Desafío 2. Prediciendo precios de propiedades

Introducción

Esta semana comenzamos a pensar en términos de modelos de forma más explícita. Empezamos con modelos de regresión lineal y su implementación en scikit-learn. También trabajamos sobre la forma de "traducir" los objetivos de negocios en un modelo. A su vez, hemos introducido formas de validación de un modelo, particularmente, utilizamos cross-validation para optimizar modelos. Ahora vamos a aplicar estos nuevos contenidos al dataset de Properati que limpiaron en el desafío anterior.

Objetivos:

- Estimar un modelo de regresión lineal que realice predicciones para el **precio por metro cuadrado**. Deberá prestar cierta atención a la **estructura espacial** de los precios. Se sugiere acotar el espacio a CABA.
- Aplicar regularización a modelos lineales (pueden hacerlo para obtener un puntaje adicional). La idea es la siguiente: estimar una **regresión Ridge y una LASSO** sobre el dataset. Para ello, deberán usar cross-validation para tunear el parámetro de regularización que maximiza R2 en tu test set. ¿Cómo son las performances entre los modelos regularizados y no regularizado? ¿Cuál funciona mejor? ¿Qué hace una regresión Ridge? ¿Y una LASSO? ¿Qué diferencias hay con la regresión lineal sin regularizar?
- Seleccionar mediante muestreo aleatorio simple una submuestra de 100 propiedades. Este será su portafolio de departamentos. En base al mejor modelo que haya encontrado determine cuáles de los departamentos, tanto en su portafolio como fuera de él, se encuentran sobrevaluados o subvaluados y en qué magnitud.
- Teniendo en cuenta que podría comprar y vender propiedades al precio de mercado, omitamos costos de transacción, con un capital inicial igual al valor de mercado de las propiedades en su portafolio. ¿Cuál es el mejor portafolio de propiedades que podría comprar?

Requisitos y material a entregar

Los modelos 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 en clase se deben armar algunos slides no técnicos para exponer en no más de 10 minutos (PPT o Google Slides). La presentación debe constar de una introducción (planteo del problema, la pregunta, la descripción del dataset, etc.), un desarrollo de los análisis realizados (análisis descriptivo, análisis de correlaciones preliminares, resultados de los distintos modelos, visualizaciones) y una exposición de los principales resultados y conclusiones.

Fecha de entrega

- El material deberá entregarse en la clase del día **Lunes 9 de Junio**

¿Cómo empezar? Sugerencias

Dado que usaremos modelos lineales, el ajuste puede ser menor a las expectativas o los modelos pueden no funcionar perfectamente. No se desanimen. Vamos a aprender más adelante técnicas que van a mejorar nuestras capacidades de predicción y de análisis. Por ahora, hagan lo mejor que puedan con las herramientas disponibles.

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 codear. 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

- Documentación de la librería Scikit-Learn
- ¿Qué es regularización?
- Valuador oficial de Properati