ITBA
Instituto Tecnológicc
de Buenos Aires

82.05 - Análisis Predictivo

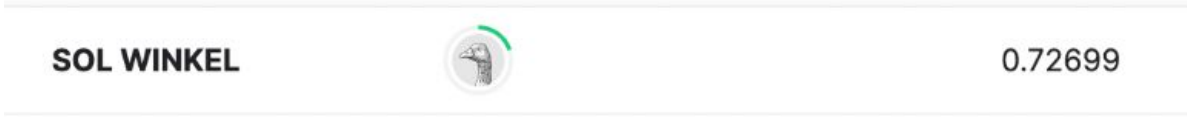
Trabajo práctico 2

–

SOL ALEJANDRA WINKEL

Score obtenido

0.67125



Análisis exploratorio

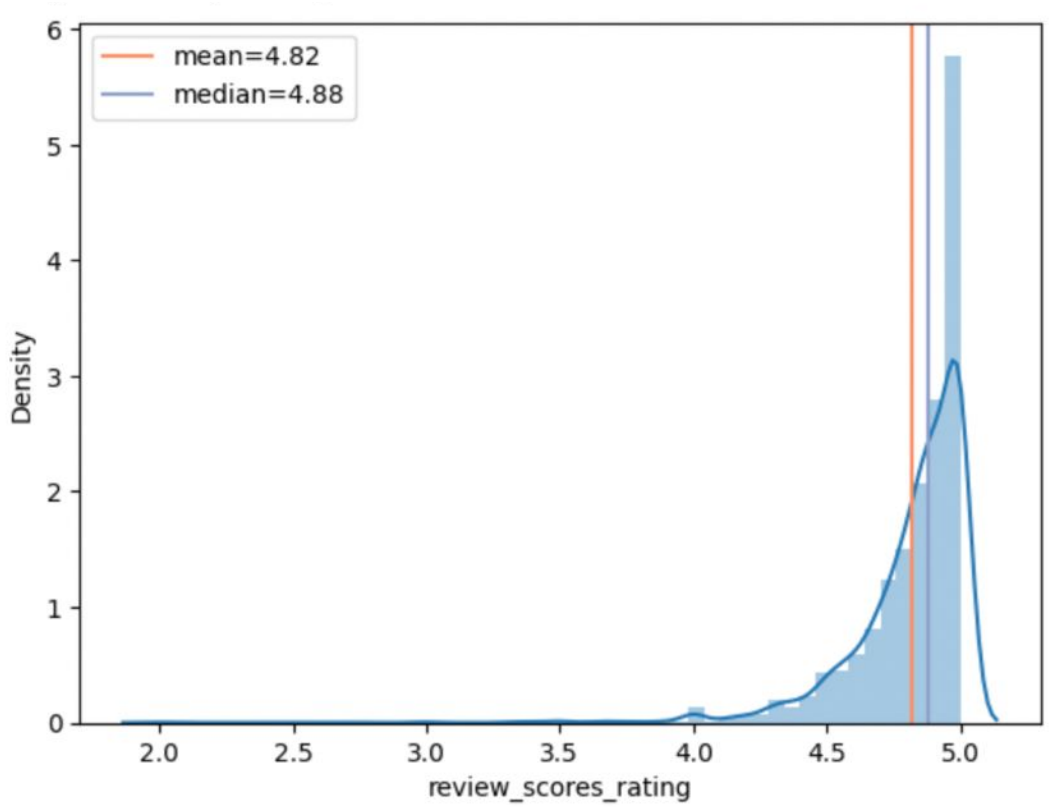
calendar_last_scraped	number_of_reviews	number_of_reviews_ltm	number_of_reviews_l30d	first_review	last_review	review_scores_rating
1	248	37	3	2016	2022	4.85
1	314	29	1	2009	2022	4.89
1	339	0	0	2010	2020	4.44
1	243	17	5	2012	2022	4.94
1	454	84	7	2010	2022	4.88
1	602	82	3	2011	2022	4.79
1	44	3	0	2010	2022	4.72
1	451	59	5	2010	2022	4.92
1	363	50	2	2011	2022	4.87
1	95	0	0	2011	2020	4.86

La **cantidad de filas** es 4928 y la **cantidad de columnas** es 68

Variable target

Objetivo: predecir el puntaje de cada uno de los alojamientos (de 0 a 5)

Variable target: review_scores_raiting



Preparando los datos

Formas de tratar las variables (modificación de las mismas para obtener información):

- Análisis de sentimiento (librería TextBlob)
- Contar caracteres (longitud)
- Reemplazar “t” y “f” por ‘0’ y ‘1’
- Reemplazar dos categorías por ‘0’ y ‘1’
- Armar diccionario
- Borrar símbolos ($ o %) y convertir de string a float
- Contar cantidad de palabras separadas por coma (características)
- Creación de nuevas variables que otorgan valor (por ej: antiguedad_host)

```
df.dtypes
```

```
id                                int64
source                            int64
name                            float64
description                       int64
neighborhood_overview           float64
host_id                           int64
host_location                     int64
host_about                      float64
host_response_time                int64
host_response_rate              float64
host_acceptance_rate            float64
host_is_superhost                 int64
host_listings_count               int64
host_total_listings_count         int64
host_verifications                int64
host_has_profile_pic              int64
host_identity_verified            int64
neighbourhood_cleansed            int64
latitude                        float64
longitude                       float64
room_type                         int64
accommodates                      int64
bathrooms_text                    int64
bedrooms                        float64
beds                            float64
amenities                         int64
price                           float64
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
has_availability                  int64
availability_30                   int64
availability_60                   int64
availability_90                   int64
availability_365                  int64
calendar_last_scraped             int64
number_of_reviews                 int64
number_of_reviews_ltm             int64
number_of_reviews_l30d            int64
first_review                      int64
last_review                       int64
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
instant_bookable                  int64
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
antiguedad_host                 float64
dtype: object
```

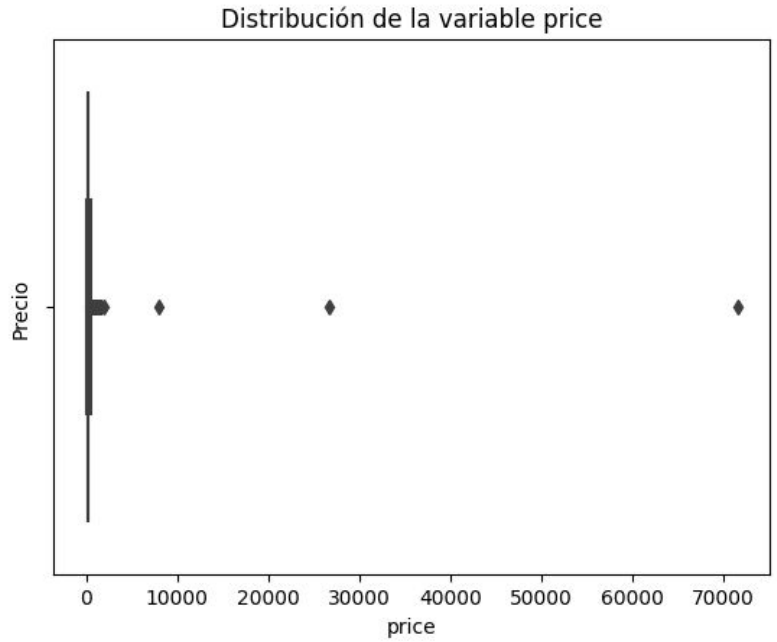
Datos faltantes (missings)

```
La columna neighbourhood_group_cleansed tiene 4928 valores nulos
La columna bathrooms tiene 4928 valores nulos
La columna calendar_updated tiene 4928 valores nulos
La columna host_neighbourhood tiene 2124 valores nulos
La columna host_about tiene 1746 valores nulos
La columna host_response_time tiene 1662 valores nulos
La columna host_response_rate tiene 1662 valores nulos
La columna neighborhood_overview tiene 1348 valores nulos
La columna neighbourhood tiene 1348 valores nulos
La columna host_acceptance_rate tiene 776 valores nulos
La columna host_location tiene 421 valores nulos
La columna bedrooms tiene 240 valores nulos
La columna license tiene 98 valores nulos
La columna beds tiene 70 valores nulos
La columna bathrooms_text tiene 9 valores nulos
La columna review_scores_accuracy tiene 4 valores nulos
La columna review_scores_cleanliness tiene 4 valores nulos
La columna review_scores_checkin tiene 4 valores nulos
La columna review_scores_communication tiene 4 valores nulos
La columna review_scores_location tiene 4 valores nulos
La columna review_scores_value tiene 4 valores nulos
La columna description tiene 3 valores nulos
La columna host_is_superhost tiene 1 valores nulos
La columna id tiene 0 valores nulos
La columna source tiene 0 valores nulos
La columna name tiene 0 valores nulos
La columna host_id tiene 0 valores nulos
La columna host_name tiene 0 valores nulos
La columna host_since tiene 0 valores nulos
```

Formas de imputar:

- Eliminación de columna (todos los valores missings)
- En variables de texto que se utiliza análisis de sentimiento se reemplazan por 0, considerando que son valores neutrales
- En variables de longitud se reemplaza por 0
- En variables categóricas se crea categoría para los valores nulos
- En variables numéricas se imputa por moda, media o mediana
- Se eliminan registros nulos de variables review ya que los nulos se encuentran en la misma fila

Outliers




```
Valor máximo: 71536.0
------------------
```

Elimino valor máximo debido a que al ser precios por noche asumo que es un error

```
------------------
Columna: minimum_maximum_nights
Valor mínimo: 1
Valor máximo: 2147483647
------------------
Columna: maximum_maximum_nights
Valor mínimo: 1
Valor máximo: 2147483647
------------------
Columna: minimum_nights_avg_ntm
Valor mínimo: 1.0
Valor máximo: 1001.0
------------------
Columna: maximum_nights_avg_ntm
Valor mínimo: 1.0
Valor máximo: 2147483647.0
------------------
```

Elimino valor máximo de las tres columnas. Este valor se encuentra en la misma fila. Asumo que es un error

Modelos utilizados

Para la evaluación de los modelos se divide la base y se destina el 15% a test.

- 4.182 registros en el conjunto de entrenamiento
- 739 registros en el conjunto de testeo

Luego, se predicen 1233 valores de la base val y se evalúan en Kaggle

Se prueban los modelos

- LinearRegression
- DecisionTreeRegressor
- RandomForestRegressor
- XGBRegressor
- CatBoostRegressor
- LGBMRegressor

Se utilizó **GridSearch** para ajustar hiper parámetros y obtener la mejor combinación junto con Cross Validation

Modelo ganador - Modelo XGBoost

Elección de hiper parámetros - Grid Search y Cross Validation

```
param_grid = {
    'nthread': [4],
    'objective': ['reg:squarederror'],
    'learning_rate': [0.1, 0.05],
    'max_depth': [2, 3, 5],
    'min_child_weight': [5, 7, 9],
    'silent': [1],
    'subsample': [0.8],
    'colsample_bytree': [0.8],
    'n_estimators': [100, 200, 300]
}

model = xgb.XGBRegressor()
grid_search = GridSearchCV(estimator=model, param_grid=param_grid, cv=5)
grid_search.fit(X, y)
```

'nthread': Número de subprocesos para entrenar el modelo

'objective': Función objetivo utilizada para la regresión

'learning_rate': Tasa de aprendizaje, que controla la contribución de cada árbol al modelo final.

'max_depth': Profundidad máxima de cada árbol

'min_child_weight': Peso mínimo necesario para crear un nuevo nodo en el árbol.

'silent': Si se imprimirán o no mensajes durante el entrenamiento

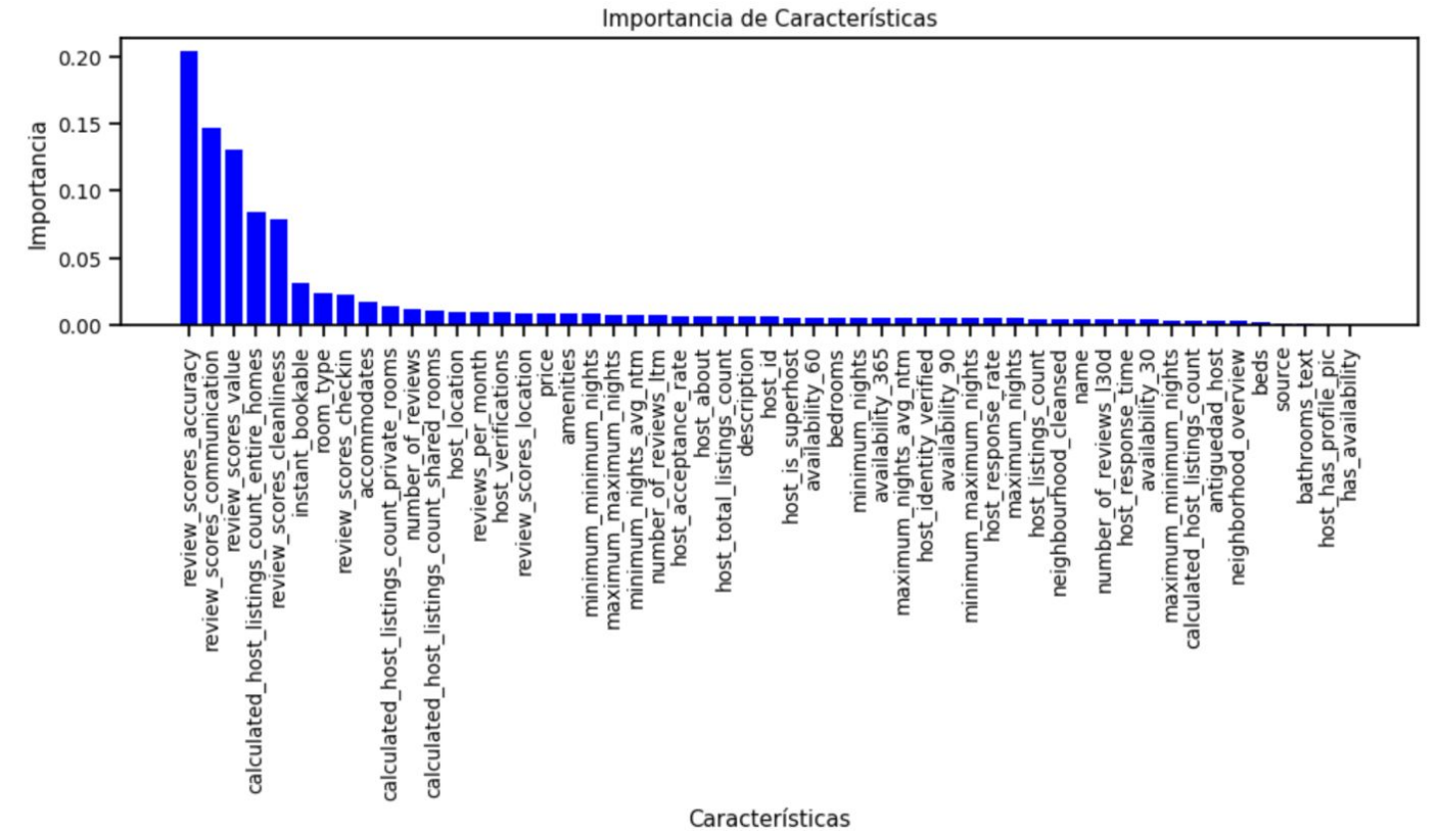
'subsample': Fracción de muestras utilizadas para entrenar cada árbol

'colsample_bytree': Fracción de columnas utilizadas para entrenar cada árbol

'n_estimators': Número de árboles (estimadores) en el modelo

Modelo ganador - Modelo XGBoost

```
model = xgb.XGBRegressor(colsample_bytree=0.8, learning_rate=0.05, max_depth=2,
                         min_child_weight=9, n_estimators=200, nthread=4, objective='reg:squarederror', silent=1, subsample=0.8)
```



Elimino features que tienen 0 de importancia y vuelvo a entrenar el modelo

Modelo ganador - Modelo XGBoost

```
[ ]  model = xgb.XGBRegressor(colsample_bytree=0.8, learning_rate=0.05, max_depth=2, min_child_weight=9,
                              n_estimators=200, nthread=4, objective='reg:squarederror', silent=1, subsample=0.8)
     model.fit(X_train_filtered, y)
     y_pred = model.predict(X_test_filtered)
```

Variables filtered son las variables sin los features que tienen 0 importancia

El modelo XGBoost, utilizando GridSearch, Cross Validation y Feature Importance fue el modelo con el que mejor score obtuve

ITBA
Instituto Tecnológicc
de Buenos Aires

Muchas gracias