PERFORMANCE ASSESSMENT / Verification

Caren Marzban
http://www.nhn.ou.edu/~marzban

Formally

Observation x
Forecast f

Continuous, Categorical, Probabilistic.

Example A - continuous x, continuous f
Example B - binary x, binary f
Example C - binary x, probabilistic f

Performance is multifaceted.
Use diagrams.

Don't bet on a "better model".

Example A

x, f = daily T highs = 32, 33, 34, ...

Scatterplot of f vs. x.
Contingency Table:

$$\begin{pmatrix} n_{32,32} & n_{32,33} & \cdots \\ n_{33,32} & n_{33,33} & \cdots \\ \cdots & & \end{pmatrix},$$

$n_{32,33}$ = no. of days with $x = 32$, $f = 33$.

Diagonal is good.

Example B

x = non/existence of tornado = 0, 1 , f = 0, 1

$$\begin{pmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{pmatrix},$$

n_{01} = number of false alarms.

Example C

x = non/existence of tornado = 0, 1
f = prob intervals = 0.0-0.05, 0.05-0.15, 0.95-1.0

$$\begin{pmatrix} n_{0,0} & n_{0,1} & \cdots & n_{0,10} \\ n_{1,0} & n_{1,1} & \cdots & \end{pmatrix},$$

$n_{0,1}$ no. nontornadoes with forecast probs 0.05-0.15

Joint Distributions

All necessary information is contained in

$$p(x=i, f=j) = \frac{n_{ij}}{n_{..}},$$

$n_{.1} = n_{01} + n_{11}$.

Useful to break down into conditional probs:

$$p(x, f) = p(x|f)p(f) = p(f|x)p(x) \quad .$$

$p(x = 0|f = 1) =$ prob of a nontornado, given that forecast is tornado.

$p(x) =$ climatological prob of x.

Each of $p(x|f), p(f|x), p(f) \longrightarrow$ performance measure.

With data at hand, these probs are computable:

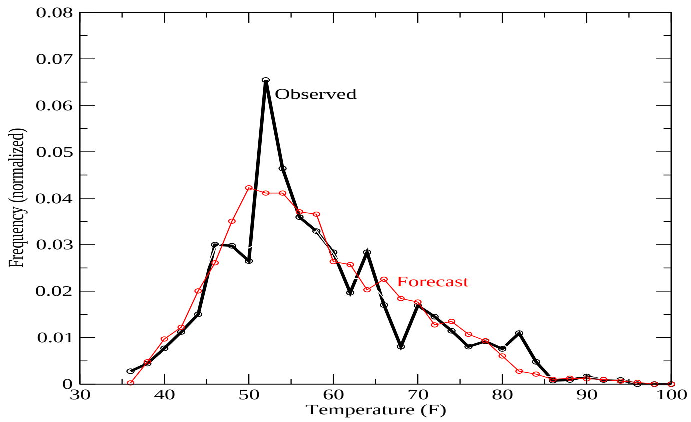
$$p(x=i|f=j) = \frac{n_{ij}}{n_{.j}} \quad , \quad p(f=j|x=i) = \frac{n_{ij}}{n_{i.}} \quad ,$$

and

$$p(x=i) = \frac{n_{i.}}{n_{..}} \quad , \quad p(f=j) = \frac{n_{.j}}{n_{..}} \quad .$$

Back to Example A

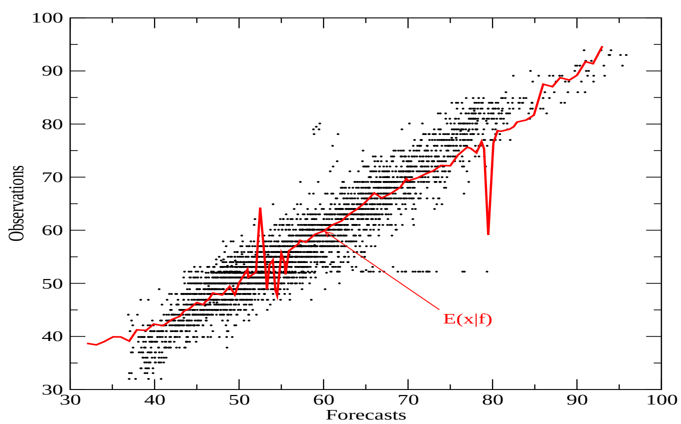
$p(x), p(f)$:



Shift $\longrightarrow$ bias.

Instead of $p(x|f)$ and $p(f|x)$, plot $E(x|f)$.

Average observations with f=32, 33, etc.



Reliable forecasts $\longrightarrow$ diagonal.

Scatterplots, residual plots, bias vs. variance.

Back to Example B

Three common (scalar) measures:

$$\text{Probability of Detection} = \frac{n_{11}}{n_{1.}}$$

$$\text{False Alarm Ratio} = \frac{n_{01}}{n_{.1}}$$

$$\text{False Alarm Rate} = \frac{n_{01}}{n_{0.}}$$

A scalar measure of *skill*:

$$\text{HSS} = \frac{2(n_{00}n_{11} - n_{01}n_{10})}{n_{0.}n_{.1} + n_{1.}n_{.0}}.$$

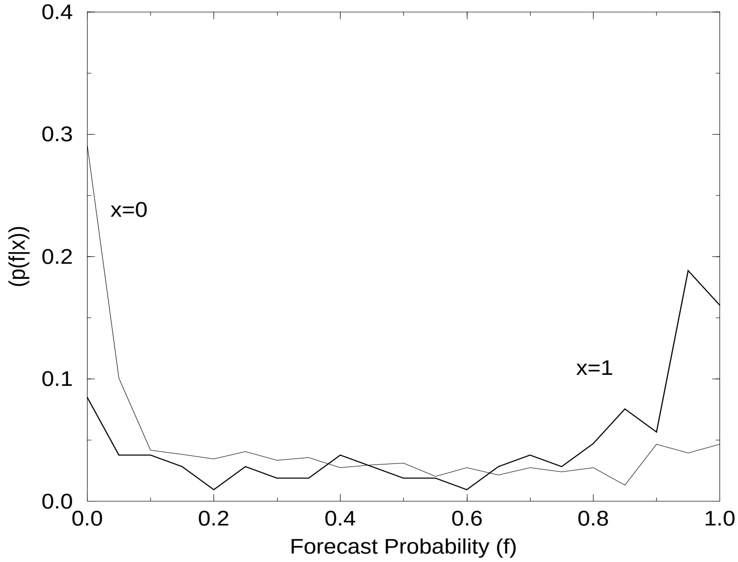
There are more, many more.
Equitable not unique.

$$FC = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}} = \frac{1 + \frac{n_{11}}{n_{00}}}{1 + \frac{n_{01}+n_{10}+n_{11}}{n_{00}}}$$

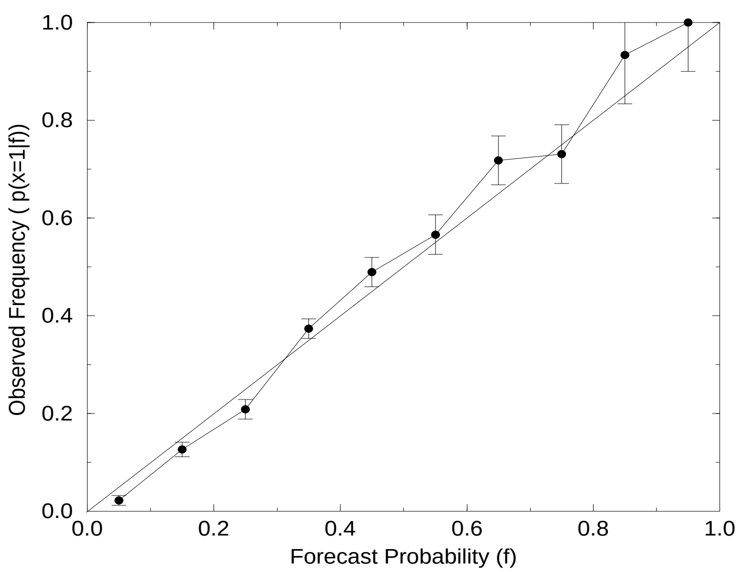
$$\rightarrow \frac{1+0}{1+0} \rightarrow 1.$$

Back to Example C

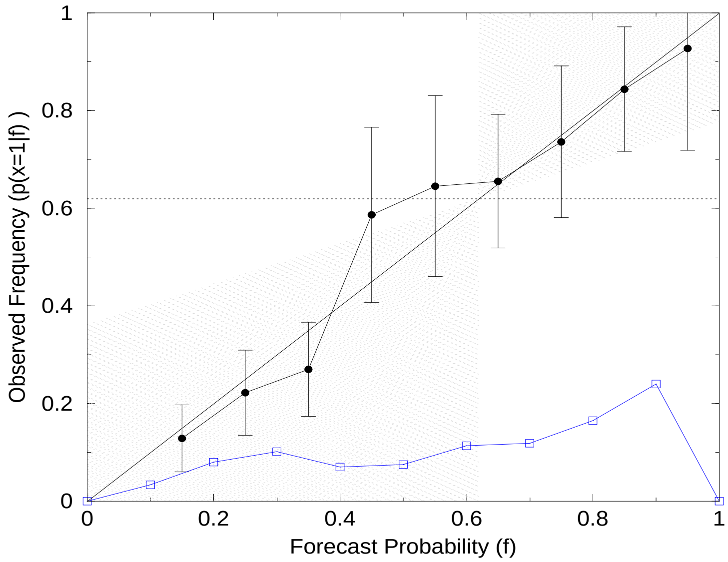
$p(f|x=0), p(f|x=1)$: discrimination



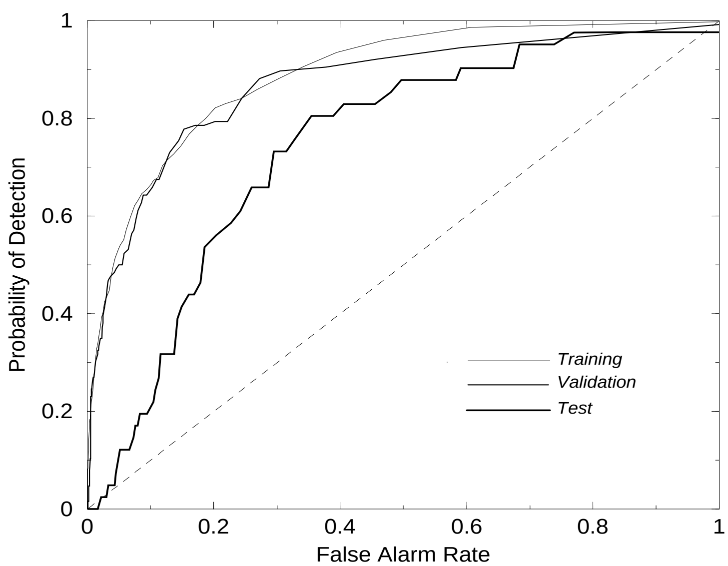
$p(x=1|f)$: reliability



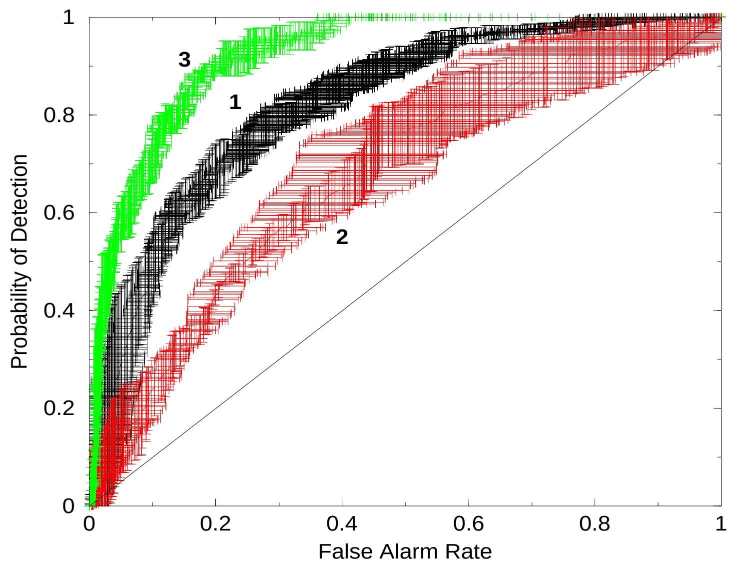
$p(f)$: refinement (sharpness)
Attributes diagram:



Finally, ROC



With error-bars:



Conclusion

- $p(x|f), p(f|x), p(f), (p(x))$ is all you need. Compute as ratios.

- Don't quantify too much (into scalars).

- Use Diagrams.

- Put error-bars.

Single Lesson: Performance is a multifaceted thing.