A Study on observing and predicting the Rental House Pricing in Manhattan

Qiuyao Liu(ql459@nyu.edu), Minqi Lu(ml4922@nyu.edu), Jiawen Liang(jl9760@nyu.edu), Guanjia Wang(gw1054@nyu.edu)

Machine learning for cities

May, 8th 2019

Abstract

The expense of buying or renting the property to live could be one of the largest portions throughout people's lives. Research has shown that the millennials are spending $93,000 on rent by age 30 [1]. Renting the right apartment at the right time can profoundly benefit the renter. In this project, our goal is to build a model that can accurately predict the price of rental apartments. Our features include the data about the apartment rooms, such as the number of bedrooms and bathrooms, as well as the apartment building itself, such as the amenities. After a series of modeling and experimentation, we discovered that the random forest regressor has the best performance and can explain the most amount of variance when predicting the rental price.

Introduction

The price and the affordability of apartments and houses are very important to us. Many people spend most of their savings on renting and purchasing a home. Due to the high cost of buying a house or apartment, a lot of people choose to rent. So, before making the

decision to rent a home, it is very important for people to get a sense of the rent of the potential home and the trend of the renting market.

On the other hand, while many individuals and governments are working on predicting the house price since it has a big influence on the GDP, the rent price of apartments also substantially affects the economy. Because the housing rental market is strongly correlated with the sales housing market, both in the short term and long term. Also, rent is considered as consumption, which is calculated in gross domestic product. There is a strong correlation between rent and GDP. So, it is only reasonable for the local government to not only focus on the housing price but also keep an eye on the rental housing market.

Therefore, the ability to accurately predict the trend of the local rental housing market is important to the residents and the government. For local government, the accurate prediction of the rental housing market can help them to make a plan accordingly, such as setting pricing regulation. For local residents, for better understanding of the trend and the variance of the rental housing market would provide more time and choices.

However, there are so many variables we have to think about in terms of rent prediction, the floor plan, the neighborhood, the amenity, etc. The good news is, as the era of big data maturing and machine learning becoming more popular, we now have massive data and powerful techniques to build models and make predictions. There are a lot of studies that are relevant to this subject, but most of them were focusing on the features that directly related to the house, like the number of bedrooms, the floor space, and have not taken the quality of neighborhood into account, for example, the rating score of nearby restaurant and coffee shop. In this project, we combine two datasets together, one is the data of apartments' information in New York City, and the other one is the data of cleaning inspection of restaurants in New York City. In addition, our housing data contains a lot of details that most

of the data do not have, such as different kinds of the doorman (full time or half time), and kinds of allowed pets (cats only, dogs only, or both). We are expecting that by adding more features in the model, we will increase the accuracy of the prediction. Our goal is to find an approach that can better predict rent and help local residents, real estate companies, and local government to take action accordingly.

Related Work

Since the rental price of houses is continuous, the researcher would choose to use the regression algorithm to build models. For example, previous research has used linear regression, Lasso regression, and Ridge regression to build their models[2]. The research, focusing on sales property, found out that among important features like room number and floor space, condition of walls and stairways also has fairly high feature importance. In addition, the initial R-squared of their model was only 0.25. Then they expanded features from 41 to 55. They also deleted some features after performing Lasso. The R-squared of the final model increased to 0.55

Another research's result suggests that before training the Support Vector Regressor (SVR) model, using PCA to perform feature extraction, or using Lasso, Lidge, RFE, random forest to perform feature selection, make no difference on the SVR model's performance[3]. It also suggests that feature reduction, parameter tuning, and log transformation can really improve the SVR model's performance.

There are also some projects that are thinking outside the box. A team has come up with a method to use spatial modeling to predict apartment rent[4]. They built the model on the point level, without aggregating any rental properties. They analyzed the rental properties individually so that they could achieve the highest resolution possible. They built several

models for each market in their study, which gave them the ability to analyze each market individually, as well as to compare them with each other.

For our project, besides linear regression and SVR, we also use random forest regressor and K-means clustering in order to get deeper insights into our data, as well as better predict the result.

Data and Preprocessing

The building data we are using for this machine learning project are provided by Easyrent, the leading real estate company in the rental market. The original dataset consists of two parts: building info and unit info. The building info contains six features, including the building ID, year of the building, address, type of the building(apartment, co-op, etc.), amenity, and nearby transportations. The amenity of the building includes the type of doorman(full time or half time), elevator, cats and dogs allowed, cats only - no dogs, and pets allowed. The unit info is a separate dataset under each building. For each unit, it contains five features, includes the unit ID, number of bedrooms, the number of bathrooms, floor space, and the unit status. The unit status contains the detail about the time that the unit was listed or off-marketed. We merge two datasets together by building ID, and only keep the features we need. We take out the transportation feature since all of the buildings have four MTA stations in the neighbor and multiple choices on the trains. Then we take out the amenities columns and convert them into dummy/indicator variables. We also clean and separate the strings under the status column as status (rent, off-market), date and pricing.

Besides, we want to see if rental housing prices could have any relationship with the restaurant in the neighborhood. We extract the restaurant inspection score from the New York City Restaurant Inspection Result, providing by DOHMH. Last but not least, we geocode our

data using Google Maps API so that the data would contain longitude and latitude that match with the address of each building. Geocoding our data enables us to understand and visualize the pattern and distribution of the rental housing on Manhattan.

Context and exploratory analysis

We look at the distribution of the building units' price, and the histogram suggests a long right tail. Interquartile range (IQR) is used to remove outliers, we only want to keep values between a lower fence and an upper fence.

$$Lower\ Fence\ =\ First\ Quartile\ (Q1)\ -\ 1.5\ \times IQR$$

$$Upper\ Fence\ =\ Third\ Quartile\ (Q3)\ +\ 1.5\ \times IQR$$

After removing the outliers, the distribution looks better although there is still a bit of long tail.

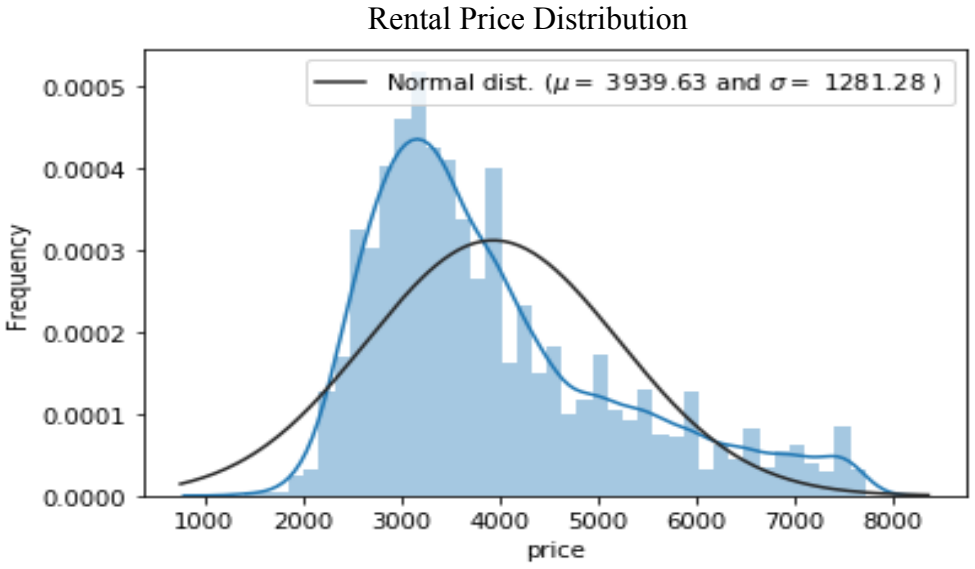

Fig. 1 the rental price distribution

We then examine the correlations among our nineteen features: price, the number of bedrooms, the number of bathrooms, area (square feet), the unit status, stories, age, score, co-op, condo, rental building, single-family home, cats only - no dogs, part-time doorman,

doorman, elevator, dogs-allowed, pets-allowed and full-time doorman. From the correlation matrix heatmap, we notice that the number of bedrooms, number of bathrooms and area are the three features with a high positive correlation with the rental price.

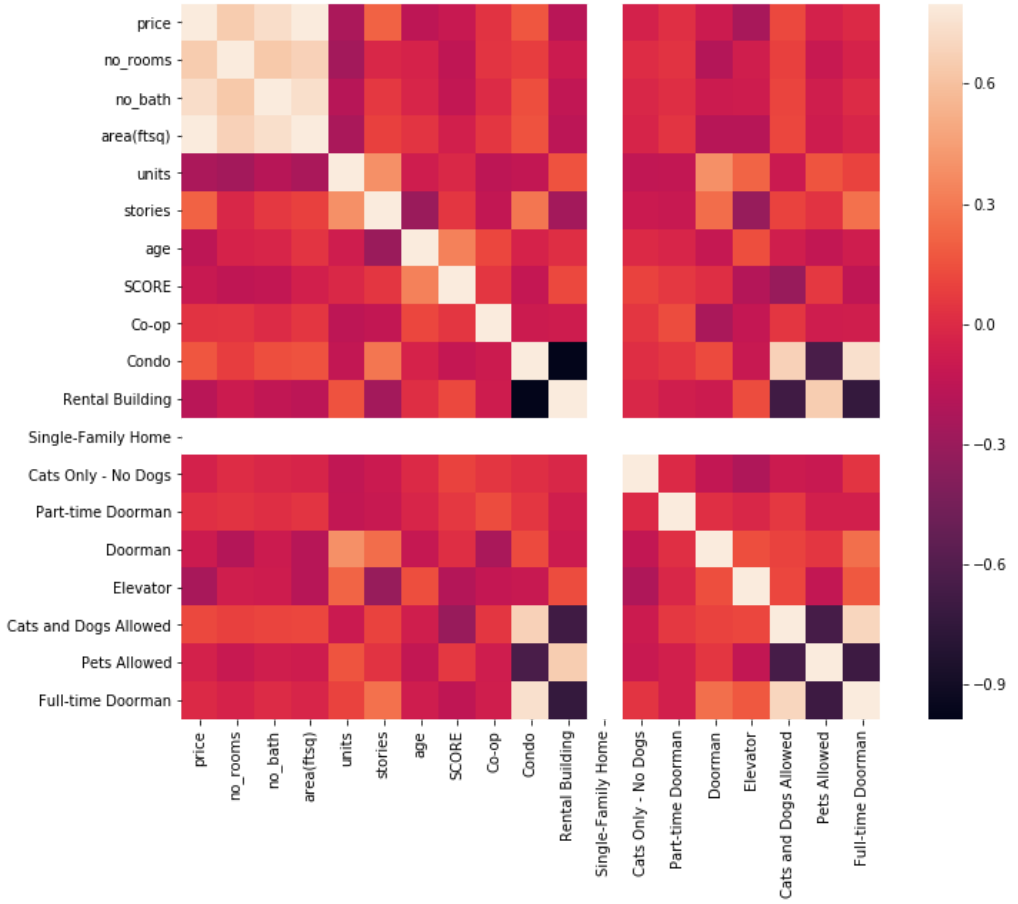

Fig. 2 the heatmap of all the feature

Methodology and Modeling

Based on the correlation matrix heatmap, we choose to include the ten most correlated features into our models. Also, we split our data into two parts: training data (80%) and test data (20%). We consider the linear regression model initially, and the in-sample R-squared is 0.7423, which is quite good. The out-of-sample R-squared also shows quite similar R-squared, 0.7407.

By using LASSO regression, we choose the most significant features among those ten features while reducing the coefficients of others to zero to see how the model fit. The results show in-sample R-squared is 0.7420 and the out-of-sample R-squared is 0.7417. There is no significant improvement in R-squared for these two models.

Linear supporting vector regression is the third model we are used. The in-sample and out-of-sample classification accuracies are 0.7379 and 0.7363 respectively. Finally, the random forest regressor is used and gives the best in-sample classification accuracy, 0.9443. Also, the out-of-sample accuracy also performs well which is 0.8931. In general, a random forest shows the best performance among the four models.

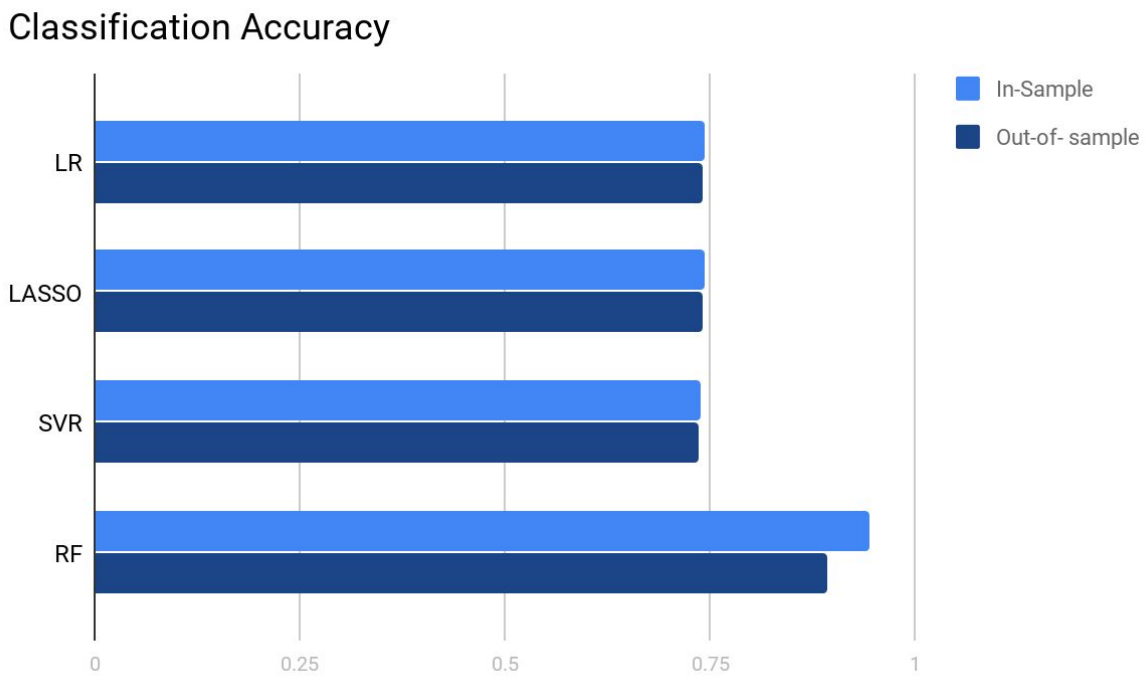

Fig. 3 The histogram of accuracy comparison on LR, LASSO, SVR, Random Forest model

Clustering Observation

We first do the feature observation through the K-means and then we clustering the rental housing on pricing. For choosing the number of clusters K, we use the Elbow method. First, we separately observe the relationship between the rooms number and the bathrooms number, the elbow point lies near k = 3. Through Fig. 4, we see that the trend of bathrooms increases as the number of rooms is increasing. This is instinct knowledge, as well as observed and used in the regression. Moreover, we perform clustering based on location coordination. The elbow point lies near k = 4. We see that k means divide Manhattan into multiple features.

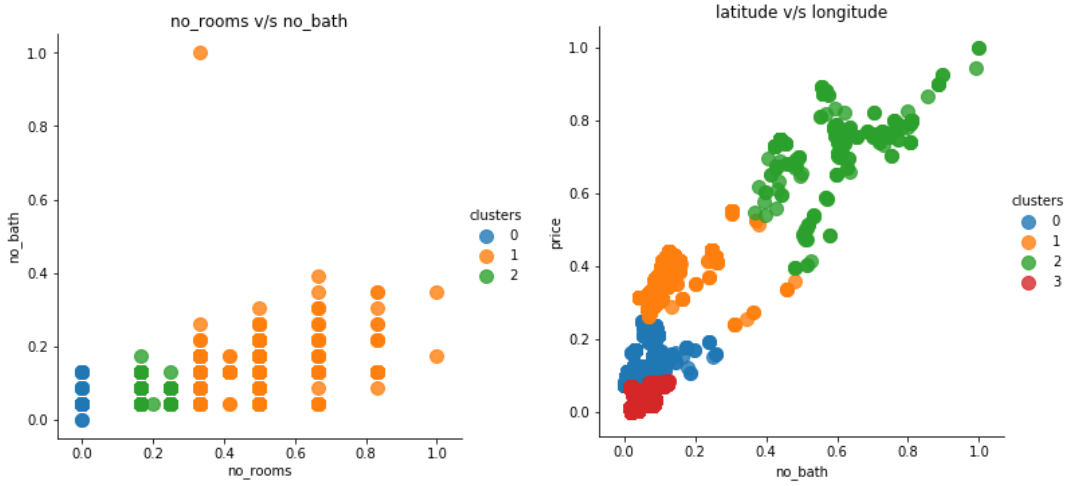

Fig. 4&5: Features Observed: rooms number and bathrooms number; latitude and longitude

K-means clustering is also used to see "Do different rent price display different trends over space?" Firstly, by looking for an elbow in the within-cluster SSE we choose the number of clusters as three and applying K-means in Python built-in package sklearn. Here is the basic summary of our three clusters:

	Cluster 0	Cluster 1	Cluster 2
Count	7757	4480	1371
Mean	$2965.03	$4705.58	$7750.28

Minimum	$1000	$3850	$6245
Maximum	$3848	$6225	$45000

The average rent prices increase from cluster 0 to cluster 2 and cluster 2 contains the highest rental price in Manhattan and with the smallest portion comparing with the other two clusters. And we can clearly see the positive correlation between price and bathroom/bedroom numbers. Moreover, we observe that cluster 0 mainly contains a studio and one bedroom while cluster 2 mainly contains one bedroom and two bedrooms apartment. Cluster 1 has a large portion of two bedrooms apartment.

To better understand the distribution of the rental housing in Manhattan, we plot the rental housing during each cluster on the NYC community Districts level through encoding the longitude and latitude with Google map API and merging with PUMA shapefile. Through Fig.6, we could observe that the lower side of Manhattan contains most of the higher pricing housings compared with other neighborhoods. Besides, the lower side contains the greatest range of pricing and the largest portion of rental housings.

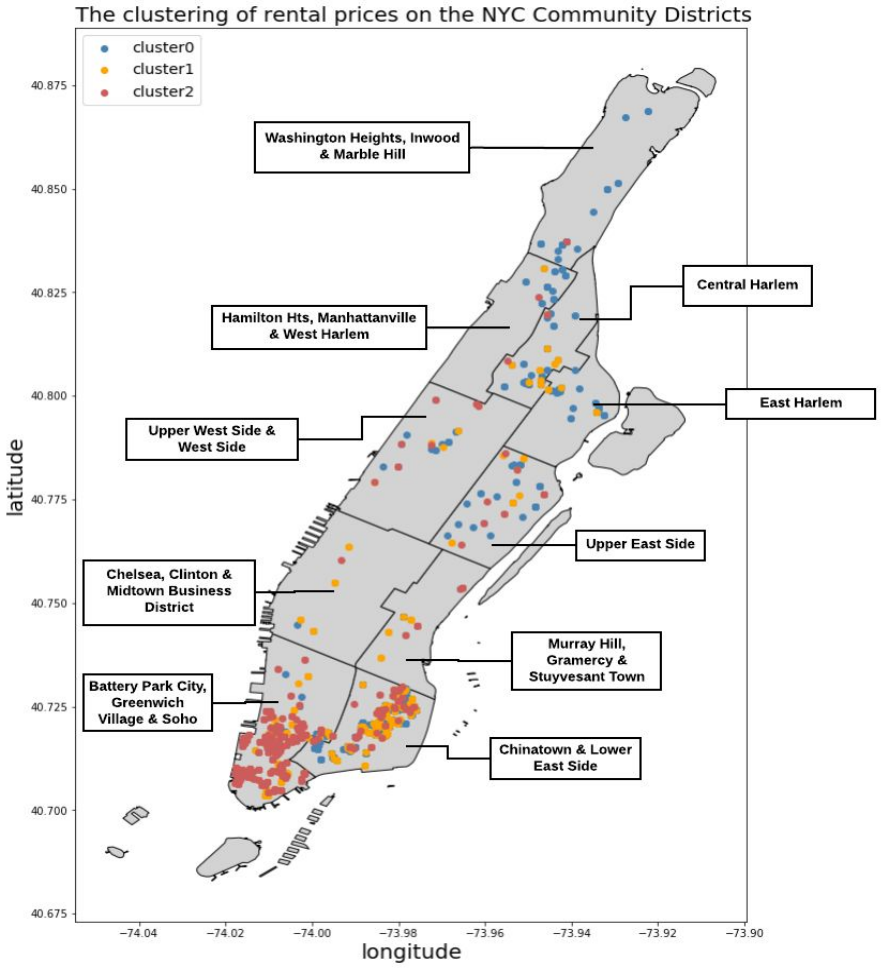

Fig. 6 clustering of rental pricing in Manhattan at NYC community Districts level, the largest portion of rental housings are located in the lower side of Manhattan

Conclusion

Rental house pricing is best predicted by random forest with ten most correlated features and the number of bathrooms/bedrooms are most pronounced features to determine rental house price.

Through K-means clustering, we find some patterns of the distribution of rental housing prices in Manhattan. For the lower side of Manhattan, rental house price above

$6245 is concentrated in Battery Park City, Greenwich Village & Soho and this area contains a large proportion of one/two bedrooms apartment. In Chinatown & Lower East Side, there are lots of low price rental houses and rental house prices within the range from $3850 to $6225. From the distribution, we can also find most of our samples are centralized on the lower side of Manhattan. For the upper side of Manhattan, all price range can be found in Upper West Side and Upper East Side, however, besides these two areas, rental house price is lower than $3848 and most of these units are studio and one bedroom apartment.

For further research, collecting as much data in Manhattan and even the rental housings data from the other four boroughs of New York City would gain us more insights about the rental housing market in NYC.

Contribution

Team member Qiuyao Liu is responsible for the data cleaning and clustering.

Team member Minqi Lu is responsible for the result analysis and visualization.

Team member Jiawen Liang is responsible for the data collecting and literature review.

Team member Guanjia Wang is responsible for model building and prediction.

All the team members worked together on writing the paper.

Code written for the project and the datasets used can be found on:

https://github.com/qiuyliu/MLC_project

Reference

[1] Sarac, Florentina. "Millennials Spend About $93,000 on Rent by The Time They Hit 30." RENTCafé Rental Blog. March 28, 2019. Accessed May 07, 2019. https://www.rentcafe.com/blog/apartment-search-2/money/millennials-spend-93000-on-rent-by-the-time-they-hit-30/.

[2] Tjokro, Moorissa. "Predicting NYC Renting Prices Using Lasso Regression." GitHub. May 26, 2017. Accessed May 07, 2019. https://github.com/moorissa/nycrentpredictor.

[3] Wu, Jiaoyang. "Housing Price Prediction Using Support Vector Regression." Housing Price Prediction Using Support Vector Regression. May 31, 2017. Accessed May 7, 2019. https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1540&context=etd_projects.

[4] Valente, James, Shanshan Wu, Alan Gelfand, and C.F. Sirmans. "Apartment Rent Prediction Using Spatial Modeling." Accessed May 7, 2019.http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.532.451&rep=rep1&type=pdf.