Feature Engineered : 0.68310

Python notebook using data from WSDM - KKBox's Music Recommendation Challenge · 7,540 views

64 Fork 171

Version 5
5 commits

Notebook
Data
Log
Comments

In [1]:

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load in

import numpy as np
import pandas as pd
import lightgbm as lgb
import datetime
import math
import gc


# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list the files in the
 input directory

from subprocess import check_output
print(check_output(["ls", "../input"]).decode("utf8"))

# Any results you write to the current directory are saved as output.
```

```
members.csv
sample_submission.csv
song_extra_info.csv
songs.csv
test.csv
train.csv
```

In [2]:

```
print('Loading data...')
data_path = '../input/'
train = pd.read_csv(data_path + 'train.csv', dtype={'msno' : 'category',
                                                    'source_system_tab' : 'category',
                                                     'source_screen_name' : 'category',
                                                     'source_type' : 'category',
                                                     'target' : np.uint8,
                                                     'song_id' : 'category'})
test = pd.read_csv(data_path + 'test.csv', dtype={'msno' : 'category',
                                                  'source_system_tab' : 'category',
                                                  'source_screen_name' : 'category',
                                                  'source_type' : 'category',
                                                  'song_id' : 'category'})
songs = pd.read_csv(data_path + 'songs.csv',dtype={'genre_ids': 'category',
                                                   'language' : 'category',
                                                   'artist_name' : 'category',
                                                   'composer' : 'category',
                                                   'lyricist' : 'category',
                                                   'song_id' : 'category'})
members = pd.read_csv(data_path + 'members.csv',dtype={'city' : 'category',
                                                       'bd' : np.uint8,
                                                       'gender' : 'category',
                                                       'registered_via' : 'category'},
                      parse_dates=['registration_init_time','expiration_date'])
songs_extra = pd.read_csv(data_path + 'song_extra_info.csv')
print('Done loading...')
```

```
Loading data...


---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-2-6e0df5b8adcf> in <module>()
      6                                                     'source_type' : 'category',
      7                                                     'target' : np.uint8,
----> 8                                                     'song_id' : 'category'})
      9 test = pd.read_csv(data_path + 'test.csv', dtype={'msno' : 'category',
     10                                                   'source_system_tab' : 'category',

/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in parser_f(filepath_or_buffer, se
p, delimiter, header, names, index_col, usecols, squeeze, prefix, mangle_dupe_cols, dtype, engi
ne, converters, true_values, false_values, skipinitialspace, skiprows, nrows, na_values, keep_d
```

```
efault_na, na_filter, verbose, skip_blank_lines, parse_dates, infer_datetime_format, keep_date_
col, date_parser, dayfirst, iterator, chunksize, compression, thousands, decimal, lineterminato
r, quotechar, quoting, escapechar, comment, encoding, dialect, tupleize_cols, error_bad_lines,
 warn_bad_lines, skipfooter, skip_footer, doublequote, delim_whitespace, as_recarray, compact_i
nts, use_unsigned, low_memory, buffer_lines, memory_map, float_precision)
    703                     skip_blank_lines=skip_blank_lines)
    704
--> 705         return _read(filepath_or_buffer, kwds)
    706
    707     parser_f.__name__ = name

/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in _read(filepath_or_buffer, kwds)
    449
    450     try:
--> 451         data = parser.read(nrows)
    452     finally:
    453         parser.close()

/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in read(self, nrows)
   1063                 raise ValueError('skipfooter not supported for iteration')
   1064
-> 1065         ret = self._engine.read(nrows)
   1066
   1067         if self.options.get('as_recarray'):

/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in read(self, nrows)
   1826     def read(self, nrows=None):
   1827         try:
-> 1828             data = self._reader.read(nrows)
   1829         except StopIteration:
   1830             if self._first_chunk:

pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader.read()

pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._read_low_memory()

pandas/_libs/parsers.pyx in pandas._libs.parsers._concatenate_chunks()

/opt/conda/lib/python3.6/site-packages/numpy/core/numerictypes.py in find_common_type(array_typ
es, scalar_types)
   1013
   1014     """
-> 1015     array_types = [dtype(x) for x in array_types]
   1016     scalar_types = [dtype(x) for x in scalar_types]
   1017

                    .
   1830             if self._first_chunk:

pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader.read()

pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._read_low_memory()

pandas/_libs/parsers.pyx in pandas._libs.parsers._concatenate_chunks()

/opt/conda/lib/python3.6/site-packages/numpy/core/numerictypes.py in find_common_type(array_typ
es, scalar_types)
   1013
   1014     """
-> 1015     array_types = [dtype(x) for x in array_types]
   1016     scalar_types = [dtype(x) for x in scalar_types]
   1017

/opt/conda/lib/python3.6/site-packages/numpy/core/numerictypes.py in <listcomp>(.0)
   1013
   1014     """
-> 1015     array_types = [dtype(x) for x in array_types]
   1016     scalar_types = [dtype(x) for x in scalar_types]
   1017

TypeError: data type not understood
   1016     scalar_types = [dtype(x) for x in scalar_types]
   1017

/opt/conda/lib/python3.6/site-packages/numpy/core/numerictypes.py in <listcomp>(.0)
   1013
   1014     """
-> 1015     array_types = [dtype(x) for x in array_types]
   1016     scalar_types = [dtype(x) for x in scalar_types]
   1017

TypeError: data type not understood
```

In [3]:

```
print('Data merging...')


train = train.merge(songs, on='song_id', how='left')
test = test.merge(songs, on='song_id', how='left')

members['membership_days'] = members['expiration_date'].subtract(members['registration_init_tim
e']).dt.days.astype(int)

members['registration_year'] = members['registration_init_time'].dt.year
```

```
   1067         if self.options.get('as_recarray'):

/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in read(self, nrows)
   1826     def read(self, nrows=None):
   1827         try:
-> 1828             data = self._reader.read(nrows)
   1829         except StopIteration:
   1830             if self._first_chunk:

pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader.read()

pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._read_low_memory()

pandas/_libs/parsers.pyx in pandas._libs.parsers._concatenate_chunks()

/opt/conda/lib/python3.6/site-packages/numpy/core/numerictypes.py in find_common_type(array_typ
es, scalar_types)
   1013
   1014     """
-> 1015     array_types = [dtype(x) for x in array_types]
   1016     scalar_types = [dtype(x) for x in scalar_types]
   1017

/opt/conda/lib/python3.6/site-packages/numpy/core/numerictypes.py in <listcomp>(.0)
```

```
test.song_id = test.song_id.astype('category')

# import gc
# del members, songs; gc.collect();

print('Done merging...')
```

```
Data merging...


---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-3-1432efb892c8> in <module>()
      2
      3
----> 4 train = train.merge(songs, on='song_id', how='left')
      5 test = test.merge(songs, on='song_id', how='left')
      6

NameError: name 'train' is not defined
```

In [4]:

```
print ("Adding new features")

        return sum(map(x.count, ['|', '/', '\\', ';'])) + 1
    return sum(map(x.count, ['|', '/', '\\', ';']))

train['lyricist'].fillna('no_lyricist',inplace=True)
test['lyricist'].fillna('no_lyricist',inplace=True)
train['lyricists_count'] = train['lyricist'].apply(lyricist_count).astype(np.int8)
test['lyricists_count'] = test['lyricist'].apply(lyricist_count).astype(np.int8)

def composer_count(x):
    if x == 'no_composer':
        return 0
    else:
        return sum(map(x.count, ['|', '/', '\\', ';'])) + 1

train['composer'].fillna('no_composer',inplace=True)
test['composer'].fillna('no_composer',inplace=True)
train['composer_count'] = train['composer'].apply(composer_count).astype(np.int8)
test['composer_count'] = test['composer'].apply(composer_count).astype(np.int8)

def is_featured(x):
    if 'feat' in str(x) :
        return 1
    return 0
```

```
test.song_id = test.song_id.astype('category')

# import gc
# del members, songs; gc.collect();

print('Done merging...')
```

```
Data merging...


---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-3-1432efb892c8> in <module>()
      2
      3
----> 4 train = train.merge(songs, on='song_id', how='left')
      5 test = test.merge(songs, on='song_id', how='left')
      6

NameError: name 'train' is not defined
```

In [4]:

```
print ("Adding new features")
```

```
TypeError                                 Traceback (most recent call last)
<ipython-input-2-6e0df5b8adcf> in <module>()
      6                                                 'source_type' : 'category',
      7                                                 'target' : np.uint8,
----> 8                                                 'song_id' : 'category'})
      9 test = pd.read_csv(data_path + 'test.csv', dtype={'msno' : 'category',
     10                                                   'source_system_tab' : 'category',

/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in parser_f(filepath_or_buffer, se
p, delimiter, header, names, index_col, usecols, squeeze, prefix, mangle_dupe_cols, dtype, engi
ne, converters, true_values, false_values, skipinitialspace, skiprows, nrows, na_values, keep_d
efault_na, na_filter, verbose, skip_blank_lines, parse_dates, infer_datetime_format, keep_date_
col, date_parser, dayfirst, iterator, chunksize, compression, thousands, decimal, lineterminato
r, quotechar, quoting, escapechar, comment, encoding, dialect, tupleize_cols, error_bad_lines,
 warn_bad_lines, skipfooter, skip_footer, doublequote, delim_whitespace, as_recarray, compact_i
nts, use_unsigned, low_memory, buffer_lines, memory_map, float_precision)
    703                     skip_blank_lines=skip_blank_lines)
    704
--> 705         return _read(filepath_or_buffer, kwds)
    706
    707     parser_f.__name__ = name

/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in _read(filepath_or_buffer, kwds)
    449
```

```
train['smaller_song'] = train['song_length'].apply(smaller_song).astype(np.int8)
test['smaller_song'] = test['song_length'].apply(smaller_song).astype(np.int8)

# number of times a song has been played before
_dict_count_song_played_train = {k: v for k, v in train['song_id'].value_counts().iteritems()}
_dict_count_song_played_test = {k: v for k, v in test['song_id'].value_counts().iteritems()}
def count_song_played(x):
    try:
        return _dict_count_song_played_train[x]
    except KeyError:
        try:
            return _dict_count_song_played_test[x]
        except KeyError:
            return 0


train['count_song_played'] = train['song_id'].apply(count_song_played).astype(np.int64)
test['count_song_played'] = test['song_id'].apply(count_song_played).astype(np.int64)

# number of times the artist has been played
_dict_count_artist_played_train = {k: v for k, v in train['artist_name'].value_counts().iterite
ms()}
_dict_count_artist_played_test = {k: v for k, v in test['artist_name'].value_counts().iteritems
train['count_artist_played'] = train['artist_name'].apply(count_artist_played).astype(np.int64)
test['count_artist_played'] = test['artist_name'].apply(count_artist_played).astype(np.int64)


print "Done adding features"
```

```
  File "<ipython-input-4-3bc4a73ec5a5>", line 116
    print "Done adding features"
                               ^
SyntaxError: Missing parentheses in call to 'print'. Did you mean print("Done adding feature
s")?
```

In [5]:

```
print ("Train test and validation sets")
for col in train.columns:
    if train[col].dtype == object:
        train[col] = train[col].astype('category')
        test[col] = test[col].astype('category')


X_train = train.drop(['target'], axis=1)
y_train = train['target'].values


X_test = test.drop(['id'], axis=1)
ids = test['id'].values


# del train, test; gc.collect();

d_train_final = lgb.Dataset(X_train, y_train)
watchlist_final = lgb.Dataset(X_train, y_train)
print('Processed data...')
```

```
Train test and validation sets


---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-5-6f773e45101b> in <module>()
      1 print ("Train test and validation sets")
----> 2 for col in train.columns:
      3     if train[col].dtype == object:
      4         train[col] = train[col].astype('category')
      5         test[col] = test[col].astype('category')

NameError: name 'train' is not defined
```

In [6]:

```
params = {
        'objective': 'binary',
        'metric': 'binary_logloss',
        'boosting': 'gbdt',
        'learning_rate': 0.3 ,
        'verbose': 0,
        'num_leaves': 108,
        'bagging_fraction': 0.95,
        'bagging_freq': 1,
        'bagging_seed': 1,
        'feature_fraction': 0.9,
        'feature_fraction_seed': 1,
        'max_bin': 256,
        'max_depth': 10,
        'num_rounds': 200,
        'metric' : 'auc'
    }

%time model_f1 = lgb.train(params, train_set=d_train_final,  valid_sets=watchlist_final, verbos
e_eval=5)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-6-f357cb29e712> in <module>()
     17     }
     18
---> 19 get_ipython().magic('time model_f1 = lgb.train(params, train_set=d_train_final,  valid_
sets=watchlist_final, verbose_eval=5)')

/opt/conda/lib/python3.6/site-packages/IPython/core/interactiveshell.py in magic(self, arg_s)
   2156         magic_name, _, magic_arg_s = arg_s.partition(' ')
   2157         magic_name = magic_name.lstrip(prefilter.ESC_MAGIC)
-> 2158         return self.run_line_magic(magic_name, magic_arg_s)
   2159
   2160     #-------------------------------------------------------------------------

/opt/conda/lib/python3.6/site-packages/IPython/core/interactiveshell.py in run_line_magic(self,
 magic_name, line)
   2077                 kwargs['local_ns'] = sys._getframe(stack_depth).f_locals
   2078             with self.builtin_trap:
-> 2079                 result = fn(*args,**kwargs)
   2080             return result
```

```
<decorator-gen-59> in time(self, line, cell, local_ns)

/opt/conda/lib/python3.6/site-packages/IPython/core/magic.py in <lambda>(f, *a, **k)
    186     # but it's overkill for just that one bit of state.
    187     def magic_deco(arg):
--> 188         call = lambda f, *a, **k: f(*a, **k)
    189
    190         if callable(arg):

/opt/conda/lib/python3.6/site-packages/IPython/core/magics/execution.py in time(self, line, cel
l, local_ns)
   1183         else:
   1184             st = clock2()
-> 1185             exec(code, glob, local_ns)
   1186             end = clock2()
   1187             out = None

<timed exec> in <module>()

NameError: name 'd_train_final' is not defined
```

In [7]:

```
params = {
        'objective': 'binary',
        'metric': 'binary_logloss',
        'boosting': 'dart',
        'learning_rate': 0.3 ,
        'verbose': 0,
        'num_leaves': 108,
        'bagging_fraction': 0.95,
        'bagging_freq': 1,
        'bagging_seed': 1,
        'feature_fraction': 0.9,
        'feature_fraction_seed': 1,
        'max_bin': 256,
        'max_depth': 10,
        'num_rounds': 200,
        'metric' : 'auc'
    }

%time model_f2 = lgb.train(params, train_set=d_train_final,  valid_sets=watchlist_final, verbos
e_eval=5)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-7-063832c731db> in <module>()
     17     }
     18
---> 19 get_ipython().magic('time model_f2 = lgb.train(params, train_set=d_train_final,  valid_
sets=watchlist_final, verbose_eval=5)')

/opt/conda/lib/python3.6/site-packages/IPython/core/interactiveshell.py in magic(self, arg_s)
   2156         magic_name, _, magic_arg_s = arg_s.partition(' ')
   2157         magic_name = magic_name.lstrip(prefilter.ESC_MAGIC)
-> 2158         return self.run_line_magic(magic_name, magic_arg_s)
   2159
   2160     #-------------------------------------------------------------------------

/opt/conda/lib/python3.6/site-packages/IPython/core/interactiveshell.py in run_line_magic(self,
 magic_name, line)
   2077                 kwargs['local_ns'] = sys._getframe(stack_depth).f_locals
   2078             with self.builtin_trap:
-> 2079                 result = fn(*args,**kwargs)
   2080             return result
   2081

<decorator-gen-59> in time(self, line, cell, local_ns)

/opt/conda/lib/python3.6/site-packages/IPython/core/magic.py in <lambda>(f, *a, **k)
    186     # but it's overkill for just that one bit of state.
    187     def magic_deco(arg):
--> 188         call = lambda f, *a, **k: f(*a, **k)
    189
    190         if callable(arg):

/opt/conda/lib/python3.6/site-packages/IPython/core/magics/execution.py in time(self, line, cel
l, local_ns)
   1183         else:
   1184             st = clock2()
-> 1185             exec(code, glob, local_ns)
```

```
   1186             end = clock2()
   1187             out = None

<timed exec> in <module>()

NameError: name 'd_train_final' is not defined
```

In [8]:

```
print('Making predictions')
p_test_1 = model_f1.predict(X_test)
p_test_2 = model_f2.predict(X_test)
p_test_avg = np.mean([p_test_1, p_test_2], axis = 0)


print('Done making predictions')
```

```
Making predictions


---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-8-dc6069237382> in <module>()
      1 print('Making predictions')
----> 2 p_test_1 = model_f1.predict(X_test)
      3 p_test_2 = model_f2.predict(X_test)
      4 p_test_avg = np.mean([p_test_1, p_test_2], axis = 0)
      5

NameError: name 'model_f1' is not defined
```

In [9]:

```
print ('Saving predictions Model model of gbdt')

subm = pd.DataFrame()
subm['id'] = ids
subm['target'] = p_test_avg
subm.to_csv(data_path + 'submission_lgbm_avg.csv.gz', compression = 'gzip', index=False, float_
format = '%.5f')

print('Done!')
```

```
Saving predictions Model model of gbdt


---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-9-2366982ed7da> in <module>()
      2
      3 subm = pd.DataFrame()
----> 4 subm['id'] = ids
      5 subm['target'] = p_test_avg
      6 subm.to_csv(data_path + 'submission_lgbm_avg.csv.gz', compression = 'gzip', index=False
, float_format = '%.5f')

NameError: name 'ids' is not defined
```

Data

Data Sources

- WSDM - KKBox's Music Recomm...
  - members.csv.7z
    - members.csv
  - sample_submission.csv.7z
    - sample_submission.csv
  - song_extra_info.csv.7z
    - song_extra_info.csv
  - songs.csv.7z
    - songs.csv
  - test.csv.7z
    - test.csv
  - train.csv.7z
    - train.csv

WSDM - KKBox's Music Recommendation Challenge

Can you build the best music recommendation system?

Last Updated: a year ago

About this Competition

In this task, you will be asked to predict the chances of a user listening to a song repetitively after the first observable listening event within a time window was triggered. If there are recurring listening event(s) triggered within a month after the user's very first observable listening event, its target is marked 1, and 0 otherwise in the training set. The same rule applies to the testing set.

KKBOX provides a training data set consists of information of the first observable listening event for each unique user-song pair within a specific time duration. Metadata of each unique user and song pair is also provided. The use of public data to increase the level of accuracy of your prediction is encouraged.

The train and the test data are selected from users listening history in a given time period. Note that this time period is chosen to be before the WSDM-KKBox Churn Prediction time period. The train and test sets are split based on time, and the split of public/private are based on unique user/song pairs.

Tables

train.csv

- msno: user id
- song_id: song id
- source_system_tab: the name of the tab where the event was triggered. System tabs are used to categorize KKBOX mobile apps functions. For example, tab `my library` contains functions to manipulate the local storage, and tab `search` contains functions relating to search.
- source_screen_name: name of the layout a user sees.
- source_type: an entry point a user first plays music on mobile apps. An entry point could be `album`, `online-playlist`, `song` .. etc.
- target: this is the target variable. `target=1` means there are recurring listening

Run Info

Succeeded	True	Run Time	3614.1 seconds
Exit Code	0	Queue Time	0 seconds
Docker Image Name	kaggle/python(Dockerfile)	Output Size	0
Timeout Exceeded	False	Used All Space	False
Failure Message			

Log

Download Log

```
Time  Line #  Log Message
           1  [{
           2    "data": "[NbConvertApp] Converting notebook __temp_notebook_source__.ipynb to html\n",
           3    "stream_name": "stderr",
           4    "time": 1.5475757209933363
           5  },{
           6    "data": "[NbConvertApp] Writing 308692 bytes to __results__.html\n",
           7    "stream_name": "stderr",
           8    "time": 1.7093351599760354
           9  }{
          10    "data": "[NbConvertApp] Converting notebook __temp_notebook_source__.ipynb to notebook\n",
          11    "stream_name": "stderr",
          12    "time": 1.5472121229977347
          13  },{
          14    "data": "[NbConvertApp] Executing notebook with kernel: python3\n",
          15    "stream_name": "stderr",
          16    "time": 1.5799874399672262
          17  },{
          18    "data": "Fontconfig warning: ignoring C.UTF-8: not a valid language tag\n",
          19    "stream_name": "stderr",
          20    "time": 2.548812124005053
          21  },{
          22    "data": "[NbConvertApp] Writing 45552 bytes to __notebook__.ipynb\n",
          23    "stream_name": "stderr",
          24    "time": 20.307915177952964
          25  }{
          26    "data": "[NbConvertApp] Converting notebook __notebook__.ipynb to html\n",
          27    "stream_name": "stderr",
          28    "time": 1.647211093979422
          29  },{
          30    "data": "[NbConvertApp] Writing 341770 bytes to __results__.html\n",
          31    "stream_name": "stderr",
          32    "time": 1.845204734010622
          33  }
          34
          36  Complete. Exited with code 0.
```

Comments (49)

Sort by: All Comments | Hotness

Click here to enter a comment...

Ferz • Posted on Latest Version • a year ago • Options • Reply

2

And one more question. Why did you use in the first lgbm 'gbdt' and in the second one 'dart'?

kmaster • Posted on Latest Version • a year ago • Options • Reply

2

Thank you for sharing.
When I try to run:

train['artistcomposer'] = (train['artistname'] == train['composer']).astype(np.int8)

I got this error:
TypeError: Categoricals can only be compared if 'categories' are the same. Categories are different lengths

However, len(train['artist_name']) == len(train['composer']) is true.
Can you please show me some directions for this?

Jingxi Xu • Posted on Latest Version • a year ago • Options • Reply

1

Why is everyone saying that it is beautiful while it actually does not work.

蜗牛爱上星星 • Posted on Latest Version • a year ago • Options • Reply

1

It works.

蜗牛爱上星星 • Posted on Latest Version • a year ago • Options • Reply

0

python 2.7, pandas 0.19

Randolph • Posted on Latest Version • a year ago • Options • Reply

0

Thank you. Is there any other EDA work can improve the model and make the prediction result better? I mean a real EDA work because some Feature work we do is not helpful.
Since i analysis the data and make some feature work which i think maybe helpful but it harms the model performance in actual. So anyone can share the useful EDA work?

Alexander Kireev • Posted on Version 2 • a year ago • Options • Reply

1

Thank you. Beautiful work.

Max Yue • Posted on Version 2 • a year ago • Options • Reply

0

Beautiful!!

Max Yue • Posted on Version 2 • a year ago • Options • Reply

0

Great feature extraction! When someone doing feature extraction, feature analysis will normally follow, so we can see which features contribute the model.

asmita vikas Kernel Author • Posted on Version 2 • a year ago • Options • Reply

1

noted! I'll add that next. thank you

Maple • Posted on Version 2 • a year ago • Options • Reply

0

great work.

Maple • Posted on Version 2 • a year ago • Options • Reply

0

why train['countsongplayed'] and train['countsongplayed'] have some negative values.

asmita vikas Kernel Author • Posted on Version 2 • a year ago • Options • Reply

1

oops! that's coz I was making the column as np.int8 but the values were of larger.
Fixed it to int64

Maple • Posted on Latest Version • a year ago • Options • Reply

0

I want to know why song language 17 or 45 are set to 1 and the others are set to 0.

SuperFrankie • Posted on Version 2 • a year ago • Options • Reply

0

Great job.

zhihuidayong • Posted on Version 2 • a year ago • Options • Reply 0

I only got 0.64 score with the same code

asmita vikas Kernel Author • Posted on Version 4 • a year ago • Options • Reply 2

hmm, that sounds weird. I did make some changes now and got 0.68
Maybe you could try again?

lystdo • Posted on Version 4 • a year ago • Options • Reply 2

You may check the version of python.

zhihuidayong • Posted on Latest Version • a year ago • Options • Reply 0

Yeah, I use version 2.6, is there something with that?

zhihuidayong • Posted on Latest Version • a year ago • Options • Reply 0

I got 0.66 score this time, why are we different so much

lystdo • Posted on Latest Version • a year ago • Options • Reply 0

There are several differences between Python 2 and Python 3. For example, "1 / 2" will give 0 in Python 2, while 0.5 in Python 3. Besides, the version of LightGBM may also bring some differences.

Shubham Gupta • Posted on Latest Version • a year ago • Options • Reply 0

I am getting a low score of 0.62, although I am using pandas=0.19 and python 2.7, and secondly, what about the remaining null values in other features, do we replace them with a constant value or leave them as it is?? can you please share the version of lightgbm and sklearn please.

DawuChen • Posted on Latest Version • a year ago • Options • Reply 0

Hey!Thx for your nice work.But can I ask how you solve the problem of "TypeError: data type not understood" when read_csv use category dtype? I also encountered the same problem. Please help.

asmita vikas Kernel Author • Posted on Latest Version • a year ago • Options • Reply 1

Well, I had written the script for Python2.7 and it ran fine.
I'm guessing its a problem with Python3. I'll try to fix it asap. Meanwhile could you try it on another version?

DawuChen • Posted on Latest Version • a year ago • Options • Reply 0

I am also using python 2.7 but got the same error. I had tried update pandas to the newest version but it still can't work. By the way , I found sometimes dtype='category' can work on numeric type but always not on object type,it's wierd...

DawuChen • Posted on Latest Version • a year ago • Options • Reply 1

Got the reason,it is because the version of pandas is too new, there is bug in version 0.21.0
see:
[https://www.kaggle.com/c/kkbox-music-recommendation-challenge/discussion/43352][1]

Fuzzylogic • Posted on Latest Version • a year ago • Options • Reply 0

Same for me.
Any workaround ..? I am also looking

DawuChen • Posted on Latest Version • a year ago • Options • Reply 1

As far, there is no workaround to use 'category' when read csv. But you can use 'astype' function to change dtype after you read data into memory.

Vinay Patlolla • Posted on Latest Version • a year ago • Options • Reply 0

downgrade pandas to 0.19

Fuzzylogic • Posted on Latest Version • a year ago • Options • Reply 0

great work..thanks a ton for sharing !

Jingxi Xu • Posted on Latest Version • a year ago • Options • Reply 0

train['artist*composer'] = (train['artist*name'] == train['composer']).astype(np.int8) does not work with the error that "Categoricals can only be compared if 'categories' are the same. Categories are different lengths". Can I have any hints on that?

Layla • Posted on Latest Version • a year ago • Options • Reply 0

I got the same error. Did you figure it out?

asmita vikas Kernel Author • Posted on Latest Version • a year ago • Options • Reply 2

it could be an issue with pandas version. Can you try downgrading pandas to 0.19?

BrianHur • Posted on Latest Version • a year ago • Options • Reply 2

Thanks! I had the same error, but downgrading to pandas 0.19 fixed it for me

WendyCui • Posted on Latest Version • a year ago • Options • Reply 0

Hi, thank you for sharing!! But there's a part of your code confusing me. Why you just considered two values(17, 45) of 'language' attribute in songs.csv? Could you please give me some hints on this?

asmita vikas Kernel Author • Posted on Latest Version • a year ago • Options • Reply 2

Well, if you look at the distributions, song_lang with values 17 and 45 have the highest percentage of 1s in response variable. So I took them as high priority

a year ago

This Comment was deleted.

asmita vikas Kernel Author • Posted on Latest Version • a year ago • Options • Reply 0

Response variable is the variable we try to predict, in this case "target". It is also called the dependent variable.

a year ago

This Comment was deleted.

a year ago

This Comment was deleted.

Felicia • Posted on Latest Version • a year ago • Options • Reply 0

TypeError: data type not understood
I got the same error.
How did you fix that? Category dtype is not recognized in read_csv

deathstar • Posted on Latest Version • a year ago • Options • Reply 0

Great work on feature engineering.
++

cowarder • Posted on Latest Version • a year ago • Options • Reply 0

Thank you for your great work! But I got some trouble while running the code "train['genreids'].fillna('nogenre_id',inplace=True)". The error message is "ValueError: fill value must be in categories",how can I solve this?

Yihong Wang • Posted on Latest Version • a year ago • Options • Reply

1

train['genreids'].cat.addcategories("nogenreid").fillna('nogenreid',inplace=True)

Ferz • Posted on Latest Version • a year ago • Options • Reply

0

Hi, thank you for sharing. Nice work. I am a beginner. Can you tell me why did you use 2 different lgbm? Did you find out that their average is better than single one?

LongYin • Posted on Latest Version • 4 months ago • Options • Reply

0

This is just for ensemble.

zouyu • Posted on Latest Version • a year ago • Options • Reply

0

good jobs,i will use some time to research your code !thank you very much!

KimchiSoup • Posted on Latest Version • a year ago • Options • Reply

0

it was helpful to me. +_+d

rbaral • Posted on Latest Version • 9 months ago • Options • Reply

0

How did you come with the parameters for the LGB models?