2020 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD) | 978-1-7281-5950-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICTMOD49425.2020.9380602

Random Forest-based Approach for Classifying Customers in Social CRM

Soumaya LAMRHARI
ENSIAS, Mohammed V University
Rabat, Morocco
soumaya_lamrhari@um5.ac.ma

Hamid ELGHAZI
National Institute of Posts and Telecommunication
Rabat, Morocco
h.elghazi@inpt.ac.ma

Abdellatif EL FAKER
ENSIAS, Mohammed V University
Rabat, Morocco
abdellatif.elfaker@um5.ac.ma

Abstract—**Social Customer Relationship Management (social CRM) has become one of the central points for many companies seeking to improve their customer experience. It comprises a set of processes that allows decision-makers to analyze customer data in order to launch an efficient customer-centric and cost-effective marketing strategy. However, targeting all potential customers with one general marketing strategy seems to be inefficient. While targeting each potential customer with a specific strategy can be cost demanding. Thus, it is essential to group customers into specific classes and target each class according to its respective customer needs. In this paper, we develop a Random Forest-based approach to classify potential customers into three main categories namely, prospects, satisfied and unsatisfied customers. The proposed model has been trained, tested, and compared to some state-of-the-art classifiers viz., Artificial Neural Network (ANN), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) based on several metrics including accuracy, sensitivity, specificity, false positive rate, and false negative rate. The reported results were satisfactory with an accuracy of 98.46%, a sensitivity of 97.69%, and a specificity of 98.84%.**

Index Terms—**Social CRM, customer experience, classification, Random Forest**

I. Introduction

Nowadays, most of the companies are constantly seeking to have a customer-centric business relationship, which is the core foundation of CRM. The evolving presence of customers in social media has created extreme pressure on companies to engage with their customers and build a long-term relationship with them [1]. Hence, they have started to focus on data collection from social media and many other peer-to-peer websites such as Facebook, Twitter, YouTube, blogs, wikis, and podcasts. Afterward, they have integrated such social data into their existing CRM systems, making it possible to understand customers and simultaneously meet their needs [2]. This inclusion of social media data in the CRM has given rise to a novel concept called social CRM, which involves social customer collaboration and engagement [3].

Building and maintaining a strong customer relationship in a cost-efficient way is one of the main concerns for any business. Companies are seeking an efficient way to identify their ideal target customers in order to take the right action for them, serve their needs more precisely, and customize the company's marketing while improving profits. Customer classification is one of the main pre-requisite phases in understanding and meeting customers' needs. A large body of research that deals with customer classification are leveraging machine learning to classify customers into various categories based on CRM data. For instance, Bahari et al. [4] proposed an efficient CRM-data mining framework for predicting customer behavior in the banking sector. They compared two classification methods, namely, Naïve Bayes (NB) and ANN, and the reported results showed that ANN outperformed NB in terms of accuracy, sensitivity, and specificity with values of 88.63%, 40.9%, and 94.85% respectively. Ahn et al. [5] proposed a customer classification approach including ANN, logistic regression (LR), decision trees (DT), and genetic algorithm for grouping customers into buyers and non-buyers using a mobile telecom dataset. D'Haen et al. [6] predict customer profitability during the acquisition process and find the optimal combination of data source and data mining technique used. DT, LR, and bagged decision trees techniques were applied on two kinds of datasets, namely: web and commercial data. The results showed that bagged decision trees provided the highest precision and web data had a higher predictive performance than commercial data. Emtiyaz et al. [7] suggested a semi-supervised learning techniques to improve CRM on banking and insurance sectors. They proposed a feed-forward ANN trained by a backpropagation algorithm to predict the category of potential customers. The results showed that their proposed model is more accurate compared to the other classification methods including ANN, SVM, KNN, and NB.

Although the satisfactory result was reported by some of the aforementioned classifiers, they have many limitations. For instance, the DT based approaches have too many instances which lead to a large decision tree [8]. This implies a low accuracy rate as compared to other classifiers. The ANN requires a long training time when it comes to a large dataset [9]. The main limitation of LR is the assumption of linearity between the dependent variable and the independent ones [10], [11]. SVM is known for its sensitiveness to the selected kernel function [12]. KNN, on the other hand, is sensitive to the irrelevant features [13].

In this paper, we develop a Random Forest (RF) based approach for classifying customers into satisfied, unsatisfied customers, and prospects. This classification is of primordial interest to companies since it allows them to have a crystal

978-1-7281-5950-8/20/$31.00 ©2020 IEEE

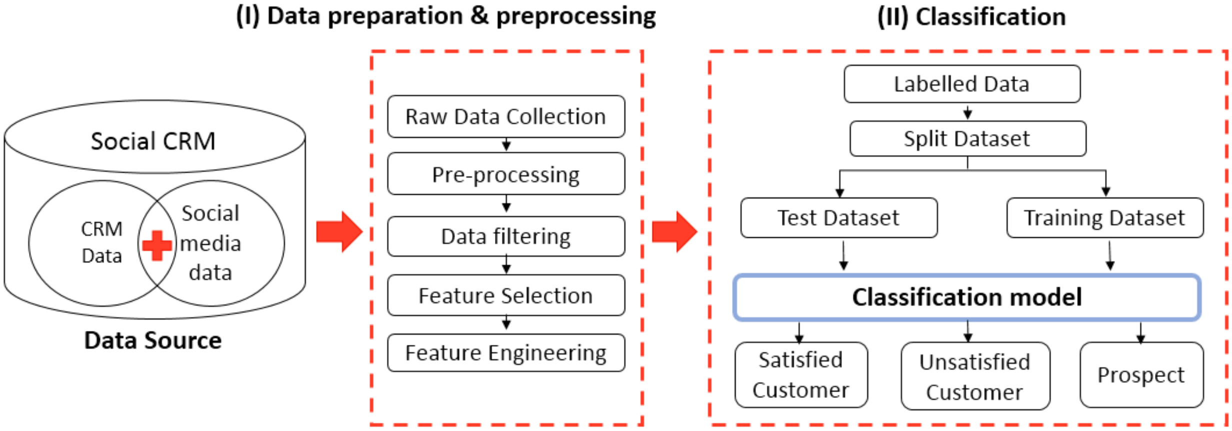

Fig. 1. Conceptual workflow of the proposed approach

clear visibility over their current customers and their prospects. The lack of attention to unsatisfied customers increases their churn rate and presents a direct entry point for competitors. Indeed, it is useful to identify unsatisfied customers and convert them into satisfied ones by using specific retention programs. On the other hand, prospecting for customers remains an important task for companies as well. To increase profit and stay ahead of competitors, companies should continually seek new customers. In this regard, identifying prospects will allow companies to customize their acquisition strategy such as running suitable promotions and advertising campaigns through the most convenient social media platform to maximize the prospects' conversion rate into customers.

The RF has been selected as it uses bagging and bootstrapping techniques which enable it to handle missing values, outliers, non-linear data [13] [14]. Additionally, to the best of our knowledge, this is the first attempt to apply the RF for a multi-classification problem in a social CRM context. Specifically, the proposed model has been trained on an e-commerce empirical dataset including the aforementioned three customer classes and compared to ANN, SVM, and KNN based on several performance metrics including accuracy, sensitivity, specificity, false positive rate, and false negative rate.

The remainder of this paper is organized as follows: Section II describes the proposed RF-based approach. Additionally, it details the training dataset along with the pertinent selected features used for classifying customers. Section III evaluates and compares the proposed model based on several performance metrics. Finally, Section IV draws some conclusions and sheds light on some future works.

II. THE PROPOSED APPROACH

A conceptual workflow for the proposed approach is shown in Fig. 1, which consists of two main phases. Phase (I), data preparation, and preprocessing. Phase (II), customer classification based on the RF model. These phases are deeply explained in the following sections.

A. Data Preparation and Pre-processing

To study the customer classification problem in social CRM, we consider real data on e-commerce provided by an online retail site over a period of three months. Our complete dataset refers to the so-called "social CRM data", which is a merging of the "CRM data" and the "social media data". CRM data consists of all information associated with the customers and products. Social media data refers to all raw insights and information collected from individuals on social media platforms. In our case of study, the social media data was about customers' comments, which was acquired from the Facebook platform after running an advertising campaign for a specific product by the e-commerce business owner.

For the subsequent classification, we consider three kinds of customer data namely: *interaction*, *behavioral*, and *attitudinal* data. Interaction data represents all digital footprints a customer or a potential customer (prospect) leaves behind as he/she consults the company store (e.g. page view, session duration). Behavioral data provides insight into customer's experience with the company's product or service (e.g. purchase, add to cart). Attitudinal data describes customers' or prospects' actual opinions about the company's product or service (e.g. positives/negatives sentiments in online reviews).

Originally, the dataset contained 1893 instances and 22 features including social media platforms' characteristics as well as customers' and prospects' data. Data filtering and missing values treatment were applied so that the dataset contains only useful instances. That is, instances with missing or incorrect records were discarded leading to 1016 instances with 22 features (categorical and numerical) and an output that includes one of the following three classes: *prospect*, *satisfied customer*, or *unsatisfied customer*.

Out of the 22 attributes, several irrelevant features adversely affect the processing time of the machine learning algorithm. A manual feature selection procedure was used to reduce the number of features to 15 based on domain knowledge. To this end, we identified the pairs of variables that are highly correlated and removed one of them to reduce dimensionality without loss of information. For instance, {"user id", "user name"}, {"item id", "item name"} are likely to carry similar information, so we dropped name variables as these are not unique and hold no significant importance for the subsequent classification.

Furthermore, all categorical features were transformed into their numerical form to speed up the convergence time of the algorithm. Besides, feature engineering is used to create new features from existing ones. Specifically, we created a feature named "sentiment_score", which was computed from the "review_text" feature that was originally existed in our dataset using our earlier LDA and Fuzzy-Kano based approach [15]. This latter study was implemented on the mobile reviews dataset, however, the current study implements the same approach on an e-commerce cosmetics dataset. Concerning data labeling, it has been done manually by the authors. The target variable *prospect* was defined depending on some attributes e.g. "event_view, "subscription", and "sentiment_score" while *unsatisfied* and *satisfied customers* variables were set according to "subscription", "event_purchase", "sentiment_score", and other attributes.

a) Feature Description: There are 15 features that have been selected for training the model. These features are described as follows: *Item id* is the product id. *User id* is the user id. *Session_duration* refers to the time spent by the user on the web site. For sake of simplicity, we refer to the customer and prospect as a user in this subsection. *Event_view* is a binomial feature that takes 1 when the user visits a product page, and 0 otherwise. *Event_addtocart* is a binomial feature that takes 1 when a user adds a product to the cart, and 0 otherwise. *Event_purchase* is a binomial feature that takes 1 when the user makes a purchase, and 0 otherwise. *No.purchase* is the number of products bought by the same user. *No.returns* is the number of times a product is returned by the same user. *Review_rating* is the rating score given by a user to a specific product. *Sentiment_scor*e is the score computed from the user's text review. *Audience_size* is the number of users on the advertising platform that is interested in a specific product. *Ad_video_avg* is the time spent, by a user, on watching a product advertising video. *Min_P* is the minimum number of users' accounts that can be reached on each advertising social media platform. *Subscription* is a binomial feature that takes 1 when the user is subscribed to the website, and 0 otherwise. *If_convert* takes 1 when a user was a prospect and has been converted to a customer after an advertising campaign, and 0 otherwise.

Table I summarizes the dataset which will be used in the classification phase.

B. Random Forest Classification

In machine learning, ensemble learning algorithms (e.g. Random Forest [16], boosting and bagging [17]) have received increasing interest due to their potential to significantly improve the classification accuracy and the generalization ability of a learning system [18], [19]. The philosophy behind ensemble algorithms is based on the fact that a set of classifiers do obtain better classification results than an individual classified does. Besides their advantage of having great generalization ability being scalable to large datasets, and benefiting of fast training and predictions, they are particularly well adapted to multi-class problems as they are inherently multi-class, and provide probabilistic output. This particular study involves classifying the customers into three classes or labels (i.e. prospects, satisfied customers, unsatisfied customers). RF, first proposed by Breiman [16], is a typical ensemble machine learning technique widely used in classification. Fig. 2 illustrates the Random Forest diagram.

TABLE I
SUMMARY OF THE DATASET USED FOR THE CLASSIFICATION

Features	**Numerical**	Item id, user id, session_duration, event_view, event_addtocart, event_purchase, no.purchase, no.returns, review_rating, sentiment_score, ausience_size, ad_video_avg, min_P
	Categorical	Subscription, if_convert
Output		Prospect, satisfied customer, unsatisfied customer
Number of features		15
Number of instances		1016

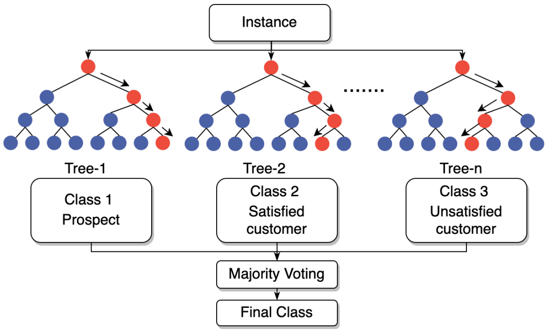

Fig. 2. The Random Forest model diagram

RF is made up of several decision trees classifiers, where each tree is trained using a different bootstrap sample of the original data [16]. Moreover, each tree contributes with a single vote for a given input data to get a class label. At each internal node of the tree, f features are randomly selected from the feature set F as candidate features (where $f \leq F$). The best feature is further selected from f candidate features for splitting the node into child nodes. The Information Gain (IG) and entropy (E) are used to decide which feature is best to choose to generate a decision tree [20]. (E) is the measure of uncertainty in a dataset and (IG) measures the difference between the entropy before and after splitting the dataset on a feature. This procedure is iterated over all trees until all non-leaf nodes are split, and the final prediction is made based on the majority votes from each of decision trees.

Assuming the node to be split is composed of a set of S, and S contains s samples and n unique class. The entropy of the node is expressed by:

TABLE II
COMPARISON BETWEEN ANN, SVM, RF, AND KNN
IN TERMS OF ACCURACY, SENSITIVITY, SPECIFICITY, FPR, AND FNR

Algorithm	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**FPR (%)**	**FNR (%)**
ANN-Tanh	57.01	35.52	67.76	32.23	64.47
ANN-Relu	77.85	66.77	83.38	16.61	33.22
SVM-Sigmoid	52.62	28.94	64.47	35.52	71.05
SVM-RFB	54.16	31.25	65.62	34.37	68.75
RF	98.46	97.69	98.84	1.15	2.30
KNN	96.92	95.39	97.69	2.30	4.60

$$E(s_1, s_2, \ldots, s_n) = -\sum_{k=1}^{n} P_k \log_2(P_k) \tag{1}$$

where s_k is sample number in class $k = (1, 2, \ldots, n)$, and $P_k = s_k/s$ denotes the probability that a sample belongs to class k. When S is completely homogeneous (i.e. contains only one class), the entropy is zero; On the other hand, if all the classes in S are equally distributed, the entropy is maximal. In our case, we consider three classes that are equally distributed i.e. $k = (1, 2, 3)$ and $n = 3$ where '1' indicates prospect, '2' indicates satisfied customer, and '3' indicates unsatisfied customer. Therefore, P_k presents the probability that a user belongs either to prospect, or satisfied customer, or unsatisfied customer. Let us assume that RF uses a feature X to split the node, S can be divided into m subsets S_j, with $j = 1, 2, \ldots, m$. The entropy with respect to the given feature X is defined as:

$$E_{split} = -\sum_{j=1}^{m} \frac{s_{1j} + \cdots + s_{nj}}{s} . E(s_{1j}, \ldots, s_{nj}) \tag{2}$$

where, s_{ij} is the number of sample of class i in the subset S_j. Hence, the IG of feature X splitting the node can be calculated according to Equations (1) and (2) as:

$$IG = E(s_1, s_2, \ldots, s_n) - E_{split} \tag{3}$$

Besides, the IG of each feature in the feature candidate set f can be calculated using Equations (1)-(3). Thus, the feature with the largest IG should be used for splitting the node.

III. RESULT AND DISCUSSION

A. Performance Metrics

In the experiments, the data was split into training (70%) and testing (30%) data. To assess the classifier model performance, the following classification metrics are used: accuracy, sensitivity, specificity, false positive rate, and false negative rate [21]. These metrics are calculated using true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Accuracy determines the number of instances that were correctly classified. It is calculated by using Equation (4):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Sensitivity corresponds to the proportion of correctly classified positive. It is given by:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

Specificity refers to the proportion of correctly classified negative. It is expressed as:

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

False Positive Rate (FPR) is the probability that a false alarm will be raised. It is given by:

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

False Negative Rate (FNR) is the probability that a true positive will be missed by the test. It is expressed as:

$$FNR = \frac{FN}{FN + TP} \tag{8}$$

B. Classification Result

To evaluate our approach, we contrast the performance of our classification model (RF) with three other classification models namely: ANN, SVM, and KNN. As some algorithms like the ANN and the SVM can perform with different parameters, a parametric study is conducted on these algorithms in order to limit the subsequent comparison to a selection of parameter-optimized classifiers. For instance, two types of activation functions are selected for ANN: Tanh and Relu, and two types of kernels are selected for SVM: Sigmoid and RBF. The comparison is carried out based on accuracy, sensitivity, specificity, FPR, and FNR.

Table II reports the comparison result of the trained and tested data using ANN-Tanh, ANN-Relu, SVM-Sigmoid, SVM-RBF, RF, and KNN models. As can be seen, the ANN with the Relu function exhibits better performance as compared to ANN with Tanh function. Hence, Relu can be considered as the optimal activation function for ANN in terms of accuracy, sensitivity, specificity, FPR, and FNR. In addition, the SVM with RBF outperforms the SVM with the sigmoid function. Thus, RBF can be considered as the optimal kernel for the SVM in terms of accuracy, sensitivity, specificity, FPR, and FNR.

Fig. 3 illustrates the comparison result between ANN, SVM, RF, and KNN in terms of accuracy, given the optimal functions. It is observed that RF provides the highest accuracy

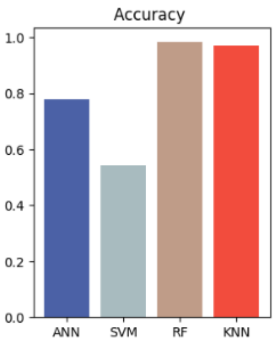

Fig. 3. Comparison between ANN, SVM, RF, KNN in terms of accuracy

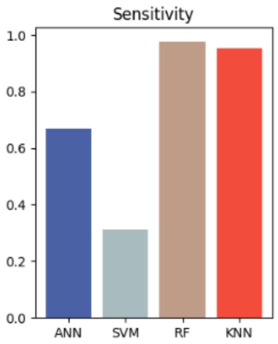

Fig. 4. Comparison between ANN, SVM, RF, KNN in terms of sensitivity

with a value of 98.46%, followed by the KNN with a value of 96.92%, ANN with a value of 77.85%, then SVM with a value of 54.16%. Similarly in Fig. 4, the four classifiers were compared in terms of sensitivity. The RF gives the best value, which is 97.69%, followed by KNN with 95.39%. Furthermore, Fig. 5 shows the comparison result in terms of specificity, where RF provides the highest value (98.84%), KNN gives a value of 97.69%, ANN gives a value of 83.38%, then SVM reports the lowest value (65.62%). Fig. 6, on the other hand, illustrates the comparison result in terms of false positive rate. The RF provides the lowest value, which is 1.15%, while SVM shows the highest one, which is 34.37%. Moreover, in Fig. 7, it is seen that RF gives the lowest value for the false negative rate (2.30%), followed by KNN (4.60%), ANN (33.22%), then SVM with the highest value (68.75%)

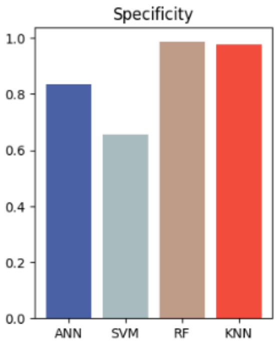

Fig. 5. Comparison between ANN, SVM, RF, KNN in terms of specificity

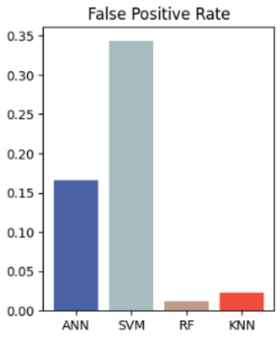

Fig. 6. Comparison between ANN, SVM, RF, KNN in terms of FPR

These findings suggest that RF is the optimal classifier comparing to ANN, SVM, and KNN in terms of accuracy, sensitivity, specificity, FPR, and FNR. Therefore, it is a good candidate for classifying accurately customers into prospects, satisfied and unsatisfied customers. Thus, each class of customers will be targeted with a suitable marketing strategy that meets their specific requirements, helping the companies to create, maintain, and develop a strong and long-term relationship with their customers.

IV. CONCLUSION

While extensive literature shows the role of CRM in the e-commerce business, low attention has been given by the academics and practitioners to customer classification in social CRM. The main contribution of this study is fulfilling this gap. This study proposed a Random Forest-based approach to classify customers into prospects, satisfied and unsatisfied customers. This model has been trained and tested on an e-commerce dataset including relevant features, which have been selected using the Information Gain technique. The obtained results show that Random Forest outperforms the other machine learning algorithms namely SVM, KNN, and ANN, with an accuracy of 98.46%, a sensitivity of 97.69%, a specificity of 98.84%, a false positive rate of 1.15%, and a false negative rate of 2.30%.

As future work, we will extend this study to address the customer acquisition and retention issues by applying optimization techniques. The customer acquisition will be

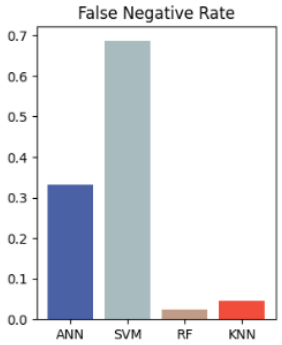

Fig. 7. Comparison between ANN, SVM, RF, KNN in terms of FNR

improved through maximizing the conversion rate of prospects into customers, while the customer retention will be improved through maximizing the customer satisfaction level, which leads to converting unsatisfied customers into satisfied customers.

References

[1] B. Halligan and D. Shah, Inbound marketing: get found using Google, social media, and blogs. John Wiley & Sons, 2009.

[2] M. Rodriguez, R. M. Peterson, and H. Ajjan, "CRM/social media technology: impact on customer orientation process and organizational sales performance," in Ideas in marketing: Finding the new and polishing the old, Springer, 2015, pp. 636–638.

[3] E. C. Malthouse, M. Haenlein, B. Skiera, E. Wege, and M. Zhang, "Managing customer relationships in the social media era: Introducing the social CRM house," Journal of interactive marketing, vol. 27, no. 4, pp. 270–280, 2013.

[4] T. F. Bahari and M. S. Elayidom, "An efficient CRM-data mining framework for the prediction of customer behaviour," Procedia computer science, vol. 46, pp. 725–731, 2015.

[5] H. Ahn, J. J. Ahn, K. J. Oh, and D. H. Kim, "Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques," Expert Systems with Applications, vol. 38, no. 5, pp. 5005–5012, 2011.

[6] J. D'Haen, D. Van den Poel, and D. Thorleuchter, "Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique," Expert systems with applications, vol. 40, no. 6, pp. 2007–2012, 2013.

[7] S. Emtiyaz and M. Keyvanpour, "Customers behavior modeling by semi-supervised learning in customer relationship management," arXiv preprint arXiv:1201.1670, 2012.

[8] V. Pohjalainen, "Predicting service contract churn with decision tree models," 2017.

[9] M. Abambres and A. Ferreira, "Application of ANN in pavement engineering: state-of-art," 2017.

[10] M. Hassouna, A. Tarhini, T. Elyas, and M. S. AbouTrab, "Customer churn in mobile markets a comparison of techniques," arXiv preprint arXiv:1607.07792, 2016.

[11] A. Hooman, G. Marthandan, W. F. W. Yusoff, M. Omid, and S. Karamizadeh, "Statistical and data mining methods in credit scoring," The Journal of Developing Areas, vol. 50, no. 5, pp. 371–381, 2016.

[12] T. Harris, "Credit scoring using the clustered support vector machine," Expert Systems with Applications, vol. 42, no. 2, pp. 741–750, 2015.

[13] A. Vanderveld, A. Pandey, A. Han, and R. Parekh, "An engagement-based customer lifetime value system for e-commerce," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 293–302.

[14] S. F. Sabbeh, "Machine-learning techniques for customer retention: A comparative study," International Journal of Advanced Computer Science and Applications, vol. 9, no. 2, 2018.

[15] S. Lamrharia, H. Elghazi, and A. El Faker, "Business intelligence using the fuzzy-Kano model," Journal of Intelligence Studies in Business, vol. 9, no. 2, 2019.

[16] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[17] R. Hao, X. Xia, S. Shen, and X. Yang, "Bank Direct Marketing Analysis Based on Ensemble Learning," in Journal of Physics: Conference Series, 2020, vol. 1627, p. 012026.

[18] Z.-H. Zhou, Ensemble learning. Encyclopedia of biometrics. Springer Berlin, 2009.

[19] N. C. Oza and S. Russell, Online ensemble learning. University of California, Berkeley, 2001.

[20] B. Sui, "Information gain feature selection based on feature interactions," PhD Thesis, 2013.

[21] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, p. 1, 2015.