For this measure, we evaluate the likelihood of a counterfactual trace by determining whether a counterfactual leads to the desired outcome or not. For this purpose, we use the predictive model, which returns a prediction for each counterfactual sequence. As we are predicting process outcomes, we typically predict a class. However, forcing a deterministic model to produce a different class prediction is often difficult. Therefore, we can relax the condition by maximising the prediction score of the desired counterfactual outcome[1]. If we compare the difference between the counterfactual prediction score with the factual prediction score, we can determine an increase or decrease. Ideally, we want to increase the likelihood of the desired outcome. We refer to this value as *delta*. However, the binary case introduces some noteworthy considerations.

Within this task setting, we have to consider multiple cases. First, the prediction score is typically limited to a domain within 0 and 1, which we can interpret as a probability distribution. Hence, if the model score is 0.6, then the model has the confidence of 60% that the input can be categorised as belonging to class 1. For instance, within a medical process, we could say the model is 75% confident that the patient can be cured. Conversely, there's a 25% confidence that the process instance belongs to class 0. We can make decisions by using a threshold. Typically, this threshold lies at 50%. Hence, we determine that a patient can make decisions by using a threshold. Typically this threshold lies at 50%. Figure 1 illustrates how this threshold behaves given the factual prediction score.

We identify 2 cases:

Case 1: A counterfactual generator *flips* the prediction score to the opposite side of the decision threshold. Then, we achieve our general aim, and the difference between the scores directly indicates the counterfactual's success. If we look at Figure 1, the two quadrants with only positive delta values cover this case.

Case 2: A counterfactual has the same decision as the factual. For instance, when the counterfactual and factual prediction scores for a patient's recovery chance are below 0.5 or above 0.5. Then, we must consider whether the counterfactual predictions score is moving towards the desired outcome or away from it.

[2.1] If the prediction for the factual decides an outcome of 0, but the predictions score for the counterfactual is even lower, then we did not change the prediction at all. In fact, we increase the chance of the actual factual outcome. That situation is worse than what we desire. For instance, a patient would not want to pursue a counterfactual

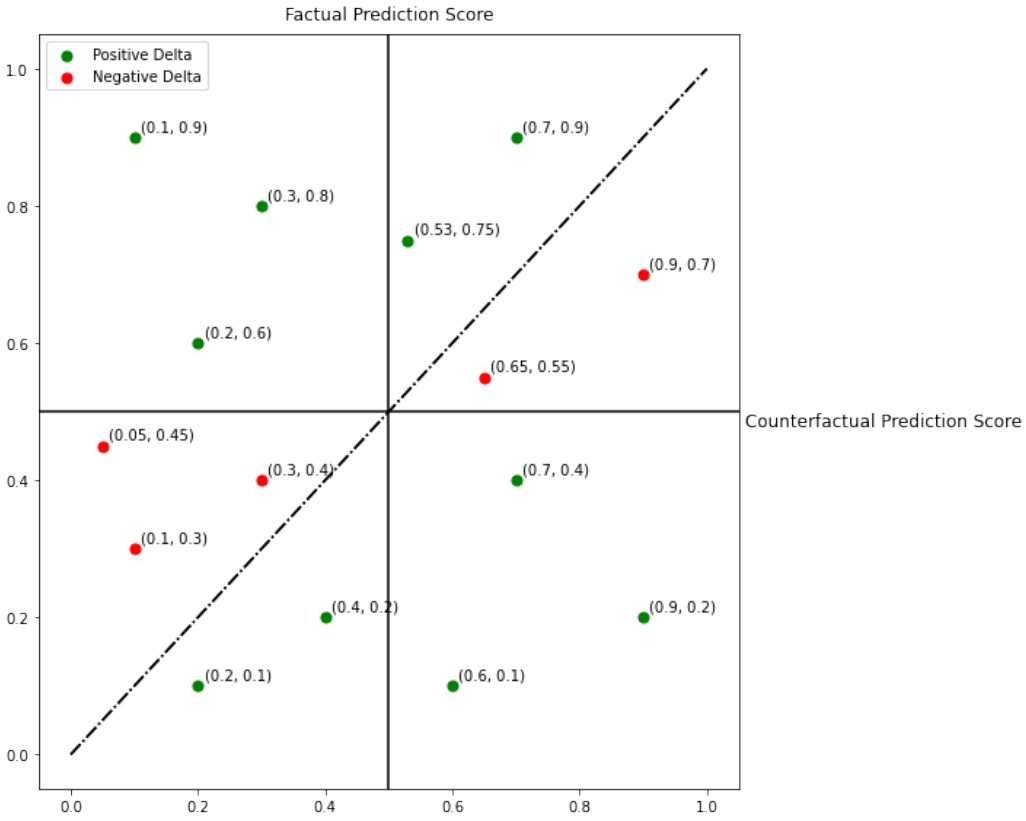

Figure 1: An example of which points yield a positive delta given the factual predictions score. Green means the delta is an improvement. Red points signify a negative improvement.

situation in which his odds of recovery are worse than his current.

[2.2] In contrast, if a prediction model's score leads to an outcome of 0 but the counterfactual returns a higher prediction score than the factual predictions score, a patient might still be interested in the counterfactual. In some situations, even a small improvement is desirable.

The sub-cases of case 2 go in both ways. Hence, we have to incorporate each case differently in the delta score. The two quadrants in Figure 1 that have positive and negative deltas reflect how we interpret these cases.

$$delta = \begin{cases} |p(o|s^*) - p(o|s)| & \text{if } p(o|s) > 0.5 \ \& \ p(o|s) > p(o|s^*) \\ -|p(o|s^*) - p(o|s)| & \text{if } p(o|s) > 0.5 \ \& \ p(o|s) < p(o|s^*) \\ |p(o|s^*) - p(o|s)| & \text{if } p(o|s) < 0.5 \ \& \ p(o|s) > p(o|s^*) \\ -|p(o|s^*) - p(o|s)| & \text{if } p(o|s) < 0.5 \ \& \ p(o|s) < p(o|s^*) \end{cases} \quad (1)$$

[0]Obviously, the domain of the application decides where this threshold lies. One can always argue that confidence of 51% is close to randomly guessing.