Final Report

Housing Price Analysis & Predictive Modeling

Joshua Ogden-Davis

For Springboard

Jan 17, 2024

1. Problem

Purchasing a home can be one of the biggest milestones in the life of an individual or a family. The location and characteristics of the home largely defines their daily experience, and the price of the home can define their financial situation for decades.

For investors, purchasing a home is a gamble with big consequences. The payout could be in the hundreds of thousands, but just one or two bad investments could bankrupt an individual investor or small company.

With so much riding on the purchase of each and every house, the ability to accurately assess the market value of residential properties is crucial not only for residents and investors, but also for the health of the economy as a whole.

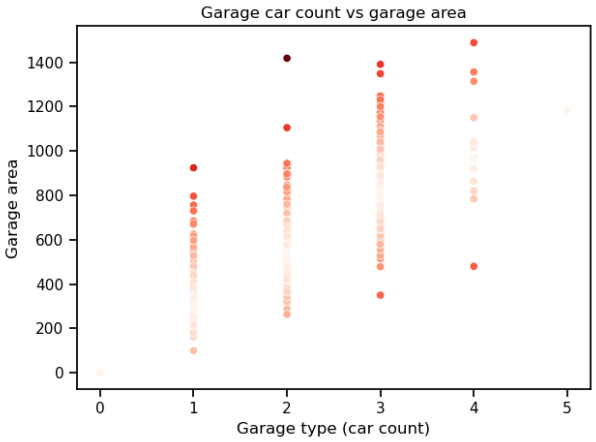
The aim of this project is to develop a model that uses a uniform set of features to predict the sale price of residential properties.

2. Data Sourcing & Wrangling

The dataset used (from Kaggle) contains 78 features and is pre-divided into a training set (1,460 records) and a test set (1,459 records). Each record corresponds to a residential property that has been sold within a 5-year period (2006-2010), and the features range from fully continuous (i.e., square footage of different parts of the property) to ordinal (i.e., quality ratings of various parts of the property on a 5-point scale) to fully categorical (i.e., the zoning of the property). The target variable is the price at which the property was sold.

Considerable wrangling was required to ensure the reasonableness and usability of the data.

Several values were obviously unreasonable and almost certainly an error. For example, the recorded garage areas and garage car counts seem to contradict each other at times. One 4-car garage is recorded has having less than 500 square feet of space, which is extremely unlikely. Unfortunately, since we cannot identify errors that don't stick out noticeably from the data, we must accept that some errors will go unnoticed.

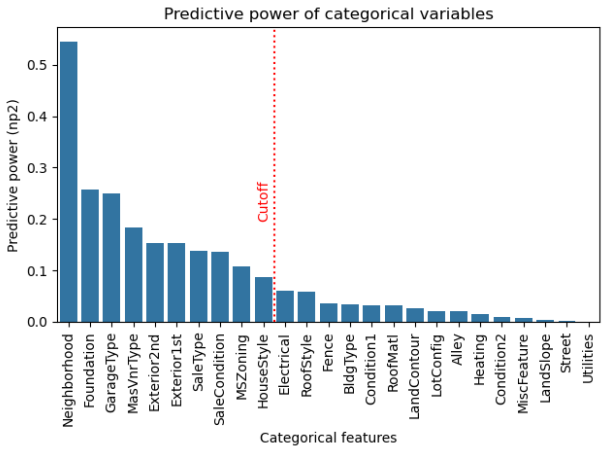


35 of the features contained one or more null values, with a total of 15,424 missing values. As many of those missing values were in the test set, I could not rely on dropping nulls. To enable the model to deal with any future dataset that includes missing values, I defined a set of rules for imputing missing values on a column-by-column basis. These rules will be applied programmatically to all datasets before prediction.

3. Preparation for Modeling

Including all 78 features in our predictive model will likely lead to overfitting and inaccuracy. Before modeling, I applied different methods to determine the features with the most predictive power and omitted the rest.

Determining the predictive power of categorical variables is especially important, as they must be one-hotted before modeling, which can be very time-consuming to undo or modify. I applied ANOVA methods to determine the ten most predictive variables and omitted the rest before continuing modeling.



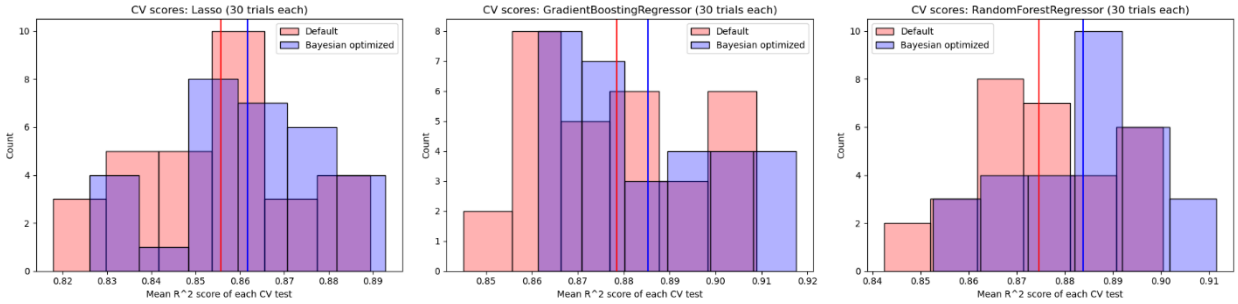
One quirk of this dataset is that some functions that are functionally identical are split across different columns. For example, the columns “Exterior1st” and “Exterior2nd” both record exterior features. Some properties only have one exterior feature recorded, in which case the second column is null. For houses that have two exterior features recorded, the order is not guaranteed. However, as long as this information is recorded in different columns, the model will consider them to be different features, leading to inaccuracy.

In order to overlap the information in both of these columns in such a way that the model will not treat them differently, I simply took the “1” values from the one-hotted “Exterior2nd” columns and pasted them into the corresponding one-hotted “Exterior1st” columns, then dropped all “Exterior2nd” columns.

I chose to center and scale all continuous and ordinal variables.

4. Modeling

I applied sklearn’s Lasso, GradientBoostingRegressor, and RandomForest models to cv splits of the training data (the target values for the test data are not publicly available). For each model, I randomly generated 30 train/test splits from the training data and ran the model on each split twice (once with default parameters, once with parameters obtained with the bayes_opt library’s BayesianOptimization function).



5. Findings

The best performing models (as measured by R^2) were Bayesian-optimized Gradient Boost (0.889) and Random Forest (0.882). While Gradient Boost seems to perform slightly better, a t-test on the arrays of 30 CV scores produces a p-value of 0.215, insufficient to confirm a significant difference between the distributions.

In order to verify a clear winner, the models may be run until a significant p-value emerges.

In order to improve the accuracy of the model overall, another round of more in-depth feature engineering could be undertaken (including a more robust selection of automated feature selection tools). Consulting an expert with deeper domain expertise could also further enlighten the feature selection process.

6. Applications

For current home buyers, the model could provide an estimate of the price a specific property could command, which could assist in identifying undervalued properties or avoid overpaying for a property.

For current homeowners, it could help with decisions about renovations or additions to the home by giving an estimate of how specific modifications to the property are likely to change the value.

For home builders, it could inform their decisions about which features to include in a new build in order to meet specific price points and optimize ROI.

7. Limitations

In addition to improvements in the modeling process, the current results suffer from the limitations of the data itself and the data collection process. As the data comes from a specific location and time period, it may not generalize to other regions or the current market situation; as the data collection process is not transparent, it is difficult to confirm the representativeness and accuracy of the data. However, interpreting the results of the model within an appropriate context is shown to provide reasonably accurate estimations of the sale price.