Introduction to Data Science

EVALUATION FOR MACHINE LEARNING: METRICS

BRIAN D'ALESSANDRO

REMINDER

You will never build the *perfect* model… but we can always have a *best* model, given the data and constraints.

So far we have discussed the following design options:

[Algorithm, Feature Set , Hyper-parameters (complexity)]

We also need to choose an evaluation metric!

THE RIGHT METRIC DEPENDS ON YOUR GOALS

- **Ranking** - Who are the top k prospects for my campaign, or what 10 items should I recommend?

- **Classification** – Is this email spam or not? Is this number a ‘1’ or a ‘7’?

- **Density Estimation** – What is the probability that this transaction is fraud? What is the expected spend of a newly acquired customer?

METRICS FOR THESE GOALS

Ranking

Area under the Receiver Operator Curve (AUC)
Area under the Cumulative Lift Curve (ACLC)

Classification

Lift (LFT)
Accuracy (ACC)
F-Score (FSC)
Precision (PRE)
Recall (RCL)

Density Estimation

Mean Absolute Error (MAE)
Mean Squared Error (MSE)
Cross-Entropy /Log-Likelihood (LL)

TRAIN VS. VAL/TEST LOSS

We don't need to use the same error/loss/risk function on our training data as we do our validation or test data.

Training Loss

$$R_{train} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{L}(f(x_i^{train}), y_i^{train})$$

Usually used because they are well posed and "easy" to actually optimize (i.e., convex and smooth).

Logistic Loss, Hinge Loss, Least Squares etc.

vs

.

Testing Loss

$$R_{test} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{L}(f(x_i^{test}), y_i^{test})$$

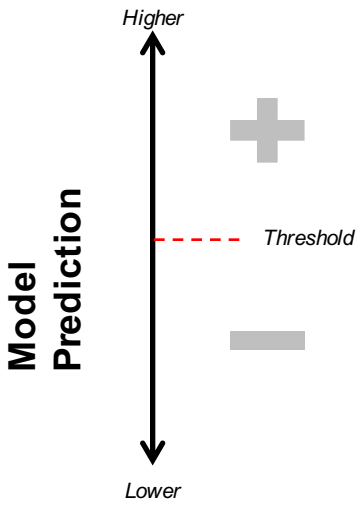
Often the most suitable loss function is not convex, not differentiable or too complex to train efficiently.

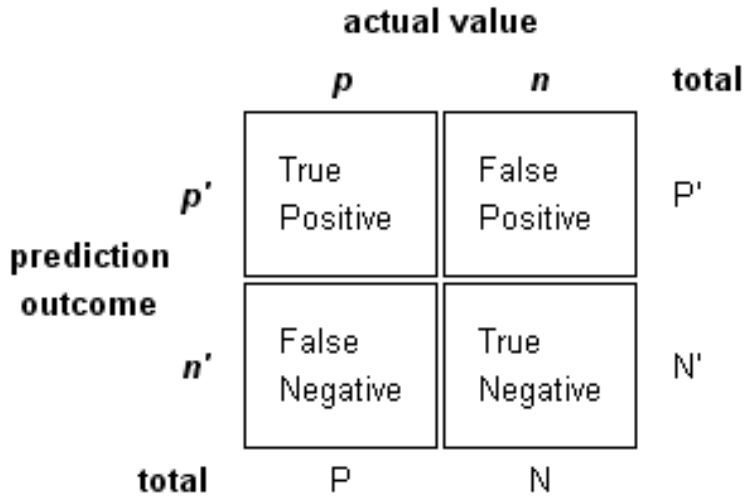
Precision, AUC, Recall.

CONFUSION MATRIX

Many of the metrics we use derive from the confusion matrix. For binary classification we assume there exists some real valued function f(x) and a decision threshold δ.

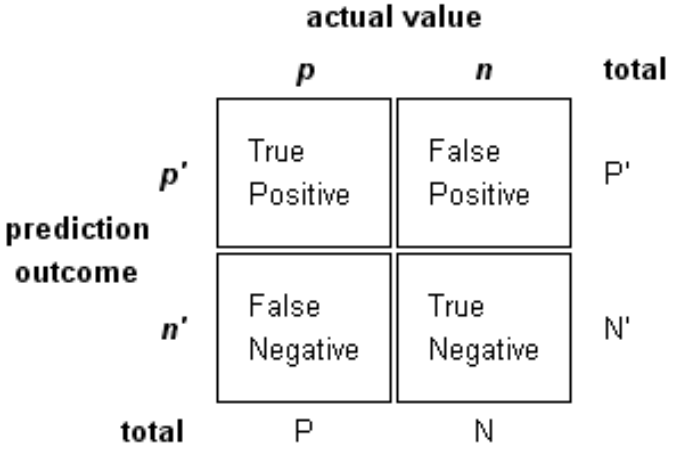
$$\hat{Y} = I(f(x) > \delta)$$





Classification Metrics

We can derive many classification metrics from the confusion matrix.



Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_curve

Terminology and derivations from a confusion matrix

true positive (TP)
eqv. with hit

true negative (TN)
eqv. with correct rejection

false positive (FP)
eqv. with false alarm, Type I error

false negative (FN)
eqv. with miss, Type II error

sensitivity or true positive rate (TPR)
eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

false positive rate (FPR)
eqv. with fall-out

$$FPR = FP/N = FP/(FP + TN)$$

accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

specificity (SPC) or True Negative Rate

$$SPC = TN/N = TN/(FP + TN) = 1 - FPR$$

positive predictive value (PPV)
eqv. with precision

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

false discovery rate (FDR)

$$FDR = FP/(FP + TP)$$

Matthews correlation coefficient (MCC)

$$MCC = (TP * TN - FP * FN)/\sqrt{PNP'N'}$$

F1 score

$$F1 = 2TP/(P + P') = 2TP/(2TP + FP + FN)$$

Source: Fawcett (2006).

SOME EXPOSITION ON COMMON METRICS

Accuracy (0/1 Loss) – The most intuitive and well known, but not commonly used outside of multi-class problems.

Precision - of all the instances I predict are positive, how many actually are positive? Best used when False Positives are relatively more expensive and need to be avoided

Recall- how many of the total positives out there did I classify as positive? Best used when False Negatives are relatively more expensive and need to be avoided

These are the most common metrics for classification problems. One shouldn't hesitate to use them, but it is important to know when they're most appropriate, as well as key caveats.

These are all dependent on the threshold used for the classifier. You can often improve one (but not all) by moving the threshold.

They are also base rate dependent. The quality of the score itself is hard to gauge without comparing to a baseline.

F1-SCORE

If there is good reason to favor both Precision and Recall, one could use the F1-score, the harmonic mean of the two:

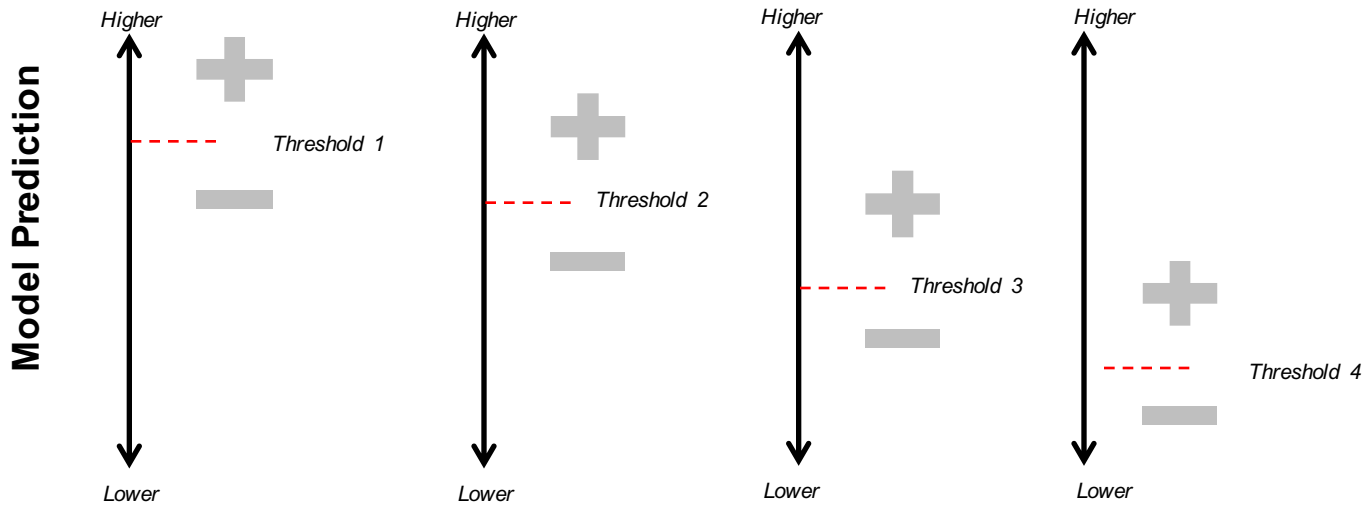
$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

A more flexible variant of the F1-Score is the F-Beta. Beta is chosen such that the Recall (controlling FN's) is considered β times more important than Precision (controlling FPs):

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}.$$

TOWARDS A RANKING METRIC

Classification metrics depend on choosing a single threshold. But what if you don't know or need the threshold?

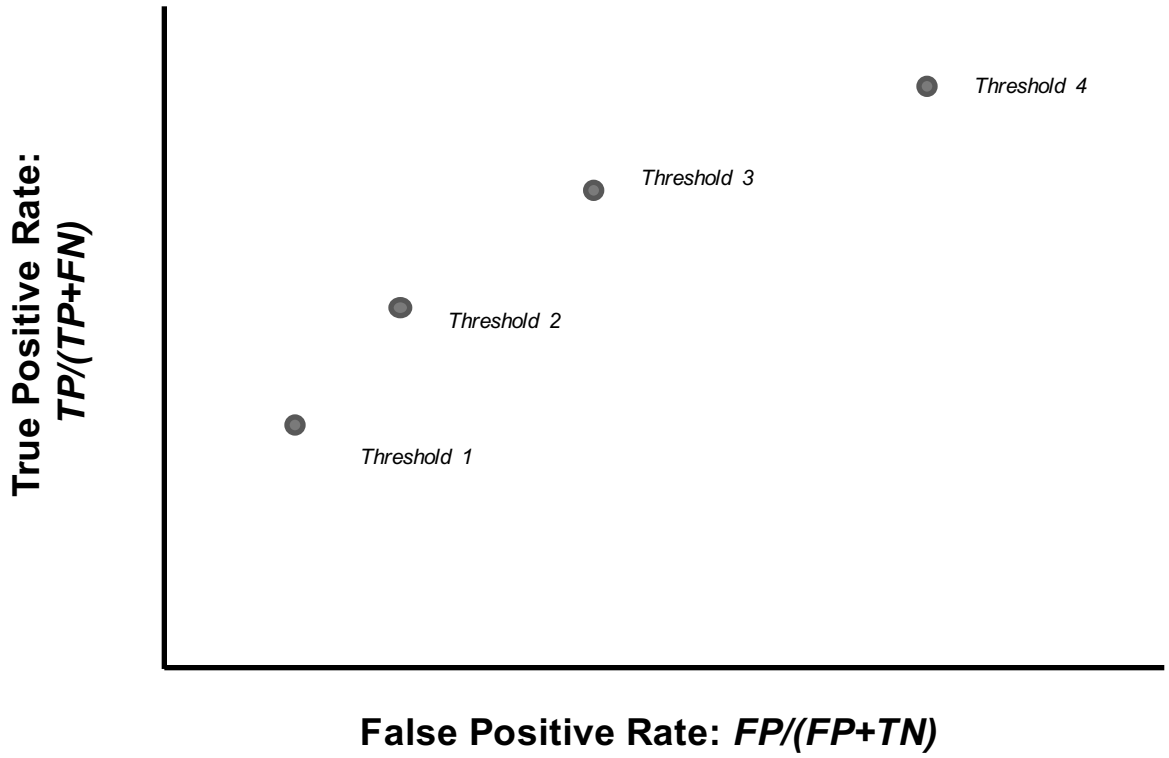


For each threshold we will get different accuracy, lift, precision and recall.

We want an evaluation method that considers the trade-off on these metrics when using different thresholds.

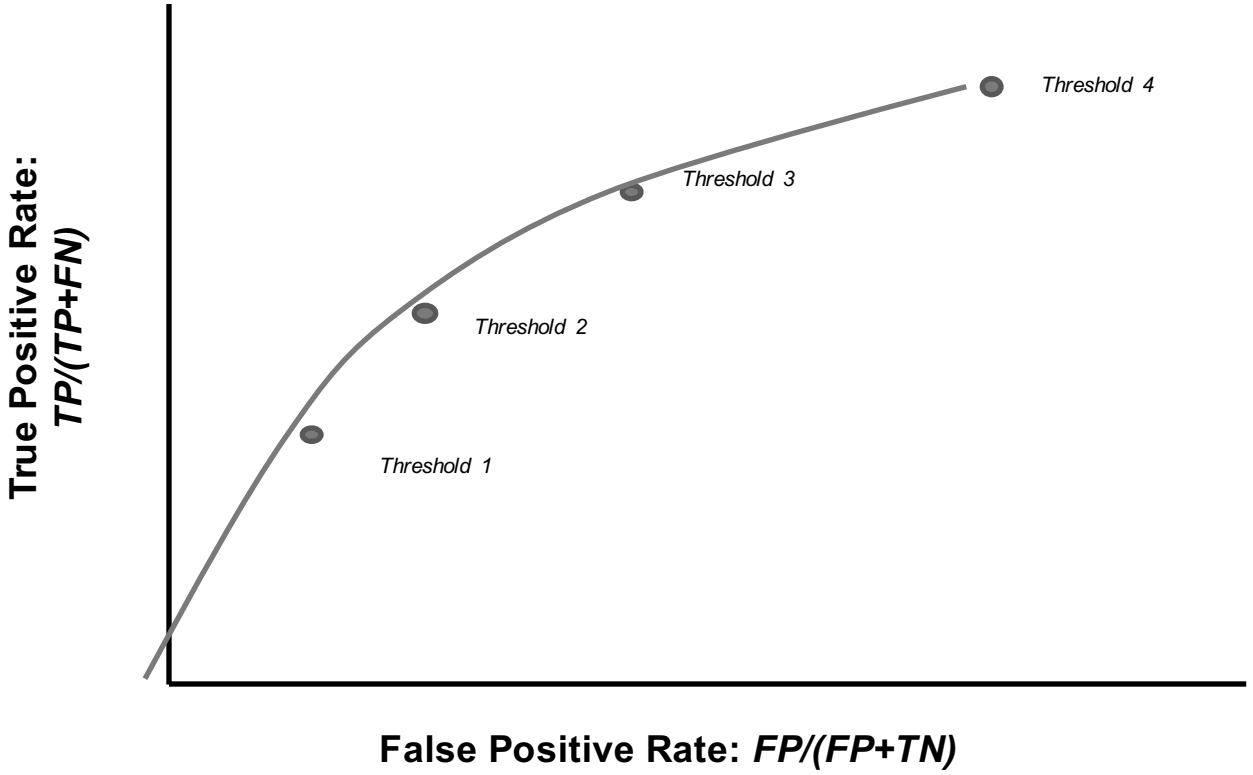
THE THRESHOLDING TRADE-OFF

Each threshold we choose creates a trade-off between false positive rate and true positive rate.



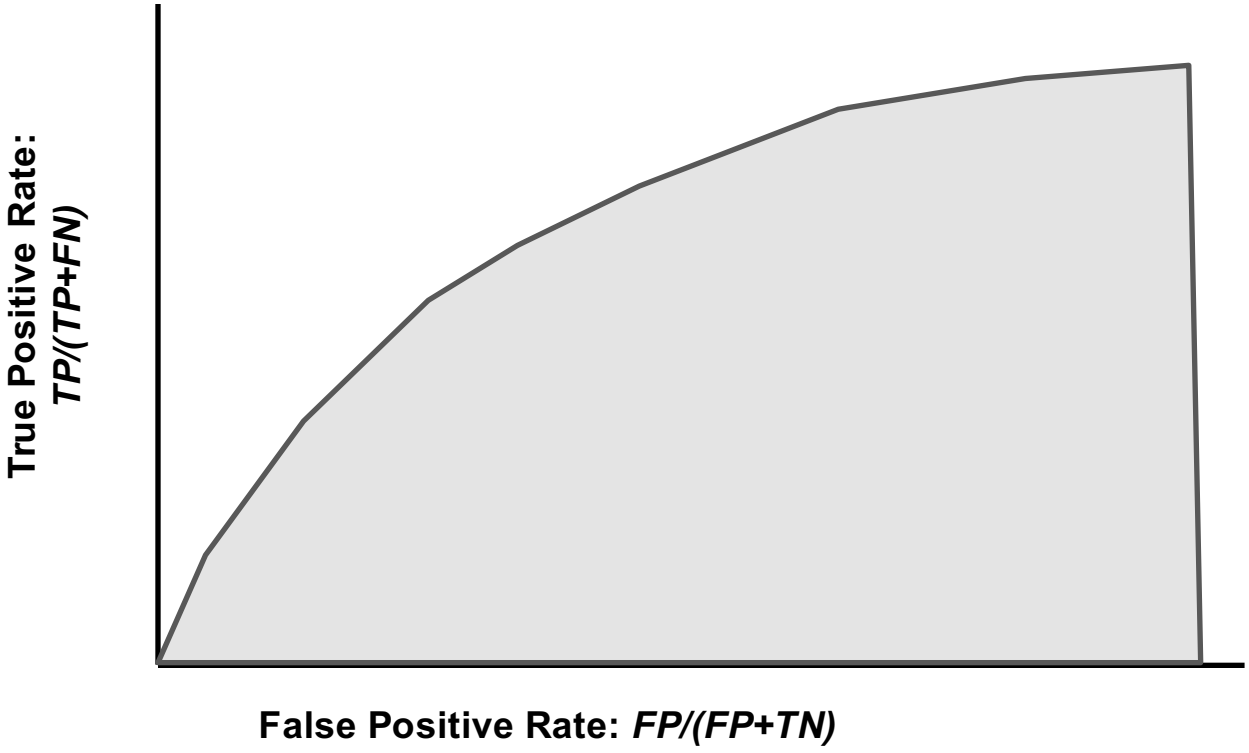
THE ROC CURVE

If we consider every threshold and plot the trade-off, we arrive at the ROC curve.



THE AREA UNDER THE ROC CURVE

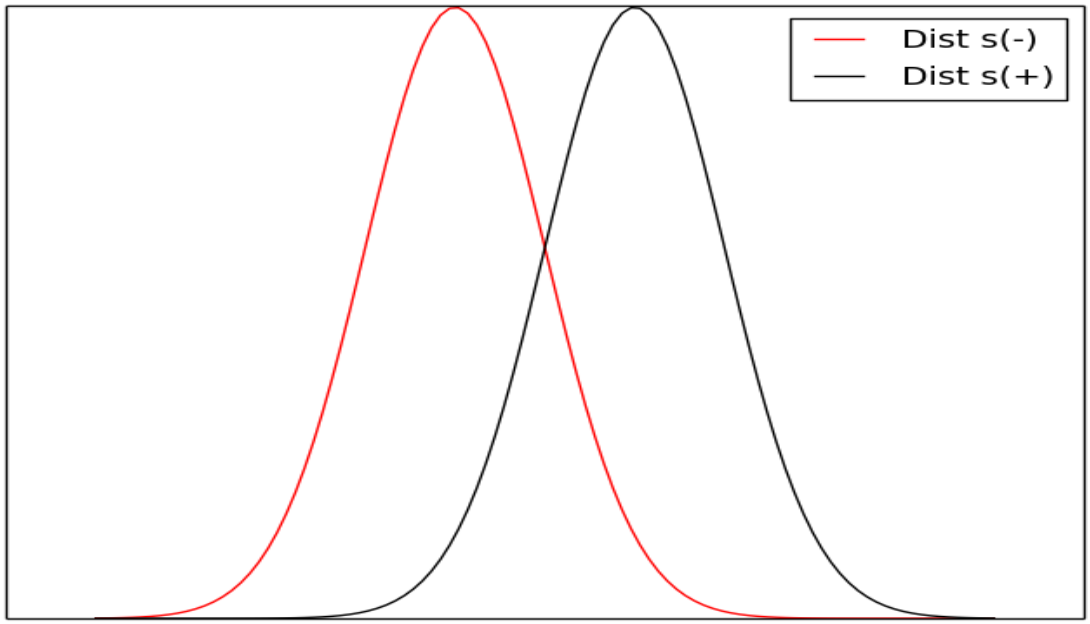
The area under this curve gives a comprehensive summary of how well your classifier ranks.



PROBABILISTIC INTERPRETATION OF AUC

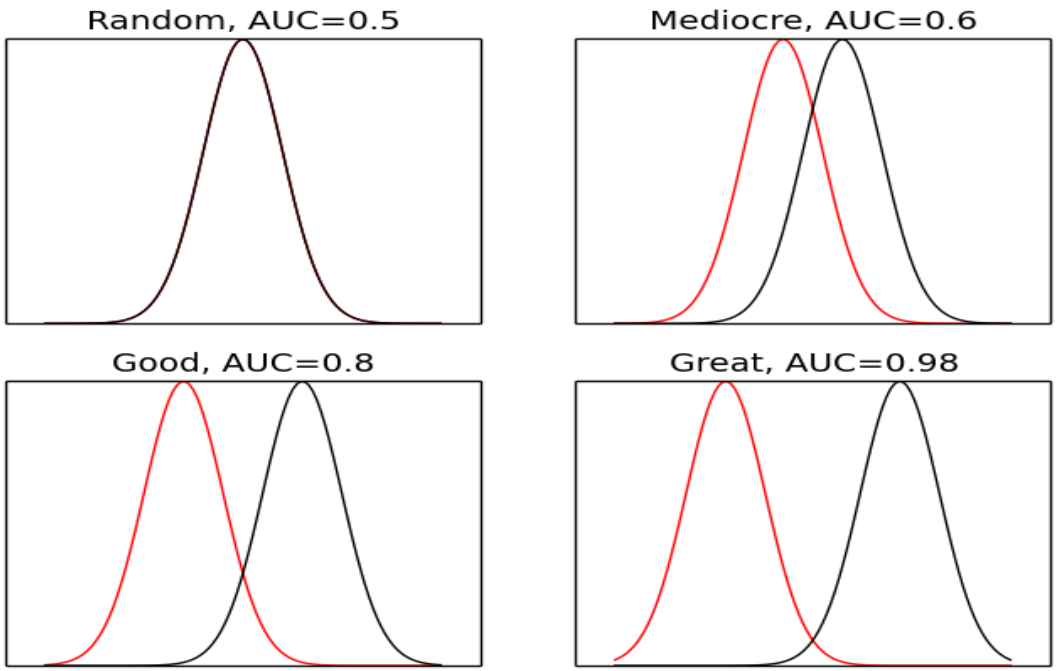
Let $s^+(x)$ and $s^-(x)$ be the PDF of $f(x)$ for positive and negative labels, respectively. Let $S^+(x)$ and $S^-(x)$ be the CDF of $f(x)$ for positive and negative labels, respectively.

$$AUC = P(f(x^+) > f(x^-)) = \int_{-\infty}^{\infty} s^+(x)\, S^-(x)\, dx = \int_{-\infty}^{\infty} s^-(x)\, (1 - S^+(x))\, dx$$



DIFFERENT EXAMPLES

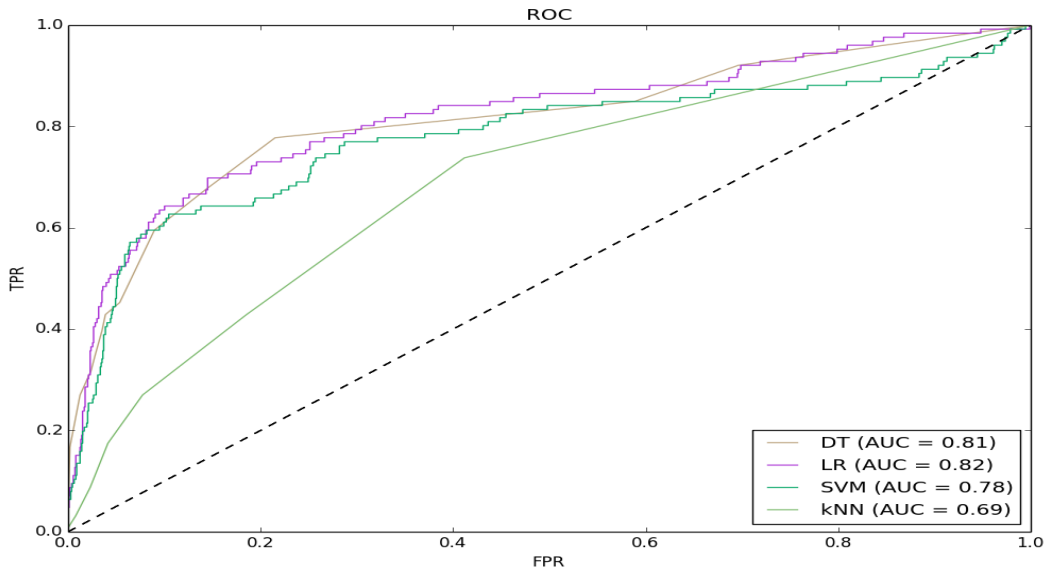
A good AUC depends on the problem. Although AUC=0.6 means reasonably bad separation of the classes, it could still create value. Also, really high AUCs are often too good to be true and should be treated with suspicion.



COMPARING AUCS

We built 4 different classifiers using the ads dataset. We can compare the models using ROC analysis.

- A universally better model has higher TPR at all FPR (LR > kNN)
- Some models overlap. Better model depends on whether you value TPR or FPR more (DT is best where FPR<0.05)



FUN AUC FACTS

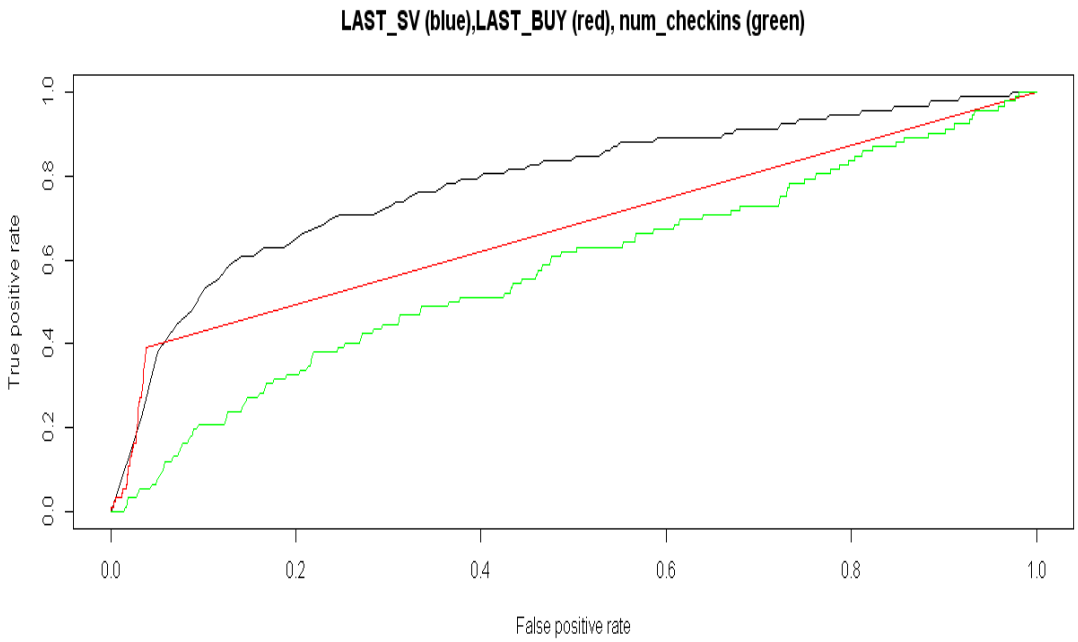
• **Nice interpretation**: gives the probability that a positive instance will have a higher score than a negative instance

• **Base Rate Invariant:** AUC is invariant to P(+) in the data set (unlike other classification metrics). Useful for doing comparisons across data sets with different base rates. Or after down sampling.

• **Is Nicely Bounded:** AUC scores range from [0,1], where 1 is a perfect classifier and 0 is a perfectly wrong classifier. A random classifier has an exact score of 0.5.

USING ROC FOR FEATURE RANKING

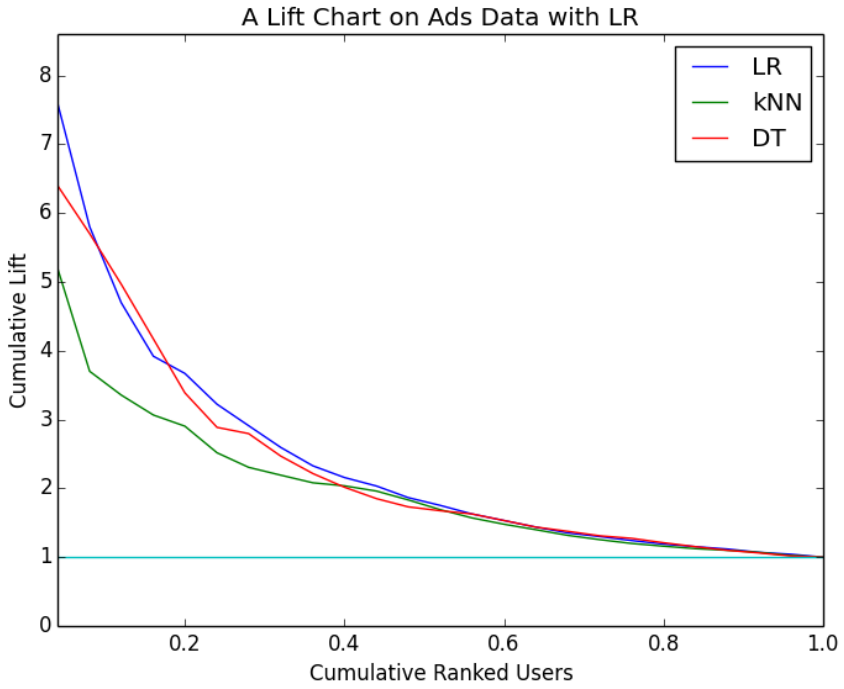
We can analyze the predictive power of individual features using AUC curves. Note the interesting shape of LAST_BUY AUC. What causes that?



To do this we don't even need a model. Let your feature value itself be the prediction, and call roc_auc_score(Y_True, X)

LIFT

Lift can be both a ranking metric and a classification metric. For ranking,
we can see which model fits the entire distribution of users better. For single classification,
we can measure lift for a desired targeting threshold.



Lift Properties

- **Nice interpretation**: the lift tells you exactly how many more positive outcomes you might expect relative to the baseline strategy. Also lends well to economic analysis

- **Not Base Rate Invariant:** Lift will change if you alter P(+). This has implications for down sampling or for comparing models from different datasets.

DENSITY ESTIMATION

Sometimes you want to evaluate how well your model estimates the underlying conditional distribution of your data: P(Y|X)

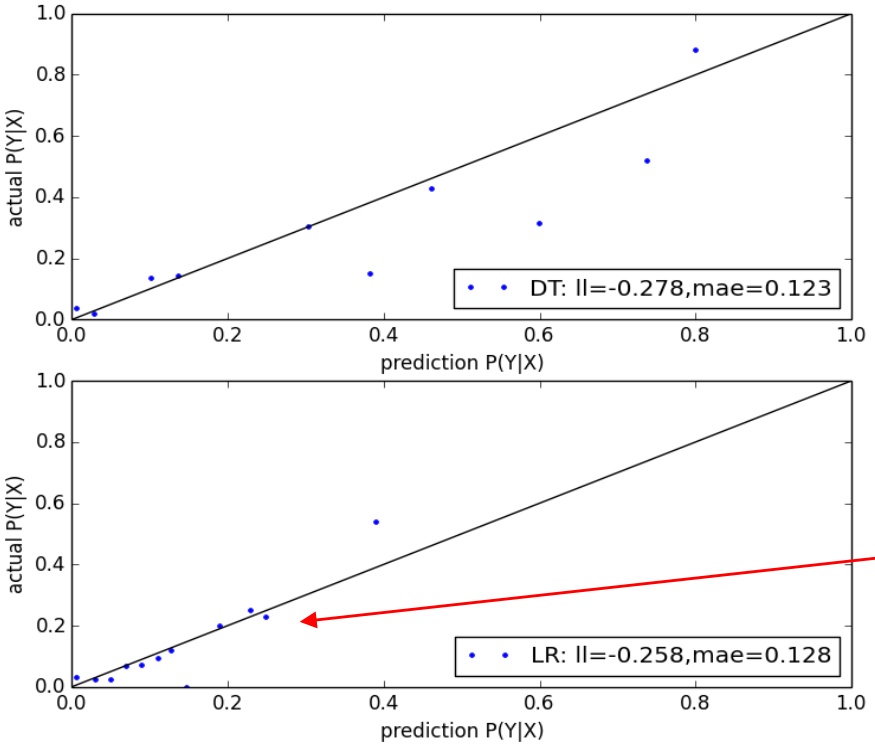
$$MAE = \frac{1}{n}\sum_{i=1}^{n} |E[y|x_i] - y_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (E[y|x_i] - y_i)^2$$

$$LL = \frac{1}{n}\sum_{i=1}^{n} y_i \, ln(P(y|x_i)) + (1 - y_i) \, ln(1 - P(y|x_i))$$

CALIBRATION PLOTS

A good way to test our ability to estimate probabilities is to generate calibration plots. To make the plot, we bin test instances by P(Y|X) and take mean(Y) against mean(P(Y|X)) for each bin.



Observations for this problem:

- DT predicts a higher range of probabilities
- LR is very well calibrated for lower valued predictions but not as good in the upper range.

We want to see our points line up against the identity line as much as possible.

METRICS DON'T ALWAYS AGREE

It is often the case that different metrics don't agree (in terms of rank) when comparing models built with different design choices.

In this case we build univariate LR models on each feature and compare AUC, LL and Gini Index (from SK Learn Decision Tree)

Feature	AUC	-LL	Gini
visit_freq	0.781	0.282	0.147
last_visit	0.780	0.306	0.528
multiple_visit	0.740	0.280	0.000
sv_interval	0.717	0.323	0.046
buy_freq	0.673	0.274	0.151
isbuyer	0.670	0.278	0.000
last_buy	0.665	0.321	0.015
buy_interval	0.581	0.310	0.000
multiple_buy	0.581	0.297	0.000
uniq_urls	0.580	0.322	0.051
num_checkins	0.567	0.326	0.062
expected_time_visit	0.564	0.329	0.000
expected_time_buy	0.518	0.327	0.000

Why does this matter?

It is important to choose the right metric for your problem. The optimal model under one metric may be different than the optimal model for the metric you care the most about.

AN ADVERTISING EXAMPLE: 1

Scenario 1:

Constraints: *a budget k and a population n (k and n on the same unit scale)*

Goal: *Maximize the ROI for the client*

Solution: *Target (k/n)% of the population, such that the selected set of k prospects maximizes the total number of conversions*

What metrics can we use to choose the best model

AN ADVERTISING EXAMPLE: 1

Scenario 1:

Constraints: *a budget k and a population n (k and n on the same unit scale)*

Goal*: Maximize the ROI for the client*

Solution: *Target (k/n)% of the population, such that the selected set of k prospects maximizes the total number of conversions*

If we know *k* and *n*: **Lift or Precision**
If we don't know *k* and *n*: **AUC**

AN ADVERTISING EXAMPLE: 2

Scenario 2:

Constraints: *each impression costs $C, for each conversion, receive $Q, unlimited budget*

Goal: *Maximize profit for the firm*

Solution: *Target every opportunity where*
E[Value]= P(Conv|X)$Q>$C*

What metrics can we use to choose the best model

AN ADVERTISING EXAMPLE: 2

Scenario 2:

Constraints: each impression costs $C, for each conversion, receive $Q, unlimited budget

Goal: Maximize profit for the firm

Solution: Target every opportunity where
E[Value]= P(Conv|X)*$Q>$C

To get a well calibrated estimate of P(Y|X), use

$$LL = \frac{1}{n} \sum_{i=1}^{n} y_i \, ln(P(y|x_i)) + (1 - y_i) \, ln(1 - P(y|x_i))$$