Intro a Data Science - Taller 4

El objetivo de este taller es, sobre la base de datos de avisos visitados en Mercado Libre, armar un sistema que prediga si un usuario comprará el producto. El taller se podrá resolver de a grupos de 3 o menos personas. La evaluación de este taller tendrá en cuenta 3 componentes (más sobre esto abajo): 1) el resultado del sistema en el leaderboard privado, 2) la calidad del informe entregado, 3) la claridad del código entregado.

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

En la página de la competencia se pueden descargar los siguientes archivos:

- *data.RDS*: 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.
- *basic_model.R*: es un script básico que genera el benchmark de la competencia. Crea un archivo para subir a Kaggle. (*basic_submission.csv*). Este script tiene un desempeño que puede mejorarse, la idea es que todos los grupos deben superar **ampliamente** su performance.

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

- La métrica de evaluación será 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%.
- Cada grupo podrá realizar a lo sumo 3 submits diarios (no dejen todo para último momento).

Criterio de evaluación del taller:

1) Performance en el leaderboard **privado** (30% de la nota): las soluciones propuestas deben alcanzar una performance buena. Esto implica superar ampliamente el benchmark propuesto y no quedar excesivamente debajo de aquellos grupos que tengan la mejor performance. **IMPORTANTE**:
   a. Se penalizará a aquellos grupos que hagan pocos submits.
   b. **Antes del 7 de junio** cada equipo debe haber realizado un primer submit (no importa que el mismo tenga una mala performance).

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:

- **o** Una sección de análisis exploratorio de datos que contenga dos gráficos de ggplot que permitan ver patrones interesantes de los datos.
- **o** Una sección que cuente qué variables armaron. Nota: crear variables adicionales es algo que se valorará positivamente.
- **o** 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.
- o Una sección que explique qué modelo predictivo usaron y cómo buscaron los mejores hiperparámetros del mismo.

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 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 tercera persona ajena al grupo.

Fecha y modalidad de entrega

Se podrán realizar submits en Kaggle **hasta el 21 de junio** (huso horario UTC), momento en que se cerrará la competencia y se harán públicos los scores del leaderboard privado. El informe y código deberá entregarse **hasta el 26 de junio**, y deberá ser enviado por correo electrónico a *introdsutdt@gmail.com* indicando los nombres y legajos de los integrantes del grupo.