Project: Bathrooms classifier

Aldo N. Cao Romero

June 2022

1 Introduction

In this project we take the cleansed data we use in the Cleaning data assignment (Figure (1), 4484 rows and 12 columns) and try to classified the houses and the amount of bathrooms that it can have. In this manner make a model that can be able to tell us if house is candidate to have an additional bathroom and offer it as a service.

	sold_price	zipcode	longitude	latitude	lot_acres	taxes	year_built	bedrooms	bathrooms	sqrt_ft	garage	fireplaces
0	2000000.0	85750	-110.848679	32.321134	0.64	11322.00	2001	5	6	7471.0	3	5.0
1	1900000.0	85750	-110.843910	32.328460	1.16	16714.00	2002	4	4	5333.0	3	2.0
2	1800000.0	85750	-110.845560	32.327714	1.32	20206.00	2002	4	7	6800.0	3	4.0
3	1950000.0	85755	-110.992676	32.464204	1.18	21063.00	2002	4	6	6622.0	6	3.0
4	1920000.0	85718	-110.910653	32.338271	1.27	24316.00	1997	5	6	7132.0	3	2.0

Figure 1: Original cleansed data

2 Creation of the model

First, the model was created using the variable bathroom as a function of the size ("sqrt_ft") of the house and the number of bedrooms ("bedrooms"). Table (2).

	sqrt_ft	bedrooms	bathrooms
0	7471.0	5	6
1	5333.0	4	4
2	6800.0	4	7
3	6622.0	4	6
4	7132.0	5	6
...	...	...	...
4479	2106.0	3	2
4480	3601.0	5	3
4481	2318.0	4	3
4482	3724.0	4	4
4483	4317.0	4	4

Figure 2: Table of the variables used to create the bathroom model.

Additionally, we use the normalization:

$$X = \frac{X - \min(X)}{\max(X) - \min(X)}. \tag{1}$$

It was trained the model with the 75% of the data by using the K-Nearest Neighbors classifier. Then, once the fit is done, we predict values with the training and the validation data and obtain the accuracy for several values of neighbors. The results can be seen in figure (4).

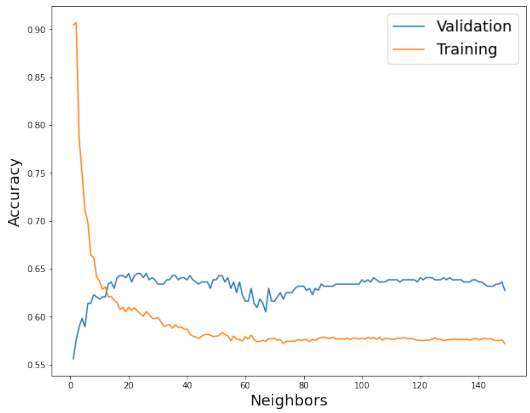

Figure 3: The accuracy as function of K-Neighbors for the validation and training data.

The intersection point of the two plots is in approximately K=14 which corresponds to an accuracy of 62%. Figure

	Neighbors	accuracy_Train	accuracy_Val
10	11	0.620536	0.628605
11	12	0.620536	0.630984
12	13	0.633929	0.620577
13	14	0.636161	0.621172
14	15	0.629464	0.617009
15	16	0.640625	0.614630
16	17	0.642857	0.607493
17	18	0.642857	0.609575
18	19	0.640625	0.605114
19	20	0.645089	0.609575

Figure 4: Table that corresponds to the accuracy for the Training and validation data as a function of the number of neighbors.

We made a prediction with the test data and obtained an accuracy of 60% for K=14 Neighbors. Even if the accuracy is not the best we wish to have, it is a good sign for our model to have similar values of it.

Finally, we show the confusion matrix of the test data. Figure (5). We observe something truly interesting, even if our model doe not have the best accuracy, it seems that it only fails with the adjacent values. That is to say, for instance, it confuses the 4 with the 3 and the 5, the 3, with the 4 and the 2 and so on. So, we can base our model on this. To have different look of this behavior, we can observe the table (6).

Now, we based our use case in the following: we took the values that the model predicts larger than the actual value and we assume that house it might have the possibility to add a new bathroom, in this manner we marked as 1 if the house is considered as a candidate to add a new bathroom or zero if is not. The results

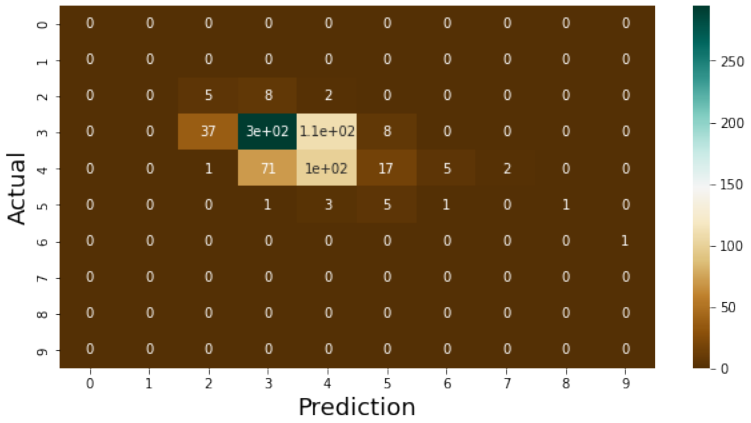

Figure 5: Table that corresponds to the accuracy for the Training and validation data as a function of the number of neighbors.

	Actual	prediction
0	3	3.0
1	3	3.0
2	3	4.0
3	2	3.0
4	3	3.0
5	3	3.0
6	4	3.0
7	3	3.0
8	2	3.0
9	3	3.0
10	3	3.0
11	3	4.0
12	3	3.0
13	3	3.0
14	2	2.0

Figure 6: Table that corresponds to the predicted value against the actual value used in the test data.

of this classification can be seen in figure (8). In overall, is was obtained a total of 734 candidate houses for adding a new bathroom

	sqrt_ft	bedrooms	bathrooms	predicted	candidate
0	7471.0	5	6	6.0	0
1	5333.0	4	4	5.0	1
2	6800.0	4	7	6.0	0
3	6622.0	4	6	6.0	0
4	7132.0	5	6	6.0	0

Figure 7: Table that shows, based on the variables used in the model and the predicted ones, if a house can be candidate of adding a new bathroom (1) or not (0).

Additionally, we made an analysis based on the coordinates of the houses, and locate in the map, the ones that are candidates. These results can be seen in the figure

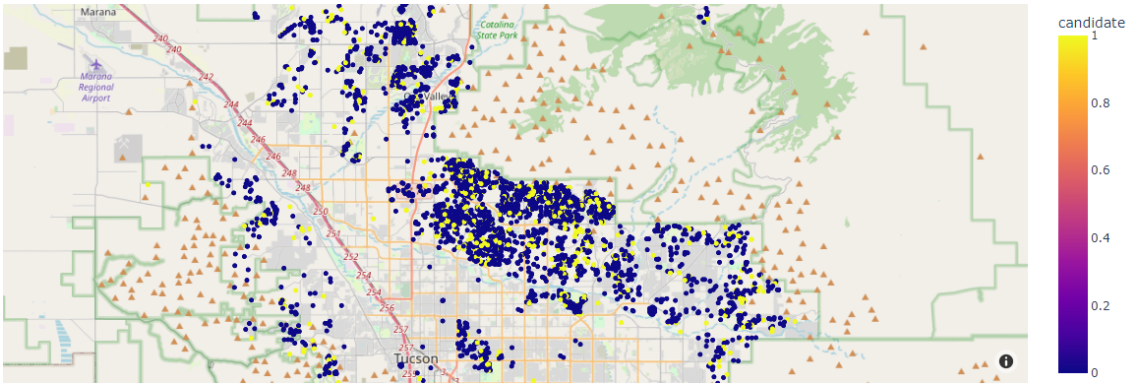

Figure 8: Map that shows the locations where are the houses that are candidates