Nicole Stone

Domain Background

This project is focusing on real estate data, specifically past real estate sales and current real estate listings in Los Angeles County. Within the domain of real estate, this project will be considering how machine learning can be used to aid in identifying quality investment properties which can be purchased below market value, generally due to the property being in a distressed state, and then renovating the property in order for it to be sold at a profit. This process is commonly referred to as a "flip" within the domain of real estate.

Problem Statement

Finding properties which present a successful margin for a "flip" makeover can be challenging. The best indicator of a profitable real estate investment is a property which can be acquired below 70% of the value at which it could potentially be sold if it were in like-new condition, also referred to as "after repair value" (ARV) [1]. The total cost of the anticipated repairs needs to be accounted for as well, but that requires a person to scope, design, and purchase. That overall renovation cost is then subtracted from the 70% of ARV to provide a maximum purchase price for this investment. A best-case scenario would be purchasing a like-new home that needs no repairs or updates and can be sold immediately as-is for a profit. This scenario would be a house at 70% or less of the ARV. Therefore, our model should target any homes that are listed at 70% or below the estimated value for that particular property.

Nicole Stone

Datasets and Inputs

This project uses data from the Redfin database of MLS listings for the Santa Clarita region of Los Angeles County. Listing information of recent transactions from the past three months will be downloaded, as well as all current listings for the region.

Solution Statement

In order to estimate the price at which a home could be sold post-repairs, a machine learning model could be fit on real sales data for the past 3 months, as reported by Redfin.com. This estimator can then be used to predict at what price a particular home should be able to sell. Then with new listings, comparisons can be made between the newly listed property's list price and the estimated sale value. Any properties that are listed at or below 70% of the estimated sale value can then be recommended for further inspection as potential "flip" investments.

Benchmark Model

Without the aid of machine learning models, list prices are determined by realtors and then "true" house values are determined by an appraiser using a thorough rubric of property characteristics. Appraiser evaluations cost several hundred dollars for each property, thus a rough evaluation is completed by the realtor using comparable recent sales or other listings in order to determine a proper list price. The most basic way a realtor determines how much to list a house is to consider current listings and recent sales for the immediate neighborhood of the property they are considering. An average price per square foot of house is determined and then multiplied by the square footage of the property they are listing. This is generally a pretty

good starting place, and many realtors list properties solely on this number. A more accurate estimate can be determined by considering lot size, number of bedrooms, and number of bathrooms in comparison to recent sales and then adjusting the list price accordingly. However, for the benchmark in this design, it will suffice to consider price per square foot as an average for the zip code within which the home is located.

Evaluation Metrics

A new model can be tested by considering how well it would be able to predict past sales prices given the listing information. This means some of the past sales data will need to be withheld from training the model in order to be used during testing the model. The accuracy of the model could then be compared to the benchmark value for that property (price per square foot) to see which is more accurate at predicting the true sale price.

Project Design

This project will implement an estimator for property values using an XGBoost model. The samples will be randomized and then split into training and testing groups. The training data will be cross validated and an accuracy score will be determined for the model's performance on the test samples. If the model performs better than the benchmark, it can be used to help find potential "flip" properties by separating out any properties that are listed at or below 70% of their predicted value for further investigation.

References

[1] https://www.biggerpockets.com/blog/2014-02-14-70-rule-bible