SUN POWER PREDICTION

Predicción de la producción eléctrica de plantas fotovoltaicas en función de los parámetros meteorológicos.

Aprendizaje supervisado
Regresión

ABRAHAN SANTANA NARANJO

1. INTRODUCCIÓN

La importancia de la predicción de la producción eléctrica, es uno de los factores más importantes en las compañías del sector. Éste hecho cobra, aún más, mayor relevancia en nuestro país debido a cómo está estructurado el mercado de la energía.

El sistema eléctrico español, y las leyes que lo regulan, obligan a que diariamente se negocie el precio de la venta del kWh del día siguiente. Este mercado es de tipo marginalista, de tal manera que cada una de las centrales eléctricas deberán negociar su precio por separado, independientemente de quien las empresas que las operen o de las agentes que la compre.

Dicho precio, se subasta de una forma un tanto peculiar, cada central debe presentar cuál cree que va a ser su producción por horas para el día siguiente y cuál es su oferta de venta de dicha producción en esa hora determinada. Una vez las empresas hayan presentado sus ofertas, están son agregadas y ordenadas de forma ascendente, conformando una curva de oferta de mercado por horas.

Por otro lado, los agentes compradores deberán indicar cuál es su oferta de compra por horas, las cuales una vez realizadas, se agregarán y ordenarán de forma descendente, creando una curva de compra de mercado por horas. Así, se aseguran de que sus consumidores tengan la potencia cubierta que van a demandar en dicho margen.

Una vez establecidas dichas curvas, se fijará el precio utilizándose el punto de intersección de ambas, de tal manera que aquellas centrales que ofrezcan su producción por encima del precio se quedarán sin conseguir vender su producción, y aquellos agentes que quieran comprar la producción por debajo del precio, se quedarán sin poder comprar su producción. Éste sistema recibe el nombre de mercado eléctrico MIBEL, el cuál queda mejor ilustrado en la siguiente imagen:

Fuente: Escuela de organización industrial

Esta previsión de producción, es fácilmente controlable en las centrales convencionales (combustibles fósiles, nucleares, etc.) o en las centrales renovables de producción controlada (centrales hidráulicas). Pero difícilmente, calculable en las centrales de energías renovables cuyo origen dependen de la meteorología como son las centrales eólicas o fotovoltaicas.

De ahí parte la motivación de la realización de dicho proyecto, la forma convencional de calcular cuanta energía se va a producir es mediante la realización de modelos matemáticos que requieren de personal experto, como matemáticos o físicos, que se encarguen de realizar predicciones de la misma. El mercado, penaliza a aquellas empresas que se comprometan a realizar una producción y que luego no la cumplan, por tanto, dichos expertos suelen utilizar modelos más conservadores que siempre arrojen datos inferiores a la producción real que pueden llegar a alcanzar.

Lo que se busca con este modelo, es reducir lo máximo posible el margen entre la producción comprometida por el mercado y la producción real que se pueda llegar a alcanzar, aportando mayores ingresos a los agentes productores, pero sin llegar a ser excesivamente optimistas para evitar las penalizaciones por incumplimiento de producción.

2. ORIGEN DE LOS DATOS

Para la realización de dicho proyecto, nos hemos apoyado en dos fuentes de datos:

1. **Datos para entrenamiento**. Para entrenar nuestro modelo y lograr encontrar el más acorde a nuestro cometido, nos hemos aprovechado del Challenge propuesto por la compañía eléctrica EDP. La cuál buscaba encontrar un modelo que realizara dicha predicción, haciendo uso de los datos recogidos por una de sus plantas ubicada en Portugal. Los datos abarcaban los años completos comprendidos entre 2014 y 2017, tomados con una frecuencia aproximada, pero variable, de 5 minutos.

   Dentro de estos datos podemos encontrar parámetros meteorológicos recogidos in situ y de forma continuada, acompañados de las producciones eléctricas suministradas por dicha planta en ese mismo momento. Se pueden encontrar información de la producción de 2 tipos de placas diferentes dispuestas en distinto ángulo.

   Estos datos pueden ser descargados a través del siguiente link: https://opendata.edp.com/pages/challenges/evaluation#description

2. **Datos destinados al modelo de producción**. Para el modelo de puesta en producción, se tomarán los datos de la Agencia Estatal de Meteorología de España, de donde obtendremos su predicción de las condiciones ambientales, y nutriremos a nuestro modelo para obtener las predicciones de producción.

   Se ha optado por este recurso, ya que es la institución con la mayor red de estaciones meteorológicas del país y, por tanto, nos puede aportar una gran información acerca de las condiciones que existirán al día siguiente. Y a partir de ahí, podemos determinar la

producción eléctrica por horas facilitando la negociación del precio de la producción. De tal manera, que dicha producción no se quede fuera del mercado, a la ves que las compañías no se queden cortas en la producción.

3. DESCRIPCIÓN DE LOS DATOS

Los datos utilizados para el entramiento del modelo, presentaban una serie de problemas los cuales debían ser subsanados antes de ser usados para dicho fin.

En primer lugar, nos encontrábamos con una gran cantidad de datos nulos o inexistentes a ciertas horas. Esto puede estar provocado por un mal funcionamiento de los instrumentos de medida, lo cual nos llevó al debate de sí prescindir de los mismos o intentar recuperarlos de alguna forma.

Además, muchos de ellos eran datos atípicos, es decir, eran datos considerados "Outlier". Como ejemplo, en muchas ocasiones nos encontrábamos que la temperatura medida en ciertas horas, superaban los 5 millones de grados centígrados, algo totalmente desorbitado para la superficie de nuestro planeta. Es por ello, que otro de los problemas que se nos presentaba era la estimación de los mismos en dicho punto, ya que se intentaban por cualquier medio evitar eliminarlos de nuestro dataset.

Por último, y debido a la alta cantidad de variables suministradas, se tenía que encontrar cuales eran las más relevantes para la producción de energía y como encarar el entrenamiento del modelo con las mismas. Tampoco podíamos olvidar el carácter estacional de los datos, ya que dichas medidas habían sido tomadas, y por tanto ofrecidas, con la hora y fecha de captura, provocando como era lógico una repetición de comportamiento en las distintas estaciones del año.

4. FASES DEL PROYECTO

El proyecto se dividió en 5 fases determinantes:

1. **Fase 1: preparación de entorno.** La primera fase, consiste en la creación de un algoritmo que preparase el entorno de trabajo para la correcta ejecución del proyecto. En el, se descargan los datos originales utilizados para el entramiento, y la instalación de las librerias necesarias para el desarrollo del mismo. Se debe mencionar que se ha utilizado el entorno de Jupyter, y como lenguaje de programación Python.

   Esta fase ha sido ejecutada en el notebook "00_Preparing_Environment.ipynb".

2. **Fase 2: reparación y limpieza de datos.** Durante esta fase, se procedió a la visualización y arreglo de los datos. Aquí se descubrió que existía un gran número de elementos NaN en las distintas variables, y se tuvo que tomar la decisión de cómo proceder ante ellos. Como última opción siempre se dejó el hecho de eliminar dichas filas, es por tanto que se realizó una búsqueda de los mismo y se ejecutó una interpolación sobre ellos de los datos anterior y posterior, ya que, al ser datos continuos de parámetros meteorológicos, dichos datos se debían encontrar en medio del dato anterior y del posterior. Por ejemplo, si a una hora se toma un dato de temperatura de 20ºC, 5 minutos después se toma un dato NaN, y posteriormente un dato de 22ºC, es lógico pensar que la temperatura que se debía haber registrado en el NaN debía estar alrededor de los 21ºC.

   Con el mismo criterio, se procedió a la sustitución de los datos atípicos. Se buscaron según la norma de que si, dicho dato superaba la media de la columna en 3 veces la desviación típica, se podía considerar un dato atípico. A partir de ahí, se actuó de la misma forma que con los datos NaN, se cambiada dicho dato por una interpolación entre el anterior y el posterior.

   A parte del tratamiento de los datos, se procedió al formateo de la fecha, ya que no estaba en unidades de UTC +0, a la concatenación de los diferentes años de forma correlativa, y la unión de los datos meteorológicos tomados con los datos de producción ya que estos se encontraban en archivos diferentes. Realizada la unión, se procedió a la separación en datos de entrenamiento y datos de test.

   Esta fase ha sido ejecutada en el notebook: “01_Fixing_data.ipynb”

3. **Fase 3: creación del modelo de predicción.** En la tercera fase, se continuo con la búsqueda el modelo de regresión más adecuada para tal fin. Para ello, buscamos entrenar dos modelos principales: Regresión lineal y árbol de decisión.

   Inicialmente, se probó con un modelo de regresión lineal simple. Tal y como se presentan los datos y el margen tan grande de los mismo, era de prever que no sería el mejor modelo, pero igualmente era el modelo básico a seguir como referente.

   Se evitó en todo momento la presentación de información que pudiera dar un componente de estacionalidad al modelo, evitando de esta forma que replicara un comportamiento inadecuado por patrón. Para ello, no se introdujo en ningún momento la variable de fecha y hora en el modelo.

   Como métrica de estudio, se utilizó el MAE, el cual nos proporcionaba la principal ventaja de que se penalizan los grandes errores, haciéndola robusta frente a los valores atípicos que previsiblemente nos íbamos a encontrar en los datos de entrenamiento, aun habiendo corregido algunos, sabíamos que seguían existiendo.

   Esta fase fue ejecutada en el notebook “02_Creating_Model.ipynb”.

4. **Fase 4: Scraping de datos de la AEMET.** En esta fase, se procedió al desarrollo de un algoritmo que se conectara a los servidores de la AEMET y que recogiera de los mismos aquellos datos de predicción de parámetros meteorológicos de las distintas estaciones repartidas por todo el territorio nacional.

   Para ello, la agencia creó una API que facilitaba la conexión a los mismos, permitiendo descargar los informes deseados. Después de descargar dichos informes, los cuales venían en formato de texto plano, se procedía a la conversión de dicho formato a un DataFrame.

   Al igual que los datos de entrenamiento, estos datos presentaban información perdida (NaN), lo que nos obligaba a seguir el mismo procedimiento que en la fase 2, corrigiendo dicha información mediante la interpolación de los datos existentes.

   Tras la limpieza y corrección de los datos descargados, nuestro cometido se centraba en la predicción de la producción eléctrica mediante el paso de dichos datos por los modelos guardados de la fase anterior. Una vez obtenidos los resultados, se creaba una tabla con toda la información y se guardaba tanto a nivel local como en la nube, usando la plataforma Dropbox.

5. **Fase 5: visualización de los datos.** En esta última fase, se usó la herramienta Tableau para crear un dashboard que permitiera visualizar la información meteorológica aportada por la agencia de meteorología y las predicciones realizadas por el modelo final seleccionado.

   Para ello, se simuló que las propias estaciones meteorológicas, eran centrales fotovoltaicas, lo cual nos permitió crear un prototipo de como quedaría el SW final en producción.

   En el nos encontraríamos con 3 dashboards diferentes:

   - Un dashboard con la información recogida de la API de la AEMET
   - Un dashboard con la producción predicha por el modelo
   - Un dashboard con la información detallada de cada central.
   -

5. METODOLOGÍA

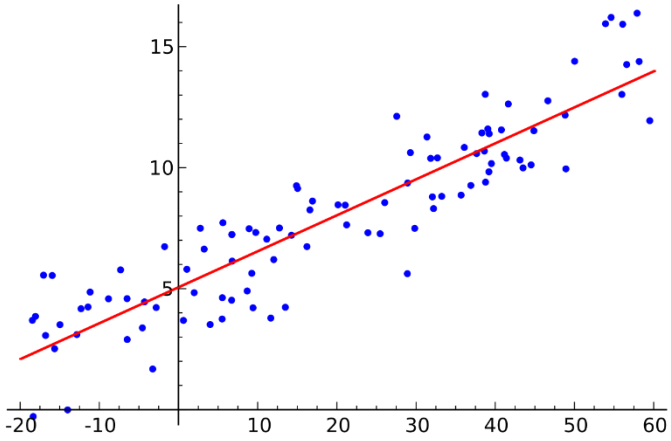
Debido a que el objetivo de este proyecto consiste en la predicción de los parámetros de producción de una planta fotovoltaica, y tenemos un dataset inicial de una planta real, se ha optado por la utilización de dos técnicas de predicción, regresión lineal y árbol de decisión. Para ambas técnicas, se entrenaron 3 modelos: uno para la predicción de la producción de tensión (VDC), otro para la predicción de la producción de corriente (CDC) y otro para la producción de potencia (PDC).

Inicialmente, se calculó la tabla de los coeficientes de correlación de Pearson, la cual nos arrojó información sobre la relación existente entre todas las variables, incluyendo las de entrada y las de salida. Podemos ver los resultados en la siguiente imagen (también la encontrarán en el notebook “02_Creating_model.ipnb”):

	Ambient_Temperature	Global_Radiation	Diffuse_Radiation	Ultraviolet	Wind_Velocity	Wind_direction	A_Optimal_VDC	A_Optimal_CDC	A_Optimal_PDC
Ambient_Temperature	1.000000	0.429854	0.068290	0.431713	0.136022	-0.031117	-0.166174	0.168704	0.144597
Global_Radiation	0.429854	1.000000	0.341474	0.987963	0.325380	-0.106952	0.163424	0.821643	0.814852
Diffuse_Radiation	0.068290	0.341474	1.000000	0.365796	0.218479	-0.052161	0.230912	0.286148	0.292225
Ultraviolet	0.431713	0.987963	0.365796	1.000000	0.341061	-0.092889	0.152905	0.802804	0.794839
Wind_Velocity	0.136022	0.325380	0.218479	0.341061	1.000000	0.156139	0.156417	0.275937	0.283009
Wind_direction	-0.031117	-0.106952	-0.052161	-0.092889	0.156139	1.000000	-0.029088	-0.120653	-0.121869
A_Optimal_VDC	-0.166174	0.163424	0.230912	0.152905	0.156417	-0.029088	1.000000	0.214218	0.255866
A_Optimal_CDC	0.168704	0.821643	0.286148	0.802804	0.275937	-0.120653	0.214218	1.000000	0.997673
A_Optimal_PDC	0.144597	0.814852	0.292225	0.794839	0.283009	-0.121869	0.255866	0.997673	1.000000

Ampliando la imagen, se puede comprobar como las variables que más afectan o más relacionadas están con las salidas, son las variables: Radiación global, radiación difusa y radiación ultravioleta. Por ello, nuestro estudio se ha centrado en ellas mismas.

Por un lado, se decidió utilizar un modelo matemático de regresión lineal, el cual consiste en establecer una relación entre una variable dependiente y varias variables independientes, mediante la aproximación de una recta que reduzca el error entre los puntos dados los resultados, tal y como podemos ver en la siguiente imagen:



Fuente: es.wikipedia.org

Para logar dicho fin, nos hemos apoyado en la librería SKLearn, la cual nos proporciona el entrenamiento y la predicción de dicho modelo de una forma sencilla y práctica. Se decide tomar este modelo como referencia, ya que es el modelo más básico de todos con el cuál podemos comparar el resto de modelos probados.

Los resultados obtenidos del test son los siguientes:

REGRESIÓN LINEAL

PARÁMETRO	MAE	R^2
VDC	1.2184	-12.3984
CDC	1.0616	0.5301
PDC	30.1481	0.5088

Atendiendo a las métricas, los resultados no son tan malos como se esperaban. Cierto es, que el parámetro de tensión (VDC) es el menos replicable, algo entendible sabiendo que físicamente no tiene una relación directa con los parámetros meteorológicos, ya que las placas solares producen corriente de forma directamente proporcional a la radiación.

A partir de aquí, se busca mejorar la predicción mediante los modelos de árbol de decisión. Para ello, inicialmente se probó con un modelo básico de profundidad 3 y con los hiperparámetros estándar que proporciona el modelo, con lo que se obtuvieron los siguientes resultados:

ÁRBOL DE DECISIÓN (profundidad = 3)

PARÁMETRO	MAE	R^2
VDC	1.1607	-4.3615
CDC	1.0698	0.5507
PDC	29.9414	0.5368

Los resultados mejoran, pero no lo suficiente como para considerar que tenemos un buen modelo. El siguiente paso que se dio, fue el de utilizar GridSearchCV para buscar optimizar los hiperparámetros mediante la técnica de Cross-Validation. En este caso, los resultados obtenidos fueron los siguientes (usando los hiperparámetros más optimizados que se encontraron):

ÁRBOL DE DECISIÓN (profundidad = 12)

PARÁMETRO	MAE	R^2
VDC	1.0990	-2.4363
CDC	0.9818	0.6357
PDC	27.3587	0.6280

En este último caso, creemos que tenemos un mejor modelo que el de regresión lineal. Aún continúa siendo difícil predecir el valor de VDC ya que no se consigue reducir el valor de R^2, el cual se debería encontrar entre 0 – 1 para ser considerado válido el modelo. Igualmente, no nos preocupa tanto ya que se busca tener una predicción de la producción (PDC), y físicamente hablando, es fácil de calcular el valor de la tensión porque, en corriente continua: P = V*I. Por lo que en el caso de querer tener un valor más preciso de la tensión dividiríamos la potencia entre la corriente producida.

6. PROTOTIPO DE PRODUCCIÓN

Una vez tenemos los modelos entrenados y validados, para completar el proyecto se ha creado un dashboard en Tableau como prototipo de como sería el proyecto en producción. Para ello, se buscaba que el propio algoritmo buscara la información meteorológica en la red, la preparara y pasara por los modelos, generando así la predicción de la producción.

Por tanto, antes de su visualización, se creó un notebook ("03_Scraping_AEMET.ipynb") que contiene un prototipo para la conexión y descarga de los datos de la radiación suministrados por la AEMET. Esto se puede llevar a cabo gracias a que la propia agencia meteorológica posee una API mediante la cual, una vez registrado y obtenido la clave, se puede descargar de forma remota los datos de predicción diarios. Como objetivo final, se buscaba que dicho código se ejecutara de forma automática diariamente, pero no ha sido posible lograrlo por falta de tiempo.

Una vez se han recopilado, limpiado y tratados los datos; estos son introducidos en los diferentes modelos, con la intención de generar un dataframe de respuesta de predicción. Tras obtenerse dichos datos, estos son guardados en archivos a nivel local, y subidos a la plataforma Dropbox.

Se ha optado por dicha plataforma, porque Google Drive que era la alternativa planteada, requiere de autorización y login antes de subir un archivo, en vez de utilizar un token como lo hace Dropbox, lo cual dificultad el trabajo con ella.

Tras la subida de dichos archivos a la plataforma, nos conectaremos a la misma a través de Tableau Desktop, que ha sido la herramienta utilizada para crear los dashboard de visualización.

LA visualización se planteo de tal manera, de que los datos descargados de la AEMET correspondientes a las diferentes estaciones meteorológicas repartidas por el territorio nacional, se simularan como si fueran centrales eléctricas de una compañía eléctrica cliente. De esta forma se puede acercar mas a la idea de lo que se buscaba solventar con este proyecto.

La visualización de los datos, se ha centrado en 3 pantallas diferentes:

- **Previsión de radiación:** aquí se muestra de forma genérica la previsión de las 3 radiaciones suministradas por la AEMET diariamente. En ella se pueden seleccionar cual de las 3 radiaciones se visualizar, de qué central/estación se quieren ver los datos y las gráficas acumulativas diarias de cada una de ellas en una gráfica de barras. Desde ella, te puedes mover a cualquiera de las otras pantallas (Alt+clic).

- **Previsión de producción:** esta pantalla, presenta un formato igual al anterior, salvo que en vez de presentar los datos de predicción meteorológicas se presentan los datos obtenidos de los modelos de predicción seleccionados. Desde ella, te puedes mover a las otras dos pantallas (Alt+clic).

- **Centrales individuales:** en esta última pantalla, se quería mostrar las gráficas completas de predicción de cada central, por falta de tiempo no se ha podido mostrar de una forma más agradable y detallada, aunque es posible ver los datos sin gran dificultad. Desde ella, te puedes mover a las otras dos pantallas (Alt+clic).

7. CONCLUSIÓN

La realización de este proyecto ha supuesto un gran reto personal y profesional. Antes de comenzar el mismo, no había tenido ningún tipo de contacto con proyectos relacionados con la ciencia de datos, no sabía como cargar y ejecutar un modelo, como funcionaba internamente, qué métricas existían, ni recordaba los conceptos estadísticos que los regían.

Tras finalizar dicho proyecto, no puedo decir que esté plenamente satisfecho del mismo, se puede mejorar en todas y cada una de las partes que lo componen, pero se puede decir que cumple gran parte de su cometido.

La limpieza de datos, el análisis de los mismos y selecciones de los más relevantes para el entrenamiento de los modelos ha supuesto la parte de mayor carga de trabajo, aún así se ha conseguido, bajo criterio personal, crear una estructura que facilita el entrenamiento de modelos predictivos destinados a dicho fin. Ciertamente, se pueden hacer grandes mejoras y sobre todo optimizar los recursos de las máquinas que deban procesarlo, ya que hemos detectado que se hace un gran uso de la memoria de los servidodes.

En lo referente al entrenamiento y generación de los modelos, se puede mejorar mucho la predicción de los datos. Hay una gran variedad de modelos que quedan sin probar y que posiblemente aporten mejores resultados de los conseguidos. Pero igualmente, se ha conseguido obtener un modelo que replica de una forma bastante significativa el comportamiento de las placas fotovoltaicas.

Para acabar, la visualización de los datos, es la parte con mayor potencial de todas. Queda mucho por desarrollar en ella, y con tiempo se puede llegar a crear grandes prototipos que mejoren y faciliten las visualizaciones de los datos y la toma de decisiones a través de los mismo. Aunque creo que se puede entender lo que se busca con ello.