Investigation of Marketing Mix Models' Business Error using KL Divergence and Chebyshev's Inequality

R. Venkat Raman	Ridhima Kumar	Pranav Krishna
venkat@arymalabs.com	ridhima@arymalabs.com	pranav@arymalabs.com
Aryma Labs Pvt. Ltd	Aryma Labs Pvt. Ltd	Aryma Labs Pvt. Ltd

Abstract

This report is an investigation into the trade-offs between Predictive Accuracy and Business Impact in Robyn, which uses the Nevergrad algorithm for optimizing (Normalized RMSE) and Business Error (Decomposed Residual Sum of Squares) as independent objectives. We examined models with the best and the worst Decomp.RSSD on the Pareto frontier, analyzing their performance through the lenses of Kullback-Leibler (KL) Divergence and Chebyshev's inequality. The aim is to explore how these models balance the dual objectives of error minimization and "Business Impact", highlighting the complexity of selecting the "Best" model when considering both statistical alignment and business relevance. Analysis revealed unexpected trends between the Best and Worst models in terms of KL Divergence and error clustering, highlighting a trade-off between minimizing business error and maintaining predictive accuracy, and pointing to the need for a nuanced model evaluation approach, and a delicate hand in choosing the final model.

Keywords: Marketing-Mix Modelling Robyn Decomp.Rssd NRMSE Pareto Frontier KL Divergence Chebyshev inequality

1. Introduction

Robyn utilises a multi-objective optimization algorithm nevergrad that simultaneously minimizes the Normalized Root Mean Squared Error (NRMSE) and the Business Error, leveraging a Decomposed Residual Sum of Squares (Decomp.RSSD) approach. The final models constructed by Robyn are those at the Pareto Frontier, i.e. the curve where the only way to improve one objective is to give up something for the other objective

This report examines the trade-offs at both the optimal and worst points of the Decomp.RSSD ("Business errors") on the Pareto frontier using the lens of Kullback-Liebler (KL) Divergence and using the error bounds given by the Chebyshev's Inequality.

2. Literature Review

Market Mix Modeling (MMM) is a technique which helps in quantifying the impact of several marketing inputs on sales or Market Share. The purpose of using MMM is to understand how much each marketing input contributes to sales, and how much to spend on each marketing input. MMM helps in the ascertaining the effectiveness of each marketing input in terms of Return on Investment. In other words, a marketing input with higher return on Investment (ROI) is more effective as a medium than a marketing input with a lower ROI. MMM uses the Regression technique and the analysis performed through Regression is further used for extracting key information/insights.

Initially developed in the 1960s, MMM has evolved significantly with advances in computing power and data availability. Early models were simple regression analyses. Over the decades, these

have transformed into more complex econometric and machine learning models that can handle multivariate scenarios across different media and sales channels. This is primarily shown by the increasing complexity of Adstock transformations, as discussed in Joseph (2006) and Fry et al. (1999). The recent developments in this field correspond to the application of machine learning and Big Data.

We have Meta's open source library for MMM, called "Robyn" Zhou et al. (2024), which uses machine learning techniques and evolutionary algorithms to estimate and forecast the effects of various marketing inputs on sales. It automates much of the model-building process, which not only speeds up the analysis but also makes it more accessible to a broader range of users. This is particularly significant in a marketing landscape that is increasingly data-driven and segmented across numerous digital platforms.

Moreover, the integration of machine learning techniques has enabled the handling of large-scale data sets and the incorporation of real-time data into MMM models. This shift allows for near-continuous optimization of the marketing mix, responding to changes in consumer behaviour and market conditions in almost real time. These advancements have greatly improved the precision of attribution models, which can now more effectively measure the impact of specific marketing activities across different channels and customer touchpoints.

As a result, organizations can tailor their marketing efforts more effectively, optimizing budget allocations and strategic decisions in real-time to maximize Return on Advertising Spend (ROAS).

3. Definitions

In this section, we elaborate upon the definitions of NRMSE, Business error as implemented in Meta's Robyn library, Pareto Frontier and KL Divergence. In this report, the "Best" model corresponds to the Model with the minimum business error while being at the pareto frontier, and the "Worst" model corresponds to the maximum Business error, whilst being on the frontier.

3.1. NRMSE

We have that NRMSE is the usual Mean Square Error, normalized by the range of the observations. Thus,

$$\text{NRMSE} = \frac{\sqrt{\text{mean}\left((y - y_{\text{pred}})^2\right)}}{\max(y) - \min(y)}$$

Observe that $\text{NRMSE}^2 \propto \Sigma_i (y_i - y_{\text{pred}})^2$, thus optimising NRMSE alone would be equivalent to Linear Regression because the normalization is equivalent to subtracting a constant in the standard LogLikelihood optimization that utilizes derivatives

3.2. Business Error

The definition of Business Error is more convoluted. It was developed and implemented by Meta, and insofar as it is available in the code, it is defined in terms of changes from the history of the Budget allocations and spending shares.

Again, from the code, we see that when we model using Robyn for the first time(initial setup and run), it is the square root of the Residual Sum of Squares between the effect share attributable to the channel and the corresponding spends' share, which corresponds to institutional knowledge being encoded in the historical spends. Thus, we have

$$\text{Decomp.RSSD} = \sqrt{\sum_{i=1}^{N} (e_i - s_i)^2}$$

where e_i is the effect share attributable to the channel for the ith observation, s_i is the spend's share, and the sum runs over all N observations.
After this first instance, Robyn utilizes the square root of the Residual Sum of Squares of effect attributed to each media variable or channel, expressed as a percentage, before and after the update of the model with newer data. Thus, we have the updated formula as follows -

$$\text{Decomp.RSSD} = \sqrt{\sum_{i=1}^{N} \left(\frac{e_{i,\text{ after}} - e_{i,\text{ before}}}{N} \right)^2}$$

where $e_{i,\text{ after}}$ and $e_{i,\text{before}}$ represent the effect share attributable to a media variable or channel for the i^{th} observation after and before the model update, respectively and the sum runs over all N observations. There is a similar term for nonmedia channels, with a weighting for time delay.
Lastly, we remark that there are some other design and implementation details, like having a penalty for models which have zero effect shares, and the possibility of incorporating weighting for non-media promotional factors. These are available in the code on GitHub at the link for the definition of Decomp.RSSD online.

3.3. Pareto Frontier

The Pareto frontier represents the set of all possible decisions that optimize two or more conflicting objectives. A point x^* is at the pareto frontier for minimising the objective functions $f_1, f_2, \ldots f_n$ if there is no point x such that -

- $\mathrm{f_i(x)} \geq \mathrm{f_i(x^*)}$ for all $\mathrm{i} \in \{1, \ldots, \mathrm{k}\}$ and
- $\mathrm{f_j(x)} > \mathrm{f_j(x^*)}$ for at least one $\mathrm{j} \in \{1, \ldots, \mathrm{k}\}$.

Thus, no point on this boundary can improve one objective without worsening another, making it a good tool to analyse trade-offs and determine the most efficient choices in multi-criteria optimization problems.

3.4. KL Divergence

KL divergence, or Kullback-Leibler divergence, is a measure from information theory that quantifies how much one probability distribution diverges from a second probability distribution. It measures the information lost when using one distribution to approximate another and is used to compare differences between two probability distributions. Though often used to measure the similarity between two distributions, it's important to note that KL divergence is not symmetric, meaning the divergence from P to Q is not the same as from Q to P. Lastly, we remark that a procedure which minimises the KL divergence is equivalent to the Ordinary Least Squares procedure in Regression.[1]
The formula in the discrete case for KL Divergence of P (base distribution) to Q (target distribution) is as follows:

[1] We remark that the Cross entropy and KL Divergence differ by a constant for a given base distribution. Furthermore, we know that optimizing the Cross Entropy Loss is equivalent to Ordinary Least Squares cro.

$$D_{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

We remark that in our case, we do not have access to the underlying distribution, and only have access to the samples. Thus, we need a non-parametric estimator for KL divergence
which preferably does not rely on accurate density estimation. Such an estimator was proposed by Boltz et al. (2009). We report that based on experiments, we see a Bias Variance trade-off as more and more neighbors are considered, with more neighbors considered leading to a lower variance higher bias scenario.

4. Methodology

We utilise the default data set in Robyn, called "dt_simulated-weekly" and utilise the outputs from the demo file available in the repository. This runs 5 instances of 2000 iterations of the optimization and allows us to choose the final model from the collection based on our judgement.
We evaluate the optimality of the "Best" fit and the "Worst" fit Business error models by analyzing the Kullback-Leibler divergence and applying the Chebyshev's inequality. The Chebyshev's inequality is used for investigating the proportion of errors corresponding to the data which is clustered near the zero-prediction error.
We give a few details about the estimator for KL Divergence for reference. The estimator utilizes a kNN framework for estimation of KL Divergence while side-stepping the need for estimating the density. As the number of neighbours increases, we see a bias-variance tradeoff, and the recommendation from the authors was to take an average of a few estimates (with varying number of neighbours being considered).
Secondly, we do not have the usual assumption that the underlying errors for the model are normal. Thus, we try to investigate the clustering of the errors using the Chebyshev's inequality. It is -

$$\mathbf{P}\left(\frac{|R_i-\mu_i|}{\sigma}\geq k\right)\leq\frac{1}{k^2}$$

where R_i is any random variable with a finite second moment and the symbols have their usual meanings. We know that the errors in OLS have expectation zero. Furthermore, they are assumed to be iid copies from an underlying distribution (which is standard normal in the 'ideal' case). In our case, we assume that the underlying distribution of the residuals after fitting both the models has a mean 0 and some finite standard deviation, which is estimated from the sample. Because we have N samples from this distribution, we can have an empirical estimate of the probability that

$$\mathbf{P}\left(\frac{|R_i-\mu_i|}{\sigma}\geq k\right)$$

by calculating

$$\widehat{\mathbf{P}}\left(\frac{|R_i-\mu_i|}{\sigma}\geq k\right):=\frac{\sum_{i=1}^{N}\mathbf{1}_{\left\{\frac{|R_i|}{\sigma}>k\right\}}}{N}$$

for all values of k. We can compare this empirical estimate from the theoretical bound of $\frac{1}{k^2}$ for values of k ≥ 1.

5. Results

5.1. Pareto Frontier

Robyn runs 5 unrelated iterations of the optimization problem and reports results from the collection. In our case, it happened so that the best and worst models were from the 2nd and the 1st iteration, respectively. The frontier shown in Figure 1 is only for representational purposes and does not correspond to the actual placement of Worst and Best model in the pareto frontier.

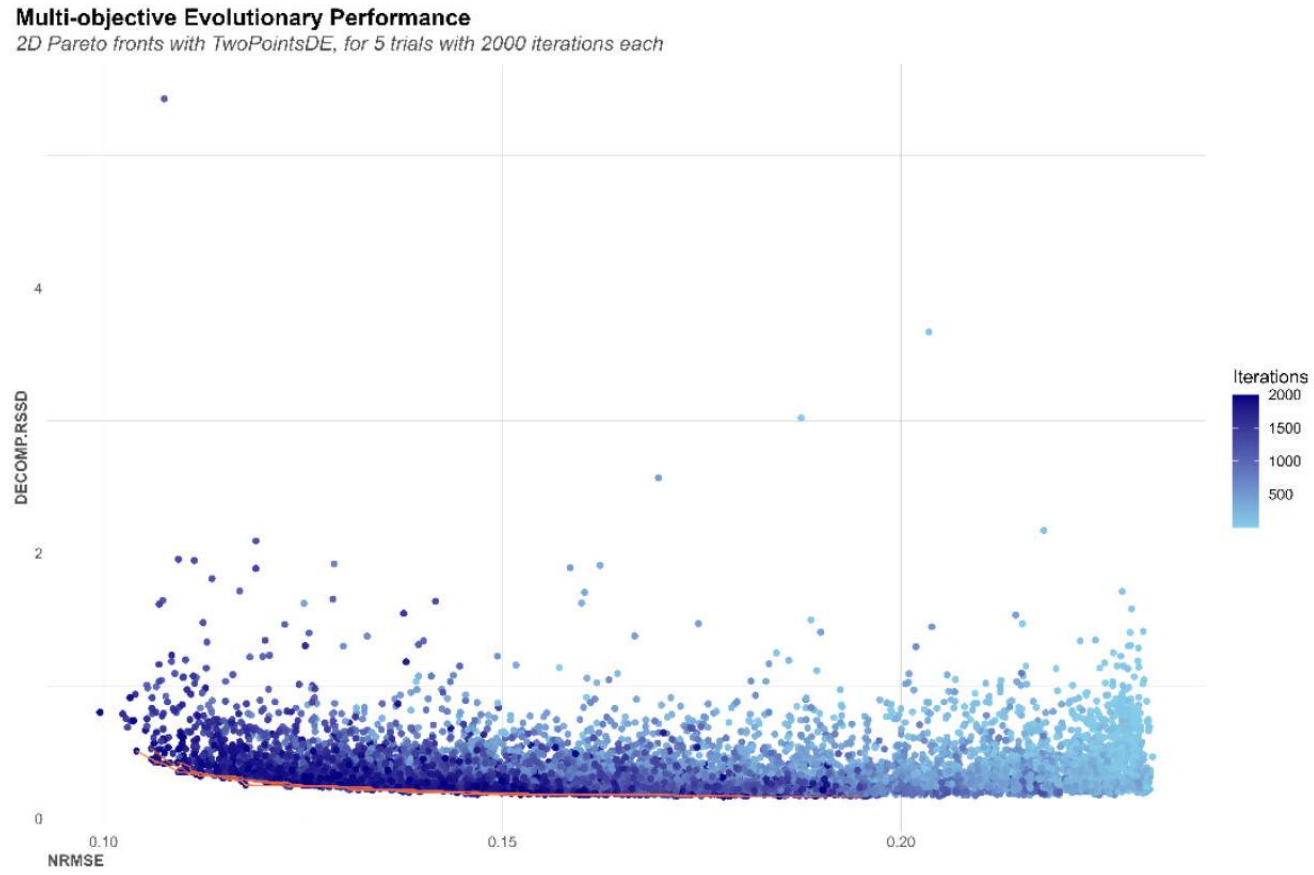

Figure 1: Representational Pareto Frontier for Decomp.Rssd and NRMSE

5.2. KL Divergence

First, we investigate the KL Divergence, and use the true values as the base distribution and the fitted values from both the models as the target distribution. We see from Table 1 that quite a lot of the values for KL Divergence are calculated to be negative (which contradicts the non-negativity from definition). Thus, any results we make using this data would be invalid, including results from taking a symmetrized KL Divergence, the Jeffrey's Divergence. We choose to interpret these negative estimates as suggesting that the modelled values are a close fit to the true values, especially for the Worst model

Thus, we change the order, to try to investigate the results of taking the true values as the target distribution and the modelled values as the base distribution. The results are shown in Table 2.

We see a few negative values, but the situation is much better than Table 1. There is a negative value for 8 and 10 neighbours for the Best model and for 3 neighbours for the Worst model.

Number of Neighbours	KL Divergence of the Best Model	KL Divergence of the Worst Model
1	0.138248	-0.113087
2	0.054990	0.148466
3	-0.060501	-0.030575
4	-0.067573	-0.064147
5	-0.056383	-0.013524
6	-0.034030	-0.028425
7	-0.000932	-0.044221
8	0.001084	-0.025689
9	0.021471	-0.045782
10	0.012554	-0.057792

Table 1: KL Divergence values with the true values as the Base Distribution

We observe from Figure 2, that the estimates seem to become stable from taking 20 neighbours for the Best Model and 24 neighbours for the Worst Model. Thus, for our purposes, we chose to take the mean of the estimates from 24 neighbours onwards to 30 neighbours to get the result, which is shown in Table 3. Clearly, we observe that the KL Divergence for the Best model is around 1.5 times as large compared to the Worst model from Table 3.

Number of Neighbours	KL Divergence of the Best Model	KL Divergence of the Worst Model
1	0.051475	0.149700
2	0.155777	0.006668
3	0.130832	-0.007519
4	0.195601	0.074805
5	0.185322	0.031313
6	0.074662	0.022191
7	0.034810	0.022777
8	-0.003185	0.040744
9	0.007811	0.026544
10	-0.015118	0.025024

Table 2: KL Divergence values with the true values as the target distribution

We see that the KL Divergence for the Best model is 3 times as large compared to the Worst model. Lastly, we remark that as soon as the estimator starts to give reasonable values for KL Divergence, we notice that the KL Divergence of the Best Model is significantly larger than the KL Divergence of the worst model, i.e., the red points are systematically above the blue points after 15 neighbours. This is further shown by the trend in the curve shown by loess smoothing, with the red curve being above the blue curve after 15 neighbours being considered

Model	KL Divergence (Model $\|$ True Values)
Best	0.0802407
Worst	0.0573257

Table 3: Results of the KL Divergence analysis

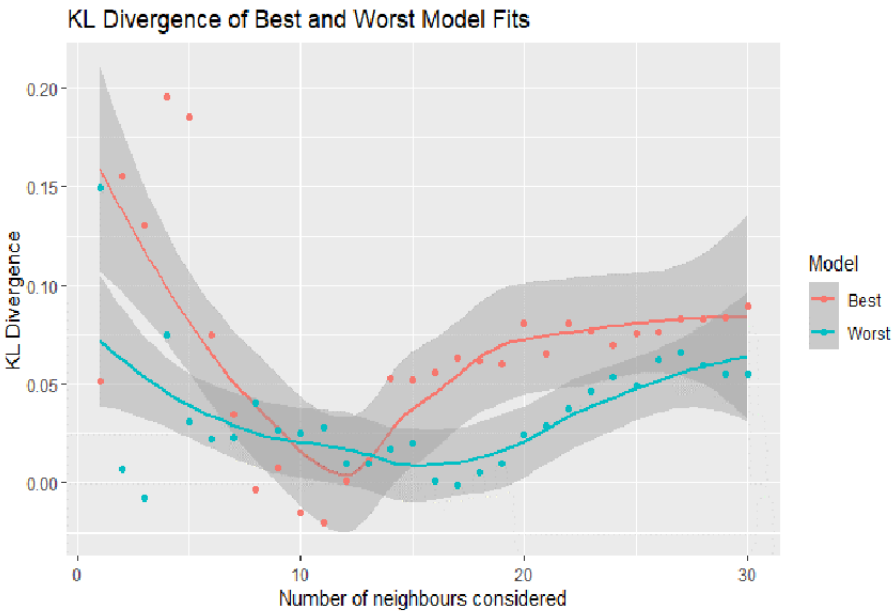

Figure 2: KL Divergence values with true values as the target distribution

5.3. Clustering of Results and Chebyshev's Inequality

From Figure 3, we see comparable performance from both models with respect to the distribution of the residuals. From Table 4, we see that for the best model, around 75% of the absolute standardized errors were beyond 0.25σ from 0 for the best model, while only 65% of the absolute standardized errors were beyond 0.25σ from 0 for the worst model.

On comparing the proportion of the residuals being close to zero, in that they are within two standard deviations (enclosed by green lines in Figure 3), we see from Table 4 that approximately 96% of the data is within 2σ for the best model whereas it is 95.5% for the worst model.

Despite the worst model having a higher percentage of residuals up to one standard deviation, the best model demonstrates its superiority by having a higher percentage of residuals beyond a 1σ threshold. This shows that while the worst model is slightly better at predicting typical cases, the best model is more consistent across the range of predictions it makes, except possibly in the most extreme cases.

Overall, Table 4 suggests that both models' residuals adhere to the expected theoretical bounds. Notably, the worst model tends to have more residuals closer to zero than the best model, indicating better performance on less unusual data points.

However, the lack of data at the far extremes (beyond 2σ) makes it difficult to assess the models' performance in those regions. It would require more data points at these extremes to fully understand the models' behavior in tail events. Without such data, any conclusions about extreme performance would be speculative.

This unexpected behavior of the residuals from the Best and Worst models indicates a complexity in model behavior and points to the importance of considering both NRMSE and Decomp.RSSD when evaluating model quality. Thus, besides saying that that both the models satisfy and validate the Chebyshev's bounds, we see also a trade-off between the prediction accuracy using NRMSE and the Business error from this table.

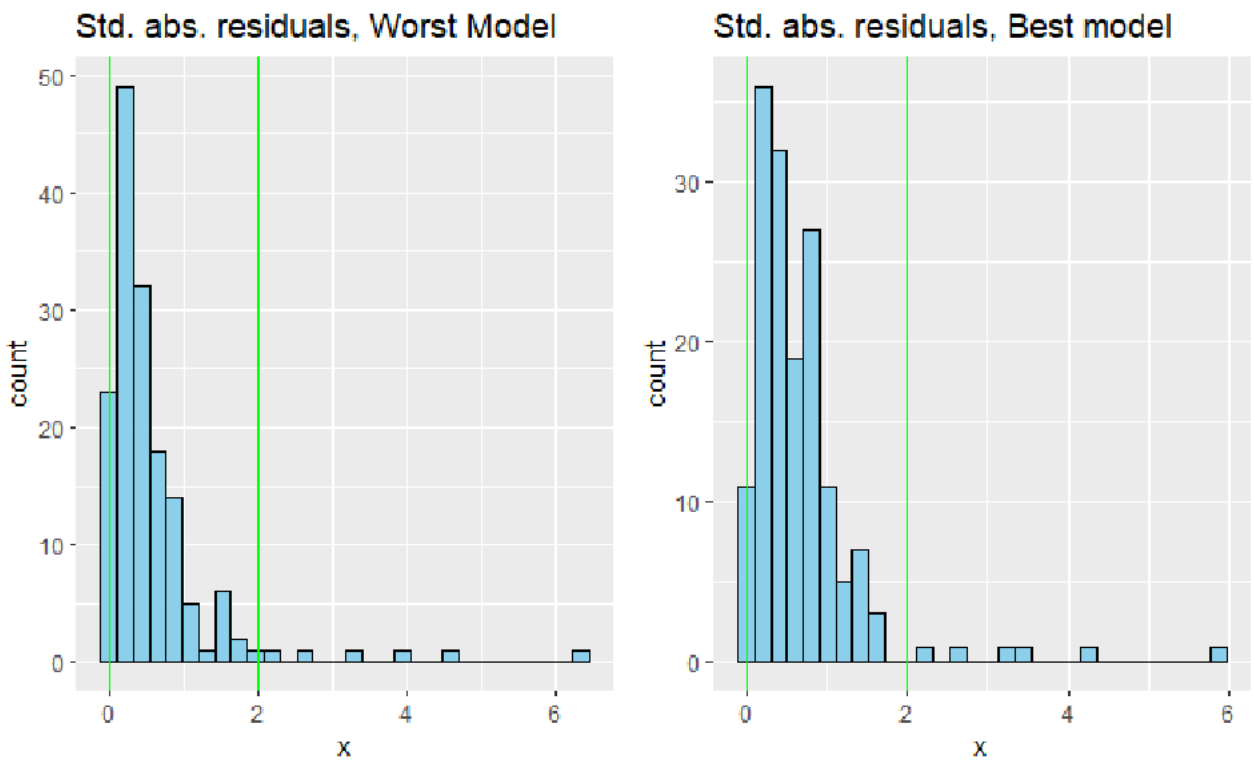

Figure 3: Standardised Absolute Residuals for both the Models

5.4. Reporting Decomp.RSSD

We report the Decomp.RSSD from the Best and Worst models on the Pareto Frontier in Table 5. While the NRMSE is comparable for both models, there is a stark increase in the Decomp.RSSD for the Worst Model compared to the Best Model, around 60 times as much.

6. Conclusion

We observe complex behaviour regarding the trends exhibited by the models, in both KL Divergence and Chebyshev's inequality style clustering. The expected results were that both the Models would satisfy the Chebyshev bounds, and the best model would outperform the Worst model in all metrics. But there was a surprising reversal. The fact of the matter is that the KL Divergence was consistently estimated to be higher for the best model compared to the Worst model, at least when the true values were taken as the target distribution. This must also be contrasted with the 60 times larger Decomp.RSSD for the Worst model, compared to the best model.
This underscores the complexity in defining the most appropriate model, emphasizing the need for a balanced approach that considers both business impacts and alignment with the data distribution. Thus, a delicate hand is needed to choose the final model, which balances both the objectives.

7. Future Work

While we only investigated the singular Best and Worst Models with respect to Business Error, it would also be interesting to conduct a sensitivity analysis to report the trade-off more congruently between the NRMSE and Business Error. This might be done by choosing the 5 best and

Region denoted by multiples of σ	Best Model (%)	Worst Model (%)	Theoretical Bound (%)
> 0.25	75.1592	65.6051	100.0000
> 0.5	50.3185	38.2166	100.0000
> 0.75	34.3949	22.2930	100.0000
> 1	15.2866	13.3758	100.0000
$> 1.41 \sim (\sqrt{2})$	7.64331	9.55414	50.0000
> 1.5	6.36943	7.00637	44.4444
> 1.75	3.82166	4.45860	32.6531
> 2	3.82166	4.45860	25.0000
> 3	2.54777	2.54777	11.1111
> 4	1.27389	1.27389	6.2500
> 5	0.636943	0.636943	4.0000

Table 4: Clustering of Residuals and Chebyshev's Bounds

Model	Decomp.RSSD	NRMSE
Best	0.006189742	0.5837057
Worst	0.3770549	0.4588901

Table 5: Values of Decomp.RSSD and NRMSE

Worst models corresponding to Decomp.RSSD, or by choosing the optimum of an appropriate two-dimensional function of the two objectives. We are also doing this analysis on a client data, and the results will be shared soon in another paper

8. Appendix

The codes and the data for recreating the Tables and the FIgures in the paper are available at the Github repository.

References

Using cross-entropy for regression problems. https://stats.stackexchange.com/questions/477152/using-cross-entropy-for-regression-problems. Accessed: 24 April 2024.

Sylvain Boltz, Eric Debreuve, and Michel Barlaud. High-dimensional statistical measure for region-of-interest tracking. IEEE Transactions on Image Processing, 18(6):1266–1283, 2009.

Tim RL Fry, Simon Broadbent, and Janine M Dixon. Estimating advertising half-life and the data interval bias. 1999.

Joy V Joseph. Understanding advertising adstock transformations. Available at SSRN 924128, 2006.

Gufeng Zhou, Igor Skokan, and Julian Runge. Packaging up media mix modeling: An introduction to robyn's open-source approach. arXiv preprint arXiv:2403.14674, 2024.