Documentation

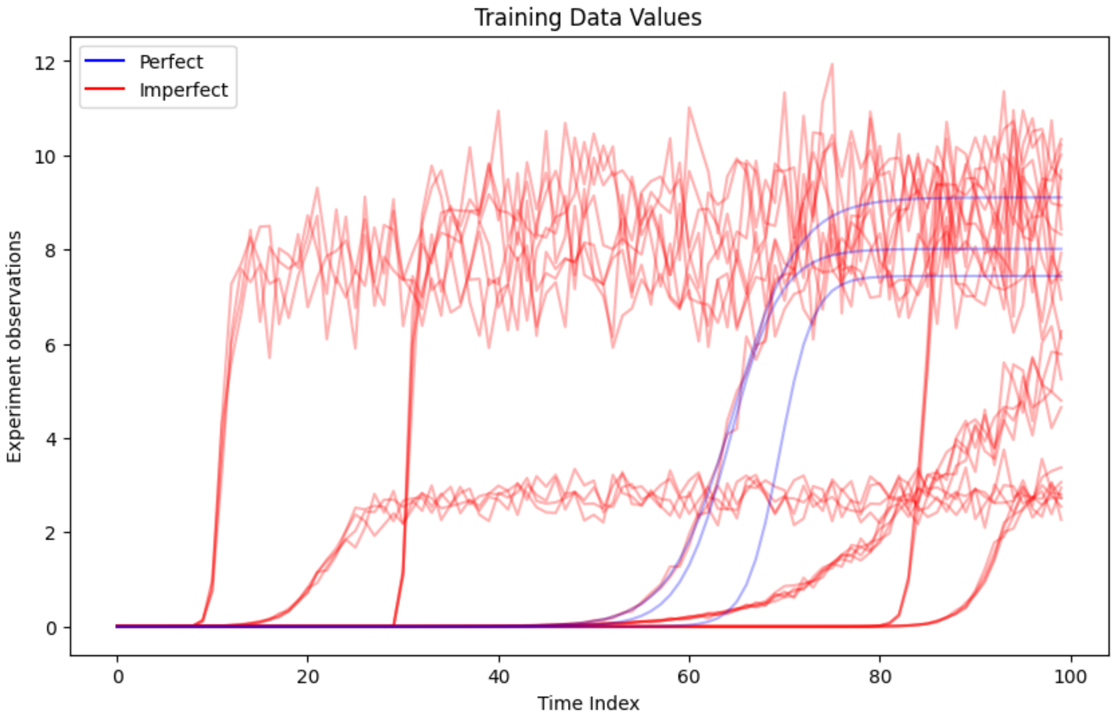
Data Challenges

As described in the instructions, the data is very imbalanced with only 4 "perfect" examples. Each experiment produce a short 100 step the time series for a total of only 39 records.

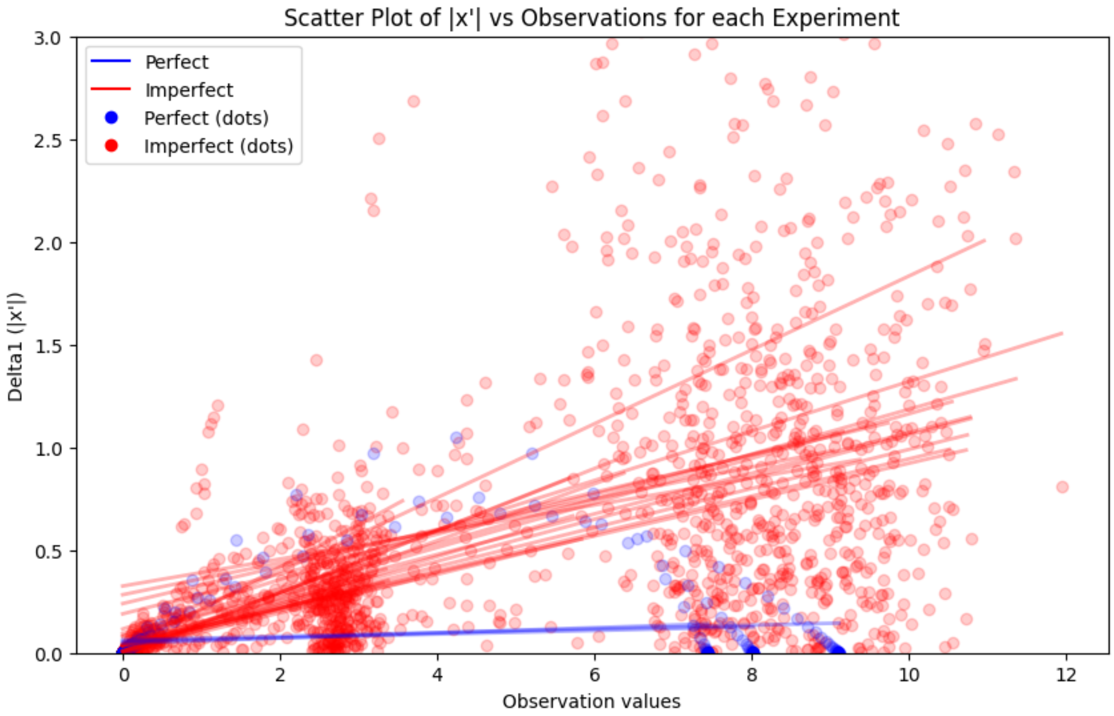
The limited data eliminates a number of ML models that require lots of data. With data of this size, feature engineering, explicit programming, and classical statistics might be preferable over ML modeling. The few "perfect" examples further limits how we conduct data splitting for model evaluations.

Normally, data exploration is done with domain expert in loop. This helps to understand how the data is collected and understand the concepts around the data production process. Without this background, it is very hard to distinquish true features from artifacts, distinquish causal relationships from correlations.

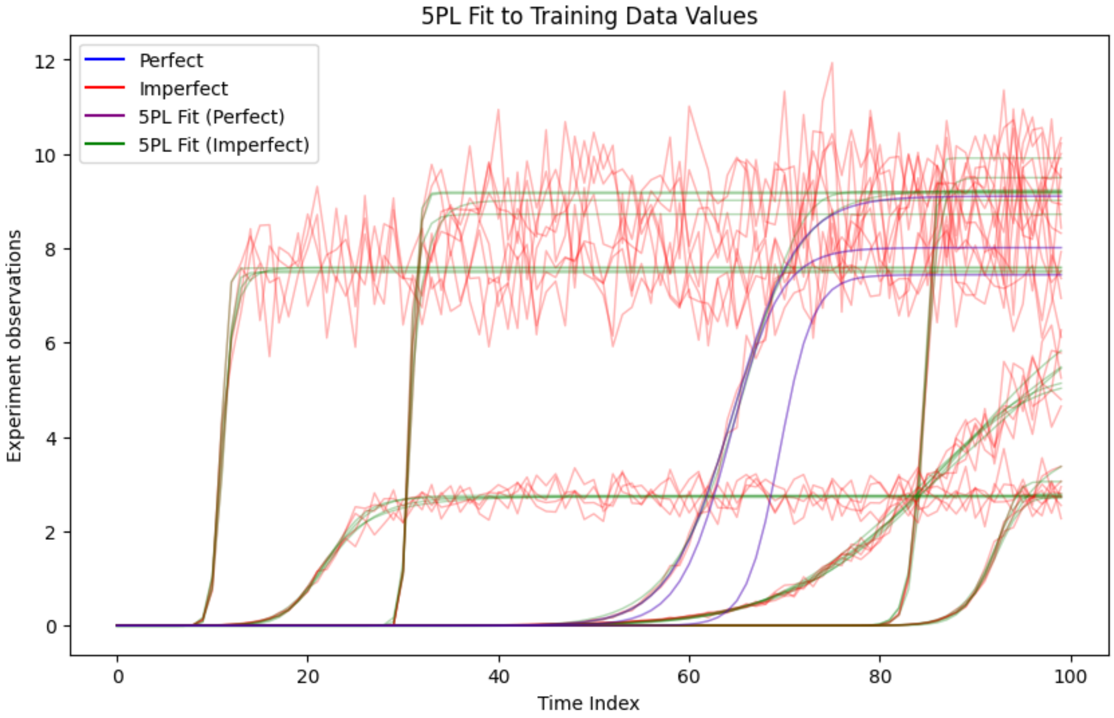
Here is a plot of the time series with the target labels in difference colors.



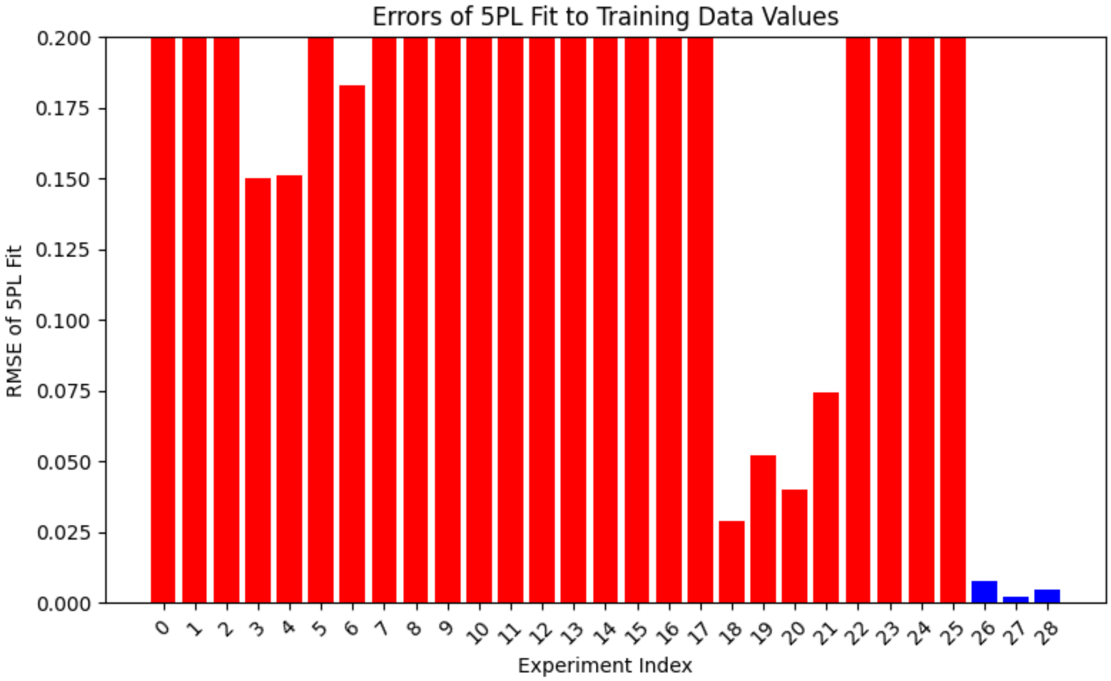
We found that the noise is associated with imperfect experiments. The amplitude of the noise increase as the observed values increase.



All of the experiments show a logistic curve and we can find the parameters of the curve using non-linear least square despite having noise in the imperfect experiments.



The root mean squared error (RMSE) can be computed between the fitted and actual curve. The perfect experiment show low RMSE compared to that of the imperfect.



Model Selection Rationale

Regression and Parametric Modeling

Based on the data exploration above, we found various features that can visually distinquish the target categories (linearly distinquishable). The time series data consistantly follow a logistic curve with various amount of noise, so they are "well-understood". These descriptions along with the fact that the data set is small and imbalanced lends itself to the application of regression and parametric modeling. Although they are not traditionally classification models, we can apply thresholds to produce boolean outputs representing imperfect and perfect. When the criteria are met and the appropriate functions are applied, the model can provide good fit (as opposed to under- or over-fit) despite the limitation mentioned.

Other Considerations

Decision Tree Classification:

Decision trees are often a good choice for determining threshold from features through data. It is often suitable in situations with smaller data sets. A decision was made not to use it. The perfect experiment category are too small, so letting the algorithm decide the cutoff threshold may lead to unnecessary overfitting even if the appropriate decision tree regularizations are applied.

Anomaly Detection:

For many imbalanced data situations, one can frame the problem as anomaly detection problem. The technique allow us to safely use various data generation techniques to "balance" the dataset. Anomaly detection usually have fewer records in the anomaly category where the anomalies are deemed outliers, but sufficiently dense data in the normal category. Based on the label names in this problem, perfect and imperfect, we have the opposit scenario: few normal outcomes (perfect) and many anomalies (imperfect). Thus, decision was made to avoid anomaly detection and proceed with curve fitting on 5PL (parametric modeling).

Choosen Features

- **c parameter**: calculated from fitting 5PL, this parameter indicate how far towards the right a particular curve has shifted. We expect perfect experiments to reside at a certain range.

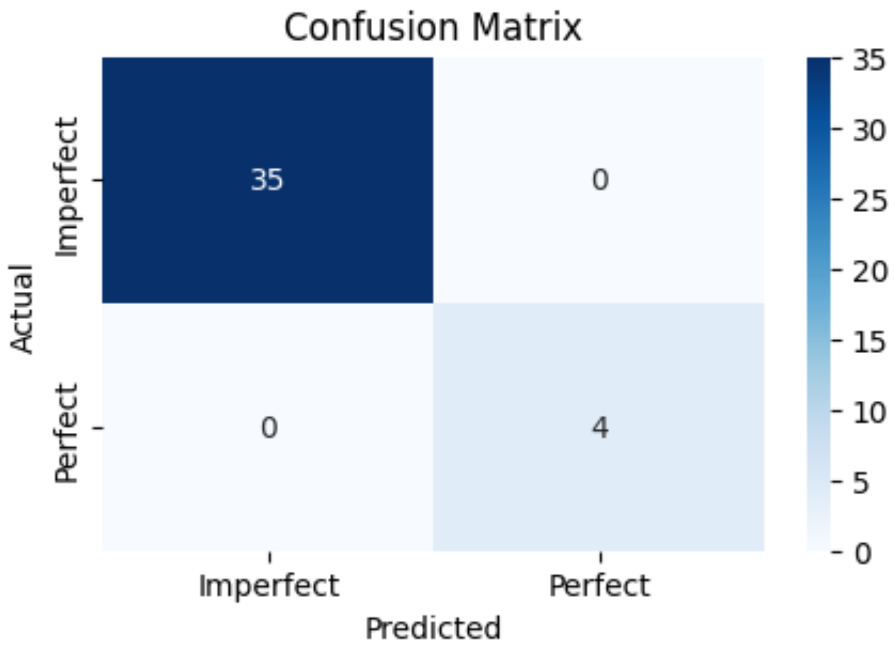
- **RMSE**: the root mean squared error (RMSE) of the 5PL fit determines the amount of noise in the observations that deviate from the 5PL curve. The perfect experiments should be a close fit with low RMSE.

In addition to the features for classification, we need to ensure that the input data can be used for curve fitting of the 5PL function. Here are the criteria for data validation:

- **min**: aggregated minimum of the time series. This must be between a range of values
- **max**: aggregated maximum of the time series. This must be between a range of values

With clear linear seperation, threshold and bounds have been set for each of these feaures to ensure good fit without overfitting.

Ensure Robust Model Performance

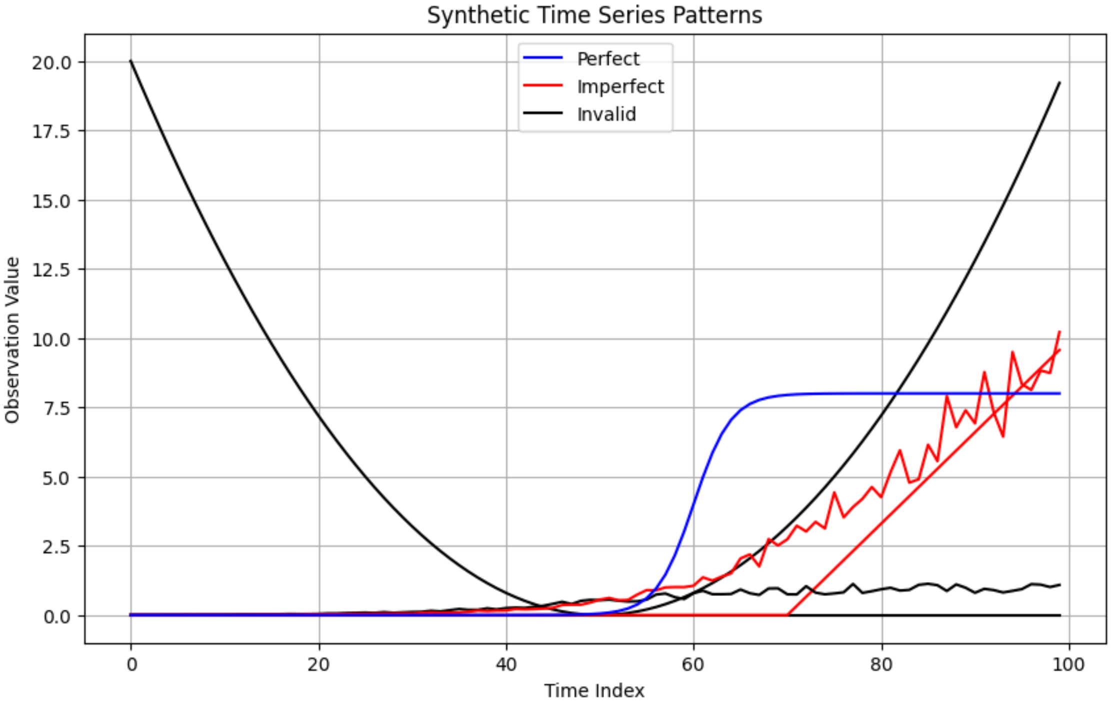


```
              precision    recall  f1-score   support

   Imperfect       1.00      1.00      1.00        35
     Perfect       1.00      1.00      1.00         4

    accuracy                           1.00        39
   macro avg       1.00      1.00      1.00        39
weighted avg       1.00      1.00      1.00        39
```

As mentioned above, data validation is applied to ensure the input are well-behaved for computing least-square. We multiple multiple features to conduct classification, eliminating potential draw back of each individual feature. Model was evaluated based on both the training and holdout set to check for overfitting. The model showed a 1.0 in precision and recall for both categories, indicating good fit. We have checked for edge cases using synthetic data to further improve our confidence about the model.


Synthetic Time Series Patterns
Perfect
Imperfect
Invalid
Observation Value
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Time Index
0
20
40
60
80
100