An Investigation of Regression Techniques to Predict Housing Prices

# Abstract

House price prediction has been an important area of research and has a practical application in real estate and real estate investment. By understanding the factors that contribute to a property's value we can accurately predict a property's value and it can allow real estate agents, buyers, and sellers to make informed decisions. Large websites such as Zillow and Redfin already employ ML methods to get an estimate on real estate values. However as shown by the “Zestimate” lawsuit[1] it is shown these techniques are not without their flaws. In this paper we investigate 3 different regression techniques, using a dataset provided by a Kaggle contest and used common metrics to see how good these techniques preformed.

# Introduction

In recent times, housing price prediction has become a classic problem for machine learning and the application of regression techniques. There are several large websites, such as Zillow and Redfin, which employ these methods to provide their user with an idea of their home, or perspective home, of its value. Zillow at one point started buying investment properties in bulk based on its own estimates, with hopes of flipping them for a quick profit. [2] In this project we will investigate 3 different regression techniques and use common regression grading methods to determine how well they performed. To accomplish this, we first will clean and analyze the data provided. Next, we will look at what features provide the best correlation with the sale price target. After we’ve determine the features we will apply linear, random forest and gradient boost regression. Finally, we will compare the performance of each method and discuss how we could improve and build upon our results.

# The Data Set

The data set was provided by Kaggle as part of their house price competition[3]. In order to reduce the size of this document, we will not include the list and description of features here. Please see the included citation for specifics about the dataset. The dataset includes both numerical values as well as “object” values. Numerical values need to be normalized and “object” need to be mapped to numerical values before we can use the dataset for regressions.

# Analyzing and Cleaning the Data

The first thing we did, as we would with any data set is look at the number of bad input values and address them. We found several columns with 80% and above bad entries. We decide to simply drop columns with greater then 40% bad entries. For the rest, non-enumerated features, we looked at the skew of the distribution to determine if we should use the mean value or the median value. For the enumerated features, we simply used the mode as the fill-in value.

Now, that the data no longer had bad entries. We decided to get an initial look at each of the features correlation with the sale price, our target. We then chose a few of the top ones and plotted these, we can visually see if there is a trend.

Chart, scatter chart

Description automatically generated

As seen, these all have a visible trend with sale price and that means it is very likely we can use them for a regression.

Lastly, as a part of the data cleaning, we attempted to remove some of the outliers from the housing prices by using the Interquartile Range method [4]. When testing later, this was found to have a negligible effect on the regression performance.

# Feature Selection

The features can be broken up into two parts the numerical and the categorical parts. The numerical parts have already been address in the data cleaning section. But we need to address the categorical data and map them to ordinal values, this way they to can be used in the regression. To do this we used the sklearn OrdinalEncoder functions on the “object” columns. Once that is complete we can then also correlate them with the sale price and see what features we may want to use. Below is the heatmap of the correlations with the categorical values.

A picture containing timeline

Description automatically generated

Here is the same heatmap for the numerical values.

A picture containing chart

Description automatically generated

This shows that the numerical values, in general, have a higher correlation with the sale price.

Now that we have correlation scores with all the available data, we can choose which we would like to use in the regressions. For this we choose features which had at least a 30% positive or negative score. This threshold was chosen via trial and error. If we set it too high few features may be chosen. If it is too low, then we may end up with a lot of noise. The chosen features are as follows:

'OverallQual', 'YearBuilt', 'YearRemodAdd', 'MasVnrArea', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'GrLivArea', 'FullBath', 'TotRmsAbvGrd', 'Fireplaces', 'GarageYrBlt', ‘GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchSF', 'ExterQual', 'Foundation', 'BsmtQual', ‘HeatingQC', 'KitchenQual', 'GarageType', 'GarageFinish'.

# Model Creation and Results

For the model creation we used libraries from sklearn; LinearRegression[5], RandomForestRegressor[6] and GradientBoostingRegressor[7]. Each of these employs a different method for curve fitting. The method is the same for each. The data is broken into 2 parts training and test data. The training data along with known prices is used to train (or fit) the model. Once the model is trained, we use the test data to see how it did. In this exercise, we did not explore changing the default parameters or other optimization methods to attempt to improve our results. The table below shows our results:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean Absolute Error | Root Mean Squared Error | R2 / EV |
| LinearRegression | ~18704.58 | ~28348.17 | ~0.771157 |
| RandomForestRegressor | ~15613.14 | ~22230.18 | ~0.85946 |
| GradientBoostingRegressor | ~15414.11 | ~21863.97 | ~0.86435 |

All of the models we employed did very well at predicting the house sale prices. However, the RandomForest and the GradientBoosting methods did slightly better. The explained variance (EV) was significantly better, which means in general it fit the data more closely. This was also reflected in the mean error values, where they had a lower expected error.

# Conclusion and Future Work

All our models did well, but they also had a fair amount of error in price estimation. These models have room for improvement, in that the parameters governing their adaptation can be tweaked. The feature set does have redundant features which could be combined, such as square foot of each level of the house and all contribute to the overall square footage. Upon looking at some of the other submissions for the Kaggle contest, it appears that others had similar results with a .96 EV score being the highest we found among the submissions we inspected.

As with any kind of market, things are worth what people are willing to pay and housing is no different. To some a large kitchen is important, to others a large garage. This personal preference will always make prices somewhat subjective and in the end, it is really about fitting the buyer to the house, which will allow the seller to ask for a higher price.

# Division of Work

|  |  |
| --- | --- |
| **Task** | **Contributors** |
| Project part 1, Proposal | Brandon and Brian |
| Creation of Notebook and Importing Data | Brandon |
| Data Cleaning and Inspection | Brandon and Brian |
| Feature Selection | Brandon |
| Model Generation: Random Forest | Brandon |
| Model Generation: Gradient Boosting | Brandon |
| Model Generation: Linear Regression | Brian |
| Report Generation | Brandon and Brian |
| Cleaning up and organizing Notebook for submission | Brandon and Brian |

# Bibliography

[1] N. Levy, “Appeals court sides with Zillow in lawsuit over Zestimate accuracy,” *GeekWire*, Feb. 12, 2019. https://www.geekwire.com/2019/appeals-court-sides-zillow-lawsuit-zestimate-accuracy/ (accessed May 02, 2023).

[2] “Zillow’s $6 billion home flipping business was a disaster. Now, a cooling housing market could foil its comeback plan | Fortune.” https://fortune.com/2022/06/02/zillow-6-billion-home-flipping-business-housing-market-fortune-500/ (accessed May 02, 2023).

[3] “House Prices - Advanced Regression Techniques.” https://kaggle.com/competitions/house-prices-advanced-regression-techniques (accessed May 02, 2023).

[4] “3.2 - Identifying Outliers: IQR Method | STAT 200,” *PennState: Statistics Online Courses*. https://online.stat.psu.edu/stat200/lesson/3/3.2 (accessed May 02, 2023).

[5] “sklearn.linear\_model.LinearRegression,” *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.linear\_model.LinearRegression.html (accessed May 02, 2023).

[6] “sklearn.ensemble.RandomForestRegressor,” *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed May 02, 2023).

[7] “sklearn.ensemble.GradientBoostingRegressor,” *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (accessed May 02, 2023).