Capstone Project
Music Box

Yajing Li

10/26/2018

Overview

- The goal of the project is through analyzing the music box user behaviors to:
  - Make Churn prediction
  - Build Recommendation system

Data description

- User log data from music box with 3 million song play records and 260K daily new user increment.
- Data size~11.2GB,including search, download and play records collected from 03/30/2017 to 05/07/2017. including uid, song id, name, type, singer, play time, song length, etc

	uid	device	song_id	song_type	song_name	singer	play_time	song_length	paid_flag	file_name
0	168920276.0	ip	619727	0.0	One Time	Justin Bieber	215	217	0.0	20170401_3_play.log
1	169038973.0	ar	0	1.0	Stellar - Marionette - 舞蹈版	未知歌手	211	211	0.0	20170401_3_play.log
2	168541028.0	ar	1.46542e+07	0.0	金大人的梦	金大人	161	161	0.0	20170401_3_play.log
3	168868797.0	ar	7.15339e+06	0.0	7妹卓越电锯(3D全景)	7妹&小可	5	234	0.0	20170401_3_play.log
4	168561111.0	ar	4.85576e+06	0.0	青春修炼手册-(动漫《燃烧的蔬菜》片尾曲/电视剧《极品家丁》插曲)	TFBOYS	3	263	0.0	20170401_3_play.log
5	168872672.0	ar	1.5825e+07	0.0	不愿红尘错过你(独白版)	吴易航	36	295	0.0	20170401_3_play.log
6	168742497.0	ar	5.25812e+06	0.0	兄弟为你	何流	252	252	0.0	20170401_3_play.log
7	168907926.0	ar	0	1.0	music	未知歌手	59	60	0.0	20170401_3_play.log
8	168788136.0	ar	2.83755e+06	0.0	夢想歌	Suara	223	247	0.0	20170401_3_play.log
9	168732607.0	ar	3.21392e+06	0.0	方圆几里	薛之谦	45	263	0.0	20170401_3_play.log
10	168949635.0	ip	4.89004e+06	0.0	祖国不会忘记	韩红	57	223	0.0	20170401_3_play.log

Churn Prediction using Spark

- Validate dataset of user activities(Search, Play, Download)
- Perform data preprocessing and feature engineering. Feature selected are activity(search,play,download) frequency, recency features (defined as days from last event), user profile features and total play time. All the features are selected over the time window of 1,3,7,14,30 days.
- Build user churn prediction model based on user behavior from downsampled user population

Algorithm comparison in churn prediction models

AUC Score	Logistic Regression	Random Forest	Gradient Boosted Tree	Naïve Bayes
Training Set	0.8957	0.9192	0.9097	0.9097
Test Set	0.8962	0.907	0.9043	0.907

- The highest test accuracy is 0.907 from Random Forest and Naïve Bayes
- From the logistic regression: top 5 impactful features on the churn predictions are: 'days_from_last_play','freq_S_last_1','freq_S_last_3','freq_P_last_1','freq_S_last_7'

Recommendation system using Spark

- Validate dataset of user play records, perform data cleaning and transformation
- Performed Feature engineering to measure the similarity between users. Selected feature is Play time ratio, defined by $\frac{user\_play\_time}{song\_length}$ is used as a measure of user's rating of the song
- Constructed Utility matrix and build music recommendation based on user listening history using ALS algorithm.