PREDICTING LIST PRICES OF HOMES IN ROUND ROCK, TEXAS

by Rebeca Ansar

BACKGROUND

Round Rock is a burgeoning city in the greater Austin, Texas area.

One of the fastest growing cities in the US with a population growth of 33.5% in the last 10 years.

Dell is Headquartered there, Apple is coming, and other major tech companies have solidified plans to build in the area

RESEARCH QUESTION

What predictors, if any, are significantly associated with the list price of a home in Round Rock, Texas?

RESEARCH MOTIVATION

Educate	Educate both homeowners and buyers about the area
Use	Sellers can use the analysis to prioritize renovations associated with a higher list price
Decide	Buyers can decide if the list price is fair for the area.

HYPOTHESES

- Significant predictors will include: number of bedrooms, the number of bathrooms, the square footage of the home, the acreage of the land, and the age of the house
- Non-significant features will include: number of heating and cooling options and the number of days the home remains on the market

METHODS

Data was sourced from the RealtyAustin website

The 100 first-available listings were used as observations of which 95 were retained

MODEL-BUILDING

- Divided into two parts: multiple linear regressions and random forest model
- 60% of the data was in the training set and 40% was in the testing set
- Models were assessed by their R-squared values on the test data

11 FEATURES USED FOR MODEL BUILDING

Feature Name	Description
Bedrooms	Number of bedrooms in the residence
Bathrooms	Number of bathrooms in the residence
Square_Feet	The square footage
Acres	The acreage of the residence
Property_Type	Whether the residence is a house or a condo
Heating	Number of heating options in the residence
Stories	Number of stories in the residence
Cooling	Number of cooling options in the residence
Days_on_Market	Number of days the residence has been listed for sale
Garages	Number of garages
House_Age	The current age of the home in years

AKAIKE INFORMATION CRITERION

- Named after Japanese statistician Hirotugu Akaike who devised it
- The AIC function was used for the model building phase when making multiple linear regression models
- The *stepAIC* function in R estimates prediction error of statistical models and returns the one with the lowest error

EXAMPLE OF AIC

```r
#use stepAIC function to improve model_1
aic_model_1 <- stepAIC(model_1, direction = 'both', trace = F)
summary(aic_model_1)
```

```
##
## Call:
## lm(formula = List_Price ~ Bedrooms + Bathrooms + Square_Feet +
##     Acres + Stories + Days_on_Market + Garages, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -140689  -30814  -11209   28375  110280
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -71750.56   46098.53  -1.556 0.126453
## Bedrooms       -28919.79   13065.74  -2.213 0.031869 *
## Bathrooms       70427.16   24600.95   2.863 0.006303 **
## Square_Feet       107.12      21.91   4.889 1.27e-05 ***
## Acres          210786.31   50575.40   4.168 0.000134 ***
## Stories2       -48881.00   19757.62  -2.474 0.017110 *
## Days_on_Market   6179.45    3060.12   2.019 0.049300 *
## Garages         57071.59   13333.31   4.280 9.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52410 on 46 degrees of freedom
## Multiple R-squared:  0.8992, Adjusted R-squared:  0.8838
## F-statistic:  58.6 on 7 and 46 DF,  p-value: < 2.2e-16
```

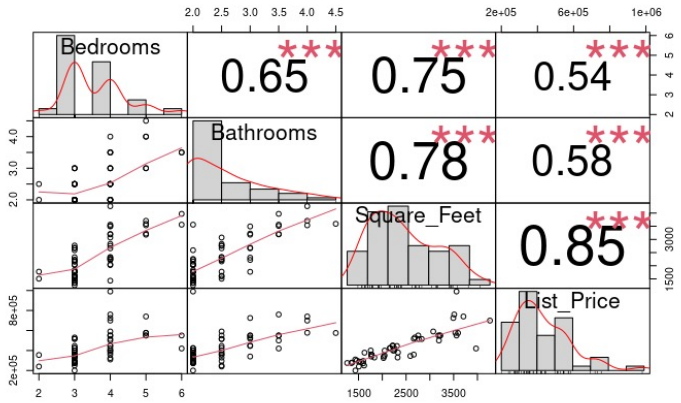
MULTICOLLINEARITY

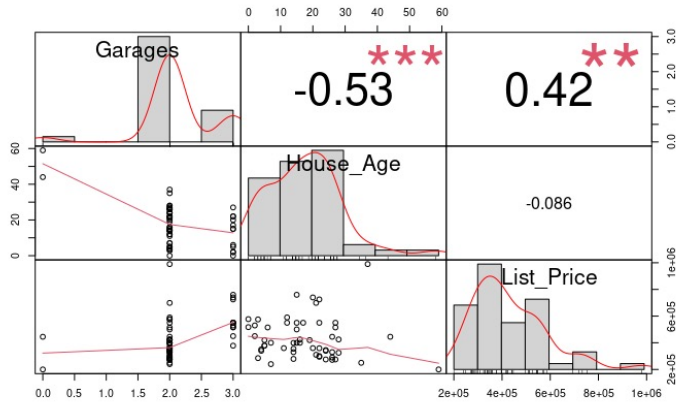
- Detecting multicollinearity was also an important part of regression model building
- Multicollinearity refers to the phenomena in which one or more of our predictor variables correlate with any of the other predictors.
- This was done by evaluating the variance inflation factors and correlation charts

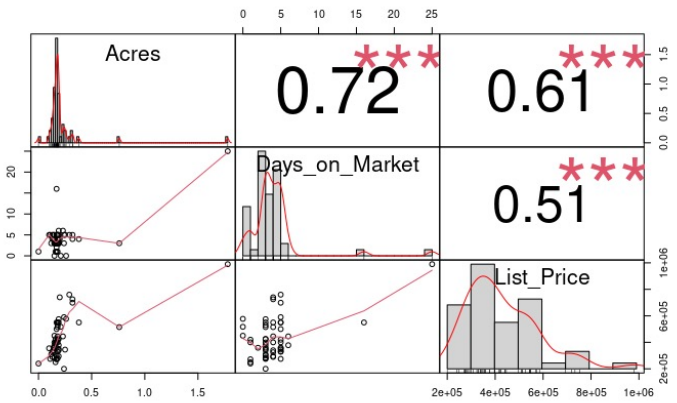
VARIANCE INFLATION FACTOR

The Variance Inflation Factor (VIF) is a statistic that is used to determine whether multicollinearity exists between any of the predictors.

In general, any VIF > 10 constitutes a high Variance Inflation Factor.



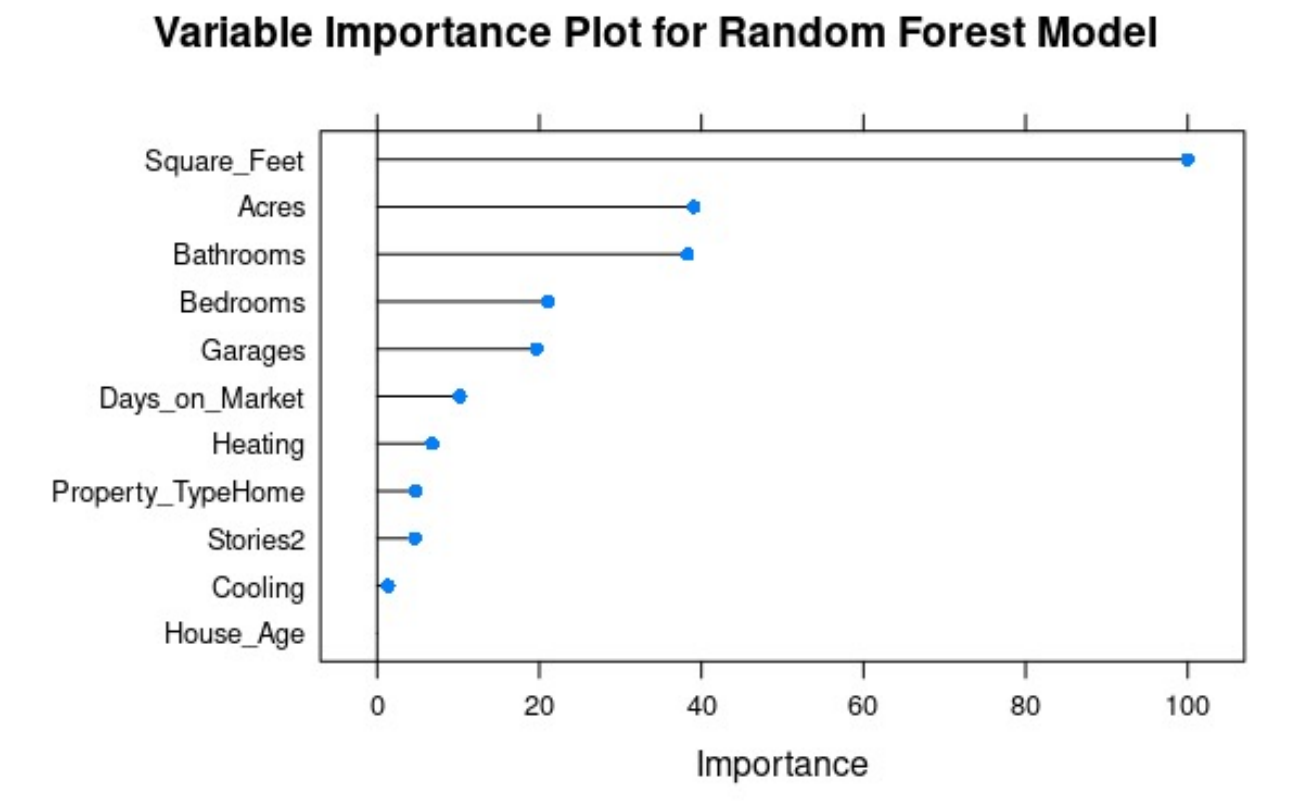




CORRELATION CHARTS

RANDOM FOREST MODEL

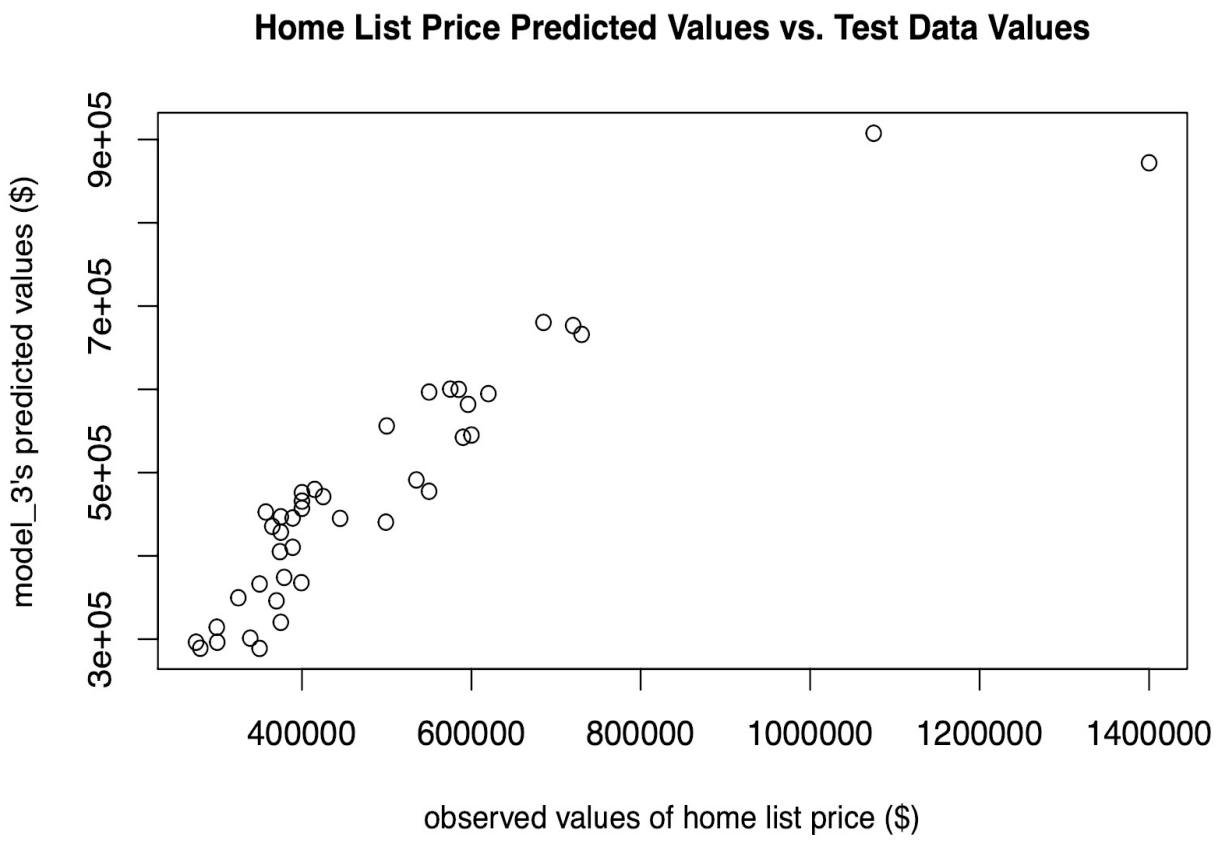
- An ensemble method was attempted to see if test accuracy could be increased
- Though this model was not optimal, it did provide a variable importance plot that could help sellers and buyers understand which home features to prioritize

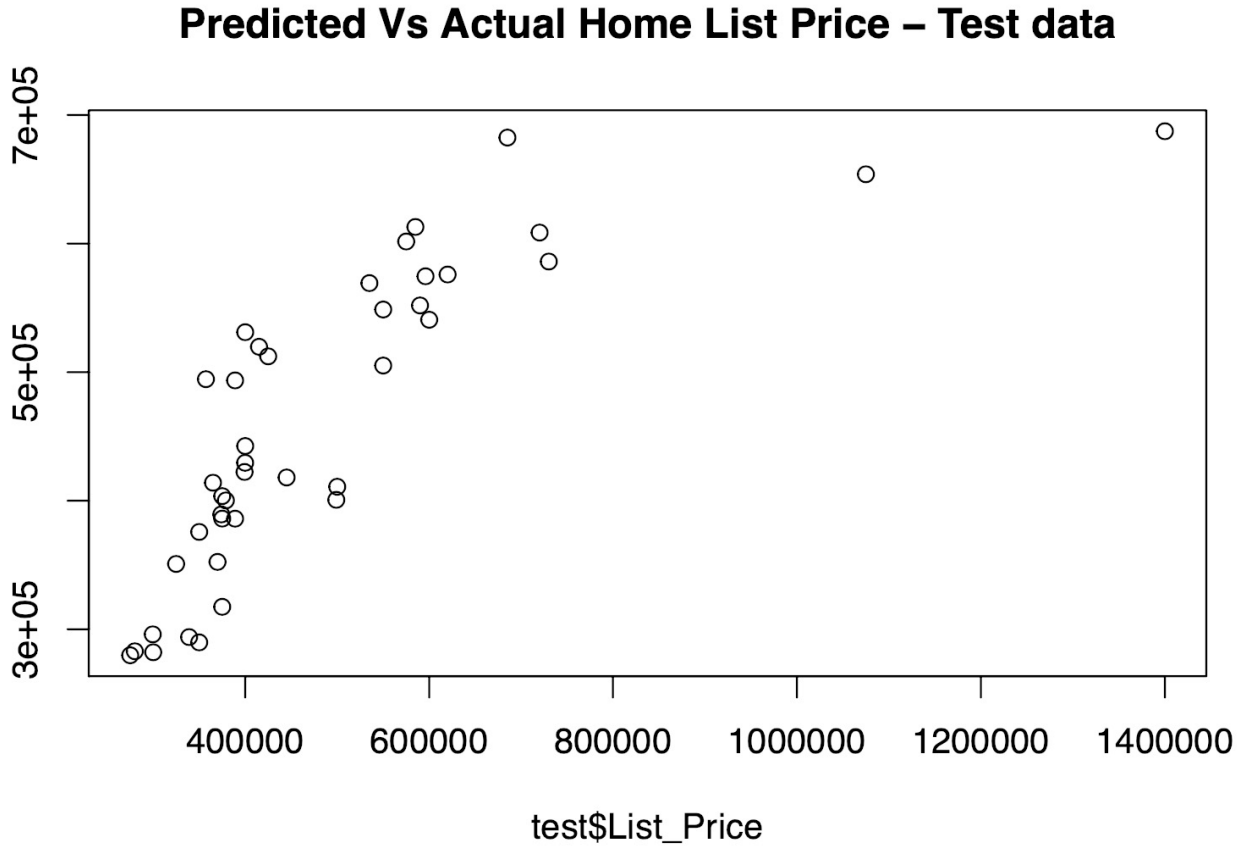


MODEL-BUILDING RESULTS

Model name	Predictors included	Multiple R-squared	Adjusted R-squared	Test R-squared
model_1	All 11 features	0.9012	0.8753	0.8242
aic_model_1	*Bedrooms, Bathrooms, Square_Feet, Acres, Stories, Days_on_Market, Garages*	0.8992	0.8838	0.8288
model_2	*Bedrooms, Bathrooms, Square_Feet, Acres, Stories, Garages*	0.8902	0.8762	0.8357
model_3	*Square_Feet, Acres, Stories, Garages*	0.8717	0.8612	0.8420
aic_model_3	*Square_Feet, Acres, Garages*	0.8703	0.8625	0.8334
forest	*Square_Feet, Acres, Bathrooms, Bedrooms, Garages,* and *Days_on_Market* have variable importance greater than 10	% Variance explained: 67.77 % proportion is 0.6777	-	0.6107

RANDOM FOREST VS. MULTIPLE REGRESSION





BEST MODEL SUMMARY

```
Call:
lm(formula = List_Price ~ Square_Feet + Acres + Stories + Garages,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-191713  -19752    -730   25142  127193

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -33034.15   33040.29  -1.000 0.322312
Square_Feet    127.93      12.44  10.281 7.96e-14 ***
Acres       246933.33   38042.00   6.491 4.08e-08 ***
Stories2    -12826.77   17837.20  -0.719 0.475494
Garages      52552.93   14312.31   3.672 0.000595 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57280 on 49 degrees of freedom
Multiple R-squared:  0.8717,        Adjusted R-squared:  0.8612
F-statistic: 83.21 on 4 and 49 DF,  p-value: < 2.2e-16
```

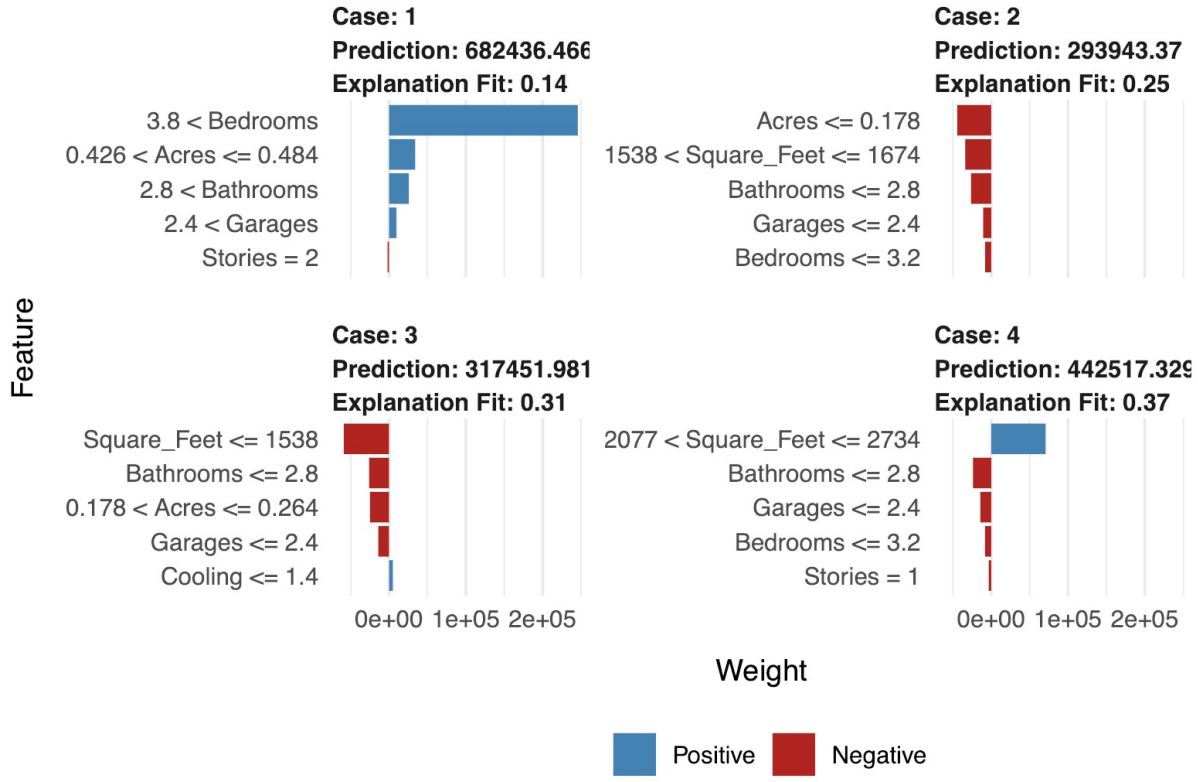
OUTCOMES

The hypotheses were partially correct.

Square footage and acreage of the home were statistically significant predictors in the best model at the alpha = 0.05 level.

MULTIPLE LINEAR REGRESSION EQUATION

$$\widehat{List\_Price} = -33034.15 + 127.93 * Square\_Feet + 246933.33 * Acres - 12826.77 * Stories2 + 52552.93 * Garages$$



CASE STUDIES WITH RANDOM FOREST MODEL

FURTHER DISCUSSION

- Process could be improved implementing web scraping so data could automatically be captured, which would lead to a bigger sample size
- Analysis could be more effective by capturing even more home features, such as presence of fireplace or square footage of backyard
- Number of garages being a significant predictor should alert us that further analysis is needed

REFERENCES

[1] *RealtyAustin*. www.realtyaustin.com. Accessed 30 March 2021.

[2] Buchanan, Taylor. "Round Rock among fastest-growing big cities in the nation, new census data shows." *Community Impact Newspaper*, 20 May 2020. https://communityimpact.com/austin/round-rock-pflugerville-hutto/data-reference/2020/05/20/round-rock-among-fastest-growing-big-cities-in-the-nation-new-census-data-shows/.