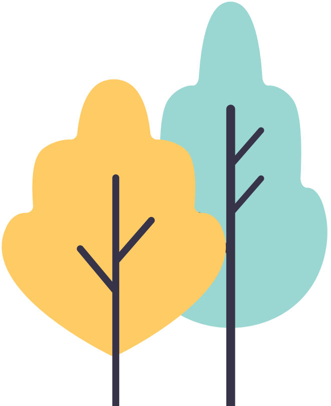
Melbourne Housing Price Prediction

Team 2

Chiebuka Onwuzurike, Tzu-Hua Huang, Yangyang Zhou, Yichi Zhang

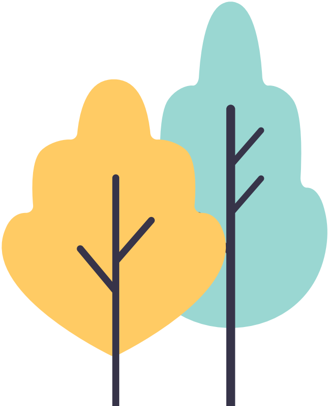
Outline

1. Business Objective
2. Dataset + Variables
3. Data Preparation
4. Exploratory Data Analysis
5. Models
    - Linear
    - Ridge
    - Lasso
    - Regression Tree
    - Bagging
    - Random Forest
    - Boosting
6. Conclusion



Objective

- Business Question/Objective
- Why it is important
- Other applications



Dataset

- Kaggle: https://www.kaggle.com/anthonypino/melbourne-housing-market
- 21 columns / 34,009 rows
- Goals:
  - Trying different models to make accurate price prediction
  - Finding important variables influence price the most
  - Giving advice to buyers and sellers
- Variables:
  - 8 characters: Address, Regionname, Type...
  - 7 integers: Rooms, Landsize, YearBuilt...
  - 6 numeric: BuildingArea, Distance...

Variables

Price: Price in Australian dollars

Type:

br - bedroom(s);

h - house,cottage,villa, semi,terrace;

u - unit, duplex;

t - townhouse;

dev site - development site;

o res - other residential.

Address: Address

Distance: Distance from Major City(km)

Latitude: Self explanatory

Longitude: Self explanatory

SellerG: Real Estate Agent

Method:

S - property sold;

SP - property sold prior;

PI - property passed in;

PN - sold prior not disclosed;

SN - sold not disclosed;

NB - no bid;

VB - vendor bid;

W - withdrawn prior to auction;

SA - sold after auction;

SS - sold after auction price not disclosed.

N/A - price or highest bid not available.

Date: Date sold

Regionname: General Region (West, North, etc)

Suburb: Suburb

Propertycount: Number of properties in the suburb.

Rooms: Number of rooms

Bedroom2 : Scraped # of Bedrooms

Bathroom: Number of Bathrooms

Car: Number of carspots

Landsize: Land Size in Metres

BuildingArea: Building Size in Metres

YearBuilt: Year the house was built

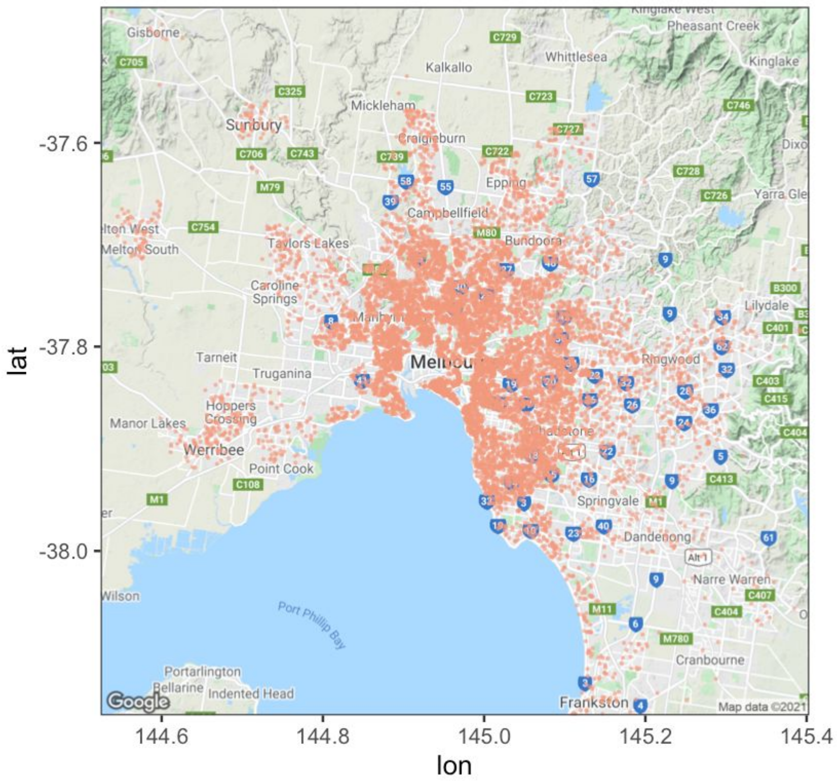
CouncilArea: Governing council for the area

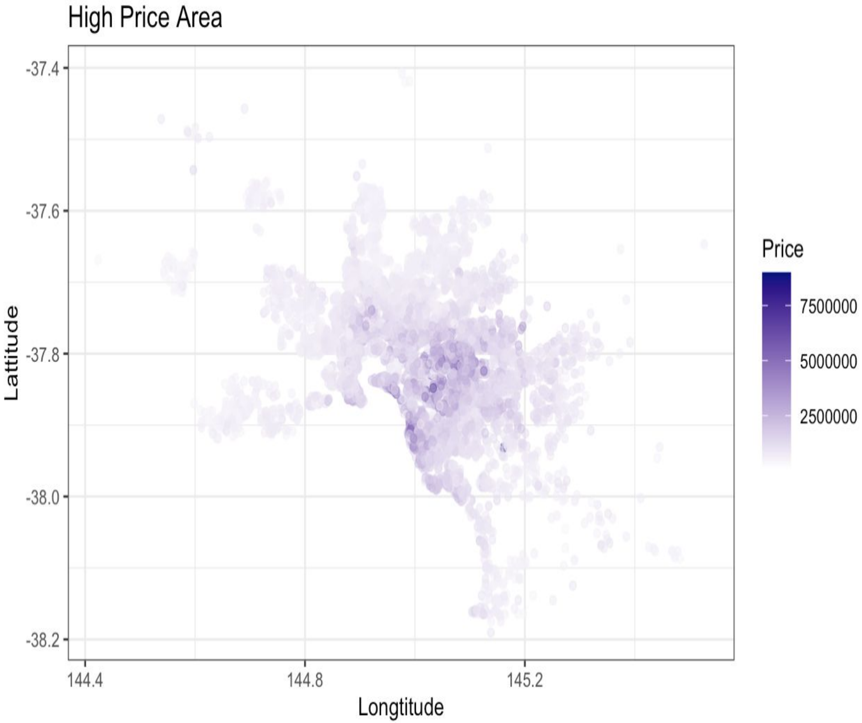
Data Preparation

- Data Formatting
- Null Values
  - Landsize, YearBuilt, Car, Latitude, BuildingArea
- Feature Engineering
  - SellYear, SellMonth
- Factored
  - Type, Method, Regionname, CouncilArea
- Scaling
- Splitting Train/Test (70/30)

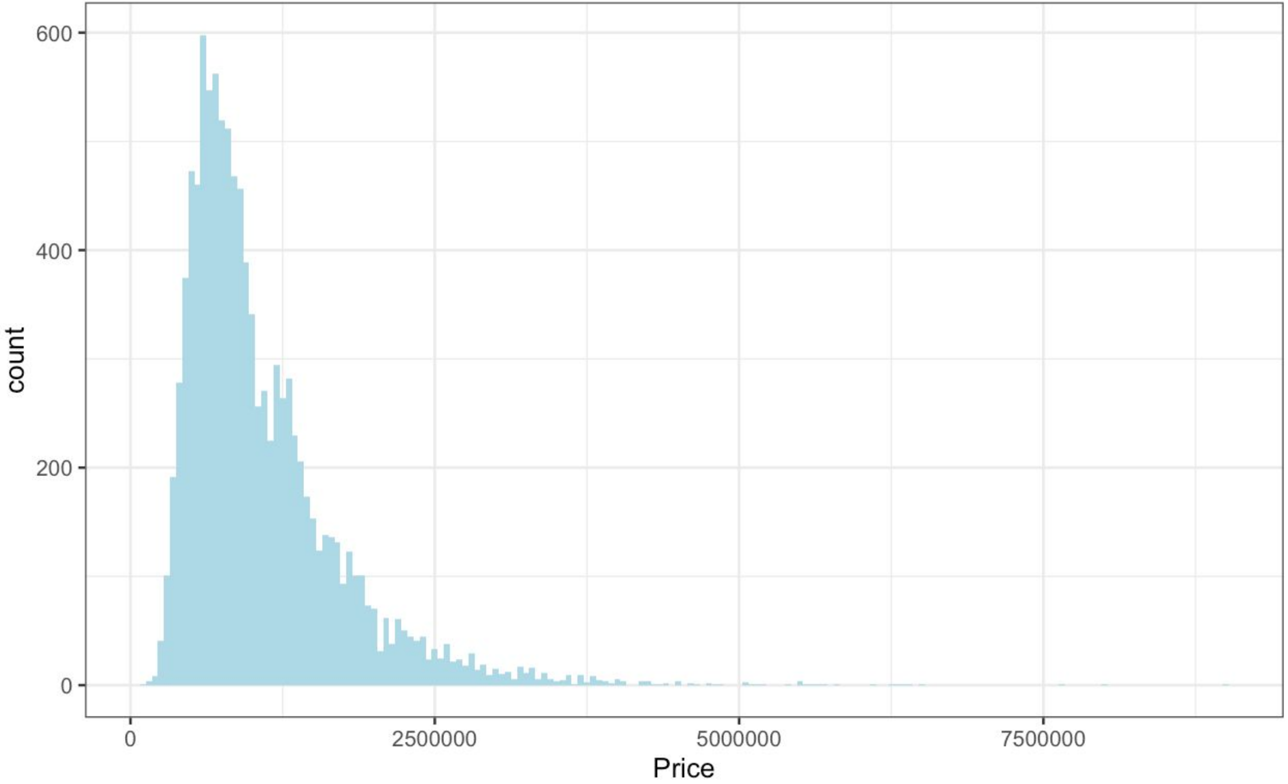
Exploratory Data Analysis

Exploratory Data Analysis



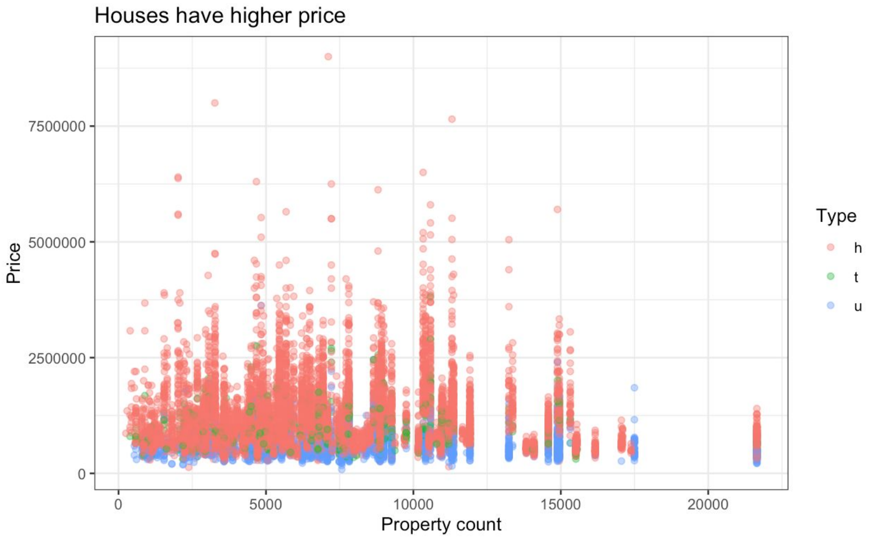


Price Range



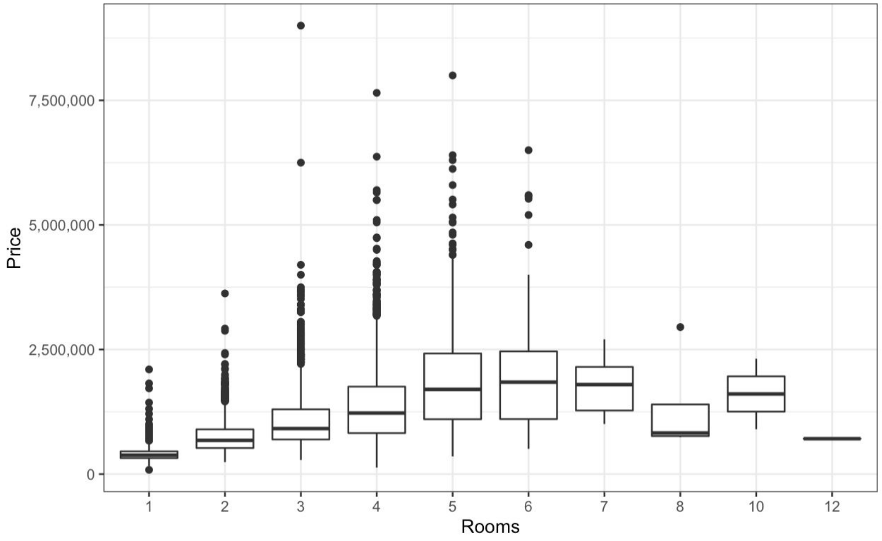
- The majority of those houses lies below $2,500,000

Price vs. Rooms/Types

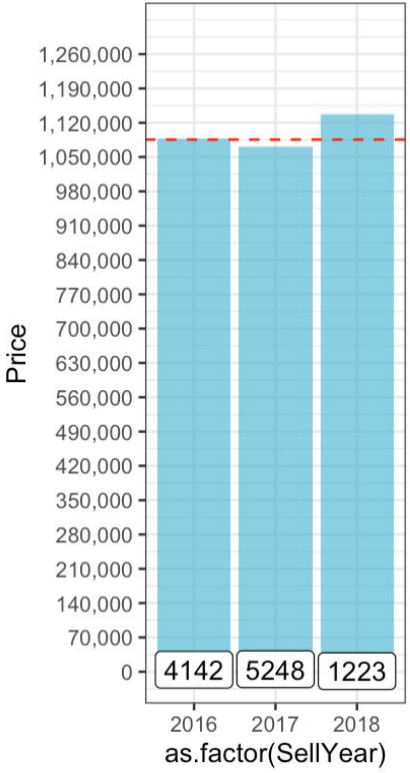


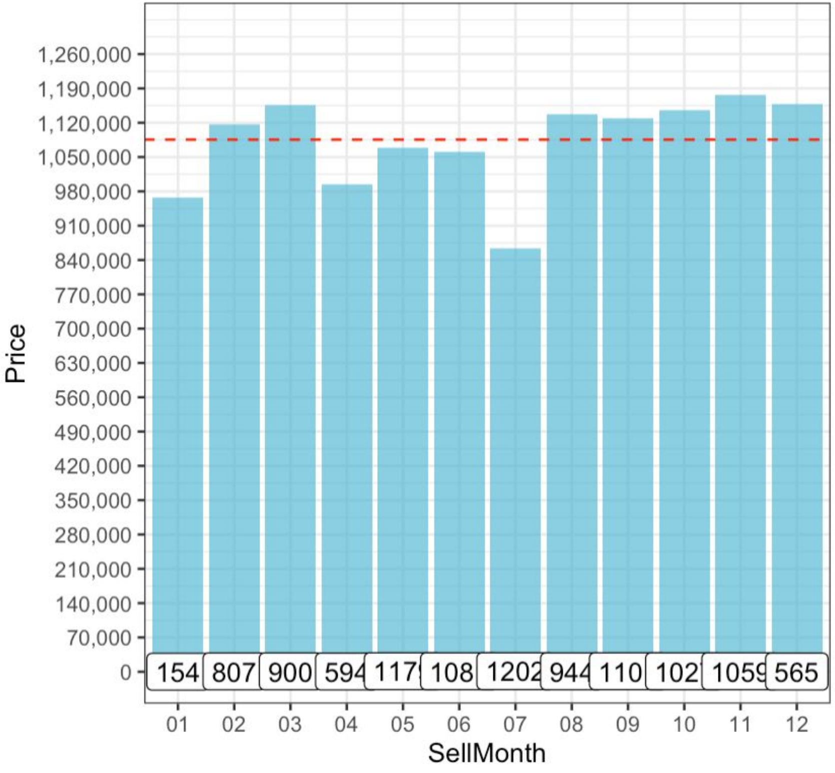
- House is the common type
- House has the highest price

- 4-6 rooms have higher price
- There are some outliers



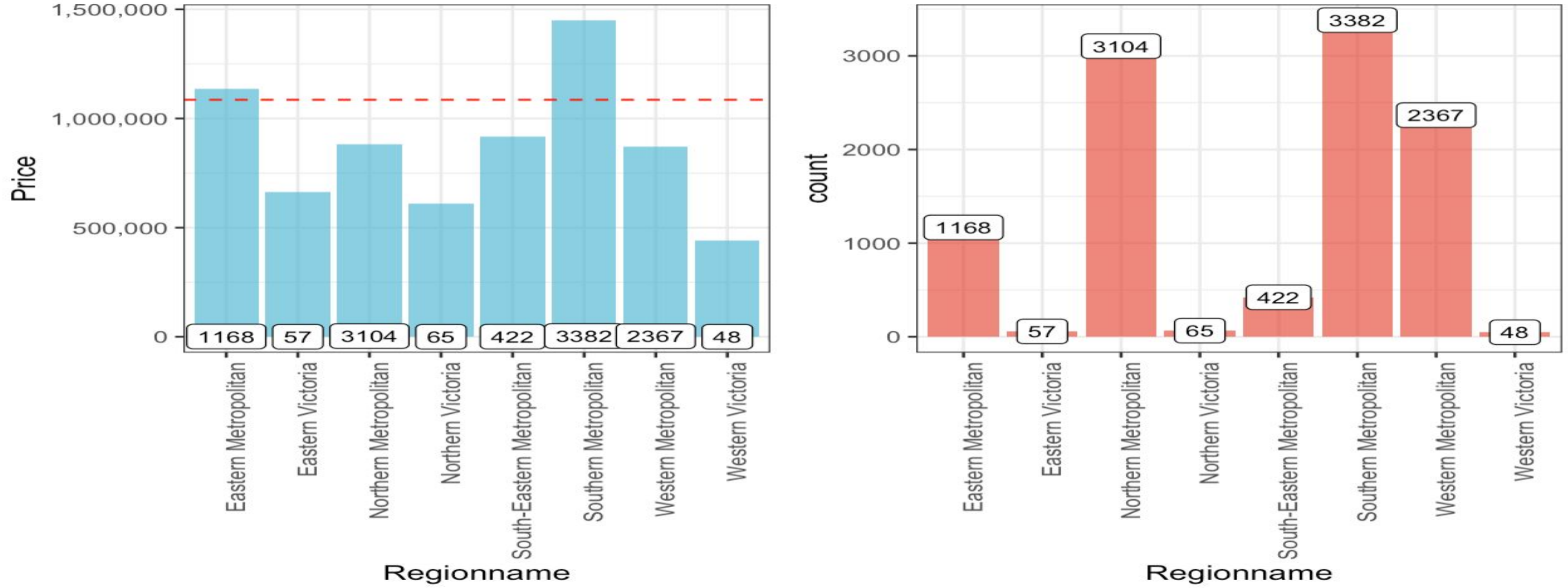
Price vs. Seasons





- January April and July have low house price
- Fall has higher house price than other seasons

Price vs. Regions



- Southern Metropolitan has the highest average price and more properties.
- Western Victoria has the lowest house price and least properties.

Linear Regression

Linear regression

- **Model 1**: Full model contains all variables

  Train MSE: 0.3314157 *Test MSE*: 0.3560723

- **Model 2**: Delete insignificant variables

  Train MSE: 0.3362878 *Test MSE*: 0.3626219

- **Model 3**: Combine 'Landsize' and 'Age' to create a interaction model

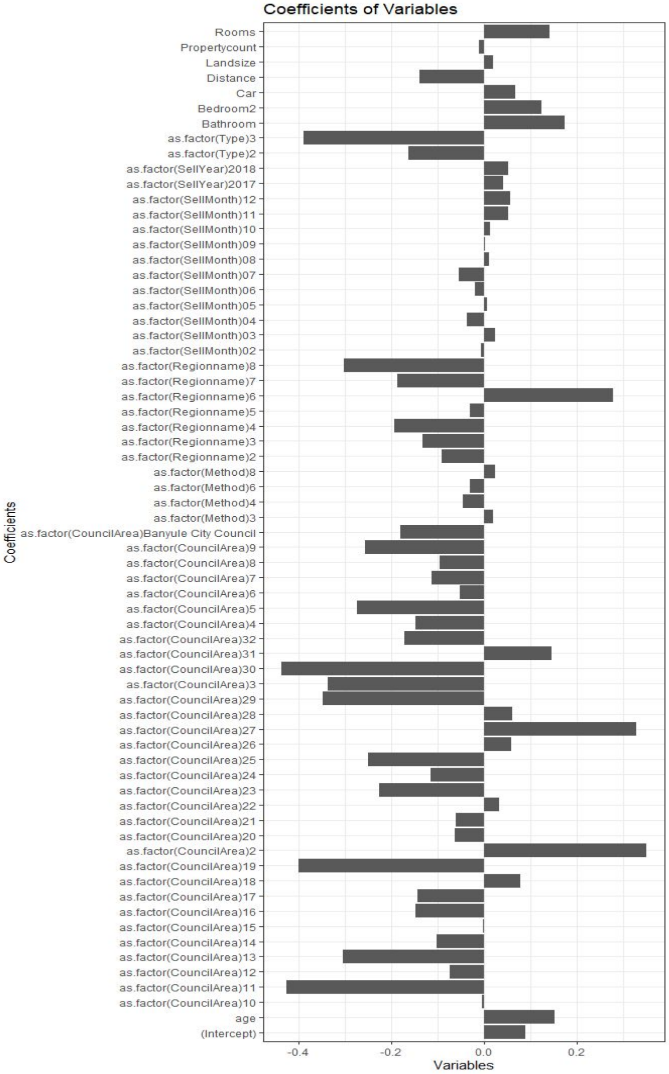
  Train MSE: 0.334264 *Test MSE*: 0.360624

Ridge & Lasso Regression

Ridge regression

- 10 fold cross validation model to find the best lambda
- CouncilArea seemed to have a strongest effect on the price of a house.
  - Boroondara City Council - 2
  - Hume City Council - 11
  - Wyndham City Council - 30

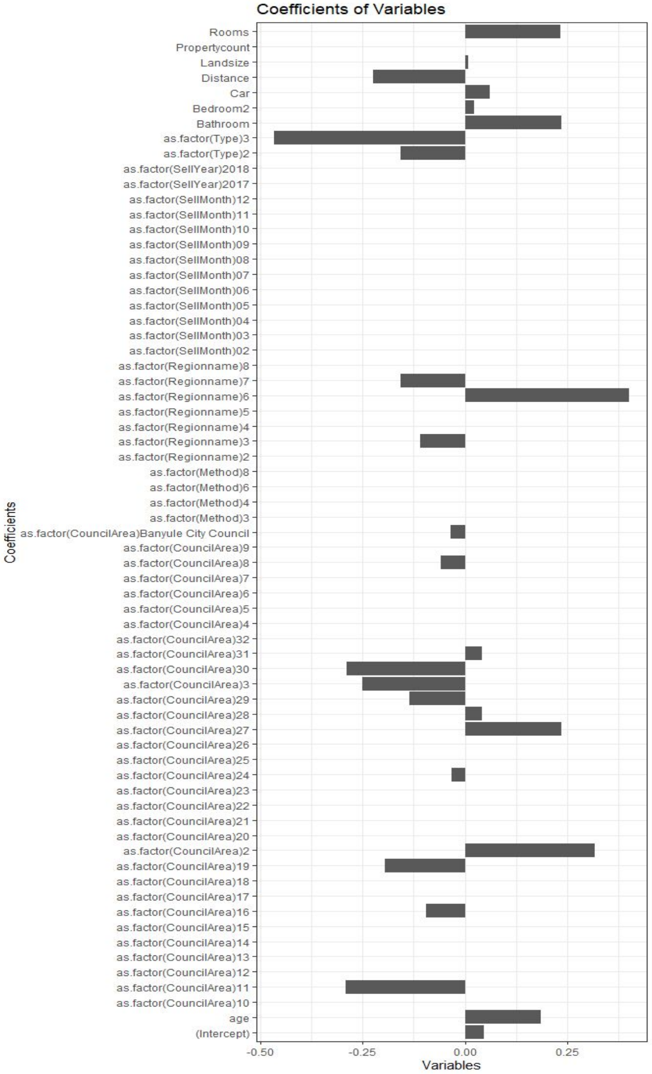
	Train MSE	Test MSE
Ridge	**0.3353068**	**0.3611925**
Lasso	TBD	TBD



Lasso regression

- 10 fold cross validation model to find the best lambda
- Again the CouncilArea, but now Type and Regionname are among the top 3
  - Boroondara City Council - CouncilArea 2
  - Unit style Home - Type 3
  - Southern Metropolitan - Regionname 6

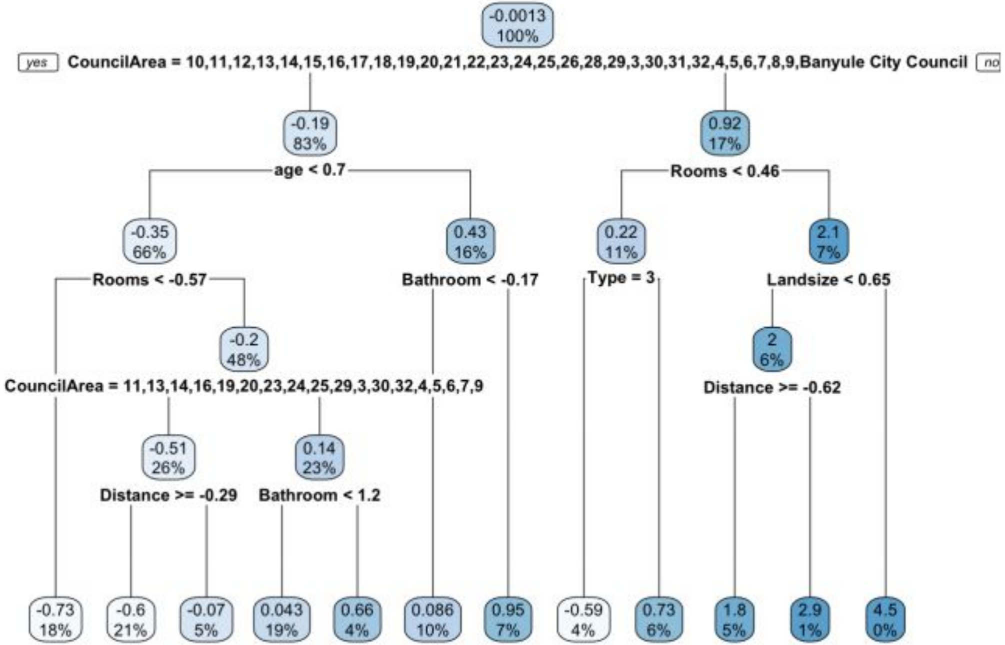
	Train MSE	Test MSE
Lasso	**0.3316269**	**0.3562078**
Ridge	0.3353068	0.3611925



Regression Trees

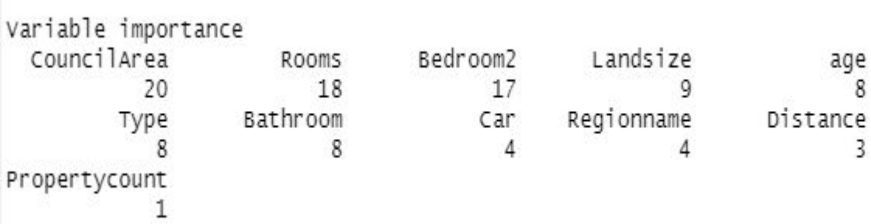
Regression Tree

- 12 variables are used to construct the tree with 11 internal nodes resulting in 12 terminal nodes.
- After 10 cross-validation : Train MSE:0.3827 ; Test MSE: 0.4011.

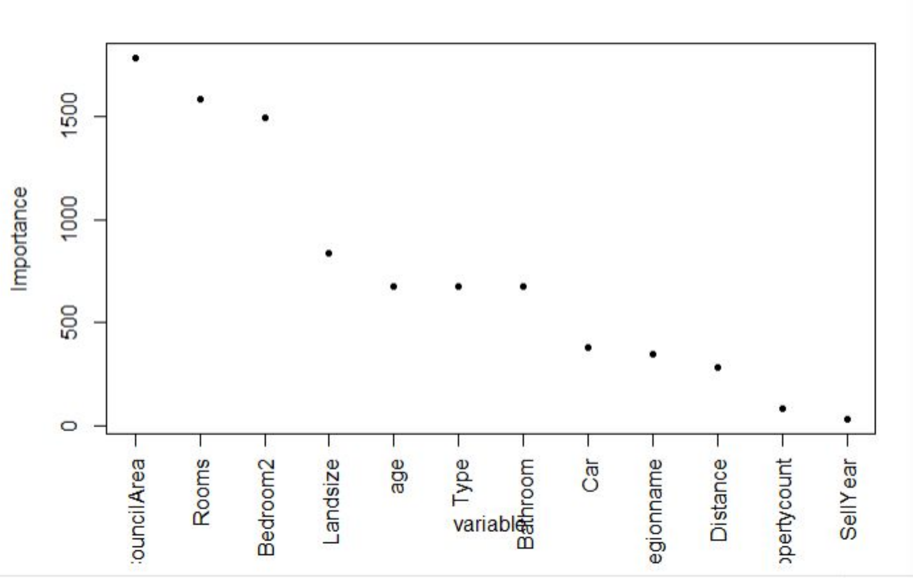


Regression Tree

- The CouncilArea , Rooms and bedrooms seem to have a stronger effect on the price of a house.



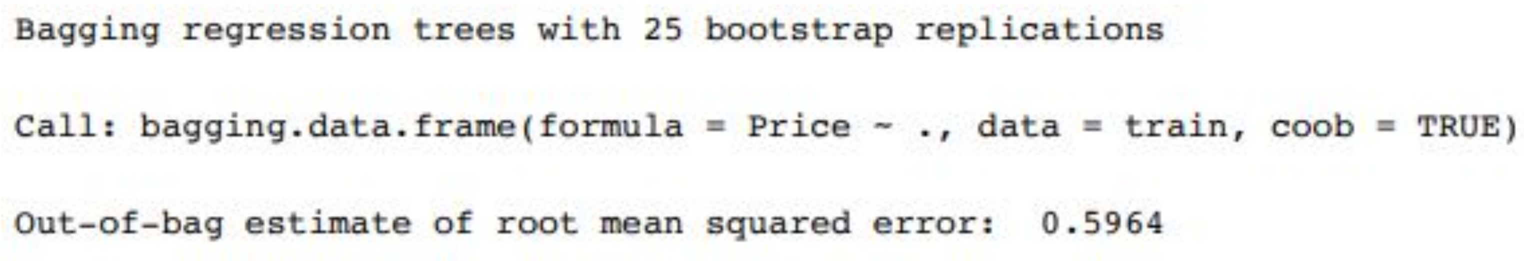
```
Variable importance
  CouncilArea          Rooms       Bedroom2       Landsize            age 
           20             18             17              9              8 
         Type       Bathroom            Car     Regionname       Distance 
            8              8              4              4              3 
Propertycount 
            1 
```



Bagging

Bagging

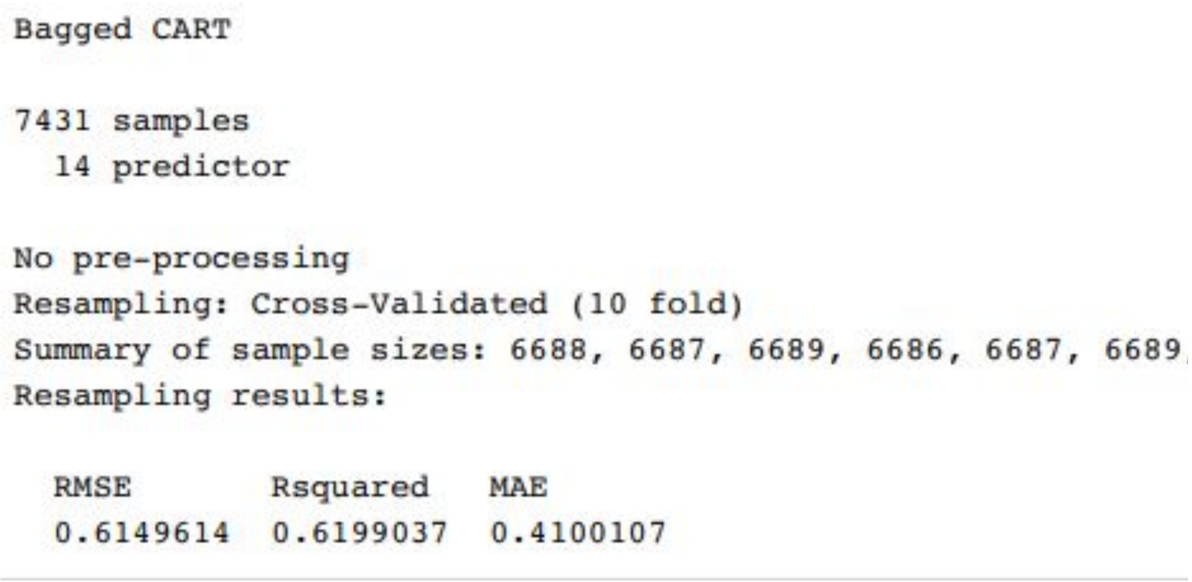
- We use bagging model and set coob = TRUE to use the OOB sample to estimate the test error.
- The test MSE is 0.3442
- The fit of Bagging is better than Regression Tree.



```
Bagging regression trees with 25 bootstrap replications

Call: bagging.data.frame(formula = Price ~ ., data = train, coob = TRUE)

Out-of-bag estimate of root mean squared error:  0.5964
```

Bagging

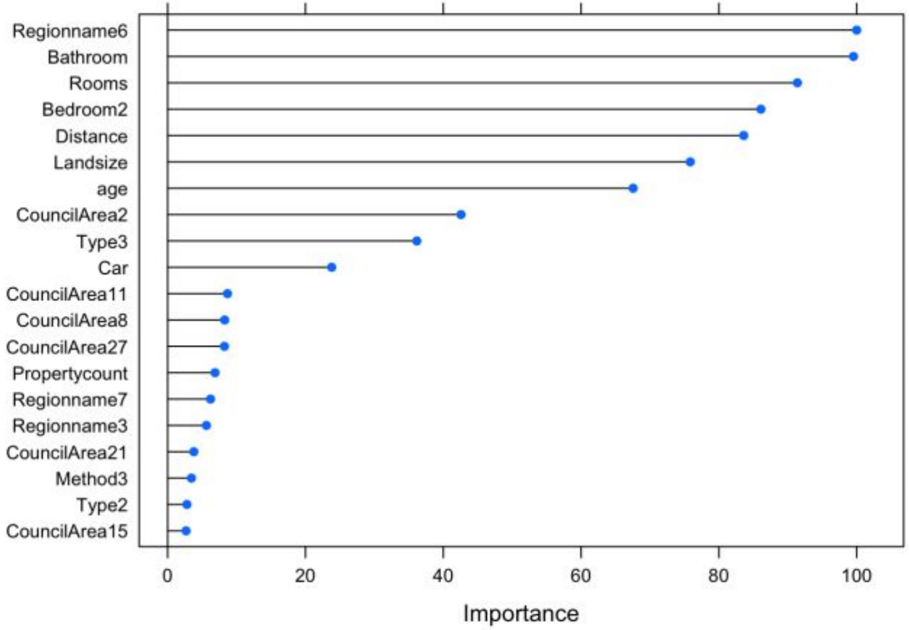
- Use caret package performing a 10-fold cross-validated by using bagging model.
- The test MSE is 0.3956.



```
Bagged CART

7431 samples
  14 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 6688, 6687, 6689, 6686, 6687, 6689,
Resampling results:

  RMSE       Rsquared   MAE
  0.6149614  0.6199037  0.4100107
```

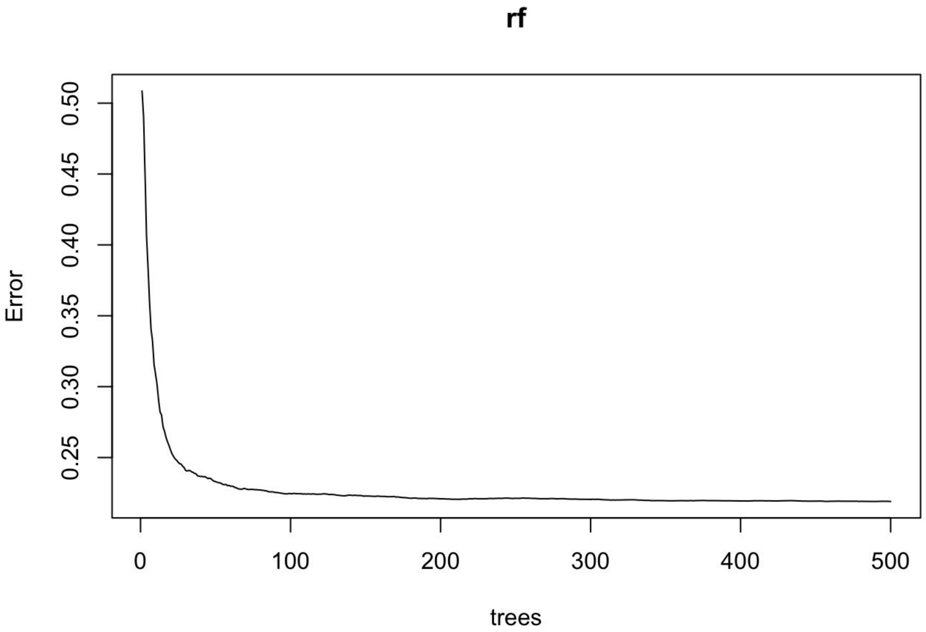
Bagging

- Feature importance:
- The Region name, bathrooms and rooms seem to have a stronger effect on the price of a house.



Random Forests

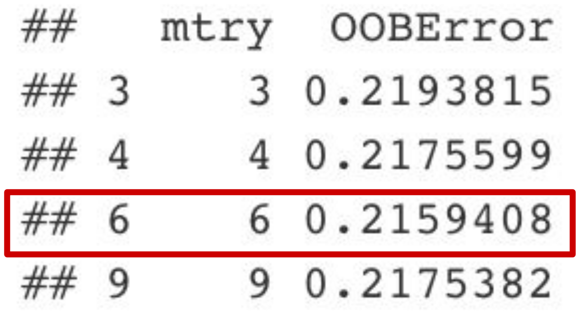
Random Forests



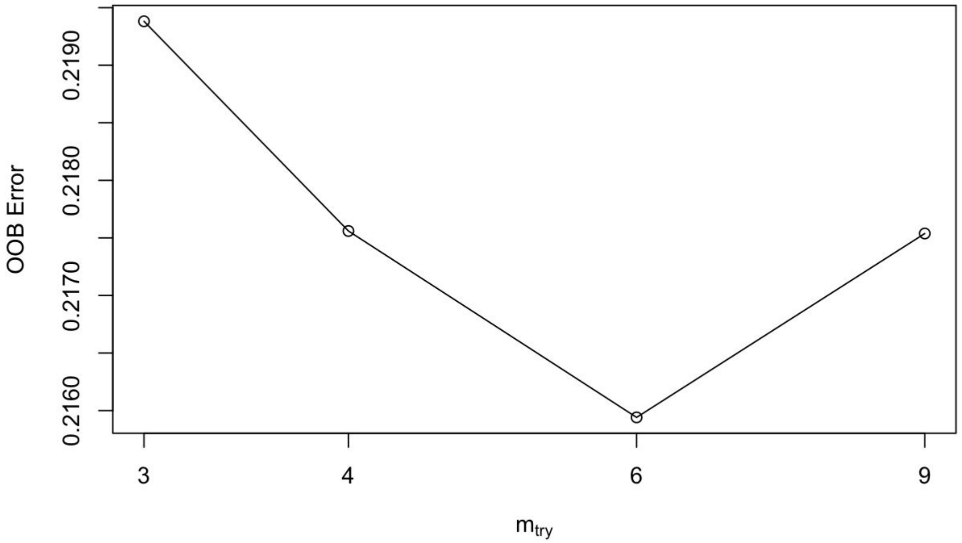
- Set ntree=500, mtry=4 (default)
- Test MSE: 0.2127

Optimal mtry Parameter

- tuneRF() function



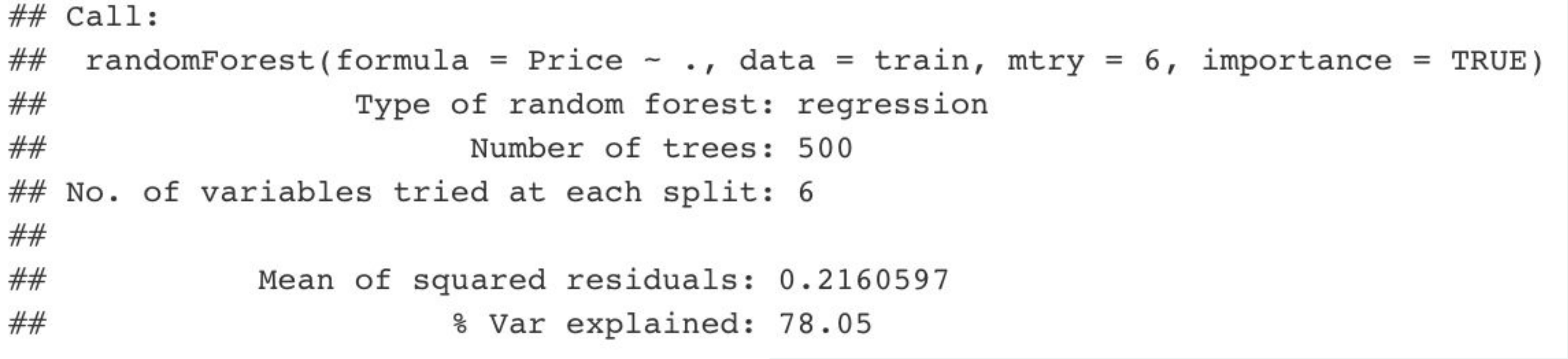
```
##     mtry   OOBError
## 3      3  0.2193815
## 4      4  0.2175599
## 6      6  0.2159408
## 9      9  0.2175382
```



Tune Using The Optimal mtry

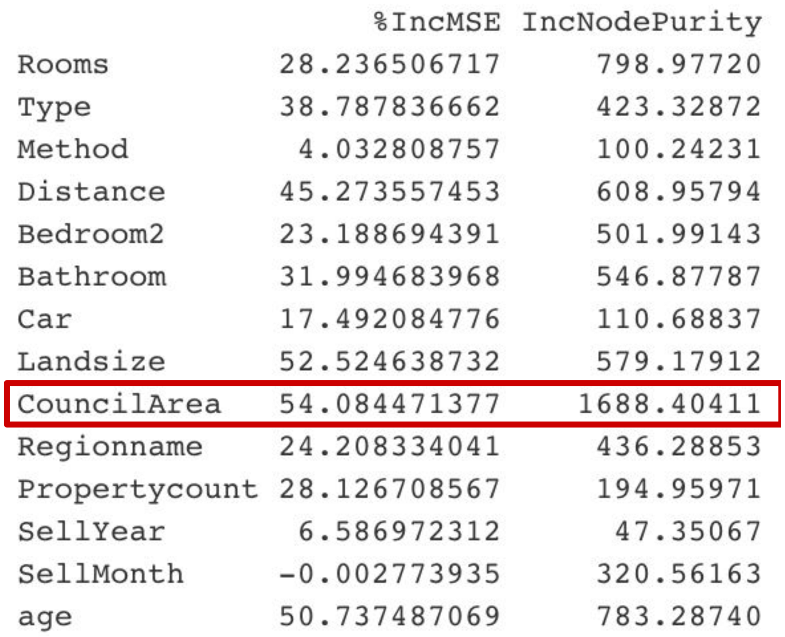
- Set ntree=500, mtry=6
- Test MSE: 0.2115

# of Trees	# of variables at each split	Test MSE
ntree=500	mtry=4	0.2127
ntree=500	mtry=6	0.2115

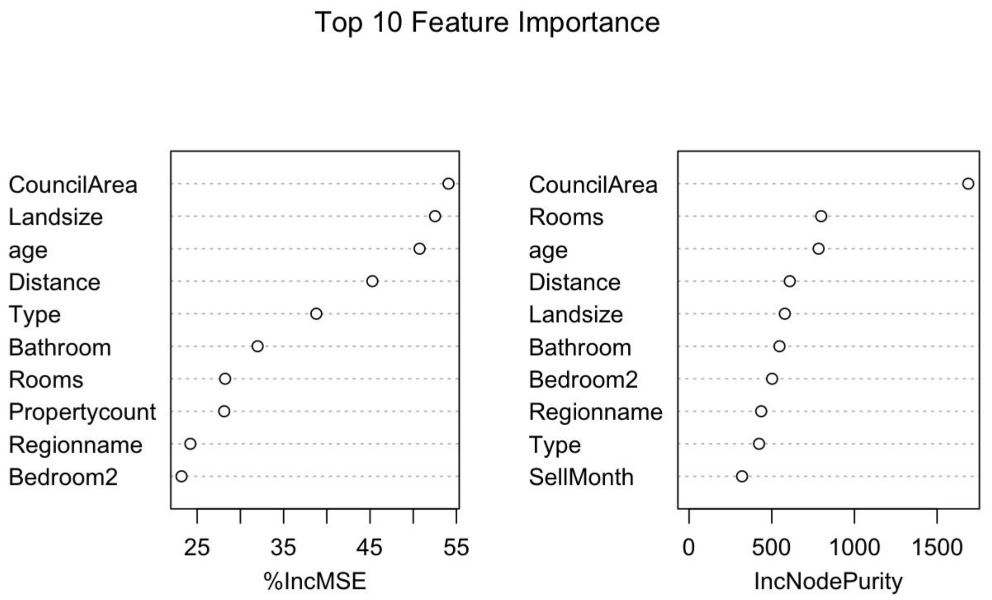


```
## Call:
##  randomForest(formula = Price ~ ., data = train, mtry = 6, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 6
##
##           Mean of squared residuals: 0.2160597
##                     % Var explained: 78.05
```

Feature Importance



```
                %IncMSE IncNodePurity
Rooms         28.236506717     798.97720
Type          38.787836662     423.32872
Method         4.032808757     100.24231
Distance      45.273557453     608.95794
Bedroom2      23.188694391     501.99143
Bathroom      31.994683968     546.87787
Car           17.492084776     110.68837
Landsize      52.524638732     579.17912
CouncilArea   54.084471377    1688.40411
Regionname    24.208334041     436.28853
Propertycount 28.126708567     194.95971
SellYear       6.586972312      47.35067
SellMonth     -0.002773935     320.56163
age           50.737487069     783.28740
```

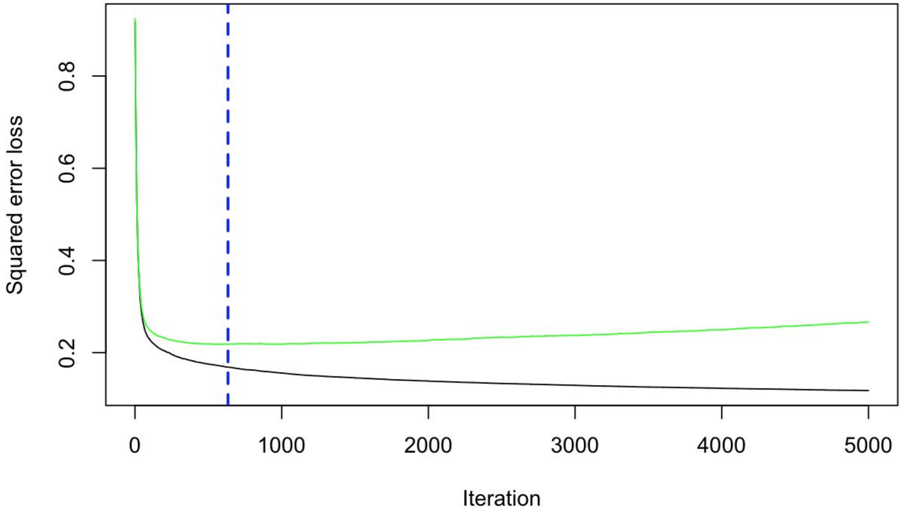


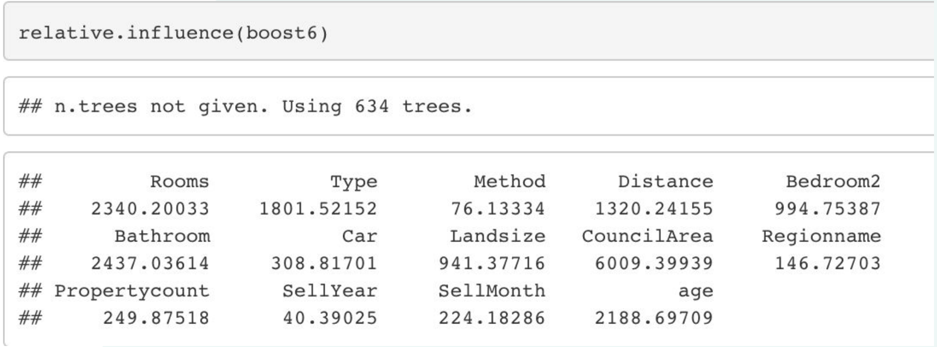
Boosting

Tune Manually

	# of Trees	shrinkage(learning rate)	Test MSE
boost1	n.trees=100	shrinkage=0.1	0.2525
boost2	n.trees=500	shrinkage=0.1	0.2072
boost3	n.trees=1000	shrinkage=0.1	0.2003
boost4	n.trees=5000	shrinkage=0.1	0.2014
boost5	n.trees=1000	shrinkage=0.2	0.2060

Optimal Iteration





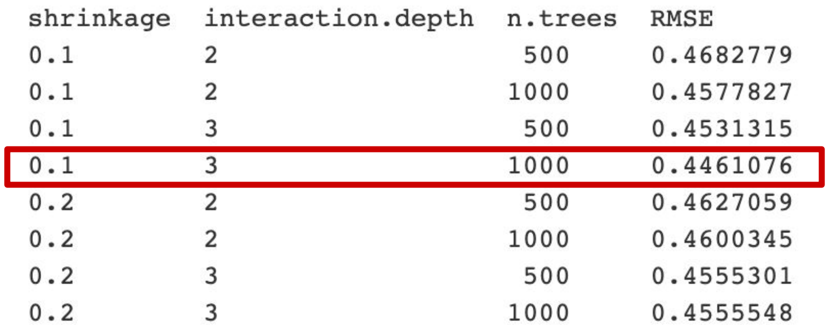
```
relative.influence(boost6)
```

```
## n.trees not given. Using 634 trees.
```

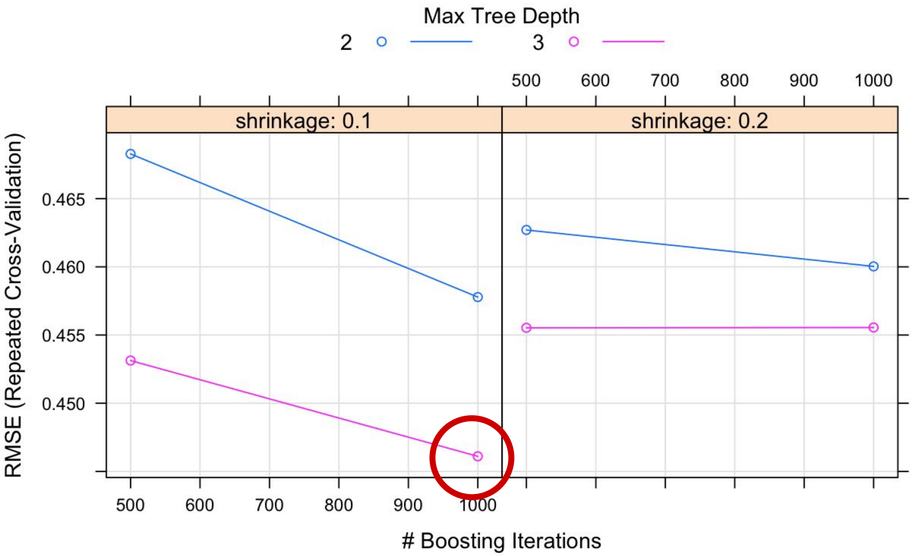
```
##          Rooms          Type        Method      Distance      Bedroom2
##     2340.20033    1801.52152      76.13334    1320.24155     994.75387
##       Bathroom           Car      Landsize   CouncilArea    Regionname
##     2437.03614     308.81701     941.37716    6009.39939     146.72703
## Propertycount      SellYear     SellMonth           age
##      249.87518      40.39025     224.18286    2188.69709
```

Tune Using Grid Search

- hyper_grid:
  - n.trees = 500, 1000
  - interaction.depth = 2, 3
  - shrinkage = 0.1, 0.2
  - n.minobsinnode = 10

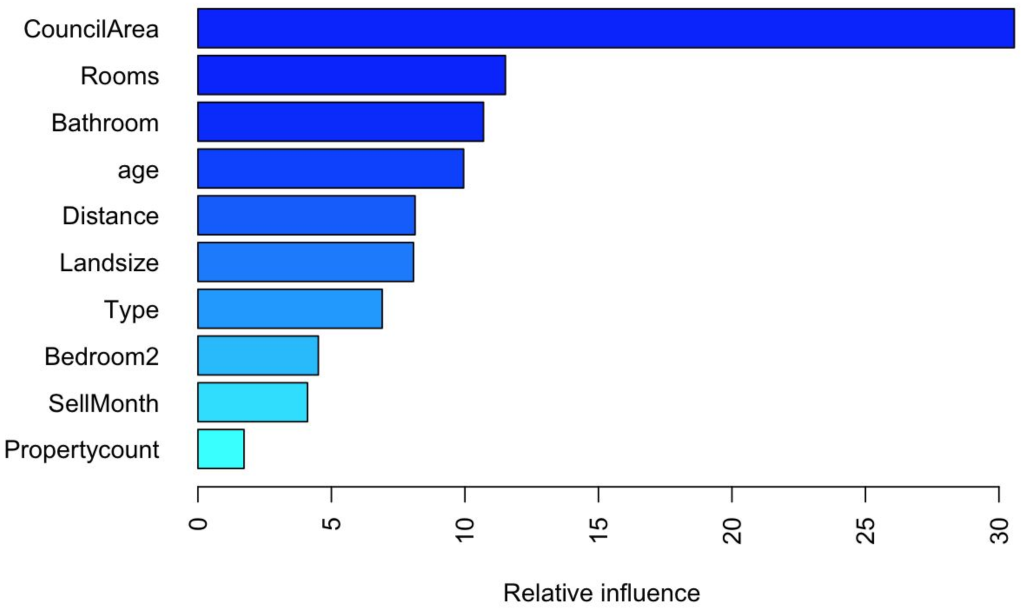


```
shrinkage  interaction.depth  n.trees  RMSE
0.1        2                   500     0.4682779
0.1        2                  1000     0.4577827
0.1        3                   500     0.4531315
0.1        3                  1000     0.4461076
0.2        2                   500     0.4627059
0.2        2                  1000     0.4600345
0.2        3                   500     0.4555301
0.2        3                  1000     0.4555548
```

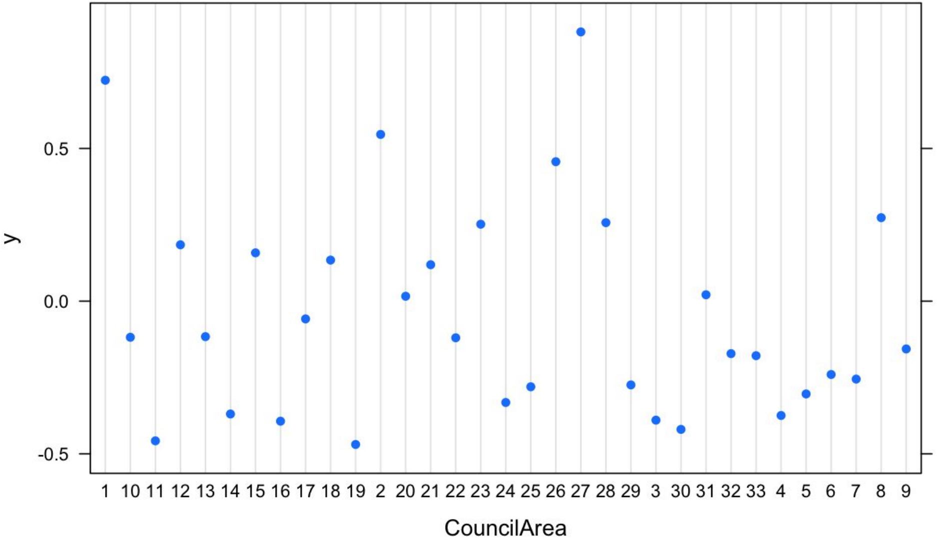


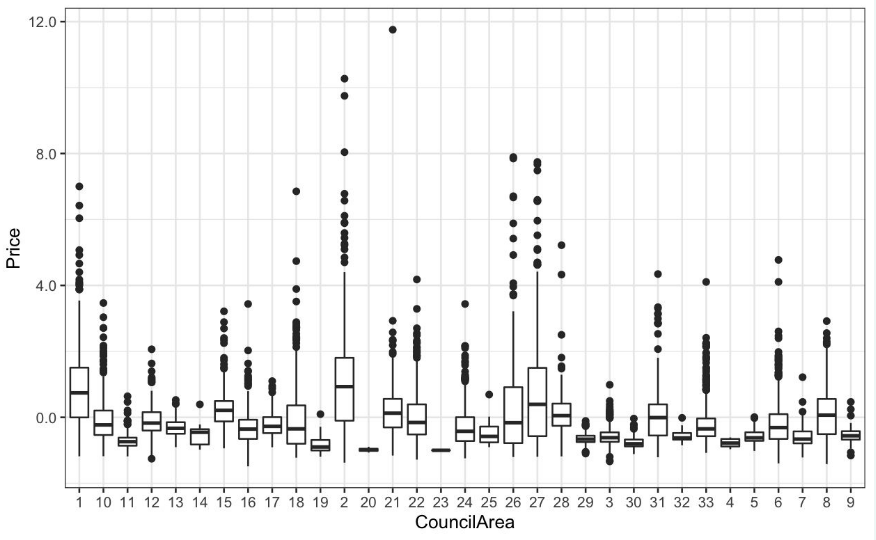
Feature Importance

- 10-fold cross validation
- Using tuned parameters:
  - n.trees = 1000
  - interaction.depth = 3
  - shrinkage = 0.1
- Test MSE: 0.1990



CouncilArea v.s. Price





- #1: Bayside City Council
- #2: Boroondara City Council
- #27: Stonnington City Council

Summary

Feature Importance:

CouncilArea

Model	Train MSE	Test MSE
Ridge Regression	0.3353068	0.3611925
Lasso Regression	0.3316269	0.3562078
Linear Regression	0.3314157	0.3560723
Regression Trees	0.3827541	0.4011495
Bagging	0.3605204	0.3956213
Random Forests	0.2012475	0.2115062
Boosting	0.1324614	0.1942090

Q&A