Michael A Moran

Capstone Project Paper

27 June 2018

What’s in a House? Predicting Housing Prices with Regression

A house is perhaps the single greatest financial investment that a couple or individual will make in their lifetimes, and housing prices are currently rising faster than both income and inflation. Understanding which features of a house are most related to the sale price of that house can provide insight into the housing market and help buyers better understand which features match their price range. The Ames Housing dataset contains “eighty variables directed related to property sales”[[1]](#footnote-1) from Ames, Iowa over a period from 2006 to 2010. The variables include a mix of continuous, discrete, and categorical data types. The data is presented with a train and a test set, with the train set containing 1460 observations. The goal is to create a regression model in R that can accurately predict the sale price. A regression model is intuitive for this problem since the goal is to predict a continuous variable.

The dataset presented a few issues that needed to be addressed before regression could be run. There were numerous NA values, and the categorical variables that were to be used in regression needed to be converted into dummy variables. To address the NA values, a new version of the dataset was created with the NAs coded as zeroes. The dummies package was utilized to create the dummy variables. Initially, this was performed by creating a second data frame of dummy variables for the variable to be converted, using cbind to join the dummy data frame with the original dataset, then removing the original variable column from the bound data set. This method was an inefficient, however, as it required these steps to be run over each variable, where the as.data.frame function allowed for the dummy variables for each original variables to created in one call, without creating a second data frame that would need to be bound as before.

Once the data was ready for analysis, an initial statistical picture of the data set was acquired using base R. It was determined that the mean sale price of a house in Ames is $180,921.20. The standard deviation of the sale price is $79,442.50 and variance is $6,311,111,264. The minimum sale price is $34,900 and the maximum is $755,000. The inner-quartile range is $84,025. A histogram of sale prices was created using base R, but the coding was time-consuming, and the visualizations were basic and plain. To create more attractive and useful graphics, the dataset was uploaded to Tableau, and graphics were created to explore the relationship between sale price and thirteen variables.[[2]](#footnote-2) Strong linear relationships were observed between sale price and the number of the bathrooms, sale price and the above ground square footage, and sale price and the overall condition of the house. A linear relationship was also observed between sale price and the number of bedrooms. This relationship was less robust over five or six bedrooms, however most of the homes in the dataset had two to four bedrooms, where a strong linear relationship between the number of bedrooms and the sale price was observed.

For categorical data that was converted into dummy variables, the relationship between sale price and each categorical variable was evaluated by plotting the distribution of sale price by each type of the variable, for each variable. Strong associations between price and variable type were observed for the garage type, the basement condition, the type of dwelling, and the neighborhood of the house. These variables, along with the earlier numeric variables that demonstrated a linear relationship with sale price, were selected for use in the regression models. The other variables did not show a relationship with sale price and were eliminated from consideration as a result.

Using the statistical analysis as a starting point, a linear regression model was created in base R to evaluate the relationship between sale price and the chosen variables. Linear regression was run over each variable individually, then over all the variables. The models were evaluated using their R-squared and p-values. The p-values indicated that the variables were significant but the R squared values, which were generally around 0.5, indicated a poor fit of the model with the data. Random Forest and Support Vector Machines were then run over the same variables to see if they might better fit the data. Random Forest was also utilized to create an importance plot of the model’s features. Using this output, both models were tuned and retuned, with the second tuning using a higher threshold for importance.

The train set from the original data was divided into a new test and train set so that the actual sale price could be compared to the predicted value created by the model. For each iteration of both the Random Forest and the Support Vector Machines models, predictions were generated for the sale price of the train set and the test set, then used to find the error of the model for each prediction. This was then used to calculate the Root Mean Squared Error and the Mean Absolute Error for the train set model and the test set model, which were compared against one another for each error type to validate the performance of the model against the test set. A third attempt to tune the model decreased the performance between the train and test sets, so the Random Forest model from the second tuning was ultimately deemed the best performing model. This model was found to be within plus or minus 15.41% of the average sale price based on its Mean Absolute Error, and the percent error between the train set and test set was found to be 25.66%. Feature engineering was attempted on the square footage variable, wherein bins were created for the distribution of continuous variables, but they were not found to improve the performance of the model.

From the best performing model, the features found to have predictive value are the square footage, the number of bathrooms, the number of bedrooms, having an attached garage, being a two-story home from 1946 or newer, the overall condition of the house, having a detached garage, and being a one-story home from 1946 or newer. By identifying these features, prospective buyers can both anticipate the cost of a potential buy based on them and can decide if certain features might be worth downsizing to control the overall sale price. One way in which these results are limited is that the data set is derived from and tested on one city. Further analysis may want to consider the performance of the model on data from different cities or create a model from data of sale prices of houses in towns across the United States and compare that model’s performance with this one.

1. Cock, D. D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistics Education, 19(3). [↑](#footnote-ref-1)
2. MSSubClass, the type of dwelling involved; FullBath, the number of full bathrooms; BedroomAbvGr, the number of bedrooms; GarargeType, the location of the garage; BldgType, another dwelling classification; SaleCond, the condition of the sale; OverallCond, the overall condition rating; LotConfig, the lot configuration; BsmtCond, the overall condition of the basement; RoofStyle, the type of roof attached; Condition1, the proximity to a main road or artery; GrLivArea, the total square footage above ground; and Neighborhood, the location within Ames, IA. [↑](#footnote-ref-2)