Proyecto final

October 28, 2024

Longevidad energética: Exploración de modelos de regresión en la predicción de la vida útil de una batería.

Nombre: Leydy Vanesa Cortés
Correo: lcortesro@unal.edu.co

Nombre: Jhon Fredy Aguirre
Correo:jhaguirreg@unal.edu.co

Universidad Nacional de Colombia

Programación II

Resumen.

En el presente artículo se analiza un conjunto de datos proporcionados por el Instituto de Energía Natural de Hawái el cual examinó 14 baterías NMC-LCO 18650 que se ciclaron más de 1000 veces a 25°C con una tasa de carga de Corriente Continua de C/2 y una tasa de descarga de 1,5C. Mediante la regresión lineal múltiple, modelaremos la relación entre la vida útil de la batería (Remaining Useful Life) y las variables predictoras como lo son la duración de carga, el voltaje y el tiempo de descarga, consiguiendo así un modelo predictivo que pueda estimar la vida útil basándose en los parámetros de entrada. Finalmente, haciendo uso de la regresión por componentes principales, reduciremos la dimensionalidad del conjunto de datos, buscando mejorar la precisión del modelo.

1. Introducción.

La Vida útil de una batería, de ahora en adelante RUL (por sus siglas del inglés Remaining Useful Life), generalmente se puede estimar por su capacidad (mAH). Sin embargo, ¿Qué pasa si esta capacidad no se puede medir? La idea detrás del proyecto es predecir la RUL midiendo otros factores como voltaje (V) y corriente (A). Para ello aplicaremos el proceso de regresión lineal múltiple a un conjunto de datos que proporciona tanto la RUL de la batería como la medida de sus otras características, así, nos enfocaremos en dos objetivos:

1. Extraer y crear funciones a partir de los conjuntos de datos de origen, que dependan del voltaje, la corriente y el tiempo.

2. Predecir la vida útil restante (RUL) de una batería mediante un modelo base de regresión multiple.

Eventualmente, buscaremos reducir por medio de PCA (Análisis de Componentes Principales) los datos mencionados, evaluando si es factible reducir la cantidad de información necesaria para ejecutar dicho modelo, visualizando progresivamente cual es el impacto en la exactitud al prescindir de algunos valores.

2. Metodología.

La metodología propuesta para una correcta solución del problema y de los objetivos propuesto consta de los siguientes pasos:

1. Recopilación y preprocesamiento de los datos: Una vez importados los datos sobre la vida útil de las baterías y sus atributos relacionados (duración de carga, tiempo de uso y otros factores relevantes), es necesario realizar un preprocesamiento antes de aplicar los modelos de regresión. Esto incluirá la limpieza de datos inconsistentes, la gestión de valores faltantes y la normalización de variables si es necesario.

   Cabe resaltar que hay variables a las cuales es imposible acceder con nuevos datos, estando fuera del experimento, por lo cual resulta importante prescindir de estas y adaptar el modelo a valores que se puedan obtener en cualquier momento en que se desee hacer una predicción de la vida útil de una batería.

2. Aplicación de regresión lineal múltiple: Después de haber encontrado y separado nuestros datos en variables dependientes e independientes, se aplicará la técnica de regresión lineal múltiple para modelar la relación entre la vida útil de la batería y las variables predictoras. Se realizará una exploración de las correlaciones entre las variables y se ajustará el modelo para obtener los coeficientes de regresión.

3. Evaluación del modelo: Se evaluará el rendimiento de los modelos de regresión utilizando métricas como el error cuadrático medio, el coeficiente de determinación (R-cuadrado) y técnicas de validación cruzada. Esto permitirá estimar la precisión del modelo. Con la finalidad de darle realismo al modelo, dividiremos los datos en dos conjuntos, uno de entrenamiento con el que se creará el modelo y uno de prueba, con el que se calcularán las métricas mencionadas.

4. Aplicación y evaluación de regresión por componentes principales: En todo modelo de regresión que trabaje con una cantidad considerable de variables podría ser útil emplear regresión por componentes principales para reducir la dimensionalidad del conjunto de datos. Por ello, se realizará un análisis de componentes principales para identificar las variables más relevantes y se construirá un modelo de regresión utilizando estas componentes. A este módelo también se le calcularán las métricas usuales para estimar que tan provechoso es usar esta técnica en nuestro conjunto de datos.

5. Análisis de resultados: Se analizarán los coeficientes de regresión y los componentes principales para comprender la influencia de cada variable en la vida útil de la batería. Por medio de gráficas y algunos valores de estimación, se extraerán conclusiones y se brindarán recomendaciones basadas en los resultados obtenidos.

3. Experimentación.

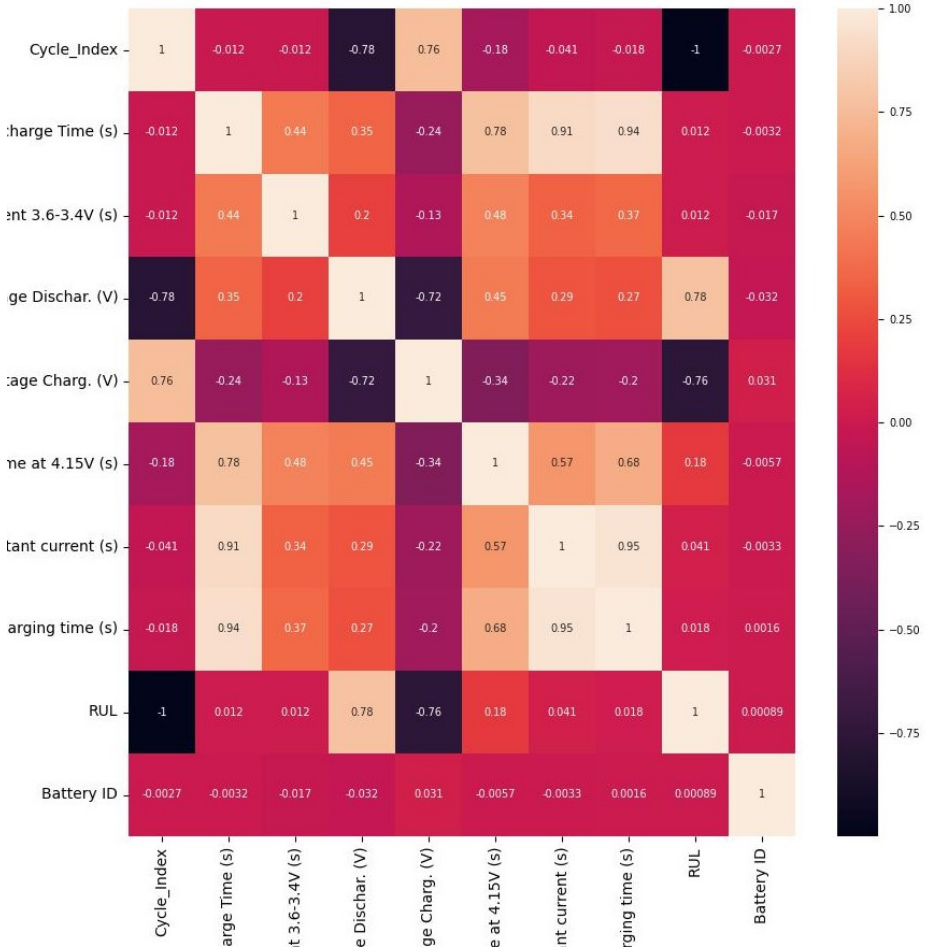
Inicialmente debemos importar los datos, para ello usaremos la librería Pandas de Python, ya que ofrece un fácil y rápido manejo de tablas. Los datos que vamos a utilizar contienen la siguiente información:

Cycle Index	Número de ciclo
Discharge Time (s)	Tiempo de descarga (s)
Decrement 3.6-3.4V (s)	Decremento 3.6-3.4V (s)
Max. Voltage Dischar. (V)	Máx. Descarga de voltaje (V)
Min. Voltage Charg. (V)	Mín. Carga de voltaje (V)
Time at 4.15V (s)	Tiempo a 4.15V (s)
Time constant current (s)	Tiempo de corriente constante (s)
Charging time (s)	Tiempo de carga (s)
RUL	Vida útil estimada

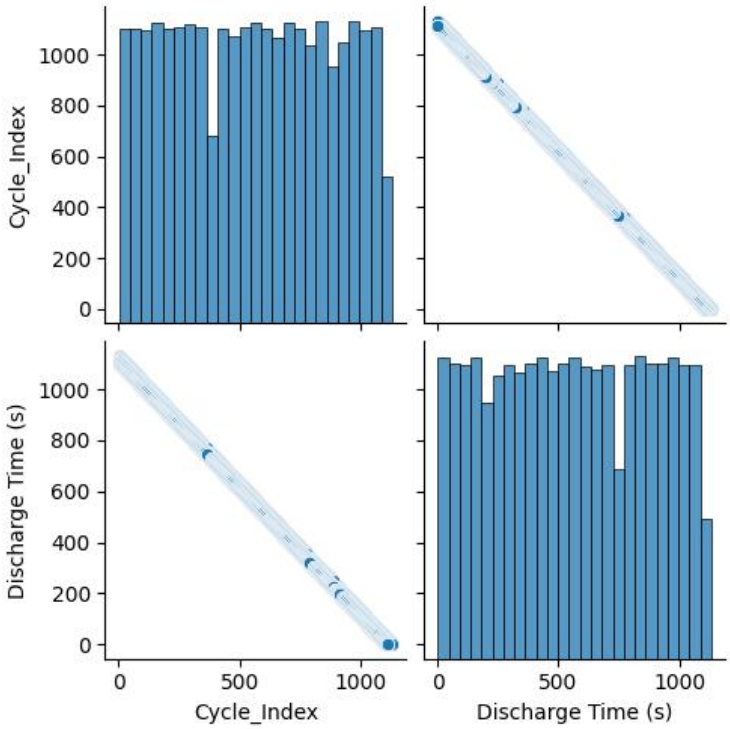
Algunos valores adicionales sobre nuestros datos se presentan a continuación:

Etiquetas	Cycle_Index	Discharge Time (s)	Decrement 3.6-3.4V (s)	Max. Voltage Dischar. (V)	Min. Voltage Charg. (V)	Time at 4.15V (s)	Time constant current (s)	Charging time (s)	RUL
count	15064.0	15064.0	15064.0	15064.0	15064.0	15064.0	15064.0	15064.0	15064.0
mean	556.0	4581.0	1239.0	3.0	3.0	3768.0	5461.0	10066.0	554.0
std	322.0	33144.0	15039.0	0.0	0.0	9129.0	25155.0	26415.0	322.0
min	1.0	8.0	-397645.0	3.0	3.0	-113.0	5.0	5.0	0.0
25%	271.0	1169.0	319.0	3.0	3.0	1828.0	2564.0	7841.0	277.0
50%	560.0	1557.0	439.0	3.0	3.0	2930.0	3824.0	8320.0	551.0
75%	833.0	1908.0	600.0	3.0	3.0	4088.0	5012.0	8763.0	839.0
max	1134.0	958320.0	406703.0	4.0	4.0	245101.0	880728.0	880728.0	1133.0

Usando el método *isna()* verificamos que no hay valores ilegibles en el DataFrame, por lo cual se prosigue encontrando las variables independientes del módelo, es importante recordar que la columna RUL es nuestra variable objetivo ya que indica la vida útil de cada batería. En este caso, tendremos en cuenta todas las variables, menos *Cycle Index* , ya que esta es una medida que solo se toma al momento de hacer el experimento, además la dependencia lineal que tiene con la variable objetivo es considerablemente alta, esto se puede evidenciar en la matriz de correlación propuesta a continuación:



Para ser más precisos, veamos la gráfica de la variable *RUL* con respecto a la variable *Cycle Index*:



Una vez determinadas las variables que se usarán en el módelo procedemos a dividir los datos en conjuntos de entrenamiento y prueba, recordemos que poseemos los datos de 14 baterias, por lo cual dejamos 9 para entrenar el modelo y 5 para evaluarlo. Para realizar la efectiva separación de estas, indexamos el DataFrame numerando cada bateria. Paso siguiente, extraemos los valores de la tabla a arrays de Numpy, donde conseguimos cuatro matrices:

Matriz	Dimensión
x_train	(9708, 7)
y_train	(9708, 1)
x_test	(5356, 7)
y_test	(5356, 1)

Completado ahora el preprocesamiento de los datos, iniciamos el módelo de regresión con los conjuntos de entrenamiento, obteniendo los siguientes coeficientes β_i como solución:

β_0	-8255.113876257205
β_1	-0.008450762912490752
β_2	-0.0006227804459565737
β_3	2822.305113930753
β_4	-631.6674588665485
β_5	0.0029057947602037972
β_6	0.0021899046237161433
β_7	0.004548388749618867

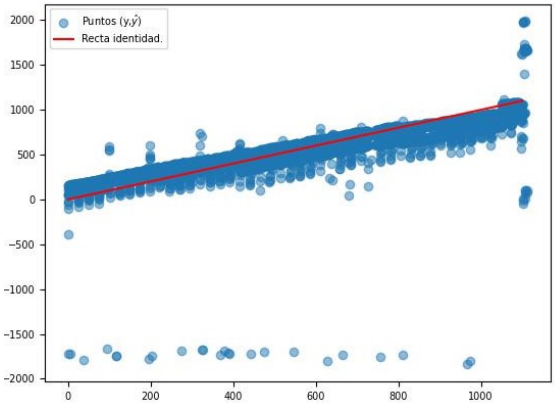
Donde:

$$RUL \approx \hat{y} = \beta_0 + \beta_1 * DT + \beta_2 * D3.6V + \beta_3 * MVD. + \beta_4 * MVC + \beta_5 * T4.15V + \beta_6 * TCC + \beta_7 * CT.$$

Podemos revisar la precisión del modelo con las siguientes métricas de ajuste:

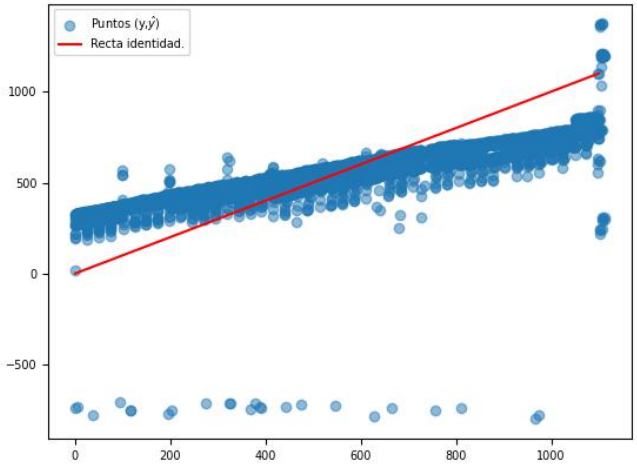
Error cuadrático	144912219.59032118
Error cuadrático medio (MSE)	37377.86377373016
Coeficiente de determinación (R-cuadrado)	0.8969645065489572
Validación cruzada	0.8626036873910552

Ya que la dimensión de los datos no nos permite presentar una gráfica de la situación, podemos visualizar que tan cercana es la aproximación $\hat{y} \approx y$, graficando los puntos $(y, \hat{y})$ en el plano cartesiano, los cuales deberían estar bastante cerca a la recta identidad:

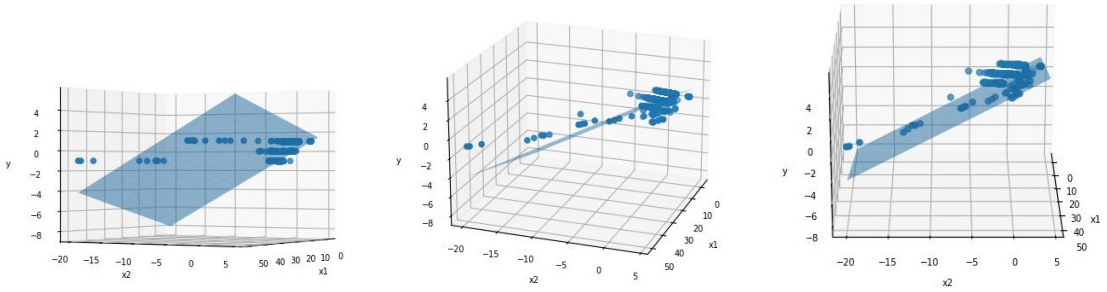


Siguiendo la misma metodoloía, reduzcamos los datos por medio del análisis de componentes principales, conservando en inicio un 97% de la información original, obtenemos los siguientes resultados:

Error cuadrático	1399.0044926044843
Error cuadrático medio (MSE)	42793.839443493336
Coeficiente de determinación ajustado (R-cuadrado ajustado)	0.5869455538996088
Validación cruzada	0.6639163500568899

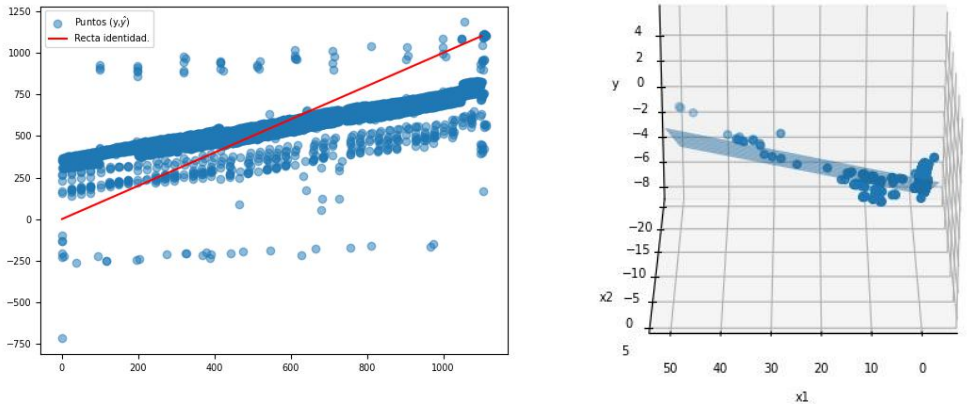


A pesar de haber hecho una reducción de la dimesión de los datos, seguimos encontrandonos en $\mathbb{R}^7$, no es hasta que decidimos conservar solo el 55% de la información cuando podemos obetener una gráfica del plano encontrado:

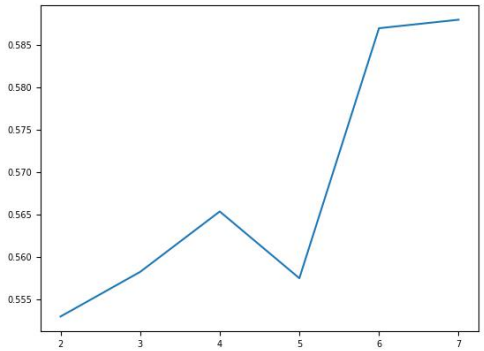


Los resultados de dicho modelo también se presentarán a continuación:

Error cuadrático	1686.365556294486
Error cuadrático medio (MSE)	46346.02042620815
Coeficiente de determinación ajustado (R-cuadrado ajustado)	0.5529934239751098
Validación cruzada	0.5949486483088842



Realizando PCR con diferentes porcentajes de varianza, obtenemos la siguiente gráfica que relaciona el crecimiento del coeficiente de determinación (R-cuadrado) con respecto al número de variables independientes que se conservan:



4. Conclusiones.

1. Se ha encontrado una correlación negativa perfecta entre la variable Cycle Index y la variable Remaining Useful Life (RUL), como indica el valor de -1 obtenido en la matriz de correlación. Esto implica que a medida que el Cycle Index aumenta, el valor de RUL disminuye de manera proporcional y predecible, es decir, una es inversamente proporcional de la otra.

2. El desarrollo del trabajo ha expuesto el análisis de un grupo de datos los cuáles se pudieron estudiar de forma más detallada a través de la regresión múltiple cumpliendo el objetivo propuesto. Además, no fue necesario reducir la dimensionalidad del conjunto de variables predictoras por medio de PCR, esto se debe a que la cantidad de variables independientes consideradas en el módelo es baja.

3. La regresión por componentes principales se utiliza para reducir la dimensionalidad de un conjunto de variables predictoras altamente correlacionadas, transformándolas en un conjunto de variables no correlacionadas llamadas componentes principales. Sin embargo, al hacer esto, se pierde la interpretabilidad directa de las variables originales y la relación con la variable de respuesta puede volverse menos clara. Evidenciamos esto en la fluctuación que tuvieron los errores de ajuste al realizarle PCR a los datos, provocando que, aunque logramos obtener un esquema ilustrativo de la situación, se perdiera gran parte de la información, haciendo inexacto el módelo.

4. Este proyecto nos permitió obtener un modelo predictivo que estima la vida útil de una batería, lo cual contribuirá al desarrollo de tecnologías de baterías más eficientes y estrategias de uso optimizadas. Las métricas de ajuste presentadas nos permiten concluir que para este conjunto de datos, el aplicar regresión múltiple genera modelos eficientes.

5. Referencias.

1. Código del proyecto: Clic aquí.

2. Base de datos: Clic aquí.

3. Proyectos guía: Clic aquí.

4. Proyectos guía: Clic aquí.

5. Proyectos guía: Clic aquí.

6. Proyectos guía: Clic aquí.

7. Información sobre los datos: Clic aquí.