

A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light greenish-blue. They are positioned diagonally, with the blue one in front of the green one.

Predicting Residential Housing Prices

By Alex Tryforos



Table of Contents

<u>Section</u>	<u>Page</u>
Introduction	4
Data Cleaning	7
Exploratory Data Analysis.....	11
Initial Modeling	17
Cross Validation & Model Tuning	21
Model Interferences	31
Test Set Performance	41
Final Thoughts	53
Appendix	57

Introduction



Introduction

- Area of Study - Residential Homes in Mandeville, Louisiana
- Statistical consulting for 'Listing Agent Referral Services' which is an independent Louisiana real estate and brokerage consulting firm whose goals include:
 - Segmenting the components of home value as determined by market transactions to better negotiate home purchases and sales.
 - Determining and advising home sellers of the best sales agent to sell their specific home based on analysis of past agent past performance.

Data Description

- 1,600 Real estate transaction records from the last ten years in Mandeville, Louisiana consisting of the following variables:
 - Sale Price
 - Condition (New, Excellent, Very Good, Average, Fair, Poor)
 - Living Area in Square Feet
 - Subdivision (Consisting of 12 separate neighborhoods)
 - Age at time of sale
 - Total Half Bathrooms
 - Total Full Bathrooms
 - Total Bedrooms
 - Foundation (Raised or Slab)

Preview of Data

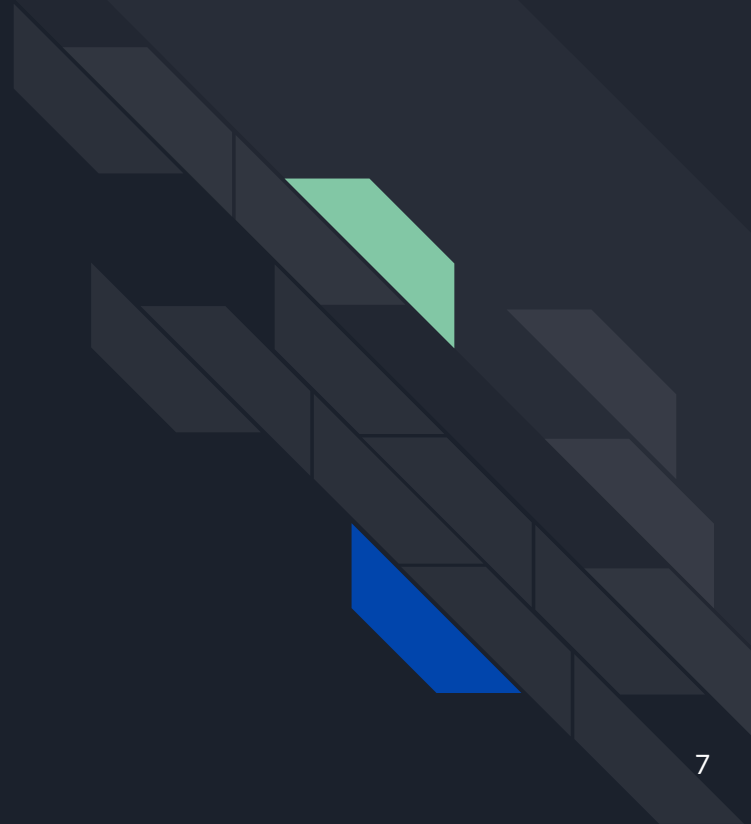
Sell_Price	Condition	Living_Area	Subdivision	Age	Baths_Half	Baths_Full	Beds	Foundation
352552.3	EXCE	2480	Audubon Lake	20	0	2	4	Slab
357000.0	EXCE	2480	Audubon Lake	28	0	2	4	Slab
405486.0	EXCE	3080	Audubon Lake	22	0	3	4	Slab
402437.5	VRGD	2825	Audubon Lake	21	0	3	4	Slab
460600.4	VRGD	3279	Audubon Lake	22	0	3	4	Slab
486238.6	EXCE	3563	Audubon Lake	16	1	3	4	Slab
456319.4	EXCE	3395	Audubon Lake	7	0	4	5	Slab
613739.0	EXCE	4489	Audubon Lake	25	0	5	5	Slab
891414.7	EXCE	5809	Audubon Lake	22	1	4	5	Slab
310843.6	VRGD	2572	Audubon Lake	19	1	2	4	Slab



Questions of Interest

- How do different attributes affect the value of a home being sold?
- How much does one subdivision sell for relative to the others all other being equal?
- When is a house for sale over or undervalued relative to the market?

Data Cleaning





Data Preprocessing

- Adjusting for inflation by updating prices at time of transaction to current USD
- Greenleaves Subdivision containing many smaller 'sub-Subdivisions'



Handling Missing Values

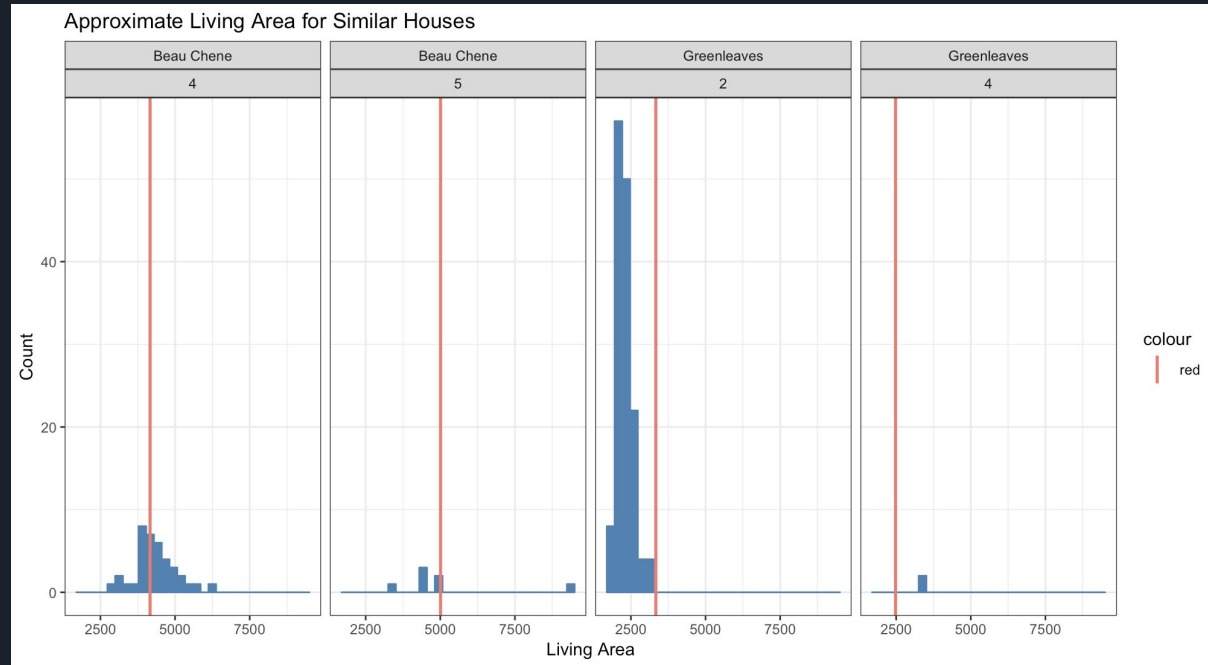
- Can fill in the missing values for half bath using half bath information from similar houses

Transactions With Missing Values

Subdivision	Approx.Living.Area	Beds.Total	Baths.Full	Baths.Half
Beau Chene	4162	5	4	NA
Beau Chene	5009	5	5	NA
Greenleaves	3335	4	4	NA
Greenleaves	2480	4	2	NA

Data Imputation

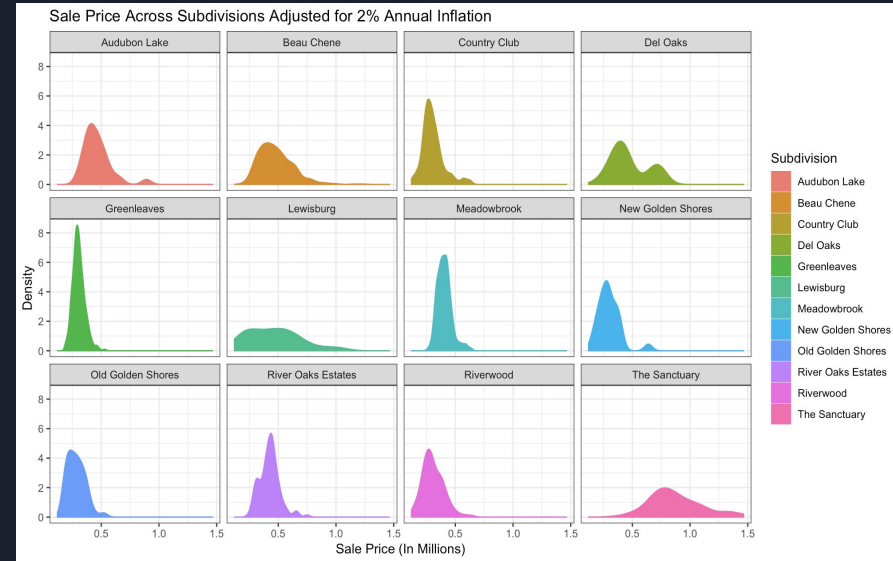
- Imputing the mode for the first observation
- Imputing the mode for the second observation
- Imputing the maximum for the third observation
- Imputing the minimum for the fourth observation



Exploratory Data Analysis

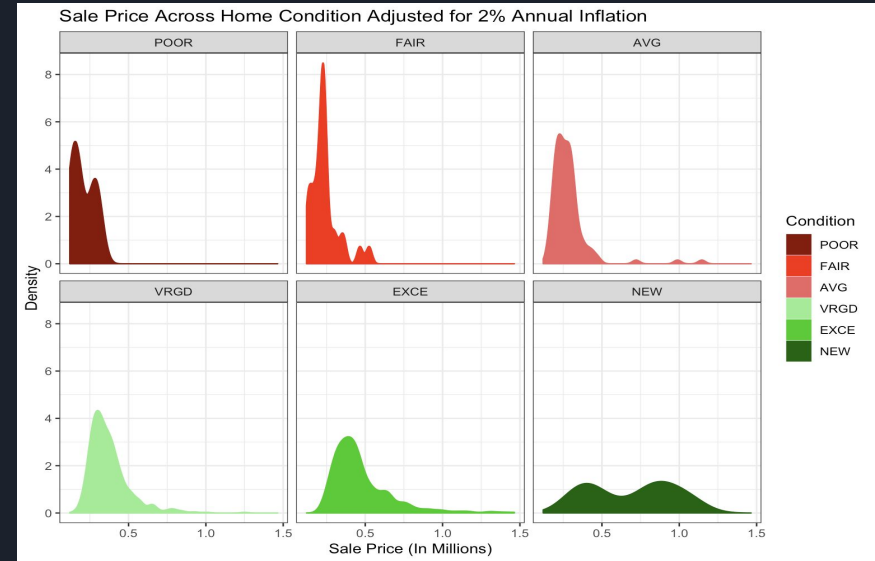
Sale Price by Subdivision

- Greenleaves, Meadowbrook, and Country Club all have very 'steep' plots suggesting that they are subdivisions which specialize in particular style homes.
 - Perhaps a single contractor worked entire neighborhood providing similar quality throughout.
- The Sanctuary has a wide variety of homes ranging from a half million to over 1.5 million.
 - People are able to buy land and build their own homes, thus, providing varying levels of quality/taste for similar sized homes.



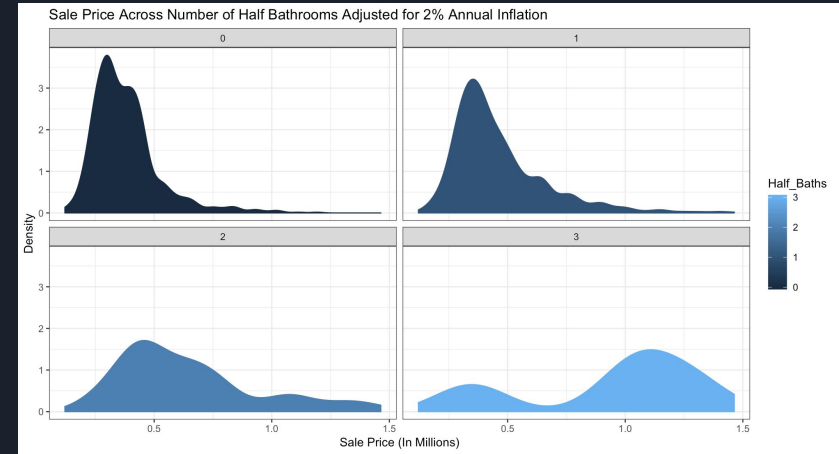
Sale Price by Condition

- As Condition of home improves, Sale Price increases.
- However, the difference between Excellent to Very Good and Very Good to Average appears marginal at best.
- Condition is measured by agents who want to promote the best possible condition of the homes they are selling. This subjectivity should be taken into consideration during analysis.



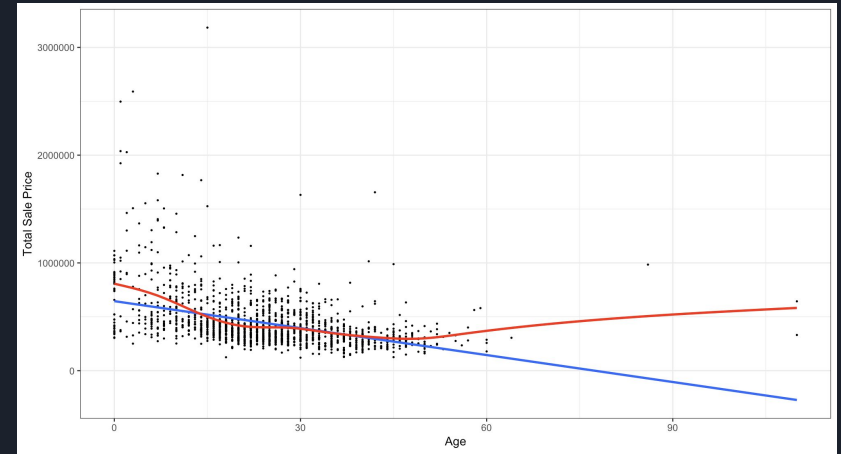
Sale Price by Number of Bathrooms

- As Half Baths are added, the distribution of Sale Price trends increasingly towards more valuable homes.
- In particular, moving from two Half Baths to three shows a significant bump in price.



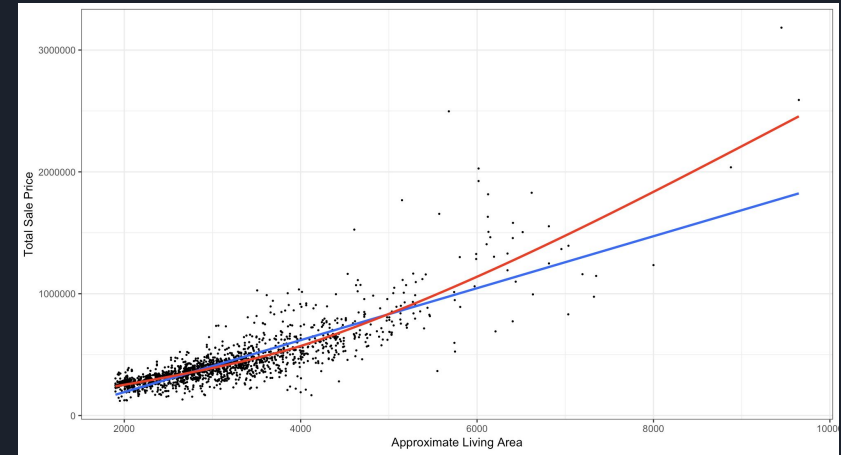
Relationship Between Age & Sale Price

- If the relationship between Age and Sale Price were truly linear, the scatter plot would track with the blue line.
- Age appears to be non-linearly related to Sale Price as the red curve appears to be more representative of the true relationship.
- Note: This is simply examining the relationship of Age and Sale Price INDEPENDENT of any other variables.

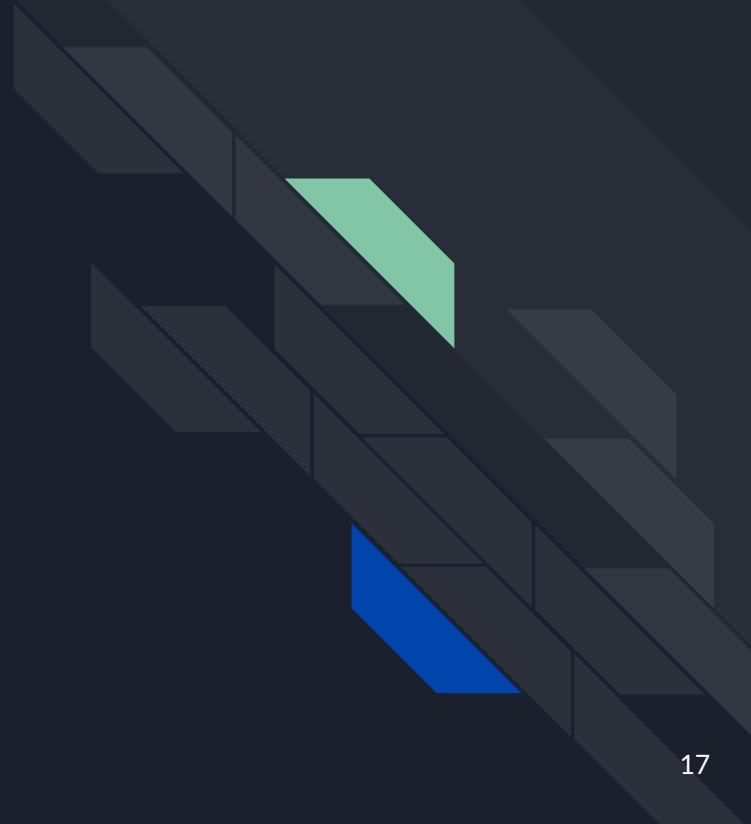


Relationship Between Living Area & Sale Price

- Similarly, Living Area and Sale Price appear to be non-linearly related.
- The red curve shows that the relationship may be truly exponential.
- Note: This is simply examining the relationship of Living Area and Sale Price INDEPENDENT of any other variables.



Initial Modeling





OLS (Ordinary Least Squares)

- While Multiple Linear Regression can be easily extended to handle non-linear relationships between X & Y , (see slide 30), first examine how modeling with a linear relationship would perform.
- Fitting OLS to the full data set yields mixed results.

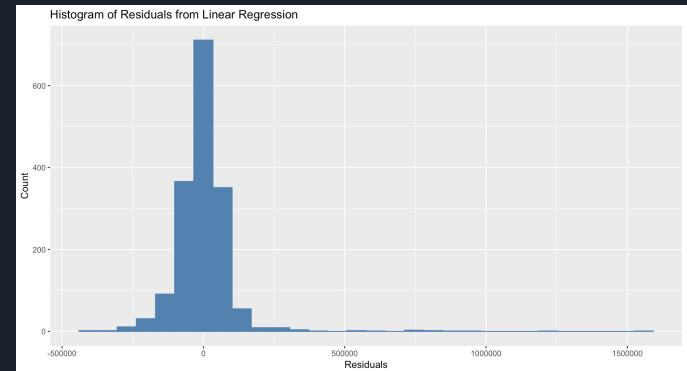
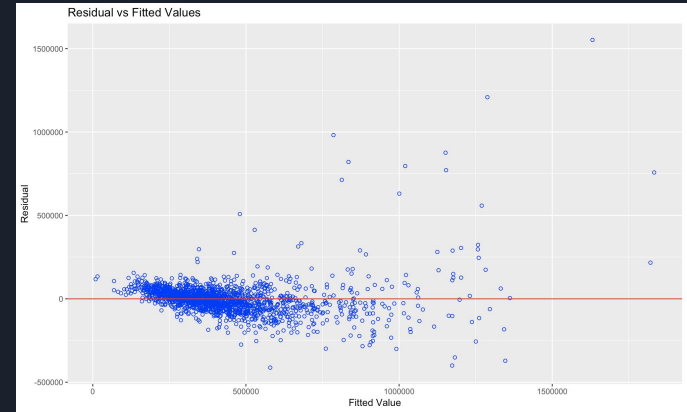


Collinearity Diagnostics

- Multicollinearity occurs when two or more of the explanatory variables are highly correlated causing unreliability in estimates.
 - If present, OLS would struggle to detect which explanatory variables are actually contributing to Sale Price.
- VIF for Living Area is 3.98. This means that 74% of Living Area can be written as a combination of other variables. Intuitively, this makes sense as living area typically increases as bedrooms and bathrooms are added.

Residual Diagnostics

- Residuals, which are the difference between actual Sale Price and predicted Sale Price, suggest that the relationship between our explanatory variables and Sale Price may be non-linear.
- Manually creating polynomial and/or interaction terms could help, although there are more preferred methods discussed in the following slides.



Cross-Validation & Model Tuning



Data Splitting

- First, split the entire data set into a classic 80% for Training and 20% for Testing.
- The **training set** will allow models to understand the underlying structure of the relationships between Sold Price and the explanatory variables.
- The **test set** will give an objective measure to determine if the models are accurately representing the true underlying structure.

Trade-off of Complexity vs. Flexibility



An Introduction to Statistical Learning

- Above is a representation of the tradeoff between flexibility and interpretability using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.



OLS:

Linear Model which seeks to estimate the average Sale Price given a set of explanatory variables

PROS

- Highly interpretable
- Easy to understand how predictions are made
 - Produces an equation exactly representing prediction process
- Typically performs well when the number of variables and observations is reasonably small

CONS

- Sensitive to extreme values
- Requires manual adjustments by analyst if attempting to appropriately model any non-linearities in the data
 - i.e. Any relationship between X & Y which is not constant



Ridge Regression:

Similar to OLS except imposes a complexity penalty on the model to stabilize estimates

PROS

- Handles multicollinearity amongst the explanatory variables
- Improves generalization to new data by stabilizing estimates of OLS
 - i.e. Small changes in data will not drastically affect predictions

CONS

- Requires manual adjustments by analyst if attempting to appropriately model any non-linearities in the data
- Similarly to OLS, Ridge Regression keeps variables in model even if they are not important.



MARS (**M**ultivariate **A**daptive **R**egression **S**plines):

Nonparametric regression extension of the linear model which repeatedly searches for cut points in data

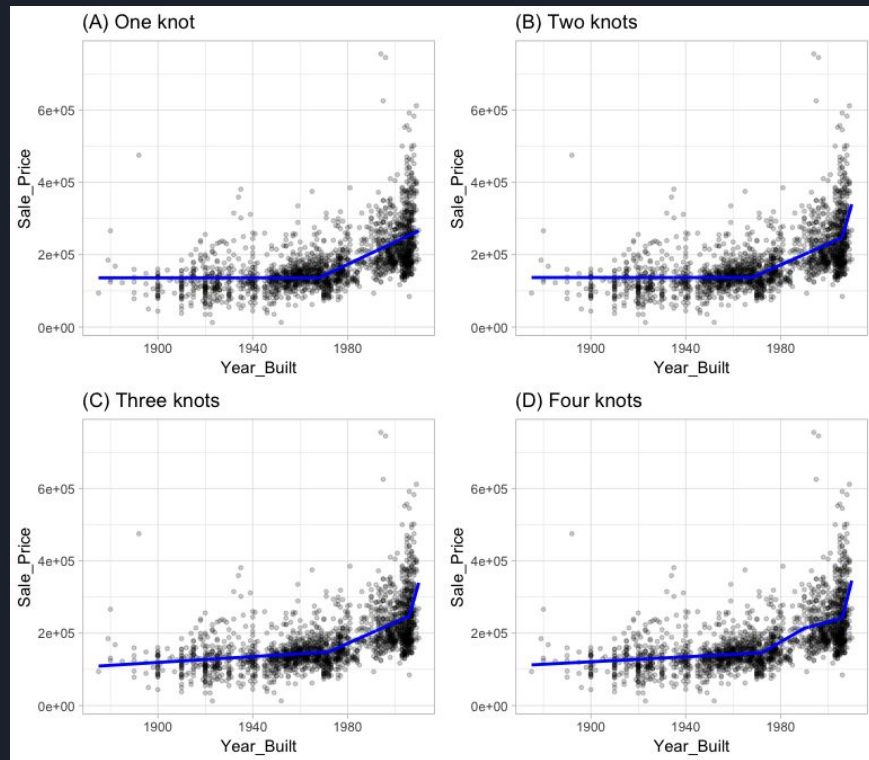
PROS

- Automatically detects interactions and non-linearities in the data
- Highly interpretable
- Automatic variables selection
 - i.e. It includes important variables in the model and excludes unimportant ones

CONS

- Multicollinearity can cause uncertainty about which variables to keep in model
- 'Greedy' process
 - i.e. Not guaranteed to find the best solution, but makes best decision at each step.

Visualizing MARS



Source: <http://uc-r.github.io/mars>



xgBoost:

Gradient boosting technique where decision trees are built sequentially with each compensating for the previous trees' weaknesses

PROS

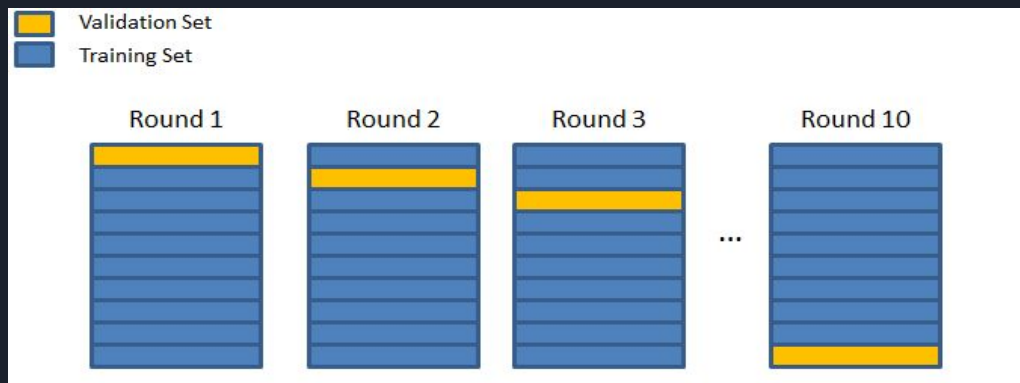
- Automatically detects interactions and non-linearities in the data
- Easily scalable to huge sets of data
- Highly flexible model which can capture the most complex data structures

CONS

- Extremely difficult to interpret directly
 - i.e. No equation to understand how predictions are made
- Many hyperparameters to tune, requiring much computation power to ensure optimal performance

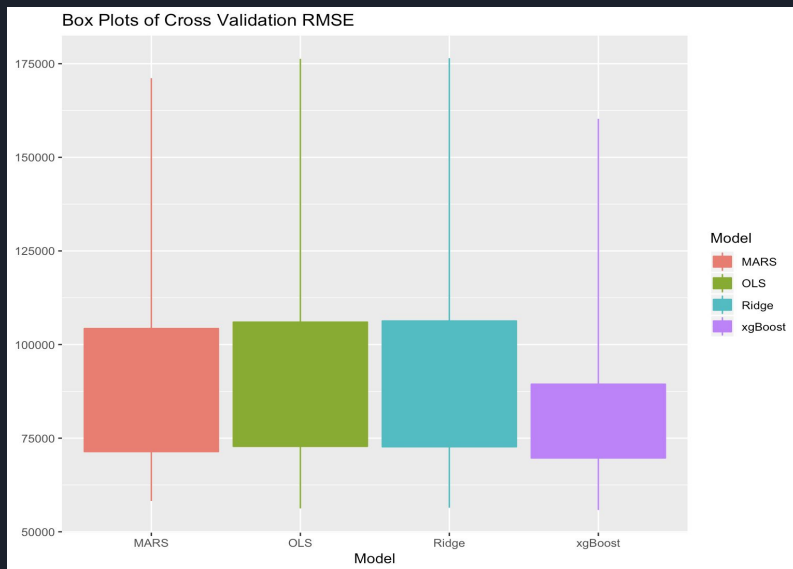
Cross Validation Comparison

- Ridge, MARS, and xgBoost have several hyperparameters that can be adjusted to ensure optimal performance. Cross Validation gives a chance to 'tune' the hyperparameters by seeing how each affects model predictions on external data.
- This is performed by repeated splitting of the training data into a smaller training set and a separate held-out validation set.
 - *Note: The Validation Set is not the Test Set, which is saved until after Cross Validation is performed.*
- Examining performance across Cross Validation gives a measure to show how each model will generalize to new data.



Cross Validation Comparison

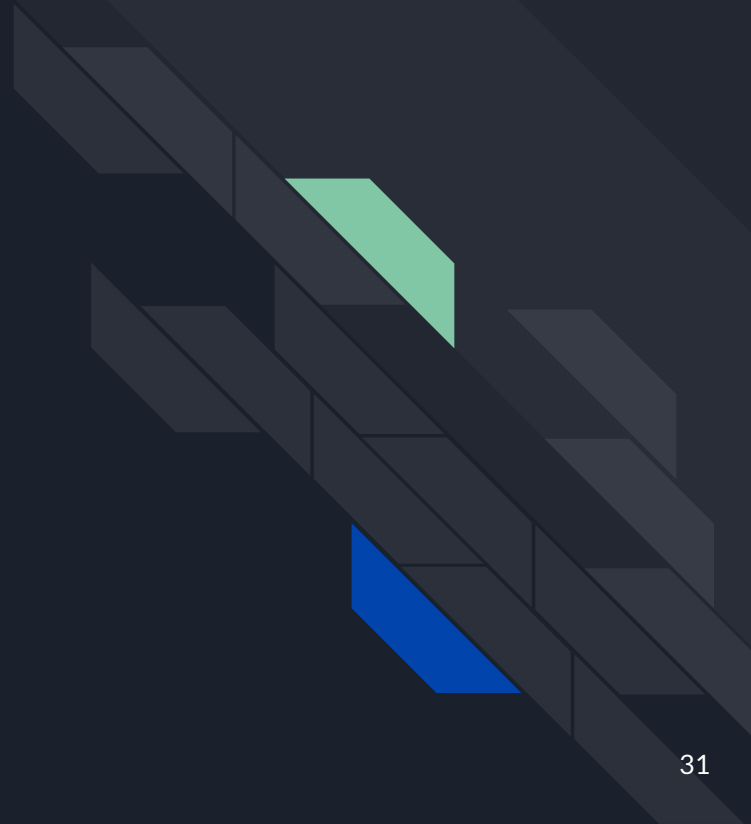
- xgBoost outperforms all other models in terms of Cross Validation with a median RMSE of 76,476.
- MARS is marginally behind xgBoost with Ridge and OLS performing with highest error.




RMSE on Cross Validation

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
OLS_Full	56241	72802	83042	93460	105995	176320
Ridge_Full	56428	72693	82869	93465	106305	176473
Mars_Full	58224	71402	82010	89757	104284	171135
XG	55808	69695	77600	85446	89426	160282

Model Inferences



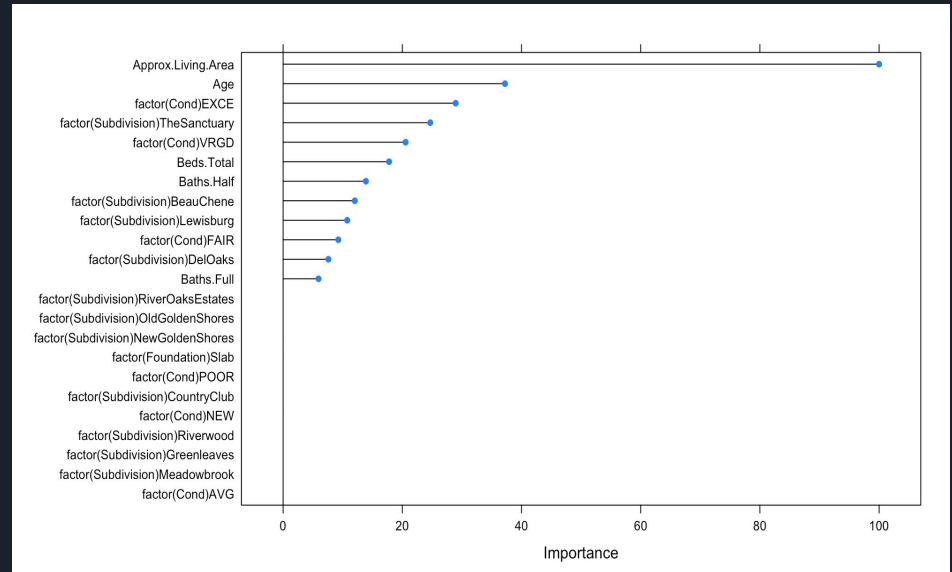


How are MARS and xgBoost making its predictions?

- Understanding how MARS/xgBoost are making decisions can also give insight to which attributes may be important in predicting real estate value.
- Perhaps, MARS/xgBoost can uncover some of the non-linearities which can be used to manually update OLS/Ridge with polynomial terms. Because of this, OLS/Ridge is likely to gain predictive power while maintaining its incredible interpretability.

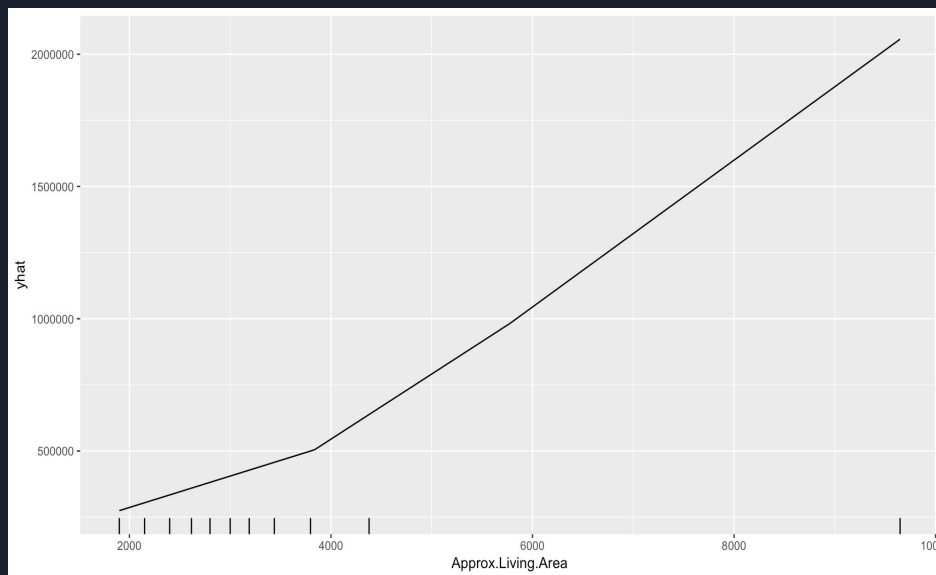
Variable Importance According to MARS/xgBoost

- Approximate Living Area and Age appear to be most important in predicting the Sale Price.
- Additionally, the relationship can be examined between variables and Sale Price by looking at *Partial Dependence Plots* produced by MARS and xgBoost.



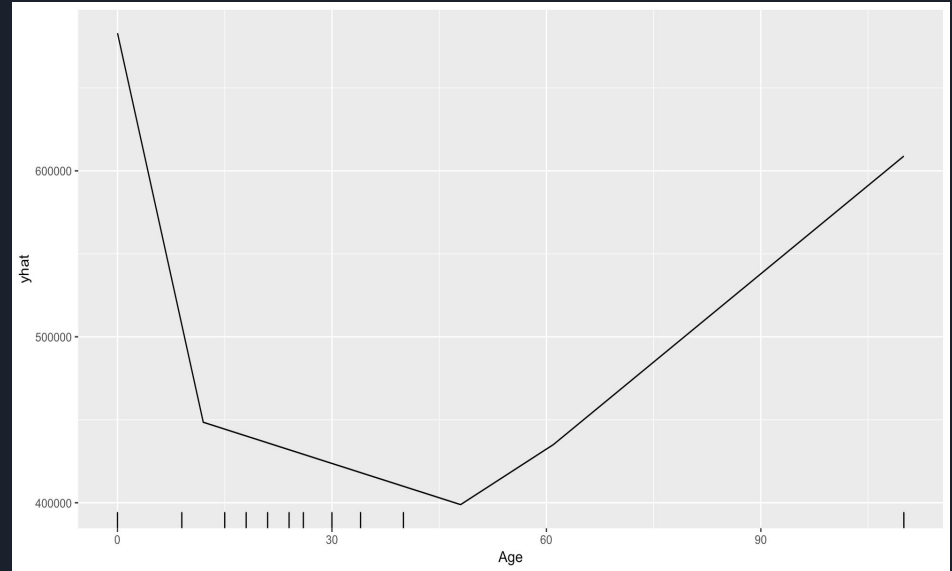
Partial Dependence of Living Area

- Partial Dependence Plots allow us to examine outcomes on predicted Sale Price while changing one variable *and controlling for all others*.
- MARS/xgBoost suggest that there is a change in effect on Sale Price when Living Area is around 4,000. This suggests a soft exponential relationship.
- *Note: All partial dependence plots were approximately the same for xgBoost and MARS.*



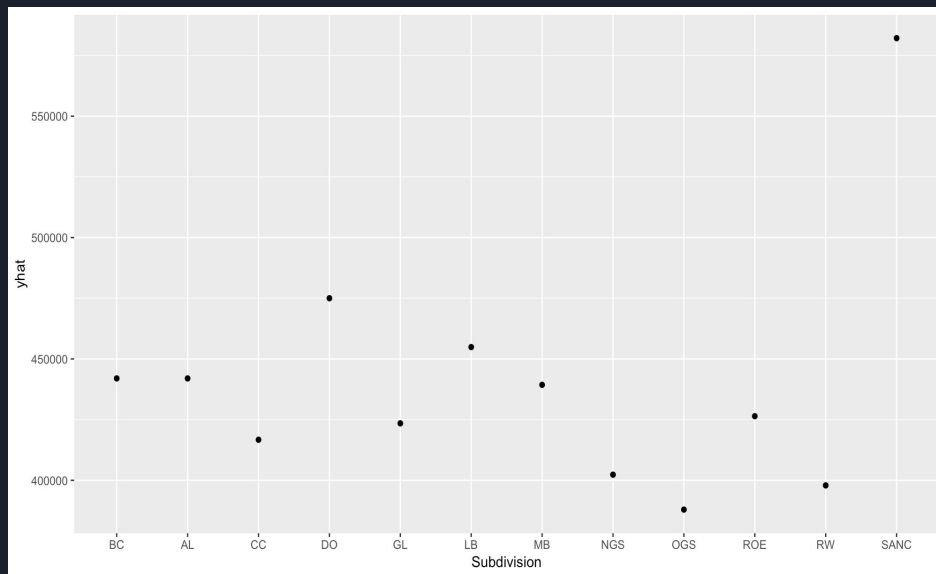
Partial Dependence of Age

- MARS/xgBoost suggest that the effect of Area on Sale Price changes around 8 years and again around 50 years. This shows a *quadratic* relationship.
- In Southern Louisiana, this is reasonable as historic architecture is very highly valued. Thus, homes with a higher age may produce a 'landmark' appeal to buyers.
- Rapid depreciation in home value after new home is built



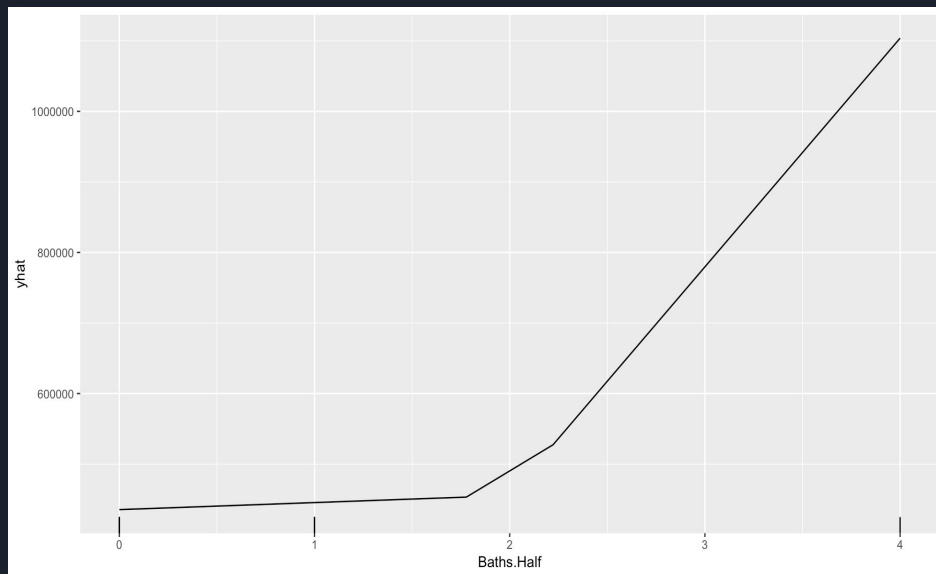
Partial Dependence of Subdivision

- MARS/xgBoost suggest that there are roughly three classes of subdivision when considering effects on home price:
 - The Sanctuary can be considered a '**high class**' neighborhood having a significant positive effect on predicted sale price.
 - Old Golden Shores, New Golden Shores, and Riverwood can be considered an '**economy class**' neighborhoods, having a marginal negative effect on predicted values.
 - The difference between the rest of the neighborhoods is much smaller..



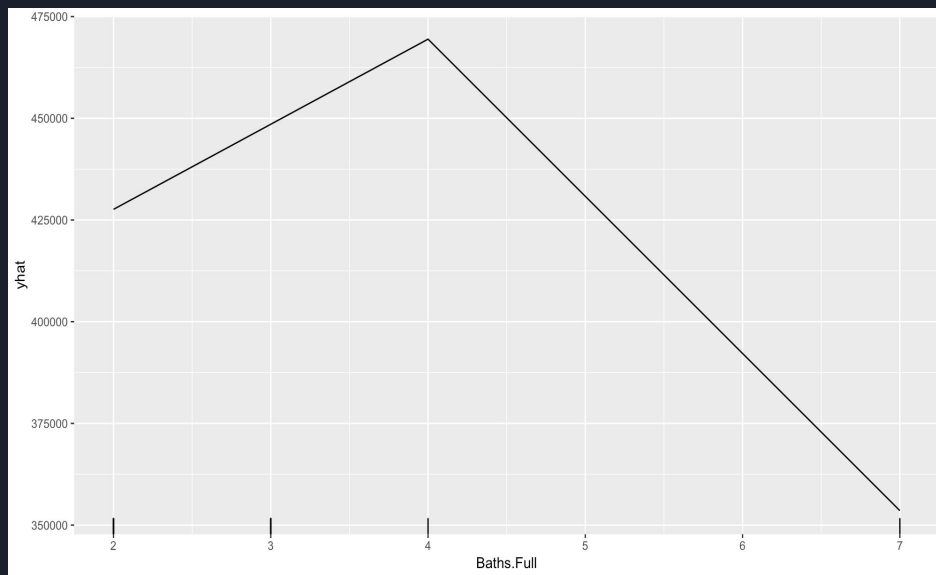
Partial Dependence of Half Baths

- MARS/xgBoost suggest that adding additional Half Baths after already having two, contribute significantly to the home's value.
- Homeowners can consider construction costs by way of adding new space to the house or converting existing space.



Partial Dependence of Full Baths

- MARS/xgBoost suggest that adding up to 4 Full Baths will add value to the home.
- However, after adding 4 full bathrooms, homeowners may prefer to focus on increasing home value in other ways (perhaps, a Half Bath).



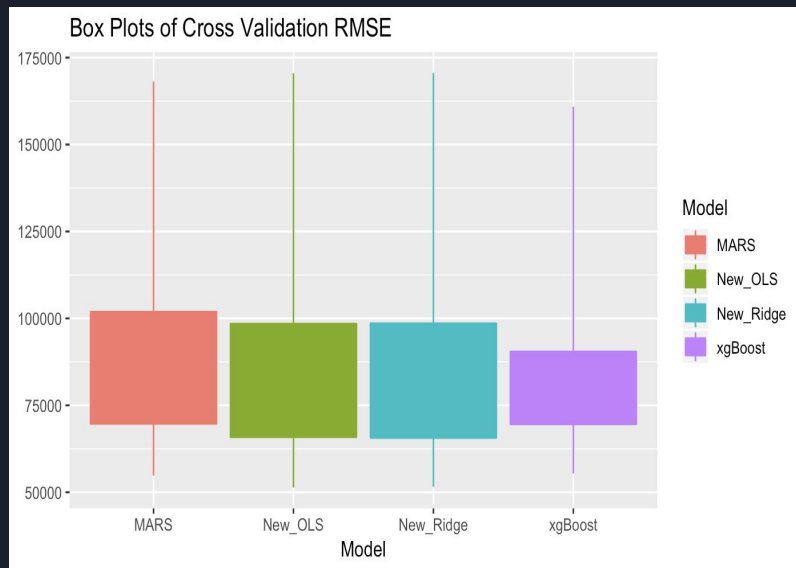


What does this mean?

- MARS/xgBoost suggests that the relationship between Age, Living Area, and Half Baths with Sale Price is non-linear. Additionally, these are three of the important variables (according to MARS/xgBoost).
- However, previously OLS and Ridge Regression modeled the relationships between Age, Living Area, and Half Baths with Sale Price as linear. This likely contributed to its poor performance during Cross Validation. It was modeling 'important' variables *inappropriately*.
- How would 'updating' the OLS and Ridge models with appropriate polynomial terms improve their performance?

Revisiting OLS and Ridge Regression

- It is important to revisit Cross Validation error with Ridge and OLS now that appropriate polynomial terms have been found.
- The mean RMSE Cross Validation error for OLS and Ridge improved by roughly 8,000.
Linear models can effectively model non-linear relationships!
- Additive structure of OLS/Ridge seems to also be advantageous.



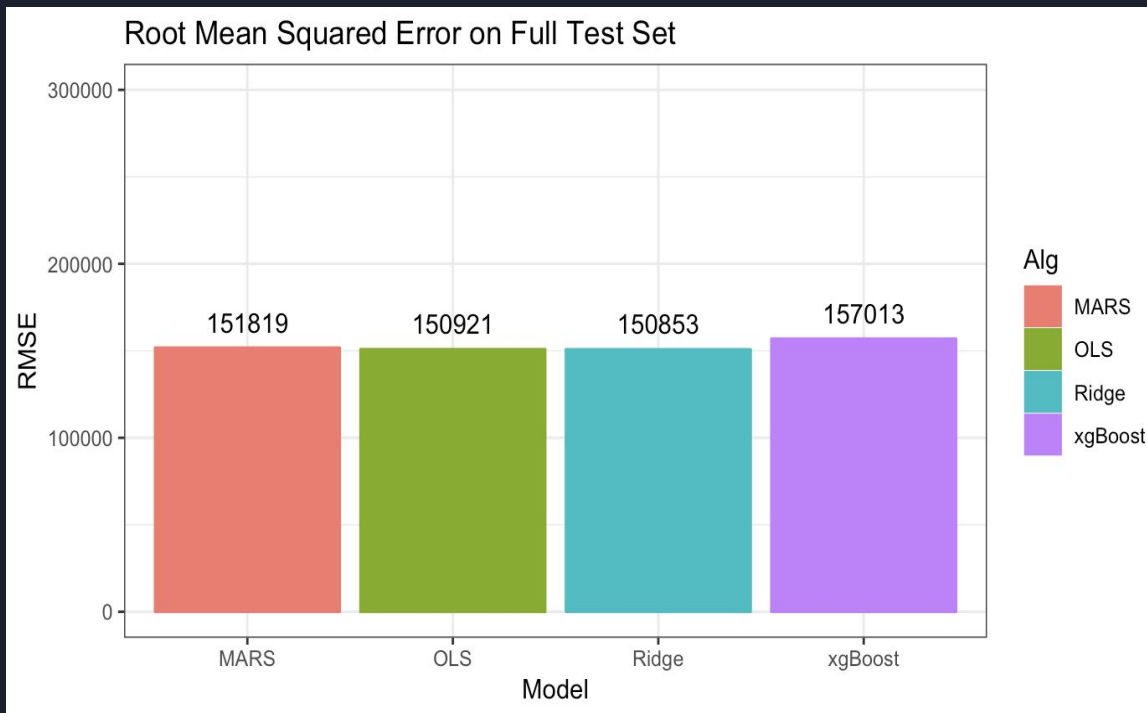
RMSE on Cross Validation

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
OLS_Full	51342	65889	79815	85871	98427	170519
Ridge_Full	51592	65659	79672	85867	98536	170622
Mars_Full	54791	69714	81235	87812	101871	168104
XG	55388	69591	76476	85282	90429	160892

Test Set Performance

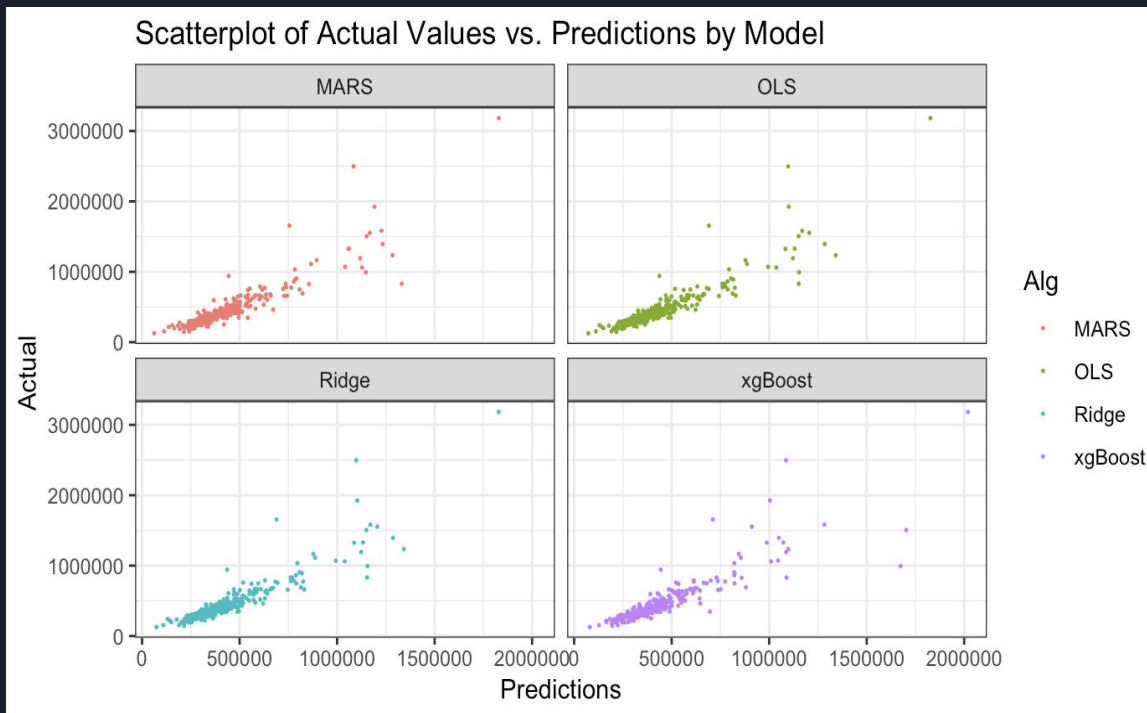
Test Set Comparison

- Comparing each model's performance on an independent test set allows comparison of how each model generalizes when making predictions about data which it has not 'seen'.
- Performance appears to be similar with xgBoost having the highest level of error.
- Test set performance is significantly worse than Cross Validation. What happened?



Examining Where Our Models Struggled

- Examining a scatterplot of each model for the *predictions* of Sale Price vs. *actual* Sale Price compares each model's predictions and true values.
- Ideally, if models predicted actual Sale Price perfectly then the scatterplot would march on a diagonal line.





Examining Where Models Struggled

- Model predictions appeared to track well with less expensive homes, yet, had tremendous difficulty in predicting more expensive homes. Why?
- Many possible reasons:
 - More expensive homes vary in ways that were not measured in the data set (i.e. tennis courts, home movie theatres, pools, etc.).
 - Unmeasured variables contributing to the real estate costs; including lot size, insurance prices, and interest rates at time of transactions.
- One approach would be to separate home value across different strata to see if any models perform particularly well.

Test Set Comparison Segmenting by Home Value

Sold Price can be split into:

1st quartile: Middle Class Homes

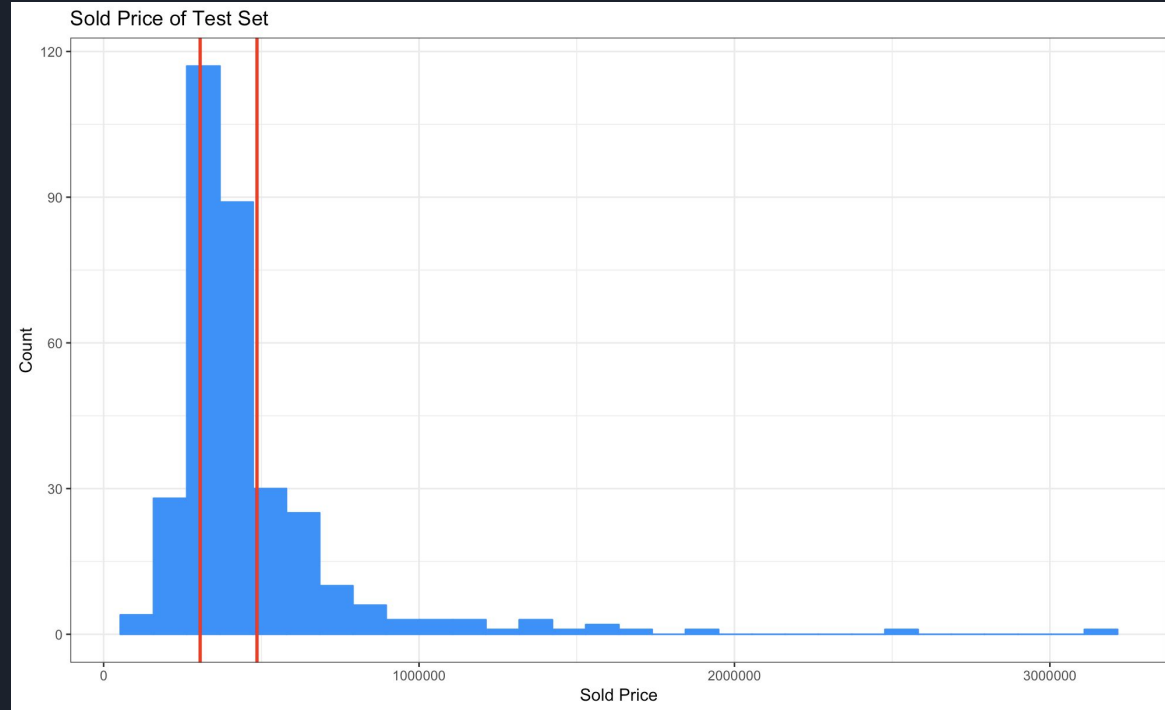
Sale Price < \$306,000

1st to 3rd quartile: Upper-Middle Class Homes

\$306,000 > Sale Price < \$486,000

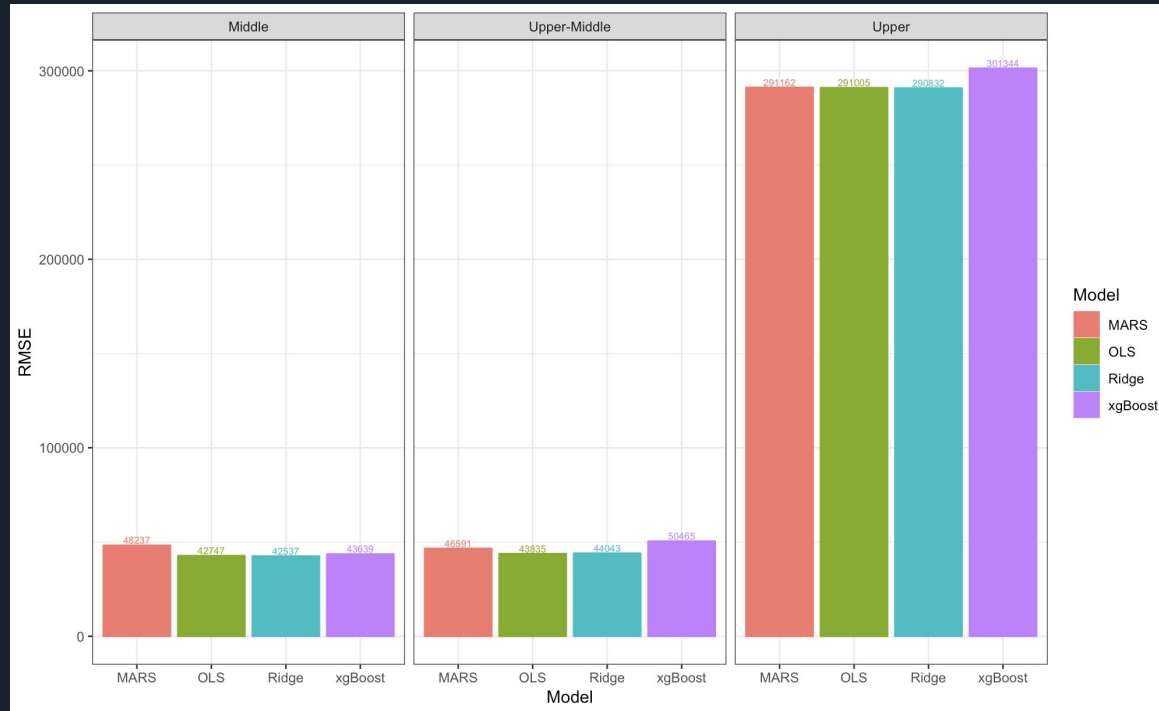
Above 3rd quartile: Upper Class Homes

Sale Price > \$486,000



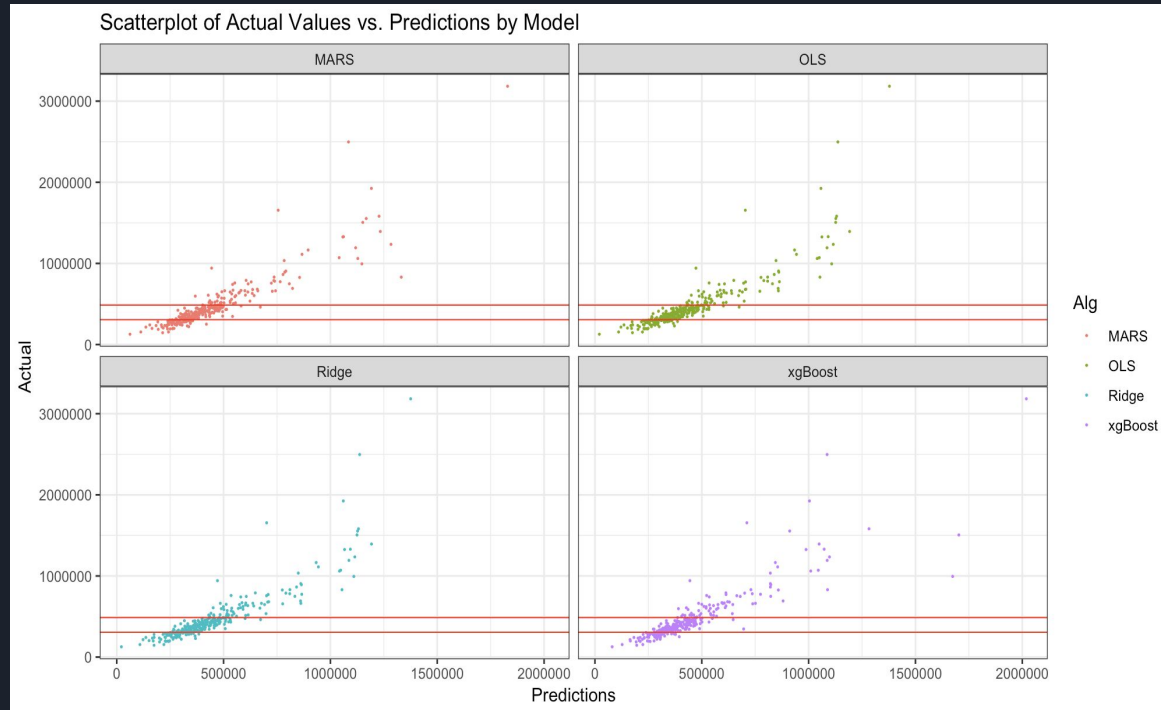
Test Set Splitting after Binning

- OLS, Ridge, and xgBoost all outperform MARS for **Middle Class Homes**.
- OLS and Ridge outperform both MARS and xgBoost on **Upper-Middle Class Homes**.
- All models struggles to accurately predict **Upper Class Homes** (likely for aforementioned reasons).
- One consideration would be to build *different models* for different price stratta of houses.



Alternate View of Model Performance

- These visual perspectives show each model's performance across the different strata of homes:
 - Below the red line represents accuracy of predictions for **Middle Class Homes**.
 - Between the red lines represents accuracy of predictions for **Upper-Middle Class Homes**.
 - Above the red line represents accuracy of predictions for **Upper Class Homes**.





Which model is 'best'?

- Depends on goals and metrics of business problem
- Cross Validation showed xgBoost had, *on average*, the best performance on new data. However, this specific test set clearly gave xgBoost problems as linear models performed best.
 - Resampling a new test set may yield more encouraging results for xgBoost
- Do we want to know *what* the prediction is or *why* the predictions was made?
 - Sometimes, knowing *why* can help in learning more about the problem, the data, and the reason why a model might fail.
 - Additionally, understanding *why* can help ensure fairness, privacy protections, and build human trust of models.

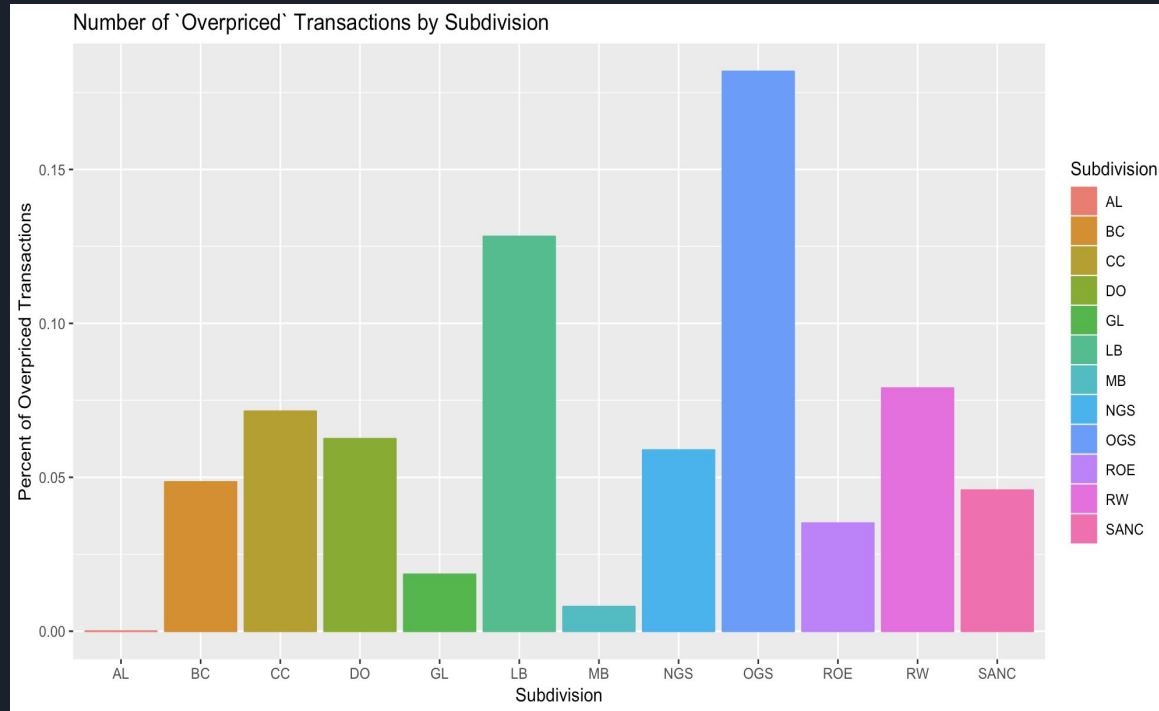


Flagging Transactions

- We can examine if any transactions were 'overpriced' or considered to be a 'bargain' where:
 - 'Overpriced' is defined as the buyer paying $>110\%$ of the home value predicted by model.
 - 'Bargain' is defined as the buyer paying $<90\%$ of the home value predicted by model.
- We can possibly discover:
 - Are certain realtors more likely to get 'bad' or 'good' deals for their clients?
 - Are certain neighborhoods more likely to have 'overvalued' or 'undervalued' homes relative to the market?
- Prior considerations:
 - Homes vary in ways outside of data measured for modeling.
 - A house could be someone's 'dream home,' it could be in a great school district, or could have a neighborhood pool. Such reasons could encourage buyers to pay a higher price.
 - Homeowners could seek to liquidate assets quickly, which leads them to sell a house below its true value.
- Inherent element of uncertainty in *every* statistical model

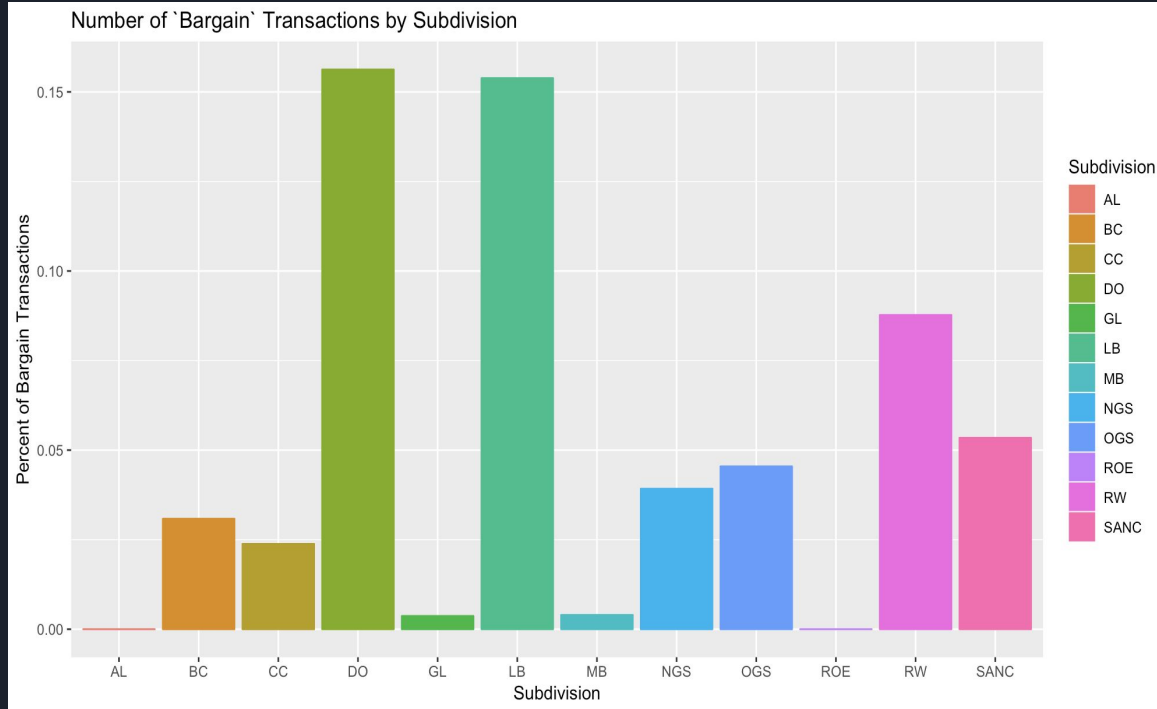
Overpriced Transactions By Neighborhood

- More likely to 'overpay' for homes in Lewisburg, Old Golden Shores, and Riverwood
 - Buyers should examine whether the extra price relative to other subdivisions is worthwhile.
- Audubon Lake and Meadowbrook appear to have little to no 'overpriced' transactions suggesting a safer market for buyers.



Bargain Transactions By Neighborhood

- More likely to get a 'bargain' for homes in Del Oaks, Lewisburg and Riverwood suggesting more for the buyer's money, all other things being equal.
- Lewisburg and Riverwood were flagged frequently in both 'bargain' and 'overpriced' transactions. This suggests their prices may deviate from true market value more often.
- Interestingly, Audubon Lakes and Meadowbrook were flagged infrequently in 'bargain' or 'overpriced' transaction suggesting their prices may be consistently close to the true value.





'Bargain' or 'Overpriced' by Selling Agent

BARGAIN

- Over seventy selling agents sold a bargain at least once.

OVERPRICED

- Four selling agents were able to sell homes at >10% of estimated prices for ***at least two*** clients.
- Over forty other selling agents were able to get homesellers an advantage at least once.

Final Thoughts



Considerations

- Model transparency is extremely important in certain circumstances. Thus, having an interpretable model that predicts as well as a complex model could be very useful.
- Data was not collected in a homogenous way. Separate realtors (with varying interests/motivations) recorded aspects of the home based on their own judgement.
 - i.e. Home Condition is *extremely* subjective.

“All models are wrong, but some are useful.”

- George E.P. box



Limitations

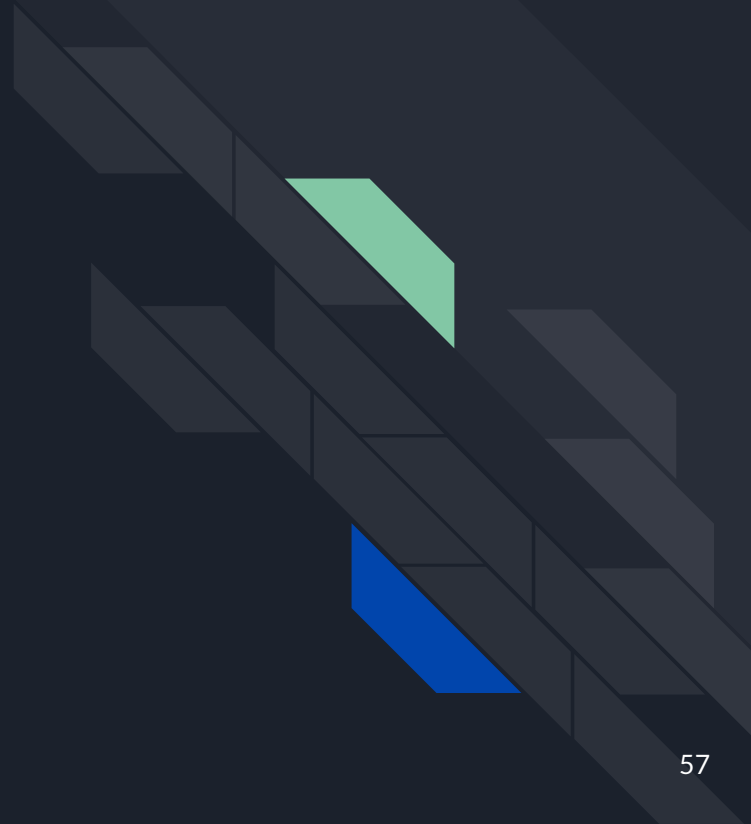
- Generalization of these models to neighborhoods outside of those presented in the data set is not recommended.
- Missing other helpful information about homes such as home attributes that improve or reduce home value



Future Work

- Creating different models for certain subgroups of real estate markets
 - For example, different models for different wealth class, cities, ages, proximity to water, etc.
- Identifying which are the 'best' selling agents for a particular homeseller based on prior transaction records

Appendix





OLS Estimated Regression Function

- Neighborhood, Condition, and Foundation can take values 0 or 1.

$$\widehat{SoldPrice} = 483398 + 3329BeauChene - 37788CountryClub + 37355DelOaks - 31565Greenleaves + 4283Lewisburg - 22087Meadowbrook - 48495NewGoldenShores - 73168OldGoldenShores - 38156RiverOaks - 45539Riverwood + 132106Sanctuary + 394584FullBaths - 356306FullBath^2 + 466625HalfBaths + 751150HalfBaths^2 - 23799Beds - 834128Age + 694015Age^2 + 6308502LivingArea + 1884190LivingArea^2 + 17907Avg + 114676Exce - 71112Fair + 211962New - 30562Poor + 72698VeryGood - 29311Slab$$



Ridge Estimated Regression Function

- Neighborhood, Condition, and Foundation can take values 0 or 1.

$$\widehat{SoldPrice} = 553039 + 4816BeauChene - 35990CountryClub + 39023DelOaks - 29930Greenleaves + 5927Lewisburg - 20583Meadowbrook - 46750NewGoldenShores - 71513OldGoldenShores - 36686RiverOaks - 43782Riverwood + 133560Sanctuary + 400097FullBaths - 358882FullBaths^2 + 466884HalfBaths + 751410HalfBaths^2 - 23826Beds - 882974Age + 694779.54574Age^2 + 6308502LivingArea + 1884190LivingArea^2 - 55466Avg + 41345Exce - 144462Fair + 138371New - 103939Poor - 543VeryGood - 29319Slab$$



MARS Estimated Function

- Neighborhood, Condition, and Foundation can take values 0 or 1.
- 'h' is a Hinge Function which takes the form $\max(0, x-c)$ or $\max(0, c-x)$ where 'c' are the respective knots.

$$\widehat{SoldPrice} = 521403 + 277h(LivingArea - 4219) - 118h(4219 - LivingArea) - 1378h(Age - 8) + 28621h(8 - Age) + 90590Excellent + 150577Sanctuary + 330962h(HalfBaths - 2) - 126308h(Beds - 5) + 12054h(5 - Beds) + 53453VeryGood + 4924h(Age - 50) + 31618BeauChene - 94202Fair - 41226h(FullBaths - 4) - 18277h(4 - FullBaths) + 63555DelOaks - 35924Slab$$