Journal of Theoretical and Applied Information Technology
15th December 2021. Vol.99. No 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TRUST SCORE MEASUREMENT METHOD FOR WEB DONOR SELECTION AND IMPUTATION OF MISSING VALUES

M. IZHAM JAYA[1,2], FATIMAH SIDI[1*], UMAR ALI BUKAR[3,4], ISKANDAR ISHAK[1], HAMIDAH IBRAHIM[1], LILLY SURIANI AFFENDEY[1], MARZANAH A. JABAR[3], NAVIN KUMAR DEVARAJ[5], AND MUSTAFA ALABADLA[1]

[1]Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor D.E., Malaysia.

[2]Department of Software Engineering, Faculty of Computing, Universiti Malaysia Pahang (UMP), 26600 Pekan, Pahang, Malaysia.

[3]Department of Software Engineering and Information System, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor D.E., Malaysia.

[4]Computer Science Unit, Department of Mathematical Science, Taraba State University Jalingo.

[5]Department of Family Medicine, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia (UPM), Serdang 43400, Malaysia.

Email: fatimah@upm.edu.my

ABSTRACT

The effects of trust score measurement is web donor selection is evaluated in this study. The performance of the proposed method is conducted by running a prediction model on the imputed dataset. Thus, several experiments were carried out to quantify the impact of the prediction model via Root Mean Squared Error (RMSE) and F-Measure. The results demonstrate that the proposed method improves the performance of existing web donor selection. The results showed that the RMSE, prediction accuracy, and F-Measure are improved when the prediction model is trained with datasets imputed using the proposed method. This research contributed to improved data quality, especially to the information system (IS) and database field, where good data quality benefited the data analysis performance.

Keywords: *Cold Deck, Missing Value, Imputation Method, Web Donors, Data Quality.*

1. INTRODUCTION

Missing values are a regular occurrence in datasets from any field of study. According to Liu et al. [1], a missing value is defined as the lack of data values in a dataset where the data records contain unwanted null values. Inability to manage missing values in a dataset can impact the performance of the analysis. For example, multiple researches [2; 3; 4] have demonstrated that the presence of missing values in a dataset might result in skewed findings in the prediction model, affecting its predictive accuracy. Similarly, classification techniques such as neural networks have the same issue. According to [5; 6], bias introduced by missing values in the training dataset might degrade the quality of the learnt pattern and thus reduce classification performance. Missing values are also connected with data quality and are quantified by the completeness dimension of the data. Data quality is defined as the state of having no faults and being 'fit for use' [7; 8; 9; 10; 11]. The presence of missing values in a dataset indicates that it is no longer defect-free, which may result in serious consequences for the entity that owns the data [12; 9]. For instance, the firm must devote more resources to rectifying missing values in the client address, as incorrect product delivery addresses might have a negative impact on the business. This case demonstrated the organization's operational cost increase as a result of poor data quality, specifically missing numbers.

Additionally, poor data quality within the company has a detrimental effect on user

perception, experience, trust, and belief in the specific application's use, such as Enterprise Resource Planning (ERP) and Business Intelligent System (BIS) [13; 14]. ERP and BIS systems are critical to the organization's success since they streamline processes and aid in decision-making. The references [15] and [16] examined the hurdles that are generated between specialised application usage and user approval as an organization's data quality degrades. Data completeness is defined as the ratio of existing data values to the total amount of data values [1; 17; 18]. When all necessary values for the data exist and there are no undesired null values [19; 20; 21; 22]. Earlier research on data completeness provided a variety of strategies for resolving the problem of missing values. These techniques are classified into two broad categories: case deletion and imputation. Imputation techniques are classified into two broad categories: multiple imputation and single imputation. Single imputation methods are further categorised into three broad categories: model-based, machine-learning-based, and data-driven. Cold deck imputation is a data-driven technique that achieves about the same imputation accuracy as multiple imputations at a reduced computing cost [23]. Cold deck imputation approaches, in contrast to multiple imputation methods, do not require several imputation processes, which can be computationally expensive. Additionally, cold deck imputation is less likely to result in model misspecification than model-based imputation. The only disadvantage of cold deck imputation is that the probability of identifying the best appropriate value to replace the missing value is low due to the small number of prospective donors. Increase the number of potential donors by collecting web donors from web data sources. [24] proposes cold deck imputation using prospective web donors from the web data source.

The proposed imputation approach by [24] was compared to three existing imputation methods for missing values: mean imputation, deletion, and K-Nearest Neighbor imputation (KNN). The results demonstrated that leveraging web donors to substitute missing variables improved the prediction model's accuracy more than existing imputation methods. Missing values were imputed during the evaluation process using the proposed imputation method, and the resulting dataset was then utilized to develop a prediction model. To evaluate the performance of each imputation approach, the prediction accuracy, root mean squared error (RMSE), and F-Measure are compared. While the proposed method in [24] achieved the highest prediction model accuracy, the proposed method is limited to a single web data source for one-time imputation. There is no guarantee that the data value provided by the web data source is correct, as data values from several web data sources sometimes contradict, even when they pertain to the same data item [25]. Thus, various web data sources should be permitted for cold deck imputation in order to obtain the most appropriate web donor. Additionally, when many web data sources are employed, an issue occurs. The strategy suggested by [24] does not include a way for measuring and determining the most suited web donor.

Additionally, the process for selecting the best acceptable web donor should identify the level of trust associated with each accessible web donor using numerous web data sources and rank the web donors based on their trust score. After that, the web donor with the highest trust score can be utilized to impute the dataset's missing values. Additionally, the trust score lets users to assess the correctness and dependability of each web donor prior to imputation. This is critical to establishing users' credibility and increasing their trust in the imputed dataset. The aforementioned constraints motivated us to conduct this study. Thus, the primary objective of this study is to boost data completeness by increasing the number of trusted web donors utilized to replace missing values in the dataset in comparison to the currently used cold deck imputation approach. As a result, this study intends to answer the research question “what is the effect of trust score measurement in web donor selection in comparison to other imputation methods.” Thus, the aim of this study is to investigate the effect of trust score measurement in web donor selection.

2. RELATED WORK

The data completeness issue arises for a variety of reasons, including human error, equipment malfunction, manual data input processes, faulty measurement, and inaccurate learning models [26; 27; 28]. Numerous approaches for imputation of missing data have been proposed in past research, and these imputation methods can be classified according to their complexity and performance. While

simpler and easier to apply, methods such as case deletion and mean imputation perform poorly in terms of bias and imputation accuracy [29]. Complex imputation approaches, on the other hand, such as multiple imputation and machine learning-based methods, improve imputation accuracy and reduce bias, but demand a significant amount of computer resources due to the repeated imputation and iteration necessary throughout the imputation [30]. The same issues occurred with model-based imputation, which need appropriate model selection in order to impute accurately [31]. A more promising approach of imputation is hot deck imputation, which achieves the same prediction accuracy as multiple imputation but at a lower computing cost [23]. However, because the donor is from the same dataset, the possibility of finding a better acceptable donor to replace the missing values is limited, much more so when the dataset is tiny. A more suitable donor can be increased by identifying potential donors to replace missing values from other data sources, most notably web data sources. However, the effectiveness of this technique is contingent upon the user's belief in the web donor's worth and the web data sources themselves [32]. Replacing missing values with untrusted data increases the risk of making incorrect decisions and performing incorrect analyses, ultimately destroying the organisational activity. Web data sources include a wealth of information that can be utilised to fill in blank values. For instance, Yahoo!Financial and Google Finance each maintained a sizable collection of financial data in order to fill in gaps in financial databases. Each web data source, on the other hand, had a unique data structure. The inability to use these data in missing value imputation is hampered by issues such as conceptual inaccuracy and terminological ambiguity.

Additionally, Reference [24] used an ontology mapping technique to overcome conceptual inaccuracies and terminological ambiguity issues in web data sources, allowing for the identification of web donors and missing values during imputation. However, the approach is confined to a single web donor value for each missing value replacement and ignores web donor value fluctuation, which is particularly important when more than one value is available to replace the missing value. As a result, the likelihood of locating the best appropriate value to replace the missing value is reduced. There are numerous sources of online donors on the web, all of which are unknown in terms of accuracy and dependability; hence, web donor values cannot be relied upon entirely. As a result, substituting web donor values for missing values may result in erroneous imputation [32]. Although they relate to the same data item, web donors from multiple web data sources may have different data values. Notably, the technique fell short of responding to critical issues such as "How much can I trust the imputed data?" & "Which data is more trustworthy, from which data source?" It is well established that data from web data sources frequently dispute with one another [25]. Thus, it is critical to address the above problems in order to boost the credibility of the analysis produced from the imputed dataset. A reliable imputed dataset is heavily dependent on the data used to replace missing values being chosen with care. According to [33] and [18], trustworthy data can only be derived from a trusted data source. For instance, if data from 'Source A' is more trustworthy than data from 'Source B,' replacing missing values with values from 'Source A' makes the imputed dataset more trustworthy than replacing missing values with data from 'Source B.' Because data selection is critical, ranking trust scores between potential online donors from numerous web data sources will aid users in determining the most trustworthy data. Thus, prior to the imputation procedure, reputable web donors can be chosen. Prior to the imputation procedure, it was necessary to assess the trustworthiness of each potential web donor. Accuracy and dependability are the needed criteria for assessing the typical qualities of trust [25; 34; 35; 36; 37; 18; 38]. The accuracy metric indicates the correspondence between the web donor's value and the reference value in the dataset. On the other side, dependability is a metric that indicates the amount to which the values claimed in an online donor's data source are accurate and trustworthy. As the precise value of such missing data is unknown, a metric for assessing correctness and reliability based on the observable data in the dataset is required [39; 25; 36; 34; 40]. A web donor provided by a web data source with the greatest accuracy and reliability score is assigned the highest trust score and is seen to be more trustworthy in replacing the missing value.

2.1 Proposed Methods

2.1.1 Trust Score Measurement For Conflicted Web Donors

The trust score is calculated using two methods; the reliability score and the accuracy score. A reliability score is necessary to quantify the discrepancy between all true values contained within the dataset and the claimed values contained within the web data source. In this scenario, the highest reliability score should go to an online data source with the smallest variance of difference. On the other hand, the accuracy score evaluates the correspondence between stated values in a web donor's data source and true values in a dataset. In our study, we derive truth values from variables with non-missing values that are related to the variable with missing values. These are the values that were obtained from the basic dataset. Reliability and accuracy scores are required to determine an online donor's trustworthiness, as correct claimed values from a web data source do not necessarily imply that the web data source is dependable, or vice versa. [34]. Additionally, reliability and accuracy are identified as critical expected traits affecting trustworthiness. The dependability and accuracy scores are weighted equally in the trust score calculation since they are equally important in determining the trust score [34; 41; 18].

Given that $\text{Accuracy}_{wd(i)}$ equals the sum of the similarity distance score and the average accurate claimed score for web donors from the web data source, $\text{Reliability}_{wd(i)}$ equals the reliability score for web donors from the web data source, Max Score equals the sum of the maximum reliability score, similarity distance score, and the average accurate claimed score.'

$$Trust\ Score = 100 \times \frac{\text{Accuracy}_{wd(i)} + \text{Reliability}_{wd(i)}}{Max\ Score} \quad (1)$$

Using Equation (1), a trust score is assigned to each web donor. A pseudocode of the trust score measurement is illustrated in the algorithm 1.

Algorithm 1: Trust Score Measurement

Input: a dataset D
Output: trust score

```
Begin
Let m be the variable with missing value
For each m in D do
Get the possible donor from the ontology
If m (has possible donor) then check for conflicts
    If (conflicted donor = yes) then
    Identify the set of variables with non-missing values which are related to m in D
    Get the truth values from the ontology
    Measure accuracy score
    Measure reliability score

    Calculate trust score: equation 1

    end if
end if
end for
End
```

3. EXPERIMENTAL DESIGN AND METHODOLOGY

The design of this research was guided by a broad theoretical framework for research [42], which incorporated the key features of research development [43] and by ensuring computations and trust dynamics issues are handled [41]. Three steps have been offered to illustrate the methodological techniques used. To meet the research's objectives, a quantitative research strategy based on an experimental research design was used to evaluate the offered approaches. During the research evaluation process, two experiments are undertaken. The imputed dataset is utilized to construct the prediction model for each experiment using ANN classification. Three performance measures are utilized to assess the prediction model's performance: root mean square error (RMSE), prediction accuracy, and F-Measure. The following sections outline the proposed solutions for achieving the research's objectives.

3.1 Implementation And Data Collection

As determined in Phase 2, the R programming language and Protege 5.2.0 are used on a PC equipped with an Intel Celeron 2.41GHz processor and 8GB RAM running the Windows operating system. The financial dataset of SP 500 firms from Standard Poor's Compustat North America was used in this study, which is available through the Compustat database. Compustat's database contained clean, consistent financial data on 56,000 businesses worldwide. However, due to the sheer volume of data, missing values in the dataset are unavoidable.

Previous study has documented the occurrence of missing values in the Compustat database (24; 44; 45). To cleanse and pre-process the data samples, Microsoft Excel and WEKA 3.8.2 are utilised as the data editor programmes. As a result of the data cleansing, 1,177 instances were retained in the dataset since they did not have zero values in inventories. From the dataset, fifteen financial variables were chosen to generate fourteen financial ratios (24). The Synthetic Minority Oversampling Technique (SMOTE) [46] is used in this study to address the issue of imbalanced classes and to ensure that the F-Measure for each class is determined during the assessment phase. SMOTE uses K-nearest neighbour analysis to build new instances from the under-represented NOCHG and UP classes. As proposed by [46], K=5 is used in this study to generate a synthetic sample. After applying the SMOTE technique, 853 synthetic cases with no missing values are added to the dataset with a balanced class. Table 1 shows the updated distribution for the Relative Change in Stock Earnings (RCSE) class. The datasets are stored in comma-separated values (CSV) format in a specific folder, and the collection is divided into two sets, as shown in Table 2.

Table 1: RCSE Class Distribution In Dataset After SMOTE Application

RCSE Class	Number of Instances	Percentage
UP	653	33%
NOCHG	653	33%
DOWN	667	34%

Table 2: Dataset Features

Set	Number of Datasets	Usage	Number of Instances each Dataset	Missing Values
1	22	Training dataset and validation dataset	1973	457
2	1	Testing dataset	197	0

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(e_{ori} - e_{est})^2} \quad (2)$$

where RMSE is calculated based on Equation 2, where e_{ori} is the observed values, e_{est} is the predicted value by the model and M is the total number of predictions.

$$\text{Prediction Accuracy } (\%) = \frac{tp + tn}{tp + tn + fp + fn} \times 100 \quad (3)$$

The prediction accuracy is calculated based on Equation (3), where the number of instances in the test dataset is given by the total number of true positive (TP), false positive (FP), and false-negative (FN). FP is the number of instances that are predicted positive but are negative, and FN is the number of instances that are predicted negative but are positive.

$$F - Measure = \frac{2 \times \text{Precision} \times \text{Recall}}{Precision + Recall} \quad (4)$$

as shown in Equation (4), the F-Measure is high when both recall and precision are high. Recall is calculated as (TP/(TP+FN)) and precision is calculated as (TP/(TP+FP)).

An ontology-based framework for financial decision-making (OFFDM) [24] performance is determined by selecting a web data source that is primarily used to replace missing values. Furthermore, the proposed OFFDM lacked the selection method to determine which web donor should replace missing values in the dataset when more than one web donor is present for each missing value replacement. Thus, the OFFDM performances are limited and depending on the first web data source visited to collect web donors, with no consideration given towards the suitability of other web donors from another web data source to replace the missing value. The suitability of web donors to replace the missing value in the dataset is unknown before the replacement, which could reduce the OFFDM performance if the web data source provides unnecessary web donors.

In this study, a trust score measurement method is introduced to treat the problem mentioned earlier and integrated into the new trust-based cold deck imputation method. The new trust-based cold deck imputation method

takes advantage of the variation of multiple web donor values from web data sources to improve the performance of OFFDM.

4. RESULTS AND DISCUSSION

A trust score is measured for each web donor from the available web data sources. Web donor with the highest trust score is then selected to replace the missing value in the dataset. In order to fairly evaluate the new trust-based cold deck imputation method against OFFDM, the web data sources, namely: Rocket Financial and Stockrow, were used to provide web donors. There are 52 remaining missing values in the dataset after web donor replacement. The remaining missing values after web donor replacement were imputed using KNN, MissForest, and PMM. As a result, three different datasets were obtained and named Trust_KNN, Trust_MissForest, and Trust_PMM accordingly. These datasets were used to train the prediction model, and the performances in terms of RMSE, prediction accuracy, and F-Measure were analyzed to validate the proposed method. The usages of these performance metrics are due to their wider application in the literature to validate web donor method (24; 44; 45). All datasets have 1973 instances and 457 imputed missing values. The proportion of training, validation, and testing are 70%, 20%, and 10%, respectively.

Table 3 presents the result of RMSE, prediction accuracy, and F-measures for dataset Trust_KNN, Trust_MissForest, and Trust_PMM.

Table 3: RMSE, Prediction Accuracy And F-Measure Between Imputation Methods

Dataset	RMSE	Prediction Accuracy	UP	NOCHG	DOWN
Trust_KNN	0.448	46.1	0.33	0.233	0.624
Trust_MissForest	0.4482	47.7	0.273	0.264	0.651
Trust_PMM	0.4512	47.2	0.456	0.282	0.578

On average, the RMSE for datasets Trust_KNN, Trust_MissForest, and Trust_PMM is 0.4491. The lowest RMSE is 0.4480 achieved when Trust_KNN is used to train the prediction model. The adoption of MissForest and PMM to substitute KNN for the remaining missing values imputation after web donor replacement does not help to improve the RMSE significantly. As shown in Figure 1, the adoption of MissForest and PMM for Trust_MissForest and Trust_PMM datasets has reduced the RMSE performance by 0.0002 and 0.0032, respectively. However, the reduction in RMSE performance is relatively low in both datasets.

The highest prediction accuracy is 47.7% achieved when Trust_MissForest is used to train the pre- diction model. On the other hand, a prediction model trained with Trust KNN has resulted in the lowest prediction accuracy, 46.1%. In general, the adoption of MissForest and PMM to replace KNN for the remaining missing values imputation after web donor replacement has increased the prediction accuracy by 1.6% and 1.1%, respectively, as shown in Figure 2. In order to determine whether there is prediction accuracy improvement for Trust_MissForest and Trust_PMM is beneficial or not, the corresponding RMSE are examined. Prediction accuracy improvement in both datasets is achieved with a relatively low reduction in RMSE performance, 0.0002 and 0.0032, respectively. Thus, MissForest and PMM for the remaining missing values imputation after web donor replacement improve prediction accuracy without a considerable reduction in RMSE performance.

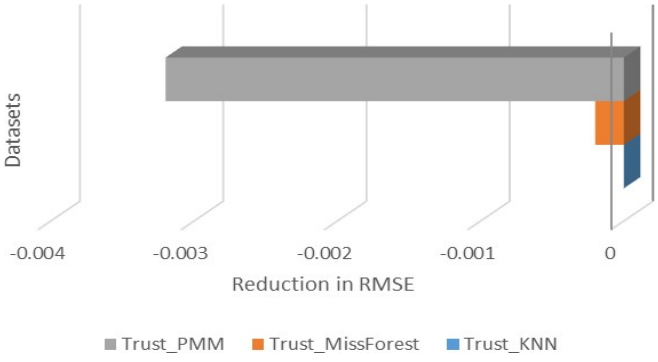

Figure 1: Reduction Of RMSE In Datasets With Respect To Trust KNN

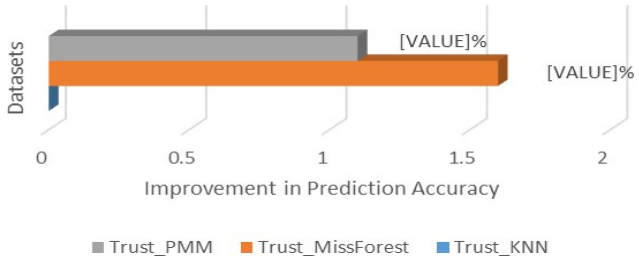

Figure 2: Prediction Accuracy Improvement In Datasets With Respect To Trust KNN

The F-Measure for NOCHG class is consistently lower than UP class, and DOWN class in each dataset used to train the prediction model as shown in Figure 3. The adoption of MissForest and PMM to replace KNN for the remaining missing values imputation has improved the F-Measure for NOCHG class by 0.031 and 0.049, respectively. However, when Trust_MissForest is used to train the prediction model, the accuracy and sensitivity of the prediction model towards the UP class are reduced by 0.057 and closer to the F-Measure of NOCHG class compared to the prediction model trained with Trust_KNN dataset. On the other hand, the accuracy and sensitivity of the prediction model towards the UP class are improved by 0.126 when Trust_PMM is used to train the prediction model compared to the prediction model trained with Trust_KNN dataset.

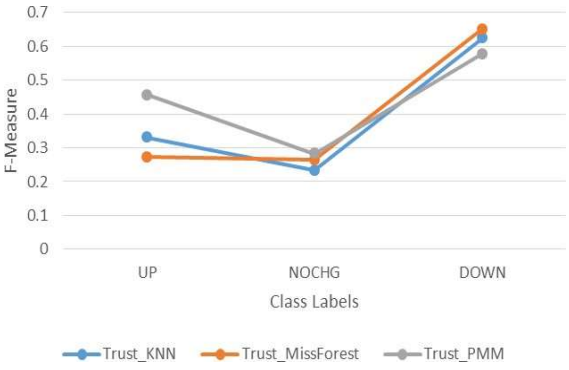

Figure 3: Comparisons Of F-Measure Between Imputed Datasets

The F-Measure in DOWN class for prediction model trained with Trust_MissForest improved by 0.027 compared to Trust_KNN. Yet, the F-Measure for the DOWN class is reduced by 0.046 when Trust_PMM is used to train the prediction model compared to the prediction model trained with Trust_KNN. This situation is expected as the reduction occurred due to the considerable improvement in F-Measure for UP class when Trust_PMM is used to train the prediction model.

From the RMSE findings, prediction accuracy, and F-Measure, it is confirmed that the adoption of PMM to replace KNN in the Trust_PMM dataset for the remaining missing values imputation has improved the performance of the trust-based cold deck imputation method. The prediction accuracy is increased by 1.1% in Trust_PMM with a 0.0032 reduction in RMSE compared to Trust_KNN. The F-Measure for prediction model trained with Trust_PMM showed the stability in terms of prediction model accuracy and sensitivity towards UP class, NOCHG class, and DOWN class compared to the F-Measure obtained from the prediction model trained with Trust_KNN and Trust_MissForest. On the other hand, the adoption of MissForest to replace KNN for the remaining missing values imputation in the Trust_MissForest dataset has improved the prediction accuracy by 1.6% with a relatively low, 0.0002 RMSE reduction. However, the reduction of F-Measure for the UP class when MissForest is adopted has reduced the accuracy and sensitivity of the prediction model towards the UP class. The F-Measure for Trust_MissForest dataset also shows that the prediction model is more biased towards the DOWN class in its prediction.

This shows that the performance is improved when the dataset imputed using a trusted web donor (Trust_PMM) is used to train the prediction model compared to the prediction model trained with dataset imputed using OFFDM (Rocket_first_PMM). Results presented in Figure 4 exhibit that the performance in terms of RMSE is better when trust score is used for web donor selection in Trust_PMM. Instead, the RMSE is worst by 0.0031 when the prediction model is trained with the Rocket_first_PMM dataset.

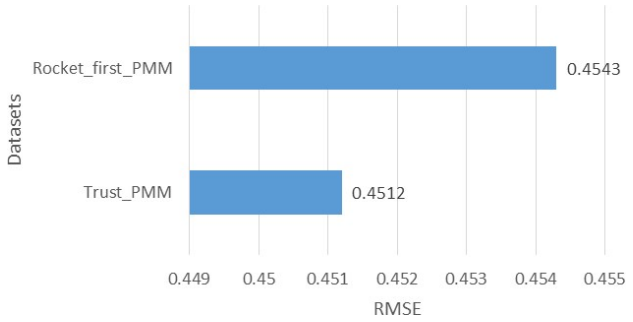

Figure 4: Comparisons Of RMSE Between Datasets Imputed Using A Trusted Web Donor From Web Data Sources And OFFDM

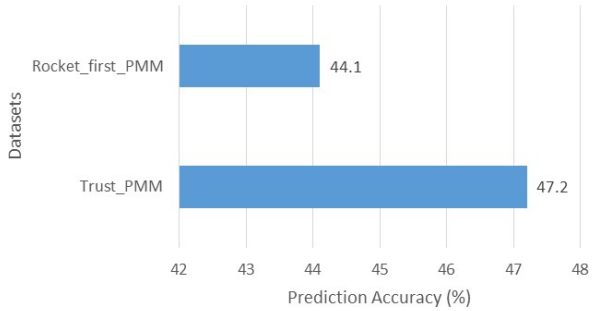

Figure 5: Comparisons of prediction accuracy between datasets imputed using a trusted web donor from web data sources and OFFDM

The same result can be observed in prediction accuracy where the prediction model trained with Trust_PMM dataset is 3.1% higher than Rocket_first_PMM as shown in Figure 5.

The best F-Measure for NOCHG class is 0.282, which is achieved when the prediction model is trained with Trust PMM as listed in Table 4. The F-Measure for Trust_PMM in UP class, NOCHG class, and DOWN class is consistently better than Rocket_first_PMM where the difference in F-Measure for each class is 0.047, 0.027, and 0.012, respectively.

Table 4: F-Measure Between Imputation Methods

Experiment	Datasets	UP	NOCHG	DOWN
2	Trust_PMM	0.456	0.282	0.278
1	Rocket_first_PMM	0.409	0.255	0.566
Performance Improvement		0.047	0.027	0.012

Similarly, the performance in terms of RMSE, prediction accuracy, and F-Measure is improved when a trusted web donor impairs the dataset's missing values compared to OFFDM. Therefore, the trust-based cold deck imputation can be considered a better cold deck imputation method when web donor from more than one web data source is available and can analyze the reliability and accuracy of each available web donor. The improvement of RMSE and prediction accuracy for Trust_PMM and Rocket_first_PMM are summarized in Table 5.

The improvement is significant as the increment in prediction accuracy is achieved without reducing the RMSE and F-Measure performance. Furthermore, OFFDM performance is determined by the selection of web data source that is primarily used to provide the web donor, and the performance of the selected web data source is unknown before the experiment.

Table 5: Performance Improvement In RMSE And Prediction Accuracy For Trust_PMM And Rocket_First_PMM

Performance Metrics	Dataset		Performance Improvement
	Trust_PMM	Rocket_first_PMM	
RMSE	0.4512	0.4543	0.0031
Prediction Accuracy	47.20%	44.10%	3.10%

The results of the RMSE, prediction accuracy, and F-Measure for the prediction model trained with Trust PMM dataset are also compared to the result of the same prediction models trained with IGN, AVG, KNN, MissForest, and PMM as shown in Figure 6, Figure 7, and Table 6. The RMSE of prediction models trained with the Trust_PMM dataset is the lowest compared to the other datasets imputed with AVG, MissForest, and PMM, as shown in Figure 6. Compared to the prediction models trained using datasets imputed with AVG, MissForest, and PMM, the RMSE for prediction model trained with Trust_PMM is better by 0.0072, 0.0024, and 0.0067, respectively. KNN and IGN are worst by 0.0093 and 0.0333 respectively compared to Trust_PMM.

In terms of prediction accuracy, the Trust_PMM dataset performed better than other datasets except for KNN, as shown in Figure 7. Trust_PMM is better by 1.1% prediction accuracy compared to the dataset imputed using MissForest. On the other hand, Trust_PMM and PMM achieved the same pre- diction accuracy percentage. The highest prediction accuracy is achieved when the dataset is imputed with KNN. There is a 1% difference in prediction accuracy between datasets imputed with KNN and Trust_PMM. However, such difference comes at the cost of 0.0093 RMSE reduction in dataset imputed with KNN compared to Trust_PMM. Compared to IGN, which has the lowest prediction accuracy, the prediction model trained with Trust_PMM is better by 14.7%.

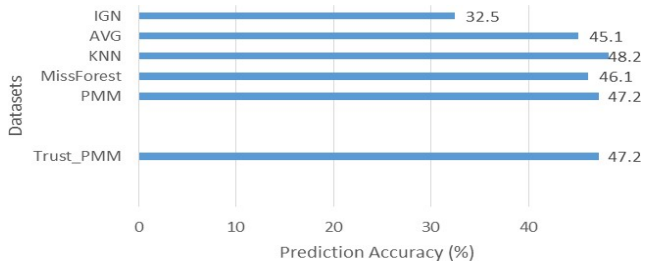

Figure 6: Comparisons Of RMSE Between Datasets Imputed Using A Trusted Web Donor From Web Data Sources And Other Imputation Methods

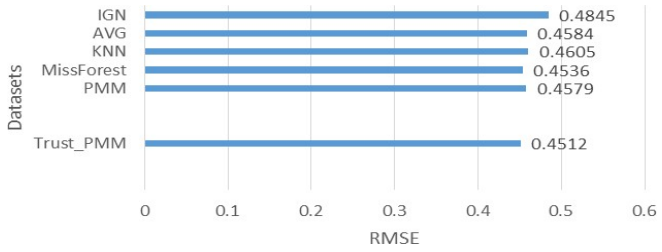

Figure 7: Comparisons Of Prediction Accuracy Between Datasets Imputed Using A Trusted Web Donor From Web Data Sources And Other Imputation Methods

As shown in Table 6, the F-Measure for NOCHG class is consistently lower than the F-Measure for the UP class and DOWN class in all datasets except IGN. The mechanism of IGN imputation influenced its F-Measure performance as 217 of 1973 instances were deleted from the dataset during the IGN process. Due to this, the F-Measure for NOCHG class in IGN dataset is increased while the F-Measure for UP class dropped. The result also shows that the highest result of F-Measure for NOCHG class is achieved in Trust_PMM and IGN. Among PMM, MissForest, KNN, and IGN, Trust_PMM also managed to achieve the highest F-Measure for the UP class. On the other hand, the F-Measure for the DOWN class in Trust PMM is 0.578.

Accordingly, the size of training datasets is reduced to highlight the performance of RMSE and prediction accuracy when the percentage of instances with missing values in the training dataset is increased.

5. CONCLUSION

This research focuses mainly on the imputation of missing values using cold deck imputation, where missing values were replaced with the most trusted web donor from web data sources. Despite the contributions made by [24] concerning cold deck imputation with web donor, the work has some disadvantages such as constraint to provide a proper selection method for web donor used to replace missing values, its unsuitability when dealing with conflicted web donor in case of more than one web data source is used, and its inability to describe the level of trust held by each web donor used to replace the missing values. Motivated by this fact, the following contributions are addressed in this research. A new method to measure the trust score for each web donor was introduced in this research. The trust score was measured based on two main components: the accuracy score and

Table 6: Comparisons of F-Measure between datasets imputed using a trusted web donor from web data sources and other imputation methods

Datasets	F-Measure		
	UP	NOCHG	DOWN
Trust PMM	0.456	0.282	0.578
PMM	0.446	0.095	0.581
MissForest	0.311	0.22	0.632
KNN	0.425	0.133	0.618
AVG	0.489	0.108	0.485
IGN	0.267	0.282	0.408

the reliability score. The accuracy score for each web donor is determined based on the average accurate claimed score and the similarity distance score between the claimed values in the web data source and the truth values in the dataset. On the other hand, the reliability score measured the difference between the claimed values in web data sources and the truth values in the user dataset. Trust score is calculated in percentage, and consequently, web donors can be ranked based on their trust score to determine their reputation. Web donor is determined as a trusted web donor if it held the highest trust score among the available web donors.

This research established a new trust-based cold deck imputation method with multiple web donors to improve data completeness in the dataset. The trust-based cold deck imputation method has success- fully utilized web donor values from multiple web data sources to perform the imputation and selected only a trusted web donor to replace the missing values. Thus, the performance of cold deck imputation is no longer restricted to which data source is primarily used to replace the missing values. Trust level in terms of accuracy and reliability is explained for each web donor before the imputation taking place and ranked accordingly. The proposed trust-based cold deck imputation method with multiple web donors achieved the highest performance in RMSE, prediction accuracy, and F-Measure compared to OFFDM during the experiments. Besides that, the trust-based cold deck imputation method also achieved better RMSE, prediction accuracy, and F-Measure than AVG and IGN imputation methods when used to impute the training datasets during the experiments. Compared to other established imputation methods in model-based and machine learning categories such as KNN, MissForest, and PMM, the proposed trust-based cold deck imputation method achieved better performance in terms of RMSE and F-Measure during the experiments. There is not much difference in terms of prediction accuracy between the trust-based cold deck imputation method, KNN, MissForest, and PMM.

Moreover, this research resolved conflicted web donor problems when more than one web data source provides web donors in a cold deck imputation. This also overcomes the issues by replacing missing values with a trusted web donor to improve data completeness. The main objective of this research has been accomplished. However, there are still existing issues that are not covered under the scope of this research and can be undertaken in the future. Firstly, the proposed trust score measurement method can be further evaluated using the dataset from other domains such as health and education. Secondly, the number of web data sources can be increased to provide more web donors. A further experiment can be conducted to analyze its effect on performance improvement. Lastly, the prediction model's performance can be improved by employing feature selection to select essential and relevant financial ratios for the prediction. On the other hand, feature selection helps to remove irrelevant and unneeded financial ratios, which did not contribute to the accuracy of the prediction model.

Funding Statement: This work was supported by the Malaysian Ministry of Higher Education under the Fundamental Research Grant Scheme (Grant No. FRGS/1/2020/ICT06/UPM/02/1), UMP Fundamental Research Grant (RDU-200317). and the Universiti Putra Malaysia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES:

[1] Y.-N. Liu, J.-Z. Li, and Z.-N. Zou, "Determining the real data completeness of a relational dataset," Journal of computer science and technology, vol. 31, no. 4, pp. 720–740, 2016.

[2] M. G. Rahman and M. Z. Islam, "Missing value imputation using a fuzzy clustering-based em approach," Knowledge and Information Systems, vol. 46, no. 2, pp. 389–422, 2016.

[3] J. D. Rubright, R. Nandakumar, and J. J. Gluttin, "A simulation study of missing data with multiple missing x's," Practical Assessment, Research, and Evaluation, vol. 19, no. 1, p. 10, 2014.

[4] P. L. Roth and F. S. Switzer III, "A monte carlo analysis of missing data techniques in a hrm setting," Journal of Management, vol. 21, no. 5, pp. 1003–1023, 1995.

[5] Z.-g. Liu, Q. Pan, J. Dezert, and A. Martin, "Adaptive imputation of missing values for incomplete pattern classification," Pattern Recognition, vol. 52, pp. 85–95, 2016.

[6] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing value estimation for mixed-attribute data sets," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 1, pp. 110–121, 2010.

[7] Y. W. Lee and D. M. Strong, "Knowing-why about data processes and data quality," Journal of Management Information Systems, vol. 20, no. 3, pp. 13–39, 2003.

[8] A. V. Levitin and T. C. Redman, "Data as a resource: Properties, implications, and prescriptions," MIT Sloan Management Review, vol. 40, no. 1, p. 89, 1998.

[9] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context," Communications of the ACM, vol. 40, no. 5, pp. 103–110, 1997.

[10] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," Journal of management information systems, vol. 12, no. 4, pp. 5–33, 1996.

[11] R. Y. Wang, "Total data quality management," Communications of the ACM, vol. 41, no. 2, pp. 58–65, 1998.

[12] A. Haug, F. Zachariassen, and D. Van Liempd, "The costs of poor data quality," Journal of Indus- trial Engineering and Management (JIEM), vol. 4, no. 2, pp. 168–193, 2011.

[13] M. I. Jaya, F. Sidi, I. Ishak, L. S. Affendey, and M. A. Jabar, "A review of data quality research in achieving high data quality within organization." Journal of Theoretical & Applied Information Technology, vol. 95, no. 12, pp. 2647–2657, 2017.

[14] M. I. Jaya, F. Sidi, L. S. Affendey, M. A. Jabar, and I. Ishak, "Systematic review of data quality research," Journal of Theoretical and Applied Information Technology, vol. 97, pp. 3043–3068, 11 2019.

[15] K. Hartl and O. Jacob, "The role of data quality in business intelligence-an empirical study in german medium-sized and large companies." in ICIQ, 2016, pp. 33–42.

[16] A. Popovicˇ, R. Hackney, P. S. Coelho, and J. Jaklicˇ, "Towards business intelligence systems suc- cess: Effects of maturity and culture on analytical decision making," Decision Support Systems, vol. 54, no. 1, pp. 729–739, 2012.

[17] A. Wechsler and A. Even, "Assessing accuracy degradation over time with a markov-chain model." in ICIQ, 2012, pp. 99–110.

[18] C. Batini and M. Scannapieca, "Data quality: Concepts, methodologies and techniques," 2006.

[19] V. Jayawardene, S. Sadiq, and M. Indulska, "An analysis of data quality dimensions," 2015.

[20] M. Bovee, R. P. Srivastava, and B. Mak, "A conceptual framework and belief-function approach to assessing overall information quality," International journal of intelligent systems, vol. 18, no. 1, pp. 51–74, 2003.

[21] B. K. Kahn, D. M. Strong, and R. Y. Wang, "Information quality benchmarks: product and service performance," Communications of the ACM, vol. 45, no. 4, pp. 184–192, 2002.

[22] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," Com- munications of the ACM, vol. 39, no. 11, pp. 86–95, 1996.

[23] U. Garciarena and R. Santana, "An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers," Expert Systems with Applications, vol. 89, pp. 52–65, 2017.

[24] J. Du and L. Zhou, "Improving financial data quality using ontologies," Decision Support Systems, vol. 54, no. 1, pp. 76–86, 2012.

[25] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang, "Knowledge-based trust: Estimating the trustworthiness of web sources," arXiv preprint arXiv:1502.03519, 2015.

[26] R. Deb and A. W.-C. Liew, "Missing value imputation for the analysis of incomplete traffic acci- dent data," Information sciences, vol. 339, pp. 274–289, 2016.

[27] C.-F. Tsai and F.-Y. Chang, "Combining instance selection for better missing value imputation," Journal of Systems and Software, vol. 122, pp. 63–71, 2016.

[28] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas, "Iterative robust semi-supervised missing data imputation,"

IEEE Access, vol. 8, pp. 90 555–90 569, 2020.

[29] B. E. Cox, K. McIntosh, R. D. Reason, and P. T. Terenzini, "Working with missing data in higher education research: A primer and real-world example," The Review of Higher Education, vol. 37, no. 3, pp. 377–402, 2014.

[30] M. Nakai and W. Ke, "Review of the methods for handling missing data in longitudinal data analysis," International Journal of Mathematical Analysis, vol. 5, no. 1, pp. 1–13, 2011.

[31] R. R. Andridge and R. J. Little, "A review of hot deck imputation for survey non-response," Inter- national statistical review, vol. 78, no. 1, pp. 40–64, 2010.

[32] H.-Z. Wang, Z.-X. Qi, R.-X. Shi, J.-Z. Li, and H. Gao, "Cosset+: Crowdsourced missing value imputation optimized by knowledge base," Journal of Computer Science and Technology, vol. 32, no. 5, pp. 845–857, 2017.

[33] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye, "Katara: A data cleaning system powered by knowledge bases and crowdsourcing," in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015, pp. 1247–1261.

[34] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han, "A confidence-aware approach for truth discovery on long-tail data," Proceedings of the VLDB Endowment, vol. 8, no. 4, pp. 425–436, 2014.

[35] B. Kitchens, C. A. Harle, and S. Li, "Quality of health-related online search results," Decision Support Systems, vol. 57, pp. 454–462, 2014.

[36] E. Asmare and J. A. McCann, "Lightweight sensing uncertainty metric—incorporating accuracy and trust," IEEE Sensors Journal, vol. 14, no. 12, pp. 4264–4272, 2014.

[37] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava, "Truth finding on the deep web: Is the problem solved?" arXiv preprint arXiv:1503.00303, 2015.

[38] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," ACM computing surveys (CSUR), vol. 41, no. 3, pp. 1–52, 2009.

[39] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," ACM Sigkdd Explorations Newsletter, vol. 17, no. 2, pp. 1–16, 2016.

[40] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," in Proceedings of the 2014 ACM SIGMOD international conference on Management of data, 2014, pp. 1187–1198.

[41] K. Govindan and P. Mohapatra, "Trust computations and trust dynamics in mobile adhoc networks: A survey," IEEE Communications Surveys & Tutorials, vol. 14, no. 2, pp. 279–298, 2011.

[42] T. Love, "Theoretical perspectives, design research and the phd thesis," Doctoral Education in Design: Foundations for the Future, pp. 237–246, 2000.

[43] P. D. Leedy and J. E. Ormrod, "Practical research: Planning and design, global edition," England: Pearson Education Limited, 2015.

[44] J. W. Kim and J.-H. Lim, "It investments disclosure, information quality, and factors influencing managers' choices," Information & management, vol. 48, no. 2-3, pp. 114–123, 2011.

[45] R.-C. Hwang, K. Cheng, and C.-F. Lee, "On multiple-class prediction of issuer credit ratings," Applied stochastic Models in business and industry, vol. 25, no. 5, pp. 535–550, 2009.

[46] N. Chawla, K. BOWYER, and L. HALL, "y kegelmeyer, wp (2002)."smote: Synthetic minority over-sampling technique"," Journal of artificial intelligence research, pp. 321–357.