

Bank Telemarketing Forecasting Model Based on t-SNE-SVM

Jianguo Che¹, Sai Zhao¹, Yongfan Li², Kai Li¹

¹Business School, Nankai University, Tianjin, China

²Mechanical Electrical Engineering School, Beijing Information Science & Technology University, Beijing, China

Email: cjg7705@nankai.edu.cn

How to cite this paper: Che, J. G., Zhao, S., Li, Y. F., & Li, K. (2020). Bank Telemarketing Forecasting Model Based on t-SNE-SVM. *Journal of Service Science and Management*, 13, 435-448.

<https://doi.org/10.4236/jssm.2020.133029>

Received: April 14, 2020

Accepted: May 15, 2020

Published: May 18, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

As a low-cost marketing model, telemarketing has always been the most important channel for banks to promote wealth management products. Traditional telemarketing has not only brought intrusiveness to many telephone access customers, but also a waste of resources for the bank itself. In order to improve the success rate of bank telemarketing, it is necessary to predict in advance which customers are most likely to purchase the wealth management product, so as to achieve precision marketing. Aiming at the complex high-dimensional nonlinear characteristics of the factors affecting the success rate of telemarketing, a t-SNE (t-distributed stochastic neighbor embedding) feature extraction method, and then take the extracted low-dimensional features as input, use nonlinear support vector machine (SVM) for training and prediction. The empirical results show that the bank phone based on t-SNE-SVM proposed in this paper. The marketing prediction model has good learning ability and generalization ability, which can provide certain decision-making reference for banks and other industries to achieve precision marketing.

Keywords

Telemarketing, t-SNE, Support Vector Machine (SVM), Forecasting, Precision Marketing

1. Introduction

As a typical strategy to promote business development, marketing activities can generally be divided into mass marketing and direct marketing. Mass marketing is the use of newspapers, radio, television and other media to promote the general public, while direct marketing is through mobile phones, fixed phones, Email, etc. directly contact customers to promote products or provide customers with

discounts. In today's highly competitive market environment, mass marketing is no longer an effective and reliable method, and marketing is moving from traditional mass marketing moving to direct marketing (Elsalamony, 2014). In direct marketing, because of its low cost and easy communication, telemarketing has gradually become one of the most widely used marketing channels. Compared to traditional account manager marketing and branch site marketing, telemarketing is promoting finance to individual customers. There are significant advantages in terms of products, and when the products are suitable, telephone marketing can directly make profits, thereby providing banks with new marketing channels and profit channels (Song, 2011). However, with the increasing use of telemarketing, companies have also received more and more customer complaints (Moro et al., 2012). For most customers who are reluctant to purchase the product, marketing calls may mean intrusiveness. On the other hand, many commercial banks have not implemented effective classified marketing strategies for their customers and often sell the same wealth management product to many customers, which is not effective. Use information to analyze customer needs, wasting a lot of manpower and other resources (Liu & Zhang, 2008). Therefore, it is necessary for companies implementing telemarketing to analyze customer data in advance using predictive models in order to select those customers who are most likely to respond to targeted marketing (Sing'oei et al., 2013), which can not only improve the marketing efficiency of bank managers, but also maximize. To reduce intrusiveness to non-target customers.

Banks usually have a large number of databases consisting of customer information and transaction information. This can not only provide banks with accurate and timely business and management information, but also perform functional query, analysis, and decision advice on information, and provide detailed information support for marketing activities (An, 2007). Data mining technology can not only explain past marketing results, but also provide decision support for future marketing activities, so it is widely used in bank marketing business. Many data mining algorithms are used to predict the success rate of telemarketing, such as support vector machines (Moro et al., 2012), deep convolutional neural networks (Kim et al., 2015), and comparative analysis of various classification algorithms (Elsalamony, 2014; Moro et al., 2011, 2014). Amponsah et al. (2016) used the J48 decision tree and Naive Bayes classifier to conduct an empirical analysis of the survey data set of a rural bank in Ghana, and provided decision suggestions for bank staff to sell a loan product to community residents (Amponsah et al., 2016). Mitik et al. (2017) constructed a data mining method based on profit-cost analysis. The empirical results show that this method will reduce a small amount of total profit, but because the total cost has dropped significantly, the total profit/cost ratio has increased significantly. Villuendas-Rey et al. (2017) proposed a naive association classifier. This classifier uses a new similarity operator, which can handle missing values as well as mixed classification and numerical variables. The classifier is verified better performance than other traditional classifiers by related financial data sets. Zakaryazad et al. (2016)

considered the cost of misclassification of each sample, modified the artificial neural network using a penalty function, and applied it to fraud detection and bank marketing. Although the empirical results are not as good as before, this method can achieve better performance when considering profit indicators. In related financial and commercial fields, Pang et al. (2009) embedded Boosting technology into the decision tree C5.0 algorithm, established a personal credit rating model based on C5.0, and performed credit rating on personal credit data of a German bank. Li et al. (2011) introduced the cross-selling sequence model to the cross-selling analysis of domestic commercial banks, constructed a cross-selling sequence model for individual customers, and used Logistic regression to predict the probability of customers buying different products. The results show that it can help banks implement cross-selling strategy effectively. Fang et al. (2014) constructed a personal credit risk early-warning model based on the Lasso-Logistic model. Through empirical analysis of credit card consumer credit default data, it showed that this model can better capture the key factors affecting consumer credit risk, and at the same time it has higher prediction accuracy. Zhang et al. (2015) applied Logistics and SVM to bank credit risk early warning. This method can fully capture and characterize the linear and non-linear complex features of influencing factors on customer defaults. The model has better generalization ability and can warn the credit risks of consumers accurately. Xiao et al. (2018) used BP neural network to evaluate the credit of network lenders. The verification of actual P2P transaction data showed that the model has strong predictive ability and can be applied to the risk control of P2P online loan platform to a certain extent. Zhang et al. (2018) constructed a credit risk assessment model of P2P online loan borrowers based on non-equilibrium fuzzy approximate support vector machines. The empirical results show that the model has better classification accuracy and better adaptability than other models. In addition, it can effectively reduce the impact of sample imbalance on classification results.

As the bank marketing data set generally contains more variables (such as many customer attributes), on the one hand it will affect the training speed of the model (dimensional disaster), it will consume more time and memory, and on the other hand it is easy to cause over-learning of the model, which makes it difficult to understand, so it is necessary to carry out the dimension reduction. The dimension reduction greatly reduces the time and memory requirements of the data mining algorithm, and can reduce the original data to 2D or 3D makes data easier to visualize (Tan et al., 2001). However, using automatic or semi-automatic feature selection methods to eliminate some input features may cause the lack of input information. To this end, this paper proposes a t-SNE-SVM-based bank telephone marketing prediction method. This method first uses the t-SNE algorithm to visually reduce the number of input attributes that may affect the success rate of telemarketing, while reducing complexity while maximizing information of original input features is retained. Then use the reduced-dimensional low-dimensional data as input to the SVM algorithm to learn and train to pre-

dict which customers will buy the wealth management products.

2. Prerequisite Knowledge

2.1. t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-distributed stochastic neighbor embedding (t-SNE) is a nonlinear dimensionality reduction and visualization algorithm proposed by Maaten et al. (2008), which maps multi-dimensional data to two or three dimensions suitable for human observation for visualization research (Maaten et al., 2008). t-SNE is derived from SNE (Stochastic Neighbor Embedding). The idea of the SNE algorithm is that while mapping high-dimensional data to low-dimensional data, try to ensure that the distribution probability between the points is unchanged, that is, similar data points in high-dimensional space have similar distances to low-dimensional space. There are two main disadvantages of the SNE algorithm: one is the large amount of gradient calculation caused by asymmetry, and the other is the crowding problem, that is, the clusters of different classes are crowded together and cannot be distinguished. In addition, the SNE algorithm only focuses on the locality of the data and ignores the globality of the data. The t-SNE algorithm is improved for the above defects. For a detailed introduction of the t-SNE algorithm, see the literature (Maaten et al., 2008).

2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) was first proposed by Corinna Cortes and Vapnik in 1995. It has many unique advantages in solving small sample, non-linear and high-dimensional pattern recognition and is widely used in industry, computer science, finance and other fields. Support vector machine provides better decision boundaries than traditional neural networks. The classification results are better than many parametric or non-parametric statistical techniques, and the problem of overfitting can be overcome through the concept of structural risk minimization (Kim et al., 2018). By non-linearly mapping the input vector to a high-dimensional feature space, the support vector machine can use a linear model to construct a nonlinear classification boundary.

Kernel functions are often used instead of dot product operations between vector pairs in transformed space. Common positive definite kernel functions that conform to Mercer's theorem are

- Linear kernel $k(x, y) = x \cdot y$, used for linearly separable cases, less applicable for practically complex problems;
- Polynomial kernel $k(x, y) = [(x \cdot y) + 1]^d$, is suitable for orthogonal normalized data. The larger the parameter d , the higher the dimension of the mapping, the more complicated the calculation, and the "over-fitting" phenomenon tends to occur.
- Radial basis kernel (Gaussian kernel) $k(x, y) = \exp(-\gamma \|x - y\|^2)$, this function is the most widely used kernel function because it has fewer parameters and is easier to learn;

- Sigmoid kernel $k(x, y) = \tanh(a(x \cdot y) + b)$, is derived from a neural network. If the Sigmoid function is used as the kernel function, the support vector machine is equivalent to a multilayer perceptron neural network.

In this paper, we use the radial basis function as the kernel function to construct a non-linear support vector machine model to learn and train the low-dimensional input after t-SNE feature extraction. Then we use the constructed model to make classification predictions on the test samples to effectively identify which customers are most likely to order bank wealth management products.

3. Data and Empirical Analysis

3.1. Problem Description

The data in this article comes from the real telephone marketing data collected by Sérgio, with 20 input variables, including 15 customer attributes such as customer age, occupation, marital status, and five social and economic attributes such as employment change rate, consumer price index, consumer confidence index (Moro et al., 2014). The output variable is whether the customer orders a certain financial product (yes/no), and the detailed introduction of each variable is shown in **Table A1** in **Appendix**.

3.2. Data Processing

The data set has a total of 41,188 customer records, of which 4640 customers successfully ordered the wealth management product, accounting for 11.27% of the total. It can be seen that the traditional telephone marketing customer response rate is very low, which brings intrusion to most customers. It is also a waste of resources for the bank itself. By reviewing the data, there are 6 attributes with missing values (“unknown”), as shown in **Table 1**.

After inspection, it is found that 79.12% of the customers in the credit default (default) field have not experienced a credit default, and the remaining 20.87% of the customers have an unknown default status, and only 3 customers (0.01%) have had a credit default (see <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing> for a more detailed understanding of the data). Therefore, this field is filtered out during modeling.

Table 1. Attributes with missing values.

Attributes	Number of valid records	Effective record ratio
Job	40,858	99.2%
Marital	41,108	99.8%
Education	39,457	95.8%
Default	32,591	79.1%
Housing	40,198	97.6%
Loan	40,198	97.6%

Considering adequate samples, in order to ensure the accuracy of the model, delete the customer records that contain missing values (unknown), and eventually retain 38,245 customer records with complete information, including 4258 (11.13%) customers successfully ordered products, 33,987 (88.87%) customers did not order products.

Due to the imbalance between the two types of output (yes/no), the learning process of the model will be troubled, resulting in the lack of application value of the model prediction results. The most commonly used method to solve the problem of class imbalance is undersampling, that is, reducing the number of samples of a larger type, making the positive and negative samples proportionally balanced. Therefore, this article keeps 4258 customer records of successfully ordered products, randomly selects 4258 samples from 33987 samples of unordered products. In the end, there were 8,516 customer records, of which 50% were customers who ordered/not ordered this product. The input variables contain 9 categorical variables, which need to be numericalized first, refer to the variable coding method used by Miguéis et al. (2017) which represents categorical variables with dummy variables. By this method, we obtain the input matrix of 8516×55 , and the output is a binary variable ($y = \text{yes/no}$).

3.3. Model Framework

Column standardization of the input variables to eliminate the influence of different variable dimensions, and then use the t-SNE algorithm to visualize the dimensionality reduction operation, the default output dimension is 2. Set the number of iterations 1000 times, the relationship between the iteration error and the number of iterations is shown in Figure 1.

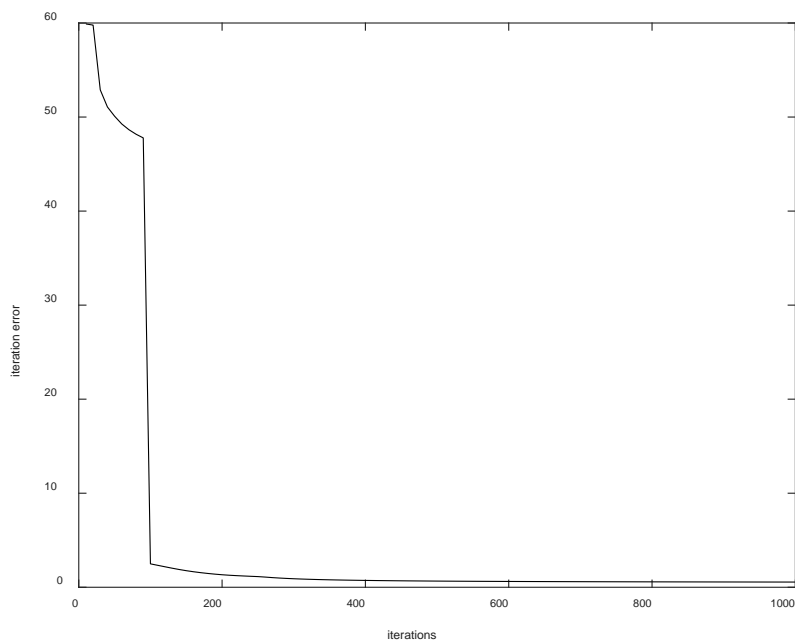


Figure 1. Iteration error of dimensionality reduction by t-SNE algorithm.

The objective function of the t-SNE algorithm is to optimize the KL divergence between the original space and the embedded space sample distribution, but the KL divergence is non-convex, so iterative iterations are required to obtain the final stable (convergent) optimal solutions. It can be seen that the error is sufficient to converge after 1000 iterations, and the output result is stable and reliable. Using t-SNE and SNE algorithms for dimensionality reduction (both target dimensions are 2 dimensions), the output results are visualized as shown in **Figure 2** and **Figure 3**.

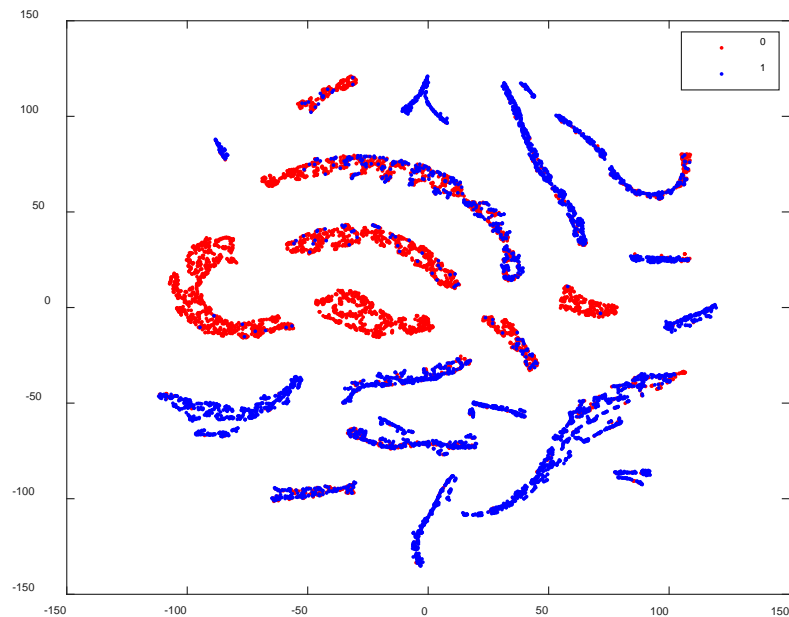


Figure 2. Two-dimensional display of raw data by t-SNE algorithm.

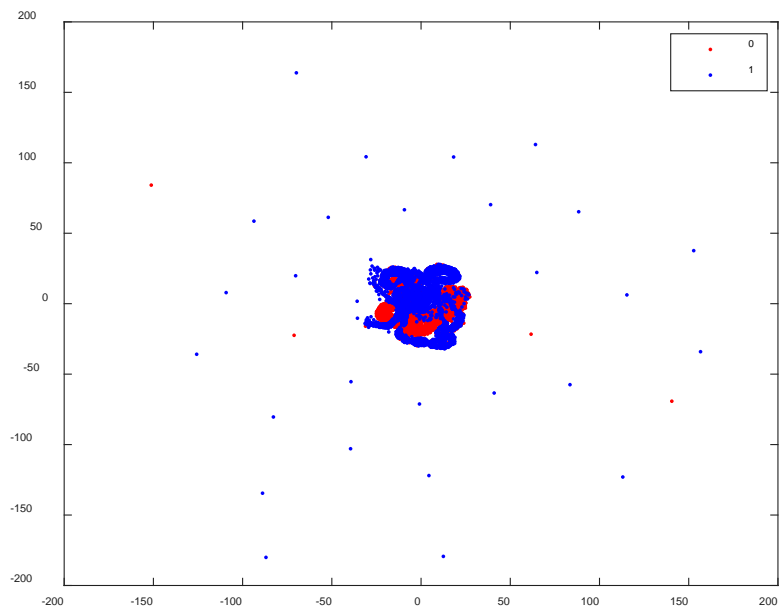


Figure 3. Two-dimensional display of raw data by SNE algorithm.

The red dot represents a sample of customers who did not order wealth management products (i.e., “ $y = 0$ ”, see the legend at the top right of **Figure 3**), and the blue dot represents a sample of customers who ordered wealth management products (i.e., “ $y = 1$ ”, see the legend at the top right of **Figure 3**). It can be seen that through the t-SNE algorithm, many input variables are compressed into a two-dimensional space, and samples of different categories appear blocky State, so it is possible to distinguish the two by using the hyperplane of the nonlinear support vector machine. The visualization effect using SNE dimensionality reduction is not ideal. A large number of samples overlap together, and it is difficult to distinguish the clusters near the center position. This may be due to the complex nonlinear non-Gaussian nature of the input variable customer attributes and social attributes. In contrast, the t-SNE algorithm uses a long-tailed t distribution to fit the distribution of data in low-dimensional space, and is more Robust, so it better captures the overall characteristics of the input data.

Using the radial basis function as a kernel function, we establish a nonlinear support vector machine classification (prediction) model. 70% of the samples are randomly selected for learning to determine the decision boundary parameters, i.e., w and b , and the rest 30% samples are used for testing to predict whether customers will order wealth management products and the likelihood of ordering wealth management products. We set that when the error between two iterations is 1×10^{-3} , stopping optimization algorithm. In addition, through several tests, taking $\gamma = 15$, can guarantee the accuracy of the classification model and as well as doesn't lead to over-fitting.

4. Evaluation and Comparative Analysis

4.1. Confusion Matrix

The confusion matrix is one of the most commonly used indicators to evaluate the quality of the prediction model. Define the positive class in the data set as P (Positive, in this case, is customers who order financial products), and the negative class as N (Negative, in this case, is customers who have not ordered financial products), the confusion matrix is shown as **Table 2**.

The relevant indicators are as follows:

Classification Rate/Accuracy: $(TN + TP)/(TP + TN + FP + FN)$;

Precision: $TP/(TP + FP)$, how many consumers who are predicted to order financial products will actually order;

Specificity: $TN/(TN + FP)$, how many customers who will not actually order financial products are successfully identified;

Recall: $TP/(TP + FN)$, how many customers who actually order financial products are successfully identified.

The confusion matrix of the test set obtained using the t-SNE-SVM prediction model in this paper is shown in **Table 3**.

Calculating from **Table 3**, we get: 1) accuracy = 86.07%, 2) precision = 83.64%, 3) specificity = 82.82%, 4) recall = 89.38%.

Table 2. Confusion matrix.

		Predicted		
		Negative	Positive	Total
Actual	Negative	TN	TP	N
	Positive	FN	FP	P
	Total	N*	P*	

Table 3. Confusion matrix of t-SNE-SVM model.

		Predicted		
		No	Yes	Total
Actual	No	1080	224	1304
	Yes	136	1145	1281
	Total	1216*	1369	2585

From a practical point of view, in the bank's marketing activities, the first task of the decision maker is to improve the customer's response rate, and hope to identify the users who are likely to order wealth management products at the least cost, so the main indicators are the accuracy rate and the recall rate. In the test set, the t-SNE-SVM model was used to predict that 1369 customers would order wealth management products, of which 1145 customers actually successfully ordered wealth management products, with an accuracy rate of 83.64%; Among the 1281 customers who actually ordered financial products, 1145 customers were successfully identified, with a recall rate of 89.38%, and the results were relatively satisfactory.

4.2. ROC Curve

The ROC curve is a curve drawn with the false positive rate as the abscissa and the true positive rate as the ordinate. The area under the ROC curve, which is the AUC (Area Under Curve) value, is commonly used to evaluate performance. The AUC value is generally from 0.5 to 1, 0.5 means completely random classification, 1 means perfect classification, the larger the value, the better the classifier performance. It is generally considered that the AUC value reaches 0.9 and the model accuracy is ideal. We draw the ROC curve using the t-SNE-SVM model in this paper, which is shown in **Figure 4**.

The ROC curves of the t-SNE-SVM model in the training set and test set have been calculated to obtain AUC values of 0.915 and 0.909 respectively, and the model classification effect is satisfactory.

The classification results of using t-SNE-SVM and SNE-SVM models respectively are shown in **Table 4** below.

By comparing the result indicators of the training set and the test set, neither of the two prediction models has overfitting. Compared with the SNE-SVM model, the learning and prediction effects of the t-SNE-SVM model are more satisfactory and can identify potential target customers more efficiently.

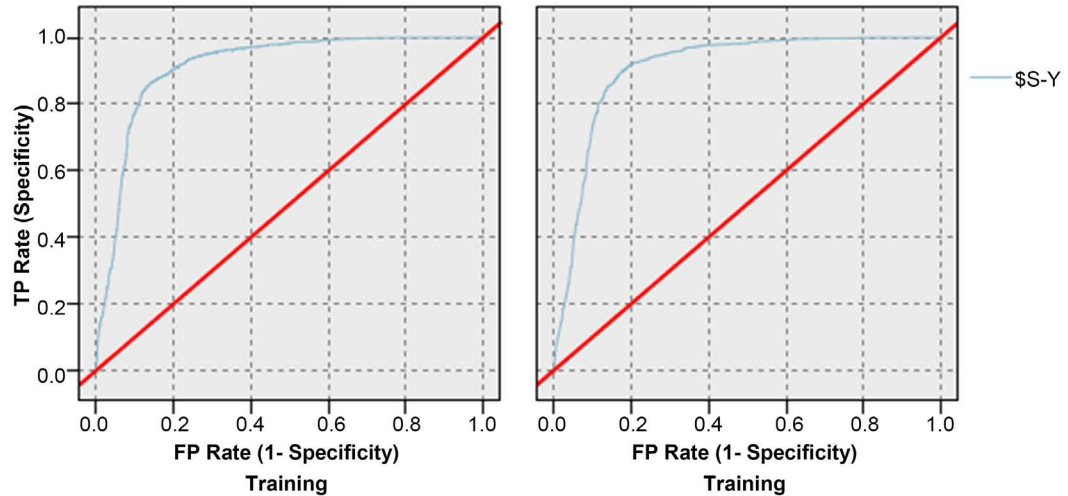


Figure 4. ROC curve of t-SNE-SVM model.

Table 4. Classification results comparison of t-SNE-SVM and SNE-SVM.

Model	Training set			Test set		
	precision	recall	AUC	precision	recall	AUC
t-SNE-SVM	84.17%	88.08%	0.915	83.64%	89.39%	0.909
SNE-SVM	70.49%	66.34%	0.751	67.28%	65.18%	0.736

4.3. Lift Curve

In the marketing field, the Lift curve is another curve used to evaluate the quality of the classification model. Unlike the ROC curve, the Lift curve considers the number of positive classes obtained using the classifier and the number of positive classes randomly obtained without using the classifier Proportion. Taking telemarketing as an example, the good or bad of the classifier is how much response to the target customer (i.e., how much final consumption) will be brought to the bank by using the predictive model in advance compared with the direct random selection of customers. This is more intuitive and effective for banks, so it is more widely used in the field of marketing.

The Lift curve of t-SNE-SVM model in this paper is drawn in Figure 5 as follows.

As shown in Figure 5, considering the cost of manpower and other costs, it is assumed that the bank manager can only select half of its customers for telemarketing. If it is not randomly selected using the classification model, it is expected to cover 50% of the target customers, and use the t-SNE-SVM model in this paper, we can get 85.1% of the customer response, and the bank benefit increased by 35.1% under the same cost (better than the 29% improvement obtained in the literature (Moro et al., 2014)). This result proves the practicability of the t-SNE-SVM model in this paper. This model allows banks to reduce costs (reducing the number of calls) while still successfully identifying most target customers.

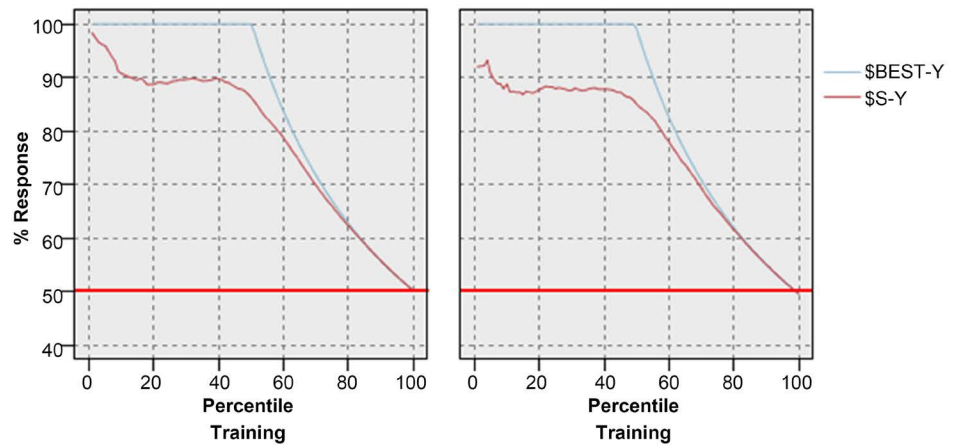


Figure 5. The Lift cumulative curve of t-SNE-SVM model.

5. Conclusion

Due to the advantages of low cost and wide coverage, telemarketing is widely used in the marketing activities of bank wealth management products. Traditional telemarketing does not distinguish between customers, and marketing staff “spray and pray” to many customers. Inefficiency also brings intrusiveness to customers of the telephone access party, which is not conducive to the maintenance and long-term development of customer relations.

Motivated by this practical situation, this paper proposes a t-SNE-SVM-based bank telemarketing prediction model. This model first reduces many input attributes that affect the success rate of telemarketing to two dimensions through the t-SNE algorithm, while maximizing the retention of original data information. It is easy to visualize, greatly reducing the complexity of input variables, and then using nonlinear support vector machines to learn and predict the dimensionality-reduced samples to effectively identify which potential customers will order financial products. This not only reduces the intrusiveness brought to many non-target customers, but also helps the bank to effectively improve marketing efficiency and reduce the cost of resources such as manpower, so that the bank can carry out more targeted marketing activities. The empirical research on the marketing data set of Portuguese banks show that, the prediction model proposed in this paper has a relatively satisfactory learning ability and generalization ability, which can provide applications for banks and other related industries to achieve precise marketing.

However, our research is certainly not without limitations. First, we only compared the feature extraction results of the two algorithms: t-SNE and SNE. In future research, more dimensionality reduction algorithms (such as PCA, LDA, LLE) can be applied to the bank tele-marketing data to verify the effectiveness of the t-SNE algorithm. In addition, due to the unavailability of data, our model only shows satisfactory results under the current data set. In the future, one can try more classification datasets to prove the versatility of the model.

Acknowledgements

This research is financially supported by Youth Fund of Humanities and Social Sciences Research Projects of Ministry of Education [(15YJC630007) “Research on Optimal Personnel Scheduling of Service Enterprise under Dynamic Situation”, 17YJC630073] and Asia Research Center Research Project of Nankai University [AS1410]. Again, more thanks are also given to the comments of the anonymous peer reviewers.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- Amponsah, A. A., & Pabbi, K. A. (2016). Enhancing Direct Marketing using Data Mining: A Case of YAA Asantewaa Rural Bank Ltd. in Ghana. *International Journal of Computer Applications*, 153, 6-12. <https://doi.org/10.5120/ijca2016912092>
- An, H.-X. (2007). On Our Commercial Banks Marketing Strategy Questions. *Journal of Central University of Finance & Economics*, 4, 42-48. (In Chinese)
- Elsalamony, H. A. (2014). Bank Direct Marketing Analysis of Data Mining techniques. *International Journal of Computer Applications*, 85, 12-22. <https://doi.org/10.5120/14852-3218>
- Fang, K.-N., Zhang, G.-J., & Zhang, H.-Y. (2014). Individual Credit Risk Prediction Method: Application of a Lasso-Logistic Model. *The Journal of Quantitative & Technical Economics*, 31, 125-136. (In Chinese)
- Kim, K. H., Lee, C. S., Jo, S. M., & Cho, S. B. (2015). Predicting the Success of Bank Telemarketing Using Deep Convolutional Neural Network. In *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPar)* (pp. 314-317). Piscataway: Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/SOCPAR.2015.7492828>
- Kim, S., Mun, B. M., & Bae, S. J. (2018). Data Depth Based Support Vector Machines for Predicting Corporate Bankruptcy. *Applied Intelligence*, 48, 791-804. <https://doi.org/10.1007/s10489-017-1011-3>
- Li, C.-Q., Li, X.-P., Tian, M., & Zhao, Q.-Y. (2011). A Research of the Cross-Selling in Bank's Individual Customer Base on the Sequence Model. *Forecasting*, 30, 13-18. (In Chinese)
- Liu, F.-J., & Zhang, S. (2008). Discussion of Financial Product Innovation and Marketing Mode. *China Soft Science*, 2, 65-75. (In Chinese)
- Maaten, L. V. D., & Hinton, G. (2008). Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9, 2579-2605. <http://www.jmlr.org/papers/v9/vandermaaten08a.html>
- Miguéis, V. L., Camanho, A. S., & Borges, J. (2017). Predicting Direct Marketing Response in Banking: Comparison of Class Imbalance Methods. *Service Business*, 11, 831-849. <https://doi.org/10.1007/s11628-016-0332-3>
- Mitik, M., Korkmaz, O., Karagoz, P., Toroslu, I. H., & Yucel, F. (2017). Data Mining Approach for Direct Marketing of Banking Products with Profit/Cost Analysis. *The Review of Socionetwork Strategies*, 11, 17-31. <https://doi.org/10.1007/s12626-017-0002-5>

- Moro, S., Cortez, P., & Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems*, 62, 22-31. <https://doi.org/10.1016/j.dss.2014.03.001>
- Moro, S., Laureano, R., & Cortez, P. (2011). Using Data Mining for Bank Direct Marketing: An Application of the Crisp-Dm Methodology. In *Proceedings of European Simulation and Modelling Conference-ESM 2011* (pp. 117-121). Guimaraes, Portugal: University of Minho,.
- Moro, S., Laureano, R., & Cortez, P. (2012). *Enhancing Bank Direct Marketing through Data Mining*. European Marketing Academy Conference.
- Pang, S.-L., & Gong, J.-Z. (2009). C5.0 Classification Algorithm and Its Application on Individual Credit Score for Banks. *Systems Engineering-Theory & Practice*, 29, 94-104. (In Chinese) [https://doi.org/10.1016/S1874-8651\(10\)60092-0](https://doi.org/10.1016/S1874-8651(10)60092-0)
- Sing'oei, L., & Wang, J. (2013). Data Mining Framework for Direct Marketing: A Case Study of Bank Marketing. *International Journal of Computer Science Issues (IJCSI)*, 10, 198.
- Song, J.-H. (2011). Research on Telemarketing of Commercial Bank. *Finance Forum*, 16, 62-66. (In Chinese)
- Tan, P.-N., Steinbach, M., & Kumar, V. (2011). *Introduction to Data Mining*. Beijing: Posts & Telecom Press, 30-31.
- Villuendas-Rey, Y., Rey-Benguría, C. F., Ferreira-Santiago, Á., Camacho-Nieto, O., & Yáñez-Márquez, C. (2017). The Naïve Associative Classifier (NAC): A Novel, Simple, Transparent, and Accurate Classification Model Evaluated on Financial Data. *Neurocomputing*, 265, 105-115. <https://doi.org/10.1016/j.neucom.2017.03.085>
- Xiao, H.-M., Hou, Y., & Cui, C.-S. (2018). Evaluation of P2P Lending Borrower's Credit on BP Artificial Neural Network. *Operations Research and Management Science*, 27, 112-118. (In Chinese)
- Zakaryazad, A., & Duman, E. (2016). A Profit-Driven Artificial Neural Network (ANN) with Applications to Fraud Detection and Direct Marketing. *Neurocomputing*, 175, 121-131. <https://doi.org/10.1016/j.neucom.2015.10.042>
- Zhang, Q., Hu, L.-Y., & Wang, J. (2015). Study on Credit Risk Early Warning Based on Logit and SVM. *Systems Engineering: Theory & Practice*, 35, 1784-1790. (In Chinese)
- Zhang, W.-G., Lu Y.-Y., & Liu, Y.-J. (2018). The Borrowers' Credit Risk Assessment in P2P Platform Based on Fuzzy Proximal Support Vector Machine and Its Application. *Systems Engineering: Theory & Practice*, 38, 2466-2478. (In Chinese)

Appendix

Table A1. Input and output variable descriptions.

	Variable name	Variable type	Variable description
Bank client data	Age	Numeric	From 17 to 98
	Job	Categorical	Type of job, including “admin”, “housemaid”, “blue-collar”, etc.
	Marital	Categorical	Marital status, “married”, “divorced”, “single”
	Education	Categorical	“Basic.4y”, “basic.6y”, “basic.9y”, “high.school”, “professional.course”, “university.degree”, “unknown”, “illiterate”
	Default	Categorical	Has credit in default? “yes”, “no”, “unknown”
	Balance	Numeric	Average yearly balance, in euros
	Housing	Categorical	Has housing loan? “yes”, “no”, “unknown”
	Loan	Categorical	Has personal loan? “yes”, “no”, “unknown”
Related with the last contact of the current campaign	Contact	Categorical	Contact communication type, “unknown”, “telephone”, “cellular”
	Month	Categorical	“Jan”, “Feb”, ..., “Nov”, “Dec”
	Duration	Numeric	Last contact duration, in seconds
	Day_of_week	Categorical	Last contact day of the week, from “Mon” to “Fri”
Other attributes	Campaign	Numeric	Number of contacts performed during this campaign and for this client
	Pdays	Numeric	Number of days that passed by after the client was last contacted from a previous campaign
	Previous	Numeric	Number of contacts performed before this campaign and for this client
	Poutcome	Categorical	Outcome of the previous marketing campaign, “failure”, “nonexistent”, “success”
Social and economic context attributes	Emp.var.rate	Numeric	Employment variation rate—quarterly indicator
	Cons.price.idx	Numeric	Consumer price index—monthly indicator
	Cons.conf.idx	Numeric	Consumer confidence index—monthly indicator
	Euribor3m	Numeric	Euribor3 month rate—daily indicator
	Nr.employed	Numeric	Number of employees— quarterly indicator
Output variable	y	Binary	Has the client subscribed a term deposit? “yes”, “no”