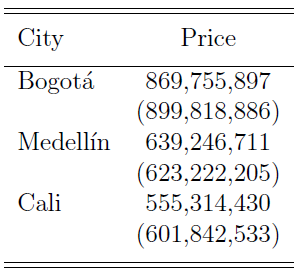
**Problem set 3: ¿Ganar dinero con ML?**

**¡Todo se trata de la ubicación, la ubicación, la ubicación!**

**1. Introducción**

Una nueva empresa dedicada a la compraventa de propiedades acaba de contratarte a ti y a tu equipo para desarrollar un modelo predictivo. Su objetivo es comprar la mayor cantidad de propiedades en Cali, Colombia, gastando lo menos posible. Sin embargo, la empresa no tiene mucha información sobre el mercado inmobiliario de Cali. La empresa tiene una muestra de datos de propiedades individuales en Bogotá y Medellín provenientes de https://www.properati.com.co y datos de precios agregados que se muestran en la Tabla 1:



Nota: La tabla muestra los precios promedio y las desviaciones estándar entre paréntesis.

Los datos, disponibles en el archivo dataPS3.zip, contienen información sobre los precios de cotización, así como las características de las propiedades. El archivo comprimido incluye otros tres archivos: datos de entrenamiento, datos de prueba y una plantilla de envío. La empresa quiere evitar el asco de Zillow.

Zillow desarrolló algoritmos para comprar casas. Sin embargo, sus modelos sobreestimaron considerablemente el precio de las viviendas. Esta sobreestimación significó pérdidas de alrededor de USD 500 millones para la empresa y una reducción aproximada del 25 % de su plantilla. Para evitar el asco, tu equipo competirá contra todos los demás grupos, y aquellos con el presupuesto total más bajo obtendrán el contrato y una bonificación.

**1.1 Instrucciones generales**

El objetivo principal es construir un modelo predictivo de precios de venta. Del artículo histórico de Rosen "Precios hedónicos y mercados implícitos: diferenciación de productos en competencia pura" (1974), sabemos que un vector de sus características, C = (c1; c2; : : : ; cn), describe un bien diferenciado. En el caso de una casa, estas características pueden incluir atributos estructurales (p. ej., número de dormitorios), servicios públicos del vecindario (p. ej., calidad de la escuela local) y servicios ambientales locales (p. ej., calidad del aire). Por lo tanto, podemos escribir el precio de mercado de la casa como:



Sin embargo, la teoría de Rosen no nos dice mucho sobre la forma funcional de f. En este conjunto de problemas, explorará diferentes modelos para producir la mejor predicción posible.

Hay dos resultados esperados:

1. Un documento .pdf.

2. Y un archivo .csv con predicciones.

El documento debe contener las siguientes secciones:

• Introducción. El resumen de introducción establece el problema y es una oportunidad para “vender” su modelo predictivo, mostrando las ventajas/desventajas de su elegido modelo y rendimiento esperado.

• Datos2. En este conjunto de problemas, debe agregar expandir las variables en sus datos (recuerde expandir los datos de entrenamiento y prueba), como mínimo debe agregar siete variables adicionales:

- Al menos 4 predictores provenientes de fuentes externas; estos pueden ser de mapas de calle abierta.

- Al menos 3 predictores provenientes del título o descripción de los inmuebles.

Trate esta sección como una oportunidad para presentar una narrativa convincente para justificar o defender sus opciones de datos y ayudar al lector a comprender sus datos y su variación. Descríbalo en consecuencia con estadísticas descriptivas, gráficos, etc. Como mínimo, debe incluir:

- Una tabla con estadísticas descriptivas

- Dos mapas, puedes elegir qué información mostrar

• Modelo y Resultados. Cuando presente su modelo predictivo, incluya:

- Una explicación de las variables utilizadas para entrenar este modelo; recuerda usar las variables que agregaste en la sección anterior.

- Una explicación detallada de cómo se entrenó, la selección de hiper parámetros y cualquier otra información relevante sobre el modelo.

- Una discusión de su medida de evaluación.

• Conclusiones y Recomendaciones. En esta sección, expone brevemente las principales conclusiones de su trabajo.

**2. Lineamientos adicionales**

Espero lo siguiente del conjunto de problemas, la omisión de cualquiera de estas pautas será penalizada.

• Convierte un documento .pdf en Bloque Neón. El documento no debe tener más de 6 (seis) páginas e incluir como máximo 6 (seis) anexos (tablas y/o figuras). La bibliografía y las exhibiciones no cuentan para el límite de páginas. Puede agregar un apéndice, pero el documento principal debe ser independiente. Específicamente, un lector debe poder seguir el análisis en el documento y estar convencido de que es correcto y coherente solo con el texto principal, sin consultar el apéndice.

• Convierte un archivo .csv en Bloque Neon. Un ejemplo de cómo debería verse el archivo de envío está en la carpeta de datos: submit\_template.csv. Este archivo incluye dos columnas, una con la variable que identifica una propiedad y otra con su precio previsto. No cambie el nombre de las columnas.

• Juzgaré las predicciones basándome en aquellos que gastan menos dinero y pueden comprar la mayor cantidad de propiedades. Es decir, si sobrevaloras una propiedad que se suma a tu ficha, si subestimas el precio, ahorras. Sin embargo, si el precio previsto está subvaluado en más de 40 millones. COP, será sancionado. En este caso, la venta no se produciría, es decir, no podrá comprar la propiedad.

• Siga la siguiente convención para su nombre de archivo .csv. El nombre del archivo debe incluir las predicciones de nombre, seguidas de los apellidos de tus compañeros, todo separado por guiones bajos, por ejemplo, predicciones\_gomez\_matinez\_sarmiento.csv

• Asignaré puntos de bonificación según las clasificaciones relativas.

• Las tablas, figuras y escritos deben ser lo más prolijos posible. Etiquete y describa todas las variables, estadísticas, etc., incluidas en sus figuras y tablas.

• El documento debe apuntar e incluir un enlace a su repositorio de GitHub.

• Siga la plantilla del repositorio.

• El repositorio debe incluir un archivo README. Un buen README ayuda a que su proyecto se destaque de otros proyectos y es el primer archivo que una persona ve cuando se encuentra con su repositorio. Por lo tanto, este archivo debe ser lo suficientemente detallado para enfocarse en su proyecto y cómo lo hace, pero no tanto como para que pierda la atención del lector. Por ejemplo, Project Awesome tiene una lista seleccionada de archivos README interesantes.

• El repositorio debe tener al menos cinco (5) contribuciones sustanciales de cada miembro del equipo.

• El código debe ser completamente reproducible. Legible e incluir comentarios, un buen estilo de codificación es crítico. Te animo a que sigas la guía de estilo de tidyverse.