Employee Satisfaction Prediction for Job Postings in Different Areas

Project

CSE 4237

Soft Computing

Submitted by

Ahmad Subaktagin Jabir	**160204061**
Aniqua Tabassum	**160204085**
Noushin Tabassum	**160204114**

Submitted to

Mr. Mir Tafseer Nayeem

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Definition

Most of the times, job postings outside of Dhaka turn out to be dissatisfactory for the employees, which leads to frequent employee turnover. This is a major problem, especially for rural areas where, suppose, doctors are allocated but due to discontentment, they leave the allocated area the first chance they get.

The model we are developing is trained to predict the satisfaction of a worker, on a scale of 1 to 5, in a given location. We take into account various factors that may influence the satisfaction of workers in a certain allocation. This will help us find a more suitable area for the workers where they will be more content, which in return will lower the workplace turnover rate, bringing more stability to the company.

However, this model, alone, will not be able to solve the allocation problem. Since if allocation is done based on this model, it is highly possible that most of the workers will be concentrated around the big cities, such as Dhaka or Chittagong in the case of Bangladesh, which is not a desired outcome, as we want the employees to be equally dispersed around the entire country. So, followed by this model, a genetic algorithm must be used in order to find the optimum allocation that ensures maximum dispersion of the employees around the country. However, the dispersion is out of the scope of this project, as it does not require Neural Networks.

In 1.1, 1.2 and 1.3, we show how the allocation is before applying the satisfaction prediction model, what happens after applying the model and what happens after maximizing dispersion (out of the scope of our project)

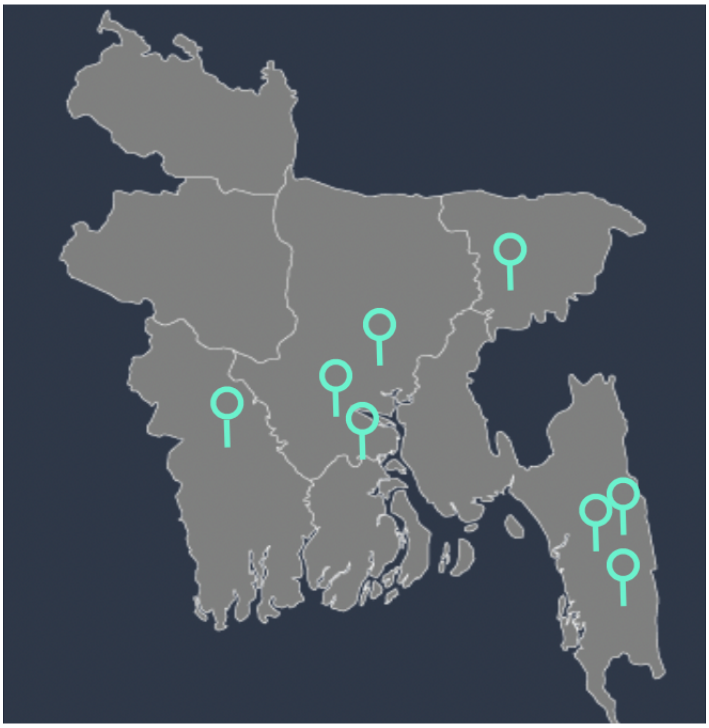


Figure 1.1: Allocation Before Applying the Model

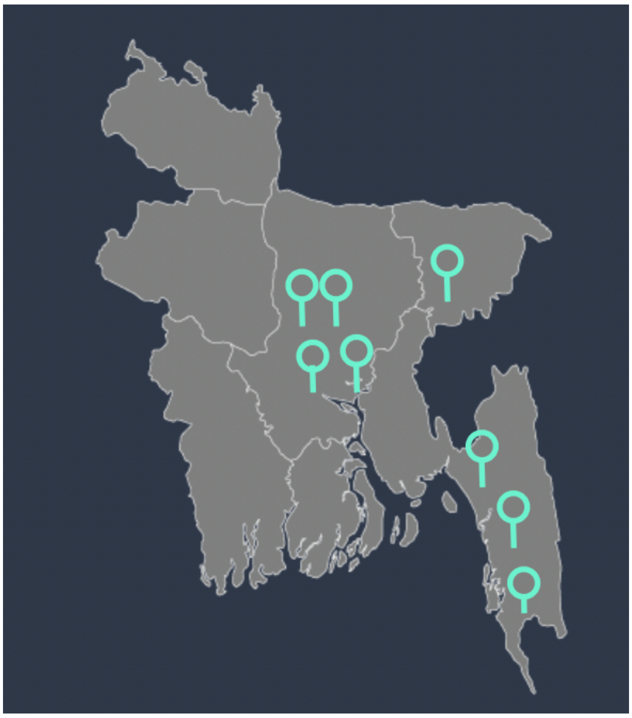


Figure 1.2: Allocation After Applying the Model

Figure 1.3: Allocation After Optimizing Dispersion (Out of Our Scope)

1.2 Motivation

Worker allocation in different locations in an optimal manner is a complicated problem for every kind of profession. Non-optimal worker-allocation leads to dissatisfaction in workers, which lowers productivity and causes dysfunction and loss of time, effort as well as monetary loss in working places. It also leads to discontentment of the employees who are designated to work in the rural areas as they feel as though they are receiving fewer benefits compared to the employees in the more developed areas. These problems are more specifically observed in the context of professionals such as Doctors, who are often forced to work in rural and often remote areas. Owing to extreme dissatisfaction, a high rate of turnover is seen at these places as most people designated to such areas are extremely eager to move away as quickly as possible. Due to such a high rate of turnover, an unstable work environment is born. Such an environment is undesirable to both the workers as well as the employers.

The importance of designing such a model, is to make sure that rural areas are getting stability from such important professionals, such as doctors. The solution to this problem is a two fold one, the first one being designing the neural network model to predict satisfaction, and the second one is to apply a dispersion function to attain optimum allocation. This model is expected to take us one step further to stabilize the workers in remote areas.

1.3 Challenges

There are a couple of challenges needing to be overcomed to solve the mentioned problem. They are mentioned below:

- **Collecting dataset:** There is no available dataset that we can use to solve this problem, so we had to collect a dataset on our own. This has been even more challenging during the Covid19 pandemic, as ideally, along with online surveys, we would be going to offices and hospitals in person to conduct our surveys. But on account of the pandemic restrictions, we had to solely depend on online circulation of our survey.

- **Improving accuracy:** This model is highly dependent on human psychology. There is no one way of predicting what will make a person content, as happiness is yet to be formulated through an equation, and there are more to play than the factors we have chosen to improve satisfaction. Each person has different thoughts, and it is especially challenging to quantify everyone's satisfaction with one model, that too created by such novice learners as us. So, the accuracy of the model is not too high.

Chapter 2

Related Works

2.1 List of Related Works

A list of similar works are mentioned below:

- A MATHEMATICAL MODEL TO MEASURE CUSTOMER SATISFACTION: Alexander C. Pereira.
- Modeling Employee Satisfaction in Relation to CSR Practices and Attraction and Retention of Top Talent: Simona Vinerean, Iuliana Cetina, Luigi Dumitrescu.

2.2 Summary of Previous Works

- To measure customer satisfaction, a Satisfaction Function (SF) has been proposed which is based on the customers' attitude regarding their attitude regarding any of the products the company is offering. The function is defined by a ratio of the customers' demands and the demands that have been met. [1]
- One of the topics discussed in [2] is emphasizing more on employee satisfaction as it is vital for attracting and retaining skilled employees to companies. They have proposed to make the work environment suitable enough so that the employees can have more decision making power which will give them a sense of fulfillment, to increase accountability, to treat jobs as products: meaning to bring in shape the jobs such that the employees will value those more and to increase sustainable practices. Their study is based on surveys from 10 multinational companies.

2.3 Our Approach

A list is given below highlighting what we are planning to do to address what these papers have missed:

- [1] does not address employee satisfaction, rather they are more concerned with customer satisfaction. Inversely, we are working the other way around.
- The Satisfaction Function (SF) proposed by [1] is, in higher level, a ratio between demands and the number of met demands. Whereas, we are planning on taking different features of a person and their work environment and training a neural network to predict employee satisfaction.
- [2] does not propose any mathematical or statistical model to increase employee satisfaction. They have taken a different approach, based on their surveys, which is to change the office environment and culture. Conversely, by using our model, allocation to areas or offices will be based on the features of the designated areas and what can be more desirable to the employees.

Chapter 3

Project Objectives

3.1 Tasks

We have divided our task to create the model into these following sub-tasks. First we are showing a flowchart of the tasks and then the descriptions will be provided.

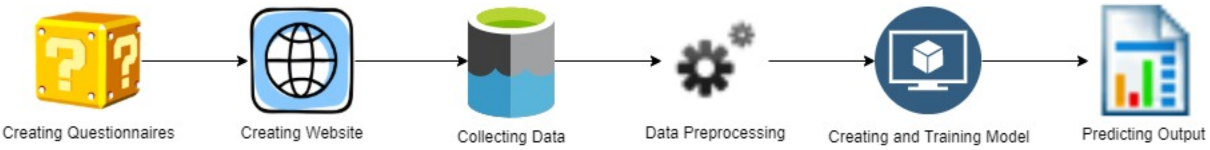


Figure 3.1: Flowchart of Tasks

3.1.1 Creating Questionnaires

Our questionnaires are comprised of questions regarding the following:

- First we ask some demographic questions, ie their age, gender, occupation, field of study, marital status, occupation of spouse and willingness of their spouse's moving in case the respondent has to move to a different city for their jobs.
- We ask questions about the 5 factors we think may have an impact on employee allocation, which are: marital status, security, schooling, house rent and distance from hometown of the designated area.
- Among the five satisfaction factors, the "marital status" one has been kept as a binary demographic question, while the rest 4 have been used to form different virtual scenarios for the respondents. These 4 factors have 3 levels: low, medium and high. The concept explained in the previous paragraph can be further clarified by 3.2

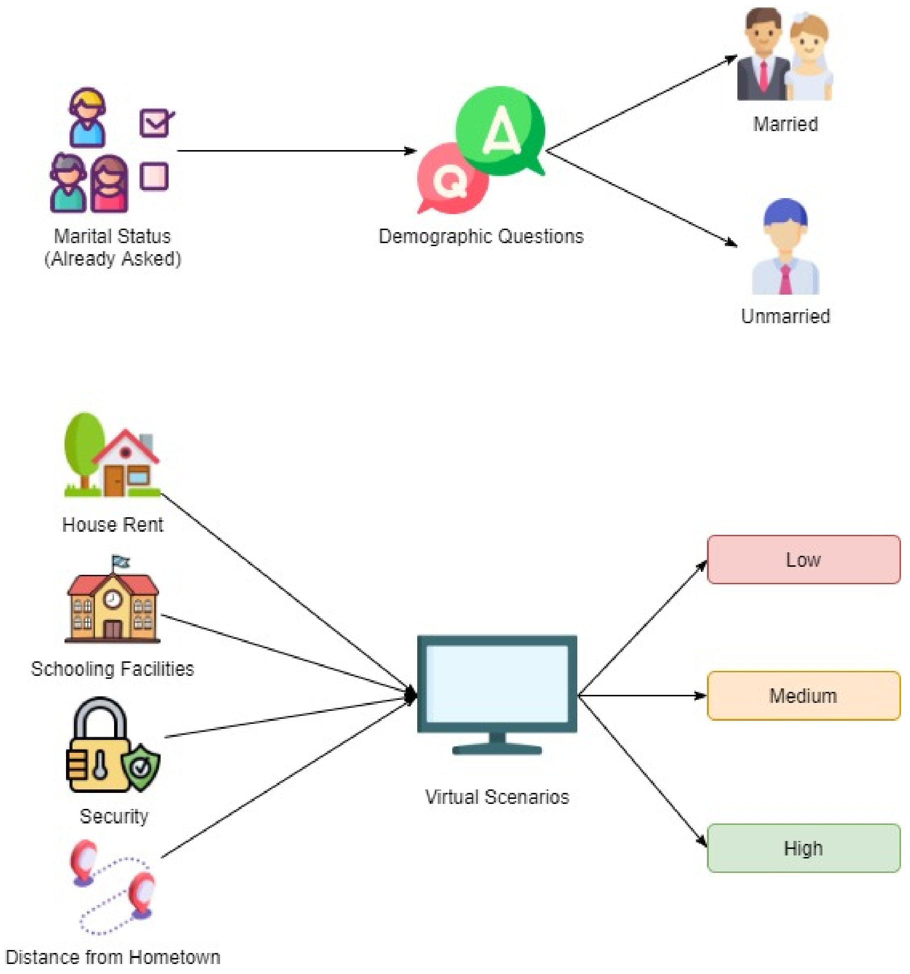
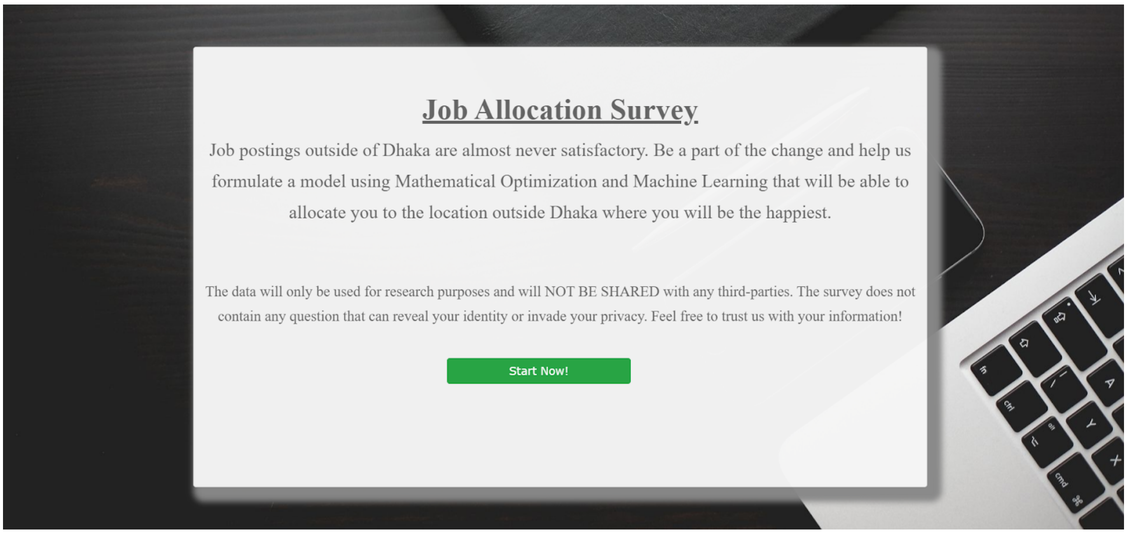


Figure 3.2: Formation of Virtual Scenarios

- So there are a total of 3^4 or 81 possible scenarios. It was not practical to ask each respondent about all of these 81 virtual scenarios, so we had to carefully partition the set of 81 scenarios to smaller sets. So, we made 27 disjoint sets of questionnaires and it was absolutely vital that these 27 sets are distributed evenly among the respondents.

3.1.2 Creating Survey Collection Website:

Since it was important for us to make sure that each of the 27 sets are being circulated evenly, we had to create a website of our own. So, we made a website using the MVC framework of ASP.NET and used an algorithm that would ensure that all of the sets are being distributed evenly. Some important pages of the website have been shown below:



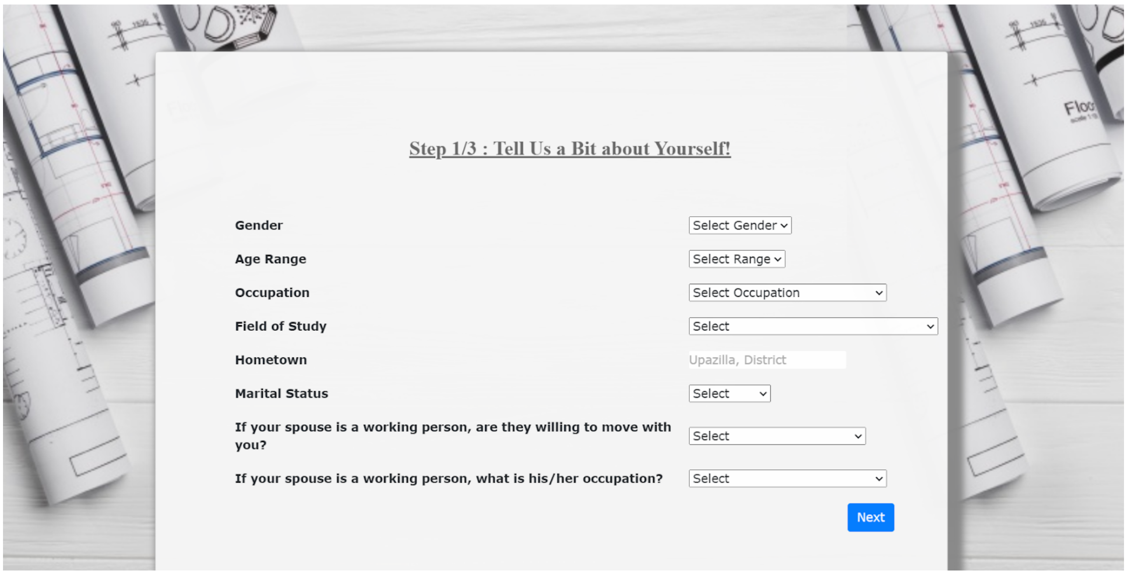
Job Allocation Survey

Job postings outside of Dhaka are almost never satisfactory. Be a part of the change and help us formulate a model using Mathematical Optimization and Machine Learning that will be able to allocate you to the location outside Dhaka where you will be the happiest.

The data will only be used for research purposes and will NOT BE SHARED with any third-parties. The survey does not contain any question that can reveal your identity or invade your privacy. Feel free to trust us with your information!

Figure 3.3: Home Page

Next we asked some demographic questions.



Step 1/3 : Tell Us a Bit about Yourself!

Gender	Select Gender
Age Range	Select Range
Occupation	Select Occupation
Field of Study	Select
Hometown	Upazilla, District
Marital Status	Select
If your spouse is a working person, are they willing to move with you?	Select
If your spouse is a working person, what is his/her occupation?	Select

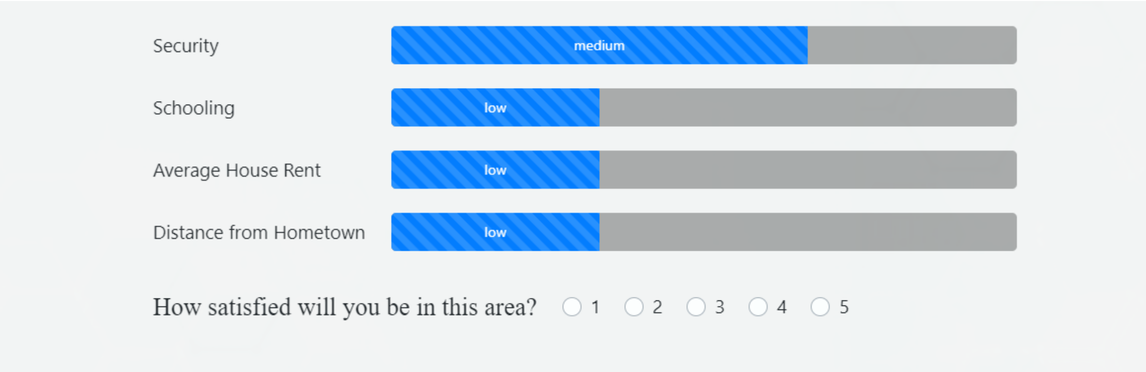
Figure 3.4: Demographic Questions

Finally, we gave the respondent 3 virtual scenarios, and asked them to rate their satisfaction on a scale of 1 to 5 under each of the three circumstances. A demo scenario is shown in this figure. Here we are telling the respondent that he will be allocated to an area that has the following features,

1. Security is Medium;

2. Schooling, Average House Rent and Distance from Hometown is Low.

We then asked him to rate his level of satisfaction if he had to live in the said area, on a scale of 1 to 5. Each respondent is presented with three such scenarios.



Security: medium

Schooling: low

Average House Rent: low

Distance from Hometown: low

How satisfied will you be in this area? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 3.5: Sample Virtual Scenario

3.1.3 Collecting Data

The next step is to collect data. We have circulated the website for about a month, and have collected 855 data points in total.

3.1.4 Data Preprocessing

Following data collection, we had to apply some pre-processing on our data. We did standardization and normalization on our dataset prior to training.

3.1.5 Creating and Training Model

Next step is to create and training neural network models using some trial and errors by tuning the hyperparameters.

3.1.6 Predicting Outputs

Finally, we can predict the outputs using the most accurate model we have made.

3.2 Dummy Input and Output

In the table below, we have prepared some dummy inputs and outputs for our model.

Our Input parameters are of two types, as following:

- **Person Specific:** Gender, Age Range, Occupation, Field of Education, Marital Status, Willingness of Their Spouses to Move with Them for a Job Posting, Spouse's Field of

Education, Spouse's Occupation.

- **Designated Area Specific:** Schooling, House Rent, Security and Distance from Hometown

Our output is the satisfaction of the employee in the said area, on the scale of 1 to 5.

We have first numerically encoded the input parameters. The mapping among actual inputs and their corresponding numerical values are presented in the following tables:

Table 3.1: Mapping of Age Range

Age Range	**Numeric Value**
20 - 25	20
26 - 30	26
31 - 35	31
36 - 40	36
41 or above	41

Table 3.2: Mapping of Gender

Gender	**Numeric Value**
Male	1
Female	2
Prefer to not disclose	3

Table 3.3: Mapping of Occupation

Occupation	**Numeric Value**
Medical Field	1
Engineering and IT	2
Business Field	3
Academia	4
Student	5
Unemployed	6
Others	7

Table 3.4: Mapping of Field of Education

Field of Education	**Numeric Value**
Medical, Biological or Chemical studies	1
Engineering and IT	2
Business Field	3
Social Studies	4
Other	5

Table 3.5: Mapping of Marital Status

Marital Status	**Numeric Value**
Married	1
Unmarried	2

Table 3.6: Mapping of Willingness of the Respondent's Spouse's Moving with Them

Spouse Willing	**Numeric Value**
Yes	1
No	2
My Spouse Does Not Work	3
I am not Married	4

Table 3.7: Mapping of Spouse Occupation

Spouse Occupation	**Numeric Value**
Medical Field (Doctor, Nurse, Nutritionist, Pharmacists and other Health Care workers etc)	1
Engineering and IT	2
Business Field (Management, HR, Banking, Marketing etc)	3
Academia (Teacher, Lecturer, Assistant/ Associate Professor, Professor)	4
Student	5
Unemployed	6
Other	7
I am not Married	8

Table 3.8: Mapping of Security of designated area

Security	**Numeric Value**
Low	1
Medium	2
High	3

Table 3.9: Mapping of Schooling Facilities of designated area

School	**Numeric Value**
Low	1
Medium	2
High	3

Table 3.10: Mapping of Rent of designated area

Rent	**Numeric Value**
Low	1
Medium	2
High	3

Table 3.11: Mapping of Distance from Hometown from designated area

Distance	Numeric Value
Low	1
Medium	2
High	3

A number of dummy input outputs are presented below:

Table 3.12: Mapping of Features with alphabet

Name of Feature	Mapped Alphabet
Age range	a
Gender	b
Occupation	c
Field of Education	d
Marital Status	e
Spouse Willing	f
Spouse Occupation	g
Security	h
School	i
Rent	j
Distance	k

Table 3.13: Dummy Inputs and Outputs

Input											Output
a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	
20	1	5	2	2	4	8	2	2	3	1	3
36	2	4	3	1	1	3	3	1	1	2	3
26	1	5	5	2	4	8	1	2	3	2	2
20	1	2	2	2	4	8	2	1	1	1	1
26	1	3	2	1	3	6	3	2	3	2	5

Chapter 4

Methodologies

4.1 How are we solving the problem

We are planning on solving this problem using Deep Neural Network. We will be experimenting with different hyperparameter such as learning rate, batch size, number of neurons per layer, optimizers, activation functions etc. The inputs of our function are shown below:

Table 4.1: Input Variables

Input number	Input
1	Gender of person
2	Age of person
3	Marital status of person
4	If the spouse of this person will move with him/ her
5	Occupation of person
6	Security of designated area
7	Schooling of designated area
8	House rent of designated area
9	Distance from hometown from the designated area

Output will be the satisfaction of the employee of the said area from 1-5. A table demonstrating the outputs are shown below.

Table 4.2: Output Variables

Output Number	Output
1	1
2	2
3	3
4	4
5	5

Description and diagram of the model we are using are explained in section 4.2.

4.2 Diagrams

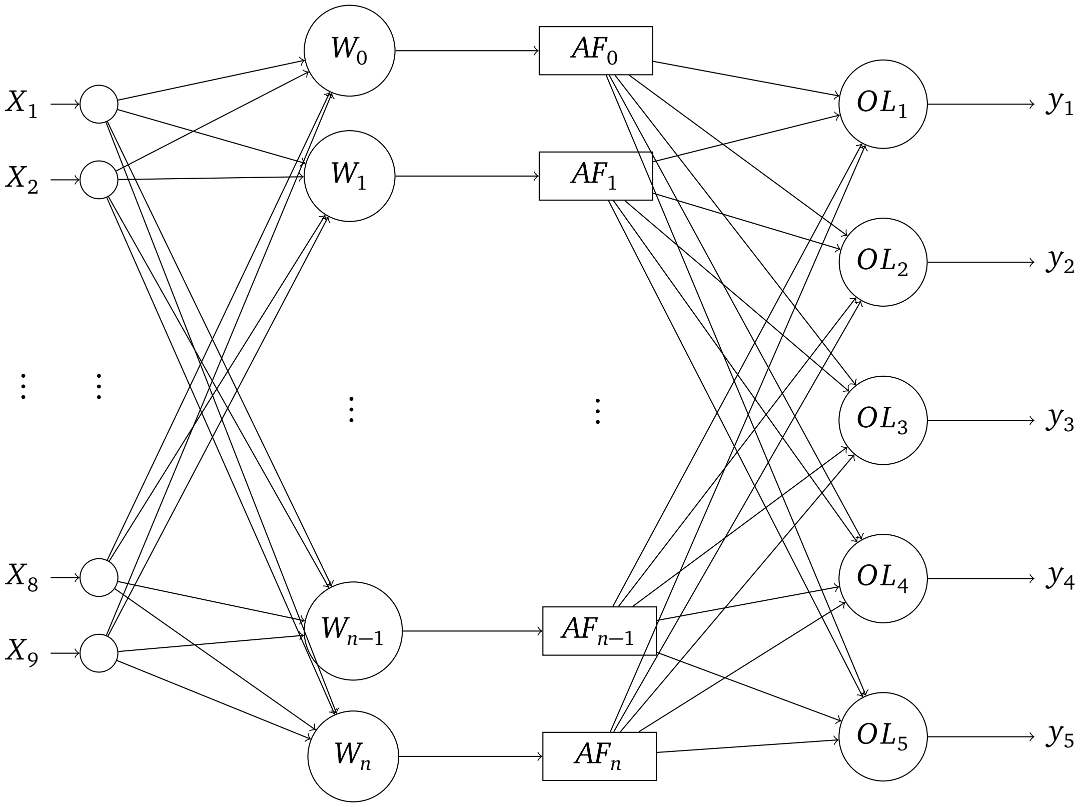


Figure 4.1: Sample Neural Network

Here, the inputs are donated by x_1 to x_9. There can be n number of neurons per hidden layer, where n can be any positive integer. However, since the dataset we are using is small, so the number of neurons per hidden layers should not be many to avoid risk of overfitting. Here, AF_n refers to activation functions, which can be ReLU, SELU, GELU etc. For our experiment, we have experimented with ReLU, SELU, GELU and ReLU6. OL refers to output layers. There are 5 OLs since we have 5 outputs. y_n refers to predictions of the neural network.

Chapter 5

Experiments

5.1 Dataset

We have a total of 855 data points, which consists of 5 classes in total, which are satisfaction rated from 1-5. The statistics are shown below:

Table 5.1: Statistics of Classes

Class	Number of Samples
1	187
2	159
3	220
4	157
5	132

Bar Chart of the Statistics of Classes is provided below:

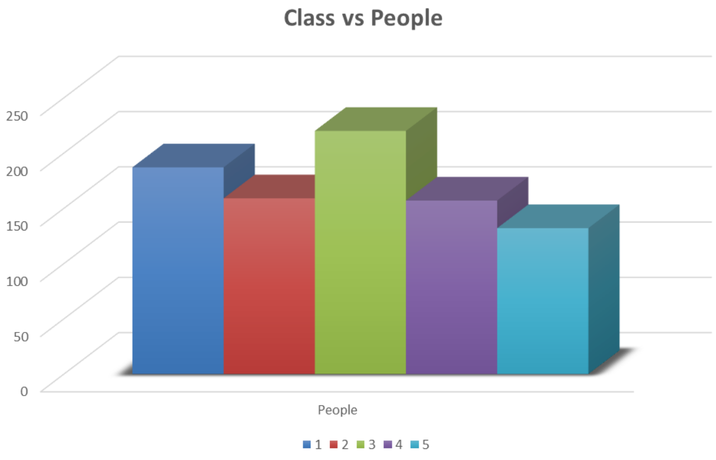
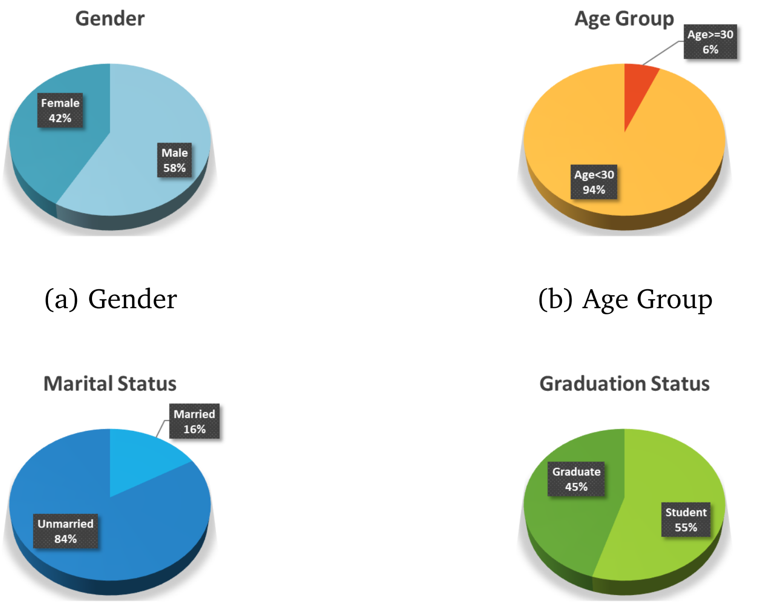


Figure 5.1: Bar Chart of the Statistics of Classes

Pie Charts of the demographics and Bar Chart based on their professions are provided below:



(a) Gender

(b) Age Group

(c) Marital Status

(d) Graduation Status

Figure 5.2: Pie Chart of Demographics

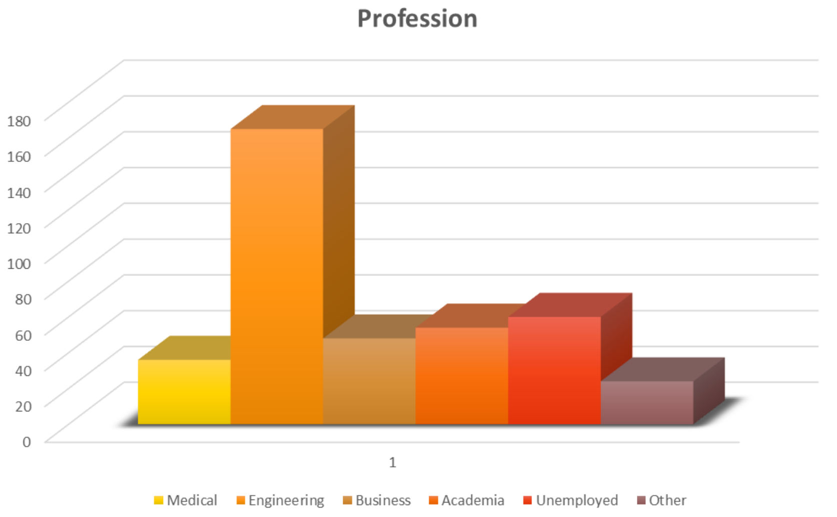


Figure 5.3: Bar Chart Based on Profession

5.1.1 Samples from Dataset

Table 5.2: Samples from the Dataset

Input											Output
a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	
20	2	5	2	2	4	8	3	2	1	3	5
26	1	2	2	2	4	8	3	3	3	3	3
31	2	4	4	1	3	6	1	2	1	3	3
20	1	5	2	2	4	8	3	1	3	1	4
41	2	6	3	1	2	3	1	3	1	1	1

5.1.2 Train, Cross Valid, Test Split

Ratio of training data : Cross Validation Data : Testing Data for our model is 60:20:20.

5.2 Evaluation Metric

- **How will we evaluate:** We will be using accuracy and loss to measure how well our model is performing.
- **List of evaluation metrics:**
  - **Loss:** For calculating loss, we are using cross entropy loss. Cross entropy loss provides a probability between 0 and 1 that indicates how close the predicted value is to the actual value. Increase of this loss value, also known as log loss, is proportionate to the divergence between actual label and predicted label. It can be denoted as following:

$$L(\hat{y}^{(i)}, y^{(i)}) = -(y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}))$$

    1. If $y^{(i)} = 1 : L(\hat{y}^{(i)}, y^{(i)}) = -\log(\hat{y}^{(i)})$ where $\log(\hat{y}^{(i)})$ and $\hat{y}^{(i)}$ should be close to 1.
    2. If $y^{(i)} = 0 : L(\hat{y}^{(i)}, y^{(i)}) = -\log(1 - \hat{y}^{(i)})$ where $\log(1 - \hat{y}^{(i)})$ and $\hat{y}^{(i)}$ should be close to 0.

  - **Accuracy:** Accuracy of the model will be predicted as following:

$$\text{Accuracy} = \frac{\text{Correctly Classified Samples}}{\text{Total Samples}} \times 100$$

5.3 Results

In table 5.3, we show the predicted output and actual actual for different scenarios. Here, scenarios refer to different inputs.

Table 5.3: Predicted and Actual Outputs

Scenario	Predicted Output	Actual Output
1	1	2
2	3	5
3	1	1
4	3	1
5	1	3

Here, the neural network that we are using is a 5-layer ReLU, that has 8 neurons per hidden layer. We took a batch size of 16 and did 10,000 iterations with 0.0001 learning rate. Since, as mentioned earlier, this problem is trying to capture human psychology, there is a lot of noise in the dataset itself, which has resulted in lower accuracy values. Accuracy on cross validation set for this setting is 32.16%, and for test set it is 35.08%. The accuracy values for cross validation dataset in different hyperparameter settings are shown in Figure 5.4. The setting we have chosen for our result belongs to Setting B of Figure 5.4.

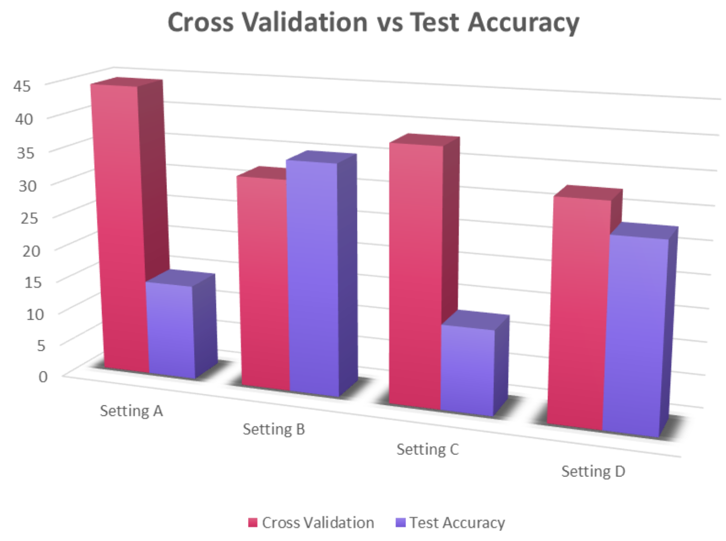


Figure 5.4: Bar Chart of Cross Validation vs Test Accuracy

Table 5.4: Accuracy Table with Different Settings

Batch Size	No. of Iterations	Optimizer	Activation Function	No. of Layers	No. of Neurons per Layer	Learning Rate	Best Accuracy
16	20,000	Adam	ReLU	5	16	0.001	43.86%
16	20, 000	Adam	ReLU	5	16	0.0001	44.44%
16	20, 000	Adam	ReLU	5	16	0.00001	32.16%
16	20, 000	Adam	ReLU	5	16	0.01	42.11%
16	20, 000	Adam	ReLU	7	16	0.0001	40.94%
16	20, 000	Adam	SeLU	5	16	0.0001	42.11%
16	20, 000	Adam	GeLU	5	16	0.0001	43.86%
16	20, 000	Adam	GeLU	7	16	0.0001	42.11%
16	20, 000	Adam	ReLU6	3	16	0.0001	42.11%
16	20, 000	Adam	ReLU6	5	16	0.0001	42.69%
16	20, 000	SGD	SeLU	5	16	0.0001	26.32%
16	20, 000	Adargrad	SeLU	5	16	0.0001	24.56%
16	20, 000	Adam	ReLU	5	8	0.0001	40.94%
16	40, 000	Adam	ReLU	7	8	0.0001	43.27%
8	20, 000	Adam	ReLU	7	8	0.0001	40.94%
16	20, 000	Adam	ReLU	7	32	0.0001	44.44%
16	15, 000	Adam	ReLU	7	8	0.0001	44.44%
16	10, 000	Adam	ReLU	5	8	0.0001	32.16%
16	10, 000	Adam	ReLU	3	8	0.0001	38.60%
16	10, 000	Adam	ReLU	3	4	0.0001	32.75%

The observation from training our model with different hyperparameter is that, how easily the model was being oveefitted. The dataset being quite small, a relatively medium neural network with just 8 neurons per hidden layer and 7 hidden layers was leading us towards overfitting. We understood that the model was being overfitted when, for some of the shown settings, the accuracy on cross validation set was around 44%, which is quite high given the complexity of our problem, but for the testing dataset, it went down to as low as 12%.

Another observation was that, without considering overfitting, on cross validation sets, minor changes in the hyperparameters did not effect the training accuracy much, with some exceptions. Such as, for the first row in table 5.4, we can see that the best accuracy was 43.86%, but if we change the learning rate to 0.0001, then the accuracy becomes 44.44%, which is not much of a change. Of course, as shown in table 5.4, some hyperparameter settings have led to drastic decrease in the accuracy. For our final result, mentioned earlier, we have chosen the second to last hyperparameter setting in table 5.4, which performs the best on the test set.

Chapter 6

Conclusion

We have completed this project to our best capability with the resources we had in our hands, but being apprentices, there are lots of scopes of improvements. We have envisioned some improvements that we are planning on executing in the future. They are mentioned below:

- **Data Collection:** Deep Neural Networks perform better with more data, but due to the limitations posed upon us due to the pandemic, we could not physically collect our dataset. We are planning on conducting in-person survey as well as online ones to enrich our dataset with more data.
- **Improved Data Pre-processing:** We are planning on mastering more data cleaning and preprocessing techniques in order to curate our dataset even further.
- **Creating a Better Model:** As we gain more and more experience in deep learning, we will be able to make more knowledgeable decisions about how to create our model more suitable for solving this problem.

We are hoping that our mistakes will be seen in the eyes of forgiveness due to our lack of experience and enthusiasm towards learning.

References

[1] A. C. Pereira, “A mathematical model to measure customer satisfaction,” *Quality Engineering*, vol. 11, no. 2, pp. 281–286, 1998.

[2] S. Vinerean, I. Cetina, and L. Dumitrescu, “Modeling employee satisfaction in relation to csr practices and attraction and retention of top talent,” *Expert Journal of Business and Management*, vol. 1, no. 1, 2013.