logistic_regression

April 25, 2023

[ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import matplotlib as mpl
import matplotlib.cm as cm
```

[ ]:
```
assessor_dir = 'datasets/assessorSequential.csv'
df = pd.read_csv(assessor_dir)
```

[ ]:
```
df = df.drop(columns=['Unnamed: 0','X11th.Draw','PIN','Township␣
↪Code','Neighborhood Code','Age','Longitude','Latitude','ZIP'])
```

[ ]:
```
draw_col = df.columns[2:12]
draw_col
```

[ ]:
```
Index(['X1st.Draw', 'X2nd.Draw', 'X3rd.Draw', 'X4th.Draw', 'X5th.Draw',
       'X6th.Draw', 'X7th.Draw', 'X8th.Draw', 'X9th.Draw', 'X10th.Draw'],
      dtype='object')
```

[ ]:
```
df['avg_draw'] = df[draw_col].mean(axis=1)
df = df.drop(columns=draw_col)
```

[ ]:
```
df_full = df.copy()
```

[ ]:
```
df_reduced = df.copy()
df_full.head(5)
```

[ ]:
```
  Date.Sampled              Address  Sale Price  Tract Median Income  \
0     9/4/2019     100XX S BELL AVE    280000.0             122727.0
1    7/16/2016   100XX S CALHOUN AVE         NaN              44423.0
2   12/17/2019   100XX S CALUMET AVE         NaN              40612.0
3   12/14/2019  100XX S CARPENTER ST         NaN              37207.0
4    7/14/2021  100XX S CARPENTER ST         NaN              37207.0

   avg_draw
0     2.433
```

```
1     9.866
2    10.399
3     8.663
4     9.280
```

[ ]:
```
df_full[~df_full.isin(['NaN', 'NaT']).any(axis=1)]
df_full.shape
df_full.dropna(inplace=True)
df_full.shape
```

[ ]:
```
(149, 5)
```

[ ]:
```
#assign 0 to <7.5 and 1 to >=7.5
df_full['avg_lead_target'] = np.where(df_full['avg_draw']<15, 0, 1)
```

[ ]:
```
df_full.head(10)
```

[ ]:
```
   Date.Sampled                   Address  Sale Price  Tract Median Income  \
0      9/4/2019          100XX S BELL AVE    280000.0              122727.0
17   12/20/2016      102XX S ARTESIAN AVE    219000.0               98281.0
20    9/17/2019       102XX S OGLESBY AVE    147400.0               30069.0
21    9/28/2016      103XX S HAMILTON AVE    330000.0              110344.0
36   10/10/2019  105XX S CENTRAL PARK AVE    280000.0              100361.0
42    11/4/2021        105XX S KEDZIE AVE    290000.0               91924.0
44    8/30/2021        105XX S SEELEY AVE    464000.0              110344.0
49   11/30/2021      106XX S EBERHART AVE     86000.0               45273.0
65    12/6/2021     108XX S EGGLESTON AVE    146000.0               41167.0
67     3/2/2022        108XX S HAMLIN AVE    273000.0              118640.0

    avg_draw  avg_lead_target
0      2.433                0
17     5.192                0
20    27.250                1
21     6.315                0
36    10.505                0
42    14.775                0
44     5.901                0
49    15.010                1
65     2.018                0
67     4.998                0
```

[ ]:
```
df_full['avg_lead_target'].value_counts()
```

[ ]:
```
0    121
1     28
Name: avg_lead_target, dtype: int64
```

[ ]:
```
df_full['sale_price'] = df_full['Sale Price']
df_full['tract_income'] = df_full['Tract Median Income']
df_full = df_full.drop(columns=['Sale Price','Tract Median Income'])
```

[ ]:
```
# class_0 = df_full[df_full['avg_lead_target']==0]
# print(class_0.shape)
# class_1 = df_full[df_full['avg_lead_target']==1]

# class_1_under = class_1.sample(class_0.shape[0])

# df_balanced = pd.concat([class_0, class_1_under], axis=0)
```

[ ]:
```
log_reg = smf.logit(formula='avg_lead_target ~ tract_income + sale_price',␣
↪data=df_full).fit()
```

```
Optimization terminated successfully.
         Current function value: 0.470499
         Iterations 6
```

[ ]:
```
print(log_reg.summary())
```

```
                           Logit Regression Results
==============================================================================
Dep. Variable:        avg_lead_target   No. Observations:                  149
Model:                          Logit   Df Residuals:                      146
Method:                           MLE   Df Model:                            2
Date:                Tue, 25 Apr 2023   Pseudo R-squ.:                 0.02627
Time:                        01:52:35   Log-Likelihood:                -70.104
converged:                       True   LL-Null:                       -71.996
Covariance Type:            nonrobust   LLR p-value:                    0.1509
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      -1.6872      0.525     -3.213      0.001      -2.716      -0.658
tract_income -6.155e-06   9.06e-06     -0.679      0.497   -2.39e-05    1.16e-05
sale_price    1.573e-06   8.63e-07      1.822      0.068   -1.19e-07    3.26e-06
==============================================================================
```

[ ]:
```
log_reg_reduced = smf.logit(formula='avg_lead_target ~ sale_price',␣
↪data=df_full).fit()
```

```
Optimization terminated successfully.
         Current function value: 0.472080
         Iterations 6
```

[ ]:
```
log_reg_reduced.summary()
```

[ ]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
```

```
                           Logit Regression Results
==============================================================================
Dep. Variable:        avg_lead_target   No. Observations:                  149
Model:                          Logit   Df Residuals:                      147
Method:                           MLE   Df Model:                            1
Date:                Tue, 25 Apr 2023   Pseudo R-squ.:                 0.02300
Time:                        01:52:35   Log-Likelihood:                -70.340
converged:                       True   LL-Null:                       -71.996
Covariance Type:            nonrobust   LLR p-value:                   0.06880
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -1.9554      0.354     -5.527      0.000      -2.649      -1.262
sale_price  1.193e-06   6.47e-07      1.845      0.065   -7.44e-08    2.46e-06
==============================================================================
"""
```

[ ]:
```
df_reduced = df_reduced[['Tract Median Income','avg_draw']]
df_reduced['avg_lead_target'] = np.where(df_reduced['avg_draw']<15, 0, 1)
df_reduced['tract_income'] = df_reduced['Tract Median Income']
df_reduced = df_reduced.drop(columns=['Tract Median Income'])
```

[ ]:
```
# class_0 = df_reduced[df_reduced['avg_lead_target']==0]
# print(class_0.shape)
# class_1 = df_reduced[df_reduced['avg_lead_target']==1]

# class_1_under = class_1.sample(class_0.shape[0])

# df_balanced1 = pd.concat([class_0, class_1_under], axis=0)
```

[ ]:
```
# df_balanced1['avg_lead_target'].value_counts()
```

[ ]:
```
log_reg1 = smf.logit(formula='avg_lead_target ~ tract_income', data=df_reduced).
↪fit()
```

```
Optimization terminated successfully.
         Current function value: 0.535740
         Iterations 5
```

[ ]:
```
log_reg1.summary()
```

[ ]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:        avg_lead_target   No. Observations:                 1217
Model:                          Logit   Df Residuals:                     1215
Method:                           MLE   Df Model:                           1
```

```
Date:                Tue, 25 Apr 2023   Pseudo R-squ.:                0.003038
Time:                        01:52:36   Log-Likelihood:                -652.00
converged:                       True   LL-Null:                       -653.98
Covariance Type:            nonrobust   LLR p-value:                   0.04622
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      -0.9078      0.169     -5.360      0.000      -1.240      -0.576
tract_income -4.755e-06   2.42e-06     -1.963      0.050    -9.5e-06   -7.22e-09
==============================================================================
"""
```

[ ]: