Proyecto Cajamar: Predicción de consumo de agua

TEAM2021IA3: Felipe Higón Martínez, Josep Año Gosp y Francisco José Iniesta.

21 de marzo de 2022

1 Resumen del trabajo desarrollado

En este proyecto tenemos datos de series temporales de 2747 contadores de agua desde el 1 de febrero de 2019 al 31 de enero de 2020 con frecuencia horaria. Como variables disponemos del número del contador (ID), la parte entera del consumo (DELTAINTEGER), la parte decimal del consumo (DELTATHOUSANDTH), la parte entera de la lectura del contador (READINGINTEGER) y la parte decimal de la lectura (READINGTHOUSANDTH).

Nuestro objetivo es predecir el consumo de agua de cada uno de los contadores para los días del 1 al 14 de febrero de 2020. También buscaremos predecir el consumo semanal de la semana del 1 al 7 y del 8 al 14 de febrero.

Para abordar este proyecto de forma que sea generalizable a más contadores hemos aplicado programación orientada a objetos con Python creando una clase denominada "Contador", que nos define la lectura de los datos del contador y el preprocesado de los datos de cada contador. En esta clase además se determina el modelo a aplicar para predecir el consumo del contador. También hemos creado otra clase denominada "Cajamar Water" que interactúa con la clase "Contador" y realiza la lectura y carga de los datos de cada uno de los contadores en función de su ID, en esta clase se separan las series temporales de los contadores en entrenamiento y validación. La validación se corresponde con las últimas dos semanas de las que disponemos. En esta clase se realiza el entrenamiento de los modelos, y también, a partir de los resultados de error obtenidos en la validación, la búsqueda de los mejores hiperparámetros para nuestro modelo. También calculamos el RMSE de la forma que se describe en el datathon. Además, se generan las gráficas de predicción de la validación de los contadores. Se puede además mostrar las gráficas de las mejores y peores predicciones según su valor de RMSE. Finalmente, se genera un archivo .csv en el formato especificado con la información de la predicción para cada uno de los contadores.

2 Resumen del análisis exploratorio

Para la realización del análisis exploratorio en primer lugar hemos realizado una descripción global de las columnas del conjunto de datos mediante diferentes parámetros estadísticos; determinando su media, desviación típica, mínimo, máximo y los percentiles 25, 50 y 75.

En este proyecto se realiza un modelo de forma individual para cada contador. Por tanto, en primer lugar, hacemos un análisis de los datos de cada contador. Buscando detectar valores atípicos (outliers, valores negativos, secuencias muy largas de valores nulos...) en la serie temporal de cada contador. También nos fijamos en la longitud de cada serie temporal después de agrupar los datos diariamente, con ello podemos detectar contadores con información dañada o poco útil para predecir su consumo.

De este modo identificamos contadores con posibles errores en sus lecturas y consumos, así como contadores de los cuales no podremos hacer predicción puesto que su serie temporal finaliza varios días (en ocasiones meses) antes de la fecha umbral para predecir, el 31 de enero de 2020 concretamente. Se muestra en la siguiente figura mediante un diagrama de sectores la cantidad de contadores de cada tipo:

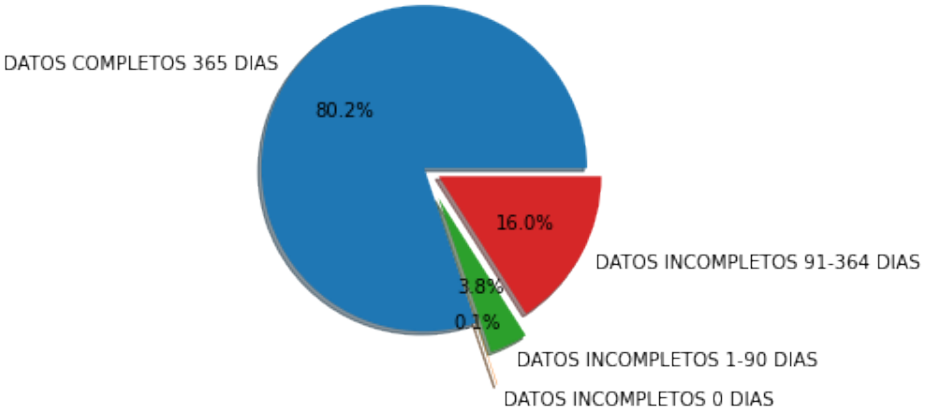

Figure 1: Diagrama de sectores según el tipo de contador

También hay que señalar el uso de varios tipos de normalización sobre los datos para identificar, de forma individual, cual es la normalización que mejor ayuda a predecir los consumos de cada contador. Hemos normalizado los datos de las siguientes formas:

- Sin normalizar, con los datos originales.
- Normalización a una distribución normal $N(0, 1)$.
- Normalización mínimo-máximo.
- Transformación de la serie temporal mediante el cálculo de medias móviles de tamaño 2, 3, 4, 5, 6, 7 y 10.

Las normalizaciones que mejor nos funcionan para predecir el consumo (en la fase de validación) son las medias móviles, ya que conseguimos suavizar la información de consumos de la serie temporal y es más sencillo para los modelos entender los datos. Decir también que al aplicar las medias móviles la longitud de la serie temporal disminuye el tamaño de media móvil utilizado menos uno, es decir, si aplicamos medias móviles tamaño n obtenemos series temporales donde los $n - 1$ primeros valores de la serie son "NaNs". Rellenamos estos valores "Nan" con ceros puesto que los modelos se fijarán fundamentalmente en los últimos valores de la serie temporal para realizar su predicción. Se aprecia en la última columna del histograma de la siguiente figura que el valor de RMSE promedio al aplicar una media móvil concreta para cada contador es significativamente mejor que aplicando una única media móvil para todos los contadores:

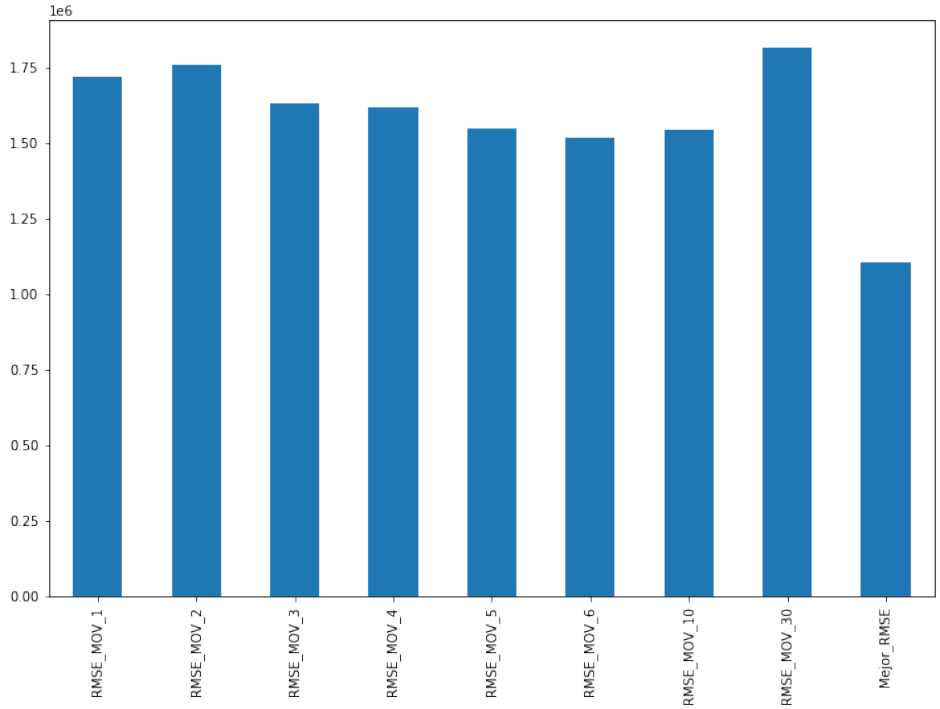

Figure 2: Histograma del valor RMSE promedio al aplicar distintas medias móviles a los datos, en la última columna se aplica de forma individualizada la media móvil a cada contador.

Con esta exploración sobre los datos ya estamos en condiciones de aplicar un modelo concreto a los datos de cada uno de los contadores de agua. Esto se explicará con detalle en el último apartado.

3 Resumen de la manipulación de variables y su argumentación

Decir que en primer lugar al analizar la columna **"ID"** hacemos una búsqueda de valores únicos para determinar la cantidad de contadores de los que disponemos y se aprecia que el valor de los IDs varía desde 0 hasta 2756 con algunos huecos, es decir, aunque llegamos al ID 2756 realmente tenemos 2747 contadores diferentes.

Las columnas de la parte decimal de la lectura y el consumo, **"READINGTHOUSANDTH"** y **"DELTATHOUSANDTH"** respectivamente. Son las únicas variables en las cuales tenemos valores "NaN", puesto que el valor medio de estas columnas es aproximadamente 0, se ha considerado adecuado rellenar estos valores "Nan" por ceros. También indicar que estas variables alcanzan su valor máximo en 100. Lo cual permite agrupar fácilmente los datos de consumo y de lectura uniendo sus partes entera y decimal.

Las variables **"READINGINTEGER"** y **"DELTAINTEGER"** en algunos contadores tienen valores negativos, los cuales son eliminados de las series temporales de los contadores que presentan estas características. Como se ha comentado en el párrafo previo, las variables **"READINGINTEGER"** y **"DELTAINTEGER"** se combinan con las variables de la parte decimal para generar dos nuevas variables a las que llamamos **"LECTURA"** y **"CONSUMO"**. Se calculan sumando la parte entera a la parte decimal dividida por 100. También se genera una nueva variable denominada **"CONSUMO NORM"** que se obtiene a partir de la normalización de la variable "CONSUMO" a una distribución

normal $N(0,1)$. Se utiliza esta variable de "CONSUMO NORM" para eliminar los outliers. En este caso para asegurarnos de que los valores que eliminamos son outliers vamos a considerar outlier a todo aquel valor que esté por encima de 8 o por debajo de -8 en la variable "CONSUMO NORM". Es decir, cualquier dato de una serie temporal de consumo de un contador que esté 8 desviaciones típicas por encima o por debajo de la media será eliminado del conjunto de datos. Se eliminará tanto el dato del consumo como el de lectura. Es importante decir que se debe realizar un test de normalidad a los datos para aplicar esta normalización de forma que tenga sentido de acuerdo a la distribución de los datos. Por falta de tiempo no se ha analizado la normalidad de los datos. Esta es mejora que tenemos pendiente de añadir.

Para la columna **"SAMPLETIME"** hemos visto en primer lugar el rango de fechas en el que se mueven las series temporales de forma general, determinando el máximo y el mínimo de esta columna, apreciamos que el valor mínimo se corresponde con el 2019-02-01 y el máximo con el 2020-01-31 . También decir que esta variable se utiliza como índice temporal de cada contador de forma individual y que realizamos una transformación de la frecuencia de muestreo de los datos. En principio este muestreo de los datos se realiza de forma horaria, por ello, los datos son agrupados a frecuencia diaria sumando el consumo realizado en un día. Para obtener la lectura diaria del contador realizamos la búsqueda del valor máximo cada día.

Una vez se muestrean las series temporales con frecuencia diaria se realizan las normalizaciones comentadas en la sección anterior sobre las nuevas variables **LECTURA** y **CONSUMO** y se guarda en un fichero cuál es la mejor normalización en base al RMSE obtenido en la fase de validación.

Un problema que ha surgido al tratar con las variables de "LECTURA" y "CONSUMO" es que no se tiene forma de saber de cual de estas dos variables provienen los errores en los datos. Concretamente, estos errores son valores negativos de consumo, lecturas de contador que no se corresponden con el consumo y lecturas de contador que son negativas de un día para otro cuando, en teoría, la sucesión de lecturas de un contador debe ser monótona creciente en el tiempo. El conjunto de decisiones que se han tomado para resolver estos problemas de la mejor forma posible han sido los siguientes:

- Eliminar filas con valores de consumo negativos.
- Buscar contadores con diferencias entre consumo de litros y lectura de contadores superiores a 100000 litros y analizar de forma individual los datos de dichos contadores.
- Analizar las lecturas de contador que son negativas de una hora a otra y eliminar esa fila si en ese caso también teníamos un valor de consumo negativo o un valor de consumo considerado outlier.
- Eliminar las filas con outliers en el consumo.

Se intenta no eliminar filas por un error únicamente en la lectura; puesto que eso nos llevaría a eliminar un consumo que podría ser correcto. Por tanto, en general, hemos priorizado la variable de consumo con la intención de tener los mejores datos de consumo posibles. Ya que, al fin y al cabo, esta será nuestra variable respuesta.

También, explicar que se ha considerado realizar un modelo que tenga en cuenta variables externas, para ello, se han buscado datos climáticos de la Comunidad Valenciana y datos del precio del agua en cada zona de la comunidad para las fechas de la serie temporal. Sin embargo, a la hora de ponerlo en práctica deberíamos utilizar datos de predicciones climatológicas y de precio. Además, no conocemos la ubicación geográfica de cada contador de agua.

4 Justificación de la selección del modelo

Para explicar la selección del modelo utilizado primero se debe tener en cuenta que las series temporales de consumos de los contadores cuya longitud es inferior a 100 días o con finalización de la serie antes del 31 de enero de 2020 tienen como respuesta para cada uno de los días de predicción el valor medio de la serie temporal.

Por tanto, a los contadores con series temporales de más de 100 días que finalizan el 31 de enero de 2020 se les ha aplicado un modelo de random forest autoregresivo, a través de la documentación encontrada de una librería personalizada denominada skforecast.

En dicha librería se nos permite realizar de forma sencilla una búsqueda de hiperparámetros para nuestro modelo. Por lo que cada contador tiene un modelo personalizado para el cual se buscan los mejores hiperparámetros.

En primer lugar, explicar que al disponer únicamente de un año a la hora de partir los datos para entrenar, validar y testear cada serie temporal no podemos realizar el testeo de los datos sobre los propios días de predicción que se nos piden en el proyecto utilizando un año diferente. En nuestro caso, lo que hemos hecho ha sido utilizar la penúltima semana para validar los modelos. Con los resultados de error en la validación, aplicamos una búsqueda de hiperparámetros para determinar cual es el mejor ajuste del modelo de cada contador. Y con la última semana disponible de las series temporales realizamos el test y de ahí calculamos un valor de RMSE que nos da indicación de como predicen los modelos. La desventaja de trabajar con los datos de este modo es que realmente estamos preparando nuestro modelo para predecir los días del 24 al 31 de enero. Sin embargo, se ha considerado que con los datos disponibles realizar esta acción es la mejor posible puesto que estamos lo más próximos a las fechas de predicción que se nos piden en el concurso.

También indicar que no se ha contemplado la idea de aplicar modelos más complejos de deep learning como podrían ser redes convolucionales unidimensionales puesto que la longitud de las series temporales es de 365 como máximo. No se considera que dicha cantidad de datos sea lo suficientemente grande como para ser útil a un modelo de aprendizaje profundo.

Añadir que no se han normalizado los datos para entrenar los modelos puesto que el modelo de random forest trabaja bien aunque los datos no estén normalizados. Por ello hemos aplicado las medias móviles con los datos originales sin normalizar.

Por último, hay que explicar que se ha buscado la aplicación más idónea de este modelo a cada contador. Pero también habría sido útil aplicar otros modelos de predicción de series temporales como pueden ser los modelos ARIMA. Se analizará esta posible mejora en caso de pasar a la siguiente fase.