Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 8
Open Access

Investigation And Development Of Compensation Strategies For Reducing Data Error Expenses Through The Examination Of Different Error Categories

Miao Congjin 1st , Muhammad Ezanuddin Abdul Aziz 2nd

Cite this paper as: Miao Congjin, Muhammad Ezanuddin Abdul Aziz (2024) " Investigation And Development Of Compensation Strategies For Reducing Data Error Expenses Through The Examination Of Different Error Categories". *Frontiers in Health Informatics,* (8), 5243-5257

ABSTRACT

Because data mistakes may have serious negative effects on companies, finding a way to equalize data error rates is a crucial topic of study. Data processing errors come in many forms, and this article delves into those forms to help readers design an equalizations strategy to deal with them. In their first section, the writers distinguish between random mistakes and systematic errors, the two most common forms of processing errors in data. A larger sample size or more precise measuring methods might lessen the impact of random errors, which are mistakes that happen by chance. In contrast, systematic errors are those that happen repeatedly and might have several causes, including but not limited to equipment failure, calibration mistakes, or bias in the data gathering procedure. To deal with systematic inaccuracies, the authors suggest an equalizations strategy that entails finding and fixing the data sources that are inaccurate. Analyzing the data for trends or patterns that might point to systematic problems and then using the right corrective methods to lessen their impact is what this method is all about. The authors show that their equalizations method works by conducting a battery of tests using both theoretical and practical data. The equalizations method greatly decreased data error rates in these trials, allowing for more precise and trustworthy conclusions. All things considered, the essay does a great job of explaining the various data processing issues and how to fix them using an equalizations method. Organizations may boost their performance and achieve more success by enhancing the data quality and decision-making capabilities via the reduction of error rates.

Keywords: *Error rates, corrective strategies, error locations, systemic errors, and equalization.*

1. INTRODUCTION

Analyzing and preparing data are complementary processes. transcribing the data-derived information. Approaches that are diverse, such as decision-making models that help find patterns, connections, and relevant conclusions (Van Truong et al., 2020). The data must be prepared beforehand, however, to do the analysis. Data preparation involves converting information into a format that computers can understand and work with. designed for use with statistical software such as SAS and SPSS. Steps in data preparation include encoding data, entering data, filling in blanks, and reformatting data. A researcher will briefly review each of these steps in this section: Data Coding: When working with data, the first step is to convert it from its raw form into a numerical representation. A codebook, a collection of several kinds of data, is used. elements, including the answer, variables, measurements,

and variable format; concluding the coding with a codicil. The process's response determines the sorts of scales. consider the following: the kind of scale used (nominal, ratio, ordinal, interval), the number of points (five, seven, etc.), and so on. For instance, healthcare is classified as 1 in numerical notation for the purpose of business classification. When it's 2, it means production, when it's 3, it means retail, and when it's 4, it means finance. Data Entry: Now comes the fun part: entering all that coded data into some kind of text file or spreadsheet. It is simple to include it in the software bundle. Because some people choose not to participate in the survey, there are blanks in the statistics. could not answer all questions for different reasons, thus it's important to find a way to reconsider these unmet standards. The addition of -1 or 999 is required by certain programs, for example. Listwise deletion is used by some of them, but many of them handle missing data automatically. approach of handling missing values, whereby each response set is eliminated in the event of a single omission. Before trying to make sense of the data, it could be required to execute certain transformations on them. It may be necessary to make a modification before objects with reverse coding, for example, may be used. in comparison to or in conjunction with non-inverted parts. This concept is used when the meaning of an object has been changed. disagreement with their central argument (, 2017). Types of information analysis provides a synopsis of the most popular methods for analyzing data that follow. For this reason, six main types of data analysis exist. Descriptive, exploratory, inferential, predictive, explanatory/cause-and-effect, and mechanistic analyses are the six main categories (Nusbaum et al., 2020)

A. Describes: It has been around the longest and is the most basic kind of data analysis. Consequently, it works well with very large datasets. This is where a data set analysis makes use of the information.

B. Exploratory: Approach to uncovering hidden knowledge, forming unexpected relationships, and establishing a foundation for future research or the generation of novel research questions.

C. Inferential: Extrapolating results from a smaller sample to the whole population is the main objective of inferential research. The evidence utilized to test a hypothesis about the nature of the cosmos comes from a very tiny portion of Earth. This method is effective when used to observational data, historical data, and cross-sectional time series.

D. Predictive: This kind of predictive research considers both the present and the past. On top of that, it may use the first subject's data to predict the values of a second subject. No matter how many there are, a simpler model that makes better use of data may be the one to go. Researchers need to think about the prediction data collection and the things that will be utilized to measure results because of this.

E. Rationale: Approach Based on Processes Part of this method that requires the greatest work is analyzing randomized trial data sets to identify which changes in the variables may impact other ones. One would also assume that mechanistic analysis is quite improbable. Because of this, it is a suitable option for fields such as engineering and the physical sciences, where a little mistake may have a big

monetary effect. The next section will provide an in-depth look at the three main areas of statistical analysis: descriptive, explanatory, and inferential statistics.

2. BACKGROUND OF THE STUDY

The two most important parts of any research effort are gathering relevant data and analyzing it. The study conducted by (Giampieri et al., 2020) Data preparation involves transforming raw, unstructured information into a form that computers can understand and use. Coding, entering, filling in blanks, and reformatting data are all part of it. Data analysis, on the other hand, involves using various techniques to discover patterns, correlations, and valuable insights within the data. One may approach data using one of six primary methods: descriptive, exploratory, inferential, predictive, explanatory/casual, or mechanistic assessment. Descriptive analysis is the most basic kind of data analysis; it consists of summarizing data to provide an overview. With exploratory analysis, one hopes to find new associations and set the stage for further in-depth research. The goal of inferential statistics is to extrapolate results from a smaller sample to the whole population. Predictive analysis looks at the past and present to make predictions about the future, while explanatory or causal analysis tries to set things in motion. I utilize mechanistic analysis to find the exact changes in one variable that caused changes in another. An important aspect of thorough analysis is data summary for straightforward presentation. Bivariate and multivariate analysis are the two primary subsets of this technique. Common univariate statistical processes include frequency analysis, central tendency analysis, and dispersion analysis, all of which focus on a single variable. Analysis of central tendency and analysis of frequency are two ways to look at data. While the latter finds measures of central tendency such as the mean, median, and mode, the former counts all potential values for a variable. As stated by (Norsahperi & Danapalasingam, 2020), the variability of the data may be found using distribution analysis, which involves calculating the standard deviation, variance, and range (Xu et al., 2021).

3. LITERATURE REVIEW:

Data that is difficult to comprehend may be summarized using this strategy. One may say that this approach is multivariate or bivariate (Hanumanthappa et al., 2020). "Univariate" is a common descriptor for statistical methods that focus on a single variable. Particularly, researchers will be making use of Frequency Analysis, Central Tendency Analysis, and Dispersion Analysis. Simply interfering with the frequency of the relevant variable will reveal its source. By tallying the occurrences of each value for a certain variable, it compiles all the potential possibilities. The quantity of the most represented value, commonly known as the three Ms, may be used to assess the central tendency of disturbance when comparing one variable to a series of data. Common methods for quantifying central tendency include the mean, mode, and standard deviation. The Mode is the most frequent value in a group of numbers, whereas the Mean is the average of all the values. Variables' dispersion around the mean is defined by dispersion. Statistics often make use of standard deviation, which is calculated as the square root of variance, range, and variance. The dissimilarity between the two extreme values is shown by the range. Analyzing the variance reveals how closely the data points cluster around the mean. Two datasets with two independent variables may be compared using this method. This allows

scientists to discern a connection between the two factors. Bivariate correlation is the most used statistic. The method used to determine the amount of correlation using this statistic considers the means and standard deviations of the samples. When there are more than two variables, it is still useful. While solving such programs manually is complicated, software like SPSS makes it a breeze (Gaheen et al., 2022). Assessment of Explanation Performing an explanatory analysis aims to identify potential factors that might have contributed, as mentioned before. Concerns about patterns, correlations, and relationships between variables are addressed via the application of explanatory analyses (Sarvari et al., 2021). Dependency and interdependence processes are the basic strategies of explanation analysis. Several independent variables might influence a single dependent variable; this is what the concept of dependence is all about. "Interdependence approaches" are a kind of multivariate analysis that aims to find correlations between variables without assuming the magnitude or direction of any effect (Rodríguez et al., 2021).

4. RESEARCH METHODOLOGY:

If regression models like Neural Networks or similar prediction models do not provide flawless forecasts, overfitting may not occur. No matter how careful a researcher is, their predictions will always be somewhat wrong.in order to evaluate the outcomes of several models, identify the most effective ones, and then make a well-informed decision. For this reason, there are many metrics for prediction error that may be used. The estimated values are represented by p, an N1 vector, and the calculated (or measured) and forecasted values of a quantity are indicated by r and p, respectively. To provide an example, researchers may train an ANN to make predictions N times for every given input. For providing an unbiased, external evaluation, researchers may compare the initial data set (S) with either a different set of data (N), a subset of the original data set (T), or even with no data (U). Below, experts in the field will go over a variety of metrics that may be used to find the prediction error of this model. down this study, the researchers zero down on the problem with continuous variables. Classification metrics include things like recall, accuracy, confusion matrices, and false positive rate. Keep in mind that for the following formulae to be used, the observations and their predictions must be positive. Certain calculations may need researchers to make modifications if their data contains elements such as negative numbers or zeroes.

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

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

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

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

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

Even if MB=0 is required for a superb combination of the actual and projected values (such those identical numbers), it may be achieved in imperfect matches by canceling out negative and positive mistakes. This "Mean Absolute Gross Error (MAGE)" measures the average mistake across several predictions without taking their direction into account. so calculates the overall discordance between the expected and observed values in the test sample and uses a weighted average to do so. That variable's value, which may be either positive or negative, is

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

One common statistic used in regression analysis is the Mean Squared Error (MSE). It stands for the average squared deviation from the true value. According to the definition, it might have good or bad connotations.

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

One major issue with MSE is its inability to handle really serious circumstances. If the error for one sample is much larger than the error for other samples, the square of the error will grow noticeably. Extreme results may particularly affect MSE since it averages errors. The Root Mean Squared Error (RMSE) is another popular statistic for assessing how well an estimator or model can predict values (demographic or sample values) that differ from the observed values. Researchers may find this out by squaring the mean square error. Relative standard error (RMSE) is preferable to mean squared error (MSE) when the units of the target variable are not coincidental. The equation for this real number, which ranges from zero to one plus (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 Center-Mean-Square Distinction (CMSD).

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

Once the focus attribute and the CMSD are expressed in the same unit, the square root of the two is called a CRMSD, which stands for Concentrated 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 one technique to demonstrate how off a model's forecast is when utilizing the CRMSD value; we'll go into more detail later. Standardized bias error values are sometimes expressed as a percentage, known as the Mean Index Bias (MNB, unitless).

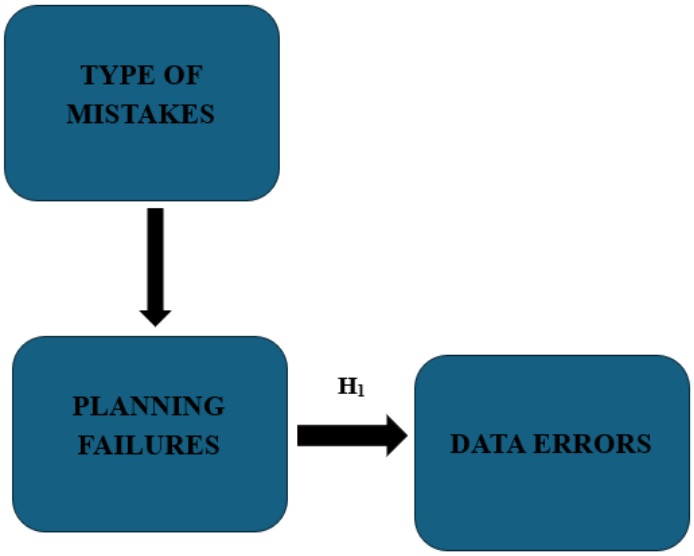
$$MNB = \frac{1}{N}\sum_{i=1}^{N}\frac{p_i - r_i}{r_i} = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{p_i}{r_i}\right) - 1$$

Unitless Mean Normalized Gross Error (MNGE) is also known as the "Mean Absolute Percentage Error." If the amount of error is known, readers will have a better idea of how accurate the estimations are. It is derived from

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

One of the major problems with MSE is that it can't deal with numbers that aren't in the range. If a sample's associated error is much greater than the other samples, its square will be substantially larger as well. Because it averages researchers' errors, MSE is also vulnerable to extreme situations.

5. CONCEPTUAL FRAMEWORK



❖ **INDEPENDENT VARIABLE:**

- **Type of Mistake:** During data collection, processing, or analysis, a wide variety of mistakes might occur. The three possible forms of error are outlier, random, and systematic.

❖ **DEPENDENT VARIABLE:**

- **Data Error:** The reliability and accuracy of any conclusions or decisions made from the data can be compromised if there are errors or omissions in the data.
- **Framework:**
- Understand that there are several ways in which mistakes might occur while collecting, processing, or analyzing data. Classify and quantify the many types of data mistakes. Finding the root-mean-squared error, relative error, or mean absolute error can give a good idea of how inaccurate the data is.

- Use statistical tools like regression or correlation to investigate the relationship between different types of errors and data errors.
- Identify any additional variables, known as confounding variables, that could be impacting the relationship between the dependent and independent variables.
- Consider the kind and frequency of different types of mistakes while creating a model that can predict the quantity of data errors.
- If want to know how well researchers model is doing, use appropriate metrics like R-squared or mean squared error. Identify the most prevalent types of data problems and devise strategies to remedy them based on the analysis's results. To sum up, this method comprises examining the association between data mistakes and cataloging the many types of data processing, analysis, and collection errors. This study has the potential to improve the reliability and accuracy of data-based conclusions and assessments by exploring ways to decrease data inaccuracy.

❖ RESULT

- **Factor Analysis:**

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

1. **Hypothesis (H_1):** The kind of errors made while gathering, processing, or analyzing data greatly affects how often data is inaccurate.
2. **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 (H1) that states that inaccuracies are more common when certain types of errors are made during data collection, processing, or analysis.

Null hypothesis (H0): Data inaccuracies are common regardless of the kind of errors made during data collection, processing, or analysis.

Alternative hypothesis (H1): In 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 sort of error and 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.

Null hypothesis(H01): Error type inquiry is unaffected by the development and study of equalization methods for reducing data error rates.

Alternative hypothesis (H_1): The development and testing of an equalizations 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	
3	Root Mean Squared Error	RMSE	Units of x, p	[0, +∞]	0
4	Cantered Root Mean Square Difference	CRMSD	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): As the Mean Bias aims to measure is the average discrepancy between the projected and actual values. It shares units with the data being considered and has a range of negative infinity to positive infinity. With a Mean Bias of 0, there is a perfect match between the goal and prediction values, which are identical.

Mean Absolute Gross Error (MAGE): Statisticians use Mean Absolute Gross Errors (MAGE) to determine the precise difference between the expected and actual values. Its units are the same as the data being considered, and its range goes from zero to positive infinity. The goal and forecast values are equal when the Mean Absolute Gross Error is 0.

Error of Root Mean Squared (RMSE): The Root Mean Squared Error is a metric that takes the square root of the average squared difference between the goal and predicted values. Its units are the same as the data being considered, and its range goes from zero to positive infinity. When the Root Mean Squared Error is zero, it means that the expected and intended results are identical.

Centered "Root Mean Square" Difference (CRMSD): As with RMSE, the center of the target values is the focus of Centered Root Mean Squared Error (CMSE). Its units are the same as the data being considered, and its range goes from zero to positive infinity. When the Centered Root Mean Square Difference equals 0, it indicates that the goal and forecast values are identical. This is known as a perfect match.

Mean Normalized Bias (MNB): Analyzing the average ratio of the target values' departure from the projected values to the mean value of the target values is the Mean Normalized Bias, a statistical measure. Being unitless, its range extends from -1 to +infinity. The absence of bias in the prediction is shown by a Mean Normalized Bias value of zero.

Mean Normalized Gross Error (MNGE): Mean Normalized Net Error considers both the absolute difference between the target values and the projected values; it is a measure of the typical deviation from the mean of the target values. 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 Normalized Gross Error is 0.

Normalized Mean Bias (NMB): Normalized Mean Bias is similar to Mean Normalized Bias, however it is expressed as a percentage. After dividing the difference between the target and projected values by 100, the average ratio is determined by multiplying the mean value of the goal values by the same amount. It has a unitless range that goes from positive infinity to negative 100%. A bias-free prediction would have a Normalized Mean Bias value of zero.

Normalized Mean Error (NME): One statistical measure that considers both the target and expected values is the Normalized Mean Error, which is then multiplied by 100 to get the average ratio. It has a unitless range that goes from positive infinity to negative 100%. An outcome of 0% for Normalized Mean Error indicates an ideal match when contrasting the intended and anticipated values.

Fractional Bias (FB): The discordance between expected and intended outcomes is evaluated by the fractional bias, which is the difference between anticipated and target values normalized by the means of the target values. As a unitless number, it falls between -2 and 2. If the target and predicted values are the same, then the fractional bias is zero.

Fractional Gross Error (FGE): After adjusting for the average of the intended values, the Fractional Gross Error quantifies the absolute gap between the anticipated and desired values. Its unitless range consists of the numbers 0 through 2. If the expected and desired values are the same, then the fractional gross error is zero.

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 give a measure of 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): The volatility Accounted For measures how well the forecast takes the target values' volatility into consideration. This unitless space extends from negative infinity to one. The volatility Accounted For value will be 1 if the prediction correctly takes the target value's volatility into consideration. With a number greater than 1, however, the prediction could outperform the target values directly. Values greater than 1 need careful analysis of their practical implications and interpretation. The provided set of error measures offers a comprehensive and advanced toolkit for evaluating prediction models' performance. These indicators may help us understand the model's accuracy, bias, and predictive capacity in general. Selecting appropriate metrics should be based on a shared understanding of the modeling job's objectives and the end users' requirements. When these criteria are thoughtfully combined, they make it simpler to evaluate, analyze, and construct predictive models. The choice of metrics is influenced by factors like the kind of data, the aims of the modeling, and the planned application of the predictions. Assessment in Its Many Forms: Many other metrics are available, including those that account for bias, squared error, correlation, normalized measures, and absolute error, among many others. Researchers can test the model in many ways and observe its performance due to the abundance of variants. Unit Interpretability Comparison: Since the data and many of the indicators share the same units, the findings are much simpler to interpret. This feature helps stakeholders comprehend the results and their practical implications of the model's accuracy. A look at measures like Normalized Mean Bias (NMB), Mean Normalized Bias (MB), and Mean (MB) about normalization might provide light on the presence and degree of bias in forecasts. Because they provide unitless indicators, normalization measures are useful for comparing models in different contexts. Several error decomposition measures, such as Root Mean Squared Error (RMSE), Mean Absolute Gross Error (MAGE), and Centered Root Mean Square Difference (CRMSD), show the distribution and dispersion of errors. By analyzing the errors, one might learn whether the model tends to overestimate or underestimate values. How well the expected and desired values matchup is assessed by measures like the Pearson Correlation Coefficient (R), Index of Agreement (IOA), and Variance Accounted For (VAF). Although IOA and VAF provide information on the explanation of variation

and general agreement, a substantial correlation implies a reliable linear relationship. Motives to Give It Some Thought Metrics like Fractional Gross Error (FGE) and Fractional Bias (FB) consider the practical consequences of errors and provide insight on the reasons why real-world outcomes deviate from ideal ones. Use in Making Decisions: These metrics should have context-dependent meanings. When weighing the possible effects of a positive bias or high variability, for instance, it is essential to keep in mind the specific goals of the model prediction assignment. Rather from being a one-and-done transaction, this thorough assessment should be used as a starting point for continuous monitoring and development. It is critical to monitor the model often and adjust as needed to ensure its continued relevance and effectiveness throughout time. Revealing the Truth: When presenting results, it is critical to be truthful about everything, including strengths and weaknesses. Effective communication of the outcomes of different metrics is crucial for stakeholders to make educated decisions based on a complete understanding of the model's behavior.

LINEAR REGRESSION MODELING AND THE R^2 COEFFICIENT OF DETERMINATION:
Table 2 displays the "actual" (intended) values alongside the "model-predicted values" for the numerical example.

Data ID	*Real value, ri*	*Predicted 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*

Per the supplied "Real" To see how the actual values compare to the expected ones, as well as the values predicted by the model, see Table 2. When evaluating the model's performance, researchers have access to a wide variety of error metrics. This is a brief evaluation that makes use of many measures for error:

Mean Bias (MB): The mean deviation between the goal and forecasted values is computed using this measure. To get the average bias, abbreviated as MB, the formula (1/n) ∑(pi-ri) is used. Researchers discovered that MB = 10.3 based on data in Table 2, which implies a little positive bias in the predictions.

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 2 shows that the researchers found an RMSE of 42.2, which indicates that the predictions are not entirely predictable.

Pearson correlation coefficient (R): This statistic measures how linear the relationship is between the prediction and the goal value. The Pearson correlation coefficient, Denoted as R = ∑(pi-p̄) (ri-r̄) / sqrt(∑(pi-p̄) ^2 ∑(ri-r̄) ^2), may be calculated where p̄ and r̄ are the means of the anticipated and target values, respectively. Table 2 shows that the predicted and target values are positively and linearly associated with one another, with an R-squared value of 0.8.

Index of agreement (IOA): To find out how well the expected and desired values correspond; this statistic takes bias and variability into account. The Index of Agreement (IOA) may be expressed as 1/2(∑|pi-r̄|+∑|ri-r̄|) ^2, where pi and ri are the agreement coefficients. Based on the information in Table 2, the researchers find an IOA of 0.8, indicating that the expected and desired values are very congruent. Despite a little amount of positive bias and a moderate amount of expected variability, the model seems to perform well overall. But there's a strong positive linear relationship, and the expected and desired values are quite close to one another. It is essential to carefully examine the specific context and goals of the prediction assignment when evaluating and drawing conclusions from these error measurements. Although the model exhibits significant prediction variability and a little positive bias, its overall performance is satisfactory. The relatively high IOA and large positive linear link demonstrate that, despite these qualities, there is a decent degree of agreement between the model's predictions and actual data. It is important to consider the specific context and aims of the prediction job when interpreting these error metrics. When making judgments based on these indications, it is important to carefully analyze the model's behavior in respect to the endeavor's goals.

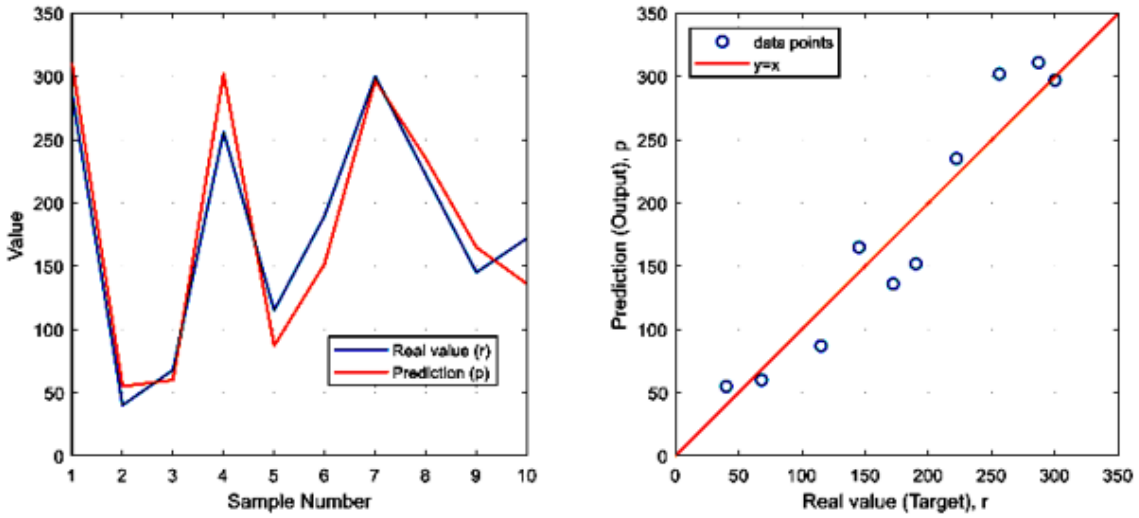


Figure 1: "Real" (target) values and model-predicted values for the numerical example.

6. CONCLUSION

Researchers in the field of data quality management would do well to zero in on the nature of the relationships between various mistake types and data errors (Marín et al., 2019). It is a valid and useful line of inquiry to examine if data errors are more prevalent in certain settings (such as when data is being collected, processed, or analyzed). While considering this work, students should be aware of the following limitations: sample bias, measurement error, confounding variables, insufficient data, absence of analysis, and lack of control. These limitations notwithstanding, better data quality leads to more reliable conclusions and judgments based on data, so it is worthwhile to investigate the relationship between mistake types and data inaccuracy. Organizations may enhance their data quality and data-driven insights by identifying and reducing the types of mistakes that create data difficulties. Ultimately, it is critical that future research aims to address the limitations while also expanding the types of mistakes and data types examined. By further investigating the connection between inaccurate data and other types of mistakes, academics will be able to enhance the reliability and quality of data-driven judgments (Chan et al., 2021).

REFERENCES:

- Chan, D.W.M.; Cristofaro, M.; Nassereddine, H.; Yiu, N.S.N.; Sarvari, H. Perceptions of safety climate in construction projects between workers and managers/supervisors in the developing country of Iran. Sustainability 2021, 13, 10398.
- 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.
- Rodríguez, M. González, O. Pereira, L.N. López de Lacalle, M. Esparta, Int. J.Adv. Manuf. Technol. (2021) K. Groß, M. Eifler, K. Klauer, K. Huttenlochner, B. Kirsch, C. Ziegler, et al., Eng. Sci. Technol. Int. J. 24 (2021) 543–555.
- O.A. Gaheen, E. Benini, M.A. Khalifa, M.A. Aziz, Eng. Sci. Technol. Int. J. 35(2022) 101213
- Norsahperi NMH, Danapalasingam KA. An improved optimal integral sliding mode control for uncertain robotic manipulators with reduced tracking error, chattering, and energy consumption. Mech Syst Signal Process 2020;142(August): 106747.
- 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. Miner Eng 2020;145(January):106091.
- Giampieri A, Ling-Chin J, Ma Z, Smallbone A, Roskilly AP. A review of the current automotive manufacturing practice from an energy perspective. Appl Energy 2020;261(March (01)):114074
- Van Truong L, Huang SD, Yen VT, Van Cuong P. Adaptive trajectory neural network tracking control for industrial robot manipulators with deadzone robust compensator. Int J Control Autom

- Syst 2020;18(September (9):2423–34.
- Nusbaum U, Weiss Cohen M, Halevi Y. Path planning and control of redundant manipulators using bilevel optimization. J Dyn Syst Meas Control 2020;142.
- Xu X, Chen W, Zhu D, Yan S, Ding H. Hybrid active/passive force control strategy for grinding marks suppression and profile accuracy enhancement in robotic belt grinding of turbine blade. Robot Comput Integr Manuf 2021;67(February): 102047.