Lambton College In Mississauga

Report on

Student Placement Prediction

Done BY: Amandeep Singh

Big Data Analytics (DSMM), Lambton college 2025S-T3 AML 3104 - Neural Networks and Deep Learning 01 (DSMM Group 1)

Ishant Gupta

Lambton College In Mississauga

Table of Contents

1. Introduction: 3
   1.1. Dataset Description: 3
2. Data Preprocessing: 4
3. Model Selection and Justification: 4
4. Model Performance Evaluation: 5
   4.1. Model Performance: 5
   4.2. Discussion of Results: 5
5. Conclusion: 6

1. Introduction:

The primary objective of this project is to develop a machine learning model capable of predicting whether a student will be recruited during campus placements. This prediction is based on a variety of academic, demographic, and specialization factors available in the dataset provided. Accurate placement prediction can significantly aid educational institutions in understanding key employability drivers and assist students by identifying areas for improvement.

1.1. Dataset Description:

The dataset used for this project is train.csv. It initially contained 215 student records and 15 features.

Key features include:

1. gender: Gender of the student.
2. ssc_p: Secondary Education percentage.
3. ssc_b: Board of Secondary Education (Central or Others).
4. hsc_p: Higher Secondary Education percentage.
5. hsc_b: Board of Higher Secondary Education (Central or Others).
6. hsc_s: Specialization in Higher Secondary Education (e.g., Commerce, Science, Arts).
7. degree_p: Degree Percentage.
8. degree_t: Under Graduation Degree Type (e.g., Sci&Tech, Comm&Mgmt).
9. workex: Work Experience (Yes or No).
10. etest_p: Employability test percentage.
11. specialisation: MBA Specialization (e.g., Mkt&HR, Mkt&Fin).
12. mba_p: MBA percentage.

The target variable for prediction is status, which indicates whether a student was 'Placed' or 'Not Placed'.

The dataset also included a sl_no (serial number) column and a salary column. These were excluded from modeling to prevent data leakage and ensure model integrity.

2. Data Preprocessing:

To prepare the data for modeling, the following preprocessing steps were performed:

1. Dropped 'sl_no' as it is an identifier.
2. Dropped 'salary' to prevent data leakage.
3. Encoded 'status' target variable using LabelEncoder.
4. Identified numerical and categorical features.
5. Scaled numerical features using StandardScaler.
6. Encoded categorical features using OneHotEncoder (drop='first').
7. Split the data into 70% training and 30% testing sets with stratification on the target.

3. Model Selection and Justification:

To predict student placement outcomes effectively, we evaluated a diverse set of machine learning models based on their strengths in handling classification tasks:

1. Logistic Regression: Selected as a baseline model due to its simplicity, speed, and ease of interpretability. It provides a solid foundation for comparison with more complex models.
2. Decision Tree Classifier: Chosen for its ability to model non-linear relationships and its intuitive, rule-based decision-making process.
3. Random Forest Classifier: An ensemble technique that combines multiple decision trees to reduce overfitting and improve overall accuracy and robustness.

To enhance the performance of the Random Forest model, we performed hyperparameter tuning using GridSearchCV with 3-fold cross-validation. The optimal parameters found are:

Best Parameters for Random Forest (Tuned):

{'classifier__max_depth': 10, 'classifier__max_features': 'sqrt', 'classifier__min_samples_leaf': 2, 'classifier__min_samples_split': 5, 'classifier__n_estimators': 100}

These parameters effectively balance model complexity and generalization by:

1. Controlling the number of trees (n_estimators)
2. Limiting the depth of each tree (max_depth)
3. Specifying thresholds for splitting (min_samples_split) and leaf creation (min_samples_leaf)
4. Selecting a random subset of features at each split (max_features)

In addition, we implemented a Voting Classifier (Hard Voting), which aggregates predictions from Logistic Regression, Decision Tree, and the Tuned Random Forest models. This ensemble approach harnesses the strengths of all three classifiers, aiming to boost predictive performance through consensus-based decision-making.

4. Model Performance Evaluation:

4.1. Model Performance:

The following metrics were used for evaluation: Accuracy, Precision, Recall, F1-Score.

Performance summary:

Model	**Accuracy**	**Precision (Placed)**	**Recall (Placed)**	**F1-score (Placed)**
Logistic Regression	0.8154	0.8235	0.9333	0.8750
Decision Tree	0.7538	0.7736	0.9111	0.8367
Random Forest (Default)	0.8615	0.8462	0.9778	0.9072
Random Forest (Tuned)	0.8615	0.8462	0.9778	0.9072
Voting Classifier (Hard)	0.8615	0.8462	0.9778	0.9072

4.2. Discussion of Results:

Based on the empirical performance metrics obtained from the model evaluation, it's evident that several models demonstrated strong capabilities in predicting student placement. Interestingly, the Random Forest (Default), Random Forest (Tuned), and Voting Classifier (Hard) all achieved the highest F1-score of 0.9072 for the 'Placed' class, indicating a very similar level of high predictive performance in this specific run.

1. Best Performing Model and Reasoning: In this evaluation, no single model distinctly outperformed the others among the top contenders. The Random Forest (Default), Random Forest (Tuned), and Voting Classifier (Hard) models all exhibited the highest F1-score. Random Forest models typically perform well due to their ensemble nature, which combines multiple decision trees to reduce individual tree biases and variances, thereby improving

overall generalization. The consistency in performance across the default and tuned versions of Random Forest, as well as the Voting Classifier, suggests that the dataset is well-suited for these robust ensemble approaches.

2. Impact of Hyperparameter Tuning on Random Forest: In this specific execution, hyperparameter tuning did not lead to an increase in the Random Forest model's F1-score. The Random Forest (Tuned) model's F1-score (0.9072) was identical to that of the Random Forest (Default) model (0.9072). This outcome suggests that the default parameters of the RandomForestClassifier were already highly effective for this dataset, or that the defined hyperparameter search space did not contain a combination that significantly improved upon the default performance. While tuning didn't yield a measurable improvement in this metric, it confirms the robustness of the Random Forest model and validates that the selected tuning process maintained its strong performance.

3. Comparison with the Voting Classifier: The Voting Classifier (Hard) also matched the top performance with an F1-score of 0.9072. This demonstrates the effectiveness of ensemble methods that combine predictions from multiple diverse models (Logistic Regression, Decision Tree, and the Tuned Random Forest). Even though it didn't surpass the best individual Random Forest model in this instance, the Voting Classifier's ability to achieve the same peak performance highlights its stability and potential for mitigating the weaknesses of individual models, providing a reliable and consistent predictive outcome. The overall high F1-scores across these top models indicate a strong capability to correctly identify "Placed" students.

5. Conclusion:

This project successfully developed and evaluated several machine learning models for predicting student placement. In this analysis, the Random Forest (Default), Random Forest (Tuned), and Voting Classifier (Hard) models collectively demonstrated the most promising performance, all achieving an F1-score of 0.9072 for the 'Placed' class. These insights can

significantly help educational institutions in understanding key employability drivers and better prepare students for successful job placements.