A Framework for Analyzing the Impact of Missing Data in Predictive Models

Fabiola Santore
Federal University of Paraná
Curitiba, Brazil
fsantore@inf.ufpr.br

Eduardo C. de Almeida
Federal University of Paraná
Curitiba, Brazil
eduardo@inf.ufpr.br

Wagner H. Bonat
Federal University of Paraná
Curitiba, Brazil
wbonat@ufpr.br

Eduardo H. M. Pena
Federal University of Technology - Paraná
Toledo, Brazil
eduardopena@utfpr.edu.br

Luiz Eduardo S. de Oliveira
Federal University of Paraná
Curitiba, Brazil
lesoliveira@inf.ufpr.br

ABSTRACT

We propose a stochastic framework to evaluate the impact of missing data on the performance of predictive models. The framework allows full control of important aspects of the data set structure. These include the number and type of the input variables, the correlation between the input variables and their general predictive power, and sample size. The missing process is generated from a multivariate Bernoulli distribution, which allows us to simulate missing patterns corresponding to the MCAR, MAR and MNAR mechanisms. Although the framework may be applied to virtually all types of predictive models, in this article, we focus on the logistic regression model and choose the accuracy as the predictive measure. The simulation results show that the effects of missing data disappear for large sample sizes, as expected. On the other hand, as the number of input variables increases, the accuracy decreases mainly for binary inputs.

CCS CONCEPTS

• **Information systems → Incomplete data**.

KEYWORDS

Data Simulation; Missing Data; Predictive Model

ACM Reference Format:
Fabiola Santore, Eduardo C. de Almeida, Wagner H. Bonat, Eduardo H. M. Pena, and Luiz Eduardo S. de Oliveira. 2020. A Framework for Analyzing the Impact of Missing Data in Predictive Models. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3412129

1 INTRODUCTION

The quality of data is key in supporting data-centric systems, machine learning routines, and predictive models. Research on data quality aims to define, identify, and repair inconsistencies in the data [9]. A common source of inconsistency is missing data, in which no data is stored for the variable in an observation, which potentially hides important information.

Although the popularity of predictive analytics using machine learning tools has been increasing, the quality of the input and output variables have been neglected on the proposition of new predictive models. Consequently, the effect of missing data in many of the standard predictive models is completely unknown. Two strategies are often used to handle missing data: i) removing missing records, and ii) data imputation. Imputation techniques replace missing values with probable values [1, 4]. However, it is hard to verify the effectiveness of the imputation techniques and their impact on data analysis.

The studies that evaluate the performance of predictive models in the presence of missing data use real data sets that already have missing data or data sets with artificially inserted missing data. In any case, the analysis requires some imputation technique to repair the missing records, followed by a predictive model's fitting. The predictive performance is usually compared with the one without missing data [2, 3, 7, 10]. Another limitation of these approaches is that they use a small number of data sets, and do not use statistical methods to analyze the results. Extensive studies about the number of data sets are presented by [8] and [6]. However, by using real data sets, they could not control the aspects concerning missing data (e.g., distribution, types). Consequently, their conclusions do not take into account the uncertainty associated with these aspects.

In this paper, we propose a stochastic framework to evaluate the impact of missing data on predictive models' performance. The framework is based on a stochastic generation of data sets with control of essential aspects of the missing data and the input variables for the predictive models. Our framework has four steps: i) generating the whole data set; ii) inserting missing observations; iii) fitting a predictive model and measuring performance, and iv) evaluating the results using a statistical tool. The framework allows us to design a wide range of simulation scenarios to evaluate the impact of important factors, such as sample size, type and predictive power of the input variables, and the correlation between variables.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412129

To generate missing data, we use the mechanisms MCAR (missing completely at random), MAR (missing at random), and MNAR (missing not at random) using different specifications of the correlation matrix of a multivariate Bernoulli distribution.

2 PROPOSED FRAMEWORK

The effect of missing data depends on data set aspects. Thus, we propose to control the sample size, number, and type of the input variables, and the correlation between these variables. We call a controlled independent variable as factor and call level the values that each factor assumes. Let $\mathbf{X}$ be a $n \times p$ matrix of input variables. We fixed the sample size at $n = 500, 1000$ and 10000, and for the input variables we fixed at $p = 10, 50$, and 200. To explore the impact of the type of input variables we designed three scenarios: i) the input variables are binaries generated from a Bernoulli distribution with success probability equals to 0.5, ii) the input variables are integers generated from a Poisson distribution with parameter equals to 10 and iii) the input variables are continuously generated from a standard Gaussian distribution. Finally, to simulate non-Gaussian correlated input variables, we adopted the NORTA (Normal to Anything) algorithm as it is implemented in the SIMCORMULTRES package for the statistical software R [11]. For increasing levels of redundancy between the input variables, we consider three correlation levels $\rho = 0, 0.5$, and 0.8.

Let $\mathbf{M}$ be an $n \times p$ indicator matrix representing the missing data process. Little and Rubin [5] proposed an extensive theory to determine the probability distribution of $\mathbf{M}$ given $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$, where $\mathbf{X}_{obs}$ contains all the elements $\mathbf{X}_{ij}$ where $\mathbf{M}_{ij} = 1$ and $\mathbf{X}_{mis}$ contains all the elements $\mathbf{X}_{ij}$ where $\mathbf{M}_{ij} = 0$. Thus, three types of missing data mechanisms are represented by the conditional distribution $f(\mathbf{M}|\mathbf{X}, \phi)$, where ϕ is a set of unknown parameters to describe the relationship between $\mathbf{M}$ and the data. We briefly discuss the missing data mechanisms, as follows:

- Missing Completely at Random (MCAR): It happens when the events that lead to the missing data are unrelated to the value of other observed or unobserved variables. We have in probabilistic terms:
$$f(\mathbf{M}|\mathbf{X}, \phi) = f(\mathbf{M}|\phi). \tag{1}$$
- Missing at Random (MAR): In this mechanism, the missing pattern is related to other observed values. One can think that the missing pattern observed for a particular input variable is a function, in general unknown, of the other input variables. We have in probabilistic terms:
$$f(\mathbf{M}|\mathbf{X}, \phi) = f(\mathbf{M}|\mathbf{X}_{obs}, \phi). \tag{2}$$
- Missing Not at Random (MNAR): In this mechanism, the variable correlated to missing data is not present in the data set. MNAR is the most complex to identify because the reasons for the missing data are unknown. We have in probabilistic terms:
$$f(\mathbf{M}|\mathbf{X}, \phi) = f(\mathbf{M}|\mathbf{X}_{mis}, \phi). \tag{3}$$

The simulation of the MCAR mechanism is done by simulating $\mathbf{M}$ from independent Bernoulli distributions, and the success probability determines the amount of missing data in the data set. On the other hand, the simulation of MAR and MNAR mechanisms are more complex. In this paper, we adopted a strategy based on a multivariate Bernoulli distribution, which in turn is specified by a vector $p \times 1$ of probabilities and a $p \times p$ covariance matrix.

Algorithm 1: Calibration algorithm

```
Input:
α: predictive power of input variables;
Output: The effect size value given simulation conditions
initialization;
begin
    for $u = 0$ to $10$ do
        generate $X$ using rnorta function given simulation
          conditions
        for $k = 0$ to $15$ do
            repeat
                $\Delta \longleftarrow$ Through the range (0,30) search the
                  root of the function that results 0.9
                  accuracy ;
                generate vector $\beta_j \sim G(\alpha)$ for $j = 1, \ldots, p$;
                $\beta' \longleftarrow \Delta * \beta_j / sum(\beta_j)$;
                generate $Y_i \sim B(p_i)$ shown in Equation 5;
                $mod \longleftarrow$ fit a logistic regression model;
                acc $\longleftarrow$ accuracy of the $mod$;
            until $acc = 0.9$;
            $eff \longleftarrow$ store each $\Delta$ value when $acc = 0.9$;
        end
        $effects \longleftarrow$ store all $eff$ vectors;
    end
    $effectSize \longleftarrow mean(effects)$;
    $return(effectSize)$;
end
```

In the MAR mechanism, the distribution of the missing values is controlled by the values of the observed input variables. Consequently, the columns of $\mathbf{M}$ should be correlated, since they are generated based on the same set of $\mathbf{X}$. Thus, we simulate $\mathbf{M}$ from a multivariate Bernoulli distribution fixing the correlation parameter at $\rho = 0.5$ and 0.8. Note that $\rho = 0$ corresponds to the MCAR mechanism. Finally, MNAR is the most challenging mechanism because the missing data pattern depends on unobserved input variables. We introduce an extra assumption: although the missing patterns depends on unobserved variables, the unobserved or latent variables induce an special pattern on the covariance matrix of the multivariate Bernoulli distribution. The idea is similar to factor analyses where the covariance pattern is interpreted as the effects of latent variables.

We specify the covariance matrix of the multivariate Bernoulli distribution as a diagonal block matrix, where each block is attributed to the effect of a latent variable. For simplicity, we divide the covariance matrix into two equal-size blocks. In the first block we use correlation $\rho = 0.5$, and in the second block we use $\rho = 0.8$. Finally, we combine all factors and their levels exhaustively to compose 2.187 simulation scenarios.

The next step is to simulate the target variable. We focus on the performance of the logistic regression model because of its popularity for classification problems. However, other predictive models could be evaluated in a similar way. In this context, given a set of input variables represented in a matrix $\mathbf{X}$ the target or response variable Y_i is simulated based on the logistic model,

$$Y_i \sim \mathrm{B}(p_i) \tag{4}$$

$$p_i = \mathrm{E}(Y_i|X_{ij}) = \mathrm{P}(Y_i = 1) = \frac{exp(X_{ij}^{\top}\boldsymbol{\beta})}{1 + exp(X_{ij}^{\top}\boldsymbol{\beta})}. \tag{5}$$

The vector $p \times 1$ of regression coefficients $\boldsymbol{\beta}$ is simulated from a geometric distribution, whose parameter α controls the predictive power of the input variables, i.e, $\beta_j \sim \Delta G(\alpha)$ for $j = 1, \ldots, p$. Also, we include a parameter Δ to control the expected accuracy of the predictive model. Algorithm 1 shows how to select Δ. It is important to keep the comparability of the results among the different simulation scenarios. We select a value Δ for each simulation scenario, such that we can then analyze the data set results in 90% of accuracy.

Given the structure of the geometric distribution, the first input variable is the most important, and the importance decreases exponentially. Δ is selected for each simulation scenario, but α was fixed at $0.2, 0.5$ and 0.8 to have one scenario where we have a uniform and asymmetric distribution for the regression coefficients, respectively.

We use the statistical software `R` and the function `glm()` to fit the logistic regression model. For each simulation scenario, we simulated 150 data sets, fit the model, and compute the accuracy. The last step can easily be done using the linear regression model, which allows us to quantify each factor's main effect on the accuracy of the logistic regression model, taking the uncertainty into account. The framework code is available on a public repository [1].

3 RESULTS

In this section, we present the results of our simulation study. Figure 1 presents an overview of our main results. For a better plot visualization, we opted to plot the results concerning the number and type of the input variables, sample sizes, and predictive power. Regarding the correlation between the input variables, we show only the most challenging scenario, i.e., the correlation of 0.8. Concerning the missing pattern mechanisms, we present one scenario for each of them, i.e., complete (COM), MCAR, MAR with correlation 0.5, and MNAR with correlations 0.5 and 0.8. For all scenarios, we show only the cases where we simulate 30% of missing data. Finally, we decided to summarise the results using bars representing the first and third quantiles. Thus, we evaluate the results taking into account the uncertainty associated with them in each scenario.

Figure 1 shows that for large sample sizes the missing data effects disappear, for all missing data mechanisms. It is also clear that the precision of the accuracy increases, i.e. narrower bars. On the other hand, as the number of input variables increases the accuracy decreases quickly, and consequently, the generalization power of the model is weak. The combination of small sample sizes and a large number of input variables is the worst scenario in terms of accuracy as well as the most affected by MAR and MNAR.

Regarding the type of input variables, we note an increase in the accuracy from Bernoulli to Poisson and Gaussian, respectively. Similarly, the accuracy tends to decrease from the MCAR to MAR and MNAR mechanisms. Overall the MNAR mechanism shows the worst results in terms of accuracy in all considered scenarios. However, when combined with large samples and a low number of input variables the accuracy is barely below the 90% level.

Parameter	Estimate	Parameter	Estimate
Intercept	79.61~(0.04)	TINPUT_Poisson	4.06~(0.03)
NINPUT_50	-4.52~(0.03)	TINPUT_Gauss	8.50~(0.03)
NINPUT_200	-19.53~(0.03)	PPINPUT_0.5	-0.07~(0.03)
SS_1000	6.31~(0.03)	PPINPUT_0.8	0.36~(0.03)
SS_10000	14.89~(0.03)	TMD_MCAR	-1.46~(0.04)
CINPUT_0.5	-0.08~(0.03)	TMD_MAR	-4.07~(0.04)
CINPUT_0.8	-0.23~(0.03)	TMD_MNAR	-4.67~(0.04)
		PMD_0.9	2.27~(0.03)

Table 1: Parameter estimates and standard errors.

The predictive power of the input variables is an important factor to determine the impact of the missing data on the predictive performance of the model. In general, when the predictive power is concentrated in a few continuous (Gaussian) input variables ($\alpha = 0.8$) the accuracy is less affected by the missing data than in the other scenarios. On the other hand, for binary inputs a more even distribution of the regression coefficients implies a smaller effect of the missing data mechanisms.

Finally, to measure the impact of the main factors considered in our simulation study we fit a multiple linear regression model. The response variable is the accuracy and the factors are: the number of input variables (NINPUT) with levels $(10, 50,$ and $200)$; sample size (SS) with levels $500, 1000,$ and 10000; correlation between the input variables (CINPUT) with levels $0, 0.5,$ and 0.8; type of the input variables (TINPUT) with levels Bernoulli, Gaussian and Poisson; predictive power of the input variables (PPINPUT) with levels $0.2, 0.5$ and 0.8; type of missing (TMD) data with levels MCAR, MAR and MNAR and proportion of missing data (PMD) with levels 0.7 and 0.9. Table 1 presents parameter estimates and standard error for the regression coefficients associated with each of the factors and their different levels. It is important to emphasise that we fit the model using only the main effects. We could fit the model using the interactive effects as well. However, only the main effects explain 67.91% of the accuracy variability, which we considered satisfactory.

Results in Table 1 quantify the effect of each factor on the expect value of the accuracy. Thus, we have that by increasing the number of input variables from 10 to 50 and 200 we expect an average decrease in the accuracy of 4.52% and 19.53%, respectively. Similarly, by increasing the sample size from 500 to 1000 and 10000 we expect an average accuracy increase of 6.31% and 14.89%, respectively. The correlation between the input variables and the distribution of their predictive power present lower impact on the accuracy. However, the type of the input variables is an important factor to explain the variability of the accuracy. The accuracy tends to increase in average 4.06% and 8.50% for Integers and continuous inputs in relation to binary inputs. Finally, concerning the type of missing data mechanism our results show that from the complete case to MCAR, MAR and MNAR we expect an accuracy decrease of 1.46%, 4.07% and 4.67%, respectively.

[1] https://github.com/fsantore/missing_data_analysis

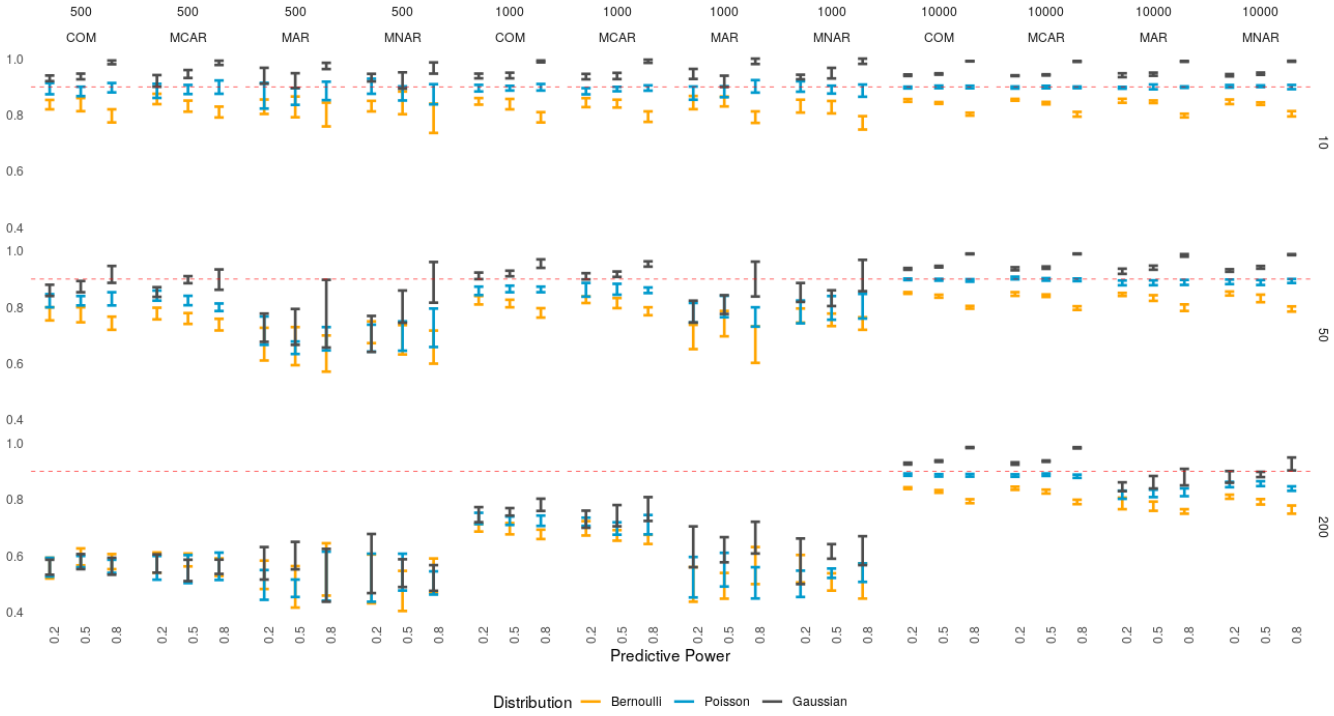

Figure 1: Accuracy quantile intervals by sample sizes and missing data mechanisms (X-Axis on the top), CINPUT and NINPUT (Y-Axis to the left and right), and predictive power of the input variables (X-Axis on the bottom).

4 DISCUSSION

This paper presents a framework to evaluate the impact of missing data in predictive models. Our framework uses a stochastic process to control different factors of the input variables: predictive power, sample size, number, type, and the correlation between each other. Besides, we consider different variations of missing data generation mechanisms, namely, MCAR, MAR, and MNAR. We use graphical tools and a linear regression model to quantify how each factor impacts the expected accuracy.

Our results show that, as the sample size increases, the missing data impact decreases drastically. On the other hand, as the number of input variables increases, the accuracy values decrease. That fact reinforces that input selection is essential in increasing the predictive power of a model. The type of input variables is also an essential factor to explain the accuracy of the model. In general continuous inputs benefit from better accuracy. Finally, the correlation between input variables and their missing mechanism presents a low impact on predictive performance.

The set of controlled factors explains the accuracy variability. Thus, we argue that our framework can effectively evaluate the impact of missing data and other essential aspects of a data set in the predictive performance of a logistic regression model. Our framework can be adopted as a practical framework to evaluate the effectiveness of new approaches that handle missing data. For future works, we suggest extending our framework to evaluate other predictive models, such as neural networks and random forest.

REFERENCES

[1] Edgar Acuna and Caroline Rodriguez. 2004. The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications*. Springer, 639–647.

[2] Gustavo E.A.P.A. Batista and Maria Carolina Monard. 2003. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence* 17, 5-6 (2003), 519–533.

[3] José M. Jerez, Ignacio Molina, Pedro J. García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine* 50, 2 (2010), 105–115.

[4] Wei-Chao Lin and Chih-Fong Tsai. 2019. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review* (2019), 1–23.

[5] Roderick JA Little and Donald B Rubin. 1987. *Statistical analysis with missing data*. Vol. 333. John Wiley & Sons.

[6] Julián Luengo, Salvador García, and Francisco Herrera. 2012. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and information systems* 32, 1 (2012), 77–108.

[7] Jason Poulos and Rafael Valle. 2018. Missing data imputation for supervised learning. *Applied Artificial Intelligence* 32, 2 (2018), 186–196.

[8] Maytal Saar-Tsechansky and Foster Provost. 2007. Handling missing values when applying classification models. *Journal of machine learning research* 8 (2007), 1623–1657.

[9] Claude Sammut and Geoffrey I. Webb. 2017. *Encyclopedia of Machine Learning and Data Mining* (2nd ed.). Springer Publishing Company, Incorporated.

[10] Qinbao Song, Martin Shepperd, Xiangru Chen, and Jun Liu. 2008. Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation. *Journal of Systems and Software* 81, 12 (2008), 2361–2370.

[11] Anestis Touloumis. 2016. Simulating Correlated Binary and Multinomial Responses under Marginal Model Specification: The SimCorMultRes Package. *The R Journal* 8, 2 (2016), 79–91.