Laboratorio 1
Predicción de Precios de Energía usando Regresión Múltiple

Objetivo:

El objetivo de esta tarea es aplicar técnicas de regresión múltiple para predecir los precios de la energía en base a la hora del día, el día del mes, el mes y el día de la semana. Se utilizarán los conceptos y métodos discutidos en clase para construir y evaluar modelos de regresión.

Datos:

En el GES se encuentra un archivo llamado POE_2023.xlsx que contiene la información de los precios de energía para el 2023. Cada pestaña es un mes del año, y contiene una tabla donde las filas son las horas del día, y las columnas son los días del mes. Esta información debe ser cargada en un dataframe usando pandas para su posterior análisis.

Requerimientos:

1. **Exploración y Visualización de Datos**
   - Cargar los datos del archivo POE_2023.xlsx en un dataframe de pandas.
   - Verificar si hay valores faltantes o errores en los datos.
   - Visualizar las relaciones entre las variables predictoras y la variable de respuesta. Utilizar gráficos apropiados, como scatterplots, boxplots, o gráficos de línea.
2. **Preprocesamiento de Datos**
   - Codificar las variables categóricas apropiadamente para su uso en un modelo de regresión (hora del día, día de la semana, mes, adicional puede crear otras variables a partir de estas como: día de la semana, horas hábiles, feriados, semana del mes, entre otras). Pueden usar `pandas.get_dummies` para crear variables dummy.
   - Normalizar o estandarizar las variables numéricas si se considera necesario.
   - Dividir los datos en conjuntos de entrenamiento y prueba.
3. **Ajuste de Modelos de Regresión**
   - Ajustar un modelo de regresión lineal múltiple usando sus predictores. Utilizar `sklearn.linear_model.LinearRegression`.
   - Considerar añadir interacciones entre variables (por ejemplo, un término de interacción entre hora, día de la semana y mes). Se puede usar `PolynomialFeatures` para crear términos de interacción.
   - Explorar relaciones potencialmente no lineales (por ejemplo, los precios pueden variar de manera no lineal a lo largo del día) y considerar transformaciones de variables.

4. **Evaluación de Modelos**
   - Calcular el RSS y el R^2 de cada modelo en los conjuntos de entrenamiento y prueba. Utilizar `sklearn.metrics.mean_squared_error` y `sklearn.metrics.r2_score`. Para responder a las preguntas o comparar modelos considerar el uso de MSE, RMSE o MAE.
   - Utilizar validación cruzada de 5-fold para obtener estimaciones más robustas del rendimiento de cada modelo. Se puede usar `sklearn.model_selection.cross_val_score`.
   - Comparar diferentes modelos de regresión (por ejemplo, con diferentes combinaciones de predictores) y seleccionar el mejor basándose en las métricas de rendimiento.
5. **Interpretación de Resultados**
   - Examinar los coeficientes del mejor modelo e interpretar su significado en el contexto de los precios de la energía.
   - Visualizar los precios reales vs los precios predichos por el mejor modelo.
   - Discutir las implicaciones de los resultados. ¿Qué factores parecen influir más en los precios de la energía? ¿Hay algunas limitaciones del modelo?
6. **Interrogantes del Modelo**
   - En promedio, ¿cuánto se equivoca el modelo en sus predicciones?
   - ¿Cuáles son las horas del día que son más predecibles para su modelo?
   - ¿Cuáles son las horas menos predecibles para su modelo?
   - ¿Cómo afectan los cambios de temporadas (invierno/verano) a su modelo?
   - ¿Cómo se comporta el modelo durante la Semana Santa u otros días festivos?
   - Demuestre que su modelo no simplemente ha memorizado los datos de entrenamiento. Utilice técnicas como la validación cruzada o la evaluación en un conjunto de datos de prueba independiente para evaluar la capacidad de generalización del modelo.
   - Si usted fuera el administrador del mercado y tuviera que predecir el precio de la energía para 2024, ¿cree que su modelo es lo suficientemente bueno? Justifique su respuesta y proporcione evidencia que respalde su conclusión.

Entregables:

- Un notebook de Jupyter documentado (código y análisis) que muestre el trabajo realizado en cada uno de los requerimientos anteriores.
- Resuma los hallazgos, interprete los resultados y discuta las implicaciones.
- En caso de ser necesario, explore técnicas de regularización como Ridge o Lasso para controlar la complejidad del modelo y evitar el sobreajuste. Compare su rendimiento con el modelo no regularizado.