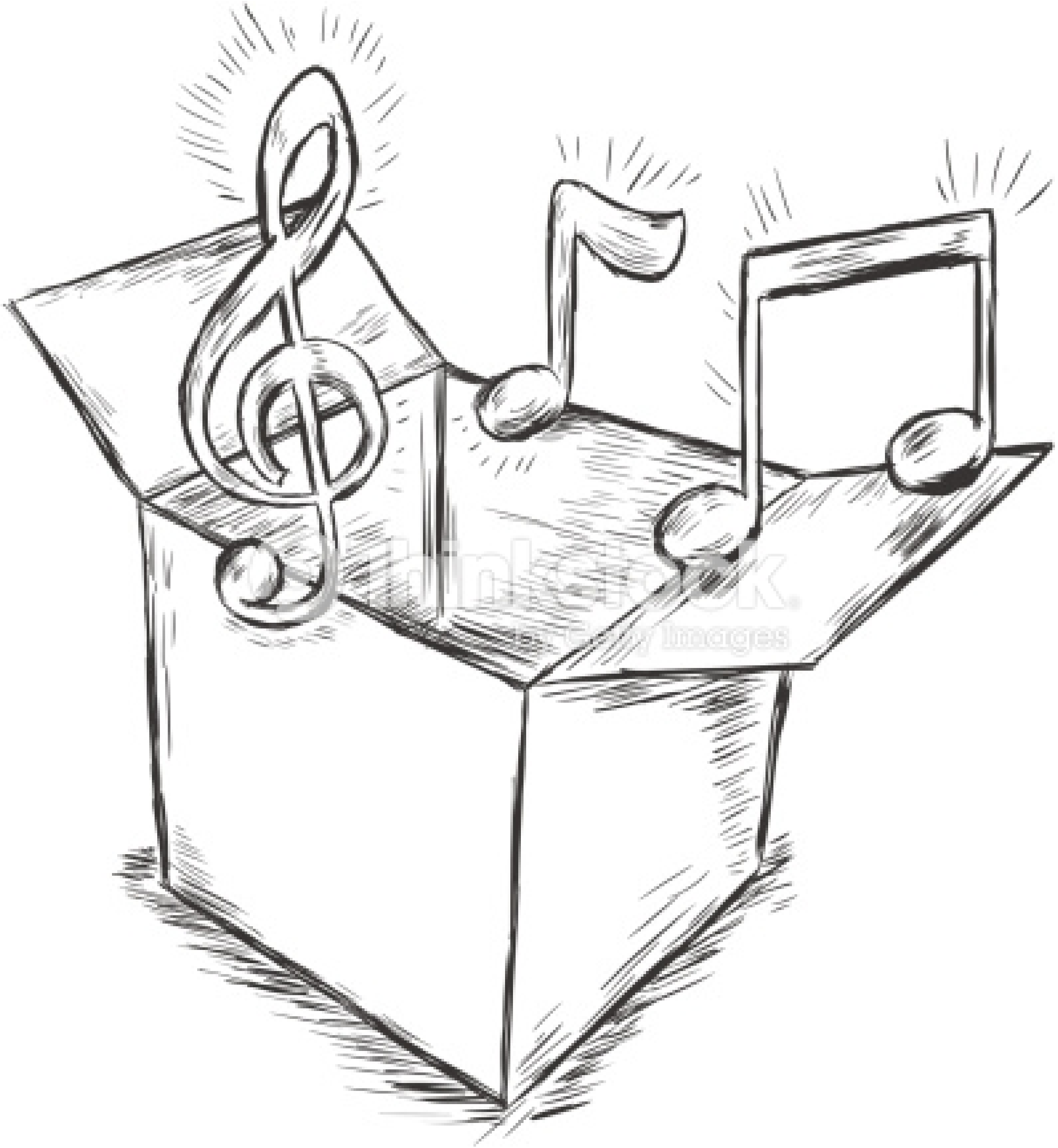
Music Box Analysis Report

Mengguo Yan

Bittiger 501-1803

2018-09-02



source: http://www.thinkstockphotos.com/image/stock-illustration-vector-sketch-illustration-box-with/460603911 (http://www.thinkstockphotos.com/image/stock-illustration-vector-sketch-illustration-box-with/460603911)

1. Goals & Results

1.1 Goals

1.1.1 User Churn Prediction

✓ Validate and clean up log-based raw datasets

✓ Data imputation, feature engineering, remove bots and outliers, downsample if possible

✓ Exploratory data analysis

✓ Build user churn prediction model based on user behavior

✓ Perform analysis, draw conclusion and give advices based on model results

1.1.2 Recommender system

✓ Validate and clean up log-based raw datasets

✓ Data imputation, feature engineering, remove bots and outliers

✓ For existing users: Use a song's played percentage as ratings to give item-item similarity recommender

✓ For new users: Popularity-based recommender

1.2 Results

1.2.1 User Churn Prediction

✓ Pipeline and grid search were used to search for the model with best hyperparameters.

✓ All the models give great results. AUC of test is between 90.5% to 92.2%.

✓ According to feature importance, the recency of play is important. In order to prevent user from churning, to boost the user activity is critical. And according to recency figure in section 3.1.4, day 7 seems to be the equilibrium point which determines whether a user is more likely to churn or not.

✓ Moreover, device type is also an important feature. Please refer to section 3.2.1, Android phone has higher churn rate than iphone or other phones.

✓ In addition, churn rate in users who enjoy different types of songs is also significantly different.

Therefore, here are advices to decrease user churn.

a. Send reminders to user who does not have activities in 7 days. It can be mobile pop-up notifications.

b. Check the user experience in Andriod phone and make some improvements.

c. Improve the recommender system. And more researches should be done in analysis user behaviors who enjoy song-type-1.

1.2.2 Recommender system

✓ For customers who have play records, collaborative filtering item-item recommender system is used. The error of mean of absolute error is 0.096.

a. Cosine similarity is used in similarity metric. And customer's play percentage is used as a rating method.

b. In this project, songs that were played more than 1000 times during the feature windows are selected in the item-item matrix.

c. More advanced model should be developed. For example, the n recommended songs contains 80% of the same type of songs the user enjoy the most and 20% of the most popular songs on the music box platform or 20% of the other types that the users might also want to try.

d. uid = '168503861' was selected as a test in this system. The list of songs heard is shown below:
['Bye Bye Bye', '剃刀边缘-(电视剧《剃刀边缘》主题曲)', '铃声多多_拥抱到最后 - Dj加快版_好听', '西海情歌', '手心里的温柔', '约定', '爱在记忆中找你', '凤凰山', '女人的选择', '为你我受冷风吹', '铃声多多_7妹 - 把你轰到太平洋_爽死了', '说一句我不走了', 'Tell Me Why', '汽车音响专用(靓音发烧女声)HIFI试音天碟(DJ版)', '北京东路的日子', '超重低音极品慢摇 煲音箱耳机专用']
And the recommender system recomends:
['你的样子', 'AINY爱你DJ(Remix)', '最后的火车站', '跑', '寻觅理想', '蝴蝶泉边(选自《五朵金花》)', '爱在2017', '爱德华时代', '夜店热播开场','雨过昔年']

✓ For new customers, popularity based recommender system can be used. Please refer to section 5 for details.

a. The way to consider popularity is to combine both ratings and counts for a song. Rating is considered the average percentage of the song played by users and count is the total number of each song played by users during the feature window

b.Top 20 the most popular songs in each type of songs can be used as a cold start for new customers

1.3 Conclusion and suggestions

1.3.1 User churn prediction

✓ In order to decrease user churn, a better recommender system should be developed.

✓ When a user stop using our music box, a gentle remider should be send.

✓ User experience on Android phone may need to be improved.

1.3.2 Recommender system

✓ In order to provide better recommendation, a more sophisticated rating system is needed. For music box platform, a simply thumbs up or favor indicator will be of great help.

✓ Song type in this data set should be carefully placed. A more detailed and accurate song type label should be generated.

2. Definitions

2.1 Churn prediction features and label window definition

All features is generated in the feature window and label as churn or not churn user is generated in label window. Churn will be generated as label 1 and not churn as 0.

✓ Feature window:
- 2017-03-30 ~ 2017-04-28 days: 30

✓ Label window:
- 2017-04-29 ~ 2017-05-12 days: 14

2.2 Population definition

Useful population in this project is considered as active users during feature windows. And their actives in the following two weeks as label window are used to generate labels.

✓ Incluse: all active users during feature time window

✓ Excluse: inactive user during feature time window and bots/outliers

2.3 Churn prediction features generation

✓ User active frequency

Defined as number of events over time windows.

✓ User active recency

Defined as number of days bewteen last event from snapshot date.

✓ User average play length percentage

Defined as average play time percentage of songs for each users.

✓ User device

2.4 Recommender

2.4.1 Item-item similar recommender

✓ Item-item similar with cosine similarity

✓ Song-recommender tank contains 1141 songs each of which has total of plays more than 1000 times during the feature window

✓ Item-item similar with cosine similarity

2.4.2 pupularity recommender

✓ Weighted Rating $(WR) = (v/(v+m).R) + (m/(v+m).C)$where
v: the number of plays for the song;
m: the plays required to be listed in the chart;

R: the average play percentage of the song;
C: the mean percentage across the whole song list.

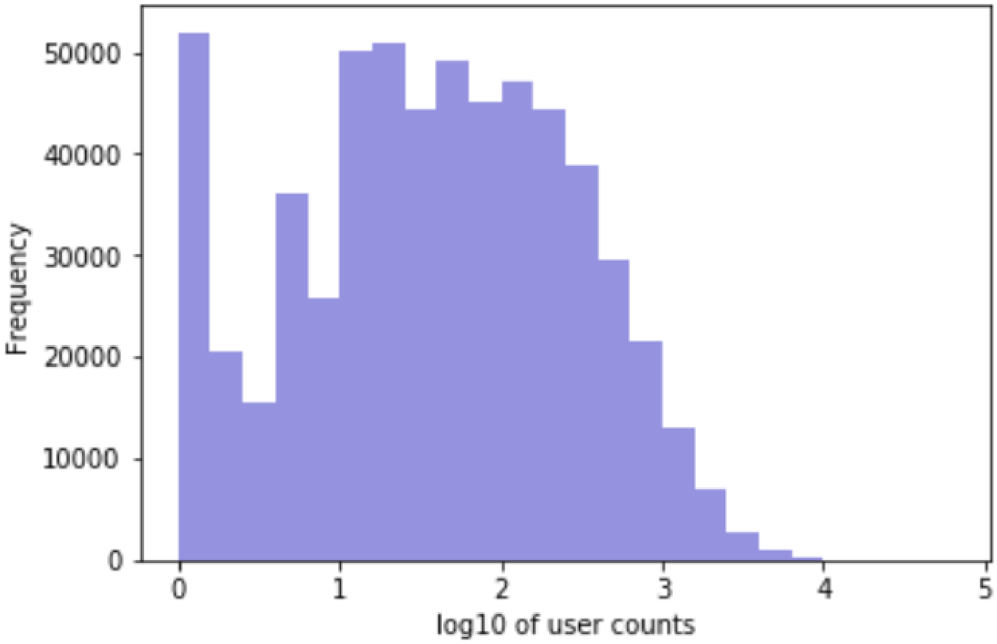
✓ Different recommendation song lists are generated for each different song types.

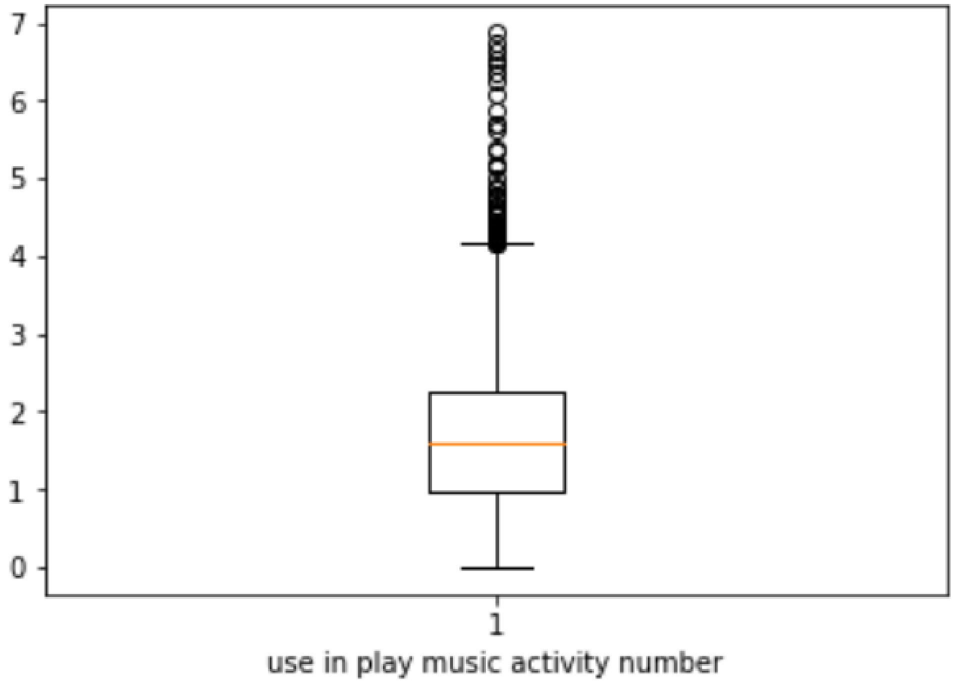
3. Exploratory data analysis

3.1 User Activity Analysis

From 2017-03-30 to 2017-05-12, the total number of users on file is 59,4720, who have activity of play/search/download music on the music box platform. However, the dataset also contains bot user, which should be removed. In addition, in order to handle large dataset calculation, downsample was applied. And the total number of users after down sample is 11,8247.

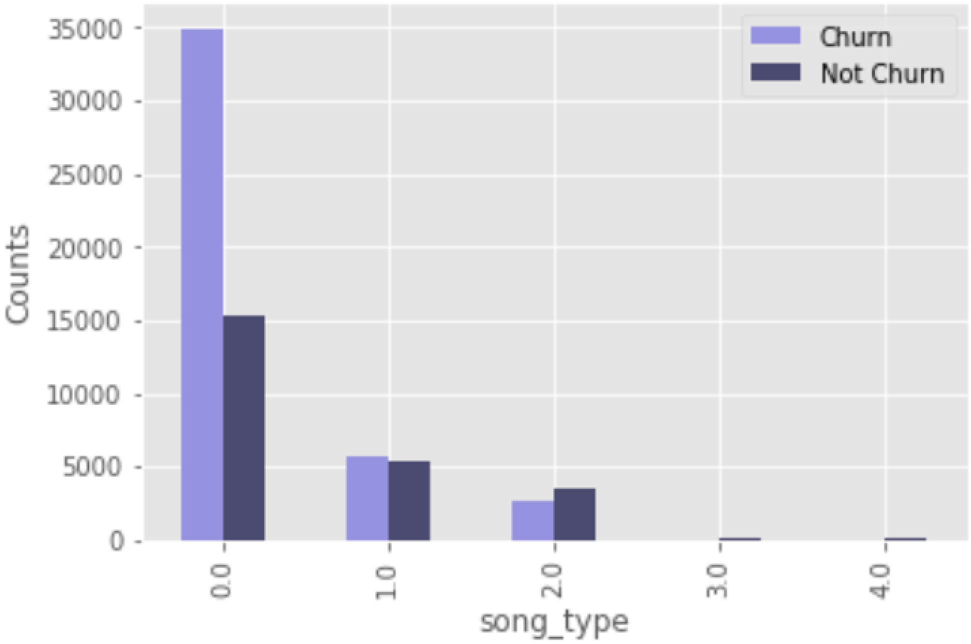
3.1.1 Distribution of number of user play during the window time





It can be easily shown that most of our users play roughly hundreds of songs during these periods of time. In average, 2~10 songs per day. Let's asume average 4 minutes per song. It means that on average, most of our customers spent less than an hour per day on our music box platform. And the lack of time spends on music box platform may lead to customer churn eventurelly.

3.1.2 Song type and churn



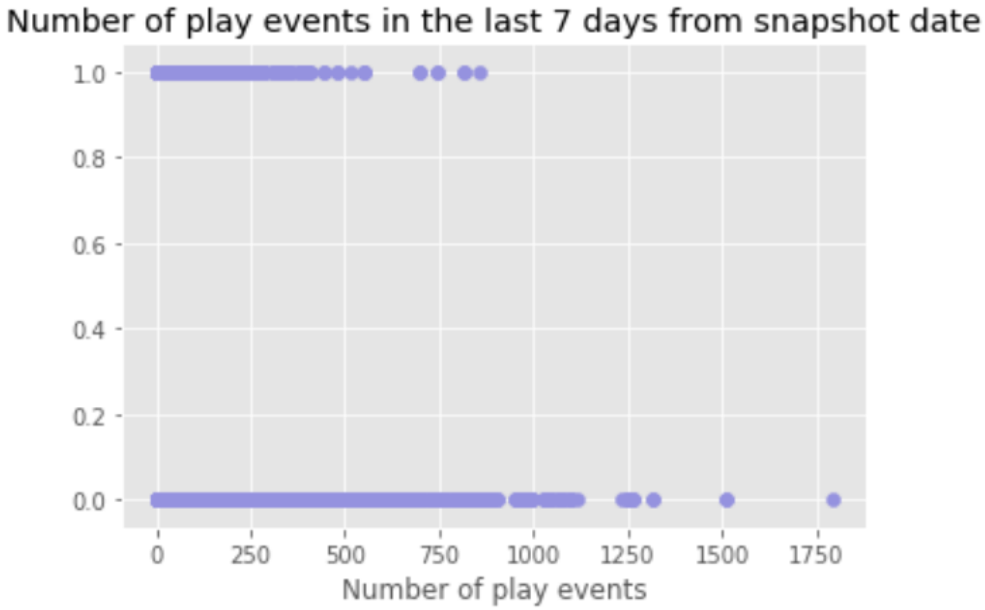
This figure interestinly shows the loyalty of our user by the type of song they like. Our music platform classifies the songs into 4 different types. Type 1 is the most popular. However, the churned users are much more than the unchurned users. There are several reasons may cause it.

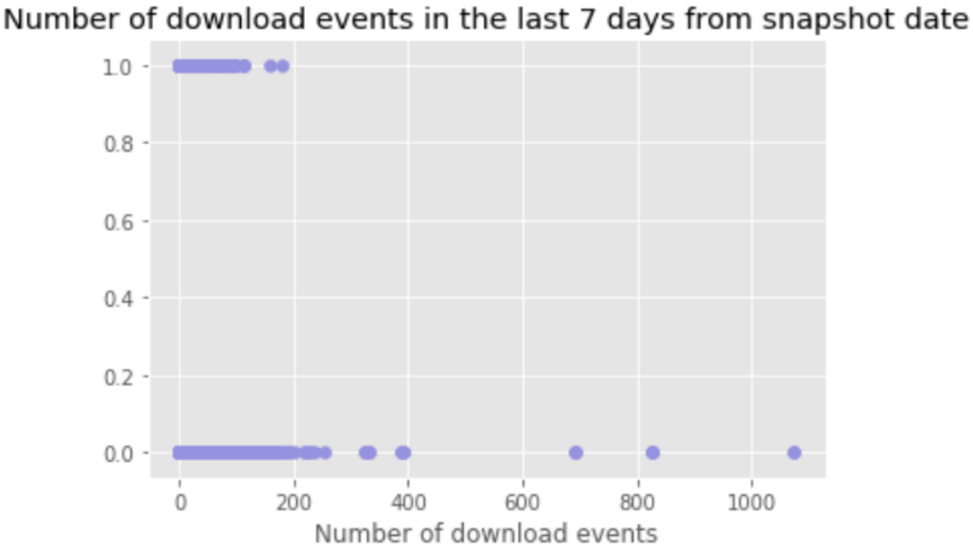
- a. Lack of type 1 songs or low quality of type 1 songs to users.
- b. Recommender system cannot provide what customers like

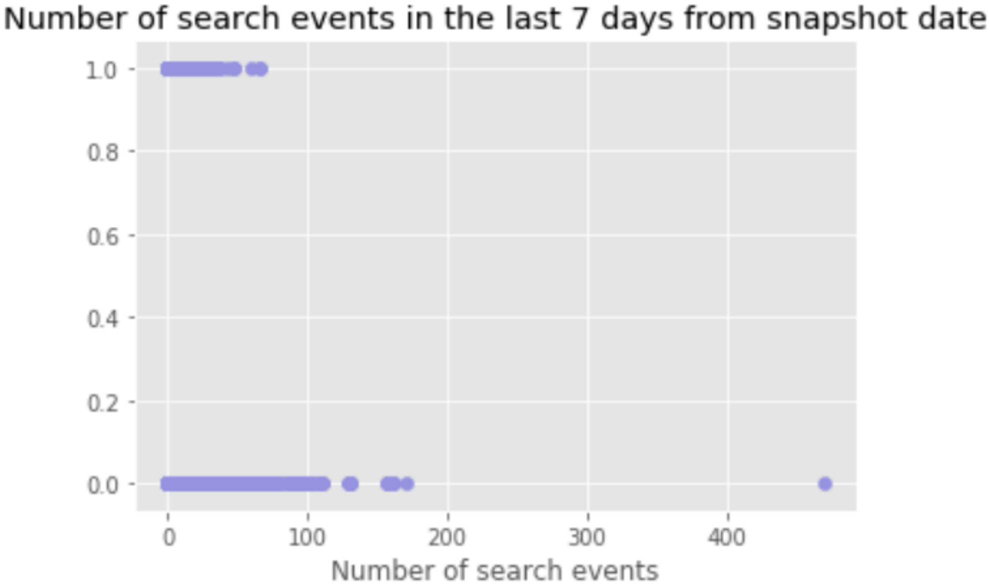
But on the other hand, users who enjoy type 1,2,3,4 are less likely to churn. It may be due to our music box can provide more of these types of songs than other platforms.

3.1.3 Event frequencies and churn

Event frequencies are another great indicators to user churn. Users who have more activities during a window time are less likely to churn and vice versa. Below shows the churn vs activities number in a 7 days window.

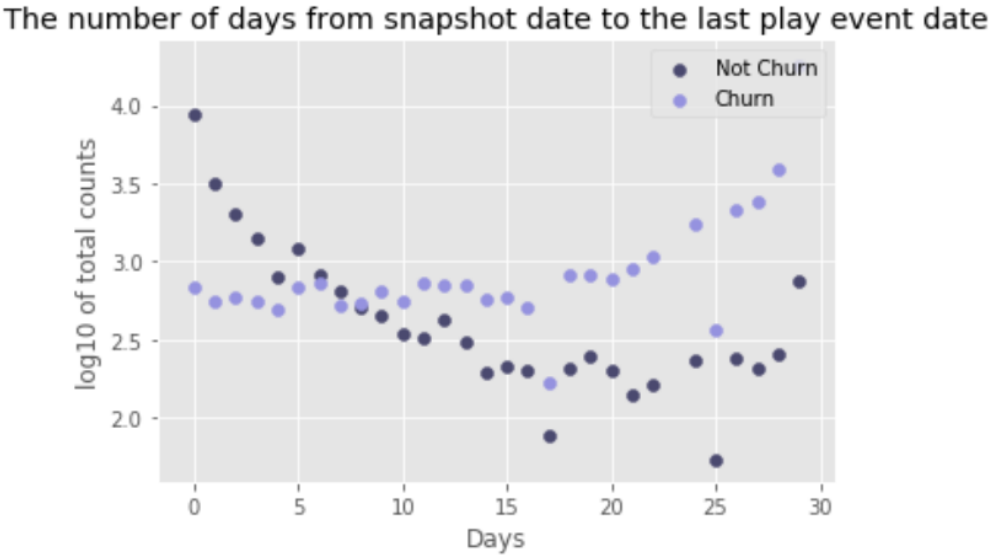






3.1.4 Event recencies and churn

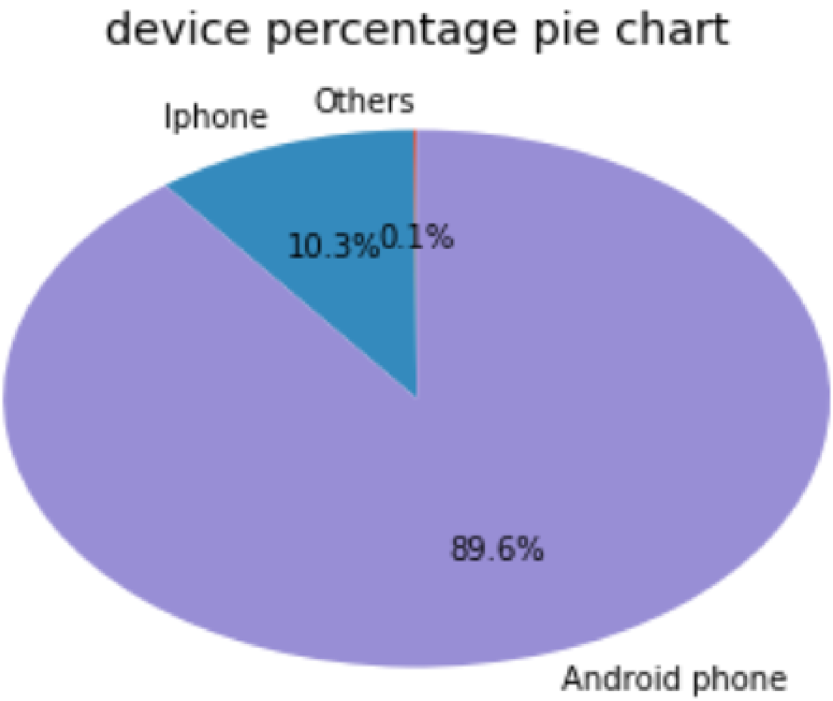
Event recencies represent the number of days from snapshot date to the last event occurred date. From the figure below, it is obvious that 7 day is a critical number. If users' last play activities date is whithin 7 days to the snapshot date, the not-churn probability of those users are higher.



3.2 User Profile Analysis

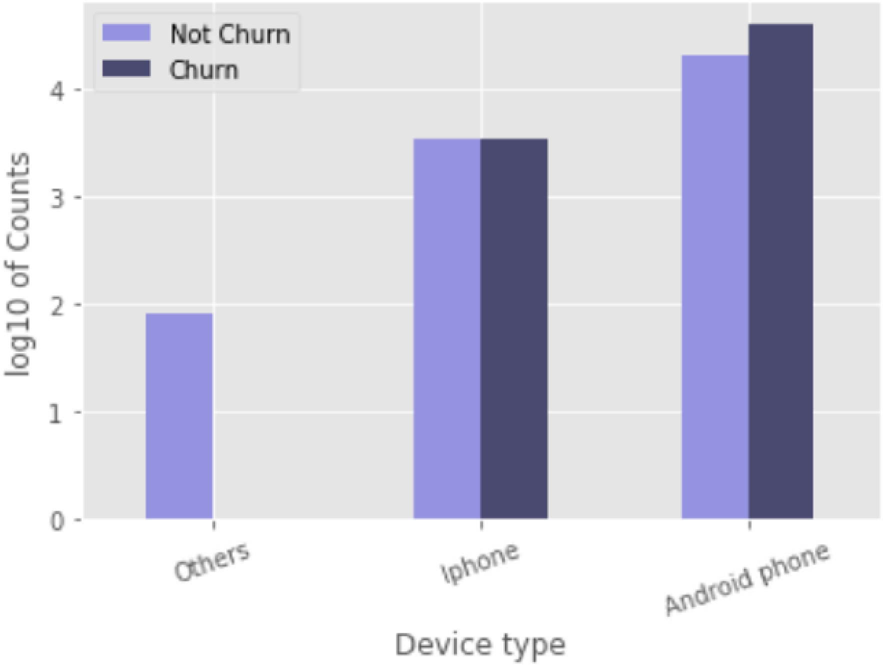
3.2.1 device type analysis

On our platform, nearly 90% of devices are with android system and 10% of IOS system. The rest includingmac and windows phone is about 0.1%.



However, according to the figure below, there are more user churns on android phone than the others. It may be due to the following reasons.

- a. The user experiences on android phone are not as good as some other music player platform.
- b. And there are more platform options on andriod phone system.

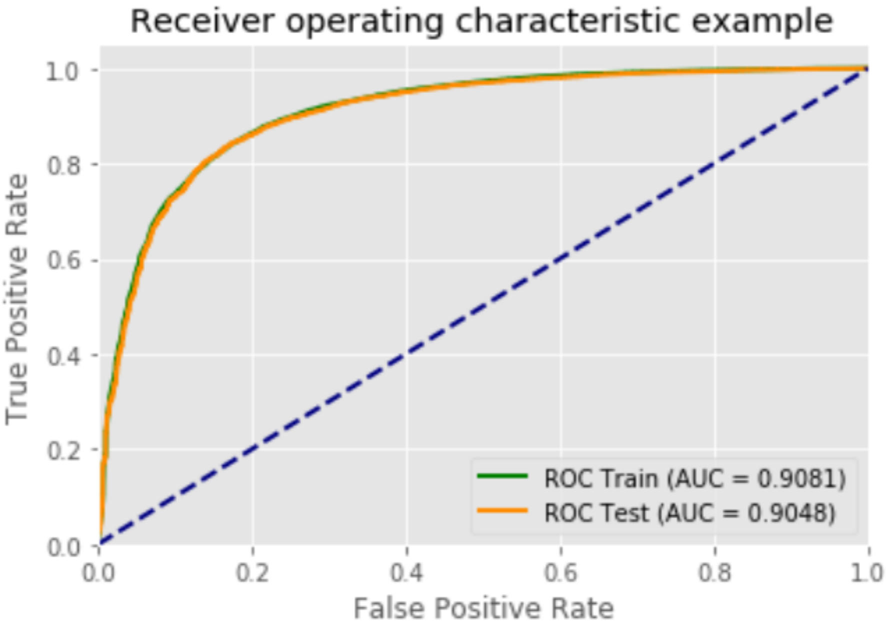

Not Churn
Churn
log10 of Counts
0
1
2
3
4
Others
Iphone
Android phone
Device type

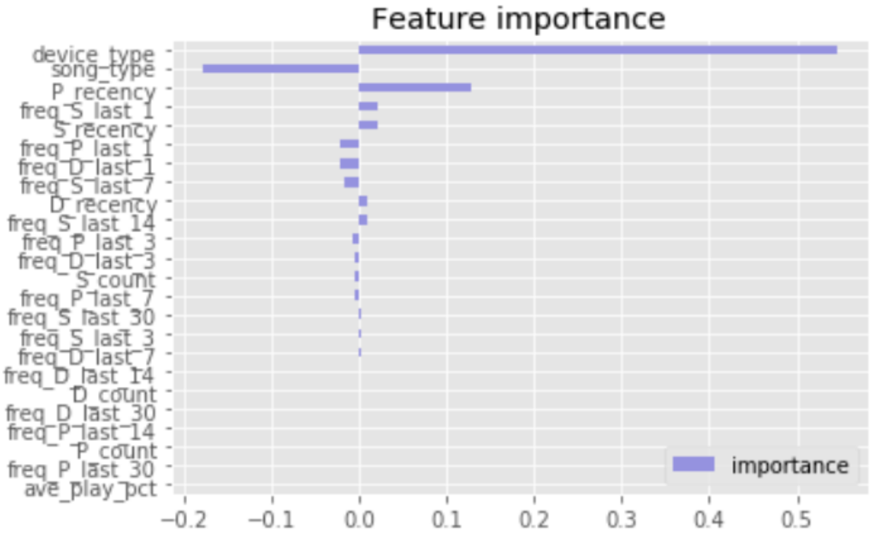
4. Prediction model detials

Use pipe line and grid search to fine tuen the model with best hyperparameters

4.1 Logistic regression

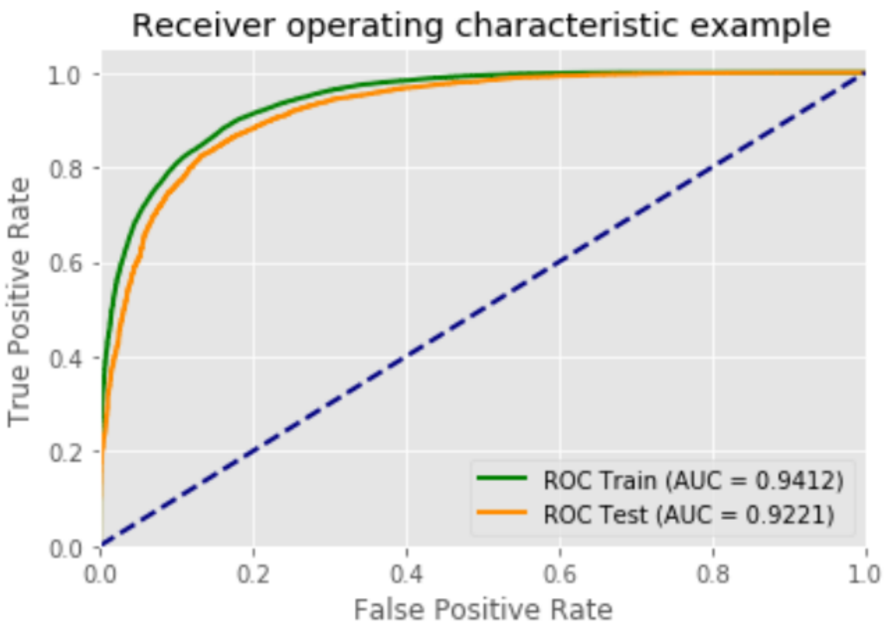
metrics	train	test
AUC	0.908097	0.904850
Accuracy	0.844160	0.842346
Precision	0.875762	0.878054
Recall	0.881238	0.875632
f1-score	0.878492	0.876842

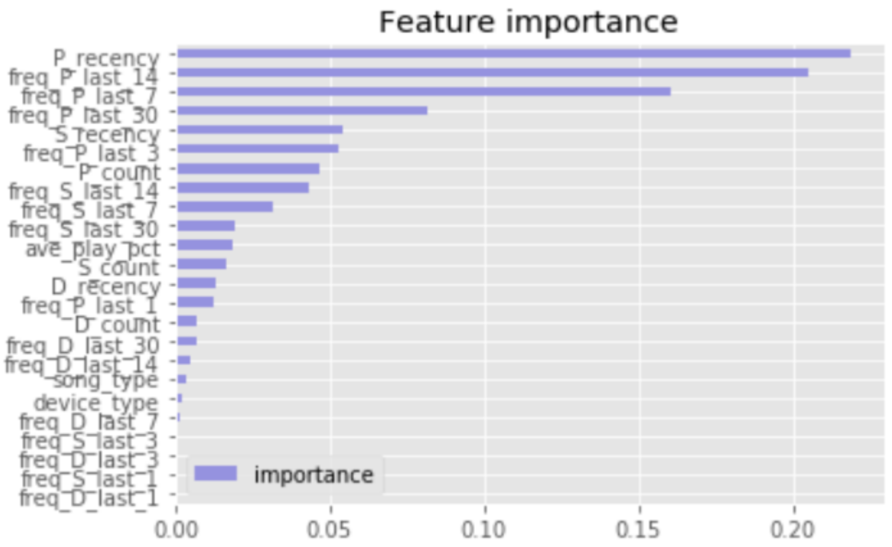




4.2 Random forest

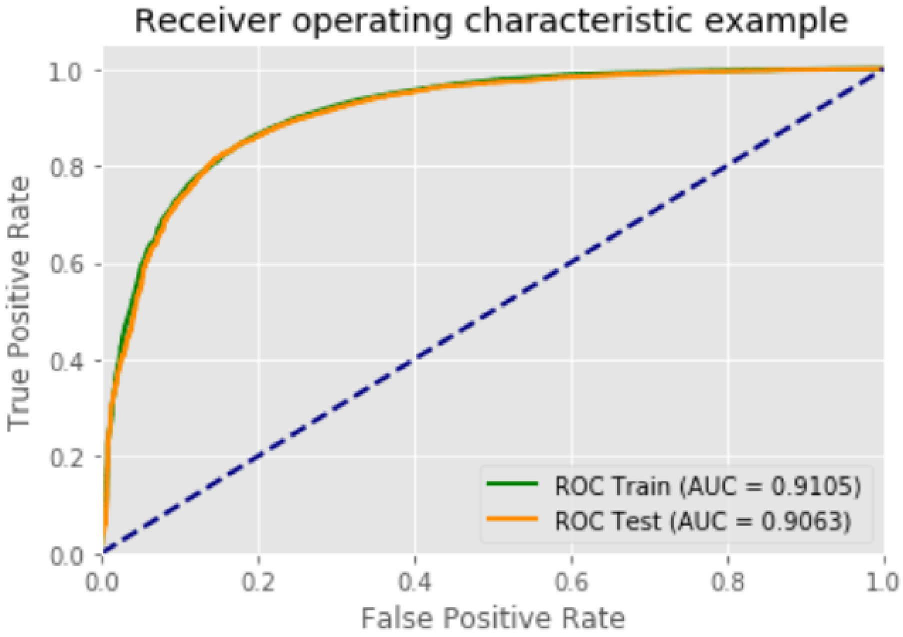
| metrics | train | test | |------------------------------| |AUC | 0.941202 | 0.922104| |Accuracy | 0.872065| 0.855680| |Precision| 0.882755| 0.873794| |Recall | 0.922381| 0.905632| |f1-score | 0.902133| 0.889428|





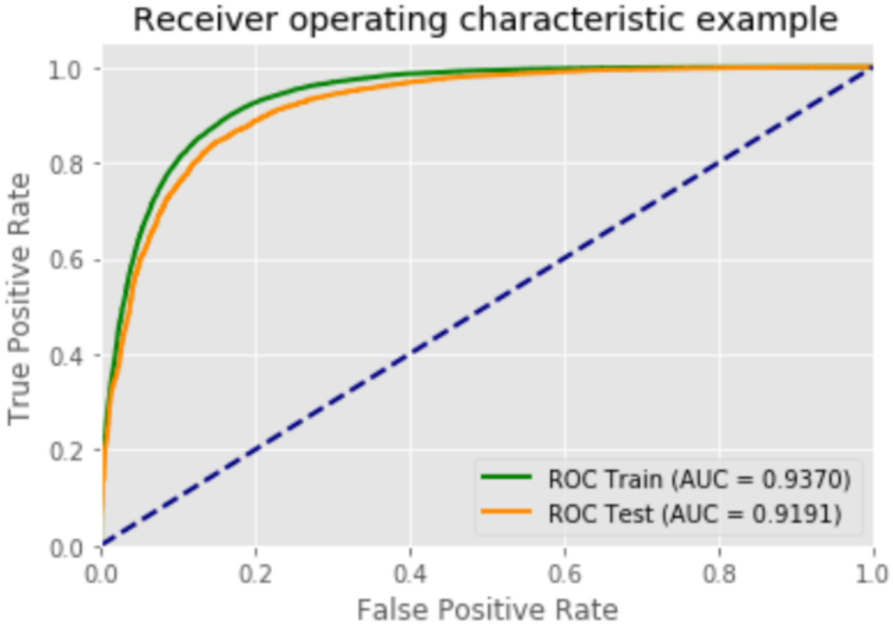
4.3 Neural network MLPClassifier

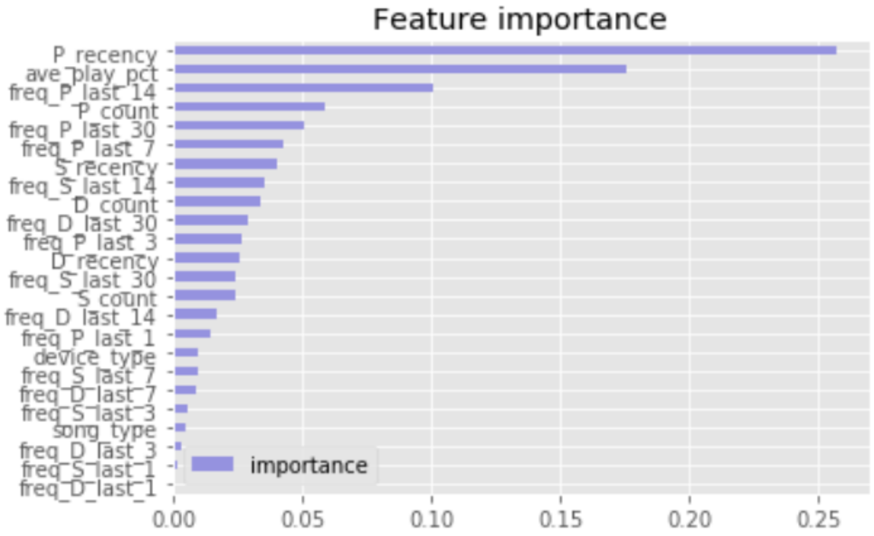
|metrics | train | test | |------------------------------| |AUC | 0.910459| 0.906268| |Accuracy | 0.844934| 0.843082| |Precision| 0.861679 | 0.863707| |Recall | 0.902270 | 0.896667| |f1-score | 0.881508 | 0.879878|



4.4 Gradient Boosting Classifier

|metrics | train | test |-------------------------------| |AUC | 0.937048| 0.919117| |Accuracy | 0.880684| 0.860027| |Precision| 0.885151| 0.874037| |Recall | 0.934626| 0.913218| |f1-score | 0.909216| 0.893198|




Feature importance
P_recency
ave_play_pct
freq_P_last_14
P_count
freq_P_last_30
freq_P_last_7
S_recency
freq_S_last_14
D_count
freq_D_last_30
freq_P_last_3
D_recency
freq_S_last_30
S_count
freq_D_last_14
freq_P_last_1
device_type
freq_S_last_7
freq_D_last_7
freq_S_last_3
song_type
freq_D_last_3
freq_S_last_1
freq_D_last_1
importance
0.00
0.05
0.10
0.15
0.20
0.25

5. Recommender system

5.1 Item-item similarity for users

- Randomly selected user with uid = '168503861' has played songs as belowed.
['Bye Bye Bye', '剃刀边缘-(电视剧《剃刀边缘》主题曲)', '铃声多多_拥抱到最后 - Dj加快版_好听', '西海情歌', '手心里的温柔', '约定', '爱在记忆中找你', '凤凰山', '女人的选择', '为你我受冷风吹', '铃声多多_7妹 - 把你轰到太平洋_爽死了', '说一句我不走了', 'Tell Me Why', '汽车音响专用(靓音发烧女声)HIFI试音天碟(DJ版)', '北京东路的日子', '超重低音极品慢摇 煲音箱耳机专用']
And the recommneder system recommends the songs:
['你的样子', 'AINY爱你DJ(Remix)', '最后的火车站', '跑', '寻觅理想', '蝴蝶泉边(选自《五朵金花》)', '爱在2017', '爱德华时代', '夜店热播开场', '雨过昔年']

- Compared to the true scores and predicted scores from the model, error of mean of absolute error is 0.096.

5.2 Popularity based recommender for new users

Song Type 0

song_id	song_name	num_of_count	ave_rate	song_type	core
219410	Harmonize	7	1285.690000	0.0	272.921077
108733	赵培、赵蕾 - 不怕	42	79.646818	0.0	49.290175
38793	社会小伙 社会嗑 上道小曲 儒小哥	11	137.239200	0.0	40.978290
46261	Let's Celebrate	49	54.472245	0.0	35.676074
219412	Minuet No. 5	4	244.287500	0.0	32.790518
66810	黄子韬 - 第一课	16	79.718269	0.0	30.525187
108747	纸弦 - 白头吟	50	34.168272	0.0	22.565515
16809	Ti Amero	30	38.897097	0.0	20.954972
107407	夜雨诉笛	49	27.336712	0.0	17.947526

Song Type 1

song_id	song_name	num_of_count	ave_rate	song_type	score
248738	罗成算卦 关小平 小豆豆	3	202.041000	1.0	21.127191
38630	社会小伙 社会曲	12	53.474118	1.0	17.059341
100176	许嵩 - 半城烟沙	44	19.339172	1.0	12.249845
86218	林俊杰 - 背对背拥抱	66	12.542433	1.0	9.069197
109963	记忆抽屉	58	10.282123	1.0	7.177723
302108	彩云追月(伴奏版)	1	152.511802	1.0	5.891754
145453	020杨家将	4	41.256488	1.0	5.719717
213433	We Are Bounce (Original Mix)	30	9.474937	1.0	5.193101

song_id	song_name	num_of_count	ave_rate	song_type	score
213443	Exploration Of Space (Cosmic Gate's Back 2 The...	20	8.421337	1.0	3.804180

Song Type 2

song_id	song_name	num_of_count	ave_rate	song_type	score
1552	命	2215	1.000000	2.0	0.991328
922	来生愿做一朵莲	1610	1.000000	2.0	0.988121
1551	不要用我的爱来伤害我	558	1.000000	2.0	0.966722
170961	让我为你唱一首歌(一起来看流星雨主题曲)	8	3.186179	2.0	0.942793
940	鸟之诗	68	1.000000	2.0	0.793250
126112	别哭 我最爱的人	38	0.982143	2.0	0.685734
42482	倾世皇妃(完整版)-林心如	51	0.868319	2.0	0.660387
73034	不安静	29	1.000000	2.0	0.646646
170899	Slow Down (BONUS TRACK)	11	1.553193	2.0	0.639207