Advertisement CTR Prediction

Data Preparation

DS5220 / Fall 2023 Semester

Team Members: Liyang Song, Qian Yin

Oct 29, 2023

Identify required transformations

Based on the exploratory data analysis conducted, we identify three required transformations.

- The first one is to **drop 9 attributes** listed below.
  - *app_score*
  - *hist_on_shelf_time*
  - *task_id*, *spread_app_id*
  - *Tags*
  - *dev_id*
  - *app_second_class*
  - *adv_prim_id*
  - *device_price*
- The third transformation is **target encoding** for categorical data.
  - Target encoding is implemented to transform data into numerical values.
- The second transformation is to perform data **standardization**.
  - We choose standardization as the method to re-scale features so that values of each feature have zero mean and unit variance.
  - It is less sensitive to outliers than normalization.

Identify unnecessary transformations

The following data transformation steps **will not be included** in our project

- **Drop outliers**

  Some data points are significantly different from the rest of dataset, but they will not be considered as ‘outliers’ based on the fact that all attributes are categorical and pre-encoded.

- **Fill/drop missingness**

  Because no missing value is found, this step is skipped.

- **Discretize continuous features**

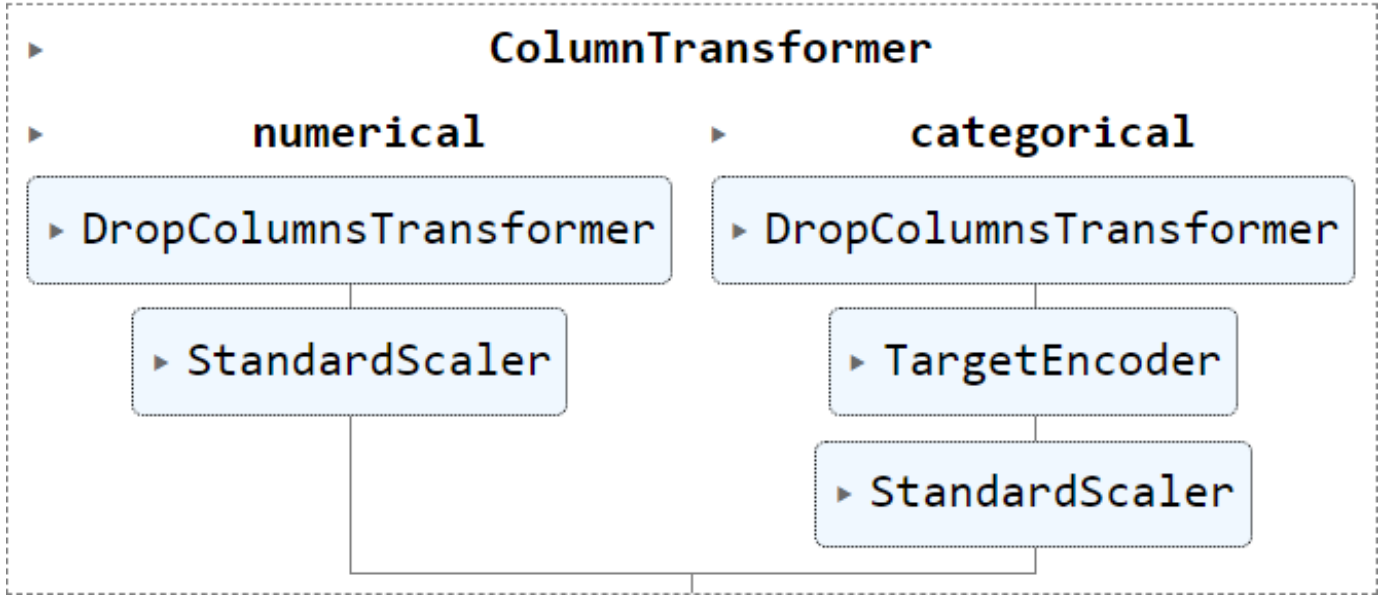
  Based on our observation, many of the attributes in the original data set are probably discretized already.

Identify unnecessary transformations - continued

- **Decompose features**
  We lack information to understand the context; therefore, decomposing these transformed features is unnecessary.
- **Transform features**
  All features are categorical; therefore, log transformation or square root transformation is not considered.
- **Aggregate features into new features**
  The actual meanings of numerous values are unclear, preventing us from constructing meaningful and useful features.

A pipeline of transformers

- A pipeline of transformers is created to perform several important transformations listed below.



EDA on transformed train set: General Information

```
General information of DataFrame:
df.shape:
(833795, 27)
```

```
 22  communication_onlinerate     833795 non-null  float64
 23  communication_avgonline_30d  833795 non-null  float64
 24  indu_name                    833795 non-null  float64
 25  pt_d                         833795 non-null  float64
 26  label                        833795 non-null  int64
dtypes: float64(26), int64(1)
```

	uid	adv_id	creat_type_cd	inter_type_cd	slot_id	app_first_class	age	city	city_rank	device_n
0	-0.234420	0.269338	0.491277	0.183077	-0.542126	-1.906517	-0.128911	0.387468	0.641769	1.12
1	0.181457	-0.694659	0.473148	0.180165	-0.253454	0.521003	-0.511085	-0.810074	-0.913338	0.53
2	0.181457	-0.263245	0.493618	1.490727	-0.229145	0.531151	-0.479514	-0.787627	-0.868357	-0.26
3	-0.234420	-0.802140	0.493618	0.193170	-0.555835	0.531151	-0.685093	-0.963011	-0.868357	-0.09
4	-0.234420	0.572652	0.493618	0.193170	-1.075084	0.531151	-0.479514	-0.660634	0.622953	-0.92

	uid	adv_id	creat_type_cd	inter_type_cd	slot_id	app_first_class	age	
count	8.337950e+05	8.337950e+05	8.337950e+05	8.337950e+05	8.337950e+05	8.337950e+05	8.337950e+05	8.3379
mean	1.146740e-12	1.527000e-13	2.110952e-11	-1.417973e-11	-4.804665e-13	1.289028e-11	-5.906684e-12	3.4174
std	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.0000
min	-2.344204e-01	-1.237773e+00	-3.411577e+00	-2.514619e+00	-1.197866e+00	-1.906517e+00	-6.850932e-01	-4.7514
25%	-2.344204e-01	-6.740555e-01	4.731478e-01	1.801647e-01	-6.563090e-01	5.210034e-01	-6.510223e-01	-6.5365
50%	-2.344204e-01	-3.130989e-01	4.778666e-01	1.874668e-01	-2.534537e-01	5.311512e-01	-2.300270e-01	-2.0957
75%	1.814570e-01	4.465355e-01	4.912766e-01	1.931700e-01	-1.385133e-04	5.353783e-01	-1.398657e-01	3.8834
max	1.181410e+01	3.473508e+01	1.130770e+00	1.725073e+00	1.825819e+00	5.376175e-01	8.993419e+00	1.0348

- Smaller data set
  - Rows: 838142 > 833795
  - Columns: 36 > 27
- Modified data types
  - The data types of most all attributes are changed from integer to float.
  - No object data type exists.
- Updated values in the descriptive statistics table
  - The dispersion and shape of data distribution is changed.

EDA on transformed train set: Duplication and missingness

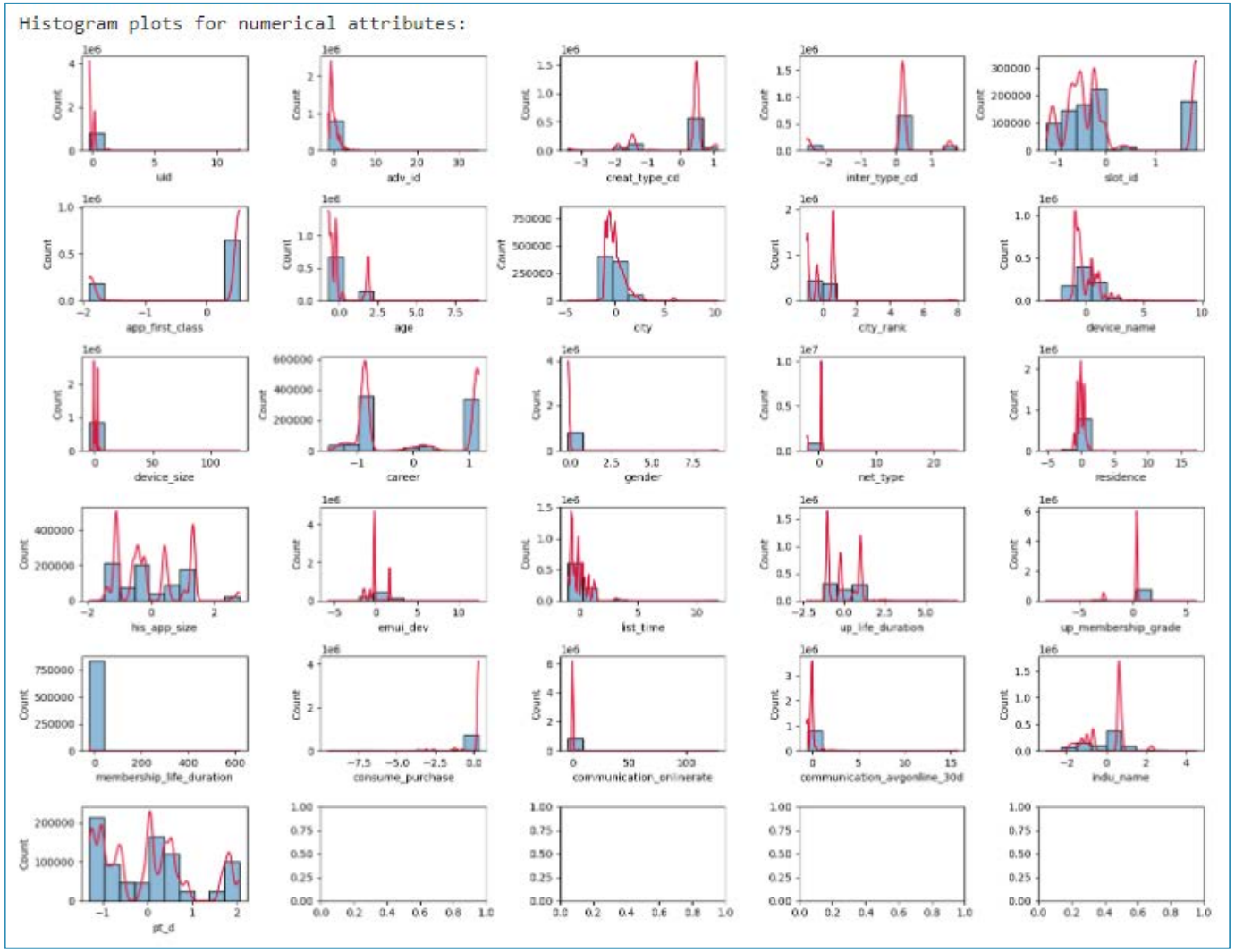
```
Check out duplicate observations:
df.shape:  (833795, 27)
drop_dup_df.shape:  (833777, 27)
Caution: data set contains duplicate observations!!!
```

- **New duplicate values** are found.
  This existence of new duplicate values is possible when the size of this data set is large and ranges of values are small.

```
Check out missingness:
No missing values in data set.
```

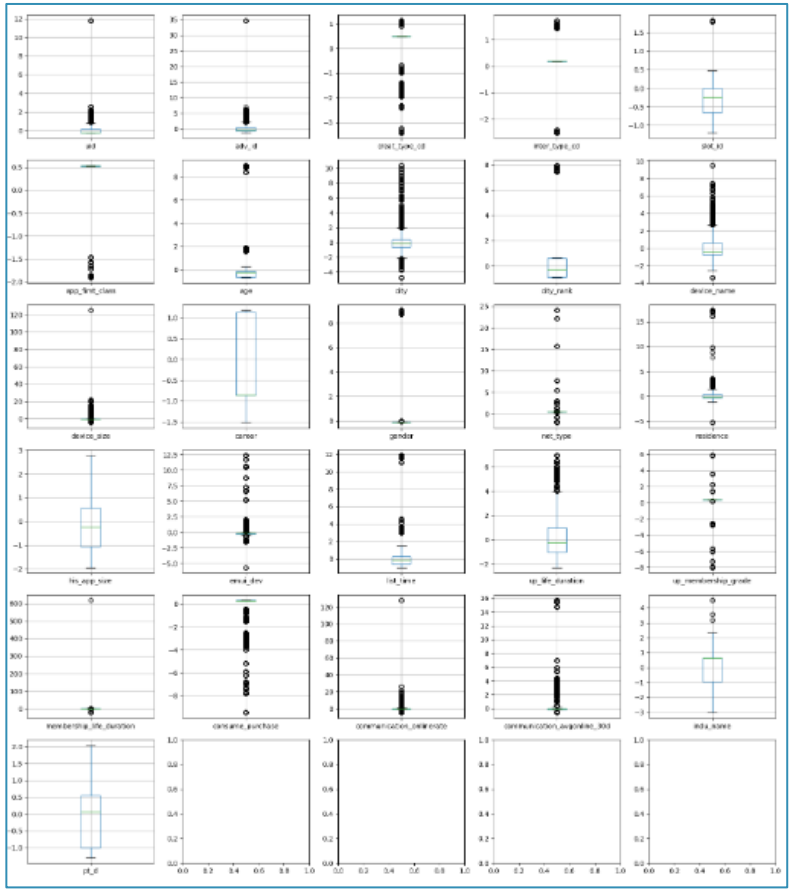
- **No missing value** exists in both features and the target variable.

EDA on transformed train set: Distribution



- The effect of standardization is shown on the histogram plots (x-axis).
- The number of plots is decreased because of dropped attributes.

EDA on transformed train set: Outlier detection



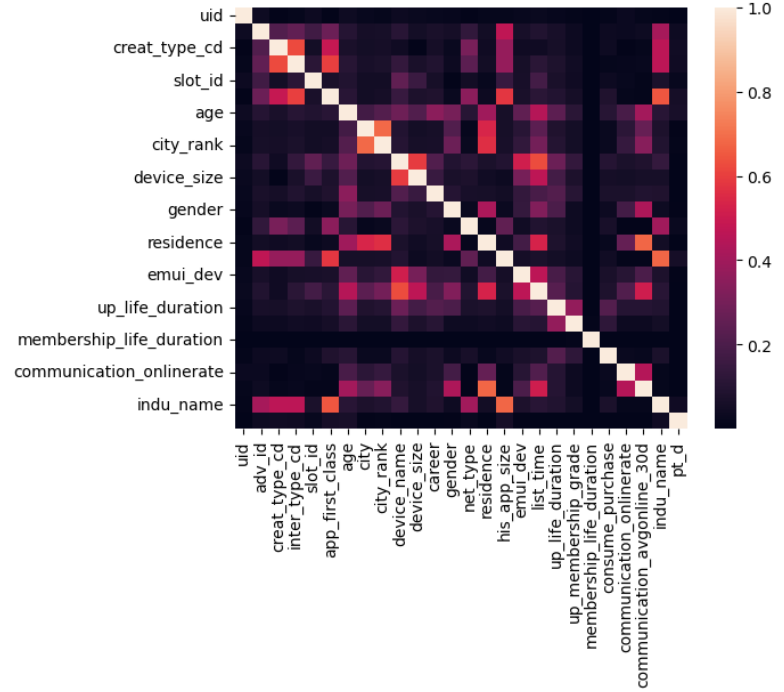
Implement Tukey's fences to identify outliers based on the Inter Quartile Range (IQR) method:

	Attribute	Outliers Prob Count	Outliers Prob Fraction	Outliers Poss Count	Outliers Poss Fraction
10	device_size	363035	0.435401	363890	0.436426
16	emui_dev	294556	0.353271	360135	0.431923
2	creat_type_cd	260334	0.312228	260334	0.312228
5	app_first_class	183892	0.220548	183892	0.220548
3	inter_type_cd	183765	0.220396	183765	0.220396
13	net_type	169724	0.203556	170334	0.204288
6	age	150420	0.180404	150420	0.180404
21	consume_purchase	103457	0.124080	103457	0.124080
19	up_membership_grade	101469	0.121695	101469	0.121695
12	gender	83805	0.100510	120719	0.144783
23	communication_avgonline_30d	39116	0.046913	224097	0.268768
22	communication_onlinerate	22330	0.026781	69947	0.083890
17	list_time	14981	0.017967	18503	0.022191
7	city	8348	0.010012	29268	0.035102
8	city_rank	7503	0.008999	7503	0.008999
0	uid	5863	0.007032	21881	0.026243
1	adv_id	4403	0.005281	32313	0.038754
14	residence	4045	0.004851	5444	0.006529
9	device_name	1832	0.002197	20227	0.024259
20	membership_life_duration	1216	0.001458	1216	0.001458
15	his_app_size	0	0.000000	0	0.000000
18	up_life_duration	0	0.000000	4534	0.005438
11	career	0	0.000000	0	0.000000
4	slot_id	0	0.000000	179788	0.215626
24	indu_name	0	0.000000	13	0.000016
25	pt_d	0	0.000000	0	0.000000

- The difference in outlier fraction between pre-transformed and post-transformed data is very small.
- Because all attributes are categorical, the existence of numerous ‘outliers’ is reasonable. No modification on these data points is added.

EDA on transformed train set: Correlation and VIF

Heatmap visualizing correlation between numerical attributes:



Matrix visualizing correlation (>0.5) between numerical attributes:

	correlation
city with city_rank	0.684963
residence with communication_avgonline_30d	0.677722
his_app_size with indu_name	0.676741
app_first_class with indu_name	0.645145
device_name with list_time	0.628226
creat_type_cd with inter_type_cd	0.621862
inter_type_cd with app_first_class	0.601278
device_name with device_size	0.591819
app_first_class with his_app_size	0.578999
city_rank with residence	0.559885
city with residence	0.535229
residence with list_time	0.527890
device_name with emui_dev	0.514740
list_time with communication_avgonline_30d	0.511252

	attribute	vif
18	list_time	3.206713
15	residence	3.150175
10	device_name	2.758776
24	communication_avgonline_30d	2.532686
25	indu_name	2.527964
6	app_first_class	2.348937
16	his_app_size	2.199281
9	city_rank	2.110852
8	city	2.089742
4	inter_type_cd	2.077706
3	creat_type_cd	1.803749
11	device_size	1.607864
7	age	1.559497
17	emui_dev	1.451711
19	up_life_duration	1.356496
2	adv_id	1.353837
13	gender	1.341934
14	net_type	1.266182
23	communication_onlinerate	1.263399
12	career	1.190188
20	up_membership_grade	1.169617
5	slot_id	1.133687
22	consume_purchase	1.065196
26	pt_d	1.011781
1	uid	1.004690
21	membership_life_duration	1.000214
0	const	1.000000

- Some attribute pairs have correlations greater than 0.5, but **no attribute** pair has a correlation **greater than 0.7** after the transformation.
- The VIFs of all attributes are **below 5**, indicating that attributes are moderately correlated.

Discussions

- Checking if the original data set is pre-encoded and if values have statistical meaning are important. We need to use the description of attributes and the distribution of dataset to identify whether attributes should be considered as categorical or numerical.

- Using the commonly used strategies to detect and handle outliers may be ineffective for pre-encoded categorical data.