TD VI: Inteligencia Artificial

Trabajo práctico 2 (2023 2do semestre)

En este trabajo práctico, crearán un sistema con el propósito de **predecir la probabilidad de que un usuario** que interactúa con un anuncio específico de un destacado *retailer online* **lleve a cabo la compra del producto anunciado**. Esto lo harán entrenando modelos de aprendizaje automático sobre la base de datos de avisos previamente visitados (algunos de los cuales fueron comprados y otros no).

El trabajo práctico se deberá resolver de a grupos de **3 integrantes**.

La **evaluación** tendrá en cuenta **3 componentes** (más sobre esto abajo): 1) el resultado del sistema en el *leaderboard* privado, 2) la calidad del informe entregado y 3) la claridad del código entregado.

Los sistemas propuestos por los diferentes grupos competirán a través de la plataforma **Kaggle**. El *link* para acceder y registrarse a la competencia se encuentra disponible en el Campus Virtual de la materia.

En el Campus también **cuentan con** un link para descargar los siguientes tres **archivos**:

- `competition_data.csv`: contiene tanto los datos de entrenamiento como los de evaluación. Para el caso de los datos de evaluación, la columna `ROW_ID` **no tiene** valores *missings* (para el caso de los datos de entrenamiento, vale siempre missing).
- `sample_submission.csv`: es un archivo CSV que sigue el formato del archivo que deben subir en la competencia.
- `basic_model.py`: es un *script* básico que genera el *benchmark* de la competencia. Concretamente, crea un archivo para subir a Kaggle (`basic_submission.csv`). Este script tiene un desempeño que debe mejorarse; la idea es que todos los grupos superen ampliamente su *performance*.

En la página de la competencia, se indica qué es cada una de las **variables** que contiene el *dataset* provisto. Allí, también se indican las **reglas** de la competencia. A saber:

- La métrica de evaluación será **ROC-AUC**.
- Un 30% elegido al azar de los datos de evaluación da lugar al puntaje del leaderboard público. Este valor les servirá de guía para evaluar su desempeño, pero la **evaluación final** se realizará sobre el restante 70%. Esta evaluación final corresponde al leaderboard privado, que podrán ver una vez cerrada la competencia.
- Cada grupo podrá realizar, a lo sumo, **3 *submits* diarios** (¡no dejen todo para último momento!).

<u>Criterios de evaluación del TP</u>

1) Performance en el **leaderboard privado** (**30%** de la nota final). Las soluciones propuestas deben alcanzar una buena performance. Esto implica superar ampliamente el benchmark propuesto y no quedar excesivamente debajo (en términos de AUC) de aquellos grupos que tengan la mejor performance.

   IMPORTANTE:

   a) Se **penalizará** a aquellos grupos que hagan pocos submits. Deberán realizar, al menos, 15 submits a lo largo de la competencia; idealmente, espaciados en el tiempo.

   b) Para el miércoles **30 de agosto** a las 23:59:59, cada equipo debe haber realizado un **primer submit**. No importa que el mismo tenga una mala performance; de hecho, puede corresponder al **mismo** benchmark provisto. El objetivo de este submit es que, para dicha

fecha, ya se encuentren **familiarizados** con el funcionamiento de Kaggle. Esto es condición necesaria para aprobar el TP.

c) Para el miércoles **6 de septiembre** a las 23:59:59, cada equipo debe haber realizado, al menos, un **submit** cuya performance sea un **5% mayor** a la del benchmark provisto. Esto es condición necesaria para aprobar el TP.

2) **Informe** que presente el sistema propuesto (**45%** de la nota final). El mismo no debe tener más que 3 carillas. Debe contener, como mínimo, las siguientes **secciones**:

a) Una sección de **análisis exploratorio** de datos que contenga **dos figuras** que permitan ver patrones interesantes de los datos.

b) Una sección que cuente qué variables armaron. Aclaración: crear **variables adicionales** (ya sea a partir de las variables que contiene el dataset provisto u otras fuentes) es algo que se valorará positivamente. Para cada variable adicional, indicar a partir de qué otra variable o fuente se creó.

c) Una sección que explique cómo armaron el **conjunto de validación** que utilizaron para entrenar el modelo. El criterio para hacer esto debe estar bien **justificado** (es decir, deben explicar por qué armaron el conjunto de validación de la forma en que lo hicieron).

regresion logistica -> pvalores

d) Una sección que explique qué **modelo(s) predictivo(s)** usaron y cómo buscaron los mejores **hiperparámetros** del / de los mismo(s). **Justificar** estas dos decisiones. En caso de haber usado más de un modelo predictivo o más de un método de búsqueda de hiperparámetros, indicar con cuál de ellos obtuvieron la mejor performance. grid search, random search, etc

árboles, xgboost, random forest -> importancia de atributos

e) Una sección que analice, sólo para su solución final, qué **atributos** resultaron ser más **importantes** o significativos. En esta sección, deben incluir las respuestas a las siguientes dos preguntas: dada una persona que se encuentra diseñando un anuncio de venta de un producto para publicar en este destacado retailer online, ¿en qué aspectos le **recomendarían** enfocarse? ¿Ven alguna **debilidad** en este análisis?

3) **Código** que lleve adelante todo lo presentado en el informe y genere la solución final propuesta (**25%** de la nota final). Se deberá entregar un único script de Python que lleve adelante todo lo presentado en el informe (gráficos, creación de variables, selección de modelo, etc.). El mismo debe ejecutarse de punta a punta sin errores y debe ser claro y legible para una persona ajena al grupo.

Fechas y modalidad de entrega

- Se podrán realizar **submits** en Kaggle hasta el miércoles **4 de octubre** (inclusive). Pasado ese momento, se cerrará la competencia y se harán públicos los *scores* del leaderboard privado.
- El **informe** y el **código** podrán entregarlos hasta el domingo **8 de octubre** a las 23:59:00.
- El **informe** debe ser entregado en formato **PDF**.
- Los archivos correspondientes al informe y al código deben ponerse dentro de una carpeta llamada `tp2-gxx`, donde `xx` sea reemplazado por el número de grupo; por ejemplo, `tp2-g01`. La versión ZIP de esta carpeta debe subirse a la tarea llamada **TP2 | Entrega** en la página de la materia en el Campus Virtual.
- **Sólo 1** integrante del grupo debe realizar la **entrega**.
- A modo de **resumen**, las principales fechas a tener en mente son las siguientes (recomendamos altamente ir revisando la siguiente *checklist*):
  - ☐ Miércoles 30 de agosto | Primer submit.
  - ☐ Miércoles 6 de septiembre | Submit +5%.
  - ☐ Miércoles 4 de octubre | Último submit.

- [ ] Domingo 8 de octubre | Informe y código.

- Recomendación: **¡no lo dejen para último momento!**
- Para realizar sus consultas sobre este TP, cuentan con el foro llamado **TP2 | Consultas** en la página de la materia en el Campus Virtual. Todas las dudas que surjan en relación al TP2 envíenlas exclusivamente a este foro; no usen ningún otro. Esto debe ser así porque este es el foro que está configurado de forma que los mensajes enviados lleguen únicamente al cuerpo docente y a sus compañeros de grupo. En otras palabras, si un integrante de un grupo envía una pregunta por acá, tanto esa pregunta como la respuesta, luego dada por el cuerpo docente, podrán ser vistas sólo por los integrantes del grupo en cuestión.

Tablero publico solo el 30% de los datos y 70% en el tablero privado.

No hacer overfitting de tablero publico

Hacer un buen conjunto de validación

La nota se va a basar sobre el tablero privado

Tenemos que comparar el public score con el validation set score, si son parecidos, tenemos un buen validation set