***Reto Cajamar Water Footprint – EXABY***

La planificación original del desarrollo del reto Cajamar Water Footprint ha ido variando conforme conocíamos más en profundidad la naturaleza de los datos e implementábamos nuestro proyecto.

Uno de los principales cambios en nuestro planteamiento inicial consistió en dejar de invertir tiempo y esfuerzos en encontrar orígenes de datos adicionales y enfocarnos únicamente en los datos proporcionados por la organización del hackathon.

Una vez tomada esta decisión, nuestro desarrollo ha tenido distintas etapas:

* Estrategia común para tratar el dataset original de gran tamaño.
* Conformar el proceso de análisis exploratorio.
* Realizar pruebas de predicciones.
* Añadir a los modelos de predicción configuraciones que puedan mejorar los resultados.
* Obtener los resultados finales en base a los resultados más precisos.

El proyecto se ha llevado a cabo progresivamente recorriendo las etapas mencionadas anteriormente dedicando bloques de tiempo acordes a las dificultades encontradas en el desarrollo.

*Análisis exploratorio*

La fase de análisis exploratorio está implementada en el notebook de jupyter denominado ‘Script\_exploracion’.

En este notebook se puede apreciar que se hace un trabajo inicial para conocer la naturaleza de los datos y variables del dataset original.

La estrategia escogida para desarrollar esta fase se basa en escoger 4 identificadores de contadores al azar para analizar la evolución temporal de los datos originales. Esto permite que el análisis de los datos se haga sumamente más ligero al no tener que trabajar con más de 2700 contadores a la vez y además nos permite investigar tendencias comunes entre los contadores escogidos al azar.

La variable que más importancia hemos dado en el análisis es la variable ‘deltainteger’, y más en concreto, la variable obtenida al sumar la variable ‘deltainteger’ con ‘deltathousand’ de cada uno de los datos temporales.

Esta nueva variable nos ha permitido analizar datos de consumo de agua por contador en distintos rangos de tiempo, al implementar agrupaciones por datos diarios, semanales y mensuales. Mediante estas agrupaciones, hemos podido realizar un análisis de datos como medias o sumas por distintos rangos de tiempo.

De esta manera, hemos podido obtener datos relativos a días atípicos en cada una de las series temporales estudiadas dependiendo de los identificadores:

Interfaz de usuario gráfica, Gráfico, Aplicación

Descripción generada automáticamente

*Figura 1: Registros de 4 contadores al azar de consumo de agua diario.*

En la ‘*Figura 1*’, podemos observar la estabilidad en el consumo diario de cada uno de los contadores y además se puede apreciar la existencia de los llamados ‘días atípicos’, los cuales se explican y se detallan más en profundidad en el script de exploración.

Por supuesto, esta misma línea de trabajo en la fase de análisis se ha seguido estudiando con diversas agrupaciones y rangos de fechas.

Además del trabajo expuesto previamente, se ha llevado a cabo en la misma fase de análisis exploratorio un trabajo para estudiar el comportamiento de la serie temporal e investigar acerca de la efectividad de la implementación de distintos modelos.

Esta tarea comienza con un trabajo de descomposición de la serie (obteniendo datos de tendencias, estacionalidad y ruido) y completa con la realización de pruebas sobre datos de contadores individuales de implementaciones de los siguientes modelos:

* SARIMAX
* Random Forest
* KNN
* Rolling Random Forest

Gráfico, Gráfico de líneas

Descripción generada automáticamente

*Figura 2: Predicción a dos semanas utilizando un modelo Random Forest de un contador al azar.*

Por supuesto, estas implementaciones vienen acompañadas de sus predicciones y métricas asociadas, para así comprobar la calidad de los modelos y ayudar en la toma de decisión del modelo que mejor se ajusta a los datos de entrada.

**Manipulación de variables**

En cuanto a manipulación de variables, hemos implementado dos cambios principales sobre el dataset original:

* Por un lado, la variable temporal ‘SAMPLETIME’ ha sufrido una eliminación de los minutos y segundos. Esto simplifica mucho el trabajo a la hora de estudiar los datos por rangos de horas, días o semanas.
* Por otro lado, se ha generado una variable previamente mencionada la cual se basa en la suma de los datos de la variable ‘DELTAINTEGER’ con los datos de la variable ‘DELTATHOUSAND’. Esta nueva variable es la pieza principal de nuestro desarrollo y sobre la cual realizamos el análisis y predicción de consumo.

Además de cambios implementados sobre los datos originales, hemos incluido nuevas variables que mejoran la calidad del modelo utilizado en la predicción como son:

* *Weekday*: Variable que denota si el dato temporal es día laborable.
* Weekend: Variable que distingue si un día es fin de semana.
* Holidays: Esta variable se utiliza en la configuración del Prophet y se ha configurado de tal manera que se incluyen los días no laborables y festivos de la Comunidad Valenciana y Madrid en el año 2019.