LLAMARADAS SOLARES

Valentina Vásquez, Amelia Hoyos, Ricardo Gandica.
Modelación y Simulación IV, Ingeniería matemática
Universidad EAFIT

---

1. Entendimiento del Negocio.

Contexto

El Sol, especialmente en lugares donde se ubican sus manchas, es muy activo, emitiendo radiación en todas direcciones. Una erupción o llamarada solar es una de estas intensas y repentinas ráfagas de radiación. Estas varían de acuerdo con el ciclo solar de once años. Se cree que una erupción solar ocurre cuando un depósito de energía magnética acelera a gran velocidad iones presentes en el plasma solar, lo que produce radiación. Esta radiación es absorbida por la ionosfera de la Tierra. No representa un peligro directo para la vida en el planeta, sin embargo, sí provoca que la atmósfera superior se ionice, lo que puede interferir con las comunicaciones por radio, dependiendo de la energía de la llamarada solar.

Algunas veces, las llamaradas solares son acompañadas por eyecciones de masa coronal (CME), ondas de radiación y viento solar. Esto puede causar tormentas geomagnéticas en la Tierra, lo que puede producir un efecto desastroso en las telecomunicaciones terrestres. Como estas tormentas están generalmente asociadas a erupciones solares, pueden detectarse antes de que sucedan. En 1859, ocurrió algo llamado el evento Carrington, durante el cual el Sol liberó una CME tan grande que en la Tierra las comunicaciones telegráficas fallaron y aparecieron auroras en toda la Tierra, provocando una conmoción generalizada. En Montería, Colombia, se dio el siguiente testimonio:

Según Exbrayat, en su libro *Historia de Montería* publicado por la Imprenta Departamental de Córdoba,

> En marzo de 1859 los habitantes de la ciudad contemplaron estupefactos un fenómeno celeste que semejaba un incendio de vastas proporciones. Negros nubarrones surcados a cada instante por candilejas y extraños fulgores; inmensas lenguas de llamas y deslumbrantes globos luminosos que deshacían para volver a condensarse segundos después, daban la impresión de cien volcanes en erupción. Y en el centro de aquellos inmensos espacios se formó, claramente dibujada, una enorme S que allí permaneció mientras duraba el fantástico espectáculo causado por la madrugada. Tan intensa era la claridad que podía confundirse con una aurora boreal y penetraba hasta el interior de las casas.
>
> Impresionadas y temerosas las gentes se precipitaron hacia la iglesia y, postradas de hinojos ante la Majestad Divina, pedían, a grandes voces, que se alejaran las calamidades presagiadas por el raro fenómeno. El Padre José Luis Pezzati públicamente tomó en llamada Oración de los Siete Derramamientos de Sangre de Nuestro Señor Jesucristo queriendo explicar: 'Sangre preciosa por mi amor vertida... Purifica mi alma de toda malicia.' Exbrayat 1971.

Si algo así sucediera en la actualidad, toda la tecnología de la que dependemos dejaría momentáneamente de funcionar. Por lo tanto, es importante prepararse para un evento así y saber con anticipación si sucederá en algún momento determinado. Aquí entra en juego la detección de llamaradas solares, que como ya sabemos, son un indicador de una posible tormenta geomagnética.

Objetivos de Negocio

Los datos que usaremos son tomados de *Kaggle*, *Solar Flares from RHESSI Mission*. Estos datos provienen de RHESSI, un observatorio de erupciones solares de la NASA. El diccionario de datos se encuentra en el Apéndice 1. Allí también se explica el significado de ciertas variables categóricas que aparecerán más adelante, denominadas Flags, y se encuentra también el link a la base datos usada. Al igual que el link al GitHub donde se encuentra el proyecto.

Nuestro objetivo es predecir el tiempo que tarda en ocurrir el siguiente pico de radiación de la siguiente llamarada solar, dado que tenemos datos de la llamarada solar actual. La predicción puede mejorar significativamente la gestión y mitigación de los impactos de las erupciones solares en la tecnología y la infraestructura crítica. Además, esta predicción es valiosa para las misiones espaciales, ya que permite a las agencias espaciales planificar actividades extravehiculares, mejorar sus mediciones y proteger a los astronautas de la radiación solar intensa. En el ámbito de la investigación científica, conocer el intervalo hasta el próximo pico puede proporcionar datos valiosos para estudiar los ciclos solares y mejorar los modelos predictivos del clima espacial, contribuyendo a un mejor entendimiento del Sol y su influencia sobre la Tierra.

Criterios de Éxito

Para considerar exitoso un proyecto como este, la predicción del tiempo hasta el próximo pico debe ser lo suficientemente precisa para demostrar la viabilidad del modelo y su aplicabilidad en situaciones reales. Un criterio de éxito concreto sería la capacidad del modelo para predecir el próximo pico de llamarada solar mejor que una predicción basada en la media o en la mediana. Esto implica que el modelo debe ser capaz de predecir el tiempo real entre picos con mayor precisión que el valor de la media o la mediana cuando se usa para este mismo propósito. Esta capacidad permitiría a los sistemas de infraestructura y comunicaciones prepararse mejor que en la actualidad para mitigar los efectos de la actividad solar.

Además, el modelo debe proporcionar predicciones con un margen de error aceptable para que sean útiles. La selección del mejor modelo se basará en el R-cuadrado (R2) score, y se evaluará esta precisión utilizando métricas como el Error Cuadrático Medio (MSE), Error Absoluto Medio (MAE) y Error Porcentual Absoluto Medio (MAPE), así como sus equivalentes calculados con la mediana en lugar del promedio (MSE mediano, MAE mediano y MAPE mediano).

Evaluación de la situación

En primer lugar, identificamos nuestros recursos disponibles, los cuales incluyen las bases de datos históricas de erupciones solares, las herramientas de análisis de datos como Python y bibliotecas específicas de machine learning, y nuestra capacidad para investigar sobre este tema relacionado con la astrofísica.

En términos de limitaciones, consideramos restricciones como la calidad y la cantidad de los datos disponibles, el tiempo limitado para completar el proyecto y la capacidad de procesamiento computacional disponible. Este útlimo resultó ser una dificultad importante. Además, se consideraron algunos supuestos imporantes sobre la naturaleza del fenómeno, como la estabilidad de los patrones de las erupciones solares a lo largo del tiempo y la suposición de que los datos disponibles son representativos de eventos futuros.

Riesgos y contingencias

En la ejecución de este proyecto, se han identificado varios riesgos y se han desarrollado planes de contingencia para abordarlos. La calidad y cantidad insuficiente de datos se tratarán mediante técnicas de limpieza de datos, enriquecimiento de datos y búsqueda de fuentes adicionales para garantizar la disponibilidad de datos de alta calidad. Las restricciones de tiempo se gestionarán priorizando las tareas críticas y ajustando el alcance del proyecto según sea necesario para cumplir con los plazos establecidos. Las limitaciones computacionales se supe-

rarán mediante la optimización del código y el uso eficiente de los recursos informáticos disponibles.

Se llevarán a cabo pruebas exhaustivas para validar y ajustar los supuestos del modelo, asegurando así su fiabilidad y precisión. Los problemas técnicos que puedan surgir se abordarán con revisiones continuas y la implementación de alternativas técnicas viables. Además, se establecerán reuniones regulares y canales de comunicación claros entre los miembros del equipo para facilitar una colaboración fluida y garantizar que todos estén alineados con los objetivos y avances del proyecto.

Estos planes de contingencia se han diseñado para garantizar que el proyecto pueda adaptarse y progresar de manera efectiva a pesar de los desafíos que puedan surgir durante su ejecución.

Costos y beneficios

En este caso hipotético, si consideráramos que este proyecto se realizara en la vida real, tendríamos que evaluar los costos asociados con cada fase del proyecto. Por ejemplo, en términos de los datos necesarios para entrenar el modelo de aprendizaje automático, los costos podrían variar significativamente dependiendo de si se necesitan recolectar nuevos datos o si ya se cuenta con un conjunto de datos existente. Además, en la fase de investigación, los costos estarían relacionados con la contratación de personal especializado y posiblemente el uso de recursos externos para ciertas tareas. Por último, en la etapa de producción, habría que considerar los costos de infraestructura tecnológica, integración de sistemas y mantenimiento continuo del modelo. Si bien estos costos no se traducen directamente en ganancias económicas, el éxito del proyecto se mediría en términos de su impacto positivo en la gestión y mitigación de los riesgos asociados con las erupciones solares, así como en su contribución al avance científico en el campo de la climatología espacial.

Hemos encontrado que bajo las condiciones más simples e ideales podría costar aproximadamente 27,000 dólares (Incze 2021). Respecto a las ganancias, al ser un proyecto enfocado en reducir daños para la población en general, podría ser cubierto por el gobierno. Los daños evitados en las telecomunicaciones podrían ser cientos de miles de dólares o incluso más, dependiendo de la región en la que se enfoque.

El objetivo de minería de datos en este proyecto es proporcionar un análisis preciso que permita predecir el tiempo hasta el próximo pico solar, dado el pico actual. Esto implica procesar y explorar datos históricos para descubrir patrones relevantes. Los resultados esperados incluyen el modelos predictivo que estimen de manera fiable el tiempo hasta el siguiente pico, así como informes y visualizaciones claras que faciliten la interpretación de los resultados. El éxito del proyecto se medirá mediante criterios técnicos específicos, explicados anteriormente.

El plan del proyecto abarcará varias etapas esenciales. La primera de ellas incluye la exploración y limpieza de los datos históricos de erupciones solares para garantizar su calidad. Posteriormente, se emplearán técnicas de minería de datos, como el uso de distintos algoritmos para modelar el problema o la identificación de datos atípicos. Más adelante se sigue con la creación y evaluación del modelo predictivo. Por último, se realiza el despliegue de la aplicación para su uso común.

2. Entendimiento de los datos

Se tiene una base de datos de 116,143 llamaradas solares. Cada una tiene algunas características asociadas a ella. Estas características incluyen un número de identificación único para cada llamarada solar, la fecha, hora de inicio, hora pico y hora de finalización de la llamarada, así como su duración en segundos. Además, se registran datos sobre la tasa y el número total de conteos de radiación detectados durante la llamarada en un rango de energía específico, junto con la banda de energía más alta observada. También se registra la posición de la llamarada solar con respecto al centro del sol en coordenadas X e Y, así como su distancia radial desde el centro del sol. Se asigna un número correspondiente a la región activa más cercana (un ID de manchas solares), si está disponible, y se registran códigos de calidad de la medición asociados con cada llamarada solar. Estos

datos proporcionan información detallada sobre las características y el comportamiento de las llamaradas solares, que se usarán en su análisis.

Análisis de la columna objetivo

Hemos graficado dos histogramas para la columna TIME TILL NEXT PEAK, que se calculó restando la hora pico de la llamarada solar siguiente a la actual (Figura 1, 2). El primero abarca un amplio rango de valores (menores de 60,000), mientras que el segundo se enfoca en valores más pequeños (menores a 4,000). Del análisis del primer histograma, observamos que la mayoría de los picos de las llamaradas solares ocurren poco tiempo después de otro pico. Esta distribución no es normal y está sesgada a la derecha. Es importante notar que hay muchos valores atípicos (outliers), los cuales pueden resultar problemáticos, por lo que más adelante aplicaremos algún tratamiento para manejar estos outliers.

El segundo histograma nos permite observar con mayor detalle la distribución de los valores más pequeños. Al igual que en el primer caso, la distribución no es normal. La distancia promedio entre picos en estos datos es de aproximadamente 700 segundos. Vemos que una gran cantidad de datos se concentran en este valor. No obstante, más adelante se encuentra un pequeño resurgimiento.

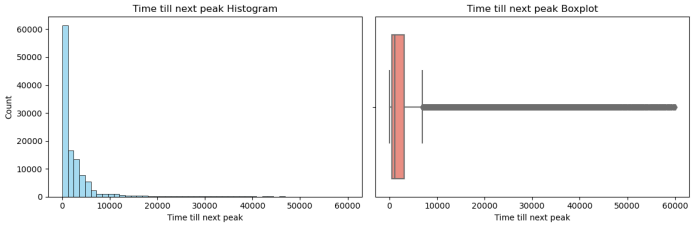

Figura 1: Histograma y Boxplot del tiempo hasta el próximo pico

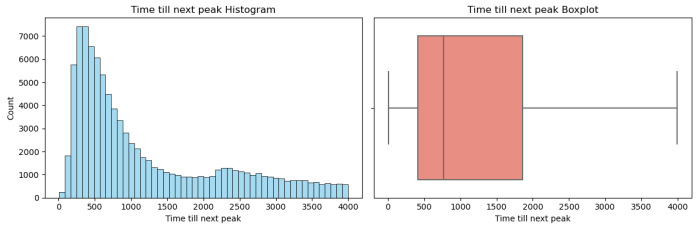

Figura 2: Histograma y Boxplot del tiempo hasta el próximo pico de tiempos cortos

Análisis de datos tipo fecha

Se inicia corrigiendo el formato de fechas para facilitar la comprensión de la base de datos. La variable DATE se ha configurado como tipo fecha y se ha concatenado con las variables START-TIME, PEAK TIME y END TIME para mejorar la legibilidad y facilitar el análisis temporal de los datos. En casos donde la hora de inicio excede la hora de finalización, especialmente alrededor de la medianoche, se ha agregado un día a la variable END TIME para garantizar la coherencia temporal de los registros. Este ajuste también se ha aplicado a la variable PEAK TIME. Los datos abarcan desde 2002 hasta 2018, cubriendo un período de 16 años. Este proceso de estandarización de fechas permite una interpretación más clara y precisa de la secuencia temporal de los eventos registrados, facilitando así el análisis y la comprensión de la actividad de las llamaradas solares a lo largo del tiempo.

Se observó la cantidad de llamaradas que ocurrieron cada mes durante la duración del proyecto RHESSI, encontrando que estas erupciones solares siguen el ciclo solar de 11 años. Se tomó como inicio del ciclo solar el año 2008.

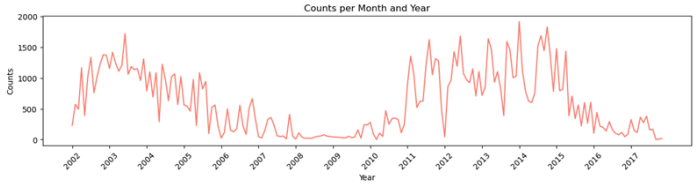

Figura 3: Ciclo solar

En función de lo descubierto anteriormente, se analiza cómo se comporta la variable TIME TILL NEXT

PEAK durante el ciclo solar (Figura 3). Se agregó una nueva variable llamada SOLAR CYCLE, que indica el año del ciclo solar al que pertenece cada registro. Para ver claramente la distribución de los datos, se tomó el rango entre 0 y 10,000 de tiempos hasta el pico siguiente.

Vemos que en la mitad del ciclo (Figura 4) (años 5-9) la cantidad de llamaradas solares es mucho mayor que a los extremos del ciclo. También es aparente que la varianza de estas llamaradas es menor en la mitad del ciclo, ya que tiene distribiciones más centralizadas que a los extremos, aunque tengan un promedio similar. Sin embargo nótese que en los años de la mitad del ciclo, hay dos modas o incluso 3, la de la izquierda dominante. Vemos que la duración promedio de las llamaradas se mantiene más o menos igual.

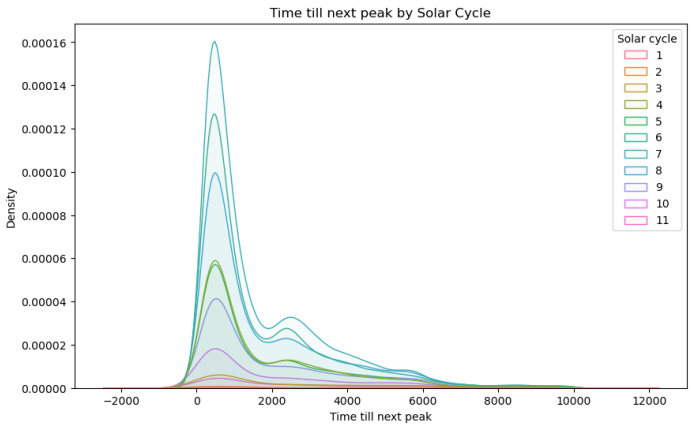

Figura 4: Tiempo hasta el siguiente pico según el ciclo solar

Luego, se calculó el tiempo que tarda en llegar al pico desde el inicio de la erupción y el tiempo que tarda en terminar la erupción desde el pico (Figura 5). Se encontró que el tiempo antes del pico es mayor que el tiempo después del pico, lo que confirma que una vez que una erupción libera una gran cantidad de energía en su punto máximo, se disipa rápidamente porque ya no tiene suficiente energía para mantenerse.

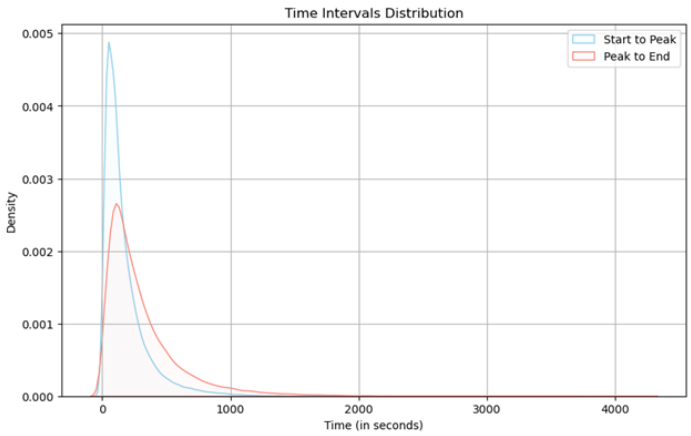

Figura 5: Distribución de intervalos de tiempo

Análisis de las variables categóricas.

Se prosigue con el análisis de las variables categóricas FLAG 1 a FLAG 5. A cada erupción se le asignaran algunas FLAGS y se guardan ahí como una lista sin un orden específico, siendo cada columna una posición de la lista. En la última columna, como ya no hay espacio para más FLAGS, algunas veces se juntan en un string, separadas por un espacio. Convertiremos cada posible FLAG en una columna binaria para analizarlas independientemente. Véase la distribución de FLAGS en la Figura 6.

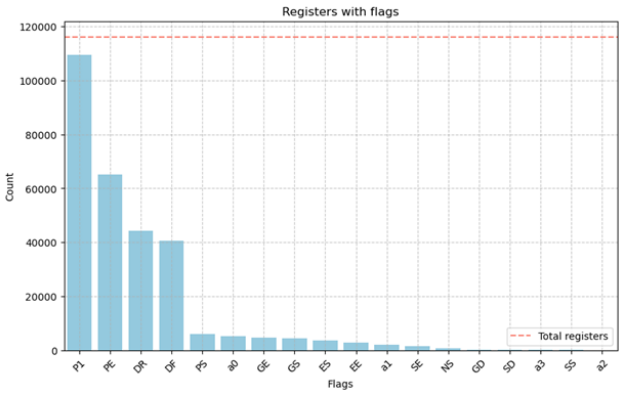

Figura 6: Flags Binarias

Entre las características más comunes está P1, que indica una posición válida para la llamarada, lo que sugiere que la mayoría de las erupciones solares registradas tienen una posición válida. Otras características incluyen PE, indicando que la llamarada

produjo materia además de energía, y DR y DF, que señalan decimación en las mediciones, disminuyendo la calidad de estas. Las definiciones completas de cada característica se encuentran en los apéndices.

Las siguientes FLAGS se analizan por separado ya que no tienen valores binarios (Figura 7). La mayoría de los datos tienen la característica A0, indicando un estado normal del atenuador en el pico del destello, seguido por A1 (estado grueso) y A3 (estado delgado). Para las FLAGS Q, que indican la calidad del registro (Q1 siendo la más alta y Q11 la más baja), las más comunes son Q1, Q2, Q3 y Q7, sugiriendo que la mayoría de los registros tienen buena calidad.

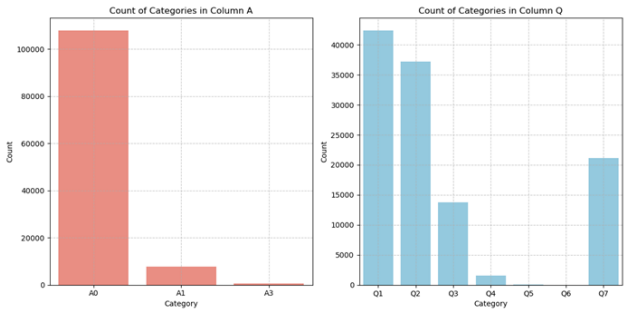

Figura 7: Flags de más categorías

En el análisis ENERGY, las categorías son 3-6, 6-12, 12-25, 25-50, 50-100, 100-300, 300-800, 800-7000 y 7000-20000 keV. Las más comunes son 3-6, 6-12 y 12-25, correspondiendo a los niveles más bajos de energía, siendo 6-12 la más predominante. Es raro encontrar muchas llamaradas solares de alta energía y también es difícil detectar llamaradas de baja energía con los instrumentos modernos.

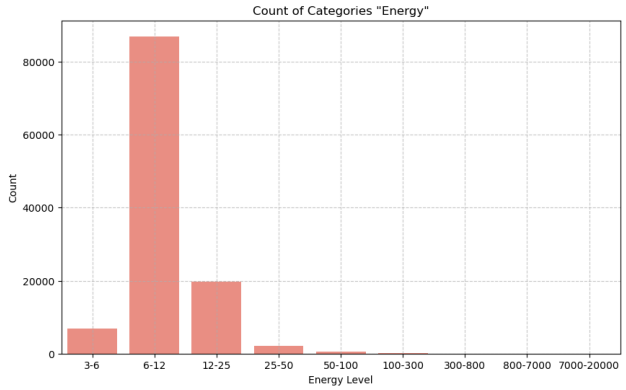

Figura 8: Energía

Para finalizar el análisis de los datos categóricos, se estudió el tiempo hasta el siguiente pico de una erupción dependiendo del nivel de energía emitido. Estas distribuciones exhiben diferentes valores medios y variabilidades. Mientras algunas muestran una forma de campana, especialmente las de niveles de energía más altos, otras presentan una marcada asimetría. Lo más importante a notar es que para energías bajas los datos están concentrados en un rango muy específico de valores, mientras que, en los rangos más altos, los datos son más dispersos.

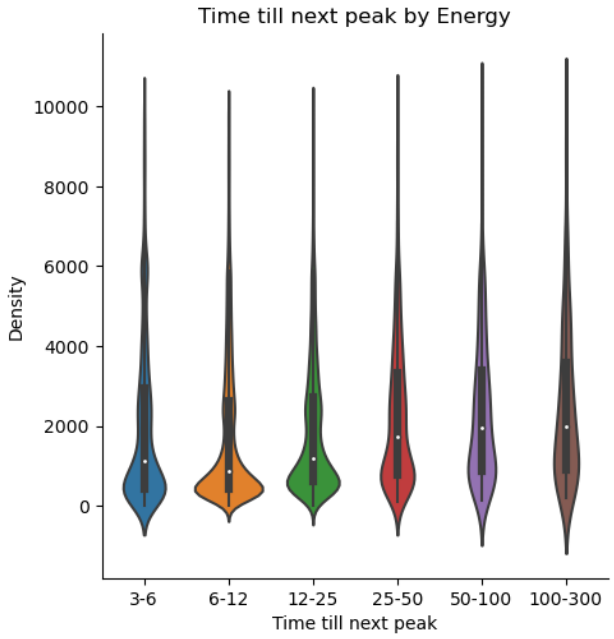

Figura 9

Análisis de las variables númericas.

Se excluyen los siguientes datos numéricos del análisis: FLARE y ACTIVE REGION, ya que son un ID y Duration, ya que no es válido utilizarla en el modelo.

Se encuentra que hay mucha variabilidad en las variables PEAK COUNTS y TOTAL COUNTS. Se utilizan histogramas, graficos de cajas de bigotes y de dispersión con TIME TILL NEXT PEAK count, para identificar visualmente los datos atípicos y ver el comportamiento con la variable a predecir (Figura 10, 11). Hemos encontrado que, debido a la gran cantidad de atípicos, no se puede ver con claridad la distribución de los datos, así que reduce el rango de datos para

graficar.

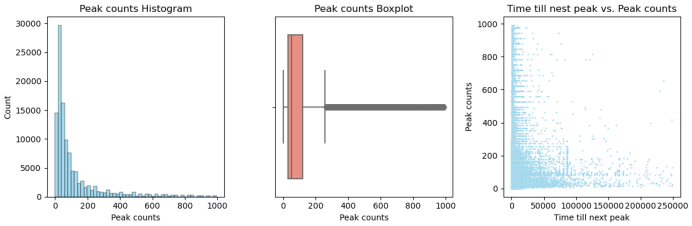

Figura 10: Histograms, Box Plot and Dispersión de Peak Counts

Para PEAK COUNTS, la mayoría de los registros se sitúan alrededor de 50, mientras que para TOTAL COUNTS, este valor se encuentra en torno a los 50000. Sin embargo, en ambas columnas, se observan numerosos datos atípicos que podrían clasificarse como outliers. Además, se nota una relación inversamente proporcional entre PEAK COUNTS y TOTAL COUNTS con TIME TILL NEXT PEAK.

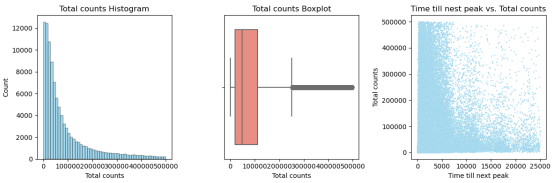

Figura 11: Histograms, Box Plot and Dispersión de Total Counts

Continuamos el análisis con la variable RADIAL, que representa la distancia radial en segundos de arco desde el centro del sol. Se observa una concentración significativa de datos alrededor de los 1000 arcosegundos (Figura 12), con una escasez de observaciones más allá de este punto. Esto sugiere que los instrumentos de medición podrían haber estado enfocados en una región específica, posiblemente el centro del Sol donde se encuentran las manchas solares. De hecho, al graficar únicamente los puntos con un radio menor a 1000 arcosegundos, se observa un patrón que se asemeja a una forma circular (Figura 13).

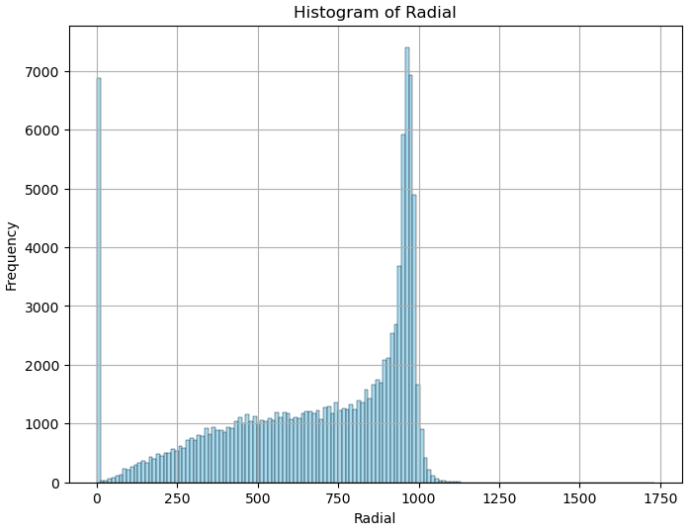

Figura 12: Histograma de Radial

Las regiones de mayor actividad solar se concentran en dos franjas centrales, con una mayor actividad en los extremos. Las ACTIVE REGION son un ID de áreas de intensa magnetización, origen de la mayoría de las llamaradas solares y eyecciones coronales de masa. Muchos datos tienen ACTIVE REGION = 0, lo que indica una falta de información.

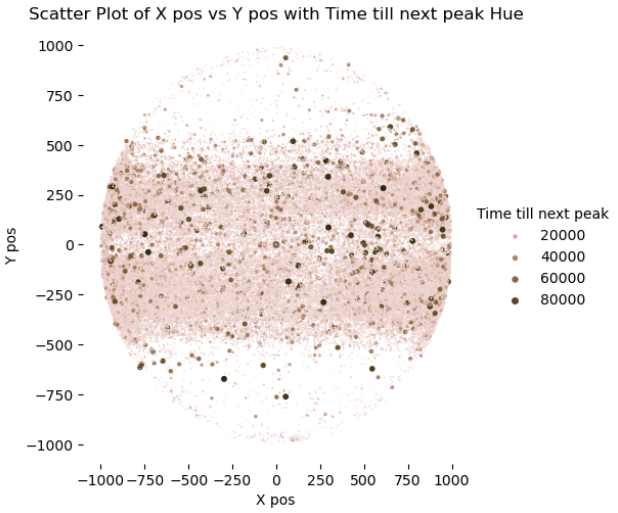

Figura 13: Regiones Activas del Sol

3. Preparación de los datos

Tratamiento de las variables.

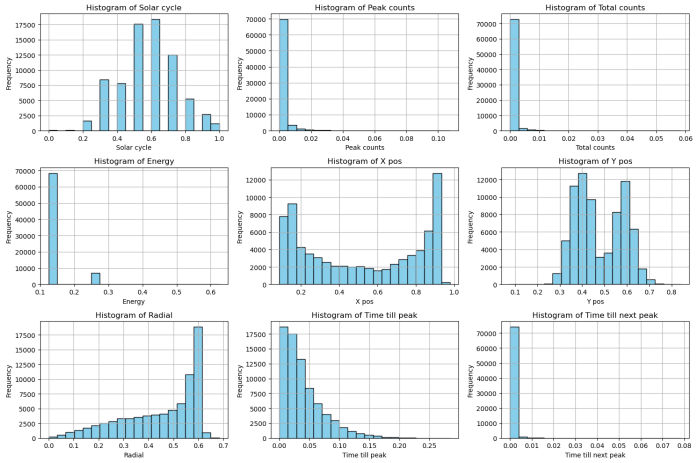

Figura 14: Distribución de datos elegidos

Eliminamos algunos datos con marcadores de mala calidad, descritos en la categoría de FLAGS, incluyendo aquellos en los que las mediciones fueron decimadas (dañadas) en la mayor parte (DF y DR), los eventos no solares (NS), la presencia de brechas de datos en las mediciones (GD, GE y GS), lo que se ha calificado como un posible destello solar en detectores frontales, pero sin posición (PS), la ocurrencia de un eclipse durante la medición (EE y ES), y momentos en los que la nave espacial estaba en SAA (SS, SD y SE). Esto nos deja con 92927 datos.

Se crearon dos nuevas variables: FLARE NUMBER, que es el número de llamaradas solares que ha habido hasta el momento en el día, y TIME TILL PEAK, que es el tiempo que toma para que nuestra llamarada solar llegue al pico.

Seleccionamos las siguientes variables para nuestro modelo: FLARE NUMBER, SOLAR CYCLE, PEAK COUNTS, TOTAL COUNTS, ENERGY, X POS, Y POS, RADIAL, PE, DURATION, TIME TILL PEAK y TIME TILL NEXT PEAK. Primero, codificamos la variable ENERGY para convertirla en una variable numérica y eliminamos los datos que están fuera del sol. Luego, aplicamos DBSCAN clustering, escalando primero los datos, para eliminar los valores outliers. Utilizamos esta técnica ya que muchos de nuestros datos no son normales. Nos quedamos con 75599 datos con la siguiente distribución.

Vemos que aún hay bastantes datos atípicos, sin embargo eliminar más afectaría la validez de nuestro modelo. Utilizamos una matriz de correlación parcial para identificar las variables que tienen más relación lineal con la variable que queremos predecir. Observamos que no hay ninguna variable que tenga una relación lineal monótona muy fuerte con TIME TILL NEXT PEAK además de FLARE NUMBER y SOLAR CYCLE. Sin embargo, como habíamos observado en nuestro análisis, la mayoría de nuestras relaciones no son lineales.

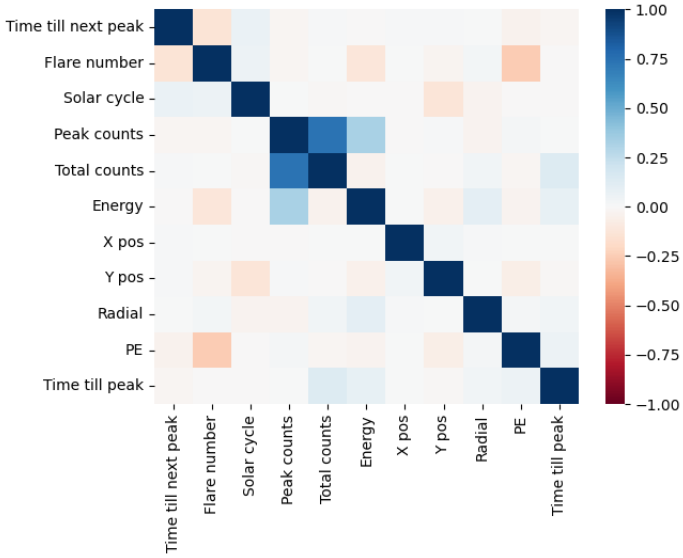

Figura 15: Mapa de correlación

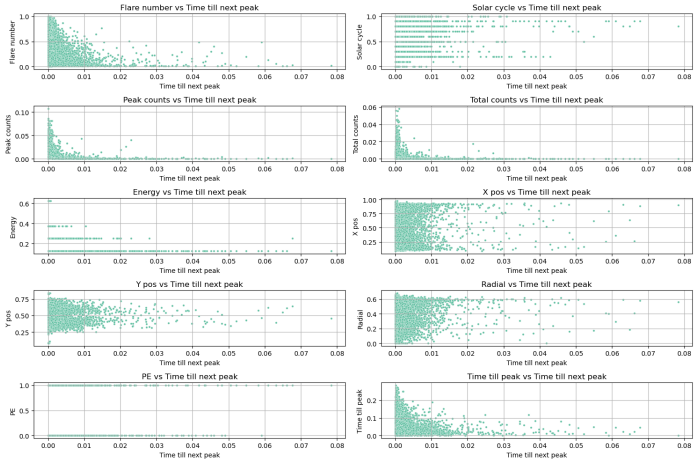

Figura 16: Gráficas de Dispersión

Nótese que para la mayoría de datos, hay una concentración de registros en puntos muy específicos. Por la gran cantidad de outliers que no eliminamos para no hacer de nuestro modelo inválido, las relaciones no se ven claramente.

4. Dimensiones de Calidad de Datos

Completitud

Al examinar los datos por primera vez, notamos numerosos valores faltantes en las columnas de FLAG N. Sin embargo, al profundizar en el análisis, determinamos que esta ausencia era más una rareza de conformidad que una verdadera falta de datos. Por otro lado, las columnas ACTIVE REGION y RADIAL contienen muchos valores en 0, lo cual indica datos faltantes. Consideramos pertinente dejar estos valores así, ya que estas columnas representan un espacio y marcarlos en el origen de los datos es adecuado (nótese que este valor también es cercano a la mediana de los datos relacionados, X POS y Y POS). En el caso de ACTIVE REGION, que representa un ID del lugar, no hay forma de reemplazar los valores, ya que un ACTIVE REGION en 0 podría indicar que ese lugar no ha sido declarado como mancha solar.

Conformidad

Mencionamos previamente que las columnas de FLAG N tenían un error de conformidad, donde los datos estaban guardados como listas a través de varias columnas. La última columna incluso contenía varias flags separadas por espacios, como si no hubiera suficiente espacio. Estos errores ya se corrigieron y los datos se almacenaron como columnas binarias o categóricas. También resolvimos un error de consistencia en las fechas, para que las llamaradas que comienzan en un día y terminan en otro se marquen correctamente.

Otros aspectos interesantes en relación con la conformidad incluyen las unidades en que se guardan algunas columnas, como los arcosegundos (una unidad común en astronomía pero rara en análisis de datos). Además, la energía se almacenó en rangos en lugar de como un dato continuo. Por último, la descripción de la columna ACTIVE REGION no es completamente clara y su explicación es insuficiente. No encontramos información más detallada sobre su significado.

Consistencia

Varios datos deben ser consistentes, como la diferencia entre START TIME y END TIME, que debe ser igual a DURATION, y se verificó que así sea. Al verificar la consistencia de las fechas, notamos que algunas no coinciden con sus números de ID, especialmente cuando la llamarada empieza a altas horas de la noche o madrugada. Utilizaremos la fecha almacenada en la columna DATE para mayor precisión. También verificamos que cuando X POS y Y POS son cero, RADIAL debe ser cero, y viceversa, lo cual se cumple.

Unicidad

Nuestros datos tienen un ID único. Sin embargo, notamos que tres de estos IDs se repiten, aunque los registros son completamente diferentes, lo cual probablemente sea un error tipográfico. Decidimos conservar estos registros a pesar de la duplicación y notamos que FLARE no es la columna más confiable.

Precisión

Con base en nuestra experiencia con los datos, hemos llegado a la conclusión de que la información es confiable y precisa. Aunque nuestra comprensión de la astronomía es limitada, hemos podido confirmar varios patrones del mundo real, como las manchas solares y el ciclo solar. Además, la entidad responsable de la publicación, la NASA, goza de una sólida reputación en el campo de la astronomía. No hemos detectado errores significativos en esta base de datos pública. Por lo tanto, concluimos que la información es lo suficientemente precisa para nuestros propósitos.

Integridad

Dado que solo contamos con una única base de datos y no existen relaciones con otras bases de datos, confirmamos que la dimensión de integridad se mantiene.

5. Modelación

Utilizamos tres algoritmos: Gradient Boosting, Random Forest y Decision Tree. Los resultados de los mejores modelos de estas tres técnicas son mostrados en el apartado siguiente. Se decidió utilizar la métrica R2, en consonancia con los objetivos de negocio. Nótese que algunos los modelos tienen R2 negativos, lo que indica que no explican bien la variabilidad de los datos.

5.0.1. Random Forest

Se realizaron 120 modelos con Random Forest, y los mejores hiperparámetros encontrados fueron 'n_estimators': 100, 'max_depth': 20, 'min_samples_split': 20, 'min_samples_leaf': 4, 'random_state': 42. El puntaje R-cuadrado (R2) en los datos de prueba fue positivo, con un valor de 0.06656528346617008, indicando una capacidad decente del modelo para explicar la variabilidad en los datos de tiempo hasta el próximo pico de llamarada solar.

Los resultados de las métricas de evaluación en los datos de prueba revelan que el Random Forest tiene un error cuadrático medio (MSE) de aproximadamente 77183599.69830665 segundos cuadrados, un error absoluto medio (MAE) de alrededor de 3178.242349896433 segundos, y un error absoluto porcentual medio (MAPE) de aproximadamente 372.17 %. La mediana del error cuadrático medio (MSE) es de alrededor de 2661282.319363131 segundos cuadrados, la mediana del error absoluto medio (MAE) es de aproximadamente 1631.3437035315228 segundos, y la mediana del error porcentual absoluto medio (MAPE) es de aproximadamente 136.60 %. Estos valores sugieren que el modelo tiene un comportamiento adecuado.

5.0.2. Gradient Boosting

Para Gradient Boosting, se entrenaron 120 modelos y se identificaron los mejores hiperparámetros como 'learning_rate': 0.1, 'max_depth': 10, 'min_samples_split': 2, 'min_samples_leaf': 2, 'random_state': 42. El mejor puntaje R-cuadrado (R2) en los datos de prueba fue negativo, con un valor de -0.005, que es bastante cercano a cero.

Los resultados de las métricas de evaluación en los datos de prueba muestran que el Gradient Boosting tiene un error cuadrático medio (MSE) de aproximadamente 83129135.54274525 segundos cuadrados, un error absoluto medio (MAE) de alrededor de 3288.7096888939686 segundos, y un error absoluto porcentual medio (MAPE) de aproximadamente 367.1353 %. La mediana del error cuadrático medio (MSE) es de alrededor de 2503785.2007315466 segundos cuadrados, la mediana del error absoluto medio (MAE) es de aproximadamente 1582.335359117549 segundos, y la mediana del error porcentual absoluto medio (MAPE) es de aproximadamente 124.11 %. Estos valores indican que el modelo tampoco logra capturar de manera precisa la relación entre las características de entrada y la variable objetivo.

5.0.3. Árbol de Decisión

Se entrenaron 120 modelos de Árbol de Decisión, y los mejores hiperparámetros identificados fueron 'max_depth': 10, 'min_samples_split': 2, 'min_samples_leaf': 12, 'random_state': 42. El mejor puntaje R-cuadrado (R2) en los datos de validación es 0.00160444215888067, un valor positivo, lo cual es buena indicación.

Los resultados de las métricas de evaluación en los datos de prueba revelan que el Árbol de Decisión tiene un error cuadrático medio (MSE) de aproximadamente 82555064.33607526 segundos cuadrados, un error absoluto medio (MAE) de alrededor de 3205.735170242685 segundos, y un error absoluto porcentual medio (MAPE) de aproximadamente 362.65 %. La mediana del error cuadrático medio (MSE) es de aproximadamente 2608156.810762765 segundos cuadrados, la mediana del error absoluto medio (MAE) es de aproximadamente 1614.9788835725676 segundos, la mediana del error porcentual absoluto medio (MAPE) es de aproximadamente 123.928 %.

6. Evaluación

Nótese que las métricas de la mediana superan en casi todo a las métricas de la media. Esto es debido a que la media es muy sensible a valores atípicos del conjunto de datos con el que trabajamos. Vemos que el mejor modelo es el de Árboles de decisión, según las métricas de la mediana. No hay un overfitting grave para estos valores. La diferencia porcentual entre los valores del primer modelo con el modelo de evaluación son las siguientes:

- La diferencia para MSE es: 11.6 %
- La diferencia para MAPE es: 1.81 %
- La diferencia para MAE es: 2.60 %
- La diferencia para MSE-mediano es: 2.00 %
- La diferencia para MAPE-mediano es: 1.00 %
- La diferencia para MAPE-mediano es: 0.12 %

Por lo tanto, el modelo se generaliza bien. Aunque las métricas muestran que el modelo tiene errores significativos, la distribución de estos errores (Figura 17, 18) revela que hay muchos datos con errores pequeños y pocos datos con errores grandes. Esto indica que, en general, el modelo es bueno, aunque tiene varios datos alejados de la media, similares a outliers. Por esta razón las métricas relacionadas a la media nos indicaban que nuestro modelo tenía mucho error.

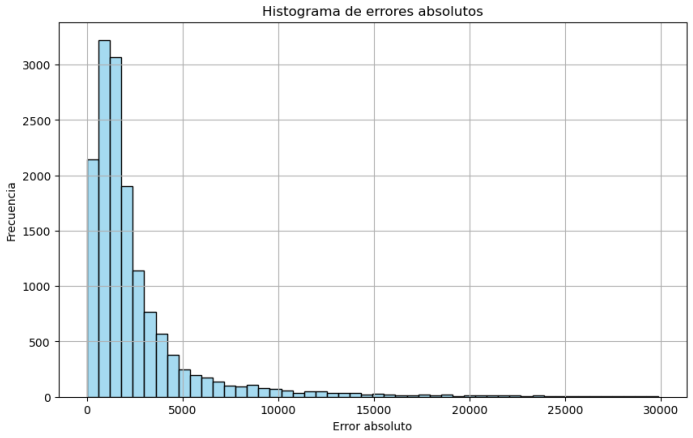

Figura 17: Histograma de Errores Absolutos, para errores menores de 20000

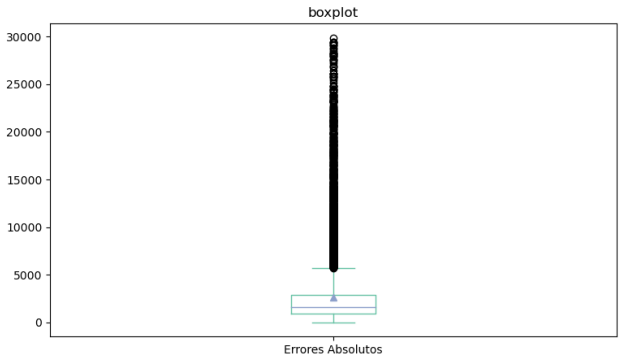

Figura 18: Boxplot de Errores Absolutos

Observamos en la Figura 19 que el modelo tiende a predecir peor los tiempos de los datos con energía 12-25 keV. Hay muy poquitos datos de estos valores en el conjunto de entrenamiento. Nótese que el modelo no contiene registros para una energía de 25-50 keV. Estos probablemente fueron eliminados con los outliers. Es importante señalar que el eje vertical de

la gráfica comienza en 1400. Para mejorar la precisión de las predicciones en este rango de energía, se requerirían más datos de llamaradas de alta energía.

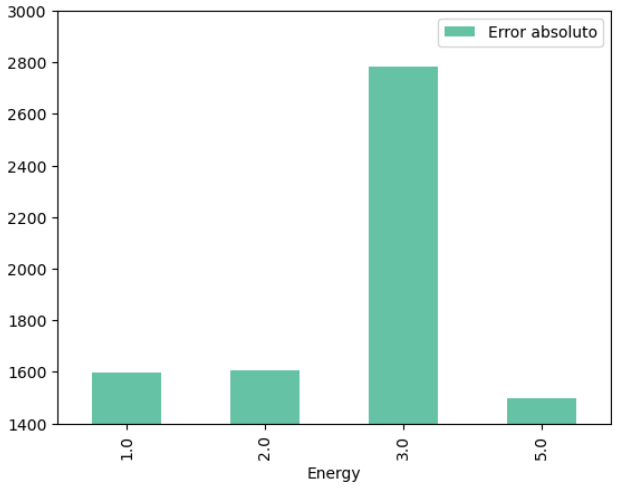

Figura 19: Error Absoluto por Energía

Además, observamos que el modelo realiza predicciones más precisas para las llamaradas que ocurren en la mitad del ciclo solar (Figura 20). Este fenómeno se debe a un desequilibrio en la cantidad de datos disponibles para entrenar el modelo en diferentes fases del ciclo solar. Algunos ciclos solares tienen una gran cantidad de datos de entrenamiento, mientras que otros tienen una cantidad mucho menor.

Como resultado, el modelo está sesgado hacia las condiciones más comunes en el conjunto de datos de entrenamiento, lo que lo hace parcial en su capacidad predictiva. Para abordar esta parcialidad, se requiere recopilar más datos de llamaradas solares en las fases extremas de los ciclos solares, asegurando así una representación más equitativa de todas las fases del ciclo solar en el conjunto de datos de entrenamiento.

Para el resto de variables se identifica un patrón similar, entre menos datos se tenga, mayor es el error.

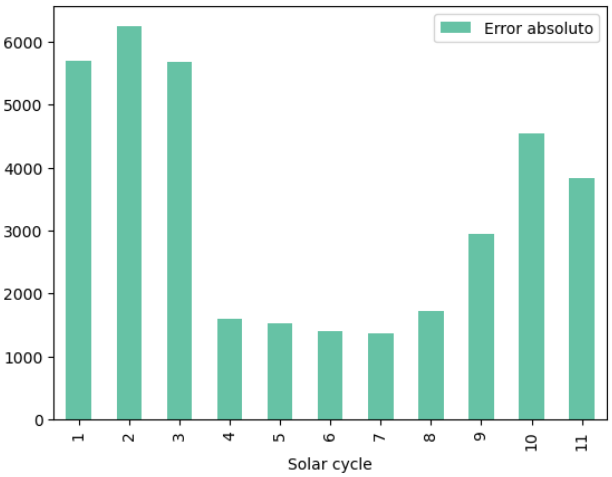

Figura 20: Error absoluto por Año del Ciclo Solar

Después de haber modelado, veremos, de nuevo cuáles han sido las variables más importantes (Figura 21). Vemos que estas resultaron ser TOTAL COUNTS y Y POS, con otras variables contribuyendo también medianamente al modelo.

Luego, procederemos a calcular los intervalos de confianza al 95 % para el MAE, MSE y MAPE, y encontramos los siguientes resultados:

- IC para el MAE es: (3021.1267791192276, 3274.259943656175)
- IC para el MSE es: (50660437.81515871, 95237823.53689645)
- IC para el MAPE es: (3.5952595356476698, 3.846701780386998)

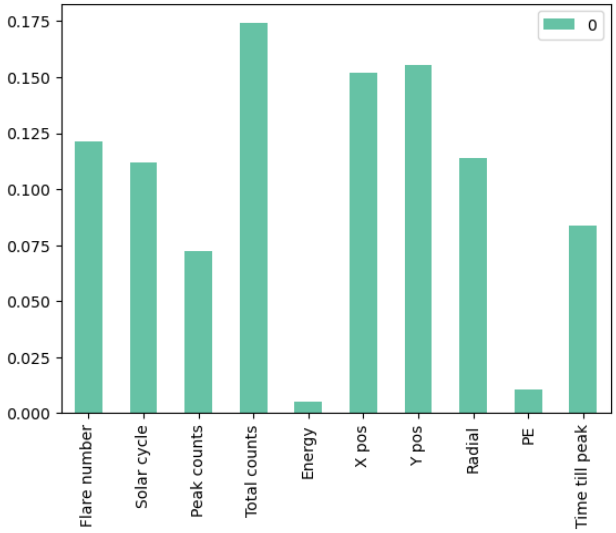

Figura 21: Importancia de variables

Normalidad de los datos Durante todo el análisis, hemos observado que los datos no siguen una distribución normal. Realizamos una prueba de Kolmogórov-Smirnov y, dado que el valor p ≈ 0, se rechaza la hipótesis nula de que los residuos siguen una distribución normal (Figura 22).

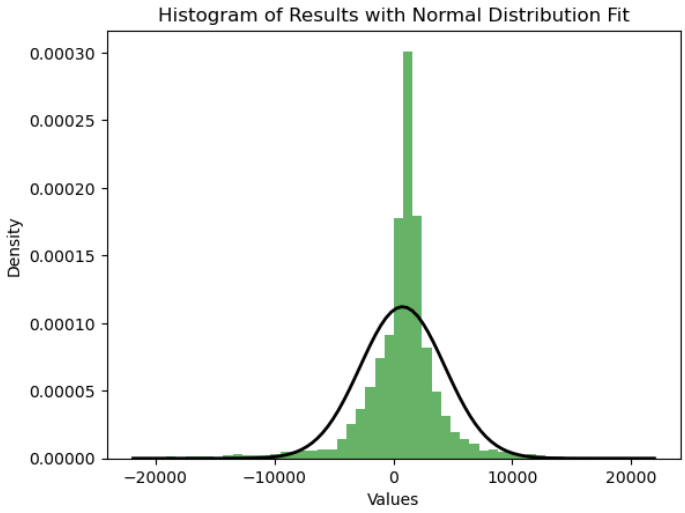

Figura 22: Histograma de Errores, para errores menores de 20000, con ajuste de curva normal

Comparación con la media simple

Procederemos a comparar al modelo propuesto con un enfoque de línea base, representado por la media simple de la variable de salida del modelo. Para esto, se utilizó el enfoque de A/B testing, donde se evaluó si el modelo propuesto superaba significativamente la media simple en términos de error absoluto.

Se calculó la media simple de la variable de salida en el conjunto de validación (y_train_test), obteniendo un valor de 3112,00. Luego, se realizó la predicción tanto del mejor modelo obtenido como de la media simple sobre el conjunto de prueba (x_val). El modelo propuesto arrojó un error absoluto medio (MAE) de 3147,69, mientras que la media simple obtuvo un MAE de 3226,71.

Para determinar la significancia de esta diferencia, se empleó el test de Wilcoxon, un enfoque no paramétrico adecuado para comparar las medias de dos poblaciones. Con un p-valor de 7.934258748868549e-233, se encontró una diferencia altamente significativa entre los dos enfoques. Este resultado proporciona una fuerte evidencia, con un nivel de confianza del 99.9999999999999 %, de que el modelo propuesto es superior a la media simple en términos de error absoluto.

Además de esta comparación puntual, se utilizó una validación cruzada repetida (Figura **??**) para obtener una estimación más robusta del error absoluto medio y sus intervalos de confianza. Se empleó un enfoque de validación cruzada con 10 folds repetidos entre 1 y 10 veces. Los resultados mostraron una tendencia consistente: el error absoluto medio se estimó en 13170.04 con un intervalo de confianza del 95 % entre (3113.72, 3226.352). Esto refuerza la confianza en la robustez y consistencia del modelo propuesto.

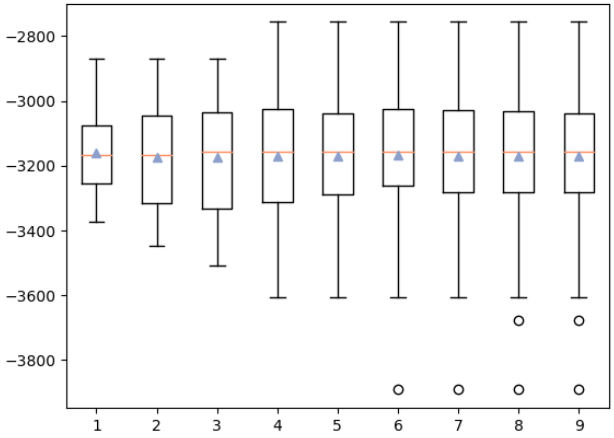

Figura 23: Repeated k-folds cross validation

En resumen, tanto el análisis comparativo inicial como la validación cruzada repetida respaldan la superioridad del modelo propuesto sobre la media simple en términos de precisión predictiva, destacando su potencial para aplicaciones futuras. Hemos concluido que nuestro modelo cumplen con los objetivos de negocio, ya que es una mejor predicción que la media simple .

7. Despliegue

Para el despliegue se utilizaron varias librerías, **SunPy** para las gráficas específicas del sol, **astropy** para las ubicaciones exactas de la posición donde se quiere predecir la llamarada, y finalmente **ipywidgets** y **Ipython** para poder hacer el despliegue en un notebook y no como aplicación aparte.

Para utilizar el modelo solo se deben rellenar los valores a predecir y de manera automática se genera la predicción.

Referencias

Exbrayat, J. (1971). *Historia de Montería*. Imprenta Departamental de Córdoba.

Incze, R. (2021). *The Cost of Machine Learning Projects*. Medium. URL: `https://medium.com/cognifeed/the-cost-of-machine-learning-projects-7ca3aea03a5c`.

A. Códigos de Calidad de cada Destello:

- **a0** - En estado de atenuador 0 (Ninguno) en algún momento durante el destello
- **a1** - En estado de atenuador 1 (Delgado) en algún momento durante el destello
- **a2** - En estado de atenuador 2 (Grueso) en algún momento durante el destello
- **a3** - En estado de atenuador 3 (Ambos) en algún momento durante el destello
- **An** - Estado del atenuador (0=Ninguno, 1=Delgado, 2=Grueso, 3=Ambos) en el pico del destello
- **DF** - Las mediciones del segmento frontal fueron decimadas en algún momento durante el destello
- **DR** - Las mediciones del segmento trasero fueron decimadas en algún momento durante el destello
- **ED** - Eclipse del vehículo espacial (noche) en algún momento durante el destello
- **EE** - El destello terminó en el eclipse del vehículo espacial (noche)
- **ES** - El destello comenzó en el eclipse del vehículo espacial (noche)
- **FE** - Destello en curso al final del archivo

- **FR** - En Modo de Tasa Rápida
- **FS** - Destello en curso al comienzo del archivo
- **GD** - Brecha de datos durante el destello
- **GE** - El destello terminó en una brecha de datos
- **GS** - El destello comenzó en una brecha de datos
- **MR** - Nave espacial en zona de altas latitudes durante el destello
- **NS** - Evento no solar
- **PE** - Evento de partículas: Partículas presentes
- **PS** - Posible destello solar; en detectores frontales, pero sin posición
- **Pn** - Calidad de la posición: P0 = Posición no válida, P1 = Posición válida
- **Qn** - Calidad de los Datos: Q0 = Calidad más Alta, Q11 = Calidad más Baja
- **SD** - La Nave Espacial Estaba en SAA en Algún Momento Durante el Destello
- **SE** - El Destello Terminó Cuando la Nave Espacial Estaba en SAA
- **SS** - El Destello Comenzó Cuando la Nave Espacial Estaba en SAA

B. Diccionario de datos

Columna	Tipo de dato	Tamaño máximo del dato	Descripción
Flare	int64	28	Un número de identificación, (y)ymmddnn, por ejemplo, 2021213 es el treceavo destello encontrado el 12 de febrero de 2002.
Date	datetime64[ns]	59	La fecha en que ocurrió el destello
Start time	datetime64[ns]	57	Hora de inicio del destello
Peak time	datetime64[ns]	57	Hora pico del destello
End time	datetime64[ns]	57	Hora de finalización del destello
Duration	int64	28	Duración del destello en segundos
Peak counts	int64	28	Tasa de conteos o picos de radiación por segundo detectados en un rango de energía de 6 a 12 kiloelectronvoltios (keV) durante la duración de la llamarada solar, promediados sobre todos los detectores, incluido el fondo
Total Counts	float64	24	Número total de conteos o picos de radiación detectados un rango de energía de 6 a 12 kiloelectronvoltios (keV) durante la duración de la llamarada solar, sumados sobre todos los detectores, incluido el fondo
Energy	object	59	La banda de energía en KeV más alta en la que se observó el destello.
X pos	int64	28	Posición del destello en segundos de arco desde el centro del sol
Y pos	int64	28	Posición del destello en segundos de arco desde el centro del sol
Radial	int64	28	Distancia radial en segundos de arco desde el centro del sol
Active region	int64	28	Número para la región activa más cercana, si esta disponible
Flags	object	69	Códigos de calidad de la medición, sin ningún orden en particular

Cuadro 1: Descripción de las columnas y sus propiedades

Base de datos

https://www.kaggle.com/code/jmquintana/solar-flares-from-rhessi-mission

Implementación

https://github.com/WonderfulAme/solar-flares/