Desafío - Prediciendo los precios de las casas

En este desafío tendrás la oportunidad de poner a prueba los conceptos aprendidos durante la sesión. Los ejercicios están diseñados para reforzar practicar lo explicado en clases y poder implementar un caso real.

Lee todo el documento antes de comenzar el desarrollo individual, para asegurarte de tener el máximo de puntaje y enfocar bien los esfuerzos. Asegúrate de seguir las instrucciones específicas en cada ejercicio y de completar los requerimientos adicionales, si los hubiera.

Tiempo asociado: 4 horas cronológicas

Descripción

Como Cientista de Datos te han contratado en una importante empresa de propiedades para analizar las diferentes características de algunas casas que se han vendido en el último tiempo, y que se encuentran en el dataset **house_data.xlsx**. Esta base de datos contiene diversas características de estas propiedades y su precio. Específicamente, se te solicita:

1. Analizar la calidad de datos, para lo que debes cargarlos y realizar un proceso exhaustivo de limpieza para eliminar valores faltantes, duplicados y atípicos que puedan afectar la calidad del modelo, si los hay.

2. Realizar un análisis descriptivo de las variables para entender la distribución de los datos y detectar posibles relaciones entre las características y los precios. En esto debes incluir un análisis de correlaciones entre las variables principalmente con el precio de las casas, comenta acerca de las variables con mayor correlación, ¿tienen sentido?

3. Divide los datos en conjuntos de entrenamiento y prueba de manera aleatoria o estratificada, dejando un 33% para test, luego de eso aplica alguna técnica de transformación de datos, como normalización o estandarización, para asegurar que las variables estén en una escala comparable.
   (hint: Recuerda que en la estandarización o normalización de los datos el fit.transform solo se aplica a los datos de train, mientras que a los datos de test se les aplica solo el transform).

4. Implementar modelos de regresión lineal con regularización, como Ridge, Lasso y Elastic Net con ajuste de hiper parámetros, según tabla sugerida, debes construir además un modelo de árboles de regresión para capturar relaciones no lineales y

complejas entre las variables predictoras y el precio de las casas, finalmente genera un cuadro comparativo con modelos y métricas.

5. Elige uno de los modelos de regresión lineal con regularización implementados en el punto anterior y gráfica cómo varían sus hiperparámetros durante el ajuste, mostrando también cómo afectan el rendimiento del modelo.

6. Utilizar métricas de evaluación de regresión, como el Error Cuadrático Medio (MSE), el Error Absoluto Medio (MAE) y el Coeficiente de Determinación (R^2), para medir el rendimiento del modelo, realizar cuadro comparativo indicado en el punto 4, finaliza con una breve conclusión de a los resultados obtenidos.

```python
lasso_params = {'alpha': [0.001, 0.01, 0.1, 1, 10]}

ridge_params = {'alpha': [0.001, 0.01, 0.1, 1, 10]}
elastic_params = {'alpha': [0.001, 0.01, 0.1, 1, 10],
                  'l1_ratio': [0.1, 0.3, 0.5, 0.7, 0.9]}

tree_params = {'max_depth': [None, 10, 20, 30],
                   'min_samples_split': [2, 5, 10],
                   'min_samples_leaf': [1, 2, 4]}
```

Requerimientos

1. Analiza y explora y prepara un conjunto de datos, considerando la intención de realizar un modelo de regresión. **(2 puntos)**
2. Implementa modelos de regresión con enfoque de Machine Learning. **(5 puntos)**
3. Evalúa e interpreta modelos de regresión, para obtener conclusiones sobre ellos. **(3 puntos)**

¡Mucho éxito!

Consideraciones y recomendaciones

- Debes entregar la solución en un archivo de Jupyter Notebook, con el código y las explicaciones necesarias