Assignment 3 - Machine Learning
Student Placement Analysis

Mohammad Hossein Basouli

May 18, 2025

1 Introduction

In this analysis, we try to predict **student placement on a campus**, based on multiple factors; such as their *Academic Data*, *Demographic Information* and *Institutional Attributes.*

2 Data

The dataset[1] has been gathered by Ben Roshan D, who is doing MBA in Business Analytics at Jain University Bangalore.
It includes 215 rows and 15 features, among these 7 are categorical and 8 are numerical.

3 Exploratory Data Analysis

3.1 Data Cleaning & Data Transformation:

- **Handling Missing Values**: Only those students that *have not been place in the campus* have missing values, only in their *salary* column (Figure 1). We will impute this by the median of this column. (Figure 2 shows the distribution of *salary* across the two classes after the imputation.)

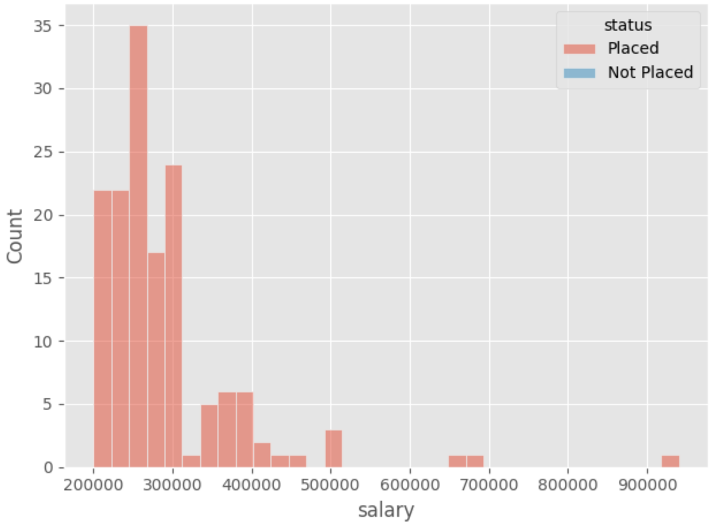

Figure 1: Distribution of Salary Across The Classes Before The Imputation

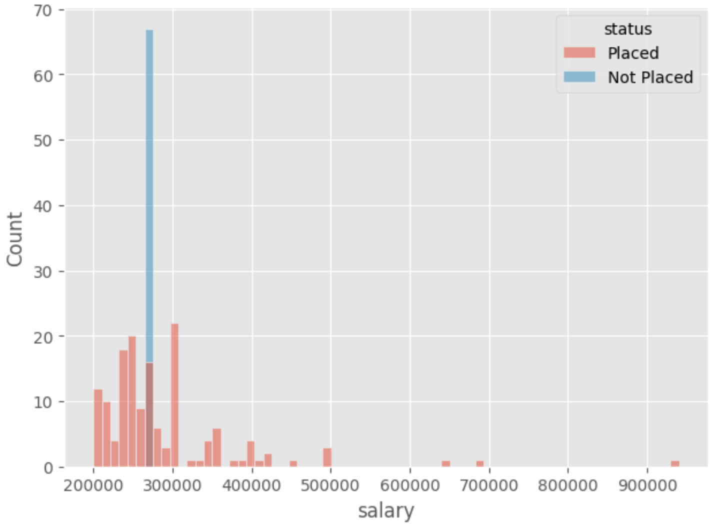

Figure 2: Distribution of Salary Across The Classes After The Imputation

- **Encoding of The Categorical Features**: We will use **One-Hot Encoding** to transform our 7 categorical features into numerical.

4 Models

4.1 Train & Test Splitting:

%80 Train - %20 Test

4.2 Threshold Tunning:

We will use **10-Fold Stratified Cross Validation** for specifying the optimal threshold for each of our models.

4.3 Logistic Regression

4.3.1 Optimal Threshold Obtained by Cross Validation:

Optimal threshold from CV: 0.50

4.3.2 Model Evaluation:

Table 1: Logistic Regression Performance Metrics

Dataset	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Train	0.9012	0.9167	0.9402	0.9283
Test	0.8837	0.9063	0.9355	0.9206

ROC Curve: The **ROC Curve** for this model, evaluated on the test set, has a **Area Under The Curve** equal to 0.84, which is not that great. Also the point that the blue lines intersect with each other, is the optimal point for **True Positive Rate** and **False Negative Rate** trade-off.

Recall - Precision Curve: The **Recall - Precision Curve** has a **Average Precision** equal to 0.98 which is great. Also the optimal point for the trade-off of **Recall** and **Precision** is the point where recall is around 0.82 and the curve is starting to fall.

Confusion Matrix: From the **Confusion Matrix** we can see that we have misclassified 5 samples from a total of 43 samples, 2 **False Positive** and 3 **False Negatives**.

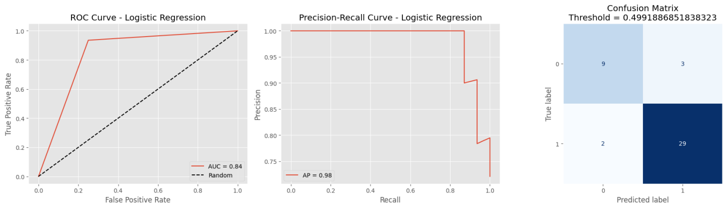

Figure 3: ROC Curve, Recall - Precision Curve and Confusion Matrix for Logistic Regression

4.4 Gaussian Naive Bayes

4.4.1 Optimal Threshold Obtained by Cross Validation:

Optimal threshold from CV: 1.0

4.4.2 Model Evaluation:

Table 2: Gaussian Naive Bayes Performance Metrics

Dataset	Accuracy	Precision	Recall	F1 Score
Train	0.9651	1.0000	0.9487	0.9737
Test	1.0000	1.0000	1.0000	1.0000

ROC Curve: The two lines intersect at a right angle, which is the optimal point for the trade-off of **True Positive Rate** and **False Negative Rate**. The **Area Under The Curve** is 1.0, which is the best that we can hope for. This indicates that the model is perfectly capable of distinguishing between the two classes.

Recall - Precision Curve: The **Recall - Precision Curve** also shows a perfect **Average Precision** of 1.0. The curve stays at maximum precision for all recall levels until the very end, suggesting excellent performance in class separation.

Confusion Matrix: From the **Confusion Matrix**, we observe that all 43 samples were classified correctly. There are **0 False Positives** and **0 False Negatives**, indicating perfect performance on the test set.

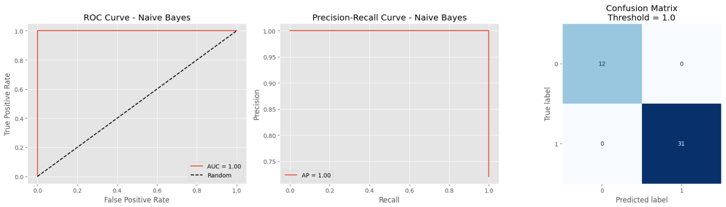

Figure 4: ROC Curve, Recall - Precision Curve and Confusion Matrix for Gaussian Naive Bayes

4.5 Linear Discriminant Analysis

4.5.1 Optimal Threshold Obtained by Cross Validation:

Optimal threshold from CV: 0.638

4.5.2 Model Evaluation:

Table 3: Linear Discriminant Analysis Performance Metrics

Dataset	Accuracy	Precision	Recall	F1 Score
Train	0.9012	0.9310	0.9231	0.9270
Test	0.8605	0.9032	0.9032	0.9032

ROC Curve: The **ROC Curve** shows an **Area Under The Curve** of 0.83, which is slightly less than ideal but still fairly strong. The curve indicates a good separation between the classes, although not as pronounced as in Naive Bayes.

Recall - Precision Curve: The **Recall - Precision Curve** has a high **Average Precision** of 0.98, suggesting the model performs well in terms of maintaining high precision even as recall increases. The curve begins to slightly fall around a recall of 0.82, which marks a reasonable trade-off point.

Confusion Matrix: The **Confusion Matrix** reveals that 6 out of 43 samples were misclassified: 3 **False Positives** and 3 **False Negatives**. While the model performs well overall, there is a small number of errors that might be worth addressing depending on the application.

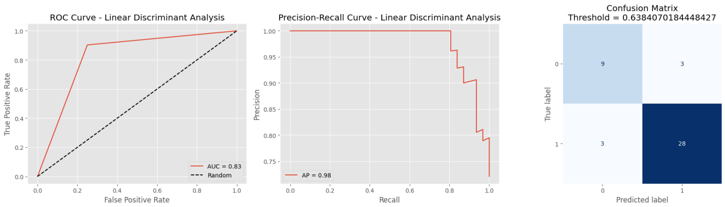

Figure 5: ROC Curve, Recall - Precision Curve and Confusion Matrix for Linear Discriminant Analysis

4.6 Best Model:

Gaussian Naive Bayes performs better in all aspects.

References

[1] Campus Recruitment, *Academic and Employability Factors influencing placement*,
Available at: `https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement/data`, 2020.