```
import pandas as pd
import jinja2
import numpy as np
from IPython.display import display
from sklearn.model_selection import train_test_split
from sklearn.compose import TransformedTargetRegressor
from sklearn.metrics import median_absolute_error, r2_score
import matplotlib.pyplot as plt
pd.set_option('display.max_columns', None)


from google.colab import drive

drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
data = pd.read_csv("/content/drive/My Drive/Colab Notebooks/df_main_final_scaled.cs


data.drop(columns=['Unnamed: 0'], inplace=True)


#data = data[data['mean_w2v']> 50]


#data.drop(columns =['Unnamed: 0',  'id', 'abstract',   'authors', 'references_x',
n_citation = list(data['n_citation'])


data.drop(columns =['n_citation'], inplace=True)
```

셀 삭제를 실행취소하려면 ⌘/Ctrl+M Z 또는 수정 메뉴의 실행취소 옵션을 사용하세요.

```
                                                                  tle_len', 'n_referenc
                    'mean_ref_ab_len', 'sd_ref_ab_len',  'ref_verb_ratio',
                    'main_journal',  'main_conf',    'main_review',  'main_meta',
                    'count_ref_conf',    'count_ref_journal',    'count_ref_meta',
                    'mean_ref_n_authors','sd_ref_n_authors', 'mean_ref_n_ref',   'sd
                    'med_ref_impact_1y','skew_ref_impact_1y',    'med_ref_n_citation
                   'mean_LDA', 'sd_LDA', 'mean_w2v', 'sd_w2v'])


fitted['n_citation'] = n_citation


from sklearn import linear_model
clf = linear_model.PoissonRegressor()


#Negative bimodal
import pandas as pd
import numpy as np

from patsy import dmatrices
```

```
import statsmodels.api as sm
from statsmodels.formula.api import glm

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from math import sqrt


mask = np.random.rand(len(data)) < 0.5

df_train = fitted[mask]
df_test = fitted[~mask]


df_train.to_csv("/content/drive/My Drive/Colab Notebooks/df_train_final.csv")
df_test.to_csv("/content/drive/My Drive/Colab Notebooks/df_test_final.csv")


df_train.shape
```

```
(1246, 35)
```

```
df_test.shape
```

```
(1293, 35)
```

```
data.head()
```

셀 삭제를 실행취소하려면 ⌘/Ctrl+M Z 또는 수정 메뉴의 실행취소 옵션을 사용하세요.

						verb_ratio	readabili
0	9	13	2550	3	140	0.107143	
1	5	4	651	1	122	0.172131	
2	2	21	2624	3	163	0.128834	
3	9	7	7019	2	235	0.068085	
4	13	5	5614	2	167	0.119760	

```
expr = """ n_citation ~  ref_verb_ratio +   count_ref_conf  + count_ref_journal + c
          + sd_ref_ab_len   + sd_ref_n_authors +    sd_ref_n_ref    + sd_ref_title_
          + mean_ref_n_ref +    mean_ref_title_len +    skew_ref_impact_1y  + skew_
            med_ref_impact_1y   + med_ref_n_citation +  med_ref_year +  sd_LDA +
            +   title_len + n_references    + ab_len    + impact_1y + readability_s
            + main_review   + main_meta + n_authors"""
```

```
y_train, X_train = dmatrices(expr, df_train, return_type = 'dataframe')

y_test, X_test = dmatrices(expr, df_test, return_type = 'dataframe')


nb_training_results = sm.GLM(y_train, X_train, family = sm.families.NegativeBinomia

result_train = nb_training_results.summary()


print(result_train)
%%capture cap
f = open("train_negative.txt", "w")
print(cap, file=f)
f.close()
```

                 Generalized Linear Model Regression Results
==
Dep. Variable:	n_citation	No. Observations:	1246
Model:	GLM	Df Residuals:	1212
Model Family:	NegativeBinomial	Df Model:	33
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-5563.4
Date:	Mon, 05 Dec 2022	Deviance:	2073.7
Time:	13:41:46	Pearson chi2:	1.55e+03
No. Iterations:	14		
Covariance Type:	nonrobust		

==

	coef	std err	z	P>|z|	[0.025
Intercept	2.4313	0.725	3.352	0.001	1.010
ref_verb_ratio	-0.5482	0.253	-2.165	0.030	-1.044
count_ref_conf	0.3185	0.542	0.587	0.557	-0.745
count_ref_journal	0.6170	0.631	0.979	0.328	-0.619
				0.542	-1.58e-14
				0.039	0.033
				0.108	-0.154
sd_ref_n_authors	1.3565	0.953	1.423	0.155	-0.511
sd_ref_n_ref	-0.2814	0.671	-0.420	0.675	-1.596
sd_ref_title_len	0.0802	0.296	0.271	0.787	-0.501
mean_ref_ab_len	-0.0776	0.320	-0.243	0.808	-0.704
mean_ref_n_authors	-0.7721	0.804	-0.961	0.337	-2.347
mean_ref_n_ref	0.0911	0.640	0.142	0.887	-1.164
mean_ref_title_len	0.3317	0.291	1.139	0.255	-0.239
skew_ref_impact_1y	-0.0262	0.253	-0.103	0.918	-0.523
skew_ref_n_citation	-0.1860	0.314	-0.593	0.553	-0.801
skew_ref_year	0.1207	0.207	0.583	0.560	-0.285
med_ref_impact_1y	-0.2437	0.285	-0.856	0.392	-0.802
med_ref_n_citation	1.3311	0.668	1.991	0.046	0.021
med_ref_year	1.7811	0.413	4.310	0.000	0.971
sd_LDA	0.2539	0.257	0.989	0.323	-0.249
sd_w2v	-2.2931	0.755	-3.037	0.002	-3.773
mean_LDA	-0.0240	0.281	-0.086	0.932	-0.574
mean_w2v	-0.0527	0.271	-0.195	0.846	-0.583
title_len	-0.4613	0.267	-1.726	0.084	-0.985
n_references	3.0257	0.607	4.989	0.000	1.837
ab_len	1.0017	0.471	2.126	0.033	0.078
impact_1y	0.7202	0.204	3.522	0.000	0.319

```
readability_score      -0.8239      0.586     -1.405      0.160      -1.973
verb_ratio             -0.1104      0.215     -0.514      0.607      -0.531
main_journal           -0.1273      0.097     -1.309      0.191      -0.318
main_conf              -0.1903      0.069     -2.761      0.006      -0.325
main_review            -0.2033      0.540     -0.377      0.706      -1.261
main_meta              -0.8097      0.370     -2.187      0.029      -1.535
n_authors               1.3552      0.464      2.922      0.003       0.446
==============================================================================
UsageError: Line magic function `%%capture` not found.
```

```
nb_test_results = sm.GLM(y_test, X_test, family = sm.families.NegativeBinomial(alph

result_test = nb_test_results.summary()


print(result_test)
%%capture cap
f = open("test_negative.txt", "w")
print(cap, file=f)
f.close()
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:             n_citation   No. Observations:                 1293
Model:                            GLM   Df Residuals:                     1258
Model Family:        NegativeBinomial   Df Model:                           34
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -5966.4
Date:                Mon, 05 Dec 2022   Deviance:                       2432.7
Time:                        13:41:59   Pearson chi2:                 2.87e+03
No. Iterations:                    15
Covariance Type:            nonrobust
==============================================================================
                          coef    std err          z      P>|z|      [0.025
------------------------------------------------------------------------------
                                                        0.102      -0.219
                                                        0.075      -0.045
                                                        0.029       0.113
count_ref_journal       0.5309      0.728      0.729      0.466      -0.896
count_ref_meta          0.5253      1.015      0.517      0.605      -1.465
count_ref_review        0.6865      0.319      2.154      0.031       0.062
sd_ref_ab_len          -0.2824      0.411     -0.688      0.492      -1.087
sd_ref_n_authors        0.5055      0.886      0.571      0.568      -1.231
sd_ref_n_ref           -0.0731      0.580     -0.126      0.900      -1.210
sd_ref_title_len       -0.1958      0.302     -0.648      0.517      -0.788
mean_ref_ab_len        -0.2717      0.303     -0.896      0.370      -0.866
mean_ref_n_authors     -1.9129      0.829     -2.307      0.021      -3.538
mean_ref_n_ref         -0.2844      0.609     -0.467      0.640      -1.477
mean_ref_title_len     -0.2148      0.290     -0.740      0.459      -0.784
skew_ref_impact_1y      0.4252      0.260      1.634      0.102      -0.085
skew_ref_n_citation     0.0394      0.295      0.134      0.894      -0.539
skew_ref_year           0.0749      0.202      0.370      0.711      -0.322
med_ref_impact_1y       0.6118      0.303      2.018      0.044       0.018
med_ref_n_citation      2.1661      0.673      3.216      0.001       0.846
med_ref_year            2.0572      0.395      5.214      0.000       1.284
sd_LDA                 -0.7747      0.260     -2.977      0.003      -1.285
sd_w2v                  0.5979      0.543      1.102      0.271      -0.466
mean_LDA                0.3003      0.269      1.118      0.264      -0.226
```

```
mean_w2v               -0.1598      0.258     -0.620      0.535      -0.665
title_len              -0.3552      0.246     -1.443      0.149      -0.838
n_references            2.4950      0.574      4.348      0.000       1.370
ab_len                  2.4968      0.416      6.006      0.000       1.682
impact_1y               1.2470      0.194      6.413      0.000       0.866
readability_score       0.2382      0.469      0.508      0.611      -0.681
verb_ratio             -0.0106      0.217     -0.049      0.961      -0.436
main_journal           -0.0318      0.102     -0.312      0.755      -0.232
main_conf              -0.2363      0.068     -3.459      0.001      -0.370
main_review             0.8466      1.027      0.824      0.410      -1.167
main_meta               0.3937      0.512      0.769      0.442      -0.610
n_authors               1.0980      0.433      2.533      0.011       0.248
==============================================================================
UsageError: Line magic function `%%capture` not found.
```

```
results_text = result_train.as_text()

import csv
resultFile = open("/content/drive/My Drive/Colab Notebooks/results_table.csv",'w')
resultFile.write(results_text)
resultFile.close()


nb_testing_results = sm.GLM(y_test, X_test, family = sm.families.NegativeBinomial(a
result_test = nb_testing_results.summary()


results_text = result_test.as_text()

import csv
resultFile = open("/content/drive/My Drive/Colab Notebooks/results_table2.csv",'w')
resultFile.write(results_text)
resultFile.close()
```



```
ain)
nb_summary_frame_train = nb_prediction_train.summary_frame()

print(nb_summary_frame_train)
```

```
             mean    mean_se  mean_ci_lower  mean_ci_upper
3       31.922090   7.055672      20.699128      49.230084
13      23.740488   3.062772      18.436382      30.570573
15      47.071333   6.032598      36.615726      60.512534
16      25.335049   2.964381      20.143044      31.865328
17      28.388792   4.522837      20.774732      38.793448
...           ...        ...            ...            ...
2533    18.359081   4.002220      11.975448      28.145574
2534    33.061609   4.001740      26.079274      41.913359
2535    52.842805  10.226035      36.162882      77.216248
2536    28.917466   3.397190      22.970057      36.404778
2537    35.641800   3.910107      28.746039      44.191756

[1246 rows x 4 columns]
```

```
predicted_counts_train = nb_summary_frame_train['mean']
```

```
predicted_counts_train
```

```
3       31.922090
13      23.740488
15      47.071333
16      25.335049
17      28.388792
          ...
2533    18.359081
2534    33.061609
2535    52.842805
2536    28.917466
2537    35.641800
Name: mean, Length: 1246, dtype: float64
```

```
# Accuracy of the train set
```

```
print("R-square of train set: ", round(r2_score(y_train, predicted_counts_train)*10
```

```
R-square of train set:  10.22 %
```

```
nb_prediction_test = nb_training_results.get_prediction(X_test)
```

```
nb_summary_frame_test = nb_prediction_test.summary_frame()
```

```
print(nb_summary_frame_test)
```

```
              mean   mean_se  mean_ci_lower  mean_ci_upper
0     31.087502  2.891884      25.906196      37.305082
1     12.837042  2.093210       9.325409      17.671037
2     49.370720  8.031297      35.892335      67.910544
4     16.190575  1.972483      12.751497      20.557172
```

셀 삭제를 실행취소하려면 ⌘/Ctrl+M Z 또는 수정 메뉴의 실행취소 옵션을 사용하세요.

```
2526  70.783289  9.793877      53.970306      92.833902
2530  20.551573  2.845563      15.667078      26.958897
2532  51.696883  7.270255      39.242609      68.103723
2538  29.958818  4.333251      22.563511      39.777975

[1293 rows x 4 columns]
```

```
predicted_counts = nb_summary_frame_test['mean']
```

```
#Accuracy of the test set
```

```
print("R-square of test set: ", round(r2_score(y_test, predicted_counts)*100, 2), "
```

```
R-square of test set:  8.32 %
```

Colab 유료 제품 - 여기에서 계약 취소

✓ 0초 오후 2:53에 완료됨

셀 삭제를 실행취소하려면 ⌘/Ctrl+M Z 또는 수정 메뉴의 실행취소 옵션을 사용하세요.