Project Proposal- Home Purchase Assistance

Shibo Zang, Yangwen Wan, Yingjie Li
{sz428, yw762, yl2552}@cornell.edu

The Problem and its Importance

When someone is looking for a house to buy, he or she would be really interested in the following questions: If the current price of the house is fair and acceptable? Should I consider buying a house recently or wait for a while? Our goal is to provide suggestions to the users that if the housing prices listed by an agent or on a website is reasonable, and would the price go higher or lower in the future.

According to statistics, there were 5,250,000 existing homes and 510,000 newly constructed homes were sold in 2015, which means there were around 11 transactions were completed on home trading per minute. In addition, large real estate marketplace like Zillow has more than 10,000,000 listings nationwide. We can see that the real estate market is very active and is critical to people's daily life. However, there is a lack of a powerful assistant tool that is able to help users to decide when to buy a house and which houses are worthy of consideration. By analyzing the data of past transactions and current local listings, we can provide suggestions to the users that if the price listed by the agent or on a website is reasonable, and would the price go higher or lower in the future.

Datasets Description

Our raw data is the annualized properties sales from the year 2003 to 2015, provided by New York City Department of Finance. So, we have 13 years of properties sales data and information on those properties for all 5 boroughs of New York. Each data entry represents one property sold and it's detailed geographical and numerical information. Geographical features include Borough, Neighborhood, Block, Lot, Address, and ZipCode. These features not only give us exact information about the properties but also provides us the chance to join with other data from outside sources. For example, ZipCode, as a categorical but numerical feature, can be used to join with income status, crime rate, and other possible social and economic data of the same area. As for the properties itself, we have Building Class (Building Class Category, Building Class at Present, and Building Class at Time of Sale), Tax Class (Tax Class at Present, and Tax Class at Time of Sale), Units (Residential Units, Commercial Units, and Total Units), Size (Land Square Feet, Gross Square Feet), and Year Built. Finally, we have the sales price and time sold for each property.

Our goal is to combine the raw properties sales data with crime, social, and economics data to build a model that predicts the price of a given property. The results of which can be used in our "Zestimates" services.