Trabajo práctico 3: Modelado lineal

Laboratorio de Datos
Facultad de Ciencias Exactas y Naturales - Universidad de Buenos Aires

Verano 2022

Este trabajo práctico debe entregarse en un `notebook` de **R**. Intercale texto código y gráficos. Asegurese de incorporar a la presentación de código lo que usted aprendió y las conclusiones que obtuvo del análisis. No todas las exploraciones necesitan estar presentes en el notebook final, sólo retenga el contenido que considere necesario. Pienselo como un informe o una historia que narra a alguien interesado en aprender del dataset.

Vamos a trabajar con un subconjunto del dataset de Properati (`datos_alquiler.csv`), que contiene la información de algunas propiedades en alquiler de algunos barrios de CABA. La columna que indica el barrio ha sido removida ya que luego será usada como herramienta de comparación.

El objetivo será analizar los datos teniendo en cuenta un subconjunto de *variables de interés*: el tipo de propiedad, su superficie (cubierta y fondo), cantidad de ambientes, precio, fecha de publicación (`start_date`) y la ubicación (`lat-lon`).

1. Repita los pasos de **EDA** que realizó en el TP anterior con este subconjunto. Observe cómo se distribuyen los valores las variables de interés. Tenga en cuenta que debe setear el tipo de dato correcto para las columnas que representan fechas (*Tip:* `?as.Date`). ¿Es igual de común publicar propiedades todos los días de la semana? ¿Y durante el mes? (*Tip:* `?weekdays`, `format(date_object, format='%d')`)

2. Ajuste un modelo constante para la variable precio, usando `lm`. Grafique la variable precio junto al modelo (*Tip:* `abline(modelo)`). ¿Qué representa esta recta?

3. ¿Cómo se relaciona (cualitativamente) el precio con las dos variables de superficie (`fondo`, `surface_covered`) en este dataset? ¿Y con la fecha de publicación?

4. Ajuste un modelo lineal usando `lm` para describir el precio en función de la superficie cubierta. ¿Cuán bien ajusta el modelo a los datos? ¿Cómo son los residuos? Grafique la recta obtenida sobre en conjunto con el precio y la superficie cubierta. Grafique los valores predichos en función de los valores observados. ¿El desempeño del modelo es igual en todo el rango de valores? Grafique los residuos (en un nuevo gráfico).

5. Ajuste un modelo lineal usando `lm` para visualizar la evolución del precio en el tiempo. Visualice el resultado junto a los datos. ¿Se observa el impacto de la inflación?

6. Ajuste un modelo lineal usando `lm` para describir el precio en función del tipo de propiedad. Visualice el resultado junto un boxplot de los datos. ¿Qué representan los coeficientes del modelo?

7. Explore modelos alternativos, sumando las variables que indican el fondo, tipo de propiedad, ubicación, fecha de publicación y cantidad de habitaciones. Empiece considerando modelos que incluyan 2 variables, y luego explore combinaciones con tres variables. Compare los modelos considerando el error de ajuste y su cantidad de parámetros. Construya una gráfica que muestre el error de ajuste en función de la cantidad de parámetros de cada modelo. ¿Cuál modelo tiene menor error de ajuste?