The Build Fellowship

BUILDFELLOWSHIP.COM

OpenAvenues

Weekly Updates

- Please provide a quick update on either:
  - Something you did/saw this week that you thought was interesting
  - What you're looking forward to about this week's workshop

(Reminder - please have your cameras on if possible)

The Build Fellowship

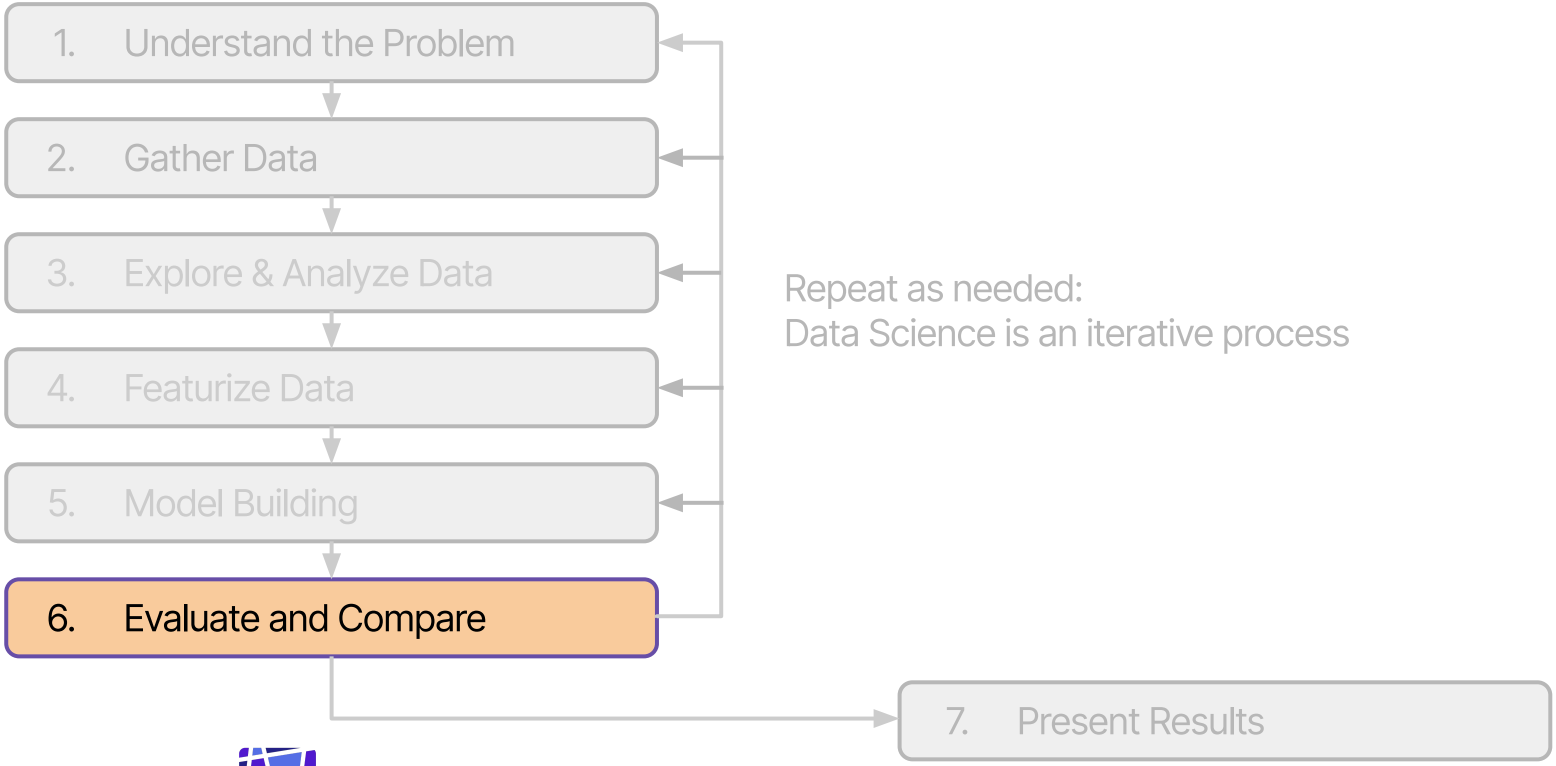
Workshop 7
Performance Evaluation

Recap

Sessions Overview

- Workshop 1 – Project Introduction & Setup
- Workshop 2 – Genomic Data (A2 Assignment)
- Workshop 3 – Data Analysis & Visualization (A3 Assignment)
- Workshop 4 – Featurization & Baseline Modeling (A4 Assignment)
- Workshop 5 – Model Training Approaches (Final Assignment Set)
- Workshop 6 – Model Tuning
- **Workshop 7 – Performance Evaluation (Final Assignment Code/Testing Due)**
- Workshop 8 – Results Presentation & Wrap up (Final Presentation Due)

The Data Science Process



1. Understand the Problem
2. Gather Data
3. Explore & Analyze Data
4. Featurize Data
5. Model Building
6. Evaluate and Compare
7. Present Results

Repeat as needed:
Data Science is an iterative process

Model Tuning & Comparison

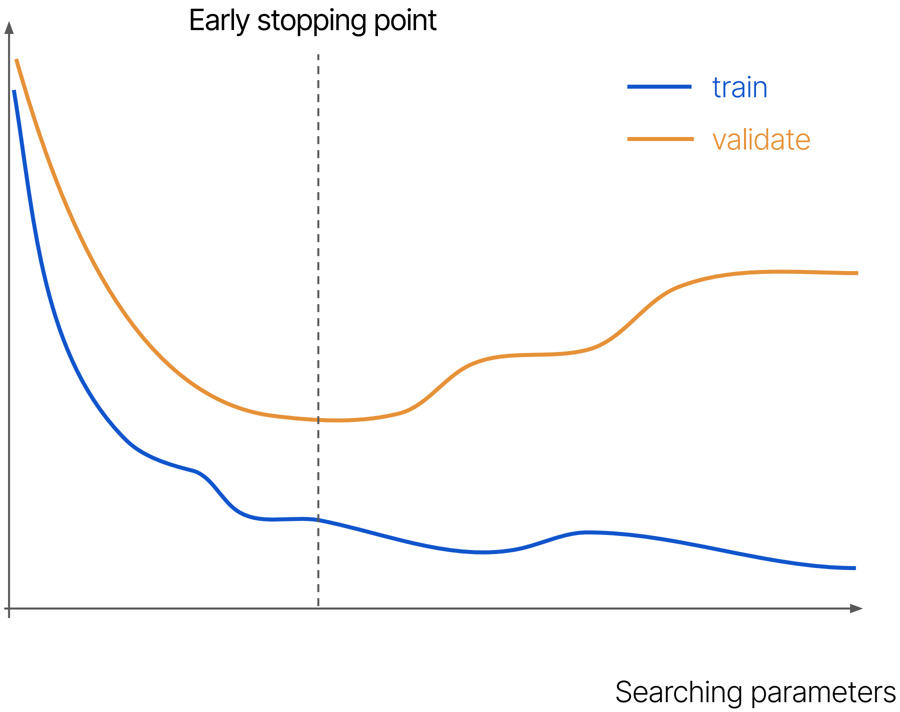
Last week we took a single Random Forest model and reviewed different methods for optimizing performance

Methods:

- Grid search
- Random search
- Bayesian Optimization

Nested Cross Validation

- Method for parameter selection & model comparison
- Higher complexity but fairer review across multiple held out datasets



The Build Fellowship

Evaluation Metrics

The Build Fellowship

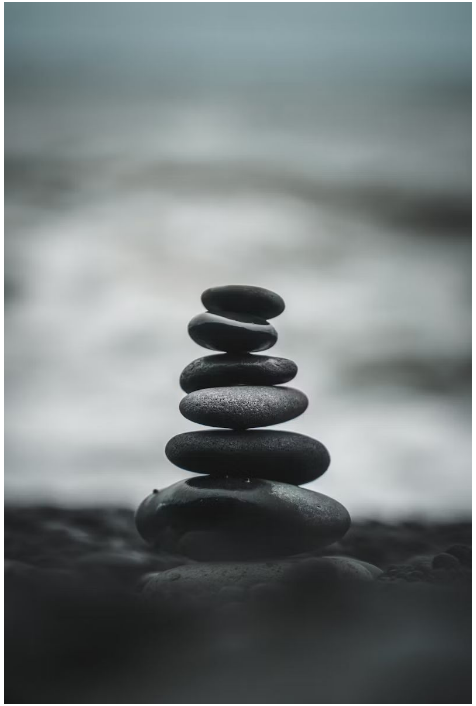
OpenAvenues FOUNDATION

What is our Definition of Good?

Model Loss/Accuracy are the most common assessments
- Loss is a continuous measurement of model fit
- What do we care about?
- Categorial targets
- Binary S vs R

Why not just accuracy?
- Remember our baseline models
- Achieved ~80% accuracy with majority predictor

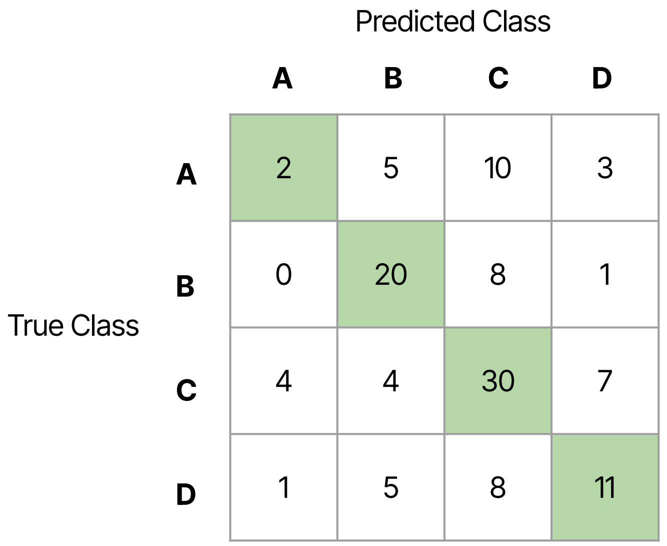
What have we used thus far?
- Balanced accuracy
- What does it actually mean and measure
- Why is it better than standard accuracy?



Confusion - let's use it

Categorical predictions

- Per class performance
- **Confusion Matrix** is an natural visual representation
- Rows = True class
- Columns = Predicted class

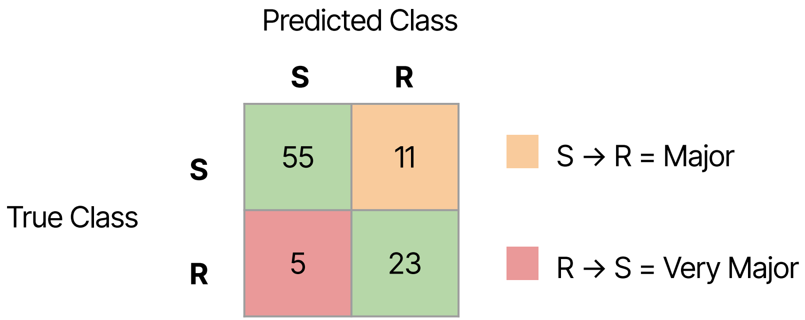


		Predicted Class			
		A	**B**	**C**	**D**
True Class	**A**	2	5	10	3
	B	0	20	8	1
	C	4	4	30	7
	D	1	5	8	11

For AMR Predictions:

- **Major** errors
- **Very** major errors

Very major errors are a worse outcome for patients



		Predicted Class	
		S	**R**
True Class	**S**	55	11
	R	5	23

🟧 S → R = Major

🟥 R → S = Very Major

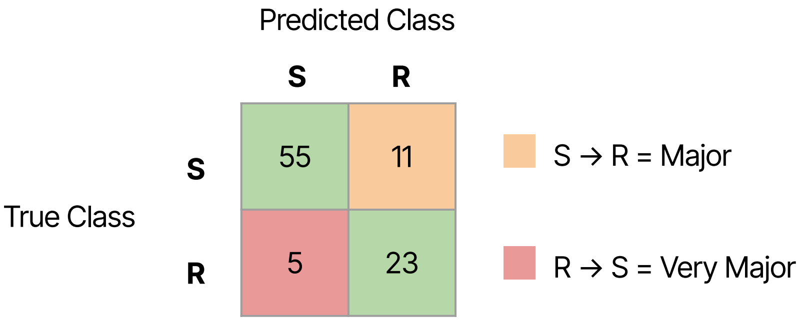
Binary Assessments

For AMR Predictions:
- **Major** errors
- **Very** major errors

In general we can think of this as Positive & Negative

- Positive Predictions = R
- Negative Prediction = S

Common Metrics:
- Sensitivity
- Specificity
- **Balanced accuracy** = average of the above

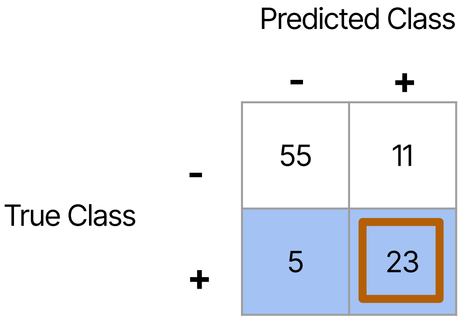


Predicted Class

True Class		S	R
	S	55	11
	R	5	23

🟧 S → R = Major

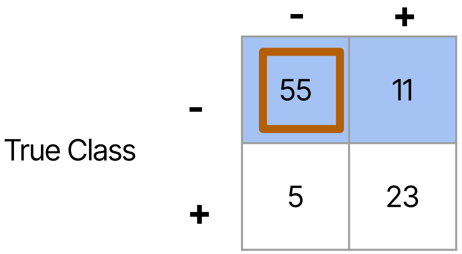
🟥 R → S = Very Major



Predicted Class

True Class		**-**	**+**
	-	55	11
	+	5	23

Sensitivity:
Fraction of Positive Class predicted as Positive



True Class		**-**	**+**
	-	55	11
	+	5	23

Specificity:
Fraction of Negative Class predicted as Negative

The Build Fellowship

OpenAvenues FOUNDATION

Binary Assessments

For AMR Predictions:
- **Major** errors
- **Very** major errors

In general we can think of this as Positive & Negative

- Positive Predictions = R
- Negative Prediction = S

Common Metrics:
- Sensitivity
- Specificity
- **Balanced accuracy** = average of the above

Baseline Majority Model

		Predicted Class	
		S	**R**
True Class	**S**	80	0
	R	20	0

Accuracy = ???

Sensitivity = ???

Specificity = ???

Balanced accuracy = ???

Binary Assessments

For AMR Predictions:
- **Major** errors
- **Very** major errors

In general we can think of this as Positive & Negative

- Positive Predictions = R
- Negative Prediction = S

Common Metrics:
- Sensitivity
- Specificity
- **Balanced accuracy** = average of the above

Baseline Majority Model

		Predicted Class	
		S	**R**
True Class	**S**	80	0
	R	20	0

Accuracy = 80 %

Sensitivity = 0 %

Specificity = 100 %

Balanced accuracy = 50 %

Uncertainty

QUIZ TIME !?

Why do we need to use statistical uncertainty?

a) To add a touch of mystery and suspense to our data analysis.
b) To make more accurate predictions on future data points.
c) To ensure that our model perfectly fits the data.
d) To provide a range of outcomes for better decision making.

Uncertainty, why?

In almost all of Data Science be it inference or predictions:

- We are never certain

We're using some sample of data to try to generalize to a population

- We took ~1,000 E coli samples from BV-BRC
- How representative are they?

In our predictive modeling problem we have three main pieces of uncertainty:

1. Sample size uncertainty - how much data?
2. Model variability - what if I trained the model again?
3. Generalizability - will this work on new data?

Sample Size - Statistical Power

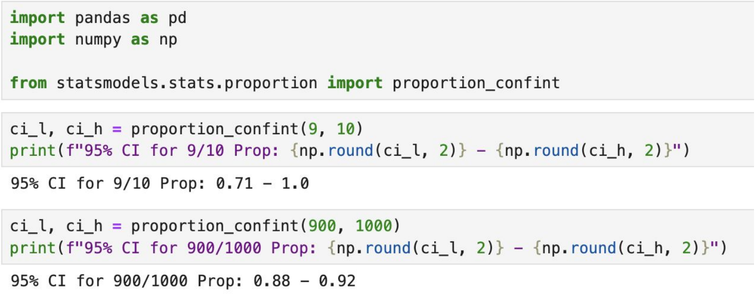
Our first layer of uncertainty. Simply put is there a difference between:

1. 9 / 10 predictions = 90% accuracy
2. 900 / 1000 predictions = 90% accuracy

Both cases have identical accuracy but we have different levels of evidence

More data points = more certainty in our assessment

Can calculate confidence intervals using common statistical tests (in this case a binomial proportion CI)



```
import pandas as pd
import numpy as np

from statsmodels.stats.proportion import proportion_confint
```

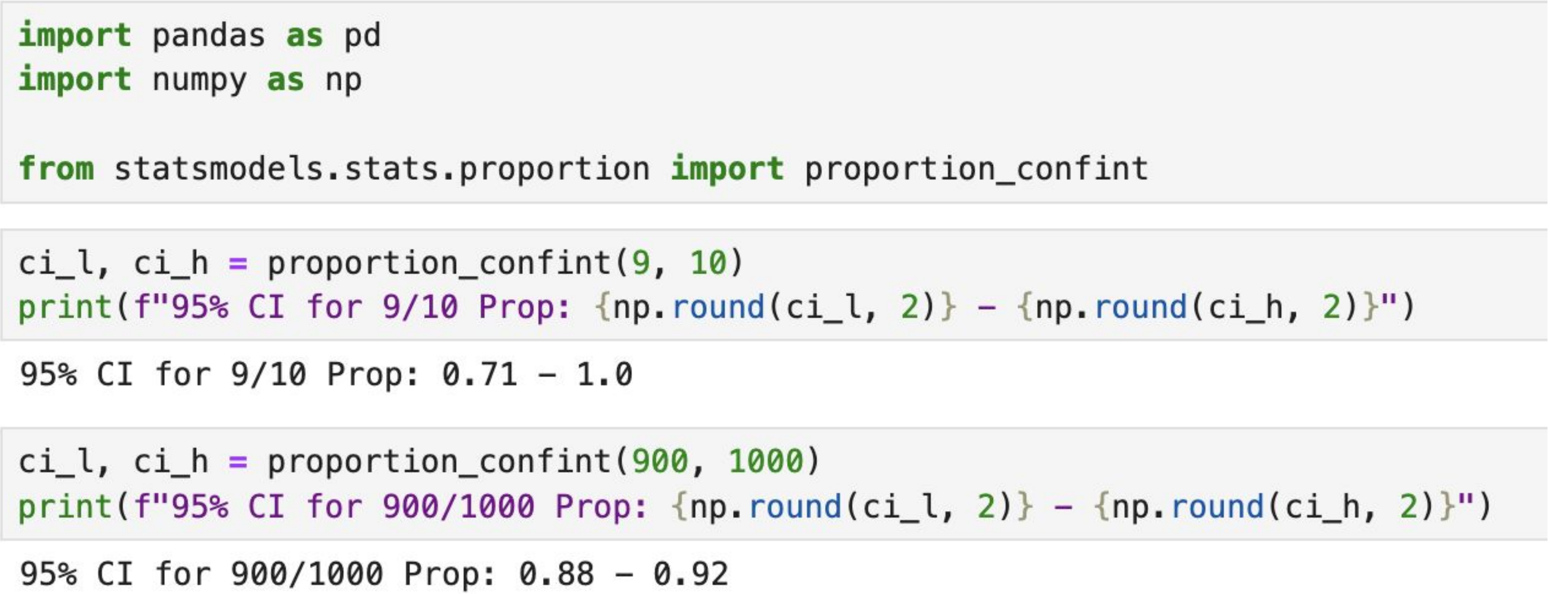
```
ci_l, ci_h = proportion_confint(9, 10)
print(f"95% CI for 9/10 Prop: {np.round(ci_l, 2)} - {np.round(ci_h, 2)}")
```

```
95% CI for 9/10 Prop: 0.71 - 1.0
```

```
ci_l, ci_h = proportion_confint(900, 1000)
print(f"95% CI for 900/1000 Prop: {np.round(ci_l, 2)} - {np.round(ci_h, 2)}")
```

```
95% CI for 900/1000 Prop: 0.88 - 0.92
```

Sample Size - Statistical Power



```
import pandas as pd
import numpy as np

from statsmodels.stats.proportion import proportion_confint
```

```
ci_l, ci_h = proportion_confint(9, 10)
print(f"95% CI for 9/10 Prop: {np.round(ci_l, 2)} - {np.round(ci_h, 2)}")
```

```
95% CI for 9/10 Prop: 0.71 - 1.0
```

```
ci_l, ci_h = proportion_confint(900, 1000)
print(f"95% CI for 900/1000 Prop: {np.round(ci_l, 2)} - {np.round(ci_h, 2)}")
```

```
95% CI for 900/1000 Prop: 0.88 - 0.92
```

Model variability

As we saw last week:

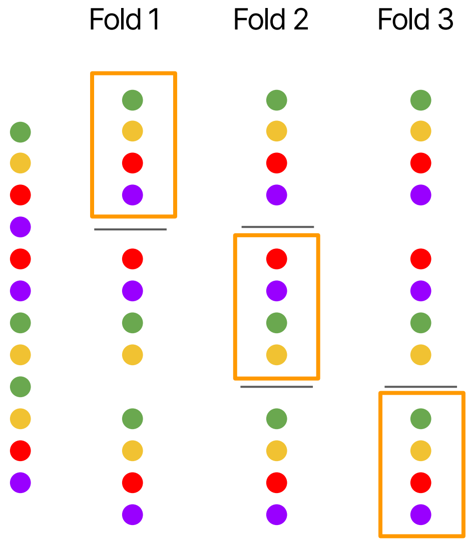
- Train a model multiple times
- Get difference assessments each time

Can we capture the variability across models?

- One common use for nested CV
- Inner folds to optimize model
- Outer fold to assess uncertainty

After completion of model assessment on CV:

- Use the final test dataset to get a single unbiased estimate



From last week's workshop, three Random Forest models:

```
Fold 1 Balanced accuracy: 80.4%
Fold 2 Balanced accuracy: 81.4%
Fold 3 Balanced accuracy: 80.6%
```

Each model was optimized on the inner fold splits

The Build Fellowship

OpenAvenues FOUNDATION

Generalizability

More of an expert knowledge question

- How representative is our data?
- Will it work on new data?

Need to review and assess your dataset carefully, in our case:

- Where did we get the E coli data from?
- Was it geographically diverse or from one state?
- Can we measure bacterial diversity? (Phylogeny)

Important to comment on your data collection!

If we train here

Can we predict here?

Summary

Not always necessary to calculate everything out

BUT:

- Vital to acknowledge and communicate uncertainty
- If not in numbers then in words

For interviews/take-home exercises I'm always looking out for candidates that are aware of and acknowledge uncertainty:

- Did they highlight places where they were unsure?
- Did they list assumptions/questions?
- Did they report any confidence levels?

The Build Fellowship OpenAvenues FOUNDATION

Workshop 7
Performance Evaluation