Índice

1. Objetivo
2. Características DataSets
   1. Análisis General
   2. Problemas con el DataSet
   3. Selección de variables
3. EDA
   1. Estacionalidad
   2. Estacionariedad
   3. Tratamiento de outliers
4. Modelización
   1. Modelo ETS
   2. Modelo ARIMA
   3. Modelo PROPHET
5. Predicción
6. Modelización con biblioteca Pycaret
7. Conclusiones

**1. Objetivo**

El objetivo del presente proyecto es analizar las series temporales de demanda de energía eléctrica de California y realizar la predicción de los primeros 7 días de diciembre de 2021 en periodicidad horaria y la predicción diaria de los 15  
primeros días de diciembre 2021.

**2. Características del DataSet**

Como Data Set se ha utilizado un archivo en formato .csv que incluye dos hojas de datos referente a la demanda de energía en el Estado de California entre el 1 de julio de 2015 y el 30 de noviembre de 2021.

**2.1 Análisis General**

Como Data Set se ha utilizado un archivo en formato .csv que incluye dos hojas de datos referente a la demanda de energía en el Estado de California entre el 1 de julio de 2015 y el 30 de noviembre de 2021.

La primera de las hojas del documento tiene por título *Hourly Data*, e incluye como variables distribuidas en columnas *Region, UTC time, Local date, Local hour, Local time, Time zone y D*, correspondiente a la demanda. Dicho dataset incluye información sobre demanda de energía según hora del día en el horizonte temporal fijado anteriormente.

La segunda hoja tiene por título *Daily Data* e incluye 2346 filas de datos con variables de datos *Region, Local date, Time Zone y D*, correspondiente a la demanda de energía distribuida por días en el mismo horizonte temporal.

**2.2 Problemas con los DataSets**

El principal problema en la arquitectura del DataSet deriva de los cambios horarios que se producen en el Estado de California en marzo y en noviembre y que tiene afección directa sobre la distribución horaria. En una primera aproximación puede inferirse de forma errónea que dichos datos se corresponden con duplicidades o con errores humanos, pero una mayor prospección a la base de datos nos permite concluir que existen días en noviembre con 1 hora más en UTC Local y días en marzo con 1 hora menos en la misma variable debido a los ajustes necesarios en el conjunto de datos para adaptarlo a los cambios de horario de invierno y de verano.

Tabla

Descripción generada automáticamenteTabla

Descripción generada automáticamente

Ilustración . Extracto documento CSV

Un segundo problema visible en una primera aproximación gráfica es la existencia de outliers o valores extremos, y que, de forma directa pueden derivar de la agrupación de los días en los que existe mayores o menores registros por el cambio horario.

**2.3 Selección de variables**

Los anteriores problemas expuestos determinan la estrategia de selección de variables. En primer lugar, se decide realizar la predicción en base a la columna UTC Time para, una vez obtenido los valores objetivo, realizar las trasformaciones horarias necesarias adaptadas al momento temporal y los horarios invernales y estivales. De la misma forma, se adopta como estrategia la predicción diaria a partir de las predicciones horarias con el fin de reducir los posibles términos de error y por la falta de información en la distribución diaria sobre si la hora 24 (12 de la noche) se computa como el último término horario del día anterior o el primero del siguiente.

En resumen, se prescinde de la hora UTL Local y se opta por seleccionar como variable referencia de la serie temporal la hora UTC. De igual forma, las predicciones horarias se utilizarán para la predicción diaria, prescindiendo completamente del dataset “Daily Data”. Esta estrategia pretende integrar criterios y reducir los posibles errores que se acumulen en la predicción.

**3. Análisis Exploratoria de Datos – EDA**

**3.1 Estacionalidad**

Aunque en las predicciones se va a utilizar la hora UTC como referencia para evitar los errores anteriormente expuestos, en el caso del análisis exploratorio, si queremos obtener conclusiones a priori sobre la relación entre la demanda y el comportamiento social, plenamente determinante de la primera, se deben elegir datos temporales que se correspondan con el horario real. Esta circunstancia lleva a realizar un análisis gráfico con la variable predictora UTC Local y que va a dar una idea más robusta y pedagógica de la demanda en términos horarios.

Gráfico

Descripción generada automáticamente

Ilustración .Box-plot horario (Hora Local)

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración .Box-plot diario (Hora Local)

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración .Box-plot mensual (Hora Local)

De forma conclusiva, se observa que los datos siguen estacionalidad horaria, semanal y mensual, siendo las más relevantes la primera y la tercera. De esta forma se puede observar como la principal demanda eléctrica se produce entre las 17:00 y las 19:00 horas local y, según variable mensual, en Julio y Agosto. Resulta llamativo que la demanda no crezca en meses fríos, como sucede con datos de otros lugares y que deriva de la climatología menos brusca y más estable de California.

**3.2 Estacionariedad**

El análisis exploratorio nos permite concluir en una primera aproximación que la serie resulta prácticamente estacionaria, es decir que resulta constante en media y en varianza con datos horarios.

Gráfico, Histograma

Descripción generada automáticamente

Ilustración . Media Móvil, Std móvil. Serie temporal

No obstante, se ha procedido a realizar análisis Dickey-Fuller, ratificando la idea anterior, siendo el p-valor 0 y rechazando la hipótesis nula y concluyendo que la serie es estacionaria.

Texto

Descripción generada automáticamente

Ilustración 6. Test Dickey-Fuller Stationarity

Finalmente, el código fuente ha permitido extraer gráficos de diferencias que, de forma gráfica parecen mejorar la serie, se ha decidido no hacer uso de ello con objeto de no perder información sobre los horarios con mayores franjas de consumo.

**3.3 Tratamiento de outliers**

Por otro lado, se han identificado outliers que han sido tratados mediante algoritmo no supervisado de clustering KNN y en el que se ha fijado un nivel de contaminación de datos extremos del 5% con el fin de no perder excesiva información y no transformar en exceso la serie. Los outliers identificados han sido sustituidos por la media agrupada según la hora. La gráfica siguiente muestra los datos horarios antes del procesamiento y tras el procesamiento de valores extremos.

Gráfico, Histograma

Descripción generada automáticamente

Ilustración . Serie temporal Sin Intervención vs. Intervención

Otro aspecto que se debe destacar del tratamiento realizado es que, debido a que se va a limitar el conjunto de entrenamiento con el objetivo de no comprometer la capacidad de cómputo, el tratamiento de dichos datos permite insesgar los datos más recientes y obtener un conjunto más homogéneo de datos.

Se ha tenido ocasión de comprobar que ciertos valores extremos repetían un patrón de ocurrencia y que derivaban de los cambios horarios de verano y de invierno, problema que también se soluciona al operar con valores horarios UTC.

**4. Modelización – *ETS, ARIMA y PROPHET***

Con objeto de obtener resultados robustos, comparables y replicables se ha optado por modelizar la serie a través de modelo ETS, ARIMA y PROHET. El objetivo ha sido reducir el término de error, realizar predicciones con los tres modelos y obtener resultados comparables. La predicción horaria y diaria plantea la posibilidad de realizar predicciones usando los datos de igual frecuencia o, en el caso de la diaria, usar los datos con frecuencia horario acumulada para predecir los datos con frecuencia diaria. Los modelos presentados han sido modelados desde sendas arquitecturas, llegando a la conclusión de que la segunda ofrece mejores resultados debido a lo siguiente:

* Reducción de riesgos en el caso de valores missing y mayor facilidad de tratamiento
* Preprocesamiento de datos efectivo al tener una mayor variabilidad horaria
* Mejor ajuste como consecuencia de la mayor complejidad del modelo y la no pérdida de información (la predicción con datos diaria expulsa del modelo la estacionalidad diaria de los datos
* Facilidad de análisis en relación con el consumo social

De esta forma, tras la modelización los errores obtenidos con los datos de entrenamiento y validación en cada uno de los modelos han resultado los siguientes:

Tabla

Descripción generada automáticamente

Ilustración . ETS, ARIMA, PROPHET Errors. (Predicciones horarias)

Tabla

Descripción generada automáticamente

Ilustración .ETS, ARIMA, PROPHET Errors (Predicciones diarias)

Los errores que se muestran se han obtenido con la modelización de modelos con los mismos datos de entrenamiento y validación, arrojando los siguientes resultados en la predicción de validación.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración . Models Validation (ETS, ARIMA, Prophet) – Datos horarios

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración .Models Validation (ETS, ARIMA, Prophet) – Datos diarios

Una vez obtenido los datos de error y de validación, se ha decidido en base a los errores obtenidos, la arquitectura de los modelos y el tratamiento que se ha dado al dataset, utilizar en las predicciones generales la modelización en base a datos horarios. En la misma línea, se observa que el modelo prophet arroja mejores resultados de ajuste; no obstante se incluyen a continuación los modelos usados de forma más detallada.

**4.1 Modelo ETS**

La modelización ETS genera un modelo “MAdM”. La descomposición conforme a las componentes obtiene lo siguiente:

* **Error**: Multiplicativo (M)
* **Tendencia**: Additive damped (Ad). Reporta que los datos muestran un resultado exponencial de crecimiento, pero de forma ralentizada.
* **Estacionalidad**: Multiplicativo (M). La amplitud de la estacionalidad aumenta con el aumento de los datos.

De esta forma, los datos que se han incorporado generan un modelo con un componente estacional multiplicativo y con tendencia aditiva ralentizada. En otros términos, se llega a la conclusión que el crecimiento de la demanda de energía en el Estado de California es aditiva y ralentizada en el tiempo y que los patrones estacionales fijos que se repiten a lo largo del tiempo resultan multiplicativos. Por otro lado el ruido o las pequeñas variaciones existentes del resto de componentes y que no pueden ser explicadas presentan un carácter multiplicativo.

Finalmente, el modelo obtiene una predicción regular en el tiempo basada eminentemente en la estacionalidad media que marca la frecuencia diaria. Aunque el ajuste es bueno, descarta fluctuaciones derivadas de la estacionalidad semanal y otros componentes.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Predicción ETS Diciembre 2021

**4.2 Modelo ARIMA**

La modelización ARIMA parte de la premisa que los datos puedan tener relaciones dinámicas a lo largo del tiempo, donde los nuevos valores dependen de los valores de las observaciones anteriores. Esta dependencia está determinada por una componente autorregresiva que explica qué datos anteriores son predictivos y una parte de media móvil, que hace referencia a qué error de los datos anteriores actúan como predictores.

El test - análisis Dickey-Fuller había concluido que la serie resultaba estacionaria por lo que no se han incorporado modificaciones de la serie respecto a la misma.

Las relaciones no solo se hacen a nivel de dato consecutivo, sino también a nivel de la estacionalidad de la serie, por lo que al modelo se le añade una componente (p,d,q) que añade esto último.

De esta forma, la modelización reporta un modelo **SARIMAX(2, 0, 2)x(1, 0, [1], 24)** que reporta las siguientes concluisiones.

-El modelo en su parte **autorregresiva AR(2**) reporta que los datos horarios están relacionados de forma dinámica con el dato horario consecutivo anterior.

- El modelo no introduce diferencias en la predicción **I(0)**

- El modelo en su componente de media movial **MA(2)** reporta que los cálculos de predicción se van a basar en el ruido de los dos datos consecutivos anteriores.

- El modelo utiliza un componente estacional diario (24 datos consecutivos anteriores) con componentes (1,0,1). En otras palabras la estacionalidad diaria y su correspondiente dato de error o ruido se utilizan en la predicción

Con estos cálculos, las predicciones basadas en frecuencias horarias han sido las siguientes:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración -Predicción SARIMAX(1,0,2)(1,0,1)

Se observa como el modelo ha predicho de forma parecida al modelo ETS, con datos parecidos en cuanto a amplitud y estacionalidad.

**4.3 Modelo PROPHET**

La modelización con PROPHET es de tipo automático, generando un modelo propio de mayor complejidad que utiliza la tendencias y más de una componente estacional. A priori, conocíamos que los datos tenían una estacionalidad semanal, determinada por los días y otra diaria, determinada por las horas; el uso del modelo prophet ratifica esta idea con la descomposición siguiente y su inclusión en el modelo.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración .Descomposición Prophet

La automatización de la librería y su complejidad, aunque permite el ajuste de hiperparámetros se remite al código fuente y no procede a su inclusión en el presente informe por no considerarse explicativo.

Gráfico

Descripción generada automáticamente

Ilustración . Predicción Diciembre 2015 (Prophet)

Como puede observarse en la predicción, el modelo obtiene valores más oscilantes y acordes con la realidad del consumo de energía. Esta circunstancia queda aprobada también por la obtención en el caso del modelo prophet del mejor score de los modelos como se ha indicado en la tabla 2 y que pueda ser elegido como opción adecuada en la predicción.

**5. Predicciones**

Las predicciones para el mes de Diciembre conjugando los tres modelos han resultado las siguientes:

Gráfico, Histograma

Descripción generada automáticamente

Ilustración . Predicción Diciembre 2021(ETS, ARIMA, Prophet)

Acumulando los datos horarios que se han obtenido en datos diarios obtenemos datos que resultan constantes debido a la modelización que se ha obtenido en los datos diarios, focalizados en el ciclo diario. Se observa como el modelo Prophet, que combina la estacionalidad horaria y diaria, consigue identificar un patrón de ciclo semanal que coincide con los datos obtenidos de estacionalidad en el análisis exploratorio previo, recogiendo el consumo máximo en días laborales de semana

**Gráfico, Gráfico de líneas

Descripción generada automáticamente**

Ilustración . Predicciones diarias Prophet, ETS, ARIMA. Datos horarios acumulados

De esta forma se pueden representar los resultados comparándolos con los valores medios de cada día de la semana, concluyendo que el modelo ha logrado capturar las características propias de la serie temporal y ajustar de modo óptimo la predicción.

**Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente**

Ilustración . Comparación Predicción vs BoxPlot

Finalmente, incluimos también las predicciones diarias que se han obtenido con los datos diarios de entrenamiento, pudiéndose observar claramente como mientras el modelo prophet captura datos equiparables a los obtenidos con la predicción a través de los datos horarios, los modelos ETS y ARIMA mejoran su rendimiento.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

**6. Modelización con pycaret**

La biblioteca Pycaret ha sido creado recientemente e incluye más de 30 algoritmos de clásicos, de aprendizaje automático y de aprendizaje profundo que permite la modelización de series temporales. No es objeto de este proyecto focalizar la atención en el mismo, pero se incluye por su reciente lanzamiento (13 de noviembre de 2021) y para cumplir con el principio de innovación continua.

La biblioteca ofrece un primer test del conjunto del dataset preseleccionado que ofrece información sobre diferentes parámetros estacionariedad, ruido, normalidad, etc.

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

Ilustración . Biblioteca Pycaret Test

Se observa como las características previas que se habían obtenido a través del EDA del presente proyecto coincide con las obtenidas por pycaret, ofreciendo un grado mayor de funcionalidad.

De la misma forma, la biblioteca realiza un análisis de diferentes algoritmos de predicción, comparando los resultados conforme a sus errores.

**Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente**

Ilustración . Selección modelos pycaret

El paquete ha seleccionado como mejor estimador el modelo Seasonal Naive con un error MAPE 0.0369, siendo menor que los estimadores expuestos anteriormente y ofreciendo la siguiente predicción.

Gráfico

Descripción generada automáticamente con confianza media

Ilustración 21. Forecast S.Naive (Pycaret)

**7. Conclusiones**