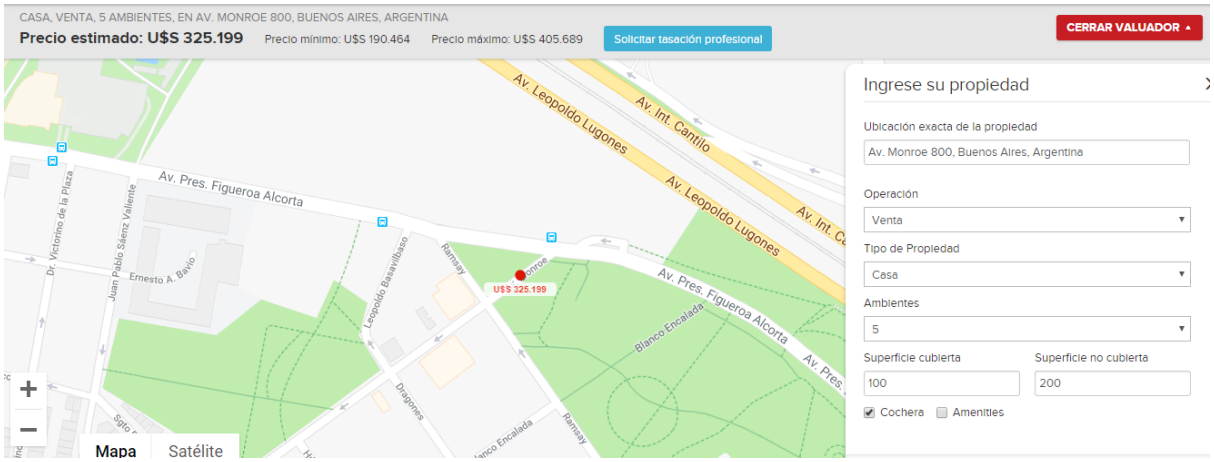
Desafío 2. Prediciendo precios de propiedades

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 validar modelos. Ahora vamos a aplicar estos nuevos contenidos al dataset de Properati que limpiaron en el desafío anterior. Properati cuenta con una herramienta de tasación en su sitio web, trataremos de construir un modelo para la misma tarea.



Objetivos:

- Estimar un modelo de regresión lineal para hacer predicciones para el precio de las propiedades. ¿Modelará el precio total o el precio por metro cuadrado?
- Usar cross-validation para validar el modelo.
- Quizá le sea útil prestar cierta atención a la estructura espacial de los precios a la hora de construir sus features.
- Aplicar regularización a modelos lineales. La idea es la siguiente: estimar una regresión Ridge, una Lasso y una Elastic Net sobre el dataset. Para ello deberán usar cross-validation para tunear los hiper parámetros de regularización.
- ¿Cómo son las performances entre los modelos regularizados y no regularizado? ¿Cuál funciona mejor? ¿Qué "hace" una regresión Ridge? ¿Y una Lasso? ¿Qué hace Elastic Net? ¿Qué diferencias hay con la regresión lineal sin regularizar?
- Seleccionar mediante muestreo aleatorio simple una submuestra de 100 propiedades. Este será su portafolio de propiedades. En base al mejor modelo que haya encontrado determine cuáles de las propiedades, tanto en su portafolio como fuera de él, se encuentran sobrevaluados o subvaluados. ¿Cómo compararía el grado de subvaluación o sobrevaluación entre dos propiedades con precios distintos?

- Teniendo en cuenta que podría comprar y vender propiedades al precio de mercado,asumamos que no hay costos de transacción ni de comprar ni de vender con un capital inicial igual al valor de mercado de las propiedades en su portafolio. ¿Cuál propiedades vendería y cuáles compraría? Piense que busca arbitrar entre estas propiedades “caras” y “baratas”.

Requisitos

Los materiales deberán ser entregados en un Notebook Jupyter que satisfaga los requerimientos del proyecto. El notebook deberá estar debidamente comentado. 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 una presentación en no más de 10 minutos.

Material a entregar

Un notebook con el código que genera los estadísticos y los gráficos debidamente comentado. El código básico y una guía de pasos fue diseñado en formato de notebook Jupyter. Pueden usar éste notebook como guía pero presentar los análisis y modelos realizados, junto con los principales resultados en un informe estructurado (ppt o google slides). 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 el jueves 11 de octubre.

¿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 ejemplo, en la predicción de clases de locales). 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 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.
- No es una competencia de performance entre grupos, le sugerimos no invertir tiempo en mejorar la limpieza de datos realizada en el desafío anterior sino en encontrar el mejor modelo dados esos datos.

Recursos útiles

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