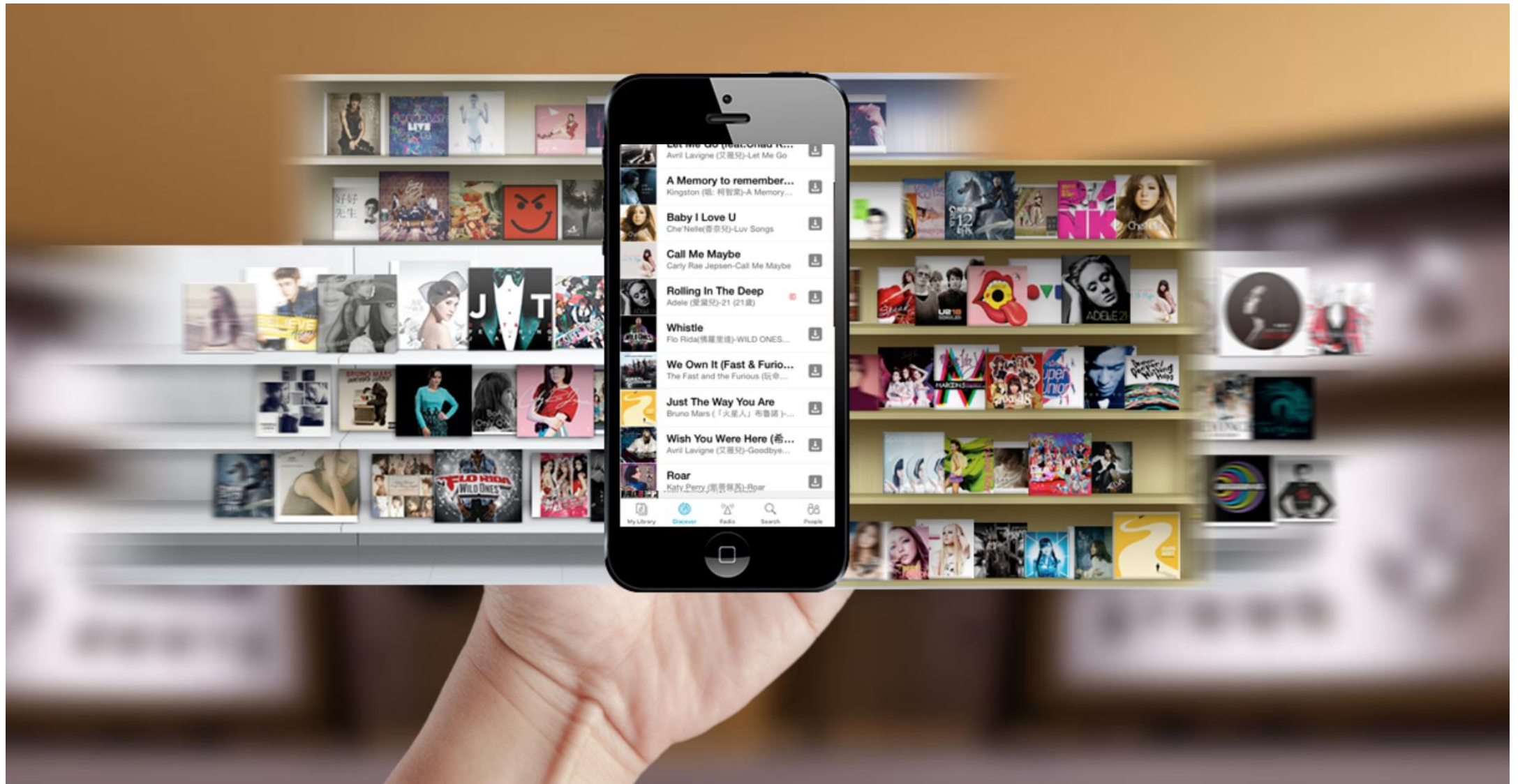


# KKBox's Churn Prediction Challenge

Can you predict when subscribers will churn?

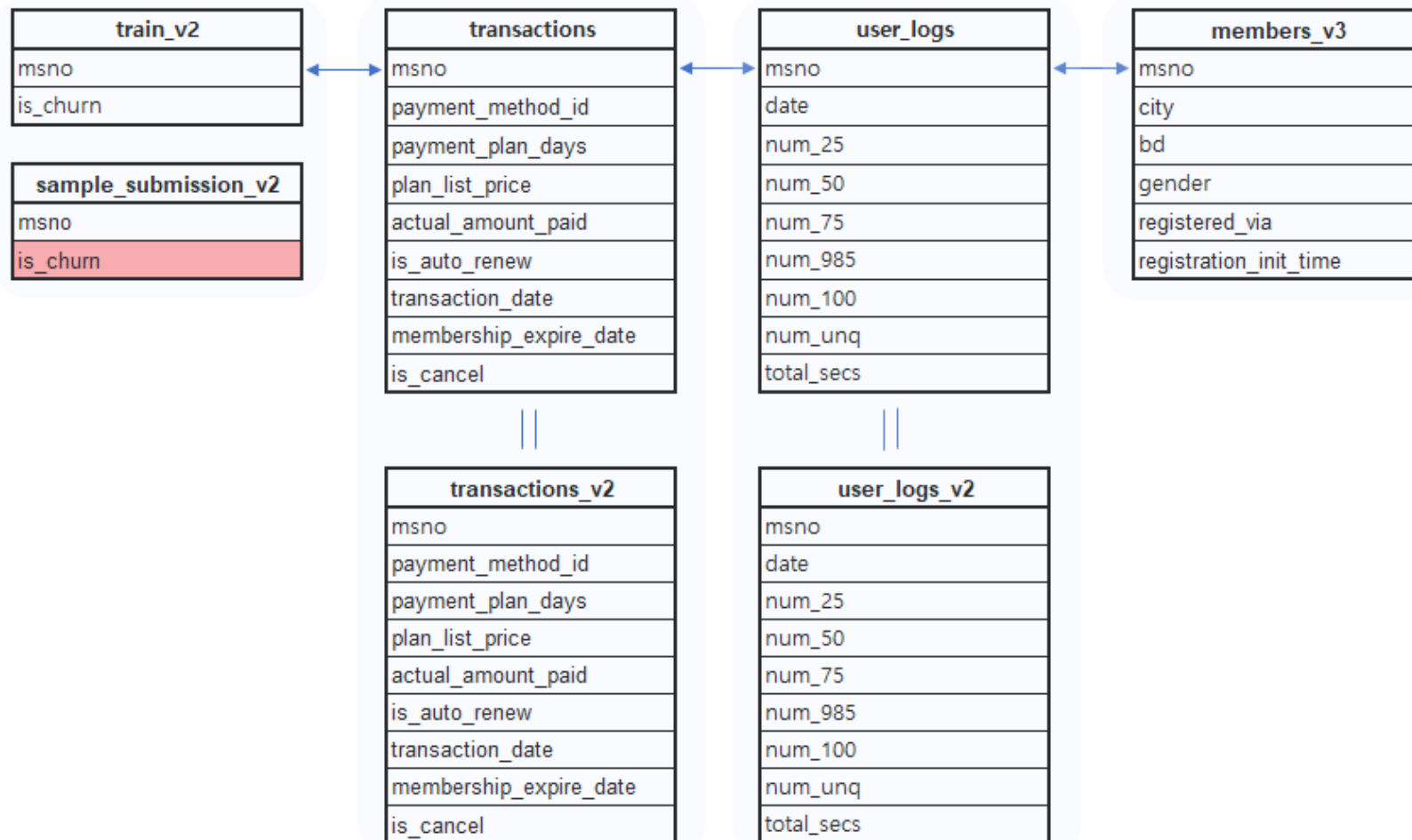
[KKBOX](#) is Asia's leading music streaming service



# 대회 설명

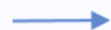
- Kkbox 유료 사용자가 3월에 이탈할 것인가? (Logloss 사용)
- 어머니? 정답셋 유출! 대회 폭파??
- 4월 이탈예측으로 변경
- 데이터 수정

# 데이터 설명



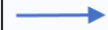
# 파생변수 생성

transactions
msno
payment_method_id
payment_plan_days
plan_list_price
actual_amount_paid
is_auto_renew
transaction_date
membership_expire_date
is_cancel



파생변수
msno
pay_30days
pay_7days
payday_sum_2015
payday_sum_2016
payday_sum_2017
avg_auto_renew
is_cancel
trans_cnt
fst_exp_date
lst_exp_date
sum_paydays
is_continuous
cancel_0_to_1_cnt
cancel_1_to_0_cnt
last_exp_month
unique_pay_method_cnt
max_pay_method
mean_diff_trans_exp
amt_per_day

user_logs
msno
date
num_25
num_50
num_75
num_985
num_100
num_unq
total_secs



파생변수
msno
sum_weight_logs_day
lst_day
log_count
mean_num_25
mean_num_50
mean_num_75
mean_num_985
mean_num_100
mean_total_secs
mean_rate_25
mean_rate_50
mean_rate_75
mean_rate_985
mean_rate_100
median_visit_days

# 최종 데이터셋

- X 변수 39개
- Y 변수 is\_churn

고객ID

target

msno
is_churn
sum_weight_logs_day
lst_day
log_count
mean_num_25
mean_num_50
mean_num_75
mean_num_985
mean_num_100
mean_total_secs
mean_rate_25
mean_rate_50
mean_rate_75
mean_rate_985
mean_rate_100
median_visit_days
pay_30days
pay_7days
payday_sum_2015
payday_sum_2016
payday_sum_2017
avg_auto_renew
is_cancel
trans_cnt
fst_exp_date
lst_exp_date
sum_paydays
is_continuous
cancel_0_to_1_cnt
cancel_1_to_0_cnt
last_exp_month
unique_pay_method_cnt
max_pay_method
mean_diff_trans_exp
amt_per_day
city
bd
gender
registered_via
registration_init_time

User\_logs  
파생변수 15개

transaction  
파생변수 19개

members  
변수 5개

# 고려해야 할 점

- Unbalanced Data

Y변수 is\_churn의 비율이 0:1 => 93:7

해결책 : upsampling, downsampling, threshold 조정

=> 0,1을 나누는 것 때문에 시행한다고 생각. Logloss score에서는 필요 X

- Feature Selection

해결책 : Stepwise, forward, backward deletion, lasso.

=> Tree based Model은 변수를 줄였을 때 성능향상이 일어나기보다는 성능이 조금이라도 떨어진다.

## 최종 모델

Gbm, Randomforest 두 모델의 조화평균

=>0.10246 (public) / 0.10389 (private) 13등!

- 왜 선택했나?

Deeplearning => 0.12212

Xgboost => 0.11408

Gbm => 0.10376

Randomforest => 0.10322



Research Prediction Competition

## WSDM - KKBox's Churn Prediction Challenge

Can you predict when subscribers will churn?

\$5,000 Prize Money

584 teams · 4 hours ago

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions **Late Submission**

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
mixing_all.csv.gz	9 hours ago	13 seconds	29 seconds	0.17733

Complete

[Jump to your position on the leaderboard](#)

Public Leaderboard Private Leaderboard

The private leaderboard is calculated with approximately 60% of the test data. [Refresh](#)

**In the money**

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲1	Bryan Gregory			0.07974	84	4h
2	▲3	Swimming			0.08926	15	4h
3	▲3	JonahWang			0.09344	25	3d
4	▲3	501		+5	0.09660	202	8d
5	▲3	Alaric			0.09664	15	7d
6	▼5	PKU Fresher			0.09743	158	10h
7	▼4	Aloisio Dourado			0.09829	79	5d
8	▲1	Cara Gu			0.09843	26	8d
9	▲1	bestfitting			0.09886	69	4d
10	▲1	Crush			0.10065	109	9h

- 0.10065 -> 10등 -> 다함께 한것
- 0.10389 -> 13등 -> 발표는 나 혼자 한 걸로
- 단, 10등에서 말소 !!!  $\pi\pi$
- 응??? 이유는?
- 계정 3개 사용
- 룰을 제대로 안 읽은 잘못.
- 그리고 10등까지 할 줄 몰랐음.
- 그래서 열심히 했지만 기록으로 남는게 없어요!
- 경험은 했지만~ 코드도 있지만~ 기록이 안남네요  $\pi\pi$

# 작업환경

## <실제 작업환경>

- RAM 8G
- CPU i5
- Macbook pro
- Rstudio
- H2O packages

## <대회 후 작업환경>

- RAM 64G
- CPU i7-8700K 12 cores
- OS Windows10
- Rstudio
- H2O packages
- docker

# 한계

Testset : 907,471  
Trainset : 10,394,394  
( 2016년 2월 - 2017년 2월)  
User\_logs : 410,502,906  
Members : 6,769,473

- 데이터가 너무 커서 시간이 너무 오래 걸린다.
- 캐글 대회 1등은 데이터 핸들링을 MS SQL로 했다고 한다.
- 데이터 핸들링 코드 짜는 거 제외 모두 돌리는데 최소 12시간은 걸렸음. 맥북으로는 transaction만 처리하는데 12시간 걸렸음.
- 모델을 한번 fitting 하는 시간도 한번에 40분 정도 걸려서 nolds=30이 한계치. 맥북으로는 2시간 정도?!(3배 차이)
- 그래서 모델 튜닝을 많이 못해본 것이 가장 아쉽다.
- 모델 튜닝하지 않고 기본 ntree=200, nolds=3을 기본으로 피팅.
- 파생변수 만드는 것에 더 집중.

# 최종 결과 및 느낀점

- 어떤 것이든 룰을 먼저 확인.
- 모델 튜닝 부분이 부족.
- Docker를 활용하여 rstudio작업환경 구성법을 배움.
- 파생변수를 만들면서 dplyr을 아주 잘 활용할 수 있게 되었음.
- 코드를 짤 때부터 주석을 잘 달아둬야 한다.