Aryan GD Singh | 2019459

AI Assignment 4 report

Job role prediction system

- Preprocessing of data
- Defining the model
- Training and testing
- Results analysis

Preprocessing

We have 20,000 samples with 38 input attributes, 1 target attribute with 34 classes.

1. First we rename the columns so they have simpler, easy-to-use names.
2. No null values found
3. The correlation matrix shows none of the attributes are highly correlated
4. Job roles are combined into 9 total classes.

    a. Design & UX, UX Designer → UX & Design
    b. Business Systems Analyst, Business Intelligence Analyst, Programmer Analyst, CRM Business Analyst, Systems Analyst, Information Security Analyst, E-Commerce Analyst → Analyst
    c. Database Manager, Information Technology Manager, Project Manager → Manager
    d. Network Security Administrator, Database Administrator, Portal Administrator, Systems Security Administrator → Administrator
    e. Software Developer, Database Developer, Web Developer, CRM Technical Developer, Applications Developer, Mobile Applications Developer → Developer
    f. Network Security Engineer, Network Engineer, Software Engineer, Software Systems Engineer, Technical Engineer → Engineer

   g. Technical Support, Technical Services/Help Desk/Tech Support → Support
   h. Solutions Architect, Data Architect → Architect
   i. Quality Assurance Associate, Information Technology Auditor, Software Quality Assurance (QA) / Testing → Quality Assurance

5. The numerical data is scaled using Standard Scaler
6. The textual data is encoded using label encoding for columns with 2/3 unique values and one hot encoding for columns with more unique values

Finally we have 75 input columns. 1 target attribute with 9 classes.

Model

Various number of hidden layers, activation functions and neuron numbers were tested, and the best results were given by the following model.

3 hidden layers with 32, 64, 16 neurons, and 1 output layer with 9 neurons.

The hidden layers use the activation function ‘tanh’ and weights are initialized with uniform values.
The output layer uses the ‘softmax’ function which outputs 9 values, that are the probabilities of the sample belonging to each class.

Categorical cross-entropy is used as the loss function and the model trains on improving accuracy.

Training

Various number of epochs and batch sizes were tested and finally a hybrid approach was chosen.
Training is done in 4 stages with changing batch size and number of epochs

	batch size	epochs
1.	256	64
2.	128	32
3.	64	16
4.	32	8

Different training methods:

1. 70:30
   Training accuracy - 39.45 %
   Testing accuracy - 15.83 %

2. 60:40
   Training accuracy - 42.08 %
   Testing accuracy - 15.39 %

3. 90:10
   Training accuracy - 35.68 %
   Testing accuracy - 14.75 %

4. No shuffling (70:30)
   Training accuracy - 38.86 %
   Testing accuracy - 15.48 %

5. Only numerical data (70:30)
   Training accuracy - 31.39 %
   Testing accuracy - 16.60 %

Analysis

None of the changes to data choice and train test split leads to good accuracy. This can mean a few things

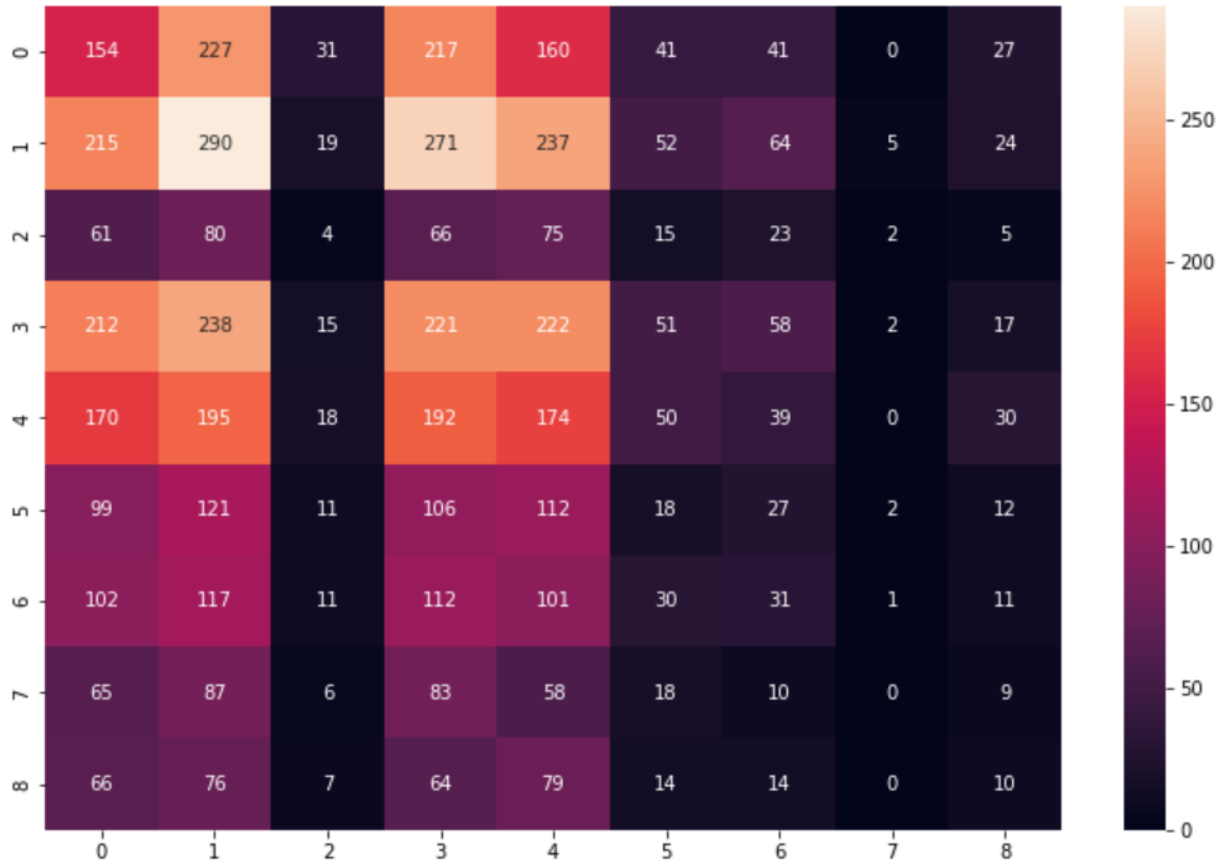
- We have not tried enough different scenarios for data modifications
- The roles are not grouped in a manner that allows accurate learning based on ML techniques
- The data does not lend itself to ML learning.

Using only numerical data leads to lower training accuracy but slightly higher testing accuracy, which suggests that using the full dataset is leading to overfitting on the training set.
But when the number of epochs was reduced in other cases it did not lead to noticeably better results, which disproves the above hypothesis.

70:30 split - Accuracy, precision, recall, f1-score

	precision	recall	f1-score	support
Administrator	.13	.17	.15	898
Analyst	.20	.25	.22	1177
Architect	.03	.01	.02	331
Developer	.17	.21	.19	1036
Engineer	.14	.20	.17	868
Manager	.06	.04	.05	508
Quality Assurance	.10	.06	.08	516
Support	.00	.00	.00	336
UX & Design	.07	.03	.04	330
accuracy			.15	6000
macro avg	.10	.11	.10	6000
weighted avg	.13	.15	.14	6000

Confusion matrix



Class-wise accuracy

Administrator :	0.1714922048997773
Analyst :	0.2463891248937978
Architect :	0.012084592145015106
Developer :	0.2133204633204633
Engineer :	0.20046082949308755
Manager :	0.03543307086614173
Quality Assurance :	0.060077519379844964
Support :	0.0
UX & Design :	0.030303030303030304