**TFM MEMORIA**

**Find the Best House Investment in Madrid (Spain)**

1- The Question: Introduction

The main idea of this project was to find the best investment in the housing market in Madrid at the moment. There are other similar studies related to the housing market in different locations but nothing related to Madrid.

We recognize this city is completely different compared to other Spanish cities like coastal or vacation cities where we historically found more buying interest. But nowadays Madrid is where investors come when they plan to invest in buildings and houses, we have a great economic growth, a mix of different potential clients (from students to business or tourists), great universities, big international and national companies and great tourist attractions.

2- Data Description

We go the data from various different websites:

* Ayuntamiento de Madrid and INE (dataset\_16\_19\_mad\_3.csv): These websites are full of historical market data, here is where we got 2016 to 2019 housing data by m2, size, number of transactions and other metrics.
* Trovit (df\_mad\_houses.csv): This data comes from the scrapper built to extract data directly from the Trovit Houses website.
* Ministerio de Educación (Addresses\_Schools\_Mad.csv): All schools data in Spain to measure distance between houses and schools (private and public).
* Ministerio de Sanidad (salud\_madrid.csv): List with all the health centres in Comunidad de Madrid (hospitals and clinics)
* Ministerio de Transportes (transport\_stops\_madrid.csv): Data from all transport stops in madrid for train and subway (Metro).
* INE (latlons\_by\_postcode\_madrid.csv): Dataset with all the postcodes in Madrid adding Lat and Long of the centre of each postcode.
* Geojson files include data from the Comunidad de Madrid including polygons and merged with our results
* x\_all.csv and y\_all.csv have features and prices to work on our model for Streamlit.

3- The Model: Tests and Results

We used two different models, one to estimate the future (2030) prices per square meter in Madrid by postcode and another one to estimate prices of the “on the market” houses using all the market data collected from Trovit.

For the first one we used the data collected on the Ayuntamiento de Madrid and INE where we can get the historical data from between 2016 to the beginning of 2020. We tested different regression models including Linear Regression, Lasso, SVR, Ridge, TheilSen and Random Forest in this case without using cross validation or pipelines because of the size of our data.

Linear regression gave us the best R2 with almost 93% accuracy so we used that one to estimate prices in 2030 for each of the Madrid postcodes.

The second one had all the market data so we could also add more information like distance to transport stops, hospitals and schools to improve our model and make it more accurate. Did a one hot encoder to change all our postcodes into columns and used them as one feature for our model. After all the data preparation we started to build and test our model.

Used different pipelines and grid searches to test different regression models, Linear, Lasso, Ridge and Random Forest with multiple parameters. In this case the Random Forest gave us the best R2 with 86% accuracy so we used that one with the best parameters from our grid search to train our final model.

4- The Answer: Final Results

With all the information merged we had our final data set with prices per square meter now, the model estimation from market data and the future estimation in 2030.

The result was that in the city centre is where you can find the highest and more estable prices per square meter in Madrid at the moment with other areas like northwest and north in continuous growth but the biggest opportunities (underpriced houses) are in the outskirts and more on the southwest, east and south areas where probably COVID financial crisis is affecting more aggressively.

5- Front-end Instructions

In the frontend you have first a table and two boxplots that gives you and idea of the market. You can select the areas you want to see and all these three will change accordingly to show different information.

Then you can find two interactive graphs where you can see how prices vary depending on the size of the house.

Below that there are two map charts, one has a slider where you can select between:

* Prices per m2 of the market
* Prices per m2 estimated by the Random Forest Regression model
* Prices per m2 estimated by the Linear Regression in 2030

The next map chart is a 3D Pydeck chart that shows the estimated profit now (underpriced areas)

Last but not least you can put data of a house (number of rooms, bathrooms, size and postcode) and the Random Forest Regression model will give you a price estimation and the average of that area. It loses accuracy over 200m2 and over 5 rooms or bathrooms because this data is not included in the model.