Prueba Autored Data Scientist

Se tiene una base de datos de publicaciones online de autos. Esta es obtenida mediante un scrapeo de un sitio web de venta de autos y luego es procesada para ordenarlos e intentar eliminar registros erróneos. Se le entrega el resultado en el archivo “autos.csv” en el cual se tienen las siguientes columnas:

* make\_id: id interna de la marca del auto
* model\_id: id interna del modelo del auto
* year: Año del auto
* cylinder\_capacity: Cilindrada del auto
* km: Kilometraje del auto
* color: Color del auto
* región: Región en la cual se ubica el vendedor del vehículo
* price: Precio de publicación del vehículo (en pesos chilenos)
* fuel\_type\_id: id interno del tipo de combustible del vehículo
* transmission\_id: id interno de la transmisión del vehículo
* traction\_id: id interno de la tracción del vehículo
* age: Antigüedad del vehículo

Para los ids de características, sus valores significan:

* fuel\_type\_id: 0: Bencina, 1: Diesel, 2: Híbrido, 3: Eléctrico
* transmission\_id: 0: Mecánico, 1: Automático
* traction\_id: 0: 4X2, 1: 4X4

1. **SQL**: Queremos obtener el precio promedio de los autos, agrupados por modelo y año, además nos interesa obtener este promedio sólo para grupos que contengan 3 o más registros de autos. Escriba la consulta en SQL que entrega el resultado solicitado.

SELECT model\_id, year, AVG(price) AS avg\_price

FROM your\_table\_name

GROUP BY model\_id, year

HAVING COUNT(\*) >= 3;

2. En esta etapa evaluaremos: Manejo de datos, Visualización, Análisis y Modelos predictivos.

Para esto use el lenguaje Python, y para las librerías debe usar Pandas para el manejo de datos, Sugerimos usar Matplotlib o Seaborn para la visualización (puede ser otra mientras sea fácil de instalar).

Para el modelo predictivo puede usar sklearn u otra parecida.

Utilice el archivo adjunto “autos.csv” para predecir el precio de un vehículo y analice el resultado de sus predicciones. Utilice gráficos para los análisis pre y post predicciones si le es útil. Justifique toda decisión que tome que no le parezca obvia.

Nota: haga todas las asunciones que considere razonables y necesarias y explique el porqué de esta. Estaremos midiendo su capacidad de razonamiento y argumentativa más que si la respuesta es correcta.

3. Los usuarios esperan que las predicciones tengan las siguientes características:

* Sean bastante precisas
* Sean justificables (explicables y entendibles)
* Predigan una gran variedad de vehículos

Suponga que para este caso uso 2 modelos predictivos. una regresión lineal multivariable y una red neuronal. La red neuronal tiene un error absoluto medio (EAM) de un valor de 20.000 menos que el error absoluto medio de la regresión

(EAMred + 20.000 = EAMregresión).

Considerando que el precio de los autos sea del orden de varios millones de pesos chilenos o más (ya sea por ejemplo 6.000.000 o 20.000.000 de pesos)

a) Usando solo el EAM ¿Cuál de estos 2 modelos preferiría y por qué?

b) Que otra métrica y/o estrategia utilizaría para tomar una mejor decisión (argumente).

a) Considerando solo el error absoluto medio (EAM), preferiría la red neuronal sobre la regresión lineal multivariable debido a que tiene un EAM inferior. Un menor error absoluto medio indica que el modelo es capaz de predecir con mayor precisión los precios de los autos. Dado que el costo de los autos es del orden de varios millones de pesos chilenos, incluso una diferencia de 20.000 pesos en el error puede tener un impacto significativo en la precisión de las predicciones y, por lo tanto, en la toma de decisiones.

b) Además del EAM, una métrica útil para evaluar la calidad de los modelos predictivos es el coeficiente de determinación (R cuadrado). El coeficiente de determinación proporciona una medida de la proporción de la variación en la variable dependiente que es predecible a partir de las variables independientes en el modelo. Un valor más cercano a 1 indica un mejor ajuste del modelo a los datos.

Otra estrategia que se podría utilizar para tomar una mejor decisión es la validación cruzada. La validación cruzada es una técnica que ayuda a evaluar la capacidad de generalización de un modelo al dividir los datos en conjuntos de entrenamiento y prueba múltiples veces y calcular métricas de rendimiento en cada división. Esto ayuda a detectar si un modelo está sobreajustando los datos de entrenamiento y si su rendimiento se mantiene consistente en diferentes conjuntos de datos.

Además, es importante considerar la complejidad del modelo y el tiempo de entrenamiento. La red neuronal puede ser más compleja y requerir más recursos computacionales y tiempo de entrenamiento en comparación con la regresión lineal multivariable. Por lo tanto, es esencial equilibrar la precisión del modelo con la eficiencia computacional y el tiempo de entrenamiento disponible.

4. Ahora en el scrapeo recibimos además el título de la publicación (Se adjunta el archivo ejemplos\_titulos.csv que contiene ejemplos de títulos scrapeados). Escriba todas las ideas que se le ocurra que sirva para mejorar sus predicciones. Justifique sus ideas.

Nota: Asuma que usted dispone de una tabla con la traducción de todos los ids usados hasta ahora, ya sea de las características del auto como de la marca y modelo

Agregar **Versión:** La versión del automóvil puede indicar características adicionales o niveles de equipamiento que podrían afectar su precio. Puedes codificar la versión como una característica categórica utilizando técnicas de codificación one-hot

Agregar agrupaciones por tipos de marca o tipos de autos.

Categorias de autos deportivos , HB , SUV, Camionetas , ETC.

Categorias de marcas , Premium, Genericas, Super Premium,