data_pipeline

September 28, 2021

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

[2]:
```
application_train = pd.read_csv('./DATA/application_train.csv')
application_train.head()
```

[2]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	\
0	100002	1	Cash loans	M	N	
1	100003	0	Cash loans	F	N	
2	100004	0	Revolving loans	M	Y	
3	100006	0	Cash loans	F	N	
4	100007	0	Cash loans	M	N	

	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	\
0	Y	0	202500.0	406597.5	24700.5	
1	N	0	270000.0	1293502.5	35698.5	
2	Y	0	67500.0	135000.0	6750.0	
3	Y	0	135000.0	312682.5	29686.5	
4	Y	0	121500.0	513000.0	21865.5	

	…	FLAG_DOCUMENT_18	FLAG_DOCUMENT_19	FLAG_DOCUMENT_20	FLAG_DOCUMENT_21	\
0	…	0	0	0	0	
1	…	0	0	0	0	
2	…	0	0	0	0	
3	…	0	0	0	0	
4	…	0	0	0	0	

	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	\
0	0.0	0.0	
1	0.0	0.0	
2	0.0	0.0	
3	NaN	NaN	
4	0.0	0.0	

	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	\
0	0.0	0.0	
1	0.0	0.0	
2	0.0	0.0	

```
3                          NaN                          NaN
4                          0.0                          0.0

   AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
0                        0.0                         1.0
1                        0.0                         0.0
2                        0.0                         0.0
3                        NaN                         NaN
4                        0.0                         0.0

[5 rows x 122 columns]
```

[3]:
```
application_train['DEF_30_RATIO_SOCIAL_CIRCLE'] =␣
↪application_train['DEF_30_CNT_SOCIAL_CIRCLE'] /␣
↪application_train['OBS_30_CNT_SOCIAL_CIRCLE']
application_train['DEF_30_RATIO_SOCIAL_CIRCLE'].fillna(0, inplace=True)
application_train['DEF_60_RATIO_SOCIAL_CIRCLE'] =␣
↪application_train['DEF_60_CNT_SOCIAL_CIRCLE'] /␣
↪application_train['OBS_60_CNT_SOCIAL_CIRCLE']
application_train['DEF_60_RATIO_SOCIAL_CIRCLE'].fillna(0, inplace=True)
```

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

[4]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
1      100003       0         Cash loans           F            N
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

  FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
0               Y             0          202500.0    406597.5      24700.5
1               N             0          270000.0   1293502.5      35698.5
2               Y             0           67500.0    135000.0       6750.0
3               Y             0          135000.0    312682.5      29686.5
4               Y             0          121500.0    513000.0      21865.5

   …  FLAG_DOCUMENT_20 FLAG_DOCUMENT_21 AMT_REQ_CREDIT_BUREAU_HOUR  \
0  …                 0                0                        0.0
1  …                 0                0                        0.0
2  …                 0                0                        0.0
3  …                 0                0                        NaN
4  …                 0                0                        0.0

  AMT_REQ_CREDIT_BUREAU_DAY AMT_REQ_CREDIT_BUREAU_WEEK  \
0                       0.0                        0.0
1                       0.0                        0.0
```

```
2                       0.0                           0.0
3                       NaN                           NaN
4                       0.0                           0.0

  AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT  \
0                       0.0                        0.0
1                       0.0                        0.0
2                       0.0                        0.0
3                       NaN                        NaN
4                       0.0                        0.0

   AMT_REQ_CREDIT_BUREAU_YEAR  DEF_30_RATIO_SOCIAL_CIRCLE  \
0                         1.0                         1.0
1                         0.0                         0.0
2                         0.0                         0.0
3                         NaN                         0.0
4                         0.0                         0.0

   DEF_60_RATIO_SOCIAL_CIRCLE
0                         1.0
1                         0.0
2                         0.0
3                         0.0
4                         0.0

[5 rows x 124 columns]
```

[5]:
```
application_train['AMT_REQ_CREDIT_BUREAU'] =␣
↪application_train['AMT_REQ_CREDIT_BUREAU_HOUR'] +␣
↪application_train['AMT_REQ_CREDIT_BUREAU_DAY'] +␣
↪application_train['AMT_REQ_CREDIT_BUREAU_WEEK'] +␣
↪application_train['AMT_REQ_CREDIT_BUREAU_MON'] +␣
↪application_train['AMT_REQ_CREDIT_BUREAU_QRT'] +␣
↪application_train['AMT_REQ_CREDIT_BUREAU_YEAR']
```

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

[6]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
1      100003       0         Cash loans           F            N
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

  FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
0               Y             0          202500.0    406597.5      24700.5
1               N             0          270000.0   1293502.5      35698.5
```

```
2               Y             0          67500.0    135000.0       6750.0
3               Y             0         135000.0    312682.5      29686.5
4               Y             0         121500.0    513000.0      21865.5

   …  FLAG_DOCUMENT_21 AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY  \
0  …                 0                       0.0                       0.0
1  …                 0                       0.0                       0.0
2  …                 0                       0.0                       0.0
3  …                 0                       NaN                       NaN
4  …                 0                       0.0                       0.0

  AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON  \
0                        0.0                       0.0
1                        0.0                       0.0
2                        0.0                       0.0
3                        NaN                       NaN
4                        0.0                       0.0

  AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR  \
0                       0.0                        1.0
1                       0.0                        0.0
2                       0.0                        0.0
3                       NaN                        NaN
4                       0.0                        0.0

   DEF_30_RATIO_SOCIAL_CIRCLE  DEF_60_RATIO_SOCIAL_CIRCLE  \
0                         1.0                         1.0
1                         0.0                         0.0
2                         0.0                         0.0
3                         0.0                         0.0
4                         0.0                         0.0

   AMT_REQ_CREDIT_BUREAU
0                    1.0
1                    0.0
2                    0.0
3                    NaN
4                    0.0

[5 rows x 125 columns]
```

[7]:
```
credit_reqs_cols = [col for col in application_train.columns if␣
↪'AMT_REQ_CREDIT_BUREAU' in col]
for col in credit_reqs_cols:
    application_train[col].fillna(0, inplace=True)
```

[8]:
```
application_train.head()
```

[8]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR
0	100002	1	Cash loans	M	N
1	100003	0	Cash loans	F	N
2	100004	0	Revolving loans	M	Y
3	100006	0	Cash loans	F	N
4	100007	0	Cash loans	M	N

	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
0	Y	0	202500.0	406597.5	24700.5
1	N	0	270000.0	1293502.5	35698.5
2	Y	0	67500.0	135000.0	6750.0
3	Y	0	135000.0	312682.5	29686.5
4	Y	0	121500.0	513000.0	21865.5

	…	FLAG_DOCUMENT_21	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY
0	…	0	0.0	0.0
1	…	0	0.0	0.0
2	…	0	0.0	0.0
3	…	0	0.0	0.0
4	…	0	0.0	0.0

	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON
0	0.0	0.0
1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0

	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
0	0.0	1.0
1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0

	DEF_30_RATIO_SOCIAL_CIRCLE	DEF_60_RATIO_SOCIAL_CIRCLE
0	1.0	1.0
1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0

	AMT_REQ_CREDIT_BUREAU
0	1.0
1	0.0
2	0.0
3	0.0

```
4                     0.0

[5 rows x 125 columns]
```

0.0.1 Fill nulls

[9]:
```
application_train['NAME_TYPE_SUITE'].fillna('Unaccompanied', inplace=True)
```

[21]:
```
application_train['CNT_FAM_MEMBERS'].
 ↪fillna(application_train['CNT_FAM_MEMBERS'].mean(), inplace=True)
```

[10]:
```
application_train['OCCUPATION_TYPE'].fillna('Unknown', inplace=True)
```

[11]:
```
social_circle_cols = [col for col in application_train.columns if␣
 ↪'SOCIAL_CIRCLE' in col]
for col in social_circle_cols:
    application_train[col].fillna(0, inplace=True)
```

[12]:
```
application_train['DAYS_LAST_PHONE_CHANGE'].fillna(0, inplace=True)
```

[13]:
```
rem_cols = [col for col in application_train.columns if 'AVG' in str(col) or␣
 ↪'MEDI' in str(col) or 'MODE' in str(col)]
add_rem_cols = ['OWN_CAR_AGE']
rem_cols.extend(add_rem_cols)
len(rem_cols)
```

[13]:
```
48
```

0.0.2 Fill category values representing null with actual null value

[14]:
```
tf = application_train['CODE_GENDER'] == 'XNA'
application_train.loc[tf, 'CODE_GENDER'] = 'M'
```

[15]:
```
tf = application_train['NAME_FAMILY_STATUS'] == 'Unknown'
application_train.loc[tf, 'NAME_FAMILY_STATUS'] = 'Married'
```

0.0.3 Drop bad columns

[22]:
```
keep_cols = [col for col in application_train.columns if str(col) not in␣
 ↪rem_cols]
final_dataset = application_train[keep_cols].copy()
```

[23]:
```
final_dataset.head()
```

[23]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
1      100003       0         Cash loans           F            N
```

```
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

  FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
0               Y             0          202500.0    406597.5      24700.5
1               N             0          270000.0   1293502.5      35698.5
2               Y             0           67500.0    135000.0       6750.0
3               Y             0          135000.0    312682.5      29686.5
4               Y             0          121500.0    513000.0      21865.5

   …  FLAG_DOCUMENT_21 AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY  \
0  …                 0                       0.0                       0.0
1  …                 0                       0.0                       0.0
2  …                 0                       0.0                       0.0
3  …                 0                       0.0                       0.0
4  …                 0                       0.0                       0.0

  AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON  \
0                        0.0                       0.0
1                        0.0                       0.0
2                        0.0                       0.0
3                        0.0                       0.0
4                        0.0                       0.0

  AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR  \
0                       0.0                         1.0
1                       0.0                         0.0
2                       0.0                         0.0
3                       0.0                         0.0
4                       0.0                         0.0

   DEF_30_RATIO_SOCIAL_CIRCLE  DEF_60_RATIO_SOCIAL_CIRCLE  \
0                         1.0                         1.0
1                         0.0                         0.0
2                         0.0                         0.0
3                         0.0                         0.0
4                         0.0                         0.0

   AMT_REQ_CREDIT_BUREAU
0                    1.0
1                    0.0
2                    0.0
3                    0.0
4                    0.0

[5 rows x 77 columns]
```

0.0.4 Add default count from bureau balances

[18]:
```
applicant_default_count = pd.read_csv('./DATA/applicant_default_count.csv')
applicant_default_count.head()
```

[18]:

	SK_ID_CURR	DEFAULT_COUNT	MONTH_COUNT
0	100001	1.0	36.0
1	100002	27.0	68.0
2	100003	0.0	0.0
3	100004	0.0	0.0
4	100005	0.0	11.0

[31]:
```
rcols = ['SK_ID_CURR','DEFAULT_COUNT']
final_dataset2 = final_dataset.merge(applicant_default_count[rcols],␣
 ↪on='SK_ID_CURR', how='left')
final_dataset2['DEFAULT_COUNT'].fillna(0, inplace=True)
final_dataset2.head()
```

[31]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	\
0	100002	1	Cash loans	M	N	
1	100003	0	Cash loans	F	N	
2	100004	0	Revolving loans	M	Y	
3	100006	0	Cash loans	F	N	
4	100007	0	Cash loans	M	N	

	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	\
0	Y	0	202500.0	406597.5	24700.5	
1	N	0	270000.0	1293502.5	35698.5	
2	Y	0	67500.0	135000.0	6750.0	
3	Y	0	135000.0	312682.5	29686.5	
4	Y	0	121500.0	513000.0	21865.5	

	…	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	\
0	…	0.0	0.0	
1	…	0.0	0.0	
2	…	0.0	0.0	
3	…	0.0	0.0	
4	…	0.0	0.0	

	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	\
0	0.0	0.0	
1	0.0	0.0	
2	0.0	0.0	
3	0.0	0.0	
4	0.0	0.0	

	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR	\
0	0.0	1.0	

```
1                       0.0                           0.0
2                       0.0                           0.0
3                       0.0                           0.0
4                       0.0                           0.0

   DEF_30_RATIO_SOCIAL_CIRCLE  DEF_60_RATIO_SOCIAL_CIRCLE  \
0                         1.0                         1.0
1                         0.0                         0.0
2                         0.0                         0.0
3                         0.0                         0.0
4                         0.0                         0.0

   AMT_REQ_CREDIT_BUREAU  DEFAULT_COUNT
0                    1.0           27.0
1                    0.0            0.0
2                    0.0            0.0
3                    0.0            0.0
4                    0.0            0.0

[5 rows x 78 columns]
```

[55]:
```
miss_cols = final_dataset2.columns[final_dataset2.isna().any()].tolist()
miss_cols
```

[55]:
```
['AMT_ANNUITY',
 'AMT_GOODS_PRICE',
 'EXT_SOURCE_1',
 'EXT_SOURCE_2',
 'EXT_SOURCE_3']
```

0.0.5 KNN Imputation

[33]:
```
imp_cols = ['SK_ID_CURR'] +␣
 ↪['AMT_ANNUITY','AMT_GOODS_PRICE','EXT_SOURCE_1','EXT_SOURCE_2','EXT_SOURCE_3']␣
 ↪+ ['AMT_INCOME_TOTAL','AMT_CREDIT']
```

[34]:
```
impute_df = final_dataset2[imp_cols].copy()
impute_df.index = final_dataset2['SK_ID_CURR']
del impute_df['SK_ID_CURR']
impute_df.head()
```

[34]:
```
            AMT_ANNUITY  AMT_GOODS_PRICE  EXT_SOURCE_1  EXT_SOURCE_2  \
SK_ID_CURR
100002          24700.5         351000.0      0.083037      0.262949
100003          35698.5        1129500.0      0.311267      0.622246
100004           6750.0         135000.0           NaN      0.555912
100006          29686.5         297000.0           NaN      0.650442
```

```
100007          21865.5         513000.0            NaN      0.322738

            EXT_SOURCE_3  AMT_INCOME_TOTAL  AMT_CREDIT
SK_ID_CURR
100002          0.139376          202500.0    406597.5
100003               NaN          270000.0   1293502.5
100004          0.729567           67500.0    135000.0
100006               NaN          135000.0    312682.5
100007               NaN          121500.0    513000.0
```

[35]:
```
from sklearn.preprocessing import MinMaxScaler
```

[36]:
```
scaler = MinMaxScaler()
```

[40]:
```
scaler = MinMaxScaler()
impute_df_scaled = pd.DataFrame(scaler.
 ↪fit_transform(impute_df),columns=impute_df.columns)
impute_df_scaled.head()
```

[40]:
```
   AMT_ANNUITY  AMT_GOODS_PRICE  EXT_SOURCE_1  EXT_SOURCE_2  EXT_SOURCE_3  \
0     0.090032         0.077441      0.072215      0.307542      0.155054
1     0.132924         0.271605      0.312933      0.727773           NaN
2     0.020025         0.023569           NaN      0.650190      0.814130
3     0.109477         0.063973           NaN      0.760751           NaN
4     0.078975         0.117845           NaN      0.377472           NaN

   AMT_INCOME_TOTAL  AMT_CREDIT
0          0.001512    0.090287
1          0.002089    0.311736
2          0.000358    0.022472
3          0.000935    0.066837
4          0.000819    0.116854
```

[43]:
```
from sklearn.impute import KNNImputer
imputer = KNNImputer(n_neighbors=5)
final_impute_df_scaled = pd.DataFrame(imputer.
 ↪fit_transform(impute_df_scaled),columns = impute_df.columns)
final_impute_df_scaled.head()
```

[43]:
```
   AMT_ANNUITY  AMT_GOODS_PRICE  EXT_SOURCE_1  EXT_SOURCE_2  EXT_SOURCE_3  \
0     0.090032         0.077441      0.072215      0.307542      0.155054
1     0.132924         0.271605      0.312933      0.727773      0.587994
2     0.020025         0.023569      0.298370      0.650190      0.814130
3     0.109477         0.063973      0.619176      0.760751      0.503014
4     0.078975         0.117845      0.661169      0.377472      0.527440

   AMT_INCOME_TOTAL  AMT_CREDIT
```

```
0          0.001512    0.090287
1          0.002089    0.311736
2          0.000358    0.022472
3          0.000935    0.066837
4          0.000819    0.116854
```

[44]:
```
final_impute_df_scaled = final_impute_df.copy()
```

[45]:
```
final_impute_df = pd.DataFrame(scaler.
 ↪inverse_transform(final_impute_df_scaled),columns=impute_df.columns)
final_impute_df.index = impute_df.index
final_impute_df.head()
```

[45]:

SK_ID_CURR	AMT_ANNUITY	AMT_GOODS_PRICE	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	AMT_INCOME_TOTAL	AMT_CREDIT
100002	24700.5	351000.0	0.083037	0.262949	0.139376	202500.0	406597.5
100003	35698.5	1129500.0	0.311267	0.622246	0.527065	270000.0	1293502.5
100004	6750.0	135000.0	0.297461	0.555912	0.729567	67500.0	135000.0
100006	29686.5	297000.0	0.601624	0.650442	0.450967	135000.0	312682.5
100007	21865.5	513000.0	0.641439	0.322738	0.472840	121500.0	513000.0

[46]:
```
final_impute_df.reset_index(inplace=True)
```

[56]:
```
final_dataset2.to_csv('./DATA/imputed_data.csv')
```

[53]:
```
(final_dataset2['SK_ID_CURR']==final_impute_df['SK_ID_CURR']).sum()
```

[53]:
```
307511
```

[54]:
```
len(final_dataset2)
```

[54]:
```
307511
```

[57]:
```
for col in miss_cols:
    final_dataset2[col] = final_impute_df[col]
```

[58]:
```
final_dataset2.columns[final_dataset2.isna().any()].tolist()
```

[58]:
```
[]
```

[59]:
```
# Write data
```

[60]:
```
final_dataset2.to_csv('./DATA/final_dataset.csv', index=False)
```

[19]:
```
import sweetviz as sv
```

[20]:
```
report = sv.analyze(final_dataset2, "TARGET", pairwise_analysis="off")
```

```
                                             |                                       ␣
↪           | [  0%]   00:00 ->…
```

[21]:
```
report.show_html('clean_data_report.html')
```

```
Report clean_data_report.html was generated! NOTEBOOK/COLAB USERS: the web
browser MAY not pop up, regardless, the report IS saved in your notebook/colab
files.
```

[ ]:
```

```