EXAMEN MINERÍA DE DATOS I

Se pretende estimar el precio de mercado (variable price) de un conjunto de inmuebles para la venta de bienes raíces en Rusia. Para ello, se proporciona un conjunto de datos que consta de 2 ficheros con información de inmuebles. Los ficheros de trabajo han sido obtenidos de la siguiente dirección de Kaggle (https://www.kaggle.com/mrdaniilak/russia-real-estate-20182021) posteriormente han sido tratados para elaborar las especificaciones de la tarea encomendada.

La muestra de entrenamiento (train) y de test contienen 76.480 y 5.873 observaciones respectivamente.

La información contenida en cada dataset es la siguiente:

id: identificador del inmueble

date: fecha de publicación del anuncio

time: la hora en que se publicó el anuncio

geo_lat: Latitud

geo_lon: Longitud

region: - Región de Rusia. Se han considerado 85 áreas en el país.

building_type: tipo de fachada.

0 - Otro

1 - Panel

2 - Monolítico

3 - Ladrillo

4 - Blocky

5 - Madera

object_type: tipo de apartamento

1 - Mercado secundario de bienes raíces

2 - Edificio nuevo

Level: Piso del apartamento

Levels: Número de plantas del edificio

rooms: el número de habitaciones

Si el valor es "-1", significa "apartamento tipo estudio".

area: el área total del apartamento

kitchen_area: Área de cocina

price: Precio en rublos -> cambio a euros (1 euro equivale a 90,06 rublo ruso)

Tareas encomendadas

Preprocesado de la información:

Dada la transcendencia de esta primera fase del análisis y de la cantidad de tiempo y recursos que conlleva (entre el 70 y el 80 por ciento del trabajo real) se considera muy importante realizar a fondo esta tarea, antes de aplicar los modelos.

- Resúmenes de la información a través de tablas y análisis gráficos.
- Análisis de correlación.
- Análisis e imputación de valores faltantes y valores extremos, si los hubiera.
- Selección de variables.
- Otros tratamientos que se consideren oportunos antes de aplicar los modelos.

Aplicación de Modelos:

Especifique y justifique cómo utiliza la muestra para la estimación de los modelos.

Métodos básicos para la REGRESIÓN:

- Árboles de decisión. (CART y Random Forest)
- Métodos de vecindad (K vecinos)
- Redes Neuronales. (Perceptron Multicapa)
- Máquinas de vectores soporte.
- Multiclasificadores: Bagging, Boosting…

Se valorará que se incluyan otros modelos y algoritmos que se consideren oportunos. Por ejemplo: Regresión Lineal, Regresión Lasso, y/o modelos de Gradient Boosting, Redes neuronales de Base Radial, etcétera.

Anexo

Definición de la muestra de desarrollo:

A la hora de definir la muestra de desarrollo del modelo se considera relevante evaluar la posible existencia de outliers en función de:

- Incoherencias: todo inmueble debe tener una cocina de mínimo 2m2)
- Estimación dimensiones de la casa siguiendo:
    - Piso:
        - Habitación (sin contar salón): 9m2
        - Salón: 12m2 (todas las habitaciones con salón, salvo los estudios)
        - Baño y/o terraza: 6m2
        - +10% por distribución de espacios
        - Umbral de 10m2 en función del área del inmueble
    - Estudio:
        - Habitación: 9 m2
        - Baño y/o terraza: 6 m2
        - + 5% por distribución de espacios
        - Umbral de 5 m^2 en función del área del inmueble
- Análisis del precio del m^2: a evaluar por alumnado en base a las técnicas clásicas (boxplot, histograma, etc.). Ayuda:
    - Precio medio del m^2 en Rusia se encuentra entre 63.000 – 90.000 rublos

La duración de publicación del inmueble puede ser relevante en la estimación del precio.

Por ello, se deja al alumnado su posible inclusión. Tomar como fecha de referencia del ejercicio, el día 31 de mayo de 2021.