Análisis Predictivo

TP-2

Juan Ignacio Scorza

2023 1Q

Score Obtenido

3	—	JUAN IGNACIO SCORZA		0.71877	21	3h

EDA

1. Observaciones generales
2. Nulls/Missings
3. Outliers
4. Target y Correlaciones
5. Creacion de Variables/ Encoding

EDA

1 Observaciones generales

Shape

```
df.shape
```

```
(4928, 68)
```

4928 filas y 68 columnas

```
duplicates = df.duplicated(subset=df.columns.difference(['id'])).sum()

print("Cantidad de filas duplicadas (excluyendo 'id'):", duplicates)
```

```
Cantidad de filas duplicadas (excluyendo 'id'): 0
```

0 duplicados

EDA

1 Observaciones generales

```
df.dtypes
```

```
id                                int64
source                           object
name                             object
description                      object
neighborhood_overview            object
host_id                           int64
host_name                        object
host_since                       object
host_location                    object
host_about                       object
host_response_time               object
host_response_rate               object
host_acceptance_rate             object
host_is_superhost                object
host_neighbourhood               object
host_listings_count               int64
host_total_listings_count         int64
host_verifications               object
host_has_profile_pic             object
host_identity_verified           object
neighbourhood                    object
neighbourhood_cleansed           object
neighbourhood_group_cleansed    float64
latitude                        float64
longitude                       float64
property_type                    object
room_type                        object
accommodates                      int64
```

```
bathrooms_text                                object
bedrooms                                     float64
beds                                         float64
amenities                                     object
price                                         object
minimum_nights                                 int64
maximum_nights                                 int64
minimum_minimum_nights                         int64
maximum_minimum_nights                         int64
minimum_maximum_nights                         int64
maximum_maximum_nights                         int64
minimum_nights_avg_ntm                       float64
maximum_nights_avg_ntm                       float64
calendar_updated                             float64
has_availability                              object
availability_30                                int64
availability_60                                int64
availability_90                                int64
availability_365                               int64
calendar_last_scraped                         object
number_of_reviews                              int64
number_of_reviews_ltm                          int64
number_of_reviews_l30d                         int64
first_review                                  object
last_review                                   object
review_scores_rating                         float64
review_scores_accuracy                       float64
review_scores_cleanliness                    float64
review_scores_checkin                        float64
review_scores_communication                  float64
review_scores_location                       float64
review_scores_value                          float64
license                                       object
instant_bookable                              object
calculated_host_listings_count                 int64
calculated_host_listings_count_entire_homes    int64
calculated_host_listings_count_private_rooms   int64
calculated_host_listings_count_shared_rooms    int64
reviews_per_month                            float64
```

EDA

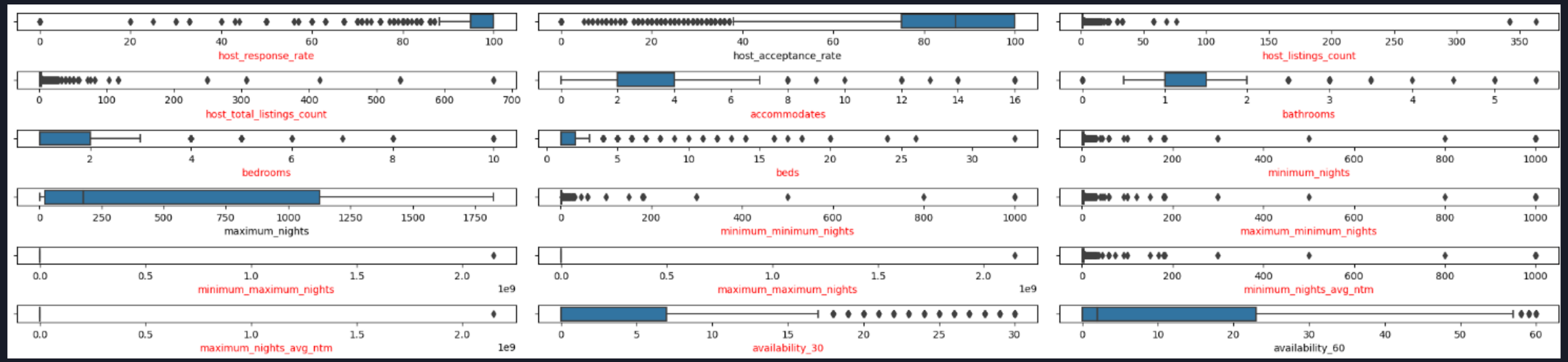
2 Nulls / Missings

	Cantidad de Nulos	Tipo de Dato
description	3	object
neighborhood_overview	1348	object
host_location	421	object
host_about	1746	object
host_response_time	1662	object
host_response_rate	1662	object
host_acceptance_rate	776	object
host_is_superhost	1	object
host_neighbourhood	2124	object
neighbourhood	1348	object
neighbourhood_group_cleansed	4928	float64
bathrooms	4928	float64
bathrooms_text	9	object
bedrooms	240	float64
beds	70	float64
calendar_updated	4928	float64
review_scores_accuracy	4	float64
review_scores_cleanliness	4	float64
review_scores_checkin	4	float64
review_scores_communication	4	float64
review_scores_location	4	float64
review_scores_value	4	float64
license	98	object

- "unknown"/ "none"
- Drop
- Imputados
  - Por variables correlacionadas
  - Por media/mediana
  - Por modelos de regresión

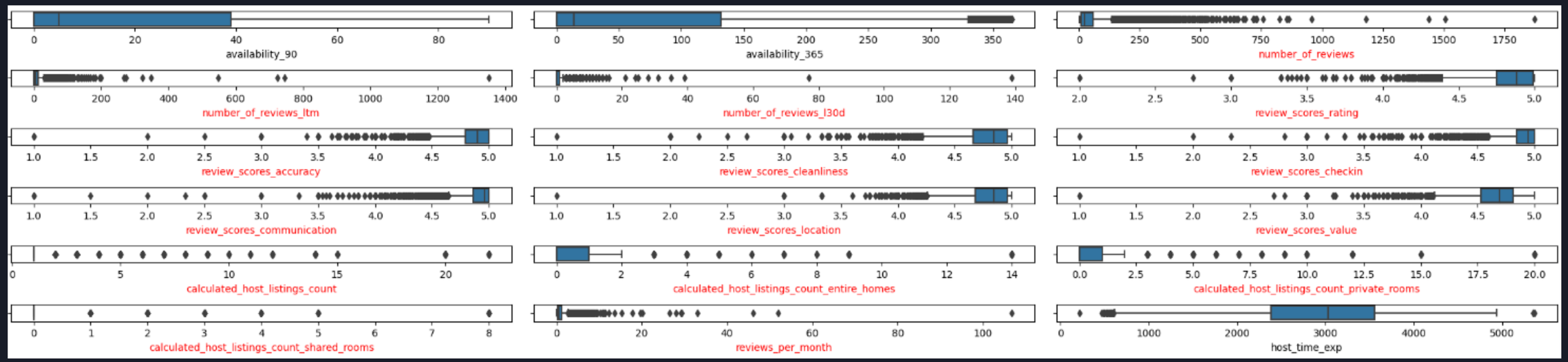
EDA

3 Outliers



EDA

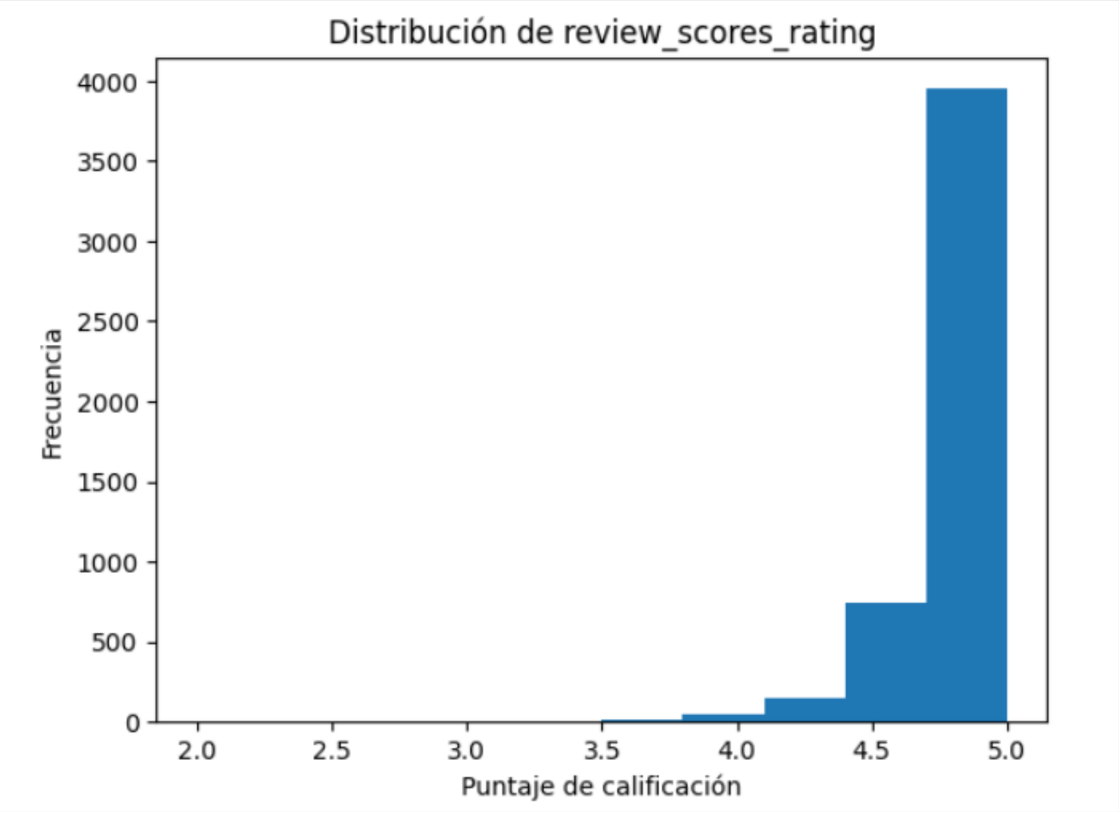
3 Outliers



- No se encuentran valores sin sentido
- Observación minimum nights

EDA

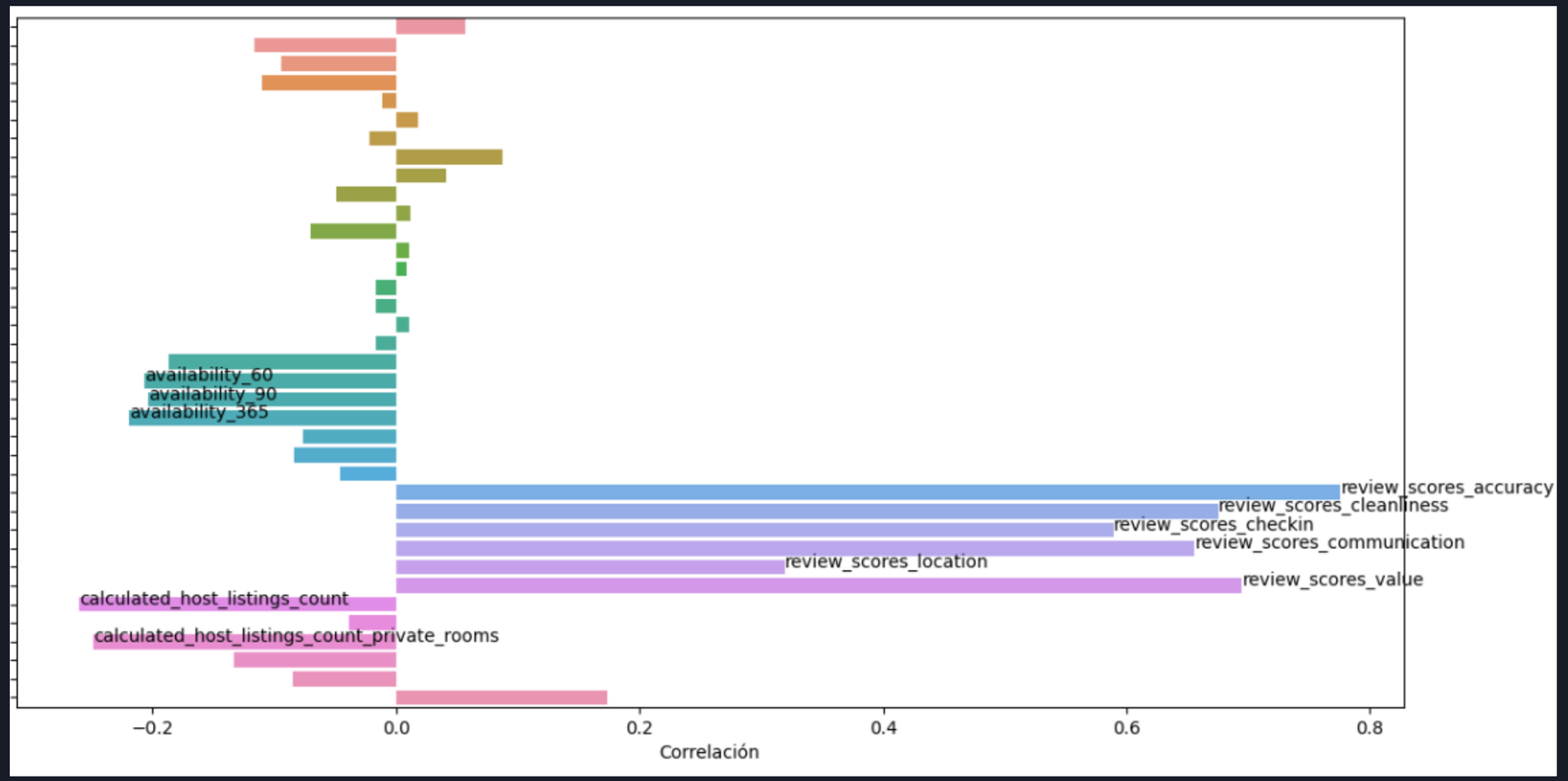
4 Target y Correlaciones



mean	4.816239
std	0.226507
min	2.000000
25%	4.750000
50%	4.880000
75%	4.990000
max	5.000000

EDA

4 Target y Correlaciones

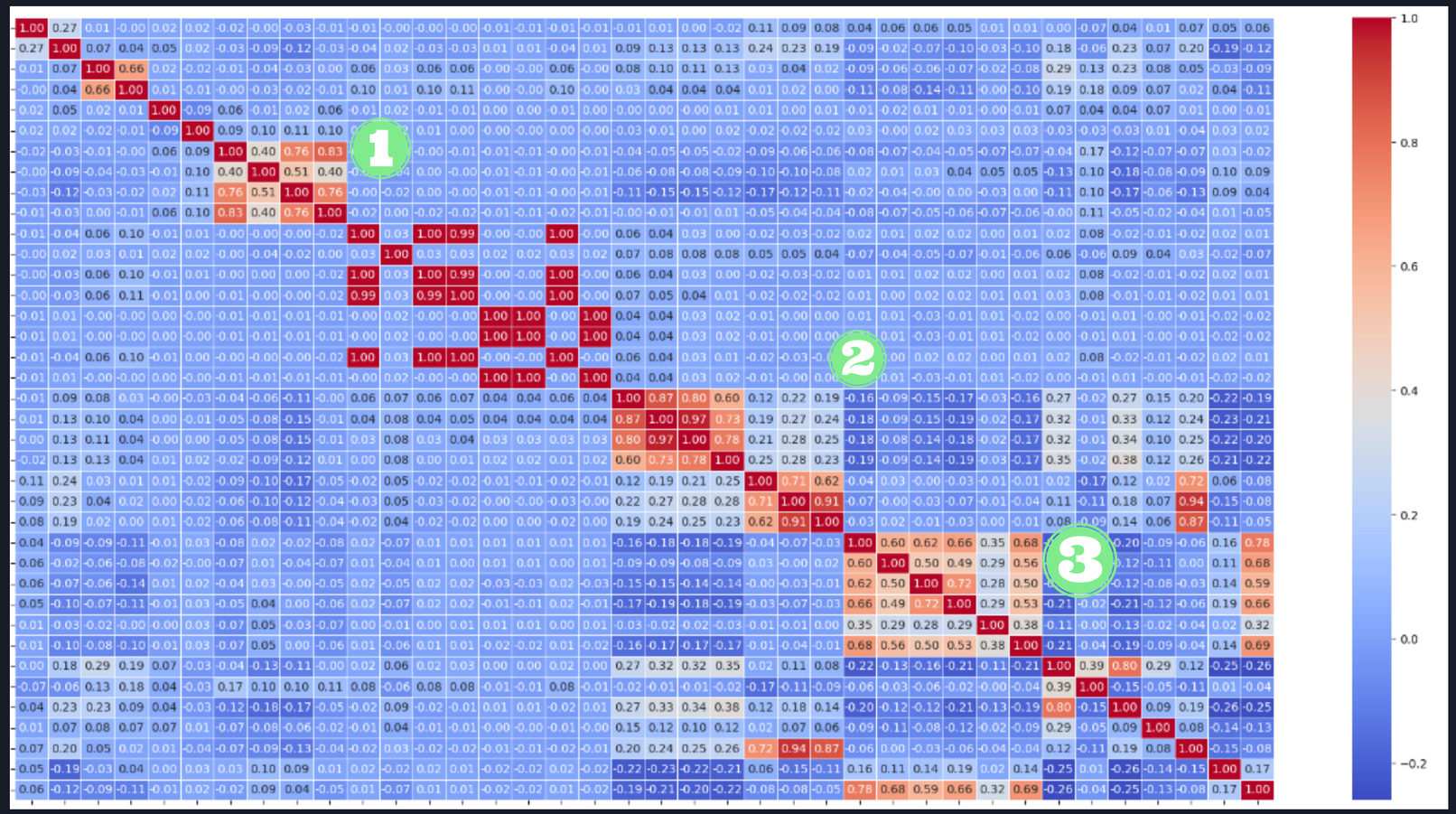


(correlación de la variable target con el resto de variables)

Solo se muestran aquellas con correlacion mayor a 0.2

EDA

4 Target y Correlaciones



1. Accommodates, bathrooms, beds
2. Availability
3. Scores

EDA

5 Creación de Variables/Encoding

Nuevas Variables

- price_num
- shared_bathrooms
- host_time_exp
- host_country
- host_city_or_state
- time_since_first_review
- time_since_last_review
- time_between_reviews

Categórica  Numérica

EDA

5 Creación de Variables/Encoding

```
id: 4924
source: 2
name: 4892
description: 4860
neighborhood_overview: 3278
host_id: 4177
host_name: 2243
host_location: 107
host_about: 2629
host_response_time: 5
host_is_superhost: 2
host_neighbourhood: 54
host_verifications: 6
host_has_profile_pic: 2
host_identity_verified: 2
neighbourhood: 46
neighbourhood_cleansed: 22
property_type: 52
room_type: 4
bathrooms_text: 20
amenities: 4821
price: 478
has_availability: 2
calendar_last_scraped: 2
first_review: 2150
last_review: 868
license: 4076
instant_bookable: 2
host_city_or_state: 106
host_country: 26
```

Variable categórica: valores únicos

Se crean variables binarias para representar las categóricas que tienen 2 valores únicos.

source_bin 1 = 'city scrape' 0 = 'previous scrape'

host_is_superhost_bin 1=t 0=f

host_has_profile_pic_bin 1=t 0=f

host_identity_verified_bin 1=t 0=f

has_availability_bin 1=t 0=f

calendar_last_scraped_bin 1= '2022-12-05' 0= '2022-12-17'

instant_bookable_bin 1= t 0= f

EDA

5 Creación de Variables/Encoding

```
id: 4924
source: 2
name: 4892
description: 4860
neighborhood_overview: 3278
host_id: 4177
host_name: 2243
host_location: 107
host_about: 2629
host_response_time: 5
host_is_superhost: 2
host_neighbourhood: 54
host_verifications: 6
host_has_profile_pic: 2
host_identity_verified: 2
neighbourhood: 46
neighbourhood_cleansed: 22
property_type: 52
room_type: 4
bathrooms_text: 20
amenities: 4821
price: 478
has_availability: 2
calendar_last_scraped: 2
first_review: 2150
last_review: 868
license: 4076
instant_bookable: 2
host_city_or_state: 106
host_country: 26
```

Variable categórica: valores únicos

Frequency Encoding.

frequency encoding para los campos en los que se tiene una cantidad de entre 3 y 110 valores posibles.

host_location
host_neighbourhood
neighbourhood
neighbourhood_cleansed
property_type
room_type
bathrooms_text
host_city_or_state
host_country

EDA

5 Creación de Variables/Encoding

```
id: 4924
source: 2
name: 4892
description: 4860
neighborhood_overview: 3278
host_id: 4177
host_name: 2243
host_location: 107
host_about: 2629
host_response_time: 5
host_is_superhost: 2
host_neighbourhood: 54
host_verifications: 6
host_has_profile_pic: 2
host_identity_verified: 2
neighbourhood: 46
neighbourhood_cleansed: 22
property_type: 52
room_type: 4
bathrooms_text: 20
amenities: 4821
price: 478
has_availability: 2
calendar_last_scraped: 2
first_review: 2150
last_review: 868
license: 4076
instant_bookable: 2
host_city_or_state: 106
host_country: 26
```

Variable categórica: valores únicos

Para las variables que contienen cadenas de texto largas,
nos limitamos a extraer información acerca de si son null o no (‘unknown’ y ‘none’ == null).

Para ello, se crean 5 variables binarias adicionales

has_description_bin

has_neighborhood_bin

has_neighbourhood_overview_bin

has_host_neighbourhood_bin

has_host_about_bin

EDA

5 Creación de Variables/Encoding

- host_response_time ordinal encoding

```
[76] response_time_mapping = {
         'within an hour': 4,
         'within a few hours': 3,
         'within a day': 2,
         'a few days or more': 1,
         'unknown': 0
     }

     df['host_response_time_ord'] = df['host_response_time'].map(response_time_mapping)
```

property_type sacamos si tiene la palabra "entire", si tiene la palabra "private" o "shared"

```
[78] df['property_type_shared_bin'] = df['property_type'].str.contains('Shared', case=False).astype(int)
     df['property_type_entire_bin'] = df['property_type'].str.contains('Entire', case=False).astype(int)
     df['property_type_private_bin'] = df['property_type'].str.contains('Private', case=False).astype(int)
```

Modelos Predictivos

1. Particiones Utilizadas.
2. Métodos de Ajuste.
3. Modelos Utilizados.
4. Mejor Modelo.

Modelos Predictivos

1 Particiones Utilizadas.

PARTICIONES

```
from sklearn.model_selection import train_test_split

model_num_columns = df.select_dtypes(include=['float64', 'int64']).columns.to_list()
model_num_columns.remove('review_scores_rating')

X = df[model_num_columns]
Y = df['review_scores_rating']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.15, random_state = 3124)
```

```
print('data Entrenamiento: ', X_train.shape)
print('data Testeo: ', X_test.shape)
```

```
data Entrenamiento:  (4185, 108)
data Testeo:  (739, 108)
```

Test Size Utilizados

- 0.10
- 0.15 ┐ Mejores
- 0.20 ┘
- 0.25
- 0.30

Modelos Predictivos

2 Métodos de Ajuste

Blackbox optimization

from ***sklearn.model_selection*** import **GridSearchCV**

from **skopt** import **BayesSearchCV**

Modelos Predictivos

3 Modelos Utilizados

Random Forest

- Modelos
  - Modelos sueltos
  - **Random forest**
  - Extra Trees Regressor
  - XGBOOST
  - CATboost
  - LGBM

from **skopt** import **BayesSearchCV**

Score: 74%

kaggle

```
Mejores hiperparámetros: OrderedDict([('max_depth', 20), ('min_samples_leaf', 1), ('min_samples_split', 2), ('n_estimators', 100)])
Mejor score R2: 0.7441140754601511
```

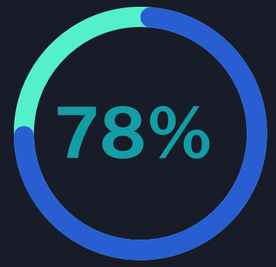
Modelos Predictivos

3 Modelos Utilizados

Modelos

- Modelos sueltos
- Random forest
- **Extra Trees Regressor**
- XGBOOST
- CATboost
- LGBM

Extra Trees Regressor (sin ajuste)

Score: 


```
ext = ExtraTreesRegressor(bootstrap=True).fit(X_train, Y_train)
print("Score-Testeo:", ext.score(X_test,Y_test))
```

```
Score-Testeo: 0.7789980861566295
```

El ajustado es el ganador

Modelos Predictivos

3 Modelos Utilizados

- Modelos
  - Modelos sueltos
  - Random forest
  - Extra Trees Regressor
  - **XGBOOST**
  - CATboost
  - LGBM

XGBOOST

from ***sklearn.model_selection*** import **GridSearchCV**

Score: 81%

kaggle

```
Mejores hiperparámetros: {'gamma': 0.1, 'learning_rate': 0.005, 'max_depth': 11, 'n_estimators': 2000}
R2:  0.815839379663388
Mean Squared Error:  0.008452383121427475
Explained Variance Score:  0.8162555580610356
```

Modelos Predictivos

3 Modelos Utilizados

- Modelos
  - Modelos sueltos
  - Random forest
  - Extra Trees Regressor
  - XGBOOST
  - **CATboost**
  - LGBM

CATBoost

from **skopt** import **BayesSearchCV**

Score: 79%

kaggle

```
Mejores hiperparámetros encontrados:
OrderedDict([('depth', 3), ('iterations', 100), ('l2_leaf_reg', 4.111115502600373), ('learning_rate', 0.15299015008733383), ('subsample', 1.0)])
R2 con los mejores hiperparámetros:
0.743502090534664
R2 del modelo final en los datos de prueba:
0.7922601011747162
```

Modelos Predictivos

3 Modelos Utilizados

LGBM

from **skopt** import **BayesSearchCV**

Modelos

- Modelos sueltos
- Random forest
- Extra Trees Regressor
- XGBOOST
- CATboost
- **LGBM**

Score: 74%

kaggle

```
Mejores hiperparámetros: OrderedDict([('colsample_bytree', 0.31725165744015615), ('learning_rate', 0.055301030608470435), ('max_depth', 5), ('min_child_samples', 43), ('n_estimators', 173), ('
Mejores r2: 0.7269851498915154
R2 de la pred:  0.7743178774488557
Mean Squared Error:  0.011264331166561366
Explained Variance Score:  0.7743662738733922
```

Modelos Predictivos

3 Mejor Modelo

ExtraTreesRegressor

from ***sklearn.model_selection*** import **GridSearchCV**

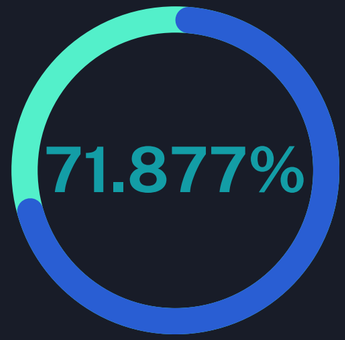
```
extra = ExtraTreesRegressor(bootstrap=True, min_samples_split=12, n_estimators=200)
extra.fit(X_train, Y_train)
extra.score(X_test, Y_test)
```

```
0.8012948716443005
```

min_samples_split = 12

n_estimators = 200

kaggle



res (6).csv Complete · 5d ago	0.71877	0.72901	

Modelos Predictivos

3 Mejor Modelo bis

ExtraTreesRegressor

from **skopt** import **BayesSearchCV**

```
0.6572751697850363
best r2 0.7534107989254445
ExtraTreesRegressor(max_depth=10, max_features=0.9296922685720606,
                    n_estimators=1000)
```

'max_depth'= 10

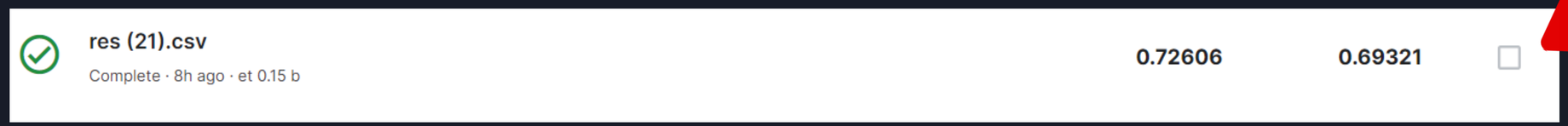
'max_features'= 0.9296922685720606

'min_samples_leaf'= 1

'min_samples_split'= 2

'n_estimators'= 1000

kaggle



No fue elegido