Bank Direct Marketing Analysis Using Supervised Learning

Introduction

In today's information age, direct marketing is a dominant strategy of a great number of banks for communicating with their customers due to intensive competitive market environment. To mitigate the rising marketing cost and decline rate, direct marketing collects customers' historical data and then utilizes predictive models to analyze them to select the targets of their marketing campaigns (Sing'oei & Wang, 2013). This report aims to provide a bank "Universal Credit" with suggestions for their marketing campaigns so that Universal Credit would correctly identify prospective client groups who are more likely to respond to its specific service offers. This report investigates the effectiveness of naïve Bayes, decision Tree, support vector machines (SVM) and Random Forest based on prior research papers and exploits historical information from last campaigns into these four methodologies. Then, different evaluation methods with interpretations would be applied to evaluate the performances of these four models. Finally, suggestions will be introduced.

Background

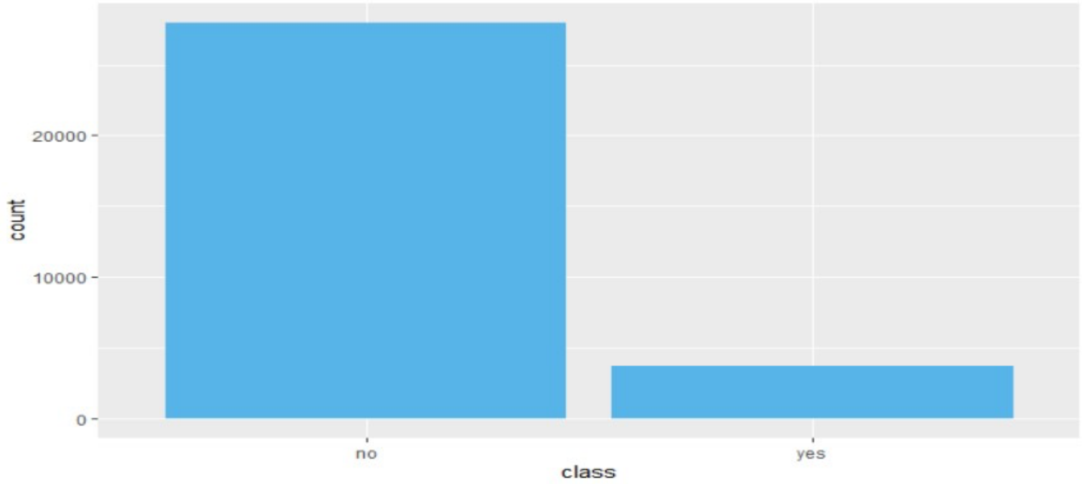

Figure 1. Proportion of Subscribe for training data

A variety of insights can be gained from prior research papers related to direct marketing and data mining. Burez and Van den Poel (2009) addressed the problems of class imbalance, stating that relative lack of data increases the difficulty of detecting regularities within the rare class and proposing that sampling can deal with rarity. This prompts us to consider the proportion of each classification of the data collected. By plotting historical Subscribe Rate (Figure 1), a significant imbalance between "yes" and "no" can be spotted. Due to this, over-sampling and both-sampling would be used to counteract imbalanced original data during data preparation. Crone, Lessmann and Stahlbock (2006) stated that the impact of data preprocessing on predictive accuracy should not be neglected and feature selection can be performed by information gain, assisting identification of the most informative variables within a dataset. Given their perspectives, means of feature selection and instance selection would be used during data preparation.

Elsalamony and Elsayad (2013) proposed that decision tree is a powerful technique of data mining to classify and predict data, because it is comparatively straightforward to understand and explain its procedures which emphasize a great quantity of data complexities, by partitioning data samples into subsets recursively until a homogenous criterion or other stopping criterions are satisfied. For instance, nonlinear and interaction simultaneously occur in real data. These merits convey an inspiration that decision tree should be chosen to optimize the list of targeted customers of this campaign. Nachev and Teodosiev (2014) suggested that SVM can be widely used in classification or regression prediction problems, because it can analyse data by a linear combination of input variables and then categorize the desired outputs by mapping a gap as wide as possible with the help of the data digging and algorithms applying. Naïve Bayes is also an effective and efficient classification algorithm during prediction for certain groups or classes. It can deal with huge database accurately in a short time with simple illustrations. In addition, a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models, which stimulates Random Forest to be included into further model building, as it consists of numerous individual decision trees that operate as an ensemble (Apampa, 2016).

However, it is hard to construct a perfect classification model that would correctly classify all examples from the test set. Therefore, there is a strong need to have more appropriate evaluation metrics to assess the models in this report. The research conducted by Karim and Rahman (2013) recommends different methods to evaluate whether predicting models perform well, such as time to build models, percentages of correct and incorrect classification, and precision, which conveys an aspiration that more evaluation methods should be applied to a further study to better evaluate performances of models from different perspectives.

Business Understanding and Data Understanding

CRISP-DM model is used in this project for data mining, which provides common approaches in mining case study. There is a non-strict sequence of six stages in this methodology. In the first stage, business understanding, focuses on determining project objectives and accomplishments from a business perspective. The data was collected from a marketing campaign run by a Portuguese banking institution. By undertaking basic data mining techniques, the business goal is to increase customers' subscription rates of a term deposit product in reducing the cost and improving returns in the following market campaigns.

	Names	Type	Description
Input attributes	Bank Clients Data		
	age	categorical	
	job	categorical	type of job
	marital	categorical	marital status
	education	categorical	educated levels
	default	binary	whether has credit in default or not
	balance	numeric	average yearly balance, in euros
	housing	binary	whether has housing loan or not
	loan	binary	whether has personal loan or not
	Data related with the last contact of the current campaign		
	contact	categorical	contact communication type
	day	numeric	last contact day of the month
	month	categorical	last contact month of year
	duration	numeric	last contact duration, in seconds
	Other		
	campaign	numeric	number of contacts performed during this campaign and for this client, includes last contact
	pays	numeric	number of days that passed by after the client was last contacted

			from a previous campaign
	previous	numeric	number of contacts performed before this campaign
	poutcome	categorical	outcome of the previous marketing campaign
Output attribute	y	binary	whether has the client subscribed a term deposit or not

Table 1. Attributes of the bank direct marketing campaign (Data Dictionary)

The second stage is data understanding. It aims to acquire data in the project resources. Then, the properties of the acquired data are described, and the quality of the data is verified. The dataset focuses on bank direct market campaigns occurred from May 2008 to November 2010 and finally published in 2011. It contains 45211 records with 16 input attributes and 1 output attribute, as shown in Table 1.

Data Preparation

To prepare data for modelling, the first step is to transform the structure of the variables according to the data understanding stage. After this was completed, we are ready to partition the dataset into training and test sets. The dataset was split into the training set (70%) and test set (30%). The training data will be used to build models and the testing data will be applied to evaluate models. As mentioned in the background, there is a significant imbalance between "yes" and "no" within the output attribute, which results in poor prediction of model. Therefore, both sample methods are implemented to balance the training data. Additionally, the comparison of F1-score will be conducted to verify this method between original training dataset and both sampling training dataset.

Modelling

Four algorithms have been chosen to construct the predictive models: decision Tree, SVM, Naïve Bayes and Random Forest. Decision tree involves creating a set of binary splits on the predictor variables in order to create a tree that can be used to classify new observations into one of two groups. SVMs are a group of supervised machine-learning models that can be used for classification because of the success in developing accurate prediction models. In this report, smallest dataset was used to build SVMs model. Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem, which assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The algorithm for a random forest involves sampling cases and variables to create a number of decision trees. For details of modelling, please refer to the RMD and HTML file attached.

Evaluation

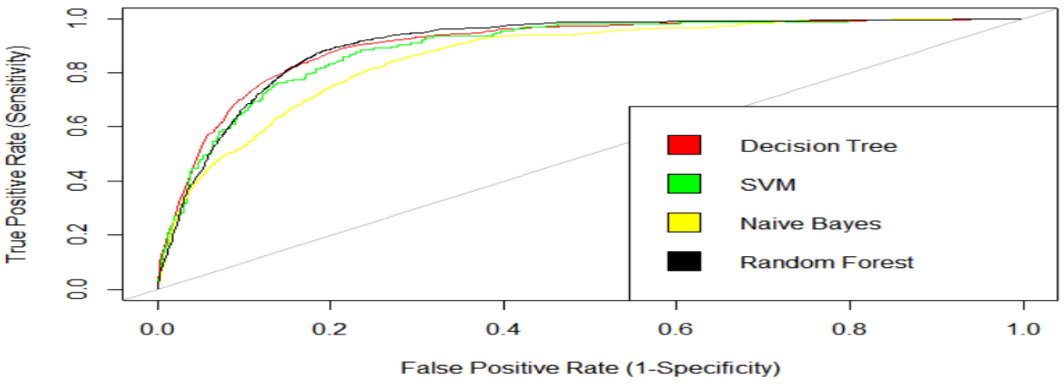

Figure 2. Receiver Operator Characteristic (ROC) for each model

Model	SVM	DT	NB	RF
AUC	0.8924599	0.9041929	0.8597964	0.9046292

Table 2. AUC for each model

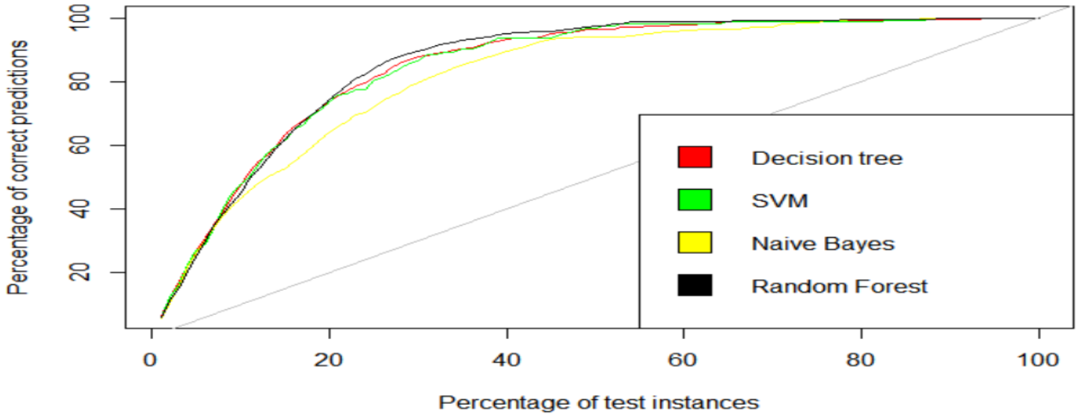

Figure 3. Cumulative Response (Gain) Chart for each model

ROC (Figure 2) and Cumulative Response Chart (Figure3) are applied in this report to initially evaluate the performances of the models we choose. Figure 2 conveys the relationship between costs of direct marketing and corresponding profits. The x-axis of the Figure 2 indicates the costs of the direct marketing and the y-axis means the profits of the direct marketing. Therefore, a better model should be the more left upper going curve, indicating higher profits with lower costs. Additionally, AUC means the area under ROC. Similar with ROC, the better model would have more area under ROC. Figure 3 illustrates the increase in the percentage of correct predictions for each increase in the percentage of data gain. As shown in the Figure 2 and Table 2, it is clear that Naïve Bayes model performs relatively worse than other models. Nevertheless, it is hard to identify which model perform the best through these methods. Therefore, other evaluation metrics would be needed.

Mode Metric	DT (Decision Tree)	SVM	NB (Naïve Bayes)	RF (Random Forest)
Accuracy	0.8188587	0.8443953	0.8208493	0.8338248
Sensitivity	0.8550725	0.7628205	0.6780088	0.8399496
Pos Pred Value	0.3786272	0.4061433	0.3592654	0.3999400
F-measure	0.5248501	0.5300668	0.4696639	0.5418699

Table 3. Metrics values for each model

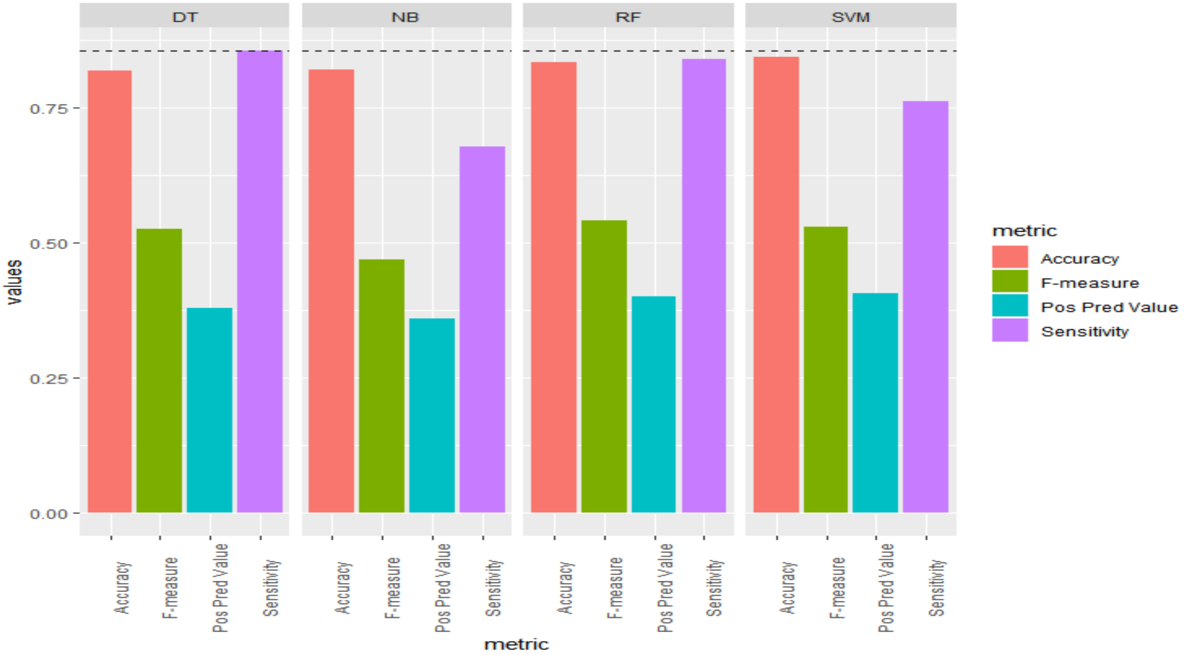

Figure 4. Metrics bar chart for each model

In this situation, metrics Calculated from Confusion Matrix of each model can be analysed for the evaluation. Table 3 shows the actual value of Accuracy, Sensitivity, Pos Pred Value (precision) and F-measure of each model and Figure 4 compares those metrics graphically. Accuracy refers to the percentage of correct prediction for customers’ behaviours. Therefore, higher accuracy means more correct classification of client response. Sensitivity indicates the proportion of the actual "yes" responses are predicted correctly. This ratio needs more consideration, as a better prediction for the actual "yes" responses would lead to higher marketing proficiency by contacting more actual "yes" responders, so as to increase profit. The result of the summary shows that decision tree model has the highest Sensitivity among these four models. Precision shows the proportion of the predicted "yes" responses are actual "yes". This is also a crucial ratio for this bank marketing case, because it is directly related to the marketing efficiency. Higher precision means less errors in predicting "yes" responders, which would lower the unnecessary marketing cost of contacting actual "no" responders. According to the result, SVM has the highest precision. F-measure is the balanced metric of precision and Sensitivity, considering both marketing proficiency and efficiency at the same time. Random forest shows the best result for the F-measure.

According to the entire evaluation, it seems ambiguous to determine which model is the most suitable approach for selecting customers for the deposit marketing. Noticeably, sensitivity method should be considered with more weights. Due to cost-efficiency of direct marketing campaign, the costs of contacting customers will have relatively low effect on the total expected value of the entire marketing than the profits of contacting prospective customers who will be most likely to say “yes” to the deposit offer (Provost & Fawcett, 2013). In other words, it is better to implement greedy direct marketing in order to obtain optimum expected value, focusing more on the profits rather than the costs. Therefore, decision tree model is determined to be the most suitable model. Even though decision tree is slightly outperformed by SVM and RF in Precision and F-measure, it has the highest Sensitivity as shown with the black dashed line in the Figure 4. Therefore, decision tree model will provide more marketing benefits than the other models, which can sufficiently offset the costs of wrong direct marketing and generate more profits.

Conclusion

To improve the outcome of direct marketing campaigns of deposit offers, it is essential for Universal Credit to pay more attention to analyze the data collected in every marketing campaign. This paper applied supervised learning method to explore the dataset of provided by a Portuguese banking institution. In order to plan this data mining project, CRISP-DM methodology is utilized. Four supervised learning methods (decision tree, SVM, naïve Bayes, random forest) were selected to construct predictive models. Decision tree is concluded to be the most suitable model to select the customers for the bank direct marketing, with which it is expected to obtain the optimal returns.

Reference List

Apampa, O. (2016). Evaluation of classification and ensemble algorithms for bank customer marketing response prediction. *Journal of International Technology and Information Management*, *25*(4), 6.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3), 4626-4636.

Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, *173*(3), 781-800.

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

Elsalamony, H. A., & Elsayad, A. M. (2013). Bank direct marketing based on neural network and C5. 0 Models. International Journal of Engineering and Advanced Technology (IJEAT), 2(6).

Karim, M., & Rahman, R. M. (2013). Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing. Journal of Software Engineering and Applications, 6(04), 196.

Nachev, A., & Teodosiev, T. (2015). Using Support Vector Machines for Direct Marketing Models. *International Journal of Engineering and Advanced Technology*, *4*(4).

Olatunji Apampa.(2016).Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction. *Journal of International Technology and Information Management Volume 25, Number 4 2016*

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.".

Sing'oei, L., & Wang, J. (2013). Data mining framework for direct marketing: A case study of bank marketing. *International Journal of Computer Science Issues (IJCSI)*, *10*(2 Part 2), 198.