Evaluating model performance

What is the desired result from data mining results?  
How would you measure that your model is any good?  
 How to measure performance in a meaningful way?

Model evaluation is application-specific.  
 We look at common issues and themes in evaluation

Frameworks and metrics for classification and instance scoring

**Bad positives and harmless negatives**

Classification terminology

Bad Outcome – something needs to be done  
Good outcome, no cause for concern

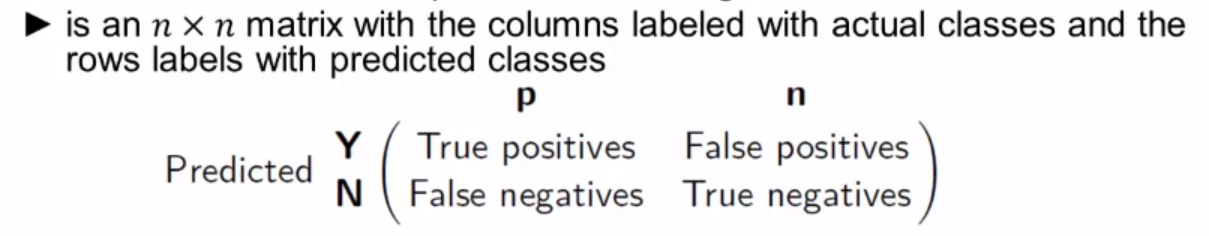
**Measuring Accuracy**

Most basic: dividing number of correct decisions / total number of decisions made

Sometimes too simplistic for applications of data mining to real business problems

**The Confusion Matrix**

An n x n matrix with the columns labeled with actual classes and the rows labels with predicted classes



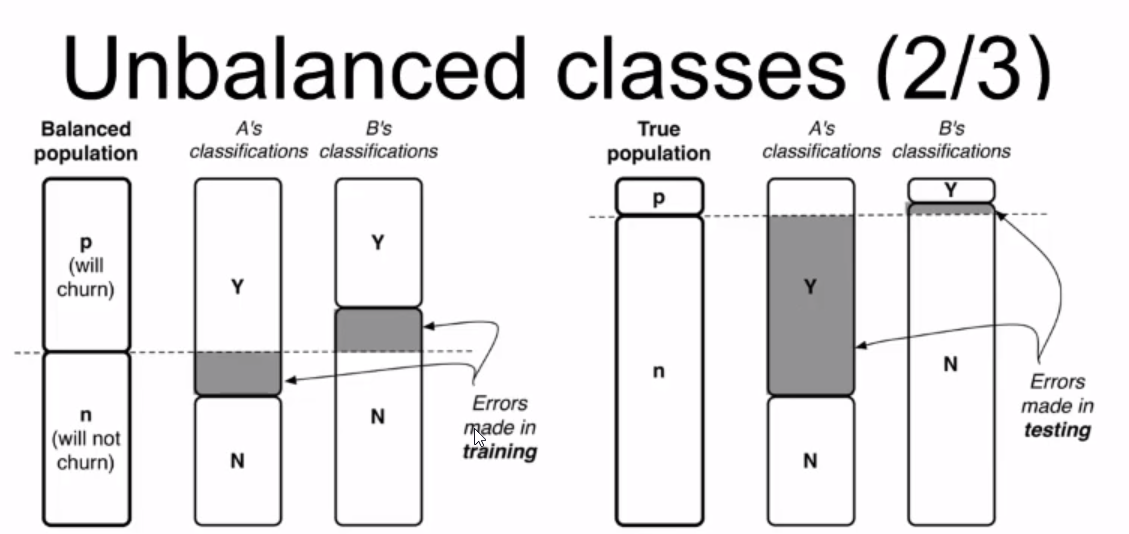
Actual class label and the class predicted by the classifier.

Confusion matrix separates the decisions made by the classifier:  
 Actual/True Classes: p(ositive), n(egative)  
 predicted class: Y9es), N(o)  
 The main diagonal contains the count of correct decisions

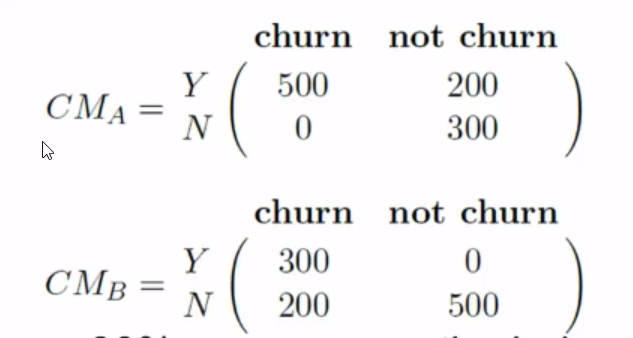
**Unbalanced Classes**

In practical classification problems, one class is often rare  
Classification is used to dins a relatively small number of unusual ones

Evaluation based on accuracy does not work:  
 ex – 999:1 ratio – always choose the most prevalent class – 99.9 accuracy  
 Fraud detection: skews of 10^2  
 IS a model with 80% accuracy always better than a model with 37% accuracy



Consider two models A and B for the churn example (1000 customers, 1:9 ratio of churning)  
 Both models correctly classify 80% of the balanced population.  
 Classifier A often falsely predicts that customers will churn  
 Classifier B makes many opposite errors



Model A achieves 80% accuracy on the balanced sample.  
Unbalanced population: A’s accuracy is 37%, B’s accuracy is 93%

**Unequal costs and benefits**

How much do we care about the different errors and correct decisions?  
Classification accuracy makes no distinction between false positive and false negative errors.  
In real-world applications, different kinds of errors lead to different consquences.

Examples for medical diagnosis:  
 A patient has cancer (although he does not)  
 False positive error, expensive, but not life threatening.  
 a patient has cancer, but she is told that she has not.  
 False negative error, more serious.

Errors should be counted separately, estimate cost or benefit of each decision.

Another example: how to measure the accuracy / quality of a regression model?  
 Predict how much a given customer will like a given movie

Typical accuracy of regression: ream-squared error

What does the mean-squared error describe?  
 Value of the target variable, e.g., the number of stars that a user would give a s a rating for the movie

Is the mean squared error a meaningful metric?

Compare the quality of different models with each other  
 Does the data-driven model perform better than a hand-crafted model  
 Does a classification tree work better than a linear discriminant model?  
 Does any of the models perform substantially better than a baseline model?

In aggregate: How well does each model do – what is its expected value?

**Cost and Benefits**

Compute cost-nenefit values for each decision pair  
A cost-benefit matrix specifies for each (predicted, actual) pair the cost or benefit making such a decision

Correct classifications correspond to b(Y,p) and b(N,n)  
Incorrect classifications (false positivies and negatives) corresponds to b(Y,n) and b(N,n), respectively [often negative benefits or costs]

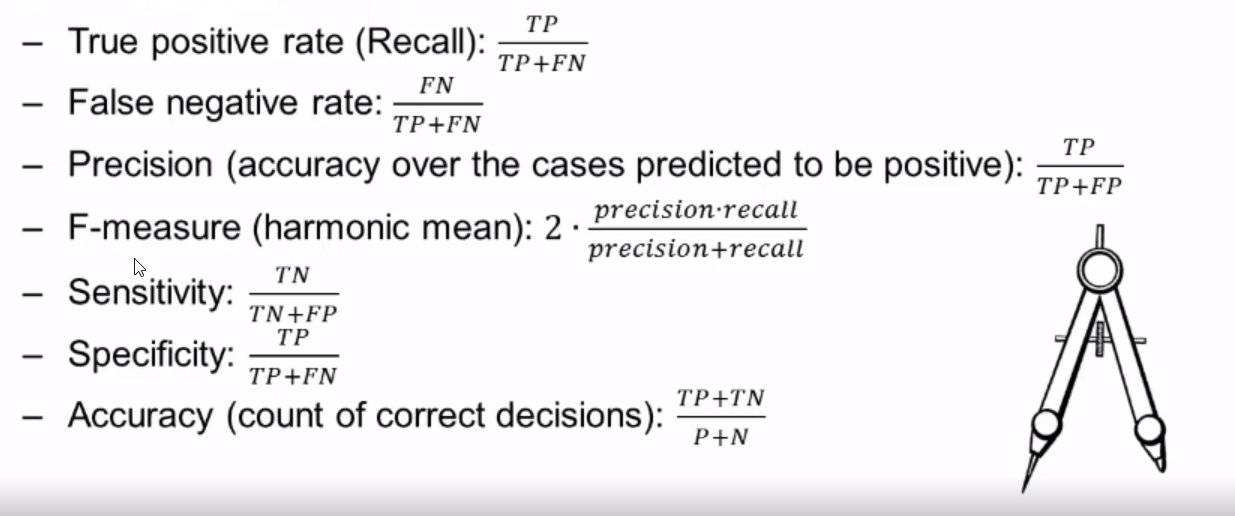
Costs and benefits cannot be estimate from data  
 How much is it really worth us to retain a customer?  
 Often use of average estimated costs and benefits

**Expected Value for use of a classifier**

Example  
 Price of product: $200, costs of product: $100  
 Targeting a consumer: $1, profit V\_r = $99, V\_nr = -$1  
 Do we make a profit? Is the expected value (profit) of targeting greater than zero?

**Other evaluation metrics**

Based on the entries of the confusion matrix, we can describe various evaluation metrics

****

**Baseline Performance**

Consider what would be a reasonable baseline against which to compare model performance.

Demonstrate stakeholder that data mining has added value or not

What is the appropriate baseline for comparison?  
Depends on the actual application

Nate silver in weather forecasting:  
 There are two basic tests that any weather forecast must pass to demonstrate its merit: (1) it must do better than what meteorologists call persistence: the assumption that the weather will be the same tomorrow and the next day as it was today. (2) It must also beat climatology, the long-term historical average of conditions on a particular date in a particular area.

Baseline performance for classification  
Compare to a completely random model  
implement a simple, but not simplistic alternative model

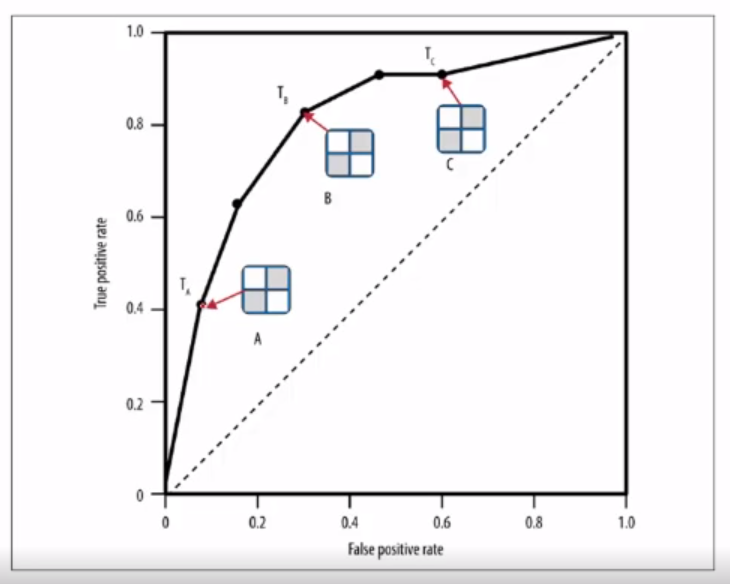
Majority Classifier = a naïve classifier that always chooses the majority class of the training data set  
 may be challenging to outperform: classification accuracy of 94%, but only 6% of the instances are positive. Majority classifier also would have accuracy of 94%

Pitfall: Don’t be surprised that many models simply predict everything to be of the majority class

Maximizing simple prediction accuracy is usually not an appropriate goal

**ROC Graph with Classifiers**

Receiver Operating Characteristics (ROC) graph is a two-dimensional plot of a classifier with flase positive rate on the x axis against true positive rate on the y axis. As such, a ROC graph depicts relative trade-offs that a classifier makes between benefits (true positives) and costs (false positives)



**Lift curves**

Essentially the same information as the cumulative response curve at a given x point divided by the diagonal line y=x value at that point.

ROC Charts can be difficult to read vs Lift Charts