Predicting Listing Prices

Given the content of a listing for a house sale in Barcelona, predict the listing price of the house.

Table of Contents

01 **EDA**

Cleaning data and exploratory data analysis

02 **Word Embeddings**

Numerically encoding textual data

03 **Modeling**

Experimented with a variety of ML models

04 **Lessons Learned**

Key Takeaways

01

EDA

Cleaning data and exploratory data analysis

Data Cleaning Process

1. Removed Currency Symbol

We removed the Euro symbol (€) symbol to convert the prices into numeric values for analysis.

2. Dropped Redundant Features

The 'subtype' column was removed as it duplicated information available in the 'type' column, simplifying our dataset. The ‘selltype’ column was removed as there was no variation, all were ‘secondhand’.

3. Feature Engineering from Text Descriptions

Extracted square meters (m²), number of bedrooms, and number of bathrooms from the 'features' text column to create distinct, quantifiable features for our model.

	price	title	loc_string	loc	features	type	subtype	selltype	desc
0	320.000 €	Piso Tallers. Piso con 2 habitaciones con asce...	Barcelona - Sant Antoni	None	[85 m2, 2 hab., 1 baño, 3.647 €/m2]	FLAT	FLAT	SECOND_HAND	Piso en última planta a reformar en calle Tall...
1	335.000 €	Piso C/ de València. Piso reformado en venta d...	Barcelona - Dreta de l´Eixample	None	[65 m2, 2 hab., 1 baño, 5.000 €/m2]	FLAT	FLAT	SECOND_HAND	Ubicado en la zona del Camp de l'Arpa, cerca d...
2	330.000 €	Piso en Dreta de l´Eixample. Acogedor piso al ...	Barcelona - Dreta de l´Eixample	None	[77 m2, 2 hab., 1 baño, 4.286 €/m2]	FLAT	FLAT	SECOND_HAND	En pleno centro de Barcelona, justo al lado de...
3	435.000 €	Piso Barcelona - corts catalanes. Soleado, cén...	Barcelona - Sant Antoni	None	[96 m2, 3 hab., 2 baños, 4.531 €/m2]	FLAT	FLAT	SECOND_HAND	Vivienda espaciosa en Sant Antoni, cerca de Pl...
4	410.000 €	Piso en Carrer de sardenya 271. Alto, reformad...	Barcelona - Sagrada Família	Carrer de Sardenya 271	[84 m2, 2 hab., 1 baño, 4.881 €/m2]	FLAT	FLAT	SECOND_HAND	En el corazón de Barcelona, en una hermosa fin...

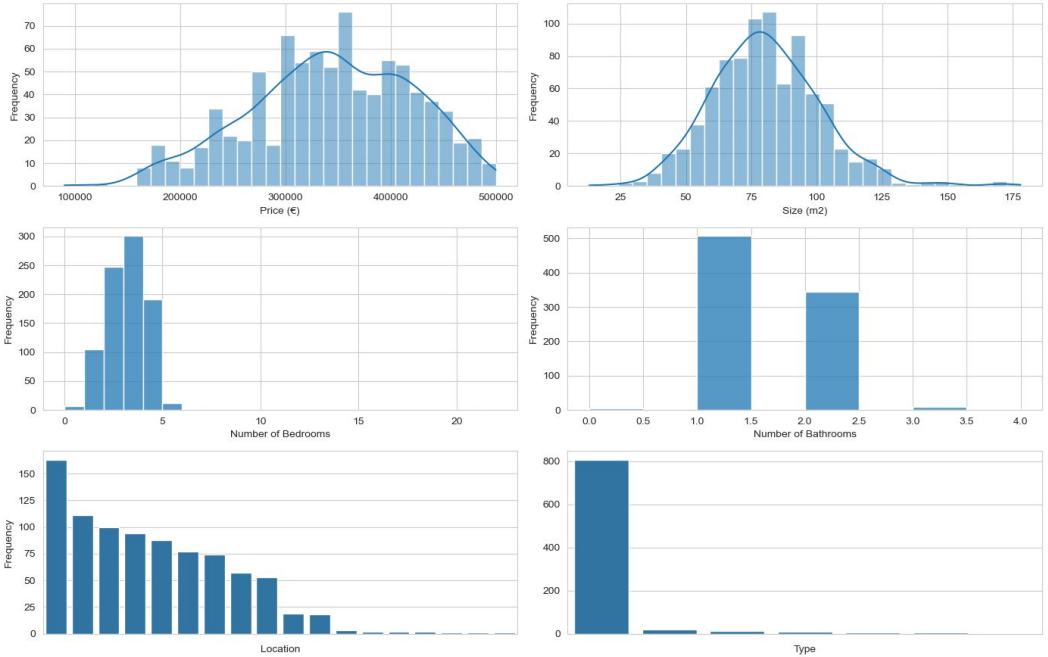
Raw data

	price	title	loc_string	loc	type	desc	m2	bedrooms	bathrooms
0	320000	Piso Tallers. Piso con 2 habitaciones con asce...	Barcelona - Sant Antoni	None	FLAT	Piso en última planta a reformar en calle Tall...	85	2.0	1.0
1	335000	Piso C/ de València. Piso reformado en venta d...	Barcelona - Dreta de l´Eixample	None	FLAT	Ubicado en la zona del Camp de l'Arpa, cerca d...	65	2.0	1.0
2	330000	Piso en Dreta de l´Eixample. Acogedor piso al ...	Barcelona - Dreta de l´Eixample	None	FLAT	En pleno centro de Barcelona, justo al lado de...	77	2.0	1.0
3	435000	Piso Barcelona - corts catalanes. Soleado, cén...	Barcelona - Sant Antoni	None	FLAT	Vivienda espaciosa en Sant Antoni, cerca de Pl...	96	3.0	2.0
4	410000	Piso en Carrer de sardenya 271. Alto, reformad...	Barcelona - Sagrada Família	Carrer de Sardenya 271	FLAT	En el corazón de Barcelona, en una hermosa fin...	84	2.0	1.0

Cleaned data

Distribution of Data

- **Property Prices:** Left-skewed distribution, indicating some low-value outliers.
- **Size Distribution:** Appears normally distributed, with the majority of properties featuring sizes between 50 to 75 square meters.
- **Bedroom Count:** Most properties have 0 to 5 bedrooms.
- **Bathroom Count:** Most properties have 1 to 2 bathrooms.
- **Location Frequency:** Indicates a varied distribution across different neighborhoods, with ‘Sagrada Familia’ having significantly more properties.
- **Property Types:** Majority of properties are ‘Flats’.



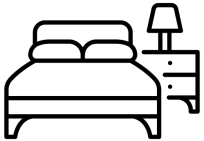
Understanding the Data

866

Listings

€343,885.86

Avg Sale Price



2.72

Avg Bedrooms

1.42

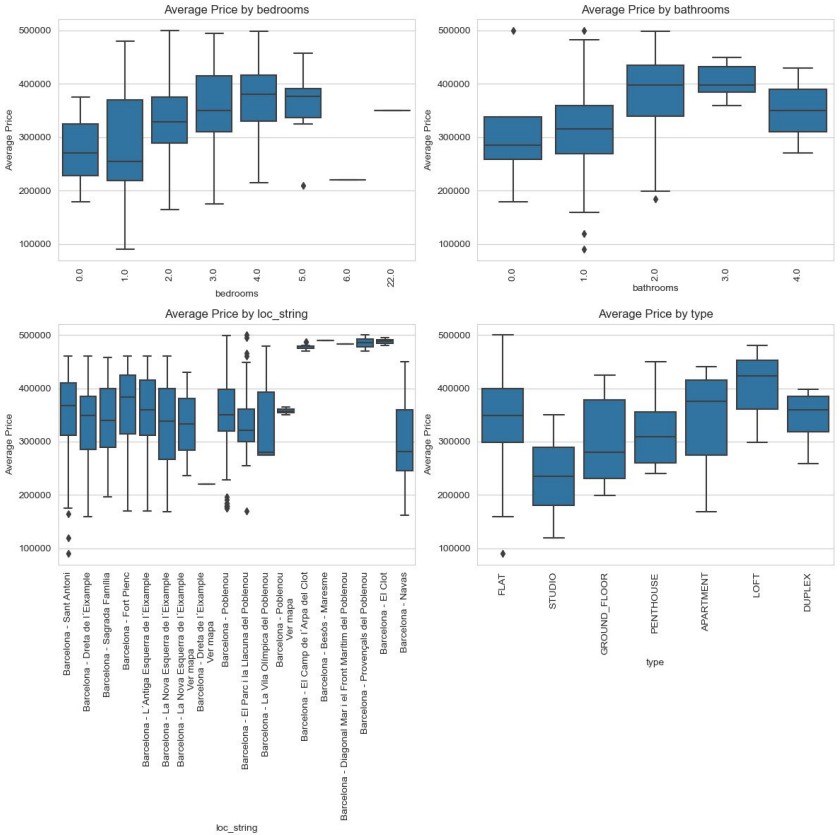
Avg Bathrooms

80.6 m²

Avg Size

Variance Across Attributes

- Price variance across bedrooms, bathrooms, location, type, and m² suggests all attributes may be predictors of property value.



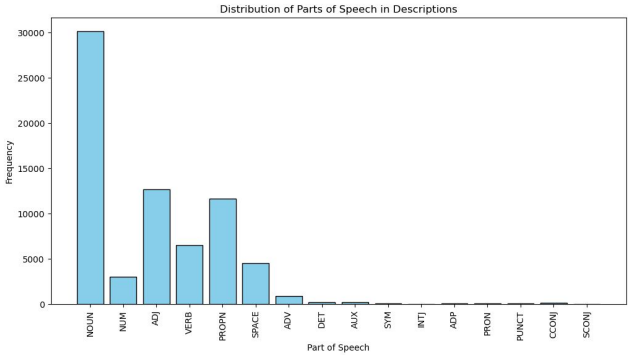
02

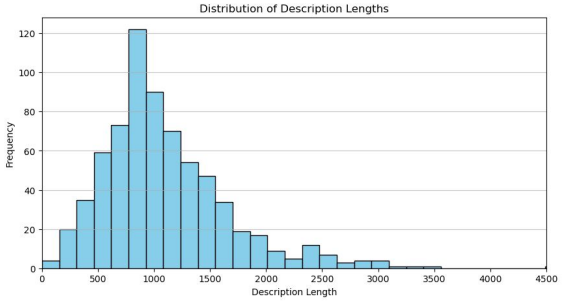
Word Embeddings

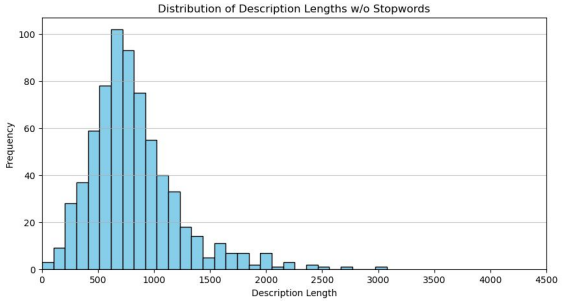
Numerically encoding description feature

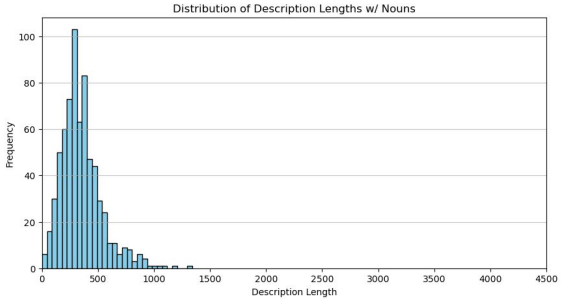
Description

- The "Description" column provides attributes of each property.
- All descriptions are provided in Spanish.
- A number of properties feature extensive descriptions.
- Nouns are the most frequent part of speech observed in the descriptions.



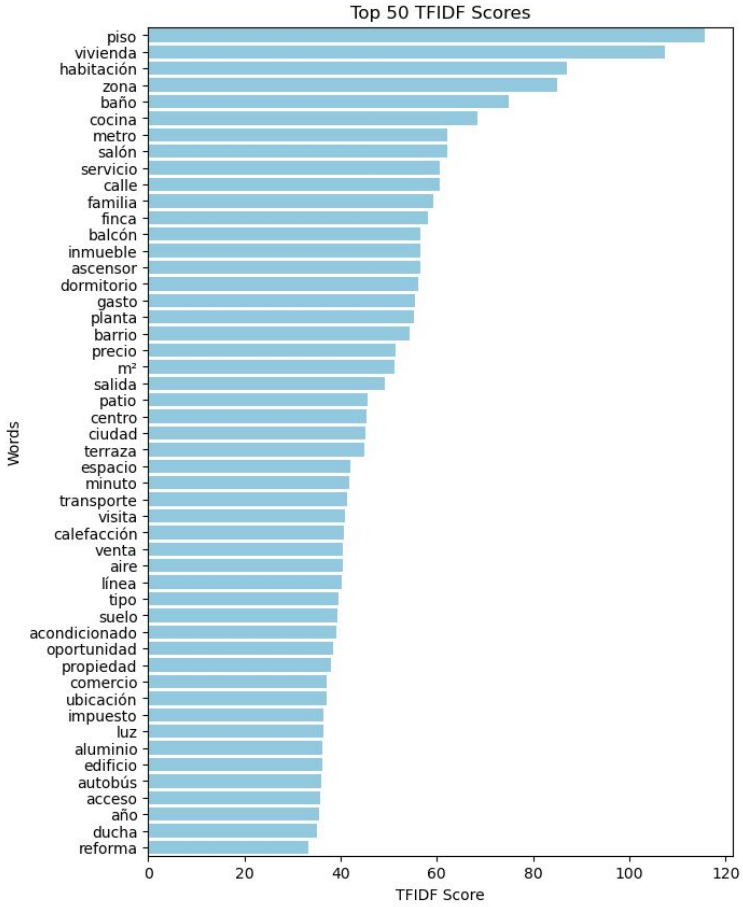






Creating Word Embeddings

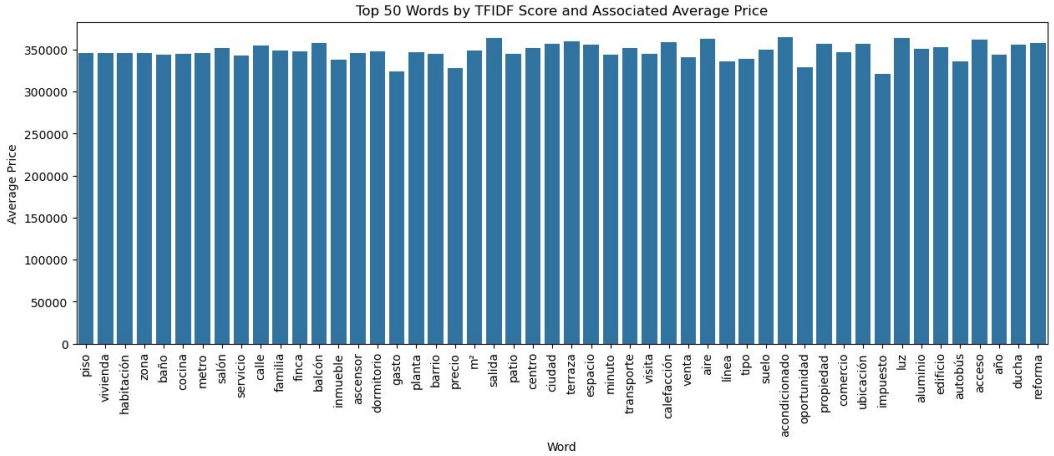
1. **Text Cleaning**
   - Remove whitespace and other non-text characters.
2. **Stopword Removal with SpaCy**
   - Eliminated common Spanish stopwords to focus on meaningful content.
3. **Focus on Key Parts of Speech**
   - Retained only nouns to emphasize actual features of the house, recognizing their importance in describing housing characteristics.
4. **Lemmatization**
   - Normalized words to their base form to reduce dimensionality and consolidate similar meanings.
5. **TF-IDF Application**
   - Calculated TF-IDF scores on the training dataset, selecting the top 50 words to highlight those most distinctive to housing.
6. **One-Hot Encoding**
   - Incorporated the top 50 TF-IDF words as one-hot encoded features into the data.



Creating Word Embeddings

	desc_piso	desc_vivienda	desc_habitación	desc_zona	desc_baño	desc_exterior	desc_cocina	desc_doble	desc_amplio	desc_metro	...
0	1	1	0	0	1	0	0	0	0	0	...
1	1	0	1	0	1	1	1	1	1	0	...
2	1	1	1	1	1	0	0	0	0	1	...
3	1	1	0	0	1	0	0	1	1	0	...
4	1	1	1	1	1	0	1	1	0	1	...
...	...	...	...	...	...	...	...	...	...	...	...

- Price variance exists across TFIDF embeddings



03

Modeling

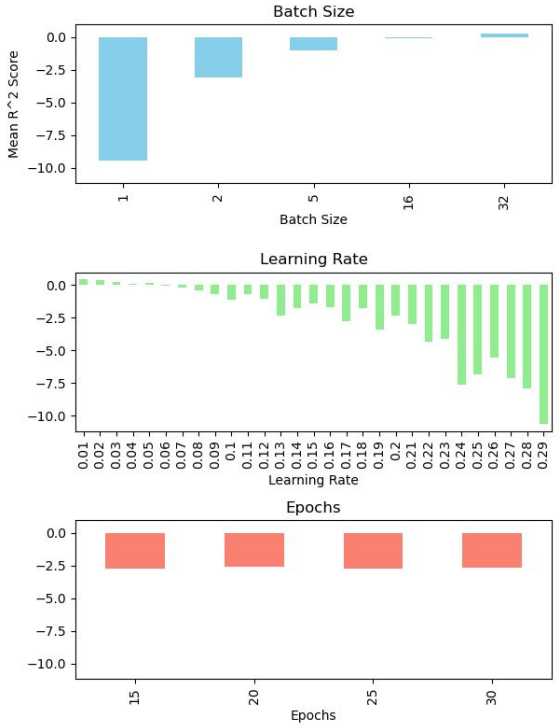
Experimented with a variety of ML models

Methodology Overview

1. **Data Preparation**
   - Split the training data into train/validation sets with an 80/20 ratio.
   - Cleaned the dataset, applied one-hot encoding to categorical variables (type and location), and normalized numerical features (square meters, bedrooms, bathrooms).
   - Dropped irrelevant features and created word embeddings for property descriptions.
2. **Model Development**
   - Developed four predictive models: Random Forest, CatBoost, Boosted Trees, and Linear Regression.
3. **Hyperparameter Tuning**
   - Initial Exploration: Investigated the impact of each model's hyperparameters against their default settings.
   - Narrowing Down: Utilized findings to refine the scope for grid search.
   - Grid Search: Conducted exhaustive grid search with 3-fold cross-validation to identify optimal hyperparameter combinations.
   - Selection Criterion: Selected the best hyperparameters based on the highest R^2 score across the 3-fold cross-validation.
4. **Ensemble Approach**
   - Combined the best-performing instances of each model into an ensemble, leveraging their collective strength to enhance prediction accuracy.

Linear Regression

1. **Hyperparameter Optimization:** Utilized grid search with 3-fold cross-validation to pinpoint optimal hyperparameters, selecting those yielding the highest R^2 Score



Note: Negative mean R^2 scores per hyperparameter region frequently indicate poor model predictability.

```
param_grid = {
    'batch_sizes': [1, 2, 5, 16, 32],
    'learning_rate': np.arange(0.01, 0.3, 0.01),
    'epoch_size': [15, 20, 25, 30]
}
```

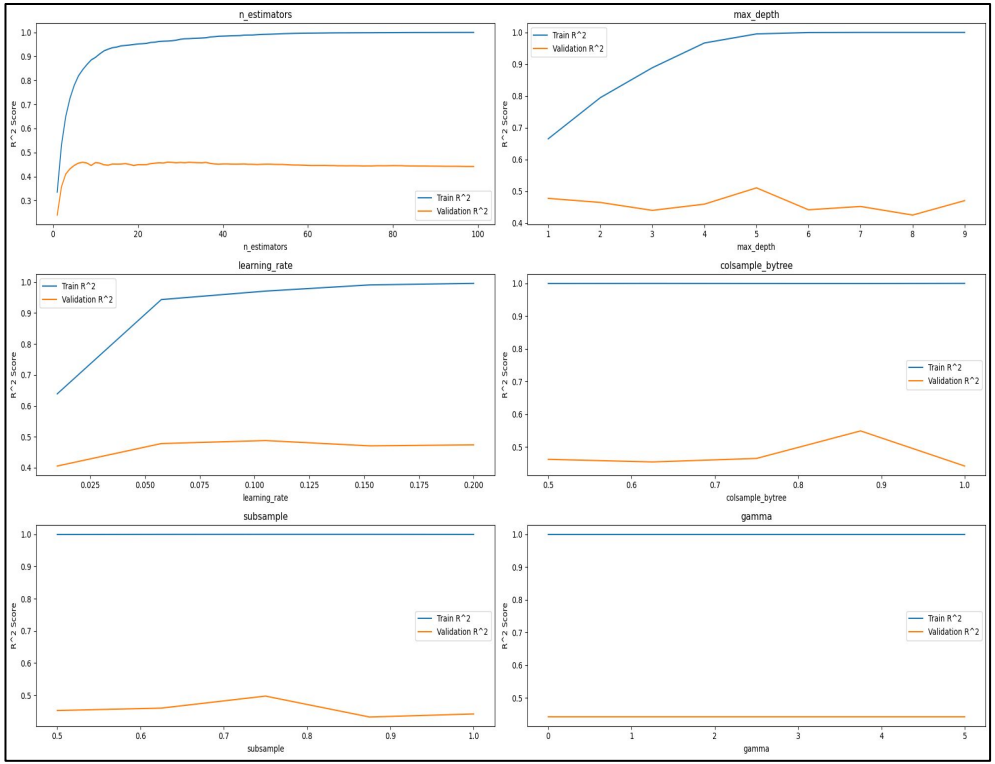
2. **Validation Performance:** Applied the best hyperparameters to evaluate the model's R^2 score on the validation set.

```
Best parameters = {
    'batch_sizes': 16,
    'learning_rate': 0.01,
    'epoch_size': 15
}

Val R^2 score: 0.4516
```

Boosted Trees

1. **Hyperparameter Preliminary Analysis:** Assessed the impact of varying hyperparameters from defaults and refined the scope for targeted grid search based on insights gained.



2. **Hyperparameter Optimization:** Utilized grid search with 3-fold cross-validation to pinpoint optimal hyperparameters, selecting those yielding the highest R^2 Score

```
param_grid = {
    'n_estimators': [15, 30, 50, 100],
    'max_depth': [2, 3, 4, 5],
    'learning_rate': np.arange(0.05, .14, 0.02),
    'colsample_bytree': [0.5, 0.85, 0.9],
    'subsample': np.linspace(0.6, 0.9, 5)
}
```

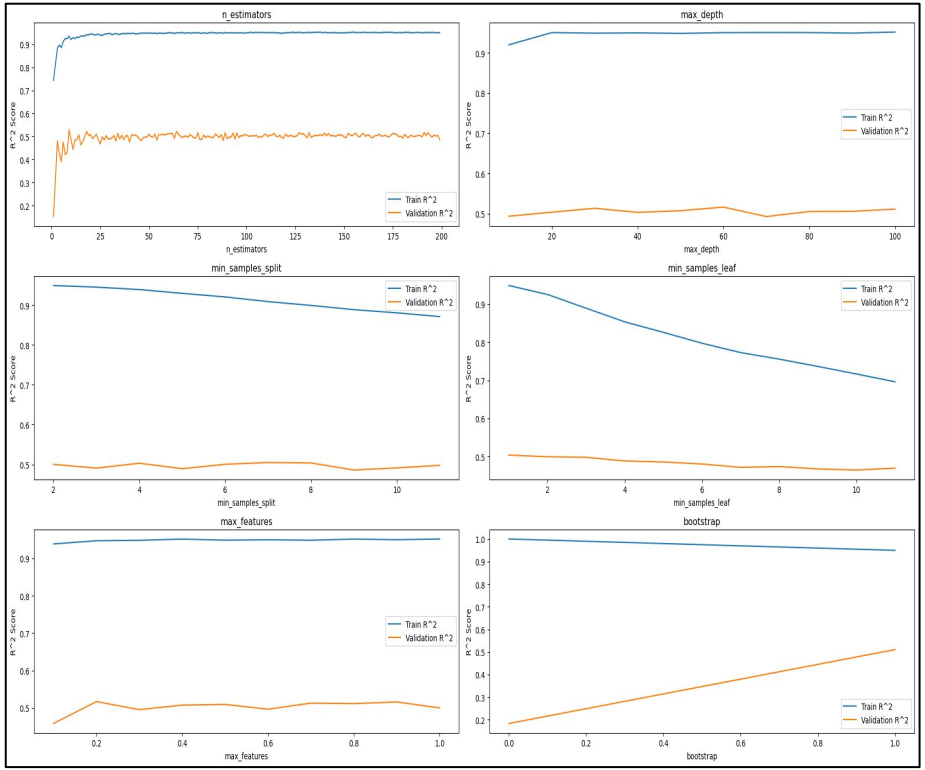
3. **Validation Performance:** Applied the best hyperparameters to evaluate the model's R^2 score on the validation set.

```
Best parameters = {
    'n_estimators': 100,
    'max_depth': 5,
    'learning_rate': 0.09,
    'colsample_bytree': 0.85,
    'subsample': 0.6
}

Val R^2 score: 0.5098
```

Random Forest

1. **Hyperparameter Preliminary Analysis:** Assessed the impact of varying hyperparameters from defaults and refined the scope for targeted grid search based on insights gained.



2. **Hyperparameter Optimization:** Utilized grid search with 3-fold cross-validation to pinpoint optimal hyperparameters, selecting those yielding the highest R^2 Score

```
param_grid = {
    'n_estimators': range(25, 150, 25),
    'max_depth': [None, 30, 45, 60],
    'min_samples_split': np.arange(2, 12, 1),
    'min_samples_leaf': np.arange(1, 5, 1),
    'max_features': np.arange(0.2, 0.6, 0.1),
    'bootstrap': [False]
}
```

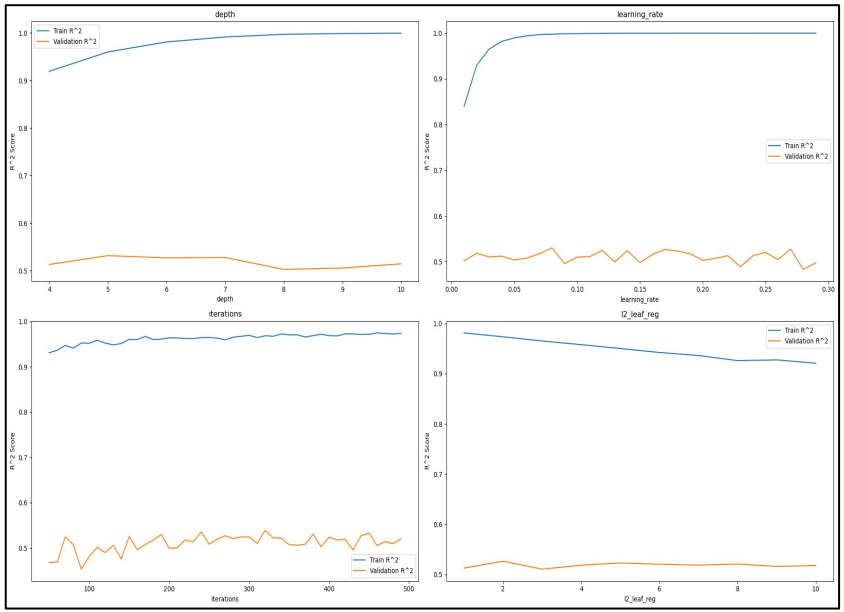
3. **Validation Performance:** Applied the best hyperparameters to evaluate the model's R^2 score on the validation set.

```
Best parameters = {
    'n_estimators': 125,
    'max_depth': 45,
    'min_samples_split': 3,
    'min_samples_leaf': 1,
    'max_features': 0.4,
    'bootstrap': False
}

Val R^2 score: 0.5244
```

CatBoost

1. **Hyperparameter Preliminary Analysis:** Assessed the impact of varying hyperparameters from defaults and refined the scope for targeted grid search based on insights gained.



2. **Hyperparameter Optimization:** Utilized grid search with 3-fold cross-validation to pinpoint optimal hyperparameters, selecting those yielding the highest R^2 Score

```
param_grid = {
    'depth': np.arange(4, 8, 1),
    'learning_rate': np.arange(0.01, 0.3, 0.02),
    'iterations': [50, 150, 300, 500],
    'l2_leaf_reg': [2, 5, 8]
}
```

3. **Validation Performance:** Applied the best hyperparameters to evaluate the model's R^2 score on the validation set.

```
Best parameters = {
    'depth': 6,
    'learning_rate': 0.05,
    'iterations': 500,
    'l2_leaf_reg': 8
}

Val R^2 score: 0.5222
```

Final Model: Optimized Ensemble Approach

- **Ensemble Composition:** Utilized a strategic blend of four distinct models:
  - Random Forest
  - CatBoost
  - Boosted Trees
  - Linear Regression
- **Methodology:** Conducted comprehensive experimentation across all potential model combinations to identify the ensemble with superior performance.
- **Selection Criteria:** Chose the final ensemble model based on achieving the highest validation R^2 score, indicative of optimal predictive accuracy.
- **Top Performing Ensemble:** The combination of **Random Forest + CatBoost** emerged as the best model, showcasing the strength of integrating diverse algorithms.

Models	Val R^2
Random Forest + CatBoost	0.532823
Boosted Trees + Random Forest + CatBoost	0.531479
Linear Regression + Boosted Trees + Random Forest + CatBoost	0.529334
Linear Regression + Random Forest + CatBoost	0.529085
Boosted Trees + Random Forest	0.528748
Linear Regression + Boosted Trees + Random Forest	0.526199
Boosted Trees + CatBoost	0.523364
Linear Regression + Random Forest	0.522421
Linear Regression + Boosted Trees + CatBoost	0.518429
Linear Regression + CatBoost	0.509765
Linear Regression + Boosted Trees	0.504244

04

Lessons Learned

Key Takeaways for future projects

Lessons Learned:

1. **Ensembling Enhances ML Models**
    - Utilizing ensembling techniques in machine learning models consistently enhanced performance.
2. **Understanding the Data is Important**
    - Focusing time exploring the data allowed us to be thoughtful about model choices
3. **Predicting Housing Prices is Challenging!**