

Trabajo Práctico 1

Desafío 1. Análisis exploratorio de un dataset de precios de propiedades

Introducción

La inmobiliaria [Properati](#) publica periódicamente información sobre ofertas de propiedades para venta y alquiler. Ud. deberá asesorar a la inmobiliaria a desarrollar un modelo de regresión que permita predecir el precio por metro cuadrado de una propiedad. El objetivo final es que el modelo que desarrollen sea utilizado como tasador automático a ser aplicado a las próximas propiedades que sean comercializadas por la empresa. Para ello la empresa le provee de un dataset correspondiente al primer semestre de 2017.

El dataset es de tamaño entre pequeño y mediano, pero tiene dos complejidades a las que deberá prestarle atención:

- Peso de missing data en algunas variables relevantes.
- Será importante tener en cuenta el problema de la influencia espacial en los precios por metro cuadrado. En efecto, es probable que existan diferencias importantes en las diferentes geografías, barrios y zonas analizadas.

Objetivos:

- Efectuar una limpieza del dataset provisto. Particularmente, deberá diseñar estrategias para lidiar con los datos perdidos en ciertas variables.
- Realizar un análisis descriptivo de las principales variables.
- Crear nuevas columnas a partir de las características dadas que puedan tener valor predictivo.

Datos

Descargar el dataset desde Playground

Requisitos

Los materiales deberán ser entregados en una Jupyter notebook que satisfaga los requerimientos del proyecto. La notebook deberá estar debidamente comentada. Además, los grupos deberán crear un repositorio para el proyecto (anonimizado) en Github. Para la presentación deben armar algunos slides no técnicos y grabar un video que exponga los resultados obtenidos en no más de 10 minutos.

Material a entregar

Presentar un informe estructurado (PPT o Google Slides). Éste debe constar en una introducción (planteo del problema, la pregunta, la descripción del dataset, etc.), un desarrollo de los análisis realizados (análisis descriptivo, análisis de correlaciones preliminares, visualizaciones preliminares) y una exposición de los principales resultados y conclusiones.

Fecha de entrega

- El material deberá entregarse en la **clase 18 del curso**.

Dataset

El dataset contiene información sobre todas las propiedades georeferenciadas de la base de datos de la empresa. La información de cada propiedad que incluye es la siguiente:

- Fecha de creación
- Tipo de la propiedad (house, apartment, ph)
- Operación del aviso (sell, rent)
- Nombre del lugar
- Nombre del lugar + nombre de sus 'padres'
- ID de geonames del lugar (si está disponible)
- Latitud, Longitud
- Precio original del aviso
- Moneda original del aviso (ARS, USD)
- Precio del aviso en moneda local (ARS)
- Precio aproximado en USD
- Superficie en m²
- Superficie cubierta en m²
- Precio en USD/m²
- Precio por m²
- N° de piso, si corresponde

- Ambientes
- URL en Properati
- Descripción
- Título
- URL de un thumbnail de la primer foto

¿Cómo empezar? Sugerencias

Agreguen toda información adicional construida a partir de los datos originales (o incluso información externa) que consideren relevante y útil para resolver los objetivos planteados.

Aprovechen las herramientas de Pandas: *groupby*, *summation*, *pivot_tables* y otras aplicaciones y métodos de los DataFrames hacen mucho más simples los cálculos y otras agregaciones de los datos.

En la presentación de los resultados tengan en cuenta que es altamente probable que la audiencia no tenga un nivel técnico así que mantengan el lenguaje en un nivel accesible.

En términos generales, recuerden las siguientes sugerencias:

- Escribir un pseudocódigo antes de empezar a escribir código. Suele ser muy útil para darle un esquema y una lógica generales al análisis.
- Leer la documentación de cualquier tecnología o herramienta de análisis que usen. A veces no hay tutoriales para todo y los documentos y las ayudas son fundamentales para entender el funcionamiento de las herramientas utilizadas.
- Documentar todos los pasos, transformaciones, comandos y análisis que realicen.

Recursos útiles

- [Github](#)
- [Documentación de la librería GeoPandas](#)
- [Cheatsheet de pandas](#)
- [Pandas user guide](#)

Evaluación

/10	Discusión del dataset: análisis exploratorio de los datos crudos
/10	Discusión del método: Creación de variables nuevas con valor predictivo

	Justificación de la elección de variables nuevas con valor predictivo
/10	Discusión del método: Detección de campos a imputar Justificación de la selección de campos a imputar
/10	Discusión del método: Estrategia para obtener un dataset de registros completos y con variables útiles para predecir el valor del m2 de las propiedades Justificación de la estrategia propuesta para obtener el dataset resultado
/10	Discusión de los resultados: análisis exploratorio del dataset resultado
/10	Adecuada profundidad del trabajo
/10	Adecuada presentación de conclusiones
/10	Formato del informe (introducción/cuerpo/conclusión, gráficos)
/10	Presentación del informe (estructura del texto, storytelling, figuras)
/10	Presentación del video (tiempo adecuado, claridad en la exposición, información relevante para la audiencia definida)
/100	Nota final