Housing Price Prediction Project

Anmol Makhija

Overview

The goal of this project is to develop a model to predict the prices of houses in Douglas County, CO during December 2012 given data on transactions that occurred between 1980 and November 2012. The raw data used for this project is publicly available here.

Data

The raw data contains transaction and property level data on the sale price of houses in Douglas County, CO. It is split up into two sets to begin with: a model training set and a test set. There are roughly 210,000 transactions for 80,000 houses in the training set and each observation records 55 variables. These variables record transaction date, price, property characteristics etc. The test set records 191 transactions.

Variable Selection

The first step taken was to prepare the raw data and remove variables that did not seem like they would logically predict the sale price of a house, such as the name of the grantor/grantee in the transaction. This was done in order to reduce the initial set of 54 variables recorded per transaction and make the computation more efficient. This led to the creation of a modified training data set that recorded 38 variables per transaction.

Several algorithms were utilized on these 38 variables to choose a predictive model with the best subset of variables.

I considered the following criterion and methods to get an optimal subset of variables:

- Criterion:
  - BIC
  - AIC
  - Adjusted R Squared
  - MRSE
- Methods:
  - Backward stepwise regression
  - Linear regression with stepwise selection (AIC)
  - Ridge regression

Out of these variations in the model development the one I thought was optimal used the Backward stepwise regression and MRSE to select a subset of variables to include in the prediction model.

This specification led me to include location_zip_code, total_net_acres, tax_district_no, style, improvement_sf, total_garage_sf, total_finished_basement_sf, total_unfinished_basement_sf, built_year, and transaction_year.

This subset was indicated to be the optimal subset by the Backward stepwise regression, MRSE criterion optimization, and was computationally more efficient than the Linear regression with stepwise selection (AIC) and Ridge regression methods. However, it did not lead to a great predictive power since only 27% of the variability in log sale price was explained by these variables. I suspected that the low predictive power could be because I was using sales price data from the early 1980s to predict housing prices in 2012 and created a trend plot to examine the yearly average sale price trend.

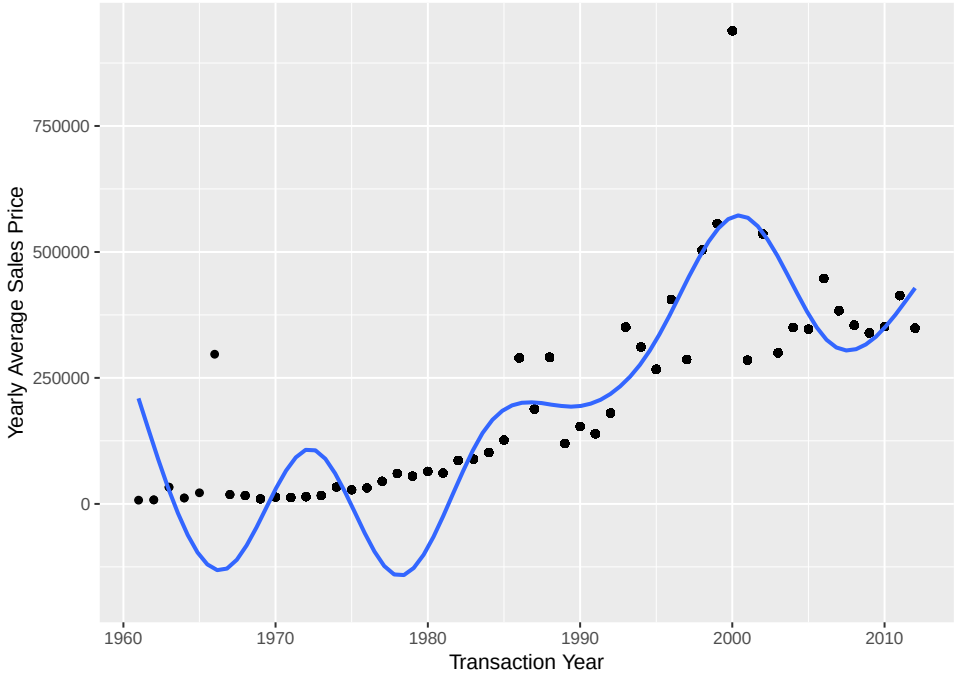

Figure 1: Trends in Yearly Average Sale Price

I then sub-set my model training set to only include observations from November 2011 onwards which increased the model's ability to accurately predict sales prices in December 2012 to explain 84% of sales price variation. However, this seemed like too small a sample size to be able to draw a generalizable predictive model for the future.

Outliers and Remedial Measures

I utilized several measures to check for x and y outliers, including DFFITS, Cook's Distance, and Deleted Studentized Residuals. Based on the outcome of these tests I determined that there were a few outliers and influential points, and remedial measures were required. I decided to run a robust regression model to dampen the effects of the outlying cases.

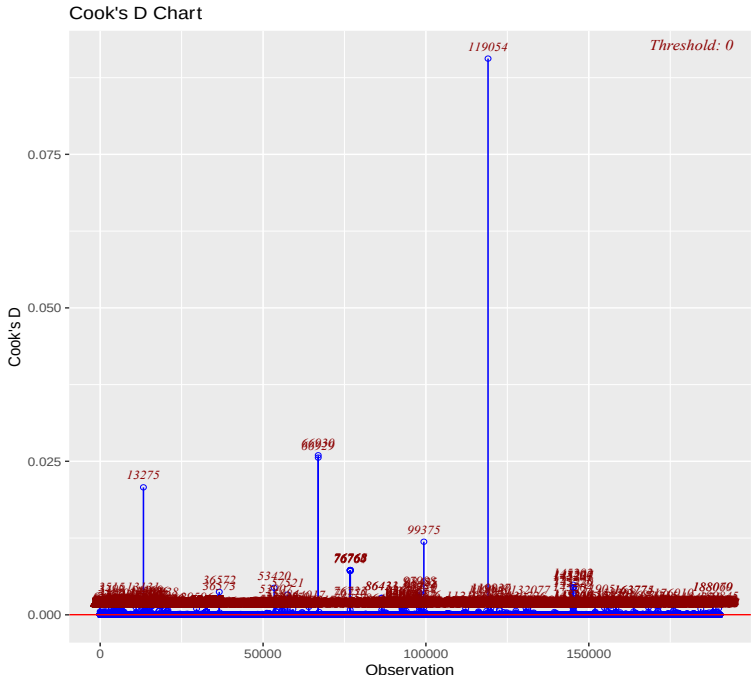

Figure 2: Cook’s Distance Chart

Figure 3: DFFITS Graph

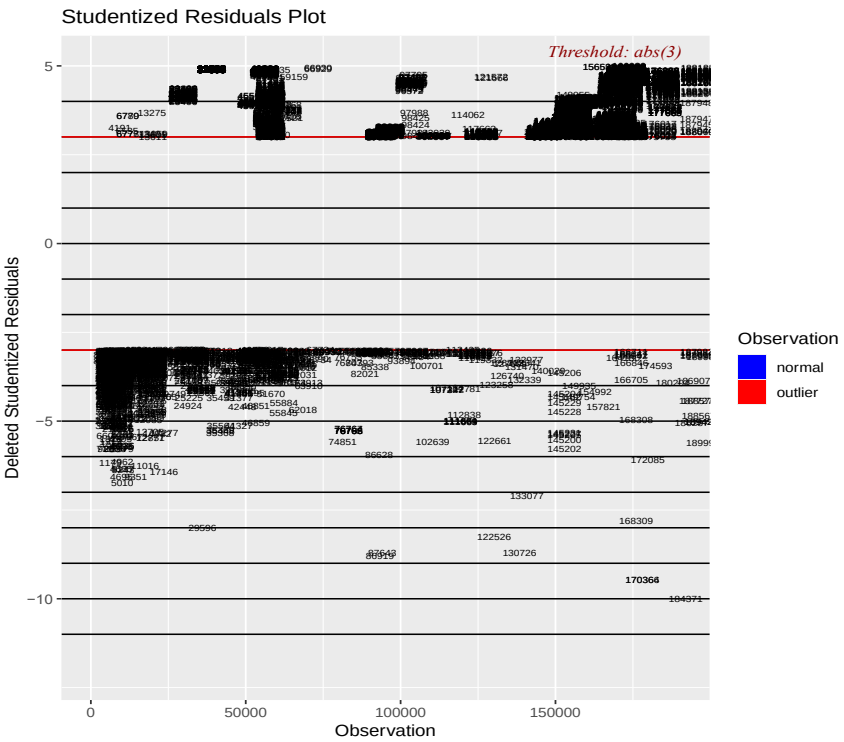

Figure 4: Deleted Studentized Residuals Plot

Analysis of Prediction Model

The model I utilized to create the predicted pricing column had an adjusted r-square of only 0.27 so the predictive power of the model was not too great. There was a lot of variation in observed sales price and predicted sales price for 2012 and I suspect this happened for two reasons:

1. Timeframe utilized for model-training: Like I had mentioned earlier, I ran one specification of the model where I utilized only observations from November 2011 onwards to predict housing prices this improved the model's predictive accuracy substantially (by a factor of 3).
2. Subset of variables included: The original dataset provided for the task had 54 variables, however, I was not able to utilize all of them since I ran out of computational power (The R vector memory exhausted). I tried working around this by providing a larger memory capacity to the R environment in my terminal. However, I was still not able to run all possible combinations of the 54 variables given to choose the best subset.

The model's predicted price was far off from the observed price for the transaction corresponding to the property ID: R0376884 on 19th December 2012. I believe that this was due to a combination of both omitted variable bias and including a broad time frame for training the model.

Overall, this model tracked the predicted prices decently well as can be seen in the figures below. However, there is still a large scope for improvement.

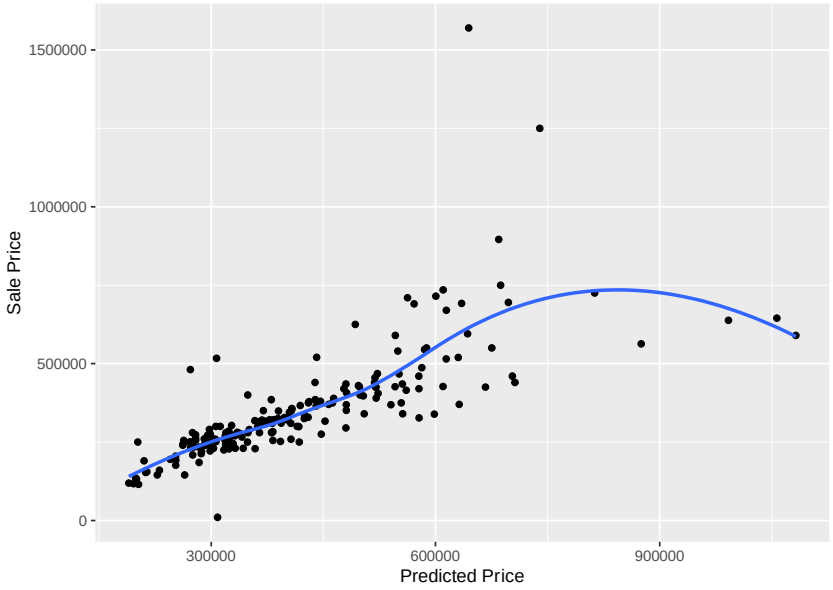

Figure 5: Predicted Vs. Observed Sales Price for Dec. 2012

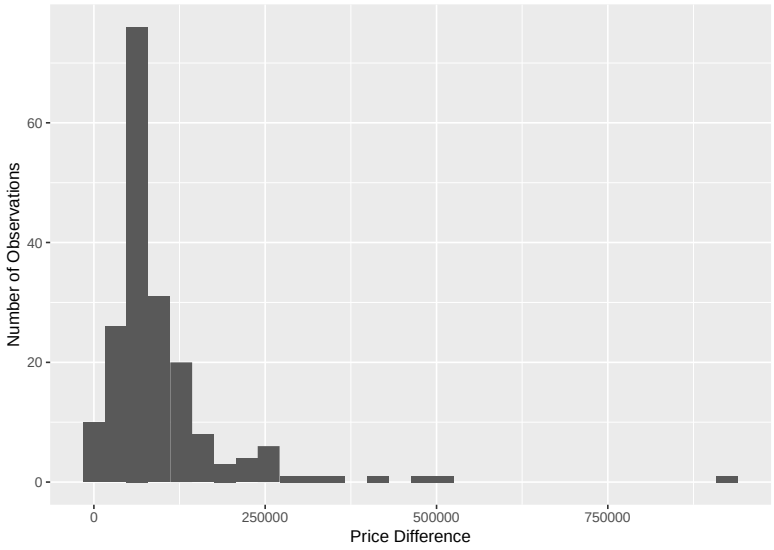

Figure 6: Histogram of Difference Between Observed and Predicted Prices