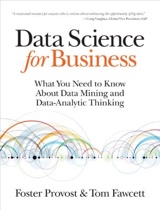
**Topic 6: Machine Learning: Performance Evaluation, Support Vector Machines & False Discoveries** Provost

* + 1. **Visualizing model performance**



1. **Describe a ranking classifier**.

**A close up of a keyboard

Description automatically generated**

﻿A ranking classifier is a classifier plus a threshold. It produces a single confusion matrix.

Whenever the threshold changes, the confusion matrix may change as well because the

numbers of true positives and false positives change.

Examples:

Web Search

Movie Rankings

Dating

1. **Define a profit curve**

**A way to compare the performance of various classifiers.**

**Profit curve -** takes into account dollar costs/benefits associated with true positives, false positives, true negatives, and false negatives.

A close up of a map

Description automatically generated

Percentage of population targeted.

* ﻿Each curve is based on the idea of examining the effect of thresholding the value of a classifier at successive points, implicitly dividing the list of instances into many successive sets of predicted positive and negative instances.
* ﻿The random classifier performs worst because it has an even chance of choosing a responder or a non-responder.
* ﻿The curves show that profit can go negative — not always, but sometimes they will, depending on the costs and the class ratio.
* The optimal threshold will ideally minimize the combined loss function you obtain by selecting a false positive or a false negative. If both are considered equally undesirable, you can generate a receiver-operating characteristic (ROC) curve and then determine the point which is closest in linear distance to the 100% sensitivity and specificity point.

1. **Calculate a confusion matrix using thresholding.**
   * + - The default value for threshold on which we generally get a Confusion Matrix is 0.50.
       - There is a tradeoff between Precision and Recall.
       - We want to optimize Precision for loan defaults. True positives from the selected group.
       - We want to optimize Recall for disease diagnosis. True positives from the relevant group.

* When you let go 100 culprits your recall is pretty low. But if you punish someone, you are absolutely sure that you are punishing only a criminal - precision is high.
  + - * Combine both with the harmonic mean F score.

1. **Describe the properties of a profit curve.**

A close up of a map

Description automatically generated

* Notice that all four curves begin and end at the same point.
* At the left, when no customers are targeted there are no expenses and zero profit.
* Notice that things can go negative. All the curves show this.
* At the ﻿right side everyone is targeted, so every classifier performs the same.
* Dotted line is random.
* ﻿If your goal was simply to maximize profit and you had unlimited resources, you should choose Classifier 2, use it to score your population of customers, and target the top half (highest 50%) of customers on the list.
* Constraint: What if you were constrained to 8,000 customers? ﻿Check the performance curves at x=8%. The best-performing model at this performance point is Classifier 1.

1. **Calculate points on the profit curve**

A close up of a map

Description automatically generated

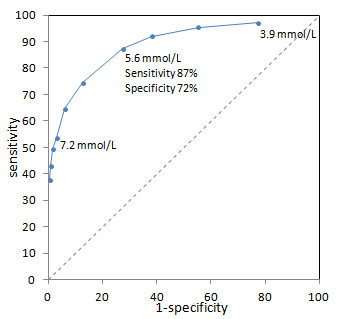
For the profit curve you need:

1. ﻿The class priors; that is, the proportion of positive and negative instances in the target population, also known as the base rate (usually referring to the proportion of positives).
2. The costs and benefits. The expected profit is specifically sensitive to the relative levels of costs and benefits for the different
3. **Describe the ROC graph. (Receiver Operating Characteristic) graph.**

 Note: Developed for military radar

* ROC curve lets you summarize a multitude of confusion matrices.
* The diagnostic ability of a [binary classifier](https://en.wikipedia.org/wiki/Binary_classifier) system with varied thresholds.
* The green diagonal line is where True Positive Rate = False Positive Rate.
* The “steepness” of ROC curves is also important, since it is ideal to maximize the true positive rate while minimizing the false positive rate.
* Lower left is 0,0. Nothing correctly identified.
* Upper right is 1,1. Correctly identify true positives but false positives all incorrect.
* ﻿The ROC graph depicts relative trade-offs that a classifier makes between benefits (true positives) and costs (false positives).

A close up of a clock

Description automatically generated

1. **Calculate points on the ROC graph using data from a confusion matrix.**

A close up of a map

Description automatically generated

* The ROC graph summarizes all of the confusion matrices which each threshold produced.
* A point at 1,1 means the threshold is set for correct classifications of True Positives but an Incorrect classification of all False Positives.

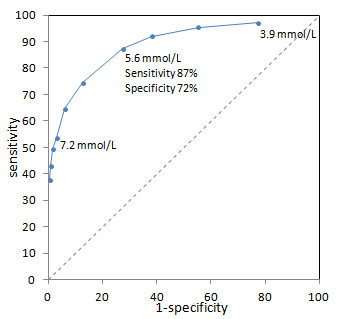
1. **Define base rate**

**﻿**The class priors; that is, the proportion of positive and negative instances in the target population, also known as the base rate (usually referring to the proportion of positives).

There is a base rate fallacy which happens with a small sample size.

1. **Describe the 4 corners and the diagonal of the ROC (Receiver Operating Characteristic) graph.**

* Upper Right: Every True Positive and Every False Positive
* Upper Left: Every True Positive and 0 False Positive. The BEST.
* Lower Right: Every False Positive and 0 True Positive. WORST.
* Lower Left: 0 True Positive and 0 False Positive



1. **Define “hit” rate (TPR) and “false alarm” rate (FPR)**

 80s hitman. ‘hit’ rate. TPR. Sensitivity.

 False Alarm. ‘false alarm’ rate. FPR.

TPR = Hit Rate

What percent of the actual positives does the classifier get right.

FPR = False Alarm Rate. Opposite of Sensitivity.

What percent of the actual negatives does the classifier get wrong.

**True Positive Rate (TPR)** **= Sensitivity (TPR) = Recall (TPR) =** **Probability of Detection (TPR)**

**A close up of a logo

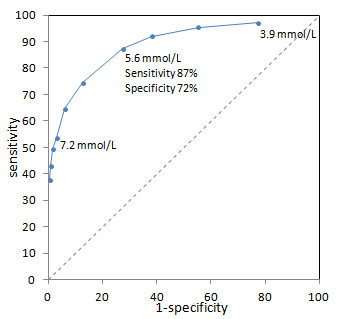
Description automatically generated**

Measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition). In epidemiology this is known as sensitivity.

1. **Describe how the ROC (Receiver Operating Characteristic) space can be used to evaluate classifiers.**

The closer a curve lies to the top left corner, the greater the area underneath it. Thus, a larger area under the ROC curve implies a better test.

This represents the intuitive trade-off between sensitivity (rising as we move up) and specificity (dropping as we move right). (1-specificity) gets larger as specificity drops.



1. **Define AUC measure**

A screenshot of a cell phone

Description automatically generated

* Area Under the Curve aka the Area Under the ROC.
* The more, the better. It includes everything under the curve to the edges.
* .5 baseline which is half the area.
* The AUC value lies between **0.5 to 1** where 0.5 denotes a bad classifier and 1 denotes an excellent classifier.
* The AUC is used to determine which method is better (e.g., Logistic Regression vs. Random forest).

1. **Describe a Cumulative Response Curve (aka Gains) also known as the Lift Curve**

Cumulative Response Curve = Gains = Lift curve. Want to maximize the area of this.

* ﻿The cumulative gains curve is an evaluation curve that assesses the performance of the model and compares the results with the random pick. It shows the percentage of targets reached when considering a certain percentage of the population with the highest probability to be target according to the model.
* The disadvantage of a profit graph is that it requires that operating conditions be known and specified exactly.
* With many real-world problems, the operating conditions are imprecise or change over time, and the data scientist must contend with uncertainty.
* When costs and benefits cannot be specified with confidence, but the class mix will likely not change, a cumulative response or lift graph is useful.
* Cumulative response curves (CRCs) display patient response rates over a continuum of possible thresholds, thus eliminating problems with a rigid threshold definition, allowing for a variety of response thresholds to be examined simultaneously, and encompassing all data.
* Cumulative response curves plot the hit rate or TPR on the y axis, (i.e., the percentage of positives correctly classified, as a function of the percentage of the population that is targeted which is on the x-axis).
* So, unlike confusion matrices with target the population as a whole, the Cumulative Response Curve (CRC) only deals with a PORTION of the population.

Cumulative Response Curve:

* Y-axis is the ‘hit’ rate. TPR = Sensitivity

**Cumulative Response Curve:**

**A close up of a map

Description automatically generated**

 Note: Hit rate on y-axis with only a portion of the data.

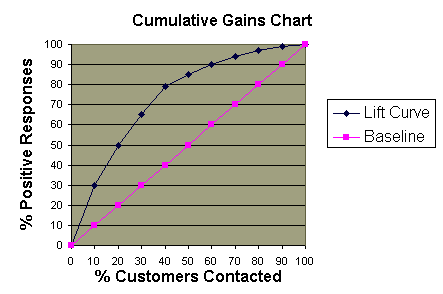
**Lift:**

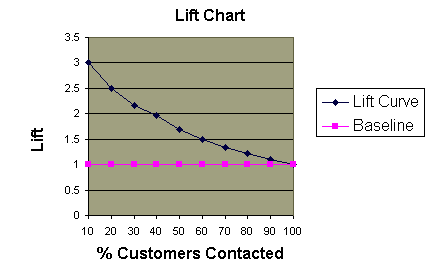
A screenshot of a social media post

Description automatically generated

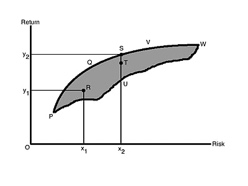
* ﻿Lift = Cumulative gains / random model (blue line).
* Select model with the greatest lift for marketing campaign.
* The lift curve is essentially the value of the cumulative response curve at a given x point divided by the diagonal line value at that point.
* ﻿The lift curve shows that the model’s targeting is twice as good as random.
* ﻿ 2X lift means that at the chosen threshold (often not mentioned), the lift curve shows that the model’s targeting is twice as good as random.

1. **Calculate points on the cumulative response curve (CRC)**





* + 1. **Back testing protocol in the Era of Machine Learning**

  Note: Markowitz

1. **Explain why cross validation may not reduce the curse of dimensionality.**

Data is limited in finance. Usually only monthly and quarterly data.

Sparse. Any kind of cross validation will not reduce the curse of dimensionality.

Just not enough data in finance.

1. **Describe research protocols and their impact on false-positives discoveries.**
2. You need to randomize your data for proper cross validation
3. Data is very limited
4. We have a strong prior that the strategy is false: if it works because of luck.
5. **Explain the role of an economic foundation when applying machine learning tools.**

* When data is limited, economic foundations become more important
* You need a hypothesis. Empirical tests are then run to prove the hypothesis wrong. The hypothesis provides a discipline which reduces the chance of overfitting.
* Without a foundation you maximize the chance of failure in live trading.

1. **Describe the winner’s curse.**

Note: Not the same case as in auction theory. When the trial is replicated, three different situations could occur.

1. The trial stands up against many replication tests.
2. The effect is much smaller than in the original finding.
3. There is no effect.
4. **Define exaggerated positive**

More common than the false positive is the exaggerated positive.

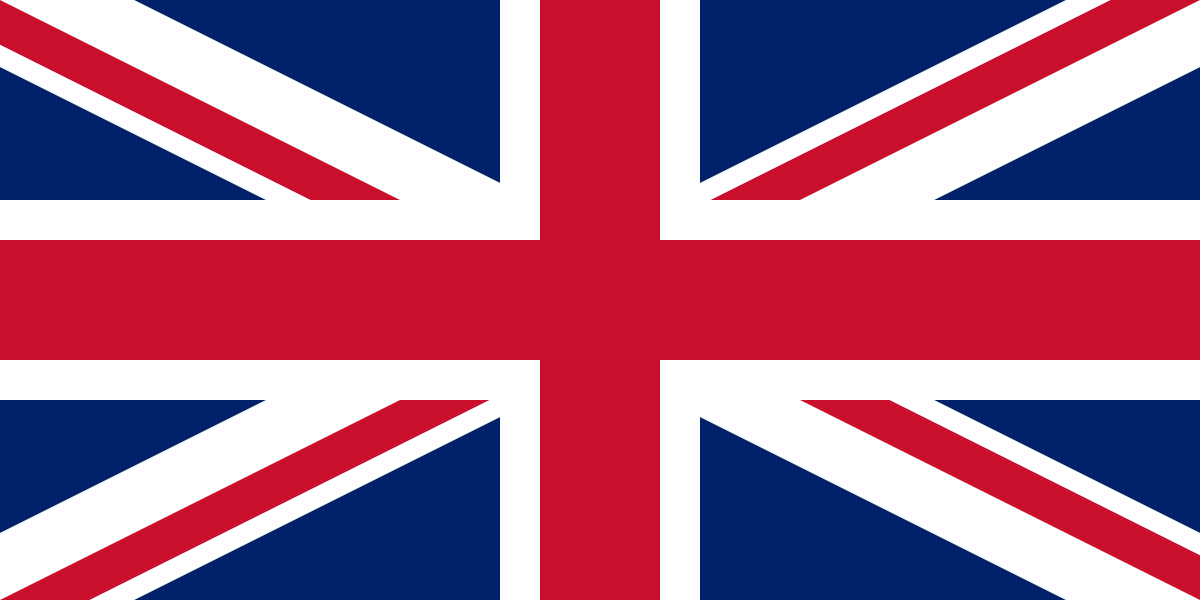
Researchers want their model to work.

Too much optimism may hurt us. We do not perceive our models accurately.

1. **Describe the 7 protocols suggested for avoiding False Positives (TYPE I Error)**

RMS CMMC. Note: “Remiss to see M&Ms crushed.”

1. Research Motivation
2. Establish an ex-ante economic foundation
3. Beware of ex post economic foundation
4. Multiple testing and statistical methods
5. Keep track of what is tried
6. Keep track of combinations of variables
7. Beware the parallel universe problem
8. Sample choice and data
9. Define the test sample ex-ante
10. Ensure Data Quality
11. Document choices in data transformation
12. Do not arbitrarily exclude outliers
13. Select Winsorization level before constructing a model. Truncating outliers.
14. Cross validation
15. Acknowledge that out of sample is not really out of sample
16. Recognize that iterated out of sample is not out of sample
17. Do not ignore trading costs and fees
18. Model dynamics
19. Be aware of structural changes
20. Acknowledge the Heisenberg principle and acknowledge Overcrowding
21. Refrain from tweaking the model
22. Model complexity
23. Beware of the curse of dimensionality
24. Pursue simplicity and regularization
25. Seek Interpretable Machine learning
26. Culture of Research
    * + 1. Establish a Research Culture that Rewards Quality
        2. Be careful with delegated research
      1. **An investigation of the false discovery rate (FDR) and the misinterpretation of p-values.** FDP site article from Royal British Academy.

 P-values Article

1. **Define sensitivity and specificity**

**A close up of a logo

Description automatically generated**

TRUE Positive Rate **TPR**. In medical diagnosis, test ***sensitivity*** is the ability of a test to correctly identify those with the disease (true positive rate). TP / (TP+TN)]

A close up of a logo

Description automatically generated

Test ***specificity*** is the ability of the test to correctly identify those without the disease (true negative rate). TN / (FP+TN).

A type I error is a false rejection of the null hypothesis.

Investment example:

A false positive more important. If we were to pour our entire net worth into this one stock, we would better hope that our model is right. **Precision would be the best** metric to use here because it determines the correctness of our model. We can afford to miss a few profitable stock investments here and there (*so recall score is not as important*), as long as our money is going to an appreciating stock correctly predicted by our model. Missed opportunities not as costly.

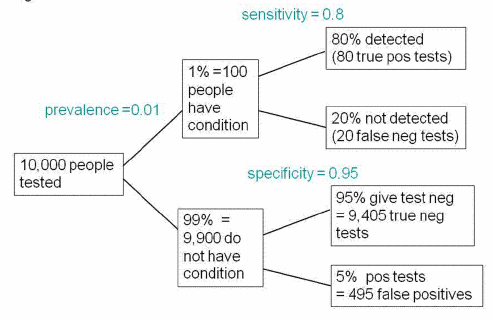
Poison Apple example:

A false negative more important. Let’s say we were trying to detect if an apple was poison or not. In this case, we would want to *reduce* the number of False Negatives because we hope to not miss any poison apples in the batch. **Recall would be the best** evaluation metric to use here because it measures how many poison apples we might have missed. We are not too concerned with mislabeling an apple as poisonous because we would rather be safe than sorry.

1. **Describe the false discovery rate (FDR) with the help of a tree diagram. Type I error.**

Just describe with tree:





Tree diagram to illustrate the false discovery rate in screening tests. This example is for a prevalence of 1%, specificity 95% and sensitivity 80%.

Prevalence in statistics is defined as the proportion of a population which has a particular

characteristic during a specified time period. Think of prevalence as your selected group.

Out of 10 000, 80 + 495 (Note: this is the p-spot) = 575 give positive tests.

Of these, 495 are false positives so the false discovery rate is 86%. 495/575 = False Positives / (the P spot)

1. **Calculate the probability of real effect given a result is significant**

Real effects on a sample vs. just using significance with p value. See example in the article.

1. **Define power of a test.**

 CORRECTLY rejecting the Null Hypothesis.

The easiest definition for students to understand is power is the probability of correctly

rejecting the null hypothesis.

The greater the power, the lower the Type II error. In statistical hypothesis testing, Type II error is given by the letter β. (Type I is α).

The "[power](https://en.wikipedia.org/wiki/Statistical_power)" (or the "[sensitivity](https://en.wikipedia.org/wiki/Sensitivity_and_specificity)") of the test is equal 1−β. The statistical power ranges from 0 to 1, and as statistical power increases, the probability of making a type II error decreases.

We need to know the probability that the test will give the right result when there is a real effect. This is called the power of the test. The power depends on the sample size, and on the size of the effect we hope to detect.

The power depends on the sample size, and on the size of the effect we hope to detect.

When calculating sample sizes, the power is commonly set to 0.8, so when there is a real

effect, we will detect it (declare the result to be ‘significant’) in 80% of tests.

We’re typically only interested in the power of a test when the null is in fact false. This

definition also makes it clearer that power is a conditional probability: the null hypothesis

makes a statement about parameter values, but the power of the test is conditional upon

what the values of those parameters really are.

An old photo of a person

Description automatically generated

Note: Sir William Blackstone. “It is better that ten guilty persons escape than that one innocent suffer.”

Precision is more important than recall when you would like to have less False Positives in

trade off to have more False Negatives. Meaning, getting a False Positive is very costly, and a

False Negative is not as much.

1. **Calculate the false discovery rate (FDR).**



FDR is a very simple concept. It is the number of false discoveries in an experiment divided by total number of discoveries in that experiment. Like precision but opposite.

FDR-controlling procedures provide less stringent control of Type I errors compared to familywise error rate (FWER) controlling procedures (such as the Bonferroni correction), which control the probability of *at least one* Type I error. Thus, FDR-controlling procedures have greater power, at the cost of increased numbers of Type I errors.

The false discovery rate is the complement of the positive predictive value (PPV) which is the probability that, when you get a ‘significant’ result there is actually a real effect. If the false discovery rate is 70%, then the PPV is 30%. The false discovery rate is a more self-explanatory term, so it is preferred.

1. **Describe underpowered study**



The power of a study is the likelihood that it will distinguish an effect of a certain size from pure luck.

An underpowered study does not have a sufficiently large sample size to answer the research question of interest.

An overpowered study has too large a sample size and wastes resources.

1. **Describe the INFLATION EFFECT in the context of false discovery (Type I error)**

A close up of a logo

Description automatically generated

Note = Low Power means High Inflation.

The inflation effect gets really serious when the POWER is LOW.

The estimated effect size is almost twice its true value with a power around 0.2. That is because the test is more likely to be positive in the small number of experiments that show a larger than average effect size.

1. **Describe what happens when we consider p=0.05 rather than p<=0.05.**

P-hacking is possible.

The outcome is that if you declare that you have made a discovery when you observe a p-value close to 0.05, you have at the least a 26% chance of being wrong, and often a much bigger chance.

Yet many results get published for which the false discovery rate is at least 30%. No wonder there is a problem of reproducibility.

1. **Describe Berger’s approach**

Note: James O. Berger

He gave a result that applies regardless of what the shape of the prior distribution might be. In effect, it chooses the prior distribution that is most favorable to the hypothesis that there is a real effect.

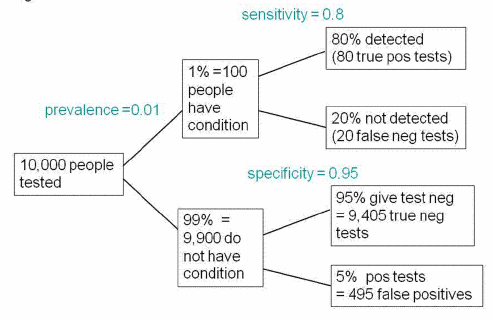
Using this, one can calculate the minimum false discovery rate that corresponds to any observed p-value.

Adopt a three-sigma policy rather than a two-sigma policy.

Berger’s approach suggests you set p = .00027 which corresponds to a false discovery rate of .042, not far from the .05 level that is customarily abused.

1. **Calculate the false discovery rate (FDR) using conditional probabilities.**





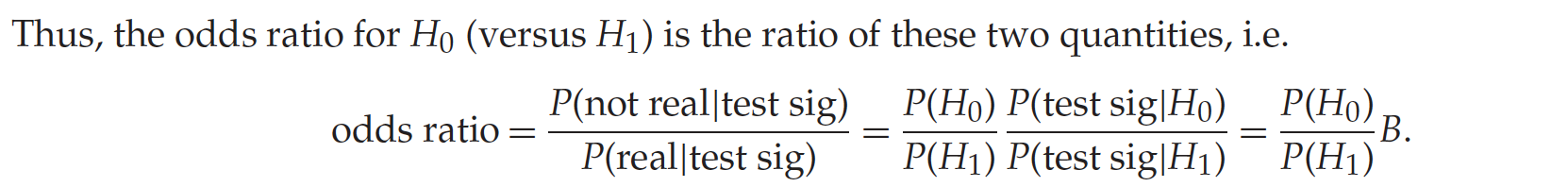
1. **Calculate the conditional probability of the real effect. New.**

**A picture containing screenshot

Description automatically generated**

1000 tests. In which the prevalence of real effects is 10%. The lower limb shows that with the conventional significance level, p= 0.05, there will be 45 false positives. The upper limb shows that there will be 80 true positive tests. ‘the false discovery rate s therefore 45/(45+80) =36%, far bigger than 5%.

1. **Calculate the odds ratio using the Bayes approach.**

****

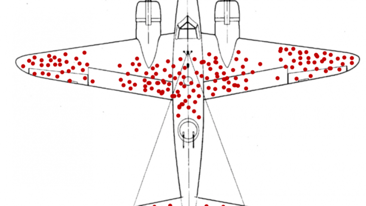
Likelihood ratio. The probability of observing the data, given a hypothesis, i.e. the probability of observing the data if that hypothesis were true.

**6.4.1 A Data Science Solution to the Multiple-Testing (SBuMT) Crisis** De Prado

**A. Define Selection Bias Under Multiple Testing (SBuMT).**

Note: Selection Bias. No runts. This is P hacking

 Note: Famous military plane study of selection bias

There is a strong possibility of performing multiple tests on the same dataset and selecting the most favorable outcome (the one that rejects the null hypothesis with the lowest false positive probability).

There is nothing wrong with running multiple tests; however, when the extent of the tests carried out is hidden from journal referees, readers, and investors, it is impossible for them to assess whether a particular result is a false positive.

* + 1. **Describe the 3 properties satisfied by trials to reduce SBuMT.**

1. Complete

The set includes every back test.

Researchers do not have the ability to delete trials.

1. Coerced

Researchers do not choose what to log or present. Terabytes of intermediate research are automatically recorded.

1. Untainted

Every batch of back tests must be preapproved by the research committee to prevent that externally preselected trials contaminate the internal trials.

* + 1. **Describe clustering of trials**

A picture containing fabric

Description automatically generated

* Prado introduced an algorithm for the optimal number of clusters.
* Prado’s article shows that the best clustering occurs when k = 4.
* The off diagonal blocks appear when k >= 4.
* Low number has to do with constraints from mathematical formulas for bond pricing.
  + 1. **Describe the implications of using an optimal number of clusters**

Compute each cluster returns applying the minimum variance allocation so that highly volatile returns do not dominate the returns time series.

Otherwise, a single volatile trial might bias the time series of returns that characterize the entire cluster.

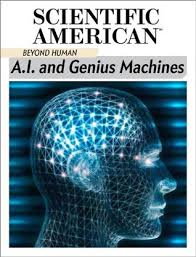
* + 1. **Describe how clustering of strategies can reduce SBuMT**

Even though the empirical evidence strongly indicates that k=4 is the optimal clustering, we choose to provide full results for all k = 2, ..., 10.

In this way referees and readers can evaluate the robustness of the conclusions under alternative solutions. K=4 still is best and is unlikely to be a false positive under SuBMT.

DSR = Deflated Sharpe Ratio which takes into account multiple testing.

* + 1. **Describe the implications for authors, journals, and financial firms**

****

Authors

1. Add to every publication an appendix explaining why the purported discovery is not a false positive caused by SBuMT.
2. Certify they have logged and recorded ALL trials that took place during their research
3. Provide to journal referees the outcomes from ALL trials

Journals

1. Demand that authors disclose ALL trials
2. Report the extent to which their findings are affected by SBuMT.
3. Evaluate the robustness of their findings to alternative scenarios of SBuMT.

Financial Firms

1. Avoid the practice of optimizing back tests (picking the winners and ignoring the losers)
2. Implement research surveillance frameworks that record, store, and curate every single research trial that takes place within the organization
3. Estimate the probability of a false positive, objectively controlling for SBuMT.