LA MANZANA DE NEWTON

1. Breve resumen del trabajo desarrollado

En este apartado se explica una visión general de los pasos que hemos seguido. En primer lugar, hemos centrado gran parte de nuestros esfuerzos en el análisis exploratorio de los datos y más en particular en los missing values que había en el dataset (horas para las cuales un determinado contador no había tomado una medida). En segundo lugar, hemos utilizado varias estrategias para inputar estos missing values atendiendo a cuan malo o bueno era un contador, entendiendo por un contador bueno aquel con casi todas las medidas y uno malo aquel al que le faltan muchas medidas. A continuación, hemos generado nuevas variables explicativas que hemos considerado que pudiesen ser relevantes para el modelo (dado que en el fondo el dataset solo tenía una variable ya que DELTA y READING aportaban información extremadamente parecida). Más adelante en este informe detallaremos qué variables hemos generado y por qué. Finalmente, con un dataset limpio, sin missing values y con nuevas variables explicativas hemos entrenado varios modelos hasta quedarnos con los dos mejores (entendiendo por mejores aquellos que obtenían un menor MSE en el conjunto de test que nos hemos fabricado con la última semana de Enero). Para la predicción final hemos tomado la media de las predicciones de estos dos modelos.

2. Resumen del análisis exploratorio llevado a cabo

El principal objetivo de nuestro análisis exploratorio ha sido detectar e interpretar los fallos en los contadores. Hemos detectado 3 tipos de fallos:

- Un contador no ha tomado una medida a una determinada hora
- Una lectura de READING es menor que la anterior (READING debería de ser siempre creciente)
- Para una medida READING y DELTA no son congruentes (DELTA debería ser la diferencia entre dos READINGs consecutivos)

Atendiendo a estos tres tipos de fallos hemos analizado cuántos contadores fallaban, por qué creíamos que fallaban y qué podíamos hacer para corregir estos fallos. Explicaremos como hemos mitigado estos fallos en el siguiente apartado.

Además de este análisis de fallos hemos hecho otros análisis como la estacionalidad y tendencia de algunos contadores o un intento de clustering para agrupar contadores con similitud entre ellos. No obstante estos otros análisis no nos han llevado a ninguna conclusión clara y por ello el principal peso de nuestro análisis exploratorio ha sido el análisis de fallos.

3. Resumen de la manipulación de variables y su argumentación

La manipulación de variables que hemos llevado a cabo se resume en tres fases claramente diferenciadas:

- Agrupación
- Imputación de missing values y corrección de fallos
- Generación de variables a partir del contador y de la fecha
- Generación de variables climatológicas

Agrupación

Primeramente, hemos agrupado los datos por días (ya que las medidas estaban tomadas cada hora, pero nos interesaba las medidas por día de cara a la predicción). Para agruparlas hemos sumado los DELTA de cada día y contador y nos hemos quedado con el máximo de los READING de cada día y contador. Esto tiene sentido ya que DELTA es el consumo de agua y READING es una medida acumulativa. No obstante, DELTA y READING aportan información muy similar por lo que hemos decidido quedarnos únicamente con la columna READING.

Imputación de missing values

Para imputar las medidas que faltaban hemos utilizado la siguiente estrategia: si las medidas faltantes se encontraban entre dos medidas READING, entonces tomamos la diferencia entre esas dos medidas READING y dividimos entre el número de días que faltan. Si las medidas faltan al principio o al final de la serie temporal, entonces sencillamente imputamos usando la media de consumo de esos contadores. Esta situación se daba en contadores con extremadamente pocas medidas por lo que no hemos conseguido dar con una solución buena para un contador que tenga medidas solamente para dos días.

Generación de variables a partir del contador y de la fecha

Para enriquecer el dataset hemos incluido algunas variables relacionadas con cada contador y con la fecha.

Variables relacionadas con el contador:

- Media de consumo diario
- Varianza del contador

Variables relacionadas con la fecha:

- Día del año
- Día de la semana
- Es fin de semana o no

Generación de variables climatológicas

Finalmente hemos considerado muy enriquecedor para los modelos incluir variables climatológicas (intuitivamente si hace más calor se consumirá más agua). Para ello hemos utilizado la API de AEMET para incluir las siguientes variables:

- Temperaturas máxima, mínima y media
- Precipitaciones
- Sol

4. Justificación de la selección del modelo.

Una vez hemos obtenido un dataset limpio con nuevas variables explicativas para enriquecerlo hemos empezado a probar modelos. En una situación ideal nos hubiese gustado dedicar más tiempo a probar modelos más complejos como redes LSTM etc. Pero dado que nuestros principales esfuerzos (al menos en tiempo) los hemos dedicado a la manipulación del dataset nos hemos conformado con probar múltiples modelos de regresión y series temporales.

Para evaluar la bondad de un modelo nos hemos fabricado un conjunto de test empleando las dos últimas semanas de Enero de 2020 (justo el final de la serie temporal). Con nuestro conjunto de train y nuestro conjunto de test hemos lanzado múltiples modelos y nos hemos quedado con los dos mejores (los que menor MSE obtenían en nuestro conjunto de test).

Estos dos modelos han sido GradientBosstingRegressor y XGBoostRegressor. Para la predicción final para las dos primeras semanas de Febrero hemos utilizado los dos modelos y hemos hecho la media de los resultados de las dos regresiones.