Symbolic Regression GECCO Competition - 2023

Track 1 - Performance

Gurushant Gurushant (gurushant.gurushant@stud.fra-uas.de)
Jatinkumar Nakrani (jatinkumar.nakrani@stud.fra-uas.de)
Rajni Maandi (rajni.maandi@stud.fra-uas.de)

---

1. Abstract

This study present a comprehensive solution designed for the Symbolic Regression GECCO Competition 2023 - Track 1 Performance, focusing on the analysis of datasets using symbolic regression. Our solution consists of various stages, including dataset exploration, model training through programming, evaluation based on accuracy and simplicity, and the selection of the best model. Our aim is to contribute to the competition by generating Sympy-compatible expressions that accurately represent the datasets and offer insights for further analysis.

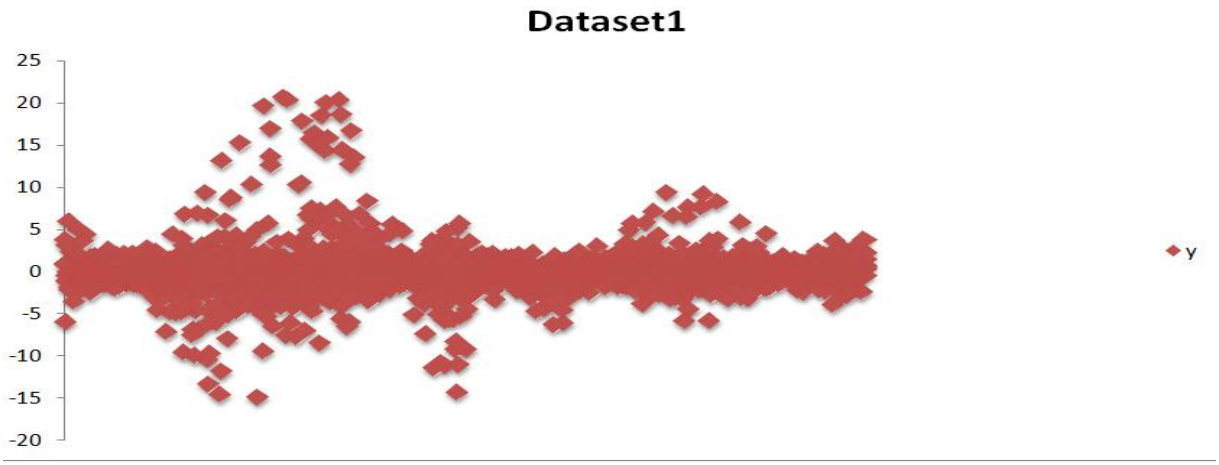
2. Introduction

- In this task, we will be given the flexibility to investigate and experiment with a provided dataset with the goal of producing best model equation, score and carrying out thorough analyses that are centered on the interpretation of these models.
- The objective of the Track1 is to demonstrate the significance of performance in machine learningand to motivate participants to use symbolic regression models to extract meaningful information fromthe existing datasets and model.
- Symbolic Regression offers an effective system for model creation and performance evaluation, making it a greattool to investigate complex connections within information.

3. Pre-Analysis

Dataset 1:

- The dataset consists of 2000 rows and 10 columns. The columns are labeled as x0, x1, x2, x3, x4, x5, x6, x7, x8, and y.
- The x0, x1, x2, x3, x4, x5, x6, x7, and x8 columns contain numerical values which are generated from an equation, where the last column y is the regression target, and the rest of the columns are the input data.
- We observed that for variable y, maximum values lies in the range from -5 to 5 which can be seen from below scatter plot.

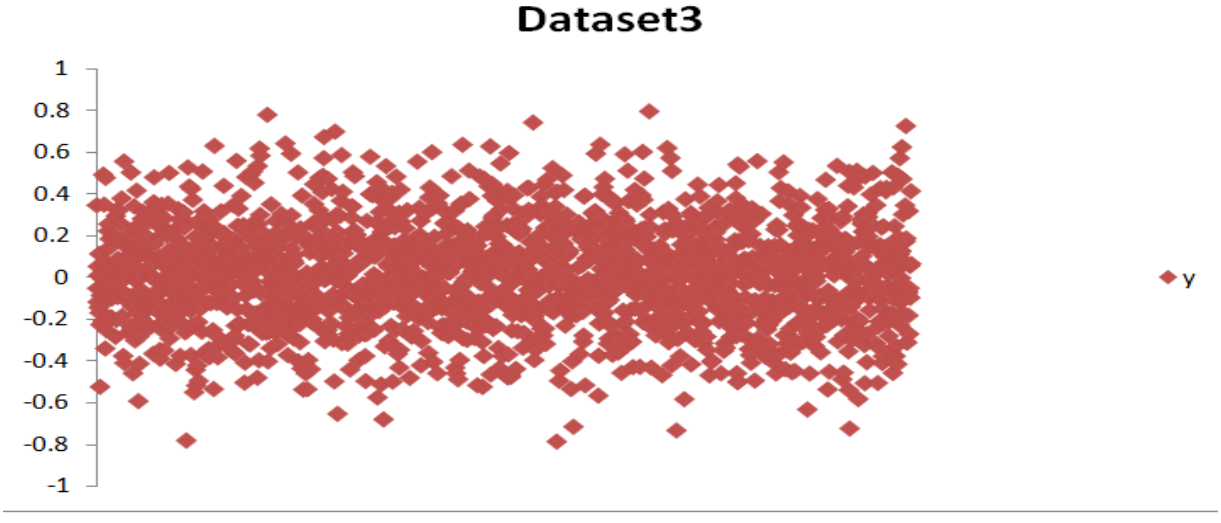


Dataset 2 :

- The dataset consists of 2000 rows and 7 columns. The columns are labeled as x0, x1, x2, x3, x4, x5, x6 and y.
- The x0-x6 columns contain numerical values which are generated from an equation, where the last column y is the regression target, and the rest of the columns are the input data.

Dataset 3 :

- The dataset consists of 2000 rows and 17 columns.The columns are labeled as x0, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15, and y.
- The x0-x15 columns contain numerical values which are generated from an equation, where the last column y is the regression target, and the rest of the columns are the input data.
- We observed that for variable y, maximum values lies in the range from -0.5 to 0.5 which can be seen from below scatter plot.



The range and distribution of values in each column may need to be further examined to identify any patterns or outliers which we have done in further analysis.

4. Algorithm/Coding

- In dataset1, we have only considered xo, x2,x4,x6 and x7 as input variable because we plotted all the input variable w.r.t y after which we observed that variable x0 has similar values as x1 , variable x2 has similar values as x3, variable x4 has similar values as x5. In order to remove the duplicity we have removed these variables.
- Since similar behavior observed in dataset3, therefore we have removed noise from dataset3 too like we did in dataset1.
- No such behavior is observed in dataset2.
- Linear Regression (dataset 1 & 3) and gplearn (dataset2) used for evaluating best equation and score.

5. Post-Analysis

- Dataset1 best model equation and score generated:
  -0.061889602053649025 + 0.08551798260799069*x0 + -0.0067068827556578 7*x2 + -0.004756698297098 2516*x4 + -0.02041331998043954*x6 + -0.03427264594150 7265*x7
- Dataset2 best model equation and score generated :
  x5/x3
- Dataset3 best model equation and score generated:
  0.04775351580625009 + -0.0019599676933630458*x1 + -0.0002870054321218585*x2 + -0.002376520502176837*x4 + 5.436209210113286e-06*x9 + -0.0005229598165321627*x10 + 7.162350305406234e-05*x13 + -0.0009928500252631746*x14

	Name	Dataset	R2	size
1	his_jsr_2023.git	1	-0.0012777831848536092	17
2	his_jsr_2023.git	2	0.09715552839364439	5
3	his_jsr_2023.git	3	0.0025825709242290884	23

6. Conclusion

We concluded that for dataset1: The best model equation suggests a linear relationship between the predicted output 'y' and the input variables ('x0', 'x2', 'x4', 'x6', 'x7'). The equation includes constant terms (-0.061889602053649025) and coefficients for each input variable. The coefficients indicate the strength and direction of the relationship between the input variables and the predicted output.

For datset2,the best model equation indicates a simple linear relationship between the input variables 'x5' and 'x3'. The predicted output 'y' is obtained by dividing 'x5' by 'x3'. This suggests that the value of 'y' is directly proportional to the ratio of 'x5' to 'x3'. It implies that changes in the values of 'x5' and 'x3' will directly affect the predicted output 'y' in a linear manner.

For dataset3, The best model equation for Dataset3 is more complex, involving multiple input variables ('x1', 'x2', 'x4', 'x9', 'x10', 'x13', 'x14') with respective coefficients. The equation implies that the predicted output 'y' is influenced by a combination of these variables, each weighted by its coefficient. Positive coefficients indicate a positive relationship, where an increase in the corresponding input variable will lead to an increase in the predicted output. Negative coefficients suggest a negative relationship, indicating that an increase in the input variable will result in a decrease in the predicted output.