FeatureEngineering_Activity_Solution

April 8, 2022

[1]:
```
import pandas as pd
```

[20]:
```
disp_url = 'https://raw.githubusercontent.com/fenago/MLEssentials/main/datasets/
 ↪disp.csv'
account_url = 'https://raw.githubusercontent.com/fenago/MLEssentials/main/
 ↪datasets/account.csv'
client_url = 'https://raw.githubusercontent.com/fenago/MLEssentials/main/
 ↪datasets/client.csv'
trans_url = './datasets/trans.csv'
# data can be found here: https://github.com/fenago/MLEssentials/tree/main/
 ↪datasets
```

[3]:
```
df_disp = pd.read_csv(disp_url, sep=';')
df_trans = pd.read_csv(trans_url, sep=';')
df_account = pd.read_csv(account_url, sep=';')
df_client = pd.read_csv(client_url, sep=';')
```

```
/opt/conda/lib/python3.8/site-packages/IPython/core/interactiveshell.py:3146:
DtypeWarning: Columns (8) have mixed types.Specify dtype option on import or set
low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

[4]:
```
df_trans.head()
```

[4]:

	trans_id	account_id	date	type	operation	amount	balance	k_symbol	\
0	695247	2378	930101	PRIJEM	VKLAD	700.0	700.0	NaN	
1	171812	576	930101	PRIJEM	VKLAD	900.0	900.0	NaN	
2	207264	704	930101	PRIJEM	VKLAD	1000.0	1000.0	NaN	
3	1117247	3818	930101	PRIJEM	VKLAD	600.0	600.0	NaN	
4	579373	1972	930102	PRIJEM	VKLAD	400.0	400.0	NaN	

	bank	account
0	NaN	NaN
1	NaN	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN

[5]:
```
df_trans.shape
```

[5]:
```
(1056320, 10)
```

[6]:
```
df_account.head()
```

[6]:

	account_id	district_id	frequency	date
0	576	55	POPLATEK MESICNE	930101
1	3818	74	POPLATEK MESICNE	930101
2	704	55	POPLATEK MESICNE	930101
3	2378	16	POPLATEK MESICNE	930101
4	2632	24	POPLATEK MESICNE	930102

[7]:
```
df_trans_acc = pd.merge(df_trans, df_account, how='left', on='account_id')
```

[8]:
```
df_trans_acc.shape
```

[8]:
```
(1056320, 13)
```

[9]:
```
df_disp.head()
```

[9]:

	disp_id	client_id	account_id	type
0	1	1	1	OWNER
1	2	2	2	OWNER
2	3	3	2	DISPONENT
3	4	4	3	OWNER
4	5	5	3	DISPONENT

[10]:
```
df_disp_owner = df_disp[df_disp['type'] == 'OWNER']
```

[11]:
```
df_disp_owner.duplicated(subset='account_id').sum()
```

[11]:
```
0
```

[12]:
```
df_trans_acc_disp = pd.merge(df_trans_acc, df_disp_owner, how='left',␣
↪on='account_id')
df_trans_acc_disp.shape
```

[12]:
```
(1056320, 16)
```

[13]:
```
df_client.head()
```

[13]:

	client_id	birth_number	district_id
0	1	706213	18
1	2	450204	1
2	3	406009	1
3	4	561201	5

```
4          5        605703            5
```

[14]:
```
df_merged = pd.merge(df_trans_acc_disp, df_client, how='left', on=['client_id',␣
↪'district_id'])
df_merged.shape
```

[14]:
```
(1056320, 17)
```

[15]:
```
df_merged.columns
```

[15]:
```
Index(['trans_id', 'account_id', 'date_x', 'type_x', 'operation', 'amount',
       'balance', 'k_symbol', 'bank', 'account', 'district_id', 'frequency',
       'date_y', 'disp_id', 'client_id', 'type_y', 'birth_number'],
      dtype='object')
```

[16]:
```
df_merged.rename(columns={'date_x': 'trans_date', 'type_x': 'trans_type',␣
↪'date_y':'account_creation', 'type_y':'client_type'}, inplace=True)
```

[17]:
```
df_merged.head()
```

[17]:

	trans_id	account_id	trans_date	trans_type	operation	amount	balance	\
0	695247	2378	930101	PRIJEM	VKLAD	700.0	700.0	
1	171812	576	930101	PRIJEM	VKLAD	900.0	900.0	
2	207264	704	930101	PRIJEM	VKLAD	1000.0	1000.0	
3	1117247	3818	930101	PRIJEM	VKLAD	600.0	600.0	
4	579373	1972	930102	PRIJEM	VKLAD	400.0	400.0	

	k_symbol	bank	account	district_id	frequency	account_creation	\
0	NaN	NaN	NaN	16	POPLATEK MESICNE	930101	
1	NaN	NaN	NaN	55	POPLATEK MESICNE	930101	
2	NaN	NaN	NaN	55	POPLATEK MESICNE	930101	
3	NaN	NaN	NaN	74	POPLATEK MESICNE	930101	
4	NaN	NaN	NaN	77	POPLATEK MESICNE	930102	

	disp_id	client_id	client_type	birth_number
0	2873	2873	OWNER	755324.0
1	692	692	OWNER	NaN
2	844	844	OWNER	NaN
3	4601	4601	OWNER	NaN
4	2397	2397	OWNER	NaN

[18]:
```
df_merged.dtypes
```

[18]:
```
trans_id             int64
account_id           int64
trans_date           int64
trans_type          object
```

```
operation            object
amount              float64
balance             float64
k_symbol             object
bank                 object
account             float64
district_id           int64
frequency            object
account_creation      int64
disp_id               int64
client_id             int64
client_type          object
birth_number        float64
dtype: object
```

[19]:
```
df_merged['trans_date'] = pd.to_datetime(df_merged['trans_date'],␣
↪format="%y%m%d")
df_merged['account_creation'] = pd.to_datetime(df_merged['account_creation'],␣
↪format="%y%m%d")
```

[ ]:
```
df_merged.dtypes
```

[ ]:
```
df_merged['is_female'] = (df_merged['birth_number'] % 10000) / 5000 > 1
```

[ ]:
```
df_merged['birth_number'].head()
```

[ ]:
```
df_merged.loc[df_merged['is_female'] == True, 'birth_number'] -= 5000
```

[ ]:
```
df_merged['birth_number'].head()
```

[ ]:
```
pd.to_datetime(df_merged['birth_number'], format="%y%m%d", errors='coerce')
```

[ ]:
```
df_merged['birth_number'] = df_merged['birth_number'].astype(str)
df_merged['birth_number'].head()
```

[ ]:
```
import numpy as np
df_merged.loc[df_merged['birth_number'] == 'nan', 'birth_number'] = np.nan
df_merged['birth_number'].head()
```

[ ]:
```
df_merged.loc[~df_merged['birth_number'].isna(), 'birth_number'] = '19' +␣
↪df_merged.loc[~df_merged['birth_number'].isna(), 'birth_number']
df_merged['birth_number'].head()
```

[ ]:
```
df_merged['birth_number'] = pd.to_datetime(df_merged['birth_number'],␣
↪format="%Y%m%d", errors='coerce')
df_merged['birth_number'].head(20)
```

[ ]:
```
df_merged['age_at_creation'] = df_merged['account_creation'] -␣
↪df_merged['birth_number']
```

[ ]:
```
df_merged['age_at_creation'] = df_merged['age_at_creation'] / np.
↪timedelta64(1,'Y')
```

[ ]:
```
df_merged['age_at_creation'] = df_merged['age_at_creation'].round()
df_merged.head()
```