Journal of International Technology and Information Management

Volume 25 | Issue 4 Article 6

2016

Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction

Olatunji Apampa
Tilburg University

Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction

Olatunji Apampa
Tilburg University
The Netherlands

ABSTRACT

This article attempts to improve the performance of classification algorithms used in the bank customer marketing response prediction of an unnamed Portuguese bank using the Random Forest ensemble.

A thorough exploratory data analysis (EDA) was conducted on the data in order to ascertain the presence of anomalies such as outliers and extreme values. The EDA revealed that the bank data had 45, 211 instances and 17 features, with 11.7% positive responses. This was in addition to the detection of outliers and extreme values.

Classification algorithms used for modelling the bank dataset include; Logistic Regression, Decision Tree, Naïve Bayes and the Random Forest ensemble. These algorithms were applied to both the balanced and original bank data using Orange 3.2 data mining application following the Cross Industry Standard for Data Mining (CRISP-DM), and the ten-fold cross-validation method. Results from the experimental methods revealed that the performance of the Random Forest ensemble improved when the data was balanced. Results also showed that the features duration, poutcome, contact, month and housing were the most important features that contribute to the success of the bank customer marketing campaign for deposit subscription. The study also revealed that the duration of call to clients, response to past promotions, and the use of cell phone contribute positively to the success of the campaign. While the months of September, November, March and April recorded higher subscription rates. Those in management cadre and technicians were found to have responded more positively to the campaign than those in other job categories.

Keywords: *Area Under Curve, classifier, cross-validation, data mining, Decision Tree, direct marketing, ensemble, exploratory data analysis (EDA), Logistic Regression, Naïve Bayes, and principal component analysis (PCA)*

INTRODUCTION

It has become pertinent in recent years for corporate organizations to explore online and direct database marketing as part of the marketing strategy in order increase sales and keep a tab on customers. This is important in order to understand and respond to the immediate needs of the consumer. According to Raorane and Kulkarni (2011), the study of consumers helps firms and organizations to improve their marketing strategies by understanding issues such as the psychology, mindset, behavior and motivation of consumers. It is clear from the last statement that corporate organizations must hold data on customers, and develop capacity to analyze and make good use of such transactional data. One of such transaction that has been automated is customer

  ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

relationship management (CRM). In this study relevant data mining techniques are used to determine which customer and what market segment would respond positively to promotions and marketing campaigns by a corporate organization. The study would rely on consumer data aggregated electronically through customer relationship management (CRM) processes.

According to Tudor, Bara and Botha (2011, p. 266), a customer relationship management system is a bucket on information technology (IT) applications and procedures whose target is to identify the main expectations and preferences of clients, and to effectively use client information to improve relationships between the enterprise and its customers. Electronic CRM (eCRM), which entails wide application of information and communication technologies (ICTs) such as personalized email for instance could be categorized into four dimensions for direct marketing purposes; (i) customer identification, (ii) customer attraction, (iii) customer retention, and (iv) customer development (Kracklauer, Mills & Seifert, 2004).

The motivation behind the study was to improve the level of confidence in the ability to accurately predict the customers and market segment that would respond positively to the marketing campaigns of an unnamed Portuguese bank. In order to improve the predictive capabilities of the developed classification models, ensembles were introduced, particularly the Random Forest ensemble. The Random Forest ensemble is known to improve the classification accuracy and is yet to be applied to the Portuguese bank dataset satisfactorily in previous studies. The study will rely on customers' bank transaction data aggregated electronically through the CRM process. By definition an ensemble is a composite model for classification based on the combination of different classification algorithms. According to Soltys, Jaroszewick and Rzepakowski (2015), ensemble methods are classes of highly successful machine learning algorithms that combines many different models to obtain an ensemble which should be more accurate than its constituent members. The tendency for higher classification accuracy makes the ensemble attractive and important.

DATA MINING AND MARKETING RESPONSE PREDICTION

In data mining there is an extensive use of statistical analysis, mathematical modelling, artificial intelligence and machine learning algorithms. According to Suman, Anuradha and Veena (2012) data mining is the process of extracting hidden patterns from data, analysing the data and summarizing it into useful information. For Tingilidou and Kirkos (2010, p. 2) data mining is the extraction of implicit, previously unknown and potentially useful information from data. Data mining in the views of Han and Kamber (2012) is the process of discovering interesting patterns and knowledge from data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are stream into the system dynamically, in order to detect possible regularities, trends and associations that are not known *a priori* (Acciani, Fucilli, & Sardaro, 2011, p. 27).

BASELINE STUDIES

Moro, Laureano and Cortez (2011), applied data mining for the analysis of direct marketing campaign of an unnamed Portuguese bank using the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. The authors collected real-world data from the bank

marketing campaign which is related with bank deposit subscriptions. In the course of the campaigns which was conducted with customers via telephone, an attractive long term deposit application with good interest rates was offered. The goal of the study was to develop a predictive model capable of improving or increasing the efficiency of directed campaign for long term deposit subscriptions by reducing the number of contacts to do; that is a reduction in the number of customers to be contacted by phone. Using Naïve Bayes (NB), Decision Tree (DT) and Support Vector Machines (SVM) classifiers, Moro *et al* used 29 features and 45,211 instances to model the bank data. Results from analysis after application of classification algorithms revealed that call *duration* is the most important feature, followed by the *month* of contact. They found the SVM to be the most reliable predictive algorithm with an Area Under Curve (AUC) value of 0.938.

Prusty (2013), split the data into two parts for easy computation (40,689, and 4,524 separate instances) while retaining all of its 17 features. He applied the Naive Bayes and Decision Tree (C4.5) algorithms to the dataset. The main thrust of Prusty's study was to compare results obtained when the data was unbalanced with results obtained by using same classification algorithms when the data was balanced with equal selections of "yes" and "no" in the response class (4,763 instances for each). Results showed that the AUC value (0.939) improved after balancing the response class.

Elsalamony (2014) used three statistical measures; classification accuracy, sensitivity and specificity on the bank dataset (17 features and 45,211 instances). He compared and evaluated the classification performance of four different data mining techniques' models; Multilayer Perception Neural Network (MPLNN), Tree Augmented Naïve-Bayes (TAN), Logistic Regression (LR) and C5.0 Decision Tree Classifier. He reported that the C5.0 model achieved slightly better performance than the MLPNN, LR and TAN.

Nachev (2015) applied cross-validation and multiple runs for the partitioning of train and test sets (70% and 30%) for the direct marketing response task. He found out that the two hidden layers architecture proposed by Elsalamony (2014) could be simplified into a single layer structure. He performed a comparative analysis of Neural Networks (NN), Logistic Regression, Naïve Bayes, Linear and Quadratic Discriminant Analysis (QDA) taking into account their performance at different levels of data saturation. Results revealed that the NN is best performer in nearly all levels of saturation with exception of poorly saturated data (10-20%), where QDA showed better characteristics, measured by AUC. There was also a comparative ROC analysis of the models (Nachev, 2015).

All of the aforementioned studies showed that classification algorithms perform better when optimized. The performance results obtained by different authors in recent years when different classification algorithms were optimized for the bank customer marketing prediction task using similar dataset is presented in Table 1. The most common metric for performance evaluation amongst authors is the AUC, but some authors returned the classification error rates as performance metric.

Table 1: State of the art performance for bank marketing response prediction.

Author(s)	Year	Instances	Features	Classification Algorithm	AUC	C.E	Remarks
Nachev	2015	45, 211	17	NN	**0.915**	-	Data saturation, 3-fold cv
Prusty	2013	9, 526	17	C4.5	**0.939**	-	Balanced dataset, test validation
Gupta *et al*	2012	45, 211	17	SVM	-	**0.22**	10 fold cross-validation
Moro *et al*	2011	45, 211	29	SVM	**0.938**	-	1/3 test validation

Note. *AUC is the Area under the ROC Curve, C.E is Classification Error, Features is the number of features/Attributes in dataset, Instances is the number of instances/records in dataset, NN is Neural Network, SVM is Support Vector Machine*

Research Question

The study attempts to use ensemble methods for improving the performance of the classification algorithms for the analysis of the particular bank customer marketing campaign. Given the lack of data mining approaches that employ an ensemble method for solving the marketing response prediction task of the unnamed Portuguese bank, this study would attempt ascertain:

> To what extent does the use of Random Forest ensemble improve/enhance the performance of the Decision Tree classification algorithm for the bank customer marketing response prediction?

APPROACH TO STUDY: CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

The Cross Industry Standard Process for Data Mining (CRISP-DM) developed in 1996 by consortium of experts and analyst in the field was followed throughout in the study. The CRISP-DM has the advantage that it was developed independent of software vendors, and is therefore non-proprietary and freely available standard process for fitting data mining for industries and research purposes. The CRISP-DM entails six critical phases; (i) research understanding phase (ii) data understanding phase (iii) data preparation phase (iv) modelling phase (v) evaluation phase (vi) deployment phase.

Research Phase

In the research understanding phase, we justify the application of three classification algorithms, namely; Logistic Regression, Classification & Regression Trees (CART) and the Naïve Bayes to

the bank dataset by exploring results attained from previous studies as espoused in section 1.2. These classification algorithms were briefly described in order to have a better insight on the research experimental setup.

Logistic Regression Algorithm

Logistic Regression (LR) algorithm is used for predicting variables with finite set of values. In LR the output is in the form of probability distribution with a value less than one. Logistic Regression is based on maximum probability estimation rather than the least squares estimation used in traditional multiple regression analysis, and hence requires more input data for better results.

Decision Tree (CART) Algorithm

Decision Tree is a structure whereby each non-terminal node represents a test or decision on the considered data feature or variable. The Decision Tree (DT) also referred to as Classification and Regression Tree (CART), and is a non-parametric classifier.

Naïve Bayes Algorithm

The Naïve Bayes classification algorithms entails the process of determining classification based on Bayes theorem of posterior probability. It is assumed that the values of each features in the train datasets are independent of one another. Bayes method learns the conditional probability of each attribute given the class label from the training data and then computes the probability of a class value given the particular instance, and predicting the class value with the highest probability.

The Portuguese Bank Dataset

Exploratory data analysis (EDA) falls under the data understanding phase. The Portuguese bank data (*bank-full*) was thoroughly explored using data visualization techniques and physical assessment of the data. Significant time was spent inspecting the dataset physically in tabular format in *R-Studio* and on the *Orange 3.2 canvas*. R Statistical software is an open source application that is available for free online. R is managed by the Comprehensive R Archive Network (CRAN), version 3.2.3 was released in December, 2015. Orange is a data mining application developed by Biolabs with Python scripts. It works by connecting widgets for data manipulation and processing on the canvas, version 3.2, 2015 was used in the study.

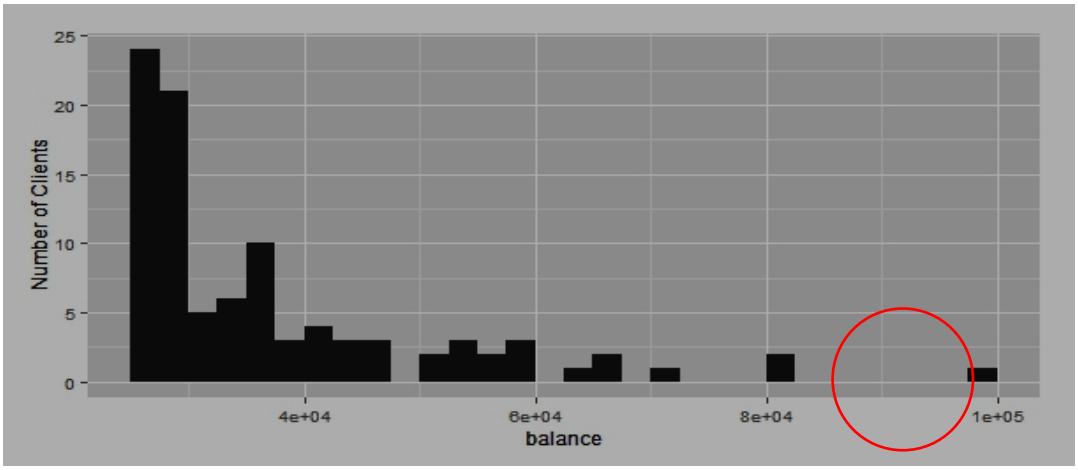
Detailed description of the composition of the Portuguese bank dataset herein referred to as the bank dataset is presented in Table 2. The dataset is available in Microsoft Excel 2013 Spreadsheet in comma separated value format (*bank-full.csv*). The dataset has ***17 features*** and ***45, 211 instances*** with seven numeric or continuous variables, and 10 categorical or discrete variables including the binary response described as y (subscription). Analysis of the data summary in R revealed that the binary response class that is the subscription (y) had 39,922 instances of "no" as response, and 5, 289 instances responded a "yes". Thus only 11.7% of the total number of customer contacted during the marketing campaign responded positively to the promotion. The data is thus unbalanced. The imbalance in the binary response class will grossly affect the experimental results. Hence the need to balance-up the instances in the response class.

Table 2: Feature description for the bank dataset.

	Feature	Description	Type
1	Age	Age of client in years	Numeric/continuous
2	Balance	Client's average annual balance in Euros (€)	Numeric/continuous
3	Day	Client's last contact day	Numeric/continuous
4	Duration	How long it takes to contact the client	Numeric/continuous
5	Campaign	Number of contacts performed for client during the current campaign	Numeric/continuous
6	Pdays	Number of days elapsed since last contact from previous campaign	Numeric/continuous
7	Previous	Number of contacts performed for the client before current campaign	Numeric/continuous
8	Job	Type of job held by client	Categorical/discrete
9	Marital	Client's marital status	Categorical/discrete
10	Education	Client's highest educational qualification	Categorical/discrete
11	Default	Is the client in default of credit facility?	Categorical/discrete
12	Housing	Does the client have housing loan?	Categorical/discrete
13	Loan	Does the client have personal loan?	Categorical/discrete
14	Contact	Client's contact communication type	Categorical/discrete
15	Month	Client's last contact month	Categorical/discrete
16	Poutcome	Outcome of the previous marketing campaign	Categorical/discrete
17	y	Subscription	Categorical/discrete

Source: http://mlr.cs.umass.edu/ml/datasets/Bank+Marketing

Figure 1: Distribution of balance showing outlier of positive bank balance of 100,000 Euros.



The Balanced Portuguese Bank Dataset

As already espoused, the binary response class, y (subscription) had 39,922 instances of "no" as response, and 5, 289 instances of "yes" as response. This is a positive response rate of 11.7% of the total number of customer contacted during the marketing campaign. This gross imbalance and

disproportionate representation of the outcomes in the response class will affect the results of our analysis adversely. Hence we balance-up the dataset in order to have equal proportions of "no" and "yes" in the binary response class. Ordinarily, an equal number of "no" response (5, 289) would have been selected to balance the "yes" response to make a total of 10, 578 instances, but because of the need for a fair and comparable result with our baseline data from Prusty (2013), we reduced the number of instances to 9, 526, that is an equal number of 4, 763 instances of "no" and "yes" in the response class. The reduction in the number of instances of the bank dataset from 45, 211 to 9, 526 notwithstanding, we retained the number of features which is 17. The 9, 526 instances of balanced response outcome were randomly selected and balanced-up manually in Microsoft Excel 2013 Spreadsheet in *comma separated values* (csv) format.

EXPERIMENTAL METHODS

Classification could be described as a two-step process consisting of a learning step in which a classification model is developed; and a classification step in which the model is used to predict class labels for a given data. Classification tasks involve two broad approaches to learning using data mining algorithms, and they are; *Supervised* and *Unsupervised* learning. Both forms of learning were applied in the study, but only the performance of the supervised learning was evaluated and compared in the final analysis.

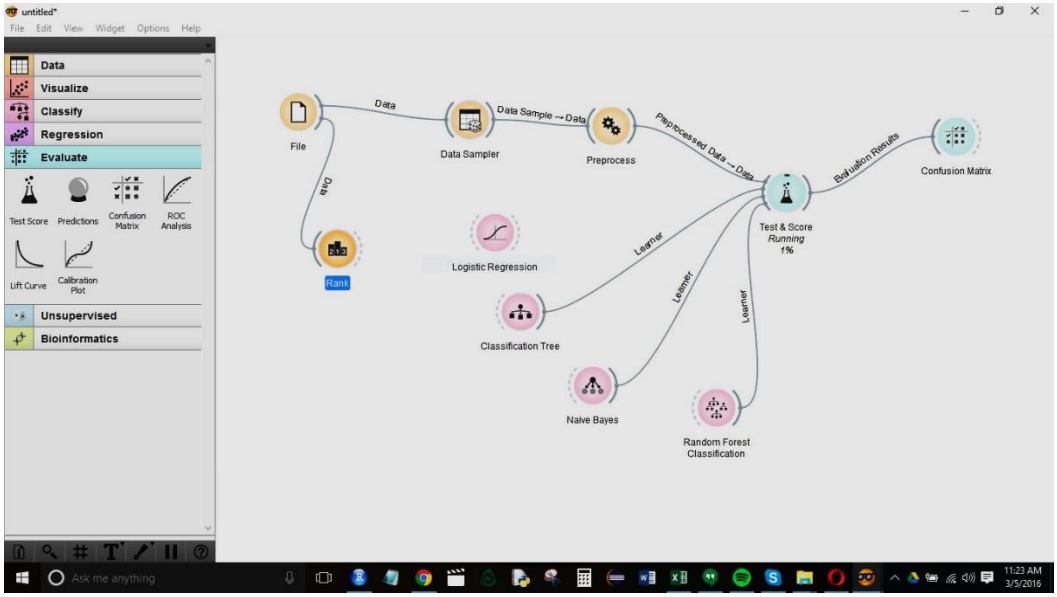
Experiment I: Performance of Random Forest Ensemble on the Bank Dataset

The Random Forest ensemble was applied to the bank dataset (*bank-full*) which has unequal numbers of "no" and "yes" in the response class, and is therefore described herein as unbalanced bank dataset. The Random Forest (RF) is a tree based classification algorithm which combines several decision trees in a logically structured manner, and is often easily interpreted.

In order to directly compare the Random Forest ensemble with other classification algorithms namely; Logistic Regression, Classification and Regression Tree and Naïve Bayes, the experiment was conducted in Orange 3.2 canvas. Using the "Data Sampler" widget, 34% of the bank dataset (*bank-full*), that is the first consecutive ***15, 372*** instances were selected and normalized using the "Normalize Feature" widget before the experiment was conducted. The three classification algorithms and the Random Forest ensemble were cross-validated using the ten-fold cross-validation method in line with the CRISP-DM. In the "Test & Score" widget in Orange, the *"leave one out"* button was checked to confirm the ten-fold cross-validation (figure 3.1). The parameters for the Random Forest were set thus; (i) the number of trees in forest was varied between 50, 100, 200 and 500, depending on the processing capability and memory limitations of the computer (ii) the maximum tree depth was set to 10; and (iii) stop splitting node was set to 2. The "Rank" widget was added and connected directly to the bank data in order to determine the contributions of each data feature to the bank customer marketing campaign.

The selected portion of the data (34%) was considered manageable for experimental purposes, and representative of the dataset. Prusty (2013) argued in favour of reduced number of instances for better performance of classification algorithms.

Figure 2: Work-flow diagram for the evaluation of Random Forest ensemble for the bank dataset.



Experiment II: Performance of Random Forest Ensemble on the Balanced Bank Dataset

For this experiment, the bank dataset was balanced such that there were equal instances of "no" and "yes" (4, 763 instances of each) in the binary response class. Thus, a total of ***9, 526 instances and 17 features*** were used in the experiment. The same parameters applied to the classification algorithms and the Random Forest ensemble in experiment I were applied in this experiment. The Rank widget was also connected directly to the data in order to determine the contributions of each data feature to the success of the bank marketing campaign.

Experiment III: Contribution of the Data Features

It is important to determine the contributory effects of the features on the bank customer marketing campaign. This gives us an idea of the impact of the data features in the overall success of the direct marketing campaign; and the hence most important and most relevant features to consider specifically when planning future marketing campaigns. The balanced bank dataset was used for this experiment.

The most relevant and most important data feature will be determined from two principal methods. Firstly, by using the *Rank* widget in Orange 3.2, the most important data features could be read-off directly from the widget as outlined in sections 3.1, experiment I, and section 3.2, experiment II. Secondly, through the scatter plot we can determine the contributions of each data feature to the success of the bank customer marketing campaign.

Unlike R which gives the coefficients of the features and residuals of each classification algorithms right away, Orange 3.2 only returns the performance metrics for the classification algorithms. In order to determine the contributory effects and importance of each data feature to the success of the bank customer marketing campaign, the CART and Logistic Regression algorithms were modelled separately in Orange 3.2, using same parameters from previous experiments, and using

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

the balanced bank dataset. For the LR, the Ridge = L2, and for strength, C = 1. In the scatter plot widget for the Logistic Regression, jittering was set to 7%, label was set to *"name"*; and size was set to *"same-size"*.

For the CART algorithm, the minimum number of instances in leaves was set to 50. The nodes with less than two instances were not split further, and the depth of trees was limited to 10. In the classification tree viewer widget, the depth of trees was limited to nine in order to have a good view of the decision trees, while the target class was set to "none" in order to have a holistic view of the decision trees.

Experiment IV: Contributions of Categorical Feature Variables

It is equally important to determine the contributions of the categorical features and their variables to the bank customer marketing campaign. For instance, the contributions of demographic variables present in the dataset could help determine which specific categories of clients to target. The binary correspondence analysis between rows and columns (i.e. instances and features) were adopted for the study. Correspondence Analysis is both a descriptive and exploratory technique designed to analyze data in a binary or multiclass problem in which the dataset contained some measure of correspondence between the instances and features. The results provide information that are similar in nature to those produced by Factor Analysis, and they allow exploration of the structure of categorical variables included in the data. Correspondence Analysis returns a scatterplot where instances and/or data features were represented as points in a sequence of low-dimensional spaces. These spaces retain a decreasing amount of the total inertia, with the first dimension capturing the highest amount, while the second will capture the second largest proportion, and so on.

The *Correspondence Analysis* widget in the unsupervised column in Orange 3.2 was connected to the balanced bank data via the File widget. Typically, the Correspondence widget selects only the categorical or non-numeric features, and measures the distance between the feature variables. It equally measures the distances between the data features, but as a rule it is inappropriate to measure the distances between instances and features (i.e. rows and columns). Scaling is an important issue in correspondence analysis because the dataset have different categorical units of measurement. Orange 3.2 automatically scales the coordinates once the correspondence widget is connected to the dataset. Settings for the correspondence analysis were such that the X-axis (horizontal) is component 1, and the Y-axis (vertical) set to component 2. It is better to select features simultaneously with the response class, y, for ease of graphical data interpretation.

RESULTS

Experiment I: Performance of Random Forest Ensemble on the Bank Dataset

Results obtained from the application of the Random Forest (RF) ensemble and the three classification algorithms (LR, CART and NB) to the unbalanced bank dataset using the ten-fold cross-validation method in Orange 3.2 revealed a maximum AUC score of 0.576 for the Random Forest ensemble. This is lower than the AUC value of 0.678 returned by the Decision Tree (CART)

algorithm which the RF was supposed to improve upon. Memory and computational limitations prevented testing of the Random Forest beyond 100 trees in the forest.

Results also showed that the Logistic Regression with an AUC of 0.657 and Classification Accuracy (CA) of 0.898, and the Naïve Bayes algorithm an AUC of 0.627 and CA of 0.885 performed below the CART algorithm which returned an AUC of 0.678 and CA of 0.9.

Precision for the Logistic Regression was found to be 0.651. The Decision Tree (CART) and Naïve Bayes algorithms returned lower values of Precision, they were 0.645 and 0.541 respectively. Precision measure could also be thought of as the percentage of instances labeled as positive that are actually positive.

Recall on the other hand refers to the fraction or percentage of relevant instances that were correctly retrieved by a classification algorithm. Better still, Recall could be thought of as the percentage of positive instances that have been returned by the classification algorithm that are labeled as such. The CART algorithm with a Recall value of 0.384 was found to have a better recall capability than the LR and NB algorithms with Recall values of 0.339 and 0.287 respectively. Findings are presented in Table 3.

Table 3: Performance of Random Forest ensemble and classification algorithms on the bank data.

Classification method	(CA)	(AUC)	F1	Precision	Recall
Random Forest (n = 50)	***0.893***	***0.576***	***0.0***	***0.0***	***0.0***
Random Forest (n = 100)	***0.881***	***0.500***	***0.0***	***0.0***	***0.0***
Logistic Regression	0.898	0.657	0.445	0.651	0.339
Decision Tree (CART)	0.900	0.678	0.482	0.645	0.384
Naïve Bayes	0.885	0.627	0.375	0.541	0.287
Ensemble (RF + LR + CART + NB)	**0.891**	**0.607**	**-**	**0.367**	**0.202**

*Where **n** is the number of trees in the Random forest, CA is Classification Accuracy, and the AUC is Area under Curve*

Experiment II: Performance of Random Forest Ensemble on the Balanced Bank Dataset

Results from experiment II, which entailed the application of Random Forest ensemble and the three classification algorithms LR, CART and NB on the balanced bank dataset revealed that the performance of the Random Forest ensemble improved when the bank data was balanced with equal instances of "no" and "yes" in the response class. The RF returned a maximum AUC of 0.742 when the number of trees in forest was set to 200. Although these AUC values were lower than that of the CART algorithm which was 0.766, it was nonetheless a marked improvement in performance when compared with results obtained from the unbalanced dataset in experiment I. Results are detailed in Table 4.

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

This is an indication that the performance of the Random Forest ensemble increases with an increase in the number of trees in the forest for the balanced dataset. But this situation requires extensive computing capabilities, otherwise it could take several hours or days to obtain results.

Table 4: Performance of the Random Forest ensemble and classification algorithms on balanced dataset.

Classification method	(CA)	(AUC)	F1	Precision	Recall
Random Forest (n = 50)	***0.732***	***0.732***	***0.735***	***0.727***	***0.744***
Random Forest (n = 100)	***0.738***	***0.738***	***0.742***	***0.731***	***0.751***
Random Forest (n = 200)	***0.742***	***0.742***	***0.748***	***0.730***	***0.766***
Logistic Regression	0.757	0.757	0.744	0.787	0.705
Decision Tree (CART)	0.766	0.766	0.769	0.760	0.779
Naïve Bayes	0.756	0.756	0.754	0.761	0.748
Ensemble (RF + LR + CART + NB)	**0.748**	**0.749**	**-**	**0749**	**0.748**

*Where **n** is the number of trees in the Random forest, CA is Classification Accuracy, AUC is Area under Curve*

Results also showed that the CART algorithm with an AUC value of 0.766 performed better than the other classification algorithms. The Logistic Regression and Naïve Bayes algorithms returned AUC values of 0.757 and 0.756 respectively.

For the Precision metric, the LR was able to retrieve more instances that are relevant than the CART and NB. The LR has a Precision value of 0.787, while the CART and NB algorithms returned Precision values of 0.76 and 0.761 respectively.

For the Recall metric, results revealed that the CART algorithm with a Recall value of 0.779 performed better than the LR and NB which returned Recall values of 0.705 and 0.748 respectively.

Thus, a balanced dataset with equal numbers of "no" and "yes" variables in the response class improved the performance metrics for the AUC, CA, Precision and Recall in the study. The classification accuracy and the AUC returned the same performance values for the RF ensemble (0.742, n = 200), LR (0.757), CART (0.766), and NB (0.756) respectively when the dataset was balanced.

Comparison of Results: Balanced and Unbalanced Datasets

Clearly results obtained when the bank data was balanced with equal number of "no" and "yes" in the response class was much better than what obtained with the original data (unbalanced dataset). All of the three classification algorithms returned improved values for the AUC in comparison to what obtained with the original dataset. It is important to recollect that the bank dataset (unbalanced) had 45, 211 instances and 17 features, while the balanced dataset had 9, 526 instances

and 17 features. The balanced bank dataset was randomly selected from the original bank data with 4, 763 equal instances of "no" and "yes" in the binary response class, y. Result obtained for the performance metrics are presented in Table 5.

Table 5: Comparison of the performance of the Random Forest ensemble and classification algorithms on balanced dataset.

Classification method	AUC (Unbalanced)	AUC (Balanced)	Precision (Unbalanced)	Precision (Balanced)	Recall (Unbalanced)	Recall (Balanced)
RF (n = 50)	***0.576***	***0.732***	***0.0***	0.727	***0.0***	***0.744***
RF (n = 100)	***0.500***	***0.738***	***0.0***	0.731	***0.0***	***0.744***
RF (n = 200)	0.742	***0.742***	-	0.730	-	***0.766***
LR	0.657	0.757	0.651	0.787	0.339	0.705
CART	0.678	0.766	0.645	0.760	0.384	0.779
NB	0.627	0.756	0.541	0.761	0.287	0.748

*Where **n** is the number of trees in the Random forest, **AUC** is Area under Curve, **Balanced** refer to the bank dataset (bank-full) with equal numbers of "no" and "yes" in the response class, CA is Classification Accuracy, and **Unbalanced** refers to the original dataset (bank-full) with disproportionate numbers of "no" and "yes" in the response class.*

The Random Forest ensemble had a jump in AUC value from 0.576 to 0.732 when the data was balanced and the numbers of tree in forest set to 50. It got better when the number of trees in forest was increased to 100; the AUC value increased from 0.5 for unbalanced data to 0.738 when the dataset was balanced in the binary response class. With n = 200 the AUC increased to 0.742.

The Logistic Regression had a 15.2% increment in AUC value when the dataset was balanced; that is from 0.657 to 0.757. Also, Precision rates were found to have increased from 0.651 to 0.787. That is the ability of the Logistic Regression algorithm to correctly retrieve relevant number of instances from the bank dataset improved by 20.89%. The Recall rates were also found to have improved significantly when the bank dataset was balanced. The LR algorithm had a jump in Recall value from 0.339 to 0.705. This is 108% increment. Thus, the percentage of relevant instances that were retrieved by the LR algorithm and labelled as such is 70.5%.

The Classification and Regression Tree algorithm also had an increment in AUC value from 0.678 for original data to 0.766 for balanced data. This was a 13% increase in performance metric. The Precision metric also improved from a value of 0.645 to 0.76 when the dataset was balanced up. This is a 17.83% increase. Thus, the CART algorithm was able to correctly retrieve relevant instances when applied to the balanced dataset with a Precision of 76%. The CART algorithm also had a significant improvement in the percentage of relevant instances that it was able to retrieve with a Recall value of 77.9%. This is an increment of 103% for the balanced dataset when compared with the original bank dataset.

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

Lastly, the Naïve Bayes classification algorithm returned an AUC value of 0.756; a jump of 20.5% when compared with the original AUC value of 0.627 for the original data. The Precision was found to increase from 0.541 to 0.761 which is an increment 40.66% in performance metric. Thus, the Naïve Bayes was found to have correctly retrieve relevant instances when applied to the balanced dataset with a Precision of 76.1%. The Naïve Bayes recorded the most significant increment in Recall metric, with an increment of 158.9%, from a Recall value of 0.287 to a value of 0.743 for the balanced dataset.

Generally, metrics such as AUC, *Precision* and *Recall* were also found to improve significantly when the bank dataset was balanced-up in the response class. The application of Random Forest ensemble did not improve the performance of the Decision Tree (CART) algorithm in this study even when the bank dataset was balanced-up.

Contribution of other Features

Experiment III: Contribution of the Data Features

Results from analysis of the classification models for the Logistic Regression and Decision Tree in Orange 3.2 revealed that the feature *duration* is the most relevant and hence most important variable that contributes to the success of the bank marketing campaign. For CART algorithm, the variable *duration* was the root node, an indication of its *purity* and importance. In the Decision Tree, a pure node is indicative of the relevance and importance of a feature to the CART algorithm. The other features selected for computations by the CART algorithm were *poutcome, month* and *contact*.

Similar results were obtained for the Logistic Regression. The outcome of previous similar marketing campaign *poutcome* was found to be the next most relevant feature after *duration*. Table 6 is a summary of the coefficients of the four most relevant features in the LR algorithm.

Table 6: Coefficients of most relevant features for Logistic Regression algorithm.

Features	***duration***	***poutcome***	***month***	***contact***
Coefficients (unbalanced)	1.1	0.40	0.28	0.65
Coefficients (balanced)	1.8	0.35	0.11	0.18

Results from the Rank widget in Orange 3.2 revealed that the features; *duration*, *poutcome, contact, month* and *housing* in that order were the most important data features that contributes to the overall success of the bank customer marketing campaign. The feature *duration* had the highest Gain Ratio and Gini index, while *default* was found to have the lowest.

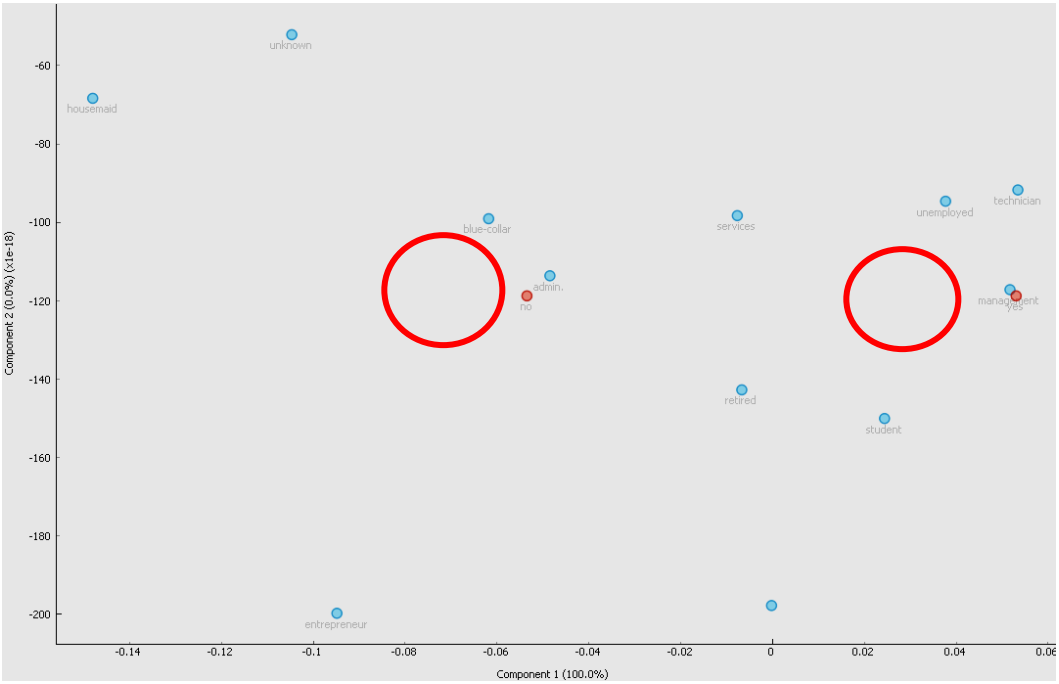
Experiment IV: Contributions of Categorical Feature Variables

Results of the Correspondence analysis revealed that for the data feature *Job*, clients in management and technician cadre corresponded with the "yes" response, hence those in the two cadres responded positively more than others to the direct marketing campaign of the unnamed

Portuguese bank. Those in administration and blue collar jobs were found to have turned down the offer more than any other job cadres. This is detailed in figure 4.1. The horizontal X-axis or component 1 represents the "no" and "yes" variables of the binary response class, y (subscription). While the vertical Y-axis in same Figure 3 represents the type of job or employment held by clients at the time of the campaign. Similarly, when the Y-axis is substituted with another data feature, and the binary response on the X-axis is retained, we can easily observe the relationship between the features variables and the response class. Results could not be determined otherwise because Orange 3.2 does not offer tabular presentation of results (distances between instances and between features) for Correspondence Analysis, hence results were read-off directly from the scatter-plot.

Further interpretation following the procedure described above revealed that singles and divorced responded positively to the campaign more than married clients. Clients with tertiary education were found to have a better subscription rates compared with clients with primary education. Expectedly, those customers without bank loans responded positively more than those with bank loans. Finally the months of September, November, March and April were found to have higher subscription rates than other months during the bank customer marketing campaigns.

Figure 3: Correspondence analysis of variables for the data feature Job in Orange 3.2.



CONCLUSIONS

For the balanced dataset, the Decision Tree (CART) algorithm was found to have performed better than the Logistic Regression (LR) and Naïve Bayes (NB) algorithms having returned an Area under Curve (AUC) and Classification Accuracy (CA) value of 76.6%. The LR and NB algorithms on the other returned AUC and CA vales of 75.7% and 75.6% respectively. The Random Forest (RF) ensemble had an AUC and CA value of 74.2% when the number of tree in the forest (n) was increased to 200. Results from experiment II showed that the performance metrics for the RF

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

ensemble increased with an increase in *"n".* It is obvious from the AUC and CA values that both the RF ensemble and classification algorithms returned same values for the AUC and CA metrics when the bank dataset was balanced.

Precision and Recall were other important metrics considered in this study. Recall values were found to increase significantly (by more than 100%) for the three classification algorithms.

Results were significantly better with the balanced dataset in comparison with the original bank data. Thus in conclusion, the balanced bank dataset improved all performance metrics (AUC, CA, Precision and Recall) for the three classification algorithms and the Random Forest ensemble. Because the Random Forest returned a lower AUC value than that of the Decision Tree (CART) algorithm, *it was concluded that the use of Random Forest ensemble does not improve or enhance the performance of the Decision Tree (CART) algorithm in this study. Thus our confidence level in the predictive abilities of the developed model did not increase according to the study.*

REFERENCES

Acciani, C., Fucilli, V., & Sardaro, R. (2011). Data mining in real estate appraisal: a model tree and multivariate adaptive regression spline approach. *AESTIMUM* ,58, 27 - 45.

Elsalamony, H. A. (2014). Bank direct marketing analysis of data mining techniques.*International Journal of Computer Applications*, 85(7), 12 – 22.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: concepts and techniques. Morgan Kaufman Publishers.

Kracklauer, A. H., Mills, D. Q., & Seifert, D. (2004). Collaborative customer relationship management. Springer Science & Business Media.

Moro, S, Laureano, R and Cortez, P (2011). Using data mining for bank direct marketing: an application of the CRISP_DM methodology. Available at:*https://repositorium.sdum.uminho.pt/bitstream/1822/14838/1/MoroCortezLaureano_DM Approach4DirectMKT.pdf*

Nachev, A. (2015). Application of data mining techniques for direct marketing. *Computational Models for Business and Engineering Domains*. Available at: *http://www.foibg.com/ibs_isc/ibs-30/ibs-30-p09.pdf.*

Prusty, S. (2013). Data mining applications to direct marketing: identifying hot prospects for banking product. Web data mining (ECT 584), Spring. DePaul University, Chicago.

Raorane, A., & Kulkarni, R. V. (2011). Data mining techniques: A source for consumer behaviour analysis. *International Journal of Database Management Systems*. 3(3), 45.

Soltys, M., Jaroszewicz, S., & Rzepakowski, P. (2015). Ensemble methods for uplift modelling. *Data Mining Knowledge Disc*. 9, 1531 – 1559.

Suman, M., Anuradha, T., & Veena, K. M. (2012). Direct marketing with the application of data mining. *International Journal of Engineering Research and Applications (IJERA)*, 2(1), 41 – 43.

Tingilidou, K., & Kirkos, E. (2010). Data mining methodologies and bankruptcy prediction: a state of the art. Eurasia Business and Economic Society Conference 2010, Athens, Greece.

Tudor, A. Bara, A., & Botha, I. (2011). Data mining algorithms and techniques research in CRM systems. *Computational Techniques, Non-Linear Systems and Control*, 265 – 269.

ABOUT THE AUTHOR

Olatunji Apampa
Tilburg University
apamps2000@gmail.com
The Netherlands