Data Exploration and Median House Value Prediction Model for 1990, California census data

Guru Prakash Pulipati

Contents

Project Overview 3

Goal 3

Hypothesis 3

Business Question 3

Data Set 3

Variables/Column Names & Description 3

Variables 3

Detail Summary () of the data set 4

Exploratory Data Analysis (EDA) and Data Visualization 4

Exploring the Data through Visualization 4

Distribution of Variables 5

Histogram plots for all the numeric variables exits in the housing data set 5

Analysis from the Histogram plots 6

Boxplots for all the numeric variables exit in the housing data set 6

Boxplots with factor variable ocean proximity 8

Data Transformation 9

Correlation Analysis on the Transformed Data Set 9

Machine Learning 10

Training and Test Set Preparation 10

Training set (70%) 10

Test set (30%) 10

Code Snippet 11

Supervised Machine Learning - Regression (randomForest ()) 11

Code Snippet 11

Model Performance Evaluation 12

Calculating Training Set RMSE 12

Making Predictions on the Test Set 12

Code Snippet 12

Calculating Test Set RMSE 13

Code Snippet 13

Comparing Training and Test Set RMSE 13

Feature Importance Analysis with varImpPlot () 13

Conclusion - Business Answer 14

Project Overview

Goal

This project aims to develop a predictive model for median house values in California using a California Housing dataset.

Hypothesis

We hypothesize that certain socio-economic factors, such as median income, and geographic factors, like proximity to desirable amenities, significantly influence house prices.

Business Question

Would it be possible to predict the median house value in California's census block groups by leveraging various socioeconomic and geographic factors?

Data Set

This data set, built using the 1990 California census data, appeared in a 1997 paper titled "*Sparse Spatial Autoregressions*" by Pace, R. Kelley, and Ronald Barry, published in the Statistics and Probability Letters journal. It contains one row per census block group (the smallest geographical unit for which the U.S. Census Bureau publishes sample data, typically with a population of 600 to 3,000 people). I.e. Each row pertains to a group of houses representing medians for groups of houses in close proximity.

Variables/Column Names & Description

Data Set name: Referred to the as housing data set in this project

Variables

longitude: Longitude of the block group.
latitude: Latitude of the block group.
housing_median_age: Median age of the houses in the block group.
total_rooms: Total number of rooms in the block group.
total_bedrooms: Total number of bedrooms in the block group.
population: Population of the block group.
households: Number of households in the block group.
median_income: Median income of the households in the block group.
median_house_value: Median house value of the block group (target variable).
ocean_proximity: Proximity to the ocean, categorical variable with values "<1H OCEAN","INLAND", "ISLAND", "NEAR BAY", "NEAR OCEAN"

Detail Summary () of the data set

```
summary(housing)
#  longitude          latitude      housing_median_age  total_rooms     total_bedrooms
#  Min.   :-124.3   Min.   :32.54   Min.   : 1.00       Min.   :    2   Min.   :   1.0
#  1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00       1st Qu.: 1448   1st Qu.: 296.0
#  Median :-118.5   Median :34.26   Median :29.00       Median : 2127   Median : 435.0
#  Mean   :-119.6   Mean   :35.63   Mean   :28.64       Mean   : 2636   Mean   : 537.9
#  3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00       3rd Qu.: 3148   3rd Qu.: 647.0
#  Max.   :-114.3   Max.   :41.95   Max.   :52.00       Max.   :39320   Max.   :6445.0
#  NA's   :207
#     population      households     median_income     median_house_value   ocean_proximity
#   Min.   :    3   Min.   :   1.0   Min.   : 0.4999   Min.   : 14999     <1H OCEAN :9136
#   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634   1st Qu.:119600     INLAND    :6551
#   Median : 1166   Median : 409.0   Median : 3.5348   Median :179700     ISLAND    :   5
#   Mean   : 1425   Mean   : 499.5   Mean   : 3.8707   Mean   :206856     NEAR BAY  :2290
#   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432   3rd Qu.:264725     NEAR OCEAN:2658
#   Max.   :35682   Max.   :6082.0   Max.   :15.0001   Max.   :500001
```

Figure 1: Summary of the raw dataset

Exploratory Data Analysis (EDA) and Data Visualization

Exploratory Data Analysis (EDA) is the initial step where we delve into the data to uncover hidden patterns, identify potential issues, and gain insights that inform our modeling strategy. It's a crucial phase that sets the stage for robust analysis. We employ various visualization techniques during EDA to explore the data from different angles.

In this project, we conducted a comprehensive EDA to understand the characteristics of our California housing dataset. We analyzed individual variables and their relationships using various visualizations.

Glimpse of the data set variables before we start the EDA process

```
Rows: 20,640
Columns: 10
$ longitude          <dbl>
$ latitude           <dbl>
$ housing_median_age <dbl>
$ total_rooms        <dbl>
$ total_bedrooms     <dbl>
$ population         <dbl>
$ households         <dbl>
$ median_income      <dbl>
$ median_house_value <dbl>
$ ocean_proximity    <chr>
```

Data types description ***<dbl>*** - Double and ***<chr>*** - Character

Exploring the Data through Visualization

Exploratory Data Analysis (EDA) heavily relies on visualization techniques to gain a deeper understanding of the data. In this project, we employed various graphical methods to explore

the distribution of variables, identify potential relationships, and uncover anomalies within the California housing dataset.

Distribution of Variables

Histograms with the mean and boxplots were used to visualize the distribution of continuous variables such as median house value, median income, and median housing age. These visualizations provided insights into the data's central tendencies, spread, and potential skewness.

Histogram plots for all the numeric variables exits in the housing data set

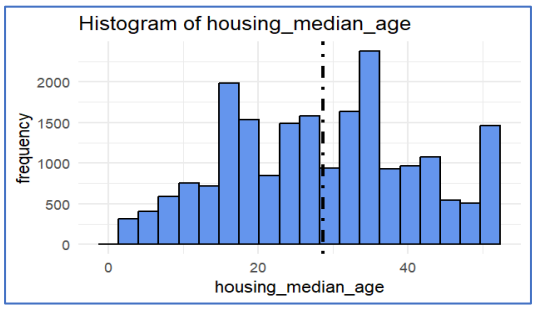

Figure 2

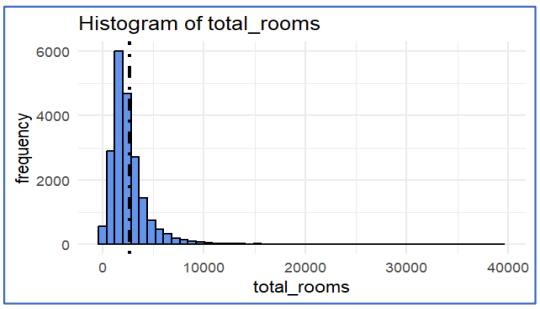

Figure 3

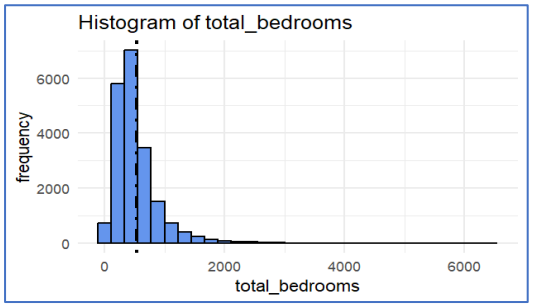

Figure 4

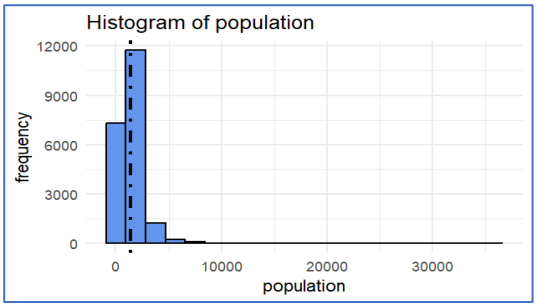

Figure 5

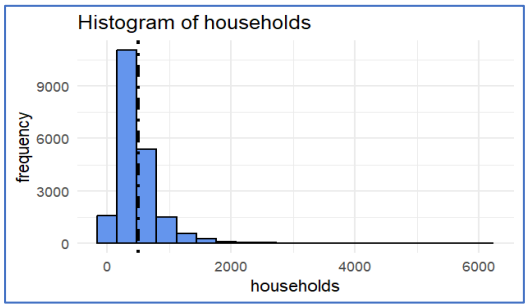

Figure 6

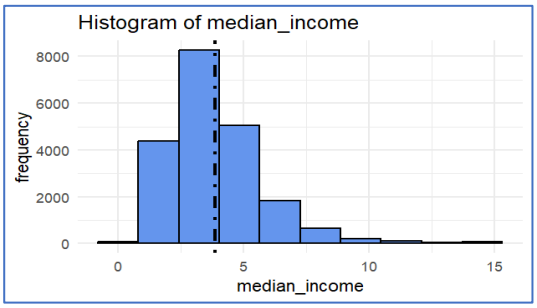

Figure 7

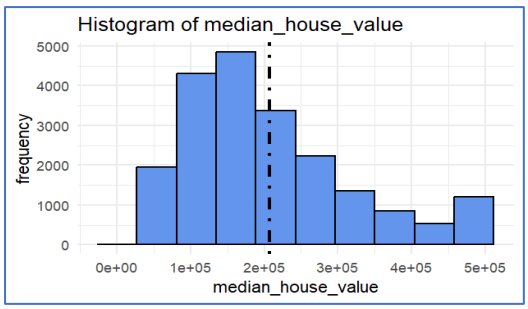

Figure 8

Analysis from the Histogram plots

'Figure 2' plotted the housing_median_age, where we could visualize the age of the houses spread across 1 to 50+ years, and as per the plot, around 40% of the houses are between 16 to 35 aged houses.

'Figure 3' & 'Figure 4' shows a trend that the majority i.e. 90% of the total rooms are within 10k and the rest of the 10% can be a very minimal count. Similarly, total bedrooms, where the maximum was below 2k.

'Figure 5' clearly shows the population across geographical units is less than 10k, where the mean is almost near 1k, as per the plot around 40% of the population could be below 1k.

'Figure 6' shows that 90% of the households across the block were less than 2k, as per the plot around 60% could be less than 500 in a block.

'Figure 7' median income on an average across the blocks was 4(normalized), where the majority income across the blocks was shown as less than 5

'Figure 8' plot shows the median house values range from 100k to 500k, where the majority were less than 250k and houses values above 400k were less.

Overall understood how the data distribution positioned across different variables in the data set.

Boxplots for all the numeric variables exit in the housing data set

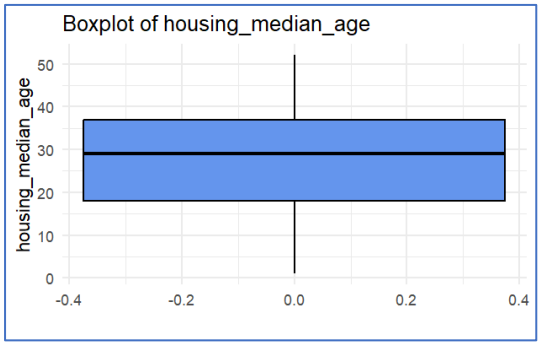

Figure 9

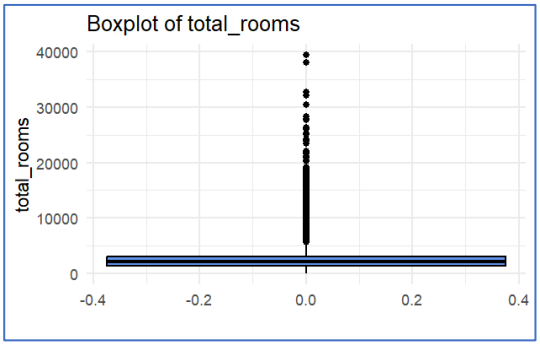

Figure 10

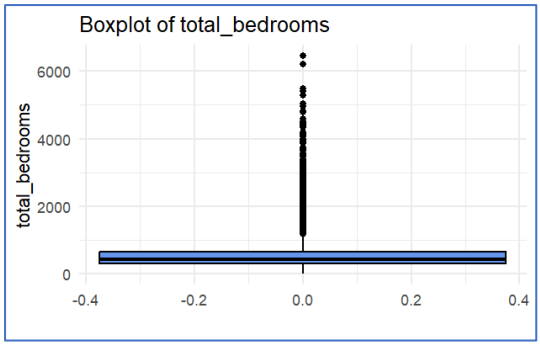

Figure 11

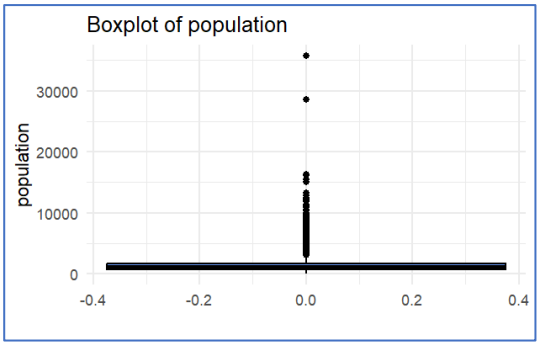

Figure 12

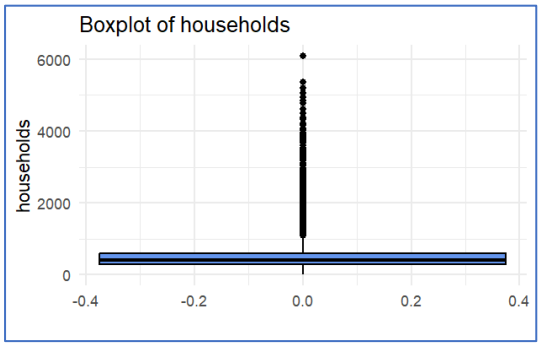

Figure 13

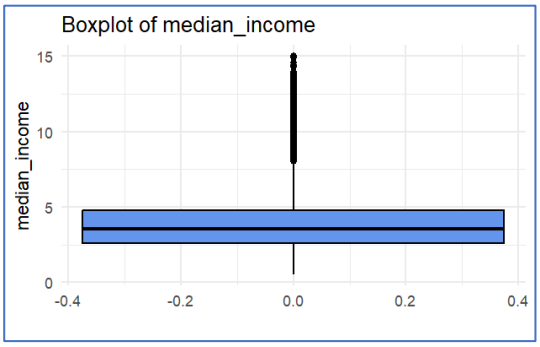

Figure 14

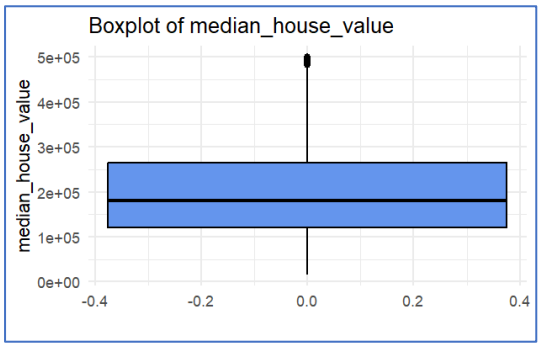

Figure 15

From *Figure 9* it is clear that 50% of the housing median age is within i.e. upper and lower quartile, where the median is inclined towards the upper quartile, and the rest of the data is evenly spread across the whiskers, cannot find any outliers.

In *Figures 10 – total rooms, Figures 11 – total bedrooms, Figures 12 – population, and Figure -13 - households,* we could observe outliers and the maximum whisker and the interquartile range are way below as observed in histogram plots. As observed skewness commonly across the data within these variables, this can affect the performance of the predictive model.

Figures 14 – median income, Figure 15 – median house value look good with 50% of the median income in the interquartile range with few outliers and maximum whisker. In median house value shows median house values majority were less than 250k. Median income also shows majority income across the blocks was less than 5.

Overall, median house value, median income, and housing median age variables data look consistent across the histogram and with boxplot i.e. with interquartile (i.e. 50%) range too.

Boxplots with factor variable ocean proximity

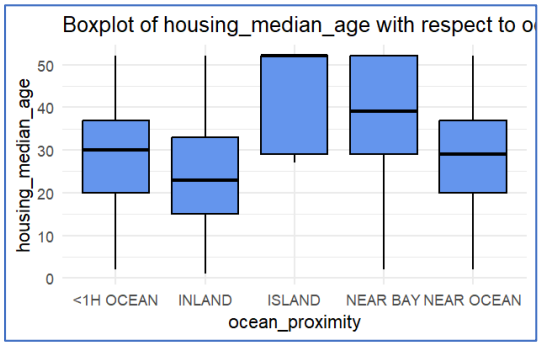

Figure 16

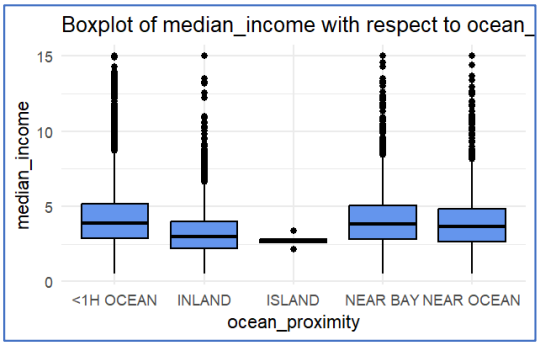

Figure 17

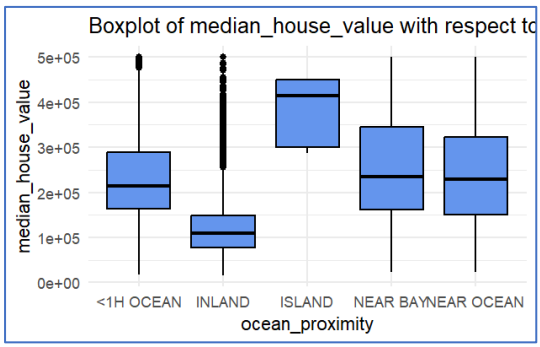

Figure 18

Boxplot with housing_median_age, median_income, and median_house_value concerning the factor variable ocean_proximity, this gives a picture of the data distribution of these variables (housing_median_age, median_income, and median_house_value) with respective to ocean proximity. *'Figure 18'* shows the median of the house value that has ocean proximity factor i.e. NEAR OCEAN, NEAR BAY and <1H OCEAN are almost the same, whereas houses with ISLAND proximity shows on the higher side of the price and proximity INLAND where the lesser price once compare to all factors.

'Figure 17' shows that 75% of the houses across all proximity were owned by households having a median income of less than 5 (neutralized) with a good number of outliers.

In *'Figure 16'* the houses with ocean proximity to the ISLAND and NEAR BAY were more than 30-year-old and the rest were spread across 16 to 37 years old.

Overall, the houses with ocean proximity to the ISLAND are most expensive and 30+ years old and 50% of them were owned by population households whose income is less than 5.

Data Transformation

In this phase, we will be transforming the raw dataset i.e. housing data set into a more refined form which will constitute to data pipeline.

- Based on the summary () details '*Figure 1*', noticed that 207 NAs in total_bedroom variable, i.e. values are missing. This is addressed by filling in the missing values using imputation. Used "statistical median" for missing total_bedrooms values. Generally, mean is used but here median is used because it's less influenced by extreme outliers. This may not be the best method, as these missing values could represent actual buildings (e.g. a warehouse) with no bedrooms, but imputation often makes the best of a bad situation.

- Ocean proximity variable i.e. ocean_proximity variable is not numerical, its factor variable, so converted the variable into a binary categorical variable consisting of 1s and 0s. Although many machine learning algorithms in R can handle categorical data stored in a factor variable, doing a split to cater to the lowest common denominator and ocean_proximity variable is removed from the data set.

- Creating and adding mean_bedrooms & mean_rooms variables using total_bedrooms and total_rooms variables as they are more accurate depictions of the house in a given group, and removing total_bedrooms and total_rooms variables.

- Performed feature scaling, scaled all numerical variables except median_house_value (as it's our response variable), and the binary categorical variables.

- Finally, with all these data transformations, achieved the final data set i.e. cleaned_housing with the following variables

"NEAR BAY" "<1H OCEAN" "INLAND" "NEAR OCEAN" "ISLAND" "longitude" "latitude" "housing_median_age" "population" "households" "median_income" "mean_bedrooms" "mean_rooms" "median_house_value"

Correlation Analysis on the Transformed Data Set

To quantify the strength and direction of relationships between multiple variables, a correlation matrix was generated. This matrix revealed a strong positive correlation between median income (median_income) and median house value(median_house_value), showing that areas with higher median incomes tend to have higher median house values and vice versa.

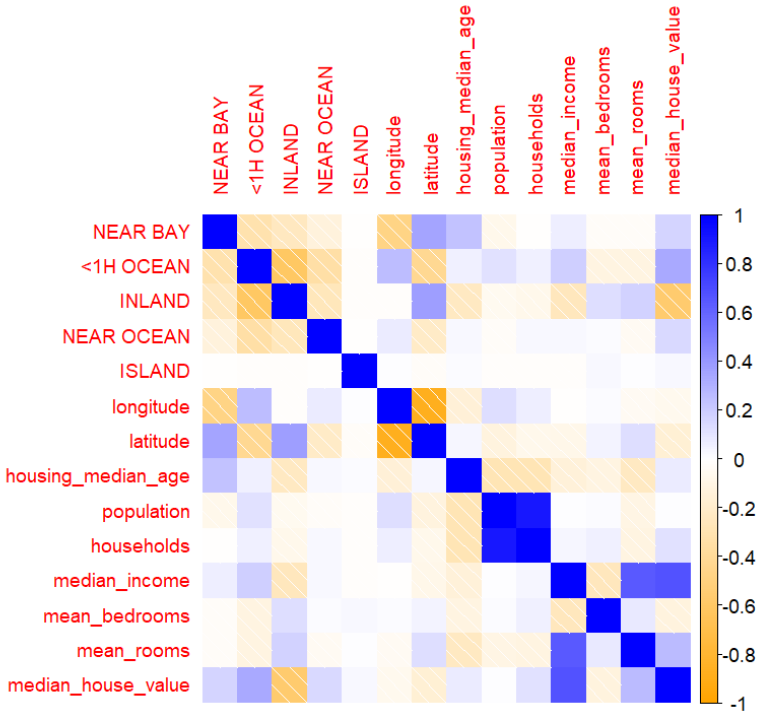

Figure 19

Machine Learning

In this phase, we will be preparing the data for training and testing the Random Forest model. It includes splitting the data into training and testing sets and implementing of supervised machine learning regression model i.e. Random Forest (randomForest ()).

Training and Test Set Preparation

To assess the model's generalization ability, we split the cleaned housing data into two sets:

Training set (70%) Used to train the Random Forest model. The model learns patterns from this data to make predictions.

Test set (30%) Used to evaluate the model's performance on unseen data. The model's predictions on the test set are compared to the actual values to assess its accuracy.

The random sample Index used for machine learning involved the creation of a sample index for the cleaned_housing data frame that involved, creating a training set named train_cleaned_housing consists 70% of the rows of the cleaned_housing data frame and 30% of the rows of the cleaned_housing as testing set named test_cleaned_housing.

Code Snippet

```
# ------------------------------------------------------------------------
# 4. TRAINING AND TEST SETS
# ------------------------------------------------------------------------
# creating
# random sample index for cleaned_housing data frame,
# training set with 70% rows
# test set with rest of the rows
n <- nrow(cleaned_housing)  # Number of observations
ntrain <- round(n*0.7)      # 70% for training set gives the min test error metric
set.seed(414)               # setting seed for reproducible results

tindex <- sample(n,ntrain)                          # sample index creation
train_cleaned_housing <- cleaned_housing[tindex,] # training set creation
test_cleaned_housing <- cleaned_housing[-tindex,] # test set creation

#validation
nrow(train_cleaned_housing) + nrow(test_cleaned_housing) == nrow(cleaned_housing)
# [1] TRUE
```

Figure 20

Supervised Machine Learning - Regression (randomForest ())

In this step, we implemented the Random Forest regression model, using the randomForest () algorithm found in the randomForest package for the training and inference. Our goal is to predict the median house value using this regression method.

First, we separated the training set train_cleaned_housing into two pieces: train_x and train_y where train_x is a data frame that has all variables except the response variable median_hous_value and train_y is a numeric vector (not a data frame) that has only the response variable values from median_house_value. If you observe we are passing data separately i.e. response variable and predictors.

Code Snippet

```
# ------------------------------------------------------------------------
# 5. SUPERVISED ML - REGRESSION - RANDOMFOREST
# ------------------------------------------------------------------------
# using randomForest algorithm for training and inference to predict
# the median house value

#training set
train_col_index <- which(colnames(train_cleaned_housing) == 'median_house_value')
train_x <- train_cleaned_housing[, -train_col_index]
train_y <- train_cleaned_housing[, train_col_index]

rf = randomForest(x=train_x, y=train_y , ntree=500, importance=TRUE)

names(rf)
# [1] "call"            "type"            "predicted"       "mse"
# [5] "rsq"             "oob.times"       "importance"      "importanceSD"
# [9] "localImportance" "proximity"       "ntree"           "mtry"
# [13] "forest"          "coefs"           "y"               "test"
# [17] "inbag"
```

the Figure 21

Model Performance Evaluation

Having trained the Random Forest model, the next phase is to assess its performance on unseen data. Here, we evaluated the model using Root Mean Squared Error (RMSE) on both the training and test sets.

Calculating Training Set RMSE

We use the model to make predictions on the training data itself. The actual median house values in the training set are compared to the predicted values from the model. The squared differences between these values are calculated and averaged. The square root of this average squared difference is the Root Mean Squared Error (RMSE) for the training set. The resulting RMSE is the prediction of the median price of a house in a given district to be within an RMSE delta of the actual median house price. The RMSE result was 49728.94.

Code Snippet

```
# calculating the set root mean squared error (RMSE)
rf
# Call:
#   randomForest(x = train_x, y = train_y, ntree = 500, importance = TRUE)
# Type of random forest: regression
# Number of trees: 500
# No. of variables tried at each split: 4
#
# Mean of squared residuals: 2472967347
# % Var explained: 81.52

train_rmse = sqrt(rf$mse[length(rf$mse)])
train_rmse
# [1] 49728.94
```

Figure 22

Making Predictions on the Test Set

In this step, we will see the model making predictions by using the test set by splitting the test set in the same manner as the training set in the previous step i.e. creating a new data frame test_x and the numeric vector test_y. Used predict () function using the trained model 'rf' along with the test_x to calculate a vector of predicted median house values.

Code Snippet

```
#predictions by using the test set
test_col_index <- which(colnames(test_cleaned_housing) == 'median_house_value')
test_x <- test_cleaned_housing[, -test_col_index]
test_y <- test_cleaned_housing[, test_col_index]

y_pred <- predict(rf, test_x)
```

Figure 23

Calculating Test Set RMSE

In this step, we compared the test RMSE with the training set RMSE, where the model scored roughly the same on the training and testing data, suggesting that it is not overfitting and this makes good predictions. The RMSE result was 48148.52

Code Snippet

```
# test set RMSE
test_rmse = sqrt(mean((test_y - y_pred)^2))
test_rmse
# [1] 48148.52

# the predictions of the median price of a house is within
# ~$49000.00 of the actual median house price.
```

Figure 24

Comparing Training and Test Set RMSE

Ideally, the training and test set RMSE values are reasonably close. i.e. training set RMSE is 49728.94 and testing set RMSE is 48148.52.

The training and test set RMSE are almost similar, it suggests the model is capturing the underlying patterns in the data without overfitting to the specifics of the training set. This implies the model might make good predictions on new, unseen data.

Feature Importance Analysis with varImpPlot ()

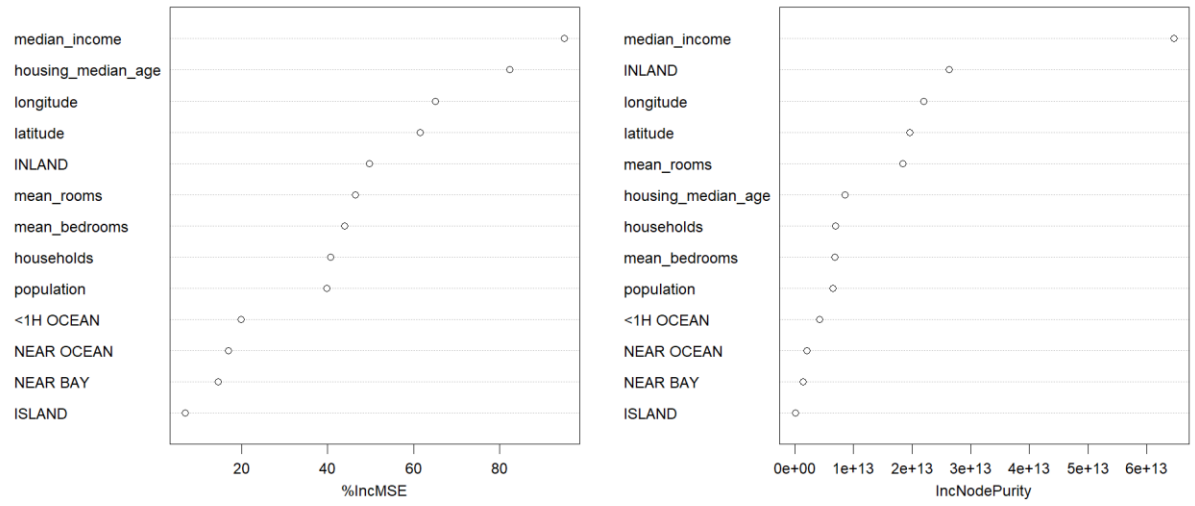

Figure 25

Running varImpPlot () confirms median_income as the most important feature, aligning with the strong correlation observed in the data. This strengthens the validity of both analyses.

Conclusion - Business Answer

Analysis using a Random Forest model shows promising results for predicting median house prices in California. The model, despite its relative simplicity (using around 500 trees), can estimate median house values within ~$49,000 of the actual price. This establishes a benchmark for further exploration.