Capturing Gentrification Early: Analyzing Alternative Data Sources for Predicting Neighborhood Changes

Cornell University
ORIE 4741 Final Project Report
Liye Zhong(lz246), Lu Li(ll565), Rong Tan(rt389)
December 10th, 2019

Abstract
Rapid changes in neighborhoods such as gentrification and displacements often impact its original dwellers negatively. The NYC Department of City Planning wants to obtain information from alternative data sets unlike the traditional demographic survey data to sense the trend of neighborhood changes early-on. We found that for each zip code area in New York City, its location (longitude and latitude), complaints about the indoor environments filed by its dwellers, and restaurant sanitary conditions in the area is predictive of rent price changes in an area.

Background
New York City neighborhoods have experienced significant changes in the past decade and many citizens struggle to afford housing as gentrification continue to disrupt their communities. Gentrification is a process of changing the characters of a neighborhood through the influx of more affluent residents and businesses to a low-income urban neighborhood. One key indicator is increase in rent prices at a feverish pace. This often comes with the costs of forcing out the low-income community and political conflicts accentuated with race, class and cultural conflicts. While the U.S. Census Bureau survey data present the neighborhood demographic and economic conditions quite well, it often lags in time and describes the after effect of gentrification and when the impacts already occurred. The data sets used in this study are updated monthly or quarterly rather than annually. Extracting the hidden information in these data with statistical models can help policy makers predict the likelihood of neighborhood changes in rent prices and/or crime rates.

Data Description
1. *Zillow New York Home Prices & Values [2018-2019]*: First the data is divided by property type. Then for each zipcode: current average home value($), year over year home value change(%), current Zillow Rent Index ($), year over year index change(%) etc[1]. This dataset will be referred to as Dataset 1: 'Home Prices'

From NYC Open Data:
2. *DOHMH Indoor Environmental Complaints*: complains to hotline 311. Each record is a single complaint. It contains information on the incident address (zip code, numeric), date receive (string), nature of the complaint (text), and a short description (text). Updated daily[2]. This dataset will be referred to as Dataset 2: 'Indoor Complaints'.

3. *DOHMH New York City Restaurant Inspection Results*: Restaurant inspection conducted by the Health Department, which include checks on food handling compliance, food temperature, hygiene, vermin control. It records the restaurant location (numeric), nature of violations (text), and score (the higher, the more violations the restaurant has). Updated daily[3]. This dataset will be referred to as Dataset 3: Restaurant Inspection'.

4. *Sidewalk Café Licenses and Applications*: This dataset features detailed information about sidewalk café license applications and, if applicable, issued licenses. It contains the applicant café location (zipcode), business name (text), license status (active/inactive), application date (string). Updated weekly[4]. This dataset will be referred to as Dataset 4: Restaurant Inspection'.

III. Initial Data Cleaning

3.1 Chosen Features & Assumptions

From Dataset 1, we chose Zillow Rent Index as the number that our model would like to predict (variable Y). We chose Zillow Rent Index rather than other features in the dataset because the index combines the prices from across various housing types (single family apartments, condos, studio, etc.). The dataset already records the index percentage change from 2018-2019 for each zip

code in NYC and has no missing value and thus we do not need to do further cleaning. From Dataset 2, we want to see whether the number of complaints changed year over year since we assume the number of complaints will increase as an area gentrifies. We recorded the total number of complaints for each zip code and their changes year over year. The changes are stored in three columns: “15-16 change”, “16-17 change”, and “17-18 change”, in percentages. From dataset 3, we assume that a gentrifying area sees improvements in restaurant qualities, which will reflect in the data as a decrease of their review scores (the more violations a restaurant has, the higher the score is). Thus, we recorded the score changes year over year and stored them in three separate columns. From dataset 4, we assumed that a gentrifying area will see an increase in café license applications. Thus, we counted the number of applications in each Zipcode and recorded their changes year over year.

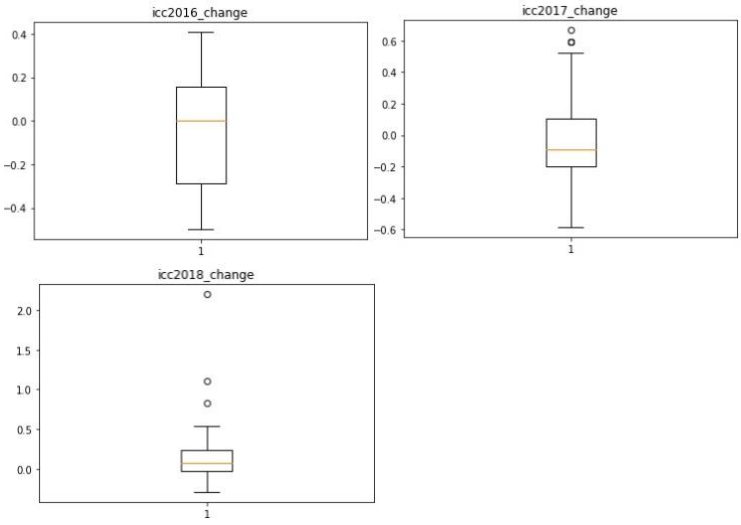
3.2 Missing Values and Zero Changes

For dataset 2 and 3, some zip codes have year over year changes are 0. Some of these zero entries are due to missing data, and some have actual value 0. Since these two datasets have only a small percentage of missing data, we replaced these entries with column mean. For dataset 4, however, there are large number of zero entries. This is due to the overall low number of café license applications. We think this may be an interesting feature for prediction but do not think it is reasonable to replace these entries with column mean since the percentage of 0’s is quite large. We want to see how the dataset performs in the learning model and decide whether we eventually want to keep it as a feature in the prediction model.

IV. Data Visualization

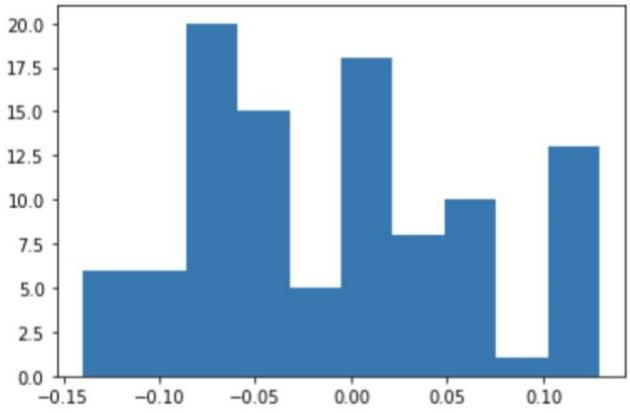
For our dataset, we created several data visualizations so that we can analyze their distribution and provide better models based on them.

The first set of visualization we created is for the indoor inspection score change, which contains the score change from 2015 to 2016, 2016 to 2017, and 2017 to 2018. The visualization is as follows:



As we can see from the three visualizations, all the boxplots are not greatly skewed, while the indoor inspection score change from 2017 to 2018 is having three outliers. But that should not be a big problem, considering that our dataset has more than 100 samples. Therefore, we are confident that these three crucial feature is suitable and validfor predicting the Zillow housing price change.

Another data we created visualization for is the data we want to predict, the Zillow housing price change. The histogram of the Zillow housing price change is as below:



As we can see from the graph above, the data is randomly distributed in the range of -0.15 to 0.15. And the labels we want to predict is not biased or following any pre-existing distribution.

Therefore, we can say that if we successfully trained our model to achieve a high accuracy rate on

the test set, our model is trained based on an accurate reflection of actual housing price change in New York City.

V. Learning Models

To build our learning models, we need to decide the feature space and the label we want to predict. From the previous data sets, we decided to use the following features to predict the housing price change for each zip code:

- average_review, which stands for the mean of restaurant review rating from the NYC 2018 restaurant week rating in each zip code
- Longitude, which stands for the longitude of the area in each zip code, which indicates the relative location of the center of the zip code.
- Latitude, which stands for the latitude of the area in each zip code, which indicates the relative location of the center of the zip code.
- icc2016_change, which stands for the indoor inspection score change from 2015 to 2016 in each zip code
- icc2017_change, which stands for the indoor inspection score change from 2016 to 2017 in each zip code
- icc2018_change, which stands for the indoor inspection score change from 2017 to 2018 in each zip code
- 2016_change, which stands for the NYC Restaurant Inspection score change from 2016 to 2017 in each zip code
- 2017_change, which stands for the NYC Restaurant Inspection score change from 2017 to 2018 in each zip code
- 2018_change, which stands for the NYC Restaurant Inspection score change from 2018 to 2019 in each zip code

The label we want to predict is the PriceChange, which stands for the Zillow price change from 2018 to 2019 in each zip code. It would be extremely hard to predict the exact percentage change of housing prices since we are predicting the year-over-year change for 102 zip codes.

We trained all our model using 75% of the dataset as training data based on the random train/test split on our existing dataset and predict on the 25% of the dataset to get the final accuracy of each prediction model. As the dataset we have has 102 entries and 9 dimensions, the train dataset has 76 entries while the test dataset has 26 entries. We did not choose a normal 80%-20% split because the limited amount of datapoints and points that need to be predicted.

5.1 Linear Regression

The first prediction method we tried is Linear Regression. We chose this model because it's intuitively the simplest model that will allow us to directly look at the relationships between the x features and y. Linear regression assumes that all our x features are equally important in predicting the result. Linear Regression is bad at predicting the categorical labels, but good at predicting the value regression with directly correlated factors. We used linear regression model to predict the numerical percentage change of Zillow housing index. With the Linear regression model trained on the training set and predicted the test set, we have the coefficient of determination R^2 of the prediction to be 0.499 for the train set, and 0.380 for the test set. As the coefficient of determination for both sets are not ideal, we can conclude that the dataset is not suitable to be predicted by a linear regression model, and the linear regression model is definitely underfitting on our dataset. Thus, we need more complex models to increase accuracy.

5.2 Ridge Regression

The second prediction method we tried is Ridge Regression. As Linear Regression doesn't differentiate "important" from "less-important" predictors in a model, and may lead to overfitting the model. Ridge regression is an improvement to linear regression as it adds a regularization term at the end of the formula. With the help of the regularization term, ridge regression shrinks coefficients towards 0 and it shrinks more in the direction of the smallest singular values of X.

The Minimization objective of Ridge Regression is as follows:

*Sum of Squared Errors + α * (sum of square of coefficients)*

From our observation and initial analyses of the housing price change, the price changes are very volatile and have a lot of variation from 2018-2019. In order to make our prediction more interpretable, we decided to mark our prediction for the test set as a correct prediction if its absolute value is less than 0.05 from the true housing price change (error margin is 0.05), which in formula can be written as

Abs(prediction - test_label) < 0.05

We tried Ridge Regression with two types of parameters: alpha as 0.01 and alpha as 100. L2 norm term in ridge regression is weighted by the regularization parameter alpha If the alpha value is 0, it means that it is just an Ordinary Least Squares Regression model. So, the larger is the alpha, the higher is the smoothness constraint. The smaller the value of alpha, the higher would be the magnitude of the coefficients.

For Ridge Regression with alpha as 0.01, our prediction model achieved an accuracy of 0.731. On the other hand, for Ridge Regression with alpha as 100, our prediction achieved an accuracy of 0.5. The Ridge Regression with alpha as 0.01 serves our purpose of the study with considerable confidence, but we still want to make the model more complex and try to yield a better result when predicting the housing price change for each zip code.

5.3 Random Forest Classifier

The Zillow housing price changes we want to predict in each zip code are also fluctuating in a small scale in an unpredictive way. No matter which loss function we use, we are going to see a large amount of discrepancy between prediction and actual number. Therefore, we use feature engineering for this feature. To make our prediction easier and the final result more interpretable, we decided to divide our data into the following data ranges based on the data distribution we discovered: [negative infinity, -0.08, -0.06, -0.04, -0.02, 0, 0.02, 0.04, 0.06, 0.08, infinity]

The third prediction method we tried is Random Forest Classifier because the Random Forest Classifier provides a reliable feature importance estimate and its predictive performance can compete with the best supervised learning algorithms at present. Compared to linear regression and ridge regression, Random Forest Classifier is advantaged at predicting categorical data, which fits our objective. In fact, from a renter's or policy maker's perspective, we do not necessarily need to know the exact percentage change of housing prices. A general range of changes should suffice the purpose of preparing a budget for next year or making corresponding urban planning schedules for the neighborhood.

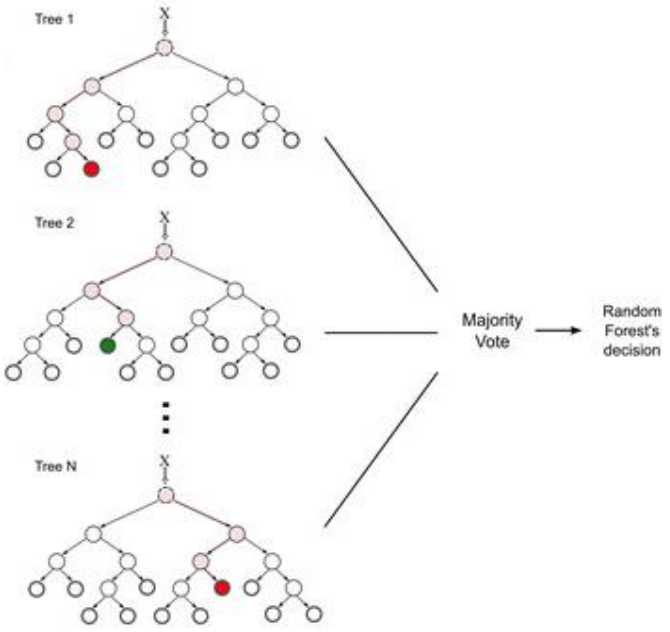


Figure 1

The Random Forest Classifier works is by creating random decision trees with various features. The algorithm splits the tree in the way that minimizes prediction errors (the entropy) so that minimum about of splits can be used in one tree. The prediction is based on the majority decision. For example, if out of 20 trees, 2/3 categorizes the change to category 3, 1/6 categorize it to category 1 and the rest categorize it to category 2, the

algorithm will spit out category 3 as the final prediction.

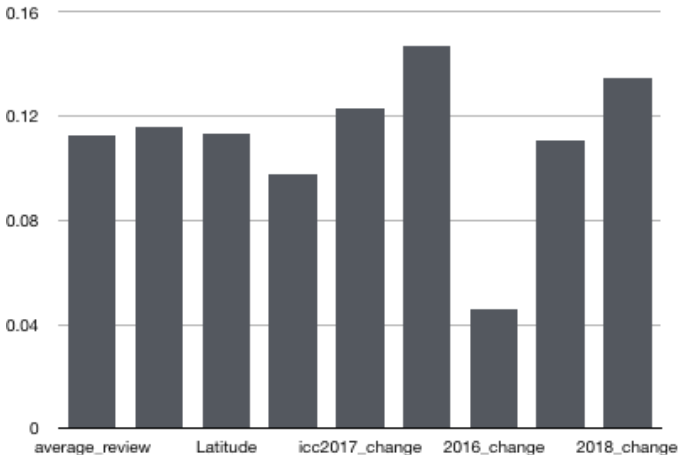
Because the Random Forest Classifier algorithm can only be used to predict labels, not numbers, we are predicting the Zillow housing price change range, which is mentioned in the beginning of the section, not the Zillow housing price change in double numbers.

After we run the random forest classifier with different parameter combinations, we found out that with parameters n_estimators as 300, max_depth as 10, and random_state as 10 will yield the best result, which has achieved the accuracy of 0.952 for the test set.
The feature importance list we discovered from this model is [0.11269174 0.11609556 0.11310713 0.09740819 0.12320665 0.14707087 0.0455877 0.1103641 0.13446806], which explains that some features, such as icc2016_change and 2016_change are not so influential for predicting the housing price change range.

From the three models we tested, we decided to move forward with the Random Forest Classifier, which yields a high accuracy of 0.952 to predict the housing price change range.

VI. Interpreting the model and results

Out of the three models we tested, the Random Forest Classifier showed the best performance. Here we want to do a further analysis of the model.
From the Random Forest Classifier, we generated a graph of the feature importance, so that we can value the importance of the features in an intuitive way.



As we can see from the graph, the housing price change is relying on the parameters in different scales.
For the two time series data, the indoor inspection score changes and NYC Restaurant Inspection score change, the closer they are to date 2018, the more important they are to predict the housing price change. This can be explained by common sense, as closer they are to date 2018, the more influential they will be to affect the housing price change from 2018 to 2019.

Which surprised us is that the longitude and the latitude plays an important role in the prediction model. To explain these two features' importance in predicting the housing price change, we would like to say that the longitude and latitude, which is derived from area of the zip code in our dataset, actually has hidden correlation with the housing price change. For example, the city area in high latitude and east longitude in the NYC area would be more likely to see a rising housing price change, because the New York City government is actively calling for gentrification process in the previously ungentrified area, which are mostly in Bronx and Brooklyn.

And we also discovered that the average_review factor, which stands for the mean of restaurant review rating from the NYC 2018 restaurant week rating in each zip code, also has similar amount of weight as the longitude and in the prediction model. This factor is affecting the housing price in an expected and explainable way. We would like to reason that if the restaurant review rating from the NYC 2018 restaurant week in a certain zip code is relatively higher, the gentrification process or the

neighborhood condition of that certain zip code us higher, thus the housing price of that are would encounter consecutive increase over the year.

VII. Future Steps and Conclusion

From the three models, we are happy to see the high accuracy in prediction from Random Forest Classifier given the limited dataset scopes. Because our model is only training on a dataset of size of 102, the model we yield is not suitable for predicting the real-life problems as it would be overfitting in a large scale. Therefore, more data is required if we want to improve our random forest classifier model.

We can further test the robustness of the model by going backwards in time. For example, we can obtain 2012-2015 features and use the same model to predict 2015-2016 changes in housing price to evaluate the accuracy of our model. We can also perform cross validation on the new dataset.

Besides finding more data, we can also add more features to the model and train more persuasive models. Every summer, NYC host restaurant week when merchants lure customers to visit their restaurant through drastic discounts. And we get to collect extra data through a short amount of time when the customers leave reviews after they actively participate in the event. Then we can correlate such data to the restaurant inspection dataset. In the current model, we didn't get to use café license data as intended due to too many entries of 0's. We would be interested in looking at alternative data sources that reflect similar information. In addition, we also did some initial analysis on income levels of different zip codes in NYC, which was not reflected in our final model. We would be interested in using the current features to predict the changes in median income of each area and see how the model performs.

References

Figure 1: Image obtained from: Machado, G., Mendoza, M. R., and Corbellini, L. G. (2015). *What variables are important in predicting bovine viral diarrhea virus? A random forest approach.* Vet Res Veterinary Research, 46(1). doi:10.1186/s13567-015-0219-7

[1] https://www.zillow.com/new-york-ny/home-values/

[2] https://data.cityofnewyork.us/Health/DOHMH-Indoor-Environmental-Complaints/9jgj-bmct

[3] https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

[4] https://data.cityofnewyork.us/Business/Sidewalk-Caf-Licenses-and-Applications/qcdj-rwhu