Housing Market Pricing Analysis – Preliminary Analysis

Scott Breitbach, Pushkar Chougule, and Rachel Nelson

Bellevue University

**Abstract**

Following the impact of the COVID-19 pandemic, the economy is looking get back on track. One of the leading indicators of recovery is the housing market demand, supply and prices. In this paper, we discuss the methods and results of a detailed preliminary analysis of the housing market and determine what all different factors, if any, contribute to the pricing of housing. We will leverage the findings of our preliminary analysis to create a housing price prediction model.

**Background**

In order to go about finding a proper data set, we looked at different options including scraping data from real estate websites like Zillow or Trulia, or reviewing various existing data sets already available online. While scraping Zillow data could provide us up-to-date information, it also presents its own challenges in obtaining a consistent set of features about the homes.

Due to having an extensive collection of factors and boasting 81 columns of data, separated into training and test sets, along with a currently active Kaggle competition, we decided to move forward with the data set found On Kaggle titled “*House Prices - Advanced Regression Techniques Predict sales prices and practice feature engineering, RFs, and gradient boostin*g”.

# Preliminary Analysis

## Methods

Using the CRISP-DM methodology to data analysis, we started with trying to understand the housing market, which has been increasing at a record setting pace (2021, Smart). We looked through the data set to understand the factors available. We hypothesized that certain factors would have an impact on the price of the home, including square feet, lot size, and neighborhood.

We combined the training and test data sets in Python in order to perform some cleansing of the data. We reviewed summary statistics including the count, mean, standard deviation, minimum, maximum, as well as quartiles on the numerical columns data. In many of the columns, the missing data represented the absence of something, like no alley access or no basement. In these instances, we simply imputed ‘None’ as a category, or ‘0’ in instances like square feet of basement when there was no basement. Remaining columns that had missing data were either filled in using mean or the record was removed from the data set.

Next, we performed a correlation analysis to identify which features correlated to the sales price of our data set. We plotted out these features on scatter plots to review the trends between numeric features. We then reviewed categorical data against sale price using box plots to see which factors contribute to the relationship and to create dummy variables. We took those dummy variables and looked at their correlation to identify which factors to keep in our data set for future modeling purposes. Histograms of these selected factors were then created to observe their distributions.

## Results

Our data set has 2,919 records with 81 different factors, including the target factor of sale price. The sale price of homes in our data set have a mean of $180,921. The minimum sale price is $34,900 and max is $755,000. The standard deviation of sale price is $79,442.50

We will be including the 14 factors which had a medium to high correlation to Sale Price. These factors include: overall quality, year built, year remodeled, total basement square feet, 1st floor square feet, living area, number of full baths, total rooms above ground, # car garage, size of garage, external quality, basement quality and kitchen quality.

## Discussion/Conclusion

Because the provided test data set from Kaggle does not contain the target variable “Sale Price”, we decided that we will use just the training set for modeling and will divide it into training and test sets. We are planning to model the data using regression, clustering and random forest methods to see which provides a better result. We will be looking at accuracy and the mean percent of accuracy to determine which model to proceed with. Additionally, we will need to be cautious of over-fitting our model.

We will be comparing the performance of both regression and decision tree models and compare the r-squared value and other performance indicators. We will also be looking at the factors and seeing how they perform in the model using their coefficient p-values, confidence and impact scores. Models are created and evaluated in R studio.

## Acknowledgements

We wish to thank all the people whose assistance was a milestone in the completion of this project with special recognition to our professor, Fadi Alsaleem.

## References

*House prices - advanced regression techniques*. Kaggle. (n.d.). Retrieved September 30, 2021, from <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview>.

Smart, T. (2021, September 28). *Home prices continue record-setting pace, rising 19.7 % in ...* USNews.com. Retrieved September 30, 2021, from <https://www.usnews.com/news/economy/articles/2021-09-28/home-prices-continue-record-setting-pace-rising-197-in-july>.