IBM Applied Data Science Capestone Project

Predicting house prices in Monza using Foursquare API

Alessandro Bonvini, 2021 February 5th

1. Introduction: business problem

The goal of this project is to study the impact of using Foursquare API on the regression problem of predicting house prices in the Italian city of Monza.

The possible advantage of using Foursquare API is related to the fact that the square meter price of the houses in Monza, like in most of cities, varies depending on the position of the house.

In particular, I will investigate if the retrieval of the top trend venues in Monza and in the immediate vicinity of each house, can improve the performance of different machine learning regression models.

The presence of one or more important monument, church or square or famous restaurants, shopping centers etc. close to a house, in fact, will likely have an impact to its square meter price.

We will check if the information given by Foursquare API could be used to automatically cluster the houses in trending neighborhoods and then use the obtained cluster information as an additional feature used by the machine learning models.

Alternatively, we will explore if adding the top trending venues directly as features to the models will lead to a better performance with respect to using the cluster information or no spatial information at all.

1.1 Interested audience

The house prices prediction can be an interesting feature to be implemented in houses selling announcements search engines.

With this feature implemented, the seller can have an idea of the reasonable value of the house that he wants to sell and he can be warned if the price of the announcement is too far from the fair price.

At the same time, the buyer can understand what a good price for the house that he wants to buy could be, obtaining an indication about the price fairness.

2. Data

2.1 Initial dataset

After searching for available data sources that could be used for the project, I realized that no houses databases are public available.

Therefore, with the help of an estate agent working in Monza, we selected what could be the most impacting features to the house prices and we extracted hundreds of houses from his management system database, compiling afterwards some chosen features for each house.

In particular, we created a csv dataset of 405 houses to be used as training set, and another csv dataset of 44 houses to be used as test set to evaluate the generalization performance of the various models.

We decided each house to have the following features:

1. **PRICE** – the price of the house, in Euros.

2. **ADDRESS** – the street of the house, in the following format: *Number Street.*

3. **ROOMS** – the number of rooms of the house. Bathrooms are not considered as rooms.

4. **METERS** – the commercial square meters of the house.

5. **BATHROOMS** – the number of bathrooms of the house.

6. **FLOOR** – this feature describe the main floor of the house. Possible values are:
    a. **GROUND:** the main floor of the house is at ground level
    b. **MIDDLE:** the main floor is between the first and the last floors
    c. **LAST:** the main floor is the last, so it is an attic or a mansard
    d. **VILLA:** the house is an independent villa

7. **FLOORS** – the number of floors of the house.

8. **YEAR** – the construction year of the house.

9. **STATUS** – the current conditions of the house. This is the most subjective and difficult to determine feature, but also one of the possible most impacting features in estate agent's opinion. Possible values are:
    a. **BAD:** the house needs to be renovated before being inhabited
    b. **GOOD:** the house don't need renovation to be inhabited
    c. **RENOVATED:** the house has been completely renovated in one of the last 3 years and it is in close-to-new conditions
    d. **NEW:** the house is a new construction

10. **TERRACE** – YES if the house has a terrace large enough to be used for eating, otherwise NO

11. **GARDEN** – YES if the house has a garden, private or common, that can be used to let the kids play, otherwise NO.

12. **GARAGE** – YES if the house has a covered place to be used for parking one or more cars, otherwise NO.

13. **ENERGY** – the certified energy class of the house. Possible values range from G, which is the lowest class, to A4 which is the highest class.

14. **NEIGHBORHOOD** – the neighborhood of the house. Possible values range from 1 to 9.

15. **GRADE** – The estate agent evaluation for the house price. Possible values:
    a. **CHEAP**: the house price is lower than the real value.
    b. **NORMAL**: the house price is average for the house value
    c. **EXPENSIVE**: the house price is too high.

Here is how the first 10 rows look like after reading with pandas:

	PRICE	ADDRESS	ROOMS	METERS	BATHROOMS	FLOOR	FLOORS	YEAR	STATUS	TERRACE	GARDEN	GARAGE	ENERGY	NEIGHBORHOOD	GRADE
0	5317000	6 viale Cesare Battisti	5	542	3	GROUND	2	1900	RENOVATED	NO	YES	NO	A	1	EXPENSIVE
1	2970000	6 viale Cesare Battisti	4	295	3	MIDDLE	2	1900	RENOVATED	YES	NO	NO	A	1	EXPENSIVE
2	280000	3 via Ambrosini	3	115	2	MIDDLE	1	1980	GOOD	NO	NO	YES	E	1	NORMAL
3	1050000	16 via Carlo Porta	5	278	3	MIDDLE	2	1800	RENOVATED	YES	NO	YES	E	1	EXPENSIVE
4	690000	1 via Bellini	5	220	3	MIDDLE	1	1970	RENOVATED	YES	NO	YES	G	1	NORMAL
5	950000	14 via Sant'Andrea	3	272	3	GROUND	1	2020	NEW	NO	YES	YES	A3	1	NORMAL
6	450000	35 via Aliprandi Pinalla	3	145	1	LAST	1	1890	RENOVATED	NO	NO	YES	G	1	NORMAL
7	510000	9 via Ramazzotti	5	220	3	MIDDLE	1	1970	GOOD	NO	NO	YES	E	1	CHEAP
8	770000	via Donizetti	4	200	2	GROUND	1	2020	NEW	NO	YES	NO	A4	1	EXPENSIVE
9	650000	20 via Francesco Frisi	5	200	2	MIDDLE	1	1900	GOOD	NO	YES	NO	E	1	NORMAL

Figure 1 - training houses dataframe

The houses have been chosen to be in the same number for each neighborhood, that is 45 for each one.

2.3 Adding meter price information and removing outliers

In addition to the initially chosen features, I decided to add the square meter price information, which will be used as target variable for the regression models.

This because the price of the house is directly computed from the square meters of the house, by multiplying them with the square meter price.

Since this is a known relationship, I decided to remove it from the prediction, in order to focus the models only to find unknown relationships.

I then added the square meter price column to the initial dataframe.

After an initial visual analysis, which will be described in the following sections, I found out that two big outliers were present in the initial dataset.

The two outliers have been removed from the dataframe.

2.2 Adding spatial information and Foursquare API

Spatial information is expected to have a key role in improving the performances of the regression models.

By means of the address information, I used geopy library to retrieve latitudes and longitudes coordinates for each house.

I used ArcGIS as geolocator, because Nominatim failed to localize the house numbers in the city of Monza.

I then added latitudes and longitudes to the initial dataframe.

Regarding the usage of Foursquare API, the idea is to find what are the top trending venues in Monza, by exploring the area using Monza's coordinates and a radius of 2500.

The search has been performed with a GET request to the *explore* Foursquare Endpoint, by specifying, in addition to the standard parameters, also to sort the venues by popularity and to choose only the *TopPicks* venues.

I appended the obtained venues as columns to create a new dataframe containing Latitude, Longitude and top trending venues.

For each house, then, I performed a new Forsquare call, retrieving this time the TopPicks venues present in a reduced radius from the house.

The obtained venues have been then compared with the top trending venues resulted from the initial call. Each column of the new dataset, corresponding to a specific venue, have been set to 1 if the venue was also present as venue of the Forsquare call of the current house.

Here is how the first 4 rows and columns of the new dataframe look like, before and after the Foursquare call for each house.

	LAT	LNG	Villa Reale	Piazza Trento e Trieste	Istituti Clinici Zucchi	Parco di Monza - Ingresso Alle Grazie	U2	Parco di Monza - Viale cavriga	Dori	Civico 1	La Rinascente	Duomo di Monza	Macellerie Monzesi	La Feltrinelli
0	45.60266	9.26639	0	0	0	0	0	0	0	0	0	0	0	0
1	45.58266	9.27903	0	0	0	0	0	0	0	0	0	0	0	0
2	45.59647	9.27031	0	0	0	0	0	0	0	0	0	0	0	0
3	45.59982	9.26604	0	0	0	0	0	0	0	0	0	0	0	0
4	45.58688	9.27912	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2 - venues dataframe before houses calls

	LAT	LNG	Villa Reale	Piazza Trento e Trieste	Istituti Clinici Zucchi	Parco di Monza - Ingresso Alle Grazie	U2	Parco di Monza - Viale cavriga	Dori	Civico 1	La Rinascente	Duomo di Monza	Macellerie Monzesi	La Feltrinelli
0	45.60266	9.26639	0	0	0	0	0	1	0	0	0	0	0	0
1	45.58266	9.27903	0	1	1	0	0	0	1	0	1	1	0	1
2	45.59647	9.27031	1	0	0	0	0	1	0	0	0	0	0	0
3	45.59982	9.26604	0	0	0	0	0	1	0	0	0	0	0	0
4	45.58688	9.27912	0	1	1	0	0	0	1	0	0	1	0	0

Figure 3 - venues dataframe after houses calls

Considering the last row in the figures, that corresponds to the fifth house, only the 2th, 3rd, 7th and 10th of the first 12 venues have been set to 1.

This means that in the immediate vicinity of the house it is possible to find only "Piazza Trento e Trieste", "Istituti Clinici Zucchi", "Dori" and "Duomo di Monza".

This dataframe has been used to cluster each house and as additional features to machine learning models as described in the following sections.