PyData : Conference Mission Statement

- PyData is a gathering of users and developers of data analysis tools in Python.
- The goals are to provide Python enthusiasts a place to share ideas and learn from each other about how best to apply our language and tools to ever-evolving challenges in the vast realm of data management, processing, analytics, and visualization.

PyData : Conference Mission Statement

- We aim to be an accessible, community-driven conference, with tutorials for novices, advanced topical workshops for practitioners, and opportunities for package developers and users to meet in person.
- A major goal of the conference is to provide a venue for users across all the various domains of data analysis to share their experiences and their techniques, as well as highlight the triumphs and potential pitfalls of using Python for certain kinds of problems.

- A predictive model is not magic and it is not rocket science. It is actually a mathematical representation of reality that offers a view based on science and statistics that uses data that is currently available to predict the risk of an adverse event (or a preferred event) at some point in the future.
- The models, and there are many, look for patterns in the data that may not always be evident and which seem linked to an outcome of some sort.

- Depending on the data and the model it can do so in either a supervised learning fashion, where the model describes a relationship between a set of independent attributes and a dependent attribute or in an unsupervised learning fashion where the model, itself, will find relationships in the data without reference to independent or dependent variables.

1. Defining the Problem
2. Processing The Data
3. Run an initial model
4. Evaluate the initial model
5. Select a final model
6. Testing the Model
7. Use the Model

Step 1 : Defining the Problem

What question(s) are you trying to answer? Once you understand that you need to then think about what data is available to you to answer the question:

- Is the data directly related to the question?
- If it is not, can you create a proxy relationship to be able to link it?

Step 1 : Defining the Problem

As part of step 1, you also need to specify the inputs and outputs of the model you are going to build as these may change as you change and tweak the model. Finally, dont forget the most commonly forgotten piece of any new initiative. Determine, up front, how you are going to measure the results.

Step 1 : Defining the Problem

- What measure of accuracy are you going to use?
- Is that level of accuracy good enough for the business?
- How will you benchmark the results?
- What criteria are you going to use to determine success or failure?

Step 2 : Processing the Data

- Step 2 is more rote and technical process the data.
- Collect the data (more is always better in this analysts mind but more does not always mean easier or better results).
- In general, more recent data is better and the data need to be consistent.
- Dont skimp on cleaning the data. While this may end up taking the most time, it is critical and erroneous data will create erroneous results.

Step 2 : Processing the Data

Transforming the Data

- Transforming the data is also worth the time and effort to improve the modeling process including such things as:
  - Converting non-numerical data to numeric (or vice versa)
  - Standardizing thing such as coding, definitions, costs, combining variables, etc.

Step 3 : Run the Initial Model

- Step 3 is running the initial model. Part of step 3 is to split the dataset into a test dataset and a validation dataset. If you really want to be able to test the accuracy of the model once it has been built, you need to do this.
- The software will walk you through this and will do this for you by default, holding back 30% of the total data for validation testing, although it does allow you to override the default values.
- This is also the step whereby you will choose the method or methods by which you want to build the model and process the data.

Step 3 : Run the Initial Model

- As you become more familiar with predictive modeling and with your own data you will find that certain types of data and certain types of analyses or problems do lend themselves more or less to certain types of modeling.
- But if you are just starting out, you can use the software to guide you in the choice of a model or simply choose to run all the models against your data.
- Once done, run the model and move on to step 4.

Step 4 : Evaluate the Initial Model

- Step 4 is to evaluate the initial results.
- *Are the results acceptable? Are they what you were expecting to see? Do you understand the results?*
- Most Importantly - Do they answer the question you are trying to answer?
- If the answer is yes, then move on to the next step.

Step 4 : Evaluate the Initial Model

If the answer is no, then consider the following:

- Try using different algorithms/models
- Try using different data elements or repackaging what you have
- Consider collecting more or different data
- Consider redefining the problem, changing the question and the means to an answer as you better understand your data and your environment

Step 4 : Evaluate the Initial Model

- Part of the learning process may be to try and not boil the ocean with your models.
- Think about setting up the model to run on a number of scenarios starting from the simple and most straightforward and then progressing to more and more complex.

Step 5

Step 5 is to select the final model.

- Dont be afraid to try a number of different models and then when you are satisfied with the results, choose the best one.
- We will talk about means of assessing the accuracy of your model in a bit.
- For now, choose the final model and consider whether you want to rerun the entire dataset against the selected model and re-examine the results

Step 6 : Testing the Model

- Step 6 is to test the final model. This is another one of those things that often seems not to get done.
- It is important to test the final model and the only way to do so is to take the selected model and run it against a second, unrelated dataset (e.g. - the validation dataset or the portion of the dataset that was held back for this purpose) and assess the results.

Step 6

- Do not tweak or change the model in any way at this point as it will invalidate any comparison to the initial results.
- If the results are similar and you are satisfied with them you can move on to the final step.
- If you are not, then go back (to step 3) to reassessing the model and the data, make any necessary or desired changes and try re-running the model again.

Step 7 : Use the Model

- Step 7 is to apply the model and run the prediction.
- There are actually two parts to this.
- One is done if you want to refine the model then you can use the output from the model to determine next steps and potential intervention or changes.
- Continue to test the model as much or as often as needed.

Step 7 : Use the Model

When you are satisfied, do two things:

1. Run the necessary measures to test the final accuracy of the model (see our next discussion points)
2. Take the output from the model and turn it back into language and output for the business that answers the initial question you set out to answer and makes it useful/usable for them

SUPERVISED LEARNING

Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems.

Supervised learning is when the data you feed your algorithm is "tagged" to help your logic make decisions.

Example: Bayes spam filtering, where you have to flag an item as spam to refine the results.

UNSUPERVISED LEARNING

In other pattern recognition problems, the training data consists of a set of input vectors x without any corresponding target values. The goal in such unsupervised learning problems may be to discover groups of similar examples within the data, where it is called clustering

Pattern Recognition and Machine Learning (Bishop, 2006)

Unsupervised learning are types of algorithms that try to find correlations without any external inputs other than the raw data. Example: datamining clustering algorithms.

Types of Predictive Analytics Problems

- What is Classification and Regression
- Binary Classification Problems
- Logistic Regression

Confusion Matrix

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Accuracy

Accuracy

Accuracy measures a fraction of the classifier's predictions that are correct.

$$\text{Accuracy} = \frac{TN + TP}{TN + FN + TP + FP}$$

Accuracy

	Predict P	Predict N
True P	100	70
True N	30	9800

$$\text{Accuracy} = \frac{9900}{10,000} = 0.99$$

Accuracy

- However, accuracy is not an informative metric if the proportions of the classes are skewed in the population. (**Class Imbalance**)
- For example, a classifier that predicts whether or not credit card transactions are fraudulent may be more sensitive to false negatives than to false positives.
- To promote customer satisfaction, the credit card company may prefer to risk verifying legitimate transactions than risk ignoring a fraudulent transaction.

Accuracy

- Because most transactions are legitimate, accuracy is not an appropriate metric for this problem.
- A classifier that always predicts that transactions are legitimate could have a high accuracy score, but would not be useful.
- For these reasons, classifiers are often evaluated using two additional measures called **precision** and **recall**.

Precision and Recall

- **Precision** (or positive predictive value (PPV))

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

- **Recall** (or Sensitivity , true positive rate (TPR))

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

- **Specificity** (or true negative rate (TNR))

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

Some Related Metrics

- Negative predictive value (NPV)

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

- False positive rate (FPR) also called "fallout"

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

- False negative rate (FNR)

$$FNR = FN/(FN + TP) = 1 - TPR$$

- False discovery rate (FDR)

$$FDR = FP/(FP + TP) = 1 - PPV$$

Precision and Recall

- Precision is the fraction of positive predictions that are correct.
- For instance, in our SMS spam classifier, precision is the fraction of messages classified as spam that are actually spam.
- Precision is given by the following ratio:

$$P = \frac{TP}{TP + FP}$$

The Confusion Matrix

- The confusion matrix indicates that there were four true negative predictions, three true positive predictions, two false negative predictions, and one false positive prediction.
- Confusion matrices become more useful in multi-class problems, in which it can be difficult to determine the most frequent types of errors.

Recall

- Recall is the fraction of the truly positive instances that the classifier recognizes. A recall score of one indicates that the classifier did not make any false negative predictions.
- For our SMS spam classifier, recall is the fraction of spam messages that were truly classified as spam.
- Recall is calculated with the following ratio:

$$\text{Recall} = \frac{tp}{tp + fn}$$

- Recall is often called sensitivity in medical contexts.

Precision and Recall

- Individually, precision and recall are seldom informative; they are both incomplete views of a classifier's performance.
- Both precision and recall can fail to distinguish classifiers that perform well from certain types of classifiers that perform poorly.
- A trivial classifier could easily achieve a perfect recall score by predicting positive for every instance.

Precision and Recall

- For example, assume that a test set contains ten positive examples and ten negative examples.
- A classifier that predicts positive for every example will achieve a recall of one, as follows:

$$10\ 1\ 10\ 0\ R = = +$$

- A classifier that predicts negative for every example, or that makes only false positive and true negative predictions, will achieve a recall score of zero.
- Similarly, a classifier that predicts that only a single instance is positive and happens to be correct will achieve perfect precision.

Akaike Information Criterion

AIC

- Akaike's information criterionis a measure of the goodness of fit of an estimated statistical model.
- The AIC was developed by Hirotsugu Akaike under the name of "an information criterion" in 1971.
- The AIC is a **model selection** tool i.e. a method of comparing two or more candidate regression models.
- The AIC methodology attempts to find the model that best explains the data with a minimum of parameters. (i.e. in keeping with the law of parsimony)

- The AIC is calculated using the "likelihood function" and the number of parameters (Likelihood function : not on course).
- The likelihood value is generally given in code output, as a complement to the AIC.
- Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best.
- (Although, a difference in AIC values of less than two is considered negligible).

Akaike Information Criterion

- The Akaike information criterion is a measure of the relative goodness of fit of a statistical model.
- It was developed by Hirotsugu Akaike, under the name of "an information criterion" (AIC), and was first published by Akaike in 1974.

$$\mathrm{AIC} = 2p - 2\ln(L)$$

Akaike Information Criterion

- p is the number of free model parameters.
- L is the value of the Likelihood function for the model in question.
- For AIC to be optimal, n must be large compared to p.

Information Criterions

We define two types of information criterion: the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). In AIC and BIC, we choose the model that has the minimum value of:

$$AIC = 2log(L) + 2m,$$

$$BIC = 2log(L) + mlogn$$

where

- L is the likelihood of the data with a certain model,
- n is the number of observations and
- m is the number of parameters in the model.

AIC

- The Akaike information criterion is a measure of the relative **goodness of fit** of a statistical model.
- When using the AIC for selecting the parametric model class, choose the model for which the AIC value is lowest.

Schwarz's Bayesian Information Criterion

An alternative to the AIC is the Schwarz BIC, which additionally takes into account the sample size n.

$$\mathrm{BIC} = p \ln n - 2 \ln(L)$$

AIC and BIC in Two-Step Cluster Analysis

(Removed from Last Week's Class due to Version Update)
Two-Step Cluster Analysis guides the decision of how many clusters to retain from the data by calculating measures-of-fit such as **Akaikes Information Criterion (AIC)** or **Bayes Information Criterion (BIC)**.

- These are relative measures of goodness-of-fit and are used to compare different solutions with different numbers of segments.(“Relative” means that these criteria are not scaled on a range of, for example, 0 to 1 but can generally take any value.)
- **Important**: Compared to an alternative solution with a different number of segments, smaller values in AIC or BIC indicate an increased fit.

SPSS computes solutions for different segment numbers (up to the maximum number of segments specified before) and chooses the appropriate solution by looking for the smallest value in the chosen criterion. However, which criterion should we choose?

- AIC is well-known for overestimating the correct number of segments
- BIC has a slight tendency to underestimate this number.

Thus, it is worthwhile comparing the clustering outcomes of both criteria and selecting a smaller number of segments than actually indicated by AIC. Nevertheless, when running two separate analyses, one based on AIC and the other based on BIC, SPSS usually renders the same results.

Once you make some choices or do nothing and go with the defaults, the clusters are formed. At this point, you can consider whether the number of clusters is "good". If automated cluster selection is used, SPSS prints a table of statistics for different numbers of clusters, an excerpt of which is shown in the figure below. You are interested in finding the number of clusters at which the Schwarz BIC becomes small , but also the change in BIC between adjacent number of clusters is small.
The decision of how much benefit accrued by another cluster is very subjective. In addition to the BIC, a high ratio of distance of measures is desirable. In the figure below, the number of clusters with this highest ratio is three.

The Coefficient of Determination

- The coefficient of determination R^2 is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information.
- It is the proportion of variability in a data set that is accounted for by the statistical model.
- It provides a measure of how well future outcomes are likely to be predicted by the model.

R-Squared

- R^2 is a statistic that will give some information about the goodness of fit of a model.
- In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points.
- An R^2 of 1.0 indicates that the regression line perfectly fits the data.
- In the case of simple linear regression, the coefficient of determination is equivalent to the squared value of the Pearson correlation coefficient. (Consider this to be co-incidental, rather than a definition)

The Adjusted Coefficient of Determination

- Adjusted R^2 (often written as and pronounced "**R bar squared**") is a modification of R^2 that adjusts for the number of predictor terms in a model.
- Adjusted R^2 is used to compensate for the addition of variables to the model.
- As more independent variables are added to the regression model, unadjusted R^2 will generally increase but there will never be a decrease.
- This will occur even when the additional variables do little to help explain the dependent variable.

- To compensate for this, adjusted R^2 is corrected for the number of independent variables in the model, increases only if the new term improves the model more than would be expected by chance.
- If too many predictor variables are being used, this will be reflected in a reduced adjusted R^2.
- The adjusted R^2 can be negative, and will always be less than or equal to R^2.
- The result is an adjusted R^2 than can go up or down depending on whether the addition of another variable adds or does not add to the explanatory power of the model.
- Adjusted R^2 will always be lower than unadjusted.

- Adjusted R square is generally considered to be a more accurate goodness-of-fit measure than R square.
- It has become standard practice to report the adjusted R^2, especially when there are multiple models presented with varying numbers of independent variables.

Error Rates

- We can evaluate error rates by means of a training sample (to construct the discrimination rule) and a test sample.
- An optimistic error rate is obtained by reclassifying the training data. (In the **training data** sets, how many cases were misclassified). This is known as the **apparent error rate**.
- The apparent error rate is obtained by using in the training set to estimate the error rates.
- It can be severely optimistically biased, particularly for complex classifiers, and in the presence of over-fitted models.

- If an independent test sample is used for classifying, we arrive at the **true error rate**.
- The true error rate (or conditional error rate) of a classifier is the expected probability of misclassifying a randomly selected pattern.
- It is the error rate of an infinitely large test set drawn from the same distribution as the training data.

Misclassification Cost

- As in all statistical procedures it is helpful to use diagnostic procedures to asses the efficacy of the discriminant analysis. We use **cross-validation** to assess the classification probability. Typically you are going to have some prior rule as to what is an **acceptable misclassification rate**.
- Those rules might involve things like, "what is the cost of misclassification?" Consider a medical study where you might be able to diagnose cancer.
- There are really two alternative costs. The cost of misclassifying someone as having cancer when they don't. This could cause a certain amount of emotional grief. Additionally there would be the substantial cost of unnecessary treatment.

There is also the alternative cost of misclassifying someone as not having cancer when in fact they do have it.
A good classification procedure should

- result in few misclassifications
- take **prior probabilities of occurrence** into account
- consider the cost of misclassification

For example, suppose there tend to be more financially sound firms than bankrupt firm. If we really believe that the prior probability of a financially distressed and ultimately bankrupted firm is very small, then one should classify a randomly selected firm as non-bankrupt unless the data overwhelmingly favor bankruptcy.

- There are two costs associated with discriminant analysis classification: The true misclassification cost per class, and the expected misclassification cost (ECM) per observation.
- Suppose there we have a binary classification system, with two classes: class 1 and class 2.
- Suppose that classifying a class 1 object as belonging to class 2 represents a more serious error than classifying a class 2 object as belonging to class 1.
- There would an assignable cost to each error. $c(i|j)$ is the cost of classifying an observation into class j if its true class is i.

Misclassification

The costs of misclassification can be defined by a cost matrix.

	Predicted Class 1	Predicted Class 2
Class 1	0	$c(2\|1)$
Class 2	$c(1\|2)$	0

Expected cost of misclassification (ECM)

- Let p_1 and p_2 be the prior probability of class 1 and class 2 respectively. Necessarily $p_1 + p_2 = 1$.
- The conditional probability of classifying an object as class 1 when it is in fact from class 2 is denoted $p(1|2)$.
- Similarly the conditional probability of classifying an object as class 2 when it is in fact from class 1 is denoted $p(2|1)$.

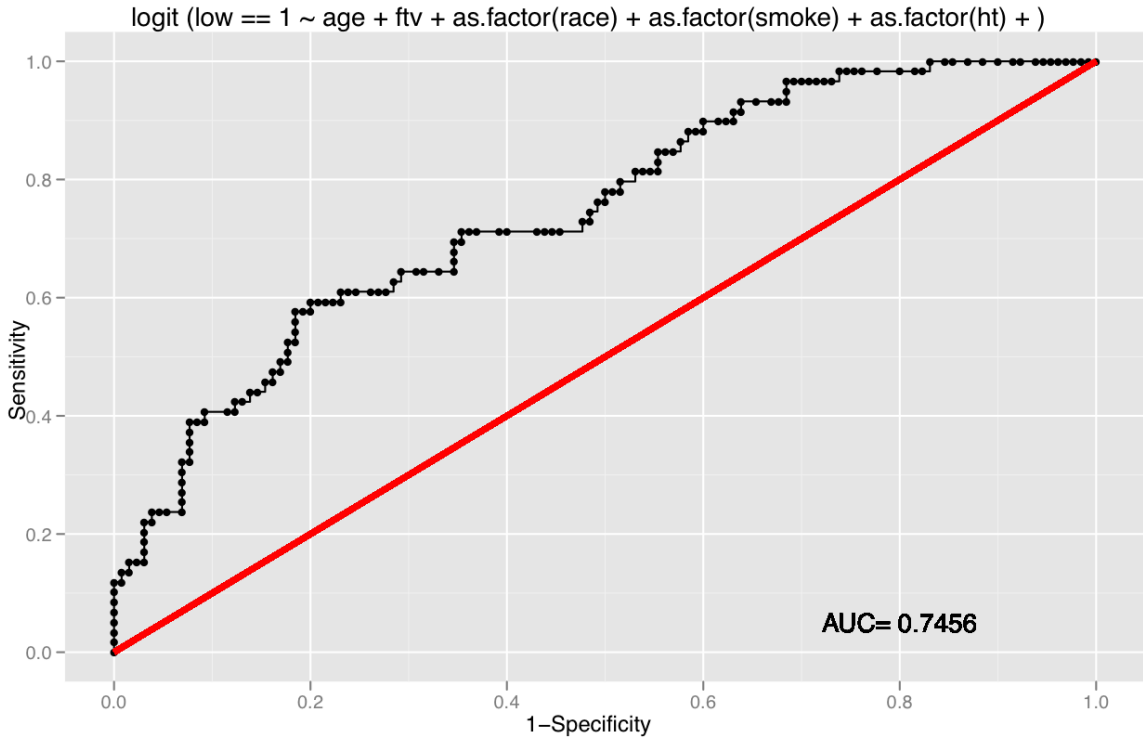
$$ECM = c(2|1)p(2|1)p_1 + c(1|2)p(1|2)p_2$$

(In other words: the sum of the cost of misclassification times the (joint) probability of that misclassification.

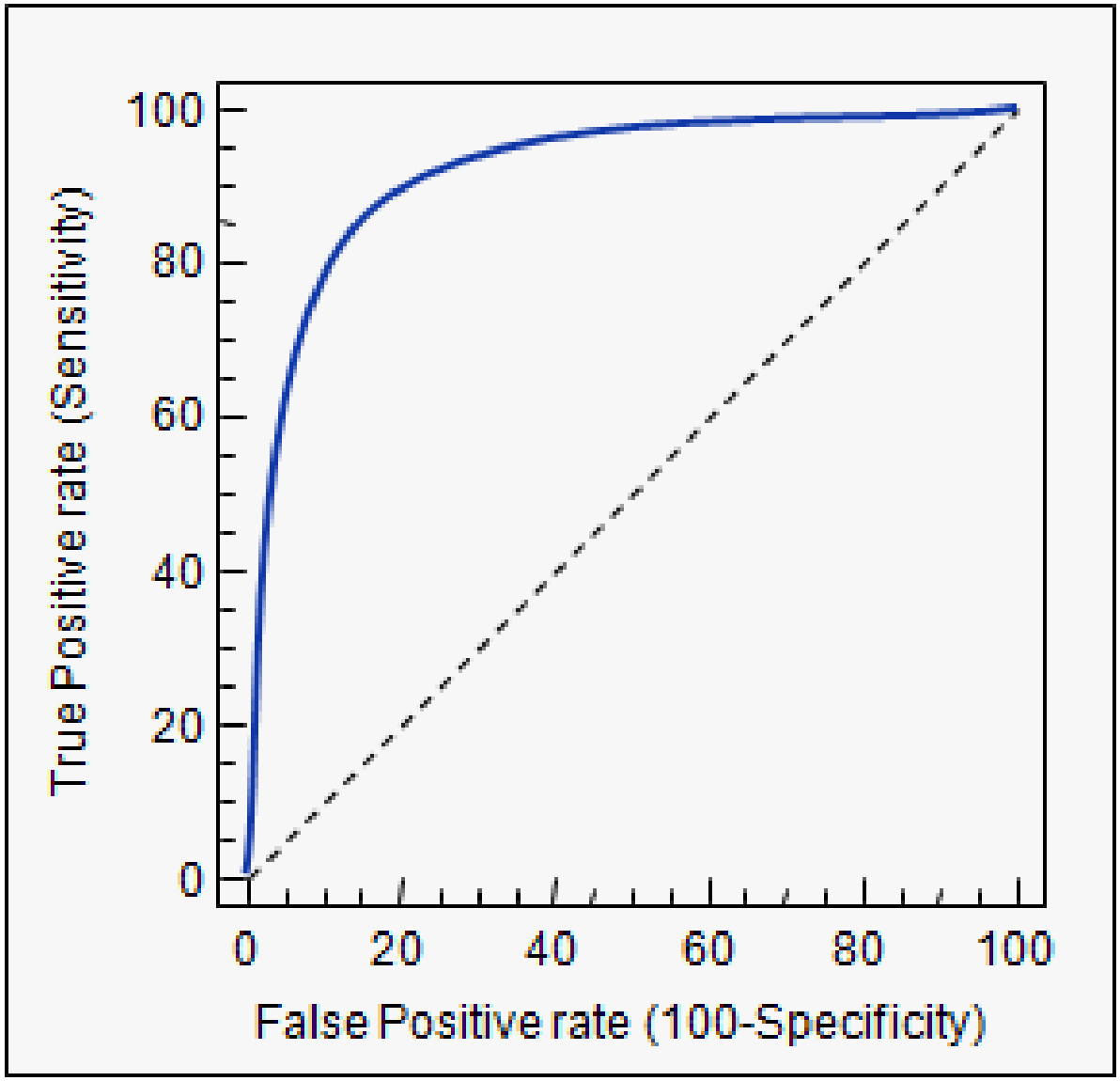
- A reasonable classification rule should have ECM as small as possible.

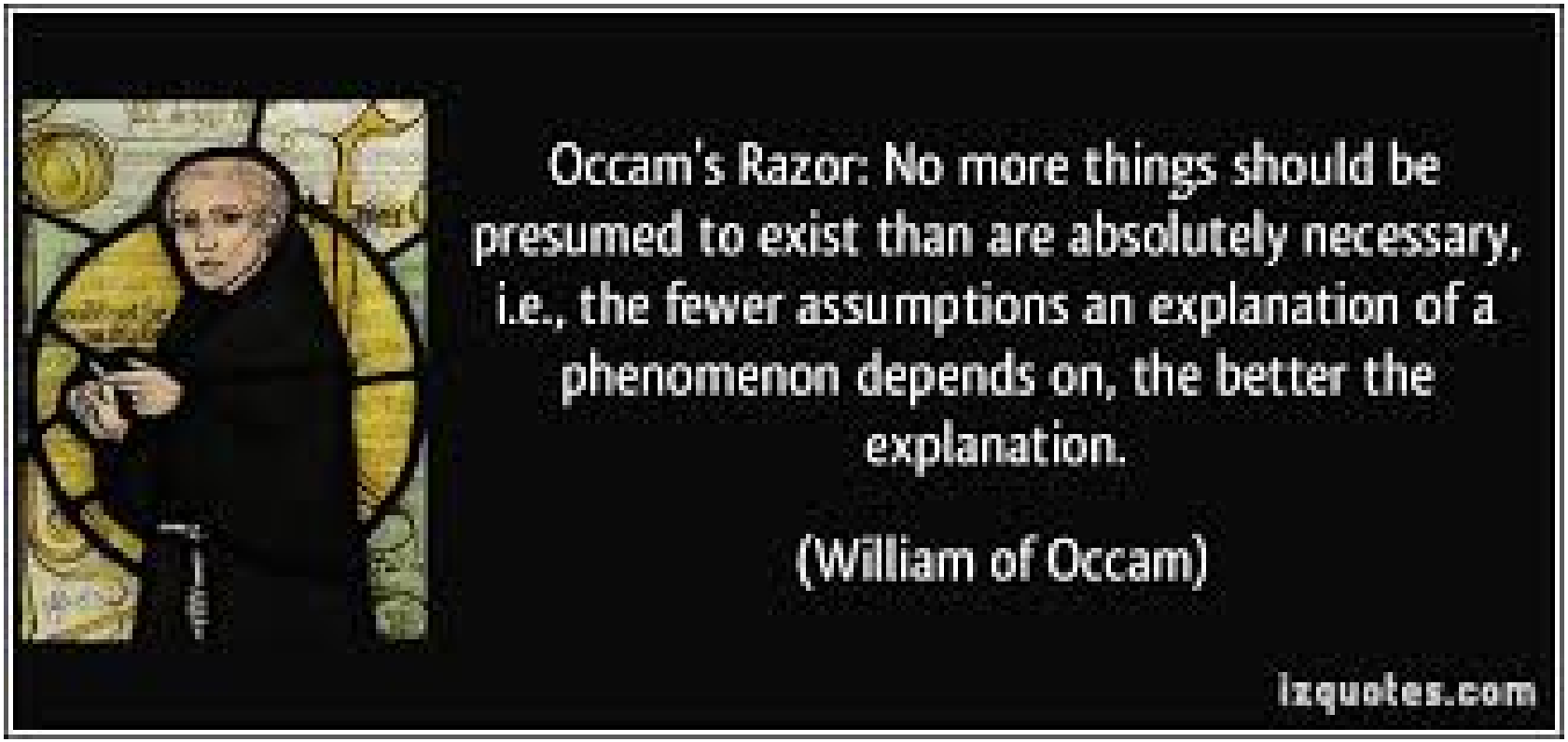
ROC curves

-
-


logit (low == 1 ~ age + ftv + as.factor(race) + as.factor(smoke) + as.factor(ht) +)
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC= 0.7456

AUC or Area Under the Curve is a single numerical measure of predictive power.




Occam's Razor: No more things should be presumed to exist than are absolutely necessary, i.e., the fewer assumptions an explanation of a phenomenon depends on, the better the explanation.
(William of Occam)
lzquotes.com

The Law of Parsimony

Ockham's Razor

- Ockham's razor, sometimes known as the law of parsimony, is simply a maxim that states that simple explanations are usually better than complicated ones.
- **Ockham's razor** was originally proposed by a monk named William of Ockham. (He did not call it "Ockham's razor" or even "my razor." This is a name that has been given to it over time.)

- Another version of this principle is the Law of Parsimony . This says that if you are choosing between two theories, choose the one with the fewest assumptions.
- Assumptions here means claims of fact that have no evidence.
- A theory that doesn't have many assumptions, and is very simple, is called a parsimonious theory.

Law of Parsimony

Law of Parsimony

In the context of statistics, the law of parsimony can be interpreted as an adequate model which requires the fewest independent variables is the preferred model.

Model building

- The traditional approach to statistical model building is to find the most parsimonious model that still explains the data.
- The more variables included in a model (overfitting), the more likely it becomes mathematically unstable, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data.

Model building

- Choosing the most adequate and minimal number of explanatory variables helps to find out the main sources of influence on the response variable, and increases the predictive ability of the model.
- As a rule of thumb, there should be more than 10 observations for each variable in the model.

Overfitting

- Overfitting occurs when a statistical model does not adequately describe of the underlying relationship between variables in a regression model.
- When overfitting happens, the model predicts the fitted data very well, but predicts future observations poorly.

Overfitting

- Overfitting generally occurs when the model is excessively complex, such as having too many parameters (i.e. predictor variables) relative to the number of observations.
- A model which has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

Variable-Selection Procedures

Variable Selection / Feature Selection

- In regression analysis, variable-selection procedures are aimed at selecting a reduced set of the independent variables - the ones providing the best fit to the model, in keeping with the Law of Parsimony.

- Occam's razor (or Ockham's razor) is a principle from philosophy. Suppose there exist two explanations for an occurrence.
- In this case the simpler one is usually better.
- Another way of saying it is that the more assumptions you have to make, the more unlikely an explanation is.
- Occam's razor applies especially in the philosophy of science, but also more generally.

Cross Validation

The confusion table is a table in which the rows are the observed categories of the dependent and the columns are the predicted categories. When prediction is perfect all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications.

Cross Validation

The cross validated set of data is a more honest presentation of the power of the discriminant function than that provided by the original classifications and often produces a poorer outcome. The cross validation is often termed a jack-knife classification, in that it successively classifies **all cases but one** to develop a discriminant function and then categorizes the case that was left out. This process is repeated with each case left out in turn.This is known as leave-1-out cross validation.

This cross validation produces a more reliable function. The argument behind it is that one should not use the case you are trying to predict as part of the categorization process.

Validation and Testing

Validation

- When you have sufficient data, you can subdivide your data into three parts called the training, validation, and test data.
- Rather than estimating parameter values from the entire data set, the data set is broken into three distinct parts. During the **variable selection** process, models are fit on the training data, and the prediction error for the models so obtained is found by using the validation data.
- Validation is the process of using part of a data set to estimate model parameters, and using the other part to assess the predictive ability of the model.
- Validation can be used to assess whether or not overfitting has occurred.

This prediction error on the validation data can be used to decide when to terminate the selection process or to decide what effects to include as the variable selection process proceeds. Finally, once a selected model has been obtained, the test set can be used to assess how the selected model generalizes on data that played no role in selecting the model.

1. The training set is the part that estimates model parameters.
2. The validation set is the part that assesses or validates the predictive ability of the model.
3. The test set is a final, independent assessment of the models predictive ability.

Model Validation

- A validation set is a portion of a data set used to assess the performance of prediction or classification models that have been fit on a separate portion of the same data set (the training set).
- Typically both the training and validation set are randomly selected, and the validation set is used as a more objective measure of the performance of various models that have been fit to the the training data (and whose performance with the training set is therefore not likely to be a good guide to their performance with data that they were not fit to).

Model Validation

It is difficult to give a general rule on how many observations you should assign to each role. One important textbook recommended that a typical split might be 50% for training and 25% each for validation and testing.