CLASS PROJECT: UNDERSTANDING HOUSING PRICES

Using a random forest model to make accurate predictions of house prices.

Jessica Krook

Introduction to Data Science COM SCI X 450.1

Table of Contents

1. Project overview: goals and hypothesis and the business question

Goals

The goal of the project is to create a model that can predict housing prices based on other variables. Of these variables, we want to identify which have the most influence on the model.

Hypothesis

A combination of geographical, economic, and demographic factors will have a significant impact on the median value of a house.

Business Question

Is it possible to create an accurate predictive model for housing prices based on geographical, economic, and demographic factors?

2. Description of the data set and variables

The dataset has been taken from the 1990 California census data and includes housing information of different locations in California. The variables are both numeric (coordinates, age, room counts, and income) and categorical indicating (ocean proximity). Each row of data is a census block group which makes up a population size of 600 to 3,000 people.

Variables in the dataset:

1. **longitude**: East-west geographic coordinate
2. **latitude**: North-south geographic coordinate
3. **housing_median_age**: The median age of households in the census block group.
4. **total_rooms**: The total number of rooms in the census block group.
5. **total_bedrooms**: The total number of bedrooms in the census block group.
6. **population**: The total population in the census block group.
7. **households**: The total number of households in the census block group.
8. **median_income**: The median income of households in the census block group.
9. **median_house_value**: The median value of houses in the census block group.
10. **ocean_proximity**: The proximity of the census block group to the ocean.

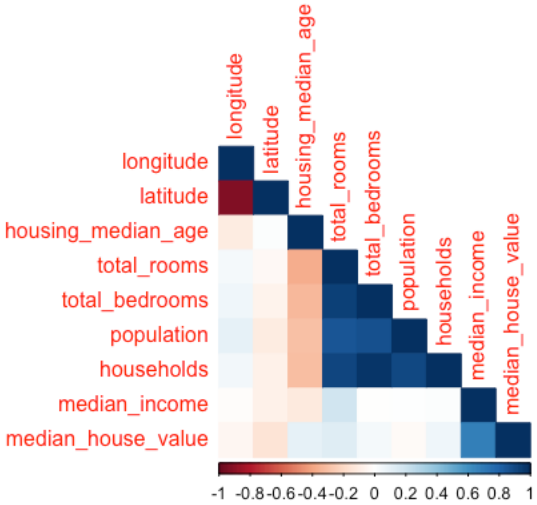
3. Discussion of EDA results and data visualization experiments

Image 1: Summary of ocean proximity

# <1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
9136	6551	5	2290	2658

A summary of the categorical variable “ocean proximity” shows most census blocks are less than 1 hour from the Ocean and only 5 census blocks are on an island.

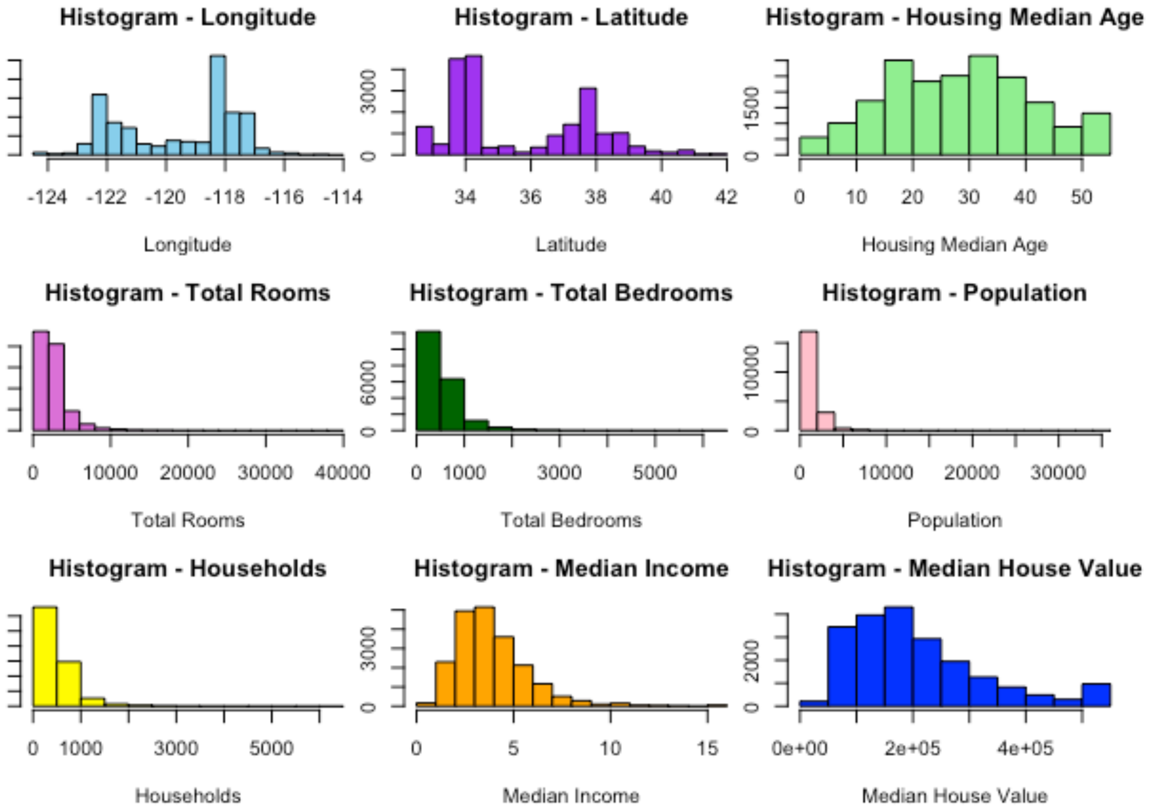
Image 2: correlation matrix



When looking at the correlation between variables we can see that there is a strong correlation between population and households.

There is also a strong correlation between households and total_bedrooms

Image 3: Numeric variable distributions



Age:

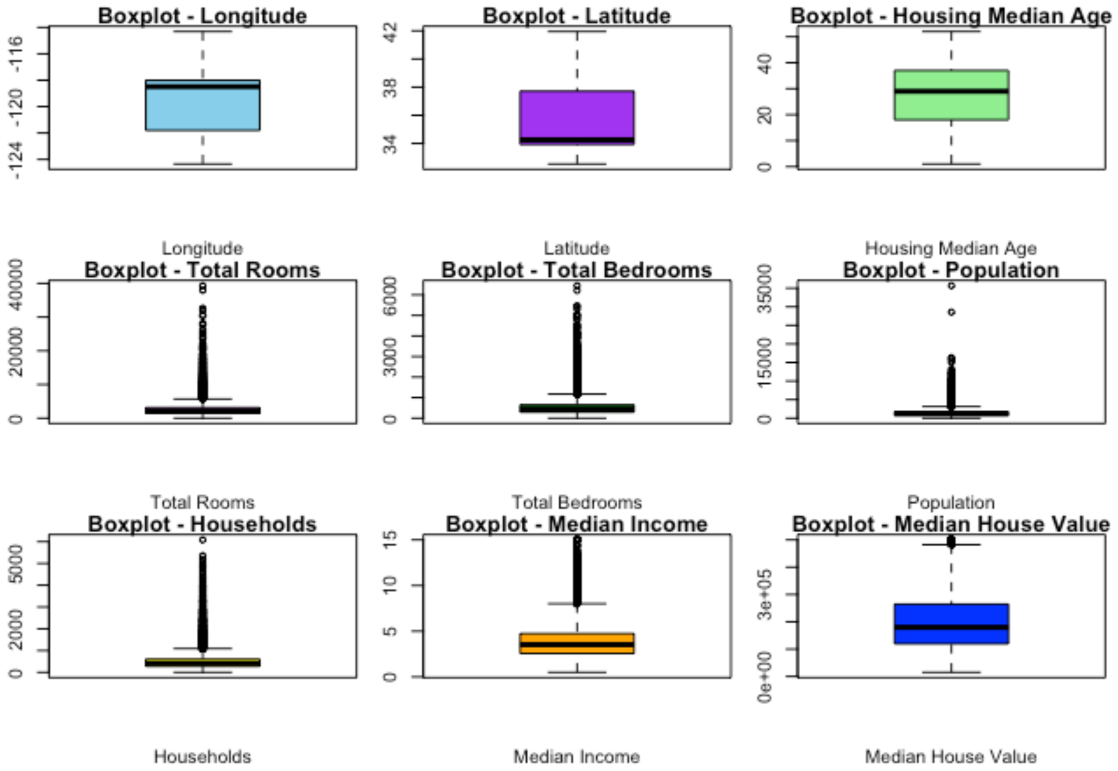
This histogram copies a general age population so we can accept it – you can see a spike from 15 -20 for teenagers- and young adults and you can see there is another spike at 30-35.

Skewed data:

We can see for total_rooms, total_bedrooms, population, and households the data is skewed right. The shape of the histogram indicates there are some extreme outliers. We will decide later how we will normalize this data.

The median_house_value is also skewed right with what looks like outliers for more expensive houses.

Image 4: Boxplots of the numeric variables



Even distribution:

Median Housing age is relatively evenly distributed, and skews left a little. There are no outliers

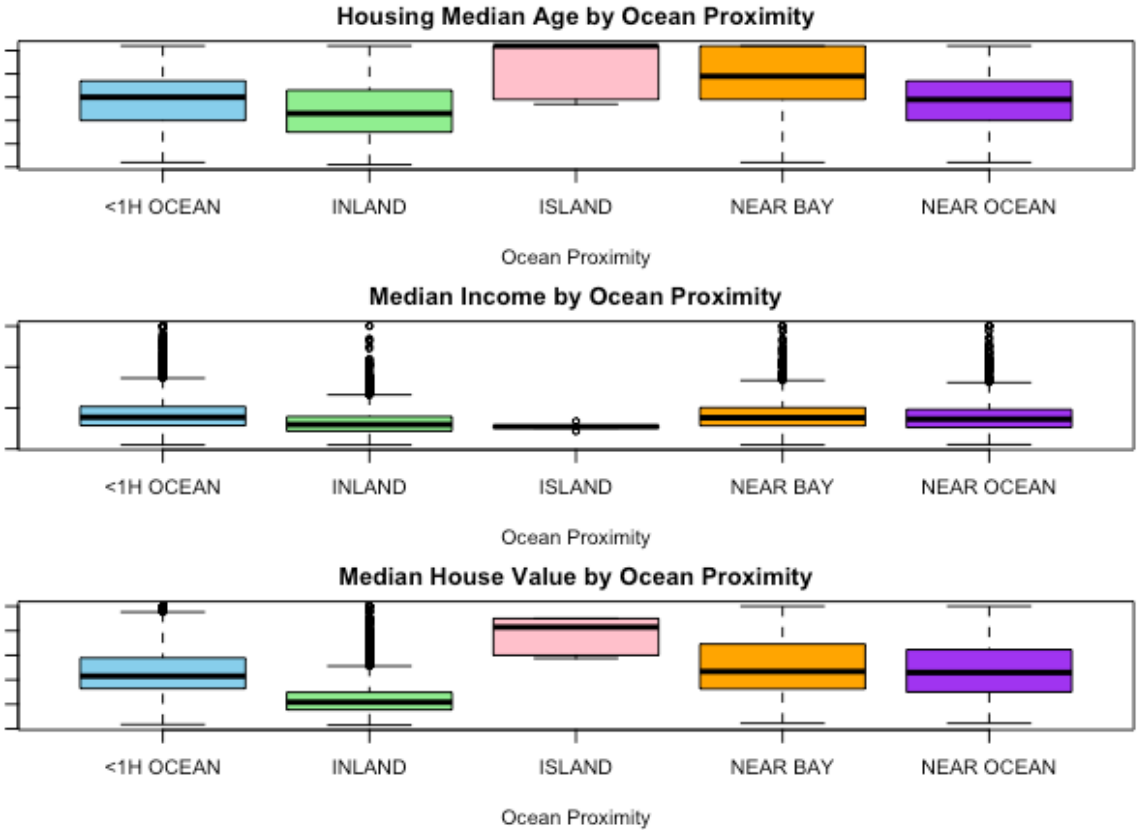
Heavily skewed data with outliers:

Total_rooms, total_bedrooms, population, households, and median_income have a lot of outliers and skew right heavily. We will need to normalize this data.

Keeping certain outliers:

Median_house_value skews right and has a few outliers, however, because the value represents 4% of the data, I have decided to keep it as removing it could result in losing the full picture of the higher-value houses in California.

Image 5: Boxplots of numeric variables relative to ocean proximity:



Median Age:

The housing_median_age by ocean proximity largely looks similar to the total median age boxplot. The two exceptions are "ISLAND" and "NEAR BAY". "ISLAND" is skewed left; its youngest median age is 30 of its oldest is 50. "NEAR BAY"'s IQR is between 30 and 50, meaning it is on average older than the general median age distribution but also includes some younger residents.

Median Income:

Most median income boxplots look similar and have a similar distribution to the general median income boxplot. However, "ISLAND" has a very small range with outliers on both sides. Its small range is likely due to there being a small population within the island category.

Median House Value:

The median house values vary widely between different ocean proximities. "<1H OCEAN", "NEAR BAY", and "NEAR OCEAN" follow the general median house value boxplot. however, "INLAND" and "ISLAND" follow different distributions. "INLAND" skews left and has a large number of outliers. Its upper quartile is also less than all the lower quartiles of "<1H OCEAN", NEAR BAY", and "NEAR OCEAN". "ISLAND" on the other hand is skewed left and has a higher IQR than all the other areas.

4. The data pipeline and data munging steps

The first step I took in cleaning the data was to address the NA values in total_bedrooms. They need to be replaced with a value in order to analyze the data effectively. I chose to calculate the median value to replace the NAs and used impute to replace the NA values with it.

My steps:

1. I loaded the library(e1071) so I could use the impute() function
2. I then used impute() to calculate the median of the total_bedrooms column and replace the NAs with the median
3. I then replaced the variables in the original column in the dataframe with the new variables including the median where NAs were

In order to run a machine learning algorithm, I needed to split the categorical variable “ocean proximity” into binary categorical variables of 1’s and 0’s.

To do this I:

1. Created binary categorical variables for each type of ocean proximity
2. I then renamed the new columns, so they appeared as "<1H OCEAN", "INLAND", "ISLAND", "NEAR BAY", "NEAR OCEAN"
3. Then I used cbind to add the new variables to the original data frame
4. Lastly, I removed the “ocean_proximity” column from the data set.

Next, I created two new variables using the mean of total_bedrooms and total_rooms. I did this because they will give more accurate depictions of houses in a group.

My steps:

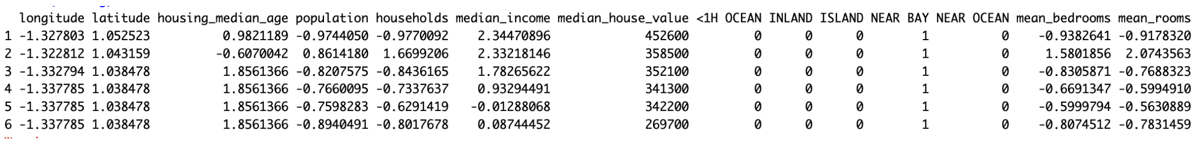
1. Used ave() to calculate the mean of total_bedrooms and rooms and created a new variable for each called mean_bedrooms and mean_rooms. I defined the means by households.
2. I then removed the total_bedrooms and total_rooms variables from the data frame

Lastly, I performed feature scaling as the dataset’s variables all have different ranges. If I did not perform feature scaling, this could lead to a biased prediction model.

To do this I:

1. Specified the numerical columns I wanted to scale
2. Performed the z-score standardization using R’s built-in scale() function

My data munging resulted in the dataframe looking like this:



```
  longitude latitude housing_median_age population households median_income median_house_value <1H OCEAN INLAND ISLAND NEAR BAY NEAR OCEAN mean_bedrooms mean_rooms
1 -1.327803 1.052523          0.9821189 -0.9744050 -0.9770092    2.34470896             452600         0      0      0        1          0    -0.9382641 -0.9178320
2 -1.322812 1.043159         -0.6070042  0.8614180  1.6699206    2.33218146             358500         0      0      0        1          0     1.5801856  2.0743563
3 -1.332794 1.038478          1.8561366 -0.8207575 -0.8436165    1.78265622             352100         0      0      0        1          0    -0.8305871 -0.7688323
4 -1.337785 1.038478          1.8561366 -0.7660095 -0.7337637    0.93294491             341300         0      0      0        1          0    -0.6691347 -0.5994910
5 -1.337785 1.038478          1.8561366 -0.7598283 -0.6291419   -0.01288068             342200         0      0      0        1          0    -0.5999794 -0.5630889
6 -1.337785 1.038478          1.8561366 -0.8940491 -0.8017678    0.08744452             269700         0      0      0        1          0    -0.8074512 -0.7831459
```

5. Statistical model - Random Forest and the feature vector and response variable.

This Random Forest model aims to predict the median house value based on a set of features, providing a robust and accurate prediction for housing prices in a given district. It does this by using a learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees.

The Random Forest model was trained on the `cleaned_housing` dataset to predict the binary class labels based on the `median_house_value`.

Feature Vector:
The feature vector in this Random Forest model includes several predictor variables, for my first attempt at a Random Forest I used all the columns in the dataset except median_house_value.

Response Variable:
The response variable is what we are trying to predict, therefore in this model, we used the median_house_value.

6. The performance metric results for the model

Model Evaluation:

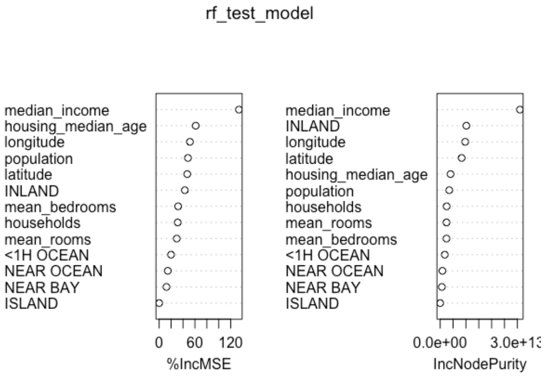
I used the Root Mean Squared Error (RMSE) on both the training set and the testing set. The performance of the Random Forest model was evaluated using Root Mean Squared Error (RMSE). The RMSE for the training and test sets should be very similar if the model's predictions are good.

After running the RMSE on the first model:

- The RMSE for the model on the training set was: 50022.14
- The RMSE for the model on the test set was: 51684.38

I decided this difference was too big (1662.24) so wanted to refine the model. I did this using the Variable Importance Plot.

Image 6: Variable Importance plot of the first model



<u>Second feature selection:</u>

Based on the Variable Importance Plot, I chose a subset of important features with significance above 40 to retrain the model:

The new variables I used are:

- longitude
- latitude
- housing_median_age
- population
- median_income
- INLAND

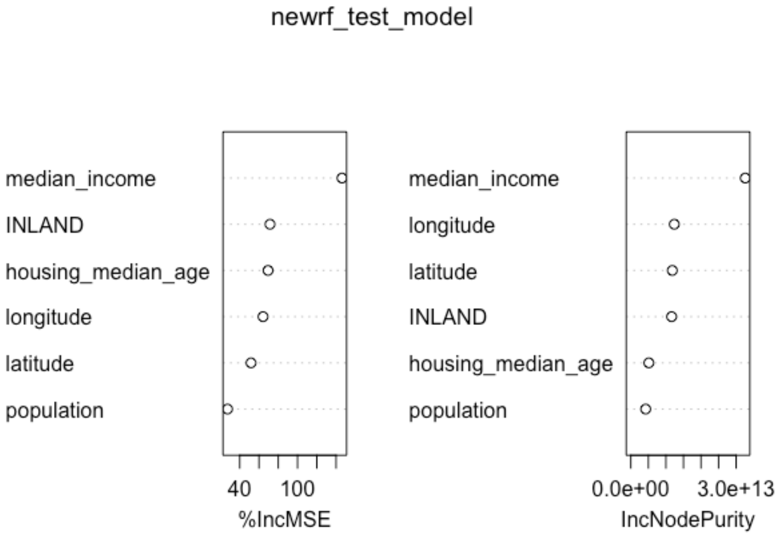
I then re-ran the Random Forest model using the new subset of data

After running the RMSE on the first model:

- The RMSE for the model on the training set was: 50022.14
- The RMSE for the model on the test set was: 50995.03

This was much closer (1022.85), and I was happier with this result as a similar performance between training and test sets means that the model is not overfitting and generalizes well to new data

Image 7: Variable Importance Plot of the Second Model



We can see that median_income is still the most influential variable along with geographical variables.

I ran a third model where I removed population, but this model was subject to overfitting, so I decided this is the optimal model for predicting housing prices.

7. The "Business Answer"

Business Answer

The Random Forest model successfully addresses the central business question: "Is it possible to create an accurate predictive model for housing prices based on geographical factors and demographics?" as the results show the model is a reliable tool for predicting median house values.

Key Findings

- Key Drivers of Housing Prices:

A Variable Importance Plot revealed that latitude and longitude, a house being classified as "INLAND", population, households, and median income are influential in determining housing prices. This shows that geographical, economic, and demographic indicators and demographics significantly influence median house prices.

- Complex Relationships:

The Random Forest algorithm considers complex relationships and non-linearities in the data, which provides a robust understanding of the factors influencing housing prices.

- Valuable Insights for Stakeholders:

Stakeholders can use the model to make informed decisions related to pricing and market trends. By understanding which features are more important, stakeholders can adapt strategies quickly when the market changes.

Business Implications

The successful creation of the model makes housing price predictions more accurate. It also provides actionable insights for stakeholders as they can use the model to improve buying and selling strategies and identify locations with good investment opportunities. The model empowers stakeholders to make data-driven decisions in the dynamic real estate market.