I used cv or bootstrap for all my models, so I am surprised that the performance dropped so much on the 'secret' sample, when I compared the ROC for the train to  the ones I was getting on the site they were lower on the site as expected, but not that low

I am not sure if the hold out test set was not representative (for all I know they selected what to hold out to make a point) or why this happened, any insight is appreciated