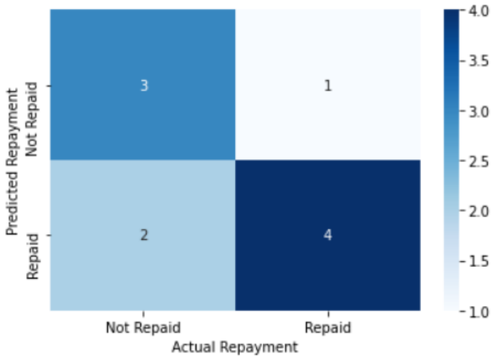
Pa.03 – Lab Assignment

Nobel Gabriel (nobelgab), von Wartburg Rebekka (vonwareb)

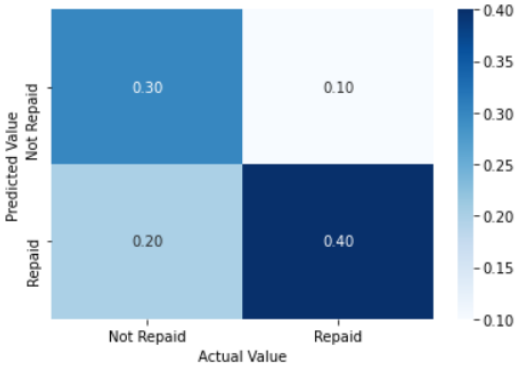
16.03.2023

1. Confusion matrices – small example

Confusion matrix of the actual and the predicted repayment with numbers:

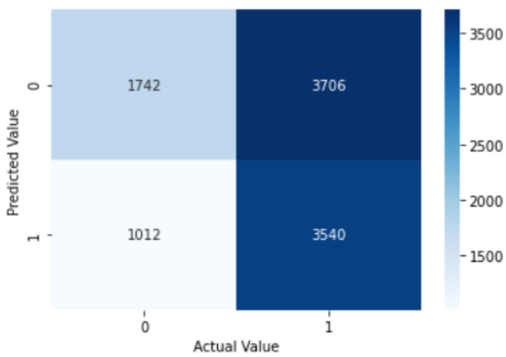


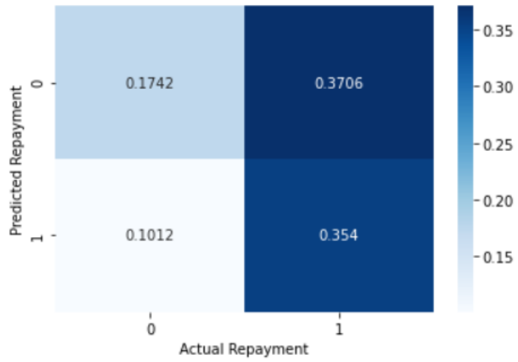
Confusion matrix of the actual and the predicted repayment with probabilities:



2. Confusion matrix – large dataset

Confusion matrix of the actual and the predicted repayment with numbers:





Confusion matrix of the actual and the predicted repayment with probabilities:

3. Calculating conditional probabilities

```
#defines
p_00 = 0.2
p_01 = 0.3
p_10 = 0.1
p_11 = 0.4
```

Answer a)

True positive rate (TPR) = P[D = 1 | Y = 1]

```
TPR = p_11 /(p_01 + p_11)
print('True positive rate (TPR) = P[D = 1 | Y = 1] = ', TPR)
```

```
True positive rate (TPR) = P[D = 1 | Y = 1] =  0.5714285714285715
```

Answer b)

False positive rate (FPR) = P[D = 1 | Y = 0]

```
FPR = p_10/(p_00 + p_10)
print('False positive rate (FPR) = P[D = 1 | Y = 0] = ', FPR)
```

```
False positive rate (FPR) = P[D = 1 | Y = 0] =  0.3333333333333333
```

Answer c)

True negative rate (TNR) = P[D = 0 | Y = 0]

```
TNR = p_00 / (p_00 + p_10)
print('True negative rate (TNR) = P[D = 0 | Y = 0] = ', TNR)
```

```
True negative rate (TNR) = P[D = 0 | Y = 0] =  0.6666666666666666
```

Answer d)

False negative rate (FNR) = P[D = 0 | Y = 1]

```
FNR = p_01 /(p_01 + p_11)
print(' False negative rate (FNR) = P[D = 0 | Y = 1] = ', FNR)
```

```
 False negative rate (FNR) = P[D = 0 | Y = 1] =  0.4285714285714286
```

Answer e)

Positive predicted value (PPV) = P[Y = 1 | D = 1]

```
PPV = p_11 /(p_10 + p_11)
print('Positive predicted value (PPV) = P[Y = 1 | D = 1] = ', PPV)
```

```
Positive predicted value (PPV) = P[Y = 1 | D = 1] =  0.8
```

Answer f)

False discovery rate (FDR) = P[Y = 0 | D = 1]

```
FDR = p_10 / (p_10 + p_11)
print('False discovery rate (FDR) = P[Y = 0 | D = 1] = ', FDR)
```

```
False discovery rate (FDR) = P[Y = 0 | D = 1] =  0.2
```

Answer g)

Negative predictive value (NPV) = P[Y = 0 | D = 0]

```
NPV = p_00 / (p_00 + p_01)
print('Negative predictive value (NPV) = P[Y = 0 | D = 0] = ', NPV)
```

```
Negative predictive value (NPV) = P[Y = 0 | D = 0] =  0.4
```

Answer h)

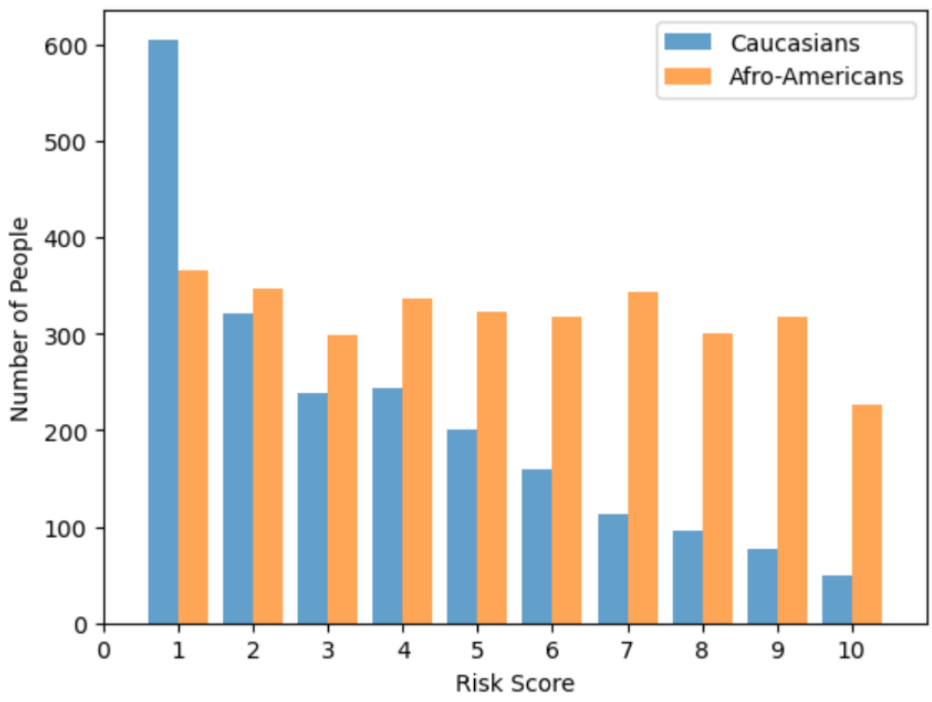
False omission rate (FOR) = P[Y = 1 | D = 0]

```
FOR = p_01 / (p_00 + p_01)
print(' False omission rate (FOR) = P[Y = 1 | D = 0] = ', FOR)
```

```
 False omission rate (FOR) = P[Y = 1 | D = 0] =  0.6
```

4. The COMPAS case

Answer a)



Findings:

Based on the histogram plotted below, we can see that caucasians are much more often classified as low risk. Another finding is that the risk score of african-american people is relatively balanced

Answer b)

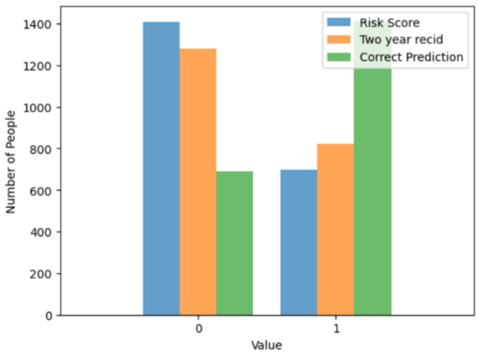
822 out of 2103 or 39.09 percent caucasians are rearrested within 2 years.

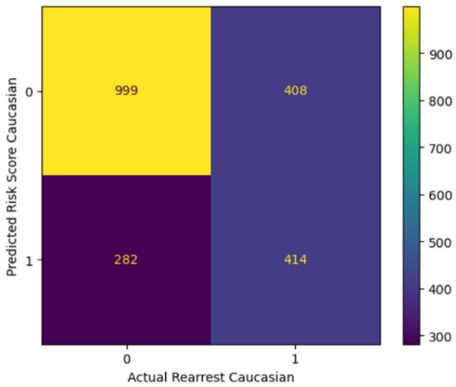
1661 out of 3175 or 52.31 percent afro-americans are rearrested within 2 years.

```
print(len(cc[cc['two_year_recid'] == 1]), 'out of', len(cc), 'or',
      np.round(len(cc[cc['two_year_recid'] == 1]) / len(cc) * 100, 2), 'percent caucasians are rearrested within 2 years.')
print(len(aa[aa['two_year_recid'] == 1]), 'out of', len(aa), 'or',
      np.round(len(aa[aa['two_year_recid'] == 1]) / len(aa) * 100, 2), 'percent afro-americans are rearrested within 2 years.')
```

Answer c)

'Caucasian'

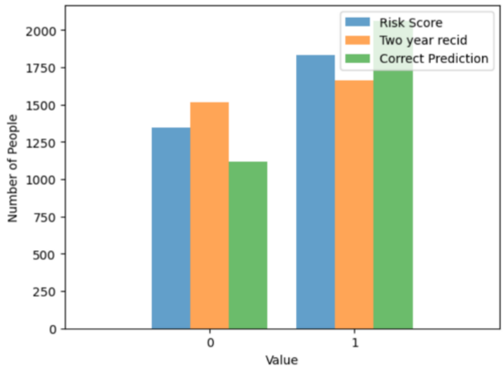


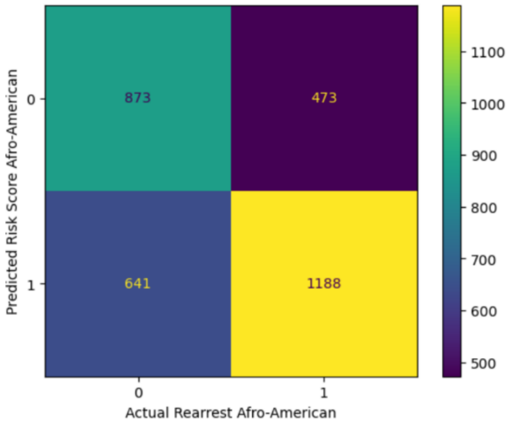


822 caucasians have been rearrested (orange = 1) from an estimated 696 (blue = 1).

The error ratio is: 126 people or 15.33 percent.

'African- American'



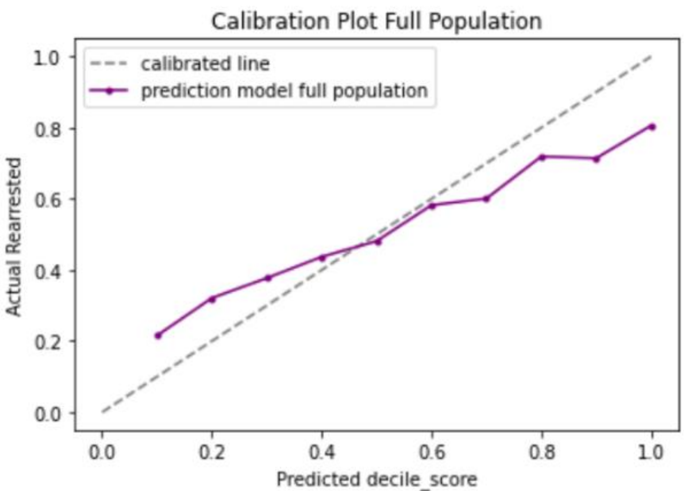


1661 African- American have been rearrested (orange = 1) from an estimated 1829 (blue = 1).

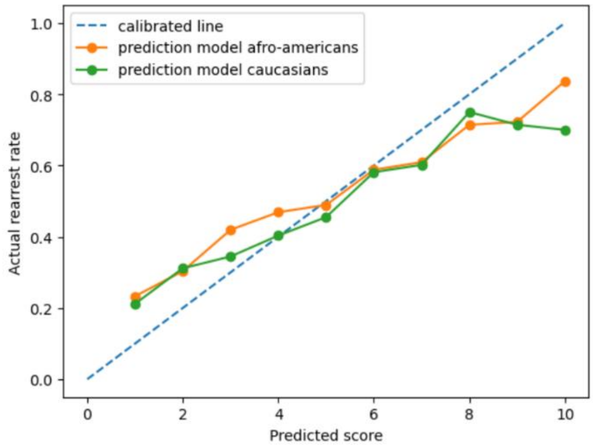
The error ratio is: 168 people or 10.11 percent.

The actual rearrest rates doesn't correspond exactly with the high-risk prediction. It is for both races close. The green pillow in the histogram shows that these statistics isn't applicable for an individual person.

Answer d)



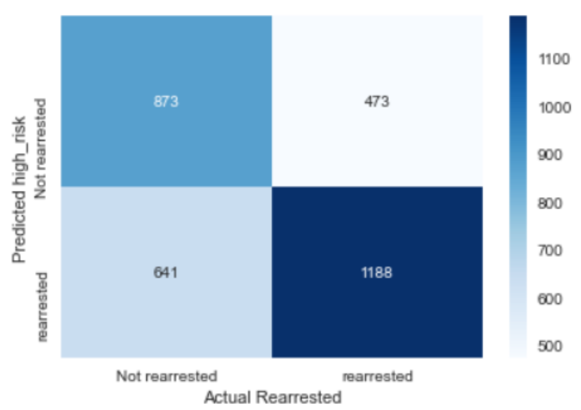
Answer e)



We can see, that for both **caucasians** and **afro-americans** the calibration line is similar. The calibration-between-groups is in our opinion achieved.

Answer f)

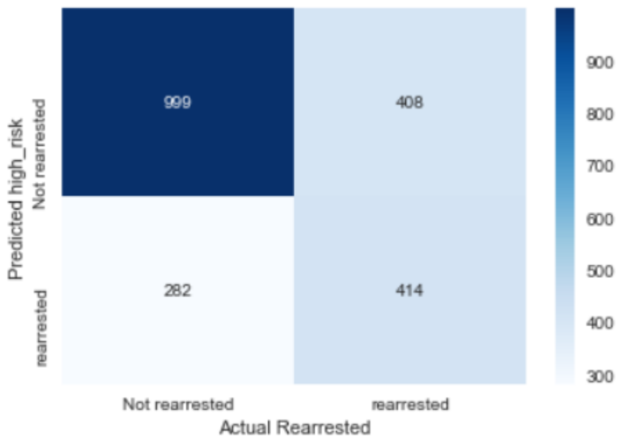
Confusion Matrix African- American defendants




```
FPR = N_10/(N_00 + N_10)
print('False positive rate for BLACK defendants (FPR) = P[D = 1 | Y = 0] = ', FPR)
```

```
False positive rate for BLACK defendants (FPR) = P[D = 1 | Y = 0] =  0.4233817701453104
```

Confusion Matrix Caucasian defendants




```
FPR = N_10/(N_00 + N_10)
print('False positive rate for WHITE defendants (FPR) = P[D = 1 | Y = 0] = ', FPR)
```

```
False positive rate for WHITE defendants (FPR) = P[D = 1 | Y = 0] =  0.22014051522248243
```

Interpret this difference: What do we learn about how the prediction algorithm works differently for Black and White defendants?

Answer: This means that more African-American who are being predicted as risk are not rearrested than Caucasians.

Answer g)

```
FNR = N_01/(N_01 + N_11)
print(' False negative rate WHITE defendants (FNR) = P[D = 0 | Y = 1] = ', FNR)
```

```
False negative rate WHITE defendants (FNR) = P[D = 0 | Y = 1] =  0.49635036496350365
```

```
FNR = N_01/(N_01 + N_11)
print(' False negative rate BLACK defendants (FNR) = P[D = 0 | Y = 1] = ', FNR)
```

```
False negative rate BLACK defendants (FNR) = P[D = 0 | Y = 1] =  0.2847682119205298
```

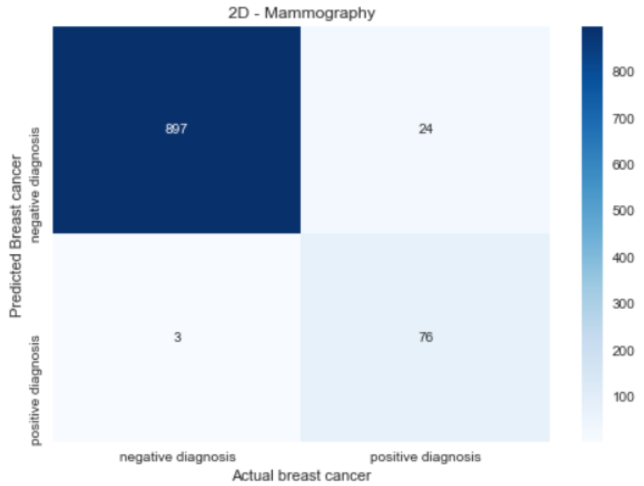
Interpret your findings as you did with the false negative rate.

Answer: This means that more caucasians who are being predicted as no risk are rearrested than african-americans.

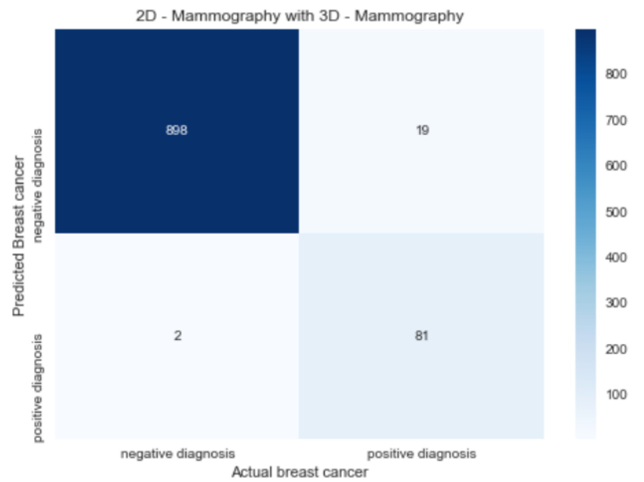
Concluding we saw that even if the prediction model seams pretty balanced the true positive rate and false positive rate show a significant racial bias.

5. Conditional probabilities for breast cancer detection

Confusion matrix for the old procedure (2D mammography):



Confusion matrix for the combination of the new procedure (2D- and 3D- Mammographie):



2D- Mammography:

- Among the patients with breast cancer, what is the share of those who were correctly diagnosed with breast cancer?

  True positive rate 2D-Mammographie (TPR) = P[D = 1 | Y = 1]

```
TPR = N_11/ (N_01 + N_11)
print('True positive rate for 2D- Mammographie (TPR) = P[D = 1 | Y = 1] = ', TPR)
```

```
True positive rate for 2D- Mammographie (TPR) = P[D = 1 | Y = 1] =  0.76
```

- Among patients diagnosed with breast cancer, what is the share of those who actually had breast cancer?

Positive predicted value 2D-Mammographie (PPV) = P[Y = 1 | D = 1]

```
PPV = N_11 / (N_10 + N_11)
print('Positive predicted value for 2D- Mammographie (PPV) = P[Y = 1 | D = 1] = ', PPV)
```

```
Positive predicted value for 2D- Mammographie (PPV) = P[Y = 1 | D = 1] =  0.9620253164556962
```

- Among patients with a negative test, what is the share of those who should have received a positive test because they had breast cancer?

False omission rate 2D-Mammographie (FOR) = P[Y = 0 | D = 0]

```
FOR = N_01 / (N_00 + N_01)
print('False omission rate for 2D- Mammographie (FOR) = P[Y = 1 | D = 0] = ', FOR)
```

```
False omission rate for 2D- Mammographie (FOR) = P[Y = 1 | D = 0] =  0.026058631921824105
```

- Among patients with a negative test, what share received the correct result?

Negative predicted value 2D-Mammographie (NPV) = P[Y = 0 | D = 0]

```
NPV = N_00 / (N_00 + N_01)
print('Negative predicted value for 2D- Mammographie (NPV) = P[Y = 1 | D = 0] = ', NPV)
```

```
False predicted value for 2D- Mammographie (FOR) = P[Y = 1 | D = 0] =  0.9792802617230099
```

2D + 3D- Mammography:

- Among the patients with a negative test, what share received the wrong diagnosis?

False omission rate 2D-Mammographie with 3D- Mammographie (FOR) = P[Y = 1 | D = 0]

```
FOR = N_01 / (N_00 + N_01)
print('False omission rate for 2D- Mammographie with 3D- Mammographie (FOR) = P[Y = 1 | D = 0] = ', FOR)
```

```
False omission rate for 2D- Mammographie (FOR) = P[Y = 1 | D = 0] =  0.020719738276990186
```

- Among patients who did not have breast cancer, what share was incorrectly diagnosed with breast cancer?

False positive rate 2D-Mammographie with 3D- Mammographie (FPR) = P[D = 1 | Y = 0]

```
FPR = N_10 / (N_00 + N_10)
print('False positive rate for 2D- Mammographie with 3D- Mammographie (FPR) = P[D = 1 | Y = 0] = ', FPR)
```

```
False positive rate for 2D- Mammographie with 3D- Mammographie (FPR) = P[D = 1 | Y = 0] =  0.0022222222222222222
```

- Among patients with a positive result, what share was actually breast-cancer-free?

False discovery rate 2D-Mammographie with 3D- Mammographie (FDR) = P[Y = 0 | D = 1]

```
FDR = N_10 / (N_10 + N_11)
print('False discovery rate for 2D- Mammographie with 3D- Mammographie (FDR) = P[Y = 0 | D = 1] = ', FDR)
```

```
False discovery rate for 2D- Mammographie with 3D- Mammographie (FDR) = P[Y = 0 | D = 1] =  0.024096385542168676
```