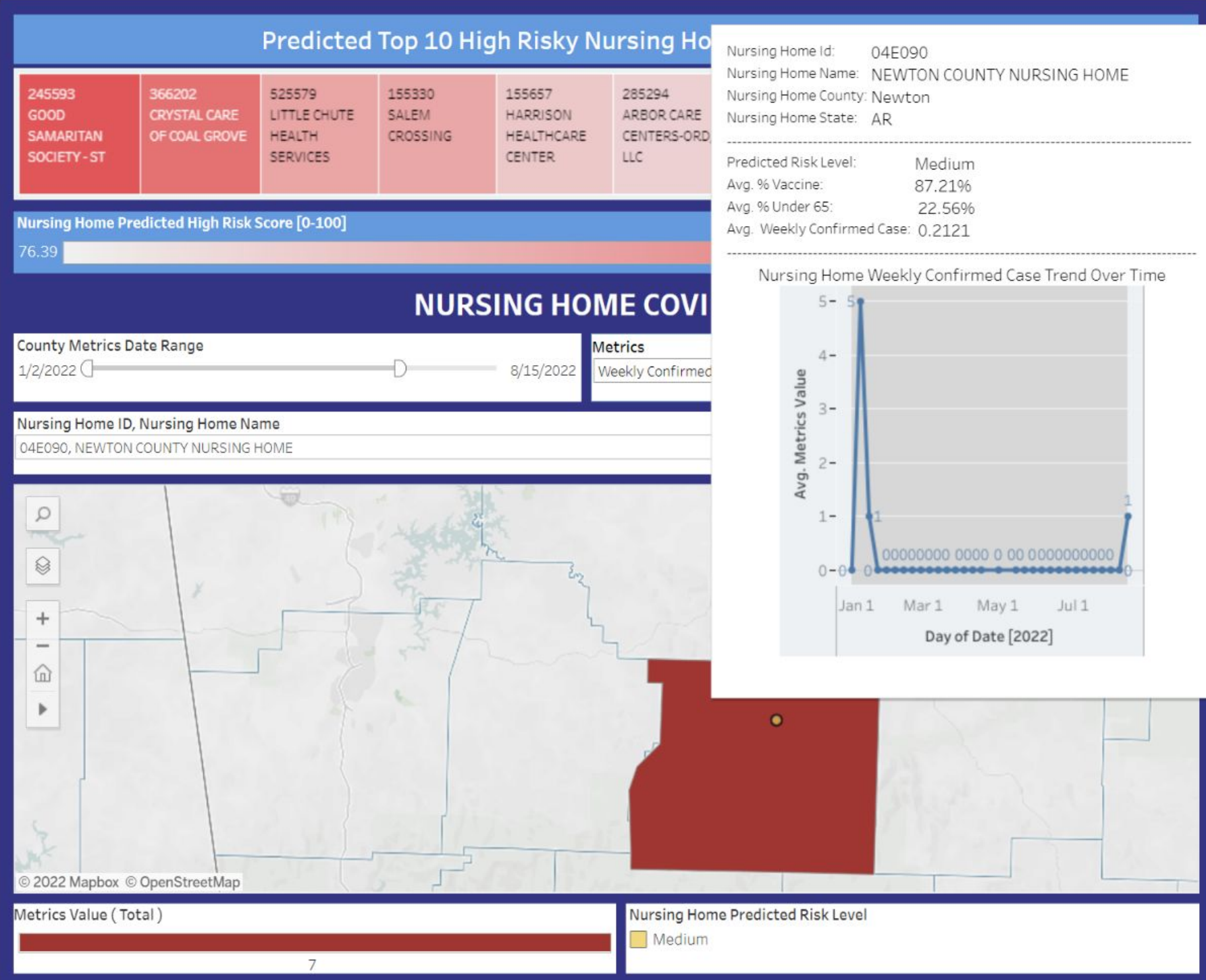
In COVID-19 map, multiple filters used to allow user to filter different COVID-19 metrics, date range etc., tooltips used to present more nursing home detailed information
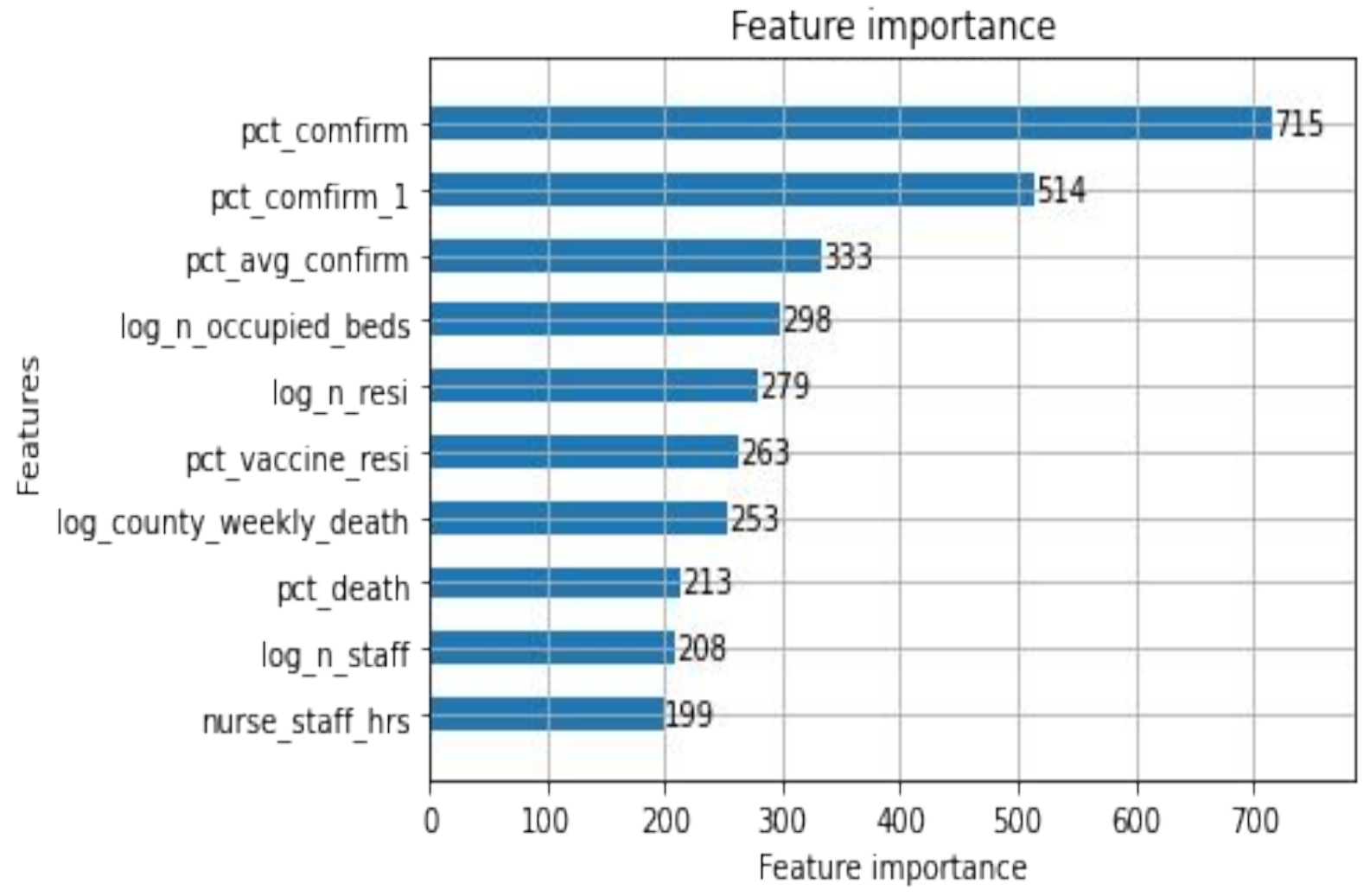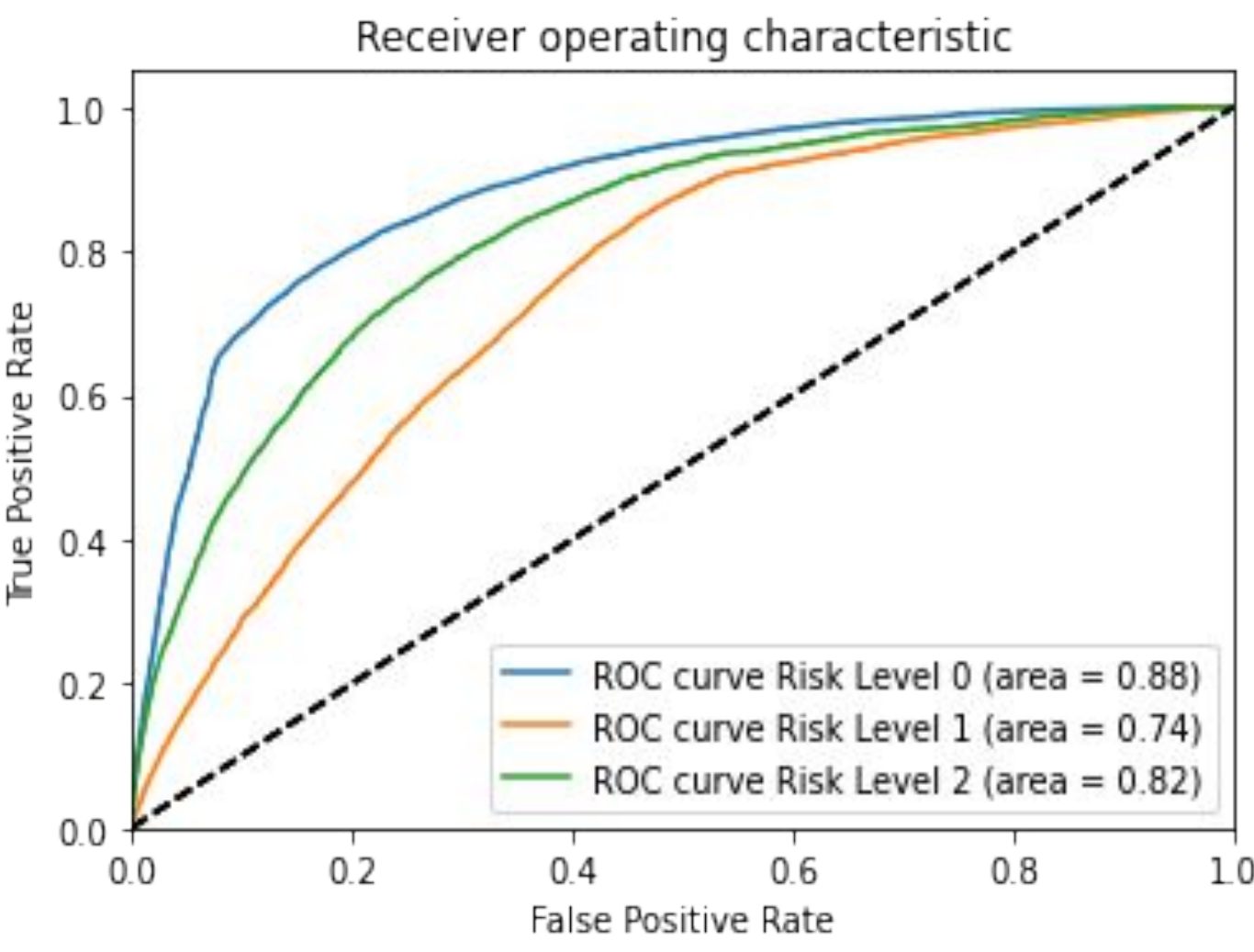
Experiments and results

How to evaluate our approaches:

Model:
- Model performance was plotted with Receiver Operating Characteristic (ROC)
- Rank the features based on feature importance

Visualization:
- Team members tested and provided user feedback
- All filters in dashboard clicked and tested functional
- Data presented in visualization dashboard validated against source data to ensure quality

What are the results:

Model:
- Model is good at identifying Risk level 0 as well as Risk level 2, while performing less accurately for Risk level 1. The model has a weighted average area under ROC of 82.6%
- Top 10 important features: infection history of the nursing home is very important, size of the nursing home, vaccination status and county level infection status are also very important

Visualization:
- Final Tableau dashboard published on Tableau server to allow all user access
https://public.tableau.com/app/profile/ruby1883/viz/team88-final-project/Dashboard1?publish=yes

Receiver operating characteristic
- ROC curve Risk Level 0 (area = 0.88)
- ROC curve Risk Level 1 (area = 0.74)
- ROC curve Risk Level 2 (area = 0.82)

Feature importance
- pct_comfirm 715
- pct_comfirm_1 514
- pct_avg_confirm 333
- log_n_occupied_beds 298
- log_n_resi 279
- pct_vaccine_resi 263
- log_county_weekly_death 253
- pct_death 213
- log_n_staff 208
- nurse_staff_hrs 199

How does the methods compare to other methods:
- Our model is unique in prediction of nursing home COVID-19 infection risk, and has a good performance
- Visualization dashboard allows nursing homeowners and health policymakers to quickly examine multiple COVID-19 metric trends and future risks at facility level and county level