10.04 Plantas Industriales
Segundo Cuatrimestre 2021
Taller de Machine Learning

Consigna

El presente trabajo práctico tiene la finalidad de evaluar la temática de Machine Learning vista en el primer módulo de la asignatura. Para esto se presenta una problemática de un caso de estudio y se suministran los datos que caracterizan y dan contexto a dicho caso. El objetivo es que los estudiantes propongan una solución a la problemática siguiendo la metodología de análisis y resolución de problemas basados en datos vista en clase, dicha solución debe estar encaminada hacia la obtención de un modelo de Machine Learning que sirva como herramienta para responder de manera adecuada a las necesidades del caso de estudio.

Para la realización del taller se deben tener en cuenta las siguientes consideraciones:

- Los alumnos deben trabajar en los equipos que conformaron al inicio del cuatrimestre.
- Se brindan dos (2) conjuntos de datos, uno que contiene las variables características y la variable de respuesta, el cual debe usarse para obtener el modelo; y otro del cual solo se conocen las variables características y cuya variable de respuesta debe predecirse utilizando el modelo obtenido.
- Los valores reales de la variable de respuesta para el segundo conjunto de datos son conocidos por la cátedra, los cuales serán comparados con las predicciones del modelo de los estudiantes para obtener así el desempeño final del modelo propuesto.
- Cada grupo debe entregar un informe en donde se explique detalladamente el proceso metodológico que se siguió para la obtención del modelo final elegido.
- La nota final del taller dependerá tanto del desempeño del modelo como de la calidad del informe entregado. Tener en cuenta que los errores de ortografía y gramática afectan negativamente la calidad del informe.

Caso

The Online Commerce (TOC.com) es una empresa de retail en línea (e-commerce) cuyo modelo de negocio se sustenta sobre una plataforma que opera únicamente de manera digital, lo que quiere decir que esta tienda no cuenta con ninguna sucursal para atención presencial y todas sus operaciones se concretan a través de internet.

A través de dicha plataforma, alojada en su sitio web, los clientes pueden acceder a las publicaciones de los diferentes artículos ofrecidos e interactuar con estas de diversas maneras. Por ejemplo, dentro de las acciones que pueden realizarse están leer la descripción de la publicación, comprar el producto, leer opiniones de otros compradores y publicar una propia sobre el producto ofrecido, hacer preguntas a la tienda y calificar el producto entre otras.

Recientemente, como producto del “boom” de la virtualidad impulsado por el contexto de la nueva normalidad han surgido una gran cantidad de pequeñas empresas emergentes cuyos modelos de negocio son similares al de TOC.com, es decir, giran alrededor de una aplicación para dispositivos electrónicos que tiene la finalidad de prestar algún servicio directamente por internet. Muchas de estas empresas, comúnmente llamadas “start-ups”, han empezado a operar en el segmento del retail en línea y por lo tanto, representan una competencia directa para TOC.com.

Debido a lo anterior, se ha creado un entorno de alta competitividad en donde las empresas del mercado se disputan por diferenciarse del resto para resultar atractivas y capturar a los clientes. Como resultado, el equipo de TOC.com ha venido observando una disminución en su base de usuarios y desea realizar un análisis que les permita obtener visibilidad e identificar clientes que puedan estar propensos a dejar de utilizar su plataforma en el corto plazo.

La idea del negocio es obtener una herramienta basada en un modelo de ML que sea capaz de identificar a tiempo los clientes que representan potenciales bajas, de manera que se puedan ejecutar acciones sobre estos que permitan retenerlos y evitar que dejen de usar la plataforma de TOC.com para irse a alguna de las de la competencia.

Para construir este modelo se cuenta con un set de datos que incluye información variada de los clientes de la empresa y en el cual se incluye una columna que informa si el cliente en cuestión dejó de utilizar la plataforma o no. En dicho set de datos cada fila representa un

cliente y cada columna es una variable que representa algún atributo o característica del cliente.

A continuación, se presenta un breve resumen de cada una de las variables (columnas) características que componen el dataset, las cuales están orientadas a describir el comportamiento de los clientes de TOC.com con la finalidad de explicar el hecho de que estos puedan o no dejar de usar la plataforma:

Variable	Descripción
CustomerID	ID del cliente
Churn	Columna que indica si el cliente dejó de usar la plataforma o no
CustomerTenure	Es el tiempo transcurrido desde el inicio de la relación con el cliente (en meses)
MainDeviceLogin	Dispositivo principal que utiliza el cliente para acceder a la plataforma
CityTier	Indicador del nivel de desarrollo de la ciudad donde vive el cliente
WarehouseToHome	Distancia desde el centro de distribución a la vivienda del cliente (en km)
MainPaymentMode	Método de pago más utilizado por el cliente
Gender	Género del cliente
HourSpendOnApp	Número de horas que el cliente ha pasado en la plataforma
DeviceRegistered	Número de dispositivos en los que el cliente ha accedido a la plataforma
PrefCategory	Categoría más común de las compras del cliente en el último mes
SatisfactionScore	Nivel de satisfacción del cliente con el servicio
MaritalStatus	Estado civil del cliente
NumberOfAddress	Número de direcciones diferentes registradas por el cliente
Complain	Si ha realizado reclamos
OrderAmountHikeFromlastYear	Incremento porcentual en la cantidad de compras con respecto al año anterior
CouponUsed	Número de cupones usados en el último mes
OrderCount	Número de compras realizadas en el último mes
DaySinceLastOrder	Cantidad de días desde la última compra
CashbackAmount	Promedio de reembolsos pedidos en el último mes

Puntos para desarrollar en el informe

1) Describa sus hipótesis iniciales respecto al problema antes de ver los datos.

2) Comente brevemente los resultados obtenidos a raíz del análisis exploratorio de datos.

3) Comente, si corresponde, acerca de los procesos de limpieza de datos ejecutados.

4) Enumere y describa brevemente, si corresponde, las nuevas variables creadas para el modelo.

5) ¿Cuál es la métrica que considera más adecuada para evaluar la performance del modelo? Justifique debidamente la elección realizada.

a) Accuracy

b) Precisión

c) Sensibilidad

d) Especificidad

e) F1 Score

6) Describa brevemente su mejor modelo, el proceso que se siguió para obtenerlo y los diferentes experimentos realizados en el camino. Finalmente, presente una tabla resumen en donde se muestren los valores de las métricas listadas en el inciso anterior para las diferentes instancias de evaluación que hayan tenido lugar para la elección de dicho mejor modelo.

7) Adjunte como resultado las predicciones hechas por su modelo para la variable de respuesta del segundo dataset. Para esto se debe enviar un archivo que contenga dos (2) columnas, la primera correponderá al ID del cliente y la segunda será la predicción realizada, la cual debe ser 0 si se espera que el cliente siga utilizando la plataforma y 1 si este es identificado como una posible baja.

Consideraciones adicionales

Los archivos que deberán ser entregados serán los siguientes:

- Informe (en formato PDF)
- Archivo con las predicciones realizadas.
- Todos los inputs que utiliza su proyecto de RapidMiner con el formato adecuado.
- El proyecto completo de Rapidminer (los dos archivos necesarios para poder abrir el RapidMiner en formato .rmp)

Ejemplo:

- Errores de ortografía y gramática serán considerados en la nota final.