Machine Learning-based Apartment Recommendation System for Student

Final Report

Yanli Sun, Huaxiang Liu, Yi Jiao, Yikai Zhang, Haobo Zhu, Junzhu Xiang

1. Introduction

The Internet has become a significant transaction platform for the real estate industry. Along with the increasing student population in Georgia Tech, many of them, without any experience with housing searching and renting, negotiate in real life, so they tend to search online for help. Rents, distance to GT and life convenience are several important factors in decision making. In this project, we will mine data from rental websites like Zillow to develop a rental property recommendation system website for Georgia Tech students. After the website is implemented and generalized, if the user feedback is positive, we will open the website to anyone who wants to rent an apartment in the Greater Atlanta Area.

2. Problem Definition

The inspiration for us choosing this project comes from our painful experiences in searching for suitable accommodations as first-year graduate students. Details about these rental properties' locations, prices, neighborhoods and residents' reviews are scattered throughout several major data sources such as Google Maps, Yelp, and rental housings' official websites, making the searching process very difficult. On the other hand, some popular websites like Zillow and Zumper are unnecessarily complicated for Georgia Tech students. Therefore, our project aims to create a rental recommendation website that combines all useful features and information in a more concise and dedicated way.

3. Survey

3.1 Data Preparation and Analysis

Vojtech [1] offers existing techniques of web scraping and introduces its development in recent years. What's more, it also gives an explanation on how to carry out web scraping by providing a specific example. Jie Yang [2] offers a novel big data processing framework to investigate a niche subset of user-generated popular culture content on Douban, which is a well-known Chinese-language online social network. The paper [3] introduces the general procedure of how to collect social media data on content, usage, and structure via direct scraping and application programming interfaces (APIs).

3.2 Machine Learning

To make our recommendation system more useful and precise, we need to use machine learning. BE Laure discussed the application of data visualization in the digital mapping field [4]. The idea inspires us to create a dynamic map showing the real-time leasing fee of different apartments. Koh and his team provide insight into the interactivity of the user and the machine learning model [5]. The machine learning decision and prediction system can be more personalized and thus can benefit the user in question more.

3.3 Data Visualization

Lodha presented several techniques for visualizing the temporal dimension of urban crime information in GIS [6]. However, it can be improved by integrating other information like traffic flow, population, bus schedules. Meanwhile, a new platform, including Google Fusion Tables and Google Map, is implemented to visualize the field data set [7]. This paper can improve our project to a full-functional application instead of a demo and tools that we can use to get data, storage data set and visualize on a map.

4. Proposed Method

4.1 Intuition

Recommendation systems are widely used in many industries including advertisements. So based on previous experience, this can be implemented in a standard way. Compared with the traditional

recommendation system, our apartment renting website will integrate different layers of information to stand out from other renting websites and be successful. A traditional renting website typically provides information like price, size, location. However, we want to integrate criminal information and commute information, like the annual occurrence of crime events and traffic conditions. If our website is successfully deployed, we can expect to provide different aspects for a potential renter to evaluate and make a better decision on which apartment to rent. The best way to measure its success is through user studies, to see whether users are satisfied with the experience.

4.2 Description of Approaches

4.2.1 Data Collection and Cleaning

Our data is collected from one primary source: Zillow. We started with data from Yelp, but accurate house prices are not showing in the dataset so we replaced it with the Zillow dataset. At first, by applying Zillow API, very limited data is retrieved, and many of them are redundant and unusable for our project. So we decide to use web scrapers to get the data response from the Zillow searching page. On the Zillow website, we limited our search scope to a 30-mile circle centered on Georgia Tech. By using web scrapers, about 2000+ data about houses and apartment information is retrieved from these 16 Atlanta counties within this circle. Each piece of data contains useful information on address, price, living area, and the number of bedrooms and bathrooms. During the time of scraping data from zillow.com, lots of unpredicted things happen. For example, many returned data rows are empty. At first, we check the status code for every request sended, there are many 404 status codes returned. And many of them return status code 200, but the information returned is still empty. After carefully checking those returned information, we find that, many times, Zillow.com requires our client to pass the robot test. So we prolong the sleep time for every request, and set timeout to a higher value, and expand the pool of user-agent lists. And this problem is settled, no empty results are returned from zillow and we retrieved the original dataset. The original dataset is not large enough for us to do machine learning training. And we notice that there are multiple house data provided in each address, so by cleaning the data, we obtain a dataset with 21337 rows and 5 columns. Since there are some bad and missing data from the origin collected, we need to filter and normalize the data. For instance, in the house _info set, there are five main factors in each row corresponding to address, price, living area(size), number of bedrooms and bathrooms. However, some rows leave one or more categories totally blank that we cannot calculate and give the overall information of that address. Therefore, we need to delete the whole row to make the dataset more reliable. For the living area, the number of bedrooms and bathrooms, there are multiple separated numbers under each category and these data are matched in order as one house. For example, in Address "1447 Akridge St NW" there is a 1360 sqft house with 2 bedrooms and 2 bathrooms. But in the sample data, it can be observed that some of these numbers remain as "null" or "\\" which means this house information is incomplete. Therefore, we need to delete the corresponding data in the other two columns to guarantee all the other houses have correct and complete information.

Address	Price
['1447 Akridge St NW']	['1800']

livingArea
['1360', '1360', '\\°1', 'null', '975', '945', '1650', '1450', '1638', '1519', 'null']

bedrooms
['2', '2', '2', 'null', '3', '2', '3', '3', 'null', '3', 'null']

bathroom
['2', '2', '2', 'null', '2', '1', '2', 'null', '2', '2', 'null']

Figure1. Raw dataset

5642 Windwood Rd, Atlanta, GA	950	627	3	2.5
5642 Windwood Rd, Atlanta, GA	950	627	3	2.5
5642 Windwood Rd, Atlanta, GA	950	1254	3	3
5642 Windwood Rd, Atlanta, GA	950	1232	2	3
5642 Windwood Rd, Atlanta, GA	950	638	2	2.5
5642 Windwood Rd, Atlanta, GA	950	638	2	2.5
5642 Windwood Rd, Atlanta, GA	950	1232	2	2.5
5642 Windwood Rd, Atlanta, GA	950	627	1	2.5
5642 Windwood Rd, Atlanta, GA	950	627	3	2.5

Figure2. Cleaned data set

There are two main goals we need to achieve in the data cleaning part. The first one is to clear out all the incomplete information and its related data to ensure that all data remaining in the dataset are complete and corresponding. The second goal is that we need to make the data reliable and believable. We first try to use Openrefine to filter this data, the missing data is easy to be filtered out but separating each number under one attribute and deleting corresponding other factors are hard to achieve using Openrefine. Therefore we wrote a python notebook to clean the csv dataset by ourselves. We first read in the whole csv and then split each row into several with only 1 number in each column. Then for each separated rows, their area, bedroom and bathroom number are matched in order. But for some rows at the end they may lack some attributes since these three attributes do not have the same quantity. Finally we delete all the rows with missing data, "null" or "\\" in any box. We clean all the dataset and one address with several kinds of rooms is split into several as shown below.

4.2.2 Google Map API

Because we need to visualize the apartment information with a local map on our website, the coordinates of each data point need to be collected. To make the map as accurate as possible, using apartments' real latitude and longitude is a good choice in visualization. Since we can only get apartment addresses from zillow, we need to use google map API to obtain the coordinates of the data set.

For each row of data, the first column is the address of the apartment which we use as input. Then we request the latitude, longitude and county information of input address and append them after each row. The form of our final data set is shown as follow:

5642 Windwood Rd, Atlanta, GA	950	627	1	2.5	33.60036	-84.419296	Clayton County
5642 Windwood Rd, Atlanta, GA	950	627	3	2.5	33.60036	-84.419296	Clayton County
8229 Queens Dr, Atlanta, GA	1475	1939	3	1.5	33.5299233	-84.3869143	Clayton County
8229 Queens Dr, Atlanta, GA	1475	1939	3	1.5	33.5299233	-84.3869143	Clayton County

Figure3. dataset with appended coordinates

The factors of each column are address, price, size, number of bedrooms, number of bathrooms, latitude, longitude and county.

4.2.3 Machine Learning Model

KNN algorithm is utilized in the machine learning model to train the data collected. Its basic idea is that when the training set data and label are known, input the test data, compare the characteristics of the test data with the corresponding characteristics of the training set, and find the first k data most similar to the training set, then the corresponding category of the test data is the one with the most frequent occurrence of K data. Its algorithm is described as:

(1) Calculate the distance between the test data and each training data;
(2) Sort them according to the increasing distance;
(3) Select K points with the smallest distance;
(4) Determine the occurrence frequency of the category of the first K points;
(5) Return the most frequent category in the first k points as the prediction classification of test data.

Based on the ideas above, the KNN algorithm is given as follows:

Input the training dataset:

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\} \tag{1}$$

In formula (1):

$$x_i \in X \subseteq R^n \tag{2}$$

Formula (2) is an instance eigenvector of N dimension.

$$y_i \in Y = \{c_1, c_2, \cdots, c_n\} \tag{3}$$

Formula (3) is the category of the instances. Formula (3) is used to predict the instances x_i, which is the output that needed. In the project, the input is designed as:

$$\begin{bmatrix} longitude_1 & \cdots & bathroom_1 \\ \vdots & \ddots & \vdots \\ longitude_i & \cdots & bathroom_i \end{bmatrix}$$

The symbol '⋯' includes the latitude, area, bedroom. And the output is expected as:

$$[price_1, price_2, \cdots, price_n]$$

According to the given distance measurement method (in order to calculate in an efficient way, use Euclidean distance), K sample points closest to X are found in training set t, and the set represented by these K sample points is recorded as $N_{k(x)}$. The Euclidean distance is described as formula (4):

$$d_{xy} = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \tag{4}$$

According to the principle of majority vote as shown below, ensure the instance x belongs to the category y:

$$y = argmax \sum_{x_i \in N_{k(x)}} I(y_i, c_j), i = 1,2, \dots, N; j = 1,2, \dots, K \tag{5}$$

In the formula (5), I is the indicator function described as:

$$I(x, y) = \begin{cases} 1, if\ x = y \\ 0, if\ x \neq y \end{cases} \tag{6}$$

There is only one super parameter K in KNN algorithm, and the determination of K value has a crucial influence on the prediction results of KNN algorithm. In the process of project, K value is selected as 150, which is shown as follows:

```
In [2]: knn_class = 150
        po_deg = 3

In [15]: # MODEL #2: KNN
         neigh = KNeighborsClassifier(n_neighbors=knn_class)
         neigh.fit(x_train, y_train)
```

Figure5. KNN model script sample

What has to be mentioned is that another two machine learning models 'Polynomial Regression' and 'Linear Regression' were also considered, however, due to the failure of the prediction, the KNN model appears to be the best choice. The specific details will be offered in section 5.

4.2.4 Data Visualization

We create a website for apartment recommendation, the tools used in this includes: Node.js, bootstrap, express, mongodb and Google map api. The website is currently on beta version, and it can satisfy the basic needs like user registration, marker a point in map and do price prediction based on fixed location. For maker point on map part, we define the Longitude and latitude of center point and zoom scale， then create a marker based on original data set, add event listener

to marks so that when the user clicks on mark, there will pop up a info window of detailed information about this apartment , including address, county, area, bedroom, bathroom and prices. if the user chooses another marker, the previous window will be closed and the corresponding window will be popped out. besides, a user can pick on any location with the flag on the map, and he will get the prediction price for the potential apartment around this area. To better protect the website from malicious attack, only registered users will have the access to query for prediction price.

5. Experiments

5.1 The Design of Experiment

(1) Will extremely large prices affect the finally predicted price?
(2) How to decide the learning method of the model.
(3) Compare predicted result among KNN model, Linear Regression model and Polynomial Regression model
(4) The loading capacity test of the website.

5.2 Results and Analysis

(1) We found in the original data that some apartments have extremely high prices that are over hundreds of thousands, which is called the bad data or fake data. We first decide to filter out these data manually by setting a reasonable upper bound, but after calculation we found that there is no need to delete them. In our 21336 total data, we get that only 1.3% of the rent prices were greater than $5000 and only 0.6% of rent prices were greater than 20000. Machine learning is based on a large learning dataset and the proportion of these abnormal data is quite low. Therefore it will have almost no effect on our machine learning model as long as the reasonable data are in large numbers.

(2) At the beginning, we wanted to divide the price into many labels, such as number 1 represents cheap, 2 represents more expensive, 5 is the symbol of the most expensive ones. During our attempt, we labeled 1000-1999 dollars as 1, 2000-2999 dollars as 2, 3000-3999 dollars as 3 and etc. But in fact, the price gap is very large and the distribution is uneven. For example, there is no price gap between 10000 and 20000, and less than 1% price gap between 20000 and more, so we give up the scheme of dividing the price into labels. However, we change the price itself into labels which are defined by the tiers of prices, which is difficult to express the inner relationship among them, the output is of better quality and more intuitive.

(3) Since KNN model, Linear Regression model and Polynomial Regression model are three most commonly used models in machine learning, we once wanted to figure out which model is most suitable for the project. We compared the MAE and MSE of the output data predicted by these different models. The results are shown as follows.

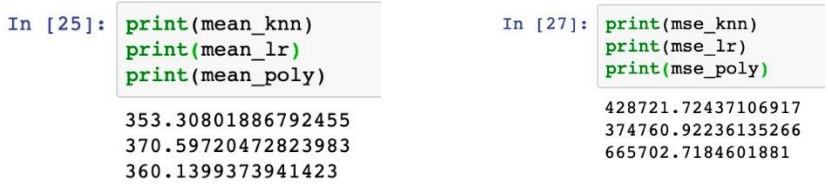

Figure6. Output of KNN, LR and PR model

According to the results and the experiments, the Polynomial Regression model (quadratic) can have the best performance yet the result is very unstable, and with the increase of power, the performance and training time worsen, so we exclude the Polynomial Regression model (power greater than 1). In the term of MAE, KNN model transparently has a better performance than Linear Regression model. Linear Regression model has a better performance in one-time comparison with KNN model in the term of MSE, however, we choose KNN model in the end. The reason is that although the Linear Regression model has

a smaller error, the KNN model is much more accurate when we have a check manually. In addition, Linear expression will have more extreme results, specifically, extremely bad. On the contrary, according to our experiment, the result offered by the KNN model is much more stable and convincing.

(4) Using testing tools provided by app.k5.io. And we do load testing, stress testing on this website. Load testing identifies the bottlenecks in the system under various workloads and checks how the system reacts when the load is gradually increased. Stress Testing determines the breaking point of the system to reveal the maximum point after which it breaks. In consideration of this website is a beta version, and it is mainly facing campus students in Atlanta, so we set UVs to 50, and peak request 50 per seconds and test loading capacity of this website for 12 minutes. During load testing time, about 27.7K requests have been made and 0 http requests failure. And the followings are a screenshot of the load testing results:

Figure7. Load testing results

And the next screenshot is the stress testing results:

Figure8. Stress testing results

6. Conclusions and Discussion

In this project we basically designed an apartment recommendation website using machine learning and data visualization on a real map. Users can register and login to our website and get a prediction price by setting desired information. All datasets were obtained on the latest zillow and google map official site. Most of these data are reliable but we still need to clean the missing and bad data. However, due to time limitation and this special period, we didn't get the criminal rate and traffic information of each location and we may optimize it later. Meanwhile we have tried several machine learning models and KNN works well currently. But we may find a better model if we add other attributes and functions in these systems. To conclude, we complete the main goal of our initial intuition, but there are still some details and functions we need developed in further working. For our website using guiding, click the location on the map that you want to predict the price around and enter the desired information. The red pinged points are apartments in our dataset and you can move the cursor on it to view the detail of that apartment.
Apartment recommendation URL: http://project6242.herokuapp.com/

Distribution of team member effort

All team members have contributed a similar amount of effort and the report was written together.

References

[1].Vojtech Draxl, “Web Scraping Data Extraction from websites”, University of Applied Sciences Technikum Wien, 04.02.2018

[2]. Jie Yang & Brian Yecies, 2016. Mining Chinese social media UGC: a big-data framework for analyzing Douban movie reviews. Journal of Big Data.

[3] Hai Liang, J. J. H. Z., 2017. Big Data, Collection of (Social Media, Harvesting).. In: The International Encyclopedia of Communication Research Methods. s.l.:s.n., pp. 118.

[4]. Laure, B. E., Angela, B., & Tova, M. (2018, April). Machine Learning to Data Management: A Round Trip. In 2018 IEEE 34th International Conference on Data Engineering (ICDE) (pp. 1735-1738). IEEE.

[5]. Koh, S., Wi, H. J., Kim, B. H., & Jo, S. (2019, June). Personalizing the Prediction: Interactive and Interpretable machine learning. In 2019 16th International Conference on Ubiquitous Robots (UR) (pp. 354-359). IEEE.

[6]. Lodha, Suresh K., and Arvind K. Verma. "Spatio-temporal visualization of urban crimes on a GIS grid." Proceedings of the 8th ACM international symposium on Advances in geographic information systems. 2000.

[7]. M. G. Lee, K. M. Yu and S. T. Chien, "Visualize Field Data in Fusion Tables - Take Chung Hua University Plant Map as Example," 2014 7th International Conference on Ubi-Media Computing and

Workshops, Ulaanbaatar, 2014, pp. 261-265.