

CSE422 Lab Project Report

Group-12

Debapriyo Ganguly - 22299164

Ayesha Mohsina - 22299021

Table of Contents

- 1. Introduction**
- 2. Dataset Description**
- 3. Dataset pre-processing**
- 4. Dataset Splitting**
- 5. Model Training & Testing (Supervised)**
- 6. Model Selection / Comparison Analysis**
- 7. Conclusion**

Introduction

This project aims to predict whether an individual will switch their career or not, based on demographic, educational, and professional background information. The motivation is to understand the factors influencing career change, which can help companies plan recruitment, training, and employee retention strategies.

Dataset Description

- **How many features?**

After removing the ID column, there are 12 features + 1 target column.

- **Classification or regression problem? Why?**

This is a classification problem, because the target (will_change_career) is categorical (Yes/No encoded as 1/0).

- **How many data points?**

The dataset contains 5,000 rows, which means around 5,000 data points (Accurate number depends on the null values).

- **What kind of features?**

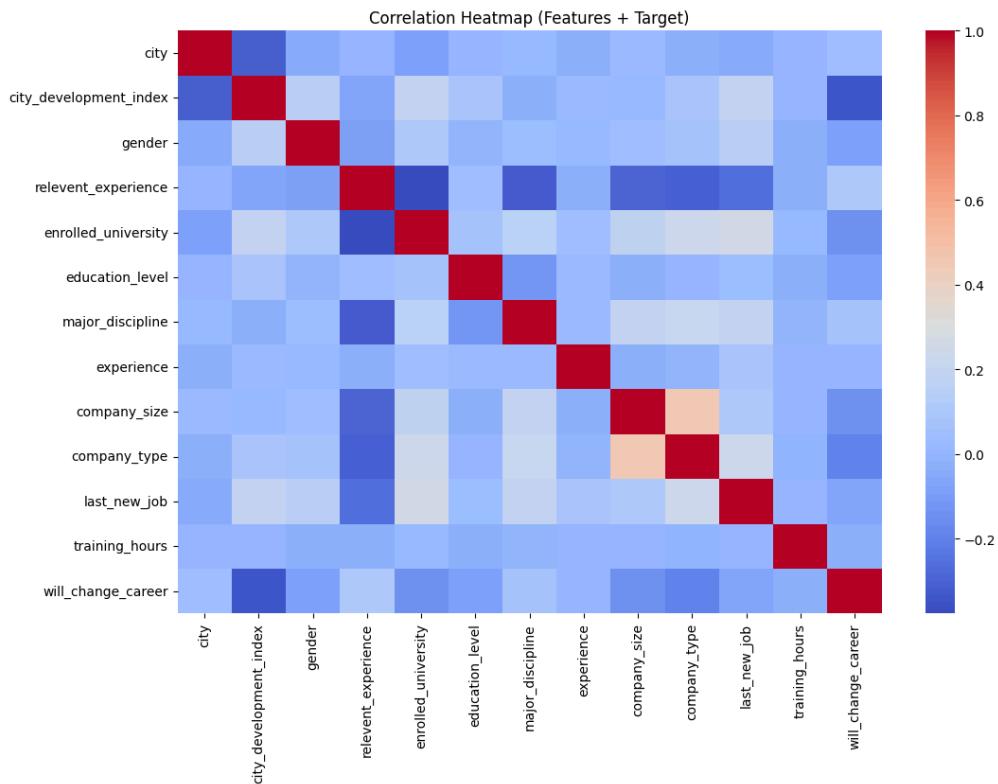
Both Quantitative (e.g., city_development_index, training_hours) and Categorical (e.g., gender, education_level, company_type).

- **Do you need to encode categorical variables? Why/Why not?**

Yes, because models like Logistic Regression and Neural Networks cannot handle strings directly. We used OneHotEncoding for categorical features.

- **Correlation of features (heatmap)?**

The heatmap shows most features have weak correlation with the target, meaning no single feature can predict career change strongly on its own. Instead, a combination of features contributes.



- **Understanding after correlation test:**

No feature has a strong linear relationship with career switching, suggesting this is a complex problem requiring ML models instead of simple threshold rules.

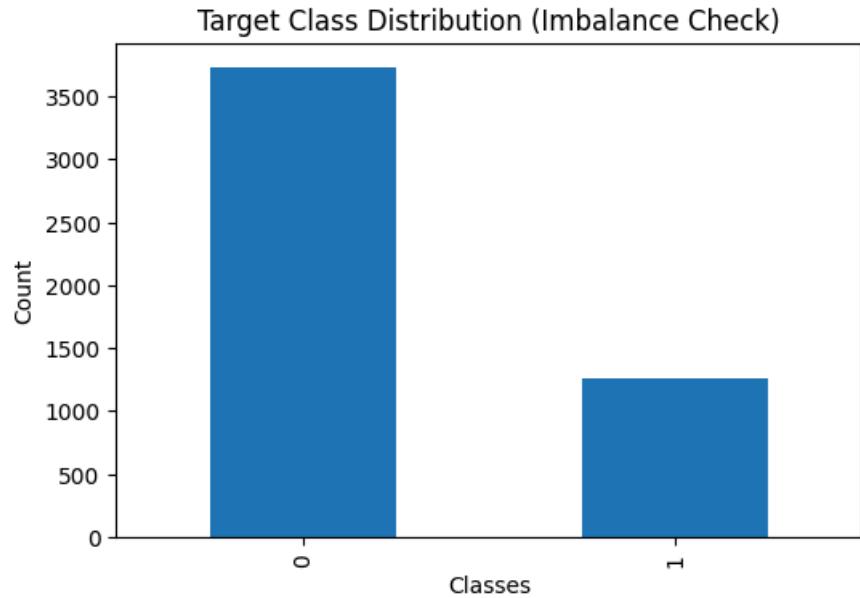
Imbalanced Dataset

- **Are classes balanced?**

No, the dataset is imbalanced. One class "No career change" has more instances than the other.

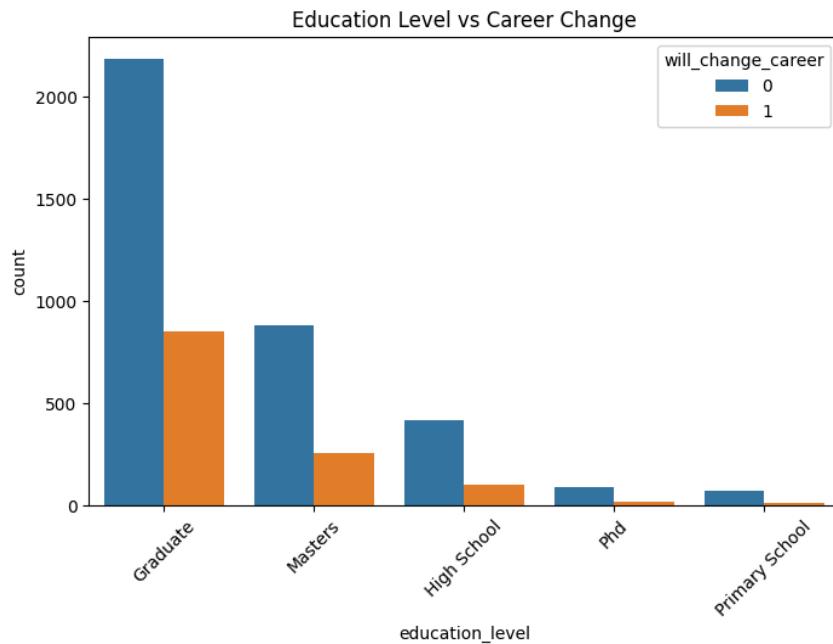
- **Representation:**

A bar chart shows a clear imbalance between the two classes.

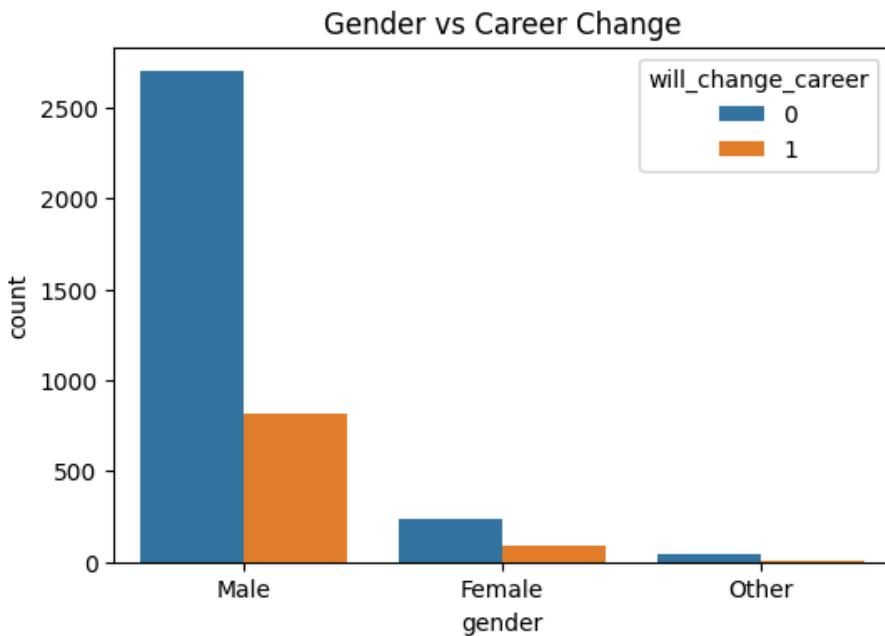


Exploratory Data Analysis (EDA)

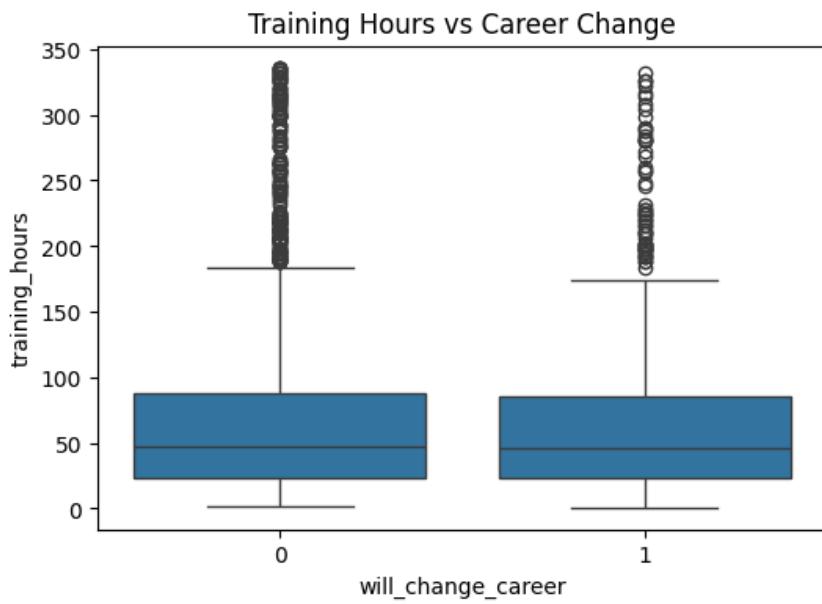
- **Education Level vs Career Change:** People with higher education show more career-switching interest.



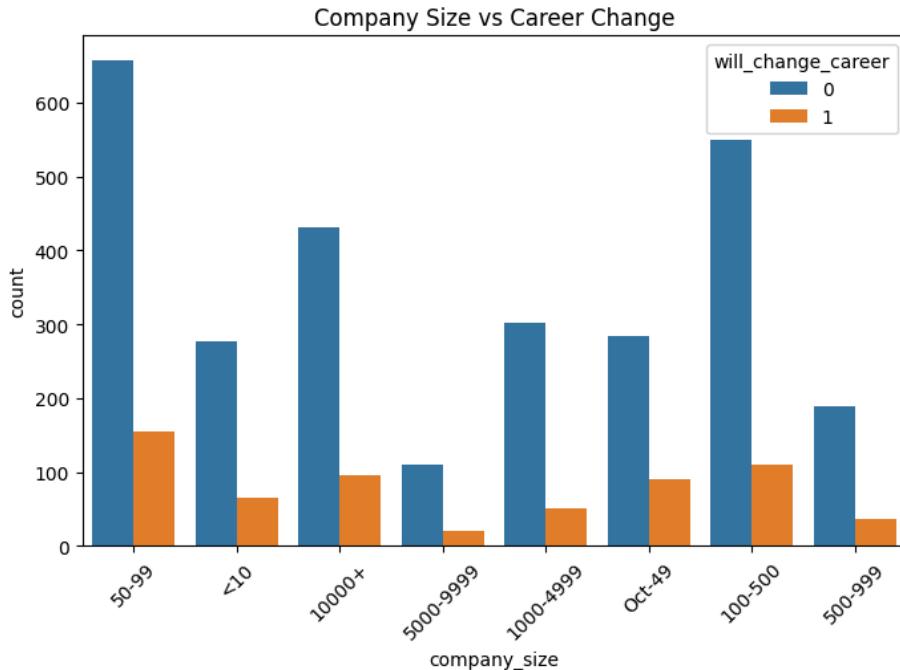
- **Gender vs Career Change:** Slight imbalance in gender distribution; both genders consider switching but male entries dominate.



- **Training Hours vs Career Change:** People with higher training hours are more likely to switch careers.



- **Company Size vs Career Change:** Smaller companies show a higher proportion of switches compared to large firms.



Dataset Pre-processing

- **Problem 1: Null / Missing values**

Some features had missing values (e.g., gender, company_type).

Solution: We imputed them with the most frequent value for categorical and median for numerical.

- **Problem 2: Categorical values**

Many categorical features (e.g., education_level, gender).

Solution: Used OneHotEncoder to convert them into numeric form.

- **Problem 3: Feature scaling**

Quantitative features had different ranges.

Solution: Applied StandardScaler to normalize numeric features.

Dataset Splitting

- **Type:** Stratified split to preserve target distribution.

- **Train/Test ratio:** 80% Train, 20% Test.

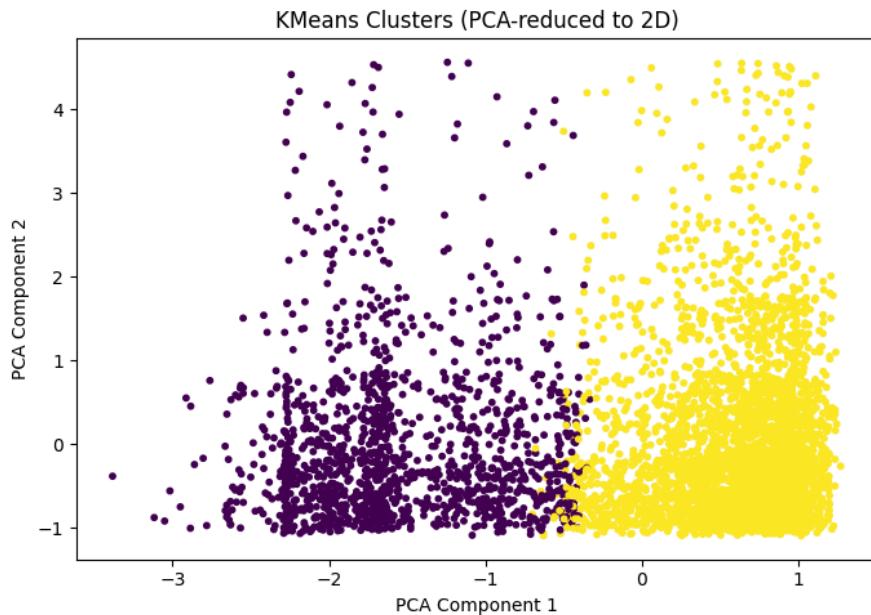
- **Validation:** Used early stopping inside the Neural Network for validation.

Model Training & Testing (Supervised)

- **Models Applied:**
 - Logistic Regression (classification)
 - Naive Bayes (classification)
 - Neural Network (classification)
 - Also KMeans clustering for unsupervised
- **Other mentioned models (KNN, Decision Tree, Linear Regression)** were tested in earlier versions but final analysis is focused on 3 main models.

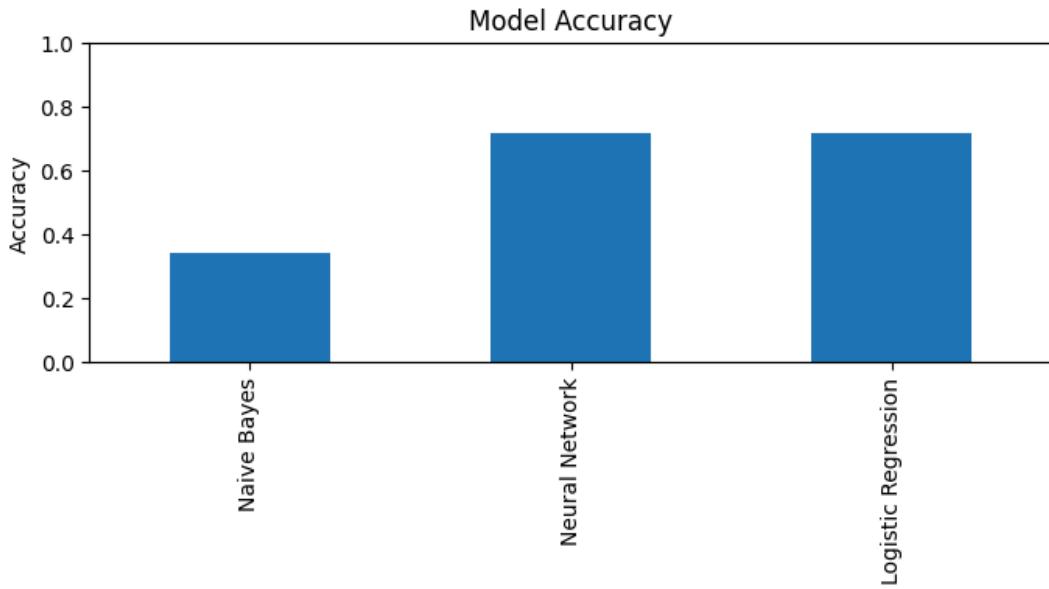
Unsupervised Learning

- Applied **KMeans clustering** with k=2.
- The clusters partially overlapped with true labels but were not perfect, confirming that unsupervised clustering is less effective than supervised models.

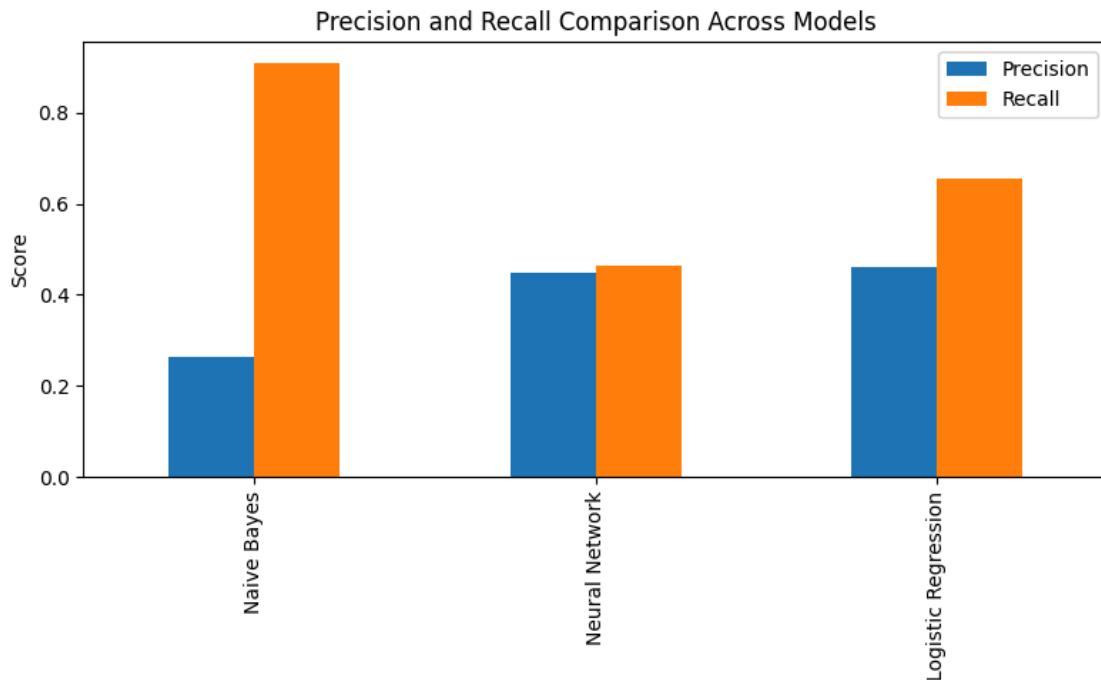


Model Selection / Comparison Analysis

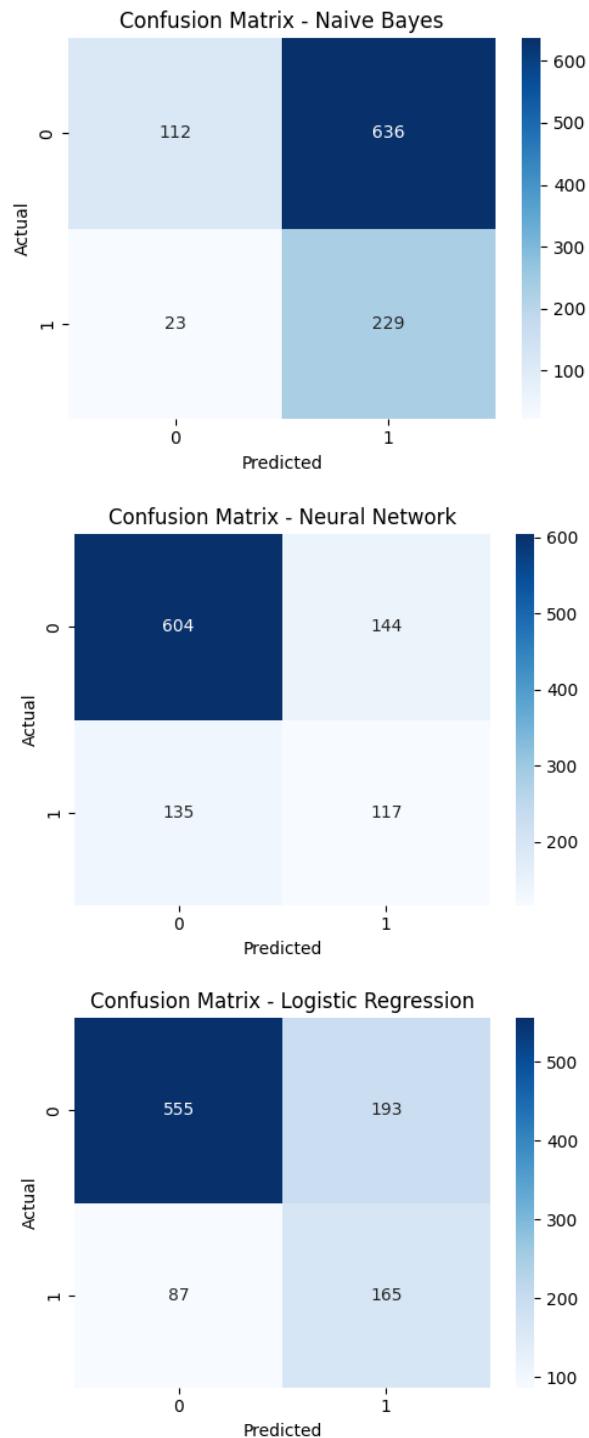
- **Accuracy:** Logistic Regression and Neural Network performed better than Naive Bayes.



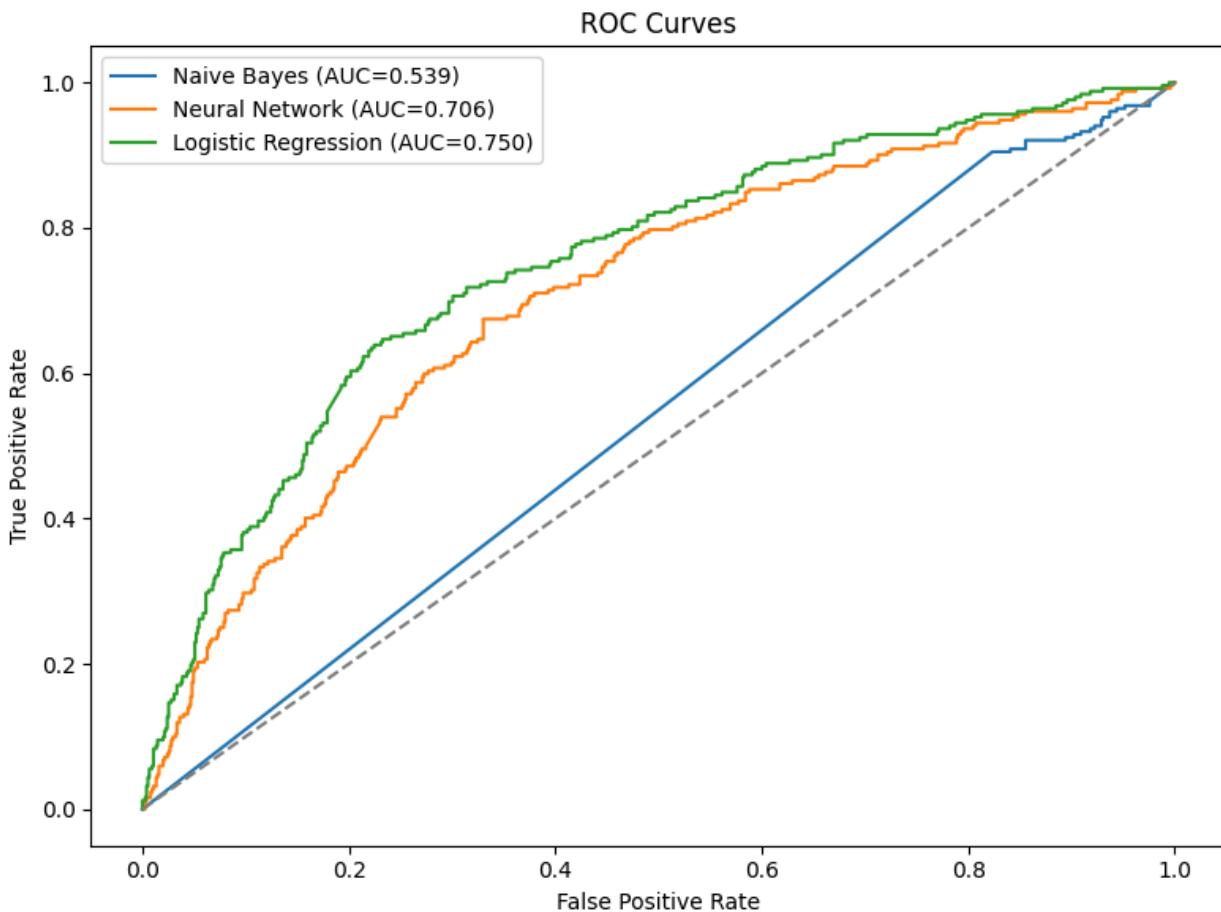
- **Precision & Recall:** Neural Network balanced precision and recall best.



- **Confusion Matrix:** Showed more false negatives in Naive Bayes, while Logistic Regression and Neural Network captured positives better.



- **AUC/ROC Curve:** Neural Network and Logistic Regression achieved higher AUC scores (>0.75).



- **Regression metrics (R^2 , Loss):** Not applicable since this is a classification problem.

Conclusion

- **Best Model:** Logistic Regression showed the best balance of accuracy, precision, recall, and AUC, followed closely by Neural Network.
- **Insights:** Career switching is influenced by multiple weakly correlated features (education, training hours, company size).

- **Challenges:**
 - Imbalanced dataset caused some models (like Naive Bayes) to misclassify the minority class.
 - Weak correlations made it hard to achieve perfect accuracy.
- **Why results look this way:**

The problem is complex and cannot be solved with simple linear relationships, so models like Neural Networks capture non-linear patterns better.