Home Value Predictions in Greater Boston Area

A home is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important. While many individual homebuyers are less sensitive to price forecasts, small errors in price prediction can have systemic negative effects in the economy as a whole. Accurate prediction makes it easier to understand which features would influence the final property price.

Problem statement

Real estate prices are very much dependent on factors that are not easy to control. Analyzing broader market conditions and specific property determinants in order to establish how property values may change over the course of time are utterly important. Massive data can be obtained about the current market situation, which demands using powerful machine learning algorithms in order to predict with high precision and in a reasonable time frame.

Goal: Build a home valuation algorithm from the ground up, using both, internal and external data sources. Compare final predictions against other popular algorithms estimated values.

Data

In this project you will be provided with a list of real estate properties in a Greater Boston Area. The dataset consists of information about 2016 and 2017 properties and contains a vast amount of features - information about the property, historical data, taxes, exclusive brokers etc. Additionally, you will be provided with the dataset which includes listings from 1/1/2018 until 1/31/2018 (approximately 23000 properties) to test your model at the end of the tracking period, that will align with the end of your final project.

Additional data challenges

1. Students will be challenged to scrape the Zestimate historical values on a properties for 2016/2017 data and compare predictions against both, Zestimate values and actual sale prices at the time.
2. In the real estate industry, pictures can easily tell people how the house looks like. Students will be challenged to use the information from provided dataset (street name, street number, and zip code) to scrape properties descriptions and pictures from Zillow web page (some listings have walkthrough videos that could be taken into an account).
3. For the given house pictures, people can easily have an overall feeling of the house - what is the construction style, how the neighboring environment looks like etc. Use Google Street View image API to submit address and scrape Google images about the building (exterior properties, street quality, neighborhood, near-by attractions).

4. Optionally, you can include/scrape the data from Zillow webpage about:

    - Market temperature (Buyers' Market vs. Sellers' Market)
    - Nearby schools
    - Potential listings (make-me-move and pre-foreclosure)
    - Comparable homes and competition for a listed property
    - Market activities

High-level Project Goals

1. Propose a method that efficiently handles all the factors in residential real estate. Train/test your model on 2016/2017 data and report your results. Compare with past Zestimate values and try to improve Zillow prediction.
2. Include the interior information about the property into your model (property images). Does including images improve the predictive power of your model?
3. Include exterior information (Google images). How does including Google images affect your predictions?
4. Your final milestone will be predicting the prices based on January 2018 data. Use the above mentioned steps and derive the best possible combination (features/models) to compare your predictions with real prices. How precisely will you be able to predict days on the market and price for every unit?

Resources and References

1. Zillow Kaggle Competition: https://www.kaggle.com/c/zillow-prize-1
2. Historical Estimate Values: https://www.zillow.com/advice-thread/Can-I-get-historical-zestimate-values-from-the-API/471215/
3. How to scrape on zillow.com: https://www.scrapehero.com/how-to-scrape-real-estate-listings-on-zillow-com-using-python-and-lxml/
4. Using Python to scrape Google Street images:https://andrewpwheeler.wordpress.com/2015/12/28/using-python-to-grab-google-street-view-imagery/
5. A. V. Dorogush, A. Gulin, G. Gusev, N. Kazeev, L. Ostroumova Prokhorenkova, A. Vorobev, Fighting biases with dynamic boosting, CoRR, 2017.
6. T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 785-794, 2016.
7. Q. You, R. Pang, L. Cao and J. Luo, Image-Based Appraisal of Real Estate Properties, in IEEE Transactions on Multimedia, vol. 19, no. 12, pp. 2751-2759, 2017.