Conjunto de problemas 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 el barrio de Chapinero en Bogotá, Colombia, gastando lo menos posible.

La empresa tiene una muestra de datos de propiedades individuales en Bogotá de https://www.propertyati.com.co. Sin embargo, falta información sobre las propiedades en Chapinero.

La empresa quiere evitar el fiasco de Zillow.1 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.

Hay dos resultados esperados:

1. Un documento .pdf.

2. Envíos con los pronósticos de tu equipo en Kaggle en el siguiente enlace.

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 locales (p. ej., delincuencia, calidad del aire, etc.). Por lo tanto, podemos escribir el precio de mercado de la casa como:

$$Pi = f(ci1, ci2, . . . , cin)$$

1Para obtener más información, consulte el siguiente artículo aquí.

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

El documento debe contener las siguientes secciones:

- Introducción. En la introducción se expone brevemente el problema y si existen antecedentes. Describe brevemente los datos y su idoneidad para abordar la pregunta del conjunto de problemas. Contiene una vista previa de los resultados y las conclusiones principales.

- 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 seis variables adicionales:

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

  – Al menos 2 predictores provenientes del título o descripción de los inmuebles.

  Al redactar esta sección, debe:

  1. Describa los datos, su idoneidad para el problema y el proceso de construcción de la muestra, incluido cómo se limpiaron y combinaron los datos y cómo se crearon nuevas variables.

  2. Incluir un análisis descriptivo de los datos. Como mínimo, debe incluir una tabla de estadísticas descriptivas y dos mapas con su interpretación. Sin embargo, espero un análisis profundo que ayude al lector a comprender los datos, su variación y la justificación de sus elecciones de datos. Utilice su conocimiento profesional para agregar valor a esta sección. No la presente como una lista “seca” de ingredientes.

- Modelo y Resultados. En esta sección se presentan los modelos presentados para evaluación. Al redactar esta sección, incluya:

  – Una explicación de las variables usadas para entrenar este modelo, recuerda usar el variables que agregó en la sección anterior.

  – Una explicación detallada de cómo se entrenó, la selección de hiperparámetros y cualquier otra información relevante.

  – Una comparación con al menos otras 4 especificaciones enviadas a Kaggle.

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

2Esta sección se encuentra aquí para que el lector pueda entender su trabajo, pero probablemente debería ser la última sección que escriba. ¿Por qué? Porque vas a hacer elecciones de datos en los modelos estimados. Y todas las variables incluidas en estos modelos deben describirse aquí.

2 Directrices adicionales

- Las predicciones deben enviarse en Kaggle. Consulte el sitio web de la competencia para obtener más información. información.

- Convierte un documento .pdf en Bloque Ne´on. El documento no debe tener más de 8 (ocho) páginas e incluir, como máximo, 8 (ocho) 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.

- El documento debe incluir un enlace a su repositorio de GitHub.
  - El repositorio debe seguir la plantilla.
  - El LÉAME debería ayudar al lector a navegar por su repositorio. 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, Proyecto Impresionante tiene una lista seleccionada de archivos README interesantes.
  - Incluya instrucciones breves para replicar completamente el trabajo.
  - La rama del repositorio principal debe mostrar al menos cinco (5) contribuciones sustanciales ciones de cada miembro del equipo.
  - El código tiene que ser:

    Totalmente reproducible.

    Legible e incluir comentarios. En la codificación, como en la escritura, un buen estilo de codificación es fundamental. Te animo a que sigas la guía de estilo de tidyverse.

- Las tablas, figuras y escritos deben ser lo más prolijos posible. Etiquete todas las variables incluidas. Si tiene algo en sus figuras o tablas, espero que se aborden en el texto. Las tablas deben seguir el formato AER.