Real estate market in Barcelona

Oscar de Dios

June 9, 2020

Introduction

Background

Barcelona is one of the world's leading tourist, economic, trade fair and cultural centres, and its influence in commerce, education, entertainment, sports, media, fashion, science, and the arts all contribute to its status as one of the major global cities. [1]

Barcelona is also one of the most densely populated cities in Europe with 1,636.762 people in 2019, a surface area of 101.35 km2 giving the city an average population density of 16,150 inhabitants per square kilometre with Eixample [2] being the most populated district.

Since 1987, the city has been divided into 10 administrative districts [3]:

- Ciutat Vella [4]
- Eixample [5]
- Sants-Montjuïc [6]
- Les Corts [7]
- Sarrià-Sant Gervasi [8]
- Gràcia [9]
- Horta-Guinardó [10]
- Nou Barris [11]
- Sant Andreu [12]
- Sant Martí [13]

Main Goal

Due to crisis of COVID-19 all the cities around the world have reported a huge impact in a lot of aspects but what is happening in the real state market? Banks do not understand of crisis and want to keep mortgages' payments up to date. One of the most relevant indicators (at least in Spain) about the situation is the real state pricing trend. Based on intuition, seems to be reasonable believing that in this critical situation the real state market should change the trend or at least should show some changes based on it.

Barcelona is divided into 10 districts. Obviously districts are very different as the people who lives in them. The current behaviour of real-state market should be different depending on purchasing power. The main goal is finding an accurate prediction about real state market situation for next months for each district and demonstrating the impact of the COVID-19 crisis. Let's analyse the consequences of COVID-19 in the real state market in Barcelona.

Data section

Data sources

In order to get relevant information I checked the public data available from different sources. I visited a lot of sites but in more cases the free information were poor. Finally I worked with this data sources:

- Related to real estate market sales I collected the information from one of the most popular real estate sites on line in Spain (https://www.idealista.com) [14]. In this site I discovered a specific place with the information about pricing evolution over years (https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/cataluna/barcelona-provincia/barcelona/) [15]. I developed a little "web scrapping" code but unfortunately this site are protected of this type of code (Congrats site admin!). Finally I decided to install a Firefox plugin called Table2Clipboard that allows table capturing and export easily from HTML code.

Data Requirements & Data Collection

In order to take the right data to afford the problem I needed to find the following information:

- Districts information
- Real estate market information

Cleansing and Transforming data

In several occasions the information needed to be modified specifically in this points:

- Replacing NaN values with correct values because the field format between the three CSV files were different
- Renaming columns' names
- Deleting unneeded columns
- Changing column values
- Deleting incomplete records
- Grouping columns by certain criteria
- Splitting datasets

Feature Selection

Before data cleaning, there were 1.720 records corresponding to the selling price of the whole districts of Barcelona. In the first approach I had several CSV files for each year and after join all the files together in one file called "Barcelona_selling_prices.csv" I had one record per month and district from Jan 2016 to 2020 May.

Due that the information comes from the same source (*idealista web site)* the information were well formatted and structured.

After that, I renamed the columns to english language. I deleted the rows corresponding to selling price with not defined values and I obtained 1.354 records.

I did some transformation converting the field date fo date type and the selling price to numeric value.

Exploratory Data Analysis

With the information of these years I proposed to put the focus in the pricing variation for all the districts in Barcelona

During last months the COVID-19 crisis collapsed the economy. All the people are worried about what will happen. In this situation one of the most relevant economic indicators of the market (at least in Spain) are the m2 pricing behaviour in the real estate market.

Global Pricing distribution

In a first approach I needed to know the distribution in price/m2 for the whole real estate market. Just as expected I found a Gauss Curve. No surprises until now. Although, there is more values in the high band between 3500€/m2 and 4500€/m2 than in the low band between 2500€/m2 and 3500€/m2. Then there is a clear trend towards high values. Therefore more expensive houses.

The first graphic shows the selling price/m2 for the years from 2016 to 2020 (see *Figure 1: Global Pricing Distribution*).
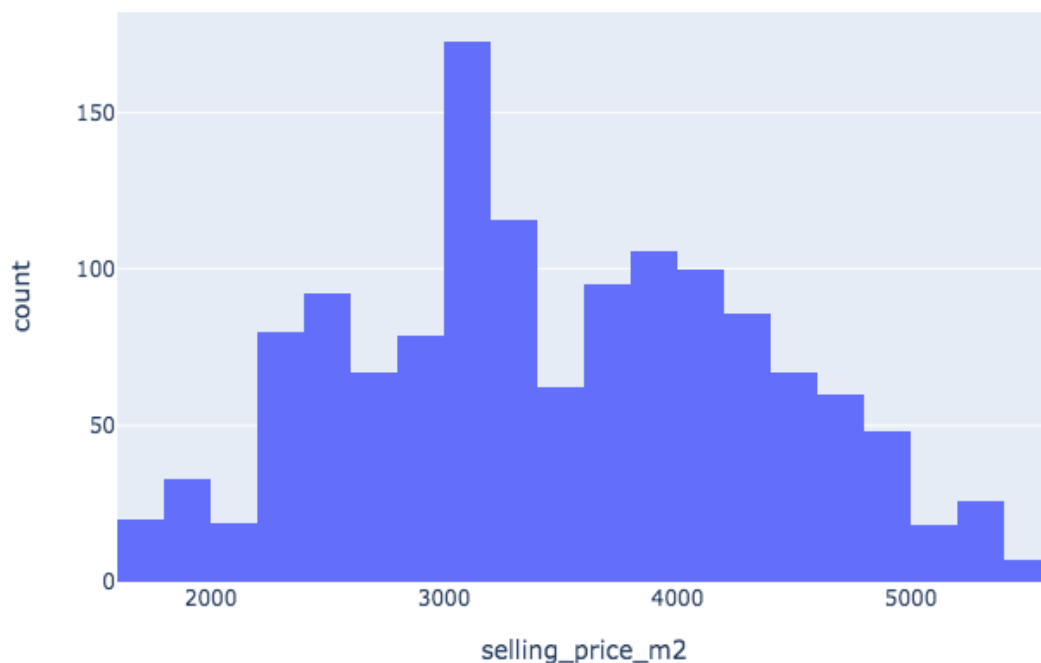
Figure 1: Global Pricing Distribution

Grouping pricing market share by district

Grouping the information by district I detected there was three groups of data:

- The first group corresponds to lowest princing band (between 2500€/m2 and 3200€/m2). Here we can found Nou Barris, Horta-Guinardó and Sant Andreu.

- The second group corresponds to mid pricing band (between 3500€/m2 and 4500€/m2). Here we cant find Sant Andreu, Sants-Montjuïc, Sant Martí, Gràcia, Ciutat Vella and Eixample

- The third group corresponds to the higher pricing band (more than 5000€/m2). Here we can find Les Corts and Sarrià-Sant Gervasi.
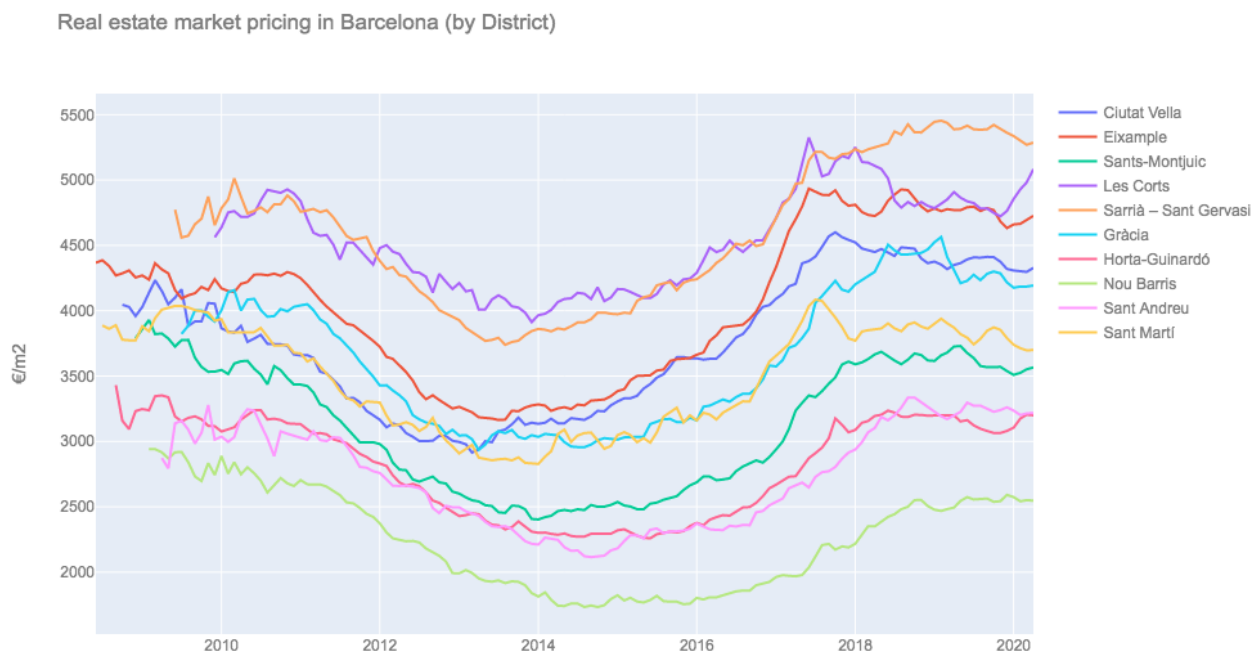

Figure 2: Real estate market pricing in Barcelona (by District)

After this steps I decided to perform a deep analysis at district level. Remember that we need to predict the future pricing for next six months based on historical data.

Pricing analysis by district

In the next section we will analyse the district information following this pattern:

1) Market price evolution & test over existent data

I have divided the rows into two windows. The first window contains all the records until -6 and the second window contains the records from -6 until the end. With the information of the first window I predicted the information of the second window. With both windows we can compare the precision of the model using root mean squared error (rmse).

I order to improve performance I used a preprocessing function to scale the all the values between 0 and 1.

After that I created and alternate time series table to store the results for the last 6 periods of time. The best practice for this type of problems related to timeframes recommends CNN LSTM architecture. The Convolutional Neural Networks Long Short-Term Memory Network or CNN LSTM is specifically designed for sequence prediction problems.

Before getting the graphic representation we must inverse the values that were scaled in previous steps and we need to show the information with readable values.

In the graphic we will see two lines in the last 6 periods. The first line (blue) comes from the start of the dataset and the second line (red) has been predicted by the CNN model. It would be desirable that both lines were overlapped. It means that the prediction and the real values are very closer and that's good.

2) Predicted future values

In the second part of the exercise I created new 6 periods of time at the end on the dataset. Once we fit the CNN model, the process returned new records with new information predicted based on previous calculations.

In the graphic we will see two lines in the last 6 periods. The first line (blue) comes from the start of the dataset and the second line (red) has been predicted by the CNN model and have been added following the blue line.

3) Root mean squared error & certainty

In both cases (Market price evolution & test over existent data and Predicted future values) I used the same performance indicator. In the first case I use the RMSE for calculation between real selling values and calculated values. In the second case I use the RMSE for calculation between previous six time period values and last six time periods calculated values. It is very important to remember that the RMSE value is evaluated in the same units of the values, in this case, €/m2.

Ciutat Vella

Predicted vs. current values comparison for Ciutat Vella
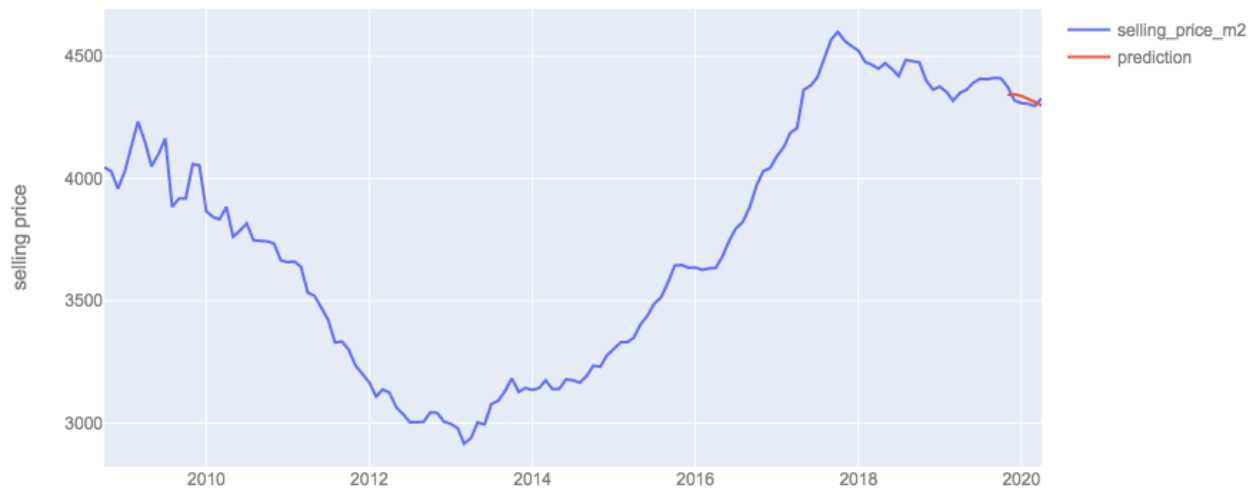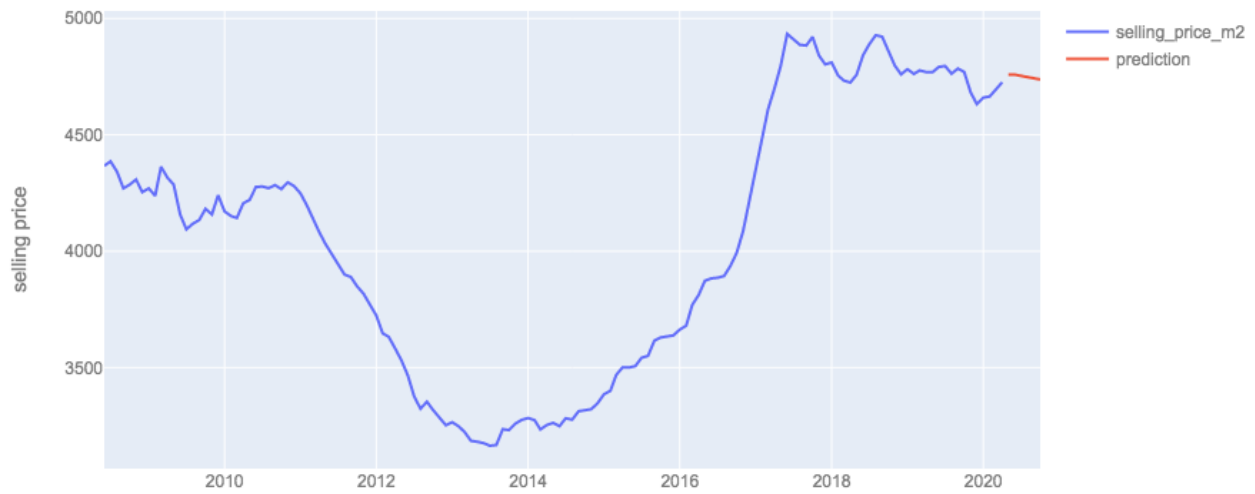

Future predicted values for Ciutat Vella

Analysis:

In the predicted values I observed a good overlapping between selling price and prediction. In the future prediction the algorithm indicates a little pricing increase at the start and it keeps flat after that. Good performance near 99% in future predictions and a deviation of 56,30€ of total price.
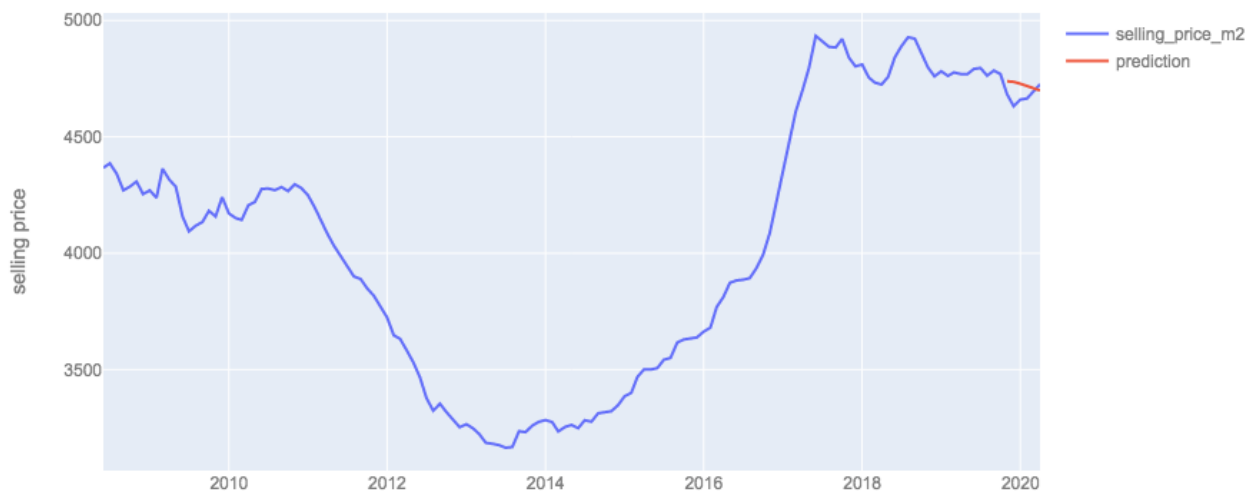
District_name	Prediction_type	RMSE(€)	Error(%)	Certainty(%)
Ciutat Vella	train-test	26.45	0.61	99.39
Ciutat Vella	future	56.30	1.29	98.71

Eixample

Future predicted values for Eixample

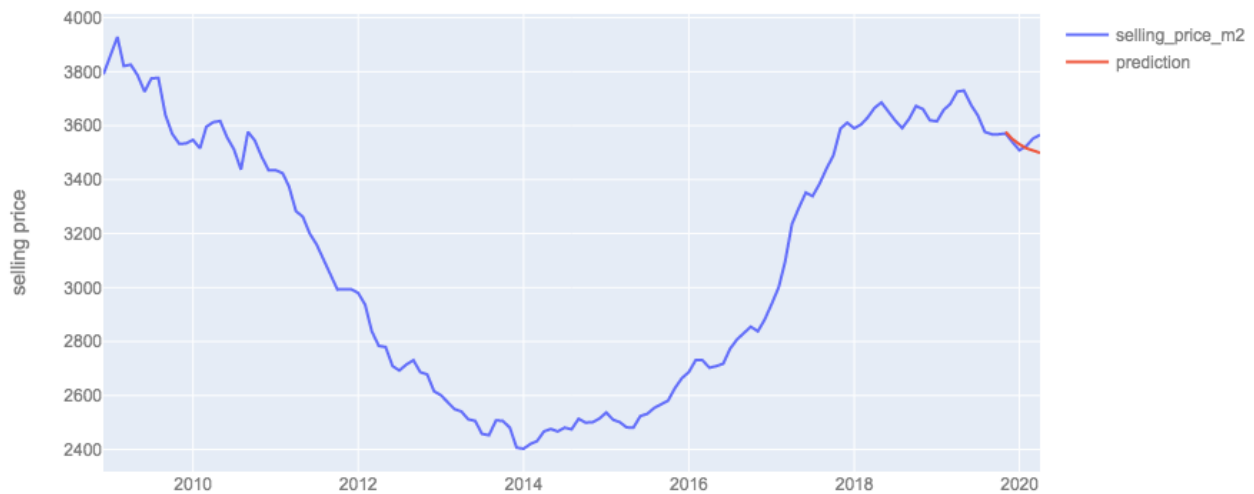Predicted vs. current values comparison for Eixample

Analysis:

In the predicted values I observed a poor overlapping between selling price and prediction. In the future prediction the algorithm indicates that the prices goes down slowly. Good performance in both values.

District_name	Prediction_type	RMSE(€)	Error(%)	Certainty(%)
Eixample	train-test	61.28	1.31	98.69
Eixample	future	81.71	1.72	98.28

Sants-Montjuïc

Predicted vs. current values comparison for Sants-Montjuic

Future predicted values for Sants-Montjuic

Analysis:

In the predicted values I observed a reasonable overlapping between selling price and prediction, at least in the firsts time series. In the future prediction the algorithm indicates that the prices goes down fast. Very good performance.

District_name	Prediction_type	RMSE(€)	Error(%)	Certainty(%)
Sants-Montjuic	train-test	34.35	0.97	99.03
Sants-Montjuic	future	29.75	0.84	99.16

Les Corts

Predicted vs. current values comparison for Les Corts
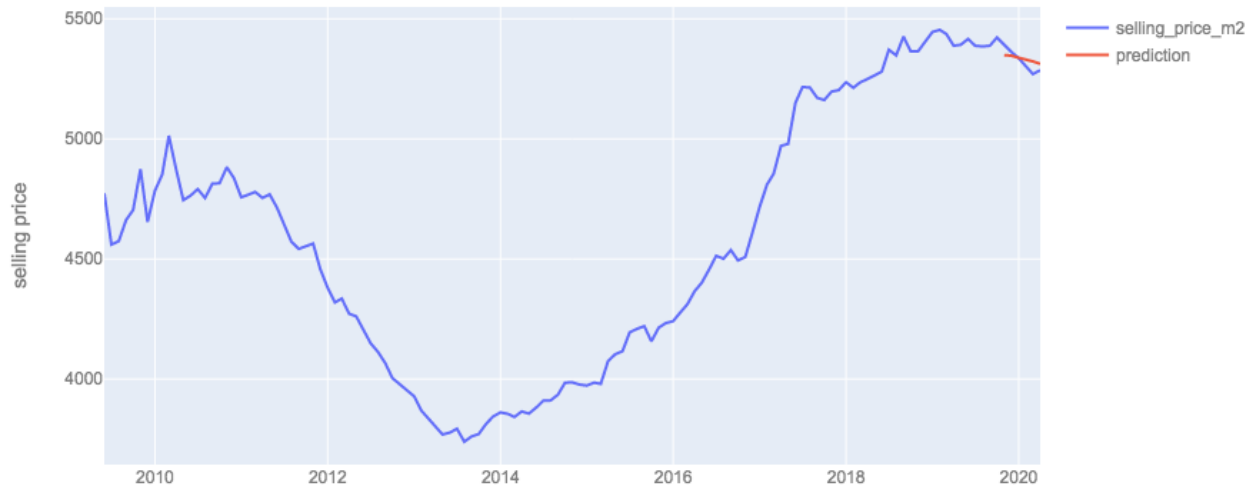

Future predicted values for Les Corts

Analysis:

In the predicted values I observed a very bad values between selling price and prediction, This behaviour is the same in the future prediction. The algorithm indicates that the prices goes down fast. Very bad performance comparing to other districts. Maybe related with the high pricing/m2 of this district.

District_name	Prediction_type	RMSE(€)	Error(%)	Certainty(%)
Les Corts	train-test	295.81	6.05	93.95
Les Corts	future	176.68	3.70	96.30

Sarrià – Sant Gervasi

Predicted vs. current values comparison for Sarrià – Sant Gervasi
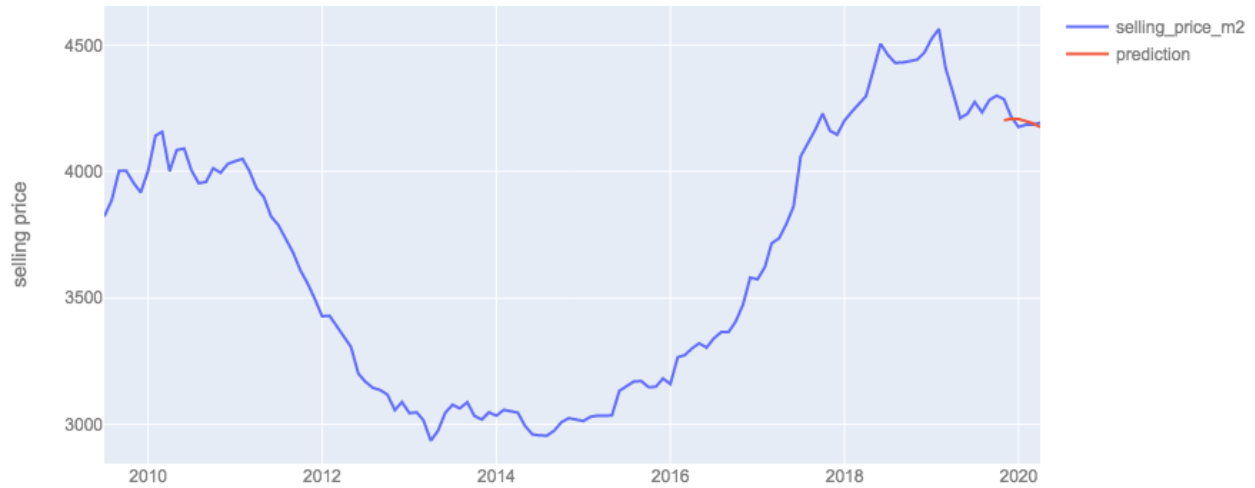

Future predicted values for Sarrià – Sant Gervasi

Analysis:

In the predicted values I observed a correct overlapping between selling price and prediction, This behaviour is the totally different in the future prediction where the algorithm indicates that the prices keeps the value. Good performance near to 99 in both cases.
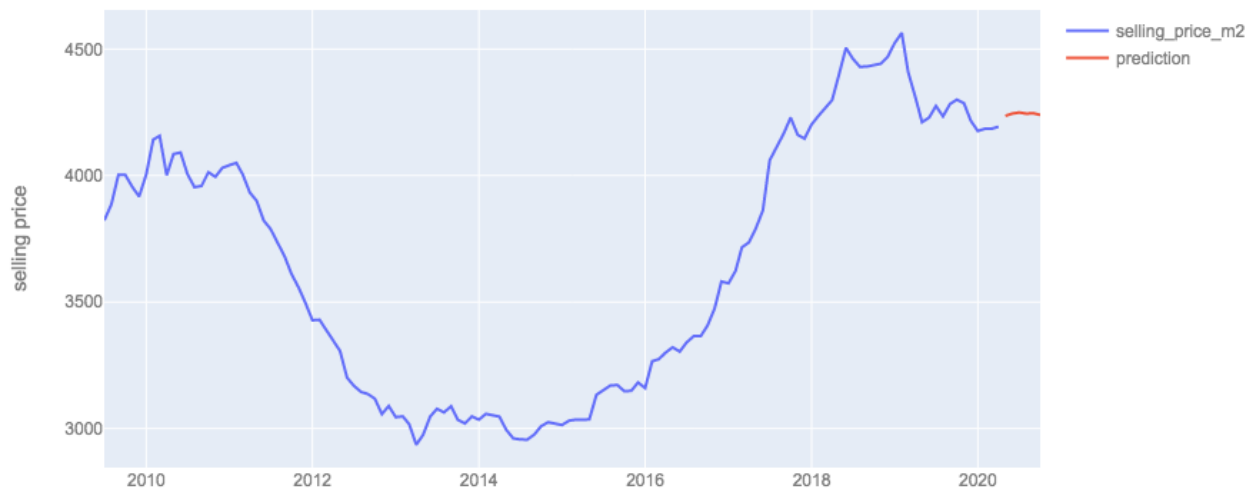
District_name	Prediction_type	RMSE(€)	Error(%)	Certainty(%)
Sarrià – Sant Gervasi	train-test	31.76	0.60	99.40
Sarrià – Sant Gervasi	future	77.82	1.44	98.56

Gràcia

Predicted vs. current values comparison for Gràcia
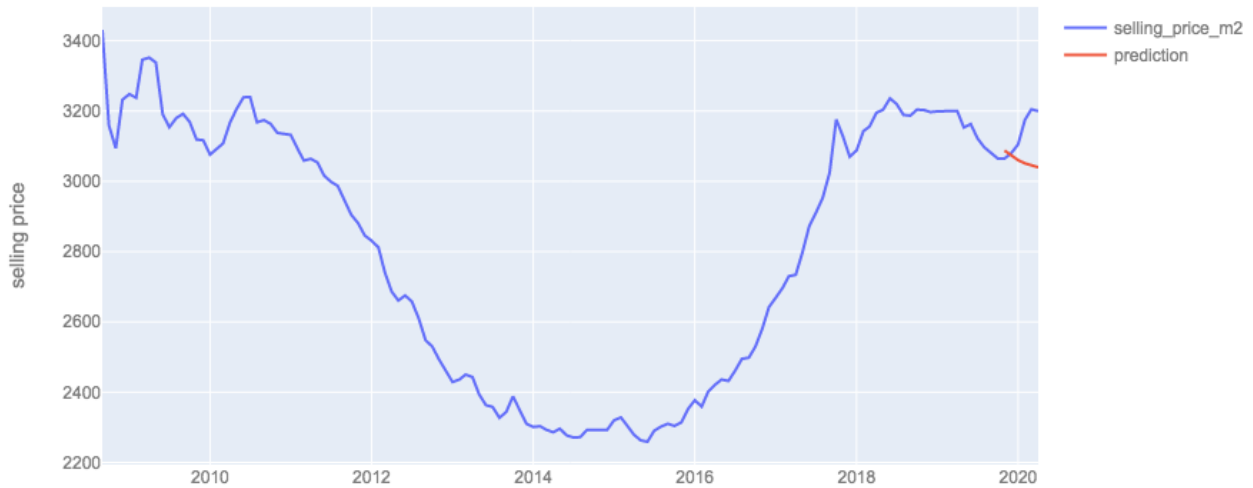

Future predicted values for Gràcia

Analysis:

In the predicted values I observed a good overlapping between selling price and prediction. As well as the future prediction offers a stable value. Very good performance of indicators.
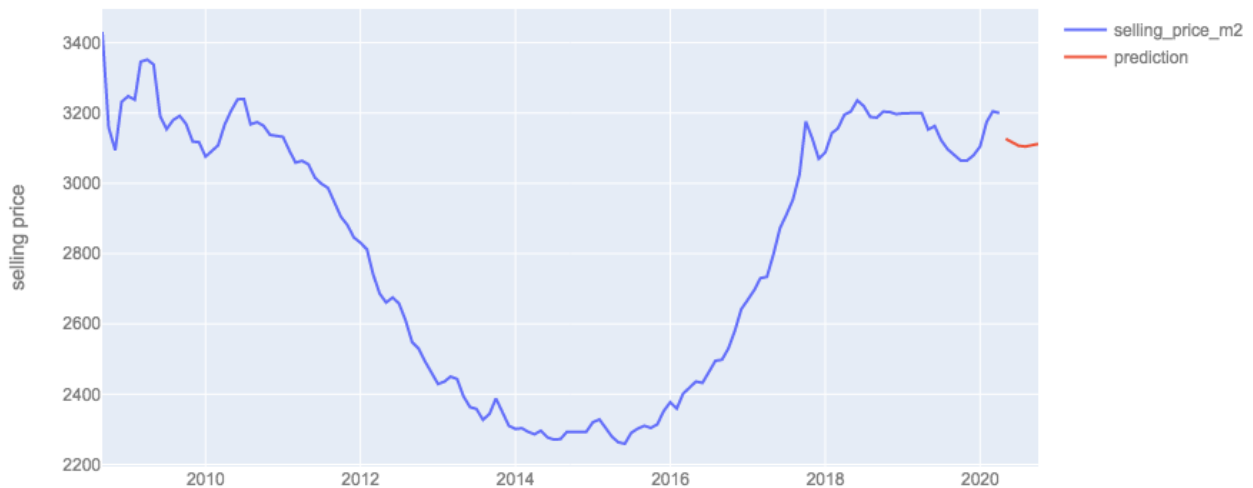
District_name	Prediction_type	RMSE(€)	Error(%)	Certainty(%)
Gràcia	train-test	37.66	0.9	99.1
Gràcia	future	55.01	1.3	98.7

Horta

Predicted vs. current values comparison for Horta-Guinardó
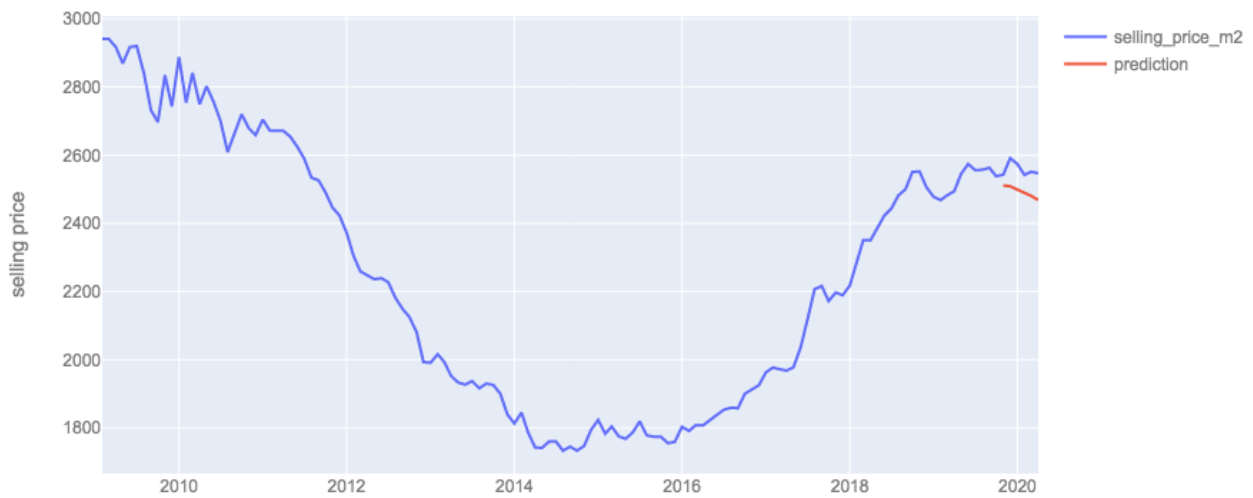

Future predicted values for Horta-Guinardó

Analysis:

In the predicted values I observed a full deviation between selling price and prediction, This behaviour is the same in the future prediction where the algorithm indicates that the prices draws a curve. Upgradable performance.
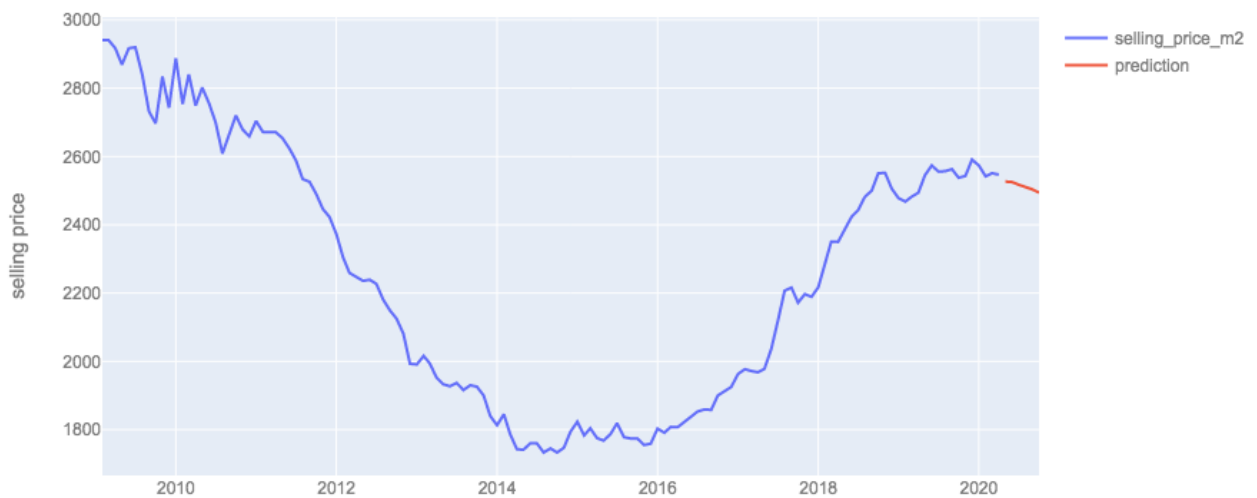
District_name	Prediction_type	RMSE(€)	Error(%)	Certainty(%)
Horta-Guinardó	train-test	107.34	3.42	96.58
Horta-Guinardó	future	67.59	2.17	97.83

Nou Barris

Predicted vs. current values comparison for Nou Barris

Future predicted values for Nou Barris

Analysis:

In the predicted values I observed an important deviation between both values. But in bot graphics I observed the same behaviour. The trend indicates decreasing values. Correct performance of the indicators.

District_name	Prediction_type	RMSE(€)	Error(%)	Certainty(%)
Nou Barris	train-test	67.44	2.64	97.36
Nou Barris	future	48.01	1.91	98.09

Sant Andreu

Predicted vs. current values comparison for Sant Andreu
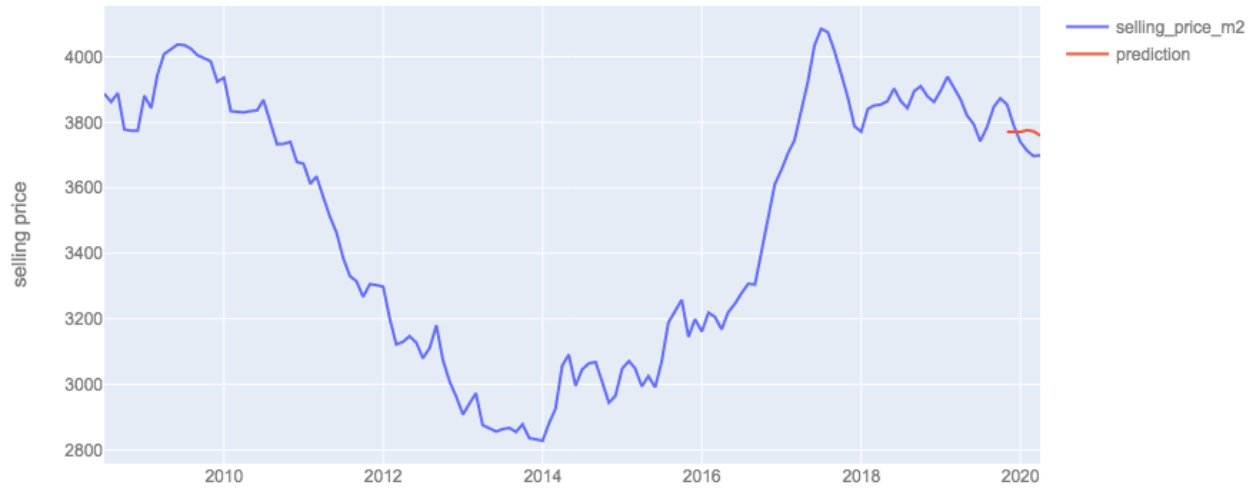

Future predicted values for Sant Andreu

Analysis:

In the predicted values I observed a good approach in trend between selling price and prediction. As well as the future prediction offers a stable decreasing value. Very good performance of indicators.

District_name	Prediction_type	RMSE(€)	Error(%)	Certainty(%)
Sant Andreu	train-test	43.21	1.34	98.66
Sant Andreu	future	13.37	0.41	99.59

Sant Martí

Predicted vs. current values comparison for Sant Martí

Future predicted values for Sant Martí

Analysis:

I appreciated that both values (predicted & current) have different trajectories. With the interactive graphic I observed that the major distance is over 80€/m2. Regarding to futures values the graphic show a clear descendent behaviour. Good performance of the indicators.

District_name	Prediction_type	RMSE(€)	Error(%)	Certainty(%)
Sant Martí	train-test	59.88	1.60	98.40
Sant Martí	future	58.19	1.54	98.46

Results

In the first part of this study, I analysed the real estate pricing evolution in Barcelona from 2009 to 2020. The main goal were to understand the relationship between historical prices, and market behaviour. Moreover, in the second part, I implemented a CNN (based on LSTM architecture) in order to predict future values for selling prices.

The requirement have been covered but for future projects I will try to develop a multivariate CNN because only one value to predict future values is less interesting that compare different rates and the correlation or interaction between them.

In the end, it all comes down to availability of quality data. The working model is easy to understand but the more quality information we have the more benefit we get.

Discussion

As I mentioned before, this project could be extended to a general use and I have a list of improvements in order to get more benefits:

- Adding more features in order to enrich the outcomes. With this action we can get a holistic vision.
- Designing a solution for real estate agents could be very useful to check the price, the trend, collateral influences in results and so on.

Conclusion

Time series prediction is one of the most typical business case. The role of CNN in this type of projects is capital. If we can enrich the inbound information we can get better combinations and awesome outcomes.

Specifically with this projects I detected a global trend for the majority of districts that indicates a stabilization of prices for the next months. Will see.

Links

[1]   https://en.wikipedia.org/wiki/Barcelona
[2]   https://www.idescat.cat/emex/?id=080193&lang=es
[3]   https://en.wikipedia.org/wiki/Guàrdia_Urbana_de_Barcelona
[4]   https://en.wikipedia.org/wiki/Ciutat_Vella
[5]   https://en.wikipedia.org/wiki/Eixample
[6]   https://en.wikipedia.org/wiki/Sants-Montjuïc
[7]   https://en.wikipedia.org/wiki/District_of_Les_Corts
[8]   https://en.wikipedia.org/wiki/Sarrià-Sant_Gervasi
[9]   https://en.wikipedia.org/wiki/Gràcia
[10]   https://en.wikipedia.org/wiki/Horta-Guinardó
[11]   https://en.wikipedia.org/wiki/Nou_Barris
[12]   https://en.wikipedia.org/wiki/Sant_Andreu
[13]   https://en.wikipedia.org/wiki/Sant_Martí_(district)
[14]   https://www.idealista.com
[15]   https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/cataluna/barcelona-provincia/barcelona/)