Principles of Urban Informatics Final Project

Cici Chen, Tao Liang, Jingjing Ge, Jiale Li, Yichen Guo, Wangtianhan Pang, Xinyu Xu.

12/10/2021

Stanislav Sobolevsky

NYC Housing price prediction

-- How feature importance change before and during COVID-19

Abstract

With the hit of the COVID-19 pandemic, we observed great fluctuations of housing prices. The main goal of this project is to analyze the impact of the COVID-19 on the housing price prediction model for NYC by observing how features' importance changes over time. While many previous studies selected features limited to the direct attributes of apartments, this project selected 31 features from 49 datasets that are potentially related to the real estate value, including the larger neighborhood characteristics and safety factors. This study used rolling sales data from 2019 to 2021 to combine with other public datasets from NYC Open Data and OpenStreetMap for 5 different machine learning models. Random Forest model was selected as the best performance on the prediction and was used to analyze the changes in feature importance. The results showed that house area dominated in the sales price prediction model and fluctuated slightly from 2019 to 2021. The impact of the distance to subways stations changed most obviously before and during the COVID19.

Key Words: Pandemics, Real Estate, Machine Learning, Big Data

Introduction

To explore the impact of the coronavirus disease since 2019 on real estate in NYC, we conducted a data-model based on research of various housing market indicators from January 2019 to November 2021. The model uses data that are correlated to citizens' health and wellbeing, housing values, and the local economy. Our expected outcome is an advanced housing price prediction model that incorporates street health-related variables that can offer a data-led view of the physical and experiential quality of the housing environments. In our project, we incorporated healthy street indicators to evaluate how the street network affects the valuation of residential properties and how the housing market patterns are influenced by the spatially varying estimates on streets. Our research also includes analyses of urban design characteristics, such as street tree density, distance to nearby schools, and public transit, and how these could affect the price of residential property for sale. The results show that street accessibility can be an important influence at a sub-market level. It found that a price premium exists in neighborhoods with an improvement in inner connectivity from smaller blocks and greater pedestrian accessibility to commercial areas.

The remainder of this report is structured as follows. "Key questions" section presents the two main questions that our research is aiming to answer. "Literature review" section includes analysis about existing house price prediction model analysis and variables review. "Data" and "Methodology" sections clarify our data selection and the empirical steps taken in the model building process. "Research results" section claims our final and factual results and model. "Discussion" section shows some challenges and risks we met during the whole research. "Conclusions" section provides the summary of our main findings and connects the significance of the results of the main points.

Key Question(s)

- Which features will people be concerned with more when finding homes in the COVID era?
- Besides the features related to the apartment's conditions, which other features will dominate the predictions of housing price before and during the COVID-19 eras?

Literature Review

According to the Healthy Streets for London project, a healthy street is defined as an "easily navigated and comfortable street with a visually appealing, activated street space" that holds the following attribute[1]:

- Protect road users
- Boost neighborhood vitality
- Increases accessibility and mobility by providing linkages to employment and services
- Provide transit options
- Promote sustainable transportation modes and physical activity

In another study conducted by Liu and Su about the impact of COVID19 on the demand for density, they discovered the decrease in housing demand in central neighborhoods and neighborhoods with higher population density. They also pointed out that the areas with higher pre-COVID19 home values suffered greater decline in housing needs.[2] In the research of Grybauskas, Andrius, Vaida Pilinkienė, and Alina Stundžienė in 2021[3], they introduced the idea

[1] "Healthy Streets for London - Transport for London," accessed October 17, 2021, https://content.tfl.gov.uk/healthy-streets-for-london.pdf.

[2] Sitian Liu and Yichen Su, "The Impact of the COVID-19 Pandemic on the Demand for Density: Evidence from the U.S. Housing Market," SSRN, August 3, 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3661052&download=yes.

[3] Grybauskas, Andrius, Vaida Pilinkienė, and Alina Stundžienė. "Predictive Analytics Using Big Data for the Real Estate Market during the COVID-19 Pandemic." Journal of Big Data 8, no. 1 (2021). https://doi.org/10.1186/s40537-021-00476-0.

of exploring which predictors are best to anticipate price changes. Inspired by what they did, this project generated the idea of exploring the feature importance changes of different factors affecting house prices in the context of COVID19.

Variable Review:

This project starts with variable review. Here is a short summary of these variables. The most commonly discussed and most frequently used in hedonic models variable, as Sirmans, Stacy et al.[4] researched in 2005, is **age**, which typically has the expected negative sign although it is seen to be positive and not significant in some studies. **Square footage and land square feet**, two quantitative variables, are the next most used characteristic and typically have a strong influence on house Price. **Year built** of property is another determined factor we used to analyze the value of it. Other urban Environmental factors created by neighborhood or location include **resident income**, **race**, **location**, **crime**, **distance**, and **urban forest density**. **Urban forest density**, encompassing the trees and shrubs in urban areas, are also seen to consistently have a positive effect on price. **Distance from or adjacent to a park** also appears to enhance sales price, although corner location does not appear to be an influence. Other environmental characteristics resulting from public services include the **school district** variable, **distance to the public transportation, distance to recent shopping center, distance to restaurant, noise control level**, **median income, race (the percent of Asian, black, white), population density** and **the percent of people over 62.** In 2019, Gao et al.[5] consider the distance to the shopping

[4] Sirmans, Stacy, David Macpherson, and Emily Zietz. 2005. "The Composition Of Hedonic Pricing Models". Journal Of Real Estate Literature 13 (1): 1-44. doi:10.1080/10835547.2005.12090154.

[5] Gao, Guangliang, Zhifeng Bao Bao, Jie Cao, A. K. Qin, Timos Sellis, and Zhiang Wu. 2019. "Location-Centered House Price Prediction: A Multi-Task Learning Approach". https://arxiv.org/abs/1901.01774.

centre as a variable in their house prediction model. **Zip Code** describes Location, generally measured as a neighborhood identifier in the real estate market.

Data

Data collection:

This research mainly uses open data from government websites and databases. The project uses rolling sales data from NYC Department of Finance as the main training, validation and testing sets for the housing price in the model. The datasets are publicly available from the nyc.gov for the year 2019 to 2021 separated by boroughs and include information about the properties sold such as sale price, location, gross square feet, and year built. For the features that would include in the models, the socioeconomic datasets (age, income, race, population) were collected from ACS 5-Year Estimates. Datasets about the neighborhood characteristics and built-in environment came from NYPD complaint data and 311 complaint data (noise) as well as the tree points, subway and bus station location datasets on NYC Open data. Other datasets related to the point of interest include park, shop, cuisine were exported from the OpenStreetMap using QGIS. All the datasets were joined based on the geographical level of census tract in 2010, for which the shapefiles were collected from the Department of City Planning official website.

Data Cleaning and Processing:

As the study focused on the change of features influencing the housing price in the model due to the COVID-19, the datasets used were first filtered to the year from 2019 to 2021. Then we cleaned the datasets by dropping missing values and extreme outliers in quantile 0.01 and 0.99. In order to merge all the features data into one dataset based on the census tract, we used spatial join on the geopandas to count the number of features (crime, noise complaints, trees, shops, etc) in each census tract and some features were normalized by the area of the census tract to be unbiased. For some point of interest variables (parks, subway, etc), the distances to the points were calculated rather than counting the number. As the rolling sale data doesn't provide the specific geometry of the property, we used the Borough Block Lot (BBL) columns from the data to join with PLUTO datasets that contain geometry information and further spatial join with the census tract dataset in order to combine with other datasets. The final merged dataset was each property's information and its neighborhood characteristics, demographics, and built-in environment based on its corresponding census tract. The datasets were further standardized with dummy variables based on the models we used.

Methods

We used 5 machine learning methods and the training dataset was fit into each model: KNN Regressor, XGBoost, Random Forest Regressor, Bagging (Bootstrap Aggregation), GradientBoost (Table 1). The final model was chosen based on the r-squared value of the fitted model. R-squared measures the goodness of fit, which represents the fraction of variance of the response variable reflected by the model.[6] We also used mean root squared logarithmic error (RMSLE) as a scoring reference. We chose RMSLE over root mean squared error (RMSE) because the RMSLE is more generous on large errors, which is expected in our dataset since housing prices range from different magnitudes. The explanation of feature importances was carried out both globally and locally. For global scope feature importance, we used the built in "feature_importances_" attributes. SHAP was used for feature importance explanations for samples chosen across the timeline. More specifically, we choose 5 samples from each month

[6] Ajitesh Kumar, "Mean Squared Error or r-Squared - Which One to Use?," Data Analytics, September 30, 2020, https://vitalflux.com/mean-square-error-r-squared-which-one-to-use/.

and plot the SHAP force plot for each observation in the same graph. In this case, we would have a feature importance explanation across the timeline, which ideally would reveal the change of feature importance from pre-COVID to COVID time.

Results

Model Selection:

Based on the evaluation output of the five models chosen (Table 2), RandomForestRegressor performed the best with the highest R2 of 0.813 and an RMSLE of 0.1, so it was chosen as our final model to fit the data.

General Trends (Overall feature importance):

Figure 1 shows the feature importance using the built in feature importance attribute of the RandomForestRegressor. It's obvious that the "GROSS SQUARE FEET" is the major contributing feature in housing price prediction, followed by the "HLINE_D", "EMPIRE_D", "LAND SQUARE FEET", "School_MinD" and "Cuisine_density".

Pre-COVID and COVID feature importances:

Figure 2 shows the overall feature importance over time using the SHAP explainer. The x-axis on the top shows the indexes of data. The data used for SHAP was chronologically ordered data with 5 observations picked from each month. For example, 60 would mean 12 months away from January 1st, 2019 (the date the data starts). We can see that there isn't a significant difference in the overall trend besides more fluctuations of feature importance that can be observed during peak COVID time (indices 60 to 120) compared to pre-COVID (indices 0 to 60).

We also looked at the feature importance changes throughout the timeline using SHAP (from January 2019 to September 2021). For the five most important features in our RandomForest regression model. Contrary to the overall feature importance fluctuations, there are fewer fluctuations in the importance of distance to High Line park, distance to the closest school, and cuisine density. While the most important feature, gross square feet, did not show any important changes. For the cuisine density, we can conclude that between early 2020 (index 60) and mid-2021 (index 140), there were no negative impacts on housing price and there were more positive correlations with higher impact (more peaks in the graph). While for the distance to Highline Park, there is a decrease in the negative impact during the COVID time. The trends start to revert back to pre-COVID around index 150 (July 2021), which has a similar trend compared to July 2019 (index 30).

Figure 4 shows other variables with obvious changes in feature importance during the COVID-19 peak period, we can see that there's a decrease in feature importance of minimum distance to closest subway, a decrease in the negative impact of the violations (level of criminal offense) and an increase in shop density (more peak values) and median income impact.

Conclusion

General trend:

The COVID-19 has dramatically impacted the real estate market in New York City. This project discovered what features people care about when they find their houses. Besides the properties of the house itself (area, built year, and location), this project also explored how socioeconomic and environmental features impacted the house sales price. The variable that has the most impact on the model's output is the gross square feet while other contributing features are related to the built environment. For example, minor contributing features include distance to major tourist attractions such as High Line Park and Empire State Building; the distance to the closest subway station and school. So we can see from the general trend that besides the size of the residential space, the location that determines the ease of accessibility to urban infrastructures is more related to higher sale prices.

Pre-COVID versus COVID period:

From the SHAP explainer, we can see that there were no persuasive general feature importance changes observed during the COVID-19 time. While looking at the feature changes from the individual level shown in Figures 3 and 4, we may conclude that during COVID-19, the need for easy accessibility to food was increased, possibly because people didn't want to go dining far away from the home due to health concerns. This coincides with the decreased importance of shorter subway distance to the house because people were less likely to go out due to lockdown and remote work/study at home. Similarly, there is less importance of the location of the house near the landmark such as Highline as people didn't consider the busy city center as a dominant factor since they were more likely to stay at home. The importance of crimes, distance to schools and median household income are relatively stable compared to previous periods.

In general, the size of the house has a deterministic effect on the housing price over time. Other features that have more impact on the housing price are usually correlated with the accessibility of urban infrastructure and convenience of living environment. Overall, based on our observation of the model result, there are some impacts of COVID 19 on housing decisions as people are less likely to have outside activities.

Discussion

It was to collect several types of data and databases for all related perspectives through the process. For example, we were not able to get the detailed attributes for each household such as the number of bathrooms, number of bedrooms. There were a couple of unrelated factors in our initial running models and results. Or there may be some indexes with a large impact, which will affect our research on other factors.

It may be difficult to preprocess and integrate a large amount of data from heterogeneous sources, including data cleaning, setting the same time frame for the data from different sources, and controlling variables. The raw datasets also had inconsistent geographical attributes which required a large amount of preprocessing before being put into use.

For the model, since there are missing household-level detailed attributes data, the model may not reach its best performance. Some features used in the model may be correlated with each other, which may make the model sensitive to minor changes in the data. In the future, it would be better to incorporate household-level data and minimize the potential correlations among features used. Another finding that requires further adjustment is the extremely high weight for the “GROSS SQUARE FEET” feature, which has a potential impact on better exploration of other features’ impact.

The SHAP explainer did not reveal an absolute pattern of general feature importance changes. While at the individual feature level, the changes varied. Given the fact that only 5 random samples were drawn from each month and running the SHAP model was time costly, it would be better to look at more samples from each month to see a generalized pattern on a better machine if time allowed.

Tables and Figures

Tables:

Table 1: 5 machine learning models used

KNNRegressor	KNNRegressor stands for k-Nearest Neighbors regressor. It is a nonparametric regression method which keeps track of the last window_size training samples. Aggregating the values of the closest n_neighbors stored-samples with respect to a query sample will obtain predictions[7].
XGBoost	XGBoost stands for eXtreme Gradient Boosting. It is an optimized distributed gradient boosting library with high-efficiency, flexibility and portability. It uses machine learning algorithms within the Gradient Boosting frame[8].
Random Forest Regressor	Random Forest Regression is a supervised learning algorithm. It uses ensemble learning methods for regression, which combines multiple machine learning methods' predictions. Thus it conducts a more accurate prediction than single model[9].
Bagging (Bootstrap Aggregation)	Bagging, also known as Bootstrap Aggregation, is a popular ensemble method that fits a decision tree on different bootstrap samples of the training dataset. Wide range of problems, and importantly, modest extensions are easy to implement and effective on the technique result in ensemble methods that are among some of the most powerful techniques such as random forest, that perform well on a wide range of predictive modeling problems[10].
GradientBoost	GradientBoost builds an additive model in a forward stage-wise fashion and it is able to optimize arbitrary differentiable loss functions[11]**.**

[7] The river developers, "Knnregressor," River, accessed December 12, 2021, https://riverml.xyz/dev/api/neighbors/KNNRegressor/.

[8]XGBoost developers. "XGBoost Documentation." XGBoost Documentation - xgboost 1.5.1 documentation. Accessed December 13, 2021. https://xgboost.readthedocs.io/en/stable/.

[9] Chaya Bakshi, "Random Forest Regression," Medium (Level Up Coding, June 9, 2020), https://levelup.gitconnected.com/random-forest-regression-209c0f354c84.

[10] Jason Brownlee, "Essence of Bootstrap Aggregation Ensembles," Machine Learning Mastery, October 21, 2021, https://machinelearningmastery.com/essence-of-bootstrap-aggregation-ensembles/.

[11] Afroz Chakure, "Random Forest Classification and Its Implementation," Medium (The Startup, November 6, 2020), https://medium.com/swlh/random-forest-classification-and-its-implementation-d5d840dbead0.

Table 2: R2 and RMSLE of 5 regression models

	RandomForestRegressor	XGBRegressor	KNeighborsRegressor	BaggingRegressor	GradientBoostingRegressor
R2	0.812532	0.771972	0.721372	0.802487	0.790581
RMSLE	0.100039	0.104740	0.117335	0.104379	0.112602

Figures:

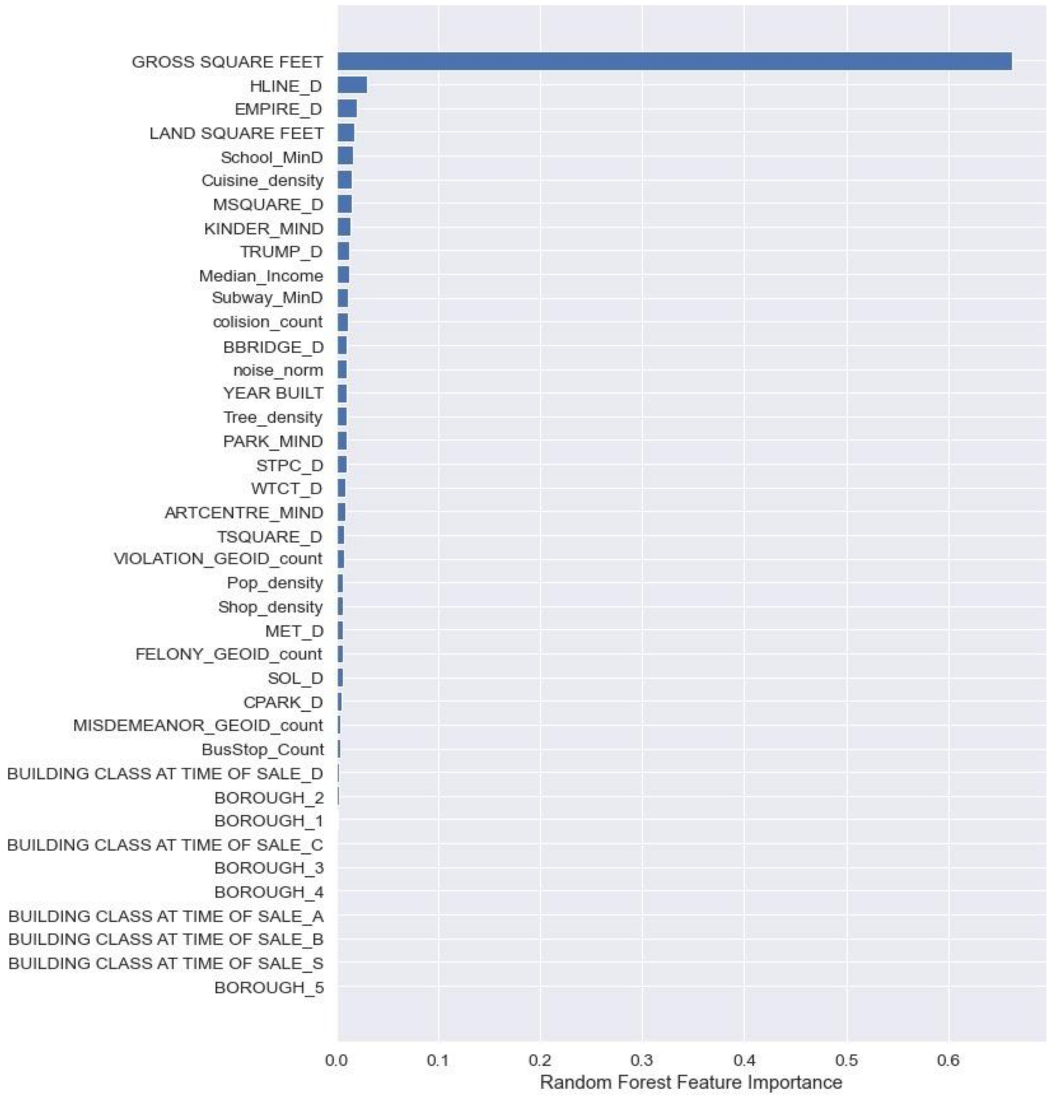

Figure 1: Feature importance from Random Forest Model

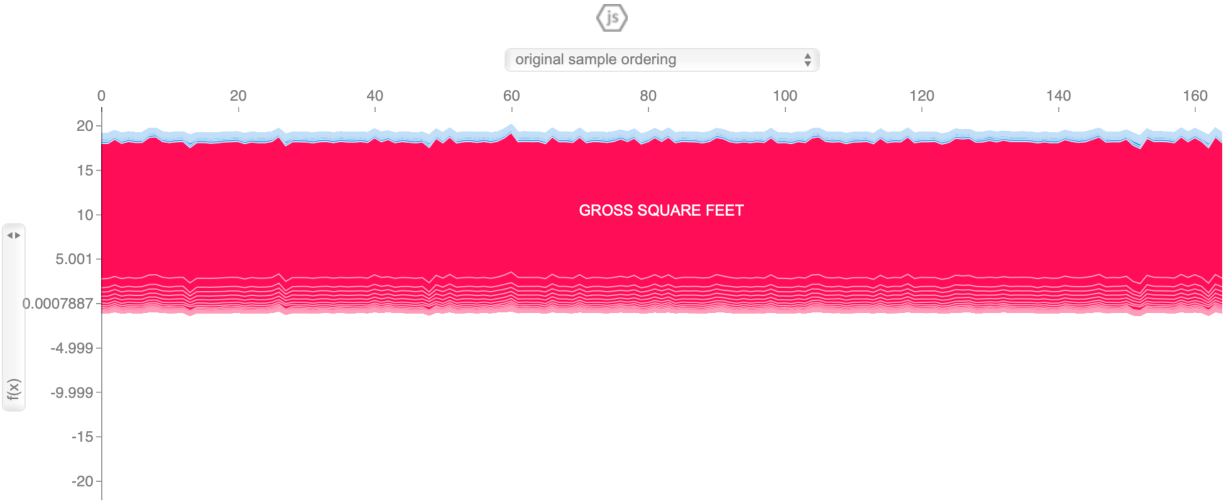

Figure 2: Overall feature importance explanation using SHAP

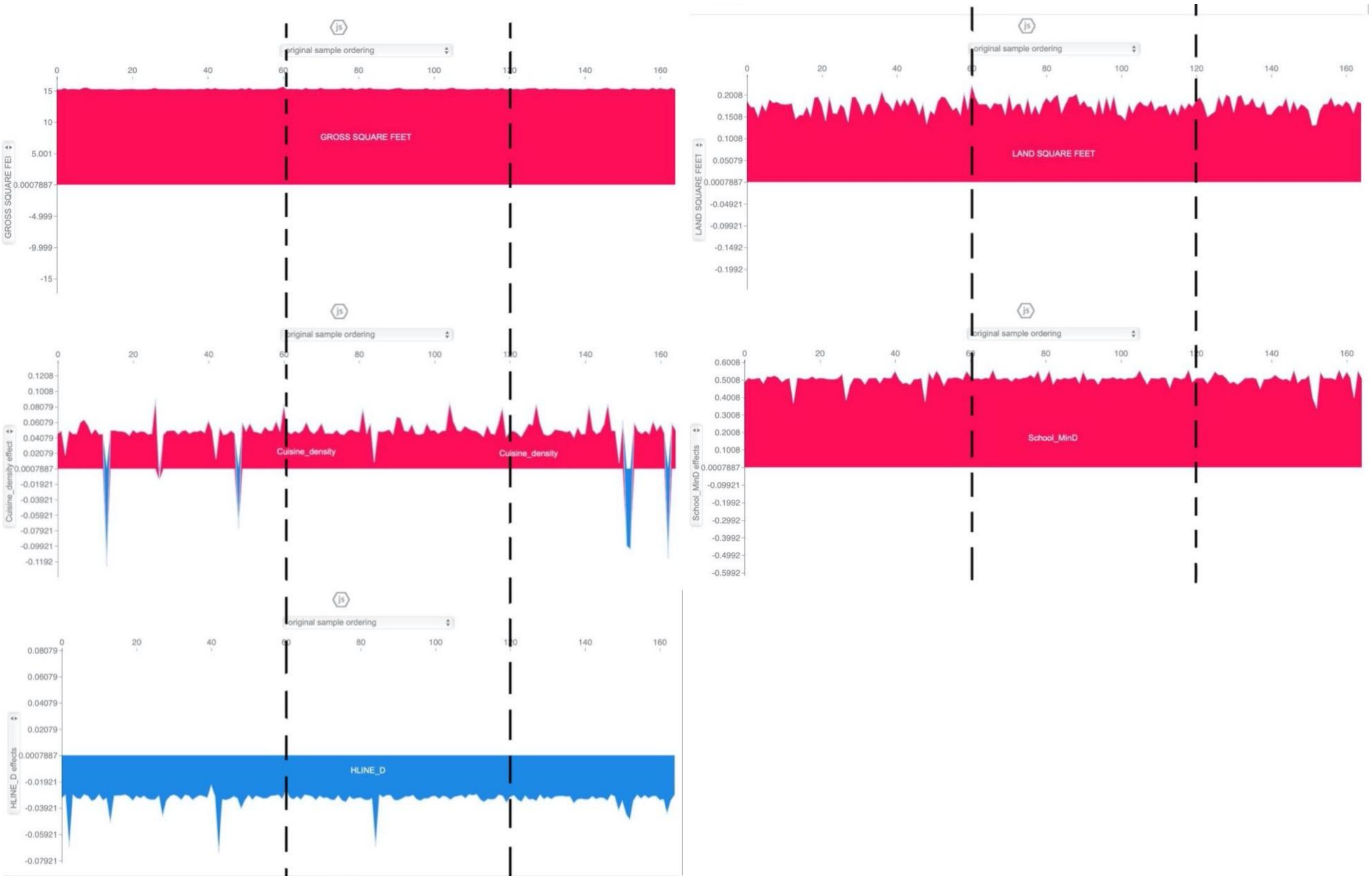

Figure 3: Feature importance explanation using SHAP for 5 most important features from RF model

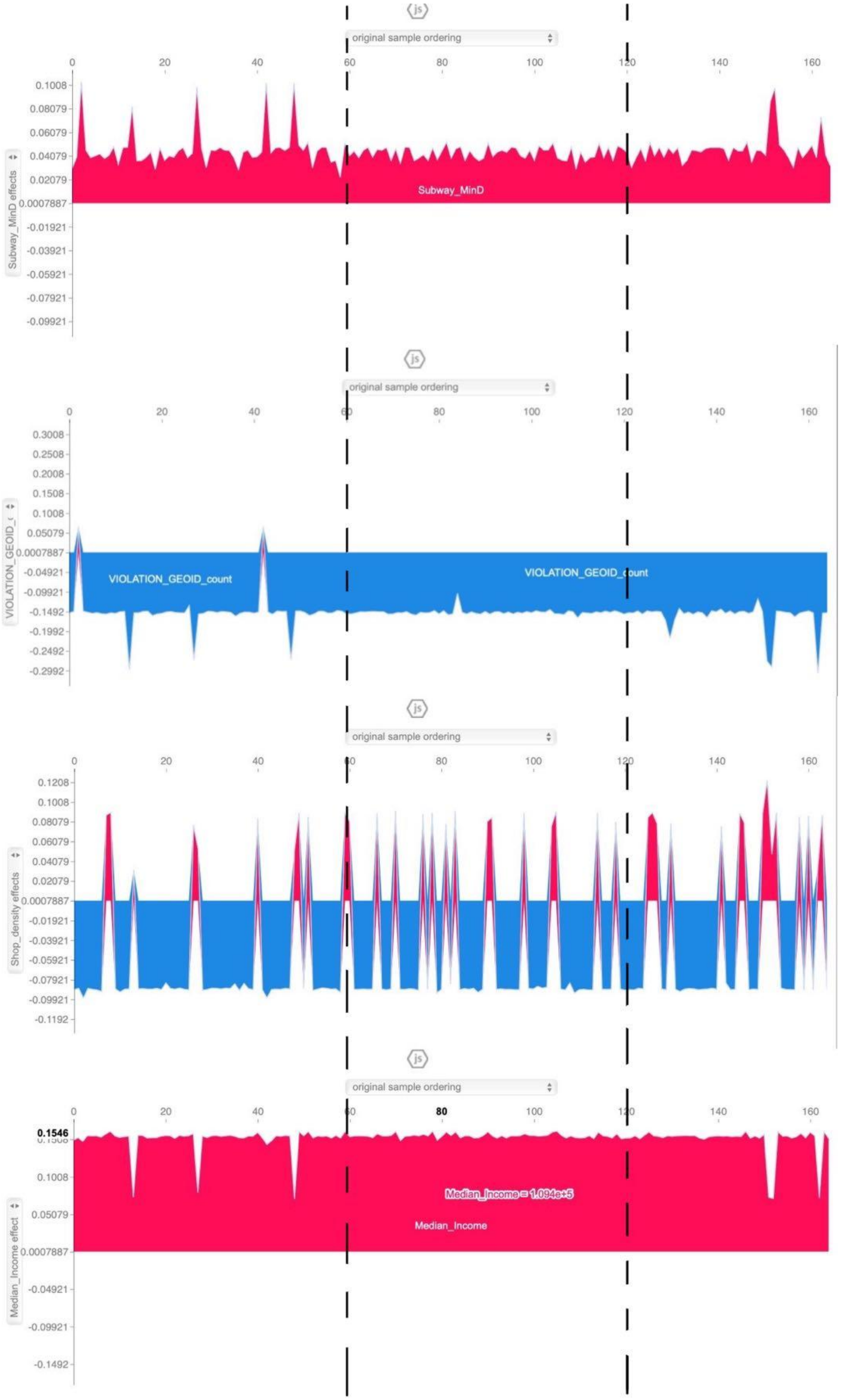

Figure 4: Features with obvious importance changes during COVID with SHAP importance explanation

Author Contributions

Cici Chen:
Data hunting, introduction, literature and variable review, discussion, data appendix

Jingjing Ge:
Data hunting, aggregation and cleaning, QGIS, abstract, data appendix and conclusion, video

Yichen Guo:
Data cleaning, SHAP model, literature review, method, results, conclusion, discussion, video

Jiale Li:
Data aggregation and cleaning, regression models running, feature and data selection

Tao Liang:
Data hunting, aggregation and cleaning, final data merge, references, data appendix

Wangtianhan Pang:
Machine Learning methods explanation, variable review, citation check

Xinyu Xu:
Data merging, abstract, literature review and data appendix, PowerPoint presentation, video

References

Bakshi, Chaya. “Random Forest Regression.” Medium. Level Up Coding, June 9, 2020. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84.

Brownlee, Jason. “Essence of Bootstrap Aggregation Ensembles.” Machine Learning Mastery, October 21, 2021. https://machinelearningmastery.com/essence-of-bootstrap-aggregation-ensembles/.

Chakure, Afroz. “Random Forest Classification and Its Implementation.” Medium. The Startup, November 6, 2020. https://medium.com/swlh/random-forest-classification-and-its-implementation-d5d840dbead0.

Gao, Guangliang, Zhifeng Bao, Jie Cao, A. K. Qin, Timos Sellis, and Zhiang Wu. “Location-Centered House Price Prediction: A Multi-Task Learning Approach.” arXiv.org, January 7, 2019. https://arxiv.org/abs/1901.01774.

Grybauskas, Andrius, Vaida Pilinkienė, and Alina Stundžienė. “Predictive Analytics Using Big Data for the Real Estate Market during the COVID-19 Pandemic.” Journal of Big Data 8, no. 1 (2021). https://doi.org/10.1186/s40537-021-00476-0.

Kumar, Ajitesh. “Mean Squared Error or r-Squared - Which One to Use?” Data Analytics, September 30, 2020. https://vitalflux.com/mean-square-error-r-squared-which-one-to-use/.

Transport for London. “Healthy Streets for London - Transport for London.” Accessed December 13, 2021. https://content.tfl.gov.uk/healthy-streets-for-london.pdf.

Sirmans, Stacy, David Macpherson, and Emily Zietz. “The Composition of Hedonic Pricing Models.” Journal of Real Estate Literature 13, no. 1 (2005): 1–44. https://doi.org/10.1080/10835547.2005.12090154.

Su, Yichen, and Sitian Liu. “The Impact of the COVID-19 Pandemic on the Demand for Density: Evidence from the U.S. Housing Market.” SSRN Electronic Journal, 2020. https://doi.org/10.2139/ssrn.3661052.

The river developers. “Knnregressor.” River, 2019. https://riverml.xyz/dev/api/neighbors/KNNRegressor/.

XGBoost developers. “XGBoost Documentation.” XGBoost Documentation - xgboost 1.5.1 documentation. Accessed December 13, 2021. https://xgboost.readthedocs.io/en/stable/.

Appendixes:

Table 3: all the features included in the initial model running

Data	Definition	Data Sources+, Year
Geographic Data		
Census Tract 2010	2010 Census Tracts boundary of New York City.	Department of City Planning, 2010
PLUTO	Primary Land Use Tax Lot Output. Geographic point data at the tax lot level.	Department of City Planning, September 2021
Housing Data		
SALE PRICE	Total house sale price	Rolling sales, Department of Finance, 2019,2020,2021
YEAR BUILT	House building time	Rolling sales, Department of Finance, 2019,2020,2021
LAND SQUARE FEET	Land area of the house	Rolling sales, Department of Finance, 2019,2020,2021
GROSS SQUARE FEET	Floor area of the house	Rolling sales, Department of Finance, 2019,2020,2021
BUILDING CLASS AT TIME OF SALE_A	One Family Dwellings	Rolling sales, Department of Finance, 2019,2020,2021
BUILDING CLASS AT TIME OF SALE_B	Two Family Dwellings	Rolling sales, Department of Finance, 2019,2020,2021
BUILDING CLASS AT TIME OF SALE_C	Walk Up Apartments	Rolling sales, Department of Finance, 2019,2020,2021
BUILDING CLASS AT TIME OF SALE_D	Elevator Apartments	Rolling sales, Department of Finance, 2019,2020,2021
BUILDING CLASS AT TIME OF SALE_S	Residence - Multiple Use	Rolling sales, Department of Finance, 2019,2020,2021
BOROUGH_1	Value=1, if the house is located in the Manhattan	Rolling sales, Department of Finance, 2019,2020,2021
BOROUGH_2	Value=1, if the house is located in the Bronx	Rolling sales, Department of Finance, 2019,2020,2021
BOROUGH_3	Value=1, if the house is located in the Brooklyn	Rolling sales, Department of Finance, 2019,2020,2021
BOROUGH_4	Value=1, if the house is located in the Queen	Rolling sales, Department of Finance, 2019,2020,2021
BOROUGH_5	Value=1, if the house is located in the Staten Island	Rolling sales, Department of Finance, 2019,2020,2021
Socioeconomic		
POP_19	Total population in the census tract	ACS 5-Year Estimates, 2019
POP_19_density	Population density in the census tract	ACS 5-Year Estimates, 2019
UNDER_18	Percentage of population under age 18	ACS 5-Year Estimates, 2019
OVER_65	Percentage of population over age 65	ACS 5-Year Estimates, 2019
WHITE	Percent of Whites in the total population	ACS 5-Year Estimates, 2019
BLACK	Percent of Blacks in the total population	ACS 5-Year Estimates, 2019

ASIAN	Percent of Asians in the total population	ACS 5-Year Estimates, 2019
Median_Income	Median Household Income	ACS 5-Year Estimates, 2019
FELONY_GEOID_count	Level of offense in crime reported: Felony	NYC Open Data, 2019,2020,2021
MISDEMEANOR_GEOID_count	Level of offense in crime reported: Misdemeanor	NYC Open Data, 2019,2020, 2021
VIOLATION_GEOID_count	Level of offense in crime reported: Violation	NYC Open Data, 2019,2020,2021
Built Environment		
BusStop_Count	Number of bus stops in the census tract	NYC Open Data, 2021
Tree_density'	Density of trees in the census tract	NYC Open Data, 2021
Shop_density	Density of shops in the census tract	Open Street Map
Cuisine_density	Density of Restaurants in the census tract	Open Street Map
Subway_MinD	Distance of the house to the nearest subway station	NYC Open Data, 2019
School_MinD	Distance of the house to the nearest school	NYC Open Data, 2019
PARK_MIND	Distance of the house to the nearest big park	Census Tract 2010, 2010
KINDER_MIND	Distance of the house to the nearest kindergarten	Open Street Map
ARTCENTRE_MIND	Distance of the house to the nearest art centre	Open Street Map
MET_D	Distance of the house to The Metropolitan Museum of Art	Open Street Map
SOL_D	Distance of the house to The Statue of Liberty	Open Street Map
EMPIRE_D	Distance of the house to The Empire State Building	Open Street Map
CPARK_D	Distance of the house to Central Park	Open Street Map
TSQUARE_D	Distance of the house to Times Square	Open Street Map
BBRIDGE_D	Distance of the house to Brooklyn Bridge	Open Street Map
MSQUARE_D	Distance of the house to Madison Square Garden	Open Street Map
TRUMP_D	Distance of the house to Trump Tower	Open Street Map
STPC_D	Distance of the house to St Patrick's Cathedral	Open Street Map
WTCT_D	Distance of the house to World Trade Center Station	Open Street Map
HLINE_D	Distance of the house to The High Line Park	Open Street Map
colision_count	The number of motor vehicle collisions (crashes)	NYC Open Data, 2019,2020,2021
noise_norm	The number of noise complaints came from 311	NYC Open Data, 2019,2020,2021