Compte-rendu sur le challenge Airbnb sur Kaggle

Grégoire MASSOT - gregoire-massot.com

15 Avril 2016

Introduction

J'ai participé au challenge proposé par Airbnb sur la plateforme Kaggle du 30 Janvier à la cloture le 11 Février. Ca a été pour moi l'occasion d'évaluer mon niveau par rapport aux Data scientists su monde entier. J'ai travaillé exclusivement avec Rstudio, l'interface graphique de programmation en R.

J'ai fini 176/1446 (12%, presque le premier décile !) pour mon second challenge.

Début du challenge, reprise d'un code existant

J'ai commencé par reprendre un code Python existant, proposé par ce Kaggler puis traduit en R par un autre Kaggler et qui a été copié par de nombreux autres participants.
Le code utilise seulement le fichier `train_users.csv` contenant les informations fournies par les utilisateurs lors de leur inscription sur Airbnb.

Ce code permet de se hisser à la 600ème place du leaderboard environ.

Exploitation de `sessions.csv`

Je décide alors d'exploiter le fichier `sessions.csv` qui est un log des actions de la plupart des utilisateurs de `train_users.csv` et de `test_users.csv`. Il faut alors insérer ces informations dans le tableau `df_all_combined`.
On procède à un *one-hot-encoding* pour chaque variable de `sessions.csv` et on reporte ces variables binaires dans `df_all_combined`.

Voici un exemple pour la variable 'Device' de `sessions.csv`

```
# On récupère les levels de la variable 'Device' de sessions.csv
vars <- levels(as.factor(df_sessions$device))

# Pour chaque level, on crée une variable dans df_all_combined et on
# l'initialise à -1 (pas d'informations)
for(i in 1:length(vars))
{
  df_all_combined[,vars[i]] <- -1
}
# Pour chaque personne présente dans sessions.csv, on va indiquer à
# df_all_combined que on a une information sur l'utilisateur et que à priori
# il n'a pas utilisé ce device
```

```r
people <- as.factor(df_sessions$user_id)
df_all_combined[df_all_combined$id %in% people,vars] <- 0

# Boucle sur les nouvelles variables one-hot. On remplit le df_all_
    combined
# de 1 pour les utilisateurs qui possèdent les différents devices.
for (i in 1:length(vars))
{
  print(i)
  sousens <- as.factor(df_sessions[df_sessions$device %in% vars[i
     ],]$user_id)
  df_all_combined[df_all_combined$id %in% sousens, vars[i]] <- 1
}
```

Ce code de *feature engineering* **permet de passer sous la barre des 250 dans le leaderboard.**

Fin du challenge, tuning du xgboost

Après avoir testé jusqu'à un certain point la méthode de tuning du package caret, j'ai utilisé mes derniers coups à gagner des places en plaçant les curseurs de xgbtrain un peu au hasard et en les testant en *submit* sur le serveur de Kaggle.

Cela m'a permis de **gagner entre 50 et 100 places** je pense.