

Prueba de conocimiento analítico

DiCAGI 2020

Gracias por su interés en participar en esta convocatoria.

Se buscan personas comprometidas, analíticas, con buena actitud de servicio y que se destaquen por realizar un trabajo oportuno y de calidad... además, personas con unos sólidos conocimientos en técnicas de modelación, estadística, matemáticas, minería de datos y bases de datos y como si fuera poco, capaces de definir adecuadamente un problema, generar las hipótesis, asumir supuestos, obtener conclusiones poderosas y comunicar los resultados de los modelos de una forma práctica y amigable para el usuario final, de tal forma que se asegure que los resultados de los modelos se utilicen en la estrategia del banco.

El propósito de la prueba es medir sus capacidades de análisis y desarrollo de modelos analíticos. La idea es que no le dedique más de 15 horas en total, incluyendo el tiempo para documentar lo que hizo.

Es posible usar cualquier herramienta que quiera (R, Python, SAS Guide/Miner, SPSS, etc.), y cualquier recurso del internet, pero no se permite consultar directamente con otras personas por ningún medio.

Puede realizar los supuestos que considere necesarios. No es necesario utilizar todos los conjuntos de datos o todas las variables. Esto depende de la forma en que usted aborde el problema. No hay una solución única. Inclusive puede darse el caso en que no se tenga un modelo viable.

1. Introducción

La estimación de los precios de la vivienda es esencial tanto para los propietarios como para los inversores, ya que ambos necesitan entender el valor de este activo inmobiliario. Para muchas personas, comprar una propiedad es una de las decisiones y compra más importantes en la vida.

Además de la asequibilidad de una vivienda, otros factores, como la conveniencia del lugar y las perspectivas de inversión a largo plazo, también afectan el proceso de toma de decisiones.

El mercado inmobiliario está expuesto a muchas fluctuaciones en los precios debido a las correlaciones existentes con muchas variables, algunas de las cuales no se pueden controlar o incluso pueden ser desconocidas. Los precios de las viviendas pueden aumentar rápidamente (o en algunos casos, también bajan muy rápido).

Algunas aplicaciones para un banco son:

- **Originación:** Establecer el valor comercial del inmueble a financiar utilizado para la aprobación final.
- **Retanqueo:** Actualizar el valor comercial de garantía ya existente para aprobación de nuevos cupos de crédito.
- **Monitoreo portafolio garantías:** Valoración del portafolio de las garantías para cumplimiento normativo y para el análisis de riesgos del colateral.
- **Normalización de cartera:** Evaluar los préstamos existentes, evaluar los acuerdos de refinanciamiento Y daciones en pago.

2. Definición de la población objetivo y variable respuesta

Se busca, inicialmente, estimar el precio total de un inmueble dadas unas características de este. En el conjunto ***entrenamiento_precios_vivienda.xlsx*** existe una columna llamada *valor_total_avaluo*, la cual es una variable numérica, que indica el precio total del inmueble que incluye áreas privadas, publicas, parqueaderos y cuarto útil en caso de tener alguno de estos.

3. Instrucciones importantes

- ***entrenamiento_precios_vivienda.xlsx***: Este conjunto de datos tiene características de los inmuebles y su precio total (incluye áreas privadas, públicas, parqueaderos y cuarto útil en caso de tener).
- ***testeo_precios_vivienda.xlsx***: contiene exactamente las mismas columnas del archivo *entrenamiento_precios_vivienda.xlsx*, exceptuando la columna *valor_total_avaluo* y los valores en uvr.
- ***base_evaluada.csv***: Este archivo contienen los mismos ids del archivo *testeo_precios_vivienda*. Usted debe poner aquí únicamente el id y el valor predicho por su modelo para cada uno de los casos de prueba. **No cambie el nombre de este archivo ni los nombres de las columnas.**
- ***Existe un identificador único por inmueble este es id.*** Todos los inmuebles son únicos, por lo cual no se tiene medido el valor de un mismo inmueble en diferentes puntos del tiempo.
- En el archivo ***metaData.pdf*** hay algunas descripciones de unas pocas variables. Se deja a interpretación la naturaleza de cada campo.
- En ***PuntosInteres.csv*** hallará puntos de interés geográfico con su respectiva longitud y latitud. La tercera columna es la categoría del punto de interés y la cuarta columna podemos encontrar el nombre si se tiene. No es necesario utilizar esta información realizar la prueba, solo si considera útil.

4. Entregables

- Se debe entregar un archivo ***base_evaluada.csv*** con las columnas de *id* y *valor_total_avaluo*. La columna *valor_total_avaluo* **debe** contener un valor real **para todos y cada uno** de los registros. No aceptaremos valores nulos, NaNs, N/A, N/D, vacíos, o mensajes de texto como, por ejemplo: “datos incompletos”. Por favor haga todo lo posible por conservar el formato del archivo (csv separado por comas, no otro carácter; el orden de las columnas; la línea de encabezado etc. El orden de las filas no es crítico).
- También nos debe entregar la implementación de su modelo (archivos de código con comentarios en caso de usar un lenguaje de programación convencional o el archivo de proyecto que incluya documentación, en caso de usar SAS Miner, Azure ML Studio, u otra herramienta parecida).
- Un archivo de texto (en .txt , .doc, .html, .rmd, .md, note books, .pdf) que contenga una descripción del proceso que siguió para generar el modelo (incluyendo exploración, transformaciones de variables, selección de variables, etc.) y luego generar sus estimaciones de precios de vivienda.

Ahora, es posible que llegue a la conclusión de que no es posible desarrollar un buen modelo predictivo a partir de la información proporcionada o dada la calidad de esta. Si este es el caso, queremos evaluar el mejor modelo que pueda producir y también que nos dé una sustentación de esa conclusión.

De manera opcional nos podría hacer saber qué otros datos o atributos añadiría idealmente al conjunto de datos, para un modelo analítico más efectivo. Aquí, tenga en cuenta la factibilidad y el costo de obtener esos datos.

Con esta información y sus propios análisis que utilidad extraería. Especifique para quien y de qué modo y el porqué es útil.

- Con el uso anterior en mente, diseñe un sistema de manera teórica que bosqueje una solución que permita hacer disponible los resultados de su modelo analítico y que sean fácilmente consumibles por servicios externos, páginas web, servicio móvil, etc. No se tiene que desarrollar. Desarrollar una aplicación o sistema de información no da ningún punto extra y no será tomada en cuenta para la calificación total por lo que recomendamos no desarrollarla sino únicamente elaborar el bosquejo y describir la estrategia.

5. Evaluación

La métrica para evaluar el modelo será el **MAPE** entre el valor real para cada inmueble, que sólo nosotros conocemos, y el valor predicho por su modelo y consignado en el archivo ***base_evaluada.csv*** por medio de la columna *valor_total_avaluo*.