California Housing Price Prediction Model

Petya Buchkova, Bogdan Iulian, Lei Xian, Tibi Mitran

March 19, 2019

Abstract. Based on our hypothesis of the `ocean_proximity` attribute is not as important, we take out the attribute to rediscover the correlation among other attributes, as well to make the housing prediction model more general applicable.

1 Introduction

The aim of this report is to present the research and analysis conducted on an already existing dataset. In Aurelien Géron's machine learning project[1], based on the California Housing dataset, we were introduced to the different approaches and steps of building a model of housing prices, based on a median income attribute. However, we decided to take a different approach on predicting the median house value. With the help of a matrix, we want to explore the relation between house price and location, more specifically its proximity to the ocean.

2 Research Question

Based on the California Housing dataset, does a correlation exist between the ocean proximity and the median housing price?

3 Method

In the following section, we describe the methods used to help us answer the research question.

3.1 Exploratory Data Analysis

Using this method, we analyzed the dataset by looking at the different fields and how they would be of use for our research.

3.2 Visual Data Analysis

We visually analyzed the data in order to form a hypothesis on our research topic.

3.3 Data Cleaning

With the data cleaning process, we identified the incompatible data in our dataset and transformed it to fit our research question needs.

4 Analysis

Deriving from Géron's machine learning project[1], we will begin by analyzing the structure of the California Housing dataset.

4.1 Exploratory Data Analysis

In order to get insights on the dataset we have at disposal we performed an exploratory analysis on the dataset. Using Jupiter notebook we prepared the workspace by importing the following python modules which will serve as tools for the project: NumPy, Pandas, Matplotlib and Scikit-learn. With the dataset already in possession we used Pandas `read_csv` function to load the dataset into the workspace. Using the `head()` function, we got a closer look on the attributes present in the dataset as well as the first five observation and made a first impression on the data we will work with.

```
housing = load_housing_data()
housing.head()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

Figure 1: Dataset overview

We observed 10 attributes from which 9 holding numerical values and one, text value, precisely the attribute we are interested in. Af-

ter a quick look over the observations we noticed that there are repetitive values for the `ocean_proximity` attribute which means that it's probably a categorical attribute.

For a closer look on the content of the dataset we used the `info()` method which returned the total number of observations, the attributes type and the number of null values in the dataset. This gave us a more clear idea of the data we will work with and outline important facts for example, the `total_bedrooms` attribute is missing values for 207 observations.

Since the `ocean_proximity` was the only non-numerical attribute, it was necessary to get more information about how it relates to the total number of the observations. Using `value_counts()` function on the attribute we could see how many observations belong to each value. Also, we noticed that there are five possible values for it.

4.2 Data visualization

We can make use of the geographical coordinates present in the dataset to visualize the density of the houses in the California area in order to understand how they are spread, as shown in Figure 2

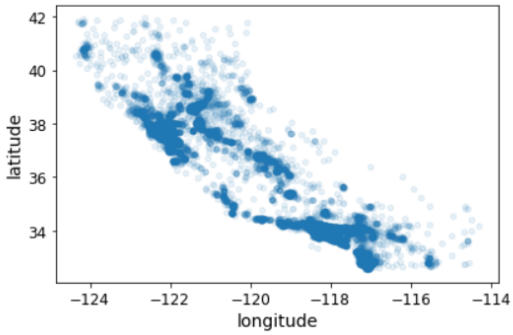

Figure 2: Density of houses

We can clearly see the highest densities are around the bay area and around big cities (Los Angeles, San Diego, Sacramento and Fresno) as well as a nice dense line around central valley. We believed that the data visuals were based on the author's assumption and therefore influencing the direction of the research therefore we decided to explore a different correlation between attributes. Looking at Figure 3 and how the median income is spread compared to the population density, we noticed the same pattern as before.

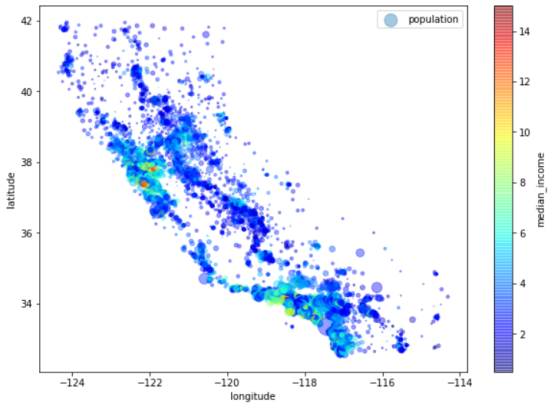

Figure 3: Median income

The areas with the highest median income are matching the areas with the highest density of population which makes us believe that there is a strong correlation between these two. In fact, if we take a look at the median income and the median house value based on population we can clearly see in Figure 4 how similar the two plots are.

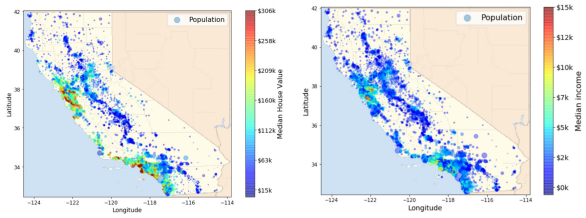

Figure 4: Visualizations comparison

Based on these facts, we formulated the following hypothesis: The ocean proximity factor should not have a high impact on the prediction and thus making the model applicable regardless of the geographical location. We observed a higher correlation between the population density and the median income therefore we believe that these attributes should play a major role in the prediction of housing prices.

4.3 Data cleaning

In the Data visualization chapter, we looked at the purpose of the ocean proximity attribute and formed a hypothesis that it can be completely disregarded from the set. In order to test our assumptions, we first need to clean up the dataset and prepare it for a machine

learning model. Therefore, we dropped the `ocean_proximity` attribute using DataFrame's `drop()` method as seen in Figure 5.

```
In [157]: sample_cleared_set = sample_incomplete_rows.drop('ocean_proximity', axis=1);
          sample_cleared_set
```

Out[157]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income
4629	-118.30	34.07	18.0	3759.0	NaN	3296.0	1462.0	2.2708
6068	-117.86	34.01	16.0	4632.0	NaN	3038.0	727.0	5.1762
17923	-121.97	37.35	30.0	1955.0	NaN	999.0	386.0	4.6328
13656	-117.30	34.05	6.0	2155.0	NaN	1039.0	391.0	1.6675
19252	-122.79	38.48	7.0	6837.0	NaN	3468.0	1405.0	3.1662

Figure 5: Drop column

Following Géron's machine learning project[1], we then proceeded with transforming the data and adding three new attributes: room and population per household, and bedrooms per room. We also standardized our data and tested the different models, covered in the machine learning project[1]. We then inspected the relevance of each attribute in the set with Listing 1 showing the results.

```
[(0.39324042349404464, 'median_income'),
 (0.11977475437785746, 'pop_per_hhold'),
 (0.10188164843322521, 'longitude'),
 (0.09941238930502191, 'latitude'),
 (0.08730795976096638, 'bedrooms_per_room'),
 (0.06333405709344203, 'rooms_per_hhold'),
 (0.055193542438925366, 'housing_median_age'),
 (0.021971208418618066, 'total_rooms'),
 (0.019646348206096002, 'population'),
 (0.019201864171771478, 'total_bedrooms'),
 (0.019035804300031437, 'households')]
```

Listing 1: Attribute relevance

Comparing our results to the machine learning project ones, we can see that the importance of the median income has increased, as well as the longitude and latitude ones. Our experiment shows that we can build a more widely applicable prediction model, by using the correlation between the median income and the location, leaving the ocean proximity out of the set.

5 Conclusion

Based on new data visualization and reproducing Geéron's machine learning project[1] steps, we were able to improve the applicability of the model, by disclaiming the strong correlation between the ocean proximity and the housing prices. Moreover, we were able to see the importance of the different attributes, once we disregard the ocean proximity.

6 Reflection

In the process of writing a scientific paper, we encountered difficulties in finding a direction for further research and a different approach to the problem. Moreover, our hypothesis would be fully confirmed, once this model is applied to datasets from diverse locations.

References

[1] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2017. ISBN: 9781491962299. URL: `http://books.google.com/books?id=W-xMPgAACAAJ`.