PROCESO DE SELECCIÓN DATA SCIENTIST

El propósito de esta prueba es medir sus capacidades para desarrollar modelos predictivos y analizar información no estructurada. Es posible usar cualquier herramienta que quiera (deseable desarrollar la solución en Python), y cualquier recurso del internet, pero no se permite consultar directamente con otras personas por ningún medio.

Modelación Data Estructurada

SITUACIÓN

Aunque las inmobiliarias son un gran aliado para aquellos que quieren arrendar su vivienda para obtener un ingreso extra, estas suelen generar una serie de malestares entre el arrendador y el arrendatario. En el país existen dos categorías que encierran los mecanismos de los cuales se puede tomar mano a la hora de alquilar un inmueble: los tradicionales y los no tradicionales.

En el segmento tradicional se encuentran las inmobiliarias y los agentes inmobiliarios. Por su parte el segmento no tradicional hay alternativas recientes en el mercado que ganan participación día a día, una de estas alternativas es colocación y distribución del inmueble a través de portales web que en la actualidad ayudan a visibilizar las viviendas que están en venta o en arriendo, estos canales funcionan perfecto para aquellos que quieren alquilar su propiedad y que cuentan con el tiempo para administrarla.

Otra alternativa es el mercado de las aseguradoras que brindan protección a los dueños por los incumplimientos en los cánones de arrendamiento o renta, en las cuotas de administración o servicios públicos por parte de los inquilinos, e incluso, algunos cubren la restitución del bien inmueble.

Un cliente del sector asegurador esta consiente de esta nueva tendencia y desea tener los más afinado posible su modelo de Provisión, sin embargo, uno de los insumos de dicho modelo es el valor del canon de arrendamiento y actualmente las estimaciones sobre esta variable están sesgando los resultados finales del modelo de Provisión, por esta razón la operación quiere desarrollar un modelo analítico para mejorar dichas estimaciones y solicita apoyo al área de Inteligencia Artificial & Data Science (AI & DS).

AI & DS propone un modelo que tiene como objetivo predecir el precio de arrendamiento a partir de información cuantitativa y cualitativa del inmueble. Este modelo permitirá determinar el valor del canon de arrendamiento y, por consiguiente, ajustar el modelo de Provisión para que este se aterrice más a la realidad y permita seguir explorando la nueva línea de negocio.

El área de AI & DS ha decidido que se comience con el desarrollo del modelo analítico por sectores, comenzando con uno de los municipios más críticos en el tema de asertividad, el municipio de Medellín.

El conjunto de datos disponible, contiene información sobre el precio del alquiler de 1575 viviendas situadas en Medellín en el año 2020. Además del precio, incluye 9 variables adicionales:

- precio: precio del alquiler.
- metros2: metros cuadrados de la vivienda.
- anyo: año de construcción.
- banyo: si tiene cuarto de baño (1) o no (0).
- calefaccion: si tiene calefacción central (1) o no (0).
- cocina: si la cocina está equipada (1) o no (0).
- Sp: si la calidad del barrio donde está situada la vivienda es superior la media (1) o no (0).
- Sm: si la calidad del barrio donde está situada la vivienda es inferior la media (1) o no (0).
- loc: combinación de Sp y Sm indicando si la calidad del barrio donde está situada la vivienda es inferior (1), igual (2) o superior (3) a la media.

La siguiente es la descripción de lo que debe hacer:

El archivo adjunto bd_train.csv contiene información anteriormente descrita. El objetivo es que usted desarrolle un modelo de inteligencia artificial (o varios) que permita, a partir de los datos del archivo bd_train.csv predecir el precio de alquiler.

El archivo adjunto *bd_test.csv* contiene exactamente las mismas columnas del archivo bd_train.csv, exceptuando la columna *precio*.

Al final del ejercicio Ud. nos debe entregar:

- **Un análisis descriptivo y exploratorio** del archivo train.csv, donde se describa a profundidad el comportamiento detallado de la base de datos.
- **Un análisis de selección de características** (deseable aplicar por lo menos 3 metodologías diferentes para la selección de variables y aplicar reducción de dimensionalidad (extracción de características) antes de entrenar un modelo si lo considera pertinente).
- La implementación de su **modelo predictivo** en un notebook de Jupyter (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).

- Se debe entregar un archivo **bd_test_evaluate.csv** con los **resultados del modelo entrenado** en el punto anterior, el archivo debe contener la predicción del precio y es un valor real positivo, para todos y cada uno de los registros. No se aceptan 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 es crítico).

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

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

CONSIGNA

La métrica con la que desee evaluar su modelo será el RMSE. Adicionalmente se evaluará integralmente el informe entregado: el análisis de las variables, su transformación, proceso de selección, y cualquier componente analítico que haya sido útil para la construcción del modelo.

Modelación Data NO Estructurada

Con el fin de medir sus capacidades usando expresiones regulares, por favor resolver los siguientes ítems:

1) Escriba un programa que lea el siguiente texto escrito link y entregue un diccionario que cuente las ocurrencias de las palabras encontradas en dicho documento.

2) Escriba un programa que pida al usuario ingresar su correo electrónico y devuelva por aparte el nombre del usuario y el nombre del dominio (Ejemplo: del correo nombre.apellido@gmail.com, nombre de usuario: nombre.apellido, nombre del dominio: gmail.com)

Al final del ejercicio Ud. nos debe entregar:

- La implementación de sus **programas** en un notebook de Jupyter (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).

Nota

Usted tendrá 48 horas para construir el análisis requerido y luego contará con 20 minutos para exponer su propuesta técnico- comercial.