GThomes

Jian Hua, MS-CS, jhua33@gatech.edu Jian Hua
Tianyu Zhan, MS-ECE ,tianyuzhan@gatech.edu
Geyu Wu, MS-ECE, wugeyu@gatech.edu
Haochen Li, MS-ECE, hli713@gatech.edu
Jing Bao, MS-CSE, jbao47@gatech.edu

December 6, 2019

1 Introduction

Currently, there are many websites, including Zillow, Trulia, Rent.com, etc., assisting individuals rent living place. However, most of these platforms offer various service to the general population instead of students. They may feel confused when facing such complicated functions. Here, we are building a succinct apartment hunting platform which is dedicated to helping GT students finding their perfect home.

2 Problem Definition

Our website will provide practical and thoughtful functions for each GT student who wants to rent a house. Most GT students would regard price as the primary factor. According to our survey, the rent has an extremely strong association with region and room type (studio, 1b1b, 2b2b, etc.). We gather approximation price by zip code and room type. Student would take a quick glance at the evaluated price of each room type from different regions. Moreover, we will provide ratings for safety, convenience and distance to GT.

3 Survey

We first examined the factors that influence renting decisions. Nathalie concludes with a review of the links between a possible connection between housing prices and real activity. It also analyses the reason of house price developments which can help us analyze rental data. We need to further visualize it.[1] Zhou et.al. conducted a survey in LA to learn transit modes for college students. The three main modes are driving, walking and public transportation. According to the study, gender, status (undergraduate vs. gradate) and age are significantly correlated to biking, walking or public transit.[2] Bowers et.al. visualize the crime data into a hot-spot maps and marked the dangerous region. We then would like to combine the crime heatmap with Google Map and Zillow to show whether a house is in dangerous region.[3]

Sean D. applies the dynamic Gordon growth model to the housing market and shows the housing premia are forecastable. He shows method on how to analyze the house price. But it doesn't connect with other factors when doing the analysis, and we will combine it with safety, surroundings and so on.[4]

We then look deeper to see how to forecast prices. Kim applies support vector machine (SVM) into stock price prediction. He also compares SVM's result with neural network's result.[5] It gives us some insight to our problem. Ali Abbasi Godarzi introduces Artificial Neural Network and dynamic Nonlinear Auto Regressive model. [6] Hanson builds an artificial neural network (ANN) to predict the house price in Turkey. It also demonstrates how to compare two different prediction model.[7] Refenes et.al. introduce arbitrage pricing theory into neural network prediction method. The special point is that they use sensitivity analysis and gain some reasonable explanation of their prediction.[8] Arlot gives a detailed guidance on how to use cross-validation procedure to select the best model for our case.[9] Also, Robin uses multiple listing data to predict house prices. In our project, we can take into consideration the

prices of neighboring houses. However, it generates errors when used stand-alone.[10]Limsombuchai shows that neural network model performs better than hedonic model on house price prediction.[11] In our project, we can adopt some of its ideas. Also, it used estimated price and we will improve this by using actual prices. Jingyi Mu compares different ML models on predicting housing price. It shows SVM performs the best. [12]

In perspective of clustering, Chan et.al performed K-means method to analysis the property price in Singapore. They considered 13 variables and used 9 clusters, rather than 6 in common rule of thumb, to get a better result. We have fewer variables thus we will simplify the process.[13]

We also studied various visualization methods to help us decide how to visualize our data. Juha introduces a visualization method called self-organizing map which works for multidimensional data. We will not use this directly because the visualization for higher dimension is problematic.[14] Muzammil lists various visualization techniques and introduce them briefly. As the methods shown in the paper each has its own shortages , we can combine several methods to show our data.[15] Lixin Li compares two spatio-temporal interpolation methods of geographic data and visualize them on a real estate data set. The visualization pictures look ugly. So we need to beautify it before utilization.[16]

4 Intuition and Innovations

The current rental platform like Zillow is hard to use. Users with little real estate knowledge might get lost at first with so much choices. And users have no where to find necessary information like safety and convenience on the website.

Our platform first provides the overview of approximated price of different types from each region (represented by zip code) so that student can quickly figure out which region they would like to live in. In this process, we scrape data from Zillow and use Google Map API to visualize. Moreover, we set up a rating method to separately evaluate safety, convenience, distance to Georgia Tech and price, then give the overall rating for each house. We also provide similar listings. Here, we will use scikit-learn library, a powerful machine learning library, to cluster data thus give the listing. More details are in following section.

5 Proposed Method

Our application used four datasets and three algorithms to develop a web app which can give the different ratings according to the importance of factor that the users choose and also give the similar recommendations according to the selected houses. In this way, it can help users find the best renting house for them.

5.1 Datasets

We used four datasets totally to help develop the web app.

5.1.1 Houses data collection and

We were trying to collect the data with Zillow API, however, there was limitation of this approach. We found that there were not current rent info of all houses but an estimated one was returned on the request. So we scrape from Zillow.com website to get the data by zipcode (30030, 30303, 30306...30354, 30363 which include 20 zipcodes). Here we employ web crawler to collect information of all currently posted houses in 20 zip code regions from Atlanta. We write our header in the program and it can automatically send the header to Zillow and receive the response HTML page, which is easily achieved by requests library. Lxml library is used to scrape the useful data from the HTML. Since HTML is a multi-layer structure, which can be interpreted as a tree structure. Lxml just treat it as a tree and introduce "path" pointing to each node, which is called xpath. Xpath to specific element are gained by Chrome. We also enable our program jump between different pages by scrape URLs. By which we can get detailed information such as area, numbers of bathrooms and bedrooms, also we can get all rent houses' information. Finally, the raw data is cleaned by regular expression. Finally we got a json file which had the houses info around Georgia Tech(1.01MB, 2685 entries)


```
{"uid": "4d36859d-cfdf-4be8-9573-b004ff609662",
  "name": "Arlo",
  "Sqft": 540,
  "bedrooms": 0,
  "bathrooms": 1,
  "price": 1418,
  "address": "245 E Trinity Pl, Decatur, GA",
  "url": "https://www.zillow.com/b/arlo-decatur-ga-65jH97/",
  "latitude": 33.77143, "longitude": -84.29308,
  "zipcode": 30030}
```


Figure 1: scraped zillow data example

5.1.2 Crime data collection

We download the crime data from Atlanta Police Department which has crime records from 01/01/2019 to 11/13/2019 as CSV file. The size of the file is 3.19MB which has 21620 crime records. Each record consists the time(the date and the morning, day, or evening time), location(street, latitude, longitude) of the crime occurrence. According to the dataset, we calculate the frequency of the crime occurrence in the same size of area. Then we use tableau to do the visualization of the crime data which is shown below. The labels on the figure mean the number of the crime events in that area.

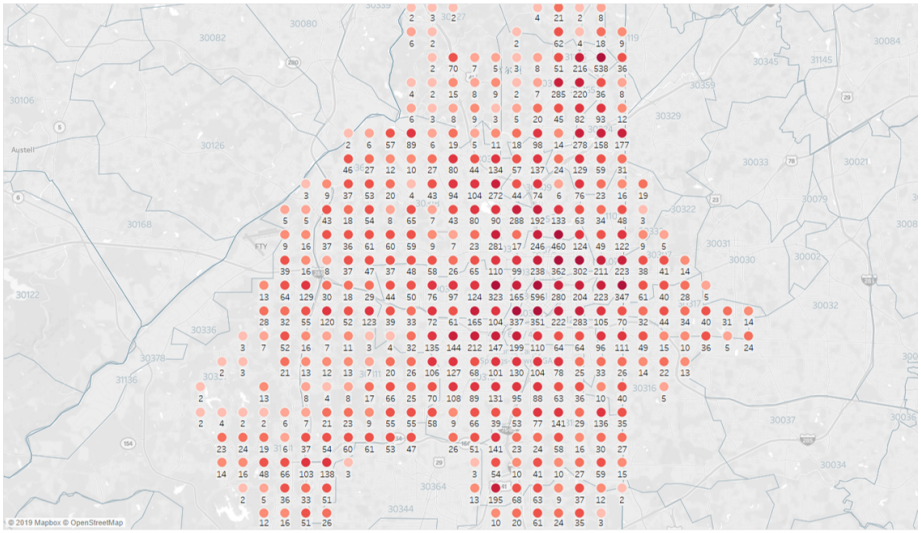


Figure 2: crime data visualization

5.1.3 Grocery data collection

We use Yelp API to get the grocery store data locations around Georgia Tech as a Json file(1.5MB, 2685 entries) And we use these data to evaluate the convenience of a specific house which will be mentioned in algorithm section. We use tableau to do the visualization of the grocery data which is shown below. The labels on the figure mean the number of the groceries in that area.

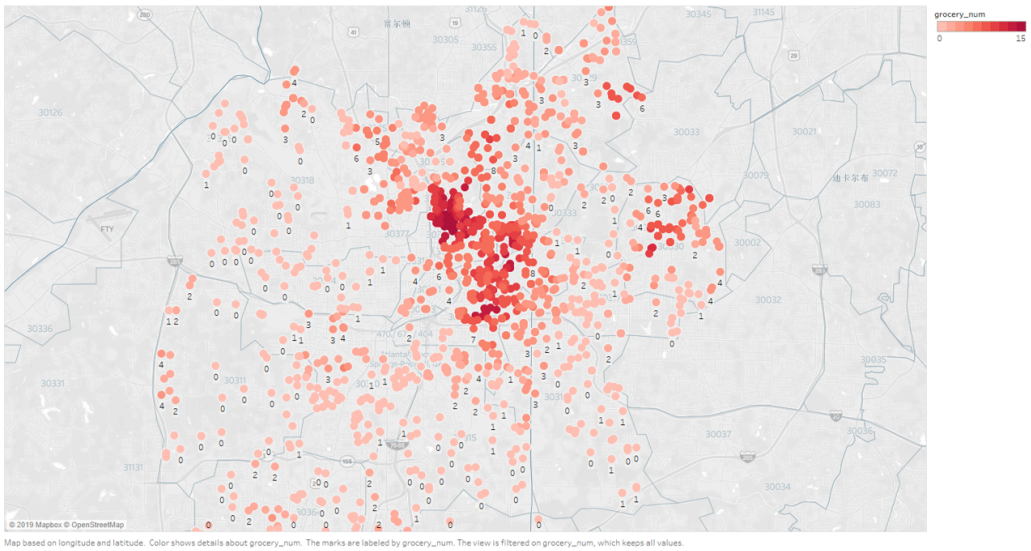


Figure 3: grocery data visualization

5.1.4 Location data

We use Google Map API to get the latitude and the longitude of the specific location points. And we can get the distance of two points We also use the location info to connect three other dataset together.

5.2 Algorithms

5.2.1 Forecasting general price by linear regression

Linear Regression algorithm is utilized to give the price prediction for given numbers of bedrooms and bathrooms. This function can be very useful for a student who has little knowledge about the real estate market in Atlanta. In our application, the user can just choose the number of bathrooms and bedrooms, say 1b1b, and application can give the predicted price in each zipcode region. Once the user choose the house type, our application just shows all corresponding houses. The calculating formula is given as below:

$$y = b1 \times n1 + b2 \times n2$$

. Where y is prediction and n1, n2 are the numbers of bedrooms and bathrooms. Although it is a quite simple model only takes price and two numbers into consideration, it can help those non-native students get the overview of each region before further explore it. For those students who have not decided which type of house they would like to rent, they can just leave these two numbers blank and our application renders all the houses.

5.2.2 Providing ratings for a specific listing

Once the user selects a region by zip code and the ideal number of bedrooms and bathrooms, the currently available listings that meet the user selection will be displayed on the map. By clicking a listing on the map, the user will be able to see the basic information including asking price, address and (maybe pictures). In addition, the user will also be able to see the safety rating, the convince rating and the distance to Georgia Tech campus.

Calculate safety rating:

We find the max number of the crime occurrence of all regions first. Then compare that number with other regions and finally calculate the safety rate. The formula is as below:

$$safty\ rating = \frac{max - number\ of\ the\ crime\ occurrence\ of\ the\ region\ in\ 2019}{the\ max\ number\ of\ the\ crime\ occurrence\ of\ all\ regions\ in\ 2019} \times 100$$

Calculate convenience rating:

We downloaded the convenience data (grocery store) from Yelp API and used it to calculate the convenience rating. The dataset consists of name of the grocery store and its location(latitude and longitude of the grocery store) and other information. The number of grocery stores around the searching area will be taken into consideration for convenience rating. The calculation formula for the convenience rate is as below:

$$convenience\ rating = \frac{number\ of\ the\ the\ grocery\ stores\ of\ the\ region\ in\ 2019}{max\ number\ of\ the\ grocery\ stores\ of\ all\ regions\ in\ 2019} \times 100$$

Calculate the distance rating to Georgia Tech campus:

The distance to Georgia Tech campus will be calculated by using Google Map API.Then we calculate the distance rating as the fomula below:

$$distance\ rating = \frac{the\ min\ distance\ between\ the\ apartments\ of\ all\ regions\ and\ the\ campus}{the\ distance\ between\ the\ apartment\ and\ the\ campus} \times 100$$

Calculate the price rating:

We get the price of the different apartments from the Zillow. Then we calculate the price rating as the formula below:

$$price\ rating = \frac{the\ min\ price\ of\ the\ apartments\ of\ all\ regions}{the\ price\ of\ the\ apartment} \times 100$$

Calculate the total rating: There are four factors in total and the user can choose the different importance for each factor. Each importance represents different ratio as below: Highly important: 100%, Very important: 75%, moderately important: 50%, slightly important: 25%, unimportant: 0%. So we calculate the the total rating as the formula below:

$$total\ rating = \frac{\sum_{i=0}^{3} rating_i \times importance_i}{the\ max\ of(\sum_{i=0}^{3} rating_i \times importance_i)\ of\ all\ regions} \times 100$$

5.2.3 Providing similar listings

Our application also enables similar house recommendation function. When a user chooses one house on the screen, we can provide three similar houses, by clicking which user can directly see those new houses' detail information. We utilize the k-nearest neighbors (KNN) algorithm to achieve this object. KNN algorithm is a simple supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm basic assumption is that similar things are close to each other. Here, we implement KNN algorithm for each zip code region so that we don't need to consider position any more. In other word, the similar recommendation must be in one region. Four features, area (Sqft), bedrooms, bathrooms and price ($) are employed and each entry is a four-dimension vector. Since the units are different, it is necessary to normalize all elements into float number between 0 to 1. We initialize K to 3, which means we choose top three nearest neighbors. In our case, the training data and test data are same dataset. Then we calculate the distance between each pair of vectors and list the three closest ones. For clustering or regression problems, the k neighbors can return the index labels or mode for a given new vector. But we just pause here since all we need is the similar recommendation. In practice, scikit-learn library is used in our code since it may directly process the data matrix efficiently.

5.3 User Interfaces

5.3.1 Home Page

Below is the home page of GTHomes. The user can input his/her ideal number of bedroom and bathroom or he/she can also leave it blank.

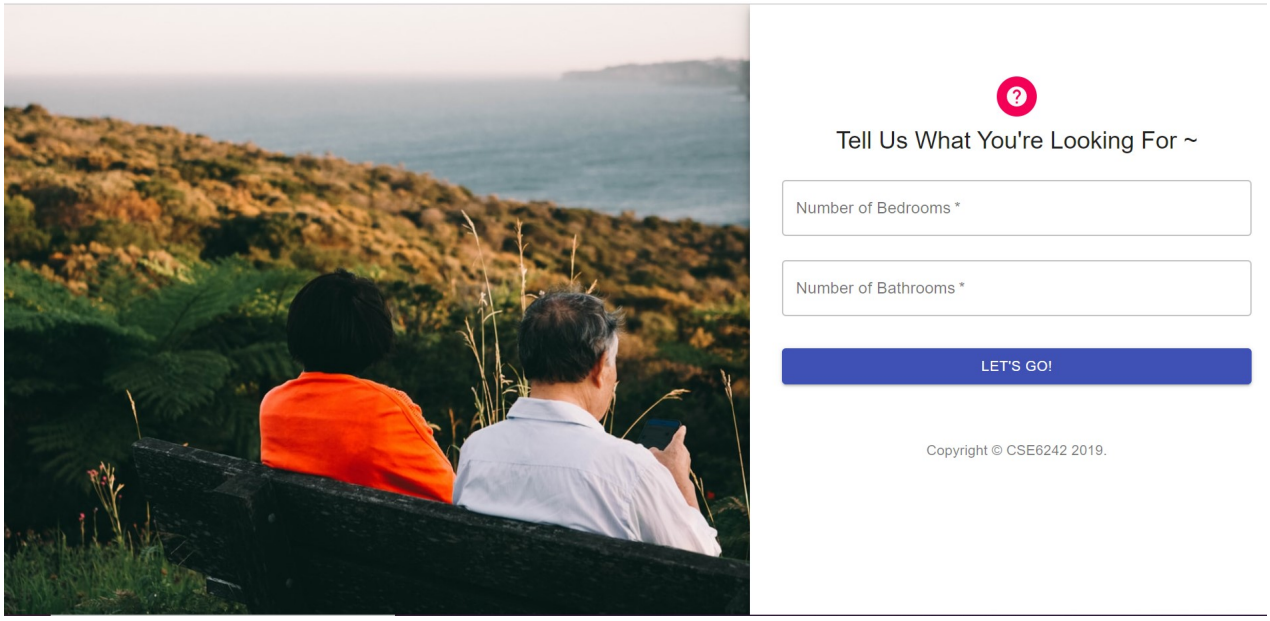


Figure 4: Home page

5.3.2 Main Page

The main page is an Atlanta map which can be zoomed in and out. On the map, there are from 1 to 10 labels which means the ratings of the house. 10 means best and 1 means worst.

Figure 5: Rating labels

For example, a user chooses the 1b1b in the home page. Then all 1b1b houses ratings will be shown in the main page.

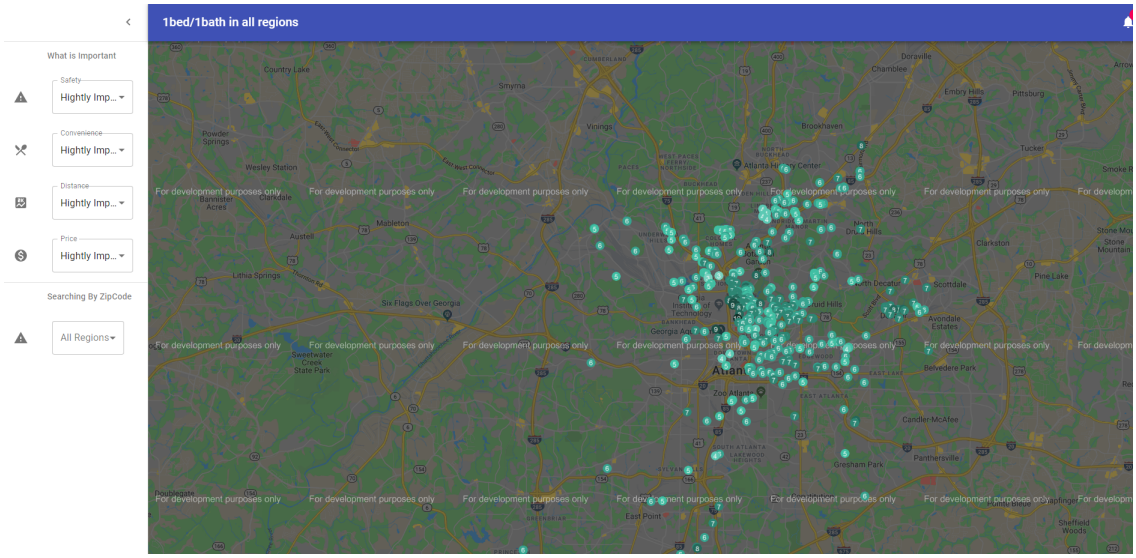


Figure 6: Main page

The left of the page is menu bar. The user can choose the ideal region which he/she wants to rent. For example, the user chooses 30309 Zipcode. Then the top of the main page will show the predicted price of this region which is 1556 dollars/Month. In this way, the user who may have little knowledge of the real estate market in Atlanta can get the approximate price to start off the searching process.

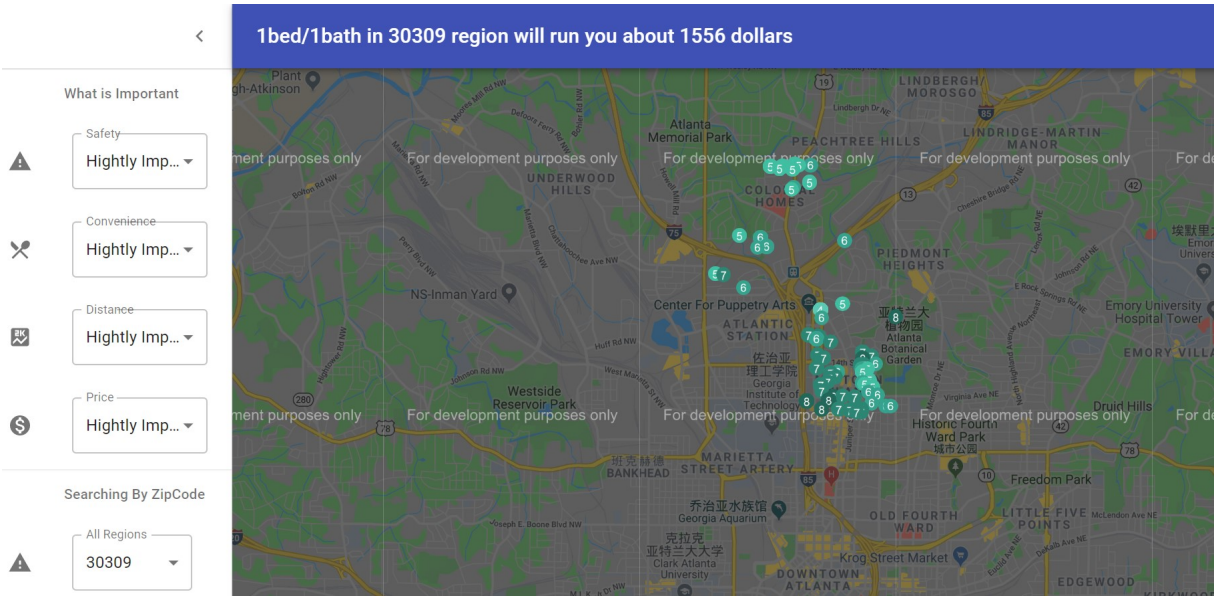


Figure 7: Main page with price prediction

In the menu bar, the user can also assign different weights on the importance of four factors: safety, convenience, distance and price. When the user changes the choice, the main page will reload and show the new ratings.

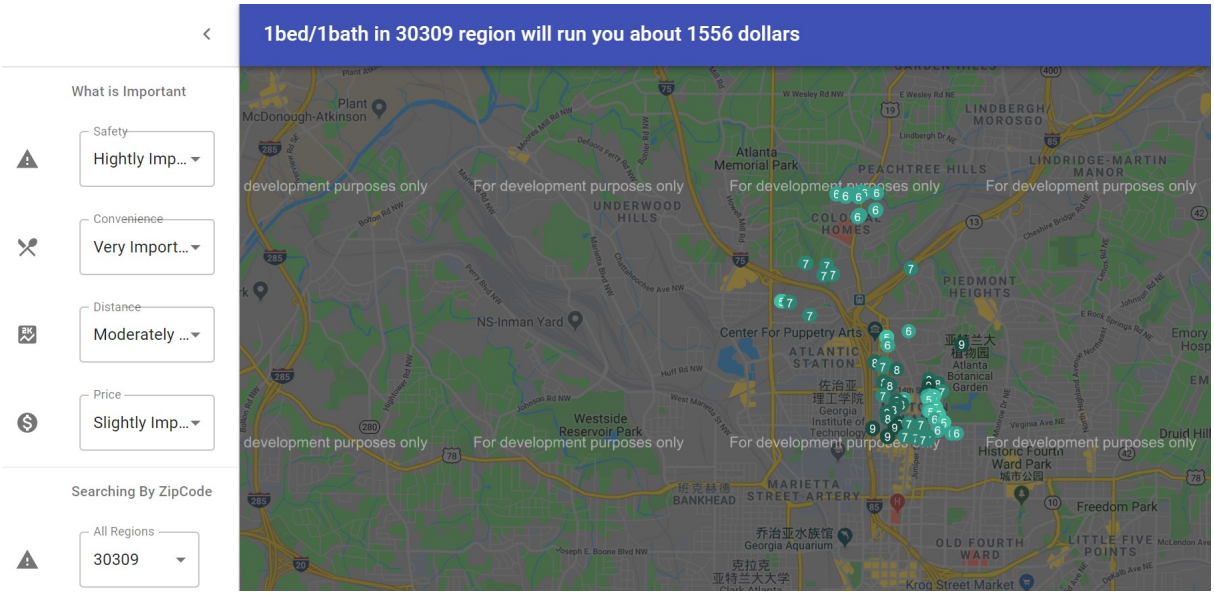


Figure 8: Main page when user changes the importance of factors

Then, user can click a specific label on the screen and then the detailed information of the house will be displayed in a popup box. In this box, user can know the basic information of the house and the safety, convenience rating of the house.

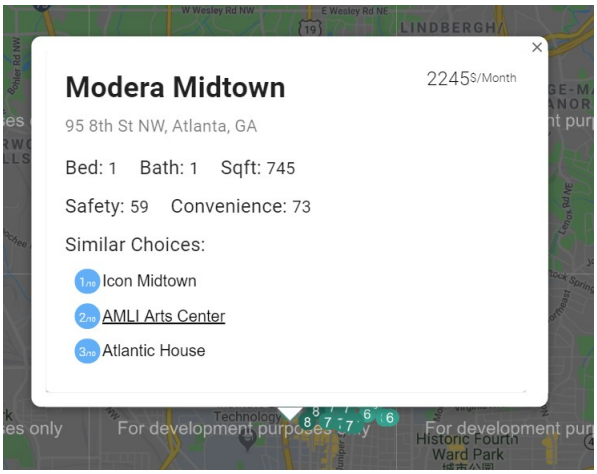


Figure 9: The popup of the detailed info for one house

Besides, there are three similar recommendation house for the user according to the house which is selected. The user can click any of the recommendation and then the popup will jump to that house. For this example, assume the user clicks the first recommendation, then the "Icon Midtown" house's detailed info will be shown.

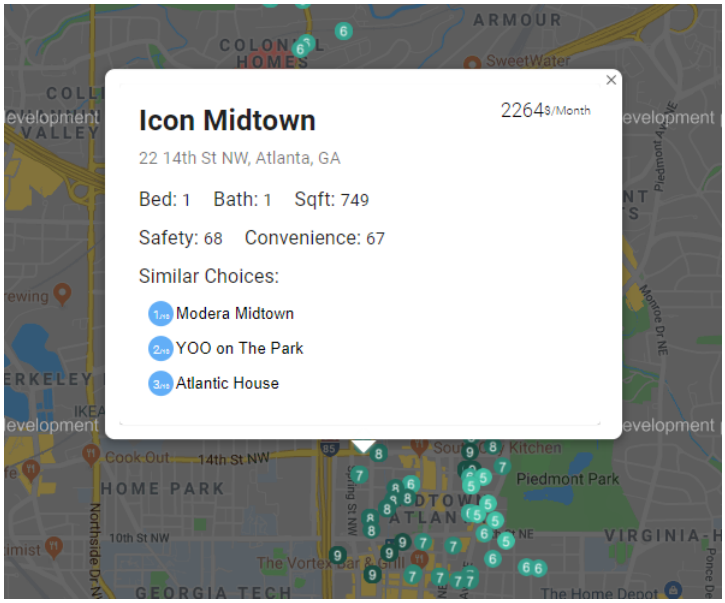


Figure 10: The popup which jump to the selected recommendation house

6 Experiments/ Evaluation

6.1 Testbed and List of Questions

To test the correctness and effectiveness of our work, we will examine our results and design experiments from the following perspectives.

- Are our predicted prices by zipcode in accordance with the reality?
- Are our potential users satisfied with our rating?
- Are the similar houses suggested really similar?

To answer question 1, we will visualize data by zipcode, if the data cluster around our prediction, our prediction is correct. To answer question 2, we will invite students to rate our application. If we receive a positive response, that means our rating system is reasonable. To answer question 3, we will invite gt students to provide a feedback on the results.

6.1.1 Recruiting Testers and Testing Sessions

We initially tested out our application within the group. Then, we started recruiting testers to participate in our experiment, and we were looking for testers who are first-year Georgia Tech students who did not live in the city of Atlanta before this semester and experienced apartment hunting at the beginning of this semester. Each team member recruited 2-3 testers for the testing sessions.

6.1.2 Survey and Interviews

After the testing session, we ask the tester to finish a quick survey where we collect both qualitative and quantitative data for our experiment. We first ask the testers to rate how accurate our rating and

prediction price are from a 1 to 10 scale. Then we collected feedback for finding similar listings, UI design and overall value of our application. And in the end, we ask the tester to comment on changes and improvements they would like to make.

6.1.3 Questions Asked

The questions we asked are as follow:

- Is the safety rating accurate?
- Is the convenience rating accurate?
- Is the distance rating accurate?
- Is the price rating accurate?
- Are the price predictions accurate?
- Are similar listings recommendations helpful?
- Is the UI easy to navigate?
- Will Georgia Tech students want to use this application?

6.2 Experiments and Observations

6.2.1 Rent Prices Visualization

As the picture shows below, the x axis refers to the number of bedroom, the y axis refers to the number of bathrooms and the z axis refers to the price. The points are the actual data from our data-set. The surface is our forecast result by linear regression. Because there are too many zipcode, we only show part of our result as example. The majority of the points cling tightly on the surface. That means that our prediction is right and reliable.

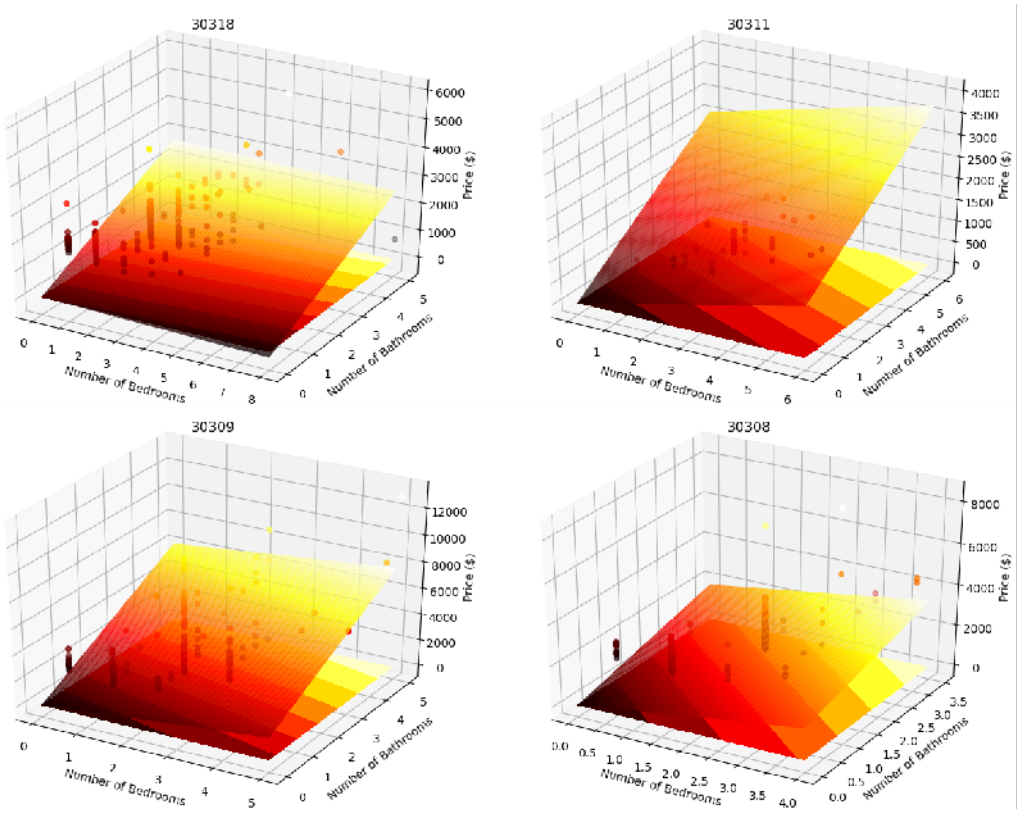


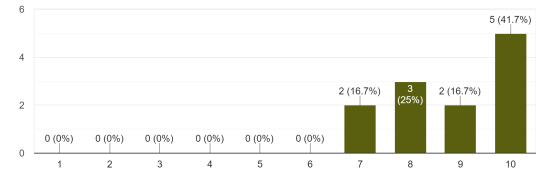
Figure 11: linear regression result visualization

6.2.2 GT Students' Reflection To The Application

We conducted 5-minute testing sessions with each tester. We first ask the tester the number of bedrooms and bathrooms in their current apartment. Then, we ask the tester to search the same number of bedrooms and bathrooms by using our application. Each tester was asked to review the ratings were provided by the application, then give feedback to these ratings during the first 3 minutes, then the tester was asked to review the prediction price in his or her current zip code and review similar listings during the last 2 minutes.

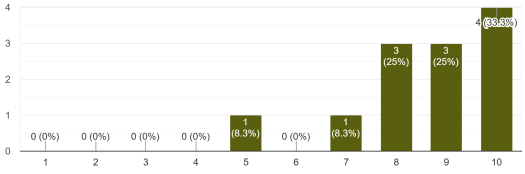
How accurate would you say the safety ratings are.

12 responses



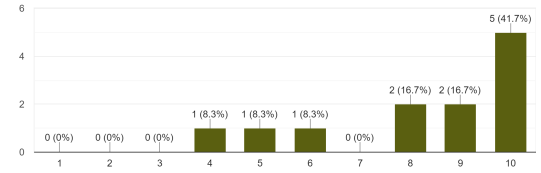
How accurate would you say the convenience ratings are.

12 responses



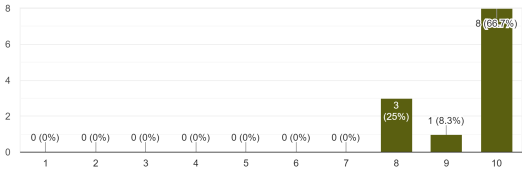
How accurate would you say the distance ratings are.

12 responses



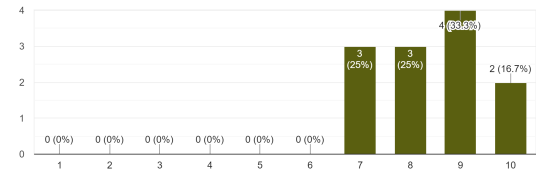
How accurate would you say the price ratings are.

12 responses



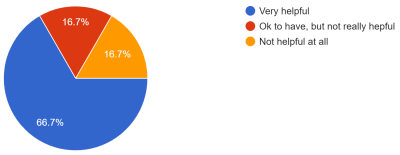
How accurate would you say the Price predictions by zip code are.

12 responses



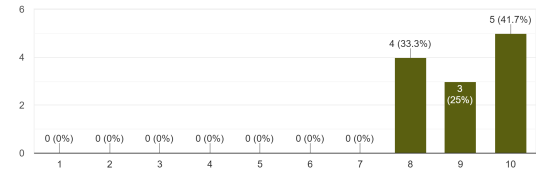
Is the similar listings recommendation helpful

12 responses



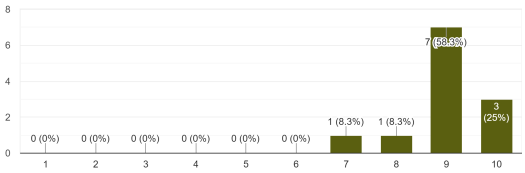
If you were to review our UI what score would you give it out of 10?

12 responses



Overall, how well did GT home meet your needs at a Georgia Tech students

12 responses



How likely are you to recommend us to a friend or colleague at Georgia Tech?

12 responses

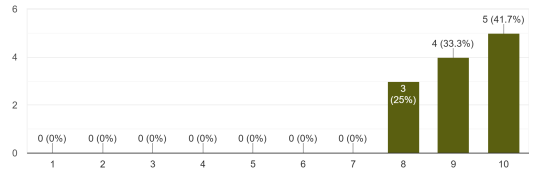


Figure 12: Survey Results

Figure 13: Suggestions

In terms of rating evaluation, the majority of the testers believe our ratings and price predictions are accurate to their experience. However; we noticed that for distance rating, we got a few lower score feedback. We believe this is because we used the absolute distance rating calculation instead of using commute distance. This problem could be resolved by using paid Google Map API to obtain the real-time commute distance instead of the free version we currently have. And for finding similar listings feature, the majority of the testers believe it is a nice feature to have. In the overall feedback, we are very pleased to see the testers gave very high scores to the overall functionality and UI design and they are willing to recommend GT home to their friends. This indicates our project has real potential business value and is ready to expand to many campuses.

In the change/improve comment feedback, we can see from the word could, the word Region and Distance appeared very frequently. We carefully reviewed all the feedback and observed the two most frequently asked improvements are supporting different regions/campuses and providing commute distance instead of absolute distance. These two features can be easily implemented when we have some funding and decide to push this project to a startup idea in future work.

7 Conclusions and discussion

7.1 Conclusions

We designed and developed a housing recommendation platform for GT students. Our platform has a very user-friendly user interface as well as a fast and stable running time. It supported many features, such as customized housing recommendations, price prediction, and similar housing recommendations.

We integrated three different algorithms into our platform:

- We provided approximated prices for different housing options based on a linear regression model, which help the users quickly narrow down their target options based on their budget.
- We provided a housing recommendation list based on four factors: distance, price, convenience, and safety, as well as the importance levels specified by the user.

- We supported a similar housing recommendation feature calculated with k-nearest neighbors models based on features such as the number of beds, number of baths, room area and price.

With the well-designed user interface and advanced algorithms, our platform provides GT students their perfect home.

7.2 Discussion

Real-time data could be integrated into the platform with more funding such that the recommendation list will be more up-to-date. Also, our platform now supports the region around Georgia Tech. It can support more regions with future development.

8 Plan of Activities

8.1 Distribution

All team members have contributed similar amount of effort.
Jian Hua: Team leader. Responsible for model evaluation and algorithm design. Mainly focus on building of models and helps to data visualization.
Geyu Wu: Front-end developer. Responsible for the web end data visualization and mainly focus on finding innovative methods of presenting data.
Tianyu Zhan: Quality controller and tester. Responsible for the correctness and robustness of the models and codes. Also works on evaluation of the performance of the algorithms.
Jing Bao: API developer. Responsible for API and interface development, data collecting and processing, analysing recommendation algorithm and helps to build models.
Haochen Li: Back-end developer. Responsible for data processing and database building. Also focus on building of models.

8.2 Timing Frame

Project Proposal (Sept. 15-Oct. 14): Brainstorming and deciding directions of discovering. Literature review and proposal writing. Start making prototypes.
Progress Report (Oct. 14-Nov. 10): Data collecting and processing. Algorithm design and model building. Visualization design and coding.
(new)Demo Making(Nov. 10 - Nov. 24): Continue coding. Finish the first demo.
Project Final Presentation (Nov. 24-Dec. 03): Model performance improvement. UI and system logic improvement. Code refactoring. Make presentation videos and posters.

8.3 Environment and Tools

Python, Javascript, Node.js, Bootstrap, D3, SQlite, etc.

References

[1] Nathalie Girouard et al. "Recent House Price Developments". In: 475 (2006). DOI: https://doi.org/https://doi.org/10.1787/864035447847. URL: https://www.oecd-ilibrary.org/content/paper/864035447847.

[2] Jiangping Zhou. "Sustainable commute in a car-dominant city: Factors affecting alternative mode choices among university students". In: *Transportation Research Part A: Policy and Practice* 46.7 (2012), pp. 1013–1029. ISSN: 0965-8564. DOI: https://doi.org/10.1016/j.tra.2012.04.001. URL: http://www.sciencedirect.com/science/article/pii/S0965856412000651.

[3] Kate J. Bowers, Shane D. Johnson, and Ken Pease. "Prospective Hot-Spotting: The Future of Crime Mapping?" In: *The British Journal of Criminology* 44.5 (May 2004), pp. 641–658. ISSN: 0007-0955. DOI: 10.1093/bjc/azh036. eprint: http://oup.prod.sis.lan/bjc/article-pdf/44/5/641/1260635/azh036.pdf. URL: https://doi.org/10.1093/bjc/azh036.

[4] Sean D. Campbell et al. "What moves housing markets: A variance decomposition of the rent–price ratio". In: *Journal of Urban Economics* 66.2 (2009), pp. 90–102. ISSN: 0094-1190. DOI: https://doi.org/10.1016/j.jue.2009.06.002. URL: http://www.sciencedirect.com/science/article/pii/S0094119009000400.

[5] Kyoung-jae Kim. "Financial time series forecasting using support vector machines". In: *Neurocomputing* 55.1 (2003). Support Vector Machines, pp. 307–319. ISSN: 0925-2312. DOI: https://doi.org/10.1016/S0925-2312(03)00372-2. URL: http://www.sciencedirect.com/science/article/pii/S0925231203003722.

[6] Ali Abbasi Godarzi et al. "Predicting oil price movements: A dynamic Artificial Neural Network approach". In: *Energy Policy* 68 (2014), pp. 371–382. ISSN: 0301-4215. DOI: https://doi.org/10.1016/j.enpol.2013.12.049. URL: http://www.sciencedirect.com/science/article/pii/S030142151301313X.

[7] Hasan Selim. "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network". In: *Expert Systems with Applications* 36.2, Part 2 (2009), pp. 2843–2852. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.01.044. URL: http://www.sciencedirect.com/science/article/pii/S0957417408000596.

[8] Apostolos Nicholas Refenes, Achileas Zapranis, and Gavin Francis. "Stock performance modeling using neural networks: A comparative study with regression models". In: *Neural Networks* 7.2 (1994), pp. 375–388. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(94)90030-2. URL: http://www.sciencedirect.com/science/article/pii/0893608094900302.

[9] Robin A. Dubin. "Predicting House Prices Using Multiple Listings Data". In: *The Journal of Real Estate Finance and Economics* 17.1 (July 1998), pp. 35–59. ISSN: 1573-045X. DOI: 10.1023/A:1007751112669. URL: https://doi.org/10.1023/A:1007751112669.

[10] Jingyi Mu, Fang Wu, and Aihua Zhang. "Housing value forecasting based on machine learning methods". In: *Abstract and Applied Analysis*. Vol. 2014. Hindawi. 2014.

[11] Visit Limsombunchai. "House price prediction: hedonic price model vs. artificial neural network". In: *New Zealand Agricultural and Resource Economics Society Conference*. 2004, pp. 25–26.

[12] Sylvain Arlot, Alain Celisse, et al. "A survey of cross-validation procedures for model selection". In: *Statistics surveys* 4 (2010), pp. 40–79.

[13] Chan Lily, Ng Heng Tiong, Rishi Ramchand, et al. "A cluster analysis approach to examining Singapore's property market". In: *BIS Papers* 64 (2012), pp. 43–53.

[14] Juha Vesanto. "SOM-based data visualization methods". In: *Intelligent data analysis* 3.2 (1999), pp. 111–126.

[15] Muzammil Khan and Sarwar Shah Khan. "Data and information visualization methods, and interactive mechanisms: A survey". In: *International Journal of Computer Applications* 34.1 (2011), pp. 1–14.

[16] Lixin Li and Peter Revesz. "Interpolation methods for spatio-temporal geographic data". In: *Computers, Environment and Urban Systems* 28.3 (2004), pp. 201–227.

[17] *sklearn.cluster.KMeans*. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[18] *sklearn.metrics.homogeneity*$_score$. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.homogeneity_score.html.

[19] *sklearn.metrics.completeness*$_score$. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.completeness_score.html.