Assignment 4

Create a ML-based prediction system to predict the job role for a new graduate

The prediction system for job role is to be done using ANN as mentioned in the assignment. Let's have a look at the data at first. There are 20000 rows and 39 columns.

The columns are as under:

```
'Acedamic percentage in Operating Systems', 'percentage in Algorithms',
'Percentage in Programming Concepts',
'Percentage in Software Engineering', 'Percentage in Computer Networks',
'Percentage in Electronics Subjects',
'Percentage in Computer Architecture', 'Percentage in Mathematics',
'Percentage in Communication skills', 'Hours working per day',
'Logical quotient rating', 'hackathons', 'coding skills rating',
'public speaking points', 'can work long time before system?',
'self-learning capability?', 'Extra-courses did', 'certifications',
'workshops', 'talenttests taken?', 'olympiads',
'reading and writing skills', 'memory capability score',
'Interested subjects', 'interested career area ', 'Job/Higher Studies?',
'Type of company want to settle in?',
'Taken inputs from seniors or elders', 'interested in games',
'Interested Type of Books', 'Salary Range Expected',
'In a Realtionship?', 'Gentle or Tuff behaviour?',
'Management or Technical', 'Salary/work', 'hard/smart worker',
 'worked in teams ever?', 'Introvert', 'Suggested Job Role'
```

Here the output variable/ target variable is "Suggested Job Role".

The information about different columns is as under:

```
#   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Acedamic percentage in Operating Systems  20000 non-null  int64
 1   percentage in Algorithms                  20000 non-null  int64
 2   Percentage in Programming Concepts        20000 non-null  int64
 3   Percentage in Software Engineering        20000 non-null  int64
 4   Percentage in Computer Networks           20000 non-null  int64
 5   Percentage in Electronics Subjects        20000 non-null  int64
 6   Percentage in Computer Architecture       20000 non-null  int64
 7   Percentage in Mathematics                 20000 non-null  int64
 8   Percentage in Communication skills        20000 non-null  int64
 9   Hours working per day                     20000 non-null  int64
 10  Logical quotient rating                   20000 non-null  int64
 11  hackathons                                20000 non-null  int64
 12  coding skills rating                      20000 non-null  int64
 13  public speaking points                    20000 non-null  int64
 14  can work long time before system?         20000 non-null  object
 15  self-learning capability?                 20000 non-null  object
 16  Extra-courses did                         20000 non-null  object
 17  certifications                            20000 non-null  object
 18  workshops                                 20000 non-null  object
 19  talenttests taken?                        20000 non-null  object
```

```
 20  olympiads                                20000 non-null  object
 21  reading and writing skills               20000 non-null  object
 22  memory capability score                  20000 non-null  object
 23  Interested subjects                      20000 non-null  object
 24  interested career area                   20000 non-null  object
 25  Job/Higher Studies?                      20000 non-null  object
 26  Type of company want to settle in?       20000 non-null  object
 27  Taken inputs from seniors or elders      20000 non-null  object
 28  interested in games                      20000 non-null  object
 29  Interested Type of Books                 20000 non-null  object
 30  Salary Range Expected                    20000 non-null  object
 31  In a Realtionship?                       20000 non-null  object
 32  Gentle or Tuff behaviour?                20000 non-null  object
 33  Management or Technical                  20000 non-null  object
 34  Salary/work                              20000 non-null  object
 35  hard/smart worker                        20000 non-null  object
 36  worked in teams ever?                    20000 non-null  object
 37  Introvert                                20000 non-null  object
 38  Suggested Job Role                       20000 non-null  object
```

The unique values for each column are captured as well.

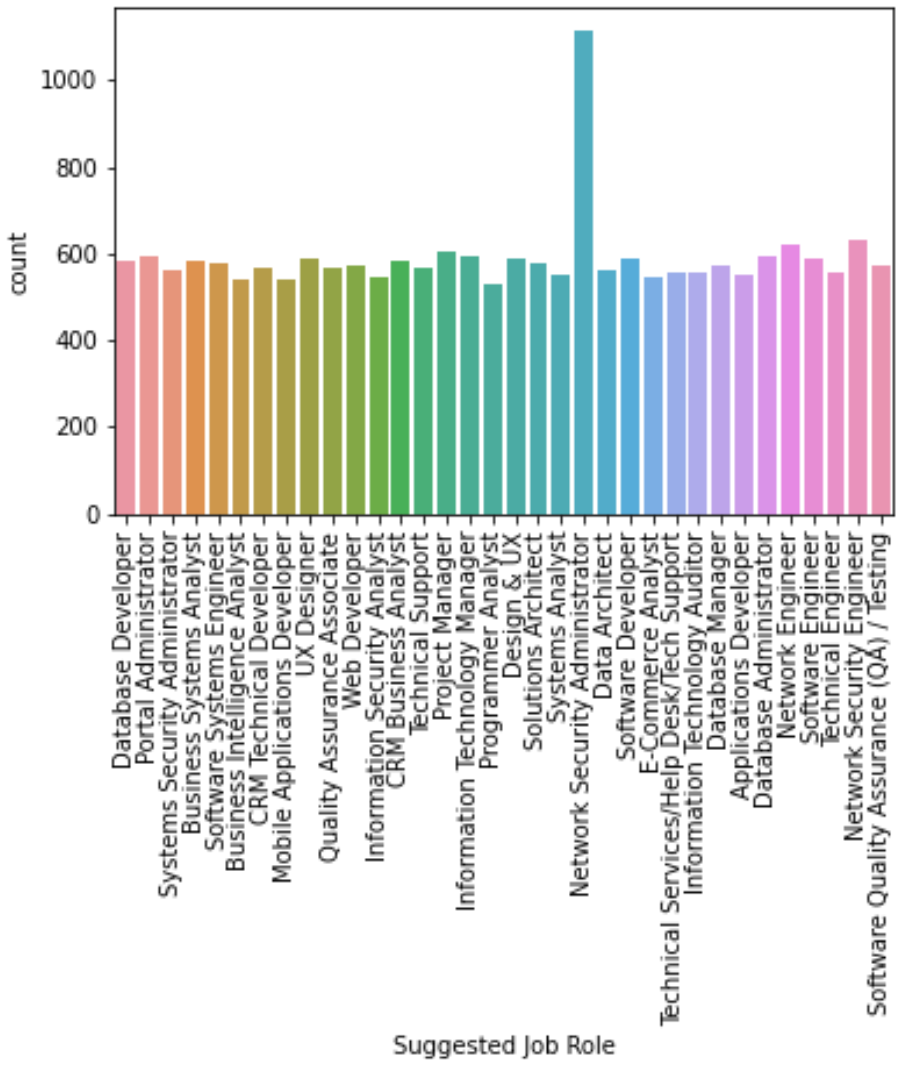
```
----------------------------------------
Unique values for Acedamic percentage in Operating Systems
----------------------------------------
[69 78 71 76 92 88 93 84 73 62 63 68 90 94 60 82 67 65 74 75 83 89 80 70
 66 85 61 81 79 86 64 91 72 77 87]
----------------------------------------
Unique values for percentage in Algorithms
----------------------------------------
[63 62 86 87 77 72 66 76 80 64 93 83 71 92 91 73 61 89 67 74 82 60 68 88
 70 85 81 78 84 69 94 75 65 79 90]
----------------------------------------
Unique values for Percentage in Programming Concepts
----------------------------------------
[78 73 91 60 90 62 69 88 66 85 70 81 61 77 63 94 68 76 75 93 64 65 84 72
 80 86 74 83 67 79 71 87 92 82 89]
----------------------------------------
Unique values for Percentage in Software Engineering
----------------------------------------
[87 60 84 67 79 62 81 91 83 90 71 74 63 86 70 75 92 93 72 78 85 64 82 65
 69 94 73 66 80 68 61 88 77 76 89]
----------------------------------------
Unique values for Percentage in Computer Networks
----------------------------------------
[94 71 61 89 93 90 66 81 82 70 77 65 62 64 78 63 67 86 69 92 84 85 87 68
 83 60 88 74 75 80 91 72 76 73 79]
----------------------------------------
Unique values for Percentage in Electronics Subjects
----------------------------------------
[94 70 81 73 89 84 93 63 69 82 72 67 65 61 88 91 74 90 80 79 75 62 76 77
 83 92 60 71 68 66 87 64 86 85 78]
----------------------------------------
Unique values for Percentage in Computer Architecture
----------------------------------------
```

```
[87 73 72 62 69 78 61 63 75 86 65 67 92 91 88 82 80 83 90 71 89 81 79 93
 70 84 85 76 74 64 77 94 60 68 66]
----------------------------------------
Unique values for Percentage in Mathematics
----------------------------------------
[84 72 88 71 63 94 87 89 64 81 62 73 65 82 60 61 80 77 78 68 76 83 92 93
 70 79 75 85 91 74 67 66 69 90 86]
----------------------------------------
Unique values for Percentage in Communication skills
----------------------------------------
[61 91 94 69 73 82 77 60 90 81 89 85 79 62 68 70 84 65 66 67 75 87 76 80
 78 63 64 88 92 71 93 86 83 74 72]
----------------------------------------
Unique values for Hours working per day
----------------------------------------
[ 9 12 11  7  4  6 10  8  5]
----------------------------------------
Unique values for Logical quotient rating
----------------------------------------
[4 7 1 5 3 2 9 6 8]
----------------------------------------
Unique values for hackathons
----------------------------------------
[0 1 4 3 2 6 5]
----------------------------------------
Unique values for coding skills rating
----------------------------------------
[4 2 1 6 8 3 5 9 7]
----------------------------------------
Unique values for public speaking points
----------------------------------------
[8 3 5 1 6 4 9 7 2]
----------------------------------------
Unique values for can work long time before system?
----------------------------------------
['yes' 'no']
----------------------------------------
Unique values for self-learning capability?
----------------------------------------
['yes' 'no']
----------------------------------------
Unique values for Extra-courses did
----------------------------------------
['yes' 'no']
----------------------------------------
Unique values for certifications
----------------------------------------
['shell programming' 'machine learning' 'app development' 'python'
 'r programming' 'information security' 'hadoop' 'distro making'
 'full stack']
----------------------------------------
Unique values for workshops
----------------------------------------
['cloud computing' 'database security' 'web technologies' 'data science'
 'testing' 'hacking' 'game development' 'system designing']
----------------------------------------
Unique values for talenttests taken?
```

```
----------------------------------------
['no' 'yes']
----------------------------------------
Unique values for olympiads
----------------------------------------
['yes' 'no']
----------------------------------------
Unique values for reading and writing skills
----------------------------------------
['excellent' 'poor' 'medium']
----------------------------------------
Unique values for memory capability score
----------------------------------------
['excellent' 'medium' 'poor']
----------------------------------------
Unique values for Interested subjects
----------------------------------------
['cloud computing' 'networks' 'hacking' 'Computer Architecture'
 'programming' 'parallel computing' 'IOT' 'data engineering'
 'Software Engineering' 'Management']
----------------------------------------
Unique values for interested career area
----------------------------------------
['system developer' 'Business process analyst' 'developer' 'testing'
 'security' 'cloud computing']
----------------------------------------
Unique values for Job/Higher Studies?
----------------------------------------
['higherstudies' 'job']
----------------------------------------
Unique values for Type of company want to settle in?
----------------------------------------
['Web Services' 'SAaS services' 'Sales and Marketing'
 'Testing and Maintainance Services' 'product development' 'BPA'
 'Service Based' 'Product based' 'Cloud Services' 'Finance']
----------------------------------------
Unique values for Taken inputs from seniors or elders
----------------------------------------
['no' 'yes']
----------------------------------------
Unique values for interested in games
----------------------------------------
['no' 'yes']
----------------------------------------
Unique values for Interested Type of Books
----------------------------------------
['Prayer books' 'Childrens' 'Travel' 'Romance' 'Cookbooks' 'Self help'
 'Drama' 'Math' 'Religion-Spirituality' 'Anthology' 'Trilogy'
 'Autobiographies' 'Mystery' 'Diaries' 'Journals' 'History' 'Art'
 'Dictionaries' 'Horror' 'Encyclopedias' 'Action and Adventure' 'Fantasy'
 'Comics' 'Science fiction' 'Series' 'Guide' 'Biographies' 'Health'
 'Satire' 'Science' 'Poetry']
----------------------------------------
Unique values for Salary Range Expected
----------------------------------------
['salary' 'Work']
----------------------------------------
```

```
Unique values for In a Realtionship?
----------------------------------------
['no' 'yes']
----------------------------------------
Unique values for Gentle or Tuff behaviour?
----------------------------------------
['stubborn' 'gentle']
----------------------------------------
Unique values for Management or Technical
----------------------------------------
['Management' 'Technical']
----------------------------------------
Unique values for Salary/work
----------------------------------------
['salary' 'work']
----------------------------------------
Unique values for hard/smart worker
----------------------------------------
['hard worker' 'smart worker']
----------------------------------------
Unique values for worked in teams ever?
----------------------------------------
['yes' 'no']
----------------------------------------
Unique values for Introvert
----------------------------------------
['no' 'yes']
----------------------------------------
Unique values for Suggested Job Role
----------------------------------------
['Database Developer' 'Portal Administrator'
 'Systems Security Administrator' 'Business Systems Analyst'
 'Software Systems Engineer' 'Business Intelligence Analyst'
 'CRM Technical Developer' 'Mobile Applications Developer' 'UX Designer'
 'Quality Assurance Associate' 'Web Developer'
 'Information Security Analyst' 'CRM Business Analyst' 'Technical Support
'
 'Project Manager' 'Information Technology Manager' 'Programmer Analyst'
 'Design & UX' 'Solutions Architect' 'Systems Analyst'
 'Network Security Administrator' 'Data Architect' 'Software Developer'
 'E-Commerce Analyst' 'Technical Services/Help Desk/Tech Support'
 'Information Technology Auditor' 'Database Manager'
 'Applications Developer' 'Database Administrator' 'Network Engineer'
 'Software Engineer' 'Technical Engineer' 'Network Security Engineer'
 'Software Quality Assurance (QA) / Testing']
```

If the count plot is drawn for the target variable it looks like (as shown below).



This shows that dataset has high entries of Network Security Administrator, which can give problem during training and testing. Also as seen above from the information of columns that some columns have values that are not numerical, so **OneHotEncoding** is done. Without it, one may not be able to fit ANN onto the data.

Code for the same is shown below.

```
from sklearn.preprocessing import OneHotEncoder
X1 = OneHotEncoder().fit_transform(X)
```

Now, grid search is done on split of 80-20 and data is not transformed any further and then ANN is fit there so as to see performance on raw data.

The grid search is performed on following parameters.

```
param_grid = [
        {
            'random_state': [1],
            'activation' : ['identity','logistic', 'tanh', 'relu'],
            'solver' : ['lbfgs', 'sgd', 'adam'],
            'hidden_layer_sizes': [
             (16,),(7,),(25,),(28,),
             (16,16),(7,7),(25,25),(28,28),
             (16,16,16),(7,7,7),(25,25,25),(28,28,28),
             ]
        }
       ]
```

The scoring is done on the basis of accuracy and then other metrics are reported for the best one. Best results came with the parameters.

```
{'activation': 'relu', 'hidden_layer_sizes': (16,), 'random_state': 1,
'solver': 'sgd'}
```

The training and test accuracy came out to be quite low.

```
Training Accuracy Score:  0.056375
Testing Accuracy Score:  0.055
```

For the best parameter, the metrics are shown below.

```
Train confusion matrix
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
Test confusion matrix
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
Train Classification Report
                                        precision    recall  f1-score   support

                  Applications Developer       0.00      0.00      0.00         1
            Business Intelligence Analyst       0.00      0.00      0.00         0
                 Business Systems Analyst       0.00      0.00      0.00         0
                     CRM Business Analyst       0.00      0.00      0.00         3
                   CRM Technical Developer       0.00      0.00      0.00         0
                            Data Architect       0.00      0.00      0.00         0
                     Database Administrator       0.00      0.00      0.00         7
```

Database Developer	0.00	0.00	0.00	0
Database Manager	0.00	0.00	0.00	0
Design & UX	0.00	0.00	0.00	0
E-Commerce Analyst	0.00	0.00	0.00	0
Information Security Analyst	0.00	0.00	0.00	0
Information Technology Auditor	0.00	0.00	0.00	0
Information Technology Manager	0.00	1.00	0.00	1
Mobile Applications Developer	0.00	0.00	0.00	1
Network Engineer	0.01	0.12	0.02	59
Network Security Administrator	1.00	0.06	0.11	15856
Network Security Engineer	0.00	0.00	0.00	1
Portal Administrator	0.00	0.00	0.00	0
Programmer Analyst	0.00	0.05	0.01	38
Project Manager	0.00	0.00	0.00	0
Quality Assurance Associate	0.00	0.00	0.00	0
Software Developer	0.00	0.00	0.00	1
Software Engineer	0.00	0.00	0.00	2
Software Quality Assurance (QA) / Testing	0.00	0.08	0.01	24
Software Systems Engineer	0.00	0.50	0.00	2
Solutions Architect	0.00	0.00	0.00	0
Systems Analyst	0.00	0.00	0.00	4
Systems Security Administrator	0.00	0.00	0.00	0
Technical Engineer	0.00	0.00	0.00	0
Technical Services/Help Desk/Tech Support	0.00	0.00	0.00	0
Technical Support	0.00	0.00	0.00	0
UX Designer	0.00	0.00	0.00	0
Web Developer	0.00	0.00	0.00	0
accuracy			0.06	16000
macro avg	0.03	0.05	0.00	16000
weighted avg	0.99	0.06	0.11	16000

Test Classification Report

	precision	recall	f1-score	support
Applications Developer	0.00	0.00	0.00	0
Business Intelligence Analyst	0.00	0.00	0.00	0
Business Systems Analyst	0.00	0.00	0.00	0
CRM Business Analyst	0.00	0.00	0.00	1
CRM Technical Developer	0.00	0.00	0.00	0
Data Architect	0.00	0.00	0.00	0
Database Administrator	0.00	0.00	0.00	0
Database Developer	0.00	0.00	0.00	0
Database Manager	0.00	0.00	0.00	0
Design & UX	0.00	0.00	0.00	0
E-Commerce Analyst	0.00	0.00	0.00	0
Information Security Analyst	0.00	0.00	0.00	0
Information Technology Auditor	0.00	0.00	0.00	0
Information Technology Manager	0.00	0.00	0.00	0
Mobile Applications Developer	0.00	0.00	0.00	1
Network Engineer	0.01	0.06	0.02	18
Network Security Administrator	0.99	0.05	0.10	3967
Network Security Engineer	0.00	0.00	0.00	0
Portal Administrator	0.00	0.00	0.00	0
Programmer Analyst	0.00	0.00	0.00	6
Project Manager	0.00	0.00	0.00	0
Quality Assurance Associate	0.00	0.00	0.00	0

```
                        Software Developer       0.00      0.00      0.00         0
                         Software Engineer       0.00      0.00      0.00         0
Software Quality Assurance (QA) / Testing       0.01      0.17      0.02         6
                Software Systems Engineer       0.00      0.00      0.00         1
                      Solutions Architect       0.00      0.00      0.00         0
                         Systems Analyst       0.00      0.00      0.00         0
           Systems Security Administrator       0.00      0.00      0.00         0
                        Technical Engineer       0.00      0.00      0.00         0
Technical Services/Help Desk/Tech Support       0.00      0.00      0.00         0
                        Technical Support       0.00      0.00      0.00         0
                              UX Designer       0.00      0.00      0.00         0
                           Web Developer       0.00      0.00      0.00         0

                                accuracy                           0.06      4000
                               macro avg       0.03      0.01      0.00      4000
                            weighted avg       0.98      0.06      0.10      4000

Train classwise accuracies
[0.                nan        nan 0.                nan        nan
 0.                nan        nan        nan        nan        nan
        nan 1.          0.         0.11864407 0.0560671  0.
        nan 0.05263158        nan        nan 0.         0.
 0.08333333 0.5               nan 0.                nan        nan
        nan        nan        nan        nan]
Test classwise accuracies
[       nan        nan        nan 0.                nan        nan
        nan        nan        nan        nan        nan        nan
        nan        nan 0.         0.05555556 0.05495337        nan
        nan 0.                nan        nan        nan        nan
 0.16666667 0.                nan        nan        nan        nan
        nan        nan        nan        nan]
```

NOTE: “nan” is for the classes that didn’t appear.

As already discussed, the training and testing accuracy were very low and there needs to be some modification. Then next try is made which included applying StandardScaler on X. Again, grid search is done on same parameters. One thing to add here is that values for hidden layers are tested based on the formulas that people recommend for number of neurons in hidden layer. There are many of them and all those are included in grid search. Best model came out to be the one with following parameters.

```
{'activation': 'logistic', 'hidden_layer_sizes': (7,), 'random_state': 1,
'solver': 'sgd'}
```

Accuracy, confusion matrix, class wise accuracy and classification reports were comparable to first try only. No, further improvement is noticed.

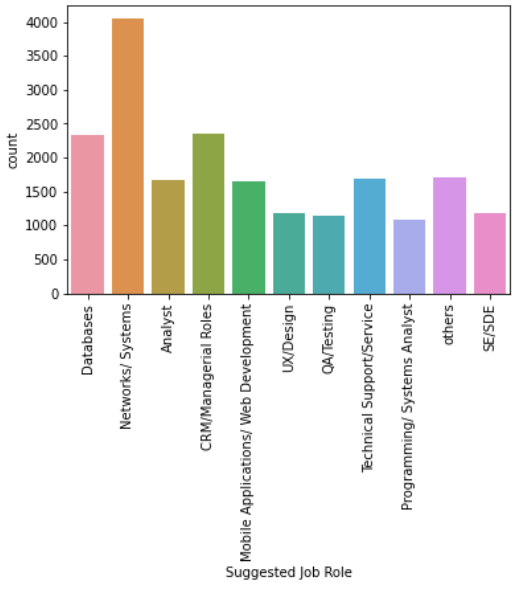
Now, next try is to club some roles. The clubbing is performed as under.

- 'Solutions Architect', 'Data Architect', 'Information Technology Auditor' → others
- 'CRM Business Analyst', 'CRM Technical Developer', 'Project Manager', 'Information Technology Manager' → Managerial roles/ CRM
- 'Business Systems Analyst','Business Intelligence Analyst','E-Commerce Analyst'→ Analyst
- 'Mobile Applications Developer','Web Developer"Applications Developer' → Mobile Applications/ Web Development
- 'Software Quality Assurance (QA) / Testing', 'Quality Assurance Associate' → QA/Testing
- 'UX Designer', 'Design & UX' → UX/Design
- 'Database Developer' ,'Database Administrator' ,'Database Manager', 'Portal Administrator' → Databases
- 'Programmer Analyst', 'Systems Analyst'→ Programming/ Systems Analyst
- 'Network Security Administrator', 'Network Security Engineer', 'Network Engineer', 'Systems Security Administrator', 'Software Systems Engineer', 'Information Security Analyst' → Networks/ Systems
- 'Software Engineer', 'Software Developer' → SE/SDE
- 'Technical Engineer', 'Technical Services/Help Desk/Tech Support', 'Technical Support' → Technical Support/Service

Now when the grid search is performed, then the model with following parameters gave the best results.

```
{'activation': 'logistic', 'hidden_layer_sizes': (16, 16), 'random_state'
: 1, 'solver': 'sgd'}
```

Now the distribution is as under:

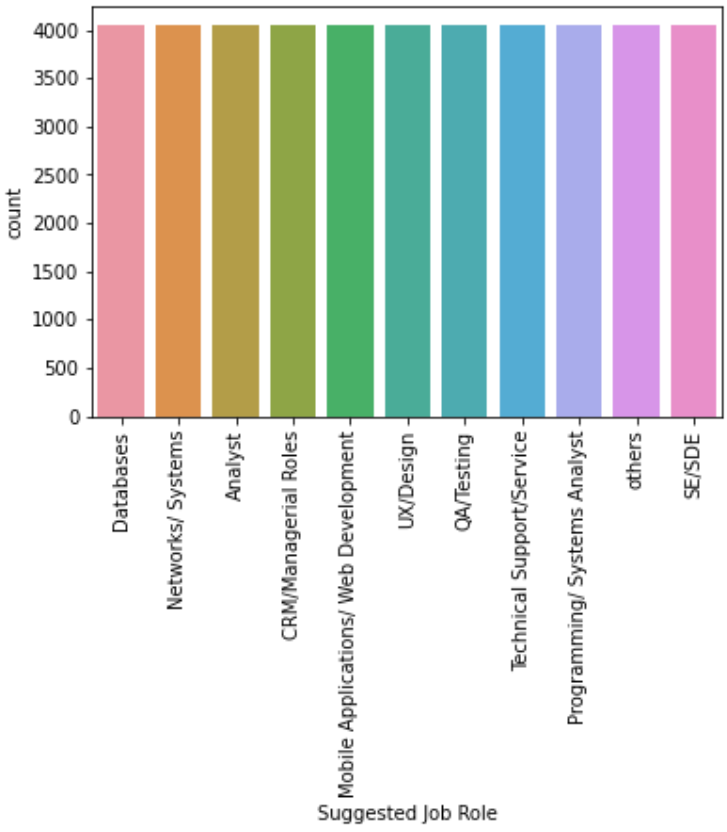


Though the accuracy and other metrics are still low, but way too better than the previous ones. The metrics are as under:

```
Training Accuracy Score:  0.201
Testing Accuracy Score:  0.20675
Train confusion matrix
[[   0    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0    0]
 [1329 1876 1869 1334 3216  875  917  935 1356  934 1359]
 [   0    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0    0]]
Test confusion matrix
[[  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [339 468 468 325 827 204 219 242 324 243 341]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]]
Train Classification Report
                                    precision    recall  f1-score   support

                           Analyst       0.00      0.00      0.00         0
             CRM/Managerial Roles        0.00      0.00      0.00         0
                         Databases       0.00      0.00      0.00         0
Mobile Applications/ Web Development       0.00      0.00      0.00         0
                  Networks/ Systems       1.00      0.20      0.33     16000
        Programming/ Systems Analyst       0.00      0.00      0.00         0
                          QA/Testing       0.00      0.00      0.00         0
                             SE/SDE        0.00      0.00      0.00         0
          Technical Support/Service       0.00      0.00      0.00         0
                          UX/Design       0.00      0.00      0.00         0
                             others       0.00      0.00      0.00         0

                           accuracy                           0.20     16000
                          macro avg       0.09      0.02      0.03     16000
                       weighted avg       1.00      0.20      0.33     16000

Test Classification Report
                                    precision    recall  f1-score   support

                           Analyst       0.00      0.00      0.00         0
             CRM/Managerial Roles        0.00      0.00      0.00         0
                         Databases       0.00      0.00      0.00         0
Mobile Applications/ Web Development       0.00      0.00      0.00         0
                  Networks/ Systems       1.00      0.21      0.34      4000
        Programming/ Systems Analyst       0.00      0.00      0.00         0
```

```
                          QA/Testing       0.00      0.00      0.00         0
                             SE/SDE        0.00      0.00      0.00         0
          Technical Support/Service        0.00      0.00      0.00         0
                         UX/Design        0.00      0.00      0.00         0
                             others       0.00      0.00      0.00         0

                           accuracy                           0.21      4000
                          macro avg       0.09      0.02      0.03      4000
                       weighted avg       1.00      0.21      0.34      4000

Train classwise accuracies
[  nan   nan   nan   nan 0.201   nan   nan   nan   nan   nan   nan]
Test classwise accuracies
[    nan     nan     nan     nan 0.20675     nan     nan     nan     nan
     nan     nan]
```

Now, to get better results, oversampling seems to be something that has to be tried. After oversampling, the distribution is as under.



Here many parameters were tried and tested and accuracy increased a lot due to oversampling. All the tried and tested parameters are shown in the notebook. Some seem to overfit and hence, trial is done for parameters on different train test split as shown on next page.

```
clf5_try1 = MLPClassifier(activation='tanh', hidden_layer_sizes = (44,44,44), solver = 'adam', random_state=1)
clf5_try1.fit(X_trains,y_trains)
print("Training Accuracy Score: ",accuracy_score(clf5_try1.predict(X_trains),y_trains))
print("Testing Accuracy Score: ",accuracy_score(clf5_try1.predict(X_tests),y_tests))
```

```
Training Accuracy Score:  0.9967556454980566
Testing Accuracy Score:  0.7545345525408484
```

```
from sklearn.model_selection import train_test_split
X_trainss, X_testss, y_trainss, y_testss = train_test_split(X_ovs,y_ovs,test_size=0.25)
```

```
clf5_try11 = MLPClassifier(activation='tanh', hidden_layer_sizes = (44,44,44), solver = 'adam', random_state=1)
clf5_try11.fit(X_trainss,y_trainss)
print("Training Accuracy Score: ",accuracy_score(clf5_try11.predict(X_trainss),y_trainss))
print("Testing Accuracy Score: ",accuracy_score(clf5_try11.predict(X_testss),y_testss))
```

```
Training Accuracy Score:  0.9966720633207411
Testing Accuracy Score:  0.7675150643043439
```

```
from sklearn.model_selection import train_test_split
X_trainsss, X_testsss, y_trainsss, y_testsss = train_test_split(X_ovs,y_ovs,test_size=0.4)
clf5_try111 = MLPClassifier(activation='tanh', hidden_layer_sizes = ((44,44,44)), solver = 'adam', random_state=1)
clf5_try111.fit(X_trainsss,y_trainsss)
print("Training Accuracy Score: ",accuracy_score(clf5_try111.predict(X_trainsss),y_trainsss))
print("Testing Accuracy Score: ",accuracy_score(clf5_try111.predict(X_testsss),y_testsss))
```

```
Training Accuracy Score:  0.9982011018251321
Testing Accuracy Score:  0.7189432265317595
```

```
from sklearn.model_selection import train_test_split
X_trainsss, X_testsss, y_trainsss, y_testsss = train_test_split(X_ovs,y_ovs,test_size=0.1)
clf5_try111 = MLPClassifier(activation='tanh', hidden_layer_sizes = (44,44,44), solver = 'adam', random_state=1)
clf5_try111.fit(X_trainsss,y_trainsss)
print("Training Accuracy Score: ",accuracy_score(clf5_try111.predict(X_trainsss),y_trainsss))
print("Testing Accuracy Score: ",accuracy_score(clf5_try111.predict(X_testsss),y_testsss))
```

```
Training Accuracy Score:  0.9313179262960649
Testing Accuracy Score:  0.7486510791366906
```

```
from sklearn.model_selection import train_test_split
X_trainn, X_testn, y_trainn, y_testn = train_test_split(X_ovs,y_ovs,test_size=0.15)
clf5_tryn = MLPClassifier(activation='tanh', hidden_layer_sizes = (44,44,44), solver = 'adam', random_state=1)
clf5_tryn.fit(X_trainn,y_trainn)
print("Training Accuracy Score: ",accuracy_score(clf5_tryn.predict(X_trainn),y_trainn))
print("Testing Accuracy Score: ",accuracy_score(clf5_tryn.predict(X_testn),y_testn))
```

```
Training Accuracy Score:  0.9451880852864928
Testing Accuracy Score:  0.7549093089491831
```

The last one seems to be best which is not overfitting and giving good accuracy. For now this seems to be best till now but still there can be experimentation by further clubbing the features, oversampling and then fit ANN. Now the labels can be as shown on next page.

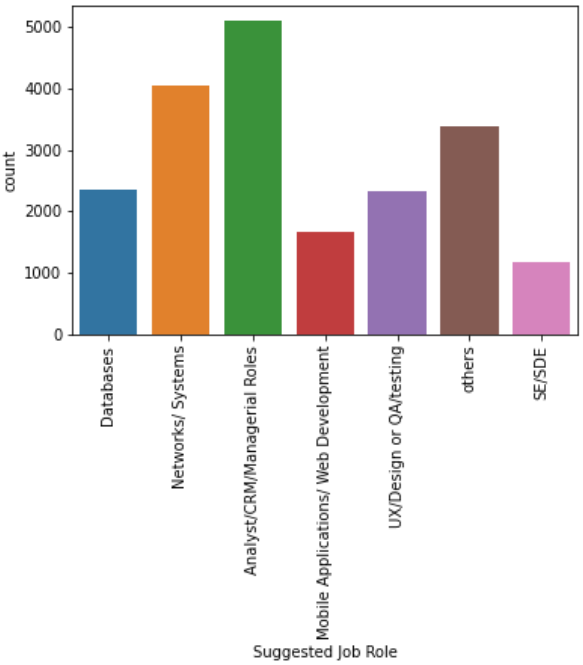


Figure: Before oversampling

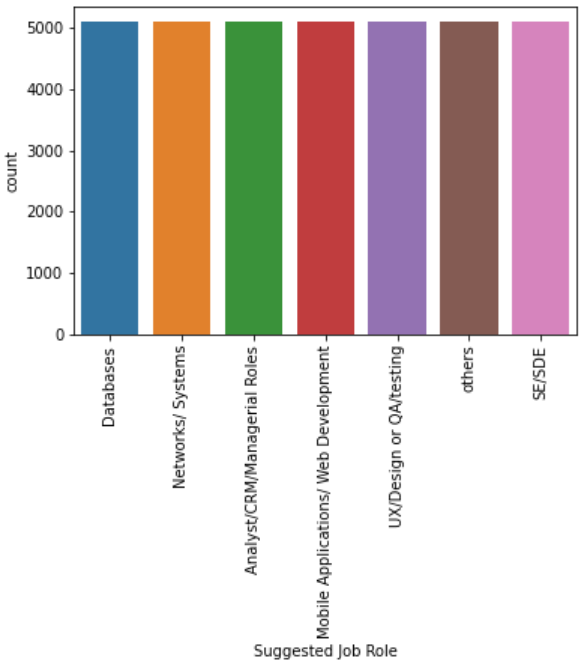


Figure: After oversampling

In this case, for different parameters, the training accuracy increased but test accuracy was a little low, so best one was the previous one. Let's report the parameters and metrics for best model so far.

CONCLUSION:

Model performs best when y is clubbed as the following

'Databases', 'Networks/ Systems', 'Analyst',
'CRM/Managerial Roles', 'Mobile Applications/ Web Development',
'UX/Design', 'QA/Testing', 'Technical Support/Service',
'Programming/ Systems Analyst', 'others', 'SE/SDE'

i.e. in total 11 classes

Clubbing technique

- 'Solutions Architect', 'Data Architect', 'Information Technology Auditor' → others
- 'CRM Business Analyst', 'CRM Technical Developer', 'Project Manager', 'Information Technology Manager' → Managerial roles/ CRM
- 'Business Systems Analyst','Business Intelligence Analyst','E-Commerce Analyst'→ Analyst
- 'Mobile Applications Developer','Web Developer"Applications Developer' → Mobile Applications/ Web Development
- 'Software Quality Assurance (QA) / Testing', 'Quality Assurance Associate' → QA/Testing
- 'UX Designer', 'Design & UX' → UX/Design

- 'Database Developer' ,'Database Administrator' ,'Database Manager', 'Portal Administrator' → Databases
- 'Programmer Analyst', 'Systems Analyst'→ Programming/ Systems Analyst
- 'Network Security Administrator', 'Network Security Engineer', 'Network Engineer', 'Systems Security Administrator', 'Software Systems Engineer', 'Information Security Analyst' → Networks/ Systems
- 'Software Engineer', 'Software Developer' → SE/SDE
- 'Technical Engineer', 'Technical Services/Help Desk/Tech Support', 'Technical Support' → Technical Support/Service

Transformation on X ==> OneHotEncoder and StandardScaler applied

Transformation on dataset as whole ==> oversampling

with following parameters:
* activation='tanh'
* hidden_layer_sizes = (44,44,44)
* solver = 'adam'
* train_test_split = 85:15

Evaluation metrics are as follows:

```
Training Accuracy Score:  0.9506110787789006
Testing Accuracy Score:  0.7609054114825363
----------------------------------
Train confusion matrix
----------------------------------
[[3361    9    6    0   36    0    0    0    3    0    4]
 [  12 3066   25   17  364    3    0    5   19    3    6]
 [  11   43 3216    2  262    0    0    0   15    0    9]
 [   3   17    5 3330   64    1    0    1    8    2    1]
 [  28  265  159   29 2538    1    0    8   41    0   31]
 [   2    3    0    9    4 3446    0    0    2    1    2]
 [   2    3   10    0   15    0 3446    0    1    1   21]
 [   1    4    1    0   11    0    0 3406    1    3    4]
 [   4   26   13    7   84    0    0    0 3332    0    5]
 [   3    3    1    2    5    0    0    0    3 3434    0]
 [   4   11   16    1   53    0    1    1    5    0 3360]]
----------------------------------
Test confusion matrix
----------------------------------
[[509  19  11  11  45   0   1   2   8   6  10]
 [ 21 349  27  16  62   3   2   6  16   4  11]
 [ 11  32 349  13  69   4   0   2  18   1  14]
 [  7  17  22 527  56   4   5  10  13   0  13]
 [ 20  69  64  22  76   3   7   7  29   1  26]
 [ 10  18  14  11  45 575   1   4   5   3   6]
 [  3  10  24  10  44   1 569   1   6   5  13]
 [  8  19  11   4  39   0   8 582  12   0  14]
 [  7  17  30  15  64   0   0   3 489   3   8]
 [  5  16  17   7  39   2   1   3   3 572   6]
 [ 11  27  22  10  68   0   2   2  14   4 479]]
----------------------------------
Train Classification Report
----------------------------------
                                    precision    recall  f1-score   support

                           Analyst       0.98      0.98      0.98      3419
              CRM/Managerial Roles       0.89      0.87      0.88      3520
                         Databases       0.93      0.90      0.92      3558
Mobile Applications/ Web Development       0.98      0.97      0.98      3432
                  Networks/ Systems       0.74      0.82      0.78      3100
        Programming/ Systems Analyst       1.00      0.99      1.00      3469
                        QA/Testing       1.00      0.98      0.99      3499
                           SE/SDE       1.00      0.99      0.99      3431
          Technical Support/Service       0.97      0.96      0.97      3471
                        UX/Design       1.00      1.00      1.00      3451
                           others       0.98      0.97      0.97      3452

                         accuracy                           0.95     37802
                        macro avg       0.95      0.95      0.95     37802
                     weighted avg       0.95      0.95      0.95     37802

----------------------------------
Test Classification Report
----------------------------------
                                    precision    recall  f1-score   support

                           Analyst       0.83      0.82      0.82       622
              CRM/Managerial Roles       0.59      0.68      0.63       517
                         Databases       0.59      0.68      0.63       513
Mobile Applications/ Web Development       0.82      0.78      0.80       674
                  Networks/ Systems       0.13      0.23      0.16       324
        Programming/ Systems Analyst       0.97      0.83      0.90       692
                        QA/Testing       0.95      0.83      0.89       686
                           SE/SDE       0.94      0.84      0.88       697
          Technical Support/Service       0.80      0.77      0.78       636
                        UX/Design       0.95      0.85      0.90       671
```

```
                              others       0.80      0.75      0.77       639

                            accuracy                           0.76      6671
                           macro avg       0.76      0.73      0.74      6671
                        weighted avg       0.80      0.76      0.78      6671

-----------------------------------
Train classwise accuracies
-----------------------------------
[0.98303598 0.87102273 0.90387858 0.97027972 0.81870968 0.99336985
 0.98485282 0.99271349 0.9599539  0.99507389 0.97334878]
-----------------------------------
Test classwise accuracies
-----------------------------------
[0.81832797 0.67504836 0.68031189 0.78189911 0.2345679  0.83092486
 0.82944606 0.83500717 0.76886792 0.85245902 0.74960876]
-----------------------------------
Train mean class accuracies:  0.9496581290460976
Test classwise accuracies:  0.7324062754444038
```