Visualize and Forecast House Value in US

Team 21: Taoyouwei Gao, Yarui Wang, Ke Zhang

1. Introduction

As the real estate bubble in China or Canada swells larger, there are more foreign investors targeting on the US real estate market. There is a need for these foreign investors to learn more about the US market, to discover more places like Atlanta, GA or Austin, TX, which has a relatively low investment cost but possesses large potential of earnings.

Currently, the most popular online real estate companies do not provide general aggregated statistical information, while the local real estate agents usually do not have the vision of the entire country.

We want to help these investors to invest real estates in the U.S. We aim to provide a summary of real estate information in terms of large scales at all aspects for real estate investors who are seeking for investment opportunities across the country and need more general guidance. We plan to use choropleth map to provide aggregate real estate information at two granularities, state-level and county-level. Once a foreign investor decides his target state based on our choropleth map for states and click the state's name, our website will display historical real estate data of this state. Furthermore, our website will have a zoom-in choropleth map for counties in this state. Similarly, our website will display historical real estate data of this county once the user clicks the county's name.

Through our website, foreign investors can get information progressively and find their most optimal investment location across the U.S

2. Problem Definition

1) Visualize house value

Our goal is to provide and visualize general information of house value in most states and counties in US. However, there are many different kinds of data to be presented. It is difficult to display thousands of data efficiently, accurately, concisely and aesthetically. To solve this problem, we use D3 choropleth map to show price distribution. Zoom-in function is also designed to show detailed county data. Historical data is shown using interactive line chart design.

2) Forecast house value:

Our goal is to forecast the growth potential of each area in US. Nonetheless, it is not easy to choose right methods or use proper features to achieve that goal. Also, the method to validate the prediction is crucial after we forecast the house value. To solve this problem, we try four different machine learning algorithms and optimize hyper-parameters. Several metrics are compared to get the best fit.

3. Survey

- **Paper 1: Redfin—A Developing Vertical E-Commerce Model**
  Paper 2: Zillow—Online Media Tycoon in US Real Estate Brokerage Industry
  These two chapters discuss how online brokerage websites like Zillow and Redfin beat the old traditional real estate brokerage industry in the U.S.
  We can have more ideas about what information we should include and what kind of user interface we should design in our website. However, these two chapters focused on data between 2005 and 2013, and authors did not notice the increasing power of foreign investors in the U.S housing market.

- **Paper 3: Factor and regional difference analysis on real estate price**
  This paper discussed what and how factors, especially location factors, like state, affect real estate price in the U.S. Authors used cross sectional data and conclude that there is a significant difference between the real estate market of Central U.S and that of the Western U.S. Results in this paper support our idea that state is a location factor that affect real estate price. However, this paper did not explore how counties affect the real estate market even for the same state. We will explore this factor in our project.

- **Paper 4: HomeSeeker: A visual analytics system of real estate data**
  Book 5: Private Real Estate Investment: Data Analysis and Decision Making
  Paper 4 and Book 5 both developed interactive visual analytics systems to serve users with different backgrounds of the local real estate market and meet different degrees of user requirements. Heterogeneous data was integrated from different channels into a location-centered integrated real estate dataset which is the reference for handling our data.

- **Paper 6: Method for determining values in real estate appraisal: comparing between linear regression model and fuzzy logic**
  This paper used Fuzzy logic method to assess the real estate values and propose a new method to model the real estate market. The inspiration of the paper is that linear regression method and Fuzzy logic method can predict the estate values. The shortcoming is the fuzzy logic method doesn't work in general situations.

- **Paper 7: Combining quantitative and logical data cleaning**
  This article designed a combined data cleaning method including both quantitative and logical approaches. Using this method, it will help us to detect subtle data problems and acquire meaningful statistical properties. The shortcoming is that this method may be difficult to apply and overreaching for our project.

- **Paper 8: Analysis of categorical data by linear models**
  This article provided an approach to analyze categorical data by linear models. We need this method in our project since part of our housing data is categorical (for example, housing type). But the result may not be specific enough to generalize useful analysis.

- **Paper 9: Visual framework for big data in d3.js**
  This paper showed a method of big data visualization using D3. Since our project deals with very large dataset, this approach will be very useful in our visualization step. However, some of the mechanisms in this article are too complicated to use in our project.

4. Proposed Method

1) **Our approach has several innovations listed below:**
   a). Two-level zooming functionality
   b). Applying Machine Learning algorithm based on large scale aggregation
   c). Interactive customizable filtering options
   The detailed information of these innovations will be mentioned below.

2) **Description of our approach:**
   a). Data cleaning:
   The data we acquired from the internet is incomplete to a certain extent. For example, not all counties in US have the complete data from 1996 to 2018. Therefore, we have to perform the cleaning process to prevent these missing data from interfering our analysis and visualization. We use python to clean the raw data. In details, Python csv package is used to read all raw data, check each house price value and remove the unreasonable missing, large or small values which are far away from average value in the same location and at the same time.

   Also, the original dataset doesn't contain ID number (FIPS code) of each state and county which is essential in D3 visualization. Therefore, we processed the data by adding the ID number to each of them with python pandas and csv package. According to the names of state and county, the custom ID in the original data is replaced with ID number (FIPS code).

   b). Database design and data loading:
   We decided to set up a backend database for our website for the following two reasons:

I. It decreases the loading time of our page and scales if our dataset grows larger in the future. The current size of our dataset is around 25 MB because we only support state and county level aggregation and we only provide median sale price data. However, the size could grow exceptionally if one day we support city level aggregation data or provide more aspects of the housing market. By that time, reading the entire dataset to memory will be unrealistic.

II. It well supports our various frontend use cases because of its queryable capability, so that we can get the data that user needs by directly querying the database instead of processing it on page.

Based on the potential use cases of our database, we designed it such that it contains the following three tables:

- Table Region (**id**, name, size_rank)
- Table Housetype (**id**, category, housetype)
- Table Median_sale_price (*region_id*, *house_type_id*, time, price)
  (Primary keys are bold and foreign keys are italic)

The benefit of this schema is that we can easily retrieve data for certain housetype of a region for a specific time frame. We further processed the data that cleaned with Pandas as we have to restructure it to fit in to those tables and parse from various date format to epoch since we want to query our time column. Then we loaded them to the above tables in SQLite.

c). Algorithm applied to predict future growth potential:
We applied Machine Learning algorithm to find the relationship between housing price and other factors, like time, location, price_per_sqft and so on. Based on the relationship, we tried to predict the housing price in next month and then get the growth potential of that area.

The algorithms we applied are linear regression, SVM, random forest and multilayer perceptron which predict future house value based to the previous dataset. The raw data is partitioned into 70% training data and 30% testing data.

The python sklearn package is imported to get above mentioned ML models within the train data. After getting ML model parameters, it's tested in testing data to check whether it's reasonable. The mean absolute error is chosen to compare the accuracy of mentioned ML models. The hyper-parameters are also optimized by GridSearchCV method.

d). Visualization - Website Design

I. Flask Microframework:

We built our website application on top of Flask Microframework, which allows us separate our view layer and controller layer. With Flask, we are able to easily pass in the user input from the front end in a URL to the controller. In the controller, we can connect to and query the database based on the user input, further process the query result and hand back the result to the front end. This is basically the workflow we adopted while populating our frontend visualization with user requested data.

II. Choropleth map by D3:

After data cleaning and analysis, we used D3 to visualize them. Create a D3 choropleth map to provide aggregate real estate information at two granularities, state-level and county-level, from April, 1996 to February, 2018. The choropleth where each state or county is filled with a color proportional to its most current housing price or growth potential based on which option is chosen in the select box.

III. Two-level zooming functionality:

We also designed a two-level zooming function to show detailed information of each county in each state. To acquire more information in one state, users can double click on any of the state and the choropleth will zoom into this state and show its counties in detail, where each county is colored by its median housing price or growth potential depending on which option is chosen in the select box.

IV. Select box functionality:

In order to give investors a full picture of house value, we decided to introduce a select box which allows them to choose display content between current price and growth potential. Simply choose different item in the select box, the choropleth map and the legend will change accordingly.

V. Pop-out line chart showing historical data:

To show the historical data of certain state and county, we decide to use pop-out line chart to show the historical price trend. By single click on any of the state or county, we will show the line charts on the bottom side of the page about housing price in this state, which is the median housing price trend in history.

VI. Customizable filtering options and interactive comparing capability:
Additionally, users can also customize what statistics they want to see by clicking the option boxes we provide. They can select the start and end date of display or different house categories or housetypes. We also provide a compare checkbox that users can check on when they want to compare the sale price trend across different housetypes.

5. Experiments and Evaluation

1) Questions our experiments are designed to answer:

a) Which machine learning algorithm is best fitted for our project requirement?
b) Which metric is best for evaluating ML performance and why?
c) Is our website functionally working and user friendly?

2) Experiment

a) Feature engineering and selection

I. The features we picked directly from Zillow Research includes: Median_sale_price, Median_listing_price, Median_listing_price_per_sqft, Days_on_Zillow, Inventory_number, Listings_with_pricecut_in_percentage, Median_price_cut_percentage.

II. From the above existing features, we also derived and experimented with the following engineered features: Growth_of_sale_price_next_month, Median_listing and median_sale_price_difference, isGrow(1, 0)

III. For feature selection, we did our experiments primarily on Microsoft Machine Learning Studio. We built up a workflow that runs linear regression model with cross validation module. Then we first ran the workflow with single predictor one by one, which gave us the linear correlation of every predictor and our target. Based on the single predictor linearity, we then tried various combination of predictors.

IV. We also tried the stepwise module in Matlab, which selected the best feature combination based on p-value. The result of running stepwise showed days_on_Zillow, sale_count, median_price_cut and median_listing_price are the best feature combination.

b) Model selection and hyper-parameter tuning

I. We experimented with Linear Regression, SVM, random forest and multilayer perceptron models. For SVM and random forest, the different parameters are compared by GridSearchCV method.

II. The hyper-parameters ‘C’ and ‘kernel’ in SVM and hyper-parameters ‘n-estimators’ and ‘'max-depth’ are selected. The options for ‘kernel’ are ‘rbf’ and ‘linear’, when the range for ‘C’ are 0.1, 1,10,100 and 1000. As for number of estimators and max depth in random forest, their range are 10,30,50,70 and 3,5,7,9,11 separately.

3) Evaluation

a) Evaluation for ML algorithms:

The mean absolute error which is expressed in $MAE = \frac{1}{n}\sum_i \left|y_{pre} - y_i\right|$ is applied to evaluated the performance of linear regression, SVM, random forest and multilayer perceptron model. Firstly, for train data error, among these four ML models, the linear SVM performs best whose MAE is only 56, while others’ errors are 4654, 2402 and 138005. Nevertheless, for test data error, the linear regression and SVM perform nearly same better whose values are 4242 and 5558 while random forest and MLP’s MAR are 10067 and 95155. From the comparison, the SVM model perform best to predict the future house price.

To improve the accuracy of SVM model, GridSearchCV method is utilized to train it for getting relative best hyper-parameters. The kernel value tested are ‘rbf’ and ‘linear’, and the C values tested are 0.1,1,10,100. The result shows that the combine best parameters are ‘linear’ and 100. Subsequently, the selected parameters are applied to SVM model to get the new train mean absolute error and test mean absolute error which are 45 and 4249. The decrease of errors indicates the optimization of GridSearchCV method works.

b) For website functionality, we did the following 2 phases of testing:

I. Functional Testing:

Basically, we manually tested all functionalities we provide with our website. We checked if all links, buttons and boxes are clickable, as well as if all onClick actions like mouse-over, mouse-click are functional. We also validated the data we displayed were as expected.

II. Usability Testing:

We listed a list of potential use cases of our website and evaluated how easy it was for us to navigate on our website under each use case. We found our website was mostly intuitive to use except for the compare box, which could be a little confusing as unchecking the compare box also serves as cleaning chart.

We then proposed and implemented an equally feasible approach, which made an additional explicit clean chart box. To decide which workflow to adopt, we conducted an A/B Testing on 8 volunteers, each of whom were asked to perform three use cases that involved using the compare box and rate their experience at a scale of ten. The average score we got from the group using with the explicit clean chart box was 8.25, while the score we got from the other group was 8.5. We concluded that both comparing workflow were equivalently feasible and decided to use the version without the explicit clean-chart as the design was cleaner.

6. Conclusions and Discussion

This project successfully applied flask and D3 to visualize house value using choropleth map and line chart. Also, machine learning algorithms were implemented to forecast house value in US.

For future estate value prediction, ML algorithm can be applied and SVM model performs best among linear regression, random forest and multilayer perceptron when using mean absolute error as basis. Besides, GridSearchCV method could address the overfitting problem in SVM model which originally had tiny training data error but large testing error. In a certain range, the experimental results show that the optimal parameters are ‘linear’ kernel and 100 ‘C’.

Although we have a good working project, there is still some room for improvement. For example, some data regarding house value is incomplete or hard to find. If we had those data, we can perform a better prediction with the application of machine learning. Also, we found the sale_price_growth data spreads in a really large range but densely distributed in the middle. It might improve model performance by applying some scaling functions. Moreover, some functions in D3 visualization, like cross state or county comparison, cannot be achieved due to high difficulty and limited time. For future development, those aspects should be most focused on.

7. Distribution of Team Member Effort

All team members have contributed similar amount of effort.

References:

Ba, Shusong, and Xianling Yang. "Redfin—A Developing Vertical e-Commerce Model." *“Internet Plus” Pathways to the Transformation of China’s Property Sector*. Springer, Singapore, 2016. 85-100.

Ba, Shusong, and Xianling Yang. "Zillow—Online Media Tycoon in US Real Estate Brokerage Industry." *“Internet Plus” Pathways to the Transformation of China’s Property Sector*. Springer, Singapore, 2016. 67-84.

Li, Li, Lei Xiao, and Jia-wen Xiao. "Factor and Regional Difference Analysis on Real Estate Price." *Technology for Education and Learning*. Springer, Berlin, Heidelberg, 2012. 317-323.

Li, Mingzhao, et al. "Homeseeker: A visual analytics system of real estate data." *Journal of Visual Languages & Computing*45 (2018): 1-16.

Brown, Roger J. *Private real estate investment: data analysis and decision making*. Elsevier, 2005.

Malaman, Carolina Scherrer, and Amilton Amorim. "Method For Determining Values In Real Estate Appraisal: comparing between Linear Regression Model and Fuzzy Logic." *Boletim de Ciências Geodésicas* 23.1 (2017): 87-100.

Prokoshyna, Nataliya, et al. "Combining quantitative and logical data cleaning." *Proceedings of the VLDB Endowment*9.4 (2015): 300-311.

Grizzle, James E., C. Frank Starmer, and Gary G. Koch. "Analysis of categorical data by linear models." *Biometrics*(1969): 489-504.

Bao, Fan, and Jia Chen. "Visual framework for big data in d3. js." *Electronics, Computer and Applications, 2014 IEEE Workshop on*. IEEE, 2014.

Appendix

Heilmeier questions

1 Object

We want to help foreign investors to invest real estates in the U.S.

2 Done work and limits

Popular real estate sources like Zillow and local agents target for people who already know which city or housing market to invest in. But for those people who don’t know which city or state to invest, like foreign investors, who don’t care about locations as long as it is a good investment, they probably need resources to get some general ideas about the real estate information for different states and cities across the U.S. We can easily find individual listings from Zillow, but we are unable to find any information in terms of a higher granularity, like the average housing price or the annual total number of listings for a city or a state.

3 Innovation

In our website, we will use choropleth map to show overall real estate values in two levels, states and counties. Once a foreign investor decides his target state based on our choropleth map for states and click the state’s name, our website will display historical real estate data of this state. Meanwhile, our website will have a zoom-in choropleth map for counties in this state. Similarly, once the user chooses his target county among the state, our website will display historical real estate data of this county. Through our website, foreign investors can get information progressively and find their most optimal investment across the U.S.

4 Target Users

This project is made for real estate investors who want to buy houses and get as high as possible earnings.

5 Difference and Impact

The complement of the project could help investors save much time to make an investment and narrow the range of selection. The more important impact of it is to provide the linear regression prediction of house value in different locations. We can utilize the users ‘comments and the future house prices to validate the project.

6 Risks and payoffs

The risks are mainly located at the incomplete data which means not all estate prices in every county are collected.

The payoffs are to provide the analysis of estate price in almost all states and counties successfully.

7 Cost

The cost is mainly time and human power.

8 Duration

It will take approximately one month to present the initial product.

9 Midterm and Final "exams" to check

The progress will be measured through the completeness of our products. The midterm is how well we acquire, clean and analyze data, the final is how well we visualize the results through D3.