Data Analysis: Measuring Model Performance

Technological University Dublin Tallaght Campus

Department of Computing

Measuring Model Performance

How the performance of a model is measured depends on

- the type of model
  - classification
  - regression
- the purpose of the model, e.g.
  - general model for use with different data sets
  - working out the profit resulting from the application of a model to a particular data set

Evaluating Classification Models

Confusion Matrix

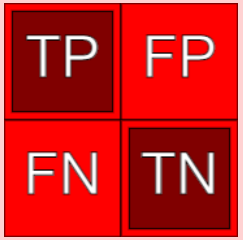
Many metrics use the values that are part of the **confusion matrix**

- number of rows and number of columns equal to number of target values
- the matrix size is $n \times n$ where n is the number of classes
- shows counts for *actual value/predicted value* combinations
- the top-left to bottom-right diagonal contains correct prediction counts
- two-value target (binary prediction) confusion matrix is the most common:

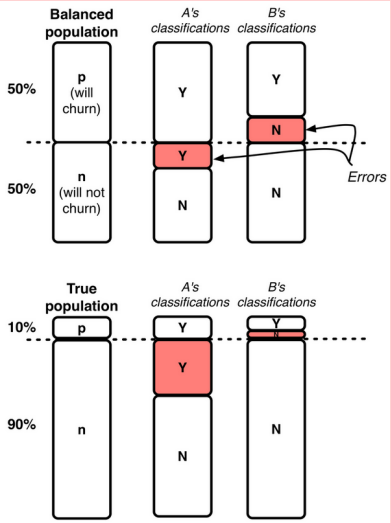
The four values here are usually denoted **TP (true positive)**, **FP (false positive)**, **FN (false negative)** and **TN (true negative)**. The 'positive' value is arbitrary, but usually indicates the value that represents 'detection' e.g. of a disease (in medical application) or a fraudulent transaction (in banking).

Measures Derived from Confusion Matrix

- **accuracy**
  - most commonly used measure for quick evaluation
  - measures the proportion of predictions that are correct
  - a 'blunt instrument'
  - but easily calculable $accuracy = \dfrac{TP + TN}{TP + FP + FN + TN}$



  - depends on the particular mix of classes in the test dataset



In the picture model A predicts positives perfectly but predicts negatives with 60% accuracy. B predicts negatives perfectly but predicts positives with 60% accuracy. The result is that when the class mix is 50-50 the overall accuracy is the same for the two models but when the mix is 10-90 model B comes out as a lot better.

- the 'rates'

true positive rate: $\dfrac{TP}{TP + FN}$

false negative rate: $\dfrac{FN}{TP + FN}$

true negative rate: $\dfrac{TN}{FP + TN}$

false positive rate: $\dfrac{FP}{FP + TN}$

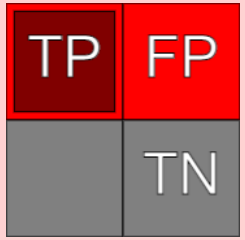
- statistics measures

sensitivity (equal to the *true positive rate*): $\dfrac{TP}{TP + FN}$

specificity (equal to the *true negative rate*): $\dfrac{TN}{FP + TN}$

- text classification and information retrieval measures

precision (or **positive predictive value**): $\dfrac{TP}{TP + FP}$



recall (equal to the *true positive rate*): $\dfrac{TP}{TP + FN}$

F-measure (harmonic mean of precision and recall):

$$F\text{–}measure = 2 \times \frac{precision \times recall}{precision + recall}$$

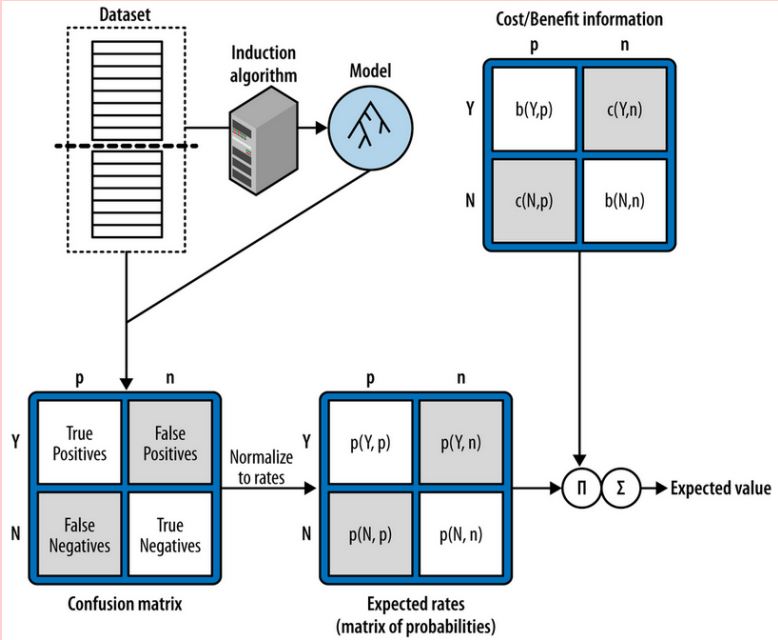
Taking into acount cost and benefit

- **Expected value** in general is a value given to a situation that has a countable number of known outcomes with known probabilities. If the value of each outcome is also known or can be estimated, then the expected value is:

$$EV = p(o_1)v(o_1) + p(o_2)v(o_2) + ... + p(o_n)v(o_n)$$

  where o_1, o_2, ... o_n are the outcomes, n is the number of outcomes, $p(o_i)$ is the probability of outcome o_i and $v(o_i)$ is the value of outcome o_i.
- In a data analytical context the values of $p(o_i)$ and $v(o_i)$ are determined as follows:
  - the **probabilities** are estimated from the data, by building a model
  - the **value** of each outcome must be determined using:
    - general knowledge of the problem domain
    - particular knowledge of the specific problem at hand

- *Expected value* can be calculated using the confusion matrix



The confusion matrix is *normalised* i.e. the counts are converted to probabilities for the different outcomes. A matrix of the same dimensions is 'filled in' with cost and benefit values. The cell-for-cell multiplication of the matrices and summation of the values in the cells of the resulting matrix yields the *expected value*. Beware of: **duplication of values** and **cost/benefit sign mix-ups**.

$$expected\_value = p(\mathbf{Y},\mathbf{p})b(\mathbf{Y},\mathbf{p}) + p(\mathbf{N},\mathbf{p})c(\mathbf{N},\mathbf{p}) + p(\mathbf{N},\mathbf{n})b(\mathbf{N},\mathbf{n}) + p(\mathbf{Y},\mathbf{n})c(\mathbf{Y},\mathbf{n})$$

- Another form of the same equation has the a priori probabilities of **p** and **n** separated out, for easy calculation for data sets with different class mixes

$$\begin{aligned} expected\_value = & p(\mathbf{p})[p(\mathbf{Y}|\mathbf{p})b(\mathbf{Y},\mathbf{p}) + p(\mathbf{N}|\mathbf{p})c(\mathbf{N},\mathbf{p})] \\ & + p(\mathbf{n})[p(\mathbf{N}|\mathbf{n})b(\mathbf{N},\mathbf{n}) + p(\mathbf{Y}|\mathbf{n})c(\mathbf{Y},\mathbf{n})] \end{aligned}$$

Example

Expected value can be used to determine whether some action would be viable and in what cases. In **targeted marketing** the action is the sending of marketing material to a particular person or class of person. When working with a predictive model that outputs the probability of a person or class of person responding to marketing, making a targeting decision based on the intuitive threshold of 0.5 would not be very useful, as all the probabilities would be a lot lower than that threshold. In such cases, it is a lot more useful to look at the expected value (i.e. benefit) of targeting rather than use a simple threshold.

$$expected\_value\_of\_targeting = p_R(x)v_R + [1 - p_R(x)]v_{NR}$$

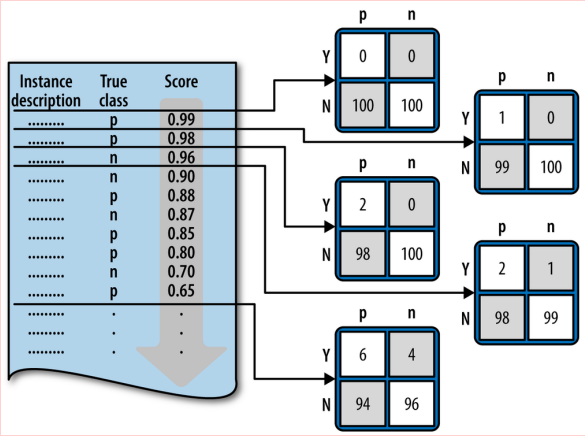
In the formula above, $p_R(x)$ is the probability of x (a person or a class of person) responding, v_R is the benefit of a response and v_{NR} is the value (which will be negative) of a non-response. Let's say that the targeted marketing is for a product that would bring a profit of €100 if the targeted person bought it (responding to the marketing action) and that the price of creating and sending the material is €1 per address. Then we would have:

$$v_R = €100 - €1 = €99, \qquad v_{NR} = -€1$$

For the values above, the expected value of targeting is greater than 0 when $p_R > 0.01$. Thus we may decide to target any person for whom this condition is met, expecting that such action would, on average, be of benefit.

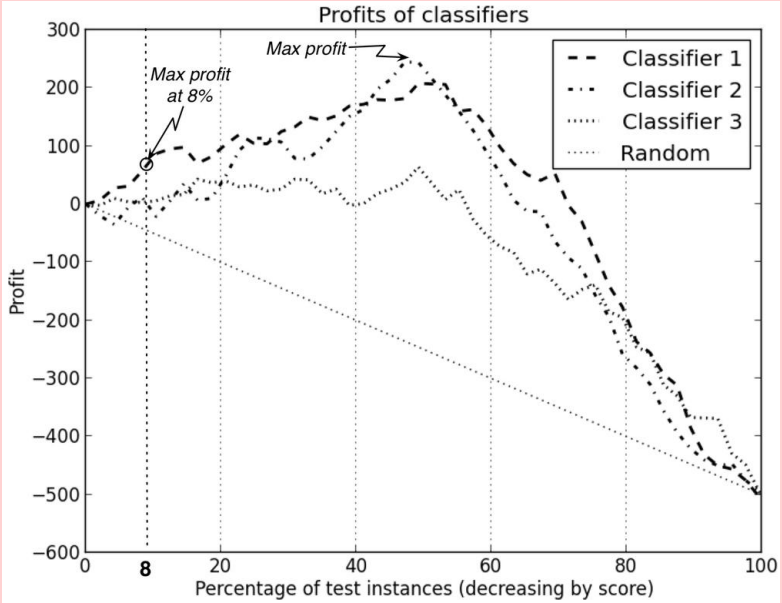
Model performance visualisation

- Ranking of items - this is something that can be achieved with most classification models, whether they produce probabilities (e.g. Naïve Bayes) or values between 0 and 1 that do not actually correspond to probabilities (logistic regression). The data instances are simply ranked from the 'most probable' to the 'least probable' based on the model output.



- Why rank instead of classify? Because the standard decision boundary of 0.5 would not make sense for many applications. E.g. the earlier described direct marketing example and the boundary that would make sense is **not known in advance**.
- Ranking allows the decision boundary to be determined depending on profit and budget

- Profit curves

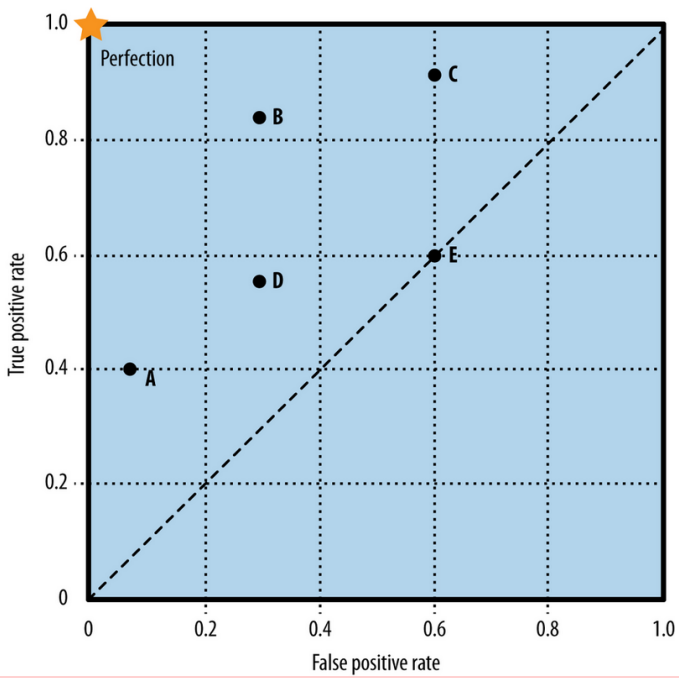


Profit curves provide information about the performance of classifiers and can help with determining the best binary decision boundary for a data set and given classification methods. A profit curve can be used when the **prior probabilities of the classes** and the **cost-benefit matrix** are stable across all data sets (those used to build the prediction model and those to be classified). Examples of questions that a profit curve would help answer are "Which one of three classifiers is the best one to use if the budget is limited (e.g. to target-marketing 8% of unknown subjects)?" [Classifier 1 in the picture] or it might be "Which one of the three classifiers, when used to rank the new subjects, would bring about the hightest profits with an unlimited budget?" [Classifier 2 in the picture]

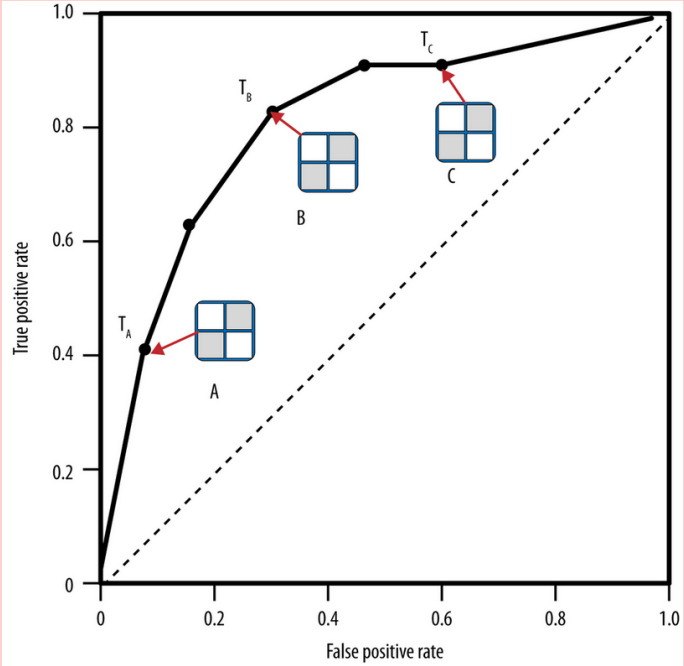
The x-axis of a profit curve represents the percentage of data instances that is found above the decision boundary as the boundary is moved, instance by instance, down the ranked (e.g. by probability to respond to targeted marketing) list of instances in a training set. The y-axis shows the profit that results from the boundary being placed at the different points in the list (e.g. targeted marketing is directed at subjects above the line). Initially the profit rises more steeply (as a working classifier would rank positives, bringing profit, at the top) and starts falling once many negatives start being included above the boundary (these are false positives, incurring cost). In the picture, classification model 1 is good at identifying the instances with the highest probabilities, while classification model 2 has better performance in identifying the most highly probable 50% of instances.

- Receiver Operating Characteristics (ROC) graphs & curves and area under the curve (AUC)

ROC has its origins in WW2, whence it was used as a measure of radar performance.



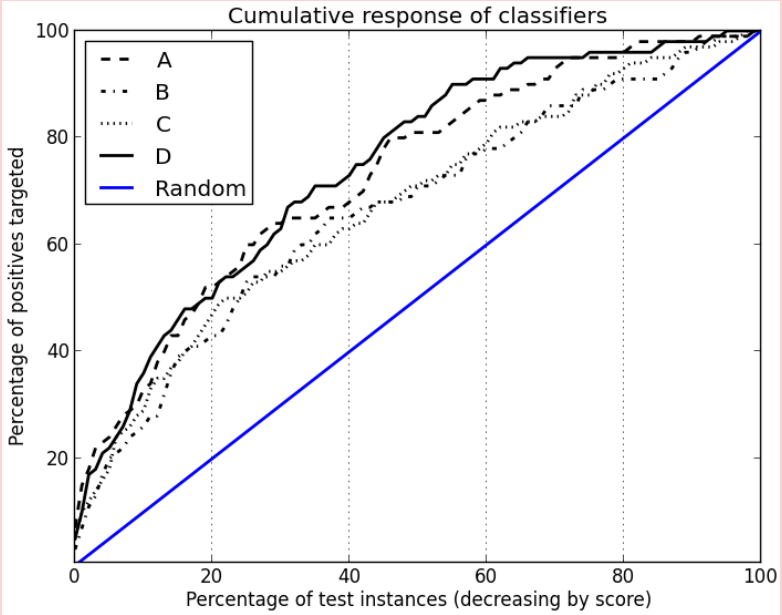
Each point in an ROC graph [left] corresponds to a contiguous portion, extending from the top, of a ranked instance list. The x-component of the point is the part (out of 1) of the data set's **true positives** contained in the list portion corresponding to the point. The y-component is the same part for **false positives**.

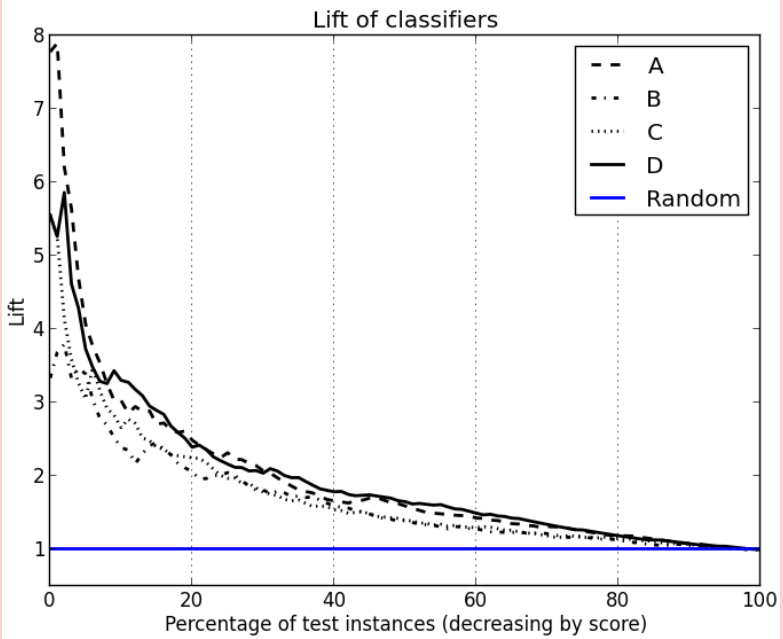


An ROC curve [right] is a systematic plot of points like those in the ROC graph together with a connecting line, for all possible contiguous portions, extending from the top, of the ranked instance list (starting with an empty list and concluding with the entire list).

ROC curves provide a measure of classifier performance that is **independent of class priors and of the cost-benefit matrix**. The closer the curve is to the two-straight-segment-line (0,0)-(0,1)-(1,1) the better the performance of the classifier, indicating more positives placed towards the top of the ranked list. The area under the ROC curve (AUC) provides a numeric measure of this quality: a value of 1 indicates a perfect classifier while a value of 0.5 correpsonds to random classification.

- Cumulative response and lift curves



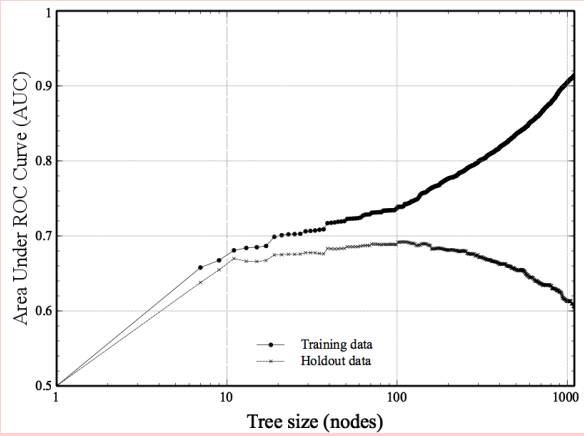


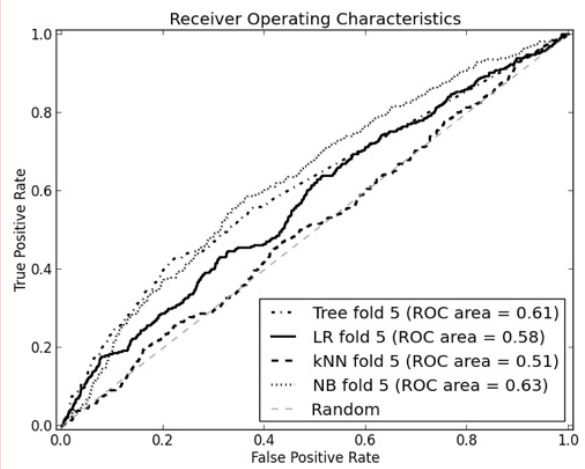
A **cumulative response curve** [left] is similar to an ROC curve but the x-axis, instead of representing the part of all false positives found in the cut-off portion of the ranked instance list, represents the percentage of the entire list of instances that is contained in the portion of the list above the cut-off point. This means that the shape of the cumulative response curve, while somewhat easier to understand for a non-technical audience, depends on the class priors and is thus not as pure a measure of a classifier's performance.

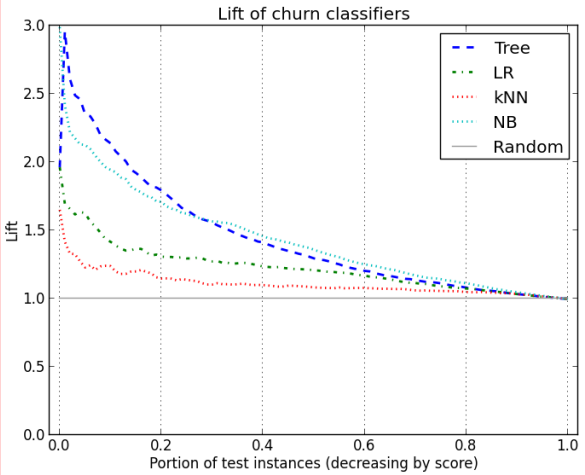
Every cumulative response curve (CRC) has a corresponding **lift curve** [right] which has the same values on the x-axis as the CRC and shows the ratios of the CRC y-values to the CRC y-values for random classification (this means that the lift curve for random classification is a constant 1).

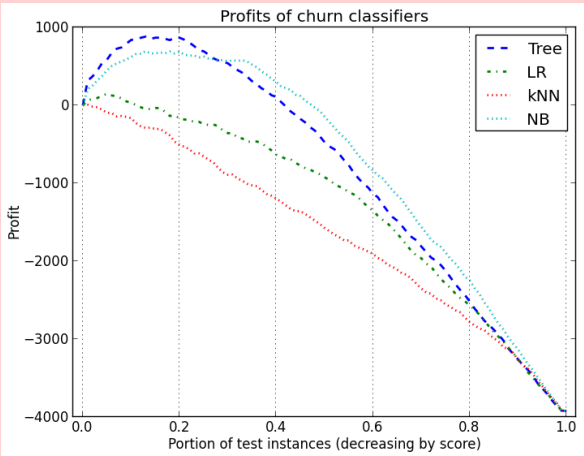
Example

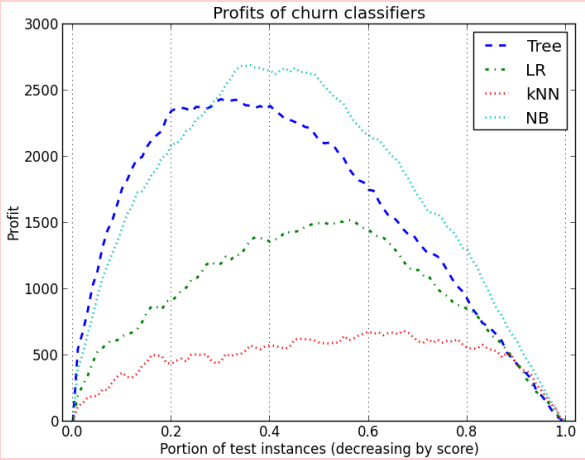
KDD Cup 2009 - French Telecom company Orange data











Model	Accuracy (%)	AUC
Classification Tree	91.8 ± 0.0	0.614 ± 0.014
Logistic Regression	93.0 ± 0.1	0.574 ± 0.023
k-Nearest Neighbor	93.0 ± 0.0	0.537 ± 0.015
Naive Bayes	76.5 ± 0.6	0.632 ± 0.019

Although kNN produces considerably better accuracy than the Naïve Bayes classifier, the confusion matrices and diagrams show that kNN acts practically as a base-rate classifier and is in fact not the better of the two models.

Naïve Bayes Confusion Matrix

	p	n
Y	127 (3%)	848 (18%)
N	200 (4%)	3518 (75%)

k-NN Confusion Matrix

	p	n
Y	3 (0%)	15 (0%)
N	324 (7%)	4351 (93%)

Evaluating Regression Models

Regression models are evaluated with some measure of how different the predicted values are from the actual values.

mean square error (MSE): $$\frac{\sum(predicted - actual)^2}{n}$$

root mean square error (RMSE): $$\sqrt{\frac{\sum(predicted - actual)^2}{n}}$$

mean absolute error (MAE): $$\frac{\sum|predicted - actual|}{n}$$

coefficient of determination (R^2): $$1 - \frac{SS_{res}}{SS_{tot}}$$

where $$SS_{res} = \sum(predicted - actual)^2 = MSE * n$$

$$SS_{tot} = \sum(actual - mean)^2 = variance * n$$

Baseline models

For the assessment of a model to make sense, we must compare it to some other reference model.

- Random model - easy to beat but good for initial exploration
- Meteorological prediction models have two important baselines: persistence and climatology
- Classification - majority class
- Regression - average or median, also averages across different data subsets
- Single feature prediction - simplest but often good as a first model (decision *stump* as opposed to tree)
- Single datasource - rather than many integrated data sources
- Baseline based on some experiential wisdom e.g. increased account activity in fraud detection

References

[DSB] *Data Science for Business: What you need to know about data mining and data-analytic thinking*, by Foster Provost and Tom Fawcett, O'Reilly Media, 2013.