UNIVERSITYHACK 2022 DATATHON:

En este Datathon se nos ha planteado el reto de la correcta estimación de la demanda de agua potable.

Nuestro principal objetivo ha sido identificar, aislar y emplear las variaciones temporales, así como otros factores económicos y naturales para lograr de esta forma la mejor predicción. Hemos partido de un amplio dataset con un histórico de consumos de la zona geográfica del litoral valenciano.

Además, hemos creado para los análisis una aplicación en la que poder visualizar mejor las conclusiones y los análisis realizados. Se puede acceder mediante el siguiente enlace: https://share.streamlit.io/jaimesz11/uni2_datathon_2022/main/app_water.py

Hemos estructurado el proyecto de la siguiente forma:

- Análisis exploratorio y descriptivo de los datos proporcionados
- Toma de decisiones respecto a la manipulación de las variables
- Análisis exhaustivo del componente temporal
- Elección del modelo de predicción
- Entrenamiento, validación del modelo elegido y predicción
- Desarrollo de aplicación web para una mejor experiencia de visualización para terceros

1. Análisis exploratorio y descriptivo

Hemos iniciado el proyecto visualizando la información básica de nuestros datos: número de registros, ids únicos de contadores, las variables con las que contábamos, etc.

Nos encontramos ante un dataset muy extenso, con un total de 21404828 registros históricos para 2756 contadores. Sin duda alguna, el componente temporal es el gran protagonista del análisis.

Tras conocer estas primeras pinceladas acerca de nuestro dataset, iniciamos un análisis más pormenorizado en torno a aquellos datos que presentaban una cierta problemática.

De esta forma, nos enfocamos en: valores NA, registros sin data y análisis de negativos y otros outliers.

NAs

Del análisis de los NAs pudimos identificar que un total de 140056 registros no tenían información en la parte decimal de la lectura (por lo tanto, ni para READING ni para DELTA).

Decidimos imputar los valores NA por ceros, viendo la distribución previa de los ceros en los contadores y que dado que solo se observan en la parte decimal, el margen de error es despreciable.

Registros con valores igual a 0

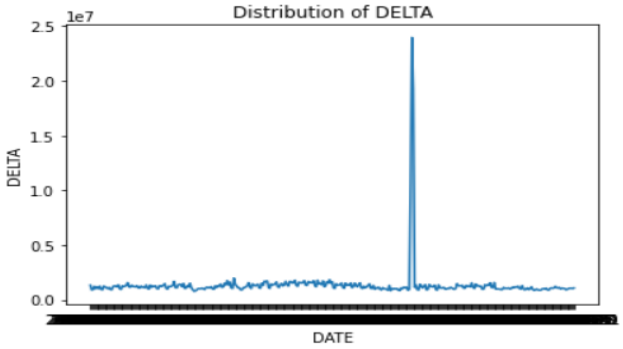
Fuimos capaces de observar que un total de 19 contadores tenían en algún momento de su vida todos los registros igual a 0. Procedimos a analizar qué proporción del total de los registros constituía para poder tomar decisiones.

Extrajimos que de entre ellos, 7 contadores (272,1896,2135,2542,2544,2545,2547) tenían todos sus registros igual a 0, por lo que lo que se procedió a eliminarlos dado que no nos aportaban ninguna información útil para el modelo. Se toma esta decisión, tomando como base la teoría de que los contadores citados pueden estar averiados o fuera de uso. Con ello perdimos un total de 43.838 registros (0.204%).

Respecto al resto de contadores, se observa que los registros igual a cero se corresponden con los primeros días de medida, confirmando nuestra teoría (el inicio de consumo de agua tiene lugar posterior a la fecha de alta del contador).

Análisis de negativos y otros outliers

Con respecto a los outliers nos saltó la alarma al ver en especial el gráfico de la tendencia temporal de la variable DELTA.



Se aprecia un clarísimo pico en el mes de octubre, el cual nos impide apreciar la tendencia del resto de días.

Al entrar a trabajar en estos datos que ya intuíamos que nos estaban alterando la tendencia temporal logramos identificar en primer lugar la primera problemática: **registros con READING y DELTA o solo DELTA negativos**.

Encontramos dos contadores (1041,2711) con READING negativo, los cuales consideramos errores de lectura puesto que no logramos encontrar un comportamiento

lógico dentro de los mismos. Se toma la decisión de eliminarlos. Se renuncia a un total de 11.529 registros.

A continuación se procedió a analizar la problemática de los que tienen el DELTA negativo en alguno de sus puntos. Se considera que esta problemática se debe a errores en la lectura, por lo que se opta por sustituir dichos deltas negativos por cero.

Tras haber solucionado la problemática de los negativos, pasamos a analizar los valores atípicos presentes en nuestro dataset, tanto por medio del método *zscore* como por el método de rango intercuartílico. Se detectan 55 contadores con outliers por el método de zscore y 2618 contadores con outliers por el método del rango intercuartílico.

Con base a los resultados y a fin de no eliminar registros buenos, se decide emplear el método de *zscore* para la eliminación de outliers. Hemos adoptado una postura en cierta forma conservadora con los outliers dado que somos conscientes de que los registros pueden corresponderse tanto a viviendas como a locales comerciales, por lo que haría falta ser conocedor de qué en caso se encontraría cada registro para saber si es o no outlier dada su condición. Es por eso por lo que hemos optado por conservar la mayor cantidad de registros.

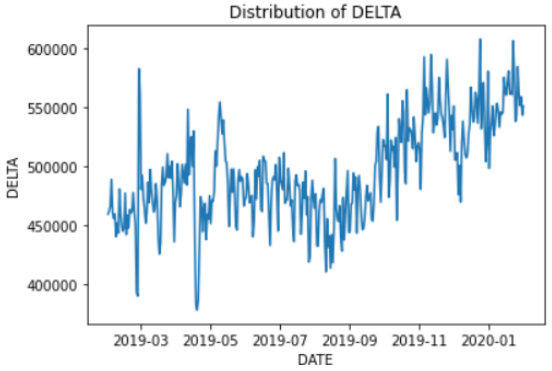
Respecto a la eliminación o no de outliers, se toman las siguientes conclusiones:

- Se mantienen los outliers que se concentran en determinadas franjas horarias.
- Se considera que los outliers presentan un patrón de comportamiento temporal cuando se suceden en un mínimo de tres días consecutivos y, en consecuencia, se mantienen.
- Los outliers alejados del comportamiento habitual del contador son suprimidos.

2. Manipulación de las variables

El resumen de los cambios que se han realizado con respecto a las variables es: imputación de los NAs por ceros, eliminación de siete contadores sin ningún registro, eliminación de dos contadores con READING negativo, eliminación de 276 registros de outliers.

Tras el tratamiento de los outliers y demás limpieza de los datos nos encontramos la siguiente representación gráfica:



Con esta limpieza preliminar podemos apreciar perfectamente la tendencia del DELTA.

Tras la limpieza, se ha procedido a la creación de diversas variables de carácter temporal capaces de aportarnos una gran cantidad de información visual de cómo se comporta la demanda de agua.

En primer lugar, nos interesó saber si una lectura era de un hogar o de una empresa, para lo cual recurrimos al consumo diario. El consumo medio de una vivienda (con los habitantes de la misma en la media de la Comunidad Valenciana) en un día sería de 132 * 2,42 = 319 litros aproximadamente. Dejando un margen de error, tomamos la decisión de que si el consumo al día es menor de 400 litros se considerará que es una vivienda, y si es de 400 litros o más, se considerará local comercial.

Asimismo, de la fecha de registro que se nos daba con los datos originales extrajimos: la hora, el día de la semana, el mes y la estación del año.

3. Análisis temporal y predicción

El objetivo de predicción que se nos planteaba era el siguiente: consumo diario del 1 al 7 de febrero incluidos, consumo de la primera semana de febrero (del 1 al 7 incluidos) y consumo de la segunda semana de febrero (del 8 al 14 incluidos).

El principal reto que nos ha supuesto este concurso ha sido el volumen de los datos.Iniciamos la predicción con la idea de correr un modelo ARIMA. Para ello, tras haber llevado a cabo los dos primeros steps y habiendo limpiado nuestro dataset original e introducido nuevas variables que nos ayudan a entender el comportamiento de consumo de agua, procedemos a agrupar la data que tenemos por día.

Para la correcta predicción mediante series temporales se requiere que los datos tengan carácter estacionario. Para comprobarlo, aplicamos el test de Dickey-Fuller (ADF) obteniendo un p-valor superior a 0,05. Por lo tanto no podemos rechazar la hipótesis nula de que los datos tengan raíz unitaria, y, en consecuencia, que los datos no sean estacionarios. Con el fin de alcanzar el objetivo de que los datos sean estacionarios, se procede a realizar varias transformaciones. Primeramente, tomamos logaritmos, pero el problema no se soluciona. Es por eso por lo que pasamos a tomar primeras diferencias, obteniendo un p-valor inferior a 0,05 y logrando, por tanto, una serie estacionaria.

Procedimos a representar los gráficos de autocorrelación y de autocorrelación parcial para identificar los valores de la parte del proceso autorregresivo (AR) y de media móvil (MA), planteándonos retardar todavía más los datos dado que pudimos observar un comportamiento persistente cada 7 días.

Optamos por lanzar dos modelos *auto-arima*, uno sobre la serie temporal retardada un periodo y otra sobre la misma serie pero retardada otros 7 periodos. Los mejores datos los obtuvimos con el primer modelo, por lo que pasa a ser el modelo elegido.

Sin embargo, a la hora de lanzar las predicciones por contador y por día nos encontramos con cierta problemática computacional.

Por lo tanto, condujimos nuestra predicción al uso de la librería Prophet. Por medio de un bucle, entrenamos y corrimos el modelo agrupado para cada id, obteniendo de esta forma la predicción presentada.

Sin embargo consideramos que existe cierto margen de mejora en el que seguiremos trabajando.

4. Aplicación web

Para facilitar la interpretación de los análisis realizados y visualizar de forma más dinámica y sencilla las conclusiones extraídas, se ha tomado la decisión de desarrollar una aplicación web. Esta plataforma, que se presenta como una herramienta visual para facilitar la comprensión de las decisiones tomadas, consta de tres apartados, en los que se puede seguir el trabajo realizado desde el análisis exploratorio de los datos hasta llegar a la predicción final.

El enlace para entrar a la aplicación es el siguiente:

https://share.streamlit.io/jaimesz11/uni2_datathon_2022/main/app_water.py

A continuación, se muestra un ejemplo de lo que se puede encontrar en la aplicación web.