Análisis del consumo energético de un centro comercial

Programa de experto en Data Science U-TAD



**Autor: Director:**

Raúl Amarelle Valera Carlos Gil Bellosta

Fecha: 03/09/2018

Índice

# Resumen ejecutivo

El presente proyecto trata del análisis de un data set que contiene, entre sus variables, datos del consumo energético de un centro comercial a lo largo de un periodo de dos años.

El objeto de este proyecto es analizar dicha serie temporal y la influencia del resto de variables, para en la medida de lo posible, realizar un modelo de predicción del consumo energético.

El proyecto se estructura en tres bloques principales:

1. Lectura y tratamiento de datos.
2. Visualización de datos y otras relaciones.
3. Casos de análisis de las series temporales y predicciones

La serie y el cálculo de predicciones se han realizado aplicando cuatro métodos:

1. Impacto causal utilizando modelos bayesianos de series de tiempo
2. Utilización del método Holt-Winters
3. Método ARIMA de predicción
4. Árboles de clasificación y regresión, RPART Tree.

El empleo de diferentes técnicas ha permitido comparar diferentes resultados de predicción y cómo hemos tratado de mejorarlo, los cuales se explicarán en el presente documento.

# Problemática y objetivos

Como sabemos, la electricidad es un bien que no puede ser almacenado, por lo que es de mucha utilidad conocer la previsión de consumo, aunque sea a corto plazo (un día) con el consumo real.

El objetivo general sería desarrollar una modelización para la previsión del consumo energético y compararla con la previsión del modelo actual.

Otros objetivos específicos:

* Análisis del dataset y generación de otra información que pudiera ser relevante.
* Analizar diferentes modelos estadísticos para la previsión de la demanda.
* Análisis comparativo de la calidad predictiva de cada modelo.

# Metodología

Tal y como hemos avanzado, el proyecto se estructura en tres bloques principales:

1. Lectura, análisis y tratamiento de datos.
2. Visualización de datos y otras relaciones entre variables.
3. Casos de análisis de las series temporales y predicciones.

La serie y el cálculo de predicciones se han realizado estudiando cuatro métodos diferentes:

1. Impacto causal utilizando modelos bayesianos de series de tiempo
2. Utilización del método Holt-Winters
3. Método ARIMA de predicción
4. Árboles de clasificación y regresión, RPART Tree.

# Tecnologías utilizadas

En la parte técnica, se ha utilizado R como lenguaje de programación.

Análisis y tratamiento de datos: uso de algunas librerías representativas como ggplot2, dplyr, data.table, zoo, etc.

Machine Learning: uso de librerías como bsts, Holt-Winters, forecast, tseries, MLmetrics, etc.

# Análisis del dataset inicial

Este dataset es una serie temporal con los siguientes datos relativos a un centro comercial.

Breve explicación de las variables:

1. Fecha: fecha de la medición, sólo hay una medición diaria
2. Estimado: sólo aparece “No”, no nos aporta información
3. Kwh: potencia real consumida
4. LB: Línea base, es decir, la predicción de potencia consumida
5. De 5 a 10. CCDD o CHDD: Estas seis columnas son de temperaturas, 3 de Cooling-Degree Day y 3 de Cooling Heating-Degree Day. Es la diferencia entre el promedio diario de temperatura y una determinada temperatura base de referencia, que suele ser la exterior.

CHDD18 = 8.5, quiere decir que sobre la temperatura de 18º, el centro comercial ha tenido que climatizar 8.5º. Por eso CHDD19=9.5º y CHDD20=10.5º, porque si la temperatura de referencia sube un grado, hay que calentar un grado más, la diferencia aumenta ese grado.

El centro comercial tiene dos modos de climatización: calefacción o refrigeración y en la práctica o funciona en un modo o en otro, por eso de las 6 columnas, son siempre 3 ceros vs 3 números.

1. Afluencia: Número de asistentes al centro comercial.

# Tratamiento de datos

De las 11 variables iniciales, podemos eliminar la 2 (“Estimado”) y 4 columnas de temperaturas. De las 3+3 columnas de temperaturas que tenemos, 2+2 son redundantes. Con quedarnos una columna de cada modo de climatización, es suficiente.

Escogemos CCDD20 y CHDD18, porque son las columnas más ventajosas. Es decir, porque si hay que calentar, CHDD18 es la línea de referencia mínima. Si hay que refrigerar, CCDD20 es la referencia mínima.

De la primera exploración de las 6 variables resultantes llama la atención un aspecto importante que conviene resaltar. La columna FECHA la reconoce como cadena de caracteres (chr) y conviene tratarla como fecha, para que pase a ser Date. Esto es importante para posteriores representaciones gráficas de la serie, porque normalmente el eje X reflejará las fechas, para ver la evolución de los valores.

Dentro de este apartado de Tratamiento de Datos, se incluyen una serie de variables creadas ad-hoc y de las que se puede extraer información relevante. Hemos realizado algunos análisis de la información disponible, pero más con la idea de ilustrar las posibilidades que ofrece el dataset. Lógicamente se podría haber profundizado más, pero el objetivo del proyecto es analizar una serie de modelos predictivos y no tanto en detallar la casuística que nos ofrece el dataset.

Caso aparte también merece el Tratamiento de Missing Values. De 729 muestras, sólo hay 4 NA, por lo que son tan pocas, que lo más práctico es eliminarlas. Además, en nuestra serie tampoco tiene sentido realizar alguna simulación para rellenar los huecos con media por columna o similar, porque las mediciones de consumo no guardan relación entre si.