Exercise 1:

In this session, you will learn to:

- Construct and interpret ROC curves.
- Calculate the Area Under the Curve (AUC).
- Analyze the impact of data imbalance and decision thresholds on model performance.
- Collaborate with peers to compare results and discuss findings.

Below a table that consist of true labels (y) and predicted probabilities of four different classifiers ($\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{\pi}_4$) generated from hypothetical models.

y	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$
1	0.99	0.10	0.01	0.7
1	0.60	0.05	0.40	0.9
1	0.95	0.07	0.05	0.2
1	0.70	0.15	0.30	0.8
0	0.80	0.01	0.20	0.5
0	0.10	0.08	0.90	0.1
0	0.30	0.02	0.70	0.3

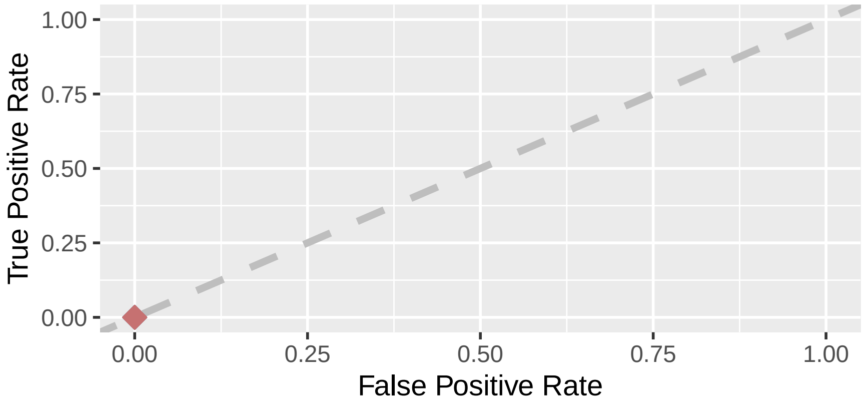
Tasks

- Step 1: Watch as the instructor demonstrates how to plot the ROC curve using $\hat{\pi}_1$ and explains the steps.
- Step 2: Form groups of 4-6 people and
  - Complete the ROC curve for $\hat{\pi}_1$.
  - Plot the ROC curves for $\hat{\pi}_2$, $\hat{\pi}_3$, and $\hat{\pi}_4$.
  - Manually calculate the AUC for each classifier and compare the results.
  - Compute the prevalence and the average of the predicted probability of each classifier across all 7 observations.
- Step 3: Within your group, discuss:
  - How the differences in predictions affect the ROC curves and AUC values.
  - The differences between average predicted probability and the prevalence.
  - Group A students: Assume you want to obtain a high partial AUC (pAUC) for low FPR values (e.g., using the constraint: FPR < 0.2). Compare the pAUC of the four classifiers.
  - Key takeaways from comparing the four classifiers.
- Step 4: Formulate 1-2 challenging TRUE-FALSE questions about ROC curves and post them into the Etherpad in Moodle. Nominate a group leader to present one question to the class and explain its relevance.

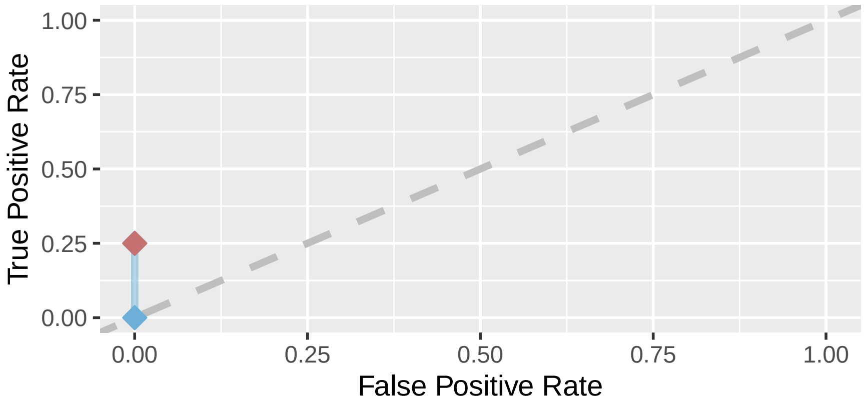
Solution 1:

1 Solution Classifer 1

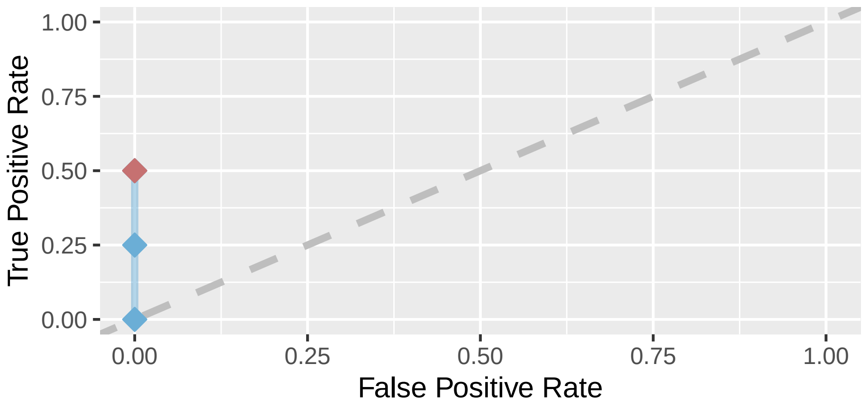
#	Truth	Score
1	Pos	0.99
3	Pos	0.95
5	Neg	0.80
4	Pos	0.70
2	Pos	0.60
7	Neg	0.30
6	Neg	0.10



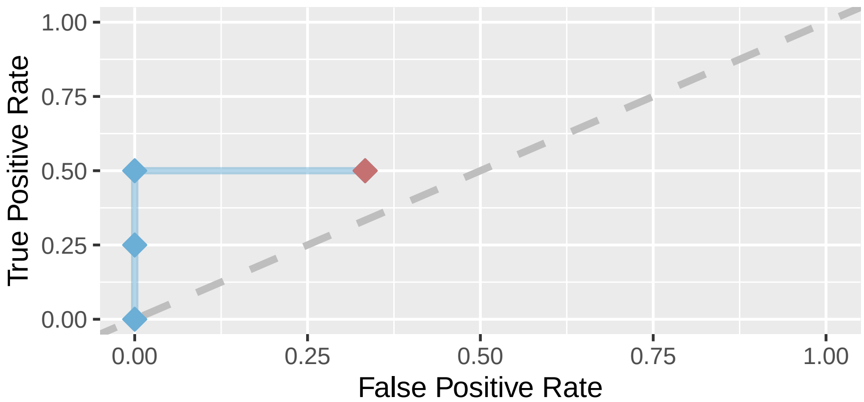
#	Truth	Score
1	Pos	0.99
3	Pos	0.95
5	Neg	0.80
4	Pos	0.70
2	Pos	0.60
7	Neg	0.30
6	Neg	0.10



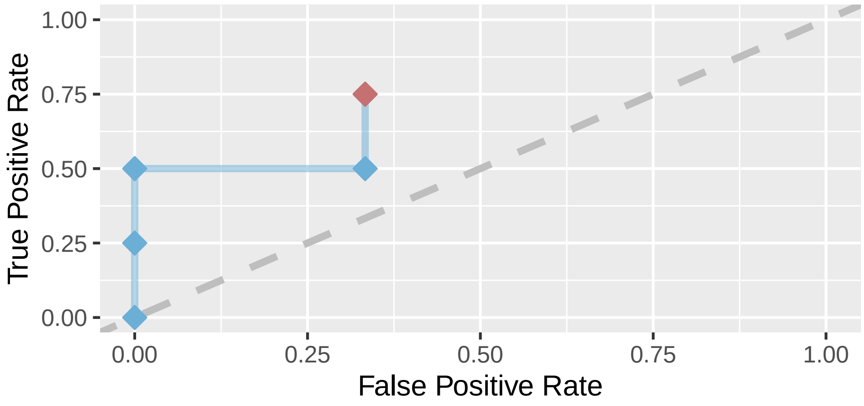
#	Truth	Score
1	Pos	0.99
3	Pos	0.95
5	Neg	0.80
4	Pos	0.70
2	Pos	0.60
7	Neg	0.30
6	Neg	0.10



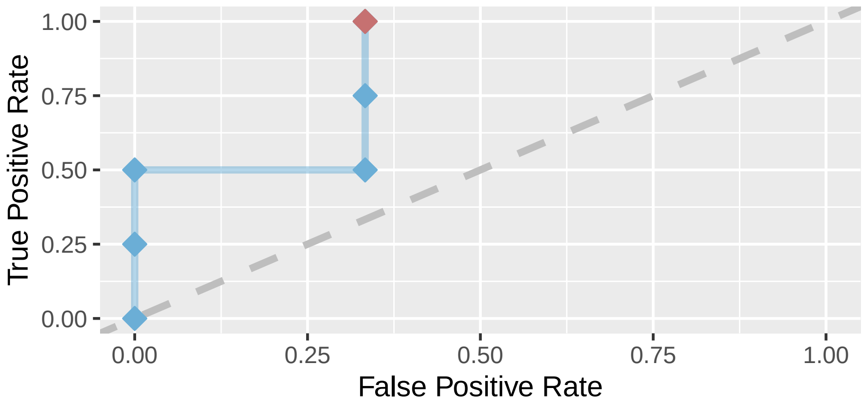
#	Truth	Score
1	Pos	0.99
3	Pos	0.95
5	Neg	0.80
4	Pos	0.70
2	Pos	0.60
7	Neg	0.30
6	Neg	0.10



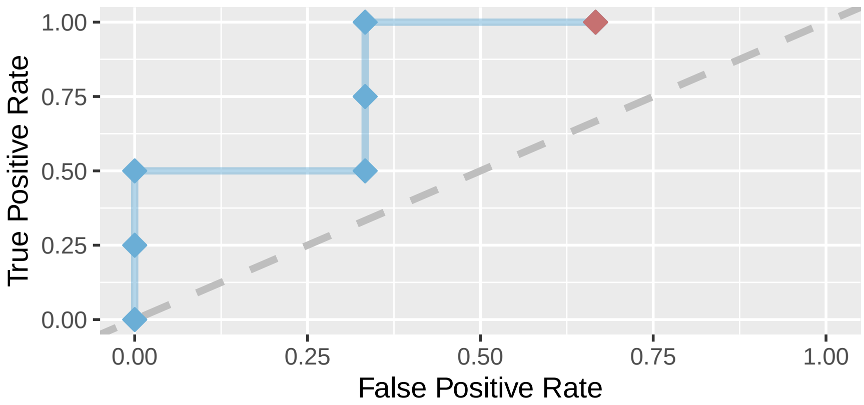
#	Truth	Score
1	Pos	0.99
3	Pos	0.95
5	Neg	0.80
4	Pos	0.70
2	Pos	0.60
7	Neg	0.30
6	Neg	0.10



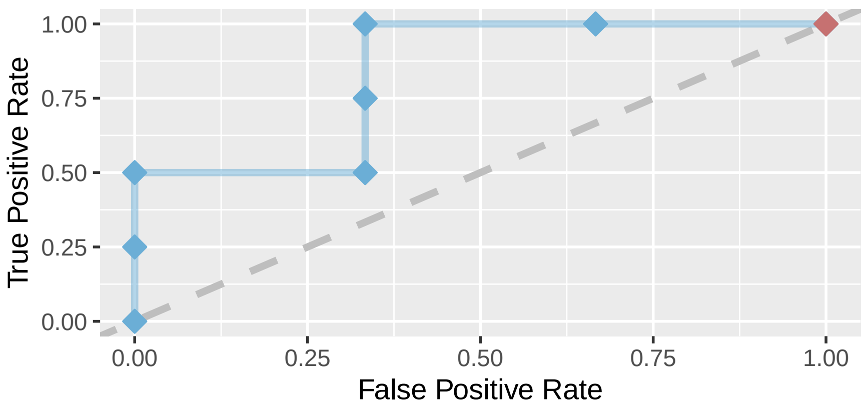
#	Truth	Score
1	Pos	0.99
3	Pos	0.95
5	Neg	0.80
4	Pos	0.70
2	Pos	0.60
7	Neg	0.30
6	Neg	0.10



#	Truth	Score
1	Pos	0.99
3	Pos	0.95
5	Neg	0.80
4	Pos	0.70
2	Pos	0.60
7	Neg	0.30
6	Neg	0.10



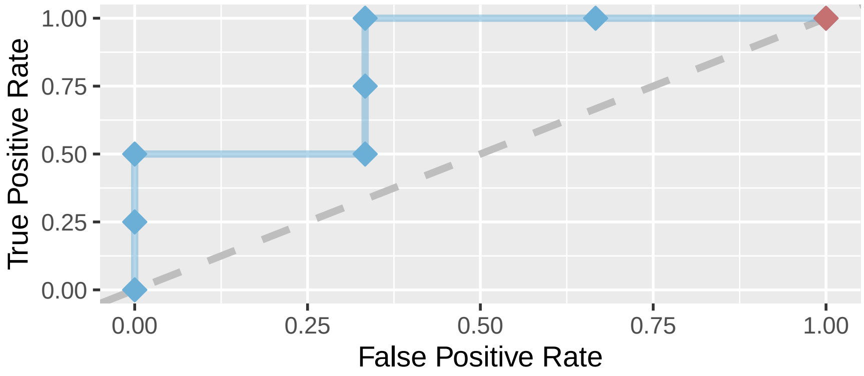
#	Truth	Score
1	Pos	0.99
3	Pos	0.95
5	Neg	0.80
4	Pos	0.70
2	Pos	0.60
7	Neg	0.30
6	Neg	0.10




```
## auc: 0.833333333333333; average predicted probability: 0.634285714285714
```

2 Solution Classifer 2

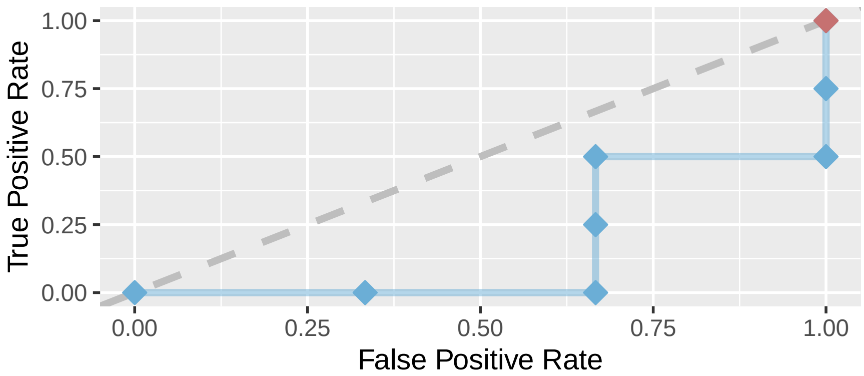
#	Truth	Score
4	Pos	0.15
1	Pos	0.10
6	Neg	0.08
3	Pos	0.07
2	Pos	0.05
7	Neg	0.02
5	Neg	0.01




```
## auc: 0.833333333333333; average predicted probability: 0.0685714285714286
```

3 Solution Classifer 3

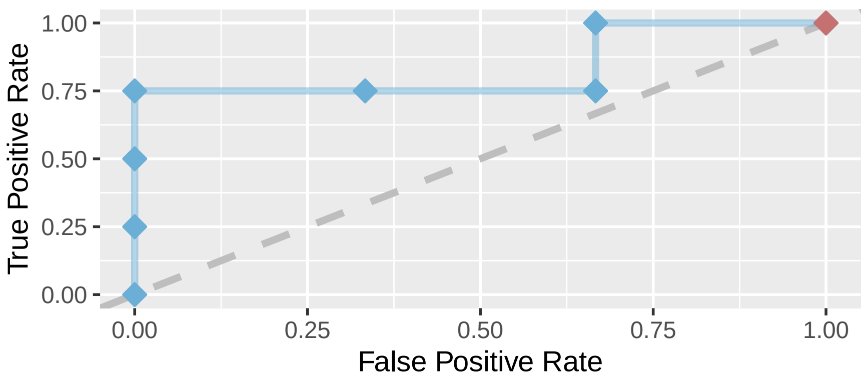
#	Truth	Score
6	Neg	0.90
7	Neg	0.70
2	Pos	0.40
4	Pos	0.30
5	Neg	0.20
3	Pos	0.05
1	Pos	0.01




```
## auc: 0.166666666666667; average predicted probability: 0.365714285714286
```

4 Solution Classifer 4

#	Truth	Score
2	Pos	0.9
4	Pos	0.8
1	Pos	0.7
5	Neg	0.5
7	Neg	0.3
3	Pos	0.2
6	Neg	0.1




```
## auc: 0.833333333333333; average predicted probability: 0.5
```

Step 3: Group Discussion

- **How the differences in predictions affect the ROC curves and AUC values:**
  - The AUC depends on the ranking of true positives ($y = 1$) versus false positives ($y = 0$):
    * π_1**:** Well-ranked probabilities result in a high full AUC (0.8333) and a steep initial ROC curve.
    * π_2**:** Effective rankings lead to the same full AUC as π_1, despite lower probability values.
    * π_3**:** Misranked probabilities (e.g., assigning higher probabilities to negatives than positives) lead to a low full AUC (0.1667) and a poor ROC curve (worse than random guessing). Using $1 - \pi_3$ would improve the classifier.
    * π_4**:** Reasonable rankings with some overlap between positive and negative probabilities yield a high full AUC (0.8333).
  - ROC curve shapes reveal separation quality:
    * Steep initial curves indicate strong separation (e.g., π_1, π_2, π_4).
    * Flat or below-diagonal curves indicate poor separation (e.g., π_3).
- **Average Predicted Probability and Prevalence:**
  - **Prevalence:** The proportion of positive cases ($y = 1$) is $\frac{4}{7} \approx 0.5714$.
  - Average predicted probability for each classifier (should ideally match prevalence for good "calibration"):
    * π_1**:** Average probability = 0.63. Closer to prevalence, with reasonable alignment.
    * π_2**:** Average probability = 0.0571. Much lower than prevalence, showing poor probability calibration despite correct rankings.
    * π_3**:** Average probability = 0.3686. Probabilities are not well aligned with the true prevalence.
    * π_4**:** Average probability = 0.5. Closer to prevalence, with reasonable alignment.
  - Takeaway: Average predicted probability reflects the alignment of the classifier's outputs with prevalence. π_1 and π_4 show better alignment, while π_2 and π_3 deviate significantly.
- **For Group A students: Partial AUC for FPR < 0.2:**
  - π_1, π_2: Moderate pAUC (0.50). Strong initial separation, but some misranked probabilities in low-FPR regions reduce performance.
  - π_3: Low pAUC (0.00). Misranked probabilities, poor performance for low FPR.
  - π_4: High pAUC (0.75). Best performance in low-FPR regions due to effective rankings.

  Example Use Case: In applications like cancer screening or fraud detection:

  - Limiting false positives is critical to avoid too many unnecessary tests or investigations.
  - π_4 would be the preferred classifier due to its superior partial AUC in low-FPR regions.
- **Key Takeaways from Comparing the Four Classifiers:**
  - Rankings drive AUC and pAUC, not the magnitude of predicted probabilities.
  - Calibration matters for aligning predictions with prevalence. π_1 and π_4 are better calibrated than π_2 and π_3.
  - π_4 excels in low-FPR regions, making it ideal for applications requiring strict control of false positives.