**SVKM’s NMIMS**

**Mukesh Patel School of Technology Management & Engineering**

Program: B.Tech/MBATech/B.Tech(CSBS)/ BTech(DS)-311/BTI IX,

**Course: Machine Learning**

**Experiment No.04**

***(Students must submit the soft copy as per following segments within two hours of the practical.)***

|  |  |
| --- | --- |
| Roll No. C026 | Name: Anirbaan Ghatak |
| Class : B | Batch : EB1 |
| Date of Experiment: 22-08-2023 | Date of Submission: 22-08-2023 |
| Grade : |  |

**B.1 Task1**

# Name: Anirbaan Ghatak

# Roll no.: C026

# Implementation of Logistic Regression. Also to find the performance using the given two datasets.

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score

from sklearn.model\_selection import train\_test\_split

insurance = pd.read\_csv("prac4-insurancedata.csv")

sns.scatterplot(data=insurance, x="age", y="Bought\_Insurance")

plt.title("Scatter plot: Age vs Bought insurance")

X = insurance.iloc[:, 0].values.reshape(-1, 1)

y = insurance.iloc[:, 1].values.reshape(-1, 1)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

    X, y, test\_size=0.2, random\_state=32)

logisticRegression = LogisticRegression()

logisticRegression.fit(X\_train, y\_train)

predictions = logisticRegression.predict(X\_test)

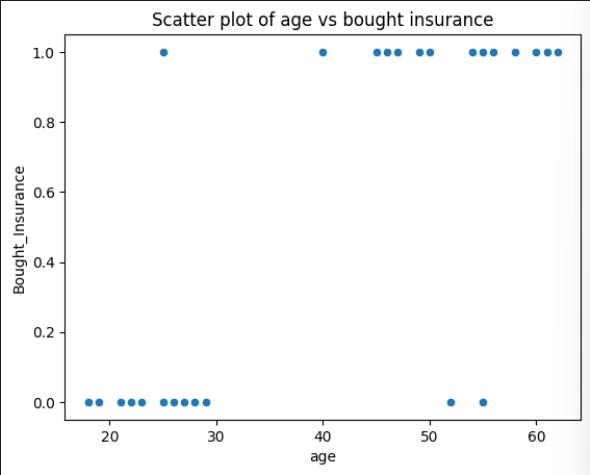
plt.scatter(X\_train, y\_train, color="g")

plt.plot(X\_test, predictions, 'k')

plt.show()

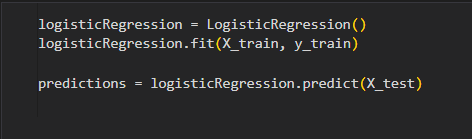
accuracy\_score(y\_train, logisticRegression.predict(X\_train))

1.Plot scatter plot, and justify the requirement of Logistic regression over Linear Regression

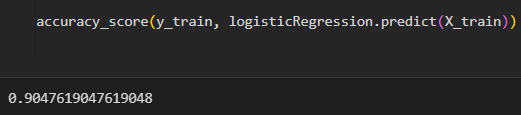


Logistic regression is suitable for binary classification and modeling probabilities due to its S-shaped curve and ability to handle categorical outcomes, while linear regression is used for predicting continuous numerical values and assumes a linear relationship between variable

2.Perform Logistic Regression



3.Measure Accuracy of the model



**B.2 Task 2**

# Name: Anirbaan Ghatak

# Roll no.: C026

# Implementation of Logistic Regression. Also to find the performance using the given two datasets.

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

hrdf = pd.read\_csv("prac4-HR\_comma\_sep.csv")

hrdf.dropna(inplace=True)

fig, ax = plt.subplots(figsize=(100, 100))

sns.histplot(hrdf['satisfaction\_level'], ax=ax)

hrdf['satisfaction\_level'] = hrdf['satisfaction\_level'].astype(

    str).apply(lambda x: x.strip())

hrdf['satisfaction\_level'] = pd.to\_numeric(

    hrdf['satisfaction\_level'], errors="coerce", downcast="float")

hrdf['satisfaction\_level'] = hrdf['satisfaction\_level'].fillna(

    hrdf['satisfaction\_level'].mean())

hrdf['salary'] = hrdf['salary'].apply(lambda x: x.strip())

hrdf['Department'] = hrdf['Department'].replace(['0', '1'], 'sales')

for i in hrdf.select\_dtypes(include='object').columns:

    sns.histplot(hrdf[i])

    plt.show()

hrdf.corr(numeric\_only=True)['left']

new\_hr = hrdf.drop(['Department', 'salary'], axis=1)

X, y = new\_hr.drop('left', axis=1), new\_hr['left']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

    X, y, test\_size=0.3, random\_state=32)

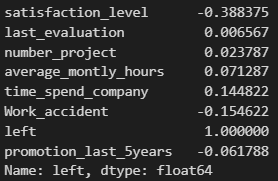
lr = LogisticRegression(max\_iter=1000)

lr.fit(X\_train, y\_train)

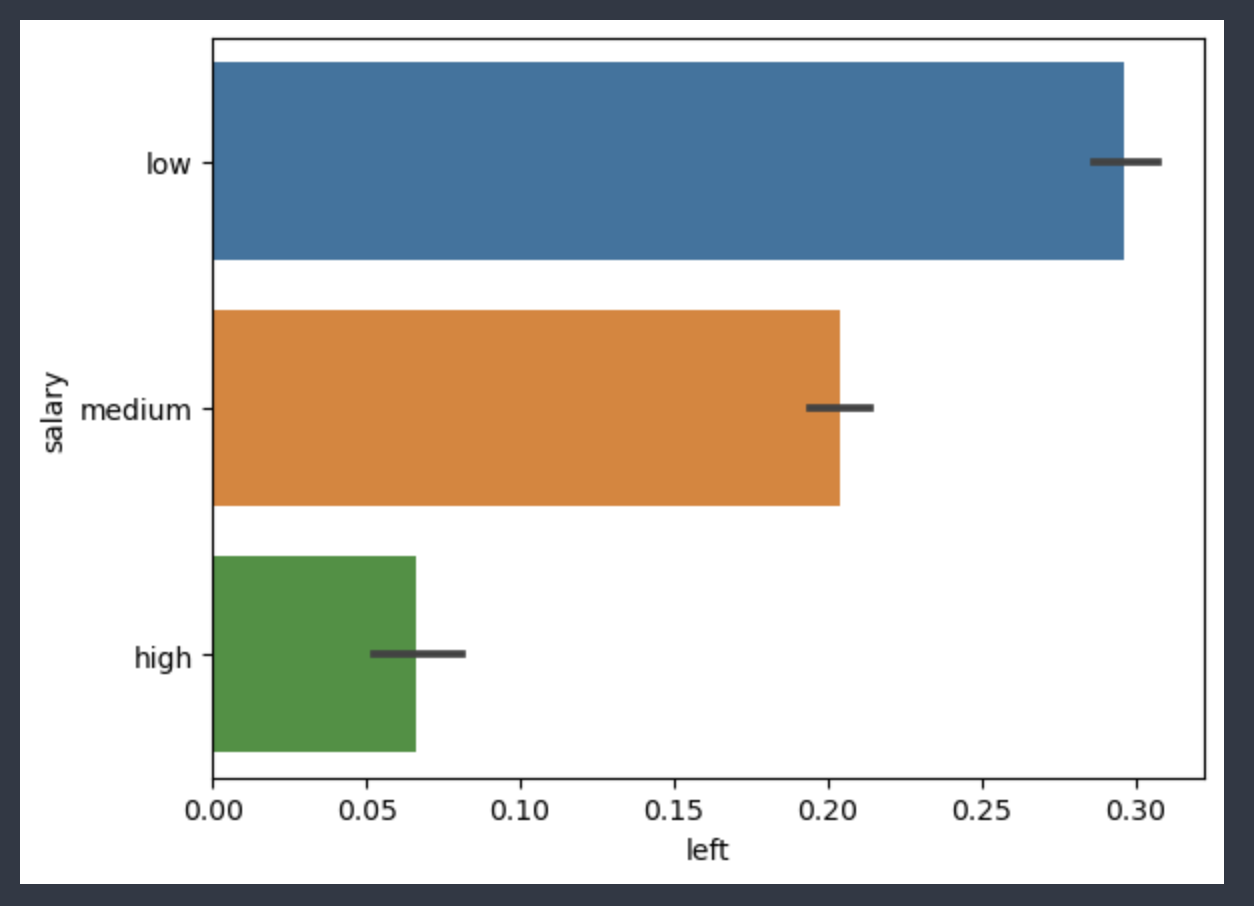
predictions = lr.predict(X\_test)

accuracy\_score(y\_test,predictions)

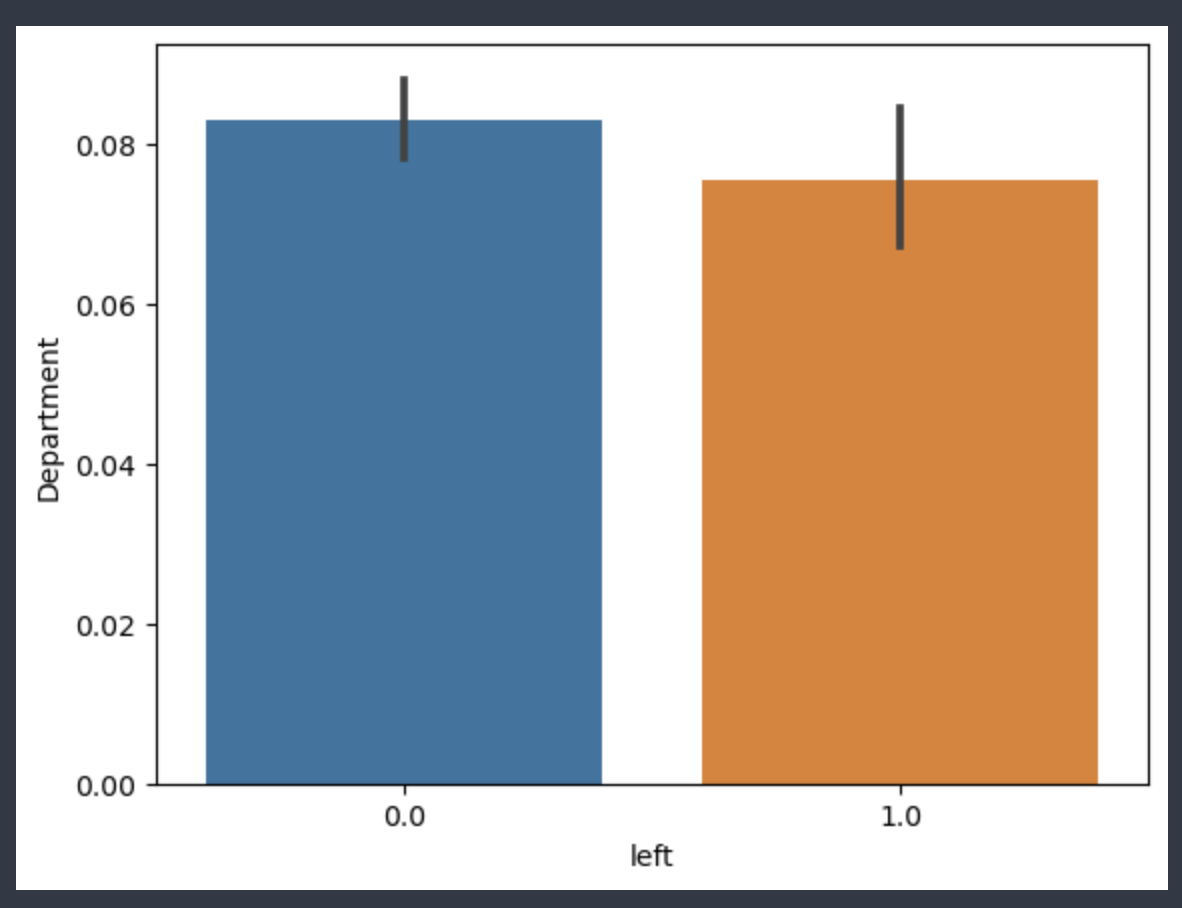
1.Exploratory Analysis to figure out which variables have direct and clear impact on employee retention (whether they leave the company or continue to work)



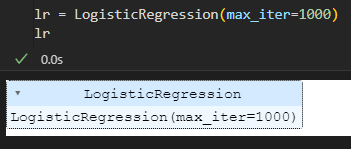
2.Plot bar charts showing impact of employee salaries on retention



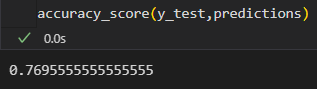
3.Plot bar charts showing correlation between department and employee retention



4.Build Logistic Regression model using variables that were narrowed down in step 1



5.Measure the Accuracy of the model

****

**B.3 Learning and Observations:**

I have gained knowledge about logistic regression, its capability to categorize inputs based on specific attributes, and how altering the shuffling method significantly impacts accuracy. Additionally, I have learned about the various hyperparameters associated with logistic regression.

**B.4 Conclusion:**

Implementation of Logistic Regression and understood various data cleaning and exploratory analysis methods.