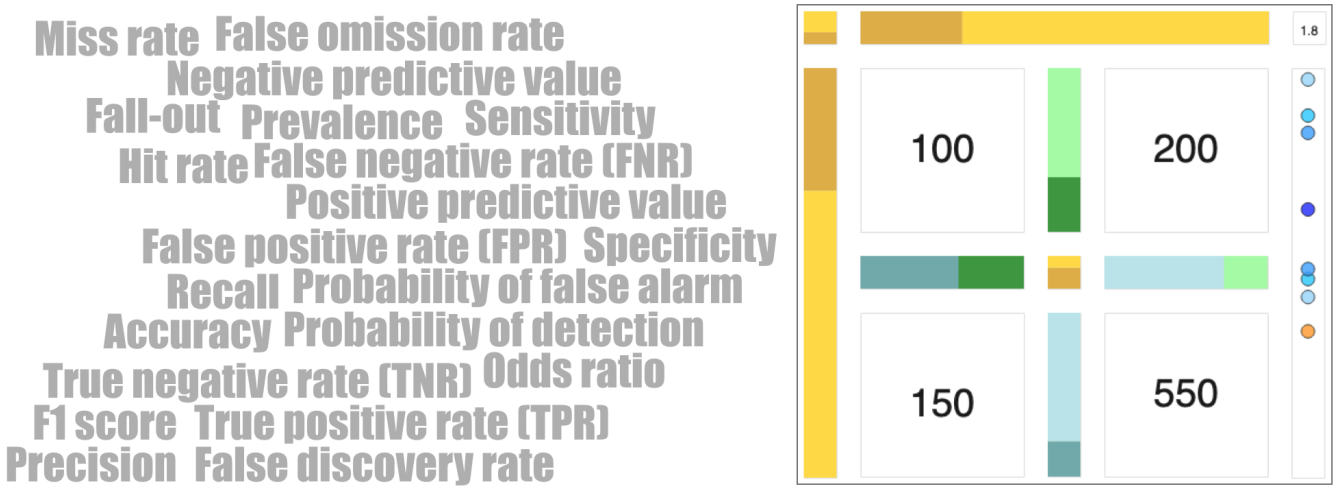


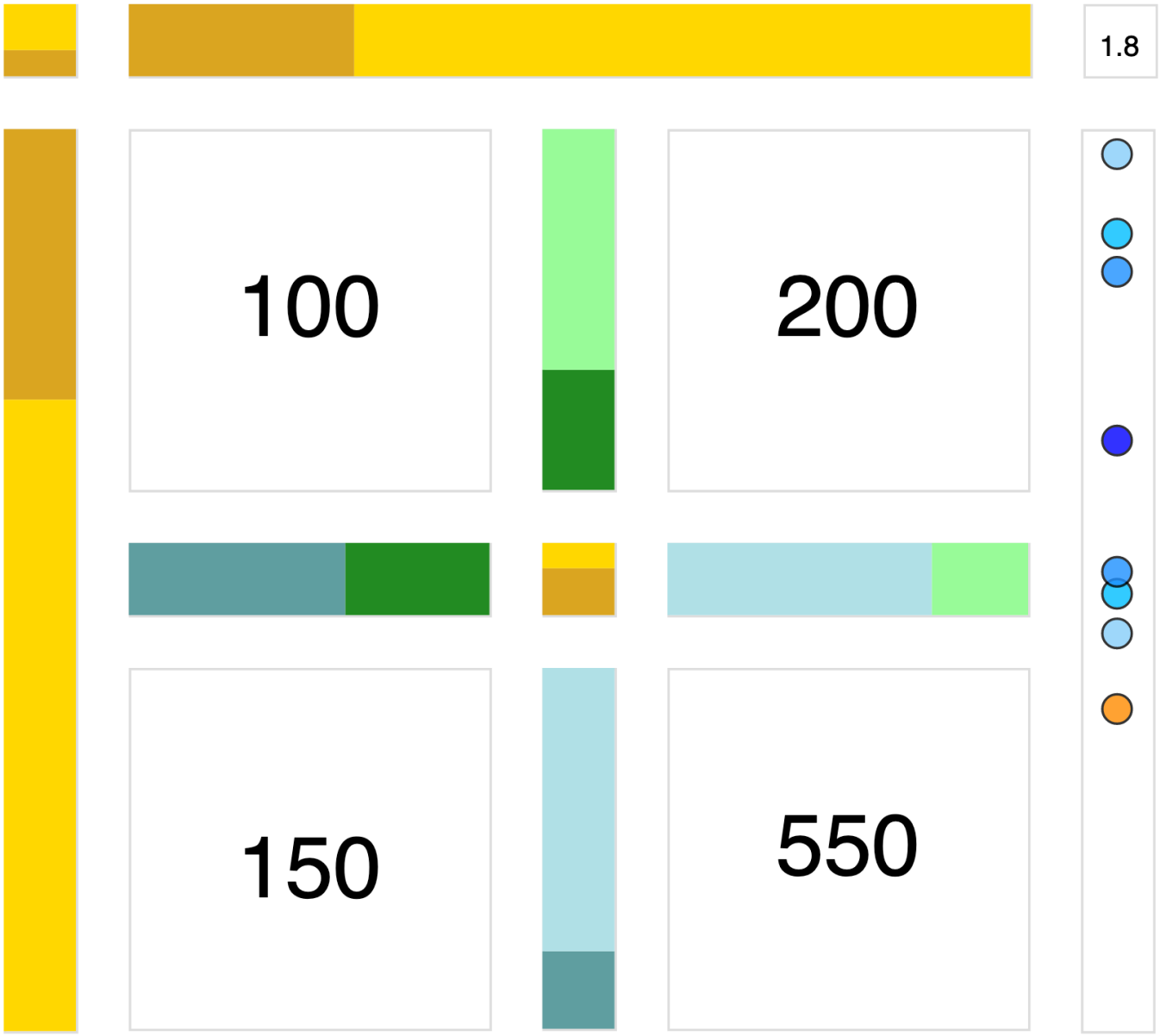
The confusion matrix visualized

A graphical approach creates insights and clarity concerning the many metrics associated with the 2×2 matrix.

Soren Laursen Follow
Mar 17 · 9 min read

I was always amazed by the richness of the 2×2 confusion matrix. After all, it's simply four numbers, how complex is that? Well, as it turns out there is an abundance of insights to be gained. But the picture is blurred by the fact that the confusion matrix is used in many areas of business, engineering, and science each with their own vocabulary.

Being inclined to visualization I created a chart that helps to illustrate and understand the many concepts. Basically, it is the confusion matrix itself with a number of decorators.

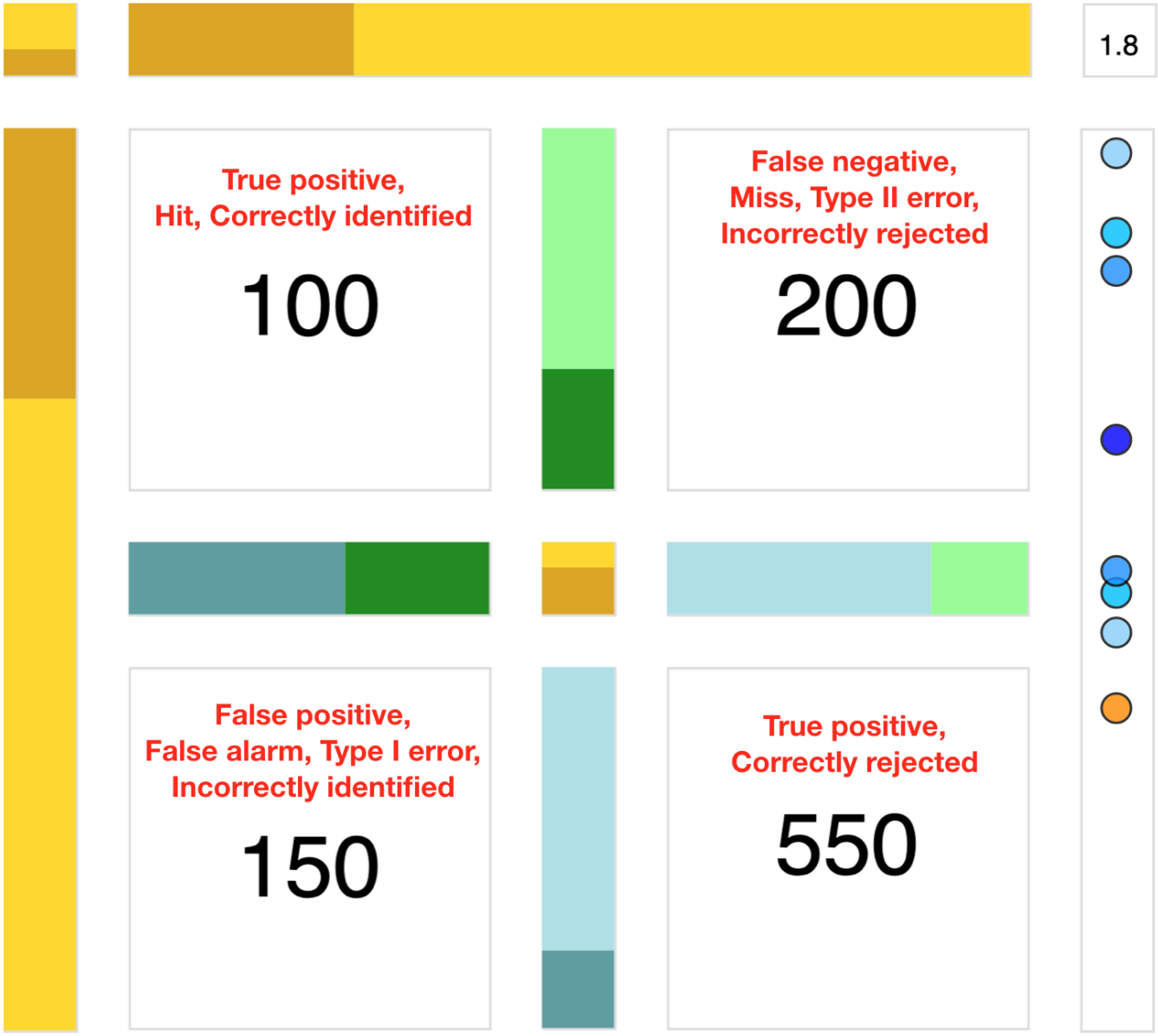


The chart is created with the python visualization package Altair. A Jupyter notebook with the chart code can be downloaded from my GitHub (https://github.com/SorenLaursen/confusion_matrix_chart). Here the chart shows the actual numerical values of all graphical items e.g. segments of bars as tooltips.

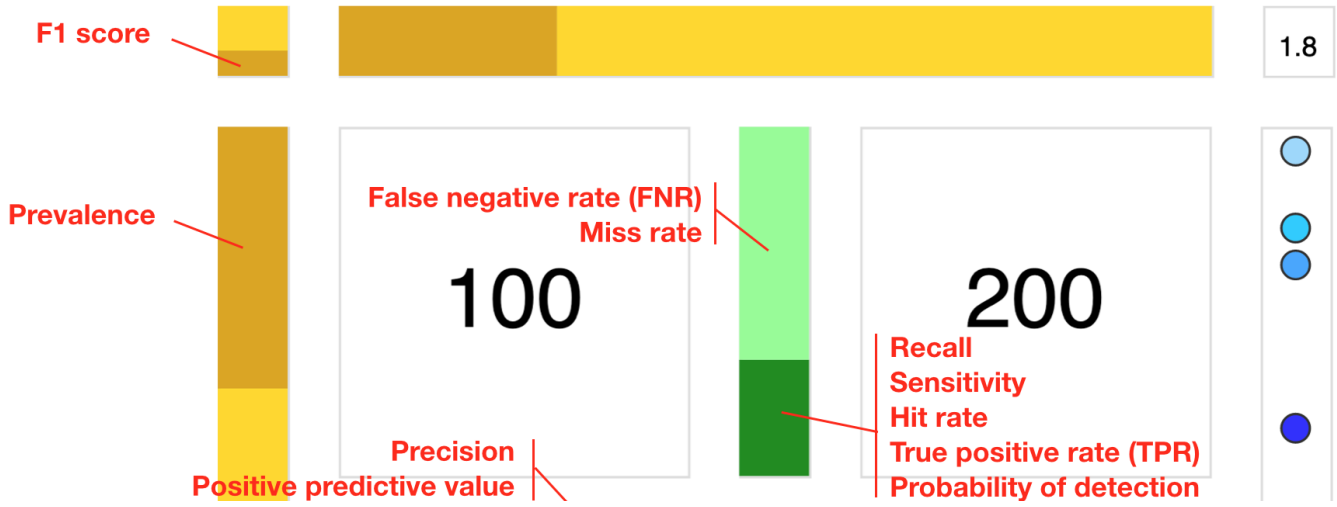
Vocabulary

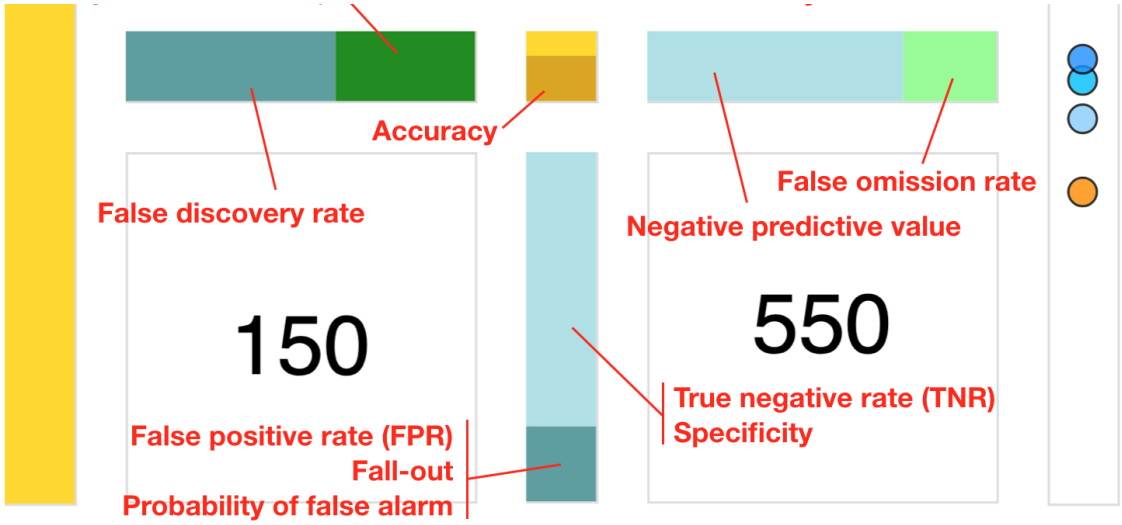
The confusion matrix maps two binary classes against each other. We shall here refer to these as (Y, N) and (y, n) for the rows and the columns respectively. Illustratively these may represent an evaluation of some criteria e.g. membership of some group (Yes, No).

Let's start exploring the vocabulary around the confusion matrix [1]. First the four numbers themselves:



Most of the alternatives are intuitively understandable. It becomes messier with the multitude of names for the many ratios derived from the four numbers:





The decorator to the right is a little more complex but still based on the four numbers.

For reference I will introduce a simpler naming convention which is also used in the implementation code:

Definitions

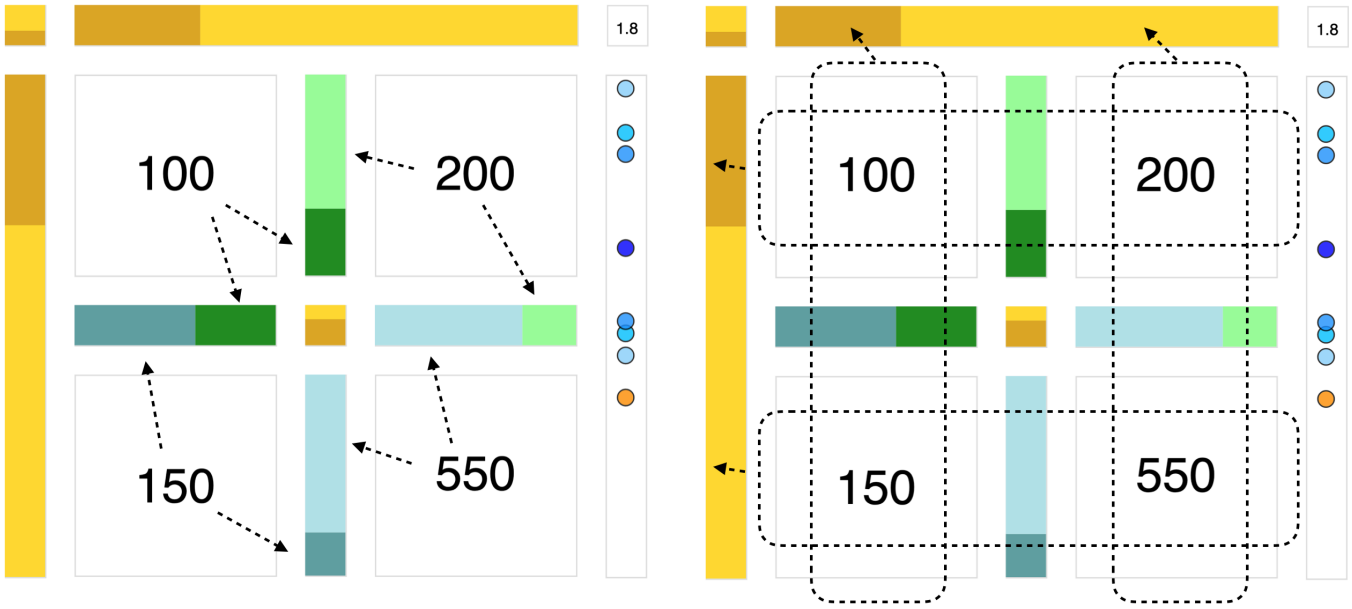
Next, let's turn to the definitions. The bars between the four numbers simply illustrate the ratio of each number to the pairwise sum. In other words, each bar is a normalized stacked bar of the two adjacent numbers.

We shall here refer to these four bars as the *pair bars*.

Of course, the $a|b$ notation for naming the segments of the pair bars hints to the understanding of these segments as conditional probabilities.

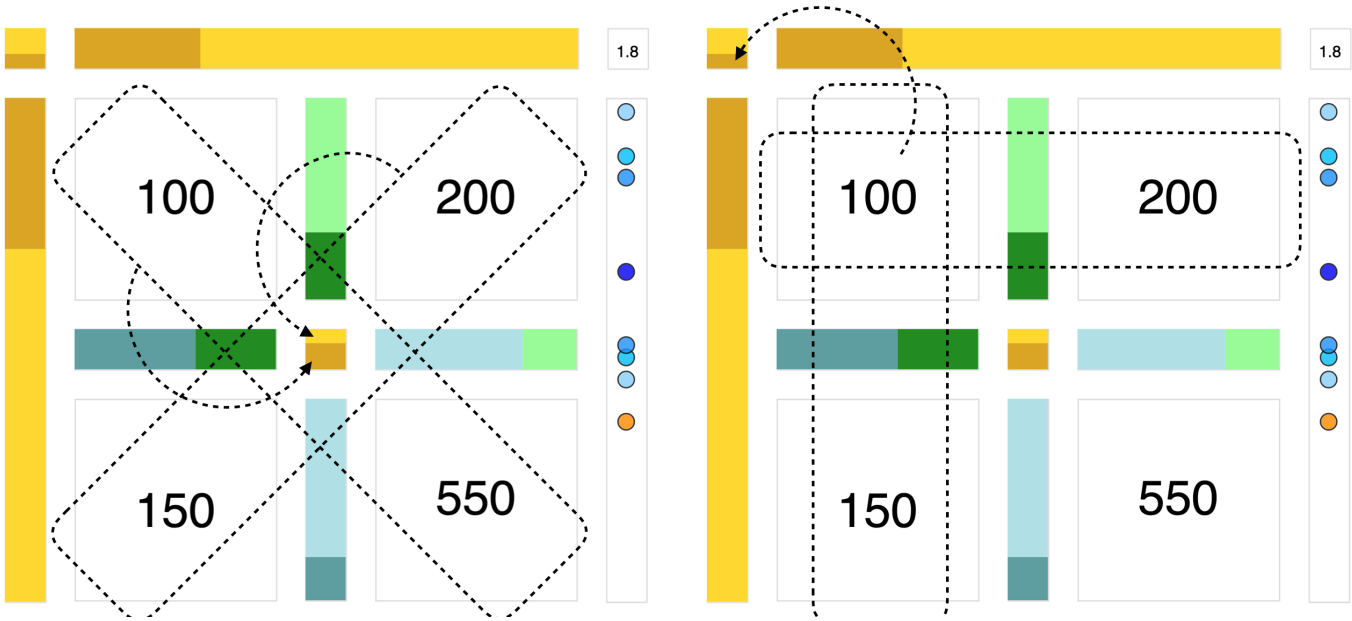
The ordering convention for the segments in the pair bars is that the sections related to the numbers in the diagonal are closest to the center of the chart.

The golden bars left and top represents the row sums and the column sums respectively. Both of these bars contains information from all four numbers. Again the bars should be considered normalized.



Similarly, the bar in the middle summarises all four numbers, this time not in rows and columns but diagonal and off-diagonal.

In contrary to these the bar in the upper left corner represents only three of the four numbers. It is the F1 score i.e. $2 \cdot (y|Y \cdot Y|y) \,/\, (y|Y + Y|y)$. In other words, it is calculated from the two dark green segments. The factor 2 implies that also this bar is normalized.



Finally, the odds ratio is defined as

$OR = (Yy \cdot Nn) / (Yn \cdot Ny)$.

The implementation in the notebook applies the small count correction (Haldane–Anscombe correction), i.e. if either of the four numbers is zero then 0.5 is added to all four numbers[2].

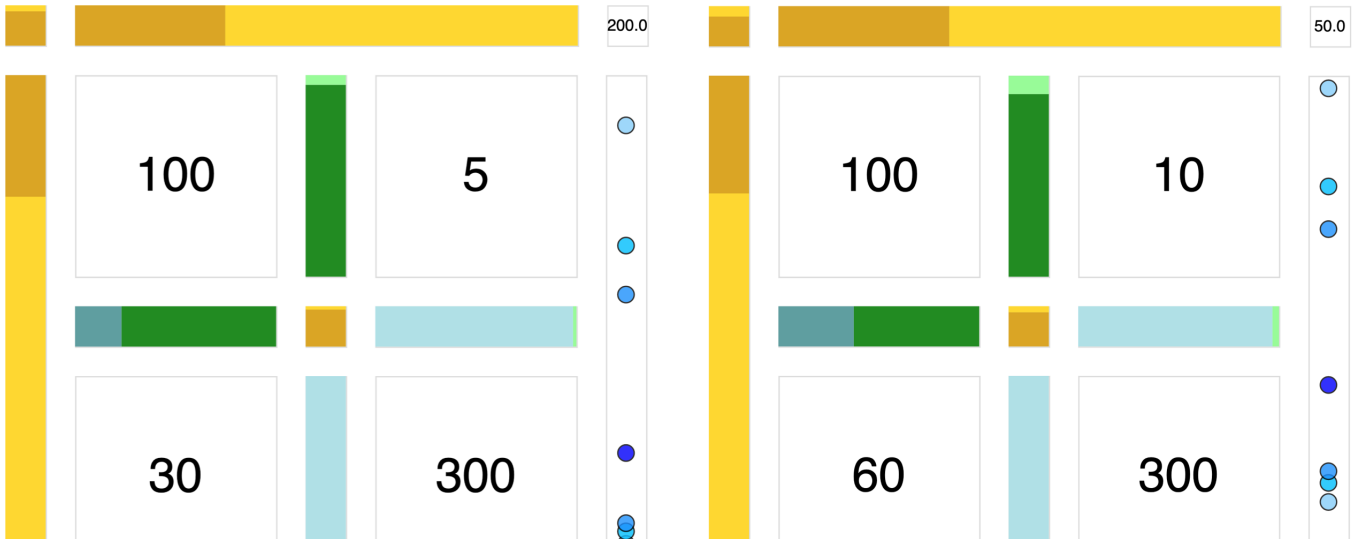
It can be shown that the confidence interval can be simply approximated:

$\exp\{ \log(OR) \pm cv \cdot \mathrm{sqrt}(1/Yy + 1/Yn + 1/Ny + 1/Nn) \}$

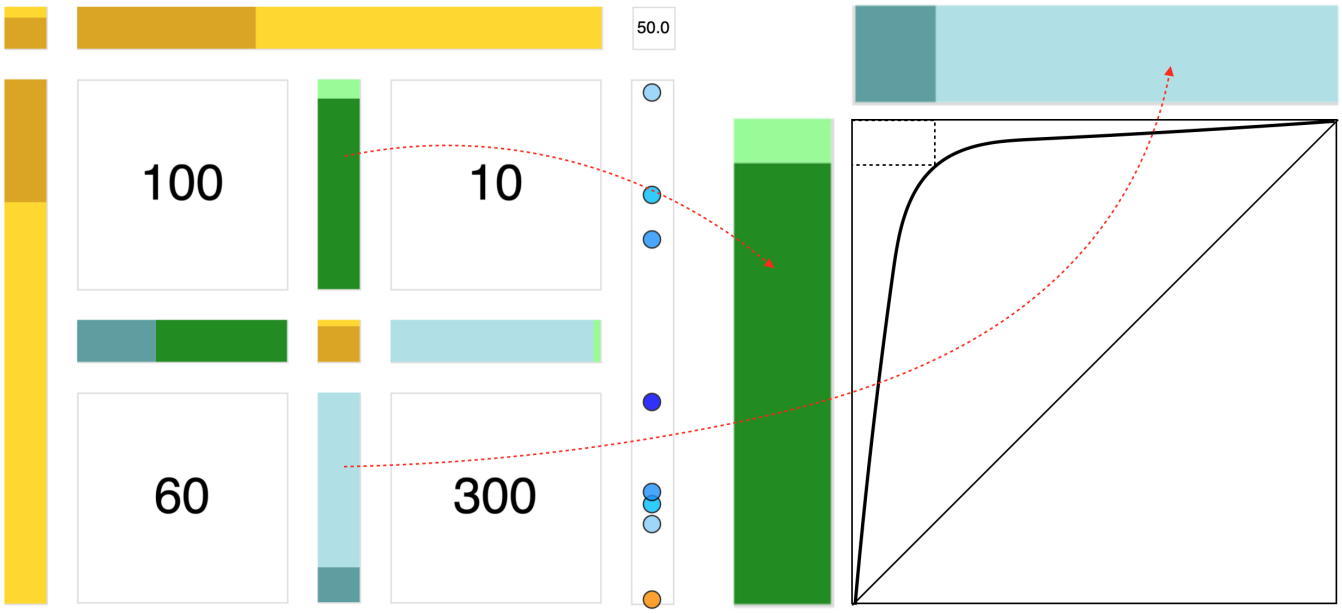
where the critical value, cv, depends on the confidence level: 90%: 1.64, 95%: 1.96, 99%: 2.58.

Applications

Let's take off with a classical application of the confusion matrix. World War II was the early days of radar. The sensitivity of the gear could be adjusted but which were the optimal settings? Turn up the sensitivity of the radar and most enemy planes are detected. But at the same time, there are many false alarms. Turn down the sensitivity and there are few false alarms but now some enemy planes are not detected which is even worse. We want to minimize false negatives. Similarly in a cancer screening program. We don't want to miss anyone who actually has cancer. This result comes at the expense of some false positives.



The analysis in this space is typically made on the vertical pair bars. These summarise information from all four numbers. Evidently, multiplying the two numbers in a row with some factor does not change the vertical pair bars. Consequently, the distribution between *Y* and *N* does not influence these two pair bars. This led to these pair bars being used in a special plot, the receiver operating characteristic curve or ROC curve.



The confusion matrix represents some sample e.g. a number of experiments, a number of patients or a number of radar measurements with some specific equipment settings i.e. operating characteristics. Thus the specific confusion matrix represents one point on the ROC curve.

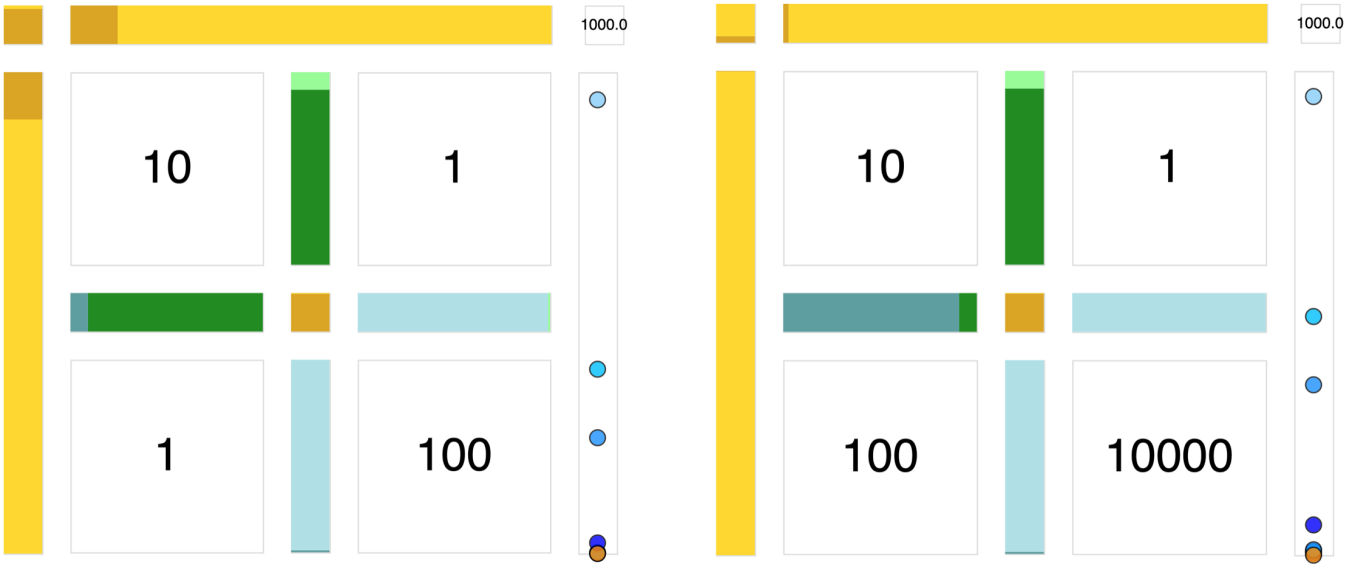
As the vertical pair bars do not depend on the *Y/N* distribution it means that the measurements from a radar system in location A can be compared to those from a radar system in location B even though there were more enemy planes in location A than in location B.

Sometimes it is desirable to minimize the false positives (*Ny*) rather than the false negatives (*Yn*). For instance, a spam filter marks a message as spam or not. We don't want a proper message to be marked as spam as we may never see it. On the other hand, it is not a big problem if a few spam messages remain in the inbox. Another example is an advertising campaign: we are predicting who is in the target group. We don't want to do

an expensive campaign directed towards a non-target audience. Thus, minimize the false positives. As a result, we might not address the entire target group, but those we address are most likely in the target group.

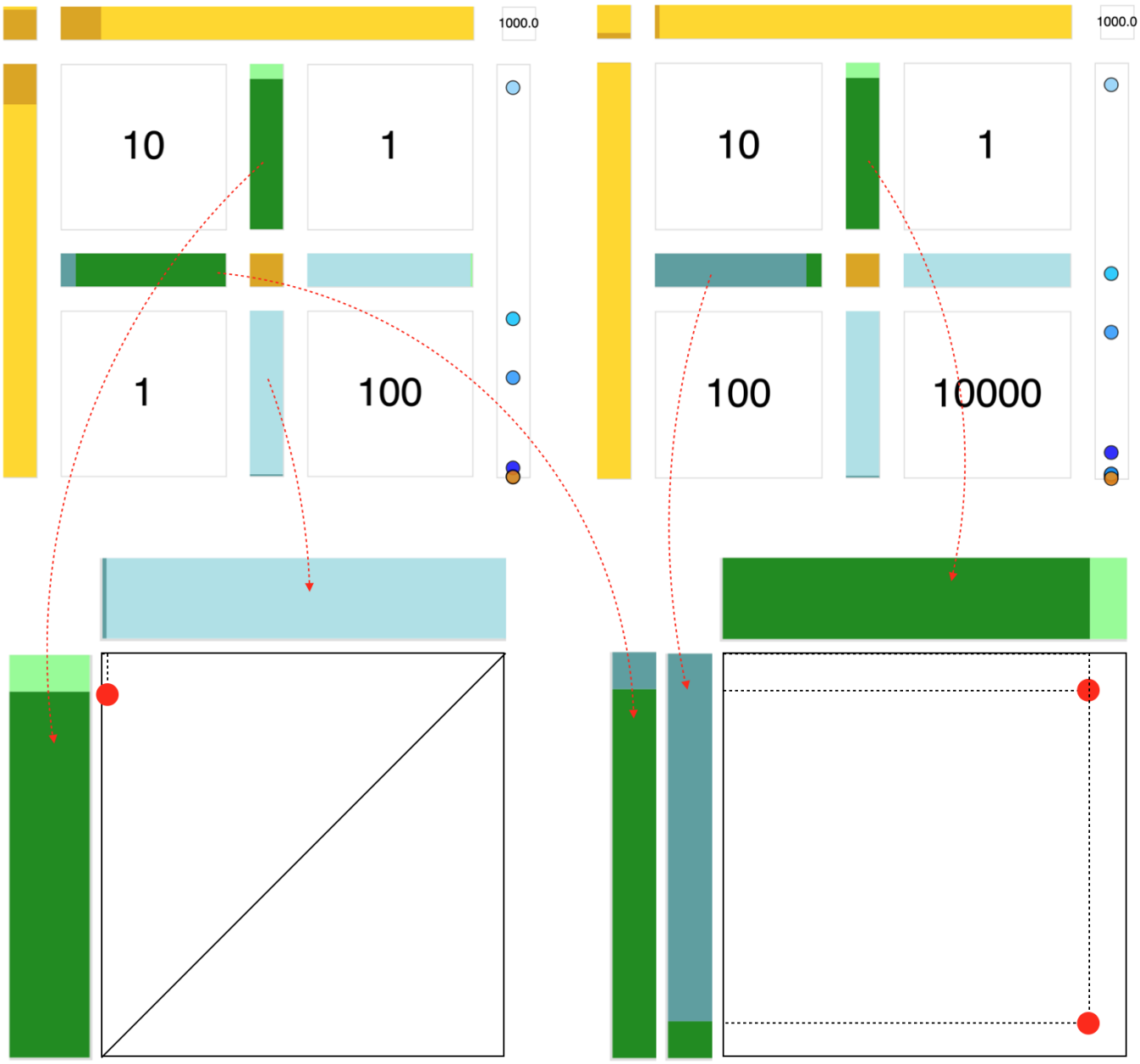
Note, whether we are minimizing the false positives or the false negatives does not tell us which measure to evaluate. We can still use the vertical pair bars, but now we will focus on $y|N$ rather than $n|Y$. But when do we use the horizontal pair bars then?

Often data sets are strongly imbalanced. For instance fraudulent credit card transactions among all credit card transactions. Take a look at the following two examples. In the right-hand case, the numbers in the second row are multiplied with 100 compared to the left-hand case. The vertical pair bars are not affected, they are independent of prevalence (Y). Also, in both cases, the accuracy — the center bar — is close to 100%. But the number of false positives compared to true positives in the second case may be devastating.



In strongly imbalanced data sets true negatives may dominate the picture and in many cases, the other three numbers are of more interest. Remember from the definition section above how the F1 score — the top left bar — is a measure based on these other three numbers, actually based on $y|Y$ and $Y|y$ i.e. the two dark green segments. As can be seen, contrary to the accuracy the F1 score is sensitive to the difference between the two cases.

The ROC plot summarised the overall picture. For the imbalanced data, there is a similar plot: the precision-recall plot [3]. Basically, both plots are two pair bars mapped against each other. For simplicity, we shall here refer to such plots as pair-pair plots. As the vertical pair bars are independent of the numbers in a row being multiplied by a common factor — they are prevalence independent — the two cases are at the same point on the ROC plot (left). In the precision-recall plot, however, the two cases separate distinctly.



In the situation where we can control prevalence (Y), both plots are applicable. This is the case for instance for model evaluation. We have a training set to train the model i.e. we can use the same set with different parameter settings. In this way, we can construct

pair-pair plots knowing the prevalence (Y) is constant. This is contrary to the radar case where the prevalence of enemy planes would differ across geography and time.

Dependency

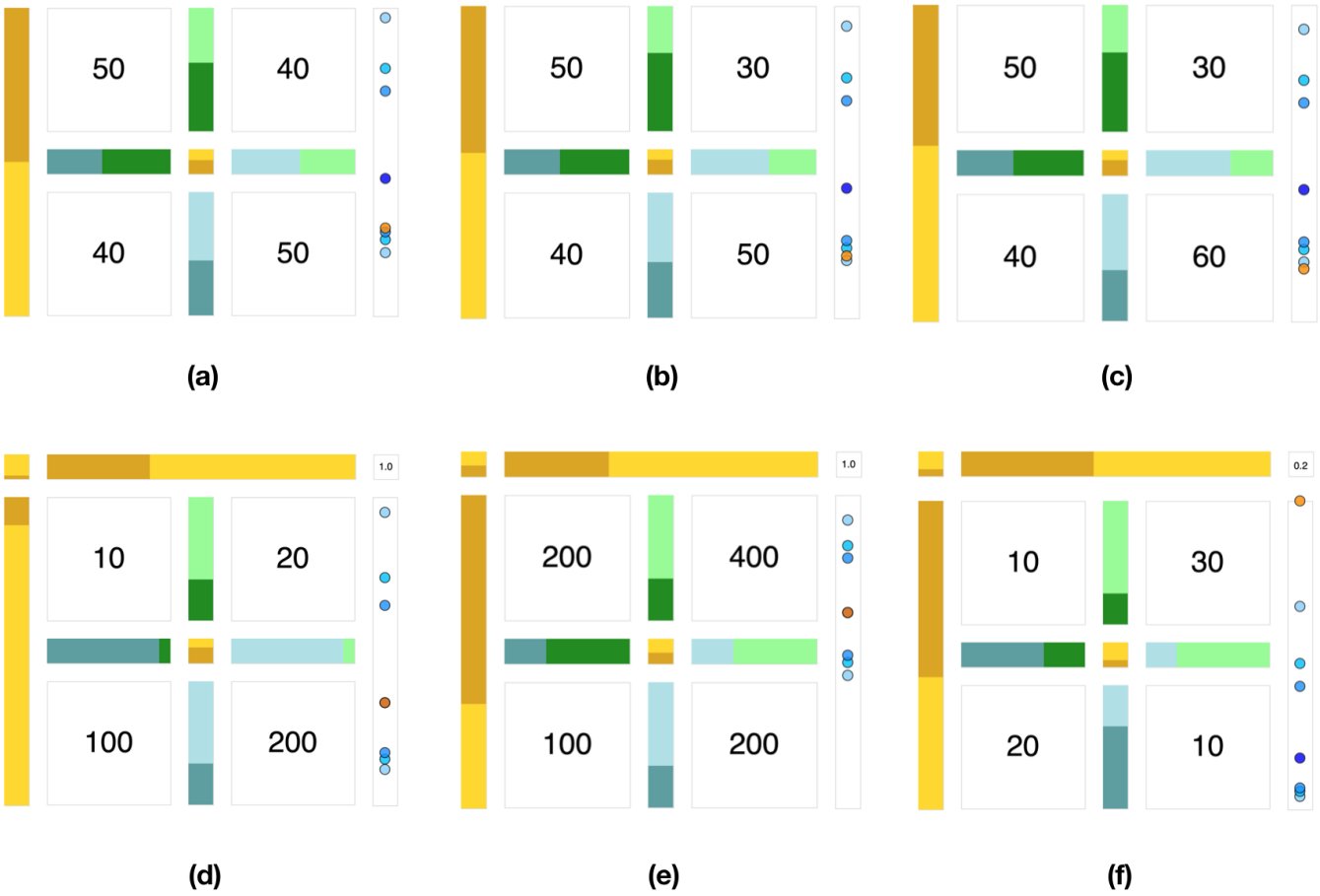
Until now we have looked at situations where there is a truth (Y, N) and some calculated results or empirical test results (y, n) and we use our confusion matrix and the corresponding derived measures to evaluate model quality, quality of medical tests, hypothesis testing, etc.

But there is another use of the 2×2 matrix with a little different mindset: are two binary classes independent or not? For instance, is the membership of the age group 20–29 years (Y, N) and being a car owner (y, n) associated? This could also be applied to all the examples above. In this area of application, the matrix is usually referred to as a contingency table.

This is where the odds ratio (OR) comes in. It was defined above as the ratio between the product of the diagonal numbers to the product of the off-diagonal numbers. If OR is 1 then the two classes are independent. In the OR decorator in the confusion matrix chart — i.e. the right-hand bar — the dark blue dot is OR and the orange dot is 1. As this OR *depiction* is not normalized its value is shown in the upper right box for convenience.

How much does OR need to differ from 1 to indicate dependency? As mentioned above, for OR it is easy to estimate confidence intervals. In the chart, three confidence intervals are shown for confidence level 90%, 95%, and 99%. Look at the examples (a)-(c) in the figure below. In (a) 1 is within all three confidence intervals i.e. the two classes are independent regardless of confidence level. In (b) The two classes are dependent or associated with a confidence level of 95% but not 99%. In (c) the classes are dependent with a confidence level of 99%.

A few illustrative examples: in (d) and (e) the numbers in the columns and the numbers in the rows are proportional. Evidently, the classes are then independent and $OR = 1$. In the last example, (f), the off-diagonal numbers are larger than the numbers of the diagonal. The dependence between the classes is strong illustrating the symmetrical nature of this analysis.



Final remarks

The motivation for creating a visualization based on the confusion matrix was to obtain some insights into the definitions and use of the measures based on the four numbers.

The elaborate writing down of the definitions of various ratios and the application of the esoteric names to the measures should be compared to the intuitive understanding of the stacked, normalized bars of the chart. Also, the concept of pair-pair plots as a simple mapping of two pair bars against each other adds clarity to the construction of the ROC plot and the precision-recall plot and the reason to use one or the other.

Finally, the addition of the odds ratio and the confidence intervals to the chart adds the ability to immediately make precise conclusions on the dependency or association of classes.

All these are insights at a glance. Feel free to download the Jupyter notebook and play around with the confusion matrix chart (https://github.com/SorenLaursen/confusion_matrix_chart).

. . .

[1] Confusion matrix, Wikipedia (https://en.wikipedia.org/wiki/Confusion_matrix)

[2] *Rafael A. Irizarry:* Introduction to Data Science (2020) — 15.10.4 The odds ratio, https://rafalab.github.io/dsbook/inference.html

[3] *Takaya Saito, Marc Rehmsmeier:* ROC and precision-recall with imbalanced datasets (2015), https://classeval.wordpress.com/simulation-analysis/roc-and-precision-recall-with-imbalanced-datasets/

Visualization Confusion Matrix Model Evaluation Classification Evaluation Metric