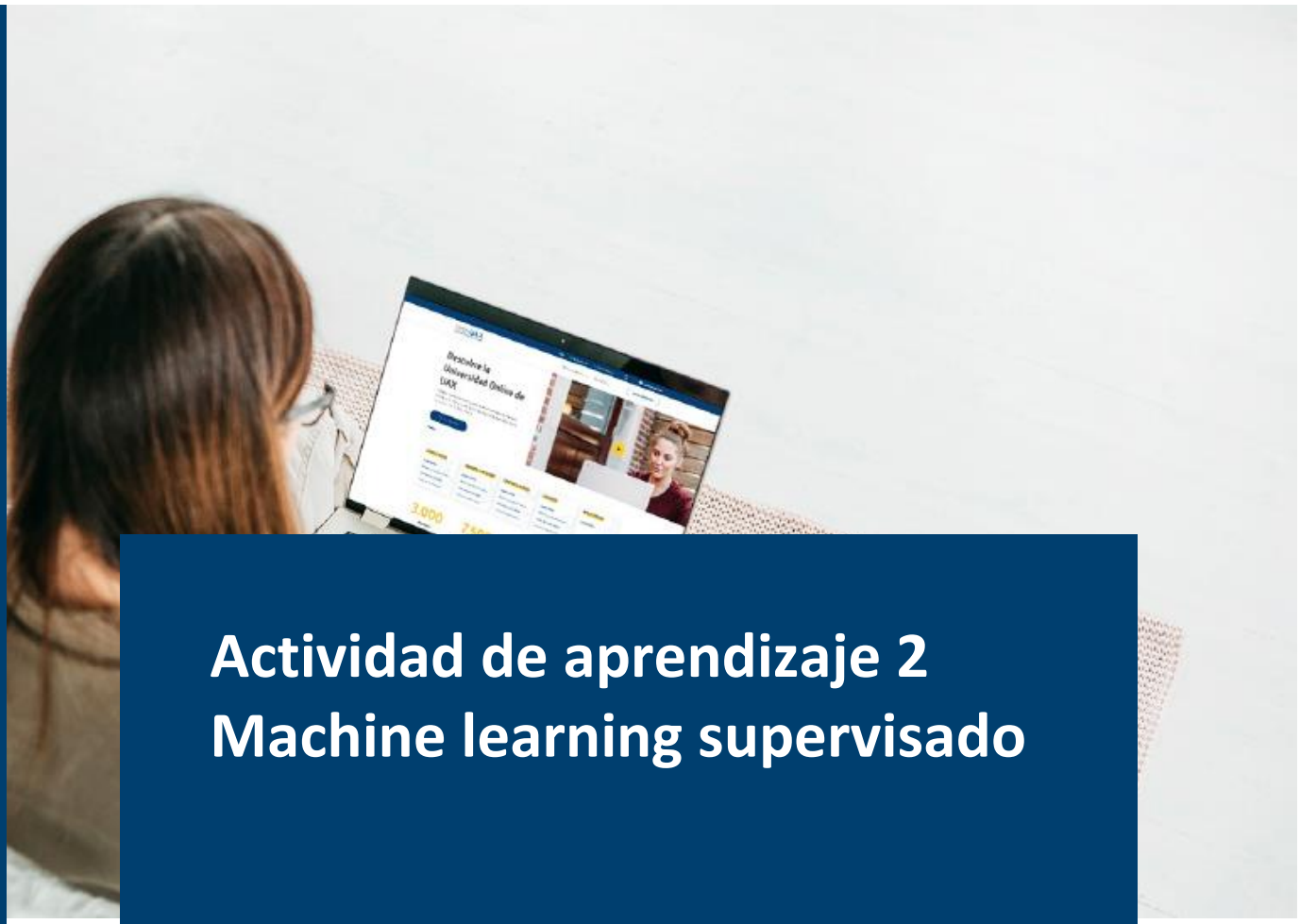

Actividad de aprendizaje 2
Machine learning supervisado

Aprendizaje Automático I

Para aplicar los conocimientos adquiridos durante la primera, segunda y tercera unidades didácticas, se propone esta actividad de aprendizaje. Deberá realizarse de forma **individual**.

1. Objetivo

Para poner a prueba los conocimientos adquiridos, vamos a practicar la formación de un dataset de entrenamiento para un problema de clasificación y a realizar un primer entrenamiento con el algoritmo de regresión logística. Para ello emplearemos dos datasets disponibles en el campus virtual. Uno de ellos contiene información relativa a los hoteles de una pequeña cadena hotelera (presente en España, Francia y Portugal) que considera que tiene un problema serio con las cancelaciones de reservas. El segundo dataset, es el histórico de reservas recibido por esta cadena durante los año 2016 y 2017, hasta la fecha en la que se quiere poner en producción, que es el **15 de junio de 2017**.

Para la resolución correcta, comprende el problema propuesto y sigue las instrucciones que se indicas en el apartado de "Enunciado" y entrega tu solución conforme a las instrucciones de entrega.

2. Enunciado

2.1. Contexto

La cancelación de reservas es uno de los principales desafíos operativos en la industria hotelera. Desarrollar un modelo predictivo de cancelaciones no solo ayuda a mejorar la eficiencia operativa y financiera de la cadena hotelera, sino que también permite una toma de decisiones más informada y centrada en el cliente, fortaleciendo su posición competitiva en el mercado.

Desde el punto de vista del negocio, predecir cancelaciones permite:

- **Optimización de Ingresos:**
  Con una predicción precisa, la cadena hotelera puede ajustar su estrategia de overbooking para maximizar la ocupación sin incurrir en excesos que perjudiquen la experiencia del cliente.
- **Reducción de Pérdidas y Costos Operativos:**
  Al anticipar cancelaciones, se pueden implementar medidas preventivas o promocionales para minimizar el impacto financiero, como ofrecer incentivos para la confirmación de reservas o reacomodar la capacidad de manera más eficiente.

En este curso, a lo largo de todas las tareas evaluables, se espera que, empleando las técnicas adquiridas en la asignatura poder predecir de manera anticipada qué clientes cancelarán su reserva en los **treinta días previos a la fecha de inicio de esta**.

Para ello debes entregar una memoria adjuntando toda la información, gráficas, explicaciones, extractos de tablas… que consideres relevante para que se comprenda fácilmente tu aproximación Además, deberás entregar archivos .py o .ipynb que contengan el código con el que se genere todo ese material de la memoria.

2.2. Explicación en detalle del problema

La cadena hotelera nos provee de dos datasets extraídos directamente de su base de datos. Básicamente su objetivo es proporcionarnos información de reservas de un subconjunto de sus 12 hoteles (los que llevan más tiempo abiertos), para utilizar esta información para predecir posibles cancelaciones en la temporada de verano en los hoteles que se han abierto durante el mismo año 2017. Es decir, NO tendremos información de reservas previas en los hoteles de nueva apertura, pero serán estos sobre los que tendrá que realizarse la predicción.

Posteriormente, a fecha 1 de septiembre de 2017, la compañía evaluará el total de aciertos que ha obtenido el modelo **durante los meses de julio y agosto**, ya sabiendo las cancelaciones que ha habido realmente.

El primero de los datasets ellos es una extracción **COMPLETA** de la tabla de hoteles, donde tenemos información estática de los 12 hoteles que integran el grupo a fecha **15 de junio de 2017**. A continuación se muestra una explicación de las variables que contiene este dataset:

Hotel_id	Id único que identifica al hotel a lo largo de la base de datos
Hotel_type	Tipo de hotel: ciudad o resort
Country	País en el que se encuentra el hotel
Parking	Indicador de si el hotel tiene instalación de parking o no
Total_rooms	Número total de habitaciones disponibles
Restaurant	Indicador de si el hotel tiene instalaciones de restauración o no
Pool_and_spa	Indicador de si el hotel tiene instalaciones de piscina y SPA o no
Avg_review	Valor promedio de reviews en Tripadvisor en la fecha indicada

El segundo dataset contiene un **HISTÓRICO PARCIAL** de reservas (de los hoteles antiguos). A continuación se muestra la explicación de las variables:

Hotel_id	Id único que identifica al hotel a lo largo de la base de datos
Board	Tipo de alojamiento: SC (soló dormir), BB (dormir + desayuno), HB (media pensión), FB (pensión completa).
Market_segment	Segmento al que pertenece el cliente. Hay varios valores como: grupos, cliente de agencia online, corporativo (cliente de negocios)…
Distribution_channel	Canal de venta de la reserva: TA/TO (agencia o tour operador), Direct (llamada o web del hotel), Corporate

	(empresa), GDS (un tipo de agencia de reservas global en tiempo real)
Room_type	Tipo de habitación con un código de letras
Required_car_parking_spaces	Indica el número de plazas de parking solicitadas por el cliente
Special_requests	Indica el número de peticiones especiales que se han hecho tras la reserva (cama adicional, bañera…)
Stay_nights	Número de noches reservadas
Rate	Precio pagado por el total de la reserva
Total_guests	Total de personas que se alojan en la reserva
Arrival_date	Fecha de inicio de la reserva (Check-in)
Booking_date	Fecha en la que se realizó la reserva
Reservation_status	Último estado conocido de la reserva: Check-Out (el cliente ha finalizado la estancia), No-Show (el cliente no se ha presentado), Booked (la reserva está realizada, esperando a a la fecha de estancia), Canceled (el cliente ha cancelado)
Reservation_status_date	Fecha del último estado de la reserva

2.3. Instrucciones

1) Elabora un dataset de entrenamiento con los archivos de datos disponibles en el campus virtual. La variable target será: “El cliente canceló la reserva en los últimos **TREINTA** días SÍ/NO”.
2) Puedes preseleccionar, cribar o hacer lo que consideres con las variables de entrada (pero mucho cuidado con cometer data leaking!).
3) Entrena una regresión logística, que usaremos como modelo baseline para las mejoras que realizarás a continuación y las siguientes entregas.
4) Entrena una regresión logística con Stochastic Gradient Descent y compara los resultados.
5) Utiliza todas las técnicas de mejora y debugging de modelos aprendidas en la UD 3:
    a. Técnicas de medición
    b. Train test split
    c. Regularización y cross validation
    d. Optimización de hiperparámetros.
    e. Desbalanceo de clases
    f. ...

En general el objetivo consistirá en mejorar el modelo baseline de regresión logística hasta tener un clasificador más robusto.

Deberás utilizar Sklearn pipelines para concatenar todos los pasos de preprocesado y entrenamiento.

Finalmente, prepara un Dockerfile para realizar inferencia batch sobre una muestra de datos desconocida. Para ello debes serializar el modelo entrenado y ser muy cuidadoso con los datos de entrada y hacer un Dockerfile lo más generalizable posible. Recuerda que debe funcionar en cualquier ordenador a partir de una muestra del dataset original.

3. Instrucciones de entrega

- **Extensión**: El fichero en formato .zip, a adjuntar, deberá contener **la memoria** .pdf, el código Python, el Dockerfile y el modelo serializado.
- **Nombre del fichero**: Actividad2_Nombre_Apellido1_Apellido2
- **Formato de entrega**: Seguir las instrucciones y fechas especificadas en el aula virtual.
- **El plagio de fuentes no referenciadas o entre compañeros significará el suspenso en la actividad de aprendizaje**.

4. Evaluación

- (3 Puntos) En la evaluación se tendrán en cuenta los aspectos conceptuales asociados a un problema de clasificación. Penalizará mucho cometer errores conceptuales graves
- (3 Puntos) Se valorará que la memoria sea generosa y autoexplicativa de qué decisiones se han tomado y por qué. Hay que guiar al lector a lo largo de los experimentos que has ido realizando de cara a llegar al modelo final
- (2 Puntos) Se valorará la portabilidad del sistema de inferencia. Debe funcionar en mi ordenador sin ningún tipo de ajuste adicional
- (3 Puntos) Parte de la calificación procederá de un ranking de la métrica f1-score sobre un dataset de test desconocido para vosotros.

UAX Universidad Alfonso X el Sabio

G R A C I A S

UAX.COM