IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 8 August 2023, accepted 12 October 2023, date of publication 19 October 2023, date of current version 25 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325892

RESEARCH ARTICLE

Exploring the Impact of Data Quality on Business Performance in CRM Systems for Home Appliance Business

YOUNGJUNG SUH
LG Electronics Inc., Yeongdeungpo-gu, Seoul 07336, South Korea
e-mail: youngjung.suh@gmail.com

ABSTRACT In customer relationship management (CRM), high-quality customer data is at the heart of reliable data analysis and is the foundation for data-driven decisions that impact business goals. To find performance indicators of data quality to maximize the effectiveness of CRM, we need to devise an approach to identifying and managing "business-relevant" information quality metrics. Therefore, this paper deals with the discovery and validation of the Data Quality Dimension (DQD) in terms of meaning and utilization value of data values other than the aspects such as syntax criteria or data format. We design the quality index and scoring logic of the customer integration profile and prove its usefulness by applying it to actual CRM data. A sample of real business operations data of approximately 1 million CRM customers was used to analyze the relevance between the DQDs and business performance indicators. As business performance indicators, we used both the company's purchasing loyalty index and the performance of past promotional campaigns. We analyzed the significant impact of each DQD on purchase loyalty and promotional campaign success rate. Next, we confirmed the effectiveness of DQDs in terms of providing analytic ease for predictive analysis such as target marketing in CRM. In addition, we showed some possibilities to consider improving data quality by analyzing the granularity of a specific attribute based on a certain DQD. Through these verification results, the validity of the DQDs of the customer profile was confirmed in the context of 'suitability for use' of customer data that affects business activities critical to the company in the CRM system.

INDEX TERMS Big data applications, data quality, customer relationship management, business performance, home appliance business.

I. INTRODUCTION

The purpose of customer relationship management (CRM) is to predict the customer's future behavior through the customer's past service use history information, and to use the result for customer management. It is a strategic approach by which companies can not only identify and attract new customers, but also strengthen and manage relationships with them with a view to sustainable company growth [1]. CRM data supports important marketing tasks such as customer segmentation, consumption forecasting, promotion management, and delivery of marketing materials [2]. It is the basis for popular marketing techniques such as customer segmentation, customer lifetime value (CLV) estimation, and RFM (Recency, Frequency, and Monetary) analysis to evaluate customer equity. Understanding customer equity based on this customer lifetime value can help optimize the balance of investments in customer acquisition and retention. In other words, CRM activities have a significant correlation with Business Performance Index and Customer Satisfaction Index [2], [3].

These CRMs have recently received more attention from academia, the public and the media. The effectiveness of CRM and the benefits gained from it depend on data resources such as customer profile, transaction history (e.g. purchases, usage, etc.), customer contact history and

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Xiao.

promotional activities. Recently, due to changes in the digital environment related to information storage, collection, and dissemination, companies manage and maintain large amounts of customer information with CRM. As a result, companies are increasingly faced with vast amounts of data contained in widely disparate and often inconsistent databases. While essential to making these businesses competitive, maintaining high-quality data sets that capture very large customer profiles and transactions is difficult and costly.

However, the company requires the sophisticated IT systems to identify the existence of incoming and stored data flaws and the extent of damage caused by these flaws in system management. It is because high-quality data is the key to interpretable and reliable customer data analysis and the basis for meaningful data-driven decision-making. However, CRM will not succeed if even the most sophisticated IT or business system relies on data of insufficient quality and is not structured for the purpose of application [1], [4]. Therefore, identifying ways to maintain the quality of customer data, which has a decisive impact on the marketing success of actual CRM, has become the most important business concern [1], [5], [6].

Most of the existing studies proposing requirements for data quality metrics have been conducted from a methodological point of view [7], [8], [9], [10], [11], [12], [13]. Recently, with the development of big data and artificial intelligence technology, it is possible to find the best CRM strategy by applying machine learning or data mining technology to customer data [14]. And the importance of customer data quality indicators is emphasized in terms of business relevance linked to company performance [11]. Therefore, in order to support the best CRM strategy, it is necessary to measure data quality issues from the perspective of contributing to the company's business performance while ensuring the ease of data analysis.

In this paper, we do not focus on generally low-level data quality issues such as data validity, standardized format, correctness, etc. In order to efficiently support machine learning-based analysis in CRM settings based on integrated customer profiles, we will discover data quality indicators, design scoring logic, and conduct sensitivity check to see if they are helpful in contributing to actual business performance improvement. Of course, previous studies provide a valuable contribution by specifying several possible requirements for data quality metrics and their values. However, studies are still insufficient on the quantification and actual application of customer profile quality indicators at the level of home appliance purchasing CRM.

First, we reviewed the various data quality dimensions of existing studies, such as accuracy, completeness, consistency, and timeliness, and the corresponding metrics developed for their quantitative evaluation. [9], [15], [16], [17], [18], [19], [20], [21], [22], [23], [25], [33]. Among the existing data quality dimension (DQD) metrics, we selectively introduced *completeness*, *timeliness*, and *plausibility*, which have relatively clear requirements and are applicable. In addition, we newly discovered *information* and *multifacetedness* metrics to enhance the interpretability of customer data attributes. Next, we conducted the sensitivity check for usefulness of the DQD metrics by applying them to the actual CRM data.

The contributions of this paper are the following three.

1) Discovery of customer data quality indicators for CRM and quality quantification through metric design: This paper deals with the semantics of data values for business performance indicators for the success of CRM, rather than DQ dimensions such as syntax standards or data formats.

2) Validation of DQD metrics in the CRM using large-scale data samples (about 1 million) of actual customer data, transaction data and contract data sets: To confirm the impact of DQDs on business performance index, we explored the mutual relevance through the purchase loyalty index, which is the business performance index of the companies that have customer purchases as their business model. We grouped customers based on the scoring results of each DQD, and identified statistically significant differences in purchase loyalty between the groups. In addition, the contribution of DQDs to the success rate of sales promotion was analyzed using customer data on the success or failure of past sales promotion campaigns. Next, in order to confirm the value of utilization in terms of ease of analysis, we performed machine learning-based RFM prediction modeling on the CRM dataset. After grouping customers based on the score of each of *information* and *plausibility*, we verified the statistically significant difference in the F-measure performance of the RFM prediction model between each group. Through this, it was confirmed that these two DQDs are useful as filtering indicators for constructing an appropriate amount of data set that guarantees the minimum performance to support ML modeling.

3) Analysis of the contribution of each profile attribute to the DQD metric score: Among the DQD metrics we defined, we analyzed *timeliness* and *information* in details. In addition, in terms of the *information* score, additional analysis was conducted on how much each attribute gives distinction to customer understanding or segmentation. The list of attributes with the detailed analysis results that can be considered for quality improvement will provide important insights for CRM data governance personnel to prioritize quality improvement efforts for data sets.

The rest of the paper is arranged as follows. In section II, we review the related work on customer data quality. In section III, we introduce the data quality dimensions and metrics. The section IV gives the experimental setting and analyzes the experimental results. Then, in section V, we discuss the additional analysis results and some insights from them. The final section concludes the paper and offers further research directions.

II. RELATED WORKS

Maintaining high-quality data is critical to customer retention, a core value of CRM, and it has become essential in many industries to prevent catastrophic losses. Therefore, related studies are currently being conducted on the discovery, application, and evaluation of big data quality metrics in various domains. It is very important not only to consider how universally the quality of customer data can be expressed and managed, but also to define and evaluate quality from the viewpoint of minimizing the loss of the company through maximizing the business utility of CRM [2]. In this section, we analyze existing studies proposing requirements for data quality metrics. Category A includes research literature on data quality metrics and requirements for them from a methodological point of view. Category B includes a study of the requirements for the general data quality assessment process (e.g. measurement frequency) required by data management organizations. Category C consists of requirements and practical recommendations for relevant data quality indicators in specific business processes.

A. CONTRIBUTION OF THE METHODOLOGICAL PERSPECTIVE TO THE REQUIREMENTS OF DATA QUALITY METRICS

Most of the existing studies proposing requirements for data quality metrics have been conducted from a methodological point of view, such as accuracy, completeness, consistency, and currency [7], [8], [9], [10], [11], [12], [13]. Most researchers have tried to improve data quality by using mathematical and programming solutions to improve the source of data, that is, lower-level data quality [7]. In addition, both researchers and practitioners have proposed requirements for data quality metrics in order to develop a more general and appropriate methodological basis from the fragmented indicators of existing studies to solve specific problems [8], [9], [10], [11], [12], [13], [35], [36].

FISHER et al. studied the error rate or accuracy percentage of the database to include random measurements. They expanded the accuracy metric including probability distribution values in order to improve the disorganization of errors and the lack of provision of probabilistic information about the distribution [15]. In [19], in order to increase the value of information at the time of decision-making, an extended metric that refers to a currency that provides an indication about the real-world information at the time of measurement, based on the stored information was developed. In addition, they proposed a quantitative approach for modelling the influence of currency on decision-making by extending the normative concept of the value of information in the field of CRM sales management and demonstrated its usefulness.

B. FRAMEWORK TO SUPPORT DATA QUALITY ASSESSMENT PROCESS (INCLUDING MONITORING AND VISUALIZATION)

Efforts toward high quality generally increase the value of data sets, but can entail high costs, such as detection and correction of defects and investment in quality monitoring tools [26]. There is a study to assist practitioners in selecting the right tool for a given use case by identifying and evaluating dedicated software tools that support these data quality measurement and monitoring functions [27]. In addition, studies on the requirements to support the efficiency of the quality control process (measurement frequency, procedure, method, etc.) to reduce these costs form an important axis [13], [23], [24]. There is the research on discussing framework implementation and data flow management across various quality management processes [23]. Using the Big Data Quality Profile concept, they proposed a BDQ management framework to enhance data control and preprocessing activities, and used the big data profiling and sampling components of the framework to support the estimation of faster and more efficient data quality before and after intermediate preprocessing steps.

In [24], the authors conducted a study on an interactive environment solution for data quality evaluation that facilitates user participation in data quality evaluation and provides reusable quality metrics with immediate visual feedback. It provides an overview visualization of these quality indicators along with error visualization that facilitates interactive exploration of the data to identify the causes of quality issues present in the data. Research on the framework supporting these data quality inspection, monitoring and control processes is an important axis of research in the sense that systematic data quality management and error predictability can be guaranteed. Some researchers focused on several metrics to measure data quality in streaming time series data in IoT applications [34]. They proposed a set of metrics for measuring data quality (DQ) in streaming time series, implemented and validated a set of techniques and tools to monitor and improve the quality of information. And the proposed techniques and tools were deployed in the Decision Support System (DSS), a data management, monitoring, and data analysis platform for decision-making on the quality of data obtained from streaming time series.

C. QUALITY METRICS AND PRACTICAL RECOMMENDATIONS TO SUPPORT BUSINESS RELEVANCE

Heinrich et.al proposed requirements to support both decision-making under uncertainty and economy-oriented data quality management for economic data quality management [28]. Based on the proposed requirements, they evaluated ‘‘fulfillment’’ and ‘‘not-fullfillment’’ for representative quality index metrics of existing studies. From a business perspective, data quality is also addressed from the perspective of understanding the type of information needed to make marketing decisions and how this information is used [29]. Recently, with the development of big data and artificial intelligence technology, it is possible to find the best CRM strategy by applying machine learning or data mining technology to customer data [14]. It uses machine learning technology to find hidden useful information in customer data and to predict trends and behaviors based on it. By using this predicted information to effectively carry out target marketing, the Business Performance Index is increased. As such, there is a research stream that discusses requirements and practical recommendations for specific performance indicators of data quality indicators (e.g. within business processes) [11], [13]. Mosley et al. introduced the requirements for data quality metrics from a practitioner's point of view, and discussed their importance in terms of business relevance [11]. The authors stated that data quality indicators should be linked to a company's performance and should be understood in the context of factors influencing identified significant business activities. Additionally, the authors called for acceptability, which means that metrics are assigned a threshold of data quality levels that meet business expectations.

The common research motivation of these studies is that poor data quality affects the achievement of business goals. Data analysts need to find and use data quality performance indicators based on the relationship between flawed data and missed business goals. Finding these indicators requires devising an approach to identifying and managing "business-relevant" information quality metrics. Therefore, in this study, we designed the DQDs and scoring logic of the customer integration profile in terms of business performance improvement. Then, we applied the discovered DQD metrics to customer profile data and confirmed its usefulness. First, we checked the sensitivity of the DQD metrics in terms of the sales contribution efficiency of the data set required for use cases related to customer promotion and marketing. Second, we conducted the verification of our DQDs from the viewpoint of the efficiency of data pre-processing for follow-up tasks such as data analysis and modeling.

III. DATA QUALITY DIMENSIONS AND METRICS

In this section, the quality indicators and metrics for diagnosing the data quality of the CRM customer profile are described. DQD is used to measure, quantify, and manage DQ, and each quality dimension has a specific metric to measure performance [23]. We explain the definition of customer data quality dimension (DQD) and the metric design for determining the quality level of data values based on the DQD.

A. DEFINITION OF CUSTOMER DATA QUALITY DIMENSION

We reviewed various data quality dimensions such as accuracy, completeness, consistency, and timeliness proposed in previous studies and corresponding metrics developed for their quantitative evaluation. In this study, it is not the purpose to define a full set of data quality indicators from the SW engineering point of view, and it is also impossible to prove the completeness and sufficiency of a set of requirements. First of all, we do not directly address DQ dimensions related to aspects such as syntactic criteria or data format (e.g. quality of data schema). In addition, we excluded the metrics which can be generated through the process of collecting answers to questions according to a standard approach for questionnaire development and application because they are difficult to apply automatically in the system. Therefore, among the DQDs reviewed, we selectively introduced *completeness*, *timeliness*, and *plausibility*, which are ones that can be applied and verified because its requirements are relatively clear. In addition, *information* and *multifacetedness* metrics, which are additional indicators to enhance the interpretability of customer data attributes, are newly discovered. Table 1 explains the meaning of each of the five DQDs.

TABLE 1. The meaning of DQDs defined in the study.

Type	DQD	Description
Basic Dimension	Completene ss	the extent to which data is not missing and is of sufficient breadth and depth for the task at hand
	Timeliness	the extent to which data is sufficiently up-to-date for the task at hand
	Information	the extent to which the information content of data is appropriate for the task at hand (information content means the amount of information you gain when an event occurs which has some probability value associated with it)
	Plausibility	the extent to which data does not fall within statistical outlier due to erroneous data generation (e.g., human error) or inconsistent sources (e.g. sensor error)
Topical-Overlap	Multifacete dness	the extent to which data has sufficient coverage for all categories in data model

B. DEVELOPMENT OF DATA QUALITY ASSESSMENT METHODOLOGY - METRIC LOGIC DESIGN OF DQD

1) COMPLETENESS

Completeness of a data set is a commonly used quality metric in the area of data quality and mainly deals with missing values. There are various types of measurements defined in the existing literature to determine missing values [30]. Basically, null values definitely degrade the overall usefulness of the dataset, and non-null values possibly provide useful information. Our basic approach is to measure the missing value within an attribute of the profile data on a per-customer basis, and set the weight as −1 for non-null and −1 for null (e.g. via a specific identifier such as NaN in Python). The metric for completeness on the level of attributes can be defined as the i^{th} customer's *completeness* score for the attributes $\mathrm{a}_{\mathrm{i}1}$, $\mathrm{a}_{\mathrm{i}2}$, …, a_{in}. in j^{th} column. After logic verification, it is possible to design a more sophisticated *completeness* metric, such as the way that reflects the ratio of the number of non-null values as a weight instead of a weight value of 1 or −1. Equation (1) below shows a metric for *completeness*.

$$completeness\_score_i = \sum_{j}^{n} a_{ij}exists$$

$$a_{ij}exists = 0\ if\ a_{ij} \in (null\ or\ NaN)\,,\ 1\ otherwise \quad (1)$$

2) TIMELINESS

Timeliness is the DQD that reflects the degree of data being up-to-date. One of representative definitions is that "*timeliness* can be interpreted as the probability that an attribute value is still up-to-date" [21]. And the metric on the level of an attribute value is defined as (2). We use this equation to define the metric for calculating *timeliness* on the level of attribute values.

$$timeliness\_score_i = \sum\nolimits_j^n \exp(-decline_j * age(i,j)) * a_{ij} exists$$
$$decline_j \text{ is the decline rate}$$
$$age\,(i,j) \text{ is the age of value } a_{ij} \tag{2}$$

The parameter $decline_j$ is the decline rate that indicates how many attribute values have become out of date on average within one period of time. Thereby $age\,(i,j)$ denotes the age of the attribute value a_{ij}, which is computed by means of two factors: the instant when DQ is quantified and the instant of data acquisition. In order to calculate the *timeliness*, it is necessary to select the criteria attributes and the corresponding attributes whose values need to be determined by the criteria attributes. Next, we need to set the decline rate parameter of the *timeliness* by estimating the statistical distribution of the actual database for each attribute or by referring to the statistical basis data of the average effective period for each attribute value (e.g. address, etc.).

3) INFORMATION

Information theory consists of a variety of pure and applied disciplines including mathematical sciences, artificial intelligence, and complexity science, which is concerned with the use, transmission, and assessment of the stochastic properties related to information. Historically, this was introduced by Shannon as the quality of information shared via a set of messages, which are affected by the overall noise [31]. The Shannon entropy of the random variable X is defined as, by definition, equal to the expected information content of measurement of X. We adopt the basic concept of the self-information of a random variable (information content), quantifying how surprising the random variable is "on average". This is the average amount of self-information which an observer would expect to gain about a random variable when measuring it. Following equation (3) is the definition of *information* score that intends to reflect the mean of *information contents* of entries in the customers profile tables.

$$information\_score_i = \sum\nolimits_j^n I_j * a_{ij} exists$$
$$entropy_j = -\sum_{x \in X_j} p\,(x) \log p\,(x) \; \left(if \;\; X_j \; is \; discrete\right) = \sum_{x \in X_j} p\,(x)\, I(x)$$
$$entropy_j = \int_{-\infty}^{\infty} p\,(x) \log \frac{p\,(x)}{q\,(x)} dx \; \left(if \;\; X_j \; is \; continuous\right),$$
$$where \; q\,(x) \; comes \; from$$
$$a \max imum \; entropy \; distribution \tag{3}$$

4) PLAUSIBILITY

An outlier is a variable value or case that distorts the results of statistical data analysis or threatens the appropriateness of data analysis. In general, data analysis obtains insight into valid, invalid, or extreme items by obtaining distribution information by column, utilizing descriptive statistical measurements of the data set to be analyzed. These items can appear in a data set for a number of reasons, such as incorrect data generation (e.g. human input) or inconsistent sources (e.g. sensor or system failure) [32]. *Plausibility* metrics help analysts find outliers by detecting anomalies using non-robust statistical measures such as mean and standard deviation, or robust statistical measures such as median-based interquartile range estimation methods. *Plausibility* DQD is defined as the metric of (4) in consideration of the variable value that distorts the central tendency value of the distribution in the descriptive statistical technique.

$$plausibility\_score_i = \sum\nolimits_j^n plaus_{ij} * a_{ij} exists$$
$$plaus_{ij} = 1 \; if \;\; a_{ij} \in (mean\left(a_j\right) - 2 * s_j, mean\left(a_j\right) + 2 * s_j), 0 \text{ otherwise}$$
$$\text{where } s_j \text{ could be}$$
$$s_j = std\left(a_j\right) \text{ (standard version)}$$
$$s_j = IQR\left(a_j\right)/1.35 \text{ (robust version)} \tag{4}$$

5) MULTIFACETEDNESS

Multifacetedness is an index intended to express the degree of coverage for the topical category of each entry in the customer profile. We defined the categories of the data model to understand customers, map the attributes of our customer profile data to each category, and measure how many categories contain the attributes of our customer profile. Currently, data integration for categories related to "pre-purchase exploration" and "actual use after purchase" has not been completed, so the customer data set is mapped into four categories: basic customer characteristics, purchase, repair/consultation, and management/care. In short, multifacetedness means a numerical value expressing the ratio of the number of variables in the category area of the customer data model, and is defined as (5) below.

$$multifaceted_{score_c} = \sum\nolimits_k^n a_{ij} exists$$
$$for \; some \; (i,j) \; s.t \; j \in Category_k \; and \; i \in Set_c$$
$$where \; n \; is \; \# \; of \;\; Category,$$
$$c \; is \; cust\_id,$$
$$and \; i \; is \; row \tag{5}$$

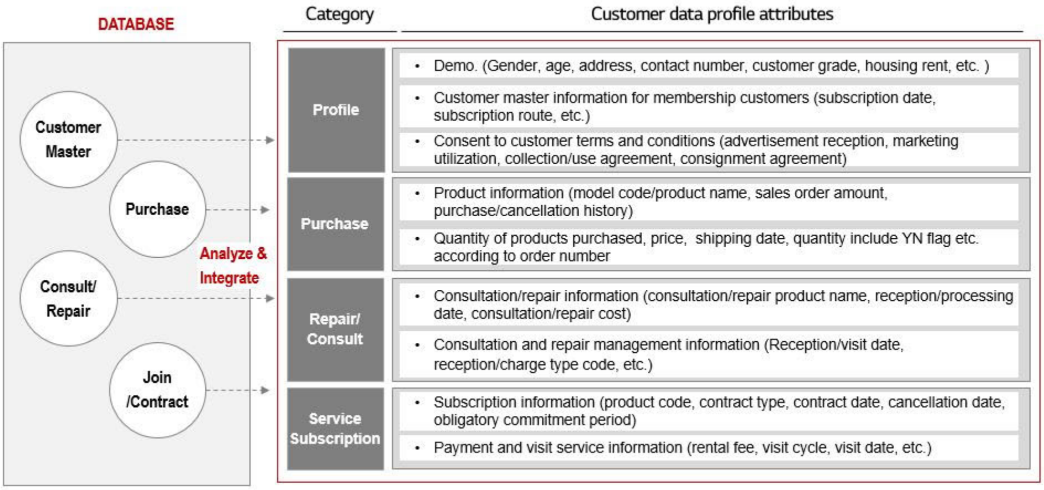
DATABASE
Category
Customer data profile attributes
Customer Master
Purchase
Consult/ Repair
Join /Contract
Analyze & Integrate
Profile
Purchase
Repair/ Consult
Service Subscription
• Demo. (Gender, age, address, contact number, customer grade, housing rent, etc.)
• Customer master information for membership customers (subscription date, subscription route, etc.)
• Consent to customer terms and conditions (advertisement reception, marketing utilization, collection/use agreement, consignment agreement)
• Product information (model code/product name, sales order amount, purchase/cancellation history)
• Quantity of products purchased, price, shipping date, quantity include YN flag etc. according to order number
• Consultation/repair information (consultation/repair product name, reception/processing date, consultation/repair cost)
• Consultation and repair management information (Reception/visit date, reception/charge type code, etc.)
• Subscription information (product code, contract type, contract date, cancellation date, obligatory commitment period)
• Payment and visit service information (rental fee, visit cycle, visit date, etc.)

FIGURE 1. Customer profile data set in CRM.

IV. ASSESSMENT OF DQD METRICS

We introduce our data set, evaluation objectives and scope, and analysis results in this section.

A. CRM DATA SET

Our CRM collects and analyzes customer interaction data from all channels of customer contact points to enable understanding of customers and various target marketing, and configures them into one integrated customer profile. Figure 1 shows a customer profile data set including customer characteristics, purchasing, repair/consultation, and rental care services. For this study, about 1 million customers (1,059,862 in total) were randomly sampled.

B. EVALUATION OBJECTIVES AND SCOPE

As mentioned in the introduction, business performance index and customer satisfaction index have a significant relevance with CRM activities [2], [3], and CRM success has a significant correlation with customer data quality [1], [4]. Therefore, this study examines the effectiveness and suitability of the DQD developed from the perspective of the business performance index. As representative business performance indices, we adopted the RFM index, which is loyalty from the buyer's point of view that implies brand loyalty, and the customer's campaign acceptance index for promotional marketing.

Statistical analysis and ML model-based analysis were performed for the analysis for verification. First, a correlation analysis was conducted between each of the 5 DQD quality score variables and RFM values. And, in order to see if there was a significant relevance between the DQD score and the business performance index (RFM and campaign acceptance), we divided the groups into High and Low groups based on the DQD score, and analyzed the statistically significant differences in the business performance index values between the groups. Next, we defined an ML problem which predicts RFM to verify analytic ease of information and plausibility and analyzed the differences in prediction performance between the DQD High Score group and the Low Score group.

C. ANALYSIS RESULTS

1) CORRELATION BETWEEN 5 DQD QUALITY SCORE VARIABLES AND RFM VALUES

As mentioned earlier, we analyzed the correlation between each of the 5 DQD quality score variables and RFM values to explore some relevance between business performance index and quality scores. Figure 2 shows the results of correlation among Data Quality Dimension scores and correlation between DQD scores and RFM values, respectively. *Completeness* showed a high correlation of about 0.71, 0.75 with *information* and *plausibility*, and there was a correlation of about 0.5 between *information* and *plausibility*. Timeliness showed no correlation with other DQD variables, and *multifacetedness* showed a correlation of about 0.46 with *completeness*. Next, looking at the correlation results of each DQD score with RFM, completeness was correlated with F and M, and there was a high correlation of about 0.75 with RFM. *Timeliness* had weak correlation with R, F, and M but relatively higher correlation with R than F, M.

2) AFTER GROUPING CUSTOMERS BY DQD SCORE VALUE, VERIFYING STATISTICAL SIGNIFICANCE OF RFM VALUES BETWEEN GROUPS

The 5 DQDs were divided into High and Low groups based on the score value, respectively, and the average value of the RFM of each group was derived. Table 2 shows the results of the RFM assessment between High DQD group and Low DQD group. As a result of the statistical significance of the differences in average RFM values between the groups, it was confirmed that the customer group with a high DQD score had a high RFM value on average, which was statistically

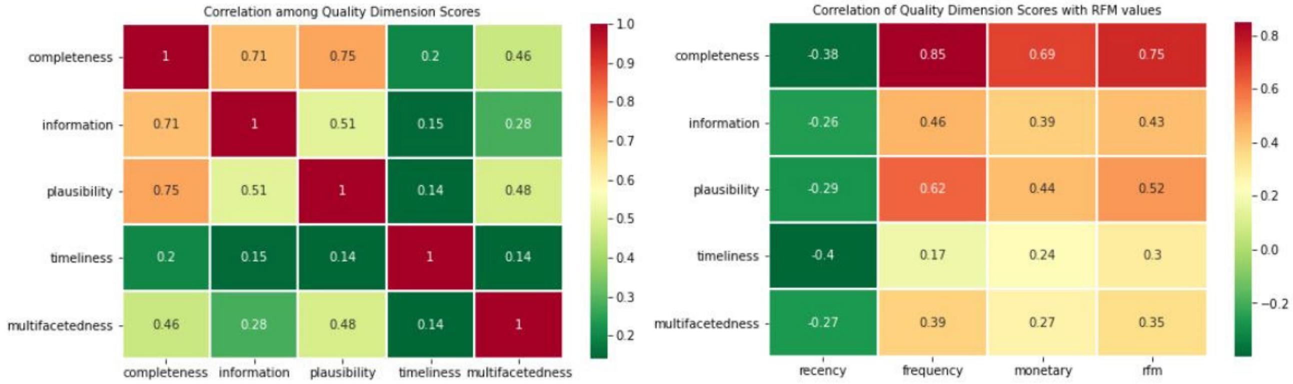

FIGURE 2. Correlation among DQD scores & Correlation btw DQD scores and RFM.

TABLE 2. RFM assessment btw high DQD and low DQD group.

DQDs	RFM in High DQD Group	RFM in Low DQD Group	statistic	p-val.
completeness	0.681506	0.296991	847.6	<0.001
information	0.599078	0.37821	403.1	<0.001
plausibility	0.622786	0.357017	502.6	<0.001
timeliness	0.57633	0.400957	311.2	<0.001
multifacetedness	0.617121	0.402688	380.7	<0.001

significant. As a result of this analysis, it can be seen that whether the customer profile data is maintained faithfully, whether it is up to date, and whether the data model has high coverage by category are important as indicators of the quality of customer data in the CRM domain. The significance of the differences in the average RFM values between the groups in the *information* and *plausibility* is partly related to the results of the previous correlation analysis showing that *information* and *plausibility* have a high correlation with *completeness*. This is because the weight for the number of rows and columns is reflected when calculating *information* and *plausibility*, but basically null values set the score to 0, so the more null values there are, the lower the score value to 0. As the two DQDs, *information* and *plausibility*, were introduced with a focus on ease of analysis, the following section describes the results of evaluating their utilization value in terms of efficiency in data pre-processing.

3) VALUATION OF "ANALYTIC EASE": INFORMATION AND PLAUSIBILITY

Information and *plausibility* are DQD findings with more emphasis on the utilization value of "ease of analysis". To verify its usefulness, customers were classified into High and Low based on the value of each DQD score variable, and the difference in predictive modeling performance between the groups was compared and analyzed. We trained with 741,903 customers, 70% of the total 1,059,862, and tested with 317,959. Customer data was classified into High and Low groups based on the *information* score, and the performance of ML models for classification with RFM as the target variable was compared between the two groups. The features used in the model are the discrete variables involved in calculating the *Information* metric. These variables include not only the original discrete variables, but also the binned continuous variables, and those variables are shown in Table 3.

TABLE 3. A list of feature variables used in the RFM classification prediction model to verify the utilization value of *information*.

Table name	Feature variables
CUST_CHTR_BAS	phn_no_chng_chnl_cd,mobl_phn_chng_chnl_cd,addr_chng_chnl_cd,age,gndr,marry_yn,ktmv_chng_yn,gdsk_apt_lettype,gdsk_house_price, gdsk_house_type
PURC_HIST	sell_amt, sell_qty,pro_lvl1_cd,qty_incl_yn
REPA_HIST	repa_rcp_yyyy, repa_rcp_tp_cd, spcl_repa_cd, pro_grp_cd,repa_amt_bill_tp_cd,repa_tp_cd,repa_amt, acty_coll_repa_amt,repa_sum_amt
CARE_SVC	ctrc_term_mm,ctrc_duty_use_term_mm,ctrc_tp_cd,ctrc_st_cd,vst_tp_cd,rntl_tot_amt,vst_cycl_mm_cont

Table 4 shows the results of the statistical significance of the difference in RFM classification prediction performance between the Low and High Groups based on the *information* score. The three classifiers including LogisticRegression, a parametric model, RandomForestClassifier, a tree-based ensemble classifier, and Adaptive Boosting (AdaBoost) that adaptively creates strong learners using weak learners were utilized. For all three classifiers, the difference in

TABLE 4. Statistical significance of differences in RFM classification prediction performance according to low & high group of *Information* score.

model name	statistic	p-val.	F1 score (Low Group)	F1 score (High Group)
Logistic Regression	-10.98	2.08E-09	0.629	0.734439
Random Forest Classifier	-2.64	1.65E-02	0.885432	0.8997
AdaBoost	-1.94	6.74E-02	0.884277	0.893559

performance values between the groups was statistically significant with a p-value of 0.05 or less.

Next, after grouping customer data based on *plausibility* score, performance comparison experiment was conducted for classification using RFM as a target variable. The features used in the model are the continuous variables involved in calculating the plausibility metric and are shown in Table 5.

TABLE 5. A list of feature variables used in the RFM classification prediction model to verify the utilization value of *plausibility*.

Table name	Feature variables
CUST_CHTR_BAS	age,marry_yn,ktmv_chng_yn,gdsk_house_price
PURC_HIST	sell_amt,sell_qty
REPA_HIST	repa_amt,acty_coll_repa_amt,repa_sum_amt
CARE_SVC	rntl_tot_amt,vst_cycl_mm_cont

Customer data was classified into High 50% and Low 50% of the sorting result based on the *plausibility* score, and a classification prediction performance comparison test was conducted between the groups. Table 6 shows the results of the statistical significance of the differences in RFM classification prediction performance of the Low & High Groups of *plausibility* scores. For all three classifiers, the difference in performance values between the groups was statistically significant with a p-value of 0.05 or less.

TABLE 6. Statistical significance of differences in RFM classification prediction performance according to low & high group of *plausibility* Score.

model name	statistic	p-val.	F1 score (Low Group)	F1 score (High Group)
Logistic Regression	-6.50	<0.001	0.716509	0.828303
Random Forest Classifier	-7.17	<0.001	0.828962	0.882083
AdaBoost	-3.86	<0.001	0.850744	0.887656

Although it is a rather simple problem of "RFM prediction", from the above two experimental results, the utilization value in terms of "analytic ease" of 2 DQDs (*information* and *plausibility*) is explained as follows. We can contribute to providing the analysts who perform various predictive modeling for target marketing using CRM data with the option to reduce the time and effort required for data preprocessing while guaranteeing minimum performance through DQD scores.

4) DETAILED ANALYSIS FOR TIMELINESS

For *timeliness*, in addition to the correlation analysis with RFM, a detailed analysis of the distribution according to the range was conducted. Figure 3 shows the percentage distribution of the number of customers according to the *timeliness* range.

As shown in Figure 3, about 64% (670,000) of all customers exist in the *timeliness* range of 0.3 or less, which means that more than 64% of the attributes involved for *timeliness* metric calculation are not updated to the latest status. Additionally, we looked at the customer distribution percentage of RFM in details by range of *timeliness*, as shown in Figure 4. In the previous correlation analysis results, *timeliness* did not show an overall correlation with RFM, but in the *timeliness* range of 0.0 to 0.1 and 0.1 to 0.2, the proportion of customers in the RFM Low Group was about 11% and 2.6% higher than those in the High Group. However, from the range of 0.2 or higher, it can be seen that the ratio of RFM High Group customers is about 2-3% higher than that of Low Group customers. In short, it can be seen that about 60% of RFM Low customers exist in a relatively low *timeliness* range (0.0∼0.2).

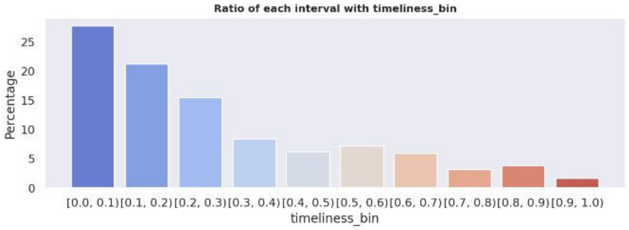

FIGURE 3. Customer proportion distribution by timeliness range.

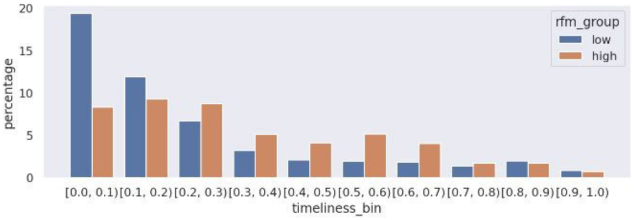

FIGURE 4. Results of comparative analysis by 'RFM Group'— proportion distribution by timeliness range.

Next, a similar analysis was conducted with only the Recency, which is estimated to be intuitively related to *timeliness*, and Figure 5 shows the result. In the *timeliness* range of 0.0∼0.1 and 0.1∼0.2, the percentage of customers in Recency Low Group was about 12% and 0.8% higher than those in High Group. However, from the range of 0.2 or more, as shown in the previous comparison results with the RFM, the ratio of Recency High Group customers is higher than that of Low Group customers. And its difference between the groups is about 9 to 3%, which is larger than that in RFM. In short, it can be seen that about 70% of customers

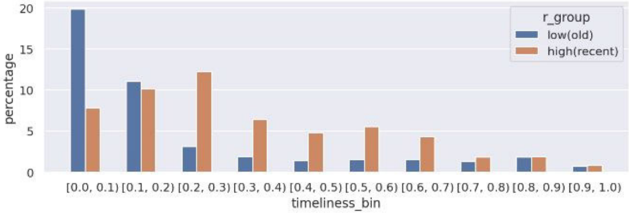

FIGURE 5. Results of comparative analysis by 'Recency Group'— proportion distribution by timeliness range.

in Recency Low Group exist in the relatively low *timeliness* range (0.0~0.2).

5) VERIFICATION OF DIFFERENCES IN THE DISTRIBUTION OF DQD SCORE VALUES ACCORDING TO CAMPAIGN ACCEPTANCE (PROMOTION SUCCESS)

In addition to RFM, which is purchase loyalty, additional analysis was conducted using campaign acceptance, which is an important business performance index in CRM. Using CRM's past promotional campaign execution history data, the contribution of DQDs to the campaign's success was analyzed. When executing promotional campaigns in the future, it can be meaningfully utilized by estimating the improved success rate for each range of DQDs and providing customer data sets above a specific DQD threshold for target marketing performing organizations. A data set of about 142,590 people was constructed by combining customer integrated IDs labeled with campaign (sales promotion) success or not with four tables (customer characteristics, purchasing, repair/consultation, and rental care services) of the profile data model. Table 7 shows the ratio of the number of customers according to PROMT_SUCCS_YN.

TABLE 7. The ratio of the number of customers according to PROMT_SUCCS_YN.

PROMT_SUCCS_YN	Count	Percentage
Y	17,878	13%
N	124,712	87%

The DQD result was derived by applying the DQD metric to the customer profile data of 140,000 customers. In the results below, "N" of the PROMT_SUCCS_YN means "Not-Purchase" and "Y" means "Purchase," respectively for customers who did not accept campaign and for those who accepted it. First, the EDA results are explained for each of DQDs. Figure 6 shows the comparison of the DQDs between the PROMT_SUCCS_YN groups. The "Purchased" customers are relatively more distributed in the range where all the DQDs are higher than those of the "Not-Purchased" customers. In the case of *completeness*, the distribution is extremely skewed, so the analysis results were derived in the range of 0.001 or less. As shown in the Figure 6, the value range in which tends to be prominent is different according to the each DQD. However, it was confirmed that "Purchased" customers have relatively higher DQD scores than those of "Not-Purchased" customers in overall.

In addition to the EDA results, the statistical significance of the difference in DQDs between Y and N groups of "PROMT_SUCCS_YN" was confirmed. As shown in Table 8, the four DQDs except for *multifacetedness* were statistically significantly higher in the customer data profile with high campaign acceptance. In the case of *multifacetedness*, since all campaign target customers had the same value, it was excluded from the analysis of significant differences between the groups.

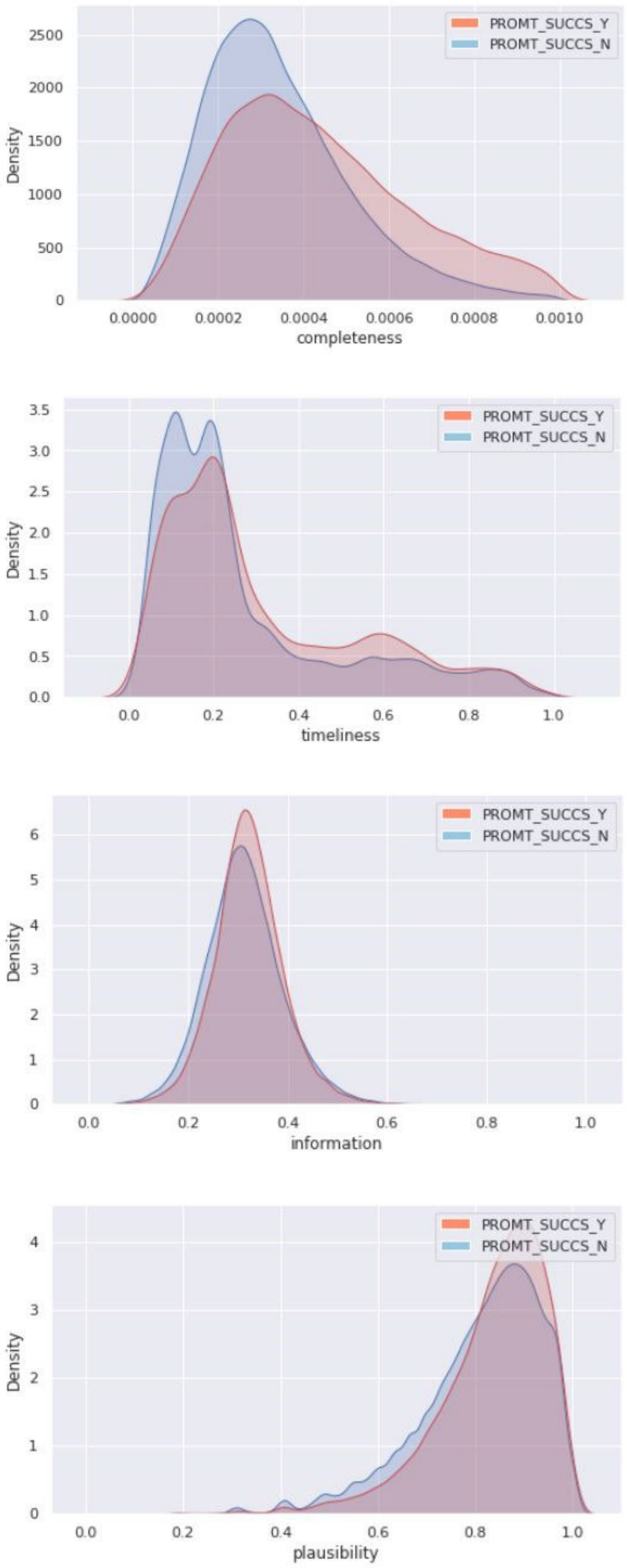

FIGURE 6. The comparison of the DQDs between the PROMT_SUCCS_YN groups.

Next, the total score was calculated from the four DQDs and the total score distribution according to Y/N of PROMT_SUCCS_YN was compared. Figure 7 shows the

TABLE 8. DQD score assessment btw purchase (PROMT_SUCCS_YN = 1) and not-purchase (PROMT_SUCCS_YN = 0).

DQDs	Purchase (17,878)	Not-Purchase (124,712)	statis tic	p-val.
completeness	0.00055	0.00035	61.5	< 0.001
information	0.32113	0.31273	13.5	< 0.001
plausibility	0.83451	0.81185	22.6	< 0.001
timeliness	0.31344	0.27201	22.8	< 0.001

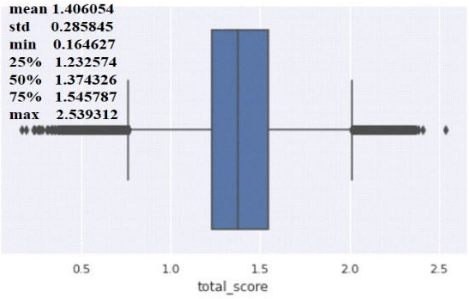

FIGURE 7. Total score distribution - the descriptive statistic.

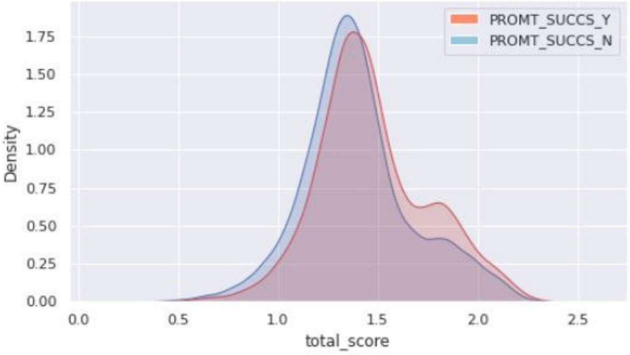

FIGURE 8. Total score distribution according to PROMT_SUCCS_YN.

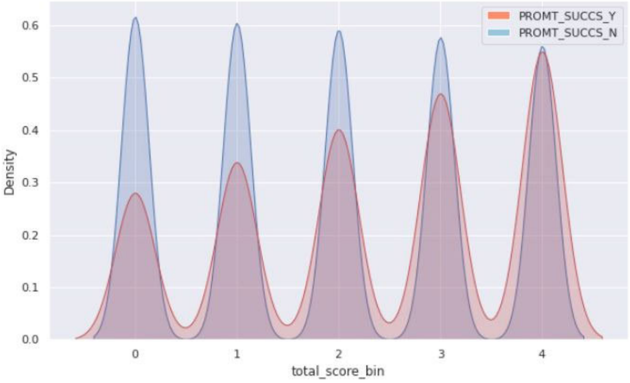

FIGURE 9. The results of the comparative analysis of the Y/N groups according to the total score bin.

descriptive statistic of the total score, and Figure 8 is the analysis result of the distribution difference of the total score generated by the combination of 5 DQDs according to Y/N. It can be seen that customers who accepted the promotional campaign are more distributed in the range with a high DQD total score.

In order to observe the total score by the detailed ranges, we divided it up into equal-sized bins and derived the ratio of the number of PROMT_SUCCS_YN customers within each bin. Table 9 and Figure 9 are the results of analyzing the ratio of the number of Y/N customers according to the binned range of the total score.

TABLE 9. Percentage of customers in PROMT_SUCCS_YN by total_score bin.

TOTAL_SCORE BIN	PROMT_SUCCS_YN	proportion
(0.164, 1.195]	0	91.3914
	1	8.608598
(1.195, 1.322]	0	89.59254
	1	10.40746
(1.322, 1.43]	0	87.66744
	1	12.33256
(1.43, 1.616]	0	85.56
	1	14.44
(1.616, 2.539]	0	83.09839
	1	16.90161

Following describes the results of the comparative analysis of the Y/N groups according to the total score bin. As can be seen in Figure 9, the percentage of campaign failure (N) customers is higher than that of campaign success (Y) customers in the range where the DQD total score is low. Also, as shown in Table 9, the ratio of successful campaign customers increases by about 2% per one bin. In the lowest score range, 8.6% of customers accepted the campaign, and in the highest score range, 17% of customers agreed to the promotion and made a purchase. It would also be worth considering estimating the campaign success rate by further subdividing the total score range or adaptively applying different weights to the importance of each DQD.

V. DISCUSSION

This section discusses some additional analysis results and considerations of DQD metrics.

A. CONFIGURATION PARAMETER FOR TIMELINESS METRIC

The *timeliness* defines the parameter age of the data value based on the time when the data value is created in the real world, and requires setting the value of the parameter *decline rate*. The *decline rate* of the criterion attribute for *timeliness* can be estimated from the sample of customer's own historical data. However, although the attribute value was replaced with the most recent data, it was not possible to directly estimate the change frequency of the attribute value because the change history data could not be accessed for reasons such as the Personal Information Protection Act. Alternatively, we could set the parameter *decline rate*of the *timeliness* metric, for example, after surveying the average lifespan of customer addresses (i.e., how long, on average, do customers live in the same place?).

Since this method incurs cost as the number of samples increases, a method using third party data was chosen. For example, the parameters of ''phone number'' and ''address'' attributes were set with reference to the generalized data from the National Statistical Office considering the mobile phone replacement frequency and moving cycle of carriers. Therefore, we set a *decline rate* of 0.02 for the ''mobile phone number'' attribute (i.e., on average, 2% of all customers change their mobile phone number) and a *decline rate* of 0.01 for the ''address'' attribute. Since customer IDs generally remain the same, we assumed a *decline rate* of 0.0 for the ''id'' attribute. In the case of ''change date of promotion management organization'', it was estimated to be 0.1 based on past data provided by the person in charge of the current CRM promotion management organization.

Finally, the age of each attribute is calculated based on the data storage time and DQD metric scoring (quantification) time of each reference attribute. Figure 10 shows the comparison of the *timeliness* score of two attributes between the Recency groups. In both cases, the ''High Recency Group'' customers are relatively more distributed in the range where all of the score values are higher than those of the ''Low Recency Group'' customers. However, the peak points of each score for each group show different aspects. For example, in the case of phn_no_drate, scores in Low Recency Group peak at 0.05 and 0.1 and scores in High Recency Group tend to be prominent at around 0.4, and 0.8. However, the addr_chng_drate shows quite a different shape. Therefore, reliability of the accuracy of the DQD metrics of *timeliness* actually requires a direct evaluation (e.g. contacting the customer) of whether the data values are still up-to-date. However, it is necessary to evaluate and verify the efficiency in terms of costs incurred in setting the configuration parameters for the accuracy of the DQD metric calculation for the actual CRM customers.

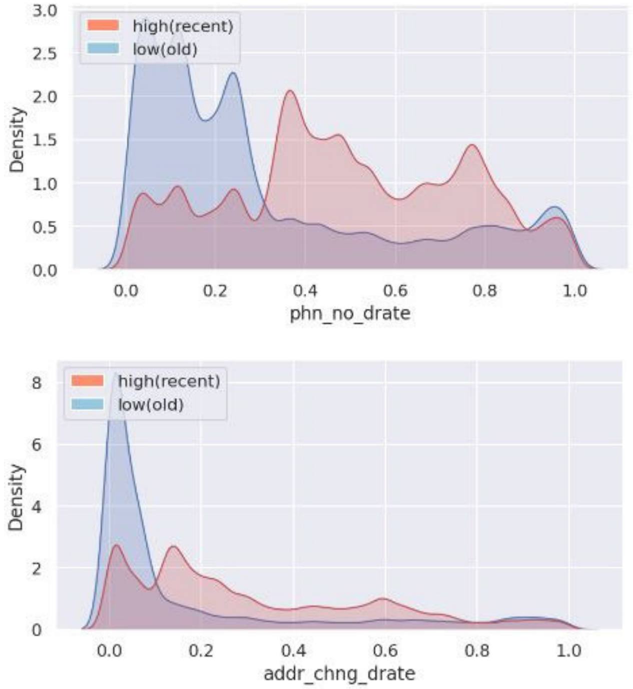

FIGURE 10. Results of comparative analysis by Recency (High vs. Low) — ''Timeliness'' score of phone number and address change attributes.

In this study, attributes such as the date of customers' agreement to terms and conditions and the date of membership subscription were not yet integrated into our CRM, so they were not involved for the application of *timeliness* metrics. In order to more sophisticatedly support customer segmentation of target marketing, it will be necessary to expand the range of criteria attributes for *timeliness*. Of course, analyzing the contribution of the update of terms and conditions or membership subscription-related dates to the *timeliness* metric will be a meaningful follow-up study. In addition to the attribute on the timing of information update in table units, it is also possible to consider adding a separate dedicated timestamp attribute to track changes in the attributes that we agreed with marketing managers to be important for the application of *timeliness*.

B. RANKING CRITICAL FACTORS OF INFORMATION METRIC SCORE

We investigated how much each attribute contributes to increasing the *information* score for each customer. To this end, the *information* contribution of each attribute in each customer data profile was derived and ranked, and the final ranking value was derived by summing the ranking values of all customers for each attribute. Following is how we rank different columns based on the amount of contribution for the information score:

$$factor_{i,j} = \sum_{i_k} I_{i_k,j} \Big/ \sum_t w_{i,t}$$
$$I_{i_k,j} = -\log\left[p_{i_k,j}\right] = \log\left(1/p_{i_k,j}\right) \qquad (6)$$

where $factor_{i,j}$ denotes proportion of contribution of j^{th} column for the i^{th} customer's *information* score. $w_{i,j} = \sum_{i_k} 1$ is the number of rows that i^{th} customer has in the t^{th} table. $I_{i_k,j}$ is the information content of i^{th} customer that is attributed to the category(class) present in the i_k^{th} row, j^{th} column. For each customer i, we can rank $factor_{i,j}$ based on its absolute value. We call it $rank_i\left(factor_{i,j}\right)$. For convention, we give higher number for higher rank. Then, $\sum_i rank_i\left(factor_{i,j}\right)$ represents relative contribution of j^{th} column on the information score.

Table 10 shows the ranking order and ranking values of the attributes derived by (6) above. An attribute with a high final ranking value can be interpreted as an attribute with a high ratio of being ranked high in the *information* score of all customers.

Figure 11 shows the ratio of the number of categories (class) and *information content* value of pro_lvl1_cd (Product level1 code) value, which is the No. 1 ranking attribute. The lower the ratio of the count of values within the category of the attribute, the higher the value of the *information content*.

TABLE 10. Results of ranking based on the amount of contribution for the *Information* score.

Rank	Attribute	Description	Ranking values
1	pro_lvl1_cd	Product level 1 code	1.04E+06
2	qty_incl_yn	Quantity include YN flag	9.20E+05
3	age	Age	8.75E+05
4	gdsk_apt_lettype	Apartment permanent lease type	7.59E+05
5	sell_amt	Sales order amount	7.52E+05
6	gdsk_house_type	General housing type	6.93E+05
7	gndr	Gender	6.73E+05
8	repa_rcp_yyyy	Year of repair	6.67E+05
9	pro_grp_cd	Product group code	6.64E+05
10	marry_yn	Marital status	6.13E+05
11	repa_amt	Repair amount	6.01E+05
12	repa_amt_bill_tp_cd	Repair amount and charge type code (repair cost type (with or without charge))	5.99E+05
13	addr_chng_chnl_cd	Address change channel code	5.51E+05
14	repa_rcp_tp_cd	Repair request type code	5.40E+05
15	ktmv_chng_yn	Whether or not KT moving is changed	4.96E+05
16	repa_tp_cd	Repair type code	4.86E+05
17	rntl_tot_amt	Total rental amount	4.81E+05
18	phn_no_chng_chnl_cd	Phone number change channel code	4.61E+05
19	ctrc_term_mm	Contract period month	4.49E+05
20	ctrc_tp_cd	Contract type code	4.38E+05
21	vst_cycl_mm_cont	The number of visits per month	4.31E+05
22	ctrc_duty_use_term_mm	Mandatory use period	4.27E+05
23	mobl_phn_chng_chnl_cd	Mobile phone change channel code	4.13E+05
24	sell_qty	Sales order quantity	4.01E+05
25	acty_coll_repa_amt	Actual collection amount	3.77E+05
26	ctrc_st_cd	Contract status code	3.65E+05
27	gdsk_house_price	General housing price	3.54E+05
28	repa_sum_amt	Total repair amount	3.54E+05
29	spcl_repa_cd	Special treatment code	3.53E+05
30	vst_tp_cd	Visit type code	1.95E+05

Of course, when the ratio of the count of values in each category is the same, the higher the number of categories, the higher the value of *information content* tends to be. The attribute of pro_lvl1_cd has a total of 42 product categories, and we can see that the number ratio for each category is different, and it can be assumed that the attribute has a large amount of information about which product group the customers purchased. On the other hand, vst_tp_cd (visit type code), the lowest ranking attribute, has a single category value, and the ratio is 43%, and the remaining 57% is confirmed as a null value.

It can be seen that these attributes provide little distinction in understanding customers, so we can do quality improvement efforts to track the cause of null values and to segment categories of the corresponding attributes. Figure 12 shows the ratio of the number of categories (class) and information content value of repa_tp_cd (repair type code) value, which is the 15th ranked attribute. "repa_tp_cd" is an attribute for the customer's product repair service type code. About 90% of customers had an "A" value (general repair) in the repair type category, followed by an "F" value (heavy repair) at 4.6%. The value of the information content is about 0.1 for the "A" category and about 4.4 for the "F" category, indicating that about 90% of customers have very poor discrimination in terms of repair type attributes.

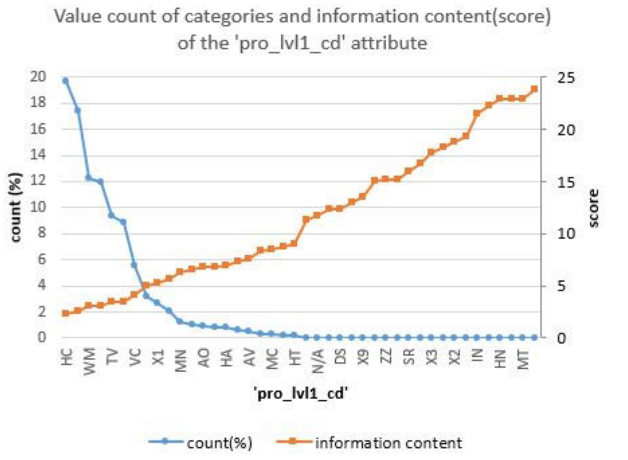

FIGURE 11. Value count of categories and information content of the 'pro_lvl1_cd' attribute.

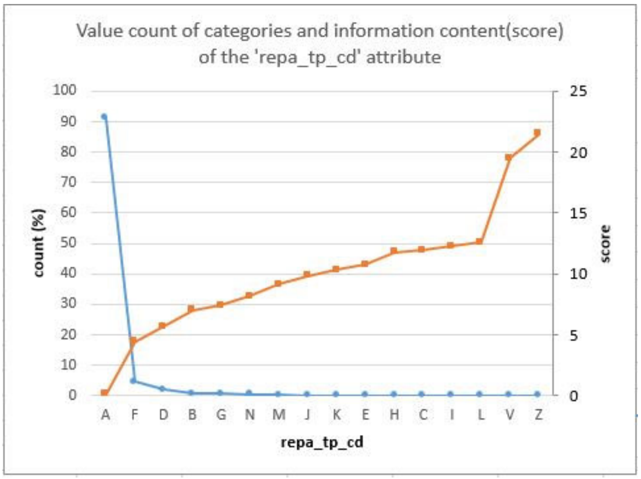

FIGURE 12. Value count of categories and information content of the 'repa_tp_cd' attribute.

In this way, it is also possible to consider improving the granularity of a specific attribute by referring to the *information* score value. If the category of "general repair" of the above repair type code is subdivided (e.g., "general repair" subdivided as C, D, E), it will be possible to increase the understanding and differentiation of customers through the corresponding attribute.

VI. CONCLUSION AND FUTURE WORKS

In this study, we explored the benefits of applying 5 DQD metrics in the context of the use case that performs prospective customer prediction modeling for target marketing using integrated customer profile data of a CRM system through RFM indicators and past campaign acceptance data. As the deliverables of this study, we can provide summary statistics (average and standard deviation of measurements) and total score of DQD metrics of CRM customer datasets for target marketing. By referring to the metric score for each DQD, quality dimensions that reduce the usefulness of specific attributes can be identified and shared with field data managers to support the establishment of quality improvement plans.

Next, the utilization value of each DQD metric is as follows. First of all, *completeness*, *timeliness* and *multifacetedness* are useful indicators for managing customer profiles in terms of business performance index. In addition, *information* and *plausibility* are meaningful in terms of ease of analysis and efficiency to support advanced analysts in identifying the minimum data set required to derive reasonable analysis results. *Multifacetedness*will be useful for management in terms of the customer interaction index, which indicates how many different traces customers have left in their interactions with a company. However, it is necessary to integrate categorical data related to ''pre-purchase search'' and ''actual use after purchase'' into our CRM, which are not currently applied, and standardization work to secure universality of the categories of the profile data model.

This study has the following limitations. In this study, we developed a DQD metric to identify and quantify problems that may exist in a CRM's customer data sets related to poor quality affecting business performance. However, it is necessary to develop a series of fixing functions tightly coupled with the results of the developed DQD metrics so that it's possible to improve data quality by actually dealing with low quality problems. Also, we confirmed the validity of data quality metrics supporting use cases for marketing purposes based on RFM, an analysis technique related to loyal customers. As a business performance index, it is necessary to verify its effectiveness by directly applying DQD metric not only to the past RFM but also to the prediction of potential future purchase possibilities. In addition, it is necessary to compare the cost of applying DQD metrics to massive customer data with the economic usefulness of marketing costs. The most challenging work in DQD metric design is that it is difficult to generalize how to measure the usefulness of data because it depends on the use cases supported. Therefore, in addition to the business performance aspect, research on DQD metrics and verification methods to support use cases in terms of customer experience satisfaction will be a meaningful follow-up study.

REFERENCES

[1] M. Petrović, ''Data quality in customer relationship management (CRM): Literature review,'' *Strategic Manage.*, vol. 25, no. 2, pp. 40–47, 2020, doi: 10.5937/StraMan2002040P.

[2] M. Cvjetković, M. Vasiljević, M. Cvjetković, and M. Josimović, ''Impact of quality on improvement of business performance and customer satisfaction,'' *J. Eng. Manage. Competitiveness*, vol. 11, no. 1, pp. 20–28, 2021, doi: 10.5937/jemc2101020C.

[3] M. Stone, *Defining CRM and Assessing Its Quality*. London, U.K.: Kogan Page, 2001, pp. 1–16.

[4] A. Negahban, D. J. Kim, and C. Kim, ''Unleashing the power of mCRM: Investigating antecedents of mobile CRM values from managers' viewpoint,'' *Int. J. Hum.-Comput. Interact.*, vol. 32, no. 10, pp. 747–764, Oct. 2016, doi: 10.1080/10447318.2016.1189653.

[5] F. C. Payton and D. Zahay, ''Understanding why marketing does not use the corporate data warehouse for CRM applications,'' *J. Database Marketing Customer Strategy Manage.*, vol. 10, no. 4, pp. 315–326, Jul. 2003, doi: 10.1057/palgrave.jdm.3240121.

[6] A. Even, G. Shankaranarayanan, and P. D. Berger, ''Evaluating a model for cost-effective data quality management in a real-world CRM setting,'' *Decis. Support Syst.*, vol. 50, no. 1, pp. 152–163, Dec. 2010, doi: 10.1016/j.dss.2010.07.011.

[7] S. Sharma, D. P. Goyal, and R. K. Mittal, ''Imperative relationship between data quality and performance of data-mining tools for CRM,'' *Int. J. Bus. Competition Growth*, vol. 1, no. 1, pp. 45–61, Apr. 2010, doi: 10.1504/IJBCG.2010.032828.

[8] L. L. Pipino, Y. W. Lee, and R. Y. Wang, ''Data quality assessment,'' *Commun. ACM*, vol. 45, no. 4, pp. 211–218, Apr. 2002, doi: 10.1145/505248.506010.

[9] A. Even and G. Shankaranarayanan, ''Utility-driven assessment of data quality,'' *ACM SIGMIS Database, Database Adv. Inf. Syst.*, vol. 38, no. 2, pp. 75–93, May 2007.

[10] B. Heinrich, M. Kaiser, and M. Klier, ''How to measure data quality? A metric-based approach,'' in *Proc. 28th Int. Conf. Inf. Syst. (ICIS)*, Dec. 2007, pp. 1–15.

[11] M. Mosley, M. Brackett, and S. Earley, Eds. *The DAMA Guide to the Data Management Body of Knowledge Enterprise Server Version*. Sydney, NSW, Australia: Technics Publications, LLC, Westfield, 2009.

[12] D. Loshin, *The Practitioner's Guide to Data Quality Improvement*. San Mateo, CA, USA: Morgan Kaufmann, 2010.

[13] K. M. Hüner, A. Schierning, B. Otto, and H. Österle, ''Product data quality in supply chains: The case of beiersdorf,'' *Electron. Markets*, vol. 21, no. 2, pp. 141–154, Jun. 2011, doi: 10.1007/s12525-011-0059-x.

[14] S. U. Natchiar and S. Baulkani, ''Customer relationship management classification using data mining techniques,'' in *Proc. Int. Conf. Sci. Eng. Manage. Res. (ICSEMR)*, Nov. 2014, pp. 1–5, doi: 10.1109/ICSEMR.2014.7043662.

[15] C. W. Fisher, E. J. M. Lauria, and C. C. Matheus, ''An accuracy metric: Percentages, randomness, and probabilities,'' *J. Data Inf. Qual.*, vol. 1, no. 3, pp. 1–21, Dec. 2009, doi: 10.1145/1659225.1659229.

[16] R. Blake and P. Mangiameli, ''The effects and interactions of data quality and problem complexity on classification,'' *J. Data Inf. Qual.*, vol. 2, no. 2, pp. 1–8, Feb. 2011, doi: 10.1145/1891879.1891881.

[17] Y. Timmerman and A. Bronselaer, ''Measuring data quality in information systems research,'' *Decis. Support Syst.*, vol. 126, Nov. 2019, Art. no. 113138, doi: 10.1016/j.dss.2019.113138.

[18] A. Wechsler and A. Even, ''Using a Markov-chain model for assessing accuracy degradation and developing data maintenance policies,'' in *Proc. 18th Amer. Conf. Inf. Syst.*, vol. 3, 2012, pp. 1–6. [Online]. Available: http://aisel.aisnet.org/amcis2012/proceedings/DataInfoQuality/3

[19] B. Heinrich and D. Hristova, ''A quantitative approach for modelling the influence of currency of information on decision-making under uncertainty,'' *J. Decis. Syst.*, vol. 25, no. 1, pp. 1–26, Sep. 2015, doi: 10.1080/12460125.2015.1080494.

[20] C. Batini and M. Scannapieco, ''Data quality dimensions,'' in *Data and Information Quality* (Data-Centric Systems and Applications). Cham, Switzerland: Springer, Mar. 2016, doi: 10.1007/978-3-319-24106-7_2.

[21] B. Heinrich, M. Kaiser, and M. Marcus, ''How to measure data quality?—A metric-based approach,'' in *Proc. ICIS*, 2007, p. 108. [Online]. Available: https://aisel.aisnet.org/icis2007/108

[22] B. Heinrich and M. Klier, ''Metric-based data quality assessment—Developing and evaluating a probability-based currency metric,'' *Decis. Support Syst.*, vol. 72, pp. 82–96, Apr. 2015, doi: 10.1016/j.dss.2015.02.009.

[23] I. Taleb, M. A. Serhani, C. Bouhaddioui, and R. Dssouli, ''Big data quality framework: A holistic approach to continuous quality management,'' *J. Big Data*, vol. 8, Dec. 2021, Art. no. 76, doi: 10.1186/s40537-021-00468-0.

[24] C. Bors, T. Gschwandtner, S. Kriglstein, S. Miksch, and M. Pohl, ''Visual interactive creation, customization, and analysis of data quality metrics,'' *J. Data Inf. Qual.*, vol. 10, no. 1, pp. 1–26, Mar. 2018, doi: 10.1145/3190578.

[25] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, ''Profiler: Integrated statistical analysis and visualization for data quality assessment,'' in *Proc. Int. Work. Conf. Adv. Vis. Interfaces*, May 2012, pp. 547–554, doi: 10.1145/2254556.2254659.

[26] A. Even, G. Shankaranarayanan, and P. D. Berger, ''Evaluating a model for cost-effective data quality management in a real-world CRM setting,'' *Decis. Support Syst.*, vol. 50, no. 1, pp. 152–163, Dec. 2010, doi: 10.1016/j.dss.2010.07.011.

[27] L. Ehrlinger and W. Wöß, ''A survey of data quality measurement and monitoring tools,'' *Frontiers Big Data*, vol. 5, Mar. 2022, Art. no. 850611, doi: 10.3389/fdata.2022.850611.

[28] B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz, ''Requirements for data quality metrics,'' *J. Data Inf. Qual.*, vol. 9, no. 2, pp. 1–32, Jun. 2017, doi: 10.1145/3148238.

[29] J. W. Peltier, D. Zahay, and D. R. Lehmann, ''Organizational learning and CRM success: A model for linking organizational practices, customer data quality, and performance,'' *J. Interact. Marketing*, vol. 27, no. 1, pp. 1–13, Feb. 2013, doi: 10.1016/j.intmar.2012.05.001.

[30] C. Batini and M. Scannapieco, ''Data quality: Concepts, methodologies and techniques,'' in *Data and Information Quality: Data-Centric Systems and Applications*. New York, NY, USA: Springer, 2006, doi: 10.1007/3-540-33173-5.

[31] C. E. A. Shannon, ''Mathematical theory of communication,'' *Bell Syst. Tech. J.*, vol. 27, no. 4, pp. 623–666, 1948.

[32] T. Gschwandtner, J. Gärtner, W. Aigner, and S. Miksch, ''A taxonomy of dirty time-oriented data,'' in *Proc. CD-ARES*, in Lecture Notes in Computer Science, vol. 7465, 2012, pp. 58–72, doi: 10.1007/978-3-642-32498-7_5.

[33] C. Cichy and S. Rass, ''An overview of data quality frameworks,'' *IEEE Access*, vol. 7, pp. 24634–24648, 2019, doi: 10.1109/ACCESS.2019.2899751.

[34] G.-O. Meritxell, B. Sierra, and S. Ferreiro, ''On the evaluation, management and improvement of data quality in streaming time series,'' *IEEE Access*, vol. 10, pp. 81458–81475, 2022, doi: 10.1109/ACCESS.2022.3195338.

[35] M. Bovee, R. P. Srivastava, and B. Mak, ''A conceptual framework and belief-function approach to assessing overall information quality,'' *Int. J. Intell. Syst.*, vol. 18, no. 1, pp. 51–74, Jan. 2003, doi: 10.1002/int.10074.

[36] I. El Alaoui, Y. Gahi, and R. Messoussi, ''Big data quality metrics for sentiment analysis approaches,'' in *Proc. Int. Conf. Big Data Eng.*, Jun. 2019, pp. 36–43, doi: 10.1145/3341620.3341629.