Churn Prediction Challenge

El objetivo de la competencia es predecir cuáles de los usuario que jugaron al juego *Castle Crush* al menos una vez durante el tercer día después de haberlo instalado, no volverán a jugarlo nunca más. Castle Crush es un popular juego online para dispositivos móviles desarrollado por Wildlife Studios.

Importante: la definición de trabajo de "**no volver a jugar más**" es que no jueguen el juego al menos una vez durante las dos semanas siguientes (14 días) al tercer día de instalación. Es decir que, habiendo jugado el tercer día después de instalarlo, no lo usen nunca en los días que van del día 4 al día 17 (inclusive) después de haberlo instalado.

Los datos fueron provistos gentilmente por Wildlife Studios. Wildlife Studios es una de las 10 compañías más grandes a nivel mundial creadora de juegos para dispositivos móviles. Fue fundada en Brasil en 2011 y creció hasta convertirse en una organización global.

Datos

Para entrenar y evaluar sus modelos cuentan con distintos conjuntos de datos. Los mismos contienen detalles de la actividad de diferentes usuarios en el juego.

- Training sets (5 archivos)

Contienen datos que van desde el día 0 de análisis hasta el día 395 de instalación (*install_date*). Por motivos de anonimidad, no podrán saber a qué día calendario se corresponde a cada valor de *install_date* pero sí es cierto que son días sucesivos. Los 5 archivos en conjunto contienen 5,547,423 registros (filas). Cada registro se corresponde a un usuario que jugó al menos una vez el juego el tercer día después de haberlo instalado.

Además de las variables predictoras (las cuales se detallarán más abajo), **sólo este conjunto de datos contiene la variable *Label_max_played_dsi***. Esta variable indica para cada jugador cuándo fue la última vez que jugó el juego. A modo de ejemplo, si para un usuario el valor es igual a 6, quiere decir que la última vez que jugó el juego fue 6 días después de haberlo instalado (y con seguridad **no** es uno de los usuario que hizo churn bajo nuestra definición). En cambio, si este valor es igual a 3, indica que la última vez que jugó el juego fue 3 días después de haberlo instalado. O sea, no abrió el juego ninguna vez después de haberlo jugado el tercer día y en consecuencia hizo churn (salvo en unos casos particulares que se explicarán más abajo). Estará a cargo de ustedes armar la variable binaría (0 si no hizo churn, 1 en caso contrario) que indique que un jugador hizo o no churn. Deberán llamar a esta variable *Label*.

- Evaluation set

Contiene 661,313 registros con datos que van desde el día 400 de análisis hasta el día 459. Este conjunto de datos tiene exactamente las mismas variables predictoras que el conjunto de entrenamiento, salvo por *Label_max_played_dsi,* la cual no está disponible. Ustedes deberán entrenar un modelo en base a los datos de entrenamiento y, para cada observación del conjunto de evaluación, predecir la probabilidad de que *Label* sea igual a 1 (de que haga churn).

- Sample submission

Es un ejemplo de cómo debe ser el archivo que se suba a la plataforma de Kaggle. Noten que

debe tener dos columnas. La primera llamada *id* (con el valor de id de cada jugador de evaluación para quien se predice, este valor no debe tener duplicados) y otra llamada *Label* (con la probabilidad de churn predicha por su modelo). Noten que el archivo está delimitado por comas, tiene nombres de columnas, y el separador decimal es el punto.

Información censurada

Noten que, con el fin de mantener el realismo en la competencia, se tuvo mucho cuidado en simular qué información estaría disponible si uno tuviera que entrenar el modelo final el día 399 para luego hacerlo predecir del día 400 en adelante. Esto tiene dos consecuencia:

1) El conjunto de datos sólo tiene registros con valores de *install_date* hasta 395, pero el de entrenamiento comienza el día 400. Noten que si alguien instala el juego el día 398, el tercer día después de instalarlo será el día 401 (día 0 desde la instalación = 398, día 1 desde la instalación = 399, día 2 desde la instalación = 400, día 3 desde la instalación = 401). Pero nosotros entrenamos el modelo justo al comienzo del día 400, de modo que a esa fecha no sabemos si esa persona jugará el juego el tercer día después de instalarlo, y por hoy no se sabe si estará en nuestra muestra de jugadores. Un argumento similar se puede pensar para los datos de los días 397, 399.[1]
2) No se puede confiar en los valores de *Label_max_played_dsi* igual a 3 para los registros con valores de *install_date* que van de 383 a 395 inclusive. Piensen qué sucede cuando alguien instala el juego el día 383 y tiene *Label_max_played_dsi* igual a 3. En ese caso el día 17 desde la instalación (el límite para considerar que no es churn) será el día 400, pero la información de ese día es del "futuro", de modo que podría llegar a jugarlo en el día 400 pero aun no lo sabemos. Por eso, aún cuando jugara el juego en el día 14 su último registro de actividad sería del día 3. Un argumento similar se puede hacer para aquellos usuarios que tienen *Label_max_played_dsi* igual a 3 y valores de install_date igual a 384, 385, ..., 395. Noten que para los registros con valores de *install_date* que van de 383 a 395 inclusive, si *LABEL_max_played_dsi* es mayor a 3, la situación es diferente (sí se puede confiar en ese valor).

Variables Predictoras

Dejando de lado *LABEL_max_played_dsi* cada conjunto de datos tiene 100 variables que pueden usarse para predecir churn. A continuación se copia la información que la Wildlife Studios dió respecto a cada una de ellas:

- Categorical features*:*

  Categorical_1 to *Categorical_7*: are categorical variables related to user acquisition information.

- Currency related features:

  The columns *soft_positive*, *soft_negative*, *hard_positive* and *hard_negative* are about the two kinds of in-game currency: Soft and Hard. They represent that total spent or earned (negative or positive) until the third day since install date.

---

[1] El día 396 es diferente. Para estas personas si podemos saber si jugó el juego el juego el tercer día después de instalarlo (el día 399). pero dado que "*Label_max_played_dsi"* únicamente puede valer 3, no se entregaron datos de personas que instalaron el juego el día 396.

- Tutorial features:

  The columns *TutorialStart*, *TutorialStartPart1*, *TutorialStartPart2*, *TutorialStartPart3*, *TutorialStartPart4*, *TutorialStartPart5*, *TutorialStartPart6* and *TutorialFinish* tell if a player reaches that tutorial step or not until the third day since install.

- Sum_dsi_X features:

  All the columns that ended with the pattern *_sum_dsiX* were calculated considering the sum of events logged at X days since install. **They are not cumulative**. Here is a brief explanation about the meaning of each event:
  StartSession: user opens the app.
  StartBattle: user starts a battle.
  WinBattle: user wins a battle.
  StartGameplayModeBattle: user starts an alternative battle mode.
  LoseBattle: user loses a battle.
  PiggyBankModifiedPoints: there is a feature called piggy bank that is modified when the player interacts with metagame.
  OpenPiggyBank: user opens a feature that gives prizes.
  OpenChest: user opens a chest. A player wins chests when he wins battles.
  EnterDeck: user enters the page that he can see its deck cards.
  UpgradeCard: user upgrades an card.
  EnterShop: user enters the app shop.
  BuyCard: user buys a card.
  ChangeArena: user changes arena. It means he is advancing in the metagame.
  JoinTournament: user start a tournament. Tournament usually occurs each 3 weeks.
  QuitTournament: user quits tournament.
  StartTournamentBattle: user starts tournament battle.
  WinTournamentBattle: user wins a tournament battle.
  LoseTournamentBattle: user loses a tournament battle.

- Other features:

  device_model: device where the app was installed.
  platform: device platform (Android, iOs).
  age: player's age.
  id: row identifier
  user_id: internal app user ID identifier.
  install_date: day when the app was installed.
  traffic_type: if the connection was made mainly by WiFi or Cell Phone network.
  site: id of the advertiser who was most shown during the game.
  country: country where the device is located.

Link de Acceso a la Competencia

https://www.kaggle.com/t/5a5f2bcdab474267bf529b653ea6cc96