Madison Single Family Housing

Ira Sharenow

2023-03-05

## Introduction

This document uses linear regression, random forests, and boosting to analyze Madison Wisconsin single family home sales, 1994-2023.

In this document, I used the R programming language for data cleaning and manipulation as well as for implementing the data science techniques.

The charts were created using the ggplot2 package. The tables were created using the flextable package.

I have also created a number of charts using Tableau and summary pivot tables and charts with Excel. Please visit my website for the links. <https://irasharenow.com/>

This document starts with a familiar pattern. The data is read in. Then the data is cleaned up as there are misspellings and other errors.

The data is broken down by high schools and school areas that had few sales were eliminated with the use of a user-defined R function.

First I did regression. Then random forest. Then boosting.

In each case, I created a model on the training set. Then I performed predictions on the test set.

The comparison table is at the end.

For the analyses, I broke the data up into a training set and a test set. I then performed a number of analyses. I plan on revisting this data set to do further analyses in the near future.

The R code is available in the alternate version of this document.

# ```{r setup, include=TRUE}
knitr::opts\_chunk$set(echo = FALSE, ft.align="left")

# Madison housing second version

# Preliminaries

library(tidyverse)
library(scales) # Graphing
library(flextable)
library(readxl) # read Excel file
library(stringr) # string manipulation
library(lubridate) # working with dates
library(leaps) # selecting best regressions
library(tree) # basic tree function
library(randomForest) # random forest
library(gbm) # gradient boosting
library(faraway) # regression techniques

Home Sales by Decade and Size

| **Madison Area Home Sales** |
| --- |
| **Decade** | **Small** | **Medium** | **Large** |
| D1990s | 7,495 | 5,915 | 4,509 |
| D2000s | 13,126 | 13,180 | 11,621 |
| D2010s | 11,525 | 14,168 | 15,697 |
| D2020s | 4,059 | 4,997 | 5,630 |
| Years: 1994-2023Small: < 1600 SqFt;Medium: 1600-2160 SqFt;Large > 2160 SqFt |

# Some exploratory charts



Boxplots of Sale Price by High School and Decade of Sale






# Regression

Split data in training and test sets
Then do regression on training set
and predict on the test set

## The regression formula is:lm.fit = lm(Sold\_Price ~ FinSqFt + Closing\_Date\_year + YearBuilt1980\_plus + Beds + LandAssess + Total\_Full\_Garage\_Stalls + Total\_Baths + High, data = housingTrain)

|  | **Estimate** | **Standard Error** | **t value** | **Pr(>|t|)** |  |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | -14,966,042.952 | 82,757.792 | -180.841 | 0.0000 | \*\*\* |
| FinSqFt | 90.118 | 0.799 | 112.784 | 0.0000 | \*\*\* |
| Closing\_Date\_year | 7,468.471 | 41.193 | 181.302 | 0.0000 | \*\*\* |
| YearBuilt1980\_plus | 15,737.545 | 862.532 | 18.246 | 0.0000 | \*\*\* |
| Beds | -36.748 | 659.499 | -0.056 | 0.9556 |   |
| LandAssess | 0.012 | 0.001 | 11.486 | 0.0000 | \*\*\* |
| Total\_Full\_Garage\_Stalls | 6,753.770 | 531.820 | 12.699 | 0.0000 | \*\*\* |
| Total\_Baths | 4,529.345 | 780.922 | 5.800 | 0.0000 | \*\*\* |
| HighEast | 5,041.789 | 4,371.555 | 1.153 | 0.2488 |   |
| HighLafollette | -15,727.800 | 4,351.389 | -3.614 | 0.0003 | \*\*\* |
| HighMcFarland | 3,651.282 | 4,628.138 | 0.789 | 0.4302 |   |
| HighMemorial | 4,964.116 | 4,312.594 | 1.151 | 0.2497 |   |
| HighMiddleton | 69,944.538 | 4,854.691 | 14.408 | 0.0000 | \*\*\* |
| HighMonona Grove | 180.184 | 4,524.939 | 0.040 | 0.9682 |   |
| HighOregon | 2,869.155 | 4,505.812 | 0.637 | 0.5243 |   |
| HighStoughton | -13,178.824 | 4,492.255 | -2.934 | 0.0034 |  \*\* |
| HighSun Prairie East | -22,503.858 | 4,338.327 | -5.187 | 0.0000 | \*\*\* |
| HighVerona | 6,654.108 | 4,425.759 | 1.503 | 0.1327 |   |
| HighWaunakee | 27,848.989 | 4,507.681 | 6.178 | 0.0000 | \*\*\* |
| HighWest | 59,651.492 | 4,362.233 | 13.675 | 0.0000 | \*\*\* |
| *Signif. codes: 0 <= '\*\*\*' < 0.001 < '\*\*' < 0.01 < '\*' < 0.05* |
|  |
| Residual standard error: 7.69e+04 on 52217 degrees of freedom |
| Multiple R-squared: 0.6711, Adjusted R-squared: 0.671 |
| F-statistic: 5609 on 52217 and 19 DF, p-value: 0.0000 |



## mean square error: 5921807030

## mean absolute difference: 51887



## Random Forest

##
## Call:
## randomForest(formula = Sold\_Price ~ FinSqFt + Closing\_Date\_year + YearBuilt1980\_plus + Beds + Full\_Baths + YearBuilt + Total\_Baths + ZIP5 + High, data = housingTrain, mtry = 4, importance = TRUE, ntree = 200)
## Type of random forest: regression
## Number of trees: 200
## No. of variables tried at each split: 4
##
## Mean of squared residuals: 3273089717
## % Var explained: 82.05

## RF MSE: 3199413497.22394

## mean abs error: 32273.7294074767

## %IncMSE IncNodePurity
## FinSqFt 77.36642 3.452654e+14
## Closing\_Date\_year 327.82376 2.650597e+14
## YearBuilt1980\_plus 26.50031 4.475491e+12
## Beds 28.77647 1.947189e+13
## Full\_Baths 12.07724 2.218624e+13
## YearBuilt 134.38530 1.246365e+14
## Total\_Baths 19.23591 1.114328e+14
## ZIP5 71.91995 3.139464e+13
## High 96.52026 4.368963e+13



## Boosting



## var rel.inf
## FinSqFt FinSqFt 42.87730738
## Closing\_Date\_year Closing\_Date\_year 26.43764245
## YearBuilt YearBuilt 11.48942343
## ZIP5F ZIP5F 10.06617681
## HighF HighF 5.03125775
## Total\_Baths Total\_Baths 2.91438632
## Beds Beds 0.79407083
## Full\_Baths Full\_Baths 0.36061158
## YearBuilt1980\_plus YearBuilt1980\_plus 0.02912345

## MSE Boosting: 3061540248.45942

## AMAD Boosting: 31876.3971474589

## Technique VarExpl MeanAbsError
## 1 Regression 0.671 51887
## 2 RF 0.821 32773
## 3 Boosting NA 31876

Performance of the 3 Data Science Techniques

| **Analysis of Madison Area Home Sales** |
| --- |
| **Technique** | **VarExpl** | **MeanAbsError** |
| Regression | 0.671 | 51,887 |
| RF | 0.821 | 32,773 |
| Boosting |  | 31,876 |