Predicting Housing Sales Prices

Shengsheng Huo, Monica Orme, Robert Yav

Agenda

01 Introduction & Project Objectives

02 Data Cleaning

03 Model Selection

04 Results

05 Conclusion

06 Discussion

Introduction

- Currently, housing prices in general are becoming increasingly out of reach for more and more people.
- Our study is conducting investigation on housing prices of residential homes in Ames, Iowa using a dataset from Kaggle.com
- According to the National Association of Realtors, house prices will be expected to climb 5.7% through the end of 2022
- In March 2022, Ames home prices were up to 4.9% compared to last year, selling for a median price of $278K

United States
CANADA
WASHINGTON
OREGON
IDAHO
MONTANA
WYOMING
NORTH DAKOTA
SOUTH DAKOTA
MINNESOTA
WISCONSIN
MICHIGAN
NEBRASKA
IOWA
ILLINOIS
INDIANA
OHIO
PENNSYLVANIA
NEW YORK
MAINE
NEVADA
UTAH
COLORADO
KANSAS
MISSOURI
KENTUCKY
WEST VIRGINIA
VIRGINIA
CALIFORNIA
ARIZONA
NEW MEXICO
OKLAHOMA
ARKANSAS
TENNESSEE
NORTH CAROLINA
SOUTH CAROLINA
MISSISSIPPI
ALABAMA
GEORGIA
TEXAS
LOUISIANA
FLORIDA
HAWAII
ALASKA
MEXICO
THE BAHAMAS
CUBA

Ames, Iowa

Project Objectives

1. Predict the house sale prices of residentials homes in Ames, Iowa
2. Identifying the leading factors may influence the housing prices of residential homes in Ames, Iowa.
3. Our study will apply parametric model such as multiple linear regression and non parametric model such as Kmeans, decision tree and random forest regression.

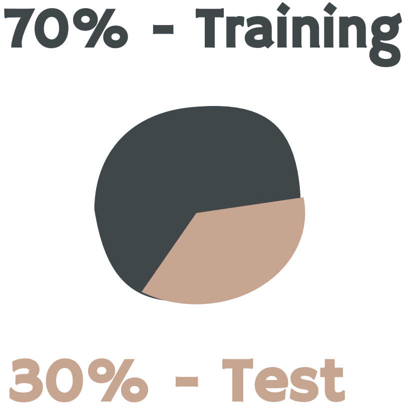
Data Cleaning

- Removing the observations(rows) containing >=20% of missing values
  - Matianing data distribution
- Removing the variables(columns) containing >=80% of missing values
  - Fence, Alley, Miscfeatures, PoolQC
- Deleting continuous variables which are strongly skewed or contained a number of outliers
  - BsmtFinSF2, ManvnrArea, and LotArea
- Removing extreme values (outliers) from Sale Price in the dataset
  - Predicting on average sale price of homes
- Applying multiple imputation for rest of missing values in the dataset
  - Maintain the distribution of the dataset
  - Kept 1386 observations and 43 columns

Model Selection Approach

Steps

1. Split the cleaned dataset into 70% training and 30% test
2. Build regression models using the training dataset
3. Use the MSE and other goodness of fit measures (if applicable) as criterion for deciding on the best models



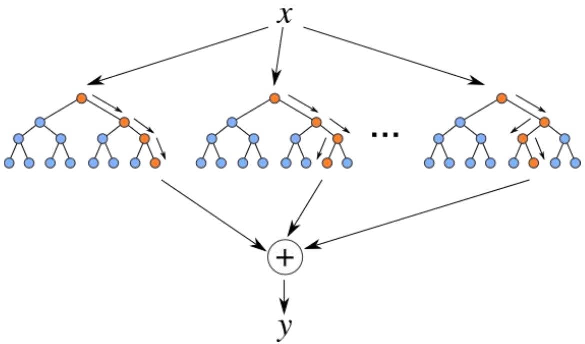
Model Selection

Model	Mean Square Error	Other Goodness of Fit Measures
Multiple Linear Regression	● Train: 373,589,470	Adjusted R^2 = 0.8175
Regression Decision Tree	● Test: 487,942,403,891	N/A
Kernel Regression	● Test: 3,479,005,530	N/A
Random Forest Regression	● Train: 93,262,941 ● Test: 486,558,431	85.07% – Variance Explained by the Model

Methodology

Random Forest Regression

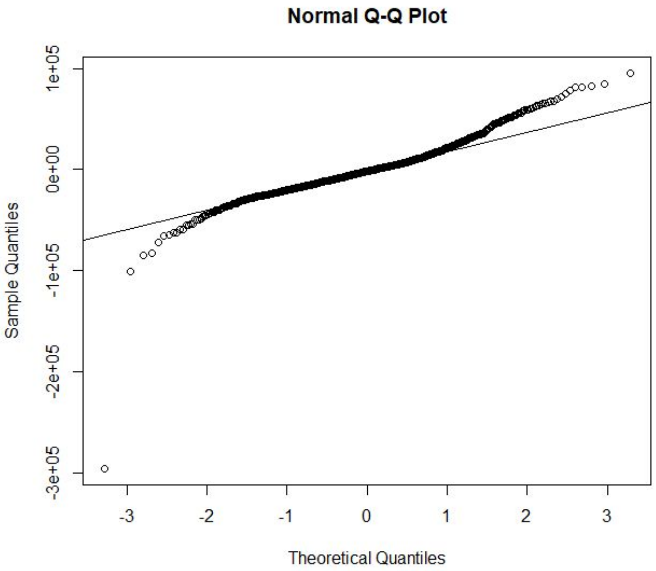
- Tune the value of m
- Non-parametric technique
- 43 predictors including continuous and categorical variables



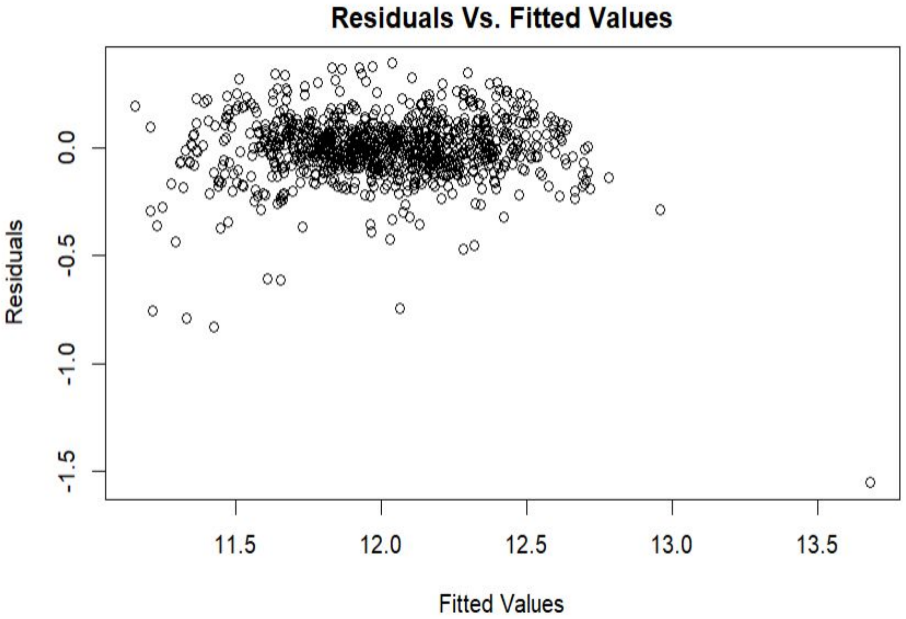
Multiple Linear Regression

- Backward Variable Selection
    - 13 continuous predictors
- Logarithmic Transformation of Sales Price
- Verification of Assumptions
    - Linearity
    - Errors have constant variance
    - Uncorrelated errors
    - Normality of the errors

Assumption Verification - Logarithmic MLR

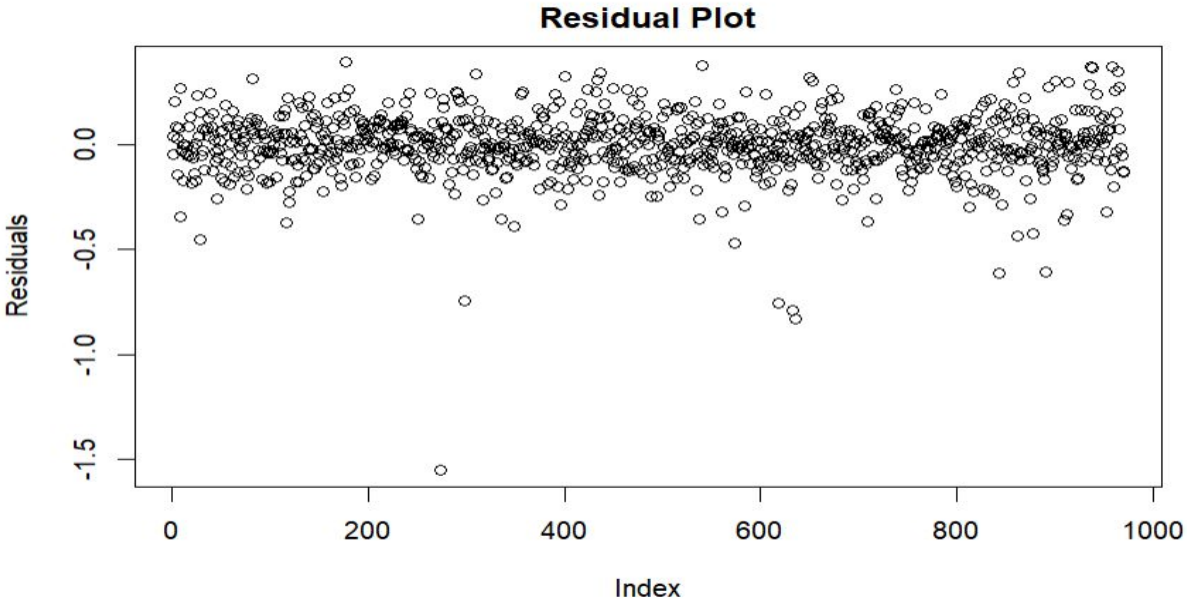


Normality Assumption



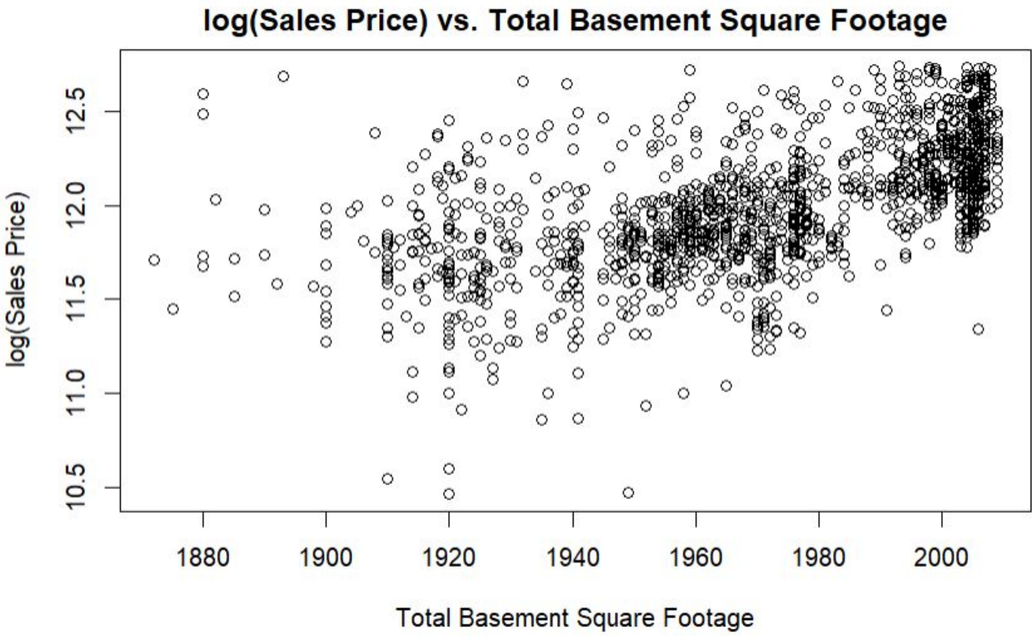
Constant Variance Assumption

Assumption Verification - Logarithmic MLR



Uncorrelated Error Assumption

Assumption Verification - Logarithmic MLR



Linearity

Results

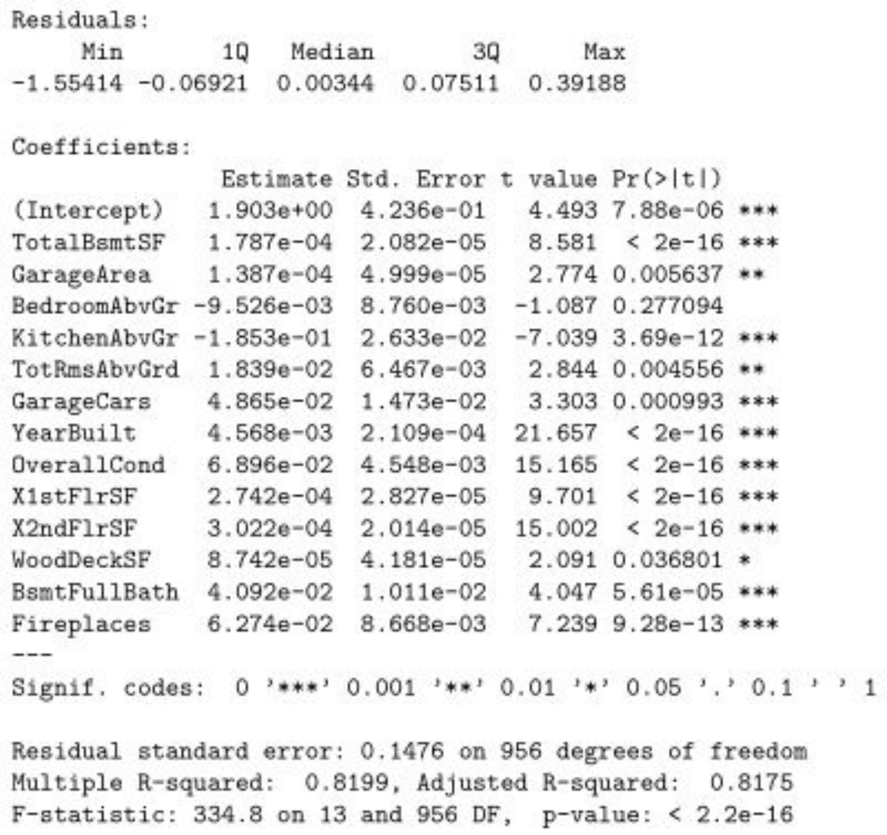
Model 1: Logarithmic Transformed Multiple Linear Regression Model

- $\log(y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8+\beta_9x_9+\beta_{10}x_{10}+\beta_{11}x_{11}+\beta_{12}x_{12}+\beta_{13}x_{13}$
- Adjusted R^2 – 0.8157
- Training MSE was 663,954,309

Model 2: Random Forest Regression Model

- 85.07% of variance explained by the model
- Test MSE was 486,558,431
- Training MSE was 93,262,941

Transformed Log MLR Model Output



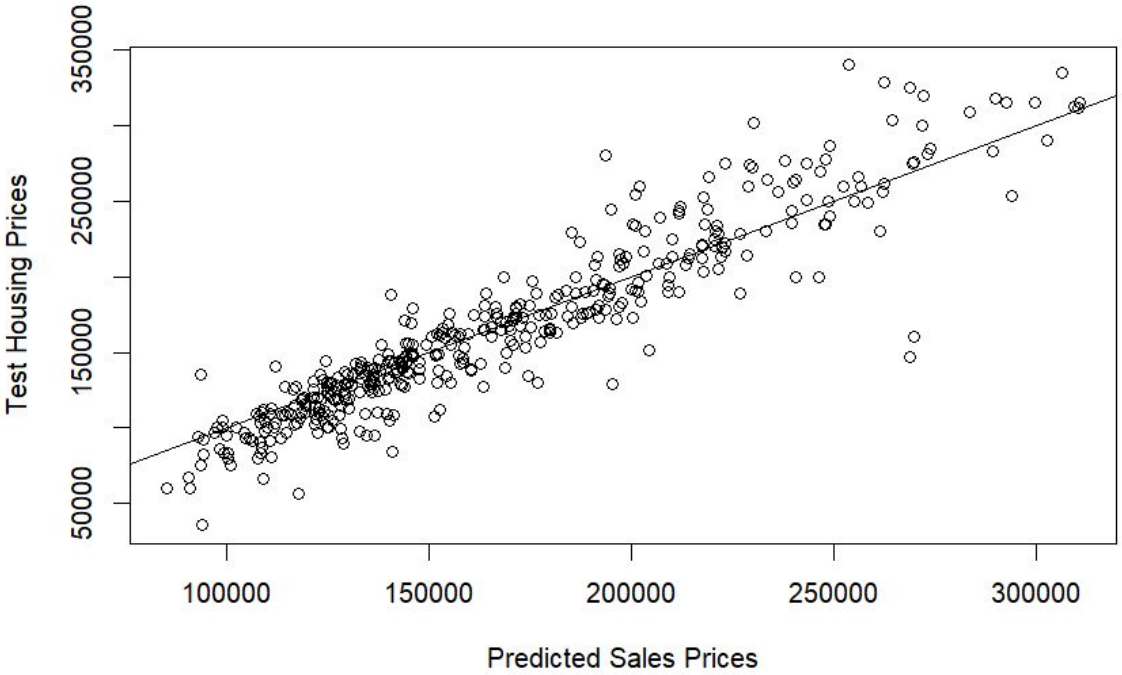
```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.55414 -0.06921  0.00344  0.07511  0.39188 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.903e+00  4.236e-01   4.493 7.88e-06 ***
TotalBsmtSF   1.787e-04  2.082e-05   8.581  < 2e-16 ***
GarageArea    1.387e-04  4.999e-05   2.774 0.005637 ** 
BedroomAbvGr -9.526e-03  8.760e-03  -1.087 0.277094    
KitchenAbvGr -1.853e-01  2.633e-02  -7.039 3.69e-12 ***
TotRmsAbvGrd  1.839e-02  6.467e-03   2.844 0.004556 ** 
GarageCars    4.865e-02  1.473e-02   3.303 0.000993 ***
YearBuilt     4.568e-03  2.109e-04  21.657  < 2e-16 ***
OverallCond   6.896e-02  4.548e-03  15.165  < 2e-16 ***
X1stFlrSF     2.742e-04  2.827e-05   9.701  < 2e-16 ***
X2ndFlrSF     3.022e-04  2.014e-05  15.002  < 2e-16 ***
WoodDeckSF    8.742e-05  4.181e-05   2.091 0.036801 *  
BsmtFullBath  4.092e-02  1.011e-02   4.047 5.61e-05 ***
Fireplaces    6.274e-02  8.668e-03   7.239 9.28e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1476 on 956 degrees of freedom
Multiple R-squared:  0.8199,	Adjusted R-squared:  0.8175 
F-statistic: 334.8 on 13 and 956 DF,  p-value: < 2.2e-16
```

Variable x_i	Variable Name	Beta Coefficient β_i
Intercept	Intercept	1.903e+00
x1	TotalBsmtSF: total basement square footage	3.365e+01
x2	GarageArea	3.311e+01
x3	BedroomAbvGr	-6.266e+03
x4	KitchenAbvGr	-3.653e+04
x5	TotRmsAbcGr	3.943e+03
x6	GarageCars	5.226e+03
x7	YearBuilt: the year that the home was built	6.477e+02
x8	OverallCond: the overall condition score of the home	8.759e+03
x9	X1stFlrSF: the square footage of the first floor	5.014e+01
x10	X2ndFlrSF: the square footage of the second floor	5.777e+01
x11	WoodDeckSF: wood deck square footage	2.086e+01
x12	BsmtFullBath: number of bathrooms in the basement	7.093e+03
x13	Fireplaces: the number of fireplaces	8.819e+03

Logarithmic Transformation MLR Interpretation

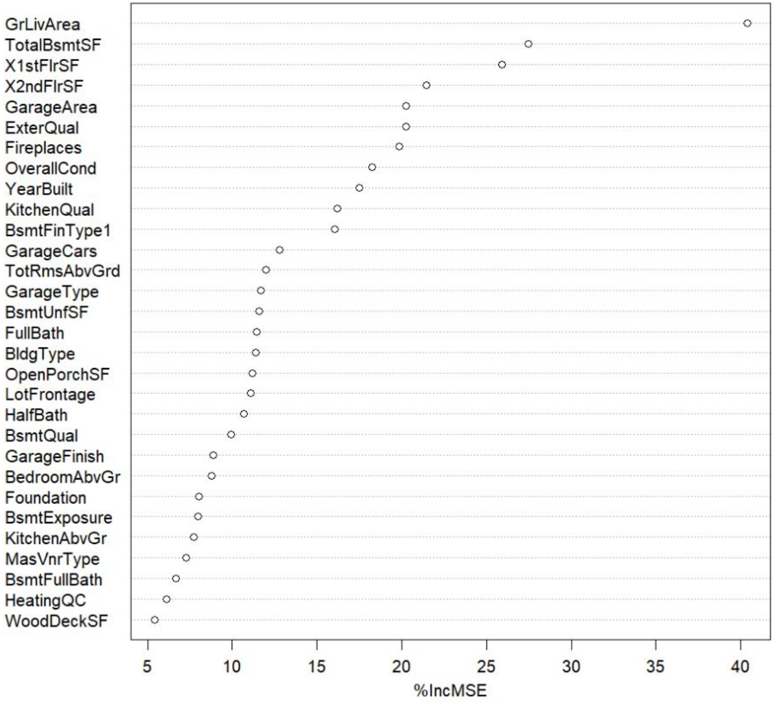
- For an increase in one unit for each predictor variable, a responsory change in beta is issued.

Results - Random Forest Regression



Predicted Sales Prices vs Test Sales Prices

Results - Random Forest Regression



Variable Importance Plot

Conclusion

Random Forest Regression is the ideal model to predict house sale prices of home

Model 1 Random Forest Regression

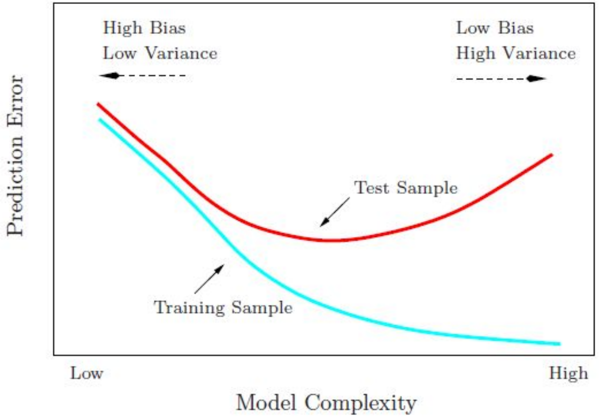
- Explained most of variance of the data – 85.07%
- Lower Training MSE: 93,262,941
- Lower Test MSE: 486,558,431

Model 2 Logarithmic transformed multiple linear regression

- Adjusted R square is high, explain most of variance of the data – 81.75%
- Violated the constant variance assumption
- Violated linearity assumption
- Violated uncorrelated variance assumption
- Higher Training MSE: 663,954,309

Discussion

- Model Selection – important to consider the train and test error for every model
- Transformation of the response and linearity in Multiple Linear Regression
- The random forest regression model is difficult to interpret
- Consider exploratory data analysis such as Principal Component Analysis
- Scaling the continuous data



Dataset

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

CREDITS: This presentation template was created by **Slidesgo,** including icons by **Flaticon,** infographics & images by **Freepik**