HOW RARE IS RARE?

THE IMPORTANCE OF VALIDATION

Dr. Aric LaBarr

Institute for Advanced Analytics

HOW GOOD ARE YOUR RESULTS?

Prediction Evaluation

- Extremely important to evaluate your predictions from a model to know how good your model **actually** is.

- **Accuracy** – predicting training data well
- **Validation** – predicting new data well

Validation is Important!

- "Simple idea underpins science: 'trust, but verify'. Results should always be subject to challenge from experiment."
- "Modern scientists are doing too much trusting and not enough verifying – to the detriment of the whole of science, and of humanity."

Crisis of False Findings

- *British Medical Journal* experiment:
  - 92% of 1,500 referees missed serious errors
- *Lancet* accepts 5% of papers, but estimates half of those are worthless
- Bayer Healthcare replicated only 25% of 67 studies
- Stan Young examined controlled experiments trying to replicate 12 data “discoveries”:
  - 0 replicated; 7 neutral; 5 reversed

Is Data Science a Science? Possibly!

science

noun

/ˈsɑɪ·əns/

the **systematic study** of the structure and behavior of the natural and physical world, or **knowledge** obtained about the world by **observing** and **experimenting**

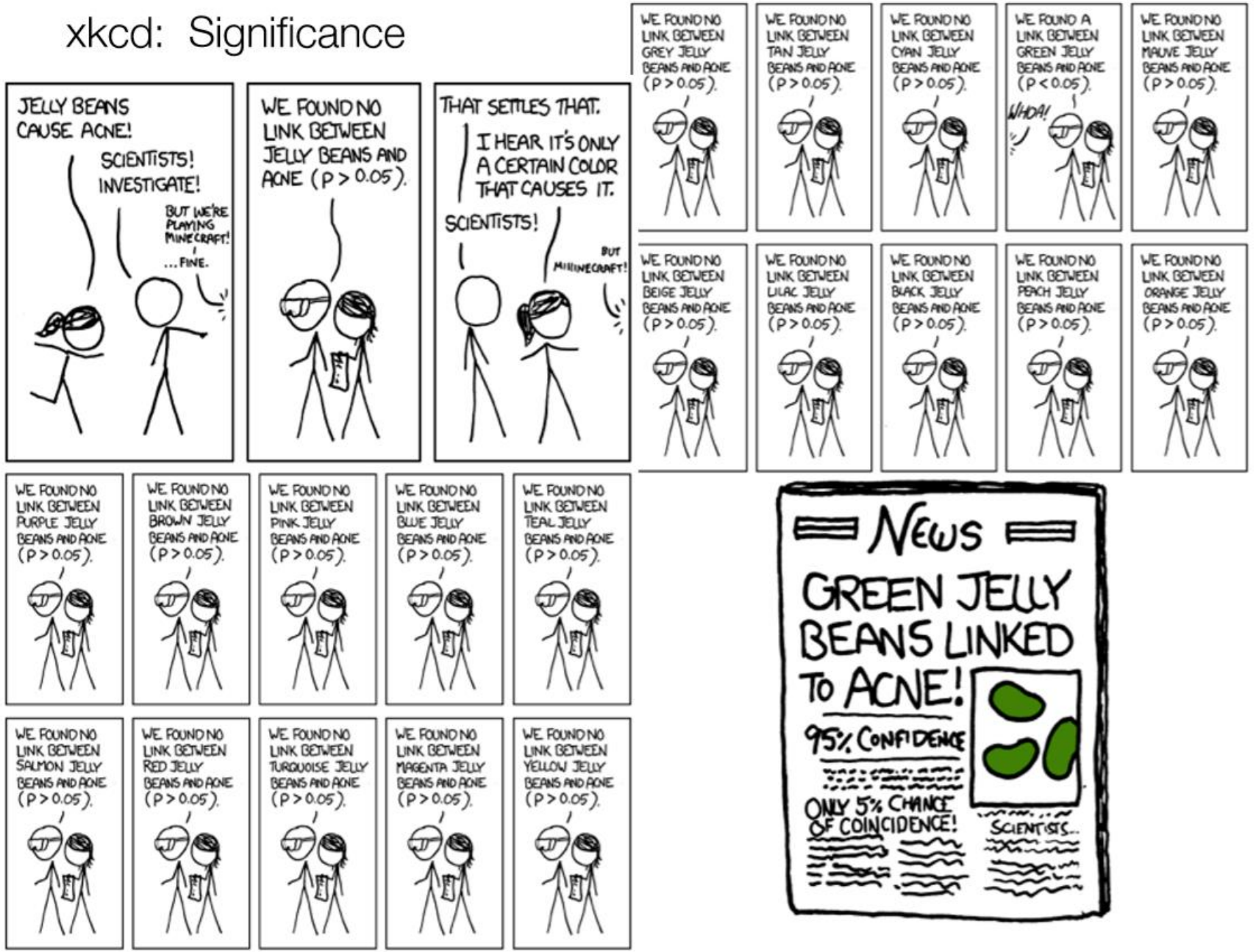
experiment

noun

/ɪkˈsper·ə·mənt/

a **test** done in order to **learn** something or to **discover** whether something **works or is true**

xkcd: Significance



Brain Naturally WANTS Patterns

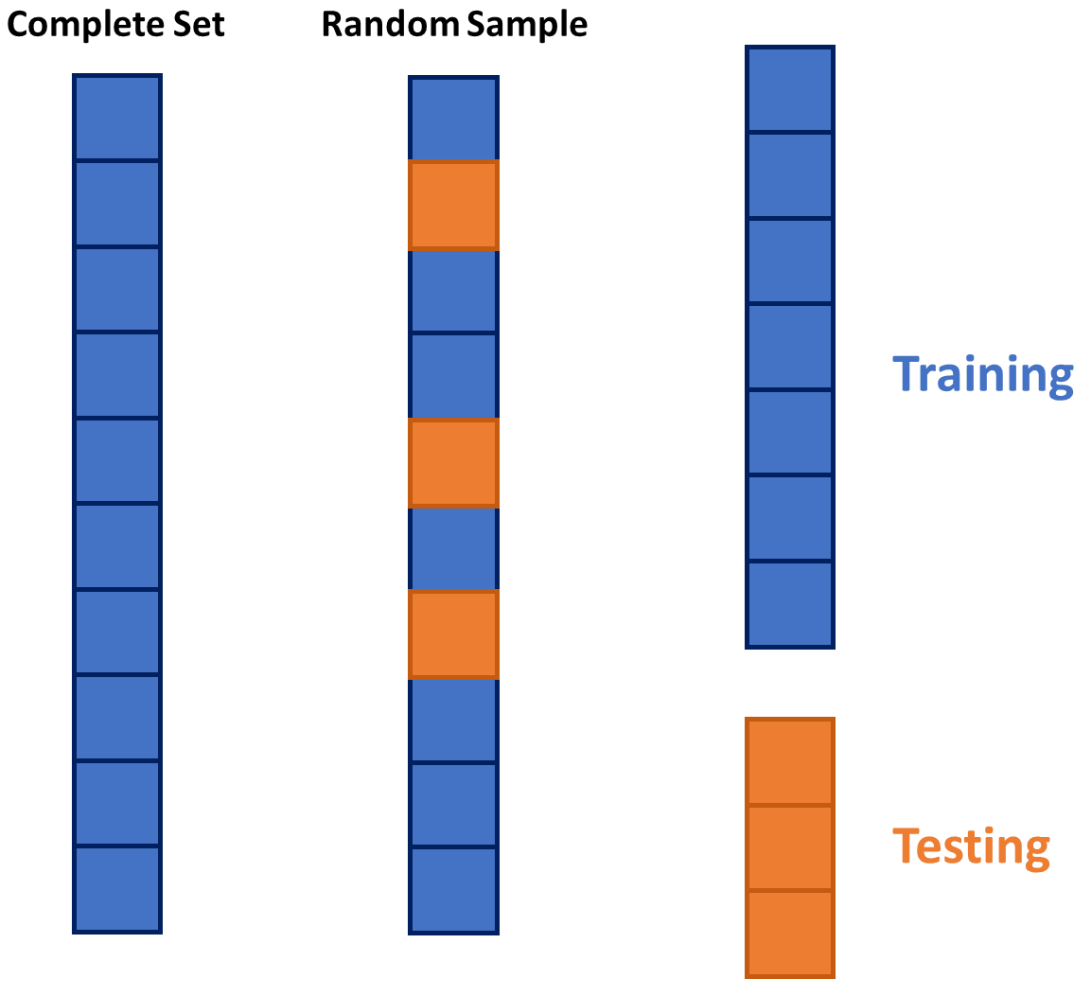
Brain Naturally WANTS Patterns

Brain Naturally WANTS Patterns

Brain Naturally WANTS Patterns

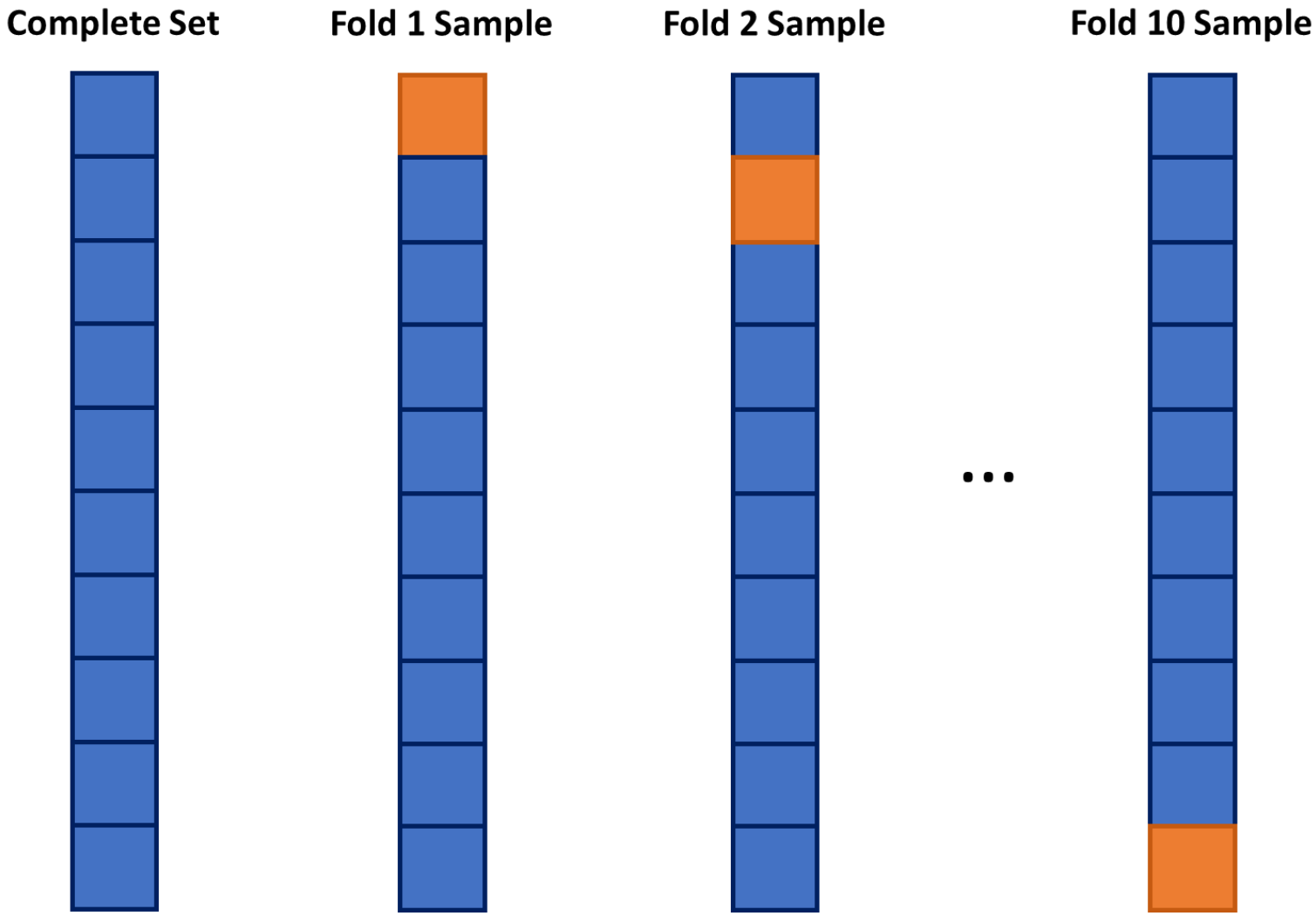
DATA SPLITTING

Train vs. Test



- Split your data into two or three sections of data
  - Training
  - Validation
  - Testing
- Common percentages:
  - 60-20-20
  - 70-20-10
  - 40-40-20
  - Etc.

Cross-Validation



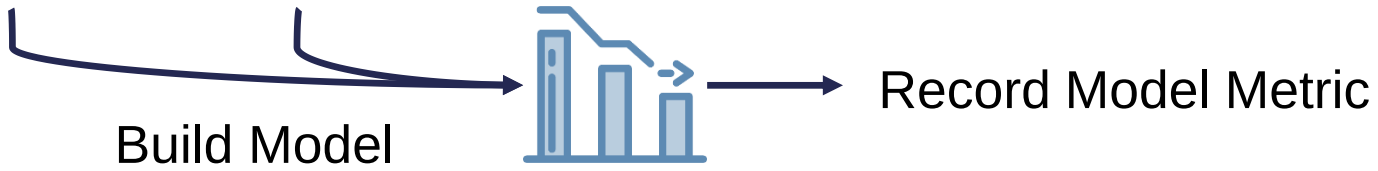
TARGET SHUFFLING

Target Shuffling

Income	Student	Default?			
$23,909	Yes	1			
$49,354	No	0			
$50,404	No	0			
$27,690	No	1			
$11,179	Yes	0			
…	…	…			
$18,475	Yes	1			

Target Shuffling

Income	Student	Default?			
$23,909	Yes	1			
$49,354	No	0			
$50,404	No	0			
$27,690	No	1			
$11,179	Yes	0			
…	…	…			
$18,475	Yes	1			



What is Target Shuffling?

1. Randomly shuffle the values of the target variable, while leaving the input variable values in the same location. This removes any possible relationship between the target variables and the inputs.

Target Shuffling

Income	Student	Default?	Y_1		
$23,909	Yes	1	0		
$49,354	No	0	1		
$50,404	No	0	1		
$27,690	No	1	0		
$11,179	Yes	0	0		
…	…	…	…		
$18,475	Yes	1	1		

What is Target Shuffling?

1. Randomly shuffle the values of the target variable, while leaving the input variable values in the same location. This removes any possible relationship between the target variables and the inputs.
2. Repeat the model building process (preferably in an automated way) to identify any possible relationships between the input variables and the newly shuffled target variable.

Target Shuffling

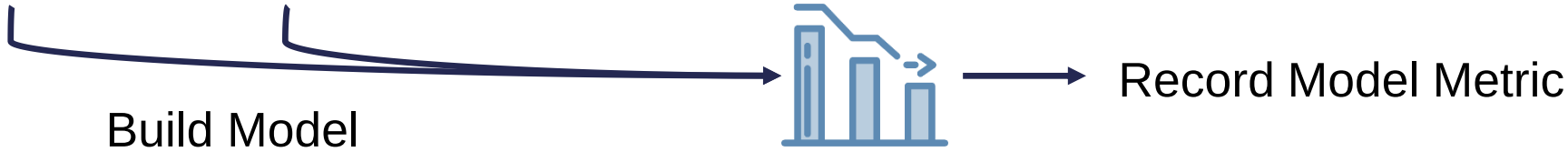
Income	Student	Default?	Y_1		
$23,909	Yes	1	0		
$49,354	No	0	1		
$50,404	No	0	1		
$27,690	No	1	0		
$11,179	Yes	0	0		
…	…	…	…		
$18,475	Yes	1	1		

What is Target Shuffling?

1. Randomly shuffle the values of the target variable, while leaving the input variable values in the same location. This removes any possible relationship between the target variables and the inputs.
2. Repeat the model building process (preferably in an automated way) to identify any possible relationships between the input variables and the newly shuffled target variable.
3. Save the “best” model’s measure of validity – RMSE, MAPE, AIC, etc.

Target Shuffling

Income	Student	Default?	Y_1		
$23,909	Yes	1	0		
$49,354	No	0	1		
$50,404	No	0	1		
$27,690	No	1	0		
$11,179	Yes	0	0		
…	…	…	…		
$18,475	Yes	1	1		



What is Target Shuffling?

1. Randomly shuffle the values of the target variable, while leaving the input variable values in the same location. This removes any possible relationship between the target variables and the inputs.
2. Repeat the model building process (preferably in an automated way) to identify any possible relationships between the input variables and the newly shuffled target variable.
3. Save the “best” model’s measure of validity – RMSE, MAPE, AIC, etc.
4. Repeat the process thousands of times.

Target Shuffling

Income	Student	Default?	Y_1	Y_2	
$23,909	Yes	1	0	1	
$49,354	No	0	1	1	
$50,404	No	0	1	1	
$27,690	No	1	0	0	
$11,179	Yes	0	0	0	
…	…	…	…	…	
$18,475	Yes	1	1	0	

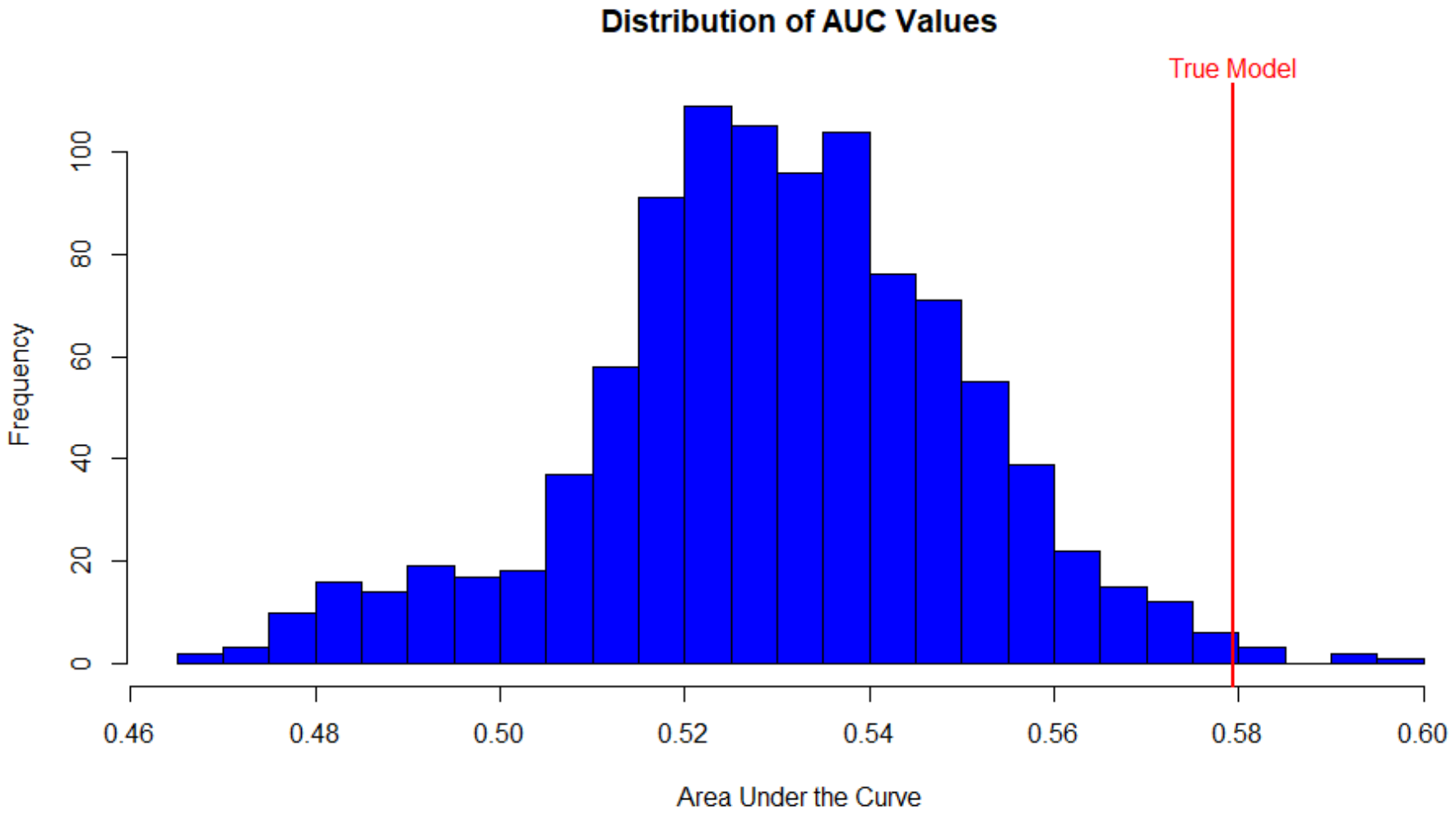
Target Shuffling

Income	Student	Default?	Y_1	Y_2	…
$23,909	Yes	1	0	1	…
$49,354	No	0	1	1	…
$50,404	No	0	1	1	…
$27,690	No	1	0	0	…
$11,179	Yes	0	0	0	…
…	…	…	…	…	…
$18,475	Yes	1	1	0	…

What is Target Shuffling?

1. Randomly shuffle the values of the target variable, while leaving the input variable values in the same location. This removes any possible relationship between the target variables and the inputs.
2. Repeat the model building process (preferably in an automated way) to identify any possible relationships between the input variables and the newly shuffled target variable.
3. Save the “best” model’s measure of validity – RMSE, MAPE, AIC, etc.
4. Repeat the process thousands of times.
5. Look at the distribution of the collection of validity measures from each iteration.

How Rare is Rare?



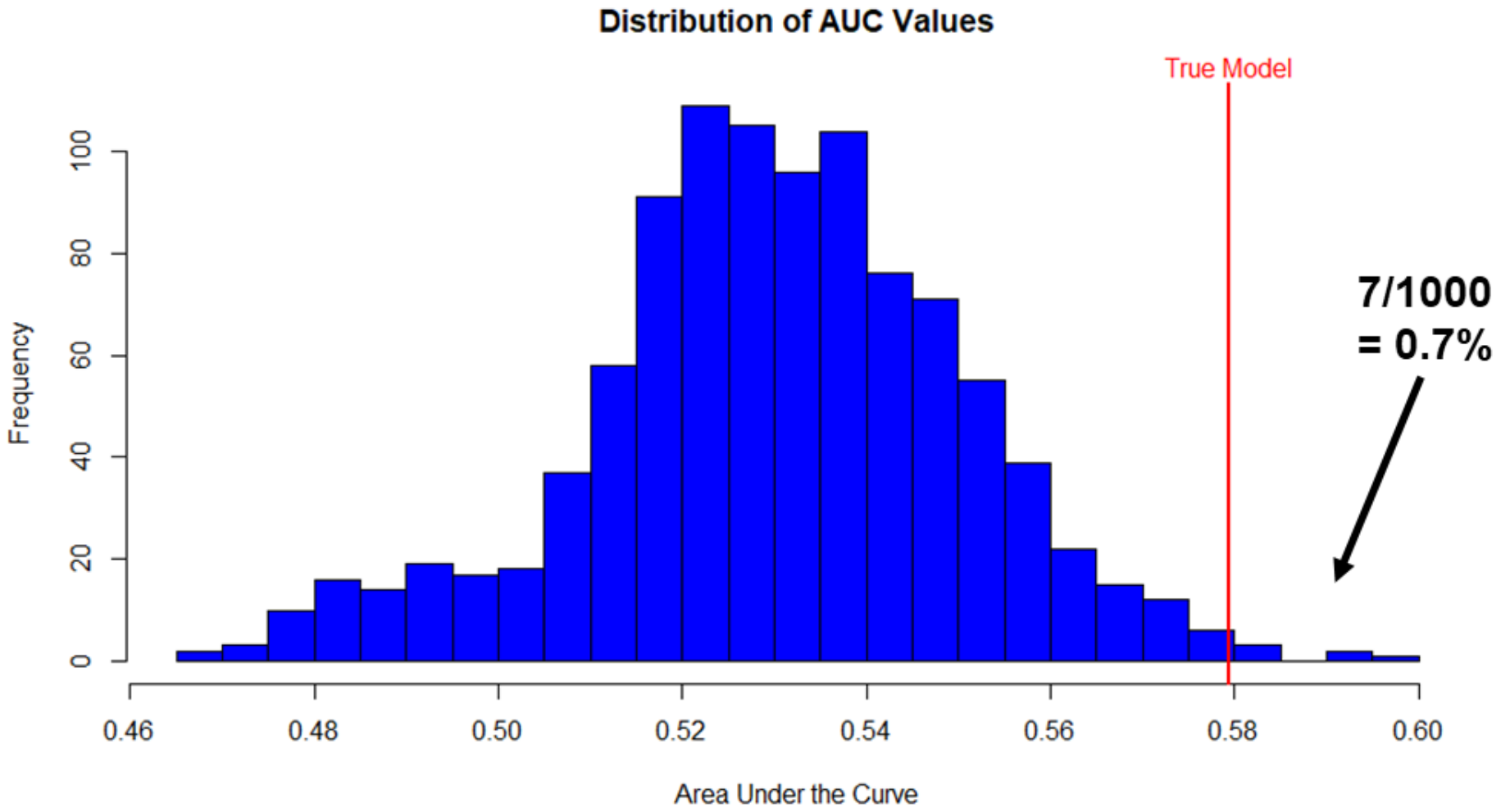
What is Target Shuffling?

1. Randomly shuffle the values of the target variable, while leaving the input variable values in the same location. This removes any possible relationship between the target variables and the inputs.
2. Repeat the model building process (preferably in an automated way) to identify any possible relationships between the input variables and the newly shuffled target variable.
3. Save the “best” model’s measure of validity – RMSE, MAPE, AIC, etc.
4. Repeat the process thousands of times.
5. Look at the distribution of the collection of validity measures from each iteration.
6. Evaluate where your original model’s validity measure falls on this distribution of validity measures.

Placebo Effect

- This should remove the pattern from the data, but some pattern may exist due to randomness.
- Look at distribution of all measurements of model success and find your value from the true model!
- What is probability your model would have occurred due to randomness?

How Rare is Rare?

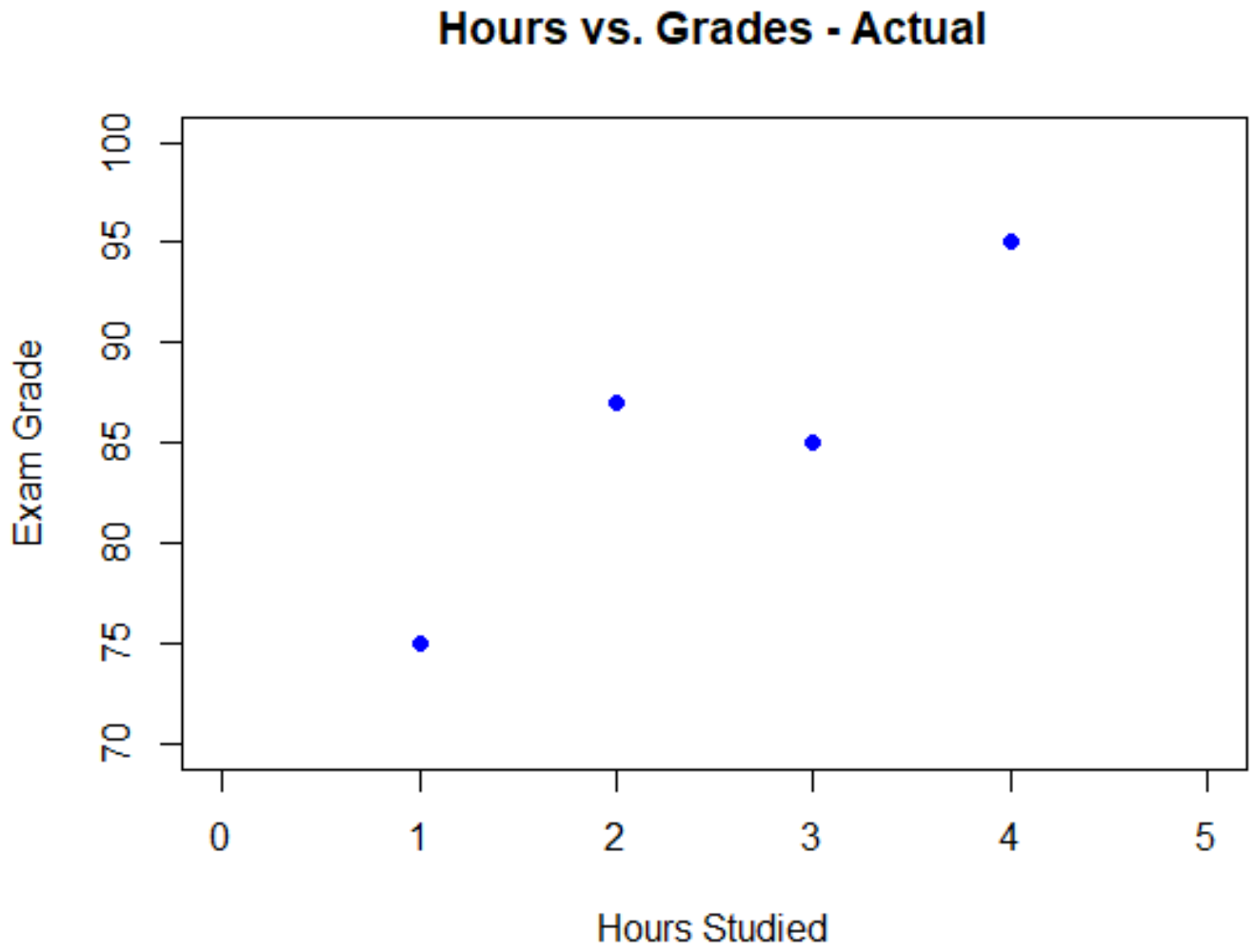


TARGET SHUFFLING

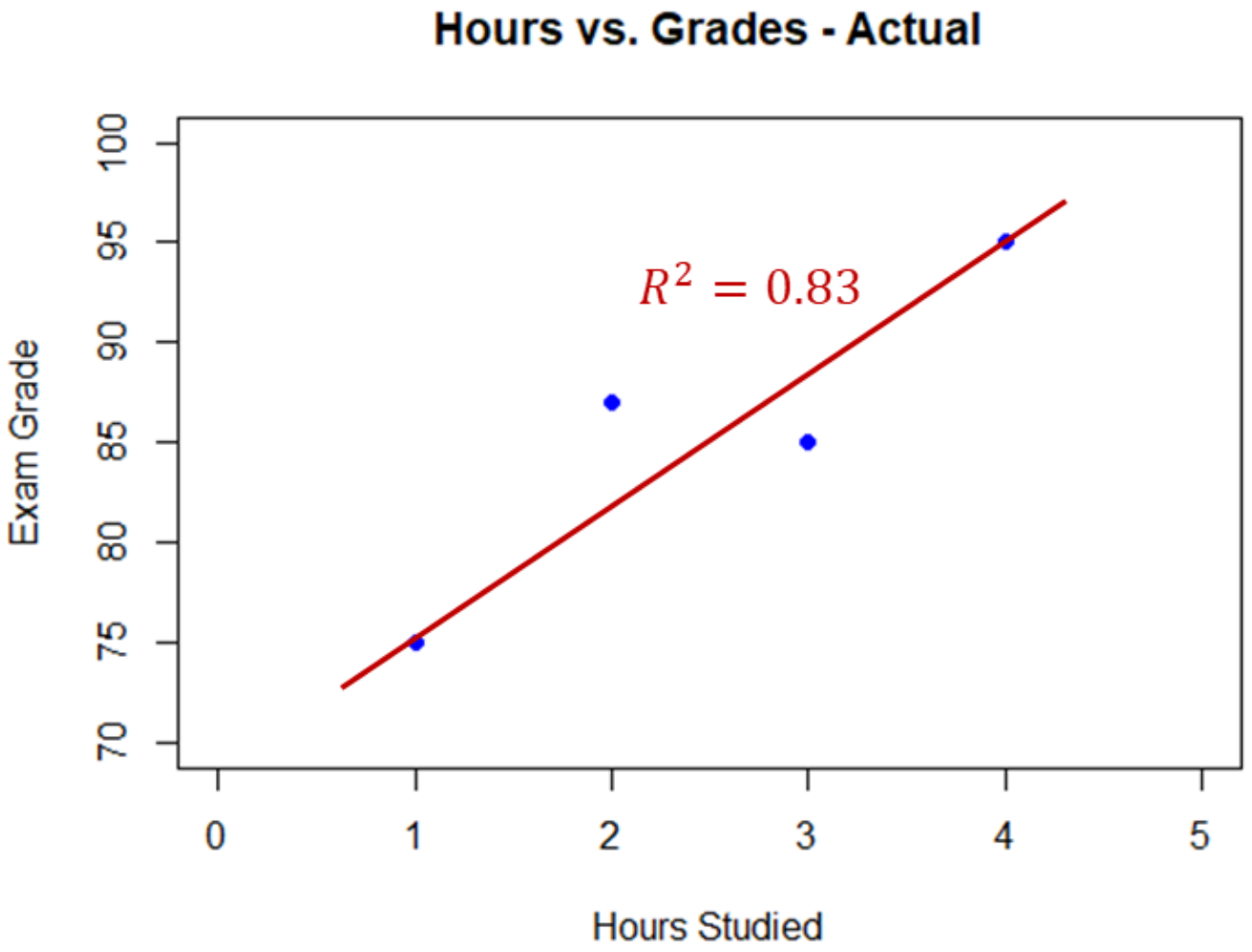
FURTHER EXAMPLES

Student Grade Analogy

Student Grade Analogy



Student Grade Analogy



Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

1	2	3	4
75	85	87	95

1	2	3	4
75	95	85	87

1	2	3	4
85	87	75	95

1	2	3	4
87	75	85	95

1	2	3	4
87	95	75	85

1	2	3	4
95	85	75	87

1	2	3	4
75	87	85	95

1	2	3	4
75	95	87	85

1	2	3	4
85	87	95	75

1	2	3	4
87	75	95	85

1	2	3	4
87	95	75	85

1	2	3	4
95	87	75	85

1	2	3	4
75	85	95	87

1	2	3	4
85	75	87	95

1	2	3	4
85	95	75	87

1	2	3	4
87	85	75	95

1	2	3	4
95	75	85	87

1	2	3	4
95	85	87	75

1	2	3	4
75	87	95	85

1	2	3	4
85	75	95	87

1	2	3	4
85	95	87	75

1	2	3	4
87	85	95	75

1	2	3	4
95	75	87	85

1	2	3	4
95	87	85	75

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!
- There are 3 possible combinations that produce a regression with an R^2 that is greater than or equal to our actual data.

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

1	2	3	4
75	85	87	95

1	2	3	4
75	95	85	87

1	2	3	4
85	87	75	95

1	2	3	4
87	75	85	95

1	2	3	4
87	95	75	85

1	2	3	4
95	85	75	87

1	2	3	4
75	87	85	95

1	2	3	4
75	95	87	85

1	2	3	4
85	87	95	75

1	2	3	4
87	75	95	85

1	2	3	4
87	95	75	85

1	2	3	4
95	87	75	85

1	2	3	4
75	85	95	87

1	2	3	4
85	75	87	95

1	2	3	4
85	95	75	87

1	2	3	4
87	85	75	95

1	2	3	4
95	75	85	87

1	2	3	4
95	85	87	75

1	2	3	4
75	87	95	85

1	2	3	4
85	75	95	87

1	2	3	4
85	95	87	75

1	2	3	4
87	85	95	75

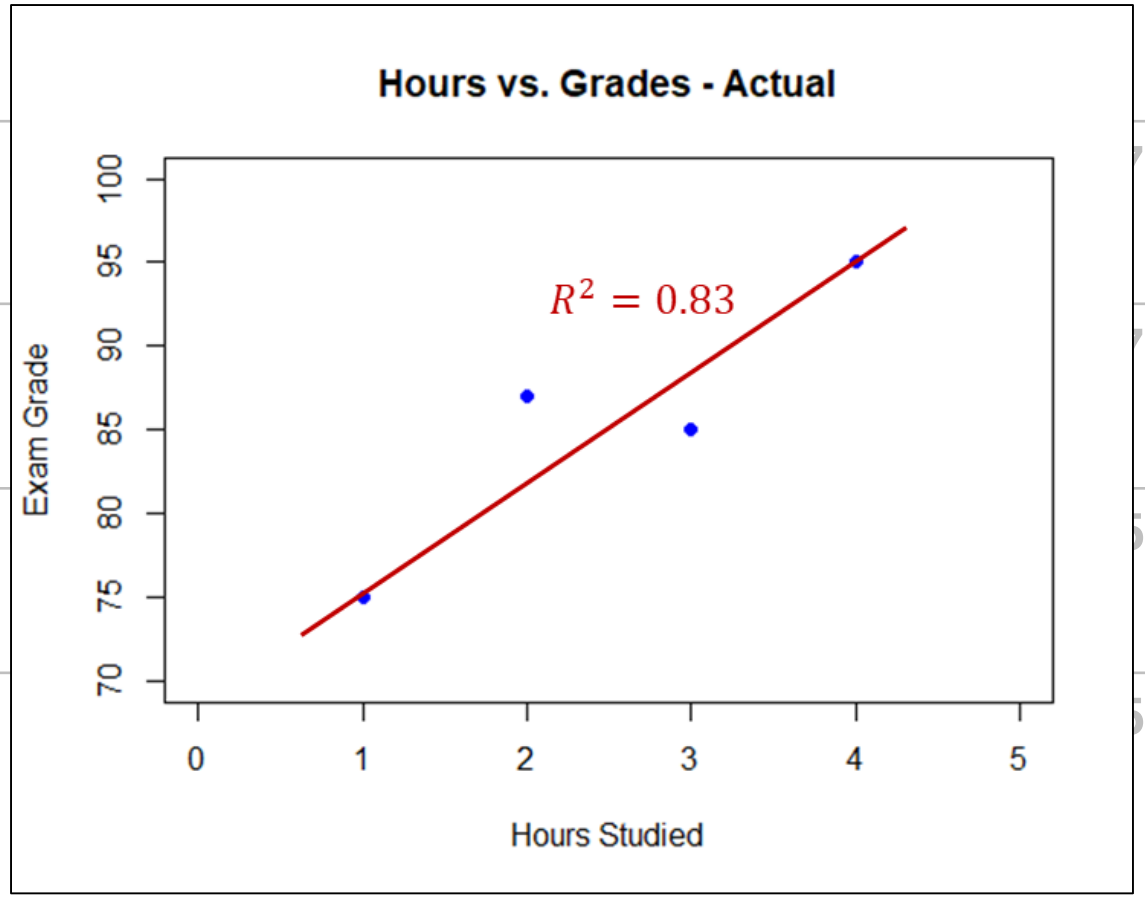
1	2	3	4
95	75	87	85

1	2	3	4
95	87	85	75

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

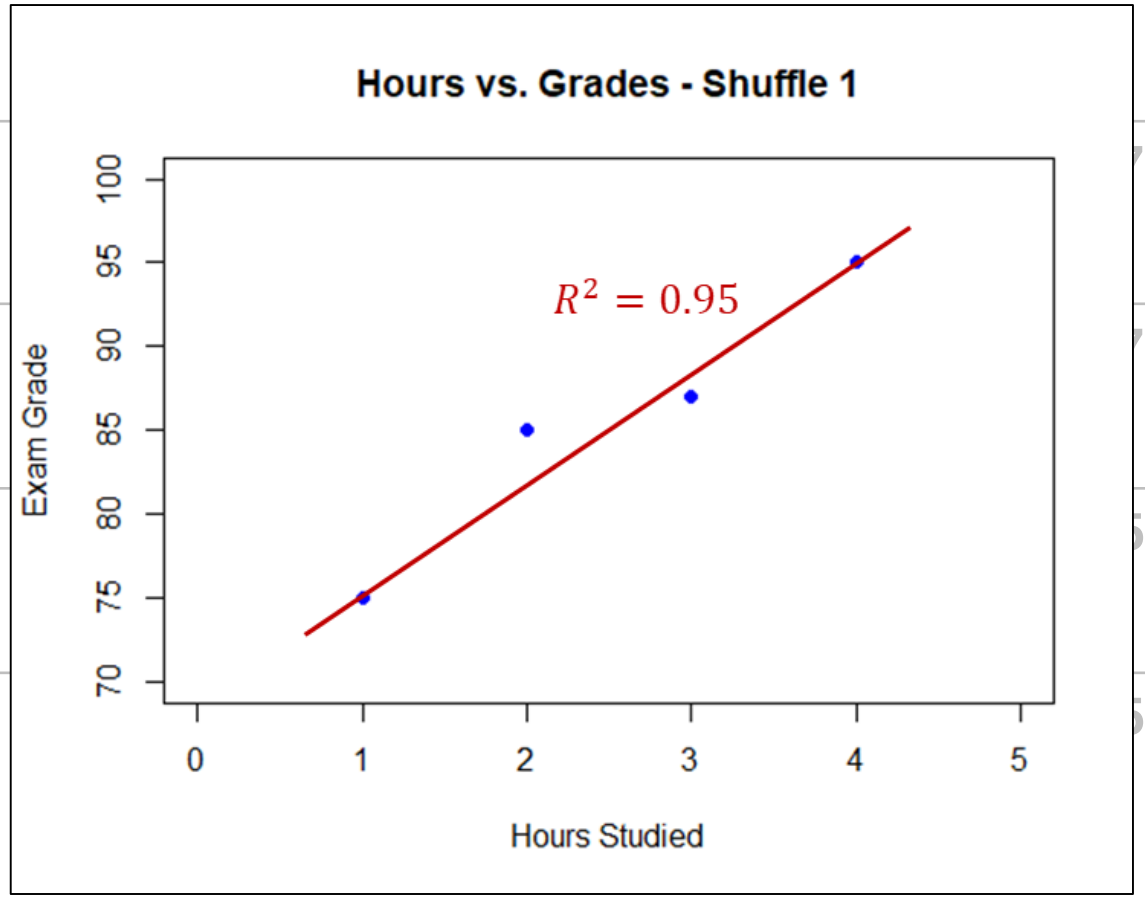
1	2	3	4
75	85	87	95
1	**2**	**3**	**4**
75	**87**	**85**	**95**
1	2	3	4
75	85	95	87
1	2	3	4
75	87	95	85



Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

1	2	3	4
75	85	87	95
1	2	3	4
75	87	85	95
1	2	3	4
75	85	95	87
1	2	3	4
75	87	95	85



Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

1	2	3	4
75	85	87	95

1	2	3	4
75	87	85	95

1	2	3	4
75	85	95	87

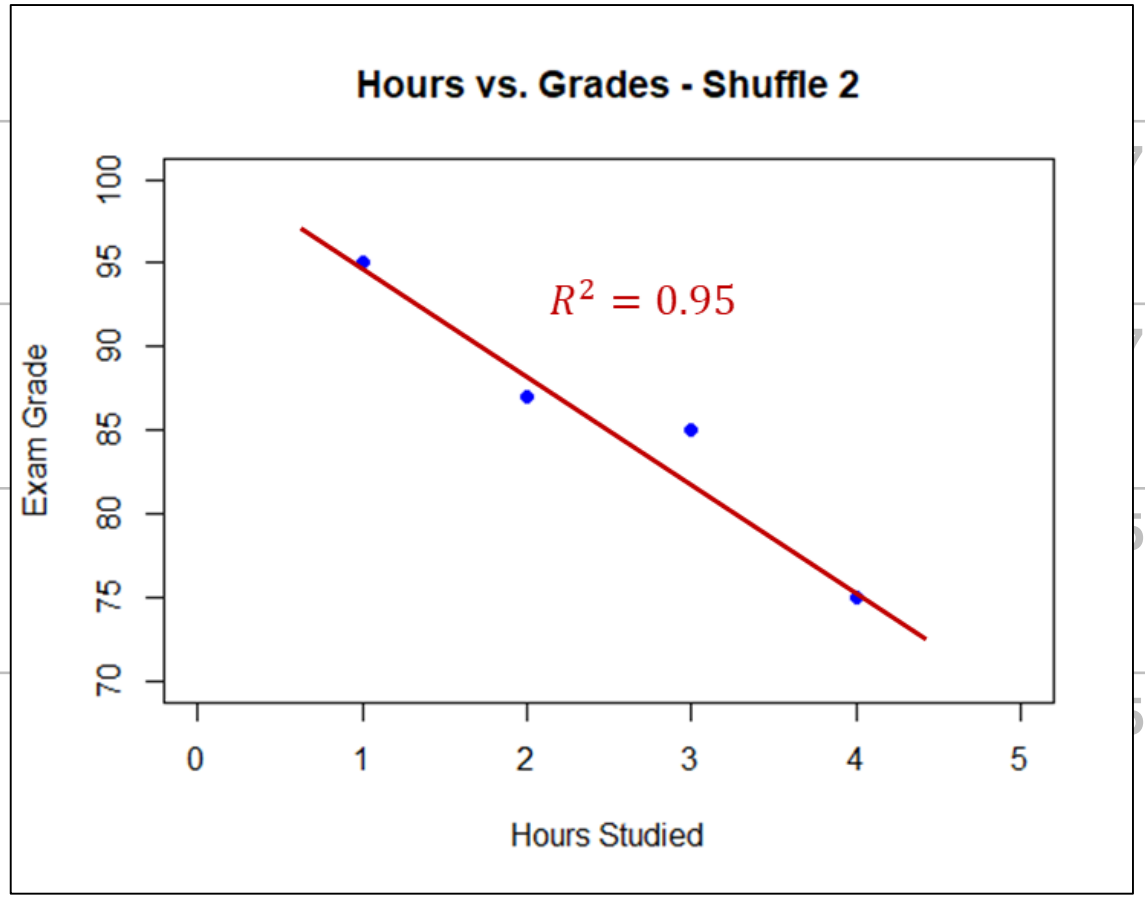
1	2	3	4
75	87	95	85

1	2
75	95

1	2
75	95

1	2
85	75

1	2
85	75



2	3	4
95	75	85

2	3	4
95	75	85

2	3	4
75	85	87

2	3	4
75	87	85

1	2	3	4
95	85	75	87

1	2	3	4
95	87	75	85

1	2	3	4
95	85	87	75

1	2	3	4
95	**87**	**85**	**75**

Fake Data Example

- Randomly generated 8 variables that follow a Normal distribution with mean of 0 and standard deviation of 8.
- Defined relationship with target variable:

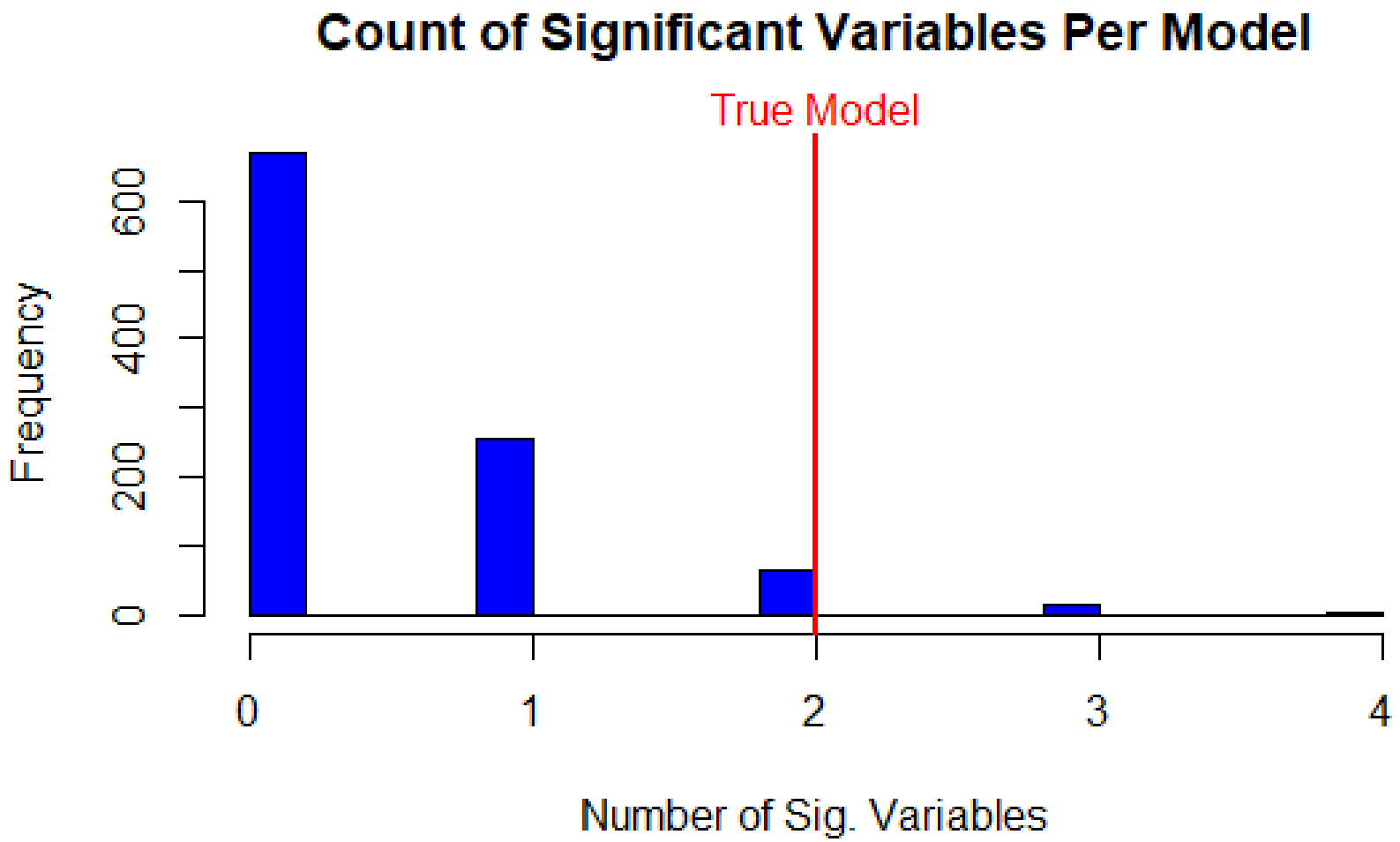
$$y = 5 + 2x_2 - 3x_8 + \varepsilon$$

- Performed target shuffle on the model.

Target Shuffle with 1000 Simulations

Variable	Times Appeared Significant ($p < 0.05$) in a Model
X1	55
X2	62
X3	47
X4	56
X5	50
X6	57
X7	58
X8	40

Target Shuffle with 1000 Simulations



CONCLUSION

Summary

- Validating models is not only important but expected.
- To accurately explain how your model performs in deployment, validation is needed.

- Target shuffling
  - What is the probability you got the model that you did?
  - Essentially, building placebo models to compare against.

THANK YOU!

Dr. Aric LaBarr

www.ariclabarr.com