International Journal of Advances in Business and Management Research (IJABMR) https://doi.org/10.62674/ijabmr.2025.v2i03.004

https://ejournal.svgacademy.org/index.php/ijabmr/

OPEN ACCESS

ANALYSIS AND FORMULATION OF COMPENSATION STRATEGIES TO MITIGATE DATA ERROR COSTS BY INVESTIGATING VARIOUS ERROR CATEGORIES

Miao Congjin*, Muhammad Ezanuddin Abdul Aziz

Original Article

Lincoln University College, 47301 Petaling Jaya, Selangor D. E., Malaysia

Corresponding Author's Email: miaocongjin004@outlook.com

Abstract

Research on methods to standardise data error rates is critical because of the potentially disastrous consequences that such errors may have on businesses. In order to assist readers in developing an equalisation approach to handle data processing mistakes, this article explores the many shapes that these errors might take. The authors begin by differentiating between the two most prevalent types of data processing errors: random and systematic. Random errors, or mistakes that occur by chance, could be mitigated by using more exact measurement techniques or a bigger sample size. Systematic errors, on the other hand, occur often and may have several sources, such as faulty equipment, inaccurate calibration, or bias in the data collection process. In order to address systematic errors, the authors propose an equalisation technique that comprises identifying and correcting the erroneous data sources. The idea behind this approach is to look for patterns in the data that might indicate systemic issues and then to implement the appropriate fixes to mitigate such problems. Through a series of experiments utilising both theoretical and practical data, the authors demonstrate the efficacy of their equalisations approach. Data error rates were significantly reduced in these experiments using the equalisation approach, enabling more accurate and trustworthy findings. In sum, the article effectively lays out the many problems with data processing and how to resolve them by using an equalising approach. Data quality and decision-making skills may be improved, and error rates can be reduced, which can lead to increased performance and success for organisations.

Keywords: Error Sites; Equalization; Rates of Mistake; Remedial Measures; Systemic Inaccuracies

Introduction

Data preparation and analysis go hand in hand. Writing up the information gleaned from the data. Various methods, such as decision-making models, can assist in identifying patterns, correlations, and relevant findings. The data has to be prepared in advance in order to do the analysis, however [1]. The process of data preparation is altering information so that computers can read and process it. Made to be used with statistical programs like SPSS and SAS. Data preparation involves a series of steps, including data encoding, data entry, blank filling, and data reformatting [2, 3]. This section provides a concise overview of each of these stages as performed by a researcher:

Received 24th January 2025; Revised 14th February 2025; Accepted 27th February 2025

Coding of Data: To begin dealing with data, it must first be transformed from its unstructured form into a numerical form. In this case, a codebook—a compilation of several types of data—is used. The codebook comprises various components such as the response, variables, metrics, and variable format, along with a codicil that concludes the coding process. The types of scales are determined by the process's reaction. Think about things like the number of points (five, seven, etc.), the kind of scale (nominal, ratio, ordinal, interval), and so on. As an example, when it comes to numerical notation for business categorisation, healthcare is categorised at 1. The numbers 2 through 4 indicate different aspects of the economy: manufacturing, retail, and finance.

Inputting Coded Data: The exciting part is about to begin inputting all that coded data into a text file or spreadsheet. Adding it to the software suite is a breeze. There are gaps in the data due to individuals' decisions not to fill out the survey. The authors were unable to provide satisfactory responses to all enquiries; therefore, it's crucial to determine how to reevaluate these unfulfilled expectations. Some applications, for instance, need an additional -1 or 999. Although many of them automatically deal with missing data, a few of them use list-wise deletion. Method for dealing with missing values, whereby every set of answers is discarded when only one is absent. The data may need to undergo certain modifications before they can be understood. For example, items with reverse coding may need a change before they can be used. This is in contrast to when they are used alongside non-inverted components. When an object's meaning changes, this idea is employed. This idea is employed when the main point of the object changes. The following is a summary of the most common approaches of data analysis, as given in Types of Information Analysis. Because of this, there are primarily six forms of data analysis. Six basic types of analyses are used: descriptive, exploratory, inferential, predictive, explanatory/cause-and-effect, and mechanistic [4].

Descriptive: It is the simplest and most fundamental kind of data analysis, and it has been around the longest. Because of this, it performs well with very massive datasets. The data is then used in a data set analysis.

Exploratory: A method for generating new research topics or laying the groundwork for future studies by revealing previously unknown information and developing unexpected connections.

Inferential: The primary goal of inferential research is to derive conclusions about the whole population from a smaller sample. Evidence used to evaluate a cosmic theory originates from a minuscule fraction of Earth's surface. This technique works well with cross-sectional time series, historical data, and observational data.

Predictive: Research of this nature takes both the past and the present into account when making predictions. Furthermore, it may extrapolate the values of a second subject from the first person's data. A more straightforward strategy that maximises data use could be the one that succeeds, regardless of the number. This means that researchers need to plan how they will gather data for predictions and how they will evaluate their findings.

Rationale: Methodology centred on procedures. Analysing randomised trial data sets to determine whether changes in the variables may affect others is the most labour-intensive part of this strategy. It is also reasonable to assume that mechanical analysis is highly unlikely.

Since even a little error may have a significant monetary impact, it is a beneficial fit for areas like engineering and the physical sciences. The next part will provide a comprehensive overview of descriptive statistics, explanatory statistics, and inferential statistics, the three primary branches of statistical analysis.

Background of the Study

Collecting pertinent data and interpreting it are the backbones of any research project. Researchers conduct the research. Preparing data is making it usable by computers by converting it from an unstructured, raw format. It includes data reformatting, coding, inputting, and filling in voids. In contrast, data analysis is the process of extracting useful information from datasets by using a variety of methods to identify patterns, correlations, and other insights. Description, exploratory, inferential, predictive, explanatory/casual, and mechanistic evaluations are the six main ways to examine data [5]. At its most fundamental level, descriptive analysis is just a summary of the data collected. The purpose of doing exploratory analyses is to discover novel relationships and provide the groundwork for more comprehensive study. Applying inferential statistics to a smaller sample allows one to draw conclusions about the whole population. To foretell

what's to come, predictive analysis considers both the present and the history, while explanatory or causal analysis seeks to initiate events [6, 7]. Researchers use mechanistic analysis to pinpoint which changes in a single variable led to shifts in other variables. A crucial component of comprehensive analysis is the preparation of data summaries for easy presentation. This method is mainly divided into bivariate and multivariate subsets. Common univariate statistical methods that concentrate on a single variable include dispersion analysis, central tendency analysis, and frequency analysis. Two methods exist for examining data: analysis of frequency and analysis of central tendency. The former counts all possible values for a variable, whereas the latter determines measures of central tendency like the mean, median, and mode [8, 9].

Literature Review

This method may be used to summarise information that is hard to understand. One may argue that this method is bivariate or multivariate. The term "univariate" is often used to describe statistical procedures that use only one variable. Scientific methods such as Dispersion Analysis, Central Tendency Analysis, and Frequency Analysis will be used extensively. Researchers can find out where the relevant variable came from only by messing with its frequency. It generates a complete set of options by counting how often each value occurs for a certain variable. When comparing one variable to a series of data points, one may use the amount of the most represented value—also known as the three Ms—to evaluate the central tendency of the disturbance. The standard deviation, mode, and mean are common ways to measure central tendency. When the authors look at a set of numbers, the most common one is called the Mode, and when they average all the values, they get the Mean. The dispersion of a variable around its mean is defined by its dispersion. Standard deviation is a common metric in statistics; it is the product of the square roots of the variance, range, and variance. The two extreme numbers are clearly different from each other, as evidenced by the range [9, 10]. The degree to which the data points cluster around the mean may be seen by examining the variance. This approach may be used to compare two datasets that have two independent variables. Because of this, the link between the two variables may be identified by the scientific community. The most common statistic is bivariate correlation. The procedure for calculating the degree of correlation using this statistic takes into account the sample means and standard deviations. It retains its utility even when dealing with more than two variables. Software such as SPSS simplifies the process of solving such problems, whereas manual methods can be challenging. Assessment of Explanation As previously stated, the goal of doing an explanatory analysis is to uncover possible variables that might have had a role. By using explanatory analysis, they may answer concerns about patterns, correlations, and connections between variables. Explanation analysis is based on processes of dependency and interdependence. The idea of dependency refers to the possibility that several independent factors could have an effect on a single dependent variable. "Interdependence approaches" are a kind of multivariate analysis that seek to identify correlations between variables without presuming the strength or direction of any impact [1, 11, 12].

Research Methodology

If regression methods, such as Neural Networks or analogous predictive models, do not provide impeccable projections, overfitting may not transpire. Regardless of a researcher's diligence, their projections will always include some inaccuracies. to assess the results of several models, choose the most successful ones, and thereafter make an educated conclusion. Consequently, several measures for prediction error may be used. The estimated values are denoted by p, a N1 vector, whereas the calculated (or measured) and predicted values of a quantity are represented by r and p, respectively. For instance, researchers may train an ANN to provide predictions N times for each specific input. The projected values are symbolised by p, an N1 vector, while the calculated (as well as measured) and forecasted values of an item are highlighted by r and p, respectively.

Researchers may train an ANN (Artificial Neural Network) to provide N predictions for each individual input. So that an outsider can give an unbiased opinion, they might put the first set of results (S) next to a different dataset (N), a

chosen part of what was in the original database (T), or nothing at all (U). This comparison enables researchers to assess the performance and precision of the ANN model across several circumstances. By evaluating the correlation between anticipated values and actual observations, they may enhance their algorithms and augment prediction accuracy for future applications.

Field experts evaluate many metrics to assess the prediction inaccuracy of models, as shown in Table 1.

Table 1: Common Prediction Error Metrics

Metric	**Formula**	**Description**
Mean Absolute Error (MAE)	(MAE = \frac{1}{n} \sum_{i=1}^{n}	p_i - r_i
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2$	Penalizes larger errors more than MAE.
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{MSE}$	Provides an error measurement in the same unit as the variable
R-squared (R2R^2R2)	$R^2 = 1 - \frac{\sum(r_i - p_i)^2}{\sum(r_i - \bar{r})^2}$	Represents how well the model explains variability in the dataset

Source: Collected by Author

To ensure an impartial external assessment, researchers may juxtapose the initial data set (S) with an alternative data set (N), a subset of the original data set (T), or with no data at all (U). Experts will examine several measures to determine this model's prediction error. In this study, the researchers focus on the issue of continuous variables. Classification metrics include recall, accuracy, confusion matrices, and false positive rate. Note that the subsequent formulas need both observations and their forecasts to be positive. For models that make categorical predictions, recall, reliability, complexity matrices, and the percentage of false positives are some of the classification measures that are used. These indicators facilitate the evaluation of the model's performance by offering insights into its advantages and deficiencies. Through the analysis of these data, practitioners may make educated judgements on necessary model enhancements and tweaks to increase forecast accuracy. These are essential for assessing models that classify data into separate categories, as shown in Table 2. Researchers may need to modify some computations if their data includes negative integers or zeroes.

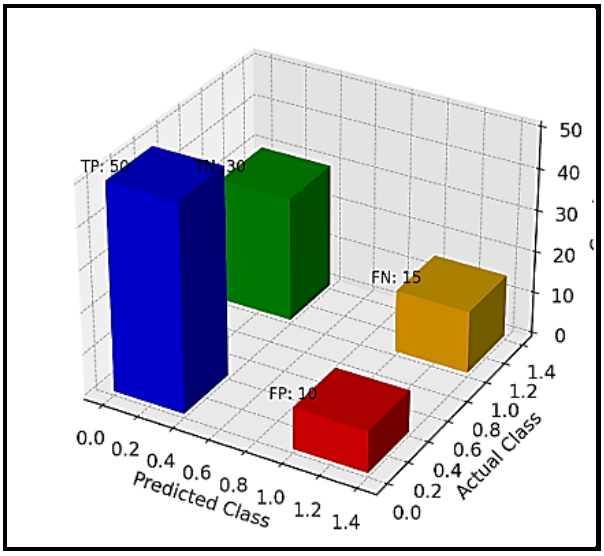
Table 2: Classification Performance Metrics

Metric	**Formula**	**Description**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Proportion of correctly classified instances
Precision	$\frac{TP}{TP+FN}$	How many predicted positives were actually positive?
Recall (Sensitivity)	$\frac{TP}{TP+FN}$	Ability to capture all actual positives.
Recall (Sensitivity)	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$	Harmonic mean of precision and recall.
F1-Score	$FPR = \frac{FP}{FP+TN}$	Probability of falsely classifying a negative as positive.

Source: Collected by Author

This paper emphasises the essential prediction error metrics used in ongoing, continuous classification models. The figure 1 and tables presented above demonstrate its use, enabling researchers to accurately assess model performance. The discourse explores the ramifications of these measurements, providing insights into their potential to enhance model correctness and dependability. By comprehending the intricacies of prediction mistakes, researchers may enhance the quality of their analysis and applications.

Figure 1: Real (Target) Model Performance



Source: Collected by Author

$$e_i = p_i - r_i \qquad (1)$$

$$MB = \bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i = \frac{1}{N}\sum_{i=1}^{N}(p_i - r_i) = \bar{p} - \bar{r} \qquad (2)$$

Where $\bar{p}$ and $\bar{r}$ are the mean values of p and r, respectively:

$$\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i \qquad (3)$$

$$\bar{r} = \frac{1}{N}\sum_{i=1}^{N} r_i \qquad (4)$$

Even though MB=0 is needed for actual and predicted values to be perfectly aligned (like having the same number), it can be reached even when correlations aren't perfect by cancelling out negative and positive errors. The "Mean Absolute Gross Error (MAGE)" quantifies the average error over several forecasts, disregarding their directional bias. Thus, it computes the total discordance between the anticipated and actual values in the test sample, using a weighted average for this purpose. The value of the variable, which may be either positive or negative, is

$$MAGE = \frac{1}{N}\sum_{i=1}^{N}|e_i| = \frac{1}{N}\sum_{i=1}^{N}|p_i - r_i|$$

A prevalent metric used in regression analysis is the Mean Squared Error (MSE). It represents the mean squared deviation from the actual value. Per the definition, it may possess positive or negative meanings.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(p_i - r_i)^2$$

A significant drawback of MSE is its incapacity to address severe situations. If the error for a single sample significantly exceeds the errors of other samples, the square of that error will increase markedly. Outliers may significantly influence the Mean Squared Error due to its averaging of discrepancies. A lot of people use the Root Mean Squared Error (RMSE) to figure out how well an estimate or model can predict things based on real data, like demographic or sample-based data. Researchers may ascertain this by squaring the mean squared error. The relative standard error (RMSE) is preferable to the mean squared error (MSE) when the units of the target variable are not aligned. The equation for this real number, which spans from zero to one (or from zero to infinity), is

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - r_i)^2}$$

One measure of this is the Centre-Mean-Square Distinction (CMSD).

$$CMSD = \frac{1}{N}\sum_{i=1}^{N}\left[(p_i - \bar{p}) - (r_i - \bar{r})\right]^2$$

When the focus attribute and the CMSD are represented in the same units, their square root is termed CRMSD, which denotes Concentrated Root Mean Square Differential.

$$CRMSD = \sqrt{CMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(p_i - \bar{p}) - (r_i - \bar{r})\right]^2}$$

Turner diagrams are a method for illustrating the accuracy of a model's prediction by using the CRMSD value; more elaboration will follow. Standardised bias error values are sometimes represented as a percentage, referred to as the mean index bias (MNB, unitless).

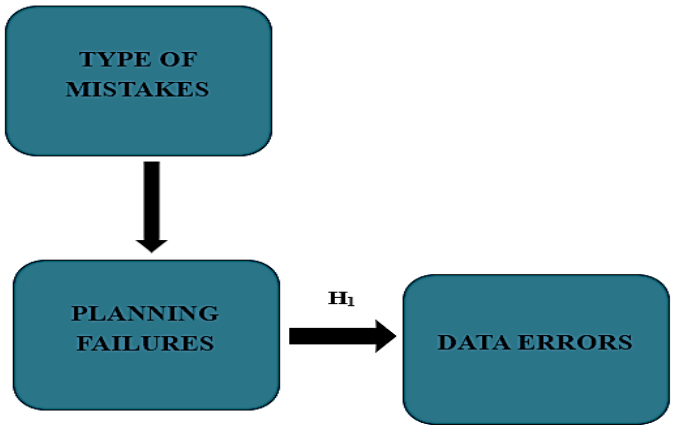
$$MNB = \frac{1}{N}\sum_{i=1}^{N}\frac{p_i - r_i}{r_i} = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{p_i}{r_i}\right) - 1$$

The Unitless Mean normalised Gross Error (MNGE) is frequently referred to as the "Mean Absolute Percentage Error." Knowledge of the error magnitude will provide readers with a clearer understanding of the accuracy of the estimates. It originates from

$$MNGE = \frac{1}{N}\sum_{i=1}^{N}\frac{|p_i - r_i|}{r_i}$$

A significant issue with MSE is its inability to accommodate numbers beyond the specified range. If a sample's associated error is much bigger than that of other samples, its square will also be considerably larger. The Mean Squared Error (MSE) is susceptible to outliers due to its tendency to average researchers' inaccuracies.

Conceptual Framework



Independent Variable:
Data collection, processing, and interpretation are all susceptible to a broad range of errors. An outlier, a random, or a systematic mistake might occur.

Dependent Variable:
Data Error: Data errors or missing information may affect the accuracy and reliability of conclusions or judgements.

Framework:
Recognise that data collection, processing, and analysis are all susceptible to human error. Many kinds of data errors should be classified and measured. Researchers can get a decent sense of the data's accuracy by finding the root-mean-squared error, relative error, or mean absolute error.

- To study the connection between various mistake types and data errors, use statistical methods such as regression or correlation.

- Find out whether there are any other factors, called confounding variables, that could influence the connection between the independent and dependent variables.

- When developing a model to forecast the number of data errors, consider the nature and frequency of various error categories. Researchers may gauge the efficacy of their models by using suitable metrics such as R-squared or mean squared error. Based on the analysis's findings, determine the most common data issues and come up with solutions. In conclusion, this approach involves cataloguing the many forms of data processing, analysis, and collecting errors and investigating the correlation between data mistakes. By looking at methods to reduce data inaccuracy, this research might help make data-based judgements and evaluations more reliable and accurate.

Results

Factor Analysis:
The paradigm for investigating the relationship between error types and data mistakes suggests many possibilities, including:

Hypothesis (H_1): The kinds of errors made while gathering, processing, or analysing data greatly affect how often the data is inaccurate.

Hypothesis (H_2): Systematic errors and outlier errors are more significant contributors to inaccurate data than random mistakes.

In this article, the researchers look at the hypothesis (H_1) that states that inaccuracies are more common when certain types of errors are made during data collection, processing, or analysis.

Null hypothesis (H_0): Data inaccuracies are common regardless of the kind of errors made during data collection, processing, or analysis.

Alternative hypothesis (H_1): When processing, gathering, or analyzing data, for example, some situations increase the likelihood of data errors. So, although the null hypothesis states that there is no link between the type of error and the data error, the alternative hypothesis suggests that there is a considerable correlation between the two variables. To test these theories, statisticians may find out how strong and which way the correlation is between data error and the kind of mistake. Not only will this analysis prove or disprove the alternative hypothesis, but it will also reveal if the null hypothesis is correct.

Hypothesis (H_{01}): Error type inquiry is unaffected by the development and study of equalization methods for reducing data error rates.

In this article, the researchers look at the hypothesis (H_1) that states that inaccuracies are more common when certain types of errors are made during data collection, processing, or analysis.

Null Hypothesis (H_0): Data inaccuracies are common regardless of the kind of errors made during data collection, processing, or analysis.

Alternative Hypothesis (H_1): When processing, gathering, or analyzing data, for example, some situations increase the likelihood of data errors. So, although the null hypothesis states that there is no link between the type of error and the data error, the alternative hypothesis suggests that there is a considerable correlation between the two variables. To test these theories, statisticians may find out how strong and which way the correlation is between data error and the kind of mistake. Not only will this analysis prove or disprove the alternative hypothesis, but it will also reveal if the null hypothesis is correct.

Alternative hypothesis (H_1): The development and testing of an equalization's technique to reduce data error rates significantly influenced studies on error types.

ID	Metric	Abbreviation	Units	Range	Perfect Match Value
1	Mean Bias	MB	Units of x, p	[-∞, +∞]	0
2	Mean Absolute Gross Error	MAGE	Units of x, p	[0, +∞]	0

ID	Metric	Abbreviation	Units	Range	Perfect Match Value
3	Root Mean Squared Error	RMSE	Units of x, p	[0, +∞]	0
4	Cantered Root Mean Square Difference	CRM/SD	Units of x, p	[0, +∞]	0
5	Mean Normalized Bias	MNB	Unitless	[-1, +∞]	0
6	Mean Normalized Gross Error	MNGE	Unitless	[0, +∞]	0
7	Normalized Mean Bias	NMB	Unitless	[-1, +∞]	0
8	Normalized Mean Error	NME	Unitless	[0, +∞]	0
9	Fractional Bias	FB	Unitless	[-2, 2]	0
10	Fractional Gross Error	FGE	Unitless	[0, 2]	0
11	Theil's UI	UI	Unitless	[0, 1]	0
12	Index of Agreement	IOA	Unitless	[0, 1]	1
13	Pearson Correlation Coefficient	R	Unitless	[-1, 1]	1
14	Variance Accounted For	VAF	Unitless	[-∞, 1]	1

Mean Bias (MB): On average, the Mean Bias will reveal how far off the actual values are from the expected ones. It ranges from negative infinity to positive infinity and shares units with the data being evaluated. The intended and anticipated values are coincident when the Mean Bias is zero.

Mean Absolute Gross Error (MAGE): When trying to pin down the exact disparity between predicted and observed values, statisticians turn to Mean Absolute Gross Errors (MAGE). Its range is from zero to positive infinity, and its units are the same as the data being evaluated. When the Mean Absolute Gross Error equals zero, the target and predicted values are identical.

Error of Root Mean Squared (RMSE): Squaring the average squared difference between the target and expected values yields the Root Mean Squared Error, a statistic. Its range is from zero to positive infinity, and its units are the same as the data being evaluated. If the Root Mean square Error is 0, then the intended and anticipated outcomes are the same.

Centred "Root Mean Square" Difference (CRMSD): When calculating CMSE, the centre of the target values is used, same as when calculating RMSE. Its range is from zero to positive infinity, and its units are the same as the data being evaluated. Assuming both the target and predicted values are zero, they say that the centred root mean squared difference is zero. They call this a perfect fit.

This is a statistical measure called Mean Normalised Bias (MNB). It looks at the average ratio of the target values' deviation from the projected values to the mean value of the target values. Being unitless, its range extends from -1 to +infinity. The absence of bias in the prediction is shown by a Mean normalised Bias value of zero.

Mean Normalised Gross Error (MNGE): Mean Normalised Net Error looks at both the absolute difference between the target values and the projected values. It shows how far the target values are usually from their mean. It is unitless and has a range from zero to positive infinity. There is no difference between the goal and forecast values when the mean normalised gross error is 0.

Normalised Mean Bias (NMB): Mean Normalised Bias and normalised Mean Bias are comparable; the only difference is that the latter is given as a percentage. Taking the mean value of the goal values and multiplying it by 100 divides the difference between the target and forecast values, allowing researchers to get the average ratio. From positive infinity to negative 100%, its unitless range extends. With a normalised Mean Bias of 0, researchers get a forecast that is devoid of bias.

Normalised Mean Error (NME): The normalised Mean Error is a statistic that takes into account both the objective and anticipated values; to get the average ratio, it is multiplied by 100. From positive infinity to negative 100%, its unitless range extends. When comparing the expected and intended values, a normalised Mean Error result of 0% indicates a perfect match.

Fractional Bias (FB): The fractional bias measures the discordance between intended and expected results. It is calculated as the difference between the target values and the anticipated values, then normalised by the means of the target values. Because it lacks a base, its value ranges from -2 to 2. There will be no fractional bias if the anticipated and goal values are identical.

Fractional Gross Error (FGE): The absolute difference between the expected and desired values may be measured by the Fractional Gross Error after compensating for the average of the intended values. In its unitless range, the integers 0–2 are contained. The fractional gross error is 0 if the anticipated and intended values are identical.

Theil's UI (UI): The Il's user interface determines the ratio of the root mean squared error of the forecast to that of the target values. Its unitless range consists of the numbers zero through one. The goal and forecast values are equal when the perfect match value is 0 in Theil's user interface.

Index of agreement (IOA): To get the agreement index, divide the mean square error of the forecast by the mean square error of the divergence from the mean value of the target values. This will measure how well two sets of forecasts agree with each other. Its unitless range consists of the numbers zero through one. When the Index of Agreement value is 1, it means that the goal and prediction values are perfectly in sync.

Pearson correlation coefficient (R): The Pearson correlation coefficient may be used to quantify the linear relationship between the target and predicted values. It is unitless and may take on values between -1 and 1. A Pearson correlation value of 1 indicates a perfect linear connection between the expected and desired results.

Variance Accounted for (VAF): This forecasting metric assesses how well the volatility of the target values is accounted for. The boundary of this space is zero, and it continues all the way to infinity. If the forecast accurately accounts for the volatility of the target value, the volatility Accounted For value will be 1. The forecast might immediately surpass the target values with a number larger than 1, however. It is important to thoroughly examine the actual consequences and interpretation of values larger than 1 before proceeding.

In order to assess the efficacy of prediction models, the offered collection of error metrics provides a full and sophisticated set of tools. It is possible that these metrics may provide light on the model's overall predictive power, bias, and accuracy. A mutual understanding of the goals of the modelling project and the needs of the end users should inform the selection of suitable metrics. It is easier to assess, analyse, and build prediction models when these criteria are carefully integrated. The data type, the modelling objectives, and the intended use of the predictions all play a role in the metric selection process.

Different Types of Evaluation: There are a plethora of more metrics to choose from, such as those that take bias into consideration: squared error, correlation, normalised measurements, and absolute error, to name a few. Due to the multitude of variations, researchers have a lot of options for testing the model and seeing how it performs.

Comparing the Interpretability of Units: The results are much easier to understand since the data and several indications use the same units. Stakeholders are better able to understand the findings and how they relate to the model's accuracy in practice thanks to this feature. Prediction bias, if any, may be better understood by examining normalisation-related metrics such as Normalised Mean Bias (NMB), Mean Normalised Bias (MB), and Mean (MB). Normalisation measures are useful when comparing models in various settings, as they provide unitless indications. Root Mean Squared Error (RMSE), Mean Absolute Gross Error (MAGE), and Centred Root Mean Square Difference (CRMSD) are some error decomposition metrics that show how mistakes are spread out and where they happen.

The tendency of the model to overestimate or underestimate values may be gleaned by examining the mistakes. Index of Agreement (IOA), Variance Accounted For (VAF), and Pearson Correlation Coefficient (R) are a few ways to evaluate the degree to which the predicted and actual values coincide. A strong correlation suggests a dependable linear connection, even if IOA and VAF provide data on the explanation of variance and general agreement. Reasons to Consider It Fractional Gross Error (FGE) and Fractional Bias (FB) are two metrics that consider the real-world repercussions of mistakes and help to explain why actual results differ from ideal ones.

Decision-Making Use: The interpretation of these measures should vary depending on the situation. For example, bear in mind the objectives of the model prediction assignment while considering the potential consequences of a high level of variability or a positive bias. Rather than being a one-time event, this thorough evaluation should act as a foundation for continuous monitoring and enhancement. To keep the model relevant and successful throughout time, it is necessary to check it periodically and adjust as required.

Unveiling the Reality: It is essential to be honest about all aspects, including strengths and faults, while presenting findings. It is critical to effectively communicate the results of various metrics so that stakeholders may make informed choices based on a comprehensive understanding of the model's behaviour.

Linear Regression Modelling and the R^2 Coefficient of Determination:

Table 3: "Actual (Intended) Values" alongside the "Model-Predicted Values"

Date ID	Real value, ri	Predicated value, pi
1	287	311
2	40	55
3	68	60
4	256	302
5	115	87
6	190	152
7	300	297
8	222	235
9	145	165
10	172	136

Source: Giampieri [14]

According to the provided "Real Value" refer to Table 3 to compare the actual values with the anticipated values and the values predicted by the model. Researchers have access to several error measures while assessing the model's performance. This is a concise assessment using many metrics for inaccuracy.

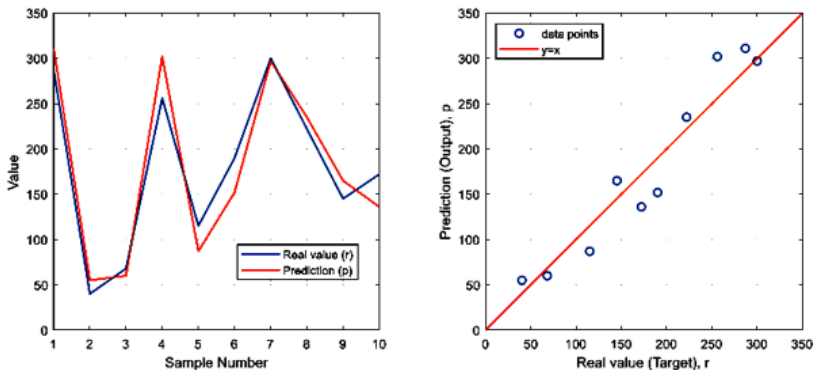
Mean Bias (MB): This metric is used to calculate the average deviation between the target and predicted values. The average bias, short for "MB," is obtained using the formula (1/n) ∑(pi-ri). Table 3 data showed that MB = 10.3, suggesting a little positive bias in the predictions, according to the researchers.

Root Mean Squared Error (RMSE): To find the average deviation from the objective and anticipated values, this statistic accounts for bias and variability. An equation for the root-mean-squared error (RMSE) may be written as RMSE = sqrt((1/n) ∑(pi-ri) ^2). Table 3 shows that the researchers found an RMSE of 42.2, which indicates that the predictions are not entirely predictable.

Pearson correlation coefficient (R): This metric assesses the degree to which the forecast and target values are linearly related. When the expected value is p̄and the desired value is r̄, the Pearson correlation coefficient, which is represented as R = ∑(pi-p̄) (ri-r̄) / sqrt(∑(pi-p̄) ^2 ∑(ri-r̄) ^2), may be determined. There is a positive and linear relationship between the anticipated and target values, as shown in Table 3, with an R-squared value of 0.8.

One measure that accounts for bias and variability is the Index of Agreement (IOA), which is used to determine the degree to which the anticipated and intended results match. Where pi and ri are the agreement coefficients, the Index of Agreement (IOA) may be defined as 1/2(∑|pi-r̄|+∑|ri-r̄|)^2. According to Table 3, the researchers have determined an IOA of 0.8, which means that the predicted and intended values are very compatible. In spite of some predicted unpredictability and a little degree of positive bias, the model seems to function adequately generally. Positive linearity exists, however, and the two sets of values—expected and desired—are quite close to one another. In order to properly assess and derive inferences from these error measures, one must meticulously investigate the particular circumstances and objectives of the prediction assignment. The model's overall performance is good, despite its high forecast variability and little positive bias. Even with these flaws, the reasonably high IOA and significant positive linear connection show that the model's predictions and actual data agree with each other quite a bit. These error measures should be interpreted with care, taking into account the prediction job's unique context and goals. Before drawing any conclusions from these signs, make sure researchers thoroughly examine the model's actions in relation to the objectives of the project (refer to Figure 2). The predicted value can be calculated using a linear regression model [13].

Figure 2: Real (target) values and model-predicted values for the numerical example



Source: Collected by Author

Discussion

Because of new developments, it is important to pick the right error measures and model evaluation methods, especially when dealing with adversarial robustness, zero-inflated data, heavy-tailed error distributions, and unbalanced regression problems. Predictive models are more trustworthy and easier to understand when these factors are included in the evaluation procedures. A lot of people have been paying attention to how to measure the accuracy of neural networks' and classification models' predictions recently, especially with regard to how to deal with zero-inflated data and how robust performance measures are. Assessing adversarial resilience is still an important part of evaluating models. This measure aids in distinguishing between adversarial and non-adversarial cases [14]. The 'residual error' measure, which evaluates a model's performance at the individual sample level, was introduced in [15]. A model's robustness to hostile perturbations may be better understood with the help of this measure.

Dealing with Zero-Inflation Datasets: Training and evaluating models may be difficult when dealing with datasets that include a large number of zero values. In order to tackle this problem, Giampieri et al. [15] presented a two-pronged

machine learning strategy, which showed better results in situations such as home appliance categorisation and airport shuttle capacity prediction. Their hierarchical approach improved accuracy, recall, and energy economy by controlling the distortion brought on by too many zeros.

Performance Metrics' Robustness: The dependability of conventional calibration statistics has been called into doubt when heavy-tailed error distributions are present. Measures like Mean Squared Error (MSE) along with Mean Variance (MV) become untrustworthy in certain situations, as pointed out [16, 17]. According to the research, a more reliable alternative to using z-scores to evaluate model performance in these types of situations is the ZMS statistic. When dealing with unbalanced data, conventional loss functions may not adequately highlight the significance of outliers. For these kinds of cases, Silva, as a researcher in 2022 proposed the Squared Error Relevance Area (SERA) loss function. Models optimised with SERA outperform those using traditional loss functions when it comes to forecasting extreme values, according to their research. Because it is critical for the interpretability of models to comprehend how well model predictions match up with real brain responses, the spectral study of neural predictions is an important tool. The authors developed a spectral framework to analyse prediction errors, considering the alignment of model Eigen spectra with brain responses [6, 12]. Differentiating across models with comparable performance measures is made easier using this method, which sheds light on the geometrical features of prediction mistakes.

Conclusion

The nature of the interactions between different forms of mistakes and data errors should be the focus of data quality management researchers. Investigating if data mistakes are more common in certain contexts (such as during data collection, processing, or analysis) is a reasonable and beneficial course of action. Sample bias, measurement inaccuracy, confounding factors, inadequate data, lack of analysis, and lack of control are some of the limitations that students should be mindful of while studying this work. Notwithstanding these caveats, studying the correlation between error and data inaccuracy is valuable since higher-quality data leads to more trustworthy findings and judgements. In the end, it's crucial that future studies try to fix the shortcomings while also looking at other kinds of errors and data. The quality and trustworthiness of data-driven decisions may be improved if scholars delve more into the link between erroneous data and other forms of error.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgement

The authors are thankful to the institutional authority for completion of the work.

References

1. Rožanec JM, Petelin G, Costa J, Bertalanič B, Cerar G, Guček M, Papa G, Mladenić D. Dealing with zero-inflated data: achieving SOTA with a two-fold machine learning approach. arXiv preprint arXiv:2310.08088. 2023 Oct 12. https://doi.org/10.48550/arXiv.2310.08088

2. Pernot P. Negative impact of heavy-tailed uncertainty and error distributions on the reliability of calibration statistics for machine learning regression tasks. arXiv preprint arXiv:2402.10043. 2024 Feb 15. https://doi.org/10.48550/arXiv.2402.10043

3. Canatar A, Feather J, Wakhloo A, Chung S. A spectral theory of neural prediction and alignment. Advances in Neural Information Processing Systems. 2023 Dec 15;36:47052-80. https://doi.org/10.48550/arXiv.2309.12821

4. Chan DW, Cristofaro M, Nassereddine H, Yiu NS, Sarvari H. Perceptions of safety climate in construction projects between workers and managers/supervisors in the developing country of Iran. Sustainability. 2021 Sep 17;13(18):10398. https://doi.org/10.3390/su131810398

5. Pernot P. Negative impact of heavy-tailed uncertainty and error distributions on the reliability of calibration statistics for machine learning regression tasks. arXiv preprint arXiv:2402.10043. 2024 Feb 15. https://doi.org/10.48550/arXiv.2402.10043

6. Canatar A, Feather J, Wakhloo A, Chung S. A spectral theory of neural prediction and alignment. Advances in Neural Information Processing Systems. 2023 Dec 15;36:47052-80. https://doi.org/10.48550/arXiv.2309.12821

7. Aboutalebi H, Shafiee MJ, Karg M, Scharfenberger C, Wong A. Residual error: a new performance measure for adversarial robustness. arXiv preprint arXiv:2106.10212. 2021 Jun 18. https://doi.org/10.48550/arXiv.2106.10212

8. Silva A, Ribeiro RP, Moniz N. Model optimization in imbalanced regression. InInternational Conference on Discovery Science 2022 Oct 10 (pp. 3-21). Cham: Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2206.09991

9. Marín, L.S.; Lipscomb, H.; Cifuentes, M.; Punnett, L. Perceptions of safety climate across construction personnel: Associations with injury rates. Saf. Sci. 2019, 118, 487–496. https://doi.org/10.1016/j.ssci.2019.05.056

10. Gonzalez M, Rodriguez A, Pereira O, Celaya A, de Lacalle LL, Esparta M. Axial-compliant tools for adaptive chamfering of sharp-edges: Characterisation and modelling. Engineering Science and Technology, an International Journal. 2023 May 1;41:101407. https://doi.org/10.1016/j.jestch.2023.101407

11. Gaheen OA, Benini E, Khalifa MA, Aziz MA. Pneumatic cylinder speed and force control using controlled pulsating flow. Engineering Science and Technology, an International Journal. 2022 Nov 1;35:101213. https://doi.org/10.1016/j.jestch.2022.101213

12. Norsahperi NM, Danapalasingam KA. An improved optimal integral sliding mode control for uncertain robotic manipulators with reduced tracking error, chattering, and energy consumption. Mechanical Systems and Signal Processing. 2020 Aug 1;142:106747. https://doi.org/10.1016/j.ymssp.2020.106747

13. Using a Linear Regression Model to Calculate a Predicted Response Value - Explanation. Study.com. https://study.com/skill/learn/using-a-linear-regression-model-to-calculate-a-predicted-response-value-explanation.html [Accessed 21 Mar. 2024].

14. Hanumanthappa H, Vardhan H, Mandela GR, Kaza M, Sah R, Shanmugam BK. A comparative study on a newly designed ball mill and the conventional ball mill performance with respect to the particle size distribution and recirculating load at the discharge end. Minerals Engineering. 2020 Jan 1;145:106091. https://doi.org/10.1016/j.mineng.2019.106091

15. Giampieri A, Ling-Chin J, Ma Z, Smallbone A, Roskilly AP. A review of the current automotive manufacturing practice from an energy perspective. Applied Energy. 2020 Mar 1;261:114074. https://doi.org/10.1016/j.apenergy.2019.114074

16. Truong LV, Huang SD, Yen VT, Cuong PV. Adaptive trajectory neural network tracking control for industrial robot manipulators with deadzone robust compensator. International Journal of Control, Automation and Systems. 2020 Sep;18(9):2423-34. https://doi.org/10.1007/s12555-019-0513-7

17. Nusbaum U, Weiss Cohen M, Halevi Y. Path planning and control of redundant manipulators using bilevel optimization. Journal of Dynamic Systems, Measurement, and Control. 2020 Apr 1;142(4):041008. https://doi.org/10.1115/1.4045976