Housing Market Pricing Analysis – Preliminary Analysis

Scott Breitbach, Pushkar Chougule, and Rachel Nelson

Bellevue University

Abstract

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

Background

In order to go about finding a proper data set, we looked at different options including scraping data from Zillow or reviewing various data sets available online. Due to having an extensive collection of factors and boasting 163 columns of data, which were 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 sqaure feet, lot size, and neighborhood.

We combined the training and test data sets 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 of data. We identified which columns had missing data and either filled in the missing data using mean or removed the record 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 scatterplots to review the trends. We then reviewed categorical data against sale price to see which factors we want 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.

## Results

Our data set has 2919 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

Due to the provided test data set from Kaggle not containing the target variable “Sale Price”, we decided that we will use the training set and divide 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 % of accuracy to determine which model to proceed with.

## Acknowledgements

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