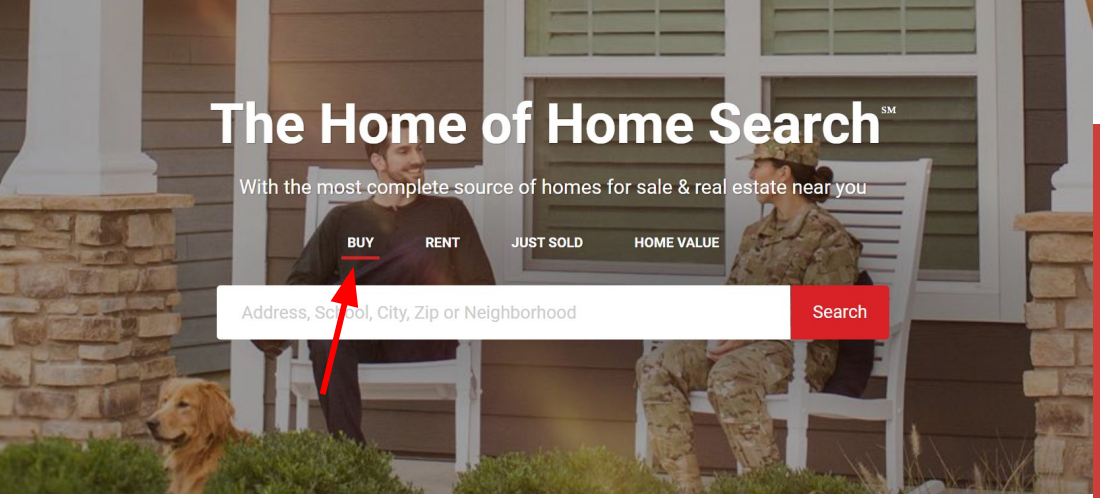
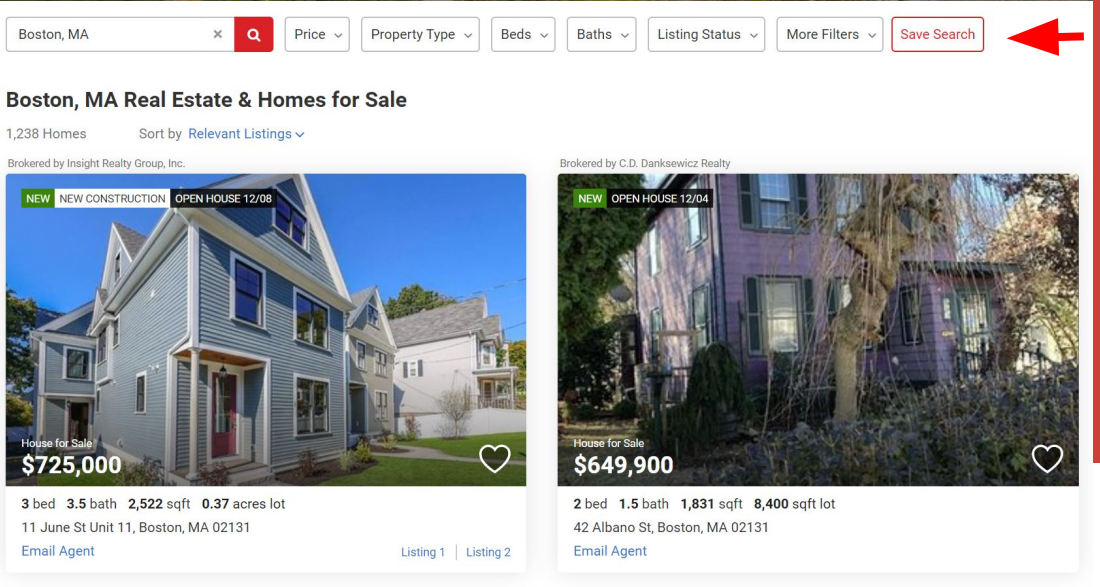
Realtor.com: Predicting Home Selling Price

Presented by
Krishna Ammini, Erica Chin, Alexandros Gounarakis,
Nicholas Lorello, Lingxiao Lyu





Data Source

Realtor.com

- Real estate listing site operated by Move, Inc.
- Founded in 1995 as the Realtor Information Network
- Allows users to browse houses for sale/rent/just sold in an area code

Sold Price

- Response variable (myresponse)
- The seller sets the price
- The buyer can negotiate, so the listed price can differ from the actual sale price

BUY

HOME BUYER

Home buyers and their agents want to make sure they are getting the best deal for the homes they are looking for. They can use the information we provide to potentially negotiate better (lower) prices

SELL

HOME SELLER

Home sellers would find price evaluation useful so they have a better idea of fair price ranges for their houses. In addition, sellers can determine if they should add an attribute to their homes to increase the price.

DATA EXPLORATION

Data preparation + cleaning

Data Overview

- 604 Observations
- 13 Variables

Variable Name	Description	Information
Sold.price:	the price in dollars that the house was sold for will be our response variable	Numeric
Zip code	the zip code (location) of the house	Factor (80)
Overview	Text description of inside and outside properties of the house	Factor (592)
Type	whether each house is a single family home or condo/townhouse	Factor (2)
Year built	the year that the house was built (year-ie 2019)	Factor (122)
Num.beds	number of beds in the house	Numeric
Num.baths	number of bathrooms in the house (.5 is only bathroom no shower)	Numeric
Living.area	size of living space within the house in sq. feet	Numeric
Lot.area	size of land outside the house on the property (sq. feet and acres)	Numeric
Masterbedroom.length	master bedroom's length in feet	Numeric
Masterbedroom.width	master bedroom's width in feet	Numeric
Kitchen.length	kitchen length in feet	Numeric
Kitchen.width	kitchen width in feet	Numeric

Data Cleaning/Preparation

Variables Omitted:

- Overview

Variables Changed:

- Kitchen Length * Width= Kitchen Sq.Ft.
- Master-Bedroom Length * Width= Master-Bedroom Sq.Ft.
- 2019- Year Built= Age of Home

Variable Modification

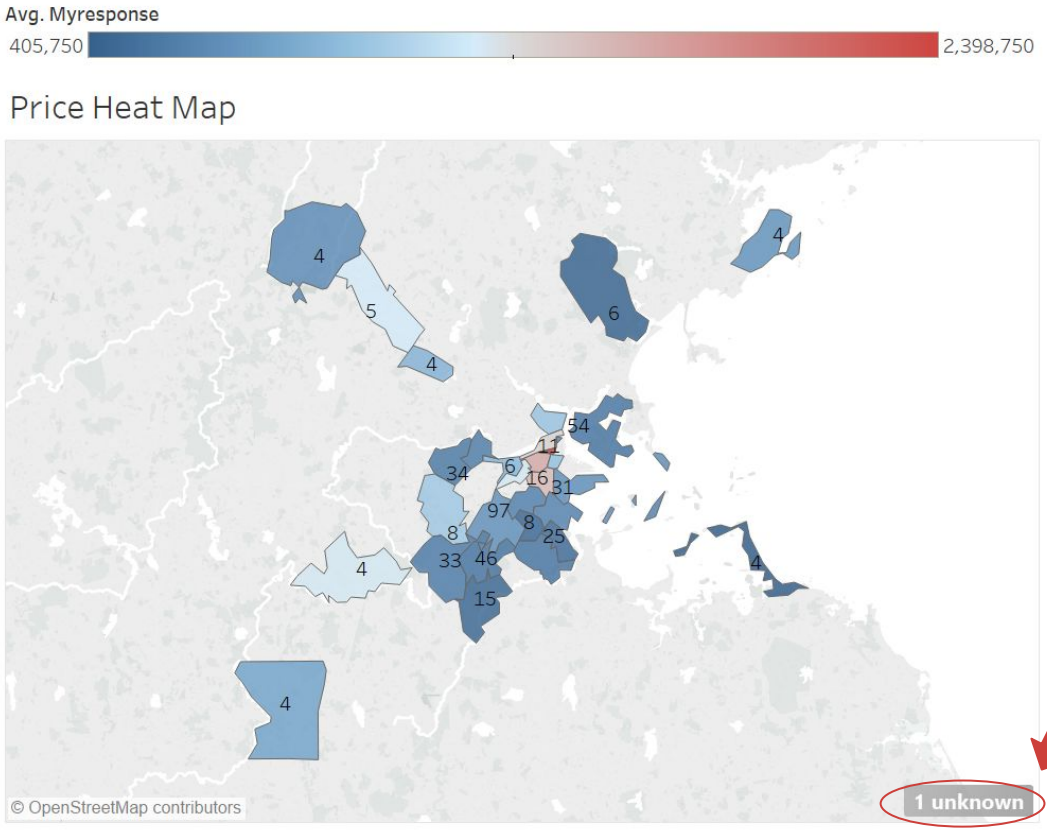
- Sold Price (remove N/A)
- Zip Code (80 to 32)
- Lot Size (all to acres and N/A to 0)

Final Dataset

- 588 Observations
- 10 Variables

Variable Name	Description	Information
Sold.price:	the price in dollars that the house was sold for will be our response variable	Numeric
Zip code	the zip code (location) of the house	Factor (32, 80)
Type	whether each house is a single family home or condo/townhouse	Factor (2)
Age of Home	Numerical value representing the home's age	Numeric
Num.beds	number of beds in the house	Integer
Num.baths	number of bathrooms in the house (.5 is only bathroom no shower)	Numeric
Living.area	size of living space within the house in sq. feet	Numeric
Lot.area	size of land outside the house on the property in acres	Numeric
Masterbedroom Sq.Ft.	master bedroom's size in squared feet	Numeric
Kitchen Sq.Ft.	kitchen size in squared feet	Numeric

Zip Code Map



Other

Price and Predictors

CART

1. Regression Tree

RF

2. Random Forest

BOOSTING

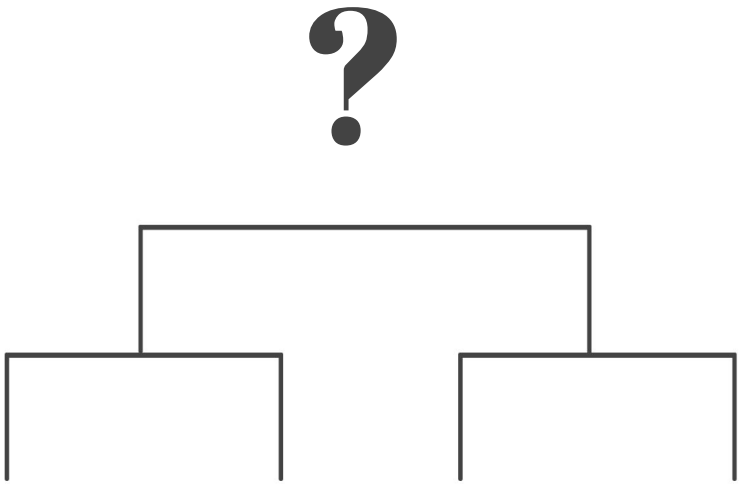
3. Boosted Trees

ANN

4. Artificial Neural Network

CART

Regression Tree

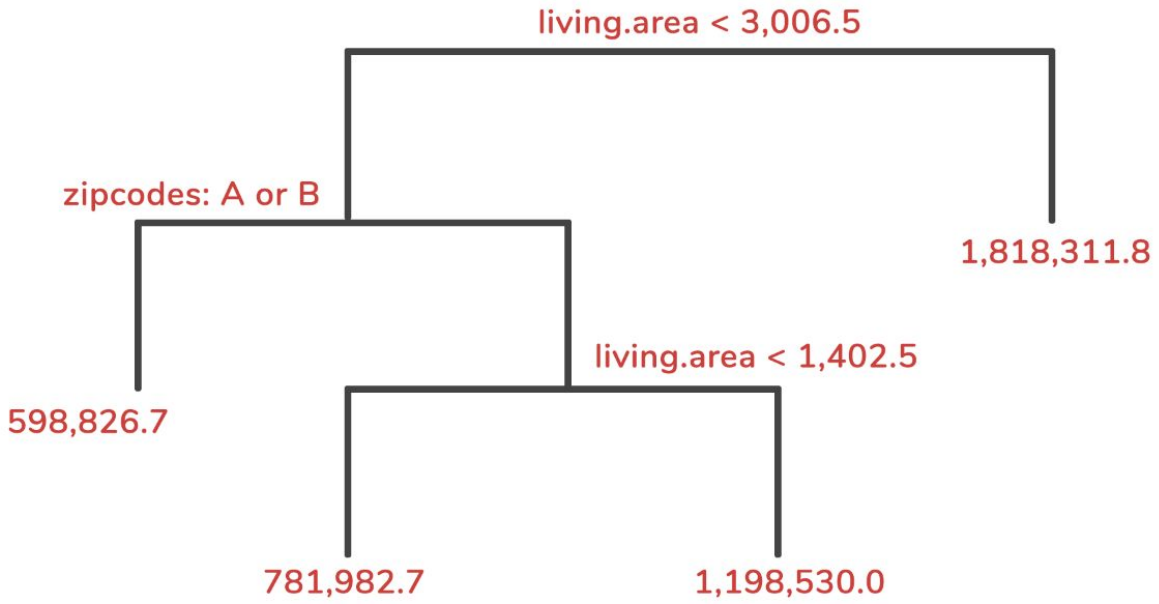


CART Tree

- Tree Type: Regression
- Minimum Leaf Size: 30
- Minimum Deviance: 0.025

see appendix for R code

CART Tree



A	01730, 01906, 01945, 02045, 02052, 02113, 02119, 02121, etc.
B	02108, 02111, 02114, 02115, 02116, 0218, 02129, 02215, etc.

see appendix for full list of zipcodes

Example: 105 Fletcher Rd, Bedford, MA 01730

Zip Code: 01730
Type: Single Family
# of Beds: 3
# of Baths: 2
Living Area: 2,000
Lot Area: 0.27
Kitchen Sq. Ft: 400
Master Bedroom Sq. Ft: 130
Age of Home: 59

Sold Price: 725,000

Sold on August 9, 2019

Map | 3 beds | 2 baths | 2,000 sq ft | 0.27 acres lot

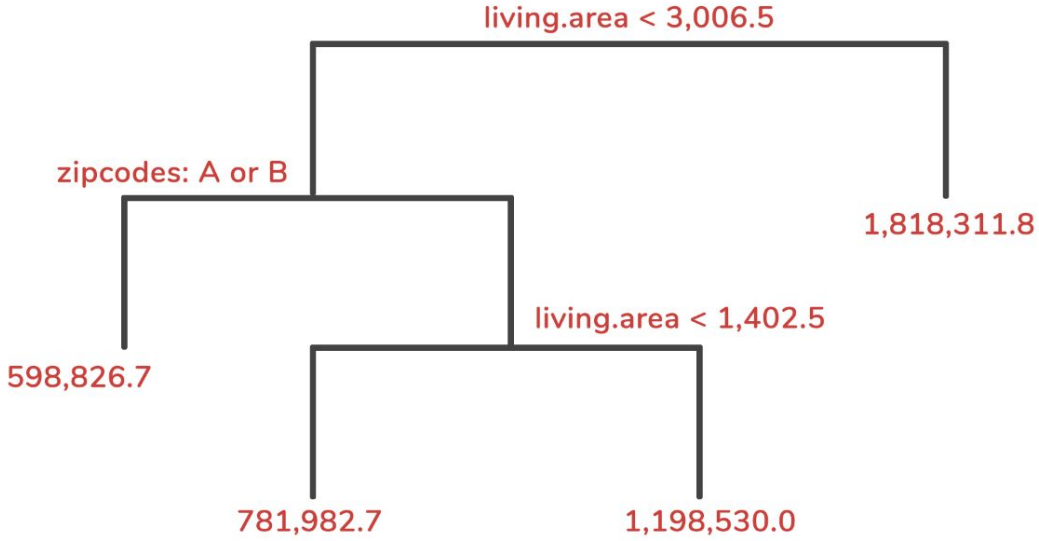
Last Sold for **$725,000**

Commute Time | 105 Fletcher Rd, Bedford, MA 01730 | Track Your Home Value

Example House

Zip Code: 01730
Type: Single Family
# of Beds: 3
# of Baths: 2
Living Area: 2,000
Lot Area: 0.27
Kitchen Sq. Ft: 400
Master Bedroom Sq. Ft: 130
Age of Home: 59

Actual Sold Price: 725,000

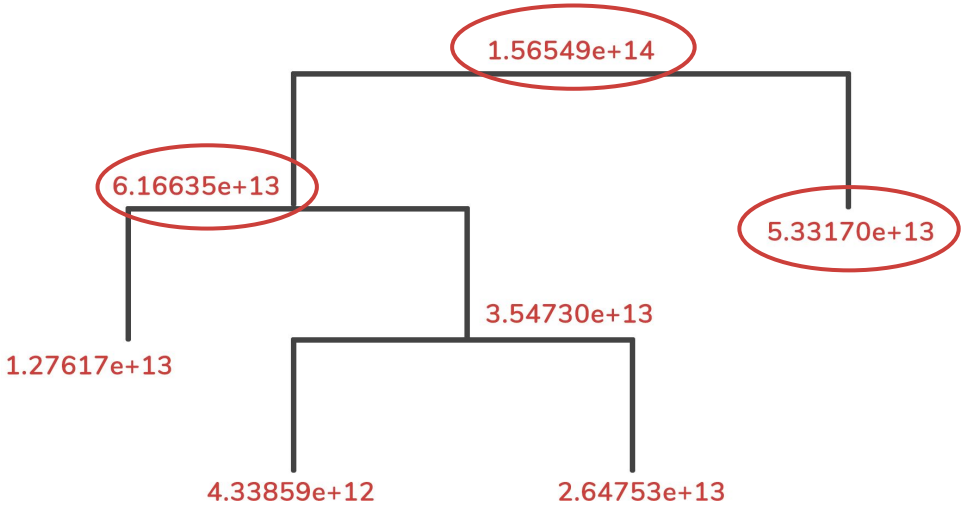


A	01730, 01906, 01945, 02045, 02052, 02113, 02119, 02121, etc.
B	02108, 02111, 02114, 02115, 02116, 0218, 02129, 02215, etc.

see appendix for full list of zipcodes

Predicted Selling Price of Home: **598,826.7** *Error: 598,826.7 - 725,000 = -126,173.3*

Deviance of nodes



How a split is determined?

➢ **Largest reduction in impurity**

Example: Root Node (Living Area < 3,006.5)

6.16635e+13 *left deviance*	+	**5.33170e+13** *right deviance*	=	**1.14980e+14** *sum of left+right*	
1.56549e+14 *root deviance*	-	**1.14980e+14** *sum of left+right*	=	**4.15685e+13** *largest reduction in impurity*	

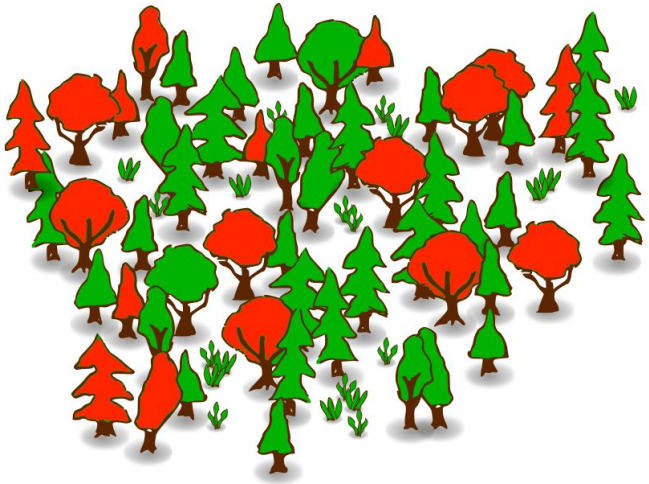
CART Model: Analysis

MAPE is 30.73%
Mean absolute percentage error

RMSE is 325,443.68
Square root of the variance of the residuals

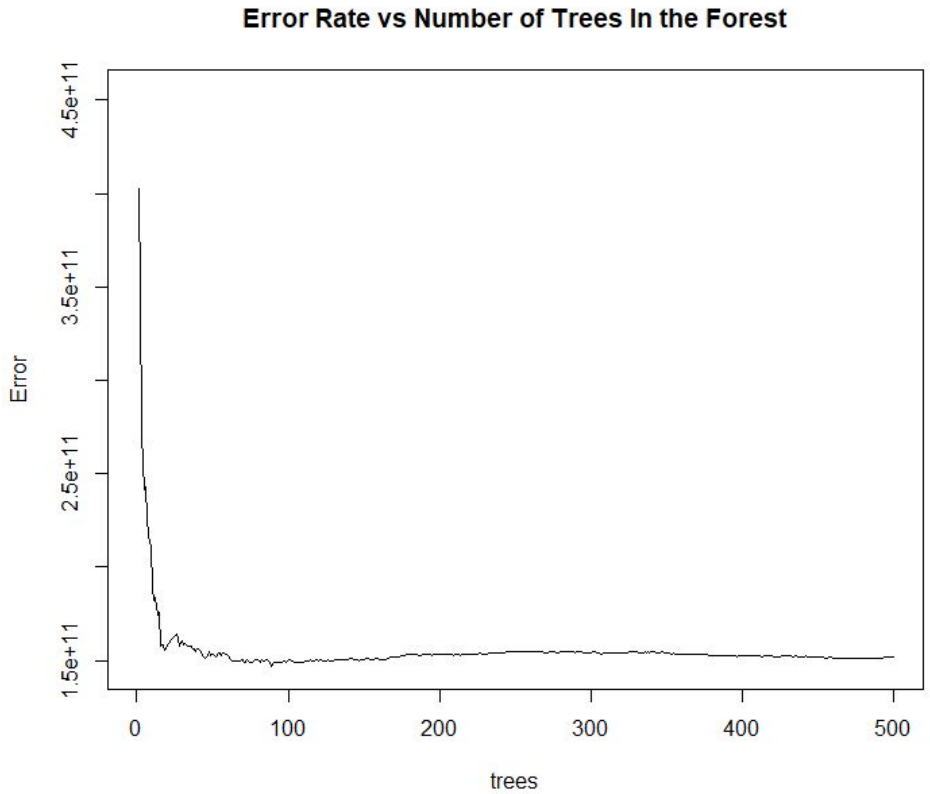
ADVANTAGES:
Easy to visualize | quickly get predicted home price with tree visual

SHORTCOMINGS:
Single house price for several house inputs | High MAPE and RMSE | Limited variables represented

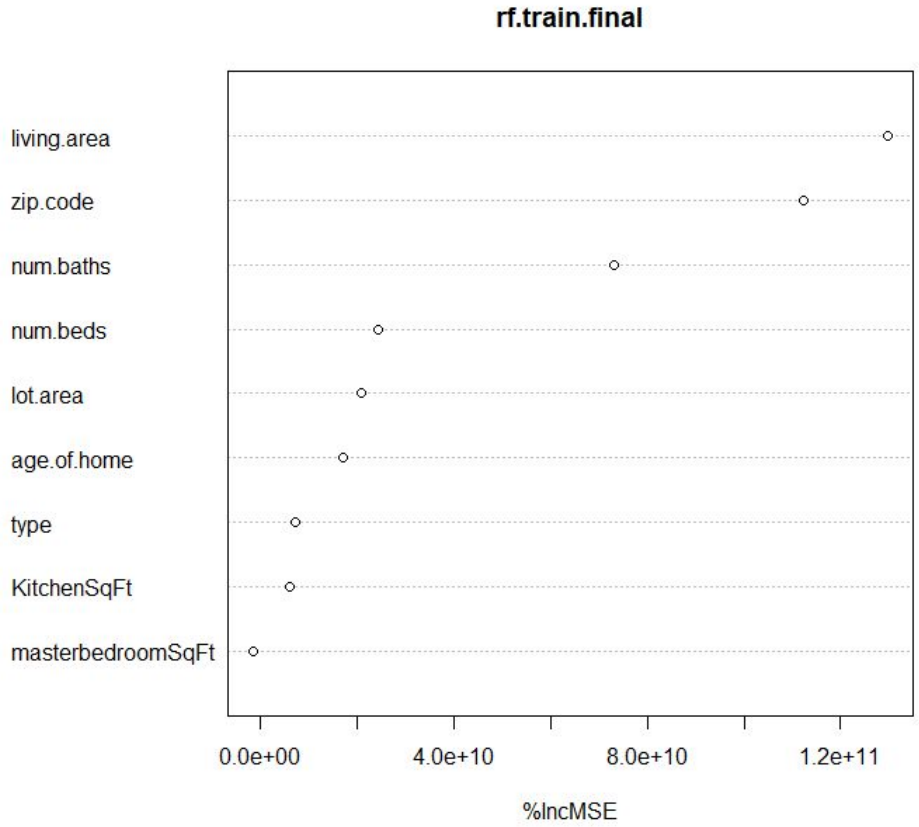


RF

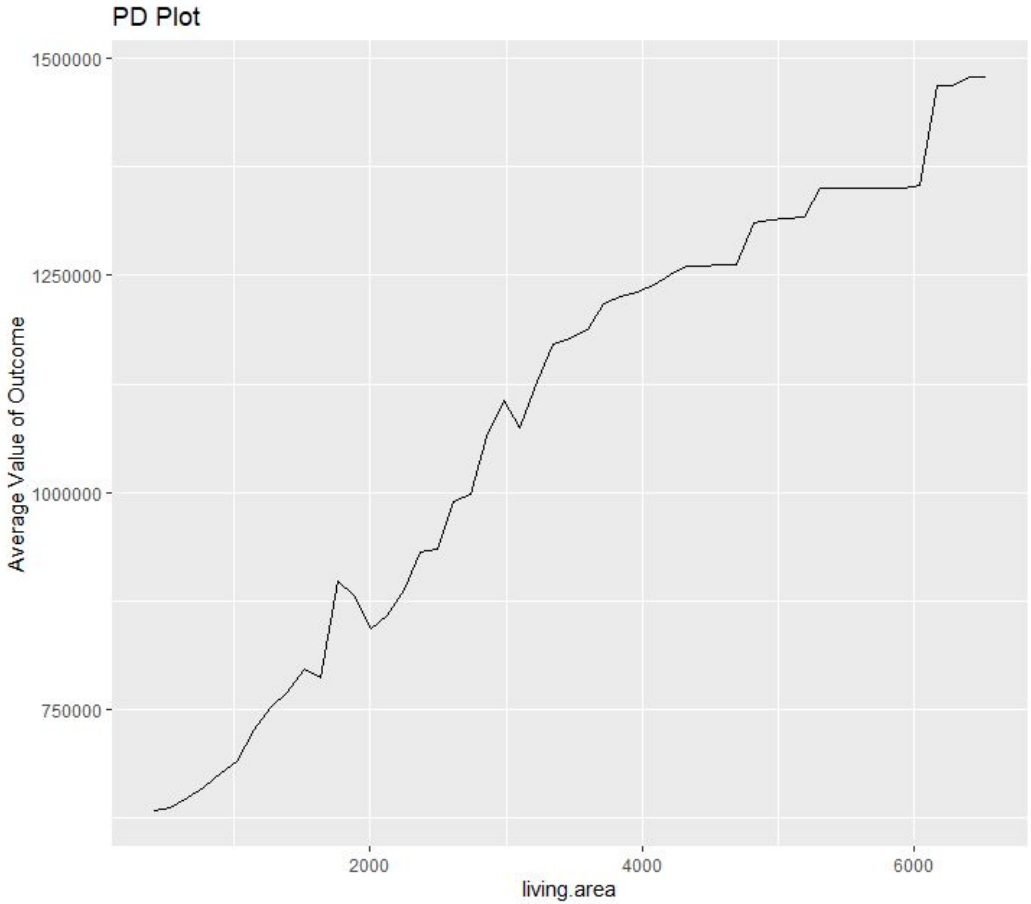
Random Forest



- Graph shows that the optimal number of trees is about 80-90, following which there is no significant reduction in error rate



- Most important variables are living area, zip code, and number of baths
- based on average change in accuracy of predicting OOB error, keeping all other predictors constant



- Example of a partial dependence (PD) plot
- dependence of the response variable on the living area predictor, holding constant all other predictors

Example House (#24) in Saugus (NE of Boston)

Zip Code: 01906
Type: Single Family
# of Beds: 3
# of Baths: 1.5
Living Area: 2556
Lot Area: 0.32
Kitchen Sq. Ft: 225
Master Bedroom Sq. Ft: 154
Age of Home: 59

Actual Sold Price: 510,000

Predicted Selling Price of Home: **671,999.3** *Error: 671,999.3- 510,000= +161,999.3*

Random Forest Model: Analysis

MAPE is 23.01
Mean absolute percentage error

RMSE is 241089.37
Square root of the variance of the residuals

ADVANTAGES:
More accurate than a single regression tree, randomly selects predictors, less overfitting than in CART

SHORTCOMINGS:
High MAPE and RMSE | Limited predictors (p/3) represented in each tree

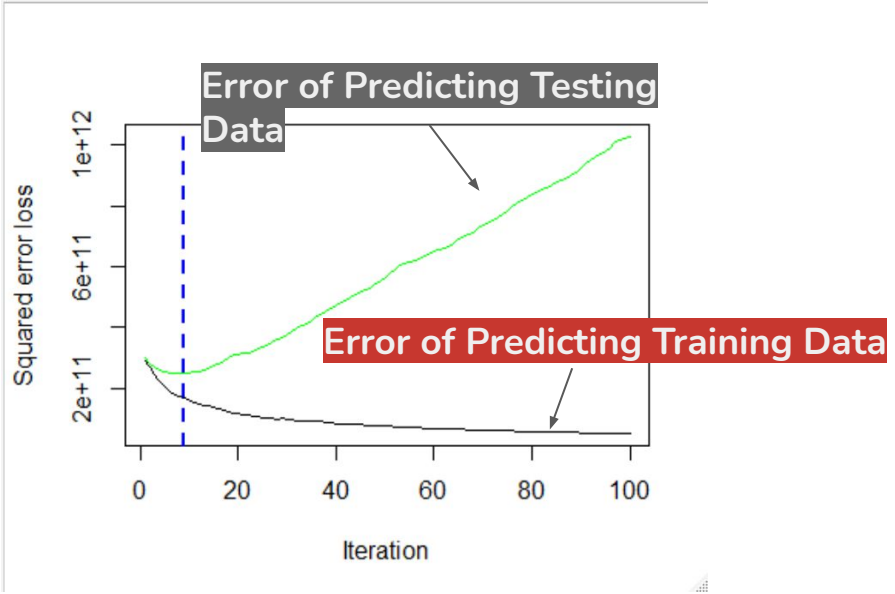
BOOSTING

Data Difference:
Keep 80 levels for Zipcode

BOOSTING Model: Analysis

Final Model:

- Recommended Number of Trees: 9
- Minimum Cross Validation Error: 246,907,111,140.441
- MAPE: 23.74%
- RMSE: 301225.34



Zip Code: 01730
Type: Single Family
# of Beds: 3
# of Baths: 2
Living Area: 2,000
Lot Area: 0.27
Kitchen Sq. Ft: 400
Master Bedroom Sq. Ft: 130
Age of Home: 59

Sold Price: 725,000

Prediction

686,890

Error: - 38,100

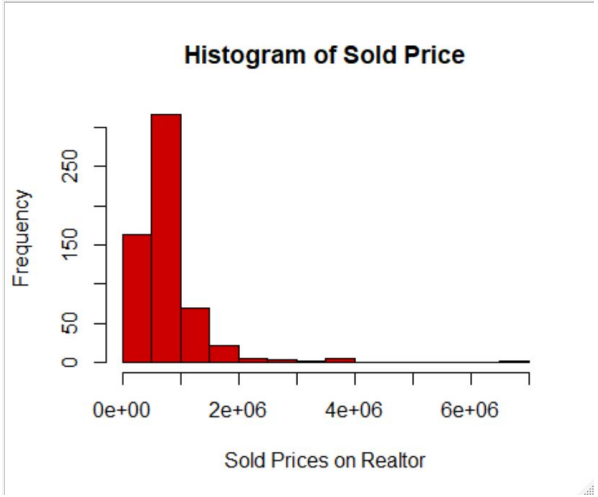
Predicted Selling Price of Home: **686,890** *Error: 686,890 - 725,000 = -38,100*

ANN

Data Difference:
Keep 80 levels in Zip Code

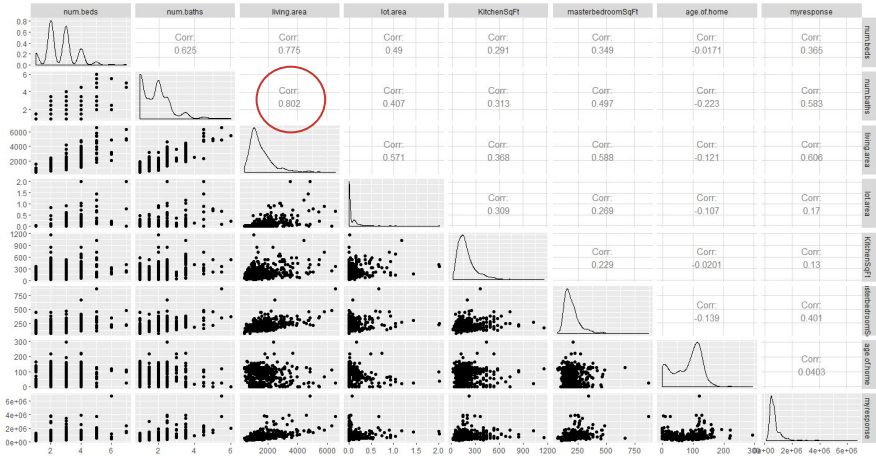
Right Skewness

Binomial Distribution



Highest Correlation → 0.802

Living Area *vs.* Number of Bathrooms



ANN Model: Analysis

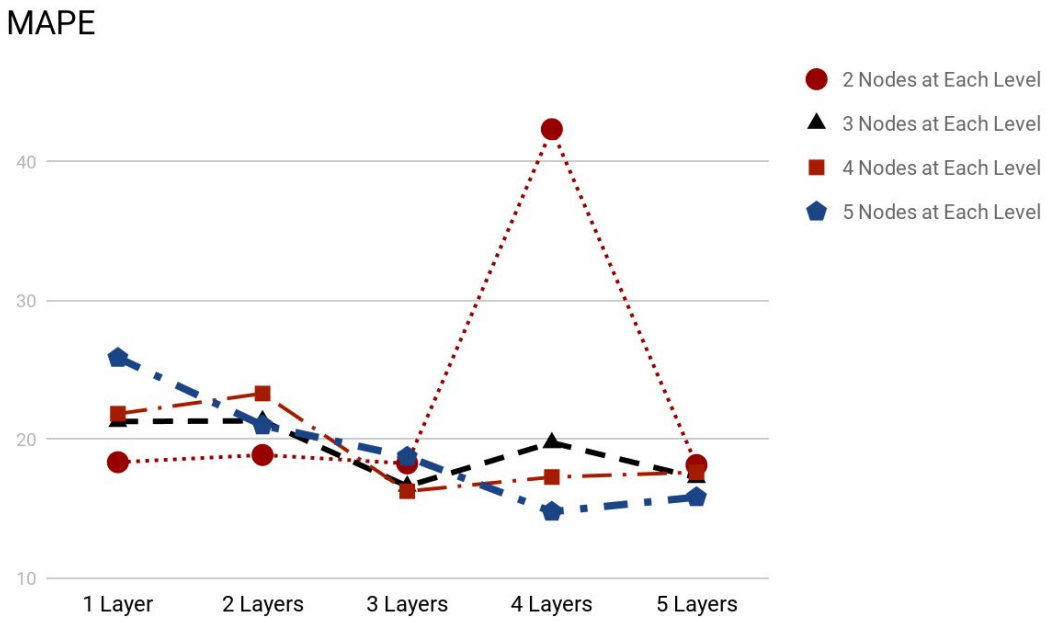
- **Negative Correlation:**

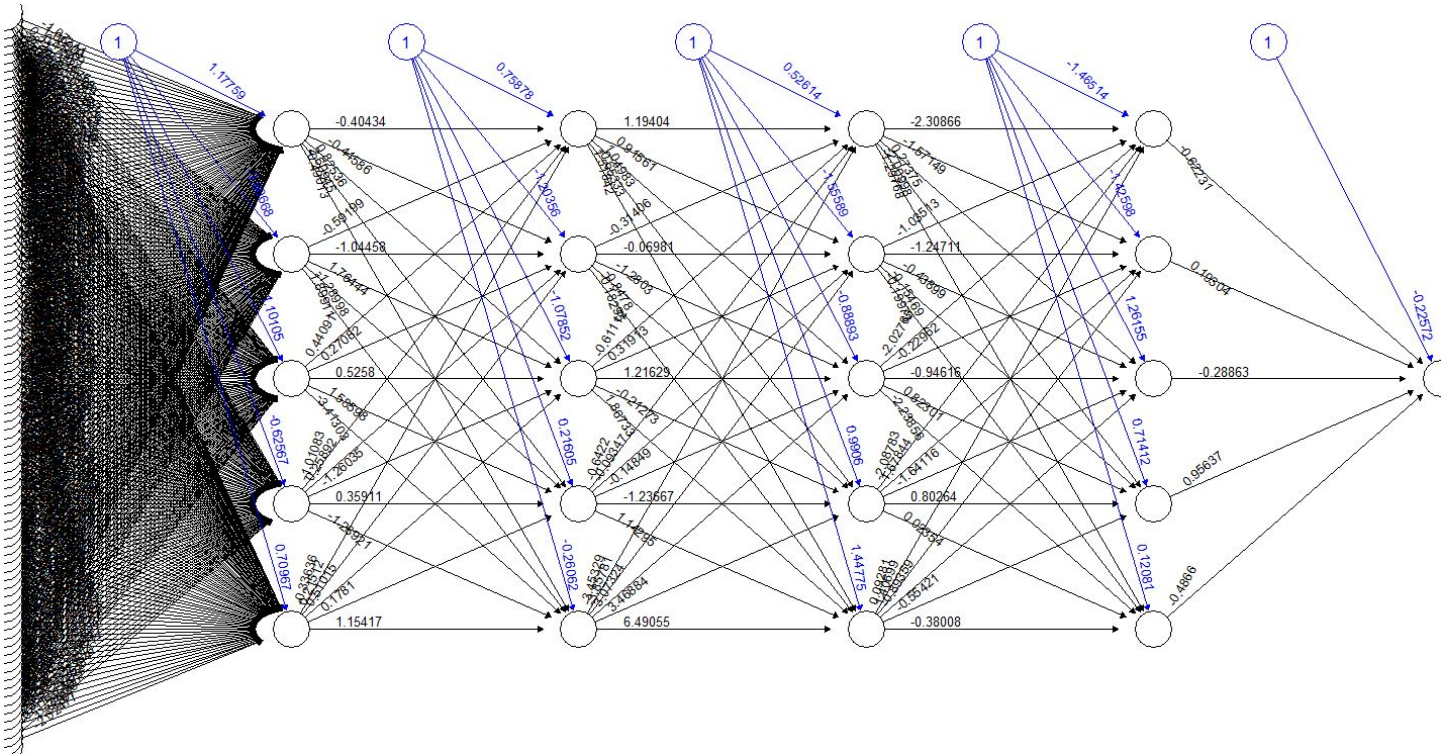
  MAPE vs. Number of Layers

- **Final Model**

  4 Layers with 5 Nodes Each

  - **Lowest MAPE: 14.79**
  - **Training SSE: 0.32**

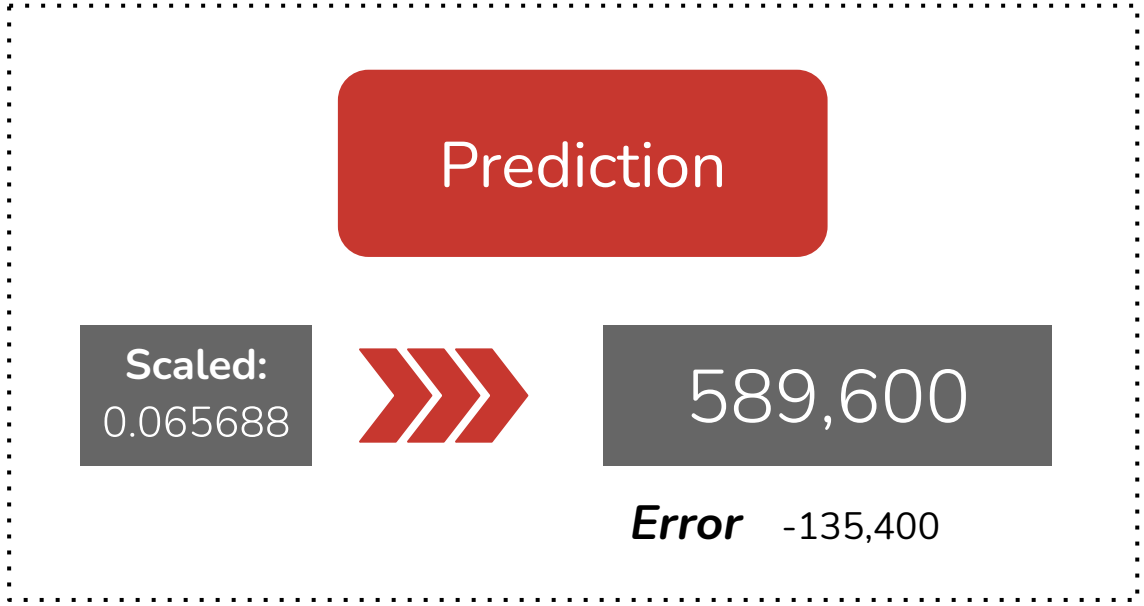



1
1.17759
0.75878
0.52614
-1.46514
-0.40434
1.19404
-2.30866
-0.92231
-1.04458
-0.06981
-1.24711
0.49804
0.5258
1.21629
-0.94616
-0.28863
0.35911
-1.23567
0.80264
0.95637
0.70967
1.15417
6.49055
-0.38008
0.4866
-0.22572

Example House

Zip Code: 01730
Type: Single Family
# of Beds: 3
# of Baths: 2
Living Area: 2,000
Lot Area: 0.27
Kitchen Sq. Ft: 400
Master Bedroom Sq. Ft: 130
Age of Home: 59

Actual Sold Price: 725,000



Predicted Selling Price of Home: **589,600** *Error: 589,600 - 725,000 = -135,400*

RESULTS

Comparing the 4 models

RESULTS: MAPE and RMSE

	MAPE	RMSE
CART	30.73%	325,443.68
RANDOM FOREST	23.24%	250,243.55
BOOSTING	23.74%	301,225.34
ANN	14.79%	N/A

Comparing the models

- CART, Boosting and Random Forest work better for classification
- ANN works more like a regression equation
  - Produced the best MAPE
- ANN exceeds all other models

Application

- Our model can give an estimate
- We would not use this model to accurately predict house prices
- Ex. $500,000 house with 14% error rate
    - Average error = $70,000

Moving Forward

- More information on income levels, school quality, crime rate etc.
- Topic modeling for overview
- More data in the dataset
- Get more accurate!

Thanks!

Any questions?

Team 2
Krishna Ammini, Erica Chin,
Alexandros Gounarakis,
Nicholas Lorello, Lingxiao Lyu

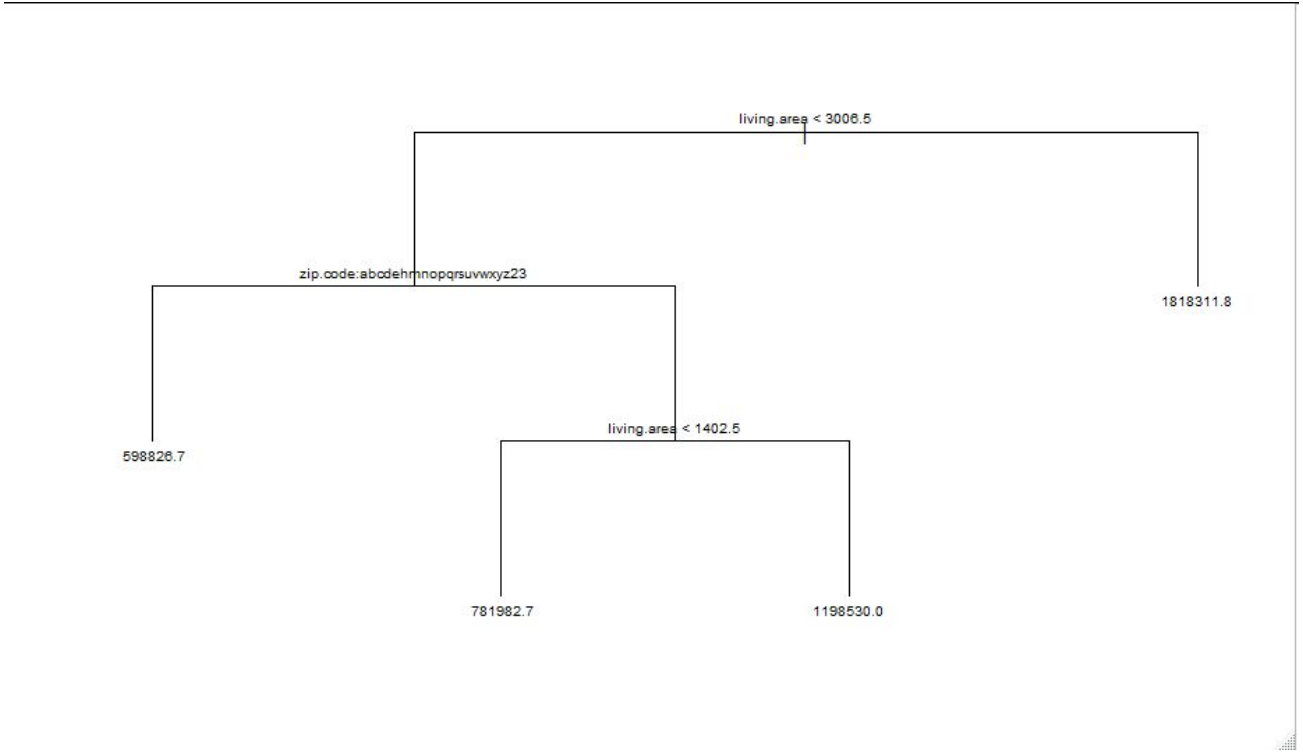
Appendix

Additional Information and outputs

CART R input

```
#Enter "R" for a regression tree and "C" for a classificaiton tree below.
tree_type="R"
#Enter the minimum number of items in each leaf
min_leaf_size=30
#Enter the minimum deviance for a node to be considered for a further split
min_deviance=0.025
```

CART Tree (R plot)



CART R output

```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 470 1.565490e+14  785724.7
   2) living.area < 3006.5 434 6.166351e+13  700072.3
     4) zip.code: 1730,1906,1945,2045,2052,2113,2119,2121,2122,2124,2125,2127,2128,2130,2131,2132,2134,2135,2136,2467,2476 326 1.276173e+13  598826.7 *
     5) zip.code: 2108,2111,2114,2115,2116,2118,2129,2215,2420,2492,other 108 3.547302e+13 1005684.0
      10) living.area < 1402.5 50 4.338590e+12  781982.7 *
      11) living.area > 1402.5 58 2.647532e+13 1198530.0 *
   3) living.area > 3006.5 36 5.331698e+13 1818311.8 *
>
>   deviance(bestcut)
[1] 9.689262e+13
```

MAPE for different ANN structures

	2 Nodes at Each Level	3 Nodes at Each Level	4 Nodes at Each Level	5 Nodes at Each Level
1 Layer	18.36	21.2928	21.8473	25.876
2 Layers	18.8821	21.3405	23.3082	21.0155
3 Layers	18.27	16.639	16.272	18.7578
4 Layers	42.3335	19.7643	17.2899	14.7973
5 Layers	18.1598	17.2585	17.6357	15.8453