Primera entrega de proyecto

POR:

Daniel Esteban Sánchez Marín

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2022

CONTENIDO

INTRODUCCIÓN

En la actualidad hay muchos ámbitos y tecnologías que han estado sobresaliendo y dando mucho de qué hablar, uno de ellos es la inteligencia artificial ya que a través de los años ha logrado permanecer en la mira de muchos ingenieros y especialistas del campo y poco a poco se ha convertido en una de las herramientas más poderosas en el mundo de la ingeniería y tecnología en general.

En este proyecto haremos uso de los algoritmos de machine learning para crear un modelo que nos ayude a predecir el déficit entre la energía generada por medio de combustibles fósiles y varias fuentes renovables, para el país de España.

1. PLANTEAMIENTO DEL PROBLEMA

El suministro de electricidad juega un papel importante en el sustento de los ciudadanos de un país. La electricidad, entre otras cosas, nos ayuda a estar conectados, calentarnos y alimentar a nuestras familias. Por lo tanto, es necesario mantener las luces encendidas para mantener y mejorar el nivel de vida mediante la inversión en infraestructura eléctrica. Sin embargo, en los últimos años, ha habido evidencia de que el uso de fuentes puramente no renovables no es sostenible. Se desea desarrollar un modelo que prediga el déficit entre la energía generada por medio de combustibles fósiles y varias fuentes renovables, para el país de España

2. DATASET

Voy a usar el dataset de Kaggle de esta competición:
https://www.kaggle.com/competitions/edsa-individual-electricity-shortfall-challenge/overview/description
En ella nos proporcionan datos que contienen información sobre las condiciones climáticas en varias ciudades españolas para el período 2015-2017. El conjunto de datos también tiene información sobre los tres déficits de carga por hora para el mismo período. En el contexto de este problema, el déficit de carga de tres horas es la diferencia entre la energía generada mediante combustibles fósiles y fuentes renovables.

Este dataset cuenta con 2 archivos que servirán para el entrenamiento de algoritmos y para las pruebas, llamados **df_test.csv** y **df_train.csv.** El primero cuenta con 48 columnas y el segundo con 49, su principales columnas son:

- **time:** Hora en la que se registraron los datos
- **{City Name}_wind_speed:** la velocidad del viento en un intervalo de tiempo específico para la ciudad respectiva.
- **{City Name}_wind_degree:** la fuerza del viento para la respectiva ciudad en un intervalo de tiempo específico.
- **{City Name}_rain_1h:** una métrica que expresa la cantidad de lluvia que ha caído en la última hora en una ciudad en particular.
- **{City Name}_rain_3h**: una métrica que expresa la cantidad de lluvia que ha caído en las últimas tres horas en una ciudad en particular.

- **{Nombre de la ciudad}_humedad:** el nivel de humedad medido en el momento definido para la ciudad específica mencionada.
- **{City Name}_clouds_all:** el nivel de cobertura de nubes medido en el momento especificado para la ciudad específica mencionada.
- **{City Name}_pression:** la presión atmosférica de la ciudad nombrada en un intervalo de tiempo específico.
- **{City Name}_snow_3h:** una métrica que expresa la cantidad de nieve que ha caído en las últimas tres horas en una ciudad en particular.
- **{City Name}_weather_id:** una métrica utilizada para explicar las condiciones climáticas de una ciudad específica en un momento específico.
- **{City Name}_temp_max:** la temperatura máxima para una ciudad específica en un momento dado.
- **{City Name}_temp_min:** la temperatura mínima para una ciudad específica en un momento dado.
- **{City Name}_temp:** la temperatura promedio de una ciudad específica en un momento dado.

También hay un archivo llamado **sample_submission_load_shortfall (1).csv** el cual contiene la siguiente información:

- **time –** Hora en la que se registraron los datos.
- **load_shortfall_3h –** La diferencia entre la energía generada por el método de fuentes de energía renovables, como solar, eólica, geotérmica, etc., y la energía generada con combustibles fósiles, dividida en ventanas de tres horas.

3. MÉTRICAS

La métrica de evaluación para esta competencia es Root Mean Square Error. El error cuadrático medio (RMSE) se usa comúnmente en el análisis de regresión y el pronóstico y mide la desviación estándar de los residuos que surgen entre los valores observados previstos y reales para un proceso de modelado.

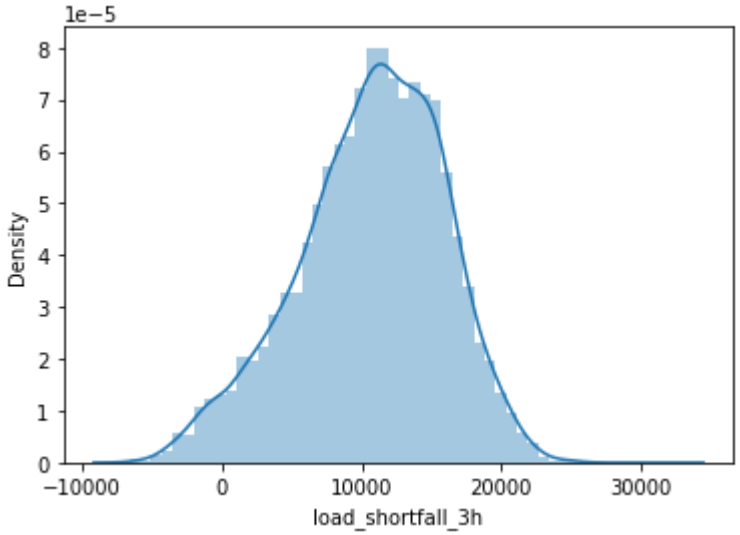
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2}$$

En cuanto a la métrica de negocio, se espera que las predicciones hechas por este modelo sean lo suficientemente convincentes para que el gobierno español se decida a realizar mayores inversiones en la infraestructura de recursos de energías renovables.

3.1 Variable Objetivo

La variable objetivo de este trabajo es “**load_shortfall_3h”** la cual nos da la diferencia entre la energía generada por el método de fuentes de energía renovables, como solar, eólica, geotérmica, etc., y la energía generada con combustibles fósiles, dividida en ventanas de tres horas.

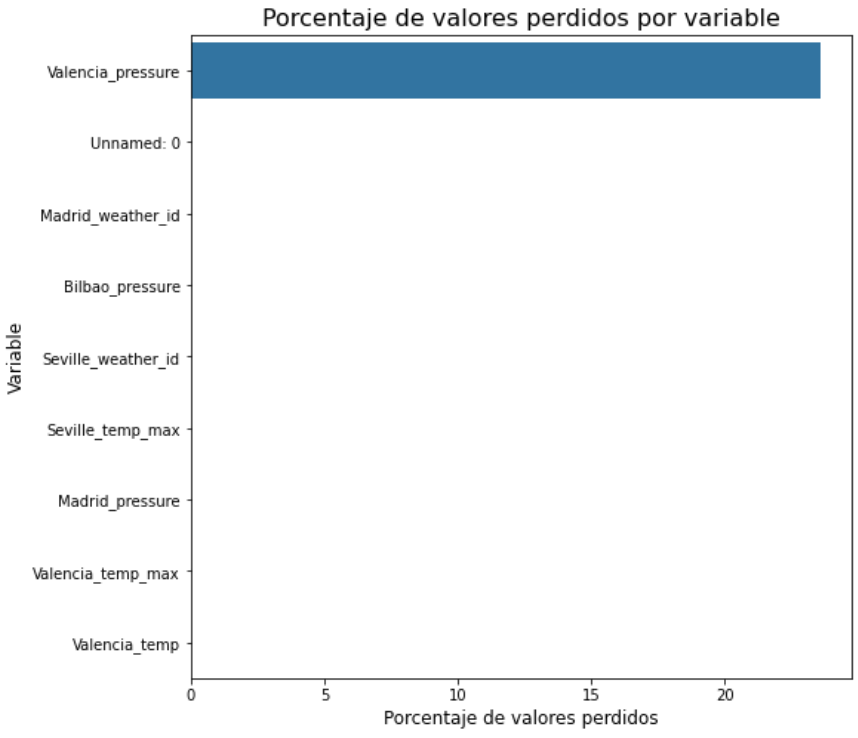
3.2. Análisis de la variable objetivo (load_shortfall_3h)



Con el gráfico mostrado podemos observar que la variable objetivo en general presenta una simetría bastante notable con una leve inclinación hacia la izquierda, además de esto queda en evidencia que el valor más recurrente en la variable objetivo es el 10000.

3.3. Datos faltantes

En el dataset que estamos utilizando para este proyecto se puede observar que Valencia_pressure es la única variable que presenta datos faltantes con un porcentaje del 23.6%



4. TRATAMIENTO DE DATOS

4.1. Eliminación de las columnas con muchos datos faltantes

Del análisis exploratorio de las variables encontramos que la única variable con datos faltantes es Valencia_pressure, por lo que procederemos a eliminarla.

4.2. Transformación de variables categóricas

Las variables categóricas, como lo son en este caso "Valencia_wind_deg" y "Seville_pressure" pueden ser útiles a la hora de realizar el análisis, sin embargo, no pueden usadas en la forma categórica, por lo que se deben convertir en variables numéricas que si podamos utilizar para entrenar un modelo.

5. RETOS Y CONDICIONES DE DESPLIEGUE DEL MODELO

Para que el desempeño del modelo pueda considerarse exitoso los datos de la energía generada por fuentes renovables deben ser mayor o por lo menos equiparable a la generada a través de combustibles fósiles. Si este es el caso podría decirse que el modelo es completamente apto para llevar a producción.

6. DESEMPEÑO

Lo que se espera de este modelo es predecir el déficit entre la energía generada por medio de combustibles fósiles y varias fuentes renovables para el país de España. Con los datos obtenidos se espera realizar análisis más acertados y confiables con respecto al uso de energías renovables y poder determinar qué tan rentable es para el gobierno español dirigir una cantidad mayor de recursos a la expansión y manejo de este tipo de fuentes de energía.

7. CONCLUSIONES

- Se recomienda aumentar la complejidad de algunos modelos ya que varios de ellos presentan problemas de overfitting.

8. BIBLIOGRAFÍA

- EDSA Individual | Electricity Shortfall Challenge | Kaggle. (2022). Retrieved 05 July 2022, from https://www.kaggle.com/competitions/edsa-individual-electricity-shortfall-challenge/overview/description