Universidad de Navarra | DATAI INSTITUTE OF DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

MÓDULO DE MACHINE LEARNING - PROYECTO

Se trata de un trabajo práctico en el cual el alumno pondrá en práctica los conocimientos aprendidos durante el módulo.

El proyecto se debe realizar en equipos de 3 a 5 personas. Los alumnos crearán los grupos, excepto si alguien se queda solo, en tal caso el profesor creará los grupos. Si alguien tiene alguna circunstancia personal y necesita hacerlo en solitario que contacte al profesor.

No se podrá usar ningún dato que no sea proporcionado por el profesor (nada de datos externos). Sí se podrán usar datos generados a partir de los proporcionados (transformaciones, feature engineering…)

Uno puede aprender de ejemplos en internet, ayudarse de manuales o blogs, ver cursos (i.e. Coursera),… para aprender y resolver el problema. Pero… todos los miembros del grupo obtendrán una calificación de 0 en la nota final del módulo cuando alguno de sus miembros copie, facilite copiar o plagie el trabajo de otros compañeros ajenos al grupo, antiguos alumnos o terceras personas. Se revisará el plagio de texto y código.

EL PROBLEMA:

Una empresa pone a disposición de sus clientes 25 productos distintos. Los clientes pueden contratar cualquiera de los productos, pero solo pueden tener un producto activo de cada tipología. Es decir:

- Un cliente puede tener activos los productos *"1", "2", "6"* y *"22"* pero no puede tener activos dos o más productos- Por ejemplo, no puede tener activos dos productos *"6"*.
- Un cliente como máximo puede tener activos 25 productos, es decir, uno por tipología.
- Un cliente como mínimo puede tener activos 0 productos.

Los productos se contratan para un periodo de un mes y los contratos se puede renovar de forma indefinida cada mes.

Se comparte una serie de datos que hacen referencia a las características asociadas a los clientes (*features*) y 25 columnas que nos indican si tienen activo o no un producto (*targets*) para cada mes. Un cliente puede hacer tres cosas con cada producto:

- **Contratar** el producto, si el cliente no lo tenía contratado el mes anterior.
- **Mantener** el producto, si el cliente ya tenía contratado el producto el mes anterior y lo mantiene el mes en curso.
- **Cancelar** el producto, si el cliente tenía contratado el producto el mes anterior y cancela el producto el mes en curso.

Tenéis que predecir los productos que contratarán los clientes (clientes existentes en el último mes del dato) el próximo mes (asumiendo que todos seguirán siendo clientes). Predecir los productos que se van a contratar no es lo mismo que predecir los productos que se van a usar el próximo mes. Los productos que los clientes usarán el próximo mes serán aquellos que se mantienen o se contratan, y lo que buscamos es predecir estos últimos. No queremos predecir que productos ya contratados en un mes anterior y que los clientes continúan usando.

En definitiva, el proyecto consiste en analizar, explicar, modelar y predecir la contratación de 25 productos por parte de cada cliente. La métrica de error que se usará es el *Mean Average Precision at 7* también llamado *MAP@7*. Se valorará la calidad del informe, las técnicas usadas, la claridad, el ingenio (feature engineering, visualizaciones…) y las métricas de error obtenidas en las predicciones a realizar.

ENTREGABLES:

INFORME: Entregar un informe en el cual cumpláis con los objetivos/contenidos planteados en no más de 10 caras (*Times New Roman* y tamaño 10). Junto a estas 10 caras, podéis añadir tantas hojas como necesitéis para las gráficas/imágenes que uséis en los anexos. Todas las gráficas/imágenes que mencionéis en las 10 primeras caras deberán estar enumeradas para facilitar la lectura y poder localizarlas en el anexo. Objetivos/contenidos planteados:

- **Resumen:** Describir brevemente los pasos, hipótesis y el proceso que habéis seguido (se recomienda redactarlo una vez terminado el trabajo).
- **Exploración:** Hacer un breve análisis exploratorio de los datos destacando los aspectos más importantes que encontréis y merezcan la pena ser contados.
- **Tratamiento del dato:** Explicar que técnicas de limpieza, imputación, transformación, gestión de las dimensiones, feature engineering... de los datos se han realizado y el motivo.
- **Modelado:** Explicar que modelo (o modelos) y recursos (entrenamiento, selección de variables, evaluación...) usáis y su justificación de cara al modelado del dato. Feature engineering puede aparecer en este apartado o en el de "Tratamiento del dato". Comentar que se puede hacer uso de técnicas de Deep Learning.
- **Conclusiones, mejoras y sugerencias:** Mejores insights obtenidos. También se valorará aportar ideas accionables por parte del negocio para incrementar ventas, evitar la fuga de clientes, predecir ciertos patrones…

PREDICCIONES: Entregar un fichero que csv con la misma estructura que el csv "**ejemplo_predicciones.csv**". Revisad que el formato de las columnas son los mismos que en el ejemplo, que se usa punto y coma (;) para separar los elementos y punto (.) para los decimales. Solo se acepta una predicción por equipo.

CÓDIGO: Tendréis que entregar un notebook en el que aparezca la exploración, transformaciones, predicciones… realizadas. El notebook entregado se debe poder ejecutar de principio a fin de tal forma que solo se tenga que cambiar el directorio del dato, obteniendo los mismos resultados, gráficas y predicciones. No es necesario adjuntar los notebooks exploratorios o de pruebas fallidas.

FICHEROS DE LOS QUE DISPONÉIS

- **Dataset para modelar**: Se os compartirá un *dataset* de entrenamiento (**dataset_para_modelar.csv**) que contiene tanto input como output. Sobre este conjunto de datos realizaréis el trabajo. Tendréis que predecir los productos que se contratarán el mes siguiente por parte de los clientes existentes en el último mes del dato.
- **Dataset ejemplo predicciones:** Ejemplo de un dataset de predicciones que tenéis que entregar "**ejemplo_predicciones.csv**". Existe una columna por producto, para que escribáis el score que predecís para cada producto (opcional, pero ayuda a entender vuestro trabajo. La última columna, obligatoria, llamada "predicted" contiene una lista con los artículos que finalmente decidís recomendar al cliente.
- **Dataset ejemplo de soluciones**: Ejemplo para jugar y entender el MAP@K junto con el dataset de ejemplo de predicciones. "**ejemplo_soluciones.csv**"

¿CÓMO SE CALCULA EL MAP@K?

Ver fichero adjunto "_map@k en Python.py"

DESCRIPCIÓN DE LOS CAMPOS:

COLUMN NAME	DESCRIPTION
cod_persona	Customer code
mes	The table is partitioned for this column
pais	Customer's Country residence
sexo	Customer's sex
edad	Customer's age
fecha1	The date in which the customer became as the first holder of a contract in the bank
xti_empleado	Employee index: A active, B ex employed, F filial, N not employee
xti_nuevo_cliente	New customer Index. 1 if the customer registered in the last 6 months.
num_antiguedad	Customer seniority (in months)
xti_rel	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
fec_ult_cli_1t	Last date as primary customer (if he isn't at the end of the month)
xti_rel_1mes	Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)
tip_rel_1mes	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
indresi	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
indext	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
des_canal	Channel used by the customer to join
xti_extra	Deceased index. N/S
tip_dom	Addres type. 1, primary address
cod_provincia	Province code (customer's address)
xti_actividad_cliente	Activity index (1, active customer; 0, inactive customer)
imp_renta	Gross income of the household
id_segmento	Segmentation: 01 - VIP, 02 - Individuals 03 - college graduated
mean_engagement	Mean customer engagement
ind_prod1	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod2	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod3	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod4	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod5	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod6	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod7	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)

COLUMN NAME	DESCRIPTION
ind_prod8	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod9	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod10	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod11	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod12	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod13	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod14	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod15	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod16	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod17	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod18	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod19	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod20	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod21	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod22	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod23	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod24	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)
ind_prod25	1 (customer uses the product this month) 0 (the customer doesn't use the product this month)