Useful Statistics and Error Analysis

Background

Suppose we have built an ML model that predicts three quantities:
(Calcium, Phosphorus, PTH)

We will look at some summary statistics of the data

We will look at some error metrics to understand the predictive perform

As we will learn, it is relatively easy to compute statistics and metrics...

...it often takes more work to measure the right things

About the Data

The data is all fictitious (simulated)

Summary Statistics

	Mean	SD
PTH	500	425
Calcium	9.1	0.7
Phosphorus	5.5	1.7

Based on these stats, which variable may be the hardest to predict?

Overall Error of Predictions

What are these metrics telling us?

Are they useful? What would be more useful?

RMSE (training set)	RMSE (test set)
0.28	0.30

Errors by Elements

What information does this add to our understanding?

What do we not know?

	RMSE (test set)
PTH	1.5
Calcium	0.2
Phosphorus	0.5

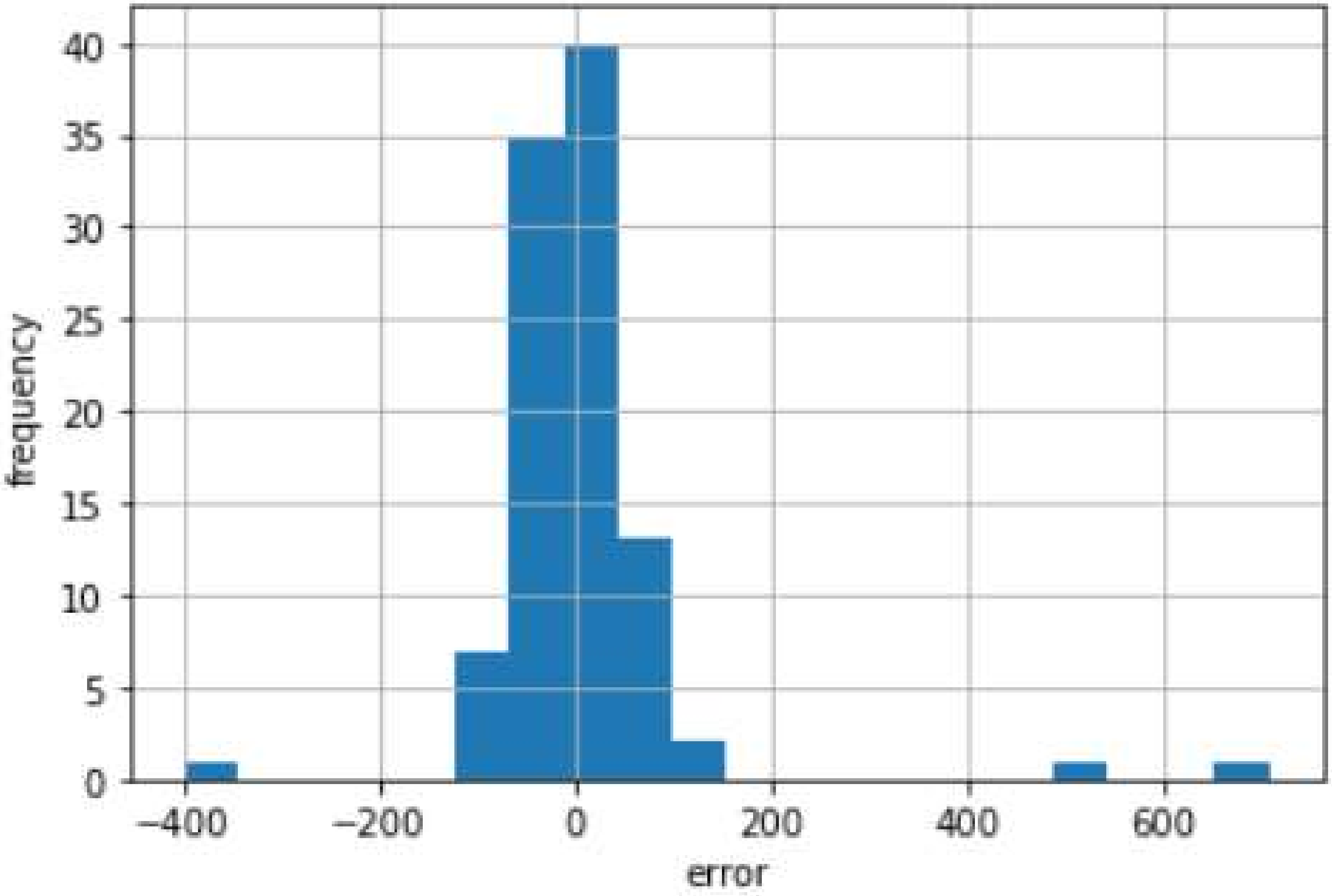
Histograms

Histograms show the full distribution

This is richer information than a statistic like a mean

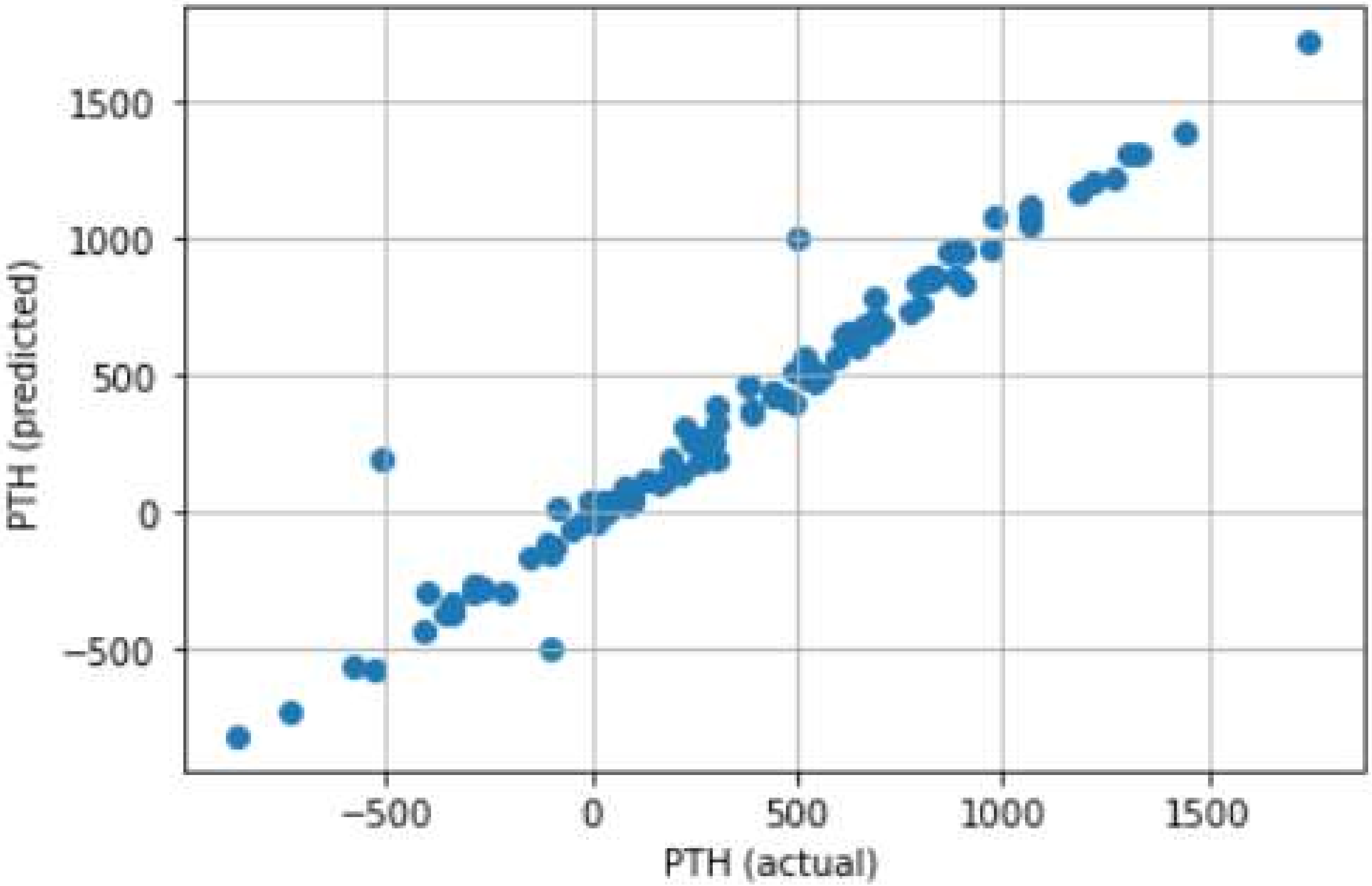
Histogram of Errors - PTH

How is this graph useful? Are there limitations?



Predicted vs Actual

How is this graph useful? Are there limitations?



Using Errors to Improve the Model

Deep diving into errors (*error analysis*) can help improve the model:

For the largest errors,

- Are there patterns?

- Are these cases due to invalid data?
  Action: Might want to repair / exclude this data.

- Are we missing predictors?
  Action: Might want to test more predictors.

Empirical Distributions

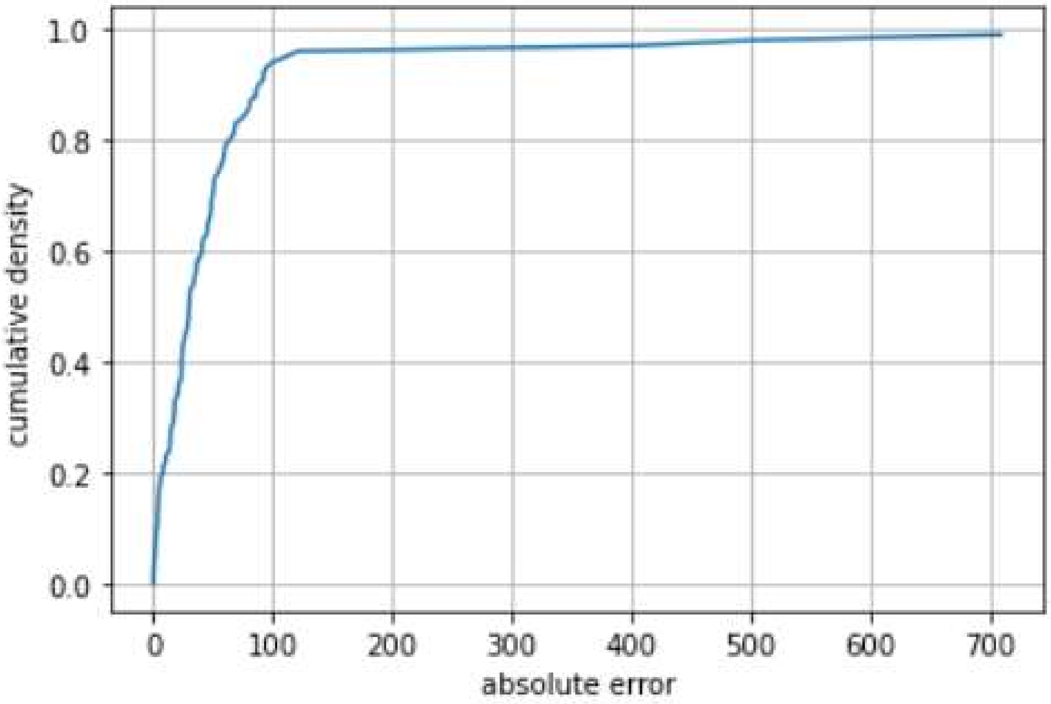
For a random variable, the **cumulative distribution function** shows th probability that the value is x or lower: F(x) = P(X <= x)

When we observe a batch of data, we can compute and graph this ove range of values for x. Called the **empirical distribution**.

This provides much richer information than the mean, as we can see th data across all percentiles.

Empirical Distributions, contd.

What information does this graph provide?



Percentiles

median	30.42
75%ile	54.90
90%ile	86.63

Takeaways

It can be tempting to:
> produce / report a lot of statistics
> try a battery of ML models

Given available packages, this is often easy to do!

But often it is better to:
- carefully consider which statistics are valuable
- carefully analyze prediction errors

Learnings from prediction errors can often improve results

No model can overcome bad data