model

June 1, 2021

[2]:
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
%matplotlib inline
```

[4]:
```
fpath = 'data/SUS_project_training_data.csv'
file = open(fpath, 'r', encoding='latin-1')

nrecords = sum(1 for line in file) - 1
sample_size = 40000
skip = np.random.choice(np.arange(1, nrecords),
                        size=nrecords-sample_size,
                        replace=False)
skip = sorted(skip)

df_X = pd.read_csv(fpath, sep=';',
                   encoding='latin-1', skiprows=skip)

df_y = pd.read_csv('data/training_targets.txt', sep=';',
                   encoding='latin-1', header=None, skiprows=skip)
```

[5]:
```
df_X
```

[5]:

	company_id	category_id	category_name	\
0	5bacd51926a9cb3d33cf4c1f	5a6f110ca0899f5ca2f7d6e9	Dania Lunch DuÅ¼e	
1	5bacd51926a9cb3d33cf4c1f	591301913dd75608a9c2ef19	Å niadania	
2	5bacd51926a9cb3d33cf4c1f	5a0033206cdc0d08a6591bfb	Dania Lunch MaÅ e	
3	5bacd51926a9cb3d33cf4c1f	5a0033206cdc0d08a6591bfb	Dania Lunch MaÅ e	
4	5bacd51926a9cb3d33cf4c1f	59005cd6c5c79d3575eb450d	PrzekÄ	
ski				
…	…	…	…	
39995	NaN	none	NaN	
39996	5c6e8878531c6b7fe27e08c8	5a6f110ca0899f5ca2f7d6e9	Dania Lunch DuÅ¼e	
39997	5c6e8878531c6b7fe27e08c8	5cd1a4f40a544c2d0d156fea	Pan Pomidor - Zupy	
39998	NaN	none	NaN	
39999	5c924f69b6cb840dcac430fe	5cd1a4f40a544c2d0d156fea	Pan Pomidor - Zupy	

```
                                          product_name  partner_product  \
0                           Indyk z pieczarkami i ryÅ¼em                0
1      FitBreak - Chlebek gryczany z kurczakiem i hum…                0
2                                   Cocido wegaÅ skie                0
3                                   Cocido wegaÅ skie                0
4                     SUPERFOOD BAR - Morela, Chlorella                0
…                                                   …              …
39995                                             NaN                0
39996                          Dyniowe curry z indykiem                0
39997                         Tajska z curry i kolendrÄ
               0
39998                                             NaN                0
39999         MarokaÅ ska z quinoÄ
, batatem i kolendrÄ
               0

      address_city                diet  size cooking_time  cooking_mv  …  \
0         Warszawa     Dieta Samuraja  500g     2-3 min.         NaN  …
1         Warszawa                none  350g          NaN         NaN  …
2         Warszawa  Kuchnia Latynoska  350g     2-3 min.         1.0  …
3         Warszawa  Kuchnia Latynoska  350g     2-3 min.         1.0  …
4         Warszawa                none   35g          NaN         NaN  …
…            …                  …   …          …          …  …
39995          NaN                 NaN   NaN          NaN         NaN  …
39996      Skawina     Dieta Samuraja  500g     2-3 min.         1.0  …
39997      Skawina                none  400g     2-3 min.         NaN  …
39998          NaN                 NaN   NaN          NaN         NaN  …
39999     Katowice                none  400g     2-3 min.         NaN  …

       prods_avail_in_cat_5a6f110ca0899f5ca2f7d6e9  \
0                                                1
1                                                0
2                                                0
3                                                0
4                                                0
…                                              …
39995                                            0
39996                                            3
39997                                            0
39998                                            0
39999                                            0

      prods_avail_in_cat_5abe0aed049e180557e22330  \
0                                               0
1                                               0
2                                               0
3                                               0
```

```
4                                                    0
…                                                  …
39995                                                0
39996                                                0
39997                                                0
39998                                                0
39999                                                0

       prods_avail_in_cat_5cb9b8eedf68013fb09db8f0  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0
…                                                  …
39995                                                0
39996                                                0
39997                                                0
39998                                                0
39999                                                0

       prods_avail_in_cat_5cd1a4d32b10792bc08dab31  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0
…                                                  …
39995                                                0
39996                                                0
39997                                                0
39998                                                0
39999                                                0

       prods_avail_in_cat_5cd1a4f40a544c2d0d156fea  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0
…                                                  …
39995                                                0
39996                                                0
39997                                                2
39998                                                0
39999                                                1
```

```
       prods_avail_in_cat_5d1b55aa5379175d45e9360a  \
0                                                0
1                                                0
2                                                0
3                                                0
4                                                0
…                                              …
39995                                            0
39996                                            0
39997                                            0
39998                                            0
39999                                            0

       prods_avail_in_cat_5d7103b8c8c4a843bc5b5706  \
0                                                0
1                                                0
2                                                0
3                                                0
4                                                0
…                                              …
39995                                            0
39996                                            0
39997                                            0
39998                                            0
39999                                            0

       prods_avail_in_cat_5d9f1dcab962ef075bd26ecb  \
0                                                0
1                                                0
2                                                0
3                                                0
4                                                0
…                                              …
39995                                            0
39996                                            0
39997                                            0
39998                                            0
39999                                            0

       prods_avail_in_cat_5e54ca889dca612f7a92cab9  prods_avail_in_cat_none
0                                                0                       0
1                                                0                       0
2                                                0                       0
3                                                0                       0
4                                                0                       0
…                                              …                       …
39995                                            0                       2
```

```
39996                                                0                      0
39997                                                0                      0
39998                                                0                      2
39999                                                0                      0

[40000 rows x 82 columns]
```

[6]:
```
categorical = df_X.dtypes[df_X.dtypes == object].index
numerical = df_X.dtypes[df_X.dtypes != object].index
```

[7]:
```
df_X[categorical].head()
```

[7]:
```
                 company_id               category_id      category_name  \
0  5bacd51926a9cb3d33cf4c1f  5a6f110ca0899f5ca2f7d6e9  Dania Lunch DuÅ¼e
1  5bacd51926a9cb3d33cf4c1f  591301913dd75608a9c2ef19         Å niadania
2  5bacd51926a9cb3d33cf4c1f  5a0033206cdc0d08a6591bfb  Dania Lunch MaÅ e
3  5bacd51926a9cb3d33cf4c1f  5a0033206cdc0d08a6591bfb  Dania Lunch MaÅ e
4  5bacd51926a9cb3d33cf4c1f  59005cd6c5c79d3575eb450d         PrzekÄ
ski

                                    product_name address_city  \
0                        Indyk z pieczarkami i ryÅ¼em     Warszawa
1  FitBreak - Chlebek gryczany z kurczakiem i hum…     Warszawa
2                                  Cocido wegaÅ skie     Warszawa
3                                  Cocido wegaÅ skie     Warszawa
4                  SUPERFOOD BAR - Morela, Chlorella     Warszawa

                diet  size cooking_time storage_temp        weekday quarter
0     Dieta Samuraja  500g     2-3 min.      2-5 Â°C         sobota      Q1
1               none  350g          NaN      2-5 Â°C         sobota      Q4
2  Kuchnia Latynoska  350g     2-3 min.      2-5 Â°C  poniedziaÅ ek      Q2
3  Kuchnia Latynoska  350g     2-3 min.      2-5 Â°C       czwartek      Q4
4               none   35g          NaN          NaN  poniedziaÅ ek      Q4
```

[8]:
```
from sklearn.preprocessing import FunctionTransformer, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.decomposition import TruncatedSVD

ct = ColumnTransformer([
    ('num', SimpleImputer(), numerical),
    ('cat', OneHotEncoder(handle_unknown='ignore'), categorical),
])
```

[7]:
```
# from sklearn.metrics import r2_score
# from sklearn.model_selection import RandomizedSearchCV
```

```
# from sklearn.linear_model import SGDRegressor
# from sklearn.linear_model import ElasticNetCV

# hyper_dist = {
#     'model__loss': ['squared_loss', 'huber', 'epsilon_insensitive',
#                     'squared_epsilon_insensitive'],
#     'model__penalty': ['l1', 'l2', 'elasticnet'],
#     'model__alpha': np.geomspace(1e-4, 1e4, 8),
#     'model__learning_rate': ['constant', 'optimal', 'invscaling', 'adaptive']
# }

# prep = Pipeline([
#     ('prep', ct),
#     ('pca', TruncatedSVD(n_components=64)),
# ])

# model = Pipeline([
#     ('model', SGDRegressor())
# ])

# # cv = RandomizedSearchCV(model, hyper_dist, n_iter=50, verbose=10, cv=5,
# #                         scoring='r2', n_jobs=4)

# prep_X = prep.fit_transform(df_X)
# # cv.fit(prep_X, df_y.to_numpy().ravel())
```

[32]:
```
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import ElasticNetCV

X_p = ct.fit_transform(df_X)
y = df_y.to_numpy().ravel()

ratios = [0.1, 0.5, 0.7, 0.9, 0.95, 0.99, 1]
cv = ElasticNetCV(l1_ratio=ratios, n_jobs=5, selection='random', verbose=1)
cv.fit(X_p, y)
```

```
[Parallel(n_jobs=5)]: Using backend ThreadingBackend with 5 concurrent workers.
…
…
…
…
…
…
…
…
…
…
…
```

```
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…[Parallel(n_jobs=5)]
: Done  35 out of  35 | elapsed:  5.1min finished
```

[32]:
```
ElasticNetCV(l1_ratio=[0.1, 0.5, 0.7, 0.9, 0.95, 0.99, 1], n_jobs=5,
             selection='random', verbose=1)
```

[35]:
```
from sklearn.linear_model import ElasticNet
from sklearn.model_selection import cross_val_score

model = ElasticNet(alpha=cv.alpha_, l1_ratio=cv.l1_ratio_)
cross_val_score(model, X_p, y, cv=5, scoring='r2', n_jobs=5)
```

[35]:
```
array([0.23505015, 0.19704505, 0.22689631, 0.30656144, 0.28083896])
```

[37]:
```
df_test_X = pd.read_csv('data/SUS_project_test_data.csv', sep=';',␣
↪encoding='latin-1')

prep_test = ct.transform(df_test_X)
df_test_y = pd.DataFrame(cv.predict(prep_test))
df_test_y[df_test_y < 0] = 0
df_test_y.to_csv('data/y.txt', index=False, header=False)
```

[ ]:
```

```