CEE 6242

Final Report

Date: Dec.5th, 2016

Final Report: Real Estate Search and Recommendation System

Prepared by:

Yunhe Song

Jin Yan

Yichao Yao

Qinlan Zhang

Xu Zhang

Prepared for:

Dr. Duen Horng Chau

1. Introduction & Problem definition

CEE Salesman focuses on the development of interactive, map-based real estate recommendation website for homebuyers in Atlanta including environmental factors, which is important in making decision, besides price and physical factors in existing search website such as Zillow. To provide valid information including environmental factors, price and physical factors, a real estate Geographic Information System (GIS) will be created after regression on physical and environmental data from Google map, Zillow, Police Department, City of Atlanta and so on, and our user can set their preference as well.

2. Survey

For our quantitative measurement model, environmental factors are proved to be important to the decision making of a home-buyer in many previous researches. *McNulty* (2000) found that the crime rate has a significant influence in housing choice. *Vega* (2009) translated the theory of economic choice behavior into models suitable for the empirical analysis of housing location. *Din* (2001) and *F.Kong* (2007) used the both linear regression and Artificial Neural Networks(ANN) to conduct a case study on value of green space and by hedonic pricing model. *D.Amenyahn* (2013) addressed the impact of locational characteristics and the impact of apartment physical characteristics. *F. Liao*(2014) compared preferences to actual residential and travel choices in two contrasting subregion. The most latest model included online reviews for the real states, a report by *Keller Center* (2012) specifically addresses the influential impact that online reviews. However, when discussing the environmental factors, previous researches neglected some important aspects. For example, *Yuan*'s (2013) and *T.Zeng*'s (2001) researches only include the distance to different locations.

As for Recommendation system, Collaborative filtering is a frequently used recommendation algorithm, while there are other recommendation algorithms which can provide better results under some specific circumstances. For example, *Xiaofang Yuan*(2013) used the case-based representation with combination of ontological information construction to set up the recommendation algorithm. *Choonho* (2003) presents another algorithm based on multi-level association rule mining with better performance. A maximum entropy web recommendation model integrates the user information with the semantic content features by Latent Dirichlet Allocation approach, performing better in accuracy and sparsity (*Jin, Zhou* 2005).

To Evaluate the recommendation system, *Mathwick* (2001) found that customers evaluate those service quality on five listed dimensions: tangibles, reliability, responsiveness, assurance. Service's performance and people's expectation are important judging parameters on benefit. *Bolton* and *Drew* (1991) suggest net value of essential services will obviously influence customers' satisfaction, and their relationship is positive proportional.

For Map visualization method, *Dean* (2008) pointed out that by restricting the programming mode, it was easy to parallelize and distribute computations with a fault-tolerant implementation. An efficient way to implement the GIS part is using map mashup, which allows for the fairly easy creation of Maps mashups by Web designers by providing extensive documentation of the interface. (*Miller* 2006)

3. Proposed Methods

3.1 Intuition

Our project has the following innovations compared with previous research:

1. Including more factors

When computing the score of different locations, the model will include a combination of environmental and physical factors such as security, schools, supermarkets and hospitals. The major search engine such as Zillow only has one factor, nearby school.

2. Customized regression result

Weight for factors will be assigned according to adjusted regression result according to user setting of priority.

With these two innovations, the search results will be more specific, enhancing the searching efficiency and accuracy and fulfilling more requirements of homebuyers. Furthermore, the results will not only satisfy the general user's preference, that is the result of regression, but also fulfill each individual homebuyer.

3.2 Description of Approaches

3.2.1 Data

1. Included Data

Firstly, physical property house data is from RealtyTrac, we developed a python crawler to retrieve recently transaction data including address, latitude, longitude, bedrooms & baths number, space area, house type and transaction price, which will be used for regression. Also, the current on sale house data including price and other factors are from Zillow. Secondly, crime rate according to areas can be achieved from atlantapd.com, which includes location (address, latitude, longitude), crime type, crime time and beat. Finally, other environmental factors data, location of schools, markets, hospitals and city hall can be directly downloaded from open data in atlantareginal.com, which includes location (address, latitude, longitude), public elements type and etc.

2. Data Processing

First all, we cleaned all collected data by removing null/error records, then using location to pair those records. To quantify the influence environmental factor, harmonic mean of distance from house to all nearby facilities will be used as index for later calculation. Besides, crime rate itself will be included as index. Using harmonic mean of distance will highlight the importance of nearest facilities for people preferring choosing them. Finally, processed data will include houses' location, their physical properties, size, number of years and indexes of chosen environmental factors for further regression and evaluation.

3.2.2 Algorithm

In this project, multivariate linear regression is used to predict the price of real estate on the market by former transaction data and set the initial weight for each environmental factor. Following is expected equation:

$$Price = \beta_0 + \beta_1 Distance + \beta_2 Physical\_Properties + \beta_3 Enviornmental\_Factors$$

Table 1

	Included Factors
Physical Properties	Size, Owner and Age
Environmental Factors	Crime Rate, Distance of Schools, Supermarket, Hospital

Our algorithm is to choose one optimal regression model that effective estimate value of houses with processed data. Find the coefficient of environmental factors for next step's evaluation that people's personal preference will alter the weight of environment factors and finally leads to their own price estimation. So our algorithm is to test the performance of different models and obtain the best one. Adjusted R-squared and t-test will be considered in model choosing.

Besides statistic methods, we also use training and test data to evaluate different regression models. Model with less error on test data will be viewed as better model. All training data and test data are randomly picked from processed data.

In regression process, we choose 80% of sold house data as training data and left 20% as test data to justify whether our model is good to explain house price.

3.2.3 User Interface

In this project, a web page will be designed to represent the results of the property recommendations. The web page mainly consists of two parts: one is the control panel to set preferences provided by the users and the other one is the results shown in the map.

In the first part, the users can input the requirements of the properties such as the location, the room-type, the working location of the user as well as the preference priorities of the factors, namely price, location, security, distance to nearby schools, hospitals and supermarkets. All the inputs of the users will be retrieved by PHP and combined with the data in the database and the algorithm to compute a score for the properties in the county specified by the users.

After all the inputs from the users are gathered, the results will be represented in the map, in which the top ten recommended properties in the specified county will be marked with a marker showing the corresponding rank, address, price and score. Furthermore, the area under investigation will be covered by the heat map layer. The colors of the heat map will indicate the level of the recommendation of purchasing the properties in a certain location according to the user's preference. Then the users can click on the marker of the properties they are interested in, and the corresponding information will show up. The above figure 3.1 shows the user interface as follow.

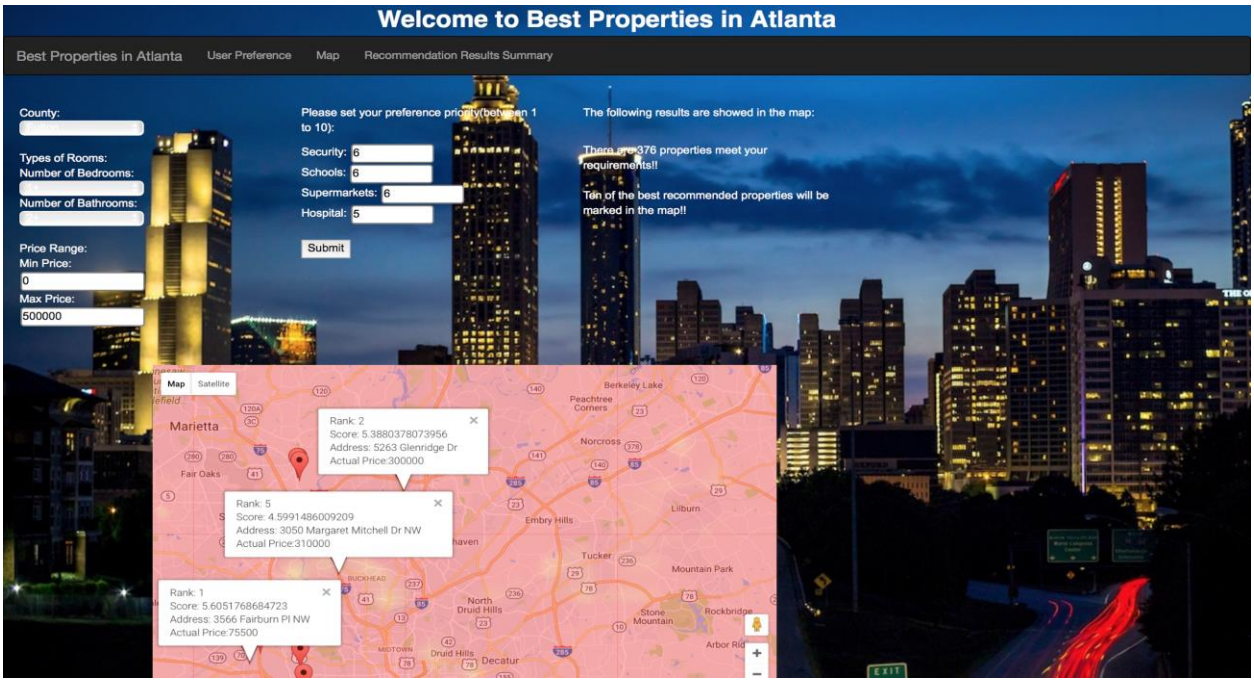

Figure 3.1 User Interface

4. Experiments and Evaluation

4.1 Test bed

We randomly choose 20% data in sold house as our test case.

4.2 Questions designed to answer

To evaluate the quality of this project. The list of questions designed to be answered:

1. Is the data correct?
    - 1.1 Is the data matched the resource house property data? Does it make sense?
    - 1.2 Is the data matched the resource environmental factor data? Does it make sense?

2. Is the algorithm correct?
    - 2.1 How does the regression model work?
    - 2.2 What is the result of this algorithm?
    - 2.3 What is the performance of this algorithm?

3. How is the visualization?
    - 3.1 Does it work properly?
    - 3.2 How fast it can do? What is the capacity?

4.3 Experiments

4.3.1 Experiments for Data

The transaction record of Atlanta in recent 10 months is retrieved. To measure the validity of data which will be used in regression, experiments are made as the table described below.

Table 4.1 Experiments Data

Dataset	Data after filtering	Data resource	match
House property	Address & Zipcode Latitude & Longitude Room & Space Price	Address & Zipcode Latitude & Longitude Room & Space Price	100% match
	Quantity	Quantity	Filter loss 4.2%
crime	Address & Zipcode Latitude & Longitude Time Crime record	Address & Zipcode Latitude &Longitude Time Crime record	100% match
	Quantity	Quantity	Filter loss 1.5%
Environmental factors	Address & Zipcode Latitude & Longitude Weight index	Address & Zipcode Latitude & Longitude	Create weight index For address etc. 100% match
	Quantity	Quantity	100% match

The data itself is quite reasonable. After filtering the records that contained *NULL* and those records obviously unreasonable (e.g. $1000 for a 300 sqft house), the database is a good match for regression in the following algorithm.

4.3.2 Experiments for Algorithm

1）Model test in testing data

To validate the correctness of our algorithm, we used test set of house sold in 2016 and observe the relationship between estimated price and price because we assume the preference of different environmental factors are on average. That is, good fit of algorithm means our estimated value has similar distribution of original data, and the model has a good performance on the test set.

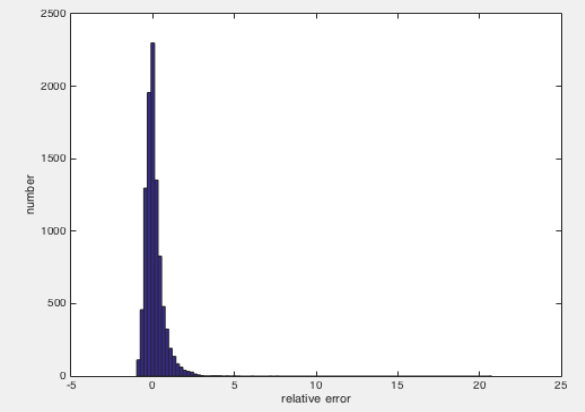

Figure 4.1 Relative error in testing data

From figure 4.1, average error between estimated value and original data is close to 0, ratio of estimated house with less than 20% error is 97%. Based on test set, former regression model is pretty good fit.

2) Regression Result

We only select size and all those environmental factors. Besides, we notice that some houses with extreme low price and size ratio that those houses contain many unexpected reasons such as no property, that will hurt the correctness of our model and those house we don't think will be good choices for our potential customers.

Core question in algorithm step is parameters and their forms in regression model. Because house with larger size will be easily associated with higher price, we use ln(size) in regression model. For environmental index, we also select their quadratic for precise estimation. Following equations are our final regression model.

$$\begin{aligned} ln\,(Price) \;=\; & 0.9 + 1.4737 ln\,(size) \;+ 0.1933 School_{Index} - 0.0033\, School_{index}^{2} \\ & - 0.2325 Hos_{index} + 0.0037 Hos_{index}^{2} + 0.0572 market_{index} \\ & - 0.0009 market_{index}^{2} - 0.0398 crime_{index} \end{aligned}$$

Adjusted R-squared 0.8070;

3) Comparing with common model (price = f(size))

Result for common estimate model is

$\ln(\text{Price}) = 7.1307 + 0.6619\ln(\text{Size})$, $R^2 = 0.333$

This model with lower R2 than our model listed before and we don't choose it. And this model is the most common method used to estimated price of house. Increasing those variables will explain more changes in house price.

4）Multicollinearity test

If there exists severe multicollinearity problem, the precise for regression model will be hurt. So it is essential to obtain the index for multicollinearity to justify how severe it is. Commonly, VIF will be used to represent it and smaller VIF value means data faces less multicollinearity problem and once VIF value larger than 20, people should take some methods to decrease possible hurt. Following table is VIF test for our model.

Table 4.2 Multicollinearity test result

VARIABLE	VIF VALUE	PROBLEM
SIZE	1.2307	No
SCHOOL	13.9830	Little
HOSPITAL	11.9594	Little
MARKET	12.4988	Little
CRIME	2.0754	No

From result table, School index, Hospital index and Market index, those data face little but acceptable multicollinearity problems.

4.3.3 Experiments for Visualization

The experiments of visualization can mainly be divided into two parts, the response time of resubmitting the priority set by users and the quality of heatmap after changing the priority.

For the response time, it highly depends on the network quality. 2500 data points will be loaded into the callback function of google map API, and the bottleneck of response time is the speed of google map API. With fast internet connection, the web will response within 1 min after setting the new priorities.

For the quality of heatmap, it cannot be judged in quantity, so that several different settings pattern are tested. Users can define their own settings in the following control panel shown in figure 4.2.

Figure 4.2 Control Panel of user interface

By default, all four preferences will be set to 5, which means user cares about all four factors on the same level (normal level). The following figure shown the result of visualization. The red areas get higher environmental score than the green areas. And ten recommendations are marked on the figure 4.3.

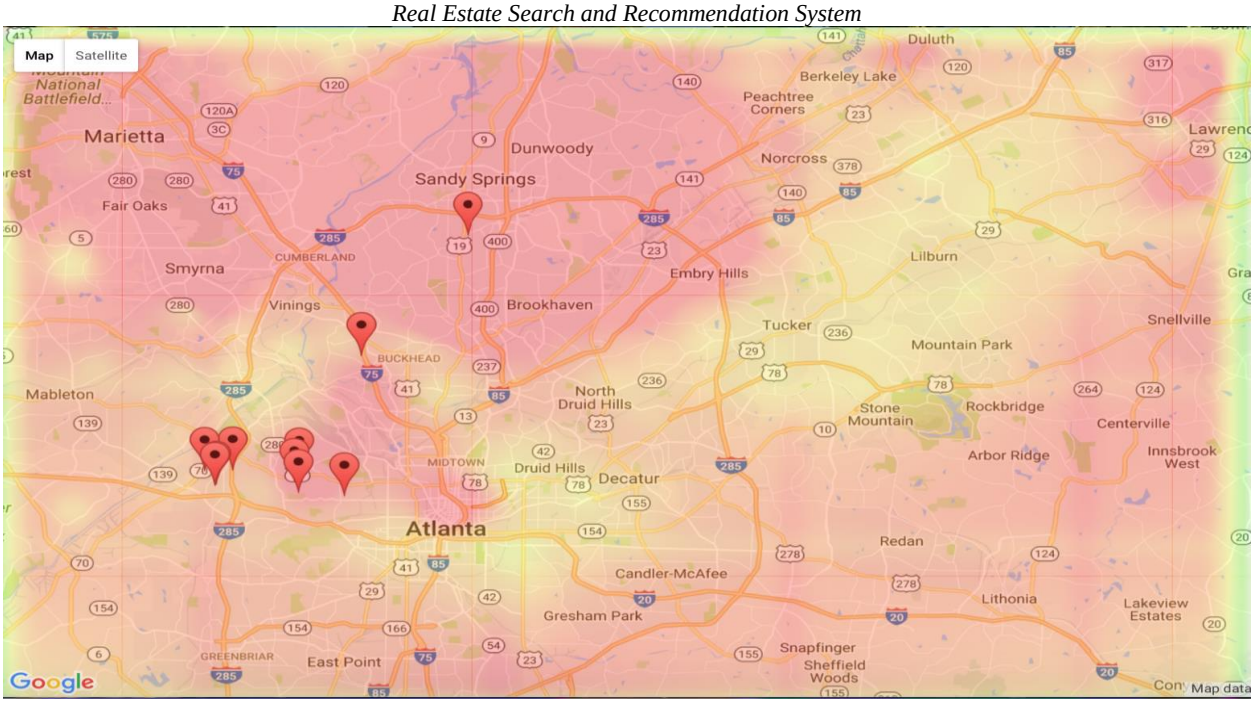

Figure 4.3 Visualization result with default priorities

On this map user can click the marker and then the detailed information of each recommended real estate will be shown. The score of each property is evaluated by comparing its listing price and the price estimated by our model. So that the marker will provide users with the information of the cost-value ratio. And the heatmap can help user find out whether this area is qualified to his priorities.

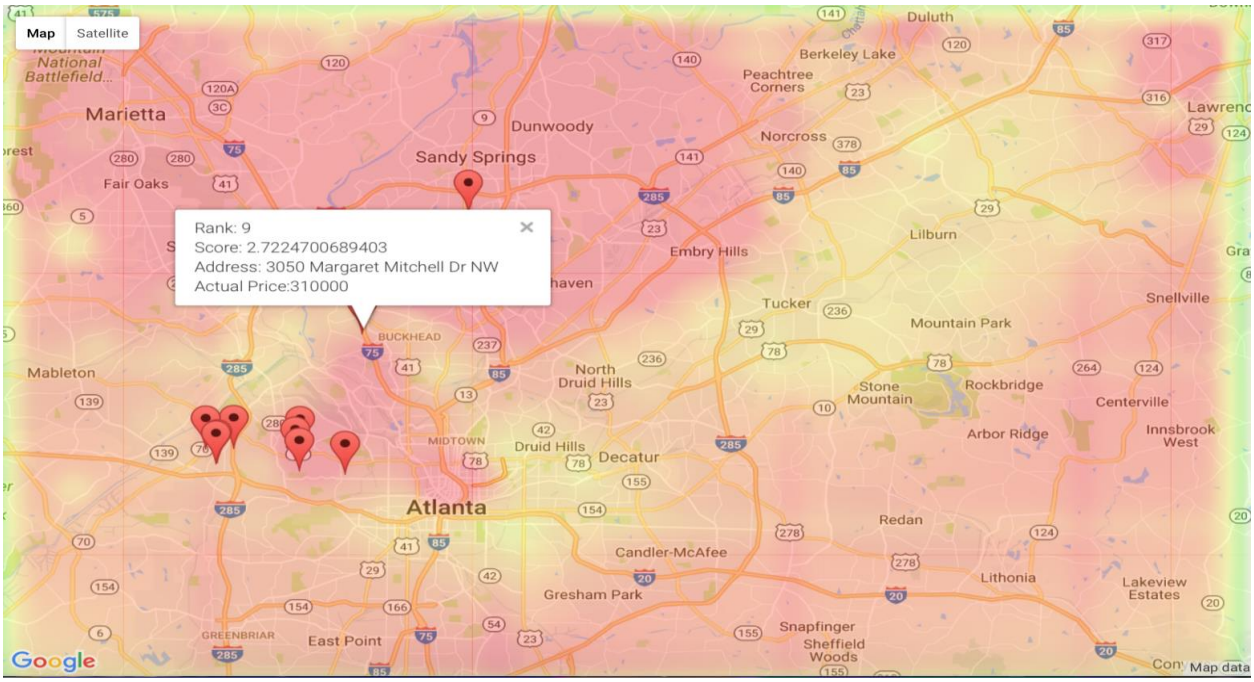

Figure 4.4 Visualization result with detailed information

And several changes are made on the priority settings, the following figures shows the result by setting different priorities.

Figure 4.5 is generated by setting security to 9 and leaving the other factors as default.

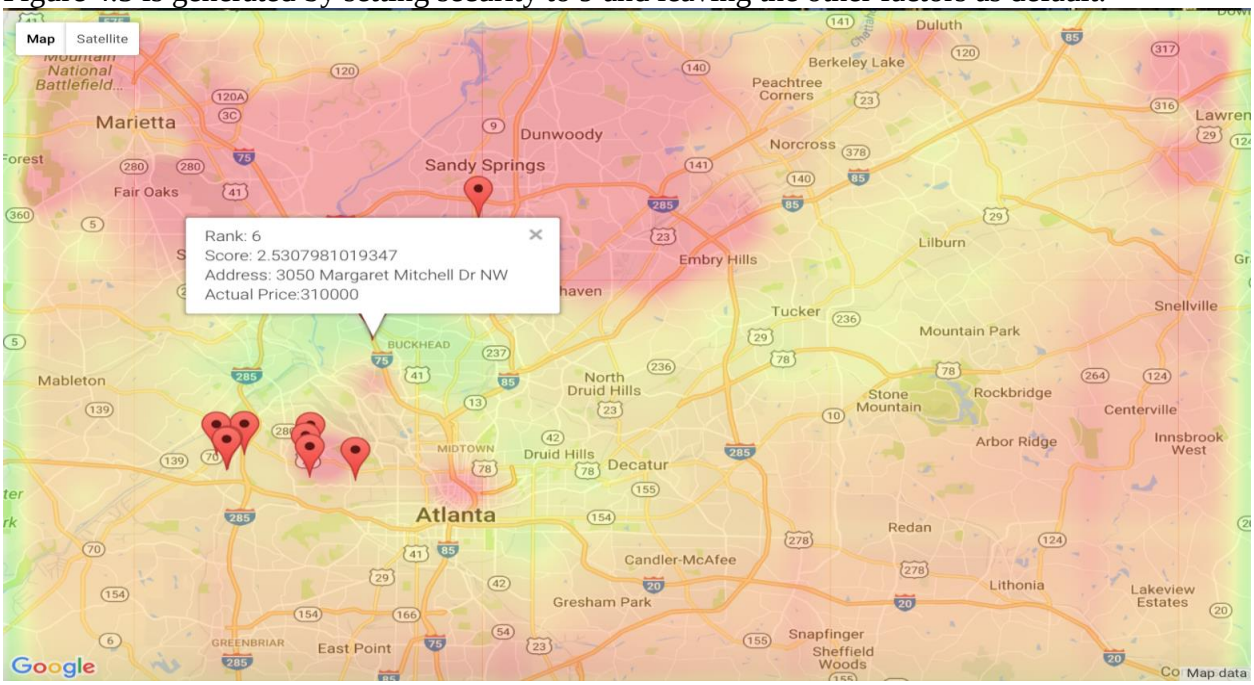

Figure 4.5 Visualization result with security = 9

Figure 4.6 is generated by setting security to 8, Schools to 3, Supermarkets 6 and hospital to 5.

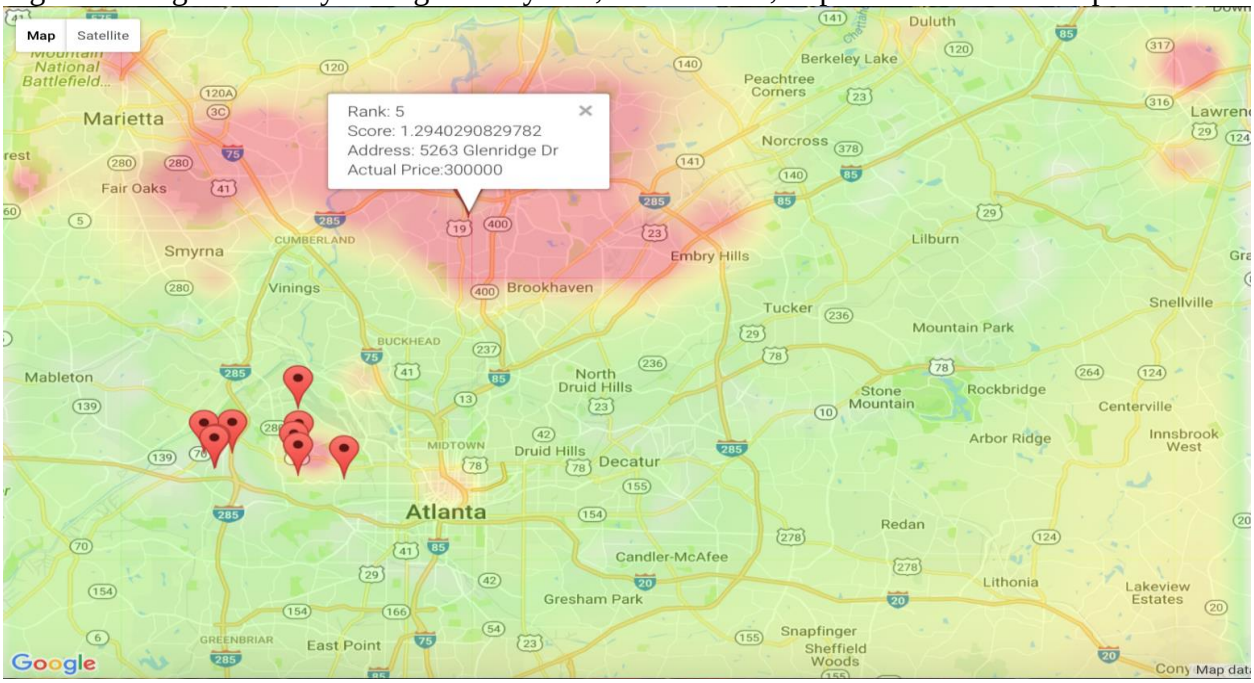

Figure 4.6 Visualization result with security = 8 schools = 3 supermarkets = 6 hospital = 5

There is clear difference between these different settings and the recommendation result will also change accordingly, which means our visualization can help users make decision by providing decent amount of information. And the information provided by this visualization is both intuitive and informative.

5. Conclusion and Discussion

Our project found a suitable model with google fit to estimate the value of a real estate. And the visualization of environmental factors is implemented by google map API. More importantly, the recommendations are not only based on the model. Users can use our interface to get the result of recommendation according to their priorities. The web page is designed well and users can get all the information they need in no time. For the future, more different models can be tested, so that result of modeling can be better.

All team member contributes similar amount of effort.

Reference:

Amenyah, I. D., & Fletcher, E. A. (2013). Factors determining residential rental prices. Asian Economic and Financial Review, 3(1), 39.

Bolton, R. N., & Drew, J. H. (1991). A multistage model of customers' assessments of service quality and value. Journal of consumer research,17(4), 375-384.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

Din, A., Hoesli, M., & Bender, A. (2001). Environmental variables and real estate prices. Urban studies, 38(11), 1989-2000.

Mathwick, C., Malhotra, N., & Rigdon, E. (2001). Experiential value: conceptualization, measurement and application in the catalog and Internet shopping environment☆. *Journal of retailing*, *77*(1), 39-56.

Jin, X., Zhou, Y., & Mobasher, B. (2005, August). A maximum entropy web recommendation system: combining collaborative and content features. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 612-617). ACM.

Kim, C., & Kim, J. (2003, October). A recommendation algorithm using multi-level association rules. In Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on (pp. 524-527). IEEE.

Kong, F., Yin, H., & Nakagoshi, N. (2007). Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China. Landscape and Urban Planning, 79(3), 240-252.

Liao, F. H., Farber, S., & Ewing, R. (2014). Compact development and preference heterogeneity in residential location choice behaviour: A latent class analysis. Urban Studies, 0042098014527138.

Vega, A., & Reynolds-Feighan, A. (2009). A methodological framework for the study of residential location and travel-to-work mode choice under central and suburban employment destination patterns. *Transportation Research Part A: Policy and Practice*, *43*(4), 401-419.

McNulty T L, Holloway S R. Race, crime, and public housing in Atlanta: Testing a conditional effect hypothesis[J]. Social Forces, 2000, 79(2): 707-729.

Miller, C. C. (2006). A Beast in the Field: The Google Maps Mashup as GIS/2. Cartographica: The International Journal for Geographic Information and Geovisualization, 41(3), 187-199. doi:10.3138/j0l0-5301-2262-n779

Smith, K. T. (2012) Using Online Reviews in Creative Selling.

Yuan, X., Lee, J. H., Kim, S. J., & Kim, Y. H. (2013). Toward a user-oriented recommendation system for real estate websites. Information Systems, 38(2), 231-243.

Parasuraman, A., & Grewal, D. (2000). The impact of technology on the quality-value-loyalty chain: a research agenda. Journal of the academy of marketing science, 28(1), 168-174.

Zeng, T. Q., & Zhou, Q. (2001). Optimal spatial decision making using GIS: a prototype of a real estate geographical information system (REGIS). International Journal of Geographical Information Science, 15(4), 307-321.