STAT 420: Data Analysis Project

Amandeep,Kumar Gaurav and Vishal

08/03/2020

Table of Contents

## Team Engineers

|  |  |
| --- | --- |
| Name | NetID |
| Amandeep Takhar | atakhar2 |
| Kumar Gaurav | Kgdubey2 |
| Vishal Agarwal | vishala2 |

## Introduction

Title - **California Housing Price Prediction** An analysis on factors contributing to determine housing price in California

### Dataset background

The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data.The dataset contains 20640 records and 9 predictors. Our goal is to explore correlation between given variables like total bedroom ,population ,ocean proximity etc in determining the price of housing in a given area.In the process we would also like to divide the dataset into test and train and test the behavior of our model.

[Source Dataset](https://www.kaggle.com/camnugent/california-housing-prices)

Reading data

data = read.csv("housing.csv")

head(data, 10)

## longitude latitude housing\_median\_age total\_rooms total\_bedrooms population  
## 1 -122.23 37.88 41 880 129 322  
## 2 -122.22 37.86 21 7099 1106 2401  
## 3 -122.24 37.85 52 1467 190 496  
## 4 -122.25 37.85 52 1274 235 558  
## 5 -122.25 37.85 52 1627 280 565  
## 6 -122.25 37.85 52 919 213 413  
## 7 -122.25 37.84 52 2535 489 1094  
## 8 -122.25 37.84 52 3104 687 1157  
## 9 -122.26 37.84 42 2555 665 1206  
## 10 -122.25 37.84 52 3549 707 1551  
## households median\_income median\_house\_value ocean\_proximity  
## 1 126 8.3252 452600 NEAR BAY  
## 2 1138 8.3014 358500 NEAR BAY  
## 3 177 7.2574 352100 NEAR BAY  
## 4 219 5.6431 341300 NEAR BAY  
## 5 259 3.8462 342200 NEAR BAY  
## 6 193 4.0368 269700 NEAR BAY  
## 7 514 3.6591 299200 NEAR BAY  
## 8 647 3.1200 241400 NEAR BAY  
## 9 595 2.0804 226700 NEAR BAY  
## 10 714 3.6912 261100 NEAR BAY

str(data)

## 'data.frame': 20640 obs. of 10 variables:  
## $ longitude : num -122 -122 -122 -122 -122 ...  
## $ latitude : num 37.9 37.9 37.9 37.9 37.9 ...  
## $ housing\_median\_age: num 41 21 52 52 52 52 52 52 42 52 ...  
## $ total\_rooms : num 880 7099 1467 1274 1627 ...  
## $ total\_bedrooms : num 129 1106 190 235 280 ...  
## $ population : num 322 2401 496 558 565 ...  
## $ households : num 126 1138 177 219 259 ...  
## $ median\_income : num 8.33 8.3 7.26 5.64 3.85 ...  
## $ median\_house\_value: num 452600 358500 352100 341300 342200 ...  
## $ ocean\_proximity : chr "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...

### Description about the variables

1. longitude: A measure of how far west a house is; a higher value is farther west 2. latitude: A measure of how far north a house is; a higher value is farther north 3. housingMedianAge: Median age of a house within a block; a lower number is a newer building 4. totalRooms: Total number of rooms within a block 5. totalBedrooms: Total number of bedrooms within a block 6. population: Total number of people residing within a block 7. households: Total number of households, a group of people residing within a home unit, for a block 8. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars) 9. medianHouseValue: Median house value for households within a block (USD Response variable) 10. oceanProximity: Location of the house w.r.t ocean/sea

## Method

### Missing Data

As a first step in data quality , we will look for missing data.

sum(is.na(data))

## [1] 207

We see 207 missing values, which we plan to remove in the below step.

data = na.omit(data)  
str(data)

## 'data.frame': 20433 obs. of 10 variables:  
## $ longitude : num -122 -122 -122 -122 -122 ...  
## $ latitude : num 37.9 37.9 37.9 37.9 37.9 ...  
## $ housing\_median\_age: num 41 21 52 52 52 52 52 52 42 52 ...  
## $ total\_rooms : num 880 7099 1467 1274 1627 ...  
## $ total\_bedrooms : num 129 1106 190 235 280 ...  
## $ population : num 322 2401 496 558 565 ...  
## $ households : num 126 1138 177 219 259 ...  
## $ median\_income : num 8.33 8.3 7.26 5.64 3.85 ...  
## $ median\_house\_value: num 452600 358500 352100 341300 342200 ...  
## $ ocean\_proximity : chr "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...  
## - attr(\*, "na.action")= 'omit' Named int [1:207] 291 342 539 564 697 739 1098 1351 1457 1494 ...  
## ..- attr(\*, "names")= chr [1:207] "291" "342" "539" "564" ...

### Categorical Variables

On taking an in depth look at each variable, we decided to make ocean\_proximity as a categorical variable, we can see below that it is broadly classified into 5 values.

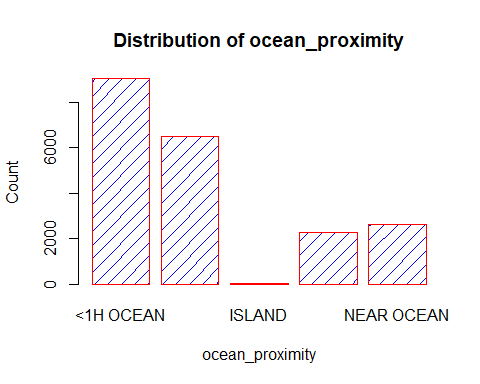
is.factor(data$ocean\_proximity)

## [1] FALSE

data$ocean\_proximity = as.factor(data$ocean\_proximity)  
levels(data$ocean\_proximity)

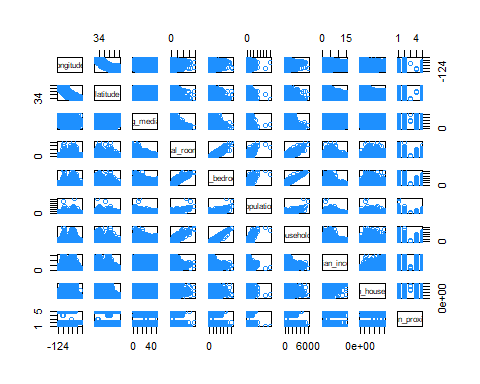
## [1] "<1H OCEAN" "INLAND" "ISLAND" "NEAR BAY" "NEAR OCEAN"

barplot(table(data$ocean\_proximity), main="Distribution of ocean\_proximity",  
 xlab="ocean\_proximity",  
 ylab="Count",  
 border="red",  
 col="blue",  
 density=10)



The distribution depicts that “island” has the least count and “1H OCEAN” has the maximum count. This data also make practical sense.

pairs(data, col = "dodgerblue")



kable(t(cor(data[,-10])))

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | longitude | latitude | housing\_median\_age | total\_rooms | total\_bedrooms | population | households | median\_income | median\_house\_value |
| longitude | 1.0000000 | -0.9246161 | -0.1093565 | 0.0454802 | 0.0696080 | 0.1002703 | 0.0565128 | -0.0155502 | -0.0453982 |
| latitude | -0.9246161 | 1.0000000 | 0.0118991 | -0.0366668 | -0.0669828 | -0.1089973 | -0.0717742 | -0.0796263 | -0.1446382 |
| housing\_median\_age | -0.1093565 | 0.0118991 | 1.0000000 | -0.3606283 | -0.3204510 | -0.2957873 | -0.3027680 | -0.1182777 | 0.1064320 |
| total\_rooms | 0.0454802 | -0.0366668 | -0.3606283 | 1.0000000 | 0.9303795 | 0.8572813 | 0.9189915 | 0.1978815 | 0.1332941 |
| total\_bedrooms | 0.0696080 | -0.0669828 | -0.3204510 | 0.9303795 | 1.0000000 | 0.8777467 | 0.9797283 | -0.0077228 | 0.0496862 |
| population | 0.1002703 | -0.1089973 | -0.2957873 | 0.8572813 | 0.8777467 | 1.0000000 | 0.9071859 | 0.0050866 | -0.0252997 |
| households | 0.0565128 | -0.0717742 | -0.3027680 | 0.9189915 | 0.9797283 | 0.9071859 | 1.0000000 | 0.0134339 | 0.0648935 |
| median\_income | -0.0155502 | -0.0796263 | -0.1182777 | 0.1978815 | -0.0077228 | 0.0050866 | 0.0134339 | 1.0000000 | 0.6883555 |
| median\_house\_value | -0.0453982 | -0.1446382 | 0.1064320 | 0.1332941 | 0.0496862 | -0.0252997 | 0.0648935 | 0.6883555 | 1.0000000 |

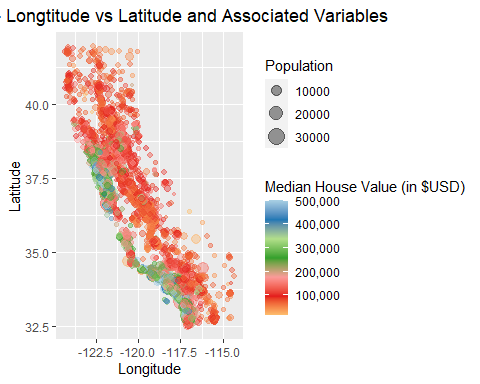
**We noticed there is collinearity between (households and total\_bedrooms) & (households and total\_rooms). We will keep this in mind and explore the data further**

### Training and Test Data

We took 80% of the data as training data and used seed to be consistent with the results.

set.seed(100)  
totalnrows = nrow(data)  
  
x = sample(totalnrows, floor(totalnrows \* .80) )  
train\_data = data[x, ]  
test\_data = data[-x, ]

plot\_map = ggplot(train\_data,   
 aes(x = longitude, y = latitude, color = median\_house\_value,   
 hma = housing\_median\_age, tr = total\_rooms, tb = total\_bedrooms,  
 hh = households, mi = median\_income)) +  
 geom\_point(aes(size = population), alpha = 0.4) +  
 xlab("Longitude") +  
 ylab("Latitude") +  
 ggtitle("Data Map - Longtitude vs Latitude and Associated Variables") +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 scale\_color\_distiller(palette = "Paired", labels = comma) +  
 labs(color = "Median House Value (in $USD)", size = "Population")  
plot\_map



The graph above shows distribution of Median house value based on population and Latitude. It gives us fair distribution of values across geographical area.

Additive Model

#Training additive Model  
model\_add = lm(median\_house\_value ~ ., data = train\_data)  
summary(model\_add)

##   
## Call:  
## lm(formula = median\_house\_value ~ ., data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -554770 -42731 -10480 28801 761094   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.274e+06 9.846e+04 -23.096 < 2e-16 \*\*\*  
## longitude -2.681e+04 1.140e+03 -23.512 < 2e-16 \*\*\*  
## latitude -2.540e+04 1.123e+03 -22.609 < 2e-16 \*\*\*  
## housing\_median\_age 1.102e+03 4.885e+01 22.557 < 2e-16 \*\*\*  
## total\_rooms -5.850e+00 8.771e-01 -6.670 2.64e-11 \*\*\*  
## total\_bedrooms 9.931e+01 7.737e+00 12.835 < 2e-16 \*\*\*  
## population -3.732e+01 1.183e+00 -31.533 < 2e-16 \*\*\*  
## households 4.817e+01 8.405e+00 5.731 1.02e-08 \*\*\*  
## median\_income 3.905e+04 3.740e+02 104.386 < 2e-16 \*\*\*  
## ocean\_proximityINLAND -3.966e+04 1.954e+03 -20.295 < 2e-16 \*\*\*  
## ocean\_proximityISLAND 1.531e+05 3.068e+04 4.990 6.09e-07 \*\*\*  
## ocean\_proximityNEAR BAY -4.041e+03 2.122e+03 -1.904 0.05691 .   
## ocean\_proximityNEAR OCEAN 5.578e+03 1.744e+03 3.199 0.00138 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 68490 on 16333 degrees of freedom  
## Multiple R-squared: 0.6471, Adjusted R-squared: 0.6469   
## F-statistic: 2496 on 12 and 16333 DF, p-value: < 2.2e-16

summary(model\_add)$adj.r.squared

## [1] 0.6468567

**By analyzing p-value of all Beta variable in Additive model, we can say that we fail to reject that Null Hypothesis that Beta value of any variable is Zero. Hence all variables are playing important role in prediction of House Median Income. And Adjusted R squared value of Model is 64.6%**

Interaction Model

model\_int = lm(median\_house\_value ~ . ^ 2, data = train\_data)  
summary(model\_int)$adj.r.squared

## [1] 0.7025208

**In interaction model we can see an increment of Model performance by Adjusted R Squared which is 70.3%**

Testing Interaction model with respect to Additive Model

anova(model\_int, model\_add)

## Analysis of Variance Table  
##   
## Model 1: median\_house\_value ~ (longitude + latitude + housing\_median\_age +   
## total\_rooms + total\_bedrooms + population + households +   
## median\_income + ocean\_proximity)^2  
## Model 2: median\_house\_value ~ longitude + latitude + housing\_median\_age +   
## total\_rooms + total\_bedrooms + population + households +   
## median\_income + ocean\_proximity  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 16277 6.4322e+13   
## 2 16333 7.6621e+13 -56 -1.2299e+13 55.575 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**P-value of test is 2.2e-16 which is very less hence we can consider Interactive models is better than additive model**

### Model Improvement Using AIC and BIC

model\_add\_aic = step(model\_add, direction = "backward", trace = 0)  
summary(model\_add\_aic)$adj.r.squared

## [1] 0.6468567

model\_add\_bic = step(model\_add, direction = "backward", trace = 0, k = log(nrow(train\_data)))  
summary(model\_add\_bic)$adj.r.squared

## [1] 0.6468567

model\_int\_aic = step(model\_int, direction = "backward", trace = 0)  
summary(model\_int\_aic)$adj.r.squared

## [1] 0.7025212

model\_int\_bic = step(model\_int, direction = "backward", trace = 0, k = log(nrow(train\_data)))  
summary(model\_int\_bic)$adj.r.squared

## [1] 0.7019587

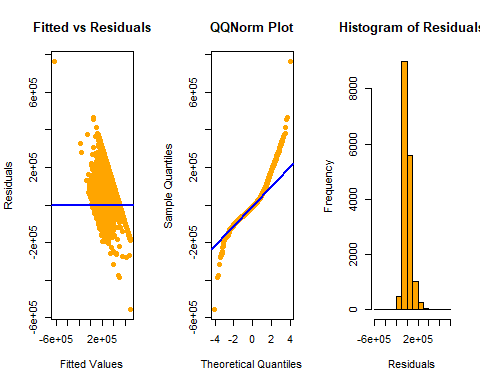
beginning\_mods\_results = data.frame(  
 "Total Predictors" =  
 c("Additive Model" = extractAIC(model\_add)[1],  
 "Interaction Model" = extractAIC(model\_int)[1],  
 "AIC\_additive Model" = extractAIC(model\_add\_aic)[1],  
 "AIC\_Int Model" = extractAIC(model\_int\_aic)[1],  
 "BIC\_additive Model" = extractAIC(model\_add\_bic)[1],  
 "BIC\_Int Model" = extractAIC(model\_int\_bic)[1]),  
 "AIC" =  
 c("Additive Model" = extractAIC(model\_add)[2],  
 "Interaction Model" = extractAIC(model\_int)[2],  
 "AIC\_additive Model" = extractAIC(model\_add\_aic)[2],  
 "AIC\_Int Model" = extractAIC(model\_int\_aic)[2],  
 "BIC\_additive Model" = extractAIC(model\_add\_bic)[2],  
 "BIC\_Int Model" = extractAIC(model\_int\_bic)[2]),  
 "Adj R-Squared" =  
 c("Additive Model" = summary(model\_add)$adj.r.squared,  
 "Interaction Model" = summary(model\_int)$adj.r.squared,  
 "AIC\_additive Model" =summary(model\_add\_aic)$adj.r.squared,  
 "AIC\_Int Model" = summary(model\_int\_aic)$adj.r.squared,  
 "BIC\_additive Model" =summary((model\_add\_bic))$adj.r.squared,  
 "BIC\_Int Model" = summary(model\_int\_bic)$adj.r.squared))  
  
kable(beginning\_mods\_results, align = c("c", "r"))

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total.Predictors | AIC | Adj.R.Squared |
| Additive Model | 13 | 364021.2 | 0.6468567 |
| Interaction Model | 69 | 361273.2 | 0.7025208 |
| AIC\_additive Model | 13 | 364021.2 | 0.6468567 |
| AIC\_Int Model | 64 | 361268.2 | 0.7025212 |
| BIC\_additive Model | 13 | 364021.2 | 0.6468567 |
| BIC\_Int Model | 56 | 361291.2 | 0.7019587 |

**We see that the model with the best (i.e., lowest) AIC is Interaction Model, with a score of 361268.2. But we will work further to enhance performance of model.**

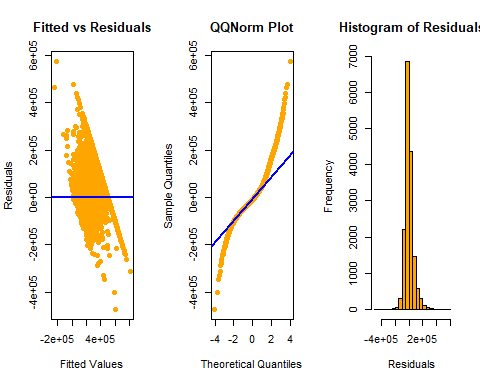
diagnostics = function(model, alpha = .05, pointcol = "orange", linecol = "blue", plots = TRUE, tests = TRUE, pointtype = 16) {  
 if (plots == TRUE) {  
 par(mfrow = c(1, 3))  
 plot(  
 fitted(model),  
 resid(model),  
 pch = pointtype,  
 xlab = "Fitted Values",  
 ylab = "Residuals",  
 main = "Fitted vs Residuals",  
 col = pointcol  
 )  
 abline(h = 0, lwd = 2, col = linecol)  
   
 qqnorm(  
 resid(model),  
 pch = pointtype,  
 main = "QQNorm Plot",  
 col = pointcol  
 )  
 qqline(  
 resid(model),  
 lwd = 2,  
 col = linecol  
 )  
 hist(  
 resid(model),  
 main = "Histogram of Residuals",  
 col = pointcol,  
 xlab = "Residuals",  
 ylab = "Frequency"  
 )  
 }  
 if (tests == TRUE) {  
 ks\_test = ks.test(resid(model),y='pnorm',alternative='two.sided')  
 bp\_test = bptest(model)  
 test\_results = data.frame(  
 "Kolmogorov-Smirnov Test" =  
 c("Test Statistic" = round(ks\_test$statistic, 5),  
 "P-Value" = ks\_test$p.value,  
 "Result" = ifelse(ks\_test$p.value < alpha, "Reject", "Fail To Reject")),  
 "Breusch-Pagan Test" =  
 c("Test Statistic" = round(bp\_test$statistic, 5),  
 "P-Value" = bp\_test$p.value,  
 "Result" = ifelse(bp\_test$p.value < alpha, "Reject", "Fail To Reject")))  
  
 kable(t(test\_results), col.names = c("Test Statistic", "P-Value", "Decision"))  
 }  
}

diagnostics(model\_add)



|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Statistic | P-Value | Decision |
| Kolmogorov.Smirnov..Test | 0.5802 | 0 | Reject |
| Breusch.Pagan.Test | 813.50284 | 2.10341410745741e-166 | Reject |

diagnostics(model\_int)

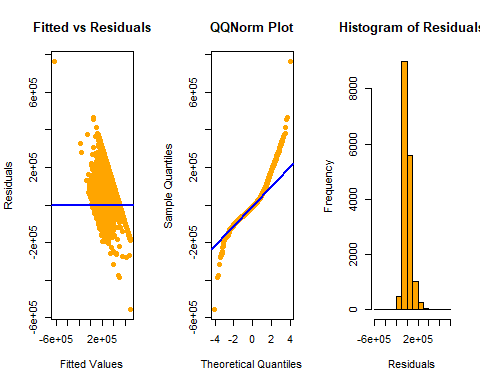


|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Statistic | P-Value | Decision |
| Kolmogorov.Smirnov..Test | 0.57806 | 0 | Reject |
| Breusch.Pagan.Test | 1698.68853 | 7.46002686123318e-310 | Reject |

diagnostics(model\_add\_aic)

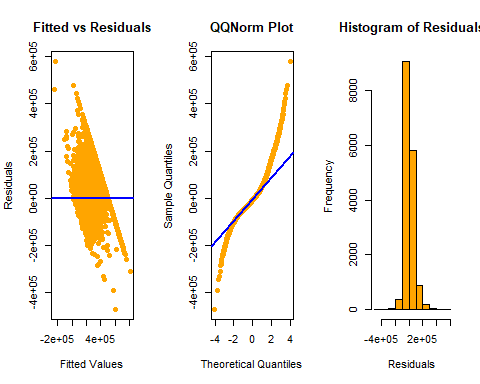
|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Statistic | P-Value | Decision |
| Kolmogorov.Smirnov..Test | 0.5802 | 0 | Reject |
| Breusch.Pagan.Test | 813.50284 | 2.10341410745741e-166 | Reject |

diagnostics(model\_add\_bic)



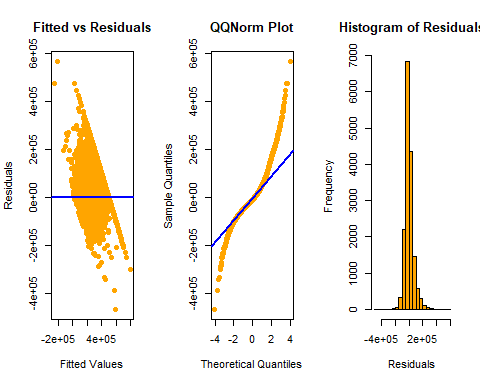
|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Statistic | P-Value | Decision |
| Kolmogorov.Smirnov..Test | 0.5802 | 0 | Reject |
| Breusch.Pagan.Test | 813.50284 | 2.10341410745741e-166 | Reject |

diagnostics(model\_int\_aic)



|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Statistic | P-Value | Decision |
| Kolmogorov.Smirnov..Test | 0.57873 | 0 | Reject |
| Breusch.Pagan.Test | 1589.03891 | 1.76669099023126e-290 | Reject |

diagnostics(model\_int\_bic)



|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Statistic | P-Value | Decision |
| Kolmogorov.Smirnov..Test | 0.57776 | 0 | Reject |
| Breusch.Pagan.Test | 1436.29723 | 3.15291874921026e-264 | Reject |

x = ks.test(x=rnorm(10^4),y='pnorm',alternative='two.sided')  
  
x$p.value

## [1] 0.9685713

We can see that all above models do not have Equal variance and residual in Normal form. Hence we need to improve model.

Kolmogorov–Smirnov test- In statistics, the Kolmogorov–Smirnov test is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution, or to compare two samples. Note- We tried using shapiro.test first, but the test that did not work considering the size of the dataset.

### Model Improvement

Now, we will calculate the cooks distance and will remove outliers and high influential values.

value = cooks.distance(model\_add)  
sum(value > 4 / length(resid(model\_add)))

## [1] 885

model\_new\_add = lm(median\_house\_value ~ ., data = train\_data, subset = value <= (4 / nrow(train\_data)))  
  
model\_new\_int = lm(median\_house\_value ~ .^2, data = train\_data, subset = value <= (4 / nrow(train\_data)))  
  
model\_new\_add\_AIC = step(model\_new\_add, direction = "backward", trace = 0)  
  
model\_new\_int\_AIC = step(model\_new\_int, direction = "backward", trace = 0)

Based on the new data data values, we will again train the models and calculate ADJ R Squared and LOOCV values (Leave-One-Out Cross-Validation)

## Results

When we initially calculated the AdjustedR2 value the results were not very convincing as we had low ADJ R Squared value for all the models. However,when we remove the outliers and high influential values using the cooks distance we got better results.

Result = data.frame(  
 "Additive Model" =c("LOOCV" = sqrt(mean((resid(model\_new\_add) / (1 - hatvalues(model\_new\_add))) ^ 2)),  
 "ADJ R Squared" = summary(model\_new\_add)$adj.r.squared,  
 "Test RMSE" = sqrt(mean((test\_data$median\_house\_value - predict(model\_new\_add, newdata = test\_data))^2)),  
 "SE" = summary(model\_new\_add)$sigma),  
   
 "Interaction Model" = c( "LOOCV" = sqrt(mean((resid(model\_new\_int) / (1 - hatvalues(model\_new\_int))) ^ 2)),  
 "ADJ R Squared" = summary(model\_new\_int)$adj.r.squared,  
 "Test RMSE" = sqrt(mean((test\_data$median\_house\_value - predict(model\_new\_int, newdata = test\_data))^2)),  
 "SE" = summary(model\_new\_int)$sigma),  
   
 "Additive Model AIC" = c( "LOOCV" = sqrt(mean((resid(model\_new\_add\_AIC) / (1 - hatvalues(model\_new\_add\_AIC))) ^ 2)),  
 "ADJ R Squared" = summary(model\_new\_add\_AIC)$adj.r.squared,  
 "Test RMSE" = sqrt(mean((test\_data$median\_house\_value - predict(model\_new\_add\_AIC, newdata = test\_data))^2)),  
 "SE" = summary(model\_new\_add\_AIC)$sigma),  
   
 "Interaction Model AIC" = c( "LOOCV" = sqrt(mean((resid(model\_new\_int\_AIC) / (1 - hatvalues(model\_new\_int\_AIC))) ^ 2)),  
 "ADJ R Squared" = summary(model\_new\_int\_AIC)$adj.r.squared,  
 "Test RMSE" = sqrt(mean((test\_data$median\_house\_value - predict(model\_new\_int\_AIC, newdata = test\_data))^2)),  
 "SE" = summary(model\_new\_int\_AIC)$sigma)  
   
)  
  
 kable(t(Result))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | LOOCV | ADJ R Squared | Test RMSE | SE |
| Additive.Model | 53496.07 | 0.7424312 | 69798.42 | 53477.11 |
| Interaction.Model | 49852.24 | 0.7847754 | 64271.04 | 48884.06 |
| Additive.Model.AIC | 53496.07 | 0.7424312 | 69798.42 | 53477.11 |
| Interaction.Model.AIC | 49773.83 | 0.7847864 | 64266.58 | 48882.81 |

Based on the results, we can say that Interaction.Model.AIC is having better ADJ R Squared(0.7847864) among all model and hence can be considered best among the given model. Also, this is also better than the previous all models discussed( without removal of outliers“) where the max adjusted R2 value was 0.7025212 for”AIC\_Int Model"

## Discussion

As shown above table, our selected model “model\_new\_int\_AIC” (AIC of Interaction Model) has lowest LOOCV RMSE in all models i.e 49773.83 and better Adjusted R squared around 78.5%. We have an average Standard Error 48882.21 that means on average, our model’s predicted housing price will be ± 48882.21 in comparison to the actual price.

Above table also shows Model performance on Test Data. “Test RMSE” columns shows root squared error for Test Data and “model\_new\_int\_AIC” showed lowest RMSE in all i.e. 64266.58.

Our aim was to predict Housing price for California Region and based on above observation we can conclude that No individual predictor determines the cost of the house however interaction of predictor make up better prediction model.

## Appendix

* Names of Team : Team Engineer
* Original Data :

head(data, 5)

## longitude latitude housing\_median\_age total\_rooms total\_bedrooms population  
## 1 -122.23 37.88 41 880 129 322  
## 2 -122.22 37.86 21 7099 1106 2401  
## 3 -122.24 37.85 52 1467 190 496  
## 4 -122.25 37.85 52 1274 235 558  
## 5 -122.25 37.85 52 1627 280 565  
## households median\_income median\_house\_value ocean\_proximity  
## 1 126 8.3252 452600 NEAR BAY  
## 2 1138 8.3014 358500 NEAR BAY  
## 3 177 7.2574 352100 NEAR BAY  
## 4 219 5.6431 341300 NEAR BAY  
## 5 259 3.8462 342200 NEAR BAY

* Outlier and high influence points removal by Cook’s Distance
* Best Model

summary(model\_new\_int\_AIC)

##   
## Call:  
## lm(formula = median\_house\_value ~ longitude + latitude + housing\_median\_age +   
## total\_rooms + total\_bedrooms + population + households +   
## median\_income + ocean\_proximity + longitude:latitude + longitude:housing\_median\_age +   
## longitude:total\_rooms + longitude:total\_bedrooms + longitude:households +   
## longitude:median\_income + longitude:ocean\_proximity + latitude:housing\_median\_age +   
## latitude:total\_rooms + latitude:total\_bedrooms + latitude:median\_income +   
## latitude:ocean\_proximity + housing\_median\_age:total\_rooms +   
## housing\_median\_age:total\_bedrooms + housing\_median\_age:population +   
## housing\_median\_age:households + housing\_median\_age:median\_income +   
## housing\_median\_age:ocean\_proximity + total\_rooms:population +   
## total\_rooms:households + total\_rooms:median\_income + total\_rooms:ocean\_proximity +   
## total\_bedrooms:population + total\_bedrooms:households + total\_bedrooms:median\_income +   
## total\_bedrooms:ocean\_proximity + population:households +   
## population:median\_income + population:ocean\_proximity + households:median\_income +   
## median\_income:ocean\_proximity, data = train\_data, subset = value <=   
## (4/nrow(train\_data)))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -237366 -30345 -5336 25048 380747   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -2.750e+06 8.339e+05 -3.298  
## longitude -9.984e+03 7.428e+03 -1.344  
## latitude 2.239e+05 2.512e+04 8.913  
## housing\_median\_age -7.573e+04 7.008e+03 -10.805  
## total\_rooms 1.336e+03 1.811e+02 7.379  
## total\_bedrooms -5.715e+03 9.617e+02 -5.942  
## population -1.851e+01 5.009e+00 -3.696  
## households -1.420e+03 4.243e+02 -3.347  
## median\_income -9.931e+05 6.630e+04 -14.980  
## ocean\_proximityINLAND -1.816e+04 2.252e+05 -0.081  
## ocean\_proximityNEAR BAY -1.746e+07 1.133e+06 -15.405  
## ocean\_proximityNEAR OCEAN -1.123e+06 2.803e+05 -4.005  
## longitude:latitude 1.489e+03 1.929e+02 7.717  
## longitude:housing\_median\_age -9.418e+02 8.100e+01 -11.627  
## longitude:total\_rooms 1.650e+01 2.146e+00 7.690  
## longitude:total\_bedrooms -7.631e+01 1.122e+01 -6.799  
## longitude:households -1.211e+01 3.605e+00 -3.360  
## longitude:median\_income -1.226e+04 7.841e+02 -15.635  
## longitude:ocean\_proximityINLAND 2.115e+03 2.668e+03 0.793  
## longitude:ocean\_proximityNEAR BAY -1.751e+05 1.004e+04 -17.436  
## longitude:ocean\_proximityNEAR OCEAN -1.081e+04 3.349e+03 -3.229  
## latitude:housing\_median\_age -1.049e+03 7.975e+01 -13.150  
## latitude:total\_rooms 1.772e+01 2.213e+00 8.006  
## latitude:total\_bedrooms -9.453e+01 1.161e+01 -8.144  
## latitude:median\_income -1.247e+04 8.105e+02 -15.385  
## latitude:ocean\_proximityINLAND 5.402e+03 2.770e+03 1.950  
## latitude:ocean\_proximityNEAR BAY -1.033e+05 7.674e+03 -13.456  
## latitude:ocean\_proximityNEAR OCEAN -4.645e+03 3.505e+03 -1.325  
## housing\_median\_age:total\_rooms -6.179e-01 7.637e-02 -8.091  
## housing\_median\_age:total\_bedrooms 4.127e+00 8.110e-01 5.088  
## housing\_median\_age:population -1.545e+00 1.102e-01 -14.023  
## housing\_median\_age:households 3.626e+00 9.027e-01 4.017  
## housing\_median\_age:median\_income 2.722e+02 2.586e+01 10.527  
## housing\_median\_age:ocean\_proximityINLAND 5.413e+02 1.343e+02 4.032  
## housing\_median\_age:ocean\_proximityNEAR BAY -7.071e+02 1.472e+02 -4.805  
## housing\_median\_age:ocean\_proximityNEAR OCEAN -1.586e+02 1.222e+02 -1.298  
## total\_rooms:population -7.594e-03 1.117e-03 -6.798  
## total\_rooms:households 2.191e-02 3.402e-03 6.438  
## total\_rooms:median\_income 3.430e+00 3.098e-01 11.071  
## total\_rooms:ocean\_proximityINLAND -1.709e+01 3.284e+00 -5.205  
## total\_rooms:ocean\_proximityNEAR BAY 1.336e+01 3.486e+00 3.832  
## total\_rooms:ocean\_proximityNEAR OCEAN 5.971e+00 2.592e+00 2.303  
## total\_bedrooms:population 2.371e-02 7.014e-03 3.381  
## total\_bedrooms:households -1.449e-01 1.549e-02 -9.353  
## total\_bedrooms:median\_income 1.636e+01 4.962e+00 3.296  
## total\_bedrooms:ocean\_proximityINLAND 2.877e+01 1.821e+01 1.580  
## total\_bedrooms:ocean\_proximityNEAR BAY -5.653e+01 1.664e+01 -3.397  
## total\_bedrooms:ocean\_proximityNEAR OCEAN -3.727e+01 1.301e+01 -2.865  
## population:households 2.690e-02 5.100e-03 5.275  
## population:median\_income -2.564e+00 7.097e-01 -3.613  
## population:ocean\_proximityINLAND 2.349e+01 2.552e+00 9.206  
## population:ocean\_proximityNEAR BAY -7.087e+00 4.641e+00 -1.527  
## population:ocean\_proximityNEAR OCEAN 1.764e+00 2.993e+00 0.590  
## households:median\_income -2.377e+01 5.783e+00 -4.110  
## median\_income:ocean\_proximityINLAND 5.920e+03 1.278e+03 4.630  
## median\_income:ocean\_proximityNEAR BAY -3.243e+03 1.202e+03 -2.697  
## median\_income:ocean\_proximityNEAR OCEAN -9.153e+02 9.691e+02 -0.944  
## Pr(>|t|)   
## (Intercept) 0.000977 \*\*\*  
## longitude 0.178903   
## latitude < 2e-16 \*\*\*  
## housing\_median\_age < 2e-16 \*\*\*  
## total\_rooms 1.67e-13 \*\*\*  
## total\_bedrooms 2.87e-09 \*\*\*  
## population 0.000220 \*\*\*  
## households 0.000818 \*\*\*  
## median\_income < 2e-16 \*\*\*  
## ocean\_proximityINLAND 0.935727   
## ocean\_proximityNEAR BAY < 2e-16 \*\*\*  
## ocean\_proximityNEAR OCEAN 6.24e-05 \*\*\*  
## longitude:latitude 1.26e-14 \*\*\*  
## longitude:housing\_median\_age < 2e-16 \*\*\*  
## longitude:total\_rooms 1.56e-14 \*\*\*  
## longitude:total\_bedrooms 1.09e-11 \*\*\*  
## longitude:households 0.000781 \*\*\*  
## longitude:median\_income < 2e-16 \*\*\*  
## longitude:ocean\_proximityINLAND 0.427824   
## longitude:ocean\_proximityNEAR BAY < 2e-16 \*\*\*  
## longitude:ocean\_proximityNEAR OCEAN 0.001246 \*\*   
## latitude:housing\_median\_age < 2e-16 \*\*\*  
## latitude:total\_rooms 1.27e-15 \*\*\*  
## latitude:total\_bedrooms 4.13e-16 \*\*\*  
## latitude:median\_income < 2e-16 \*\*\*  
## latitude:ocean\_proximityINLAND 0.051162 .   
## latitude:ocean\_proximityNEAR BAY < 2e-16 \*\*\*  
## latitude:ocean\_proximityNEAR OCEAN 0.185183   
## housing\_median\_age:total\_rooms 6.37e-16 \*\*\*  
## housing\_median\_age:total\_bedrooms 3.65e-07 \*\*\*  
## housing\_median\_age:population < 2e-16 \*\*\*  
## housing\_median\_age:households 5.91e-05 \*\*\*  
## housing\_median\_age:median\_income < 2e-16 \*\*\*  
## housing\_median\_age:ocean\_proximityINLAND 5.57e-05 \*\*\*  
## housing\_median\_age:ocean\_proximityNEAR BAY 1.56e-06 \*\*\*  
## housing\_median\_age:ocean\_proximityNEAR OCEAN 0.194304   
## total\_rooms:population 1.10e-11 \*\*\*  
## total\_rooms:households 1.24e-10 \*\*\*  
## total\_rooms:median\_income < 2e-16 \*\*\*  
## total\_rooms:ocean\_proximityINLAND 1.97e-07 \*\*\*  
## total\_rooms:ocean\_proximityNEAR BAY 0.000128 \*\*\*  
## total\_rooms:ocean\_proximityNEAR OCEAN 0.021267 \*   
## total\_bedrooms:population 0.000723 \*\*\*  
## total\_bedrooms:households < 2e-16 \*\*\*  
## total\_bedrooms:median\_income 0.000982 \*\*\*  
## total\_bedrooms:ocean\_proximityINLAND 0.114236   
## total\_bedrooms:ocean\_proximityNEAR BAY 0.000684 \*\*\*  
## total\_bedrooms:ocean\_proximityNEAR OCEAN 0.004174 \*\*   
## population:households 1.35e-07 \*\*\*  
## population:median\_income 0.000304 \*\*\*  
## population:ocean\_proximityINLAND < 2e-16 \*\*\*  
## population:ocean\_proximityNEAR BAY 0.126741   
## population:ocean\_proximityNEAR OCEAN 0.555492   
## households:median\_income 3.98e-05 \*\*\*  
## median\_income:ocean\_proximityINLAND 3.68e-06 \*\*\*  
## median\_income:ocean\_proximityNEAR BAY 0.007008 \*\*   
## median\_income:ocean\_proximityNEAR OCEAN 0.344960   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 48880 on 15404 degrees of freedom  
## Multiple R-squared: 0.7856, Adjusted R-squared: 0.7848   
## F-statistic: 1008 on 56 and 15404 DF, p-value: < 2.2e-16