A dynamic ensemble selection method for bank telemarketing sales prediction

Yi Feng[1], Dujuan Wang[1,*], Yunqiang Yin[2], Lalitha Dhamotharan[3]

[1] Business School, Sichuan University, Chengdu 610064, China

[2] School of Economics and Management, University of Electronic Science and Technology of China, Chengdu 610064, China

[3] Centre for Simulation, Analytics and Modelling (CSAM), University of Exeter Business School, EX4 4PU, UK

Abstract: We propose a dynamic ensemble selection method, META-DES-AAP, to predict the success of bank telemarketing sales of time deposits. Unlike existing machine learning-based marketing sales prediction methods focusing only on prediction accuracy, META-DES-AAP considers the accuracy and average profit maximization. In META-DES-AAP, to consider both accuracy and average profit in the framework of dynamic ensemble selection using meta-training, a multi-objective optimization algorithm is designed to maximize the accuracy and average profit for base classifiers selection. Base classifiers suitable for each test telemarketing campaign are integrated according to the dynamic-based base classifiers integration method. Experimental results on bank telemarketing data show that META-DES-AAP achieves the best accuracy and the largest average profit when compared across several state-of-the-art machine learning methods. In addition, the factors influencing telemarketing on the average predicted probability of telemarketing success and average profit obtained by META-DES-AAP are analyzed.

Keywords: Time deposits; Multi-objective; Dynamic ensemble selection; Telemarketing sales; Marketing strategy

1. Introduction

For contemporary financial institutions, highly liquid deposits serve as the main funding source, accounting for more than two-thirds of the total asset sources (Mei et al., 2016). In particular, large retail banks, an integral part of the globalized financial system, are key intermediary players in international trade and investment. These large retail banks require sufficient deposits to thrive and expand their businesses (Dincer et al., 2019). The performance of most retail banks depends largely on the success of their marketing campaigns (i.e., whether customers subscribe to their financial products). Therefore, stronger stability and lower cost of financial products and time deposits inform the bank's future expansion capacity on loans. However, conducting successful marketing campaigns in a competitive financial market is undoubtedly difficult, as retail banks are constrained to invest in 30% of customers below the economic profit threshold (Méndez-Suárez & Crespo-Tejero, 2020). Hence, banks attempt to target marketing campaigns efficiently in their profit and social welfare maximization strategies.

One of the main ways banks reach their customers is through telemarketing services. As a marketing method, this covers the promotion of products by making calls out of a curated customer list (Kotler, 2016), which has the advantage of lower cost and higher yield than other methods (Mei et al., 2016; Yan et al., 2020) when a consumer is accurately targeted. In other words, conducting telemarketing without a strategy can result in unnecessary cost expenditures (Yan et al., 2020). Furthermore, several customers face discomfort when they experience telemarketing campaigns because they are not interested in products or think telemarketing is an invasion of their privacy (Ritsema & Piëst, 1990; Yan et al., 2020). Thus, some customers' opinions, critics, and telemarketing complaints are easily found on social media (Maulana & Nurulfirdausi, 2015). To better regulate telemarketing and reduce its negative perceptions, state governments and regulators have launched regulations and laws for telemarketing in their jurisdictions to ensure the legality of telemarketing campaigns and protect customers' privacy (Rita, 1995). Moreover, companies employ self-regulation in telemarketing, professional training, and telemarketing strategies to regulate telemarketing practices and the behavior of telemarketers (Ritsema & Piëst, 1990). There has been continuous advancement in machine learning methods that provide opportunities to strategize marketing and sales needs (Reis et al., 2020). Therefore, there is a great need to conduct telemarketing sales prediction studies using machine learning methods to further guide marketers in developing telemarketing strategies.

Machine learning is a method of data mining, and a precursor to artificial intelligence (Bose & Mahapatra, 2001; Antons & Breidbach, 2018; Lismont et al., 2017) provides a way to scale complex problems by building predictive models based on historical datasets and using them to make predictions (Chen et al., 2012). At present, machine learning has led to several advancements in healthcare (Chatterjee et al., 2020), finance (Zhu et al., 2019), education (Fernandes et al., 2019), and other data-intensive fields. Companies that use machine learning methods and data analytics have seen a 5% increase in productivity and a 6% increase in profits compared to those that do not use them (Mcafee & Brynjolfsson, 2012). An increasing number of companies invest in advanced marketing sales predictive technologies to better address their market needs (Chen et al., 2015; Jimenez-Marquez et al., 2019).

Nonetheless, machine learning methods in marketing management are still under development (Salminen et al., 2019). Specifically, current machine learning-based marketing sales predictive studies often lack a careful consideration of the tradeoff between cost and profitability in the model construction (Jacobs et al., 2016; Martínez et al., 2018; Alemán Carreón et al., 2019; Ren et al., 2020). The purpose of a marketing sales predictive model should include the accurate prediction of future campaigns that would lead to marketing success and support marketers to intuitively understand how much profit can be achieved. Hence, telemarketing sales predictive models considering accuracy and average profit (AP) can help marketing managers turn future marketing decisions into benefits. The profit-maximizing behavior of enterprises is one of the oldest, most fundamental, and widely used assumptions in all fields of economics (Levitt, 2016). Just as shareholders demand that management maximize the profits of the enterprise, marketers are under increasing pressure from management to demonstrate their contribution to the benefits through the effective and efficient use of company resources (Hanssens et al., 2009; Kim & Mcalister, 2011; Markovitch et al., 2020).

This study proposes a dynamic ensemble selection method that considers the accuracy and AP with meta-training, which we refer to as META-DES-AAP for telemarketing sales prediction. In META-DES-AAP, we explore the accuracy of the model in its construction and the AP maximization. Compared with the mainstream machine learning methods, the predictive performance of the proposed method is systematically evaluated from both machine learning and economic metrics, and the experimental results verify the superiority of the proposed method. In addition, many machine learning techniques that lack interpretability and understandability are considered black-box models (Martens et al., 2011; Najafabadi et al., 2015; Albrecht et al., 2021). We provide a post hoc explanation of the predicted results for marketing managers to better understand the effects on the average predicted probability of telemarketing success (TSP) and the AP obtained by META-DES-AAP in the bank telemarketing sales of time deposits. This allows marketers to understand the mechanisms and dynamics that lead to the prediction results.

In summary, this study addresses the following research questions. 1) How can we turn telemarketing sales predictions into benefits using advanced machine learning methods? 2) How can we consider both model performance (accuracy) and economic performance (AP) in the construction of a telemarketing sales predictive model simultaneously? 3) How does the proposed method perform better than mainstream machine learning methods? 4) What affects the trend of the TSP and AP obtained by META-DES-AAP?

The remainder of the paper is organized as follows. In Section 2, we briefly review the related literature on machine learning in marketing sales prediction and classification methods. In Section 3, we introduce META-DES-AAP in detail. In Section 4, the experimental design and results are elaborated and analyzed based on a real bank telemarketing dataset. The theoretical implications, managerial implications, and limitations of this study and directions for future studies are discussed in Section 5. Finally, the conclusions are drawn in the last section.

2. Literature review

This section first reviews the literature on machine learning methods in marketing sales prediction. Next, we introduce the formulation of classification problems and classification methods in machine learning. Finally, we conclude with a summary of the contributions of this study.

2.1 Machine learning in marketing sales prediction

Machine learning is a general term that covers various computer-based data mining techniques to discover complex patterns in data, especially big data (Pereira et al., 2018). The literature on marketing topics suggests that machine learning methods can provide effective decision support for both direct marketing (Cui & Man, 2004; Adyyński et al., 2019) and strategic marketing (Orriols-Puig et al., 2013; Salminen et al., 2019).

The prediction of customers' purchase intentions has always been an interesting research issue in marketing. Adyyński et al. (2019) proposed a credit product prediction method based on random forest classification and deep neural networks, which predicts customers' willingness to take out a personal loan, i.e., whether they are interested in buying a credit product or not, and the experimental results proved that the proposed method has good recall and accuracy. Ren et al. (2020) proposed a two-stage hybrid model to predict consumers' willingness to purchase (purchase or not) using e-coupons, where the first stage uses a clustering algorithm to obtain groups of consumers with different levels of the campaign, followed by a predictive model based on the behavioral characteristics of consumers in the different groups. Martínez et al. (2018) developed a machine learning framework for the given historical transaction data to predict whether potential customers would perform the purchase in the future. In the proposed framework, the performance of logistic LASSO regression, extreme learning machine, and gradient boosting tree is evaluated, and the experimental results suggest that the gradient boosting tree performs the best. To accurately predict what products customers would purchase next, Jacobs et al. (2016) developed a novel prediction method containing latent dirichlet allocation (LDA) and mixtures of Dirichlet-Multinomials (MDM). They found that the proposed method has greater scalability for future large product purchase predictions. Based on TV advert exposure time data, Alemán Carreón et al. (2019) applied machine learning methods including support vector machines (SVM), extreme gradient boosting (XGBoost), and logistic regression (LR) to predict customers' purchase intentions.

Bank telemarketing research is relatively small compared to other marketing studies. Many scholars have also tried to use machine learning methods to investigate the bank telemarketing sales prediction problem in recent years. Yan et al. (2020) proposed an improved whale optimization algorithm (IWOA) to optimize the weights between the input and competing layers of the S_Korhonen network. They applied the developed method to predict the bank telemarketing problem. The experimental results showed that the proposed method has good classification accuracy. Moro et al. (2014) applied four machine learning methods (LR, decision tree, neural networks, and SVM) to develop predictions on the success of bank telemarketing calls and found that the neural networks had the best AUC values. Moro et al. (2018) proposed a data mining approach based on a divide-and-conquer strategy, whereby sensitivity analysis was performed on the bank telemarketing dataset to extract variable correlations, and expert analysis was conducted to find the best subset of features. Call direction (inbound or outbound) was considered to be the most important feature. As explained in the introduction, a highly accurate predictive model does not provide a marketer or management for an intuitive understanding of future benefits. In this work, accuracy and profit are considered equally important, and they are all considered in the construction of the telemarketing predictive model.

2.2 Classification methods

In a classification problem, a dataset is provided in which each sample in the dataset consists of multiple variables or attributes. One of the variables, called the decision label, should be categorical, indicating the category to which each sample belongs. The classification task aims to build a classifier that can predict the decision label after inputting the variables for a particular sample (Zhang et al., 2019). Classification methods can be broadly classified into two categories: dynamic and static. Both static and dynamic classification methods are further subdivided in terms of classifier selection methods and ensemble selection methods (Cruz et al., 2018; Zhang et al., 2019), the differences of which are depicted in Figure 1.

[Insert Figure 1 about here]

The main difference between static and dynamic classification methods is whether the same classifier is used to predict all test samples. The main difference between classifier selection and ensemble classifier selection is whether a single classifier or an ensemble classifier integrated by multiple base classifiers is used to predict test samples, where the different combinations between them result in a variety of classification techniques. However, static classification methods are usually unable to achieve the best performance on a given sample because different classifiers have different performances in different regions (Zhang et al., 2019); it is necessary to try our best to find the best classifier to predict a given sample. In addition, several studies have shown that the performance of ensemble classifiers is better than that of single classifiers (Hwang et al., 2020; Papouskova & Hajek, 2019; Zhu et al., 2019). Therefore, this study chooses the dynamic ensemble selection method to develop a classification prediction. In what follows, we perform a brief review of the dynamic ensemble classification methods.

One of the most promising classification techniques is the dynamic ensemble selection (Cruz et al., 2018), which dynamically determines the most appropriate classifier for each new sample to be classified. The most critical issue in the dynamic ensemble selection methods is selection, i.e., how to select the most competent base classifiers from a base classifier pool for any given query sample, thus guiding subsequent base classifiers integration. Among the existing studies related to the dynamic ensemble selection method, various criteria are used to assess the competence of the base classifiers. However, most existing studies, such as the K-nearest oracles eliminate (Ko et al., 2008), K-nearest oracles union (Ko et al., 2008), ensemble selection performance (Woloszynski et al., 2012), and k-nearest output profiles (Cavalin et al., 2013), only use a single criterion (accuracy or probability) to evaluate the ability of the underlying classifier. It is easy to make mistakes when using only a single standard to measure the ability level of the base classifier. Therefore, in the selection process, multiple criteria should be considered to measure the ability level of the base classifiers to reduce the error of the ability level assessment of the base classifier and realize a powerful dynamic ensemble selection method. To the best of our knowledge, only a few studies (Cruz et al., 2015; Cruz et al., 2017; Hou et al., 2020) have been conducted in this way. For example, Cruz et al. (2015) proposed a dynamic ensemble selection method using meta-training, namely META-DES and used several criteria (including probability, accuracy, behavior, and others) to assess whether the base classifier can correctly classify a certain test sample. In this study, we chose to use the META-DES framework to construct a predictive model to predict the success of bank telemarketing sales for time deposits.

2.3 Contribution

By reviewing the above literature, we summarize the machine learning methods in marketing sales prediction from seven aspects, as listed in Table 1.

[Insert Table 1 here]

Different prediction results will bring a certain profit or loss to the company in different marketing campaigns, i.e., a correct (resp., wrong) prediction will bring a certain profit (resp., loss) to the company. However, from Table 1, we can observe that most scholars focus only on the predictive model's performance in terms of machine learning metrics, and few have considered accuracy and AP simultaneously in constructing the predictive model. Furthermore, although the dynamic ensemble selection method is proven to be one of the most promising classification techniques (Cruz et al., 2018), most studies still apply static classification methods to conduct predictions in marketing, limiting the further improvement of predictive model performance. In addition, few studies extract interpretable information from predictive models to support the development of marketing strategies. Our study aims to narrow the above gaps, advancing the development of machine learning methods in the prediction of telemarketing sales of time deposits, and our main contributions can be summarized as follows:

1) We propose a novel model termed META-DES-AAP, which considers both the accuracy of the model and AP maximization in predicting the success of bank telemarketing sales of time deposits.
2) We embed a multi-objective evolutionary algorithm in META-DES-AAP to form an accuracy- and profit-based optimal base classifier pool and use a dynamic-based base classifiers integration method to integrate these base classifiers. The predictive results include TSP and AP, which can help marketing managers improve cost reductions.
3) To support actionable insights, the model supports a post hoc explanation of META-DES-AAP on TSP and AP for important factors, which helps marketers better understand the

impact and dynamics of the underlying factors that led to the predicted results.

4) We conduct numerical experiments on the telemarketing data of a bank in Portugal. The results demonstrate that META-DES-AAP achieves the best accuracy and largest AP compared to several other state-of-the-art machine learning methods.

3. Model development

In the bank telemarketing process, marketers call on customer lists to sell time deposits, resulting in a binary result for each call, i.e., successful or failed. The focus of this study is to predict whether telemarketers will be successful in selling time deposits, and we propose a dynamic ensemble selection method called META-DES-AAP for bank telemarketing sales of time deposits.

Figure 2 presents the framework of the proposed method. First, all base classifiers in the accuracy- and profit-based optimal sub-pool are trained; they are determined by executing a multi-objective evolutionary algorithm. Second, dynamic ensemble selection using meta-training is conducted. Specifically, in META-DES-AAP, the level of competence of the base classifiers is measured, and the competent base classifiers are selected for each test sample from the accuracy- and profit-based optimal sub-pool; thus, the optimal pool for each test sample can be determined, where the optimal pool for each test sample is composed of all competent base classifiers corresponding to that sample. Finally, for all competent base classifiers in the optimal pool of each test sample, a dynamic-based base classifiers integration method is employed to integrate them and output the predicted results. Moreover, it is necessary to extract valuable information from META-DES-AAP to guide marketers in developing marketing strategies; thus, we use the statistical method to conduct post hoc interpretability visually.

[Insert Figure 2 here]

As mentioned above, finding the accuracy- and profit-based optimal sub-pool is the key to achieving maximum accuracy and AP in the proposed method. We treat this problem as a multi-objective task and propose a multi-objective base classifiers selection method based on the binary multi-objective nondominated sorting based genetic algorithm (NSGA-II) (Deb et al., 2002). In META-DES-AAP, NSGA-II performs multi-objective base classifiers selection for all base classifiers in the initial pool by optimizing the two objectives of classification accuracy and AP simultaneously to find an accuracy- and profit-based optimal sub-pool.

Figure 3 illustrates the evolutionary process for determining this sub-pool with optimal accuracy and AP, where g denotes the gth iteration ($g = 1,2, \ldots, G$), G represents the maximum number of iterations, p signifies the pth individual among the population, and P is the size of the population. Initially, a heterogeneous base classifier pool is generated. Next, NSGA-II is carried out on the generated pool and generates the initial individuals, where each individual represents a sub-pool. Subsequently, dynamic ensemble selection using meta-training is performed in each sub-pool to calculate and evaluate the competence of the base classifiers in each sub-pool. After the evaluation, the competent classifiers are selected in each sub-pool for a certain test sample. All the selected competent classifiers are integrated among each sub-pool using the dynamic-based base classifiers integration method as the final ensemble to predict a certain test sample. The predicted results of all sub-pools are evaluated based on the accuracy and AP, the values of which are used to guide the new iteration. Finally, when G is reached, an accuracy- and profit-based optimal sub-pool is obtained. All the key parts of the proposed method, META-DES-AAP, are described in detail.

[Insert Figure 3 here]

(1) Generation of the heterogeneous base classifiers pool: It is widely accepted that the key factor for constructing a good ensemble is the diversity among the base classifiers in the pool, and existing studies have shown that diversity among heterogeneous base learners is higher than that among homogeneous base learners (Ghaderi Zefrehi & Altınçay, 2020;

Papouskova & Hajek, 2019). Specifically, the potential bias of each base classifier owing to its inherent assumptions can be reduced by using heterogeneous integration, which leads to better generalization to unseen samples. As alternative diversity methods, the random subspace selection method and bootstrap sampling method have also been validated in existing studies. To generate a base classifier pool with diversity described above, META-DES-AAP incorporates heterogeneous ideas and a random subspace selection method and bootstrap sampling method to generate the initial base classifier pool. First, the random subspace selection method and bootstrap sampling method are used on the training dataset to generate a training subset with different features and different samples for each base classifier. Immediately following this, the heterogeneous base classifiers are trained on their respective training subsets, and the heterogeneous base classifiers pool is generated at this point.

In META-DES-AAP, the base classifiers include six different mainstream machine learning methods: decision trees (DT) (Mitchell, 1997), logistic regression (LR) (Mitchell, 1997), random forests (RF) (Breiman, 2001), gradient boosting decision trees (GBDT) (Friedman, 2001), eXtreme gradient boosting (XGBoost) (Chen & Guestrin, 2016), and light gradient boosting machine (LightGBM) (Ke et al., 2017). The reasons for choosing these methods are as follows.

First, DT, LR, and RF are popular in marketing sales prediction and are frequently used as the main or important components of predictive models (Moro et al., 2014; Adyyński et al., 2019; Ren et al., 2020). Second, XGBoost, GBDT, and LightGBM outperform SVM, artificial natural network (ANN), and other methods in many existing studies (Chen & Guestrin, 2016; Ke et al., 2017; Hwang et al., 2020). Third, the above six methods have outstanding performance characteristics and increase diversity due to the theoretical differentiation of the different methods. Finally, as mentioned earlier, efficiency is an important consideration for META-DES-AAP. Telemarketing data must be updated regularly; therefore, it is very important to consider the computational cost in constructing telemarketing sales, predictive models. Studies have shown that nonlinear SVM and ANN as base classifiers will inevitably increase the computational cost; therefore, these two methods are unsuitable for use as base classifiers (Xia et al., 2020). In contrast, XGBoost, LightGBM, RF, and GBDT are tree-based ensemble learning methods with low computational cost, especially LightGBM, which supports GPU computing and distributed computing with satisfactory computational cost (Ke et al., 2017). Therefore, these methods can better handle large amounts of data. In summary, the heterogeneous classifier pools generated based on these six classification methods can achieve better performance and lower computational costs than other methods.

(2) Individual generation based on NSGA-II: Based on the generated heterogeneous base classifiers, NSGA-II is performed to find the optimal sub-pool that maximizes the accuracy and AP, where NSGA-II achieves multi-objective optimization by finding a set of non-dominated solutions (sub-pools) that form an effective Pareto frontier. Individual generation is the first step of NSGA-II, and an initial population containing n individuals is generated using random gene values. In the case of base classifiers selection, we encode an individual with a set of binary genes, each representing whether a particular base classifier is selected. Specifically, for each trained base classifier, the value "1" represents the selected base classifier, and "0" represents the unselected base classifier. Therefore, each individual is decoded to obtain a sub-pool, where it consists of the base classifiers with gene value "1" in that individual. All the obtained sub-pools will be meta-trained separately afterward.

(3) Dynamic ensemble selection using meta-training: Dynamic ensemble selection using meta-training aims to train the meta-classifier based on the meta-feature and is used to determine whether the base classifier C_j is competent enough to classify the test sample $X_{i,test}$. The meta-features consist of the behavioral information associated with C_j for measuring its level of competence. The meta-classes of meta-classifiers are either "competent" or "incompetent" to classify $X_{i,test}$ for C_j. It consists of four main steps: un-consistent samples selection, meta-feature extraction, meta-classifier training, and competent base classifiers selection, which are elaborated as follows.

Step 1 *Un-consistent samples selection*. The low consensus of the base classifiers in the sub-pool is one of the problems in dynamic ensemble selection when classifying test samples (Cruz et al., 2015; Hou et al., 2020). Therefore, it is necessary to select un-consistent samples from the training dataset and focus on dealing with them for each sub-pool. In META-DES-AAP, the specific selection method for each sub-pool is as follows: first, the base classifiers in each sub-pool predict for each sample in the training dataset, and the degree of consensus of each sample in the training dataset is calculated. If the sample consensus degree value is lower than the consensus threshold h, then the sample is selected, where h is pre-set by the decision-maker. For different sub-pools, the corresponding un-consistent data that contain the selected samples can be obtained.

Before conducting the mate-feature extraction, K nearest neighbours of each sample in un-consistent data should be determined from the un-consistent data using K nearest neighbours algorithm (KNN), and the region of competence (RC) of each sample can also be obtained. Let $\theta_t = \{x_1, x_2, \dots, x_K\}$ be the RC for the tth sample in the un-consistent data. Conversely, each sample in un-consistent data should also be converted into an output profile, consisting of the predicted labels of all base classifiers in the sub-pool on that sample. Based on all transformed output profiles, K_d most similar output files of each sample can be determined and called the decision space-based competence region (DSCR), and $\phi_t = \{x_1, x_2, \dots, x_{K_d}\}$ is used to denote the DSCR for the tth sample.

Step 2 *Meta-feature extraction*. For each sub-pool, the meta-features can be viewed as different attributes of the behavior of the base classifiers in the sub-pool. After determining θ_t and ϕ_t, the meta-features of tth sample can be extracted according to the performance of C_j in

θ_t and ϕ_t. In META-DES-AAP, five different meta-features are used to measure the competence level of C_j from five different aspects, described as follows:

f_1- Neighbors' hard classification: a K-dimensional binary vector, where the kth element is equal to 1 if and only if C_j predicts correctly for the sample x_k in θ_t;

f_2- Posterior probability: a K-dimensional binary vector, where the kth element is equal to the posterior probability $P(class\ 1|x_k)$ of C_j for the sample x_k in θ_t;

f_3- Overall local accuracy: the accuracy of base classifier C_j over whole θ_t;

f_4- Output profiles classification: A K_d-dimension binary vector, where each element is equal to 1 if and only if C_j predicts correctly for the sample x_{kd} in ϕ_t;

f_5- Classifier's confidence: the perpendicular distance between the tth input sample and decision boundary of C_j.

The Min-Max normalization method is used to normalize f_5. Thus, for the tth sample on C_j, the meta-vector $v_{jt} = (f_1, f_2, f_3, f_4, f_5)$ can be obtained, as shown in Figure 4. Moreover, we denote by L_{jt} the local accuracy of C_j on tth sample, which is a binary parameter equal to 1 if and only if C_j correctly predicts for the tth sample. The obtained v_{jt} and L_{jt} are then merged to form the meta-training dataset.

[Insert Figure 4 here]

Step 3 *Meta-classifier training*. The following step is the meta-classifier training, and the obtained meta-training dataset is used to train the meta-classifier in this step. The selection of the meta-classifier can be determined using the method developed by Cruz et al. (2015). Many existing studies selected multi-layer perception (MLP), SVM, and multinomial naïve Bayes (MNB) as meta-classifiers. However, as stated above, MLP, SVM, and MNB do not perform well in terms of performance and computational cost (Xia et al., 2020). Therefore, in META-DES-AAP, we choose XGBoost as the meta classifier because of its excellent performance and low computational cost.

Step 4 *Competent base classifiers selection*. For each test sample $X_{i,test}$ from the test dataset, the competent base classifiers are selected from all base classifiers in each sub-pool. Specifically, according to the description of *step 1*, the θ_i of $X_{i,test}$ and the ϕ_i of $X_{i,test}$ can be found. Next, the meta-feature vectors v_{ji} are calculated according to the competent performance of each base classifier C_j in θ_i and ϕ_i. Subsequently, the meta-feature vectors are input to the meta-classifier to predict whether C_j has the competence to correctly predict $X_{i,test}$. Finally, we use a threshold value of 0.5 to judge the competence of C_j on $X_{i,test}$ as in Cruz et al. (2018) and Hou et al. (2020). Specifically, if the predicted probability is greater than 0.5, the label L_{ji} is set to 1 and 0 otherwise. After competent base classifiers selection, the competent base classifiers set for each $X_{i,test}$ can be obtained, which forms the ground for the integration described as follows.

(4) Dynamic-based base classifiers integration: In each sub-pool, for the competent base classifiers set of $X_{i,test}$, a dynamic-based base classifiers integration is used to integrate these competent base classifiers. First, the predicted posterior probability of each competent base classifier with $L_{ji} = 1$ is obtained using the above operations. Second, for all the competent base classifiers of $X_{i,test}$, a weighted majority voting scheme is used to integrate all competent base classifiers based on their competence level. The higher the predicted posterior probability of the competent base classifier, the higher the weight it takes in the integration. Finally, the ensemble classifier for $X_{i,test}$ is applied to make predictions and output the prediction results.

(5) Accuracy- and profit-based evaluation: After obtaining the prediction results on the test dataset for each sub-pool, the prediction results will be evaluated for the sub-pools using the fitness functions of NSGA-II. In META-DES-AAP, we aim to find the maximum AP in the sub-pools with the best accuracy to better guide marketing decisions. The bank telemarketing sales predictive problem that we concentrate on in this study is a two-class problem, i.e., the predicted result of each sample belongs either to *failed marketing* (class 0) or *successful marketing* (class 1). Specifically, the classifiers give the probability of *successful marketing* to each telemarketing campaign. A telemarketing campaign is considered *successful* if the corresponding probability value exceeds a cutoff value α, and vice versa.

[Insert Table 2 here]

Table 2 provides a confusion matrix with benefits and costs, where N_{test} denotes the number of samples in the test dataset; π_1 and π_0 represent the prior probabilities of samples belonging to *successful marketing* and *failed marketing*, respectively; $F_1(\alpha)$ and $F_0(\alpha)$ are the cumulative density functions for *successful marketing* and *failed marketing*, respectively. Generally speaking, the results of classification prediction are used as inputs to marketing strategies, resulting in benefits for correct classification and incorrect classification costs. We denote by $P(a|b)$ the cost or benefit associated with the classifier that classifies a sample from class b to class a with $a, b \in \{0,1\}$. Thus, in Table 2, B_0 (resp., B_1) is the benefit for classifying a sample from class 0 (resp., 1) to class 0 (resp., 1), and C_0 (resp., C_1) is the cost for classifying a sample from class 0 (resp., 1) to class 1 (resp., 0). The specific values for B_0, B_1, C_0 and C_1 can be determined by managers according to the bank's situation. In marketing management practice, there are generally the following two relationships: $B_1 \gg B_0$ and $C_1 \gg C_0$. The former inequality means that the profit from a correct prediction of a telemarketing campaign that will be successful is much greater than that from a correct prediction of a telemarketing campaign that will fail. The latter indicates that the cost of an incorrect prediction for a telemarketing campaign that will actually be successful is much greater than that from an incorrect prediction for a telemarketing campaign that will be failed.

Let TB_1 (resp., TB_0) be the total profit from the correct prediction of telemarketing campaigns that will actually be successful (resp., failed), and TC_1 (resp., TC_0) be the total cost from the incorrect prediction of telemarketing campaigns that will actually be successful (resp., failed). Then, we have $TB_1 = \pi_1(1 - F_1(\alpha))N_{test} \times B_1$, $TB_0 = \pi_0 F_0(\alpha)N_{test} \times B_0$, $TC_1 = \pi_1 F_1(\alpha)N_{test} \times C_1$, and $TC_0 = \pi_0(1 - F_0(\alpha))N_{test} \times C_0$.

Therefore, the accuracy and AP for each sub-pool can be calculated and evaluated according to the following formulations:

$$Accuracy = \pi_0 F_0(\alpha) + \pi_1\big(1 - F_1(\alpha)\big) \tag{1}$$

$$Average\ profit(AP) = \frac{TB_1 + TB_0 - TC_1 - TC_0}{N_{test}}$$

$$= B_0\pi_0 F_0(\alpha) + B_1\pi_1\big(1 - F_1(\alpha)\big) - C_0\pi_0\big(1 - F_0(\alpha)\big) - C_1\pi_1 F_1(\alpha) \tag{2}$$

(6) Evolution based on NSGA-II: In the following stage, the genetic operators of NSGA-II, including selection, crossover, and mutation, are performed on the parental population based on the accuracy and AP to generate offspring population. For the selection operator, a binary tournament method based on non-dominated sorting and crowding distance is used. First, the population is sorted by the non-domination rank according to the individual's objective values (accuracy and AP). Individuals in the same frontier, i.e., they have the same non-dominance rank, are sorted according to their crowding distance, which can effectively maintain the diversity of these individuals. Uniform crossover and bit flip mutations are applied for crossover and mutation operators. The gene values of the two individuals are flipped with a preset crossover probability and mutation probability. Specifically, each gene value is flipped from zero to one or from one to zero with a predetermined probability.

The offspring population can be obtained after performing genetic operators. The parental and offspring populations with P individuals are merged to form a new population with $2P$ individuals to ensure elitism. Subsequently, the non-dominated sorting and crowding distance computation is conducted again on the new population, and only the top P individuals are selected to enter the next iteration. When the maximum number of iterations reaches G, a non-dominated solution set (an optimal solution set) can be obtained, and each solution in the set can be decoded into a sub-pool, which can be seen as a trade-off between accuracy and AP. This study chooses a solution with optimal AP as the final solution from the non-dominated solution set. The base classifiers with gene value 1 form the final accuracy- and profit-based optimal sub-pool.

After the accuracy- and profit-based optimal sub-pool has been found, the telemarketing sales predictive model with the highest accuracy and maximum AP simultaneously is realized.

4. Experiments

4.1 Data introduction and pre-processing

In this study, five years of real marketing dataset, including 41188 telemarketing campaigns for time deposits, were collected from a retail bank in Portugal, which can be downloaded from the UCI website (https://archive.ics.uci.edu/ml/index.php). In the 41188 telemarketing campaigns, 4639 telemarketing campaigns were successful, and the remaining 36549 campaigns failed. Each campaign consists of 20 variables and a label, where the 20 variables include bank customer information (e.g., "age"), last contact information (e.g., "month"), social and economic background information (e.g., "consumer confidence index"), and other information (e.g., "campaign"). The label "y" is a binary variable, i.e., a label value equal to 1 means that the telemarketing sale of time deposits is successful, and 0 otherwise. The specific descriptions and types of each variable can be found in Table A1 in the Appendix.

Not all samples will be used for model construction because the model can only achieve better performance on a high-quality dataset. After acquiring the dataset, we performed three aspects of data preprocessing before model construction, as follows:

1) *Sample preprocessing*. Note that the sample size of telemarketing campaigns failing is 7.9 times higher than that of telemarketing campaigns succeeding, and there are many outlier samples observed by plotting the box plots. It follows that the data are characterized by an imbalance and multiple outlier samples. Therefore, constructing a telemarketing sales

predictive model of time deposits directly on these data would lead to poor accuracy and AP. In this study, outlier samples are detected and filtered based on the quartile and interquartile distance information, and a balanced dataset can be obtained after performing the undersampling method. After performing the above operations, the 2000 samples obtained are used for model construction, with half the number of samples in each of the 1 and 0 classes. The obtained data are divided into training, dynamic selection, and test datasets at ratios of 33%, 33%, and 33%, respectively, following Hou et al. (2020).

2) *Variable preprocessing.* Different variable types should be treated differently in preprocessing. The original dataset contains many nominal variables, such as the variable “job”. Since there is no correlation between the attributes of such variables, one-hot coding should be used to encode the attributes of such variables instead of numerical mapping coding because training the model on the data after taking numerical mapping coding will produce incorrect judgments. In addition, the original data contain some ordered variables, such as the variable “education”. The attributes of such variables are correlated with each other, e.g., there is a higher or lower level of education. Therefore, we encode them using a numerical mapping, the higher the education level, the larger the value. Finally, for the label “y” the attribute value of “yes” is assigned to 1 and vice versa.

3) *Training dataset normalization.* In the original dataset, some variables have different types and ranges of attribute values. If the differences are too large, the training model will give high weights to the attributes with high values, which will cause the model to provide wrong perceptions and incur high computational costs. Therefore, it is necessary to normalize the training dataset before model construction, which helps to speed up the model’s convergence. This study standardizes the training dataset by applying the max-min standardization method, and the subsequent training of the model will be implemented on the standardized dataset.

4.2 Performance evaluation

To perform a comprehensive and reliable evaluation of the proposed method and the mainstream machine learning methods in terms of machine learning metrics and economic metrics, we conduct several comparative experiments based on five popular machine learning metrics: accuracy, recall, precision, specificity, and area under the receiver operating characteristic (ROC) curve (AUC), and four economic evaluation metrics: total benefits (TB), total cost (TC), net profit (NP), and AP. As a binary classification problem, these evaluation metrics can be calculated based on the confusion matrix, as shown in Table 2.

Specifically, accuracy is defined in Eq. (1), and the other four machine-learning metrics are calculated as follows:

$$Recall = Sensitivity = 1 - F_1(\alpha) \tag{3}$$

$$Specificity = F_0(\alpha) \tag{4}$$

$$Precision = \frac{\pi_1(1-F_1(\alpha))}{\pi_1(1-F_1(\alpha))+\pi_0(1-F_0(\alpha))} \tag{5}$$

$$AUC = \int_{x=0}^{1} TPR\left(FPR^{-1}(X)\right)dx \tag{6}$$

The ROC curve can be plotted with the false positive rate (FPR, which equals $1 - Specificity$) and true positive rate (TPR, i.e., $Recall$) tabulated as x and y axes, respectively, which represent the tradeoffs between FPR and TPR. The area under the ROC curve is the AUC, which can effectively assess the generalization ability of the model. The closer the ROC curve is to the upper left corner, the larger the AUC value, indicating a stronger generalization ability of the prediction method.

The AP is defined in Eq. (2), and the other three economic metrics are computed as follows:

$$TB = TB_0 + TB_1 = B_0\pi_0N_{test}F_0(\alpha) + B_1\pi_1N_{test}\left(1 - F_1(\alpha)\right) \tag{7}$$

$$TC = TC_0 + TC_1 = C_0\pi_0N_{test}\left(1 - F_0(\alpha)\right) + C_1\pi_1N_{test}F_1(\alpha) \tag{8}$$

$$NP = TB - TC \tag{9}$$

4.3 Experimental design and parameters tuning

In the experimental design, we attempt to perform a comprehensive analysis of META-

DES-AAP and explore the impact of important factors on the TSP and AP. First, to verify the superiority of the proposed method for dynamic ensemble selection, META-DES-AAP is compared with several static mainstream machine learning methods, including single classifiers (DT, KNN, and others) and ensemble learning classifiers (RF, GBDT, and others), in terms of machine learning metrics. Second, to verify that the proposed method can better help companies maximize AP, META-DES-AAP is compared with several mainstream machine learning methods in terms of economic metrics. Finally, to analyze the impact of important factors on the TSP and AP, a statistical analysis of some important factors, including bank customer information, last contact information, socio-economic information, and other information, is performed visually.

The experiments were conducted on an experimental device with Intel CoreTM i7-8656 Processor @2.40GHZ, Windows10, 16G operating system. Python (version 3.6) and Python third-party libraries (numpy, pandas, and others) were used for modeling. To better evaluate the performance of each method involved in our experiments, the entire experimental process is repeated 20 times independently, and the average result of 20 times for each metric is used to evaluate the performance of each method.

The parameters that need to be predetermined in META-DES-AAP are listed in Table 3. In the initial pool of heterogeneous base classifiers, 30 base classifiers are included, of which the numbers of DT, RF, GBDT, LightGBM, LR, and XGBoost are all set at 5. The economy-related parameters can be set according to different companies' situations, and the values of relevant profits and costs in this study are pre-set, as shown in Table 3. The most suitable parameter values for the evolutionary parameters in NSGA-II after performing extensive experiments are also summarized in the table. Finally, the average optimal number of heterogeneous base classifiers in the accuracy- and profit-based optimal base classifiers pool is listed in Table 4. Compared with the 30 base classifiers in the original pool, 15 base classifiers constitute the accuracy- and profit-based optimal sub-pool, which greatly reduces the model's complexity. Furthermore, compared with the dynamic ensemble selection on the original pool, the accuracy and AP are greatly improved in the accuracy- and profit-based optimal sub-pool, where the accuracy and AP of the predictive model built on the accuracy- and profit-based optimal sub-pool are 1.6% and 4.7% higher than those built on the original pool (accuracy: 0.8779; AP:90.0096).

[Insert Tables 3 and 4 here]

4.4 Experimental results

4.4.1 Performance among META-DES-AAP and mainstream methods

Previous studies on marketing sales prediction based on machine learning (Salminen et al., 2019) have shown the performance of RF, DT, KNN, naïve Bayes (NB), LR, and SVM studied extensively. In addition, adaptive boosting (Adaboost), XGBoost, LightGBM, and GBDT are popular in various fields because of their superior performance. In this subsection, we compare and analyze the performance of META-DES-AAP with the above ten mainstream static machine learning methods in terms of the five machine learning metrics. The results are shown in Table 5.

[Insert Table 5 here]

For single classifiers, KNN and LR perform better than NB and DT, and SVM has the worst performance among them according to their performance of five metrics. For ensemble classifiers, their overall performance in terms of five metrics has an absolute advantage over the single classifiers, where META-DES-AAP has the best performance, followed by XGBoost, and RF is the worst of all ensemble classifiers. META-DES-AAP is 2.12%, 3.39%, and 2.14% higher than XGBoost in terms of accuracy, recall, and AUC, respectively. In addition, META-DES-AAP performs less well in terms of specificity and precision than RF. However, compared with the performance of methods in terms of precision and specificity, marketers may pay more attention to the model's performance on recall because recall reflects the percentage of all successful telemarketing campaigns correctly predicted by the model. META-DES-AAP has the highest recall among all the compared machine learning methods, with a value of 92.62%, further proving the superiority of META-DES-AAP.

4.4.2 Economic performance among META-DES-AAP and mainstream methods

For the predictive model in the field of marketing and sales, it should not only achieve high accuracy but also bring a large AP. In this subsection, we conduct extensive experiments to verify that META-DES-AAP performs the best in telemarketing over the ten mainstream machine learning methods mentioned in Subsection 4.4.1 in terms of economic metrics. The experimental results are summarized in Table 6 and analyzed as follows.

[Insert Table 6 here]

META-DES-AAP achieves the maximum TB, NP, and AP compared to other machine learning methods, where the values of TB, NP, and AP are 65980, 62200, and 94.2424, respectively. Meanwhile, META-DES-AAP reaches the minimum TC, with a value of 3780, compared with other machine learning methods. In addition to META-DES-AAP, XGBoost performs the best, while SVM performs the worst among the compared machine learning methods for TB, NP, AP, and TC. Therefore, we compare the performance of META-DES-AAP and XGBoost for TB, TC, NP, and AP, respectively. In terms of TB, NP, AP, META-DES-AAP is 2260, 3450, and 5.227 higher than XGBoost, respectively. For TC, META-DES-AAP is 1190 lower than that of XGBoost. The above analysis demonstrates that META-DES-AAP can bring greater AP to banks and has a higher application value.

4.5 Impact of important factors on the TSP and AP

Note that verifying the superior performance of the proposed method is not sufficient to support marketers in making actions. Thus, in this subsection, we demonstrate the important factors affecting the trends of the TSP and AP. In practice, marketers are more interested in knowing the relationship between TSP and AP obtained from META-DES-AAP and the values of important factors. In this way, marketers can further understand *what* important factors affect the TSP and AP, *how* to change the important factors affecting the trend of TSP and AP, and reduce the average probability of telemarketing failure and maximize AP, which can better guide them in developing telemarketing strategies. Therefore, we attempt to explore the factors influencing telemarketing on the TSP and AP given by META-DES-AAP and visually display them.

The important factors we consider include bank customer information, last contact information, socio-economic information, and other information. The results are shown in Figures 5-8. In Figures 5-8, the horizontal axis represents the values or attributes of different factors, and the left (resp., right) vertical axis represents the average predicted probability of telemarketing (resp., AP). The types of all factors include discrete-type and continuous-type. For discrete-type factors (job, education, etc.), we use pink (resp., dark blue) bars to denote the average predicted probability of telemarketing sales success for time deposits (resp., failure). For continuous-type factors (age, duration, etc.), we use the red (resp., blue) smooth line to denote the TSP (resp., AP) and the red (resp., blue) dotted line to denote the corresponding overall trend line. In what follows, we formally analyze the impact of the factors we consider.

(1) *Impact of bank customer information*

The relationship between the four factors belonging to the bank customer information, i.e., “type of job”, “education degree”, “marital status”, and “age”, and the average predicted probability and AP are displayed in Figure 5 (a)-(d).

As can be seen from Figure 5 (a), customers with job type of “retired”, “administrator”, “student”, and “unemployed” have higher TSP and AP, while customers with job types of “blue-collar” and “manager” have lower TSP and AP.

Figure 5 (b) indicates that the higher the education level of the customer, the higher the

TSP. It is worth noting that when the education level is raised from "basic nine years" to "professional course", AP increases significantly.

As for the factor "marital status" depicted in Figure 5 (c), there is no significant difference in the TSP and AP between customers with the marital status of "divorced" and "married", but customers with the marital status of "single" have a higher TSP and AP in comparison.

From Figure 5 (d), we can observe that for the factor "age", both TSP and AP tend to decrease before 43-year-old and then increase as the age increases, which is consistent with the above analysis that retired or single customers would have a better TSP and AP. This observation is also consistent with the result found in Guo et al. (2020) that marketers are more likely to promote products to customers older than 55 years.

[Insert Figure 5 about here]

(2) *Impact of last contact information*

The relationship between the three factors belonging to the last contact information, i.e., the "last contact month of year", "last contact day of week", and "last contact duration" and the average predicted probability and AP are displayed in Figure 6 (a)-(c).

As can be seen from Figure 6 (a), marketing managers can obtain a maximum TSP of approximately 0.8 and an AP of approximately 170 for the last customer contact in "Sep", "Oct" and "Dec".

From Figure 6 (b), we can observe that there is no significant difference in the TSP and AP in terms of the factor "last contact day of week". Only the AP on "Tue" will be slightly more than on the other days of the week.

Figure 6 (c) indicates that the TSP and AP increase significantly, as the last contact duration becomes longer, consistent with the result found in Moro et al. (2014) that more time spent on calls in telemarketing will increase the probability of telemarketing sales success for time deposits. For customers with the last contact duration of calls after approximately 430 seconds, the overall trend lines for both the TSP and AP show a leveling off (as indicated by the red and blue dotted lines). This may be because a long last contact duration means that the customer is more interested in the product being promoted.

[Insert Figure 6 about here]

(3) *Socio-economic information*

The relationship between the four factors belonging to the socio-economic information, i.e., “Euribor 3-month rate”, “consumer confidence index”, “employment variation rate” and “number of employees,” are displayed in Figure 7 (a)-(d).

Note that the factor “Euribor 3-month rate” reflects the tightness of funds in the market in the short term. In general, the higher the value of the “Euribor 3-month rate”, the tighter the funds in the market. In Figure 7 (a), the TSP presents a significant downward trend, whereas the AP first presents a downward trend, followed by an upward trend, with an increase in the Euribor rate. Specifically, the AP tends to increase when the Euribor rate is greater than 4.034. This observation is also consistent with the result obtained in Guo et al. (2020) that there may be a negative correlation between the Euribor rate and the probability of long-term deposit subscription. Factor “consumer confidence index” can be regarded as a measure of consumers' attitudes toward the economy. The “consumer confidence index” of the test samples is negative, indicating that consumers are pessimistic about their investments. Nonetheless, Figure 7 (b) indicates that the TSP will improve slightly, and the AP will remain stable as the value of the “consumer confidence index” increases. Hampson et al. (2020) pointed out that the consumer confidence index would impact consumers’ behavior, which was fully reflected here.

From Figure 7 (c)-(d), we can observe that for the factor “employment variation rate” (Figure 7 (c)) and “number of employees” (Figure 7 (d)), the TSP and AP show a sharp decline trend, with an increase of employ variation rate and the number of employees. It also suggests that customers are less likely to order long-term deposits in an environment with a high employment variation rate or high employment.

[Insert Figure 7 about here]

(4) *Impact of other information*

The relationship between the three factors belonging to the other information, i.e., "number of contacts performed during this campaign" and "number of contacts performed before this campaign", and the average predicted probability and the AP are displayed in Figure 8 (a)-(b).

As shown in Figure 8 (a), the TSP and AP show a steep declining trend as the number of contacts made by the marketer for this customer during this campaign increased. Conversely, Figure 8 (b) indicates that the TSP and AP have a sharp upward trend, as the number of contacts the marketer made for the customer before the campaign increased.

These observations are interesting. The number of times a marketer contacts a customer has an impact on customers' emotions. These emotions, in turn, determine customers' behavior and purchase decisions (Babin & Attaway, 2000; Machleit & Eroglu, 2000). During a telemarketing campaign, repeated contact with customers can help telemarketing, but it may also have adverse effects. This is probably because repeated contact during the campaign makes customers prone to adverse and boring emotions, which greatly reduces the probability of telemarketing sales success for time deposits. In contrast, contacting customers several times before the telemarketing campaign positively impacts TSP and AP. Therefore, the marketer should try to contact the customer before the campaign rather than during the campaign.

[Insert Figure 8 about here]

5. General discussion

5.1 Theoretical implications

Machine learning has a strong potential that surpasses humans on complex problems and prediction (Antons & Breidbach, 2018) because by leveraging the insights of advanced machine learning methods, companies can identify campaigns that will be marketed successfully in the future and develop response strategies based on expectations (Kumar et al., 2020). This study constructs a predictive model with high accuracy and AP that predicts whether telemarketers will be successful in selling time deposits. Using a retail bank dataset from Portugal, we compare the performance of the proposed method META-DES-AAP and advanced machine learning methods and attempt to explain the impact of important factors on TSP and AP trends derived from META-DES-AAP. These comparisons and explanations provide several important theoretical and managerial implications.

First and foremost, much of the prior literature related to bank telemarketing sales prediction using machine learning methods mainly focus on obtaining highly accurate marketing predictive models (Moro et al., 2014; Mei et al. 2016; Yan et al., 2020). Few studies have taken benefits into account in the model construction and transformed the marketing predicted decisions into benefits. This study is a pioneering attempt to consider both accuracy and AP in constructing telemarketing predictive models. For each query telemarketing campaign, META-DES-AAP selects competent base classifiers from the accuracy- and profit-based optimal sub-pool and integrates them dynamically, resulting in telemarketing sales prediction results showing both higher accuracy and larger AP. The results indicate that META-DES-AAP outperforms the mainstream machine learning methods in several dimensions, and the accuracy of prediction using META-DES-AAP reaches 89.39%, with an AUC of 89.44% and an AP of 94.2424.

Second, as shown in Table 1, there is limited literature investigating how to explain the marketing predicted results of the predictive models and how to use marketing predictive models to guide the development of marketing strategies, making it challenging for the predictive model to guide banks in their businesses. In contrast to prior literature, the results of this study are a trigger point to explain the impact of important factors on the TSP and AP for the proposed META-DES-AAP. We reveal how important factors (i.e., bank customer information, last contact information, socio-economic information, and other information) affect the trends of the TSP and AP for META-DES-AAP, which makes several distinctive theoretical contributions to the extant literature on the development of telemarketing strategy from the predictive model.

Finally, not every machine learning method can have excellent predicted performance and achieve satisfactory and stunning results. The most existing predictive model for marketing topics still suffered from low accuracy (Moro et al., 2014; Mei et al., 2016). Since the classifier that makes the best prediction for each test sample may differ, the static classification methods may not achieve the best performance on a given test sample (Cruz et al., 2018). In contrast, the proposed META-DES-AAP bridges this gap. META-DES-AAP combines a multi-objective optimization algorithm and the framework of dynamic ensemble selection using meta-training, which improves the predictive performance of telemarketing sales prediction. The results obtained in this study further demonstrate that META-DES-AAP outperforms the mainstream machine learning methods in business research (Coussement & De Bock, 2013; Liang et al., 2020; Tsai et al., 2021). Moreover, the results indicate that the performance of the ensemble classifiers (resp., dynamic classification methods) is generally superior to that of single classifiers (resp., static classification methods), which is consistent with previous studies on machine learning (Hou et al., 2020; Hwang et al., 2020; Albrecht et al., 2021). Thus, our study enriches the literature on applying the dynamic ensemble selection method in the business field.

5.2 Managerial implications

From a practical point of view, our study collectively contributes to advance bank telemarketing practices and offers insights for telemarketing managers to develop telemarketing strategies. First, banks can use our proposed method to construct their dynamic ensemble selection models for telemarketing sales prediction. Specifically, the profit and cost parameters in the model can be set by the bank according to their situations and demands, thus assisting them in maximizing their AP. Moreover, the way machine learning methods work can affect corporate and individual decisions, and the lack of transparency due to depth can lead to tactical and strategic errors (De Bruyn et al., 2020; Mustak et al., 2020). Consequently, in Subsection 4.5, we investigate the impact of customer information, last contact information, socio-economic information, and other information on TSP and AP derived from META-DES-AAP. The results of this study can further enable telemarketers to understand the investment intentions of their customers so that they can develop better telemarketing strategies.

Second, based on our findings, younger and retired customers tend to save for future expenses. Moreover, customers with higher levels of education are more financially literate and likely to invest their income in the market. Thus, we suggest that an experienced bank marketer should target older or younger customers and highly educated customers rather than middle-aged ones, consistent with the findings of Guo et al. (2020). In addition, at the end of the year, customers are more inclined to save for the coming year's expenses. This means that an experienced bank marketer should make one last contact with a customer at the end of the year. Furthermore, our findings indicate that marketing managers organize telemarketers to expand their customer groups during periods when the socio-economic environment is characterized by the following four characteristics: low Euribor rate, high customer confidence index, low employment variation rate, and low employment.

Third, META-DES-AAP suggests that when the last contact duration between the marketer and the customer on the telephone is greater than about 430 seconds, the customer's willingness to subscribe to time deposits at this time is very strong. Thus, this finding suggests that an experienced telemarketer should introduce products they are interested in, which will have a high probability of telemarketing sales success for time deposits. In addition, contacting customers several times before a campaign may generate positive emotions and thus increase the probability of success, whereas contacting them several times during the campaign may generate adverse emotions. This means that an experienced marketer should contact customers before the event rather than during the event.

Finally, although state governments and regulators have introduced telemarketing regulations and laws to the implementation of bank telemarketing campaigns (Rita, 1995), banks that use aimless and widely distributed telemarketing strategies may still cause disturbances to some customers who are not interested in the time deposit (Yan et al., 2020). In actual bank telemarketing campaigns, banks can use the proposed telemarketing sales prediction method META-DES-AAP to identify target customers and predict the success probability for future telemarketing campaigns accurately, thus reducing the disturbance to those customers who are not interested in the subscription of time deposits. Furthermore, bank marketers can implement the telemarketing strategies extracted from this study in their campaigns to judge customers' intention to subscribe to time deposits. Marketers can take certain measures to reduce discomfort for customers with low intention in telemarketing campaigns to reduce customer complaints.

5.3 Limitation and future research directions

Although this study provides several meaningful insights into the marketing field, there are some concerns and limitations. Some interesting topics can be explored in the future based on these limitations. First, although we try to reduce the computational time in training the model, i.e., developing efficient methods to choose the optimal base classifiers, but META-DES-AAP has, on average, a longer training time than other machine learning models. This is obvious since superior performance is usually at the cost of a longer computational time, which is also highlighted by Coussement & Bock (2013). To reduce the training of our developed method, we attempted to apply the parallel computing method and tested the effectiveness of the proposed method and marketing strategy in bank telemarketing sales of time deposits. In addition, deep learning is recognized for its superior predicted performance compared to machine learning methods (Ebadi et al., 2016; Shrestha et al., 2021; Zhu et al., 2021); hence, deep learning methods can be applied to bank telemarketing sales prediction.

Second, clearly, the benefits gained by the bank are directly related to the amount of customer time deposits; the more the time deposits, the greater the benefits gained by the bank. However, since the dataset of banks involves much private information related to customers, the data collected in this study did not contain data related to the deposit amount for each customer who chooses to order a time deposit, which limits us to construct a more accurate telemarketing sales predictive model. It is interesting to investigate the impact of different time deposit amounts for different customers with successful telemarketing on bank profits in our model. Hence, it is expected that more banks will disclose the desensitized and relevant datasets to attract more machine learning researchers to investigate bank marketing to provide more targeted marketing suggestions.

Finally, META-DES-AAP is constructed on the balanced dataset, which follows that META-DES-AAP may perform poorly on another unbalanced, the performance of which usually depends on the proportion of data imbalance. In addition, we use bank telemarketing data from Portugal, making the applicability of the proposed method to telemarketing sales in other countries to be further validated. One valuable improvement of our study is to obtain data on more types of marketing, not limited to bank telemarketing, such as online bank marketing and bank service marketing, and verify our proposed method on these data.

6. Conclusion

This study uses machine learning methods to help banks improve the probability of success of bank telemarketing and maximize the AP brought by telemarketing. To consider both accuracy and AP, we propose a novel model, META-DES-AAP, to predict whether bank telemarketing sales of time deposits would be successful. In META-DES-AAP, a multi-objective optimization algorithm is designed to find an accuracy- and profit-based optimal sub-pool based on meta-training, and a dynamics-based base classifiers integration method is used to integrate base classifiers and make a prediction based on future telemarketing data. Extensive experiments on real-world bank telemarketing data prove the superiority of META-DES-AAP in bank telemarketing sales prediction, which has the best accuracy and achieves maximum AP. Moreover, the impact of four types of information (customer information, last contact information, socio-economic information, and other information) on TSP and AP are analyzed, supporting many marketing strategies. The results of this study can serve as a stepping stone for future research on telemarketing sales predictions.

Acknowledgments

This paper was supported in part by the National Natural Science Foundation of China under grant numbers 71871148, 71971041, and 71501024; and by the Outstanding Young Scientific and Technological Talents Foundation of Sichuan Province under grant number 20JCQN0281.

Appendix

Table A1. The description of bank telemarketing dataset.

Type	No	Variable name	Description
Bank customer information	1	age	age
	2	job	type of job
	3	marital	marital status
	4	education	education degree
	5	default	has credit in default?
	6	housing	has a housing loan?
	7	loan	has personal loan?
Last contact information	8	contact	contact communication type
	9	month	last contact month of the year
	10	day_of_week	last contact day of the week
	11	duration	last contact duration
Socio-economic information	12	emp.var.rate	employment variation rate (quarterly indicator)
	13	cons.price.idx	consumer price index (monthly indicator)
	14	cons.conf.idx	consumer confidence index (monthly indicator)
	15	euribor3m	Euribor 3-month rate (daily indicator)
	16	nr. employed	number of employees (quarterly indicator)
Other information	17	campaign	number of contacts performed during this campaign and for this customer
	18	pdays	number of days that passed by after the customer was last contacted from a previous campaign
	19	previous	number of contacts performed before this campaign and for this customer
	20	poutcome	the outcome of the previous marketing campaign

Target	21	y	has the customer subscribed to a term deposit?

Reference

Adyyński, P., Bikowski, K., & Gawrysiak, P. (2019). Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications*, *134*, 28–35.

Albrecht, T., Rausch, T. M., & Derra, N. D. (2021). Call me maybe: Methods and practical implementation of artificial intelligence in call center arrivals' forecasting. *Journal of Business Research*, *123*. DOI：10.1016/j.jbusres.2020.09.033

Alemán Carreón, E. C., Nonaka, H., Hentona, A., & Yamashiro, H. (2019). Measuring the influence of mere exposure effect of TV commercial adverts on purchase behavior based on machine learning prediction models. *Information Processing & Management*, *55*(4), 1339–1355.

Antons, D., & Breidbach, C. F. (2018). Big data, big insights? Advancing service innovation and design with machine learning. *Journal of Service Research*, *21*(1), 17–39.

Babin, B. J., & Attaway, J. S. (2000). Atmospheric affect as a tool for creating value and gaining share of customer. *Journal of Business Research*, *49*(2), 91–99.

Bose, I., & Mahapatra, R. K. (2001). Business data mining—A machine learning perspective. *Information & Management*, *39*(3), 211–225.

Breiman, L. (2001). Random forest. *Machine Learning*, *45*, 5–32.

Cavalin, P. R., Sabourin, R., & Suen, C. Y. (2013). Dynamic selection approaches for multiple classifier systems. *Neural Computing & Applications*, *22*(3–4), 673–688.

Chatterjee, S., Goyal, D., Prakash, A., & Sharma, J. (2020). Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. *Journal of Business Research*, S0148296320307049.

Chen, D. Q., Preston, D. S., & Swink, M. (2015). How the use of big data analytics affects value creation in supply chain management. *Journal of Management Information Systems*, *32*(4), 4–39.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *Mis Quarterly*, *36*(4), 1165–1188.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 785–794.

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, *66*(9), 1629–1636.

Cruz, R. M. O., Sabourin, R., & Cavalcanti, G. D. C. (2017). META-DES.Oracle: Meta-learning and feature selection for dynamic ensemble selection. *Information Fusion*, 38, 84–103.

Cruz, R. M. O., Sabourin, R., & Cavalcanti, G. D. C. (2018). Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, *41*, 195–216.

Cruz, R. M. O., Sabourin, R., Cavalcanti, G. D. C., & Ing Ren, T. (2015). META-DES: A dynamic ensemble selection framework using meta-learning. *Pattern Recognition*, *48*(5), 1925–1935.

Cui, G., & Man, L. W. (2004). Implementing neural networks for decision support in direct marketing. *International Journal of Market Research*, *46*(2), 235-254+263.

De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, K. U., & Von Wangenheim, F. (2020). Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, *51*.

Dincer, H., Hacioglu, U., Tatoglu, E., & Delen, D. (2019). Developing a hybrid analytics approach to measure the efficiency of deposit banks. *Journal of Business Research*, *104*, 131–145.

Ebadi Jalal, M., Hosseini, M., & Karlsson, S. (2016). Forecasting incoming call volumes in call centers with recurrent Neural Networks. *Journal of Business Research*, 69(11), 4811–4814.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. V. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*, 335–

343.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Ghaderi Zefrehi, H., & Altınçay, H. (2020). Imbalance learning using heterogeneous ensembles. *Expert Systems with Applications*, *142*, 113005.

Guo, M., Zhang, Q., Liao, X., Chen, F. Y., & Zeng, D. D. (2020). A hybrid machine learning framework for analyzing human decision-making through learning preferences. *Omega*, 102263.

Hampson, D. P., Gong, S., & Xie, Y. (2020). How consumer confidence affects price conscious behavior: The roles of financial vulnerability and locus of control. *Journal of Business Research*, S0148296320306937.

Hanssens, D. M., Rust, R. T., & Srivastava, R. K. (2009). Marketing strategy and Wall Street: Nailing down marketing's impact. *Journal of Marketing*, *73*(6), 115–118.

Hou, W., Wang, X., Zhang, H., Wang, J., & Li, L. (2020). A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment. *Knowledge-Based Systems*, *208*, 106462.

Hwang, S., Kim, J., Park, E., & Kwon, S. J. (2020). Who will be your next customer: A machine learning approach to customer return visits in airline services. *Journal of Business Research*, *121*, 121–126.

Jacobs, B. J. D., Donkers, A. C. D., & Fok, D. (2016). Model-based purchase predictions for large assortments. *Erim Report*, *35*(3), 389–404.

Jimenez-Marquez, J. L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Towards a big data framework for analyzing social media content. *International Journal of Information Management*, *44*, 1–12.

K Deb, A Pratap, S Agarwal, & T Meyarivan. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*. *6*(2): 182–197.

Ke, G., Meng, Q., Finely, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems 30 (NIP 2017)*.

Kim, M. C., & Mcalister, L. M. (2011). Stock market reaction to unexpected growth in marketing expenditure: Negative for sales force, contingent on spending level for advertising. *Journal of Marketing*, *75*(4), 68–85.

Ko, A. H. R., Sabourin, R., & Britto, Jr., A. S. (2008). From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, *41*(5), 1718–1731.

Kotler, P. T. (2016). A framework for marketing management. *Sloan Management Review*, *32*(2), 94–104.

Kumar, V., Ramachandran, D., & Kumar, B. (2020). Influence of new-age technologies on marketing: A research agenda. *Journal of Business Research*.

Levitt, S. D. (2016). Bagels and Donuts for sale: A case study in profit maximization. *Research in Economics*, *70*(4), 518–535.

Liang, D., Tsai, C.-F., Lu, H.-Y. (Richard), & Chang, L.-S. (2020). Combining corporate governance indicators with stacking ensembles for financial distress prediction. *Journal of Business Research*, *120*, 137–146.

Lismont, J., Vanthienen, J., Baesens, B., & Lemahieu, W. (2017). Defining analytics maturity indicators: A survey approach. *International Journal of Information Management*, *37*(3), 114–124.

Machleit, K. A., & Eroglu, S. A. (2000). Describing and Measuring Emotional Response to Shopping Experience. *Journal of Business Research*, *49*.

Markovitch, D. G., Huang, D., & Ye, P. (2020). Marketing intensity and firm performance: Contrasting the insights based on actual marketing expenditure and its SG&A proxy. *Journal of Business Research*, *118*.

Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems*, *51*(4), 782–793.

Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2018). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, *281*(3).

Maulana, A. E., & K. Nurulfirdausi. (2015). Permissive, aggressive or apathetic? Indonesian telemarketing customer. *Procedia - Social and Behavioral Sciences*, *169*:69-74.

Mcafee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review, 90*(10), 60–68.

Mei, R., Xu, Y., & Wang, G. (2016). Telephone marketing forecast of bank time deposits based on LASSO-SVM model. *Statistics Application*, *05*(03), 289–298.

Méndez-Suárez, M., & Crespo-Tejero, N. (2020). Why do banks retain unprofitable customers? A customer lifetime value real options approach. *Journal of Business Research*, *122*, 621–626.

Mitchell, T. M. (1997). Machine learning. *McGraw-Hill*.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31.

Moro, S., Cortez, P., & Rita, P. (2018). A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. *Expert Systems*, *35*(3), e12253.1-e12253.13.

Mustak, M., Salminen, J., Plé, L., & Wirtz, J. (2020). Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research*, S0148296320307165.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21.

Orriols-Puig, A., Francisco J. Martinez-Lopez, Casillas, J., & Lee, N. (2013). A soft-computing-based method for the automatic discovery of fuzzy rules in databases: Uses for academic research and management support in marketing. *Journal of Business Research*, *66*(9), 1332–1337.

Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*, *118*, 33–45.

Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. C. (2018). Correlation analysis of performance measures for multi-label classification. *Information Processing & Management*, *54*(3), 359–369.

Reis, C., Ruivo, P., Oliveira, T., & Faroleiro, P. (2020). Assessing the drivers of machine learning business value. *Journal of Business Research*, *117*, 232–243.

Rita, M., C.. The state of telemarketing regulation in the states. *Journal of Direct Marketing*, *9*(4), 76-83.

Ritsema, H., Piëst, B. (1990). Telemarketing: The case for (self) regulation? *European Management Journal*, *8*(1):63-66.

Ren, X., Cao, J., Xu, X., & Gong, Y. (2020). A two-stage model for forecasting consumers' intention to purchase with e-coupons. *Journal of Retailing and Consumer Services*, 102289.

Salminen, J., Yoganathan, V., Corporan, J., Jansen, B. J., & Jung, S.-G. (2019). Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type. *Journal of Business Research*, *101*, 203–217.

Shrestha, Y. R., Krishna, V., & von Krogh, G. (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Journal of Business Research*, *123*, 588–603. DOI: 10.2139/ssrn.3701592

Tsai, C.-F., Sue, K.-L., Hu, Y.-H., & Chiu, A. (2021). Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. *Journal of Business Research*, *130*, 200–209. DOI: 10.1016/j.jbusres.2021.03.018..

Woloszynski, T., Kurzynski, M., Podsiadlo, P., & Stachowiak, G. W. (2012). A measure of competence based on random classification for dynamic ensemble selection. *Information Fusion*, *13*(3), 207–213.

Xia, Y., Zhao, J., He, L., Li, Y., & Niu, M. (2020). A novel tree-based dynamic heterogeneous ensemble method for credit scoring. *Expert Systems with Applications*, *159*, 113615.

Yan, C., Li, M., & Liu, W. (2020). Prediction of bank telephone marketing results based on improved whale algorithms optimizing S_Kohonen network. *Applied Soft Computing*, *92*, 106259.

Zhang, Z.-L., Chen, Y.-Y., Li, J., & Luo, X.-G. (2019). A distance-based weighting framework for boosting the performance of dynamic ensemble selection. *Information Processing & Management*, *56*(4), 1300–1316.

Zhu, J. J., Chang, Y.-C., Ku, C.-H., Li, S. Y., & Chen, C.-J. (2021). Online critical review classification in response strategy and service provider rating: Algorithms from heuristic processing, sentiment analysis to deep learning. *Journal of Business*

Research, *129*, 860–877. DOI: 10.1016/j.jbusres.2020.11.007.
Zhu, Y., Zhou, L., Xie, C., Wang, G.-J., & Nguyen, T., V. (2019). Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics*, *211*, 22–33.

Vitae

Yi Feng is currently pursuing a Ph.D. degree in Management Science and Engineering from Sichuan University, Chengdu, China. Her current research interests include data mining, machine learning, smart marketing, and she is also interested in medical data mining and health care. She has participated in the journal review of Information Sciences, Computers & Industrial Engineering and soft computing as an assistant.

Dujuan Wang received the B.S. and M.S. degrees in Computer Science and Technology, and the Ph.D. degree in Management Science and Engineering from the Dalian University of Technology (DUT), Dalian, China. She is currently a professor at Business School, Sichuan University, Chengdu, China. Her research interests are in machine learning, service operation management and optimization, and artificial intelligence. She has published more than 40 papers in various international journals including Naval Research Logistics, Omega, European Journal of Operational Research, IEEE Transactions on SMC, International Journal of Production Economics, Computers & Operations Research, Knowledge-Based Systems, etc.

Yunqiang Yin received a Ph.D. degree from Beijing Normal University, Beijing, China, in 2009. He is currently a Professor in the School of Management and Economics, University of Electronic Science and Technology of China, Chengdu. His research interests are in production and operations management, medical operations management, and machine learning. He has published more than 80 papers in various international journals including Naval Research Logistics, Omega, European Journal of Operational Research, International Journal of Production Research, Computers & Operations Research, Journal of Scheduling, IEEE Transactions on SMC, etc.

Lalitha Dhamotharan is a Lecturer of Entrepreneurship and Business Analytics at the University of Exeter Business School. She holds a Ph.D. in Econometrics with a focus on International Finance. Before joining the University of Exeter, she was a Postgraduate Researcher as well as Hub of All Things (HAT) Research and Policy Manager for an EPSRC funded project at the University of Warwick. She also has experience in teaching for the Innovation and Entrepreneurship course at the University of Warwick. Lalitha worked as a Senior Analyst dealing with research and policies issues on the various sector in South East Asia in Malaysia.

Figure

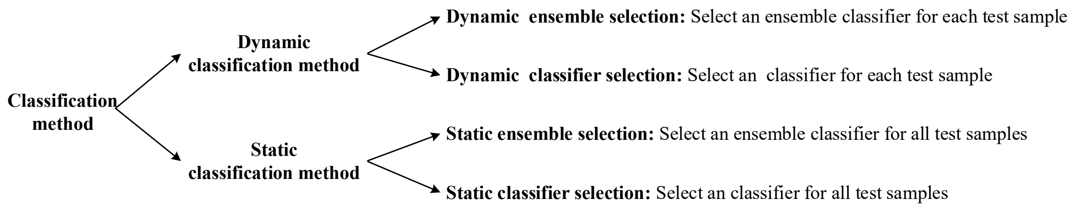


Figure 1. The framework of the classification method.

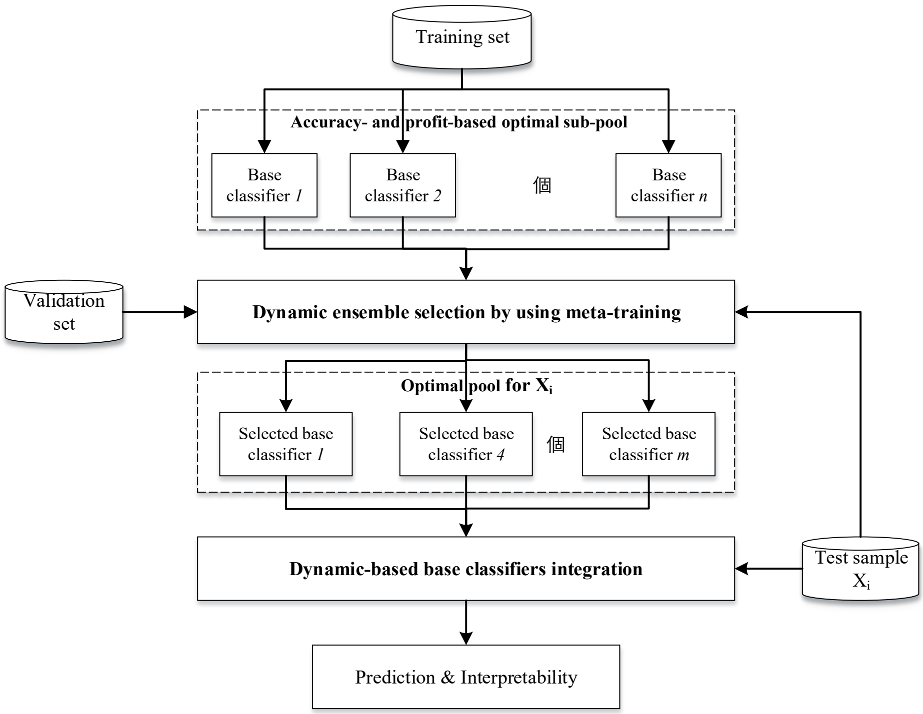


Figure 2. The framework of META-DES-AAP.

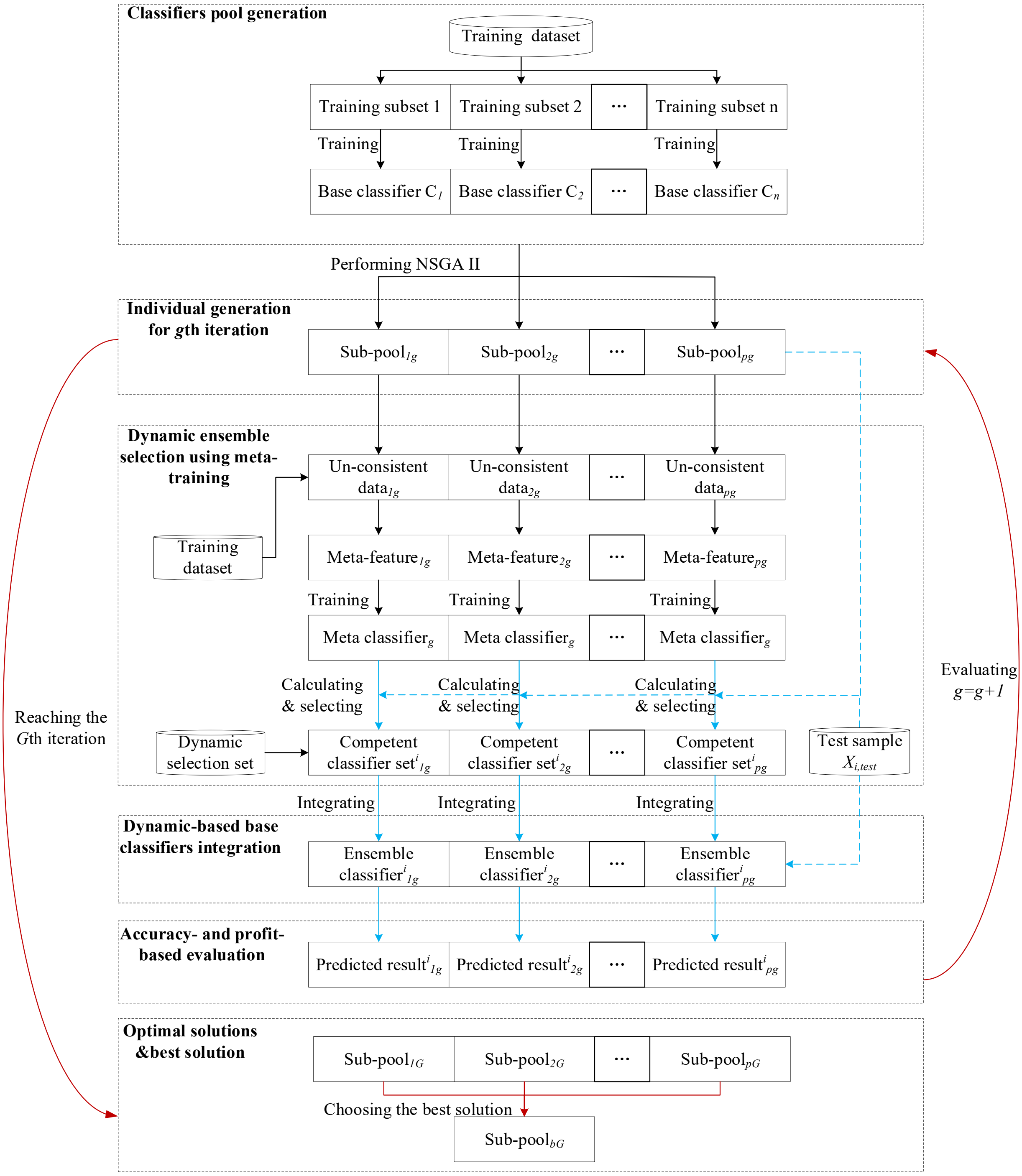


Figure 3. Multi-objective evolutionary process.

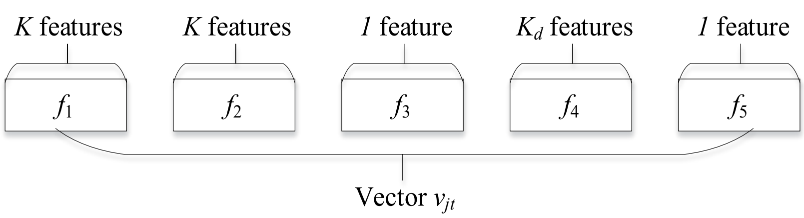
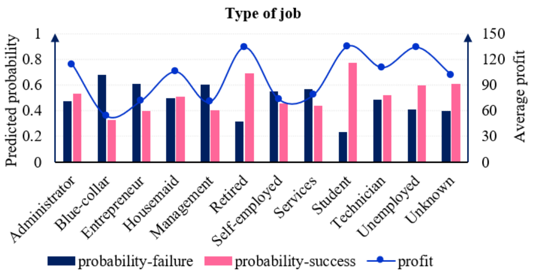
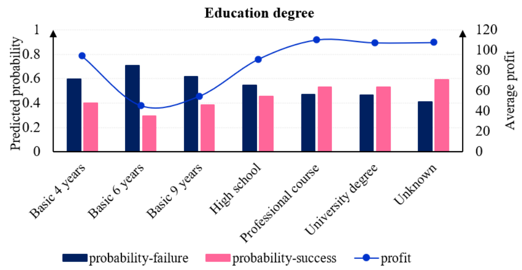


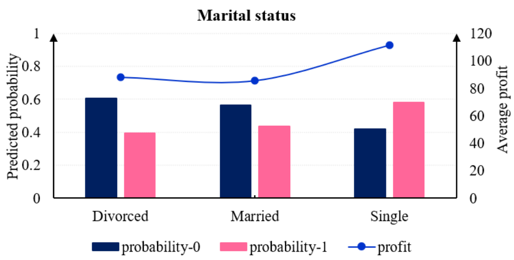
Figure 4. The structure of the feature vector.



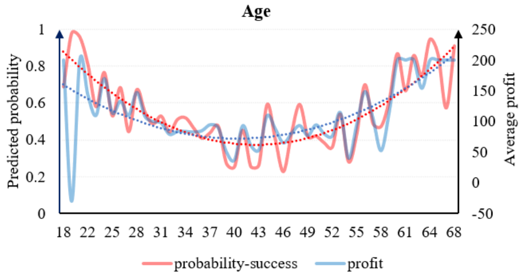
(a)



(b)

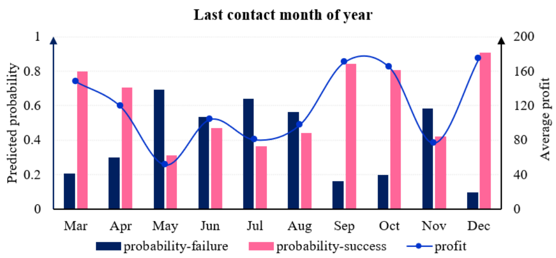


(c)

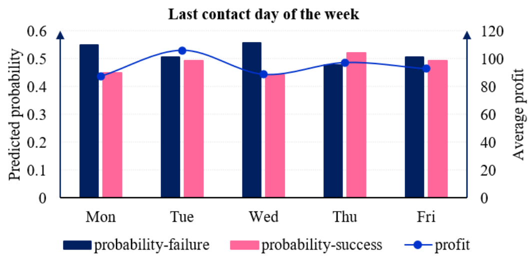


(d)

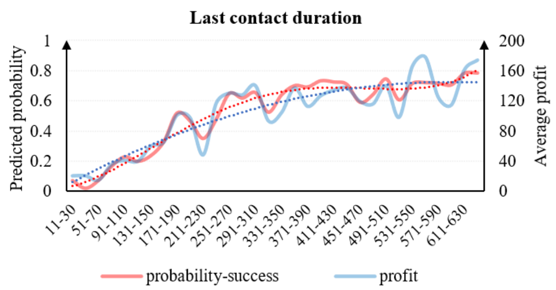
Figure 5. Relationship between bank customer information factors and average predicted probability and AP.



(a)



(b)



(c)

Figure 6. Relationship between last contact information factors and average predicted probability and AP.

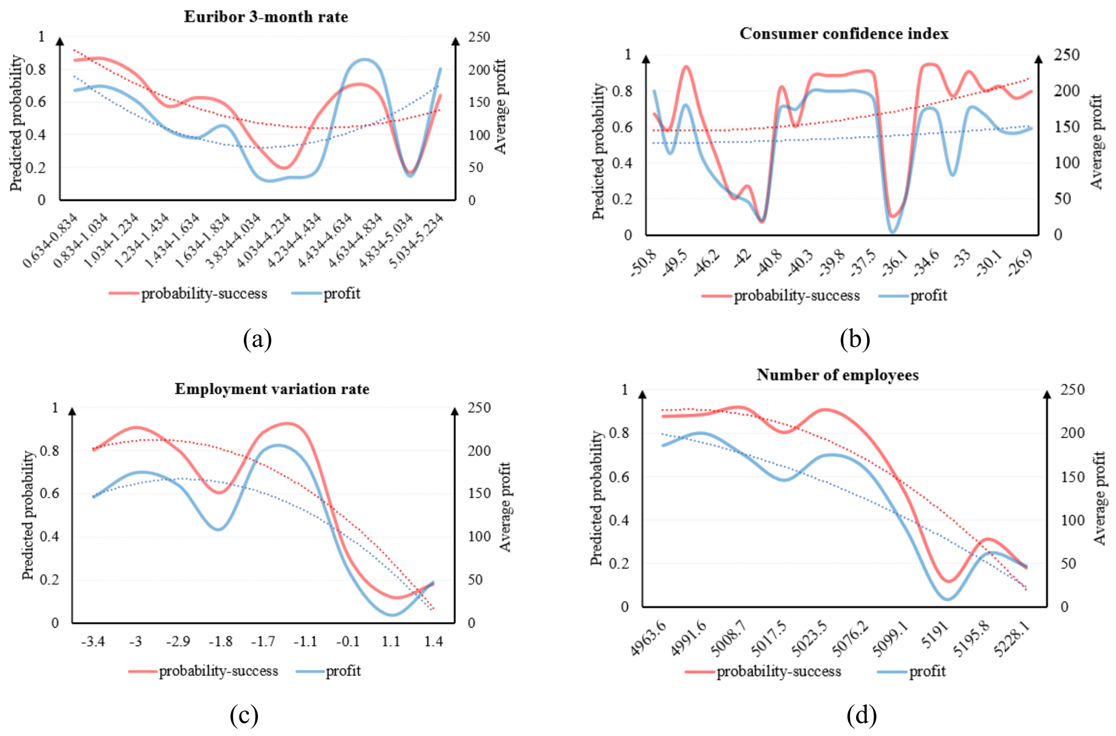


Figure 7. Relationship between socio-economic information factors and average predicted probability and AP.

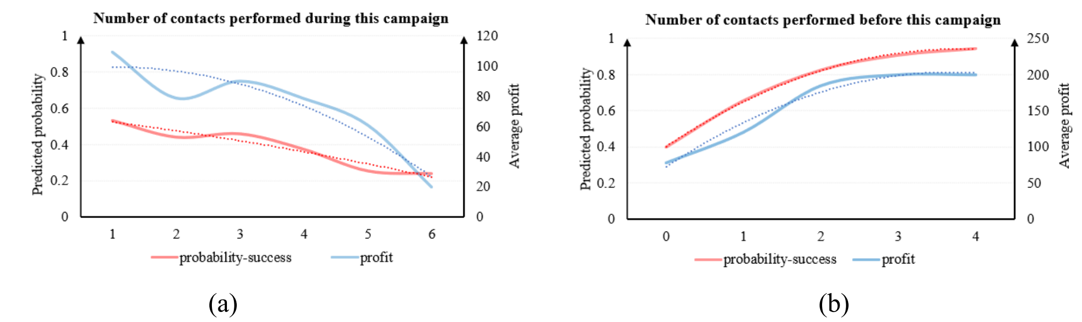


Figure 8. Relationship between other information factors and average predicted probability and AP.

Table

Table 1. Summary of machine learning research on marketing fields.

Literature	Problem	Method[a]	Type[b]	Accuracy[c]	AP[d]	Metrics[e]	Marketing strategy[f]
Adyyński et al. (2019)	Credit product marketing	DL, RF	Static	×	×	ML	×
Ren et al. (2020)	E-coupons marketing	C, IF, LR	Static	×	×	ML	×
Martínez et al. (2018)	Transactional marketing	LLR, GBDT, ELM	Static	×	×	ML	×
Alemán Carreón et al. (2019)	TV marketing sales	SVM, XGBoost, LR	Static	×	×	ML	×
Mei et al. (2016)	Telemarketing sales	SVM, Lasso	Static	×	×	ML	×
Yan et al. (2020)	Telemarketing sales	SKN	Static	×	×	ML	×
Moro et al. (2014)	Telemarketing sales	DT, SVM, NN	Static	×	×	ML	√
Moro et al. (2018)	Telemarketing sales	NNE	Static	×	×	ML	×
Our work	Telemarketing sales	META-DES-AAP	Dynamic	√	√	ML & E	√

Notes: [a] DL: deep learning; RF: random forests; C: cluster; LR: logistic regression; IF: isolation forest; LLR: logistic LASSO regression; ELM: extreme learning machine; GBDT: gradient tree boosting; SVM: support vector machine; XGBoost: an extreme gradient boosting; SKN: SKohonen network; DT: Decision tree; NN: Neural network; NNE: neural network ensemble. [b] Type of predictive model (static classification or dynamic classification method). [c] Accuracy is considered in the model construction. [d] Whether AP is considered in model construction. [e] Metrics for model evaluation (ML: machine learning metrics, E: economic metrics). [f] Extract interpretable information from predictive models and develop marketing strategies

Table 2. Confusion matrix with benefits and costs.

Actual label	Predicted label	
	Failed marketing	Successful marketing
Failed marketing	TN: $\pi_0 F_0(\alpha) N_{test}$ Benefit: $[P(0\|0) = B_0]$	FP: $\pi_0(1 - F_0(\alpha)) N_{test}$ Cost: $[P(1\|0) = C_0]$
Successful marketing	FN: $\pi_1 F_1(\alpha) N_{test}$ Cost: $[P(0\|1) = C_1]$	TP: $\pi_1(1 - F_1(\alpha)) N_{test}$ Benefit: $[P(1\|1) = B_1]$

Note: TN: True Negative; TP: True Positive; FN: False Negative; FP: False Positive

Table 3. Parameters of META-DES-AAP.

Information	Parameter	Value
Base classifiers parameters	Initial number of DT	5
	Initial number of RF	5
	Initial number of GBDT	5
	Initial number of LightGBM	5
	Initial number of LR	5
	Initial number of XGBoost	5
Economic parameters	Benefit B_0	20
	Benefit B_1	200
	Cost C_0	30
	Cost C_1	100
	a cut value α	0.5
Evolutionary parameters	Size of population	30
	Crossover probability	0.7
	Mutation probability	0.01
	Maximum number of iterations	50

Table 4. The optimal number of heterogeneous base classifiers.

Parameter	Value
The optimal number of DT	2
The optimal number of RF	2
The optimal number of GBDT	2
The optimal number of LightGBM	3
The optimal number of LR	3
The optimal number of XGBoost	3
Size of accuracy- and profit-based optimal sub-pool	15

Table 5. Performance overview of the different machine learning methods.

Method		Performance metric				
		Accuracy	Recall	Precision	Specificity	AUC
Single classifier	NB	0.7939	0.7938	0.7890	0.7940	0.7939
	DT	0.8182	0.8308	0.8060	0.8060	0.8184
	KNN	0.8348	0.8154	0.8439	0.8537	0.8346
	LR	0.8318	0.8246	0.8323	0.8388	0.8317
	SVM	0.7348	0.6892	0.7517	0.7791	0.7342
Ensemble classifier	RF	0.8515	0.8154	**0.8746**	**0.8866**	0.8510
	Adaboost	0.8455	0.8338	0.8495	0.8567	0.8452
	XGBoost	0.8727	0.8923	0.8555	0.8537	0.8730
	GBDT	0.8712	0.8862	0.8571	0.8567	0.8714
	LightGBM	0.8712	0.8800	0.8614	0.8627	0.8713
	META-DES-AAP	**0.8939**	**0.9262**	0.8674	0.8627	**0.8944**

Table 6. Economic performance overview of the different machine learning methods.

Method	Confusion matrix		Profit performance							
	TP	TN								
	FP	FN								
META-DES-AAP	301	289	**60200**	5780	1380	**2400**	**65980**	**3780**	**62200**	**94.2424**
	46	24								
RF	265	297	53000	**5940**	**1140**	6000	58940	7140	51800	78.4848
	38	60								
Adaboost	271	287	54200	5740	1440	5400	59940	6840	53100	80.4545
	48	54								
DT	270	270	54000	5400	1950	5500	59400	7450	51950	78.7121
	65	55								
XGBoost	290	286	58000	5720	1470	3500	63720	4970	58750	89.0152
	49	35								
KNN	265	286	53000	5720	1470	6000	58720	7470	51250	77.6515
	49	60								
NB	258	266	51600	5320	2070	6700	56920	8770	48150	72.9545
	69	67								
GBDT	288	287	57600	5740	1440	3700	63340	5140	58200	88.1818
	48	37								
LightGBM	286	289	57200	5780	1380	3900	62980	5280	57700	87.4242
	46	39								
LR	268	281	53600	5620	1620	5700	59220	7320	51900	78.6364
	54	57								
SVM	224	261	44800	5220	2220	10100	50020	12320	37700	57.1212
	74	101								