Project Part 2

Chakrika Vemireddy

Mis 432

Toni Garcia

5/6/25

**Introduction**

The dataset used for the project analysis is found from the website Kaggle, which is a popular data science platform that contains datasets and hosts competitions for other data scientists. It allows for them users to communicate with each other as well as to look at each other’s code. The dataset is data that displays whether a candidate has received a job offer based on several attributes that contribute to the result. There are both continuous and categorical variables in the original dataset. Two neural network models are used on the preprocessed data to predict whether or not a candidate got the job.

**Explanatory Variables – Original Data Information**

Serial Number: numerical ID number that is used for each applicant.

Gender: A categorical variable indicating the gender

Python\_exp: A categorical variable that shows whether or not the applicant has experience with the Python language.

Experience\_Years: A numeric variable representing the number of work experience in years.

Education: A categorical variable indicating whether they graduated from college.

Internship: A categorical variable that indicates whether or not the applicant had an internship prior.

Score: A numeric variable that appears a test score outcome.

Offer\_History: A numeric variable that shows if the applicant had a offer previously.

Location: A categorical variable displaying where the applicant lives.

Salary: A numeric variable indicating the applicant’s expected or offered salary.

**Response Variable** - **Original Data Information**

Recruitment\_Status: This is a categorical feature that is the target variable, and states whether or not the person was hired.

**Preprocessing Steps**

Before fitting the data into the neural network models that original data went through some preprocessing steps to ensure that the models would predict the best results possible. First, missing values were handled, the variables Internship and Gender had some missing entries, which were then filled using its common output. No was filled in for Internship and Male was filled in for Gender. Next, numerical variables Experience\_Years, Score, and Offer\_History, were scaled between 0 and 1. This was necessary to ensure that variables did have any bias influence on the neural network models more than any other feature.

Categorical variables such as Gender, Python\_exp, Internship, Education, and Location were converted into dummy variables to allow to make sure the neural network would be able to read them. The target variable Recruitment\_Status was also converted into a binary numeric format and renamed as Target, where 1 means the applicant was hired and 0 represents the applicant who was not hired. Finally, the Salary variable was removed during preprocessing because it contains a lot of missing values and was in scientific notation making it difficult to clean properly and reliably. Another reason that salary was dropped is due to Salary not mattering until after a person was hired or not which is the target variable. The preprocessed dataset was then saved as a CSV file.

**Model 1: 1 Hidden Layer with 3 Nodes**

The first neural network model was built using one hidden layer with three nodes. The model was then trained on 60% of the preprocessed dataset. The input variables were ExperienceS, ScoreS, OfferS, GenderM, PythonExpY, InternshipY, Edu\_Grad, and Loc\_Urban due to these all being the explanatory variables that make an impact on the Target Variable.

A diagram of a data flow

AI-generated content may be incorrect.

The model based on the confusion matrix seemed to have performed well. The model was evaluated on the 40% validation dataset. Predictions were converted to labels using a threshold of 0.5. Anything over 0.5 was considered 1 which indicated hired while anything under 0.5 was considered 0 and indicated not hired.

This model had a relatively high accuracy at 76.53% meaning it was effective at identifying applicants who were hired but struggled more with identifying applicants who were not hired. However, the model did have strong accurate results which indicate that it is still a reliable model to use.

A screenshot of a computer

AI-generated content may be incorrect.

**Model 2: 2 Hidden Layer with 2 Nodes**

The second neural network model used two hidden layers, each with two nodes. The same input variables from Model 1 were used for model 2, ExperienceS, ScoreS, OfferS, GenderM, PythonExpY, InternshipY, Edu\_Grad, and Loc\_Urban. This model was also trained on 60% of the dataset and the predictions were done on the remaining 40% validation set.

A screenshot of a computer

AI-generated content may be incorrect.

Like before, the predictions were converted to labels using a threshold of 0.5. Any prediction greater than 0.5 was classified as 1 and anything under 0.5 was classified as 0. 1 indicated hired while 0 indicated not hired. The accuracy was 79.81% for this model, which shows the model made more correct predictions. Similar to Model 1, model 2 is better at predicting people who are hired rather than those who were not hired. Model 2 has better predictability than model 1.

A screenshot of a computer

AI-generated content may be incorrect.

**Model Comparison**

Model 1 had one hidden layer with three nodes, while Model 2 used two hidden layers with two nodes each. Model 2 had a higher accuracy of 79.81% compared to Model 1’s accuracy of 76.53%. This means that model 2 performed better than model 1. Based on the prediction results, both Model 1 and Model 2 are better at predicting applicants who were hired than applicants that weren’t hired but overall based on accuracy results Model 2 is better than Model 1.

**Conclusion**

After preprocessing the dataset and creating two different neural network models, it was shown that both models performed relatively well with high accuracy score but the model with two hidden layers performed better. This indicates that complexity within the neural network model is better for this dataset. However, there is no guarantee that two hidden layers with two nodes is the best till more versions of the neural network are tested and run. Part of the reason that the two models had good accuracy results were due to the preprocessing steps taken when cleaning the data before the model fitting. The scaling features made sure that all the features were the same impact on the model and didn’t add any unnecessary bias. Overall, being a senior graduating soon, this dataset helped me get the hands on experience for neural networks while also looking and exploring a dataset that will be beneficial in my future job search.

**Source**

https://www.kaggle.com/datasets/rafunlearnhub/recruitment-data