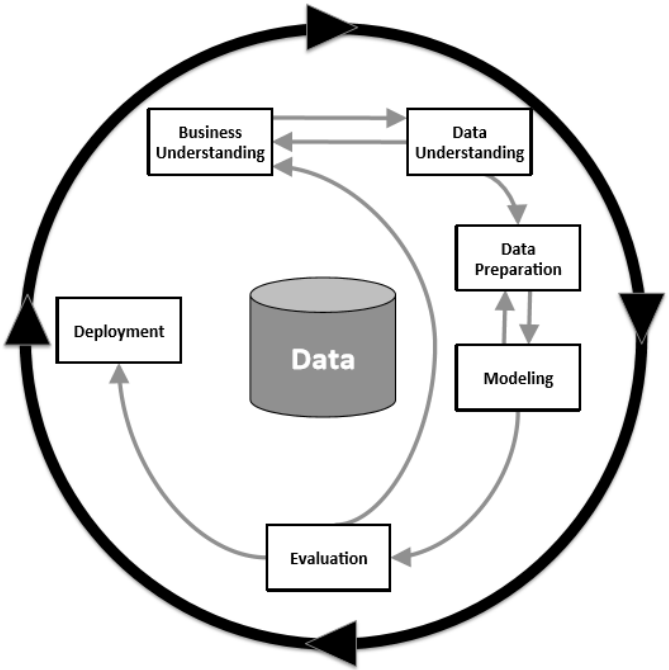
Advanced Data & Network Mining

Evaluation

2023-24

terri.hoare@dbs.ie

Model Evaluation
Crisp-DM Methodology



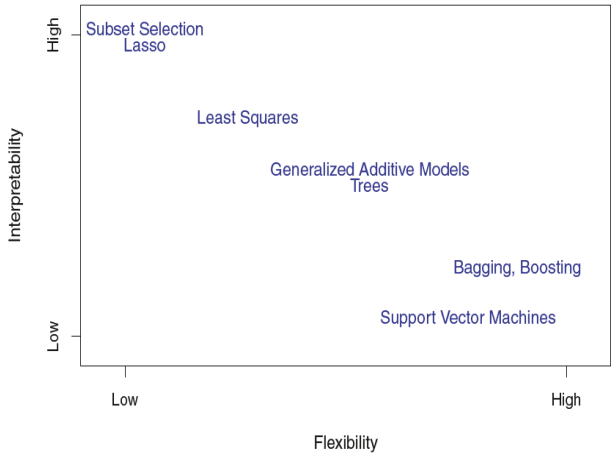
In building a model, the data preparation portion may be considered "pre-processing" while the evaluation portion may be considered "post-processing". Before deployment we need to ensure model validity by evaluating the model.

Model Evaluation
Modelling for Prediction and Inference

Models may be built for prediction (accuracy) or for inference (which factors are important). For prediction accuracy, we need to evaluate how well does the model built using training data “generalise” on unseen test data (real world data). Further, in classification, the predictions are ranked (confidence scores).

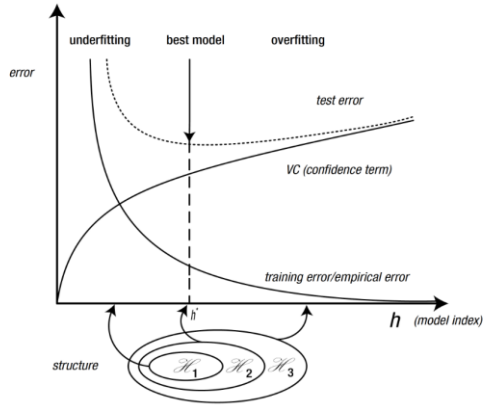
Model Evaluation
Modelling for Prediction and Inference

In statistical learning, the space of hypotheses C is a first prior assumption. There is no “free lunch”, every hypothesis or algorithm will work well on some dataset. Further “Every model is wrong, some are useful” – George Box. Model complexity is typically traded against accuracy, less complexity allowing more interpretability and more complexity achieving greater accuracy. Refer illustration below. (ISLR Figure 2.7).



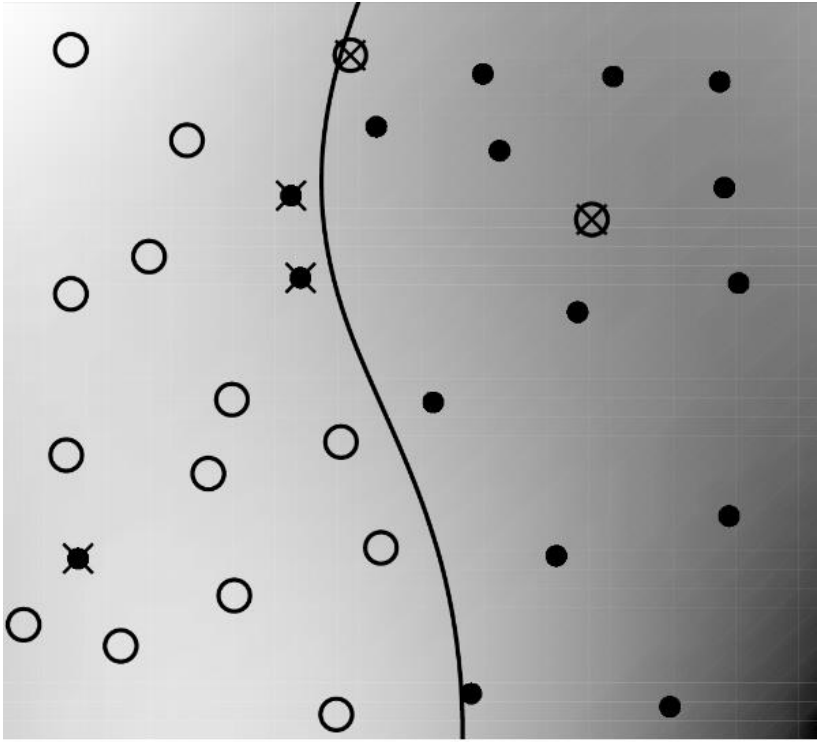
Model Evaluation
Minimise Error (Underfitting-Bias² and Overfitting-Variance)

Typically in supervised learning labelled data is split roughly 2/3 training and 1/3 testing. Assuming a true function f, the aim when learning is to estimate the "best" function $\hat{f}$ (regression) or g (classification). A good learning algorithm should give a solution that behaves similarly to the target function f and predicts or classifies well on new data. If we achieve this, we say that the algorithm **generalises**. Generalisation is a key requirement in Supervised Learning. There is a sweet spot between underfitting a model (high bias²) and overfitting a model (high variance – sensitivity to small changes in data) where the error is minimised.



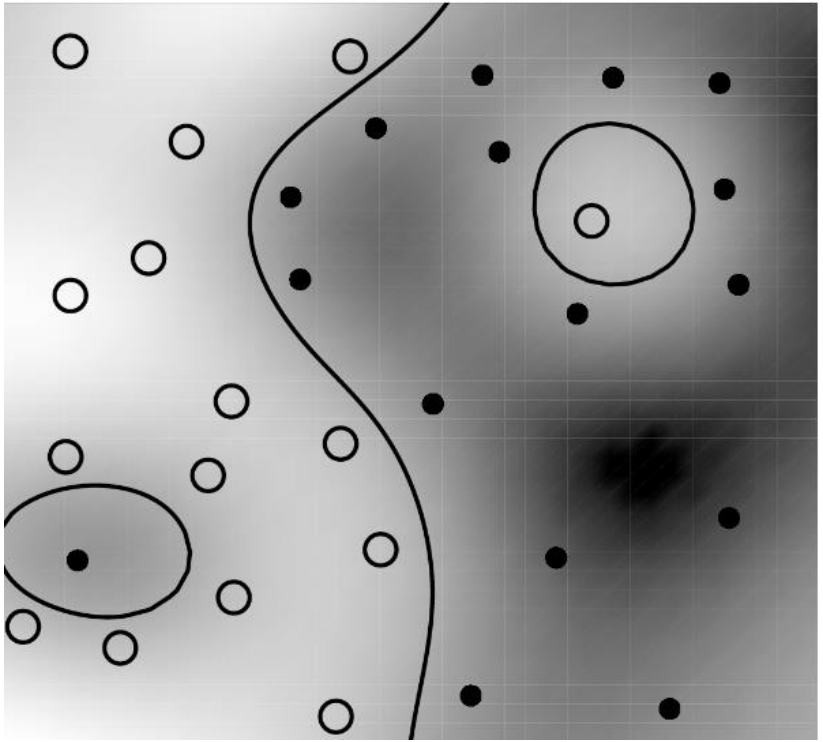
Model Evaluation
Erring towards the over simple – Underfitting (Classification)

A binary classification boundary which is over-simple (underfit) and fails to correctly classify many data items.



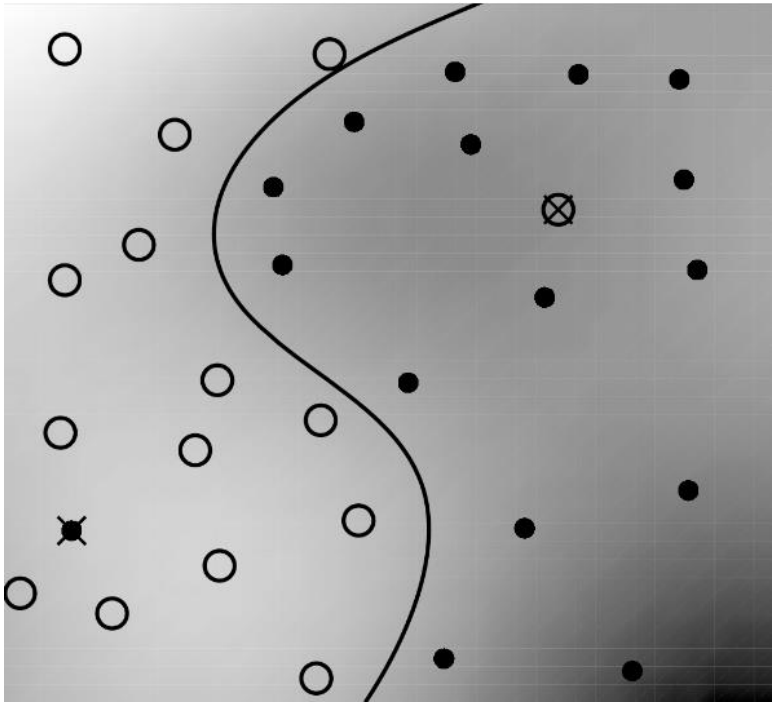
Model Evaluation
Erring towards the over complex – Overfitting (Classification)

A binary classification boundary which is overfitted to the training data and will fail to generalise well.



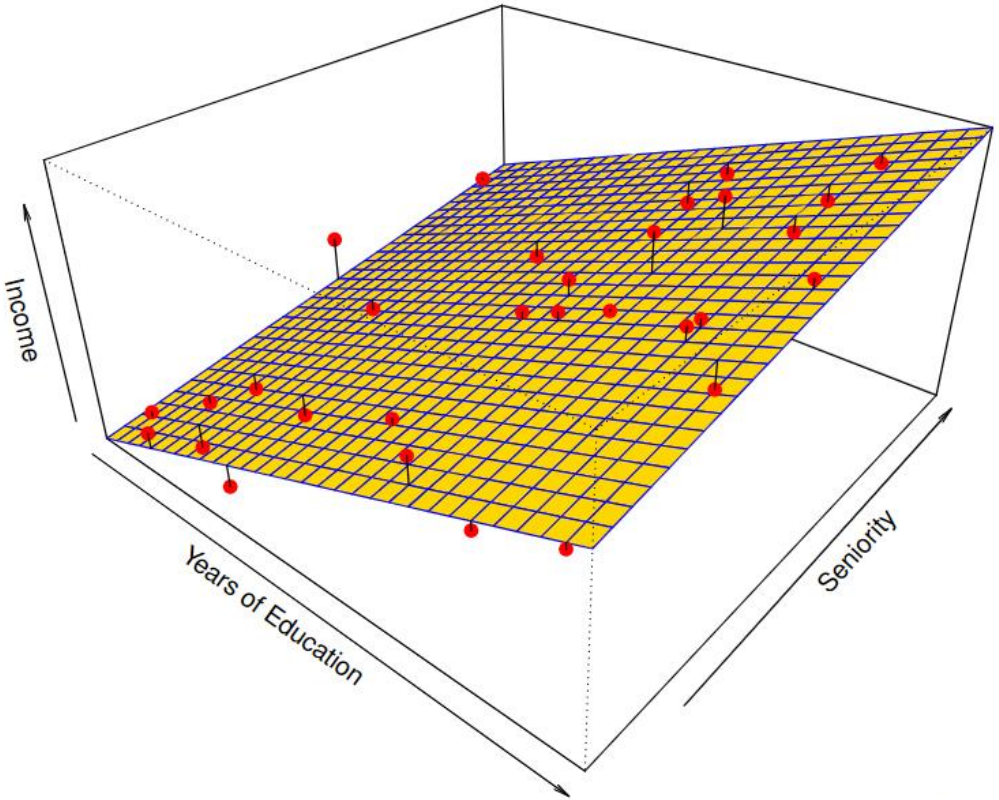
Model Evaluation
“Goldilocks” balance between simplicity and goodness of fit

A binary classification boundary which strikes a balance between simplicity and goodness of fit. It misclassifies a small number of data items but can generalise well to out-of-sample data.

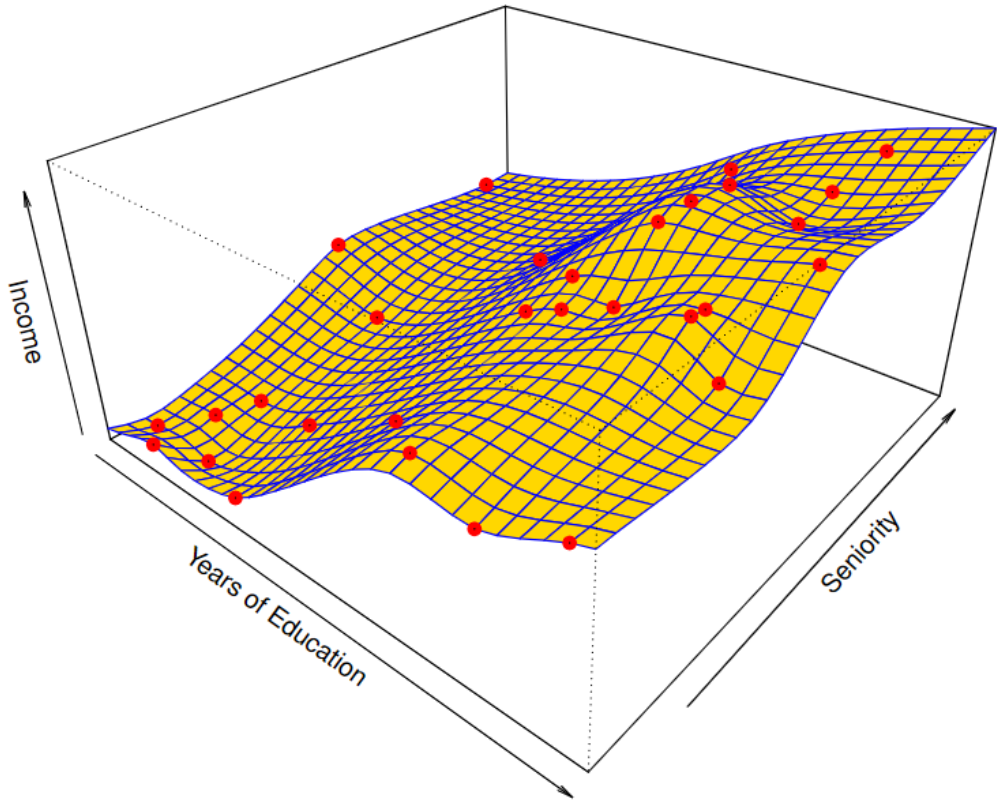


Model Evaluation

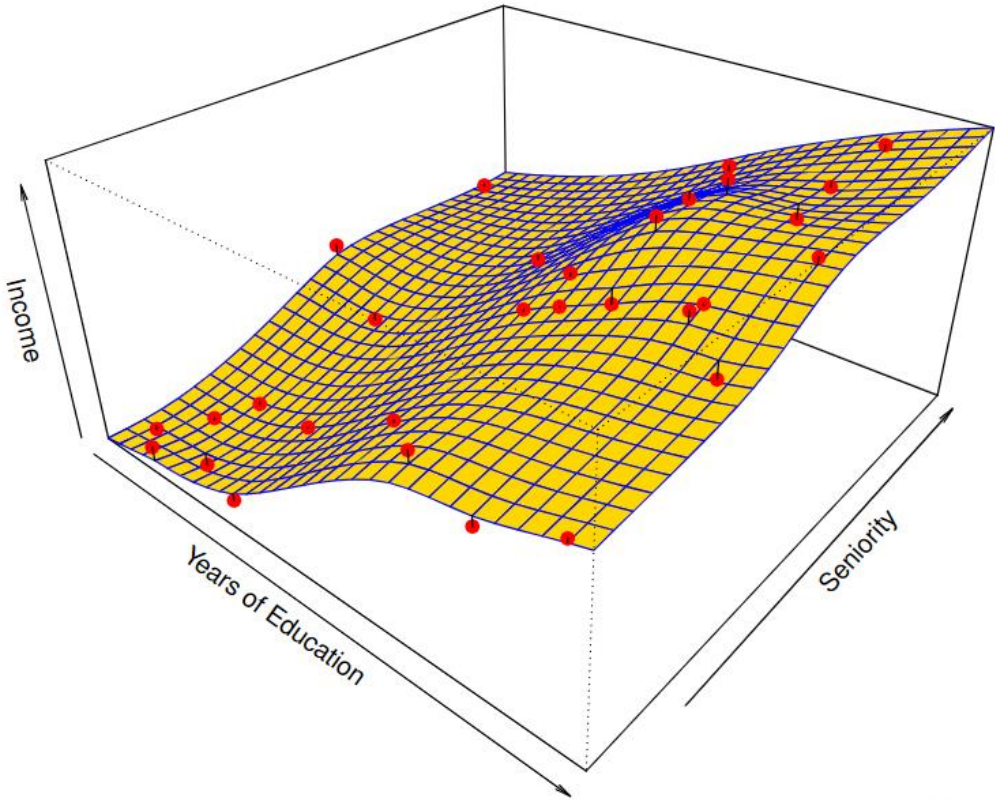
A very restrictive linear regression fit (high bias-underfitting)



Model Evaluation
A very flexible rough spline fit (high variance-overfitting)



Model Evaluation
A fairly flexible spline fit that can generalise well



Model Evaluation
Underfitting

Underfitting happens when a model is overly simple: too rigid a structure, too few features. This makes the model inflexible in learning from the dataset.

Simple models / learning machines tend to have lower variance but higher bias. A larger training set does not generally help with underfitting. Instead, we should increase the number of predictors/features as this makes the hypothesis space bigger and so allows for more possible functions.

Model Evaluation
Overfitting

Overfitting happens because the learning method is looking too hard for patterns in the training data and may be picking up some random patterns rather than true properties of the unknown function, overfitting “noise" in the training data that does not carry over to the test data.

Even a high-bias method like linear regression can overfit if the number of parameters (number of dimensions or number of predictors) is greater than or equal to the training set size. Analogously, in a high-dimensional situation, even though it is possible to perfectly fit the training data, the linear regression model we get will perform very poorly on unseen test data and so is not useful. The problem is that if the number of parameters is greater than or roughly equal to the training set size, a least squares regression is actually too flexible and so overfits the data.

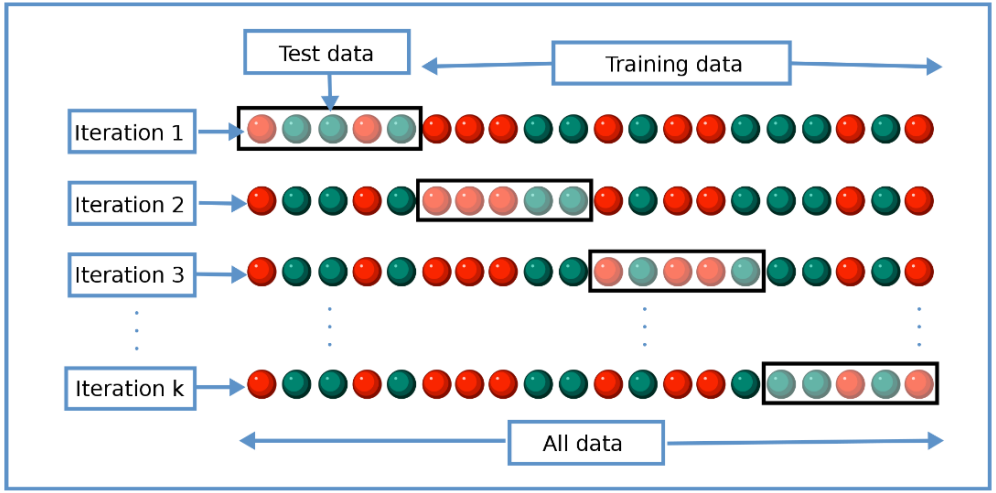
Model Evaluation
Addressing Overfitting

We need methods to address overfitting and enhance the robustness (against fitting noise) of the model.

- increase the training set size; Although a larger training set can help with overfitting, this is not always achievable.
- cross-validation
- early-stopping
- feature selection
- ensemble methods
- Regularisation (statistical learning)

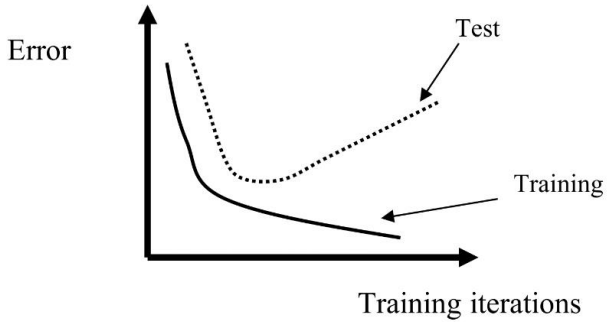
Model Evaluation
Addressing Overfitting – k-fold Cross Validation

Multiple rounds are carried out, typically 5 or 10 rounds. In each round, the sample is partitioned into two complementary subsets; the training is carried out on one subset (the training set), and it is validated on the other subset (the validation set). To insulate against variability resulting from division into training and validation sets, a different partition is used in each round. The validation results are averaged over the rounds to estimate the model's performance.



Model Evaluation
Addressing Overfitting – Early Stopping

For learners which are trained incrementally (e.g. , sequentially, in batches or in epochs), the method of early-stopping can address overfitting. In this method, the dataset is divided into three components: training data, validation data, and out-of-sample (test) data. The model is constructed using the training dataset; but, periodically during this process, the performance of the model is tested against the validation (test) dataset. The model performance on the validation dataset is used to determine when to stop the learning process and the best model is defined as that which produces the minimum error on the validation dataset. Often used during training of a neural network. Below the model starts to perform more poorly on out-of-(training) sample test data, which indicates possible overfitting



Model Evaluation
Addressing Overfitting – Feature Selection

Feature selection is the removal of predictors (features). For it to work, we assume that the data contains some predictors that are either redundant or irrelevant. Irrelevant means that the predictor has no substantial effect on the response. A relevant predictor can still be redundant if it is strongly correlated to one or more other relevant predictors. Just as adding features can address underfitting, so removing them (feature selection) can address overfitting.

As well as reducing overfitting (by reducing model complexity and variance), this can also have other uses.

- Simplifying the model, making it easier to interpret, aiding inference;
- reducing the curse of dimensionality by reducing the number of dimensions;
- reducing training time

Model Evaluation
Addressing Overfitting – Ensemble Methods

Ensemble methods work by combining predictions from multiple separate models. The two most common approaches are:

Bagging (Bootstrap AGGregatING) tries to reduce the chance of overfitting by training a large number of flexible (decision tree / neural network) models independently of each other. It then averages out the predictions of these models, on the principle that averaging a set of observations reduces variance.

Boosting tries to enhance the flexibility of simple models. Each learner is a “weak" or “slow" learner: it is a constrained model which does not learn quickly; in fact, it may be only slightly better than random guessing. Boosting trains a large number of these “slow" learners sequentially. Each learner is built using information from the one before it. Boosting then combines all the weak learners into a single strong learner, usually weighting them according to their strength.

Model Evaluation
Classification Models : Evaluation Tools

In order to build a **Classification** Model, we need labelled data. The data is split into a training (and validation) set and a testing set (ratio for example 70:30, 80:20). The training set is used to build the model and the testing set is used to evaluate the performance of the model

There are three main tools to test a **classification** model's quality (performance):-

- Confusion matrices (or truth tables)
- Lift charts
- ROC (receiver operator characteristic) curves

These measures are used to fine-tune models and/or select the algorithm(s) that best fit the problem

Model Evaluation
Classification Models : An Example : Direct Marketing

Direct Marketing companies were one of the early pioneers in adopting predictive analytics techniques.

If typical response for a direct mail campaign is 10% and considering the cost of mailing, it would make sense to only send to this 10%.

We can use classification models to rank or score prospects by their likelihood to respond to the mailers. Predictive analytics is about converting future uncertainties into usable probabilities. We can order these probabilities and send out mailers to only those who score above a particular threshold (say 85% chance of response)

How do we compare different methods by their performance? What are the metrics we can use to select the best performing methods?

Model Evaluation
Classification Models : Confusion Matrices

Confusion Matrix			
		Actual Class(Observation)	
		Y	**N**
Predicted Class (Expectation)	Y	TP (true positive) Correct result	FP (false positive) Unexpected result
	N	FN (false negative) Missing result	TN (true negative) Correct absence of result

True Positives (TP) – the predicted class is Y AND the actual class is Y

False Positives (FP) – the predicted class is Y BUT the actual class is N

False Negatives (FN) – the predicted class is N BUT the actual class is Y

True Negatives (TN) – the predicted class is N AND the actual class is N

A perfect classification would only have entries along main diagonal with off diagonal elements zero

Model Evaluation
Classification Models : Confusion Matrices cont.

		Actual Class(Observation)	
		Y	N
Predicted Class (Expectation)	Y	TP (true positive) Correct result	FP (false positive) Unexpected result
	N	FN (false negative) Missing result	TN (true negative) Correct absence of result

	true no response	true response	class precision
pred. no response	1231	146	89.40%
pred. response	394	629	61.49%
class recall	75.75%	81.16%	

Confusion matrix for validation set of direct marketing data set.

Evaluation Measures

Term	Definition	Calculation
Sensitivity	Ability to select what needs to be selected	TP/(TP+FN)
Specificity	Ability to reject what needs to be rejected	TN/(TN+FP)
Precision	Proportion of cases found that were relevant	TP/(TP+FP)
Recall	Proportion of all relevant cases that were found	TP/(TP+FN)
Accuracy	Aggregate measure of classifier performance	(TP+TN)/(TP+TN+FP+FN)

Model Evaluation
Classification : Receiver Operator Characteristic Curves

Sensitivity and specificity are statistical measures of the performance of a binary classification test. Sensitivity (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition). Specificity (also called the true negative rate) measures the proportion of negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

Model Evaluation
Classification : Receiver Operator Characteristic Curves

Thus, sensitivity quantifies the avoiding of false negatives, as specificity does for false positives. For any test, there is usually a trade-off between the measures. For instance, in an airport security setting in which one is testing for potential threats to safety, scanners may be set to trigger on low-risk items like belt buckles and keys (low specificity), in order to reduce the risk of missing objects that do pose a threat to the aircraft and those aboard (high sensitivity). This trade-off can be represented graphically as a receiver operating characteristic curve. A perfect predictor would be described as 100% sensitive (e.g., all sick are identified as sick) and 100% specific (e.g., no healthy are identified as sick); however, theoretically any predictor will possess a minimum error bound.

Model Evaluation
Classification : ROC : Worked Example (Wikipedia)

A diagnostic test with sensitivity 67% and specificity 91% is applied to 2030 people to look for a disorder with a population prevalence of 1.48%. (Next slide).

Related calculations

- False positive rate (α) = Type I Error = 1 − specificity = FP / (FP + TN) = 180 / (180 + 1820) = 9%
- False negative rate (β) = Type II Error = 1 − sensitivity = FN / (TP + FN) = 10 / (20 + 10) = 33%
- Power = sensitivity = 1 − β
- Likelihood Ratio positive = sensitivity / (1 − specificity) = 0.67 / (1 − 0.91) = 7.4
- Likelihood ratio negative = (1 − sensitivity) / specificity = (1 − 0.67) / 0.91 = 0.37

Hence with large numbers of FP and few FN, a positive screen test is in itself poor at confirming the disorder (PPV = 10%) and further investigations must be undertaken; it did, however, correctly identify 66.7% of all cases (the sensitivity). However as a screening test, a negative result is very good at reassuring that a patient does not have the disorder (NPV = 99.5%) and at this initial screen correctly identifies 91% of those who do not have cancer (the specificity).

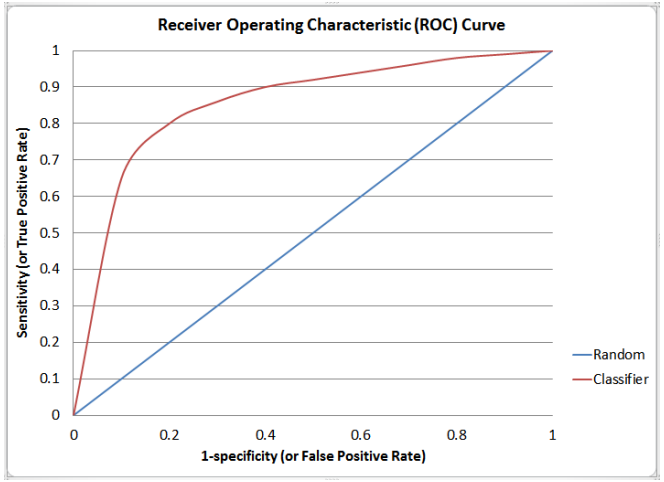
Model Evaluation
Classification : ROC : Worked Example (Wikipedia)

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition positive	Condition negative	
Fecal occult blood screen test outcome	Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Model Evaluation
Classification : Receiver Operator Characteristic Curves

ROC graphs have long been used in signal detection theory to depict the **trade off** between **hit rates** and **false alarm rates** of classifiers. **Hit rates** = sensitivity

False alarm rates = 1-specificity. A **ROC chart** is constructed with 1-Specificity (False Positives rate) on the X-axis and Sensitivity (True Positives rate) on the Y-axis. Thus a ROC curve simply helps one quantify how many true positives are detected by the algorithm for every false positive.



Model Evaluation
Classification : ROC Curves cont.

Consider a classifier that predicts whether a website visitor is likely to click on a banner ad. The model would most likely be built using historic click-through rates based on pages visited, time spent on certain pages, and other characteristics of site visitors.

In order to evaluate the performance of the model, we generate a table sorting the predicted data in decreasing order of confidence. We keep a running count of the TP's and FP's and also calculate the fraction of TP's and FP's.

Model Evaluation
Classification : ROC Curves cont.

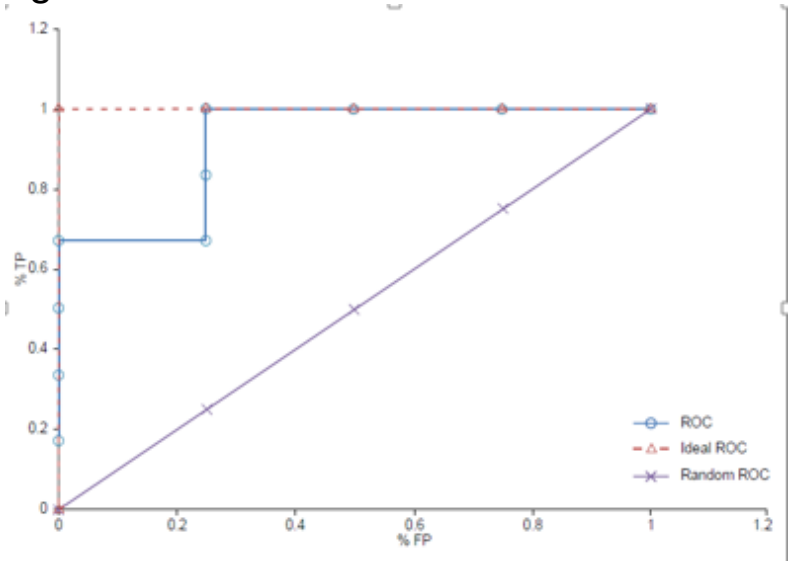
Note that the model had identified nearly all the TP's (4/6 = 67%) before it hits the first FP and all TP's have been identified before it hits the next FP

Classifier Performance Data Needed for Building an ROC Curve

Actual Class	Predicted Class	Confidence of "response"	Type?	Number of TP	Number of FP	Fraction of FP	Fraction of TP
response	response	0.902	TP	1	0	0	0.167
response	response	0.896	TP	2	0	0	0.333
response	response	0.834	TP	3	0	0	0.500
response	response	0.741	TP	4	0	0	0.667
no response	response	0.686	FP	4	1	0.25	0.667
response	response	0.616	TP	5	1	0.25	0.833
response	response	0.609	TP	6	1	0.25	1
no response	response	0.576	FP	6	2	0.5	1
no response	response	0.542	FP	6	3	0.75	1
no response	response	0.530	FP	6	4	1	1
no response	no response	0.440	TN	6	4	1	1
no response	no response	0.428	TN	6	4	1	1
no response	no response	0.393	TN	6	4	1	1
no response	no response	0.313	TN	6	4	1	1
no response	no response	0.298	TN	6	4	1	1
no response	no response	0.260	TN	6	4	1	1
no response	no response	0.248	TN	6	4	1	1
no response	no response	0.247	TN	6	4	1	1
no response	no response	0.241	TN	6	4	1	1
no response	no response	0.116	TN	6	4	1	1

Model Evaluation
Classification : ROC Curves cont.

Plotting the ROC curve %FP (X axis) vs %TP (Y axis). Note that the 45 degree line would be a ROC for a random classifier (with no better than a coin toss chance of getting it right). The Area Under the Curve (AUC) in this case 0.5. A perfect Classifier would be the solid line Y=1 with AUC=1. The best classifier will have a ROC curve closest to perfect with an AUC ideally higher than 0.8.



Model Evaluation
Classification : Lift Curves

Often the measure of overall effectiveness of the model is not enough. It may be important to know if the model does increasingly better with more data. Is there any marginal improvement in the model's predictive ability if for example, we consider 70% of the data versus only 50%?

Gain (and Lift) charts were developed to answer this question. The focus is on the true positives and thus it can be argued that they indicate the sensitivity of the model. These types of charts are common in Direct Marketing where the problem is to identify if a particular prospect was worth calling.

Basis for building Gains charts

Randomly selecting x% of data (prospects) would yield x% of targets (to call or not). Gain is the improvement over this random selection that a predictive model can potentially yield.

Model Evaluation
Classification : Lift Curves cont.

Scoring a list of prospects by their propensity to respond to an ad campaign. When we sort (decreasing propensity to respond) the prospects by this score, we obtain a mechanism to systematically select the most valuable prospects right at the beginning and thus maximise our return. Thus rather than mailing out to a random group of prospects, we can send our ads to the first batch of “most likely responders”, followed by the next batch and so on.

Without classification, the “most likely responders” are distributed randomly throughout the data set. Suppose we have a data set of 200 prospects and it contains a total of 40 responders or TP’s. If we break the dataset up into deciles (10 batches), the likelihood of finding TP’s in each batch is also 20% that is 4 samples in each decile will be TP’s.

Model Evaluation
Classification : Lift Curves cont.

However, when we use a predictive model to classify the prospects, a good model will tend to pull these “most likely responders” into the top few deciles. Thus, we find that the first two deciles will have all 40 TP’s and the remaining eight deciles will have none.

Model Evaluation
Classification : Lift Curves : Response Ad campaign

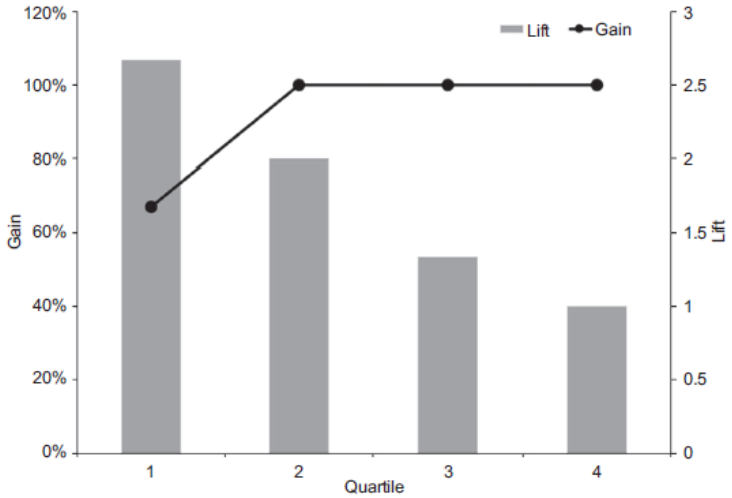
In the data set of 20 there are 6 TPS's. Using quartiles rather than deciles, we can randomly expect there to be 25% of 6 or 1.5 in each quartile. Sort the scored data set by confidence of responses. Cumulative TP then provides a way of counting the number of TPs in the first 25%, then 50%, and so on. ***Gain*** is the cumulative number TPs in that quartile to the total number in the data set (1st quartile 4/6 = 67%). ***Lift*** is the ratio of gain to the random expectation at a given quartile level (1st quartile 4/1.5 = 2.667)

Scoring Predictions and Sorting by Confidences is the Basis for Generating Lift Curves

Actual Class	Predicted Class	Confidence of "response"	Type?	Cumulative TP	Cumulative FP	Quartile	Gain	Lift
response	response	0.902	TP	1	0	1st	67%	2.666667
response	response	0.896	TP	2	0	1st		
response	response	0.834	TP	3	0	1st		
response	response	0.741	TP	4	0	1st		
no response	response	0.686	FP	4	1	1st		
response	response	0.616	TP	5	1	2nd	100%	2
response	response	0.609	TP	6	1	2nd		
no response	response	0.576	FP	6	2	2nd		
no response	response	0.542	FP	6	3	2nd		
no response	response	0.530	FP	6	4	2nd		
no response	no response	0.440	TN	6	4	3rd	100%	1.333333
no response	no response	0.428	TN	6	4	3rd		
no response	no response	0.393	TN	6	4	3rd		
no response	no response	0.313	TN	6	4	3rd		
no response	no response	0.298	TN	6	4	3rd		
no response	no response	0.260	TN	6	4	4th	100%	1
no response	no response	0.248	TN	6	4	4th		
no response	no response	0.247	TN	6	4	4th		
no response	no response	0.241	TN	6	4	4th		
no response	no response	0.116	TN	6	4	4th		

Model Evaluation
Classification : Lift Curves : Response Ad campaign cont.

Plotting the Gain and Lift across four quartiles. Note typically deciles are used. ***Gain*** is the cumulative number TPs in that quartile to the total number in the data set (1st quartile 4/6 = 67%). ***Lift*** is the ratio of gain to the random expectation at a given quartile level (1st quartile 4/1.5 = 2.667)



Lift and gain curves.

Model Evaluation
Classification Models : Discussion Points

- Relying on a single measure like accuracy may be misleading for example in the case of credit card fraud (1% transactions), the model may be 99.9% accurate and still not identify the critical fraudulent case. For highly unbalanced data sets we rely on several measures such as class recall and specificity in addition to accuracy

- ROC curves are frequently used to compare several algorithms side-by-side

- Both AUC and ROC should be used to rate a model's performance

- Lift and Gain charts are commonly used for scoring applications where we need to rank-order the examples in a data set by their propensity to belong to a particular category (response to a mailer)

Model Evaluation
Regression Models Summary

Methods for evaluating Regression models **(Performance Clustering):** -

- **R^2 (squared correlation)**

*V*alues [0,1] with values closer to 1 indicating a better model

- **Squared error output**

Further differentiates models, confidence intervals of predictions

- **ANOVA F**

At least one of the coefficients is significant (non-zero)

- **t-stat and p-values**

Results of hypothesis tests on the coefficients.

A higher **t-stat** signals the **NULL Hypothesis** *(*assumes coefficient is zero), can safely be rejected. The corresponding **p-value** indicates the probability of wrongly rejecting the null hypothesis

Model Evaluation
Clustering Models Summary

Two methods for evaluating models *(Performance Clustering):* -

- **SSE**

SSE is the average within cluster distance and can be calculated for each cluster and averaged across clusters. Good models will have low SSE both within and across clusters

- **Davies-Bouldin Index**

Function of the ratio of within cluster separation to the separation between the clusters. The lower the index, the better the cluster

(Note however, both SSE and Davies Bouldin Index have the limitation of not guaranteeing better clustering when they have lower scores)

Model Evaluation
Association Rule Models Summary

By Inspection. Two methods for selecting relevant rules : -

Association Rule :- {**antecedent}** → **{consequent}**

- The **support of a rule** is a measure of how all the items in a rule are represented in overall transactions. Worth considering.

- The **confidence of a rule** measures the likelihood of occurrence of the consequent of the rule out of all the transactions that contain the antecedent of the rule.

References

- Berry, M. A. (1999). Mastering data mining: The art and science of customer relationship management. New York: John Wiley and Sons.
- Black, K. (2008). Business statistics for contemporary decision making. New York: John Wiley and Sons.
- Green, D. S. (1966). Signal detection theory and psychophysics. New York: John Wiley and Sons.
- Kohavi, R., & Provost, F. (1998). Glossary of terms. Machine Learning, 30, 271274. Rud, O. (2000). Data mining cookbook: Modeling data for marketing, risk and customer relationship management. New York: John Wiley and Sons.
- Taylor, J. (2011). Decision management systems: A practical guide to using business rules and predictive analytics. Boston, Massachusetts: IBM Press.