Logistic Regression on Attrition Data Set

Loading Data and Data Treatment:

```
import pandas as pd


Attrition_dataset = pd.read_csv("general_data.csv")


Attrition_dataset.head(2)
Out[100]:
  Age Attrition  ... YearsSinceLastPromotion YearsWithCurrManager
0  51       No  ...                       0                    0
1  31      Yes  ...                       1                    4

[2 rows x 24 columns]


Attrition_dataset.isnull().sum()
Out[102]:
Age                        0
Attrition                  0
BusinessTravel             0
Department                 0
DistanceFromHome           0
Education                  0
EducationField             0
EmployeeCount              0
EmployeeID                 0
Gender                     0
JobLevel                   0
JobRole                    0
MaritalStatus              0
MonthlyIncome              0
```

```
NumCompaniesWorked        19
Over18                    0
PercentSalaryHike         0
StandardHours             0
StockOptionLevel          0
TotalWorkingYears         9
TrainingTimesLastYear     0
YearsAtCompany            0
YearsSinceLastPromotion   0
YearsWithCurrManager      0
dtype: int64

Attrition_dataset = Attrition_dataset.fillna(Attrition_dataset.mean().round())

Attrition_dataset.columns
Out[104]:
Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',
       'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',
       'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',
       'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')

Attrition_dataset.dtypes
Out[105]:
Age                       int64
Attrition                 object
BusinessTravel            object
Department                object
DistanceFromHome          int64
```

```
Education                 int64
EducationField            object
EmployeeCount             int64
EmployeeID                int64
Gender                    object
JobLevel                  int64
JobRole                   object
MaritalStatus             object
MonthlyIncome             int64
NumCompaniesWorked        float64
Over18                    object
PercentSalaryHike         int64
StandardHours             int64
StockOptionLevel          int64
TotalWorkingYears         float64
TrainingTimesLastYear     int64
YearsAtCompany            int64
YearsSinceLastPromotion   int64
YearsWithCurrManager      int64
dtype: object
```

Encoding Categorical Features:

```
from sklearn import preprocessing

label_encoder = preprocessing.LabelEncoder()

Attrition_dataset["Attrition"] = label_encoder.fit_transform(Attrition_dataset["Attrition"])

Attrition_dataset["BusinessTravel"] =
label_encoder.fit_transform(Attrition_dataset["BusinessTravel"])
```

```
Attrition_dataset["Department"] = label_encoder.fit_transform(Attrition_dataset["Department"])

Attrition_dataset["EducationField"] =
label_encoder.fit_transform(Attrition_dataset["EducationField"])

Attrition_dataset["Gender"] = label_encoder.fit_transform(Attrition_dataset["Gender"])

Attrition_dataset["JobRole"] = label_encoder.fit_transform(Attrition_dataset["JobRole"])

Attrition_dataset["MaritalStatus"] = label_encoder.fit_transform(Attrition_dataset["MaritalStatus"])
```

Model Generation:

```
X = Attrition_dataset[['Age','BusinessTravel', 'Department', 'DistanceFromHome',
    'Education', 'EducationField','Gender',
    'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',
    'NumCompaniesWorked','PercentSalaryHike',
    'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
    'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']]

Y = Attrition_dataset[['Attrition']]

import statsmodels.api as sm

X1 = sm.add_constant(X)

Logistic = sm.Logit(Y, X1)

result = Logistic.fit()
Optimization terminated successfully.
      Current function value: 0.393008
      Iterations 7
```

result.summary()

Out[122]:

<class 'statsmodels.iolib.summary.Summary'>

```
                           Logit Regression Results
==============================================================================
Dep. Variable:              Attrition   No. Observations:                 4410
Model:                          Logit   Df Residuals:                     4390
Method:                           MLE   Df Model:                           19
Date:                Sun, 09 Aug 2020   Pseudo R-squ.:                  0.1102
Time:                        14:43:40   Log-Likelihood:                -1733.2
converged:                       True   LL-Null:                       -1947.9
Covariance Type:            nonrobust   LLR p-value:                 3.276e-79
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                        0.0752      0.414      0.182      0.856      -0.736       0.886
Age                         -0.0309      0.007     -4.523      0.000      -0.044      -0.018
BusinessTravel              -0.0177      0.065     -0.270      0.787      -0.146       0.111
Department                  -0.2422      0.081     -2.980      0.003      -0.402      -0.083
DistanceFromHome            -0.0013      0.005     -0.247      0.805      -0.012       0.009
Education                   -0.0628      0.043     -1.474      0.140      -0.146       0.021
EducationField              -0.0966      0.033     -2.895      0.004      -0.162      -0.031
Gender                       0.0860      0.090      0.960      0.337      -0.090       0.261
JobLevel                    -0.0241      0.040     -0.609      0.542      -0.102       0.054
JobRole                      0.0377      0.018      2.108      0.035       0.003       0.073
MaritalStatus                0.5888      0.063      9.324      0.000       0.465       0.713
MonthlyIncome           -1.873e-06   9.56e-07     -1.960      0.050   -3.75e-06    2.22e-10
NumCompaniesWorked           0.1164      0.018      6.342      0.000       0.080       0.152
PercentSalaryHike            0.0118      0.012      1.005      0.315      -0.011       0.035
StockOptionLevel            -0.0639      0.052     -1.238      0.216      -0.165       0.037
TotalWorkingYears           -0.0574      0.012     -4.822      0.000      -0.081      -0.034
TrainingTimesLastYear       -0.1466      0.035     -4.170      0.000      -0.216      -0.078
YearsAtCompany               0.0120      0.018      0.657      0.511      -0.024       0.048
YearsSinceLastPromotion      0.1322      0.020      6.459      0.000       0.092       0.172
YearsWithCurrManager        -0.1395      0.022     -6.307      0.000      -0.183      -0.096
============================================================================================
```

Inference:

Features **'Age', 'Department', 'EducationField','Gender', 'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsSinceLastPromotion', 'YearsWithCurrManager'** are significantly important for analysis the attrition in the company.

The probability for Attrition P(Y/N) is calculated as follows

P(Y/N) = 1/(1+e^(-k))

Where k = -0.0752 +(-0.0309)(Age) +(-0.0177)('BusinessTravel') +(-0.2422)('Department') +(-0.0013)('DistanceFromHome') +(-0.0628)('Education') +(-0.0966)('EducationField') +(0.0860)('Gender') +(-0.0241)('JobLevel') +(0.0377)('JobRole') +(0.5888)('MaritalStatus') +(-1.873e^(-06))('MonthlyIncome') +(0.1164)('NumCompaniesWorked') +(0.0118)('PercentSalaryHike') +(-0.0639)('StockOptionLevel') +(-0.0574)('TotalWorkingYears') +(-0.1466)('TrainingTimesLastYear') +(0.0120)('YearsAtCompany') +(0.1322)('YearsSinceLastPromotion') +(-0.1395)('YearsWithCurrManager')