BCSE209L - Machine Learning Project

# Placement and Salary Bracket Predictor

Team Members:

Srikar Ganesh (22BAI1080)

Meera R Deepu (22BAI1142)

Sarath Chander (22BAI1148)

Index:

## Introduction:

This project is dedicated to utilizing data analysis to provide valuable insights into students' post-graduation career prospects and potential salary levels. By examining a range of factors including academic performance, demographic characteristics, and work experience, we aim to develop predictive models that forecast the likelihood of students securing employment and the corresponding salary bands they may anticipate. This project holds significant relevance as it equips both students and educational institutions with actionable intelligence to make informed decisions regarding career pathways and academic strategies. Through rigorous analytical methodologies, this project aims to serve as a pivotal tool in optimizing career planning initiatives and enhancing academic guidance practices.

## Dataset and Models:

1. Dataset:
    a. Description:
    The dataset utilized in this study comprises various attributes of graduating students, including academic performance metrics such as secondary and higher secondary examination scores, as well as demographic information like gender and work experience. Additionally, it encompasses data pertaining to the students' performance in aptitude tests, their specialization choices, and eventual placement outcomes, including salary bands.
    b. Pre-processing:

| Attribute | Pre-Processing |
|---|---|
| Gender | Encoded using LabelEncoder |
| Secondary Education (SSC) Percentage | Scaled using MinMaxScaler to normalize between 0 and 1 |
| Secondary Education Board | Encoded using LabelEncoder |
| Higher Secondary Education (HSC) Percentage | Scaled using MinMaxScaler to normalize between 0 and 1 |
| Higher Secondary Education Board | Encoded using LabelEncoder |
| Higher Secondary Education Specialization | Encoded using LabelEncoder |
| Degree Percentage | Scaled using MinMaxScaler to normalize between 0 and 1 |
| Degree Type | Encoded using LabelEncoder |
| Work Experience | Encoded using OneHotEncoder |
| Aptitude Test Percentage | Scaled using MinMaxScaler to normalize between 0 and 1 |

| | |
|---|---|
| MBA Percentage | Scaled using MinMaxScaler to normalize between 0 and 1 |
| Specialization in MBA | Encoded using OneHotEncoder |
| Placement Status | Encoded using LabelEncoder |
| Salary Band | Transformed based on predefined salary bands to categorical labels |

2. Models:
    a. Logistic Regression:
        i. Logistic Regression is a statistical method used for binary classification tasks.
        ii. It models the probability that a given input belongs to a certain class using the logistic function.
        iii. The algorithm optimizes the parameters to minimize the difference between predicted and actual class labels.
    b. Support Vector Machine (SVM):
        i. SVM is a supervised learning algorithm used for classification and regression tasks.
        ii. It works by finding the hyperplane that best separates different classes in the feature space.
        iii. SVM aims to maximize the margin between classes while minimizing classification errors.
    c. Naive Bayes:
        i. Naive Bayes is a probabilistic classifier based on Bayes' theorem.
        ii. It assumes that the features are conditionally independent given the class.
        iii. Naive Bayes calculates the probability of each class for a given input and selects the class with the highest probability.
    d. XGBoost:
        i. XGBoost is a powerful ensemble learning algorithm used for classification and regression tasks.
        ii. XGBoost builds a series of decision trees sequentially, where each subsequent tree corrects the errors made by the previous ones.
        iii. It employs a gradient descent optimization technique to minimize a specified loss function while adding trees to the ensemble.
    e. Random Forest:
        i. Random Forest is an ensemble learning method based on Decision Trees.
        ii. It constructs multiple decision trees on random subsets of the data and features.
        iii. The final prediction is made by aggregating the predictions of individual trees, often resulting in improved accuracy and robustness.

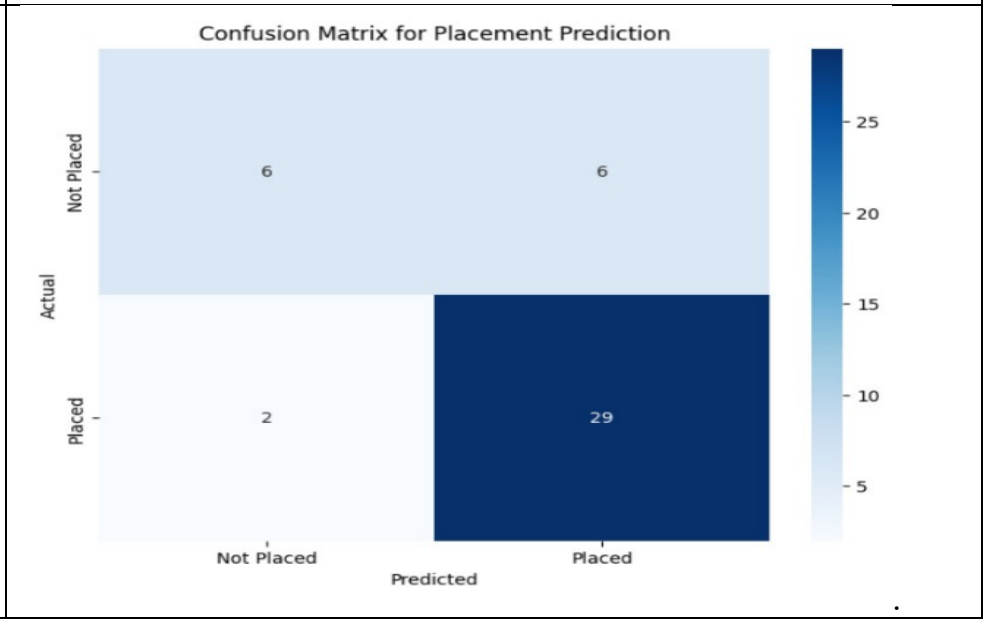
f. Neural Network:
   i. Neural Network is a biologically inspired computational model composed of interconnected nodes (neurons).
   ii. It consists of an input layer, one or more hidden layers, and an output layer.
   iii. Neural networks use mathematical functions to process input data and learn complex patterns by adjusting the weights between neurons through backpropagation algorithm to minimize prediction errors.

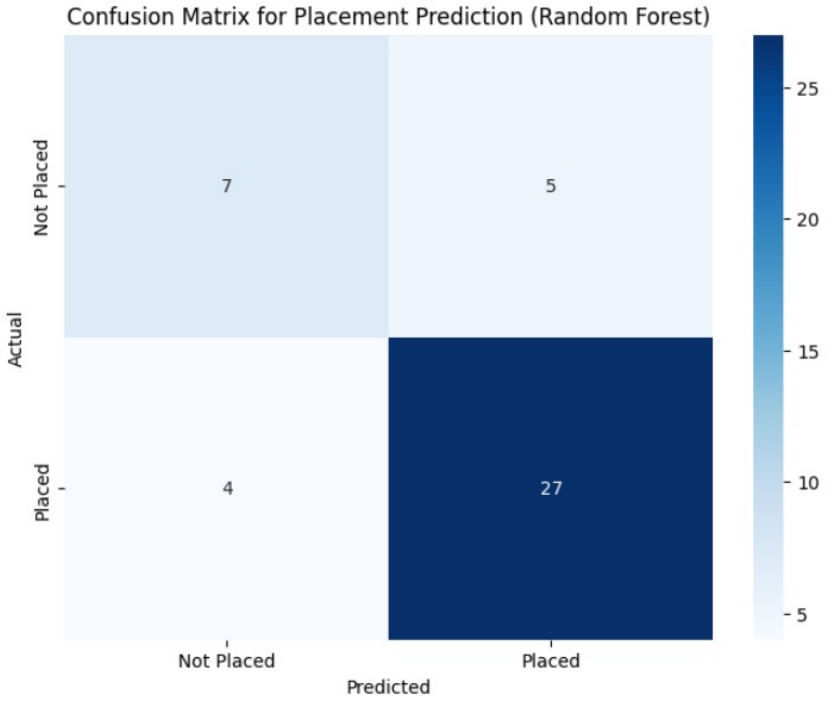
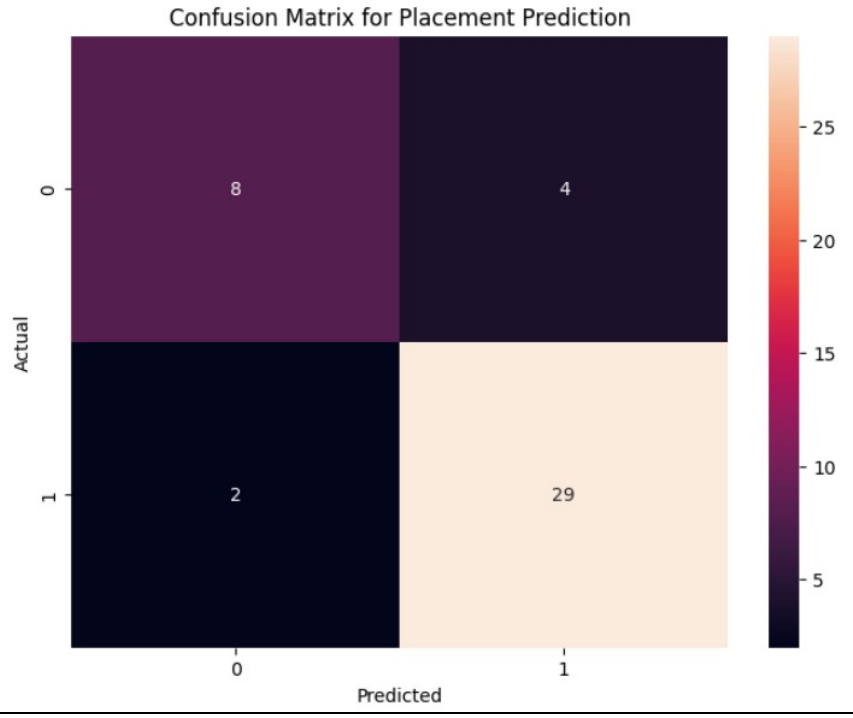
## Comparing Results:

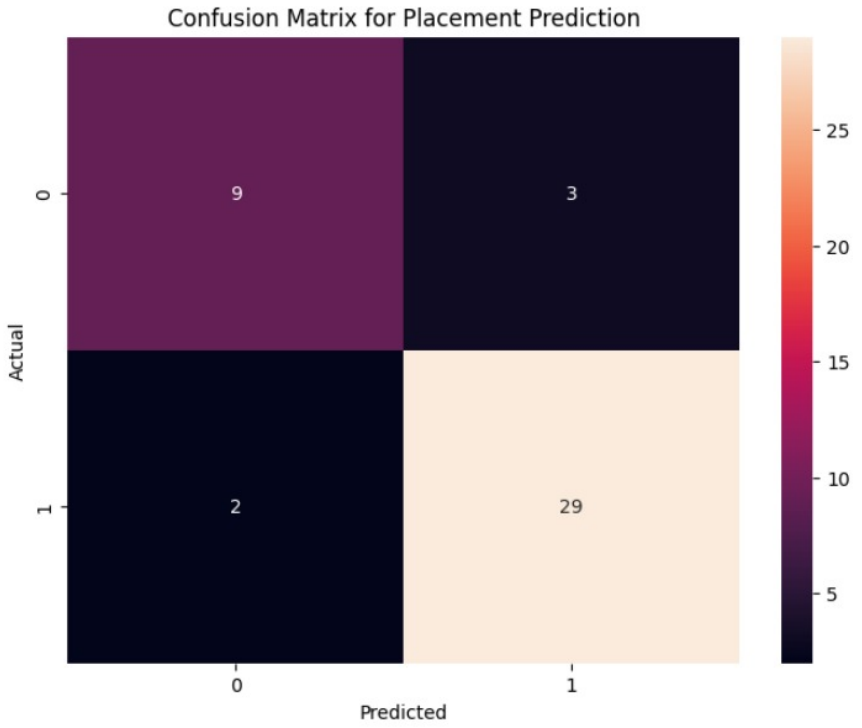
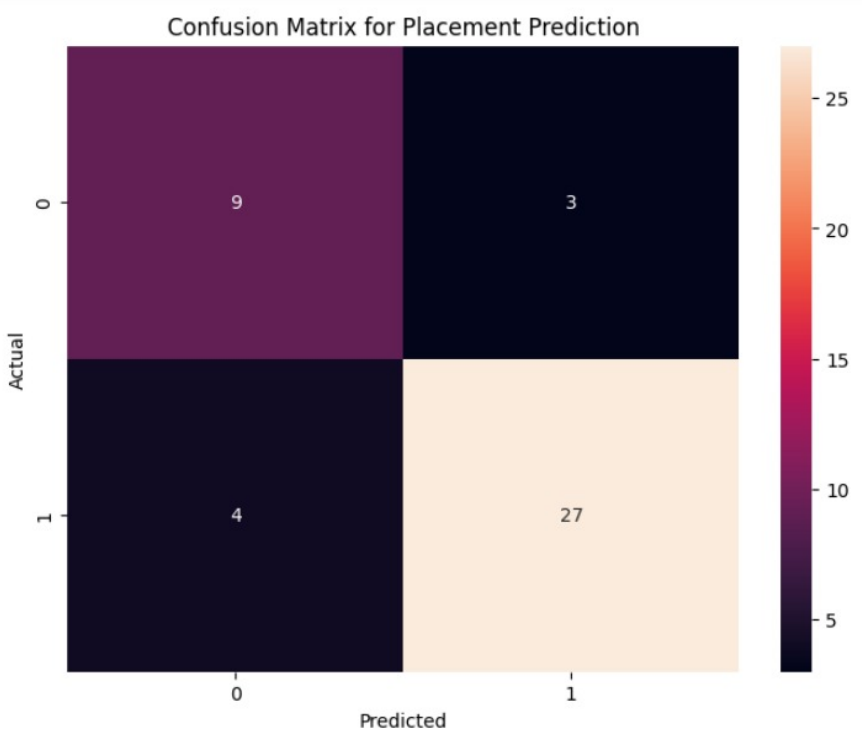
1. Accuracy and RMSE value:

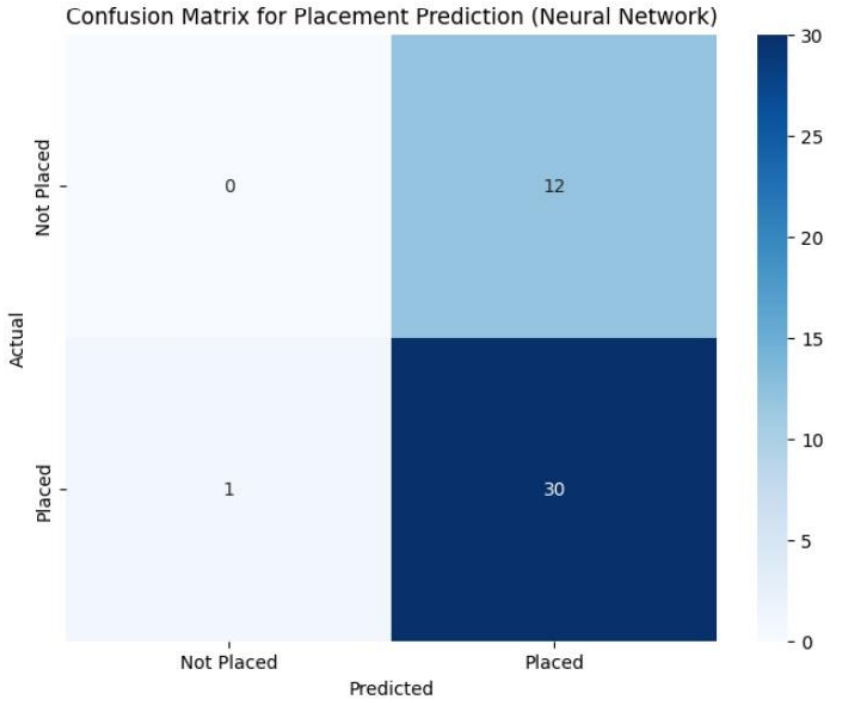
| Model | Accuracy (Placement) | RMSE value (Salary Band) |
|---|---|---|
| XGBoost | 0.883 | 0.0001 |
| RandomForest | 0.796 | 0.086 |
| Logistic Regression | 0.8604 | 0.3049 |
| SVM | 0.813 | 0.152 |
| Naïve Bayes | 0.8372 | Not Applicable |
| Neural Networks | 0.697 | 0.0523 |

2. Confusion Matrix, ROC Curve, Feature Importance:

| Model | Confusion Matrix |
|---|---|
| XGBoost |  |

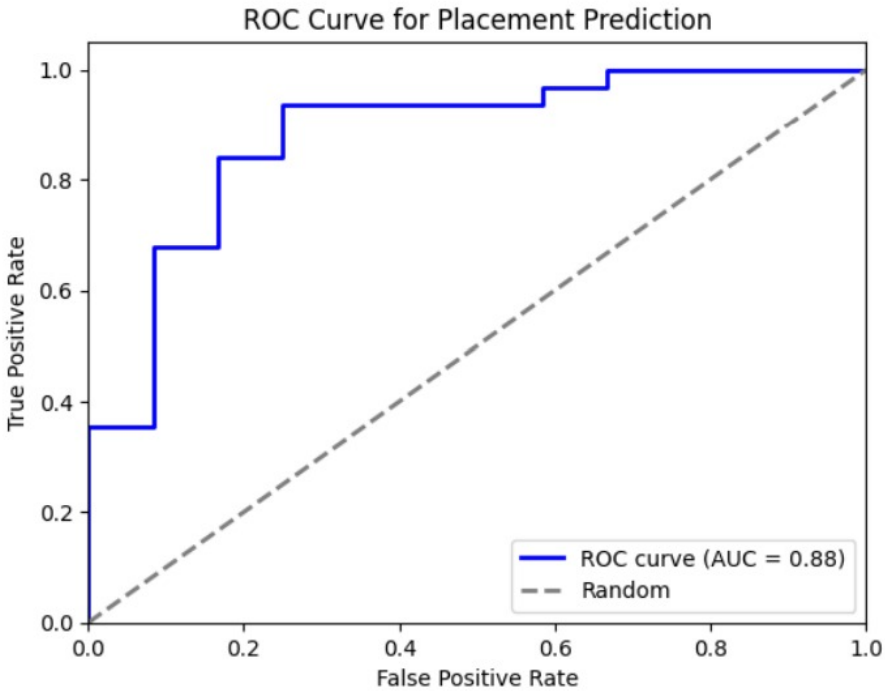
| | |
|---|---|
| RandomForest |  |
| Logistic Regression |  |

| | |
|---|---|
| SVM |  |
| Naïve Bayes |  |

| | |
|---|---|
| Neural Networks |  |

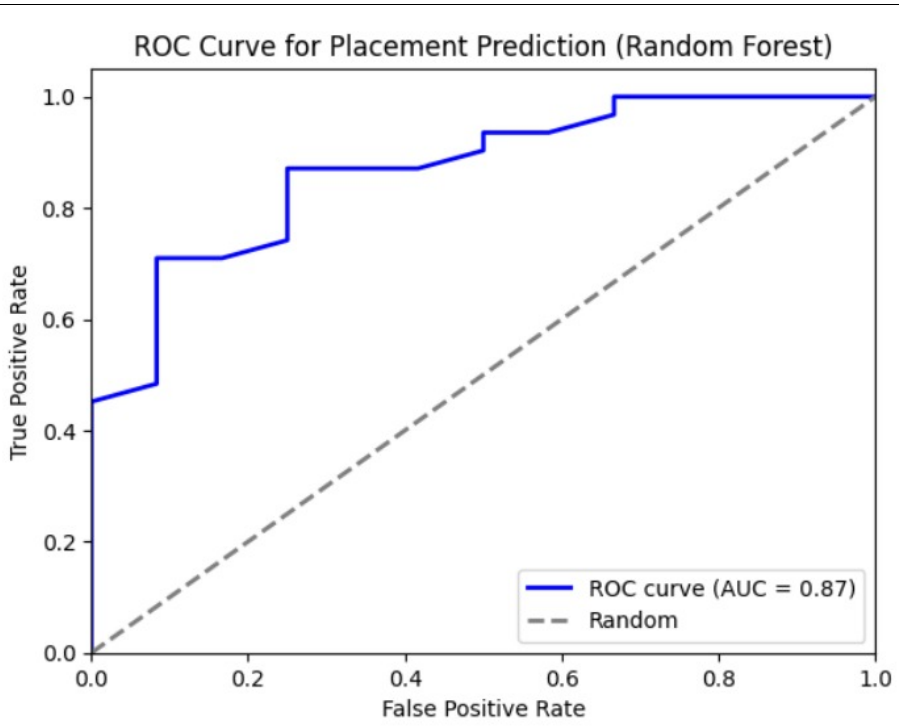
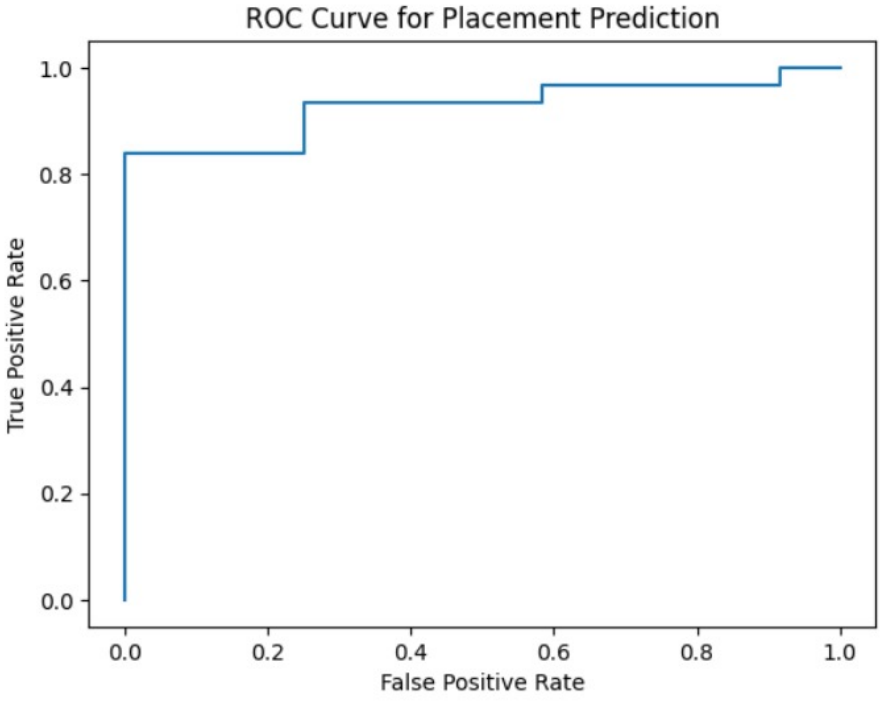


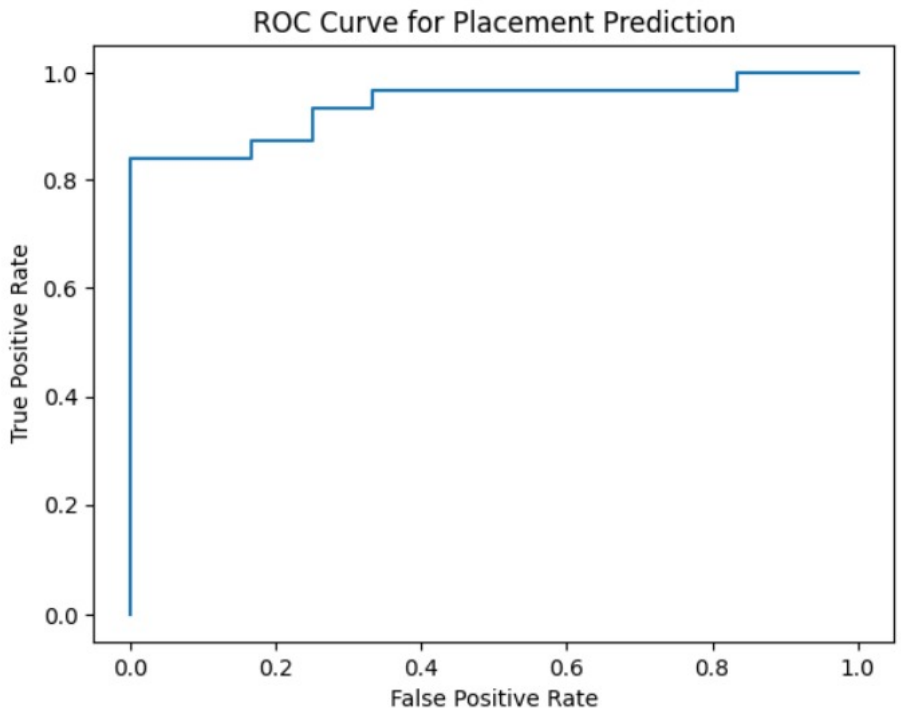
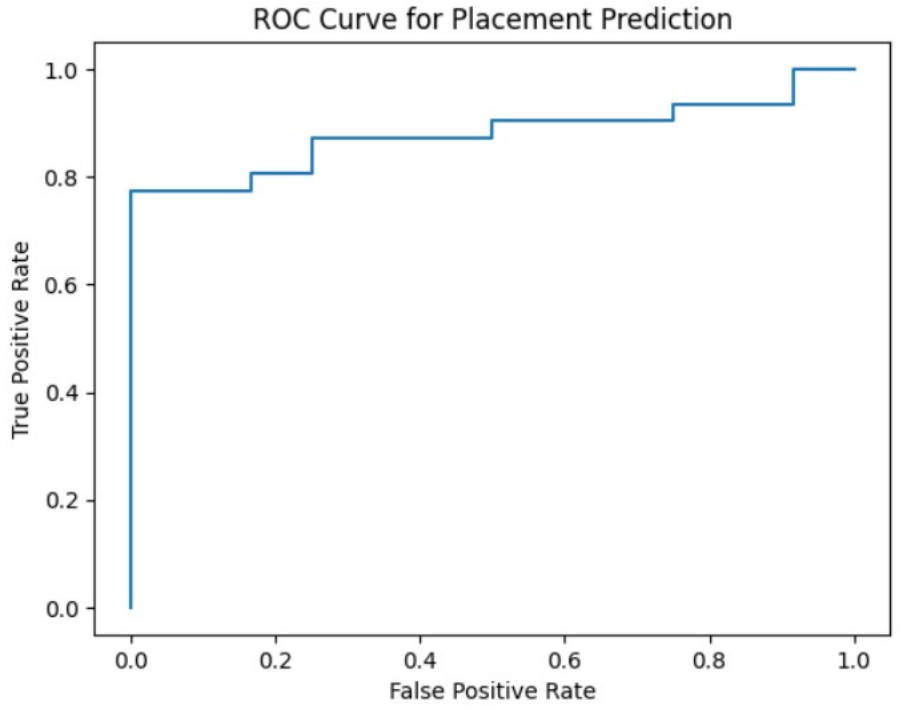
3. ROC Curve:

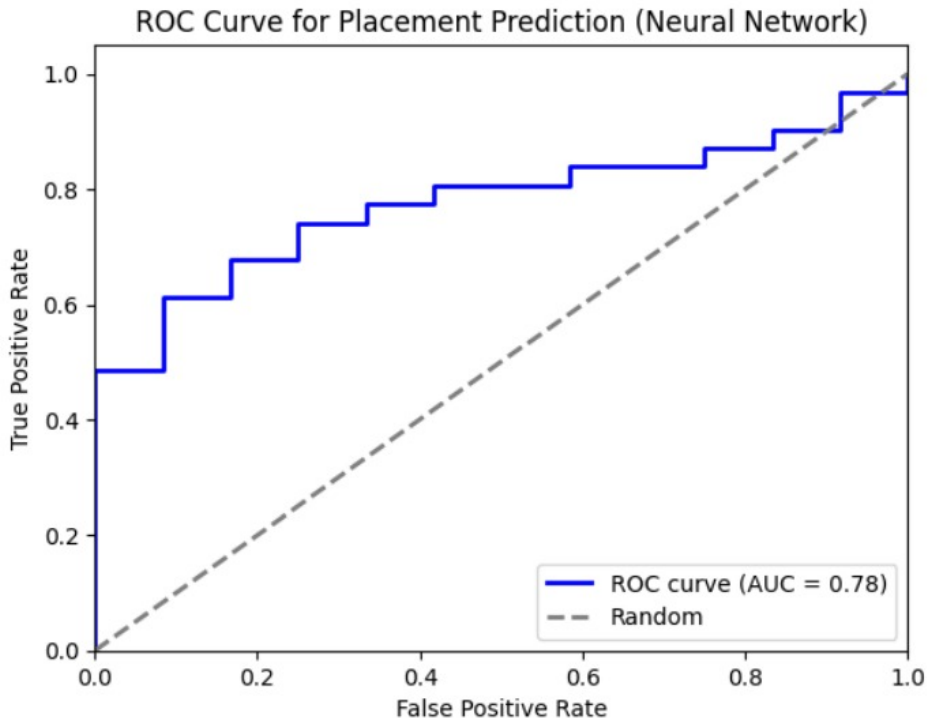
| Model | ROC Curve |
|---|---|
| XGBoost |  |



.

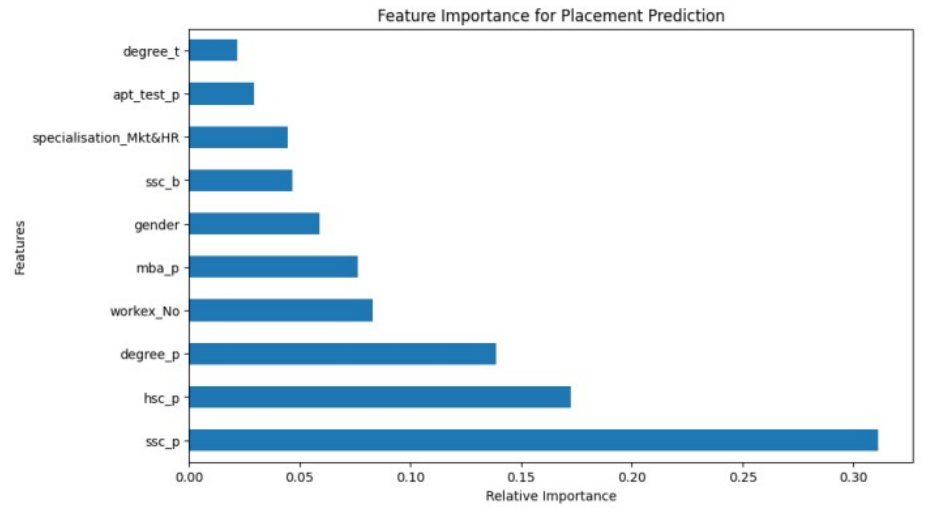
| | |
|---|---|
| RandomForest |  |
| Logistic Regression |  |

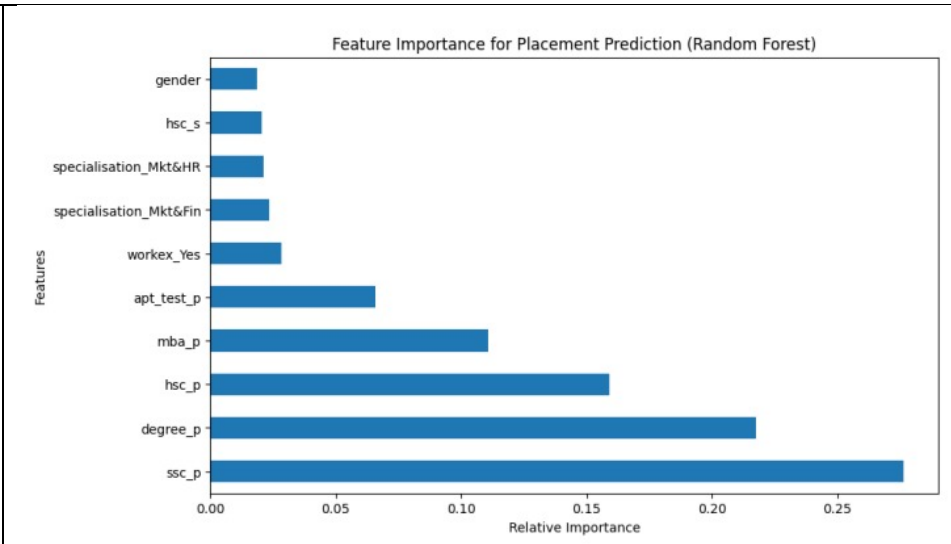
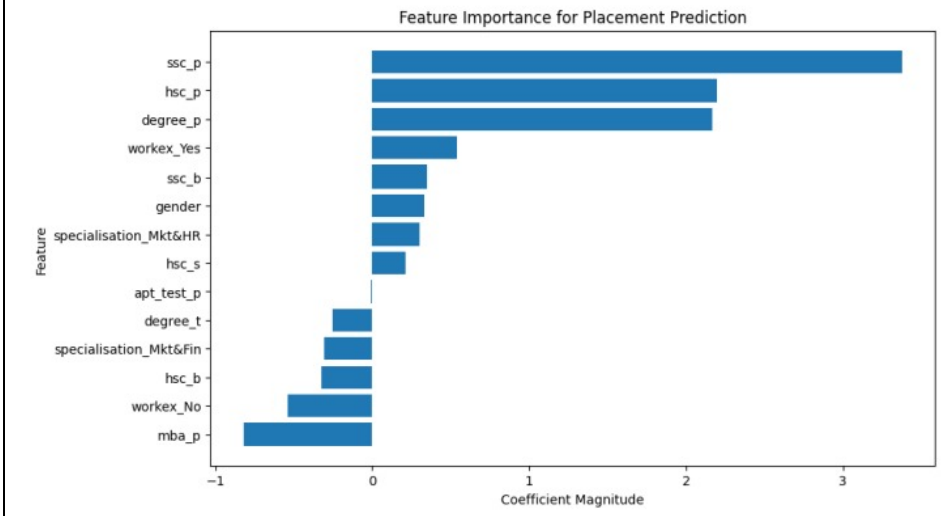
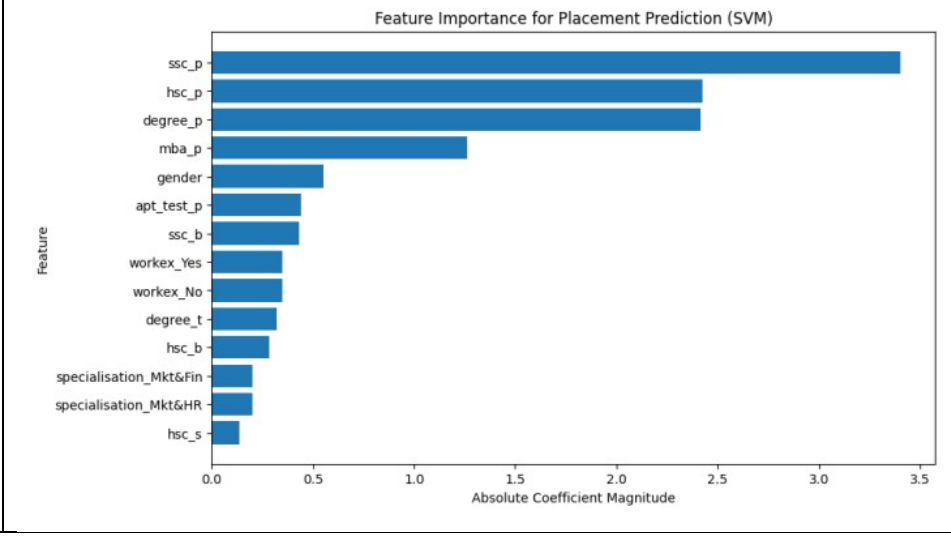
| | |
|---|---|
| SVM |  |
| Naïve Bayes |  |

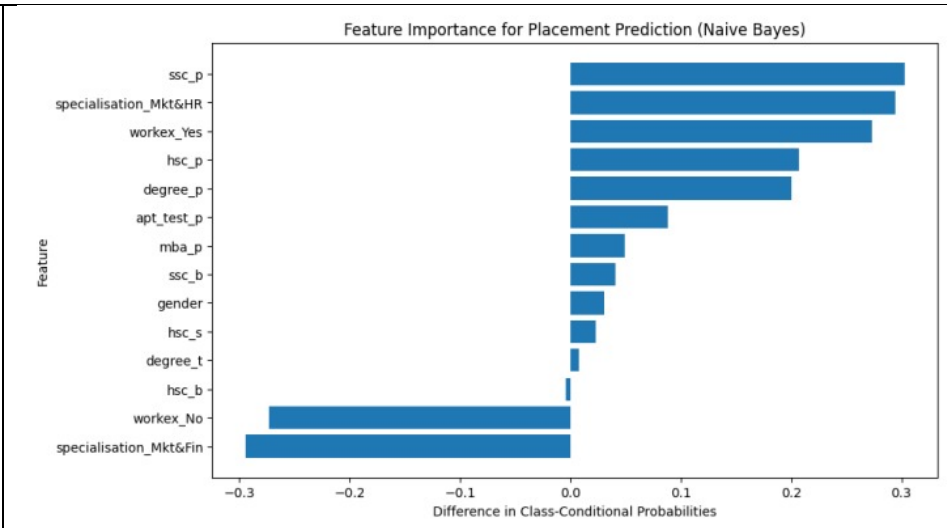
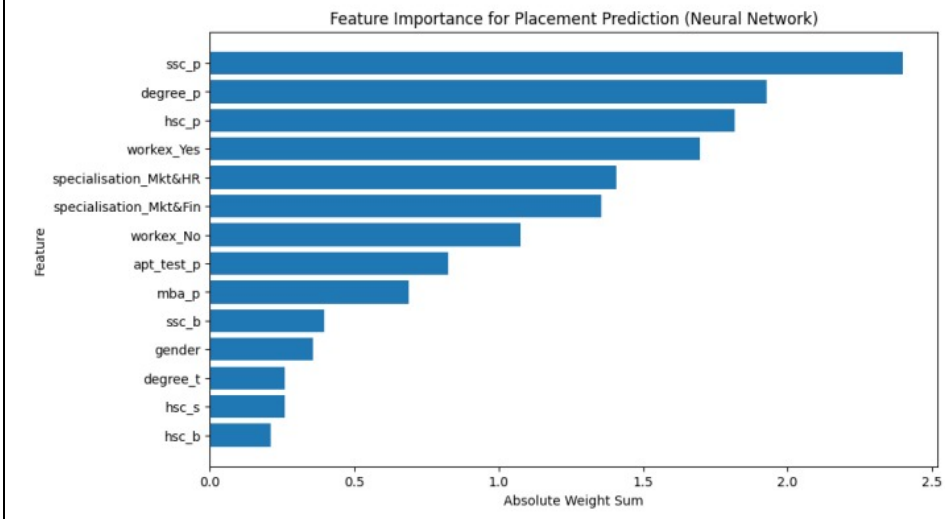
| | |
|---|---|
| Neural Networks |  |



4. Feature Importance:

| Model | Feature Importance |
|---|---|
| XGBoost |  |

| | |
|---|---|
| RandomForest |  |
| Logistic Regression |  |
| SVM |  |

| | |
|---|---|
| Naïve Bayes |  |
| Neural Networks |  |

## Conclusion:

The best fit model for the given dataset can be found by finding the model that gives the highest accuracy for placement prediction and lowest RMSE value for salary bracket prediction.

According to the observed results, which are recorded in the tables above, it is found that XGBoost model has the necessary conditions to be the model of best fit. As observed, the confusion matrix has the highest true positives and true negatives, and the ROC curve tends most to the upper-left corner, indicating high sensitivity and specificity across many threshold values.

XGBoost demonstrates superior performance on this dataset due to its ability to handle complex relationships in the data, resulting in higher accuracy and lower RMSE compared to other models. Its optimized implementation of gradient boosting ensures robust predictions and scalability, making it the top choice for both placement and salary

bracket prediction tasks. Additionally, XGBoost's ensemble approach mitigates overfitting

Another feature we can observe that, all models unanimously point that ssc_p is the attribute that most affects the result of placements. This can be observed from the feature importance graph generated by each model.

In conclusion, this project has successfully generated the results for placement and salary bracket prediction, and has also uniquely identified the best fit model for the data.

----------The End----------