## LOGISTIC REGRESSION (WITH SENSITIVE FEATURES)

### Chỉ số tối ưu

Threshold = 0.3

Print value: False 465

True 328

Name: probability, dtype: int64

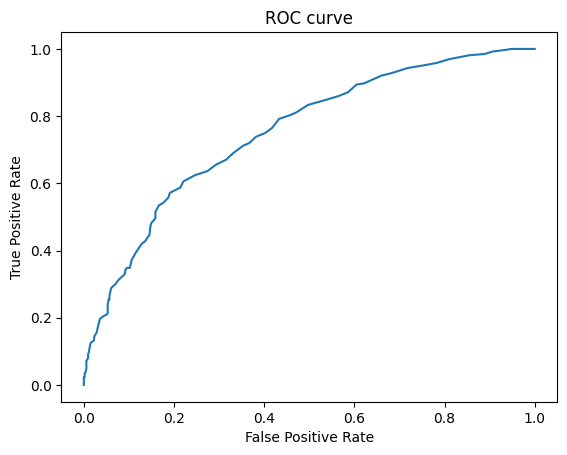
Print value: False 529

True 264

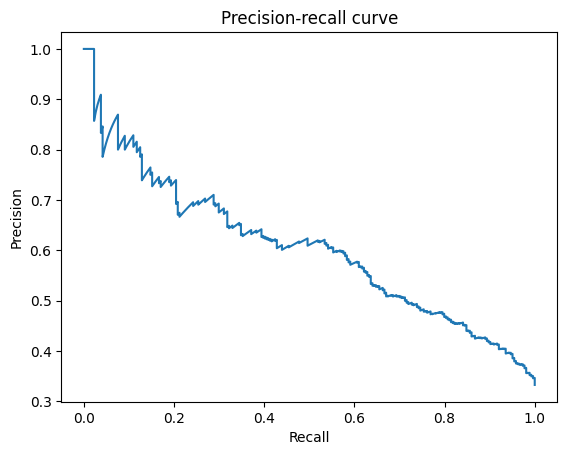
Name: decision, dtype: int64

F1 score: 0.5844594594495869

Accuracy: 0.6897856242118537



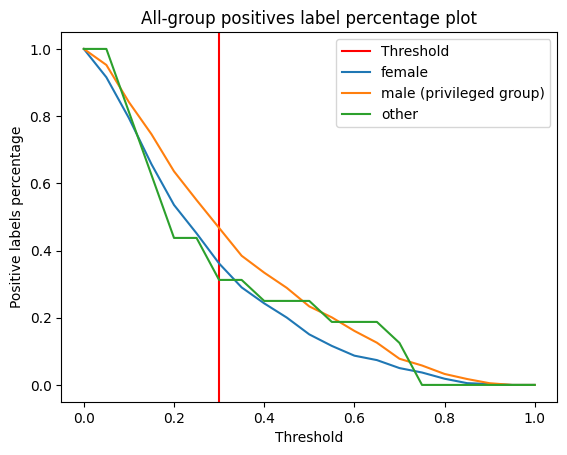
AUC-ROC score: 0.7517037533622939



AUC-PR score: 0.6040892783891119

### Chỉ số ngưỡng

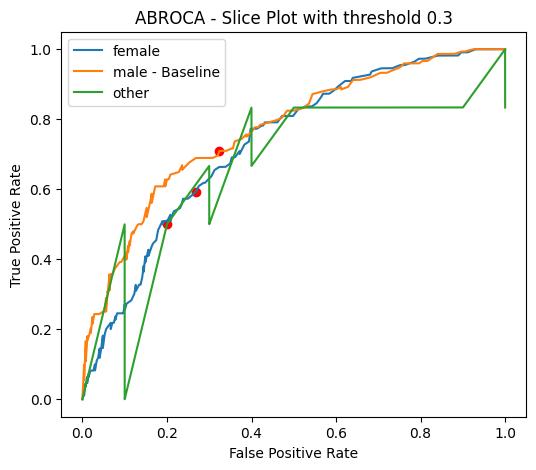
For feature gender:



Demographic parity score: 0.15483668341708545

ABROCA score between male and female: 0.04350382740477098

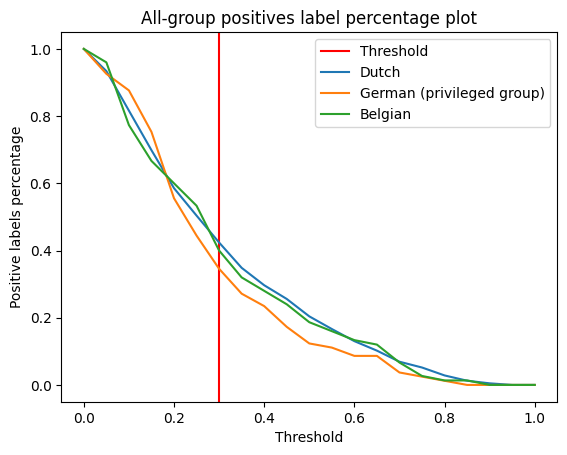
ABROCA score between male and other: 0.09907929397553969



Maximum ABROCA score: 0.09907929397553969

Maximum equalized odd difference: 0.33345945945945943

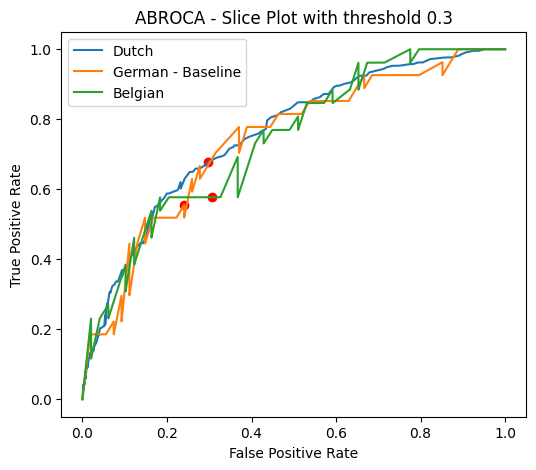
For feature nationality:



Demographic parity score: 0.07818284008760201

ABROCA score between German and Dutch: 0.02848218705150626

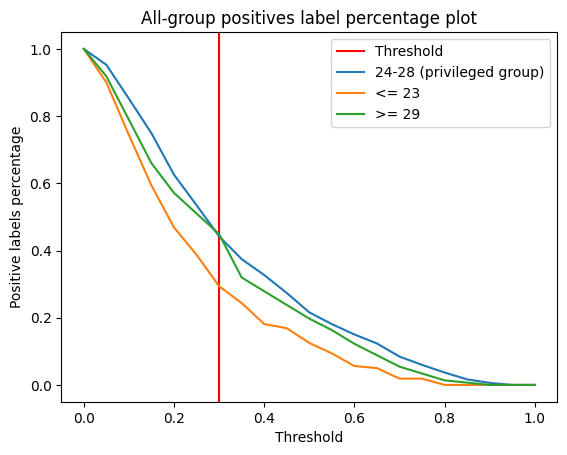
ABROCA score between German and Belgian: 0.04277811117537801



Maximum ABROCA score: 0.04277811117537801

Maximum equalized odd difference: 0.17955088791481555

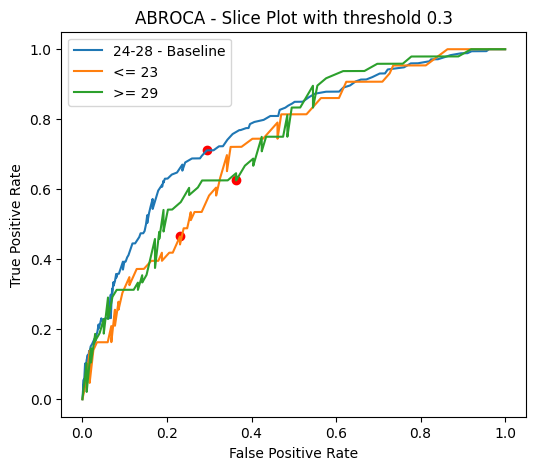
For feature age\_group:



Demographic parity score: 0.1486368312757202

ABROCA score between 24-28 and <= 23: 0.05446581996714603

ABROCA score between 24-28 and >= 29: 0.050830526826591324



Maximum ABROCA score: 0.05446581996714603

Maximum equalized odd difference: 0.3090268615806034

## LOGISTIC REGRESSION (WITHOUT SENSITIVE FEATURES)

### Chỉ số tối ưu

Threshold = 0.3

Print value: False 472

True 321

Name: probability, dtype: int64

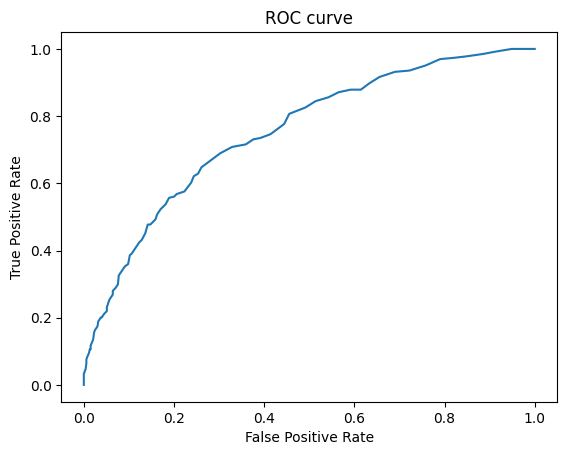
Print value: False 529

True 264

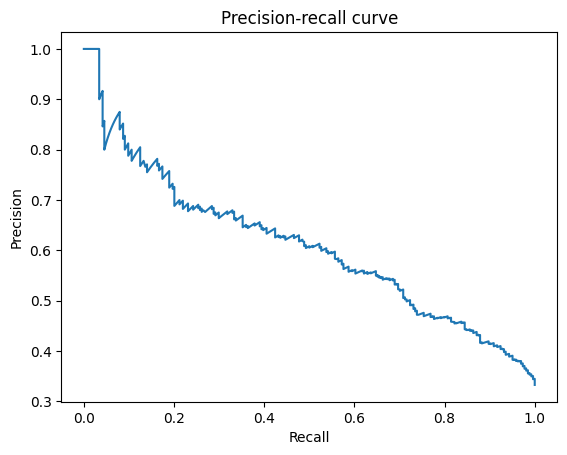
Name: decision, dtype: int64

F1 score: 0.5982905982803711

Accuracy: 0.7036569987389659



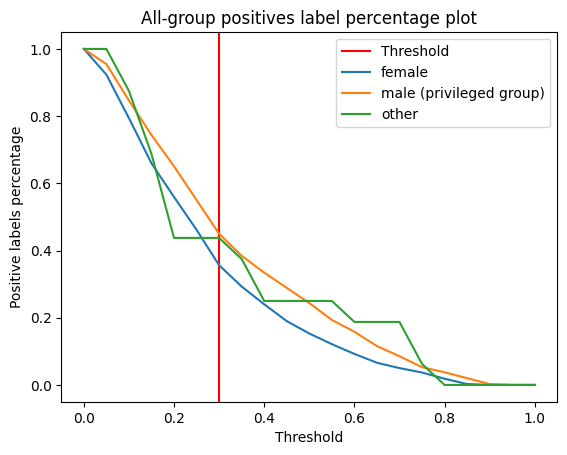
AUC-ROC score: 0.7524513897963231



AUC-PR score: 0.6084459081410444

### Chỉ số ngưỡng

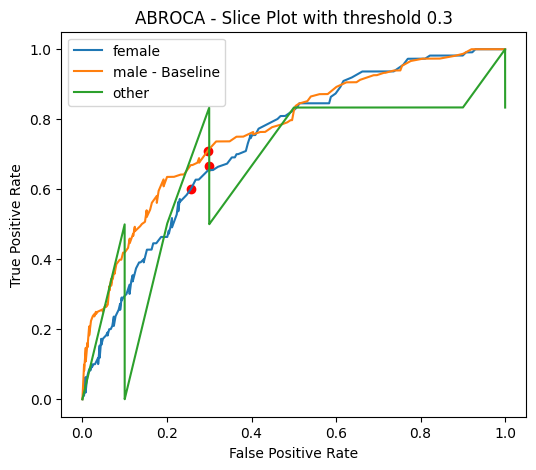
#### For feature gender:



Demographic parity score: 0.09354821601410751

ABROCA score between male and female: 0.04282506164261081

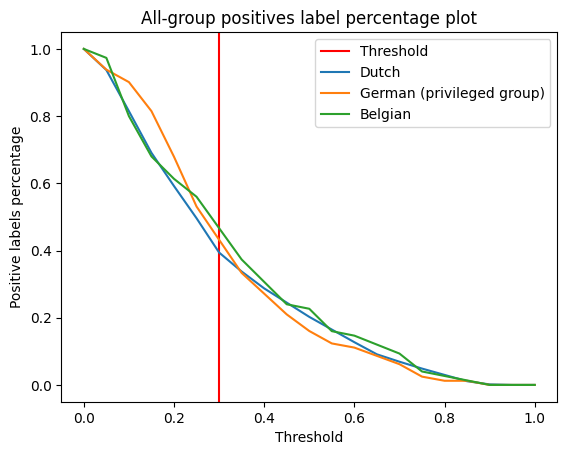
ABROCA score between male and other: 0.1033621100967336



Maximum ABROCA score: 0.1033621100967336

Maximum equalized odd difference: 0.148953883251281

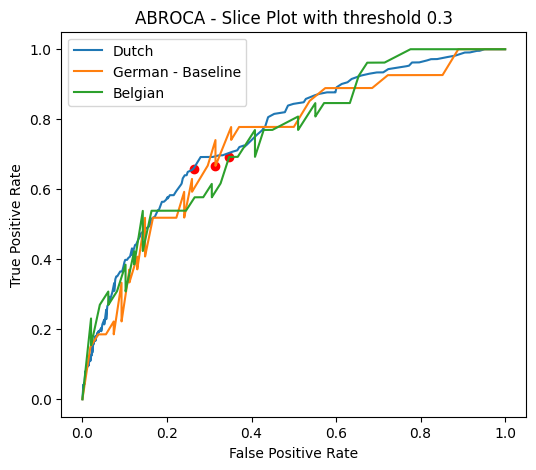
#### For feature nationality:



Demographic parity score: 0.03806422854041891

ABROCA score between German and Dutch: 0.03300931753100507

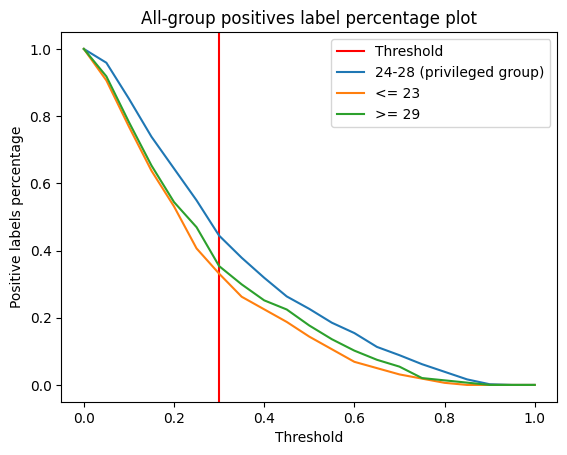
ABROCA score between German and Belgian: 0.041934912951343585



Maximum ABROCA score: 0.041934912951343585

Maximum equalized odd difference: 0.05980291084756728

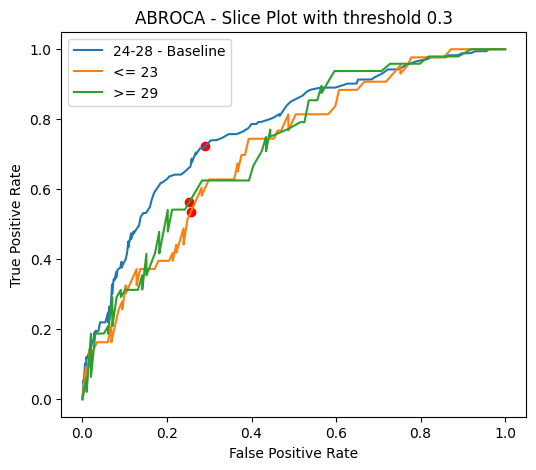
#### For feature age\_group:



Demographic parity score: 0.11319444444444449

ABROCA score between 24-28 and <= 23: 0.06586120340658158

ABROCA score between 24-28 and >= 29: 0.059401602980892663



Maximum ABROCA score: 0.06586120340658158

Maximum equalized odd difference: 0.22198419954181736

## SVM (WITH SENSITIVE FEATURES)

### Chỉ số tối ưu

### Chỉ số ngưỡng

## SVM (WITHOUT SENSITIVE FEATURES)

### Chỉ số tối ưu

### Chỉ số ngưỡng

## DECISION TREE (WITH SENSITIVE FEATURES)

### Chỉ số tối ưu

### Chỉ số ngưỡng

## DECISION TREE (WITHOUT SENSITIVE FEATURES)

### Chỉ số tối ưu

### Chỉ số ngưỡng

## RANDOM FOREST (WITH SENSITIVE FEATURES)

### Chỉ số tối ưu

### Chỉ số ngưỡng

## RANDOM FOREST (WITHOUT SENSITIVE FEATURES)

### Chỉ số tối ưu

### Chỉ số ngưỡng

## GRADIENT BOOSTING (WITH SENSITIVE FEATURES)

### Chỉ số tối ưu

### Chỉ số ngưỡng

## GRADIENT BOOSTING (WITHOUT SENSITIVE FEATURES)

### Chỉ số tối ưu

### Chỉ số ngưỡng

Trong hồi quy logistic, chúng ta sử dụng hàm logistic để chuyển đổi tổng trọng số của các đặc trưng đầu vào thành một giá trị xác suất. Hàm logistic được định nghĩa bởi công thức:

Trong đó:

- là xác suất để là true, hay ở đây là ứng viên đỗ

- là các trọng số của mô hình mà ta cần tối ưu

- là các feature (đặc trưng) của bộ dữ liệu.

Để tìm các trọng số tối ưu, chúng ta sử dụng phương pháp tối ưu hóa hàm chi phí (cost function), thường là hàm log-likelihood.

Hàm chi phí (cost function) trong logistic regression được định nghĩa như sau:

Trong đó:

- là số lượng samples trong bộ dữ liệu.

- là giá trị dự đoán xác suất của .

- là ground truth của mẫu .

Về mặt lý thuyết, ta có thể sử dụng các thuật toán tối ưu hóa như gradient descent để điều chỉnh các trọng số và làm giảm giá trị của hàm chi phí.