Developing Machine Learning Models To Predict House Prices

Critical Thinking Group 5

Bivek **A**dhikari ◆ Junlan **Z**hou ◆ **M**organ **K**isselburg ◆ **T**yler **V**iolillo

Northwestern University

Predict 422 – Practical Machine Learning

Winter 2017

Outline

- Outline
- Assignment
- Introduction
- Resources
- Modeling Process
- Model Development Plan
- Data Preparation
- Exploratory Data Analysis & Transformation
- Supervised Learning

Assignment

1) Select a real-world data set of interest (from national database, from work, from Kaggle, etc.).

2) Conduct exploratory data analysis.

3) Perform supervised and/or unsupervised learning.

4) Give a presentation during the Thursday sync session and submit the slides in PDF format on Canvas.

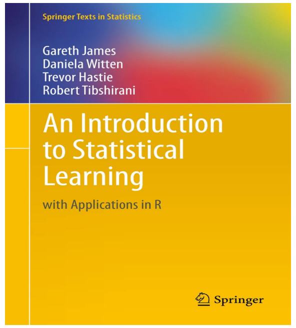
Each group will give a 10-15 minute presentation to the class at the sync meeting on Thursday, March 2, 2017.

Introduction

- Exploration of developing supervised learning predictive models in R
- Will utilize structured data
- Data source is www.kaggle.com
- Data set is titled **House Sales in King County, USA**
- Objective is to predict house prices
- Identify the value of predicting house prices

Resources

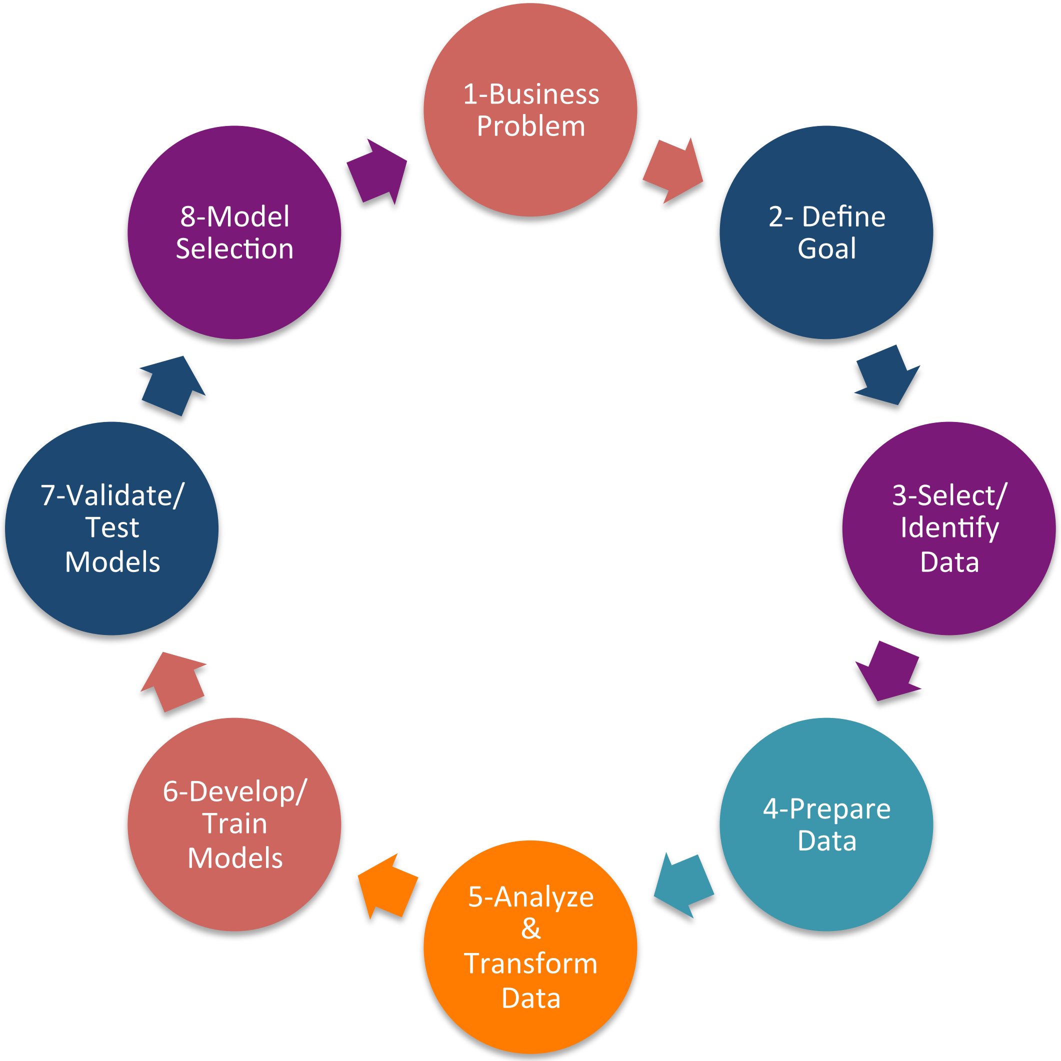
- An Introduction to Statistical Learning, with Applications in R



- R Studio

- Kaggle

Modeling Process



Model Development Plan

Plan:

- Analyze the effects of 19 predictors (bedrooms, bathrooms, sqft_living, sqft-lot etc.) on the house price (the response variable) by building following predictive models:
- OLS regression model
- Comparison of OLS regression model and shrinkage method (if p>n) – Lasso in the same dataset
- GAM model to address the limitations of non-linear data
- PCA model
- Rpart model

Limitations

- The machine learning models used are not exhaustive.
- Equations excluded due to assumption that class has subject matter expertise of calculations.

Background

21613 observations

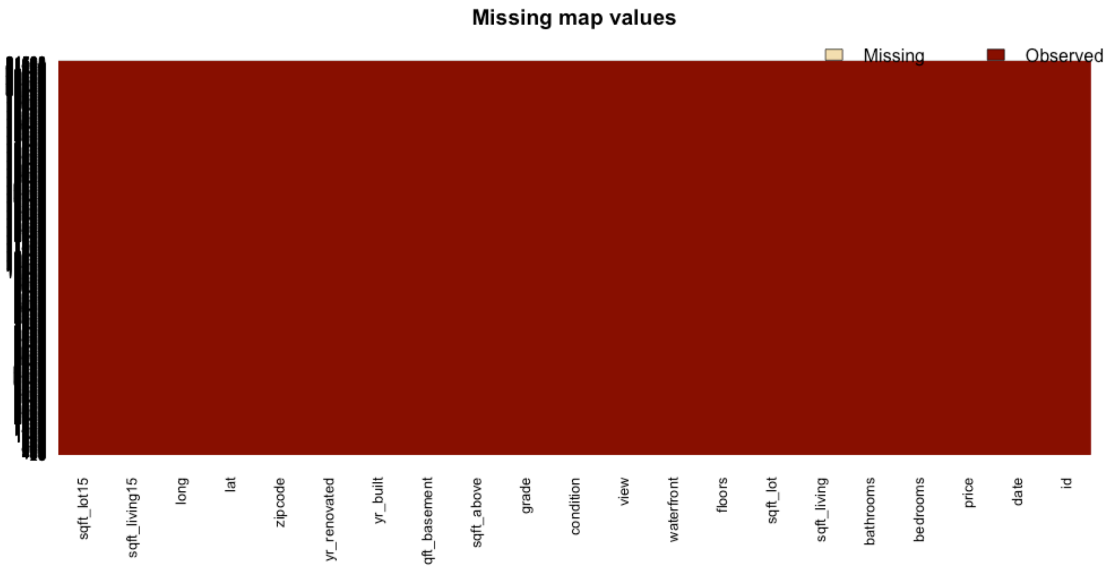
21 variables

```
'data.frame':   21613 obs. of  21 variables:
 $ id           : num  7.13e+09 6.41e+09 5.63e+09 2.49e+09 1.95e+09 ...
 $ date         : Factor w/ 372 levels "20140502T000000",..: 165 221 291 221 284 11 57 252 340 306 ...
 $ price        : num  221900 538000 180000 604000 510000 ...
 $ bedrooms     : int  3 3 2 4 3 4 3 3 3 3 ...
 $ bathrooms    : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
 $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
 $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
 $ floors       : num  1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ view         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ condition    : int  3 3 3 5 3 3 3 3 3 3 ...
 $ grade        : int  7 7 6 7 8 11 7 7 7 7 ...
 $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
 $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
 $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
 $ zipcode      : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
 $ lat          : num  47.5 47.7 47.7 47.5 47.6 ...
 $ long         : num  -122 -122 -122 -122 -122 ...
 $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
 $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
```

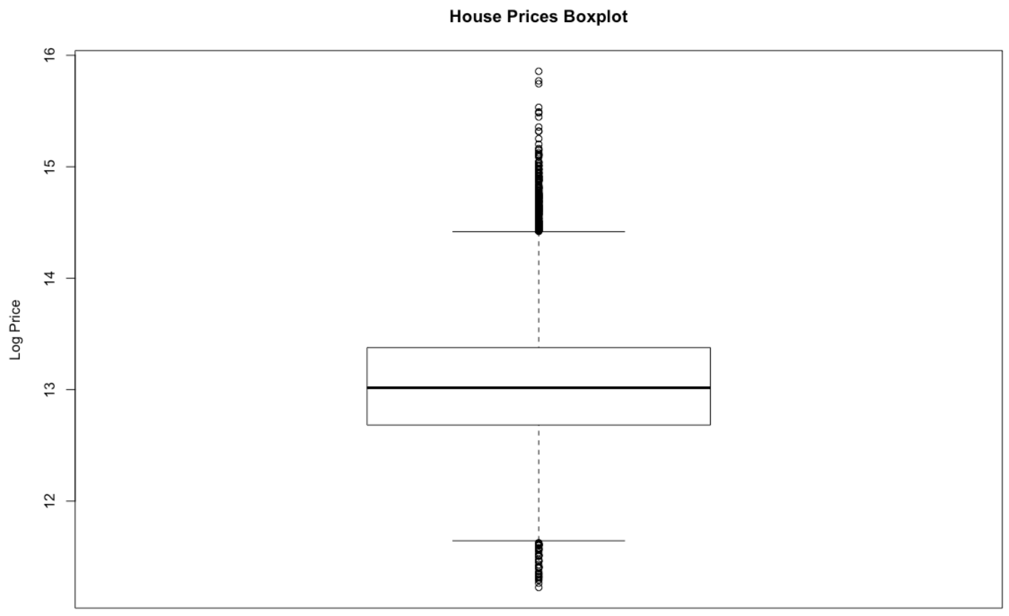
Exploratory Data Analysis

Checking missing values:

```
> dim(House2)
[1] 21613    19
> sum(is.na(House2))
[1] 0
> summary(House2)
     price            bedrooms         bathrooms      sqft_living       sqft_lot      
 Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290   Min.   :    520  
 1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040  
 Median : 450000   Median : 3.000   Median :2.250   Median : 1910   Median :   7618  
 Mean   : 540088   Mean   : 3.371   Mean   :2.115   Mean   : 2080   Mean   :  15107  
 3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688  
 Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :13540   Max.   :1651359  
     floors        waterfront            view          condition          grade       
 Min.   :1.000   Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000  
 1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000  
 Median :1.500   Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000  
 Mean   :1.494   Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657  
 3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000  
 Max.   :3.500   Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000  
   sqft_above    sqft_basement       yr_built     yr_renovated       zipcode     
 Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0   Min.   :98001  
 1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0   1st Qu.:98033  
 Median :1560   Median :   0.0   Median :1975   Median :   0.0   Median :98065  
 Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4   Mean   :98078  
 3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0   3rd Qu.:98118  
 Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0   Max.   :98199  
      lat             long         sqft_living15    sqft_lot15    
 Min.   :47.16   Min.   :-122.5   Min.   : 399   Min.   :   651  
 1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490   1st Qu.:  5100  
 Median :47.57   Median :-122.2   Median :1840   Median :  7620  
 Mean   :47.56   Mean   :-122.2   Mean   :1987   Mean   : 12768  
 3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360   3rd Qu.: 10083  
 Max.   :47.78   Max.   :-121.3   Max.   :6210   Max.   :871200  
```

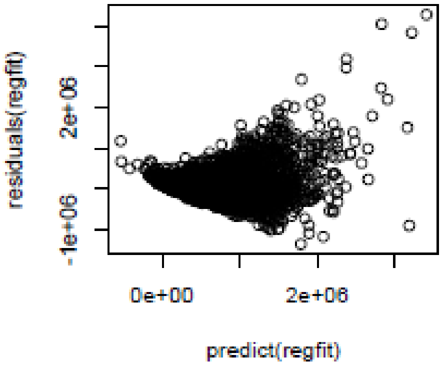


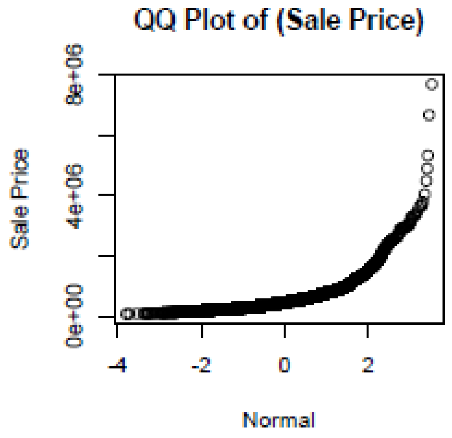
Exploratory Data Analysis



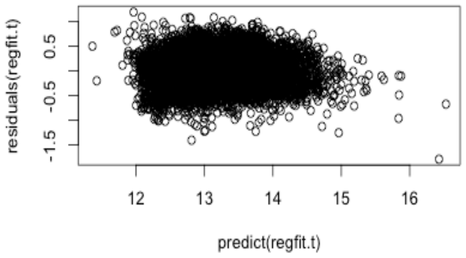
Exploratory Data Analysis

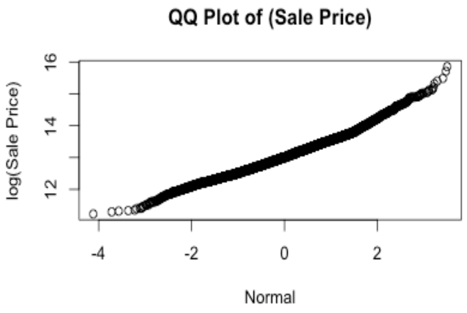
Before(transformation)

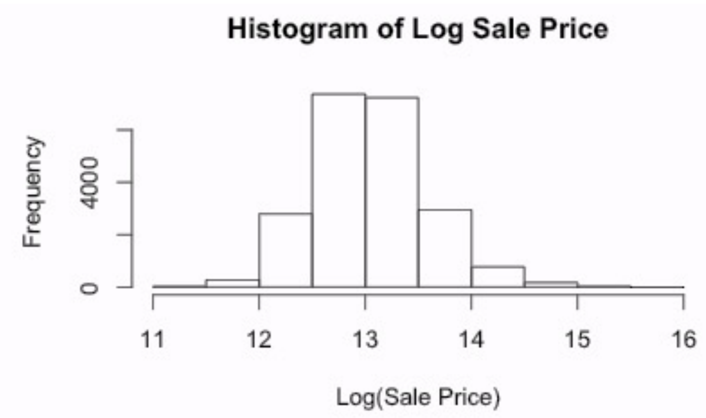




After(transformation)







Note: The residual plot indicates that there are non-linear associations in the data

Exploratory Data Analysis

Forward Selection

```
Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.411e+00  4.220e+00  -0.571 0.567864
bedrooms     -1.367e-02  2.704e-03  -5.054 4.37e-07 ***
bathrooms     6.667e-02  4.748e-03  14.040  < 2e-16 ***
sqft_living   1.578e-04  6.372e-06  24.756  < 2e-16 ***
sqft_lot      6.186e-07  7.524e-08   8.221  < 2e-16 ***
floors        7.642e-02  5.196e-03  14.707  < 2e-16 ***
waterfront    3.612e-01  2.553e-02  14.148  < 2e-16 ***
view          6.148e-02  3.134e-03  19.620  < 2e-16 ***
condition     6.223e-02  3.390e-03  18.357  < 2e-16 ***
grade         1.570e-01  3.116e-03  50.401  < 2e-16 ***
sqft_above   -1.724e-05  6.301e-06  -2.736 0.006217 **
sqft_basement        NA         NA      NA       NA
yr_built     -3.394e-03  1.052e-04 -32.266  < 2e-16 ***
yr_renovated  3.900e-05  5.330e-06   7.316 2.67e-13 ***
zipcode      -6.847e-04  4.755e-05 -14.400  < 2e-16 ***
lat           1.395e+00  1.554e-02  89.722  < 2e-16 ***
long         -1.706e-01  1.881e-02  -9.069  < 2e-16 ***
sqft_living15 9.602e-05  5.011e-06  19.164  < 2e-16 ***
sqft_lot15   -4.071e-07  1.105e-07  -3.685 0.000229 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2519 on 16192 degrees of freedom
Multiple R-squared:  0.7693,    Adjusted R-squared:  0.769
F-statistic:  3175 on 17 and 16192 DF,  p-value: < 2.2e-16
```

```
> CV(mylm)
           CV           AIC          AICc           BIC
 6.356497e-02 -4.468222e+04 -4.468218e+04 -4.453605e+04
        AdjR2
 7.690204e-01
```

Backward Selection

```
Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.411e+00  4.220e+00  -0.571 0.567864
bedrooms     -1.367e-02  2.704e-03  -5.054 4.37e-07 ***
bathrooms     6.667e-02  4.748e-03  14.040  < 2e-16 ***
sqft_living   1.578e-04  6.372e-06  24.756  < 2e-16 ***
sqft_lot      6.186e-07  7.524e-08   8.221  < 2e-16 ***
floors        7.642e-02  5.196e-03  14.707  < 2e-16 ***
waterfront    3.612e-01  2.553e-02  14.148  < 2e-16 ***
view          6.148e-02  3.134e-03  19.620  < 2e-16 ***
condition     6.223e-02  3.390e-03  18.357  < 2e-16 ***
grade         1.570e-01  3.116e-03  50.401  < 2e-16 ***
sqft_above   -1.724e-05  6.301e-06  -2.736 0.006217 **
sqft_basement        NA         NA      NA       NA
yr_built     -3.394e-03  1.052e-04 -32.266  < 2e-16 ***
yr_renovated  3.900e-05  5.330e-06   7.316 2.67e-13 ***
zipcode      -6.847e-04  4.755e-05 -14.400  < 2e-16 ***
lat           1.395e+00  1.554e-02  89.722  < 2e-16 ***
long         -1.706e-01  1.881e-02  -9.069  < 2e-16 ***
sqft_living15 9.602e-05  5.011e-06  19.164  < 2e-16 ***
sqft_lot15   -4.071e-07  1.105e-07  -3.685 0.000229 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2519 on 16192 degrees of freedom
Multiple R-squared:  0.7693,    Adjusted R-squared:  0.769
F-statistic:  3175 on 17 and 16192 DF,  p-value: < 2.2e-16
```

```
> CV(mylm2)
           CV           AIC          AICc           BIC         AdjR2
 6.356497e-02 -4.468222e+04 -4.468218e+04 -4.453605e+04  7.690204e-01
```

Methodology

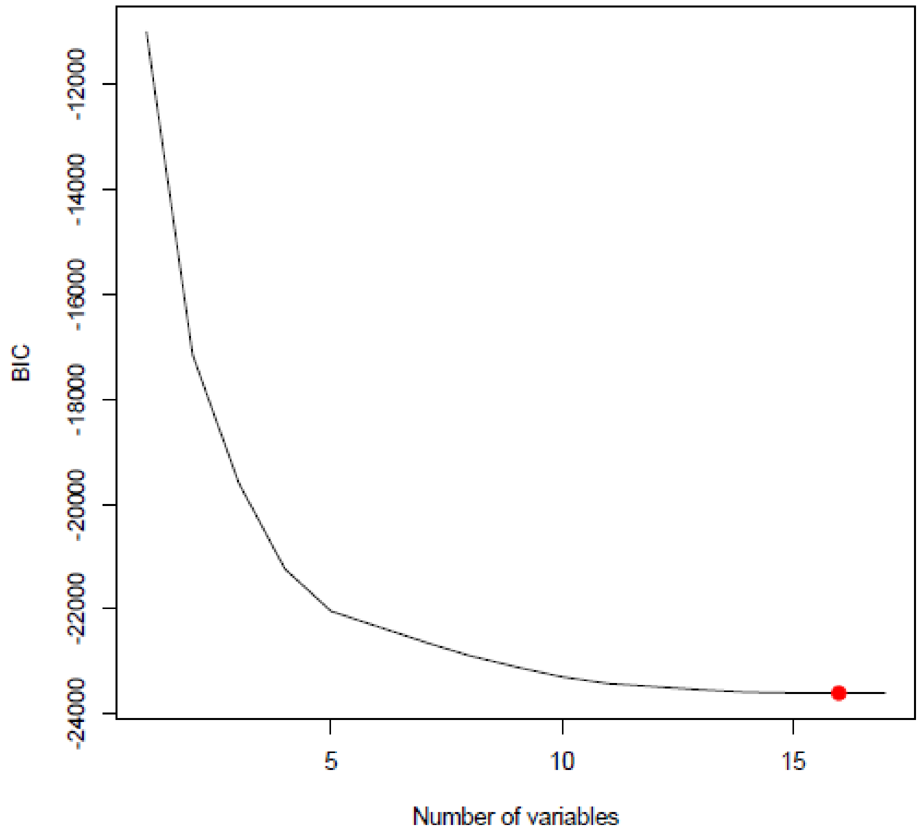
- Split the observations into a training data set (75%) and a test data set (25%).
- Choose the best least squares regression model with the lowest BIC and lowest test error by using 10-fold cross validation approaches.
- Used 10-fold cross validation with the largest value of tuning parameter (λ) for Lasso.
- Tested different components for PCA model, and nodes for Rpart model.

Results

Model #1:
Least squares regression model

16 Predictors with Lowest BIC of -23599.28

MSE in test set is 0.06903407



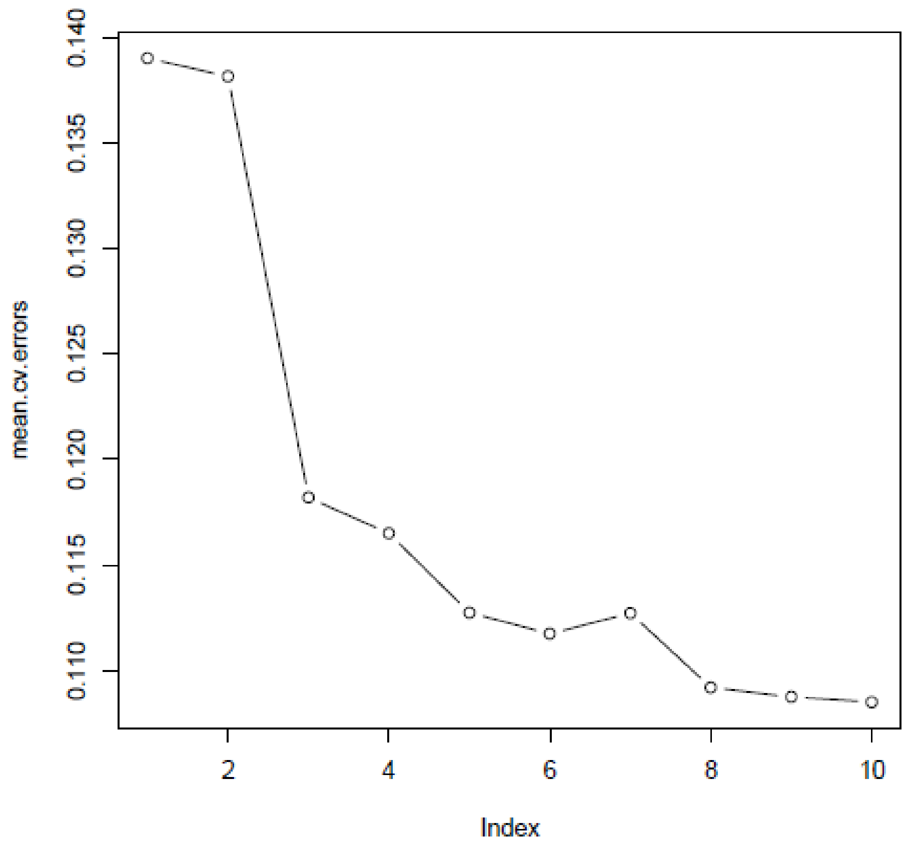
Best subset selection using BIC:
The best model with 16 predictors was chosen because it had the lowest value of BIC

Results

Model #1 Continued…
Least squares regression model

Lowest value of Mean CV Error is 0.1085001

MSE in test set is 0.1123473



Model selection by 10-fold cross validation:
The best model with 10 predictors was chosen because it had the lowest value of mean.cv.error.

Results Analysis

Model #1: Least squares regression model Analysis

- The formula is intuitive for the most part because good performance by the factors (bathroom, sqft-living, sqft-lot etc.) are all rewarded with a positive coefficient indicating that the results would be associated with a higher price.

- Likewise, yr_built and long have negative coefficients indicating that they would be associated with a lower price.

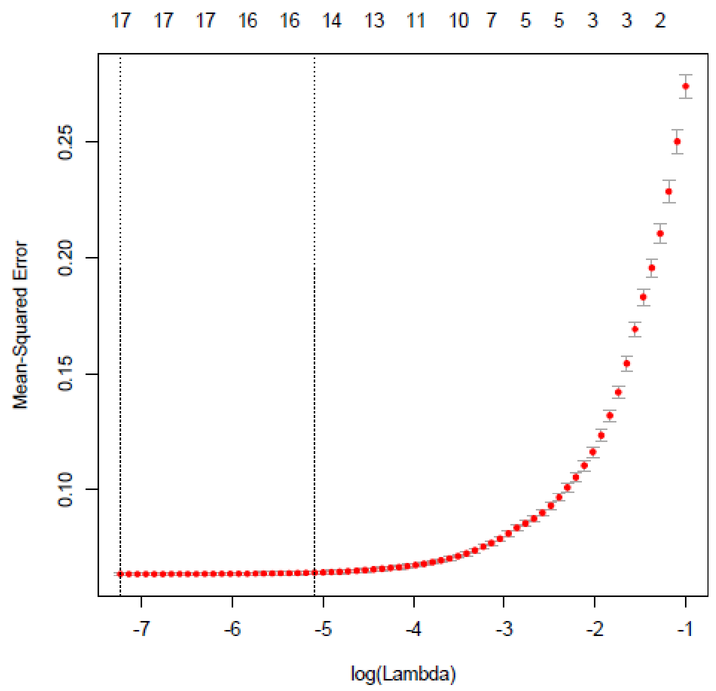
- In this model is not free from 'sign' issues -- bedrooms, sqft_above and sqft_lot15 seemingly have good effects on the price of the houses, that would eventually result in a lower price. This issue needs further investigation.

Results

Model #2:
Lasso model : Model selection by 10-fold cross validation (the largest λ)

The largest value of λ is 0.006144808

MSE in test set is 0.06579507



The best model with 14 predictors was chosen because it had the largest value of λ (0.006144808)

Results Analysis

Model #2: Lasso model analysis

- The factors such as: bathroom, sqft-living, sqft-lot etc. are all rewarded with a positive coefficient indicating that the results would be associated with higher prices.

- Likewise, yr_built and long have negative coefficients. Indicating that they would be associated with lower prices.

- There are no 'sign' issues in this model; the reason is that lasso model shrinks the estimated coefficients toward zero relative to the OLS estimates.

Results

Model #3: GAM model
The best model with 17 predictors was chosen (ANOVA)

MSE in test set is 0.06474506

```
> gam.house$coef
  (Intercept)      bedrooms     bathrooms   sqft_living      sqft_lot        floors
waterfront          view      condition
-2.410679e+00 -1.366792e-02  6.666953e-02  1.577548e-04  6.186054e-07  7.641639e-02
3.612074e-01  6.148077e-02  6.222554e-02
        grade    sqft_above sqft_basement      yr_built  yr_renovated       zipcode
lat           long sqft_living15
 1.570469e-01 -1.724299e-05            NA -3.394271e-03  3.899675e-05 -6.847221e-04
1.394565e+00 -1.705634e-01  9.602424e-05
   sqft_lot15
-4.071429e-07
```

Coefficients of the GAM model

Results

Model #4: PCA model

Lowest MSE in test set is 0.06337 with 17 and 18 components

```
> MSEP(pcamodel)
(Intercept)      1 comps      2 comps      3 comps      4 comps
    0.27463      0.27158      0.27158      0.14160      0.13542
    5 comps      6 comps      7 comps      8 comps      9 comps
    0.13534      0.13343      0.12897      0.12542      0.11740
   10 comps     11 comps     12 comps     13 comps     14 comps
    0.10759      0.09697      0.09675      0.09546      0.09517
   15 comps     16 comps     17 comps     18 comps
    0.06622      0.06416      0.06337      0.06337
```

MSE of different components in PCA model

Results

Model #5: Rpart model

Lowest MSE in test set is 0.05106359 in Node number 22

```
Node number 1: 16210 observations,
  mean=13.05015, MSE=0.2746303

Node number 2: 13024 observations,
  mean=12.90178, MSE=0.1787177

Node number 3: 3186 observations,
  mean=13.65668, MSE=0.208835

Node number 4: 5424 observations,
  mean=12.60436, MSE=0.1150723

Node number 5: 7600 observations,
  mean=13.11404, MSE=0.1159545

Node number 6: 2454 observations,
  mean=13.53138, MSE=0.1391945

Node number 7: 732 observations
  mean=14.07674, MSE=0.213217

Node number 8: 3377 observations
  mean=12.46247, MSE=0.08511315

Node number 9: 2047 observations
  mean=12.83843, MSE=0.07649384

Node number 10: 3502 observations
  mean=12.93061, MSE=0.08762608

Node number 11: 4098 observations,
  mean=13.27079, MSE=0.08683814

Node number 12: 556 observations
  mean=13.15002, MSE=0.08266812

Node number 13: 1898 observations
  mean=13.6431, MSE=0.1006683

Node number 22: 1024 observations
  mean=13.06853, MSE=0.05106359

Node number 23: 3074 observations
  mean=13.33817, MSE=0.08058761
```

MSE of different Node numbers in Rpart model

Model Performance

Models	MSE in test set
Best subset selection using BIC	0.06903407
Least squares regression model by 10-fold cross validation	0.1123473
Lasso model by 10-fold cross validation (the largest λ)	0.06579507
GAM Model (ANOVA)	0.06474506
PCA	0.06337
Rpart	0.05106359

With the lowest MSE of **0.05106359**, Rpart model is our best model.


Data
Collaborative
Medical
Risk Assessment
ProviderReimbursement
DiseaseManagement
QualityImprovement
PrimaryCareDelivery
Predictive Modeling
Preventive
CaseManagement
Analytics
Needs Assessment
Coordinated
Healthcare
Integrated
Comprehensive
Partnership


USED
REGRESSION
VARIABLES
MODELING
DEPENDENT
PREDICTIVE
ANALYTICS
CLASSIFICATION
ANALYSIS
VARIABLE
SERIES
DATA
MODELS
TIME
TECHNIQUES
DISTRIBUTION
DECISION
MODEL
CUSTOMER
LOGISTIC
LOGIT
NETWORKS

Thank You!


CLUSTERING
STRUCTURE
INFORMATION
ANALYSIS
WAREHOUSING
CLASSIFICATION
ANOMALY
RECORDS
DATA
QUANTITY
PROCESS
DETECTION
LARGE
MINING
MODELING
ASPECT
DATABASE
SET
SUMMARIZATION
UNDERSTANDABLE
COLLECTION
PATTERNS
REGRESSION
STATISTICS
MANAGEMENT
EXTRACTION


SOFTWARE
LARGE
STORAGE
BIG
INFORMATION
DATA
SETS
ANALYSIS
RESEARCH
SEARCH
PETABYTES
MANAGEMENT
VOLUME
TECHNOLOGIES
TECHNOLOGY
DEVELOPMENT