Model Calibration

Inbar Naor
Data Scientist @ Taboola

A Tale of Two Classifiers

Model A: 90% accuracy, 0.91 confidence in each prediction

Model B: 90% accuracy, 0.99 confidence in each prediction

Which is better?

Calibration:

Post processing a model to improve probability estimate

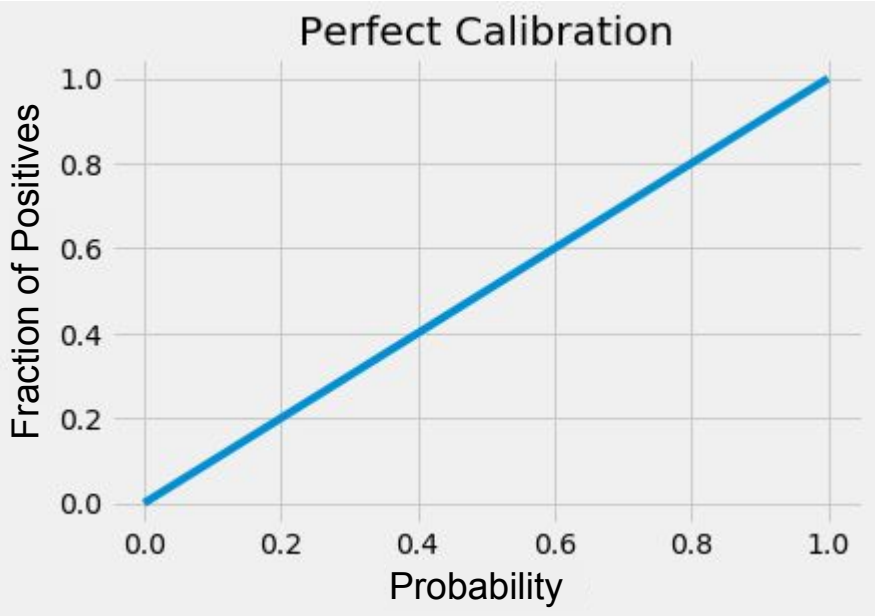
Among the samples that got a 0.8 probability of being positive, we expect 80% to be positive.

Intuition

$$\mathbb{P}(\hat{Y} = Y|\hat{P} = p) = p,$$

$$\forall p \in [0, 1]$$

A model is perfectly calibrated if, for any p, a prediction of a class with confidence p is correct 100*p percent of the time.



Agenda

- Why is that important?
- How do we know if our model is calibrated?
- Calibration of different models
- Calibration methods:
  - Platt Scaling
  - Isotonic Regression

Why is calibration important?

Calibration is important only if probabilities are important.

- High risk applications
- Combining with other probability models, thresholding
- Improving our models:
  - Mistakes with high probabilities
  - True labels with low probabilities

Calibration != accuracy

- Well calibrated model can have low accuracy
  - Example: random coin
- Models can have high accuracy and bad calibration

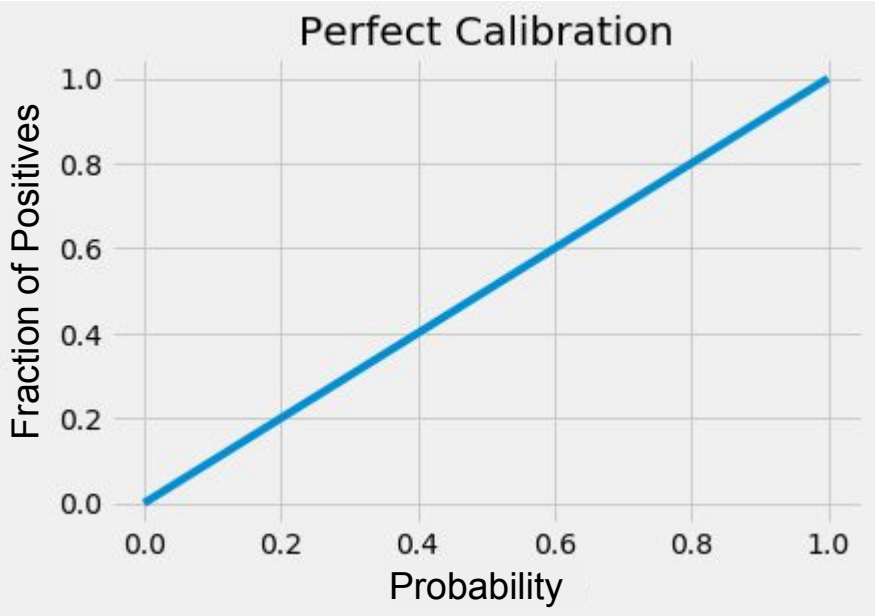
How Do We know If Our Model Is Calibrated?

Intuition

$$\mathbb{P}(\hat{Y} = Y|\hat{P} = p) = p,$$

$$\forall p \in [0, 1]$$

A model is perfectly calibrated if, for any p, a prediction of a class with confidence p is correct 100*p percent of the time.



Reliability Plots

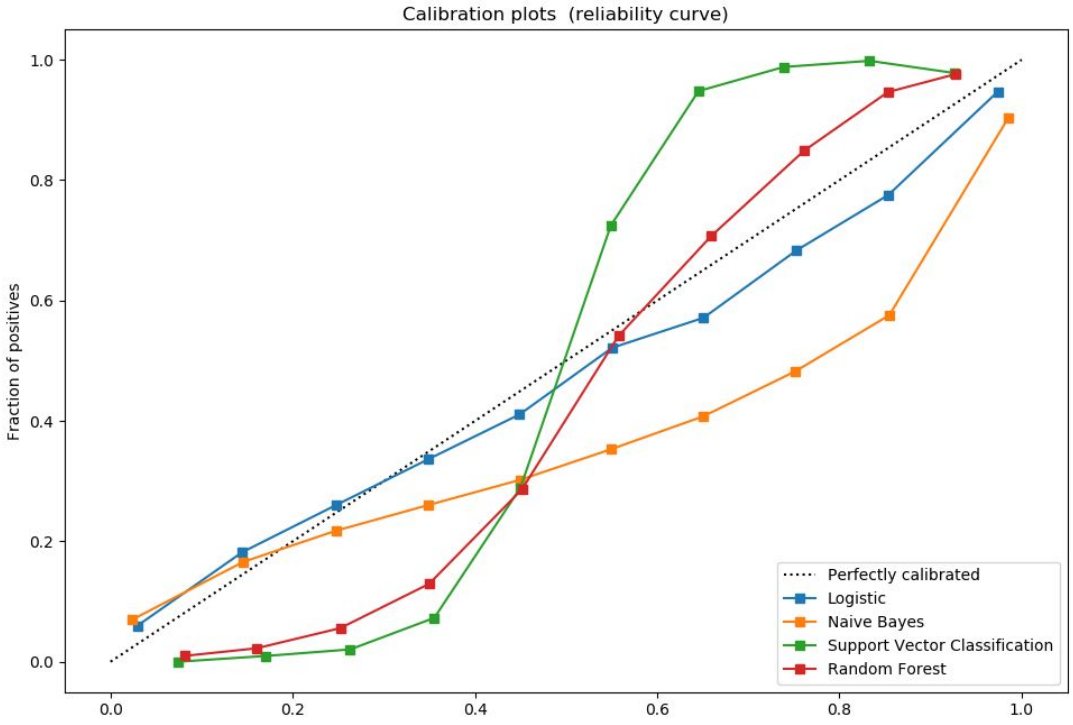
1. Group predictions into bins
2. Fraction of positives per bin
3. Confidence per bin:

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$

Reliability Plots

1. Group predictions into bins
2. Fraction of positives per bin
3. Confidence per bin:

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$



What If I want a Number?

Expected Calibration Error

$$\mathop{\mathbb{E}}_{\hat{P}}\left[\left|\mathbb{P}\left(\hat{Y}=Y \mid \hat{P}=p\right)-p\right|\right]$$

Use approximation

$$\mathrm{ECE}=\sum_{m=1}^{M}\frac{|B_m|}{n}\left|\mathrm{acc}(B_m)-\mathrm{conf}(B_m)\right|$$

Maximum Calibration Error

$$\max_{p\in[0,1]}\left|\mathbb{P}\left(\hat{Y}=Y \mid \hat{P}=p\right)-p\right|$$

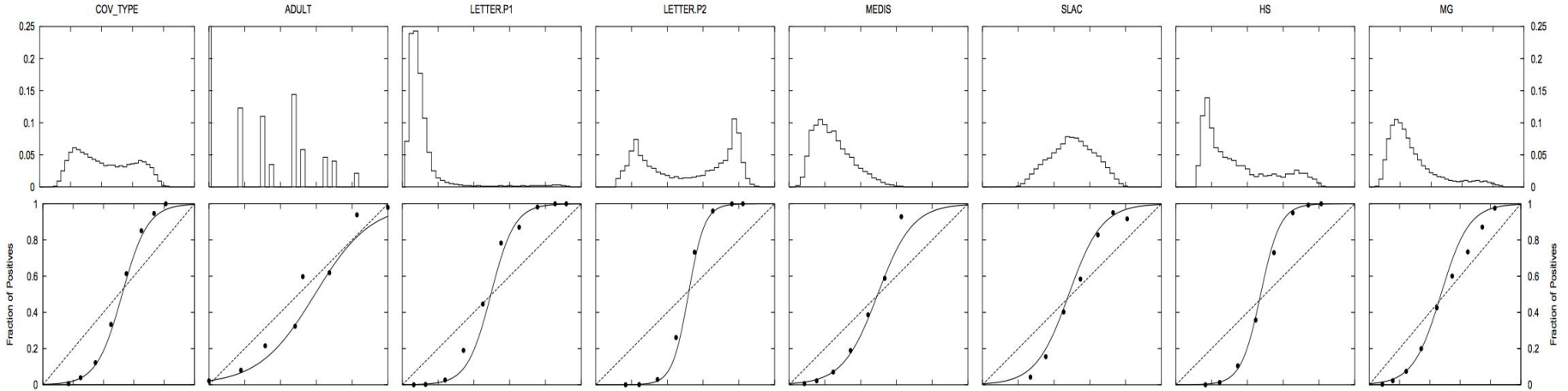
Use approximation

$$\max_{m\in\{1,\ldots,M\}}\left|\mathrm{acc}(B_m)-\mathrm{conf}(B_m)\right|$$

Calibration Of Different MOdels

Predicting Good Probabilities with Supervised Learning,
Niculescu-Mizil & Caruana

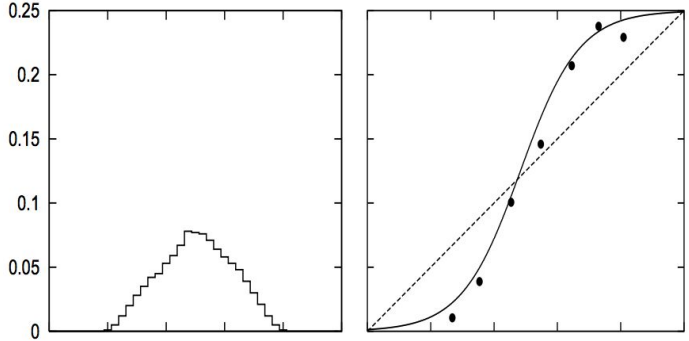
Boosted Trees



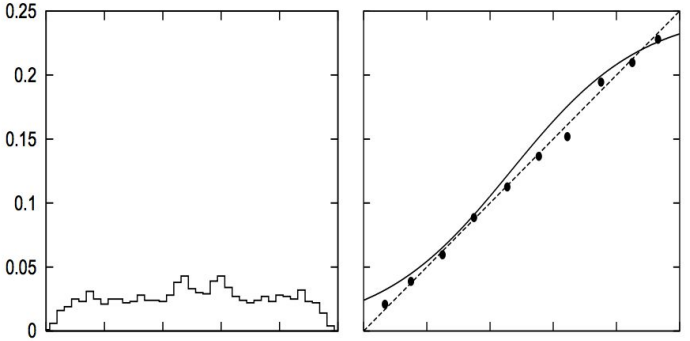
- Sigmoidal shape reliability plots
- Values pushed away from 0-1

Calibration Of DifferenT Classifiers

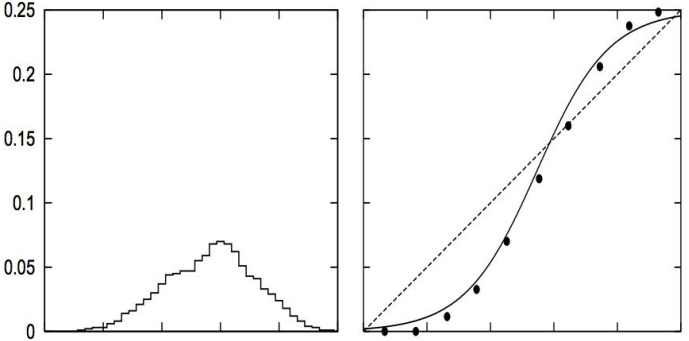
Boosted Trees



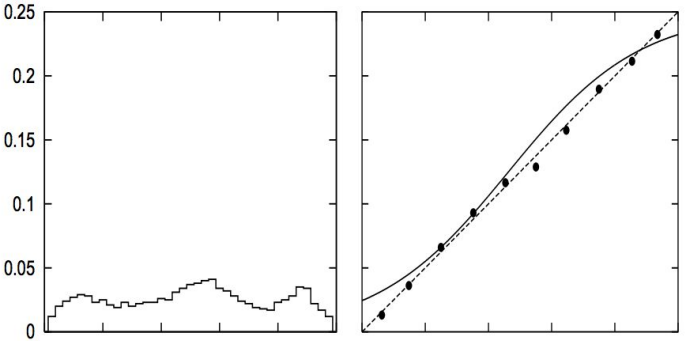
Logistic Regression



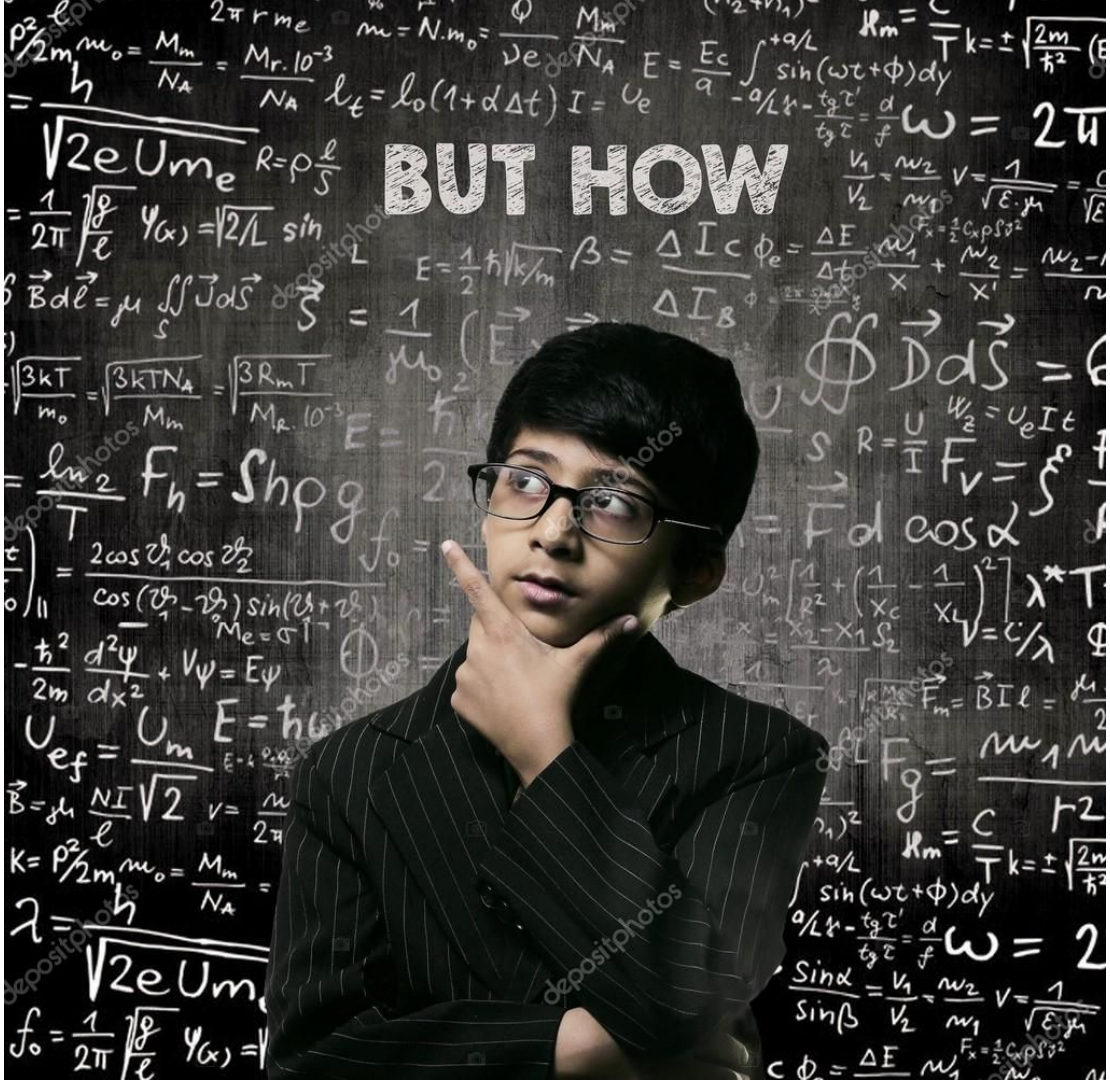
SVM



Neural Network *



* before 2006


BUT HOW

Platt Scaling

- New data set
- Fit logistic regression on the output of the model

```
from sklearn.linear_model import LogisticRegression as LR

lr = LR()
lr.fit(p_holdout, y_holdout)
calibrated_p = lr.predict_proba(p_test)[:, 1]
```

Isotonic Regression

Learn a monotonic piecewise constant function f to transform uncalibrated outputs.

```
from sklearn.isotonic import IsotonicRegression as IR

ir = IR()
ir.fit(p_holdout, y_holdout)
calibrated_p = ir.transform(p_test)
```

Summary

- Calibration is important if you care about probabilities
- Reliability plots
- Different methods for calibration

Questions?