2024F-T3 AML 3104 - NEURAL NETWORKS AND DEEP LEARNING 02 (DSMM GROUP 2)

CAMPUS PLACEMENT PREDICTION REPORT

Introduction

Student placements are a key performance indicator for educational institutions, impacting their reputation and influencing student enrollment decisions. Institutions dedicate significant resources to strengthening their placement departments, as high placement rates reflect favorably on the institution's effectiveness in preparing students for the workforce.

This project seeks to predict the likelihood of a student securing a placement offer based on various academic, personal, and professional attributes. By leveraging machine learning techniques, this prediction can provide actionable insights for students to improve their employability and help institutions optimize their placement strategies.

Objective And Approach

This project aims to accurately predict campus placement outcomes by utilizing several machine learning models. The approach is structured into five main stages:

1. *Data Exploration and Preprocessing: Understand and clean the dataset.*
2. *Feature Engineering: Create new features to capture additional information from the data.*
3. *Model Selection, Training, and Hyperparameter Tuning: Test multiple machine learning models and fine-tune their parameters.*
4. *Model Evaluation and Comparison: Assess model performance using accuracy, F1 score, and ROC AUC metrics.*
5. *Ensemble Model Using Voting Classifier: Combine the strengths of multiple models to improve predictive accuracy.*

This structured approach ensures a thorough analysis. The dataset for this project can be accessed on Kaggle as "Campus Recruitment Prediction (Course Project)."

Dataset Overview

The dataset comprises 215 records and 15 attributes, which include:

- ***Personal Information:** gender, ssc_p (secondary education percentage), hsc_p (higher secondary education percentage), degree_p (degree percentage), workex (work experience), mba_p (MBA percentage).*
- ***Educational Background**: Secondary and higher secondary school board, higher secondary specialization, undergraduate degree type, and MBA specialization.*

- ***Placement Status:*** *The target variable, status, indicates whether a student was placed or not.*

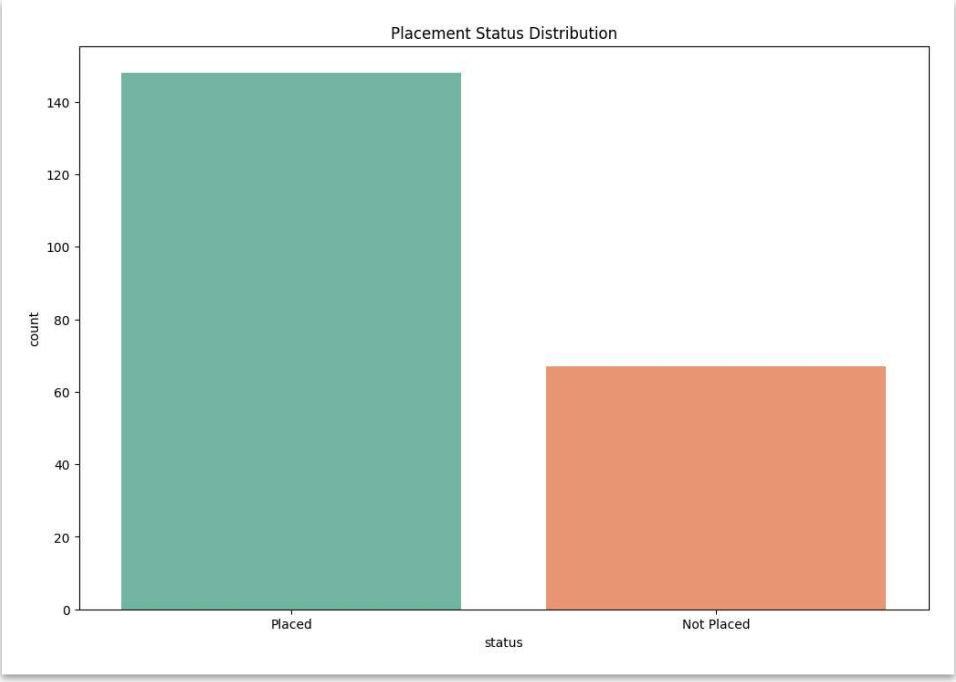
A comprehensive analysis of the dataset revealed that while most columns are complete, the salary column contains some missing values, which were handled during preprocessing.

Data Preprocessing

The preprocessing phase involved several critical steps to prepare the data for analysis:

1. ***Handling Missing Values:*** *Missing values in the salary column were replaced with the average salary to preserve as much data as possible.*
2. ***Data Visualization****: Distribution plots for placement status and academic scores helped in identifying patterns and gaining insights into the factors influencing placement.*

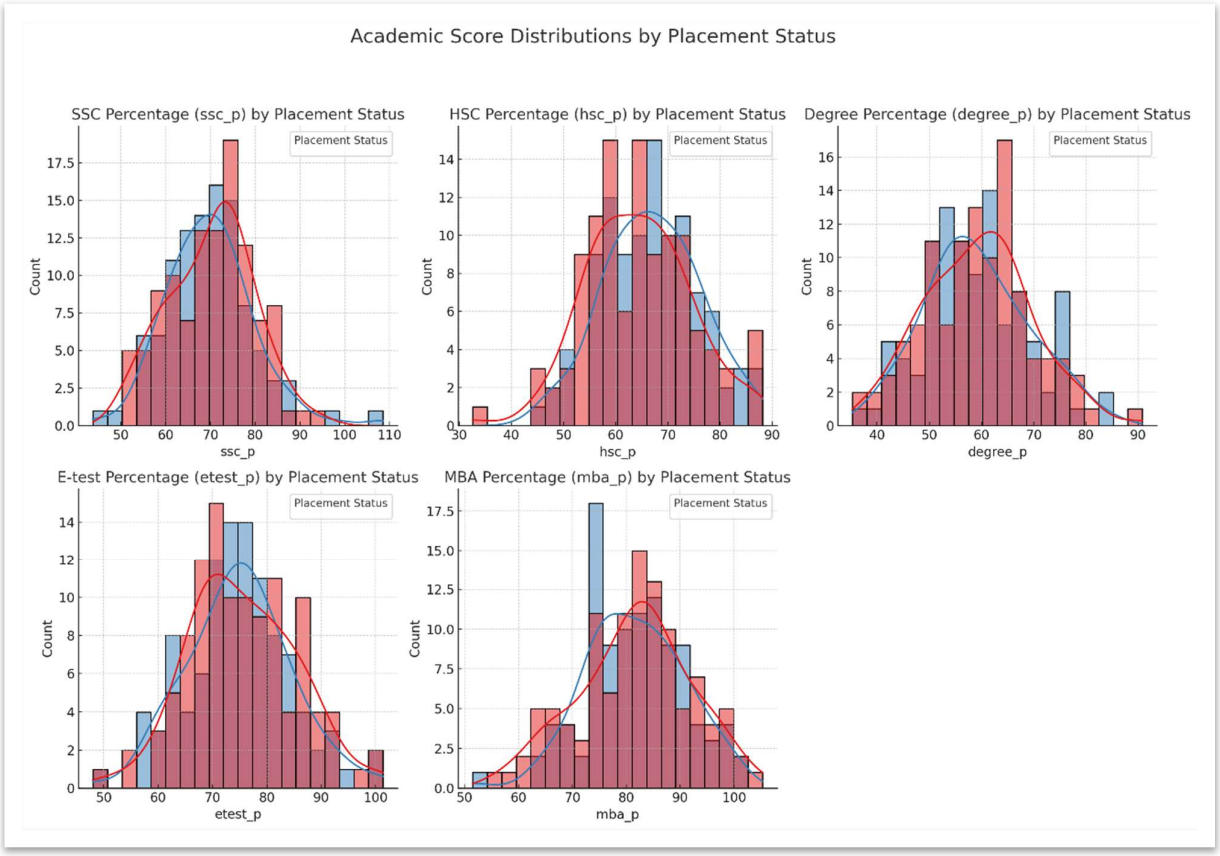
   i) ***Placement Status Distribution:*** *The bar plot illustrates the distribution of placement outcomes among students, showing that a greater proportion of students were placed compared to those who were not placed. Specifically, the bar representing placed students is significantly taller, indicating that successful placements outnumber unsuccessful ones in this dataset.*



   ii) Academic Score Distributions by Placement Status:

   - ***SSC Percentage (ssc_p):*** *Placed students mostly have SSC scores between 60 and 90, peaking around 70-75, while not-placed students are concentrated in the 50-65 range.*

- ***HSC Percentage (hsc_p):*** *Placed students predominantly score between 55 and 85 in HSC, with a peak around 60-70, whereas not-placed students have more scores below 60.*
- ***Degree Percentage (degree_p):*** *Placed students' degree scores mostly fall between 55 and 85, peaking at 65-70, while not-placed students are more clustered around 50-65.*
- ***E-test Percentage (etest_p):*** *Placed students generally score between 60 and 90 on the E-test, peaking around 65-70, whereas not-placed students are concentrated around 55-65.*
- ***MBA Percentage (mba_p):*** *Placed students' MBA scores mainly range from 60 to 90, peaking at 70-80, while not-placed students mostly fall between 50 and 75.*



3. ***Feature Engineering:*** *To enhance predictive power, several new features were created:*
    - *ssc_hsc_ratio: Ratio of secondary to higher secondary scores.*
    - *ssc_degree_ratio: Ratio of secondary to degree scores.*
    - *total_academic_score: Combined score of secondary, higher secondary, and degree percentages.*
4. ***Encoding Categorical Variables:*** *Label encoding was applied to categorical columns, converting categories into numerical values compatible with machine learning algorithms.*
5. ***Splitting and Scaling:*** *The data was divided into training and testing sets with a 70-30 split, followed by feature scaling to standardize the range of numerical features.*

These preprocessing steps provided a clean, standardized dataset suitable for accurate model training and evaluation.

Model Selection And Training

For this binary classification task, several machine learning algorithms were selected to cover a range of approaches. Each model was trained and optimized using hyperparameter tuning to achieve the best possible performance:

1. ***Logistic Regression:*** *Known for its simplicity and effectiveness in binary classification.*
2. ***Decision Tree Classifier:*** *Suitable for capturing non-linear patterns and providing feature importance insights.*
3. ***Random Forest Classifier:*** *An ensemble model that combines multiple decision trees for improved accuracy.*
4. ***Support Vector Machine (SVM):*** *Effective in high-dimensional spaces, using linear and RBF kernels for flexibility.*
5. ***k-Nearest Neighbors (k-NN):*** *A simple, non-parametric algorithm using distance-based classification.*
6. ***Gradient Boosting Classifier****: An ensemble boosting method known for strong performance on complex datasets.*

Hyperparameter Tuning

Each model is associated with a specific set of hyperparameters that are to be tested during training. Here's a breakdown of each model and its hyperparameters:

1. ***Logistic Regression****: Uses the C parameter, which controls the regularization strength. Values to test are [0.01, 0.1, 1, 10], where lower values indicate stronger regularization.*
2. ***Decision Tree Classifier:*** *Tests different max_depth values [3, 5, 7, 10], which control the maximum depth of the tree. Shallower trees reduce complexity but might limit accuracy, while deeper trees capture more detail but risk overfitting.*

3. ***Random Forest Classifier:*** *Uses two parameters:*
    - ***n_estimators:*** *Number of trees in the forest, with values [50, 100, 200].*
    - ***max_depth:*** *Maximum depth of each tree, with values [5, 10, 15].*
4. ***Support Vector Machine (SVM):*** *Tests both the regularization parameter C with values [0.1, 1, 10] and the kernel type (kernel) with options ['linear', 'rbf']. The kernel defines the transformation applied to the data before classification.*
5. ***k-Nearest Neighbors (k-NN):*** *Uses n_neighbors, the number of nearest neighbors considered for classification, with values [3, 5, 7].*
6. ***Gradient Boosting Classifier:*** *Tests two parameters:*
    - ***n_estimators:*** *Number of boosting stages, with values [50, 100, 200].*
    - ***learning_rate****: Controls the contribution of each tree, with values [0.01, 0.1, 0.2]. Lower values increase the number of boosting stages required for good performance.*

Model Evaluation

Each model's performance was evaluated using several standard metrics:

- ***Accuracy:*** *The proportion of correct predictions.*
- ***Precision and Recall:*** *Precision indicates how effectively the model avoids false positives, while recall shows the model's ability to identify true positives.*
- ***F1 Score:*** *A balance between precision and recall, particularly useful for imbalanced classes.*
- ***ROC AUC:*** *The area under the ROC curve, indicating the model's ability to distinguish between positive and negative classes.*

The table below summarizes each model's performance:

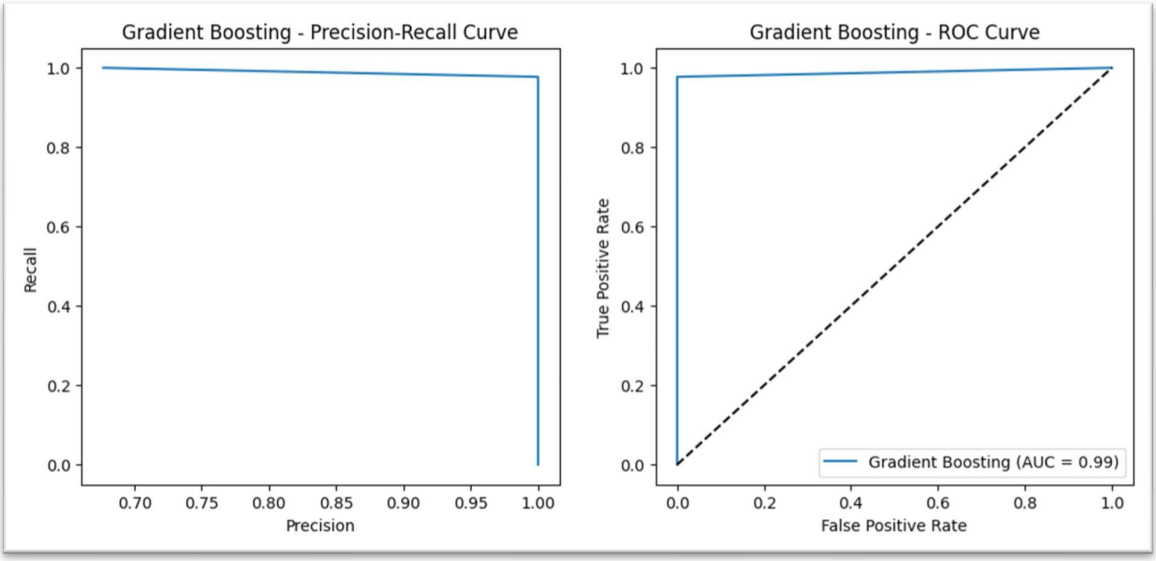
Model Performance Summary:

	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.753846	0.759259	0.931818	0.836735	0.92316
Decision Tree	0.984615	1.0	0.977273	0.988506	0.988636
Random Forest	0.892308	0.877551	0.977273	0.924731	0.983766
SVM	0.8	0.792453	0.954545	0.865979	0.922078
k-NN	0.784615	0.767857	0.977273	0.86	0.88961
Gradient Boosting	0.984615	1.0	0.977273	0.988506	0.988636

Precision-Recall and ROC Curves

Precision-Recall and ROC curves were used to visualize each model's classification performance. These visualizations helped highlight the trade-offs between precision and recall and allowed for a deeper analysis of each model's effectiveness in handling imbalanced data.

- ***Precision-Recall Curve:*** *The curve shows high recall (close to 1.0) across the precision range, indicating that the Gradient Boosting model is effective at identifying positive cases (placed students) with minimal false negatives.*
- ***ROC Curve:*** *The ROC curve is close to the top-left corner with an AUC of 0.99, signifying strong model performance in distinguishing between placed and not-placed students. The high AUC indicates that the model effectively differentiates between the two classes.*



Ensemble Model: Voting Classifier

To capitalize on the strengths of each model, a weighted Voting Classifier was constructed. The soft voting approach was applied, where predictions from the top-performing models were combined based on their F1 scores. The Voting Classifier achieved the following results:

- *Accuracy: 0.9846*
- *Precision: 0.9778*
- *Recall: 1.0*
- *F1 Score: 0.9888*
- *ROC AUC: 1.0*

The Voting Classifier outperformed individual models in accuracy and provided a balanced precision-recall trade-off, illustrating the value of ensemble techniques for complex classification tasks.

Campus Placement Analysis Dashboard Report

The Campus Placement Dashboard provides an interactive overview of key statistics and visualizations related to student placement outcomes. This dashboard leverages data on academic scores, degree types, and placement status to offer insights into factors influencing

student placements. The dashboard was developed using Streamlit, with Plotly visualizations for enhanced interactivity.

- ***Summary Statistics***
  The first section, Summary Statistics, displays the counts of students in each placement category (Placed vs. Not Placed). This summary provides an immediate sense of the placement distribution across the dataset, helping users understand the overall success rate in campus placements.
- ***Placement Rate by Degree Type***
  The second section, Placement Rate by Degree Type, shows the proportion of students placed within each degree type. This section uses a bar chart to visualize the placement rate across different degree categories, giving insights into how academic backgrounds might correlate with placement outcomes.
  Insight: Higher placement rates for specific degree types can indicate a trend where certain academic programs may better align with industry requirements, enhancing employability for students in those fields.
- ***SSC Score Distribution by Placement Status***
  The third section, SSC Score Distribution by Placement Status, displays a histogram of secondary school (SSC) scores, overlaid by placement status. This chart allows users to observe how SSC scores vary between placed and not-placed students, helping to identify if early academic performance has a significant influence on placement outcomes.
  Insight: Variations in the SSC score distributions by placement status can reveal patterns, such as whether higher SSC scores correlate with a greater likelihood of placement.

Conclusion

The Voting Classifier demonstrated the best overall performance, combining the predictive strengths of multiple models. Its high accuracy, precision, and recall make it an ideal choice for campus placement prediction. This ensemble approach can be further refined and tested on additional data to improve its generalization capabilities, helping institutions and students make data-driven decisions about placement opportunities.

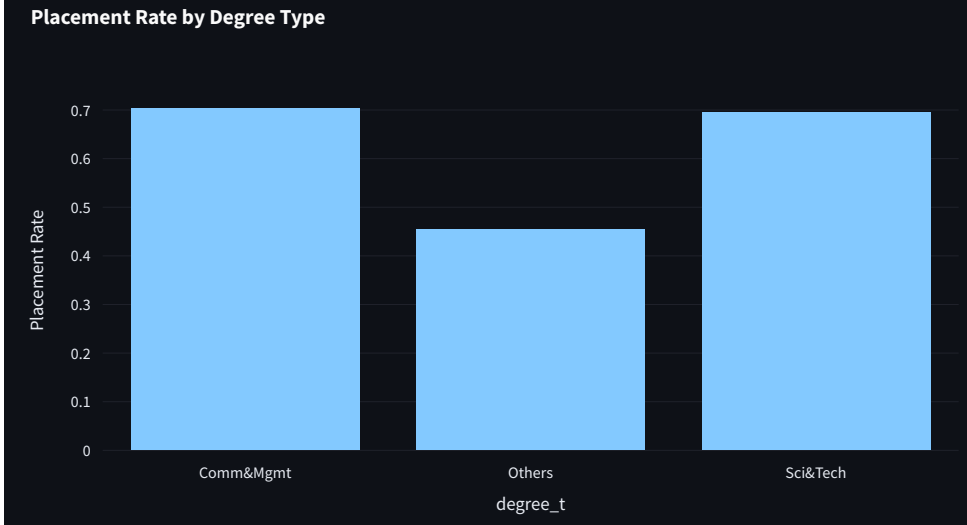
Campus Placement Dashboard

Summary Statistics

Placement Status Counts:

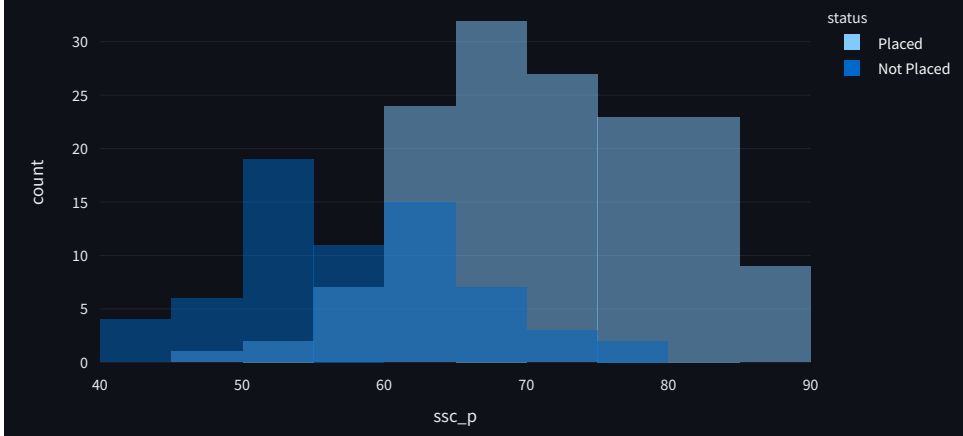
	status
Placed	148
Not Placed	67

Placement Rate by Degree Type



SSC Score Distribution by Placement Status

SSC Score Distribution



Made with Streamlit