Predicción de Precio/m2

Santiago Fernández del Castillo Sodi

1. Preprocesamiento
2. Análisis exploratorio
3. Modelos Probados
4. Modelo extra
5. Mejores Modelos
6. Limitaciones
7. Conclusiones

Preprocesamiento y limpieza de datos

Limpieza:

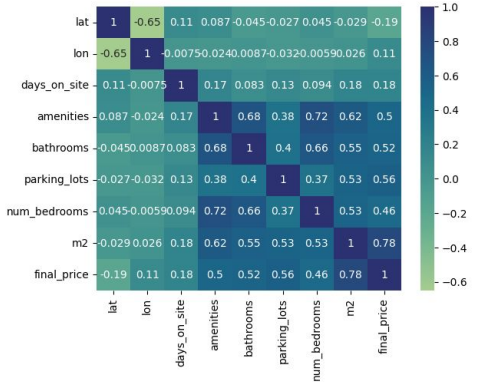
- Se eliminaron variables con un solo valor, variables no imputables, no cuantificables, poco informativas y repetitivas.
- Se eliminó la variable precio por metro cuadrado en favor de predecir el precio y de ahí calcular el precio/m2.
- Variables no imputables y nulas, se eliminan.

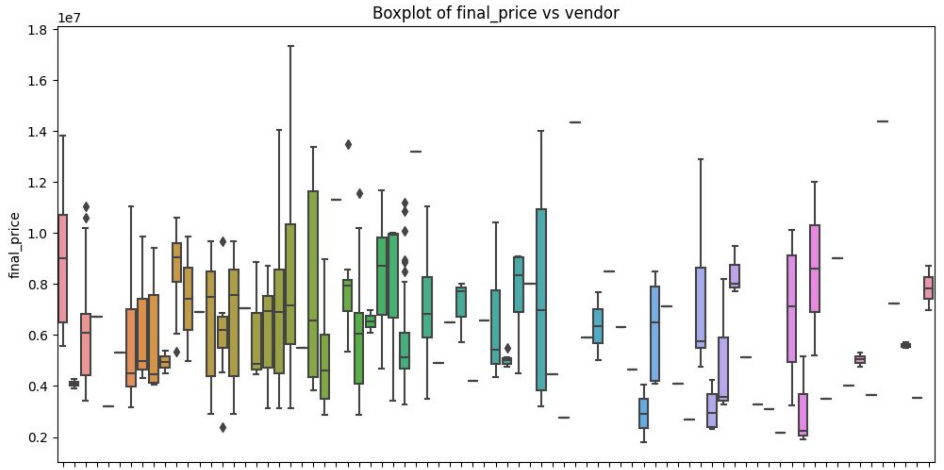
Preprocesamiento:

- Utilizando la columna address y regex, se obtienen la colonia y estado a los que pertenece el departamento.
- Utilizando colonia, estado y coordenadas, se imputa la colonia y estado de aquellos apartamentos sin datos en ese rubro.

Análisis exploratorio

- Visualizar las coordenadas para facilitar comprenderlas.
- Visualizar las distribuciones y correlaciones para identificar outliers y técnicas para eliminarlos.
- Visualizar distribución de la variable objetivo para cada variable categórica.





Modelos Probados (y descartados)

- Utilizando GridSearchCV con sus limitaciones

	model	best_score	best_params	RMSE	MAPE
0	RandomForest	0.7655	{'max_depth': 10, 'min_samples_leaf': 1, 'n_estimators': 50}	1125958	10.93
1	Ridge	0.7094	{'alpha': 1}	1254854	14.20
2	XGBoost	0.7702	{'learning_rate': 0.2, 'n_estimators': 100}	1110950	11.93

Vendor

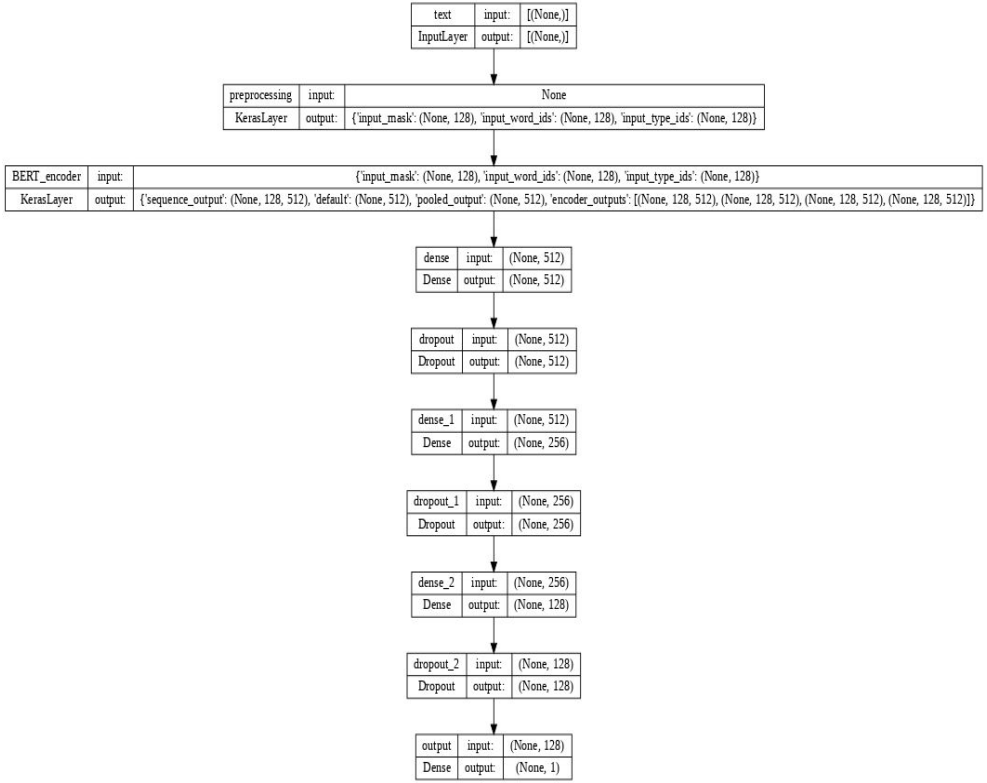
	model	best_score	best_params	RMSE	MAPE
1	Ridge	0.6472	{'alpha': 1}	1739388	20.53
3	SGD	0.6627	{'max_iter': 10000, 'tol': 0.0001}	1733861	20.43

No Vendor

Modelo Extra: NeuReal Estate

- BERT:
  - Predicción del precio por m2 a partir de la descripción, el texto con la dirección, el nombre del anuncio y del vendedor.
  - Embeddings del texto.
  - Activación: ReLu, Sigmoide
  - Optimizador: AdamW

```
Mean Squared Error (MSE): 2090950441833.2131
Mean Absolute Error (MAE): 973244.4410
R-squared (R^2) Score: 0.7334
```



Mejores Modelos

- Sin usar vendedores/anunciantes
- RandomForest:
  - n_estimators=150
  - criterion='squared_error',
  - max_depth=None
  - min_samples_split=2
  - min_samples_leaf=1
- xGBoost:
  - loss='squared_error'
  - learning_rate=0.1,
  - n_estimators=100,
  - criterion='friedman_mse'
  - min_samples_split=2
  - min_samples_leaf=1

model	best_score	best_params	RMSE	MAPE
XGBoost	0.8042	{'learning_rate': 0.1, 'n_estimators': 200}	1593697	18.28
RandomForest	0.7912	{'max_depth': None, 'min_samples_leaf': 1, 'n_estimators': 150}	1644395	18.30

Limitaciones

- Pocos datos
- Datos incompletos
- Datos poco imputables/aumentables
- Poco conocimiento del problem domain
- Mucho desbalance
- Distribuciones
- Baja interpretabilidad

Mejores Modelos: Intepretación

Random Forest

```
Feature ranking:
Feature m2, Importance: 0.709518093242232
Feature lon, Importance: 0.06632210821790104
Feature lat, Importance: 0.061022563266130224
Feature days_on_site, Importance: 0.03872989584200082
Feature amenities, Importance: 0.03745586305572415
Feature parking_lots, Importance: 0.023692265747461312
Feature bathrooms, Importance: 0.015528931423818252
Feature num_bedrooms, Importance: 0.011587384282209205
Feature roma norte, Importance: 0.010552725454537823
Feature distrito federal, Importance: 0.007865861953335388
Feature roma, Importance: 0.0065388329824567085
Feature nuevo leon, Importance: 0.005794599973684581
Feature roma_sur, Importance: 0.004559157917662823
Feature otra, Importance: 0.0005679365650354097
Feature baja california, Importance: 0.00024202532070954062
Feature baja california, Importance: 2.1754755100681833e-05
```

xGBoost

```
Feature ranking:
Feature m2, Importance: 0.6983350146232654
Feature lat, Importance: 0.08197443807860617
Feature lon, Importance: 0.043346726583167794
Feature parking_lots, Importance: 0.04120724901833418
Feature amenities, Importance: 0.03861466230523558
Feature num_bedrooms, Importance: 0.019147254714339583
Feature days_on_site, Importance: 0.0185704303112272
Feature bathrooms, Importance: 0.018157972427827008
Feature roma norte, Importance: 0.017456061100212267
Feature nuevo leon, Importance: 0.009203313459046491
Feature distrito federal, Importance: 0.008563296555985209
Feature roma, Importance: 0.0045064068044730395
Feature roma_sur, Importance: 0.000887214736197684
Feature otra, Importance: 2.9959282082431492e-05
Feature baja california, Importance: 0.0
Feature baja california, Importance: 0.0
```