UDC 519.245+004.896
DOI: 10.20535/SRIT.2308-8893.2025.1.03

EFFICIENCY COMPARISON OF MISSING DATA IMPUTATION METHODS IN PREDICTIVE MODEL CREATION

A. POPOV

Abstract. Missing data is a common issue in data analysis and machine learning. This article analyzes the impact of missing data imputation methods during the data preprocessing stage on the quality of forecasting models. Selected methods are listwise deletion, mean imputation, and two implementations of the multiple imputation method in Python and R languages. Selected classifiers are Logistic Regression, Random Forest, Support Vector Machine, and Light Gradient Boosting Machine. The performance quality of forecasting models is estimated using accuracy, precision, and recall metrics. Two datasets were used as binary classification problems with different target metrics. The highest performance was achieved when the R implementation of the multiple imputation method was combined with RF and LGBM classifiers.

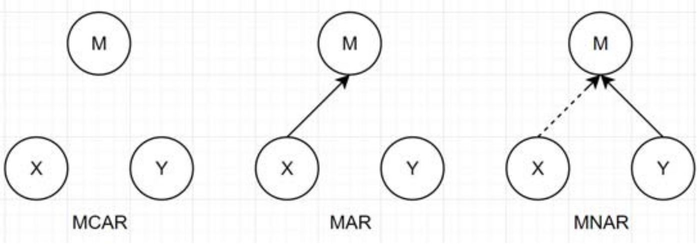
Keywords: missing data, imputation methods, forecasting models, machine learning.

INTRODUCTION

Today, every forecasting task involves processing large amounts of information. One of the key aspects of preparing data for creating predictive models is handling missing values, as machines learning algorithms mostly require complete data. In real-world datasets, it is common to find gaps that can occur for a variety of reasons, such as technical issues, human errors, the specifics of the research in which the data was collected, and other factors. Missing information in a dataset can distort statistical parameters, which can have a serious impact on the quality and reliability of the model and lead to incorrect conclusions. With proper handling of missing data prior to model training, the probability of successful training of a predictive model can be increased, which will positively affect its quality.

MISSING DATA MECHANISMS

To describe the logic behind the occurrence of missing data, the concept of a missing data mechanism was created. A mechanism is a term that is meant to describe in a general way the relationship between missing and observed data. According to the most common classification, there are three types of mechanisms based on what determines the probability of missing a particular variable in the observation: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR; Rubin 1976 [1]). MCAR is the case when the missingness is completely random, i.e., independent of the values of the variables in the set. This mechanism usually poses the least problems for imputation, since the statistical parameters of the data are by definition not biased. MAR — the absence of data on a particular variable depends on the values of other variables in the set, but does not depend on the value of the variable itself. This mechanism inherently contains bias and requires more careful handling than MCAR. MNAR — the absence of a variable depends on the value of the variable itself. This is the most complicated mechanism that does not have a clearly described solution, and in this case, the processing of missing data requires a specialised approach. Figure shows a simplified diagram of the mechanisms, where Y is the variable in question, M is an indicator of missing data for Y, X is other variables, a solid arrow is an existing dependency, and a dashed arrow is a possible dependency.



Simplified diagram of missing data mechanisms

KEY STAGES OF IMPUTATION

Data analysis. Determining statistical parameters of the available data, analysing relations between variables, correlations, identifying missing data, analysing them if possible, and determining the mechanism of their formation. The purpose of this stage is to gain an understanding of the available data and, ideally, the missing data, which will greatly facilitate the process of filling in the data.

Method selection for processing missing data. The large variety of available methods allows choosing the most suitable option for a particular task. The choice may depend on the amount of missing data, the mechanism behind it and complexity. Simple single-imputation methods (such as mean, mode, interpolation) are very popular and generally accepted, they are easier to understand and implement, but have disadvantages that limit their use. Sophisticated methods usually provide better imputation because they are able to take into account the relations between the data and do not skew the statistical parameters as a result, thus there are fewer limitations to their use. [2] In many cases, it makes sense to choose several methods and compare the results to choose the most appropriate one for the task at hand.

Performing the imputation. Application of the selected methods to fill in the missing values based on the observed data. This stage results in a complete dataset. The statistical quality of the imputation may depend on the nature of the gaps, the number of gaps, and the selected method. The wrong choice of method can lead to significant distortion of the results.

RELATED WORK

The topic of missing data processing is addressed in a large number of different studies, since it appears in any field and can be solved in a variety of ways with different levels of efficiency. With the accelerating development of artificial intelligence and machine learning technologies, the topic of missing data processing has become even more discussed — high-quality process modelling in any field requires high-quality data, which creates the necessity of efficient processing of missing values. Research on methods is diverse, and depends on the goal of the researchers: some papers address general issues, review methods, and propose solutions [3–7]. In other works, there is a specific problem and methods for solving it are considered.

In particular, in the paper "The impact of imputation quality on machine learning classifiers for datasets with missing values" [8], the authors study the impact of imputation methods on the predictive ability of models. The methods studied are mean imputation, multiple imputation by chained equations (MICE), MissForest, generative adversarial imputation networks (GAIN), and missing data importance-weighted autoencoder (MIWAE); the selected datasets include both complete datasets with MCAR gaps of 25–50% and datasets with intrinsic MNAR gaps. The models under study are logistic regression, random forest, XGBoost, and artificial neural network. The selected datasets are used to compare the results between the trained models. The paper uses a multivariate ANOVA model to evaluate the impact of the imputation on the quality of the models. The results show that the quality of predictive models depends on the amount of missing data and training on imputed datasets usually produces lower quality results compared to training on complete datasets. At the same time, for the same dataset, the quality ranking of the models usually does not change for different amounts of missing data, i.e. a model that performs better on 25% of missing data will also perform better on 50% of the missing data. Different methods perform better depending on the dataset, but some imputation methods have less variation in quality across datasets, with MIWAE consistently performing well across the study. In some cases, logistic regression, which typically has the worst quality metrics, was also able to achieve high quality metrics.

Another paper by Jale Bektas, Turgay Ibrikci, and Ismail Turkay Ozcan [9] investigates the impact of imputation methods on the quality of classifiers in the task of diagnosing coronary artery disease. In this paper, three imputation methods based on machine learning techniques (K-means, multilayer perceptron, and self-organising maps) are presented and their performance is compared with the conventional mean imputation method and listwise deletion. The selected classification methods were Logistic Model Trees (LMT), multilayer perceptron, random forest method, and support vector machine. The developed imputation methods showed significantly better results than the mean imputation method, which was ranked fourth in terms of model quality, surpassed only by the listwise deletion. The best results were achieved when using self-organising maps (88.23% accuracy), and the most stable results were obtained when using a multilayer perceptron.

The papers "Do we really need imputation in AutoML predictive modelling?" [10] and "Does imputation matter? Benchmark for predictive models" [11] investigate the necessity of using complex imputation methods in machine learning processes. In the first study, 6 imputation methods were used to process data in 25 datasets with natural missing data and 10 datasets with artificial missing data. In the second one, 7 imputation methods were used on 13 classification tasks. The conclusions of both papers are that simple methods usually perform slightly worse than more complex methods, while gaining considerably in computational power. The first paper found that using a binary indicator with simple mean/mode imputation (for continuous and categorical data, respectively) performed well and was significantly more efficient than more complex methods. In the second paper, simple methods also achieved good results, although even with similar predictive quality of the models, more complex methods produced more statistically accurate imputations.

In summary, the use of imputation methods at the stage of data preprocessing is a common subject in machine learning, with wide application regardless of the specific field of study.

STATEMENT OF THE RESEARCH PROBLEM

The purpose of this paper is to investigate the impact of missing data processing method on the quality of predictive machine learning models. In the process, we take complete datasets and using them as basis we artificially create datasets with different missing data configurations to study the effect of imputations on the predictive ability of models. All datasets are taken from the public domain.

The research algorithm consists of the following general steps: selection of imputation methods for the study, selection of prediction methods, search and research of datasets, creation of datasets with missing data, processing missing data, training models on the obtained datasets, and analysis of the results.

Four imputation methods were selected for the study:

1. Listwise deletion.
2. Mean imputation.
3. Multiple imputation using Python library scikit-learn (Iterative Imputer).
4. Multiple imputation using R library MICE.

The following 4 algorithms were chosen as forecasting algorithms:

1. Logistic regression.
2. Support vector machine.
3. Random Forest.
4. LGBM (Light Gradient Boosting Machine).

The first selected dataset is the Churn dataset of bank customers, the task of classification is to determine customer churn, i.e. to identify customers who are likely to cancel their bank services based on the available data. The selected dataset consists of 10.000 records and 10 variables, including 4 continuous and 6 categorical variables. The continuous variables are: *CreditScore* — customer's credit score, *EstimatedSalary* — customer's estimated salary, *Age* — customer's age, and *Balance* — customer's balance. Categorical variables include: *Geography* — country of origin of the customer, *Gender* — gender of the customer, *Tenure* — number of years the customer has been with the bank, *NumOfProducts* — number of bank products used by the customer, *HasCrCard* — indicator of whether the customer has a bank credit card, *isActiveMember* — indicator of customer activity, and *Exited* — target variable reflecting the churn/retention status of the customer.

To handle missing data, we use only continuous variables. In total, 12 datasets with different types and numbers of gaps were created, including 4 datasets with only MCAR gaps, 2 datasets with only MAR gaps, 1 dataset with only MNAR gaps, 2 datasets with mixed MCAR and MAR gaps (such datasets are considered in the MAR category), and 2 datasets with mixed gaps using MNAR gaps. As a result, 48 datasets were obtained after completion of the imputation. [12] Datasets with mixed gaps are considered in the category of a less strong assumption — for example, for mixed MAR and MCAR gaps, the dataset is considered in the MAR category.

The second selected dataset is a set of characteristic parameters of wine for the purpose of wine quality classification. The selected dataset consists of 1599 records and 12 variables, of which 11 are continuous and 1 is a categorical variable. The categorical variable is the target variable quality. The continuous variables are: *fixed acidity* — fixed (nonvolatile) acids, *volatile acidity* — the amount of volatile acids, *citric acid* — the amount of citric acid, *residual sugar* — the amount of residual sugar after the fermentation process is stopped, *chlorides* — the amount of salt, *free sulfur dioxide* — the free form of SO_2 that exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ion, *total sulfur dioxide* — the amount of free and bound SO_2, *density* — the density (the density of wine is almost the same as that of water, depending on the alcohol and sugar content), *pH* — an indicator of the acidity/alkalinity of wine from 0 to 14 (most wines are between 3–4 on this scale) and *sulphates* — additives to wine that can contribute to the level of SO_2.

In total, 9 datasets with different types and numbers of gaps were created, including 3 datasets with exclusively MCAR gaps, 2 datasets with exclusively MAR gaps, 1 dataset with exclusively MNAR gaps, 1 dataset with mixed MCAR and MAR gaps, and 2 datasets with mixed gaps using MNAR gaps. As a result, 36 datasets were obtained after the completion of the imputation.

PERFORMANCE METRICS

To evaluate the quality of the obtained predictive models, we used the accuracy, precision and recall metrics based on the confusion matrix.

- Accuracy is a metric of the overall classification accuracy of the model, calculated as the ratio of correct predictions to all predictions.
- Precision is a metric that shows how many positive predictions were correct.
- Recall is a metric that shows how many elements of a positive class were detected by the model.

For each dataset, one of the metrics is the target metric, i.e. the main quality criterion in the context of a particular task. The quality comparison was performed for the values of the target metrics for the models trained on the imputed datasets.

In the Churn dataset, the target metric is recall, since the most important ability of the model should be the ability to correctly identify customers who will leave. In the Wine dataset, the target metric is accuracy, since the accuracy of classification is equally important for both classes.

CREATION OF ARTIFICIAL MISSING DATA

Missing data was created with different combinations of mechanisms and quantities. For the purpose of more accurate comparison, the gaps were created exclusively in the training dataset — this was done in order to compare the classification quality of different models on the same test set. Before creating missing data, the full datasets were split into training and test samples in the ratio of 80 to 20.

The number of MCAR missing data for each selected variable ranges from 5% to 20%, and the total number of observations with gaps in the MCAR datasets ranges from 9.72% to 47.54%. To create MAR missing data two variables were selected and the values of the first variable were removed for records that had values for the second variable below or above the selected percentile. The selected percentiles ranged from 5% to 20% for values below them and 90–95% for values above them. The total number of observations with missing data ranged from 21.59% to 48.83%. To create the MNAR type of missing data, a variable was selected and those values below or above the selected percentile were removed, which ranged from 7 to 13% and from 90% to 93%, respectively. The total number of observations with missing data ranged from 23.77% to 45.97%.

The total number of records with missing data for the datasets derived from the first dataset ranged from 9.72% to 48.82%. The average number was 30.7%. The size of the full training dataset was 8000 records and the test dataset was 2000 records.

The total number of records with missing data for the datasets derived from the second dataset ranged from 18.14% to 45.97%, with an average of 33%. The size of the full training set was 1279 records, and the test set was 320 records.

MODEL TRAINING RESULTS

The performance metrics of models trained on complete datasets are presented in Table 1. The highest predictive quality was achieved with LGBM model for the Churn dataset and RF model for Wine dataset.

Table 1. Results of training on complete datasets

Churn				**Wine**			
Model	Accuracy	Precision	Recall	**Model**	Accuracy	Precision	Recall
LR	0.725000	0.386397	0.679389	**LR**	0.784375	0.392157	0.851064
SVC	0.799500	0.492982	0.715013	**SVC**	0.859375	0.513514	0.808511
RF	0.808500	0.508711	0.743003	**RF**	0.8625	0.518072	0.914894
LGBM	0.820500	0.530357	0.755725	**LGBM**	0.840625	0.476744	0.87234

Tables 2, 3, 4 present the performance metrics of models trained on imputed Churn datasets with MCAR, MAR and MNAR missing data respectively.

Table 2. Results of training on imputed Churn datasets with MCAR missing data

MCAR Churn													
	Model	9.72%			18.46%			22.03%			47.54%		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Listwise	**LR**	0.724	0.383602	0.666667	0.678	0.348247	0.732824	0.678	0.348247	0.732824	0.7285	0.387048	0.653944
	SVC	0.7945	0.484211	0.70229	0.805	0.502712	0.707379	0.79	0.477686	0.735369	0.805	0.502879	0.666667
	RF	0.8085	0.508865	0.73028	0.8015	0.496587	0.740458	0.8005	0.494845	0.732824	0.8135	0.518587	0.709924
	LGBM	0.812	0.514834	0.750636	0.804	0.500846	0.753181	0.8095	0.510345	0.753181	0.812	0.515371	0.725191
	Avg	0.78475	0.472878	0.712468	0.772125	0.462098	0.733461	0.7695	0.457781	0.73855	0.78975	0.480971	0.688932
Mean	**LR**	0.7285	0.390988	0.684478	0.7245	0.386494	0.684478	0.7275	0.389535	0.681934	0.721	0.382646	0.684478
	SVC	0.795	0.48532	0.715013	0.799	0.492091	0.712468	0.8025	0.498246	0.722646	0.812	0.515315	0.727735
	RF	0.802	0.497453	0.745547	0.7925	0.481544	0.73028	0.803	0.499145	0.743003	0.7995	0.493151	0.732824
	LGBM	0.817	0.52356	0.763359	0.8145	0.518966	0.765903	0.812	0.51463	0.760814	0.805	0.502555	0.750636
	Avg	0.785625	0.47433	0.727099	0.782625	0.469774	0.723282	0.78625	0.475389	0.727099	0.784375	0.473417	0.723918
Iterative	**LR**	0.6745	0.345694	0.735369	0.6745	0.345694	0.735369	0.678	0.348247	0.732824	0.68	0.349206	0.727735
	SVC	0.809	0.509874	0.722646	0.8115	0.514235	0.735369	0.8115	0.514337	0.73028	0.818	0.527938	0.697201
	RF	0.8385	0.573529	0.694656	0.8355	0.568085	0.679389	0.8415	0.581197	0.692112	0.8245	0.542339	0.684478
	LGBM	0.841	0.585421	0.653944	0.8415	0.584071	0.671756	0.8365	0.571121	0.6743	0.818	0.527619	0.704835
	Avg	0.79075	0.50363	0.701654	0.79075	0.503021	0.705471	0.791875	0.503726	0.707379	0.785125	0.486776	0.703562
MICE	**LR**	0.728	0.390421	0.684478	0.726	0.387844	0.681934	0.7265	0.387755	0.676845	0.7005	0.36658	0.720102
	SVC	0.795	0.485062	0.70229	0.798	0.490435	0.717557	0.8005	0.4947	0.712468	0.813	0.517625	0.709924
	RF	0.801	0.495798	0.750636	0.8085	0.508961	0.722646	0.8005	0.495	0.755725	0.814	0.519626	0.707379
	LGBM	0.8175	0.524735	0.755725	0.8125	0.515789	0.748092	0.8155	0.521053	0.755725	0.824	0.538752	0.725191
	Avg	0.785375	0.474004	0.723282	0.78625	0.475757	0.717557	0.78575	0.474627	0.725191	0.787875	0.485646	0.715649

Table 3. Results of training on imputed Churn datasets with MAR missing data

MAR Churn													
	Model	21.59%			26.16% (+ MCAR)			39.99%			48.73% (+ MCAR)		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Listwise	**LR**	0.7235	0.385714	0.687023	0.6605	0.336758	0.750636	0.7315	0.395349	0.692112	0.737	0.399698	0.6743
	SVC	0.791	0.479132	0.73028	0.7915	0.480198	0.740458	0.7755	0.454984	0.720102	0.7725	0.449838	0.707379
	RF	0.787	0.473083	0.737913	0.798	0.490566	0.727735	0.8075	0.507143	0.722646	0.7835	0.466667	0.712468
	LGBM	0.803	0.499139	0.737913	0.8085	0.509158	0.707379	0.8105	0.512411	0.735369	0.805	0.502636	0.727735
	Avg	0.776125	0.459267	0.723282	0.764625	0.45417	0.731552	0.78125	0.467472	0.717557	0.7745	0.45471	0.705471
Mean	**LR**	0.7235	0.384058	0.6743	0.716	0.376934	0.681934	0.7215	0.381159	0.669211	0.7225	0.383285	0.676845
	SVC	0.7965	0.487931	0.720102	0.792	0.480475	0.720102	0.797	0.488927	0.73028	0.795	0.485114	0.704835
	RF	0.802	0.497427	0.737913	0.814	0.519409	0.715013	0.7855	0.471061	0.745547	0.795	0.485904	0.745547
	LGBM	0.8135	0.517241	0.763359	0.8175	0.525926	0.722646	0.81	0.511149	0.75827	0.801	0.495881	0.765903
	Avg	0.783875	0.471664	0.723919	0.784875	0.475686	0.709924	0.7785	0.463074	0.725827	0.778375	0.462546	0.723283
Iterative	**LR**	0.68	0.349206	0.727735	0.6765	0.346618	0.73028	0.68	0.349206	0.727735	0.689	0.356336	0.722646
	SVC	0.811	0.513711	0.715013	0.814	0.518717	0.740458	0.8125	0.515845	0.745547	0.8095	0.511111	0.70229
	RF	0.8435	0.58658	0.689567	0.837	0.572043	0.676845	0.835	0.567452	0.6743	0.834	0.564482	0.679389
	LGBM	0.845	0.595402	0.659033	0.8355	0.568966	0.671756	0.837	0.574944	0.653944	0.818	0.527619	0.704835
	Avg	0.794875	0.511225	0.697837	0.79075	0.501586	0.704835	0.791125	0.501862	0.700382	0.787625	0.489887	0.70229
MICE	**LR**	0.7235	0.384058	0.6743	0.7265	0.388081	0.679389	0.722	0.382055	0.671756	0.7265	0.387755	0.676845
	SVC	0.791	0.478336	0.70229	0.8025	0.498258	0.727735	0.7895	0.476271	0.715013	0.8075	0.50738	0.699746
	RF	0.8075	0.506849	0.753181	0.813	0.516522	0.755725	0.7965	0.488294	0.743003	0.8155	0.522556	0.707379
	LGBM	0.8125	0.515625	0.755725	0.813	0.516579	0.753181	0.812	0.51468	0.75827	0.822	0.533821	0.743003
	Avg	0.783625	0.471217	0.721374	0.78875	0.47986	0.729008	0.78	0.465325	0.722011	0.792875	0.487878	0.706743

Table 4. Results of training on imputed Churn datasets with MNAR missing data

MNAR Churn													
	Model	24.01%			31.74% (+ MAR)			35.81% (+ MCAR)			42.43% (+ MCAR/MAR)		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Listwise	**LR**	0.7205	0.382436	0.687023	0.6605	0.336758	0.750636	0.65	0.331873	0.770992	0.65	0.331873	0.770992
	SVC	0.779	0.460292	0.722646	0.79	0.47644	0.694656	0.7925	0.480903	0.704835	0.779	0.459098	0.699746
	RF	0.8015	0.496503	0.722646	0.812	0.515654	0.712468	0.817	0.525424	0.709924	0.813	0.517691	0.707379
	LGBM	0.807	0.506087	0.740458	0.815	0.521024	0.725191	0.794	0.484034	0.732824	0.777	0.459168	0.75827
	Avg	0.777	0.46133	0.718193	0.769375	0.462469	0.720738	0.763375	0.455559	0.729644	0.75475	0.441958	0.734097
Mean	**LR**	0.7185	0.381616	0.697201	0.715	0.377931	0.697201	0.7125	0.375683	0.699746	0.7135	0.377049	0.70229
	SVC	0.769	0.445498	0.717557	0.7665	0.440895	0.70229	0.774	0.453249	0.727735	0.7755	0.454397	0.709924
	RF	0.771	0.450077	0.745547	0.781	0.464115	0.740458	0.7935	0.482759	0.712468	0.797	0.488774	0.720102
	LGBM	0.799	0.492487	0.750636	0.806	0.504488	0.715013	0.803	0.499086	0.694656	0.8055	0.503663	0.699746
	Avg	0.764375	0.44242	0.727735	0.767125	0.446857	0.713741	0.77075	0.452694	0.708651	0.772875	0.455971	0.708016
Iterative	**LR**	0.671	0.342823	0.735369	0.671	0.342823	0.735369	0.671	0.342823	0.735369	0.671	0.342823	0.735369
	SVC	0.8005	0.494505	0.687023	0.801	0.495379	0.681934	0.797	0.48816	0.681934	0.799	0.491682	0.676845
	RF	0.8265	0.545276	0.704835	0.822	0.535783	0.704835	0.83	0.555324	0.676845	0.8215	0.535714	0.687023
	LGBM	0.8265	0.546939	0.681934	0.826	0.546012	0.679389	0.8255	0.546025	0.664122	0.8255	0.545267	0.6743
	Avg	0.781125	0.482386	0.702290	0.78	0.479999	0.700382	0.780875	0.483083	0.689568	0.77925	0.478872	0.693384
MICE	**LR**	0.7225	0.385593	0.694656	0.7245	0.387464	0.692112	0.719	0.380481	0.684478	0.72	0.383543	0.699746
	SVC	0.792	0.480207	0.709924	0.7885	0.474832	0.720102	0.785	0.468908	0.709924	0.793	0.48199	0.715013
	RF	0.7925	0.482315	0.763359	0.79	0.478049	0.748092	0.7995	0.493197	0.737913	0.8035	0.5	0.709924
	LGBM	0.806	0.504303	0.745547	0.804	0.500855	0.745547	0.8185	0.527372	0.735369	0.798	0.490787	0.745547
	Avg	0.77825	0.463105	0.728372	0.77675	0.4603	0.726463	0.7805	0.46749	0.716921	0.778625	0.46408	0.717558

Tables 5, 6, 7 present the performance metrics of models trained on imputed Wine datasets with MCAR, MAR and MNAR missing data respectively.

Table. 5. Results of training on imputed Wine datasets with MCAR missing data

MCAR Wine										
	Model	18.14%			30.73%			45.35%		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Listwise	**LR**	0.796875	0.408163	0.851064	0.81875	0.438202	0.829787	0.809375	0.425532	0.851064
	SVC	0.85625	0.506667	0.808511	0.865625	0.529412	0.765957	0.86875	0.537313	0.765957
	RF	0.86875	0.533333	0.851064	0.84375	0.481013	0.808511	0.865625	0.532258	0.702128
	LGBM	0.853125	0.5	0.87234	0.8625	0.519481	0.851064	0.84375	0.478873	0.723404
	Avg	0.84375	0.487041	0.845745	0.847656	0.492027	0.81383	0.846875	0.493494	0.760638
Mean	**LR**	0.803125	0.418367	0.87234	0.79375	0.405941	0.87234	0.796875	0.41	0.87234
	SVC	0.859375	0.513514	0.808511	0.8375	0.46988	0.829787	0.85	0.493506	0.808511
	RF	0.84375	0.483146	0.914894	0.85	0.494253	0.914894	0.85	0.494253	0.914894
	LGBM	0.8375	0.47191	0.893617	0.8375	0.471264	0.87234	0.834375	0.465909	0.87234
	Avg	0.835938	0.471734	0.872341	0.829688	0.460335	0.87234	0.832813	0.465917	0.867021
Iterative	**LR**	0.784375	0.392157	0.851064	0.7875	0.39604	0.851064	0.809375	0.427083	0.87234
	SVC	0.86875	0.534247	0.829787	0.8625	0.519481	0.851064	0.875	0.547945	0.851064
	RF	0.859375	0.512195	0.893617	0.86875	0.530864	0.914894	0.865625	0.526316	0.851064
	LGBM	0.840625	0.476744	0.87234	0.834375	0.464286	0.829787	0.840625	0.475	0.808511
	Avg	0.838281	0.478836	0.861702	0.838281	0.477668	0.861702	0.847656	0.494086	0.845745
MICE	**LR**	0.7875	0.398058	0.87234	0.784375	0.392157	0.851064	0.809375	0.425532	0.851064
	SVC	0.86875	0.534247	0.829787	0.859375	0.512821	0.851064	0.875	0.547945	0.851064
	RF	0.85	0.494253	0.914894	0.865625	0.52439	0.914894	0.86875	0.534247	0.829787
	LGBM	0.83125	0.460674	0.87234	0.853125	0.5	0.87234	0.8625	0.518519	0.893617
	Avg	0.834375	0.471808	0.87234	0.840625	0.482342	0.872341	0.853906	0.506561	0.856383

Table. 6. Results of training on imputed Wine datasets with MAR missing data

MAR Wine										
	Model	23.53%			34.48% (+ MCAR)			42.30%		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Listwise	**LR**	0.828125	0.45122	0.787234	0.815625	0.428571	0.765957	0.809375	0.420455	0.787234
	SVC	0.853125	0.5	0.659574	0.821875	0.434211	0.702128	0.80625	0.363636	0.425532
	RF	0.8625	0.525424	0.659574	0.84375	0.477612	0.680851	0.846875	0.48	0.510638
	LGBM	0.86875	0.538462	0.744681	0.8375	0.467532	0.765957	0.85625	0.507692	0.702128
	Avg	0.853125	0.503777	0.712766	0.829688	0.451982	0.728723	0.829688	0.442946	0.606383
Mean	**LR**	0.790625	0.397959	0.829787	0.8	0.414141	0.87234	0.8	0.412371	0.851064
	SVC	0.85625	0.507042	0.765957	0.8375	0.469136	0.808511	0.8375	0.467532	0.765957
	RF	0.853125	0.5	0.87234	0.84375	0.481481	0.829787	0.8625	0.519481	0.851064
	LGBM	0.8375	0.47191	0.893617	0.825	0.448276	0.829787	0.828125	0.455556	0.87234
	Avg	0.834375	0.469228	0.840425	0.826563	0.453259	0.835106	0.832031	0.463735	0.835106
Iterative	**LR**	0.821875	0.445652	0.87234	0.8125	0.43299	0.893617	0.815625	0.434783	0.851064
	SVC	0.865625	0.527027	0.829787	0.83125	0.453333	0.723404	0.85	0.493151	0.765957
	RF	0.85625	0.506329	0.851064	0.859375	0.512821	0.851064	0.86875	0.535211	0.808511
	LGBM	0.853125	0.5	0.893617	0.8375	0.469136	0.808511	0.859375	0.512821	0.851064
	Avg	0.849219	0.494752	0.861702	0.835156	0.46707	0.819149	0.848438	0.493992	0.819149
MICE	**LR**	0.8125	0.430108	0.851064	0.8125	0.43299	0.893617	0.81875	0.43956	0.851064
	SVC	0.8625	0.52	0.829787	0.85625	0.507042	0.765957	0.84375	0.479452	0.744681
	RF	0.85625	0.506173	0.87234	0.859375	0.513158	0.829787	0.859375	0.513514	0.808511
	LGBM	0.834375	0.464286	0.829787	0.84375	0.481928	0.851064	0.846875	0.488095	0.87234
	Avg	0.841406	0.480142	0.845745	0.842969	0.48378	0.835106	0.842188	0.480155	0.819149

Table 7. Results of training on imputed Wine datasets with MNAR missing data

MNAR Wine										
	Model	23.77%			32.60% (+ MCAR)			45.97% (+ MCAR/MAR)		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Listwise	LR	0.796875	0.404255	0.808511	0.796875	0.404255	0.808511	0.828125	0.452381	0.808511
	SVC	0.8375	0.463768	0.680851	0.809375	0.397059	0.574468	0.859375	0.52	0.553191
	RF	0.853125	0.5	0.702128	0.825	0.44	0.702128	0.853125	0.5	0.553191
	LGBM	0.834375	0.461538	0.765957	0.83125	0.454545	0.744681	0.8375	0.454545	0.531915
	Avg	0.830469	0.45739	0.739362	0.815625	0.423965	0.707447	0.844531	0.481732	0.611702
Mean	LR	0.79375	0.402062	0.829787	0.790625	0.395833	0.808511	0.809375	0.423913	0.829787
	SVC	0.83125	0.452055	0.702128	0.834375	0.460526	0.744681	0.8625	0.52381	0.702128
	RF	0.83125	0.455696	0.765957	0.8375	0.469136	0.808511	0.85	0.493151	0.765957
	LGBM	0.825	0.445783	0.787234	0.825	0.448276	0.829787	0.834375	0.464286	0.829787
	Avg	0.820313	0.438899	0.771277	0.821875	0.443443	0.797873	0.839063	0.47629	0.781915
Iterative	LR	0.803125	0.418367	0.87234	0.80625	0.421053	0.851064	0.809375	0.423913	0.829787
	SVC	0.853125	0.5	0.787234	0.8625	0.521739	0.765957	0.8625	0.52381	0.702128
	RF	0.846875	0.486842	0.787234	0.84375	0.481013	0.808511	0.85	0.493151	0.765957
	LGBM	0.840625	0.47561	0.829787	0.83125	0.455696	0.765957	0.834375	0.464286	0.829787
	Avg	0.835938	0.470205	0.819149	0.835938	0.469875	0.797872	0.839063	0.47629	0.781915
MICE	LR	0.796875	0.40625	0.829787	0.80625	0.419355	0.829787	0.803125	0.416667	0.851064
	SVC	0.840625	0.474359	0.787234	0.85	0.493333	0.787234	0.878125	0.560606	0.787234
	RF	0.88125	0.561644	0.87234	0.86875	0.534247	0.829787	0.859375	0.513889	0.787234
	LGBM	0.8375	0.46988	0.829787	0.83125	0.45679	0.787234	0.834375	0.460526	0.744681
	Avg	0.839063	0.478033	0.829787	0.839063	0.475931	0.808511	0.84375	0.487922	0.792553

DISCUSSION

Listwise deletion, which is a highly problematic method from the statistical point of view and is rarely recommended, has proven in some cases to be able to produce datasets that yield prediction quality that is as good as when sophisticated imputation methods are used. The method performs better with the Wine dataset, where the target metric is accuracy, and performs worse with the Churn dataset when the target metric is recall. In particular, for Churn on datasets with MCAR missing data, the method showed mixed and unpredictable results. In some cases, the obtained value of the target metric was not worse than the results obtained using other methods, but the same model could have very different metric values on different datasets, which made it difficult to predict the result. On datasets with MAR missing data, the recall value was as good as the other methods, but it also often increased at the cost of the accuracy value, so the overall quality of the models was lower. Similar results can be seen on the datasets with MNAR missing data. In addition, on these datasets, the highest recall score was achieved using logistic regression, which usually shows the worst results. Thus, for the recall as target metric, good results using this method are not uncommon, but the best quality models are obtained on datasets that have exclusively MCAR mechanism — otherwise, the overall quality of the model decreases.

For the Wine dataset with the accuracy target metric, the method's performance was significantly higher. Despite the loss of a large amount of information, the predictive quality of the obtained models was not inferior to other methods. It is worth noting that the value of the recall metric was significantly lower than that of the other methods, especially as the number of missing data increased, which resulted in lower overall model quality. The most balanced models were obtained on datasets containing only MCAR missing data: accuracy ranged from 79.7% to

86.9%, recall from 76.7% to 87%, which matches the quality of models obtained using more complex methods. On the datasets with MAR and MNAR missing data, the trained models showed good results in terms of accuracy (79.7–86.9%), but the values of the recall metric were significantly lower (51–80.1%).

In summary, in both problems, it was observed that the method is not a reliable choice for the recall metric, as satisfactory and predictable results were obtained only on MCAR missing data. At the same time, the method is able to show very good results when working with the accuracy metric, but is still limited by the MCAR missing data mechanism and the percentage of missing data to obtain balanced models for the metrics. Due to the general unreliability and unpredictability of the method, it can be concluded that it is not the best choice for solving such problems, but its use does not necessarily mean obtaining unsatisfactory results, because predictive models can often learn to correctly identify the features of the target classes of the problem even using datasets with biased statistical parameters.

Mean imputation generally showed more reliable results on most datasets than listwise deletion, as the quality of trained models fluctuated less regardless of the type and number of missing data. For this method, the best results were achieved with SVC, RF or LGBM classifiers, while the method performed worse with logistic regression. From a statistical point of view, a significant problem with this method is the reduction of data variability and weakening of correlations between variables (which was observed for the datasets imputed with this method), but, as in the case of listwise deletion, machine learning models are able to learn to identify features of the target class even with statistically skewed data, and they do so with greater success for the mean imputation method. Using this method, satisfactory results were obtained on MCAR and MAR missing data for both datasets for three out of four methods: for Churn, the accuracy was in the range of 71–81% in almost all cases, recall was 71–76%, and only when using logistic regression were the results worse (recall 67–68%); for the Wine dataset, the accuracy was 79–86.2%, recall was 80–91.5% (exceptions are two cases of SVC method on MAR, where recall was 76.6%). The training results on MNAR missing data were of lower quality, with a noticeable decrease in recall compared to other methods: for Churn, the metric had results of 69.4–75%, and for Wine — 70–83%. The best results in these cases were achieved using LGBM (Churn) and RF or SVC (Wine). In summary, using mean imputation method is a relatively good choice, as the models trained on these datasets were of high quality more often and had more predictable results than those using listwise deletion. In addition, the method also performs better because the range of values obtained for the metrics, even for the worst outcomes, is smaller, making the results more predictable.

Multiple imputation in the Python implementation of IterativeImputer from the scikit-learn library showed unsatisfactory results for the Churn dataset. The models often did not meet the minimum required classification quality. Across all the missing data mechanisms, it was observed that this method worked best with logistic regression and support vector machine — in particular, logistic regression repeatedly showed significantly better results on the recall metric when using this method compared to the complete data (up to 73.5%) — but fell short on other metrics (67–68% accuracy). The SVC models performed relatively well (accuracy 80–81%, recall 70–74.5%), except for datasets with MNAR-type missing data (accuracy 80%, recall 68%). The RF and LGBM models consistently had low recall values in combination with this method (67–70%), regardless of the mechanism and number of gaps.

On the Wine dataset, by contrast, the method performed quite well on MCAR and MAR missing data, especially when SVC and RF models were used (accuracy 83–87.5%, recall 82.9–91.5% with two exceptions on MAR data for SVC). Logistic regression generally performed worse than the other models, but often had the highest recall, which was also observed on the Churn dataset. On the MNAR missing data, the metrics were also quite high and did not fall short of other imputation methods.

In general, this method proved to be quite unpredictable and data-dependent, as there was a significant difference in quality between models trained on different groups of datasets. In addition, specifically for the case of maximising recall, the method showed unsatisfactory results, although it was able to create powerful models for the Wine task with the target accuracy metric.

Multiple imputation in the implementation of the R MICE library proved to be the best, providing the most consistently high results for all metrics, which were closest to the performance after training on the complete datasets. The method performed well on all datasets regardless of the type, combination and number of gaps. For the Churn dataset, it worked best when combined with the RF and LGBM algorithms, with the RF algorithm even performing better in some cases using this imputation method than after training on the full dataset (MAR dataset). On the Wine models, the method also showed excellent results, delivering high scores on both the target and recall metrics. The method worked best with RF and SVC models. In general, the method had the highest level of reliability and predictability of results, and the models trained on the datasets with this imputation had consistently high prediction quality with the least fluctuations. Overall, this particular implementation of the multiple imputation method proved to be the most successful choice among studied methods for solving the problem of processing missing data.

CONCLUSIONS

The widespread problem of missing data is becoming especially relevant today due to the rapid development of artificial intelligence and machine learning technologies, which create a growing need for large amounts of high-quality data, as most algorithms require complete datasets. A large number of different methods for processing missing data have been created to solve the problem of missing data, while preserving the statistical parameters of the data for the success of further modelling. An important issue is the compatibility of imputation methods and predictive models, as different methods have different levels of quality and predictability of modelling results.

In this paper, an impact of the selected imputation methods on the quality of forecasting models is analysed. The best results were obtained using the multiple imputation method in the implementation from the R MICE library. Training on data using this method most reliably produced results that had high scores on quality metrics and were characterised by smaller quality fluctuations compared to other methods. The Python implementation of the multiple imputation method was less reliable, as its effectiveness strongly depended on the target metric and the specifics of the available data.

It has also been observed that statistically unreliable imputation methods, such as mean imputation or listwise deletion, do not necessarily lead to poor prediction results, as quite often predictive models are able to learn to recognise the target class even in the case of biased parameters. Therefore, their use, although riskier and more dependent on the characteristics of the available data, can also produce satisfactory results, which may not be inferior in quality to training using more complex methods.

REFERENCES

1. Donald B. Rubin, "Inference and Missing Data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.
2. Craig K. Enders, *Applied Missing Data Analysis*; 1 ed. The Guilford Press, 2010, 377 p.
3. Therese D. Pigott, "A review of methods for missing data," *Educational Research and Evaluation*, vol. 7, no. 4, pp. 353–383, 2001.
4. Luke Oluwaseye Joel, Wesley Doorsamy, and Babu Sena Paul, "A Review of Missing Data Handling Techniques for Machine Learning," *International Journal of Innovative Technology and Interdisciplinary Sciences (IJITIS)*, vol. 5, no. 3, pp. 971–1005, 2022. doi: https://doi.org/10.15157/IJITIS.2022.5.3.971-1005
5. Helen Bridge, Thomas Schindler, "The perils of the unknown: Missing data in clinical studies," *Medical Writing*, 27(1), pp. 56–59, 2018.
6. Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, 8(1), article no. 140, 2021. doi: 10.1186/s40537-021-00516-9
7. Hyun Kang, "The prevention and handling of the missing data," *Korean Journal of Anesthesiology*, 64(5), pp. 402–406, 2013. doi: 10.4097/kjae.2013.64.5.402
8. Tolou Shadbahr et al., "The impact of imputation quality on machine learning classifiers for datasets with missing values," *Communications medicine*, vol. 3, article no. 139, 2023. doi: 10.1038/s43856-023-00356-z
9. Jale Bektas, Turgay Ibrikci, and Ismail Ozcan, "The impact of imputation procedures with machine learning methods on the performance of classifiers: An application to coronary artery disease data including missing values," *Biomedical Research*, 29(13), pp. 2780–2785, 2018. doi: 10.4066/biomedicalresearch.29-18-199
10. George Paterakis, Stefanos Fafalios, Paulos Charonyktakis, Vassilis Christophides, and Ioannis Tsamardinos, "Do we really need imputation in AutoML predictive modeling?" *ACM Transactions on Knowledge Discovery from Data*, 18(6), 2024. doi: 10.1145/3643643
11. Katarzyna Woźnica, Przemyslaw Biecek, *Does imputation matter? Benchmark for predictive models*, 2020. doi: 10.48550/arXiv.2007.02837
12. A. Popov, O. Makarenko, and P. Bidyuk, "Rozv'iazannia zadachi zapovnennia propuskiv danykh alternatyvnymy metodamy pry stvorenni prohnoznykh modelei [Solving missing data imputation problem using alternative methods in predictive model creation]," *Proceedings of the II All-Ukrainian Scientific and Practical Conference "System Sciences and Informatics", December 4–8, 2023, Kyiv: National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"*, pp. 201–206.

Received 10.04.2024

INFORMATION ON THE ARTICLE

Andrii Yu. Popov, ORCID: 0009-0001-4783-7401, Educational and Research Institute for Applied System Analysis of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: popovandrii1403@gmail.com

ПОРІВНЯННЯ ЕФЕКТИВНОСТІ МЕТОДІВ ЗАПОВНЕННЯ ПРОПУЩЕНИХ ДАНИХ ПІД ЧАС РОЗРОБЛЕННЯ МОДЕЛЕЙ ПРОГНОЗУВАННЯ / А.Ю. Попов

Анотація. Наявність пропущених даних є поширеною проблемою в аналізі даних та машинному навчанні. У роботі проаналізовано залежності якості прогнозування моделей машинного навчання від використаних методів оброблення пропущених даних на етапі підготовки даних до навчання моделей. Досліджуваними методами є аналіз повних спостережень, заповнення середнім та дві реалізації методу множинного заповнення — мовами Python та R. Обраними класифікаторами є логістична регресія, метод випадкового лісу, метод опорних векторів та Light Gradient Boosting Machine (LGBM). Якість прогнозних моделей оцінюється за метриками accuracy, precision та recall. Розглянуто два набори даних із задачами класифікації, що мають різні цільові метрики. Найкращі результати досягнуто з використанням алгоритму множинного заповнення у реалізації мовою R у поєднанні з класифікаторами випадкового лісу та LGBM.

Ключові слова: пропущені дані, методи заповнення, прогнозні моделі, машинне навчання.