FINAL PROJECT

1

Predict the likelihood of working in the nonprofit sector by sex and education

2

Identify how demographic patterns align with nonprofit employment

Project 1

Dataset and Target Variable

- U.S. Census PUMS microdata used for individual-level analysis
  - Restricted to employed adults age 18 and older
  - Target variable: nonprofit_worker (binary classification)
  - Only about 10–11% of workers are nonprofit employees
  - Strong class imbalance reflects real-world labor patterns

Train/Test Split and Class Balance

- Data split into training (75%) and testing (25%) sets
  - Stratified sampling preserves nonprofit proportions
  - Training data contains ~89% non-nonprofit and ~11% nonprofit workers
  - Test set mirrors real-world class distribution

Data Cleaning and Feature Preparation

- Key demographic features selected: age, education, sex, disability, region
  - Missing numeric values replaced using median imputation
  - Missing categorical values replaced using most common category
  - Categorical variables converted using one-hot encoding
  - Numeric variables standardized to comparable scales

Project 2

Supervised Learning Models

- Supervised learning predicts outcomes using labeled data
  - Logistic Regression used as an interpretable linear baseline
  - Random Forest used to capture nonlinear relationships
  - Model evaluation focused on precision, recall, and F1-score
  - Accuracy avoided due to strong class imbalance

Handling Class Imbalance: Balanced Models

- Default models failed to detect nonprofit workers (recall ≈ 0)
  - Class weighting applied to emphasize minority-class errors
  - Balanced Logistic Regression achieved recall ≈ 0.635
  - Balanced Random Forest achieved recall ≈ 0.666
  - Balanced Random Forest produced the strongest overall F1-score

ROC Curves and Model Discrimination

- ROC curves evaluate performance across all thresholds
  - Logistic Regression AUC ≈ 0.673
  - Random Forest AUC ≈ 0.682
  - Both models perform better than random guessing but remain limited

Project 3

Unsupervised Learning and PCA

- Unsupervised learning explores structure without labels
  - Goal: test whether demographics naturally cluster
  - Principal Component Analysis (PCA) reduced dimensionality
  - 95% of variance retained after PCA
  - PCA improved clustering stability but not label alignment

Clustering Methods and Quantitative Results

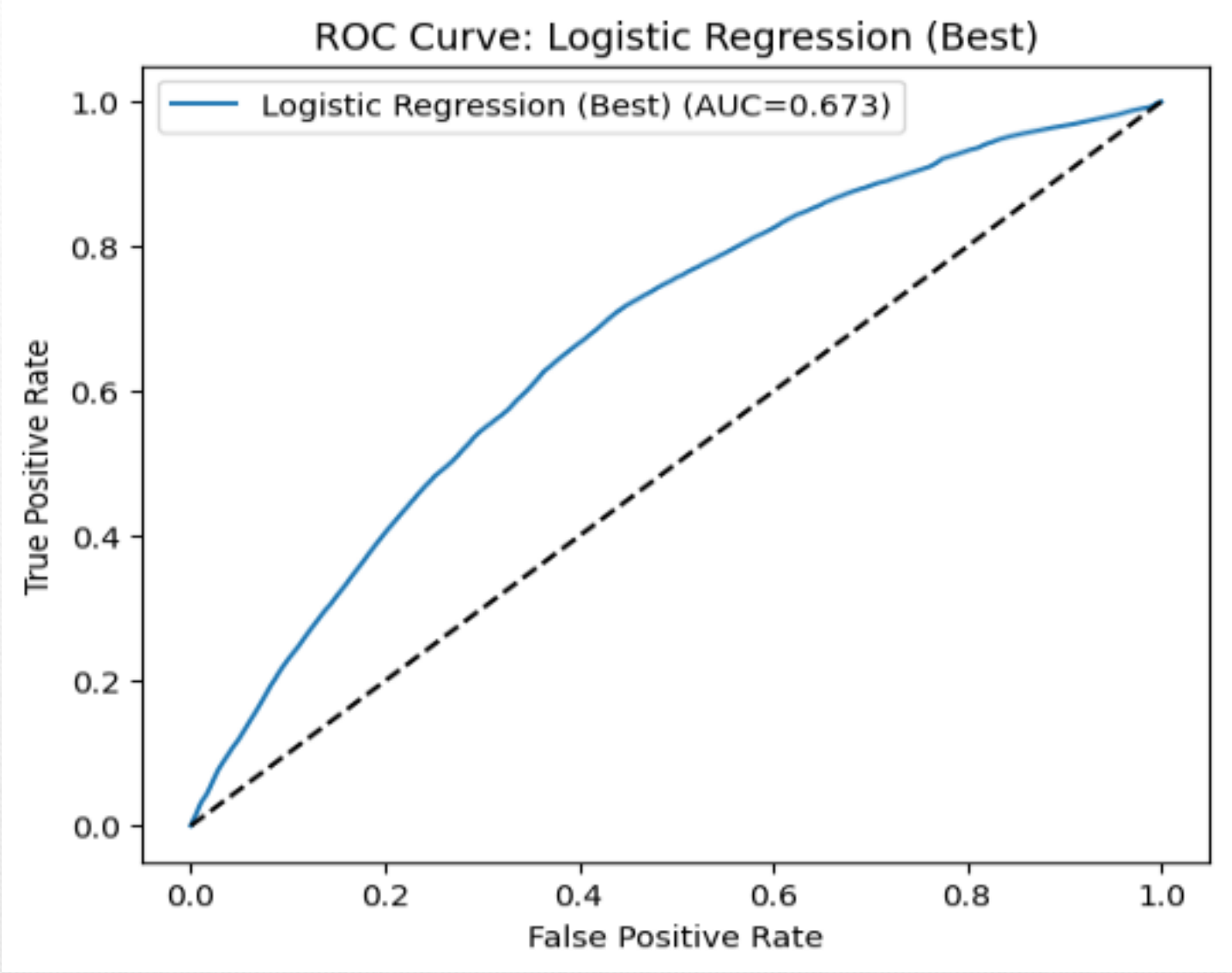
- Three clustering algorithms tested: k-Means, Agglomerative, DBSCAN
  - k-Means achieved the highest silhouette score (~0.33)
  - Adjusted Rand Index values were near zero across all methods
  - PCA slightly improved DBSCAN ARI but not overall conclusions
  - Clusters did not strongly align with nonprofit employment

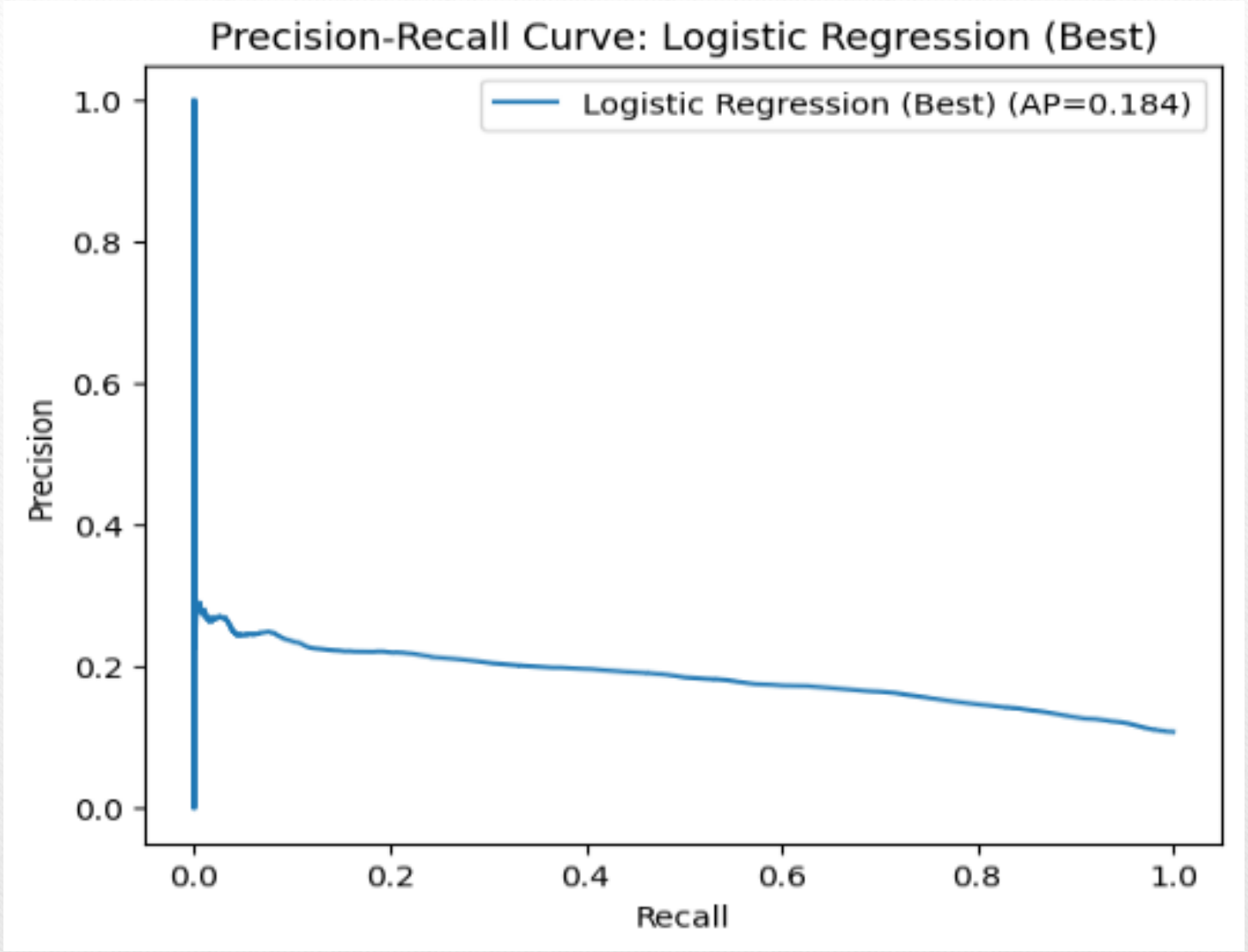
Limitations and Future Work

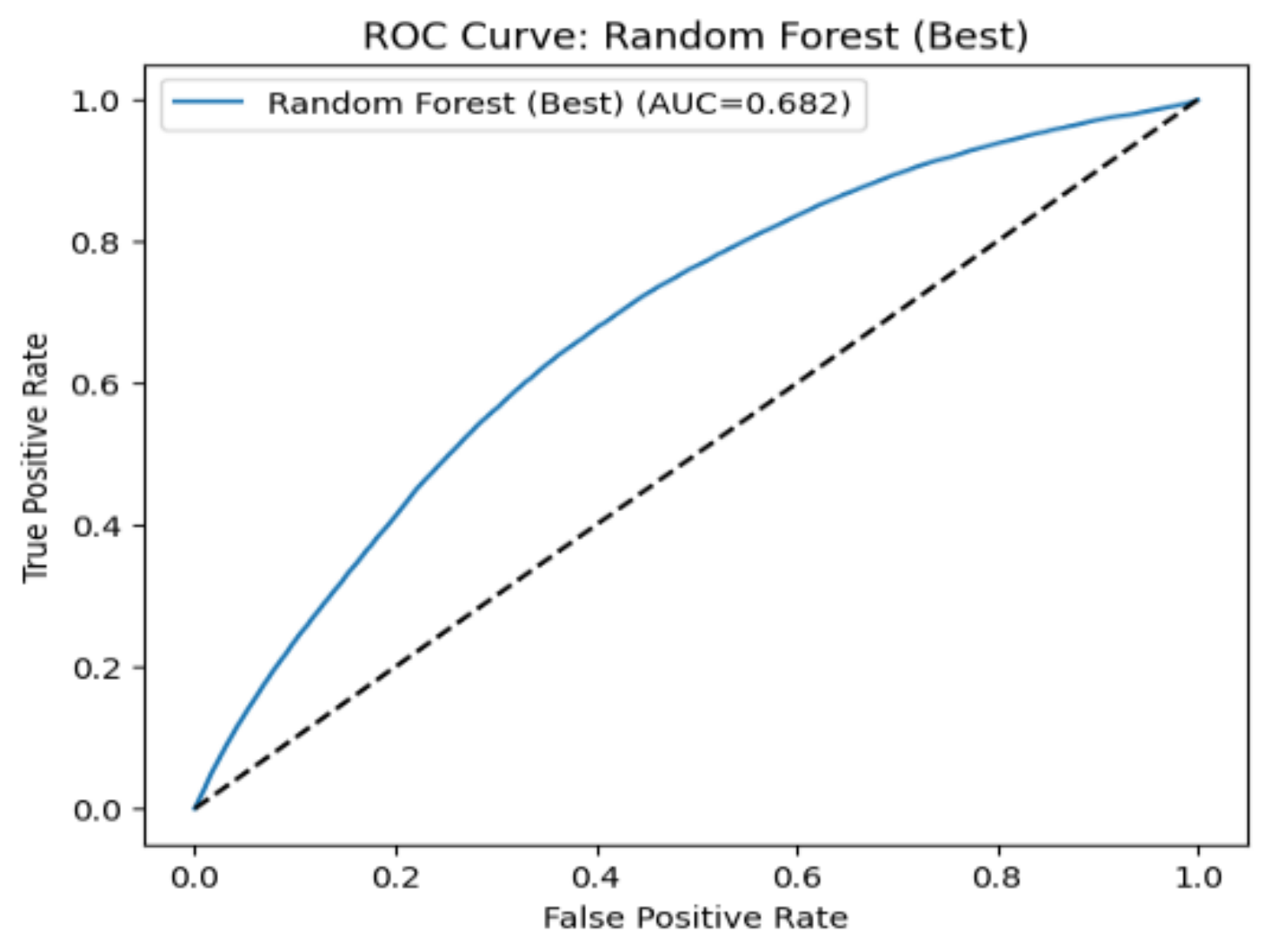
- Features used alone may be insufficient to predict nonprofit employment
  - Strong class imbalance limits predictive performance
  - Clustering results sensitive to feature choice and distance metrics
  - Future work could include income, occupation, or industry variables
  - Additional methods could include resampling or threshold tuning

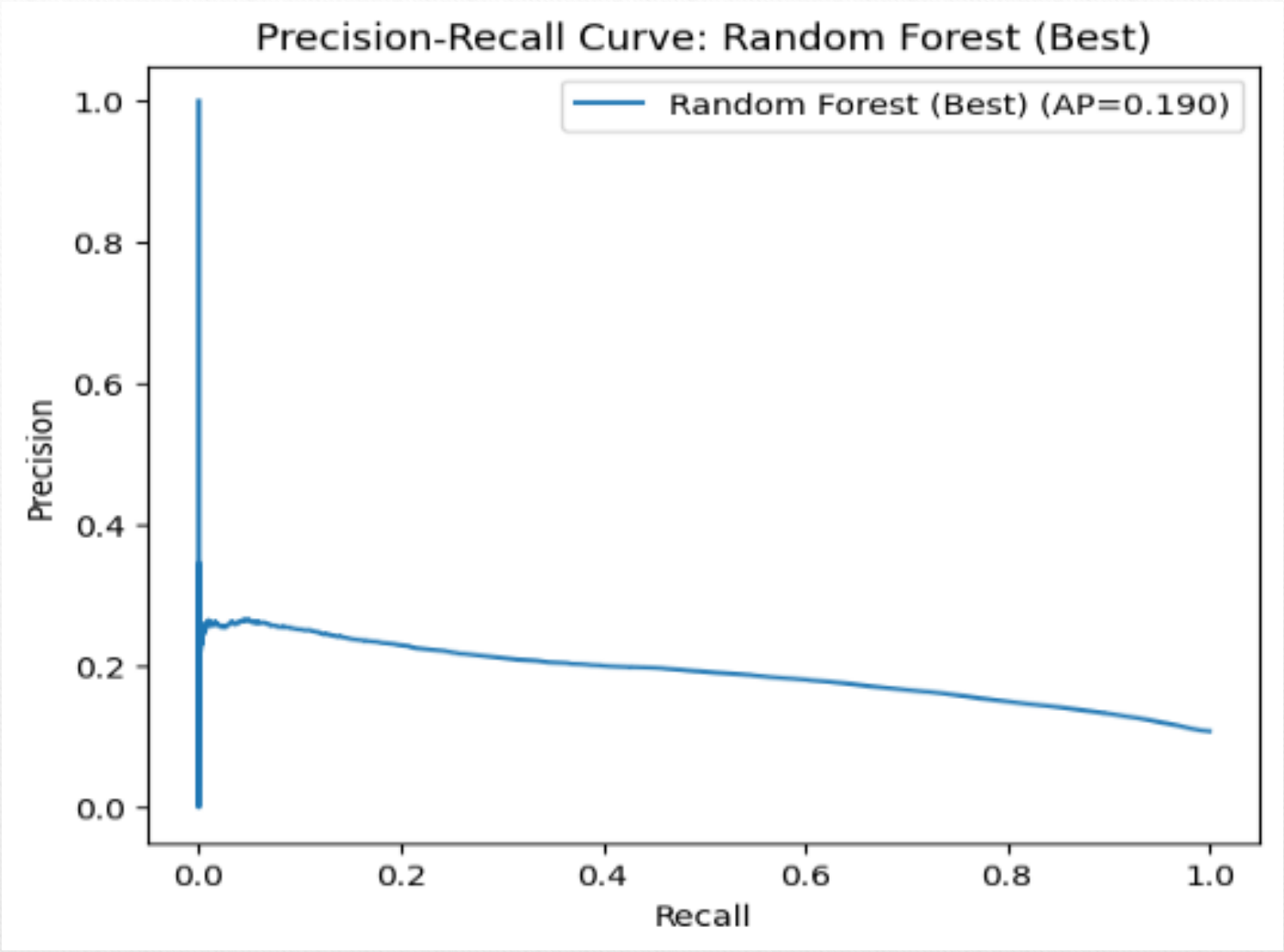
Thank you

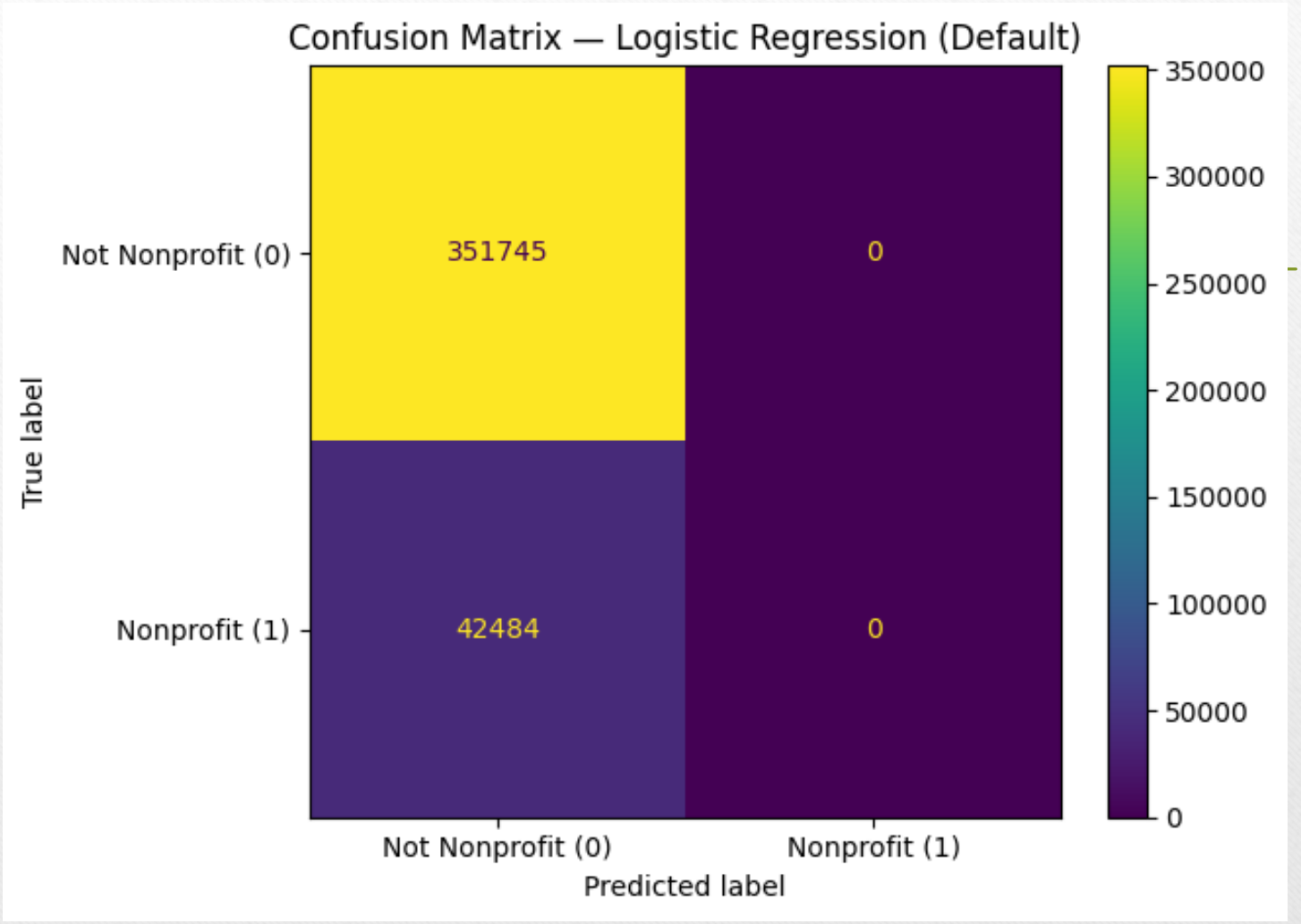
Appendix

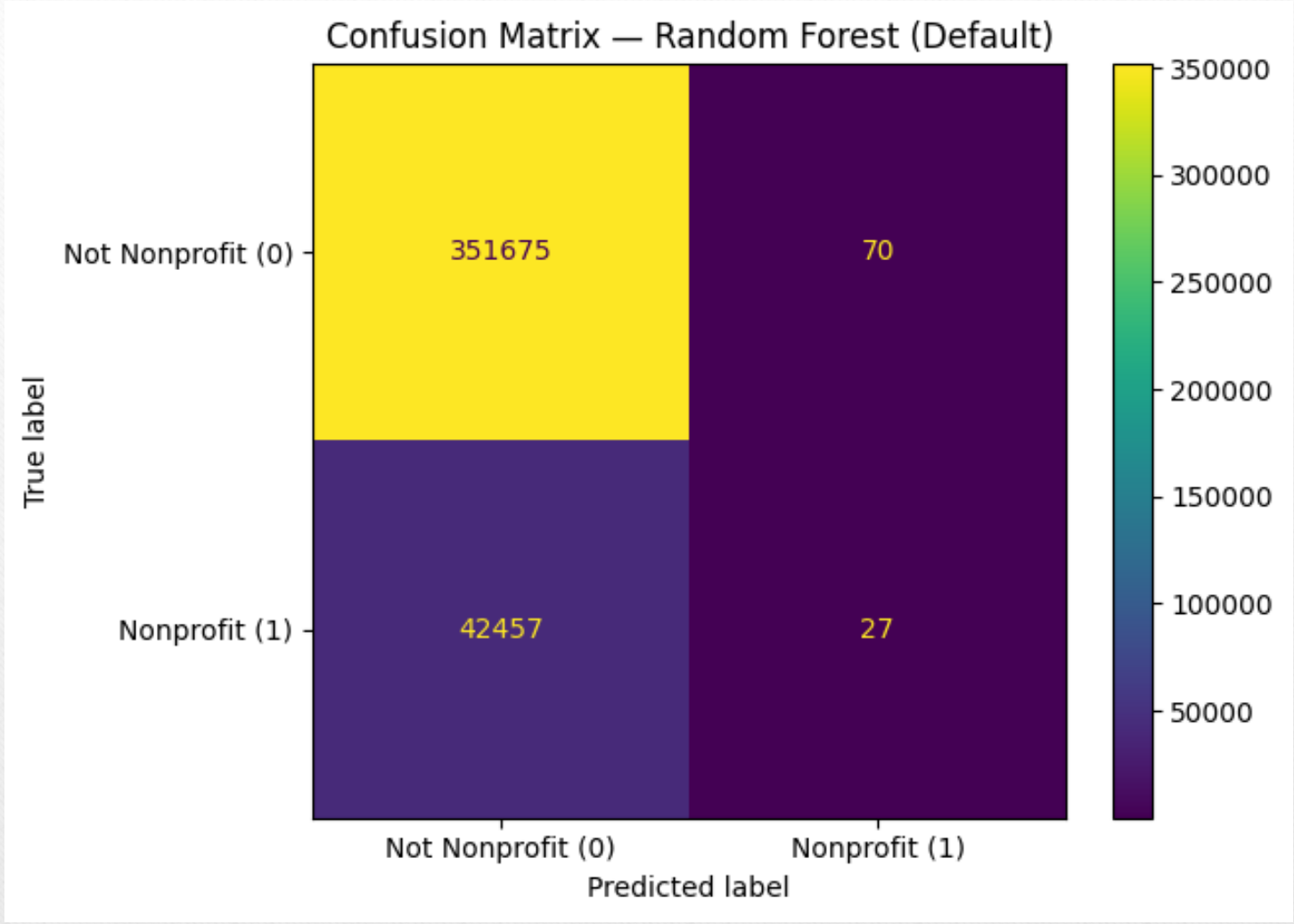

ROC Curve: Logistic Regression (Best)
Logistic Regression (Best) (AUC=0.673)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

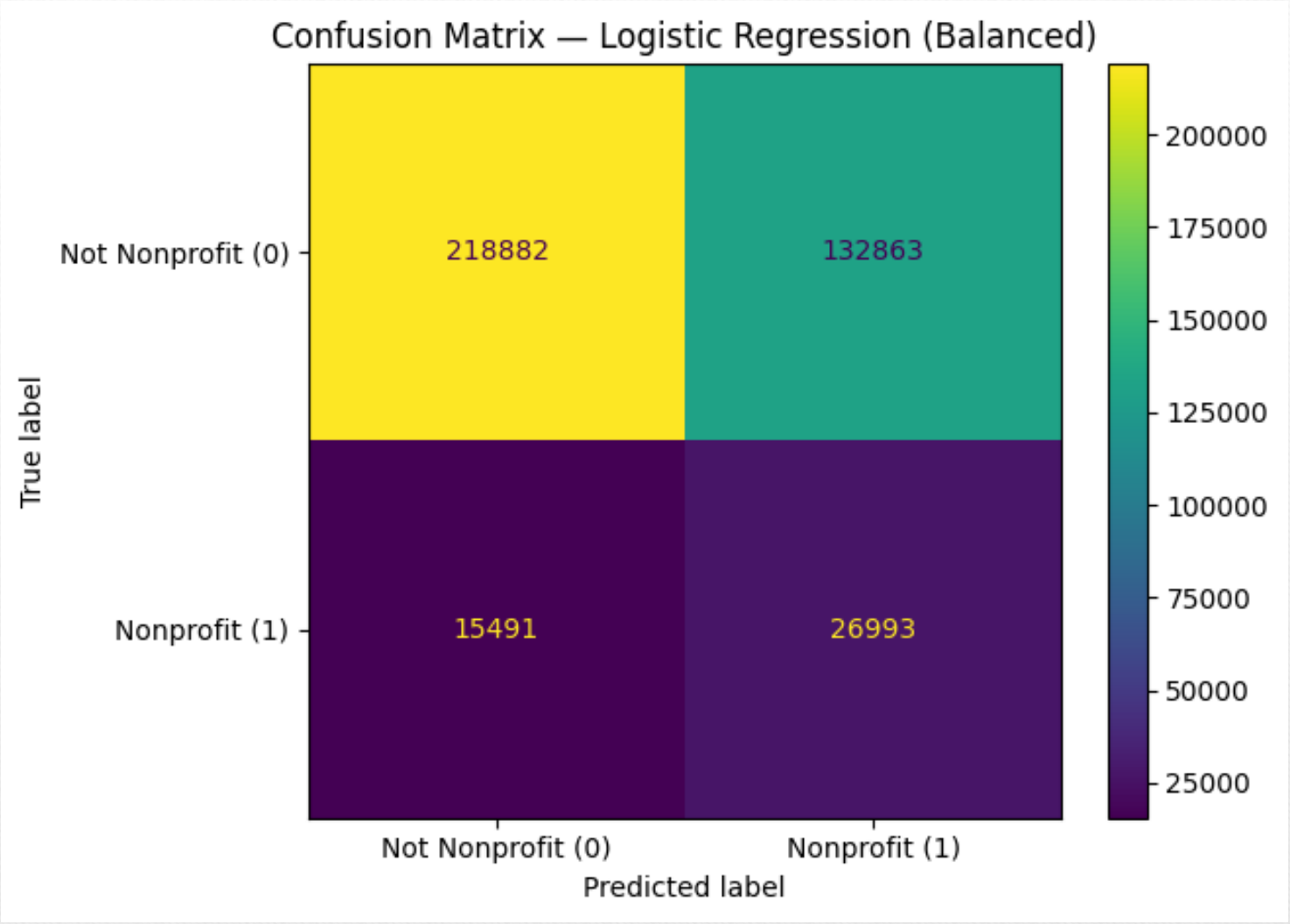

Precision-Recall Curve: Logistic Regression (Best)
Logistic Regression (Best) (AP=0.184)
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

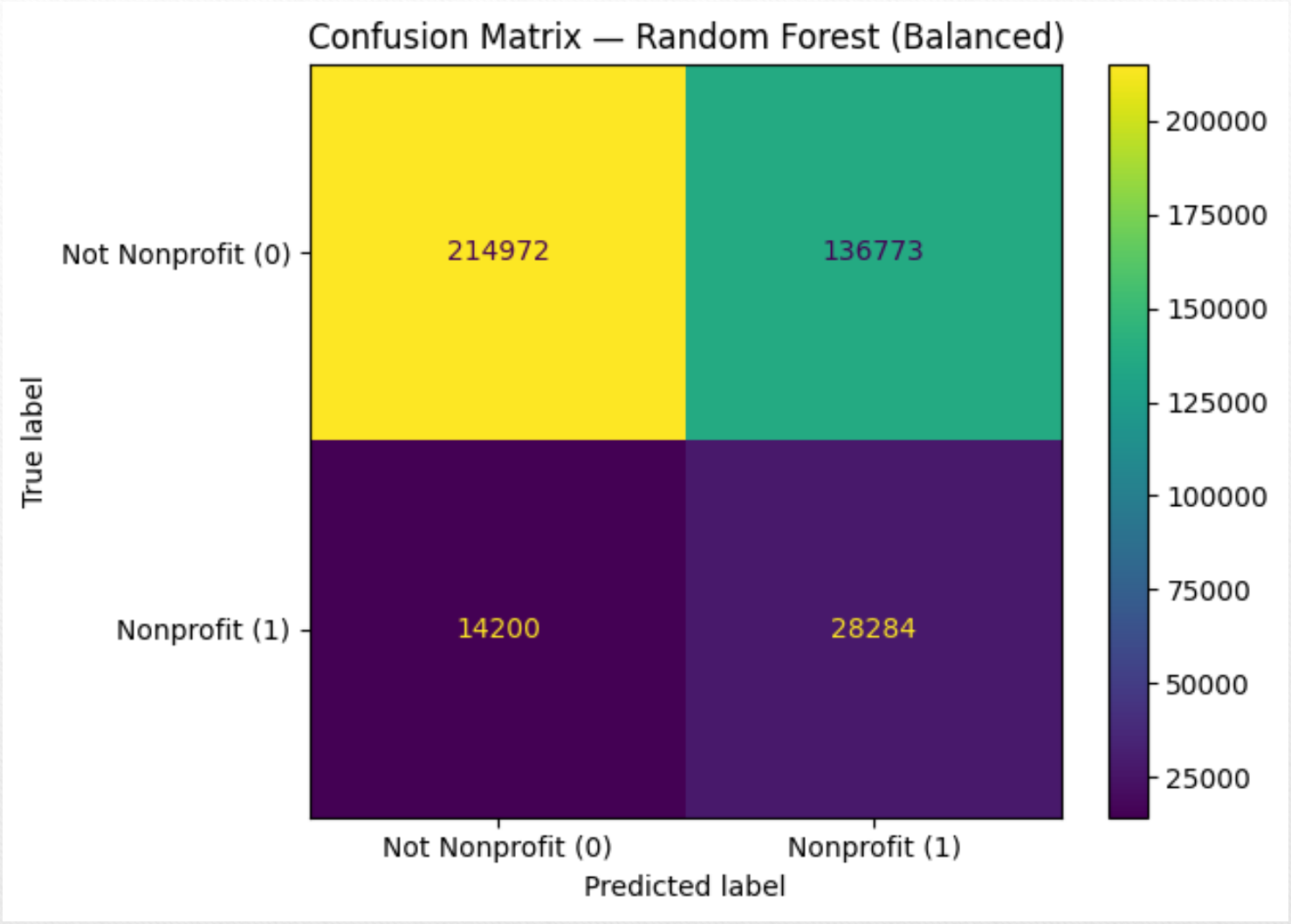

ROC Curve: Random Forest (Best)
Random Forest (Best) (AUC=0.682)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall Curve: Random Forest (Best)
Random Forest (Best) (AP=0.190)
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

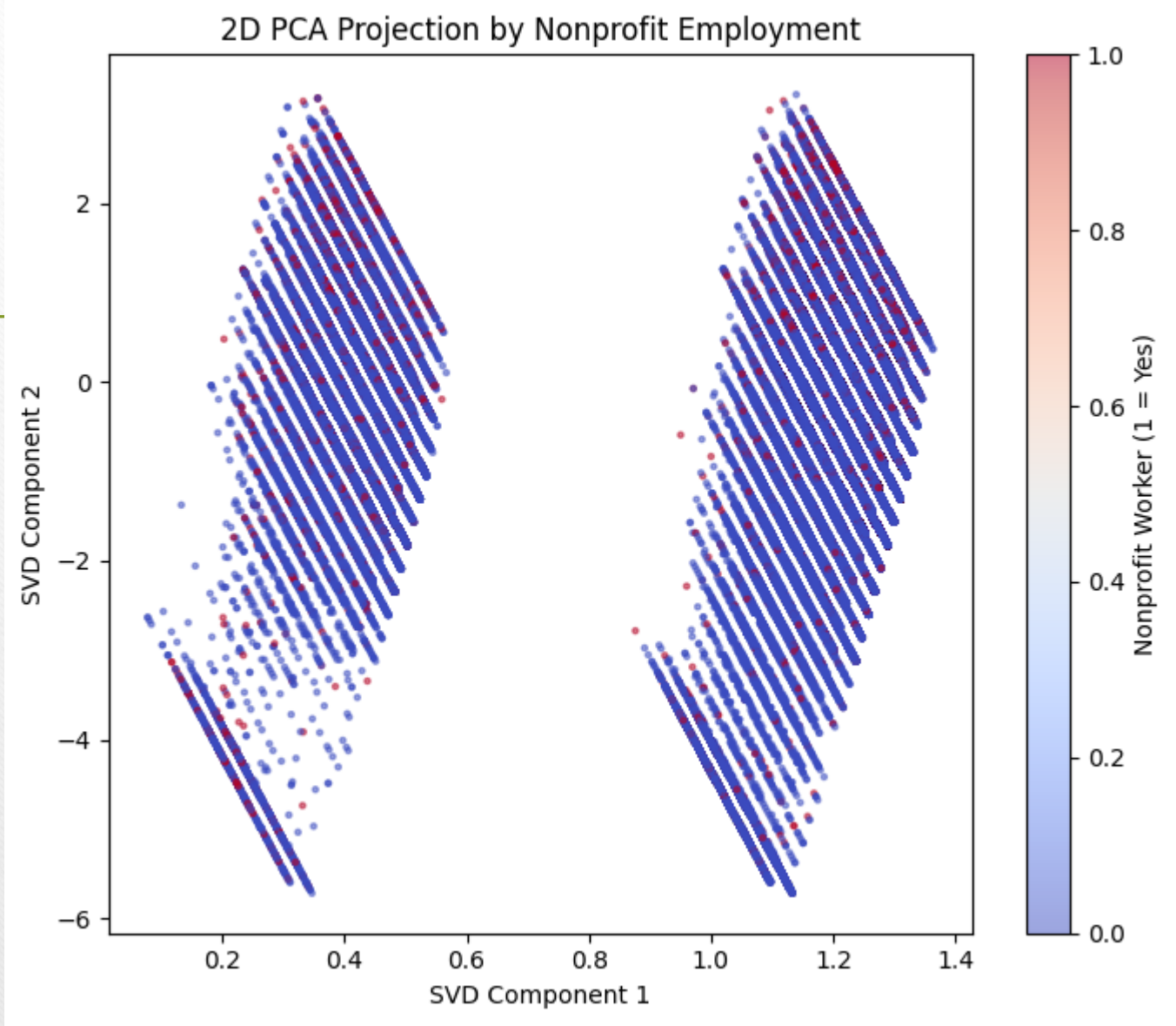

Confusion Matrix — Logistic Regression (Default)
Not Nonprofit (0)
Nonprofit (1)
True label
351745
0
42484
0
Not Nonprofit (0)
Nonprofit (1)
Predicted label
350000
300000
250000
200000
150000
100000
50000
0

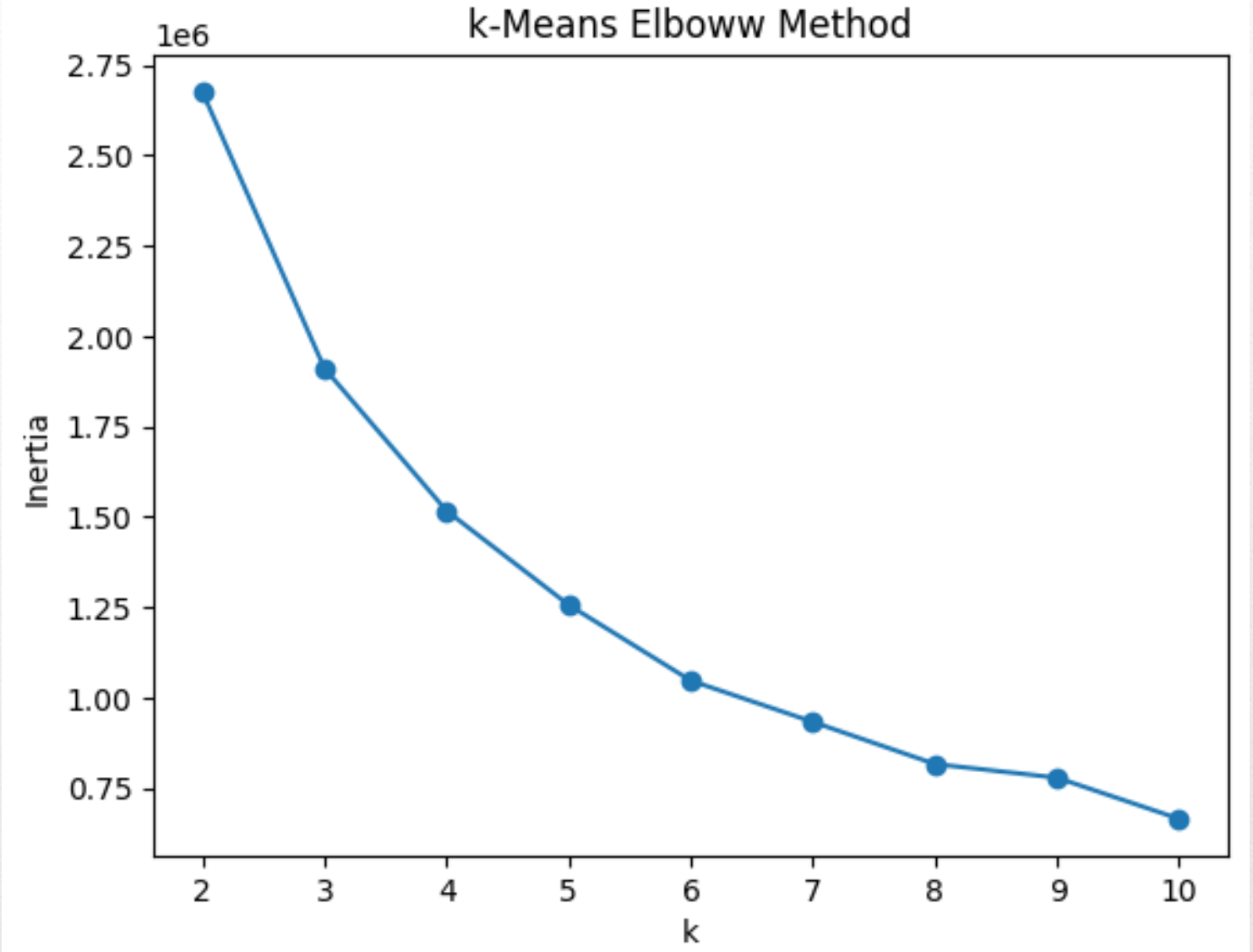

Confusion Matrix — Random Forest (Default)
Not Nonprofit (0)
Nonprofit (1)
True label
351675
70
42457
27
Not Nonprofit (0)
Nonprofit (1)
Predicted label
350000
300000
250000
200000
150000
100000
50000

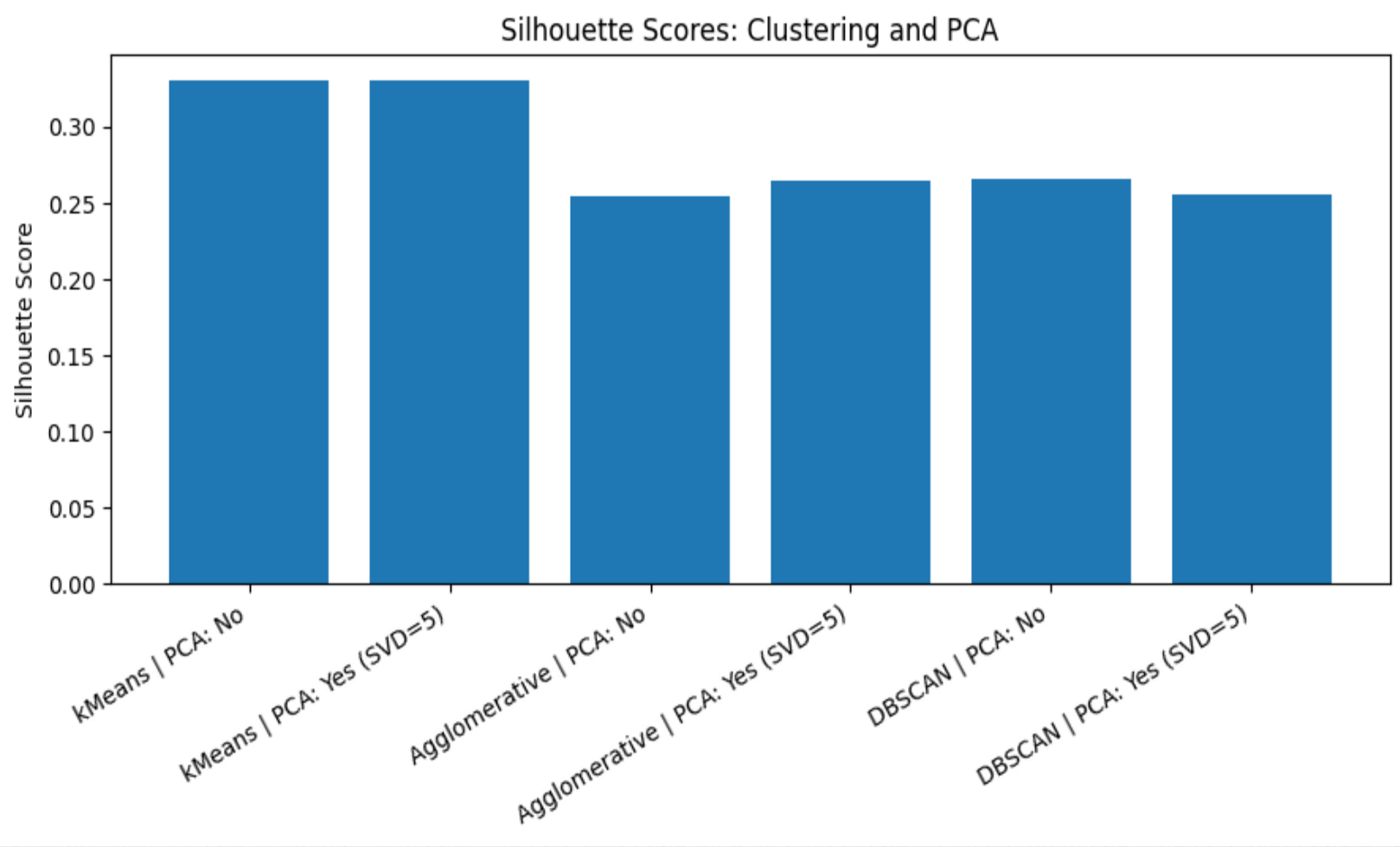

Confusion Matrix — Logistic Regression (Balanced)
Not Nonprofit (0)
Nonprofit (1)
True label
218882
132863
15491
26993
200000
175000
150000
125000
100000
75000
50000
25000
Not Nonprofit (0)
Nonprofit (1)
Predicted label


Confusion Matrix — Random Forest (Balanced)
Not Nonprofit (0)
Nonprofit (1)
True label
214972
136773
14200
28284
Not Nonprofit (0)
Nonprofit (1)
Predicted label
200000
175000
150000
125000
100000
75000
50000
25000


2D PCA Projection of Training Data (All Observations)
SVD Component 2
SVD Component 1
2
0
−2
−4
−6
0.2
0.4
0.6
0.8
1.0
1.2
1.4


2D PCA Projection by Nonprofit Employment
SVD Component 2
SVD Component 1
Nonprofit Worker (1 = Yes)
2
0
−2
−4
−6
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.0
0.8
0.6
0.4
0.2
0.0


k-Means Elboww Method
1e6
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
Inertia
2
3
4
5
6
7
8
9
10
k


Silhouette Scores: Clustering and PCA
Silhouette Score
0.00
0.05
0.10
0.15
0.20
0.25
0.30
kMeans | PCA: No
kMeans | PCA: Yes (SVD=5)
Agglomerative | PCA: No
Agglomerative | PCA: Yes (SVD=5)
DBSCAN | PCA: No
DBSCAN | PCA: Yes (SVD=5)