Impact of COVID-19 on the US Housing Market

Grace Wang, Jianan Song, Yifan Tang, Jiebei Luo, Po-Han Lai, Mingzhi Dai

ABSTRACT

The study is to quantitatively and visually explore the COVID-19 effects on key US housing market metrics. Starting from a demographic analysis of confirmed cases and reported deaths. The analysis will be performed at sate level with cumulative data over time. Machine learning techniques will be utilized to market trend prediction as if no pandemic happened. An interactive dashboard will be created to show how the the impacts on a map when predicted model parameters changed. This study will help to answer housing related questions during pandemic. The study can also provide evidence for policy makers to have more effective stimulus plan for economic recovery.

1 INTRODUCTION

In early 2020, a new respiratory disease, commonly known as the coronavirus disease (COVID-19), gripped the world([1] and [2]). Despite worldwide government shutdown orders, and practice social distancing, the disease inevitably triggered recession ([16], [7], [13]).

The pandemic directly impacts the economy through sharp decreases in consumer spending ([9], [15]) increase in unemployment ([3], [5], [8]), and volatility in equity returns ([17]). [8] shows that this pandemic disproportionately affected older workers of ages 65 and older, particularly for women, whose unemployment rate is significantly higher than men across all age or ethnicity groups. Meanwhile, a rapidly growing literature on the economic effects of COVID-19 investigate whether recent policy interventions have been successful ([6], [19], [16]).

Housing-related issues attracted immediate attention with shelter-in-place orders, self-isolation, and quarantine as the primary global responses to the COVID-19 pandemic. shows that countless people are struggling paying for house mortgages or rents due to the heterogeneous impact on industries and areas brought by the pandemic. With the pandemic, housing is more than a place of residence – it becomes an office, a classroom, and a shelter ([17] and ([12])). It would be particularly interesting to study how housing market has been affected by coronavirus and make predictions on how it may evolve in the near future.

All the literature above provide a global overview for our project on the economic with the COVID-19 situation which trigger our interest in the impact of house value during COVID-19 pandemic ([19] and [20]). However, we would like to know if uncertainty situations of COVID-19 affect the house price, house value and house inventory, while most literature about the socioeconomic consequences of COVID-19 focus on the labor market, health outcomes, gender and racial inequality, and environmental outcomes, the consequences of COVID-19 on housing market is almost an unexplored topic.

Among literature about effects of COVID-19 on the housing market, most of those articles either don't have much data of the correlations between housing market and COVID-19 or the data is only from one or two month time period ([2], [10] and [4]). Not to mention, there was very limited data visualizations to directly show the relations of the two from different perspective, especially when the housing market measures are limited to housing with federally backed mortgages ([18], [14] and [11]).

2 PROPOSED METHODS

2.1 Housing Price Prediction Model

We predict how the housing market perform during Covid-19 based on the housing data in the past three years. The innovation aspect of our approach is to train our housing market model based on housing data prior to Covid-19 and use the trained model to predict the housing performance during Covid-19.

The objective is to predict the total sold value in the housing market when the Covid-19 started, thus we choose ARIMA for time series analysis. ARIMA is a generalization of an autoregressive moving average (ARMA) model and widely applied in cases where data show evidence of non-stationarity.

To start with, we will first run time series model, autoregressive integrated moving average model (ARIMA),

as our baseline model and then improve them using Seasonal ARIMA (SARIMA) and Long Short-term memory (LSTM) models.

Here are steps we apply the ARIMA model: 1. Tune AR (Auto Regressive), I(difference), MA (Moving Average) parameters which are denoted as p, d, q. The initial values will be obtained using exploratory data analysis on the sold housing value. For the parameters AR and MA (p and q), we will use partial autocorrelation (PACF) plot and autocorrelation plot to decide initial values. For the difference (parameter d), we will use Augmented Dickey Fuller test which is a common statistical test used to test whether a given time series is stationary or not. 2. After having initial values for AR, I and MR (parameters p, d, and q), we will apply the grid-search to find the optimal values for the AR and MR combinations. We execute the search by adding one unit for each step and there are 9 possible combinations in our project. The best ARIMA performed model is determined by the minimum Root-mean-square deviation (RMSE) on the training dataset.

We extend it by using the Seasonal ARIMA to account for the seasonal effects in the housing market. Moreover, we plan to apply the Long Short-term memory (LSTM) model of the recurrent neural networks (RNN). RNN is a powerful type of neural network designed to handle the sequence dependence since it persists all previous information as a short-term memory and make decisions based on all stored information in the LSTM cells. In this project, we will use the Vanilla Long Short-term memory (LSTM) and multiples layer LSTM (stacked LSTM) models. Here are steps we implement the LSTM model:

1. Transfer our time series data into a supervised learning problem with input as X and output as Y, which is how Keras implements the LSTM model. Specifically, we use the house value in (t-1) period as our input X and the house value in t period as our output Y. Besides, we use the MinMaxScaler function in the sklearn package to transform the observations to have a specific scale since the neural networks expect data to be within the scale of the activation function.

2. After the data processing, we will first fit our model using the Vanilla LSTM, which has a single hidden layer of LSTM units and then extend to use the stacked LSTM, which has Multiple hidden LSTM layers, to forecast more complex data trends. For each LSTM layer, we construct inputs as samples, time steps and feature. Specifically, for our case, samples are our rows of house data. There is only one feature as we mainly use total sold housing value.

3. Tune parameters for the number of epochs, neurons, number of memory units. We will use the grid search to minimize the RMSE to determine the optimal values for the above parameters.

2.2 Housing Data Clustering

To evaluate the housing market performance during COVID-19, we apply clustering analysis to combine similar states as a group and investigate the effect on the group level. Clustering algorithm is an unsupervised Learning type.

For clustering that offers excellent data visualizations, K-means can be used to find states that may experience similar levels of COVID-19 impacts. Our clustering features include four COVID-19 metrics and one from housing market defined as “combined sold value” (CSV) to represent the overall housing price status within a state. The five features will be normalized into the same scale within 0 to 1 under the assumption that each feature is equally important. The results will then be validated using Gaussian Mixture model and this helps to build another evidence to confirm the groups are reasonable.

How K-means is utilized for the clustering? To perform K-means clustering at the state level, we randomly pick K centers and assign the states to each group according to the similarity defined as the distance. We will then recalculate the centers (the average of all data in each group) given the grouping results. We will keep updating the grouping and centers until converge. We define converge when all states are assigned to one of the K groups and the overall within-group variation is minimized. This overall variation is the sum of the distances between each data point and its assigned center.

Why and how Gaussian mixture model can be used to validate our clustering results? For the Gaussian Mixture Model, the underlying assumption is that the data are multivariate Gaussian distributed, and the data is the sum of them. It’s mathematical equivalent to K-means but the algorithms and metrics to deliver the results are different. We randomly pick K centers (the means and the standard deviations for the multivariate Gaussian distributions) and assign states to each group according to the probability of each data point based

on the distributions. We then recalculate the mean and standard deviation of the multivariate Gaussian distributions using the Estimated Maximum Likelihood (E-M) algorithm. Finally, we keep updating grouping, means and standard deviation until converge.

How to choose the parameter K? We plan to use gap-statistics to determine K. Gap-statistics compares the difference between the overall variation from our data (all states) with the variation from the same number of data with random values. Then we test out all K numbers and check the gap-statistics, the highest difference is associated with the right number of K.

The whole clustering steps including K-means and Gaussian Mixture Model are repeated with monthly Covid19 data, we will use the most common K as the final K values to generate the clustering results for each month. The reason to have common K is because the optimized K in each month could be different. The description of each group could be arranged as cluster: high new cases, high deaths with increasing housing price trend; cluster 2 low new cases and low deaths with stable housing price trend, etc. The weight of features may need to be fine tuned based on results.

2.3 Data Visualizations

Tableau is employed in our project as the primary visualization presentation platform. Essentially, the descriptive statistics of U.S. housing marking (2017-2020), COVID-19 data, and the analysis results of our model are realized by utilizing multiple visualization tools in Tableau with interactive features designed to demonstrate the impact of COVID-19 on the U.S housing marking in the past 10 months.

2.3.1 Visualization of U.S housing market. The line plot is chosen to demonstrate the change in house sales. Two variables are selected as the indicators, i.e., total number of houses sold by week, and mean sales price changes over time by week. Given the potential impact of the seasonality, the line plot is created using a year-over-year growth fashion that uses the last two years as the reference. Secondly, two filters are applied to two of the geographic parameters (i.e., region and state) in the housing data. The region-level filter provides the comparison of the housing market between areas with different population densities. The second filter applied at state-level creates a dynamic visualization that compares the U.S. housing market with and without the impact of COVID-19 across different states.

2.3.2 Visualization of COVID-19 cases and deaths. Bar charts and smoothing fitted lines are employed to visualize COVID-19 new cases, new deaths, cumulative cases and death trends in our dashboard. With the inclusion of time series case data (weekly), the bar charts with the trend line will be able to clearly demonstrate the development of the pandemic over time. Similarly, two filters are applied respectively to the two geographic parameters (i.e., region and state) that will create an interactive visualization demonstrating the differences of COVID-19's outbreak and development among states and in the areas with different population densities.

More importantly, four heatmaps will illustrate the four c decided by the case number. The number of cases relates to the Marks-color function, which will show the case number differences by color.

2.3.3 COVID-19 case clustering and U.S. housing market prediction results. In this part, we combine the COVID-19 clustering results with the heatmaps to interactively show the impact of COVID-19 on the housing market. Based on the clustering result from the previous section, we will mark each cluster with different color. For example, high new cases, high deaths with increasing housing price trends will be red, low new cases and low deaths with stable housing price trends will be blue.

The innovative aspect in our visualization approach is that the color scale will be determined by the percentage of change between the actual values and our predicted values (absent of the pandemic) for a variety of metrics of interest, i.e., total homes sold and median sale price, the results obtained from the ARIMA/LSTM/RNN models, etc. In particular, when moving the mouse pointer over the trend graph, the below heatmap will show the data for that month. And when moving the mouse from left to right, we can see the evolution of how the housing market deviates from the normal as the pandemic evolves as well as how the cluster changes over time.

3 DESIGN OF EXPERIMENTS

3.1 Data Sources and Processing

Dataset 1 is the housing market data by Redfin[1]: It provides housing market daily data for metropolitan areas, cities, neighborhoods and zip codes in the US. It includes more than 30 housing metrics(columns) with a combined data size 800MB that covers the data from 2017 to 2020. Dataset 2 is the COVID-19 data released by NY Times on Github[2]. This dataset shows daily cumulative counts of coronavirus cases in the US, at the state and county level. Dataset 3 is populated (inner join) from housing data and Covid-19 data discussed above with time frame March to October 2020. Python Pandas package is used for data process and cleaning. Keras, Statsmodels will be utilized for housing market predition. Skelearn will be applied for clustering construction. Tableau will be the main tool for data visualizations. The team will use Jupyter Notebook for coding interactively.

3.2 Prediction Evaluation

We split our data into three sub-datasets, including training, validation and test. Training data is designed for training the ARIMA, SARIMA, and LSTM models. Validation data is created for model comparisons. Test data is reserved only to be compared with predictions and visualizations. In our experiment, we used 03/2020 as a split point. Data after 03/2020 is test data. We will use 80% data before 03/2020 for training (01/2017 - 07/2019), and 20% for validation (08/2019 - 02/2020).

We apply the ARIMA model by preparing and tuning parameter on a training dataset and evaluating predictions with the validation dataset. After training our ARIMA model, we make a prediction on the validation data by taking a time series dataset as input as well as a tuple with the p, d, and q parameters required by the model, and calculate error score for predictions towards expected values. The combination of p, d and q with lowest RMSE will be used to finalize the ARIMA model.

In LSTM modelling, four different experiments will be performed with the four different numbers of time steps. We first use a representation with 1 time step as the default representation in a stateful LSTM and then use 2 to 4 time steps to see if the results can be improved. The hope is that the additional context from the lagged observations may improve the performance of the predictive model.Similarly, the LSTM model with lowest RMSE in validation data is our final LSTM model.

Finally, we compare the different models, including ARIMA, SARIMA, and LSTM, to select our final house value model and apply in our test data to get the true value without COVID-19.

3.3 Clustering Evaluation

Gap statistics will be used to determine the K, which compares the total within-cluster variation for different values of k with their expected values under the null reference distribution of the data (i.e., a distribution with no obvious clustering). In our context, it compares the difference between the overall within-cluster variation among all states with the variation from the data with random values. Next, we test out all K values and check the difference (Gap-statistics) to identify the right number of K according to the highest difference.

Next, we repeat the same steps for each month in our data. Given that the optimized K in each month could be different, we adopt the most common K as the final K value to generate the clustering results for each month. Once the grouping of clusters by state is obtained, all 50 states will be assigned to different clusters with the grouping arrangement, e.g., cluster 1: high new cases, high deaths with increasing housing price trend; cluster 2 low new cases and low deaths with stable housing price trend, etc. The final clustering is applied in our visualization, namely, in section 2.3.3: COVID-19 case clustering and U.S. housing market prediction results.

3.4 Survey

During the process of designing visualization dashboard, a survey will also be conducted to solicit feedback and suggestions. Particularly, based on audience's perspective, we will refer to the housing market trend graph and build the initial version of line plots and share among our team members and their friends.

4 INNOVATION LIST

(1) Various machine learning techniques including K-means, Gaussian Mixture model, ARIMA, LSTM, RNN have been used for prediction and clustering.
(2) In clustering, gap-statistics have been utilized to determine K with higher confidence.

[1]https://www.redfin.com/news/data-center/
[2]https://github.com/nytimes/covid-19-data

(3) In data visualization, we use parameters as filters that enable dynamically column switching.
(4) Interactive dashboard will be created based on user's preference e.g showing cluster only when moving mouse over the monthly trend of COVID-19, the heat map will be filtered to show the data and cluster for the month selected above.

REFERENCES

[1] Dave Altig, Scott Baker, Jose Maria Barrero, Nick Bloom, Phil Bunn, Scarlet Chen, Steven J Davis, Julia Leather, Brent Meyer, Emil Mihaylov, et al. 2020. Economic uncertainty before and during the COVID-19 pandemic. *Journal of Public Economics* (2020), 104274.

[2] Mariana C Arcaya, Yael Nidam, Andrew Binet, Reann Gibson, and Vedette Gavin. 2020. Rising home values and Covid-19 case rates in Massachusetts. *Social Science & Medicine* (2020), 113290.

[3] Louis-Philippe Beland, Abel Brodeur, and Taylor Wright. 2020. COVID-19, stay-at-home orders and employment: Evidence from CPS data. (2020).

[4] Nicholas Biddle, Ben Edwards, Matthew Gray, and Kate Sollis. 2020. COVID-19 and mortgage and rental payments: May 2020. *ANU Centre for Social Research and Methods, Canberra, June, viewed* 8 (2020).

[5] George J Borjas and Hugh Cassidy. 2020. *The adverse effect of the covid-19 labor market shock on immigrant employment.* Technical Report. National Bureau of Economic Research.

[6] Mike Brewer and Laura Gardiner. 2020. The initial impact of COVID-19 and policy responses on household incomes. *Oxford Review of Economic Policy* 36, Supplement_1 (2020), S187–S199.

[7] Abel Brodeur, David M Gray, Anik Islam, and Suraiya Bhuiyan. 2020. A Literature Review of the Economics of COVID-19. (2020).

[8] Truc Thi Mai Bui, Patrick Button, and Elyce G Picciotti. 2020. *Early Evidence on the Impact of COVID-19 and the Recession on Older Workers.* Technical Report. National Bureau of Economic Research.

[9] Raj Chetty, John N Friedman, Nathaniel Hendren, Michael Stepner, et al. 2020. *How did covid-19 and stabilization policies affect spending and employment? a new real-time economic tracker based on private sector data.* Technical Report. National Bureau of Economic Research.

[10] Vincenzo Del Giudice, Pierfrancesco De Paola, and Francesco Paolo Del Giudice. 2020. Covid-19 infects real estate markets: Short and mid-run effects on housing prices in Campania region (Italy). *Social Sciences* 9, 7 (2020), 114.

[11] Augusto Ricardo Delgado Narro and Yuya Katafuchi. 2020. COVID-19, state of emergency, and housing market. (2020).

[12] Walter D'Lima, Luis A Lopez, and Archana Pradhan. 2020. Covid-19 and housing market effects: Evidence from us shutdown orders. *Available at SSRN 3647252* (2020).

[13] Nuno Fernandes. 2020. Economic effects of coronavirus outbreak (COVID-19) on the world economy. *Available at SSRN 3557504* (2020).

[14] Laurie Goodman and Dan Magder. 2020. Avoiding a COVID-19 Disaster for Renters and the Housing Market. *Washington, DC: Urban Institute* (2020).

[15] Akos Horvath, Benjamin Kay, and Carlo Wix. 2020. The COVID-19 Shock and Consumer Credit: Evidence from Credit Card Data. *Available at SSRN 3613408* (2020).

[16] Peterson K Ozili. 2020. Covid-19 pandemic and economic crisis: The Nigerian experience and structural causes. *Available at SSRN 3567419* (2020).

[17] Peterson K Ozili and Thankom Arun. 2020. Spillover of COVID-19: impact on the Global Economy. *Available at SSRN 3562570* (2020).

[18] Yichen Su and Sitian Liu. 2020. The Impact of the COVID-19 Pandemic on the Demand for Density: Evidence from the US Housing Market. *Available at SSRN 3661052* (2020).

[19] Rob Wallace, Alex Liebman, Luis Fernando Chaves, and Rodrick Wallace. 2020. COVID-19 and Circuits of Capital. *Monthly Review* 72, 1 (2020), 1–13.

[20] Pengpeng Yue, Aslihan Gizem Korkmaz, and Haigang Zhou. 2020. Household financial decision making amidst the COVID-19 pandemic. *Emerging Markets Finance and Trade* 56, 10 (2020), 2363–2377.