International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

Improving Prediction of Potential Clients for Bank Term Deposits using Machine Learning Approaches

Prabodh Wankhede[1], Rohit Singh[2], Rutesh Rathod[3], Jayesh Patil[4], T.D. Khadtare[5]

[1,2,3,4]BE Student, SITS, Pune
[5]Prof., Dept. of Computer Engineering, SITS, Pune, India

Abstract - *Novel results for bank telemarketing have been obtained to enhance bank term deposits by employing machine learning approaches. Acquiring term deposits in banking is always an essential business for them and it is good marketing campaign always plays an essential role in financial selling. More often than not, it is always a challenge to financial institutions to identify the group of customers for this purpose. Since a very few customers respond positively to direct telemarketing campaigns, there is a class imbalance in the data. This inherent problem of class imbalance in response modeling brings some additional difficulties into response prediction. As a result, the prediction models are generally biased towards non-respondent customers and this adds a further challenge. In this paper, the focus is on critical feature engineering and investigating the predictive machine learning approaches like Logistic Regression, Support Vector Machine, Random Forest, Neural Networks, and XGBoost. The best performing predictive model have been identified which can be deployed in bank telemarketing campaigns.*

Key Words: **Bank, term deposit, telemarketing, imbalance, prediction, classification**

1. INTRODUCTION

While banks do not need the deposits to offer loans, they do need it to balance it by attracting client deposits. Banks can borrow funds from the government treasury at very minimal interest rate, but it's still more expensive than borrowing from the bank's own depositors. How does the deposit have an effect on stability in banking has become a crucial issue in the growth of banking institutions. In banking business, the question of attracting the customers for term deposits has been around for some time and has arisen again after the global economic crisis. In response to the crisis, a number of countries considerably extended the coverage in their protection nets on the way to repair marketplace confidence and to ward off potential contagious runs on their banking sectors. Through a marketing campaign about contacting clients on telephone directly, the bank intends to select the favourable set of clients that are highly inclined to deposit. It is beneficial for narrowing the range of potential customers, elevating the rate of success as well as reducing the cost of marketing process. The remaining paper is organized as follows. In Section 2, we review the related literature. We describe the dataset in Section 3 and data pre-processing in Section 4. Traditional machine learning approaches are discussed in Section 5. This is followed by implementation and their results. Finally, we conclude in Section 6.

2. LITERATURE REVIEW

We surveyed papers relevant to bank telemarketing campaigns based on machine learning approaches and identified four distinct approaches the researchers worked on. The following is an exploration of literature review in each category. The problem of predicting target customers using bank telemarketing has been addressed using neural networks [1]. But since the data used by the author [1] is unbalanced (because 11.07% of customers subscribe to a certificate of deposit) and the methodologies attributed to neural networks, they had to deal with the data in either of the three ways — ignoring the problem, undersampling the majority class, and oversampling the minority class. The disadvantage of undersampling is that we may risk removing the majority class instances which is more representative, thus discarding useful information. The disadvantage of oversampling is that, since it simply adds replicated observations in original dataset, it ends up adding multiple observations of several types, thus leading to overfitting. Logistic Regression tend to be easily understood by humans, have the advantage of fitting models and providing good predictions in classification task [1]. Logistic Regression is suitable for a small amount of feature space, which implies such model is generally ill-fitting with multiple variables. For logistic regression models, unbalanced training data affects only the estimate of model intercept — it skews all the predicted probabilities, which in turn compromises the predictions. Consequently, the authors clearly implant the suggestion for improvement [2]. In another approach focused on sensitivity analysis, the authors have shown with rminer package [3] how a model is influenced by each of its input attributes in percentage of the remaining. The performance of support vector machine (SVM) on the same dataset [1] has also been tested [1] showing better results compared to logistic regression. While SVMs do not perform well on highly skewed/imbalanced data sets, authors have made no comment on class imbalance in the paper. For unbalanced datasets, the separating hyperplane found by the SVM can be biased in a manner to favor predictions of the majority class on the test samples. Better results are portrayed by Random Forest Classifier with an accuracy of 87% and it is the most promising classifier with respect to predictive ability [4]. To handle class imbalance, 'Spread-subsample' was used to

balance the class values. The class values were suitably balanced as 5255 instances being 'No' and 5,255 instances being 'Yes' totaling 10,510 instances. Although the training accuracy of such model is high, accuracy on unseen data will be worse. Another caveat is that Random forests are built on decision trees, and decision trees are sensitive to class imbalance because each tree tends to be biased in the same direction and magnitude (on average) by class imbalance [4]. Gradient Boosting classifier performed slightly worse than other classifiers [5]. Though gradient based methods generally give better results, gradient boosted trees are harder to fit than random forests. Gradient boosting algorithms generally have three parameters which can be fine-tuned — shrinkage parameter, depth of the tree, and the number of trees. The proper training of these parameters is required for good fit, failing which may result in overfitting [5]. Bagging and boosting are ensemble-based meta-learning algorithms that pool decisions from multiple classifiers [5]. They have their caveats. First, the sample size of the training data affects the performance of both bagging and boosting in the ensemble design. Second, boosting was designed for an inherently poor classification model given a large amount of data while bagging tends to be more useful for classification problems with a limited amount of available data. Third, the degree of class imbalance may play a role in affecting the model performance [5]. Rotation Forest is another method for generating classifier ensembles based on feature extraction [6]. Rotation Forest ensemble was executed on a random selection of 33 benchmark datasets from the UCI repository and compared it with Bagging, AdaBoost, and Random Forest. The results were favorable to Rotation Forest to a great extent. The accuracy level that was achieved for the dataset in the first run was in the order of approximately 85%. The literature review shows that there is scope for further improvement of the models by improving further feature engineering and hyperparameter tuning. Since XGBoost is popular in handling class imbalance and existing data is inherently imbalanced, we are encouraged to investigate the results with XGBoost.

3. DATASET

The dataset was downloaded from UCI Machine Learning Repository [1]. The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. It consists of 41,188 customers' data. Total 21 features are recorded for each customer. These features include:

- Customer features - age, job, marital status, education, housing, default and loan
- Phone call features - contact, month, day of the week and phone call duration
- Social and economic factors - employment variation rate, consumer price index, consumer confidence index, 3 month Euribor rate and number of employees
- Other attributes - campaign, pdays, previous and poutcome

4. DATA PREPROCESSING

The dataset is semi-colon separated (;). We converted it into comma separated values (csv). The output variable named "y" is string which has values either "yes" or "no". Therefore, we converted it into Boolean values - True for "yes" and False for "no". 6 out of 21 features have missing values default, education, housing, loan, job, and marital. Missing values have been imputed by using multivariate imputation by chained equation. This method is based on fully conditional specification, where each incomplete variable is imputed by a separate model. "Age" variable is binned into 4 bins — Teens, Young Adults, Adults, and Senior Citizens. Feature binning is a method of turning continuous variables into categorical values. Three features are removed to increase model interpretability. "euribor3m" is removed because it is highly correlated with other variables leading to severe mutli-collinearity. "duration" is removed despite being a strong predictor to make the model realistic because after the end of call, "y" is obviously known. "pdays" is removed because it had constant values. Synthetic Minority Oversampling TEchnique (SMOTE) is applied to modify imbalanced data into balanced distribution. SMOTE generates a random sample of minority class observations using bootstrapping and k-nearest neighbors to shift the classifier learning towards minority bias [7].

5. MODELING, IMPLEMENTATION AND EVALUATION

In data preprocessing, we found that there is class imbalance. Class imbalance is encountered in most of the real world classifications. In some of the cases, class imbalance is not just common, it is expected. To handle class imbalance, different models' researchers have worked on [6] [2] [3] [1] [5], have their own approaches. In the following sub-sections, we explore the theoretical aspects and implementation issues of the models.

A. Logistic Regression Analysis

During analysis of data, we found that customer deposit rate was highly dependent on social and economic factors. We applied logistic regression model for predicting decision based on social and economic factors and customer features.

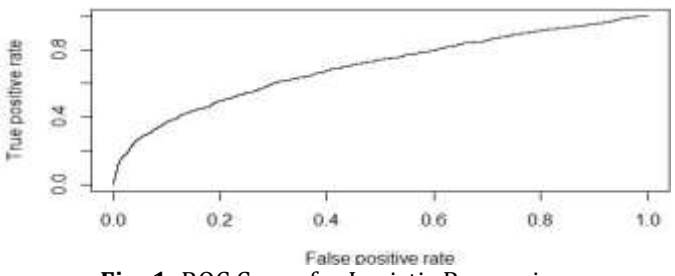

Fig.-1: ROC Curve for Logistic Regression

Optimal cut-off point was found for ROC curve (Fig. 2) to include few false positives because banks can afford to contact few customers who are unlikely to deposit rather than losing potential customers who have proclivity to deposit. The scores above optimal cut-off point are classified as positive, and those below as negative. The area under ROC curve is obtained with optimal cut-off point as 0.6162.

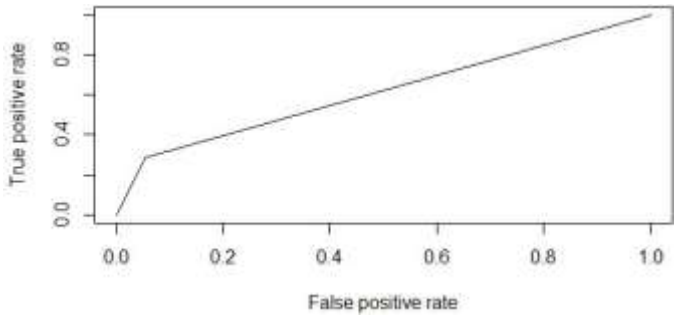

Fig -2: ROC Curve for Logistic Regression with Cut-off point

B. Support Vector Machine

Analysis Support Vector Machine is implemented to check if we could get better accuracy than logistic regression. Since SVM is a computationally demanding algorithm, we used the smaller and stratified version of original dataset [1]. The dataset is preprocessed in the same way as done for logistic regression.

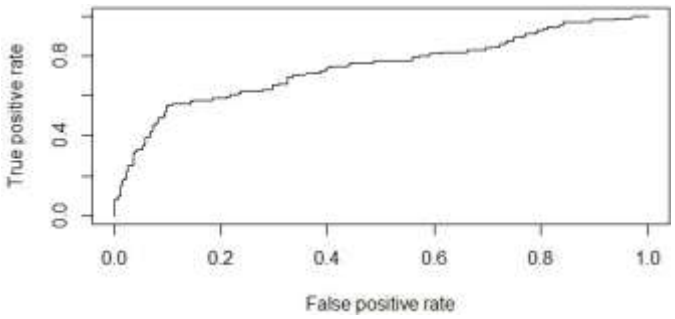

Fig.3: ROC Curve for SVM

Optimal cut-off point was found using cost function. The value of area under ROC curve (Fig. 4) came out to be 0.7182.

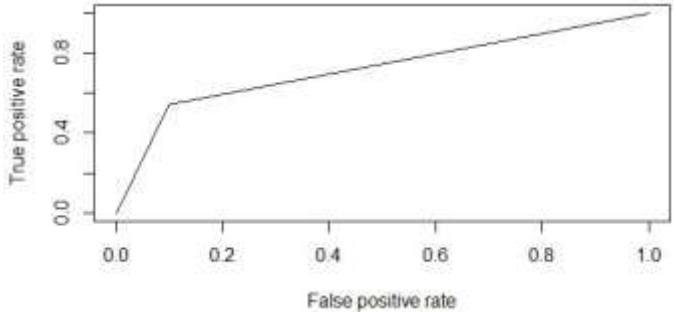

Fig.4: ROC Curve for SVM with Cut-off point

C. Random Forest

Analysis Another approach that we tested is to execute random forest on the dataset. Random forests do not have the issue of overfitting as decision trees, and they are an effective tool for prediction. The area under ROC curve (Fig. 5) is better than that of SVM.

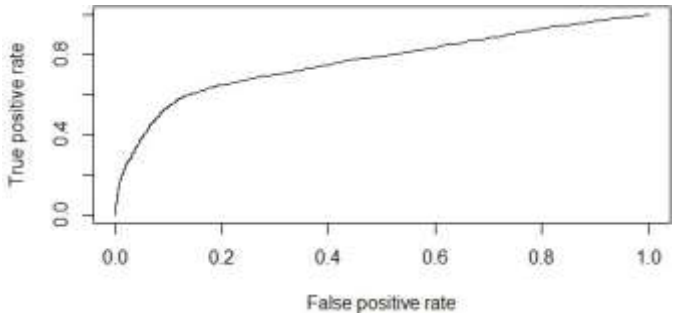

Fig.5: ROC Curve for Random Forest

Optimal cut-off point is obtained by means of cost function. This cut-off point is "optimal" in the sense that it weighs both sensitivity and specificity equally. After cut-off point, the area under ROC curve (Fig. 6) is 0.6989, which is slightly less than that of SVM.

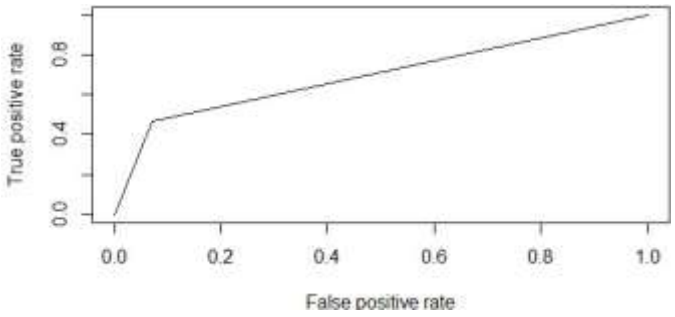

Fig.6: ROC Curve for Random Forest with Cut-off point

D. Extreme Gradient Boosting (XGBoost) Analysis

Unlike gradient boosted machine, where tree pruning stops once a negative loss is encountered [8], XGBoost grows the tree up to max depth and then prune backward until the improvement in loss function is below a threshold. The data preprocessing for XGBoost is little bit different from other models. Age is binned into 4 bins - Teens, Young Adults, Adults, Senior Citizens. Missing values are not required to impute, hence, not imputed. One hot encoding and label encoding is performed on categorical variables and target variables respectively. Unlike other algorithms, XGBoost provides special data structure called DMatrix to create training matrix and test matrix. A dictionary of default parameters is created which needs to be passed for training. Using default parameters and cross validation, the best iteration is found in order to avoid overfitting. The area under ROC curve (Fig. 7) obtained is 0.79.

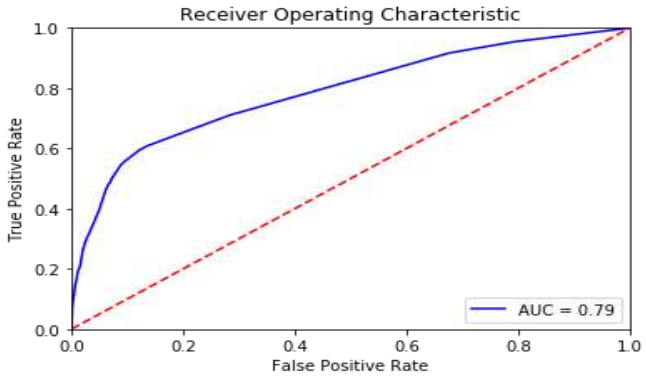

Fig.7: ROC Curve for XGBoost

The optimal cut-off value is found using cost function and scores are classified above it as positive, those below as negative. This optimized model is executed again on test matrix. The new area under ROC curve (Fig. 8) is 0.7368.

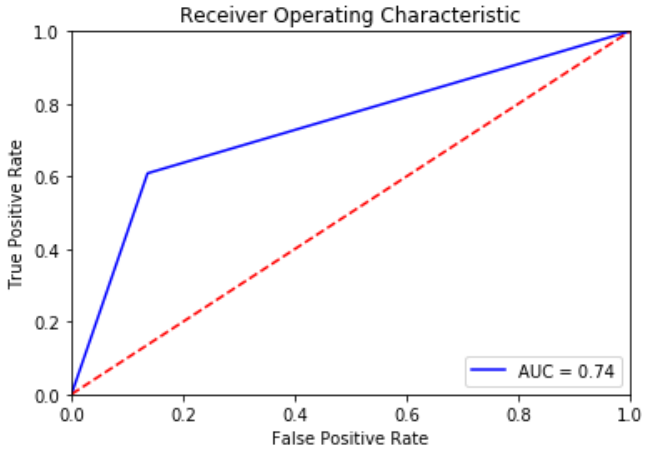

Fig.8: ROC Curve for XGBoost with Cut-off point

E. Results and Discussion

Test accuracy is not a very good indicator of model performance in case of class imbalance. We are more concerned about the customers who say "yes" rather than those who say "no". We decided to choose ROC curve as performance metric. Like precision and recall, accuracy is divided into sensitivity and specificity and models are chosen based on the balance thresholds of these values. After comparing results of algorithms with three aforementioned performance metrics, we observed that XGBoost is the top performing algorithm. XGBoost gives better performance than all other algorithms as is shown by Area under curve (AUC).

	Logistic Regression	Random Forest	Support Vector Machine	XGBoost
AUC	0.6162	0.6989	0.7182	0.7368
F1 Score	0.9295	0.9306	0.9206	0.9291
Test Accuracy	0.8726	0.8771	0.8610	0.8351

Table -1: Summary of Results

We have also predicted realistic results with the exclusion of "duration" variable whereas other researchers have clearly ignored it despite of it being a very good indicator. According to Hosmer and Lemeshow goodness of fit (GOF) test, logistic regression model doesn't fit the data well. The results of Hosmer and Lemeshow goodness of fit (GOF) test are: X-squared = 219.7, df = 8, p-value < 2.2e-16 The p-value is very low (<0.05), so logistic regression model should be rejected.

6. CONCLUSION

The problem of identifying best performing predictive model that predicts potential clients for term deposits have been tackled. With the help of critical literature review and data preprocessing, we characterized the dataset and accordingly implemented the different predictive models and compared their results with AUC of ROC as performance metric. The challenge was to tackle class imbalance and severe multicollinearity in the dataset. We found that XGBoost (AUC = 0.7368) performs better than rest of the algorithms. Though AUC of ROC curve is not very high, but the predictions are reliable owing to the novelty of data preprocessing. The value of AUC can be improved with a larger dataset having more instances of minority class.

REFERENCES

[1] Sergio Moro, Paulo Cortez, Paulo Rita, "A data-driven approach to predict the success of bank telemarketing", Elsevier, June 2014.

[2] Yiyan Jiang, "Using Logistic Regression Model to Predict the Success of Bank Telemarketing", International Journal on Data Science and Technology, June 21, 2018.

[3] Sergio Moro Paulo Cortez Raul M. S. Laureano, 2013. "A data mining approach for bank telemarketing using the rminer package and r tool", Working Papers Series 2 13-06, ISCTE-IUL, Business Research Unit (BRU-IUL).

[4] Justice Asare-Frempong, Manoj Jaybalan, "Predicting Customer Response to Bank Direct Telemarketing Campaign", 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T), 18-20 Sept. 2017

[5] Y. Pan, Z. Tang, "Ensemble methods in bank direct marketing", 2014 11th International Conference on Service Systems and Service Management (ICSSSM), 25-27 June 2014.

[6] J.J. Rodrguez, L.I. Kuncheva, C.J. Alonso, "Rotation Forest: A New Classifier Ensemble Method", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 28, Issue 10, October 2006.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Oversampling Technique", Journal of Artificial Intelligence Research, Volume 16, June 2002.

[8] J. Friedman. "Greedy function approximation: a gradient boosting machine. Annals of Statistics", 29(5):1189–1232, 2001.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.