A/B Testing

Md Emdadul Sadik

Md Enamul Huq Sarker

Summer 2020

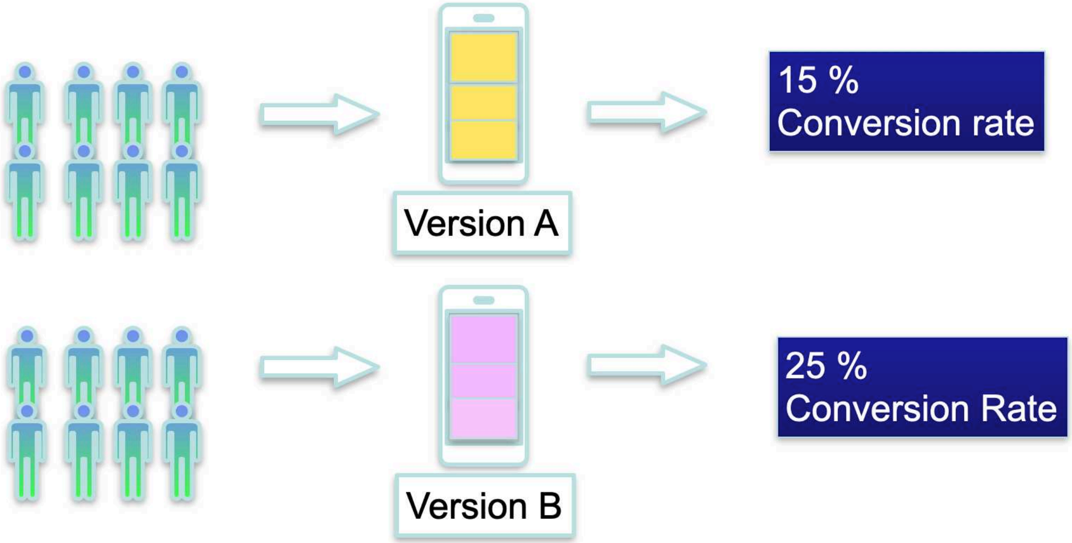
TECHNISCHE
UNIVERSITÄT
DARMSTADT

Overview

1. A/B testing
    - What is it?
    - Why is that used?
    - When (or not) to use A/B test?
    - Hypothesis testing & p-value
    - Type I & Type II error
2. Multivariate testing
3. A/B Testing of ML Models

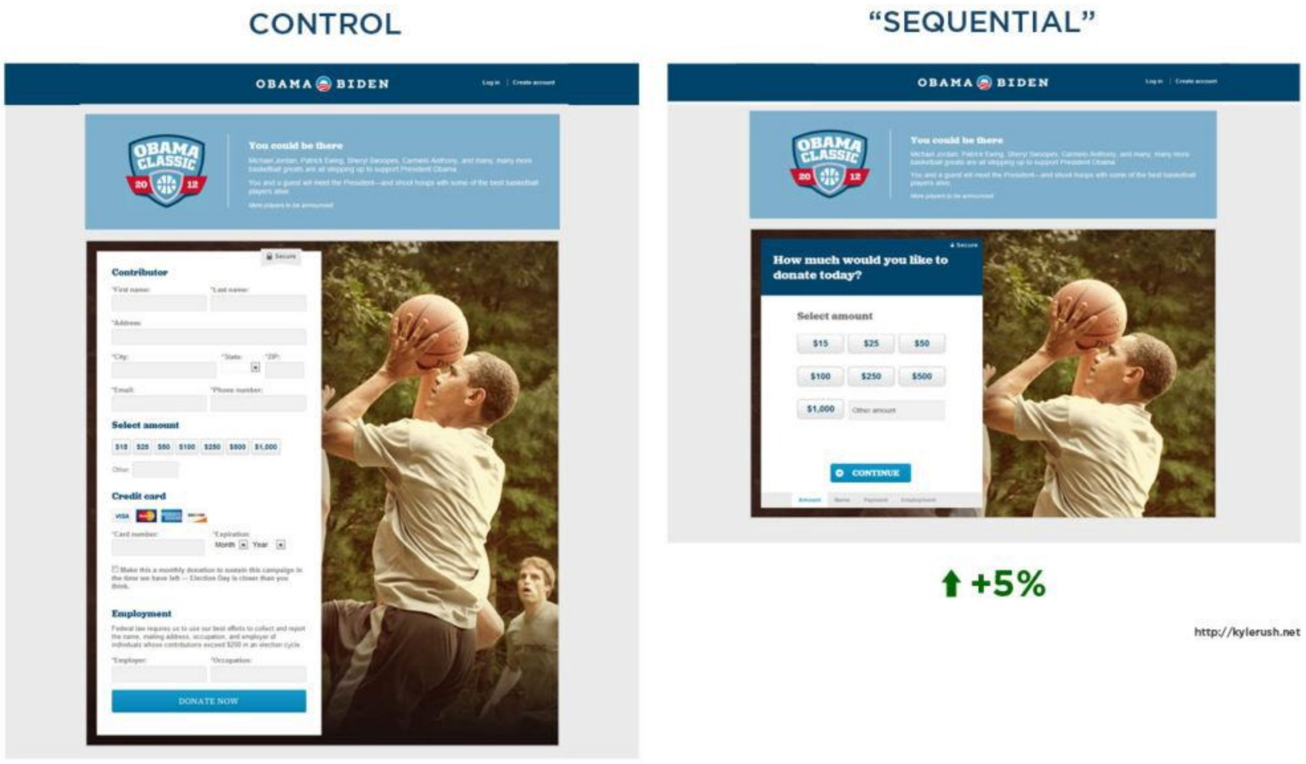
What is A/B Testing?

- A user experience research methodology.
- Compares two versions of design alternatives (i.e two versions of a single variable)



Obama campaign 2012

- A/B testing in Obama's 2012 presidential campaign
- 165 team digital team
- 500+ experiments
- Over 20 months
- $190 million extra



CONTROL

OBAMA BIDEN — Log in | Create account

You could be there

Michael Jordan, Patrick Ewing, Sheryl Swoopes, Carmelo Anthony, and many, many more basketball greats are all stepping up to support President Obama.

You and a guest will meet the President—and shoot hoops with some of the best basketball players alive.

More players to be announced

Secure

Contributor

*First name: *Last name:

*Address:

*City: *State: *ZIP:

*Email: *Phone number:

Select amount

$15 $25 $50 $100 $250 $500 $1,000

Other:

Credit card

VISA

*Card number: *Expiration: Month Year

□ Make this a monthly donation to sustain this campaign in the time we have left — Election Day is closer than you think.

Employment

Federal law requires us to use our best efforts to collect and report the name, mailing address, occupation, and employer of individuals whose contributions exceed $200 in an election cycle.

*Employer: *Occupation:

DONATE NOW

"SEQUENTIAL"

OBAMA BIDEN — Log in | Create account

You could be there

Michael Jordan, Patrick Ewing, Sheryl Swoopes, Carmelo Anthony, and many, many more basketball greats are all stepping up to support President Obama.

You and a guest will meet the President—and shoot hoops with some of the best basketball players alive.

More players to be announced

Secure

How much would you like to donate today?

Select amount

$15 $25 $50

$100 $250 $500

$1,000 Other amount

CONTINUE

Amount Name Payment Employment

+5%

http://kylerush.net

Image source

Should I use A/B Test?

- All the big companies use A/B testing. But why?
  - Intuition can be often wrong! Reading user mind is complex.
  - Higher risk to roll out a features to all users.

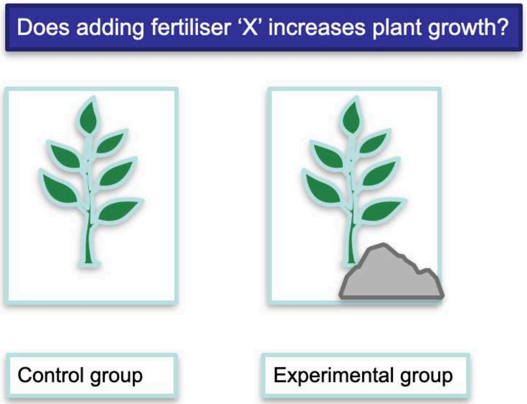
- Think, if you should use A/B testing in below cases?
  - Changing colour or theme of a website
  - Changing company logo
  - Car sellers website
  - Movie preview

When A/B test shouldn't be used?

- You shouldn't go for A/B test if
  - You don't have meaningful traffic
    - Statistically significant sample size is important.
  - You can't spend the mental bandwidth.
  - You don't have a solid hypothesis to start with.
    - Ex: Adding a 'Finish purchase' button will increase purchase by 20 percent.
  - Risk is too low to immediate action.
    - Implementation is preferable instead of wasting time on A/B testing

Common terms

- What is a hypothesis?
  - Claim or idea to be tested
- Control group
  - Doesn't get special treatment.
- Experimental group
  - Gets special treatment.
- Null hypothesis (H_0)
  - Outcome from control and treatment are identical.
- Alternate Hypotheis (H_a)
  - Outcome from treatment is different.



Hypothesis Testing

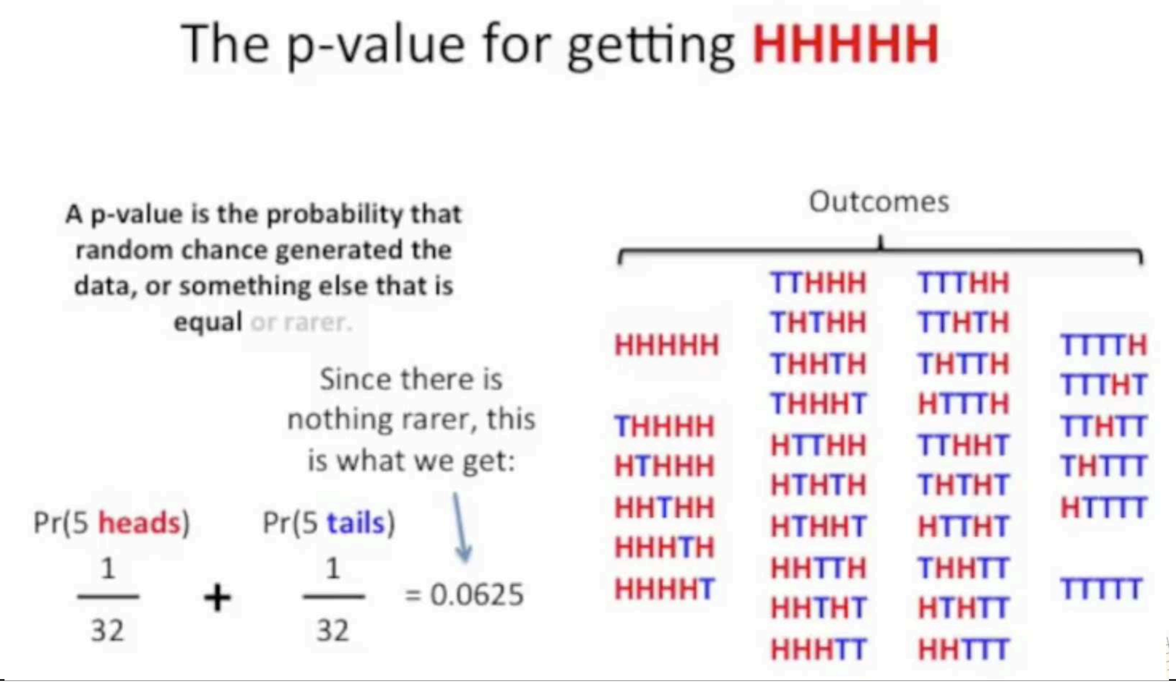
- Average session time is 20 minutes
- Change website background colour from Blue to Orange
- How to do the hypothesis testing?
  1. Null hypothesis (H_0) : mean = 20 minutes after the change
  2. Alternate hypothesis (H_a) : mean > 20 minutes after the change
  3. Significance level (p-value threshold): $\alpha = .05$
  4. Take sample, for example, $n = 100$, sample mean $\bar{X} = 25$ minutes.
  5. p-value: $P(\bar{X} >= 25 \text{ minutes} \mid H_0 \text{ is true})$
     - If p-value < α then reject $H_{0,}$ suggest H_a
     - If p-value >= α then don’t reject $H_{0,}$ (doesn’t mean accept H_0)

Hypothesis Testing (cont.)

- If p-value < α then reject $H_{0,}$ suggest H_a
- If p-value >= α then don’t reject $H_{0,}$ (doesn’t mean accept H_0)
- Example:
  - p_value is 0.03, reject $H_{0,}$ suggest H_a
  - p_value is 0.05, Fail to reject reject H_0
- Why should you set significance value prior to the experiment?
  - Ethical reason

How to calculate P-value

- P-Value means probability value which indicates how likely a result occurred by chance alone
- P-value is calculated as probability of the random chance that generated the data or (+) something else that is equal (probability) or (+) something rarer (less probability)



The p-value for getting HHHHH

A p-value is the probability that random chance generated the data, or something else that is equal or rarer.

Since there is nothing rarer, this is what we get:

$$\underbrace{\frac{1}{32}}_{\Pr(5\ \text{heads})} + \underbrace{\frac{1}{32}}_{\Pr(5\ \text{tails})} = 0.0625$$

Outcomes

HHHHH

THHHH
HTHHH
HHTHH
HHHTH
HHHHT

TTHHH
THTHH
THHTH
THHHT
HTTHH
HTHTH
HTHHT
HHTTH
HHTHT
HHHTT

TTTHH
TTHTH
THTTH
HTTTH
TTHHT
THTHT
HTTHT
THHTT
HTHTT
HHTTT

TTTTH
TTTHT
TTHTT
THTTT
HTTTT

TTTTT

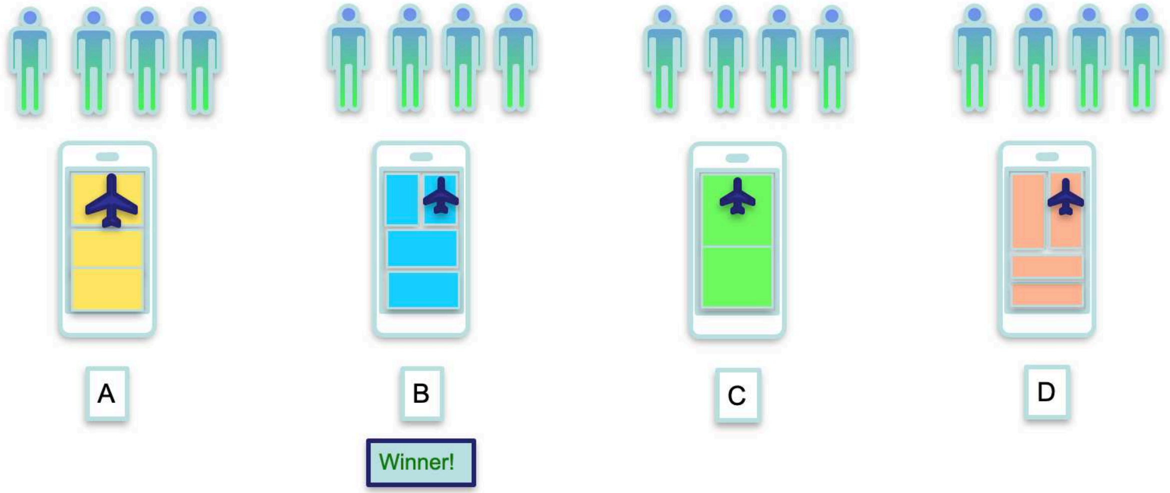
Type I and Type II error

	Fail to Reject	Reject
H_0 is true	Correct conclusion	Type I error
H_0 is false	Type II error	Correct conclusion

- How to reduce Type I error?
    - Lower the value α
    - Reducing value of α, increases type II error
- How to to reduce Type II error?
    - Increased sample size
    - Less variability
    - True parameter far from from H_0

Multi variate & A/A testing

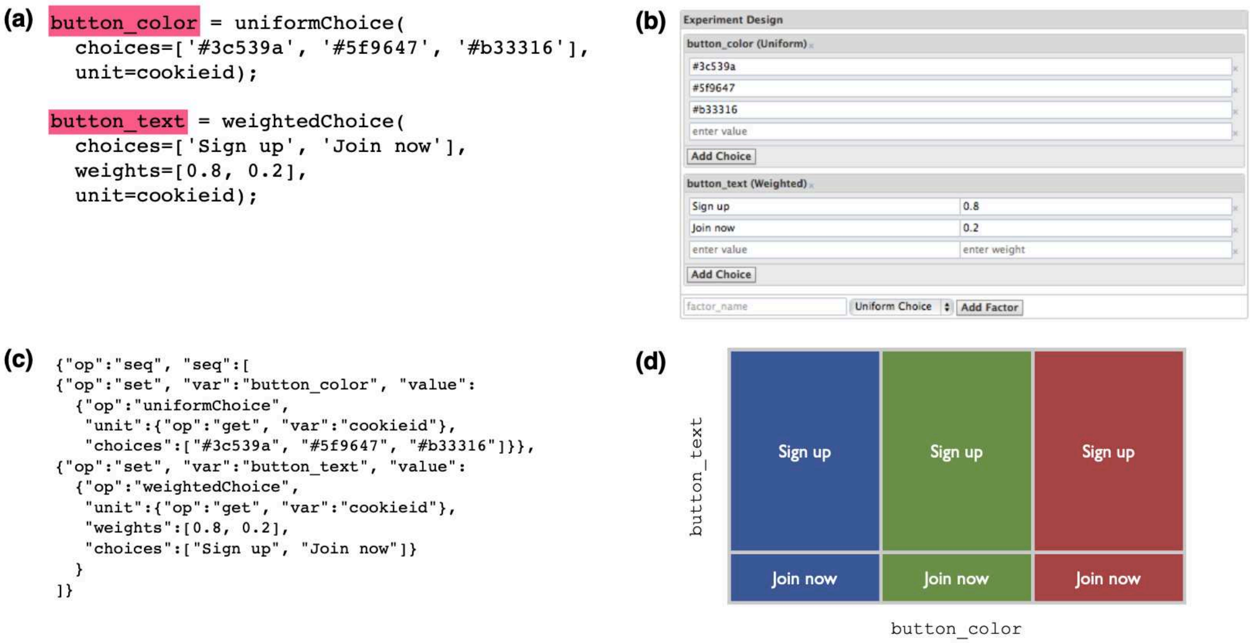
- **Multivariate testing :** Multiple variables are modified, also called full factorial testing.
  - Advantage: A lot of combinations can be tested
  - Limitation: Bigger sample size, complex, needs better understanding of interactions



- **A/A Testing:**
  - Identical version is compared against each other.
  - Used to validate the tool(s) being used.

Factorial testing with PlanOut

- Factorial test is complex to realise and implement.
- Planout (https://facebook.github.io/planout/) a framework for online field experiment



(a)

```
button_color = uniformChoice(
  choices=['#3c539a', '#5f9647', '#b33316'],
  unit=cookieid);

button_text = weightedChoice(
  choices=['Sign up', 'Join now'],
  weights=[0.8, 0.2],
  unit=cookieid);
```

(b)

Experiment Design

button_color (Uniform)

#3c539a

#5f9647

#b33316

enter value

Add Choice

button_text (Weighted)

Sign up	0.8
Join now	0.2
enter value	enter weight

Add Choice

factor_name | Uniform Choice | Add Factor

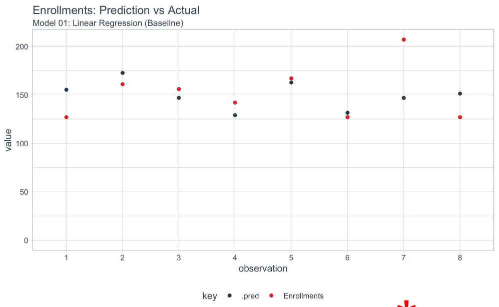
(c)

```
{"op":"seq", "seq":[
{"op":"set", "var":"button_color", "value":
  {"op":"uniformChoice",
   "unit":{"op":"get", "var":"cookieid"},
   "choices":["#3c539a", "#5f9647", "#b33316"]}},
{"op":"set", "var":"button_text", "value":
  {"op":"weightedChoice",
   "unit":{"op":"get", "var":"cookieid"},
   "weights":[0.8, 0.2],
   "choices":["Sign up", "Join now"]}
  }
]}
```

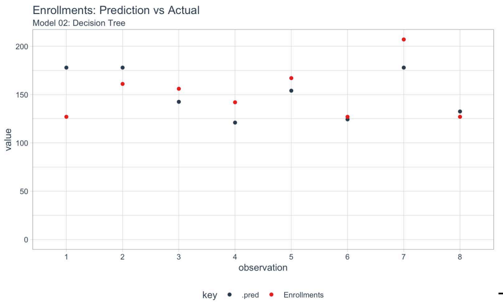
Figure 1: A factorial experiment in PlanOut. (a) PlanOut language script (b) a graphical interface for specifying simple PlanOut experiments (c) a JSON representation of serialized PlanOut code (d) an illustration of the proportion of `cookieids` allocated to each parameterization. Note that because we use `weightedChoice()` to assign `button_text`, more cookies are assigned to "Sign up" than ''Join Now".

Machine learning with A/B Testing

- Only relying on outcome from A/B testing sometimes doesn’t lead to best decision.
- Applying machine learning, better insight on user behaviour
  - Possible to achieve alternate suggestion. I.e In order to achieve A, instead of adding a button ‘X’ focus on Y.



*



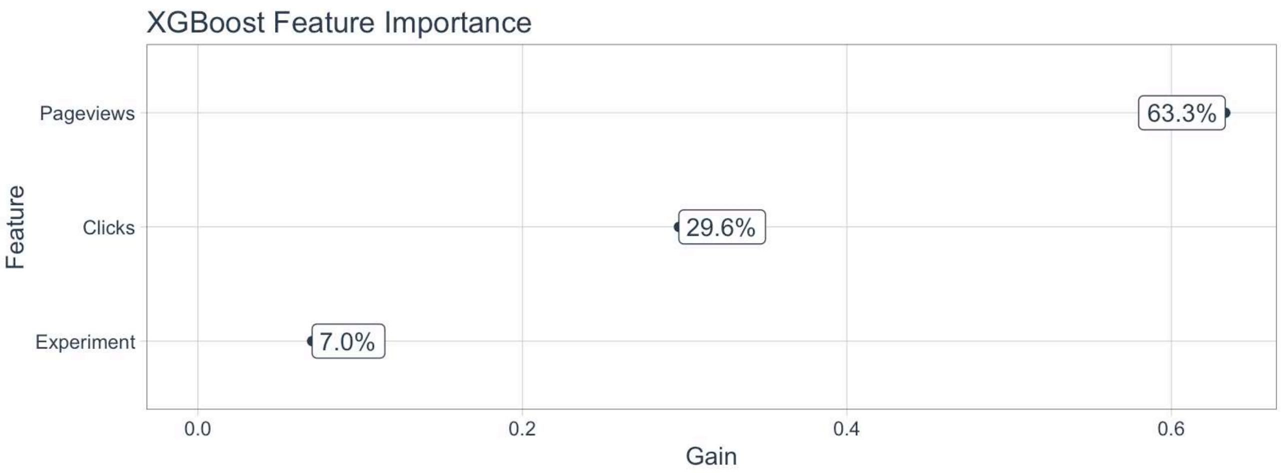
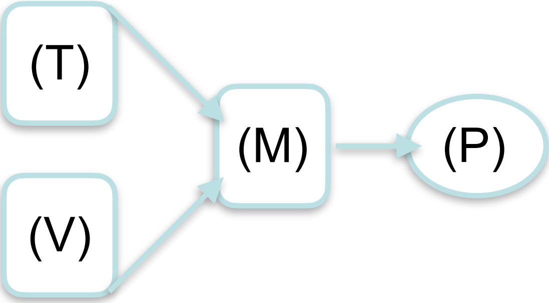


Image source

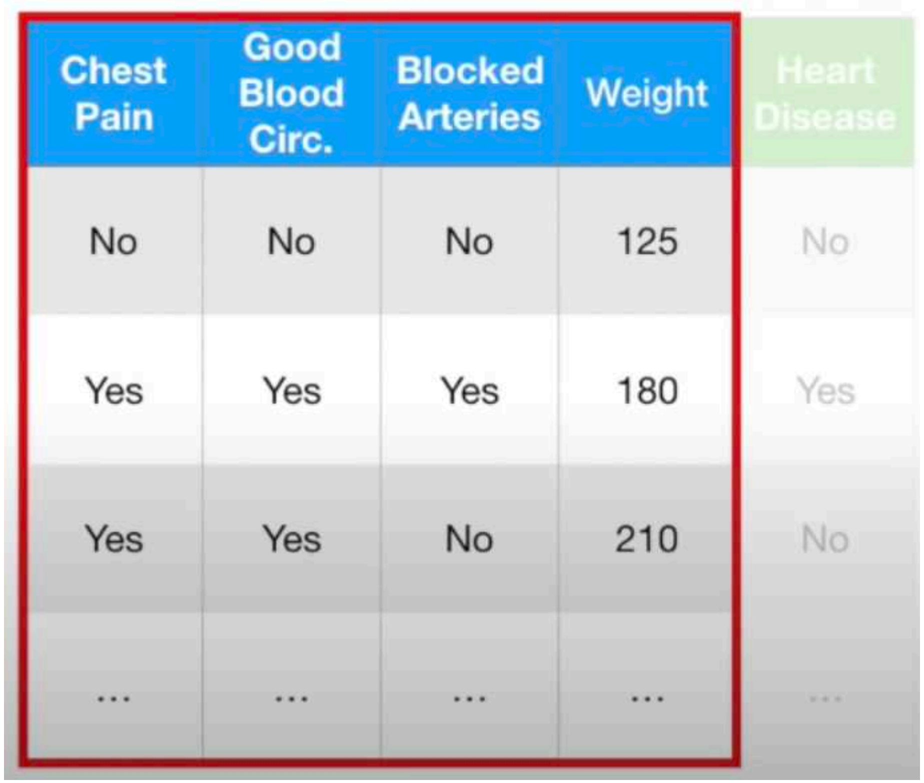
A/B Testing of ML Models

- Model (M)
  - A model is artefact(s) created (trained) by AI creation algorithm(s). Example: MS ONNX file.
- Model Predictions (Brings Output)
  - Predictions, (P) are the output of a model, (M) trained using AI algorithm(s).
- Model Deployment (Brings Outcome)
  - Means that model predictions are being consumed by an application that is directly affecting business operations.
- Predictive models are trained on historical data set (experiences), (T)
- Models are tested on holdout/validation data set (V). Presumably best performant model is deployed.
- Finding the best model post-deployment is the purpose.



The Two Variants

Imagine, we have some clinical data that helps deciding whether a patient has heart disease or not.



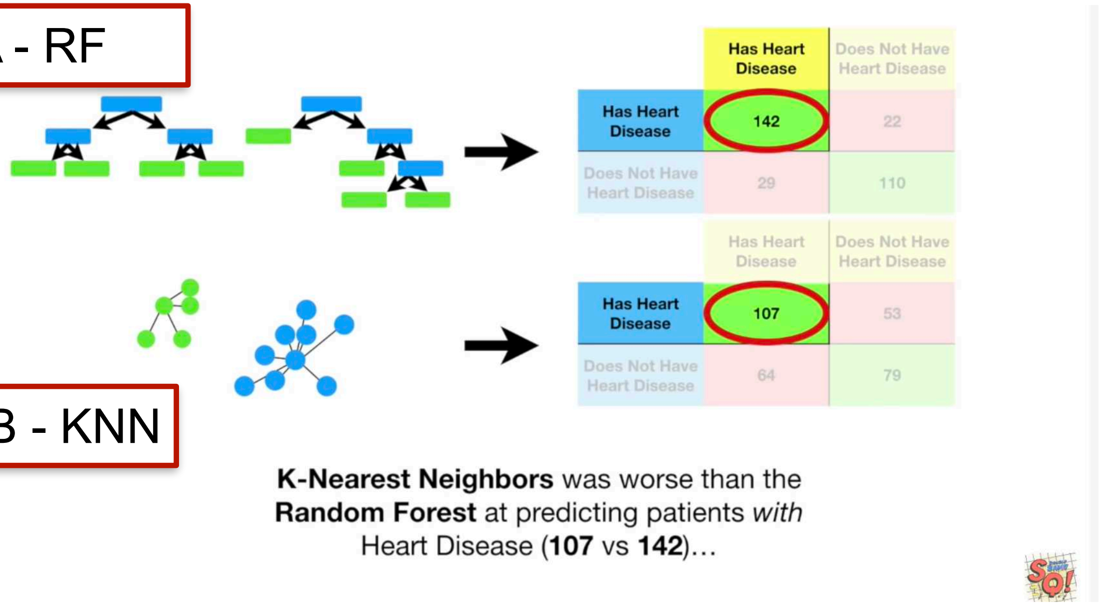
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
...	...	...	...	...

The Two Variants

We deploy Random Forest (**model A**) and K-Nearest (**model B**) and to find out. TP looks good for **model A.**

Model A - RF

Model B - KNN



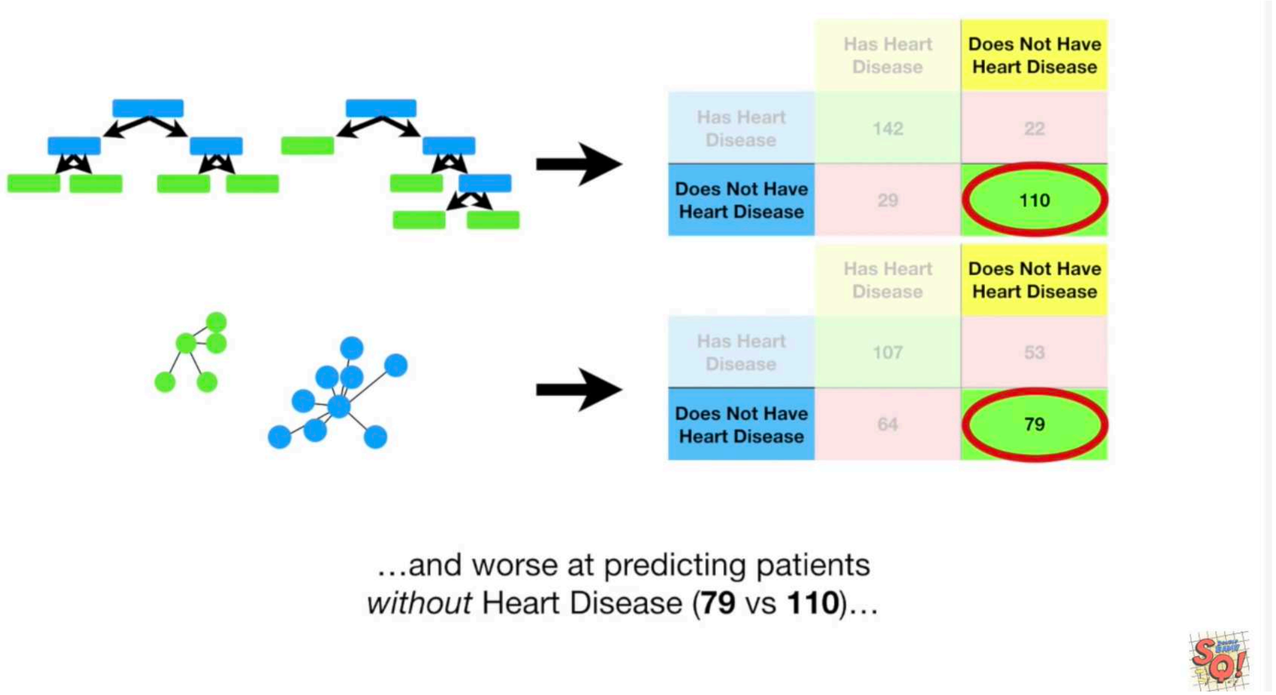
	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	142	22
Does Not Have Heart Disease	29	110

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	107	53
Does Not Have Heart Disease	64	79

K-Nearest Neighbors was worse than the **Random Forest** at predicting patients *with* Heart Disease (**107** vs **142**)…

The Two Variants

We deploy Random Forest (**model A**) and K-Nearest (**model B**) to find out.
TN also looks good for **model A.**



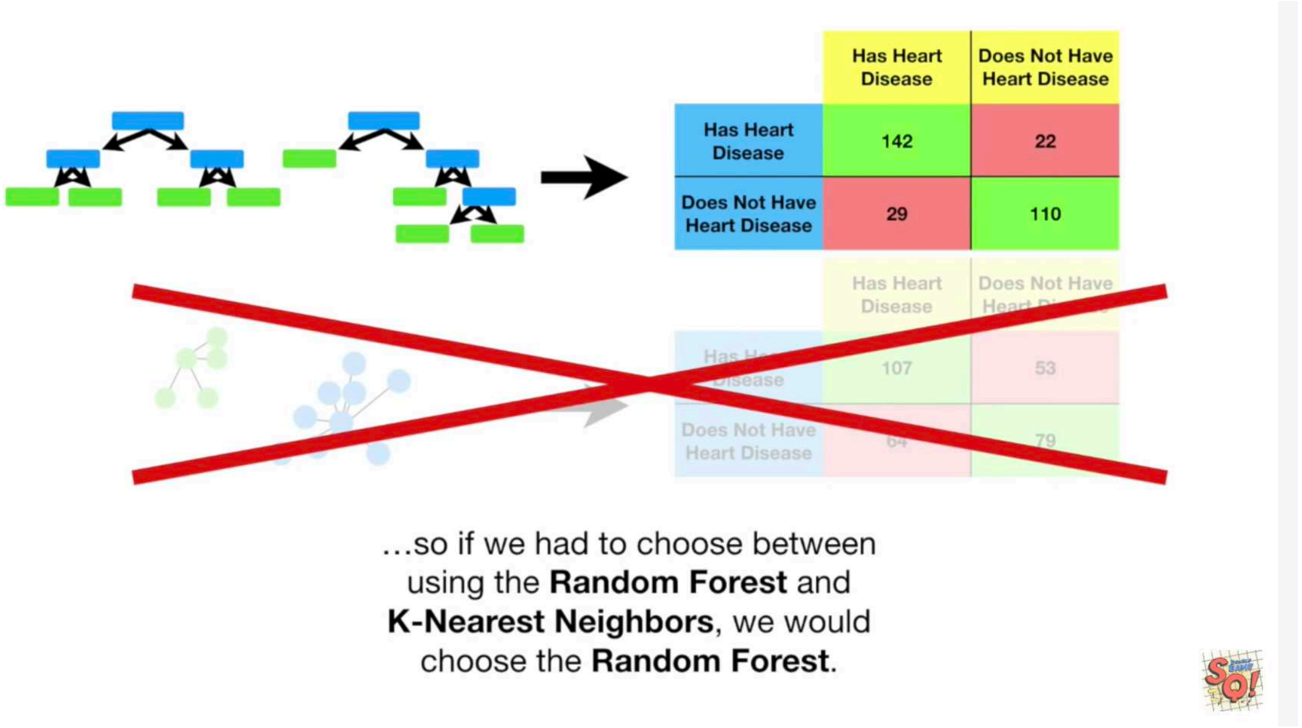
	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	142	22
Does Not Have Heart Disease	29	110

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	107	53
Does Not Have Heart Disease	64	79

…and worse at predicting patients *without* Heart Disease (**79** vs **110**)…

The Two Variants

We deploy Random Forest (**model A**) and K-Nearest (**model B**) to find out.
Model A wins!



	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	142	22
Does Not Have Heart Disease	29	110

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	107	53
Does Not Have Heart Disease	64	79

…so if we had to choose between using the **Random Forest** and **K-Nearest Neighbors**, we would choose the **Random Forest**.

Model Quantification - MQ

- **Hypothesis Test** (between models A, B to find a winner)
- **model A** (control) is deployed and predicting sth. i.e **Null Hypothesis H_0**
- **model B** (test), challenging **model A**, predicts sth. even better i.e **Alternative H_a**

$$Sensitivity = TPR = \frac{TP}{TP + FN} = \frac{TP}{Actual\ Positives} \qquad Specificity = TNR = \frac{TN}{TN + FP} = \frac{TP}{Actual\ Negatives}$$

$$ConfidenceLevel, CL = The\ probability\ of\ correctly\ retaining\ the\ H_0\ ;\quad 95\,\%$$

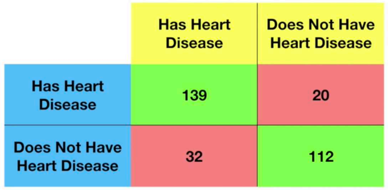
$$Statistical\ significance\ \alpha = 1 - CL$$

$$Accuracy = \frac{Total\ Correct\ Predictions}{Total\ Data\ Set}$$

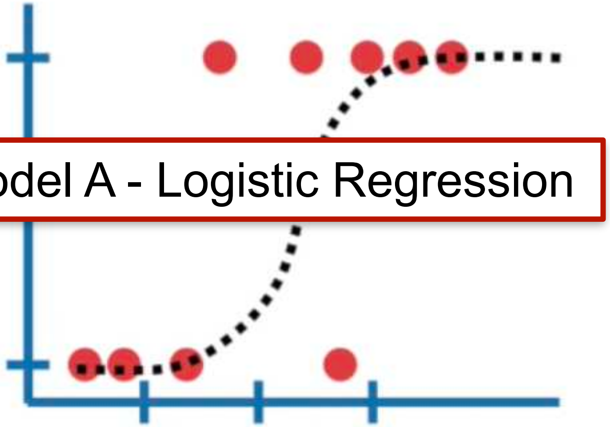
Effect Size: The difference between the two models' performance metrices.

MQ: Sensitivity & Specificity

Again we have a confusion matrix from that clinical data we saw. This time we apply LR (A) and RF (B) to measure models' performance w/ Sensitivity and Specificity.

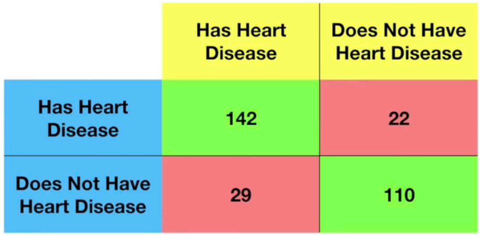


	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	139	20
Does Not Have Heart Disease	32	112



Model A - Logistic Regression

Src: StatQuest



	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	142	22
Does Not Have Heart Disease	29	110

Model B - Random Forest

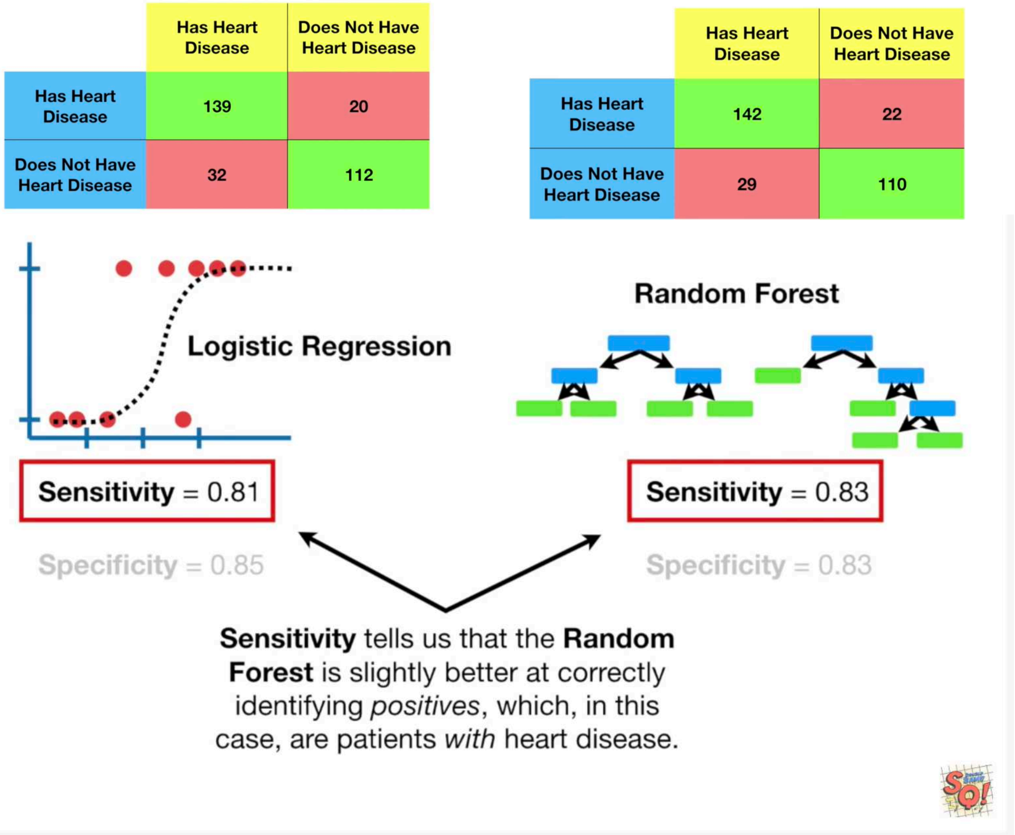
Src: StatQuest

MQ: Sensitivity & Specificity

$$Sensitivity = TPR = \frac{TP}{TP + FN} = \frac{TP}{Actual\ Positives}$$

$$Sensitivity(LR) = \frac{139}{139 + 32} = 0.81$$

$$Sensitivity(RF) = \frac{142}{142 + 29} = 0.83$$



	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	139	20
Does Not Have Heart Disease	32	112

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	142	22
Does Not Have Heart Disease	29	110

Sensitivity = 0.81

Specificity = 0.85

Sensitivity = 0.83

Specificity = 0.83

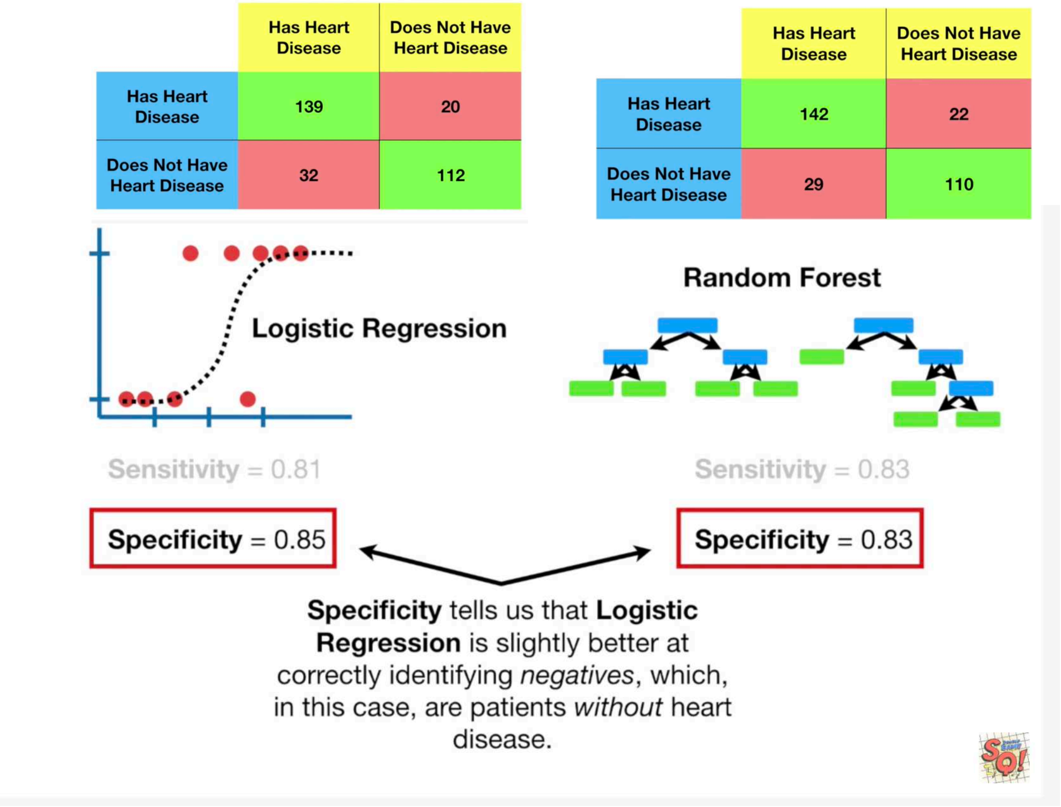
Sensitivity tells us that the **Random Forest** is slightly better at correctly identifying *positives*, which, in this case, are patients *with* heart disease.

MQ: Sensitivity & Specificity

$$Specificity = TNR = \frac{TN}{TN + FP} = \frac{TP}{Actual\ Negatives}$$

$$Specificity(LR) = \frac{112}{112 + 20} = 0.85$$

$$Specificity(RF) = \frac{110}{110 + 22} = 0.83$$



	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	139	20
Does Not Have Heart Disease	32	112

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	142	22
Does Not Have Heart Disease	29	110

Sensitivity = 0.81

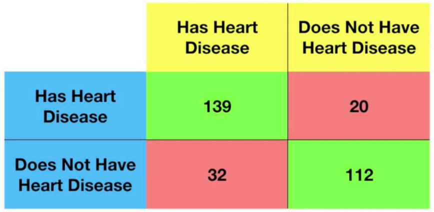
Sensitivity = 0.83

Specificity = 0.85

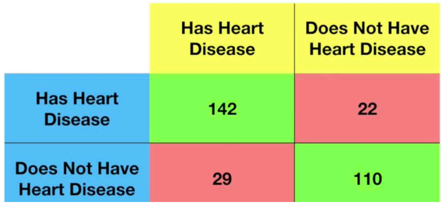
Specificity = 0.83

Specificity tells us that **Logistic Regression** is slightly better at correctly identifying *negatives*, which, in this case, are patients *without* heart disease.

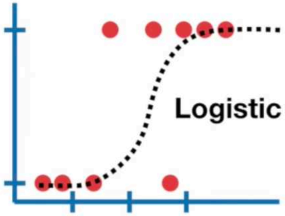
MQ: Sensitivity & Specificity



	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	139	20
Does Not Have Heart Disease	32	112



	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	142	22
Does Not Have Heart Disease	29	110



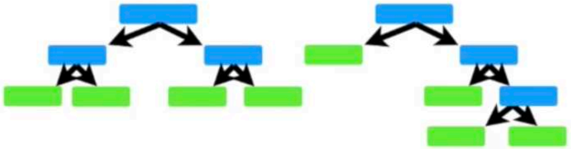
Logistic Regression

Sensitivity = 0.81

Specificity = 0.85

We would choose the **Logistic Regression** model if correctly identifying patients **without** heart disease was more important than correctly identifying patients **with** heart disease.

Random Forest



Sensitivity = 0.83

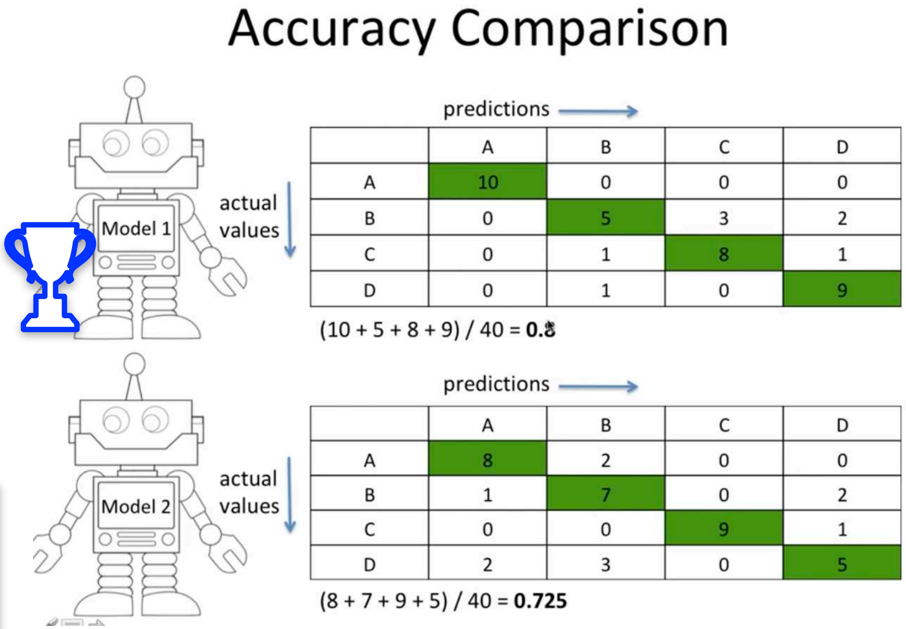
Specificity = 0.83

Alternatively, we would choose the **Random Forest** model if correctly identifying patients **with** heart disease was more important than correctly identifying patients **without** heart disease.

MQ: Accuracy

$$Accuracy = \frac{Total\ Correct\ Predictions}{Total\ Data\ Set}$$

For balanced data accuracy could alone answer for the best model.
But the reality is not always ideal!



predictions →

	A	B	C	D
A	10	0	0	0
B	0	5	3	2
C	0	1	8	1
D	0	1	0	9

(10 + 5 + 8 + 9) / 40 = **0.8**

predictions →

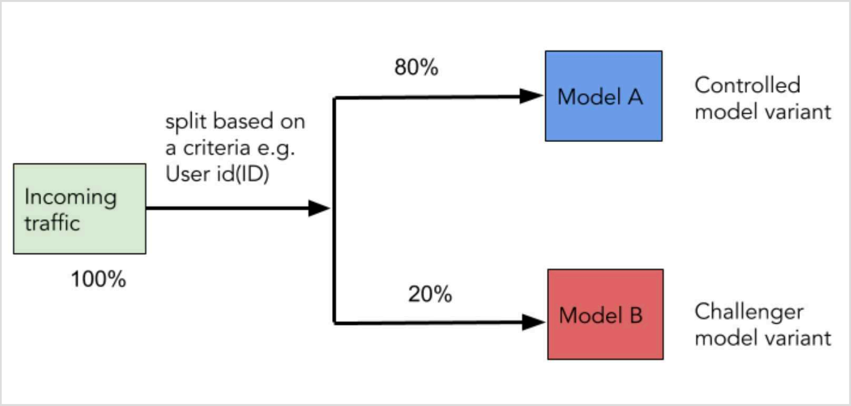
	A	B	C	D
A	8	2	0	0
B	1	7	0	2
C	0	0	9	1
D	2	3	0	5

(8 + 7 + 9 + 5) / 40 = **0.725**

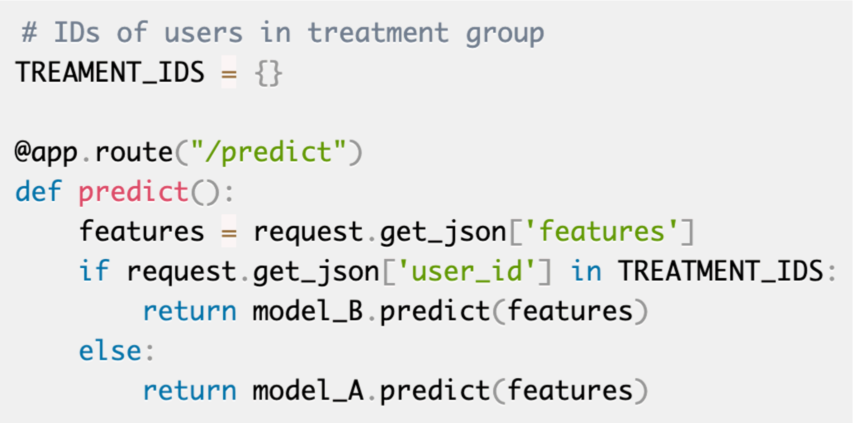
Image Courtesy : Minsuk Heo

The picture shows two models deployed to classify multiple classes (A-D). By comparing the accuracies one could decide that Model 1 wins.

Model Deployment - MD



Orcale White Paper on Model Testing



```
# IDs of users in treatment group
TREAMENT_IDS = {}

@app.route("/predict")
def predict():
    features = request.get_json['features']
    if request.get_json['user_id'] in TREATMENT_IDS:
        return model_B.predict(features)
    else:
        return model_A.predict(features)
```

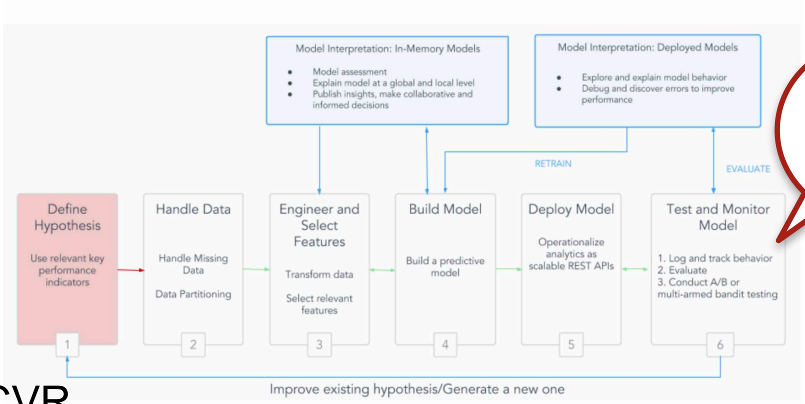
mlinproduction.com

The picture shows an A/B testing of two models. If we add more models C,D.. N in the same way the test would become a A/B/n or multivariate test.

A Trivial model deployment example using Python Flask http endpoint.

MD - Post Deployment Discrepancies

- Predictors (features) changing
  - e.g. a CTR model sees a new acquisition channel.
- Performance Metrics may differ
  - e.g. Training set was measured against
  - Balanced data —> AUC, Accuracy
  - Imbalanced data —> F1-score
  - With which do we measure the winner?
- Experiments of models may hurt UX
  - shouldn’t be the case in anyway.
- Deployed to measure a business KPI
  - e.g customer churn rate or to increase CVR.
  - But now it measures performance with AUC


Model Interpretation: In-Memory Models
• Model assessment
• Explain model at a global and local level
• Publish insights, make collaborative and informed decisions
Model Interpretation: Deployed Models
• Explore and explain model behavior
• Debug and discover errors to improve performance
RETRAIN
EVALUATE
Define Hypothesis
Use relevant key performance indicators
1
Handle Data
Handle Missing Data
Data Partitioning
2
Engineer and Select Features
Transform data
Select relevant features
3
Build Model
Build a predictive model
4
Deploy Model
Operationalize analytics as scalable REST APIs
5
Test and Monitor Model
1. Log and track behavior
2. Evaluate
3. Conduct A/B or multi-armed bandit testing
6
Improve existing hypothesis/Generate a new one

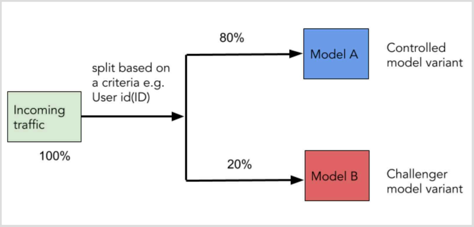
Discrepancies reveals post deployment.

Orcale White Paper on Model Testing

MD - Deploy an A/B Test

Designing a Model A/B Test

At a high level, designing an A/B test for models involves the following steps



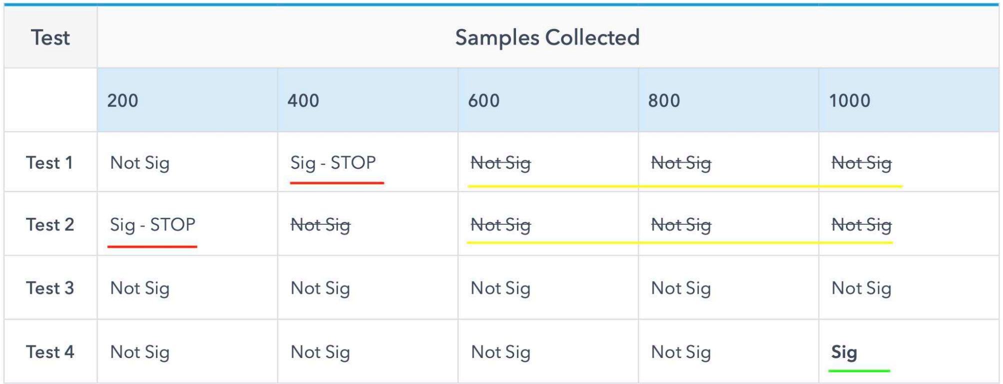
- Deciding on a performance metric. It could be the same as the one used during the model training phase (e.g., F1, AUC, RMSE, etc.)
- Deciding on test type based on your performance metric.
- Choosing a minimum effect size you want to detect.
- Determining the sample size N, based on your choice of selected minimum effect size, significance level, power, and computed/estimated sample variance.
- Running the test until N test units are collected.

Effect Size: The difference between the two models' performance metrices.

MD - Mistake of Early Declaration

It’s a mistake, don’t you pull the plug!

Declaring a model a resounding success before collecting N units of sample. The early significance could also be achieved by random chance!



Test	Samples Collected				
	200	400	600	800	1000
Test 1	Not Sig	Sig - STOP	~~Not Sig~~	~~Not Sig~~	~~Not Sig~~
Test 2	Sig - STOP	~~Not Sig~~	~~Not Sig~~	~~Not Sig~~	~~Not Sig~~
Test 3	Not Sig	Not Sig	Not Sig	Not Sig	Not Sig
Test 4	Not Sig	Not Sig	Not Sig	Not Sig	**Sig**

- If we pick a significance level α=0.05, we’d expect to see significant results (Sig) in one of 20 independent tests for a fixed and identical N (N=1000).
- If we stop as soon as significance is reached, we preferentially select spurious false positives.

MD - Holy Grails of Model A/B Testing

- Perform an A/A test. At α=0.05, a significant result should be seen 5% of the time.
- Do not turn off the test as soon as you detect an effect. Stick to your pre-calculated sample size. Often, there is a novelty effect in first few days of model deployment and a higher risk of false positives.
- Use a two-tailed test instead of a one-tailed test. (look both for H_0 and H_a)
- Control for multiple comparisons. (use Bonferroni correction, stringent α to avoid Type I / FPs.)
- Beware of cross-pollination of users between experiments. (same user does not get both a/b)
- Make sure users are identically distributed. Any segmentation (traffic source, country, etc.) should be done before randomisation is needed.
- Run tests long enough to capture variability such as day of the week seasonality.
- If possible, run the test again. See if the results still hold. B
- Beware of Simpson's paradox. (Changing experiment during intervention settings skews result. Rollout new model instead.)
- Report confidence intervals; they are different for percent change or non-linear combinations of metrics.

References

A / B Testing: The Most Powerful Way to Turn Clicks Into Customers. John Wiley & Sons. ISBN 978-1-118-65920-5.

Designing and Deploying Online Field Experiments
By Eytan Bakshy, Dean Eckles, Michael S. Bernstein

When A/B Testing Isn't Worth It

Oracle Whitepaper - Testing Predictive Models in Production
By Ruslana Dalinina Jean-René Gauthier, and Pramit Choudhary

A/B Testing Machine Learning Models (Deployment Series: Guide 08)
ML in production

Khan Academy - Unit: Significance tests (hypothesis testing)

Confusion Matrix _ StatQuest on Youtube
Sensitivity and Specificity - StatQuest on Youtube.
Statistical Significance in A/B Testing – a Complete Guide

Licence