A picture containing drawing

Description automatically generated

­­­­­­Large orange blocky "S" placed on the bottom right corner of the slide. 

IST 687 Group Project

Final Report  
  
March 29, 2022

Prepared by:

John Jefferson

Tanya MacKinnon

John O’Rourke

Contents

[1. Summary 4](#_Toc99473961)

[Measuring Results 4](#_Toc99473962)

[Conclusion 4](#_Toc99473963)

[2. Data Set Overview 5](#_Toc99473964)

[Oddities 5](#_Toc99473965)

[Fun facts 6](#_Toc99473966)

[3. Data Cleaning 6](#_Toc99473967)

[Null Value Handling 6](#_Toc99473968)

[Factor Variable 6](#_Toc99473969)

[Unit Adjustments 6](#_Toc99473970)

[House vs Block 6](#_Toc99473971)

[4. Data Exploration 7](#_Toc99473972)

[a. Examining Variables 7](#_Toc99473973)

[Observations 7](#_Toc99473974)

[Histograms 8](#_Toc99473975)

[Box plots 9](#_Toc99473976)

[b. Examining Variables Relationships 10](#_Toc99473977)

[Observations 10](#_Toc99473978)

[Correlation Matrix 10](#_Toc99473979)

[Scatter Plot with Best Fit Line 11](#_Toc99473980)

[Maps 13](#_Toc99473981)

[5. Model Development 15](#_Toc99473982)

[Outcome variable selection 15](#_Toc99473983)

[Predictor variables selection 15](#_Toc99473984)

[Answering the questions 15](#_Toc99473985)

[Stepped Approach 15](#_Toc99473986)

[Train and Test Data Sets 15](#_Toc99473987)

[Measuring Prediction Results 16](#_Toc99473988)

[Cross Validation 16](#_Toc99473989)

[Comparing Model Performance 18](#_Toc99473990)

[R2 values by Model Name 18](#_Toc99473991)

[RMSE values by Model Name 18](#_Toc99473992)

[6. Experimentation 19](#_Toc99473993)

[Will removing outlier from the data set improve the accuracy of the model? 19](#_Toc99473994)

[7. Conclusion 20](#_Toc99473995)

[8. Next Steps - Continuing the Analysis 23](#_Toc99473996)

[9. Sources 23](#_Toc99473997)

[Appendix 1 – R Notebook 23](#_Toc99473998)

# Summary

The project team began this analysis by asking, what best predicts California housing prices?

The team drilled down further and initial question evolved to these six questions

1. Location, location, location, it’s a common saying in real estate but is it true? Is location the best predictor?
2. Is true the more money you have the more house you buy? Is household income the best predictor?
3. Is urban more desirable than suburban or rural? Is population density the best predictor?
4. Would a stepped approach to variable selection produce the best results?
5. Would using all available variables produce the most accurate prediction?
6. Would results improve using if outliers were removed from the data set?

This paper analyzes these questions using a data set of housing prices from 1990 California census by attempting to predict the median house value.

The programing language, R was used to develop features to execute this data science endeavor. The code developed was organized into a R markdown notebook which is found in Appendix 1.

The team chose to organize and execute the project using a phased approach.

|  |  |
| --- | --- |
| Phase | Goal |
| Project Initiation | Data set selection and outcome variable identification |
| Data Loading | Programmatic features to initialize the environment and load data |
| Data Cleaning | Identify data transformations and develop features |
| Data Exploration | Examination of variables and their relationships |
| Model Development | Features to train and test linear models |
| Experimentation | Data transformation to improve accuracy |
| Results | Analyze results |

## Measuring Results

Prediction results were compared to the actual results in the test data set. Two measures were selected to evaluate the quality of predictions:

* Root mean square error
  + Measures the square root of variance of the residuals
* R2 predict vs test
  + Measures the relative fit of the model

## Conclusion

The most accurate model produced in this analysis used all available variables to predict median\_house\_value. The results were further improved when k fold cross validation techniques were introduced.

* Adjusted R2 on train data – 0.6528
* RMSE on predict vs test data – 7.00
* R2 on predict vs test data – 0.6283

# Data Set Overview

* This a data set is California housing prices based on 1990 Census t
* It includes 20640 observations of 10 variables
* It is a widely published data set within Machine Learning literature
  + The data set contains intricacies suitable for experimenting with various Machine Learning techniques
  + The published body of work provides guidance on model development

| Field Name | Description | Min | Max | Mean | Skewness | Std Deviation |
| --- | --- | --- | --- | --- | --- | --- |
| median\_house\_value | Median house price in USD within a block capped at 500001 | 14999 | 500001 | 206856 | 0.98 | 1.15 |
| median\_income | Median income for households within a block in 10000 US$ capped at 15 | 0.4999 | 15.001 | 3.8707 | 1.65 | 1.90 |
| longitude | longitude | -124.3 | -114.3 | -119.6 |  | 2.00 |
| latitude | latitude | 32.54 | 41.95 | 35.63 |  | 2.14 |
| housing\_median\_age | Median age of a house within a block | 1 | 52 | 28.64 | 0.06 | 1.29 |
| total\_rooms | Total rooms within a block | 2 | 39320 | 2636 | 4.15 | 2.18 |
| total\_bedrooms | Total bedrooms within a blocks  *207 null observations* | 1 | 6445 | 537.9 | 3.46 | 4.2 |
| population | Total population within a block | 3 | 1425 | 35682 | 4.94 | 1.13 |
| households | Households within a block | 1 | 6082 | 499.5 | 3.41 | 3.82 |
| ocean\_proximity | Description of proximity to ocean - <1H OCEAN , NEAR OCEAN, NEAR BAY, INLAND, ISLAND | | | | | |

## Oddities

* + Blocks vs House
    - Some variables describe the median home while others describe the block where the median home is found
      * total\_rooms, total\_bedrooms, and population describe the entire block
  + Capped values
    - Both median\_income and median\_house\_value are capped at seemingly arbitrary values. Certainly, there were both incomes and house prices above the maximum values in this data set in California in 1990.
      * It is unclear how this distortion will affect the reliability of the model constructed using this data set. Any efforts to adjust the data set to account for this distortion appear cumbersome and beyond the skill set of the project team.
      * Is this data set biased against the wealthy?
  + US$ units
    - median\_house\_value and median\_income are both report in US dollars. However, the unit on median\_income is in $10000 units while median\_house\_value is in whole dollars.

## Fun facts

* + Location of mean latitude and longitude is Lost Hills, California
    - <https://www.google.com/maps/place/35%C2%B037'48.0%22N+119%C2%B036'00.0%22W/@34.9309945,-117.0087231,7.42z/data=!4m5!3m4!1s0x0:0x8002b69cd9b5a3e4!8m2!3d35.63!4d-119.6>
    - It is inland between Bakersfield and Fresno
  + This data set was first published by R. Kelly Pace and Ronald Barry in “Sparse Spatial Autoregression”, Statistics and Probability Letters, 33, no.3 (1997)
    - <https://www.sciencedirect.com/science/article/pii/S016771529600140X>

# Data Cleaning

## Null Value Handling

* Replace null values in total\_bedrooms
  + Using the imputeTS library, the na\_interpolation function was used to replace the null values in 207 records
  + There are no other null values within the dataset

## Factor Variable

* A factor variable was created using ocean\_proximity
  + A new field, ocean\_factor, was append to the dataset using the as.factor function on the ocean\_proximity field

## Unit Adjustments

### House vs Block

* Units were adjusted to measure population, rooms and bedrooms to express these value in relation to a household rather than a block
  + New fields were appended to the dataset to store these values

|  |  |  |  |
| --- | --- | --- | --- |
| New Field | Min | Max | Mean |
| rooms\_per\_house | 0.8461 | 141.9091 | 5.4290 |
| bedrooms\_per\_house | 0.05394 | 34.06667 | 1.10199 |
| population\_per\_house | 0.6923 | 1243.3333 | 3.0707 |

US Dollar Units

* Median\_house\_value was adjusted to express the variable in units of $10000 to agree with the representation of median\_income
  + Revised values:

Text

Description automatically generated

# Data Exploration

# Examining Variables

Histograms and box plots were used to visually examine the variables.

## Observations

* Outliers
  + median\_income and median\_house\_value both include outliers which can be intuitively explained by the economic stratification of society
  + Values describing the median home such as rooms, bedroom and population extreme values which are beyond intuition.
    - It is hard to imagine a household where 1200 people live or a home with 140 rooms.
* Left Skew
  + Many variables are heavily left skewed particularly: median\_house\_value, median\_income, rooms\_per\_house, bedrooms\_per\_house, population\_per\_house and households

## Histograms

Chart, bar chart, waterfall chart

Description automatically generated

## Box plots

Chart, box and whisker chart

Description automatically generated

# Examining Variables Relationships

Correlation matrix, scatter plots with best fit lines and maps were used to visually examine the relationship between the variables.

## Observations

* median\_income has greatest correlation coeffiecient of 0.6881 with median\_house\_value
* No other variable has significant correlation with median\_house\_value
* Variables for bedrooms, rooms, and population were found to have significant correlation with each other
* Median\_income was the only variable found to have a significant linear relationship with median house value, with R2 of 47%
* P-values on the linear relationships between median\_house\_value and the other variables were all statistically significant
* Data is concentrated in California’s biggest urban centers including San Francisco, Los Angeles Sacramento and San Diego
* Homes near the borders of Arizona and Nevada had the least amount of room, bedrooms, median income and population

## Correlation Matrix

A picture containing table

Description automatically generated

## Scatter Plot with Best Fit Line

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## Maps

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

# Model Development

## Outcome variable selection

The intuitive question proposed by a real estate data set, is what is the price? As such, median\_house\_value was as the selected as the outcome Y variable.

## Predictor variables selection

The dataset includes numerous options for predictor variable selection including the different numeric variables and the categoric variable for ocean proximity.

### Answering the questions

The team began the project by proposing several questions:

* Is location the best predictor of housing prices?
* Is income the best predictor of housing prices?
* Does urban density best predict housing prices?
* Would using all variables produce the best results?

Variables were selected and models were constructed to evaluate these questions

### Stepped Approach

The team also tried another approach. Numeric variables were ordered by their coefficients of correlation and determination. Models were constructed using stepped variable selection where all variables were selected, and then one by one each variable was removed from the selection in ascending order.

|  |  |  |
| --- | --- | --- |
| variable | R | R2 |
| median\_income | 68.8% | 47.3% |
| rooms\_per\_house | 15.2% | 2.3% |
| latitude | -14.4% | 2.1% |
| housing\_median\_age | 10.6% | 1.1% |
| households | 6.6% | 0.4% |
| longitude | -4.6% | 0.2% |
| bedrooms\_per\_house | -4.3% | 0.2% |
| population\_per\_house | -2.4% | 0.1% |

## Train and Test Data Sets

The createDataPartition function in the Carat library was used to segment the data into a train and test set. The data was segregated on outcome variable, median\_house\_value and a ratio of 2/3 into the train set and 1/3 into the test set.

Train datasets were utilized in model construction. Test dataset were used to predict results.

## Measuring Prediction Results

Prediction results were compared to the actual results in the test data set. Two measures were selected to evaluate the quality of predictions:

* Root mean square error
  + Measures the square root of variance of the residuals
  + A measure of how accurately the model predicts the actual outcome result
  + Root mean square error can be calculated using the formula below

Text, application

Description automatically generated

* + The code block below demonstrates the RMSE calculation

Text

Description automatically generated

* R2 predict vs test
  + Measures the relative fit of the model
  + The code block below demonstrates the R2 calculation



## Cross Validation

Train feature of the model was adjusted to utilize the k-fold cross validation. The models were executed again with these revised settings and RMSE and R2 values were measured.

The code block below demonstrates the use of the k-fold cross validation.

Graphical user interface, text

Description automatically generated with medium confidence

| Model Name | Predictor Variables | Adjusted R2 All data | Adjusted R2 Train data | RMSE  predict vs test | R2  predict vs test | Adjusted R2  Train data  K fold | RMSEpredict vs test K fold | R2  predict vs test  K fold |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| All | latitude  longitude  population\_per\_house  rooms\_per\_house  bedrooms\_per\_house  households  housing\_median\_age  median\_income  ocean\_factor | 0.6174 | 0.6238 | 7.238 | 0.6027 | 0.6528 | 7.00 | 0.6283 |
| Location | latitude  longitude  ocean\_factor | 0.27 | 0.2679 | 9.784 | 0.2739 | 0.2679 | 9.784 | 0.2739 |
| Income | median\_income | 0.4734 | 0.4587 | 8.526 | 0.4486 | 0.4857 | 8.5268 | 0.4486 |
| Density | population\_per\_house  rooms\_per\_house  bedrooms\_per\_house  households | 0.0955 | 0.1048 | 11.050 | 0.078 | 0.1048 | 11.050 | 0.078 |
| All | latitude  longitude  population\_per\_house  rooms\_per\_house  bedrooms\_per\_house  households  housing\_median\_age  median\_income  ocean\_factor | 0.6174 | 0.6238 | 7.238 | 0.6027 | 0.6528 | 7.00 | 0.6283 |
| Step 1 | median\_income  rooms\_per\_house  latitude  housing\_median\_age  households  longitude  bedrooms\_per\_house  population\_per\_house | 0.6057 | 0.6128 | 7.356 | 0.589 | 0.6128 | 7.356 | 0.589 |
| Step 2 | median\_income  rooms\_per\_house  latitude  housing\_median\_age  households  longitude  bedrooms\_per\_house | 0.6048 | 0.6118 | 7.36 | 0.588 | 0.6118 | 7.36 | 0.588 |
| Step 3 | median\_income  rooms\_per\_house  latitude  housing\_median\_age  households  longitude | 0.6018 | 0.6071 | 7.34 | 0.59 | 0.6071 | 7.34 | 0.59 |
| Step 4 | median\_income  rooms\_per\_house  latitude  housing\_median\_age  households | 0.5308 | 0.5402 | 8.043 | .0509 | 0.5402 | 8.043 | .0509 |
| Step 5 | median\_income  rooms\_per\_house  latitude  housing\_median\_age | 0.5189 | 0.5297 | 8.17 | 0.4937 | 0.5297 | 8.17 | 0.4937 |
| Step 6 | median\_income  rooms\_per\_house  latitude | 0.4857 | 0.5009 | 8.52 | 0.45 | 0.5009 | 8.52 | 0.45 |
| Step 7 | median\_income  rooms\_per\_house | 0.4794 | 0.4954 | 8.59 | 0.441 | 0.4954 | 8.59 | 0.441 |

## Comparing Model Performance

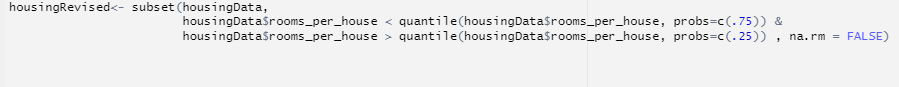
### R2 values by Model Name

### RMSE values by Model Name

# Experimentation

## Will removing outlier from the data set improve the accuracy of the model?

As seen in the box plot below, the variable rooms\_per\_house, has significant outliers. The data set was transformed using the code below to remove outliers from the rooms\_per\_house variable.



A picture containing scatter chart

Description automatically generated

The transformation removed approximately 10,000 records leaving 10319 observations. After the transformation the data took on the shape of normal distribution.

Chart, histogram

Description automatically generated

This new data set was segment between train and test sets using the same parameters as earlier. The train data set was used to retrain the model using all variables. The test data set was used to predict results and RMSE and R2 values were measured against the predicted values.

The results degraded when outliers were removed from the dataset

* Adjusted R2 on train data – 0.5622
* RMSE on predict vs test data – 7.421
* R2 on predict vs test data – 0.5888

# Conclusion

This project began by posing several questions all which strived to uncover best predictors California housing prices.

In summary, we asked:

* Is location is the best predictor?
  + No, we found location could only explain 27% of variability in median\_house\_value. However, examining the coefficients of this model, 2 variables have p-values greater than 0.05 which makes them statistically insignificant.

Text

Description automatically generated

Further, we found predictions of based on location had the 2nd worse error score, measured by RMSE.

* Is household income the best predictor?
  + No, the median\_income variable looked promising due high value of correlation coefficient, 68.8% and the clear linear relationship between median\_income and median\_house value.

Chart, scatter chart

Description automatically generated

The model was able to explain 45% of variability of median\_house\_value. Furthermore, the quality of predictions generated by this model of 8.526 RMSE was bested 6 other variable selections.

* Is population density the best predictor?
  + No, density could only account for approximately 10% of change in median\_house\_value. It seems intuitive that higher density would drive higher demand and therefore higher prices, however; the results of this analysis does not support that conclusion. Unfortunately, the measure of error in the predictions generated by this model was 11.05, the highest RMSE score observed in this analysis.

Chart, scatter chart

Description automatically generated

* Would a stepped approach to variable selection work best?
  + No, but it is worth noting the ocean\_factor variable was not included in this model. The stepped approached did demonstrate clear a degradation of results as each variable was removed.

|  |  |  |  |
| --- | --- | --- | --- |
| Step | Variable Count | Train R2 | RMSE |
| Step 1 | 8 | 0.6128 | 7.356 |
| Step 2 | 7 | 0.6118 | 7.36 |
| Step 3 | 6 | 0.6071 | 7.34 |
| Step 4 | 5 | 0.5402 | 8.043 |
| Step 5 | 4 | 0.5297 | 8.17 |
| Step 6 | 3 | 0.5009 | 8.52 |
| Step 7 | 2 | 0.4954 | 8.59 |

* Would using all available variables produce the most accurate prediction?
  + **Yes.** The best results produced in this analysis used all available variables. The results were further improved by incorporating k fold cross validation techniques. However, one variable, ocean\_factor NEAR\_BAY was consistently observed as not being statistically significant.
  + Using the train data set this model explained of 65.28% of the change in median\_house\_value. Text

    Description automatically generated
  + Predictions generated from test data set, produced the lowest RMSE value observed in this analysis. This indicates the prediction generated by this model had the fewest errors of all models considered in this project.

Text

Description automatically generated

* Would more accurate results be produced if the outliers were removed?
  + No. A data transformation was executed to remove the tail from the rooms\_per\_house variable. This transformation removed nearly 50% of the observations from the data set.
  + Using the train data set created after this transformation and the model utilizing all variables, an Adjusted R2 score of 56.22% was observed
  + Predictions generated from test data set, produced the lowest RMSE value observed in this analysis. This indicates the prediction generated by this model had the fewest errors of all models considered in this project.

In conclusion, the most accurate model produced in this analysis used all available variables to predict median\_house\_value. The results were further improved when k fold cross validation techniques were introduced.

* Adjusted R2 on train data – 0.6528
* RMSE on predict vs test data – 7.00
* R2 on predict vs test data – 0.6283

# Next Steps - Continuing the Analysis

The project team considered how the continue this analysis and further improve the accuracy of model predictions. The team would take the following next steps:

* Removing ocean\_factor NEAR BAY variable from dataset
* Execute RESET test for linearity regression assumption which may lead to data transformation on median\_income
* Combine corelated variables – rooms, bedrooms and population variable
* Experimentation with additional Machine Learning and modeling techniques

# Sources

<https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html>

<https://www.kaggle.com/datasets/camnugent/california-housing-prices>

<https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/>

<https://en.wikipedia.org/wiki/Root-mean-square_deviation>

Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow by Aurelien Geron

# Appendix 1 – R Notebook

Please see file: IST687\_FinalProject\_GroupG.html