Contenido de la actividad

**Enunciado**

Trabajamos como científicos de datos para una empresa de *retail* que, debido al cambio en los hábitos de consumo de los clientes, está potenciando ampliamente el servicio de venta *online*. La empresa quiere realizar un modelo de aprendizaje automático para clasificar a los clientes en función de la probabilidad de generar ingresos al comprar en la web. El objetivo es realizar una serie de acciones específicas para los clientes que es más probable que hagan compras en la web.

Para ello, la empresa ha ido recopilando datos con su herramienta de Google Analytics. Cuenta con datos de sesiones, cada una de ellas de un cliente distinto en un periodo de un año.

**Se pide**

**Tareas que nos pide la empresa:**

* Realización de un análisis de las variables del *dataset* de Google Analytics como pueden ser histogramas, *boxplots*, etc. Cualquier otro análisis es bienvenido, siempre explicándolo y con un sentido de negocio.
* Tratamiento de los valores faltantes, *outliers*, etc., en caso de que los hubiese. Si hay valores con *missings*, habrá que eliminarlos con el método de Pandas llamado *Dropna().a*.
* Tratamiento de categóricas, pasándolas a numéricas por medio de *dummies*, mapeándolas o utilizando un *label encoder*. Hay que justificar las operaciones que se realizan.
* Si existe alguna variable que se necesite borrar, habrá que borrara y justificarlo.
* Estandariza los datos.
* Dividir los datos en *train* y en test. Con los datos de *train* se pretende ajustar modelos con CrossValidation y GridSearch.
* Utilizar un modelo lineal. Entre los modelos lineales están las regresiones logísticas, las regresiones lineales, etc.
* Utilizar un modelo de redes neuronales.
* Utilizar cualquier otro modelo de clasificación.
* Optimizar algún parámetro de cada modelo utilizando CrossValidation y GridSearch, o de la forma que se estime oportuna, siempre justificándolo.
* Elegir el mejor modelo de los tres según la métrica ROC en CrossValidation. Predecir Test y obtener una métrica estimada.
* Umbralizar las probabilidades utilizando el umbral que maximice el área bajo la curva ROC.
* El entregable final será un Jupyter Notebook en el que se realicen todos los análisis y los modelos.

**NOTAS:**

Lo que se pide es un modelo *end-to-end* como los vistos en prácticas en el módulo. En este caso, hay que aplicar tres modelos y escoger el mejor, pero la parte previa es común a los tres.

Se tendrán en cuenta la estructura del código, las interpretaciones y las justificaciones.

El esquema que se pone en las tareas es un esquema general. Si se necesita hacer alguna tarea, se puede llevar a cabo. También es posible aplicar cualquier otro algoritmo visto en el módulo, siempre justificando y haciendo las cosas con un sentido.

**Información de los datos:**

El conjunto de datos consta de diez atributos numéricos y ocho categóricos.

* El atributo ***revenue*** puede ser usado como la etiqueta de la clase.
* ***Administrative***, ***administrative duration***, ***informational***, ***informational duration***, ***product related*** y ***product******related******duration*** representan el número de diferentes tipos de páginas visitadas por el visitante en esa sesión y el tiempo total dedicado a cada una de estas categorías de páginas. Los valores de estas características se derivan de la información del URL de las páginas visitadas por el usuario y se actualizan en tiempo real cuando el usuario realiza una acción, por ejemplo, pasar de una página a otra.
* Las características ***bounce rate***, *exit rate* y *page value* representan las métricas medidas por Google Analytics para cada página del sitio de comercio electrónico. El valor de la característica *bounce rate* de una página web se refiere al porcentaje de visitantes que entran en el sitio desde esa página y luego salen (*rebote*) sin activar ninguna otra solicitud al servidor de análisis durante esa sesión. El valor de la característica ***exit rate*** para una página web específica se calcula como para todas las visitas a la página, el porcentaje que fueron las últimas en la sesión. La función ***Page Value*** representa el valor medio de una página web que un usuario ha visitado antes de completar una transacción de comercio electrónico.
* La característica de ***special day*** indica la cercanía de la hora de visita del sitio a un día especial específico (por ejemplo, el Día de la Madre, San Valentín) en el que es más probable que las sesiones finalicen con una transacción. El valor de este atributo se determina teniendo en cuenta la dinámica del comercio electrónico, como la duración entre la fecha del pedido y la fecha de entrega. Por ejemplo, para San Valentín, este valor toma un valor distinto de 0 entre el 2 y el 12 de febrero, 0 antes y después de esta fecha a menos que esté cerca de otro día especial, y su valor máximo de 1 el 8 de febrero.
* El conjunto de datos también incluye el sistema operativo, el navegador, la región, el tipo de tráfico, el tipo de visitante como visitante que regresa o como nuevo visitante, un valor booleano que indica si la fecha de la visita es de fin de semana, y el mes del año.

Los datos se han obtenido de la siguiente web, aunque se pueden descargar de la plataforma de IMF Business School: <https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>