Comparative study of the classification models for prediction of bank telemarketing

Elzhan Zeinulla
Nazarbayev University,
Computer Science
Astana, Kazakhstan
E-mail:
elzhan.zeinulla@nu.edu.kz

Karina Bekbayeva
Nazarbayev University,
Computer Science
Astana, Kazakhstan
E-mail:
karina.bekbayeva@nu.edu.kz

Adnan Yazici
Nazarbayev University,
Computer Science
Astana, Kazakhstan
E-mail:
adnan.yazici@nu.edu.kz

***Abstract* — This research paper has evaluated various classification models for prediction of bank telemarketing campaign results regarding the probability of the subscription of the customer to the deposit. The effectiveness of these algorithms has been evaluated by Receiving Operator Characteristic (ROC) and Cumulative Accuracy Profile (CAP) curve analysis, the accuracy of the algorithm and variance of the predictions. According to the results of the research, the best model for bank telemarketing effectiveness prediction are Random Forest and Deep Artificial Neural Network. The Logistic Regression and Naive Bayes are not as suitable as the other classification methods for this kind of problems due to the poor accuracy and overfitting issues.**

***Keywords* — bank telemarketing; kNN; ANN; Random Forest; Naive Bayes; SVM; Logistic Regression, CAP, ROC.**

I. Introduction

The marketing campaign is widely applied in entrepreneurship. It includes promoting an object or a service via mass communication. For instance, radio, television, telephone, print, social media and other ways to contact with the targeted audience. When organization contacts with customers through these communication methods, it is called direct marketing. The data which was used for conducting data analysis in this study is about direct marketing campaigns, specifically telemarketing, of a Portuguese banking institution. Telemarketing builds the relationships with a potential customer by calls by mobile and/or fixed-line telephones [1].

The relevance of this technique nowadays is justified by several advantages. First of all, it gives an opportunity to build mutual relations and immediately understand needs of the customer. Receiving such a feedback allows to evaluate a given situation and make possible compromises. In addition, reminders about the existence of a given service or a product take a significant role in enhancing an efficiency of the sales. It contributes to the increase of an awareness among the population. Moreover, in comparison with other marketing technique, such as prime TV advertisement, telemarketing is a cheaper way to make new customers interested as well as support relationships with already existed clients [2].

On the other hand, when the target audience is very wide, calling all clients takes a huge amount of time. The possible solution, which is described in this work, is to construct classification models, compare them between each other for identifying the most accurate one. The classification models aim to recognize a person with an increased probability to make a purchase. It allows to focus on potential buyers and not to spend time on inactive population. The contribution of this paper is to apply the machine learning algorithms, such as Deep Artificial Neural Network (DANN), Naive Bayes, Logistic Regression, k-Nearest Neighbors, Decision Trees, and Support Vector Machines, for identifying the most suitable method for evaluating whether the target customers of the bank telemarketing campaign subscribes to the deposit or not.

Additionally, the performance results of this study are much better than the solution previously provided by S. Moro in [3]. Using the same dataset, he used four classifiers: Logistic regression, Decision Trees, Neural Network, and Support Vector Machine. These models were compared among each other by two metrics: ROC curves analysis and LIFT cumulative curves analysis. According to Moro's work, the best classification model was NN and presented an accuracy of 0.80 by ROC curve. However, DANN used in our paper demonstrated an accuracy of 0.86 using the same analysis method and the same dataset. Additionally, another classifier, Random Forest, is used in this study and produces the highest accuracy results of 0.90 by ROC curve analysis.

Regarding the organization of this paper, we start with the evaluation of the previous studies related to the bank telemarketing classification problem and then we continue with application of the various machine learning techniques to the field of study in Section 3. Next, the research provides the brief explanation of each machine learning algorithm used in the paper with some assumptions about the performance of the algorithms in this application. Then it provides the results of each algorithm with a detailed description of the performance with the CAP and ROC curve analysis. Finally, in Section 5, the paper provides some possible better boosting approaches and alternative algorithms that should be considered in the future study.

II. Related Work

The similar problem is addressed in the article written by S. Moro and called "A data-driven approach to predict the success of bank telemarketing" [3]. The authors use the same dataset as we use in this research work of 5-year period, between 2008 and 2013, of the Portuguese retail bank. They implement four classification learning models for identifying the clients with increased probability of opening a long-term deposit: Logistic Regression, Decision Trees, Neural Network and Support Vector Machine. The best result is obtained by Neural Network.

A telecommunication industry problem based on customer churn prediction is solved by T. Vafeiadis [4]. Authors compare most popular classification methods. It includes Logistic Regression, Naive Bayes, Decision Trees learning, Support Vector Machine, and Artificial Neural Network. These models are compared by precision, recall, accuracy, and F-measure criterias.

Similarly, classification models are compared in the article "Islamic versus conventional banks in the GCC countries" [5]. The paper aims to predict suitability of financial ratios for distinguishing between conventional and Islamic banks. The paper again uses Neural Network, Classification tree, Logistic Regression, and Linear discriminant analysis. Authors consider the period between 2003 and 2010 years.

III. Research Methodology

For prediction of deposit subscription, we use the following most popular and effective machine learning algorithms for classification of telemarketing campaign for banks. We will very briefly summarize them in this section for self-reliability purpose.

A. *Naive Bayes (NB):*

Naive Bayes is simple and powerful probabilistic classifier. The main assumption while evaluating the model in NB is that the variables are strongly independent of each other. However, in reality, this algorithm still performs well if there is a correlation between the independent variables of the dataset. The worst accuracy of the NB algorithm is between the extremes of full independence and functionally dependence of the variables [6].

B. *Artificial Neural Network (ANN)*

Artificial Neural Network is the engineering model of the biological neurons, consisting of multiple interconnected and layered elements. After receiving the signal from the input, weights are adjusted and according to the activation function (rectifier, tanh, sigmoid, etc.) output is produced [7]. This paper employs the backpropagation technique to train the neural network. The approach is to minimize the cost function by adjusting the weights after each epoch of the ANN [8]. One of the main advantages of the ANN is that once trained it can make prediction fast.

C. *Random Forest (RF)*

Random forest classifier is the assemblage of the simple Decision Tree classifiers [9]. Each tree in the forest is trained on the random subset of instances and features of the training data. The probability of each tree is averaged and adjusted according to their contribution. Eventually, the highest probability score is assigned to the target variable [10]. It takes a relatively small amount of time to train the model, but much more time is needed for making the prediction, especially if there is a huge amount of random decision trees.

D. *Support Vector Machine (SVM)*

Support Vector Machine is a supervised learning algorithm which applies kernel functions. In this paper RBF gaussian (Radial Basis Function) kernel is implemented. It classifies for determining the hyperplane which corresponds to the maximum margin between two categories (distance between hyperplane and nearest data points). Support vectors are critical points closest to the hyperplane which can be non-linear. If training dataset has some bias, SVM can be robust by generalizing it with appropriately chosen C parameter. The C parameter is responsible for the adjustment of misclassification level [11].

E. *Logistic Regression (LR)*

Logistic Regression is a statistical method with a binomial response variable (binomial outcome) which is used in the analysis of data with one or more explanatory variables [12]. The model identifies the best fitting model for describing the relationships between dependent variables and a predictor variable [13]. In comparison with multiple regression analysis, logistic regression does not require the independent variables to be normally distributed and is able to handle nonlinear relationships between dependent and independent variables, which makes it a flexible model.

F. *k-nearest neighbors (kNN)*

kNN is a non-parametric method of classification, which determines a belonging of a data point to one of the groups depending on data points around it and distance between them for optimal values of k, where k is a number of nearest training samples [14]. This classifier is simple for implementation and understanding. However, during kNN analysis, every feature is computed similarly. For that reason, when the subset of features is small, it can lead to the classification errors [11].

G. *Dataset*

According to the evaluation of various papers, we created the following tables [Table 1 and Table 2], describing when the mentioned algorithms are suitable [10, Table 3], [15], [16]. The following table is applicable only to the datasets which are similar to that used in this paper.

The dataset is based on real data taken from the University of California, Irvine (UCI) Machine Learning Repository [3], which refers to the records of the Portuguese retail bank from 2008 till 2013. It contains overall 45211 records, with 11.58% (5235 contacts) of campaign's success. For evaluation purposes, the dataset is divided into the training (36168 contacts) and test (9043 contacts) sets with the ratio of 8:2.

1. Input features:

Personal information: Age, marital status, education and area of work (management, marketing, entrepreneur, self-employed, retired, etc.);

Billing: billing status of the target person (information about the existence of credit card, loan, and housing).

Previous campaign: Information about the results of the previous marketing campaign: the amount of previous contacts before the current campaign, number of days passed from the previous campaign and result if exists (success or failure).

Current campaign: Information about current campaign: date of the last contact, call duration, communication type and amount of contacts performed during this campaign.

2. Output variable:

Prediction: Binary value signifying whether the client subscribed for the deposit.

H. Assumptions

In order to find out predictions which algorithms will give the best accuracy of the prediction, it is crucial to characterize the input dataset. Due to the fact that term "Volume" of the dataset is relative, this paper evaluates the size of the dataset according to the research of Mayer [17] and Lowe [18], which shows the large variance of the datasets.

First of all, according to the research papers mentioned above, dataset of Portuguese retail bank can be characterized as "large". Considering the purpose of the research, the prediction of the result of the marketing campaign will not be in real time, so it is acceptable to neglect the training and prediction speeds of the algorithms. Definitely the input dataset has some "noisy" features, so the ability of the algorithm to handle these values is a plus. Finally, the main criterion in the evaluation of the best predictive model is the accuracy of algorithm. Not less important criterion, which is not specified in the table due to the complexity of the evaluation - variance of the accuracies.

It is clear from the table [Tab. 3] that the best models for classification are Deep Artificial Neural Network and Random Forest. The worst performance among the chosen algorithms is Logistic Regression. The performances of the algorithms are evaluated with Cumulative Accuracy Profile (CAP) and Receiving Operator Characteristic (ROC) curves. To improve and tune the predictive model, we use GridSearch, which is brute-force method to find the best combination of the hyperparameters [19].

The general strategy to avoid the overfitting is k-fold cross validation, where k is 10 folds. The dataset is divided into the k equally sized folds and for each fold, the model learns on k-1 subsets, and the remaining k is considered as the validation data. This process is repeated k - times (number of folds) and for each fold k-th validation data can be used only once. Then the average of the accuracies of each fold is taken to estimate the overall accuracy of the model [20].

IV. RESULTS

All algorithms are executed, tested and evaluated on Linux Dell xps15, with CPU intel-core i7-8559U, 32gb RAM and 2.2GHz of frequency.

According to the results in the table [Tab. 3], the Random Forest algorithm reaffirms our prediction with the accuracy of 90.884%. However, the CAP curve shows the extremely well performance of this algorithm (95.83% positive observations at 50% of data), which is unusual. In order to check whether the model was overfitted or not, an additional ROC curve [Fig 1] has been constructed and k-fold (k=10) cross validation accuracies of train and test set has been compared. Values of accuracies for train and test sets are 89.574% and 90.170% respectively, which proves that the model is not overfitted. Additionally, as Breiman [9] states that the Random Forest algorithm always converge, so it is impossible that the model is overfitted.

DANN and SVM also provide good accuracies and CAP graph performances. The variances of the predictions of both DANN and SVM are remarkably small signifying that the overfitting problem has been avoided. Analysis of the table [Tab. 3] shows that Naive Bayes algorithm is not suitable for the telemarketing campaigns for banks due to the high variance of accuracies. The problem may be caused by the fact the input dataset is too large for this algorithm, and it is overfitted. Although Logistic Regression has shown acceptable results of accuracies and variance, it provides much worse performance on the CAP graph than the other techniques.

The results of these series of experiments are obtained via searching of the best combination of the hyperparameters. The

TABLE I. RELEVANCE OF ALGORITHMS TO THE DATASET

Parameter	*NB*	*KNN*	*RF*
Effectiveness on	Small data	Small data	Large data
Training speed	Fast	Intermediate	Intermediate
Prediction speed	Intermediate	Very slow	Slow
Can deal with noise data	Yes	No	Yes
Accuracy if requirements satisfied	High	High	Very high

TABLE II. RELEVANCE OF ALGORITHMS TO THE DATASET

Parameter	*SVM (rbf)*	*DANN*	*LR*
Effectiveness on	Intermediate data	Any data	Intermediate data
Training speed	Slow	Very slow	Slow
Prediction speed	Fast	Very fast	Intermediate
Can deal with noise data	Yes	Yes	No
Accuracy if requirements satisfied	Very high	Very high	Intermediate

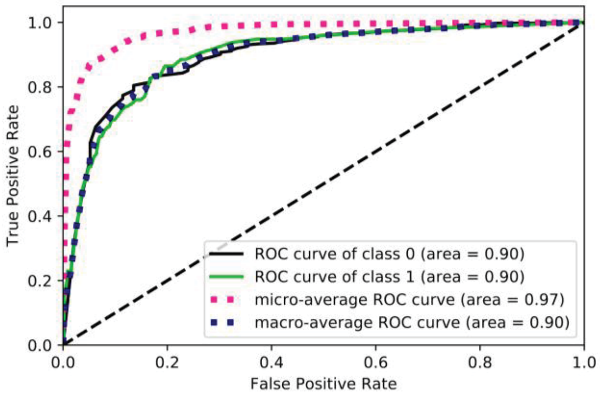

Fig. 1. ROC Curves analysis for Random Forest.

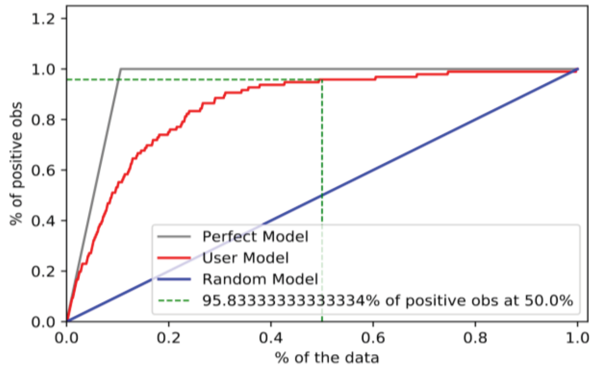

Fig. 2. CAP Curves analysis for Random Forest.

specific choice of the method is GridSearch cross validation due to the simplicity of implementation and low-cost computational resources. As Bergstra claims [21], random search of the parameter space might be much more effective than GridSearch algorithm, so the more sophisticated methods of boosting should be applied as a future work. Good alternatives to these methods can be AdaBoost, Gradient Boosting and GentleBoost.

V. CONCLUSION AND FUTURE WORK

The effectiveness of the bank telemarketing for customer attraction is one of the weakest links of the overall marketing campaigns. The success rate of telemarketing campaign for customers deposit subscription is extremely low, so the accurate probabilities of the subscription can save the huge amount of financial and human resources.

This paper has evaluated six different machine learning algorithms for the classification problem: probability of the acceptance rate for the deposit agreement by the target audience. The dataset is taken from Portuguese trade bank and consists of the personal information and billing status of the customer, reports about previous and current campaigns. The machine learning algorithms used for evaluation are

TABLE III. PERFORMANCE ANALYSIS OF ALGORITHMS

	Accuracy %	Variance %	% of positive observations at 50% (CAP curve)	k-fold cross validation accuracy %	
				Train set	Train set
KNN	86.229	1.03	78.41	88.563	87.447
RF	90.884	1.72	95.83	89.574	90.233
LR	86.185	5.23	60.185	86.863	86.186
ANN	90.286	1.57	92.41	90.007	91.210
NB	86.868	10.94	88.03	82.580	86.395
SVM	89.670	1.35	93.27	89.351	90.121

Random Forest (Decision tree ensemble), k-nearest neighbours, DANN, Naive Bayes, SVM and classical Logistic Regression. According to the results of our experiments, the best algorithm for the prediction in this particular problem is Random Forest which has the highest accuracy rate among all the algorithms and it presents very good performance on CAP graph with the 95.83% of positive observation at 50% of data.

This algorithm was not considered in Moro's [3] study, where he showed the Area Under the Curve of ROC graph with coefficient of 0.8 compared to our finding, which we found 0.9 using Random Forest, which signifies the remarkable increase of the performance compared to the previous studies in literature. On the other hand, the classical Logistic Regression and Naive Bayes has provided poor results, which lead to the conclusion that they are not suitable for this problem. Additionally, The SVM and DANN has provided excellent results as well, which make them competitive with the Random Forest. These performance results lead to the fact that the best algorithms for the telemarketing classification problems are SVM, DANN and RF, and they might be used interchangeably depending on the conditions and the size of the dataset. We think that DANN might provide better performance for a bigger dataset, with appropriate hyper parameter tuning and using some optimization techniques.

As future works the more sophisticated boosting algorithms should be used to increase the performance of the algorithms, such as AdaBoost, GentleBoost and Gradient boosting. Moreover, in order to better estimate the models,

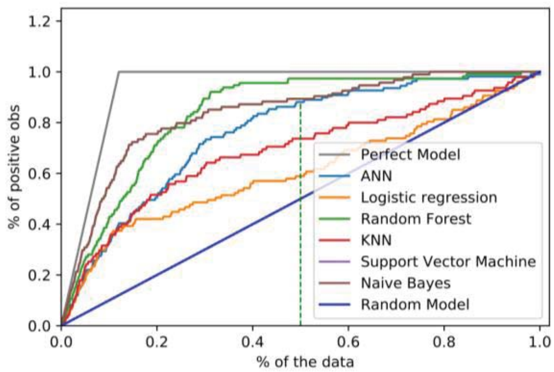

Fig. 3. CAP Curves for all algorithms.

area under the LIFT cumulative curve can be applied to increase the precision of the evaluation. Also, it is interesting to consider such algorithms as Xgboost, LIghtGBM and Extra Trees to find a better approach for prediction of the telemarketing effectiveness.

REFERENCES

[1] S. Palaniappan, A. Mustapha, C. F. M. Foozy, and R. Atan, "Customer Profiling using Classification Approach for Bank Telemarketing" *International journal on informatics visualization,* vol. 1, no. 4, pp. 214, December, 2017. [Online]. Available: http://joiv.org/index.php/joiv/article/view/68/48. [Accessed: 02- May-2018].

[2] K. C. Schneider, "Telemarketing as a promotional tool - its effects and side effects" *Journal of Consumer Marketing,* vol. 2, no. 1, pp. 29-39, 1985.

[3] S. Moro, P. Cortez and P. Rita, "A data-driven approach to predict the success of bank telemarketing", *Decision Support Systems*, vol. 62, pp. 22-31, 2014.

[4] T. Vafeiadis, K. Diamantaras, G. Sarigiannidis and K. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction", *Simulation Modelling Practice and Theory*, vol. 55, pp. 1-9, 2015.

[5] K. Khediri, L. Charfeddine and S. Youssef, "Islamic versus conventional banks in the GCC countries: A comparative study using classification techniques", *Research in International Business and Finance*, vol. 33, pp. 75-98, 2015.

[6] I. Rish, *An empirical study of the naive Bayes classifier,* T. J. Watson Research Center [Online]. Available: https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf. [Accessed: 04- May- 2018].

[7] V. Sharma, S. Rai, and A. Dev, "A Comprehensive Study of Artificial Neural Network" *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 2, no. 10, pp. 279, October, 2012. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.468.9353&rep=rep1&type=pdf. [Accessed: 06- May- 2018].

[8] K. Miroslav, *An Introduction to Machine Learning.*, 1st ed. Zurich: Springer, 2015, pp. 91-111.

[9] L. Breiman, "Random Forests" *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[10] S. D. Jadhov and H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques" *International Journal of Science and Research,* vol. 5, no. 1, pp. 1842-1845, January, 2016. [Online]. Available: https://www.ijsr.net/archive/v5i1/NOV153131.pdf. [Accessed: 09- May-2018]

[11] A. Tomar and A. Nagpal, "Comparing Accuracy of K-Nearest-Neighbor and Support-Vector-Machines for Age Estimation" *International Journal of Engineering Trends and Technology,* vol. 38, no. 6, pp. 327-329, August, 2016. [Online]. Available: http://www.ijettjournal.org/2016/volume-38/number-6/IJETT-V38P260.pdf. [Accessed: 11- May- 2018].

[12] "Lesson 6: Logistic Regression | STAT 504", *Newonlinecourses.science.psu.edu*, 2018. [Online]. Available: https://newonlinecourses.science.psu.edu/stat504/node/149/. [Accessed: 17- May- 2018].

[13] S. Sperandei, "Understanding logistic regression analysis", *Biochemia Medica*, pp. 12-18, February, 2014.

[14] T. Cho, R. W. Conners, and P. A. Araman, *A Comparison of Rule-Based, K-Nearest Neighbor, and Neural Net Classifiers for Automated Industrial Inspection.* [Online]. Available: https://pdfs.semanticscholar.org/2284/b9265cb010c530b06d140c006476c440d7b8.pdf. [Accessed: 17- May- 2018].

[15] M. Cleasen, F. D. Smet, J. A. K. Suykens, and B. D. Moor, *Fast Prediction with SVM Models Containing RBF Kernels,* vol. 3, October, 2014. [Online]. Available: https://arxiv.org/abs/1403.0736. [Accessed: 18- May- 2018]

[16] B. K. Raghavendra and J. B. Simha, "Evaluation of Logistic Regression Model with Feature Selection Methods on Medical Dataset" *International Journal on Computational Intelligence,* vol. 1, no. 2, December, 2010.

[17] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, and A. Dosovitskiy, *A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation,* Computer Vision Foundation, 2016, pp. 4040-4047.

[18] R. Lowe, N. Pow, I. Serban, and J. Pineau, *The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems,* vol. 3, February, 2016. [Online]. Available: https://arxiv.org/abs/1506.08909. [Accessed: 02- Jun- 2016]

[19] X. Ma, Y. Zhang and Y. Wang, "Performance evaluation of kernel functions based on grid search for support vector regression", *IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, 2015.

[20] P. Refaeilzadeh, L. Tang and H. Liu, "Cross-Validation", *Encyclopedia of Database Systems*, pp. 1-7, 2016.

[21] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization" *Journal of Machine Learning Research,* 2012. [Online]. Available: http://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra12a.pdf. [Accessed: 05- Jun- 2018].