Real estate annual profitability prediction

# 

[**1. Declaració del problema 2**](#_n67msyvt6nf)

[**2. Part 1: Evolució i predicció del mercat immobiliari a Barcelona 2**](#_4ihp2ws6d62t)

[**3. Part 2: Predicció del preu d’una vivenda 2**](#_tfmtrnh3lljs)

[Descripció 2](#_hoz223vcskc0)

[Exploració de les dades 3](#_nmgtpzwqismr)

[Model predictiu 3](#_sf8zuxxpag52)

[Explicabilitat del model 4](#_4yjlvaz5nzob)

[**4. Metodologia (no se si posar-ho aqui o abans de les explicacions) 4**](#_98lztkba5f9c)

[Eines 4](#_ycd898f0n4x5)

[**5. Conclusions 4**](#_323ebxyvsvba)

# 

# 

# 1. Declaració del problema

L’objectiu inicial del projecte era predir el comportament del mercat immobiliari català durant els propers anys per poder saber quina era la millor zona de Catalunya per invertir en vivenda.

Dels varios datasets que hem trobat, no n’hi ha cap de perfecte. Així doncs hem decidit provar-ne dos i veure quin ens funciona.

Finalment, el projecte consta de dues parts, una amb cada dataset. La primera és un anàlisi del mercat a Catalunya el novembre del 2020, amb una eina de predicció del preu d’una vivenda segons el tipus d’habitatge, el nombre d'habitacions, banys, m2 i la localització del mateix, juntament amb l’explicació de les prediccions del model.

La segona és un anàlisi centrat en la ciutat de Barcelona on es pot veure, per districtes o barris, l’evolució del preu mitjà de compra i de lloguer juntament amb una predicció pels propers anys, d’aquesta manera estarem responent a la pregunta inicial del projecte.

# 2. Part 1: Evolució i predicció del mercat immobiliari a Barcelona

En aquesta part del projecte, hem analitzat l'evolució dels preus de compra i lloguer d'habitatges a la ciutat de Barcelona, diferenciant per districtes i barris. Hem utilitzat dades històriques (del Gencat) des de l'any 2000 per districtes i des de l'any 2012 per barris, per tal de proporcionar una visió detallada de les tendències del mercat immobiliari. A més, hem aplicat el model de predicció Prophet per estimar els preus futurs fins a l'any 2034. Els resultats mostren una clara tendència de creixement en els preus de compra i lloguer, amb variacions significatives entre diferents districtes i barris. Aquesta anàlisi permet identificar les zones amb major potencial de rendibilitat per a inversions futures. Els gràfics de predicció i comparació de preus de compra i lloguer proporcionen una eina visual poderosa per a la presa de decisions informades en el mercat immobiliari de Barcelona.

Un insight interessant que hem extret de les dades és que els districtes amb major rendibilitat anual prevista són aquells amb preus de lloguer relativament alts en comparació amb els preus de compra. Per exemple, districtes com Ciutat Vella i Sant Martí mostren una rendibilitat significativa, fent-los atractius per a inversors. A més, la nostra aplicació web desenvolupada amb Streamlit permet als usuaris explorar aquestes dades de manera interactiva, visualitzant les prediccions de preus i rendibilitats per a cada districte i barri, facilitant així la presa de decisions basada en dades actualitzades i prediccions futures.

# 3. Part 2: Predicció del preu d’una vivenda

Font: <https://zenodo.org/records/4263693> (TODO: Citar bé)

## Descripció

El primer dataset és una imatge del mercat immobiliari espanyol al novembre de 2020. Tot i no ser dades molt recents, l’hem escollit perquè ens brindava l’oportunitat de fer un model tipus “Quant val casa meva?”.

En la primera fase de neteja ens hem quedat només amb les zones de Catalunya, filtrant manualment localització per localització. Un cop fet, hem arreglat i preparat el dataset per poder fer l’exploració de les dades (EDA).

## Exploració de les dades

Abans de l’EDA, ens hem adonat que hi havien moltes files duplicades, entenem que el dataset és un recull de varies pàgines web d’ofertes i que cada vivenda pot estar ofertada en varies webs. Un cop eliminats els duplicats passem d’uns 170k pisos a uns 60k, una reducció significativa que ens farà baixar el rendiment del model.

Durant l’exploració de les dades hem aprofitat per eliminar outliers per cada una de les variables, aquests son els rangs que utilitzarem de cada variable en el model:

* Preu: menys d’1M €
* Preu m2: entre 50€ i 8000€
* Area: entre 30m2 i 600m2
* Nombre d’habitacions: 7 o menys
* Nombre de banys: 5 o menys
* Tipus de vivenda: Pis, casa, dúplex, àtic i finca rústiques.

També hem eliminat del dataset totes aquelles localitzacions amb menys de 10 entrades.

Tot l’EDA es pot veure i entendre en el notebook “main\_predicció\_catalunya.ipynb”

## Model predictiu

El model utilitza RandomForestRegressor per predir el preu d’una vivenda donades les seguents característiques:

* Ubicació (poble o barri)
* Superfície de la vivenda
* Nombre d’habitacions
* Nombre de banys
* Tipus de vivenda

Disposant d’unes 60k entrades al dataset, hem entrenat el model amb 48k i reservat 12k per la validació, és a dir, hem fet servir un 80/20 del dataset. Amb aquestes dades de validació, hem obtingut:

* R2 score: 0.77 +- 0.01
* Error Absolut Mitjà: 57841 +- 824

En quant a error concret en cada predicció, hem observat que el model funciona molt correctament per a localitzacions i vivendes en properes a la mitjana però perd precisió en localitzacions molt cares o molt barates o si introdueixes dades poc comuns dins de certa localització. Per exemple, si fem la valoració d’un pis de 250m2 en una zona on la majoria dels pisos son de 60m2, l’error de la predicció serà alt, i així amb les altres variables.

## Explicabilitat del model

Per a treure millors i més precises conclusions sobre el model, hem aplicat SHAP, una llibreria de python que ens permet explicar el model variable per variable i fila per fila.

De les 5 variables presents en la predicció, les més rellevants son la superfície de la vivenda, el nombre de banys i la localització. Sorprèn veure que el model no dona gaire rellevància al nombre d’habitacions, un cop finalitzat el projecte ho compararem amb anuncis actuals per entendre aquesta poca rellevància en el preu.

Per a un anàlisi més en profunditat, el notebook conté les explicacions de la resta de conclusions extretes amb l’ajuda de SHAP.

# 4. Metodologia

Hem utilitzat diverses eines i llibreries de Python per a l'anàlisi de dades, la construcció de models predictius i la visualització de resultats. Les principals llibreries utilitzades inclouen Pandas per a la manipulació de dades, Scikit-Learn per a la construcció del model RandomForestRegressor, MatPlotLib i Seaborn per a la visualització de dades, i SHAP per a l'explicabilitat del model. A més, hem utilitzat Prophet per a la predicció de preus futurs. Per a la creació de l'aplicació web interactiva, hem utilitzat Streamlit, que permet als usuaris explorar les dades i les prediccions de manera intuïtiva. També hem utilitzat Tableau per a la visualització avançada de dades. Les dades han estat preprocessades i netejades per eliminar duplicats i outliers, assegurant així la qualitat dels resultats obtinguts.

# 5. Conclusions

Aquest projecte ha permès obtenir una visió detallada de l'evolució i predicció del mercat immobiliari a Catalunya, amb un enfocament especial en la ciutat de Barcelona. Hem demostrat que és possible utilitzar models predictius per estimar els preus futurs de compra i lloguer d'habitatges, identificant les zones amb major potencial de rendibilitat. Els resultats mostren una tendència de creixement en els preus, amb districtes com Ciutat Vella i Sant Martí destacant per la seva alta rendibilitat.

L'aplicació web desenvolupada amb Streamlit facilita la presa de decisions informades, permetent als usuaris explorar les dades i les prediccions de manera interactiva. Tot i les limitacions dels datasets utilitzats, els models han demostrat ser útils per a la predicció de preus, tot i que amb certa pèrdua de precisió en casos extrems. A més, el projecte inclou una predicció de taxació de vivenda per a Catalunya, permetent als usuaris estimar el valor d'una vivenda segons les seves característiques específiques. Creiem que aquest projecte proporciona una eina valuosa per a inversors i professionals del sector immobiliari, oferint una anàlisi complet i prediccions fiables del mercat immobiliari.