Importing the requried libraries

In [1]:
```
import nltk
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import re
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
```

Importing the data set

In [2]:
```
JOB = pd.read_csv("C:/Users/naikc/Downloads/Job titles and industries.csv")
```

Exploratory Data Analysis

In [3]:
```
### Finding the first 10 columns and row's of the data set ###
```

In [4]:
```
JOB.head(10)
```

Out[4]:

	Job Title	Industry
0	technical support and helpdesk supervisor - co...	IT
1	senior technical support engineer	IT
2	head of it services	IT
3	js front end engineer	IT
4	network and telephony controller	IT
5	privileged access management expert	IT
6	devops engineers x 3 - global brand	IT
7	devops engineers x 3 - global brand	IT
8	data modeller	IT
9	php web developer £45,000 based in london	IT

In [5]:
```
### Finding the last 10 columns & row's of the given data set ###
```

In [6]:
```
JOB.tail(10)
```

Out[6]:

	Job Title	Industry
8576	marketing & social media specialist	Marketing
8577	senior php developer	Marketing
8578	social media graphic designer	Marketing
8579	sponsorship sales executive	Marketing
8580	marketing specialist	Marketing
8581	data entry clerk	Marketing
8582	content creator	Marketing
8583	sales & marketing manager	Marketing
8584	marketing & digital marketing consultant	Marketing
8585	creative copywriter (arabic/english)	Marketing

In [7]:
```
### Find the total number of colmun's & row's in the data set ###
```

In [8]:
```
JOB.shape
```

Out[8]:
```
(8586, 2)
```

In [9]:
```
## Number of Columns and Row's ##
```

In [16]:
```
print('Count of columns in the data is:  ', len(JOB.columns))
print('Count of rows in the data is:  ', len(JOB))
```

```
Count of columns in the data is:   2
Count of rows in the data is:   8586
```

In [11]:
```
### Find the data types of the given data set ###
```

In [12]:
```
JOB.dtypes
```

Out[12]:
```
Job Title    object
Industry     object
dtype: object
```

In [13]:
```
## Checking the unique variables in data set ##
```

In [14]:
```
JOB.nunique()
```

Out[14]:
```
Job Title    3890
Industry        4
dtype: int64
```

In [15]:
```
## Chekcing the unique variable in Industry vaiable ##
```

In [17]:
```
JOB['Industry'].unique()
```

Out[17]:
```
array(['IT', 'Marketing', 'Education', 'Accountancy'], dtype=object)
```

In [18]:
```
## Chekcing the unique variable in Job Title vaiable ##
```

In [19]:
```
JOB['Job Title'].unique()
```

Out[19]:
```
array(['technical support and helpdesk supervisor - county buildings, ayr soa04086',
       'senior technical support engineer', 'head of it services', ...,
       'sales & marketing manager',
       'marketing & digital marketing consultant',
       'creative copywriter (arabic/english)'], dtype=object)
```

In [20]:
```
## Checking the number of counts in Job Title variable ##
```

In [22]:
```
JOB['Job Title'].value_counts()
```

Out[22]:
```
marketing executive                                  91
php developer                                        54
trainee network technician                           53
software developer                                   53
marketing manager                                    49
                                                     ..
data analyst - sql                                    1
accounts payable specialist - temp to perm            1
project support co-ordinator - £19/£20 per hour paye  1
product development manager                           1
business analyst - it                                 1
```

```
Name: Job Title, Length: 3890, dtype: int64
```

In [23]:
```
## Checking the number of counts in Industry variable ##
```

In [24]:
```
JOB['Industry'].value_counts()
```

Out[24]:
```
IT             4746
Marketing      2031
Education      1435
Accountancy     374
Name: Industry, dtype: int64
```