Campus Placement Prediction Report

Student: Devansh Patel (C0928483)

Course: AML-3104 Neural Networks and Deep Learning

Submitted to: Ishant Gupta

Introduction

Predicting a student's likelihood of being placed on campus based on a variety of academic and demographic criteria is the aim of this project. Employability test scores, job experience, specialisation, degree %, secondary education percentage, and higher secondary education percentage are among the features included in the dataset. This report outlines the dataset, preprocessing steps, model selection, evaluation, and results.

DataSet Description

The dataset contains the following features:

- Numerical Features:
    - **ssc_p:** Secondary education percentage (10th grade).
    - **hsc_p:** Higher secondary education percentage (12th grade).
    - **degree_p:** Degree percentage (undergraduate).
    - **etest_p:** Employability test percentage.
    - **mba_p:** MBA percentage.
- **Categorical Features:**
    - **gender:** Gender of the student (Male/Female).
    - **ssc_b:** Board of education for secondary education (Central/Others).
    - **hsc_b:** Board of education for higher secondary education (Central/Others).
    - **hsc_s:** Stream in higher secondary education (Commerce/Science/Arts).
    - **degree_t:** Type of degree (Sci&Tech/Comm&Mgmt).

- **workex:** Work experience (Yes/No).
- **specialisation:** Postgraduate specialization (Mkt&HR/Mkt&Fin).

- **Target Variable:**
  - **status:** Placement status (Placed/Not Placed).

Preprocessing Steps

The following preprocessing steps were applied to prepare the dataset for modeling:

1. **Handling Missing Values:**
   - Since it was irrelevant for predicting placement status, the pay column—which included missing values for students who were not placed—was removed.
2. **Encoding Categorical Variables:**
   - To transform them into numerical format, categorical features including gender, ssc_b, hsc_b, hsc_s, degree_t, workex, and specialisation were one-hot encoded.
3. **Splitting the Dataset:**
   - To assess model performance, the dataset was divided into training (70%) and testing (30%) sets.
4. **Feature Scaling:**
   - To guarantee that every feature was on the same scale, numerical features were standardised.

Model Selection

Three models were chosen for this project:

1. **The Logistic Regression Model:**
   - An easy-to-understand paradigm that works well for binary classification tasks.

- It helps comprehend the connection between attributes and the goal variable and offers a baseline performance.

2. **Random Forest:**
   - An ensemble approach that manages feature interactions and non-linear relationships.
   - It performs well with both numerical and categorical data and is resistant to overfitting.

3. **XGBoost:**
   - a strong gradient boosting technique with a good accuracy rate.
   - It offers feature importance and manages unbalanced datasets effectively.

4. **Voting Classifier:**
   - XGBoost, Random Forest, and Logistic Regression combined with soft voting.
   - It enhances overall performance by utilising the advantages of all three models.

Model Evaluation

The following measures were used to assess the models:

- **Accuracy:** The percentage of placements that were accurately anticipated.

- **Precision:** The percentage of accurately predicted placements.

- **Recall:** The percentage of actual placements that were accurately anticipated is known as recall.

- **F1-Score:** The precision and recall harmonic mean.

Results:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**0.85**	**0.86**	**0.90**	**0.88**
Random Forest	**0.88**	**0.89**	**0.92**	**0.90**
XGBoost	**0.89**	**0.90**	**0.93**	**0.91**
Voting Classifier	**0.90**	**0.91**	**0.94**	**0.92**

Confusion Matrices:

- To visualise true positives, true negatives, false positives, and false negatives, confusion matrices were plotted for every model.

ROC Curves:

- The true positive rate (TPR) and false positive rate (FPR) for every model were compared using ROC curves. The Voting Classifier's AUC score was the highest.

Conclusion

- With an F1-score of 0.92 and an accuracy of 90%, the Voting Classifier outperformed the rest.
- Specialisation, etest_p, and degree_p are important variables that affect placement forecasts.
- Based on student information, the model can be used as a Streamlit app to forecast placement status.