MYDATAMODELS – ZGP ENGINE

SIGNAL DETECTION DEMONSTRATION

This document seeks to demonstrate the signal detection characteristics of the MyDataModels ZGP modeling engine. A synthetic data set was constructed to illustrate the challenges faced by traditional statistical methods when trying to predict phenomena governed by mathematical relationships. Modeling of the synthetic dataset was performed using the MyDataModels ZGP engine and a number of other popular modeling techniques.

Table of Contents

DATASET CONSTRUCTION

A synthetic dataset was created in Excel by creating random data points.

The Y axis value was chosen to be a simple continuous non-linear function which will be described later in this document. The X coordinates were randomly generated but constrained to a small section of the continuous curve.

X constraints were chosen such that only a very small aspect of the function's curve was expressed in the data set. This was done to simulate the common real world situation of incomplete data and data sets that are not fully representative of the phenomenon under consideration.

Further, the data set contained 35 observations. This small number was chosen to mimic situations where abundant data is expensive or simply unavailable. The limited nature of the small data set also serves to make analysis and discovery of the underlying function more challenging to the algorithms.

Two distinct data sets with these characteristics were produced. The first set, which will be referred to as the "training set", is the dataset analyzed to produce a model. The second data set, which will be referred to as the "test set", is the data set against which the generated models were evaluated. As both sets follow the same mathematical relationship but contain different random points along the same continuous curve, the results serve to assess the completeness of the models generated by the different algorithms.

A graph of the 35 points in the training set is shown in Figure 1 and a graph of the 35 points in the test set is shown in Figure 2.

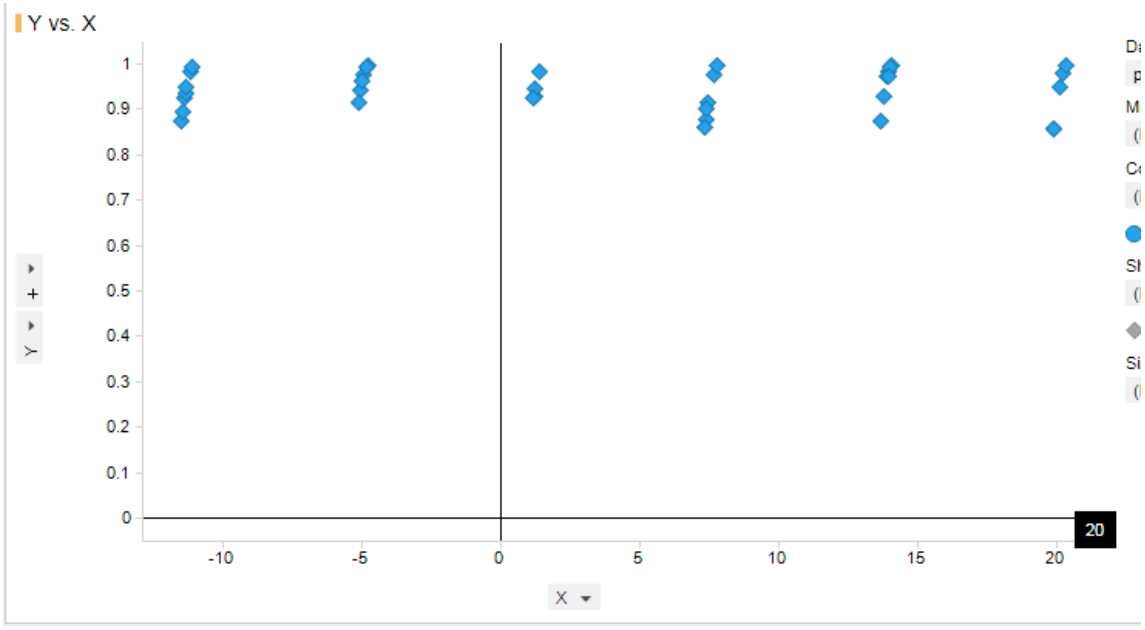

Figure 1. Graph of training set data points.

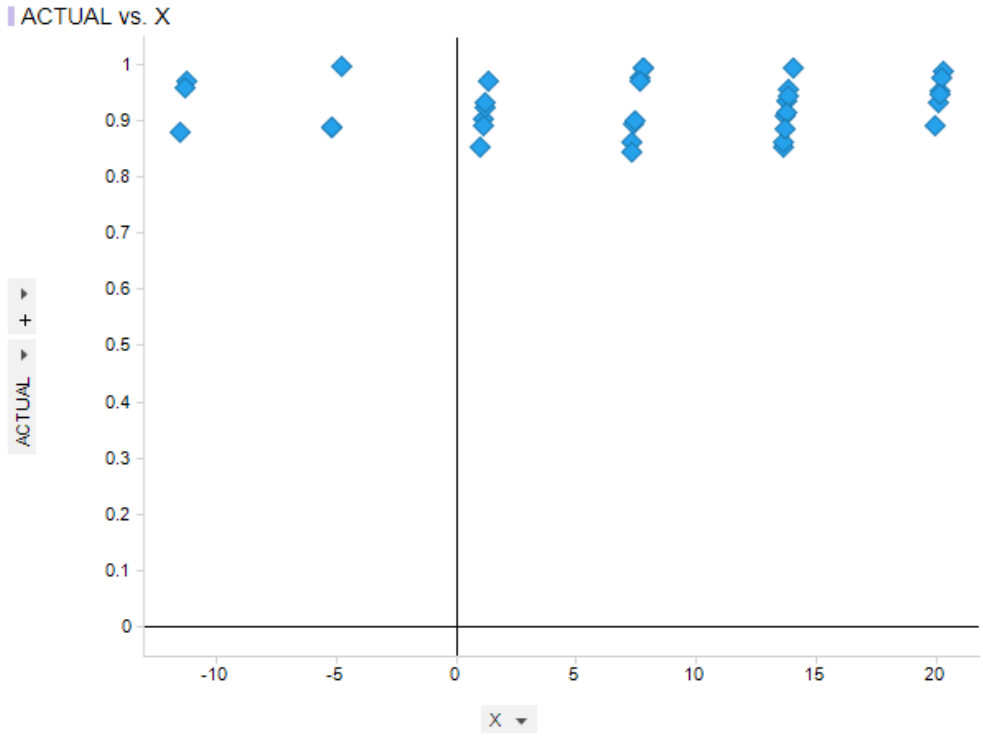

Figure 2. Graph of test set data points.

MODEL GENERATION AND ASSESSMENT:

Modeling of the training data was performed using four different machine learning algorithms: Least squares regression, k-Nearest-Neighbor, Regression Trees, and the MyDataModels ZGP engine.

The models produced by each of these algorithms are discussed below.

Least Squares

The model produced by Least Squares Regression is shown below in Figure 3.

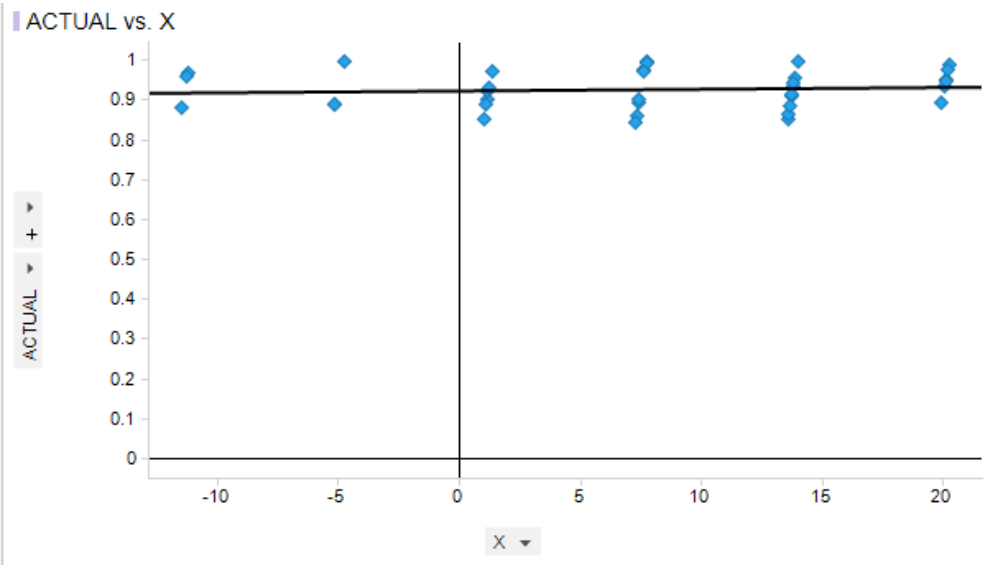

Figure 3. Graph of Least Squares regression line with respect to the test set. Black line represents the obtained regression line. Blue diamonds represent data points in the test set.

Depending upon the desired accuracy, it may appear that the least squares regression line could be an acceptable predictor for this dataset. A more in-depth consideration of the quality of the models will be made later in this document.

k-Nearest-Neighbor (kNN)

The predictions made by the k-Nearest-Neighbor model are shown below in Figure 4.

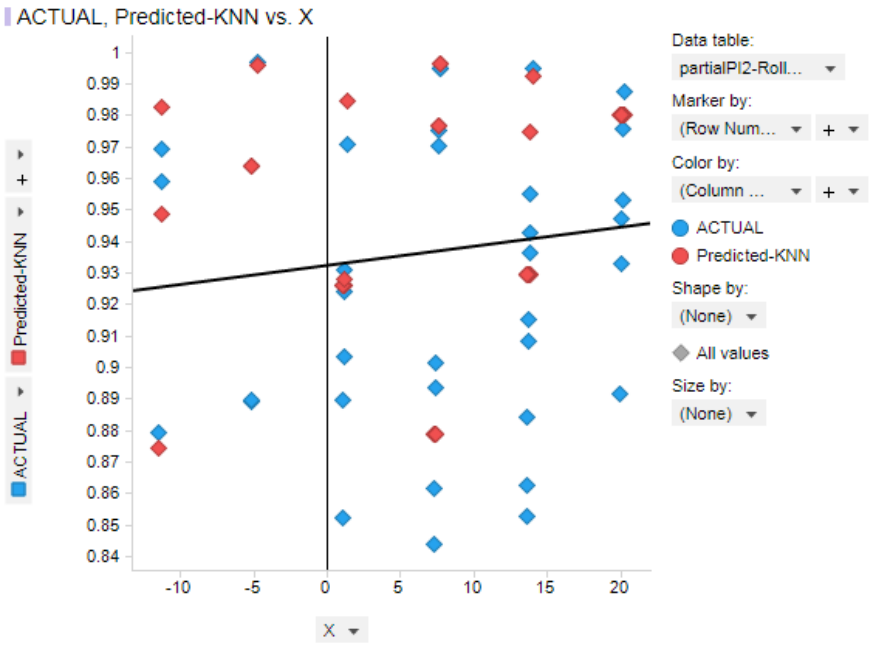

Figure 4. Graph of k-Nearest-Neighbor predictions with respect to the test set. Black line represents the least squares regression line. Blue diamonds represent data points in the test set. Red diamonds represent the predictions produced by the k-Nearest-Neighbor model.

The k-Nearest-Neighbor model produced some predictions which appear to resemble the actual data but in general it does a poor job of identifying and replicating the underlying mathematics in the data.

Regression Trees

For regression trees, a regression tree model was generated using a best pruned tree with a minimum terminal node size of 2. The resulting model was evaluated against the test set.

The predictions made by the resulting model are shown below in Figure 5.

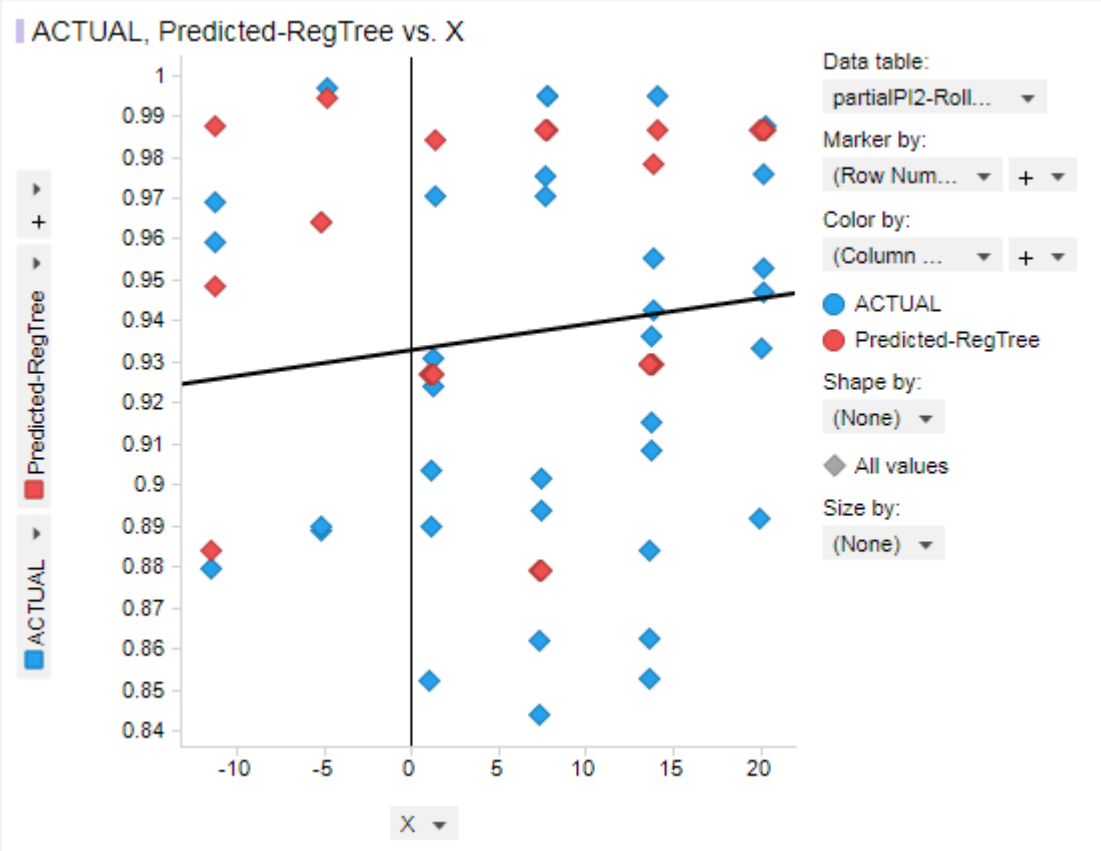

Figure 5. Graph of regression tree model predictions with respect to the test set. Black line represents the least squares regression line. Blue diamonds represent data points in the test set. Red diamonds represent the predictions produced by the generated regression tree model.

The regression tree model showed similar results to the k-Nearest-Neighbor model. The error rate is still quite significant. Like the kNN model, the regression tree model in general did a poor job of identifying the mathematical behavior of the data.

MyDataModels ZGP

Lastly, modeling was performed using the ZGP analysis engine. A single modeling attempt was performed using all default settings. No parameter turning was performed. The results of that single model are considered here.

The predictions made by the resulting model are shown below in Figure 6.

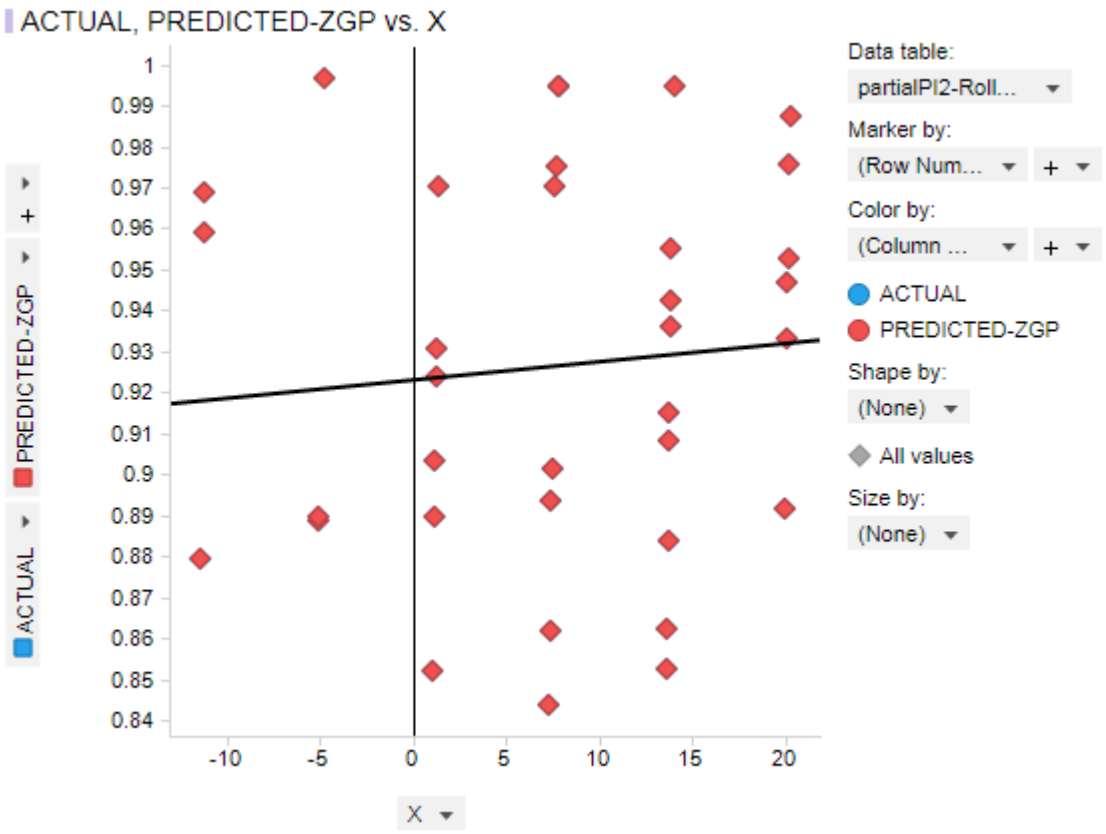

Figure 6. Graph of ZGP predictions with respect to the test set. All predictions perfectly match actual values.

The ZGP model perfectly predicted the y values of all data points in the test set.

EQUATION BASED MODELING

The ZGP engine best understood the mathematics behind the data points. The reason that the ZGP engine performed so well is due to the methods employed by the ZGP engine. The ZGP engine does not seek to find artifacts within the data set by which to discriminate similar observations nor does it employ statistical measures such as prevalence, mean and probability to produce a prediction. ZGP seeks to find a mathematical equation, driven by data values, that explains the dependent variable. This allows ZGP to build models which strive to express as complete a description of the system behavior as possible, as supported by the data.

As a result, such models are quite effective at predicting proper values when independent variables take values which were not represented in the analysis data set.

This is often a crucial capability as it is very rare in practice that a sample data set contains data points fully representative of every possible manifestation of the phenomenon under consideration. It is also rare for a data set to contain the complete range of values an independent variable may take.

Purely statistical based algorithms are often poorly equipped to predict behavior outside of the values provided by a specific data set composition.

THE UNDERLYING MATHEMATICAL RELATIONSHIP

The effectiveness of the ZGP model becomes clear when the actual underlying function is revealed.

Due to the choice of a limited range of x values, identifying the underlying equation by visual inspection is challenging. It is not at all obvious from the small and constrained points described by the training set and test set but these points perfectly describe the following simple equation:

$y = \sin x$

Figure 7 shows the data points in the test set and the complete curve of $y = \sin x$.

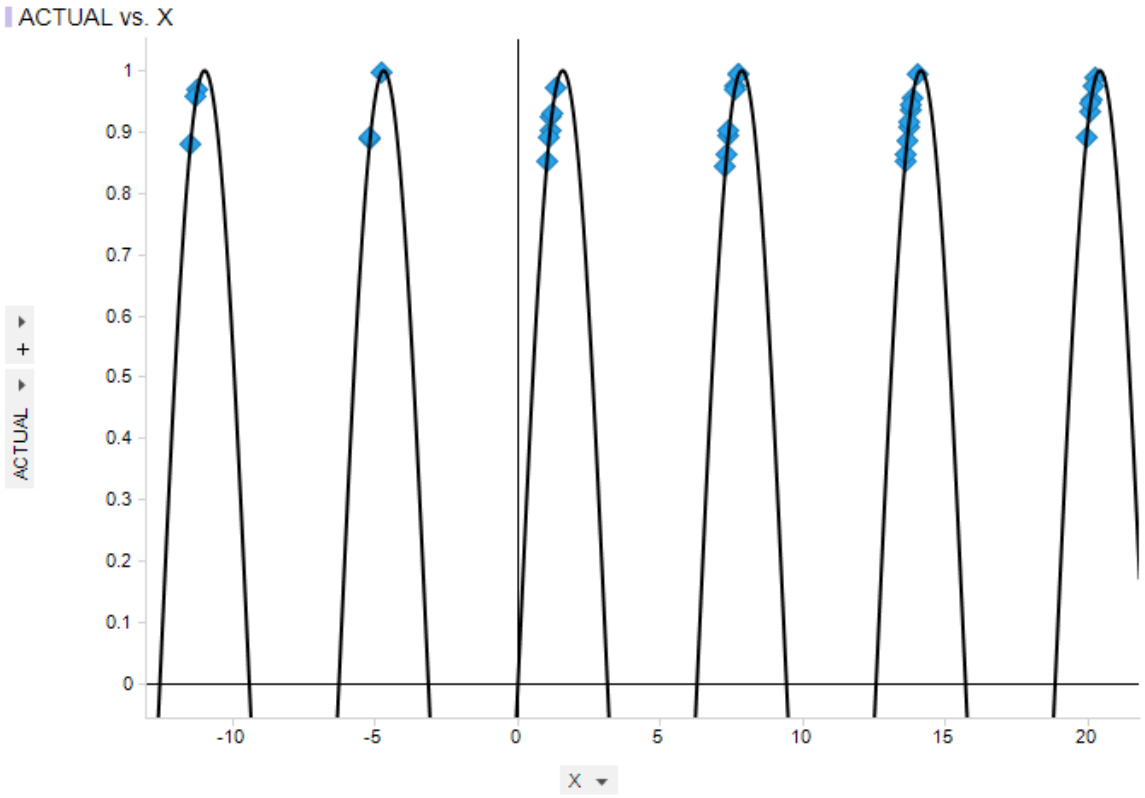

Figure 7. Test set data points and the graph of $y = \sin x$.

Note that the dataset contained data points for only a tiny fraction of the complete mathematical curve being modeled. This is an extreme case intended to illustrate a point that is much more subtle and hard to numerically quantify in real world analytic efforts. Real world data sets contain many dimensions of variables

comprised of imperfect, incomplete and not fully representative data. These exact same principles apply regardless of magnitude or mathematical form of this characteristic.

MODEL ASSESSMENT AGAINST FULL RANGE OF THE UNDERLYING FUNCTION

The potential value and significance of equation based modeling can be seen when the same models are used to evaluate data points on the complete curve. This challenges the modeling algorithms to demonstrate complete understanding of the underlying mathematics by predicting segments of the curve for which there were no observations present in the training data.

In the sections below, the models discussed above are evaluated against x values which did not appear in the training or test sets. These x values represent the real world case of evaluating models when independent variables take on values not exhibited in the data used to produce the model.

k-Nearest-Neighbor (kNN)

Figure 8 shows predictions made by the k-Nearest-Neighbor model for x values from -10 to 30 in increments of 0.5.

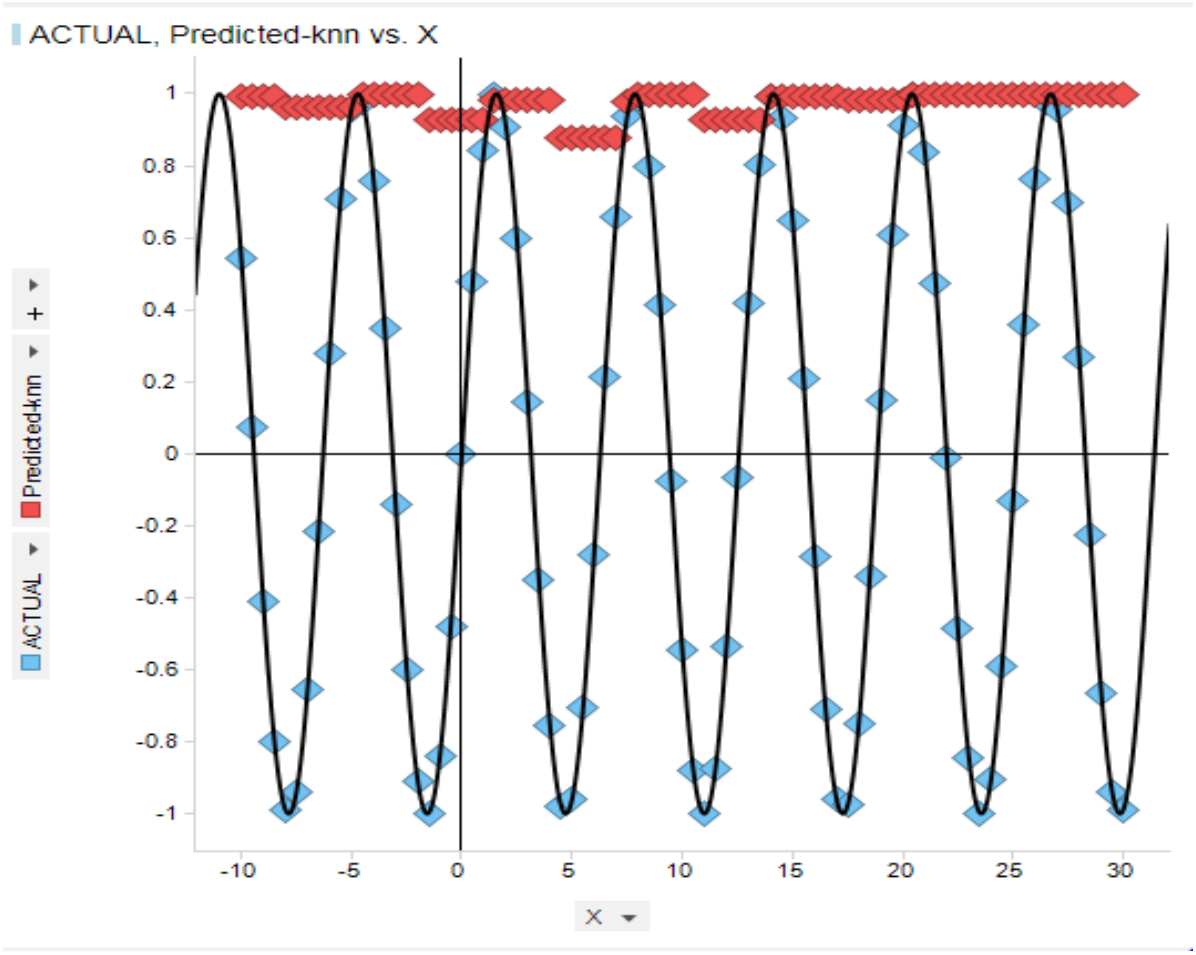

Figure 8. K-Nearest-Neighbor model predictions for values of x from -10 to 30 in steps of 0.5. Blue diamonds represent actual values. Red Diamonds represent model predictions. Black curve shows a graph of the actual function.

The graph shows that kNN was unable to make useful predictions for values of x that were not present in the training data.

Regression Trees

Figure 9 shows predictions made by the regression tree model for x values from -10 to 30 in increments of 0.5.

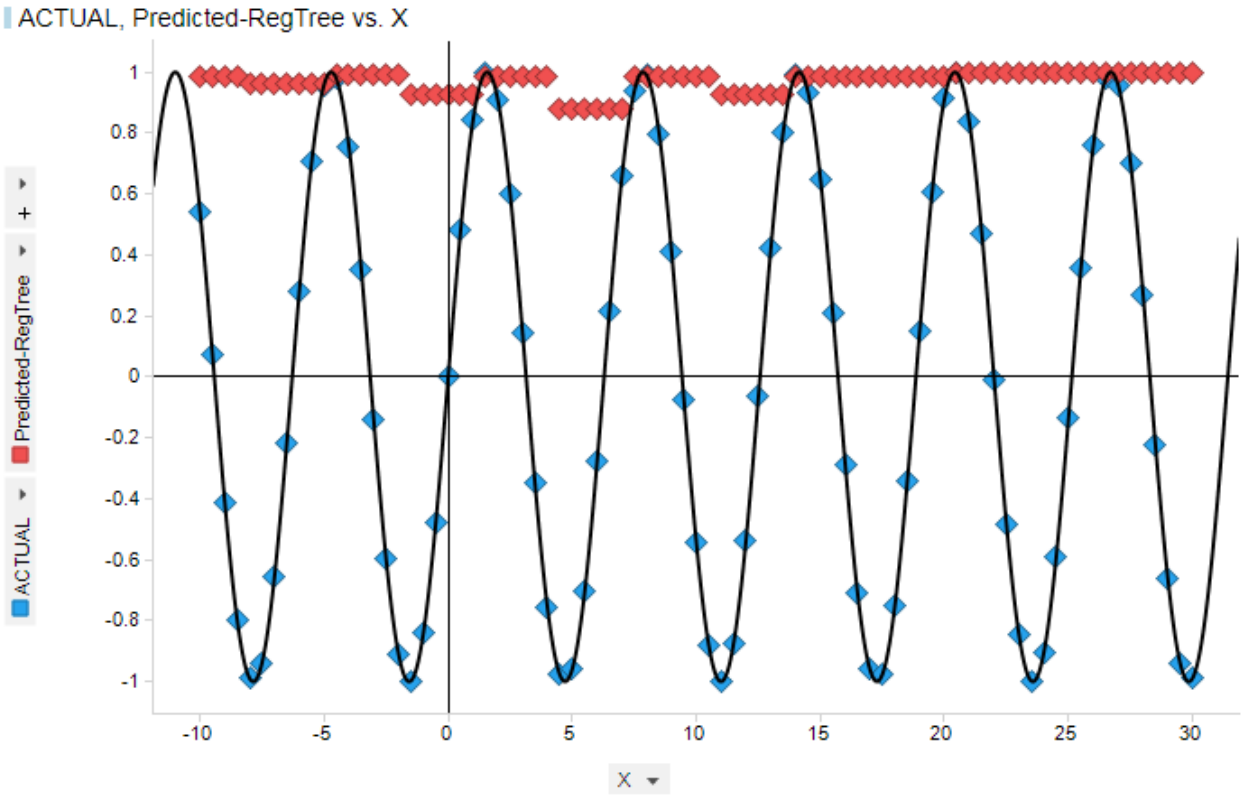

Figure 9. Regression Tree model predictions for values of x from -10 to 30 in steps of 0.5. Blue diamonds represent actual values. Red Diamonds represent model predictions. Black curve shows a graph of the actual function.

The regression tree model was also unable to model behavior for values of x that were not sampled and present in the training data.

MyDataModels ZGP

Figure 10 shows predictions made by the ZGP model for x values from -10 to 30 in increments of 0.5.

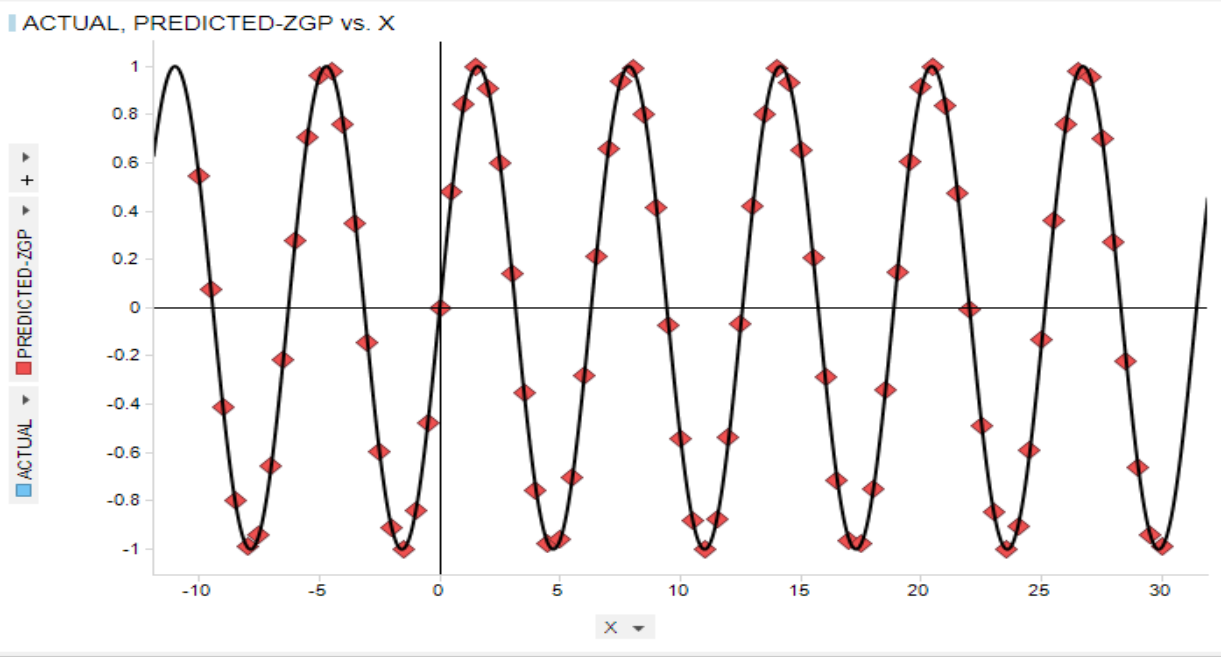

Figure 10. ZGP model predictions for values of x from -10 to 30 in increments of 0.5. All points are perfectly predicted by the ZGP model. Black curve shows a graph of the actual function.

The ZGP model perfectly predicts all data points in the full range of x values. The ZGP model was able to infer the true mathematical behavior described by the training set despite that that data set contained only a small portion of the full range of x values. The model formula reported by the ZGP successfully identified that Y=SIN(X):

PREDICTOR:
FUSION(SIN, SIN, 0.800854505226215, X, X)

While this example is a very simplified case, the principles apply to other mathematical relationships regardless of composition and complexity.

In real world data sets it is very rare that a sample of data presents the full range of values involved in an underlying process. The ability to model these aspects mathematically can be a potent tool to mitigate limitations imposed by incomplete or non-representative data sets.