Data Preparation

April 8, 2024

1 Data Preparation

[ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import gc
pd.options.mode.chained_assignment = None

root = './data/'
```

Reading all data

[ ]:
```
orders = pd.read_csv(root + 'orders.csv',
                     dtype={
                            'order_id': np.int32,
                            'user_id': np.int64,
                            'eval_set': 'category',
                            'order_number': np.int16,
                            'order_dow': np.int8,
                            'order_hour_of_day': np.int8,
                            'days_since_prior_order': np.float32})


order_products_train = pd.read_csv(root + 'order_products__train.csv',
                                   dtype={
                                          'order_id': np.int32,
                                          'product_id': np.uint16,
                                          'add_to_cart_order': np.int16,
                                          'reordered': np.int8})

order_products_prior = pd.read_csv(root + 'order_products__prior.csv',
                                   dtype={
                                          'order_id': np.int32,
                                          'product_id': np.uint16,
                                          'add_to_cart_order': np.int16,
                                          'reordered': np.int8})
```

```
product_features = pd.read_pickle(root + 'product_features.pkl')

user_features = pd.read_pickle(root + 'user_features.pkl')

user_product_features = pd.read_pickle(root + 'user_product_features.pkl')

products = pd.read_csv(root +'products.csv')

aisles = pd.read_csv(root + 'aisles.csv')

departments = pd.read_csv(root + 'departments.csv')
```

merging train order data with orders

[ ]:
```
train_orders = orders.merge(order_products_train, on = 'order_id', how =␣
↪'inner')
train_orders.head()
```

[ ]:

	order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	\
0	1187899	1	train	11	4	8	
1	1187899	1	train	11	4	8	
2	1187899	1	train	11	4	8	
3	1187899	1	train	11	4	8	
4	1187899	1	train	11	4	8	

	days_since_prior_order	product_id	add_to_cart_order	reordered
0	14.0	196	1	1
1	14.0	25133	2	1
2	14.0	38928	3	1
3	14.0	26405	4	1
4	14.0	39657	5	1

removing unnecessary columns from train_orders

[ ]:
```
train_orders.drop(['eval_set', 'add_to_cart_order', 'order_id'], axis = 1,␣
↪inplace = True)
```

unique user_ids in train data

[ ]:
```
train_users = train_orders.user_id.unique()
train_users[:10]
```

[ ]:
```
array([ 1,  2,  5,  7,  8,  9, 10, 13, 14, 17])
```

keeping only train_users in the data

[ ]:
```
user_product_features.shape
```

[ ]:
```
(13307953, 11)
```

[ ]:
```
user_product_features.head()
```

[ ]:
```
   user_id  product_id  total_product_orders_by_user  \
0        1         196                            10
1        1       10258                             9
2        1       10326                             1
3        1       12427                            10
4        1       13032                             3

   total_product_reorders_by_user  user_product_reorder_percentage  \
0                               9                         0.900000
1                               8                         0.888889
2                               0                         0.000000
3                               9                         0.900000
4                               2                         0.666667

   avg_add_to_cart_by_user  avg_days_since_last_bought  last_ordered_in  \
0                 1.400000                   17.600000               10
1                 3.333333                   19.555555               10
2                 5.000000                   28.000000                5
3                 3.300000                   17.600000               10
4                 6.333333                   21.666666               10

   is_reorder_3  is_reorder_2  is_reorder_1
0           1.0           1.0           1.0
1           1.0           1.0           1.0
2           0.0           0.0           0.0
3           1.0           1.0           1.0
4           1.0           0.0           0.0
```

[ ]:
```
df = user_product_features[user_product_features.user_id.isin(train_users)]
df.head()
```

[ ]:
```
   user_id  product_id  total_product_orders_by_user  \
0        1         196                            10
1        1       10258                             9
2        1       10326                             1
3        1       12427                            10
4        1       13032                             3

   total_product_reorders_by_user  user_product_reorder_percentage  \
0                               9                         0.900000
1                               8                         0.888889
2                               0                         0.000000
3                               9                         0.900000
4                               2                         0.666667
```

```
   avg_add_to_cart_by_user  avg_days_since_last_bought  last_ordered_in  \
0                 1.400000                   17.600000               10
1                 3.333333                   19.555555               10
2                 5.000000                   28.000000                5
3                 3.300000                   17.600000               10
4                 6.333333                   21.666666               10

   is_reorder_3  is_reorder_2  is_reorder_1
0           1.0           1.0           1.0
1           1.0           1.0           1.0
2           0.0           0.0           0.0
3           1.0           1.0           1.0
4           1.0           0.0           0.0
```

[ ]:
```
df = df.merge(train_orders, on = ['user_id', 'product_id'], how = 'outer')
df.head()
```

[ ]:
```
   user_id  product_id  total_product_orders_by_user  \
0        1         196                          10.0
1        1       10258                           9.0
2        1       10326                           1.0
3        1       12427                          10.0
4        1       13032                           3.0

   total_product_reorders_by_user  user_product_reorder_percentage  \
0                             9.0                         0.900000
1                             8.0                         0.888889
2                             0.0                         0.000000
3                             9.0                         0.900000
4                             2.0                         0.666667

   avg_add_to_cart_by_user  avg_days_since_last_bought  last_ordered_in  \
0                 1.400000                   17.600000             10.0
1                 3.333333                   19.555555             10.0
2                 5.000000                   28.000000              5.0
3                 3.300000                   17.600000             10.0
4                 6.333333                   21.666666             10.0

   is_reorder_3  is_reorder_2  is_reorder_1  order_number  order_dow  \
0           1.0           1.0           1.0          11.0        4.0
1           1.0           1.0           1.0          11.0        4.0
2           0.0           0.0           0.0           NaN        NaN
3           1.0           1.0           1.0           NaN        NaN
4           1.0           0.0           0.0          11.0        4.0

   order_hour_of_day  days_since_prior_order  reordered
0                8.0                    14.0        1.0
```

```
1                8.0                     14.0        1.0
2                NaN                      NaN        NaN
3                NaN                      NaN        NaN
4                8.0                     14.0        1.0
```

for order_number, order_dow, order_hour_of_day, days_since_prior_order, impute null values with mean values grouped by users as these products will also be potential candidate for order.

[ ]:
```
df.order_number.fillna(df.groupby('user_id')['order_number'].transform('mean'),␣
 ↪inplace = True)
df.order_dow.fillna(df.groupby('user_id')['order_dow'].transform('mean'),␣
 ↪inplace = True)
df.order_hour_of_day.fillna(df.groupby('user_id')['order_hour_of_day'].
 ↪transform('mean'), inplace = True)
df.days_since_prior_order.fillna(df.
 ↪groupby('user_id')['days_since_prior_order'].\
                                                          transform('mean'),␣
 ↪inplace = True)
```

Removing those products which were bought the first time in last order by a user

[ ]:
```
df.reordered.value_counts()
```

[ ]:
```
reordered
1.0    828824
0.0    555793
Name: count, dtype: int64
```

[ ]:
```
df.reordered.isnull().sum()
```

[ ]:
```
7645837
```

[ ]:
```
df = df[df.reordered != 0]
```

[ ]:
```
df.shape
```

[ ]:
```
(8474661, 16)
```

Now imputing 0 in reordered as they were not reordered by user in his/her last order.

[ ]:
```
df.reordered.fillna(0, inplace = True)

df.isnull().sum()
```

[ ]:
```
user_id                          0
product_id                       0
total_product_orders_by_user     0
total_product_reorders_by_user   0
user_product_reorder_percentage  0
```

```
avg_add_to_cart_by_user            0
avg_days_since_last_bought         0
last_ordered_in                    0
is_reorder_3                       0
is_reorder_2                       0
is_reorder_1                       0
order_number                       0
order_dow                          0
order_hour_of_day                  0
days_since_prior_order             0
reordered                          0
dtype: int64
```

[ ]:
```
df.head()
```

[ ]:
```
   user_id  product_id  total_product_orders_by_user  \
0        1         196                          10.0
1        1       10258                           9.0
2        1       10326                           1.0
3        1       12427                          10.0
4        1       13032                           3.0

   total_product_reorders_by_user  user_product_reorder_percentage  \
0                             9.0                         0.900000
1                             8.0                         0.888889
2                             0.0                         0.000000
3                             9.0                         0.900000
4                             2.0                         0.666667

   avg_add_to_cart_by_user  avg_days_since_last_bought  last_ordered_in  \
0                 1.400000                   17.600000             10.0
1                 3.333333                   19.555555             10.0
2                 5.000000                   28.000000              5.0
3                 3.300000                   17.600000             10.0
4                 6.333333                   21.666666             10.0

   is_reorder_3  is_reorder_2  is_reorder_1  order_number  order_dow  \
0           1.0           1.0           1.0          11.0        4.0
1           1.0           1.0           1.0          11.0        4.0
2           0.0           0.0           0.0          11.0        4.0
3           1.0           1.0           1.0          11.0        4.0
4           1.0           0.0           0.0          11.0        4.0

   order_hour_of_day  days_since_prior_order  reordered
0                8.0                    14.0        1.0
1                8.0                    14.0        1.0
2                8.0                    14.0        0.0
```

```
3                8.0                    14.0        0.0
4                8.0                    14.0        1.0
```

Merging product and user features

[ ]:
```
product_features.head()
```

[ ]:
```
   product_id  mean_add_to_cart_order  total_orders  total_reorders  \
0           1                5.801836          1852            1136
1           2                9.888889            90              12
2           3                6.415162           277             203
3           4                9.507599           329             147
4           5                6.466667            15               9

   reorder_percentage  unique_users  order_first_time_total_cnt  \
0            0.613391           716                         716
1            0.133333            78                          78
2            0.732852            74                          74
3            0.446809           182                         182
4            0.600000             6                           6

   order_second_time_total_cnt  is_organic  second_time_percent  …  \
0                          276           0             0.385475  …
1                            8           0             0.102564  …
2                           36           0             0.486486  …
3                           64           0             0.351648  …
4                            4           0             0.666667  …

   department_std_add_to_cart_order  department_total_orders  \
0                          7.692492                 2887550
1                          7.875241                 1875577
2                          6.711172                 2690129
3                          7.393502                 2236432
4                          7.875241                 1875577

   department_total_reorders  department_reorder_percentage  \
0                    1657973                       0.574180
1                     650301                       0.346721
2                    1757892                       0.653460
3                    1211890                       0.541885
4                     650301                       0.346721

   department_unique_users  department_0  department_1  department_2  \
0                   174219             0             0             0
1                   172755             0             0             0
2                   172795             0             0             0
3                   163233             0             0             1
4                   172755             0             0             0
```

```
   department_3  department_4
0             0             1
1             1             0
2             1             1
3             0             0
4             1             0

[5 rows x 35 columns]
```

[ ]:
```
user_features.head()
```

[ ]:
```
   user_id   avg_dow   std_dow    avg_doh   std_doh  avg_since_order  \
0        1  2.644068  1.256194  10.542373  3.500355        18.542374
1        2  2.005128  0.971222  10.441026  1.649854        14.902564
2        3  1.011364  1.245630  16.352273  1.454599        10.181818
3        4  4.722222  0.826442  13.111111  1.745208        11.944445
4        5  1.621622  1.276961  15.729730  2.588958        10.189189

   std_since_order  total_orders_by_user  total_products_by_user  \
0        10.559065                    10                      59
1         9.671712                    14                     195
2         5.867396                    12                      88
3         9.973330                     5                      18
4         7.600577                     4                      37

   total_unique_product_by_user  total_reorders_by_user  \
0                            18                      41
1                           102                      93
2                            33                      55
3                            17                       1
4                            23                      14

   reorder_propotion_by_user  average_order_size  reorder_in_order  orders_3  \
0                   0.694915            5.900000          0.705833         6
1                   0.476923           13.928571          0.447961        19
2                   0.625000            7.333333          0.658817         6
3                   0.055556            3.600000          0.028571         7
4                   0.378378            9.250000          0.377778         9

   orders_2  orders_1  reorder_3  reorder_2  reorder_1
0         6         9   0.666667        1.0   0.666667
1         9        16   0.578947        0.0   0.625000
2         5         6   0.833333        1.0   1.000000
3         2         3   0.142857        0.0   0.000000
4         5        12   0.444444        0.4   0.666667
```

[ ]:
```
df = df.merge(product_features, on = 'product_id', how = 'left')
df = df.merge(user_features, on = 'user_id', how = 'left')
df.head()
```

[ ]:

	user_id	product_id	total_product_orders_by_user	\
0	1	196	10.0	
1	1	10258	9.0	
2	1	10326	1.0	
3	1	12427	10.0	
4	1	13032	3.0	

	total_product_reorders_by_user	user_product_reorder_percentage	\
0	9.0	0.900000	
1	8.0	0.888889	
2	0.0	0.000000	
3	9.0	0.900000	
4	2.0	0.666667	

	avg_add_to_cart_by_user	avg_days_since_last_bought	last_ordered_in	\
0	1.400000	17.600000	10.0	
1	3.333333	19.555555	10.0	
2	5.000000	28.000000	5.0	
3	3.300000	17.600000	10.0	
4	6.333333	21.666666	10.0	

	is_reorder_3	is_reorder_2	…	total_reorders_by_user	\
0	1.0	1.0	…	41	
1	1.0	1.0	…	41	
2	0.0	0.0	…	41	
3	1.0	1.0	…	41	
4	1.0	0.0	…	41	

	reorder_propotion_by_user	average_order_size	reorder_in_order	orders_3	\
0	0.694915	5.9	0.705833	6	
1	0.694915	5.9	0.705833	6	
2	0.694915	5.9	0.705833	6	
3	0.694915	5.9	0.705833	6	
4	0.694915	5.9	0.705833	6	

	orders_2	orders_1	reorder_3	reorder_2	reorder_1
0	6	9	0.666667	1.0	0.666667
1	6	9	0.666667	1.0	0.666667
2	6	9	0.666667	1.0	0.666667
3	6	9	0.666667	1.0	0.666667
4	6	9	0.666667	1.0	0.666667

[5 rows x 69 columns]

The dataframe has null values because the product was never bought earlier by a user

[ ]:
```
df.shape
```

[ ]:
```
(8474661, 69)
```

[ ]:
```
df.isnull().sum().sort_values(ascending = False)
```

[ ]:
```
user_id                     0
department_unique_users     0
avg_dow                     0
department_4                0
department_3                0
                           ..
aisle_reorder_percentage    0
aisle_unique_users          0
aisle_0                     0
aisle_1                     0
reorder_1                   0
Length: 69, dtype: int64
```

[ ]:
```
df.to_pickle(root + 'Finaldata.pkl')
```

[ ]:
```
df2 = pd.read_pickle(root +'Finaldata.pkl')
df2.head()
```

[ ]:
```
   user_id  product_id  total_product_orders_by_user  \
0        1         196                          10.0
1        1       10258                           9.0
2        1       10326                           1.0
3        1       12427                          10.0
4        1       13032                           3.0

   total_product_reorders_by_user  user_product_reorder_percentage  \
0                             9.0                         0.900000
1                             8.0                         0.888889
2                             0.0                         0.000000
3                             9.0                         0.900000
4                             2.0                         0.666667

   avg_add_to_cart_by_user  avg_days_since_last_bought  last_ordered_in  \
0                 1.400000                   17.600000             10.0
1                 3.333333                   19.555555             10.0
2                 5.000000                   28.000000              5.0
3                 3.300000                   17.600000             10.0
4                 6.333333                   21.666666             10.0

   is_reorder_3  is_reorder_2  …  total_reorders_by_user  \
```

```
0           1.0           1.0  …                    41
1           1.0           1.0  …                    41
2           0.0           0.0  …                    41
3           1.0           1.0  …                    41
4           1.0           0.0  …                    41

   reorder_propotion_by_user  average_order_size  reorder_in_order  orders_3  \
0                   0.694915                 5.9          0.705833         6
1                   0.694915                 5.9          0.705833         6
2                   0.694915                 5.9          0.705833         6
3                   0.694915                 5.9          0.705833         6
4                   0.694915                 5.9          0.705833         6

   orders_2  orders_1  reorder_3  reorder_2  reorder_1
0         6         9   0.666667        1.0   0.666667
1         6         9   0.666667        1.0   0.666667
2         6         9   0.666667        1.0   0.666667
3         6         9   0.666667        1.0   0.666667
4         6         9   0.666667        1.0   0.666667

[5 rows x 69 columns]
```

Yayyyyy. Ready for some cool modeling now :p