Desafío 3. 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 a un dataset que tiene cierta complejidad. 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 a lo que va del año 2017.

El dataset es de entre mediano y pequeño pero tiene dos complejidades a la 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 de 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.
- Estimar un modelo de regresión lineal que realice predicciones para el precio por metro cuadrado
- Usar cross-validation para validar el modelo. Deberá prestar cierta atención a la estructura espacial de los precios.
- *Bonus*: Aquellos osados que sientan que pueden entender y 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 el modelo lineal sin restricciones?

Requisitos

Los materiales deberán ser entregados en un Notebook Jupyter que satisfaga los requerimientos del proyecto. El notebook deberá estar debidamente comentado:

Material a entregar

Una notebook con el código que genera los estadísticos y los gráficos debidamente comentados. El código básico y una guía de pasos fueron diseñados en formato de notebook jupyter. Pueden usar éste notebook como guía pero presentar el código generado, los análisis y modelos realizados, junto con los principales resultados en un informe estructurado (también en formato de jupyter notebook). El mismo 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, modelos estimados) y una exposición de los principales resultados y conclusiones.

Fecha de entrega

- El material deberá entregarse en la **clase 28** del curso. (Fecha: 27/09 o 28/09 según la fecha de cursada correspondiente).

Dataset

Este dataset contiene información sobre todas las propiedades georeferenciadas de la base de datos de la empresa. La información de cada propiedad que incluye el dataset 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 otra información construida a partir de los datos originales (o incluso información externa) que consideren relevante y útil para resolver los objetivos planteados.

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 ejemplo, en la predicción de clases de locales). Por ahora, hagan lo mejor que puedan con las herramientas disponibles.

Aprovechen las herramientas de pandas: *gropuby*, *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 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 uses. 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 realices.

Recursos útiles

- Documentación de la librería SKLearn
- ¿Qué es regularización?

Evaluación

Los profesores usarán la siguiente escala para calificar tu trabajo y las habilidades técnicas adquiridas en módulo:

Puntaje	Descripción
0	Incompleto
1	No cumple con las expectativas
2	Cumple con las expectativas. Buen trabajo!
3	Excede con creces las expectativas!