



# AirBed&Breakfast



Evaluador automático de precios  
Machine learning: Modelos supervisados

# Team



**Raúl Arellano**

Data Scientist



**María Sánchez Yuste**

Data Scientist



**Sofía García-Baquero**

Data Scientist

# Introducción

Airbnb es una compañía que ofrece una plataforma digital dedicada a la oferta de alojamientos a particulares y turísticos

Es importante distinguir las diferencias entre los alojamientos arrendados a **corto plazo** y a **largo plazo**

Se analizará la predicción de precios de Airbnb en **Canada**



# Preprocesamiento de datos



## Análisis

A través de Pandas Profiling

## Encoding

- OneHot Encoder
- TargetEncoder
- KNN Imputer

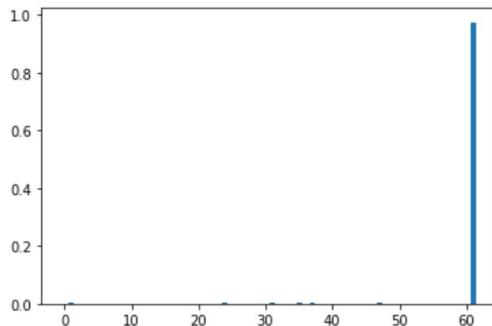
## PCA

Para reducción de la dimensión

A su vez, hemos tratado  
valores missings y outliers

# Random Forest inicial

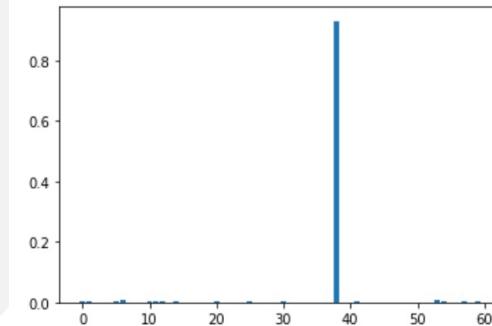
Alojamientos de corta duración:



	param_max_depth	param_n_estimators	mean_test_score	std_test_score	mean_train_score	std_train_score
0	6	200	0.954283	0.022997	0.986547	0.001259

- Hemos probado cual es el óptimo de árboles que tiene el Random Forest en el rango de 75 y 1500 y hemos comprobado que el número óptimo para este modelo es de 200 con un cross validation.
- Asimismo, hemos probado los árboles en su profundidad en un rango de 2 a 100 alturas, siendo el óptimo de 6.

Alojamientos de larga duración:



	param_max_depth	param_n_estimators	mean_test_score	std_test_score	mean_train_score	std_train_score
0	8	70	0.84504	0.176793	0.176793	0.017776

- Hemos probado cual es el óptimo de árboles que tiene el Random Forest en el rango de 75 y 1500 y hemos comprobado que el número óptimo para este modelo es de 200 con un cross validation.
- Asimismo, hemos probado los árboles en su profundidad en un rango de 2 a 100 alturas, siendo el óptimo de 6.



# Random Forest final

Alojamientos de corta duración:

	param_max_depth	param_n_estimators	mean_test_score	std_test_score	mean_train_score	std_train_score
0	8	250	0.959579	0.019691	0.992211	0.000876

- Hemos probado cual es el óptimo de árboles que tiene el Random Forest final en el rango de 75 y 1000 y hemos comprobado que el número óptimo para este modelo es de 250 con un cross validation.
- Asimismo, hemos probado los árboles en su profundidad en un rango de 2 a 40 alturas, siendo el óptimo de 8.

Alojamientos de larga duración:

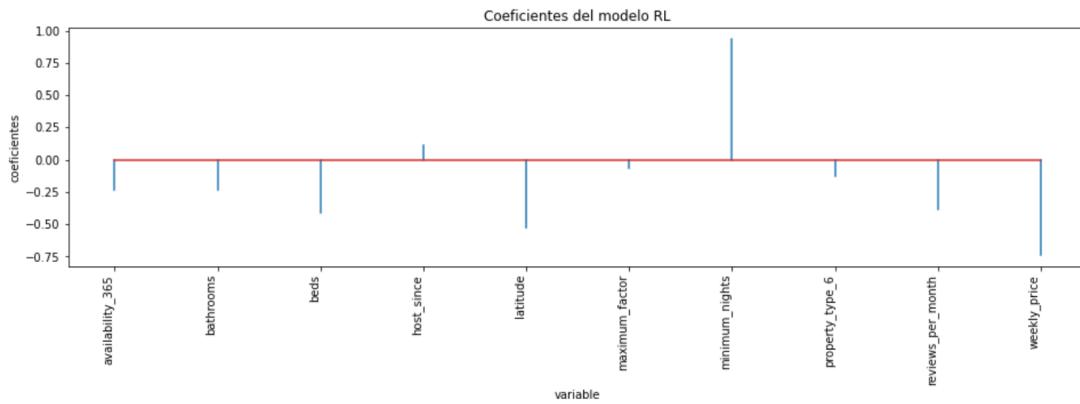
	param_max_depth	param_n_estimators	mean_test_score	std_test_score	mean_train_score	std_train_score
0	6	80	0.863525	0.164408	0.960188	0.014277

- Hemos probado cual es el óptimo de árboles que tiene el Random Forest final en el rango de 75 y 1000 y hemos comprobado que el número óptimo para este modelo es de 80 con un cross validation.
- Asimismo, hemos probado los árboles en su profundidad en un rango de 2 a 40 alturas, siendo el óptimo de 6.

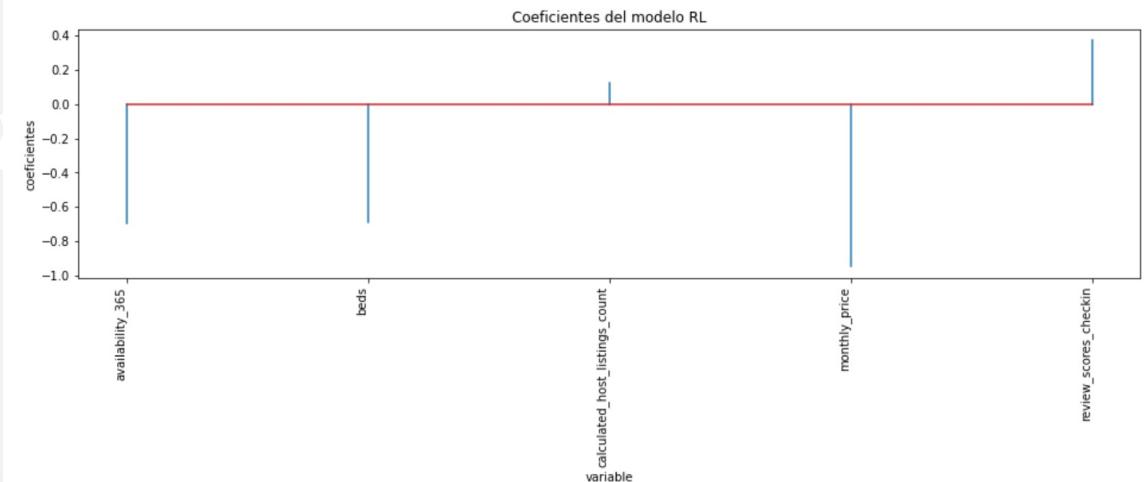


# Regresión Logística

Alojamientos de corta duración:



Alojamientos de larga duración:



Para hacer la regresión logística de los alojamientos hemos seleccionado los datos más relevantes que hemos hallado anteriormente con el Random Forest, y hemos visto la importancia, -coeficiente- que tienen estas variables para la regresión.



# XGBoost

Alojamientos de corta duración:

El modelo de XGBoost es muy similar al Random Forest ya que utiliza árboles de decisión, sin embargo se diferencia al poder crecer de manera más especializada en función de los datos.

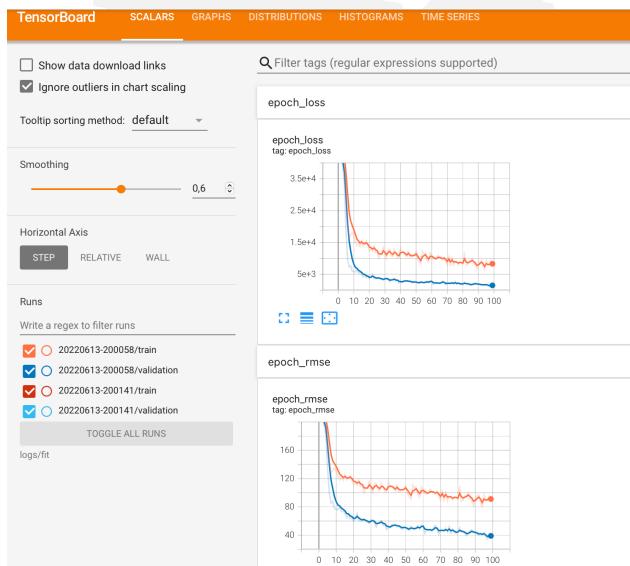
	param_learning_rate	param_max_depth	param_max_features	param_n_estimators	mean_test_score	std_test_score	mean_train_score	std_train_score
0	0.1	2	auto	100	0.952574	0.024714	0.985764	0.002497

Alojamientos de larga duración:

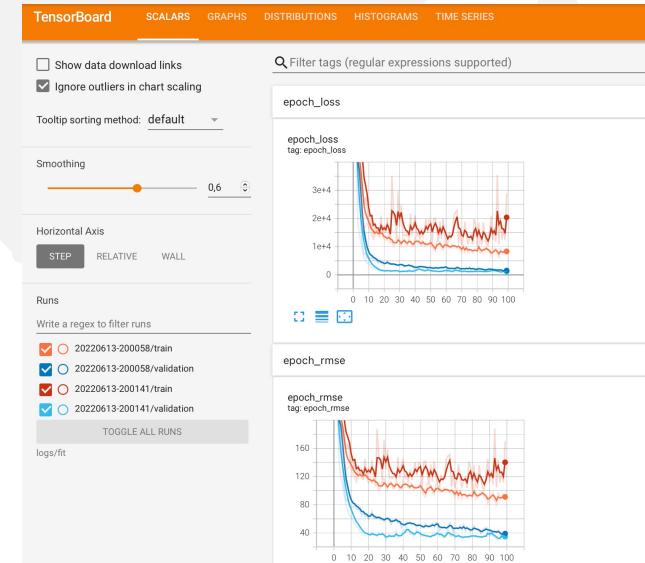
	param_learning_rate	param_max_depth	param_max_features	param_n_estimators	mean_test_score	std_test_score	mean_train_score	std_train_score
0	0.08	4	auto	70	0.817163	0.207026	0.994056	0.002653

# Redes Neuronales

## Alojamientos de corta duración:



## Alojamientos de larga duración:



- Para seleccionar el óptimo de capas nos hemos guiado por la lógica y la teoría aprendida en clase. Sin embargo, para seleccionar la función de activación óptima hemos seguido la publicación **indicada en el notebook**. Nos hemos dado cuenta que de todas las funciones de activación disponibles en la **librería Keras** aquellas basadas en una distribución exponencial son las que mejor se adaptan a la composición de nuestros datos.
- Finalmente nos hemos quedado con Adam Max frente a Adam, que es el que vimos en clase por defecto.



# SME y R<sup>2</sup>



¡Buenos modelos!



En conclusión, se observa que los mejores modelos son los de **Random Forest** y **Redes Neuronales** por su alta precisión. Presentan un bajo error cuadrático medio y un alto R cuadrado.

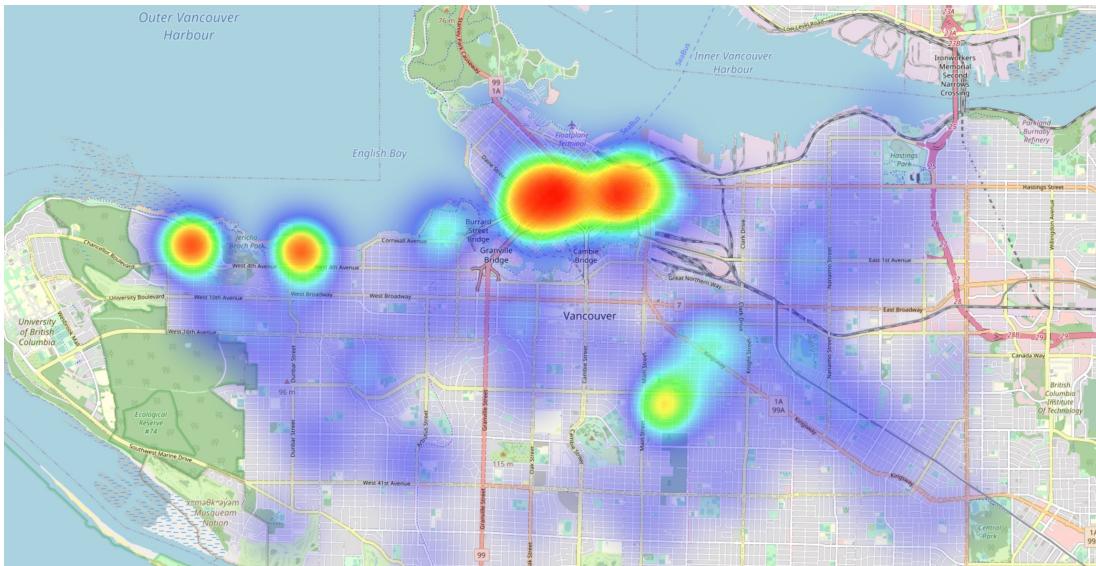
No obstante, vamos a mostrar la tabla que indica todos los modelos con sus errores cuadráticos medios y R-cuadrados correspondientes.

Modelo	RMSE	R2
<b>Random Forest corta duración inicial</b>	21.993893	0.973903
<b>Random Forest larga duración inicial</b>	20.753872	0.986590
<b>Random Forest corta duración final</b>	21.260027	0.975616
<b>Random Forest larga duración final</b>	20.976073	0.986301
<b>Regresión Logística corta duración</b>	62.658731	0.788191
<b>Regresión Logística larga duración</b>	70.223776	0.846464
<b>XGBoost corta duración</b>	21.984043	0.973927
<b>XGBoost larga duración</b>	32.501039	0.846464
<b>SVM poly corta duración</b>	53.190214	0.847368
<b>SVM poly larga duración</b>	141.837419	0.373642
<b>Red Neuronal corta duración</b>	39.092416	0.917555
<b>Red Neuronal larga duración</b>	32.479573	0.967156

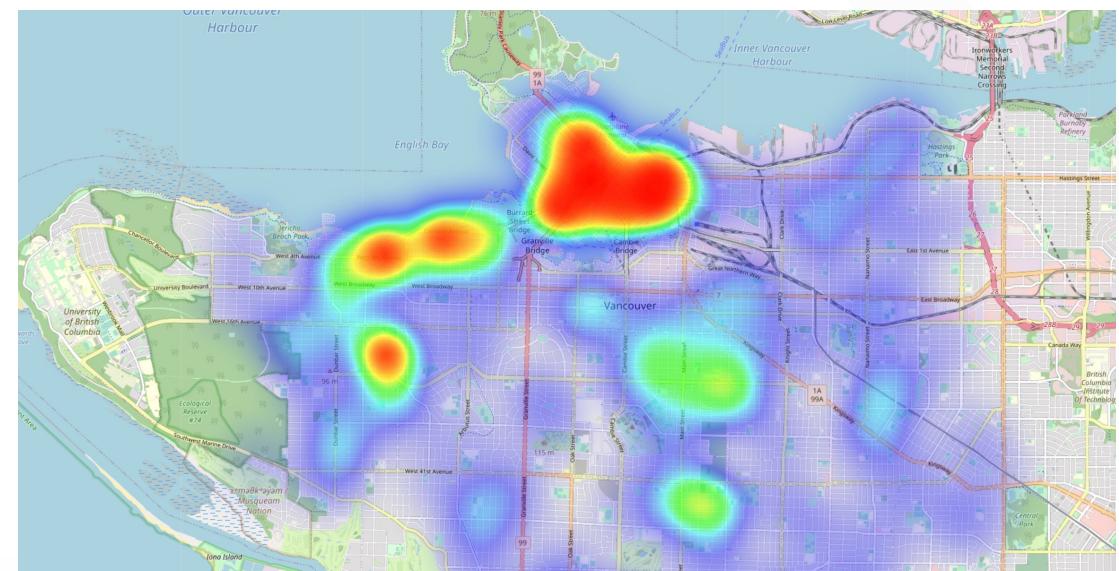


# Ubicación

Corta Duración



Larga Duración



Search By Canada

