UNIVERSITÀ DEGLI STUDI DI MILANO

LA STATALE

Department of Economics, Management and Quantitative Methods

Università degli Studi di Milano

Data science and Economics

STATISTICAL LEARNING PROJECT

CALIFORNIA HOUSING PRICE PREDICTION

REPORT

Author:

MURAT AYDIN

Academic year 2020-2021

ABSTRACT

Immovable properties have been seen as a shield against inflation by many people. Today, all around the world millions of people using them as an invesment tool. A report done by MSCI says that t*he size of the professionally managed global real estate investment market increased from $8.9 trillion in 2018 to $9.6 trillion in 2019.* Housing prices prediction are therefore an important reflection of many world economies and are also of great interest for both buyers and sellers and for many other sub-sectors. It is of a high importance for prospective homeowner because they do not want to make a bad investment. It is important for an real estate agency because they want to run their business efficiently. It is also important to banks, mortgage lenders and insurers. It is eventually an economic decision that should be made carefully through the eyes of homo-economicus. Traditional house price prediction is based on cost and sale price comparison lacking of an accepted standard and a certification process. Therefore, the availability of a house price prediction model helps filling up an important information gap and improve the efficiency of the real estate market and many other sub-sectors related to it. The objective of this paper is to empirically compare the predictive power of various regression and tree-based models on house price prediction with its perfomance metric being Mean Squared Error. Among the regression models used, the best one turned out to be a second-degree polynomial regression while with the tree based methods, we got the lowest MSE with Random Forest.

1.DATASET

Our dataset is composed of 10 features with over 20.000 examples containing information about houses in California. Factors like location, median house age, proximity to the ocean, median income, physical condition of house are considered. The describtion of features on kaggle as follows:

We have 20640 observations with 10 attributes. The description of each of these attributes are given on kaggle as follows:

* longitude: A measure of how far west a house is; a higher value is farther west

* latitude: A measure of how far north a house is; a higher value is farther north

* housing_median_age: Median age of a house within a block; a lower number is a newer building

* total_rooms: Total number of rooms within a block

* total_bedrooms: Total number of bedrooms within a block

* population: Total number of people residing within a block

* households: Total number of households, a group of people residing within a home unit, for a block

* median_income: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

* median_house_value: Median house value for households within a block (measured in US Dollars)

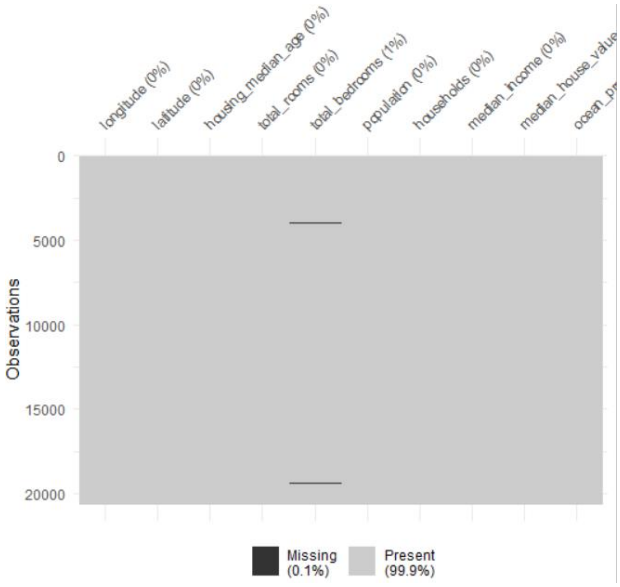
* ocean_proximity: Location of the house w.r.t ocean/sea

* Among these, “median_house_value” is the response variable and others will be the covariates.

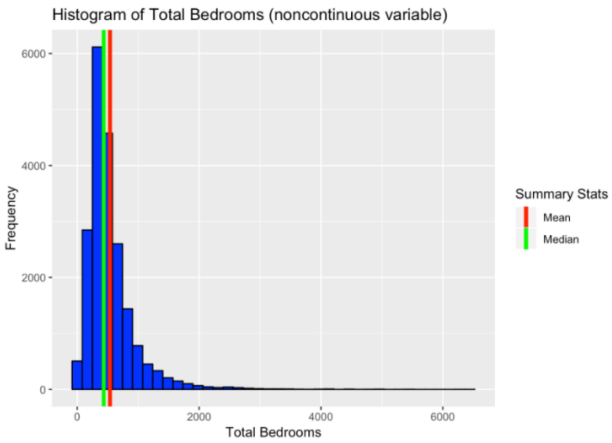
2. PRE-PROCESSING AND EXPLORATORY ANALYSIS

At this part of the analysis, we will deal with pre-processing steps and study the missing values, multicollinearity and scale the data to make the results more interpretable. We scaled it because the scaled data gives a MSE that is always lower than 1 so the comparison and interpretation is easier. Note that the data has been scaled after the exploratory analysis.

2.1 NAs Handling



Approximately 1% of the whole dataset is missing. To be precise, there are 207 missing values all belonging to the total_bedrooms variable.

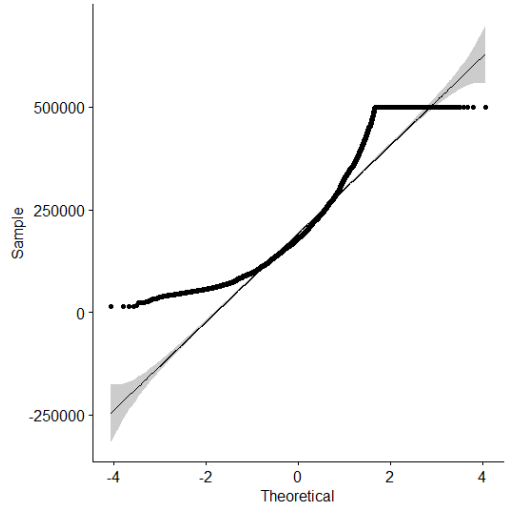


The distribution of the Total_bedrooms has a positive skew. The mean is biased by the values at the far end of the distribution. Therefore, the median is a better representation for the majority of values in the variable. The missing values are imputed by the median of the distribution.

```
  ocean_proximity
<1H OCEAN :9136
INLAND    :6551
ISLAND    :   5
NEAR BAY  :2290
NEAR OCEAN:2658
```

Another point that we want to mention here about the ocean proximity. Ocean proximity is the only categorical variable with five levels. Most of the houses are located INLAND. Only 5 of them is ISLAND. So, 5 data points in a set of 20.000 observation is not important. We will thus simply drop them

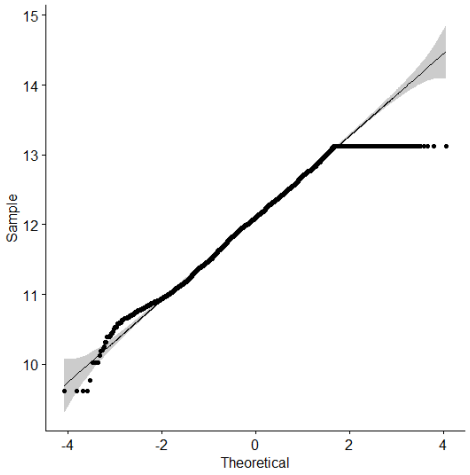
2.2 Normality



We know that one of the assumptions of linear regression is that the response variable should have a normal distribution. This is a Q-Q plot where we have theoretical and sample quantiles and we can see that our response variable is not normally distributed. Overall, we have problems here but especially in the lower and upper tails.

statistic	p.value	method
501.7764	0.00000000000000000000000037	Anderson-Darling normality test

Anderson-Darling normality test empirically shows us that that the response variable is not normal. However, since our dataset is large enough, we can assume that it is asymptotically normal.



With log transformation, we actually highly recover the normally assumption however we have still a big problem on the upper tail which corresponds houses having the highest price.

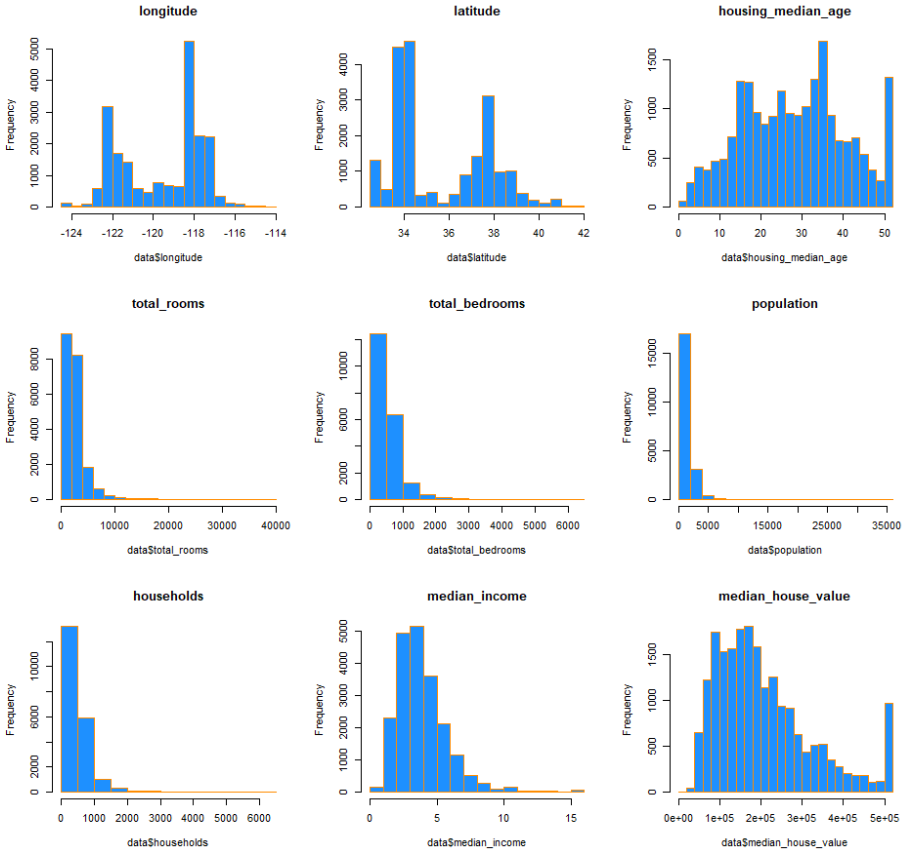
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
14999  119600  179700  206814  264700  500001
```

As we can see, the max value for median house price is $500.001 and we have 965 houses having this value. So, they are not really outliers, they correspond to almost %10 of the dataset.

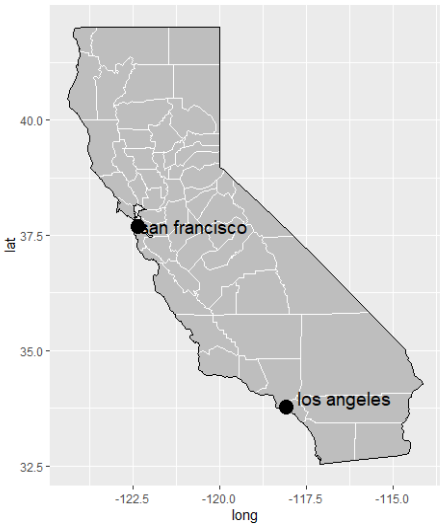
The statistics value highly decreased while P value basically remained zero. We can empirically still say that our response variable is not normal.

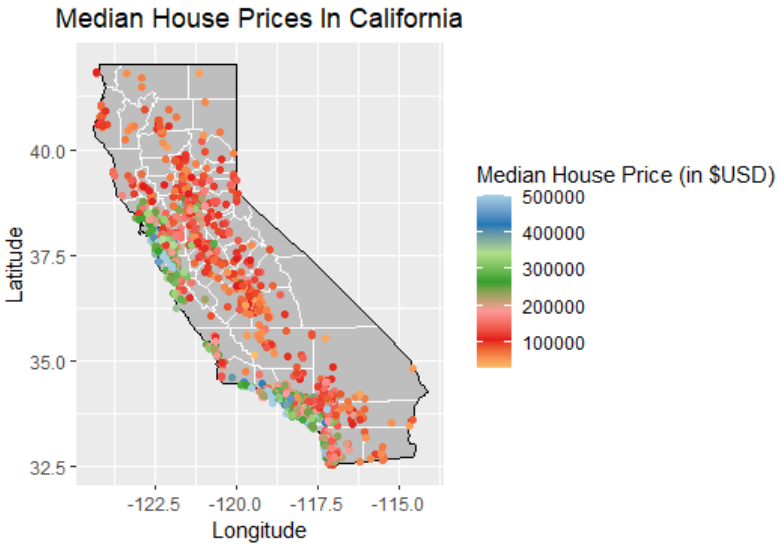
statistic	p.value	method
30.91679	0.0000000000000000000000037	Anderson-Darling normality test

2.3 Distribution of Features



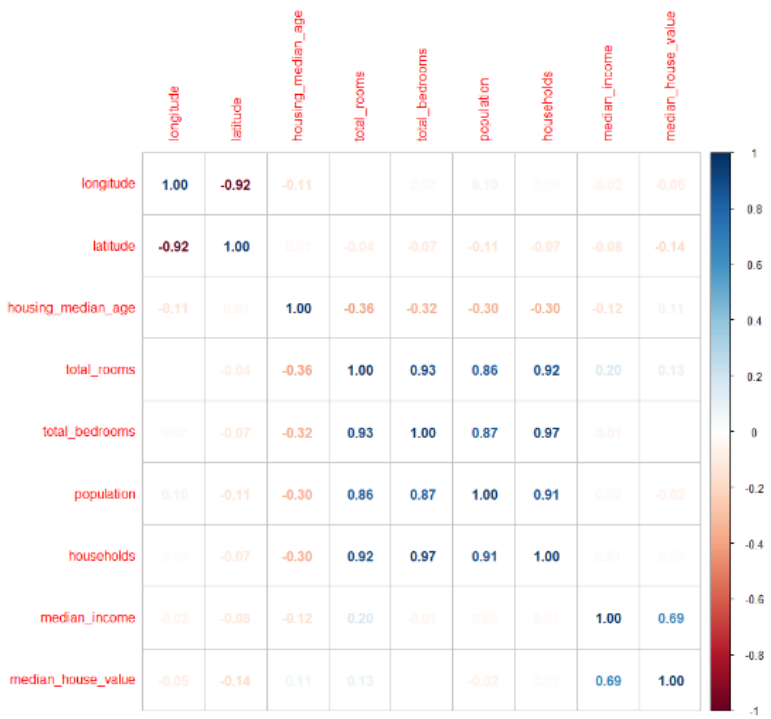
- Most of the houses are located around longitude -118 and around latitude 34 which corresponds to Los Angeles which is located near the ocean. Another dense location is San Francisco that is around longitude -123 and latitude 37. Since prices are higher in these locations, we expect a good linear predictor to have its projection aligned with those places.





As we can see, the most expensive houses are located in San Francisco and Los Angeles which implies that Ocean proximity variable will play an important role in predicting the prices. On the other hand, the geographical shape of California will probably cause a multicollinearity because as Latitude goes up, Longitude decreases so we expect a negative correlation.

3. MULTICOLLINEARITY AND VARIABLE SELECTION

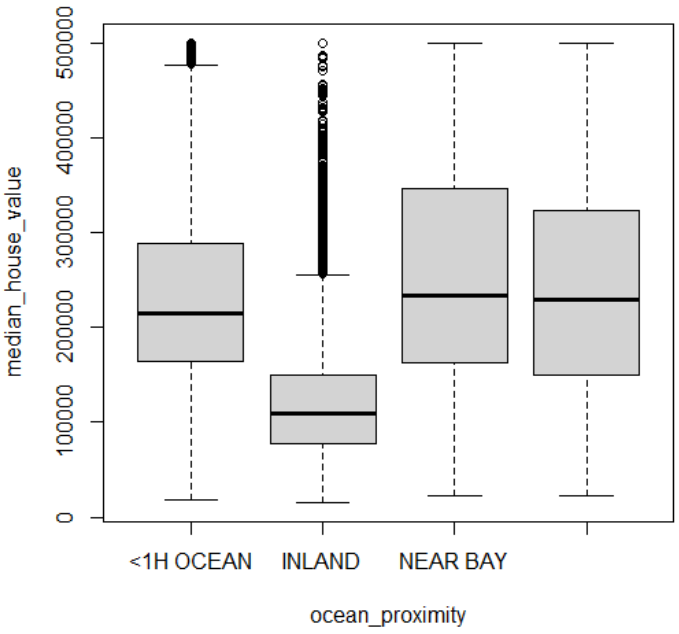


High correlation is present among some variables. In particular, total rooms and total bedrooms are highly correlated as expected. Households (number of people living in the house) is highly correlated with total rooms and bedrooms as well. As expected, we have a high negative correlation between longitude and latitude. Longitude and latitude should be highly important here since they imply the location of the house. That is why we cannot drop them. As for the features about the number of rooms, the reason of the correlation is clear. As number of rooms increases, there is a high chance that the number of bedrooms will increase as well. However, for a house having rooms and bedrooms have different implications on the price. It is true that they cause multicollinearity but we know that they imply different things and having higher number of bedrooms or normal rooms might have different effects on the price.

row	column	cor	p
median_house_value	longitude	-0.04620776	0.000000000031190606
median_house_value	latitude	-0.14383687	0.000000000000000000
median_house_value	housing_median_age	0.10527182	0.000000000000000000
median_house_value	total_rooms	0.13437325	0.000000000000000000
median_house_value	total_bedrooms	0.04956145	0.000000000001052936
median_house_value	population	-0.02442076	0.000450936770433064
median_house_value	households	0.06606909	0.000000000000000000
median_house_value	median_income	0.68856266	0.000000000000000000

The highest correlation between the response variable and the features come from the median income with a p-value being 0. Therefore, median income will be our starting point in estimating the response variable among many other models.

We now want to check in a descriptive way if median house value differs with respect to the ocean proximity which we will do through an ANOVA test because we want to check the hypothesis of all the means are equal in a given group. Since ocean proximity has four levels, we can not apply t-test.



The box plot here shows the distribution of house prices with respect to the ocean proximity. As we can see INLAND prices differ relative to others. We expect to reject the null hypothesis because it is enough that only one pair is different.

ocean_proximity	mean	sd	n
<1H OCEAN	240084.3	106124.29	9136
INLAND	124805.4	70007.91	6551
NEAR BAY	259212.3	122818.54	2290
NEAR OCEAN	249434.0	122477.15	2658

This is the starting point for the ANOVA. The conditional means, standard deviations and number of units.Note that houses apparently are more expensive in Near Ocean locations however it has the highest standard deviation. Standard deviation of inland is way lower than the others and the houses there are way cheaper. We expect that being located in INLAND makes a difference with respect to the median house price.

```
                  Df          Sum Sq       Mean Sq F value              Pr(>F)
ocean_proximity    3  65286462190192 21762154063397    2144 <0.0000000000000002 ***
Residuals      20631 209368866526315    10148265548
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- In ANOVA, we have within variance, between variance and sum of squares.
- DF is 3 because we have four groups, K-1.
- For within variance, DF is N-K that is 20635-4.
- The F stats which ranges from 0 to infinity is given by the ratio of mean squares of between and mean squares of within and it is highly large here.
- P value is basically zero, so we reject the hypothesis that the means are equal as we discussed above. However, note that the fact that we reject the null hypothesis does not mean that all the groups differ, it is enough if only one of them differs.

4.RESULTS

We first applied simple linear regression with the most important variables that is median income and ocean proximity.

```
Call:
lm(formula = median_house_value ~ median_income + ocean_proximity,
    data = scaled_data[train, ])

Residuals:
    Min      1Q  Median      3Q     Max
-3.0883 -0.4078 -0.1133  0.2521  4.1111

Coefficients:
                           Estimate Std. Error t value             Pr(>|t|)
(Intercept)                0.172457   0.009600  17.964 < 0.0000000000000002 ***
median_income              0.606174   0.006516  93.022 < 0.0000000000000002 ***
ocean_proximityINLAND     -0.666865   0.015119 -44.108 < 0.0000000000000002 ***
ocean_proximityNEAR BAY    0.164704   0.020938   7.866  0.00000000000000402 ***
ocean_proximityNEAR OCEAN  0.164512   0.019989   8.230 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6446 on 10418 degrees of freedom
Multiple R-squared:  0.5807,    Adjusted R-squared:  0.5806
F-statistic:  3608 on 4 and 10418 DF,  p-value: < 0.00000000000000022
```

All the variables selected are highly significant and they have the expected signs. Having a high median income increases median house prices. Being closer to the Ocean also has a positive effect. As we saw in the exploratory analysis above, INLAND houses should have lower prices and in fact it has a negative sign. The interpretation of dummies should be with respect to the baseline. Here the baseline is 1H OCEAN automatically chosen by the model. So, being located INLAND with respect to the houses that are in 1H OCAN, NEAR BAY and NEAR OCEAN decreases the prices. R square is around %58, however since we will do comparisons between models, we should use Adjusted R square that penalizes the model for each variable added which is almost %58 in this case. The mean squared test error of this model is 0.4089.

4.1 Model Selection

Here, we will try to identify a subset of P predictors that we believe to be related to the response. We will then fit a model using least squares on the reduced set of variables.

4.1.1 Step-Wise Selection

Here we used both forward and backward approach and the result turned out to be the same. Both methods choose the full model where all the variables are included. All of the variables are highly significant at %1 level expect the NEAR BAY. Positive values of longitude and latitude seems to effect negatively prices. That makes sense, recall the map of California above. As longitude increases, we move to the INLAND. As latitude increases, we move away from two big important cities San Francisco and Los Angeles.

```
Call:
lm(formula = median_house_value ~ longitude + latitude + housing_median_age +
    total_rooms + total_bedrooms + population + households +
    median_income + ocean_proximity, data = scaled_data[train,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-3.2779 -0.3747 -0.0908  0.2637  4.2944

Coefficients:
                          Estimate Std. Error t value             Pr(>|t|)
(Intercept)               0.101676   0.010231   9.938 < 0.0000000000000002 ***
longitude                -0.494888   0.024672 -20.059 < 0.0000000000000002 ***
latitude                 -0.511891   0.026012 -19.679 < 0.0000000000000002 ***
housing_median_age        0.113355   0.006733  16.835 < 0.0000000000000002 ***
total_rooms              -0.048352   0.020835  -2.321               0.0203 *
total_bedrooms            0.234609   0.033035   7.102  0.00000000000131133 ***
population               -0.439099   0.016193 -27.116 < 0.0000000000000002 ***
households                0.276248   0.034160   8.087  0.00000000000000068 ***
median_income             0.623560   0.007711  80.868 < 0.0000000000000002 ***
ocean_proximityINLAND    -0.316895   0.021259 -14.906 < 0.0000000000000002 ***
ocean_proximityNEAR BAY  -0.046465   0.022991  -2.021               0.0433 *
ocean_proximityNEAR OCEAN  0.033150   0.019095   1.736               0.0826 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.597 on 10411 degrees of freedom
Multiple R-squared:  0.6406,    Adjusted R-squared:  0.6402
F-statistic:  1687 on 11 and 10411 DF,  p-value: < 0.00000000000000022
```

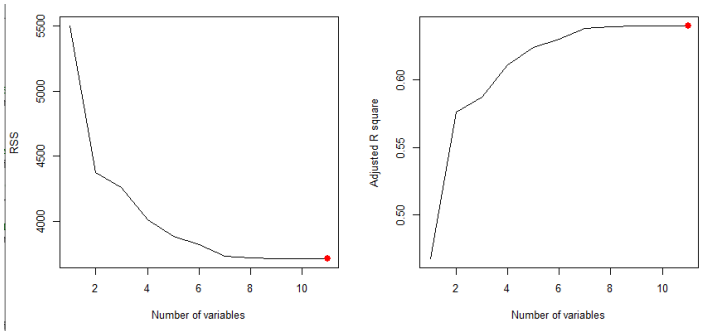
Here the only weird result comes from total rooms. It has a negative sign which does not make sense. Probably it is because it is correlated with total bedrooms. Adjusted R square increased to %64. The mean squared test error of this model is .35496, we highly recovered the accuracy of the model however lost a little bit of interpretability.

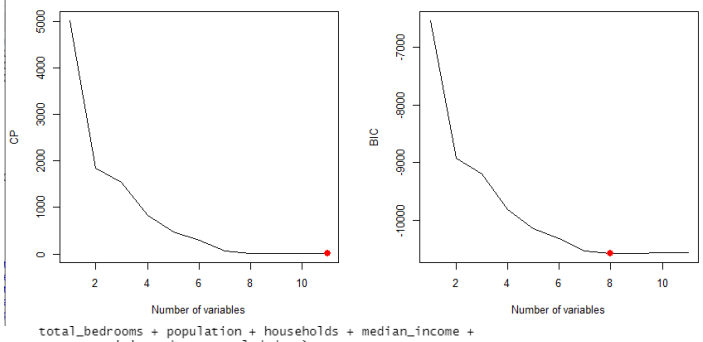
```
Call:
lm(formula = median_house_value ~ longitude + latitude + housing_median_age +
    population + households + median_income + ocean_proximity,
    data = no_bed)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2233 -0.3765 -0.0909  0.2622  4.3247

Coefficients:
                          Estimate Std. Error t value             Pr(>|t|)
(Intercept)               0.102594   0.010197  10.061 <0.0000000000000002 ***
longitude                -0.474102   0.024395 -19.435 <0.0000000000000002 ***
latitude                 -0.495128   0.025767 -19.216 <0.0000000000000002 ***
housing_median_age        0.110364   0.006717  16.431 <0.0000000000000002 ***
population               -0.464030   0.015457 -30.021 <0.0000000000000002 ***
households                0.482643   0.014974  32.233 <0.0000000000000002 ***
median_income             0.610411   0.006225  98.055 <0.0000000000000002 ***
ocean_proximityINLAND    -0.323942   0.021156 -15.312 <0.0000000000000002 ***
ocean_proximityNEAR BAY  -0.038847   0.022994  -1.689               0.0912 .
ocean_proximityNEAR OCEAN  0.037401   0.019085   1.960               0.0501 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5985 on 10413 degrees of freedom
Multiple R-squared:  0.6388,    Adjusted R-squared:  0.6385
F-statistic:  2046 on 9 and 10413 DF,  p-value: < 0.00000000000000022
```

We next tried the model by dropping total number of bedrooms since it is already included in the total number of rooms. This time, interestingly, Step-Wise approach did not choose the full model but it also dropped the total number of rooms.

So, the model does not use any information about the number of rooms. Adjusted R squared remained almost the same %63.8 and the test MSE of this model is .3518. We are now more accurate in terms of predictions and our model is more interpretable in the sense that it does not give a result that is incompatible with the reality as before. This model includes seven variables and we will see that this model turns out be the best one among other and so we will check its diagnostics. We call this model no_rooms model since we will mention it later.

4.1.2 Best-Subset Selection





Best subset selection, as Step-Wise Regression, goes for the full model. All of the indicators except BIC advises to use the full model. However, we can be smarter. With the plots provided here, we can see the trade-off clearly. Best two-coefficient model here, is income and INLAND. We expected it because we checked the correlation matrix and did an ANOVA test and saw that prices differ with INLAND. The RSS decreases dramatically until a five-variable model and then this decrease becomes almost stable at eight-variable model as BIC selects.

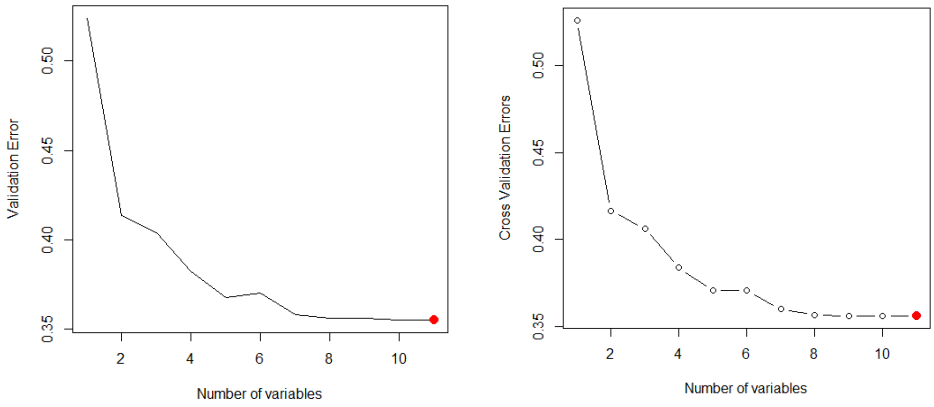
```
total_bedrooms + population + households + median_income +
ocean_proximity, data = scaled_data)

Coefficients:
                (Intercept)                   longitude                    latitude          housing_median_age
                    0.11236                    -0.46721                    -0.47641                     0.11710
                 population                  households               median_income          ocean_proximityINLAND
                    -0.40237                     0.25084                     0.61846                    -0.35584
                total_bedrooms          ocean_proximityNEAR BAY  ocean_proximityNEAR OCEAN
                     0.19808                    -0.03595                     0.03570
```

The best eight variable model is given by the formula here.

However, note that regsubsets function in R considers dummies as different variables so it only suggests to include the dummy INLAND, however, we know that we should include all of the dummies even if they are not significant. This model is similar to reduced Step-Wise model where we dropped total number of bedrooms. Here the only difference is that total number of rooms is dropped while bedrooms is kept. The test MSE of this model .3531. This model turned out to have a slightly lower accuracy (with a change in the third digit) with respect to reduced Step-Wise while keeping the same interpretability power. We now have a MSE almost as good as the full model with a more meaningful outcome.

4.1.3 VALIDATION SET AND CROSS-VALIDATION

Best subset selection, Step-wise selection result in the creation of a set of models, each of which contains a subset of the p predictors. In order to implement these methods, we need a way to determine which of these models is the best. We know that the model containing all of the predictors will always have the smallest RSS and the largest R square, since these quantities are related to the training error. Instead, we want to choose a model with a low test error. Therefore, RSS and R square is not the best thing we could use to select the best model. That is why here we will try both Validation set and Cross-Validation approach. We choose these two approaches over BIC, Adjusted R square and Cp because they provide a direct estimate of the test error.



As shown in the plots, both methods give the same result. The errors are minimized with the full model however we know that the full-model had problems in terms of interpretability. In particular, in the full model we had a negative sign for total number of rooms which is not meaningful. Instead, in the eight-variable model we have a MSE as good as the full model and it is more interpretable.

DIAGNOSTICS

Here we will check the diagnostics of the no_room model where we did not use any information about number of rooms. Recall that this model had only seven variables.

4.2 Multi-Collinearity

Since we have a good amount of predictors, the first thing we will do is to test multicollinearity. The way we do it is through Variance Inflation Factor that for each variable X calculates a linear model of all other X variables. It tries to express each single X as a linear combination of others.

	GVIF	Df	GVIF^(1/(2*Df))
longitude	4.18442254966404	1	2.0455861139693
latitude	4.39520064604139	1	2.09647338309872
housing_median_age	1.1440905839292	1	1.06962170131743
population	2.51323619267598	1	1.58531895613343
households	2.50703116087124	1	1.58336071723131
median_income	1.05894394361399	1	1.02905001997667
ocean_proximity	1.99976182381294	1.73205080756888	1.12243976855959

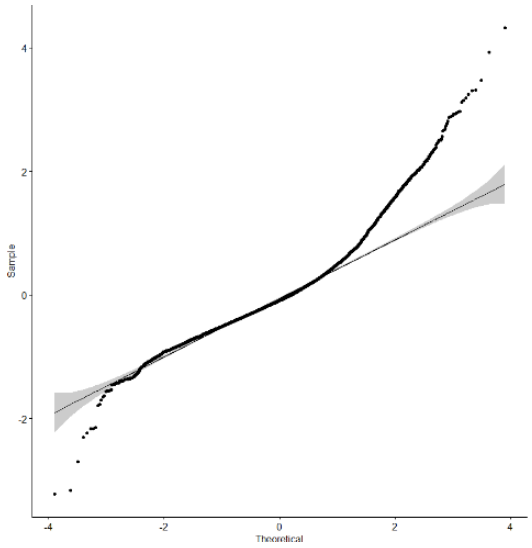
This table provides the square root VIF scores of the no_rooms model where we did not use any predictor about the number of rooms. Notice that values having a higher value than 2 is accepted to cause multicollinearity.

We already knew that our data has some multicollinearity. As we see here, longitude and latitude causes multicollinearity if we just ignore population and households and say that they are on the border. Number of rooms and bedrooms were causing it as well but we dropped the number of bedrooms and Step-wise regression dropped number of rooms so we somehow shrunk the problems of collinearity. However, we cannot drop longitude and latitude since they provide the location of the houses and are essential in pricing.

	GVIF	Df	GVIF^(1/(2*Df))
housing_median_age	1.13720727816546	1	1.06639921144263
population	2.50039055430974	1	1.58126232937794
households	2.50342698837323	1	1.58222216783018
median_income	1.04923680178437	1	1.02432260630349
ocean_proximity	1.11853967592211	1.73205080756888	1.01884604893593

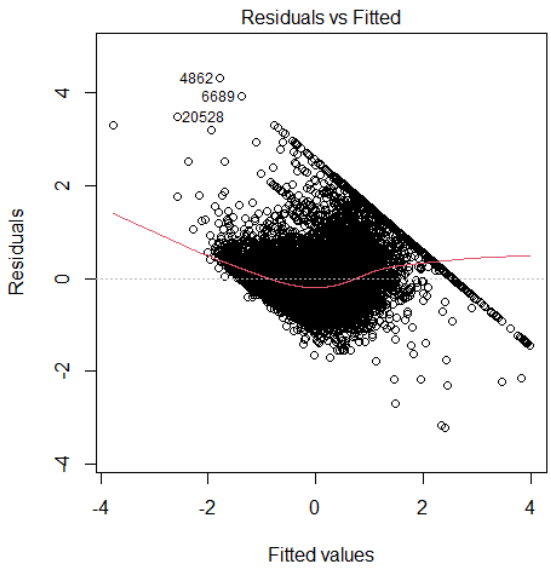
When we dropped those variables, the test MSE increased to .366 and we still have the problem with population and households. When we dropped them as well, we have the first model where we used only two

predictors chosen through correlation matrix and ANOVA test which had substantially higher test MSE error than this model. Recall that first model had a MSE .4089 while no_rooms model had .3518.




```
        Anderson-Darling normality test

data:  mod.step_no_bed$residuals
A = 164.66, p-value < 0.00000000000000022
```

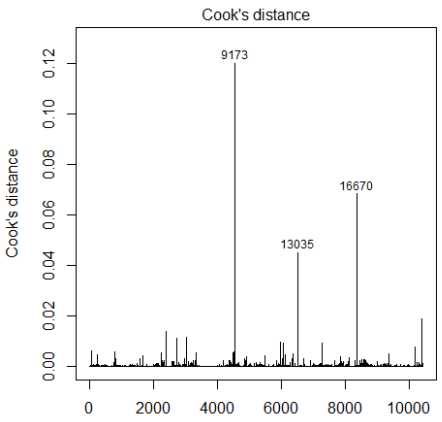
The plot is about the residuals. They are not normal as proven by the test. We have a really big problem in the upper tail. So, one of the most important assumptions of linear regression is not met. Probably, this is not a linear problem. We should probably go for a non-linear model.

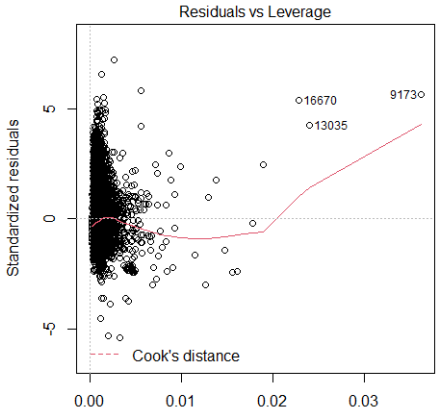


As we can see, the residuals vs Fitted plot is slightly non-linear in nature. The hard black line is probably because the observations that constantly had median house value $500.001. Recall that there were around 1000 observations having this value.

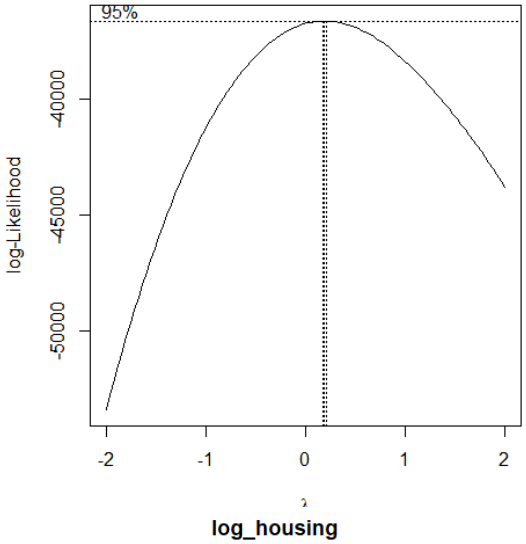
```
        studentized Breusch-Pagan test

data:  mod.step_no_bed
BP = 363.17, df = 9, p-value < 0.00000000000000022
```

WE also applied Breusch-Pagan test to check if the model is homoscedastic. The null hypothesis is that our model homoscedastic and we reject it. The P value is basically zero. So our model does not respect the assumptions of normality and homoskedasticity.

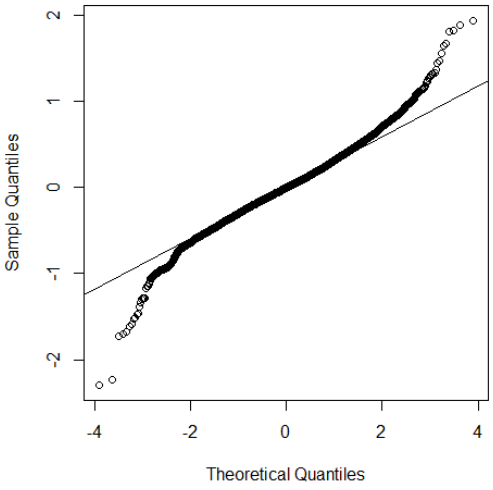


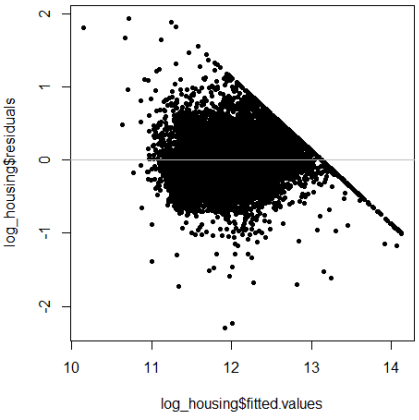


There are hardly any points outside the Cook's distance line. This implies that there are no influential outliers in the data.



Box-Cox transformation required positive values for the response variable. So here unscaled data used for this. The value of gamma is close to 0 indicating that a log transformation might be good for this model.





As we can see, there is no significant effect of the transformation on the distribution of residuals. We will try a polynomial model.

4.3 MODEL RE-SPECIFICATION

In performing a polynomial regression, we must decide on the degree of the polynomial to use. One way to do this is by using hypothesis tests which we check through an ANOVA test. ANOVA test suggested to use 7-degree polynomials which make the model almost impossible to interpret. On the other hand, they incredibly overfit the model and so they had really large test MSE sometimes reaching to over 100. Recall that we were jumping around 0.35.

As an alternative to using hypothesis tests, we choose to use CV and then decide about the degree. We did 10-fold Cross-Validation up to 5 degree on the scaled data set and got the following results through a prediction on the test set:

0.329041904992433	0.288198678846708	0.254616009774949	0.341689786214486	3.24560161977969
0.346966799236859	0.311577001765944	0.270613457435366	0.272220185076729	31.3934523212951
0.36746713343559	0.626233825395996	5.82811657413556	1446.71613155353	5752806.62171872
0.329440740820192	0.286951496179545	0.256421540070852	0.447182816764756	70.9378521908948
0.345020691415938	0.295633985993346	0.26364946408678	0.45864963699981	6.39258346657965
0.357467236725693	0.310290638109535	0.32811390840905	37.8656164247016	11668.4505310459
0.357385774346659	0.306775745908654	0.28788155160524	0.564008513821525	17.2876889188651
0.35833388427871	0.306627290162988	0.286148493089868	0.274378505838184	2.92303775592952
0.379659707536883	0.309809656539861	1.28512026616743	4.34631021160215	8403.37625831732
0.38642848494806	0.327884652102378	0.294491394738844	0.911973960407937	19.1486187595886

	First.Degree	Second.Degree	Third.Degree	Fourth.Degree	Fifth.Degree
1	0.355721235773702	0.336998297100495	0.935517265951394	149.219816159495	577302.977734312

These are column-wise means of CV errors. We can see that the smallest CV error is given by the second degree polynomial, the mean CV is .336 which is way better than the result provided by linear regression. Notice that in general CV errors in higher orders, in particular Third and Fourth degree is substantially smaller compared to the Second Degree. However, here we can see the power of CV approach. Once in a while, we see huge MSE errors in both terms. In particular, in the fourth degree it reaches to a max around 1446, in the fifth degree it reaches to 57528. This shows that higher degrees overfit the training data. The result of the polynomial regression is given on the next page.

```
Call:
lm(formula = median_house_value ~ (.)^2, data = scaled_data[train,
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-2.9457 -0.3220 -0.0785  0.2282  5.0369

Coefficients:
                                            Estimate Std. Error t value             Pr(>|t|)
(Intercept)                                0.0931162  0.0227254   4.097  0.00004208994738626 ***
longitude                                 -0.7178942  0.0629730 -11.400 < 0.0000000000000002 ***
latitude                                  -0.8151658  0.0739548 -11.022 < 0.0000000000000002 ***
housing_median_age                         0.0623151  0.0117382   5.309  0.00000011266033366 ***
total_rooms                               -0.0881432  0.0444176  -1.984             0.047235 *
total_bedrooms                             0.2562272  0.0731517   3.503             0.000463 ***
population                                -0.6452336  0.0286060 -22.556 < 0.0000000000000002 ***
households                                 0.5042492  0.0779118   6.472  0.00000000010105193 ***
median_income                              0.5594574  0.0158970  35.193 < 0.0000000000000002 ***
ocean_proximityINLAND                     -0.3090312  0.0266205 -11.609 < 0.0000000000000002 ***
ocean_proximityNEAR BAY                   -2.8919459  0.3269664  -8.845 < 0.0000000000000002 ***
ocean_proximityNEAR OCEAN                  0.0005875  0.0360795   0.016             0.987008
longitude:latitude                         0.0694980  0.0114140   6.089  0.00000000117778850 ***
longitude:housing_median_age              -0.2259394  0.0273374  -8.265 < 0.0000000000000002 ***
longitude:total_rooms                      0.4303749  0.1102544   3.903  0.00009542294723312 ***
longitude:total_bedrooms                  -0.2987024  0.1352303  -2.209             0.027207 *
longitude:population                       0.1174694  0.0624556   1.881             0.060021 .
longitude:households                      -0.1979603  0.1235672  -1.602             0.109176
longitude:median_income                   -0.3000587  0.0378303  -7.932  0.00000000000000238 ***
longitude:ocean_proximityINLAND            0.2163603  0.0724437   2.987             0.002828 **
longitude:ocean_proximityNEAR BAY         -4.0108020  0.2660664 -15.074 < 0.0000000000000002 ***
longitude:ocean_proximityNEAR OCEAN       -0.1946835  0.0921995  -2.112             0.034749 *
latitude:housing_median_age               -0.2768215  0.0286907  -9.648 < 0.0000000000000002 ***
latitude:total_rooms                       0.5431827  0.1196653   4.539  0.00000571087250675 ***
latitude:total_bedrooms                   -0.4408441  0.1447807  -3.045             0.002333 **
latitude:population                        0.2113793  0.0652082   3.242             0.001192 **
latitude:households                       -0.2615942  0.1233367  -2.121             0.033947 *
latitude:median_income                    -0.3697581  0.0416245  -8.883 < 0.0000000000000002 ***
latitude:ocean_proximityINLAND             0.2838266  0.0801794   3.540             0.000402 ***
latitude:ocean_proximityNEAR BAY          -2.4006658  0.2191378 -10.955 < 0.0000000000000002 ***
latitude:ocean_proximityNEAR OCEAN        -0.0762380  0.1026616  -0.743             0.457732
housing_median_age:total_rooms            -0.0780824  0.0251547  -3.104             0.001914 **
housing_median_age:total_bedrooms          0.0276397  0.0454660   0.608             0.543253
housing_median_age:population             -0.1745258  0.0166871 -10.459 < 0.0000000000000002 ***
housing_median_age:households              0.2636475  0.0477967   5.516  0.00000003550900218 ***
housing_median_age:median_income           0.0130146  0.0072802   1.788             0.073857 .
housing_median_age:ocean_proximityINLAND   0.0904877  0.0227361   3.980  0.00006940923142757 ***
housing_median_age:ocean_proximityNEAR BAY -0.0463737  0.0245191  -1.891             0.058609 .
housing_median_age:ocean_proximityNEAR OCEAN -0.0025817  0.0209786  -0.123             0.902059
```

Now, the model is way more difficult to interpret however the test MSE of this model is 0.297 which is substantially lower than all of the models we applied above. We gained a lot in terms of predictions but jumped into a very complex model. This model should be chosen if the aim is prediction however if one tends to do inferences or a simpler model then the previous model should be chosen.

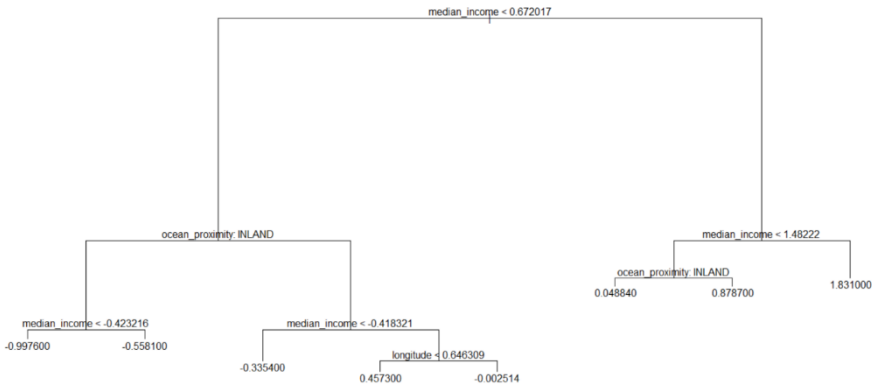
```
total_rooms:total_bedrooms                -0.0084002  0.0223251  -0.376             0.706725
total_rooms:population                    -0.0927972  0.0238044  -3.898  0.00009747323620663 ***
total_rooms:households                     0.0835214  0.0317799   2.628             0.008598 **
total_rooms:median_income                  0.0435281  0.0158481   2.747             0.006033 **
total_rooms:ocean_proximityINLAND         -0.3284469  0.0798032  -4.116  0.00003889828081613 ***
total_rooms:ocean_proximityNEAR BAY        0.2273567  0.0947839   2.399             0.016472 *
total_rooms:ocean_proximityNEAR OCEAN     -0.0407981  0.0673919  -0.605             0.544936
total_bedrooms:population                  0.0690817  0.0390238   1.770             0.076715 .
total_bedrooms:households                 -0.1106444  0.0195922  -5.647  0.00000001672134271 ***
total_bedrooms:median_income               0.0213901  0.0388331   0.551             0.581769
total_bedrooms:ocean_proximityINLAND       0.2099734  0.1261064   1.665             0.095933 .
total_bedrooms:ocean_proximityNEAR BAY     0.0806432  0.1652130   0.488             0.625478
total_bedrooms:ocean_proximityNEAR OCEAN   0.2383712  0.1300271   1.833             0.066795 .
population:households                      0.0603131  0.0332099   1.816             0.069381 .
population:median_income                  -0.0881980  0.0236580  -3.728             0.000194 ***
population:ocean_proximityINLAND           0.2840767  0.0522926   5.432  0.00000005684358447 ***
population:ocean_proximityNEAR BAY        -0.2263434  0.0776029  -2.917             0.003545 **
population:ocean_proximityNEAR OCEAN       0.2150246  0.0466107   4.613  0.00000401274629216 ***
households:median_income                   0.1095946  0.0446812   2.453             0.014191 *
households:ocean_proximityINLAND          -0.1344271  0.1207804  -1.113             0.265740
households:ocean_proximityNEAR BAY        -0.1234343  0.1819523  -0.678             0.497541
households:ocean_proximityNEAR OCEAN      -0.3763846  0.1374051  -2.739             0.006169 **
median_income:ocean_proximityINLAND        0.1573498  0.0294592   5.341  0.00000009425179101 ***
median_income:ocean_proximityNEAR BAY      0.0475396  0.0268998   1.767             0.077210 .
median_income:ocean_proximityNEAR OCEAN    0.0201519  0.0224832   0.896             0.370108
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5515 on 10359 degrees of freedom
Multiple R-squared:  0.6948,    Adjusted R-squared:  0.693
F-statistic: 374.4 on 63 and 10359 DF,  p-value: < 0.00000000000000022
```

5. DECISION TREES

Differently from regression, trees are non-parametric methods which can be used both for classification and regression purposes. From a geometrical point of view, a tree is a predictor that stratifies the predictors into hyperboxes. In order to make a prediction for a given observation, we typically use the mean or the mode of the training observations in the region to which it belongs. They are simpler compared to linear regression and are more interpretable. We know that linear regression was not the best idea for our dataset because the assumption of linearity was not met. In fact the best result came from a non-linear model. Trees are generally not comparable to linear regression when the relationship is well approximated by a linear model but in this case, we expect the tree methods to outperform the linear ones.

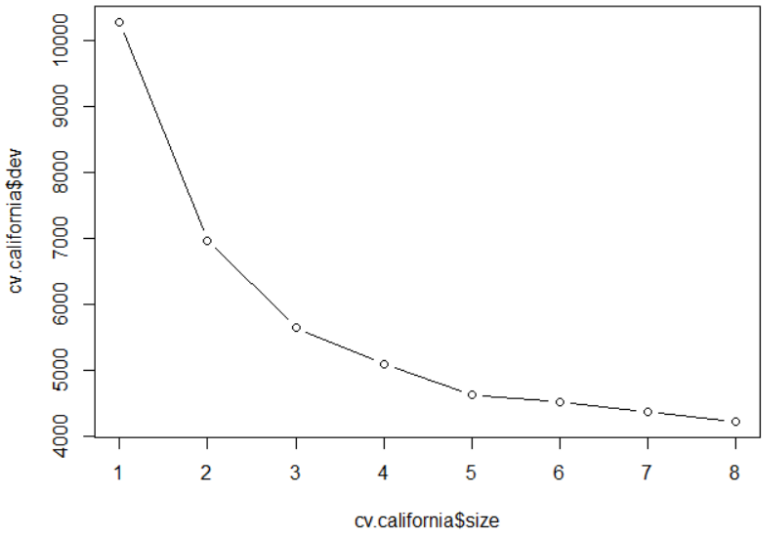
```
Regression tree:
tree(formula = median_house_value ~ ., data = scaled_data, subset = train)
Variables actually used in tree construction:
[1] "median_income"   "ocean_proximity" "longitude"
Number of terminal nodes:  8
Residual mean deviance:  0.4059 = 4184 / 10310
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.5210 -0.3893 -0.1041  0.0000  0.2888  3.5390
```

Notice that only three out of nine variables are used. In fact, the selected variables are in line with our previous analysis. We already stated that median income and ocean proximity were the most important variables in explaining the median house price but now we have also the longitude. Longitude differs in giving exact location with respect to ocean proximity, since locations are highly important in pricing houses, the idea behind the choice of longitude is very intuitive. We now take a closer look to the tree.



The median_income turned out to be the most important variable in explaining the prices as before. So, as we can see, lower median income leads to lower median house value. Then comes the ocean proximity as expected, INLAND houses leads to lower median house price. The split of longitude in the unscaled dataset corresponds to -118.275 that is the exact longitude of Los Angeles. So the tree at the last node is doing the split whether a house is located around Los Angeles or not because if we recall the map, longitude lower in absolute value than -118.275 corresponds to houses located next to the ocean while houses having longitude than -118 becomes more INLAND.

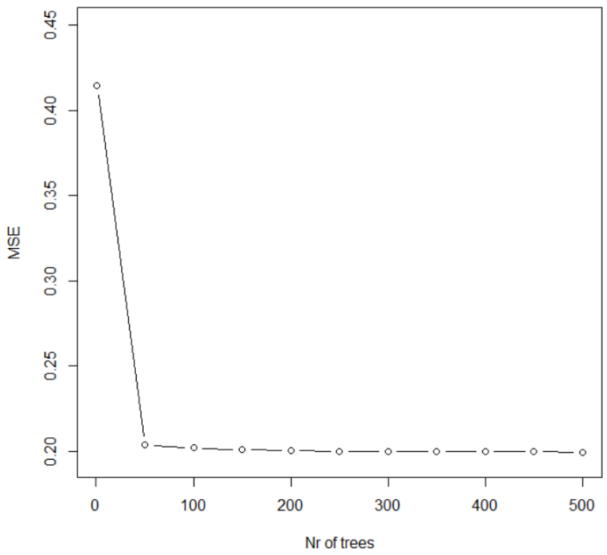
Notice that the tree is way too simpler than the polynomial regression with which we got the best result. Given any example, one can easily determine the median price of a house with the above structure.



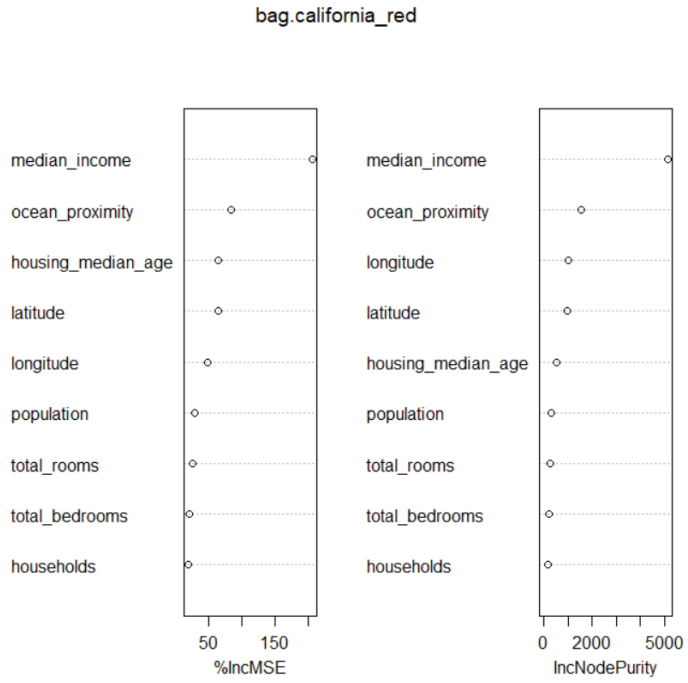
In this case, the most complex tree is selected by cross-validation. So, we do not have to prune the tree. The test MSE of this model is 0.447.

5.1 BAGGING

Decision trees suffer from high variance. One main method to improve the performance of tree based methods is bagging, which is based on bootstrap. Basically, we perform bootstrapping to create several different training sets from our original dataset, then a tree is grown on each of this sets and average the resulting predictions. By averaging the predictions, we successfully reduce the variance of the prediction, just like the variance of the sample mean of a random variable is lower than the variance of the random variable itself. We therefore now will analyse how the training error changes as the number of bootstrapped sample increases.



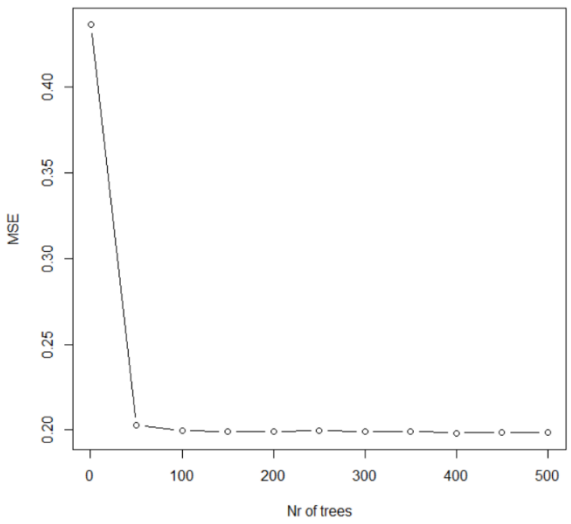
The plot shows how powerful ensemble methods are! MSE errors decrease dramatically as number of sampled trees increase. The lowest test MSE is reached with 500 trees which turns out to be 0.1994. This is way lower than any of the regression models' test errors applied on this dataset achieved so far. However, notice that the error becomes stable after around 250 trees. We therefore grown a model with 250 trees whose Test MSE turned out to be 0.1994. That's very good result!



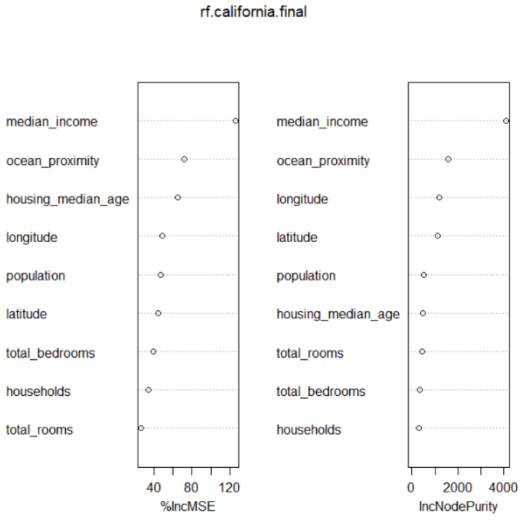
Two measures of variable importance are reported. The first one is based on the mean decrease of accuracy in predictions when a given variable is excluded. The second one is a measure of the total decrease in the node impurity that results from splits over that variable. Median income and ocean proximity still plays the heading role here.

5.2 RANDOM FOREST

We now apply Random Forest to our dataset. This approach is expected to improve the performance of bagging, especially in the situations where some of the available predictors are more correlated than others. In fact, what would happen in this case is that most of the trees built on different training sets will use the same (most correlated) set of variables, resulting in a forest of very similar and therefore correlated trees. Unfortunately, averaging correlated predictions does not reduce the variance of the classifier, rather the opposite. Random Forests therefore provide a way of decorrelating the grown trees. They do so by considering only a random sample of the available predictors for each split. In other words, at each split the algorithm will consider different predictor so that the final set of trees will not be similar as in the case of bagging. The typical choice for the number of predictors to be considered at each split is the square root of the number of predictors, so in our case we select p=3.

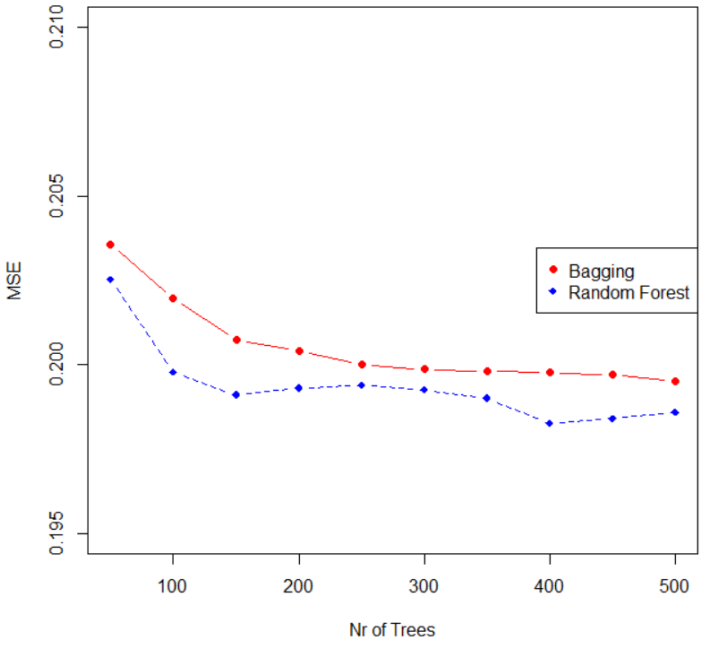


Test MSE is reported here. As we can see, with 400 trees, we get the lowest test MSE. As bagging, there is a sharp decrease in 50 trees, after that it stabilizes and reaches the lowest in 400. Therefore, we again apply Random Forest with 400 tree and the test MSE turns out to be 0.1986 which is slightly lower than bagging MSE as expected.



Again, the importance of variables are almost the same. Median income and ocean proximity leads the team.

5.3 Results



The plot reports the performances of Bagging and Random Forest. As the plot suggests, their performances do not differ significantly. However, even if slightly, Random Forest always outperforms Bagging.

CONLUSION

We first applied two variable model which we selected through correlation matrix and ANOVA test hoping to do statistical inferences. However, it had a very high MSE. We then applied model selections and decided to choose the seven-variable model(where we did not use any information about the number of rooms and bedrooms) because it provided a better MSE with a less complex, more parsimonious model. However, we saw that none of the assumptions of linear regression are met. Our response variable was not normal, there was the problem of multicollinearity, homoskedasticity and our residuals were not normal either. In residual analysis, we realized that a second degree polynomial would be better suited for this dataset. In fact, we got the lowest MSE with that. However, the interpretability of the polynomial model was super low. Therefore, one should make this trade-off between interpretability and accuracy.

	MSE
Two Variable	0.4089
Step-wise Full	0.35496
Step-wise-No_rooms	0.3518
Best Subset-eightvar	0.3531
Polynomial-Reg	0.297
Single-tree	0.447
Bagging	0.1994
Random-Forest	0.1986

We then moved to non-parametric models. In particular, we applied tree-based models. Trees returned a very interpretable result with respect to, especially, polynomial regression. The single tree had the highest MSE of all the models applied. We then applied ensemble methods hoping to decrease the test error which turn out to work really well. In fact, we got the lowest MSE with Random Forest.