Predicting COVID-19 burden in NYC neighborhood

Chengda Zhang
5/16/2020

1. Introduction

1.1 Background

The New York City (NYC) has faced an outbreak of COVID-19 in the past 2 months. According to the data presented by NYC Health Department, different modified zip code areas (MODZCTA) are affected differently by COVID-19. Distribution of venues in NYC may directly affect the outbreak or indirectly associated with COVID-19 through close connection with local social economic situation.

1.2 Problem

More US cities are facing outbreak of COVID-19. Modeling case burden of COVID-19 in US cities will help us allocate proper resources and polices to certain areas based on available location data.

1.3 Goal

I aim to use location data from Foursquare to model the case burden of COVID-19 in different MODZCTAs of NYC.

2. Data acquisition and cleaning

2.1 Data source

COVID-19 case burden by MODZCTA and geographic boundary of each MODZCTA are obtained from NYC Health Department. NYC venues category information is queried from Foursquare.

2.2 Data cleaning

After downloading the data, we use geographic boundary to determine which MODZCTA a venue belongs to. We have discarded all venues that are not physically located in an MODZCTA.

I use the total number of venues in each category instead of a mean. Because this would better reflect the density of venues in an area, and should be better correlate with COVID-19 case numbers.

2.3 Exploratory analysis

I used category data to fit linear regression model, which returned an R2 of 1. However, when applying model to testing set, or using cross validation, the R2 drop to 0.3. This is clearly a sign of over-fitting.

In order to avoid over-fitting, we combined sub-categories of venues into 9 large groups. This combining is based on category tree from Foursquare. The 9 large groups that are included are: 'Food', 'College & University', 'Event', 'Arts & Entertainment', 'Nightlife Spot', 'Outdoors & Recreation', 'Professional & Other Places', 'Shop & Service', 'Travel & Transport', 'Residence'.

3. Results

When using linear regression, the number of venues in each groups could partly explain the difference in COVID-19 cases between MODZCTAs, with an R2 score of training set of 0.25, and testing set of 0.21. When using cross validation, the mean R2 score is -1.29. The distribution of testing set and predicted testing set is shown in Figure 1.

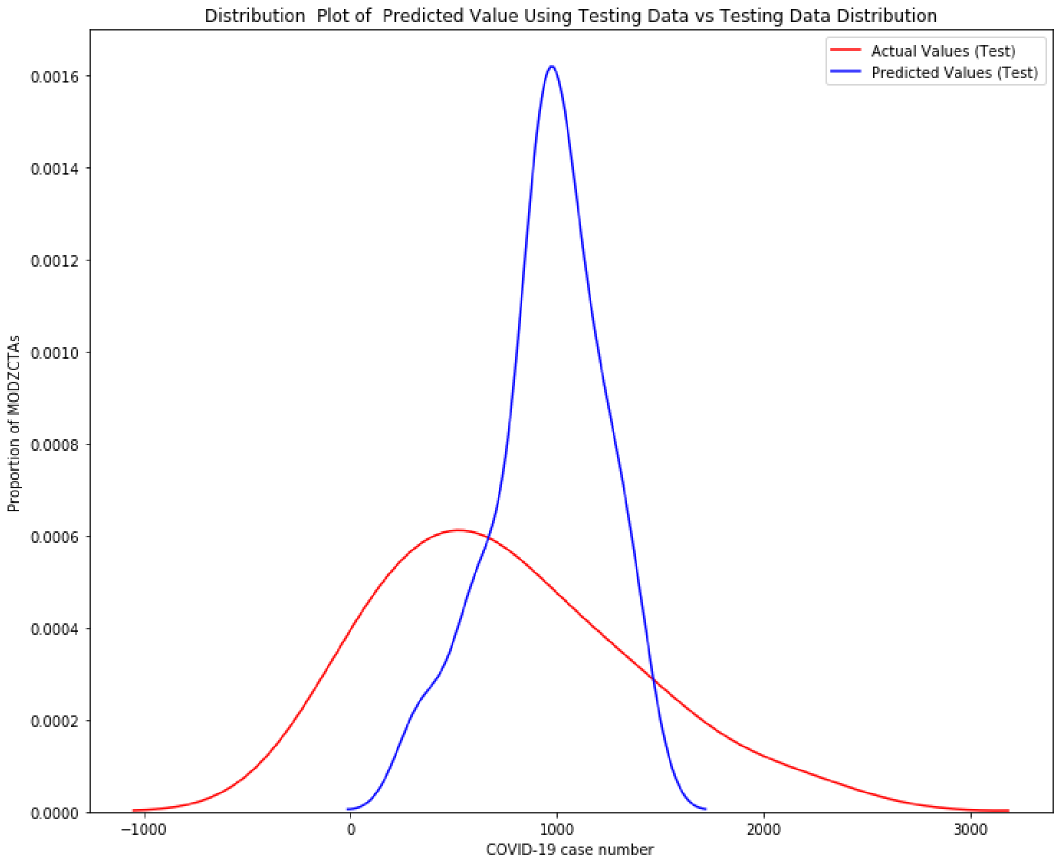

Figure 1 Distribution Plot of Predicted Value Using Testing Data vs Testing Data Distribution, in linear regression

When using Degree 2 Polynomial Regression, the model did not improve much, with an R2 score of training set of -39, and testing set of -491. When using cross validation, the mean R2 score is -44.3. The distribution of testing set and predicted testing set is shown in Figure 2.

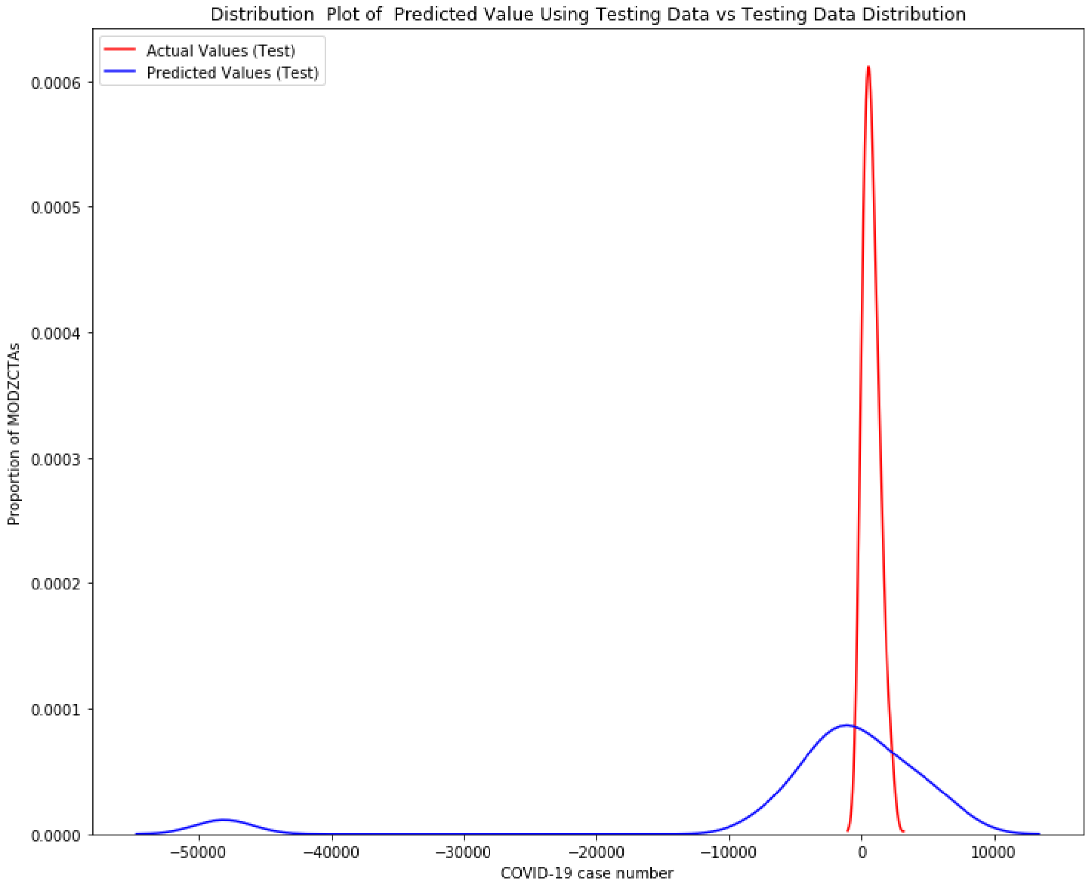

Figure 2 Distribution Plot of Predicted Value Using Testing Data vs Testing Data Distribution, in polynomial regression

4. Conclusion

The distribution of venues may partly explain the COVID-19 case distribution, however is not an ideal predictor for COVID-19 case burdens.