Submitter study group: -

Vikas Tunwal
Anuj Kumar Tiwari
rang Thuy Van

Summary Overview

Methodology:

1. Library and Data Ingestion:

The process kicked off by incorporating necessary libraries for analysis, followed by loading the dataset for examination.

2. Data Scrutiny and Understanding:

The initial step was devoted to gaining a comprehensive understanding of the data by inspecting the dataset's structure and attributes.

3. Data Preprocessing and Exploratory Data Analysis (EDA):

A series of procedures were carried out to prepare the data for analysis. We identified outliers, checked for duplicate entries, replaced 'Select' values with 'nan' as they appeared to be default values left unchanged by the user, and addressed columns with null entries exceeding 40%. Univariate and Bivariate analyses were also conducted for a deeper understanding of individual variables and their relationships.

- Univariate Analysis:

Some columns such as 'Country', 'what is your current occupation', and 'What matters most to you in choosing a course', were found to be highly skewed towards 'India', 'Unemployed', and 'Better Career Prospects' respectively. As such, these columns were dropped. Columns with single values or extreme skewness towards one category (>95%) were also eliminated.

- Bivariate Analysis:

The distribution of categorical columns was inspected. Columns with negligible category distribution were identified and grouped under a new column. We then created dummy variables and excluded the first category to avoid multicollinearity.

4. Splitting the Data:

The data was divided into a training set and a test set to facilitate model validation.

5. Feature Scaling:

In order to prevent variables with large magnitudes from overpowering the model, the variables were scaled to be on the same plane.

6. Model Construction & Refinement:

A model was built using Recursive Feature Elimination (RFE) on the selected columns. The model was further optimized by removing columns based on their p-values.

7. ROC Curve Plotting:

An ROC curve was plotted to visualize the performance of the model at different classification thresholds.

8. Determining the Optimal Cut-off Point:

We identified the optimal cutoff point to maximize model performance.

9. Test Set Predictions & Model Validation:

Predictions were made on the test set to validate the model's performance.

Final Model Metrics:

The ultimate model showed promising results with an accuracy of 89%, a sensitivity of 88%, and a specificity of 90%. These metrics provide a solid foundation for the model's applicability and robustness.