IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 June 2024, accepted 29 July 2024, date of publication 2 August 2024, date of current version 13 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3437648

RESEARCH ARTICLE

Repurchase Prediction Using Survival Ensembles in CRM Systems for Home Appliance Business

YOUNGJUNG SUH
Department of Computer Science and Engineering, Kongju National University, Cheonan-si, Chungcheongnam-do 31080, Republic of Korea
e-mail: youngjung.suh@kongju.ac.kr

This work was supported by Kongju National University in 2024.

ABSTRACT In order for company's promotions to continue to have a beneficial impact on sales, it is important for companies to identify which of the interested buyers can be converted into repeat buyers. By targeting these potential loyal customers, companies can significantly reduce promotional costs and increase return on investment. The existing studies related to repurchase prediction in the e-commerce area have focused on the statistical techniques and more common binary classification models. In this paper, we propose a survival analysis-based machine learning/deep learning model to predict TV repurchase time of customers using home appliance company's CRM data. The prediction model is verified based on actual operational data such as customer profile, purchase, counseling, and repair history for approximately 1.45 million customers in electronics company's CRM. As a deep learning method, Algo 6-1 (DeepHit with the feature set selected from Cox regression and preprocessed with multiple imputation) achieved the best performance (c-index 0.828). Algo3 (Random Survival Forest with the feature set selected from Cox regression and preprocessed with multiple imputation), a machine learning method, not only showed similar performance to deep learning (c-index 0.823), but also provided insights in key features that influenced repurchase. In addition, we provided a utility function that provides TV repurchase probability over time so that marketers can cost-effectively determine the timing to provide promotional events or benefits to customers.

INDEX TERMS Big data applications, repurchase prediction, predictive models, customer relationship management, ensemble learning, home appliance business.

I. INTRODUCTION

Customer retention refers to the rate at which customers stay with a business in a given period of time and is a key metric for practically all B2B and B2C businesses. To enhance corporate competitiveness through extending the customer retention period, customer churn should first be predicted to reduce the possibility of churn, to bring economic benefits to the enterprise [1], [2], [3], [4]. Other related studies have emphasized the need for strategies to maintain existing customers arguing that customer maintenance costs are lower than the cost of attracting new customers [5], [6], [7], [8].

In the case of predicting customer churn to maintain customer retention, the method is different in contractual and non-contractual business settings. First of all, customer churn models applicable to the contract settings-based companies are generally developed based on a binary classifier for whether to churn or not. Such companies develop their own prediction models with predictive good power by using well-known models such as logistic regression, random forests, gradient boost approaches, and other classifiers to estimate who will churn [9].

However, in a non-contractual business setting, there is uncertainty about both the target and the timing of churn. "Customer churn" for companies whose business model is customer purchases is defined as a case in which a user who has made a transaction at least once does not make a repeat purchase for a certain period of time [10], [11]. In order for product sellers (non-contractual business settings) to secure loyal customers, it is important to simply

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

prevent existing customers from churn, but it is even more important to predict the likelihood of repurchase. In other words, a more preemptive and active method than directly estimating customer churn to retain loyal customers is ''inducing repurchase'' [11], [12].

Research on repurchase has received widespread attention, especially in the marketing field. The most representative technique for predicting customer repurchase is the BTYD (Buy Till You Die) model [14], [15], [16]. Unlike the BTYD model, a distinctly different approach to predicting customer repurchase is machine learning [17]. This class of models includes a wide variety of computational and statistical learning algorithms [18], [19], [20], [21], [22]. There has been also researches on comparison between the BTYD model (para-metric model) and machine learning algorithms (observation-driven models) in customer-based analysis [23], [24].

Most of these studies quantifying repurchase likelihood have focused on predicting repurchase behavior in terms of what to buy in the e-commerce sector. However, repurchase prediction studies can help marketing strategists by estimating not only whether existing customers are likely to make follow-up purchases but also when those purchases are likely to occur, providing optimal resale timing. It is because it may be worthwhile to maintain customer relationships if the likelihood of a repurchase is high and the duration is short, but it may be less rewarding if the probability is low and there is some distance in the future. In other words, a cost-effective marketing strategy can be established depending on the length of the subsequent purchase period. However, researches on establishing customer retention strategies by predicting repurchase times using actual CRM data from home appliance companies, remain largely unexplored.

Therefore, this study examines repurchase time prediction models for proactive customer retention management in home appliance sales businesses. We utilize an event-time analysis technique called survival analysis, which is a type of modified regression task but works well even with partially incomplete (censored) data [25]. Specifically, the survival ensemble approach was used to estimate the time until the next TV purchase to model the prediction of repurchase time. To this end, we conducted an integrated analysis of customer-company interaction data including customer purchase history, demographic information, counseling history, and repair history. In the TV sales business of the electronics company that inspired this study, the level of promotion timing was differentiated not by customer but by customer cluster (e.g., sales strategy for large categories such as number of months of use or product lines). Therefore, the purpose of this study is to propose a promotional campaign tool based on survival analysis machine learning that enables customized promotional activities for customers by effectively learning and predicting the likelihood and expected timing of customer repurchase.

The contributions of this paper are the following two.

1) Rather than the statistical techniques and binary classification models of existing studies related to repurchase prediction in the e-commerce domain, labeling logic that reflects the customer's purchase cycle characteristics of the home appliance business domain was applied to Survival Analysis-based predictive modeling. We developed a survival ensemble model that predicts the timing of customer repurchase by analyzing information on factors that affect customer repurchase. Additionally, the approach proposed in this paper not only provides a list of possible repeat customers, but also provides a survival probability function (a utility function that estimates the probability of repurchase within N days) that tells how the repurchase probability for a target customer changes as a function of time. Our model allows us to distinguish between planned re-purchasers, near and distant future re-purchasers, and the variables that influence this repurchase behavior. In particular, we can create a cost-effective marketing strategy based on the follow-up purchase period: if the predicted probability of purchase is high and the period is short (e.g. within 3 months), intensive active marketing is carried out, and if the predicted probability of purchase is low and there is some distance in the future (e.g. after 1 year), passive marketing is carried out to reduce costs.

2) In this study, the repurchase prediction model was verified using customer-company interaction data such as customer purchase history, demographic information, and counseling and repair history for products from actual CRM products, rather than a benchmark data set. We conducted modeling to predict the timing of repurchase for each customer based on survival analysis machine learning using an actual dataset (1,452,316 TV purchase customers). The performance of the repurchase prediction model was verified by using the c-index score, with a performance of 0.823 for the Random Survival Forest model (with selected features and multiple imputation) and about 0.828 for the Deep Learning model (with selected features and multiple imputation). Our survival ensemble modeling with features selected based on empirical research on home appliance customer repurchase behavior showed similar performance to more complex neural networks. Survival ensembles are not completely overwhelmed by deep learning approaches and can improve their performance. Our findings are highly beneficial to numerous marketers who prefer to adopt simple, reliable, and interpretable predictive models for consumer marketing analysis.

The remainder of this paper is organized as follows. In ''Related work'' Section, the related studies on repurchase prediction are reviewed. In ''Survival analysis-based prediction'' Section, survival analysis for modeling is introduced. In ''Modeling'' Section, the details of the modeling for customer repurchase prediction are presented. In ''Model validation'' Section, the experimental setting is described and an analysis of the experimental results is presented. The

final section concludes the study and offers further research directions.

II. RELATED WORKS

In this section, we investigate existing researches on customer repurchase prediction. Category A includes research literature on traditional statistical and machine learning techniques from a methodological perspective for repurchase prediction. The ML-based repurchase prediction studies mentioned above are mainly binary classification studies. Meanwhile, even if a user incorrectly predicted as non-purchasing does not repurchase at the accurately labelled specific point in time, he or she may repurchase after a certain period of time. Conversely, there is also the possibility of losing the loyalty of current repeat customers at some point in the future. Survival Analysis is a representative technique that overcomes the limitations of binary classification and estimates the time point [25]. Category B includes studies on survival analysis.

A. REPURCHASING PREDICTION

Customer repurchase prediction has received extensive research attention in the fields of marketing, operations, statistics, and computer science. In marketing field, the most representative technique for predicting customer repurchase is the BTYD (Buy Till You Die) model [14], [15]. BTYD models are certainly powerful in that they can extract information from only a small number of high-dimensional customer features (i.e. recency and frequency). Inspired by the BTYD model, there is also a study developed for modeling repeat purchase recommendations in the e-commerce sector that recommends repeat purchase products to customers based on their purchase history [16]. The authors demonstrated 7% increase in click through rate for products on the Amazon.com personalized recommendations page.

A distinctly different approach from the BTYD model to predictive modeling for customer repeat business is machine learning [17]. This type of approaches includes a wide variety of computational and statistical learning algorithms. In the ''Amazon'' research mentioned just before, the authors presented as their future works the plans of investigating the methods to help in improving the quality of their recommendations. They planned to explore some recent models that are BTDs like the BG-NBD model, and supervised learning models like Logistic Regression, neural networks. Unlike the BTYD model, which seeks to explicitly model behavioral processes through probability distributions, machine learning-based methodologies take a data-driven approach to predictive modeling [18], [19], [20], [21], [22].

In [20], the authors designed a two-layer fusion ensemble machine learning based on GBDT (TMFBG), and applied it to repurchase prediction in E-business. They showed the results that the TMFBG has greater robustness and more accurate prediction results than the single base classifier. The algorithm was validated on the data obtained from the behavior of certain customers of the yearly ''Double 11'' on Tmall platform. Machine learning ensembles have been also used in [21] to propose an online shopping behavior analysis and prediction system in China's e-commerce industry. They adopted linear model logistic regression and decision tree based XGBoost model. After optimizing the model, it was found that the nonlinear model can make better use of these features and get better prediction results. The authors in [22] studied how to use edge computing to collect customer shopping data accurately. Then, they established a mathematical model by a joint model of Long-Short Term Memory neural network model and convolutional neural network model. Based on this model, a method of information segmentation processing was proposed to further improve the prediction accuracy of the neural network model for consumer shopping behavior. They demonstrated the prediction accuracy of a variety of neural network models by more than 2%, and the one of the models based on Extreme Gradient Boosting by 5.4%.

Meanwhile, there also have been researched on comparative analyses between BTYD and ML-based methodologies. In the proposal of [23], the authors conducted a dynamic rolling comparison between the Pareto/NBD model (parametric model) and machine learning algorithms (observation-driven models) in customer-based analysis, which the literature related to this has not comprehensively investigated before. The authors presented their findings from those comparisons in terms of assisting both in defining the comparative edge and implementation timing of these two approaches and in modeling and business decision making. Similarly, the authors of [24] presented predictive analytics for customer repurchase by interdisciplinary integration of Buy Till You Die modeling and machine learning. Using a large online retail data, they empirically assessed the prediction performance of BTYD modeling and machine learning. More importantly, they investigated how the two approaches could complement each other for repurchase prediction. They used the BG/BB model given the discrete and non-contractual problem setting and incorporated BG/BB estimates into high-dimensional Lasso regression. They showed the proposed Lasso-BG/BB outperforms two sophisticated recurrent neural networks, validating the complementarity of machine learning and BTYD modeling. Their work can be said to be meaningful in that it explains how the interdisciplinary integration of the two modeling paradigms contributes to the theory and practice of predictive analytics.

B. SURVIVAL ANALYSIS

The studies on ML-based repurchase prediction mentioned above are mainly studies on binary classification. Meanwhile, Survival Analysis is a representative technique that estimates the timing and overcomes the risk of losing the loyalty of future repeat customers depending on the predicted timing. Time-to-event analyses are important methods to help us analyze problems with a temporal component to our research question. As the name suggests, these are used when we are interested in understanding the relationship between time and some event. Survival Analysis was originally mainly

used and developed in biological research (i.e., clinical, pharmaceutical), but has been used in various industrial domains such as IT and business administration. A time-to-event analysis is a type of modified regression task, but it's unique because a portion of the data is incomplete (censored) [25]. In the IT field, it has been mainly used for customer churn analysis (time to membership cancellation), machine failure, etc. [25], [26], [27], [28], [29], [30], [31], [32], [33], [34].

First of all, until recently, in addition to the binary classification method for predicting whether to leave, the studies on predicting the time to leave using Survival Analysis techniques have been conducted focusing on the insurance, finance, and gaming fields. In [25], the authors estimated the average survival period for a claim to occur and to be settled in the automobile insurance company by applying survival analysis techniques in order to secure sufficient reserves for insurance claims. They statistically compared the Kaplan Meier survival plots of various covariate groups and the time it takes for a particular vehicle to incur a loss after the majority of the insured risk occurred, and tested it using cox-regression. In a study on the financial sector with Greek bank data, the determinants of the increase in churn rate were analyzed using the risk proportionality model and survival analysis [26]. The study in [27] aimed to investigate the issue of supply overhang of affordable homes and financial exclusion in the Malaysian housing market. By employing survival analysis via Kaplan-Meier survival estimates for the period covering 2009 to 2014, they discovered that higher inflation rate and lower house price volatility may reduce the likelihood for home loans exclusion and thus allow banks to allocate higher loan disbursements. There are also studies that apply Survival Analysis techniques to predict the timing of customer churn in the gaming field [28], [29]. Recently, there have been research activities that hold an international competition on game data mining using commercial game log data and introduce cases of applying Survival Analysis techniques to game log data [30].

In addition to customer churn analysis, survival analysis is conceptually largely consistent with research on predicting machine failures [31]. The model proposed in [32] predicted the probability of survival for welded pipes using a tree-like accident theory and Bayesian survival analysis model. Using Bayesian, Kaplan–Meier, and Weibull curves, the authors constructed staged Bayesian distribution, which was then used to make predictions about the time-to-failure of the pipes. In the proposal of [33], the authors suggested a new approach for predicting the remaining service life of water mains by combining machine learning and survival statistics. Similarly, the authors of [34] suggested the similar approach as [33].

As mentioned earlier, survival analysis research started in the medical field and was applied to customer churn analysis and machine failure prediction in IT fields such as finance and games. Meanwhile, most studies quantifying repurchase likelihood have focused on predicting repurchase behavior in terms of what to buy in the e-commerce sector. In this study, we aim to support the establishment of customer retention strategies by applying survival analysis to predicting the timing of repurchase. For example, survival analysis can help marketers reduce wasted marketing efforts by understanding when customers are most likely to be receptive to a marketing communications plan and when additional efforts are likely to be ineffective. Additionally, since customers who purchase home appliances generally do not buy new products frequently, predicting repurchase timing in the home appliance sales area requires a different strategy from that of e-commerce retail customers. Therefore, rather than analyzing repurchase behavior of daily necessities based on periodicity, this study comprehensively analyzed customer integrated data from the company's CRM, ranging from customer demographics to purchase history and use history of repair and counseling services. And we applied the analysis results to labeling of status and time for the survival ensemble model and the creation of features, which are factors that affect repurchase.

III. SURVIVAL ANALYSIS-BASED PREDICTION

A more sophisticated research direction is to apply a regression model that predicts the user's repurchase point instead of a classifier that predicts whether or not to repurchase. However, there is a problem of not being able to accurately label the life expectancy for the training data set because there is a censoring problem that indicates that observations do not include complete information about the occurrence of the event of interest. For a certain number of customers, that means we do not know the time of repurchase experience because they have not repurchased it yet. To resolve this problem, we used survival analysis that assimilates censored data in studying the time until an event of interest happens and its relationship with various factors. Originally in medical field, an event refers to a case in which a patient fails or dies, however in our case it is the moment when a customer repurchases TV.

A. SURVIVAL ANALYSIS

Survival Analysis is a statistical analysis and prediction technique based on Kaplan-Meier estimation, a non-parametric method of estimating the survival function by considering the probability of an event occurring along with the variable time [25]. It is piecewise constant and can be thought of as an empirical survival function for censored data. For example, if 20% of the 1 billion new customers who signed up so far have shown a tendency to churn within a month, you can simply predict that 20% of the 10 million people who signed up today will churn within a month. However, this prediction result does not take any feature variable of each target into account. Therefore, prediction performance is generally improved by applying semi-parametric or parametric methods that consider the characteristics of the survival time distribution and the influence of various features on the prediction results.

Survival Analysis is based on the probability that an event of interest has not occurred at time t, and a survival function over time S(t) is usually used to represent that probability. As shown in equation (1), S(t) is the probability of survival after time t, and T is the random life expectancy taken from the population. S(t) is between 0 and 1, and is a decreasing function of t.

$$\mathrm{S}\,(\mathrm{t}) = P(T > t) \tag{1}$$

The hazard function is defined as the possibility that a subject will experience an event of interest within a small time interval if the individual has survived until the start of that period. It is rather an instantaneous rate calculated over a period of time than a probability. It can also be regarded as a risk of experiencing the event of interest at time t. The goal is to find the risk of an event and is shown in equation (2) below.

$$\lim_{\delta t \to 0} \frac{\Pr\left(t \leq T \leq t + \delta t \mid T > t\right)}{\delta t} \tag{2}$$

Concordance index (c-index) is the most commonly used accuracy index in Survival Analysis [49]. It is an indicator that does not evaluate the exact survival time of a subject, but instead compares the survival time (or risk) of several subjects relatively. In our case, it determines whether it is good at predicting the order of repurchase. Below is the equation that compares the survival time of a pair of subjects. y_i is the actual time when the event occurred, and $\hat{y}_i$ is the time predicted by the model.

$$\mathrm{c} = \Pr\left(\hat{y_1} > \hat{y_2} \mid y_1 \geq y_2\right) \tag{3}$$

Based on equation (3) above, c-index can be calculated as equation (4).

$$\hat{c} = \frac{1}{P'} \sum_{i:\delta_i=1} \sum_{j:y_j<y_j} I\left[S\left(\hat{y}_i|X_i\right) < S\left(\hat{y}_j|X_j\right)\right] \tag{4}$$

P' is the number of pairs to be evaluated, and I is a function that extracts cases where the given condition is true. In other words, among the total set of pairs of evaluation objects, the ratio of pairs that predict a greater survival function of object j, which survived longer than object i, is calculated, and this is between 0 and 1. Here, the condition of $\hat{y}_i$, which means that an event must occur for the target, indicates that the censored i is excluded from the comparison due to lack of certainty that the target j survived longer.

B. SURVIVAL PREDICTION TECHNIQUES

Survival prediction-related techniques include non-parametric methods, semi-parametric methods, and machine learning-based methods [35], [36], [37].

1) NON-PARAMETRIC METHODS

The Kaplan-Meier estimator is used to estimate the survival function, which measures the proportion of subjects surviving for a specific survival time t. This function represents the probability of an event in a specific time interval (e.g. survival) through a Kaplan-Meier curve.

$$\hat{S}(\mathrm{t}) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{5}$$

n_i represents the number of subjects at risk before time t, and d_i represents the number of events of interest at time t. Non-parametric methods do not use features and survival time distribution information. This is useful when distribution information is unknown, but predictions may be inaccurate.

2) SEMI-PARAMETRIC METHODS

Cox Proportional Hazards Model was introduced by Cox and considers the influence of several variables at once and explores the relationship of the survival distribution to these variables. It is similar to multiple regression analysis, but the difference is that the dependent variable at a given time t is a hazard function. It is based on a very small intervals of time containing at most one event of interest and is a semi-parametric approach for estimating weights in a proportional hazards model. The equation for the Cox proportional hazards regression model is as follows:

$$\mathrm{h}\,(t|x) = b_0\,(t) \exp \sum_{i=1}^{n} b_i\,(x_i) \tag{6}$$

Here, t represents survival time and risk may vary over time. $h(t)$ is a hazard function determined by a set of n covariates. $b_0\,(t)$ is the baseline risk function and is defined as the probability of experiencing the event of interest when all other covariates are zeros. $\exp \sum_{i=1}^{n} b_i(x_i)$ is a partial risk, a time-invariant scalar factor that increases or decreases only the baseline risk. These semi-parametric methods such as Cox Proportional Hazard, utilize feature information but do not use survival time distribution information and assume a fixed relationship between the output and the variables. It has difficulties to scale with big data problems, and alternative regularized versions of Cox regression [38] have been proposed to tackle this. Nevertheless, they are still based on restrictive assumptions that are not easy to fulfill. Thus, parametric approaches, such as the accelerated failure time models [39], assume the existence of a survival time distribution (e.g. Weibull, lognormal, exponential) and predict survival time using a regression model.

3) MACHINE LEARNING-BASED METHODS

There is a methodology that addresses the shortcomings of the above-mentioned methods by applying various machine learning algorithms to survival analysis based on censored data. One of the most famous and widely used machine learning algorithms is the SVM algorithm. As an extension to the standard support vector machine (SVM), the survival SVM separates classes based on linear or non-linear relationships between our features and survival [40]. Then, there are non-parametric machine learning techniques such as

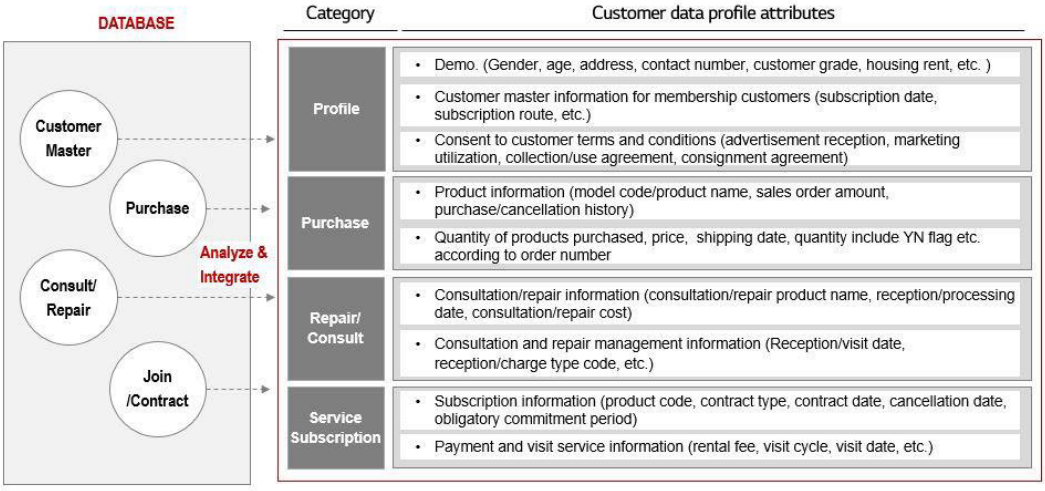

FIGURE 1. Customer profile data set in CRM.

classification and regression trees. The first survival tree was presented in [41], where a Kaplan-Meier estimator survival function was computed at every node.

Despite a powerfulness which is able to model censored data, using a single tree can produce instability in its predictions. Survival forests are ensemble-based learning methods where the underlying algorithm is a kind of survival tree. The two main survival ensemble techniques are random survival forest and gradient boosting survival analysis [42], [43]. Recently, there have been survival analysis approaches using deep neural networks such as continuous-time model (DeepSurv) [44] and discrete-time model (DeepHit) [45].

IV. MODELING

We introduce our data set, labeling for survival analysis, and feature engineering in this section.

A. CRM DATA SET

In order to understand customers and enable various target marketing, our CRM collected and analyzed customer interaction data from all channels of customer contact and organized it into one integrated customer profile. Our CRM has the infrastructure configuration that collects and pseudonymizes all identification data in AWS and transmits all pseudonymized data to a GCP-based customer data analysis platform. Because customer identification information was pseudonymized due to privacy issues in the data analysis platform, attributes such as customer age could not be used. Figure 1 shows a customer profile data set including data on customer characteristics, purchases, repair/counseling, and rental care services. For this study, we extracted target customers with the goal of predicting the period it takes for customers who had purchased a TV at least once before to repurchase, that is, how soon they would repurchase after their first purchase.

B. CUSTOMER RATIO ANALYSIS ACCORDING TO TV REPURCHASE PERIOD

From the CRM data set, we extracted approximately 85,588 customers who purchased a 2nd TV from B2C customers with a TV purchase history from 2016 to 2021. First, we checked how long it took for these customers to purchase their second TV. It was confirmed that the percentage of customers repurchasing within 1 month was approximately 27%, the one within 2 months was 31%, the one within 3 months was 34%, and the one within 12 months was 52%. Figure 2 shows a histogram of data on the time taken to purchase a 2^{nd} TV.

A surprising and interesting result here is that among people who held two TVs, about 27% of them repurchased within one month after purchasing the first TV, and about 50% of them repurchased within one year. Therefore, for target marketing for customers who are likely to purchase 2 or more TVs, modeling to predict the timing of 2nd TV purchase will be very effective. In other words, efficient target marketing will be possible according to the 2nd TV prediction time for each customer by using the prediction model as follows: promotional marketing within 1 to 2 months for about 30% of customers, promotional marketing based on the number of months within 1 year for 50% of customers, etc.

C. LABELING FOR SURVIVAL ANALYSIS PREDICTION

In the CRM data set, we performed labeling of training data for survival analysis based on whether customers with TV purchase history from 2016 to 2020 purchased a 2^{nd} TV in 2021. As described in the section III SURVIVAL ANALYSIS-BASED PREDICTION above, repurchase times is subject to right-censoring, therefore, we need to consider a customer's status in addition to repurchase times. Generally, *status* and *survival_in_days* need to be extracted with the first field indicating whether the actual survival time was

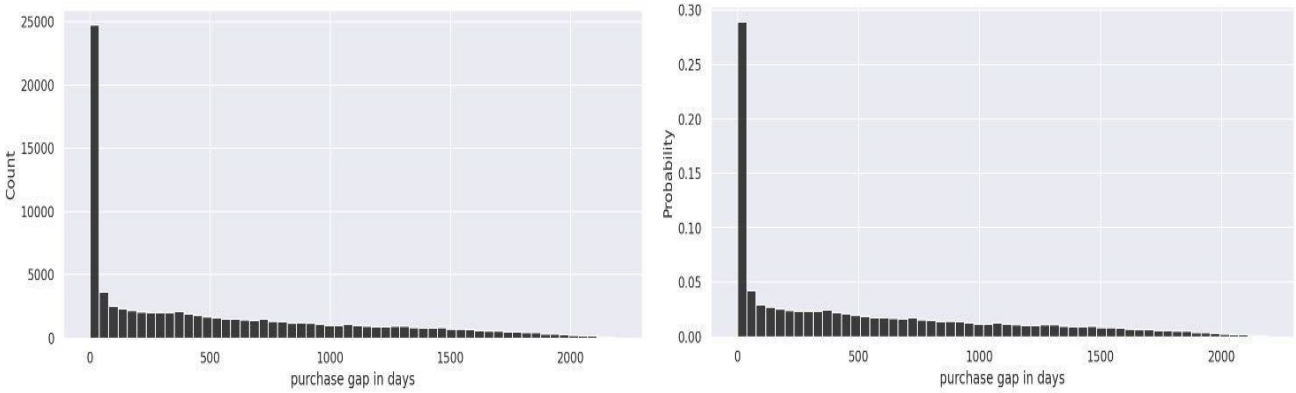

FIGURE 2. Histogram of time taken to purchase 2nd TV.

observed or censored, and the second field denoting either the observed survival time which corresponds to the time of death (if Status = True) or the last time that the person was contacted (if Status = False). Based on the purchase history of the past 5 years, our model learns the differences of various behavioral information between customers who, within 1 year since then, made a 2nd TV purchase and those who did not. For labeling the training data of this prediction model, we set ''current date'' to Jan. 1, 2021 and ''maximum date'' to Dec. 31, 2021. If there is a purchase after ''current date'', 'Status' is set to True, and the 'duration' field is set to the number of days since the previous purchase date. And if there is no purchase after ''current date'', 'Status' is set to False, and the 'duration' field is set to the number of days since the customer's last purchase date to ''maximum date'' of Dec. 31, 2021.

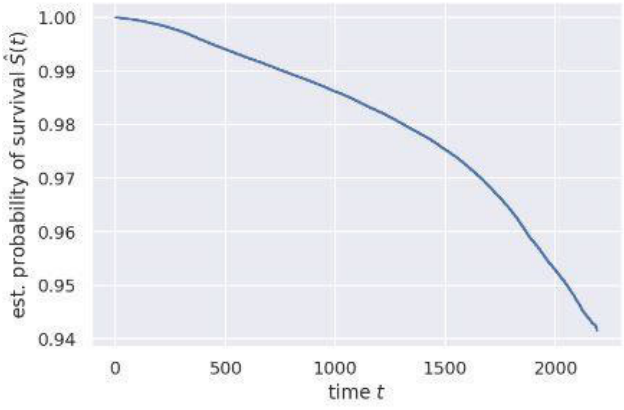

FIGURE 3. Kaplan-Meier estimation of the labeling data set.

As a result of labeling, 96% of a total of 1,452,316 people did not purchase a 2nd TV, and about 4% did. Figure 3 shows the Kaplan-Meier estimation, a non-parametric method for estimating the survival function of our labeled data set. As time goes by, we can see that the number of repeat buyers roughly increases and the survival function value decreases, but further analysis is needed to make more accurate predictions using the features of the data.

D. FEATURE ENGINEERING

Our CRM consists of an integrated customer profile that is analyzed based on customer interaction data collected for various target marketing purposes. The main attributes of the CRM data set for features included information on user demographics, purchase history, repair history, and counsel history. Then, we investigated some smart TV watching-dependent attributes such as viewing time, channels, connected devices, contents, service usage history, etc. However, we ultimately did not include them as the features in our model. The reason is that, unfortunately, there were only a limited number of TV products released equipped with this logging module, and since it has been released for less than a year, there was not enough logging data. In follow-up research, we plan to analyze smart TV viewing logs and CRM customer integration profiles in combination, which will improve the model's predictive power.

To create the first feature set, we analyzed the data related to past purchase behavior of 1,452,316 labeled customers from 2016 to 2020. For preprocessing, we looked through data containing the attribute ''QTY_INCL_YN'', which indicates whether it corresponds to the TV body. Most of the data with NULL value was removed as it was confirmed to be the products not for sale such as employee free gifts, remote controls, HDMI connection cables, etc. We also reviewed data in which SELL_AMT, an attribute corresponding to the purchase amount, had a negative value. It was confirmed to be a refund case and a pre-processing logic was applied to exclude cases where a refund was made from the purchase history. After preprocessing, RFM features directly related to ''purchase'' were created: Frequency (purchase frequency), Monetary (total purchase amount), Days_since_last_purchase (elapsed time from the last purchase date). To create the second feature set, we analyzed customer characteristic data and created the related features. Table 1 shows a list of feature sets related

to customer purchase history and customer characteristic information.

TABLE 1. The feature set related to customer characteristics and purchase history.

Attribute	Description
IDAP_ID	Unified Customer ID
FREQ	frequency in RFM
SELL_TOT	monetary in RFM
DAYS_SINCE_LAST_PURCHASE	recency in RFM
MARR_YN	Marital status
GNDR_CD	Gender
BAS_ADDR	Customer address
ZPCD	Zip code
APRT_FLSP	Apartment size
APRT_PMNC_LEAS_YN	Apartment permanent lease type
APRT_DELG_PRC	Apartment sales price
LGE_MSHP_JOIN_PATH_NM	LG Electronics membership subscription path name
TH01_SMIL_CUST_TP_NM	SMILE customer type
SMIL_SGNT_CUST_YN	SMILE signature customer Y/N
NCRM_CUST_GRD_NM	NEWCRM customer grade
ACCR_POSS_YN	Whether you have an affiliate card or not
PTNS_JOIN_YN	Affiliate registration status

The third feature set was created using the attributes related to customer counseling history. For the history table that exists as 1: N by customer id, an analysis table was created by generating a derived variable to which the summary logic of the corresponding attribute was applied. We applied the ''derived frequency variable'' creation logic to these attributes as follows. Unique categorical values for each original attribute are created as derived frequency variables, and the derived frequency variables are summed for each customer. Table 2 shows a list of derived frequency variables created through analysis of counseling history data. Next, as the other feature of the counseling history, a variable was developed to specify COUNSEL types for each customer. The counseling history-based attributes consist of the following three hierarchical levels:

- 1st level (CONS_TP_LARG_CLSS_CD): Product counseling, service inquiry, simple inquiry
- 2nd level (CONS_TP_MIDD_CLSS_CD): How to use, other inquiries, repair related, reception related, care solution inquiry, delivery/installation, pre-purchase inquiry, center inquiry, parts reservation, etc.
- 3rd level (CONS_TP_SMAL_CLSS_CD): Action guide, function guide, others, simple complaint, payment information change, receipt confirmation, specifications/functions, location, delivery date, reservation request, etc.

The types of counseling history can be distinguished by the combination of each of the above levels. For example, ('Product Counseling, 'How to Use', 'Action Guide') is one possible combination. The first derived variable as shown below was created by combining the variables made up of these hierarchical classes, and the second derived variable was created by summing them for each customer. Table 3 shows the additional feature set related to counseling history.

For the fourth feature set, features related to product repair history were created. We generated the derived frequency variables of repair history attributes in the same way as the counseling history-related feature creation method. Table 4 shows the list of resulting features.

We merged customer characteristics, purchase history, counseling history, and repair history features to create a final set of 122 features.

V. MODEL VALIDATION

This section discusses evaluation objectives and scope, and analysis results.

A. EXPERIMENTAL DESIGN

Using 100,000 selected randomly from 1,452,316 customers, we tested several different survival analysis algorithms based on an ensemble and verified their performance. Specifically, we constructed several algorithm sets according to methodology of survival analysis, method of constructing the feature set, and method of imputation technique for missing values in the feature set. Then, we built our model by dividing the training data and test data of the sample data at a ratio of 7 to 3, respectively.

1) PREDICTION MODEL

a: ALGORITHM SET

We selected several representative algorithms among machine learning algorithms for survival analysis and applied them to our prediction model. We basically tested Cox Proportional Hazard model as a semi-parametric method of survival analysis [35]. And we tested linear survival support vector machine as a machine learning method and then random survival forests and gradient boosting survival analysis as a survival ensembles method. Additionally, survival deep neural networks were tested with a continuous-time model (DeepSurv) [44] and a discrete-time model (DeepHit) [45].

b: FEATURE SET

The feature sets consisted of two groups for comparative analysis. Among the total 122 features, the first feature set (feature_set_1) was constructed by excluding attributes that meet the following two conditions: (1) the attributes where the number of unique values for a categorical variable is just 1 (2) the attributes in which the attribute's variance is 0 in either ''Repurchase '' group or ''Not-Repurchase'' group.

The criterion for the first feature set selection is related to known problems associated with convergence of Cox proportional hazards models. Since the estimation of the coefficients in the Cox proportional hazard model is done using the Newton-Raphson algorithm, there are sometimes problems with convergence. If attributes have very low variance depending on whether a ''repurchase'' event exists or not, this may harm convergence. That is because the very low variance means that the attribute completely determines

TABLE 2. The feature set related to counseling history.

Original attribute	Derived frequency variable
CONS_TP_LARG_CLSS_DESC (Counseling type major classification code)	Product counseling, end of call, service inquiry, purchase inquiry, hygiene/health, simple inquiry, care solution inquiry, service reception
PROD_GRP_NM (PRODUCT_GROUP_CODE)	Healthcare, cooking appliances, air conditioners, audio, smart ThinQ, vacuum cleaners, other products, robots, communication devices, office equipment, PCs, signage, storage devices, heaters, telephones, washing machines, TVs, small home appliances, LED lighting, refrigerators, Home-net, PLS lighting, monitor
CONS_RCP_CHNL_NM (Counseling reception channel code)	Call center, homepage (smartphone connection), PC, Chatbot (PC), homepage (PC connection), Chatbot (Mobile), Chatbot (Kakao), homepage (QR connection)
CONS_ST_NM (Processing status)	Completed, Incomplete, Other, Reminder, Repair Request, Recall
RNTL_CONS_YN (Rental counseling Y/N)	RNTL_CONS_YN_Y
MDIA_QT_ISUE_YN (Whether there is a media quality issue or not)	MDIA_QT_ISUE_YN_Y

TABLE 3. The additional feature set related to counseling history.

Original attribute	Derived frequency variable
cons_type_01	('Simple inquiry', 'Other inquiry', 'Other')
cons_type_02	('Simple inquiry', 'Center inquiry', 'Location')
cons_type_03	('Service Inquiry', 'Delivery/Installation', 'Delivery Date')
cons_type_04	('Service Inquiry', 'Repair Related', 'Simple Complaint')
cons_type_05	('Service inquiry', 'Reception related', 'Reception confirmation')
cons_type_06	('Product counseling', 'Inquiries before purchase', 'Specifications/functions')
cons_type_07	('Product counseling ', 'Parts reservation', 'Reservation request')
cons_type_08	('Product counseling', 'How to Use', 'Function Guide')
cons_type_09	('Product counseling', 'How to Use', 'Guidance on Action Methods')
cons_type_10	('Product counseling', 'Care solution inquiry', 'Change payment information')

TABLE 4. The feature set related to repair history.

Original attribute	Derived frequency variable
REPA_SDAY_REPA_YN (Same day processing Y/N)	REPA_SDAY_REPA_YN_Y
REPA_RCP_CHNL_NM (Reception location code (a: CIC, b: center, c: agency, d: web))	Call center, homepage (PC connection), IVR unmanned, center, visible ARS, Chatbot (Mobile), homepage (smartphone connection), Chatbot (PC), Chatbot (Kakao), homepage (QR connection), ThinQ PCC
REPA_RCP_TP_NM (Repair application type name)	General case, pre/post-inspection, circuit case/pre-inspection, small accessories, active case, dealer return request, dealer unsold, dealer take-in, general case/flood damage, caretaker, delivery of optional items, circuit case-general, circuit case/ Flood damage, B2B maintenance, B2B regular inspection, tour case/sales event
REPA_BAD_TP_NM (Repair defect type code)	Product defects, environmental problems, customer negligence, inexperience in use, emotional complaints, installation problems, distribution problems
REPA_DECI_KND_NM (Repair Confirmed Type)	Free B, paid B, free A, paid A, C (including agency fee)
REPA_SVC_TP_NM (Service type code)	Business trip, internal, B2B business trip, agency (internal), agency (business trip)
REPA_VST_TP_NM (Repair visit type code)	Appointment within business hours, visit at convenient time, request outside the office on the same day, appointment outside of business hours
REPA_TP_NM (Repair type code)	General repairs, heavy repairs, specification repairs, mobile terminal simple repairs, manufacturing division processing, minor processing, new/relocation product installation, mobile phone events, natural disasters/lightning, third party products, mobile terminal board repair, mobile phone product review (within 14 days)
REPA_GRD_NM (Repair grade code)	explanation processing/no visit, product refund, parts out of stock, customer postponement, parts arrival, parts not present, delayed repair, transfer to sales/logistics/installation, exchange/refund processing, non-delivery, customer absence, transfer for heavy repair, out of stock management, Center bring-in, business division request/support request, parts requirement, product exchange, adjustment repair, explanation processing/visit, repair failure/visit, repair failure/non-visit

whether a person repurchases or not. The second feature set (feature_set_2) was composed of the features selected from feature_set_1 only for variables with a small p-value (<0.05) through univariate Cox Regression fitting. Table 5 shows a list of the top 20 features selected through Cox Regression fitting.

c: IMPUTATION METHODS

Datasets in the analysis tables may contain values that are often missing due to data corruption or failure to record. Various imputation techniques have been applied to solve this problem of missing data. There are two main types of imputation techniques: single imputation and multiple

TABLE 5. Top 20 features with p-values from Cox regression.

Feature list	p-value
NCRM_CUST_GRD_NM_Leaders	<0.001
NCRM_CUST_GRD_NM_C	<0.001
freq	<0.001
sell_tot	<0.001
days_since_last_purchase	<0.001
NCRM_CUST_GRD_NM_A	<0.001
TH01_SMIL_CUST_TP_NM_General customer	<0.001
SMIL_SGNT_CUST_YN_N	<0.001
SMIL_SGNT_CUST_YN_Y	<0.001
NCRM_CUST_GRD_NM_B	<0.001
REPA_TP_NM_General case	<0.001
APRT_DELG_PRC	<0.001
REPA_RCP_TP_NM_General case	<0.001
REPA_SVC_TP_NM_Business trip	<0.001
REPA_VST_TP_NM_Appointent within business hours	<0.001
REPA_RCP_CHNL_NM_Call Center	<0.001
APRT_PMNC_LEAS_YN_N	<0.001
LGE_MSHP_JOIN_PATH_NM_BEST	<0.001
REPA_SDAY_REPA_YN_Y	<0.001
REPA_GRD_NM_Explanation Processing-Visit	<0.001

imputation [46]. The single imputation approach estimates missing values in the data only once. On the other hand, the multiple imputation approach creates multiple data sets, each containing approximations/estimates of missing values, and integrates the results of all imputations in the final step to generate the inferred values of missing values. Single imputation approaches can be broadly classified as follows [47]: (1) univariate single imputation approaches; (2) Multivariate single imputation approaches, such as k-Nearest Neighbors (KNN) and Random Forests (RF)-based imputation. The univariate imputation approach uses observations from the same column to impute missing values in a column, while the multivariate imputation approach uses observations from the other columns of the data to estimate missing values in a column. MICE is a commonly used multiple imputation approach to generate imputations based on a set of imputation models for each variable with missing values [48]. In this study, we compare the univariate single imputation approach and the MICE method, a representative multiple imputation method, by applying them to the preprocessing of the feature set.

2) PERFORMANCE METRIC

For performance measurement, c-index, the most commonly used accuracy index in survival analysis, was used [49]. It is a method that does not evaluate the exact survival time of a subject, but instead relatively compares the survival time (or risk) of several subjects. C-index is an indicator that verifies the superiority of the relative risk ranking of survival analysis that can be compared with AUC, which measures whether stable predictions can be made to distinguish labels while being less sensitive to decision boundaries in general classification models [50].

The process for calculating c-index is as follows. We look at all possible customer pairs in the test data set. If one of the two customers experienced an event (e.g. repurchase) sooner, we check whether the model assigned a higher risk to that customer. We repeat this for all customer pairs and calculate the proportion of correct predictions made by the model. For example, a C-index of 0.8 means that the model correctly predicted who would experience an event sooner for 80% of customer pairs. In other words, it is an indicator that determines whether the order of occurrence of events of interest is well predicted, and it means rank correlation with the predicted risk score. If the perfect prediction is 1, the random guess is 0.5.

Concordance intuitively means that two samples were ordered correctly by the model. More specifically, two samples are concordant, if the one with a higher estimated risk score has a shorter actual survival time. Based on the model's prediction results, we would like to support a repurchase promotion campaign as follows. First, for n current CRM customers, marketers obtain the predicted risk

TABLE 6. Experimental designs of predictive models.

Prediction models	Imputation techniques	Feature sets	Survival Analysis algorithms
Base-line Algo.	Simple Imputation	Feature set 1	Cox Proportional Hazard model
Algo-1.	Simple Imputation	Feature set 1	Fast Survival SVM Random Survival Forest Gradient Boosting Survival Analysis
Algo-2	Simple Imputation	Feature set 2 (selected features)	Fast Survival SVM Random Survival Forest Gradient Boosting Survival Analysis
Algo-3	Multiple Imputation (MICE)	Feature set 2 (selected features)	Random Survival Forest Gradient Boosting Survival Analysis
Algo-4-1.	Simple Imputation	Feature set 1	DeepHit - 2 layers 32 Nodes (2 multi-layer perceptrons, each consisting of 32 nodes)
Algo-4-2.	Simple Imputation	Feature set 1	DeepHit - 3 layers 64 Nodes
Algo-5-1.	Multiple Imputation (MICE)	Feature set 1	DeepHit - 2 layers 32 Nodes
Algo-5-2.	Multiple Imputation (MICE)	Feature set 1	DeepHit - 3 layers 64 Nodes
Algo-6-1	Multiple Imputation (MICE)	Feature set 2 (selected features)	DeepHit - 2 layers 32 Nodes
Algo-6-2	Multiple Imputation (MICE)	Feature set 2 (selected features)	DeepHit - 3 layers 64 Nodes
Algo-7-1	Multiple Imputation (MICE)	Feature set 2 (selected features)	DeepSurv - 2 layers 32 Nodes
Algo-7-2	Multiple Imputation (MICE)	Feature set 2 (selected features)	DeepSurv - 3 layers 64 Nodes

scores for repurchase for each customer, sort them from 1st to nth, and execute promotional activities for the top *m* customers. Next, by utilizing a survival function-based repurchase probability for each customer's time (months), they can carry out promotional activities for the *m* customers with a high repurchase probability after the desired target month (e.g. 3 months).

B. EXPERIMENTAL RESULTS

The proposed prediction methods were evaluated with cross-validation (10-fold) and all combinations of each algorithm-feature set-imputation are shown in table 6.

1) COX FITTING AND ML LEARNING WITH SINGLE IMPUTATION

This section describes the experimental results of the prediction model through cox fitting and ML learning with single imputation. For preprocessing of imputation, a univariate single imputation approach was used. Depending on the combination of feature set and ML algorithm, we constructed and compared three prediction models, Base-line Algo, Algo-1, and Algo-2. Table 7 shows the design details of the prediction models.

TABLE 7. The design details of the prediction model.

Prediction models	Feature set	Survival Analysis method	Survival Analysis algorithms
Base-line Algo.	feature set 1	Semi-parametric methods	Cox Proportional Hazard model (CPH)
Algo-1.	feature set 1	Machine learning methods	Fast Survival SVM (FSSVM) Random Survival Forest (RSF) Gradient Boosting Survival Analysis (GBSA)
Algo-2	feature set 2	Machine learning methods	Fast Survival SVM (FSSVM) Random Survival Forest (RSF) Gradient Boosting Survival Analysis (GBSA)

- feature set 1: Features that excludes the following two cases: (1) the attributes where the number of unique values for a categorical variable is just 1 (2) the attributes in which the attribute's variance is 0 in either "Repurchase " group or "Not-Repurchase" group.
- feature set 2: Features selected through cox-regression (p-value 0.05 or less)

Table 8 shows the prediction performance comparison results of 10-fold cross validation of Baseline, Algo1 and Algo2. As shown in the table, the Cox Proportional Hazard model as a baseline model showed the lowest performance with a c-index of approximately 0.58. In the case of machine learning methods, we compared the performance of the same ML model in Algo-1 and Algo-2. This aims to investigate the impact of the feature set selected based on the p-value of cox-regression on the performance of the prediction model. Only Random Survival Forest was confirmed to have better performance in Algo-2 than Algo-1 with statistical significance (t-test statistic = -2.94, p-value = 0.01). As a result of comparing the performance of each ML algorithm regardless of the feature set, SVM showed the lowest performance, followed by Gradient Boosting Survival Analysis and Random Survival Forest.

TABLE 8. Comparisons of predictive performance of the models I.

Prediction models	Survival Analysis algorithms	Feature set	c-index
Base-line Algo.	CPH	Feature set 1	0.586
Algo-1.	FSSVM	Feature set 1	0.619
	RSF		0.769
	GBSA		0.752
Algo-2	FSSVM	Feature set 2	0.619
	RSF		0.787
	GBSA		0.754

The results of checking the statistical significance of these performances are shown in Table 9. One of the limitations of the survival SVM is the inability to compare it with the Random Survival Forest or Gradient Boosting Survival Analysis in details. This is due to the lack of ''standard'' metrics for time-to-event analyses, such as the survival function and cumulative hazard function. This made our comparison restrict to the c-index score.

TABLE 9. The design details of the prediction model.

Comparison target	Algorithms	Statistical verification
ML algorithms in Algo-1	FSSVM vs. RSF	statistic=-23.86, pvalue<0.001
	RSF vs. GBSA	statistic=3.40, pvalue=0.007
ML algorithms in Algo-2	FSSVM vs. RSF	statistic=-40.32, pvalue<0.001
	RSF vs. GBSA	statistic=6.49, pvalue<0.001
Prediction Model (Algo-1 vs Algo-2)	FSSVM in Algo-1 vs. FSSVM in Algo-2	statistic=-1.14, pvalue=0.28
	RSF in Algo-1 vs. RSF in Algo-2	statistic=-2.94, pvalue=0.01
	GBSA in Algo-1 vs. GBSA in Algo-2	statistic=-1.17, pvalue=0.27

2) ML LEARNING WITH MI IMPUTATION

Here, we describe the experimental results of applying feature sets using different imputation methods to each ML model. The MICE method, one of the multiple imputation techniques mentioned above, was applied. The MICE imputation is performed M times (m = 5 here) based on tree-based ML. Considering memory and speed issues, we resampled only 50,000 out of 100,000 customers to efficiently verify the methodology. Feature set 2 obtained through previous cox-regression (p-value 0.05 or less) was used as the feature set. For comparative analysis, the MICE imputation technique was applied to two tree-based ensemble algorithms, models that showed relatively excellent performances in the previous

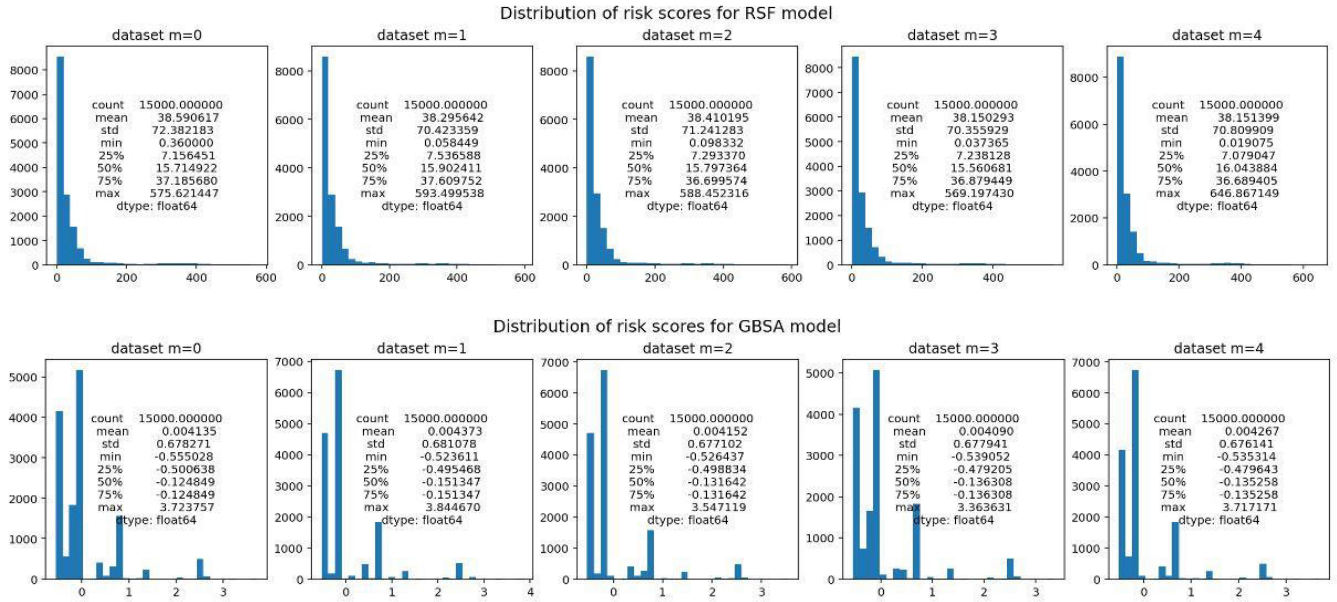

FIGURE 4. Distribution of predicted risk scores (RSF & BGSA model).

experiment results. We defined the combination of MICE and feature set 2 as Algo3. Table 10 shows the null value proportion of this feature set.

TABLE 10. Null value proportion of feature set 2.

Features	Null (%)
freq	2.844
sell_tot	2.844
days_since_last_purchase	2.844
MARR_YN	0.056
GNDR_CD	18.082
BAS_ADDR	0.718
APRT_FLSP	30.468
APRT_PMNC_LEAS_YN	38.102
APRT_DELG_PRC	30.468
LGE_MSHP_JOIN_PATH_NM	1.306
TH01_SMIL_CUST_TP_NM	27.8
SMIL_SGNT_CUST_YN	27.774
NCRM_CUST_GRD_NM	0.06
ACCR_POSS_YN	0.06
PTNS_JOIN_YN	0.06

a: POOLED RISK SCORE ESTIMATES AND CONCORDANCE INDEX

We conducted performance verification using pooled risk score estimates and concordance index. In the case of Algo-3 (feature set 2 and MI imputation), MI imputation is performed on a data set that is randomly divided into 5 data sets. First, a prediction model is created for each imputed data set (m = 1∼5). Then, a risk score estimate is obtained for each generated prediction model. Finally, we derived the final pooled c-index by pooling the risk score estimates. Figure 4 shows the distribution of predicted risk scores for each imputation data set of Algo3's Random Survival Forest and Gradient Boosting Survival Analysis.

Table 11 shows the prediction performance of the Prediction Model of Algo-3 (feature set 2 and MI imputation). By running 10 experiments of creating 5 data sets for MI imputation, we obtained pooled c-index, which is pooled risk score estimates. Although RSF showed about 1.6% higher performance than GBSA, the difference was not statistically significant. (t-test statistic = −2.62, p-value = 0.058). Next, to investigate the impact of MI imputation on prediction performance, we compared the performance of each ML model in Algo-2. In the case of RSF, there was a performance improvement of about 3.6% (t-test statistic = −3.36, p-value = 0.02) compared to the result of Algo2 (0.787) using a single imputation method.

Also, in the case of GBSA, a performance improvement of approximately 5.3% (t-test statistic = −9.97, p-value <0.001) was confirmed compared to the result of Algo2 (0.754). In other words, the performance difference between Ensemble algorithms within the same imputation method was not large, but it was confirmed that the MI imputation showed superior performance compared to the SI imputation.

b: CUMULATIVE DYNAMIC AUC METRIC

The c-index provides us with information for the whole model, but it's also useful to examine how well the model performed at various time points. For this, we used the cumulative dynamic AUC metric and visualized it [50]. We derived the pooled mean score of the AUC scores 'up to M days at an interval of N-days'. As shown in Table 12, in the case of RSF, the AUC mean score 'up to 500 days at an interval of 30-days' is high, and in the case of GBSA, the AUC mean score 'up to 180 days (6 months) at an interval of one-week' is high.

TABLE 11. Comparisons of predictive performance of the models (feature set 2 + MI imputation -Algo3).

Prediction Model	Pooled Risk Score Estimates (Pooled c-index)
RSF	0.823 (approximately 3.6% improvement compared to 0.787 for Algo2 with single imputation result)
GBSA	0.807 (approximately 5.3% improvement compared to 0.754 for Algo2 with single imputation result)

TABLE 12. Pooled AUC mean score.

Comparison target	Algorithms	Mean score
Up to 2300 days at an interval of 90-days	RSF	0.839
Up to 500 days at an interval of 30-days	RSF	0.874
Up to 180 days at an interval of 7-days	RSF	0.836
Up to 2300 days at an interval of 90-days	GBSA	0.837
Up to 500 days at an interval of 30-days	GBSA	0.846
Up to 180 days at an interval of 7-days	GBSA	0.872

By looking at the average AUC value, it is possible to determine whether the model performed well throughout the study (minimum AUC = 0.80). However, through Figure 5 and 6 below, we observed the periods when AUC values peaked, which could also be useful in determining the timing of promotional campaigns. For example, for 'up to 2300 days at an interval of 90-days', GBSA model had a pooled mean AUC score of 0.84, which is similar to that obtained using the Random Forest model. In Figure 6, we observed that, although the model performed well throughout the study (min AUC = 0.81), it had two lower points around days 500 and 2000 where it had AUC values < 0.82.

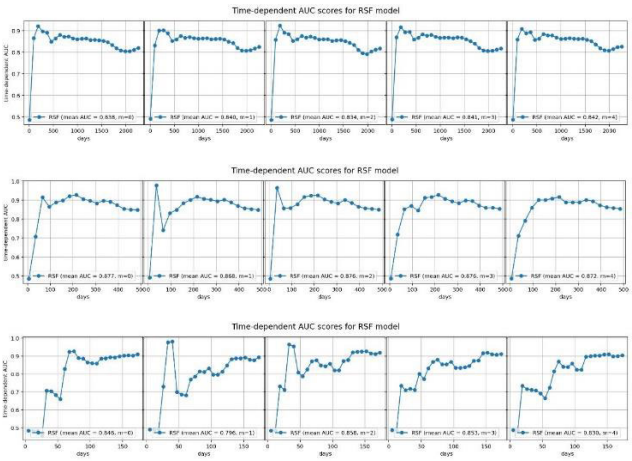

FIGURE 5. Time-dependent AUC scores for the Random Forest Survival model.

In details, we can figure out that discriminating whether survival or not became worse with time until the 500th day, with a small improvement toward around the 1500th, and finally it became worse again towards the end.

c: PERMUTATION-BASED FEATURE IMPORTANCE

Next, we looked at how various features contributed to the prediction model. Table 13 shows the top 15 results of permutation importance of the RSF model and GBSA model.

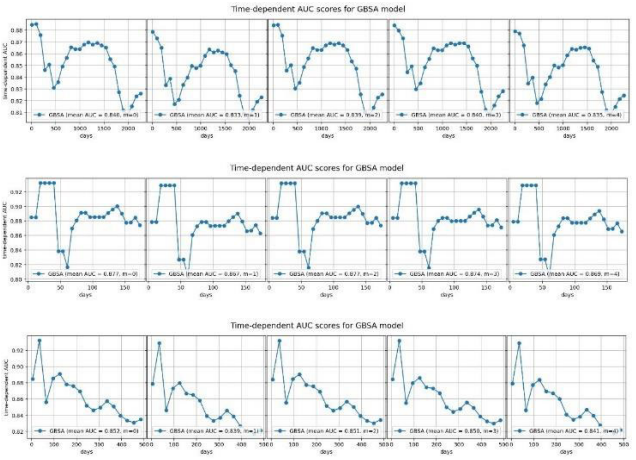

FIGURE 6. Time-dependent AUC scores for the Gradient Boosting Survival Analysis model.

As for feature importance, although the order of the top 15 features was different in the two models, the same top 4 groups of features were obtained: (1) features related to customer loyalty grade (NCRM_CUST_GRD_NM), (2) RFM features (days_since_last_purchase, sell_tot), (3) repair grade code (REPA_GRD_NM: explanation not processed/no visit, product refund, parts out of stock, etc.), (4) repair defect type code (REPA_BAD_TP_NM: product defect, emotional dissatisfaction, installation problem, etc.). Despite the differences between our models in both overall fit (c-index) and time-dependent fit (time dependent AUC), these results provide more confidence in estimating the common factors that influence a customer's repurchase probability. In other words, we confirmed that it is possible to select the superior model based on performance among those models, but there is an advantage of obtaining consistent insight in selecting important features through those models.

d: UTILITY FUNCTION

We sampled seven instances from the test data to examine repurchase estimates over time for each customer. Each instance was sampled in the following intervals: within 1 month, 1 to 3 months, 3 to 12 months, 12 to 24 months, 24 to 36 months, 36 to 48 months, and 48 to 60 months. By extracting the predict survival function and hazard function, the survival probability and cumulative risk over time are depicted in Figure 7 and Figure 8. The survival function refers to the probability that an instance survives after time t. For example, the survival probability starts at 1 and the point at which it drops is different for each instance, providing an appropriate option for each customer at what point to take promotional action.

TABLE 13. Permutation importance of RSF model and GBSA model.

RSF		GBSA	
Weight	feature	Weight	feature
0.0696 ± 0.0072	days_since_last_purchase	0.0831 ± 0.0105	days_since_last_purchase
0.0407 ± 0.0112	NCRM_CUST_GRD_NM_C	0.0357 ± 0.0082	NCRM_CUST_GRD_NM_C
0.0133 ± 0.0046	NCRM_CUST_GRD_NM_Leaders	0.0088 ± 0.0043	NCRM_CUST_GRD_NM_Leaders
0.0052 ± 0.0027	NCRM_CUST_GRD_NM_A	0.0054 ± 0.0014	sell_tot
0.0028 ± 0.0037	NCRM_CUST_GRD_NM_B	0.0011 ± 0.0017	NCRM_CUST_GRD_NM_A
0.0027 ± 0.0029	sell_tot	0 ± 0.0000	REPA_BAD_TP_NM_Customer negligence
0.0015 ± 0.0011	ACCR_POSS_YN_N	0 ± 0.0000	REPA_GRD_NM_Repair Failed
0.0011 ± 0.0019	REPA_BAD_TP_NM_Environmental Issues	0 ± 0.0000	GNDR_CD_F
0.0009 ± 0.0028	freq	0 ± 0.0000	REPA_VST_TP_NM_Visit at convenient time
0.0009 ± 0.0009	REPA_GRD_NM_Repair Failed	0 ± 0.0000	REPA_DECI_KND_NM_Free
0.0008 ± 0.0018	ACCR_POSS_YN_Y	0 ± 0.0000	GNDR_CD_*
0.0008 ± 0.0005	PTNS_JOIN_YN_Y	0 ± 0.0000	REPA_DECI_KND_NM_C (including agency fee)
0.0008 ± 0.0006	BAS_ADDR_Jeonnam	0 ± 0.0000	REPA_TP_NM_Serious Repair
0.0007 ± 0.0019	APRT_DELG_PRC	0 ± 0.0000	REPA_BAD_TP_NM_Inexperienced in use
0.0007 ± 0.0011	REPA_RCP_TP_NM_Pre/Post Inspection	0 ± 0.0000	REPA_VST_TP_NM_Request other than same day

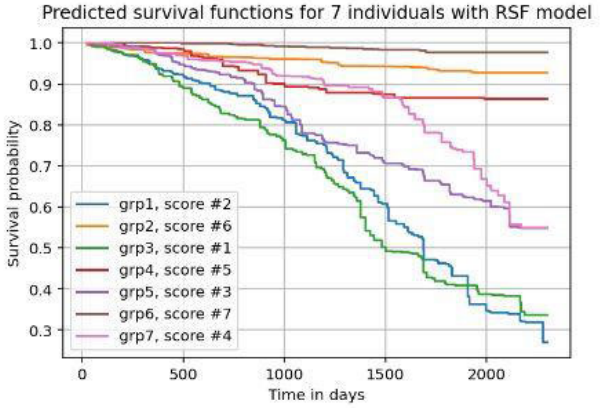

FIGURE 7. Predictive survival probability according to time in days (seven customers sampled from test data, grp1: within 1 month, grp2: 1 to 3 months, grp3: 3 to 12 months, grp4: 12 to 24 months, grp5: 24 to 36 months, grp6: 36 to 48 months, grp7: 48 to 60 months).

Conversely, a hazard function or hazard rate h(t) refers to the probability that an individual will survive until time t and experience an event of interest exactly at time t. In other words, it is possible to identify the point in time when the probability of repurchase increases rapidly for each customer and effectively introduce it into the strategy of a promotional

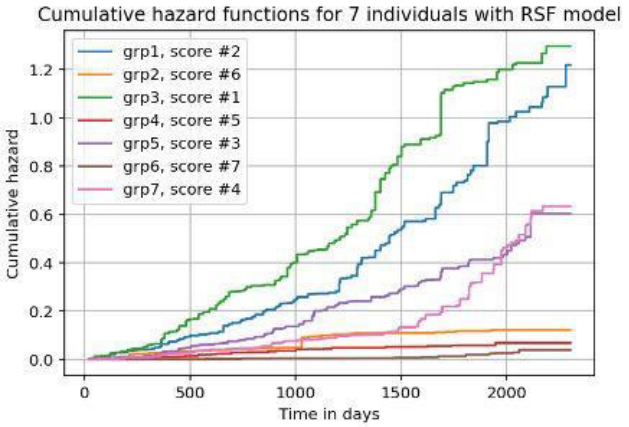

FIGURE 8. Cumulative hazards according to time in days (seven customers sampled from test data, grp1: within 1 month, grp2: 1 to 3 months, grp3: 3 to 12 months, grp4: 12 to 24 months, grp5: 24 to 36 months, grp6: 36 to 48 months, grp7: 48 to 60 months).

campaign, similar to how the survival function is used. Of course, the proportion of customers who repurchase TVs is very small, and thus logic to determine the risk rate within 6 months or 1 year with a fine-grained cut-off will be needed

We provide utility functions to improve usability when marketers establish promotional strategies based on the predicted results. First, marketers can obtain the predicted risk score for repurchase through a utility function, sort them from 1st to *n-th*, and run a repurchase promotional campaign for the top *m* people based on this result. In addition, by providing a probability of repurchase within *n* days based on the survival function, we support the execution of repurchase promotional activities for *m* customers with a high repurchase probability in the desired target month.

The first utility function provides the probability that a specific customer will not repurchase for a specific time period. The second is the cumulative risk function, which provides the cumulative repurchase risk of a specific customer up to a certain point in time (here, risk means repurchase).

Table 14 explains the definition of our utility function and its implications for use. Table 15 shows the labeled data, predicted scores, and results of utility function of the seven sampled customers mentioned above. It exemplarily shows the probability of repurchase within 30 days and cumulative risk values for the seven customers sampled above.

The first utility function works as follows. When T is the time when an event occurs and t is a random time point during observation, survival S(t) is the probability that T is greater than t. In other words, the survival function here is the probability of an individual surviving after time t (after 30 days), i.e., the probability of not repurchasing the TV. We can consider customers with the lowest probability of not repurchasing as target candidates. Here, it would be customer index 3 in Table 15.

The second utility function returns the value obtained by subtracting the predicted value based on the survival function

TABLE 14. Definition and interpretation of utility functions.

Comparison target	Algorithms	Mean score
(1) Survival (non-repurchase) probability *get_survival_prob(model, X_test, time)*	Probability that a specific customer (X_test) will NOT repurchase a TV during a specific period (time) based on a prediction model	Predicted value based on survival function
(2) Repurchase probability *get_purchase_prob(model, X_test, time)*	Probability of repurchasing a TV within a certain period of time for a specific customer (X_test) based on a prediction model	Value obtained by subtracting the predicted value based on survival function from 1
(3) Probability of experiencing an event of interest (TV repurchase) *get_cum_hazard(model, X_test, time)*	Probability of experiencing an event of interest (TV repurchase) within a specific period (time) of a specific customer (X_test) based on a prediction model	Predicted value based on hazard function

TABLE 15. The labeled data, predicted scores, and results of utility function of the seven sampled customers.

	Actual Data		Predicted Risk Scores		Utility function (Based on RSF model, e.g. within 30 days)		
index	2nd TV purchase	days	RSF	score rank	(1) Survival (non-repurchase) probability	(2) Repurchase probability	(3) Probability of experiencing an event of interest (TV repurchase)
1	TRUE	14	341.45	2	0.998363152	0.001636848	0.001645018
2	TRUE	58	42.52	6	0.994469333	0.005530667	0.005530667
3	TRUE	195	411.28	1	0.99984127	0.00015873	0.00015873
4	FALSE	629	86.55	5	1	0	0
5	FALSE	1014	224.08	3	1	0	0
6	FALSE	1200	9.70	7	1	0	0
7	TRUE	1608	124.30	4	1	0	0

from 1. Customers with the highest probability of repurchase within 30 days, that is, customers with index number 2 above, can be considered as target candidate customers.

The third utility function makes the result based on the hazard function. The hazard function, or hazard rate h(t), is the probability that an individual will survive until time t and experience the event of interest exactly at time t. In other words, it means the probability of being indifferent until 30 days and then experiencing an event of interest on the 30th. Customers with the highest probability of purchasing within 30 days, that is, those with index number 2 above, can be considered as target candidate customers. In this way, it is possible to select customers based on a specific threshold of the result of the utility function and use it for target marketing.

3) DEEP LEARNING

Finally, we tested survival analysis approaches using deep neural networks: the continuous-time model (DeepSurv) [44] and the discrete-time model (DeepHit) [45]. Since DeepHit is a discrete-time model, we need to define discrete times to evaluate. We adopted a quantile discretization in which intervals are defined by the proportion of events (repurchase). We preprocessed features with categorical embedding and performed batch normalization following each layer as well as a 20% dropout. As an optimizer, we selected the cyclic Adam (WR), which is a weight decay regularized version of the Adam optimizer. Table 16 shows the performance comparison results for each algorithm in deep learning prediction models.

According to the experimental results in Table 16, Algo6-1 (DeepHit, Feature set 2, MICE) shows the best performance (c-index 0.828) among deep learning methods. However, this is almost similar to the performance (c-index 0.823) of the machine learning method Algo3 (RSF, feature set 2, MICE). It showed meaningfulness that survival analysis, which was once seen as the domain of statistical analysts, can also be applied to deep learning, but the model we would like to recommend for use in actual target marketing is the ML-based model (Algo3). The reason lies in the interpretability of the model. Deep learning successfully classifies problems related to nonlinear decision-making, but lacks the ability to interpret, while tree-based machine learning predicts customer repurchase likelihood in an intuitive and easy-to-interpret manner based on customer service usage history. Unlike problems where superior performance is a priority, such as computer vision or signal processing problems, it is essential to understand the importance of feature variables in survival analysis problems. For example, if a customer has a frequent history of repairs due to defective products, the probability of the customer repeating a purchase is low. In other words, when performing target marketing using prediction models, the performance of the model is important, but it is more important to discover important variables that marketers can intuitively understand, utilize, and interpret, as shown in Table 13.

VI. REPURCHASE PROBABILITY-BASED TARGET PROMOTION STRATEGY

Since the proportion of customers who actually repurchase TVs is very small, we plan to target both customers at the top and bottom of the repurchase risk prediction score derived through the utility function. In particular, the target campaign

TABLE 16. Comparisons of predictive performance of the deep learning models.

Prediction models	Survival Analysis algorithms	Network configuration	Feature set	Imputation	c-index
Algo-4-1	DeepHit	2 layers 32 Nodes (2 multi-layer perceptrons, each consisting of 32 nodes)	Feature set 1	Simple Imputation	0.8106
Algo-4-2	DeepHit	3 layers 64 Nodes	Feature set 1	Simple Imputation	0.8164
Algo-5-1	DeepHit	2 layers 32 Nodes	Feature set 1	Multiple Imputation (MICE)	0.8109
Algo-5-2	DeepHit	3 layers 64 Nodes	Feature set 1	Multiple Imputation (MICE)	0.8115
Algo-6-1	DeepHit	2 layers 32 Nodes	Feature set 2 (selected features)	Multiple Imputation (MICE)	0.8281
Algo-6-2	DeepHit	3 layers 64 Nodes	Feature set 2 (selected features)	Multiple Imputation (MICE)	0.8117
Algo-7-1	DeepSurv	2 layers 32 Nodes	Feature set 2 (selected features)	Multiple Imputation (MICE)	0.7998
Algo-7-2	DeepSurv	3 layers 64 Nodes	Feature set 2 (selected features)	Multiple Imputation (MICE)	0.7894

TABLE 17. A concrete example of possible operation strategies.

Type of campaign activity	Criteria for subject extraction	Campaign (Care) activity
CRM promotion offering combined product purchases to customers with high affordability (NCRM_CUST_GRD_NM-based loyal customer)	- Common criteria: Top m people based on repurchase risk prediction score derived through utility function - Detailed criteria: Top n people based on repurchase probability derived through utility function according to target campaign date (within k days). (Target campaign cycles can be weekly or monthly, and every few weeks or months)	- Conduct an email campaign linking target customers' TV product purchase history to suggest repurchase or purchase of replacement products - Test the effectiveness of the campaign by monitoring purchase status after the campaign runs.
Follow-up management for customers with a history of specific types of repair grade code/repair defect type code	- Common criteria: Bottom m people based on repurchase risk prediction score derived through utility function - Detailed criteria: Bottom n people based on repurchase probability derived through utility function according to target campaign date (within k days). (Target campaign cycles can be weekly or monthly, and every few weeks or months)	- Divide into 2 groups of 2/n each. Group 1 provides outbound phone calls, and Group 2 sends text messages to provide emotional care support for inconveniences while using the repair service and to recommend repurchase using discount coupons. - Test the effectiveness of activities by monitoring customer behavior patterns (satisfaction surveys, purchase status) after performing care activities.

cycle (weekly or monthly, and several weeks or months, etc.) needs to be carefully set up and customer segments are created based on feature values analyzed as dominant influencing factors. Thus, the effectiveness of the model can be verified by deriving the utility function result (repurchase probability) for each customer segment and implementing differentiated strategies for top or bottom ones.

As for feature importance, although the order of the top 15 features was different in the two models, the same top 4 groups of features were obtained: (1) features related to customer loyalty grade (NCRM_CUST_GRD_NM), (2) RFM features (days_since_last_purchase, sell_tot), (3) repair grade code (REPA_GRD_NM: explanation not processed/no visit, product refund, parts out of stock, etc.), (4) repair defect type code (REPA_BAD_TP_NM: product defect, emotional dissatisfaction, installation problem, etc.).

Potentially, we can design the type of campaign activity with customer segments using feature values (1) and (2) above and customer segments using feature values (3) and (4) above. Next, subject extraction is performed based on the following criteria.

By providing a probability of repurchase within n days based on the survival function, we support the execution of repurchase promotional activities for m customers with a high repurchase probability in the desired target promotion time. It can be done both by every week or month and by every few weeks or months. Table 17 shows a concrete example of possible operation strategies for target marketing that can be executed first using the dominant features derived from our prediction model and utility functions.

VII. CONCLUSION AND FUTURE WORKS

In this study, we applied statistical and machine learning techniques based on survival analysis to predict TV repurchase and analyzed the results. Our work has academic significance in the following respects. First, we created features by comprehensively analyzing customer-company interaction data, such as customer purchase history, demographic

information, and counseling and repair history for actual CRM products in operation. In addition, we developed a survival ensemble model by applying the labeling logic that reflects the customer's purchase cycle characteristics of the home appliance business domain, rather than the statistical techniques and binary classification models of existing studies related to repurchase prediction in the e-commerce area.

Specifically, we applied survival analysis-based predictive modeling to tv repurchase prediction and verified the performance of it. Last but not least, we developed a utility function that provides useful information about what decisions to make at what point in time by estimating survival functions and risk functions. For example, marketers can group customers by target month (e.g. within 3 months after first purchase, 3 to 6 months, within 12 months, etc.) by referring to each customer's monthly survival function-based repurchase probability information. Then, for each group, a customized promotional campaign can be implemented for the m customers with a high probability of repurchase at the target time. In particular, if the predicted purchase probability is high and the period is short (e.g. within 3 months), intensive active marketing is carried out, and if the predicted probability of purchase is low and there is some distance in the future (e.g. after 1 year), passive marketing that takes more cost into consideration is conducted. They can create a cost-effective marketing strategy based on the follow-up purchase period.

A limitation of this study is that features other than CRM's customer data-related features, such as actual TV device usage history or external data, have not yet been applied.

In addition, local interpretation method which analyzes how much individual observation units contributed to the model prediction value for each feature variable, is also meaningful. Since implementation in the data-based marketing area must be based on an understanding of the basis for AI decision-making, the sensitivity analysis is very important as a way to increase the explanatory power of AI. I learned through literature survey that SurvLIME and SurvSHAP, explainable artificial intelligence models specialized in survival prediction, had been developed. Thus, the task of increasing explainability will be a meaningful future study that provides important insight in introducing a repurchase prediction model in the digital marketing area.

Furthermore, when using the utility function that provides the probability of TV repurchase within N days based on the survival function, the ratio of the number of customers who repurchase TV is very small, so the additional selection logic based on the fine-grained thresholds is required. (Especially in the case of selecting target customers based on the probability of purchase (risk rate) within 3 months). Additionally, it is necessary to apply the developed prediction model to an actual repurchase promotion marketing campaign to analyze the contribution of the prediction model to the marketing success rate. Here, what is more important than the accuracy of repurchase prediction is actually increasing service quality and sales while reducing marketing costs. Therefore, performance judgments regarding the accuracy of repurchase prediction may also vary depending on the detailed strategy of the promotional campaign to be proposed to customers. In the future, researches to optimize the thresholds of risk probability and survival period using marketing profits and costs will also be worthwhile.

REFERENCES

[1] C.-P. Wei and I.-T. Chiu, "Turning telecommunications call details to churn prediction: A data mining approach," *Exp. Syst. Appl.*, vol. 23, no. 2, pp. 103–112, Aug. 2002, doi: 10.1016/s0957-4174(02)00030-1.

[2] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," in *Proc. 8th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Islamabad, Pakistan, Sep. 2013, pp. 131–136, doi: 10.1109/ICDIM.2013.6693977.

[3] E. Ascarza, R. Iyengar, and M. Schleicher, "The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment," *J. Marketing Res.*, vol. 53, no. 1, pp. 46–60, Feb. 2016, doi: 10.1509/jmr.13.0483.

[4] M. Zhao, Q. Zeng, M. Chang, Q. Tong, and J. Su, "A prediction model of customer churn considering customer value: An empirical research of telecom industry in China," *Discrete Dyn. Nature Soc.*, vol. 2021, Aug. 2021, Art. no. 7160527.

[5] F. F. Reichheld and W. E. Sasser, "Zero defections: Quality comes to services," *Harv. Bus. Rev.*, vol. 68, no. 5, pp. 11–105, 1990.

[6] T. O. Jones and W. E. Sasse, "Why satisfied customers defect," *Harv. Bus. Rev.*, vol. 73, no. 6, pp. 88–99, 1995.

[7] M. R. Colgate and P. J. Danaher, "Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution," *J. Acad. Marketing Sci.*, vol. 28, no. 3, pp. 375–387, Jul. 2000, doi: 10.1177/0092070300283006.

[8] S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," *J. Marketing Res.*, vol. 43, no. 2, pp. 204–211, 2006, doi: 10.1509/jmkr.43.2.204.

[9] J. Ahn, J. Hwang, D. Kim, H. Choi, and S. Kang, "A survey on churn analysis in various business domains," *IEEE Access*, vol. 8, pp. 220816–220839, 2020, doi: 10.1109/ACCESS.2020.3042657.

[10] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, pp. 1–24, Mar. 2019, doi: 10.1186/s40537-019-0191-6.

[11] V. L. Miguéis, D. Van den Poel, A. S. Camanho, and J. F. E. Cunha, "Modeling partial customer churn: On the value of first product-category purchase sequences," *Exp. Syst. Appl.*, vol. 39, no. 12, pp. 11250–11256, Sep. 2012, doi: 10.1016/j.eswa.2012.03.073.

[12] W. Buckinx and D. Van den Poel, "Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," *Eur. J. Oper. Res.*, vol. 164, no. 1, pp. 252–268, Jul. 2005, doi: 10.1016/j.ejor.2003.12.010.

[13] D. C. Schmittlein, D. G. Morrison, and R. Colombo, "Counting your customers: Who-are they and what will they do next?" *Manage. Sci.*, vol. 33, no. 1, pp. 1–24, Jan. 1987. [Online]. Available: http://www.jstor.org/stable/2631608

[14] P. S. Fader, B. G. S. Hardie, and K. L. Lee, "'Counting your customers' the easy way: An alternative to the Pareto/NBD model," *Marketing Sci.*, vol. 24, no. 2, pp. 275–284, May 2005, doi: 10.1287/mksc.1040.0098.

[15] P. Fader, B. Hardie, and J. Shang, "Customer-base analysis in a discrete-time noncontractual setting," *Marketing Sci.*, vol. 29, no. 6, pp. 1086–1108, Nov. 2010, doi: 10.2139/ssrn.1373469.

[16] R. Bhagat, S. Muralidharan, A. Lobzhanidze, and S. Vishwanath, "Buy it again: Modeling repeat purchase recommendations," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 62–70, doi: 10.1145/3219819.3219891.

[17] A. Martínez, C. Schmuck, S. Pereverzyev, C. Pirker, and M. Haltmeier, "A machine learning framework for customer purchase prediction in the non-contractual setting," *Eur. J. Oper. Res.*, vol. 281, no. 3, pp. 588–596, Mar. 2020, doi: 10.1016/j.ejor.2018.04.034.

[18] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *Eur. J. Oper. Res.*, vol. 269, no. 2, pp. 760–772, Sep. 2018, doi: 10.1016/j.ejor.2018.02.009.

[19] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Exp. Syst. Appl.*, vol. 38, no. 3, pp. 2354–2364, Mar. 2011, doi: 10.1016/j.eswa.2010.08.023.

[20] D. Xu, W. Yang, and L. Ma, "Repurchase prediction based on ensemble learning," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Oct. 2018, pp. 1317–1322, doi: 10.1109/SMARTWORLD.2018.00229.

[21] C.-J. Liu, T.-S. Huang, P.-T. Ho, J.-C. Huang, and C.-T. Hsieh, "Machine learning-based e-commerce platform repurchase customer prediction model," *PLoS ONE*, vol. 15, no. 12, Dec. 2020, Art. no. e0243105, doi: 10.1371/journal.pone.0243105.

[22] C. Zhu, M. Wang, and C. Su, "Prediction of consumer repurchase behavior based on LSTM neural network model," *Int. J. Syst. Assurance Eng. Manage.*, vol. 13, no. 3, pp. 1042–1053, Aug. 2021, doi: 10.1007/s13198-021-01270-0.

[23] S.-M. Xie, "Comparative models in customer base analysis: Parametric model and observation-driven model," *J. Bus. Econ. Manage.*, vol. 21, no. 6, pp. 1731–1751, Oct. 2020, doi: 10.3846/jbem.2020.13194.

[24] P. Chou, H. H.-C. Chuang, Y.-C. Chou, and T.-P. Liang, "Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning," *Eur. J. Oper. Res.*, vol. 296, no. 2, pp. 635–651, Jan. 2022, doi: 10.1016/j.ejor.2021.04.021.

[25] J. A. Bogonko, G. Orwa, and A. Wanjoya, "Modeling of average survival time for a loss to be handled in insurance company," *Amer. J. Math. Comput. Model.*, vol. 5, no. 1, pp. 18–21, Feb. 2020, doi: 10.11648/j.ajmcm.20200501.13.

[26] M. Mavri and G. Ioannou, "Customer switching behaviour in Greek banking services using survival analysis," *Managerial Finance*, vol. 34, no. 3, pp. 186–197, Feb. 2008, doi: 10.1108/03074350810848063.

[27] R. M. Yusof, S. Aliyu, S. J. M. Khan, and N. H. A. Majid, "Supply overhang of affordable homes: A survival analysis on housing loans application," *Planning Malaysia*, vol. 17, no. 1, pp. 250–266, 2019, doi: 10.21837/pm.v17i9.603.

[28] Á. Periáñez, A. Saas, A. Guitart, and C. Magne, "Churn prediction in mobile social games: Towards a complete assessment using survival ensembles," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Montreal, QC, Canada, Oct. 2016, pp. 564–573, doi: 10.1109/DSAA.2016.84.

[29] M. Viljanen, A. Airola, T. Pahikkala, and J. Heikkonen, "Modelling user retention in mobile games," in *Proc. IEEE Conf. Comput. Intell. Games (CIG)*, Santorini, Greece, Sep. 2016, pp. 1–8, doi: 10.1109/CIG.2016.7860393.

[30] E. Lee, Y. Jang, D.-M. Yoon, J. Jeon, S.-I. Yang, S.-K. Lee, D.-W. Kim, P. P. Chen, A. Guitart, P. Bertens, Á. Periáñez, F. Hadiji, M. Müller, Y. Joo, J. Lee, I. Hwang, and K.-J. Kim, "Game data mining competition on churn prediction and survival analysis using commercial game log data," *IEEE Trans. Games*, vol. 11, no. 3, pp. 215–226, Sep. 2019, doi: 10.1109/TG.2018.2888863.

[31] G. Moat and S. Coleman, "Survival analysis and predictive maintenance models for non-sensored assets in facilities management," in *Proc. IEEE Int. Conf. Big Data*, Orlando, FL, USA, Dec. 2021, pp. 4026–4034, doi: 10.1109/bigdata52589.2021.9671625.

[32] Q. Feng, S. Sha, and L. Dai, "Bayesian survival analysis model for girth weld failure prediction," *Appl. Sci.*, vol. 9, no. 6, p. 1150, Mar. 2019, doi: 10.3390/app9061150.

[33] D. Papathanasiou, K. Demertzis, and N. Tziritas, "Machine failure prediction using survival analysis," *Future Internet*, vol. 15, no. 5, p. 153, Apr. 2023, doi: 10.3390/fi15050153.

[34] B. Snider and E. A. McBean, "Combining machine learning and survival statistics to predict remaining service life of watermains," *J. Infrastruct. Syst.*, vol. 27, no. 3, Sep. 2021, Art. no. 04021019, doi: 10.1061/(ASCE)IS.1943-555X.0000629.

[35] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Feb. 2019, doi: 10.1145/3214306.

[36] J. Tobin, "Estimation of relationships for limited dependent variables," *Econometrica*, vol. 26, no. 1, pp. 24–36, Jan. 1958, doi: 10.2307/1907382.

[37] E. Liu, R. Y. Liu, and K. Lim, "Using the Weibull accelerated failure time regression model to predict time to health events," *Appl. Sci.*, vol. 13, no. 24, p. 13041, Dec. 2023, doi: 10.3390/app132413041.

[38] S. Mittal, D. Madigan, R. S. Burd, and M. A. Suchard, "High-dimensional, massive sample-size cox proportional hazards regression for survival analysis," *Biostatistics*, vol. 15, no. 2, pp. 207–221, Apr. 2014, doi: 10.1093/biostatistics/kxt043.

[39] E. Marubini and M. G. Valsecchi, *Analysing Survival Data From Clinical Trials and Observational Studies*. Hoboken, NJ, USA: Wiley, Jul. 2004.

[40] S. Pölsterl, N. Navab, and A. Katouzian, "Fast training of support vector machines for survival analysis," in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science), vol. 9285, Jan. 2015, pp. 243–259, doi: 10.1007/978-3-319-23525-7_15.

[41] L. Gordon and R. Olshen, "Tree-structured survival analysis," *Cancer Treat. Rep.*, vol. 69, no. 10, pp. 1065–1069, Oct. 1985.

[42] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 841–860, Sep. 2008, doi: 10.1214/08-aoas169.

[43] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. van der Laan, "Survival ensembles," *Biostatistics*, vol. 7, no. 3, pp. 355–373, Jul. 2006, doi: 10.1093/biostatistics/kxj011.

[44] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Med. Res. Methodol.*, vol. 18, no. 1, pp. 1–12, Feb. 2018, doi: 10.1186/s12874-018-0482-1.

[45] C. Lee, W. Zame, J. Yoon, and M. Van der Schaar, "DeepHit: A deep learning approach to survival analysis with competing risks," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1–8, doi: 10.1609/aaai.v32i1.11842.

[46] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *J. Clin. Epidemiol.*, vol. 59, no. 10, pp. 1087–1091, Oct. 2006, doi: 10.1016/j.jclinepi.2006.01.014.

[47] Y. He, "Missing data analysis using multiple imputation: Getting to the heart of the matter," *Circulation, Cardiovascular Quality Outcomes*, vol. 3, no. 1, pp. 98–105, Jan. 2010, doi: 10.1161/circoutcomes.109.875658.

[48] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statist. Med.*, vol. 30, no. 4, pp. 377–399, Feb. 2011, doi: 10.1002/sim.4067.

[49] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statist. Med.*, vol. 15, no. 4, pp. 361–387, Feb. 1996, doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[50] Y. Li, J. Wang, J. Ye, and C. K. Reddy, "A multi-task learning formulation for survival analysis," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, Aug. 2016, pp. 1715–1724, doi: 10.1145/2939672.2939857.

[51] M. S. Kovalev, L. V. Utkin, and E. M. Kasimov, "SurvLIME: A method for explaining machine learning survival models," *Knowl.-Based Syst.*, vol. 203, Sep. 2020, Art. no. 106164, doi: 10.1016/j.knosys.2020.106164.

[52] M. Krzyziński, M. Spytek, H. Baniecki, and P. Biecek, "SurvSHAP(t): Time-dependent explanations of machine learning survival models," *Knowl.-Based Syst.*, vol. 262, Feb. 2023, Art. no. 110234, doi: 10.1016/j.knosys.2022.110234.

YOUNGJUNG SUH received the B.S. degree in computer engineering from Chonnam National University, in 2001, and the M.S. and Ph.D. degrees in information and communication engineering from Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea. In 2011, as a Research Engineer, she joined LG Electronics, Seoul, South Korea. Since March 2024, she has been with Kongju National University, where she is currently an Assistant Professor with the Department of Computer Science and Engineering. Her research interests include big data analytics, business analytics, machine learning, and context awareness.