Appendix 1 - King County house price prediction

## Part 2 - Decision Tree Regression Model

### Step 1 : Collecting Data

Same as of Part 1 : Step 1

### Step 2 : Exploring, preprocessing and cleaning the data

Primary setup

knitr::opts\_knit$set(root.dir = '/Users/sobil/Documents/MSC/Sem 1/Data Mining & Machine Learning/Project/Kc\_house/')  
remove(list = ls())  
set.seed(1)

adding alll the libraries required

library(RWeka)  
library(rpart)  
library(rpart.plot)  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

Rest is same as of Part 1 : Step 2

source("Kc\_Import\_Explore\_Clean.R")

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

## '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 : chr "20141013T000000" "20141209T000000" "20150225T000000" "20141209T000000" ...  
## $ 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 ...

### Step 3 - Data transformation & preparation

creating training and testing dataset from exisitng sample

indx <- createDataPartition(kc\_house$price, p = 0.8, list = FALSE)  
  
house\_train <- kc\_house[indx,]  
house\_test <- kc\_house[- indx,]

### Step 4 : Training a model

modeling a regression tree

house.rpart <- rpart(price ~ ., data = house\_train)

### Step 5 : Evaluating the model

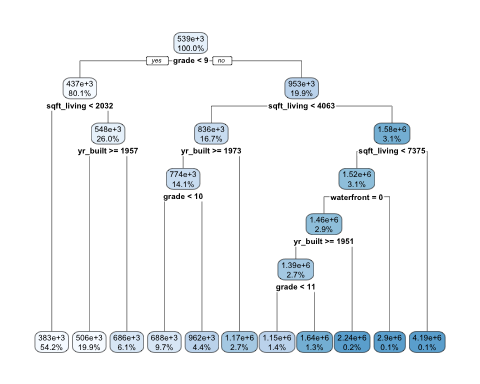
Summary of the model and description

#house.rpart  
summary(house.rpart)

## Call:  
## rpart(formula = price ~ ., data = house\_train)  
## n= 17291   
##   
## CP nsplit rel error xerror xstd  
## 1 0.32282388 0 1.0000000 1.0000754 0.04302218  
## 2 0.11069233 1 0.6771761 0.6772682 0.03386504  
## 3 0.03693204 2 0.5664838 0.5715869 0.02547574  
## 4 0.03632547 3 0.5295517 0.5720321 0.02594972  
## 5 0.02628165 4 0.4932263 0.5331213 0.02476824  
## 6 0.01922944 5 0.4669446 0.4885913 0.01712738  
## 7 0.01733461 6 0.4477152 0.4749996 0.01683980  
## 8 0.01179755 7 0.4303806 0.4546777 0.01584424  
## 9 0.01150364 9 0.4067855 0.4469484 0.01577295  
## 10 0.01000000 10 0.3952818 0.4319201 0.01521171  
##   
## Variable importance  
## grade sqft\_living sqft\_above bathrooms yr\_built   
## 33 26 20 9 4   
## sqft\_basement waterfront bedrooms view yr\_renovated   
## 4 2 1 1 1   
##   
## Node number 1: 17291 observations, complexity param=0.3228239  
## mean=539193.4, MSE=1.312817e+11   
## left son=2 (13855 obs) right son=3 (3436 obs)  
## Primary splits:  
## grade < 8.5 to the left, improve=0.3228239, (0 missing)  
## sqft\_living < 3087.5 to the left, improve=0.3101538, (0 missing)  
## sqft\_above < 2829 to the left, improve=0.2211162, (0 missing)  
## bathrooms < 3.125 to the left, improve=0.2104690, (0 missing)  
## view < 0.5 to the left, improve=0.1290508, (0 missing)  
## Surrogate splits:  
## sqft\_above < 2495.5 to the left, agree=0.885, adj=0.420, (0 split)  
## sqft\_living < 2915 to the left, agree=0.880, adj=0.394, (0 split)  
## bathrooms < 3.125 to the left, agree=0.838, adj=0.184, (0 split)  
## sqft\_basement < 1615 to the left, agree=0.807, adj=0.031, (0 split)  
## view < 2.5 to the left, agree=0.807, adj=0.027, (0 split)  
##   
## Node number 2: 13855 observations, complexity param=0.03632547  
## mean=436673.5, MSE=3.827863e+10   
## left son=4 (9365 obs) right son=5 (4490 obs)  
## Primary splits:  
## sqft\_living < 2032 to the left, improve=0.15547930, (0 missing)  
## grade < 7.5 to the left, improve=0.15026070, (0 missing)  
## sqft\_above < 1417 to the left, improve=0.08633045, (0 missing)  
## view < 0.5 to the left, improve=0.07755471, (0 missing)  
## bathrooms < 1.625 to the left, improve=0.06831389, (0 missing)  
## Surrogate splits:  
## sqft\_above < 2035 to the left, agree=0.845, adj=0.523, (0 split)  
## sqft\_basement < 725 to the left, agree=0.770, adj=0.292, (0 split)  
## bedrooms < 3.5 to the left, agree=0.769, adj=0.286, (0 split)  
## bathrooms < 2.375 to the left, agree=0.760, adj=0.259, (0 split)  
## grade < 7.5 to the left, agree=0.737, adj=0.188, (0 split)  
##   
## Node number 3: 3436 observations, complexity param=0.1106923  
## mean=952584.9, MSE=2.93025e+11   
## left son=6 (2896 obs) right son=7 (540 obs)  
## Primary splits:  
## sqft\_living < 4062.5 to the left, improve=0.2495652, (0 missing)  
## grade < 10.5 to the left, improve=0.2215809, (0 missing)  
## bathrooms < 3.625 to the left, improve=0.1625699, (0 missing)  
## sqft\_above < 4165 to the left, improve=0.1616369, (0 missing)  
## view < 0.5 to the left, improve=0.1142973, (0 missing)  
## Surrogate splits:  
## sqft\_above < 4062.5 to the left, agree=0.925, adj=0.524, (0 split)  
## grade < 10.5 to the left, agree=0.884, adj=0.265, (0 split)  
## bathrooms < 3.875 to the left, agree=0.883, adj=0.256, (0 split)  
## sqft\_basement < 1585 to the left, agree=0.862, adj=0.124, (0 split)  
## bedrooms < 5.5 to the left, agree=0.846, adj=0.020, (0 split)  
##   
## Node number 4: 9365 observations  
## mean=383255.9, MSE=2.356423e+10   
##   
## Node number 5: 4490 observations, complexity param=0.01150364  
## mean=548088.9, MSE=5.060422e+10   
## left son=10 (3440 obs) right son=11 (1050 obs)  
## Primary splits:  
## yr\_built < 1956.5 to the right, improve=0.11492820, (0 missing)  
## view < 0.5 to the left, improve=0.08605582, (0 missing)  
## waterfront splits as LR, improve=0.08538260, (0 missing)  
## sqft\_living < 2557.5 to the left, improve=0.06326877, (0 missing)  
## grade < 7.5 to the left, improve=0.05369639, (0 missing)  
## Surrogate splits:  
## yr\_renovated < 972.5 to the left, agree=0.797, adj=0.130, (0 split)  
## bathrooms < 1.625 to the right, agree=0.781, adj=0.064, (0 split)  
## condition < 4.5 to the left, agree=0.775, adj=0.037, (0 split)  
## grade < 6.5 to the right, agree=0.773, adj=0.030, (0 split)  
## sqft\_above < 1095 to the right, agree=0.768, adj=0.007, (0 split)  
##   
## Node number 6: 2896 observations, complexity param=0.02628165  
## mean=835812.1, MSE=1.265127e+11   
## left son=12 (2437 obs) right son=13 (459 obs)  
## Primary splits:  
## yr\_built < 1972.5 to the right, improve=0.16283370, (0 missing)  
## sqft\_living < 3155 to the left, improve=0.11520040, (0 missing)  
## grade < 9.5 to the left, improve=0.11088580, (0 missing)  
## view < 2.5 to the left, improve=0.09847294, (0 missing)  
## waterfront splits as LR, improve=0.09561904, (0 missing)  
## Surrogate splits:  
## yr\_renovated < 978 to the left, agree=0.877, adj=0.224, (0 split)  
## condition < 4.5 to the left, agree=0.864, adj=0.142, (0 split)  
## bedrooms < 5.5 to the left, agree=0.844, adj=0.015, (0 split)  
## bathrooms < 1.875 to the right, agree=0.844, adj=0.013, (0 split)  
## sqft\_basement < 1920 to the left, agree=0.843, adj=0.011, (0 split)  
##   
## Node number 7: 540 observations, complexity param=0.03693204  
## mean=1578833, MSE=7.207079e+11   
## left son=14 (528 obs) right son=15 (12 obs)  
## Primary splits:  
## sqft\_living < 7375 to the left, improve=0.2154143, (0 missing)  
## waterfront splits as LR, improve=0.1584211, (0 missing)  
## view < 3.5 to the left, improve=0.1430094, (0 missing)  
## sqft\_above < 6115 to the left, improve=0.1417749, (0 missing)  
## grade < 11.5 to the left, improve=0.1282572, (0 missing)  
## Surrogate splits:  
## sqft\_above < 6625 to the left, agree=0.989, adj=0.500, (0 split)  
## bathrooms < 6.125 to the left, agree=0.983, adj=0.250, (0 split)  
## sqft\_basement < 3815 to the left, agree=0.981, adj=0.167, (0 split)  
## grade < 12.5 to the left, agree=0.980, adj=0.083, (0 split)  
##   
## Node number 10: 3440 observations  
## mean=505955.9, MSE=3.61513e+10   
##   
## Node number 11: 1050 observations  
## mean=686124.5, MSE=7.308504e+10   
##   
## Node number 12: 2437 observations, complexity param=0.01733461  
## mean=773522.1, MSE=9.185468e+10   
## left son=24 (1675 obs) right son=25 (762 obs)  
## Primary splits:  
## grade < 9.5 to the left, improve=0.17578480, (0 missing)  
## sqft\_living < 3155 to the left, improve=0.14302750, (0 missing)  
## bathrooms < 3.125 to the left, improve=0.11465870, (0 missing)  
## view < 3.5 to the left, improve=0.09042453, (0 missing)  
## waterfront splits as LR, improve=0.08946337, (0 missing)  
## Surrogate splits:  
## sqft\_above < 3325 to the left, agree=0.734, adj=0.148, (0 split)  
## sqft\_living < 3425 to the left, agree=0.732, adj=0.144, (0 split)  
## bathrooms < 3.625 to the left, agree=0.699, adj=0.038, (0 split)  
## view < 3.5 to the left, agree=0.693, adj=0.018, (0 split)  
## waterfront splits as LR, agree=0.690, adj=0.009, (0 split)  
##   
## Node number 13: 459 observations  
## mean=1166532, MSE=1.805485e+11   
##   
## Node number 14: 528 observations, complexity param=0.01922944  
## mean=1519433, MSE=5.177857e+11   
## left son=28 (506 obs) right son=29 (22 obs)  
## Primary splits:  
## waterfront splits as LR, improve=0.1596640, (0 missing)  
## view < 3.5 to the left, improve=0.1358158, (0 missing)  
## grade < 10.5 to the left, improve=0.1061851, (0 missing)  
## yr\_built < 1944.5 to the right, improve=0.1057782, (0 missing)  
## sqft\_living < 5015 to the left, improve=0.0856600, (0 missing)  
##   
## Node number 15: 12 observations  
## mean=4192458, MSE=2.663e+12   
##   
## Node number 24: 1675 observations  
## mean=687816.1, MSE=4.943799e+10   
##   
## Node number 25: 762 observations  
## mean=961918, MSE=1.334538e+11   
##   
## Node number 28: 506 observations, complexity param=0.01179755  
## mean=1459479, MSE=4.272531e+11   
## left son=56 (467 obs) right son=57 (39 obs)  
## Primary splits:  
## yr\_built < 1950.5 to the right, improve=0.11954010, (0 missing)  
## grade < 10.5 to the left, improve=0.11449980, (0 missing)  
## sqft\_living < 5015 to the left, improve=0.07874945, (0 missing)  
## sqft\_lot < 35070.5 to the right, improve=0.05617333, (0 missing)  
## bathrooms < 4.625 to the left, improve=0.04976594, (0 missing)  
##   
## Node number 29: 22 observations  
## mean=2898364, MSE=6.159149e+11   
##   
## Node number 56: 467 observations, complexity param=0.01179755  
## mean=1394170, MSE=3.803135e+11   
## left son=112 (236 obs) right son=113 (231 obs)  
## Primary splits:  
## grade < 10.5 to the left, improve=0.15606020, (0 missing)  
## sqft\_living < 5015 to the left, improve=0.08001487, (0 missing)  
## bathrooms < 4.625 to the left, improve=0.05171274, (0 missing)  
## sqft\_above < 4227.5 to the left, improve=0.05108456, (0 missing)  
## view < 0.5 to the left, improve=0.04756573, (0 missing)  
## Surrogate splits:  
## sqft\_above < 4255 to the left, agree=0.664, adj=0.320, (0 split)  
## sqft\_living < 4655 to the left, agree=0.640, adj=0.273, (0 split)  
## sqft\_lot < 11499.5 to the left, agree=0.612, adj=0.216, (0 split)  
## yr\_built < 1987.5 to the left, agree=0.591, adj=0.173, (0 split)  
## sqft\_basement < 890 to the right, agree=0.587, adj=0.165, (0 split)  
##   
## Node number 57: 39 observations  
## mean=2241513, MSE=3.266734e+11   
##   
## Node number 112: 236 observations  
## mean=1153143, MSE=2.157654e+11   
##   
## Node number 113: 231 observations  
## mean=1640415, MSE=4.284349e+11

visulazing decision tree

rpart.plot(house.rpart, digits = 3)



Evaluating the model on testing data

house.predict.rpart <- predict(house.rpart, house\_test)  
summary(house.predict.rpart)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 383256 383256 383256 542083 686124 4192458

summary(house\_test$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 82500 322200 450000 543646 645000 7700000

checking correlation

cor(house.predict.rpart, house\_test$price)

## [1] 0.7713274

cheking performace with MAE creating a function mae to calculate Mean Absolute error

mae <- function(actual, pred) {  
 mean(abs(actual-pred))  
}

MAE with respect to predicted values

mae(house\_test$price, house.predict.rpart)

## [1] 158941.5

MAE with respect to mean values

mae(mean(house\_train$price), house.predict.rpart)

## [1] 187045.4

RMSE - Root Mean Squared Error - log values to compare to multi-linear model

RMSE(log(house\_test$price), log(house.predict.rpart))

## [1] 0.3780299

### Step 6 : Improving the model

modeling a model tree

house.m5p <- M5P(price ~ . , data = house\_train)  
#house.m5p  
summary(house.m5p)

##   
## === Summary ===  
##   
## Correlation coefficient 0.6249  
## Mean absolute error 8705723.0228  
## Root mean squared error 25833978.3661  
## Relative absolute error 3735.2195 %  
## Root relative squared error 7129.9947 %  
## Total Number of Instances 17291

house.predict.m5p <- predict(house.m5p, house\_test)  
summary(house.predict.m5p)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 241503 639213 798946 9295952 1286236 283228764

cor(house.predict.m5p, house\_test$price)

## [1] 0.6551532

mae(house\_test$price, house.predict.m5p)

## [1] 8756021

RMSE(log(house\_test$price), log(house.predict.m5p))

## [1] 1.794102