1. Null and Alternative hypotheses.

A null hypothesis (H0): MSE, MAPE (MetaModel1 || MetaModel2) >= MSE, MAPE (Benchmark)

An alternative hypothesis (Ha): MSE, MAPE (MetaModel1 || MetaModel2) < MSE, MAPE (Benchmark)

**The Level of Significance**: 5% (α = 0.05).

(b, c) The tests used and why? And The results of each test.

Note: most of the tests are determined after deciding that the data is non-parametric (not normally distributed).

1. Run two statistical test that consider the three models together.

1. Ordering data without magnitude of difference.

**Why the Friedman Test:**

* The **Friedman test** is a non-parametric (it does not make any assumptions about the distribution of the data) statistical test that evaluates **if there are significant differences between the rankings of several models across multiple datasets.**
* I used it to test whether there are significant differences in **MAE** and **MAPE** across three models: **Benchmark**, **MetaModel1**, and **MetaModel2**.
* The Friedman test **does not** consider the magnitude of differences in errors, but instead **compares the ranks** of the models' performance across datasets.

**Resul**t:

FriedmanchisquareResult(statistic=11.0322580645161, p-value=0.004021384501613288) FriedmanchisquareResult(statistic=11.0322580645161, p-value=0.004021384501613288)

* + Since both **p-values < 0.05**, it means that there are **significant differences** between the models in terms of their rankings for both **MAE** and **MAPE** across the datasets.
  + we can **reject the null hypothesis** that all three models perform similarly in rankings.
  + This suggests that at least one model performs significantly differently from the others in the rankings.

2. Magnitude of error difference.

There were limited options for statistical test for more than 2 groups that considers Magnitude of error difference.

I tried the **Permutation Test** and **Repeated Measures ANOVA** , although **Permutation Test** is better due to the distribution of the data but, it have that there is no significant difference, while i used **Repeated Measures ANOVA** and it's supposed to be for normally distributed data but, according to other tested i made, it gave there is a significant difference and its similar to the other tests.

WHY?

* + **Repeated Measures ANOVA** relies on several assumptions, such as normality, sphericity (homogeneity of variances), and independence of observations. When these assumptions are violated (as is the case with your non-normal data), the test can produce **misleading results**, particularly an **inflated Type I error rate**. This means that ANOVA might **overstate the significance** of differences when the assumptions aren’t met.
  + The **Permutation Test**, being a **non-parametric test**, doesn't rely on these assumptions. It directly compares the observed differences in the data with randomly permuted versions. As such, the **Permutation Test** is more reliable for **non-normal data** like yours.

Results:

**Permutation Test** (Non-parametric)

Observed MAE difference: 57759.24200696599

Permutation test p-value for MAE: 1.0

Observed MAPE difference: 2168.097651776

Permutation test p-value for MAPE: 0.999

* + Both **MAE** and **MAPE** differences between the three models are **not significant**. The **high p-values** (1.0 for MAE and 0.999 for MAPE) indicate that there is no strong evidence that the models perform differently in terms of MAE or MAPE.
  + Fail to reject null hypo.

**Repeated Measures ANOVA** (Parametric)

**MAE Results**: **p-value = 0.7763**, meaning that there is **no significant difference** in MAE across the three models.

**MAPE Results**: **p-value = 0.0180**, meaning that there is a **significant difference** in MAPE between the models.

The **ANOVA result** suggests that there is a statistically significant difference in **MAPE** between the models, based on the distribution of the data it's not accurate.

**Conclusion**:

* Since **Permutation Tests** are **distribution-free** and the data is **not normally distributed**, we should **trust the results from the Permutation Test** over those from ANOVA.
* The **Permutation Test's non-significant result** suggests that the observed differences between the models could simply be due to random chance, and you should be cautious about over-interpreting ANOVA's significant result.

2. Run two statistical test that consider the three models together in pairs.

1. Ordering data without magnitude of difference.

**Why Friedman and Nemenyi Post-Hoc test?**

**Where** the **Friedman Test** is a **non-parametric alternative** to repeated measures ANOVA, The Friedman test ranks the performance of the models across the datasets without considering the magnitude of the differences **as we mentioned before,** However, the Friedman test only tells you that **at least one model is different**. It doesn’t tell you **which specific models** are different from each other. The **Nemenyi Post-Hoc Test** does pairwise comparisons to show **which specific pairs of models** differ significantly.

* **After the Friedman test** shows a significant result (p-value < 0.05).
* To determine **which pairs of models** are significantly different from each other.
* The **Nemenyi test** compares the differences in ranks between all pairs of models to see if the differences are statistically significant.

Results:

**MAE, MAPE Friedman Test AS before**: **p-value = 0.004**. Since **p-value < 0.05**, it indicates that there is a **significant difference** in the ranks of **MAE** across the three models.

**MAE** and **MAPE**, the Nemenyi post-hoc test results provide p-values for pairwise comparisons between the three models:

| **Models Compared** | **MAE p-value** | **MAPE p-value** |
| --- | --- | --- |
| **Benchmark vs MetaModel1** | **0.0209** (significant) | **0.0209** (significant) |
| **Benchmark vs MetaModel2** | **0.900** (not significant) | **0.900** (not significant) |
| **MetaModel1 vs MetaModel2** | **0.0065** (significant) | **0.0065** (significant) |

Conclusion:

In terms of **ranking**, **MetaModel1** consistently ranks differently compared to both **Benchmark** and **MetaModel2** in terms of **MAE** and **MAPE**, meaning we can reject null hypo.

**Wilcoxon Signed-Rank Test** (**non-parametric)**

**Why?**

It is an appropriate choice as its used to compare **paired data** and assess whether the ranks of the **differences** between paired observations is statistically significant.

**Result:**

* + **MetaModel1 vs Benchmark*:* (MAE Statistic = 120.0**, **p-value = 0.0110*), (*MAPEStatistic = 99.0**, **p-value = 0.0027**.**)** The p-value < **0.05**, meaning that the **magnitude of error difference** in MAE between **MetaModel1** and **Benchmark** is **statistically significant**.
  + **MetaModel2 vs Benchmark**: **(MAE** **Statistic = 212.0**, **p-value = 0.4919**), **(MAPE statistic = 241.0**, **p-value = 0.9001** the p-value is **greater than 0.05**, meaning that there is **no significant difference** in MAE between **MetaModel2** and **Benchmark**.
  + **MetaModel1 vs MetaModel2**: **(MAE** **Statistic = 103.0**, **p-value = 0.0036**), (**MAPE** **Statistic = 110.0**, **p-value = 0.0058**) The p-value is **less than 0.05**, showing a **significant difference** in MAE between **MetaModel1** and **MetaModel2**.

### Conclusion:

* **MetaModel1** stands out from both **Benchmark** and **MetaModel2** with statistically significant differences in both MAE and MAPE.
* **MetaModel2** performs similarly to the **Benchmark**, showing no significant differences in error magnitudes.
* **We can reject null hypo.**

2. Magnitude of error difference.

**Bootstrap Test and Permutation Test are non-parametric,** both tests evaluate the **actual magnitude of differences** in error metrics and are designed to compare two models at a time.

**Bootstrap Test**

**Result:**

* **MSE:** None of the pairwise comparisons show a **statistically significant** difference in MAE between the models (p-values > 0.05), indicate that the models perform similarly in terms of MAE.
* **MAPE**: the p-values suggest that this difference is **not statistically significant**.

**Permutation Test**

**Result:**

* **MSE:** None of the pairwise comparisons show a **statistically significant** difference in MAE between the models (p-values > 0.05), indicate that the models perform similarly in terms of MAE.
* **MAPE**: The only statistically significant result is for **Benchmark vs MetaModel1** (p-value = 0.0157), indicating that **MetaModel1** performs significantly better than **Benchmark** in terms of MAPE (lower error).
* **Fail to reject null hypo in Bootstrap Test**, but we can reject it in **Permutation Test.**

**Conclusion**:

**MetaModel1** shows some promise in improving performance over the Benchmark in terms of **MAPE** but **all models perform similarly** in terms of **MAE** (absolute errors).