temp

by Anshu Singh

Submission date: 07-Dec-2022 08:31PM (UTC+0530)
Submission ID: 1931178625
File name: Report_-_Data_Mining_on_Bank_Marketing.docx (148.27K)
Word count: 3146
Character count: 16441

Data Mining on Bank Marketing

First Author[#], Second Author[*], Third Author[#]

[#]*First-Third Department, First-Third University*
Address
[1]first.author@first-third.edu
[3]third.author@first-third.edu

[*]*Second Company*
Address Including Country Name
[2]second.author@second.com

Abstract— **A bank is a type of company that deals with saving, money circulation, deposits, and other services. Banks offer a wide range of services, which vary according to each bank's capabilities. The more capable and superior the bank, the more services it will provide. Direct product introduction is a popular practise in several businesses, including the banking industry. Banks can undertake market study by employing the information technology area to aid in decision making when directly offering goods. It is possible to decide the sort of marketing to undertake by reviewing bank marketing data. Marketing campaigns can be carried out by email, phone, and direct email to prospective consumers, allowing them to select whether to purchase the goods supplied. The volume of incoming data continues to increase as time passes. With this growing data, one bank found it impossible to forecast whether its consumers would sign up for a term deposit. As a result, in this work, the data mining process will be carried out utilising classification methods (Gradient boost classifier, Random Forest, XGB Classifier, and Logistic regression classifier) to forecast if the client would subscribe to a term deposits.**

Keywords— **Data mining, Bank Marketing, Term deposit, Random Forest, XGB, Logistic regression.**

I. Introduction

In our study, we examined Bank Marketing Data Set data from the UCI Machine Learning Repository. The information is connected to a Portuguese financial institution's direct marketing efforts. Phone calls were used in the marketing activities. More than one contact with the same consumer was frequently necessary to determine if the product (bank term deposit) will be subscribed to ('yes') or not ('no'). We choose the bank-full.csv data collection, which contains all samples from the previous edition of this data set. In the first section, we will look at data description and visualisation, and in the second, we will look at data categorization models.

The practise of retrieving previously undiscovered information from a huge dataset is known as data mining. Data mining is being employed in a variety of industries, including finance and banking. Data mining may be used by the bank's marketing department to evaluate customer information and create statistical profiles of individual customer preferences for product and service. Several data mining approaches may be utilised to classify marketing services in the bank direct marketing industry.

To classify the bank client's data, exploratory data analysis on variables will be used to uncover the relationship between the variables and the class variable, the relationship between two variables according to the class variable, and data mining techniques classification. The categorization purpose is to anticipate whether the client will sign up for a term deposit (variable outcome). Classification is the process of identifying a model (or function) that explains and separates data classes or ideas to use the model to predict the class of unknown items. The model developed is based on an examination of a set of training data (data objects whose class label is known).

II. Problem and Data Set(s)

The information is connected to a Portuguese financial institution's direct marketing efforts.

Phone calls were used in the marketing activities. More than one contact with the same consumer was frequently necessary to determine if the product (bank term deposit) will be subscribed to ('yes') or not ('no'). There are four datasets available:
1) bank-additional-full.csv contains all cases (41188) and 20 inputs arranged by date (May 2008 to November 2010), which is quite like the data evaluated in [Moro et al., 2014].
2) bank-additional.csv containing 10% of the cases (4119), chosen at random from 1), and 20 inputs.
3) a full bank csv file containing all samples and 17 inputs, sorted by date (older version of this dataset with less inputs). Here our problem is related to a Portuguese banking institution and we are going to analyse how direct marketing campaigns of a Portuguese banking institution are supposed to be and how it is now the data set is taken from the UCI machine learning repository and it has 21 columns and 41000 records of information so I spot of this analysis we are going to explore the data and understanding customers base and finding insights and make recommendations for improved marketing performance and bill comma evaluate machine learning models for successful prediction of customer subscription.

III. Methods

Here to do the analysis we are selecting different machine learning algorithms: logistic regression random forest gradient boosting and extreme gradient boosting classifiers.

XGboost is also known as extreme gradient boosting, is it distributed gradient-boosted decision tree machine learning library which provides parallel boosting for the trees, and it is the leading machine learning library that is available in the market for doing classification tasks and solving ranking problems. It is suitable to solve larger data sets and it sequentially built shallow decision trees to provide accurate results and highly scalable training methods that avoid overfitting also this method is more accurate than other algorithms like SVM and random forest and the gradient of the data is considered for each tree so the calculation is faster and precision is accurate than random forest this makes users depend on forest algorithm also the sexy boost is more complex in terms of model development and in understanding as well. Basically bursting is an ensemble modelling technique that tries to build a strong classifier from the given number of weak classifiers that already exist and it will update wait for the weak classifiers for a number of iterations this procedure will be continued and all models are added until either they complete the training data set predicted correctly or a maximum number of models that added and an important fact in XG boost is the two trees try to complement each other as the prediction scores of each individual decision tree will be some of to get great results. Also, this module is written in c++ and it helps ml model algorithms by training a gradient boosting the reasons for getting more fame for the sexy boost are execution and speed and centre calculation is parallelizable as we can build multiple models in parallel and it relieved out flyings are their technical equations as it shows better output on many a benchmark data sets and also it has a wide assortment of tuning boundaries this makes the XG boost method most effective and to meet expectations of each data set.

Gradient boosting is another & technique in which the most popular technology is used to build predictive models for various complex regression and classification tasks. Before discussing about gradient posting it is necessary to discuss about boosting in machine learning basically boosting considers popular and simple modelling techniques to build a strong classified from various weak classifiers so, first will build a primary model which is make on top of training data and will identify errors and in upcoming things we will resolve this Airways and we will introduce more models in the process until we get complete training data said by which the end model will predict correctly as the others being rectified in each iteration the different steps involving in boosting algorithms are first we need to consider a data set that you are having different data point and we have to give equal weight to each data point then why are you doing this input as a weight we have to pass it to model then we need to calculate the data points that were incorrectly and we need to increase the weight for data points which

were wrongly classified and this process need to be repeated until will receive a proper expected outputs from our classification there are different boosting algorithms available those are gradient boost machine learning extreme gradient post machine learning classifier light grade and boost machine learning classified and categorical boosting classifier and this gradient boosting utilizes forward learning ensemble method in machine learning and helps to get a predictive model in form of ensemble method.

Logistic regression:

One of the most used machine learning techniques, logistic regression uses input label data to classify the output label. In essence, logistic regression is employed to address classification issues, and it will use several independent variables together with one dependent variable as the output for training and performing classification. Regression essentially forecasts the results of categorical dependent variables; therefore, the results will be expressed as either yes or no, 0 or 1, or binary output. Therefore, it will return a numerical value and, using the Sigmoid function, transform that value to either 0 or 1, rather than only returning 0 or 1. Except for the final layer, which consists of this probability layer that turns ordinal values into binary values, logistic regression is nearly identical to the linear regression employed in regression procedures. Instead of fitting a regression line, we will fit a sigmoid curve for logistic regression; this sigmoid curve will help us generate probabilities from regression values and explains the likelihood of our problem statement, such as whether the image is a cat or not or the patient is pregnant or not. Logistic regression assumes that the dependent variable must be categorical and that the independent variables must not be multicollinear. The only difference between linear regression and logistic regression is logistic regression uses the concept of predictive modelling like linear regression but it is used to classify samples so it will fall under classification algorithm. The main distinction between logistic regression and linear regression is that the latter employs the idea of predictive modelling, much like the former, while the former uses it to classify data to fall under a classification procedure.

Random Forest classifier:

As we will create multiple decision trees under each algorithm, ran over is essentially the process of combining multiple classifiers to solve a complex problem and improve the performance of the model. Random forest is a popular machine learning algorithm that belongs to the branch of supervised machine learning and can be used for both classification and regression problems in machine learning. The more decision trees we build, the more accurate the classifier will be since random forest is described as a classifier that consists of varying numbers of decision trees that belong to various subsets of the provided data and averages out the predictive accuracy of the data set. We'll receive There are a few presumptions associated with this random forest; they are that each decision tree's predictions will have the fewest core relationships and that there will be some actual values in the data set's future variables so that the random forest can produce accurate results rather than random ones. The output produced by random forest is having high accuracy even for larger data sets, and it maintains stability in the output, which are the reasons for choosing random for s for this problem. The working of random forest algorithm is as follows: first will select random K point for training set and random K point for output.

We'll receive There are a few presumptions associated with this random forest; they are that each decision tree's predictions will have the fewest core relationships and that there will be some actual values in the data set's future variables so that the random forest can produce accurate results rather than random ones. The output produced by random forest is having high accuracy even for larger data sets, and it maintains stability in the output, which are the reasons for choosing random for s for this problem. The working of random forest algorithm is as follows: first will select random K point for training set and random K point for output. We will construct decision trees for the subsets of data that are provided, then select the number of decision trees that should be constructed and repeat the process several times. Finally, we will export the model, and for new data points, we can find the predictions of new data points to the category that it belongs to base on the majority vote that it received. The few advantages of the random forest algorithm are that it

can perform both classification and regression tasks, and it can also be used to perform regression tasks.

IV. Experimental setup

Before proceeding to the training of the models we need to understand what type of data where having and different patterns that exist in our data so as what of that first we checked our target variable and it seems to be more unbalance as very few people out of 41000 only 11% of customers are subscribed for term deposit where as 89% of customers are non-subscribers so if he will model with this data it will be more nevi and model will predict like all customers will not subscribe to a town deposit so we could expect an accuracy of around 90% but we need more recall here instead of accuracy so we need to optimise our models for best accuracy and true positive while minimising falls positive and we need to do data balance in also and in our data where having 11 columns that are string objects and 10 columns that are integer and floor data types for each column wear having 41000 normal entries so there is no missing data available so if we see the age variable majority of bank customers are in the range of age between 21 to 60 so Bank would benefit from increasing market towards individual ages of 17 to 21 or greater than 60 and a majority of customers who are having less than age then 21 have subscribed to bank home deposit and if we look at the second variable which is job the few inside from this column are students are more likely then individuals with other job titles to subscribe down deposit students are 30% likely to subscribe Tom deposit and retirees are 25% likely to subscribe to a term deposit and all other individuals are less than 50% likely to subscribe to a termed deposit and also customers whose marital status is unknown or single or more likely to subscribe for the term deposit then people who are married and the absolute difference in percentage is between divorce married and unknown is likely to be around 0.05 so we can drop these variables and we can consider single for our training and customers for study more or likely to subscribe to a term deposit who studied High school, professional degree and unknown and who are illiterate are extremely low and their not subscribed and apart from above variables we are having some other data related to customer contact details and health related things and his educational status. We analysed them clearly and we move to the data scaling part data scaling part is important to make models understand and perform better when data is normalized so before normalizing waste plated our data into training and testing with help of the train test split method and later we rescaled our data with help of standard scalar method and as our data set is highly imbalance concerning target variable, we are balancing our data set with help of SMOTETomek library which will resemble the data which cause funds to our target variable and dependent data so after balancing we are having 2700020 to records of persons who are not subscribed and same count of persons who are subscribed. Few visualizations can be checked from below figures.

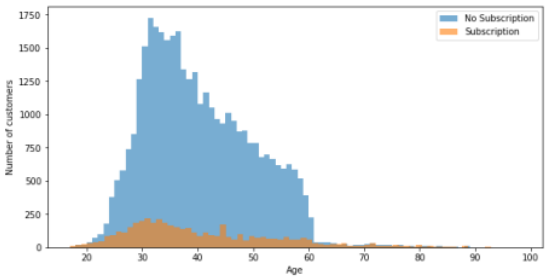

Fig: Frequency distribution of age column

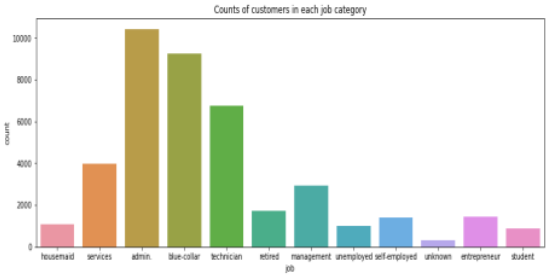

Fig: Bar chart of different jobs that bank customers do.

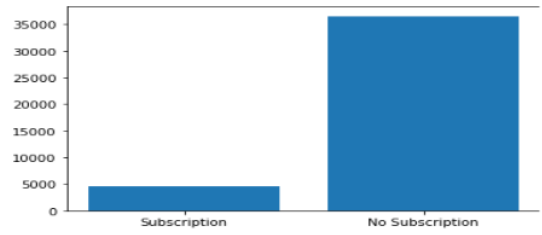

Fig: Bar chart of Target variable which show imbalance in given data.

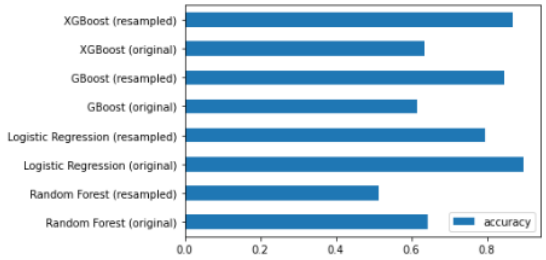

Fig: Different classification algorithms performance graph

V. Results

Before proceeding to result section it is worth to mention that we are applying for different algorithms and those are random for his classified logistic regression classifier gradient boosting classified and extreme gradient boosting classifier so the reasons for selecting these algorithms is our data set is big as it is consist of 40000 + records and the number of columns are also more which were close to 20 + columns and we did two types of analysis here for evaluating model performance that is without sampling our data and with something our data and for each algorithm that we selected we did performance evaluation with help of classification report and we did hyper parameter turning as well so from the results out of all models original logistic regression model is performing good as it's scored 89% accuracy and extreme gradient boosting performing good as it's good 86% accuracy but as our data set is slightly imbalanced so we need to check with recall of these two good performing models for available performance Metrics, so for logistic regression original classifier the recall value is 0.59 whereas for XGB resampled classifier it is 0.73 and normal extreme gradient boosting classifier it is 0.63. so, with help of these recall Metrics, we can select XGB Classifier as a final model as our classification model. The model AUC curves, and accuracy scores can be checked from below figures.

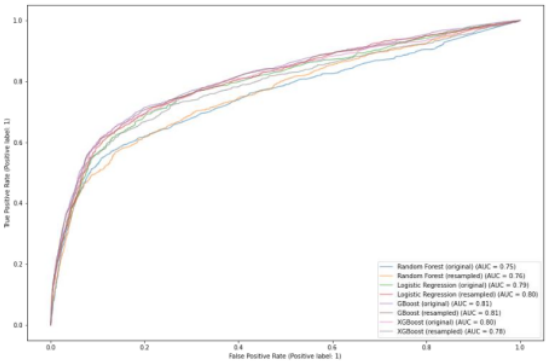

Fig: AUC Curves for different machine learning classifiers.

VI. Conclusions

Bank direct marketing and business decisions are more vital than ever for retaining the greatest customer connection. Customer service and marketing techniques are essential for the business's success and survival. Such marketing methods can benefit from data mining and predictive analytics. Its applications have a significant impact on every industry that contains complicated data and lengthy procedures. It has been shown to lower the number of false positives and false negatives. We were able to analyze the bank marketing dataset; we created several models that assisted us in appropriately analyzing the dataset, and we classified the dataset according to the data preparation description. Banks should use targeted marketing to reach out to new clients based on study findings. This list of clients may then be filtered using the classifier we created in this notebook to get the best results and boost revenues through term deposits with the least amount of overhead and maximum efficiency.

VII. References

[1] H. A.Elsalamony, "Bank Direct Marketing Analysis of Data Mining Techniques," *International Journal of Computer Applications*, vol. 85, no. 7, pp. 12–22, Jan. 2014, doi: 10.5120/14852-3218.

[2] M. C. Keshava, "Predictive Analysis on Bank Marketing Campaign using Machine Learning Algorithms," *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, no. 4, pp. 2664–2668, Apr. 2019, doi: 10.22214/ijraset.2019.4483.

[3] Y. Wu, "Machine Learning Approaches for Retail Bank Marketing Practice," *BCP Business & Management*, vol. 23, pp. 931–937, Aug. 2022, doi: 10.54691/bcpbm.v23i.1475.

[4] "Predicting Loan Approval of Bank Direct Marketing Data Using Ensemble Machine Learning Algorithms," *International Journal of Circuits, Systems and Signal Processing*, vol. 14, Dec. 2020, doi: 10.46300/9106.2020.14.117.

temp

ORIGINALITY REPORT

30%	14%	10%	26%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Coventry University Student Paper	14%
2	Submitted to Rowan University Student Paper	3%
3	Submitted to University of Teesside Student Paper	3%
4	edmontonrealestateinvestmentblog.com Internet Source	1%
5	Submitted to University of Kent at Canterbury Student Paper	1%
6	Submitted to SASTRA University Student Paper	1%
7	www.geeksforgeeks.org Internet Source	1%
8	Youngkeun Choi, Jae Choi. "How does Machine Learning Predict the Success of Bank Telemarketing?", Research Square Platform LLC, 2022 Publication	1%

9	Submitted to UOW Malaysia KDU University College Sdn. Bhd Student Paper	1 %
10	www.educba.com Internet Source	1 %
11	www.coursehero.com Internet Source	1 %
12	medium.com Internet Source	1 %
13	pdfcoffee.com Internet Source	1 %
14	pdfs.semanticscholar.org Internet Source	1 %
15	www.javatpoint.com Internet Source	<1 %
16	nottingham-repository.worktribe.com Internet Source	<1 %
17	www.ijirset.com Internet Source	<1 %
18	"Machine Learning and Data Mining in Pattern Recognition", Springer Science and Business Media LLC, 2018 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		