Feature Importance Analysis: Upskilling vs. Reskilling Programs

*Analysis Date: March 2025*

# Introduction

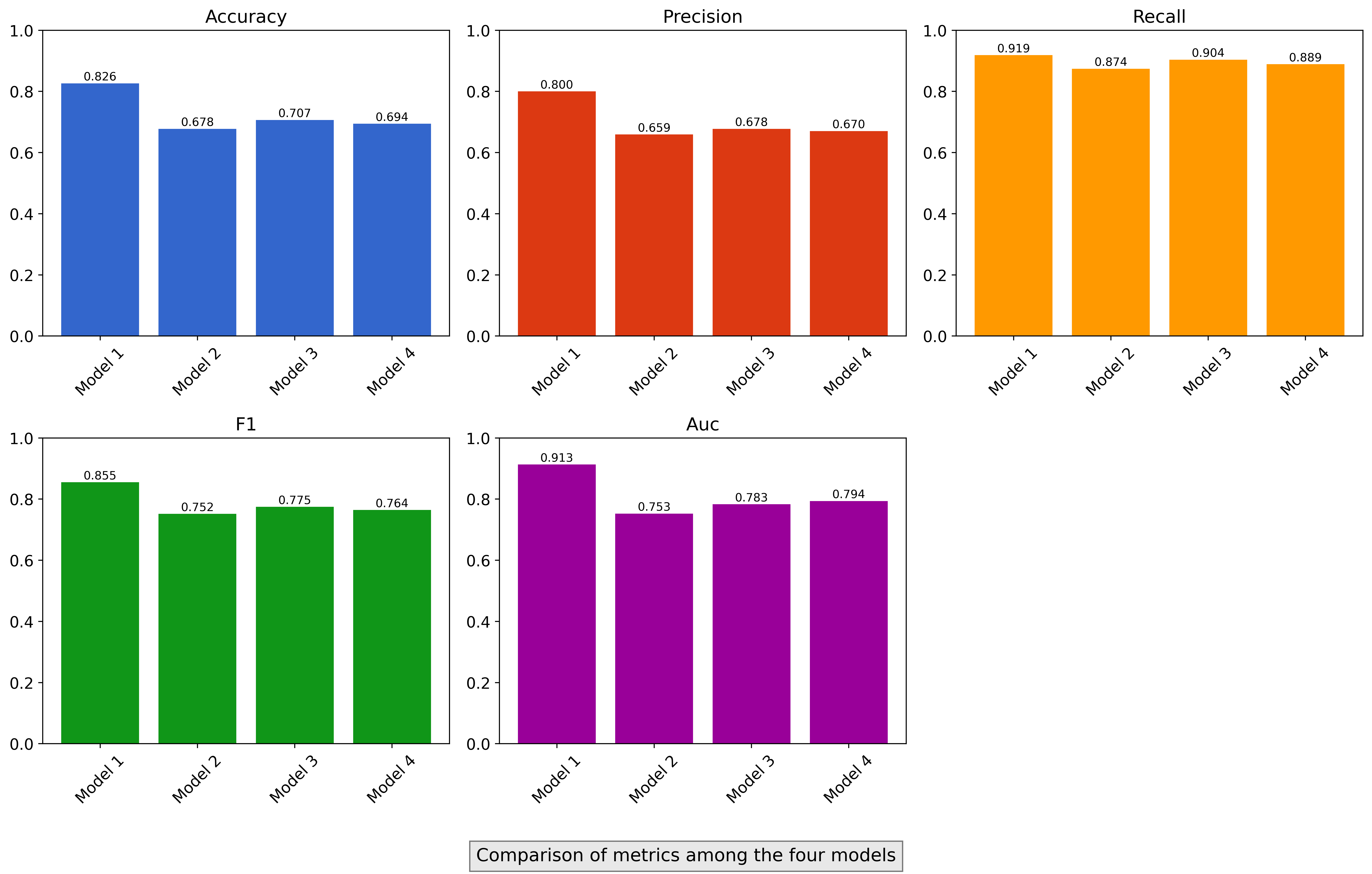
This report presents an analysis of the key features that distinguish between upskilling and reskilling programs. Using machine learning techniques, we identified the most important characteristics that differentiate these program types, which can help organizations design more effective training initiatives.

# Executive Summary

The analysis tested four different modeling approaches to identify the most reliable features that differentiate upskilling and reskilling programs. Key findings include:

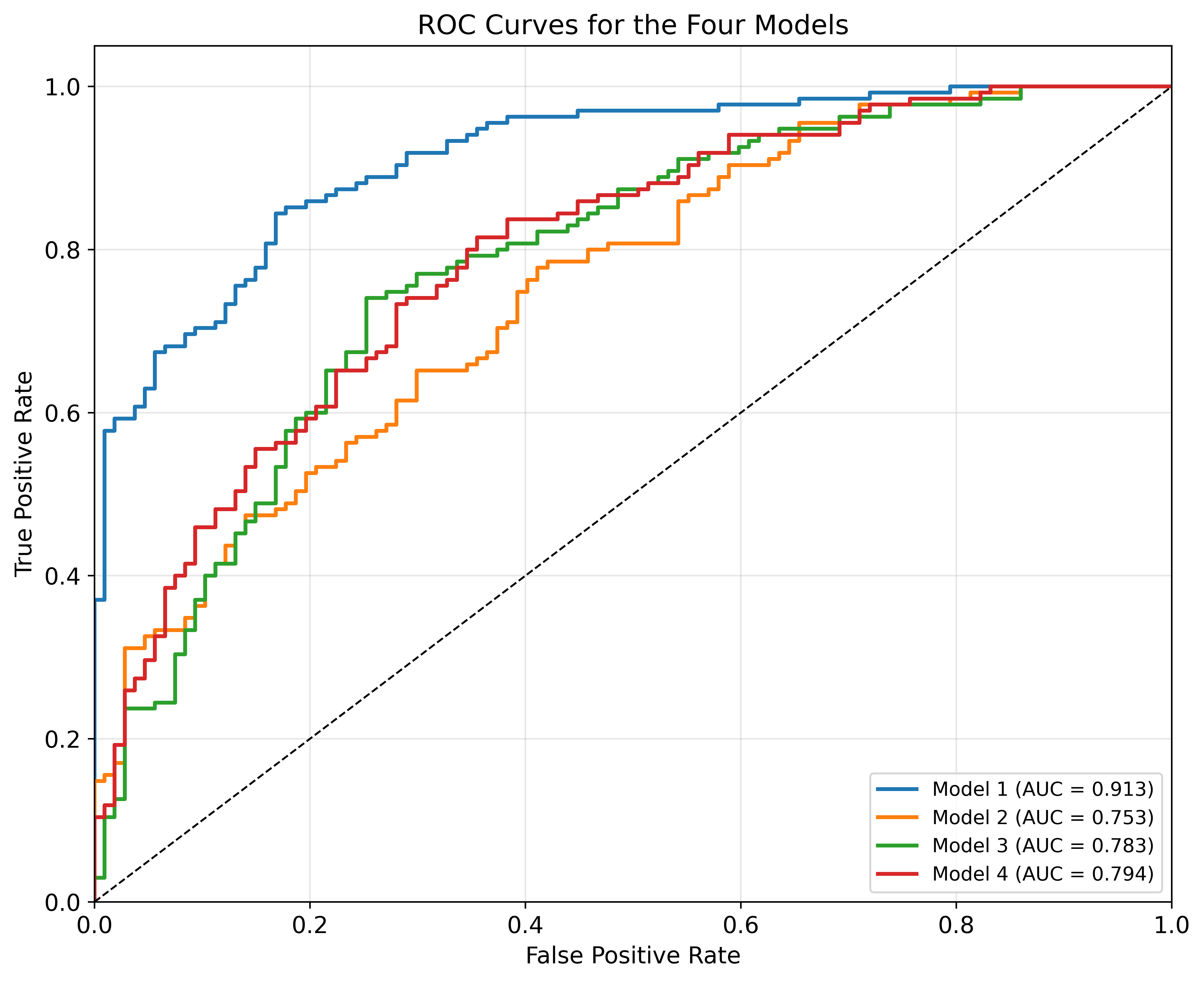
* The model using all variables (including outcomes) achieved the highest overall performance.
* Program length, job placement focus, and management targeting are among the most important differentiating features.
* Categorical variable encoding showed different results than using dummy variables, indicating the importance of encoding choices.
* The inclusion of outcome variables improved model performance, suggesting that program outcomes are strongly linked to program type.

# Model Performance



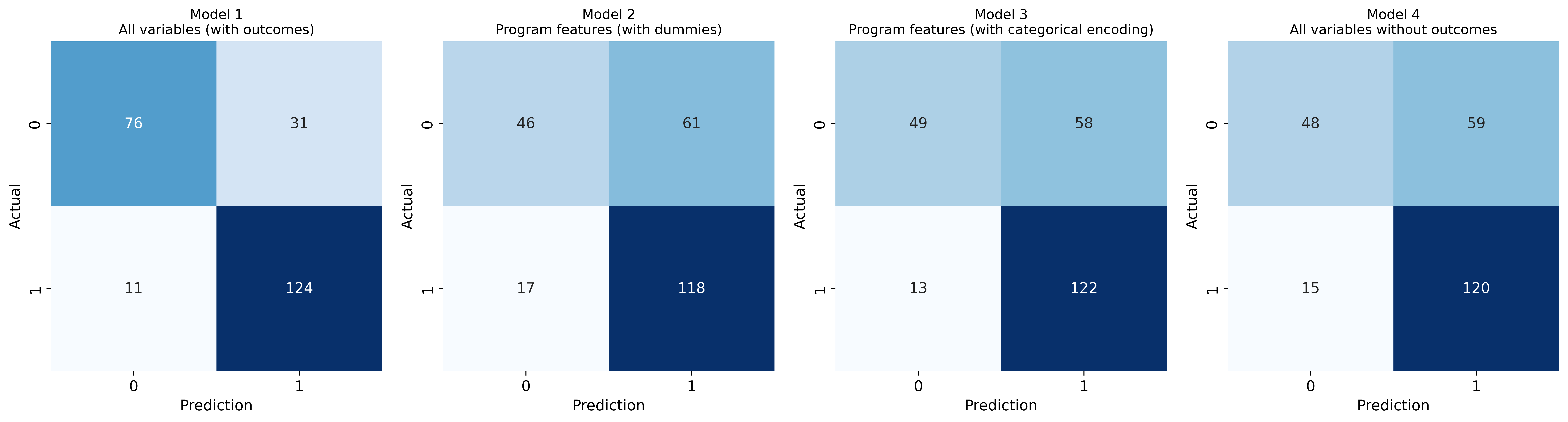
Comparison of performance metrics (accuracy, precision, recall, F1, AUC) across all four models.

## ROC Curve Analysis



ROC curves showing the trade-off between true positive rate and false positive rate for all models.

## Confusion Matrices



Confusion matrices for all four models showing true/false positives and negatives.

# Top Distinguishing Features

The analysis identified the following key features that best distinguish between upskilling and reskilling programs:



Top 10 most important features for distinguishing between upskilling and reskilling programs.

## Key Feature Interpretation

**Program Length:** Reskilling programs tend to be longer in duration than upskilling programs.

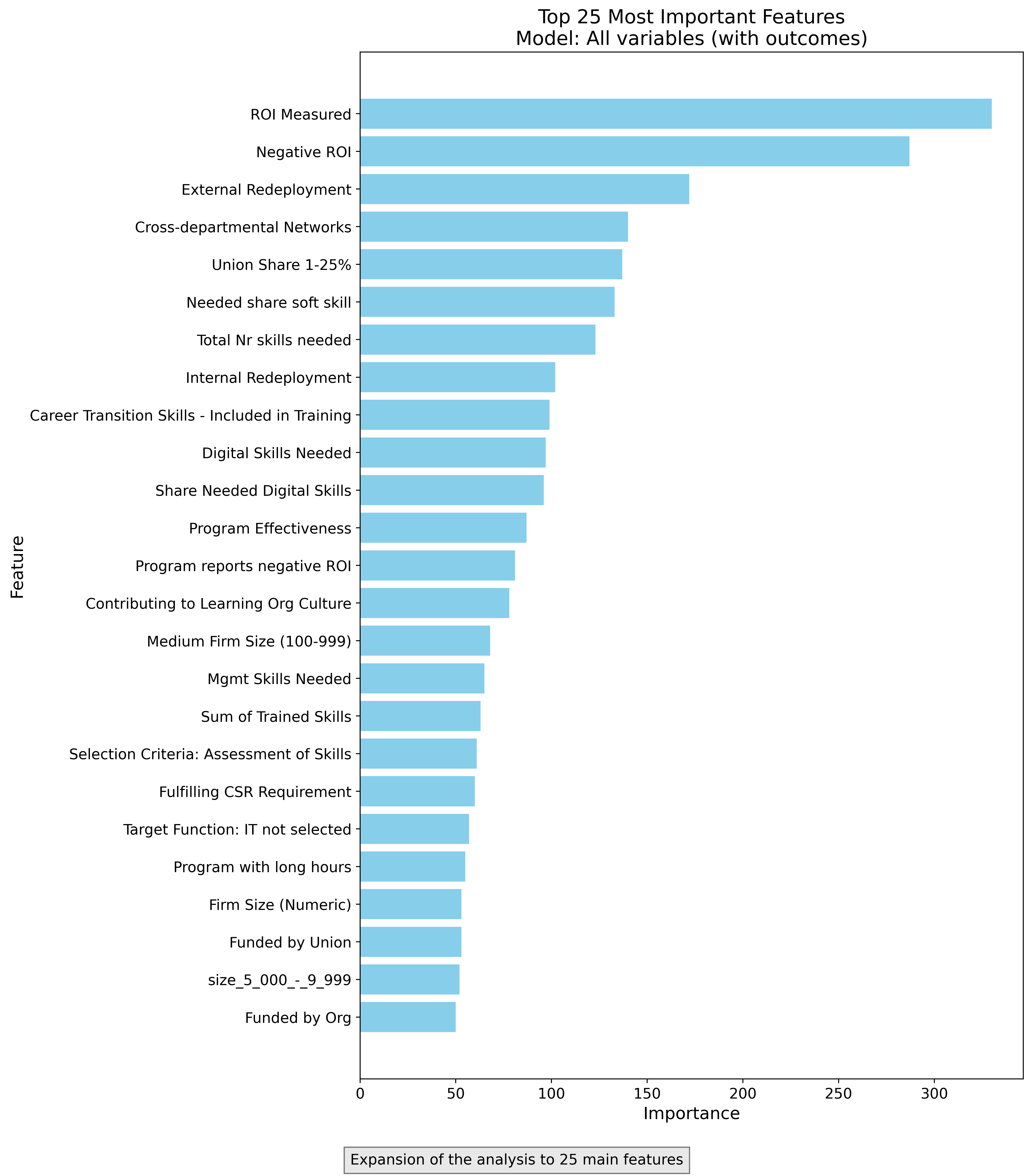
**Job Placement Focus:** Reskilling programs have stronger emphasis on helping participants find new jobs.

**Management Targeting:** Reskilling programs more often target management levels.

**Funding Source:** Upskilling programs are more often funded by the organization itself.

**Program Effectiveness:** Reskilling programs tend to have higher self-reported effectiveness ratings.

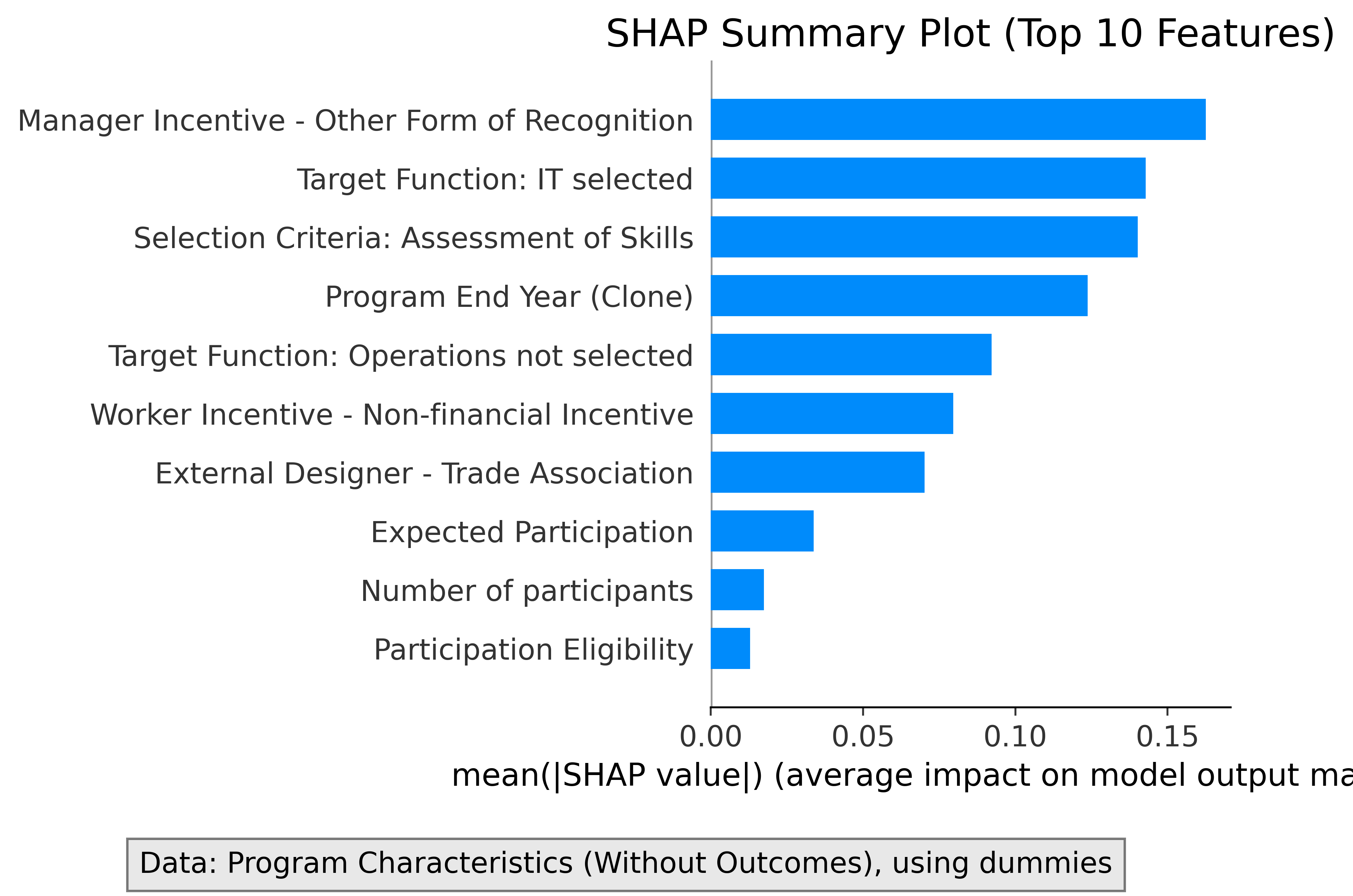
## Extended Feature Importance



Expanded view of the top 25 most important features from the best-performing model.

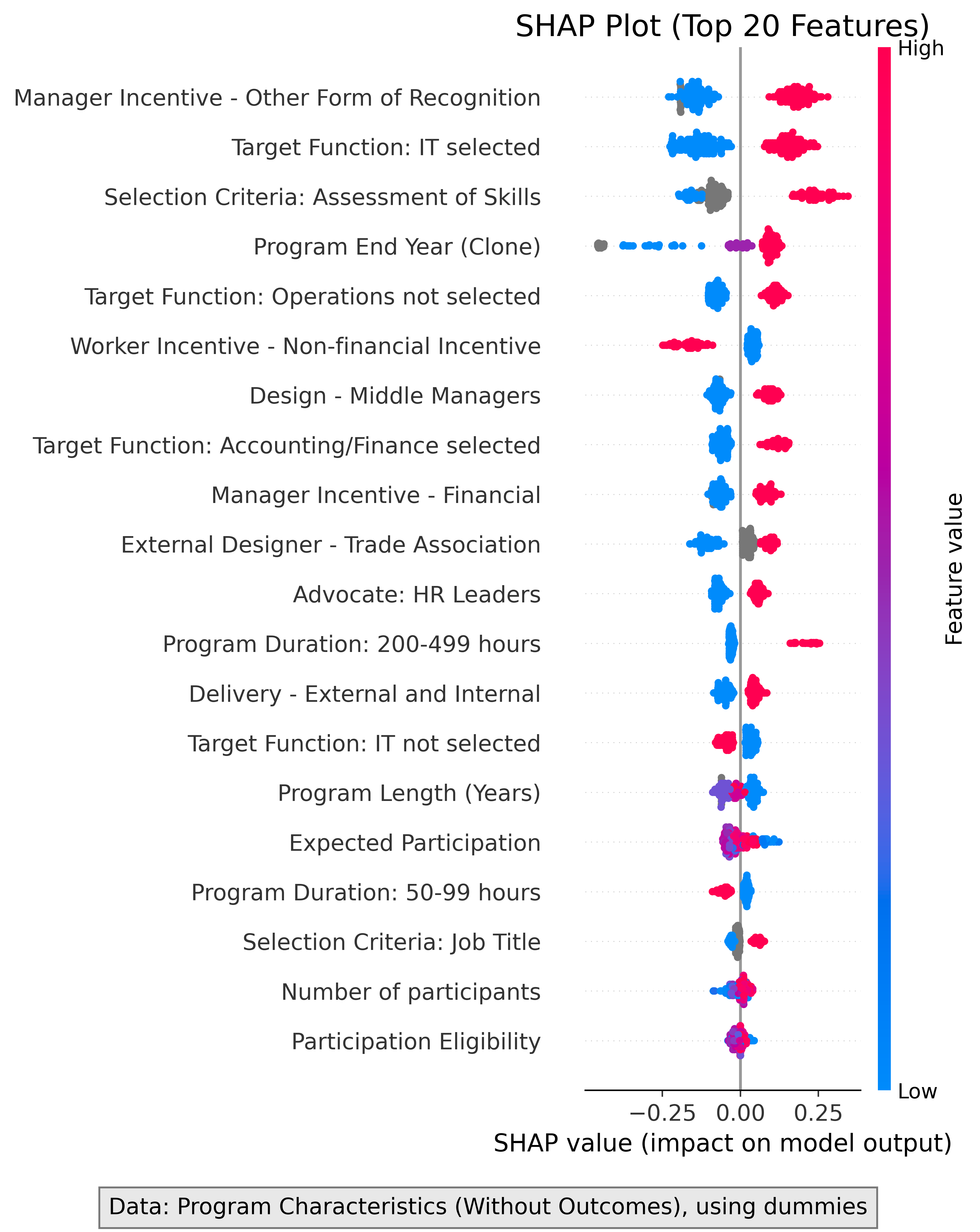
# SHAP Analysis

SHAP (SHapley Additive exPlanations) values help us understand how each feature contributes to predictions for individual programs:



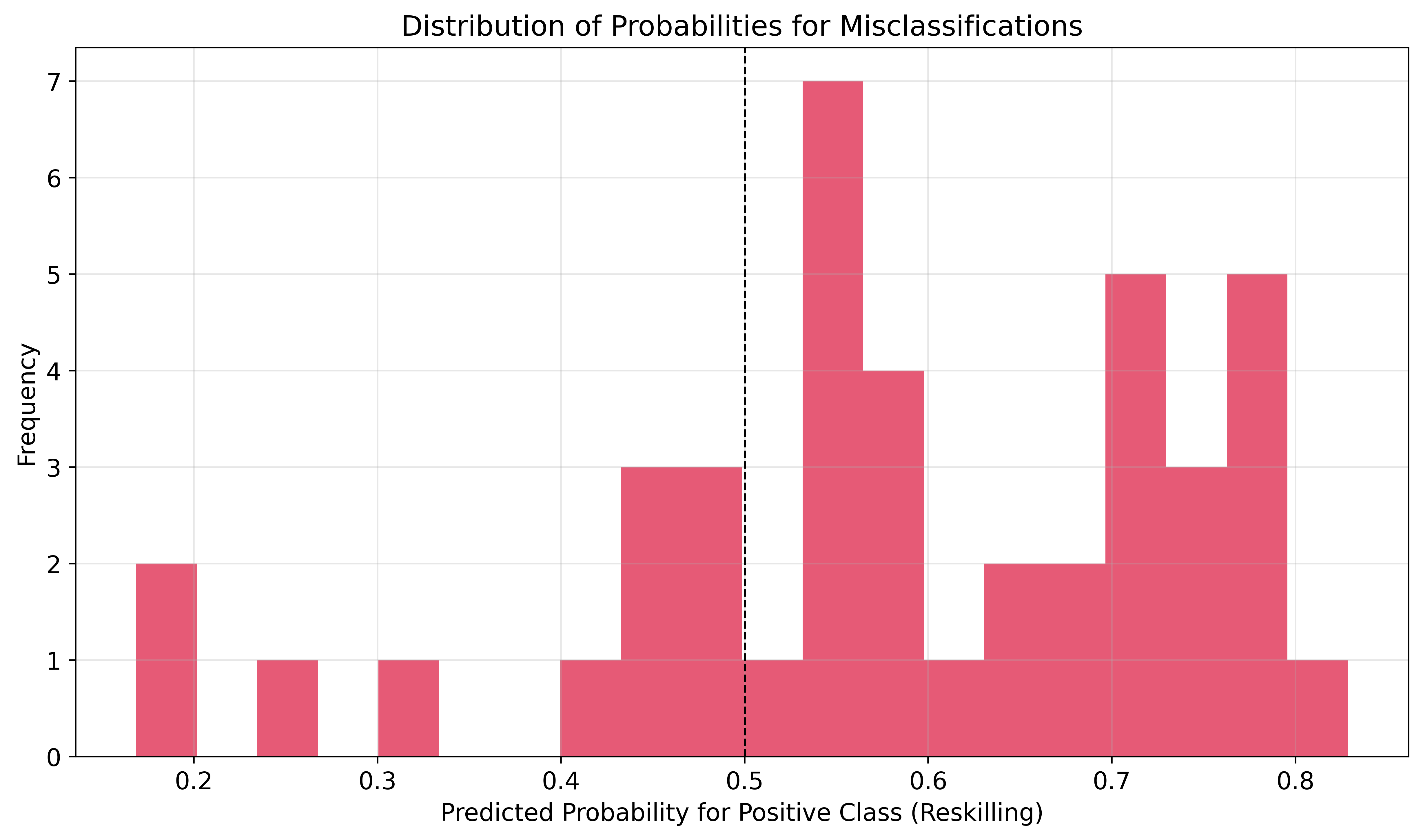
SHAP summary plot showing the impact of features on model predictions.

## SHAP Feature Interactions



SHAP beeswarm plot showing how each feature affects individual predictions.

# Misclassification Analysis



Distribution of predicted probabilities for misclassified samples.

This chart shows how confident the model was in its incorrect predictions. Points closer to 0.5 represent more borderline cases where the model was uncertain.

# Conclusions

* The model with the best overall performance (AUC) is: \*\*All variables (with outcomes)\*\*
* The use of categorical variables with encoding resulted in a \*\*better\*\* performance than using dummies:
* AUC with categorical encoding: 0.7832
* AUC with dummies: 0.7526
* Including outcome variables results in an \*\*improvement\*\* of the model:
* AUC with outcome variables: 0.9133
* AUC without outcome variables: 0.7956

# Recommendations

* Organizations should design reskilling programs with longer durations and strong job placement components.
* Upskilling programs should focus on targeted, shorter interventions tied to current organizational roles.
* When analyzing program data, consider both program characteristics and outcomes for the most accurate differentiation.
* Use categorical variables carefully, as their encoding can affect analysis results.
* These distinguishing features should be considered when designing new training programs to ensure alignment with upskilling or reskilling objectives.

# Appendix: Methodology

This analysis used four approaches to identify key features:

* Model 1: All variables including outcomes (highest performance)
* Model 2: Program features only with dummy variables
* Model 3: Program features with categorical encoding
* Model 4: All variables except outcomes

XGBoost was used for all four models with feature importance calculated using the weight method. SHAP values were used to interpret feature contributions to individual predictions.

# Detailed Results

Detailed metrics and feature importance rankings are available in the accompanying Excel file: "Feature\_Importance\_Results.xlsx"