Random Forests Prediction

Anusha Raisinghani

- UW ID: 20823986
- Kaggle public score: 0.20693
- Kaggle submission count/times: 16

Summary

To obtain the final model, the following steps were applied:

- Preprocessing: This step involved creating some new variables and converting the existing variables into a more desirable form.
- Data Analysis: This step involved plotting the predictors against the response variable to better guess the relationship between the two. Additionally, some weird data points were identified and fixed in this step.
- Imputing Missing Data: The train and test data had missing values for some predictors such as `ayb`, `yr_rmdl`, `kitchens`, `stories` and `quadrant` which were imputed by intuition as well as by identifying the trends in those predictors.
- Model Building: To build the model, the `ranger` function in the `ranger` library was used. To improve the accuracy of the model, the parameters, `mtry`, `min.node.size` and `split.rule` was tuned using grid search. The best model was selected depending on which model had the highest R^2 value.

Preprocessing

The following steps were applied during preprocessing:

- `heat`: This variable was converted to a categorical variable.
- `ac`: This variable was converted to a categorical variable.
- `style`: This variable was converted to a categorical variable.
- `grade`: This variable was converted to a categorical variable.
- `cndtn`: This variable was converted to a categorical variable.
- `saledate`: This variable was converted to a numerical value representing the number of days since 1970/01/01.
- `extwall`: This variable was converted to a categorical variable.
- `roof`: This variable was converted to a categorical variable.
- `intwall`: This variable was converted to a categorical variable.
- `nbhd`: This variable was converted to a categorical variable.
- `ward`: This variable was converted to a categorical variable.
- `quadrant`: This variable was converted to a categorical variable.

Transformation (if any)

- `price`: I took the logarithm of the response variate price.

New Variables

- `yr_since_rmdl`: Difference in years between `saledate` and `yr_since_rmdl`
- `yr_since_imprv`: Difference in years between `saledate` and `eyb`

- `yr_since_b`: Difference in years between `saledate` and `ayb`
- `ysb1`: Boolean variable for whether the `yr_since_b` > 0
- `ysb2`: Boolean variable for whether the `yr_since_b` > 125

Other Preprocessing

Upon plotting the `bathrm` vs the `price`, it was found that there was a data point with 0 bathrooms. This data point had "No Data" listed for heat, 0 rooms, 0 bedrooms, etc. Hence, this data point was removed from the dataset. Similarly, for stories, there was an observation with 25 stories. The style for this house was "2.5 Story Fin" and hence the 25 was likely a typo. This was changed to 2.5. Lastly, there was a datapoint with 0 rooms, 0 bedrooms, 1 bathroom and 1 kitchen. This data point was dropped as well.

Model Building/Tuning

Main package used: `ranger`

Parameters tuned and their optimal values:

- mtry: 13
- min.node.size: 1
- splitrule: 'variance'

1.Preprocessing

1.1 Loading data

```
library(ranger)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

1.1 Loading data

```
load('RF.Rdata')

summary(dtrain)
```

```
##      bathrm         hf_bathrm          heat               ac
##  Min.   : 0.000   Min.   :0.0000   Length:6000        Length:6000
##  1st Qu.: 2.000   1st Qu.:0.0000   Class :character   Class :character
##  Median : 2.000   Median :1.0000   Mode  :character   Mode  :character
##  Mean   : 2.575   Mean   :0.8138
##  3rd Qu.: 3.000   3rd Qu.:1.0000
##  Max.   :12.000   Max.   :4.0000
##
##      rooms            bedrm             ayb          yr_rmdl          eyb
##  Min.   : 0.000   Min.   : 0.000   Min.   :1754   Min.   :1910   Min.   :1928
##  1st Qu.: 6.000   1st Qu.: 3.000   1st Qu.:1926   1st Qu.:2002   1st Qu.:1963
```

```
##  Median : 8.000   Median : 4.000   Median :1937   Median :2007   Median :1969
##  Mean   : 7.986   Mean   : 3.808   Mean   :1942   Mean   :2004   Mean   :1973
##  3rd Qu.: 9.000   3rd Qu.: 4.000   3rd Qu.:1951   3rd Qu.:2013   3rd Qu.:1978
##  Max.   :26.000   Max.   :15.000   Max.   :2018   Max.   :2018   Max.   :2018
##                                    NA's   :17     NA's   :2410
##     stories         saledate             price              gba
##  Min.   : 1.000   Length:6000        Min.   :   22000   Min.   :  480
##  1st Qu.: 2.000   Class :character   1st Qu.:  350000   1st Qu.: 1445
##  Median : 2.000   Mode  :character   Median :  641995   Median : 1870
##  Mean   : 1.975                      Mean   :  803310   Mean   : 2138
##  3rd Qu.: 2.250                      3rd Qu.:  980000   3rd Qu.: 2526
##  Max.   :25.000                      Max.   :15000000   Max.   :15902
##  NA's   :4
##     style              grade              cndtn             extwall
##  Length:6000        Length:6000        Length:6000        Length:6000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      roof             intwall             kitchens      fireplaces
##  Length:6000        Length:6000        Min.   :0.00   Min.   : 0.000
##  Class :character   Class :character   1st Qu.:1.00   1st Qu.: 1.000
##  Mode  :character   Mode  :character   Median :1.00   Median : 1.000
##                                        Mean   :1.04   Mean   : 1.163
##                                        3rd Qu.:1.00   3rd Qu.: 2.000
##                                        Max.   :3.00   Max.   :11.000
##                                        NA's   :1
##     landarea        latitude       longitude           nbhd
##  Min.   :  255   Min.   :40.63   Min.   :-74.26   B2     : 941
##  1st Qu.: 4300   1st Qu.:40.73   1st Qu.:-74.22   A2     : 446
##  Median : 5469   Median :40.75   Median :-74.19   F8     : 390
##  Mean   : 6286   Mean   :40.74   Mean   :-74.17   B9     : 312
##  3rd Qu.: 7300   3rd Qu.:40.77   3rd Qu.:-74.12   A7     : 295
##  Max.   :73771   Max.   :40.80   Max.   :-74.05   D2     : 274
##                                                   (Other):3342
##      ward             quadrant
##  Length:6000        Length:6000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
# CONVERTING CATEGORICAL VARIABLES TO FACTORS
dtrain$heat <- as.factor(dtrain$heat)
dtrain$ac <- as.factor(dtrain$ac)
dtrain$saledate <- as.numeric(as.Date(dtrain$saledate))
dtrain$style <- as.factor(dtrain$style)
dtrain$grade <- as.factor(dtrain$grade)
dtrain$cndtn <- as.factor(dtrain$cndtn)
dtrain$extwall <- as.factor(dtrain$extwall)
```

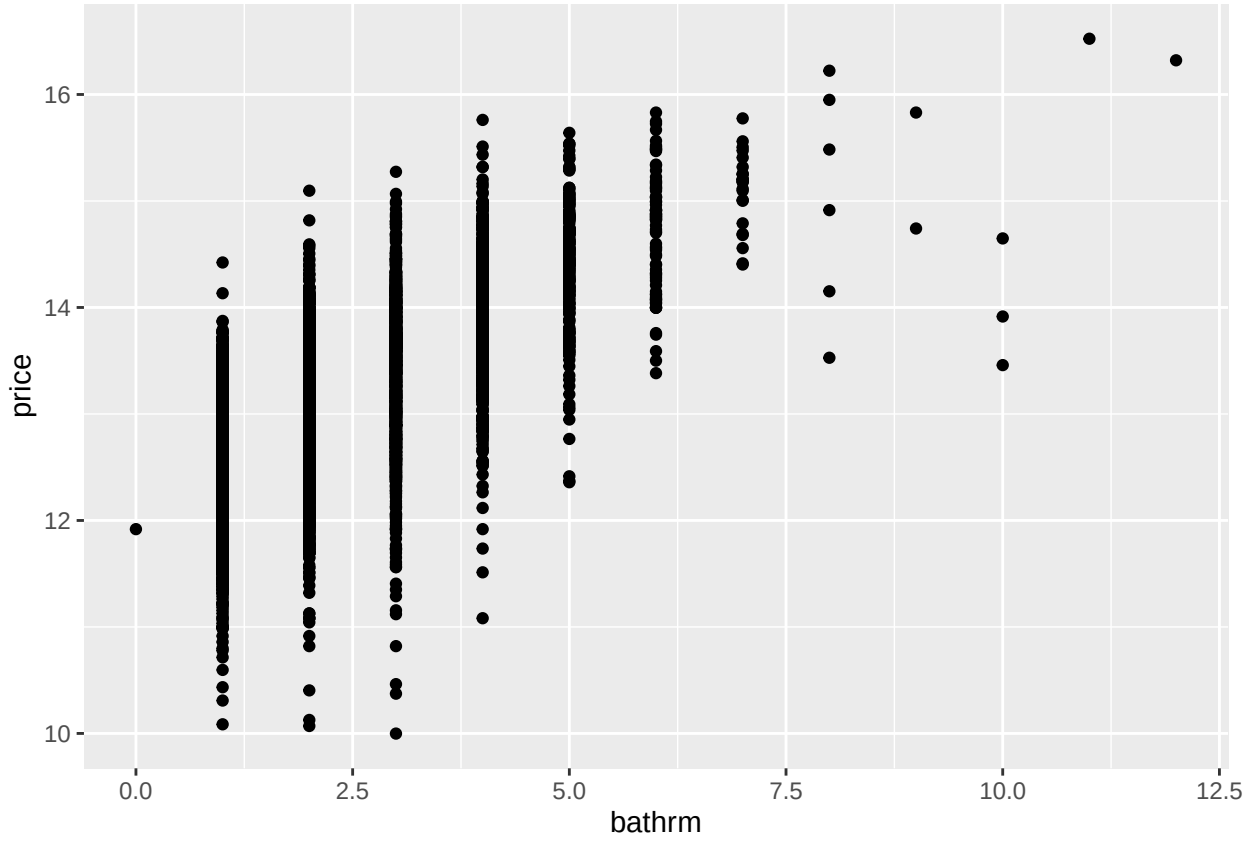
```r
dtrain$roof <- as.factor(dtrain$roof)
dtrain$intwall <- as.factor(dtrain$intwall)
dtrain$nbhd <- as.factor(dtrain$nbhd)
dtrain$ward <- as.factor(dtrain$ward)
dtrain$quadrant <- as.factor(dtrain$quadrant)
dtrain$price <- log(dtrain$price)
# CREATING NEW VARIABLES
dtrain$yr_since_rmdl <- 1970+dtrain$saledate/365.25-dtrain$yr_rmdl
dtrain$yr_since_imprv <- 1970+dtrain$saledate/365.25-dtrain$eyb

summary(dtrain)
```

```
##      bathrm         hf_bathrm                 heat      ac
##  Min.   : 0.000   Min.   :0.0000   Warm Cool     :2343   N: 958
##  1st Qu.: 2.000   1st Qu.:0.0000   Forced Air    :1908   Y:5042
##  Median : 2.000   Median :1.0000   Hot Water Rad :1659
##  Mean   : 2.575   Mean   :0.8138   Ht Pump       :  64
##  3rd Qu.: 3.000   3rd Qu.:1.0000   Water Base Brd:   7
##  Max.   :12.000   Max.   :4.0000   Elec Base Brd :   5
##                                    (Other)       :  14
##      rooms            bedrm            ayb          yr_rmdl          eyb
##  Min.   : 0.000   Min.   : 0.000   Min.   :1754   Min.   :1910   Min.   :1928
##  1st Qu.: 6.000   1st Qu.: 3.000   1st Qu.:1926   1st Qu.:2002   1st Qu.:1963
##  Median : 8.000   Median : 4.000   Median :1937   Median :2007   Median :1969
##  Mean   : 7.986   Mean   : 3.808   Mean   :1942   Mean   :2004   Mean   :1973
##  3rd Qu.: 9.000   3rd Qu.: 4.000   3rd Qu.:1951   3rd Qu.:2013   3rd Qu.:1978
##  Max.   :26.000   Max.   :15.000   Max.   :2018   Max.   :2018   Max.   :2018
##                                    NA's   :17     NA's   :2410
##     stories          saledate         price             gba
##  Min.   : 1.000   Min.   : 8039   Min.   : 9.999   Min.   :  480
##  1st Qu.: 2.000   1st Qu.:12258   1st Qu.:12.766   1st Qu.: 1445
##  Median : 2.000   Median :14890   Median :13.372   Median : 1870
##  Mean   : 1.975   Mean   :14241   Mean   :13.285   Mean   : 2138
##  3rd Qu.: 2.250   3rd Qu.:16605   3rd Qu.:13.795   3rd Qu.: 2526
##  Max.   :25.000   Max.   :17725   Max.   :16.524   Max.   :15902
##  NA's   :4
##              style                grade            cndtn
##  2 Story        :3440   Good Quality :1638   Average  :1947
##  2.5 Story Fin  :1136   Above Average:1581   Excellent:  85
##  1 Story        : 623   Very Good    :1140   Fair     :  27
##  1.5 Story Fin  : 380   Average      : 850   Good     :3261
##  3 Story        : 179   Excellent    : 390   Poor     :   8
##  2.5 Story Unfin: 116   Superior     : 230   Very Good: 672
##  (Other)        : 126   (Other)      : 171
##          extwall               roof               intwall        kitchens
##  Common Brick:3217   Comp Shingle:3472   Hardwood     :5001   Min.   :0.00
##  Wood Siding : 618   Slate       :1922   Hardwood/Carp: 661   1st Qu.:1.00
##  Brick/Siding: 544   Built Up    : 213   Wood Floor   : 192   Median :1.00
##  Vinyl Siding: 496   Metal- Sms  : 128   Carpet       : 130   Mean   :1.04
##  Stucco      : 383   Shake       : 106   Lt Concrete  :   7   3rd Qu.:1.00
##  Shingle     : 159   Clay Tile   :  76   Ceramic Tile :   3   Max.   :3.00
##  (Other)     : 583   (Other)     :  83   (Other)      :   6   NA's   :1
##    fireplaces        landarea        latitude       longitude
##  Min.   : 0.000   Min.   :  255   Min.   :40.63   Min.   :-74.26
```

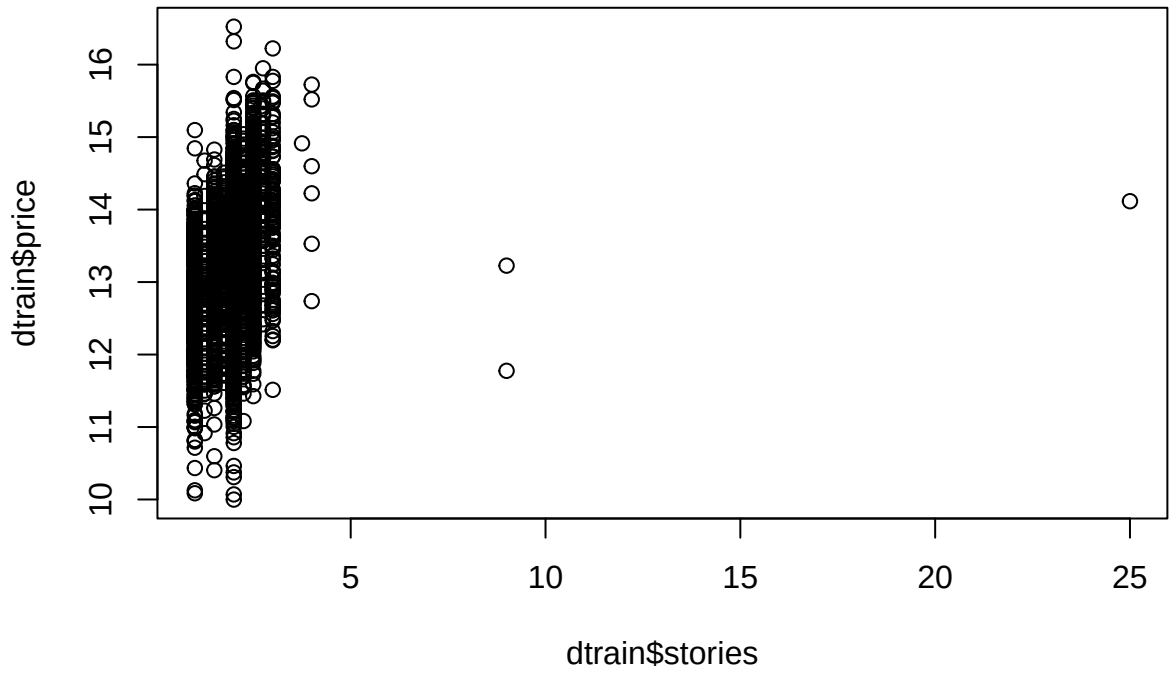
```
##  1st Qu.: 1.000   1st Qu.: 4300   1st Qu.:40.73   1st Qu.:-74.22
##  Median : 1.000   Median : 5469   Median :40.75   Median :-74.19
##  Mean   : 1.163   Mean   : 6286   Mean   :40.74   Mean   :-74.17
##  3rd Qu.: 2.000   3rd Qu.: 7300   3rd Qu.:40.77   3rd Qu.:-74.12
##  Max.   :11.000   Max.   :73771   Max.   :40.80   Max.   :-74.05
##
##       nbhd           ward      quadrant    yr_since_rmdl     yr_since_imprv
##  B2     : 941   Ward 3 :2271   NE  :1230   Min.   :-21.927   Min.   :-14.52
##  A2     : 446   Ward 4 :1742   NW  :3995   1st Qu.:  0.205   1st Qu.: 29.49
##  F8     : 390   Ward 7 : 830   SE  : 730   Median :  1.425   Median : 40.23
##  B9     : 312   Ward 5 : 782   SW  :  13   Mean   :  5.652   Mean   : 36.43
##  A7     : 295   Ward 8 : 241   NA's:  32   3rd Qu.:  9.505   3rd Qu.: 47.17
##  D2     : 274   Ward 2 :  84               Max.   :108.370   Max.   : 78.26
##  (Other):3342   (Other):  50               NA's   :2410
```

```
ggplot(dtrain, aes(x = bathrm, y = price)) +
  geom_point()
```



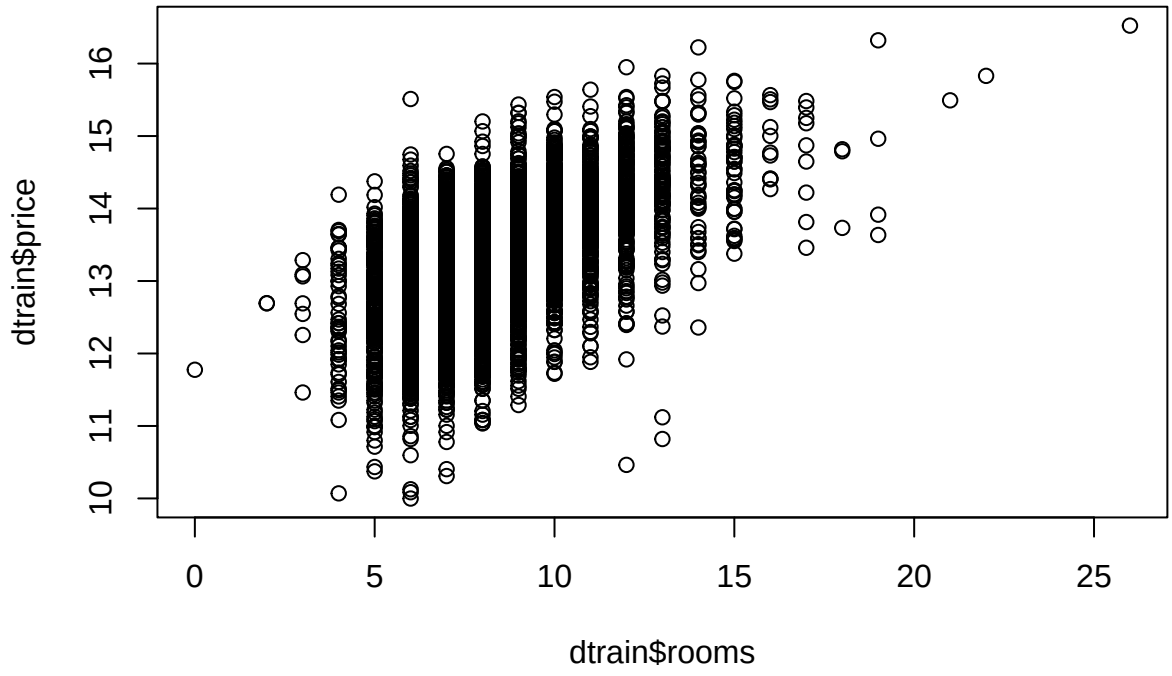

```
# There's a house with 0 bathrooms. Upon close inspection of this data point,
# we see weird data, there's "No Data" for heat, no rooms, and ayb > eyb.
# We get rid of this data point:
dtrain <- dtrain[dtrain$bathrm > 0,]

plot(dtrain$stories, dtrain$price)
```




```
# This is prolly a typo
dtrain[296, 'stories'] <- 2.5

plot(dtrain$rooms, dtrain$price)
```



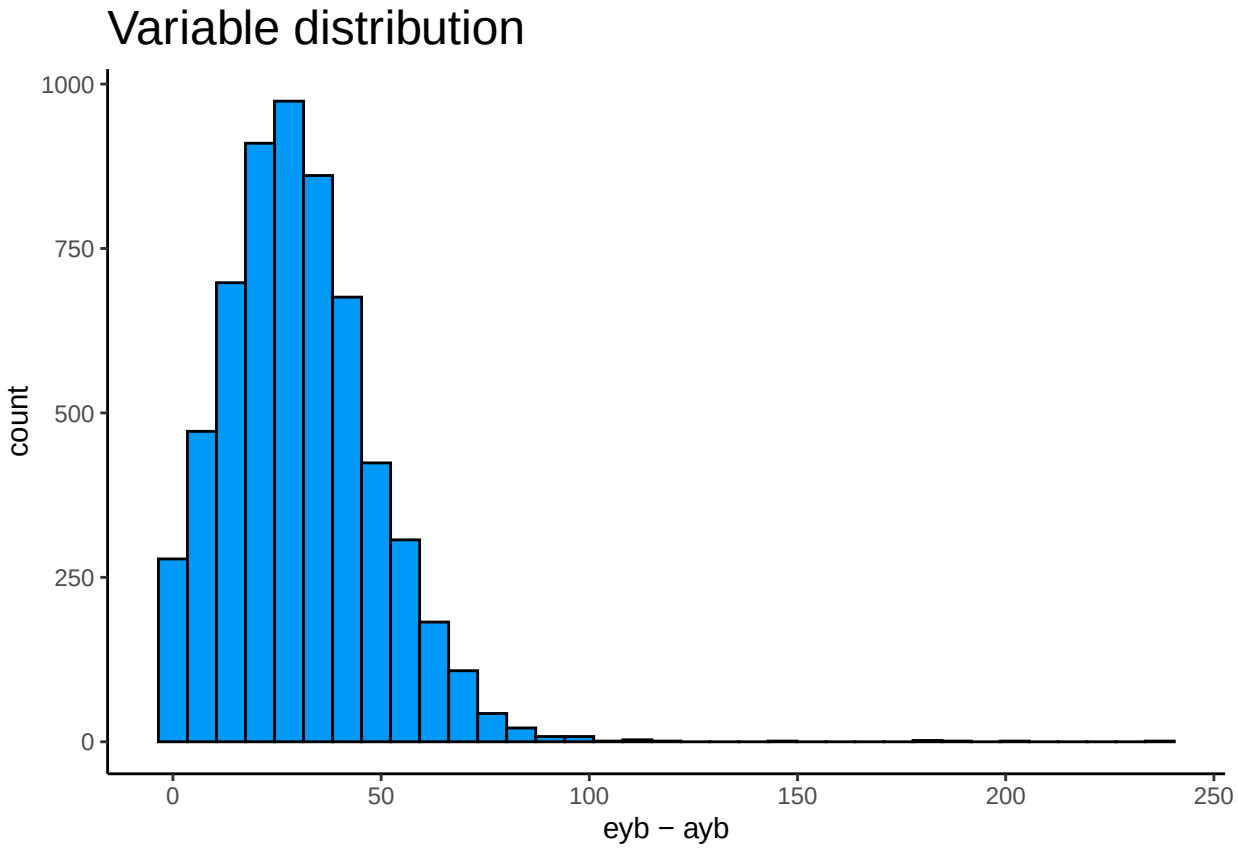

```
# This house has 0 rooms, 0 bedrooms, 1 bathrm and 1 kitchen, maybe drop it?
dtrain <- dtrain[dtrain$rooms > 0,]
```

1.2 Missing data handling

For `ayb`:

Upon plotting the distribution of `eyb - ayb`, we get a bell-shaped curve indicating a normal distribution. Hence, the `ayb` is imputed by using the mean of `eyb-ayb`.

```
# This follows a normal distribution
ggplot(dtrain[!is.na(dtrain$ayb),], aes(eyb - ayb)) +
  geom_histogram(color = "#000000", fill = "#0099F8", bins=35) +
  ggtitle("Variable distribution") +
  theme_classic() +
  theme(plot.title = element_text(size = 18))
```




```
# We do mean imputation
mean_ayb_eyb <- round(mean(dtrain$eyb - dtrain$ayb, na.rm=TRUE))
dtrain[is.na(dtrain$ayb),'ayb'] <- dtrain[is.na(dtrain$ayb),'eyb'] - mean_ayb_eyb

# Create yr_since_b variable
dtrain$yr_since_b <- 1970+dtrain$saledate/365.25-dtrain$ayb
```

For `yr_rmdl` (`yr_since_rmdl`)

We are using the `yr_since_rmdl` in our model, and it has missing values due to missing values in `yr_rmdl`. These houses were never remodeled and so for imputing the `yr_since_rmdl`, the 2 plus the maximum of the `yr_since_rmdl` is taken.

```
impute_value_rmdl <- max(dtrain$yr_since_rmdl, na.rm=TRUE) + 2
dtrain[is.na(dtrain$yr_since_rmdl), 'yr_since_rmdl'] <- impute_value_rmdl
```

For `stories`

This was missing for 4 observations. The styles of these houses are listed as "2 Story" and "2.5 Story". Hence, these houses were imputed by taking the mean of the stories of "2 Story" and "2.5 Story" style houses.

```
dtrain[is.na(dtrain$stories), ]
```

```
##      bathrm hf_bathrm       heat ac rooms bedrm  ayb yr_rmdl  eyb stories
## 893       2         1 Forced Air  Y     8     4 1940      NA 1940      NA
## 1493      3         1 Forced Air  Y     7     4 2014      NA 2015      NA
## 2777      5         1  Warm Cool  Y    13     5 2016      NA 2017      NA
## 4644      5         1 Forced Air  Y    12     5 2014      NA 2016      NA
##      saledate    price  gba         style        grade     cndtn      extwall
## 893     17626 12.25486 2124       2 Story  Low Quality      Poor Brick/Stucco
## 1493    16323 13.54419 2816       2 Story Good Quality Very Good Common Brick
## 2777    16932 14.71160 4013       2 Story    Excellent Excellent Stone/Siding
## 4644    15237 13.88036 7163 2.5 Story Fin Good Quality Very Good Stone/Siding
##              roof       intwall kitchens fireplaces landarea latitude longitude
## 893      Built Up Hardwood/Carp        1          0     1062 40.69787 -74.14344
## 1493 Comp Shingle      Hardwood        1          1     5565 40.77102 -74.14692
## 2777 Comp Shingle      Hardwood        1          1     8878 40.73053 -74.24346
## 4644 Comp Shingle Hardwood/Carp        1          2    11925 40.77241 -74.20657
##      nbhd   ward quadrant yr_since_rmdl yr_since_imprv yr_since_b
## 893    B1 Ward 6       NE      110.3696     78.2573580 78.2573580
## 1493   F3 Ward 4       NE      110.3696     -0.3100616  0.6899384
## 2777   E6 Ward 3       NW      110.3696     -0.6427105  0.3572895
## 4644   B2 Ward 4       NW      110.3696     -4.2833676 -2.2833676
```

```
# Get the mean of stories grouped by style
tapply(dtrain$stories, dtrain$style, mean, na.rm=TRUE)
```

```
##         1 Story   1.5 Story Fin 1.5 Story Unfin         2 Story   2.5 Story Fin
##        1.018489        1.475658        1.428571        1.996144        2.443965
## 2.5 Story Unfin         3 Story         4 Story        Bi-Level         Default
##        2.368534        2.861453        3.550000        1.500000        5.750000
##     Split Foyer     Split Level
##        1.232558        1.500000
```

```
# Add the mean of stories as a column (avg_stories) to dtrain
dtrain <- dtrain %>%
  group_by(style) %>%
  mutate(avg_stories = mean(stories, na.rm = TRUE))
# Set the missing stories to the mean grouped by style
dtrain <- dtrain %>%
  mutate(stories = ifelse(is.na(stories), avg_stories, stories))
# Drop the avg_stories column
dtrain <- dtrain %>%
  select(-avg_stories)
```

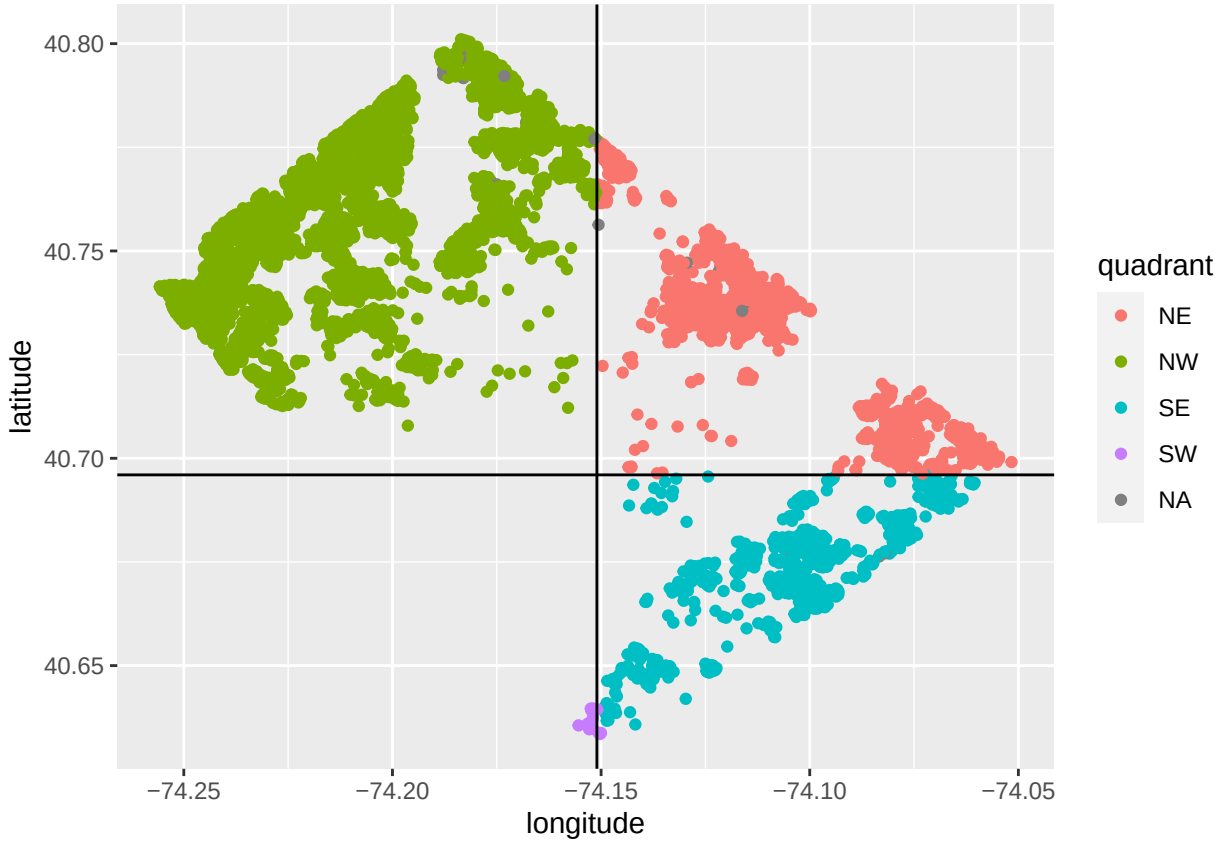
For `kitchens`

```
dtrain[which(is.na(dtrain$kitchens)), 'kitchens'] <-
  dtrain[which(is.na(dtrain$kitchens)), 'rooms'] - (dtrain[which(is.na(dtrain$kitchens)), 'bedrm']
                                                    + dtrain[which(is.na(dtrain$kitchens)), 'bathrm'])
```

For `quadrant`

When we plot the graph of latitude and longitude grouped by quadrant, we see clear distinctions between each quadrant. So the quadrant is imputed on the basis of the latitude and longitude of the house.

```
ggplot(dtrain, aes(x = longitude, y = latitude, color = quadrant)) +
  geom_point() + geom_vline(xintercept=-74.151) + geom_hline(yintercept = 40.696)
```




```
dtrain[(dtrain$longitude < -74.151) & (dtrain$latitude > 40.696) & is.na(dtrain$quadrant), 'quadrant'] <
dtrain[(dtrain$longitude > -74.151) & (dtrain$latitude > 40.696) & is.na(dtrain$quadrant), 'quadrant'] <
dtrain[(dtrain$longitude > -74.151) & (dtrain$latitude < 40.696) & is.na(dtrain$quadrant), 'quadrant'] <
```

2. Model building

For model building, the `ranger` function in the `ranger` library was used. For this, the parameters tuned were `mtry`, `min.node.size` and `split.rule` using grid search.

```
hyper_grid_expanded <- expand.grid(
  mtry = seq(1, 26, by=1),
  nodesize = seq(1, 20, by = 1),
  splitrule = c("variance", "extratrees", "maxstat")
)

for (i in 1:nrow(hyper_grid_expanded)) {
  price.bag <- ranger(price ~ bathrm + hf_bathrm + heat + ac + rooms + bedrm
                      + stories + saledate + gba + style + grade + cndtn + extwall +
                        roof + intwall + kitchens + fireplaces + landarea + latitude +
                        longitude + nbhd + ward + quadrant + yr_since_rmdl + yr_since_imprv
```

```
                      + yr_since_b,
                      data = dtrain, mtry = hyper_grid_expanded$mtry[i],
                      min.node.size=hyper_grid_expanded$nodesize[i]
                      , splitrule = hyper_grid_expanded$splitrule[i], seed=20823986)
  hyper_grid_expanded$rsq[i] <- price.bag$r.squared
}
```

After tuning, the model with the highest value for R^2 was selected.

```
hyper_grid_expanded %>%
  arrange(rsq) %>%
  tail(10)
```

```
##      mtry nodesize splitrule       rsq
## 1551   14        6  variance 0.9381632
## 1552   16        5  variance 0.9381636
## 1553   13        6  variance 0.9382327
## 1554   16        3  variance 0.9382454
## 1555   16        2  variance 0.9382569
## 1556   11        3  variance 0.9382727
## 1557   13        4  variance 0.9382950
## 1558   13        3  variance 0.9384243
## 1559   13        2  variance 0.9384531
## 1560   13        1  variance 0.9384718
```