Maestría en Ciencia de Datos

Data Mining en Economía y Finanzas

Comisión Jueves

Competencia Kaggle

DMEyF 2024 Primera

Robinson Galvis Marin

robins_gama@mail.com

UBA

Instructivo para la ejecución y carga de la primera competencia en Kaggle

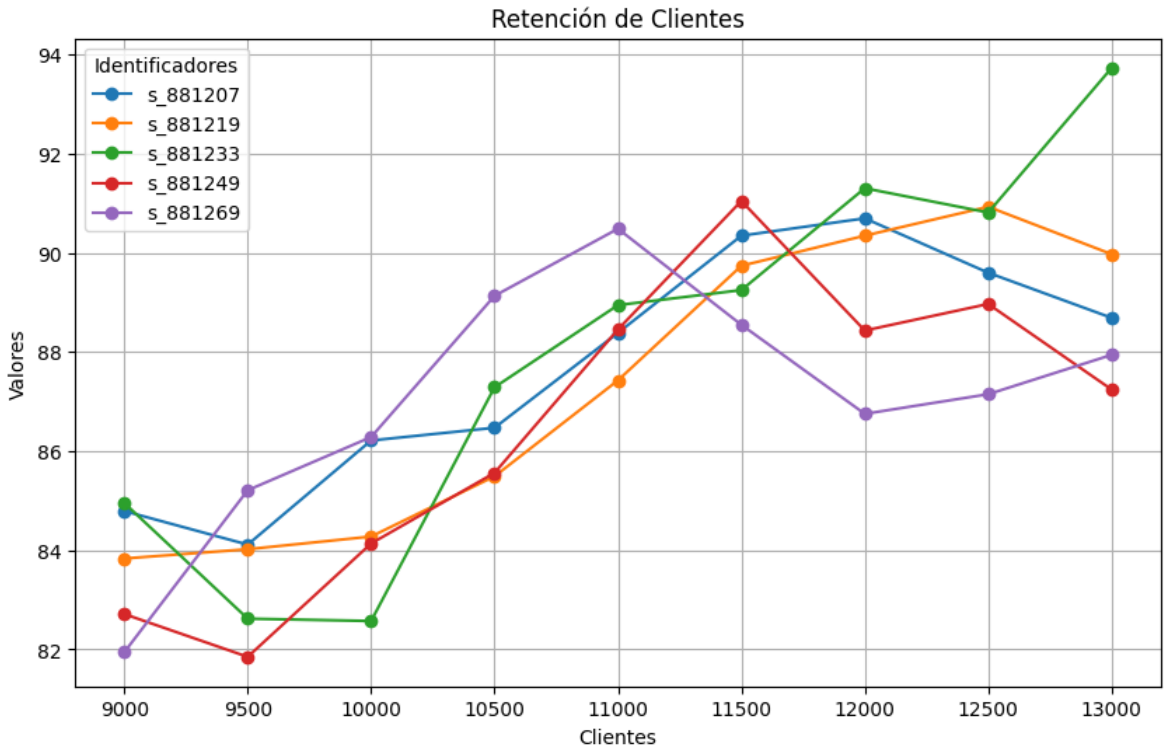
Se deben seguir los siguientes pasos, con el fin de llevar a cabo la correcta ejecución, predicción y carga del archivo en Kaggle.

1. Se debe contar con el dataset inicial, llamado **competencia_01_crudo.csv**.
2. Ejecutar en Python el script llamado **script_clase_ternaria.ipynb** para la obtención del atributo **clase_ternaria**.
    a. Aquí se debe ajustar la ruta en la que se dejó el archivo indicado anteriormente.
    b. Con este script se calcula la clase ternaria y se genera el archivo **competencia_01.csv**, el cual se utilizará para los pasos posteriores.
3. **Opcional:** Ejecución de la optimización Bayesiana:
    a. Cargar en R el siguiente script **422_lightgbm_binaria_BO.r**
    b. La ejecución de este script se realiza con la semilla 881269.
    c. Los hiperparámetros utilizados en esta ejecución, con sus respectivos rangos, son los siguientes:
        i. *("learning_rate", lower = 0.01, upper = 0.1)*
        ii. *("num_leaves", lower = 8L, upper = 1024L)*
        iii. *("feature_fraction", lower = 0.4, upper = 1.0)*
        iv. *("min_data_in_leaf", lower = 1L, upper = 3000L)*
        v. *("envios", lower = 5000L, upper = 13000L)*
        vi. *("max_depth", lower = 7L, upper = 15L)*
        vii. *("lambda_l1", lower = 0.0, upper = 10.0)*
        viii. *("lambda_l2", lower = 0.0, upper = 10.0)*
    d. Conservando la estructura indicada en la cátedra para el manejo del repositorio, la ejecución de este script creará una subcarpeta **HT4220** en la carpeta **exp**.
    e. Aquí se genera el archivo **HT4220.txt**, el cual se debe acceder para ordenar los datos por el campo **ganancia** de manera descendente.
    f. Seleccionar los hiperparámetros de la mayor ganancia.
    g. Los hiperparámetros seleccionados son:
        i. *PARAM$finalmodel$num_iterations <- 1425*
        ii. *PARAM$finalmodel$learning_rate <- 0.01*
        iii. *PARAM$finalmodel$feature_fraction <- 0.624*
        iv. *PARAM$finalmodel$min_data_in_leaf <- 1347*
        v. *PARAM$finalmodel$num_leaves <- 674*
        vi. *PARAM$finalmodel$max_depth <- 10*
        vii. *PARAM$finalmodel$lambda_l1 <- 7*
        viii. *PARAM$finalmodel$lambda_l2 <- 1*

4. Cargar y ejecutar en R el script **421_lightgbm_final.r**.
    a. Con la ejecución de este script, la carga de los archivos a Kaggle y la validación de los resultados obtenidos, el mejor resultado es:
        i. Semilla: 881269
        ii. Número de clientes a retener: 12000
        iii. Public score: 86,751
    b. Para la validación de los datos obtenidos anteriormente, los siguientes pasos son opcionales (incluyendo el paso 5)
    c. Con la optimización de los hiperparámetros, se realiza la ejecución inicial de este script con la semilla **881269**.
    d. Conservando la estructura indicada en la cátedra para el manejo del repositorio, la ejecución de este script creará una subcarpeta **KA4210** en la carpeta **exp**.
    e. Dentro de esta carpeta se genera una serie de archivos que indica el número de clientes a los cuales se les realizará la retención. La partición del número de clientes va entre 9.000 y 13.000 en deltas de 500.
    f. Cargar en Kaggle cada uno de estos archivos generados.
    g. Repetir los pasos a., b. y c. para cada una de las siguientes semillas:
        i. 881249
        ii. 881233
        iii. 881219
        iv. 881207
5. En Python, ejecutar el script **script_selección_semilla.ipynb**, en el cual se carga la siguiente tabla y nos brinda la gráfica para la selección del score final (indicado en el ítem 4.a)

	s_881207	s_881219	s_881233	s_881249	s_881269
9000	84,791	83,834	84,954	82,714	81,944
9500	84,114	84,021	82,621	81,851	85,211
10000	86,214	84,277	82,574	84,137	86,284
10500	86,471	85,491	87,287	85,561	89,131
11000	88,384	87,427	88,944	88,454	90,484
11500	90,344	89,737	89,247	91,044	88,547
12000	90,694	90,344	91,301	88431	86,751
12500	89,597	90,927	90,811	88,967	87,147
13000	88,687	89,971	93,727	87,241	87,941

    a. Gráfica obtenida


Retención de Clientes
Identificadores
s_881207
s_881219
s_881233
s_881249
s_881269
Valores
Clientes
94
92
90
88
86
84
82
9000
9500
10000
10500
11000
11500
12000
12500
13000