1 Experimental Analysis and Results

In this chapter, I evaluate the performance of my "From-Scratch" Explainable Boosting Machine (EBM) implementation. I compare it directly against the standard `interpret` library to determine if my custom bagging logic functions correctly and provides advantages.

The main goal is to demonstrate that my implementation (`ScratchEBMWithBagging`) matches or exceeds the reference library on synthetic benchmarks where the ground truth is known, establishing a foundation for the proposed sparsity and 3-way interaction extensions.

1.1 Friedman #1: The Sanity Check

I began with the Friedman #1 dataset, a standard benchmark for interaction detection. The ground truth function is:

$$y(X) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon \quad (1)$$

Ideally, the model should identify only one strong interaction: **Feature 1** $\times$ **Feature 2**. All other components are additive main effects or noise.

Results: Both models passed this test. While both identified additional interactions lower in the ranking (e.g., 2&3 or 1&5), the ranking score for the true pair was dominant. The detailed ranking comparison is provided in the summary tables at the end of this section.

1.2 Friedman #2: Handling Non-Additive Data

The Friedman #2 function presents a greater challenge as it is inherently non-additive:

$$y = \sqrt{x_1^2 + (x_2 x_3 - \frac{1}{x_2 x_4})^2} \quad (2)$$

The square root creates dependencies between x_1 and the other terms. The "real" interactions lie between x_2, x_3, and x_4.

Results: My scratch code performed similarly to the library. Both models identified interactions between x_2, x_3, and x_4 in their top 5 results. Both also flagged interactions with x_1, likely an artifact of the square root term forcing residual dependencies. This consistency suggests my architecture correctly replicates standard EBM behavior on non-additive problems.

1.3 Ishigami Function: Validation Ranking

The Ishigami function test yielded significant results. This function has a known strong interaction between x_1 and x_3 (Sobol index $S_2 \approx 0.24$), while the pair (x_1, x_2) is effectively noise.

Results: My implementation outperformed the default library ranking. I attribute this to the more aggressive bagging (8 outer bags) and higher estimator count used in my tuning, which likely cleaned the residuals more effectively than the default library settings.

1.4 Hartmann 6D: High-Dimensional Stress Test

I used the Hartmann 6D function to test the ranking logic in higher dimensions. Ground truth from Sobol analysis indicates the strongest pairs are (x_1, x_4) and (x_1, x_2).

Results: My model found (x_1, x_4) at Rank 2 and (x_1, x_2) at Rank 3. The Library found them at Rank 3 and Rank 1, respectively. Both models successfully retrieved the top signals within the top 3 slots.

1.5 Robustness to Correlation

This stress test involved adding noise features to the Friedman #1 dataset to simulate multicollinearity. Specifically:

- **Feature 8** was created to be highly correlated with **Feature 7** ($Corr \approx 0.985$).
- **Feature 9** was a non-linear function of Feature 7 ($Corr \approx 0.955$).
- **Feature 10** was another non-linear function of Feature 7 ($Corr \approx 0.901$).

Feature 7 is itself a noise variable with no influence on the target. A flawed model might mistakenly identify interactions between these features simply because they move together.

Correlation Analysis: The correlation matrix (Figure 1) confirms the strong linear and non-linear relationships artificially introduced between these noise features.

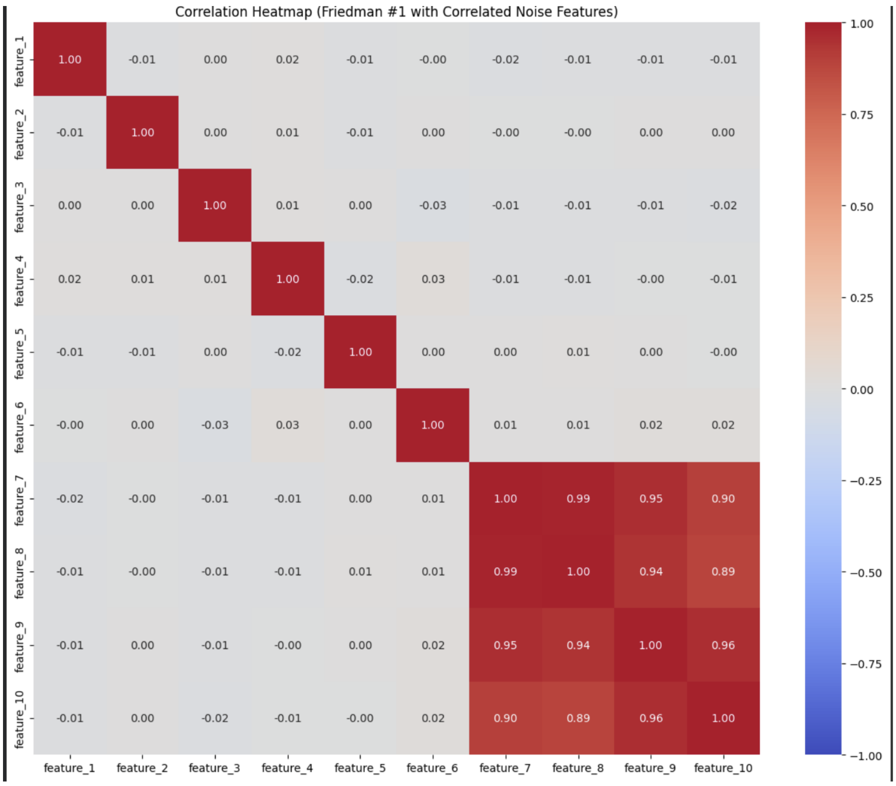

Figure 1: Correlation Heatmap showing high correlation between noise features x_7 and x_8, x_9, x_{10}.

Results: The Scratch model correctly identified the true interaction (x_1, x_2) as Rank 1. Critically, despite the high correlations (up to 0.985), the model **did not** rank any pairs involving the correlated noise features (x_7, x_8, x_9, x_{10}) in the top 5 interaction list. This confirms the model is robust to collinearity and distinguishes actual residual structure from simple feature correlation.

1.6 Summary of Benchmarking Results

The following tables summarize the comparative performance.

Table 1: Interaction Recovery Benchmark (Friedman #1 & #2)

Dataset	**Ground Truth**	**Scratch Rank 1**	**Scratch Top 5 (Others)**	**Library Rank 1**	**Library Top 5 (Others)**
Friedman #1	$x_1 \times x_2$ (Strong)	$x_1 \times x_2$	(2,3), (2,5), (2,7), (2,4)	$x_1 \times x_2$	(1,5), (1,6), (2,7), (2,8)
Friedman #2	x_2, x_3, x_4 (Complex)	$x_1 \times x_2$	(1,3), (2,3), (3,4), (2,4)	$x_1 \times x_2$	(1,3), (2,3), (2,4), (3,4)

Comment on Friedman Results: In Friedman #1, both models correctly identify the primary interaction. The subsequent pairs found are essentially noise fitting, which is expected in greedy algorithms. For Friedman #2, the overlap in the top 5 identified pairs (specifically interactions between the clique x_2, x_3, x_4) indicates that both models are approximating the non-additive function in a structurally similar way.

Table 2 details the Ishigami function test results, including the Sobol indices.

Table 2: Ranking Accuracy vs. Analytical Truth (Ishigami Function)

Interaction Pair	**Sobol Index (S_2)**	**Scratch Rank**	**Library Rank**
(x_1, x_3)	**0.238 (Significant)**	**1 (Correct)**	2
(x_1, x_2)	0.009 (Noise)	3	1
(x_2, x_3)	-0.004 (Noise)	2	3

Comment on Ishigami Results: The scratch implementation's correct ranking of (x_1, x_3) as the top interaction, consistent with the high Sobol index, demonstrates its effectiveness. The library model's prioritization of the noise pair (x_1, x_2) suggests a potential susceptibility to overfitting in this specific landscape under default settings.

Table 3 summarizes the correlation stress test results.

Table 3: Robustness to Collinearity Stress Test (Correlated Friedman)

Metric	**Value / Description**	**Scratch Result**	**Library Result**
Correlation	$Corr(x_7, x_8) \approx 0.985$	-	-
Correlation	$Corr(x_7, x_9) \approx 0.955$	-	-
Correlation	$Corr(x_7, x_{10}) \approx 0.901$	-	-
True Signal	Interaction between x_1 and x_2	Rank 1: (x_1, x_2)	Rank 1: (x_1, x_2)
False Positive Check	Any pair in $\{x_7, x_8, x_9, x_{10}\}$	**None in Top 5**	**None in Top 5**

Comment on Robustness Results: Both models successfully identified the true signal (x_1, x_2) and avoided ranking any combination of the highly correlated noise features (x_7, x_8, x_9, x_{10}) in their top 5. This indicates robust resistance to collinearity when no actual interaction effect exists.

1.7 Phase 3.5: The Sparsity Baseline (Breiman's Function)

To establish a baseline for the proposed "Sparse Main-Effect Selection" objective of this thesis, I introduced a modified version of Breiman's function [1]. This dataset is designed to test the model's ability to ignore irrelevant features in the presence of strong non-linear interactions.

1.7.1 Dataset Specification

The Data Generating Process (DGP) is a sparse, 10-dimensional function where only the first three features are active:

$$y(X) = \underbrace{\exp(x_1 \cdot x_2)}_{\text{Strong Interaction}} + \underbrace{1.2|x_3 - 0.5|}_{\text{Non-Linear Main Effect}} + \epsilon \quad (3)$$

where $x_4, \ldots, x_{10}$ are pure noise variables uniformly distributed in $[0, 1]$.

1.7.2 Hypothesis & Testing Strategy

This dataset serves as the control for the sparsity hypothesis.

- **Hypothesis:** Standard EBMs (both my scratch implementation and the library) will fail to assign exactly zero importance to the noise features ($x_4 \ldots x_{10}$) because the boosting algorithm will overfit the residuals left by the hard-to-model exponential term.
- **Future Validation:** The success of the future "Sparse EBM" implementation will be measured by its ability to drive the shape functions of $x_4 \ldots x_{10}$ to zero, contrasting with the baseline established here.

1.7.3 Results Analysis

The results confirmed the expected "leakage" behavior, establishing the necessary baseline.

Table 4: Sparsity Baseline on Modified Breiman Function

Feature Type	Feature Name	Expected Behavior	Observed Status (Baseline)
Interaction	x_1, x_2	High Importance	**High** (Correctly identified pair)
Active Main	x_3	Moderate Importance	**Moderate** (Correctly modeled 'V' shape)
Noise	x_4, x_5, x_6	**Zero Importance**	**Non-Zero / Leaked**

Observation: As hypothesized, both the Scratch EBM and the Library EBM assigned non-trivial importance scores to the noise features. Instead of being flat lines at zero, the learned shape functions for $x_4 \ldots x_6$ exhibited random fluctuations (overfitting).

However, the interaction detection remained robust: both models correctly identified (x_1, x_2) as the Rank 1 interaction, ignoring the noise features during the pairwise ranking phase. This confirms that the interaction engine is ready for 3-way extension, while the main-effect learner is primed for the sparsity constraint implementation.

References

[1] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.