Experiment Analysis

Content

Data Exploration

When we tried to explore the datasets, we found two very interesting things.One is that all the train data was provided based on time series, which means all the play behaviors were sorted by date.So in order to establish a better split of the train data to build a local validation strategy we need to maintain at least 20% samples which were located at the tail of the train dataset.This indicates that we use the passed play behaviors to predict the future play.This strategy will uniform the local score and the online score regularly.

The other thing is about “cold start” problems.After the analysis, we also found out that about 18% samples of the test set were “cold start” samples so it’s a big problem that may lead to the differences of the result.

Data Preprocessing

1.Merge of the metadata and the train&test data

We merged the train&test dataset with songs,members,song_extra_info datasets using the primary key[‘msno’] and [‘song_id’].Then we got the alpha train&test datasets.

2.The **alpha data features**

Categorical features:

artist_name
composer
gender
genre_ids
lyricist
msno
song_id
source_screen_name
source_type
source

All these categorical features were applied the “fillna” function with “Unknown”

numerical features:

rest of the features

song_length NaN was filled with 200000
bd NaN was filled with mode number and constrain the range to [0,75]
song_year was filled with median number

3.The **beta data features**

Adding song basic features:

membership_days
registration_year
registration_month
registration_date
expiration_year
expiration_month
expiration_date

All these features were about **time series**

Feature Engineering

In this part, we tried several different new features and then we drop them or reserve them considering the contribution of the online score.

1.lyricist_count
Calculate the number of lyricist of each song

2.composer_count
Calculate the number of composer of each song

3.**count_song_played**
Calculate the number of play of each song

4.count_artist_played
Calculate the number of play of each artist

5.m_c
Calculate every member's number of **play behaviors**

6.m_a_c
Calculate every member's number of play behaviors of **each artist**

7.m_s_c
Calculate every member's number of play behaviors of **each source_screen_name**

8.m_st_c
Calculate every member's number of play behaviors of **each source_type**

9.m_c_c
Calculate every member's number of play behaviors of **each composer**

10.m_g_c
Calculate every member's number of play behaviors of **each gener_id**

10.m_a_s_c
Calculate every member's number of play behaviors of **each source_screen_name and artist**

11.m_a_c_ratio
Calculate every artist's occupation of all play records of one member

12.m_s_c_ratio
Calculate every source_screen_name's occupation of all play records of one member

13.m_st_c_ratio
Calculate every source_type's occupation of all play records of one member

14.m_c_c_ratio
Calculate every composer's occupation of all play records of one member

15.m_a_s_c_ratio
Calculate every source_screen_name and artist's occupation of all play records of one member

Training with LightGBM

Parameters:

'objective': 'binary',
'boosting': 'gbdt',
'learning_rate': 0.05 ,
'verbose': 0,
'num_leaves': 128,
'bagging_fraction': 0.9,
'bagging_freq': 1,
'bagging_seed': 2017,
'feature_fraction': 0.9,
'feature_fraction_seed': 2017,
'max_bin': 512,
'max_depth': 16,
'num_rounds': 1500,
'metric' : 'auc'

In this part we found that when we set the num_rounds with more than 1500, overfitting will occur. And when the rounds was set less than 1500, the model couldn't touch the limitation of it's performance.Other phenomena was when we use the whole train dataset as the **watchlist** to calculate the AUC value, the closer the final AUC was to 0.84, the better the online result could get. That also supported that about 18% cold start samples would influence the result.

Key parameters:

1.learning_rate
2.num_leaves
3.feature_fraction
4.max_depth

Because the training time was more than 1h(num_rounds = 1500), GridSearch is not a proper method to attune the parameters.