# House Price Prediction in Atlanta

## Objectives

(Q1 in Heilmeier questions)

Motivated by various applications in the real estate market, this project aims at developing a novel predicting model for house price prediction in Atlanta. This model will be implemented in a web application to serve as a functional additive tool in the prediction of future housing prices and the establishment of real estate policies.

## Literature Review

(Q2)

The modeling of house price can start with linear models. Two common methods for calibrating linear regression are least squares estimation (LSE) or maximum likelihood estimation (MLE) [1]. House price depends on both location and time. Spatial effects include two parts, spatial dependence and spatial heterogeneity. For spatial effects, locations are grouped into sub-regions [2]. STAR model [3], which is based on Hedonic regression and considers the spatial and temporal factors, could perform better than ordinary least square model [3].

The hedonic pricing models, which are commonly used to estimate house prices, regard each house as integrated goods of bundles of attributes. By estimating the value of each attribute embodied and employing regression analysis, the total property value can be estimated [4]. This model can reveal the effects of attributes in the house price changes. However, one limitation of this method is that it is hard to efficiently capture the independent variables. To improve the prediction accuracy, more recent studies have applied machine learning approaches to build hedonic models and select independent variables [5-7], which provide a guide for this project.

The study of house price is also dependent on the commuting cost. One classical model is standard urban model (SUM), which assumes one center for a city and correlate the house price with the distance to the center of the city [8]. Larger cities have higher house price volatility due to lower elasticity of resources and there is higher spatial heterogeneity for the submarkets in larger cities [9]. In suburban regions, the house supply has high elasticity and the price tends to be stationary [10]. In this project, the goal is to predict house price in Atlanta, which is expected to have high price volatility. To have an accurate prediction, as mentioned earlier, it is necessary to have a proper submarket grouping, assuming that the house supply is consistent in each submarket.

House prices might vary even under the same conditions. Thus fuzzy linear regression is introduced to handle the fuzziness of such systems[11]. Peters [12] improved this model by introducing fuzzy intervals to ensure algorithm's robustness even with outlier. Gerek [13] proposes an adaptive-fuzzy inference system (ANFIS) for house price forecasting in which the fuzzy-based models are constructed by implementing the grid partition and subtractive clustering identification approaches.

A variety of machine learning algorithms are proposed to develop a prediction model for housing prices. Park and Bae [14] studied development of housing price prediction models with four classifier methods; C4.5, RIPPER, Naïve Bayesian, and AdaBoost.

Support vector machine (SVM) is another machine learning technique for problems with limited sample learning and nonlinear relation. Wang et al. proposed a real estate price predicting model based on particle swarm optimization (PSO) and SVM to describe the real estate price prediction model as a mathematical mapping problem on pattern matching [15].

The house price prediction model can be implemented in a location-based recommendation system (RS). Such system provides predicted values based on location information and hence increases the accuracy. Major steps for establishing the system include determining the locations, as well as concerned variables of spots near the location, performing collaborative filtering on the spots, and giving predictions based on results [16]. Park et al. [17] developed a RS taking in spatial variation using Bayesian Network [18, 19] which can efficiently provide the conditional probability distributions of results.

All the above-mentioned references are related to this project since they address various aspects of the objectives.

## Approach

(Q3)

A novel comprehensive model will be proposed in this project to take some additional parameters, such as crime rate and neighborhood amenities, into account for house price prediction, and implemented as a web application.

Accordingly, proper machine learning libraries will be used to develop the back-end in Python. Data are obtained from various sources like Kaggle. Further data scraping is needed to gather sale prices and property locations by using Zillow and Google map API. OpenRefine will be used for data cleaning. As the user interface for this web application, a front-end will be developed in HTML, CSS, and JavaScript, with D3 used for visualizing the outputs. Back-end will be developed in Flask.

This combination can serve as coherent prediction tool.

## Applications

(Q4, Q5)

The deliverable web application can be very useful for a wide range of stakeholders in Atlanta real estate market including real estate agents, appraisers, policy makers, mortgage lenders, investors, property developers, and existing and potential homeowners. Such tool will reduce the amount of manual works for users who consider buying or selling a house and want to consider more details in their predictions.

## Challenges

(Q6, Q7, Q8)

Risks: The risks include failure in gathering proper data and the over-fitting issues. The potential problems in employing Google API may slow down the progress.

Payoffs: A successful product can be offered to available price predicting systems or launched as an independent house price predicting tool for commercial purposes.

Costs & Time: It is anticipated that by spending 300 man-hours.

**Evaluation**
(Q9)

To evaluate the back-end and proposed predictive model, cross validation approach will be employed in this project. The front-end will be continuously tested to ensure the desirable performance.

**Schedule and Work Distribution**

| | Group Member | Responsibilities | Time (hour) |
|---|---|---|---|
| 1 | Zichen Wang | Machine Learning | 60 |
| 2 | Wenqing Shen | ML Implementations | 40 |
| 3 | Yixing Li | Back-end | 60 |
| 4 | Dong Gao | Data Cleaning, Scraping | 40 |
| 5 | Xiangyi Yan | Data Visualization | 40 |
| 6 | Shahrokh Shahi | Front-end | 60 |

**References**

1. Seber, G.A. and A.J. Lee, *Linear regression analysis*. Vol. 329. 2012: John Wiley & Sons.
2. Liu, X., *Spatial and temporal dependence in house price prediction.* The Journal of Real Estate Finance and Economics, 2013. **47**(2): p. 341-369.
3. Pace, R.K., et al., *Spatiotemporal autoregressive models of neighborhood effects.* The Journal of Real Estate Finance and Economics, 1998. **17**(1): p. 15-33.
4. Limsombunchai, V. *House price prediction: hedonic price model vs. artificial neural network*. in *New Zealand Agricultural and Resource Economics Society Conference*. 2004.
5. Bajari, P., et al., *Demand estimation with machine learning and model combination*. 2015, National Bureau of Economic Research.
6. Yoo, S., J. Im, and J.E. Wagner, *Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY.* Landscape and Urban Planning, 2012. **107**(3): p. 293-306.
7. Yu, D., & Wu, C. (2006). *Incorporating remote sensing information in modeling house values.* Photogrammetric Engineering & Remote Sensing, 72(2), 129-138.
8. Muth R. *Cities and housing: The spatial patterns of urban residential land use.* University of Chicago, Chicago. 1969;4:114-23.
9. Bogin, A., W. Doerner, and W. Larson, *Local house price dynamics: New indices and stylized facts.* Real Estate Economics, 2016.

10. Glaeser, E.L., et al., *Housing dynamics: An urban approach.* Journal of Urban Economics, 2014. **81**: p. 45-56.
11. Asai, H.T.-S.U.-K. and S. Tanaka, *Linear regression analysis with fuzzy model.* IEEE Transaction Systems Man and Cybermatics, 1982. **12**(6): p. 903-07.
12. Peters, G., *Fuzzy linear regression with fuzzy intervals.* Fuzzy sets and Systems, 1994. **63**(1): p. 45-55.
13. Gerek, I.H., *House selling price assessment using two different adaptive neuro-fuzzy techniques.* Automation in Construction, 2014. **41**: p. 33-39.
14. Park, B. and J.K. Bae, *Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data.* Expert Systems with Applications, 2015. **42**(6): p. 2928-2934.
15. Wang, X., et al., *Real estate price forecasting based on SVM optimized by PSO*. Optik-International Journal for Light and Electron Optics, 2014. **125**(3): p. 1439-1443.
16. Horozov, T., N. Narasimhan, and V. Vasudevan, *Location-based recommendation system*. 2006, Google Patents.
17. Park, M.-H., J.-H. Hong, and S.-B. Cho. *Location-based recommendation system using bayesian user's preference model in mobile devices*. in *International Conference on Ubiquitous Intelligence and Computing*. 2007. Springer.
18. *Bayesian network.* February 28, 2018]; Available from: https://en.wikipedia.org/wiki/Bayesian_network.
19. Friedman, N., D. Geiger, and M. Goldszmidt, *Bayesian network classifiers.* Machine learning, 1997. **29**(2-3): p. 131-163.