Potential Customer Prediction of Telecom Marketing based on Machine Learning

Yuxin Dong *

Department of Beijing Institute of Technology, Beijing, China

* Corresponding author: 1120212913@bit.edu.cn

Abstract. Telemarketing has an important application in commercial promotion, and blind product recommendation has a high failure rate. However, product recommendation to potential users can effectively reduce marketing costs and increase revenue. In this paper, 41,188 data on telemarketing from a Portuguese banking institution are selected with the classification objective of predicting whether a customer will subscribe to a time deposit account or not. The paper first preprocesses the data to fill in missing data. Secondly, this paper describes the four models used in this paper: Logistic Regression, K-Nearest Neighbor, Decision Tree and Random Forest Classifier. As well as the five-assessment metrics used to evaluate these models: accuracy, AUC value, KS value, model lift and profit. In the experimental stage, this paper uses the above four models to predict the effectiveness of bank telemarketing. And the five evaluation indexes are combined to judge the prediction effect of the models. The results show that Decision Tree and Random Forest Classifier have better prediction effect.

Keywords: Telemarketing, Customer forecasting, Machine learning.

1. Introduction

Potential customer predicting, as an important part of enterprise management, has a direct impact on enterprise marketing strategy. Merchants can divide different customer groups through customer segmentation. And forecasting for different customer segments to realize precise product recommendation. In commercial applications, the complexity of customer and transaction information greatly increases the difficulty of customer segmentation. With the development of e-commerce, how to improve the accuracy of Potential customer forecasting has greatly attracted extensive interest from researchers.

In the early stage of the development of potential customer forecasting, researchers performed simple customer segmentation to predict customers by dividing users by gender and geography. This method is simple to operate. But it is less effective in practical use. With the development of machine learning, some researchers proposed to apply machine learning in the field of potential customer forecasting to obtain good performance. The customer-oriented age and growing bank competitiveness have led to the widespread application of machine learning in the banking and finance industry today to identify target customers and enhance bank sales. [1].

However, in the practical application of predicting customers in the field of banking and finance, due to the different effects and complexity of various algorithms, there are fewer studies on the comparison of categorization effects between different algorithms. Therefore, this paper will analyze and compare several commonly used machine learning algorithms to provide reference for researchers in the field of banking and finance.

This paper firstly introduces the Bank Marketing dataset of Portuguese Banking Institution. Then it introduces the Logistic Regression, K-Near Neighbors, Decision Tree and Random Forest algorithm. After that, it conducts an experiment that attempts to segment the bank's customers in order to predict whether or not they will subscribe to fixed-term deposit products [2]. The different algorithms of the experiment are analyzed and compared using data indicators (Accuracy, AUC values and KS values) and business indicators (Model Lift and Profit). Finally, it concludes with a summary and a vision of future directions for bank marketing.

2. Dataset

Bank Marketing Dataset is a commonly used dataset in the field of customer forecasting, which was proposed by S. Moro, P.Rita and P.Cortez in 2012 [3] and selected from UC Irvine Machine Learning Repository. The dataset consists of prior obtained by the research team through crawling the marketing campaign data of a Portugeuse bank. These data include customers' jobs, ages, education level, marital status etc. Each type of data is described in Table 1.

The classification goal is to predict if the client will subscribe to a term deposit account or not by telemarketing. It contains 41188 clients and covers 21 fields, 10 of which are numerical fields (age, duration, campaigns, etc) and 11 of which are categorical fields (job, marital, education, etc). As shown in Table 2, the target variable is a binary (y) "Yes" (the client subscribed, positive, 1) or "No" (the client did not subscribe, negative, 0).

Table 1. Numerical variable table

Field	Description
age	Age
duration	Duration of last touch, in seconds
campaign	Quantity of connections made for this customer and for this campaign
pdays	Number of days (numeric; 999 indicates the client was not previously contacted) that elapsed since the client was last contacted from a previous campaign
previous	Quantity of connections made both for this customer and prior to this campaign
emp.var.rate	Variation in employment rate - a quarterly measure
cons.price.idx	Index of consumer prices: a monthly measure
cons.conf.idx	Indicator of consumer confidence: a monthly measure
euribor3m	Three-month euribor rate: a daily measure
nr.employed	Employee count: a quarterly measure

Table 2. Categorical variable table

Field	Description	Filed	Description
job	Type of job	default	Is credit in default?
education	Educational level	loan	Possesses a personal loan?
housing	Has housing loan?	month	Month of the year's final contact
contact	Type of contact interaction	poutcome	Outcome of the previous marketing campaign
day_of_week	The final day of contact for the week	y	Has the client subscribed a term deposit?
marital	Marital status		

In data processing, subscription is set to 1 and unsubscription to 0 for the target variable y. The positive sample (i.e., y=1) accounts for 11.265% of the total number of samples in the entire data level. The next step is to explore and clean other variables. After removing the leakage variable "duration", the relative frequencies of the numerical and categorical variables are plotted based on the numerical variables and categorical variables.

For the missing values in the categorical features, this paper adopts a data estimation method, which uses other independent variables to estimate these missing values and infer the missing values. Although not all missing data can be recovered, most of the data can be restored. For example, there is a certain correlation between "work" and "education level" - the higher the education level, the higher the technicality of the work. Therefore, this paper infers the work of an individual based on their education level, and vice versa. The same approach is used to fill in the missing data for "work" and its relationships with "age", "house" and "loan". During the filling process, this paper pays attention to the rationality of these relationships in the real world. If they are not reasonable, the data will not be changed.

3. Method

This chapter focuses on the application of Logistic Regression, K-Near Neighbors, Decison Tree and Random Forest algorithms to customer predicting.

3.1. Logistic Regression

One effective statistical technique for solving classification difficulties is logistic regression, which is very easy to apply to binary classification problems. This method is used for classification and models the probability of response variable given independent variables using a logistic function. In logistic regression models, the results of linear regression are converted to probability values by log-odds [4]. The input features are summed linearly weighted and converted by a sigmoid function to a probability value between 0 and 1. Typically, statistical methods such as residual analysis and goodness of fit tests are used to evaluate the performance and fit of logistic regression models.

3.2. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a nonparametric statistical method for classification and regression, the basic idea of which is to classify and predict observations by measuring their distances in feature space. KNN is an intuitive classification and regression method that is intuitive in its principles and easy to apply as a statistical method for multiclass problems.

The KNN operates by computing the distance, identifying the nearest neighbours, and casting votes for labels. The method's controlling variable is K, which stands for the number of nearest neighbours. There is a specific optimal value for K in each dataset. If K is large, the noise would significantly affect the result and the calculation would be expensive. Research indicates that a small K is more flexible with low bias and great variance, while a huge K is not flexible with lower variation and increased bias. [5].

In this paper, the performance of KNN is evaluated on the training set using cross-validation for different values of k. The average roc_auc score and standard deviation corresponding to each value of k are output. From the results, it can be seen that when k=60, the roc_auc score is the highest and the standard deviation is small, when the model performance is optimal. Therefore, in the subsequent experiments, this paper sets the k value to 60.

3.3. Decision Tree

A Decision Tree is a representation of a decision set or a method of classifying data based on many qualities that uses a tree form. It is a method of efficiently supervising inductive learning by using data to create rules. Put another way, a decision tree is a method that illustrates rules for values to be acquired under identical circumstances by using a tree structure to describe decisions. An event often has the potential to trigger two or more events with distinct outcomes.

The decision tree's top-down structure is derived from this property. The flowchart and this structure are comparable. The decision tree's root node symbolizes the tree's beginning. Every branch of the tree indicates a new choice outcome and the progeny nodes on each node of the tree show the relevant attribute test. This child node's number has been transmitted. It is derived from decision-making algorithms. Fig. 1 displays the decision tree model [6].

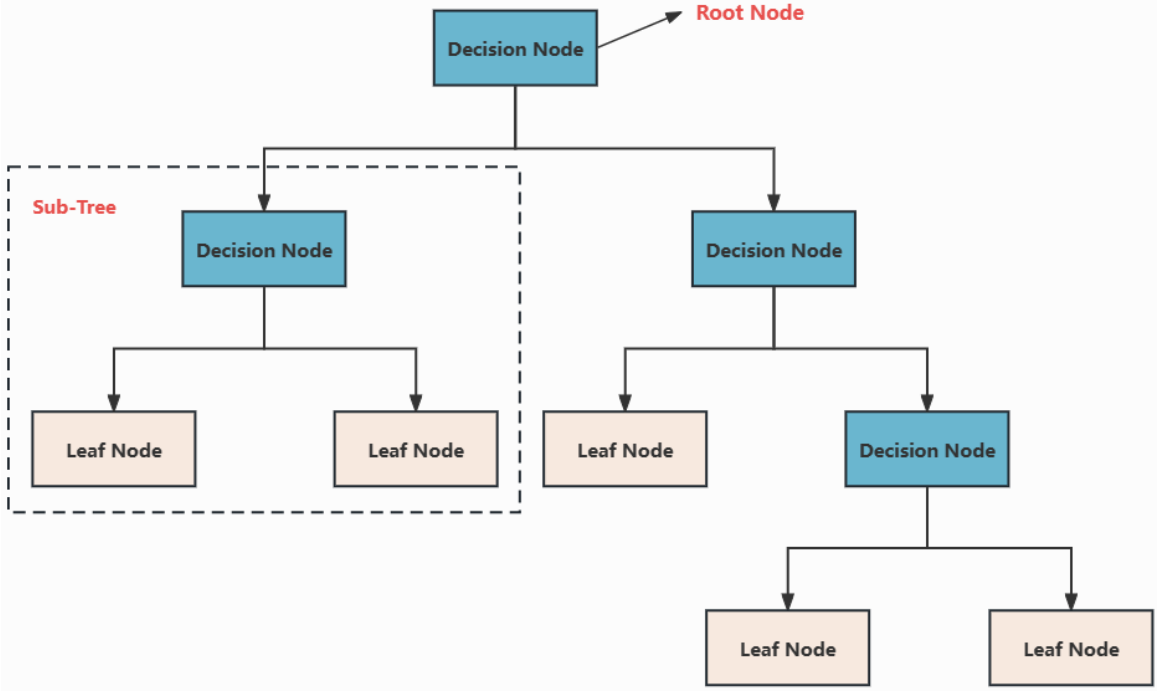

Figure 1. Principle of Decision Tree

This approach is useful for feature selection and variable screening. Additionally, with little to no user data preparation, it can handle large dimensional data consistently. Compared to previous methods, this one requires less data cleaning since it is not affected by outliers. [5].

3.4. Random Forest

RF is a decision tree based algorithm for learning regression models [7] . Each tree is grown using a randomly selected subset of features. And then the mean of the predictions obtained at the last node of each tree is computed, making up for the shortcomings of the low bias but very high variance deficiency exhibited by a single decision tree [8]. It combines the simplicity of decision trees with the advantages of integrated learning. And it is able to effectively deal with complex datasets and high-dimensional features. It's an effective machine learning technique with good predictive performance, robustness and interpretability. It is widely used in practical applications to solve a variety of classification and regression problems and still performs well for large-scale datasets and complex feature spaces.

4. Performance Metrics

This paper uses a number of standard performance metrics to evaluate the effectiveness of the various classifiers for predicting bank marketing, which are data indicators of AUC and KS and business indicators of Model Lift and Profit. These metrics may be utilized to assess the effectiveness of any model [9]. The following describes the metrics.

4.1. Accuracy

It refers to "the quantity of samples that were accurately predicted ÷ the total quantity of samples", which is a relatively simple and intuitive indicator. It is usually used when the data set is balanced [10]. The formula is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where:
TP: number of true 1s and predicted 1s

FN: number of true 1's and predicted 0's
FP: number of true 0's and predicted 1's
TN: number of true 0's and predicted 0's

4.2. AUC

AUC is a metric for assessment for binary classification models. The area under the ROC curve, which ranges from 0 to 1, is the value of AUC. One may use the AUC as a numerical value to infer a classifier's level of performance. The better, the higher the value.

AUC is expressed as a probability value. The AUC value indicates the likelihood that, in the event that a positive sample and a negative sample are randomly chosen, the current classification algorithm will rank the positive sample higher than the negative sample based on the calculated Score value, which represents the probability that each sample belongs to the positive sample. The likelihood that the positive sample will be ranked ahead of the negative sample by the present classification algorithm increases with the size of the AUC value. It can thus classify it more accurately.

The formula is as follows:

$$AUC = \frac{\sum_{i \in positiveClass} rank_i - \frac{M(1+M)}{2}}{M \times N} \quad (2)$$

The scores are firstly sorted from highest to lowest. Assume that the sample with the highest score has rank n, the sample with the second-highest score has rank n-1, and so on. Subtract M-1 combinations of two positive samples from the rankings of all positive samples. The number of pairings in which the positive samples have scores higher than those of the negative samples across all samples is what we obtain. Then divide the result by M × N, where M and N represent the number of positive and negative samples, respectively.

4.3. KS

The empirical cumulative distribution function serves as the foundation for the KS statistic. Two numbers are required to determine KS: the true positive rate (TPR), which is the percentage of positive cases detected by the classifier over all positive instances and is computed as TPR=TP/(TP+ FN). The other is the false positive rate (FPR), which is the ratio of negative cases to all negative instances that the classifier incorrectly reports as positive. It is computed as FPR= FP / (FP + TN). And KS=max(TPR-FPR).

The KS value is typically multiplied by 100% and falls between 0 and 1. In general, the degree of positive and negative sample discrimination improves with increasing KS. The following Table 3 displays the KS operational evaluation criteria:

Table 3. The operational evaluation criteria of KS

KS value range	Model effect
KS<0.2	No ability to differentiate
0.2≤KS<0.3	Some ability to differentiate, barely acceptable
0.3≤KS<0.5	Ability to differentiate
0.5≤KS<0,75	Strong ability to differentiate
0.75≤KS	Possible anomalies (so effective that there may be problems)

4.4. Model Lift

Lift diagrams, which show the ratio of outcomes produced "with the model" to "without the model," are frequently used in model evaluation to evaluate a model's efficacy. The ratio of the outcomes "with the model" to the outcomes "without the model" is known as lift. It quantifies the degree to which the model's predictive power outperforms that of the model alone (that is, the model's ability to forecast a target's "response" outperforms a randomly selected multiple of the average

response). The larger the lift, the better the model performs. The model or rule works independently of random selection if the lift is larger than 1, greater than 1 indicates that the model or rule captures more "responses" than random selection, and less than 1 indicates that the model or rule captures fewer "responses" than random selection.

4.5. Profit

The performance of the model is assessed by measuring how much profit there is relative to different marketing efforts using different models. The greater the profit, the better.

5. Experiment

In this experiment, 80% of the data is used as the training set and the remaining 20% as the test set. The training set is divided into 5 subsets for cross validation and model training and testing is performed on each subset. Accuracy, AUC value, KS value, Model Lift and predicted Profit values are calculated separately for each model and based on that comparison is done.

5.1. Accuracy

As seen in the Table 4, the Random Forest Classifier model has the maximum accuracy, while the Decision Tree Classifier model also has a higher accuracy, which indicates that these two models have higher prediction accuracy.

Table 4. Accuracy of different models on training set

Models	Accuracy
Random Forest Classifier	0.811649
Decision Tree Classifier	0.810310
K-Near Neighbors	0.808749
Logistic Regression	0.807186

5.2. AUC

From the results, the Random Forest Classifier model appears to have the greatest AUC value, while the Decision Tree Classifier model also has a higher AUC value, which indicates that these two models have a better prediction (Table 5).

Table 5. AUC value of different models

Models	Training set	Test set
Random Forest Classifier	0.856	0.849
Decision Tree Classifier	0.852	0.846
K-Near Neighbors	0.850	0.836
Logistic Regression	0.830	0.833

5.3. KS

The results are shown in the Table 6. As can be observed, the Random Forest Classifier model has the highest KS value, followed by the Decision Tree Classifier model, suggesting that these two models are more predictive than the other.

Table 6. KS value of different models

Models	Training set	Test set
Random Forest Classifier	0.576128	0.585947
Decision Tree Classifier	0.551563	0.573445
K-Near Neighbors	0.566656	0.571043
Logistic Regression	0.546686	0.571838

5.4. Profit and Model Lift

In this paper, the intensity of telemarketing is divided into 100 intervals. And the results of the experiment show that telemarketing is most effective at the 50th interval, when the profits predicted using different models are maximized. At this point, the values of profits predicted using different models and their corresponding cumulative lift are as follows.

From the results, it can be seen that the profit predicted using the Decision Tree model is the largest at $9,184, which corresponds to a cumulative uplift of 1.73. While the profit predicted using the Random Forest Classifier model is slightly lower than that predicted using the Decision Tree at $9,164, which also corresponds to a cumulative uplift of 1.73 (Table 7).

Table 7. Profit and Model Lift

Models	Profit	Model Lift
Random Forest Classifier	9164	1.728713
Decision Tree Classifier	9184	1.728713
K-Near Neighbors	9009	1.703102
Logistic Regression	8937	1.690297

6. Conclusion

Optimizing telemarketing objectives is a key issue in the banking industry under constant pressure to increase profits and reduce costs. In this context, the use of machine learning models to predict the outcome of telemarketing for long-term deposits is a valuable approach to support banks in their customer decision making.

In this paper, we select 41,188 marketing campaign datas from Portugeuse bank, whose goal is to use known customer attributes prior to the execution of a telemarketing call to predict whether the customer will subscribe to a long-term deposit. In this paper, four models, Logistic Regression, K-Nearest Neighbors, Decision Tree and Random Forest Classifier, are used for prediction and five metrics, Accuracy, AUC value, KS value, Model Lift and Profit are used to evaluate the prediction effect and compare the prediction effect of the four models. The experimental results show that the prediction effect of Random Forest Classifier and Decision Tree models is more satisfactory. In the future prediction work for bank marketing, these two models may continue to be considered for use. And other machine learning models, such as Support Vector Machine and Neural Network, can also be considered.

References

[1] Wang D. Research on Bank Marketing Behavior Based on Machine Learning. Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture, 2020.

[2] Moro S, Cortez P, Rita P. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, 2014, 62: 22 – 31.

[3] Moro S, Rita P, Cortez P. UCI Machine Learning Repository, 2012. http://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

[4] Walker S H, Duncan D B. Estimation of the Probability of an Event as a Function of Several Independent Variables. Biometrika, 1967, 54 (1/2): 167.

[5] Wang Y, Zhang Y, Lu Y, et al. A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a Case Study of Bank Loan Data. Procedia Computer Science, 2020, 174: 141 – 149.

[6] Chen C, Geng L, Zhou S. Retraction Note: Design and Implementation of Bank CRM System Based on Decision Tree Algorithm. Neural Computing and Applications, 2022, 35 (6): 4803 – 4803.

[7] Breiman L. Random Forests. Machine Learning, 2001, 45: 5 - 32. DOI: 10.1023/A: 1010933404324.

[8] Han Yajuan, Gao Xin. E-commerce merchandise sales prediction based on machine learning combinatorial model. Computer System Applications, 2022, 31 (01): 315 - 321.

[9] Umayaparvathi V, Iyakutti K. A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics. 2016.

[10] Cao C, Wang P, Huang H, et al. A Review of Methods for Telecom Customer Churn Prediction in Imbalanced Data. In: Proceedings of the 17th Chinese Academy of Automation System Simulation Technology and its Application Academic Annual Meeting (CCSSTA 2016). Chinese University of Science and Technology Press, 2016: 5.