Can Scraping Twitter aid in helping Local Governments better predict Covid Outbreaks?

Bernardo Oviedo, Blessing Oguyemi, Reese Roberts, Samuel Sorensen, Desi Warren II

Highlights

- Our model was able to beat the baseline and predict 66.65% of the variance in new COVID-19 cases in NY counties in a 7 day window.
- Higher levels of stringency and economic support were correlated with higher case numbers for more populous counties (medium sized and bigger counties)
- Counties with the highest number of new cases were also the counties with the highest rates of unemployment
- Government officials will be able to utilize our app to determine the effect decisions such as shutdowns and work / school closings will have on cases and deaths

Background

New York State was one of the earliest states to be hit the hardest by the Coronavirus with a record surge in hospitalizations, COVID cases, and deaths. Because of a huge lack of a coordinated and unified response from the United States federal government, New York government officials were left to address the incoming public health crisis with their own resources and within their own limits. In addition, this large void left by the federal government also left residents perplexed and confused, which was often reflected on Twitter. With the Twitter data, we are seeking to show a connection between Tweets , which will be in various counties in New York State, and policy initiatives, economic activity, and other measures of behavior.

Model	7-Day New Cases (R^2)	14-Day New Cases (R^2)	30-Day New Cases (R^2)
Always Predicting Mean	-0.0047	-0.0048	-0.0051
Always Predicting Median	-0.0048	-0.0625	-0.065
Decision Tree Regressor	**0.6665**	**0.5137**	**0.3**

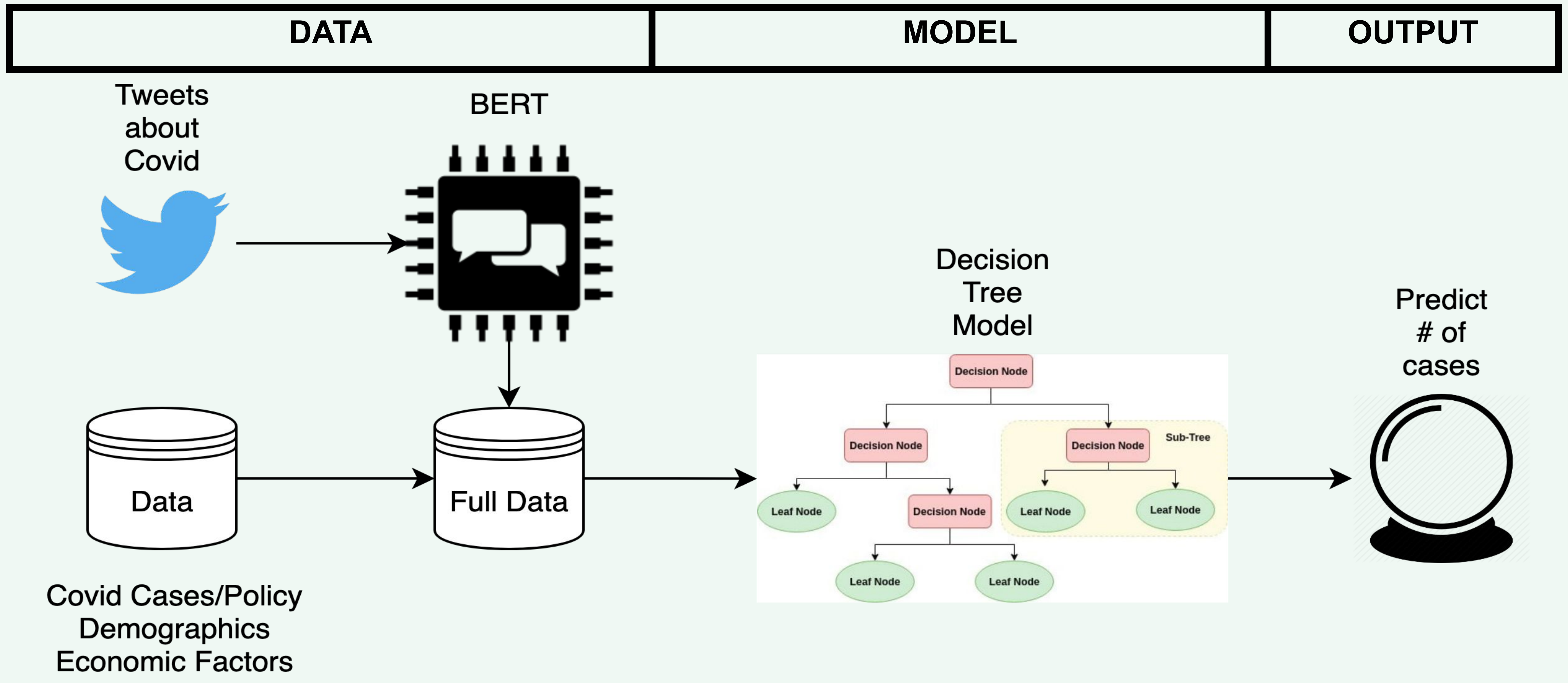
Data

- Daily Covid Cases for each County in NY
- Demographic Data by County
- Economic factors at a County and State level
- Indexes for rating government action during Covid i.e. economic aid, lockdown strictness
- Tweets for each county

Model

We chose a DecisionTreeRegressor because it can model nonlinear relations in the data without previous presuppositions on the degree of the "line" of best fit. Decision Tree models also usually perform well in various scenarios. The performance metric that we prioritized was the coefficient of determination (R^2).

This metric tells us the proportion of the variance in the dependent variable (daily number of new cases) that is predictable from the input features. This measure is generally used to test models whose main purpose is either the prediction of future outcomes or testing hypotheses. Since we are trying to predict the future number of daily new cases this measure made the most sense. Baseline results of the model are presented in table x.



DATA

MODEL

OUTPUT

Tweets about Covid

BERT

Data

Full Data

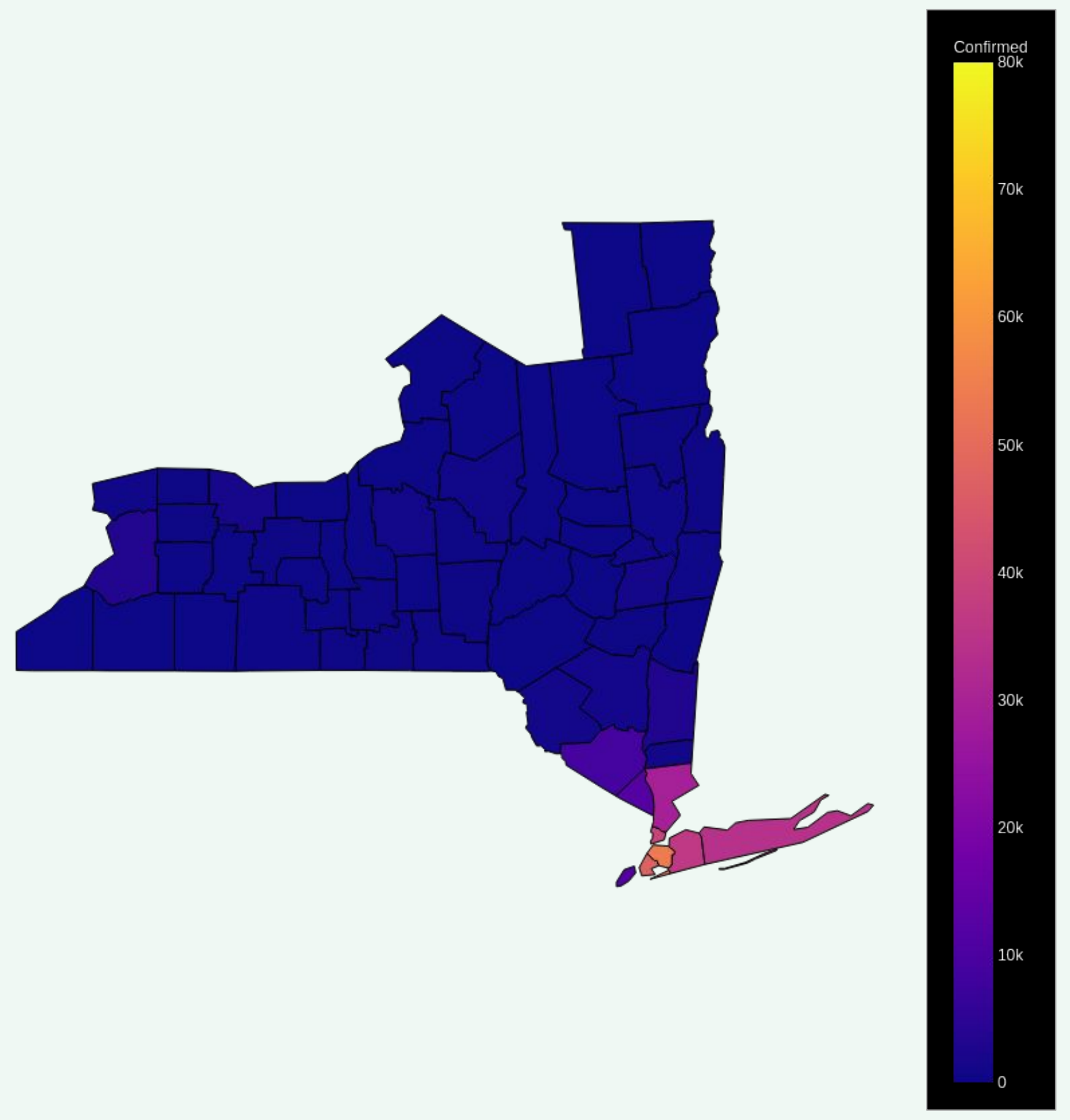
Covid Cases/Policy
Demographics
Economic Factors

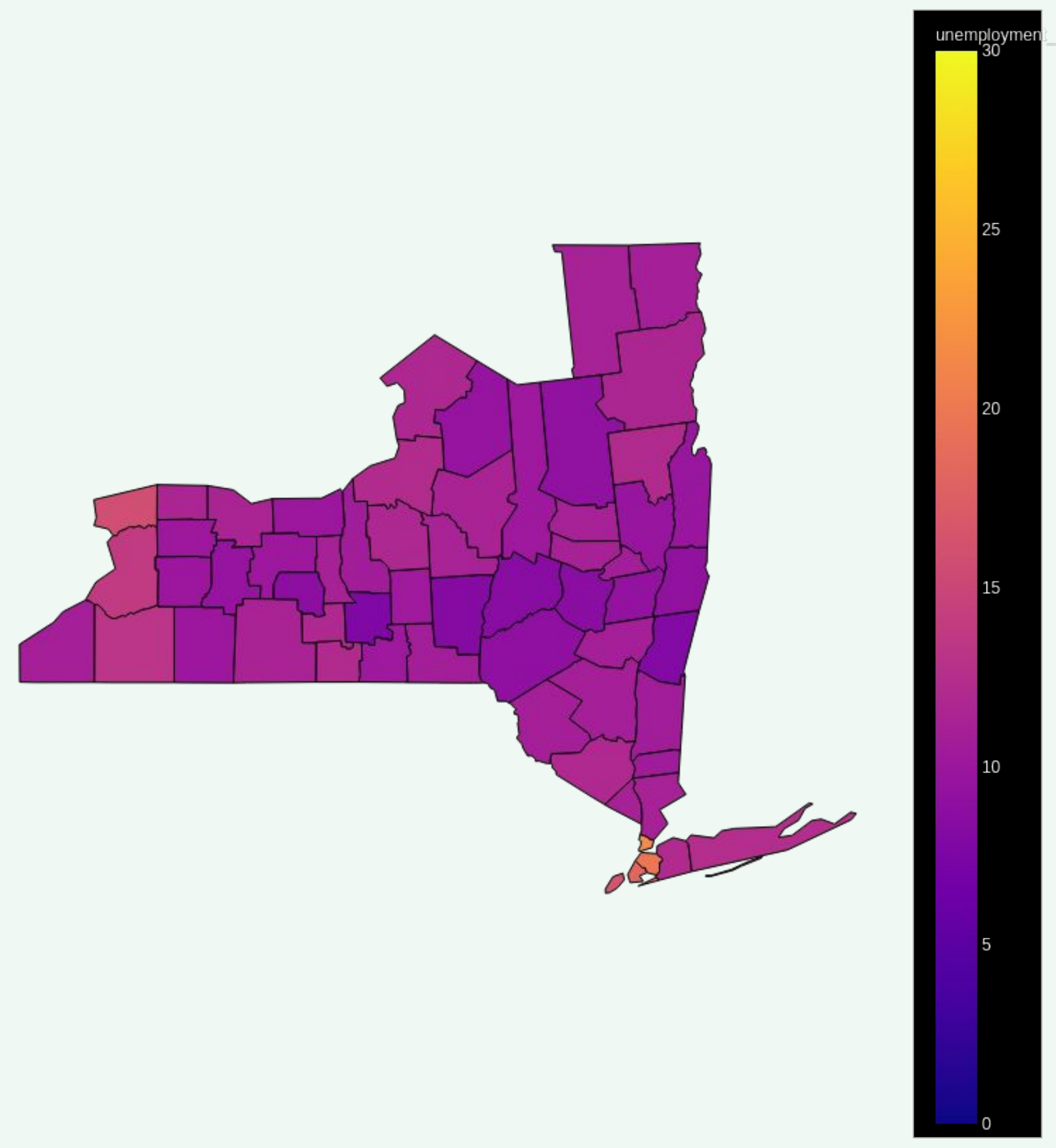
Decision Tree Model

Predict # of cases

Visualizations & Analysis

The visualizations below show how specific counties behaved differently during the pandemic. The figure on the left shows variations in confirmed COVID-19 cases across the state of New York, and the figure on the right shows the variation in unemployment rates for each county. There is a clear correlation between these factors as seen by similar 'hotspots' in areas such as Queens and Kings County. Correlations between behavior and demographics allowed for the division of all of the counties into three clusters for further analysis.

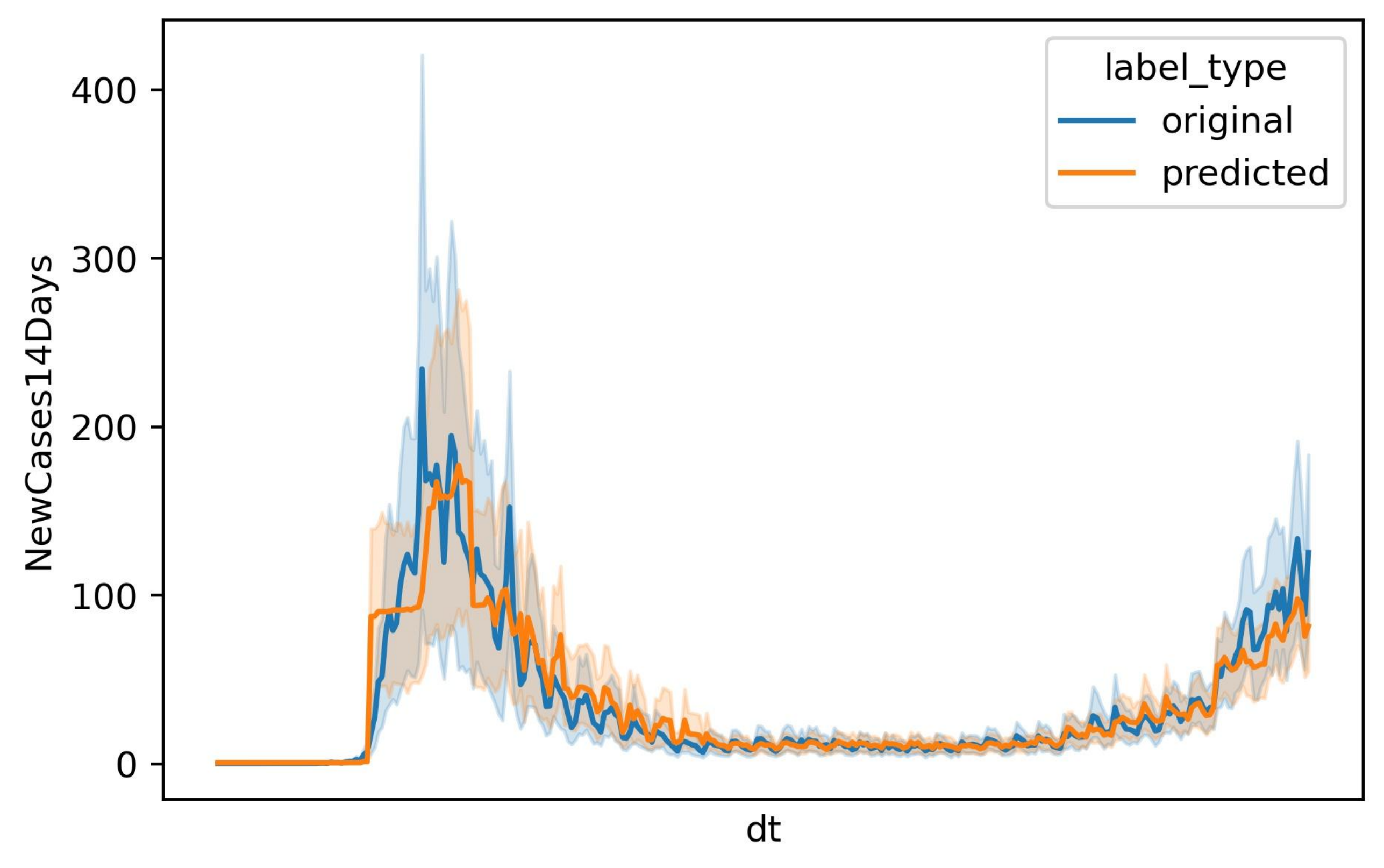




Model Performance

This figure shows the accuracy of the predictions of our model. The solid lines showcase the mean value, and the larger bands surrounding the mean line signify the 80% confidence interval of the model. There is a clear correlation between the predicted and original new cases found, therefore indicating that our model in accurate.

Due to its high performance, we recommend our model to New York policy makers or citizens who wish to predict the number of COVID-19 cases in their area resulting from policy changes.



label_type
original
predicted
NewCases14Days
dt